EasyChair Preprint

№ 5763

Damegender: Towards an International and Free Dataset about Name, Gender and Frequency

David Arroyo Menéndez

January 30, 2023

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, etc. 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 greater than 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 Introduction

The United Nations has a goal to address the gender gap[2], but "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.

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

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

Figure 1: Grey (Countries with Open Data provided by Statistical Institutions), White (Countries without Open Data or don't reproduced by the author)

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 Genderapi, 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 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, etc.

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, AZ09], 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.

Many articles related are about to apply machine learning techniques, but using only Open Data is possible to acquire very good accuracies. So, this work is augmenting the base of truth of the state of art related to Open Data, so this work is helping to look for datasets checking accuracies with several data sources in statistical institutions and other Open Data trust sources as Wikipedia, Project Gutenberg, Amazon, sport institutions, Forbes, etc.

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

[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

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. Global Gender Gap Report [CPSZ22] has been proposed economy, health, education and politics. The United Nations site[7] is showing indicators 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, etc.

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

Related to measuring the gender gap in social research, Bimber [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.

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, etc. 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.

Burger and others [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.

Understanding the demographics of twitter users [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.

Wagner and others [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. Arjona and others [RRGBD16] published in 2016 a survey of 2000 contributors where showed that the female participation would be around 2% to 5%. Izquierdo and others [IHSR18] revealed that few females contribute code or take political responsability in the OpenStack community. Recently, Zacchiroli [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. Vasilescu and others [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. Vasilescu and others [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.

[7] https://www.unwomen.org/

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

Related to the gender gap in science, Cassidy R. Sugimoto and colleagues [LNG+13] present a very good bibliometric analysis confirming that gender imbalances persist in research output worldwide. Holman and others [HSFH18] presented a code in R using genderize API and providing 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.

Liwicki and others [LSB11] presented a method inferring gender from hand written texts with a 67.5 % accuracy.

Gallagher and others [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.

Argamon and others [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, Koppel and others [KAS02] presented a document classification system with accuracy of approximately 80 %. Cheng and others [CCS11] exposes a feature selection and a model built using machine learning resulting in 85.1 % accurate rate for identifying gender from text.

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.

Liu and others [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.

Mueller and others [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

Ambekar and others [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 guided by Rodríguez Pérez and others [NRN21], in which presented NamPrism giving fresh ideas classifying races and being applied to massive software repositories.

Bollegala and others [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]

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, etc. The first name can be useful to infer another stuff such as race, ethnicity or culture, too.

[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

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 accepts that using name, gender and frequency is ok because there are people paying for or downloading a product. Typically, this is an acceptable assumption, although the consumer may purchase a bad product due to a good marketing strategy, a monopoly or there is a fraud, etc. Consumers may also trust in the government statistics regarding the economy, demography, or democracy. Therefor the people may trust the data for names, gender, and frequency. The Damegender[12] point of view is to trust in: the market tools and the official statistics.

Sometimes there are problems downloading official statistics, but there are people who have retrieved these data with web scraping. These files with another idea about truth.

Another problem arises when the government changes the data, sometimes communicating it to the users and other times not. This may be problematic for upgrades, but not with the truth, since changes can be traced.

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

3.2 Gender, language, nation and diversity

There are rules and exceptions in different languages to predict if a name is about male or female. For example, in Spanish or English, there are more names ending with 'a' classified as females than classified as males. However, Andrea is female in Spain and male in Italy. So, it is 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. Spanish is the main language in Spain. and in other countries such as Argentina, Mexico, Ecuador and Bolivia, Thus, it is helpful to know the language and nation a name or surname has originated from to help to detect gender.

Some countries, such as Spain, are providing free datasets for surnames but we need more efforts from many countries on this objective. However, Wikipedia and machine learning are working to relate names and surnames with ethnicity [AWM+09].

3.3 Damegender open datasets collection

Damegender[13] unified the different formats for name, gender, and frequency from statistical offices of the following countries: Argentina, Austria, Australia, Belgium, Canada, Switzerland, Germany, Denmark, Spain, Finland, France, Great Britain, Ireland, Iceland, Norway, New Zealand, Mexico, Portugal, Russia, Slovenia, Sweden, the United States of America and Uruguay. There are datasets as Italy, Brazil or China that has been found scraped from statistical institutions, we are evaluating the quality of these datasets and will be included in the future if it possible demonstrate the quality of these datasets.

It has been developed several python commands to collect datasets (downloadjson.py, get-wikidata-names.py, get-wikidata-surnames.py, orig2.py)

orig2.py calls a single sh script per country with statistical office using open data for names, gender and frequency and makes the preprocessing.

downloadjson.py is to retrieve data from external API sources (genderapi, genderize, namsor, nameapi, ...) from a csv file.

get-wikidata-names.py and get-wikidata-surnames.py is to retrieve names and surnames from wikidata giving the country. I have developed scripts to retrieve scientists from wikidata, so these scripts will give support to some ocupations in a near future.

[12] https://damegender.davidam.com
[13] https://github.com/davidam/damegender

Dataset	SSA	namdict	NLTK	Damegender
males	91.320	48.821	2.943	257.925
females	91.320	48.821	5.001	304.553

Table 1: Comparison of 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

The test datasets, wikidata sources and API sources could be unified with orig2.py in a single command, with a single interface, perhaps inspired in apt (debian command), or pip (python command).

It has been applied the criteria that one person usign a name is a gender vote (male or female)[14]. So, for each country has been choosen births or total of people using names in the country. Later, it has been merged these datasets building a free and international dataset.

Damegender uses surnames given by statistical institutions (Spain, Russia, the United States of America and Argentina in this moment). However, there are few statistical institutions that provide surnames and the lingüistic diversity is a good point, has been added surnames for all countries from Wikidata. Perhaps in the future Wikidata will become the best source of data for names and surnames, but now there are few elements.

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 with a scientific approach.

A possible criticism about this approach is the Leslie Problem[BM15]: the match between gender and name depends of the year. The solution to this problem is to introduce the age of the person in question. The most common use case states that the input is the name and the output must be gender, frequency and percentage. Therefore, a decision about gender must be made without data on the age or surname in most cases. This dataset was designed for the most of used use cases. 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 that have been classified by years. Therefore, this problem can be fixed with open data, too.

We have measured about the international DameGender dataset, using the dataset test made by Santamaria and Mihaljevic [SM18] reaching: accuracy (0.8756), precision (0.9638), recall (1.0) and f1-score (0.925). With other test datasets, similar and better results:

3.4 Free APIs for free datasets?

There area number of open data websites that allow the user to retrieve structured open data without cost. These sites include Wikipedia with SPARQL and OpenStreetMap with API rest, among others.

The open datasets for names, gender, and frequency are being modified once a year, at most 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 %"
}]
```

Therefore, it may be possible to have free API rest about names, gender, and frequency with reduced costs to fix the gender gap in a collaborative way similar to Wikipedia, OpenStreetMap or many free software projects.

[14]On the future, laws about non binary could be including other options, but only male and female was found as valid options in the open datasets when this article is being written

Dataset	Accuracy
Scientists Wikipedia	0.93
FIFA soccer	0.93
WTA tenis	0.91
National League	0.91
Baby Names New York	0.98
Conseil Garonne	0.97
P Sebo[Seb21]	0.88
Santamaria & Mihaljevic[SM18]	0.88
Average	**0.92**

Table 3: Accuracies using several test datasets about the DameGender international dataset

4 Measuring gender gap. GNU/Linux as use case

With a trust open dataset for names, gender, and frequency it is easy to measure the gender gap. Students and academics could measure the gender gap inexpensively and meet the fifth Sustainable Development Objective of the United Nations, which is to erase the gender gap completely.

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

It has been obtained the csv files using different methods to determine the names about the people in these communities.

In the Debian community all members must collaborate with a gpg key, so we can count males and females from the keyring. The keyring was imported with gpg commands and later the keyring was placed in a csv file.

GNU[15] and Linux[16] have collaborative websites for these projects. Therefore, making web scraping scripts has been downloaded the people and processed the people to csv files in this experiment.

In DameGender, has been developed csv2gender, a software with a csv file as input and deployed 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 a modern version of Damegender.

```
python3 csv2gender.py files/gnu.csv
 --first_name_position=0
 --title="GNU maintainers grouped by gender"
 --dataset="inter"
 --outcsv="files/gnu.gender.csv"
 --outimg="files/gnu.gender.png"

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"
 --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"
 --delete_duplicated
```

The inter dataset was created by merging several open datasets downloaded from official statistics sites from different nations, being a good representation of the Western World and the free software world is populating this world's area[GBRAIG08].

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

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

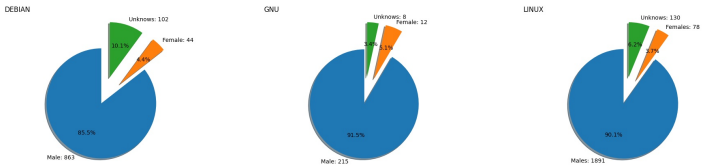

Figure 2: Males (blue), Females (orange) and Unknows (green) in Debian, GNU and Linux

Linux divides the developers in 1891 males (90.1%), 78 females (3.7%) and 130 unknowns (6.2%). 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 215 males (91.5%), 12 females (5.1%) and 8 unknowns (3.4%) Richard Stallman, the GNU founder returned to be president apologizing by his personal behaviour with the females.[17]

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 863 males (85.5%), 44 females (4.4%) and 102 unknowns (10.1%).

The researcher using Damegender finds the advantages of the Free Software for researchers. That is that they can to use, to read, to modify and to redistribute each step of trust in software or in data helping to the peer review, science reproducible, etc. Many researchers was trusting on US census, namdict (generally distributed in genderguesser), or commercial APIs. With the Damegender effort the researchers can obtain easily many new datasources extending the base of truth of the research works.

5 Conclusions and Future Works

Data feminism[DK20] is an area of growing interest. It has been explained the DameGender application, the motivations (reproducible research, fix gender gap to solve the United Nations objective, fields of application, including linguistic, social sciences, software engineering, natural language processing and journalism).

An improvement would be to build an international, universal, and free dataset of names, gender and frequency for the right design with the current state of the job, attending to the diversity (LGBT options, cultural minorities, etc.).

This article has explained the technologies involved in reducing costs related to studying the gender gap about gender gap (gender detection from the names, API rest, semantic web, etc.).

Augmenting the number of countries with statistical institutions that provide names, gender, and frequencies with open data involved addresing these data and giving more attention to 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. This data may provide an accurate real-world picture being a solution with a low number of unknowns.

Future works will address changes in the big software industry, improving user experiences in search engines, software repositories, and match sites due to detect gender from the name.

Acknowledgments

We would like to thank: the statistical institutions by release of the open datasets about names, gender and frequency. Luz Galvis for 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.

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

[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.

[AZ09] Abdul Wahed Qasem Ghaleb Al-Zumor. A socio-cultural and linguistic analysis of yemeni arabic personal names. *GEMA: Online Journal of Language Studies*, 9(2):15–27, 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.

[CPSZ22] Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. *World inequality report 2022*. Harvard University Press, 2022.

[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.

[DK20] Catherine D'ignazio and Lauren F Klein. *Data feminism*. MIT press, 2020.

[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.

[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.

[LNG+13] Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R Sugimoto. Bibliometrics: Global gender disparities in science. *Nature*, 504(7479):211–213, 2013.

[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.

[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.