A Novel Dataset for Historical Innovation Studies: Linking USPTO Patents and US Census Data from 1840 to 1950

VERY PRELIMINARY AND INCOMPLETE VERSION FOR NBER 3I MEETING ONLY - PLEASE DO NOT DISSEMINATE

Stefano Breschi[*] Zenne Hellinga[†] Andrea Morrison[‡]
Jay Prakash Nagar[§] Gianluca Tarasconi[¶]

Abstract

Patents and inventor data are vital for studying innovation, but often lack socio-demographic information, such as nationality, gender, and occupational history, limiting their research potential. Moreover, the availability of digitized patent data has historically been restricted to recent decades, constraining long-term analyses of technological change. The emergence of digitized patent records (e.g., Google Patents) and accessible census micro-data has created new opportunities for integrating innovation data with demographic information. This has sparked interest in the academic community, with few attempts to produce patent census datasets. By building on these efforts, this work compiles a patents-census linked dataset for the US over the period 1840 to 1950. It contributes to the existing literature by enhancing the scale and quality of linked patent-census datasets. First, we compile a patent-inventor dataset with over 2.5 million records by triangulating multiple data sources. Second, we create a patent-census linked dataset containing approximately 1.4 million records—substantially larger than those used in prior research.

[*]Bocconi University, Email: stefano.breschi@unibocconi.it
[†]Utrecht University, Email: z.hellinga@uu.nl
[‡]University of Pavia and Icrios, Bocconi University, Email: andrea.morrison@unipv.it
[§]Duke University, Email: jayprakash.nagar@duke.edu
[¶]IPQuants AG, Email: gt@ipquants.com

1 Introduction

Innovation and technological change are fundamental drivers of economic growth and essential factors in explaining the long-term prosperity of countries (Romer 1990). Patents, and inventor data more broadly, have been extensively used in economic literature to measure innovation, and despite certain well-documented limitations (J. Acs and Audretsch 1989), they remain a unique and almost irreplaceable source of information to analyse how innovation processes unfold over time and across space.

The reason can be attributed to several factors. First, patent documents provide long-time series with standardized information that allows to make comparisons over time and across geographical locations. Second, patent records detail inventors' full names, addresses (work or home), and locations, enabling precise geolocalization of innovative activity. Likewise, data on the invention's owner (i.e. the patent assignee) is also included (Morrison, Riccaboni, and Pammolli 2017). Third, patent documents contain detailed descriptions of technological features, including references to prior inventions or scientific publications, allowing researchers to track knowledge linkages (Marx and Fuegi 2020). Despite patent documents provide extremely rich information, they also lack certain relevant pieces of information. For example, patents documents lack information on inventors' nationality, country of birth and gender - with few exceptions (Fink, Miguélez, and Raffo 2013) -, as well as occupational history before or after patenting. Consequently, researchers are limited in their possibilities when investigating the interplay between socio-demographic factors—such as migration, fertility, gender, and racial discrimination—and innovation processes. To address such questions, researchers must match inventor and assignee data with external sources, such as employer-employee linked administrative data. However, both administrative micro-data and patent data have been available only for recent time periods until very recently. For instance, digitized patent data (e.g. from the USPTO or EPO) was only available from the early 1970s onward until recently, limiting any long-term analysis of technological change. Historical studies, therefore, have typically relied on self-collected datasets from archival records, which restricts them in temporal range and technological scope (Nuvolari and Vasta 2017; Sokoloff 1988).

The release of digitized patent data via Google Patents, along with accessible population census micro-data for numerous countries, particularly the USA, has opened new avenues for research and sparked significant scholarly interest. Recent years have seen original efforts to construct long-term patent datasets and link

them with other databases, including census data. Two main strands of this literature are discernible: one focusing on constructing historical patent databases and the other on linking patent data to historical census micro-data. This paper seeks to contribute to the latter by presenting a preliminary version of a dataset linking USPTO patent-inventor data to U.S. census micro-data for the period 1840–1950. As said, there have been already some efforts that have generated historical patent datasets, and some of them have linked patents to census data. For example, Petralia, Balland, and Rigby (2016) produced a pioneering dataset, i.e. "HistPat", by geolocating USPTO patent documents for U.S.-resident inventors from 1790 to 1978. HistPat is a publicly available dataset containing over 1.5 million records. Another recent example is "PatentCity" by Bergeaud and Verluise (2022). PatentCity expanded existing datasets to include also the patenting activity of France, the UK, and Germany. By employing advanced NLP techniques, they enriched the dataset for example with information on occupation and citizenship [1].

To our knowledge, only four datasets have linked U.S. patents to U.S. census data: Akcigit, Grigsby, and Nicholas (2017b), Sarada, Andrews, and Ziebarth (2019), Arkolakis, Lee, and Peters (2020), and Hartog et al. (2024) (see Section 1.1 for an overview). Akcigit, Grigsby, and Nicholas (2017b) use a patent-census dataset to investigate factors underpinning the emergence of the U.S. technological leadership in the twentieth century, examining inventor life cycles and socio-economic processes like spatial and social mobility and inequality. Sarada, Andrews, and Ziebarth (2019) provide an overview of inventors' socio-demographic characteristics, comparing them to general census populations regarding factors such as gender, spatial mobility, and race. Arkolakis, Lee, and Peters (2020) supplement the patent-census dataset with immigration records and passenger lists between Europe and the US, allowing insights into the skill profiles of immigrant inventors. Finally, Hartog et al. (2024) examine the emergence of research labs in US companies and analyse how research activity shifted from an individual to a collective endeavour by integrating multiple datasets.

These studies, while addressing distinct research questions, share a core methodology: linking U.S. census data with historical patent data. Nonetheless, their dataset construction strategies differ to some extent. First, temporal coverage varies: Akcigit, Grigsby, and Nicholas (2017b) examine 1880–1940, whereas Hartog et al. (2024) begin in 1850 and Arkolakis, Lee, and Peters (2020) conclude their analysis in 1920. Second, they draw on different patent data sources. For instance, Sarada, Andrews, and Ziebarth (2019) and Hartog et al. (2024) use inventor data from

[1] for an review of earlier studies see Andrews (2021)

the Annual Reports by the Commissioner of Patents , while Akcigit, Grigsby, and Nicholas (2017b) and Arkolakis, Lee, and Peters (2020) rely on USPTO documents. Third, the methodologies used for matching inventors to census records vary: for example all studies use some kind of sub-national geographical scale (e.g., state, county) to identify potential matches, except Hartog et al. (2024), who match each inventor to all census records in the US for a certain vintage. Additional differences arise in matching procedures, such as matching patents to one or multiple census waves or employing various inventor name fields (e.g. surname, initials, forename). Finally, some studies implement matching scores and accuracy measures to select potential candidates (e.g., Hartog et al. (2024); Arkolakis, Lee, and Peters (2020); Sarada, Andrews, and Ziebarth (2019)), while others do not (e.g., Akcigit, Grigsby, and Nicholas (2017b)).

We claim that our work contributes to these efforts in two main ways. First, we expand the patent dataset, achieving higher-quality data extracted from Google Patents. This is done by collecting more patent documents and by triangulating multiple data sources (e.g. OCR of patent images and information from Google Patent pages), resulting in a patent-inventor dataset with over 2.5 million records. Second, our matching procedure produce a larger patent-census linked dataset, containing approximately 1.4 million records—significantly more than in prior studies. Three factors have contributed to this outcome: a) a larger initial patent-inventor dataset; b) the use of two (and occasionally three) census waves to match patents; c) and a more comprehensive candidate-matching methodology employing a wider predictors set than other studies.

Our matching rates range from 28,7% to about 45,2%, with lower rates observed for early census waves, where evidence suggests some under-reporting by enumerators, especially in rural areas of the US. We also find variation by nationality, suggesting potential inaccuracies in the census reporting of foreign names, as also noted in Akcigit, Grigsby, and Nicholas (2017b). We find also that States with higher patenting rates generally show higher matching rates as well, which is perhaps not surprising.

Our findings indicate that inventors were predominantly middle-aged white men, consistent with prior studies. For example, Sarada, Andrews, and Ziebarth (2019) and Akcigit, Grigsby, and Nicholas (2017b) report that 95–97.9% of inventors were male and approximately 97% were white, with an average age of around 41. As in Akcigit, Grigsby, and Nicholas (2017b) we also found that immigrants accounted for about 20% of inventors. These results suggest that, despite some discrepancies,

our data align well with previous findings, providing a coherent view of historical inventor demographics.

The paper is structured as follows. Section 1.1 provides a brief overview of previous studies that have attempted to match US patent data with US census data. Section 2 discusses the methodology used to construct our patent dataset. It details the data sources (Section 2.1), the methodology used to extract information from the patent documents (Section 2.2 and subsections therein), and how gender was assigned to the inventor (Section 2.3). Section 3 discusses the differences between our dataset and other publicly available patent datasets. In Section 4, we present the full count US Census dataset of Ipums, which was used to match our patent-inventor dataset. The matching procedure is presented in Sections 5 and 6 and their subsections. In particular, Section 5 discusses the blocking procedure used to generate candidate matches. Section 6 discusses in detail the matching strategy. Section 7 discusses the machine learning model used to select among the many multiple candidates found during the matching procedure. Section 8 presents some demographics of the inventors with the sole purpose of validating our methodology. Section 9 illustrates some robustness checks of the ML model. Section 10 concludes.

1.1 Other attempts to match USPTO patents to US Census data

In recent years, four other studies have attempted to match historical USPTO patents to US census data. In this section, we aim to provide as complete an overview as possible of their strategies and outcomes.

The earliest attempt is made by Akcigit, Grigsby, and Nicholas (2017a) and Akcigit, Grigsby, and Nicholas (2017b), who match the first inventor appearing on patents granted in the same year as the census was conducted for each census year between 1880 and 1940. Their initial matching criteria require inventors to have the same first name, last name, and state and county of residence as census individuals. If multiple candidates remain for an inventor after this initial match, they apply additional filters sequentially: (1) only individuals with the same middle initial (if available) are retained, (2) only those residing in the same city or township are kept, (3) only those aged 16 to 85 are retained, and (4) only those aged 16 to 65 are retained. Any inventors with duplicate matches after all these filters are removed from the dataset. This methodology yields a matching rate of 29–46%.

Rather than using patent documents, Sarada, Andrews, and Ziebarth (2019) begin their analysis with lists of patent grantees from the Annual Report of the

Commissioner of Patent. Similar to Akcigit, Grigsby, and Nicholas (2017a), they restrict their matches to inventors granted patents in the same census years, covering 1870 to 1940. Their matching process requires an exact match on state of residence and initials of both the first and last name, supplemented by a fuzzy match based on full first and last names and the name of the town of residence. Each of these characteristics are scored for quality and weighted using a subset of manually matched inventors from Vermont in 1900. To be matched an inventor must have an overall score above the average match score of all possible matches. This results in a match rate varying around 10%, with most matches being unique around 70% of them being perfect. Since the Annuals do not report the abstract of the patent or the technological class, the dataset lack such information.

Arkolakis, Lee, and Peters (2020) Arkolakis et al. (2020) link patents (only for male inventors) to census data for the years from 1880 to 1920. Patents are matched to a census record if they were filed within the same decade and State Economic Area (SEA), using the Jaro-Winkler distance to assess similarity between the first and last names of the patent inventor and the census individual. A Jaro-Winkler score above 0.8 is required, and the individual's age at the time of filing must be over 16. To ensure unique matches, they exclude cases with multiple candidates who have a Jaro-Winkler score above 0.85 or are over 80 years old. Their matching rate varies between 8% and 60%.

The most recent attempt is by Hartog et al. (2024). Similar to Sarada, Andrews, and Ziebarth (2019), they begin with yearly lists of patent grantees, matching these to the two closest census waves based on each patent's grant date for census years between 1850 and 1940. For each inventor, they generate an initial set of census candidates by identifying all individuals in a census wave with similar last names, yielding an average of 326,697 candidates per inventor. Next, to establish a ground truth for inventor-census matches, they incorporate additional information on a subset of inventors (i.e., birth date, birthplace, and family ties) obtained from Wikidata. Using this ground truth, they employ a two-stage XGBoost model. In the first stage, they predict an initial match quality score based on name similarity (first and last name, initials) and geographic distance between the inventor's and census individual's counties of residence. In the second stage, they apply another XGBoost model that uses the first-stage quality scores along with a different set of ground truth observations to generate a refined matching score for each inventor-candidate pair. The final match is made by selecting the pair with the highest score from the second stage, provided the candidate is over 16 years old. This approach allows them to successfully match 46% of inventors.

2 Historical USPTO data

Since the release of digitized patents, a number of attempts have been made to construct long time series of historical patent data, particularly using US patent data. Building on some of these attempts (Petralia, Balland, and Rigby (2016), Bergeaud and Verluise (2024), we present our historical patent dataset in the following sections. We discuss the methodological steps taken to extract and standardise relevant information from patent documents. We also compare our dataset with other publicly available datasets along some key dimensions (e.g. number of patents).

2.1 Data sources

The dataset for our study is based on USPTO patent documents filed between 1836 and 1950.[2] We obtained these documents as PDF files from Google Patents. Since these files are image-based and do not contain selectable text, we used the Python library `PDFMiner`, specifically the `pdf2txt.py` function, to extract text data.[3]

The files were downloaded and processed in batches of 20.000 patents, since the overall amount of data (2.521.306 patents) could not be stored on the working machine. Since it was only possible to work with publications, the application was iterating on a progressive number ranging from one to the last published patent in the time-frame we were interested to (end of 1950 in our exercise).

Aside PDF download, we enriched the dataset by downloading the main pieces of information relative to the patents in exam as assignees, cpcs, references, citations, concepts, priorities and the timeline of events related to that patents.

2.2 Extracting inventor data

This section describes the procedure used to extract inventor information (e.g. name, address) from patent documents. Since the format in which this information is reported in the patent changes over time, we had to first parse the document to extract the relevant text where the information was plausibly found (see Section 2.2.2), and then analyse it to extract the relevant inventor information (see Section 2.2.3).

[2]The USPTO was formally established on July 4, 1836.
[3]Note that Google Patents PDFs now often come pre-OCRed.

2.2.1 Pre-processing

To link patent inventors with individuals in census records, we parsed the patent documents to extract relevant inventor information. Until approximately 1925, patent documents typically began with one or more technical drawings signed by the inventor and two witnesses, followed by a detailed description of the invention (Bazerman 1997). The first page of text was headed by "United States Patent Office" and included subheadings with the inventor's name and the title of the invention. This was followed by a formal phrase, *Specification of Letters Patent [number], dated [date]*, introducing the document. The main text often began in the style of a formal letter with the salutation *To all whom it may concern*, leading into an introductory declaration (see Figure 1a):

Be it known that I, [name], of [city, county and state], have invented [...]

In 1925, with the transition of the Patent Office from the Department of the Interior to the Department of Commerce, a modernization phase began, aimed at fostering industrial growth and innovation in the United States.[4] As part of this transition, the patent document format was simplified. In the revised layout, as shown in Figure 1b, the inventor's name and location were placed directly below the "United States Patent Office" heading, eliminating the formal preamble and streamlining the overall document structure.

[4] https://www.commerce.gov/about/history/evolution

Figure 1: Headers of patents US10001 and US1683266

UNITED STATES PATENT OFFICE.

THOS. ALLISON, OF MILTON, NEW YORK.

STRAW-CUTTER.

Specification of Letters Patent No. 10,001, dated September 6, 1853.

To all whom it may concern:

Be it known that I, THOMAS ALLISON, of Milton, in the county of Ulster and State of New York, have invented a new and useful Improvement in Straw-Cutters; and I do hereby declare that the following is a full, clear, and exact description of the same, reference being had to the accompanying drawings, forming part of this specification.

C, is the obliquely arranged roller. The journals of the shaft D, of the same, resting in movable boxes *b*—which rest on springs *c*. By securing these journals in the movable sliding boxes the roller can be adjusted to any height and be kept in that position by springs *c* and set screw *d;* which screw serves to lower and raise the oblique roller. The springs upon which the movable boxes

(a) Header of patent US10001, filed on September 6, 1853

Patented Sept. 4, 1928. 1,683,266

UNITED STATES PATENT OFFICE.

LEWIS H. SHIPMAN, OF BOSTON, MASSACHUSETTS.

SOLAR HEATING APPARATUS.

Application filed August 5, 1925. Serial No. 48,323.

My invention relates to solar heating apparatus and it has for its object to provide an apparatus of this kind which will be of simple and efficient construction.

To these ends I have provided a solar heating apparatus having the peculiar features of construction and mode of operation set forth in the following description, the novel features of the invention being particularly pointed out and defined in the claims at the close thereof.

from the line of incidence of this median line with its reflector to the pipe 17 being, in the case of each of the six lenses, the same.

It will be clear from the description given above that with the group of lenses disposed perpendicularly to solar rays, those rays passing through each lens will be condensed and directed upon the reflectors, and thence directed, still converging to a focal line at the pipe 17. The middle lens 4 is constructed so as to focus directly upon the pipe 17. It

(b) Header of patent US1683266, filed on August 5, 1925

Mirroring these evolving document structures, we adopted flexible *regular expressions* to identify and extract paragraphs containing bibliographic information on inventors, while excluding extraneous elements such as witness signatures. The approach also addressed common formatting inconsistencies and corrected typical OCR errors. The box below provides a snapshot of the code used.

```
if pat_no <=1583766:
    start_pars = [
        'Know all men by these presents:',
        'That I,', 'Beit known that I', 'Beit known that, I',
        'That we,', 'Beit known that we', 'Beit known that, we',
        'To all whom it may concern:',
        'To all, whom it inval / concern: '
        ]
    endparagraph = re.compile('([a-z][a-zA-Z]( |)[.]| [0-9][.])$')
    MisspelledFormula = re.compile("^(7|T)OALL((A|)
    WOHM|OTHON|TUHOT|WHOM)(,|)IT(MAY|INCTLY)
    (CONCERN|6ON6ERN|CONCE'N)(:|;|)")
    alternative_ends = ['WITNESS', 'WHATICLAIM', 'WHATDOICLAIM' ]
    alternative_begin = 'LETTERSPATENT'
else:
    start_pars = ['UNITED STATES',
                  'UNITED STATES PATENT OFFICE',
                  'UNITED STATES PATENT OFF']
    endparagraph = re.compile('Application Filed')
    MisspelledFormula = re.compile(
    "^(7|T)OALL((A|)WOHM|OTHON|TUHOT|WHOM)(,|)
    IT(MAY|INCTLY)(CONCERN|6ON6ERN|CONCE'N)(:|;|)")
    alternative_ends = [
        'APPLICATIONFILED', 'SERIALNO',
        'APPLICATIONFLED', 'INVENTIONRELATES', 'FOREXAMPLE']
     alternative_begin = 'UNITED STATES PATENT OFF'
```

The code differentiates patents by number, adjusting the parsing rules accordingly:

- For earlier patents (numbered $\leq$ 1583766), the code searches for the initial legal phrases (e.g., "Know all men by these presents") and ends with phrases such as "WITNESS" or "WHAT I CLAIM."
- For later patents, the code searches for text beginning with "UNITED STATES PATENT OFFICE" and ending with phrases like "APPLICATION FILED" or "SERIAL NO."

For example, in the case of patent US10001 (see Figure 1a), the code extracted

the following paragraph:

> *Be it known that I, THOMAS ALLISON, of Milton, in the county of Ulster and State of New York, have invented a new and useful*

In contrast, for patent US1683266, the code extracted:

> *UNITED STATES 1,683,266 PATENT OFFICE. LEWIS H. SHIPMAN, OF BOSTON, MASSACHUSETTS, HEATING APPARATUS. Application filed August 5, 1925. Serial No. 48,323.*

2.2.2 Post-processing with SpaCy

After extracting paragraphs containing data on inventors, the texts were processed to eliminate noise introduced by OCR technology. Subsequently, these texts were analyzed using SpaCy, an open-source software library dedicated to advanced natural language processing (Honnibal and Montani 2017).

For this purpose, we constructed a training set utilizing the HistPat dataset developed by Petralia, Balland, and Rigby (2016).[5] The training set comprises a dictionary associating each paragraph with the start and end character indices of the inventor's name within the text. The training set includes approximately 21,000 annotated cases. Below is a sample record from the annotated training set:

```
TRAIN_DATA = [
    ("Be it known that I, JoHN ALLEN CLARK,
    a citizen of the United States, residing at Columbus,
    in the county of Franklin and State of Ohio,
    have invented a new and",
    "entities": [(20, 36, "PERSON")]),
    (...),
]
```

After training the model, we leveraged the Named Entity Recognition (NER) capabilities of SpaCy to extract the names of inventors, identified by SpaCy as a PERSON, from all paragraphs outlined in section 2.2.1.

Similarly, we utilized SpaCy to identify the locations of inventors, which are recognized as geopolitical entities (GPE). For American inventors, addresses are detailed at the city, county, and state levels, whereas for non-US inventors, only city

[5] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BPC15W

and country are specified. We cross-verified and supplemented the information on inventors' locations using the dataset created by Bergeaud and Verluise (2022) and Bergeaud and Verluise (2024).[6]

Finally, we enriched the dataset with information on the patents' technological classes, retrieved from Google Patents and PATSTAT. Where available, citizenship information was also obtained by leveraging the dataset developed by Petralia, Balland, and Rigby (2016) and Diodato, Morrison, and Petralia (2021).

It is important to note that our unit of analysis is the "patent-inventor name" pair. We did not attempt to disambiguate individual inventors before matching them with census records.

Table 1: Number of patents in the data set by location of inventors

	Located in the US	Located elsewhere	Total
Patent-inventor pairs	2,559,849	524,455	3,084,304
Patents	2,237,296	428,222	2,519,800

Note: This table reports the number of patents published between 1840 and 1950 (inclusive) in our dataset, categorized by the location of the inventors. Note that the sum of patents with inventors located in the United States and those with inventors in other countries exceeds the total reported in the last column, as some patents include inventors from both the United States and abroad.

2.3 Our set of historical patents

The methodology described in the previous sections resulted in a final dataset of 3,084,304 patent-inventors and 2,519,800 patent documents, as shown in Table 1. At this stage, we have not undertaken any effort to disambiguate the names of inventors. The identification of individual inventors will indeed be possible with greater accuracy after we have matched them with the census records. For this purpose, we will only use patent-inventor pairs located in the US, which are the only ones we can expect to find in the US census. This results in a total of 2,559,849 patent-inventor records, which is more than 80% of the total. It is worth noting that the number of records in our dataset is larger than those already in the public domain. For example, Petralia, Balland, and Rigby (2016) have about 1.5 million patent inventors, while Bergeaud and Verluise (2024) have approximately 2.8 millions patent-inventors [7].

[6] https://cverluise.github.io/projects/patentcity/

[7] This number refers to the US patent data, excluding duplicate co-inventors

2.4 Assigning gender to inventors

While census data (see below, section 4) provides gender information for individuals, this information is not available for USPTO inventors. Since gender is a relevant feature in our matching and blocking strategy (see below, section 5), we assigned gender to inventors by utilizing two sources of information. The first source is the list of the 1,000 most common male and female names from 1880, as reported by the U.S. Social Security Administration.[8] The second source used is the World Gender-Name Dictionary (WGND) developed by Lax-Martinez et al. (2021).[9] Specifically, we used the `WGND_nocountry` version of the dictionary, which contains 177,042 unique names that are non-conflicting across sources and countries. We supplemented this list with a manually curated list of names to cover cases not captured by these sources.

Tables A1 and A2 in the Appendix report the number and percentage distribution of patent-inventor pairs by decade and inventor gender. Despite our use of an extensive name list, a significant share of inventors' genders could not be determined, particularly in the decades before 1870. This is likely due to OCR issues in transcribing names or the presence of obsolete names from the early 19th century that are not included in our name dictionary.

3 Comparing with other data sets

As discussed in the introduction, several historical USPTO patent datasets have been compiled over the past decade. To contextualize our dataset (BHMNT) within this landscape, it is essential to evaluate how it compares to these alternative sources. Our comparison focuses on three publicly available datasets.[10]

- The first dataset, HISTPAT, was compiled by Petralia, Balland, and Rigby (2016).[11] HISTPAT provides geographic information on patents granted by the United States Patent and Trademark Office (USPTO) from 1836 to 1975. Similar to our approach, HISTPAT is constructed using digitized records of

[8] https://www.ssa.gov/OACT/babynames/decades/names1880s.html

[9] For details on the methodology used by the authors, please see https://github.com/IES-platform/r4r_gender/blob/main/wgnd/README.md

[10] Another data set publicly available is the KPSS dataset, created by Kogan et al. (2017), which provides patent-level data from 1926 to 2023. Data is available at https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Extended-Data. KPSS is distinct in that it focuses specifically on patents assigned to publicly traded firms, and therefore it does not contain the full universe of patents. For this reason, we leave it outside this comparison.

[11] Specifically, we used version 8.0 of this dataset, which is available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BPC15W.

original patent documents that are publicly accessible.

- The second dataset, PATENTCITY, was developed by Bergeaud and Verluise (2022) and Bergeaud and Verluise (2024). Like HISTPAT, PATENTCITY is based on digitized images of USPTO patents.[12] PATENTCITY includes U.S. patents dating back to 1836 and provides extracted information such as inventors' names, locations, occupations, and citizenship.
- The third dataset, SAZ, was compiled by Sarada, Andrews, and Ziebarth (2019).[13] Differently from the other existing data sets, it takes patent data from the Annual Reports of the Commissioner of Patents and Annual Indices of Patents. The time span covered goes from 1870 to 1942. The data set provides information for each patent on inventor names, inventor location at the state and town levels, invention title, and type (i.e., utility patent, design patent, plant patent, and so on). For this comparison, we selected only utility patents.

3.1 Patent counts

The most straightforward dimension for comparison is coverage in terms of the number of patents. Figure 2a shows the number of patents in our dataset (BHMNT) alongside HISTPAT, PATENTCITY, and SAZ.

All datasets exhibit similar trends, showing cyclical patterns in patenting activity consistent with previous literature. Patent counts tend to increase over time, with noticeable declines during periods of economic downturns and wars, such as World Wars I and II. The BHMNT and PATENTCITY datasets are closely aligned in terms of patent counts, whereas HISTPAT and SAZ generally contain fewer patents than these two in most years.

To further evaluate the coverage of the BHMNT dataset, we benchmark it against aggregate patent counts provided by the USPTO. Figure 2b illustrates the difference between patent publications in our dataset and the official counts reported by the USPTO.[14] Our dataset aligns closely with official USPTO counts for most years, with only a few exceptions in 1881, 1900, 1926, and 1928. In these years, the

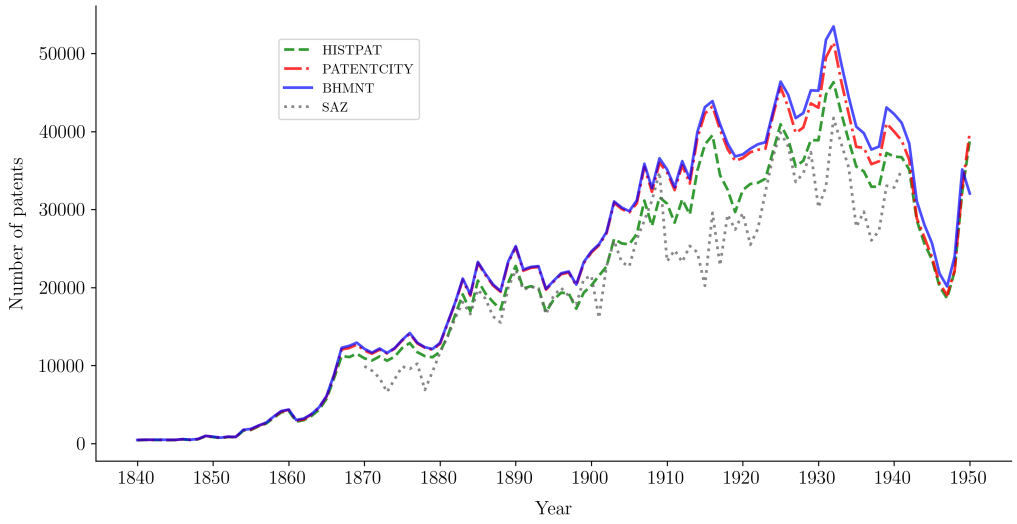
[12]The dataset is publicly available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PG6THV, and we used version 2.1. Notably, this dataset also includes historical patent records for Germany, France, and the UK.

[13]The data set is publicly available at https://www.openicpsr.org/openicpsr/project/120556/version/V1/view?path=/openicpsr/120556/fcr:versions/V1/READ_ME.txt&type=file.

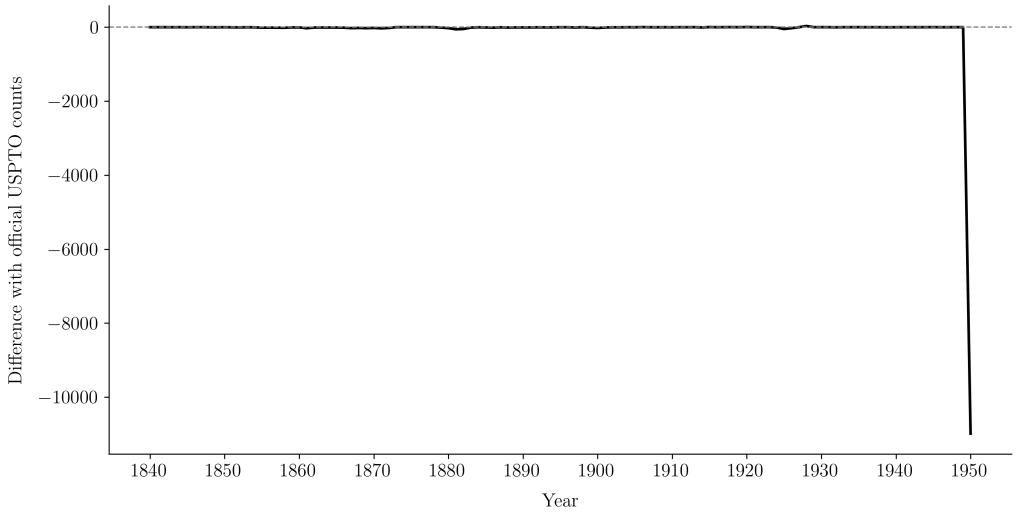
[14]See https://www.uspto.gov/web/offices/ac/ido/oeip/taf/h_counts.htm.

Figure 2: Number of patents

(a) Number of patents in different data sets



(b) Difference between BHMNT and official USPTO counts



Note: Figure 2a compares the count of patents for each year in our data set (BHMNT) with the data set built by Petralia, Balland, and Rigby (2016) (HISTPAT), that built by Sarada, Andrews, and Ziebarth (2019) (SAZ) and that developed by Bergeaud and Verluise (2022) and Bergeaud and Verluise (2024) (PATENTCITY). Figure 2b illustrates the difference in terms of patent publications between our data set and the official patent counts reported by the USPTO (See `https://www.uspto.gov/web/offices/ac/ido/oeip/taf/h_counts.htm.`).

differences never exceed 57 patents in absolute terms or 0.9% of the official counts, making the discrepancies negligible for practical purposes. However, 1950 presents a significant exception: our dataset is missing approximately 10,986 patents, or 25% of the official total for that year. This gap occurs because our dataset's time series ends at patent US2524969, with a publication date of 1950-10-10.

3.2 Number of inventors

Another comparative dimension relates to the identification and count of inventors. The HISTPAT datasets does not provide inventor names, whereas PATENTCITY and SAZ do. Although HISTPAT does not record inventor names, it does allow for the identification of patents with multiple inventors. In this dataset, each inventor's location is listed as a separate row in the data file, so patents with multiple rows indicate multiple inventors.

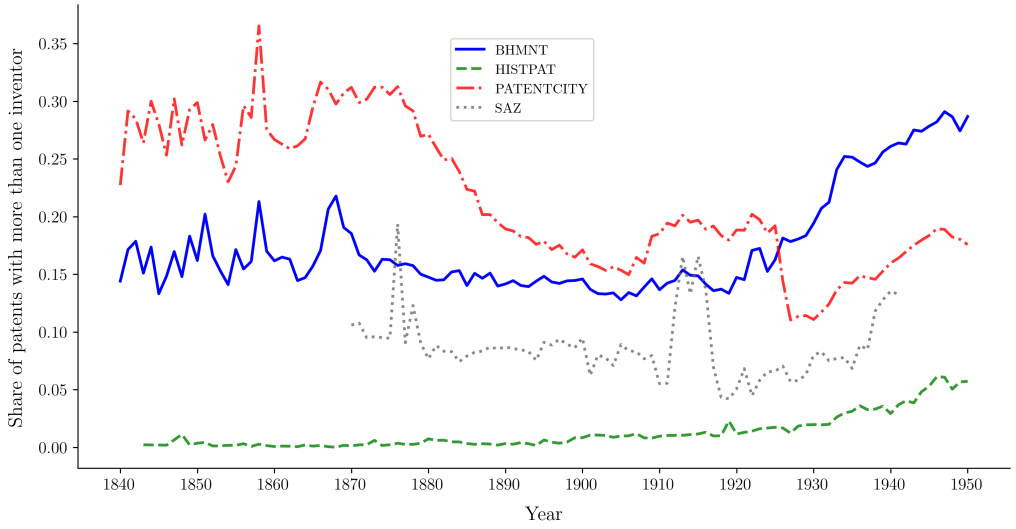
Figure 3a reports the share of patents with more than one inventor across the three datasets. The differences are notable. According to HISTPAT, nearly all patents before 1920 have only a single inventor. This could be due to a limitation in recording only the first inventor's name before 1920 or because the dataset aggregates inventors from the same location into a single record. In contrast, PATENTCITY shows that approximately 25-30% of patents before 1880 were produced by teams of inventors. This share then declines sharply until 1925, after which there is a further significant drop. In our dataset, however, the share of patents produced by teams remains relatively stable at around 15% until 1920, after which it begins to rise significantly. The SAZ database stands in between our data set and HISTPAT, reporting a share of patents made by teams around 8-10% until 1920.

A similar trend is observed in Figure 3b, which shows the average number of inventors per patent. In our dataset, the average number of inventors per patent remains steady, with minor fluctuations, at around 1.2 until 1920, after which it starts to increase. On the other hand, PATENTCITY reveals an opposite trend, with an average of approximately 1.3 inventors per patent before 1880, followed by a gradual decline until 1926.

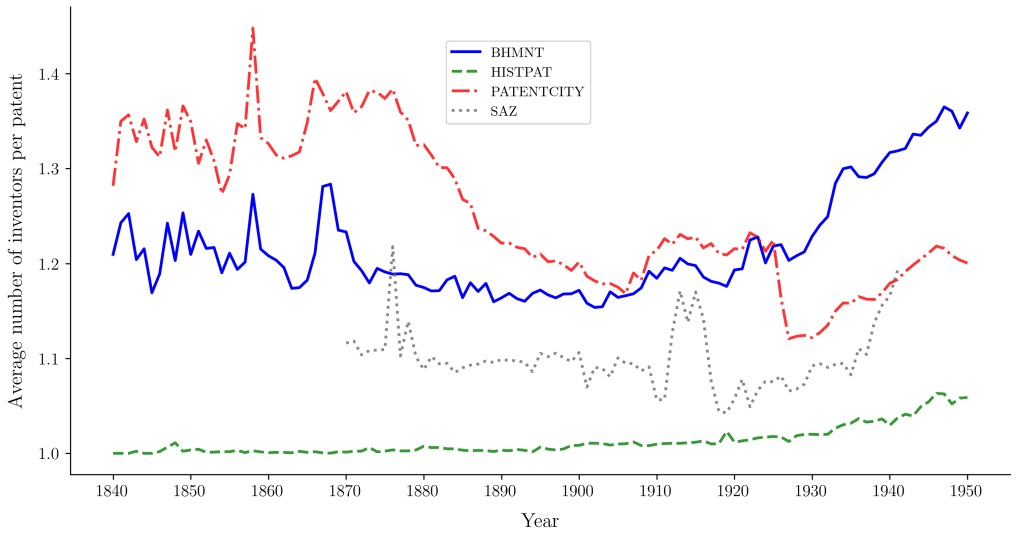
To investigate the likely causes of these discrepancies, we randomly selected a few patents where the number of inventors differed between our dataset and PATENTCITY, and inspected the inventor names reported in each. An illustrative example is patent US209237 (see Appendix, Figure A1). As shown in Table 2, this patent has 7 inventors listed in PATENTCITY but only 2 in our dataset. Some of the inventor names in PATENTCITY are duplicated, and these duplications are marked

Figure 3: Inventors per patent

(a) Share of patents filed by teams of inventors



(b) Average number of inventors per patent



Note: Figure 3a illustrates the yearly trends in the share of patents with multiple inventors, comparing data from our dataset (BHMNT) against the datasets compiled by Petralia, Balland, and Rigby (2016) (HISTPAT), Sarada, Andrews, and Ziebarth (2019) (SAZ), and Bergeaud and Verluise (2022) and Bergeaud and Verluise (2024) (PATENTCITY). Figure 3b displays the average number of inventors per patent across these four datasets.

by the variable `is_duplicate`. However, even after excluding all entries marked as duplicates (i.e., where `is_duplicate` is set to `True`), PATENTCITY still lists 4 inventors for this patent.

This discrepancy arises for two main reasons: (a) `Ep@ar PULASKI Da- VIS` and `EDGAR P. DAVIS` are not identified as the same person due to OCR and name recognition errors, and (b) `Hiram A. STURGES` is incorrectly labeled as an inventor, even though he is actually a *witness* (see Figure A1).

Table 2: Comparing inventor names, US Patent 238207

PATENCITY		**BHMNT**
name_text	is_duplicate	name
Hiram A. STURGES	False	
WALTER JOHN GODFREY	True	
WALTER J. GODFREY	False	WALTER J. GODFREY
WALTER J. GODFREY	True	
Ep@ar PULASKI Da- VIS	False	
EDGAR PULASKI DAVIS	True	EDGAR PULASKIDAVIS
EDGAR P. DAVIS	False	

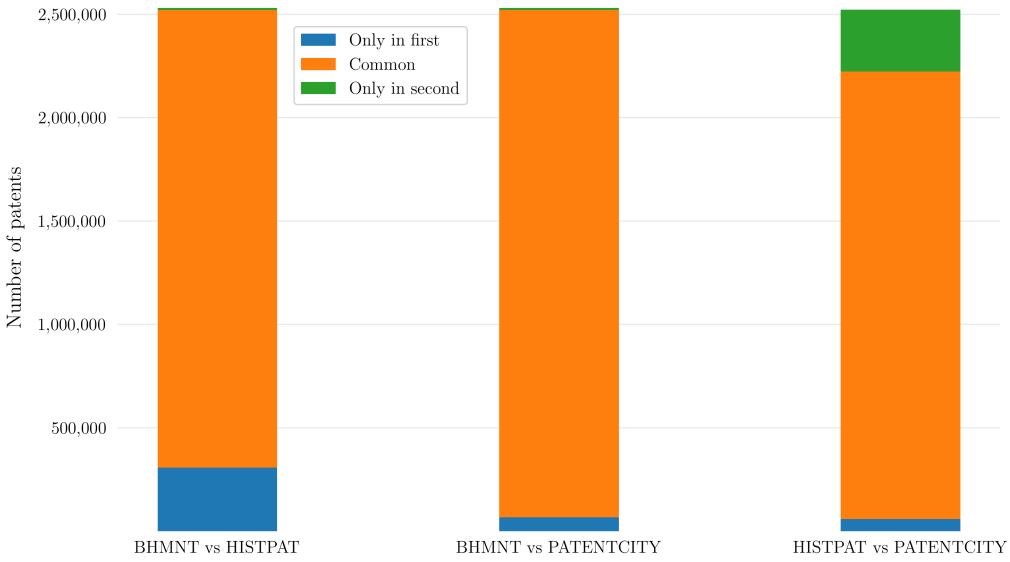
Therefore, for patents filed before 1926, PATENTCITY appears to overestimate the number of inventors for two key reasons. First, by searching for inventor names throughout the entire patent document, it leads to duplicate entries for the same inventor. Second, the named entity recognition system used in PATENTCITY sometimes misclassifies *witnesses* as inventors, further inflating the inventor count.

3.3 Intersection of data sets

In this section, we examine whether the datasets under review contain identical patents—that is, whether patents included in one dataset are also present in the others. Since each dataset records patent numbers, identifying patents that appear in one dataset but not in another is straightforward. Figure 4 and Table A5 in the Appendix present pairwise comparisons among the BHMNT, HISTPAT, and PATENTCITY datasets.

The results show that the majority of patents are common across all three datasets. However, BHMNT and PATENTCITY exhibit a substantial degree of overlap, while HISTPAT is missing a notable number of patents relative to these two datasets.

Figure 4: Intersection between data sets



Note: Figure 4 reports the number of patents appearing in one data set but not in the other.

4 Census data

We obtained the full count U.S. Census microdata from 1840 to 1940 from IPUMS (Integrated Public Use Microdata Series) at the University of Minnesota. Specifically, we were given access to the *restricted* version of decennial censuses.

Compared to the IPUMS Full Count dataset, the additional variables in the restricted-use files include:

- `namefrst`: 16-character first name (and possibly middle initial)
- `namelast`: 16-character last name
- `histid`: 36-character person identifier of the individuals included in each census for matching across IPUMS versions (but not census decades)
- `street`: street address.

All census tables were converted to `parquet` format to facilitate querying data.

We also use the information included in the IPUMS Full Count dataset, which provides both person and household level variables. The dataset contains over 800 million individual-level records for the period 1850-1940[15]. Person-level variables provide information on the status of the household members, such as demography and family relationships, income and occupation, and education. Household-level

[15]See https://usa.ipums.org/usa/full_count.shtml

variables provide information on the composition of the household, its geographical location, and its economic characteristics. It should be noted that not all the variables are available for the period 1850-1940, in particular those relating to economic characteristics of the household, personal income and work, which are missing from most (if not all) census waves in the time period 1850-1940. As an illustration, Table 3 reports some of the most relevant variables with a brief description of their content[16].

Furthermore, with the IPUMS Multigenerational Longitudinal Panel (MLP) project , individuals' records have been linked across different censuses. We plan to exploit this feature of the MLP using the records that have links across the censuses of 1850 through 1940.[17]

[16] A detailed description of the variables and how they are coded can be found here: https://usa.ipums.org/usa-action/variables/group

[17] More details can be found here: https://usa.ipums.org/usa/mlp/mlp.shtml

Table 3: IPUMS Full Count US Census (1850-1940): description of selected variables

Type	**Category**	**Name**	**Description**
P	Demography	RELATE	Relationship to household head: e.g., head, spouse, child
		SEX	Sex: male/female
		AGE	Age of the person
		MARST	Marital status: e.g., married, separated, divorced
		BIRTHYR	Year of birth
	Race, Ethnicity, and Nativity	RACE	Race: e.g., White, Black/African American, American Indian, Chinese
		BPL	Birth place: name of a U.S. state or a foreign country
		CITIZEN	Citizenship status: e.g., born abroad of American parents, naturalized citizen, not a citizen
		YRIMMIG	Year of immigration
		SPEAKENG	Speaks English: e.g., No/Yes
	Education	SCHOOL	School attendance: e.g., No, not in school; Yes, in school
		LIT	Literacy: e.g., No, illiterate/Yes, literate
	Work	LABFORC	Labor force status: e.g., No/Yes
		OCC1950	Occupation class (1950 basis): e.g., accountants and auditors
	Occupational standing	OCCSCORE	Occupational income score: 2-digit numeric variables
		ERSCOR50	Occupational earnings score (1950 basis): 4-digit numeric variable
H	Geography	STATEFIP	State: 2 digit State FIPS code
		COUNTYFIP	County: 3-digit numeric variable
		CITY	City: 4-digit numeric variable
		CITYPOP	City population
	Economic characteristics	FARM	Farm status: e.g., Not a farm; Yes, farm
	Household composition	NFAMS	Number of families in household
		NFATHERS	Number of fathers in household

4.1 Cleaning census names

Both the USPTO and census datasets underwent rigorous cleaning and standardization to enable accurate comparisons between inventors and census individuals. For the census data, one of the primary challenges was handling variations in how names were recorded, including differences in spelling, capitalization, and punctuation. To address this, we normalized the names by converting all text to lowercase and removing non-alphanumeric characters (except for spaces). Digits were also removed where necessary to ensure consistency. This step facilitated more effective matching by ensuring that minor differences in how names were recorded did not interfere with identifying potential matches.

In addition, to allow for more granular comparisons during the matching process, we maintained the separation of first and last names in the census data. This approach helped refine the matching process, as it enabled more flexibility in handling discrepancies between how names were entered across the two datasets.

Lastly, we enhanced the census data by attaching county and state of residence information to each individual using the crosswalks developed by Berkes, Karger, and Nencka (2023) as part of the Census Place Project[18]. These crosswalks provide standardized, geocoded locations for individuals based on raw place names found in the census data, significantly improving the geographic precision of our matches. This geolocation data was critical for enabling more refined and accurate comparisons, especially for common names that might otherwise produce ambiguous results.

This pre-processing step yielded a set of tables, one for each decennial census, with the following variables:

`histid`	36-char identifier of individual
`clean_fn`	Lower-case, cleaned first name
`clean_ln`	Lower-case, cleaned last name
`county_fips_code`	County FIPS code
`age`	Age
`birthyr`	Year of birth
`sex`	Sex (M, F, U)
`city`	City

As for the `county_fips_code` this is a 5-digit code obtained from joining state and county FIPS code of individuals from Berkes, Karger, and Nencka (2023). It

[18]Data is publicly available at https://www.openicpsr.org/openicpsr/project/179401/version/V2/view

is also worth noting that gender is not available for all individuals in decennial censuses. In a relatively limited number of cases, gender is unknown (`U`).

5 Blocking USPTO and Census

After collecting and standardizing data on USPTO inventors and individuals from the census, the first step in the matching process involved implementing a *blocking* technique to generate *candidate matches*. This initial step is crucial for reducing the computational complexity of the matching task, which involves billions of potential pairs between the two datasets. Blocking involves partitioning the data into smaller subsets based on shared attributes such as geographic location, gender, or age so that comparisons are made only within these subsets rather than across the entire dataset.

In what follows, we describe in detail the steps taken.

Step 1: We first divided the set of inventors into three subsets based on assigned gender: female (F), male (M), and unknown (U). Similarly, we splitted individuals in the Census into gender-specific tables. This approach allows us to reduce the overall number of pairs to evaluate by matching inventors and Census individuals within the same gender groups. Specifically, male (M) inventors were matched only with male (M) individuals in the Census, and female (F) inventors only with female (F) individuals. For the relatively small number of inventors with unknown (U) gender, we matched them with both male and female individuals in the Census. This division of data enabled us to create separate matching pipelines for each gender group, thereby drastically reducing the computational complexity.

Step 2: Inventors were further divided into separate sets based on 21-year moving windows aligned with each decennial Census. For a given Census year t, we retained for matching only those inventors with patents filed within the window $t - 10$ to $t + 10$—that is, 10 years before and 10 years after the Census year.

Step 3: For each decennial Census, we retained only individuals aged between 8 and 80 at the time of the Census. Formally, for a given Census year t, we included individuals who reported an age between 8 and 80. This age-based filtering is designed to ensure that individuals could plausibly match with inventors in the surrounding 21-year window.

The rationale for this rule is that an individual who was 8 years old at time t would be 18 years old at $t + 10$, the upper end of the matching window. For example, an individual who was 8 years old in the 1860 Census could potentially match with an inventor who filed a patent in 1870, when the individual would be 18. Conversely, an individual who was 80 years old at time t would have been 70 years old at $t - 10$, the lower end of the matching window. Thus, an individual who was 80 years old in the 1860 Census could plausibly be matched with an inventor who filed a patent in 1850, when the individual would have been 70 years old.

Step 4: Finally, inventors and Census individuals were further divided into subsets based on county and state of residence, as reported in the patent document and the Census, respectively. To achieve this, we utilized FIPS codes assigned to each inventor and Census individual, as detailed in Section 4.1.[19] The purpose of this geographic blocking rule is to reduce the number of pairs to compare and thereby decrease computational complexity.

While this approach effectively reduces processing requirements, it also introduces a potential limitation: we may miss true matches if an inventor's address on the patent reflects a state different from the one reported in the Census, for example, due to a move across state lines before or after the patent was filed. Nevertheless, we believe the gains in computational efficiency outweigh this drawback.

To summarize, this blocking strategy effectively reduces computational complexity by segmenting inventors and Census individuals according to gender, age, and geographic location. In the following section, we provide a detailed overview of the specific matching algorithms applied to each of these segmented groups.

6 Identifying candidates

6.1 Matching algorithms

To match inventors with census records, we employed two primary techniques:

(1) Cosine similarity using `FAISS`

(2) Fuzzy string matching with `RapidFuzz`

[19] Note that this approach excludes inventors located outside the United States when matching with the Census (see Table 1).

These complementary approaches offer distinct methods for comparing names across the two datasets. In the following sections, we provide a detailed discussion of each technique.

6.2 Cosine Similarity with `FAISS`

The first method used FAISS (Facebook AI Similarity Search), a highly efficient tool for large-scale similarity search.[20]. For each of the segmented groups outlined above (section 5), inventor and census names were converted to lowercase, special characters were removed, and standardized (e.g., titles such as "Dr." or "Prof" were stripped).

The cleaned names were then vectorized using TF-IDF (Term Frequency-Inverse Document Frequency) with character-based 3-grams. In text processing, 3-grams are sequences of three consecutive characters extracted from a string. They capture local context and are particularly useful for comparing names where spelling variations or minor typos might exist. For example, the names `William` and `Wiliam` would result, respectively, in the following trigrams:

`wil`, `ill`, `lli`, `lia`, `iam`

`wil`, `ili`, `lia`, `iam`

Both names share three 3-grams: `wil`, `lia`, and `iam`. Despite the missing "l" in the second version, these common 3-grams would yield a higher similarity score in vectorized space. As mentioned above, this approach is suitable to the context of our data in which transcription of names through OCR can result in variations, typos, and other errors.

The vectorized names from Census and patent records were indexed in FAISS, with each inventor's name compared against Census names within the same county to identify the top 30 nearest neighbors by cosine similarity.[21] Limiting the number of potential matching candidates to the top 30 most similar individuals struck a balance between maintaining manageable computational complexity and maximizing recall.[22] This step produced 27 tables (9 Census years × 3 gender groups), each

[20] See https://ai.meta.com/tools/faiss/

[21] This method was particularly effective for large datasets, as GPU acceleration allowed for parallel similarity computations, significantly speeding up the matching process.

[22] Although the code was executed on an HPC server equipped with two GPUs, the process still required several weeks to complete. This computational complexity motivated the decision to limit the computation of cosine similarity to the county level. Extending this comparison to the

containing potential Census matches for every patent-inventor pair.

6.3 Fuzzy string matching with `RapidFuzz`

The second matching technique utilized the `RapidFuzz` library, which offers a wide range of algorithms for string matching.[23] Specifically, we implemented fuzzy string matching based on the normalized Levenshtein distance. The Levenshtein distance calculates the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into another, making it particularly suited to handling name variations, typographical errors, and other inconsistencies present in our data.

For example, consider the names (a case from our sample): `george h rodgers` and `george h rogers`. To convert the first name into the second, only one edit is required—the deletion of the `d` in `rodgers`. Thus, the normalized Levenshtein distance is calculated as follows:

$$\text{Normalized Levenshtein distance} = \frac{\text{Number of edits}}{\text{Length of longest string}} = \frac{1}{16} = 0.0625$$

In calculating the normalized Levenshtein distance, we applied the following parameters:

- A score cutoff of 0.2, meaning that only matches with a similarity score of 0.2 or lower (indicating closer matches) were retained. This threshold was chosen to ensure that only the most similar candidate matches from the Census were kept for each inventor while maximizing the possible recall rate.
- A maximum limit of 500 matches per inventor. In other words, if an inventor had more than 500 potential matches with a similarity score below 0.2, only the top 500 closest matches were retained.

The Levenshtein distance was computed for each of the segmented groups described in Section 5. Additionally, this score was calculated separately at both the state and county levels. This step produced 54 tables (9 Census years $\times$ 3 gender groups $\times$ 2, i.e. state and county), each containing potential Census matches for every patent-inventor pair

state level—or worse, to the entire country—would have dramatically increased the computational time.

[23] See https://rapidfuzz.github.io/RapidFuzz/

6.4 Initial set of matching candidates

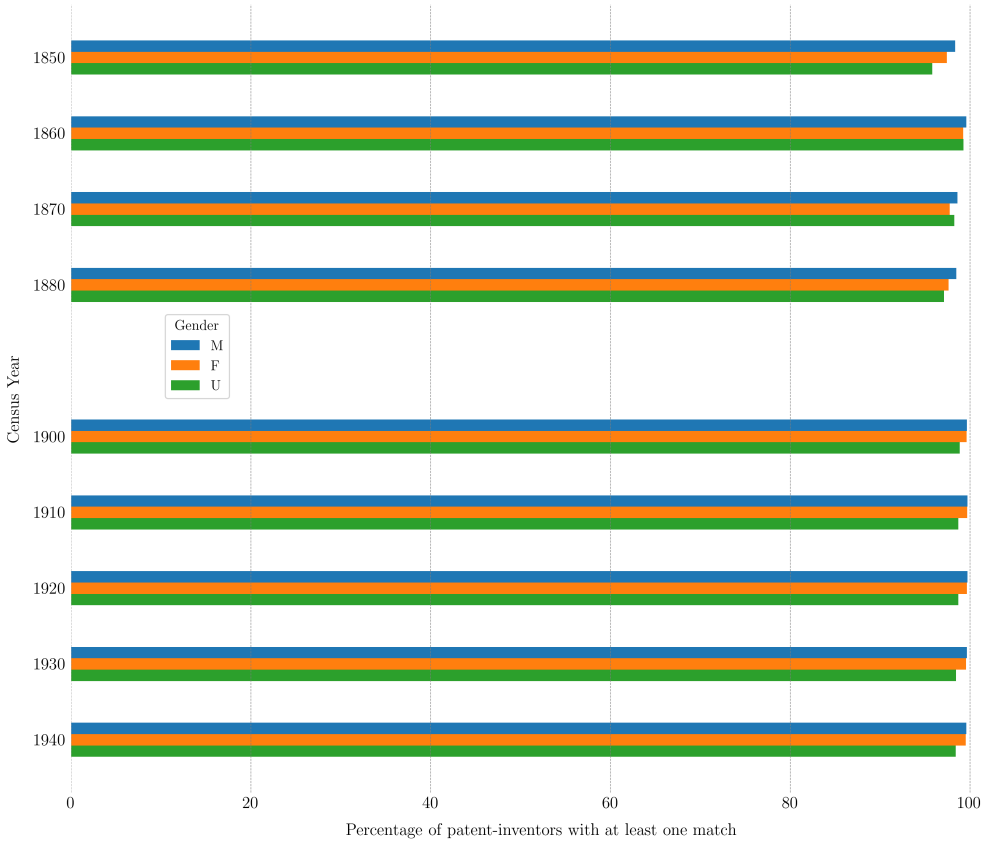
The procedure outlined above generated an initial, broad set of potential matching candidates. The metrics and cutoffs were deliberately chosen to maximize recall, even at the expense of precision. For all patent-inventor pairs where the inventor reports a FIPS code within the United States and the patent has a publication date between 1840 and 1950 (inclusive), at least one candidate match is found in the Census (see Figure 5 and Table A6 in the Appendix). Our matching strategy ensures that each inventor has some corresponding matches, consisting of the nearest neighbors based on cosine similarity (see Section 6.2) and the closest census names according to Levenshtein distance, with a cutoff of 0.2 and a maximum of 500 matches.[24]

Figure 6 presents the mean and median number of matches per inventor (left y-axis), along with a whisker plot for each Census year showing the distribution of matched historical IDs (HISTID) per inventor (right y-axis). The data on the right y-axis are log-transformed to normalize and improve visualization. This chart highlights the broad scope of our initial matching strategy, with the average number of matched individuals per inventor going from around 80 in the 1850 Census to approximately 110 in the 1940 Census. Additionally, the considerable gap between the mean and median points to substantial variability, indicating that some inventors have an exceptionally high number of matches (see also the right y-axis).

Finally, Table 4 provides a summary of the results from the initial high-recall matching stage. In total, we identified at least one candidate match in the Census for 2,546,332 patent-inventors, representing approximately 99.4% of the population of patent-inventors located in the US during the period 1840–1950. This exceptionally high matching rate reflects the design of this initial stage, which aimed to maximize recall. The primary objective at this stage was to narrow the pool of candidates to a manageable size, smaller than the entire Census population, while ensuring that any true matches would remain within the set of candidates. In the next section, we describe the filtering procedures applied to this initial set to remove the most obvious false positives.

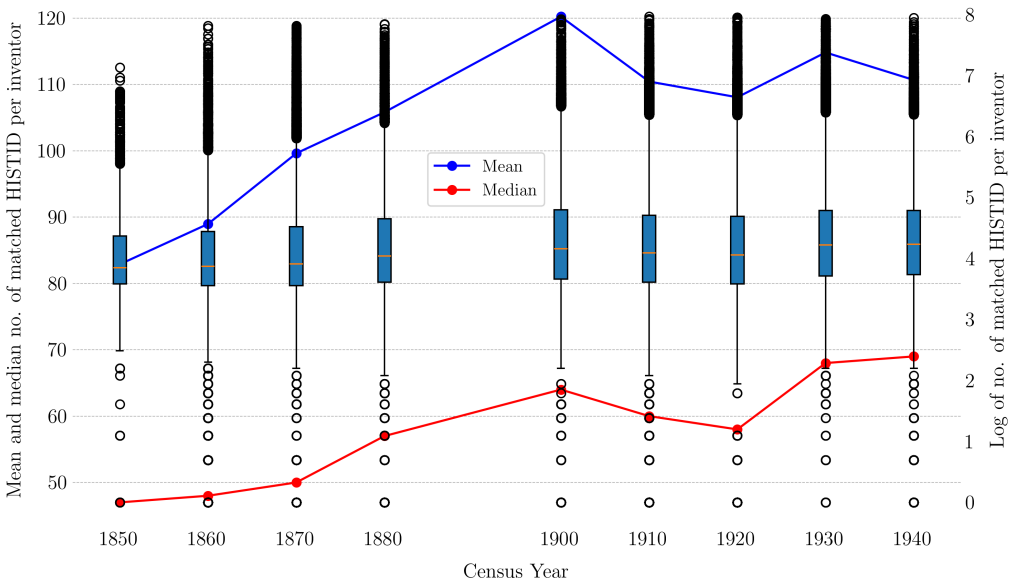
[24]It should be noted, however, that 13,517 patent-inventor pairs, despite reporting a location within the United States, did not receive a preliminary match with any Census individual. This discrepancy arises because the FIPS codes associated with these patents are absent from the Census. Consequently, due to the blocking strategy used to partition inventors and Census participants, these inventors could not be matched to any individuals.

Figure 5: Patent-inventor pairs with at least one match (%)
Initial matching (High Recall)



Note: This graph shows the percentage of patent-inventor pairs with at least one matching candidate, obtained by retrieving the top 30 nearest neighbors based on cosine similarity (Section 6.2) and the closest census names according to Levenshtein distance, with a cutoff of 0.2 and a maximum of 500 matches (Section 6.3). Table A6 reports the absolute numbers.

Figure 6: Distribution of matched Census HISTID per patent-inventor
Initial matching (High Recall)



Note: This graph displays the mean and median numbers of matched Census individuals (HISTID) per inventor on the primary (left) y-axis. The lines representing mean and median values are plotted against specific census years. On the secondary (right) y-axis, the graph presents the distribution of the number of matched HISTIDs per inventor, which has been log-transformed to normalize the data and enhance visualization. Each boxplot corresponds to a specific census year and visually summarizes the distribution for that period. The median of the distribution is represented by the central line in each box. The box itself, indicating the interquartile range (IQR), extends from the first quartile (25th percentile) to the third quartile (75th percentile). The whiskers extend from the edges of the box to the furthest points within 1.5 times the IQR from the quartiles, thus illustrating the range of the majority of the data. Data points that appear beyond the whiskers are considered outliers, emphasizing observations that deviate significantly from the general distribution.

Table 4: Initial matching stage: summary

Description	Observations
Start of process	
Number of patent-inventors (1840-1950)	3,084,304
Number of US patent-inventors (1840-1950)	2,559,849
Initial matching stage (High recall)	
US patent-inventors unmatched in any Census	13,517
US patent-inventors matched in at least one Census	2,546,332
of which:	
Matched in exactly one Census	700,303
Matched in more than one Census	1,846,029
Matched patent-inventors × Census pairs	4,527,402

Note: This table summarizes the steps undertaken in the matching process, reporting the number of patent-inventors at the initial matching stage (high recall). Due to the blocking strategy adopted (see Section 5), a single patent-inventor can potentially be matched in multiple Censuses. For instance, a patent-inventor with a publication year of 1901 could theoretically be matched with both the 1900 and 1910 Censuses. To account for this, the rows labeled "Matched patent-inventors × Census pairs" report the total number of patent-inventors multiplied by the number of Censuses in which they are matched. For example, a patent-inventor matched in only one Census is counted once, while a patent-inventor matched in two Censuses is counted twice, and so forth.

6.5 Refining the set of matching candidates

As discussed, the initial matching process was intentionally broad to maximize the recall rate, which resulted in a number of obvious false positives. To refine the initial set of potential matches, we applied a series of filters as described below.

For each inventor–Census name pair, we computed the following string distance metrics using the `RapidFuzz` library:

1. The `normalized Levenshtein` distance (see Section 6.3 for the definition). This metric ranges from 0 (identical strings) to 1 (maximum distance).
2. The `fuzz.ratio` similarity score, computed as:

$$(1 - \text{normalized Levenshtein distance}) \times 100$$

   This metric ranges from 0 (minimum similarity) to 100 (identical strings).
3. The `token sort ratio` similarity score, which is robust against differences in word order. It first splits strings into tokens (typically by spaces), sorts them alphabetically, and then re-joins them before comparing. For example, `"Doe, John"` and `"John Doe"` would both be tokenized, sorted to `["Doe,", "John"]`, and then rejoined for comparison.

4. The `Double Metaphone` similarity.[25] Double Metaphone is a phonetic algorithm designed to match similar-sounding names despite spelling differences. It returns primary and secondary phonetic codes, which are especially useful for matching names that may vary across languages or cultures. For example:

   `Catherine Zeta-Jones` and `Kathryn Zita Johns`

   The algorithm produces the codes:

   ```
   Catherine Zeta-Jones: ('KORNSTJNS', 'KTRNSTJNS')
   Kathryn Zita Johns: ('KORNSTJNS', 'KTRNTSTJNS')
   ```

   Since both names share at least one phonetic code, they are considered phonetically equivalent despite differences in spelling. This measure was implemented for all inventor–Census candidate pairs.

Using these metrics, we applied a series of sequential filters to progressively eliminate unlikely matches, focusing on name similarity and consistency to enhance the accuracy of our results.[26]

Filter 0: Remove cases where the normalized Levenshtein distance is greater than 0.2 **and** the token sort ratio is less than 90. This step removes records that are dissimilar both in direct comparison and when rearranged. While one matching algorithm used a normalized Levenshtein cutoff of 0.2, another retrieved the top 30 most similar Census names based on 3-gram cosine similarity. This filter ensures that low-similarity cases are excluded, reducing false positives.

Filter 1: If both names have three parts, check for matching middle initials. If the initials differ and consist of a single letter, the pair is removed. For example:

`John A. Smith` vs. `John B. Smith`

If the initials match or do not meet the single-letter condition, further evaluation is conducted.

Filter 2: Remove matches where both names contain only two tokens (first and last names) and the last names have a low similarity score or differ phonetically. For example:

`John Smith` vs. `John Johnson`

[25] See https://github.com/dedupeio/doublemetaphone

[26] The code and thresholds for determining name similarity are available upon request. For readability, specific threshold values are omitted here.

This filter ensures that the last names are nearly identical or phonetically consistent, helping to improve accuracy for common first names.

Filter 3: Exclude pairs where the last names are identical but the first names have a low similarity score or differ phonetically, specifically when both names have two tokens. For example:

`John Smith` vs. `James Smith`

This filter refines matches, ensuring that first names closely match or are phonetically identical.

Filter 4: Remove pairs where the inventor's name has three tokens (first name, middle initial, surname) and the Census name has two tokens, focusing on cases where the last names differ significantly. For example:

`John A. Smith` vs. `John Saratoga`

This filter targets mismatches involving middle initials and differing last names.

Filter 5: Similar to Filter 4, but applies where the inventor's name has two tokens and the Census name has three tokens (including a middle initial), focusing on cases where the last names differ significantly.

`John Smith` vs. `John A. Saratoga`

Filter 6: Remove pairs where the last names are identical but the first names are substantially different, specifically for names where the inventor name consists of three tokens (first name, middle initial, surname) and the Census name has only two tokens. For example:

`John A. Smith` vs. `James Smith`

Filter 7: Exclude matches where the last and middle names are the same, but the first names differ significantly, specifically when both names contain three tokens.

`John A. Smith` vs. `James A. Smith`

Filter 8: Remove pairs where the first name matches, but the middle initial and last names differ, applicable to names with three tokens.

`John A. Smith` vs. `John B. Johnson`

Filter 9: Filter based on the inventor's age at the time of patent publication, retaining only pairs where the inventor's age is plausible (between 18 and 80). Pairs with ages outside this range are removed.

Filter 10: For cases where first names match but middle initials differ, both consisting of a single character, the pair is removed.

`James A. [...]` vs. `James B. [...]`

Filter 11: Exclude pairs where the first two elements of each name are initials that differ between the inventor and Census records.

`J.D. Butler` vs. `E.A. Butler`

Filter 12: Finally, exclude pairs where first names differ substantially, applying a threshold of 40 for the Levenshtein distance to retain only records with similar first names.

`James Butler` vs. `John Butler`

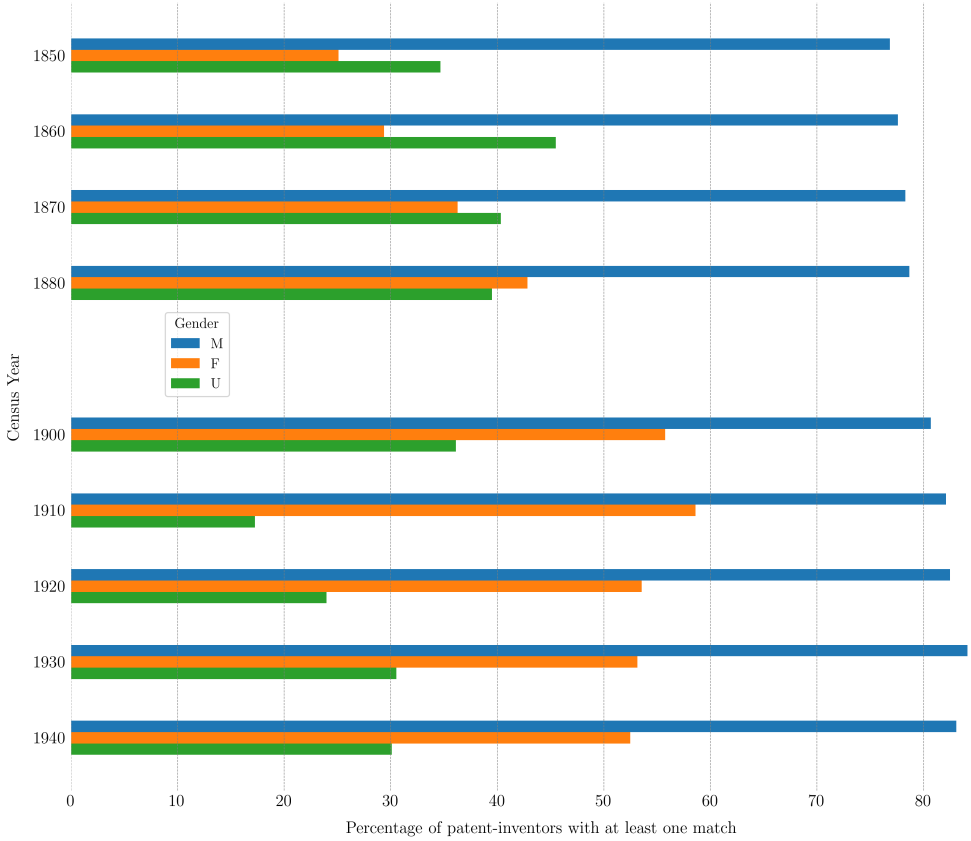
This filter reduces false positives by eliminating pairs with highly different first names.

6.6 Matching candidates after filtering

The application of the filters described above significantly reduced the number of potential matching candidates per inventor, as well as the percentage of patent-inventor pairs with at least one match (see Figure 7). Interestingly, the decrease in matched inventors was relatively moderate for male patent-inventors: the percentage of pairs with at least one match declined from approximately 97-99% before filtering to around 76-83% afterward (compare Figures 5 and 7). In contrast, the reduction was notably more pronounced for female and unknown-gender patent-inventors. This discrepancy may be attributed to several factors, including under-enumeration of females in the census, surname changes due to marriage, or name misspellings for inventors with unknown gender.

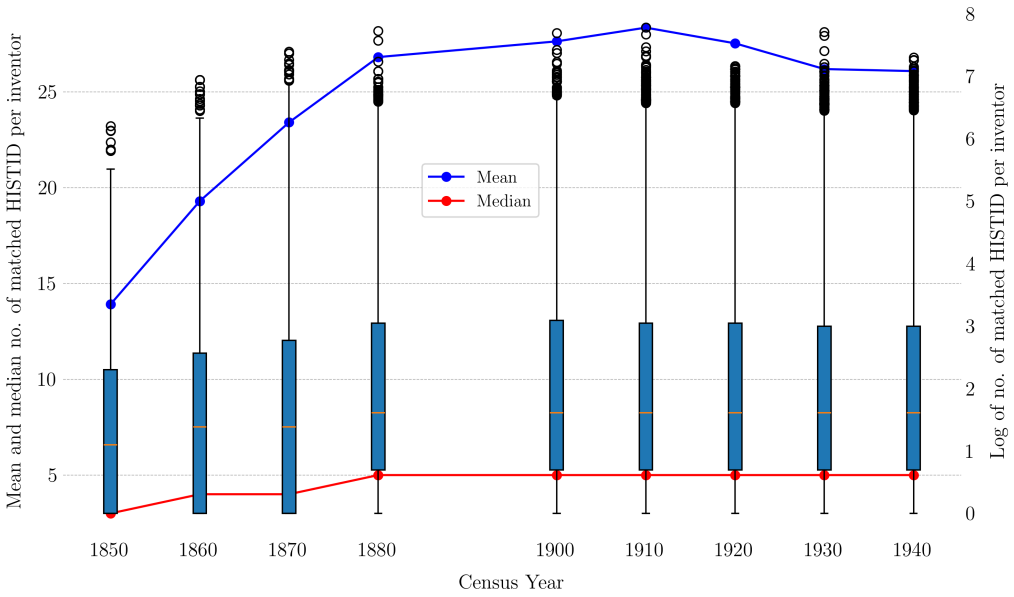
For the remaining inventors with at least one matching candidate, Figure 8 presents the mean and median number of matches per inventor (left y-axis), along with a whisker plot for each Census year showing the distribution of matched historical IDs (HISTID) per inventor (right y-axis). The data on the right y-axis are log-transformed to normalize and enhance visualization. The chart highlights the substantial reduction in the number of candidates per inventor after filtering. On average, the number of potential matches per inventor ranges from approximately 14 in 1850 to about 28 in 1910. As before, the median number of matches is considerably lower then the mean due to the presence of some inventors with a large number of potential matches, ranging from 3 in 1850 to 5 for the censuses after 1900.

Figure 7: Percentage of patent-inventor pairs with at least one match
After filtering



Note: This figure shows the percentage of patent-inventor pairs with at least one matched candidate in the Census by gender of inventors after applying the filters described in Section 6.5. Table A7 in the Appendix reports the absolute numbers.

Figure 8: Distribution of matched Census HISTID per patent-inventor
After filtering



Note: This graph displays the mean and median numbers of matched Census individuals (HISTID) per inventor on the primary (left) y-axis, after applying the filters described in Section 6.5. The lines representing mean and median values are plotted against specific census years. On the secondary (right) y-axis, the graph presents the distribution of the number of matched HISTIDs per inventor, which has been log-transformed to normalize the data and enhance visualization. Each boxplot corresponds to a specific census year and visually summarizes the distribution for that period. The median of the distribution is represented by the central line in each box. The box itself, indicating the interquartile range (IQR), extends from the first quartile (25th percentile) to the third quartile (75th percentile). The whiskers extend from the edges of the box to the furthest points within 1.5 times the IQR from the quartiles, thus illustrating the range of the majority of the data. Data points that appear beyond the whiskers are considered outliers, emphasizing observations that deviate significantly from the general distribution.

Since the filtered set of matches serves as the foundation for our ML model, we further examined the factors influencing the inclusion or exclusion of inventors from this set. Table 5 compares the characteristics of inventor names with and without at least one matching candidate after applying the filters described in Section 6.5. The table reveals some distinct patterns. Patent-inventor pairs without valid matching candidates in the Census after filtering tend to have names that are longer, contain a higher number of tokens, or begin with a single letter. Due to errors in parsing inventor names from OCRed patent documents (see Section 2.2.1), these names may result in excessively long strings or include numerous tokens, often bearing little resemblance to the actual inventor's name. Conversely, patent-inventor pairs with commonly occurring first or last names among inventors are more likely to be retained after applying the filters.

Table 5: Name features by matched status
After filtering

Variable	Matched	Mean	Std	Min	Max	Obs	T-stat
Number of characters	No	16.394	4.75	1	469	435029	184.32***
	Yes	15.026	2.58	4	37	2111303	
Number of tokens	No	2.904	0.85	1	81	435029	83.87***
	Yes	2.792	0.48	1	7	2111303	
Starts with single letter	No	0.036	0.19	0	1	435029	99.98***
	Yes	0.007	0.08	0	1	2111303	
Frequency of first name	No	19776.008	44394.04	1	182660	435029	-460.28***
	Yes	56960.022	64914.33	1	182660	2111303	
Frequency of last name	No	569.253	1893.76	1	24475	435029	-227.8***
	Yes	1402.233	3290.19	1	24475	2111303	

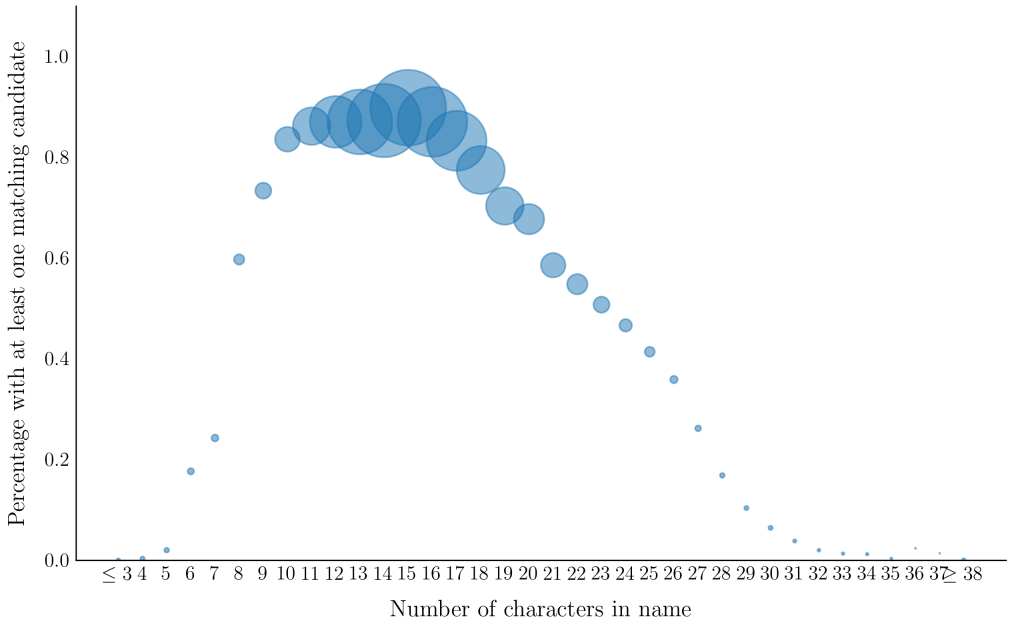
Note: The table presents a comparison of name features based on whether they were included in the matching set after applying the filters described in Section 6.5. The *Number of characters* refers to a count of all characters in the inventor's name, including spaces. The *Number of tokens* counts the number of words in the inventor's name. The variable *Starts with a single letter* is binary and takes the value one if the inventor's name begins with a single letter. The *Frequency of first name* represents the count of occurrences of the first word (or token) of the inventor's name within the inventor population. Lastly, the *Frequency of last name* represents the count of occurrences of the last word (or token) of the inventor's name within the inventor population.

As a further illustration, Figures 9a and 9b present bubble charts showing the percentage of inventors with at least one matching candidate after filtering, by name length and number of tokens, respectively. The bubble sizes in the charts are proportional to the number of inventors in each category. These charts clearly show that the probability of having a match is high (above 0.9-0.95) for names with 10 to 15 characters and with 2 to 3 tokens. This probability declines rapidly for names with fewer or more characters or tokens.

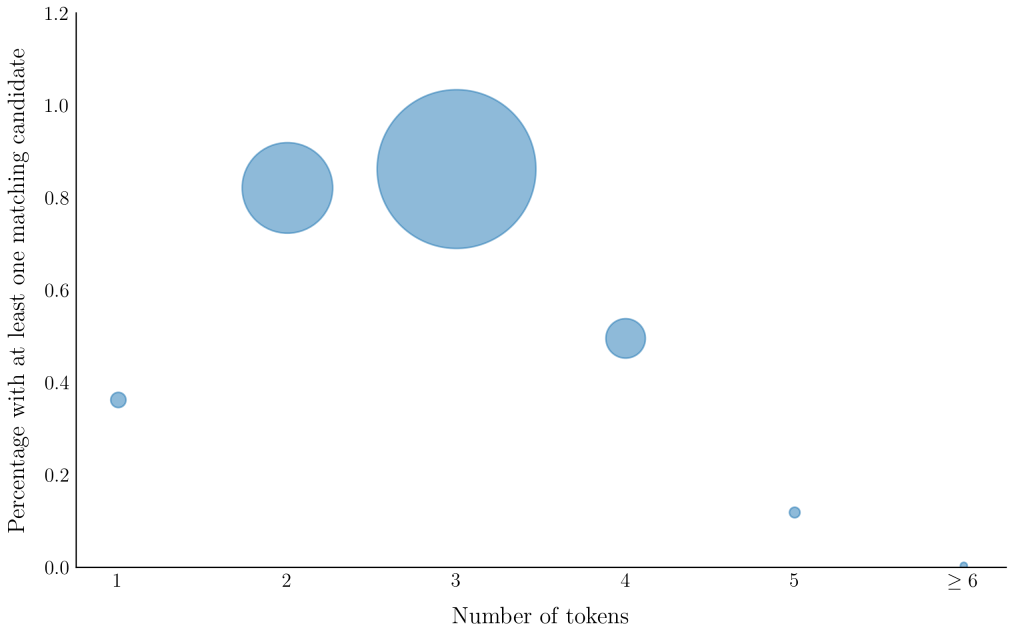
Finally, in Table 6, we present the results of a logit regression where the de-

pendent variable equals one if a patent-inventor name has at least one matching candidate after applying the filtering procedure, and zero otherwise. The results indicate an inverted U-shaped relationship between name length and the number of tokens, on the one hand, and the probability of being matched, on the other, with turning points at approximately 13 characters and 2.15 tokens, respectively. Additionally, the results confirm that names with more frequently occurring first and last names are more likely to have at least one match in the Census, while names starting with a single letter are less likely to match. Finally, the analysis shows that male inventors are significantly more likely to be matched than female or unknown-gender inventors.

Figure 9: Percentage with matching by features of name



(a) Percentage with matching by number of characters in name



(b) Percentage with matching by number of tokens in name

Note: The charts report the percentage of inventors with at least one matching candidate after the filtering process described in Section 6.5 by character (Figure 9a) and token counts (Figure 9b) within names. The size of bubbles is proportional to the number of inventors in each category.

Table 6: Probability of having at least one match
After filtering

Variable	Model 1	Model 2
Number of characters	0.545***	
	(0.004)	
Number of characters squared	-0.021***	
	(0.000)	
Number of tokens		2.792***
		(0.021)
Number of tokens squared		-0.650***
		(0.004)
Starts with single letter	-1.584***	-1.377***
	(0.015)	(0.015)
Frequency of first name (log)	0.297***	0.310***
	(0.001)	(0.001)
Frequency of last name (log)	0.247***	0.275***
	(0.001)	(0.001)
Male inventor	0.205***	0.137***
	(0.011)	(0.011)
Unknown gender	-0.023*	-0.051***
	(0.012)	(0.012)
Patents 1840-1849	-1.021***	-1.056***
	(0.033)	(0.034)
Patents 1941-1950	-0.284***	-0.230***
	(0.006)	(0.006)
Constant	-5.170***	-4.790***
	(0.034)	(0.031)
Number of Observations	2546332	2546332
Number of 1s	2111303	2111303
Number of 0s	435029	435029
Log Likelihood	-878240.00	-882950.00

Note: This table presents the results of logit regressions where the dependent variable equals 1 if a patent-inventor name is included in the refined set of inventors with at least one matching candidate in any Census after the application of filters described in Section 6.5. The regressors include the total number of characters in the inventor's name (*Number of characters*), the number of words in the inventor's name (*Number of tokens*), a binary variable set to one if the inventor's name begins with a single letter (*Starts with a single letter*), a variable representing the frequency of the first word (or token) of the inventor's name within the inventor population (*Frequency of first name*), and a variable representing the frequency of the last word (or token) of the inventor's name within the inventor population (*Frequency of last name*). Additionally, two dummy variables, $d1840$ and $d1940$, take the value one if the patent was published before 1850 or after 1940, respectively, to account for the fact that these inventors had only one Census against which to be matched.

6.7 Identifying candidates: summary

The methodology used to filter out false positives resulted in 2,111,303 patent-inventors with at least one match to the census, representing approximately 83% of the matches obtained in the pre-filtering stage and 82% of the population of patent inventors resident in the US during the period 1840-1950 (see Table 7). As shown in Figure 7, the reduction mainly affected female inventors for the reasons discussed above (e.g. name change after marriage). This procedure has allowed the pool of candidates to be narrowed down reasonably, but there is still a significant number of matches that have multiple candidates. In the next section, we describe a procedure that uses machine learning tools to build predictions and help us select from multiple candidates.

Table 7: Post-filtering stage: summary

Description	Observations
Start of process	
Number of patent-inventors (1840-1950)	3,084,304
Number of US patent-inventors (1840-1950)	2,559,849
Initial matching stage (High recall)	
US patent-inventors unmatched in any Census	13,517
US patent-inventors matched in at least one Census	2,546,332
of which:	
Matched in exactly one Census	700,303
Matched in more than one Census	1,846,029
Matched patent-inventors × Census pairs	4,527,402
Post-filtering stage	
US patent-inventors unmatched in any Census	435,029
US patent-inventors matched in at least one Census	2,111,303
of which:	
Matched in exactly one Census	825,829
Matched in more than one Census	1,285,474
Matched patent-inventors × Census pairs	3,500,614

Note: This table summarizes the steps undertaken in the matching process, reporting the number of patent-inventors at each stage. Due to the blocking strategy adopted (see Section 5), a single patent-inventor can potentially be matched in multiple Censuses. For instance, a patent-inventor with a publication year of 1901 could theoretically be matched with both the 1900 and 1910 Censuses. To account for this, the rows labeled "Matched patent-inventors × Census pairs" report the total number of patent-inventors multiplied by the number of Censuses in which they are matched. For example, a patent-inventor matched in only one Census is counted once, while a patent-inventor matched in two Censuses is counted twice, and so forth.

7 Machine learning

In this section we present the machine learning model used to select between candidate matches. Specifically, we illustrate the main stages of this process, beginning with the construction of a manually curated dataset that will be used to train the ML model (Section 7.1). After presenting the characteristics of the ML model and the features used to predict matches (Section 7.2), we test the performance of the model (Section 7.4) and perform a series of robustness checks (Section 9).

7.1 Developing a ground truth dataset

To train our machine learning model to identify the most likely true match, we manually curated a subset of patent-inventor-census matches. Ensuring the accuracy of this subset is critical, as false matches could negatively impact the training of the model. To achieve high confidence in the accuracy of these matches, we gathered additional inventor information from two sources: the 1906 and 1921 edition of American Men of Science (Cattell 1906; Cattell and Brimhall 1921), and 19th and 20th century American inventors listed on Wikipedia (Wikipedia, 2024a; Wikipedia, 2024b).

7.1.1 American Men of Science

The American Men of Science (AMS) is a biographical reference work on notable scientists in the United States and Canada, published by Gale. It includes professionals from fields such as physical and biological sciences, engineering, mathematics, statistics, computer science, and public health. We randomly selected 500 scientists from AMS who had complete information on first and middle names, surnames, date and place of birth, and city and state of residence. Where available, we also collected additional details, including field of study, educational background, degrees, and positions held with corresponding dates. We identified all patents filed by these scientists published between 1911 and 1929, retaining only those for which we had high confidence.

7.1.2 Wikipedia pages on American inventors

Wikipedia offers category pages that organize links to individuals or topics within specific categories, including dedicated pages for 19th- and 20th-century American inventors. For each inventor listed on these pages, we scraped their personal Wikipedia page to collect relevant biographical details and references to patent pub-

lications. Biographical information is generally well-structured within the ”infobox” on the right side of a Wikipedia page. It includes a person's full name, date and place of birth and death, and occasionally details about their occupation, field of study, or affiliated companies are provided. Inventors for which we were unable to extract information on place or date of birth were discarded.

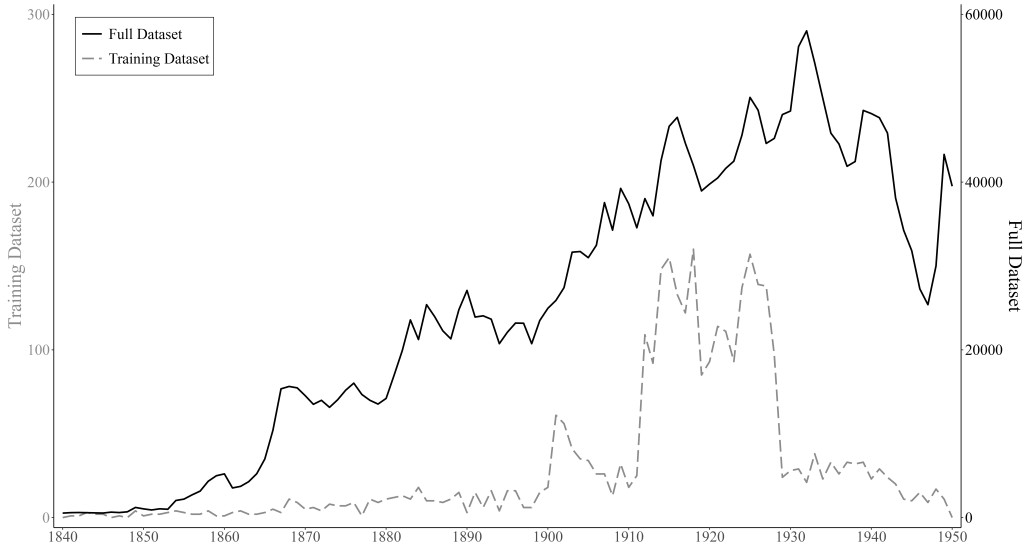
Identifying references to an inventor's patented inventions (if applicable) was less straightforward, as patent publication numbers are not listed in a standardized way on Wikipedia. Therefore, we searched the full text for hyperlinks to Google Patents or phrases such as “patent no.” or “patent US” followed by a number, to identify patent publication numbers. When patent number(s) were identified, we linked the Wikipedia inventor to our patent database, while making sure that the patent-inventor's name closely matched the name of the Wikipedia inventor. If an inventor's page did not mention a patent number, we manually searched for relevant patent publications using the inventor's name, biographical information, and field of expertise. Wikipedia inventors without any patent references after this search were excluded from the training dataset, as well as patents published before 1840 or after 1950.

7.1.3 Ground truth vs. full patent dataset

The retrieval of inventors from AMS and Wikipedia led to the construction of a ground truth dataset of 4,595 patent-inventor-census matches in total, where a patent-inventor pair have been matched where possible to multiple histids in different census editions. Thus in total 2,808 distinct patent-inventor pairs were successfully matched to the census.

Figure 10 displays the temporal distribution of the ground truth dataset compared to the full patent dataset. As shown, the ground truth dataset places slightly more emphasis on the period between 1910 and 1930, due to the inclusion of the AMS. The emphasis on this specific period could to some extent explain the observed differences in distributions across technological classes between the two datasets, as illustrated in Figure 11. The full dataset seems to have higher shares of patents in relatively older technological classes, such as Mechanical or Apparel & Textile, whereas the ground truth dataset seem to have higher shares of patents in more modern technologies, such as Electrical & Electronic or Computers & Communications.

Figure 10: Temporal distribution of the ground truth dataset and the full patent dataset



Note: Figure 10 reports the number of patents published per year for the full dataset (right y-axis) and those included in the ground truth dataset (left y-axis.

7.1.4 Training dataset

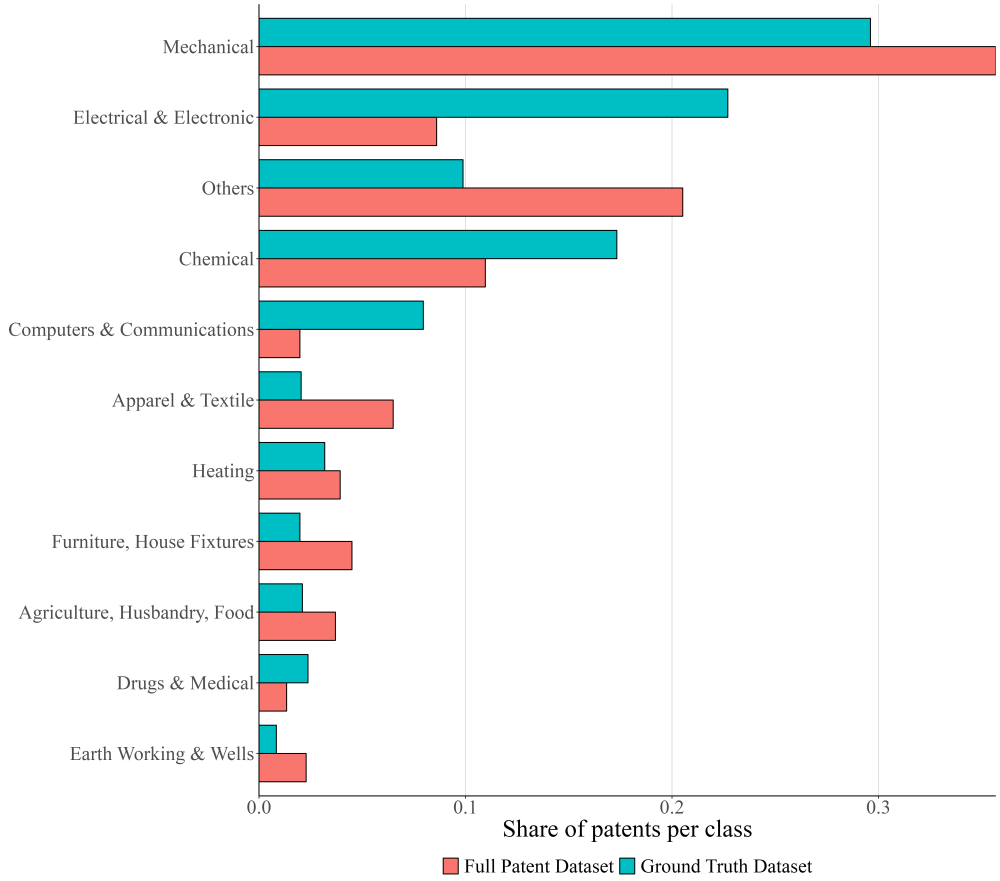
Currently, the ground truth dataset contains only true matches. However, to use it as a training dataset for the machine learning model, it also needs to include false matches. Specifically, for each patent-inventor in the ground truth dataset, we add the census candidates identified using the procedure outlined in Section 6.

For 879 patent-inventors in the ground truth dataset, either no candidates remained after filtering, or the correct match was not included among the candidates. These cases were excluded from the training dataset. As a result, the final training dataset comprises 1,929 true matches and 96,478 false matches.

7.2 ML models

In this section we illustrate the characteristics of the ML model and its performance. We referred to the Scikit-Learn model selection guide available at `https://scikit-learn.org/stable/machine_learning_map.html` to determine the most suitable machine-learning model. Given the size of our dataset and the predictive task, we tested both regression-based models (XGBoost and Random Forest) and a classification model (SVC). This approach is appropriate as our task can be viewed both as a classification problem (match vs. non-match) and as a problem of estimating the probability of a match.

Figure 11: Distribution across technological classes of the ground truth dataset and the full patent dataset



Note: Figure 11 shows the share of patents per technological class for both the ground truth and full patent dataset.

Our problem mainly involves $1:N$ type linkages, where identifying the highest probability match is crucial. The Random Forest model was particularly well-suited, as it provides probabilities for potential matches between Census IDs (HISTID) and patent-inventor IDs (PatInvID).

To validate the model selection, we used the `TPOT` Python library (Olson et al. 2016), which confirmed the suitability of the Random Forest model for estimating match probabilities in our dataset. The Random Forest model has also characteristics that are very relevant for our exercise:

1. **Robustness**: It handles outliers and nonlinear relationships well.
2. **Interpretability**: It provides insights into which features are most important.
3. **Performance**: The ensemble technique often results in strong predictive performance.

In our analysis we used 22 features that can be roughly grouped in four categories (see more details in Table 8). First, some of the features relate to measures of string similarity between inventor and census names, namely: the normalized Levenshtein distance, the token sort ratio, and the interaction between the two.

Second, we included geographical features, such as the distance in kilometers between the centroids of the counties of the inventor and census respondents' residences. Third, characteristics of the inventor, such as the inventor's age at the time of patent publication, defined as the difference between the patent's publication year and the birth year reported in the Census.

Finally, we added features that can describe the activity of the inventor. To this aim we created 14 occupational groups, by using the `occ1950`, the `ind1950`, and the `occstr` variables available in the Census. We defined occupational categories that could be easily linked to broad technological classes (i.e. mechanical, chemical, medical), or occupation for which we assumed they are likely held by patent applicants (i.e. engineer, inventor, manager). The definition of these occupational categories is reported in the Appendix in Table A8.

To refine the assignment of occupations to certain categories, we created both broad and specific versions of the occupational categories. The broad version included all occupations with even a slight connection to a category, while the specific version applied stricter criteria for inclusion. We evaluated the model's performance with both versions and ultimately selected the broad version based on its superior results. Additionally, using a subset of these occupational features, we created a binary feature that indicates whether the technological class of the patent is related

Table 8: Features included in the ML model

Feature	Description
`diff_length`	Difference in length between census and inventor names
`leven_simil`	Levenshtein similarity score between census and inventor names
`inter`	Product of Levenshtein similarity and `token sort ratio`
`tkratio`	Token ratio (`Rapidfuzzy`)
`tksortratio`	Token sort ratio (`Rapidfuzzy`)
`county_distance`	Geographical distance between inventor's and census counties
`age_at_pat`	Age of the individual at the time of patent
`occ` × `tech`	Indicator if occupation is related to patent technology
Occupational features (1 if related, 0 otherwise):	
`agriculture`	Agriculture field
`apparel`	Apparel industry
`chemistry`	Chemistry field
`civil_eng`	Civil engineering
`comm_comp`	Commercial computing
`electricity`	Electrical engineering
`mechanical`	Mechanical engineering
`medical`	Medical field
`engineering`	General engineering
`physics`	Physics field
`science`	Scientific research
`inventors`	Inventive professions
`managers`	Management occupations
`manufacturing`	Manufacturing industry

Note: This table reports the features used in the ML models and their descriptions. For the definition of occupational variables, see Table A8 in the Appendix. To define the occupational features, we employed the following variables from the Census: `ind1950`, the industry where the person was employed, `occ1950`, the occupational code of the person, and `occstr`, the original unedited occupational string entry from the census manuscript. See, https://usa.ipums.org/usa-action/variables/group/hist_tech fore details.

to the occupation of the matched census candidate.

For each patent-inventor (`PatInvID`), we selected the individual in the Census (`histid`) with the highest probability of match. We only matched records with a probability greater than a specified threshold, which is detailed in Table 9. The choice of this threshold was strategically determined with the objective of creating two versions of the matched data: one emphasizing recall rate (High Recall) and the other emphasizing precision (High Precision). The former aims to capture as many true positives as possible, thus reducing the risk of missing relevant matches, while the latter aims to ensure that the matches made are highly likely to be correct, thus minimizing the inclusion of false positives. This dual approach allows us to assess the trade-offs between capturing all potential matches and ensuring the accuracy of those

matches, which is crucial in fields where the consequences of false identifications can be significant. By employing this dual-threshold strategy, we effectively tailor the matching process to meet diverse analytical needs and stakeholder requirements, ensuring a flexible and robust approach to patent-inventor linkage. In the following sections, we will focus exclusively on presenting statistics for the "High Recall" scenario. For training purposes we used 80% of the sample, keeping 20% of the records for testing the results.

Finally, we paid particular attention to the tuning of hyperparameters, a crucial step for enhancing the performance of machine learning models. For the Random Forest model, key hyperparameters include the number of trees (`RFC_estimators`), the maximum depth of each tree (`RFC_depth`), the function to measure the quality of a split (`RFC_criterion`), and the number of features to consider when looking for the best split (`RFC_features`). We conducted a randomized search to optimize these hyperparameters and subsequently fine-tuned them manually. The parameters for each of the two models, High Recall and High Precision, are summarized in Table 9.

Table 9: Model parameters: high precision and high recall scenarios

Parameter	**High precision**	**High recall**
ML model	Random forest	Random forest
Probability threshold	0.5	0.2
RFC estimators	800	300
RFC criterion	Entropy	Entropy
RFC features	15	13
RFC depth	20	20

Note: This table reports the parameters used in the ML models for scenarios focusing on high precision and high recall, respectively. The *Probability threshold* represents the threshold for making a decision, such as whether to accept or reject a match. *RFC estimators* refer to the number of trees in the forest; higher numbers generally lead to better performance but require more computation. *RFC criterion* indicates the function used to measure the quality of a split. *RFC features* denotes the number of features considered when looking for the best split. *RFC depth* limits the maximum depth of the trees, preventing overfitting by restricting how complex the models can become.

7.3 ML matching summary

Table 10 reports a summary of the ML model. As far as the post-ML model application is concerned, there is a significant number of unmatched patent-inventors—980,734 out of a total of 2,111,303—highlighting a substantial reduction in the pool of patent-inventors matched to census records when compared to the initial and

post-filtering stages. This underscores the stringency of the ML model in refining matches to ensure higher reliability and accuracy.

Of the patent-inventors matched to at least one census, 788,959 were matched in exactly one census, while a smaller number, 341,610, were matched in more than one census. This distribution suggests that the ML model effectively minimized redundant or less certain matches, thereby enhancing the precision of single census linkages.

The total number of matched pairs, 1,508,780, also indicates a focused and stringent selection by the ML model, which likely emphasizes precision and relevance of the linkages over mere quantity. This reduction in total matched pairs, from over 3.5 million in the post-filtering stage to about 1.5 million, illustrates a significant tightening in matching criteria.

7.4 ML model performance

The performance of the machine learning model, as configured for both High Precision and High Recall scenarios, is reported in Table 11. This is crucial for validating the effectiveness of the matching process.

The High Precision model achieved a precision of 0.899 and a recall of 0.741, with an F1 score of 0.813. These metrics indicate that the model is highly effective in ensuring that the identified matches are likely correct but at the cost of missing a quite significant number of potential true matches (lower recall). The detailed classification report reported in the Appendix, Table A12 reinforces this, showing that for class 1.0 (true matches), the precision is very high at 0.899, which means that when a match is predicted, it is very likely to be a true match. However, the recall for this class is lower, suggesting that the model fails to capture all true matches (see also, the confusion matrix reported in Table A10 in the Appendix).

Conversely, the High Recall model shows a different trade-off, prioritizing the capture of true matches as indicated by a recall of 0.907 and a precision of 0.805, leading to an F1 score of 0.853. This model configuration aims to minimize the risk of missing true matches but accepts a higher rate of false positives, as evidenced by the slightly lower precision. The classification report for the High Recall scenario highlights an excellent ability to identify true matches (high recall for class 1.0), which is crucial in scenarios where missing a true match carries significant consequences (see Tables A11 and A9 in the Appendix).

Table 10: Results of ML matching: summary

Description	Observations
Start of process	
Number of patent-inventors (1840-1950)	3,084,304
Number of US patent-inventors (1840-1950)	2,559,849
Initial matching stage (High recall)	
US patent-inventors umatched in any Census	13,517
US patent-inventors matched in at least one Census	2,546,332
of which:	
Matched in exactly one Census	700,303
Matched in more than one Census	1,846,029
Matched patent-inventors × Census pairs	4,527,402
Post-filtering stage	
US patent-inventors unmatched in any Census	435,029
US patent-inventors matched in at least one Census	2,111,303
of which:	
Matched in exactly one Census	825,829
Matched in more than one Census	1,285,474
Matched patent-inventors × Census pairs	3,500,614
After the ML model	
US patent-inventors unmatched in any Census	980,734
US patent-inventors matched in at least one Census	1,130,569
of which:	
Matched in exactly one Census	788,959
Matched in more than one Census	341,610
Matched patent-inventors × Census pairs	1,508,780

Note: This table presents the number of patent-inventors matched with Census HISTIDs following the application of the ML model. Due to the blocking strategy employed (see Section 5), a single patent-inventor can be matched to individuals in multiple Censuses, although within each Census, they are matched to only one individual. For instance, a patent-inventor with a publication year of 1901 could be matched to an individual in the 1900 Census and another in the 1910 Census. To account for this, the rows labeled "Matched patent-inventors × Census pairs" report the total number of patent-inventors multiplied by the number of Censuses in which they are matched. For example, a patent-inventor matched in one Census is counted once, while a patent-inventor matched in two Censuses is counted twice, and so on.

7.5 Match rates

Figure 12 shows the match rates by decennial Census year from 1850 to 1940. The match rate ranges from a low of 28.7% in 1860 to a high of 45.2% in 1940. Several factors contribute to this wide variation, with two primary explanations: potential biases in the matching procedure or issues related to census data collection. Regarding the latter, it is well-documented that census enumeration methods were inconsistent during this period, with significant developments over time that influenced the quality of the data collected in each census. Interestingly, the trend in match rates, as shown in Figure 12, aligns well with certain historical events and

Table 11: Summary of model performance

Metric	High precision	High recall
Precision	0.899	0.805
Recall	0.741	0.907
F1 Score	0.813	0.853

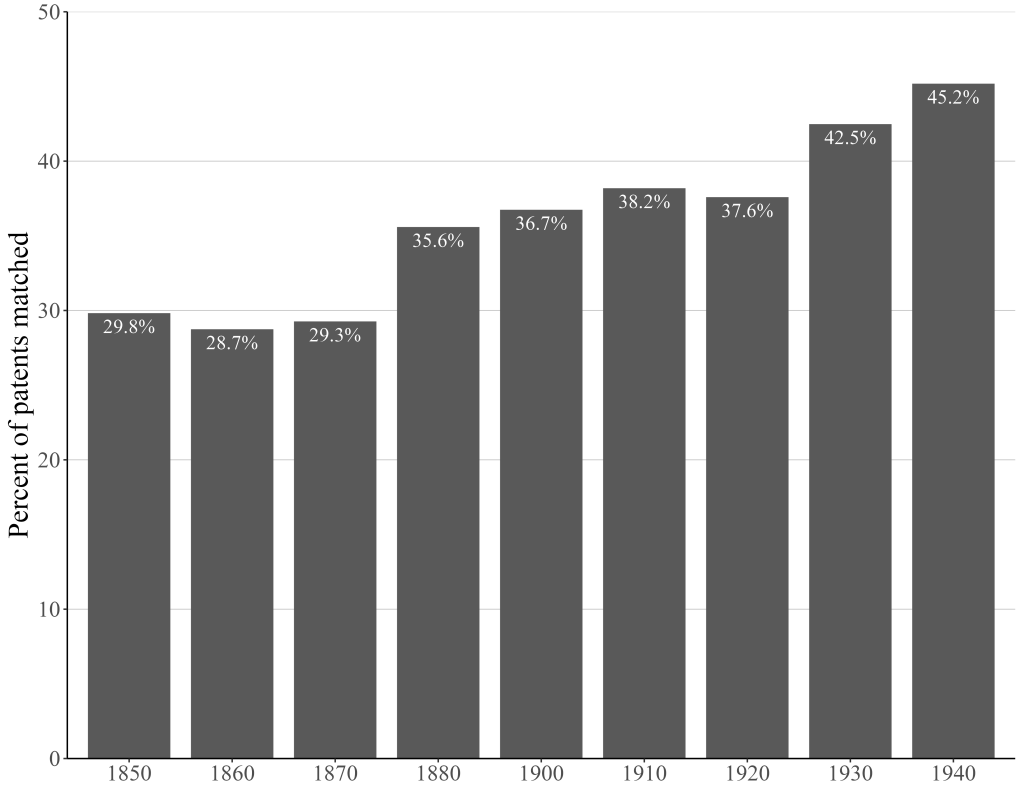
Note: This table displays metrics of model performance, including Precision, Recall, and F1 Score. Precision measures the accuracy of positive predictions, Recall assesses how well the model can identify all relevant cases, and the F1 Score balances the trade-offs between Precision and Recall. The first column is optimized for higher Precision, while the second column is focused on maximizing Recall. Higher scores, closer to 1.0, indicate better performance across these metrics, demonstrating the model's efficacy in distinguishing and retrieving relevant instances.

procedural changes in census enumeration. For instance, the match rate remains relatively stable between 1850 and 1870 but rises notably in 1880—a year often regarded as a "turning point" in census enumeration practices (King and Magnuson 1995, p. 28). Although subsequent improvements were less profound, the enumeration process continued to evolve from 1890 to 1940, with refinements introduced in each new census that likely contributed to a steady increase in match rates. In addition, the lack of an increase in the match rate in the 1920 census may be due to specific challenges during that census wave. As noted by Akcigit, Grigsby, and Nicholas (2017a), the 1920 census was conducted in winter, which may have led to an undercount. In addition, the end of the First World War and its impact on the mobility of individuals may have further affected the match rate for that year. In support of this interpretation, we note that Hartog et al. (2024) report a similar jump in 1880 and a more stable trend around 1920.

While the overall trend in match rates across years aligns well with findings from other studies, the percentage of patent inventors we match with the census population differs from some prior work. In particular, compared to Hartog et al. (2024) we tend to have lower match rates. For example, for the census years 1900 to 1930 their match rate is around 50%, while ours is close to 40%. Instead, our rates are closer to those found by Akcigit, Grigsby, and Nicholas (2017b), who report an average match rate of about 40% for the census years 1880 to 1910.

Figure 13 shows the match rates by state. As in Akcigit, Grigsby, and Nicholas (2017b) we observe considerable heterogeneity across states. At first glance, we see that match rates tend to be lower in predominantly rural areas (e.g., the Midwest), while coastal areas such as California or Massachusetts have significantly higher match rates. This may partly reflect the fact that patents are relatively scarce in

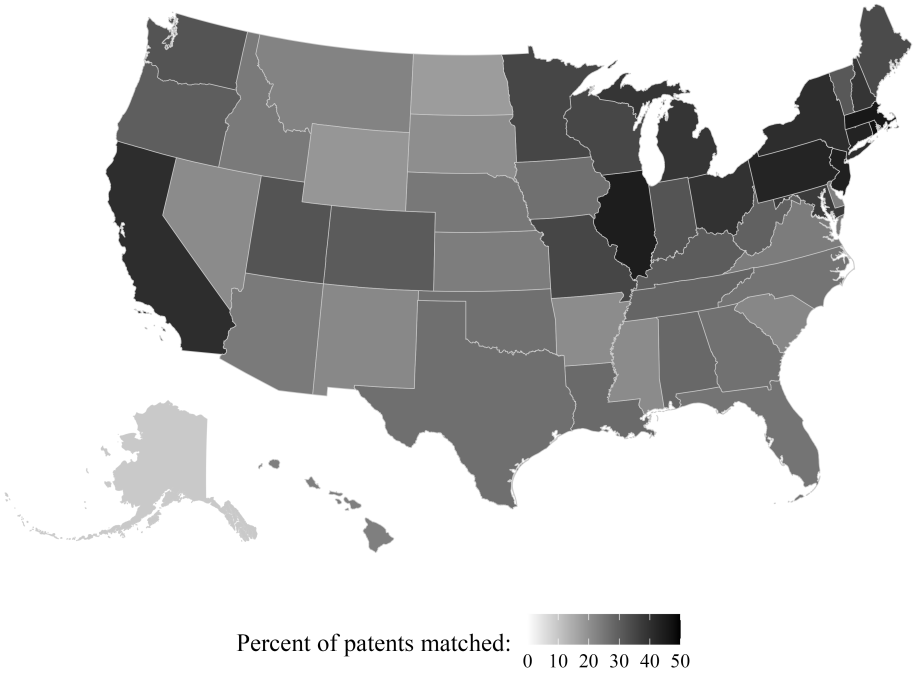
Figure 12: Match rate by census wave



Note: Figure 12 reports for each census year the percentage of patents that were matched to an individual in the census.

these rural areas compared to the others.

Figure 13: Match rate by state



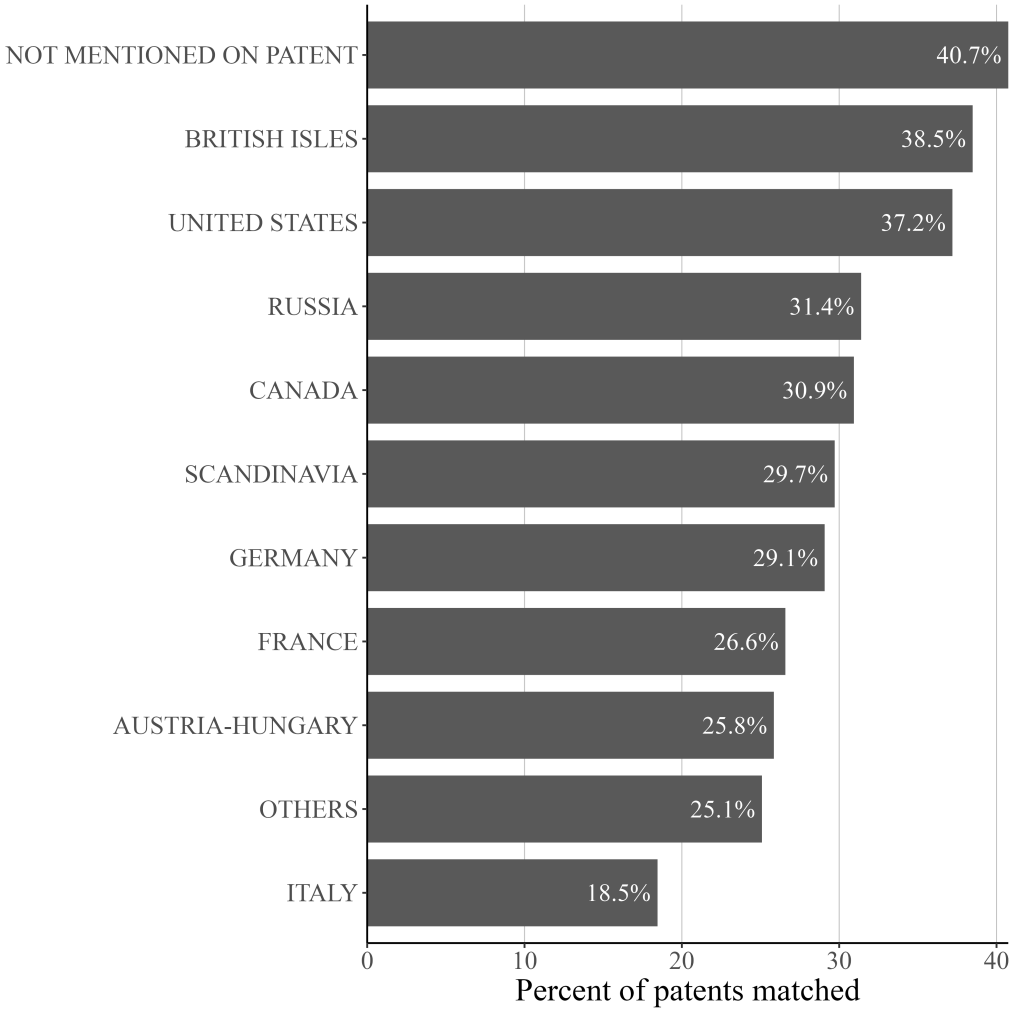
Note: Figure 13 reports for each state the percentage of patents that were matched by compiling all years.

Figure 14 shows the match rate by the origin of the inventors. Match rates are higher for Anglo-Saxon citizens, possibly reflecting a bias towards this group, as they were more likely to have popular names, as also found by Akcigit, Grigsby, and Nicholas (2017b). Other differences, such as the lower match rates for German names as compared to Russian names, may indicate a bias towards ethnic groups that were more likely to live in urban areas.

Figure 15 presents the match rates according to the distance between the publication year of the patent and the census year to which the patent was matched. We find that the closer the distance, the higher the match rate. This finding may suggest that inventors were mobile, so the further away from the date of publication of the patent, the more likely we are not match the patent because they changed their home address.

Finally Figure 16 shows the match rates according to the technological classes in which the inventor patented. Ideally, in this case we would like to see an even distribution, as there should be no particular reason why match rates should be

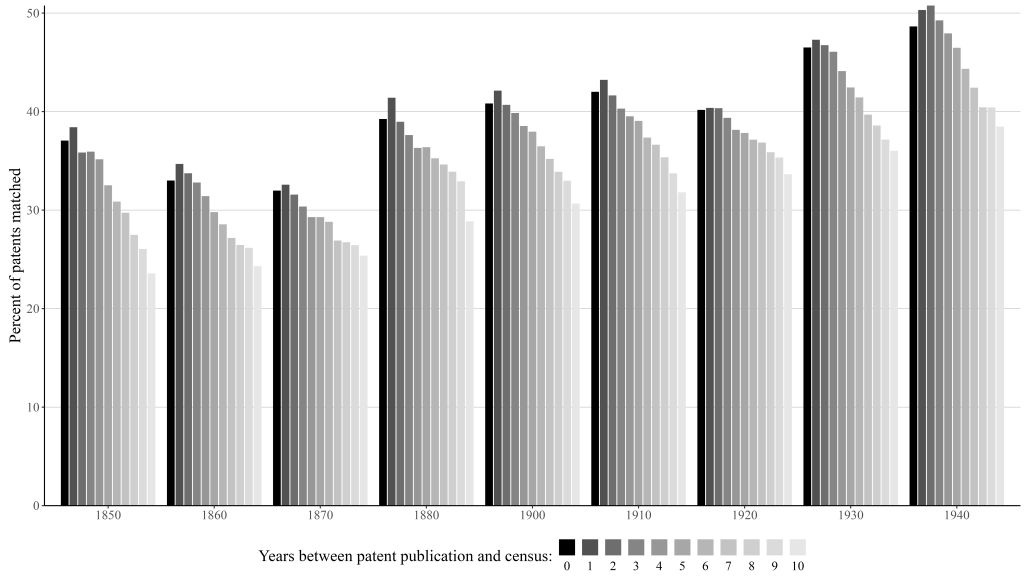
Figure 14: Match rate by nationality of inventor



Note: Figure 14 reports the match rate by the nationality of the inventor. Here we use the nationality reported on the patent document. Information on nationality is mainly available for patents published before 1926, thereafter it was rarely mentioned.

higher for certain technologies than for others. The findings show some variability between classes (from a low of 45,8% to 63.4%), though many are in a smaller range. Differences may reflect the over-representation of new technological classes (e.g. communication),as also found in the ground truth.

Figure 15: Match rate by nr. of years between patent grant year and census year



Note: Figure 15 reports the match rate by the number of years between patent publication and the census year it was matched to.

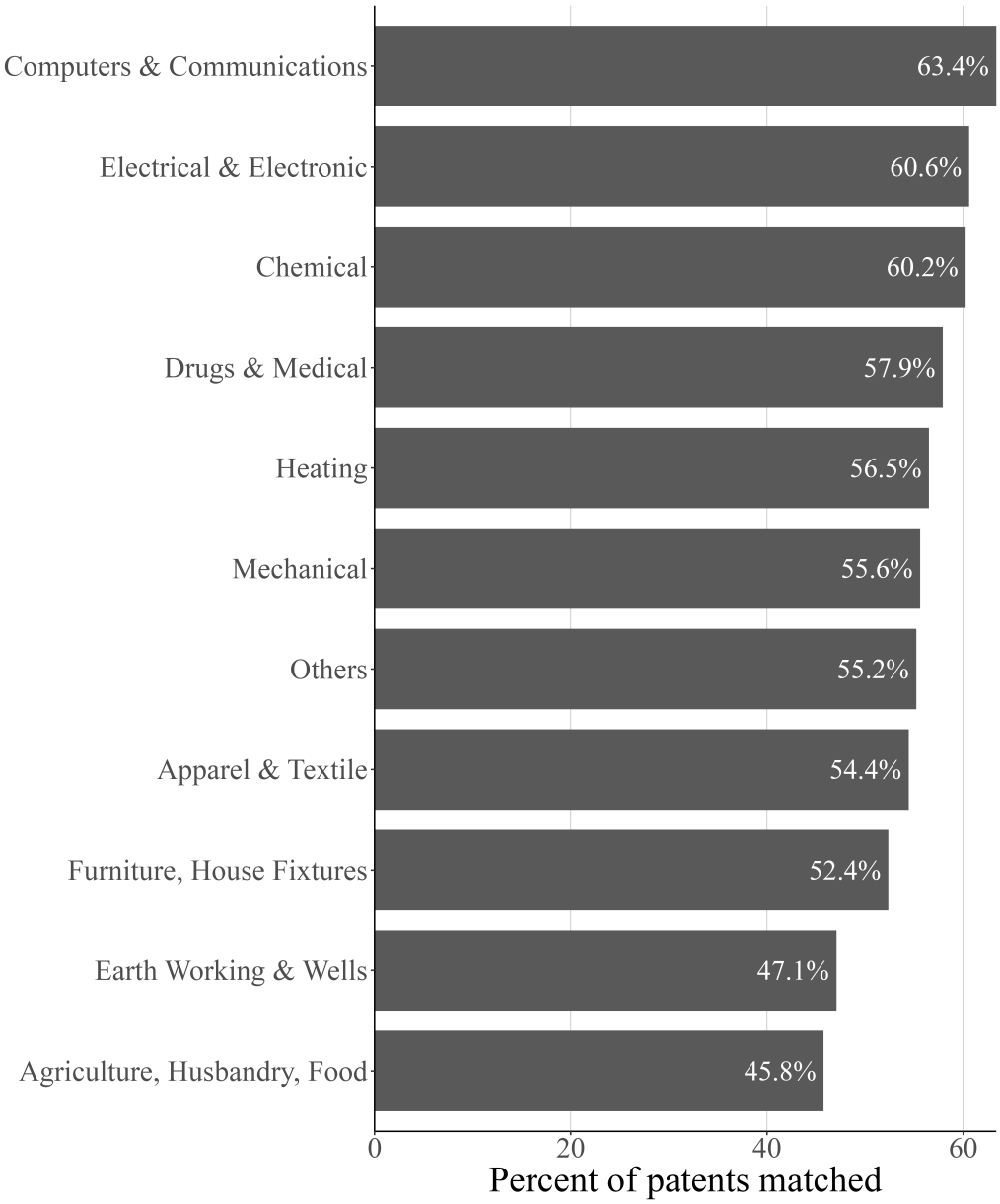
8 Inventor demographics

In this section we present some demographics of patentees. We compare them with the general census and also with findings of previous studies. Our exercise serves primarily as a validation of our matching process. If inventor demographics had closely mirrored those of the general population, it would have indicated that our procedure was essentially assigning inventors to individuals at random.

Table 12 presents the general characteristics of a typical inventor in the early 19th century compared to the average person in the 1900 census[27]. A typical inventor was a white male born in the U.S., though a significant portion—approximately 25% —were foreign-born. Since the immigrant population was about 14% in the total population, this result indicates that immigrants were overrepresented in the inventor population. The most prominent immigrant groups were British and German, each accounting for about 6% of the inventor population. These findings align

[27]The statistics are computed using the patents matched to the 1900 census

Figure 16: Match rate by technological class



Note: Figure 16 shows the match rate for each technological class, i.e. the percentage of patents matched to at least one census.

with the results of Akcigit, Grigsby, and Nicholas (2017b) and Sarada, Andrews, and Ziebarth (2019).

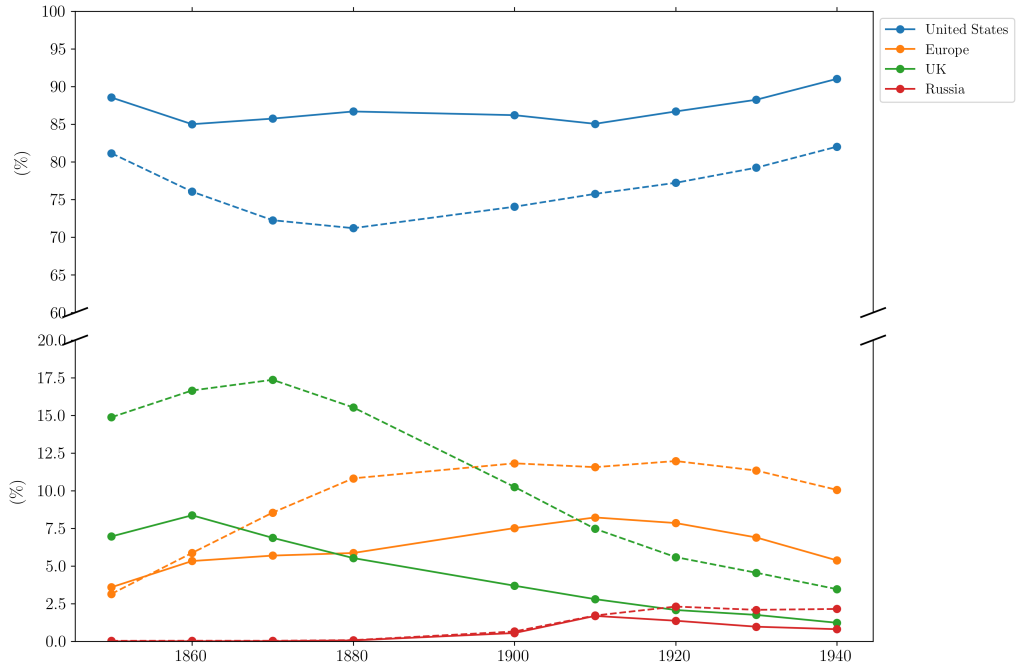
Figure 17 illustrates the distribution of inventors' nationalities compared to the census population. The solid line represents the census data, while the dotted line corresponds to inventors. The two lines exhibit similar, nearly parallel trends over the period. However, each nationality shows distinct dynamics reflecting well-documented migration patterns. For instance, Russian migration increased toward the end of the 19th century, coinciding with a decline in British inflows. These patterns suggest, as noted in previous studies, that inventors tended to follow the migration patterns of their respective ethnic groups.(Diodato, Morrison, and Petralia (2021))

Table 12: Characteristics of Inventors (1900 Census)

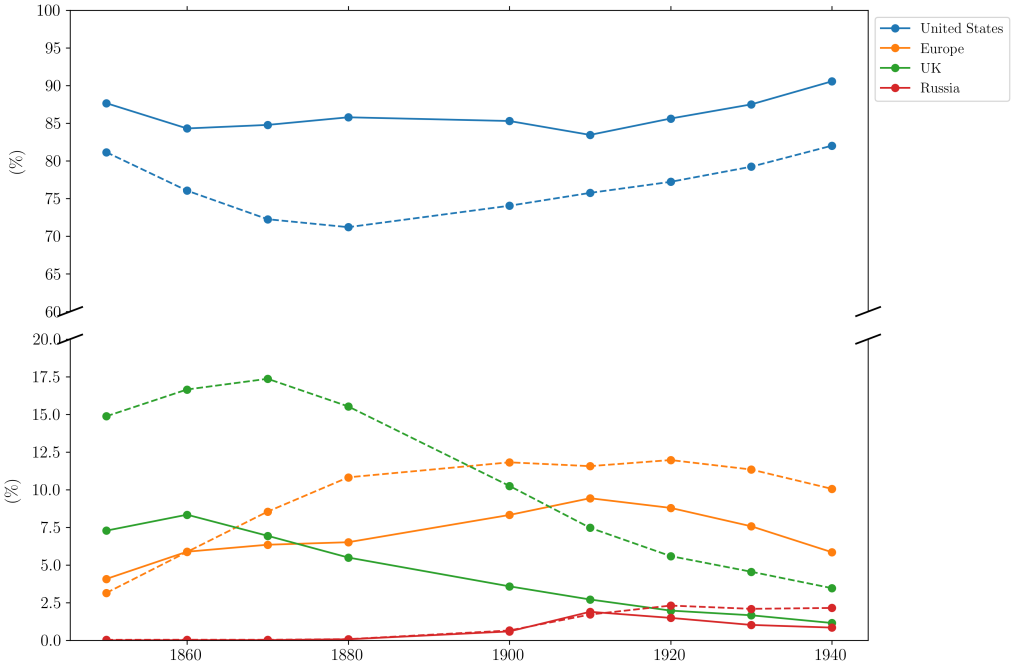
	Inventors	**Full U.S.**
Percent of Population	0.11%	99.89%
Percent Male	98.24%	51.08%
Percent White	98.23%	87.91%
Percent Native Born	74.05%	86.39%
Percent Foreign Born	25.95%	13.61%
Percent Born in Canada	3.04%	1.55%
Percent Born in Great Britain	6.38%	1.54%
Percent Born in Germany	6.77%	3.50%

Note: Statistics are computed using all inventors matched to the 1900 census and the full population in the 1900 census.

Figure 17: Distribution by birthplace



(a) Population vs. inventors

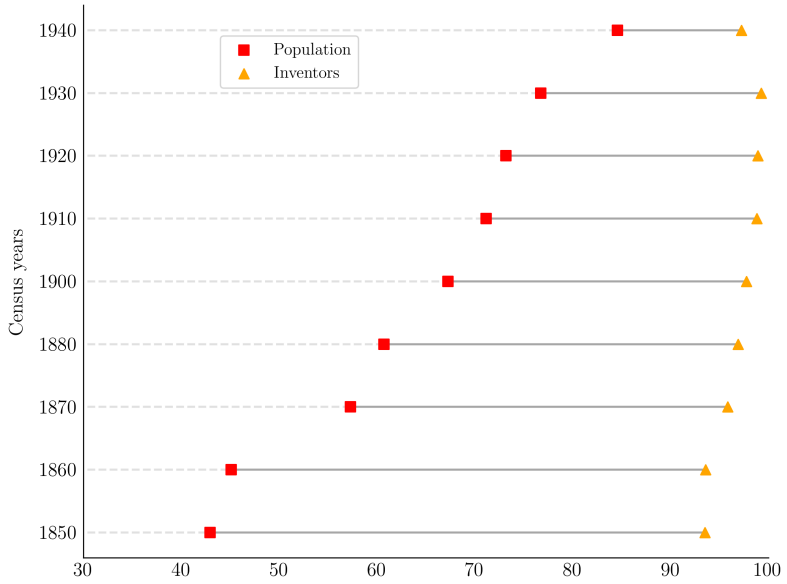


(b) Male population vs inventors

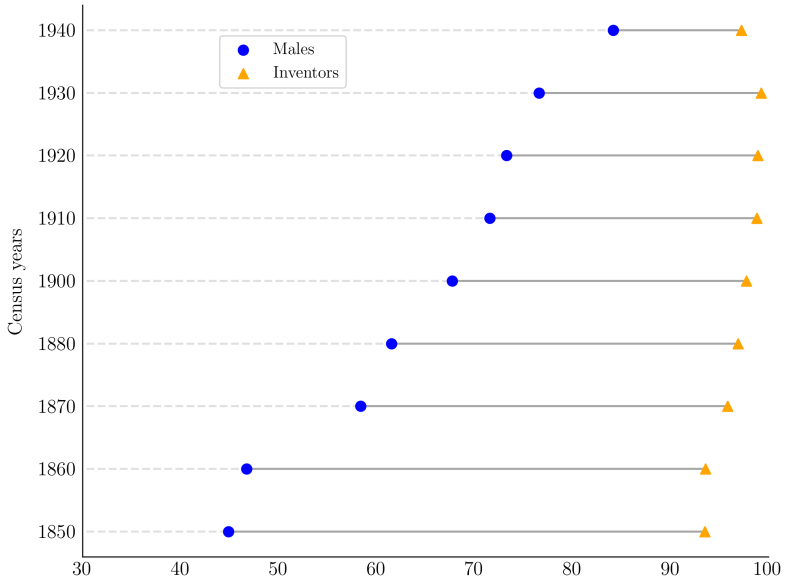
Note: The figure displays the distribution of Census respondents (solid lines) and inventors (dashed lines) across Census years by birthplace (`BPL`). Only respondents aged 18 to 70 were included in the general population.

Figure 18 shows the literacy rates of inventors and compares them with the total population in the census (panel a) and with the male population (panel b) over the period 1850 to 1940. As expected, the vast majority of inventors, particularly male inventors, are literate - over 90% of inventors throughout the time span analyzed. In contrast, literacy rates for the total population are significantly lower, ranging from approximately 40% in 1850 to about 80% in 1940.

Figure 18: Literacy rates (% distribution)



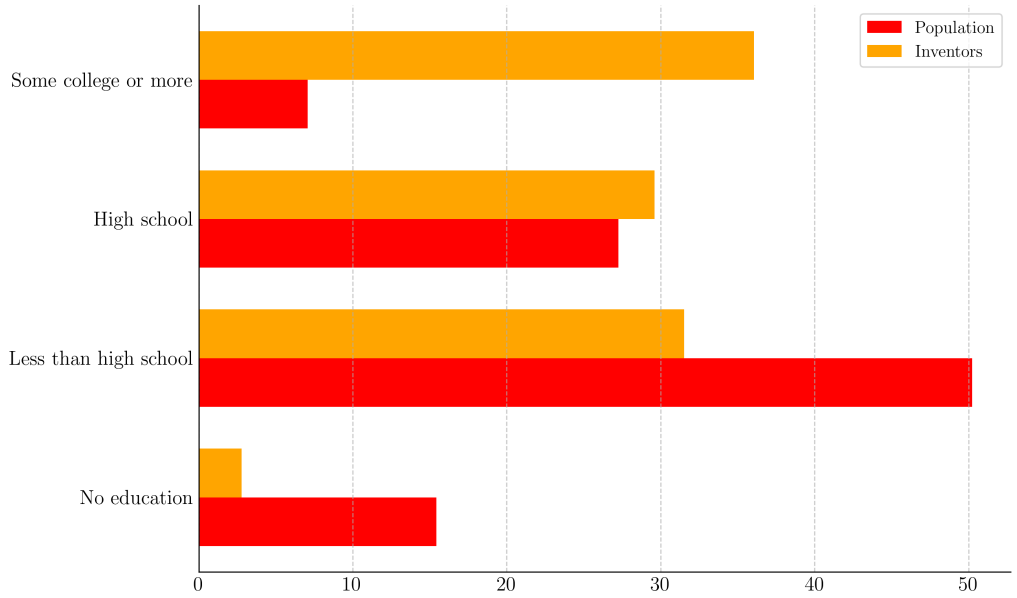
(a) Population vs. inventors



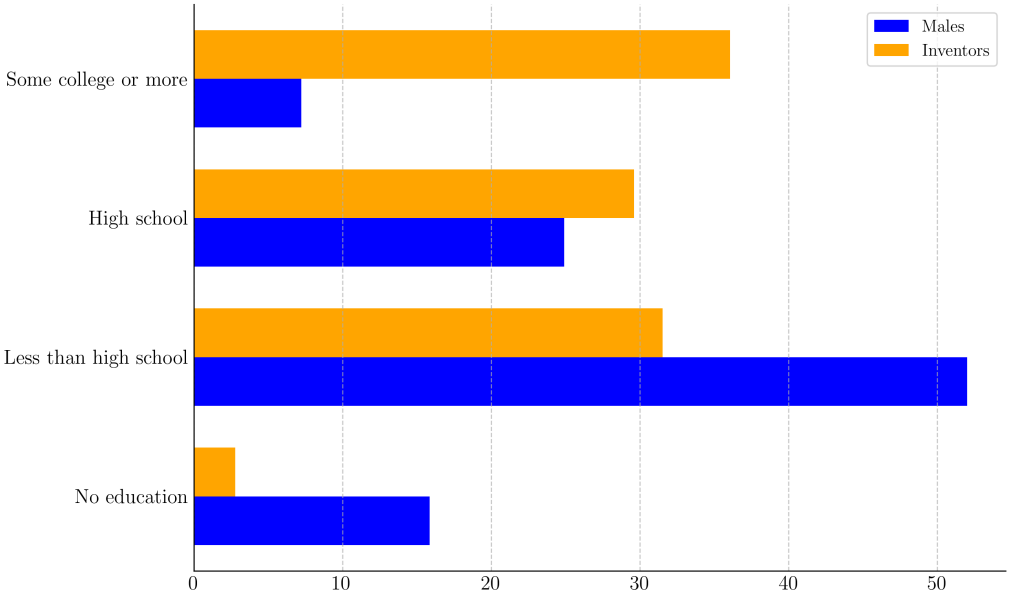
(b) Male population vs inventors

Note: This figure illustrates the distribution of education levels (`EDUC`) across three groups: the total population, the male population, and inventors. It shows the percentage of individuals in each group across different levels of educational attainment. The `EDUC` variable, which quantifies educational attainment, is available starting with the 1940 Census. For a detailed description of the `EDUC` variable, see https://usa.ipums.org/usa-action/variables/EDUC#description_section. Inventors were categorized into the following education levels based on their `EDUC` codes: 'Less than high school' (codes 10–26), 'High school' (codes 30–64), and 'Some college or more' (codes 65–997). Respondents with codes outside these ranges were classified as having 'No education.' Only individuals aged 18 to 70 were included in the general and male populations.

Figure 19: Education rates, 1940 (% distribution)



(a) Population vs. inventors



(b) Male population vs inventors

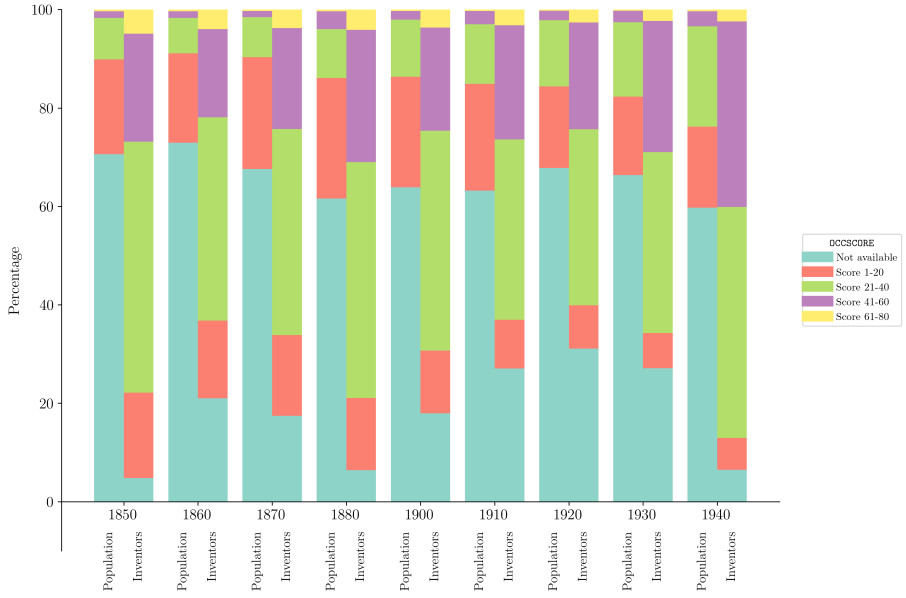
Note: The figure displays literacy rates (`LIT`), defined as the ability of Census respondents to read and/or write in any language. It shows the percentage of literate individuals across three groups: the total population, the male population, and inventors. In 1940, literacy is represented by the percentage of individuals with an education level (`EDUC`) equivalent to grade 14 or higher. The analysis includes only respondents aged 18 to 70 for the general and male populations.

Figure 19 compares the education levels across the three groups, i.e. total population, male population, inventors. It is important to note that this data are avail-

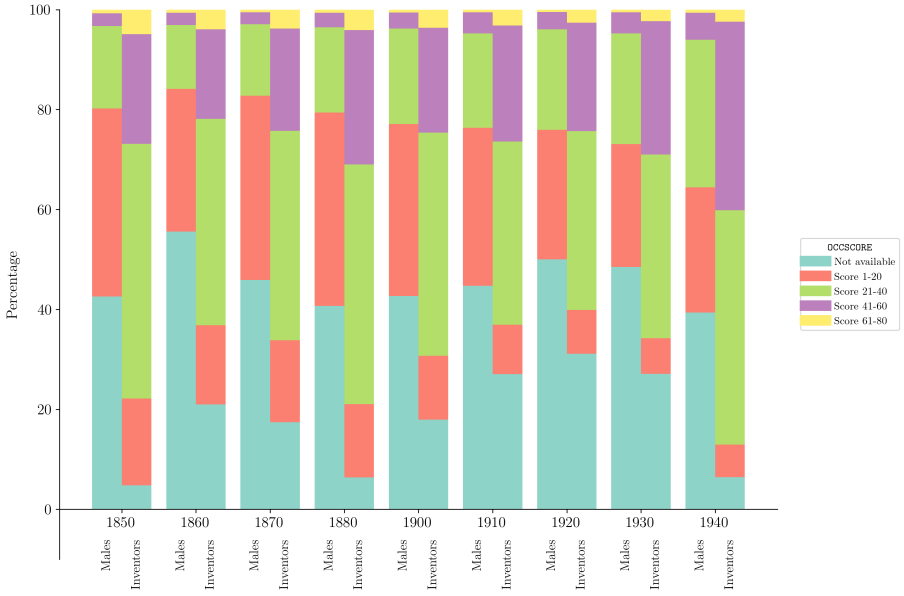
able only for the 1940 census year. As expected, inventors generally exhibit higher levels of educational attainment. For example, over 35% of inventors had a college degree or higher compared to less than 10% of the total population. Conversely, fewer than 5% of inventors lacked any formal education, in contrast to 15% of the total population. The latter findings are comparable with those shown by Akcigit, Grigsby, and Nicholas (2017b), though we have more educated inventors, which is perhaps what should be expected. For instance, they found that approximately 40% of inventors had a college degree or higher, like us, but also about 40% with less than high school.

In Figures 20a and 20b are reported the occupational scores for two groups: the census population, the inventor population. We use occupational scores solely to compare the occupational standing of the average US citizen with that of the average inventor. The ranking can be interpreted such that higher scores correspond to better-rewarded occupations. It is clear that inventors tended to occupy relatively higher-paying occupations. For example, between 20% and 40% of inventors were in the top two occupational groups (i.e., purple and yellow) throughout the period.

Figure 20: Occupation scores (% distribution)



(a) Population vs. inventors



(b) Male population vs inventors

Note: The figure shows the distribution by occupational scores (`OCCSCORE`) across three groups: the general population, the male population, and inventors. Only respondents aged 18 to 70 were included in the general and male populations.

9 ML model: robustness checks

In the previous section we presented the performance measures of the ML model. In this section, we further examine the results of the ML to identify potential biases that need to be addressed in subsequent versions of the dataset. The overall aim of the robustness checks is to demonstrate that the dataset is reliable for a variety of research purposes. To check whether the ML model identifies true matches, we perform three different types of analysis. First, we use the IPUMS crosswalks to compare our matched pairs for adjacent census years with those in the crosswalk (see Table 13). Second, we use a logit model to show the most important features of the ML model (see Table 15). Third, we compare citizenship information from USPTO patent records with place of birth from the census (see Table 16).

Table 13 shows the number of matches in both the ML and the census crosswalk (Column 2) as well as the pairs that matched in the ML but not in the crosswalk (Column 7). It also shows inconsistencies of the ML model across adjacent census years in columns 3 to 6. The first thing we notice is that about a third of the ML matches differ from the crosswalk (see Column 7 and Column 8). This proportion is stable in the early adjacent census, but it tends to fall from the 1870-1880 onward. In the last census period it falls to one fifth of the total number of pairs.

Table 13: Inconsistencies between ML model and IPUMS crosswalks

High Recall

Adjacent census years	ML model and IPUMS crosswalk match consistently	ML model matches histid in year t with a different histid in year $t+10$ compared to IPUMS crosswalk		ML model matches histid in year $t+10$ with a different histid in year t compared to IPUMS crosswalk		Pairs matched by ML model not in IPUMS crosswalk	Overall pairs where ML model matches an inventor with two adjacent census histids
		Crosswalk Steps		Crosswalk Steps			
		Step 1	Step 2	Step 1	Step 2		
1850-1860	1053	424	124	319	65	1129	3114
1860-1870	3640	1559	332	1092	256	3456	10335
1870-1880	7870	3019	609	2243	476	5767	19984
1900-1910	22511	10471	1905	7079	1280	13606	56852
1910-1920	29792	13403	2055	8629	1448	17215	72542
1920-1930	35872	16716	2101	10013	1558	18200	84460
1930-1940	46758	19396	2256	12206	1408	18834	100858

Note: The table presents the matched pairs and counts for adjacent census years, including detailed metrics for matches and mismatches based on the ML model and IPUMS crosswalk data. For a detailed description of IPUMS Crosswalks, see Helgertz et al. (2022).

Table 14 presents the descriptive statistics of the features used in the ML model. Note that name features between matched and unmatched in Table 14 are more similar now than in Table 5, despite a significant t-test due to a large number of observations.

Table 15 shows the estimates of the features that are used in the ML model that predicts true matches. It is worth noting that the signs of the coefficients are largely consistent with the estimates presented in Table 6, where it was tested the same model after filtering. There are two main differences. First, the magnitude of the coefficients in Table 15 is smaller and second, the curvilinear relationship found in the Model after filtering for the token variable (see Table 6) disappears in the ML model.

Table 14: Name features by matched status
After ML model (High recall)

Variable	Matched	Mean	Std	Min	Max	Obs	T-stat
Number of characters	No	15.011	2.62	4	37	980734	-7.83***
	Yes	15.039	2.54	6	36	1130569	
Number of tokens	No	2.817	0.47	1	7	980734	72.25***
	Yes	2.77	0.48	1	7	1130569	
Starts with single letter	No	0.012	0.11	0	1	980734	74.96***
	Yes	0.003	0.05	0	1	1130569	
Frequency of first name	No	50119.813	62310.64	1	182660	980734	-143.96***
	Yes	62893.693	66523.9	1	182660	1130569	
Frequency of last name	No	1102.791	2808.67	1	24475	980734	-125.85***
	Yes	1661.989	3636.98	1	24475	1130569	

Note: The table presents a comparison of name features based on whether they have been matched with a Census HISTID after the ML model. The *Number of characters* refers to a count of all characters in the inventor's name, including spaces. The *Number of tokens* counts the number of words in the inventor's name. The variable *Starts with a single letter* is binary and takes the value one if the inventor's name begins with a single letter. The *Frequency of first name* represents the count of occurrences of the first word (or token) of the inventor's name within the inventor population. Lastly, the *Frequency of last name* represents the count of occurrences of the last word (or token) of the inventor's name within the inventor population.

Table 15: Probability of having at least one match
After ML model (High Recall)

Variable	Model 1	Model 2
Number of characters	0.020***	
	(0.004)	
Number of characters squared	-0.001***	
	(0.000)	
Number of tokens		-0.846***
		(0.023)
Number of tokens squared		0.078***
		(0.004)
Starts with single letter	-0.993***	-0.755***
	(0.023)	(0.023)
Frequency of first name (log)	0.086***	0.099***
	(0.001)	(0.001)
Frequency of last name (log)	0.118***	0.132***
	(0.001)	(0.001)
Male inventor	0.457***	0.433***
	(0.013)	(0.013)
Unknown gender	0.075***	-0.145***
	(0.016)	(0.017)
Patents 1840-1849	-0.684***	-0.865***
	(0.032)	(0.033)
Patents 1941-1950	-0.062***	-0.020***
	(0.004)	(0.004)
Constant	-1.905***	-0.173***
	(0.034)	(0.033)
Number of Observations	2111303	2111303
Number of 1s	1130569	1130569
Number of 0s	980734	980734
Log Likelihood	-1419800.00	-1410100.00

Note: This table presents the results of logit regressions where the dependent variable equals 1 if a patent-inventor name was matched with a Census HISTID after the application of the ML model. The regressors include the total number of characters in the inventor's name (*Number of characters*), the number of words in the inventor's name (*Number of tokens*), a binary variable set to one if the inventor's name begins with a single letter (*Starts with a single letter*), a variable representing the frequency of the first word (or token) of the inventor's name within the inventor population (*Frequency of first name*), and a variable representing the frequency of the last word (or token) of the inventor's name within the inventor population (*Frequency of last name*). Additionally, two dummy variables, $d1840$ and $d1940$, take the value one if the patent was published before 1850 or after 1940, respectively, to account for the fact that these inventors had only one Census against which to be matched.

Table 16: Distribution of matched inventors by citizenship and birthplace

Citizenship	Birthplace: Austria-Hungary	Canada	Denmark	France	Germany	Great Britain	Italy	Norway	Poland	Russia	Sweden	Switzerland	United States	Others	Total
Austria-Hungary	424	9	4	6	172	26	11	0	81	77	6	6	307	608	1737
Canada	1	439	7	2	5	74	0	5	0	8	10	3	228	9	791
Denmark	0	4	182	1	10	4	0	24	0	0	11	0	58	5	299
France	1	2	0	182	29	7	1	0	1	4	0	4	79	14	324
Germany	72	4	7	20	1666	24	5	5	18	62	20	9	565	105	2582
Great Britain	23	1454	10	8	108	3491	17	14	6	47	38	8	2771	191	8186
Italy	3	4	0	1	7	9	471	0	1	3	1	0	38	9	547
Norway	0	3	18	1	17	11	0	191	1	19	99	0	66	27	453
Poland	21	1	0	0	18	3	1	0	161	32	1	0	62	15	315
Russia	48	2	5	6	59	16	9	3	108	704	11	1	212	79	1263
Sweden	1	4	15	1	19	18	0	50	0	6	803	0	309	38	1264
Switzerland	7	3	0	12	36	9	2	1	4	12	5	294	93	2	480
Total	601	1929	248	240	2146	3692	517	293	381	974	1005	325	4788	1102	18241

Note: This table presents a cross-tabulation of inventors by their citizenship (rows) and birthplace (columns). The analysis specifically excludes inventors with US citizenship and focuses on the most significant non-US citizenships reported in patent documents. The column labels represent the birthplaces of census respondents who have been matched with these inventors.

Finally, Table 16 presents a further robustness check. Specifically, we extracted the inventors who reported a US address and foreign citizenship in the patent document. We then looked at the birthplaces reported in the Census by the respondents who matched with these inventors. It is reasonable to expect that citizenship and birthplace should overlap to some extent for these inventors, to the extent that the ML matching model predicted matches accurately. For example, one could reasonably expect that an inventor reporting Italian citizenship would also be born in Italy. The findings reported in Table 16 indicate that this expectation is met for some nationalities, but not for many. For example 86% of inventors with an Italian citizenship are matched with individuals in the census that are also born in Italy. However, this percentages drops to around 55% for Canadians and around 42% for Norwegians. All in all, this latter finding, as well as the one reported in Table 13 indicate that there are clearly some issues that deserves further investigation.

10 Conclusion

In this paper we have compiled a novel patent-census linked dataset for the US over the period 1840 to 1950. The methodology employed allows for the construction of a more comprehensive dataset than has previously been available. The dataset is the result of two distinct efforts: first, we have built a novel patent dataset using USPTO patent documents. We have compared our work with previous attempts and discussed its merits and potential drawbacks. The patent dataset forms the basis for the second effort, which matches patent-inventors to census data. To do this, we have implemented a multi-stage methodology to identify potential candidates. We have also trained a machine learning model to select matches with multiple candidates. Several robustness checks were carried out to test the merits and potential biases of the model. Overall, our results regarding the demographic characteristics of inventors are reassuring and in line with previous work. Despite this achievement, our work is still very preliminary and not without limitations. One key challenge lies in the trade-off between the model's recall and precision. The methodology currently presented in this work prioritizes high recall to maximize the dataset's coverage. While this ensures the inclusion of as many inventor-census matches as possible, it may introduce a higher rate of false positives, potentially affecting the accuracy of downstream analyses. Addressing this limitation will require refining the machine learning model to balance precision and recall more effectively. Second, we aim at improving the accuracy of the machine learning model. For that we plan to explore additional predictors that will improve the reliability of matches without

compromising scale. Our ultimate goal is to produce a dataset that is openly shared to the academic community for research purposes.

References

Akcigit, Ufuk, John Grigsby, and Tom Nicholas (2017a). "Immigration and the Rise of American Ingenuity". *American Economic Review* 107.5, pp. 327–331.

— (Jan. 2017b). *The Rise of American Ingenuity: Innovation and Inventors of the Golden Age*. Working Paper 23047. National Bureau of Economic Research.

Andrews, Michael J (2021). "Historical patent data: A practitioner's guide". *Journal of Economics & Management Strategy* 30.2, pp. 368–397.

Arkolakis, Costas, Sun Kyoung Lee, and Michael Peters (June 2020). *European Immigrants and the United States' Rise to the Technological Frontier*. Tech. rep. Yale University.

Bazerman, Charles (1997). "Performatives constituting values: the case of patents". *The Construction of Professional Discourse*. Ed. by Britt-Louise Gunnarsson, Per Linell, and Bengt Nordberg. Language in Social Life. Longman, pp. 42–53.

Bergeaud, Antonin and Cyril Verluise (2022). *A new dataset to study a century of innovation in Europe and in the US*. CEP Discussion Papers dp1850. Centre for Economic Performance, LSE.

— (2024). "A new dataset to study a century of innovation in Europe and in the US". *Research Policy* 53.1, p. 104903.

Berkes, Enrico, Ezra Karger, and Peter Nencka (2023). "The census place project: A method for geolocating unstructured place names". *Explorations in Economic History* 87, p. 101477.

Cattell, J. McKeen, ed. (1906). *American Men of Science: A Biographical Directory*. English. New York: The Science Press.

Cattell, J. McKeen and Dean R. Brimhall, eds. (1921). *American Men of Science: A Biographical Directory*. English. 3rd. Garrison, NY: The Science Press.

Diodato, Dario, Andrea Morrison, and Sergio Petralia (2021). "Migration and invention in the Age of Mass Migration". *Journal of Economic Geography* 22.2, pp. 477–498.

Fink, Carsten, Ernest Miguélez, and Julio Raffo (2013). "The global race for inventors' brains". *WIPO Economic Research*.

Hartog, Matte et al. (2024). *Inventing modern invention: the professionalization of technological progress in the US*. Tech. rep. 2408. Utrecht University, Department of Human Geography and Spatial Planning, Group Economic Geography.

Helgertz, Jonas et al. (2022). "A new strategy for linking U.S. historical censuses: A case study for the IPUMS multigenerational longitudinal panel". *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 55.1, pp. 12–29. eprint: https://doi.org/10.1080/01615440.2021.1985027.

Honnibal, Matthew and Ines Montani (2017). "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing".

J. Acs, Zoltan and David B Audretsch (1989). "Patents as a measure of innovative activity". *Kyklos* 42.2, pp. 171–180.

King, M. L. and D. L. Magnuson (1995). "Perspectives on Historical U.S. Census Undercounts". *Social Science History* 19.4, pp. 455–467.

Kogan, Leonid et al. (2017). "Technological Innovation, Resource Allocation, and Growth". *The Quarterly Journal of Economics* 132.2, pp. 665–712.

Lax-Martinez, Gema et al. (July 2021). *Expanding the World Gender-Name Dictionary: WGND 2.0*. Economic Research Working Paper Series No. 64. World Intellectual Property Organization (WIPO).

Marx, Matt and Aaron Fuegi (2020). "Reliance on science: Worldwide front-page patent citations to scientific articles". *Strategic Management Journal* 41.9, pp. 1572–1594.

Morrison, Greg, Massimo Riccaboni, and Fabio Pammolli (2017). "Disambiguation of patent inventors and assignees using high-resolution geolocation data". *Scientific data* 4.1, pp. 1–21.

Nuvolari, Alessandro and Michelangelo Vasta (2017). "The geography of innovation in Italy, 1861–1913: evidence from patent data". *European Review of Economic History* 21.3, pp. 326–356.

Olson, Randal S. et al. (2016). "Automating Biomedical Data Science Through Tree-Based Pipeline Optimization". *Applications of Evolutionary Computation*. Ed. by Giovanni Squillero and Paolo Burelli. Cham: Springer International Publishing, pp. 123–137.

Petralia, Sergio, Pierre-Alexandre Balland, and David L. Rigby (2016). "Unveiling the geography of historical patents in the United States from 1836 to 1975". *Scientific Data* 3, p. 160074.

Romer, Paul M (1990). "Endogenous technological change". *Journal of political Economy* 98.5, Part 2, S71–S102.

Sarada, Sarada, Michael J. Andrews, and Nicolas L. Ziebarth (2019). "Changes in the demographics of American inventors, 1870–1940". *Explorations in Economic History* 74, p. 101275.

Sokoloff, Kenneth L (1988). "Inventive activity in early industrial America: evidence from patent records, 1790–1846". *The Journal of Economic History* 48.4, pp. 813–850.

Appendix

Figure A1: US Patent 209237

UNITED STATES PATENT OFFICE.

EDGAR P. DAVIS AND WALTER J. GODFREY, OF OMAHA, NEBRASKA.

IMPROVEMENT IN MOLDS FOR CASTING SASH-WEIGHTS.

Specification forming part of Letters Patent No. **209,237,** dated October 22, 1878; application filed August 6, 1878.

To all whom it may concern:

Be it known that we, EDGAR PULASKI DAVIS and WALTER J. GODFREY, of Omaha, in the county of Douglas and State of Nebraska, have invented a new and Improved Mold for Casting Sash-Weights, of which the following is a specification;

The object of our invention is to provide a permanent mold for casting sash-weights, which will be available for use at any time, and suited for various sized weights.

Our invention consists in an iron or steel mold divided in two parts and constructed so as to be adjustable in length. It is also made with lugs and pins to form the eye for the cord, and a number of weights may be cast in the same mold.

In the accompanying drawing, Figure 1 is top view of our improved mold closed. Fig. 2 is a longitudinal section of the same at line *x x* of Fig. 1, and Fig. 3 is a cross-section at the line *y y*.

Similar letters of reference indicate corresponding parts.

The mold is made in two portions, *a b*, that are cored out in the form of two or more sash-weights, as seen at *c d*. *e e* are studs on the under part *b* of the mold, that pass through openings in the part *a*. These studs *e* cause the parts *a b* to go together correctly, and they are provided with a hole where they project above *a*, for the insertion of a wedge-pin, as seen at *f*, to lock the parts of the mold together. The upper end of the mold is formed so that when the parts *a b* are together a space, *g*, is left, which serves as a pouring hole or gate, and communicates with each space *c d*. The ends *a' b'* of the mold are made separate from the other parts and attached by screws *h h* and *i i* to the parts *a* and *b*, respectively. This construction permits the head of the mold to be extended or longer parts attached to lengthen the spaces *c* and *d*, and consequently increase the length of the sash-weight cast in the mold. *k k* are plugs attached to the under part, *b*, of the mold in the spaces *c d*, so as to close the ends of said spaces when the mold is secured together. These plugs *k* are held in place by screws *l*, which pass through slots *p* in the under side of *b*, to permit the adjustment of plugs *k*. *m m* are center-pins projecting from the inner end of plugs *k*. When the mold is together these pins *m* abut against the lugs *n* that project downward from the part *a*. The pins *m* and lugs *n* form the cord-hole when the weight is cast; and by the adjustment of plugs *k* by the screws *l* the pin *m* can be set closely against *n*, so that a clean hole will be formed.

o is a handle attached to the upper part of the mold.

The mold above described is made of iron or steel, and may be constructed so as to cast any number of weights at once. It is always ready for use, and will turn out smoother and better weights than the ordinary sand mold. It also does away with the labor of preparing sand molds.

Having thus described our invention, we claim as new and desire to secure by Letters Patent—

1. The ends *a' b'*, detachably secured to the parts *a b* of the metal mold by lap-joints and screws, substantially as and for the purpose described.

2. The part *a* of the metal mold, provided with the detachable end *a'* and projecting lug *n*, in combination with the part *b*, provided with the detachable end *b'* and adjustable plug *k* and pin *m*, substantially as and for the purpose described.

EDGAR PULASKI DAVIS.
WALTER JOHN GODFREY.

Witnesses:
HIRAM A. STURGES,
HUGH KILLEN.

Table A1: Gender distribution of inventors by decade

Decade	Female	Male	Unknown	Total
1830	22	1,908	304	2,234
1840	57	6,067	1,106	7,230
1850	217	21,093	6,694	28,004
1860	1,145	72,979	23,687	97,811
1870	2,114	134,647	12,133	148,894
1880	3,741	220,862	18,586	243,189
1890	5,472	263,732	17,060	286,264
1900	8,600	379,816	16,753	405,169
1910	11,557	441,183	33,696	486,436
1920	14,609	521,822	53,527	589,958
1930	12,274	454,863	53,496	520,633
1940	7,379	235,321	27,459	270,159
Total	67,187	2,754,293	264,501	3,085,981

Note: The table reports the number of patent-inventor pairs by inventor's gender and decade. Decades are defined by excluding the decade's first year and including the last. For example, the decade 1840 refers to patents filed from 1841 through 1850, inclusive.

Table A2: Gender distribution of inventors by decade (%)

Decade	Female	Male	Unknown	Total
1830	1.0	85.4	13.6	100.0
1840	0.8	83.9	15.3	100.0
1850	0.8	75.3	23.9	100.0
1860	1.2	74.6	24.2	100.0
1870	1.4	90.4	8.1	100.0
1880	1.5	90.8	7.6	100.0
1890	1.9	92.1	6.0	100.0
1900	2.1	93.7	4.1	100.0
1910	2.4	90.7	6.9	100.0
1920	2.5	88.5	9.1	100.0
1930	2.4	87.4	10.3	100.0
1940	2.7	87.1	10.2	100.0

Note: The table reports the percentage of patent-inventor pairs by the inventor's gender and decade. Decades are defined by excluding the decade's first year and including the last. For example, the decade 1840 refers to patents filed from 1841 through 1850, inclusive.

Table A3: Gender distribution of Census participants

Census	Female	Male	Total
1850	9,738,515	10,242,920	19,981,435
1860	13,378,480	14,091,791	27,470,271
1870	19,048,383	19,473,433	38,521,816
1880	24,633,505	25,506,977	50,140,482
1900	37,180,647	38,644,065	75,824,712
1910	44,792,106	47,611,905	92,404,011
1920	51,883,728	53,861,405	105,745,133
1930	60,700,907	62,089,070	122,789,977
1940	65,767,108	66,136,802	131,903,910

Note: This table reports the count of individuals by gender from various decades of the US decennial Census. It includes all individuals, regardless of age.

Table A4: Gender distribution of Census participants (%)

Census	Female	Male	Total
1850	48.7	51.3	100.0
1860	48.7	51.3	100.0
1870	49.4	50.6	100.0
1880	49.1	50.9	100.0
1900	49.0	51.0	100.0
1910	48.5	51.5	100.0
1920	49.1	50.9	100.0
1930	49.4	50.6	100.0
1940	49.9	50.1	100.0

Note: This table reports the percentage distribution of Census participants by gender and US decennial Census. It includes all individuals, regardless of age.

Table A5: Intersection of data sets

Comparison	Common patents	Only in first	Only in second
BHMNT vs HISTPAT	2,212,421	307,379	10,722
BHMNT vs PATENTCITY	2,453,373	66,427	10,537
HISTPAT vs PATENTCITY	2,165,804	57,339	298,106

Note: This table reports the number of patents shared between each pair of data sets, as well as those exclusive to each set.

Table A6: Initial set of matched pairs

Census	Gender	Original data set		With at least one matching		Unique Census HISTID matched
		Patent-inventors	Patents	Patent-inventors	Patents	
1850	M	25,883	23,730	25,401	23,336	1,027,132
	F	233	231	227	225	12,587
	U	7,254	6,629	6,926	6,351	215,353
1860	M	87,271	79,300	86,851	78,979	2,518,729
	F	1,190	1,175	1,181	1,166	59,237
	U	27,723	25,333	27,477	25,147	807,685
1870	M	193,094	174,559	190,178	172,093	4,666,559
	F	2,912	2,875	2,846	2,810	140,003
	U	33,149	30,828	32,533	30,282	975,042
1880	M	332,297	300,716	326,824	296,095	6,691,285
	F	5,286	5,231	5,158	5,105	250,127
	U	27,671	26,868	26,858	26,089	803,951
1900	M	527,615	477,029	525,653	475,391	10,834,226
	F	11,098	10,939	11,056	10,897	533,898
	U	23,071	22,494	22,791	22,233	619,348
1910	M	684,551	621,597	682,504	619,862	13,540,110
	F	15,714	15,393	15,661	15,342	743,779
	U	28,791	27,289	28,401	26,933	593,689
1920	M	783,930	710,864	781,630	708,858	15,033,778
	F	19,288	18,887	19,221	18,821	873,666
	U	48,136	45,285	47,479	44,688	1,123,283
1930	M	821,047	732,817	817,919	730,116	15,858,472
	F	19,679	19,331	19,591	19,245	852,196
	U	49,302	46,980	48,512	46,248	1,301,316
1940	M	725,923	632,692	722,794	630,083	14,413,892
	F	16,881	16,624	16,796	16,542	722,372
	U	35,518	34,680	34,934	34,121	985,731

Note: The table reports, for each Census year t, the number of patent-inventor pairs located in the US and relative patents in our dataset for the period $t-10$ to $t+10$. It also shows the number of patent-inventor pairs located in the US and relative patents in our dataset that have at least one matched census candidate during this period, as well as the count of unique matched candidates in the Census.

Table A7: Matched pairs after filtering

Census	Gender	Original data set		With at least one matching		Unique Census HISTID matched
		Patent-inventors	Patents	Patent-inventors	Patents	
1850	M	25,883	23,730	19,304	18,234	220,093
	F	233	231	59	58	995
	U	7,254	6,629	2,337	2,297	8,296
1860	M	87,271	79,300	65,695	61,523	594,600
	F	1,190	1,175	347	345	6,802
	U	27,723	25,333	11,796	11,521	52,883
1870	M	193,094	174,559	146,986	136,663	1,161,079
	F	2,912	2,875	1,051	1,043	18,860
	U	33,149	30,828	12,664	12,437	53,243
1880	M	332,297	300,716	255,067	236,570	1,796,765
	F	5,286	5,231	2,258	2,241	39,615
	U	27,671	26,868	10,717	10,612	54,181
1900	M	527,615	477,029	417,075	384,932	3,139,539
	F	11,098	10,939	6,156	6,099	99,388
	U	23,071	22,494	8,164	8,121	41,322
1910	M	684,551	621,597	550,756	510,406	3,862,672
	F	15,714	15,393	9,108	9,020	136,797
	U	28,791	27,289	4,737	4,707	27,326
1920	M	783,930	710,864	631,211	586,274	4,074,428
	F	19,288	18,887	10,241	10,116	142,112
	U	48,136	45,285	10,995	10,856	68,013
1930	M	821,047	732,817	674,155	616,529	4,215,594
	F	19,679	19,331	10,407	10,276	134,493
	U	49,302	46,980	14,513	14,340	81,735
1940	M	725,923	632,692	585,679	525,612	3,574,473
	F	16,881	16,624	8,839	8,725	113,381
	U	35,518	34,680	10,559	10,440	54,502

Note: The table reports, for each Census year t, the number of patent-inventor pairs and patents in our dataset for the period $t-10$ to $t+10$. It also shows the number of patent-inventor pairs and patents in our dataset that have at least one matched census candidate during this period after applying the filters described in Section 6.5, as well as the count of unique matched candidates in the Census.

Table A8: Definition of occupations for ML model

occupation	ind1950	occ1950	occstr
Agriculture	105, 116, 126, 356, 406, 407, 409, 416, 619, 636, 637	12, 13, 26, 52, 53, 61, 62, 69, 98, 100, 123, 500, 510, 532, 555, 640, 643, 644, 810, 820, 830, 840, 910, 950	FARM, AGRICULTUR, BIOLOG, VETERINARY, DAIRY, MEAT
Apparel	436, 437, 438, 439, 446, 448, 449, 466, 487, 488, 489, 608, 656, 657, 847, 848	525, 543, 582, 590, 593, 633, 634, 645, 675, 684	TEXTILE, APPAREL, WEAVER, CLOTH, COTTON, DRESS MAKER, DRESSMAKER, DRESS MAKING, DRESSMAKING, TAILOR, FASHION, SEAMSTRESS, SEAMSTER, SEWER, SEWING
Chemistry	466, 469, 476, 477, 478, 607, 618	7, 14, 26, 42, 49, 69	CHEM
Civil Engineering	206, 216, 226, 236, 239, 246, 317, 318, 326, 379, 506, 516, 546, 556, 567, 587, 588, 596, 687	3, 16, 17, 33, 35, 43, 48, 49, 63, 95, 96, 203, 240, 504, 505, 510, 511, 513, 541, 544, 553, 585, 601, 611, 620, 622, 624, 650, 660, 661, 681	CIVIL ENG, CIVIL-ENG, CIVIL MINING ENGINEER, ARCHITECT, CONSTRUCT, RAILROAD ENG, GEOPHYSIC, GEOLOG
Commercial Computing	357, 387, 578, 579, 856	16, 18, 23, 25, 26, 44, 49, 67, 68, 69, 74, 76, 83, 95, 96, 360, 365, 370, 540, 562, 563, 571, 671	COMPUT, TELEGRAPH ENG, TELEPHONE ENG, RADIO ENG
Electricity	367, 586, 588, 616	16, 18, 23, 26, 33, 35, 44, 49, 67, 68, 69, 76, 95, 96, 360, 365, 370, 515, 540, 551, 552, 603, 672	ELECTRIC, ELECTRO, X RAY, X-RAY, XRAY, PHYSICS
Engineering	206, 216, 226, 236, 239, 898, 587	2, 3, 16, 17, 26, 33, 35, 41, 42, 43, 44, 45, 46, 47, 48, 49, 69, 95, 96, 240, 503, 515, 520, 530, 531, 534, 540, 541, 544, 545, 550, 551, 552, 553, 554, 561, 583	ENGINEER, MACHIN, MANUFACT, MFR, TECHNIC, MECHANIC
Inventors	-999	-999	INVENTOR, INVENTIVE, INVENTION, INVENTING
Managers	-999	290	MANAGER, PROPRIETOR, OWNER
Manufacturing	306, 307, 308, 309, 316, 317, 318, 319, 326, 336, 337, 338, 346, 347, 348, 356, 357, 358, 367, 376, 377, 378, 379, 386, 387, 388, 399	-999	MANUFACT, MFR
Medical	467, 607, 669, 868, 869	8, 19, 26, 32, 34, 58, 59, 62, 69, 70, 71, 73, 75, 94	MEDIC, PHYSICIAN, SURGEON, DOCTOR, PHARMA, DRUGGIST, DENTIST, DENTAL, NURSE
Mechanical	307, 336, 337, 338, 346, 347, 348, 356, 357, 358, 376, 377, 378, 379, 388, 399, 606, 617, 816, 817	2, 16, 23, 33, 35, 41, 45, 46, 47, 48, 49, 69, 74, 76, 95, 96, 240, 360, 365, 370, 503, 531, 534, 540, 541, 544, 545, 550, 551, 552, 553, 554, 555, 560, 561, 563, 571, 574, 575, 582, 583, 585, 591, 592, 600, 604, 605, 610, 612, 613, 622, 624, 635, 641, 642, 662, 674, 685	MECHANIC, TECHNIC
Physics	-999	2, 16, 17, 23, 26, 41, 48, 63, 68, 69, 70, 74, 76, 360, 365, 370, 540, 545, 552, 562, 563, 571, 572, 671	PHYSICIST, PHYSICS
Science	888	12, 13, 14, 15, 16, 17, 18, 19, 23, 24, 25, 26, 27, 28, 29, 61, 62, 63, 67, 68, 69, 81, 82, 83, 84	SCIENCE, SCIENTIST, RESEARCH, PROFESSOR, ACADEMIC, SCHOLAR, LABORATORY, ANALYST, TEACHER, EDUCAT

Table A9: Confusion matrix for the Random Forest model
High recall

		Predicted class	
		0	1
Actual class	0	18,036	85
	1	36	350

Note: The table shows the confusion matrix for the Random Forest model in the high-recall scenario. Rows represent the *actual* class labels (0 and 1), and columns represent the *predicted* class labels. *Actual Class 0*: Instances of the negative class. *Actual Class 1*: Instances of the positive class. Diagonal values (18,036 and 350) indicate correct predictions, while off-diagonal values (36 and 85) indicate misclassifications.

Table A10: Confusion matrix for the Random Forest model
High precision

		Predicted class	
		0	1
Actual class	0	18,089	32
	1	100	286

Note: The table shows the confusion matrix for the Random Forest model in the high-precision scenario. Rows represent the *actual* class labels (0 and 1), and columns represent the *predicted* class labels. *Actual Class 0*: Instances of the negative class. *Actual Class 1*: Instances of the positive class. Diagonal values (18,089 and 286) indicate correct predictions, while off-diagonal values (32 and 100) indicate misclassifications.

Table A11: Classification report for the Random Forest model (**High Recall**)

Class	Precision	Recall	F1-Score	Support
0.0	0.998	0.995	0.997	18,121
1.0	0.805	0.907	0.853	386
Accuracy		0.993 (18,507 samples)		
Macro Avg	0.901	0.951	0.925	18,507
Weighted Avg	0.994	0.993	0.994	18,507

Note: This table presents the classification metrics for the Random Forest model in the high recall scenario. *Precision*: Fraction of true positives among predicted positives. *Recall*: Fraction of true positives among actual positives. *F1-Score*: Harmonic mean of precision and recall. *Support*: Number of true instances for each class. *Accuracy*: Overall fraction of correct predictions. *Macro Avg*: Arithmetic mean of metrics across classes. *Weighted Avg*: Weighted mean of metrics, accounting for support.

Table A12: Classification report for the Random Forest model (**High Precision**)

Class	Precision	Recall	F1-Score	Support
0.0	0.995	0.998	0.996	18,121
1.0	0.899	0.741	0.813	386
Accuracy		0.993 (18,507 samples)		
Macro Avg	0.947	0.870	0.904	18,507
Weighted Avg	0.993	0.993	0.993	18,507

Note: This table presents the classification metrics for the Random Forest model. *Precision*: Fraction of true positives among predicted positives. *Recall*: Fraction of true positives among actual positives. *F1-Score*: Harmonic mean of precision and recall. *Support*: Number of true instances for each class. *Accuracy*: Overall fraction of correct predictions. *Macro Avg*: Arithmetic mean of metrics across classes. *Weighted Avg*: Weighted mean of metrics, accounting for support.