A public exploratory data analysis of gender bias in teaching evaluations

Benjamin M. Schmidt

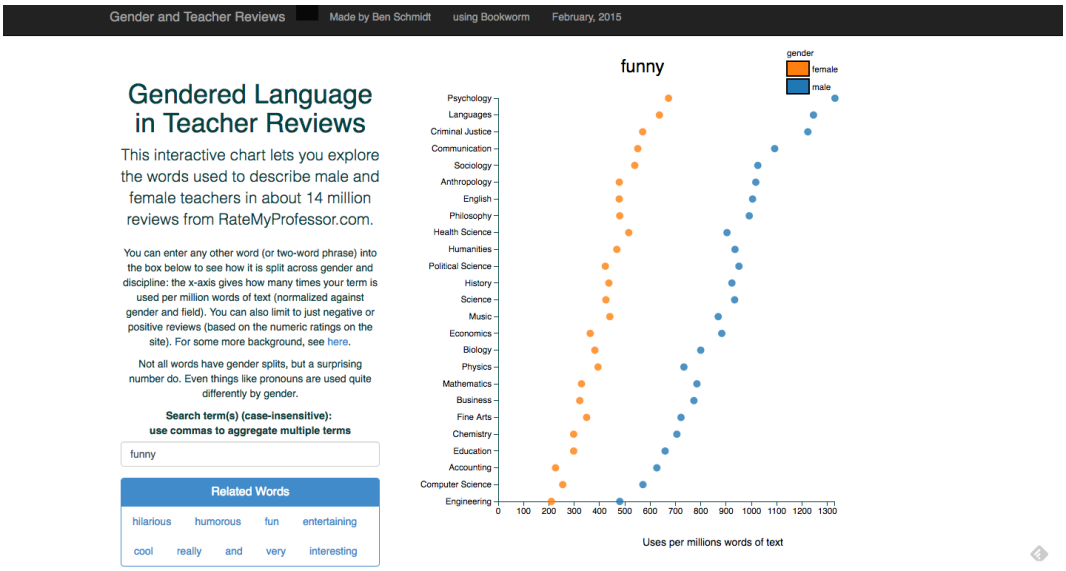

Fig. 1. Website at benschmidt.org/profGender. Users enter terms in the box at left: circles update at right to show the usage of the entered word or term.

Abstract—Humanists frequently want to engage in exploratory data analysis without the stage of confirmatory analysis that, in statistical practice, it traditionally precedes. This paper explores the possibilities and limits of purely exploratory data analysis through a case study in visualizing a a corpus of approximately 14 million teacher evaluations from the website "RateMyProfessors.com." The visualization provides an exploratory interface that, while problematic as a vehicle for causal inference, allows users to explore a rich data set in the light of their own experience at a granular level and promotes a high level of user engagement. Experiments on public traffic show that user engagement seems to be relatively fixed in the face of improvements or detriments to the user experience; interactive exploratory analysis can play a useful role in validating individual experience. On the other hand, there is some reason to fear that exploration ceases when users' presumptions are not met. Embedded interactive narratives may be a better solution for exploring data sets like this than purely exploratory views lacking any interpretive framework.

Index Terms—Gender bias, teaching evaluations, exploratory data analysis, Digital Humanities, Text Visualization

1 Introduction

1.1 Exploratory data analysis without confirmation

John Tukey's concept of "exploratory data analysis" (EDA) appears again and again, with or without attribution, in the apparatus of the digital humanities. It has a sort of natural affinity for humanities scholars. While statistics proper can carry an unfortunate connotation of positivism that discomforts scholars in hermeneutic disciplines, "exploration" is a virtue shared across the academy. Visualization, as the leading tool of exploratory data analysis, is the data analysis tool that finds far and away the least resistance in humanities departments.

This is a fortunate occurrence, but humanists and statisticians alike should be careful not to assume too easy a rapprochement between statistical and humanistic practice. For Tukey and his successors EDA was a complement to confirmatory data analysis (CDA) rather than an alternative, and perhaps as likely to refute established models as to generate new ones. [7] [1] Exploration in the humanities, on the other hand, often serves as almost an end in itself; particularly in the well established public history tradition of digital humanities, a commitment to openness and sharing can have a fraught relationship to attempts to proof or "argument-driven scholarship." [2]

One of the more interesting possibilities for intersection between the digital humanities and data visualization, then, is the degree to which humanities practices may offer an approach to visualization that does not presume any attempt to ultimately end in confirmation or argument. Digital humanities data visualizations may be more radically exploratory because their primary end can be simply to make an archive accessible; in so doing, it is possible that they can help refine a point of view in which exploratory data analysis rests an end in itself for public consumption, rather being subordinated to a search for truth.

In this environment, "exploration" in the digital humanities may be something significantly different than exploration that ultimately leads to confirmatory data analysis. It may be an attempt to represent the past "as it really was," in Leopold von Ranke's 19th century formulation–a

• *Benjamin M. Schmidt is Assistant Professor of History at Northeastern University and core faculty in the NuLab for Texts, Maps, and Networks. E-mail: b.schmidt@northeastern.edu*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

call recognized in the field as problematic, but nonetheless lurking, perhaps, in the continuing attempts to build virtual worlds, pedagogical tool, and elaborate models of physical historical spaces. Or it may, following the injunctions of Johanna Drucker [6] calls to humanities data analysis, be more resolutely perspectival or personal. In either case, visualization may be straightforward but targeted at making a public *experience* rather than helping an individual researcher to *understand* a cultural phenomenon.

This tension presents a potentially useful area of conversation between practitioners of information visualization and humanities scholarship. A fully public exploratory data analysis not geared towards confirmatory models might provide a useful bevy of techniques towards heightening a rift between EDA and CDA that statisticians might want to paper over. On the other hand, exercises in data visualization unmoored from confirmatory data analysis may represent in some ways the worst of both worlds from the humanities and the sciences. Without the hard check of statistical proof, they can overwhelm with the weight of their size while not providing any new evidence.

This paper investigates this tension through a case study; the author's visualization of gender bias in 14 million teaching evaluations from the website RateMyProfessors.com. (URL: benschmidt.org/profGender.) As detailed below, the underlying data presents severe enough strains on straightforward proof that inferential statistics are generally unwarranted. Instead, I chose to adopt a presentation-only approach drawing on a tradition of public digital humanities in which a single hermeneutic framework is provided to the user–that he or she will explore the differing rates of usage of words by gender of the teacher and by discipline of the course–and no strict inferences are made.

This paper describes those design choices. At the same time, the visualization has been widely shared since it was released in February 2015. I thus also explore here the ways the visualization has been *used* through analysis of logfiles and some A-B testing. I conclude that there is both substantial public interest and real social benefit in exposing archives through data visualization and hermeneutical approaches alone; but that even when proof through confirmatory data analysis is an inappropriate use of a data set, humanists bear some responsibility to guide users through data. In the conclusion, I suggest the pure interactive exploration may not be the best option for sharing data, and that humanists might better consider joining it with a tradition shared by information visualization's journalistic wing and the native traditions of the humanities: narrative presentation of evidence.

2 Exploring gender bias in teaching evaluations through textual data

A variety of recent studies have argued that teaching evaluations show substantial gender bias against female teachers ([9]; [10]; [4]; [11]). These studies use a variety of methods through experiments or natural experiments to control for differences in ratings given to men and women in similar teaching situations. Insofar as it is possible to prove that ratings are intrinsically biased or not, such methods present the soundest means to do so.

Bias, though, is not generally *experienced* as a gap of some number of tenths of a point that one's gender deducts from a Lickert scale. While that gap is important to understand from the stance of public policy or tenure and promotion, the free-form portion of course evaluations can be as more nettling to teachers than the raw numbers. Some individual words may be obviously gendered in their usage ("shrill"); other words that crop up frequently in evaluations can be the source of nagging doubts for teachers. Is a frequent criticism of "unfair grading" a universal feature of evaluations, or does it show up more in evaluations of women because students more easily chuck aside their respect for female authority figures than for male ones?

2.1 Data overview

To investigate questions about the niceties of language like these requires a corpus of evaluations significantly larger than is likely to be generated experimentally. The data set visualized here is, instead, a collection of approximately 14.1 million evaluations of 1.88 million individual teachers scraped from the website RateMyProfessors.com (hereafter abbreviated as "RMP"). RMP is probably the most widely used of a number of online sites which allow students to rate faculty members on a variety of criteria (including clarity, helpfulness, and easiness). For each review I collected both the ratings and the full text of the "comment" on the evaluation, averaging 43.1 words per evaluation. These were ingested into the Bookworm text-as-data visualization platform to facilitate computation and visualization.

The vast majority of reviews span a period from 2003 to 2014; the highest use of the website was in 2005-2007, but it continued to add over a million reviews a year through 2013. (For comparison, there were 11 million full-time and 6 million part-time college students in the United States in 2010.) Usage is elective by students, and penetration seems to be quite different across roughly comparable colleges; for example, Rutgers University in New Jersey contributed 75,000 evaluations, while the University of Maryland's flagship campus in College Park contributed only 12,000. Evaluations are generally positive and somewhat biased towards the extremes: for the "Helpful" score, 48.6% of reviews are '5', 16.6% are 4, 10.2 are '3', 9.2% are '2', and 15.2% are '1'.

2.2 Visualization strategy

The data set was visualized using a dynamic D3 visualization embedded into a single-page webpage. I have written a library using the d3 data visualization library [5] that allows interactive visualization of texts stored using Bookworm. For the charts in the web page, I used a two-color scatterplot with a categorical y axis (Fig. 1). Users enter any word or two-word phrase into a text box, and the chart updates to show the relative frequency of the word in evaluations of men and women for the most common fields of teaching. Academic department is reported on the RMP website. The default search was for the term "funny," which shows a male skew in all disciplines; other than that and the suggestion that users try pronouns, the site makes no statements about what the nature of the bias revealed might be. The reason for the use of discipline is as a sort of control on the fact that women are more likely to teach in humanities fields than in the sciences; by allowing comparisons across a variety of disciplines, users can see both whether a gap persists in many settings or likely the result of random noise, as well as correcting to a first approximation the fact that men and women teach different things. For the display, the results were limited to the 25 most common dsicplines (about 9.2 million reviews.) Gender was automatically assigned using an R package by Lincoln Mullen [3] that guesses gender based on first name frequency in the US Social Security database. Names with few examples or a less than 95% gender skew were excluded from the visualization.

User engagement Since its creation, the visualization has seen consistent use from a variety of traffic sources. To date, I have logged about 1.5 million queries from about 130,000 distinct IP addresses. These logs show many examples of sustained engagement; over 1,000 individual IP addresses have each generated over 100 queries apiece. These logs allow more extensive exploration of user engagement. User queries seem to focus on terms that are strongly evaluative or that users suspect to have a strong gender bias. Female-biased words are slightly more common. In order from most male-biased to most female-biased, the 50 most commonly searched terms are:

he, handsome, jerk, dick, arrogant, genius, gay, hilarious, entertaining, funny, sexy, clever, brilliant, old, cool, ass, intelligent, smart, fat, cute, interesting, weird, engaging, idiot, boring, great, knowledgeable, best, knowledgeable, inspiring, awesome, good, challenging, pretty, excellent, passionate, fun, lazy, difficult, hot, good teacher, hard, stupid, fair, dumb, bad, competent, easy, approachable, kind, hate, useless, tough, demanding, attractive, creative, bad teacher, enthusiastic, angry, amazing, love, clear, confusing, ugly, condescending, crazy, happy, understanding, worst, terrible, nice, friendly, helpful, mean, biased, harsh, horrible, unfair, awful, aggressive, organized, rude, incompetent, annoying, caring, strict, disorganized, evil, bossy, beautiful, sweet, she.

2.3 User engagement experiments

Visualizations that experience steady public use like this one provide opportunities for controlled experiments in the form of A-B testing

on regular web traffic. Two A-B testing experiments of this sort were conducted to see how visualization implementations might change user engagement. These were performed on the ordinary stream of web traffic to the page; the first over about 6 months in early 2016, and the second over a few weeks in July 2016. In both cases, the javascript on the page randomly assigned users to a random test group when they navigated to the page; the test group for a user was kept in local storage on their browser. This means that testing group was persistent by browser (possibly including several IP addresses) but not across different browsers or computers.

The first experiment (on two samples of 15,000 users apiece) added a "suggested queries" box that, for most words in the corpus, used a word2vec model to generate a list of 10 words used in similar circumstances. (For "funny," users can click to see the usage of "hilarious," "humorous," "entertaining," etc.) For users not sure what to type, this was intended to give a point of entry into the visualization; the box updates with each query, and generates synonyms for each term as they are entered.

The second experiment (on a much smaller pair of test groups with roughly 500 members apiece) studied the effect of animated transitions on user engagement. Upon entry of a new search term, the visualization engages in two delayed transitions; first to adjust the positions of the circles for men and women in each bar, and second to reorder disciplines so the disciplines with the highest use of a term appear at the top of the graph. These transitions are intended link each query to another; visualization research has shown that animated transitions of this sort can improve user perception. [8] It was thought that this improved perception of patterns in the data might increase user engagement by making the differences between the charts more clear after updates. For the experimental test group, transitions were disabled and the chart area simply refreshed.

Neither experiment showed significant effects on the number of queries entered by users into the website. This was somewhat surprising; the lack of animated transitions, in particular, seems subjectively to greatly diminish the user experience. An optimistic spin on the lack of change is that viewers are more engaged with the substance than the form of data visualization in this particular case. For exploratory data analysis to be publicly useful, the users need to be able to have questions they find interesting to bring to the data archive. It is possible that persistence in time at this site is driven through interest in the data strongly enough that differences in data visualization quality are less important. Insofar as persistence is an adequate metric (rather than any of the other outcomes that were not studied here: e.g. understanding of results, sharing of the visualization), this might suggest that humanists can feel more confident in building data visualizations on interesting data without worrying about perfecting the user experience.

2.4 Data limitations

Since the use of an exploratory data analysis is heavily limited by the constraints on the data, it is worth exploring some of those here. (A link from the webpage gives some of the restrictions of the data. Log data indicates viewership on the caveats is about 10% that of the main page.) This data does not, for a variety of reasons, easily rise to the standards of providing useful proof of gender bias along the lines of the controlled studies cited earlier. The testing corpus is large enough that it would be trivial to create large numbers of statements along the lines of "women are described as shrill more often than men in these evaluations." (Although the algorithmic gender assignment process is imperfect, misclassifications simply bias the evidence of an effect from gender toward zero; the real gaps are almost certainly slightly larger than displayed in the chart because of (for instance) female professors named "Chris" or male professors named "Ashley."

As a heterogeneous collection of postings by a self-selecting population over time, though, it is hazardous to generalize from the population of RMP reviews to broader patterns in course evaluations. Students who rate professors online are a different population than those who assess their professors in class; there may be a greater bias towards strongly positive or strongly negative reviews. (The majority of reviews on the site are positive). There are certain confounding factors of demography; "old" is a more frequent descriptor of men than women, but this is at least partially because women are particularly under-represented in the professoriate at higher ages. And words are only an approximate proxy for language: some uses of the word "brilliant" may actually be phrases like "not brilliant." (On the other hand, one of the advantages of an exploratory framework like this is that users can search for "not brilliant" and see that it is only about one in 500 uses of the word "brilliant.")

A second class of problems is slightly more subtle; there may be differences in the evaluations of men and women based on the demographics of *evaluators*. For example, different academic fields have different gender splits that tend to exist both for instructors and for students. (For example, computer science has a strikingly high percentage of men both as majors and as faculty members.) Disentangling discipline in the primary chart acts as a sort of control on this; but within subfields, it is likely the pattern persists. Within history departments, for example, it is conceivable that more men teach military history than women's history, and that more women take women's history courses than military history courses. This means that evaluations of men and women are almost certainly drawn from different groups. The base visualization shows greater use of pronouns in evaluations of women than of men; but rather reflecting different ways that some abstract ideal type of the "student" talks about the same people, this phenomenon likely reflects differing uses of pronouns by men and women in their writing.

In this latter case, the evidence generated by the chart may still be useful in learning to read evaluations, because changes in student composition are legitimately important differences in the evaluations that teachers receive. If (for instance) female students are generally harsher evaluators, female faculty members would be systemically disadvantaged in evaluations even if no individual students showed any bias.

None of these challenges to statistical verification are necessarily insurmountable. They do, however push the assessment of the data well beyond the simple model implied by the visualization; that frequency of use of a word is a function of discipline, word, and faculty gender. Subsequent to publication of the visualization, some scholars have attempted to use the visualization as evidence for more statistically rigorous projects. [12]

3 The ends of our exploring

Even supposing that no strict proofs of bias were possible, though, there are still a great number of possible benefits to pull from an exploratory humanistic data visualization like this. One is exploration of a corpus for discovery. Private versions of the site make it possible to click on any segment to see examples of reviews of (for instance) students describing female physicists as "unfair." Because of privacy concerns, however, this functionality was disabled in the public-facing website.

Another valuable use of publicly-facing exploratory visualizations to aggregate public knowledge about what terms might be useful; they can help to crowd-source the uses of exploratory data visualization, or suggest avenues for future research by others. The visualization of teaching evaluations has connections to many other areas in the humanities, where subject matter expertise may be widely distributed and relatively independent from the ability to analyze data at all. (This may differ from the science, where it is more reasonable to expect a baseline of quantitative literacy out of domain experts). Online exploratory interfaces like these can act as tools to rapidly generate and preliminarily test hypotheses. For example, a collection of Twitter users quickly identified on of the most interesting patterns in the RMP data set: that frequently opposite pairs of words show the *same* gender skew. Women are both more frequently "nice" and more frequently "mean," while men are both more frequently "brilliant" and more frequently "idiots." This suggests that in many cases, not just individual words but whole standards for assessment show a gender bias.

The goal of these findings could be to inspire further quantitative studies down the road; but they could also be simply to help people who have to read teaching evaluations. To be aware that a given set of terms *may* be gendered in its application is useful information to have

for persons who have to make decisions based on them; an expectation of strict proof may be unnecessary for such uses.

Finally, well designed exploratory data analyses can conceivably act in a mode of *affirming experience.* Although most programmatic remedies for bias are will require structural change and thus should be based on solid evidence, some effects may be usefully remediated simply by allowing individuals to see that their experiences are not unique. As one user of the visualization put it on Twitter, "In a way it's… a relief? To know it's NOT all in your head?" The ends of humanistic data visualization like this may be less in the public policy or decision-making that is the aspiration of much scientific and social-scientific data visualization but in allowing individuals to place their own experience in a broader framework that can help them to be affirmed in some understandings of bias that are frequently denigrated.

3.1 Exploring Evidence without proof

Still, while affirming experience may be a valuable goal, it is directly contrary to scholarship's drive to create new knowledge. One user succinctly summarized their interpretation of the site: "Teacher evals tell us exactly what we know about gender biases in higher ed." Very few users have reported finding firmly unexpected things; those unwilling to admit bias exists can find grounds to persist in doing so in the admittedly unscientific style of the visualization and data collection. Anecdotally, users seem to engage much more strongly with terms that do show bias than those that don't. For example: are evaluations biased in the sense that students more often describe the *appearance* of female faculty? The frequent query lists shows users entering terms that will be interesting from a gendered standpoint; a large number are terms like "sexy," "hot," or "handsome." I suspect that users may expect these terms to show a significant gender bias; they do not.

In fact, among the most frequent queries, there is some minor reason to think that users may be more likely to exit on encountering a word that shows less of a gender split (Fig. 2). For each word in the server logs, I have looked at its position in each individual user's path through the archive and calculated the percentage of times that a user exited after viewing the results for that query; I compare this to how consistent the gap between male and female scores is. (I choose to use number of disciplines, rather than average gap, because the most striking visual feature on the charts is how frequently all disciplines show the same direction of a gender split). Although the effect is not especially strong, a linear model is slightly suggestive of a slightly increased probability to depart the page after viewing a null result (p=.014, but R^2 = .06).

In fact, a large number of words related to physical appearance, dress, and attractiveness are either skewed male or show no major bias in the plot. That men's appearance is talked about as much as women's doesn't necessarily mean there is no bias in attitude or importance; it may be that negative physical assessments are more common of women, or that criticisms of dress in women are more highly linked with an overall negative opinion. But it is still striking that, insofar as this visualization does give evidence of differing standards for men and women vis-a-vis categories like "intelligence" or "professionalism" or "care," it does not do so for physical appearance.

This highlights one key problem with detaching exploratory from confirmatory data analysis; that users may simply use exploration to confirm models they already hold, without being able to convince others *and* without being fully receptive to their own preconceptions being disproven.

3.2 Conclusion: narrativity as non-positivistic exploration

While exploratory data analysis is exciting from the traditional digital humanities activities of access and engagement, it still may fall short of the highest aspirations of humanists to promote critical engagement with sources. Confirmatory data analysis to high statistical standards would certainly provide one alternative; is there any other?

The exploratory possibilities of interactive visualization are incredibly valuable for humanists; but humanists can do more than the RMP visualization described here to lead users to an interpretation of the data even without full-on confirmatory data analysis. One possibility worth further exploration is *narrative assemblage.* The same tools that make

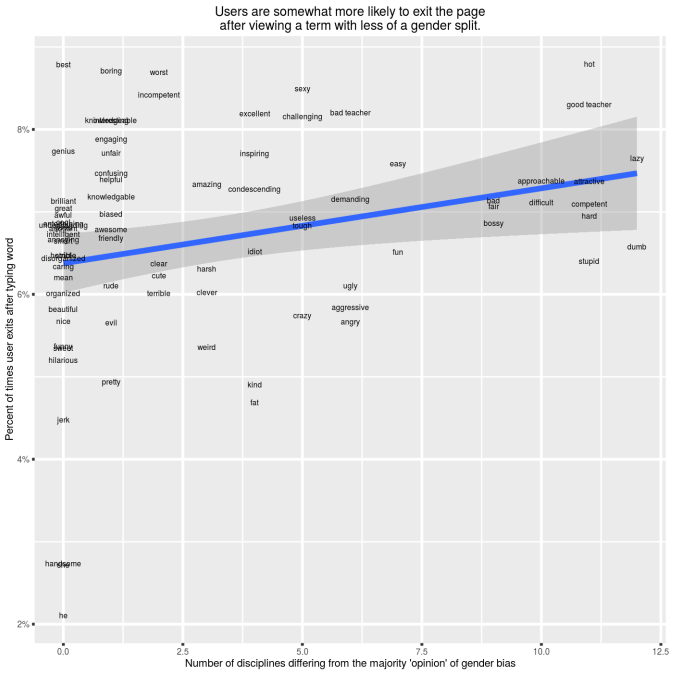

Fig. 2. Scatterplot of gender bias in the visualization and rate of users leaving the site. In a linear model, p=.0139

it possible to build richly interactive visualizations with transitions for the open web like this one also make it possible to embed a changing visualization in a narrative story. This is, in fact, a leading form of information visualization, particularly in the section of the information visualization community around digital journalism. Combining interactive visualization with narrative description provides one particularly useful avenue for further work in data visualization in the digital humanities.

References

[1] J. T. Behrens. Principles and procedures of exploratory data analysis. 2:131.

[2] C. Blevins. *Digital History's Perpetual Future Tense*. University of Minnesota Press.

[3] C. Blevins and L. Mullen. Jane, john... leslie? a historical method for algorithmic gender prediction. 9.

[4] A. Boring and others. Gender biases in student evaluations of teachers. 13.

[5] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. 17:23012309. doi: 10.1109/TVCG.2011.185

[6] J. Drucker. *Graphesis: visual forms of knowledge production*.

[7] A. Gelman. Exploratory data analysis for complex models. 13:755–779. doi: 10.1198/106186004X11435

[8] J. Heer and G. Robertson. Animated transitions in statistical data graphics. 13:12401247. doi: 10.1109/TVCG.2007.70539

[9] L. MacNell, A. Driscoll, and A. N. Hunt. Whats in a name: Exposing gender bias in student ratings of teaching. 40:291–303. doi: 10.1007/s10755-014-9313-4

[10] L. Martin. Gender issues and teaching.

[11] G. Potvin and Z. Hazari. Student evaluations of physics teachers: On the stability and persistence of gender bias. 12:020107. doi: 10.1103/PhysRevPhysEducRes.12.020107

[12] D. Storage, Z. Horne, A. Cimpian, and S.-J. Leslie. The frequency of Brilliant and Genius in teaching evaluations predicts the representation of women and african americans across fields. 11:e0150194. doi: 10.1371/journal.pone.0150194