MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

# Examining the Role of Training Data for Supervised Methods of Automated Record Linkage: Lessons for Best Practice in Economic History

James Feigenbaum
Boston University and NBER

Jonas Helgertz†
University of Minnesota and Lund University

Joseph Price
Brigham Young University and NBER

November 2023

Working Paper No. 2023-03
DOI: https://doi.org/10.18128/MPC2023-03

†Address correspondence to Jonas Helgertz, University of Minnesota, Minnesota Population Center, 50 Willey Hall, 225 19th Ave S., Minneapolis, MN 55455 (email: helgertz@umn.edu). Support for this work was provided by the National Institute on Aging (R01 AG057679) and the Minnesota Population Center at the University of Minnesota (P2C HD041023).

## Abstract

During the past decade, scholars have produced a vast amount of research using linked historical individual-level data, shaping and changing our understanding of the past. This linked data revolution has been powered by methodological and computational advances, partly focused on supervised machine-learning methods that rely on training data. The importance of obtaining high-quality training data for the performance of the record linkage algorithm largely, however, remains unknown. This paper comprehensively examines the role of training data, and---by extension---improves our understanding of best practices in supervised methods of probabilistic record linkage. First, we compare the speed and costs of building training data using different methods. Second, we document high rates of conditional accuracy across the training data sets, rates that are especially high when built with access to more information. Third, we show that data constructed by record linking algorithms learning from different training-data-generation methods do not substantially differ in their accuracy, either overall or across demographic groups, though algorithms tend to perform best when their feature space aligns with the features used to build the training data. Lastly, we introduce errors in the training data and find that the examined record linking algorithms are remarkably capable of making accurate links even working with flawed training data.

## 1. Introduction

Recent advances in machine learning make it possible to link individuals across multiple historical record collections, creating longitudinal datasets that include millions of people over extended periods of time. A number of datasets currently exist, having been generated over the course of several decades and relying on more or less sophisticated methods of record linkage, both manual and automated. During the past decade, several large-scale efforts in the United States have resulted in the linkage of individuals across federal census records (Abramitzky et al. 2021; Helgertz et al. 2022; Price et al. 2021), as well as between federal and i) state census records (Feigenbaum 2018), ii) different vital records (Bailey et al. 2022) and iii) death records (Goldstein et al. 2021; Halpern-Manners et al. 2020). Several of these efforts are based on supervised probabilistic methods of record linkage, where the scholar calibrates an algorithm to learn from an external source of data---typically known as training data---under which circumstances a pair of records to be linked are uniquely similar enough to be considered a match. Training data is generally developed through the manual assessment of record pairs carried out by humans---or hand linkers, thus ultimately affected by factors such as their amount of training, supervision and the amount of information about the record pairs they are tasked with evaluating.

While much progress in the field of record linkage has resulted from aforementioned efforts, the role played by training data in the generation of linked individual-level data using supervised methods of probabilistic linking is largely unexplored. In fact, in many applications, the researcher's access to training data of high or at least sufficient quality is often taken for granted. Intuitively, of course we want to work with high-quality training data; the computing principle of "garbage in, garbage out" known at least since Babbage (1864) certainly applies to record linkage. Even the most sophisticated algorithms could generate data plagued by linking error if calibrated on poor quality training data. Worse, those bad links could possibly yield incorrect or biased empirical findings. Often, scholars take steps to ascertain that the training data is as accurate as possible. But how do we generate high quality training data and what costs should researchers be willing to pay to do so? What are the trade-offs?

Training data is important not only for the calibration and implementation of record linkage algorithms, but also in our assessment of algorithmic performance. Using metrics like recall and precision, we "ask" our training data how accurate the links declared by the algorithm are, as well as about the share of matches in the training data that the algorithm is able to successfully identify. Such statistics are used to compare across efforts and consequently of significant importance for our understanding of the relative success of a set of linked data. Despite the central role played by training data throughout the entire process of implementing a supervised, probabilistic method of record linkage, many assumptions concerning its importance remain poorly investigated.

We address this gap in the literature by comprehensively investigating how choices concerning how to generate training data affect various key characteristics of this data, as well as the extent to which these choices influence the downstream performance of a record linkage algorithm calibrated using this training data. We generate training data between the 1900-1910 and the 1900-1930 U.S. full-count censuses using three different methods, for the first two relying on a team of research assistants to manually match records, resembling methods encountered within various efforts within economic history. For the third set of training data, we exploit crowd-sourced genealogies consisting of links across records beyond the decennial censuses obtained through the Family Search platform, representing our ground truth data. Besides comparing the speed and costs of building training data using the different methods, we also assess differences across the data sets both concerning the share of links that can be obtained as well as their accuracy. The training data sets are used to calibrate and implement two well-documented supervised record linkage algorithms, linking the complete male U.S. population aged 0-50 in 1900 full-count census to 1910 and 1930, investigating the extent to which training data linkage rates and accuracy influences the downstream performance of the record linkage algorithm, again benefiting from data from the Family Search platform to assess out-of-sample linking accuracy. Lastly, we demonstrate the implications of training data errors for standard performance metrics used in record linkage, as well as investigating how the introduction of additional error---presumed to resemble mistakes typically made by a human hand linker---influences the ability of the algorithm to accurately declare matches in the data.

## 2. Generating Training Data

The results of this paper are based on training data sets spanning two different time periods; between 1900-1910 (10 years) and between 1900-1930 (30 years). The shorter 10-year period reflects a typical census interval for many countries and data spanning such an interval is of interest to scholars studying a range of short run processes, including the effects of historical shocks to workers (Feigenbaum and Gross 2022). Shorter 10-year links are also necessary for scholars using multiple observations to deal with measurement errors in census variables like occupation (Ward 2023). The longer 30-year period reflects an interval typically used within research on the intergenerational transmission of socio-economic status (Long and Ferrie 2013), but also used for the examination of any longer-term process, such as the lasting effects of childhood shocks (Choi 2022; French 2022).

In generating the training data for the 1900-1910 and the 1900-1930 periods, we began by selecting two random samples of 1,000 men aged 0-50 from the 1900 census. We proceeded to match the records to their respective universe of potential matches from the 1910 and 1930 census. Potential matches were restricted to individuals sharing the same place of birth, race, year of birth (+/- 3 years) and having a first and last name similarity Jaro-Winkler score above 0.7, a rather standard initial screening used in the recent literature (Feigenbaum 2016). While a record from the 1900 census in theory could be a match for any record in 1910 or 1930, the aim of imposing aforementioned blocking criteria is to limit the universe of potential matches to a set that is both computationally feasible and practical for a human to evaluate, while at the same time maximizing the likelihood that it contains the true match. The blocking criteria causes some of our individuals in the 1900 census sample to not have any potential matches in the other census, ultimately resulting in a sample of 977 and 968 individuals to try and link to 1910 and 1930, respectively.

We proceeded to train a four-person team of research assistants to find matching census records in the data, using two different methods. Both datasets were split into 20 separate files, with each file containing fifty 1900 census individuals and all of their potential matches. Each file was independently linked by two individuals. Two of the hand-linkers were tasked with finding matches in files showing only basic information on the individual to be linked, including names, age, and birthplace (Method I), whereas the other two were provided files containing an extended set of individual, household, and contextual information directly derived from the census records[1] (Method II). Potential matches for each case were presented to the hand-linker in descending order according to a simple similarity score we created[2]. We also implemented a macro adding a timestamp to whenever a hand-linker declared a match, providing information about how long each task took.[3]

Our process guaranteed that every file was linked using both methods, in addition to no file being linked twice by the same individual. Additionally, hand-linkers worked on both types of training data files, both in terms of the linking information they contained and whether the period spanned 10 or 30 years, and in a random sequence. Prior to commencing their work, the team of undergraduates were provided with three one-hour-long training sessions lead by the authors[4], where the focus was to inform the students of how the data was collected and digitized and how to interpret the information they were provided in the files. No feedback on the hand-linkers work was provided throughout its duration. We did this to avoid our

[1] See Figures A1 and A2, Appendix, for examples of the information provided to the hand-linkers in the two different types of hand-linked datasets.

[2] The similarity score was calculated as follows: 0.778 + (-7.5*first name JW distance) + (-14.5*last name JW distance) + (-0.8*year of birth difference), consistent with Feigenbaum (2016).

[3] An infrastructure was set up where each hand-linker was provided an individual Google Drive folder where the training data files were placed as spreadsheet files.

[4] The training took place over Zoom and the recorded sessions are available at bit.ly/linking-training.

views and opinions from influencing the students' performance and also to make our approach more scalable for other groups wanting to use our instruction materials[5].

Key characteristics of the sets of training data are displayed in Table 1, with the 1900-1910 linked data in the first two columns and the 1900-1930 linked data in the final two columns. As Method I and Method II contain the exact same cases---the only difference was the amount of information the hand-linker was provided---some rather striking discrepancies emerge. Requiring the hand-linker to consider more information when distinguishing between potential matches results in a considerably longer average time required per completed file[6]. As a result, having two individuals hand link 1,000 cases requires 1,756 minutes following Method II, a full 54 percent more than Method I (1,140 minutes).

***Table 1: Characteristics of Method I and Method II training data***

| | 1900-1910 | | 1900-1930 | |
|---|---|---|---|---|
| | Method I | Method II | Method I | Method II |
| Cases linked | 977 | 977 | 968 | 968 |
| Avg minutes per file | 28.5 | 43.9 | 26.6 | 37.4 |
| Linkage rate, L1 | 32.6 | 40.4 | 29.3 | 28.5 |
| Linkage rate, L2 | 38.0 | 50.4 | 37.7 | 39.2 |
| Linkage rate, L3 | 28.6 | 37.6 | 26.6 | 24.0 |
| Linkage rate, L4 | 44.3 | 50.3 | 38.6 | 39.5 |
| Share of cases linked by both linkers | 27.4 | 36.4 | 24.1 | 24.3 |
| Conditional agreement rate | 99.3 | 98.9 | 97.9 | 96.2 |
| Links, Method I | 266 | | 228 | |
| Linkage rate, Method I | 27.2 | | 23.6 | |
| Links, Method II | 352 | | 226 | |
| Linkage rate, Method II | 36.0 | | 23.3 | |
| Cases linked by both methods | 198 | | 142 | |
| *% identifying same match* | *98.5* | | *98.6* | |
| Cases uniquely identified by Method I | 68 | | 86 | |
| Cases uniquely identified by Method II | 154 | | 84 | |

[5] This part of our approach is a notable difference from the approach used by LIFE-M, which involved a large amount of feedback to the trainers over the course of the project (Bailey et al 2022)

[6] The time is estimated as the time elapsed between the first and last declared match for the file in question, representing a slight underestimation of the actual time required. To only include actual working time when calculating this quantity, periods of time between declared matches greater than 15 minutes were excluded.

In contrast, when creating the data for 1900-1930, we find that the files were completed more quickly but also with a smaller difference between the methods. The quicker pace was not a result of hand-linkers gaining experience; the linkers worked on batches of 10-year and 30-year links simultaneously. Instead, we suspect a greater prevalence of cases that can straightforwardly be dismissed through the absence of plausible potential matches, either because a greater share of 1900 census individuals would have died or---among the foreign-born---emigrated from the U.S. by 1930 compared to by 1910. As for the smaller difference between Methods I and II, we suspect that may be due to the lower value of the extra contextual information provided by the Method II data when trying to distinguish between potential matches over the longer time period. For example, the share of potential matches representing 1900 and 1930 census individuals who both live in households with parents will be trivial. Hand-linkers, we expect, would realize this and then in many cases ignore the extra data, thus expediting the process.

Table 1 also provides data on each hand linker's respective linkage rate, for which there is considerable variation, both within and across linkers. For 1900-1910 training data, all hand-linkers declare up to a ten-percentage point higher share of links when using more comprehensive information (Method II). Consistent with the presumed limited added value of extended household and contextual characteristics when tracking individuals over a more extended time period, no large difference in the share of linked cases can be observed for the 1900-1930 period.

We consider a pair of records to be confidently linked only when both hand-linkers are in agreement. We believe this to provide us with sets of training data which better highlight the differences across the different methods' abilities to confidently declare matches with a minimum of instruction and supervision. The following two rows in the table illustrate a great degree of similarity in the hand-linkers' *conditional agreement rate*, in no case less than 96 percent and for the 1900-1910 Method I training data over 99 percent. In other words, provided that both hand linkers declared *any* match for a given case and method, they very consistently declared the same match.

Proceeding to investigate how the links compare between Methods I and II, the overall linking rates over the 1900-1910 period differ considerably, with 27.2% linked via Method I and 36.0% linked via Method II. This translates to 32 percent more links in the Method II training data. Since the same population was linked across both methods within a given time period, the added links were made possible through providing the hand linker with additional information. However, the differences between Method I and Method II not only pertain to extra links made in the Method II data. The additional information also enabled the hand linker to eliminate a number of links that were declared in the Method I data. More specifically, only 198 cases (74% of Method I links) were linked in both Method I and Method II data. Yet when matches were made by both Method I and Method II, the conditional agreement rate was very high (98.5 percent). Thus, for about a quarter of the Method I matches, there was information in the Method II data that dissuaded the hand-linkers from declaring a match. It should be emphasized that while in certain cases this may be information that directly rejects the match, it is also possible that a lack of confirmatory evidence from the extended information made the hand-linker reluctant to declare a match.

The overall linkage rates for the 1900-1930 data initially indicate trivial differences between linking methods: 23.6% matched via Method I and 23.3% matched via Method II. The similar link rates, however, hides the fact that over a third of all matches declared by each respective method are unique to that method. Consequently, despite our expectation that the additional information provided to the Method II hand-linkers would be of limited value given the long duration of the links, the contextual information nevertheless prevented 86 matches being declared and promoted the declaration of 84 others. Again, the conditional agreement rate was extremely high, underscoring that when both methods yielded a match it was almost always the same match.

### 3. Evaluating Training Data against the Ground Truth

We complement these data sets with a third set of training data (Method III) which we believe to be as close as possible to the ground truth for historical record linking. We use data from the *Family Tree* obtained from FamilySearch.org, a large free wiki-style genealogical platform. The *Family Tree* allows users to create profiles for deceased individuals and gather to that profile information about life events, historical records, and family connections to other relatives. Historical records are attached to profiles as sources which help confirm information about life events and relationships (Price et al. 2021). For the purpose of this paper, we focus specifically on US census records that are attached as sources to these profiles. When a profile has multiple years of census records attached to a single profile, we can use pairs of these census records as ground truth data. The individuals attaching these census sources have access to additional records and frequently have personal information that allow them to verify the linkages with census records.

FamilySearch also provides various search tools that allow users to identify possible record matches for profiles. These search tools make it possible for users to adjust different search parameters to broaden their search to find even more possible matches that they can evaluate. These search parameters include the use of wild card search terms that can help identify matches even when names have been incorrectly indexed. For the random samples that we employ in this paper, we use both the data that already existed on the *Family Tree* as well as links to census records created by our research assistants using the same tools normal users on familysearch.org would use to find records for their ancestors. The data from the *Family Tree* not only allows us to better understand what the achievable match rate is, but also to identify incorrectly declared matches in the training data generated by our team of hand-linkers, as well as to identify matches they were unable to declare.

*1900-1910*

For the 1,000 men we extracted from the 1900 census, the *Family Tree* provided 581 links to a 1910 census record. However, 129 of the links are to a 1910 census record that lie outside the set of potential matches based on our blocking strategy. The most common reasons that a match on the *Family Tree* fell outside our blocking strategy was because of a low first name Jaro Winkler score or a mismatch on birth year (+/- 3 years) or place of birth. For the Method III training data, the sample contains the 457 matches (46.8 percent linkage rate) that lie within our universe of potential matches.

***Table 2: 1900-1910 training data comparison***

| | | 1900-1910 | |
|---|---|---|---|
| | Method I | Method II | Method III |
| Links | 266 | 352 | 457* |
| *Linkage rate* | *27.2* | *36.0* | *46.8* |
| Link universe covered by FS data | 85.0 | 91.5 | |
| ***Accuracy*** | | | |
| **Overall** | 91.6 | 98.1 | |
| *Lower- Upper bound* | *[77.8 - 92.9]* | *[89.8 - 98.3]* | |
| **Only Method I links (n = 68)** | 80.9 | | |
| *Lower bound* | *38.2* | | |
| **Only Method II links (n = 154)** | | 98.1 | |
| *Lower bound* | | *85.7* | |

With the help of this data, we are able to compute what we argue to be more meaningful measures of training data accuracy, expressed as the share of declared links in the training data that are confirmed by the *Family Tree*. Table 2 indicates that the *Family Tree* covers 85 percent of the links declared in the Method I training data and 91.5% of those declared in the Method II training data. Links declared in both sets of data are highly accurate, ranging from 92% for Method I to 98% for Method II. These percentages represent the *conditional accuracy*, which is why we also present upper and lower bounds, since the universe of links covered by the *Family Tree* may not be representative. The bounds are calculated assuming that all links not covered by the *Family Tree* are either i) accurate (upper bound) or ii) inaccurate (lower bound). The upper bound estimates only change marginally, as only the denominator is affected by the Method I cases that are not covered by the *Family Tree*. This is in contrast to the lower bound, where cases are added to both numerator and denominator. As a result of the adjustment, both datasets emerge as potentially considerably less accurate, with Method I data possibly containing as much as 22% of incorrectly declared matches, compared to 10% for Method II. Lastly, the conditional accuracy rate for the links found only by

Method I is consistent with these links often being ones that would be rejected if the linker had more information (as in Method II or on FamilySearch). The lower bound accuracy estimates for Method I links is 38 percent, compared with 86 percent for Method II, the latter only differing marginally from the overall accuracy estimate.

*1900-1930*

Unsurprisingly, the share of individuals out of the baseline sample of 1,000 males that are linked to a 1930 census record on the *Family Tree* is considerably lower than when the censuses only are separated by a decade. Part of this discrepancy is due to census records separated by a longer time period on average being more likely to differ in key identifying information, making it more difficult for the genealogist to identify the correct matching census record, among other important sources of attrition such as mortality and emigration. The *Family Tree* provides us with 344 confirmed links between 1900-1930, again, however, with a nontrivial share of these links (22 percent) lying outside our standard way of defining the universe of potential matches, resulting in a linkage rate conditional on being within this universe amounting to 27.8 percent. Again, the main characteristics causing a de-facto link not to lie within the universe of potential matches as it is defined here are the first name Jaro Winkler score, along with year and place of birth mismatch.

***Table 3: 1900-1930 training data comparison***

| | | 1900-1930 | |
|---|---|---|---|
| | Method I | Method II | Method III |
| Links | 228 | 226 | 269* |
| *Linkage rate* | *23.6* | *23.3* | *27.8* |
| Link universe covered by FS data | 63.6 | 66.8 | |
| ***Accuracy*** | | | |
| **Overall** | 91.7 | 96.7 | |
| *Lower- Upper bound* | *[58.3 - 94.7]* | *[64.6 - 97.8]* | |
| **Only Method I links (n = 86)** | 89.5 | | |
| *Lower bound* | *40.7* | | |
| **Only Method II links (n = 154)** | | 97.6 | |
| *Lower bound* | | *57.1* | |

Table 3 presents the Method III data, showing that the overall conditional accuracy in both Method I and II training data is similar to training data for the 1900-1910 period, at 92 and 97 percent, respectively. However, the proportion of linked individuals that are on the *Family Tree* is considerably lower, which is why the lower bound accuracy estimates are adjusted downward. This reflects the unlikely possibility that all matches uniquely declared in either the Method I or Method II training data are inaccurate. This is also reflected in the links uniquely declared by only one of the hand-linked training data sets again on average suggesting Method II training data being more accurate.

## 4. Calibrating Machine Learning Algorithms with Different Training Data Methods

We use our six different sets of training data to train two different and well-documented supervised algorithms to link the complete population of men aged 0-50 in the US full count census of 1900 to the 1910 and 1930 censuses. Our aim is to examine the degree to which different supervised algorithms' performance is affected by the training data used for its calibration. While the previous sections indicated important differences across the training data sets---not only in terms of the time required to generate them but also in the number of declared links and their accuracy---we do not yet know what (if any) are the implications of these different characteristics of training data for the links made downstream by a supervised method of record linkage. In this section we detail how we calibrate the algorithms. We focus on two straightforwardly implementable regression based linking algorithms whose optimum performance can be calibrated in an identical manner.

The first algorithm replicates Feigenbaum's (2016) probit machine learning model, training the algorithm to identify matches in the data using a selection of characteristics capturing various aspects of similarity between the 1900 census record and the 1910/1930 census potential match. The characteristics are described in greater detail in Feigenbaum (2016), but all are based on the individual in the census and does not use any information about other household members or the context. The primary motivation for this limited set of features is to focus on characteristics of the individual that should not change over time and to avoid features that might change endogenously. We will refer to this algorithm as the *basic features algorithm*.

The second algorithm is based on the Helgertz et al (2021) method that was used to produce the Multigenerational Longitudinal Panel links published by IPUMS, with a few minor modifications. First, while the original method implements two separate algorithms (linking stages), we limit our exercise to the first linking stage in order to maximize the comparability to the basic features algorithm. Second, while we replicate the first stage algorithm in terms of the linking characteristics used, it is calibrated and implemented using a probit estimator, again in order to resemble the basic features algorithm to the greatest extent possible. This ensures that the linking characteristics used by each respective algorithm is the only fundamental difference between the two, with Helgertz et al (2021) proposing a considerably more extensive set of characteristics intended to capture the likelihood of two records corresponding to the same individual, including characteristics at the household and contextual level. This algorithm is hereafter referred to as the *extended features algorithm*.

For both pairs of census years, we train the algorithms using each of the three training data sets, resulting in six uniquely calibrated and implemented algorithms. We calibrate all algorithms using *hlink*[7] developed at the Institute for Social Research and Data Innovation at the University of Minnesota. The goal of the calibration stage is to use the training data to train the algorithm to avoid declaring incorrect matches while at the same time maximizing the number of accurate matches declared. In practice, matches are determined based on the predicted probability that is assigned to each combination of records in the data, estimated from model parameters and the characteristics of the record pair. The predicted probabilities produce two essential threshold values for each 1900 census individual in the data; $\alpha$, representing the predicted probability for the best possible match to a 1910/1930 census record, and $\beta$, representing the relative superiority of the best match compared to the second-best match. Holding everything else constant, a higher $\alpha$ value implies a greater degree of record similarity, whereas a higher $\beta$ value indicates an increasing degree of uniqueness characterizing the best possible match. Matches are determined based on the $\alpha$ and $\beta$ threshold values that are selected, from which it follows that the higher the selected values of $\alpha$ and $\beta$, the more certain and unique the matches will be, at the risk of declaring an ever-smaller number of matches.

The algorithm's optimum performance is found through a train-test-split procedure, where model parameters are estimated on one half of the data and matches predicted using the other half of the data. For a given set of $\alpha$ and $\beta$ threshold values, *predicted* matches will be compared to *actual* matches in the training data, allowing for the calculation of four key quantities: true positives (negatives) represent matches (non-matches) both declared by the algorithm and in the underlying training data. Conversely, false positives (negatives) are matches (non-matches) declared by the algorithm, whereas the underlying training data indicates that they are non-matches (matches). Using these quantities, we calculate Matthew's Correlation Coefficient (MCC), representing a single metric of the algorithm's overall performance that is particularly well-suited for unbalanced data like ours, ranging from zero to one. To obtain stable MCC values, we use the mean value of MCCs obtained from a ten-fold train-test-split procedure, each time split in a different way. For each split, the estimation, prediction, and calibration step were carried out across a full range of plausible $\alpha$ and $\beta$ values. The results from the calibration stage are presented in the Appendix (Table A1), showing key parameters at the optimum performance for all combinations of algorithm type and training data.

## 5. Evaluating Differences in Linked Samples Based on Different Training Data Methods

[7] github.com/ipums/hlink

In this section, we show how training data matters to the performance of the different supervised algorithms. We first show in subsection 5.1 that linking rates vary dramatically depending on the method to build the training data but that differences in conditional accuracy of those links are relatively small across training data methods. We then show in subsection 5.2 that the stability of conditional accuracy across training data methods also holds for specific groups in the population who may be harder to link.

### 5.1 Linking Accuracy Overall

Following the calibration stage described in the previous section, we link the full population of men aged 0-50 in 1900 to the 1910/1930 complete-count population, using the same blocking criteria as when generating the training data. We use *hlink* along with the optimal threshold values reported in Table A1, thus uniquely calibrated for each combination of algorithm and training data type.

Summary statistics concerning these links are displayed in Table 4. We see that there are large differences both *across* linking algorithms calibrated on the same training data set and *within* linking algorithms calibrated on different training data sets. While differences between the linking algorithms might be of some interest, we focus on the differences within algorithms across training data sources as the other issue has already been examined comprehensively elsewhere (Bailey et al. 2020; Helgertz et al. 2022). As before, we mainly focus on the conditional accuracy of the links, using the *Family Tree* as an external source of validity.[8]

Our first finding from Table 4 is that the match rate (and corresponding size of the final dataset) depends greatly on the training data method. To the extent that the accuracy and quantity of matches (and mismatches) in a certain set of allows the parameters of the algorithm to be estimated with greater precision and better distinguish between the two outcomes in the data, it should increase the number of matches it declares, and/or increase the precision of the matches declared by better weighing the relative importance of the features used by the algorithm. Comparing first training data built with Method I (basic features only) and Method II (extended features), we see larger shares of matches within both algorithms when training on Method II data and linking the 1900 to the 1910 census. Thus, the greater number of matches in the Method II training data, along with its greater precision allows both the extended and the basic features algorithm to declare a considerable larger number of matches. This is further emphasized when combining

[8] While a nontrivial share of the links – between 43 and 69 percent – are covered by the *Family Tree*, it should be emphasized that groups that are more difficult to link are overrepresented among the records not covered. Thus, when using FamilySearch as a source of ground truth, we consequently overestimate the algorithms' performance as the error rate in the universe of uncovered links is likely higher. To approximate the magnitude of this, we randomly extracted 100 cases from the universe of links not covered by the *Family Tree* for each combination of training data and algorithm, presented in Appendix B.

the Method III training data and the extended features algorithm, indicating that the information it provides to the algorithm effectively can be used to declare matches in the data.

***Table 4: Links obtained and linking accuracy***

| Training data | Method I | Method II | Method III |
|---|---|---|---|
| *1900-1910* | | *Basic features algorithm* | |
| Total links | 7,256,432 | 10,111,957 | 9,635,842 |
| Added links (%, compared with Method I column) | - | 39 | 33 |
| Family Tree coverage | 63.5% | 60.7% | 61.1% |
| Family Tree accuracy | 93.1% | 90.9% | 91.4% |
| *1900-1910* | | *Extended features algorithm* | |
| Total links | 8,895,153 | 13,220,247 | 14,146,603 |
| Added links (%, compared with Method I column) | - | 49 | 59 |
| Family Tree coverage | 68.5% | 65.9% | 64.8% |
| Family Tree accuracy | 97.6% | 98.8% | 98.6% |
| *1900-1930* | | *Basic features algorithm* | |
| Total links | 6,233,463 | 6,569,392 | 7,700,836 |
| Added links (%, compared with Method I column) | - | 5 | 24 |
| Family Tree coverage | 46.2% | 46.0% | 43.4% |
| Family Tree accuracy | 95.5% | 95.4% | 93.9% |
| *1900-1930* | | *Extended features algorithm* | |
| Total links | 7,355,297 | 7,881,922 | 8,848,253 |
| Added links (%, compared with Method I column) | - | 7 | 20 |
| Family Tree coverage | 45.3% | 46.2% | 45.4% |
| Family Tree accuracy | 96.2% | 97.5% | 96.6% |

Turning to the 1900-1930 period, while the benefits of using Method III persist (albeit of a lesser magnitude), using either an optimally calibrated basic or extended features algorithm combined with Method II or Method I training data has only a marginal influence on the number of matches declared. This largely reflect differences between the training data types in terms of their respective linkage rates. It would therefore appear that the challenges experienced by the hand linkers in using the additional household and contextual information when generating the Method II training data also affected the subsequent algorithm. Interestingly, similar issues do not appear to affect the algorithms trained on Method III training data, despite the genealogists generating this data also relying on external information that never is used when calibrating the algorithm, however evidently providing the algorithm with information it was able to meaningfully use.

Turning to the conditional accuracy, Table 4 suggests that this tends to be highest for both algorithms when trained on data from the method at most closely follows the features used by the algorithm.

Although these the differences in conditional accuracy are not dramatic, the basic features algorithm is most accurate using Method I and the extended features algorithm is most accurate when using Method II data. Consequently, the higher share of links obtained using Method III training data for the extended features algorithm (or Method II and III training data for the basic features algorithm) seems to come at the price of lower accuracy due to the algorithm's inability to observe the extra information used by human linkers to effectively distinguish between potential matches. In short, giving an algorithm better data but without the features to understand why the data is better may be counterproductive, as it confuses the algorithm. This is true both for the basic features algorithm (it does not observe differences in spouse name or geography that pushed a human linker one way or another) but also true for the extended features algorithm (it does not observe the other non-census records or family history and lore that pushed a FamilySearch user to make or not make the link). Arguably more importantly, these differences in conditional accuracy are rather marginal compared to what would be expected based on the calibration statistics, a point we will return to later.

### 5.1 Linking Accuracy across Subgroups

The previous subsection illustrated how all forms of training data yield overall high conditional accuracy, regardless of the algorithms or the time period separating the censuses. The overall accuracy statistic may, however, hide differences between subgroups, with nontrivial consequences for conducting empirical research. For example, it is quite likely that conclusions about intergenerational mobility will be biased upwards for any group with higher rates of linking error, because measurement error from bad links will push the persistence parameter lower. Minorities and migrants are especially likely candidates for inaccurate links due to more frequent under-enumeration, poor enumeration, higher mortality, name changes, age heaping, and more.

In Table 5, we examine heterogeneity within our linking, focusing on i) race, ii) interstate migrant states between the two censuses, and iii) U.S. versus foreign born individuals. As before, we rely on the *Family Tree* to adjudicate the conditional accuracy of the declared links. As expected, conditional accuracy is consistently lower for African Americans, inter-state migrants, and the foreign born.

***Table 5: Linking accuracy, by selected subgroups***

| | Basic features algorithm | | | Extended features algorithm | | |
|---|---|---|---|---|---|---|
| | Method I | Method II | Method III | Method I | Method II | Method III |
| | *1900-1910* | | | | | |
| Race: | | | | | | |
| White | 93.4 | 91.0 | 91.5 | 97.7 | 98.9 | 98.6 |
| Black | 87.5 | 87.3 | 87.6 | 95.1 | 97.1 | 98.4 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 1900-1900 interstate stayer | 95.3 | 93.8 | 94.1 | 98.3 | 98.9 | 98.7 |
| 1900-1910 interstate migrant | 78.4 | 72.2 | 73.4 | 91.9 | 97.3 | 97.2 |
| | | | | | | |
| U.S. born | 93.4 | 91.1 | 91.6 | 97.7 | 98.8 | 98.6 |
| Foreign born | 88.1 | 86.1 | 86.7 | 96.8 | 98.5 | 98.6 |
| *1900-1930* | | | | | | |
| Race: | | | | | | |
| White | 95.5 | 95.5 | 94.0 | 96.2 | 97.5 | 96.6 |
| Black | 91.4 | 91.2 | 89.3 | 92.2 | 95.2 | 97.4 |
| | | | | | | |
| 1900-1930 interstate stayer | 96.8 | 96.7 | 95.7 | 97.0 | 97.8 | 97.1 |
| 1900-1930 interstate migrant | 91.5 | 91.4 | 88.6 | 93.0 | 96.1 | 95.0 |
| | | | | | | |
| U.S. born | 95.6 | 95.5 | 94.1 | 96.2 | 97.5 | 96.6 |
| Foreign born | 92.3 | 92.4 | 90.6 | 94.3 | 96.8 | 97.2 |

But does the proportion of accurate links by group differ across training data methods? Yes, but as we show in Table 5, for the most part *within* an algorithm and subgroup, the differences tend to be small. We also find more evidence suggesting that algorithms perform the best (at least in terms of accuracy) when using training data that was generated while providing the hand linker with the same information that subsequently will be used by the algorithm.

Reading Tables 4 and 5 together, we conclude that while some groups are harder to link, these differences in linking-difficulty are mostly orthogonal to the choice researchers face when deciding how to construct their training data.

## 6. Implications for the use of Standard Performance Metrics in Historical Record Linkage

When scholars describe their record linkage efforts using supervised, probabilistic methods, an emphasis is often placed on performance statistics that are directly derived from the training data used to calibrate and implement the record linkage algorithm. These scholars usually report the overall share of declared matches (often called a linking rate) along with two other common metrics: a version of recall (what share of the matches identified in the training data does the algorithm successfully link?) and a version of precision (what share of the matches made by the algorithm were also made in the training data?). This is hardly surprising, as these statistics have a straightforward interpretation and allow comparison between research projects. But even when these statistics are based on test sets or via cross-fold validation (to avoid problems of over-fitting) caution may be warranted. In short, these statistics are based on (potentially) imperfect training data, as opposed to ground truth. Rather than asking whether the algorithm agrees with ground truth links (or vice versa), our training data-based measures of recall and precision compare the algorithm to the training data. In this section, we explore how much this distinction, between ground truth and training data, matters in the record linkage setting.

Earlier, we demonstrated that what constitutes a match in a given set of training data can differ considerably depending on the method used to generate it. More specifically, a substantial number of record pairs that we know to be *actual* matches on the *Family Tree* (Method III training data) were coded as non-matches according to the two hand linking methods explored in this paper. There is also a nontrivial number of matches declared by the hand-linkers that we know to be inaccurate from the *Family Tree*, though this is rarer.[9] As the calibration of a machine learning model's precision and recall are derived from the underlying training data, key performance metrics may therefore not only be misleading, but also impair the ability to meaningfully compare linked datasets generated based on different training data sets.

An algorithm's out-of-sample performance is estimated based on its ability to accurately predict outcomes in the training data (precision) and the share of matches in the training data that it is able to predict (recall). A standard scenario from a set of hypothetical training data (N=26) is presented in Figure 1. We include the *algorithm's predicted outcome* in column (A) and the *actual outcome observed in the hand linked training data* in column (B). Precision, is calculated as TP/(TP+FP) which amounts to 2/(2+2) or 50%. Recall is calculated as (TP/(TP+FN) which amounts to 2/(2+2), which is also 50%.

While the Method III training data displayed in column C does not represent ground truth data in its strictest sense, there should be no doubt that it is generated through a more comprehensive process that yields both superior accuracy and coverage. By contrasting column (A) with column (C), we therefore obtain a more accurate assessment of how well the algorithm performs than the usual comparison scholars would make between (A) and (B). As highlighted in the *adjusted classification* column (E), the implication is not only that observations that previously were classified as true positives/negatives may be reclassified as false positives/negatives, but also vice versa. The reclassification of observations results in an *adjusted precision* of 75% and an *adjusted recall* of 38%. Note that we limit the calculation of (E) to Census A individuals who are covered by the *Family Tree* data.

Does the same issue characterize Method I and Method II training data? In short, yes. Standard precision metrics based on training data constructed via methods like Method I or II (that is, non-genealogically) are likely to consistently underestimate actual precision and dramatically overestimate recall. In the following paragraphs, we work through the details of this exercise. To calculate the adjusted precision and recall measures associated with both algorithm types and the Method I and II training data

[9] While we think this is rarer (see the very high conditional accuracy rates we have reported throughout), it is possible, especially for certain subgroups.

sets (1900-1910: N=164,265, 1900-1930: N=137,697[10]), we first obtain the predicted outcomes vectors (A)[11].

We begin by calculating *standard* precision and recall measurements by contrasting the predicted outcome (A) with the underlying training data set (B)[12]. Using the Method I training data for 1900-1910 as an example, 266 matches were declared by the hand-linkers, from which it follows that 163,999 (164,265-266) observations in the training data set (B) are non-matches. When paired with the basic features algorithm, the exercise outlined above yields a predicted outcome vector (A) containing 261 declared matches and 164,004 non-matches. 187 of those matches overlap with the training data---classified as true positives---with remaining 74 declared matches classified as false positives, yielding an estimated precision of 0.72, identical to the value obtained from the model calibration stage (Table A1). 79 of the matches in the training data were not declared by the algorithm, thus classified as false negatives, allowing for the estimation of a recall of 0.70, compared to the 0.71 obtained during the model calibration stage.

Proceeding to the calculation of the adjusted measures, we contrast the predicted outcome vector with the Method III---ground truth---training data (C). From earlier, we know that this dataset contains considerably more matches, in addition to knowing that there are individuals for whom there exists a match that is outside the universe of potential matches of the training data. Since the universe covered by the *Family Tree* data is likely not to be fully representative of the entire training data set, we conduct the exercise in two different steps. We first calculate measures of precision and recall again using the (A) and (B) vectors, but limited to the universe covered by (C), referred to as *FT universe precision*. We then proceed to calculate the *adjusted precision*, in the same sample, but by using the (A) and (C) vectors. While we emphasize that the specific values should be interpreted with caution, we also expect that the direction and magnitude of the difference between the various measures provide a meaningful approximation of the degree of over or underestimation of standard performance metrics.

[10] These refer to the full training data sets. Thus, the 977 and 968 individuals in the 1900-1910 and 1900-1930 training data sets are on average linked to 168.1 and 142.2 potential matches, respectively.

[11] These are generated through another feature of *hlink*, through a 100-fold train-test-split procedure at the optimum $\alpha$ and $\beta$ threshold values for each respective combination of training data and algorithm type. For each split, the algorithm is trained on one half of the data, with point estimates subsequently used to classify outcomes in the *entire* training data set[11]. Since the parameter estimates for each of the 100 splits is generated based on a different subset of data, these will also differ slightly between the runs, as will the predicted probability that is assigned to any given observation. Although this is only the case for less than 200 observations of any training data set (<1%), the predicted outcome for any given observation is not necessarily identical across all 100 splits of the data. For outcomes that were consistently classified as a match or a non-match, we consider this to be the outcome that always will result from a given combination of training data, algorithm, and threshold values. For remaining cases, we use the most frequent outcome.

[12] While we do not necessarily expect these measures to perfectly overlap with those obtained during the model calibration stage, we do expect them to be largely similar. This is indeed the case, with estimated values across both Method I and II training data sets, algorithm types, and periods all being within a few percentage points of the ones obtained during the model calibration stage (Table A1).

Again using the Method I training data combined with the basic features algorithm for the 1900-1910 period as an example, the baseline precision and recall measures illustrated in Figure 2 (Method I TD, Basic) are those already reported earlier in this section, 0.72 and 0.70, respectively. Turning to the *FT universe precision* and *recall* measures, both are a few percentage points higher, at 0.76 and 0.75, consistent with it being represented by comparatively more straightforward matches. Finally, we observe a considerably higher *adjusted precision* (0.87), indicating that the matches declared during the training stage accurately identifies matches to a much greater extent than what the standard measure would suggest. More specifically, several of the matches declared are matches according to our ground truth data, despite the algorithm being calibrated using data (Method I) in which some of these matches were not identified. We also see that *adjusted recall* is considerably lower than what both the standard measure and the *FT universe* measure would suggest. Given that the hand linkers failed to declare many ground truth matches in the data, this should come as no surprise. In spite of this, the magnitude is still noteworthy, with an *adjusted recall* of 0.42, suggesting that only about four out of ten ground truth matches in the training data were identified by the algorithm.

The pattern observed for the other combinations of training data and algorithm type is very similar, suggesting that standard precision metrics consistently underestimate actual precision, but dramatically overestimate recall. Systematically, the difference between the *FT universe precision* and the *adjusted precision* measures range between 0.11 percentage points for the basic features algorithm using Method I training data, to almost 0.18 percentage points for the extended features algorithm regardless of training data set. The differences for recall are even larger, up to 33 percent for either algorithm when combined with Method II training data.

Figure 3 shows the results for the exercise on the 1900-1930 data, which display a similar story. The standard measure of precision consistently underestimates true precision, whereas the opposite is true for recall.

## 7. Algorithm Sensitivity to Training Data Error

As scholars construct training data, a natural question to ask is how important is each individual link? Or, how much time and effort should be spent debating whether to consider a given combination of records as a match? As a final exercise, we proceed to investigate how an algorithm's estimated and actual out-of-training data performance is influenced by deliberately introducing error in the training data. We do this by changing a given---but randomly selected---share of matches in the Method I and II training data sets that we know to be accurate according to the ground truth to the second best---but inaccurate---potential match. Thus, the altered training data sets contain the same exact amount of *declared matches*, but with a given share of the matches now changed into an inaccurate link. For all cases changed to an inaccurate match, we select the second-best candidate[13] intending to illustrate a situation where the hand linker mistakenly chooses the incorrect candidate when selecting between two or more quite similar potential matches. For example, this could result in a change from an accurate match between a 1900 "Thomas Farrell" and a 1910 "Thomas Farrell" record to instead selecting an incorrect but---based on name similarity only---perfect 1910 substitute also named "Thomas Farrell". It could also, however, result in changing from an accurate match between two "James Eudey" census records to instead selecting a matching record with the name "James Emery". Consequently, there is reason to believe that the noise introduced by this procedure exceeds what would result from the work of a typical hand linker.

We introduce error to approximately 10, 25 and 50 percent of the ground truth verified matches in each set of training data. For example, the 1900-1910 Method I training data set contains 207 matches that perfectly overlap with the Method III training data, so we randomly select 20, 50 and 102 correctly matched individuals, respectively, and change them from an accurate to an inaccurate match. After altering each set of training data, we repeat the same model calibration procedure as earlier, finding the $\alpha$ and $\beta$ thresholds that maximize the MCC. We again run a 100-fold train-test-split procedure to obtain the predicted matches in the training data set (column A in Figure 1). Similar to the earlier exercise, for over 99 percent of the record pairs in each training data set, the algorithm consistently predicts the outcome as a match or a non-match in all 100 runs. When this is not the case, we again select the most commonly observed outcome, proceeding to use this data to calculate analogous precision and recall measures as in the previous section. Figures 4 and 5 show the results for 1900-1910 and 1900-1930, respectively, for each combination of training data and algorithm. Thus, the statistics for "Method I TD, Basic" in Figure 4 is identical to that already presented in the previous section, with the corresponding recall numbers presented in the right-hand panel.

We focus on the second row of Figure 4 as an example, presenting the basic features algorithm and Method I training data with 10% introduced error. Since the algorithm was optimized for the particular training data used, the baseline precision should be interpreted as the share of declared matches that were deemed as true positives based on training data containing error. We find that the noise introduced to the training data causes the basic features algorithm to predict a lower proportion of matches in the training data (estimated precision falls to 0.64). However, the adjusted precision is virtually unchanged to the exercise using training data without introduced error. Thus, an almost identical share of declared matches are accurate when compared to the ground truth data. Turning to the recall, we again observe that while the standard metric indicates a worse performing algorithm in the presence of 10% induced error, the adjusted

[13] The second-best match was determined using the formula described in footnote 5

recall indicates that the algorithm nevertheless predicts an almost identical share of the ground truth matches.

This quite remarkable finding is further emphasized when we increase the fraction of errors introduced into our training data to 25% or 50%. The impact on our measures of precision based on the training data are dramatic, dropping to a precision of 0.5 when 25% error rate is introduced and 0.4 when a 50% error rate is introduced. However, when we use our adjusted precision measures (comparing our algorithm links to ground truth data instead of intentionally flawed training data) we find virtually no reduction in precision as the amount of error in the training data increases. Similarly, the adjusted recall also remains robust to the increased presence of training data error, with a gradually diminishing discrepancy between standard and adjusted measures.

Turning to the other combinations of algorithm and training data method, we again see that the standard precision measures systematically underestimate the actual precision. Across most methods, the adjusted precision is only slightly lower than the corresponding statistic obtained based on training data without any error. As the amount of error reaches 50%, both algorithms using Method II training data display a declining adjusted precision exceeding 0.1 compared to the baseline; however, this is still consistent with an algorithm performing well beyond what could otherwise be expected.

For recall, a pattern that is overall consistent with that previously demonstrated can again be observed. More specifically, increasing the share of incorrect matches in the training data negatively impacts the algorithm's ability to predict the (increasingly incorrect) matches in the data that was used for its calibration. When instead comparing the predicted matches to the ground truth matches through the adjusted recall, the results indicate that the algorithm predicts a rather consistent share of the ground truth matches regardless of the degree of error. Using the extended features algorithm paired with 50% error Method II training data as an example, the universe overlapping with the *Family Tree* contains 322 training data matches and 457 ground truth matches, respectively. When comparing the matches declared by the algorithm to the training data, 161 are successfully identified, translating to a recall of 0.5. However, when we benchmark them against the ground truth matches, 366 are correctly identified, resulting in an adjusted recall of 0.8.

The results for the 1900-1930 period display a similar pattern, as illustrated in Figure 5. Across all applications and introduced error, the adjusted precision lies around 0.9, which is dramatically above the standard precision measure. For example, the optimally calibrated basic features algorithm based on Method I training data without error yields 134 matches in the universe that overlaps with the *Family Tree* data. Of these, 121 are accurate, translating to an adjusted precision of 0.9. When using the training data containing 50% error, the algorithm predicts 131 matches, of which 118 are accurate, as well as an identical adjusted precision. For comparison, the FT universe precision are 0.81 and 0.42, respectively, based on

109/55 accurate matches when compared to the 0/50% error training data. For recall, we again observe standard measures overestimating the algorithm's performance at no or little training data error, generally reversing at 25 % introduced error or above.

***Table 6: Links obtained and linking accuracy in the presence of training data error***

| | Basic features algorithm | | Extended features algorithm | |
|---|---|---|---|---|
| | Method I | Method II | Method I | Method II |
| **1900-1910** | | | | |
| *10% error introduced in training data* | | | | |
| Total links | 7,686,295 | 9,829,073 | 6,539,192 | 12,628,308 |
| Family Tree coverage | 62.6% | 61.0% | 74.5% | 67.8% |
| Family Tree accuracy | 92.4% | 90.5% | 98.8% | 99.0% |
| *25% error introduced in training data* | | | | |
| Total links | 7,613,731 | 11,165,220 | 9,084,358 | 12,686,277 |
| Family Tree coverage | 62.5% | 58.7% | 68.2% | 67.1% |
| Family Tree accuracy | 92.2% | 88.5% | 97.2% | 98.4% |
| *50% error introduced in training data* | | | | |
| Total links | 7,266,669 | 11,729,666 | 9,890,277 | 12,669,743 |
| Family Tree coverage | 62.7% | 56.5% | 64.1% | 60.3% |
| Family Tree accuracy | 92.6% | 86.9% | 94.9% | 95.4% |
| **1900-1930** | | | | |
| *10% error introduced in training data* | | | | |
| Total links | 6,197,235 | 6,626,951 | 7,780,620 | 8,156,150 |
| Family Tree coverage | 45.9% | 45.9% | 45.1% | 46.1% |
| Family Tree accuracy | 95.0% | 95.4% | 96.2% | 97.4% |
| *25% error introduced in training data* | | | | |
| Total links | 6,354,444 | 7,869,705 | 7,395,830 | 8,475,627 |
| Family Tree coverage | 45.5% | 43.3% | 45.7% | 44.8% |
| Family Tree accuracy | 94.7% | 93.6% | 95.8% | 96.7% |
| *50% error introduced in training data* | | | | |
| Total links | 6,154,331 | 6,970,535 | 7,872,481 | 8,656,806 |
| Family Tree coverage | 45.9% | 44.9% | 43.7% | 43.2% |

| | | | | |
|---|---|---|---|---|
| Family Tree accuracy | 94.8% | 94.3% | 94.9% | 95.8% |

These results suggest that the examined algorithms display quite a remarkable ability to separate signal from noise in training data. This ability is not limited to the small set of records in the training data, but extends to the application of the algorithm on the full census population. To show this, we use the training data containing introduced error to link the full male population aged 0-50 in 1900 to the 1910/1930 complete-count U.S. census, subject to the same blocking criteria as earlier and each training data's optimum $\alpha$ and $\beta$ thresholds. The results are presented in Table 6 and overall confirm the indications provided by the adjusted precision presented earlier. More specifically, the conditional accuracy statistics obtained from overlapping the full count linking runs with the *Family Tree* data resemble the adjusted precision much more closely than the standard precision measurement, with very marginal changes as the share of incorrect matches in the training data increase. Compared to the baseline links (Table 4), the largest reduction in conditional accuracy is observed for the basic features algorithm combined with Method II training data, amounting to 3.9 percentage points, or from 90.9 to 86.9. Indeed, when using Method I training data, the accuracy of the basic features algorithm never drops by more than a percentage point, regardless of time period between the censuses. It is also worth noting that the share of the links on which the accuracy measurement is based is quite consistent within each combination of training data, algorithm and time period, increasing our confidence in this comparison.

## 8. Discussion

Efforts to link various sources of historical data have been fundamental for pioneering scientific advances within the social sciences during the past decade. This has been made possible through innovative methods of automated record linkage, in combination with the digitization of historical source material and improvements in the processing capabilities of computers. In all likelihood, however, we are only witnessing the beginning of this revolution, with countless future efforts to link historical source material bound to reshape our understanding of the past. For the research project which entails generating new data through record linkage, a choice will need to be made between whether to opt for a deterministic or a probabilistic method. Deterministic methods offer several nontrivial advantages, including a more straightforward and direct implementation, also typically yielding a linked dataset in shorter time and with less demands on computational power. Perhaps somewhat ironically, herein lies arguably also the main weakness of the deterministic method; the brute force nature characterizing how matches are identified. We argue that this particularly is the case when the datasets contain a relative abundance of information that could effectively be used by a probabilistic linking algorithm as well as when the datasets are characterized by considerable noise, ranging from nontrivial changes in how names are spelled to the year or state of birth. Indeed, the main advantage of a probabilistic method of record linkage is the ability to better account for nuance in the data due to noise, in addition to better being able to distinguish between potential matches through accounting for a more extended set of information.

For many empirical applications, despite the greater investment in terms of effort, time and financial cost, the implementation of a probabilistic method of record linkage is likely not only to yield larger datasets in terms of sample size, but also greater linking accuracy. Our point of departure is thus the considerable opportunities offered by probabilistic record linkage, but where most methodological research to date has focused on analyzing the output associated with any given method in terms of the linkage rates as well as performance indicators such as precision and recall, in this paper, we have taken a step backwards to carefully examine the role of training data for supervised methods of record linkage.

This study has provided useful insights into the best practice for record linkage within economic history and our findings range from the expected to the unexpected. First, and this was expected, generating training data while providing the human hand linker with more information on the records can have a substantial impact on the characteristics resulting data. This comes at a (time) cost, that we have documented, but this may be a cost that researchers are eager to pay for higher quality links and more links. And though this may be obvious, the type of information provided to hand linkers is important: only information that could be useful in distinguishing between matches will make a difference. For example, showing our hand linkers extra information about households in a 30-year link was of little value. However, our hand linkers recognized this and the marginal time costs of extra useless information were minimal as well. Consequently, the researcher who strives to generate high quality training data should carefully consider how to maximize the amount of *relevant* information they are able to provide to the hand linker on the records they are tasked with evaluating. The resulting training data will be characterized by much higher quality, both in terms of linkage rate and accuracy of the declared links.

We want to make an additional point about the marginal time costs of extra linking features. While it is beyond the scope of this article to assess the impact of the size of the training data set needed to optimize the performance of the algorithm, we know from previous research that both algorithms used in this paper can be trained effectively with small sets of data. Thus, for most applications, the greater amount of time required for a hand linker to generate training data while evaluating a broader set of information should not represent a significant obstacle; in our case with 1000 records to train, moving from Method I to Method II for the 10-year links would “cost” only 13 additional linking hours (twice that for double entry).

What about the costs of genealogical data? That depends greatly on what a researcher can access. Genealogical data, as examined in this paper, remains unsurpassed in terms of quality. This is particularly the case if the researcher is able to access a source similar to the Family Tree, where the data has already been generated and just needs to be collected and formatted. In contrast, the task of generating new data through genealogical methods is at least an order of magnitude more time consuming and costly than less sophisticated methods.

An unexpected insight of this paper is the important consequence of using training data versus ground truth data to calculate standard algorithm performance metrics. Our results show how the standard measurement of training-data-based precision consistently is underestimated because the algorithm can and will successfully declare matches that were not found in the training data but are demonstratively accurate when we examine the ground truth. Conversely, recall estimated based on hand linked training data will tend to overestimate the share of matches that the algorithm is able to declare, again due to the existence of a substantial number of unknown true matches in the data. The difference between the standard and adjusted performance measures is furthermore often nontrivial. Yet, the differences are not necessarily correlated with the quality of the training data, due to how this affects the training of the algorithm. This evidence obviously complicates making comparisons across different linking efforts, as it is impossible to a priori assess the extent of the precision/recall bias. While we have no reason to suspect it being inappropriate to rely on standard performance metrics in the quest to optimize one's own supervised record linkage algorithm, we do caution against using it to compare across efforts.

Another significant conclusion that may be surprising relates to the sometimes quite limited importance of training data quality for the algorithm's ability to accurately declare matches. We see this in two ways. First, the performance of both examined algorithms were only marginally affected by the choice of training data, judging by the conditional accuracy of the links. In fact, our results overall indicate that an algorithm performs best when the information used by the hand linker reflects that subsequently used by the algorithm, even though we know Method I training data is objectively worse than Method II training data. Second, the ability of the supervised record linkage algorithm to effectively exploit training data even when it is filled with errors and noise was demonstrated by our analysis in Section 7. More specifically, even in the presence of considerable error, we observed virtually unchanged conditional accuracy compared to the baseline training data. It is, however, important to recall that the errors we introduced were always the next best possible match. In many cases, these incorrect links were therefore very similar across many characteristics to the correct link. Thus, even when there are inaccurate matches in the training data, the machine-learning algorithm was quite able to appropriately use the provided information to make accurate out-of-sample predictions.

These results are not a license to scholars to produce sloppy training data as fast as possible. Supplementary analyses (not shown) where we instead replace the accurate matches with randomly selected matches (not the second-best choice) crash our estimates of accuracy as they rapidly venture into the realm of "garbage in, garbage out." But, given the marginal differences in downstream accuracy between training data types and the robustness of linking algorithms to trainers selecting only the second-best links, we encourage the prudent scholar to generate training data through a good-faith effort to achieve high quality data, but without excessive agonizing over the specific method used and without the fear that a few poor match choices by research assistants doing linking could spell disaster.

While the choice of training data throughout our analyses only influences linking accuracy at the margin, our results indicate more substantial differences when it comes to the number of links that are obtained. These higher match rates are driven by two factors. First, training data with more links should yield algorithms with more links mechanically; if an algorithm is trained on data where only 1 in every 1000 records links, it will learn that very low rate of linking and apply it, conversely for training data with higher rates of linking. But, second, we also think larger number of (good) matches that were declared in the training data were able to provide the algorithm with extra examples it was able to use to better learn how to discriminate between matches and non-matches.

We conclude with a few caveats of our study. Overall, we caution against any oversimplified extrapolations that motivate quick and dirty approaches to record linkage. There are still a lot of ways record linkage and training data construction can go wrong if researchers are not careful. We also want to emphasize our context. We linked between US censuses in the early 20th century. Many linking applications are also using US censuses (this motivated our choice) but how far beyond the application in this article our conclusions should extend is unclear. Another important caveat pertains to our use of the *Family Tree* data. When constructing measures like precision and recall, have we simply substituted one flawed "truth" (researcher-built training data) for another (FamilySearch genealogical links)? We do not think this is the case, but we also know that the universe covered by this data is characterized by a certain degree of selection. We can therefore not be certain regarding the extent to which to which the observed conditional accuracy reflects the unconditional accuracy.

The revolution in economic history that complete count data and automated record linkage at scale have ushered in is exciting. But the toolkits scholars are using to create longitudinal historical data are still evolving. In this paper, we examined one aspect of the data construction process---the creation of training data to "teach" supervised record linking algorithms. In other work, other scholars are interrogating other parts of the record linkage pipeline, from census enumeration accuracy and transcription (Ghosh, Hwang and Squires 2023; Hwang and Squires 2023) to sample bias (Bailey et al 2020) and performance across algorithms (Abramitzky et al 2021; Bailey et al 2020; Helgertz et al 2022). We hope other researchers will continue this trend, both improving the tools we all use to record link and deepening our understanding of how our new historical data is created.

## References

Abramitzky, R., L. Boustan, K. Eriksson, J. Feigenbaum, and S. Pérez. 2021. "Automated linking of historical data." *Journal of Economic Literature* 59(3):865-918.

Bailey, M.J., C. Cole, M. Henderson, and C. Massey. 2020. "How well do automated linking methods perform? Lessons from US historical data." *Journal of Economic Literature* 58(4):997-1044.

Bailey, M.J., P.Z. Lin, A.R.S. Mohammed, P. Mohnen, J. Murray, M. Zhang, and A. Prettyman. 2022. "The Creation of LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database Project." California Center for Population Research, UCLA.

Choi, J. 2022. "The Effect of Deindustrialization on Local Economies: Evidence from New England Textile Towns."

Feigenbaum, J.and D.P. Gross. 2022. "Answering the Call of Automation: How the Labor Market Adjusted to the Mechanization of Telephone Operation." National Bureau of Economic Research.

Feigenbaum, J.J. 2016. "A Machine Learning Approach to Census Record Linking."

—. 2018. "Multiple measures of historical intergenerational mobility: Iowa 1915 to 1940." *The Economic Journal* 128(612):F446-F481.

French, J. 2022. "Technological Change, Inequality, and Intergenerational Mobility: The Case of Early 20th Century Agriculture."

Ghosh, A., S.I.M. Hwang, and M. Squires. 2023. "Links and legibility: Making sense of historical US Census automated linking methods." *Journal of Business & Economic Statistics*:1-12.

Goldstein, J.R., M. Alexander, C. Breen, A.M. González, F. Menares, M. Osborne, M. Snyder, and U. Yildirim. 2021. "CenSoc Mortality File: Version 2.0." Berkeley: University of California.

Halpern-Manners, A., J. Helgertz, J.R. Warren, and E. Roberts. 2020. "The effects of education on mortality: Evidence from linked US Census and administrative mortality data." *Demography* 57(4):1513-1541.

Helgertz, J., J. Price, J. Wellington, K.J. Thompson, S. Ruggles, and C.A. Fitch. 2022. "A new strategy for linking US historical censuses: A case study for the IPUMS multigenerational longitudinal panel." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 55(1):12-29.

Hwang, S.and M. Squires. 2023. "Linked Samples and Measurement Error in Historical Us Census Data." *Available at SSRN 4519619*.

Long, J.and J. Ferrie. 2013. "Intergenerational occupational mobility in Great Britain and the United States since 1850." *American Economic Review* 103(4):1109-1137.

Price, J., K. Buckles, J. Van Leeuwen, and I. Riley. 2021. "Combining family history and machine learning to link historical records: The Census Tree data set." *Explorations in Economic History* 80:101391.

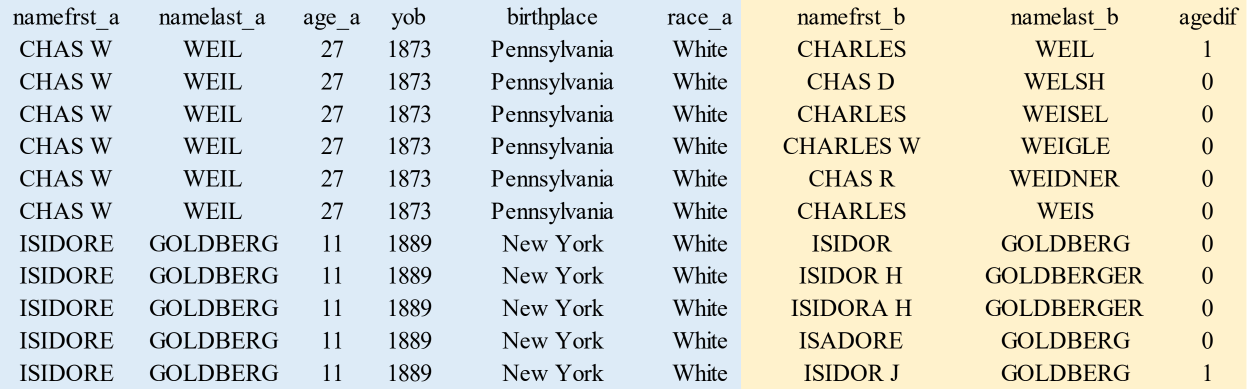
Ward, Z. 2023. "Intergenerational mobility in American history: Accounting for race and measurement error." National Bureau of Economic Research.

## Appendix A: Tables and Figures

*Table A1: Machine learning algorithm calibration statistics*

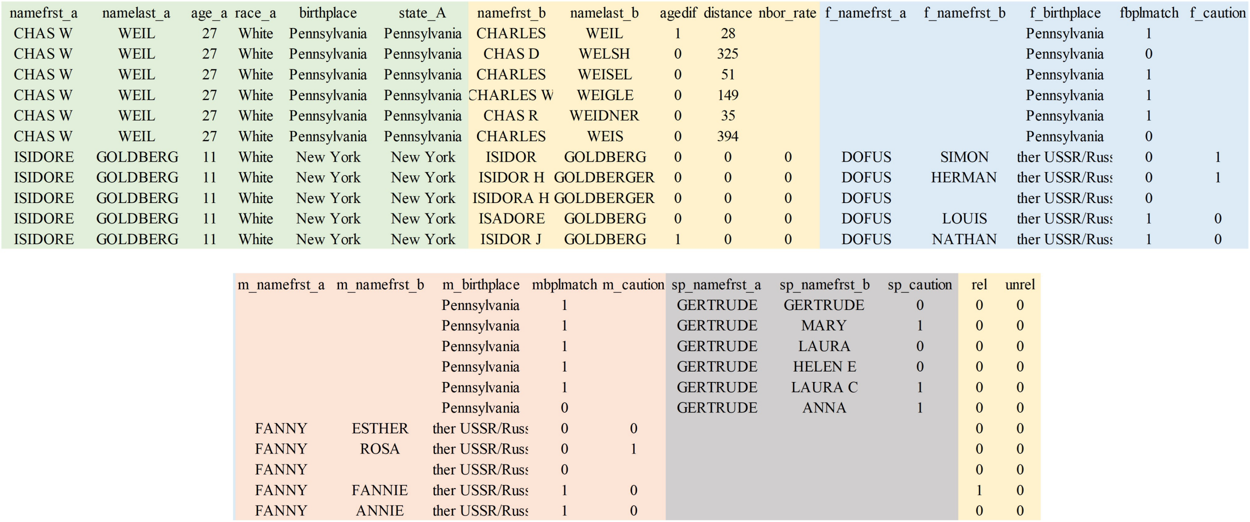
| | 1900-1910 | | | 1900-1930 | | |
|---|---|---|---|---|---|---|
| **Training data** | **Method I** | **Method II** | **Method III** | **Method I** | **Method II** | **Method III** |
| Basic features algorithm | | | | | | |
| MCC | 0.71 | 0.58 | 0.58 | 0.73 | 0.57 | 0.51 |
| Precision | 0.72 | 0.50 | 0.61 | 0.71 | 0.58 | 0.45 |
| Recall | 0.71 | 0.67 | 0.56 | 0.73 | 0.57 | 0.57 |
| $\alpha$ | 0.30 | 0.13 | 0.20 | 0.31 | 0.28 | 0.14 |
| $\beta$ | 2.50 | 2.05 | 2.90 | 2.60 | 2.35 | 2.45 |
| Extended features algorithm | | | | | | |
| MCC | 0.72 | 0.79 | 0.84 | 0.70 | 0.68 | 0.57 |
| Precision | 0.73 | 0.74 | 0.84 | 0.64 | 0.65 | 0.53 |
| Recall | 0.71 | 0.85 | 0.84 | 0.77 | 0.71 | 0.62 |
| $\alpha$ | 0.36 | 0.20 | 0.26 | 0.19 | 0.23 | 0.15 |
| $\beta$ | 2.05 | 1.75 | 1.20 | 3.00 | 2.75 | 1.80 |

***Figure A1: Examples of Method I training data***



| namefrst_a | namelast_a | age_a | yob | birthplace | race_a | namefrst_b | namelast_b | agedif |
|---|---|---|---|---|---|---|---|---|
| CHAS W | WEIL | 27 | 1873 | Pennsylvania | White | CHARLES | WEIL | 1 |
| CHAS W | WEIL | 27 | 1873 | Pennsylvania | White | CHAS D | WELSH | 0 |
| CHAS W | WEIL | 27 | 1873 | Pennsylvania | White | CHARLES | WEISEL | 0 |
| CHAS W | WEIL | 27 | 1873 | Pennsylvania | White | CHARLES W | WEIGLE | 0 |
| CHAS W | WEIL | 27 | 1873 | Pennsylvania | White | CHAS R | WEIDNER | 0 |
| CHAS W | WEIL | 27 | 1873 | Pennsylvania | White | CHARLES | WEIS | 0 |
| ISIDORE | GOLDBERG | 11 | 1889 | New York | White | ISIDOR | GOLDBERG | 0 |
| ISIDORE | GOLDBERG | 11 | 1889 | New York | White | ISIDOR H | GOLDBERGER | 0 |
| ISIDORE | GOLDBERG | 11 | 1889 | New York | White | ISIDORA H | GOLDBERGER | 0 |
| ISIDORE | GOLDBERG | 11 | 1889 | New York | White | ISADORE | GOLDBERG | 0 |
| ISIDORE | GOLDBERG | 11 | 1889 | New York | White | ISIDOR J | GOLDBERG | 1 |

## *Figure A2: Example of Method II training data (split across two sections)*



| namefrst_a | namelast_a | age_a | race_a | birthplace | state_A | namefrst_b | namelast_b | agedif | distance | nbor_rate | f_namefrst_a | f_namefrst_b | f_birthplace | fbplmatch | f_caution | m_namefrst_a | m_namefrst_b | m_birthplace | mbplmatch | m_caution | sp_namefrst_a | sp_namefrst_b | sp_caution | rel | unrel |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CHAS W | WEIL | 27 | White | Pennsylvania | Pennsylvania | CHARLES | WEIL | 1 | 28 | | | | Pennsylvania | 1 | | | | Pennsylvania | 1 | | GERTRUDE | GERTRUDE | 0 | 0 | 0 |
| CHAS W | WEIL | 27 | White | Pennsylvania | Pennsylvania | CHAS D | WELSH | 0 | 325 | | | | Pennsylvania | 0 | | | | Pennsylvania | 1 | | GERTRUDE | MARY | 1 | 0 | 0 |
| CHAS W | WEIL | 27 | White | Pennsylvania | Pennsylvania | CHARLES | WEISEL | 0 | 51 | | | | Pennsylvania | 1 | | | | Pennsylvania | 1 | | GERTRUDE | LAURA | 0 | 0 | 0 |
| CHAS W | WEIL | 27 | White | Pennsylvania | Pennsylvania | CHARLES W | WEIGLE | 0 | 149 | | | | Pennsylvania | 1 | | | | Pennsylvania | 1 | | GERTRUDE | HELEN E | 0 | 0 | 0 |
| CHAS W | WEIL | 27 | White | Pennsylvania | Pennsylvania | CHAS R | WEIDNER | 0 | 35 | | | | Pennsylvania | 1 | | | | Pennsylvania | 1 | | GERTRUDE | LAURA C | 1 | 0 | 0 |
| CHAS W | WEIL | 27 | White | Pennsylvania | Pennsylvania | CHARLES | WEIS | 0 | 394 | | | | Pennsylvania | 0 | | | | Pennsylvania | 0 | | GERTRUDE | ANNA | 1 | 0 | 0 |
| ISIDORE | GOLDBERG | 11 | White | New York | New York | ISIDOR | GOLDBERG | 0 | 0 | 0 | DOFUS | SIMON | ther USSR/Russ | 0 | 1 | FANNY | ESTHER | ther USSR/Russ | 0 | 0 | | | | 0 | 0 |
| ISIDORE | GOLDBERG | 11 | White | New York | New York | ISIDOR H | GOLDBERGER | 0 | 0 | 0 | DOFUS | HERMAN | ther USSR/Russ | 0 | 1 | FANNY | ROSA | ther USSR/Russ | 0 | 1 | | | | 0 | 0 |
| ISIDORE | GOLDBERG | 11 | White | New York | New York | ISIDORA H | GOLDBERGER | 0 | 0 | 0 | DOFUS | | ther USSR/Russ | 0 | | FANNY | | ther USSR/Russ | 0 | | | | | 0 | 0 |
| ISIDORE | GOLDBERG | 11 | White | New York | New York | ISADORE | GOLDBERG | 0 | 0 | 0 | DOFUS | LOUIS | ther USSR/Russ | 1 | 0 | FANNY | FANNIE | ther USSR/Russ | 1 | 0 | | | | 1 | 0 |
| ISIDORE | GOLDBERG | 11 | White | New York | New York | ISIDOR J | GOLDBERG | 1 | 0 | 0 | DOFUS | NATHAN | ther USSR/Russ | 1 | 0 | FANNY | ANNIE | ther USSR/Russ | 1 | 0 | | | | 0 | 0 |

***Figure A3: Examples of adjusted classification of training data through ground truth data***

| | | (A) | (B) | (C) | (D) | (E) |
|---|---|---|---|---|---|---|
| Census A | Census B | Predicted outcome | Training data | Ground truth | Unadjusted classification | Adjusted classification |
| John Smith | John Smith | 0 | 0 | 1 | TN | FN |
| John Smith | John Smith | 0 | 0 | 0 | TN | |
| John Smith | Jon Smith | 0 | 0 | 0 | TN | |
| John Smith | Johnny Smith | 0 | 0 | 0 | TN | |
| Steve Corcoran | Steven Corcoran | 1 | 1 | 1 | TP | |
| Steve Corcoran | Steve Camillo | 0 | 0 | 0 | TN | |
| Steve Corcoran | Steve Carmelo | 0 | 0 | 0 | TN | |
| J.S. Little | J Little | 1 | 0 | 1 | FP | TP |
| J.S. Little | John Lidle | 0 | 0 | 0 | TN | |
| J.S. Little | Sol Little | 0 | 0 | 0 | TN | |
| Karl Nordstrom | Charles Nordstroı | 0 | 0 | 0 | TN | |
| Karl Nordstrom | Carl Nordstrom | 0 | 1 | 1 | FN | |
| Karl Nordstrom | Karl Nord | 0 | 0 | 0 | TN | |
| James Arnold | James Arnold | 0 | 0 | 0 | TN | |
| James Arnold | James Arnold | 0 | 0 | 0 | TN | |
| James Arnold | Jimmy Arnold | 0 | 0 | 1 | TN | FN |
| Jacob Rovesky | Jacob Roveskij | 0 | 0 | 0 | TN | |
| Jacob Rovesky | Jake Roveski | 1 | 1 | 0 | TP | FP |
| Jacob Rovesky | Jacob Rogenbeck | 0 | 0 | 0 | TN | |
| Jacob Rovesky | Jacob Kovesky | 0 | 0 | 1 | TN | FN |
| Charles Arrington | Charles Arington | 1 | 0 | 1 | FP | TP |
| Charles Arrington | Charles Livingstoı | 0 | 0 | 0 | TN | |
| William Kraft | William Kratz | 0 | 0 | 0 | TN | |
| William Kraft | William Kraft | 0 | 1 | 0 | FN | TN |
| William Kraft | Bill Krafth | 0 | 0 | 1 | TN | FN |
| William Kraft | William Kurth | 0 | 0 | 0 | TN | |

## Appendix B: Manual check of links not covered by the FamilyTree

To investigate the extent to which the accuracy of links covered by the *FamilyTree* differs from those that do not, we extracted 100 random cases from each linked dataset presented in Table 4 that were not covered by the *FamilyTree*. We first had all cases manually checked at the Record Linking Lab at Brigham Young University (BYU). The procedure for doing so involves finding or creating a profile on FamilySearch for one of the census records and then using all of the information, records and search tools on FamilySearch to find the correct match in the census year. Based on this investigation, each case was allocated into one out of three outcomes: a correct match, an incorrect match, or not enough information to determine. For a total of 136 (1900-1910) and 131 (1900-1930) cases, the BYU team was unable to conclusively determine the outcome. We therefore tasked a pair of hand linkers already employed within the Multigenerational Longitudinal Panel project at the University of Minnesota (UMN) to evaluate these cases. From previously performing related tasks, the UMN team already had ample experience in evaluating the accuracy of linked records by using the wealth of information provided by Ancestry.com. Especially useful is its "suggested records" feature, which often allows the hand linker to be more confident in declaring a match or a mismatch. For example, two census records being linked to a marriage record may assist the hand linker in tracking an individual across census records when remaining household members offer no supporting evidence in favor of or against a record pair being a match. These supplementary sources are particularly important when an individual transitions from being a child, residing with their parents and siblings, to being a married adult, residing with their spouse and children.

For 46 (1900-1910) and 64 (1900-1930) cases, the two UMN hand linkers were in agreement that the match was accurate, leaving 16/16 unresolved and 74/51 mismatched cases. We proceed to calculate two measures of accuracy, with the first one representing the share of accurate matches after adjustment for the cases where the UMN hand linkers were able to confirm a match. The second measure represents an upper bound accuracy, where the aforementioned unresolved cases also are considered as accurate matches. The results are presented in the table below, illustrating a nontrivial decline in accuracy across all linked datasets, as was also expected, given the *FamilyTree* universe of links not being representative of the underlying population of possible links. While the magnitude of none of the differences between the accuracy statistics presented here and in Table 4 should not be underestimated, it is at the same time evident that the same conclusions fundamentally hold.

| | Basic features algorithm | | | Extended features algorithm | | |
|---|---|---|---|---|---|---|
| | Method I | Method II | Method III | Method I | Method II | Method III |
| **1900-1910** | | | | | | |
| Accuracy | 60% | 60% | 53% | 77% | 88% | 86% |
| *Upper bound* | *66%* | *62%* | *55%* | *80%* | *90%* | *87%* |
| **1900-1930** | | | | | | |
| Accuracy | 86% | 77% | 74% | 81% | 85% | 84% |
| *Upper bound* | *86%* | *79%* | *78%* | *86%* | *88%* | *86%* |