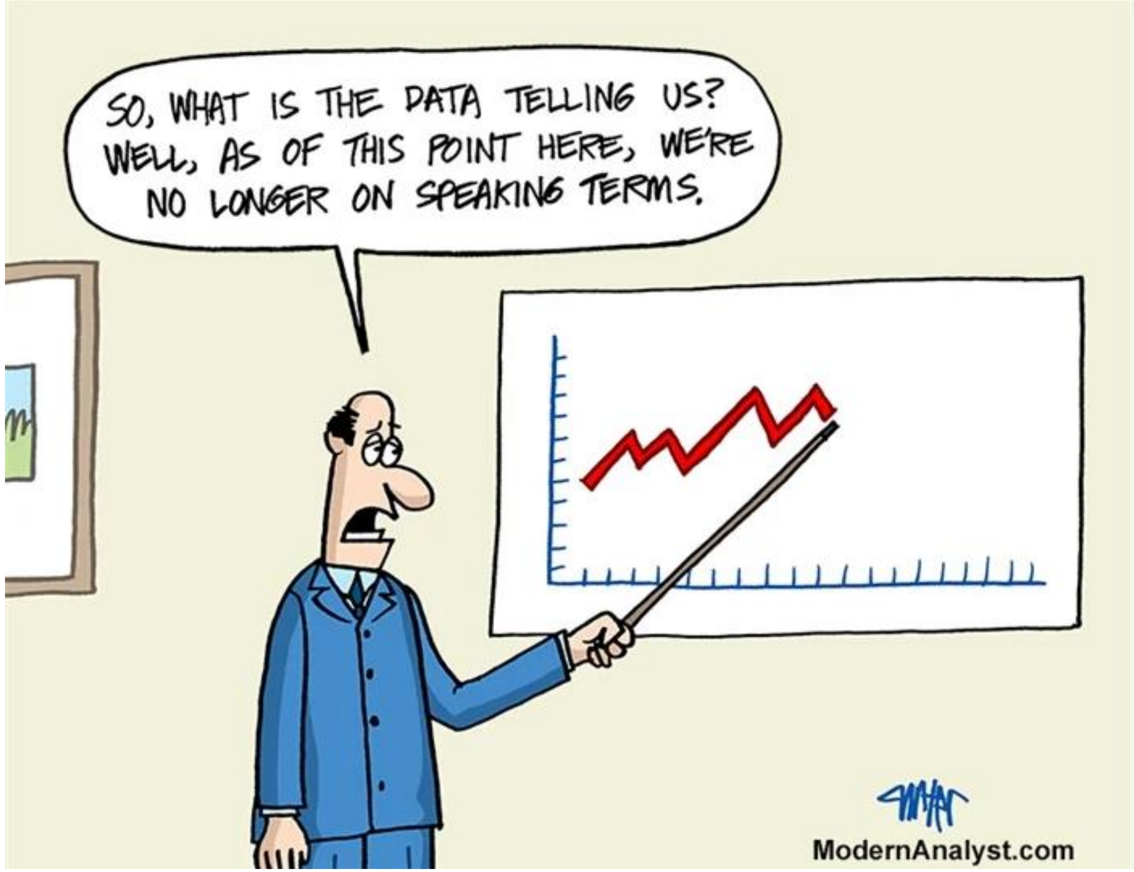
Data cleaning, ethics, and conclusions


SO, WHAT IS THE DATA TELLING US? WELL, AS OF THIS POINT HERE, WE'RE NO LONGER ON SPEAKING TERMS.
ModernAnalyst.com

Overview

Scrollytelling documents with Closeread

Ethics

Conclusions

Announcements

There will be no late penalty for the final project that are turned in before the end of reading period

ALL YOU NEED IS MOTIVATION
FALSE: YOU NEED FEAR AND AN APPROACHING DEADLINE
memegenerator.net

Highly recommend you turn it in at the original deadline so that you have plenty of time to study for the final exam

- The final exam is weighted significantly more than the project

I will have extra office hours this Friday from 1:30-3pm

Exam review session on Tuesday December 10th at 4pm

- In this classroom

Scrollytelling with Quarto and Closeread

Scrollytelling

"Scrollytelling" is a web design technique that uses visual and textual elements to tell a story as the reader scrolls through a page

- Closeread examples, and other examples

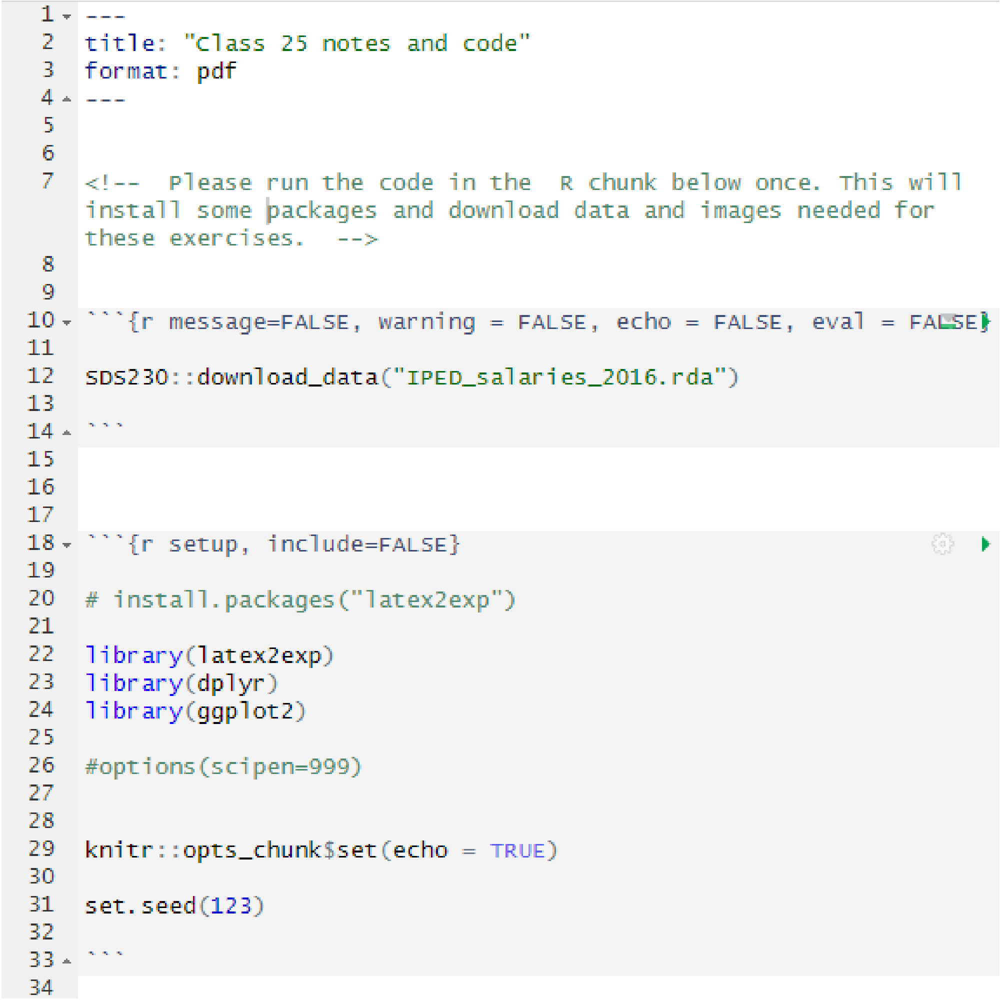
Quarto is an open-source publishing system that combines text, code, and output in one document. It is a successor to R Markdown.

We can creating scrollytelling documents by using the Closeread is a custom format in Quarto

Creating a Closeread Quarto document

To create a Closeread Quarto document, we start by specifying the appropriate quarto header

```
---

title: My Closeread story

format: closeread-html

---
```



~~~
---
title: "Class 25 notes and code"
format: pdf
---


<!--  Please run the code in the  R chunk below once. This will
install some packages and download data and images needed for
these exercises.  -->


```{r message=FALSE, warning = FALSE, echo = FALSE, eval = FALSE}

SDS230::download_data("IPED_salaries_2016.rda")

~~~

```{r setup, include=FALSE}

# install.packages("latex2exp")

library(latex2exp)
library(dplyr)
library(ggplot2)

#options(scipen=999)


knitr::opts_chunk$set(echo = TRUE)

set.seed(123)

```

```
```

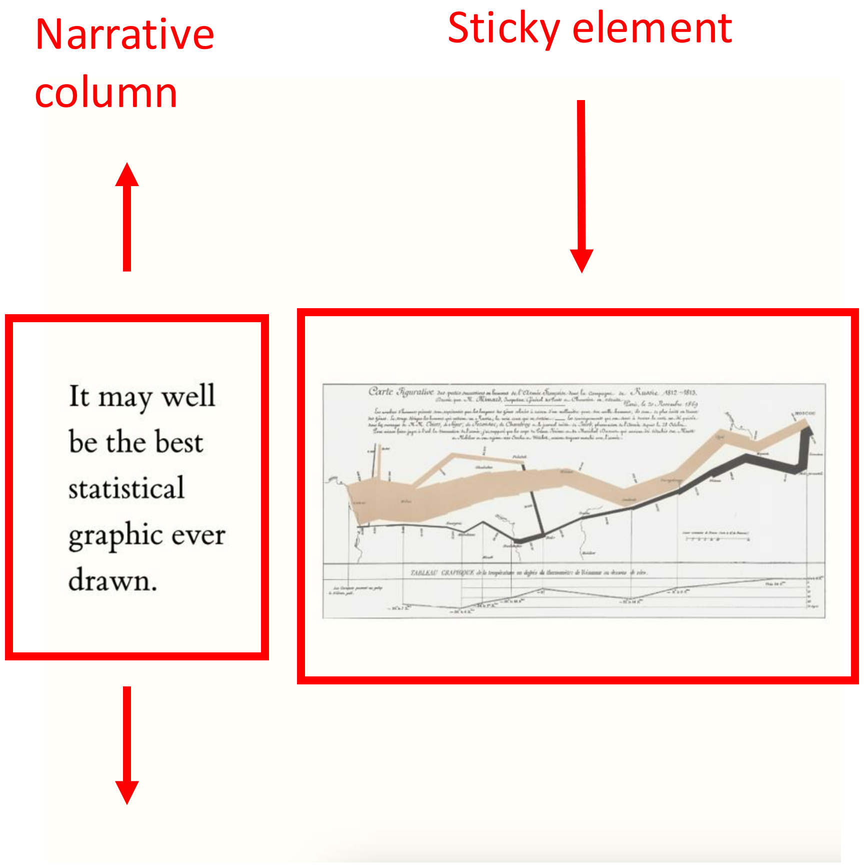
Closeread components

Every Closeread document consists of three components:

1. A **section** of the document flagged as a Closeread section

2. Within the section, a "**sticky**" **element** flagged to appear in the main column of the Closeread section

3. **An element** (often a paragraph of text) that appears in the "narrative column" will serve to trigger the appearance of the sticky element



Narrative column

Sticky element

It may well be the best statistical graphic ever drawn.

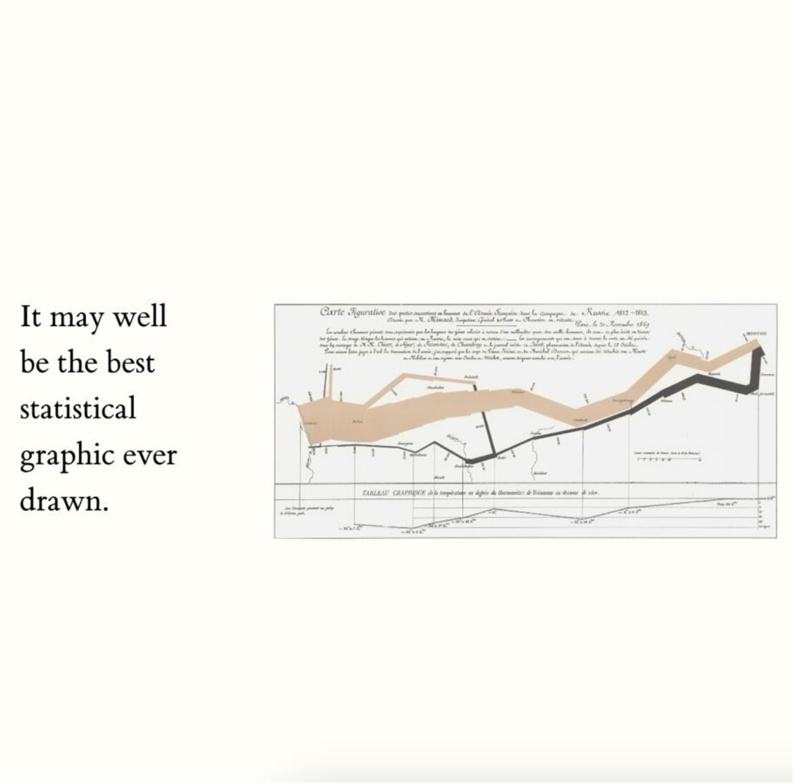
Closeread sections

We can add Closeread sections which specify what is part of our scrolling story (scrolly?)

- Elements outside of a section will appear as a normal Quarto HTML

```
::::{.cr-section}

                Text and images/code go in here…

::::
```

Must have at least 3 colons
Can have more to make sections stand out



It may well be the best statistical graphic ever drawn.

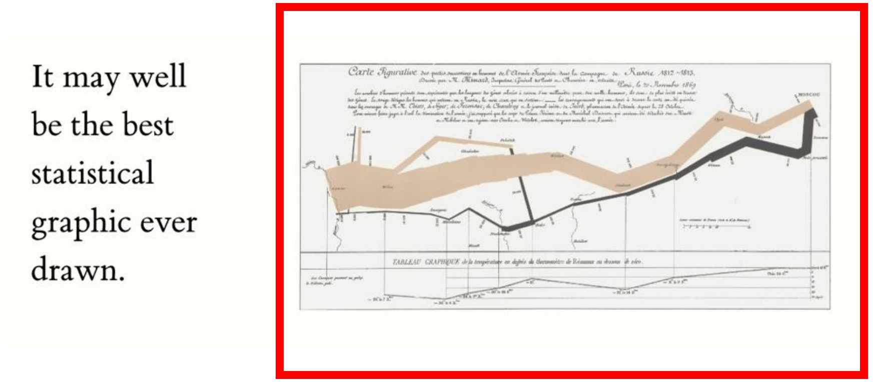
Adding stickies

Elements that are pinned as one scrolls are called ”stickies”

We can add stickies using:

```
:::{#cr-stickyname}
      sticky content
:::
```

Must start with #cr-



It may well be the best statistical graphic ever drawn.

```
:::{#cr-myimage}
   ![](path-to-myimage.png)
:::
```

~~~
:::{#cr-myplot}
    ```{r}
          hist(rnorm(15))
    ```
:::
~~~

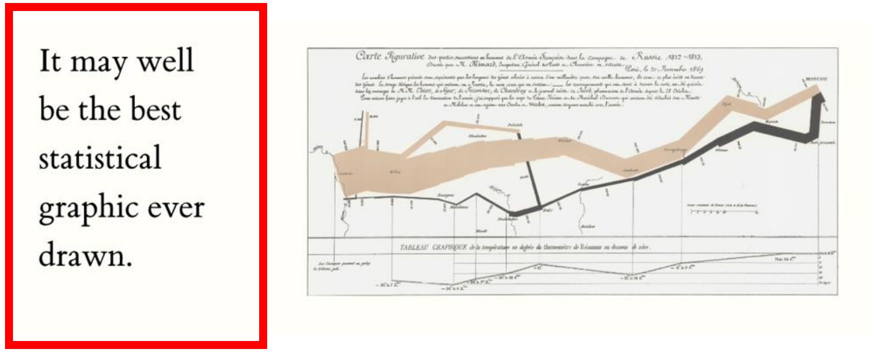
Triggers

Any text (or other elements) in a Closeread section that are not stickies will be placed in the narrative column

You can set a paragraph to trigger focus on a particular sticky by using Quarto’s cross-referencing syntax: @cr-mysticky

It may well be the best statistical grapic every drawn

@cr-myimage



It may well be the best statistical graphic ever drawn.

Let’s now show a histogram @cr-myplot

Focus effects

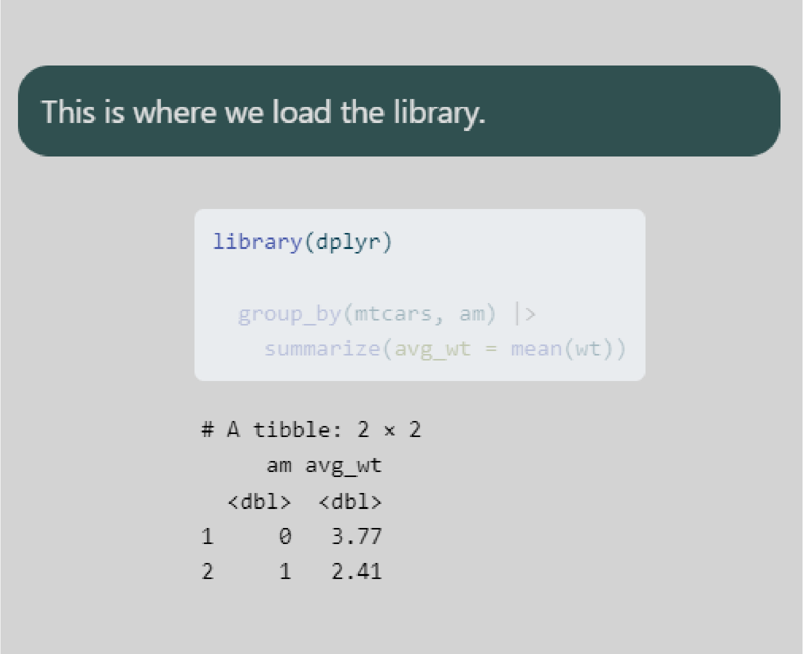
We can add “focus effects” which can do the following:

- Scale, pan, or zoom in on images (or other elements)
- Highlight lines of text

Example:

This is where we load the library. [@cr-dplyr]{highlight="1"}
This calculates summary statistics. [@cr-dplyr]{highlight="2-3"}

~~~
:::{#cr-dplyr}
          ```{r}
          library(dplyr)
          group_by(mtcars, am) |>
                    summarize(avg_wt = mean(wt))
         ```
:::
~~~



This is where we load the library.

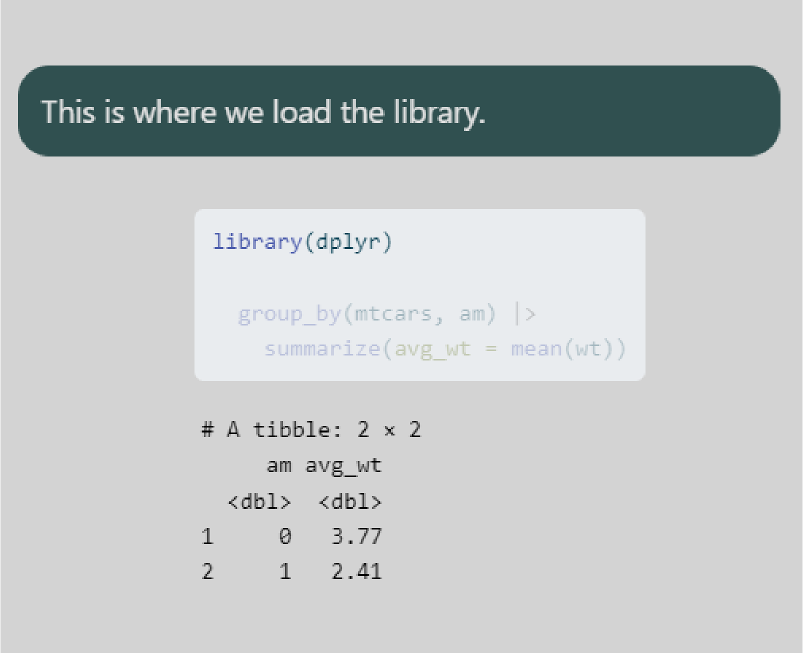
```
library(dplyr)

  group_by(mtcars, am) |>
    summarize(avg_wt = mean(wt))
```

```
# A tibble: 2 × 2
     am avg_wt
  <dbl>  <dbl>
1     0   3.77
2     1   2.41
```

Layouts

We can also change the layout of where the narration text and stickies appear by modifying the Quarto header

```
---
  format:
        closeread-html:
              cr-section:
                    layout: "overlay-center"
---
```



This is where we load the library.

```
library(dplyr)

  group_by(mtcars, am) |>
    summarize(avg_wt = mean(wt))
```

```
# A tibble: 2 × 2
     am avg_wt
  <dbl>  <dbl>
1     0   3.77
2     1   2.41
```

Options are:

- sidebar-left   sidebar-right
- overlay-right   overlay-right   overlay-center

Contest

Closeread is running a contest for the best example scrollytelling

Judging criteria will be based on a combination of:

- Educational value or newsworthiness
- Effective use of scrollytelling to deliver the story
- Artistic and design merit
- Technical accomplishment

Awards

- **Honorable mentions**: Posit hex stickers
- **Three runners up**: a free year of Posit Connect Cloud
- **The grand winner**: Posit swag prize pack valued at $200

Deadline: Dec 15

Installing the Closeread Quarto extension

To use Closeread you need to install the Quarto extension by running the following at the terminal tab:

```
quarto add qmd-lab/closeread
```

To get to work you might have to open a new terminal in RStudio using:

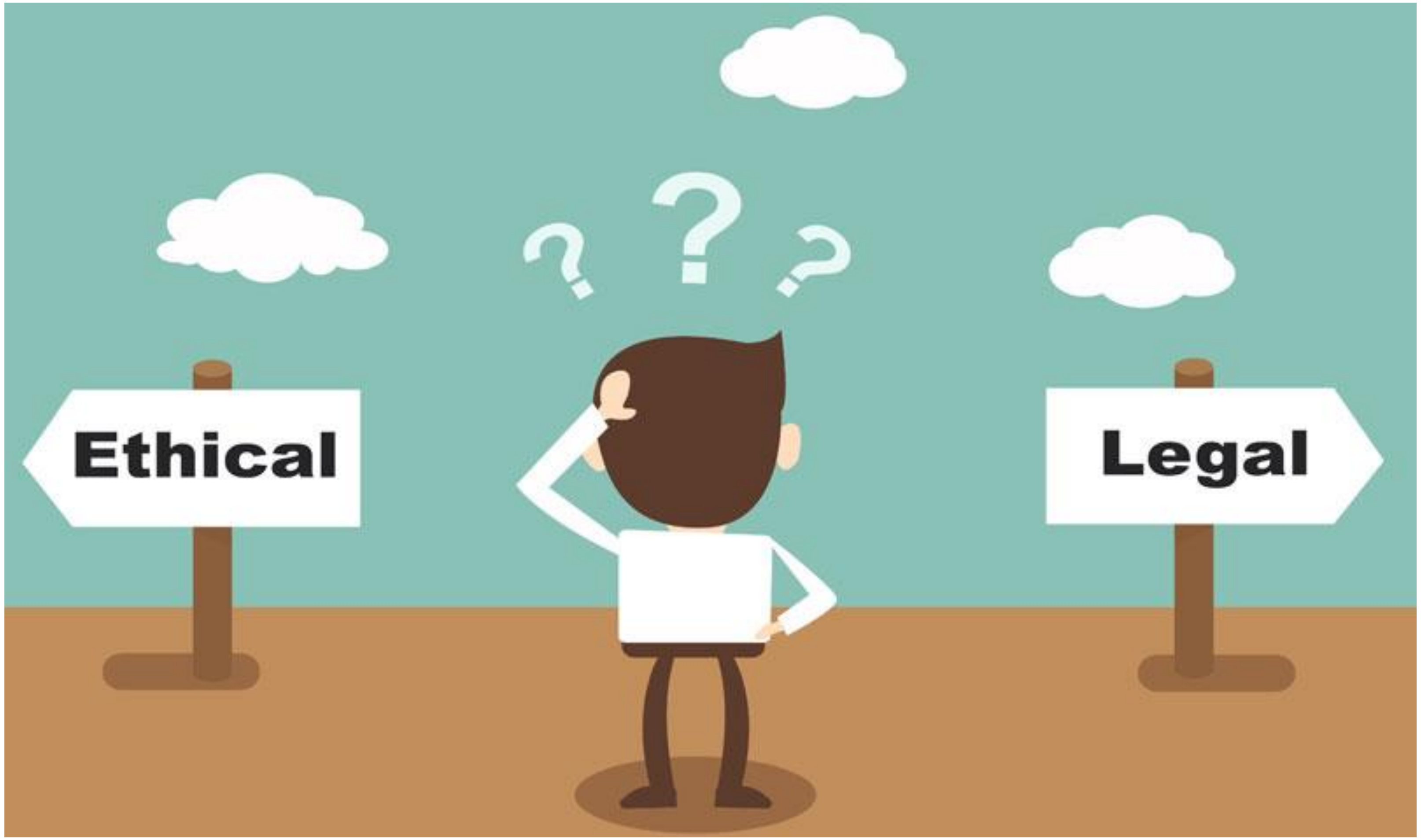
- Tools -> Terminal -> New Terminal

```
SDS230::download_any_file("class_code/closeread.qmd")
```

- (You can cut and paste this from Canvas)

Let’s try it in R!

Ethics in Statistics and Data Science



Ethics in Data Science

Ethics of:

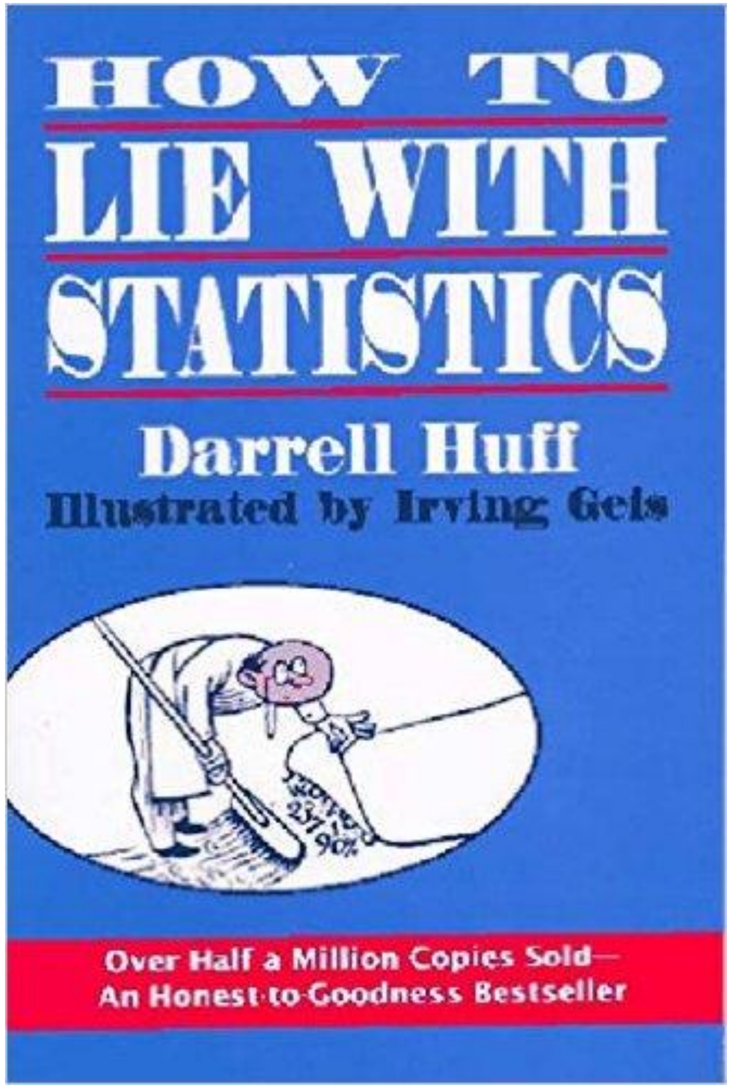
1. Data presentation
2. Using valid data
3. Data scraping TOS and privacy
4. Reproducibility
5. Citations/peer review
6. Disclosure
7. Ethics in Statistical analyses
8. Ethics of creating powerful tools

1. Ethics of data presentation

Data should be displayed in an honest way that gives an accurate picture of trends

Darrell Huff wrote a classic book in the 1950’s pointing out ways that people lie with statistics

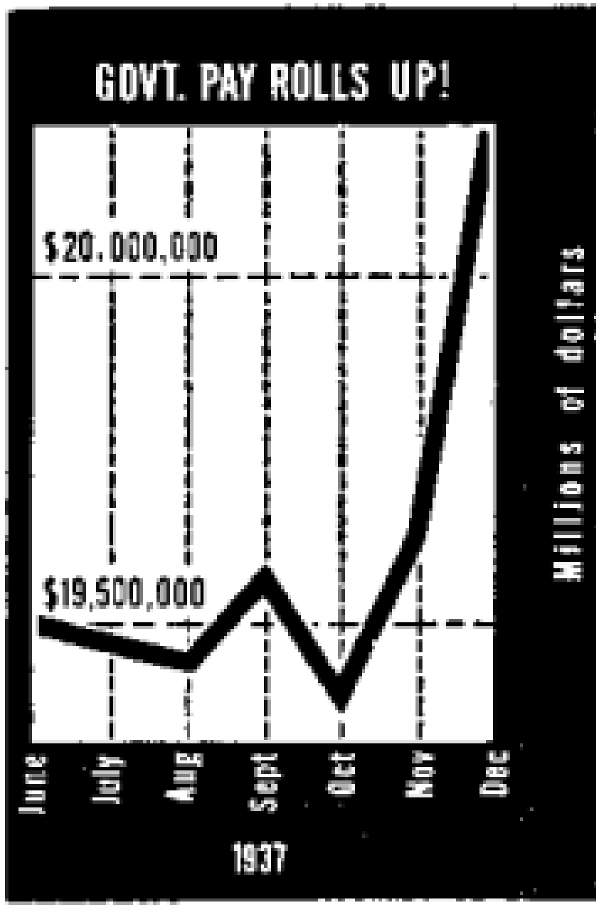
The book was banned as training material at the VA


HOW TO
LIE WITH
STATISTICS
Darrell Huff
Illustrated by Irving Geis
Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller

Ethics of data presentation

What is potentially misleading with this figure?

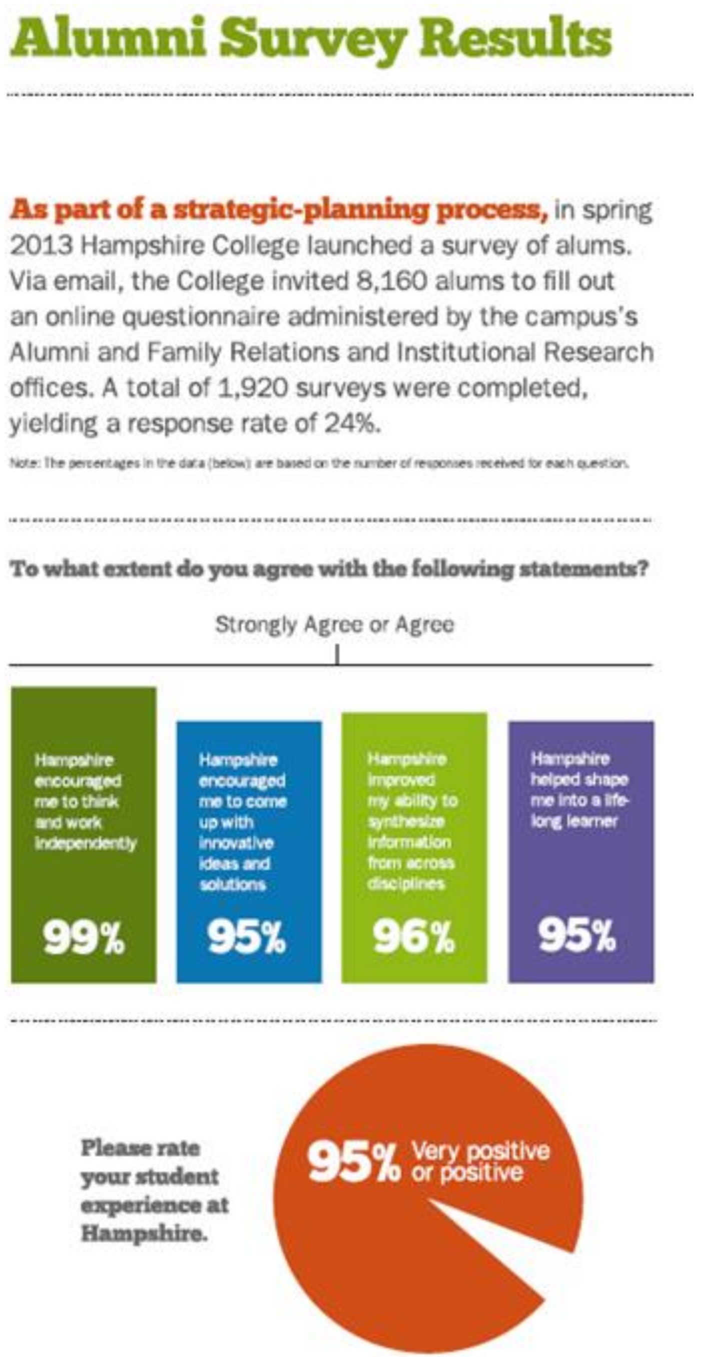
Only a 4% increase in payroll



From a 1938 article in Dun's Review titled 'GOVERNMENT PAY ROLLS UP!'

2. Using valid data

Is almost everyone satisfied with Hampshire College?



Alumni Survey Results

As part of a strategic-planning process, in spring 2013 Hampshire College launched a survey of alums. Via email, the College invited 8,160 alums to fill out an online questionnaire administered by the campus's Alumni and Family Relations and Institutional Research offices. A total of 1,920 surveys were completed, yielding a response rate of 24%.

Note: The percentages in the data (below) are based on the number of responses received for each question.

To what extent do you agree with the following statements?

Strongly Agree or Agree

Hampshire encouraged me to think and work independently	Hampshire encouraged me to come up with innovative ideas and solutions	Hampshire improved my ability to synthesize information from across disciplines	Hampshire helped shape me into a life-long learner
99%	95%	96%	95%

Please rate your student experience at Hampshire.

95% Very positive or positive

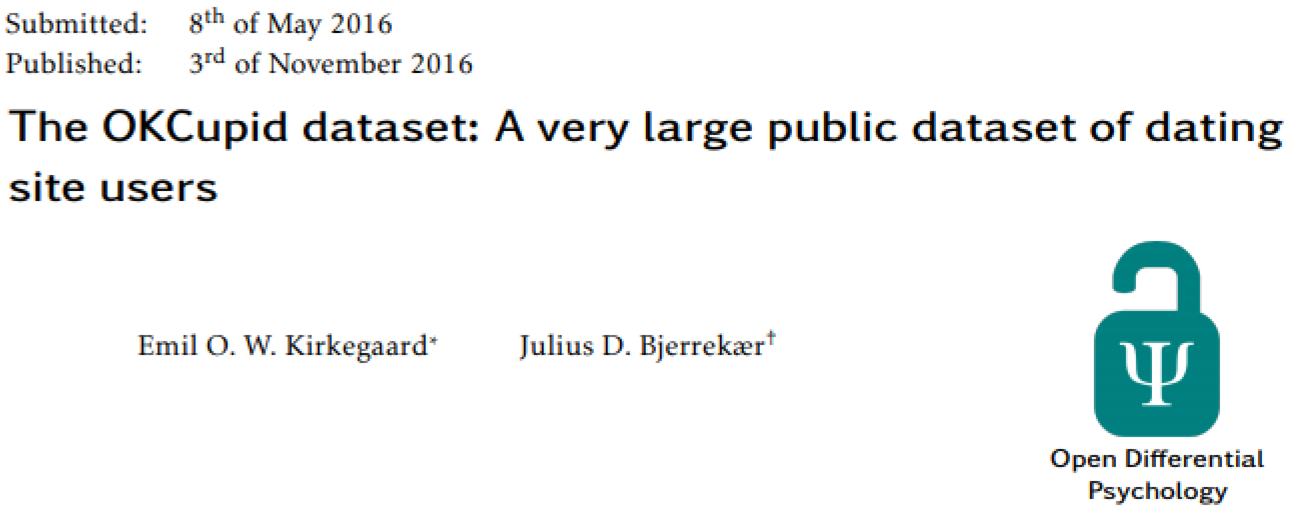
3. Data scraping, terms of service and privacy

Scraping publicly available data is fine (e.g., Wikipedia) but what about scraping data if:

- It violates a website's Terms of Service?
- User privacy?

Kirkegaard and Bjerrekaer scraped okcupid and data on 68,371 users publicly available including usernames, dating preferences, etc.

- Is this ok?



Submitted: 8th of May 2016
Published: 3rd of November 2016

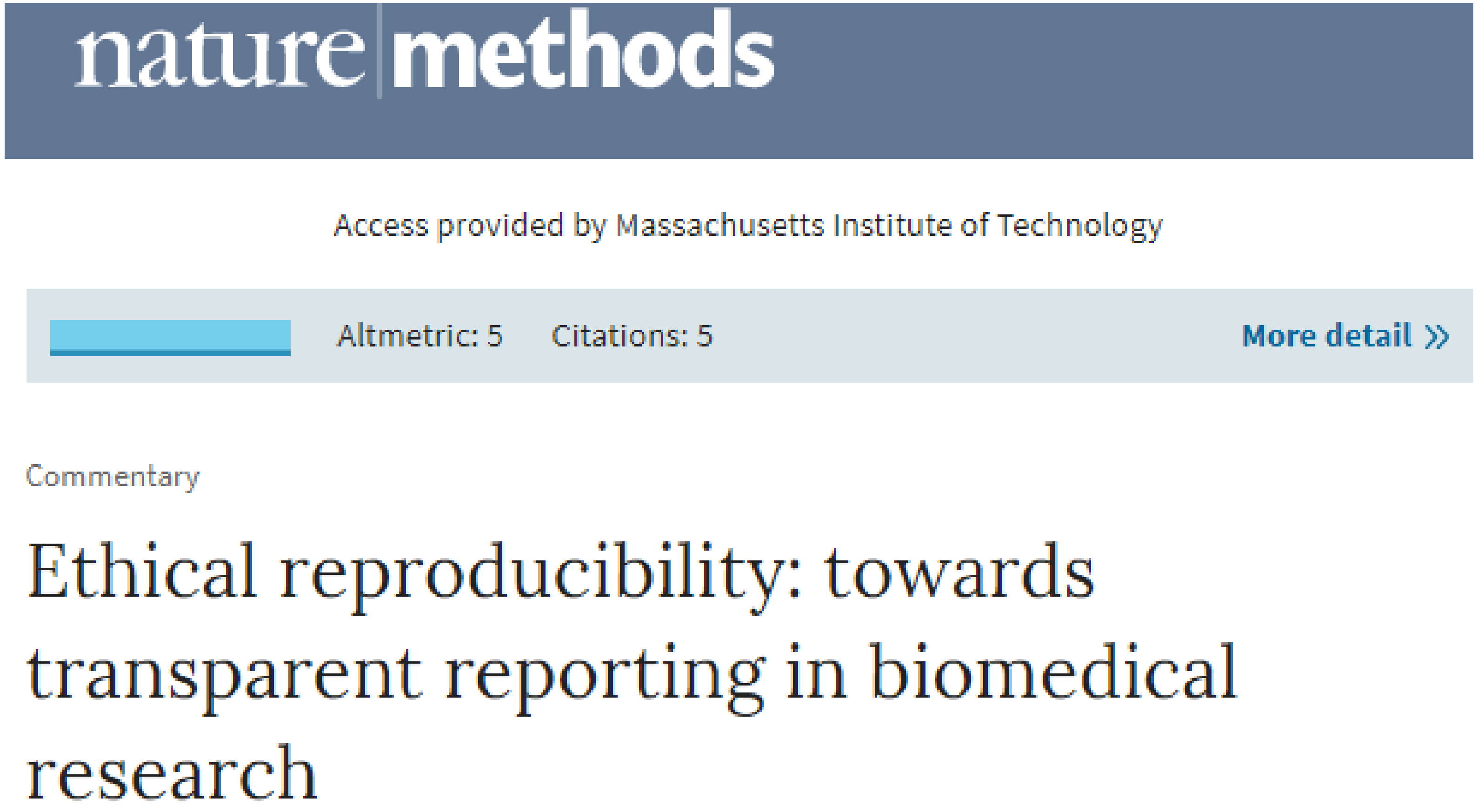
The OKCupid dataset: A very large public dataset of dating site users

Emil O. W. Kirkegaard[*] Julius D. Bjerrekær[†]

Open Differential Psychology

4. Reproducibility

Do scientists have an ethical obligation to make sure their research is reproducible?



Access provided by Massachusetts Institute of Technology

Commentary

Ethical reproducibility: towards transparent reporting in biomedical research

Reproducibility

Do scientists have an obligation to share data/code?

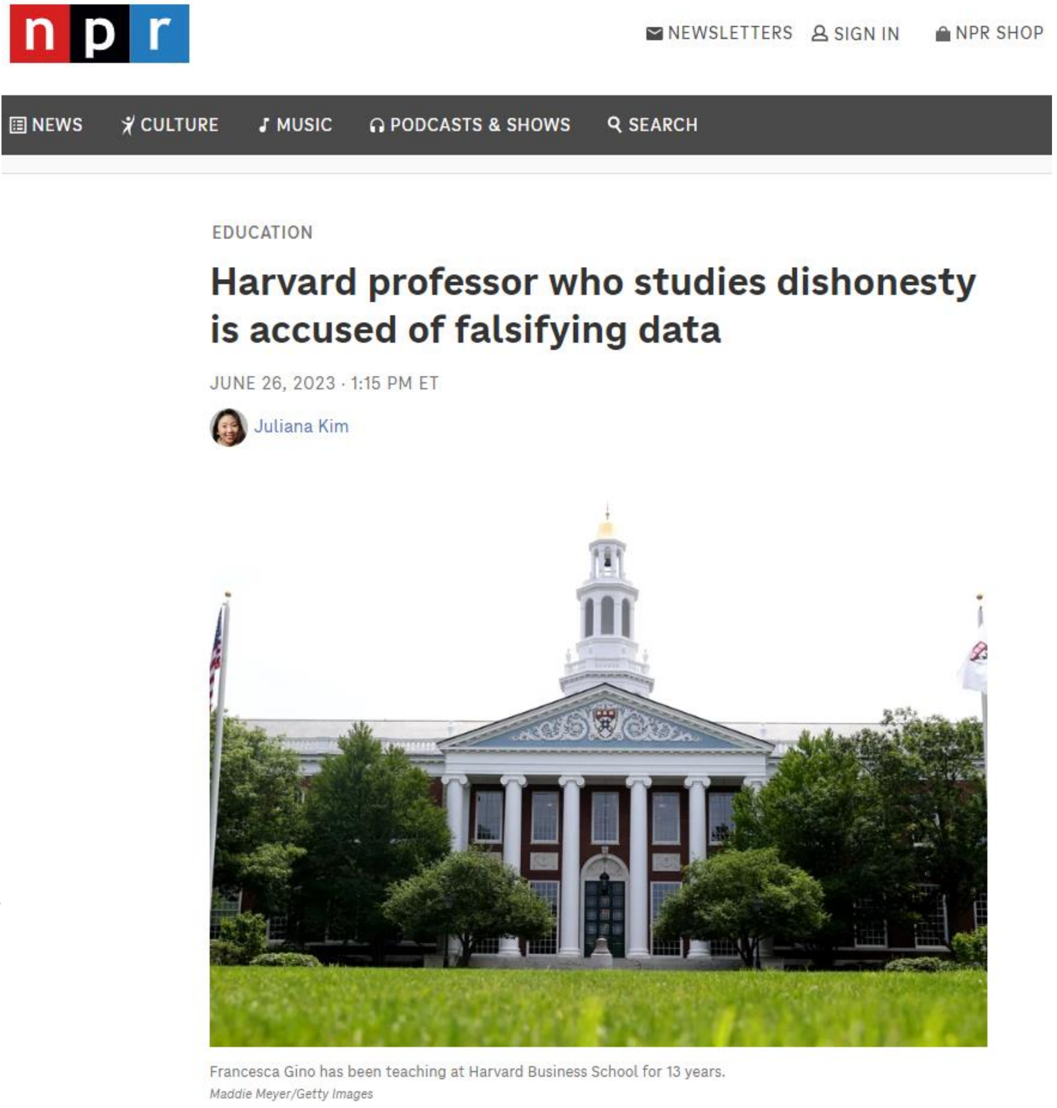
- What if it could hurt your career?
  - Others could prove you wrong, make new findings on your own data, etc.

What should you do if you find one of your papers is wrong?

- You need to retract the paper!

Retraction Watch

DATA COLADA



EDUCATION

Harvard professor who studies dishonesty is accused of falsifying data

JUNE 26, 2023 · 1:15 PM ET

Juliana Kim

Francesca Gino has been teaching at Harvard Business School for 13 years.

Maddie Meyer/Getty Images

5. Citations

If you got an idea from someone else you should always cite their work!

- What is the term for failing to do this?

You should also cite other background work that is relevant

- What if they didn’t cite you?

What about citing someone because they will be a reviewer of your paper?

- How do you deal with someone else’s questionable behavior?

6. Disclosure of conflicts of interest

If you have a conflict of interest you should always disclose it

- Even if you think it doesn’t affect your judgement it might

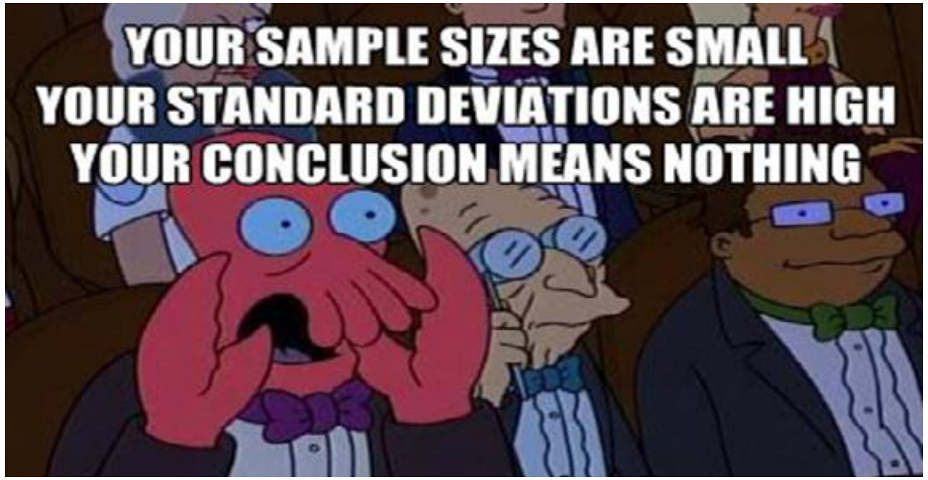
7. Ethics in Statistics

P-hacking (data dredging):

Keep trying different hypothesis tests on a data set until you reach ‘statistical significance’ ($p < 0.05$)

File drawer effect:

- Try a million studies until one is significant


YOUR SAMPLE SIZES ARE SMALL
YOUR STANDARD DEVIATIONS ARE HIGH
YOUR CONCLUSION MEANS NOTHING

DATA COLADA

8. Ethics of creating powerful tools

Some prominent people are concerned about job loss due to machine learning, or even computers posing an existential threat to humans

- Is this something we should be concerned with as Data Scientists?

Ethics in machine learning

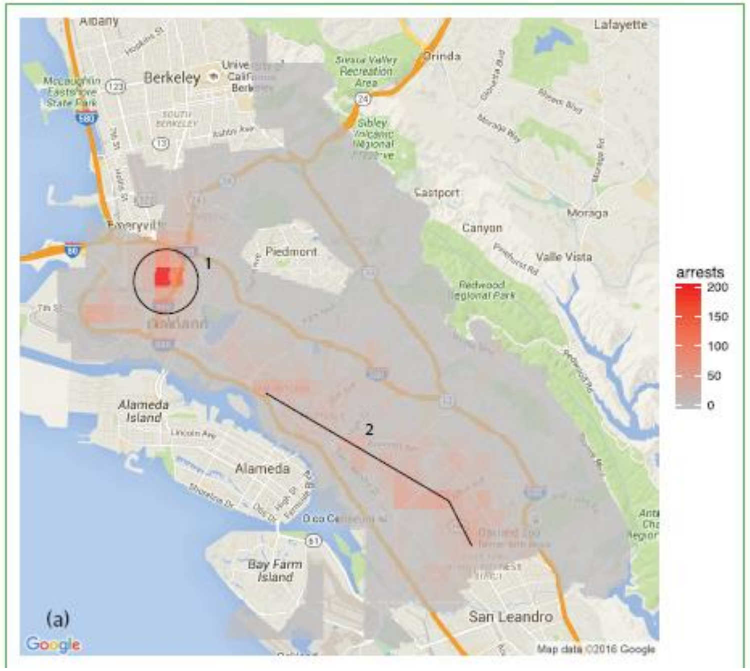
Idea: use ML to police areas with most crimes

- E.g. more police where most drug arrests were made

Possible results

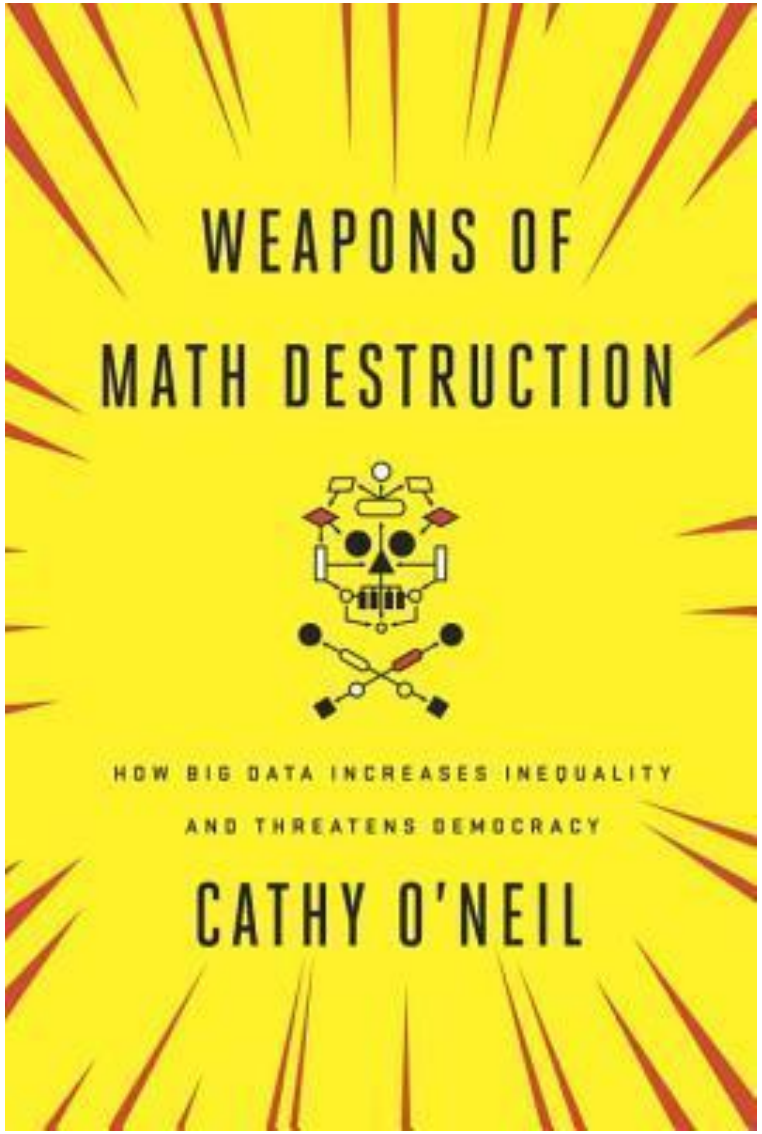
- Higher arrest rates for drugs found in these areas seemingly showing the ML algorithm is working

Any potential problems with this?



Additional reading

https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end



WEAPONS OF MATH DESTRUCTION

HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL

Wrap up and conclusions

Topics we covered

R and descriptive statistics/plots: Base R, fundamental concepts in Statistics

Review confidence intervals: Sampling and bootstrap distributions

Review of hypothesis tests: Permutation and parametric tests, theories of testing

Data wrangling: filtering and summarizing data, joining data sets, reshaping data

Data visualization: grammar of graphics

Regression: simple/multiple, non-linear terms

ANOVA: one-way/factorial, interactions

Statistical learning: cross-validation, logistic regression

Scrollytelling

Course objectives

Extend and solidify concepts and method learned in intro stats

Learn how to use the R programming language to analyze, visualize and wrangle data

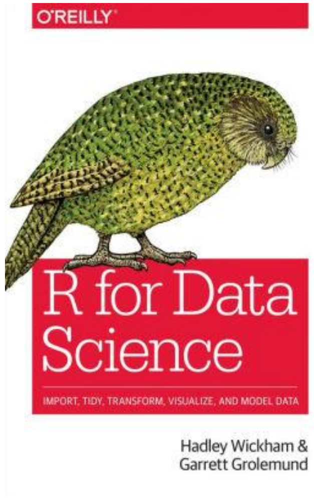
Gain experience extracting insights from real data

Learn how to find patterns in a large noisy data sets and convincingly convey the results to others!

Next steps

Take more advanced Statistics and Data Science classes offered at Yale!

There are many good online resources to learn more R



Good luck with the end of the semester!

Good luck finishing your final projects!

Exam review session on Tuesday December 10th at 4pm

- In this classroom

The final exam is on Tuesday December 17th at 9am

- In SPL 59