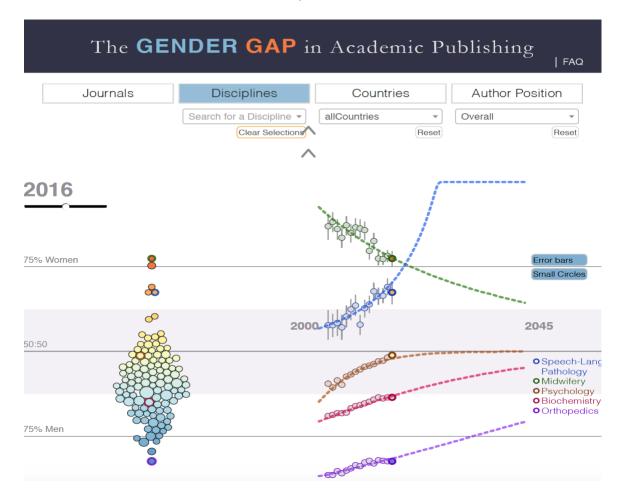# Data Visualization Critique on "The Gender Gap in Academic Publishing"

Siyi Fan

https://lukeholman.github.io/genderGap/

This visualization allows readers to view the past, present, and estimated future gender ratio of authors on academic publications listed on PubMed in different fields of Science, Technology, Engineering, Mathematics, and Medicine (STEMM). The readers could select a specific discipline, country, and author position they are interested in by clicking on the four buttons at the top of the page, which represent the subgroups of all data in journal, research discipline, the author's country of affiliation, and position in the author list. The left-hand plot displays the estimated author gender ratio for each subset of the data in a

certain year, which is controlled by the slider. The gender ratio was estimated by fitting a data. Clicking on a data point on the left plot will bring up a right-hand curve showing estimate of the past, present, and future gender ratio, as well as the author gender ratio and its 95% confidence limits, which is controlled by an "error bars" button that could be toggled on or off. The "Small Circles" button could adjust the size of the data point. Hovering the mouse cursor over a data point will present the sample size in terms of the number of male and female authors, and the number of papers.

This interactive gender gap plot is one of the most iconic and inspiring visualizations I have ever seen. First, it is relatively truthful. The study began by downloading the author list for every single publication on the PubMed and arXiv databases. They then used the genderize.io database to assign genders to authors based on their given names and, where possible, determined the country in which each author was based from their academic affiliation. For the PubMed data, each journal was assigned to an academic discipline, using PubMed's own classifications where possible. For each journal, discipline, and arXiv category, they fit a model to the data to estimate the present-day author gender ratio, its rate of change, and when or if gender parity will be reached. The complete dataset is archived as a SQLite3 database at the Open Science Framework (https://osf.io/bt9ya/) and scripts used to collect and analyze the data could be archived at https://github.com/lukeholman/genderGapCode. However, there is one limitation on the data collection. Some authors in the PubMed dataset used initials rather than their full name, so there could be a gender difference in the tendency to publish under one's initials. In addition, gender assignment is doubtful for authors with affiliations in some East Asian

countries, because of a high frequency of given names that are commonly used by both men and women. Therefore, this visualization is based on an honest and relatively thorough research.

Secondly, this visualization is functional. In this plot, the readers could not only view the past and present data, but also explore the future tendency. The interactive visualization brings the main results to the readers, while at the same time, helps readers interact with the dataset to uncover the findings themselves. The most functional part to me is that even if I select a specific year to visualize (left-handed plot), I could still view a tendency of the change from 2000 to 2045 under my selected subgroup shown on the right hand. Additionally, the lines of 50%, 75% women, and 75% Men present readers a straightforward visualization of the data distribution. Take the psychology in year 2010 as an example. On the left distribution plot, we can see psychology is below the 50% line (45.5% female). Among all disciplines, it has a relatively high female ratio. To be note, there are only 11 disciplines passing over the 50% line, with highest 85.5% for Midwifery. On the right tendency plot, we can see the female ratio in psychology is increasing since 2000. Till 2016, there were 13615 women and 14296 men (48.8%) publishing in the academic journals. Based on the prediction, this ratio will still go up and reach 50% around year 2025. For the comparison, we could also click the highest data point on the left – Midwifery. The tendency of female ratio in this discipline goes down dramatically from 2000 to 2016 and is estimated to reach 65% in 2045, which would still be higher than psychology. Thus, this interactive visualization presents multiple levels of information

simultaneously, providing depiction of the data and their change in time in a functional way.

Third, this visualization is beautiful. It is informational and clear to both intended scientists and general public. The dots represent different data points, from red to blue representing female-dominated to male-dominated. It is also very lovely (cute!) when all data points piling up and bouncing around. The colored lines and corresponding notes clearly show the trends of gender ratio of authors on academic publications. However, I think it would be more neat if the designers separate information into different areas or boxes by adding some frames. For example, the area1/box1 contains buttons for different subgroups; the area2/box2 includes year slider, button error bars, and button small circles; the last area/box covers all main information, including data points and colored trend lines.

Fourth, this visualization is insightful. When I first saw this interactive plot, I experienced a eureka moment ("wow"). I have never thought of a huge gender gap in academic publishing for most STEMM disciplines. Compared to other countries, this gender gap in China is more significant, which made me shocked. Moreover, interactive visualizations often lead to knowledge-building insight. Following the instance I mentioned before, it seems like the prediction for all disciplines will reach 50% around year 2025. However, if you click the discipline speech-language pathology, then you will be struck to see a boost in its change, which could possibly reach 100% in 2025, opposite to the other disciplines. This will probably shift from one gender gap (men>women) to the other gender gap (women>men). The promotion of gender diversity is a matter of national and global urgency. This interactive visualization will be of great help with that.

Finally, this visualization is enlightening and inspiring. It implies for most disciplines, males have more publications than females and the good news is that this gender gap is shrinking. Importantly, the aggregated changes in different subgroups could inspire readers to think further about the interactions of those subsets and design interventions to produce a diverse academic environment.

To sum up, this interactive visualization is generally truthful, functional, intriguing, insightful, and changes our minds for the better.