Women in Neuroscience: A Neuronal Representation

Mio Akasako

Parsons School of Design
Independent Study, Fall 2019
Advisor: Katayoun Chamany

Introduction

Women in Neuroscience: Contemporary Environment

It has been known for decades that women have a significant disadvantage in STEM academia, a historically male dominated environment. The National Academy of Sciences reported in 2018 that the rate of sexual harassment in STEM is second only to that of the miliatry.[1] Representation of women in STEM is to this day, is still meager. Women make up half of the college-educated workforce in the US, but only 29% of the science and engineering workforce.[2] If we probe even further, we find that a mere 15% of engineers, and 25% of computer scientists are women. In STEM, there exists a unique condition where there is a staggering lack of female presence, let alone female empowerment. In the absence of such oversight, support network, and/or mentorship, sexism becomes normalized, and toxic environments are left to bloom.

Academic institutions do little to help the issue. Sexual misconduct, harassment, and discrimination are extremely common in STEM academia, and more often than not fail to be addressed properly. Around 45% of publicly reported cases of sexual misconduct end with little to no action by the institution, and/or with a voluntary resignation or retirement from the perpetrator.[3] Frequently, even when the institutions do suspend the perpetrators, it is under paid leave. Some lawsuits are settled by the institution on behalf of the perpetrator. Because of such harsh conditions,, retention rates of women in STEM academia are low. Female faculty in the sciences are consistently lower than that of social sciences or humanities. As one moves up the academic ladder, there is an obvious drop in the number of women within the field. Though the percentages of women majoring in the sciences at the undergraduate level can go as high as 50%, the percentage of female tenured/tenure-track professors in the sciences has never gone above 22% (in NYU or Columbia). [4]

Yet, women have persisted in the field, and despite the setbacks, have slowly gained ground towards more representation in STEM academia. I wanted to depict this progress with data. Perhaps selfishly, as I myself have a background in neuroscience and work in the field, I wanted to narrow my focus to neuroscience. With this project, I aimed to

---

[1] *Sexual Harassment of Women: Climate, Culture, and Consequences in Academic Sciences, Engineering, and Medicine*(Washington D.C.: National Academies Press, 2018).

[2] *Science & Engineering Indicators 2016. National Science Board*(Place of Publication Not Identified: Distributed by ERIC Clearinghouse, 2016).

[3] Geocognition Research Laboratory, "The Academic Sexual Misconduct and Violations of Relationship Policies Database," GEOCOGNITION RESEARCH LABORATORY, September 15, 2018, , accessed March 26, 2019, https://geocognitionresearchlaboratory.com/2018/08/20/the-academic-sexual-misconduct-database/.

[4] Carol Shoshkes Reiss and Andre Fenton, Faculty of Arts and Sciences Equity Committee Executive Summary of Data to End of 17/18 Academic Year and Recommendations, Report, Faculty of Arts and Sciences Equity Committee, New York University.

visualize the trends in female representation in academic neuroscience by using journal publishing as a metric for success and advancement in the field.

Humanizing Data & Designing for Play

Research exploring the gender gap in the workforce, and even STEM, is not novel. The Wall Street Journal recently published a piece on the wage gap that reports that out of 446 major occupations in the US, women still earn less than men in 439 of them.[5] In 2018, Luke Holman published a very comprehensive paper on the gender gap in STEM academia, noting that the gender gap in journal publishing will likely persist for generations.[6] However, in all of these pieces, the data is conveyed in a conventional method using statistical graphs and charts, and rightly so, as rigorous statistics was intended. I wanted to take a different approach--to showcase this data in a more relatable way. I took inspiration from Giorgia Lupi, an information designer who advocates for data humanism--a practice to "humanize" data in order to reveal the more qualitative aspects of it that we as people can relate to more. She "challenges the impersonality that data communicate, designing engaging visual narratives that re-connect numbers to what they stand for: stories, people, ideas".[7] The data I am working with is inherently human--it is about the progress that women have achieved and continue to achieve, and therefore is attached intimately to human emotions and desires. Though mapping this in graphs gets the point across, I wanted my depiction of the data to be more playful, energizing, and celebratory of women in neuroscience.

Methods

Data Collection

Initially, I had several datasets that I started collecting for the purposes of this project. My intention was to use each one to map a different aspect of representation of women in neuroscience. I used the Springer Nature API to collect the metadata of ~300,000 journal articles published between 1868 and 2019 by Springer, under the subject "Neurosciences".[8] I also obtained the Neurotree database from the admins in hopes of mapping networks among female neuroscientists across time.[9] Neurotree is a volunteer-run website & database that keeps track of the neuroscience academic genealogy. Information is collected on connections amongst PIs, post docs, and grad students within the field. For each person, we can obtain publication data, mentors,

---

[5] WSJ.com News Graphics. 2010. "What's Your Pay Gap?" WSJ. 2010. https://graphics.wsj.com/gender-pay-gap/.

[6] Holman, Luke, Devi Stuart-Fox, and Cindy E. Hauser. 2018. "The Gender Gap in Science: How Long until Women Are Equally Represented?" Edited by Cassidy Sugimoto. *PLOS Biology* 16 (4): e2004956. https://doi.org/10.1371/journal.pbio.2004956.

[7] giorgialupi, ABOUT. 2014. "Giorgialupi." Giorgialupi. 2014. http://giorgialupi.com/about.

[8] "Springer API." 2019. Springernature.Com. 2019. https://dev.springernature.com/.

[9] "Neurotree." 2019. Neurotree.Org. 2019. https://neurotree.org/neurotree/.

mentees, and affiliated institutions. Although by no means comprehensive, the database is composed of a total of ~700,000 data points (people in the database). I was also exploring the possibility of highlighting female neuroscientists in the past who have not been given the recognition they deserve. Some names I researched include Maria Manasseina, a Russian pioneer in the study of sleep, Thanjavur Santhanakrishna Kanaka, Asia's first female neurosurgeon, and Lina Solomonova Stern, an early contributor to blood brain barrier research.[10]

It soon became clear that I needed to focus on narrowing the breadth of data I was to work with, rather than trying to use everything that I had. Due to time constraints, I decided to focus on journal publication data. However, instead of using the Springer Nature data, I decided to pull the metadata from MEDLINE, a U.S. National Library of Medicine bibliographic database that contains more than 25 million references to journal articles in life sciences with a concentration on biomedicine.[11] I switched datasets because I was made aware that the Springer Nature database only contains literature published by Springer, and moreover, that many investigators have boycotted it. I used a package called RISmed in R to pull metadata from all neuroscience related journals from MEDLINE. The categories I pulled for each paper are as follows:

- Year
- Abstract
- Title
- Affiliation
- Country
- Journal
- Author (Forename, Last Name, Order, Gender)
- MeSH Terms

In total, I extracted the metadata of around ~1,000,000 papers published between 1945 and 2019.

Data Processing, Cleaning, and Organization

Now that I had a dataset to work with, I needed to tailor it to my needs. Though I had the names of the authors, I had to acquire the probable gender associated with those names. To do this, I ran all of the forenames through the Genderize API, an API that gives the probability that a name is male or female.[12] In order to do this, I used Python. I cleaned all of the forenames in my dataset by taking out the names that only had initials, and taking

---

[10] "WiNEu – European Women in Neuroscience." 2018. Wineurope.Eu. 2018. http://wineurope.eu/.

[11] "About MEDLINE® and PubMed®: The Resources Guide." 2019. Nih.Gov. U.S. National Library of Medicine. 2019. https://www.nlm.nih.gov/bsd/pmresources.html.

[12] Genderize.io. 2019. "Genderize.Io | Determine the Gender of a Name." Genderize.Io. 2019. https://genderize.io/.

out middle initials if there were any attached. Then I called the API over a few days in order to get the gender and the gender probability of each name in my dataset. To determine an acceptable probability for a name to be assigned a gender, I used a version of poor man's bootstrapping--I hand-scored 1000 random names in the dataset by hand in groups of 100, and looked up the probabilities of those that I got wrong. I took those probabilities and determined the average probabilities and standard deviation of "wrongly determined gender names" from each set of 100. I then calculated the probability at which I am 95% confident that I will guess the gender of the name correctly . This came out to be 93%. I used this probability to assign genders to the names in my dataset. If they were below 93% probability, or if the names were initials, I assigned them the label "uncertain".

After obtaining the gender data, I created several different data tables in order to accomodate my analysis: one for general metadata, one for author gender data, and one for MeSH terms (terms used to categorize the topic covered in the paper). All have PMID numbers as their keys. The information is separated into tables as follows:

Data Table (with example)

pmid	abstract	year	title	journal	country
31722223	BACKGROUND: Zebrafish are a popular animal model for investigations into the neurobiological mechanisms of learning...	2019	Classical Fear Conditioning in Zebrafish (Danio rerio) Using an Upgraded Version of the Goldfish Conditioned Withdrawal Preparation.	Journal of neuroscience methods	Netherlands

Authors Table (with example)

pmid	ForeName	LastName	Initials	order	Affiliation	gender
25588974	Fernanda Cristina Rueda	Lopez	FC	1.0	Federal University of Rio de	female

					Janeiro, Radiology	

Mesh Table (with example)

pmid	mesh
30412317	DNA Mutational Analysis

[Reference: gender_extraction.ipynb, medline_data.ipynb]

Data Analysis

Some examples of questions I wanted to answer with my dataset are the following:

- How has the proportion of female contributors to neuroscience changed over time?
- How has the proportion of female last/first authors changed over time?
- How has the focus of neuroscience changed over time? In what fields/areas are women more prevalent?
- Do women thrive more under a female PI? How many of those relationships exist (where both first and last authors are women)?
- Are there differences in the proportions of women publishing in each country?

Although there were many more questions I wanted to explore, these were the main ones I wanted to answer first. All of the data analysis was done in Python, and required some going back and forth between all of the tables. Though in the end, I was unable to get to some of the proposed questions (ie women publishing in each country/what subfields are women prevalent in), I was able to obtain sufficient data for visualization.

Gender Ratios

To calculate the overall gender ratios of journal contributors over time, I obtained the PMIDs of all of the papers published each year, then obtained the gender counts of journal authors who published each year. To calculate first author ratios, I looked at the gender assignment of all authors with 'order' == 1.0 and quantified female counts vs male counts for each year. To calculate the percentage of female last authors per year, I reordered PMID entries by descending order of 'order', then kept just the first entries for each PMID (the 'order' category goes from '1.0' to however many authors there are, so the last number is not the same for each PMID). To calculate the percentage of female single authors per year, I created dataframes for all first authors and last authors for a given year. I determined whether a paper had only a single author by calculating if the first author 'order' was the same as the last author 'order'. If so, I counted it towards a paper with a single author. I then determined the gender associated with the authors of these papers, and got the ratio of female single authors vs total single authors for every year. Similarly, to calculate the papers where both first and last authors were female, I calculated the number of papers that had multiple authors, whose genders matched, for every year. Then, I determined whether those genders were female to get the percentage of papers which had female first and last authors.

These ratios should be taken with a grain of salt, especially those from earlier years where there are not enough data points to offset the number of "uncertain" gender assignments. There seems to have been a general rule where forenames were given as initials rather than full names until the 90s. Ratios from later years are more accurate, given that full names were provided, and that there are enough data points for both male and female assigned names to counter the uncertain assignments.

Though this did not go in the current visualization, I also started analyzing the papers whose first and last authors were both female. I wanted to determine whether these women who were mentored by female principal investigators would also go on to mentor other women. I took the PMIDs of the mentioned papers and ordered the results by forename, then last name to see if the same names would pop up multiple times with different "order" values.

General Data Points

I determined the total number of papers published in the neuroscience field and the total number of countries that these authors were affiliated with per year. To do this, I counted the number of unique PMIDs and unique countries in the dataset for every given year.

Trending Neuroscience Topics

I wanted to explore what topics were trending in the neuroscience research world over each year, and how they changed over time. I mainly used the MeSH dataset for this. I aggregated a list of all MeSH terms used in the dataset, and for each term, went through and counted how many times the term was used in each year, and stored all of the values in a dataframe. I then took the average number of times each term was used during the span of the dataset (1945-2019), and calculated the ratio of the number of times each term was used every year over the average. The higher the ratio, the more times a given term was used in that year compared to other years. I took out any term that had less than 100 papers associated with it (to take out the effects of "one paper wonders"), and then determined the terms with the highest ratios for each year. I used the top 10 from each year for my final visualization.

[Reference: medline_analysis.ipynb]

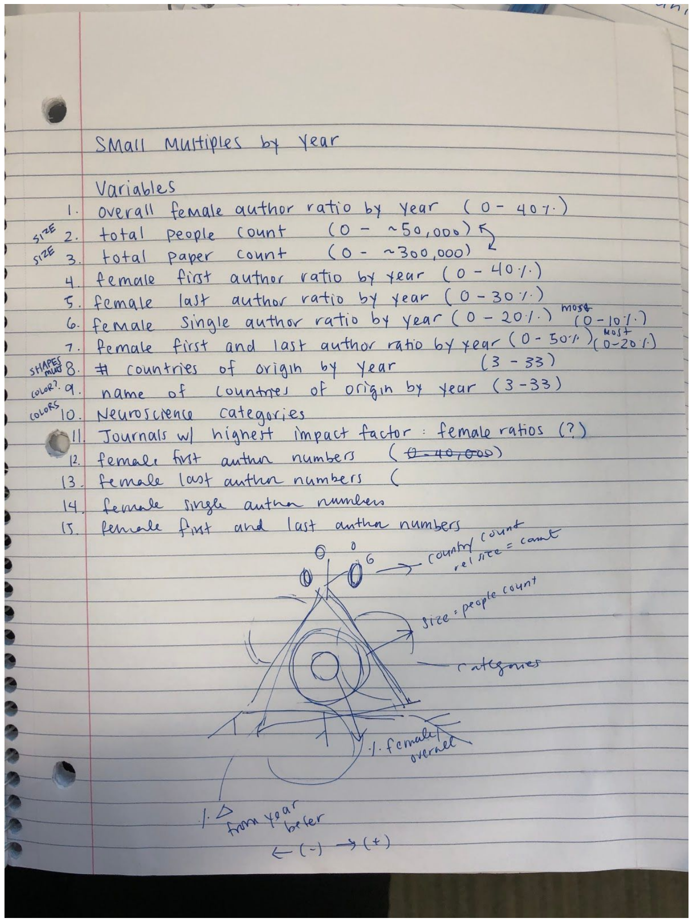
Data Modeling and Design Methods

Ultimately, I wanted to produce a visualization that was more visually friendly rather than technical, yet communicated important data information--even if the output was creative, everything needed to be data-driven and have some meaning behind it. I was inspired by Ingrid Ching's small multiple abstract paintings, in which she painted abstract shapes with watercolor.[13] I decided to create my visualizations in a similar manner, using small multiples, watercolors, and abstract shapes. In my case, these abstract shapes put together would form an abstract neuron. Each shape would be associated with a variable from my data analysis. In order to convert the results of my analysis into a visualization, I used P5.js to draw shapes from the values. The main circle (soma) signifies the overall percentage of female authors per year, by size. The main triangle (cell body) signifies the overall paper count per year, by size and color. The lines and arrows (axons and dendrites) signify various ratios of female authorship by color: first authors, last authors, single authors, and multiple authors (first and last authors both female). The length of the lines represent the total count of people in each group, while the arrows at the bottom of the lines represent the percentage of women authors in a given category. The background of each small multiple is a word cloud of the trending neuroscience terms for each year, categorized by color. I visualized the values for every 5 years.

Fig. Ingrid Ching - small multiples

[13] "Ingrid and Ching Originals | The 'O' Series." 2019. Ingrid and Ching. 2019. https://ingridandching.com/collections/theoseries.

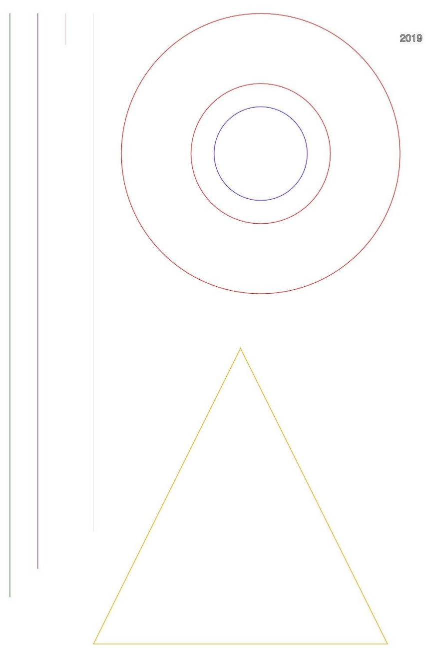
Here is a sketch of what I was envisioning, and an example of a P5 sketch of values:



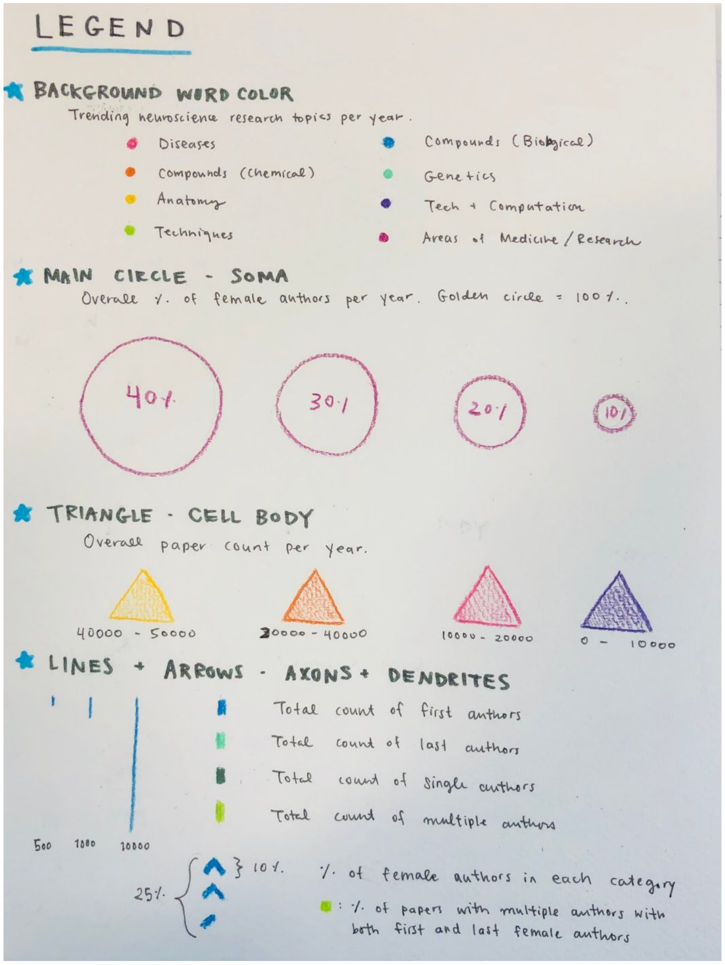
SMall Multiples by Year

Variables

1. overall female author ratio by year (0 - 40%)
2. total people count (0 - ~50,000) — SIZE
3. total paper count (0 - ~300,000) — SIZE
4. female first author ratio by year (0 - 40%)
5. female last author ratio by year (0 - 30%)
6. female single author ratio by year (0 - 20%) most (0 - 10%)
7. female first and last author ratio by year (0 - 50%) most (0 - 20%)
8. # countries of origin by year (3 - 33) — SHAPES
9. name of countries of origin by year (3 - 33) — COLOR?
10. Neuroscience categories — COLORS
11. Journals w/ highest impact factor : female ratios (?)
12. female first author numbers (~~0 - 40,000~~)
13. female last author numbers (
14. female single author numbers
15. female first and last author numbers



Here is a legend of the visualization for reference:



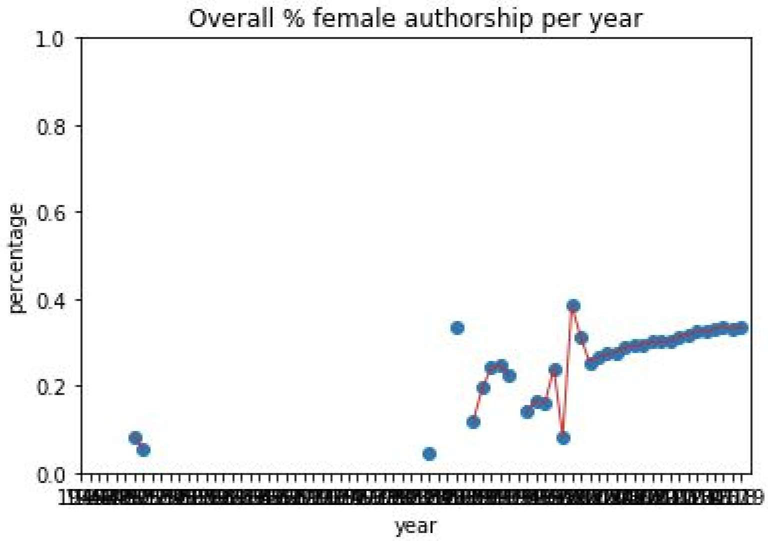
[Reference: sketch.js]

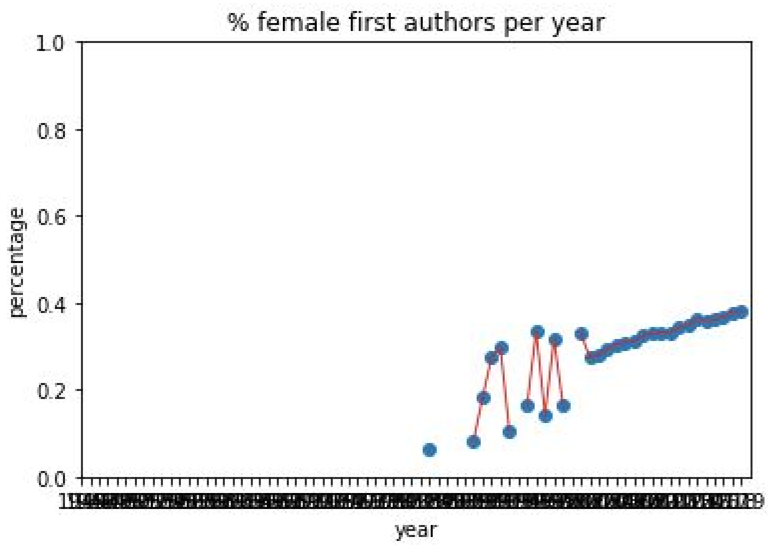
Results

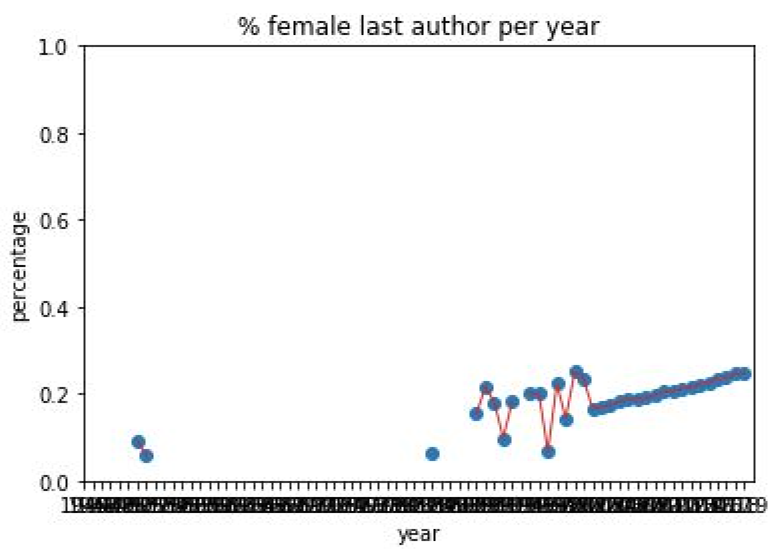
General Results

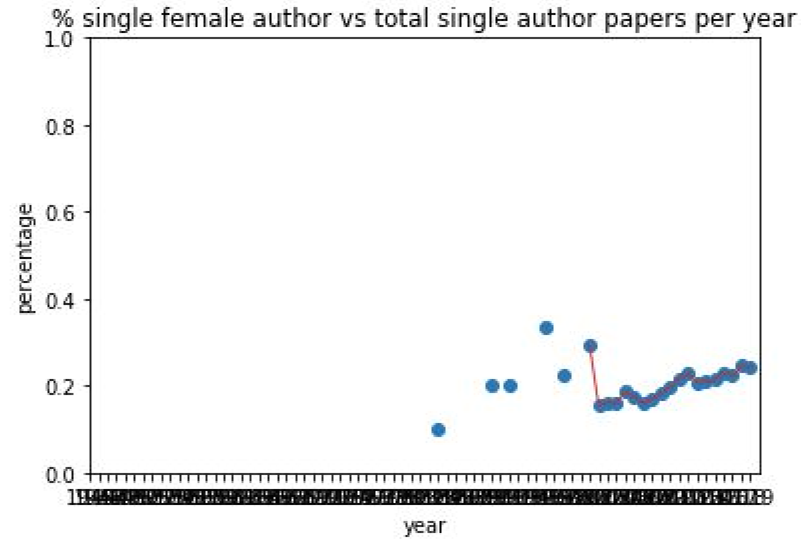
Unsurprisingly, the gender gap in publishing still exists in academic neuroscience. However, both overall and within each authorship category, the percentages of women contributors are going up. The overall gender ratio of papers written by women is around 33% in 2019 compared to 10-20% in the 90s. The percentage of female first authors has risen steadily and is at 38% in 2019 . The percentage of female last authors are lower, as expected, peaking at 25%. Papers with single female authors have steadily increased in percentage and is at 24% in 2019. Papers with both first and last female authors peak at around 13% in 2019. The first identifiable papers with both first and last female authors came out in 1990 from France in The European Journal of Neuroscience (all five of them).

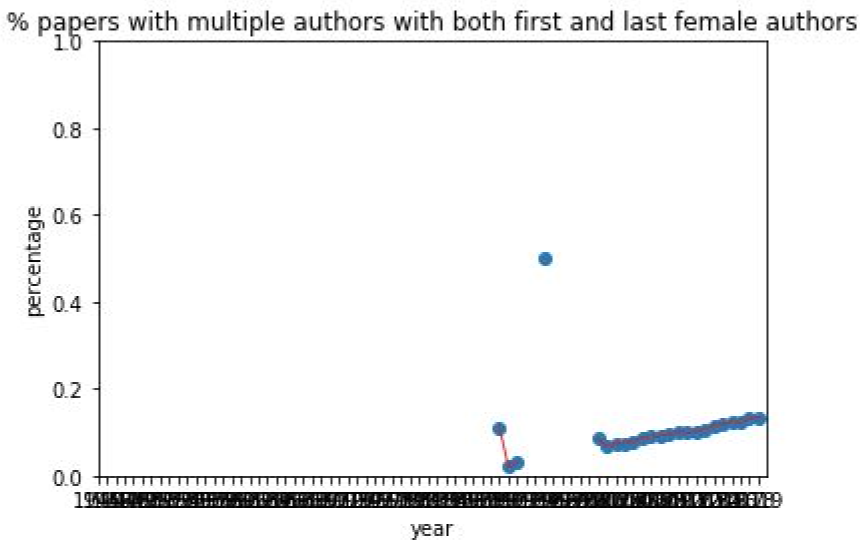
It was frustrating to have so many uncertain gender assignments, especially for the earlier years, which have resulted in some inaccurate percentages that display as outliers. Some values seem to be following a linear relationship while others are not--however, overall, the values are improving over time.


Overall % female authorship per year
1.0
0.8
0.6
0.4
0.2
0.0
percentage
year


% female first authors per year
1.0
0.8
0.6
0.4
0.2
0.0
percentage
year


% female last author per year
1.0
0.8
0.6
0.4
0.2
0.0
percentage
year





Researching neuroscience trends gave an interesting output--the terms very much reflected the events that were occurring during the time period. For example, “military medicine” was trending in 1946, as well as “zika virus” in 2016 and 2017. Moreover, trends in discovery were reflected as well--for example, “neuronal tract-tracers” in 2009, and “machine learning” in 2018. This in a way validated my methodology for finding the terms most relevant for each year in the academic neuroscience field.

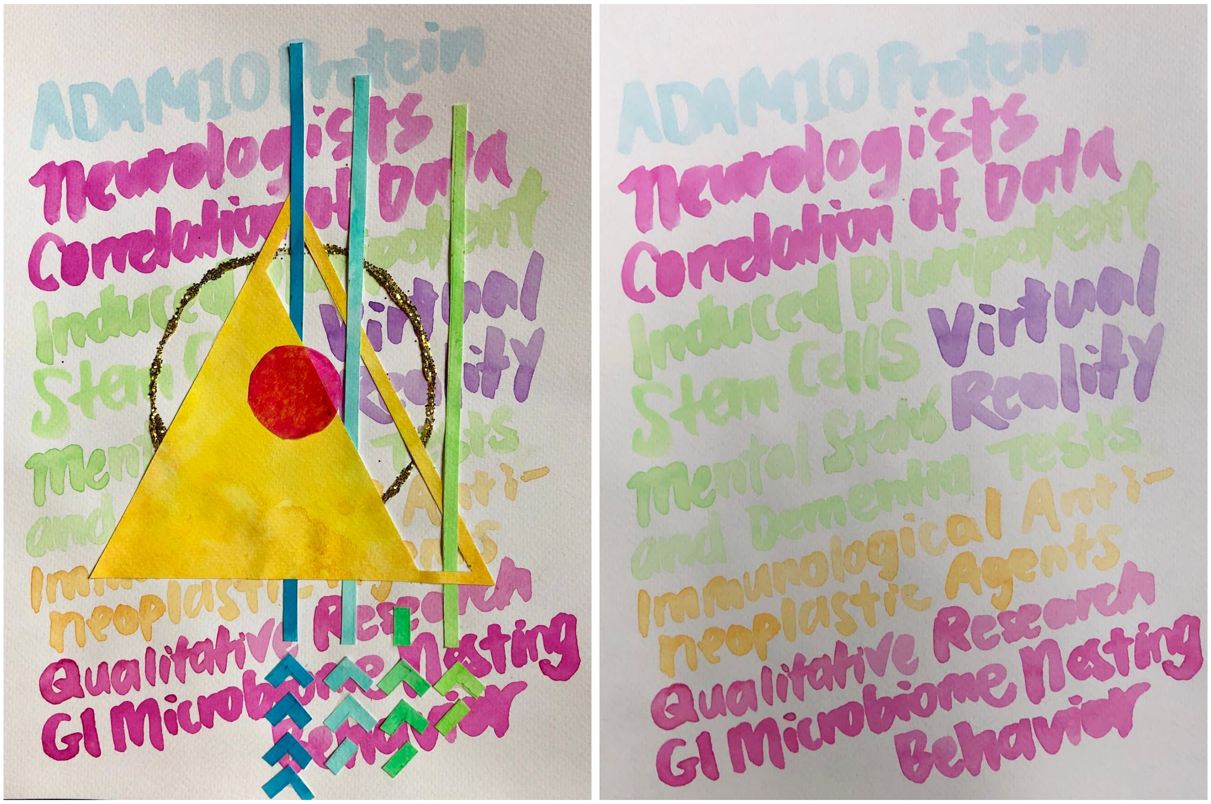


Fig. Examples of 2019 - with abstract neuron and just background trending terms

In addition to within Google Drive and GitHub, you can find my final images hosted here:

https://neuronal.fun

Discussion

In general, I was not incredibly surprised by the data I obtained--I was aware of the general ratios of female authorship in STEM academia, and those in neuroscience do not stray too much from that of the whole. It is to be expected that female PIs are not as prevalent as female grad students, or postdocs (higher percentage of first author females than last author females). However, there is potential to take this exploration further in more interesting directions. In addition to answering the questions I did not have time to answer in the first pass (countries where women publish more, how gender ratios in journals with high impact factors compare to the average), I would like to explore more in depth the networks amongst women who publish in neuroscience. I have already come up with the initial dataframe for papers with both first and last female authors. I would like to map out the connections between these women over time, and create a network map of women who have facilitated other women in the research process. I think this would be another great way to highlight women neuroscientists who have encouraged other women to pursue the same path. I would also like to fit the various author ratios to models that predict when, if at any time point, the publishing gender gap will be closed in neuroscience. This should be fairly doable by trying different models using scikit-learn.

I would also like to further humanize the design by adding stories and names of significant women who contributed to the field each year, whether it is through adding interactive elements to each image online, or writing little notes on each image. Though right now it is hosted online as a simple slideshow, I would like to create a web platform that showcases all of the images as small multiples, where a curious user could hover over each image and get actual numbers associated with the data.

References

"About MEDLINE® and PubMed®: The Resources Guide." 2019. Nih.Gov. U.S. National Library of Medicine. 2019. https://www.nlm.nih.gov/bsd/pmresources.html.

Geocognition Research Laboratory. "The Academic Sexual Misconduct and Violations of Relationship Policies Database." GEOCOGNITION RESEARCH LABORATORY. September 15, 2018. Accessed March 26, 2019. https://geocognitionresearchlaboratory.com/2018/08/20/the-academic-sexual-misconduct-database/.

Genderize.io. 2019. "Genderize.Io | Determine the Gender of a Name." Genderize.Io. 2019. https://genderize.io/.

giorgialupi, ABOUT. 2014. "Giorgialupi." Giorgialupi. 2014. http://giorgialupi.com/about.

Holman, Luke, Devi Stuart-Fox, and Cindy E. Hauser. 2018. "The Gender Gap in Science: How Long until Women Are Equally Represented?" Edited by Cassidy Sugimoto. *PLOS Biology* 16 (4): e2004956. https://doi.org/10.1371/journal.pbio.2004956.

"Ingrid and Ching Originals | The 'O' Series." 2019. Ingrid and Ching. 2019. https://ingridandching.com/collections/theoseries.

Johnson, Paula A., Sheila E. Widnall, and Frazier F. Benya, eds. 2018. *Sexual Harassment of Women*. Washington, D.C.: National Academies Press. https://doi.org/10.17226/24994.

"Neurotree." 2019. Neurotree.Org. 2019. https://neurotree.org/neurotree/.

Science & Engineering Indicators 2016. National Science Board. Place of Publication Not Identified: Distributed by ERIC Clearinghouse, 2016.

Sexual Harassment of Women: Climate, Culture, and Consequences in Academic Sciences, Engineering, and Medicine. Washington D.C.: National Academies Press, 2018.

Shoshkes Reiss, Carol, and Andre Fenton. Faculty of Arts and Sciences Equity Committee Executive Summary of Data to End of 17/18 Academic Year and Recommendations. Report. Faculty of Arts and Sciences Equity Committee, New York University.

"Springer API." 2019. Springernature.Com. 2019. https://dev.springernature.com/.

"WiNEu – European Women in Neuroscience." 2018. Wineurope.Eu. 2018. http://wineurope.eu/.

WSJ.com News Graphics. 2010. "What's Your Pay Gap?" WSJ. 2010. https://graphics.wsj.com/gender-pay-gap/.