Data Science Practice

STATS 369 Coursebook: Week 2.2 & 2.3

Plan

[L05 & L06] Ethics (part 1)

- Getting data
- Analyzing data
- Making decisions with data

This topic introduces or recaps a range of ideas about ethical professional practice for data scientists. Think of this week as just a very high-level survey of these ideas; we aren't going very deep on any of them, but some will set up conversations/tasks later in the course.

- We'll revisit some of these concepts in later classes.
- We'll touch on **web scraping** and **APIs** in the lab this week.

Ethical professional practice for statisticians

The following sections consider the *whys* and *whats* of ethical practice organized under the categories of:

- getting data: confounding and study designs, human research ethics, web scraping and APIs, Indigenous data sovereignty);
- analyzing data: HARKing & multiple testing;
- making decisions with data: algorithmic bias and transparency.

There is a lot of ground covered, though lightly, and it may feel intimidating or overwhelming—there is so much to consider, so much we could get wrong! So, here is some advice from a teaching assistant who worked with me on developing this content:

> *developing the judgment and skills related to these topics takes time and no analysis is ever perfect—we rely on collaboration with other scientists and statisticians who all try to think critically to avoid harm and to improve our collective knowledge and understanding.* -- Sonia Markes

Why should statisticians be ethical?

I hope this questions seems a bit silly to most of you, shouldn't we all strive to be ethical in our professional and personal lives? Well regardless of your personal philosophy, there are additional considerations with respect to our professional ethics.

- **Professional societies:** You may have heard of the Hippocratic oath ("first do no harm") in Medicine and there are similar codes of conduct for Statisticians (see the links to the Canadian and US version below) through our own professional societies.
- **Research ethics boards:** If you are conducting research within universities, it will often have to be approved by a research ethics board. You need to have the skills and knowledge to help appropriately design studies that can be approved under these rules.
- **Legal and business considerations:** Unethical behaviour can also get you are your future employers into trouble. There are of course reputational risks, but there can also be legal risks for creating algorithms that discriminate against protected attributes like gender or ethnicity.

Ethical practice

Statistical tools are used to create knowledge. Therefore, anyone using statistical tools is responsible for the knowledge they create through their data collection and analyses.

As ethical data scientists it is important to:

- be accurate in our analyses and conclusions
- be alert to possible consequences of our results/recommendations on others
- be honest in reporting results, even when we don't get the results we hoped for
- be respectful of other reasonable results (based on well-conducted research) even if they differ from our own
- be mindful of what your data represents, especially if it represents people or their behaviour (a study on people has *subjects*, not objects)
- share credit when our work is based on the ideas of others
- and more...

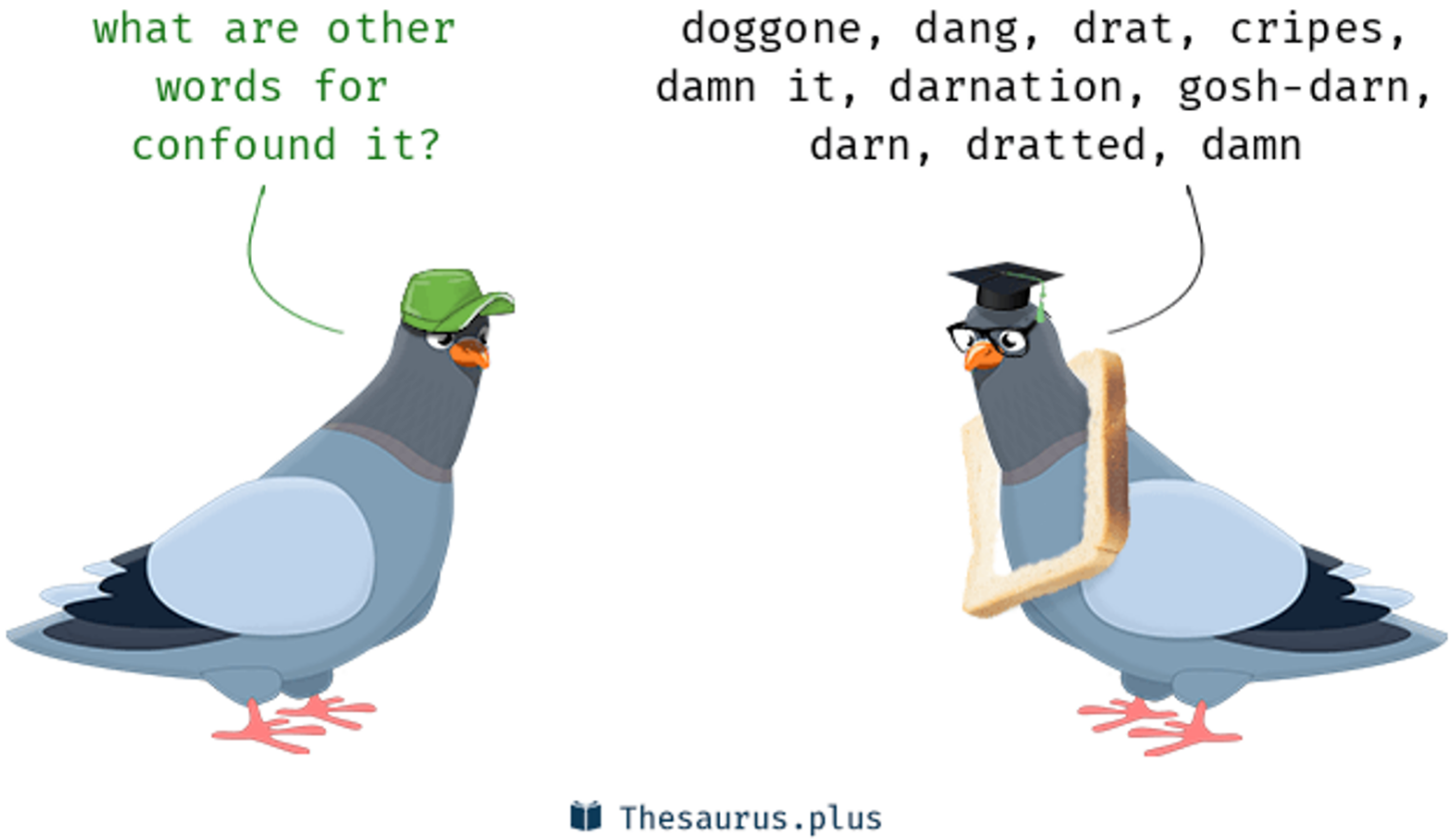
📚 **[Optional] Code of conduct for the American Statistical Association {#ethicscodes}**

American Statistical Association: https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx

Getting data

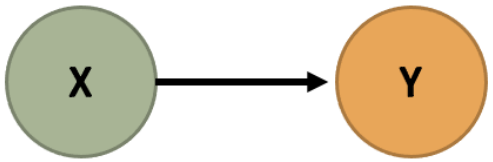
Confounding and study design

Confounding and study design

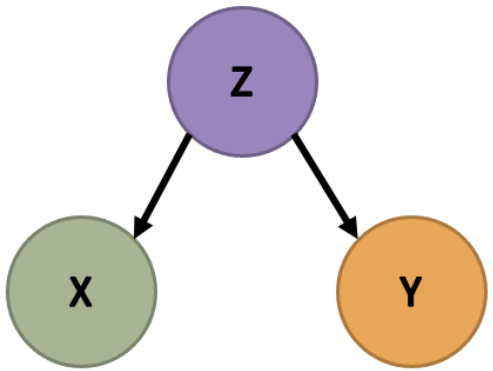


Confounders (i.e. confounding factors or confounding variables)

Suppose you are interested in the association between a explanatory variable, **X**, and a response variable **Y**.



A **confounding variable** (or just 'confounder'), Z, is a variable that influences BOTH the explanatory variable and the response variable.



If we fail to account for our confounding variable, either by not measuring it or not including it, we can come to incorrect conclusions.

Study types

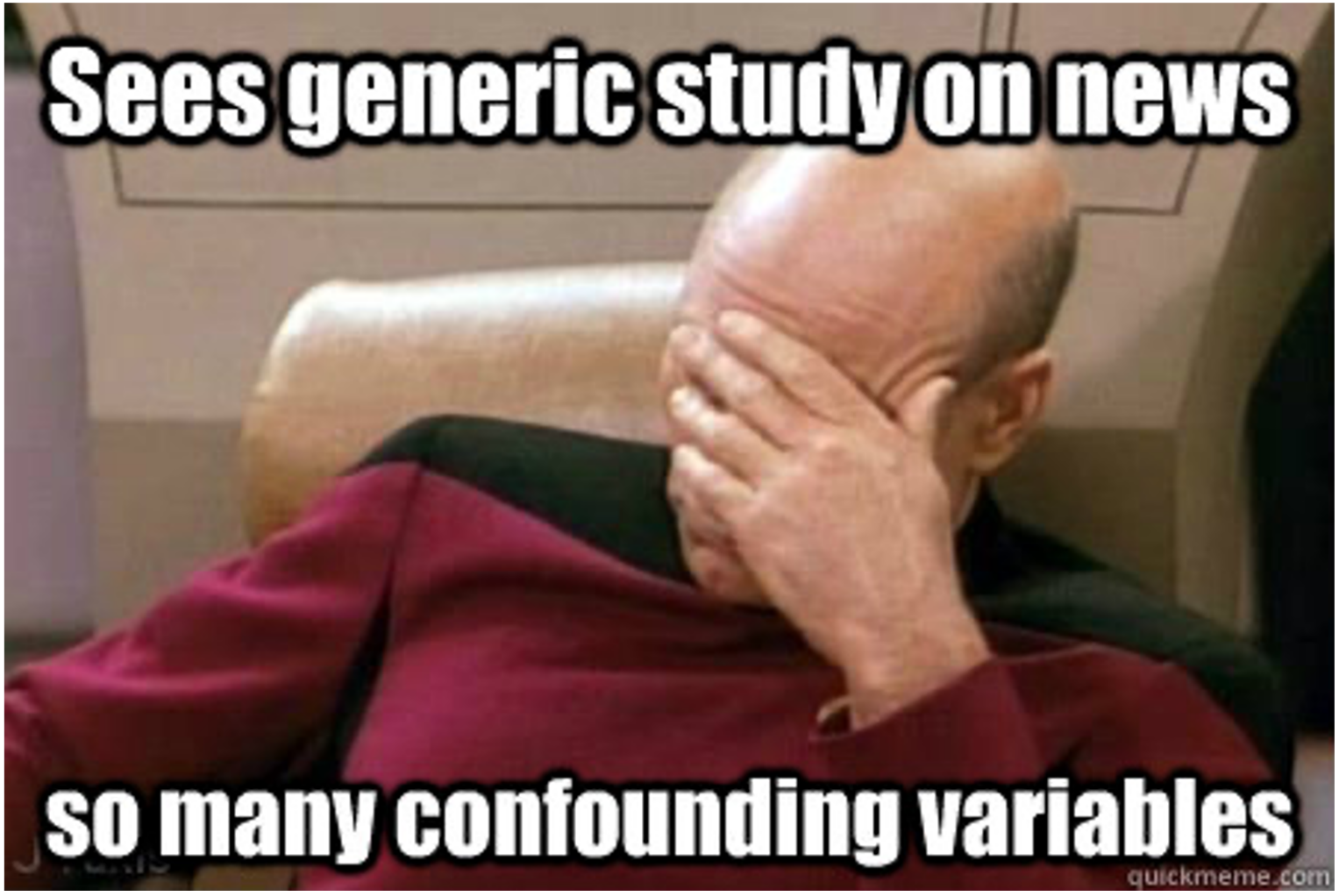
In an **observational study**, variables are "observed" (measured and recorded) without manipulation of variables or conditions by the researcher.

Two variables are **confounded** if their effects on the response variable are mixed together and there is no way to separate them out. If this is the case, we have no way of determining which variable is causing changes to the response.

Example: As ice creams sales rise, so do drownings. But are people drowning because of ice cream? What would be a plausible confounder for this relationship? (Hint: ☀️🌡️)

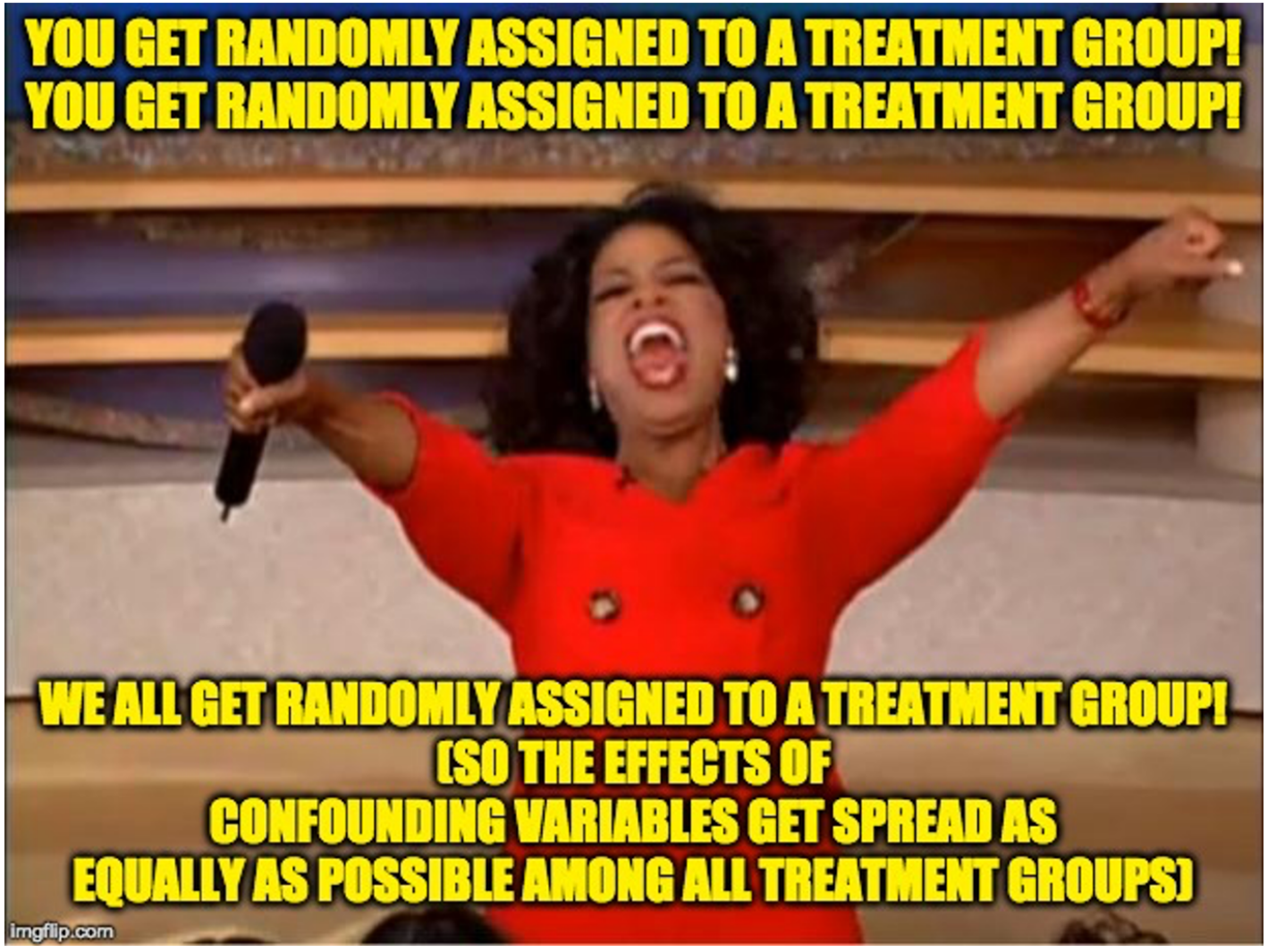
There is no 'test' for confounding. But this is a great thing if you don't want the robots to take your job anytime soon. We need smart people who can think well about confounding.

When we have data from an observational study, we can often only conclude association between variables, not causation. It is worth noting that methods of causal inference are significant area of research for statisticians and others.


Sees generic study on news
so many confounding variables
quickmeme.com

So, do we give up on claiming **causation**?

No!


YOU GET RANDOMLY ASSIGNED TO A TREATMENT GROUP!
YOU GET RANDOMLY ASSIGNED TO A TREATMENT GROUP!
WE ALL GET RANDOMLY ASSIGNED TO A TREATMENT GROUP!
(SO THE EFFECTS OF
CONFOUNDING VARIABLES GET SPREAD AS
EQUALLY AS POSSIBLE AMONG ALL TREATMENT GROUPS)
imgflip.com

Designing studies to avoid confounding

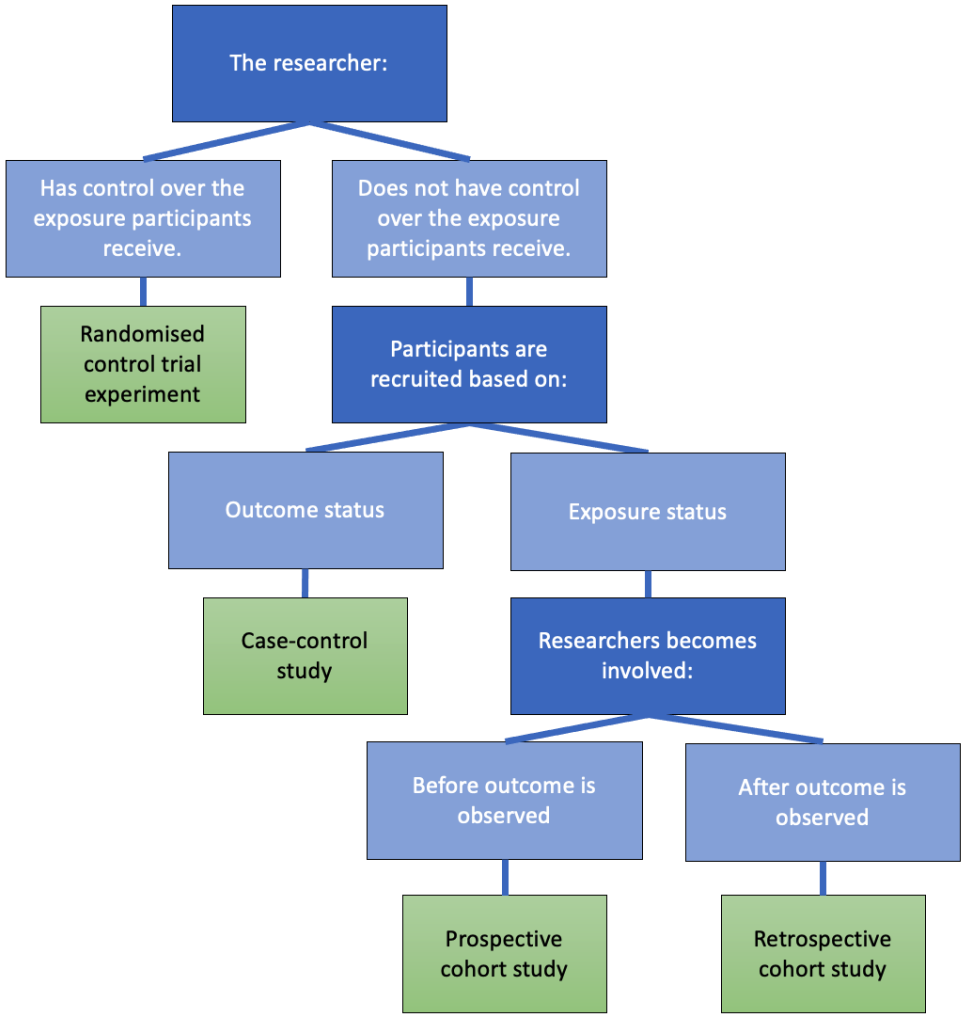
In an **experiment** (or **randomized trial** or **randomized control trial**) variables and/or conditions are manipulated by the researcher and the impact on other variable(s) is measured and recorded.

The key is to randomly assign some individuals to one treatment (or condition) and randomly assign others to another treatment (sometimes this other treatment is a **control**) *Note: you can have more than two treatments groups too—what is important is that individuals are randomly assigned to them!*

The groups (before treatments are applied) should be very similar to each other with respect to the other variables. Any differences between individuals in the treatment and control groups would just be due to random chance!

If there is a **significant difference** in the **outcome** between the two groups, we may have evidence that there is a **causal relationship** between the treatment and the outcome.

Common study designs and how to recognize them



Causation from observational studies?

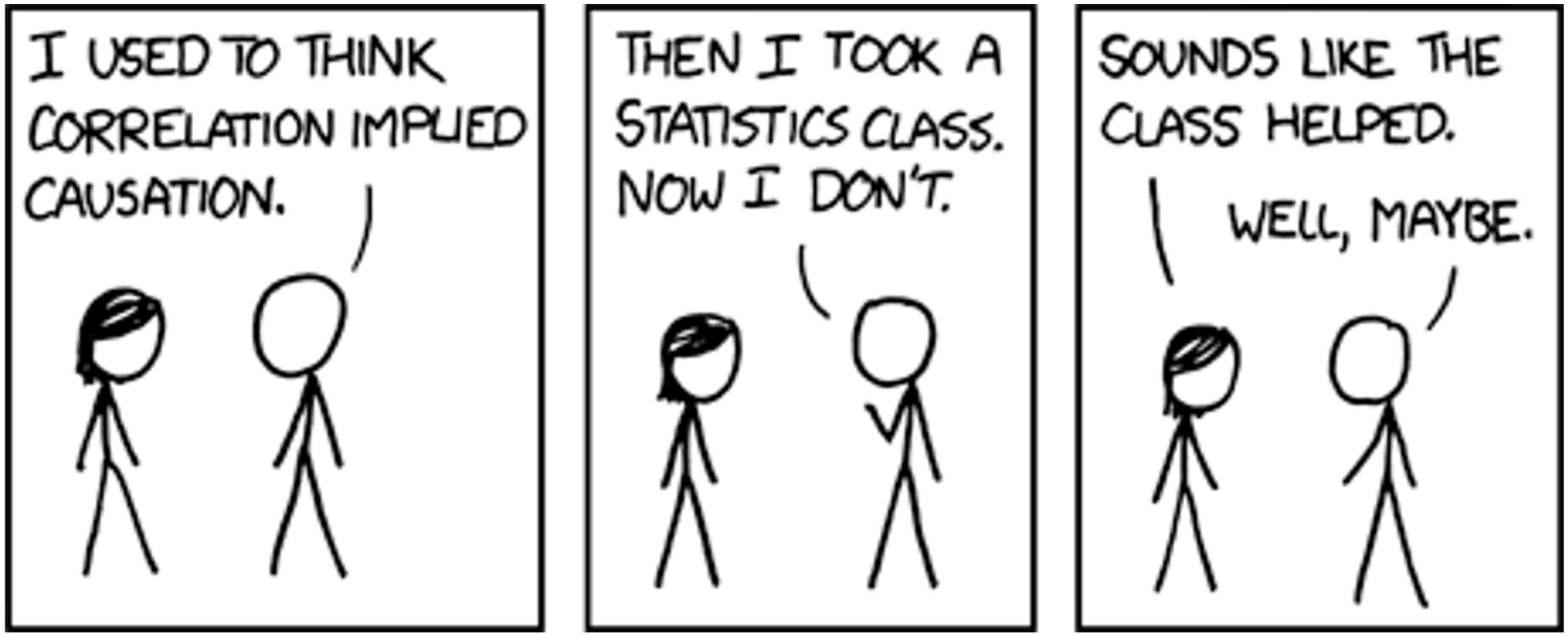
Although well-designed randomized trials are the gold standard way to establish a causal relationship, observational studies can also help build **evidence** for causation.

Bradford Hill criteria (not assessed)

- Strength of association
- Consistency
- Specificity
- Temporality
- Biological gradient
- Plausibility
- Coherence
- Experiment
- Analogy

> Science is not a magic wand that turns everything it touches to truth. Instead, “science operates as a procedure of uncertainty reduction,” said Nosek, of the Center for Open Science. “The goal is to get less wrong over time.” This concept is fundamental — whatever we know now is only our best approximation of the truth. We can never presume to have everything right.

From the *Science isn't broken* (Aschwanden, 2015).



Source: https://xkcd.com/552/

Human research ethics

Ethical codes often emerge out of crisis events.

The Nuremberg code was formulated in August 1947 in Nuremberg, Germany, by American judges* sitting in judgment of Nazi doctors accused of conducting murderous and torturous human experiments in concentration camps during the war.

The Nuremberg code codified many of our standard principles of ethical research, including:

- research must appropriately balance risk and potential benefits
- researchers must be well-versed in their discipline and ground human experiments in animal trials.

- Did this judgment mean Americans were always getting research ethics right? Definitely not. (Optional) YouTube video about the Tuskegee Study

Principles of free and informed consent

Information

The research procedure, risks and anticipated benefits, alternative procedures (where therapy is involved), and a statement offering the participant the opportunity to ask questions and to withdraw at any time from the research.

Comprehension

The manner and context in which information is conveyed is as important as the information itself. For example, presenting information in a disorganized or rapid manner (with too little time to think about it or ask questions), may limit a participant's ability to make an informed choice.

Voluntariness

An agreement to participate in research constitutes a valid consent only if it is voluntary; this requires conditions free of coercion and inappropriate influence.

Web scraping and APIs

Web scraping (also known as web harvesting, web crawling or web data extraction) is any method of copying data from a webpage, usually to then store it in a spreadsheet or database.

Downloading a ready-made .csv file hosted by a site wouldn't be considered web scraping. (Although you might find a programmatic way to download many of these could be.)

What do you need to web scrape?

- Some knowledge of URLs, HTML and CSS
  - URL - Universal Resource Locator
  - HTML - HyperText Markup Language
  - CSS - Cascading Style Sheets
- The `rvest` and `polite` packages in R (or `Rcrawler`, there may be others too) or `Beautiful Soup` for Python
- Professional ethics!

Image description: Abstract painting of an eye symbol wiht lots colours and the text: "Just because you CAN do something, should you?

The Ethical Scraper

I, the web scraper will live by the following principles:

- If you have a public API that provides the data I'm looking for, I'll use it and avoid scraping all together.
- I will always provide a User Agent string that makes my intentions clear and provides a way for you to contact me with questions or concerns.
- I will request data at a reasonable rate. I will strive to never be confused for a DDoS attack.
- I will only save the data I absolutely need from your page. If all I need is OpenGraph meta-data, that's all I'll keep.
- I will respect any content I do keep. I'll never pass it off as my own.
- I will look for ways to return value to you. Maybe I can drive some (real) traffic to your site or credit you in an article or post.
- I will respond in a timely fashion to your outreach and work with you towards a resolution.
- I will scrape for the purpose of creating new value from the data, not to duplicate it.

Source: James Densmore, https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01

Terms and Conditions and Robots.txt

Many sites give instructions about what you're allowed and not allowed to do on them. One way is through the Terms and Conditions and another is through a file called robots.txt.

T&Cs

Ideally, we should all be reading all the Terms and Conditions of all the websites we use...and of course I'm sure you dooooo.

But when in a hurry, search (CTRL+F or CMD+F) “scrape”, “harvest” “crawl” and if none of those come up then “data” and “copied” more generally and that can give you a sense if they prohibit certain uses.

Robots.txt

Most large websites have a robots.txt page to give instructions about what 'robots' are and aren't allowed to visit the page. This is most often used for search engines, but we can check them too. Bad bots can still do what they want. More on these protocols (and templates you could add to your own site here).
http://www.robotstxt.org/robotstxt.html (optional)

An ethical scraper...

- ...follows the site's terms and conditions and/or robots.txt.
- ...uses an API when provided.
- ...rate limits their requests.
  - I.e., respect a 'crawl limit' suggested by the site, 5 seconds is a polite default if not told otherwise.
- ...credits their sources.

Using an API

API stands for **a** pplication **p** rogramming **i** nterface.

It is a structured way for data (broadly) requests to be made and fulfilled with computers.

I like this comparison to a restaurant menu. You don't need to know HOW to make crème brûlée to be able to know you WANT it.

If you are using an API, there still may be rules about things like how many requests you can make in a certain time frame and rate limiting. Make sure you're aware of these rules and behave in the spirit of them!

Optional reading (not assessed): https://beanumber.github.io/mdsr2e/ch-ethics.html#sec:terms-of-use

Indigenous data sovereignty

Countries and nations tend to want data collected and stored about them/their people to be subject to their laws. You might see examples of this in how government agencies require any cloud storage they use to have the servers be based within their boundaries. This area of thinking is often called **data sovereignty**.

New Zealand is one of many countries with a history of colonization by settler peoples and the displacement of, discrimination against, and in many cases mass murder of the Indigenous peoples. As countries like New Zealand (and Canada, my birth country) go through processes of truth and reconciliation to address these violent and oppressive histories, *indigenous data sovereignty* has also become a growing area of thought.

Indigenous data sovereignty

Why do data scientists needs to know about this? Because we must move from data gathering and analysis as further tools of oppression and be part of honouring the sovereignty of Indigenous peoples and nations over their own data.

1. Be aware of Indigenous rights and interests in relation to data.
2. Understand protocols for consulting with Indigenous peoples about data collection, access and use.
3. Ensure data for and about Indigenous peoples we are given access to is safeguarded and protected.
4. Support quality and integrity of Indigenous data and its collection.
5. Advocate for Indigenous involvement in the governance of data repositories.
6. Support the development of Indigenous data infrastructure and security systems.

Several of these points are generalized from: https://www.temanararaunga.maori.nz/kaupapa.

Selection bias

Selection bias can occur in a range of ways, but the key feature is that your sample is not representative of the population.

Example: Suppose I want to email out a survey to investigate if U of T students think statistics is important for their future career. I only have the emails for students I teach...in statistics courses. If I randomly sample from this list of students, can I make claims about the population of all U of T students? No.

Example: *The healthy migrant effect*. It has been noticed in many countries that migrants have mortality advantages over local-born populations. While there are several possible things going on and being researched in this area, these findings likely show a component of 'selection bias' in that usually only healthy people can migrate, either due to health screening requirements in the country to which they are migrating, or by people with health complications self-selecting out due to inability/disinclination.

Survivorship bias is a specific type of selection bias.

Survivorship bias

Note: There is a risk that I'm about to ruin a bunch of 'inspirational' internet content for you.

If you've spent any time on LinkedIn, and probably lots of other social media sites, you've probably seen an image like this one. Is it inspiring? Sure, maybe.... BUT as an attempted claim about the value of failure it commits the logical error of focusing on just the people who eventually succeeded. I'm sure there are plenty of unemployed, divorced, university dropouts *not* writing *Parry Hotter and the Windows OS*...

Click for source.

Analyzing data

📚 Optional readings

The following optional readings discuss on the reproducibility crisis, P-hacking and HARKing. I won't cover P-hacking or the reproducibilty crisis any further, as these readings are great introductions, but the next few slides talk a little more about **HARKing** and **multiple testing problems**.

- Motulsky, H.J., (2014). *Common misconceptions about data analysis and statistics.* https://doi.org/10.1007/s00210-014-1037-6
- Aschwanden, C. (2015). *Science Isn't Broken: It's just a hell of a lot harder than we give it credit for*. Retrieved from https://fivethirtyeight.com/features/science-isnt-broken

HARKing

HARKing is "**H** ypothesizing **A** fter the **R** esults are **K** nown".

I sometimes talk about this as the 'no peeking rule' in setting up hypotheses. For example, you should never pick a one-tailed hypothesis test because of your data, it should only be based on findings from previous studies or a physical theory of a phenomenon.

There is a good introduction to this idea in the Motulsky paper. (Think the XKCD jelly bean comic!)

There are critiques about whether HARKing is as harmful to science as is sometimes claimed and a lot of yummy philosophy of science that we won't go into. If you're interested in this area, try reading Rubin (2017), "When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress" (optional).

Multiple testing problem

One of my favourite examples of positive academic trolling is the dead salmon study. The study used methodology for exploring animal reactions to human emotions expressed in photographs through fMRI scans.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

If the researchers had ignored the issue of multiple comparisons (there are thousands of areas for which brain activation is measured) they might have ended up with test results that claimed the (dead) salmon was engaging in 'perspective-taking' when shown the photos of the humans. (It was not).

See the poster here: http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf

Multiple testing: Salmon

This is the reason we would want to run an ANOVA when we have more than two groups to compare the means of, instead of doing multiple t-tests between every pair of levels and it is mentioned in conjunction with HARKing in Motulsky paper.

But what is actually going on here?

Suppose you've picked a significance level of $\alpha = 0.05$. When conducting just one test this means we're accepting a 5% risk of making a Type 1 Error, that is, rejecting the null hypothesis when we shouldn't. BUT, if we are conducting several tests at the same time, then we need to think about our **family-wise error rate**, which is our chance of making at least one Type 1 Error across all our tests.

So, if you are making a Type 1 Error 5% of the time, the idea is that 95% of the time you're not. And while 95% is pretty good, if you're doing m independent tests, the not-making-an-error rate becomes 0.95^m, e.g., if you're doing 20 tests, $0.95^{20} = 0.359$, meaning the chance of making at least one 'false discovery' is now $\sim 64\%$

Correcting for multiple comparisons?

There are multiple methods investigators employ in an effort to have their Type 1 Error across multiple tests actually reflect the error rate they are comfortable with.

The simplest but most conservative of these is the Bonferroni correction where you just divide your significance level (e.g., 0.05) by the number of tests you are conducting and use that as the new significance cut off.

E.g., If you'd usually use a 5% threshold and are doing 20 tests, your new threshold is

$$\alpha_{adjusted} = \frac{0.05}{20} = 0.0025.$$

As you can see, this is now a much stronger level of evidence we're requiring against our null hypothesis than when doing a single test.

Making decisions with data

Algorithmic bias

Prediction models are taught what they "know" from training data. Training data can be incomplete, biased, or skewed. This can result in **algorithmic bias**.

Proxy variables

There can also be situations where we know we DON'T want to use a variable as part of an algorithm, for ethical and often legal reasons (anti-discrimination laws about gender, race, health status, e.g. American's with Disabilities Act means you can't discriminate against people with mental health conditions). BUT there might be other variables in your data, like certain types of hobbies/memberships, home address, 'personality' quiz questions, that act as 'proxies' for these things, meaning they end up determining outcomes even when you don't want them to.

📚 Optional readings

- Amazon scrapped 'sexist AI' tool, BBC, 2018. https://www.bbc.com/news/technology-45809919
- Amazon discreetly abandoned gender-biased AI-based recruiting tool, HRK News, 2018 https://www.hrkatha.com/recruitment/amazon-discreetly-abandoned-gender-biased-ai-based-recruiting-tool/

Should algorithms be transparent?

Some predictive algorithms give us more than just a prediction: they also give us some insight as to what factor(s) influenced the prediction. Examples you might have encountered in your studies already include linear regression models and classification trees.

Other algorithms yield predictions, but no information about how it got from the inputs to the prediction, such as neural networks (you may see these in future courses). These are sometimes called 'black box' algorithms and many machine learning tools fall into this. We'll briefly talk more about this when we discuss generalized additive models later in the course.

What is more important—getting the most accurate predictions, or understanding the factor(s) which influence a prediction?

When you approach a new statistical problem, try to figure out if you are aiming to make **predictions**, make **inferences** about relationships between variables, or provide informative **summaries and descriptions** of the data.