Jane, John … Leslie? A Historical Method for Algorithmic Gender Prediction

Cameron Blevins and Lincoln Mullen

### Abstract

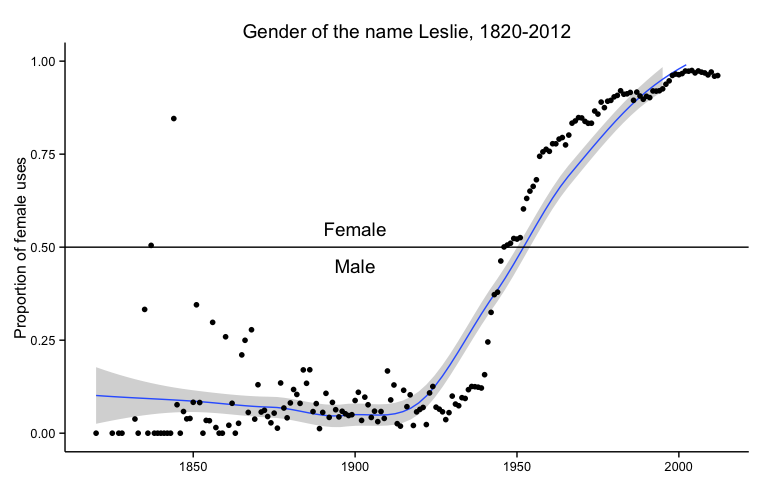
This paper addresses two trends in the humanities: the use of gender as an analytical framework and the proliferation of large, machine-readable datasets. It outlines a new method for algorithmically inferring the gender of a name using historical data. In contrast to existing methods of predicting gender, this method takes into account how naming practices change over time and space. It provides a higher level of accuracy and a precise statement of the likelihood that a name was a given gender, especially for datasets where names can be associated with dates. The article describes the methodology as implemented in the [Gender](https://github.com/ropensci/gender) package for the [R programming language](http://www.r-project.org/). It goes on to apply the method to a case study in which we examine gender and gatekeeping in the history profession by inferring the gender of authors of history dissertations and of authors and reviewers in the *American Historical Review*. The Gender package illustrates how humanities scholars can make concrete methodologial contributions to computer science and related fields.

### Introduction: The "Leslie Problem"

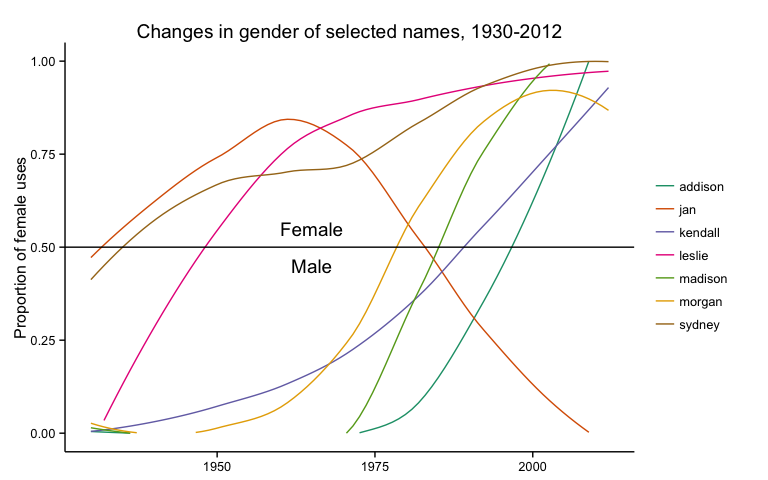
As humanities researchers grapple with larger and larger datasets they will increasingly be able to not just extract existing information from that data but to also infer additional information from it [Goldstone and Underwood 2014].[[1]](#endnote-1) Many of these datasets deal with lists of people: company payrolls, military rosters, passenger bookings, records of correspondence, or lists of published works. Some of these records might contain associated information such as a person's address, age, or rank, while others contain little more than a list of personal names. A common question in the history of American religion, for example, is whether more women than men participated in a church or religious group. A simple membership roll might contain names, but it is unlikely to identify the gender of the members. One might also ask a question about changing patterns in male and female authorship of books. Library catalogs contain enormous amounts of data about publication, but they do not record the sex of authors.

Given large but sparse datasets, researchers have to use their ingenuity to derive additional kinds of information. What can researchers infer from just a name? On a basic level, first names often imply a person's gender. "Jane Fay" is almost certainly a woman, while her brother "John Fay" is almost certainly a man. Given a dataset matching names to genders, a straightforward computer program can easily infer their respective genders.

The problem with inferring gender from names is that the link between gender and naming practices, like language itself, is not static. What about Jane and John Fay's sibling, Leslie? If Leslie was born in the past sixty years, chances are good that Leslie would be their sister. But if the three siblings were born in the early twentieth century, chances are good that Leslie would be their brother. That is because the conventional gender for the name Leslie switched over the course of one hundred years. In 1900 some 92% of the babies born in the United States who were named Leslie were male, while in 2000 about 96% of the Leslies born in that year were female.

 **Figure 1: Gender of the name Leslie, 1820-2012**

Other names used in the United States such as Madison, Morgan, Sydney, and Kendall, have also flipped their gender over the past century. Such changes usually follow set patterns. In general, names tend to shift from being predominantly male to being predominantly female; the shift rarely happens in the other direction. In general, names that transition from male to female also became increasingly popular over time. For example, the name Madison changed from male to female quite suddenly after 1985. Whereas before it had been an infrequently used male name, after 1985 it skyrocketed in popularity and by the early 2000s was one of the ten most frequently used names for baby girls. Broadly speaking, in the late twentieth century Americans girls became increasingly more likely than American boys to receive a "trendy" name, and once that name was associated with women it was unlikely to be given to male babies. These names present a moving target for researchers hoping to use them to infer additional information about the people under study. In addition, the prevalence of changes in the gender of names in the last half of the twentieth century implies that gendered lists of names drawn from contemporary data are anachronistic and so of little use for earlier periods.[[2]](#endnote-2)

 **Figure 2: Proportion of female uses**

For those working on contemporary subjects the "Leslie problem" is not an especially pressing one. An education policy analyst studying the demographics of different urban school districts, for instance, can turn to a variety of tools that use current databases of names, and these tools will likely identify the gender of Leslie Fay correctly. One online service, [Genderize.io](http://genderize.io/), determines the gender of first names "from user profiles across major social networks." [Genderize.io n.d.] Checking the service for the name Leslie returns a prediction that the name is female, along with an estimate of male/female proportions of the uses among members of unnamed social media sites.[[3]](#endnote-3)

gender("Leslie", method = "genderize")

## $name  
## [1] "Leslie"  
##   
## $gender  
## [1] "female"  
##   
## $proportion\_female  
## [1] 0.89  
##   
## $proportion\_male  
## [1] 0.11

But researchers studying a longer timespan need to take into account the changing nature of naming practices. If that same education researcher wanted to compare the demographics of contemporary urban schools with those from the early twentieth century, existing tools like Genderize.io would misidentify Leslie Fay's gender in historical datasets. The "Leslie problem" is not just for researchers who think of themselves as historians. Anyone studying a period longer than a few years, or anyone studying a group whose demographics do not match the groups used by a contemporary tool such as Genderize.io will also encounter this problem. As of 2012, the average American lifespan was nearly seventy-nine years---more than enough time for naming practices to change quite dramatically between their births and deaths [World Bank 2014]. Predicting gender from first names therefore requires a fundamentally historical method.

Our solution to the "Leslie problem" is to create a software package that combines a predictive algorithm with several historical datasets suitable to various times and regions. The algorithm calculates the proportion of male or female uses of a name in a given birth year or range of years. It thus can provide not only a prediction of the gender of a name, but also a measure of that prediction's likelihood. For example, our program predicts that a person named Leslie born in 1950 is female, but only just barely: 52 percent of the babies named Leslie born in the United States in 1950 were girls and 48 percent were boys. Including these probability figures allows researchers to conduct more nuanced analyses and to determine for themselves what level of certainty is acceptable for predicting the gender of names.

We recognize that gender is a fluid concept, both today and in the past, and that this method glosses over the complexities of historical definitions of gender and of the relationship between biological sex and gender [Butler 1990, Butler 1993]. In some sense this is unavoidable since the large historical datasets which make this kind of analysis possible were created by state agencies such as the U.S. Census Bureau. For the category of gender (as for race) states have offered a limited number of boxes to check, and those boxes exert a tremendous power to define state subjects [Brubaker and Cooper 2000, 14-17, Pascoe 2009, Canaday 2011, Landrum 2014]. This method does not presuppose that the gender binary imposed by governments and their wider societies reflect how an individual identifies oneself and is neutral as to what the categories "male" or "female" actually mean in any given historical context. Researchers are free---indeed, obligated---to make that determination for themselves, and this package provides access to the underlying dataset so that researchers can interrogate every assumption that comes with it. This method does not argue against the complexities of historical understandings of gender, but as with any project it does require researchers who use the method to think through how the data was gathered, what it implies, and how it reflects social and cultural practices.

The remainder of this article is divided into two sections. First, we describe our method in more detail, compare it to existing methods, and explain how to use it. Second, we apply the method to a case study of gatekeeping in the historical profession in order to demonstrate its usefulness.

### I. The Method

Why do scholars need a method for inferring gender that relies on historical data? Let's start with a comparison. One existing method for predicting gender is available in the [Natural Language Toolkit](http://www.nltk.org/) for the Python programming language [Bird et al. 2009, Chapter 2].[[4]](#endnote-4) The NLTK is an influential software package for scholarship because it provides an extraordinary range of tools for analyzing natural language.[[5]](#endnote-5) Included in the NLTK are two lists of male and female names created by the computer scientist Mark Kantrowitz with assistance from Bill Ross.[[6]](#endnote-6) These lists include 7,576 unique names. Using the Kantrowitz names corpus in the NLTK, one could look up a name like Jane or John and find out whether it is male or female.

The Kantrowitz corpus provides the list of names.[[7]](#endnote-7)

genderdata::kantrowitz

## Source: local data frame [7,579 x 2]  
##   
## name gender  
## 1 aamir male  
## 2 aaron male  
## 3 abbey either  
## 4 abbie either  
## 5 abbot male  
## 6 abbott male  
## 7 abby either  
## 8 abdel male  
## 9 abdul male  
## 10 abdulkarim male  
## .. ... ...

One can then easily write a function which looks up the gender of a given name.

gender("abby", method = "kantrowitz")

## $name  
## [1] "abby"  
##   
## $gender  
## [1] "either"

The most significant problem with the Kantrowitz names corpus and thus the NLTK implementation is that it assumes that names are timeless. As pointed out above, this makes the corpus problematical for historical purposes. Furthermore the Kantrowitz corpus includes other oddities which make it less useful for research. Some names such as Abby are overwhelmingly female and some such as Bill are overwhelmingly male, but the corpus includes them as both male and female. The Kantrowitz corpus contains only 7,576 unique names, a mere 8.3% of the 91,320 unique names in a dataset provided by the Social Security Administration and 2.23% of the 339,967 unique names in the census records provided in the Integrated Public Use Microdata Series (IPUMS) USA dataset. There are therefore many names that it cannot identify. Assuming for the moment that our method provides more accurate results, we estimate that 4.74% percent of the names in the Kantrowitz corpus are classified ambiguously when a gender could be reasonably predicted from the name, that 1.24% percent of the names are classified as male when they should be classified as female, and that 1.82% are classified as female when they should be classified as male. This error rate is a separate concern from the much smaller size of the Kantrowitz corpus.[[8]](#endnote-8)

We mention the Kantrowitz name corpus as implemented in NLTK because the Natural Language Toolkit is rightly regarded as influential for scholarship. Its flaws for predicting gender, which are a minor part of the software's total functionality, are also typical of the problems with most other implementations of gender prediction algorithms. The Genderize.io API is, for example, a more sophisticated implementation of gender prediction than the NLTK algorithm. Besides predicting male or female for gender, it also reports the proportion of male and female names, along with a count of the number of uses of the name on which it is basing its prediction. Genderize will also permit the user to customize a prediction for different countries, which is an important feature. Genderize.io reports that its "database contains 142848 distinct names across 77 countries and 85 languages." Genderize is unsuitable for historical work, however, because it is based only on contemporary data. According to the documentation for its API, "It utilizes big datasets of information, from user profiles across major social networks." It would be anachronistic to apply these datasets to the past, and Genderize.io provides no functionality to filter results chronologically as it does geographically. In addition, Genderize.io does not make clear exactly what comprises the dataset and how it was gathered, which keeps scholars from interrogating the value of the source [Genderize.io 2014-].[[9]](#endnote-9)

R and Python are two of the most commonly used languages for data analysis [Piatesky 2014, Muenchen 2014-].[[10]](#endnote-10) Python's existing packages for gender prediction all implement a method similar to the NLTK or Genderize.io.[[11]](#endnote-11) To our knowledge, [CRAN](http://cran.rstudio.com/) (the central repository for R packages) does not include any packages dedicated to gender prediction. Thus two of the most popular data analysis languages used in the digital humanities currently have no satisfactory method for predicting gender from names for historical and humanities research.

To that end we have created the [Gender](https://github.com/ropensci/gender) package for R which includes both the predictive algorithm and an associated [genderdata](https://github.com/lmullen/gender-data-pkg) package containing various historical datasets. This R implementation is based on an [earlier Python implementation](https://github.com/cblevins/Gender-ID-By-Time) by [Cameron Blevins](http://www.cameronblevins.org/) and [Bridget Baird](http://www.conncoll.edu/directories/emeritus-faculty/bridget-baird/). The Gender package is affiliated with [rOpenSci](http://ropensci.org/), an initiative that supports reproducible research and open data for scientists using R. The rOpenSci advisory board has provided code review and publicity for the project.[[12]](#endnote-12)

The possibilities for inferring gender from names depends on two things. First, it needs a suitable (and suitably large) dataset for the time period and region under study. Unsurprisingly such datasets are almost always gathered in the first instance by governments, though their compilation and digitization may be undertaken by scholarly researchers. Second, it depends on a suitable algorithm for estimating the proportion of male and female names for a given year or range of years, since often a person cannot be associated with an exact date. It is especially important that the algorithm take into account any biases in the data to formulate more accurate predictions. Development of the R package has had two primary aims. The first is to abstract the predictive algorithm to the simplest possible form so that it is usable for a wide range of historical problems rather than depending on the format of any particular data set. The second has been to provide as many datasets as possible in order to localize the predictions to particular times and places. To that end, [Benjamin Schmidt](http://benschmidt.org/) contributed a dataset of US Census Data and accompanying code, as well as provided a significant correction to predictions based on Social Security Administration data.

The Gender package currently uses two datasets which make it suitable for studying the United States from the first federal census in 1790 onwards.[[13]](#endnote-13) The first dataset contains names of applicants for Social Security and is available from [Data.gov](https://www.data.gov/). The second dataset contains names gathered in the decennial censuses and is available from the IPUMS-USA (Integrated Public Use Microdata Series) service from the Minnesota Population Center at the University of Minnesota.

The Social Security Administration (SSA) [Baby Names](http://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data) dataset was created as a result of the Social Security Act of 1935 during the New Deal.[[14]](#endnote-14) This dataset contains a list of first names along with how many times each name was assigned to each gender in every year beginning with 1880 and ending with 2012. The SSA list includes any name which was used more than five times in a given year, thereby capturing all but the most infrequently used names from each year. The description "baby names" provided by the SSA is a serious misnomer. When Social Security became available during the New Deal, its first beneficiaries were adults past or near retirement age. The dataset goes back to 1880, the birth year for a 55 year-old adult when Social Security was enacted. Even after 1935, registration at birth for Social Security was not mandatory until 1986. As we will demonstrate below, the way in which the data was gathered requires an adjustment to our predictions of gender.[[15]](#endnote-15)

The IPUMS-USA dataset, contributed by Benjamin Schmidt, contains records from the United States decennial census from 1790 to 1930. This dataset includes the birth year and numbers of males and females under the age of 62 for all the years in that range. This data has been aggregated by IPUMS at the University of Minnesota and is released as a sample of the total census data. Unlike the SSA dataset, which includes a 100% sample for every name reported to the Social Security Administration and used more than five times, the IPUMS data contains 5% or 10% samples of names from the total census data. Because the gender() function relies on proportions of uses of names, rather than raw counts of people with the names, the sampling does not diminish the reliability of the function's predictions [Ruggles et al. 2010].

These two SSA and IPUMS datasets (and any future datasets to be added to the package) thus contain a simple table of data.[[16]](#endnote-16) The columns contain the name, year, and number of female and male uses of that name in a particular year.

genderdata::ssa\_national

## Source: local data frame [1,603,026 x 4]  
##   
## name year female male  
## 1 aaban 2007 0 5  
## 2 aaban 2009 0 6  
## 3 aaban 2010 0 9  
## 4 aaban 2011 0 11  
## 5 aaban 2012 0 11  
## 6 aabha 2011 7 0  
## 7 aabha 2012 5 0  
## 8 aabid 2003 0 5  
## 9 aabriella 2008 5 0  
## 10 aadam 1987 0 5  
## .. ... ... ... ...

Our method for predicting gender is best understood through a series of examples. First we will use it to predict the gender of a single name in order to demonstrate a simplified version of the inner workings of the function. We will then apply it to a small sample dataset to show how a researcher might use it in practice.

**Example #1: A Sample Name**

The method for predicting gender from a name using the package's datasets is simple. Let's begin by assuming that we want to predict the gender of someone named Sidney who was born in 1935 using the Social Security Administration dataset. Because the dataset contains a list of names for each year, we can simply look up the row for Sidney in 1935. Using the [dplyr](https://github.com/hadley/dplyr) package for R, which provided a grammar for data manipulation, this can be expressed with the action "filter":

genderdata::ssa\_national %>%  
 filter(name == "sidney", year == 1935)

## Source: local data frame [1 x 4]  
##   
## name year female male  
## 1 sidney 1935 93 974

Thus, according to the Social Security Administration, there were 974 boys and 93 girls named Sidney born in 1935. We can add another command to calculate the *proportion* of females and males ("mutate" in the [dplyr](https://github.com/hadley/dplyr) package vocabulary) rather than raw numbers.

genderdata::ssa\_national %>%  
 filter(name == "sidney", year == 1935) %>%  
 mutate(proportion\_female = female / (male + female),  
 proportion\_male = 1 - proportion\_female)

## Source: local data frame [1 x 6]  
##   
## name year female male proportion\_female proportion\_male  
## 1 sidney 1935 93 974 0.08716 0.9128

In other words, there is an approximately 91.3% percent chance that a person born in 1935 named Sidney was male. In 2012, for comparison, there was an approximately 60.7% percent chance that a person born named Sidney was female.

The method is only slightly more complex if we do not know the exact year when someone was born, as is often the case for historical data. Suppose we know that Sidney was born in the 1930s but cannot identify the exact year of his or her birth. Using the same method as above we can look up the name for all of those years.

genderdata::ssa\_national %>%  
 filter(name == "sidney", year >= 1930 & year <= 1939)

## Source: local data frame [10 x 4]  
##   
## name year female male  
## 1 sidney 1930 48 1072  
## 2 sidney 1931 48 940  
## 3 sidney 1932 57 958  
## 4 sidney 1933 77 949  
## 5 sidney 1934 78 930  
## 6 sidney 1935 93 974  
## 7 sidney 1936 81 952  
## 8 sidney 1937 63 902  
## 9 sidney 1938 89 875  
## 10 sidney 1939 63 861

Next we can sum up the male and female columns ("summarize" in [dplyr](https://github.com/hadley/dplyr) package vocabulary) and calculate the proportions of female and male uses of "Sidney" during that decade.

genderdata::ssa\_national %>%  
 filter(name == "sidney", year >= 1930 & year <= 1939) %>%  
 group\_by(name) %>%  
 summarize(female = sum(female),  
 male = sum(male)) %>%  
 mutate(proportion\_female = female / (male + female),  
 proportion\_male = 1 - proportion\_female)

## Source: local data frame [1 x 5]  
##   
## name female male proportion\_female proportion\_male  
## 1 sidney 697 9413 0.06894 0.9311

In other words, for the decade of the 1930s, we can calculate that there is a 93.2% percent chance that a person named Sidney was male. This is roughly the same as the probability we calculated above for just 1935, but our method also returns the figures it used to calculate those probabilities: 1,067 instances of "Sidney" in 1935 versus 10,110 total instances for the decade as a whole.

**Example #2: A Sample Dataset**

The method's real utility stems from being able to process larger datasets than a single name. Let's use, for example, a hypothetical list of editors from a college newspaper to illustrate how a researcher might apply it to their own data. The package's prediction function allows researchers to choose which reference datasets they would like to use and the range of years for making their predictions.[[17]](#endnote-17)

editors

## Source: local data frame [6 x 2]  
##   
## name birth\_year  
## 1 Madison 1934  
## 2 Madison 1990  
## 3 Morgan 1948  
## 4 Morgan 1970  
## 5 Jan 1965  
## 6 Jan 1998

The simplest way to use the package would be to call the gender() function, specifying the method, in this case the Social Security Administration dataset since all the names are after 1930, and convert the results to a data frame (one of R's native data structures).

gender(editors$name, method = "ssa") %>% do.call(rbind.data.frame, .)

## name proportion\_male proportion\_female gender year\_min year\_max  
## 2 Madison 0.0162 0.9838 female 1932 2012  
## 21 Madison 0.0162 0.9838 female 1932 2012  
## 3 Morgan 0.1467 0.8533 female 1932 2012  
## 4 Morgan 0.1467 0.8533 female 1932 2012  
## 5 Jan 0.2680 0.7320 female 1932 2012  
## 6 Jan 0.2680 0.7320 female 1932 2012

This basic call of the gender() function would be suitable if we did not know the birth year for each editor. But we can also specify that information in the gender() function by using this call:

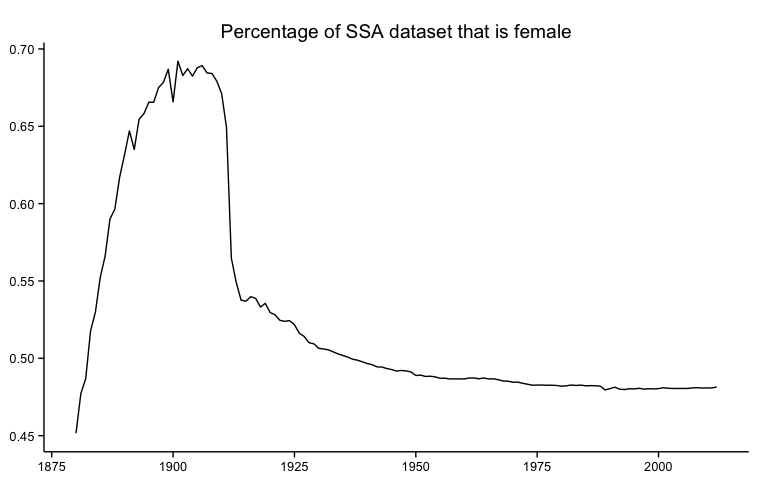
Map(gender, editors$name, years = editors$birth\_year) %>%   
 do.call(rbind.data.frame, .) %>% print.data.frame(row.names = FALSE)

## name proportion\_male proportion\_female gender year\_min year\_max  
## Madison 1.0000 0.0000 male 1934 1934  
## Madison 0.0870 0.9130 female 1990 1990  
## Morgan 1.0000 0.0000 male 1948 1948  
## Morgan 0.8074 0.1926 male 1970 1970  
## Jan 0.1813 0.8187 female 1965 1965  
## Jan 0.8604 0.1396 male 1998 1998

By taking into account the year of birth, we find that four of our six names were likely male, whereas we might otherwise have predicted that all six were female. We also now know the approximate likelihood that our predictions are correct: at least 80% for all of these predictions.

It is also possible to use a range of years for a dataset like this. If our list of editors contained the year in which the person served on the newspaper rather than the birth year, we could make a reasonable assumption that their ages were likely to be between 18 and 24. We could then calculate a minimum and a maximum year of birth for each, and run the prediction function on that range for each person. The exact code to accomplish these types of analysis can be found in the Gender package's vignette.

As previously mentioned, the history of how the Social Security Administration collected the data effects its validity. Specifically, because the data extends back to 1880 but the first applications were gathered after 1935, the sex ratios in the dataset are skewed in the years before 1930. For example, this dataset implies that thirty percent of the people born in 1900 were male.

 **Figure 3: Percentage of SSA dataset that is female**

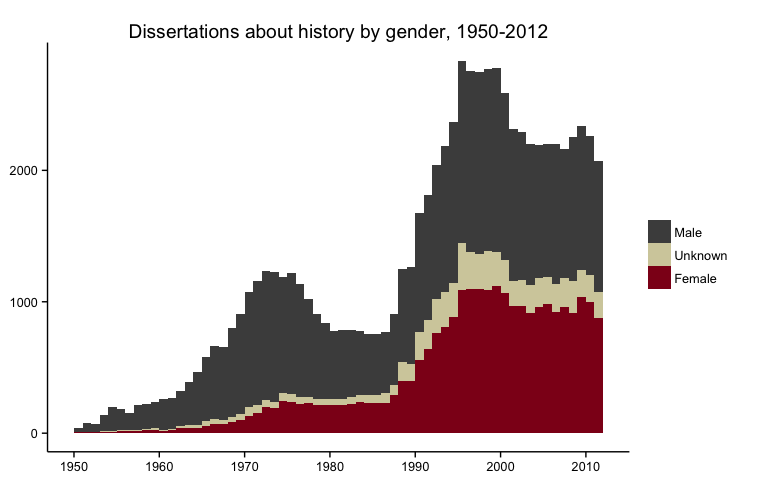
It is extremely improbable that nearly seventy percent of the people born in 1900 were female. Exactly why the dataset has this bias is unclear, though we speculate that it is because the applicants for Social Security in the early years of the program were approaching retirement age, which was set much closer to the average life expectancy in 1935 than it is today. Since women tend to live longer than men, they were overrepresented in Social Security applications.

The solution to this problem is two-fold. First, we recommend that researchers use the IPUMS-USA dataset to make predictions for years from 1790 to 1930 (which avoids the "bubble" in the SSA data) and that they use the SSA dataset for years after 1930.[[18]](#endnote-18) Second, we have built in a correction to the SSA dataset when using the gender() function. If we assume that the secondary sex ratio (that is, the ratio of male to female births) in any given year does not deviate from 0.5 (that is, equality), it is possible to calculate a correction factor for each year or range of years to even out the dataset. We apply this correction factor automatically when using the SSA dataset.

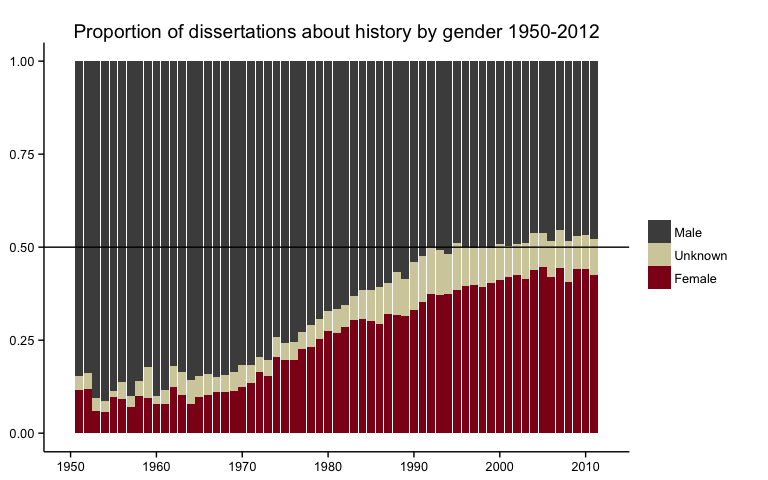
### II. Measuring Gatekeeping in the Historical Profession

In a 2005 report for the American Historical Association, Elizabeth Lunbeck acknowledged "a sea change in the [historical] profession with respect to gender" before going on to describe the painful limitations of this sea change for female historians: ongoing personal discrimination, lower salaries, and barriers to securing high-ranking positions. Lunbeck's report drew in part on a survey of 362 female historians that produced a rich source of responses detailing the deep and multi-faceted challenges facing women in the profession [Lunbeck 2005]. What follows is a quantitative supplement to Lunbeck's analysis that uses our program to analyze gender representation amongst historians across a much larger scale and a much longer time period. We have found that although women have achieved rough equality in terms of the number of dissertations written, there is still a significant gap between male and female authorship of books in history.

Our analysis focuses on one of the bedrocks of the historical profession: scholarly research. We begin with the history dissertation, often the defining scholarly output of a historian's early career. The completion of a dissertation marks a standardized moment of transition out of the "training" phase of the historical profession. To study dissertations about history, we used data supplied by ProQuest for roughly eighty thousand PhD dissertations completed in the United States.[[19]](#endnote-19) Identifying the gender of their authors identifies the approximate number of women and men who are completing PhD-level training in history each year. Our program uses the year a historian wrote his or her dissertation to estimate a period for when they might have been born: we have assumed that a historian was between 25 and 45 years old when completing the dissertation. Using this temporal information, we are better able to infer their gender and chart how the larger representation of women and men changes year-by-year.

 **Figure 4: Dissertations about history by gender, 1950-2012**

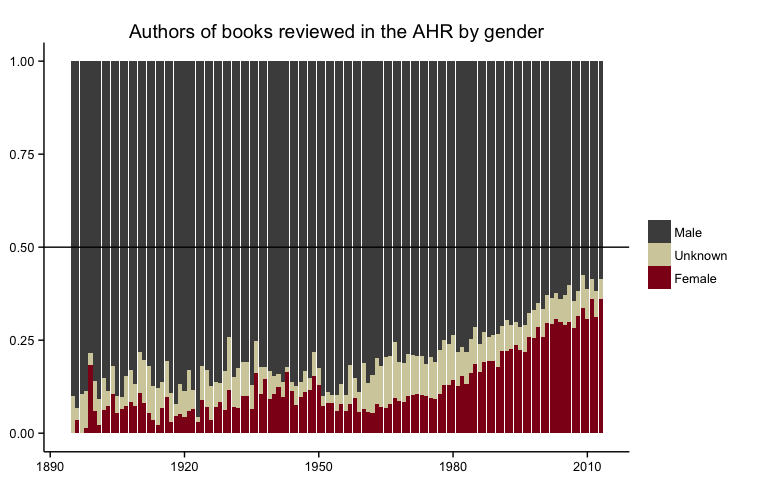
Another way to examine this trend is to look at the proportion of male and female dissertation authors over time, which smoothes out changes in the absolute number of dissertations produced. The proportion of dissertations written by women has steadily increased over the past half-century, a change that began in the late 1960s and continued through the early 2000s. Since that point, the proportion of dissertations written by women has largely plateaued at a few percentage points below the proportion written by men. Female historians have achieved something approaching parity with male historians in terms of how many women and women complete dissertations each year.

 **Figure 5: Proportion of dissertations about history by gender 1950-2012**

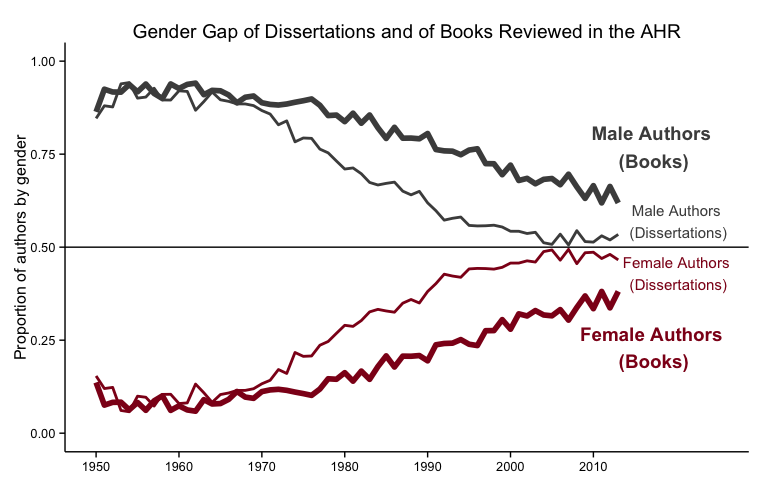
But what happens after the dissertation? It is, after all, only the first major stage of a historian’s research output. The "coin of the realm" for many historians remains the publication of a scholarly monograph to be read and evaluated by peers. Once published, the impact of a book is often shaped by its reviews in academic journals. One of the leading journals in the historical profession is the *American Historical Review*, the flagship journal of the American Historical Association. Along with a few dozen articles published in its five issues each year, the *AHR* publishes roughly one thousand books reviews per year covering (in its own words) "every major field of historical study." [American Historical Review n.d.] The *AHR* is not only one of the widest-ranging journals in the profession, it is also the oldest; the journal has been publishing continuously since 1895. The range, scope, longevity, and reputation of the *AHR* makes it an ideal proxy to study gatekeeping within the historical discipline.

A few caveats are in order. First, a book that never appears in the *AHR* at all may still go on to have a significant impact on the profession. Second, a book's appearance in the *AHR* is not necessarily correlated with the *quality* of the book, as the journal prints negative reviews as well as positive ones. Nevertheless, when a book appears in the *AHR* it serves as a signal that other historians in positions of power have "approved" it for consideration. Even if it garners a negative review, the fact that it appears in the journal at all is a measure of the fact that the profession's gatekeepers have deemed it important enough to review. It is precisely this professional gatekeeping dimension that we can analyze using the Gender package.

Scraping the table-of-contents of every issue of the American Historical Review results in a dataset of close to 60,000 books reviewed by the journal since it began publication in 1895. Our program then inferred the gender of the authors of these books, which we could in turn use to plot the proportion of female and male authors over time. The temporal trajectory of gender representations roughly resembles that of history dissertations: women began making inroads in the late 1960s and have made steady gains over the past four decades. By the twenty-first century the proportion of female authors reviewed in the *AHR* had more than tripled.

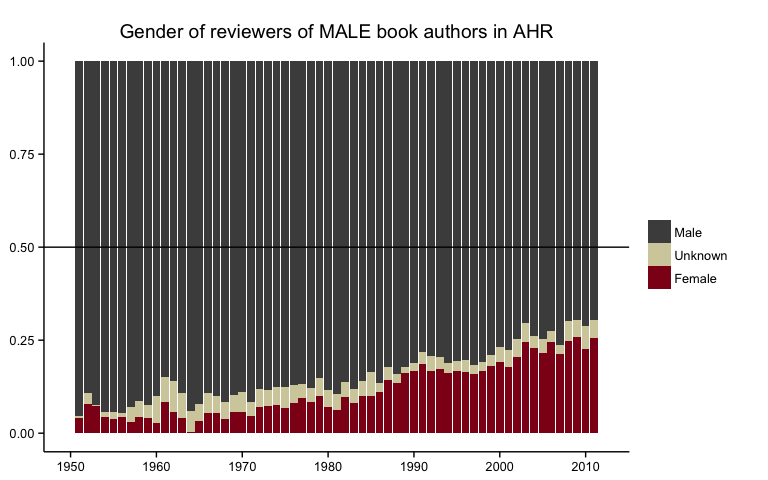
 **Figure 6: Authors of books reviewed in the AHR by gender**

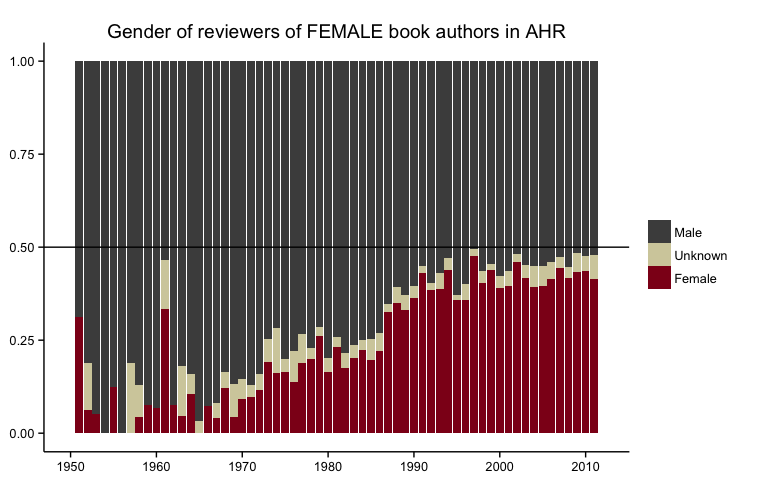
But a closer look shows important differences between the proportion of women writing history dissertations and the proportion of women appearing in the *AHR*. Although both have trended upwards since the 1970s the slope is a lot steeper for dissertations than it is for the *AHR*. Female historians have very nearly closed the gap in terms of newly completed dissertations, but the glass ceiling remains stubbornly low in terms of what happens from that point onwards. In book reviews published in the *AHR* male authors continue to outnumber female authors by a factor of nearly 2 to 1. The long-term gains made by female doctoral students do not carry over into the pages of the historical profession's leading journal.[[20]](#endnote-20)

 **Figure 7: Gender Gap of Dissertations and of Books Reviewed in the *AHR***

The path from a dissertation to a monograph to a book review in the *American Historical Association* isn't automatic. Not all dissertations are published as books, not all books are submitted to the *American Historical Review*, and not all book submissions are accepted by the *AHR* editors for review. As of September 2014 the journal's website reminds readers that "the AHR receives over 3,000 books a year; we have the resources to publish at most 1,000 reviews a year (approximately 200 per issue)." [American Historical Review n.d.] But fewer and fewer female historians make it through each successive step. Establishing a single causal reason for this pattern is all but impossible, and instead a murky concoction of interrelated factors likely contributes to the successively widening gap. A 2011 survey by the American Historical Association revealed that female historians dedicated substantially more time to child and elder care than their male colleagues, leaving them with less time for research. [Townsend 2012] This gender gap is more likely to impact the production of a book than a dissertation, as graduate students are generally younger and less likely to be juggling child-rearing responsibilities while writing than their older peers. Their male colleagues, meanwhile, enjoy more time to publish and receive more accolades when they do. As of 2014, the Pulitzer Prize for History has been awarded to exactly one female historian in the past nineteen years. The Bancroft Prize in American History has been awarded to almost four times as many male historians as female historians over that same period. ["Bancroft Prize"] The *American Historical Review* is, unfortunately, far from alone in its gatekeeping practices.

We went on to use our program to infer the gender of not just book authors reviewed by the *AHR*, but the gender of the reviewers themselves. The story is much the same for reviewers as it is for authors: more than twice as many men as women appear as reviewers in the journal. But gender inflects book reviews in less direct ways by shaping who writes reviews of which book authors. About three times as many men write reviews of male-authored books as do women. In the case of female-authored books the ratio of female-to-male reviewers is much closer to 50/50. In short, women are much more likely to write reviews of other women. And while men still write the majority of reviews of female-authored books, they tend to gravitate more towards male authors – who are, of course, already over-represented in the *AHR*.

 **Figure 8: Gender of reviewers of MALE book authors in *AHR***

 **Figure 9: Gender of reviewers of FEMALE book authors in *AHR***

One 2013 analysis of 2,500 recent history Ph.D.'s found that "gender played little role in employment patterns across particular professions and industries" [Wood and Townsend 2013]. Our own analysis reveals that the story is different for book publishing than for employment. The flagship journal of professional historians continues to publish twice as many reviews of books by male authors as female authors. This disparity is put into even sharper relief when set against the relative parity achieved by women in producing history dissertations. Women might be getting hired at comparable rates to men, but discriminatory gatekeeping remains alive and well in the profession. As explicit instances of sexism fade from view the disadvantages women face are growing more and more opaque. But that does not make them any less damaging. As discrimination becomes less blatant and more systemic, digital methods like the Gender package are a way to a shine light on these cloaked disadvantages, to reveal the true extent of the problem and the work that remains to be done.

### Conclusion

Our algorithm for predicting gender from first names, when combined with historical datasets in the Gender package, provides a powerful new method for researchers interested in the study of gender. The digital turn has made available very large datasets, yet in the midst of abundance many humanists continue to deal with the problem of incompleteness and scarcity within their sources. Our method uses *abundant* existing datasets to infer additional information from *scarce* datasets. The temporal dimension of our work matches historical questions to a historical method, providing much improved results over simple, anachronistic, or ahistorical lookup methods. This method is also extensible to different times and places, provided that a researcher has a suitable reference dataset of names and genders from the period or area under study. The Gender package reverses the traditional disciplinary direction between the humanities and computer science. Rather than adapting existing computational methodologies, our method uses a humanistic starting point---the historical relationship between gender and naming practices---to contribute a new method to computer science.

The analytical category of gender has transformed humanities research over the past half-century. More recently, the digitization movement has opened up opportunities to apply this category to new sources and in new ways. Projects in the digital humanities have ranged from recovering the voices of early-modern women writers to tracing the linguistic styles of male and female authors [Women Writers Project n.d., Argamon et. al. 2009]. The proliferation of machine-readable datasets and digital texts brings with it a vast trove of material with which to study gender. But unlocking its full range of meaning necessitates not just accessing the data that *is* there, but inferring information that *isn't* there. The Gender package is a powerful tool to do just that.

### References

[American Historical Review n.d.] American Historical Review. "Book Reviewing in the AHR." <http://www.historians.org/publications-and-directories/american-historical-review/book-review-guide>. Accessed 29 September 2014.

[Argamon et. al. 2009] Argamon, Shlomo, Jean-Baptiste Goulain, Russell Horton, and Mark Olsen. "Vive la Différence! Text Mining Gender Difference in French Literature" *Digital Humanities Quarterly*. 3.2 (2009). <http://www.digitalhumanities.org/dhq/vol/3/2/000042/000042.html>

["Bancroft Prize"] "Bancroft Prize." *Wikipedia* <http://en.wikipedia.org/wiki/Bancroft_Prize>. Accessed 5 October 2014.

[Bird et al. 2009] Bird, Steve, Edward Loper, and Ewan Klein. *Natural Language Processing with Python* (O’Reilly, 2009).

[Brubaker and Cooper 2000] Brubaker, Rogers, and Frederick Cooper. "Beyond 'identity.'" *Theory and Society* 29.1 (2000): 1–47.

[Butler 1989] Butler, Judith. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge, 1990.

[Butler 1993] Butler, Judith. *Bodies That Matter: On the Discursive Limits of Sex*. New York: Routledge, 1993.

[Canaday 2011] Canaday, Margot. *The Straight State: Sexuality and Citizenship in Twentieth-Century America*. Princeton, N.J.: Princeton University Press, 2011.

[Genderize.io n.d.] *Genderize.io*. <http://genderize.io/>. Accessed 15 September 2014.

[Goldstone and Underwood 2014] Goldstone, Andrew, and Ted Underwood. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* (Forthcoming 2014).

[Healy 2013] Healy, Kieran. "Using Metadata to Find Paul Revere." *kieranhealy.org*. 9 June 2013, <http://kieranhealy.org/blog/archives/2013/06/09/using-metadata-to-find-paul-revere/>.

[Landrum 2014] Landrum, Shane. "The State's Big Family Bible: Birth Certificates, Personal Identity, and Citizenship in the United States, 1840--1950." PhD dissertation, Brandeis University, August 2014.

[Lunbeck 2005] Lunbeck, Elizabeth, et al. "The Status of Women in the Historical Profession." American Historical Association Committee on Women Historians. May 2005, <http://www.historians.org/Documents/About%20AHA%20and%20Membership/CWH-Report_5.20.05.pdf>.

[Muenchen n.d.] Muenchen, Robert A. "The Popularity of Data Analysis Software," *r4stats.com*. <http://r4stats.com/articles/popularity/>. Accessed 12 September 2014.

[Pascoe 2009] Pascoe, Peggy. *What Comes Naturally: Miscegenation Law and the Making of Race in America*. Oxford University Press, USA, 2009.

[Piatetsky 2013] Piatetsky, Gregory. "Top Languages for analytics, data mining, data science," *KDN Nuggets*. 27 August 2013, <http://www.kdnuggets.com/2013/08/languages-for-analytics-data-mining-data-science.html>.

["The Pulitzer Prizes: History" 2014-] "The Pulitzer Prizes: History." 2014-. <http://www.pulitzer.org/bycat/History> Accessed 5 October 2014.

[R Core Team 2014-] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. <http://www.R-project.org/>

[Ruggles et al. 2010] Ruggles, Steven, J. Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek. Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]. Minneapolis, MN: Minnesota Population Center [producer and distributor], 2010.

[Social Security Administration 2014-a] Social Security Administration. "Baby Names from Social Security Card Applications-National Level Data". Data.gov dataset ID US-GOV-SSA-338. Accessed 31 August 2014, <http://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data>.

[Social Security Administration 2014-b] Social Security Administration. "Baby Names from Social Security Card Applications-Data by State and District of Columbia." Data.gov dataset ID US-GOV-SSA-339. Accessed 31 August 2014, <http://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-data-by-state-and-district-of->.

[Townsend 2012] Townsend, Robert B. "What Makes a Successful Academic Career in History? A Field Report from the Higher Ranks." *Perspectives on History*. December 2012, <http://www.historians.org/publications-and-directories/perspectives-on-history/december-2012/what-makes-a-successful-academic-career-in-history>

[Weingart 2014] Weingart, Scott. "The Moral Role of DH in a Data-Driven World." *the scottbot irregular*. 9 September 2014, <http://www.scottbot.net/HIAL/?p=40944>

[Wickham 2014] Wickham, Hadley. "Tidy Data." *Journal of Statistical Software*, 50 (2014): <http://www.jstatsoft.org/v59/i10>.

[Wickham and Francois 2014] Wickham, Hadley, and Romain Francois. *dplyr: dplyr: a grammar of data manipulation*. R package version 0.2 (2014): <http://CRAN.R-project.org/package=dplyr>.

[Women Writers Project n.d.] Women Writers Project. "About." <http://www.wwp.northeastern.edu/about/>. Accessed 5 October 2014.

[Wood and Townsend 2013] Wood, L. Maren, and Robert B. Townsend. *The Many Careers of History PhDs: A Study of Job Outcomes*. American Historical Association, 2013: <http://historians.org/Documents/Many_Careers_of_History_PhDs_Final.pdf>.

[World Bank 2014-] World Bank. "Life expectancy at birth, total (years)." *World Bank Open Data*. <http://data.worldbank.org/indicator/SP.DYN.LE00.IN>. Accessed 12 September 2014

1. The availability of large-scale datasets dealing with personal information raises important ethical issues for digital humanists, particularly in light of the revelations about domestic spying at the NSA brought forth by Edward Snowden in 2013. [Weingart 2014] Kieran Healy has used network analysis to demonstrate how researchers can infer information from datasets while also touching on the troubling ethical issue of surveillance [Healy 2013]. [↑](#endnote-ref-1)
2. The few exceptions of names which flip from being female to male, such as "Jan", have a different causal mechanism. We speculate that Jan as an English name is female but that Jan as a Scandinavian name is male. Because the name has remained relatively uncommon, other causes besides changing gender preferences can explain the change. [↑](#endnote-ref-2)
3. In this example the Genderize.io service has been accessed through an API wrapper provided in our Gender package described below. [↑](#endnote-ref-3)
4. Chapter 6 of [Bird et al. 2009] also describes a method of predicting gender from the last letter of the name, but this example is given for the sake of explaining supervised classification and not as an end in itself, so we do not have reason to address this as a historical method. [↑](#endnote-ref-4)
5. [Bird et al. 2009] has been cited 812 times according to [Google Scholar](http://scholar.google.com/scholar?cites=3735380965977150091&as_sdt=5,47&sciodt=0,47&hl=en). Because in the humanities (unlike in other fields) the practice of citing software essential to performing scholarly analysis is regrettably rare, even nonexistent, such metrics are difficult to determine. Yet searching journals such as *Digital Humanities Quarterly* and the *Journal of Digital Humanities* reveals that natural language processing, for which the NLTK is the foremost implementation, is one of the key methodologies of the digital humanities. [↑](#endnote-ref-5)
6. The [original lists of names](http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names/0.html) are available at the website of the computer science department at Carnegie Mellon University, and the NLTK implementation is described in its [documentation](http://www.nltk.org/howto/corpus.html). [↑](#endnote-ref-6)
7. The examples below is drawn from our Gender package which implements the Kantrowitz dataset for sake of comparsion, though its use is not recommended for actual research. The Gender package combines the Kantrowitz corpus's two lists into a single list, identifying names that are ambiguous because they appear on both lists as "either". [↑](#endnote-ref-7)
8. This error rate was calculated using the Social Security Administration database, assuming a range of years between 1932 and 2012, which is the default setting within the Gender package. Of course, the error rate would fluctuate for different time periods and regions. [↑](#endnote-ref-8)
9. Quotations are taken from the [Genderize.io website](http://genderize.io), with correspondence from the site's developer for confirmation. [↑](#endnote-ref-9)
10. Surveys and analysis of useage point out the growth of R and the continuing popularity of Python for "data science" generally. [↑](#endnote-ref-10)
11. These Python libraries include [estimate.gender](https://pypi.python.org/pypi/estimate.gender/0.4), [genderize](https://github.com/SteelPangolin/genderize), and [gender.py](https://github.com/block8437/gender.py) [↑](#endnote-ref-11)
12. The gender package depends especially on [R Core Team 2014-, Wickham and Francois 2014]. [↑](#endnote-ref-12)
13. The authors of the package are both American historians, which explains why the initial release of the package focuses on that country. We have plans, however, to extend the underlying data to six other European and North American countries included in the [North Atlantic Population Project](https://www.nappdata.org/napp/) for the nineteenth century. We also welcome contributions of datasets for other regions and times from other scholars. [↑](#endnote-ref-13)
14. The Social Security Administration also provides a related dataset which contains the names of applicants broken down by state and year since 1910 [Social Security Administration 2014-b]. [↑](#endnote-ref-14)
15. Shane Landrum has studied the history of birth registration in the United States [Landrum 2014]. It is well-known among American historians that the Social Security Act was written to exclude farmworkers as a means of also excluding African Americans. In examining the dataset we have not found any reason to believe that this policy noticably skewed the prediction of gender from names. [↑](#endnote-ref-15)
16. The format of this data departs from the more generally desirable "tidy data" framework proposed by Hadley Wickham for performance reasons. By keeping the data in a "wider" format instead of a tidy format as defined by Wickham, the function is able to make its predictions faster [Wickham 2014b]. [↑](#endnote-ref-16)
17. See the package vignette for more details: vignette("predicting-gender", package = "gender"). [↑](#endnote-ref-17)
18. The IPUMS-USA dataset includes only apparently random fluctuations in the gender ratio in any given year. [↑](#endnote-ref-18)
19. ProQuest's dataset does not permit us to limit dissertations just to dissertations produced in history departments. These dissertations include dissertations *about* history. For a justification of this approach, see a [series of blog posts](http://lincolnmullen.com/research/history-dissertations/) exploring the dataset. The American Historical Association's dataset of dissertations relies on reporting from departments and does not have the chronological range of the ProQuest data. Nevertheless, applying the same method to the AHA's dataset does not provide any reason for thinking that the proportion of authorship by gender over time is substantially different when limited to just history department dissertations. American Historical Association, *Directory of History Dissertations*, <https://secure.historians.org/pubs/dissertations/>. Data export provided to author by Liz Townsend, March 3, 2014. [↑](#endnote-ref-19)
20. The chart below includes only authors whose gender we successfully inferred; it does not include authors of unknown gender as in the previous charts. [↑](#endnote-ref-20)