EMORY
UNIVERSITY
SCHOOL OF
MEDICINE

Department of
Biomedical Informatics

BMI 500

Introduction to Biomedical Informatics

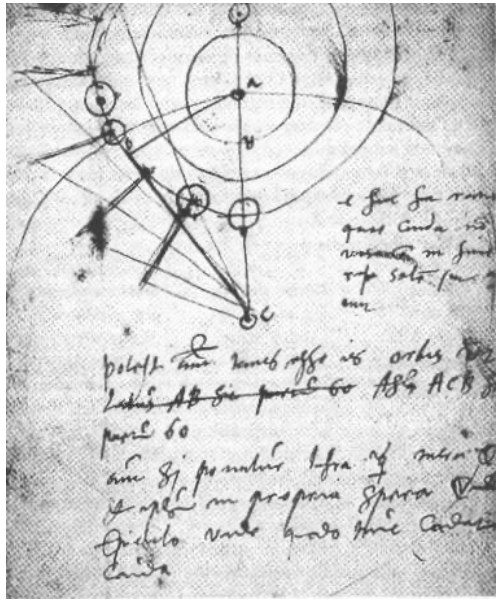
Lecture 4: (Better) Data Treatment

https://tinyurl.com/bmi500

14 September 2022

Matthew Reyna
Department of Biomedical Informatics, Emory University, Atlanta, GA, USA

Data Treatment

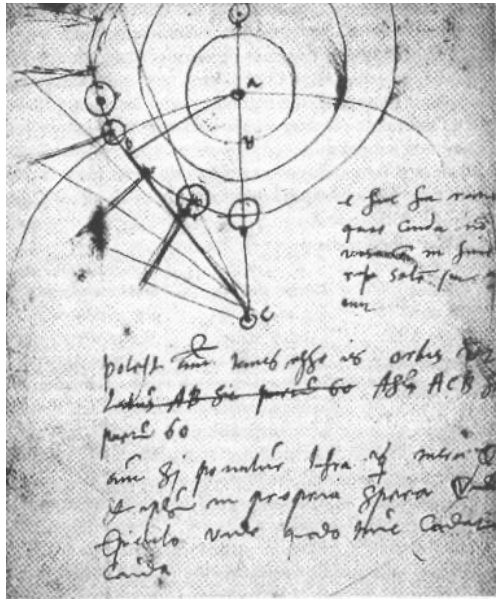


Tycho Brahe's observations of the great comet of 1577. See https://en.wikipedia.org/wiki/Tycho_Brahe.

Information, often numeric, collected through observation

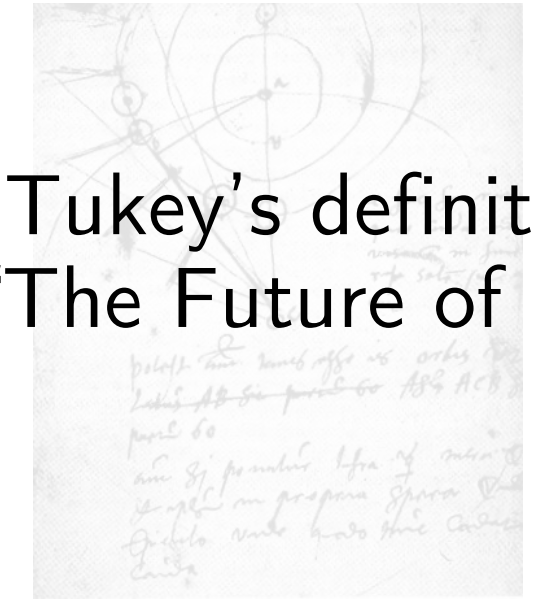
Storage, processing, analysis, interpretation, summarization, reporting, and other uses

Data Treatment



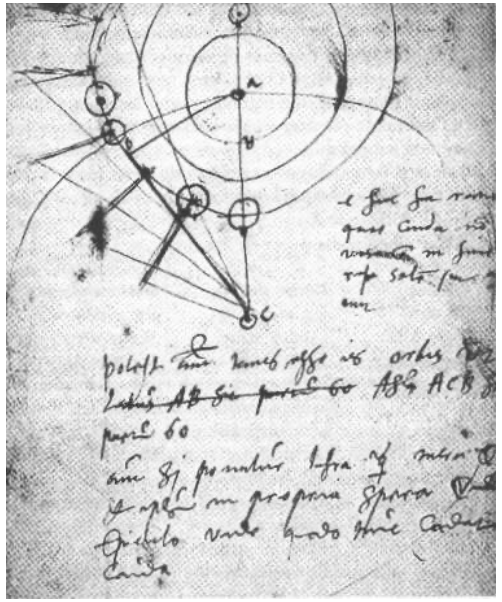
"Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

Data Treatment

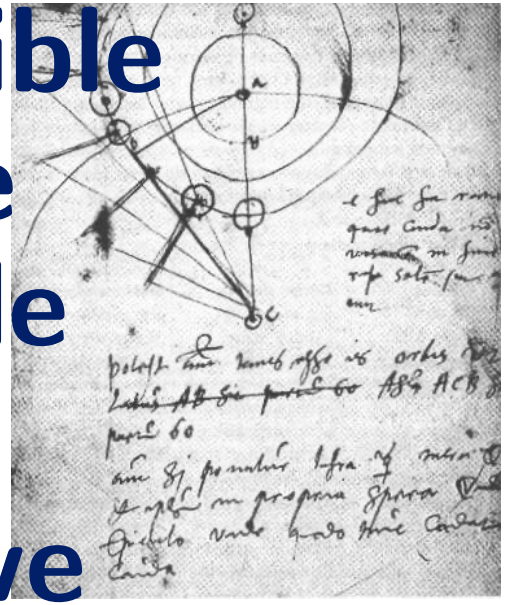


John Tukey's definition of data analysis in "The Future of Data Analysis" (1961)

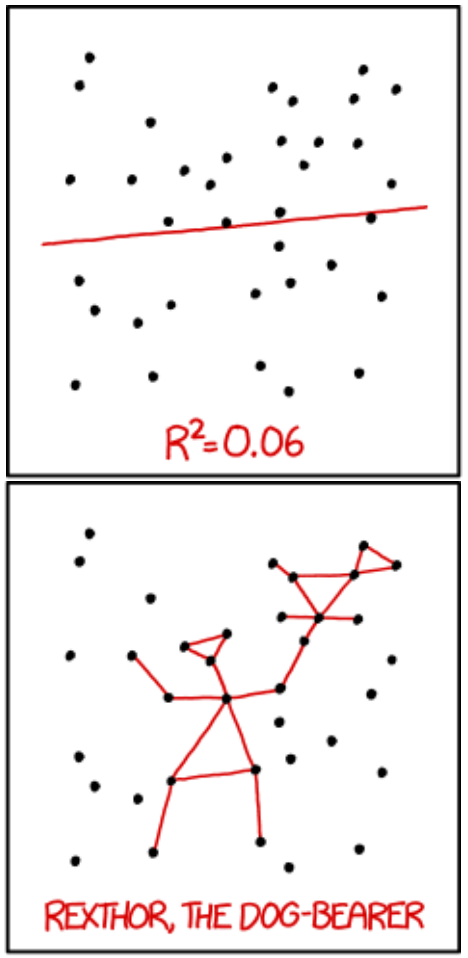
Better Data Treatment

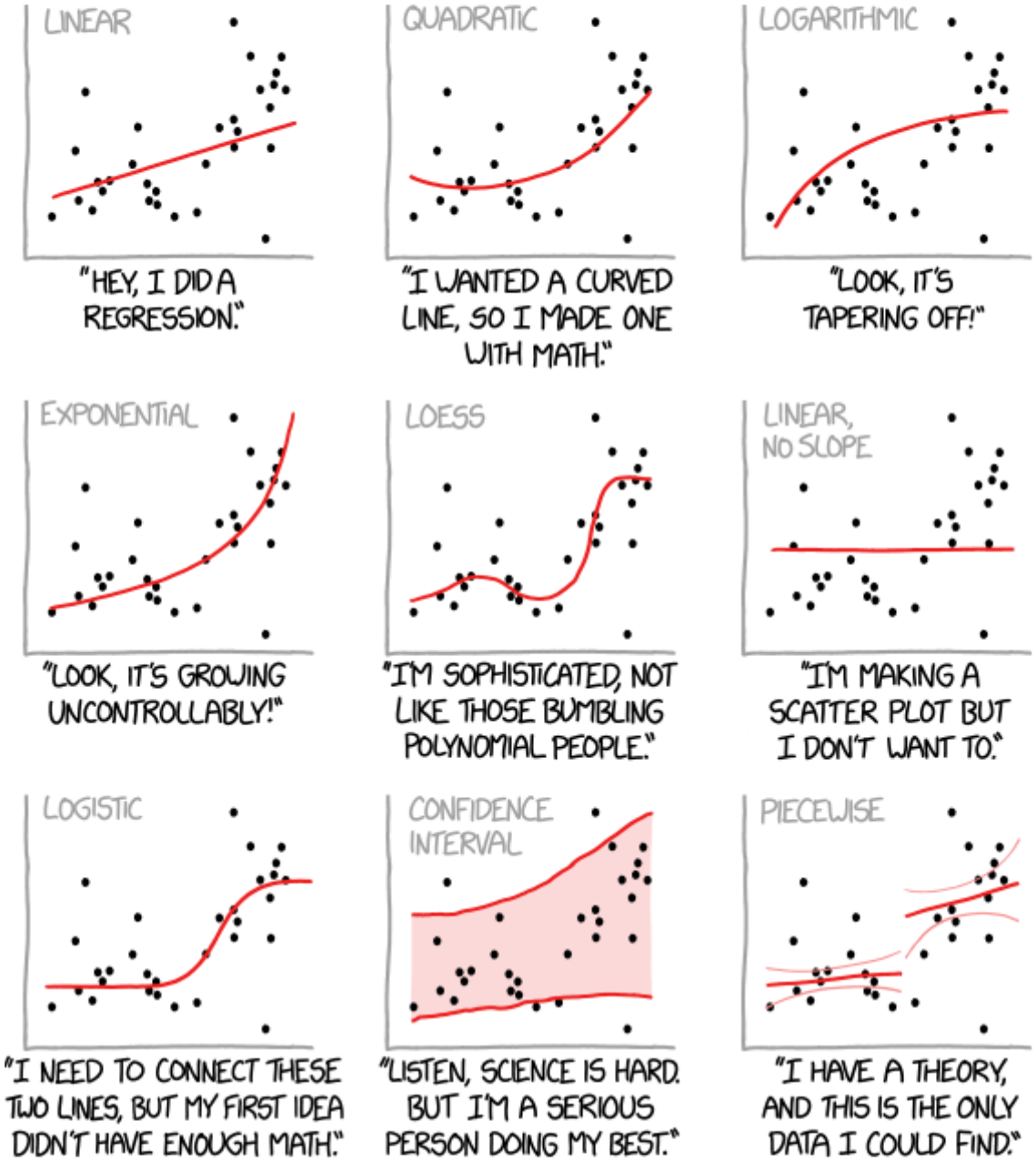


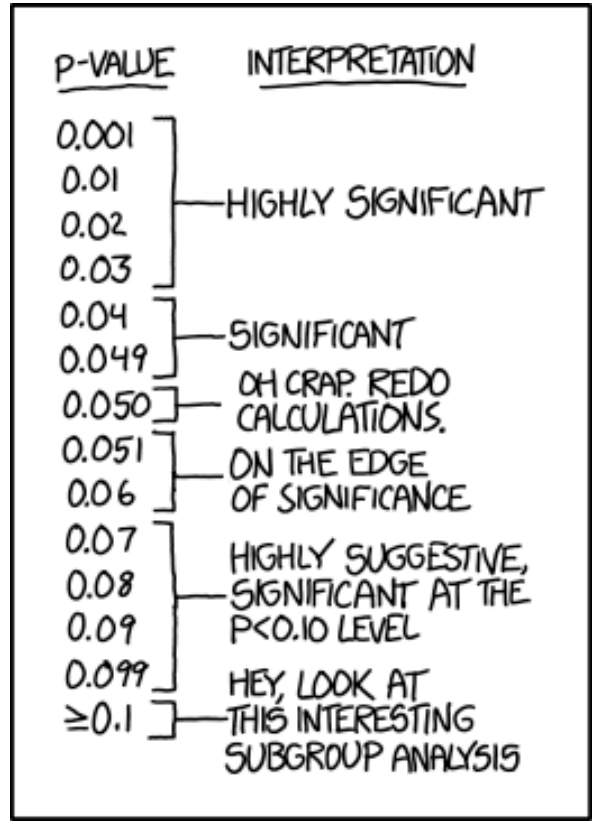
Ethical
Responsible
Accurate
Effective
Objective
Better Data Treatment
Reproducible
Replicable
Repeatable
Quick
Inexpensive



We know that good data treatment is important...







Randall Munroe and XKCD
https://xkcd.com/1725/
https://xkcd.com/2048/
https://xkcd.com/1478/

… but bad data treatment is still common.

Sections The Harvard Crimson

More than 60 Fall CS50 Enrollees Faced Academic Dishonesty Charges

By Hannah Natanson, Crimson Staff Writer

May 3, 2017

The New York Times

Online Cheating Charges Upend Dartmouth Medical School

The university accused 17 students of cheating on remote exams, raising questions about data mining and sowing mistrust on campus.

By Natasha Singer and Aaron Krolik

Published May 9, 2021 Updated June 10, 2021

The New York Times

Top Cancer Researcher Fails to Disclose Corporate Financial Ties in Major Research Journals

By Charles Ornstein and Katie Thomas

Sept. 8, 2018

FiveThirtyEight abc NEWS

Politics Sports **Science & Health** Economics Culture

APR. 7, 2016, AT 2:00 PM

How Two Grad Students Uncovered An Apparent Fraud — And A Way To Change Opinions On Transgender Rights

By Christie Aschwanden and Maggie Koerth-Baker

Filed under Scientific Method

THE NEW YORKER

HOW A GAY-MARRIAGE STUDY WENT WRONG

By Maria Konnikova May 22, 2015

The New York Times Magazine

When the Revolution Came for Amy Cuddy

As a young social psychologist, she played by the rules and won big: an influential study, a viral TED talk, a prestigious job at Harvard. Then, suddenly, the rules changed.

By Susan Dominus

Oct. 18, 2017

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands

RESEARCH ARTICLE

Questionable research practices in ecology and evolution

Hannah Fraser[1]*, Tim Parker[2], Shinichi Nakagawa[3], Ashley Barnett[1], Fiona Fidler[1,4]

1 School of BioSciences, University of Melbourne, Parkville, VIC, Australia, **2** Biology Department, Whitman College, Walla Walla, WA, United States of America, **3** School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW, Australia, **4** School of Historical and Philosophical Studies, University of Melbourne, Parkville, VIC, Australia

* hannahsfraser@gmail.com

Abstract

We surveyed 807 researchers (494 ecologists and 313 evolutionary biologists) about their use of Questionable Research Practices (QRPs), including cherry picking statistically significant results, *p* hacking, and hypothesising after the results are known (HARKing). We also asked them to estimate the proportion of their colleagues that use each of these QRPs. Sev-

REPORT

Evaluating replicability of laboratory experiments in economics

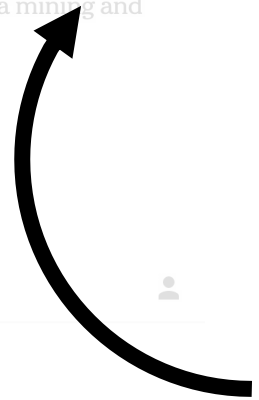
Colin F. Camerer[1,*,†], Anna Dreber[2,†], Eskil Forsell[2,†], Teck-Hua Ho[3,4,†], Jürgen Huber[5,†], Magnus Johannesson[2,†], Michael …

+ See all authors and affiliations

Science 25 Mar 2016:
Vol. 351, Issue 6280, pp. 1433-1436
DOI: 10.1126/science.aaf0918

… but bad data treatment is still common.

"'I P-hacked like crazy all through my time at Princeton, and I still couldn't get interesting results,' [Professor Joseph] Simmons says."



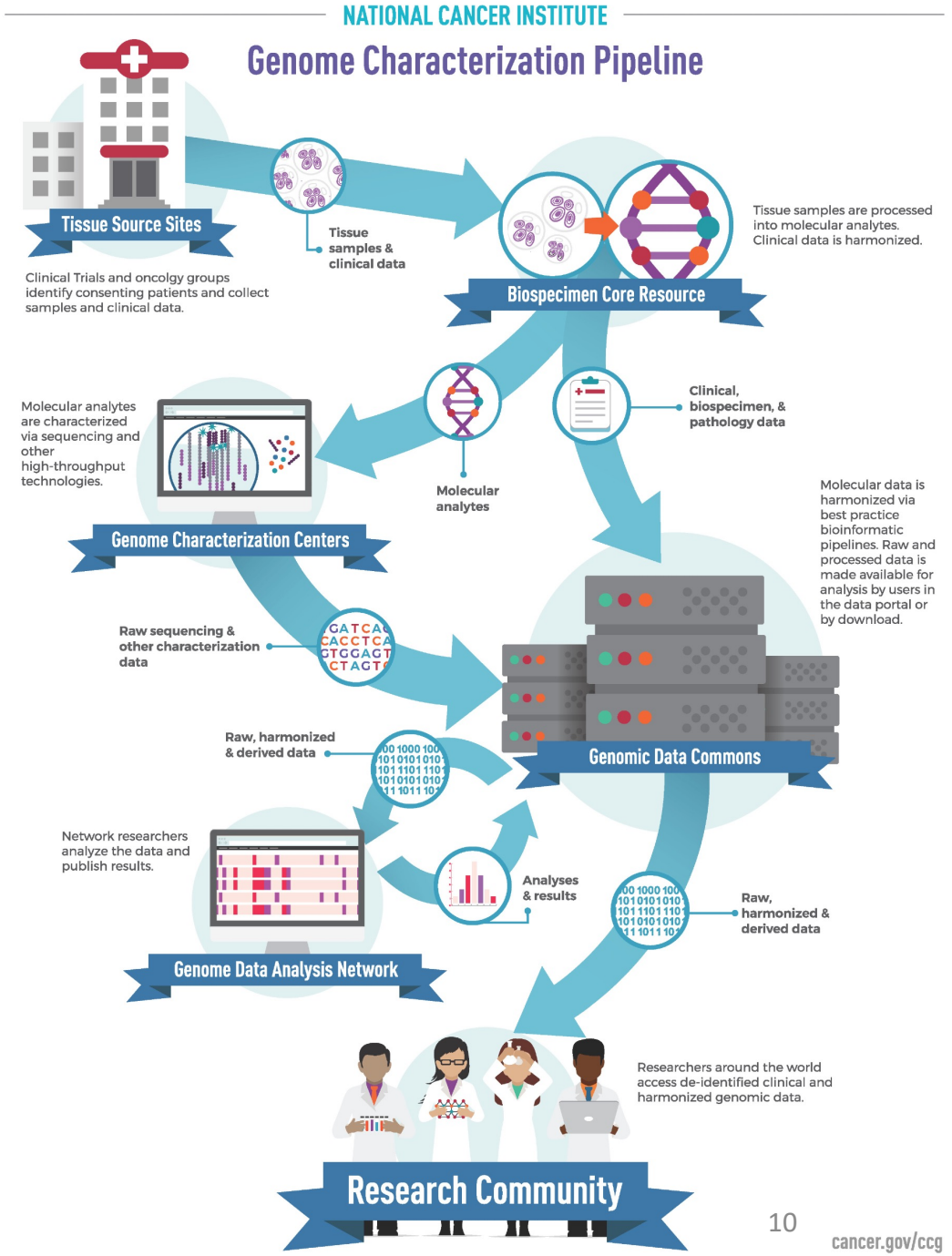
Better data treatment

Each type of data and each research area has its own best practices, and each researcher develops their own preferences.

We can't possibly cover all **best** practices, but we can discuss **better** practices for data treatment that are common to different types of data and different research areas.

During today's lecture, we will **skim**:

- data acquisition, storage, security, and privacy
- data processing, including repeatability, replicability, and reproducibility
- statistics
- figures and tables.



Better data treatment

This is (only) a **single** lecture.

Want to learn more?

- Take courses.
- Read articles and attend talks.
- Read (reputable) websites.
- Talk to your advisor, other lab members, other faculty, and/or other trustworthy individuals.

Better practices, a healthy suspicion of data and code, and good-faith attempt to do the right thing go a long way toward better data treatment.

Calling Bullshit:
Data Reasoning in a Digital World

Logistics

Course: INFO 270 / BIOL 270. University of Washington
Next offered: Autumn Quarter 2019
Credit: 3 credits, graded
Enrollment: To be determined
Instructors: Carl T. Bergstrom and Jevin West
Synopsis: Our world is saturated with bullshit. Learn to detect and defuse it.

Learning Objectives

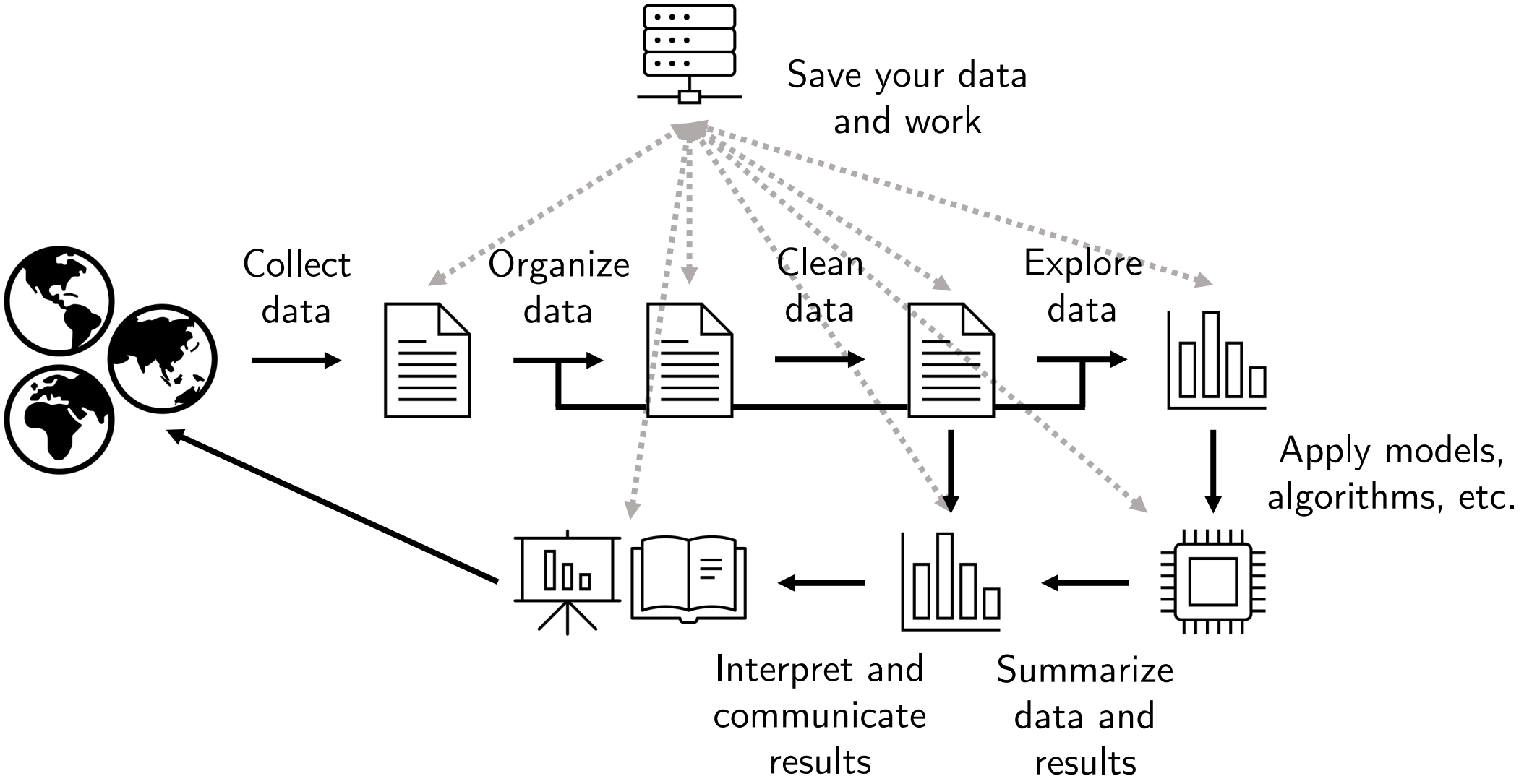
Our learning objectives are straightforward. After taking the course, you should be able to:

- Remain vigilant for bullshit contaminating your information diet.
- Recognize said bullshit whenever and wherever you encounter it.
- Figure out for yourself precisely *why* a particular bit of bullshit is bullshit.
- Provide a statistician or fellow scientist with a technical explanation of why a claim is bullshit.
- Provide your crystals-and-homeopathy aunt or casually racist uncle with an accessible and persuasive explanation of why a claim is bullshit.

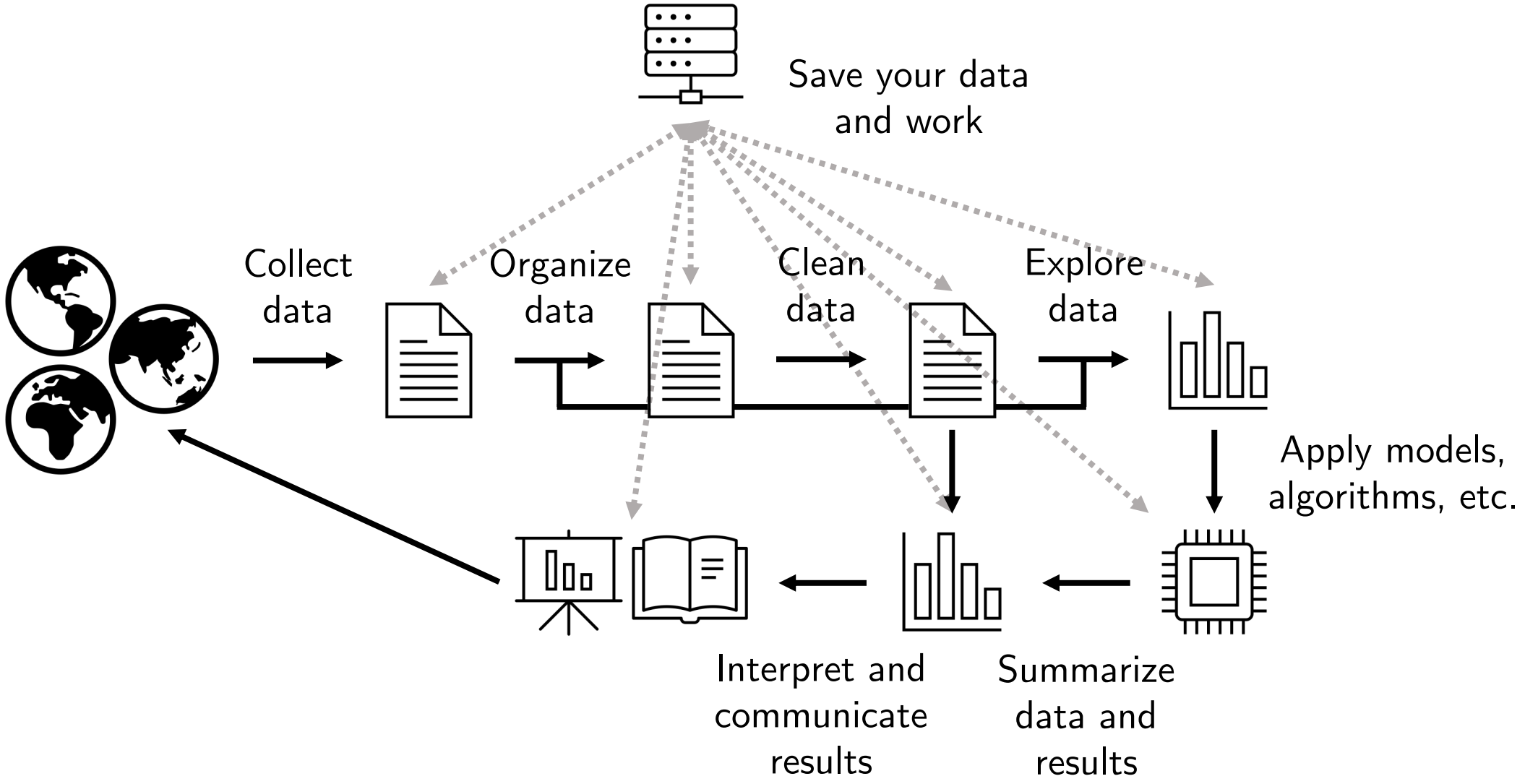
We will be astonished if these skills do not turn out to be among the most useful and most broadly applicable of those that you acquire during the course of your college education.

callingbullshit.org

A generic data treatment pipeline



A generic data treatment pipeline



You should build an ***automated*** data treatment pipeline that checks and transforms the raw results into the statistics, tables, and figures. Shell scripts and notebooks? Great. Manual steps? Not reproducible.

Data collection

Different fields can have **very** different standards for data collection, and **how** you collect the data and **what** data you collect **will** affect your results, so **ask your advisor and be careful!**

Humans? Be careful!

Tiny humans? Be very careful!

Dinosaurs? Awesome, and be careful!

Primates? Be careful!

Vertebrates? Be careful!

Animals? Be careful!

Plants? Be careful!

Alive or "alive"? Be careful!

Biological? Be careful!

Laboratory? Be careful!

Physical? Be careful!

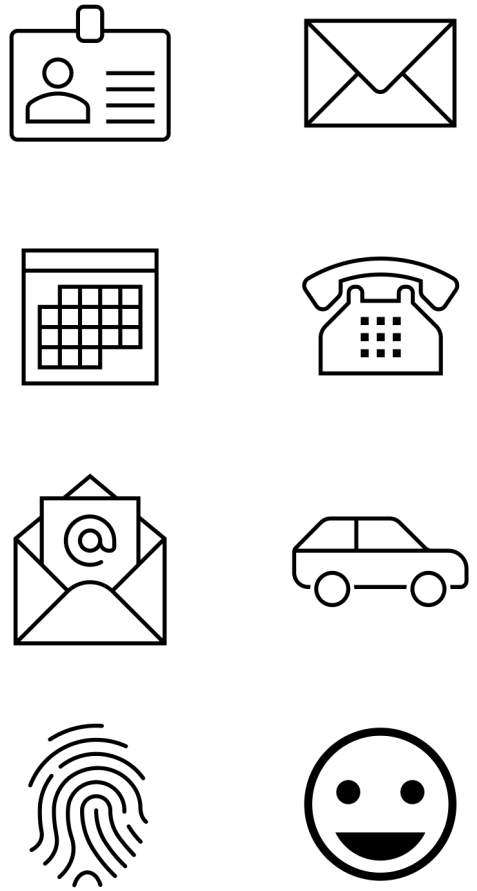
Digital? Be careful!

Something else? Be careful!

From last week: the HIPAA Safe Harbor method

1. Names
2. Addresses (including parts of some ZIP codes)
3. Dates (including some years and ages)
4. Telephone numbers
5. Fax numbers
6. Email addresses
7. Social security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers
13. Device identifiers and serial numbers
14. Web URLs
15. IP addresses
16. Biometric identifiers (fingerprint, voice, etc.)
17. Full-face photographs
18. Other unique identifiers, except for allowed identifier

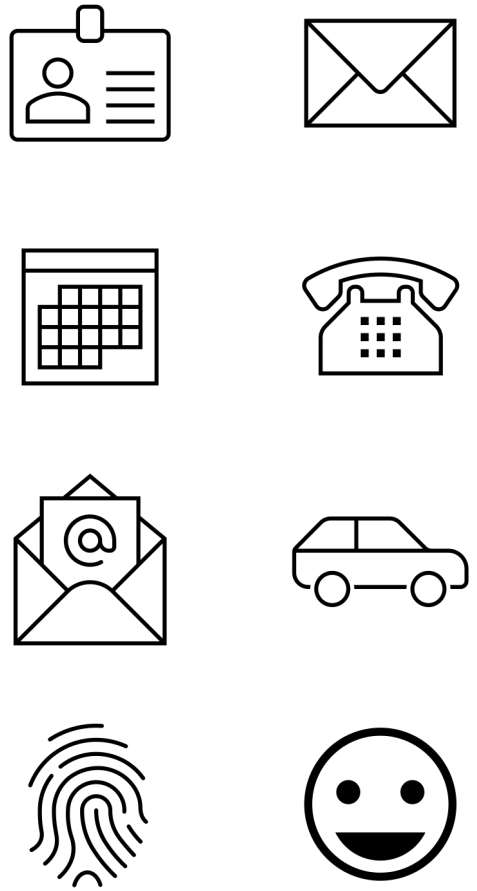
Deidentification of Private Health Information (PHI)



From last week: the HIPAA Safe Harbor method

1. Names
2. Addresses (including parts of some ZIP codes)
3. Dates (including some years and ages)
4. Telephone numbers
5. Fax numbers
6. Email addresses
7. Social security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers
13. Device identifiers and serial numbers
14. Web URLs
15. IP addresses
16. Biometric identifiers (fingerprint, voice, etc.)
17. Full-face photographs
18. Other unique identifiers, except for allowed identifiers

Deidentification of Private Health Information (PHI)



Data collection

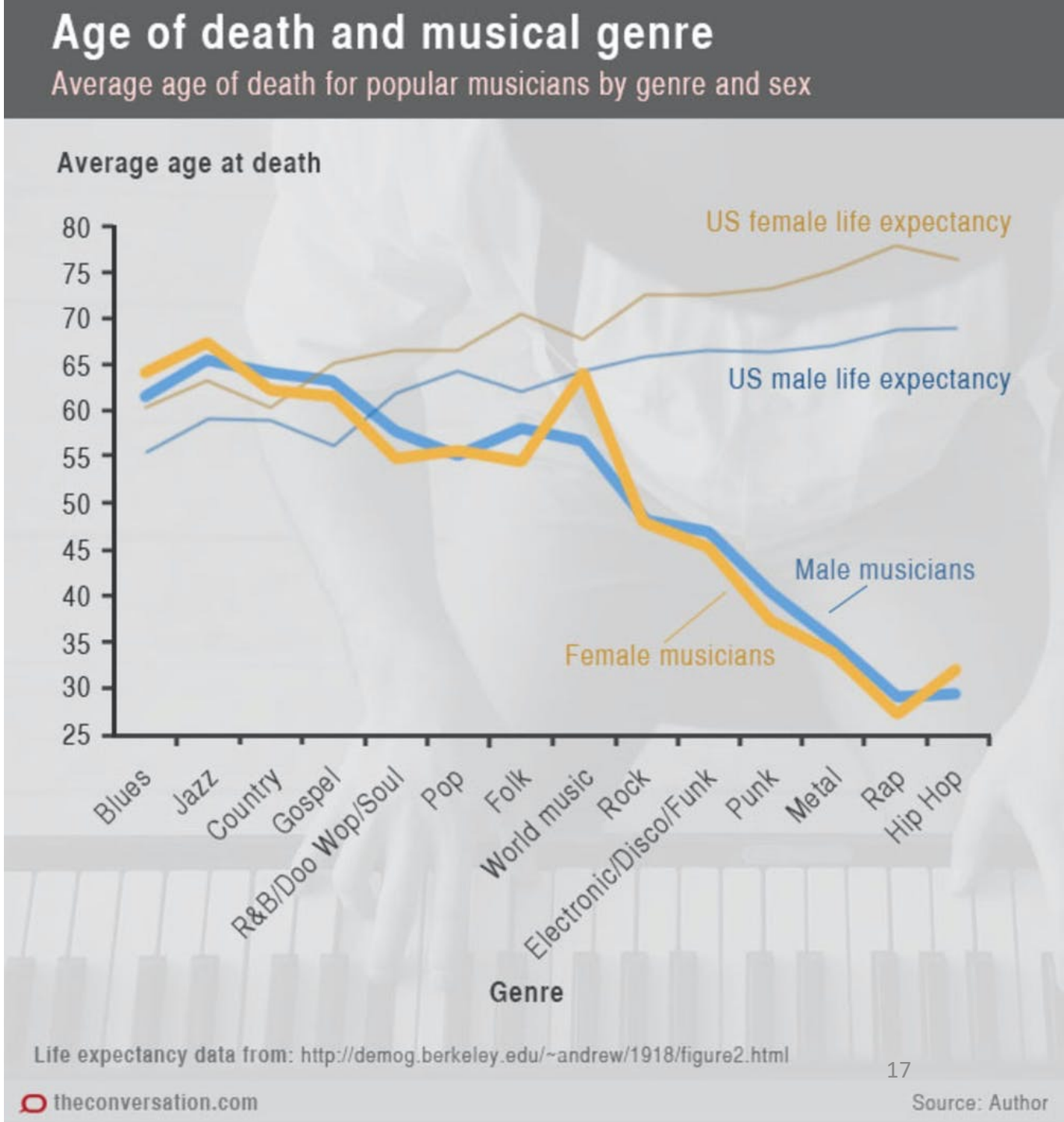
This plot is from the following article:

> Kenny, D.T. "Music to die for: How genre affects popular musicians' life expectancy." *The Conversation* (2015).

The plot suggests that musicians who play different types of music have very different life expectancies.

This plot has **many** problems, including at least one problem related to data collection. What are they?

https://theconversation.com/music-to-die-for-how-genre-affects-popular-musicians-life-expectancy-36660

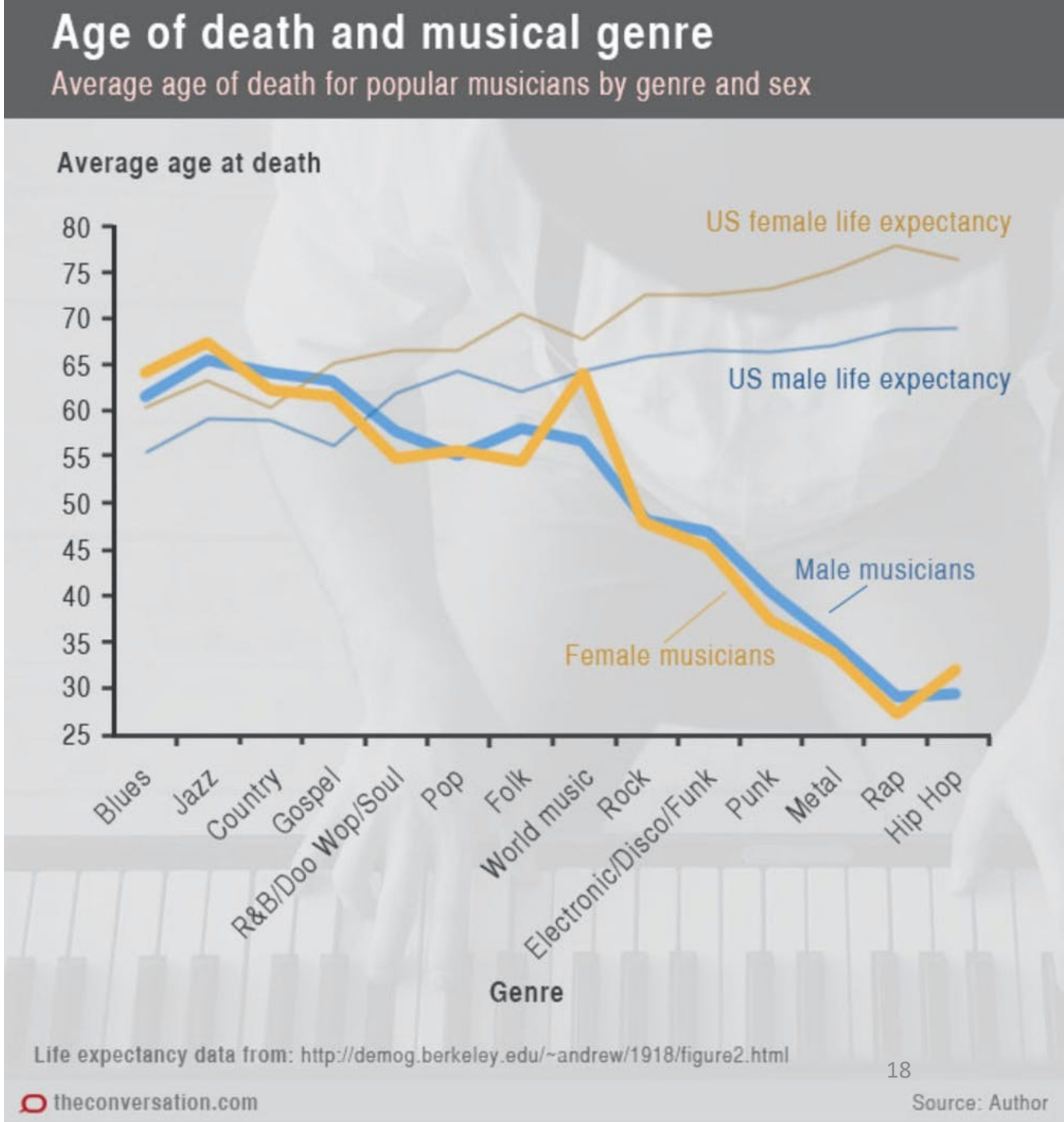


Data collection

Several issues:

1. **What happens with musicians who are still alive? How are censored data handled?**
2. How are the genres defined? How are "popular musicians" defined? Are the musicians from the U.S.?
3. Is average age a meaningful statistic?
4. How are musician ages and life expectancy data matched?
5. Why are lines used to connect data points between different genres? The genres are categorical variables. What does it mean to interpolate between categorical variables?

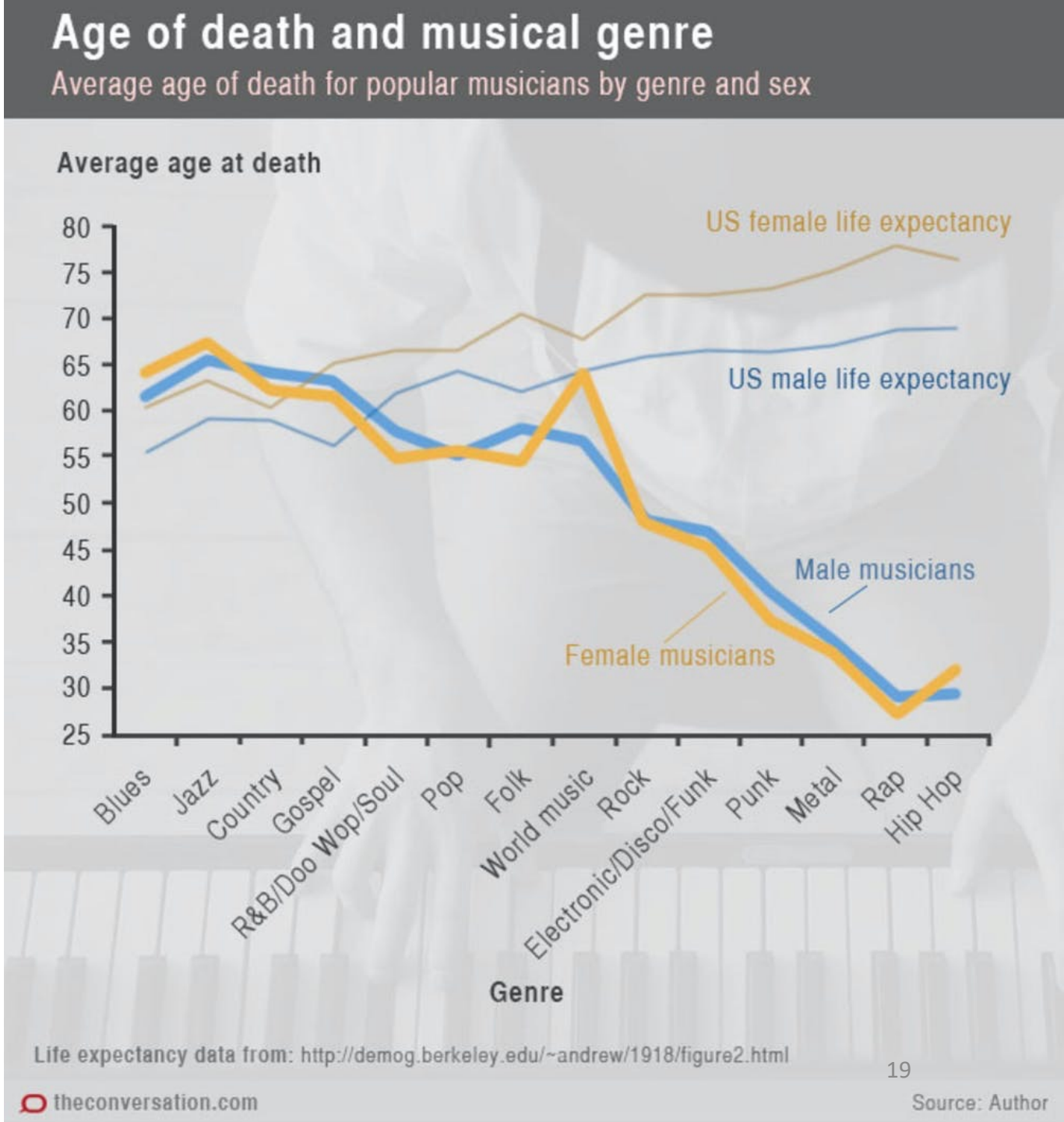
https://theconversation.com/music-to-die-for-how-genre-affects-popular-musicians-life-expectancy-36660



Data collection

Several more issues:

6. Why are the labels for the x axis separated by tick marks and rotated 45°?
7. Should the range of the y axis be determined by the smallest and largest values in the plot?
8. Where is the figure caption? Why is there information in the figure that should be in the figure caption?
9. How are the colors chosen?
10. Why is there picture of a piano player in the background of the figure?
11. Why is a lossy JPEG image format used for the figure?



https://theconversation.com/music-to-die-for-how-genre-affects-popular-musicians-life-expectancy-36660

Data organization

Like data collection, data organization can be domain specific.

Basic principles:

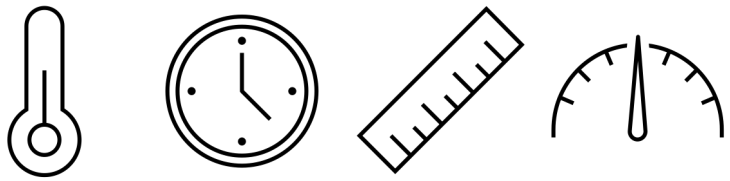
1. We do **not** want to change the data (yet).
2. We do **not** want to answer any research questions (yet).
3. We can save data in text and binary formats:
   a. Text formats: less efficient, but more readable.
      CSV/TSV/DSV and JSON are text common formats.
   b. Binary formats: more efficient, but less readable.
      HDF5 is a common binary format, and there are many specialized binary formats.
      Don't use Pickle.

Data cleaning

Like data collection and data organization, data cleaning can be domain specific.

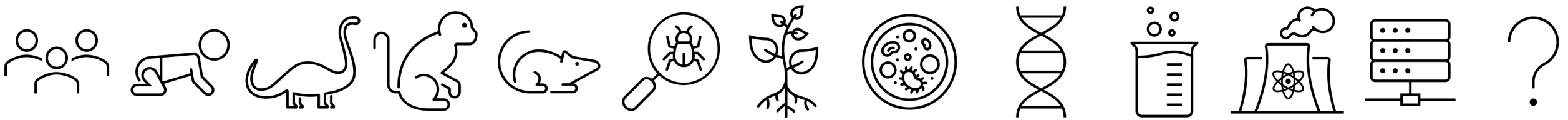
Basic principles:

1. We **do** want to change the data, but only in defensible, deliberate, and documented ways.
2. We do **not** want to answer any research questions (yet).
3. What do we do about missing data? Can/should we impute it?
4. What do we do about duplicate data? Can/should we remove it?
5. What do we do about errors? Can/should we use what we know about the problem domain to impute, remove, or correct it?



Exploratory data analysis: statistics, tables, and figures

Different fields can have **very** different standards for statistical analysis. What does your advisor do? What do papers in your field do? What do **good** papers in your field do?



Exploratory data analysis: **statistics**, tables, and figures

Different fields can have **very** different standards for statistical analysis. What does your advisor do? What do papers in your field do? What do **good** papers in your field do?

You statistics course(s) will give you tools to perform appropriate statistical analyses and to identify inappropriate statistical analyses, but still consider the standard of your field.

Common statistics issues:

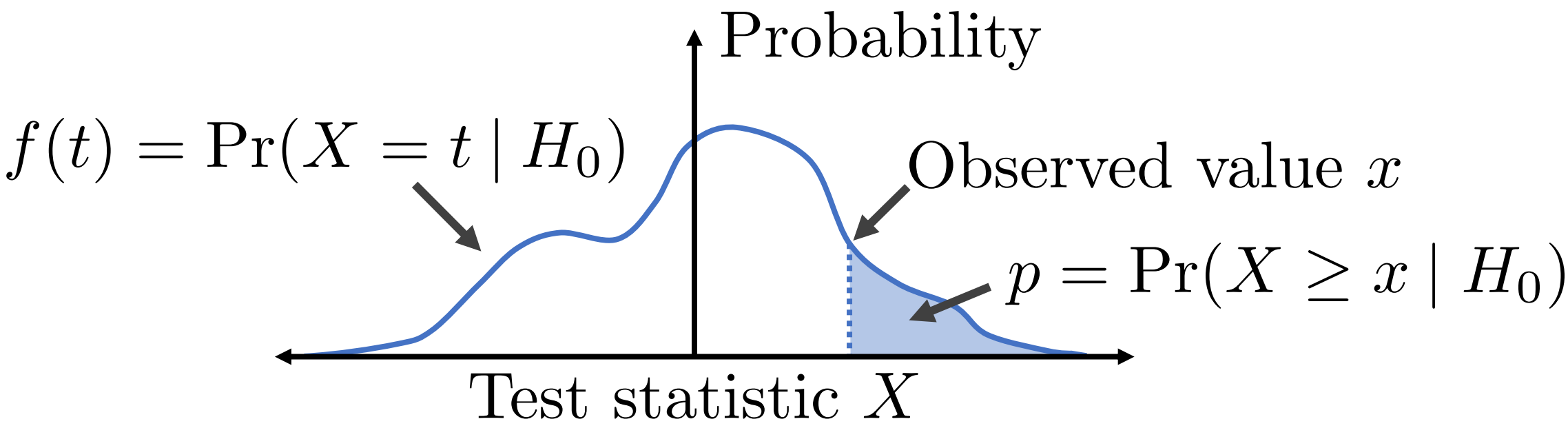
1. Cherry picking: suppressing data that does not support your hypothesis
2. Fudging the data: changing data that does not support your hypothesis
3. Data dredging or p-hacking: using data to find hypotheses and not testing them on different data or using multiple testing corrections (FWER, FDR, etc.)
4. "Significance": statistical significance does not mean that results are interesting or meaningful, tests can be overpowered, the null hypothesis cannot be confirmed, etc.

Fundamental issue: what is a p-value?

A p-value p is...

Fundamental issue: what is a *p*-value?

A *p*-value p is the probability of observing a result X that is at least as extreme as the observed result x under the assumption that the null hypothesis H_0 is true.

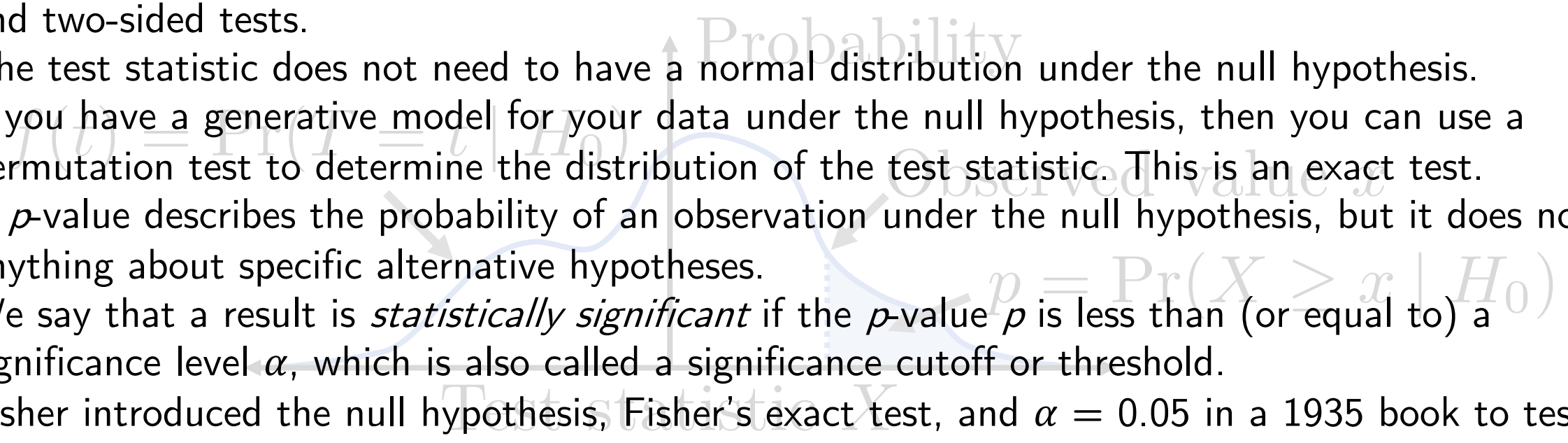


Fundamental issue: what is a *p*-value?

A *p*-value p is the probability of observing a result X that is at least as extreme as the observed result x under the assumption that the null hypothesis H_0 is true.

Some notes:

1. This diagram illustrates a one-sided right-tailed test, but there are also one-sided left-tailed tests and two-sided tests.
2. The test statistic does not need to have a normal distribution under the null hypothesis.
3. If you have a generative model for your data under the null hypothesis, then you can use a permutation test to determine the distribution of the test statistic. This is an exact test.
4. A *p*-value describes the probability of an observation under the null hypothesis, but it does not say anything about specific alternative hypotheses.
5. We say that a result is *statistically significant* if the *p*-value p is less than (or equal to) a significance level α, which is also called a significance cutoff or threshold.
6. Fisher introduced the null hypothesis, Fisher's exact test, and $\alpha = 0.05$ in a 1935 book to test whether someone could tell whether tea or milk was first added to a cup.
7. What happens when you have many or too many *p*-values? Any other questions?



Exploratory data analysis: statistics, **tables**, and figures

Tables can include raw data, statistics, and other descriptions of numerical or categorical data and/or results.

A TABLE of the *Apertures* of *Object-Glasses*.

The Points put to some of these Numbers denote Fractions.

Lengths of Glasses. Feet, Inches.	For excellent ones. Inch. Lines.	For good ones. Inch. Lines.	For ordinary ones. Inch. Lines.	Lengths of Glasses. Feet, Inches.	For excellent ones. Inch. Lines.	For good ones. Inch. Lines.	For ordinary ones. Inch. Lines.
4	4·	4	3	25	3 4	2 10	2 4·
6	5·	5	4	30	3 8	3 2	2 7
9	7	6	5	35	4 0	3 4·	2 10
1 0	8.	7	6	40	4 3	3 7	3 ·
1 6	9	8.	7	45	4 6	3 10	3 2.
2 0	11	10	8	50	4 9	4 0	3 4.
2 6	1 0	11	9	55	5 0	4 3	3 6.
3 0	1 1	0	10	60	5 2	4 6	3 8.
3 6	1 2	1	11	65	5 4	4 8	3 10
4 0	1 4	1 2	1 0	70	5 7	4 10	4 ·
4 6	1 5	1 3	1 ·	75	5 9	5 0	4 2.
5 0	1 6	1 4	1.	80	5 11	5 2	4 5
6	1 7.	1 5	2	90	6 4	5 6	4 7.
7	1 9	1 6	3	100	6 8	5 9	4 10
8	1 10	1 8	4	120	7 5	6 5	5 3
9	1 11	1 9	5	150	8 0	6 7	5 11
10	2 1	1 10	6	200	9 6	8 6	6 9
12	2 4	2 0	8	250	10 6	9 2	7 8.
14	2 6	2 2	9.	300	11 6	10 0	8 5
16	2 8	2 4	11.	350	12 6.	10 9	9 0
18	2 10	2 6	1	400	13 4	11 6	9 8
20	3 0	2 7	2.				

Auzout, "A TABLE of the Apertures of Object-Glasses", *Philosophical Transactions* (1665).

Summarizing data: too much data...

x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

(a) I

x	y
10.0	9.14
8.0	8.14
13.0	8.74
9.0	8.77
11.0	9.26
14.0	8.1
6.0	6.13
4.0	3.1
12.0	9.13
7.0	7.26
5.0	4.74

(b) II

x	y
10.0	7.46
8.0	6.77
13.0	12.74
9.0	7.11
11.0	7.81
14.0	8.84
6.0	6.08
4.0	5.39
12.0	8.15
7.0	6.42
5.0	5.73

(c) III

x	y
8.0	6.58
8.0	5.76
8.0	7.71
8.0	8.84
8.0	8.47
8.0	7.04
8.0	5.25
19.0	12.5
8.0	5.56
8.0	7.91
8.0	6.89

(d) IV

Table X: Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Summarizing data: ... which has been summarized...

7.0	4.82
5.0	5.68

(a) I

7.0	7.26
5.0	4.74

(b) II

7.0	6.42
5.0	5.73

(c) III

8.0	7.91
8.0	6.89

(d) IV

Table X: Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

	I	II	III	IV
Mean of x	9.00	9.00	9.00	9.00
Sample variance of x	11.00	11.00	11.00	11.00
Mean of y	7.50	7.50	7.50	7.50
Sample variance of y	4.13	4.13	4.13	4.13
Linear regression line	$y = 0.5x + 3.0$	$y = 0.5x + 3.0$	$y = 0.5x + 3.0$	$y = 0.5x + 3.0$
Coefficient of determination R^2	0.67	0.67	0.67	0.67
Correlation ρ	0.816	0.816	0.816	0.816

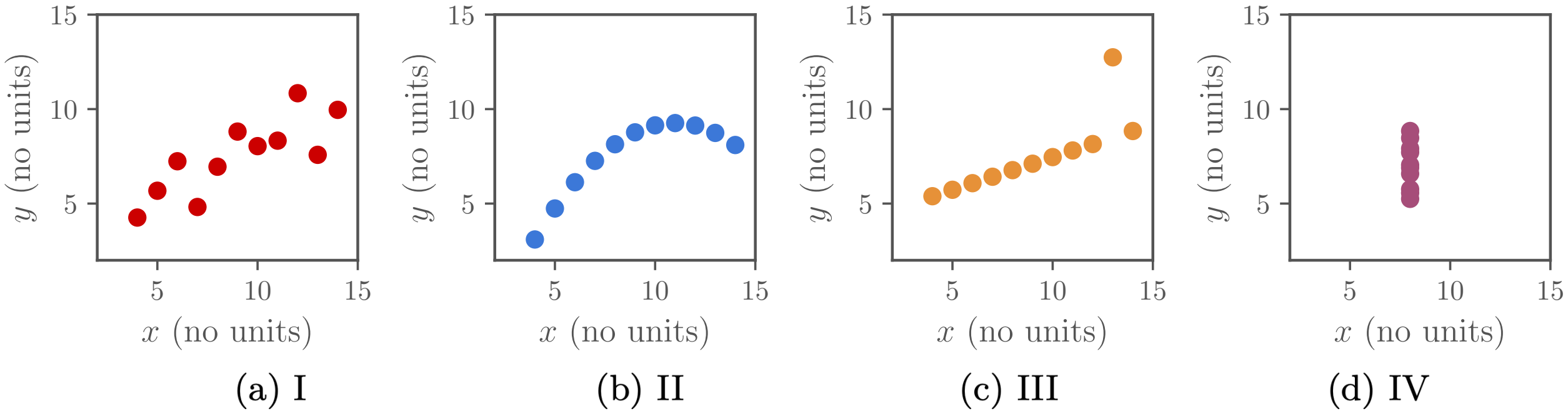
Table Y: Relevant summary statistics for Table X.

Since datasets I, II, III, and IV are sampled from the same distribution, we conclude that ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat; see Supplemental Figure Z for plots of datasets I, II, III, and IV.

Summarizing data: ... but not appropriately.

Since datasets I, II, III, and IV are sampled from the same distribution, we conclude that ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat; see Supplemental Figure Z for plots of datasets I, II, III, and IV.

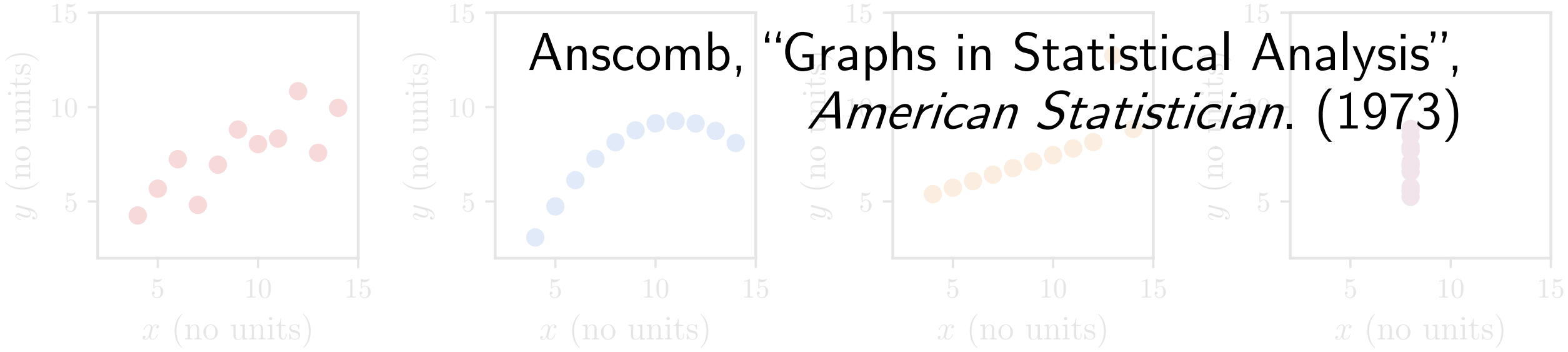
[... many pages omitted ...]



Supplemental Figure Z: Plots of datasets I, II, III, and IV in Table X.

Summarizing data: a picture is worth a thousand words

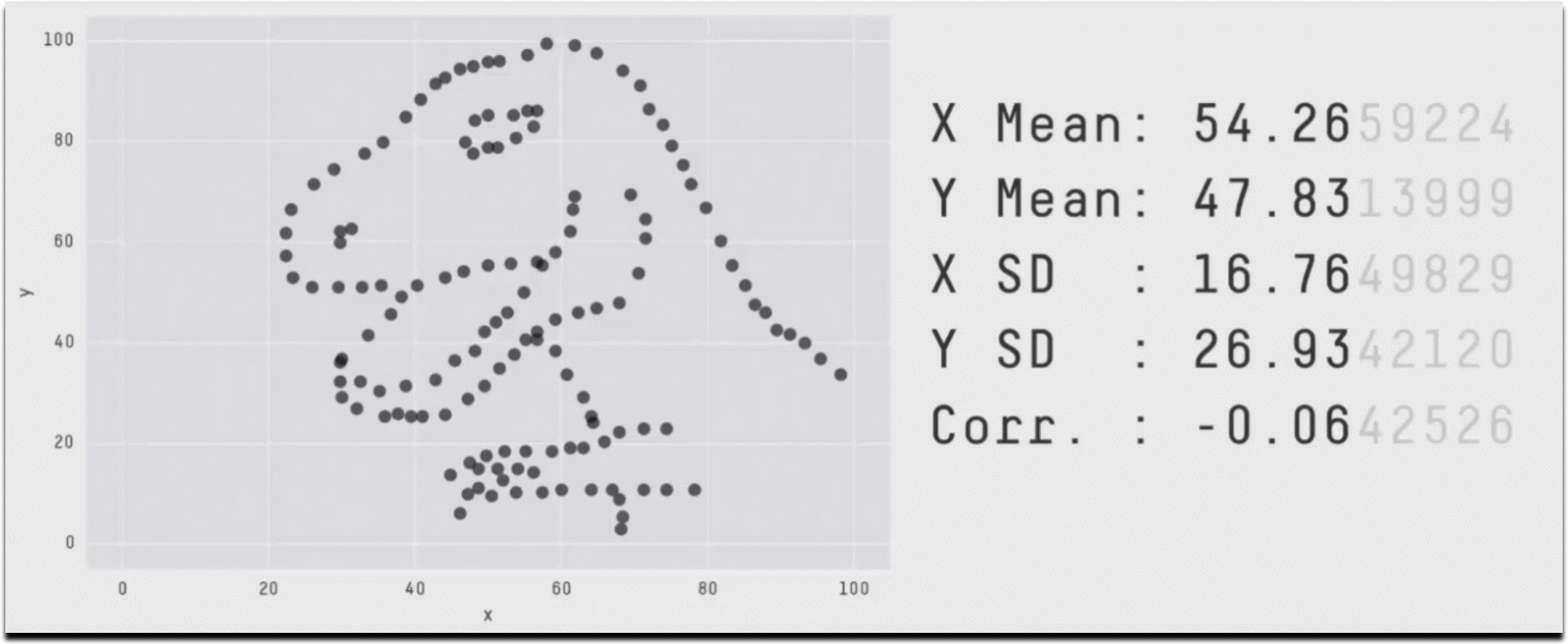
This is **Anscomb's quartet**, which is a collection of four data sets with very similar descriptive statistics but very different distributions. This example counters the idea that "numerical calculations are exact, but graphs are rough."



Anscomb, "Graphs in Statistical Analysis", *American Statistician*. (1973)

Summarizing data: the Datasaurus dozen

These twelve datasets have the same means, standard deviations, and correlation coefficients but very different distributions. What are better ways to summarize these data?



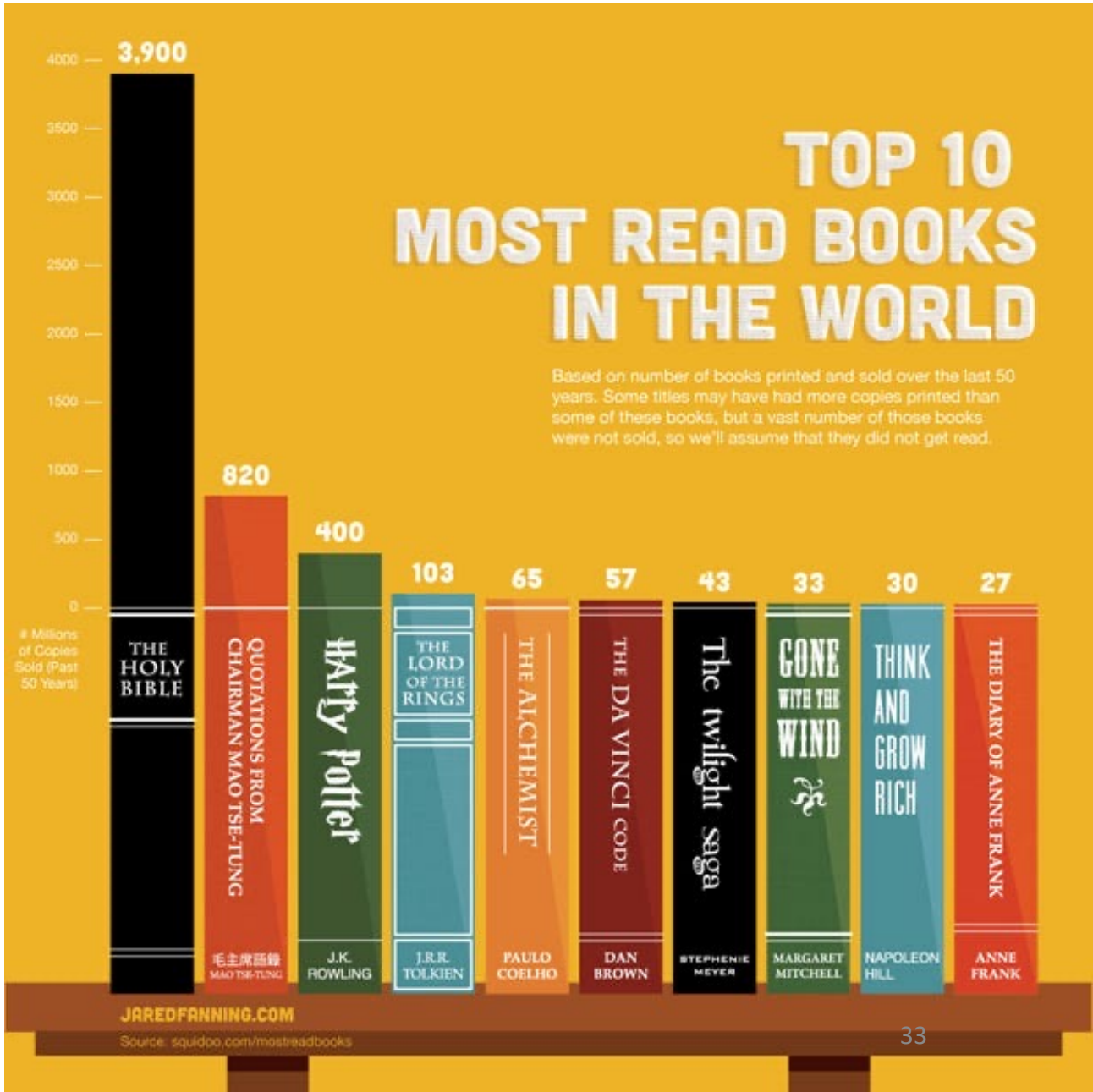
Datasaurus Dozen: https://www.autodesk.com/research/publications/same-stats-different-graphs

Exploratory data analysis: statistics, tables, and **figures**

"[A misleading figure] is vastly more effective, however, because it contains no adjectives or adverbs to spoil the illusion of objectivity, there's nothing anyone can pin on you.
– D. Huff, *How to Lie with Statistics* (1954).

What are common ways that figures mislead, and how can we avoid misleading figures?

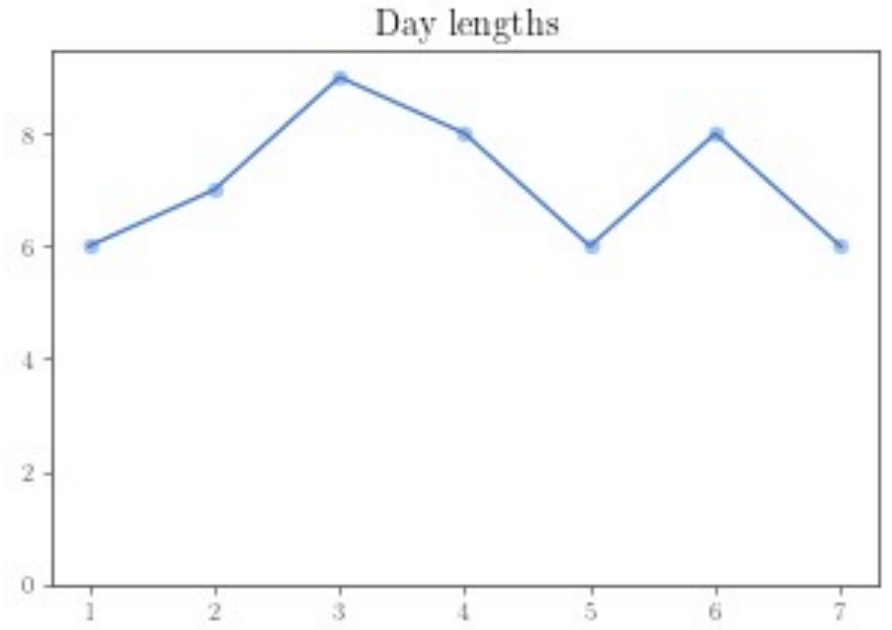
More generally, what are common issues with figures, and how can we make better figures?

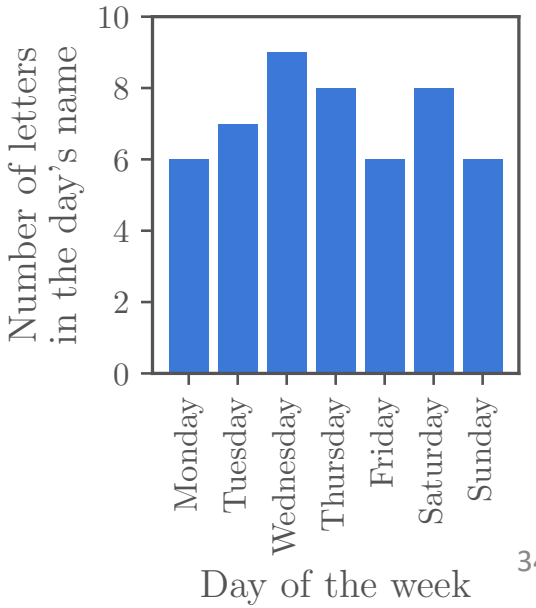


https://www.businessinsider.com/the-top-10-most-read-books-in-the-world-infographic-2012-12

Advice for making better figures

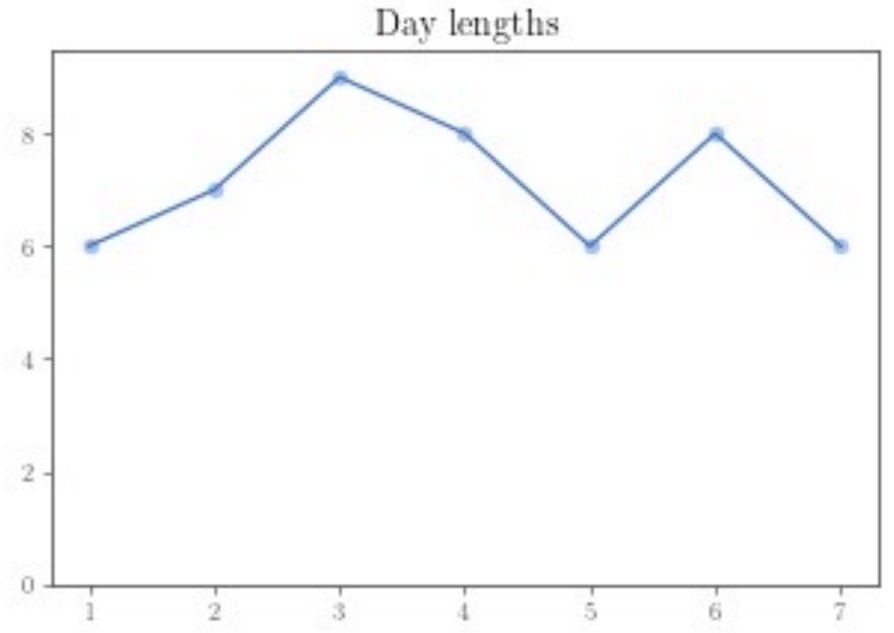
1. Label your axes and other plot elements.
2. Add units, even if there are no units (say "no units").
3. Spell words correctly.
4. Use proper abbreviations.
5. Use proper mathematical notation. Use LaTeX!
6. Make all text, lines, points, etc. legible. You can reduce the plot size before you make the plot to make the plot elements larger.
7. Choose an appropriate plot type.
8. Use colors, markers, lines appropriately. Be sensitive to colorblindness.
9. Use appropriate axis limits with consistent axis limits for subplots.
10. Follow the principle of proportional ink. When in doubt, use relative axis limits for line plots and absolute axis limits for histograms, but this is not a hard-and-fast rule.
11. Use logarithmic axes when appropriate.
12. Don't include a title in the figure image (unless you need it).
13. Use clear and descriptive figure captions.
14. Use the figure caption to explain the figure but not to interpret the figure.

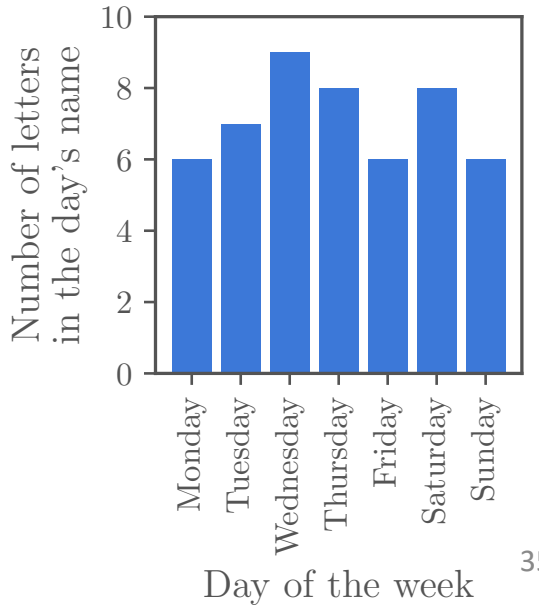




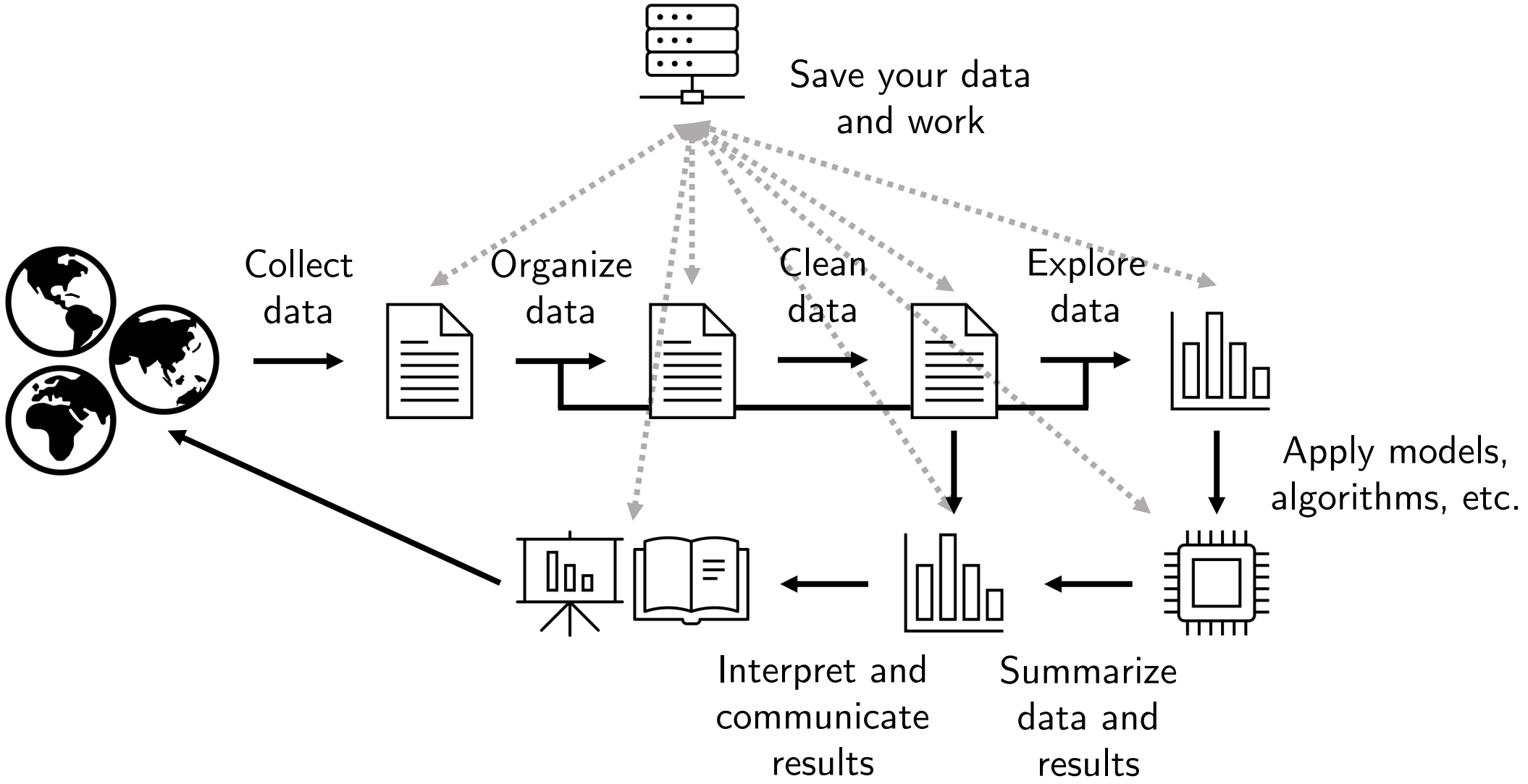
Advice for making better figures

15. Cite figures and tables that you didn't make – and even figures and tables that you did make.
16. Don't use lossy JPEGs for figures. When in doubt, use PDFs for vector-based graphics and PNGs with at least 300 DPI for raster-based graphics.
17. Save the data for your plots so that you can quickly redraw them as needed. Many journals are starting to require authors to supply the data shown in their plots, so you may need to do this anyway.
18. Save the code for making your plots so that you can reuse it. Use version control.
19. Make meaningful figures. A good figure can effectively replace a long paragraph or a large table, but a bad figure can have very little information content.
20. Beware of the law of diminishing returns. Don't spend too much time making figures.
21. Make sure that your advisor can use your figures.
22. Follow your advisor's advice about figures.





A generic data treatment pipeline



You should build an ***automated*** data treatment pipeline that checks and transforms the raw results into the statistics, tables, and figures. Shell scripts and notebooks? Great. Manual steps? Not reproducible.

Lab: find problematic plots and remake one of them

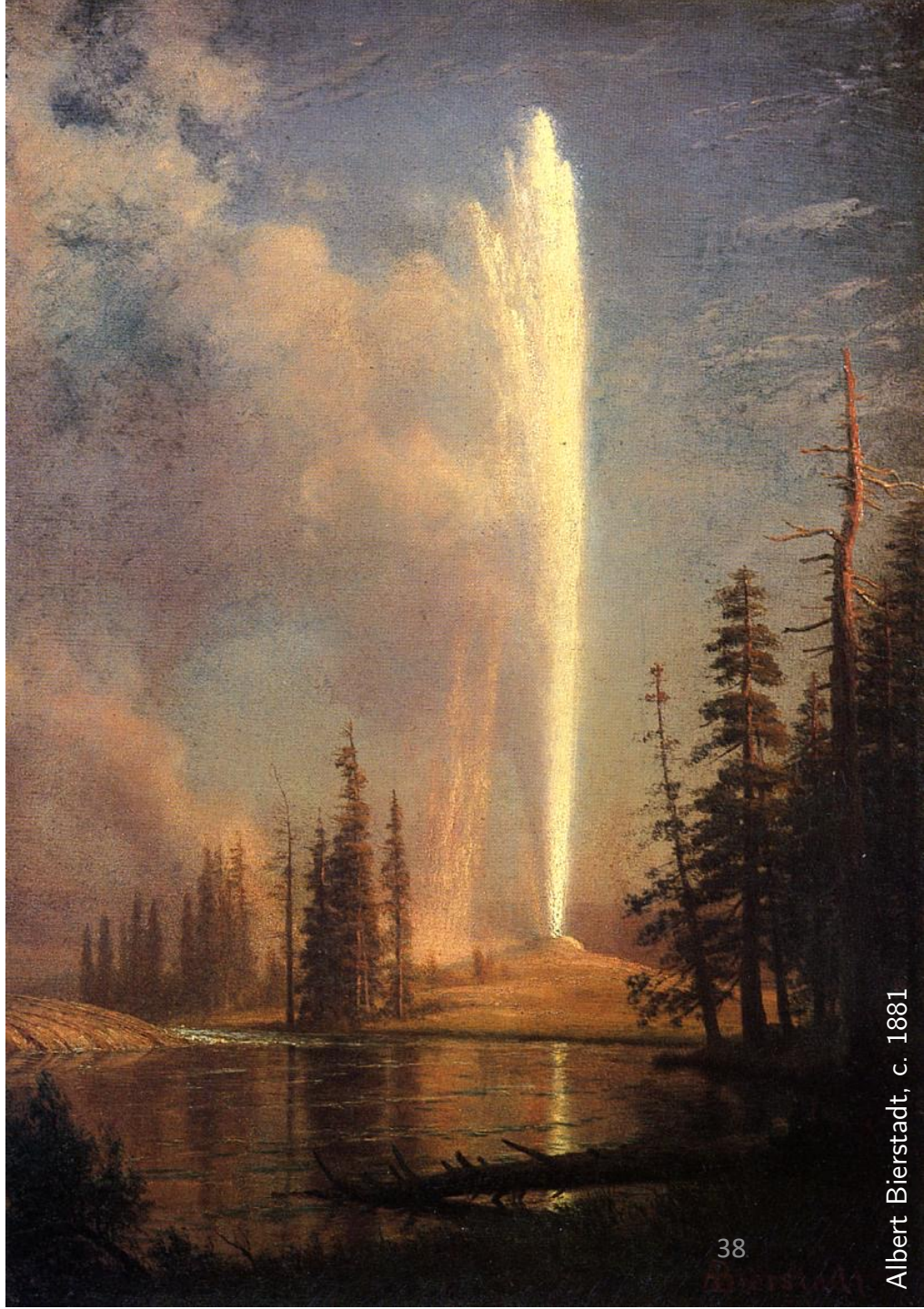
1. Go to https://www.cinc.org/cinc-papers-on-line/ for the conference proceedings for Computing in Cardiology* (or another conference of your choosing).
2. Find one plot in the conference proceedings articles that have one or more of the issues that we discussed in this class (or other issues that we have not discussed). Which plots did you choose? What are the issues?
3. Extract, reproduce, or guess the data from one of these plots and remake the plot to address the issues that you found.
4. Don't have a favorite plotting program? Try Excel, ggplot2, gnuplot, Google Sheets, MATLAB, Matplotlib, pgfplots, seaborn, or something else.
5. Can you add summary information to your plot to improve it? Try a few things.
6. Make a slide, a document, an image, or something else to show us at the end of the lab.

*This is a conference that Gari and I help to organize. Some of the contributors are still learning the very topics that we are discussing in this lecture. We are not making fun of them.

Homework: data treatment project

Old Faithful is a geyser in Yellowstone National Park, Wyoming, United States. It erupts at predictable intervals. Over one million eruptions have been recorded.

1. Download an "Old Faithful" dataset here: https://reynalab.org/teaching/bmi500_fall2022/geyser.csv
2. Collect, organize, and clean this dataset.
3. Make a "bad" plot of the data and a "good" plot of the data.
4. Interpret the results.
5. Write a short 2-page report. Make sure to
   a. clearly and concretely describe and defend your data treatment;
   b. include your plots with clear captions;
   c. interpret the data and results;
   d. include your code in the supplement (not part of the page limit); and
   e. add anything else that a good report needs.
6. Use this rubric to improve your report: https://reynalab.org/teaching/bmi500_fall2022/rubric.pdf
7. Send your full report in PDF format to your TA by 5:00pm ET on Monday, 19 September 2022.



Albert Bierstadt, c. 1881