NBER WORKING PAPER SERIES

THE AGE OF INVENTION: MATCHING INVENTOR AGES TO PATENTS BASED ON WEB-SCRAPED SOURCES

Mary Kaltenberg
Adam B. Jaffe
Margie E. Lachman

Working Paper 28768
http://www.nber.org/papers/w28768

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2021

This research was supported by a grant from the Working Longer Program (WLP) of The Alfred P. Sloan Foundation. We thank the WLP director, Kathleen Christensen, for her encouragement and support. Special thanks to Gary King for help in thinking about missing and mis-matched inventors, Lee Fleming, Kenneth Lai, Qiang Guo, Victoria Sorrentino, and Jenna DeFrancisco for their contributions, and to the students who helped research ages:, Kumba Gaye, Aishwarya Khanna, Michael Leven, Peggy Zhang, Samantha Barrett, Ruxuan Zhao, Wanchen, Mina Antic, Kacy Ninteau, Melody Wilkenfeld, and Becca Feenstra. We also would like to thank Bingyu Xu for her work on assigning gender to inventors. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Age of Invention: Matching Inventor Ages to Patents Based on Web-scraped Sources
Mary Kaltenberg, Adam B. Jaffe, and Margie E. Lachman
NBER Working Paper No. 28768
May 2021
JEL No. O31,O34

ABSTRACT

This paper overviews the data collection procedures and resulting data for inventor ages and associated death dates. We use information about inventors from patents (name and location) and search for age and date of death information from publicly available online web directories and build a scoring system to indicate the quality of information that we collect. After applying a variety of heuristics and robustness checks, we are confident of 1,508,676 inventor ages associated with patents granted between 1976 and 2018. We also find the death dates of 206,589 inventors, though we are not as confident of the accuracy of the death information. The datasets and associated replication files are freely available at: https://doi.org/10.7910/DVN/YRLSKU

Mary Kaltenberg
Pace University
and Brandeis University
mkaltenberg@brandeis.edu

Margie E. Lachman
Brandeis University
lachman@brandeis.edu

Adam B. Jaffe
Brandeis University
and Queensland University of Technology
and also NBER
adam.jaffe@motu.org.nz

A Complete dataset and documentation is available at https://doi.org/10.7910/DVN/YRLSKU

1 Introduction

We have created a new public-access database that merges publicly available patent information with information on the dates of birth and death of inventors scraped from web sources (https://doi.org/10.7910/DVN/YRLSKU). In this paper, we overview the data generation process and provide an overview of the database. A companion paper uses these data to explore how patenting changes over the life course of inventors.

2 Data

Our starting point to collect ages of inventors is information found on granted patent records. The patent office requires individuals to include their home address when filing for a patent. We utilize the name of an inventor and the location at which the inventor resided to search for information about an inventor's age on directory websites. This is particularly important because public information about individuals is typically related to their home address rather than their place of work. Directory websites are constrained to information about the US, and thus, we constrain our search to patents in which the inventor lived in the US at the time of application. We web scrape distinct combinations of the inventor name and location in the US of which there were 2,343,555 name-location combinations.

A recent feature in the USPTO dataset provided on patents view is the inclusion of inventor disambiguated data. The inventor-disambiguated data were built from the corpus of the United States Patent Office (USPTO) patent records using a hierarchical coreference (Monath, N. & McCallum, A., 2015). The persistent inventor ids associated with our dataset is from the 'patent 20180528' database. In our dataset, we have 1,858,516 unique inventor ids that are associated with 3,648,663 patents granted between 1976-2018. However, because inventors move and change addresses, there are multiple inventor-location pairs for many inventors.

3 Scraping Methodology

We search for ages primarily from three directory websites, Radaris, Spokeo, Beenverified, and a smaller subsample is scrape from Peoplefinders. For each website, we use the inventor's name and location (city and state) to generate a website URL. The program will then open and read a text representation of that webpage. If the website has information about that person with that name in that location, that information will appear in a certain section of the website, specifically, inside a div tag labeled with item type http://schema.org/Person. For each person, we extract the first

and last names (including any aliases), middle name or initial, city, state, and age, and compute a similarity score for each result to the information in our database.

When scraping information from directory websites, errors can occur in regard to matching the inventor based on name and location alone. We created a scoring system that provides a confidence rating of the age that is collected as well as, providing a basis of disambiguation in cases where a website returns multiple results for a single inventor. Higher scores are associated with higher accuracy of the collected information. High scores indicate a high level of confidence in the results, while low scores indicate a high probability of mismatch. The scoring system considers three criteria: middle name or initial, city, and state and the score is defined if it is able to find one or more of these. *Table 1* describes the scores and their criteria:

Table 1: Scoring system for accuracy of inventor information

Score	Item Match
1	only state
2	only city
3	city and states
4	only middle name
5	only middle name and state
6	middle name and city
7	all criteria

In some cases, we may not be able to match any information. In those cases, we document the type of error that occurs, but do not collect age information. *Table 2* describes these errors:

Table 2: Scoring system for intervention information errors

Error	Item Match
0	first and/or last name fails to match
-1	cannot find age information
-3	multiple matches with the same score
-4	unexpected error

3.1 Scraped Data Results

We searched for age information using 2,343,555 name-location combinations in three

web-directory websites, Radaris, Spokeo, and Beenverified. Within each of these websites, our ability to capture an age had the minimum requirement of matching on at least the first and last name of the inventor. *Table 3* shows our ability to successfully capture an age based on this minimum requirement for each website. All three websites had similar rates of success ranging from 64%-73% with Radaris having the highest success rate. These rates do not reflect any disagreements between these sources or our scoring system, it is only a reflection of our ability to collect an age within the web directory using the same location-inventor pair information.

Table 3: Age success rates by website, patent-inventor combination, n = 8,080,135

Radaris	Spokeo	Beenverified
72.46%	64.54%	66.43%

In addition to these primary web directory websites, we also ran the same web scraping technique for a fourth website, Peoplefinder, on the subset of data for inventors where we could not find reliable age information from Beenverified, Radaris and Spokeo. We deemed ages unreliable if there were multiple disagreements between ages or where we could not find an age on any of the three web directories. The subset contains 292,786 inventors, and the webscrape resulted in 75,187 additional inventors associated with an age, which is 30.04 % of the subset, and represents 4.04% of the complete inventor dataset.

Table 4 shows the extent of matching success across all four websites for inventor-location pairs. After web scraping our primary web directory websites and the subset of data from Peoplefinder, we were unable to collect at least one age on any website for 9.65 % of the inventor-location pairs. However, since inventors may be searched multiple times given that there can be multiple locations associated with a single inventor, we were able to associate at least one age to an inventor for 92.6% of inventors in our dataset. We found age agreement across all three sources of data for location-inventor pairs for approximately 30% of the data; 8.61% of inventor-location pairs had different ages on each of the three the web directories. Keep in mind that agreement across 4 sources is rare since Peoplefinder find ages for the small subset of the data.

Table 4: Age success rates overall, patent-inventor combination, n = 8,080,135

Source Combination	% Inventors with Age
Only 1 source	23.18
Only 2 sources agree	28.29
Only 3 sources agree	30.24

All 4 sources	0.03
No Age	9.65
Disagreement across sources	8.61

One concern may be that the success rate may depend on when a person published a patent. At first glance, we do not find any success rate bias due to patent application year, as evident in *Table 5*, which provides the percent of inventors where we have found at least one age by application year. Though this provides some evidence that the ages we collected do not depend on the year of patent application, it does not verify if the age collected is indeed found for the person who patented. Given this potential error, we provide a variety of robustness checks which are described in Section 5.

Table 5: Percent of inventors found at least one age by application year

App. Year	pct	App. Year	pct
1974	85.52	1997	87.65
1975	86.32	1998	87.41
1976	86.64	1999	87.28
1977	86.81	2000	86.63
1978	87.08	2001	86.43
1979	87.44	2002	86.25
1980	87.68	2003	86.14
1981	87.92	2004	86.06
1982	88.01	2005	85.93
1983	87.54	2006	85.57
1984	86.95	2007	85.27
1985	87.34	2008	85.23
1986	87.58	2009	85.29
1987	87.57	2010	85.29
1988	88.04	2011	84.78
1989	87.91	2012	84.56
1990	87.71	2013	84.15
1991	87.84	2014	84.17
1992	88.17	2015	84.12
1993	88.15	2016	83.81
1994	87.60	2017	83.73
1995	87.41	2018	83.21
1996	87.74		

3.2 Scraped Data Result Scores

In our web scraping methodology we also created a scoring system based on how accurately the patent information matched to our web directory scraping. We analyze the score results to better understand the quality of the ages that we scraped. *Table 6* shows the percentages of patent-inventor combinations with a specific score for each website where an age was found. *Table 7* displays the score results as a percent of inventor-location pairs for peoplefinders, where we used a subset of data on inventors. The most common error for peoplefinders was the inability to find age information. Spokeo provided the largest percentage of correctly matched information with 37% correctly matched on all criteria. Radaris and Beenverified did similarly well as they correctly matched over 33% for all the criteria provided. Those that matched city and state (in addition to first and last name) on Radaris was 18%, with Spokeo and Beenverified were at around 15%. In terms of errors, the most common was multiple matches for a given name, with Beenverified having the highest percentage at 32%. Radaris had the highest error percentage of not being able to find any age information. Across the three websites, an unidentifiable error occurred in half of a percent of the cases.

Table 6: Winning Scores by Website in % of total inventor-location pairs

Score	Beenverified	Spokeo	Radaris
-4	0.05	0.05	0.05
-3	31.58	27.41	9.58
-1	4.48	10.85	21.39
0	2.30	1.10	1.75
1	2.69	2.22	4.22
2	0.12	0.02	0.01
3	15.35	14.85	18.32
4	2.89	0.56	2.90
5	6.08	5.79	7.88
6	0.22	0.04	0.02
7	34.23	37.09	33.88

Table 7: Winning Scores for Peoplefinders in percent of inventor-location pairs

Score	Percent
-4.0	0.01
-3.0	10.58
-1.0	2.66
0.0	0.47
1.0	0.22
2.0	0.01
3.0	1.31
4.0	0.07
5.0	0.54
6.0	0.00
7.0	2.82

4 Data Cleaning

While we provide inventor birth year and matching scores from all of the sources we collected, we also provide a preferred birth year based on a set of heuristics. We first filter birth year based on agreement of birth year information that we collect. In cases where we only find one birth year for an inventor, we use that birthyear. In cases of multiple birth years collected, the birth year that we use depends on how many birth years we were able to find for that inventor. In general, if there is agreement on a birth year, we will use the agreed birth year among the sources. Where there are only two birth years found, if there is a difference between them, we only keep birth years that have less than a 3 year difference and use the average the discrepancy rounded to the nearest integer. If 3 birth years are collected, we keep birth years where 2 out of 3 birth years agree. If all three birth years disagree, we only keep a birth year for the inventor if the difference between each of the birth years are less than 3 years. In some cases, we may have four ages on file, in this case, we keep a birth year in which there are 3 birth years that are the same. In the case that there is disagreement between the 4 birth years, we only keep a birth year that has less than 3-year difference between all of the birth years on file. Finally, since birth years are searched by inventor-location pairs, some inventors may have had their age searched multiple times. We look for

disagreements between these searches, and only keep ages that imply less than a 3-year difference in the date of birth. This process leaves us with 1,643,968 inventors. For subsequent analysis, we further restrict the dataset to exclude inventors whose web-sourced age implies that they applied for patents before age 15 or after age 89, reducing the total number of high-confidence inventor ages to 1,508,676. A reproduction file of our process is included on dataverse and the data on dataverse includes all matched inventors so that researchers can make their own choices as to exclusions they may wish to impose.

4.0.1 Application Year and State Data

Within the original patent application forms there are errors due to the filer. To include as many patents as possible, we clean up errors related to application dates and state abbreviations. To do this, we asked a team to review patents with potential errors and make a best approximation of the correct date or state. For example, the application year "2794" may actually refer to the application year "1974". Researchers would review the grant date and other relevant information and correct the date, if possible. In some cases, the error remains and thus, the patent is not used in our analysis. In regards to state abbreviations, some errors occur, such as "NB" may actually be referring to "NE" for Nebraska. Researchers reviewed these addresses, their coauthor addresses and the assignee addresses that guided their decision on the correct state, if possible. These additional files are provided in the reproduction files on dataverse.

4.0.2 Gender

In addition to ages, we also include information about gender for our analysis, and in our dataset. We identify the gender of an inventor using the first name and birth year (using the information collected as described earlier) of the inventor via the R package, The Gender Package by Lincoln Mullen. This R package uses historical datasets from the U.S. Social Security Administration (SSA) and the U.S. Census Bureau to provide predictions of gender for first names of particular time periods. This method addresses what is commonly referred to as the, "Leslie problem," in which the popularity of a name and corresponding gender associated with a first name can change over time. We assign the birth year of inventors using the age information we scraped from web directories as discussed earlier and assign the first names from the disambiguated data provided in patentsview. In our analysis, we choose the cutoff to be 85% probability that the inventor is male or female. The gender package is only valid for the United States, but given that we only use U.S. based inventors within our dataset, we avoid

problems associated with gender attribution that may vary by language (commonly known as the "Andrea" problem).

The USPTO has also provided a dataset identifying the gender of inventors via patentsview. However, we continue using gender identification from the Gender Package since we can incorporate information about an inventor's birth year into our gender prediction and are able to identify more cases of the gender of an individual. We have compared the two datasets and find differences in gender identification (when both datasets identify a gender) to be minimal. We have provided a comparison of these differences in *Table 7.* In most cases, the difference between the datasets was that we were able to identify the gender of more inventors. Disagreement only occurs in 0.31% of names.

Table 7: Gender Identification Comparison of Results between USPTO and Mullen Algorithm

USPTO	Our DS	Diff	% Diff
F	Miss	62055	4.29
F	M	2714	0.19
Agree	Agree	889440	61.55
M	F	1661	0.12
Miss	F	6752	0.47
Miss	Miss	13937	0.96
Miss	M	457020	31.62
M	Miss	11596	0.80

4.1 General Description of Age of Inventor

We combine information about patents with our birthyear information to understand new insights about the relationship of aging over the life course of an inventor. We exclude inventors for whom the age we found suggests that they applied for a patent before age 15 or above 89, as patents at ages outside that range would be highly unlikely. After this procedure, we are left with 1,508,676 inventors with age information holding 3,383,594 patents. Previous research typically uses experience, which is noted as the number of years since first patent, as a proxy for age of invention. With this new dataset, age can be disentangled from years of experience. The start of invention can occur throughout anytime of an inventor's life course, though this typically occurs in the late twenties or early 30s. Interestingly, the pattern of the age of first invention has changed over the past decades. *Figure 1* reveals this changing pattern in that the

age of first patent occurred in the early 30s in 1996-2005, but the decade following shows an earlier shift where inventors typically first patent in their mid-20s.

Figure 1: Age at First Patent by Decade of Patenting

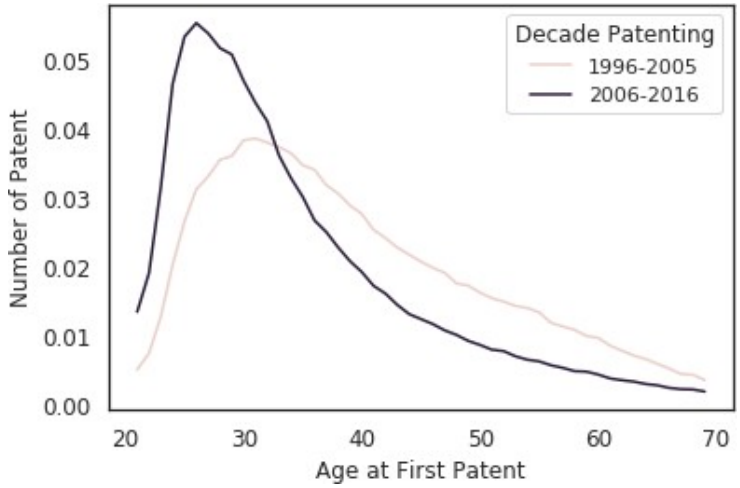
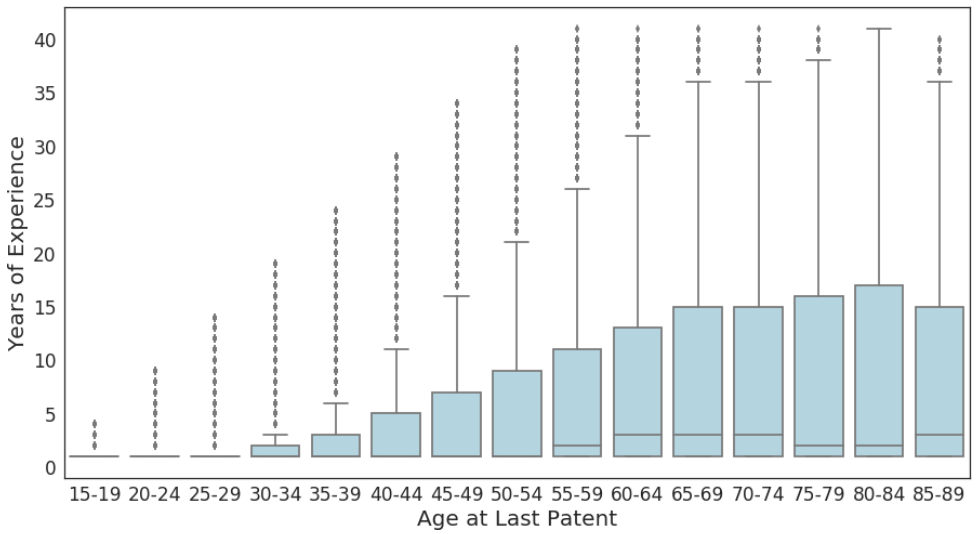


Figure 2 shows boxplots of the years and experience and age at which the inventor last patented. Mean years of experience rises with age until the 60s, after which it plateaus. In general, there is a lot of variation within age groups on the years of experience and age of which an inventor patents.

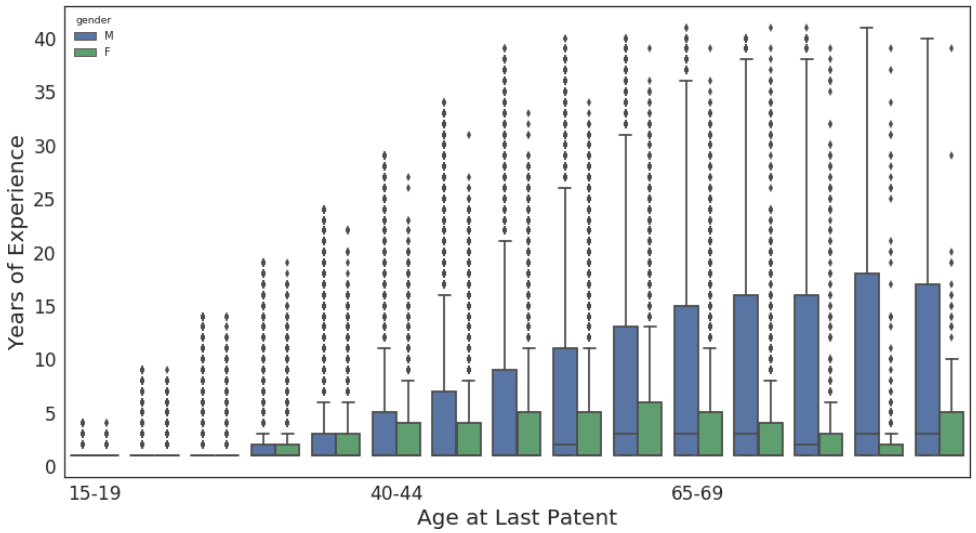
Figure 2: Years of Experience by Age



When looking at the relationship of experience by age and gender more interesting patterns emerge, as seen in *Figure 3*. Experience increases as men age, but for women,

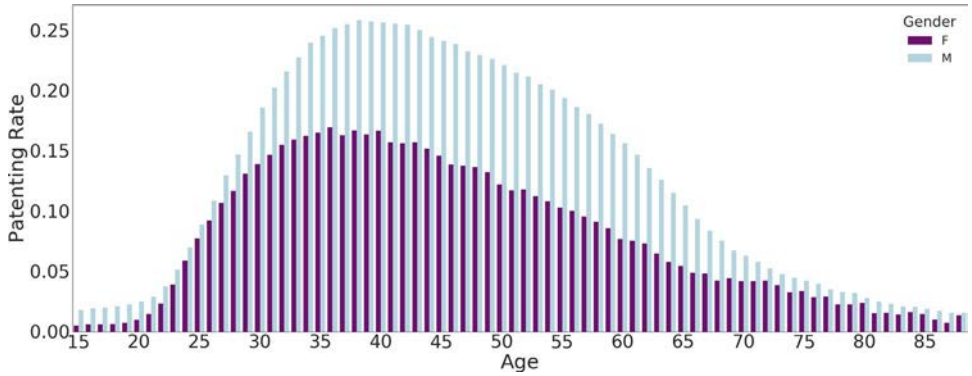
years of experience peaks in their early 60s, and further, patterns of experience start to differentiate between men and women when inventors are in their 40s.

Figure 3: Years of Experience by Age and Gender



Differences in patenting between men and women is further seen in patenting rates in *Figure 4.* Patenting rates of women are relatively similar to men until their late 20s, where rates start to diverge.

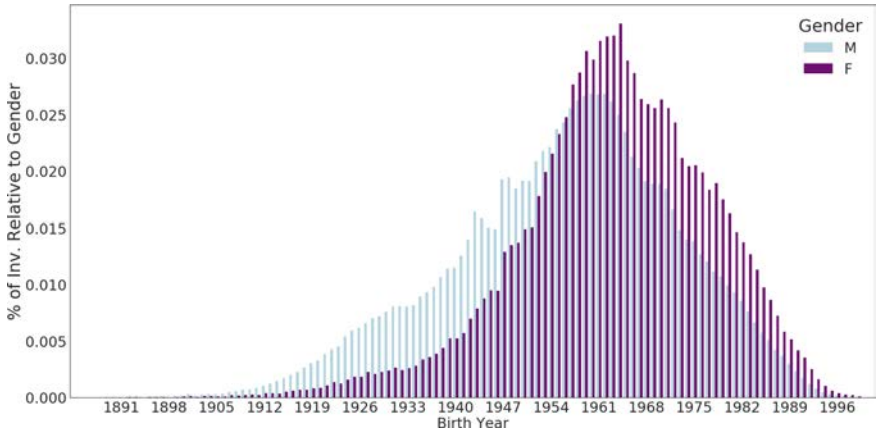
Figure 4: Patenting Activity by Gender



Another interesting viewpoint that our data provides is a deeper dive of female participation in invention through cohorts. Previous research has found that female patent participation has increased over time. *Figure 5* shows the percent of inventors by birth year and gender. While most inventors in our dataset tend to be born in the late 50s and early 60s, we can see there is a shift in the distribution of female versus male inventors. There are many more men who are born prior to the 1940s as compared to

women. These patterns suggest that increased female inventors is likely due to younger generations patenting.

Figure 5: Birth Year by Gender



5 Data Verification and Reliability

5.1 Student Searched Data

To compliment the web-scraping of inventors, we also researched birthyears of inventors through an in-depth web search. We randomly selected 4,060 inventors based on the consistency of scraped birthyear information of inventors, which included the categories: multiple birthyears found, no birthyear found, one birthyear found, three birthyears found that disagreed, and two birthyears found that disagreed. Students who searched for birthyear related information used a variety of sources at their discretion and a copy of their instructions is reproduced in the Appendix. They indicated their subjective confidence of the birthyear that they found from 1-3 with 3 meaning that they were highly confident in the birthyear that they found. Where there was additional accompanying information found on the internet, such as newspaper articles, obituaries, university information, LinkedIn profiles, or other detailed information that could confirm the birthyear of an inventor, students considered that the birthyear they collected for the inventor as "highly confident."

Overall, we could not find 947 inventors' birthyear (23.3% of subsample). However, students found exact birthdates for 36.8% of the sample, death dates for 13.4% of the sample and about 11.65% of the sample with both death and birth information. We also analyzed differences in the percent of birthyears found by our birthyear searched categories. We find that multiple consistent birthyears found in our web scraping methods correspond to the highest likelihood of finding a birthyear with the most confidence through a manual web search. Where there were no birthyears found,

manual web searching was also difficult, but were able to recover some birthyears through manual searching. When there were multiple disagreeing birthyears or only one birthyear found, students found birthyears, but they were less confident about the birthyear they found. *Table 8* describes the results of the manual search, which includes the total number of inventors searched in that category, percent that we found a birthyear and the percent of birthyears that we found that we are highly confident is the true birthyear of the inventor.

Table 8: Descriptives of Student Researched Data

Type	Total	Pct Birthyear Found	Pct. Birthyear Confidently Found
Mult. Cons. Ages	986	86.95	73.91
No ages	658	65.23	45.03
One Age Found	1012	73.69	56.02
3 Ages Disagree	614	79.58	62.15
2 Ages Disagree	785	76.73	58.39

We also looked at the manually searched results based on the decade that inventor patented and the subset of data to see if there were time-varying differences in finding an age. For example, there might be less information available online for older inventors. Overall, such differences are small. Those who patented in 1996-2005 do seem to have the highest success rate, corresponding to ten percentage points success rate of finding an age compared to those who patented in 1976-1985. These results are consistent with our logistic regression to predict if an age is found in Table 10. *Table 9* describes the results of the manual search by decade.

Table 9: Descriptives of Student Researched Data by Decade

Decade	1976-1985		1986-1995		1996-2005		2006-2015	
	% Type	% Conf	% Type	% Conf	% Type	% Conf	%Type	% Conf
Mult. Ages	82.98	71.92	89.74	78.89	93.20	81.35	86.95	73.90
No ages	62.32	45.08	72.33	54.85	77.90	56.04	65.23	45.03
One Age	70.23	55.34	79.712	64.21	85.18	67.95	73.69	56.02
3 Ages Disagree	73.33	56.96	86.031	68.25	88.71	71.15	79.58	62.15
2 Ages Disagree	74.36	57.82	80.41	64.69	84.11	66.82	76.73	58.39

After completing the web search tasks, we interviewed the student about their experiences in finding ages. Some students noted that there are cases where a pool of patents were assigned to one inventor id by the disambiguation method, however they believe there are actually multiple people associated with that inventor id. This was only the case for about 2.4% of the random sample. They also noted that unique and foreign names are more likely to be unresolved and common names are more likely to have multiple ages in the search process and therefore those inventors were harder to verify. Older patents associated with an inventor have lower certainty of age given that there is less information (inability to find an age), however, ages were easily resolved if the individual has passed (obituaries provide a rich source of information).

5.2 Randomness of missing data

The consequences of our failure to match some of the inventors to find their ages will depend on the extent to which the inventors or patents that were matched differ in systematic ways from those that were not matched. To address this issue, we compare those patent-inventor pairs that were matched to those that were not, using all available information. *Table 10* presents a logistic regression of age found/not found for each patent-inventor pair, using as regressors all observable attributes of the patent-inventor pair, and displays the average partial effects[3]. The results suggest that the only variable that has an impact on predicting our ability to find an age for an individual is gender. Other variables--patent attributes, year of application, team size or field of the patent—have effects that are statistically distinguishable from zero because of the large sample size, but which are small in magnitude. Finding the age of women, in particular, can be particularly difficult because women are more likely to change their name throughout their lifetime. In addition, we rely on two algorithms – one that pools patents to a unique inventor, and another that identifies gender. It is possible that a woman patents under two different names and consequently, that algorithm creates two separate inventor ids. As a result, there may be more difficulty in finding ages for women as there may be more public information for one name than a another.

Table 10 Logistic regression for Selection (Displaying Average Partial Effects)

Variable	APE/SE	Variable	APE/SE	Variable	APE/SE
Lifetime Tot Pat	-0.003***	Yr: 1976	0.007	Yr: 1996	0.066***
	0.00		-0.013		-0.01
Female	1.247***	Yr: 1977	0.055***	Yr: 1997	0.091***
	-0.007		-0.013		-0.009

[3] The regression predicts if there is an age found (found age =1)

Cmp & Cmm	-0.066***	Yr: 1978	0.031*	Yr: 1998	0.110***
	-0.003		-0.013		-0.009
Drgs & Med	-0.034***	Yr: 1979	0.007	Yr: 1999	0.104***
	-0.004		-0.014		-0.009
Elec	-0.006	Yr: 1980	0.016	Yr: 2000	0.105***
	-0.004		-0.013		-0.009
Mech	0.057***	Yr: 1981	0.02	Yr: 2001	0.099***
	-0.004		-0.012		-0.008
Other Field	0.041***	Yr: 1982	0.003	Yr: 2002	0.084***
	-0.004		-0.013		-0.008
Missing Field	-0.028***	Yr: 1983	0.016	Yr: 2003	0.084***
	-0.008		-0.013		-0.008
Team: 2	-0.010**	Yr: : 1984	0.02	Yr: 2004	0.083***
	-0.003		-0.012		-0.008
Team: 3	-0.032***	Yr: 1985	0.000	Yr: 2005	0.080***
	-0.003		-0.012		-0.009
Team: 4	-0.045***	Yr: 1986	-0.008	Yr: 2006	0.082***
	-0.004		-0.012		-0.008
Team: 5	-0.041***	Yr: 1987	0.021	Yr: 2007	0.070***
	-0.004		-0.011		-0.008
Team: 6	-0.040***	Yr: 1988	0.002	Yr: 2008	0.068***
	-0.005		-0.012		-0.008
Tean: 7	-0.026***	Yr: 1989	0.027*	Yr: 2009	0.037***
	-0.006		-0.011		-0.008
Team: 8+	-0.013**	Yr: 1990	0.037***	Yr: 2010	0.036***
	-0.005		-0.011		-0.008
Forward Cit.	-0.001*	Yr: 1991	0.040***	Yr: 2011	0.025**
	0.00		-0.011		-0.008
Backward Cit.	0.004***	Yr: 1992	0.055***	Yr: 2012	0.015*
	0.00		-0.01		-0.008
N. Claims	0.008***	Yr: 1993	0.075***	Yr: 2013	0.005
	-0.001		-0.01		-0.007
Disruptiveness	-0.023***	Yr: 1994	0.072***	Yr: 2014	0.002
	-0.004		-0.01		-0.007
		Yr: 1995	0.071***		
			-0.01		
N: 6656934	Log likelihood	-3095173.966			

6 Death Data

Following a similar process of web scraping for information of ages of inventors, we also scraped data related to the death of an inventor using the same data starting point of inventor-location pairs from the USPTO. Information about the death of inventor is likely to be more unreliable for several reasons. An inventor who stops patenting earlier in their life may have moved from the last location listed on the patent, so there is likely a longer lag between the last patent and when they die. Nonetheless, we web scrape information about the death of an inventor as best we can and develop a scoring system of the information that we collect. We use two main websites as a source of our information, ancestry.com and familysearch.com. Within these websites there are subsets of data banks which we utilize. For ancestry.com, we use the data banks of find a grave index (fgi), social security death index (ssdi), and obituaries. For familysearch.com, we use the data banks of find a grave index, social security death index, and genealogy bank. Ancestry and Family Search share two data banks, but websites have various search algorithms and data coverage, which is why we chose to web scrape the same data banks at different websites.

We impose some restrictions in data collection, namely we drop death information where the first and last name do not match, we drop data whose birth year (where available) is later than the first patent application date listed for the inventor, and finally, we drop death related information if the death date is earlier than the last patent application date. In addition to this filtering process, we also created a scoring system designed to be a proxy for accuracy of date of death related information. The score system includes matches of the birth year (within two years as identified by the birth year on file from our cleaned data set) and matches of the state (this may be related to a residence, burial or obituary place depending on the data availability for the data bank). We chose to search a broader location area to be state rather than city and state because individuals may not necessarily be buried in the same town of which they resided. *Table 11* provides an overview of the matching requirement and score assigned. Genealogy Bank includes a variety of additional information, and thus the scoring system includes more combinations and overviewed in *Table 12*.

Table 11: Scores for Inventor Death Dates

Score	Matched Items	
	Birth Year	Residence of death location
24	✓	✓
16	✓	
8		✓
0		

Table 12: Scores for familysearch.com: Genealogy Bank Inventor Death Dates

Score				Matched Items
	Birth year	Residence place	Burial place	Obituary place
30	✓	✓	✓	✓
28	✓	✓	✓	
26	✓	✓		✓
24	✓	✓		✓
22	✓	✓		
20	✓	✓		
18	✓			✓
16	✓			
14		✓	✓	✓
12		✓	✓	
10		✓		✓
8	✓	✓		
6		✓		✓
4		✓		
2				✓
0				

We provide a broad overview of how many inventors we were able to identify death date, as well as how many inventors matched on some aspects of our scoring system in *Table 13*. We find that the source of the data bank does make a difference in terms of

how many matches we are able to find. Looking at the same original source of data, for example, the social security death index (ssdi), we are able to find more inventors that met our minimum requirements and matched on more additional information through Ancestry.com than through Family Search.

Table 13: Overview of Scraped Death Data

Data Bank	N Records	N Unq. Inv.	N Rec. w/ Death Date	N Rec. w/ Score > 0	N Unq. Inv. w/ Score > 0
Anc.: fgi	1,908,251	354,456 (7.80%)	1,763,147 (92.40%)	148,786	102496 (28.92%)
Anc.: ssdi	1,965,873	290,234 (6.23%)	1,965,535 (99.98%)	122,382	81,378 (28.04%)
Anc.: obit.	5,701,817	465,932 (6.26%)	1,853,691 (32.51%)	356,954	137,824 (29.58%)
Fam. S.: fgi	1,854,947	354,589 (6.89%)	1,531,381 (82.56%)	129,388	92,803 (26.17%)
Fam. S.: ssdi	514,920	81,590	514,920 (100%)	225 (0.04%)	223 (0.27%)
Fam. S.: gen.	391,914	319,914	188,435 (48.08%)	44,736 (11.41%)	44,736 (11.41%)

We combine all of the scraped results and present results only when a date of death was captured by our scraping algorithm. Some date of deaths had data errors where the year was larger than 2020, so we removed those death dates from our summary statistics. We find a death date for 535,120 inventors. *Table 14* summarizes a few statistics about the data set. Inventors are searched multiple times through the various sources, but also if they have multiple addresses, we provide a summary of the number of death dates captured by inventor id, and the total number of unique death dates captured by inventor. We also provide a descriptive age at death given using the birth year that we collected from the age dataset described earlier, and the maximum and minimum death year on file for each inventor. Broadly, there are mismatches between the age on file and the death year, but after some data cleaning processes, we can recover some information of the date of death of inventor. More work on correctly matching this information should be done in the future, but this provides a first step in garnering date of deaths for inventors.

Table 14: Description of Raw Scraped Death Data

Descriptive	Tot. Ct.	Tot. Uniq. Ct.	Death Max	Death Min
mean	64.73	15.13	2015.66	1995.23
std	56.03	9.62	3.81	10.51
min	1.00	1.00	1941.00	1846.00
25%	20.00	7.00	2015.00	1988.00
50%	51.00	14.00	2017.00	1996.00
75%	95.00	22.00	2018.00	2003.00
max	451.00	83.00	2020.00	2019.00

In some data banks, additional information can be acquired beyond the date of death of an individual, such as birth dates. When this information was available, we also collected birth dates associated with death dates listed on file when available. To provide a descriptive snapshot of the data we collected, we clean the data to present some summary statistics of the viable death data collected. The first cleaning step we provide is to limit information to where we collect both a birth and death date. We then match death date to birth years collected from the birthyear web scrapped information from section 3 and keep death dates only for inventors whose birth year matches (within 2 years). We are left with 216,802 inventors after this cleaning process. *Table 15* displays a crosstab of the birth years and death years for this subsample and shows that majority of inventors in this dataset are born in 1920-1950 and die in the 1990s. This corresponds to age of death peaking between 40-70 years old. The average life expectancy in the US is 54 (for those born 1920) - 68 years old (for those born 1950), which roughly corresponds to the data we collect (NCHS, 2017).

Table 15: Crosstab of Birth and Death Years of Subset

Death Year Birth Year	1970-1979	1980-1989	1990-1999	2000-2009	2010-2020
1880-1889	7	15	0	0	0
1890-1899	32	196	78	3	0
1900-1909	66	715	1,234	392	15
1910-1919	84	1,403	5,088	5,840	1,216
1920-1929	79	1,339	7,579	18,628	9,387
1930-1939	33	667	4,593	18990	14,089
1940-1949	28	386	3,081	19,154	18,613
1950-1959	15	209	1,984	16,164	20,692
1960-1969	4	116	985	9,271	14,303
1970-1979	2	19	240	2,370	4,951
1980-1989	0	0	22	433	1,614

1990-1999	0	0	0	19	146

Note: One inventor excluded from this table who was born in 1910 and died in 1959

Figure 6 displays the distribution of the age of death of inventors in this subset of data, where you can see that most inventors in this dataset die between the ages of 60-80 years old.

Figure 6: Age at Death for Inventors

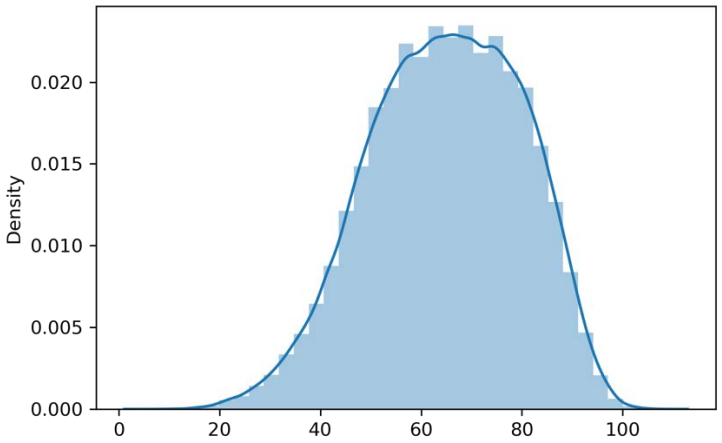
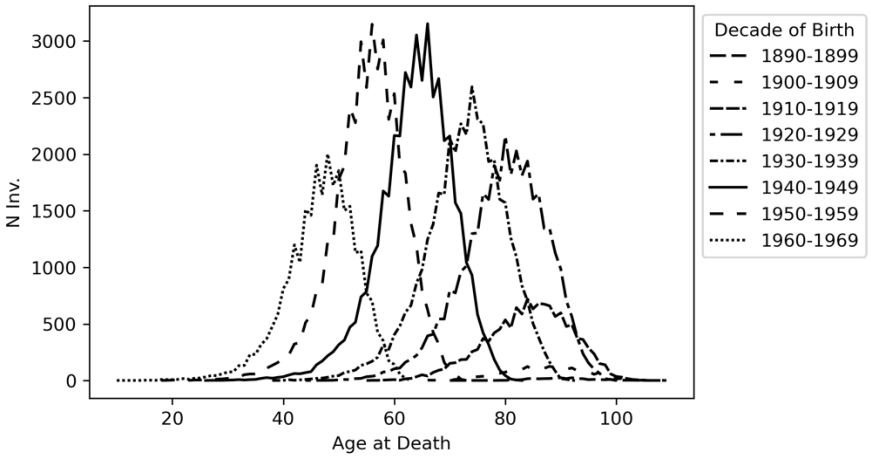


Figure 7 depicts the distribution of age of death of inventors by birth cohorts. The three largest cohorts represented in our dataset are 1950-1959, 1940-49, 1930-1939, and 1920-1929, which reflects the cohorts where we capture most of their career patenting. The cohort patterns match expected age of death for cohorts. Younger age groups are likely not to have as many deaths, and the age of death increases with cohorts reflecting the overall trend of longer life expectancy in the US over this time period.

Figure 7: Age at Death by Birth Year



7 Conclusion

We reviewed the data collection methods for finding birth years and death years of inventors by utilizing patent records and web scraping public information on web directory websites. We found 1,643,968 consistent birth years using information just from web scraping and with additional information from patents we are highly confident of the birth year of 1,508,676 inventors, which we use to provide descriptive statistics of patenting across the lifecourse. We also find death dates of 535,120 inventors and with a set of heuristics and utilizing information from patent data we are moderately confident of the death year for 216,802 inventors. These two datasets are the first publicly available data of the birth and death year of a large and representative group of inventors spanning multiple decades. We hope this will spur additional research on inventors and their innovative activities across their lifetime.

Given that the data collected cannot be verified from linked census data records, there remains some uncertainty of matching of birth and death years to inventors. For individuals who have common names, matching is likely to be more unreliable. This problem persists for associating patents to inventor ids as well as for collecting birth years and death years. We identify a birth or death year for a person associated to a city or state, which does not mean we have identified the birth or death year for the true inventor.

We analyze whether the inventors who were matched differ systematically from those who were not. In terms of all observable characteristics of inventors and their patents, the only characteristic that is associated with likelihood of age matching is inventor gender. This non-random selection into the dataset should be considered if it is used to explore gender-related issues. Observable attributes related to the timing of the patent, the composition of the inventor team and the technology field are not associated with likelihood of matching to a meaningful degree. We cannot, of course, determine what *unobservable* attributes might be associated with failure to match or likelihood of mismatch. Other researchers using the data should consider what kinds of unobservable characteristics that might be in play are relevant in their context, and consider caveats or robustness checks as appropriate.

References

Kaltenberg, Mary, Adam Jaffe and Margie Lachman (2021), "Invention and the Life Course: Age Differences in Patenting" National Bureau of Economic Research Working Paper No. 28769

Monath, Nicholas and Andrew McCallum. *Discriminative Hierarchical Coreference for Inventor Disambiguation*. PatentsView Inventor Disambiguation Technical Workshop. September 2015

NCHS. 2017. United States life tables, 2014. Table 19. National Vital Statistics Reports 66(4). https://www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_04.pdf (PDF)

Appendix

Student instructions for randomized search of ages is provided below:

Sloan Inventors Search Best Practices

We're creating a dataset of inventor ages to verify an existing dataset that web scraped inventor ages from a variety of web directories. The priority is find the age of an inventor.

Step One – Setting Up

1. **Email Lifespan Lab manager** when you are about to start working so she can keep track of your hours.
2. **Save a new file**

   Make a copy of the original and save it with you initials and manual search as a CSV file (not xlsx or any other format), like:
   originalfilename_LL_manualsearch.csv
   If you already have a file started just open it and begin working where you left off. When you save it and upload it to Box it will be uploaded as a new version.
   You'll download it either to your flash drive or a computer, but make sure to update your flash drive ~once a week.
3. **Add Columns**

   You shouldn't have to do the following steps. I will make the datafile and you will simply have to fill out the variables based on your searches. However, please review all of the below variables so you know exactly what you are looking for and what you need to complete on the spreadsheet.

   For each spreadsheet, add 9 columns after the final existing column on the Excel sheet:
   1. Age – the age you found manually
   2. Gender—M/F if clear, put U if not specified (only fill this in if you come across during your search)
   3. Birth date—if found
   4. Date of death—if found
   5. Source—a link to the website where you found the information
   6. Certainty – 1,2, or 3, (See Step Three in documentation below)
   7. Resolved—Y/N used mainly to keep track of the ones we are not certain about or can't find
   8. Comments – indicate what ultimately helped you decide this was the right person/right age

   Only fill out the below if you come across it in your search

   9. Flags_location – put a 1 here if the location in your file does NOT match the location on the patent
   10. Flags_LatLong – Google the latitude and longitude. Put a 1 if this does NOT match Justia/Google patents
   11. Location_edits – Put the correct city (from Justia/Google patents)

Step Two – Search

Preface: If you spend 15 minutes on a single inventor and cannot find their age, mark the age as -1, mark the case as unresolved, and move onto the next inventor.

A. **FIRST** check if the location in your file matches the location on the patent. If not, flag and comment (see above)

B. Look up the inventor's **exact name** (including initials, spelling and any suffix (Jr./Sr./III, etc) and location. Most cases can be resolved with this information by confirming the same age on **at least two websites/sources.**

Websites that have been the most successful in confirming ages are:
Intelius (includes job history)
Instantcheckmate
Pipl
Beenverified

More resources that were also very helpful:
Ancestry (username: xubingyu97 password: undertherose3022)
MyLife
Verifinder
TruthFinder
Whitepages
DOBSearch
USA people Search
Peoplefinders
Peoplefinder
Veripages
https://beebom.com/best-people-search-engines/
http://www.ebizmba.com/articles/people-search
http://www.zabasearch.com/

C. In cases where this information cannot be confirmed, use the workplace (assignee) and patent ID associated with the inventor.

Websites that have been helpful in resolving more difficult cases are:
Justia Patents (for patents)
Linkedin (work history)
ResearchGate (published work)
Google their name and workplace (can find company website or published work)
Findagrave (death info; cemeteries may be nearby, not exact location)
Obituaries (try to google name and obituary for every person to see if you can collect death info where applicable)

D. Death information can be successfully found through:
Ancestry.com (use all category search, but limit location when possible; use the exact name option)
Familysearch.com
Other options:
Legacy
ObitTree

Helpful Tips:

- Open up a few sites and confirm across all of them
- Find an age on two or three sites, then look at workplace, location, and patent info to try to confirm this is in fact the inventor
- Check that locations match database, use Google maps to see if mismatched city is near listed city, or if cities match company city or patent city
- Use the location related to the most recent patent of the inventor in your initial search.

- If instant checkmate is younger age, check ancestry.com for death info
- Patent numbers can be found on google to see dates, other inventors, match to company
    - Consider the year of the patent – there is a chart on Box with birth years versus patent dates, but we assume a person would not have a patent before age 18
    - Ex) you find a person who is either 35 or 65, but his patent is from 1980 – he can't be 35
    - Patent year is most helpful with young people to see if the patent is before their time
    - Can use patent location to match with inventor location
- LinkedIn can be helpful to confirm person with company
    - College years minus ~20 years to estimate birth year (use this as confirmation of an age you found)
    - See if patent content is related to their field of study
- Sometimes a different city might show up – you can check Google Maps to see if it is the next closest city or the county name instead if you are finding someone who seems correct, or check the database to see if that inventor ID is listed in multiple cities
- Father/son teams – pay attention to Jr./Sr./III. Some people have the same name but ~25 years apart – we want to determine which one we need
- In some cases, preposition can be searched differently than recorded. For example, "John von Lastname" might not get a hit, but try the preposition without the space attached to the lastname, like, "John vonLastname"
- For death records:
    - Use as many details as you can
    - If only birth year is provided, allows +/- error
    - Some collections provide only county names, so be sure not to treat the address specific to a county as a different address

Step Three - Record and Document

A. Input your decisions, including when you searched and couldn't confirm. Take caution in entering the information correctly (it's very easy to include a typo)
    a. Rate your decisions that you've identified the inventor
        1 not confident, 2 somewhat confident, 3 very confident
        - Opt to spend a little more time on a case to get a confident age than leaving it as a questionable age with a low rating
    b. When you have exhausted your search strategies **and can't confirm have missing data put -1** so we know an attempt was made. It is very important that if you can't find the age you mark the age variable as -1 for that inventor. Age is the most important piece of information we are looking for. Also, please remember to mark cases as resolved or unresolved based on your search process.
    c. List the sources you used and any interesting notes about the case
        - Write a note if you feel there is incorrect information in the data, like wrong city assigned to someone with the same name as another owner, or patent had the name/date recorded incorrectly
        - Include what makes you confident that this is the right person and right age. **Remember** The goal is to confirm that the INVENTOR is a certain age. The computer found that A PERSON of that name is that age, but we want to try to confirm that this person is in fact the inventor we are looking for.

Step Four – Upload to Box

- When you are done working for the day, save your file and upload it to Box in your folder. It is very important for you to remember to do so, because I do keep track of how

much is getting done per day, so I can tell how many cases we are on track to complete by the end of the semester. If you forget to upload your file every time you work I cannot tell how much is getting done every day/work session.

- When you finish working, email me to "clock out" so that I can record how much you work per week. You should still be working 13 hours per week.
- Make sure to log and submit your hours on Workday.

Important things we have established since beginning the search process

- One inventor may have multiple locations. You can tell if it is all the same person by the inventor id. If it has the same inventor id, it belongs to what we believe is the same person.
- It is okay if you can't confirm every location an inventor has lived in. We just need a general idea of the inventor's age and that the inventor lived in one or two of the places listed.
- Prioritize age, once you have found age move on without confirming additional information
- If there is an obituary, it closes the search faster.
- IMPORTANT: If you find yourself spending more than 15 minutes on a single inventor and cannot find an age, mark age as -1 and mark the case as unresolved.