Introduction to data science

Patrick Shafto

Department of Math and Computer Science

Plan for today

- HW 3
- Asking and answering questions (Part 1 of N)

HW 3

- Write a short tutorial on numpy and scipy
- Cover basic functionality with worked examples
- Due Sunday by 11:59 pm
- Also read the Betchel test notebook and come prepared to answer questions
- Evaluations due Tuesday by 11:59pm
    - Use the rubric!
    - Justify your scores!

HW 3: Examples

Asking and answering questions (Part 1 of N)

The need for openness in data journalism
https://nbviewer.jupyter.org/github/brianckeegan/Bechdel/blob/master/Bechdel_test.ipynb

Analyzing the structure of a report

What is the goal?

The Need for Openness in Data Journalism

Brian Keegan, Ph.D. (@bkeegan) College of Humanities and Social Sciences, Northeastern University

Do films that pass the Bechdel Test make more money for their producers? I've replicated Walt Hickey's recent article 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.

- 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.

- 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.

- **Second, for those readers who are willing to pay what one might call the "Iron Price of Data Journalism", I go "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..."

- 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.

- **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.

- 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.

- **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" that you can download and execute the code to verify that it works. Furthermore, I'm also "hosting" this data on GitHub 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.

Analyzing the structure of a report

What is the goal?

- Walk Hickey published an article on April 1 on FiveThirtyEight, titled The Dollar-And-Cents Case Against Hollywood's 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. 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.

- Hickey's article makes two central claims:

  - 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.

  - 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.

- Go back to the source to analyze:

- https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/

Analyzing the structure of a report

What is the goal?

What are the data?

- Our analysis relies on two data sets: BechdelTest.com and The-Numbers.com. The site BechdelTest.com is operated by committed moviegoers who analyze films and ascertain if they pass the Bechdel test. The site has detailed, coded information for about 5,000 films.

- To find financial information on these films, we went to The-Numbers.com, a leading site for box office and budget data. It inventories financial information for roughly 4,500 films.

- The intersection of The-Numbers and BechdelTest was a set of 1,615 films released between 1990 and 2013. When considering the financial information, we adjusted all numbers for inflation, using 2013 dollars. While hardly a complete record of contemporary films, this gave us a sample that has both rigorous evaluations of female character involvement as well as the most accurate financial data available online.

Analyzing the structure of a report

What is the goal?

What are the data?

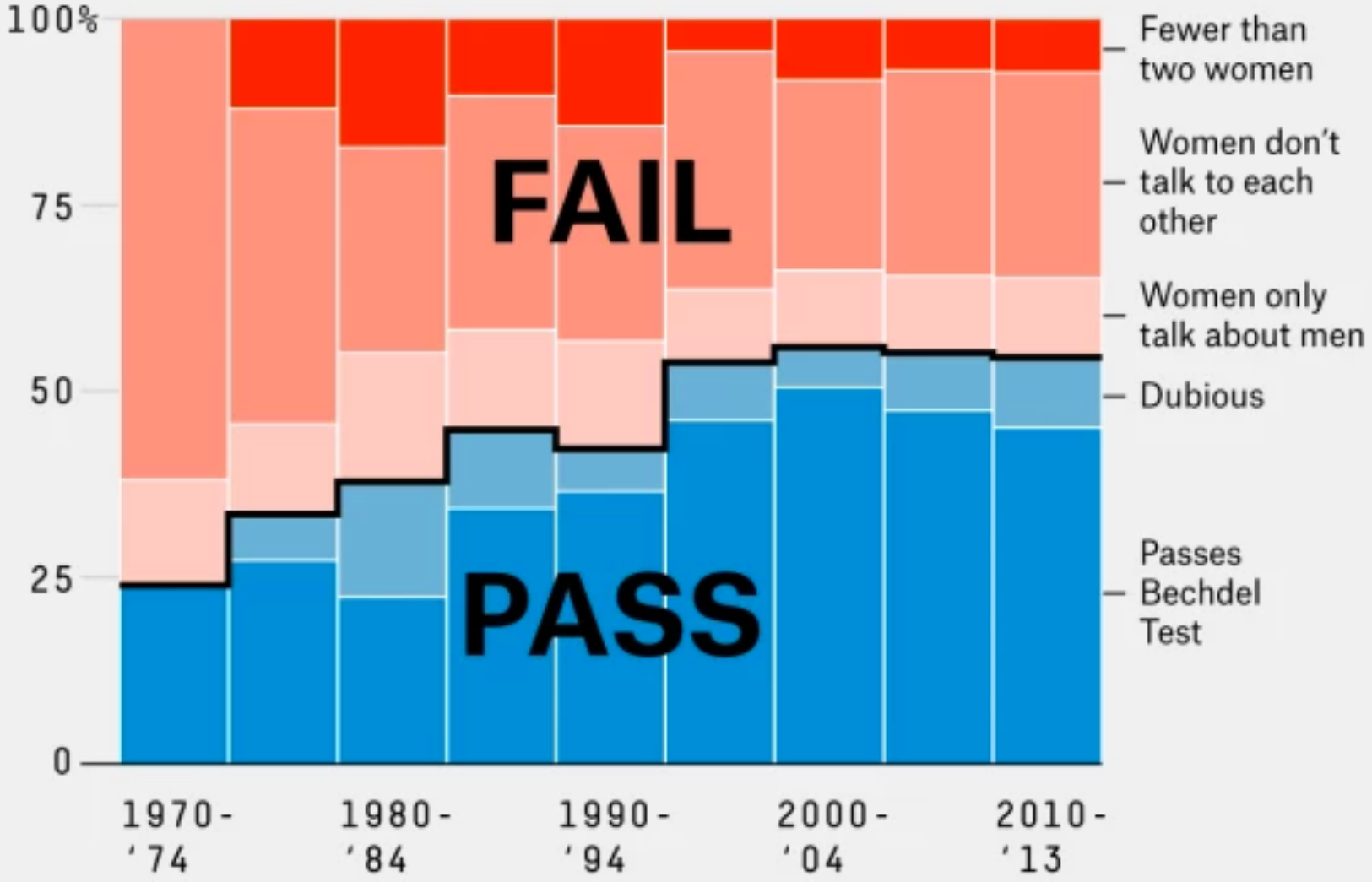
What is the analysis/logic?

- In order to get a consistent look at budget data, we're going to focus on films released from 1990 to 2013, since the data has significantly more depth since then.

- We ran a statistical test analyzing the inflation-adjusted median budgets of films, and found that films passing the Bechdel Test had a median budget that was 16 percent lower than the median budget of all films in the set.3 Smaller budgets constrain filmmakers in production, in landing talented performers and in promoting and advertising their films. Funding can make a big difference when it comes to how a movie performs at the box office, and how many people see it.

The Bechdel Test Over Time

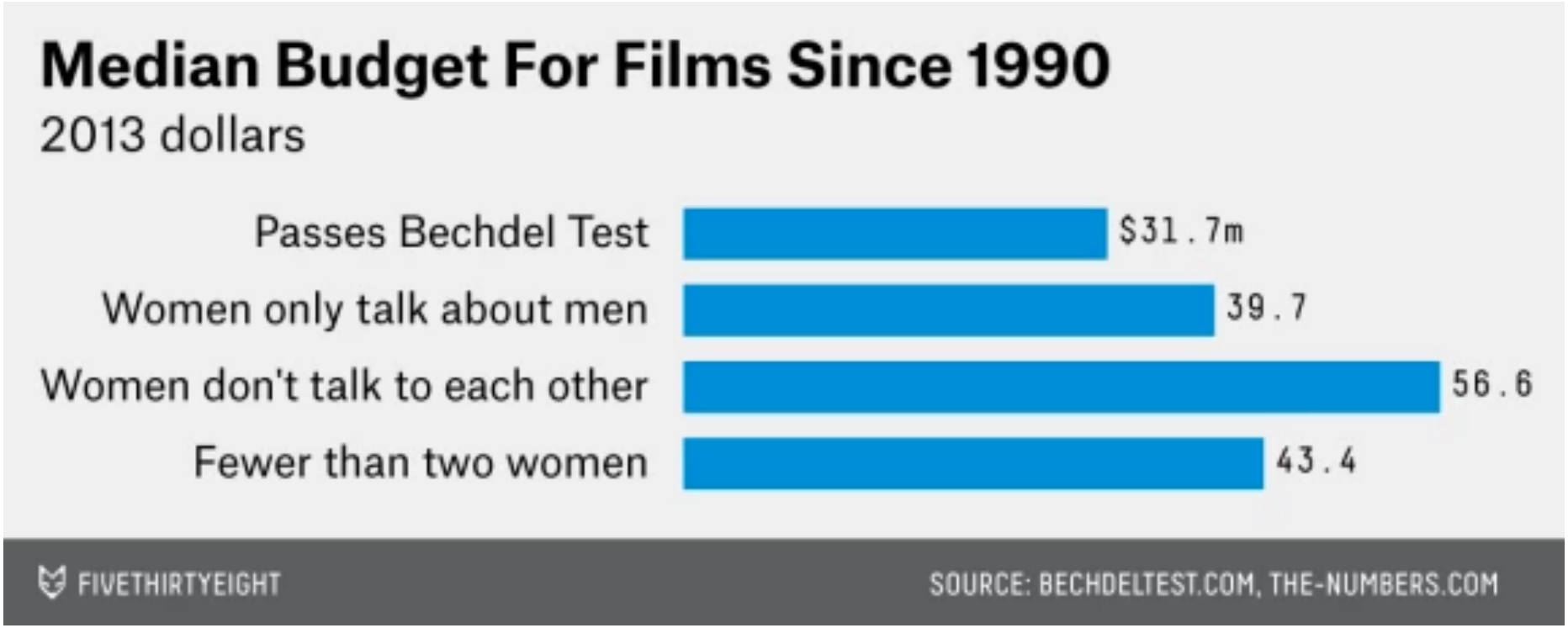
How women are represented in movies



ALLISON MCCANN

SOURCE: BECHDELTEST.COM

- The funding distinctions were even more remarkable when comparing the set of films that passed and the set of films that failed. The median budget of a film that failed the test was $48.4 million. The median budget of a film that passed was $31.7 million, or 35 percent less.



Analyzing the structure of a report

What is the goal?

What are the data?

What is the analysis/logic?

What are the conclusions?

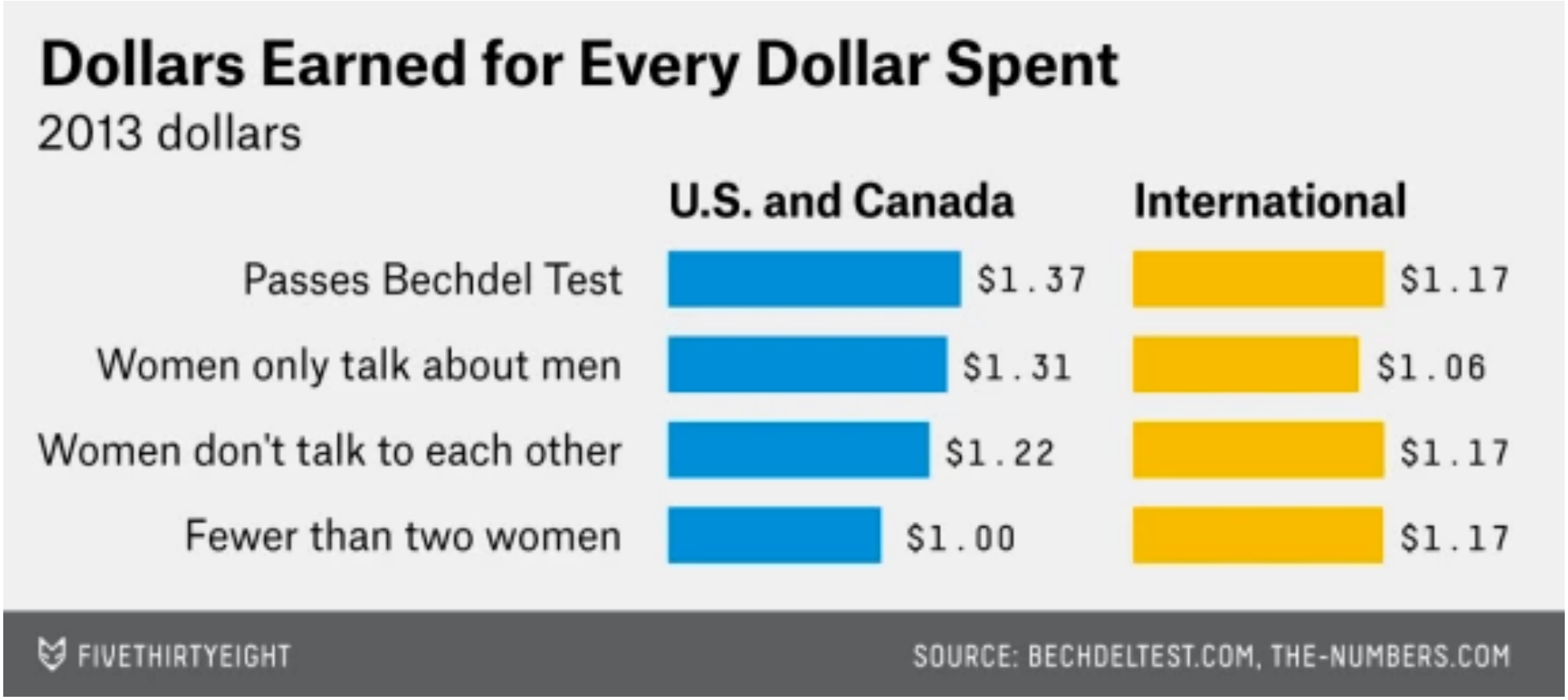
- In trying to explain the difference in funding levels between movies that pass the Bechdel test and your typical movie, Hollywood insiders referred to a culture in which men control the creative process and the purse strings, and a pervasive belief that audiences — both in the U.S. and internationally — just don't like films with strong female characters. Yet we found no evidence in the data to support the idea that films with women perform any worse at the box office than films without them, and what's more, films with women appeared to outperform expectations.

- 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,7 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. And while this might be a side effect of films with lower budgets tending to have higher returns on investment than films with higher budgets, it's still a strong indicator that films with women in somewhat prominent roles are performing well.

- 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,8 once again passing the Bechdel test did not have any effect on a film's gross profits.



- Hollywood is the business of making money. Since our data demonstrates that films containing meaningful interactions between women do better at the box office than movies that don’t, it may be only a matter of time before the data of dollars and cents overcomes the rumors and prejudices defining the budgeting process of films for, by and about women.

Asking and answering questions (Part 1 of N)

The need for openness in data journalism
https://nbviewer.jupyter.org/github/brianckeegan/Bechdel/blob/master/Bechdel_test.ipynb

HW 3

- Write a short tutorial on numpy and scipy
- Cover basic functionality with worked examples
- Due Sunday by 11:59 pm
- Evaluations due Tuesday by 11:59pm
    - Use the rubric!
    - Justify your scores!