Chapter 8: Data Ethics

- Ethics in data science and statistical analysis is critical to maintaining the trust of your clients and producing analyses that reveal scientific truth.
- The book mentions three tenets of the *Hippocratic Oath* traditionally followed by physicians that are especially relavant to data scientists:
  1. "I will not be ashamed to say 'I know not,' nor will I fail to call in my colleagues when the skills of another are needed for a patient's recovery."
  2. "I will respect the privacy of my patients, for their problems are not disclosed to me that the world may know."
  3. "I will remember that I remain a member of society, with special obligations to all my fellow human beings, those sound of mind and body as well as the infirm."
- Some ethical guidelines for data scientists are founded in legal responsibilities.
- Others are principles and guidelines suggested by professional societies (like the American Statistical Association).

History of Data Ethics

- Ethical considerations in statistical practice have been considered for many years.
- Famous 1954 book: *How to Lie with Statistics* by Darrell Huff.
- Warned about using graphics to present actual data in a deceptive way.
- Similar book from 1983: *How to Tell the Liars from the Statisticians* by Robert Hooke.
- Discussed only graphics but how improper statistical reasoning can produce deceptive conclusions in various contexts.

Misleading Graphics

- An unethical data scientist can use graphics to give the wrong impression of what can be learned from accurately recorded data. Some ways to produce deceptive graphics include:
  1. Using poorly chosen or unconventional axis labels and values
  2. Using color to draw attention to the wrong information in the data
  3. Presenting a misleading ordering of data values
  4. Choosing an inappropriate graphic for the type of data
- See Figures 8.1, 8.2, and 8.3 for some example of misleading plots. What makes them deceptive?
- Other examples of misleading graphs are seen in the Moore and Notz book *Statistics: Concepts and Controversies*.

Figure 8.1

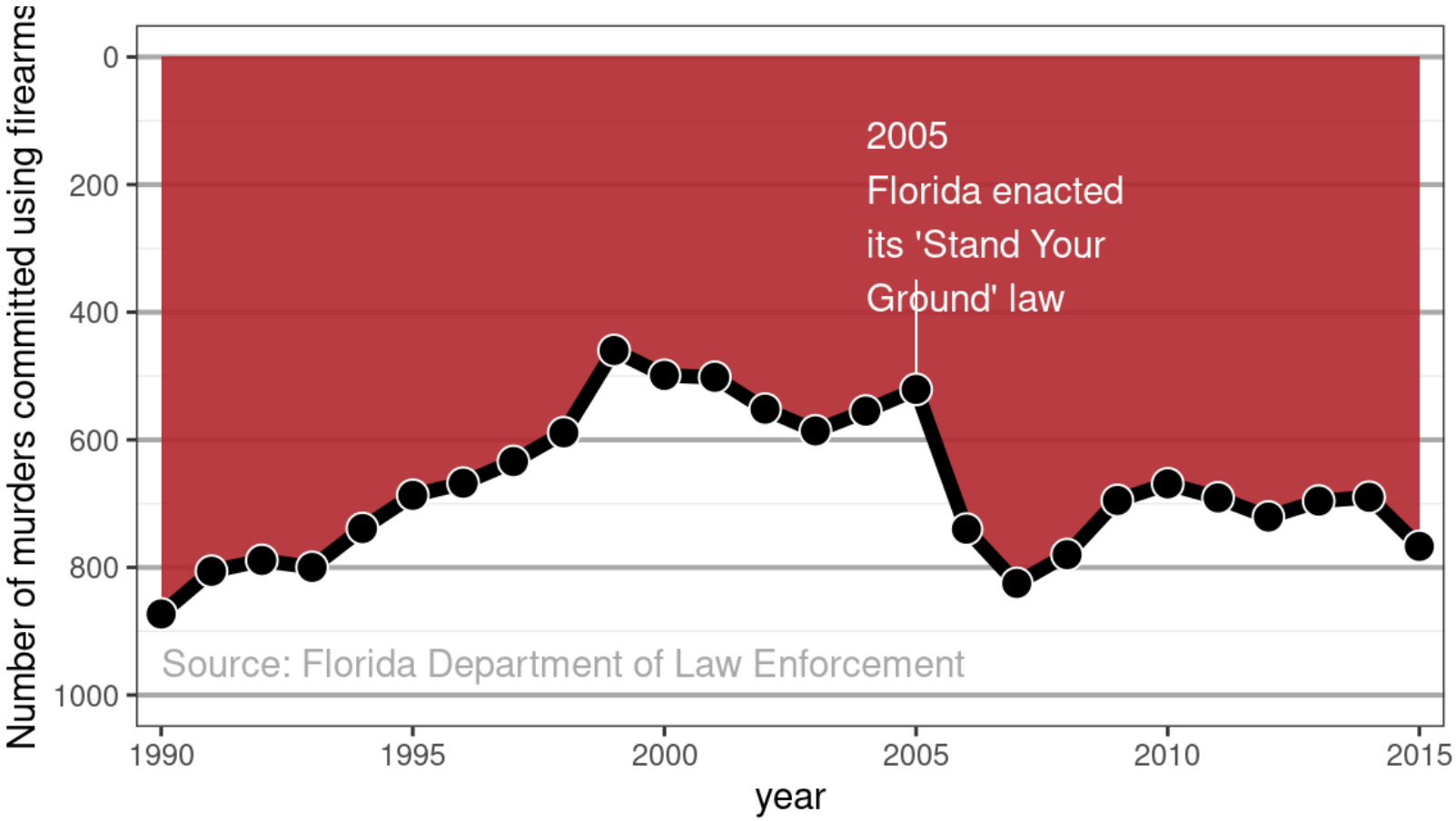

Figure 8.1 from MDSR textbook

Figure 8.2

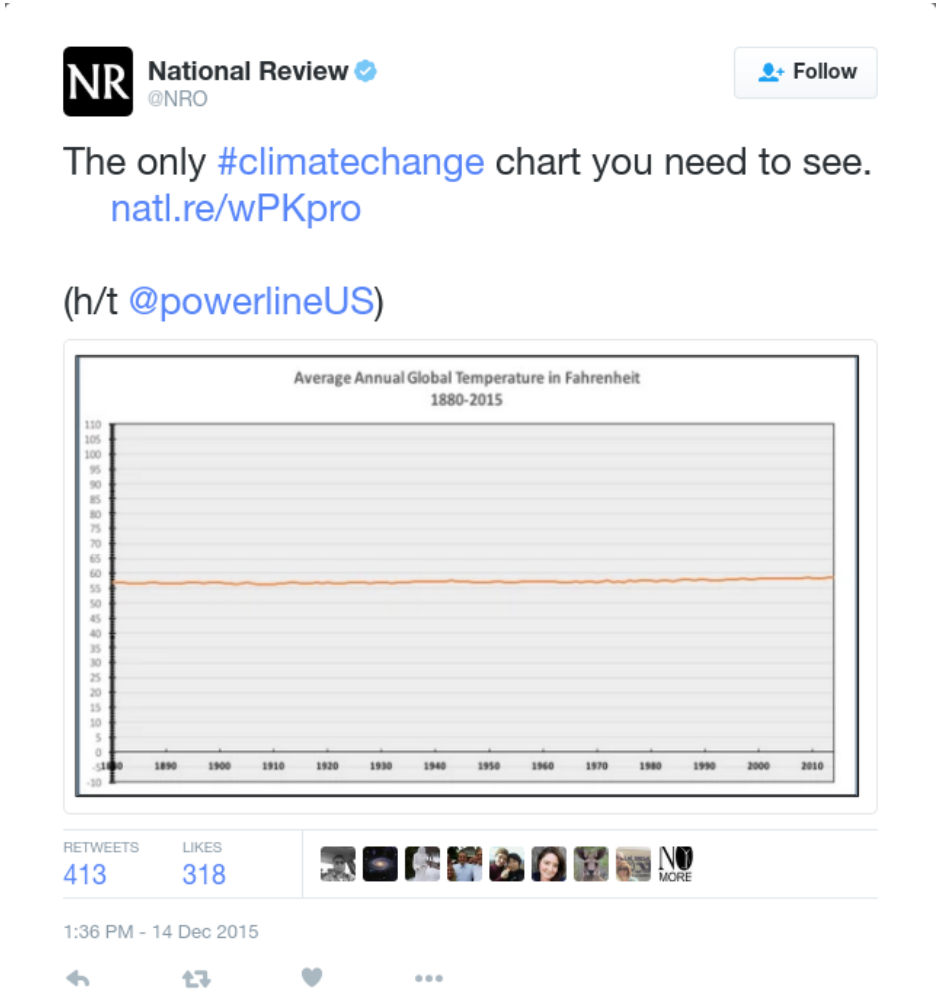

Figure 8.2 from MDSR textbook

Figure 8.3

Top 5 Counties with the Greatest Number of Confirmed COVID-19 *(Reproduction of Figure)*

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.

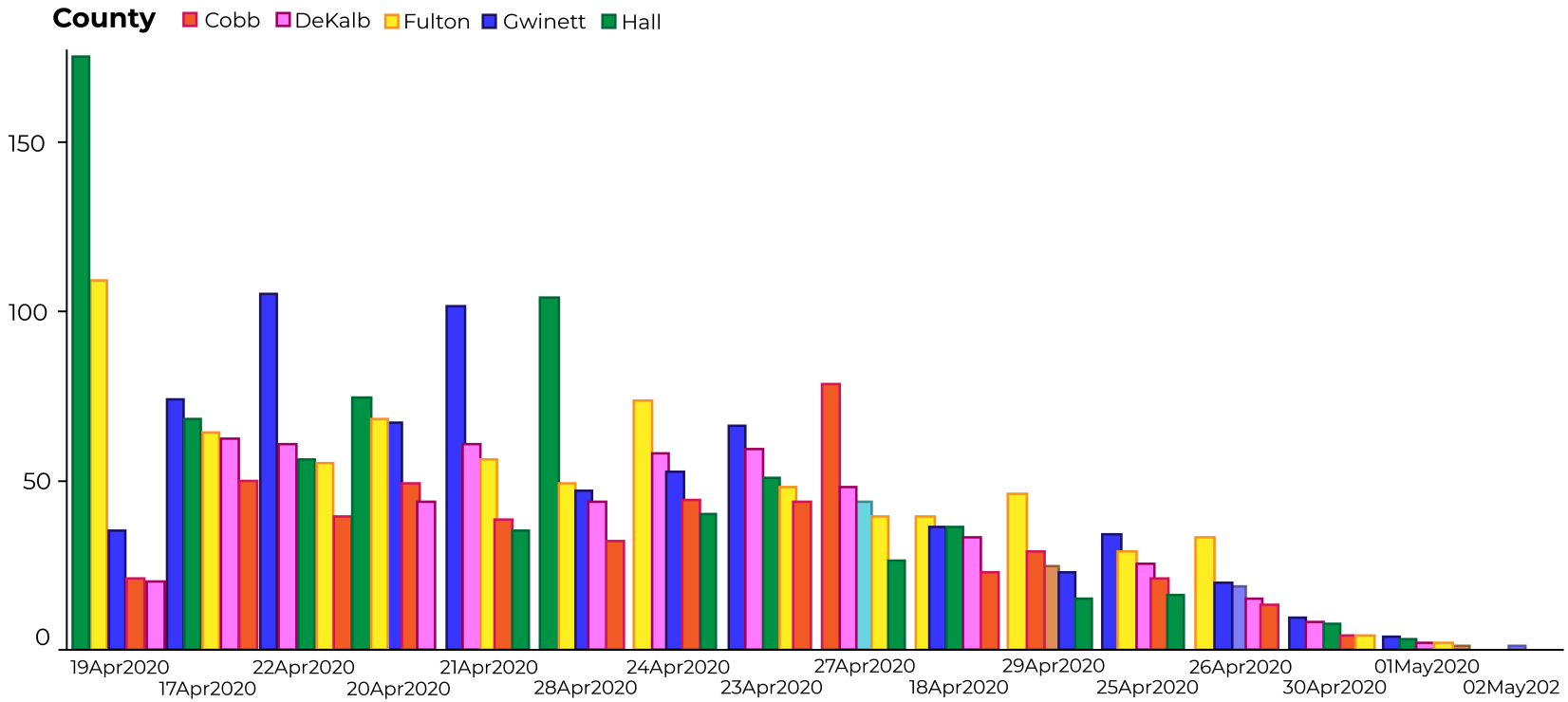

Figure 8.3 from MDSR textbook

Importance of Correct Data Reporting

- Data scientists have a responsibility to report data correctly, and not in a biased way to favor one particular conclusion.
- Old saying:

Importance of Correct Data Reporting

- Data scientists have a responsibility to report data correctly, and not in a biased way to favor one particular conclusion.
- Old saying: "It's wrong to use statistics like a drunk uses a lamppost: For support rather than for illumination."
- In the three examples previously, the data reporting may be a large effect on public opinions about important policy issues:
  1. Firearm laws in the wake of the death of Trayvon Martin
  2. Belief in climate change, especially among consumers of conservative media
  3. Reopening plans during the first few months of the COVID pandemic

Some Real Data Ethics Situations

- Some specific real ethical situations described in Section 8.4:
- CEO wants statisticians to "fudge" some coefficients to better represent company's values
- Employers submitting hiring data could be accused of discrimination by the U.S. Office of Federal Contract Compliance Programs (OFCCP) because of questionable classification method used by OFCCP
- Researchers use photos from dating site to train model to predict sexual orientation
- Researchers use last name (or last name and address) to predict a person's race based on Census data (see R examples)

More Real Data Ethics Situations

- More specific real ethical situations described in Section 8.4:
- Scraping data (2620 variables, including usernames, gender, and dating preferences for 68K people) from an OKCupid dating site and making it a publicly accessible data file
- Publishing a flawed analysis about austerity and economic growth (several data analysis mistakes were found thanks to requests for spreadsheet of data)
- Study showing Vioxx had increased risk of heart attacks after FDA had approved the drug to reduce severe gastrointestinal events
- Statisticians on both sides of legal trials asked to estimate damages — major incentive to estimate in direction of client!

Principles to Guide Ethical Action: Data Science Oath

- Three principles in the *Data Science Oath* published by the National Academy of Sciences in 2018:
  1. I will not be ashamed to say, "I know not," nor will I fail to call in my colleagues when the skills of another are needed for solving a problem.
  2. I will respect the privacy of my data subjects, for their data are not disclosed to me that the world may know, so I will tread with care in matters of privacy and security.
  3. I will remember that my data are not just numbers without meaning or context, but represent real people and situations, and that my work may lead to unintended societal consequences, such as inequality, poverty, and disparities due to algorithmic bias.

Principles to Guide Ethical Action: Data Values and Principles (1-6)

1. Use data to improve life for our users, customers, organizations, and communities.
2. Create reproducible and extensible work.
3. Build teams with diverse ideas, backgrounds, and strengths.
4. Prioritize the continuous collection and availability of discussions and metadata.
5. Clearly identify the questions and objectives that drive each project and use to guide both planning and refinement.
6. Be open to changing our methods and conclusions in response to new knowledge.

Principles to Guide Ethical Action: Data Values and Principles (7-12)

7. Recognize and mitigate bias in ourselves and in the data we use.
8. Present our work in ways that empower others to make better-informed decisions.
9. Consider carefully the ethical implications of choices we make when using data, and the impacts of our work on individuals and society.
10. Respect and invite fair criticism while promoting the identification and open discussion of errors, risks, and unintended consequences of our work.
11. Protect the privacy and security of individuals represented in our data.
12. Help others to understand the most useful and appropriate applications of data to solve real-world problems.

Ethical Principles in the Eight Case Studies

- Each of the eight case studies in Section 8.4 relate to one or more of the 12 guiding principles presented here. For example:
- CEO pushback: Principles 8 and 12
- Employment Discrimination: Principle 10
- Prediction of Sexual Orientation: Principles 1, 3, 7, 9, and 11
- Prediction of Race: Principles 3, 7, and 9.

More Ethical Principles in the Eight Case Studies

- Data scraping of OKCupid members: Principles 1 and 11
- Reproducible Spreadsheet: Complied with Principle 10, but maybe not with Principle 2
- Vioxx dangers: Principle 6
- Statisticians on legal cases: Principle 8, but note difference between role on a *legal team* in an adversarial situation and role as an expert witness

Algorithmic Bias

- Many data science methods rely on algorithms which are fairly opaque and difficult for the user to fully understand.
- These are used in many high-tech applications, such as determining medical treatments, evaluating criminals being considered for parole, allocating policing resources, even the operation of self-driving cars.
- If input data contains some bias against a demographic group, this bias could be reinforced since these models get updated in part based on *feedback* from their predictions.
- Even if variables like race are not explicitly part of the prediction model, *proxy variables* like number of interactions with police or family history of arrest may be.
- Great book on algorithmic bias and fairness: *Weapons of Math Destruction* by Cathy O'Neil.
- Data scientists should be aware of potential algorithmic biases to identify and counteract bias and maximize fairness.

Data Confidentiality and Disclosure Issues

- It's important to ensure that personal data is not disclosed, even unintentionally.
- Some data sets without identifying personal information have enough details that can be linked with other data sets in order to identify the subjects.
- HIPAA regulations prohibit making personal medical information public.
- Sometimes broad geographic identifiers (state or territory of the subject) can be published to help answer research questions.
- For rare diseases or conditions, even broad information can be enough to identify the subject.
- Those with access to personal information must use reasonable safeguards to prevent intentional or unintentional disclosure.

Data Storage and Data Scraping

- Sometimes protected data is inadvertently accessible on the Internet, causing harm to those whose sensitive information is stolen.
- Safe computing and good database management can prevent this problem.
- Other times, companies allow some public usage of their data, with certain restrictions.
- These include terms of use that prohibit using data for direct marketing.
- Some companies restrict how much data you can scrape or how frequently you can access the data.
- Publishers sometimes restrict text data mining (note that LLMs like ChatGPT are typically trained on text published on the Internet).

Reproducibility Issues

- For an alarming amount of research, the analysis cannot be replicated later by others or even by the same researchers.
- This is often because of preprocessing or preliminary analysis like sorting, filtering cases, or selecting columns that were done via menu-driven steps rather than by code that would reproduce those steps.
- In addition, the results and graphics given in a report are often disconnected from the statistical analysis that created those numbers and graphs.
- The reader must take it on faith that the numbers and graphs presented are correct — usually they are, but there is no way to detect if there's an incorrect table or plot.

How to Ensure Reproducibility

- The best way to make research reproducible is to record and make available every step in the process, from data organization to preprocessing to statistical analysis to creation of graphics.
- Researchers should make available:
  1. All data files in their original form
  2. Metadata/codebooks helping to understand the data
  3. Computer code for extracting and transforming the data and performing analyses, fitting models, generating graphics
  4. A file that connects the computer output to the results in the report
- Tools like *R Markdown* and `knitr` are powerful helpers to achieve these goals — see Appendix D of the textbook for more details.
- *Quarto* is a newer software that some people now prefer to R Markdown.

Data Merging

- In Chapter 5, we discussed how to merge two data tables (e.g., with `inner_join`).
- It can be hard to know whether data tables have been merged incorrectly, especially with large tables.
- That makes it important to keep all the code used to merge the tables.
- Book gives example of research paper showing a link between immune response and depression that was not really true: The conclusion came about because the lab results and some survey data had been merged improperly.

Collective Ethics

- Ethical issues are the responsibility of not just individual researchers, but of the whole scientific community.
- The "publish-or-perish" paradigm for academic researchers can incentivize people to act unethically, and trust in science suffers (Stanford president's research on Alzheimer's).
- Publication bias: The need to get *statistically significant results* before journals will publish the research can lead to the "file-drawer problem".
- Imagine 100 parallel research studies to establish some significant association of interest (but the association doesn't really exist). About how many will produce significant findings at the 0.05 level?

Multiple Testing Issues

- Multiple testing problem: When you do many simultaneous inferences on the same data set using formal hypothesis tests, you really must adjust the tests to account for the multiple tests, or else the probability of Type I error will be inflated.
- Report *all* the tests you were doing, not just the ones that yielded significant results.
- This is a major issue in observational studies when many variables are being measured and the researchers are "looking for which one(s) have an effect" on some outcome variable.
- Section 9.7 in the textbook has some excellent related discussion on the "perils of p-values" and the "garden of forking paths".

Statistical Studies with Human Subjects

- Studies with human subjects (like drug trials, for example) must be approved in advance by an *institutional review board*.
- Subjects must be warned about potential side effects and give their *informed consent* ahead of time.
- All information about subjects must be kept confidential.
- Researchers should weigh the future benefits of the experiment against the welfare of the subjects involved (Tuskegee syphilis study example).
- For clinical trials, issues of when to stop the trial once the drug has proved effective arise as ethical questions.