Analysis of science journalism reveals gender and regional disparities in coverage

This manuscript (permalink) was automatically generated from greenelab/nature_news_manuscript@e0b1ead on March 15, 2024.

Authors

- **Natalie R. Davidson**
  0000-0002-1745-8072 · nrosed · n_rose_d
  University of Colorado School of Medicine, Aurora, Colorado, United States of America; Center for Health AI · Funded by The Gordon and Betty Moore Foundation (GBMF 4552)

- **Casey S. Greene** ✉
  0000-0001-8713-9213 · cgreene · GreeneScientist
  University of Colorado School of Medicine, Aurora, Colorado, United States of America; Center for Health AI · Funded by The Gordon and Betty Moore Foundation (GBMF 4552)

✉ — Correspondence possible via GitHub Issues or email to Casey S. Greene <casey.s.greene@cuanschutz.edu>.

Abstract

Science journalism is a critical way for the public to learn about and benefit from scientific findings. Such journalism shapes the public's view of the current state of science and legitimizes experts. Journalists can only cite and quote a limited number of sources, who they may discover in their research, including recommendations by other scientists. Biases in either process may influence who is identified and ultimately included as a source. To examine potential biases in science journalism, we analyzed 22,001 non-research articles published by Nature and compared these with Nature-published research articles with respect to predicted gender and name origin. We extracted cited authors' names and those of quoted speakers. While citations and quotations within a piece do not reflect the entire information-gathering process, they can provide insight into the demographics of visible sources. We then predicted gender and name origin of the cited authors and speakers. We compared articles with a comparator set made up of first and last authors within primary research articles in Nature and a subset of Springer Nature articles in the same time period. In our analysis, we found a skew toward quoting men in Nature science journalism. However, quotation is trending toward equal representation at a faster rate than authorship rates in academic publishing. Gender disparity in Nature quotes was dependent on the article type. We found a significant over-representation of names with predicted Celtic/English origin and under-representation of names with a predicted East Asian origin in both in extracted quotes and journal citations but dampened in citations.

Introduction

Science journalism is an indispensable part of scientific communication and provides an accessible way for everyone from researchers to the public to learn about new scientific findings and to consider their implications. However, it is important to identify the ways in which its coverage may skew towards particular demographics. Coverage of science shapes who is considered a scientist and field expert by both peers and the public. This indication of legitimacy can either help recognize people who are typically overlooked due to systemic biases or intensify biases. Journalistic biases in general-interest, online and printed news have been observed by journalists themselves [1,2,3,4], as well as by independent researchers [5,6,7,8,9,10,11]. Researchers found a gap between men and women subjects or sources, with independent studies finding that between 17-40% of total subjects were women across multiple general-interest printed news outlets between 1985 and 2015 [5,6,10]. One study found 27-35% of total subjects in international science and health related news were women between 1995 and 2015, and 46% in print, radio, and television in the United States in 2015 [10]. While gender disparities in news coverage have been extensively researched, our research is different because it focuses on science journalism and comparing it against the demographics of actively publishing scientists. Additionally, our work focuses on research into disparities with respect to name origins, a focus which is currently lacking in the literature.

It should be noted that scientific news coverage is confounded by the existing differences in gender and racial demographics within the scientific field [12,13]. However, we are interested in quantifying disparities with respect to observed demographic differences in the scientific field, using academic authorship as an estimate for the existing demographics. This is similar to other studies that have quantified gender or racial disparities in science as observed in citation [14,15] and funding rates [16,17,18,19,20].

In researching a story, a journalist will typically interview multiple sources for their opinion, potentially asking for additional sources, thus allowing individual unconscious biases at any point along the interview chain to skew scientific coverage broadly. In addition, the repeated selection of a small set of field experts or the approach a journalist takes in establishing a new source may intensify existing

biases [3,4,21]. While disparities in representation may go unnoticed in a single article, analyzing a large corpus of articles can identify and quantify these disparities and help guide institutional and individual self-reflection. In the same vein as previous media studies [5,6,7,8,9,10], we sought to quantify differences in representation across predicted of gender and name origin beyond the existing demographic differences in the scientific field. Our study focused solely on science journalism, specifically content published by *Nature*. Since *Nature* also publishes primary research articles, we used these data to determine the demographics of the expected set of possible sources. This is not a perfect comparator since journalists will not cover every research article presented in the journal. However, we assume it is a reasonable baseline and that large deviations are worthy of investigation as reflecting potential journalistic biases. For clarity, throughout this manuscript we will refer to journalistic articles as *news* and academic, primary research articles as *papers*. Furthermore, when we refer to "authors" we mean authors of academic papers, not journalists. In our analysis, we identified quoted and cited people by analyzing the content and citations within all news articles from 2005 to 2020, and compared this demographic to the academic publishing demographic by analyzing first and last authorship statistics across all *Nature* papers during the same time period.

Through our analysis of 22,001 news articles, we were able to identify >88,000 quotes and >15,000 citations with sufficient speaker or author information. We also identified first and last authors of >10,000 *Nature* papers. We then identified possible differences in predicted gender or name origin using the extracted names. The extracted names were used to generate three data-types: quoted, mentioned, and cited people. We used computational methods to predict gender and identified a trend towards quotes from people predicted to be men in news articles when compared to both the general population and authors predicted to be men in papers. Within the period that we examined, the proportion of quotes predicted to be attributed to men in news articles went from initially higher to currently lower than the proportion first and last authors predicted to be men in *Nature* papers. Furthermore, we found that the quote difference was dependent on article type; the "Career Feature" column achieved gender parity in quoted speakers.

We also used computational methods to predict name origins of quoted, mentioned, and cited people. Through our analysis, we found a significant over-representation of names with predicted Celtic/English origin and under-representation of names with a predicted East Asian origin in both quotes and mentions. To our knowledge, our work is the first to identify a substantial under-representation of names with a predicted East Asian origin in scientific journalism.

While we focused on news from *Nature*, our software can be repurposed to analyze other text. We hope that publishers will welcome systems to identify disparities and use them to improve representation in journalism. Furthermore, our approach is limited by the features we were able to extract, which only reflects a portion of the journalistic process. Journalists could additionally track all sources they contact to self-audit. However, auditing is only part of the solution; journalists and source recommenders must also change their source gathering patterns. To help change these patterns, there exist guides [21], databases [23], and affinity groups that can help us all expand our vision of who can be considered a field expert.

Results

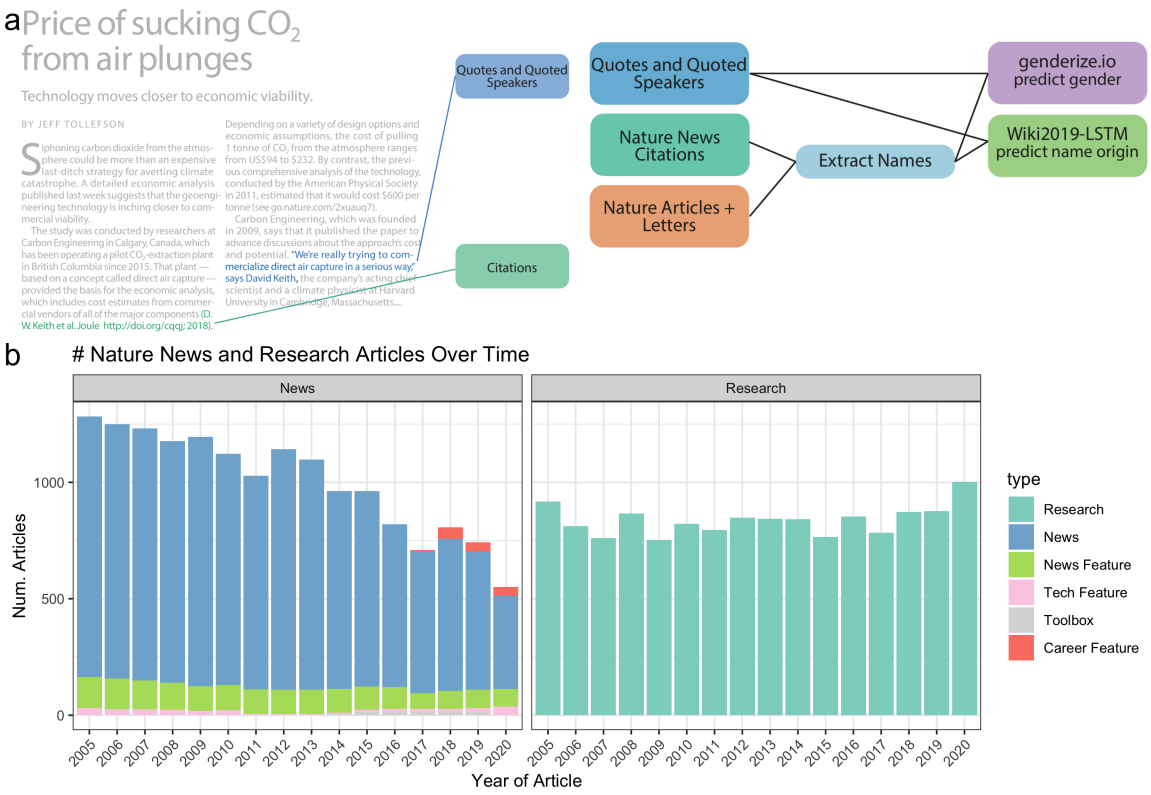

Figure 1: Data and Processing Pipeline Overview Panel **a**, left, depicts an example news article and the type of data extracted from the text. Green and blue highlighted text depicts all quotes, and associated speakers identified by the coreNLP pipeline. A custom script described in section Methods identifies all citations. Panel **a**, right, charts the analyses done on the extracted names and locations from news articles and papers published by *Nature*. Panel **b** shows the types and amounts of articles that we have used for analyses.

Creation of an Annotated News Dataset

We have analyzed the text of 22,001 news-related articles hosted on "www.nature.com" that span 15 years from 2005 to 2020. Our primary focus is on 16,080 articles written by journalists which include the following five article types: "Career Feature", "News", "News Feature", "Technology Feature", and "Toolbox". "Career Feature" generally focuses on the career-related aspects of being a scientist. "News" and "News Feature" focuses on current events related to science as well as new scientific findings. It should be noted that the types of articles contained in "News" changed over time which may induce content shifts in a subset of the articles within our corpus. "Technology Feature" also covers current events and scientific findings, but additionally focuses on how science intersects with technology, such as apps, methodologies, tools, and practices. Lastly, "Toolbox" is similar to "Technology Feature", but is more centered on technology, especially the tools used to perform science. We also include one analysis of the scientist-written news articles, "Career Column" and "News and Views", as an additional set of 5,921 articles. "Career Column" is similar to "Career Feature", except it is not written by journalists, but individuals in the scientific field. "News and Views" is similar to a review article, where a field expert writes an article relating to a recently written article within Nature.

The top three observed article frequencies are "Research" (including "Letters" and "Articles"), "News", and "News Feature". Since *Nature* merged "Letters" and "Research" papers in 2019, we combined them in our analysis. We observed substantial variability in the number of *Nature* news articles by type between 2005 and 2020 (Figure 1b). The changing classification of article types partially explains temporal changes in the frequency of news articles within each category. Over time, the frequency of "News" articles decreased; however, more specific news-related article types increased, including the introduction of the new categories "Career Feature", "Toolbox", and "Career Column".

Extracted Data Used for Analysis

The text and citations were then uniformly processed as depicted in Figure 1a to identify: 1) quotes and quoted speakers (blue box) and 2) cited authors (green box). The extracted names from the text were used to generate three data types for downstream processing: quoted, mentioned, and cited people. Quoted names are any names that were attached to a quote within the article. Mentioned names are any names that were stated within the article. Cited names are all author names of a scientific paper that was cited in the news article. A summary of frequencies for each data type at each point of processing is provided in Tables 1 - 4. We scraped the text using the web-crawling framework Scrapy [24], processed, and ran it through the coreNLP pipeline (Methods). To identify quotes and speakers, we used the coreNLP quote extraction and attribution annotator. We performed multiple name formatting processes (Methods) to identify the speaker's full name for gender and name origin prediction. All names where we could identify two name parts, assumed to be a first and last names exclusive of titles, were used for gender prediction and checked against the genderize.io database. Since names used in the name origin analysis were computationally analyzed and not checked for existence in an existing database, we used additional filters (Methods).
All names excluded from the name origin analysis of quotes and mentions are provided on our github in the file "data/author_data/all_mentioned_fullname_excluded.tsv" and "data/author_data/all_speaker_fullname_excluded.tsv". We found that most names were excluded because two name parts, assumed to be a first and last names exclusive of titles, were not found. We scraped the citations using an independent scraper to the text scraper, but still utilizing the Scrapy framework. All identified DOI's were queried using the *Springer Nature* API to attain all authors' names, positions, and affiliations.

Comparator Datasets

Next, we determined if the quoted speakers and cited authors in news articles have a similar demographic makeup as the scientists who publish their primary research in *Nature*. To make this determination, we used all authors' names, positions, and affiliations of papers published by *Nature* over the same time period (Figure 1a, dark orange box). The author metadata of *Nature* papers from 2005 to 2020 totaled 13,414. To more broadly represent overall science authorship, we also separately analyzed 38,400 *Springer Nature*-published papers from English-language journals over the same time. It should be noted that extracted quotes may come from multiple types of people, such as academic scientists, clinicians, the broader scientific community, politicians, and more. However, through anecdotal observation we believe that most sources come from either academic scientists or those actively involved in science. Similarly, author names were uniformly processed and then used to predict both gender and name origin.

Quoted Speakers and Primary Research Authors in *Nature* are More Often Men

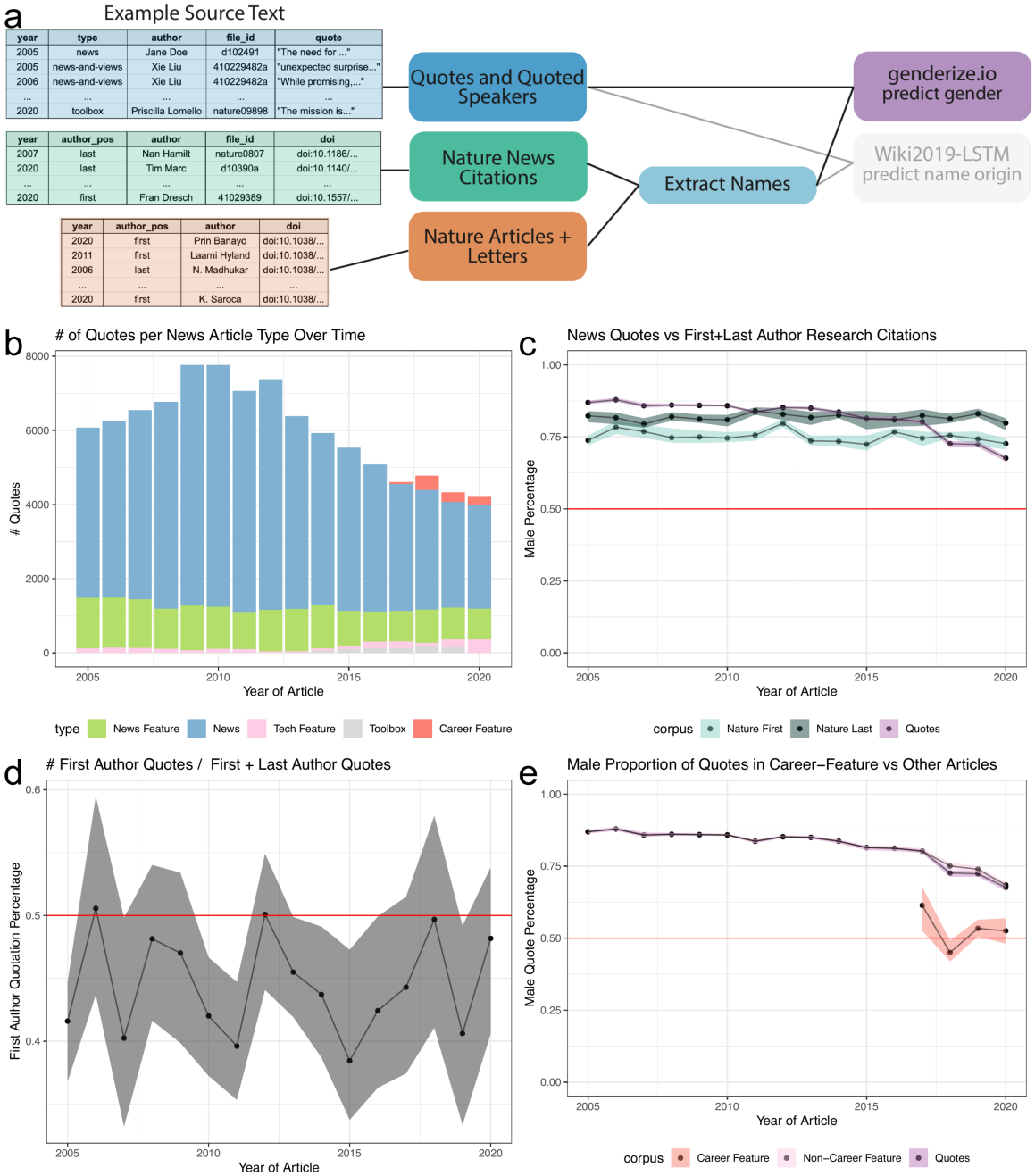

Figure 2: Speakers predicted to be men are sometimes overrepresented in quotes, but this depends on the year and article type. Panel **a**, left, depicts an example of the names extracted from quoted speakers in news articles and authors in papers. Panel **a**, right, highlighted the data types and processes used to analyze the predicted gender of extracted names. Panel **b** shows an overview of the number of quotes extracted for each article type. Panel **c** depicts three trend lines: Purple: Proportion of quotes for a speaker estimated to be a man; Light Blue: Proportion of first author papers estimated to be a man; Dark Blue: Proportion of last authors predicted to be a man. We observe that the proportion of quotes estimated to come from a man is steadily decreasing, most notably from 2017 onward. This decreasing trend is not due to a change in quotes from the first or last authors, as observed in Panel **d**. Panel **d** shows a consistent but slight bias towards quoting the last author of a cited article than the first author over time. Panel **e** depicts the frequency of quote by article type highlighting an increase in quotes from "Career Feature" articles. Panel **e** depicts that the quotes obtained in this article type have reached parity. The colored bands represent a 5th and 95th bootstrap quantiles in all plots, and the point is the mean calculated from 1,000 bootstrap samples.

To quantify and compare the gender demographic of quoted people and authors, we analyzed their names. While we could have analyzed the proportion of unique predicted men speakers, we were interested in measuring the overall participation rates by gender and analyzed the proportion of total quotes, e.g. a single speaker may have more than one quote in an article. Furthermore, we assume that a majority of quoted speakers are typically involved in scientific research and therefore primary research authors is a comparable demographic. Figure 2 shows an overview of the process and example input data for this analysis: 1) quotes and quoted speakers (blue box), 2) first and last

authors' names of papers published by *Nature* (dark orange box). These analyses relied upon accurate gender prediction of both authors and speakers. To predict the gender of the speaker or author, we used the package genderizeR [25], an R package wrapper to access the genderize.io API [26] to get binary gender predictions for each identified first name. We unfortunately cannot identify non-binary gender expression with the tools we used. Performance of binary prediction was evaluated on a benchmark data set of thirty randomly selected news articles, ten from each of the following years: 2005, 2010, 2015 (Figure 1 – figure supplemental 1). In addition, genderize.io has been found by independent researchers to have an error rate comparable to other published gender prediction methods, with a error-rate on predicted names below 6% [27,28]. However, it should be noted that the error rate varies by name origin with the largest decrease in performance on names with an Asian origin [27,28]. In our analysis, we did not observe a large difference in names predicted to come from a man or woman between predicted East Asian and other name origins (Table5).

We first examined the number of quotes identified within each type of science-news article (Figure 2b), totaling 105,457 quotes with 96,390 of them containing a gender prediction for the speaker. Quote frequencies vary by article type. We compared the number of quotes from predicted men to the number of predicted first and last author men published in *Nature*. The total number of authors with a gender prediction was 10,601 first authors and 10,572 last authors across a total of 11,161 publications. As denoted by the red line, we found that the predicted genders of authors and source quotes were far from gender parity (Figure 2c). We found this result consistent for articles written by either a predicted man or woman journalist (Figure 2 - figure supplemental 1a,b). Additionally, we observed a difference in the predicted genders between first and last authors, with the last authors more frequently predicted to be a man. In our supplemental analyses, we provide an additional comparator, a selection of articles from English language journals published by *Springer Nature* (Figure 2 - figure supplemental 2a). The predicted gender gap between first and last authors was larger in our selection of *Springer Nature* papers; however, both first and last authors were predicted to be closer to parity than for *Nature* authors. Overall, predicted men were more frequently quoted than predicted women in *Nature* news articles and first and last authors in *Nature* and *Springer Nature* papers over the same time period.

Career Feature Articles Reach Gender Parity

The gender proportions of authorship were relatively stable over time for both *Nature* and *Springer Nature* papers. In contrast, we found that the rate of quotes predicted to be from men noticeably decreased over time. In 2005, the fraction of quotes predicted to be from men was 87.09% (5,291/6,075), whereas in 2020 it was 68.86% (2,870/4,168). We identified that a large decrease occurred in *Nature* between 2017 and 2018. We explored the possible reasons for this decrease. First, we looked at the authorship position of speakers who were quoted about their published paper (Figure 2d). We identified 6,545 quotes with an associated citation (2,871 first author and 3,674 last author quotes). We found that quotes are slightly biased towards last authors from 2005 to 2020, but because the fraction of last authors predicted to be men remained stable over time both for *Nature* and the selection of *Springer Nature* papers, which likely does not explain the downward trend. We then analyzed the breakdown of gender in quotes by article type. Interestingly, one article type, "Career Feature", achieved gender parity in its quotes (Figure 2e and Figure 2 - figure supplemental 2b). In this article type, we identified a total of 898 quotes (449 predicted women's and 449 predicted men's quotes), which only slightly pulled the overall quote gender ratio closer to parity from 2018 onward. In general, we found that each article type independently trended towards gender parity.

Predicted Celtic English Name Origins are over-represented in quoted and mentioned people, while predicted East Asian name origins are under-represented

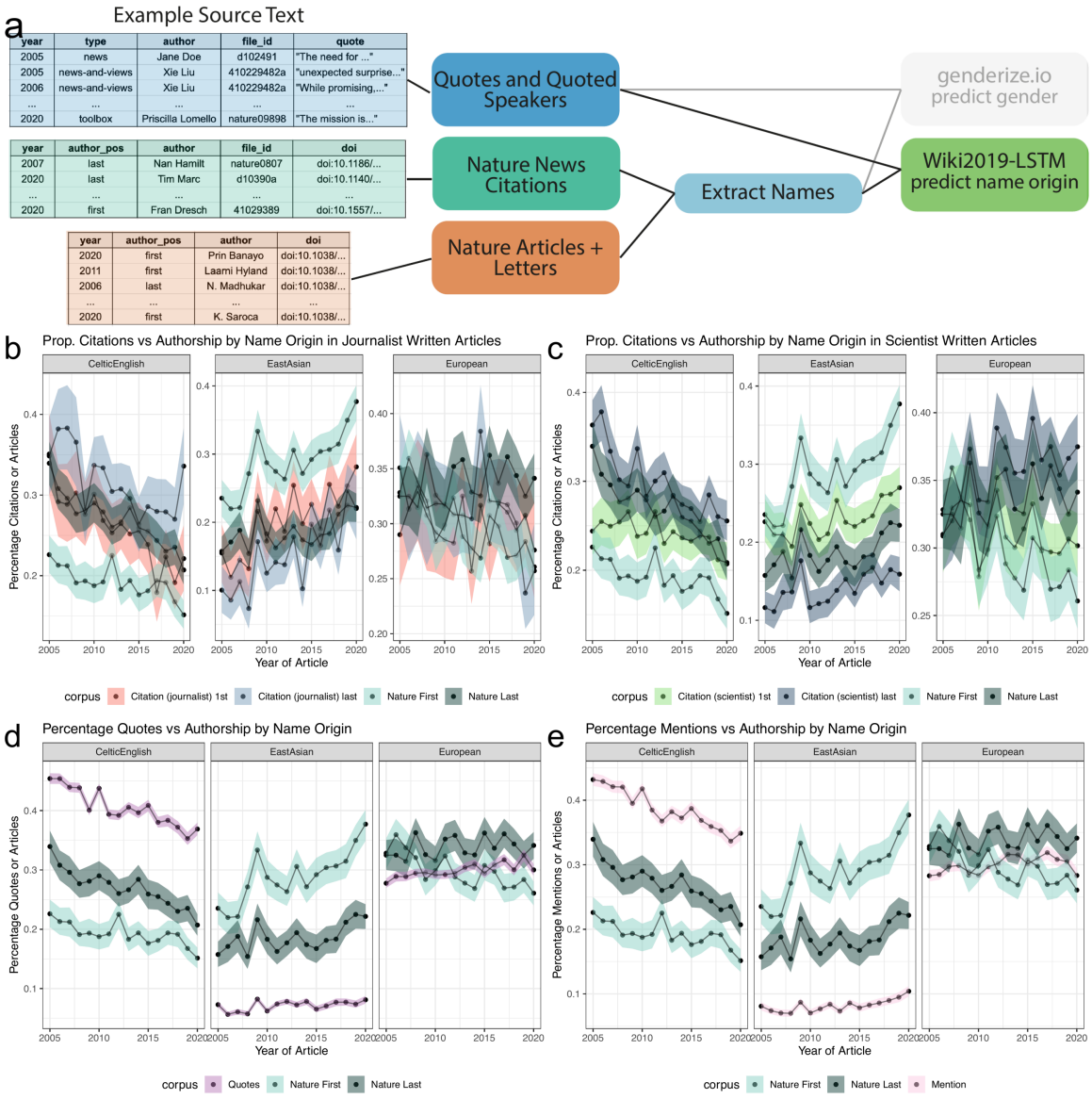

Figure 3: Analysis of Quotes and Citations found Over-representation of Celtic/English and under-representation of East Asian predicted name origins. Panel **a**, left, depicts an example of the names extracted from quoted speakers and citations found within news articles and authors in papers. Panel **a**, right, highlights the data types and processes used to analyze the predicted origin of extracted names. Panels **b** and **c** depict a comparison between the predicted name origins of last authors in *Nature* and cited papers in the news. Panel **b** and **c** differ in the news article types. Panel **b** calculates the predicted name origin proportion using only journalist-written articles, whereas Panel **c** only uses scientist-written articles. The distinction between scientist- and journalist-written articles are defined by the article appearing in either the "Career Column" or "News and Views" sections, or another section, respectively. Similarly, Panels **d** and **e** depict two possible trend lines, comparing predicted name origins of either quoted or mentioned people against name origins of last authors of *Nature* research papers. For more precise numerical comparisons, the mean yearly fold-change for each comparison is provided in Table 6. .

To identify possible disparities with respect to name origin, we again used the extracted names of quoted speakers from *Nature* news articles and last authors of published papers in *Nature*. In addition, we also identified the last authors of all papers cited by a *Nature* news article. All processed names were then input into Wiki2019-LSTM and assigned one of ten possible name origins (Methods). In our analysis, we use name origin to estimate the perceived ethnicity of a primary source by a journalist or fellow scientists who might recommend the individual as a source. Our prediction is not intended to assign ethnicity to an individual, but to be used broadly as a tool to quantify representational differences in a journalist's sociologically constructed perception of a primary source's ethnicity. Figure 3a shows an overview of the process and example input data for this analysis: 1) quotes and quoted speakers (blue box), 2) names of cited first and last authors in news articles (green) 3) first and last authors' names of papers published by *Nature* (dark orange box). We divided our analysis into three parts: firstly, quantifying the proportions of predicted name origins of

first and last authors cited in *Nature* news articles. Secondly, calculating the proportion of quotes from speakers with a predicted name origin. Thirdly, calculating the proportion of unique names mentioned within an article with a predicted name origin. As a comparator set, we again used the first and last author names in *Nature* papers for all three analyses. Additionally, in our supplemental analyses, we compared against the first and last authorship in a selection of *Springer Nature* papers. We found that the number of quotes and unique names mentioned dramatically outnumbered the number of cited authors in *Nature* news articles, as well as first and last authors within *Nature* papers (Figure 3 – figure supplemental 1a). Still, since we have more than one hundred observations per time point for each data type, we believe this is sufficient for our analysis. Minimum and median per data type over all years: *Nature* papers, (568, 684); *Springer Nature* papers, (1332, 1710); *Nature* quotes, (3788, 5696); *Nature* mentions, (3225, 4752); citations in journalist-written *Nature* article, (142, 268) citations in a scientist-written *Nature* article, (512, 664).

News citation rates across name origin predictions nearly match *Nature* authorship

In comparing the citation rate of first and last author name origins in news articles, we decided to additionally analyze scientist-written articles. Though fewer in number, scientist-written news articles have many citations, making the set sufficient for analysis and providing an opportunity to measure differences in citation patterns between journalists and scientists. In both journalist- and scientist-written articles, we found that most cited name origins were predicted Celtic/English or European, both with a bootstrapped estimated citation rate between 17.9-39.6% (Figure 3 – figure supplemental 1b,c). East Asian predicted name origins are the third highest proportion of cited names, with a bootstrapped estimated citation rate between 7.3-28.1%. All other predicted name origins individually account for less than 8.1% of total cited authors.

We analyzed how these distributions compare to the composition of the first and last authors in *Nature* (Figure 3 – figure supplemental 2), by examining the top three most frequent predicted name origins (Figure 3b,c, Table 6). When considering only first authors, we found a slight over-representation for predicted Celtic/English name origins and a small under-representation for predicted East Asian name origins in scientist-written and journalist-written news articles when compared to the composition of first authors in *Nature* (Figure 3b, c). When considering last authors, this pattern no longer exists. Furthermore, we found no substantial difference for European or other predicted name origins when comparing against first and last authorship within *Nature* (Figure 3 – figure supplemental 3a, Table 6). We also observed the predicted Celtic/English over-representation and East Asian under-representation when considering our subset of *Springer Nature* papers (Figure 3 – figure supplemental 3b, Table 7) for both journalist- and scientist-written news articles. In contrast to *Nature*, in the *Springer Nature* set, we see a difference in predicted European name origins, with a growing over-representation. Additionally, we see a difference in predicted Arabic/Turkish/Persian name origins frequencies between cited authors and *Springer Nature* authors, however the absolute difference is lower than observed for Celtic/English and East Asian predicted name origins.

News quotation rates are over-represented for predicted Celtic-English and under-represented for East Asian name origins.

We then sought to determine whether or not the quoted speaker demographic replicated the cited authors' over- and under-representation patterns. We found a much stronger Celtic/English over-representation in comparison to citation patterns, with quotes from those with Celtic/English name origins at a much higher frequency than quotes from those with European name origins (Figure 3 – figure supplemental 1d, Table 6). We also found a much stronger reduction of quotes from people with predicted East Asian name origins (Figure 3 – figure supplemental 1b), with never more than 8.2% of quotes (Figure 3d, Table 6). This reveals a large disparity when considering that people with a

predicted East Asian name origin constitute between 7.3-24.6% of last authors cited in either journalist- or scientist-written news articles (Figure 3b,c, Table 6). When we compare against first and last authorship in *Nature* across all predicted name origins, we find that for all other name origins except for East Asian and Celtic/English, the quote rates closely matches the predicted name origin rate of first and last authors in *Nature* (Figure 3 – figure supplemental 3c, dark grey and light blue lines compare to the purple lines).

To further understand the source of Celtic/English over-representation and East Asian under-representation, we selected a subset of quotes from people whose works were also cited in the news article. The purpose of this additional comparison of quoted speakers versus quoted *and* cited speakers was to reveal source gathering patterns beyond cited works. We found that the proportion of predicted East Asian name origins was closer to the expected rate after considering only quoted speakers with citations, more closely matching the analysis on citations alone (Figure 3 - figure supplemental 4a,b). This indicates that expert opinions gathered beyond manuscript authors is responsible for a large proportion of the observed name disparities.

Next, we sought to determine if predicted journalist name origin had any effect on quote disparities. We found that journalists with a predicted East Asian name origin had a higher rate of East Asian quoted speakers (24.3%) in comparison to journalists with Celtic/English (3.8%) or European (8.6%) predicted name origins (Table 8). To examine if this was again driven by source gathering beyond manuscript authors we again subsetted the quotes by adding two constraints: 1) the quotes must be from a cited first or last author in the same news article (Table 9) and 2) that the cited article must have a US affiliation (Table 10). We found that differences between journalists with different predicted name origins was nearly eliminated when restricting to quoted and cited speakers, and with the additional restriction of US affiliated citations, as evidenced in the predicted East Asian column of Table 10. The differences between Table 8 and Tables 9 and 10, indicate that the predicted name origin of a journalist has some association with sources gathered outside of directly cited works.

When comparing *Nature* articles against the *Springer Nature* set of first or last authors, we again find the same patterns in quoted speakers with East Asian, Celtic/English, and Arabic/Turkish/Persian predicted name origins as we did in the previous citation analysis (Figure 3 – figure supplemental 3d, green and purple lines). In addition, we find an under-representation of predicted Hispanic, South Asian, and Hebrew name origins when comparing against the predicted name origin rate of first and last authors in our *Springer Nature* set.

News mention rates are over-represented for predicted Celtic-English and under-represented for East Asian name origins.

Since many journalists use additional sources that are not directly quoted, we also analyzed likely paraphrased speakers, e.g. a case in which the person was a source and mentioned in the story but not directly quoted. To do this, we identified all unique names that appeared in an article, which we term *mentions*. We found the same pattern of over-representation for predicted Celtic/English name origins and under-representation for East Asian name origins when comparing against both *Nature* and *Springer Nature* first and last authorships (Figure 3e, Figure 3 – figure supplemental 1d,e, Figure 3 – figure supplemental 3e,f, Table 6, Table 7). Similar to the quote analysis, we selected a subset of mentions from people that were also cited in the news article. We again found that the disparity was greatly reduced (Figure 3 - figure supplemental 4c,d).

Discussion

Science journalism is the critical conduit between the academic and public spheres and consequently shapes the public's view of science and scientists. However, as observed in other forms of recognition

in science, biases may shift coverage away from the known demographics within science [30]. Ideally, scientific journalism is representative of academic papers. Though it would be best for news coverage to promote equitable representation, at a minimum, quotes and citations would ideally match the predicted name origin and gender demographics of scientific academia. To examine this last point, we analyzed 22,001 news articles published in *Nature*, to identify quoted, mentioned, and cited people. We then compared this to the authorship statistics from *Nature*'s papers and a subset of *Springer Nature*'s English language papers.

We first looked at possible gender differences in quotes and found a large, but decreasing, gender gap when compared to the general population in all but one article type. Additionally, this result was consistent in articles written by journalists predicted to be women or men. We found that one column, "Career Feature", has an equal number of quotes from both genders, showing that gender parity is possible in science journalism. This finding, coupled with the near equal number of articles written by journalists predicted to be men or women, argues for more diversity in topical coverage. "Career Feature" articles highlight current topics relevant to working scientists and frequently highlight systemic issues with the scientific environment. This column allows space for marginalized people to critique the current state of affairs in science or share their personal stories. This type of content encourages the journalist to seek out a diverse set of primary sources. Including more content that is not primarily focused on recent publications, but all topics surrounding the practice of science, can serve as an additional tool to rapidly achieve gender parity in journalistic recognition.

When considering the relative proportion of authors and speakers predicted to be men, we only find a slight over-representation of men. This overrepresentation is dependent on the authorship position and the year. Before 2010, quotes predicted as from men are overrepresented in comparison to both first and last authors, but between 2010 and 2017 quotes predicted from men are only overrepresented in comparison for first authors. In 2020, we find a slight over-representation of quotes predicted to be from women relative to first and last authors, but still severely under-represented when considering the general population. The choice of comparison between first and last authors can reveal different aspects of the current state of academia. While this does not hold in all scientific fields, first authors are typically early career scientists and last authors are more senior scientists. It has also been shown that early career scientists tend to be more diverse than senior scientists [31,32]. Since we find that quotes are only slightly more likely to come from a last author, it is reasonable to compare the relative rate of predicted quotes from men to either authorship position. Comparison with last authorships may reveal more how gender bias currently exists whereas comparison with early career scientists may reveal bias in comparison to a future, more possibly diverse academic environment. We hope that increased representation and recognition of women in science, even beyond what is observed in authorship, can increase the proportion of women first and last authors such that it better reflects the general population.

To further our analysis of possible coverage disparities, we looked at differences in predicted name origins of quoted and cited authors across all the processed news articles. Our use of name origins is a proxy for a journalist's or referring scholarly peer's potential perceptions of the ethnicity of a primary source as signaled by an individual's name. We do not intend to assign an identity to an individual, but to generate a broad metric to measure possible bias for particular ethnicities during journalists' primary source gathering. Our findings provide additional support for previous studies that identified under-citation [33] and under-recognition [30] of East Asian people. Interestingly, we found under-citation of people with predicted East Asian name origins to be much less pronounced than under-quotation. We do not believe that the under-quotation is driven by paraphrasing sources, which may occur more frequently with non-native English speakers. We also found that the disparity observed in quotes and mentions was almost eliminated when only considering people who were additionally cited within the same article. This suggests that the source of the disparity may lie in the search for additional expert opinions.

Either way, the clear disparity of predicted East Asian researcher quotes and mentions argues for including a broader set of voices when seeking opinions beyond the academic papers being covered in the article. One solution could be to have region-specific journalists. While we were not directly able to examine the regions journalists lived in, this potential strategy is supported by our analysis of journalists with a predicted East Asian name origin. When considering quotes from people with a predicted East Asian name origin, we found that journalists who themselves have a predicted East Asian name origin include a higher proportion of these quotes than journalists with European or Celtic/English predicted names. When considering only people who were both quoted and cited, the effect of the predicted name origins of journalists was substantially dampened. We are unable to identify if this is a geographic bias of the reporters in this analysis, since we do not know the location of the journalist at the time of writing the article. As a proxy for measuring possible geographical bias of a journalist, we attempted to identify if there was any geographical bias of cited authors. To do this, we identified the affiliation of each cited author and identified their affiliated country. Unfortunately, we could not robustly extract a large enough number of cited authors from different countries to make any conclusive statements. Expanding our work to other science journalism outlets could help identify possible ways in which geographic region, genders, and perceived ethnicity interact and affect scientific visibility of specific groups. While we are unable to identify that journalists have a specific geographical bias, having reporters explicitly focused on specific regional sources will broaden coverage of international opinions in science.

In our analysis, we also find that there are more first authors with predicted East Asian name origin than last authors. This is in contrast to predicted Celtic/English and European name origins. Furthermore, we see that the amount of first author people with predicted East Asian name origins is increasing at a much faster rate than quotes are increasing. If this mismatched rate of representation continues, this could lead to an increasingly large erasure of early career scientists with East Asian name origins. As noted before, focusing on increasing engagement with early career scientists can help to reduce the growing disparity of public visibility of scientists with East Asian name origins.

Through our comprehensive analysis, we were able to quantify how recognized persons in news journalism vary by name origin and gender, then compare it to scientific publishing background rates. While we found a significant gender disparity compared to the general population, the rate of women's representation in scientific news is increasing and outpacing first and last authorships on scientific papers. Furthermore, we identified a significant reduction of quotes from scientists with a predicted East Asian name origin when compared to paper authorship and a significant but smaller reduction of cited authors with a predicted East Asian name origin in news content.

Computational tools enabled us to automatically analyze thousands of articles to identify existing disparities by gender and name origin, but these tools are not without limitations. Our tools are unable to identify non-binary people and rely on gender predictors that are known to have region-specific biases, with the largest decrease in performance on names of an Asian origin [27,28]. Furthermore, name origin is only a proxy for externally perceived racial or ethnic origins of a source or author and is not as accurate as self-identified race or ethnicity. Self-identification better captures the lived experience of an individual that computational estimates from a name can not capture. This is highlighted in our inability to distinguish between Black and White people from the US by their names. As the collection of demographic data by publication outlets grows, we believe this will enable a more fine-grained and accurate analysis of disparities in scientific journalism

Previous anecdotal studies from journalists have shown that awareness of their bias can help them to reduce it [2,3,4]. Once a bias is identified an individual can seek resources to help them find and retain diverse sources, such as utilizing international expert databases like gage [22] and SheSource [23]. Additional tips for journalists to achieve and maintain a diverse source pool is described by Christina Selby in the Open Notebook [21].

It should also be mentioned that we were only able to analyze the data provided through scraping "www.nature.com". This is a major limitation, because the only measures that we have of demographics of sources are people who have their name mentioned or research cited within the article. Journalists do not quote or mention all of the sources that they interviewed or cite all of the papers that they read when researching an article. For example, a person may not be mentioned or quoted in the article because of length limitations, because they do not want to be named, or if they provide information that is not directly quotable but that still shapes the content of the article. A more accurate reflection of journalists' sources would be a self-maintained record of people they interview. Our work examines disparities with respect to recognition within articles, which can be measured by mentions, quotes, or citations of people.

Furthermore, the news articles present on "www.nature.com" are intended for a very specific readership that may not be reflective of more broad scientific news outlets. In a separate analysis, we took a cursory look into a comparison with *The Guardian* and found similar disparities in gender and name origin. However, it is not clear which publications should be used as a comparator for science-related articles in *The Guardian*, and difficult to compare relative rates of representation. While other science news outlets may not have a direct comparator, it would be useful to take a broad comparison across multiple science news outlets to compare against one another. Our existing pipeline could be easily applied to other science news outlets and identify if there exists a consistent pattern of disparity regardless of the intended readership.

Another major limitation of our study, is that we only used articles published by *Nature* or *Springer Nature* as a comparator. Not all papers are interesting to the general public and likely to be covered by journalists. In this work, we assume that the demographics of scientists publishing work that is likely to be covered by journalists matches the demographics of all scientists publishing articles in *Nature*, *Springer Nature* or other publishers. Our work could be extended to include additional publishers and pre-print servers. To reveal more scientific-field-specific biases, analyses could be performed on individual topics versus our aggregate analysis.

Furthermore, many journalists are limited by who responds to their requests for an interview or recommendations from prominent scientists. Scientists fielding reporter inquiries can also audit themselves to examine the extent to which there are disparities in the sets of experts they recommend. Journalists and the scientists they interview have a unique opportunity to shape the public and their peers' perspectives on who is a scientific expert. Their choice of coverage topics and interviewees could help to reduce disparities in the outputs of science-related journalism.

Methods

Data Acquisition and Processing

Text Scraping

We scraped all text and metadata from *Nature* using the web-crawling framework Scrapy [24] (version 2.4.1). Scrapy is a tool that applies user-defined rules to follow hyperlinks on webpages and return the information contained on each webpage. We used Scrapy to extract all web pages containing news articles and extract the text. We created four independent scrapy web spiders to process the news text, news citations, journalist names, and paper metadata. News articles were defined as all articles from 2005 to 2020 that were designated as "News", "News Feature", "Career Feature", "Technology Feature", and "Toolbox". Using the spider "target_year_crawl.py", we scraped the title and main text from all news articles. We character normalized the main text by mapping visually identical Unicode codepoints to a single Unicode codepoint and stripping many invalid Unicode characters. Using an additional spider defined in "doi_crawl.py", we scaped all citations within news articles. For simplicity,

we only considered citations with a DOI included in either text or a hyperlink in this spider. Other possible forms of citations, e.g., titles, were not included. The DOIs were then queried using the *Springer Nature* API. The spider "article_author_crawl.py" scraped all articles designated "Article" or "Letters" from 2005 to 2020. We only scraped author names, author positions, and associated affiliations from research articles, which we refer to as *papers*. It should be noted that "News" article designations changed over time and partially explain the changing frequency of news articles within each category. The frequency of "News" articles decreased, but more specific news-related article types increased. Additionally, scraping for journalist names was performed months after the initial scraping of the text, and some aspects of the *Nature* website changed. Website changes caused us to lose unique file mappings between the scraped journalist name and other article metadata for 137 articles. Less than thirty articles per year were impacted.

coreNLP

After the news articles were scraped and processed, the text was processed using the coreNLP pipeline [34] (version 4.2.0). The main purpose for using coreNLP was to identify named entities related to mentioned and quoted speakers. We used the standard set of annotators: tokenize, ssplit, pos, lemma, ner, parse, coref, and additionally the quote annotator. Each of which respectively performs text tokenization, sentence splitting, part of speech recognition, lemmatization, named entity recognition, division of sentences into constituent phrases, co-reference resolution, and identification of quoted entities. We used the "statistical" algorithm to perform coreference resolution for speed. Each of these aspects is required in order to identify the names of quoted or mentioned speakers and identify any of their associated pronouns. All results were output to json format for further downstream processing.

Springer Nature API

Springer Nature was chosen over other publishers and search engines for multiple reasons: 1) ease of use; 2) it is a large publisher, second only to Elsevier; 3) it covers diverse subjects, in contrast to PubMed, which focuses on the biomedical and life sciences literature; 4) its API has a large daily query limit (5000/day); and 5) it provided more author affiliation information than found in Elsevier. We generated a comparative background set for supplemental analysis with the *Springer Nature* API by obtaining author information for papers cited in news articles. We selected a subset of papers to generate the *Springer Nature* background set. These papers were the first 200 English language "Journal" papers returned by the *Springer Nature* API for each month, resulting in 2400 papers per year from 2005 through 2020. These papers are the first 200 papers published each month by a *Springer Nature* journal, which may not be completely random, but we believe to be a reasonably representative sample. Furthermore, the *Springer Nature* articles are only being used as an additional comparator to the complete set of all *Nature* research papers used in our analyses. To obtain the author information for the cited papers, we queried the *Springer Nature* API using the scraped DOI. For both API query types, the author names, positions, and affiliations for each publication were stored and are available in "all_author_country.tsv" and "all_author_fullname.tsv".

Name Formatting

Name Formatting for Gender Prediction in Quotes or Mentions

We first pre-filter articles that have more than 25 quotes, which results in the removal of 2.69% (433/16,080) of the total articles. This was done to ensure no single article is over-represented and to avoid spuriously identified quotes due to unusual article formatting. To identify the gender of a quoted or mentioned person, we first attempt to identify the person's full name. Even though genderizeR, the computational method used to predict the name's gender, only uses the first name to

make the gender prediction, identifying the full name gives us greater confidence that we correctly identified the first name. To identify the full name, we take the predicted speaker by coreNLP. Unfortunately, this is not always the full name and is only either the first or last name, with the full name occuring somewhere else in the article. In order to get the full name for all names that coreNLP is unable to identify, we match the coreNLP-identified partial name to the longest matching name within the same article. We match names by finding the longest mentioned name in the article with minimal edit (Levenshtein) distance. The name with the smallest edit distance, where character deletions have zero cost, is defined as the matching name. Character deletion was assigned a zero cost because we would like exact substring matches. For example, the calculated cost, including a cost for character deletion, between John and John Steinberg is 10; without character deletion, it is 0. Compared with the distance between John and Jane Doe, with character deletion cost, it is 7; without it is 2. If we are still unable to find a full name, or if coreNLP cannot identify a speaker at all, we also determine whether or not coreNLP linked a gendered pronoun to the quote. If so, we predict that the gender of the speaker is the gender of the pronoun. We ignore all quotes with no name or partial names and no associated pronouns. A summary of processed gender predictions of quotes at each point of processing is provided in Table 1.

Name Formatting for Gender Prediction of Authors

Because we separate first and last authors, we only considered papers with more than one author. Roughly 7% of all papers were estimated to be single authors and removed from this analysis: 1113/15013 for cited Springer articles, 2899/42155 for random Springer articles, 955/12459 for Nature research articles. As for quotes, we needed to extract the first name of the authors. We cast names to lowercase and processed them using the R package humaniformat [35]. humaniformat is a rule based program that uses character markers to identify if names are reversed (Lastname, Firstname), find middle names and titles. This processing was not required for quote prediction because names written in news articles did not appear to be reversed or abbreviated. Since many last or first authorships may be non-names, we additionally filtered out any identified names if they partially or fully match any of the following terms: “consortium”, “group”, “initiative”, “team”, “collab”, “committee”, “center”, “program”, “author”, or “institute”. Furthermore, since many papers only contain first name initials (for example, “N. Davidson”), we remove any names less than four letters (length includes punctuation) and containing a “.” or “-”, then strip out all periods from the first name. This ensures that hyphenated names are not changed, e.g. Julia-Louise remains unchanged, but removes hyphenated initials, e.g. J-L. A summary of processed author gender predictions at each point of processing is provided in Tables 2 - 4.

Name Formatting for Name Origin Prediction

In contrast to the gender prediction, we require the entire name in all steps of name origin prediction. For names identified in the *Nature* news articles, we use the same process as described for the gender prediction; we again try to identify the full name. For author names, we process the names as previously described for the gender prediction of authors. For all names, we only consider them in our analyses if they consist of two distinct parts separated by a space and excluding titles (e.g. Mrs., Prof., Dr., etc.). All names that were filtered out in the analysis of quotes and mentions are provided on our github in the file “data/author_data/all_mentioned_fullname_excluded.tsv” and “data/author_data/all_speaker_fullname_excluded.tsv”. A summary of processed name origin predictions of quotes and citations at each point of processing is provided in Tables 1 - 4.

Gender Analysis

The quote extraction and attribution annotator from the coreNLP pipeline was employed to identify quotes and their associated speakers in the article text. In some cases, coreNLP could not identify an

associated speaker's name but instead assigned a gendered pronoun. In these instances, we used the gender of the pronoun for the analysis. The R package genderizeR [25], a wrapper for the genderize.io API [26], predicted the gender of authors and speakers. We predicted a name as indicating a man if the first name was predicted by genderizeR to come from a man with at least a probability of 50%. To reduce the number of queries made to genderize.io, a previously cached gender prediction from [29] was also used and can be found in the file "genderize.tsv". All first name predictions from this analysis are in the file "genderize_update.tsv". To estimate the gender gap for the quote gender analyses, we used the proportion of total quotes, not quoted speakers. We used the proportion of quotes to measure speaker participation instead of only the diversity of speakers. The specific formulas for a single year are shown in equations 1 and 2. We did not consider any names where no prediction could be made or quotes where neither speaker nor gendered pronoun was associated.

$$\text{Prop. Quotes from Men} = \frac{|\text{Speaker Quotes from Men}|}{|\text{Speaker Quotes from Men or Women}|} \tag{1}$$

$$\text{Prop. First Author Men} = \frac{|\text{First Authors Men}|}{|\text{First Author Men or Women}|} \tag{2}$$

Name Origin Analysis

We used the same quoted speakers as described in the previous section for the name origin analysis. In addition, we also consider all authors cited in a *Nature* news article. In contrast to the gender prediction, we need to use the full name to predict name origin. We submitted all extracted full names to Wiki-2019LSTM [29] to predict one of ten possible name origins: African, Celtic/English, East Asian, European, Greek, Hispanic, Hebrew, Arabic/Turkish/Persian, Nordic, and South Asian. While a full description of Wiki-2019LSTM is outside the scope of this paper, we describe it here breifly. Wiki-2019LSTM is trained on name and nationality pairs, using 3-mers of the characters in a name to predict a nationality. To ensure robust predictions, nationalities were grouped together as described in NamePrism [36]. NamePrism chose to exclude the United States, Australia, and Canada from their country groupings and were therefore excluded during training of Wiki-2019LSTM. This choice was justified by NamePrism in stating that these countries had a high level of immigration. The treemap of country groupings defined in the NamePrism manuscript are found in figure 5 of the publication [36].

After running the pre-trained Wiki-2019LSTM model, we used the probability origin for each name instead of a hard assignment to a single class. Hard assignment was not used because it has been shown to reproduce biases due to the underreporting of Black and overprediction of White individuals [37]. Similar to the gender analyses, quote proportions were again directly compared against publication rates, except using the probability of assignment instead of the count of hard assignments. For citations, quotes, and mentions, we calculated the proportion for a given year for each name origin. This is shown in Eq. 3 to, for example, calculate the citation rate for last authors with a Greek name origin for a single year.

$$\begin{aligned}&\text{Prop. Greek Last Author Cited}\\ &= \frac{\Sigma(\text{Probability Greek Name for each Cited Last Author})}{|\text{Cited Last Authors w/any Name}|}\end{aligned} \tag{3}$$

$$\text{Prop. Greek Quotes} = \frac{\Sigma(\text{Probability Greek Name for each Quoted Speaker})}{|\text{Quotes w/any Named Speaker}|} \quad (4)$$

$$\text{Prop. Greek Names Mentioned} = \frac{\Sigma(\text{Probability Greek Mentioned Name})}{|\text{Unique Names w/any Origin Mentioned}|} \quad (5)$$

Bootstrap Estimations

We used the boostrap method to construct confidence intervals for each of our calculated statistics. For all analyses related to equations 1 - 5, we independently selected 1,000 bootstrap samples for each year. We sampled with replacement of size equal to the cardinality of the complete set of interest. Bootstrap estimates for equations 1 - 5 were performed by sampling the denominator set. The mean, 5th, 95th quantiles across the estimates are reported as the estimated mean, lower, and upper bounds.

Data availability

Due to copyright, we are unable to provide the unprocessed scraped data used in this analysis. To ensure reproducability without violating copyright, we provide the word frequencies for each news article, the coreNLP output, all analyzed names with an article identifier, as well as any other associated data used in the analyses such as quotes and citations. We provide all of this data on github. We also provide data descriptions in our github README, under the header “Quick data folder overview”.

Code availability

This manuscript was written using Manubot [38] and is available on github: manuscript repository link. All code and metadata is also available on github, full analysis repository link, under a BSD 3-Clause License. The code to generate all main and supplemental figures are available as R markdown documents within our main analysis github, in the following subfolder: notebooks. We provide a docker image that can re-run the analysis pipeline using intermediate, pre-processed data and produce all the main and supplemental figures. To re-run the entire pipeline (including scraping), the docker image contains all necessary packages and code. Scraping code is available on our main analysis github, in the following subfolder: scraper. The shell scripts to re-run the entire analysis are provided in the README file in the github repository.

Acknowledgements

We would like to thank Jeffrey Perkel for asking thoughtful questions that spurred this line of research, and providing feedback and insight into the news-gathering process during the course of this project.

Supplemental Figures

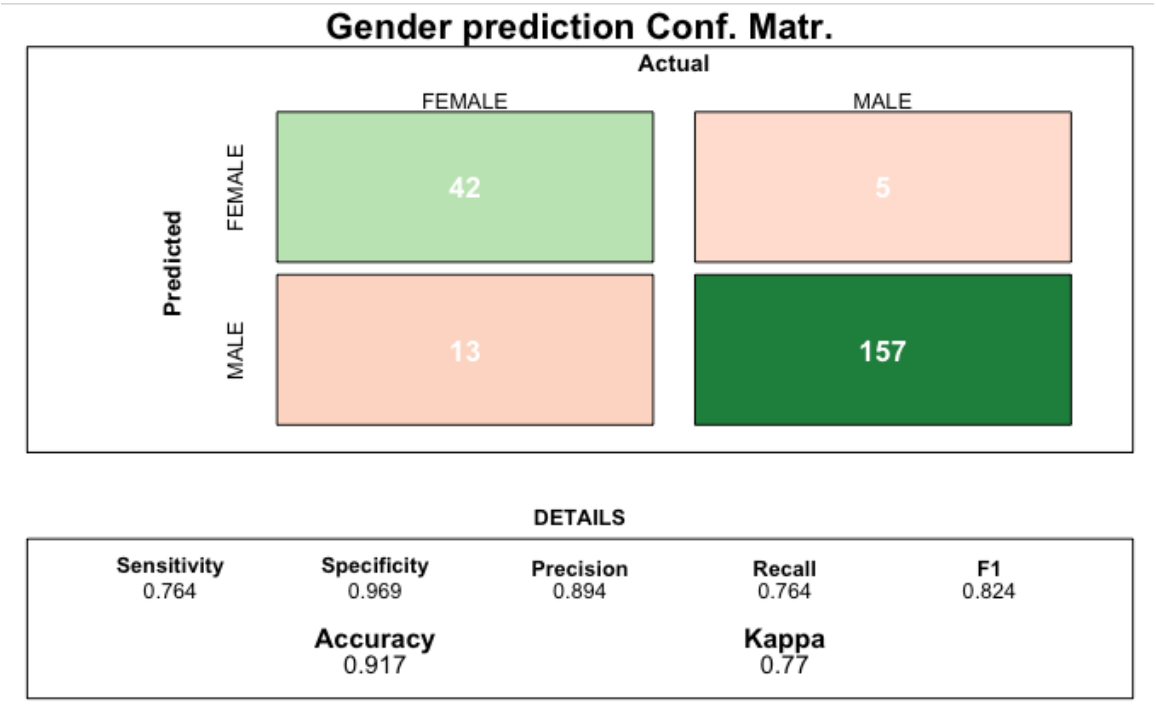

Figure 1 – figure supplemental 1: Benchmark Data The performance of gender prediction for pipeline-identified quoted speakers.

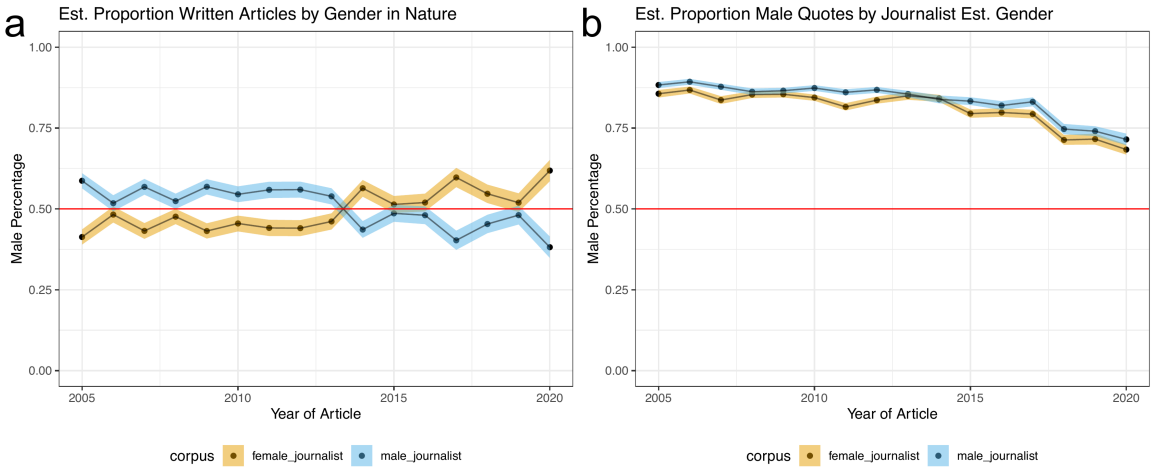

Figure 2 - figure supplemental 1: Speakers predicted to be men are overrepresented in news quotes regardless of predicted journalist gender Panel a depicts two trend lines: Yellow: Proportion of *Nature* news articles written by a predicted women journalist; Blue: Proportion of *Nature* news articles written by a predicted men journalist. We observe a moderate gender difference in the number of articles written by men and women journalists. Panel b depicts two trend lines: Yellow: Proportion of quotes predicted to be from men in an article written by a journalist predicted to be a woman; Blue: Proportion of quotes predicted to be from men in an article written by a journalist predicted to be a man. In all plots, the colored bands represent the 5th and 95th bootstrap quantiles and the point is the mean calculated from 1,000 bootstrap samples.

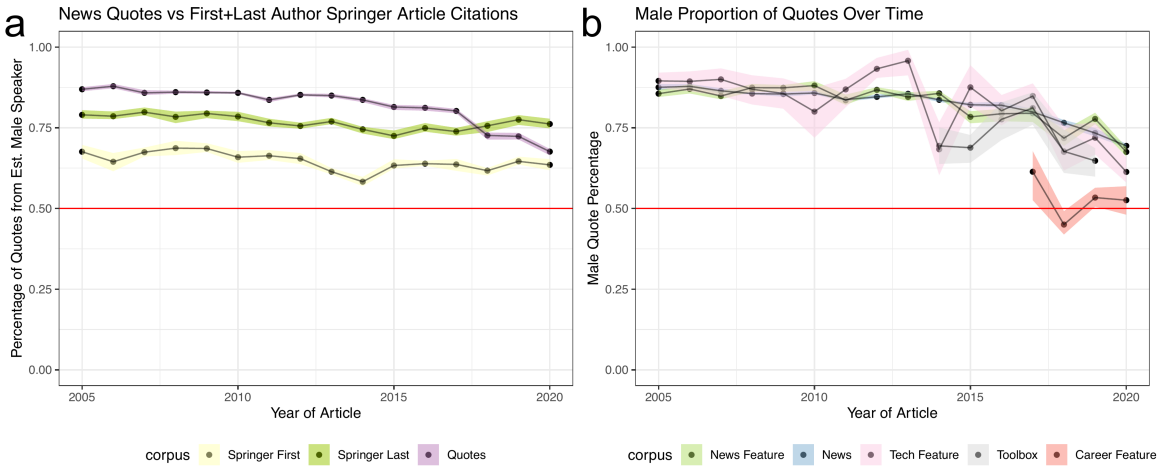

Figure 2 - figure supplemental 2: Speakers predicted to be men are overrepresented in news quotes when compared against *Springer Nature* authorship Panel a depicts three trend lines: Purple: Proportion of *Nature* quotes for a speaker estimated to be a man; Light Grey: Proportion of *The Guardian* quotes for a speaker estimated to be a man; Yellow: Proportion of first author articles from an author estimated to be a man in *Springer Nature*; Dark Mustard: Proportion of last author articles from an author estimated to be a man in *Springer Nature*. We observe a larger gender difference between first and last authors in *Springer Nature* articles, however the proportion of speakers estimated to be men is less than observed in *Nature* research articles. Panel b depicts the proportion of quotes from predicted men broken down by article type. In all plots the colored bands represent the 5th and 95th bootstrap quantiles and the point is the mean calculated from 1,000 bootstrap samples.

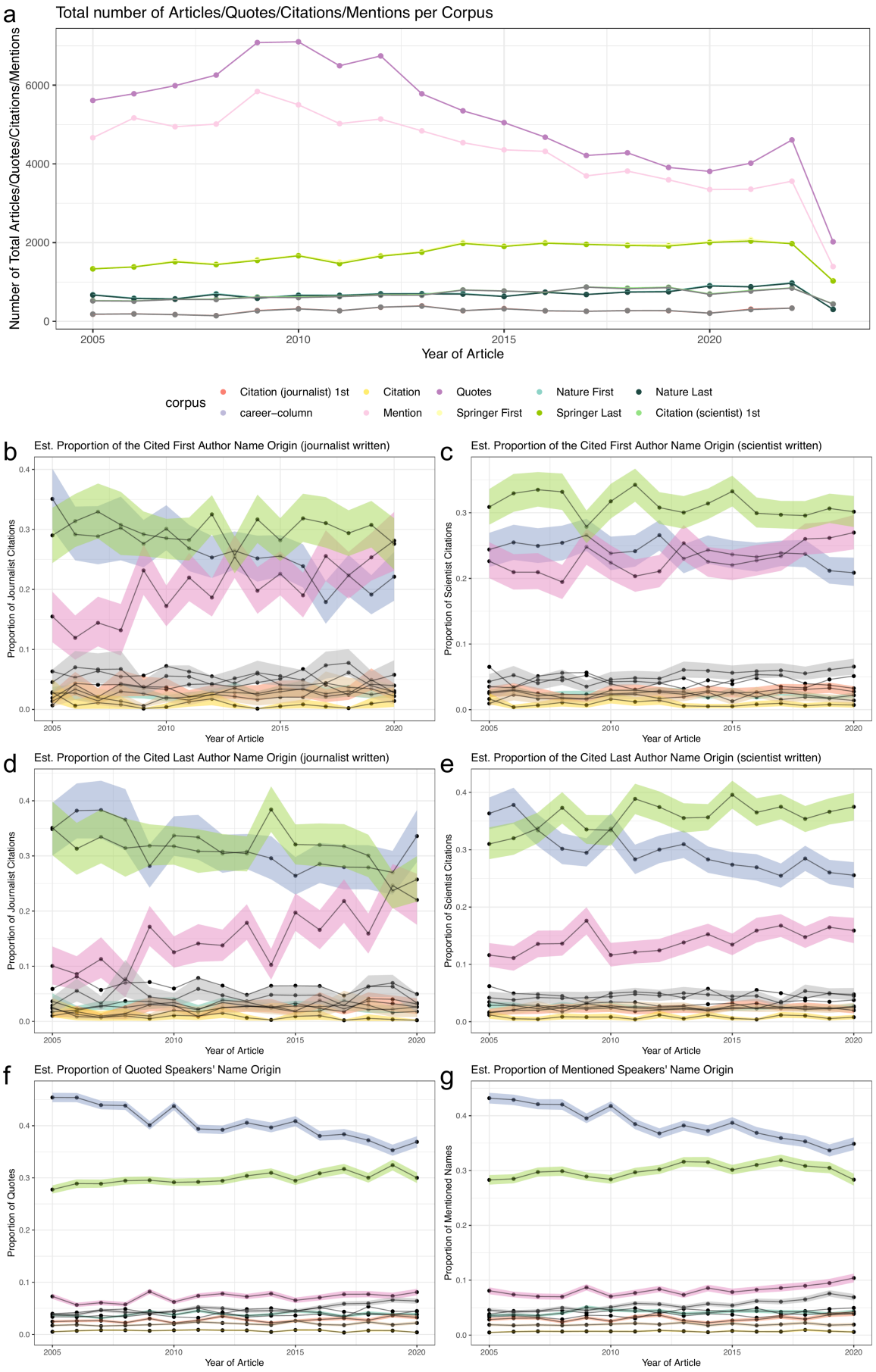

Figure 3 – figure supplemental 1: Predicted Celtic/English, and European name origins are the highest cited, quoted, and mentioned Panel a, depicts the number of quotes, mentions, citations, or research articles considered in the name origin analysis. Panels b-g depicts the proportion of a name origin in a given dataset, citations in articles written by journalists or writers, quoted speakers or mentions. In all plots the colored bands represent the 5th and 95th bootstrap quantiles and the point is the mean calculated from 1,000 bootstrap samples.

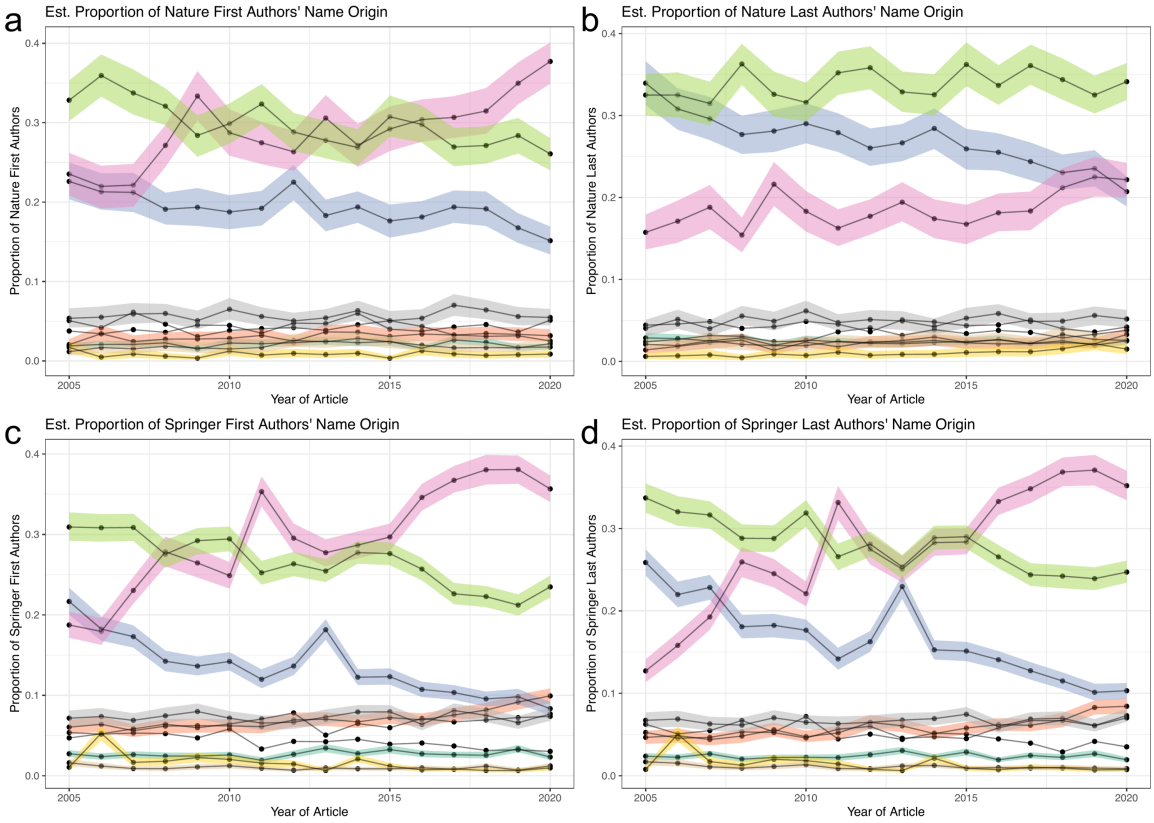

Figure 3 – figure supplemental 2: Distribution of name origins *Nature* and *Springer Nature* articles Panels a-d depicts the predicted name origins of first and last authors in our background sets. Panel a and b show the predicted name origins of *Nature* first and last authors, respectively. Panel c and d show the predicted name origins of *Springer Nature* first and last authors, respectively.

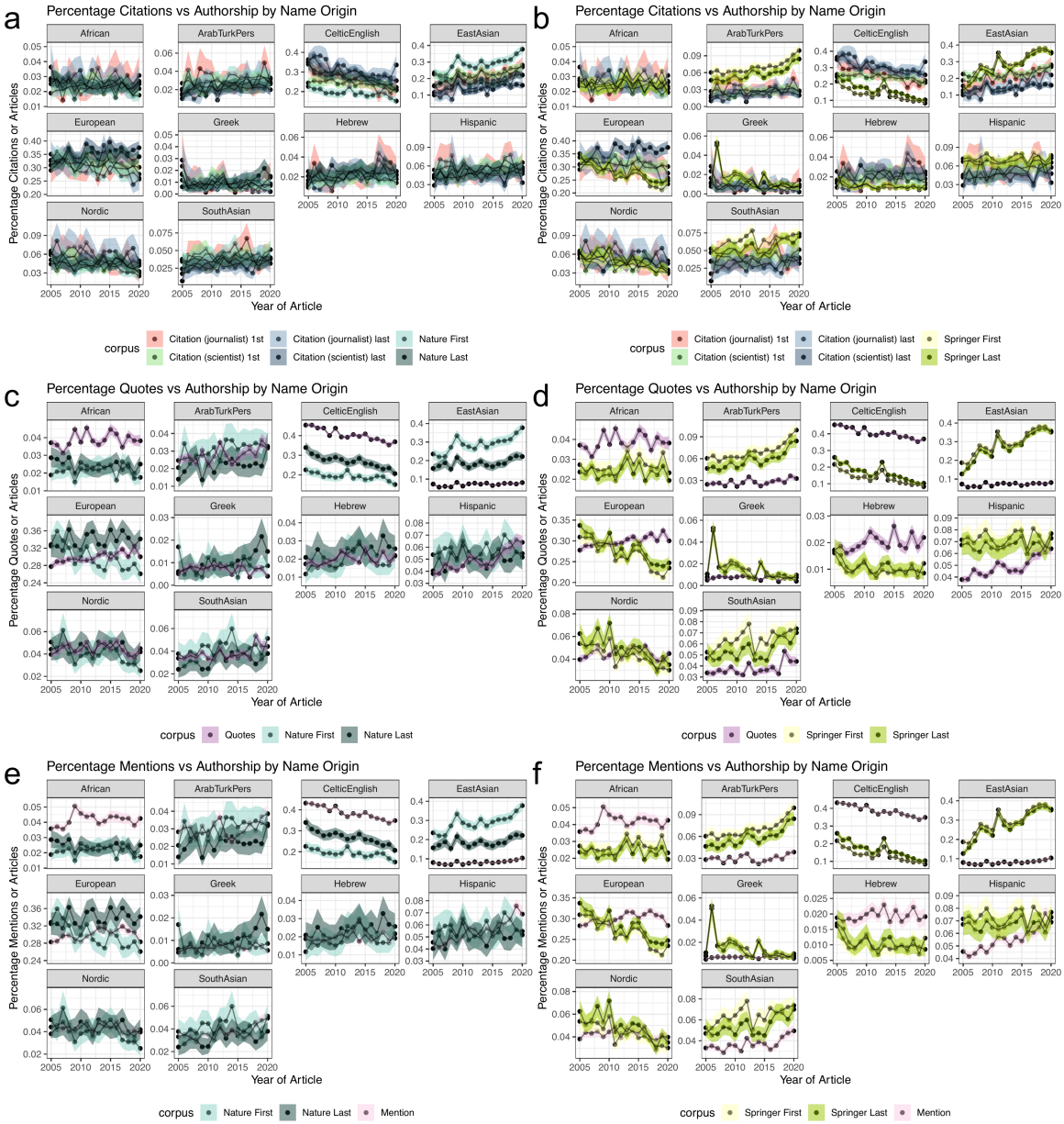

Figure 3 – figure supplemental 3: Over-representation of predicted Celtic/English and under-representation of East Asian name origins is also found in comparison to *Nature* and *Springer Nature* articles Panels a-f depicts ten plots, each for a possible name origin comparison against a background set. Panel a, c, and e compare the citation (a), quote (c), or mention (e) rate against *Nature* first and last author name origins. Panel b, d, and f compare the citation (a), quote (c), or mention (e) rate against *Springer Nature* first and last author name origins. Panels a and b additionally partition the citation rates by journalist-written articles and scientist-written articles, each further divided into first or last author position. For c-f, only journalist written articles are considered.

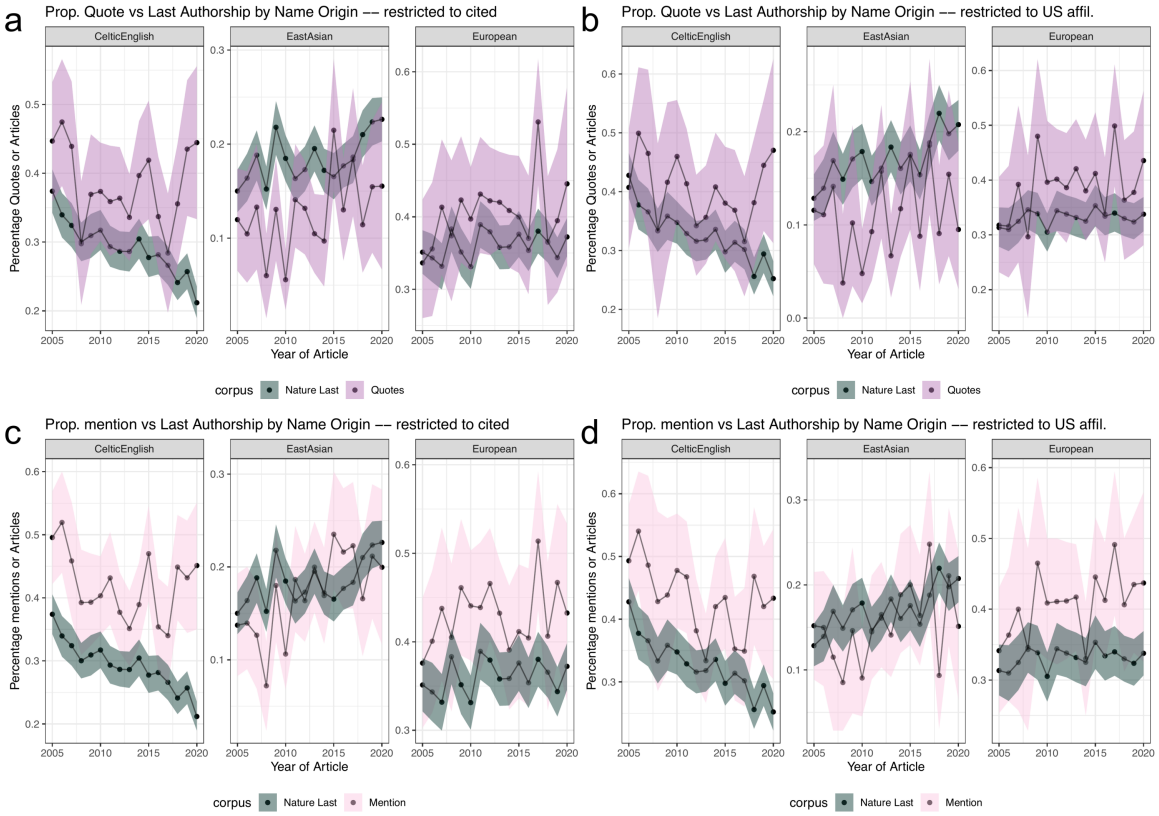

Figure 3 - figure supplemental 4: Over-representation of predicted Celtic/English and under-representation of East Asian quotes and mentions are reduced when additionally considering citation Panels a-d depicts twelve plots, each for a possible name origin comparison against a background set. Panels a and b compare name origin proportions of quotes from people that were also cited in the same article. Panels c and d compare name origin proportions from mentions of people that were also cited in the same article. In all plots the colored bands represent the 5th and 95th bootstrap quantiles and the point is the mean calculated from 1,000 bootstrap samples.

Table 1: Breakdown of quotes at major processing steps

Processing Step	Frequency
Total Quotes	105457
Quotes with a full name or pronoun associated	96620
Quotes with a gender prediction	96390
Quote with a full name	88535
Quotes with a name origin prediction	100457

Table 2: Breakdown of citations at major processing steps

Writer of Article	Total citations	Total Springer Nature citations	First author citations with a full name	Last author citations with a full name	First author citations with a name origin prediciton	Last author citations with a name origin prediciton
Journalist	15713	5736	4452	4464	4449	4447
Scientist	40707	14597	11276	11170	11276	11152

Table 3: Breakdown of all Springer Nature papers at major processing steps

Processing Step	Frequency
# Springer Nature Articles	38400

Processing Step	Frequency
# First + last authors with a full name in Springer Nature Articles	55370
# First + last authors with a gender prediction in Springer Nature Articles	51686
# First + last authors with a name origin prediction in Springer Nature Articles	55197

Table 4: Breakdown of all Nature papers at major processing steps

Processing Step	Frequency
# Nature Articles	13414
# First + last authors with a full name in Nature Articles	21996
# First + last authors with a gender prediction in Nature Articles	21173
# First + last authors with a name origin prediction in Nature Articles	21996

Table 5: Quoted speaker gender by name origin

	Women	Men	Proportion Men
African	270	1554	0.8519737
ArabTurkPers	346	1765	0.8360966
CelticEnglish	6399	33329	0.8389297
EastAsian	1090	4438	0.8028220
European	4788	22844	0.8267226
Greek	73	445	0.8590734
Hebrew	213	1303	0.8594987
Hispanic	760	2450	0.7632399
Nordic	593	2397	0.8016722
SouthAsian	465	2019	0.8128019

Table 6: Mean fold change comparison with Nature from bootstrap samples with 95% CI

	CelticEnglish	EastAsian	European
citation_journalist_first vs. nature_first	1.36 (0.96, 1.74)	0.7 (0.46, 0.91)	1.01 (0.8, 1.25)
citation_journalist_last vs. nature_last	1.18 (0.93, 1.54)	0.82 (0.42, 1.27)	0.93 (0.71, 1.19)
citation_scientist_first vs. nature_first	1.26 (1.05, 1.5)	0.81 (0.66, 1.02)	1.05 (0.88, 1.22)
citation_scientist_last vs. nature_last	1.11 (0.95, 1.31)	0.77 (0.58, 0.99)	1.06 (0.93, 1.19)
quote vs. nature_first	2.12 (1.77, 2.51)	0.25 (0.2, 0.32)	1.01 (0.81, 1.22)
quote vs. nature_last	1.52 (1.32, 1.75)	0.39 (0.3, 0.49)	0.89 (0.79, 1.01)
mention vs. nature_first	2.03 (1.67, 2.39)	0.29 (0.23, 0.36)	1.02 (0.81, 1.22)
mention vs. nature_last	1.44 (1.26, 1.67)	0.45 (0.35, 0.54)	0.89 (0.79, 1)

Table 7: Mean fold change comparison with Springer Nature from bootstrap samples with 95% CI

	CelticEnglish	EastAsian	European
citation_journalist_first vs. springer_first	1.99 (1.42, 2.64)	0.69 (0.47, 0.96)	1.14 (0.89, 1.47)

	CelticEnglish	EastAsian	European
citation_journalist_last vs. springer_last	2.01 (1.31, 3.08)	0.56 (0.3, 0.82)	1.12 (0.91, 1.37)
citation_scientist_first vs. springer_last	1.54 (0.95, 2.17)	0.91 (0.62, 1.64)	1.13 (0.91, 1.33)
citation_scientist_last vs. nature_last	1.11 (0.95, 1.31)	0.77 (0.58, 0.99)	1.06 (0.93, 1.19)
quote vs. springer_last	2.58 (1.74, 3.6)	0.28 (0.2, 0.54)	1.08 (0.84, 1.35)
quote vs. nature_last	1.52 (1.32, 1.75)	0.39 (0.3, 0.49)	0.89 (0.79, 1.0)
mention vs. springer_last	2.45 (1.65, 3.42)	0.32 (0.23, 0.59)	1.08 (0.85, 1.32)
mention vs. nature_last	1.44 (1.26, 1.67)	0.45 (0.35, 0.54)	0.89 (0.79, 1)

Table 8: Quoted speaker name origin, by journalist name origin

Journalist Name Origin	African	Arab Turk Pers	Celtic English	East Asian	European	Greek	Hebrew	Hispanic	Nordic	South Asian
CelticEnglish	0.020	0.025	0.484	0.038	0.319	0.006	0.016	0.033	0.035	0.022
EastAsian	0.018	0.017	0.354	0.243	0.250	0.004	0.016	0.026	0.036	0.035
European	0.022	0.023	0.420	0.086	0.326	0.005	0.016	0.043	0.032	0.027

Table 9: Quoted + cited speaker name origin, by journalist name origin

Journalist Name Origin	African	Arab Turk Pers	Celtic English	East Asian	European	Greek	Hebrew	Hispanic	Nordic	South Asian
CelticEnglish	0.016	0.027	0.368	0.070	0.363	0.008	0.017	0.023	0.083	0.025
EastAsian	0.002	0.077	0.377	0.143	0.167	0.000	0.012	0.133	0.019	0.080
European	0.014	0.028	0.363	0.116	0.352	0.006	0.030	0.026	0.035	0.030

Table 10: Quoted speakers (with US affiliated citation) name origin, by journalist name origin

Journalist Name Origin	African	Arab Turk Pers	Celtic English	East Asian	European	Greek	Hebrew	Hispanic	Nordic	South Asian
CelticEnglish	0.011	0.023	0.378	0.086	0.361	0.010	0.021	0.029	0.056	0.025
EastAsian	0.000	0.066	0.340	0.148	0.209	0.000	0.005	0.148	0.033	0.049
European	0.021	0.030	0.410	0.111	0.300	0.012	0.023	0.019	0.030	0.046

References

1. **The enduring whiteness of the American media**
   Howard W French
   The Guardian (2016-05-25) https://www.theguardian.com/world/2016/may/25/enduring-whiteness-of-american-journalism

2. https://medium.com/ladybits-on-medium/i-analyzed-a-year-of-my-reporting-for-gender-bias-and-this-is-what-i-found-a16c31e1cdf

3. **I Analyzed a Year of My Reporting for Gender Bias (Again)**
   Adrienne LaFrance
   The Atlantic (2016-02-17) https://www.theatlantic.com/technology/archive/2016/02/gender-diversity-journalism/463023/

4. **I Spent Two Years Trying to Fix the Gender Imbalance in My Stories**
   Ed Yong
   The Atlantic (2018-02-06) https://www.theatlantic.com/science/archive/2018/02/i-spent-two-years-trying-to-fix-the-gender-imbalance-in-my-stories/552404/

5. **A Paper Ceiling**
   Eran Shor, Arnout van de Rijt, Alex Miltsov, Vivek Kulkarni, Steven Skiena
   American Sociological Review (2015-09-30) https://doi.org/f7tzps
   DOI: 10.1177/0003122415596999

6. **Time Trends in Printed News Coverage of Female Subjects, 1880–2008**
   Eran Shor, Arnout van de Rijt, Charles Ward, Aharon Blank-Gomel, Steven Skiena
   Journalism Studies (2013-09-12) https://doi.org/gj3z8b
   DOI: 10.1080/1461670x.2013.834149

7. **Women and news: A long and winding road**
   Karen Ross, Cynthia Carter
   Media, Culture & Society (2011-11) https://doi.org/ccxhvz
   DOI: 10.1177/0163443711418272

8. **Women Are Seen More than Heard in Online Newspapers**
   Sen Jia, Thomas Lansdall-Welfare, Saatviga Sudhahar, Cynthia Carter, Nello Cristianini
   PLOS ONE (2016-02-03) https://doi.org/f8q47g
   DOI: 10.1371/journal.pone.0148434 · PMID: 26840432 · PMCID: PMC4740422

9. **Lack of female sources in NY Times front-page stories highlights need for change**
   Alexi Layton and Alicia Shepard
   Poynter (2013-07-16) https://www.poynter.org/reporting-editing/2013/lack-of-female-sources-in-new-york-times-stories-spotlights-need-for-change/

10. **Who Makes the News | GMMP 2015 Reports** https://whomakesthenews.org/gmmp-2015-reports/

11. **The gender gap in science: How long until women are equally represented?**
    Luke Holman, Devi Stuart-Fox, Cindy E Hauser
    PLOS Biology (2018-04-19) https://doi.org/gdb9db
    DOI: 10.1371/journal.pbio.2004956 · PMID: 29672508 · PMCID: PMC5908072

12. **Diversity and STEM: Women, Minorities, and Persons with Disabilities 2023 | NSF - National Science Foundation** https://ncses.nsf.gov/pubs/nsf23315/

13. **Why we need to increase diversity in the immunology research community**
Akiko Iwasaki
Nature Immunology (2019-08-19) https://doi.org/gkmwwv
DOI: 10.1038/s41590-019-0470-6 · PMID: 31427777

14. **Men Set Their Own Cites High: Gender and Self-citation across Fields and over Time**
Molly M King, Carl T Bergstrom, Shelley J Correll, Jennifer Jacquet, Jevin D West
Socius: Sociological Research for a Dynamic World (2017-01-01) https://doi.org/ddzq
DOI: 10.1177/2378023117738903

15. **Bibliometrics: Global gender disparities in science**
Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, Cassidy R Sugimoto
Nature (2013-12) https://doi.org/qgf
DOI: 10.1038/504211a · PMID: 24350369

16. **Fund Black scientists**
Kelly R Stevens, Kristyn S Masters, PI Imoukhuede, Karmella A Haynes, Lori A Setton, Elizabeth Cosgriff-Hernandez, Muyinatu A Lediju Bell, Padmini Rangamani, Shelly E Sakiyama-Elbert, Stacey D Finley, … Omolola Eniola-Adefeso
Cell (2021-02) https://doi.org/ghvqv5
DOI: 10.1016/j.cell.2021.01.011 · PMID: 33503447

17. **NIH peer review: Criterion scores completely account for racial disparities in overall impact scores**
Elena A Erosheva, Sheridan Grant, Mei-Ching Chen, Mark D Lindner, Richard K Nakamura, Carole J Lee
Science Advances (2020-06-05) https://doi.org/gjnjbz
DOI: 10.1126/sciadv.aaz4868 · PMID: 32537494 · PMCID: PMC7269672

18. **Topic choice contributes to the lower rate of NIH awards to African-American/black scientists**
Travis A Hoppe, Aviva Litovitz, Kristine A Willis, Rebecca A Meseroll, Matthew J Perkins, Blan Hutchins, Alison F Davis, Michael S Lauer, Hannah A Valantine, James M Anderson, George M Santangelo
Science Advances (2019-10-11) https://doi.org/gghp8t
DOI: 10.1126/sciadv.aaw7238 · PMID: 31633016 · PMCID: PMC6785250

19. **The Gender Gap in NIH Grant Applications**
Timothy J Ley, Barton H Hamilton
Science (2008-12-05) https://doi.org/frdj6k
DOI: 10.1126/science.1165878 · PMID: 19056961

20. **Race, Ethnicity, and NIH Research Awards**
Donna K Ginther, Walter T Schaffer, Joshua Schnell, Beth Masimore, Faye Liu, Laurel L Haak, Raynard Kington
Science (2011-08-19) https://doi.org/csf8j8
DOI: 10.1126/science.1196783 · PMID: 21852498 · PMCID: PMC3412416

21. **Including Diverse Voices in Science Stories**
Christina Selby
The Open Notebook (2016-08-23) https://www.theopennotebook.com/2016/08/23/including-diverse-voices-in-science-stories/

22. **gage. Discover Brilliance** https://gage.500womenscientists.org/

23. **WMC SheSource - Women's Media Center** https://womensmediacenter.com/shesource

24. **Scrapy | A Fast and Powerful Scraping and Web Crawling Framework** https://scrapy.org/

25. **Gender Prediction Methods Based on First Names with genderizeR**
Kamil Wais
The R Journal (2016) https://doi.org/gf4zqx
DOI: 10.32614/rj-2016-002

26. **Determine gender from a name - Accurate gender prediction API - Genderize.io**
https://genderize.io/

27. **Comparison and benchmark of name-to-gender inference services**
Lucía Santamaría, Helena Mihaljević
PeerJ Computer Science (2018-07-16) https://doi.org/ggpbn6
DOI: 10.7717/peerj-cs.156 · PMID: 33816809 · PMCID: PMC7924484

28. **Using genderize.io to infer the gender of first names: how to improve the accuracy of the inference**
Paul Sebo
Journal of the Medical Library Association (2021-11-22) https://doi.org/gn94vp
DOI: 10.5195/jmla.2021.1252 · PMID: 34858090 · PMCID: PMC8608220

29. **Analysis of ISCB honorees and keynotes reveals disparities**
Trang T Le, Daniel S Himmelstein, Ariel A Hippen Anderson, Matthew R Gazzara, Casey S Greene
Cold Spring Harbor Laboratory (2020-04-14) https://doi.org/ggr64p
DOI: 10.1101/2020.04.14.927251

30. **Underrepresentation of Asian awardees of United States biomedical research prizes**
Yuh Nung Jan
Cell (2022-02) https://doi.org/gpgvs2
DOI: 10.1016/j.cell.2022.01.004 · PMID: 35120660

31. **Why scientific societies should involve more early-career researchers**
Adriana Bankston, Stephanie M Davis, Elisabeth Moore, Caroline A Niziolek, Vincent Boudreau
eLife (2020-09-23) https://doi.org/gtg9gw
DOI: 10.7554/elife.60829 · PMID: 32965217 · PMCID: PMC7511228

32. **Examining trends in the diversity of the U.S. National Institutes of Health participating and funded workforce**
Silda Nikaj, Deepshikha Roychowdhury, PKay Lund, Marguerite Matthews, Katrina Pearson
The FASEB Journal (2018-06-19) https://doi.org/gdqrqh
DOI: 10.1096/fj.201800639 · PMID: 29920223

33. **Racial and ethnic imbalance in neuroscience reference lists and intersections with gender**
Maxwell A Bertolero, Jordan D Dworkin, Sophia U David, Claudia López Lloreda, Pragya Srivastava, Jennifer Stiso, Dale Zhou, Kafui Dzirasa, Damien A Fair, Antonia N Kaczkurkin, … Danielle S Bassett
Cold Spring Harbor Laboratory (2020-10-12) https://doi.org/gj7mdc
DOI: 10.1101/2020.10.12.336230

34. **The Stanford CoreNLP Natural Language Processing Toolkit**

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, David McClosky
Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (2014) https://doi.org/gf3xhp
DOI: 10.3115/v1/p14-5010

35. **humaniformat: A Parser for Human Names**
Oliver Keyes
(2016-04-24) https://cran.r-project.org/web/packages/humaniformat/index.html

36. **Nationality Classification Using Name Embeddings**
Junting Ye, Shuchu Han, Yifan Hu, Baris Coskun, Meizhu Liu, Hong Qin, Steven Skiena
Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (2017-11-06) https://doi.org/ggjc78
DOI: 10.1145/3132847.3133008

37. **Avoiding bias when inferring race using name-based approaches**
Diego Kozlowski, Dakota S Murray, Alexis Bell, Will Hulsey, Vincent Larivière, Thema Monroe-White, Cassidy R Sugimoto
PLOS ONE (2022-03-01) https://doi.org/gthgs3
DOI: 10.1371/journal.pone.0264270 · PMID: 35231059 · PMCID: PMC8887775

38. **Open collaborative writing with Manubot**
Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter
PLOS Computational Biology (2019-06-24) https://doi.org/c7np
DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653