Data Visualization

First Steps: Reproducible Data Visualization

Ciara Zogheib

Data Sciences Institute, University of Toronto

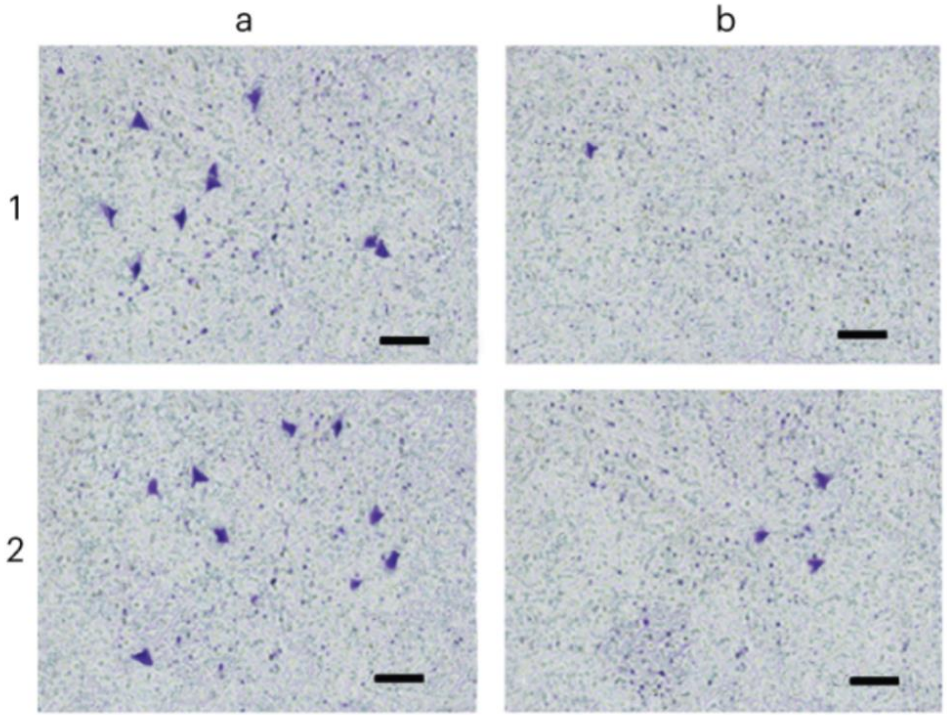
We’re going to...

- Explore why reproducibility in data visualization matters
- Understand why reproducible data visualization practices are both ethical and practical
- Discuss practices we can implement to make our data visualizations reproducible

Case Study: Why should we care about reproducible data visualization?

Image Manipulation

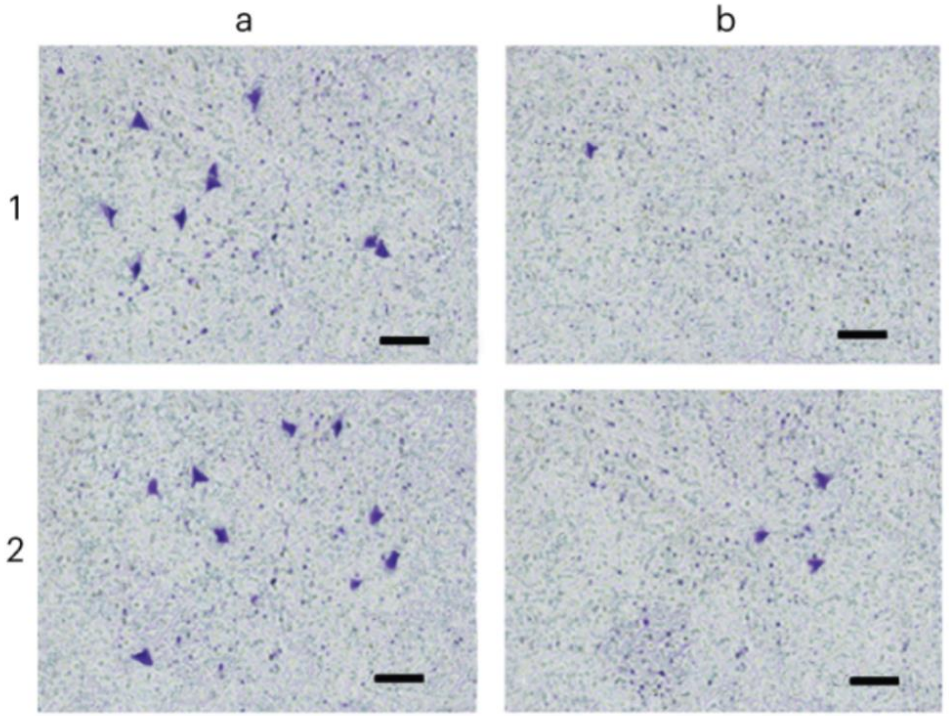
- In 2016, a research article showing the effectiveness of a particular molecular compound as a potential cancer treatment was published
- Cell staining images were used to support authors' conclusions



(Shen, 2020)

Image Manipulation

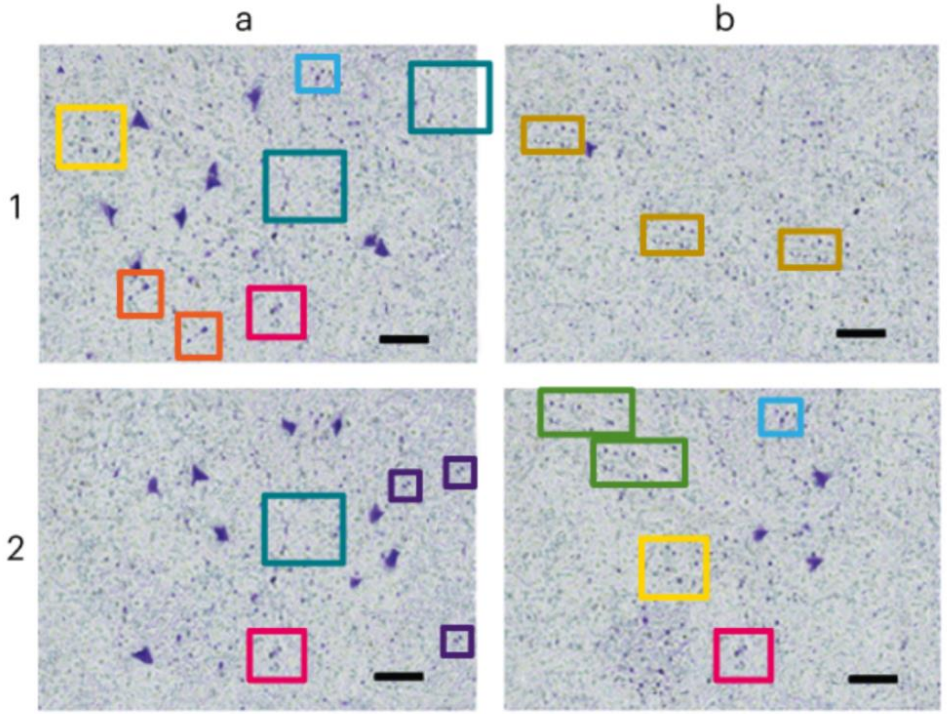
- Dr. Elisabeth Bik, a microbiologist, noticed that certain features of the published images contained “problematic duplications”
- **Can you spot the duplications in the original images?**



(Shen, 2020)

Image Manipulation

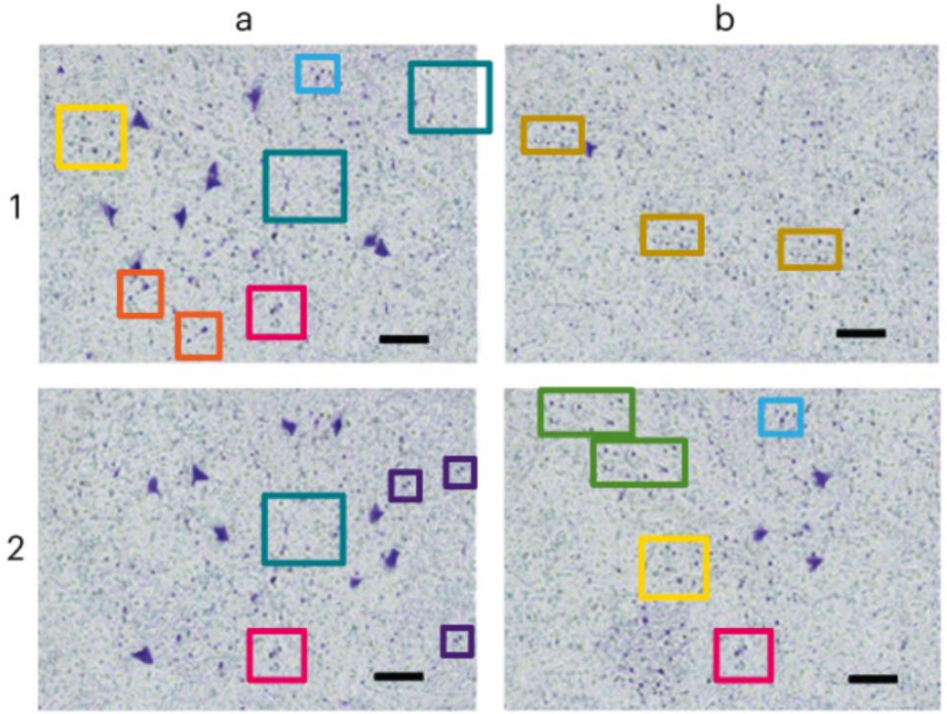
- Areas enclosed in same-coloured boxes show signs of deliberate duplication
- If the images were manipulated, can we still trust the authors' conclusions?



(Shen, 2020)

Image Manipulation

- The paper was ultimately retracted because of Dr. Bik's findings
- Study authors said that the images were generated by a third party company whose involvement was not declared in the initial publication



(Shen, 2020)

Image Manipulation - Dr. Elisabeth Bik

- Dr. Bik's work spotting manipulated figures and images in scientific publications has led to 172 retractions and more than 300 corrections
- She and her colleagues have examined >20,000 biomedical publications and found that over 3.8% contain problematic figures, and that numbers of these figures are rising

(Shen, 2020; Bik, 2016)

Why is this so important?

- Images and visualizations are often the best way to communicate our data and insights...
- **But** if we can't trust that visualizations are representing the 'real' data, we can't trust what their creators are trying to communicate...
- **So** how can we ensure that the visualizations we create are representing our data with integrity?

Reproducibility

What is reproducibility?

- Reproducible work is “capable of being checked because the data, code, and methods of analysis are available to other researchers”
  - That is: someone could repeat the steps we took to generate a particular result or image from our data
- The figures in the case study we saw were **not** reproducible, because they were made by an unknown third party using unclear methods
- Reproducibility of data is a hot topic across professional and academic research contexts, and increasingly a requirement for publication

(Allison et al., 2016)

Reproducibility is ethical

- The American Statistical Association's Ethical Guidelines for Statistical Practice state that "Good statistical practice is fundamentally based on transparent assumptions, **reproducible results**, and valid interpretations"
- Recall from last class that visualizing data means making decisions
- Reproducibility helps to hold us accountable for those decisions

(ASA, 2018)

Reproducibility is practical

- Making data visualizations reproducibly is **practical** as well as **ethical:**
  - Makes it easier to make changes if we have to edit a plot or image weeks or months after its creation
  - Helps us to draw on previous work to make new graphics more easily
  - Useful for version control because we can easily see where exactly we made changes

(Healy, 2018)

NOTE:

The ability to reproduce a result does not necessarily indicate correctness, nor does the inability to do so mean a result is incorrect.

But “science is incremental: it is only through transparency and by enabling reproducibility that scientific knowledge evolves.”

(Fekete & Freire, 2020)

So how can we make our data visualizations reproducible?

Work programmatically

- We want to do as much work on our figures as possible programmatically
- This means we want to make our images in code (eg. ggplot in R) rather than in programs like Adobe Illustrator, where changes and data sources are harder to trace

(Healy, 2018)

Work in plain text

- Code should be written in a simple, plain-text format (eg. R scripts or .txt files)
- Code should **not** be written in a word processor (eg. Microsoft Word)
- Ideally, our ‘pretty’ final products (images, graphs, charts) can be procedurally (and reproducibly!) generated just by running our code

(Healy, 2018)

Comment your code

- When others (or us, at a later date) want to look back at the code we used to make our data visualizations, comments can help us to make sense of what choices were made and why
- Comments make code easier to understand, maintain, and update

(Steidl et al., 2013; Healy, 2018)

Activity: Comment our code

- Let’s return to our saved code from practicing making ggplots in R
- We can add comments to our file by preceding them with ‘#’
- Take a few minutes to comment your code, and then return and discuss what you did
  - What information is it helpful to include?
  - How do you write a ‘good’ comment?

Datasheets for datasets

- Gebru et al. (2020) propose that datasets be accompanied by a datasheet that “documents its motivation, composition, collection process, recommended uses”, etc
- Datasheets can facilitate connections between the underlying data and the final analytical products (in our case, images and figures)

(Gebru et al., 2018)

Datasheets for datasets - Sample datasheet

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity: given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. It was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided though five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are movie reviews extracted from newsgroup postings, together with a sentiment rating for whether the text corresponds to a review with a rating that is either strongly positive

these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example "negative polarity" instance, taken from the file `neg/cv452_tok-18656.txt`.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and alter fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

Is there a label or target associated with each instance? If so, please provide a description.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe

(Gebru et al., 2018)

Next...

- Graphing our Data
- Building a deeper understanding of how ggplot works
- Making new kinds of graphs, better preparing our data for plotting
- Choosing the right graph for a given situation (Professional skills)