**The Need for Openness in Data Journalism**

Do films that pass the Bechdel Test make more money for their producers? I’ve replicated Walt Hickey’s [recent article](http://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/) in FiveThirtyEight to find out. My results confirm his own in part, but also find notable differences that point the need for clarification at a minimum. While I am far from the first to make this argument, this case is illustrative of a larger need for journalism and other data-driven enterprises to borrow from hard-won scientific practices of sharing data and code as well as supporting the review and revision of findings. This admittedly lengthy post is a critique of not only this particular case but also an attempt to work through what open data journalism could look like.

**The Angle: Data Journalism should emulate the openness of science**

New data-driven journalists such as FiveThirtyEight have faced criticism from many quarters and the critiques, particularly around the naïveté of assuming credentialed experts can be bowled over by quantitative analysis so easily as the terrifyingly innumerate pundits who infest our political media [[1](http://krugman.blogs.nytimes.com/2014/03/18/sergeant-friday-was-not-a-fox/),[2](http://pressthink.org/2014/03/review-and-comment-on-the-launch-of-nate-silvers-fivethirtyeight-com-for-espn/),[3](http://gigaom.com/2014/03/17/nate-silvers-new-fivethirtyeight-site-launches-but-can-his-data-driven-journalism-find-a-market/),[4](http://www.vanityfair.com/online/daily/2014/03/nate-silver-vs-the-experts-five-thirty-eight)]. While I find these critiques persuasive, I depart from them here to instead argue that I have found this “new” brand of data journalism disappointing foremost because *it wants to perform science without abiding by scientific norms.*

The questions of demarcating what is or is not science are fraught, so let’s instead label my gripe a “failure to be open.” By openness, I don’t mean users commenting on articles or publishing whistleblowers’ documents. I mean “openness” more in the sense of “open source software” where the code is made freely available to everyone to inspect, copy, modify, and redistribute. But the principles of open-source software trace their roots more directly back norms in the scientific community that [Robert Merton](https://en.wikipedia.org/wiki/Robert_K._Merton#Sociology_of_science) identified and came to known as “[CUDOS](https://en.wikipedia.org/wiki/Mertonian_norms)” norms. It’s worth reviewing two of these norms because Punk-ass Data Journalism is very much on the lawn of Old Man Science and therein lie exciting possibilities for [exciting adventures](http://viewhdwallpaper.com/wp-content/uploads/2014/02/Pixar%E2%80%99s-UP-Movie-HD-Wallpaper.jpg).

The first and last elements of Merton’s “CUDOS” norms merit special attention for our discussion of openness. **Communalism** is the norm that scientific results are shared and become part of a commons that others can build upon — this is the bit about “standing upon the shoulders of giants.” **Skepticism** is the norm that claims must be subject to organized scrutiny by community — which typically manifests as peer review. Both of these strongly motivated philosophies in the open source movement, and while they are practiced imperfectly in my experience within the social and information sciences (see my colleagues’ recent work on the “[Parable of Google Flu](http://bit.ly/NHIgqM)“), I nevertheless think data journalists should strive to make them their own practice as well.

1. *Data journalists should be open in making their data and analysis available to all comers*. This flies in the face of traditions and professional anxieties surrounding autonomy, scooping, and the protection of sources. But how can claims be evaluated as true unless they can be inspected? If I ask a data journalist for her data or code, is she bound by the same norms as a scientist to share it? Where and how should journalists share and document these code and data?
2. *Data journalists should be open in soliciting and publishing feedback*. Sure, journalists are used to clearing their story with an editor, but have they solicited an expert’s evaluation of their claims? How willing are they to publish critiques of, commentary on, or revisions to their findings? If not, what are the venues for these discussions? How should a reporter or editor manage such a system?

*The Guardian*‘s [DataBlog](http://www.theguardian.com/news/datablog" \t "_blank) and [ProPublica](http://www.propublica.org/tools/" \t "_blank) have each doing exemplary work in posting their datasets, code, and other tools for several years. Other organizations like the [Sunlight Foundation](https://sunlightfoundation.com/tools/) develop outstanding tools to aid reporters and activists, the [Knight Foundation](http://www.knightfoundation.org/what-we-fund/innovating-media) has been funding exciting projects around journalism innovation for years, and the [Data Journalism Handbook](http://datajournalismhandbook.org/1.0/en/) reviews other excellent cases as well. My former colleague, Professor [Richard Gordon](http://www.medill.northwestern.edu/experience/people/faculty/rich-gordon.html) at Medill reminded me ideas around “[computer assisted reporting](https://en.wikipedia.org/wiki/Computer_assisted_reporting)” have been in circulation in the outer orbits of journalism for decades. For example, [Philip Meyer](https://en.wikipedia.org/wiki/Philip_Meyer) has been (what we would now call) evangelizing since the 1970s for “[precision journalism](http://www.unc.edu/~pmeyer/book/)” in which journalists adopt the tools and methods of the social and behavioral sciences as well as its norms of sharing data and replicating research. Actually, if you stopped reading now and promised to read his [2011 Hedy Lamarr Lecture](http://www.nieman.harvard.edu/reports/article-online-exclusive/100044/Precision-Journalism-and-Narrative-Journalism-Toward-a-Unified-Field-Theory.aspx), I won’t even be mad.

The remainder of this post is an attempt demonstrate some ideas of what an “open collaboration” model for data journalism might look like. To that end, this article tries to do many things for many audiences which admittedly makes it hard for any single person to read. Let me try to sketch some of these out now and send you off in the right path.

* First, I use an article Walt Hickey of FiveThirtyEight published on the relationship between the financial performance of films that the extent to which they grant their female characters substantive roles as a case to illustrate some pitfalls in both the practice and interpretation of statistical data. This is a story about having good questions, ambiguous models, wrong inferences, and exciting opportunities for investigation going forward. If you don’t care for code or statistics, you can start reading at “The Hook” below and stop after “The Clip” below.
* Second, for those readers who are willing to pay what one might call the “[Iron Price](http://gameofthrones.wikia.com/wiki/Iron_price) of Data Journalism”, I go “[soup to nuts](https://en.wikipedia.org/wiki/Soup_to_nuts)” and attempt to replicate Hickey’s findings. I document all the steps I took to crawl and analyze this data to illustrate the need for better documentation of analyses and methods. This level of documentation may be excessive or it may yet be insufficient for others to replicate my own findings. But providing this code and data may expose flaws in my technical style (almost certainly), shortcomings in my interpretations (likely), and errors in my data and modeling (hopefully not). I actively invite this feedback via email, tweets, comments, or pull requests and hope to learn from it. I wish new data journalism enterprises adopted the same openness and tentativeness in their empirical claims. You should start reading at “Start Your Kernels…”
* Third, I want to experiment with styles for analyzing and narrating findings that make both available in the same document. The hope is that motivated users can find the detail and skimmers can learn something new or relevant while being confident they can come back and dive in deeper if they wish. Does it make sense to have the story up front and the analysis “below the fold” or to mix narrative with analysis? How much background should I presume or provide about different analytical techniques? How much time do I need to spend on tweaking a visualization? Are there better libraries or platforms for serving the needs of mixed audiences? This is a meta point as we’re in it now, but it’ll crop up in the conclusion.
* Fourth, I want to experiment with technologies for supporting collaboration in data journalism by adopting best practices from open collaborations in free software, Wikipedia, and others. For example, this blog post is not written in a traditional content-management system like WordPress, but is an interactive “[notebook](http://ipython.org/notebook.html)” that you can download and execute the code to verify that it works. Furthermore, I’m also “hosting” this data on [GitHub](https://github.com/brianckeegan/Bechdel) so that others can easily access the writeup, code, and data, to see how it’s changed over time (and has it ever…), and to suggest changes that I should incorporate. These can be frustrating tools with demoralizing learning curves, but these are *incredibly* powerful once apprenticed. Moreover, there are amazing resources and communities who exist to support newcomers and new tools are being released to flatten these learning curves. If data journalists joined data scientists and data analysts in sharing their work, it would contribute to an incredible knowledge commons of examples and cases that is lowering the bars for others who want to learn. This is also a meta point since it exists outside of this story, but I’ll also come back to it in the conclusion.

In this outro to a very unusual introduction, I want to thank Professor Gordon from above, Professor [Deen Freelon](http://www.american.edu/soc/faculty/freelon.cfm" \t "_blank), [Nathan Matias](http://natematias.com/), and [Alex Leavitt](http://alexleavitt.com/) for their invaluable feedback on earlier drafts of this… post? article? piece? notebook?

**The Hook: The Bechdel Test article in FiveThirtyEight**

Walk Hickey published an article on April 1 on FiveThirtyEight, titled [The Dollar-And-Cents Case Against Hollywood’s Exclusion of Women](http://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/). The article examines the relationship between movies’ finances and their portrayals of women using a well-known heuristic call the [Bechdel test](https://en.wikipedia.org/wiki/Bechdel_test" \t "_blank). The test has 3 simple requirements: a movie passes the Bechdel test if there are (1) two women in it, (2) who talk to each other, (3) about something besides a man.

Let me say at the outset, *I like this article*: It identifies a troubling problem, asks important questions, identifies appropriate data, and brings in relevant voices to speak to these issues. I should also include the disclaimer that I am not an expert in the area of empirical film studies like[Dean Keith Simonton](http://psychology.ucdavis.edu/faculty_sites/simonton/) or [Nick Redfern](http://nickredfern.wordpress.com/). I’ve invested a good amount time in criticizing the methods and findings of this article, but to Hickey’s credit, I also haven’t come across any scholarship that has attempted to quantify this relationship before: **this is new knowledge about the world**. Crucially, it speaks to empirical scholarship that has exposed how films with award-winning female roles are significantly less likely to win awards themselves [[5]](http://link.springer.com/article/10.1023/B:SERS.0000029097.98802.2c), older women are less likely to win awards [[6]](http://www.amsciepub.com/doi/abs/10.2466/pr0.2000.86.1.175?journalCode=pr0), actresses’ earnings peak 17 years earlier than actors’ earnings [[7]](http://jmi.sagepub.com/content/early/2014/01/27/1056492613519861.abstract), and differences in how male and female critics rate films [[8]](http://baywood.metapress.com/app/home/contribution.asp?referrer=parent&backto=issue,5,8;journal,6,63;linkingpublicationresults,1:300310,1). I have qualms about the methods and others may be justified in complaining it overlooks related scholarship like those I cited above, but this article is in the best traditions of journalism that focuses our attention on problems we should address as a society.

Hickey’s article makes two central claims:

1. *We found that the median budget of movies that passed the test…was substantially lower than the median budget of all films in the sample.*
2. *We found evidence that films that feature meaningful interactions between women may in fact have a better return on investment, overall, than films that don’t.*

I call Claim 1 the “Budgets Differ” finding and Claim 2 the “Earnings Differ” finding. The results, as they’re summarized here are relatively straightforward to test whether there’s an effect of Bechdel scores on earnings and budget controlling for other explanatory variables.

But before I even get to running the numbers, I want to examine the claims Hickey made in the article. The interpretations he makes about the return on investment are particularly problematic interpretations of basic statistics. Hickey reports the following findings from his models (emphasis added).

*We did a statistical analysis of films to test two claims: first, that films that pass the Bechdel test — featuring women in stronger roles — see a lower return on investment, and second, that they see lower gross profits. We found no evidence to support either claim.*

*On the first test, we ran a regression to find out if passing the Bechdel test corresponded to lower return on investment. Controlling for the movie’s budget, which has a negative and significant relationship to a film’s return on investment,passing the Bechdel test had no effect on the film’s return on investment. In other words, adding women to a film’s cast didn’t hurt its investors’ returns, contrary to what Hollywood investors seem to believe.*

*The total median gross return on investment for a film that passed the Bechdel test was $2.68 for each dollar spent. The total median gross return on investment for films that failed was only $2.45 for each dollar spent.*

*…On the second test, we ran a regression to find out if passing the Bechdel test corresponded to having lower gross profits — domestic and international. Also controlling for the movie’s budget, which has a positive and significant relationship to a film’s gross profits, once again passing the Bechdel test did not have any effect on a film’s gross profits.*

Both models (whatever their faults, and there are some as we will explore in the next section) apparently produce an estimate that the Bechdel test has no effect on a film’s financial performance. That is to say, the statistical test could not determine with a greater than 95% confidence that the correlation between these two variables was greater or less than 0. *Because we cannot confidently rule out the possibility of there being zero effect, we cannot make any claims about its direction.*

Hickey argues that passing the test “didn’t hurt its investors’ returns”, which is to say there was no significant negative relationship, but neither was there a significant positive relationship: *The model provides no evidence of a positive correlation between Bechdel scores and financial performance.* However, Hickey switches gears an in the conclusions, writes:

*…our data demonstrates that films containing meaningful interactions between women do better at the box office than movies that don’t…*

I don’t know what analysis supports this interpretation. The analysis Hickey just performed, again taking the findings at their face, concluded that “passing the Bechdel test did not have any effect on a film’s gross profits” ***not*** “passing the Bechdel test increased the film’s profits.” While Bayesians will cavil about frequentist assumptions — as they are wont to do — and the absence of evidence is not evidence of absence, the “Results Differ finding” is not empirically supported in any appropriate interpretation of the analysis. The appropriate conclusion from Hickey’s analysis is “there no relationship between the Bechdel test and financial performance,” which he makes… then ignores.

What to make of this analysis? In the next section, I summarize the findings of my own analysis of the same data. In the subsequent sections, I attempt to replicate the findings of this article, and in so doing, highlight the perils of reporting statistical findings without adhering to scientific norms.

**The Clip: Look in here for what to tweet**

I tried to retrieve and re-analyze the data that Hickey described in his article, but came to some conclusions that were the same, others that were very different, and still others that I hope are new.

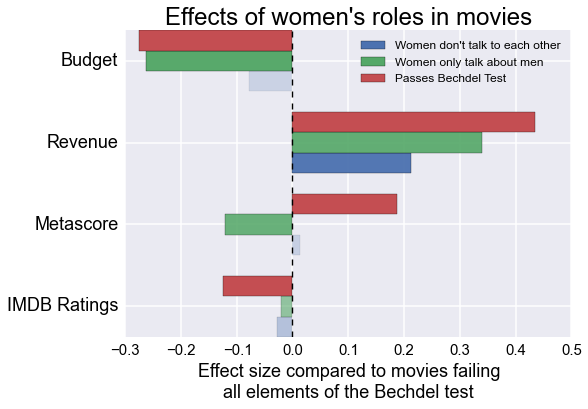
I was able to replicate some of his findings, but not others because specific decisions had to be made about the data or modeling that dramatically change the results of the statistical models. However, the article provides no specifics so we’re left to wonder when and where these findings hold, which points to the need for openness in sharing data and code. Specifically, while Hickey found that women’s representation in movies had no significant relationship on revenue, I found a positive and significant relationship.

But the questions and hypotheses Hickey posed about systematic biases in Hollywood were also the right ones. With a reanalysis using different methods as well as adding in new data, I found statistically significant differences in popular ratings also exist. These differences persist in the face of other potential explanations about differences arising because of genres, MPAA ratings, time, and other effects.

In the image below, we see that movies that have non-trivial women’s roles get 24% lower budgets, make 55% more revenue, get better reviews from critics, and face harsher criticism from IMDB users. Bars that are faded out mean my models are less confident about these findings being non-random while bars that are darker mean my models are more confident that this is a significant finding.

Movies passing the Bechdel test (the red bars):

* …receive budgets that are 24% *smaller*
* …make 55% *more* revenue
* …are awarded 1.8 *more* Metacritic points by professional reviewers
* …are awarded 0.12 *fewer* stars by IMDB’s amateur reviewers

[](http://nbviewer.ipython.org/github/brianckeegan/Bechdel/blob/master/Bechdel_test.ipynb)

**Read the entire replication**[**here**](http://nbviewer.ipython.org/github/brianckeegan/Bechdel/blob/master/Bechdel_test.ipynb)**.**

**Conclusions**

The four main findings from this analysis of the effects of women’s roles in movies are summarized in the chart above. These four points point to a paradox in which movies that pass an embarrassingly low bar for female character development make more money and are rated more highly by critics, but have to deal with lower budgets and more critical community responses. Is this definitive evidence of active discrimination in the film industry and culture? No, but it suggests systemic prejudices are contributing to producers irrationally ignoring significant evidence that “feminist” films make them more money and earn higher praise.

The data that I used here was scraped from public-facing websites, but there may be reasons to think that these data are inaccurate by those who are more familiar with how they’re generated. Similarly, the models I used here are simple Stats 101 ordinary least squares regression models with some minor changes to account for categorical variables and skewed data. There are no Bayesian models, no cross-validation or bootstrapping, and no exotic machine learning methods here. But in making the data available (or at least the process for replicating how I obtained my own data), others are welcome to perform and share the results such analyses — and this is ultimately my goal of asking data journalism to adopt the norms of open collaboration. When other people take their methodological hammers or other datasets and still can’t break the finding, we have greater confidence that the finding is “real”.

But the length and technical complexity of this post also raise the question of, who is the audience for this kind of work? Journalistic norms emphasize quick summaries turned around rapidly with opaque discussions of methods and analysis and making definitive claims. Scientific norms emphasize more deliberative and transparent processes that prize abstruse discussions and tentative claims about their “truth”. I am certainly not saying that Hickey should have the output of regression models in his article — 99% of people won’t care to see that. But in the absence of soliciting peer reviews of this research, how are we as analysts, scientists, and journalists to evaluate the validity of the claims unless the code and data are made available for others to inspect? Even this is a higher bar than many scientific publications hold their authors to (and I’m certainly guilty of not doing more to make my own code and data available), but it *should* be the standard, especially for a genre of research like data journalism where the claims reach such large audiences.

However, there are exciting technologies for supporting this kind of open documentation and collaboration. I used an “open notebook” technology called IPython Notebook to write this post in such a way that the text, code, and figures I generated are all stitched together into one file. You’re likely reading this post on a website that lets you view any such Notebook on the web where other developers and researchers share code about how to do [all manner of data analysis](https://github.com/ipython/ipython/wiki/A-gallery-of-interesting-IPython-Notebooks). Unfortunately, this was intended as a word processing or blogging tool, so the the lack of features such as more dynamic layout options or spell-checking will frustrate many journalists (apologies for the typos!). However, there are tools for customizing the CSS so that it plays well (see [here](http://danielfrg.com/blog/2013/02/16/blogging-pelican-ipython-notebook/) and [here](http://jakevdp.github.io/blog/2012/10/04/blogging-with-ipython/)). The code and data are [hosted on GitHub](https://github.com/brianckeegan/Bechdel), which is traditionally used for software collaboration, but its features for others to discuss problems in my analysis (issue tracker) or propose changes to my code (pull requests) promote critique, deliberation, and improvement. I have no idea how these will work in the context of a journalistic project, and to be honest, I’ve never used them before, but I’d love to try and see what breaks.

Realistically, practices only change if there are incentives to do so. Academic scientists aren’t awarded tenure on the basis of writing well-trafficed blogs or high-quality Wikipedia articles, they are promoted for publishing rigorous research in competitive, peer-reviewed outlets. Likewise, journalists aren’t promoted for providing meticulously-documented supplemental material or replicating other analyses instead of contributing to coverage of a major news event. Amidst contemporary anxieties about information overload as well as the weaponization of [fear, uncertainty, and doubt](https://en.wikipedia.org/wiki/Fear,_uncertainty_and_doubt) tactics, data-driven journalism could serve a crucial role in empirically grounding our discussions of policies, economic trends, and social changes. But unless the new leaders set and enforce standards that emulate the scientific community’s norms, this data-driven journalism risks falling into traps that can undermine the public’s and scientific community’s trust.

This suggests several models going forward:

* **Open data**. Data-driven journalists could share their code and data on open source repositories like GitHub for others to inspect, replicate, and extend. But as any data librarian will rush to tell you, there are non-trivial standards for ensuring the documentation, completeness, formatting, and non-proprietaryness of data.
* **Open collaboration**. Journalists could collaborate with scientists and analysts to pose questions that they jointly analyze and then write up as articles or features as well as submitting for academic peer review. But peer review takes time and publishing results in advance of this review, even working with credentialed experts, doesn’t imply their reliability.
* **Open deliberation**. Organizations that practice data-driven journalism (to the extent this is different from other flavors of journalism) should invite and provide empirical critiques of their analyses and findings. Making well-documented data available or finding the right experts to collaborate with are extremely time-intensive, but if you’re going to publish original empirical research, you should accept and respond to legitimate critiques.
* **Data omsbudsmen**. Data-driven news organizations might consider appointing independent advocates to represent public interests and promote scientific norms of communalism, skepticism, and empirical rigor. Such a position would serve as a check against authors making sloppy claims, using improper methods, analyzing proprietary data, or acting for their personal benefit.

I have very much enjoyed thinking through many of these larger issues and confronting the challenges of the critiques I’ve raised. I look forward to your feedback and I very hope this drives conversations about what kind of science data-driven journalism hopes to become.