A Novel Dataset for Historical Innovation Studies: Linking USPTO Patents and US Census Data from 1850 to 1940

September 2024

Zenne Hellinga[a], Jay Prakash Nagar[b], Stefano Breschi[c], Andrea Morrison[acd], & Gianluca Tarasconi[e]

Recent quantitative research on innovation has increasingly examined historical events to uncover the driving forces behind innovation and the origins and contributions of the inventors involved. A significant focus has been placed on the United States from 1850 to 1940, a century marked by technological enthusiasm (Hughes, 2004). This era coincided with the Age of Mass Migration, when more than 30 million Europeans migrated to the US (Bandiera et al., 2013), bringing with them a large influx of knowledge that triggered important technological change (Moser & San, 2020; Diodato et al., 2022). To better comprehend these developments, we present a new longitudinal micro-level dataset linking inventors on USPTO patents to individuals in US census records.

Several prior efforts have been made to construct datasets that connect inventors to US census data (see Table 1 for an overview). The earliest attempt is made by Akcigit et al. (2017), who link the first name appearing on a patent to US census data between 1880 and 1940, using the inventor's full name, geography (state, city, and county), and age as criteria for matching and filtering. After removing all multiple matches, their match rate ranges from 29% to 46%. Similarly, Sarada et al. (2019) matches patent grantees listed on the *Annual Reports of the Commissioner of Patents* to Census individuals from 1870 to 1940. They apply a fuzzy match based on name, surname, and the name of the town mentioned on the patent, and an exact match on first letter of the name and surname, and the state of residence. This results in a match rate varying around 10%, with the majority being single matches. Arkolakis et al. (2020) match male inventors on US patents to US census data covering 1880 to 1920. Similar to the approach of Akcigit et al. (2017), they match and filter based on name, geography, and age. To distinguish the correct match in cases of multiple matches, they apply a machine-learning approach based on a training dataset, although they do not provide any information on how the training dataset was created. Lastly, in a recent working paper, Hartog et al. (2024) present a similar dataset covering patents from 1856 to 1945. They first perform a fuzzy match on last name only, and subsequently filter out individuals under the age of 16. Next, they build a training dataset with additional matching features derived from inventors mentioned on Wikipedia, which they use to train a machine learning model that

[a] Department of Human Geography and Planning, Utrecht University
[b] Fuqua School of Business, Duke University
[c] Department of Management and Technology and ICRIOS, Bocconi University
[d] Department of Political and Social Sciences, University of Pavia
[e] ipQuants AG

selects the most likely match among the multiple matches. This results in a match rate varying between 20% and 60%.

***Table 1**. Overview of other datasets matching historical patents to US census data.*

Paper	Level of analysis	Patent years	Census years	Match rate	Match on	Filters	Treatment multiple matches	Training data	ML approach
Akcigit et al. (2017)	Patent-inventor (disambiguated)	1880-1940 (only patents filed in census years)	1880-1940	34% – 62%	First- and last name, and state	County, city, township, age	After filtering only single matches are kept	-	-
Sarada et al. (2019)	Inventor	1870-1940	1870-1940	±10%	Fuzzy match first- and last name, and name town	First letter first- and last name, state, and age	ML model	Manually matched inventors in Vermont	Informal support vector machine
Arkolakis et al. (2020)	Patent-inventor	1836-1975	1880-1920	8% – 48%	Fuzzy match on first- and last name, and state	JW > 0.85 & age < 80	ML model	Not reported	Random forest classification
Hartog et al. (2024)	Patent-inventor	1856-1945	1850-1940	20% – 60%	Fuzzy match last name	Age > 16	ML model	Wikipedia inventors	XGBoost

We build upon these earlier efforts with an approach that aims to provide a more comprehensive and precise method for linking inventors to census records. We link all inventors on USPTO patents to individuals in US census data from 1850 to 1940, matching them to the two adjacent census years around the granting year of a patent. Matching and filtering is done based on the similarity of the full name, gender, age, and state. We then train a machine-learning model (random forest) to select the correct match among all multiple matches. The model, based on a training dataset, applies weights to various predictors that are constructed using variables from both patent and census data. These predictors include several measures of name similarity, geographical distance between patent-county and census-county, age at patenting, literacy, labor force participation, several indicators built from occupation and industry entries in the census data, and a concordance between technological class of the patent and occupation/industry of the census individual.

To construct the training dataset, we manually curated a subset of patent-inventor-census matches. To this purpose we exploited two sources. The first source is *American Men (and Women) of Science* (AMS), a biographical reference work on notable scientists in the US and Canada published by Gale. We utilized the third edition of AMS (McKeen Cattell & Brimhall, 1921) to gather information on inventors from 1911 to 1929. The AMS entry for each scientist includes their full name, field of study, city and state of residency, place and date of birth, education, degrees, and positions held with dates. The second source of data is derived from 19th and 20th century American inventors listed on Wikipedia (Wikipedia, 2024a; Wikipedia, 2024b). Each personal page of the listed inventors was scraped for their birth date, birthplace, and any reference to a patent publication (for instance a hyperlink to GooglePatents). All references to patent publication numbers were manually checked or manually added in case the Wikipedia page made no reference to a patent. For each inventor reported in AMS

and Wikipedia, we tried to find a corresponding individual in the census, using the birth date, birthplace, and occupation as major factors to confirm the correctness of the link. Overall, we were able to create a training dataset of 4,595 patent-inventor-census links. This training dataset not only guides our machine learning model but can also serve as a validity check for our and other attempts to match inventors to census records.

It is important to be aware of any potential sampling biases in the training dataset, since this could also bias our machine learning model. Therefore, in Figure 1 (see Appendix) we display the temporal distribution of the training dataset compared to the full dataset. As shown, the training dataset places slightly more emphasis on the period between 1910 and 1930, due to the inclusion of the 1921 edition of the AMS. This bias may also account for the differences in technology distributions between the two datasets, as illustrated in Figure 3 (see Appendix).

As we are currently finetuning the hyperparameters of our machine learning model, we can only provide preliminary matching results based on our training dataset. For inventors included in the training dataset, our model predicts the true match with a precision of 80%, and a recall of 91%. In addition to this model, we are working on a second model that maximizes precision. We aim to finalize the full dataset in the upcoming weeks, therefore the presentation during the working group meeting will include more details about our matching results as well as some demographic and economic statistics on inventors obtained from the new dataset.

References

Akcigit, U., Grigsby, J., & Nicholas, T. (2017). Immigration and the Rise of American Ingenuity. *The American Economic Review, 107*(5), 327–331. https://doi.org/10.1257/aer.p20171021

Arkolakis, C., Peters, M., & Kyoung Lee, S. (2020). European Immigrants and the United States' Rise to the Technological Frontier. Working Paper. Retrieved from https://economics.indiana.edu/documents/European-immigrants.pdf

McKeen Cattel, J. & Brimhall, D. R. (1921) *American Men of Science: A Biographical Directory.* The Science Press. New York.

Diodato, D., Morrison, A., & Petralia, S. (2021). Migration and invention in the Age of Mass Migration. *Journal of Economic Geography, 22*(2), 477–498. https://doi.org/10.1093/jeg/lbab032

Hall, B., Jaffe, A., & Trajtenberg, M. (2001). *The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools*. https://doi.org/10.3386/w8498

Hartog, M., Gomez-Lievano, A., Hausmann R. & Neffke, F. (2024). Inventing modern invention: the professionalization of technological progress in the US. *Papers in Evolutionary Economic Geography, 24.08*. https://peeg.wordpress.com/2024/04/22/24-08-inventing-modern-invention-the-professionalization-of-technological-progress-in-the-us/

Hughes, T. P. (2004). *American Genesis: A Century of Invention and Technological Enthusiasm, 1870–1970*. University of Chicago Press.

Moser, P., & San, S. (2020). Immigration, Science, and Invention. Lessons from the Quota Acts. *Social Science Research Network*. https://doi.org/10.2139/ssrn.3558718

Sarada, S., Andrews, M. B., & Ziebarth, N. R. (2019). Changes in the demographics of American inventors, 1870–1940. *Explorations in Economic History*, 74, 101275. https://doi.org/10.1016/j.eeh.2019.05.003

Wikipedia (2024a). *Category: 19th-century American inventors*. https://en.wikipedia.org/wiki/Category:19th-century_American_inventors

Wikipedia (2024b). *Category: 20th-century American inventors*. https://en.wikipedia.org/wiki/Category:20th-century_American_inventors

Appendix

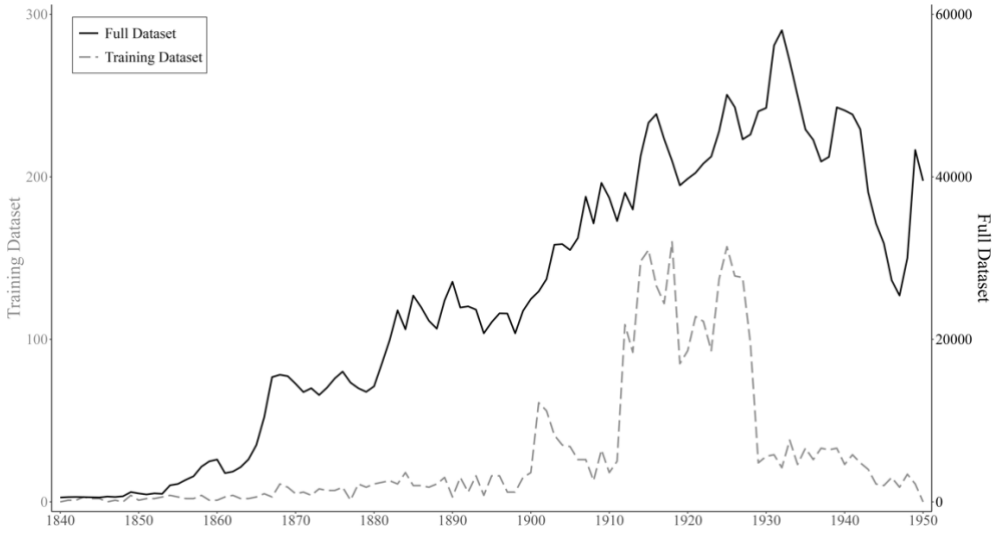


Figure 1. *Temporal distribution of the training dataset and full dataset.*

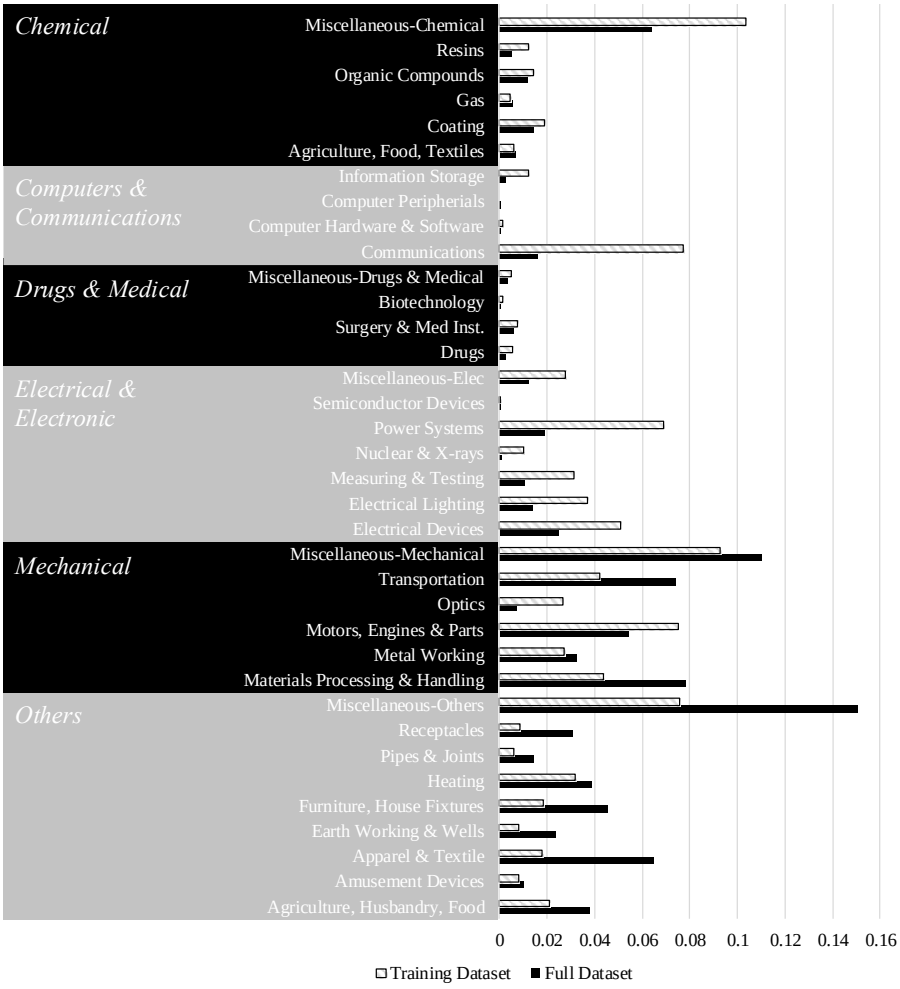


Figure 2. *Each sample's share of patents per technological field. Based on the technological classification by Hall, Jaffe, Trajtenberg (2001).*