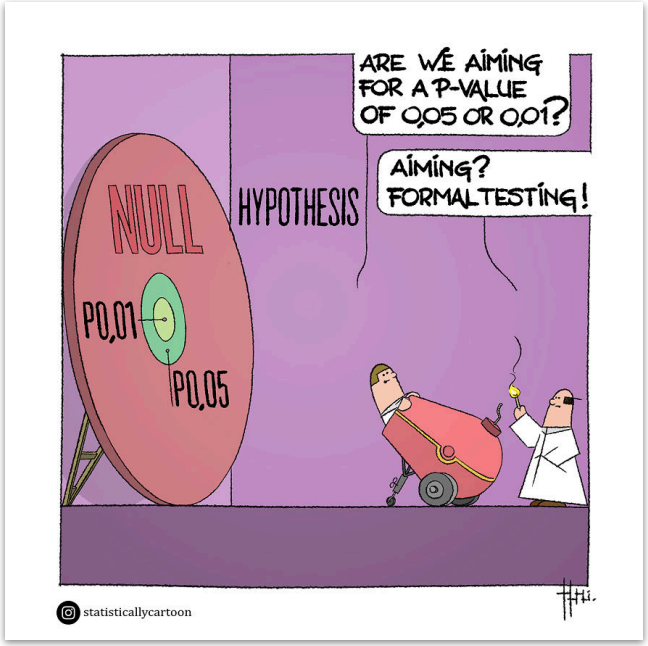
PSYCH 252: Statistical Methods for Behavioral and Social Sciences

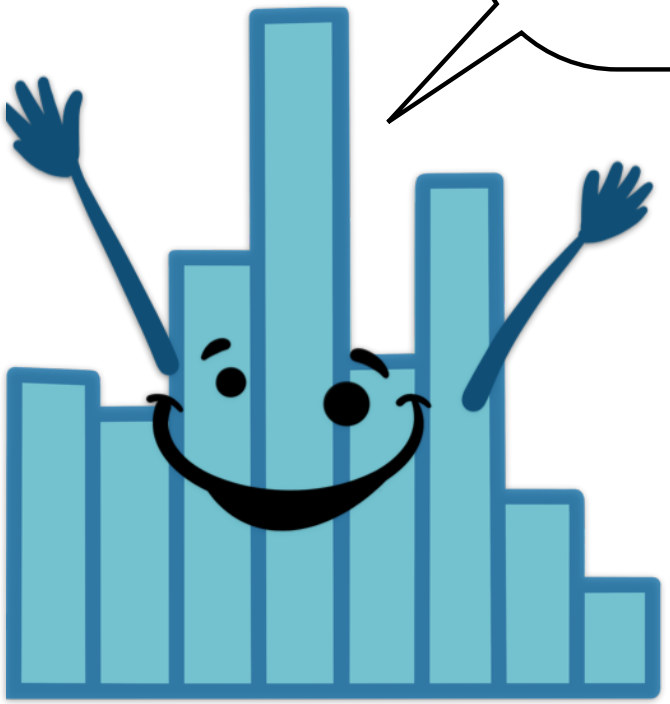
Introduction



add songs that get you into stats mood 📈

01/05/2026

Welcome!



many of the illustrations we use in class are by

Allison Horst
allisonhorst

https://allisonhorst.com/

Outline

- Introduction
- **What** will we learn?
- **How** will we learn?
- Tools we will use in class
- Some general thoughts on learning Stats in Social Science
- Feedback

Introduction

Introduction to the class

master the tools that empower you to do your research

develop statistical literacy

Introduction to the team

Tobi Gerstenberg

Pronouns: he/him

Department: Psychology

Interests:

- running the Causality in Cognition Lab
- computational modeling

Fun facts:

- I like surfing and R

Nilam Ram

Pronouns: he/him

Department: Psychology / Communication

Interests:

- running The Change Lab @ Stanford
- longitudinal design and data analysis

Fun facts:

- Trying out new ways to make data into music (data sonification)

Justin Yang

Pronouns: he/him

Department: Psychology

Interests:
- mental simulation
- learning

Fun facts:
- I like cooking

Verity Lua

Pronouns: she/her

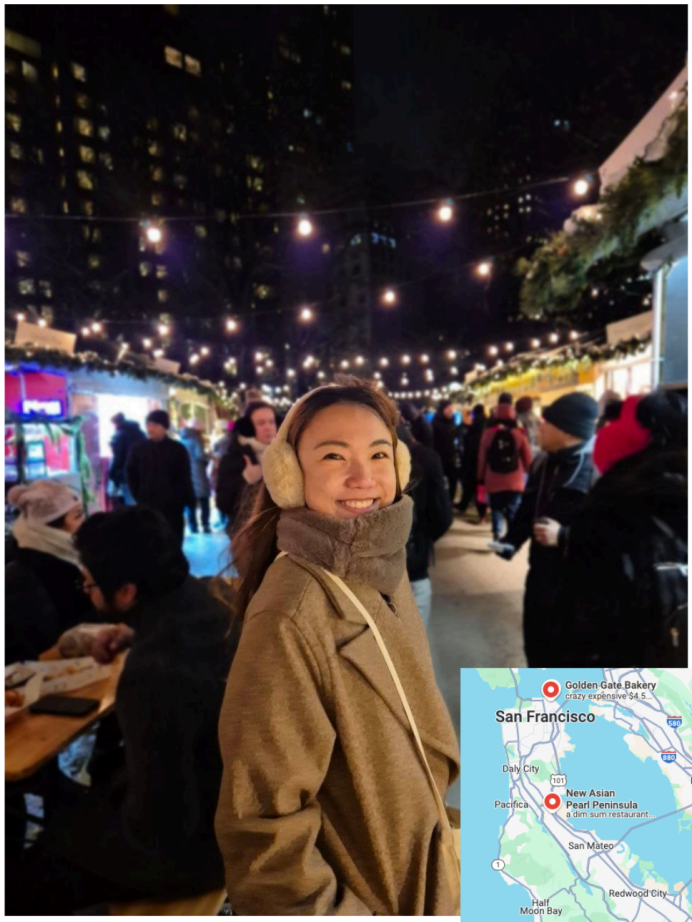
Department: Psychology

Interests:

- culture and emotions

Fun facts:

- I'm on a quest to find the best egg custard tarts around



Alvin Tan

Pronouns: he/him

Department: Psychology

Interests:

- language acquisition
- pragmatics
- machine learning dynamics

Fun facts:

- I really dislike eating the same meal multiple times in a row

Shashanka Subrahmanya

Pronouns: he/him

Department: Psychology

Interests:

- AI (LLMs), longitudinal methods
- Mental health, wellbeing

Fun facts:

- Likes chai, anime, hiking (in that order)

Cynthia Wu

Pronouns: she/her

Department: Psychology

Interests:

- Emotions & decisions

Fun facts:

- I did a lot of baking and ended up filling my fridge instead of my stomach

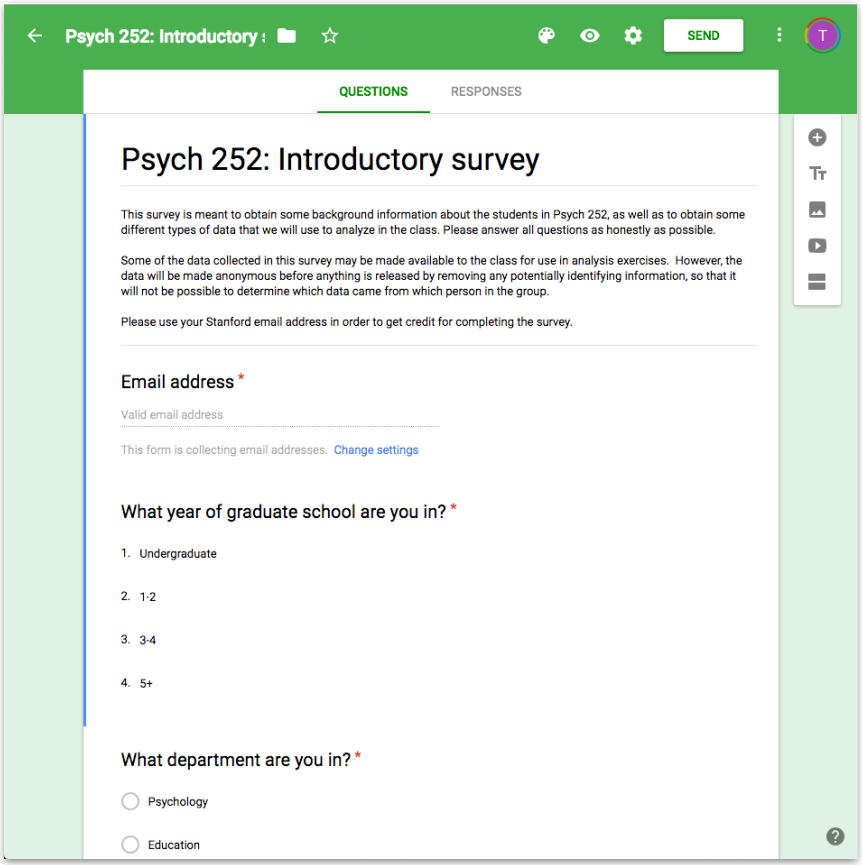
What did you do during the winter break?

What are you hoping to get out of this class?

(5 minutes)


05:00

Introduction



https://tinyurl.com/psych252survey26

What will we learn?

Learning goals

1. You will learn how to **use R** to ...
   - read, wrangle, simulate and analyze data
   - make publication-ready plots

2. Understand the philosophy behind **null hypothesis significance testing (NHST)** and **Bayesian statistics** through ...
   - running computer simulations and visualizing the results

3. Formulate **research questions as statistical models** and ...
   - determine which models work best for different situations

4. **Communicate** what you have learned about your data ...
   - in short presentations in class, showcasing your visualization and analysis
   - in written reports

5. Contribute to open and **reproducible science** through ...
   - adopting good coding practices
   - sharing your data and research reports online

What will we learn?

Weeks 1-3

1. Use R!

- Data visualization
- Data manipulation/wrangling
- Understand key statistical concepts
  - Simulation, manipulation, visualization

2. Build models

Weeks 4-7

- Formulate hypotheses as statistical models

Weeks 8-9

- Bayesian statistics

all the time (& Week 10)

3. Report results

- Reproducible research

What will we learn?

1. Use R!

4 Reasons for why  ?

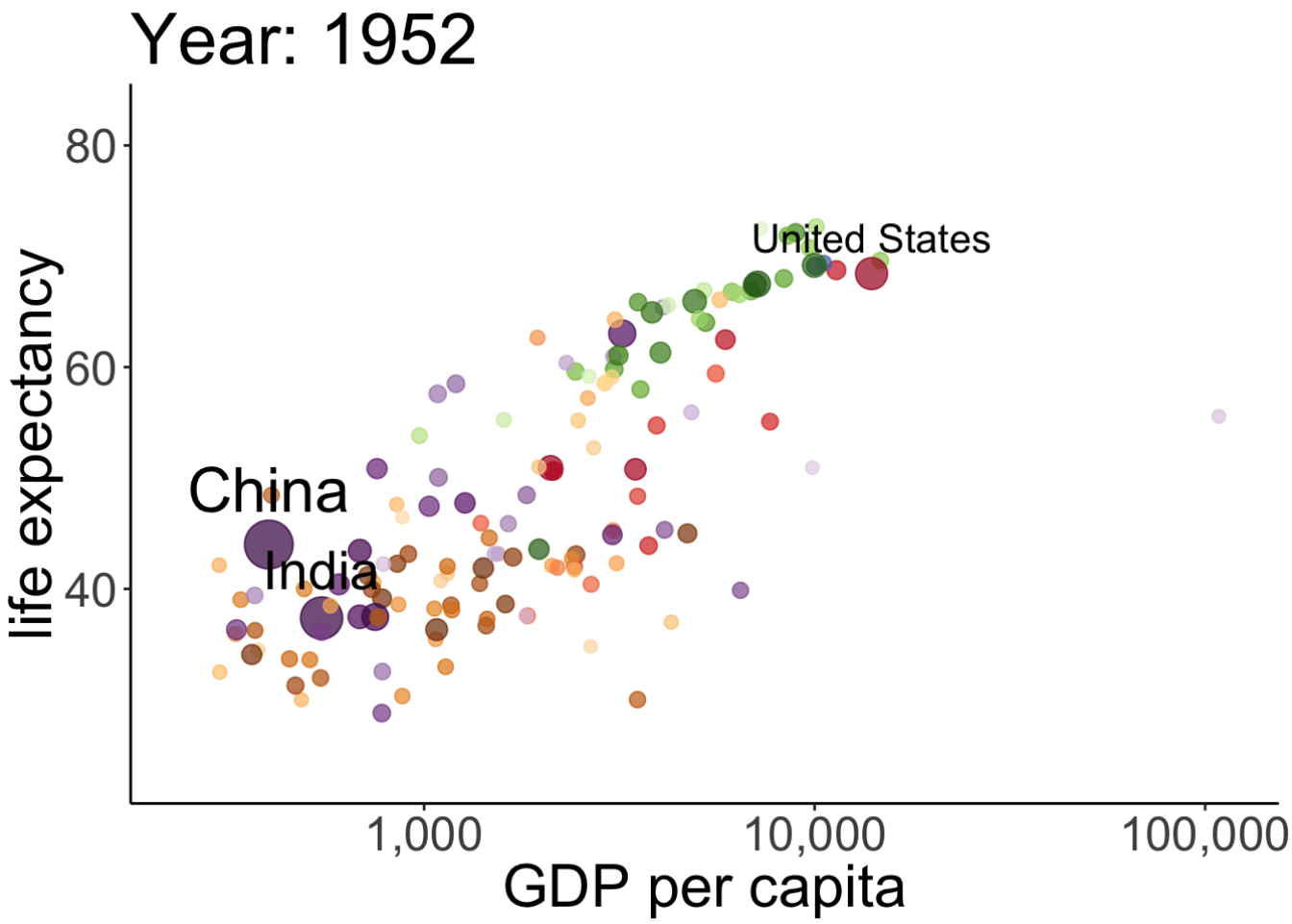
a. Visualization

b. Data manipulation

c. Statistical modeling

d. Communicating results

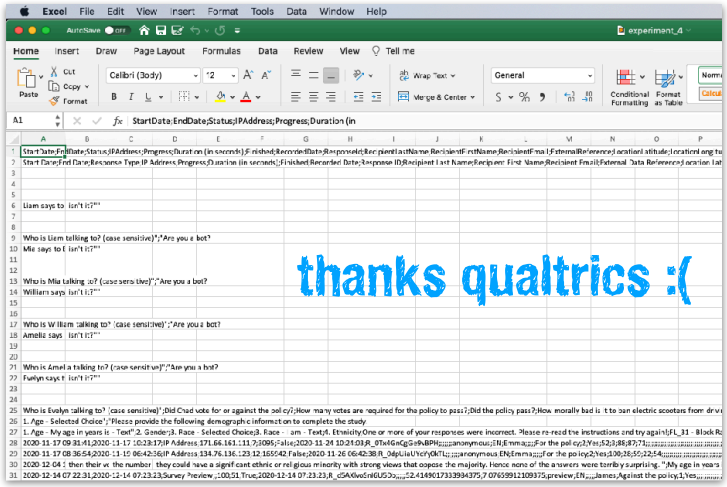
Why R ? **a. Visualization**

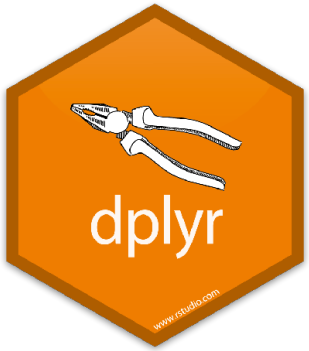


Why  ? a. Visualization

```
ggplot(gapminder, mapping = aes(x = gdpPercap, y = lifeExp, size = pop, color = country)) +
  geom_point(alpha = 0.7, show.legend = FALSE) +
  geom_text(data = gapminder %>% filter(country %in% c("United States", "China", "India")),
            mapping = aes(label = country),
            color = "black",
            vjust = -0.75,
            show.legend = FALSE) +
  scale_colour_manual(values = country_colors) +
  scale_size(range = c(2, 12)) +
  scale_x_log10() +
  labs(title = "Year: {frame_time}", x = "GDP per capita", y = "life expectancy") +
  transition_time(year) +
  ease_aes("linear")
```

Why R ? b. Data manipulation

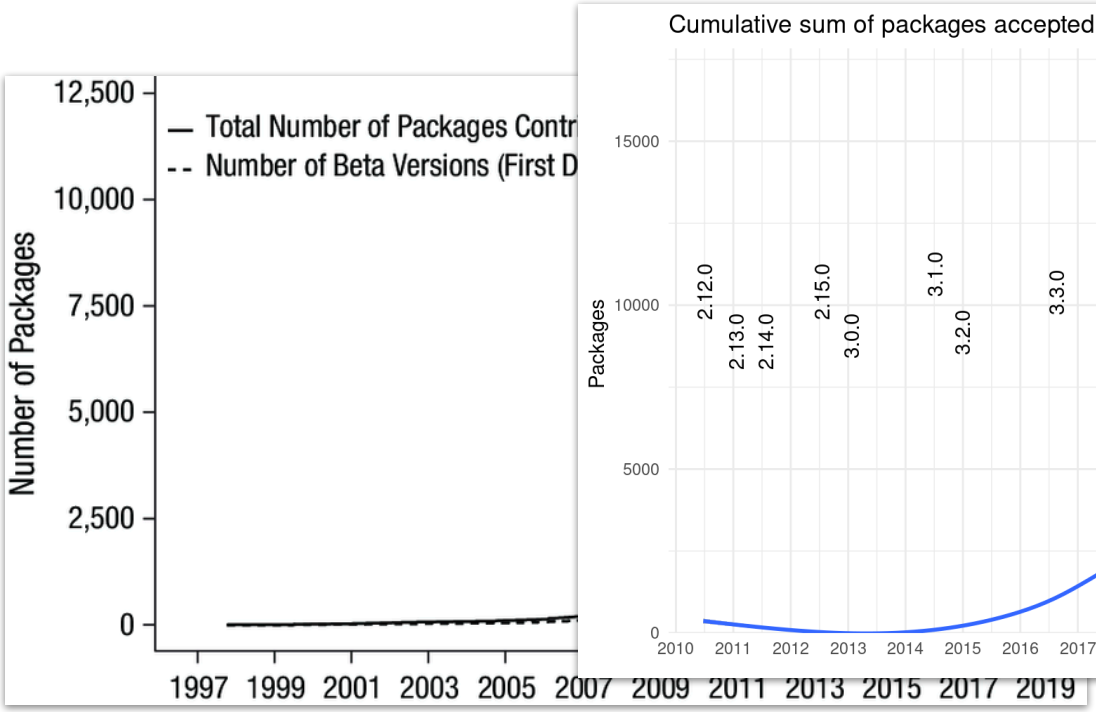


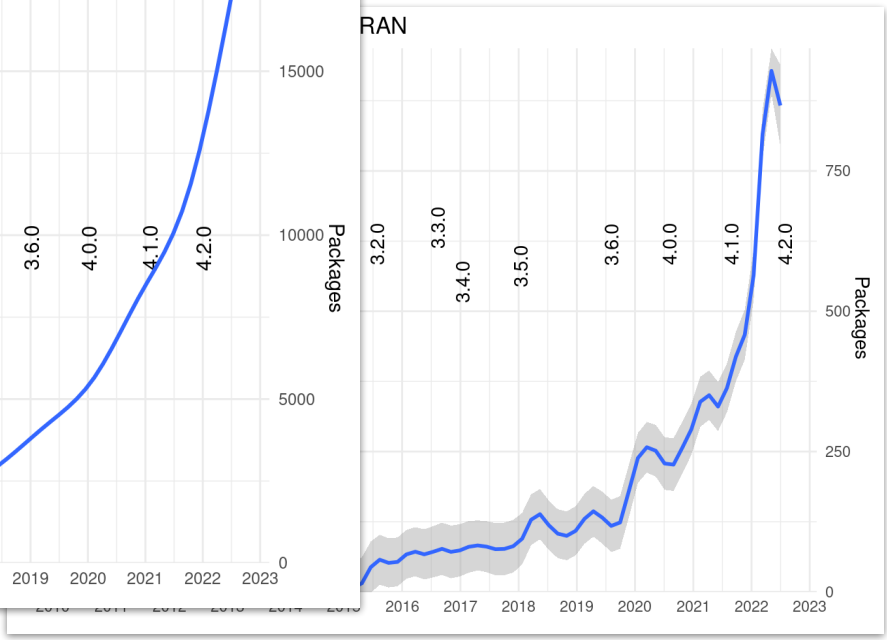


Making Data Wrangling Suck Less

Why R ? **c. Statistical modeling**

CRAN = The comprehensive R network





This is what R was developed to do.
(and it continues to develop!!)

Why R? **d. Communicating results**

01-intro.Rmd

```
# Introduction {#intro}

You can label chapter and section titles using `{#label}` after them,
do not manually label them, there will be automatic labels anyway, e.

Figures and tables with captions will be placed in `figure` and `tabl


```{r nice-fig, fig.cap='Here is a nice figure!', out.width='80%', fi
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.
can reference tables generated from `knitr::kable()`, e.g., see Table

```{r nice-tab, tidy=FALSE}
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

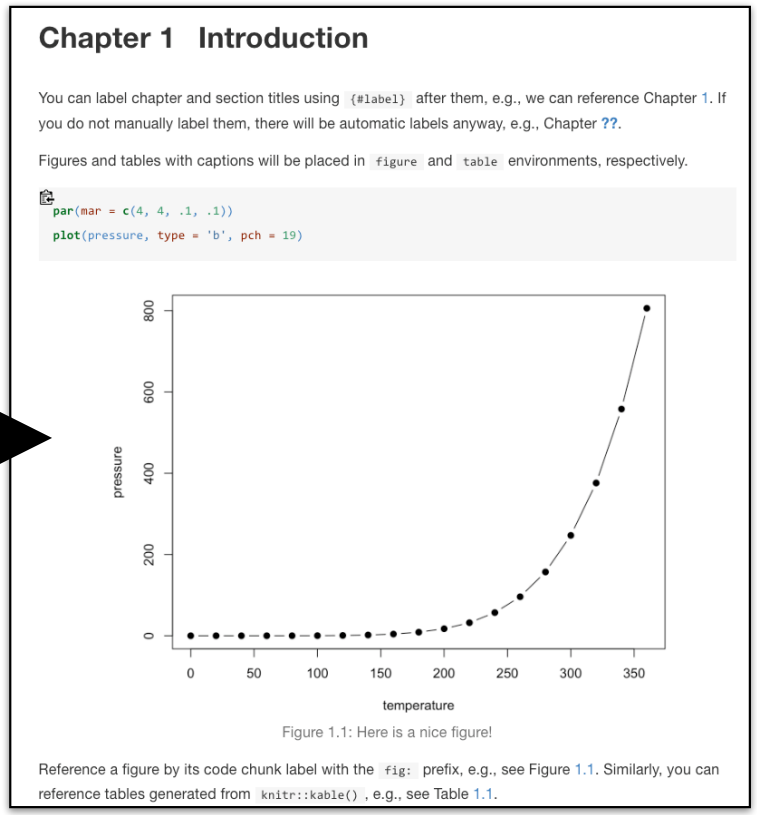
You can write citations, too. For example, we are using the **bookdow
was built on top of R Markdown and **knitr** [@xie2015].

```{r stats-help, fig.cap='Stats cheatsheet',fig.align='center',echo=
knitr::include_graphics('figures/cheatsheets/stats-help.jpg')
```

See figure \@ref(fig:stats-help)


```{r klippy, echo=FALSE, include=TRUE}
klippy::klippy()
```
```

R Markdown



Chapter 1 Introduction

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 1. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter ??.

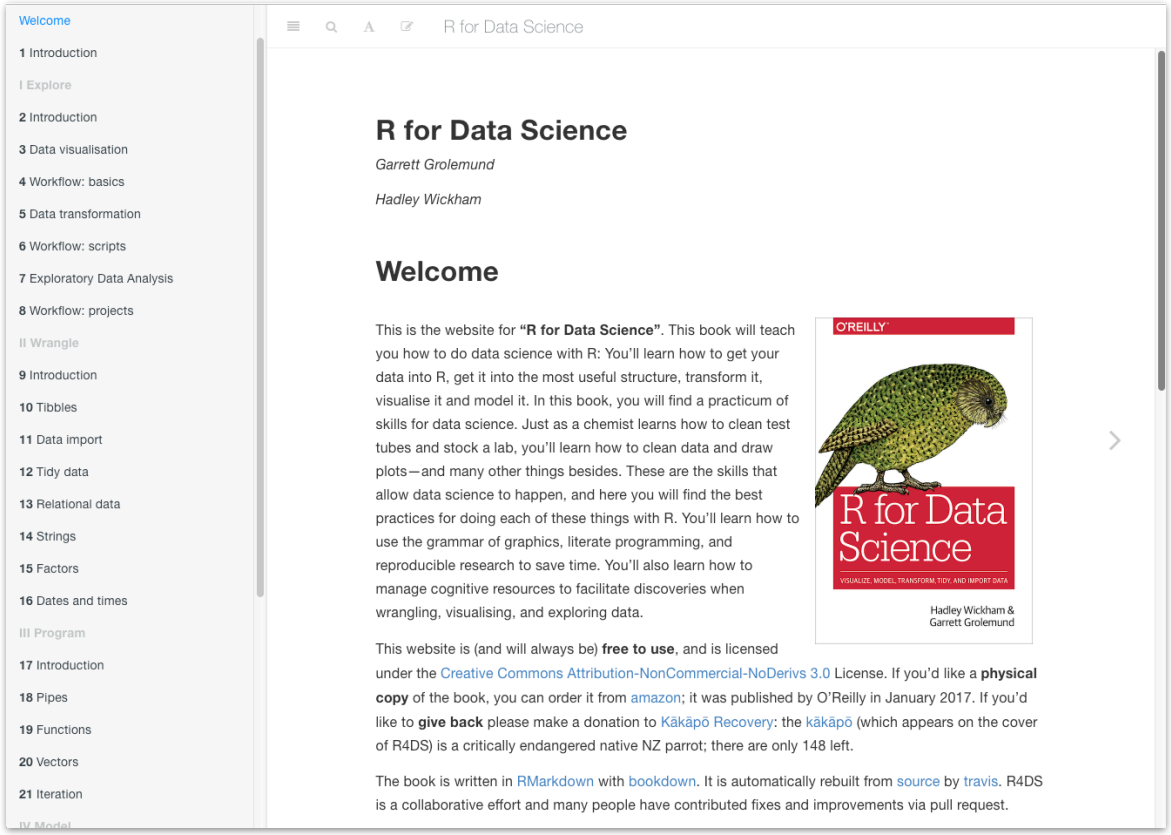
Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```


Figure 1.1: Here is a nice figure!

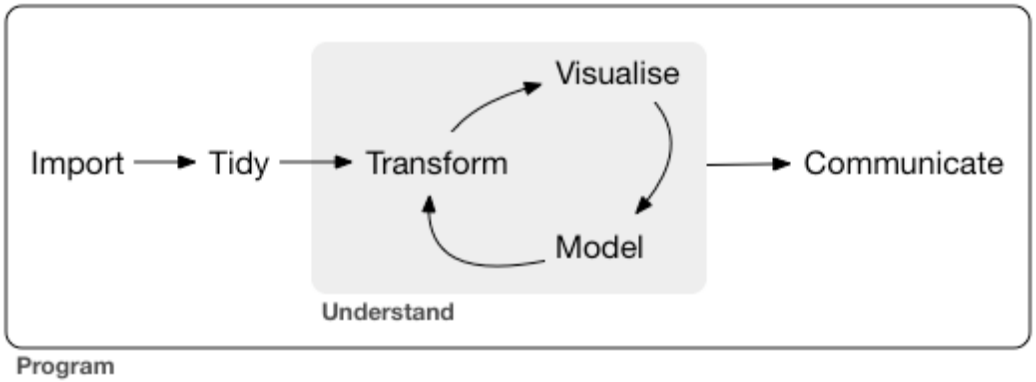
Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 1.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 1.1.

Why R ? **d. Communicating results**



https://r4ds.had.co.nz/

Why R ? **Tidyverse**



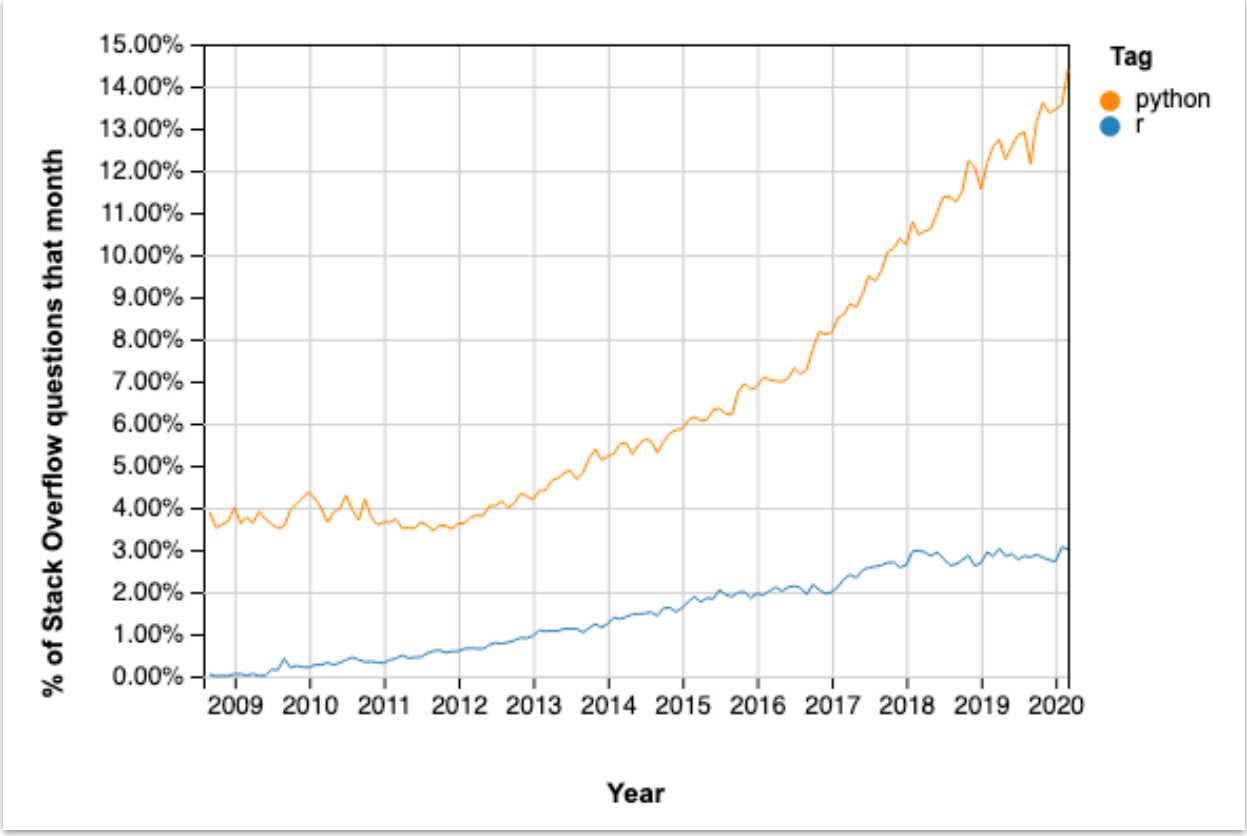
- *Import*: get data into R
- *Tidy*: clean and format the data
- *Transform*: select variables, create new ones, group and summarize
- *Visualize*: "look" at the data in different ways
- *Model*: answer questions about the data
- *Communicate*: write reproducible research reports

Grolemund & Wickham (2016). R for data science: import, tidy, transform, visualize, and model data

Why  ? **Also ...**

- many researchers use R (including many of us at Stanford)
- R has a nice online community. #rstats
- RStudio is a great IDE (Integrated Development Environment)!

But what about python™ ?



But what about  ?

- Yes, you should learn python, too! :)
- Python is great for machine learning (e.g. deep learning, LLMs, ...).

visualization

data manipulation

statistical modeling

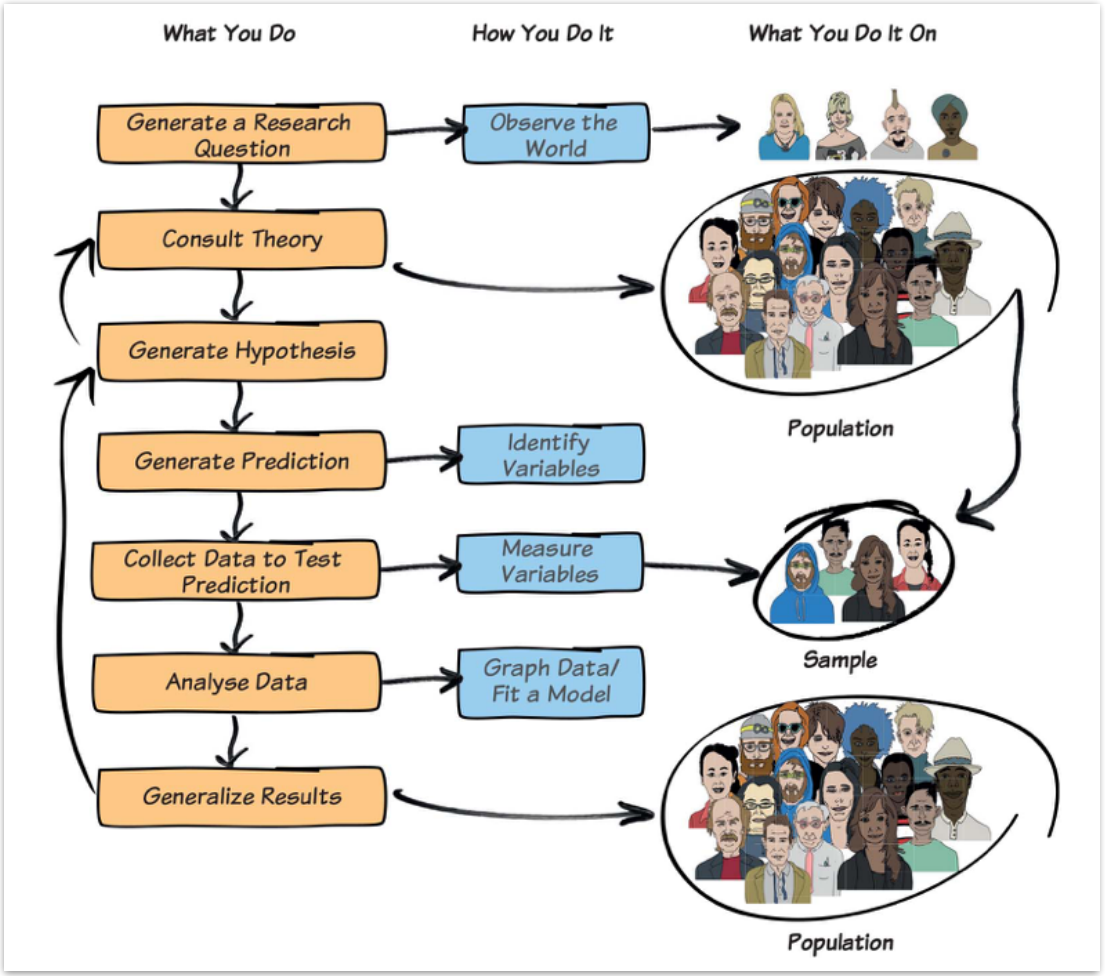
reporting

python™

What will we learn?

2. Build models

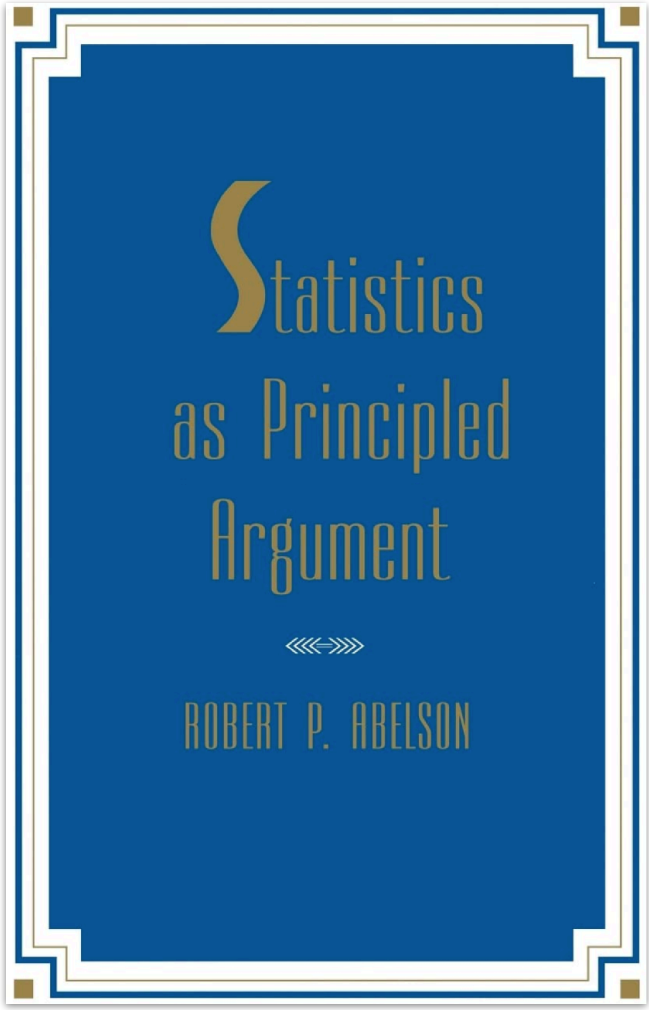
The research process



Statistics as argument

Researchers use empirical observations (data) for **making arguments about research questions**.

Statistics is part of a narrative and requires **figures**, context, rhetoric, prose, ...



Abelson, R. P. (1995). *Statistics as principled argument*. New York: Psychology Press.

helps us with building models

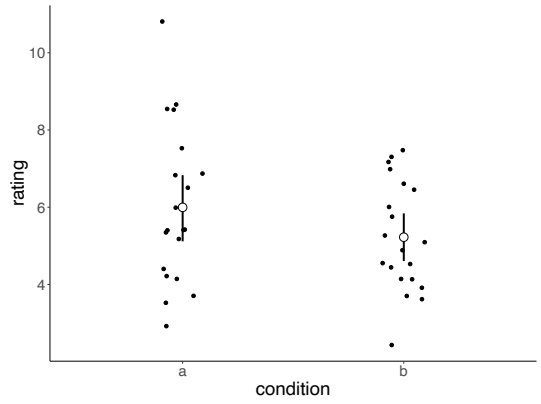
instead of

$$\sum_{\mu,\nu,\lambda} \hat{M}^{(3)}_{\mu\nu\lambda} F_{\mu\nu\lambda} = \hat{M}^{(3)}_{000} F_{000} + 3 \sum_i \hat{M}^{(3)}_{00i} F_{00(i)} + 3 \sum_i \hat{M}^{(3)}_{0ii} F_{0ii} + 3 \sum_{i<j} \hat{M}^{(3)}_{0(ij)(ij)} F_{0(ij)(ij)}$$
$$+ 6 \sum_{i<j} \hat{M}^{(3)}_{0(ii)(ij)} F_{0(ii)(ij)} + 3 \sum_{i<j} \hat{M}^{(3)}_{0(ij)(ij)} F_{0(ij)(ij)} + 3 \sum_{i \neq j} \hat{M}^{(3)}_{ii(ij)} F_{ii(ij)} + 6 \sum_{i<j} \hat{M}^{(3)}_{ij(ij)} F_{ij(ij)} + \cdots$$
$$= S^{(0)} F_{000} - \frac{2S^{(1)}}{d} \frac{\sigma_1^2}{\sigma_0^2} \left(2 \sum_i F_{00(ii)} - \sum_i F_{0ii} \right) - \frac{2S^{(2)}}{d^2(d+2)} \frac{\sigma_1^4}{\sigma_0^4}$$
$$\times \left[3(d-1) \sum_i F_{0(ii)(ii)} - 6 \sum_{i<j} F_{0(ij)(ij)} + \frac{3d}{2} \sum_{i<j} F_{0(ij)(ij)} + (d+2) \sum_{i \neq j} F_{ii(ij)} - (d+2) \sum_{i<j} F_{ij(ij)} \right]$$
$$+ \frac{3S_2^{(2)}}{d(d+2)} \frac{\sigma_1^4}{\sigma_0^4} \left(3 \sum_i F_{0(ii)(ii)} + 2 \sum_{i<j} F_{0(ii)(jj)} + \sum_{i<j} F_{0(ij)(ij)} \right) + \cdots . \quad (85)$$

(google search "complicated statistical equation")

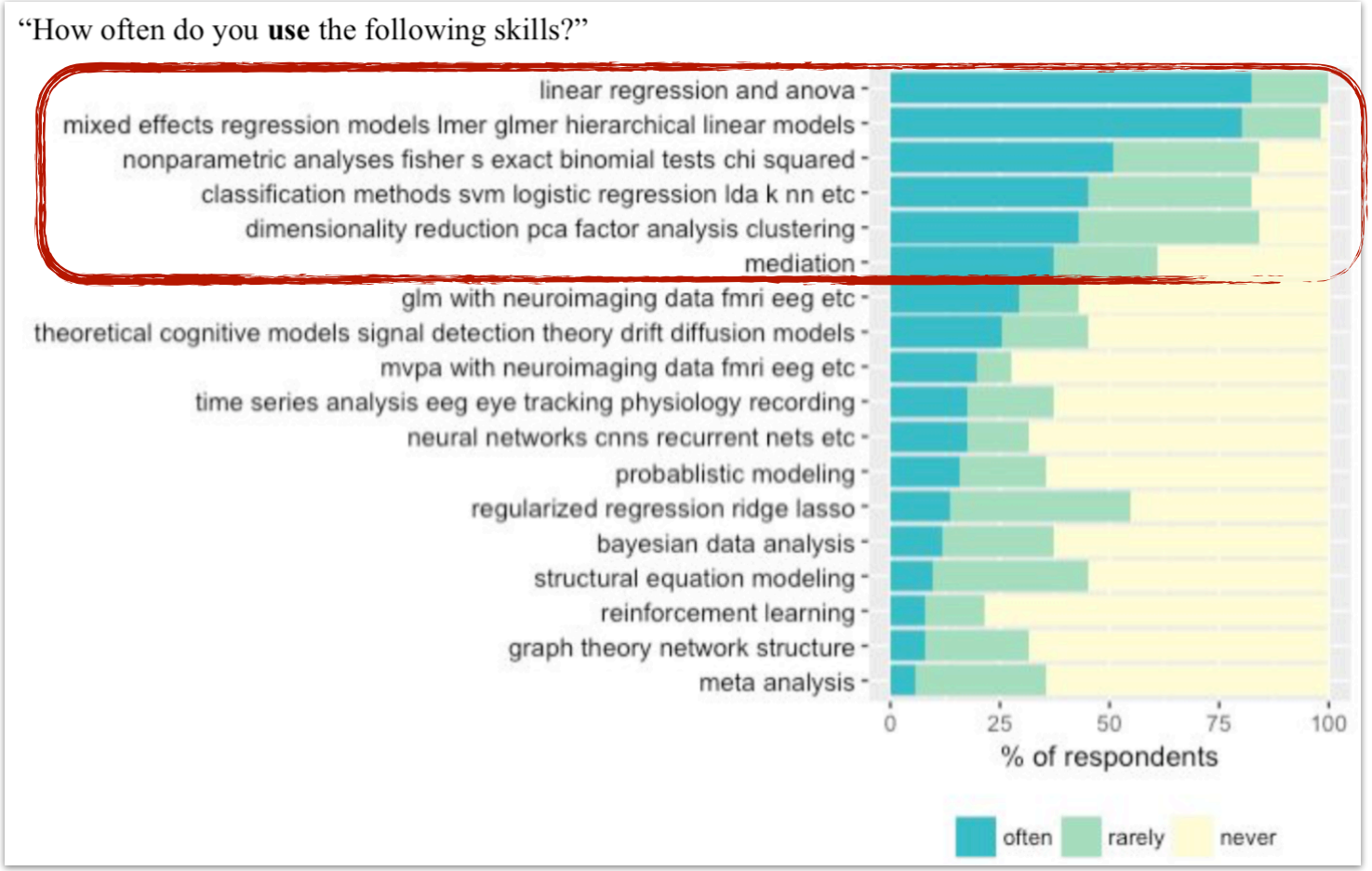
we'll do

```
1  set.seed(0)
2  df.data = tibble(
3    a = rnorm(20, mean = 6, sd = 2),
4    b = rnorm(20, mean = 5, sd = 2)
5  ) %>%
6    gather("condition", "rating")
7
8  df.data %>%
9    group_by(condition) %>%
10   summarize(rating.mean = mean(rating),
11             rating.sd = sd(rating)) %>%
12   kable()
13
14 # calculate the difference between conditions
15 difference.actual = df.data %>%
16   group_by(condition) %>%
17   summarize(rating.mean = mean(rating)) %>%
18   pull(rating.mean) %>%
19   diff() %>%
20   -.
```

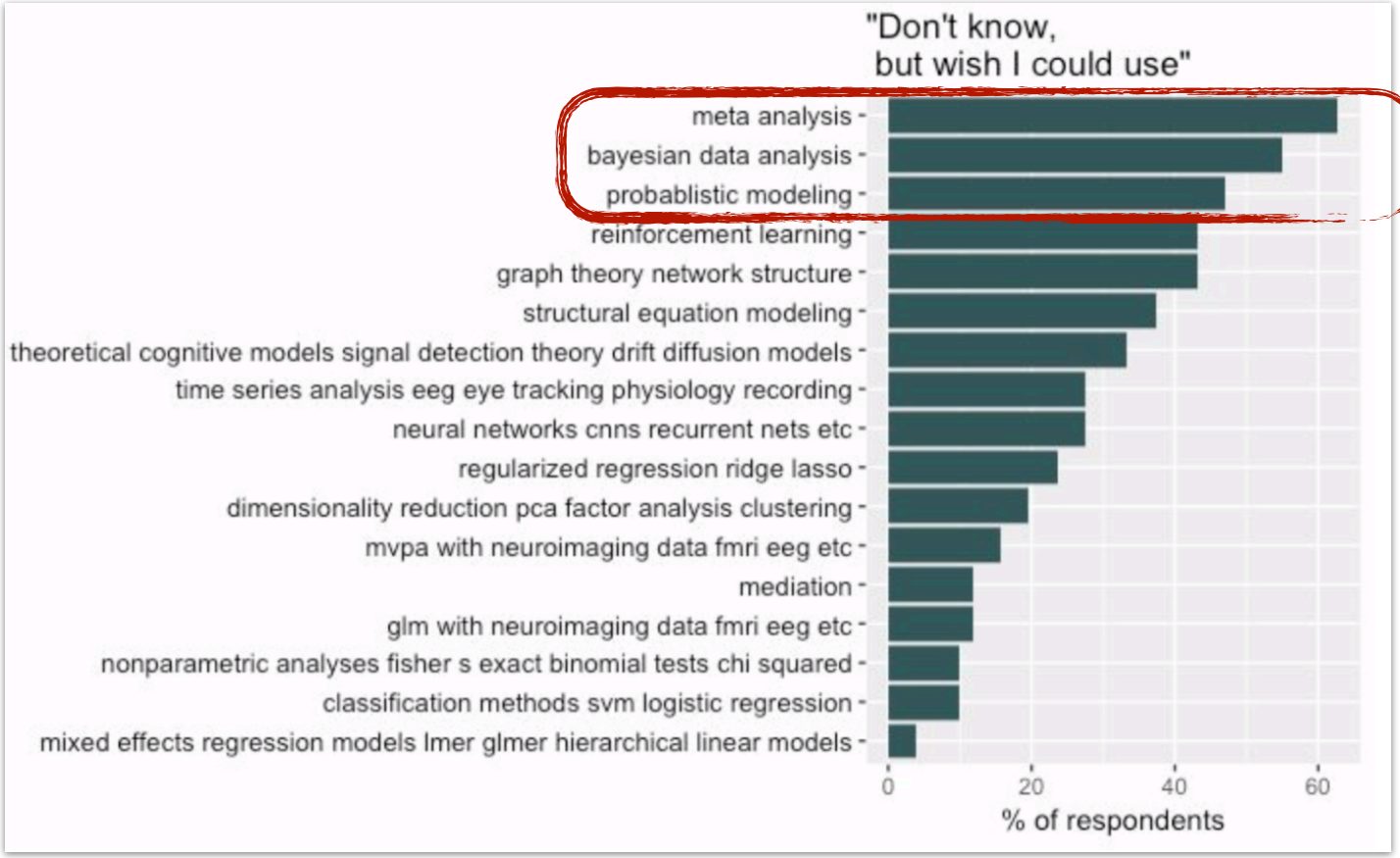


(although Nilam does often require learning the equation …)

Psychology grad student survey



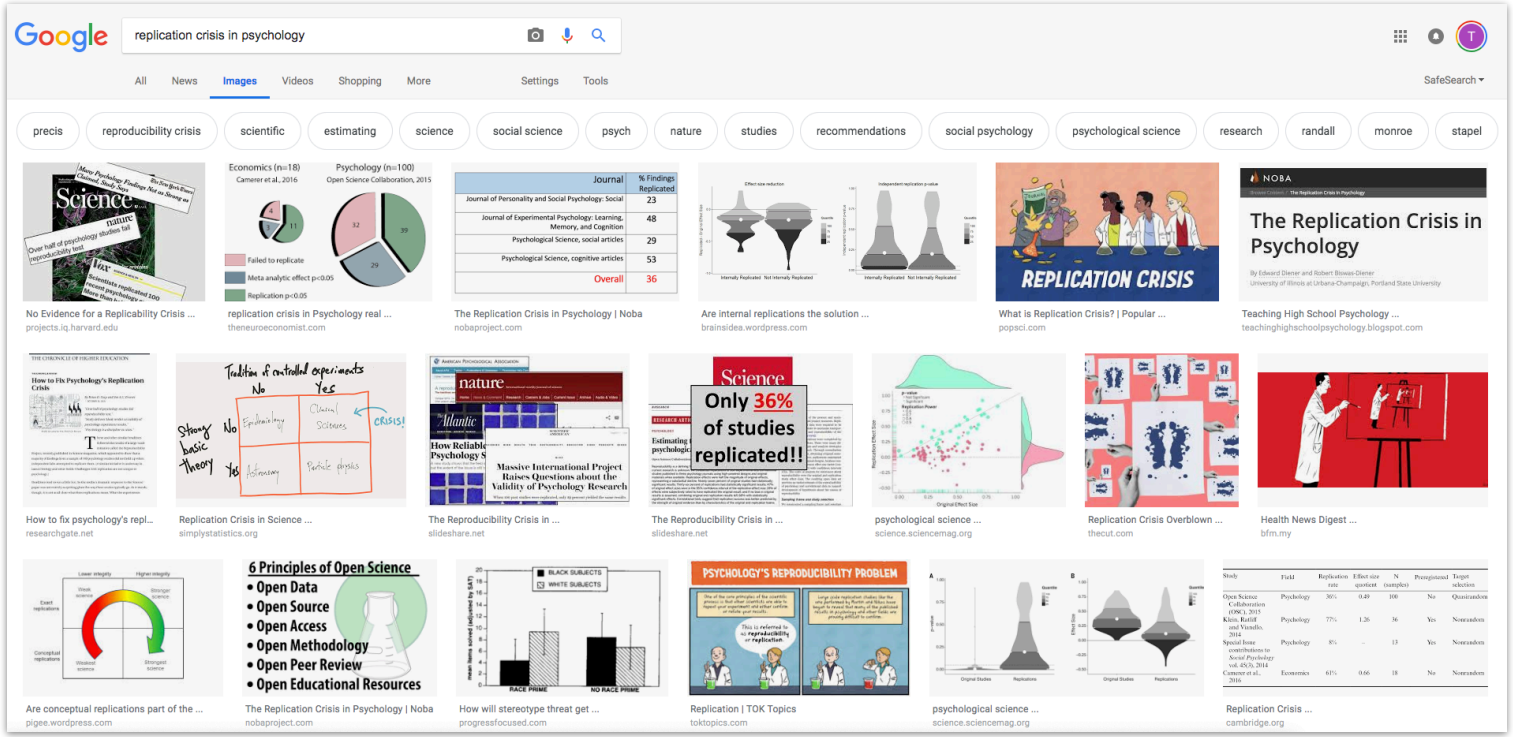
Psychology grad student survey



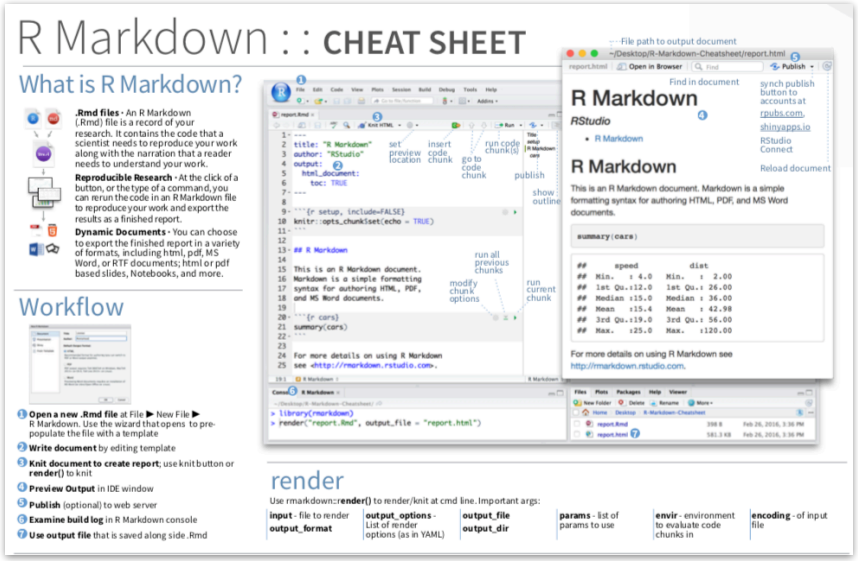
What will we learn?

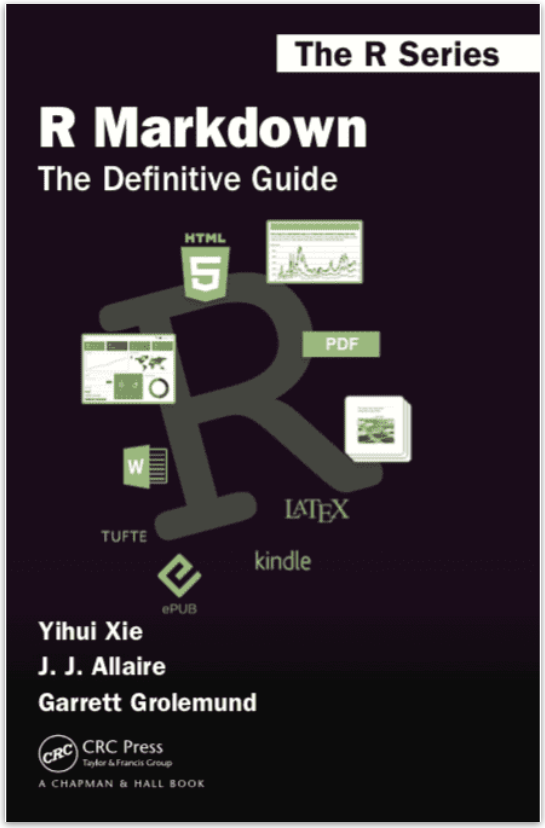
3. Report results

Replication crisis



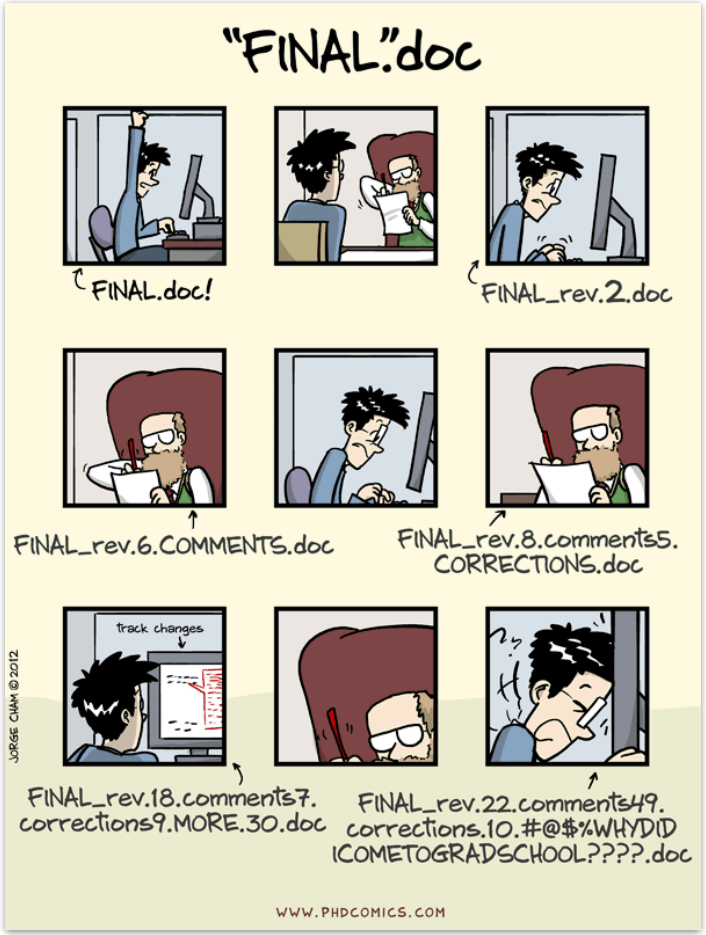
R Markdown





https://bookdown.org/yihui/rmarkdown/

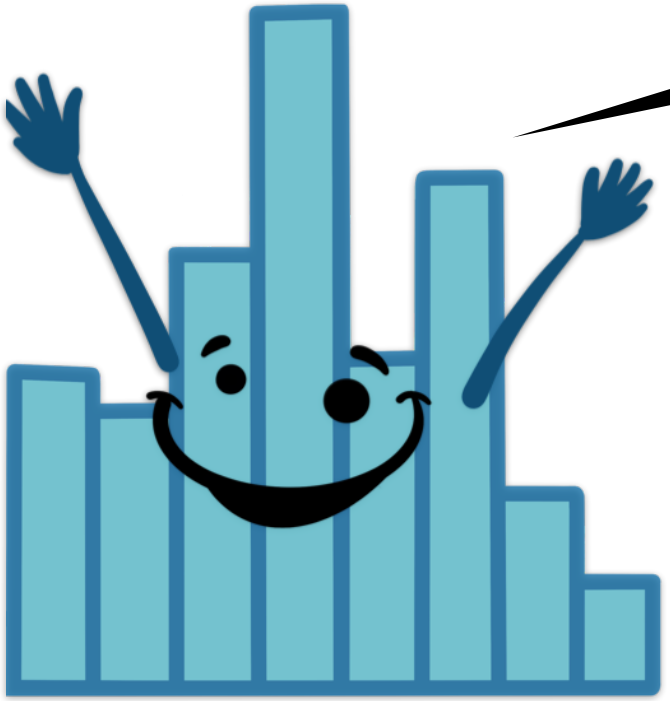
R Markdown



- if we use a word document ... :
  - figures change
  - results change
  - copy and paste is error prone
- in R Markdown ... :
  - figures and statistics are updated
  - no need for copy and paste
  - everything in one place
  - even better with version control (e.g. via github)

01:00

stretch break!



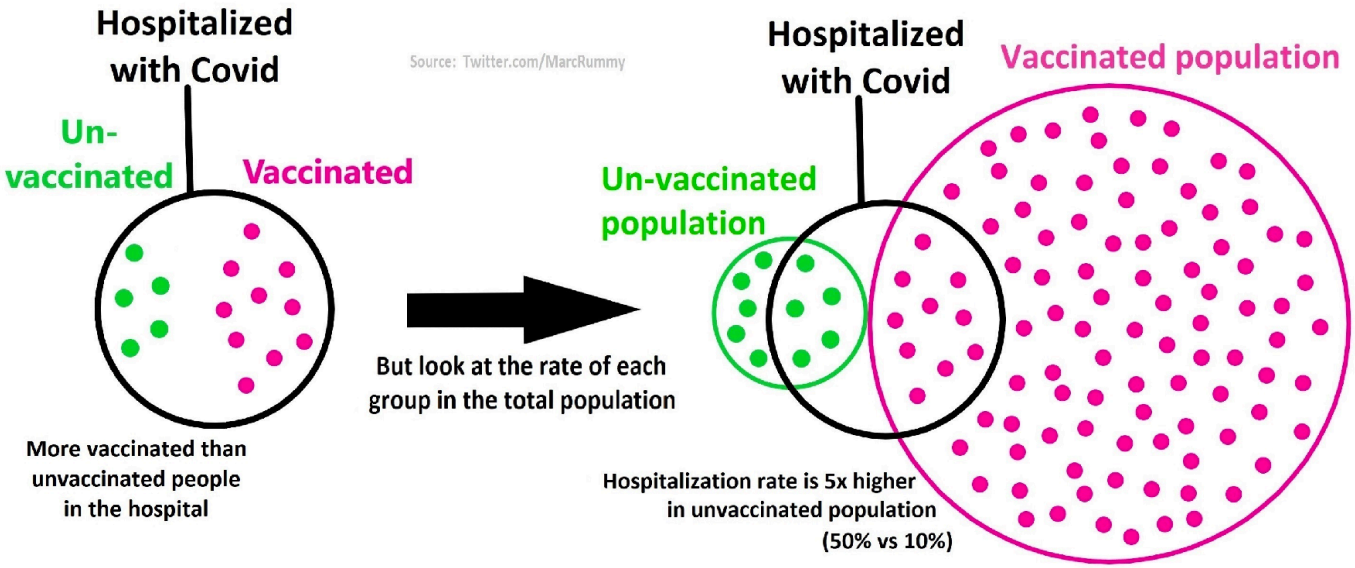

Ray
@ray4tesla
Subscribe
80% of Covid patients put on ventilators died. This is probably most shocking revelation unknown to the public.
0:37
2:38 PM · Oct 31, 2023 · 10.1M Views
2.2K
2.5K
4.5K
565

← **Post**

Daniël Lakens
@lakens

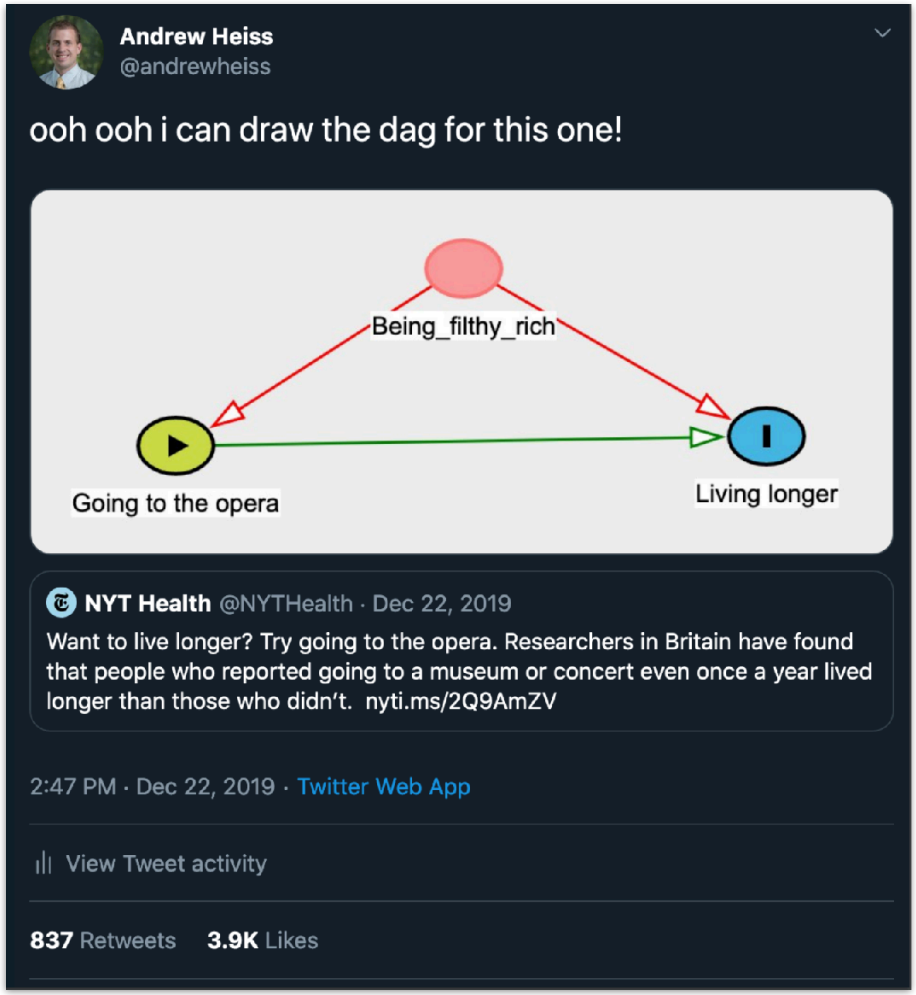
This is why people need basic training in logic (in this case about the positive predictive value). After selecting patients on 'needing to go on a ventilator' many will die, whatever you do.

More vaccinated people are hospitalized with COVID than unvaccinated people!



Note: The ratios presented are made to illustrate the concept of the base rate fallacy when the vaccination rate is high


NYT Health
@NYTHealth
Want to live longer? Try going to the opera. Researchers in Britain have found that people who reported going to a museum or concert even once a year lived longer than those who didn't.
Another Benefit to Going to Museums? You May Live Longer
Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.
nytimes.com
9:19 AM · Dec 22, 2019 · SocialFlow
336 Retweets 1.3K Likes


Andrew Heiss
@andrewheiss
ooh ooh i can draw the dag for this one!
Being_filthy_rich
Going to the opera
Living longer
NYT Health @NYTHealth · Dec 22, 2019
Want to live longer? Try going to the opera. Researchers in Britain have found that people who reported going to a museum or concert even once a year lived longer than those who didn't. nyti.ms/2Q9AmZV
2:47 PM · Dec 22, 2019 · Twitter Web App
View Tweet activity
837 Retweets 3.9K Likes

How will we learn?

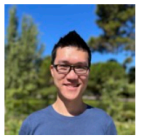
How will we learn?

- Lectures
- Sections
- Office hours
- Homework / Quizzes
- Midterm exam
- Final project
- Grading

How will we learn?

- **Lectures**
  - Monday, Wednesday, Friday **10:30-11:50am**
  - get familiar with R and RStudio
    - visualization
    - data manipulation
    - simulation
  - learn statistical methods
  - insights into ‘statistical thinking’

How will we learn?

Alvin Tan	Cynthia Wu	Justin Yang	Shashanka Subrahmanya	Verity Lua
				
TA	TA	TA	TA	TA
he/him	she/her	he/him	he/him	she/her
tanawm	cwu0701	justin.yang	ssbrahma	vyqlua

- **Sections**
  - section times are **Tuesday and Thursday 3:30-4:20pm**
  - **R tutorial** in week 1 ← recommended if you have little R experience
  - work on the homework assignments together
  - ask the TAs questions
  - attending sections is **optional** (but **recommended**)

How will we learn?

- **Office hours**

	Tobi Gerstenberg	Nilam Ram
		
Role	Instructor	Instructor
Pronouns	he/him	he/him
Email (@stanford.edu)	gerstenberg	nilamram
Office hours	Wednesday 1:30-2:30pm	Monday 1:30-2:30pm

Thursday

How will we learn?

- **Homework**
  - one assignment per week (6 in total)
  - you can work in groups
    - clearly indicate who you worked with on your submission
    - each group member should write their own code and written responses
  - download and submit via Canvas
  - you'll submit **two** files:
    - raw .Rmd file
    - rendered .pdf file
  - homework will be available **after class on Fridays**, and is due **Thursday 8pm** the week after

How will we learn?

- **Homework**
  - late submission policy:
    - you have 5 slip days in total
    - if you submit a homework within 24h of the deadline, this costs 1 slip day (or 2 slip days if it's submitted within 48h, ...)
    - once you have used up all your slip days, a late homework will receive a 0 score

How will we learn?

- **Quizzes**
  - Series of in-class quizzes
    - reinforce some fundamentals we think are important
    - reward in-person/human engagement
    - missed quizzes receive a 0 score
    - average of all quizzes = one homework

How will we learn?

- **Midterm exam**
  - like a homework assignment, but:
    - you have to work on it **on your own**
    - will be made available on Friday (2/6) and is due on **Friday (2/13) at 8pm**
      - no class on Wednesday that week so you have more time

How will we learn?

- **Final project**
  - you can work in groups of up to 3 members
  - everyone in the group will receive the same grade
  - the expectations for the project scale with the size of the group
  - you can find examples on canvas under Files > final_project

How will we learn?

- **Final project**
  - structure:
    1. short written project proposal
    2. oral presentation
    3. written report

How will we learn?

- **Final project**

  1. short written project proposal

     - due Thursday 2/19
     - we'll provide an RMarkdown template
     - examples from prior years are on Canvas

How will we learn?

- **Final project**

  2. oral presentation

     - during exam week
     - short slideshow of the project
     - groups present together

How will we learn?

- **Final project**

  3. written report

     - due **Friday 3/20**
     - ~ 2000 words per group member
     - answer an interesting research question

  - demonstrate what you've learned in class:
     - data wrangling
     - visualization
     - statistical modeling
     - reporting
  - you'll be using github to publish your final project online
  - examples are on Canvas

How will we learn?

- **Grading**
  - Homework/Quizzes: 40%
  - Midterm: 20%
  - Final project: 40%
    - Proposal: 5%
    - Presentation: 10%
    - Report: 25%
  - Bonus:
    - Ed discussion: 2%

Tools we will use in class

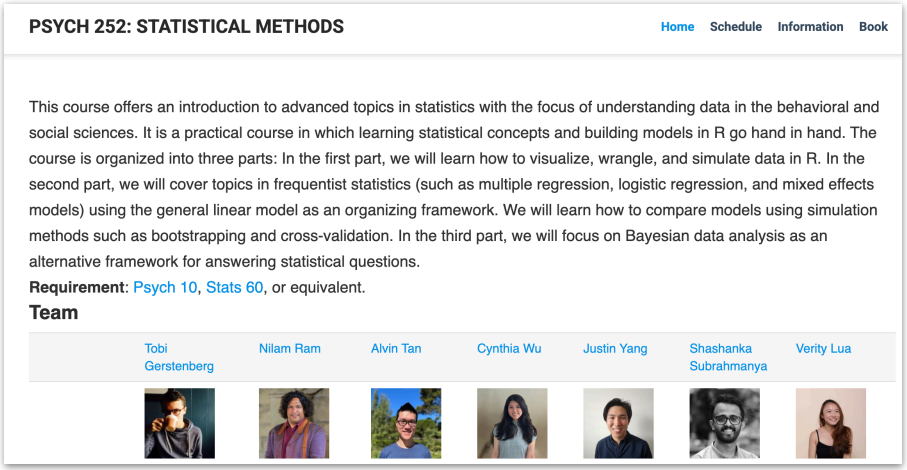
- Canvas
- Course website
- PollEverywhere
- Datacamp
- Ed Discussion
- Slack
- Github
- Free online books
- chatGPT

Tools we will use in class

- **Canvas**
  - **We** will:
    - send announcements
    - upload:
      - slides
      - class notes / code files
      - homework assignments
  - **You** will:
    - read announcements :)
    - upload homework

Tools we will use in class

- **Course website**
  - general information
  - schedule with links to materials
  - links to relevant datacamp classes



https://psych252.github.io/

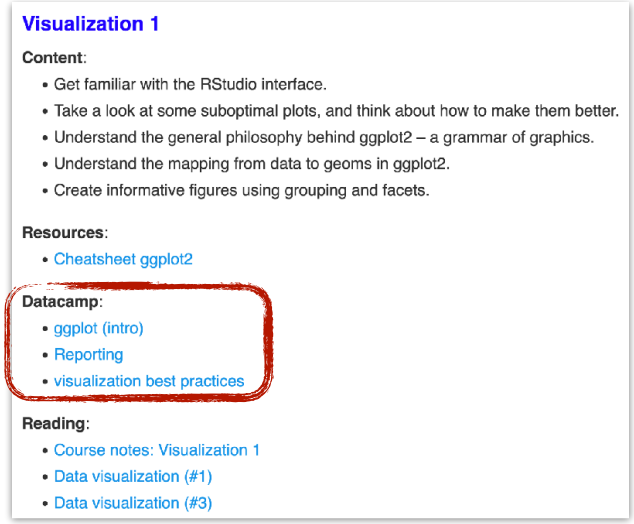
Tools we will use in class

- **PollEverywhere**
  - quick polls in class
  - feedback at the end of class
  - address: **pollev.com/psych252**
  - address: **pollev.com/psyc252**

How are you feeling today?

Tools we will use in class

- **DataCamp**
  - use your **stanford.edu** address to sign up!
  - if you haven't already, sign up here: https://tinyurl.com/psych252datacamp25



Visualization 1

Content:

- Get familiar with the RStudio interface.
- Take a look at some suboptimal plots, and think about how to make them better.
- Understand the general philosophy behind ggplot2 – a grammar of graphics.
- Understand the mapping from data to geoms in ggplot2.
- Create informative figures using grouping and facets.

Resources:

- Cheatsheet ggplot2

Datacamp:

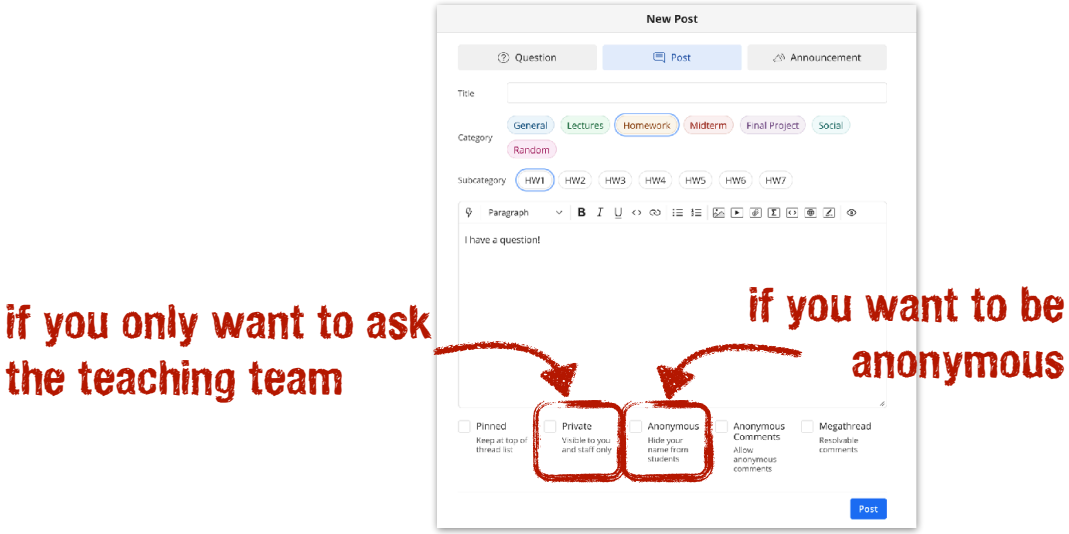
- ggplot (intro)
- Reporting
- visualization best practices

Reading:

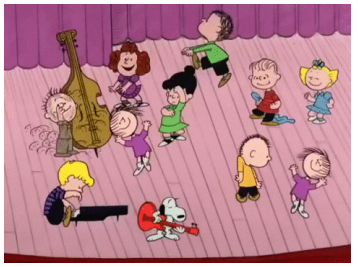
- Course notes: Visualization 1
- Data visualization (#1)
- Data visualization (#3)

Tools we will use in class

- **Ed Discussion**
  - post your own questions and answer those of your colleagues



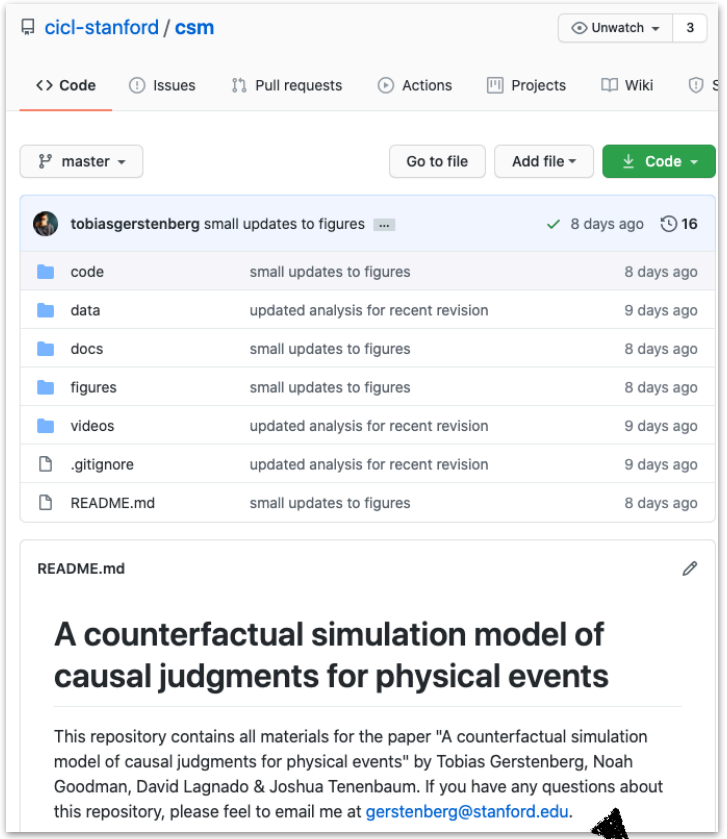
Tools we will use in class



- **Slack**
  - just a space for you to build community
  - the teaching team won't monitor what's happening on slack
  - you can join via the Slack tab on Canvas

Tools we will use in class

- **Github**
  - we will use github for the final project
  - it's an essential collaboration tool and critical for contributing to open science



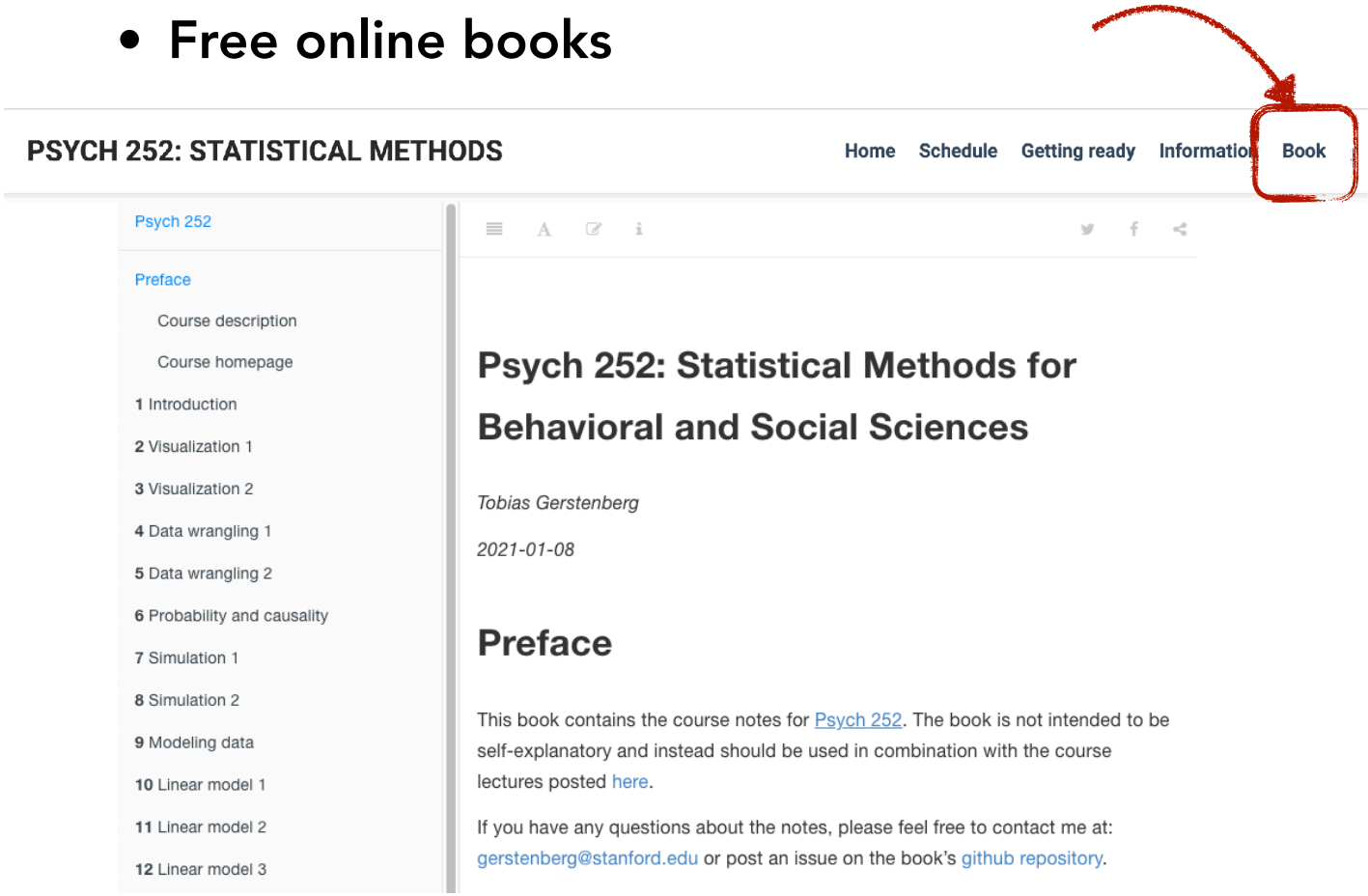
you want one of these for each paper you publish!

Tools we will use in class

- **Free online books**
  - we won't use a text book in class
  - statistics and data science are developing fast and we didn't find a single book that fits the bill
  - **but**: many great, free books online
  - **and**: we will point out suggested readings as we go along
  - **also**: we'll update our course notes as we go along

Tools we will use in class

- **Free online books**



https://github.com/psych252/psych252book

Tools we will use in class

- **GPT, Gemini, Claude, etc.**
  - coding is thinking
  - there is a lot of value in coding yourself (particularly when you're just learning a new language)
  - if you don't know how to code, it will be difficult for you to check generated code
  - helpful tool once you've mastered the language (just like autocomplete on your phone ...)

Ed discussion

I have a question about the homework.

Canvas

I would like to submit my homework on time.

Slack

I want to connect with others in my class.

Course website

I don't remember the schedule.

Datacamp

I would like to learn more about a topic.

with help from

Section & Office hours

I'd like to discuss something in person.

PollEverywhere

I have an idea how to make this better!

Anonymous feedback form

Something bothers me ...

Some **general thoughts**

Stats has a very troubled past

JOURNAL OF STATISTICS AND DATA SCIENCE EDUCATION
2024, VOL. 32, NO. 1, 108–119
https://doi.org/10.1080/26939169.2023.2224407

Teaching the Difficult Past of Statistics to Improve the Future

Lee Kennedy-Shaffer

Department of Mathematics and Statistics, Vassar College, Poughkeepsie, NY

ABSTRACT

In recent years, the discipline of statistics has begun reckoning with its difficult history. Institutions are reconsidering names that have honored key historical figures in statistics who have deep ties to eugenics movements and racial and class prejudice. These names, however, continue to appear in our classrooms, where we teach the methods created by these individuals, raising the question of how instructors should address their legacies. Three examples of famous statisticians and their work—Francis Galton's use of conditional probabilities to demonstrate "hereditary talent," Karl Pearson's attempt to quantify the intelligence of Jewish immigrant students, and Ronald A. Fisher's creation of the analysis of variance to de-emphasize environment in human development—highlight the intimate ties between statistics and eugenics. These examples, along with a discussion of the context of these men, eugenics movements, and the statisticians and scientists who opposed their eugenic programs, can humanize the field for students, teach them about the challenges in accurate and unbiased data collection and analysis, and connect historical mistakes to contemporary ethical issues. Confronting this history in the classroom can both improve the teaching of the statistical methodologies themselves and begin a broader conversation about the role of statistics in the world. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2023
Accepted May 2023

KEYWORDS
Correlation; Ethics; Eugenics; Genetics; History of statistics; Interaction

you can find the article under Files > papers

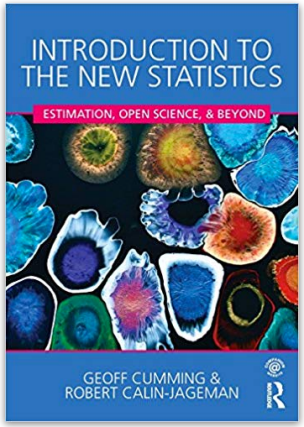
Vision for this class

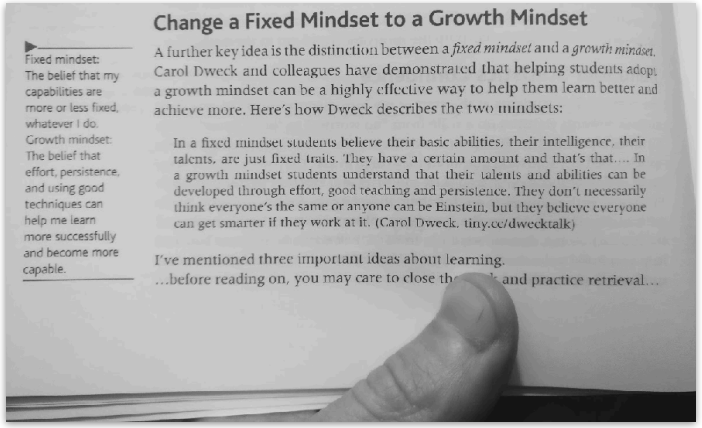
In "A Vision for Stanford", it says that Stanford wants to be

"an inspired, inclusive and collaborative community of diverse scholars, students and staff, where all are supported and empowered to thrive."

Let's try our best together in this class to make this happen!

Fear of statistics





Tobi's thumb

Carol Dweck

Try to adopt a growth mindset!

fixed mindset:
students believe their basic abilities, their intelligence, their talents, are just fixed traits.

growth mindset:
students *understand* that their talents and abilities can be developed through effort, good teaching and persistence

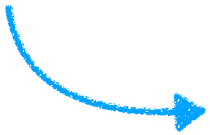
Stats is important but ...

- it's not the only thing that matters for being a good scientist
- generating research ideas matters
- writing matters
- presenting matters
- ...
- (nobody is very good at all of these things)

Feedback

Teaching philosophy

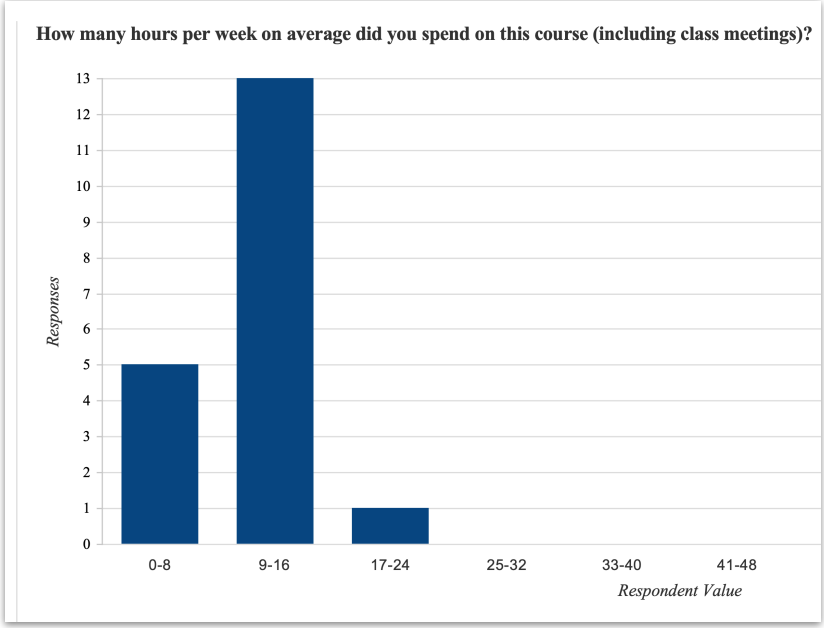
"The one who does the work does the learning." (Doyle, 2008)

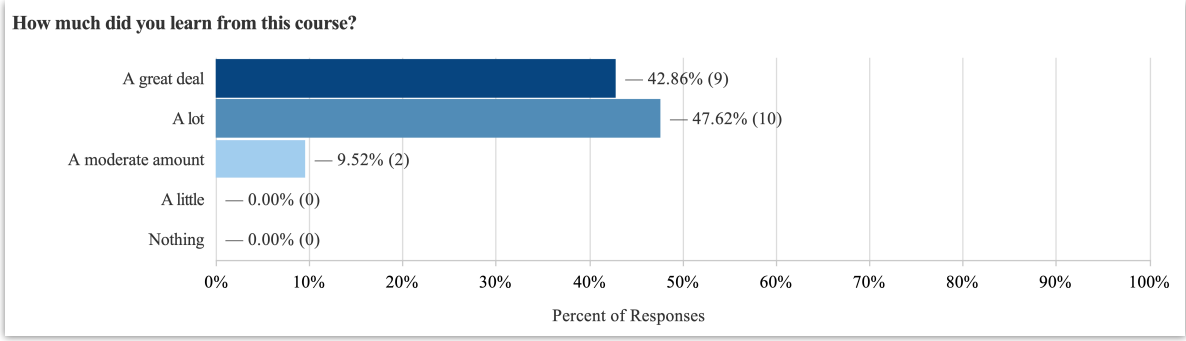


we will try our best to make you do the work!

Doyle, T. (2008). Helping students learn in a learner-centered environment: A guide to facilitating learning in higher education. Stylus Publishing, LLC.

Doing the work takes time ...





... but it's worth it

What would you like to say about this course to a student who is considering taking it in the future?

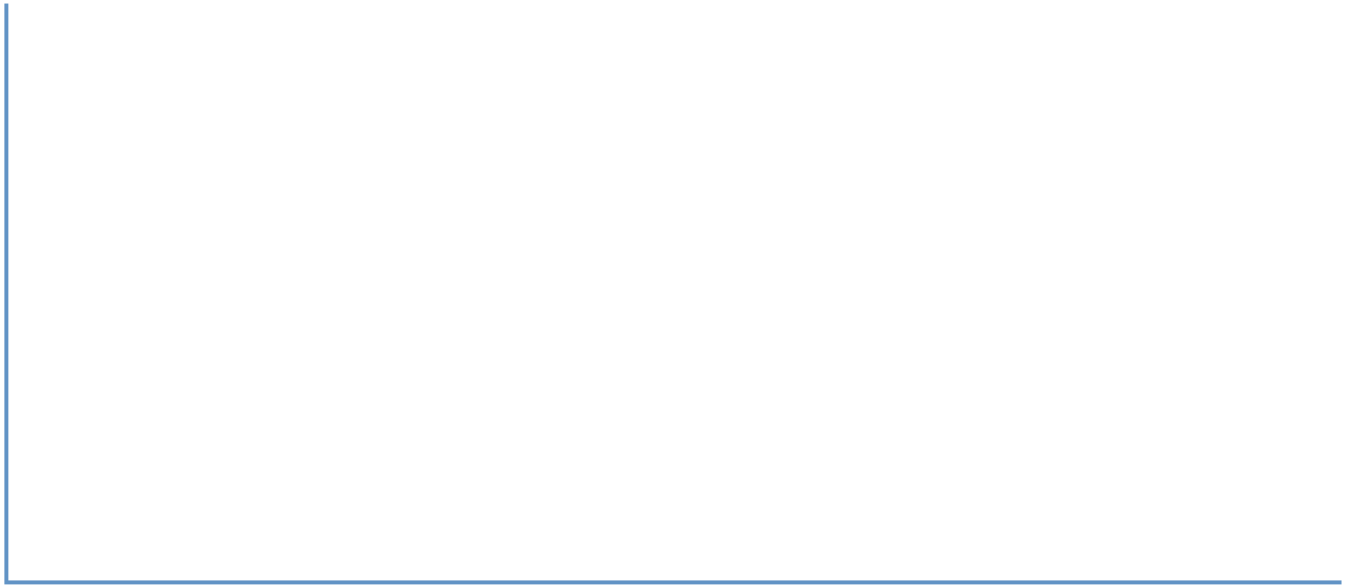
- "It will be a lot of work, but it will teach you a lot!"
- "I would absolutely recommend taking this class. I came into this class with random/haphazard stats knowledge (ex. STATS60) and no coding experience, and I left as pretty much a convert to R and model comparison approaches."
- "It can sometimes feel like a ton of work but I advise that **you use office hours at least once**. Even if you think you don't need it. It will help you during the midterm (I struggled a bit with this) and the problem sets."
- "This is a great course to get more comfortable running your own stats and to secure proficiency in R. There is a ton of course material, though, and sometimes that felt really overwhelming. It was **hard to balance actually understanding the stats with bringing my coding skills up** to where they need to be to do this course."
- "Don't take it unless you are an R wizard before the start of the class. It is impossible to get the R skills in 2 weeks. And the class gets REALLY high level just before the midterm (and it just gets worse...)."

Feedback

Help us help you!

- ask questions in class and stick around after class
- come to sections
- post questions on Ed discussion (you can post anonymously)
- send us an email (but use Ed discussion for questions about homework or class!)
- use anonymous feedback form (link at the bottom of the course website)

How was the pace of today's class?



How happy were you with today's class overall?

What did you like about today's class? What could be improved next time?

Thank you to ...

Alexandra Chouldechova
Allison Horst
Andrew Heiss
Ben Baumer
Benoit Monin
Bodo Winter
David Lagnado
Ewart Thomas
Henrik Singmann
Julian Jara-Ettinger
Justin Gardner
Kevin Smith
Lisa DeBruine
Maarten Speekenbrink
Matthew Kay
Matthew Salganik
Mika Braginsky
Mike Frank
Mine Çetinkaya-Rundel
Nick C. Huntington-Klein
Kevin Grimm
Patrick Mair
Paul-Christian Bürkner
Peter Cushner Mohanty
Richard McElreath
Russ Poldrack
Stephen Dewitt
Solomon Kurz
Tom Hardwicke
Tristan Mahr

Thanks!

see you on **Wednesday**

make sure to have
R and **RStudio** up to date

post on Ed discussion if you experience any problems