DameGender: Towards an international and free dataset about name, gender and frequency

David Arroyo Menéndez, Madrid, Spain
{davidam@gmail}.com

DAMEGENDER

Abstract

Equality of gender is the fifth objective of sustainable development for United Nations[1].

This equality can be reached by measuring and analyzing data and making good politics with the results. Many gender studies count males and females based on their names, for instance, research papers, job positions, streets, ... The traditional research method is to use commercial APIs with proprietary data without idea about how the data was collected. Data may also be gathered from Wikipedia, lingüistic studies, scientific sites, or statistical offices.

This approach is based collecting Open Datasets regarding name, gender and frequency from many statistical institutions. So, we need a scientific discussion about unifying formats and processing data easily.

Therefore, Damegender (Free and Open Source Software) to retrieve and make calculus with these data.

The dataset we used covers more than 20 countries in the occidental world encompassing many names with an accuracy of approximately 90with it. This will create to measure gender gap to students and academics interested on the phenomenon without costs and on a reproducible way and more people will be contributing to fix the gender gap.

Free software and the data provided by statistical institutions make it possible to produce reproducible research for peer review. Thus, semantics and diversity can be more easily addressed.

[1]https://www.un.org/sustainabledevelopment/gender-equality/

1 Introduction

The United Nations has a goal to address the gender gap[2]. You must remember that "if you cannot measure it, you cannot improve it" [Tho33] and "Software Engineering Economics is an invaluable guide to determining software costs, applying the fundamental concepts of microeconomics to software engineering [B+81]". Free software and open data lead to a reduction in costs, for example, many people and institutions is using LibreOffice and Ubuntu (GNU/Linux) to avoid paying the fees with similar products such as Microsoft Windows and Microsoft Office. Gender detection tools based on the user name is based on API solutions, providing a free software and open data solution. This will createit competition in a market without a very strong leader, avoiding payments and strategizing profits from a trademark, such as, Firefox or Chrome.

Through the use of personal names, one may infer gender on academical papers, books, newspapers and many interactions on Internet. So, detecting gender from the names may be a strategical way to measure gender gap.

Many users today are using APIs such as Gender-api, Genderize, Namsor, or NameApi, Wikipedia, or Free Software solutions (NLTK[LB02], R Gender, Gender Detector and Gender Computer[3].

Traditional open source solutions has a few number of names due to use files of a single country or being software not maintained in the long time. And Wikipedia is storing few names per country.

However, the gender gap is a problem recognised in United Nations and the IT market is leading big

[2]https://www.un.org/sustainabledevelopment/gender-equality/

[3]https://github.com/tue-mdse/genderComputer

inequalities in economic and gender gap. This article presents data collected to assist in finding the solution to a number of problems (search engine, infering gender in csv files, names in different countries, wide dataset) faced by the industry as well as other problems not solved in an industrial way such as counting males and females in GitHub repositories, mailing lists, ...

A previous study [KWL+16] dicussed these datasets as a way to improve their accuracy, comparing tools that use different public datasets (SSA [4], IPUMS [5], namdict [6], etc)

So, a goal is to augment the number of names using official statistics and taking into account diversity goals such as non binary gender and cultural minorities.

With DameGender we will make science reproducible[Pen11] in fields with similar works such as Natural Language Processing (gender detection from the name [SGT+19]), social sciences or journalism (gender gap [HSFH18, MLA+11, NP17, dBA14]), linguistic [Hut16, vdWRvdW+20, Oka18], software engineering [VCS12], among other fields.

The remainder of this article is structured as follows:

Section 2 presents the main research measuring the gender gap and gender detection tools using name.

Section 3 gives vocabulary and philosophy about to choose sources and to face the diversity troubles building a dataset.

Section 4 explains an application for this dataset: to measure gender gap in GNU/Linux.

Section 5 points a summary about this approach and future works.

The contributions of this article are:

1. An integrated solution in the different applications field relative to inferring gender from the name.
2. A collection of open datasets retrieved from statistical sources and standardized in an unique format.
3. A new study applying DameGender to count males and females in GNU/Linux.
4. An approach based on reproducible results.

2 State of the Art

2.1 About Gender Gap

To reduce gender gap refers to equality between males and females, and non discrimination policies. Gender refers to the sex of a person determined in the moment of the birth, although it can be changed throughout life. Discussions about gender definitions refers to these problems. However, there is a consensus determining gender, frequency and names with official statistics released by the institutions in the states.

Measuring the gender gap requires set indicators. [For21] has been proposed economy, health, education and politics. And in United Nations[7] there are indicators used to measure disparities such as laws, education, maternal mortality, political participation, poverty, domestic work, gender parity in the work, to access to the economy, youth issues (access to studies and/or work), violence against women, climate justice, access to the justice, health, ...

It's possible to make impactful decisions on an issue through research results that have taken these indicators into consideration. For example, [MKSF+10] concluded that making affirmations about ethical values reduced the gender achievement gap in colleges.

Measuring the gender gap in social research, such as the survey. For example, in [Bim00] presented two factors affecting the gender gap on the Internet (access and use) by socioeconomic and gender reasons in a survey that collect data over several years. Internet access is vital to today's in the economy, education. [RRGBD16] is using a survey of 2000 contributors.

2.2 Counting males and females on the Internet. Why? Where?

This work focused on retrieving data from secondary sources such as GitHub, Wikipedia, APIs, websites in general, mailing lists, etc. Previous research works about factors modifying several gender gap indicators (economy, education, politics) were obtained from secondary sources.

For example, a social scientist studying gender gap in journalism [ÁACS12] can count males and females on Twitter. These metrics are important because the journalism is evaluating gender gap in political, education, or the economy, ... Meanwhile, Computer Science making research about how to count males and females in Twitter [BHKZ11]. In these studies the name, nickname, photo, and identifying gender are retrieved from these data.

[BHKZ11] presented several configurations of a language-independent classifier for predicting the gender of Twitter users. The large dataset used for the construction and evaluation of these classifiers was drawn from Twitter users who also completed blog profile pages.

[MLA+11] analyzed the Twitter population, including the gender. The gender was inferred making queries from the names to the dataset provided by the United States Census Bureau.

[4] https://www.ssa.gov/oact/babynames/limits.html
[5] https://usa.ipums.org/usa-action/variables/NAMEFRST
[6] https://raw.githubusercontent.com/lead-ratings/gender-guesser/master/gender_guesser/data/nam_dict.txt
[7] https://www.unwomen.org/

[WGJS15] analyzed the gender gap in Wikipedia, showing evidence of more subtle forms of gender inequality explaining how to solve these evidences. To measure gender inequality has been developed the next bias: coverage, structural, lexical (ex: discriminatory words for women), and visibility.

Computer Science is generating many Forbes billionaires and the public code may help to understand the gender gap in this field, which may have some importance to the economy. Public repositories can be used to build indicators about the economy in Computer Science with more factors, such as job positions, value of companies, etc. [Zac20] conducted the first large-scale longitudinal study of gender imbalance among authors of collaboratively developed, publicly available code, where contributions by female authors remain scarce less that 8 % of commits was able to be detected were from women, confirming decades of gender imbalance in Free/Open Source Software (FOSS). Steffano used to namdict [8] dataset with genderguesser to infer gender from the name. [VPR+15] determined that women programmers are in the minority in OSS and other technical fields, although increased gender and tenure diversity is associated with greater productivity.

[VCS12] explored the popular Q&A about technological issues called StackOverflow, which summarizes that the percentage of women engaged in SO is greatly imbalanced, and men represent the vast majority of contributors.

[IHSR18] revealed that data few females contribute code or take political responsability in the OpenStack community.

Related to the gender gap in science, [HSFH18] presented a code in R using genderize API and provides a good approach about how to calculate gender gap inferring gender from an author names retrieved from arXiv.

2.3 Automatic approaches to infer gender

There are several ways to infer gender from Internet sources: hand written, images, documents and names.

[LSB11] presents a method inferring gender from hand written texts with a 67.5 % accuracy.

[GC08] combines image based gender and age classifiers with the cultural information provided by first names to recognize people with no labeled examples with results near to 60 % accurate.

[AKFS03] explains that females use many more pronouns, while males use many more noun specifiers, in a large subset of the British National Corpus covering a range of genres. Therefore, [KAS02] presented a document classification system with accuracy of approximately 80 %. [CCS11] exposes a feature selection and a model built using machine learning resulting in 85.1 % accurate rate for identifying gender from text.

[8] https://raw.githubusercontent.com/lead-ratings/gender-guesser/master/gender_guesser/data/nam_dict.txt

2.4 Infering gender from name

The tools used to infer gender from a name are tipically based on datasets that, at a minimum, is include gender and name as minimum.

[LR13] presented a method to infer gender from first names in Twitter, the dataset was hand coded by agreement between three Amazon workers with 50,000 Twitter users select at random with only 12,681 gender labels. The goal of this study was to determine the incremental value of using the user name as a feature in gender inference based on tweets.

[MS16] presented how to infer gender in Twitter. They used namdict and the United States census as datasets. The features were 'number of consonants', 'number of vowels', 'number of syllables', 'number of bouba consonants', 'number of bouba vowels', 'number of kiki consonants', 'number of kiki vowels'. The classification model was created using SVM.

2.5 Related ideas

[AWM+09] presented a system to classify name and ethnicity from open sources using machine learning to extract a name list from Wikipedia. A more recent work is [NRN21], in which presented NamPrism was applied to massive software repositories.

[BMI10] presented another approach that used a lexical-pattern-based approach to extract aliases of a given name, with a set of names and their aliases as training data to extract lexical patterns. The candidates are ranked using various ranking scores. Support vector machines were used to construct the ranking function.

2.6 Related Standards

ISO/IEC 5218 proposes the following norm about coding gender: "0 as not know", "1 as male", "2 as female" and "9 as not applicable".

The RFC 6350 (vCard) [9] has these categories: "m as male", "f as female", "o as other", "n as not applicable" and "u as undefined". Based on this standard, those conducting web publishing can use CSS classes using a web standard such as h-card [10] microformats in the context of to write forms in web interfaces consider w3 lectures [11]

[9] https://datatracker.ietf.org/doc/html/rfc6350

[10] https://github.com/microformats/h-card

[11] https://www.w3.org/International/questions/qa-personal-names

2.7 Summary

The first name of a subject is the is the key factor used to determine gender in the State of Art gender inference tool. However, in many contexts there are more features: surnames, text, images, nicknames, ... The first name can be useful to infer another stuff such as race, ethnicity or culture, too.

Machine learning and the previous features selection is being used in many works, although there is an open discussing as to which is the best approach

The datasets can be built by human experts, although there are some open datasets used several times in these researches, such as namdict, or the United States census.

3 Design

3.1 Truth and Falsehood in names, gender and frequency

The current idea in the field is the data about name, gender and frequency is ok because there are people who is paying by it, or many people is downloading a product. This intuition is right generally, although sometimes the people is paying by a bad product due to a good marketing strategy, a monopoly or there is a fraud, ... Another idea is the people trust in the government about statistics such as economy, demography, democracy, ... So the people can trust on names, gender and frequency. In Damegender, we are trusting in both notions about truth: the market's point of view and the official statistic's point of view.

Sometimes there are problems downloading the official statistics, but there are people who has retrieved these data, for example, with web scraping. We want classify these files with another idea about truth.

Another problem arises when the government does little chances in the data, sometimes communicating it to the users and other times not. That could be a problem about upgrades, but it's not a problem with the truth, although it's possible make a trace about these chances.

With an international free dataset about names, gender and frequency we can build reproducible science in fields such as Natural Language Processing (gender detection from the name), social sciences or journalism (gender gap [HSFH18, MLA+11, NP17, dBA14]), linguistic [LN05, Kru62, vdWRvdW+20, Agy06, FMO+87], software engineering [VCS12], ...

3.2 Gender, Language, Nation and Diversity

There are rules and exceptions in the languages to predict if a name is about male or female when you don't know the name. For example, in Spanish or English there are more names ending with 'a' classified as females than classified as males. And Andrea is female in Spain and male in Italy. So, it's useful to understand the language and culture associated with a name. Language is close to nation, but there are differences, for example, in Spain there are several languages basque, catalan, castillian, ... or the Spanish is the main language in Spain and in other countries such as Argentina, Mexico, Ecuador, Bolivia, ... So, it would be useful to detect the language and nation from names and surnames to help to detect gender.

Some countries, such as Spain, are providing free datasets about surnames but we need more efforts from many countries on this objective. On other hand, there are previous works to relate name and surnames with ethnicity using Wikipedia and Machine Learning [AWM+09].

3.3 Damegender Open Datasets Collection

In Damegender, we have unified the different formats to name, gender and frequency from official sources in these countries: Argentina, Austria, Australia, Belgium, Canada, Denmark, Germany, Spain, Finland, France, Great Britain, Ireland, Mexico, New Zealand, Norway, Russia, Portugal, Slovenia, United States of America and Uruguay.

We have found 2 main criteria counting males and females: number of births in a year and people using the name in a year. So, we have divided the files being to able of make merging. The criteria making the count is: the average of the last twenty years where the data has been provided. Both criteria are good indicators to understand how many people is using a name as male or as female.

Later, we have merged these datasets building a free and international dataset.

We have found open datasets about countries such as Turkey and China retrieved by other open source developers that is being included in Damegender, but not in the international dataset. In Turkey the data has been retrieved using web scraping. And in China the data has been built by a company in collaboration with the China government and contributed to R language program. We want compare precision about this dataset with the commercial solutions to understand the truth about these datasets.

We have found surnames given by statistical institutions in Spain, Russia, United States of America and Argentina. So few statistical institutions is giving surnames. Understanding the diversity problem with this fact, we have retrieved surnames for all countries from Wikidata, these datasets contains few surnames being compared with datasets provided by statistical institutions. Although in names the diversity problem would be a minor problem we are giving names retrieved with

Dataset	SSA	namdict	NLTK	Damegender
males	91.320	48.821	2.943	278.928
females	91.320	48.821	5.001	299.870

Table 1: Comparison about the number of names between Open Data solutions

Dataset	Accuracy	Precision	Recall	F1-Score
Damegender	0.8756	0.9638	1.0	0.925

Table 2: Several precision measures about the Damegender International Dataset

Wikidata to the public. We are releasing a free dataset about surnames and frequencies, too.

When the work is finished, we could to rebuild machine learning models to predict new names and nicknames in any language and culture. The results is the longest list of public names.

A possible criticism about our idea is the Leslie Problem[BM15]: the match between gender and name has been changing in some years. And the answer is about you need introduce the age of the person to solve it. The most used use case is the input is the name and the output must be gender, frequency and percentage. So, we are deciding without age, surname, ... in the most of use cases. The idea about this dataset is to be designed for the most used use case. Although, we can take into account other inputs, such as surname or age to improve the accuracy. There are many Open Datasets with names and frequencies classified by years. So, this problem can be fixed with Open Data, too.

We have made measurements about the international DameGender dataset, using as base of truth the dataset explained in [SM18] reaching accuracy (0.8756), precision (0.9638), recall (1.0) and f1-score (0.925). With other test datasets we are finding similar results:

- Females scientists in Wikipedia (accuracy: 0.89)
- Males scientists in Wikipedia (accuracy: 0.98)
- FIFA football dataset (accuracy: 0.93)
- Scientific test Dataset designed[Seb21] (accuracy: 0.88)

3.4 Free APIs for Free Datasets?

Many websites about Open Data is delivering methods to retrieve Open Data with a structured format and without costs for the user, such as, Wikipedia with SPARQL, OpenStreetMap with API rest, ...

We are detecting that the Open Datasets about names, gender and frequency are being modified one time per year as maximum for each statistical institution.

Damegender contains python scripts designed to create the different datasets and publish json files that could be used as a Free API Rest publishing the json files in sites as github pages, gitlab pages, or similar sites with free uploads.

```
$ cat DAVID_all.json
[{
    "name": "DAVID",
    "frequency": 4856689,
    "males": "99.73267796229078 %",
    "females": "0.26732203770922947 %"
}]
```

So, we could think that could have Free API Rest about names, gender and frequency reaching reduce costs to fix the gender gap on a collaborative way similar to Wikipedia, OpenStreetMap or many Free Software projects.

4 Measuring Gender Gap. GNU/Linux as Use Case

With a trust open dataset about names, gender and frequency is too easy to measure gender gap. Doing cheap to measure gender gap more students and academic people could work in the fifth Objective Development Sustainable of United Nations: to delete the gender gap.

This section is divided counting males and females in Debian, GNU and Linux.

We have reached the csv files from different ways to know the names about the people in these communities.

When this paper was being wrote in the Debian community all members must be collaborating with a gpg key, so we can count males and females from the keyring. The keyring was imported with gpg commands and later was dumped the keyring in a csv file.

In the moment to write this paper GNU[12] and Linux[13] has websites with the people collaborating in these projects. So, making web scraping scripts we have downloaded the people and processed the people to csv files

In Damegender, we have developed csv2gender, a software with a csv file as input and deploy a statistics graph and/or return the result of males, females and unknowns about the input.

To make easy to reproduce the experiment we are pasting the commands used with the version 0.3.4 of Damegender.

```
python3 csv2gender.py files/gnu.csv
```

[12]https://www.gnu.org/people/

[13]https://www.kernel.org/doc/html/latest/process/maintainers

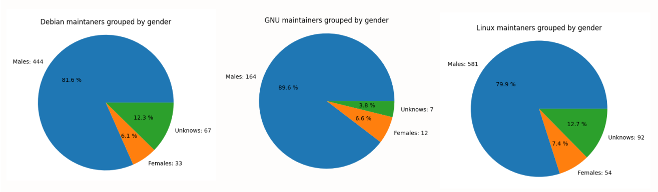

Figure 1: Males (blue), Females (orange) and Unknows (green) in Debian, GNU and Linux

```
 --first_name_position=0
 --title="GNU maintainers grouped by gender"
 --dataset="inter"
 --outcsv="files/gnu.gender.csv"
 --outimg="files/gnu.gender.png"
 --noshow --delete_duplicated

python3 csv2gender.py files/linux.csv
 --first_name_position=0
 --title="Linux maintaners grouped by gender"
 --dataset="inter"
 --outcsv="files/linux.gender.csv"
 --outimg="files/linux.gender.png"
 --noshow --delete_duplicated

python3 csv2gender.py files/debian.csv
 --first_name_position=0
 --title="Debian maintaners grouped by gender"
 --dataset="inter"
 --outcsv="files/debian.gender.csv"
 --outimg="files/debian.gender.png"
 --noshow --delete_duplicated
```

The inter dataset was created merging several open datasets downloaded from official statistics sites from different nations: Austria, Australia, Belgium, Canada, Germany, Denmark, Spain, Finland, Ireland, Iceland, Mexico, New Zealand, Portugal, Slovenia, United States of America, Uruguay and France. That's a good representation of the Western World and the Free Software world is populating this world's area[GBRAIG08].

Linux divides the developers in 537 males (73.9%), 98 females (13.5%) and 92 unknowns (12.7%). The number of unknowns is due to different reasons, but it's so common in Linux that the developer is a company and not a name of a person.

GNU divides the developers in 164 males (89.6%), 12 females (6.6%) and 7 unknowns (3.8%)

The GNU people has a number lowest in females, they are the founder of the Free Software philosophy, the Debian principles and the Open Source philosophy was invented later influenced by GNU with very similar practical decisions (for example: deciding licenses for the software). Richard Stallman returned to be president recently apologizing by his personal behaviour with the females.[14]

Debian is a distribution, the project who makes the CD/DVD and the software ready to be downloaded from Internet with the dependencies. There are many distributions, such as, Ubuntu or RedHat so it is not representative, but it's interesting to understand that the numbers are similar in Debian dividing the developers in 408 males (75%), 69 females (12.7%) and 67 unknowns (12.3%).

5 Conclusions and Future Works

This paper is explaining the application about Damegender, the motivations (reproducible research, fix gender gap to reach an objective of United Nations, fields of application: linguistic, social sciences, software engineering, natural language processing, journalism, ...)

A good improvement is to build an international, universal and free dataset about names, gender and frequency with the right design with the current state of the job, attending to the diversity (LGBT options, cultural minorities, ...).

This paper has explained what technologies is involved on reduce costs about gender gap (gender detection from the names, api rest, semantic web, ...)

Augmenting the number of countries with statistical institutions giving names, gender and frequencies with Open Data will be augmenting the accuracies and giving more attention to the diversity.

The current state of work is the longest Open Dataset about names, gender and frequency with more than 20 countries representing the Western World, being a solution with low number of unknowns in the real world.

The future works is about changes in the big software industry.

Making searches with strings about personal names (ex: Leticia) in search engines such as Google, these strings are not being classified as personal names, one solution will be data structured such as JSON-LD, microdata, microformats, rdfa ... Another solution will be store in the servers the Open Data Collection about names, gender and frequency and identify the context about the string is a personal name, that's an easy problem in popular sites such as Wikipedia, academic websites, ...

If the search engine identify the string as personal name, it can help to the user about the gender. That is similar than other problems such as streets, products, ... where you are giving additional information such as maps in streets or prices in products.

[14] https://www.fsf.org/news/rms-addresses-the-free-software-community

Other sites such as Github or Gitlab could be giving data about gender of developers in the site or in the software project with these datasets.

Another industry is about match sites (Meetic, Tinder, ...) where only is important photos, age and gender generally. It could be possible to give to users gender, photo and interests from personal names our open data collection and information related in Internet.

Acknowledgments

We would like to thank: Statistical institutions by release Open Datasets about names, gender and frequency. Luz Galvis by the software contributions, Daniel Izquierdo and Laura Arjona for starting this research field at URJC all those working with Jesús González Barahona and Gregorio Robles.

References

[ÁACS12] Pilar Carrera Álvarez, Clara Sainz De Baranda Andújar, Eva Herrero Curiel, and Nieves Limón Serrano. Journalism and social media: How spanish journalists are using twitter/periodismo y social media: cómo están usando twitter los periodistas españoles. *Estudios sobre el mensaje periodístico*, 18(1):31, 2012.

[Agy06] Kofi Agyekum. The sociolinguistic of akan personal names. *Nordic journal of African studies*, 15(2):206–235, 2006.

[AKFS03] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Text & Talk*, 23(3):321–346, 2003.

[AWM+09] Anurag Ambekar, Charles Ward, Jahangir Mohammed, Swapna Male, and Steven Skiena. Name-ethnicity classification from open sources. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 49–58, 2009.

[B+81] Boehm Barry et al. Software engineering economics. *New York*, 197, 1981.

[BHKZ11] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[Bim00] Bruce Bimber. Measuring the gender gap on the internet. *Social science quarterly*, pages 868–876, 2000.

[BM15] Cameron Blevins and Lincoln Mullen. Jane, john... leslie? a historical method for algorithmic gender prediction. *DHQ: Digital Humanities Quarterly*, 9(3), 2015.

[BMI10] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Automatic discovery of personal name aliases from the web. *IEEE Transactions on Knowledge and Data Engineering*, 23(6):831–844, 2010.

[CCS11] Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

[dBA14] Clara Sainz de Baranda Andújar. El género de los protagonistas en la información deportiva (1979-2010): noticias y titulares/the gender of the main characters in sports reporting (1979-2010): News and headlines. *Estudios sobre el mensaje periodístico*, 20(2):1225, 2014.

[FMO+87] Peter Marshall Fraser, Elaine Matthews, Michael J Osborne, Sean G Byrne, Richard WV Catling, J-S Balzat, E Chiricat, Thomas Corsten, and Fabienne Marchand. *A lexicon of Greek personal names*, volume 5. Lexicon of Greek Personal Name, 1987.

[For21] World Economic Forum. Global gender gap report 2021: Insight report. World Economic Forum Cologny, Switzerland, 2021.

[GBRAIG08] Jesus M Gonzalez-Barahona, Gregorio Robles, Roberto Andradas-Izquierdo, and Rishab Aiyer Ghosh. Geographic origin of libre software developers. *Information Economics and Policy*, 20(4):356–363, 2008.

[GC08] Andrew C Gallagher and Tsuhan Chen. Estimating age, gender, and identity using first name priors. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[HSFH18] Luke Holman, Devi Stuart-Fox, and Cindy E Hauser. The gender gap in science: How long until women are equally represented? *PLoS biology*, 16(4):e2004956, 2018.

[Hut16] ¿ Matthew Hutson. The gender of names. *Scientific American Mind*, 27(4):14–14, 2016.

[IHSR18] Daniel Izquierdo, Nicole Huesman, Alexander Serebrenik, and Gregorio Robles. Openstack gender diversity report. *IEEE Software*, 36(1):28–33, 2018.

[KAS02] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412, 2002.

[Kru62] John R Krueger. Mongolian personal names. *Names*, 10(2):81–86, 1962.

[KWL+16] Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International conference companion on World Wide Web*, pages 53–54, 2016.

[LB02] Edward Loper and Steven Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[LN05] Edwin D Lawson and Natan Nevo. Russian given names: Their pronunciation, meaning, and frequency. *Names*, 53(1-2):49–77, 2005.

[LR13] Wendy Liu and Derek Ruths. What's in a name? using first names as features for gender inference in twitter. In *2013 AAAI Spring Symposium Series*, 2013.

[LSB11] Marcus Liwicki, Andreas Schlapbach, and Horst Bunke. Automatic gender detection using on-line and off-line information. *Pattern Analysis and Applications*, 14(1):87–92, 2011.

[MKSF+10] Akira Miyake, Lauren E Kost-Smith, Noah D Finkelstein, Steven J Pollock, Geoffrey L Cohen, and Tiffany A Ito. Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, 330(6008):1234–1237, 2010.

[MLA+11] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. Understanding the demographics of twitter users. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[MS16] Juergen Mueller and Gerd Stumme. Gender inference using statistical name characteristics in twitter. In *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016*, pages 1–8, 2016.

[NP17] Mari K Niemi and Ville Pitkänen. Gendered use of experts in the media: Analysis of the gender gap in finnish news journalism. *Public Understanding of Science*, 26(3):355–368, 2017.

[NRN21] Reza Nadri, Gema Rodriguezperez, and Meiyappan Nagappan. On the relationship between the developer's perceptible race and ethnicity and the evaluation of contributions in oss. *IEEE Transactions on Software Engineering*, 2021.

[Oka18] Benard Odoyo Okal. A linguistic overview of the patronymic and gender names amongst the selected african communities. 2018.

[Pen11] Roger D Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.

[RRGBD16] Gregorio Robles, Laura Arjona Reina, Jesús M. González-Barahona, and Santiago Dueñas Domínguez. Women in free/libre/open source software: The situation in the 2010s. In Kevin Crowston, Imed Hammouda, Björn Lundell, Gregorio Robles, Jonas Gamalielsson, and Juho Lindman, editors, *Open Source Systems: Integrating Communities*, pages 163–173, Cham, 2016. Springer International Publishing.

[Seb21] Paul Sebo. Performance of gender detection tools: a comparative study of name-to-gender inference services. *Journal of the Medical Library Association: JMLA*, 109(3):414, 2021.

[SGT+19] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

[SM18] Lucía Santamaría and Helena Mihaljević. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4:e156, July 2018.

[Tho33] W Thompson. Electrical units of measurement, popular lectures, 1833.

[VCS12] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. Gender, representation and online participation: A quantitative study of stackoverflow. In *2012 International Conference on Social Informatics*, pages 332–338. IEEE, 2012.

[vdWRvdW+20] Jeroen van de Weijer, Guangyuan Ren, Joost van de Weijer, Weiyun Wei, and Yumeng Wang. Gender identification in chinese names. *Lingua*, 234:102759, 2020.

[VPR+15] Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark GJ van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov. Gender and tenure diversity in github teams. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3789–3798. ACM, 2015.

[WGJS15] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media*, 2015.

[Zac20] Stefano Zacchiroli. Gender differences in public code contributions: a 50-year perspective. *IEEE Software*, 38(2):45–50, 2020.