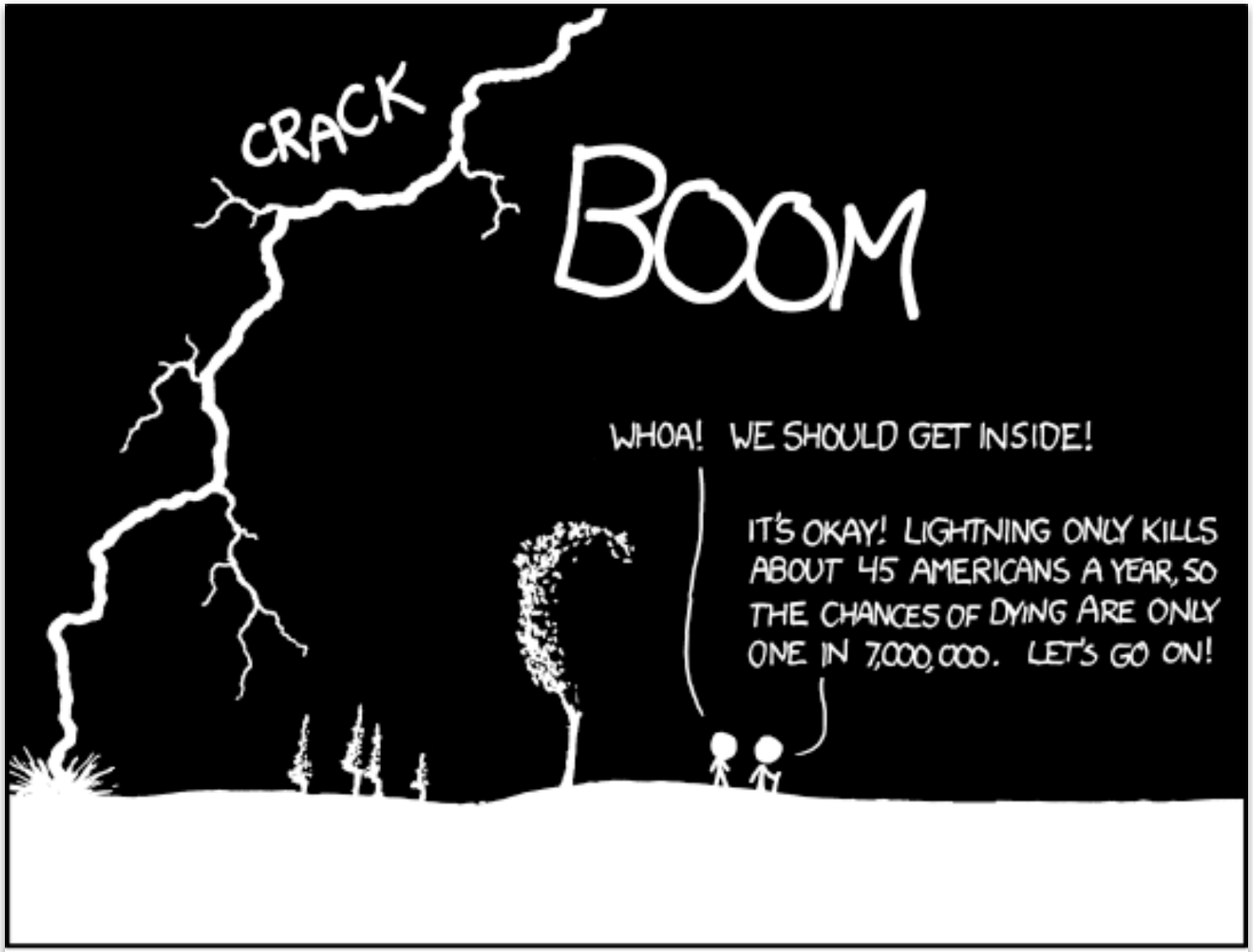
PSYCH 252: Statistical Methods for Behavioral and Social Sciences

Introduction



xkcd

01/06/2020

Outline

- Introduction
- **What** will we learn?
- **How** will we learn?
- Some general thoughts
- Visualization
- Feedback

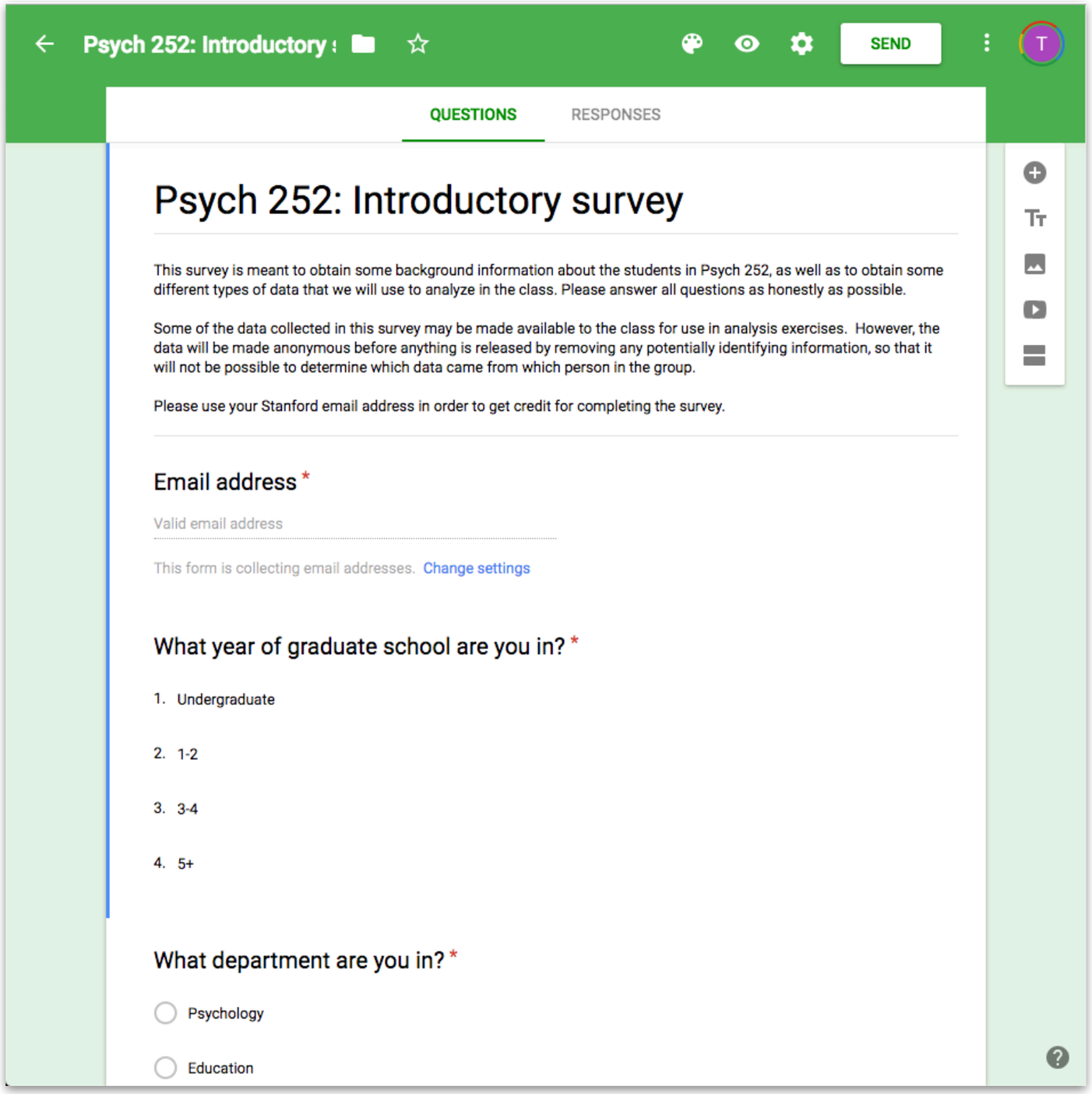
Introduction

Introduction

1. What did you do during the winter break?

2. What are you hoping to get out of this course?

Introduction



(announcement soon via Canvas)

Tobias Gerstenberg (he/him/his)

Role:	Instructor
Email:	gerstenberg@stanford.edu
Office:	420-302
Office hours:	Monday 4:00-5:00pm

- I lead the

  (http://cicl.stanford.edu/).

- The lab studies the role of causality in our understanding of the world, and of each other.

1. How does the mind learn to represent the causal structure of the world?
2. What is the relationship between causal thinking and counterfactual simulation?
3. How do we hold others responsible for the outcomes of their actions?

We formalize people's mental models as **computational models**.

We use a combination of large-scale **online experiments**, **interactive experiments** in the lab, and **eye-tracking experiments**.

@tobigerstenberg

Tyler Bonnen (he/him/his)

- My current work centers on the relationship between perception and memory

- I use machine learning + non-parametric statistical tools to build and evaluate biologically inspired computational models

- I'm co-advised by Anthony Wagner and Dan Yamins

Tyler Bonnen

Role:	Teaching assistant
Email:	bonnen@stanford.edu
Office:	420-406
Office hours:	Wednesday 3:00-4:00pm

Andrew Nam (he/him/his)

- 2nd year PhD working with Jay McClelland. Interested in abstract/symbolic reasoning, meta-learning, and mathematical cognition.

- I use neural networks and empirical human data to understand both human cognition and AI systems.

- Previously software engineer at Salesforce, BA in economics and computer science at UC Berkeley.

Andrew Nam

Role:	Teaching assistant
Email:	ajhnam@stanford.edu
Office:	420-326
Office hours:	Friday 12:30-1:30pm

Jinxiao Zhang (he/him/his)

- 2nd year PhD student working with Dr. James Gross

- I am interested generally in how the 'emotional' system and 'cognitive' system interact in the brain.

- In particular, I study how one use cognitive techniques to regulate emotions and how emotion influences memory.
  In addition, I study the involvement of sleep in those processes.

Jinxiao Zhang

Role:	Teaching assistant
Email:	jinx.zhang@stanford.edu
Office:	420-425
Office hours:	Tuesday 3:00-4:00pm

What will we learn?

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

3. Report results

all the time (& Week 10)
- Reproducible research

Learning goals

You will learn how to **use R** to …

- read, wrangle, simulate and analyze data
- make publication-ready plots

Use R!

Why  ?

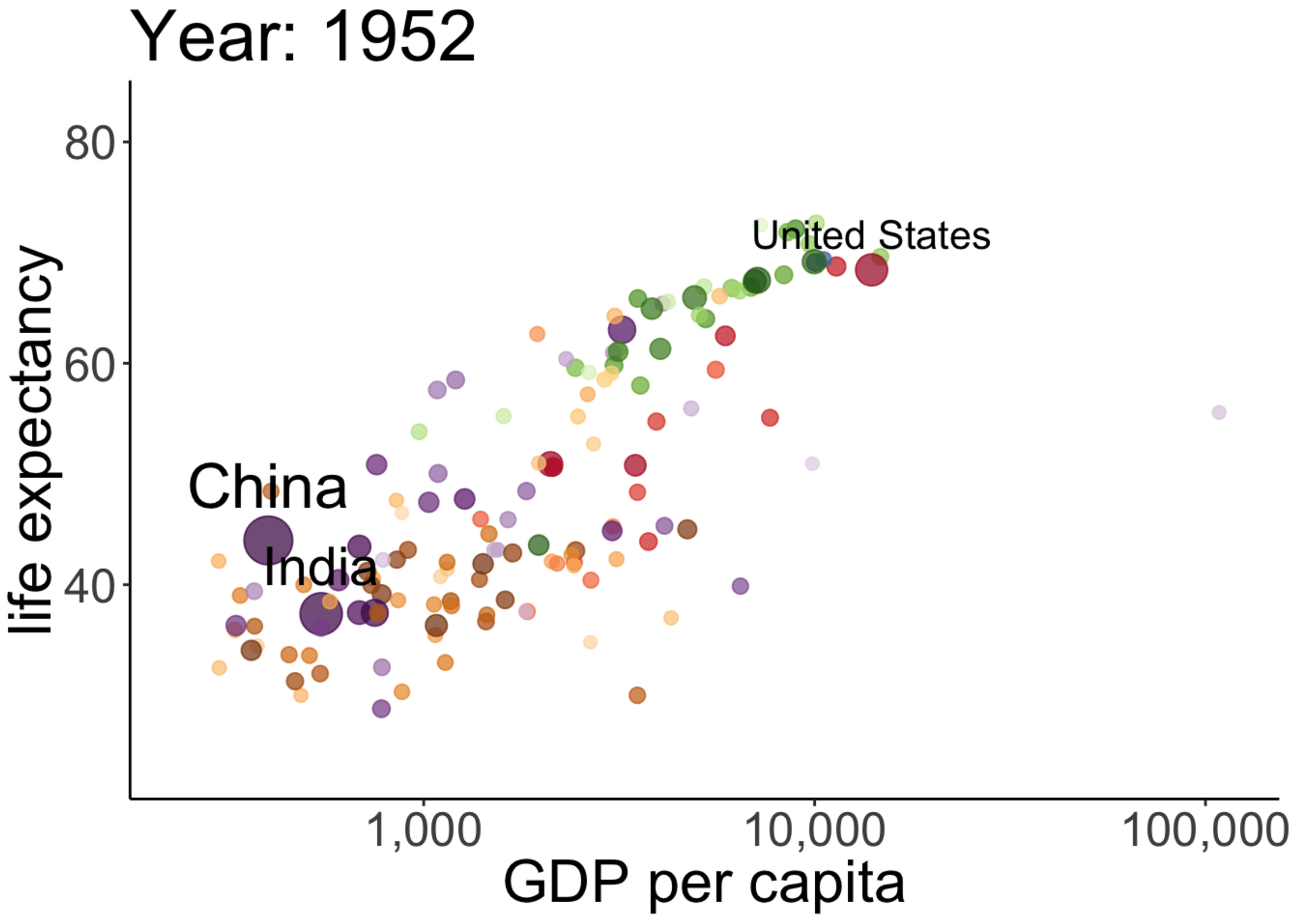
Visualization

Data manipulation

Statistical modeling

Communicating results

Why R ? **Visualization**



Why R ? 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 ? Data manipulation

Data Transformation with dplyr : : CHEAT SHEET

dplyr functions work with pipes and expect **tidy data**. In tidy data:

Each **variable** is in its own **column**

&

Each **observation**, or **case**, is in its own **row**

pipes

x %>% f(y) becomes **f(x, y)**

Summarise Cases

These apply **summary functions** to columns to create a new table of summary statistics. Summary functions take vectors as input and return one value (see back).

summary function

summarise(.data, …)
Compute table of summaries.
summarise(mtcars, avg = mean(mpg))

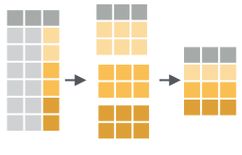
count(x, ..., wt = NULL, sort = FALSE)
Count number of rows in each group defined by the variables in … Also **tally**().
count(iris, Species)

VARIATIONS

summarise_all() - Apply funs to every column.
summarise_at() - Apply funs to specific columns.
summarise_if() - Apply funs to all cols of one type.

Group Cases

Use **group_by()** to create a "grouped" copy of a table. dplyr functions will manipulate each "group" separately and then combine the results.



```
mtcars %>%
group_by(cyl) %>%
summarise(avg = mean(mpg))
```

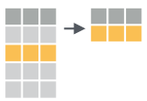
group_by(.data, ..., add = FALSE**)**
Returns copy of table grouped by …
g_iris <- group_by(iris, Species)

ungroup(x, …**)**
Returns ungrouped copy of table.
ungroup(g_iris)

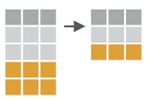
Manipulate Cases

EXTRACT CASES

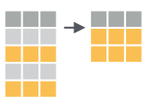
Row functions return a subset of rows as a new table.



filter(.data, …**)** Extract rows that meet logical criteria. *filter(iris, Sepal.Length > 7)*

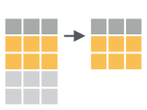


distinct(.data, ..., .keep_all = FALSE**)** Remove rows with duplicate values.
distinct(iris, Species)



sample_frac(tbl, size = 1, replace = FALSE, weight = NULL, .env = parent.frame()**)** Randomly select fraction of rows.
sample_frac(iris, 0.5, replace = TRUE)

sample_n(tbl, size, replace = FALSE, weight = NULL, .env = parent.frame()**)** Randomly select size rows. *sample_n(iris, 10, replace = TRUE)*



slice(.data, …**)** Select rows by position.
slice(iris, 10:15)

top_n(x, n, wt**)** Select and order top n entries (by group if grouped data). *top_n(iris, 5, Sepal.Width)*

Logical and boolean operators to use with filter()

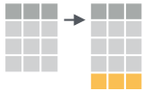
<	<=	is.na()	%in%	|	xor()
>	>=	!is.na()	!	&	

See **?base::logic** and **?Comparison** for help.

ARRANGE CASES

arrange(.data, …**)** Order rows by values of a column or columns (low to high), use with **desc()** to order from high to low.
arrange(mtcars, mpg)
arrange(mtcars, desc(mpg))

ADD CASES



add_row(.data, ..., .before = NULL, .after = NULL**)**
Add one or more rows to a table.
add_row(faithful, eruptions = 1, waiting = 1)

Manipulate Variables

EXTRACT VARIABLES

Column functions return a set of columns as a new vector or table.

pull(.data, var = -1**)** Extract column values as a vector. Choose by name or index.
pull(iris, Sepal.Length)

select(.data, …**)**
Extract columns as a table. Also **select_if()**.
select(iris, Sepal.Length, Species)

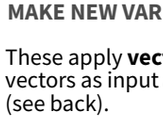
Use these helpers with select (),
e.g. select(iris, starts_with("Sepal"))

contains(match**)**
ends_with(match**)**
matches(match**)**
num_range(prefix, range**)**
one_of(…**)**
starts_with(match**)**
:, e.g. mpg:cyl
-, e.g, -Species

MAKE NEW VARIABLES

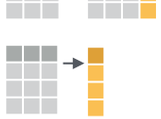
These apply **vectorized functions** to columns. Vectorized funs take vectors as input and return vectors of the same length as output (see back).

vectorized function

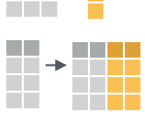


mutate(.data, …**)**
Compute new column(s).
mutate(mtcars, gpm = 1/mpg)

transmute(.data, …**)**
Compute new column(s), drop others.
transmute(mtcars, gpm = 1/mpg)

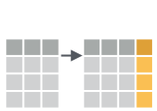


mutate_all(.tbl, .funs, …**)** Apply funs to every column. Use with **funs()**. Also **mutate_if()**.
mutate_all(faithful, funs(log(.), log2(.)))
mutate_if(iris, is.numeric, funs(log(.)))



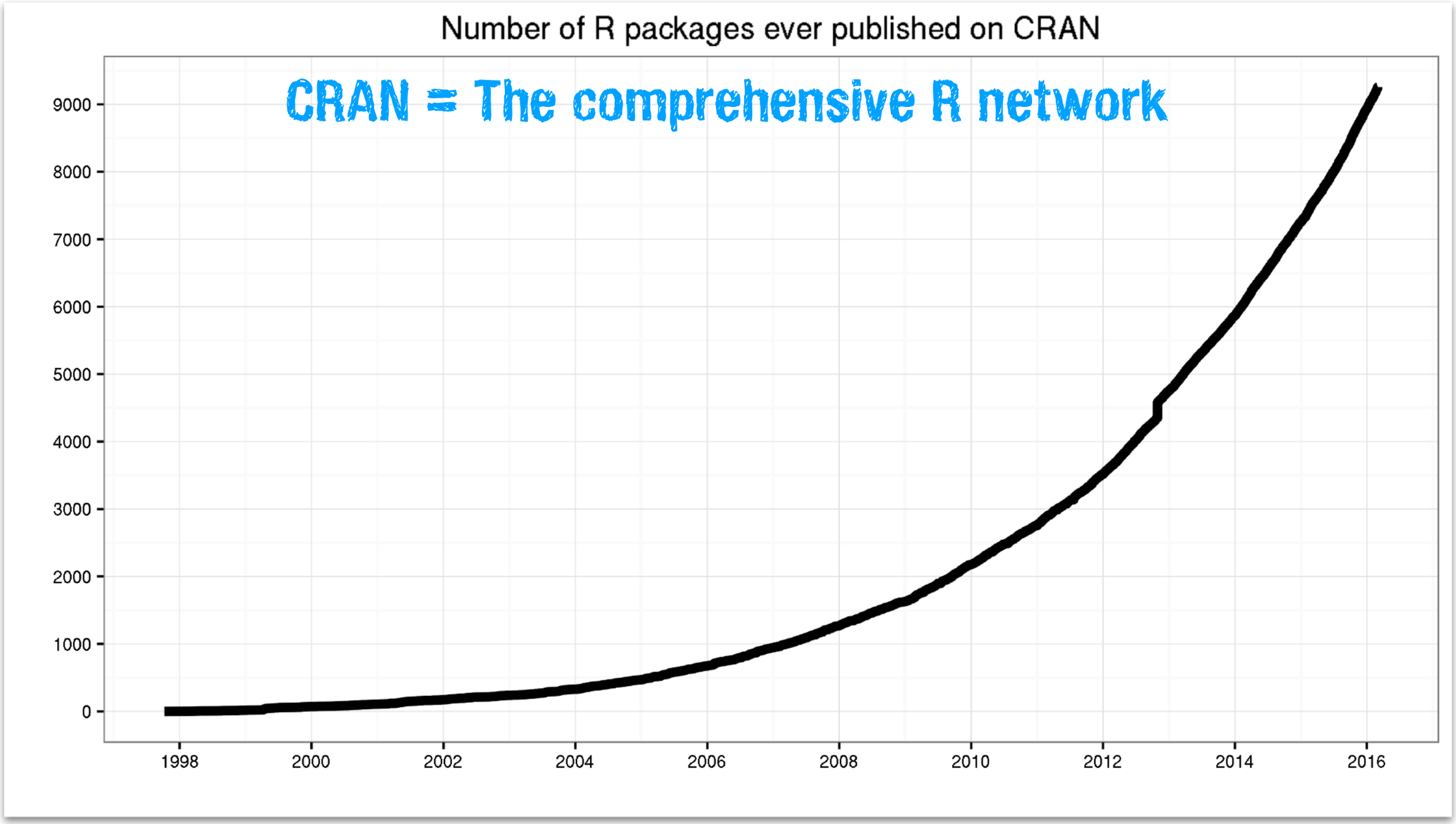
mutate_at(.tbl, .cols, .funs, …**)** Apply funs to specific columns. Use with **funs()**, **vars()** and the helper functions for select().
mutate_at(iris, vars(-Species), funs(log(.)))

add_column(.data, ..., .before = NULL, .after = NULL**)** Add new column(s). Also **add_count()**, **add_tally()**. *add_column(mtcars, new = 1:32)*



rename(.data, …**)** Rename columns.
rename(iris, Length = Sepal.Length)

Why R ? Statistical modeling



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

Why R ? Communicating results

01-intro.Rmd

Knit

```
# 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)
```

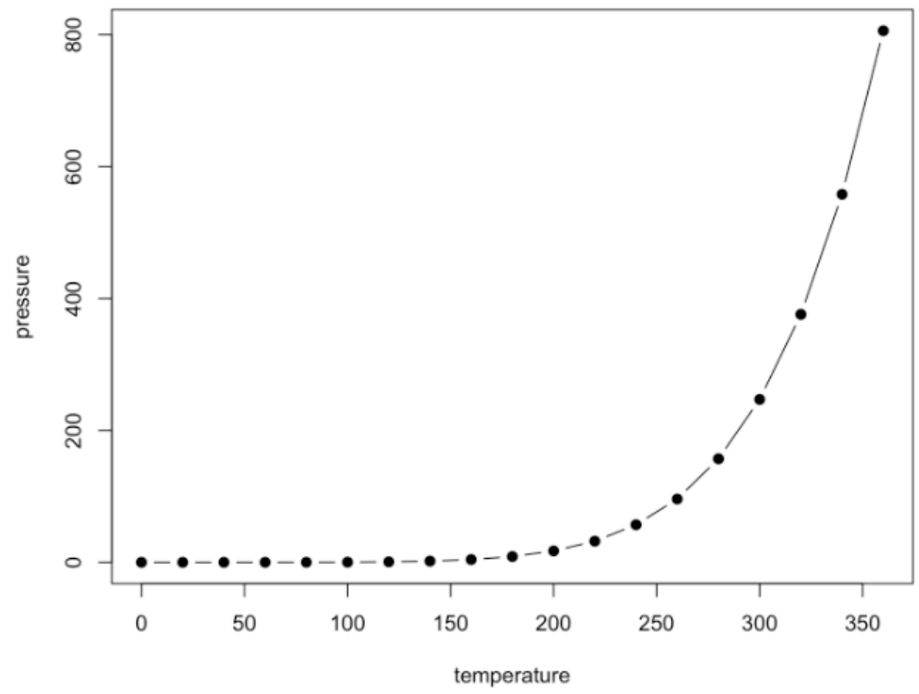
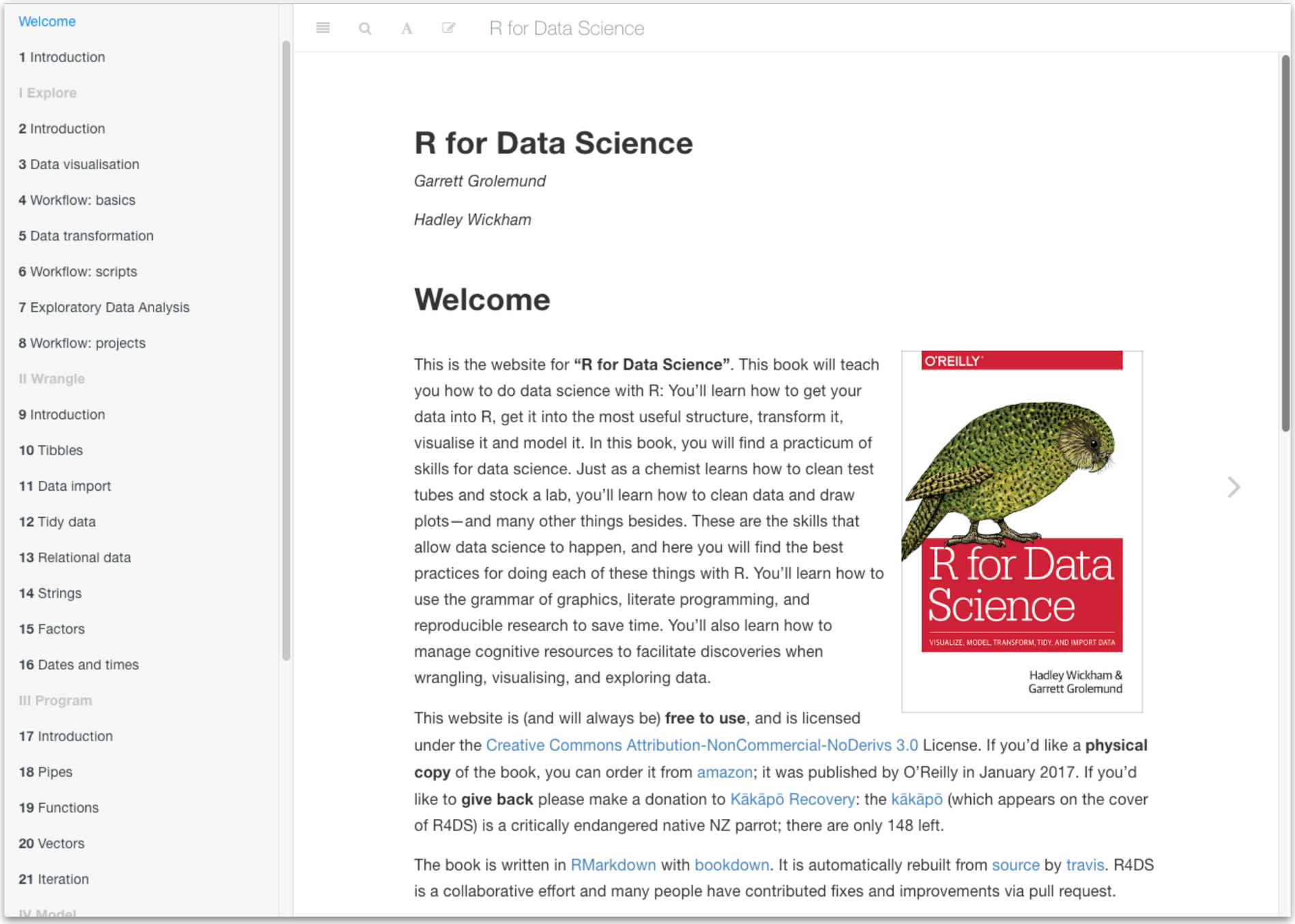


Figure 1.1: Here is a nice figure!

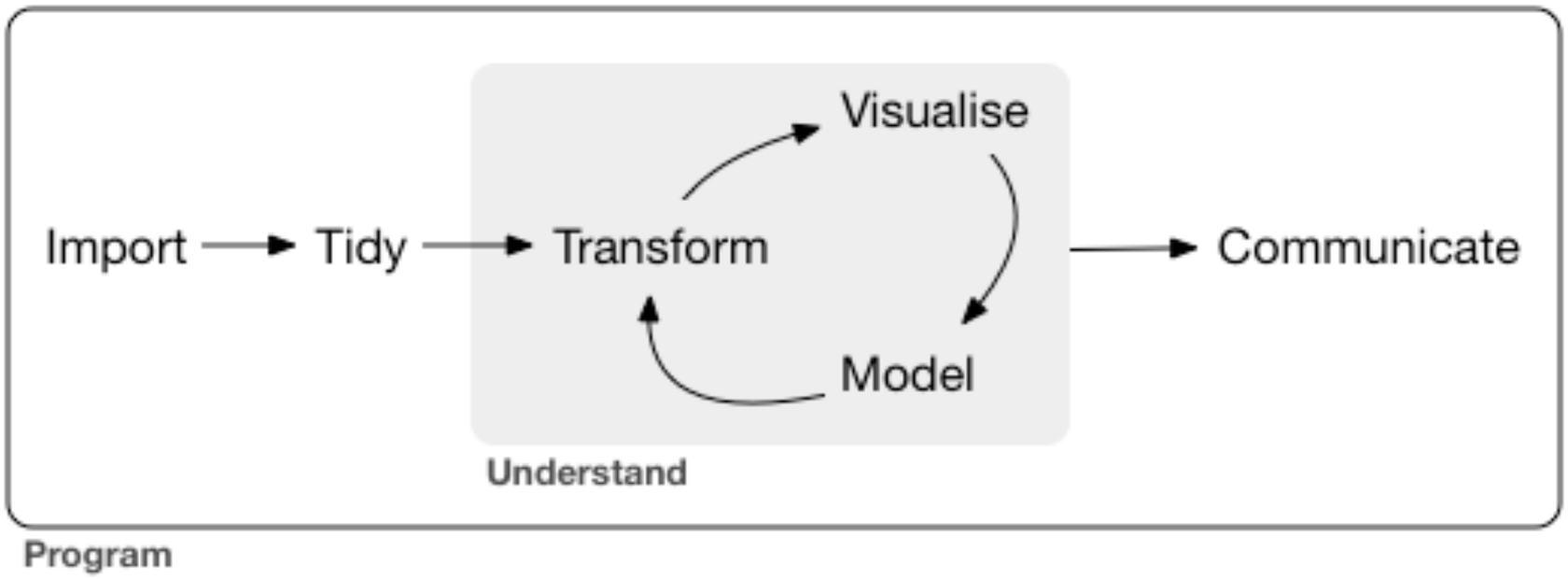
Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 1.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 1.1.

Why R ? Communicating results



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

Why  ? **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 of us here at Stanford use it.
- R has a nice online community. **#rstats**
- RStudio is a great IDE (Integrated Development Environment)!
- We use R to better understand statistics

Why  ? **Also ...**

instead of

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

(google search "complicated statistical equation")

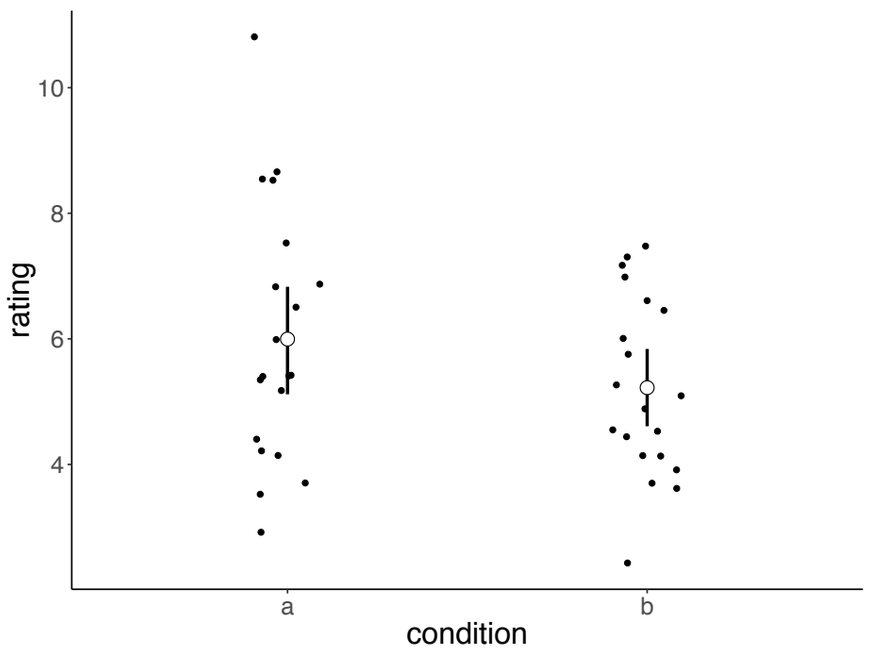
d_f	Level of Significance for Directional Test (t_{crit})					
	∝=0.10	∝=0.05	∝=0.025	∝=0.01	∝=0.005	∝=0.0005
	Level of Significance for Non-directional Test (t_{crit})					
	∝=0.20	∝=0.10	∝=0.05	∝=0.02	∝=0.01	∝=0.001
1	3.0780	6.3140	12.7100	31.8200	63.6600	636.6000
2	1.8860	2.9200	4.3030	6.9650	9.9250	31.6000
3	1.6380	2.3530	3.1820	4.5410	5.8410	12.9200
4	1.5330	2.1320	2.7760	3.7470	4.6040	8.6100
5	1.4760	2.0150	2.5710	3.3650	4.0320	6.8690
6	1.4400	1.9430	2.4470	3.1430	3.7070	5.9590
7	1.4150	1.8950	2.3650	2.9980	3.4990	5.4080
8	1.3970	1.8600	2.3060	2.8960	3.3550	5.0410
9	1.3830	1.8330	2.2620	2.8210	3.2500	4.7810
10	1.3720	1.8120	2.2280	2.7640	3.1690	4.5870
11	1.3630	1.7960	2.2010	2.7180	3.1060	4.4370
12	1.3560	1.7820	2.1790	2.6810	3.0550	4.3180
13	1.3500	1.7710	2.1600	2.6500	3.0120	4.2210
14	1.3450	1.7610	2.1450	2.6240	2.9770	4.1400
15	1.3410	1.7530	2.1310	2.6020	2.9470	4.0730

we'll do

```r
set.seed(0)
df.data = tibble(
  a = rnorm(20, mean = 6, sd = 2),
  b = rnorm(20, mean = 5, sd = 2)
) %>%
  gather("condition", "rating")

df.data %>%
  group_by(condition) %>%
  summarize(rating.mean = mean(rating),
            rating.sd = sd(rating)) %>%
  kable()

# calculate the difference between conditions
difference.actual = df.data %>%
  group_by(condition) %>%
  summarize(rating.mean = mean(rating)) %>%
  pull(rating.mean) %>%
  diff() %>%
  -.
```



But what about python™ ?

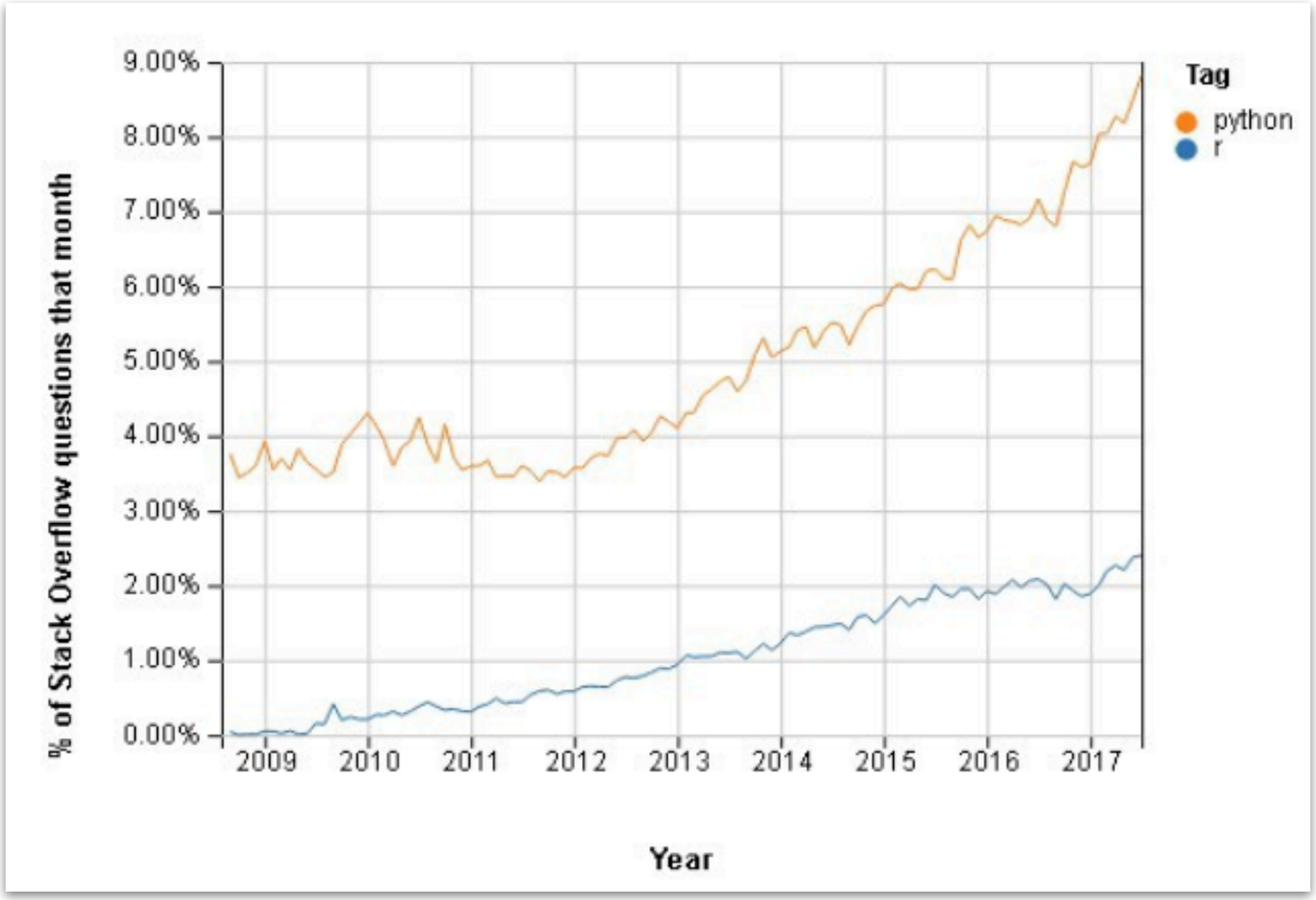
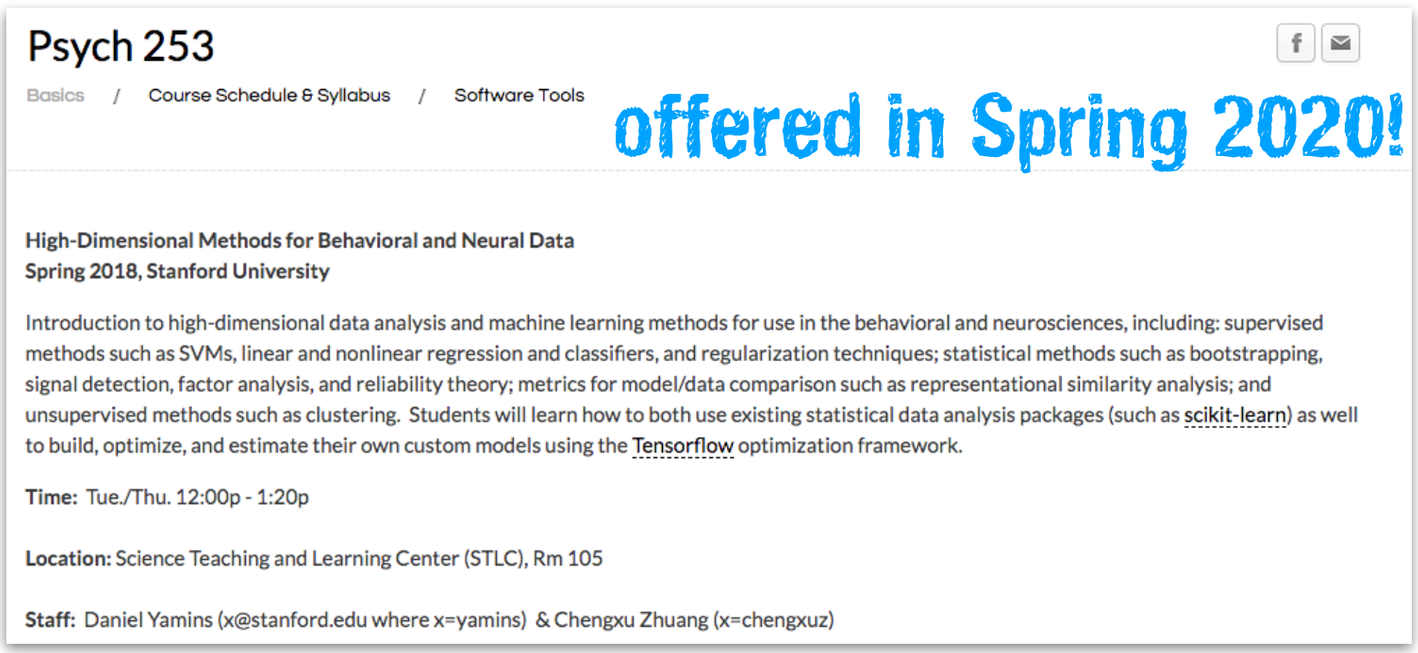


Image Source: https://dzone.com/articles/r-or-python-data-scientists-delight

But what about python™ ?

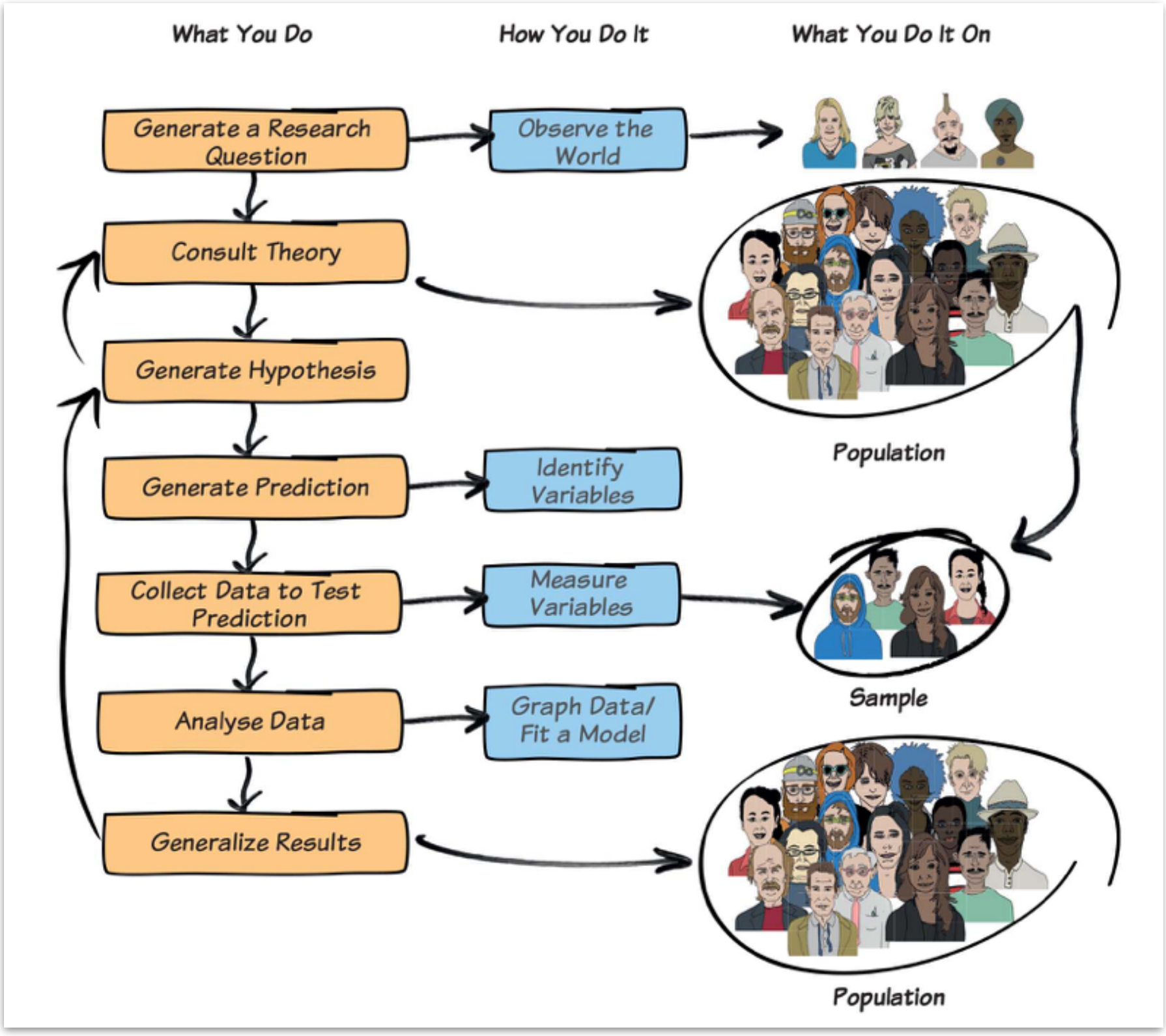
- You should learn python, too! :)
- Python is the more general purpose programming language.
- Python is great for machine learning (e.g. neural networks and deep learning).

- You can learn about all of these things in **Psych 253**!



Build models

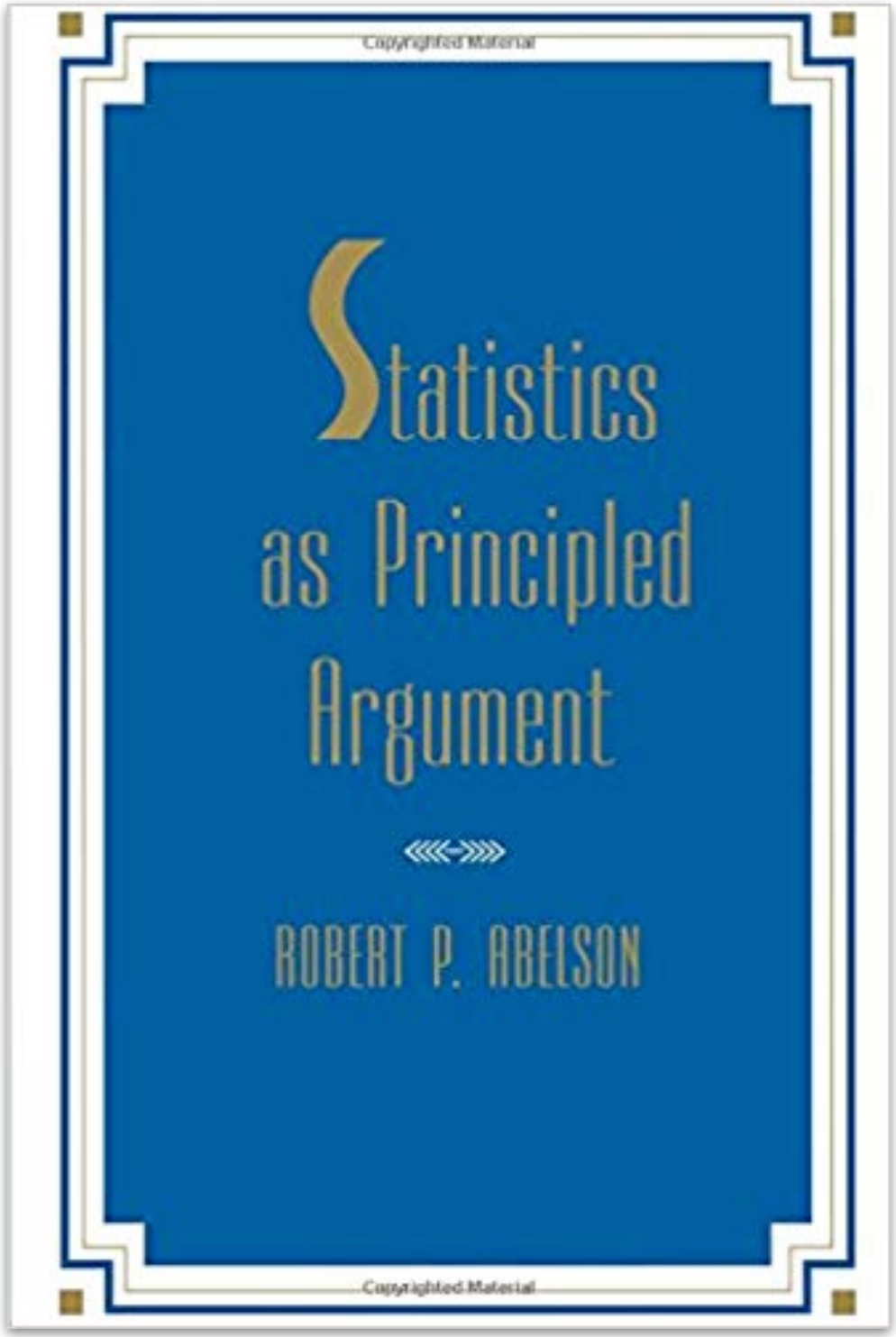
The research process



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, ...

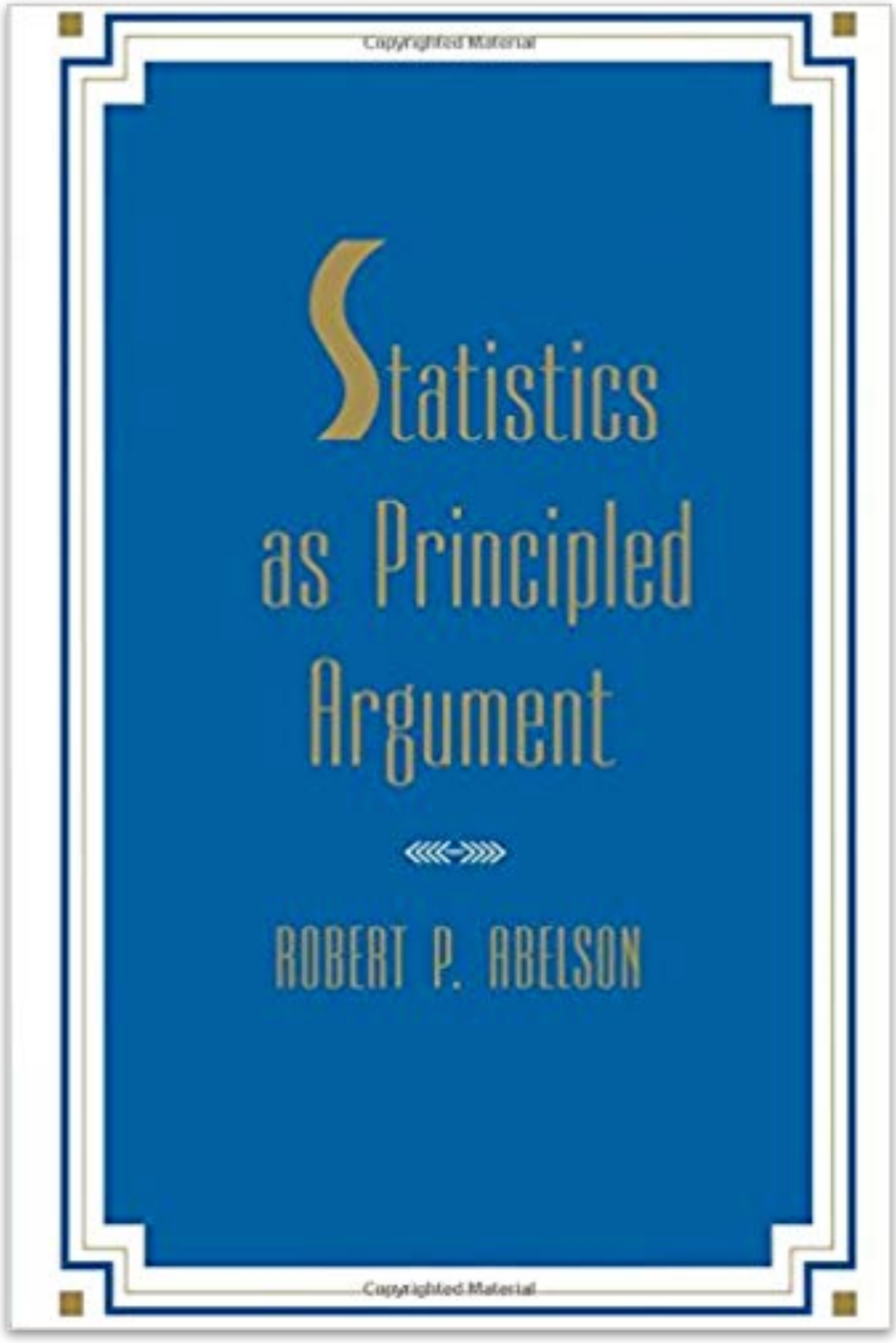


Abelson, R. P. (1995). *Statistics as principled argument*. New York: Psychology Press.

Statistics as argument

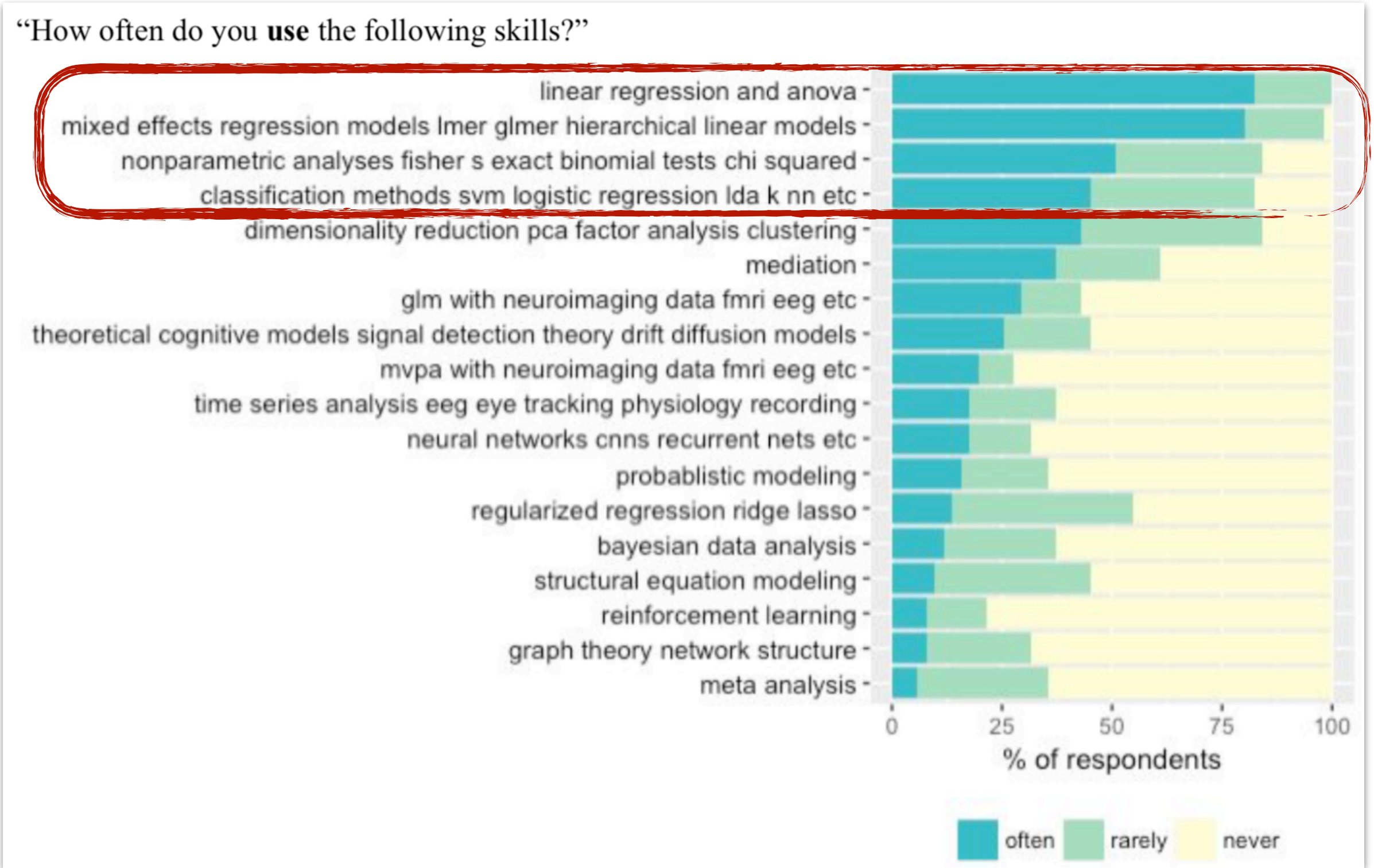
5 criteria make statistical arguments persuasive, or **MAGIC**:

- ***Magnitude***: Larger effects are more persuasive
- ***Articulation***: Understandability of results
- ***Generality***: How much does the effect depend on contextual factors
- ***Interestingness***: Change what people believe about an important issue
- ***Credibility***: Believability of a research claim, methodological soundness, and theoretical coherence

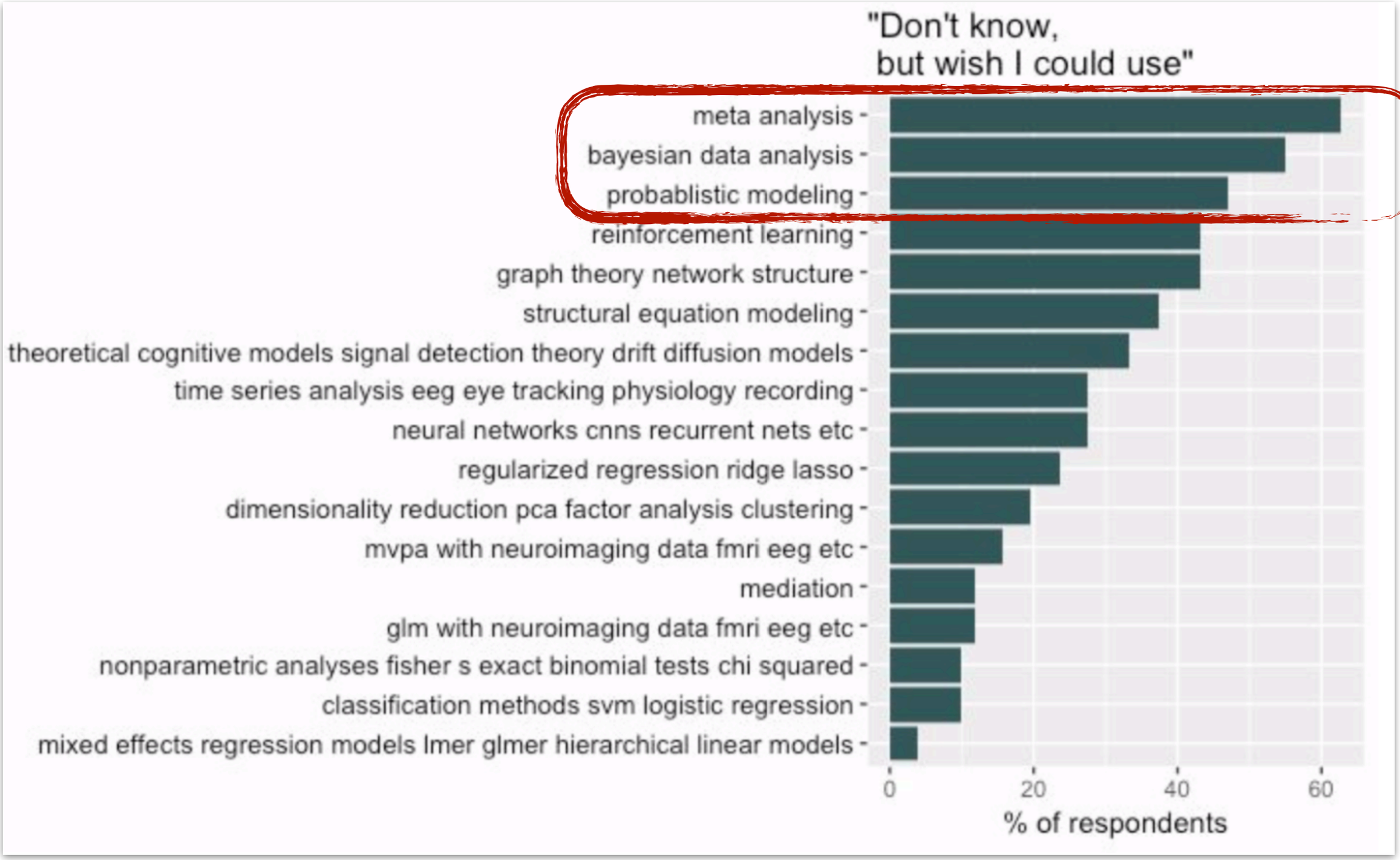


Abelson, R. P. (1995). *Statistics as principled argument*. New York: Psychology Press.

Psychology grad student survey

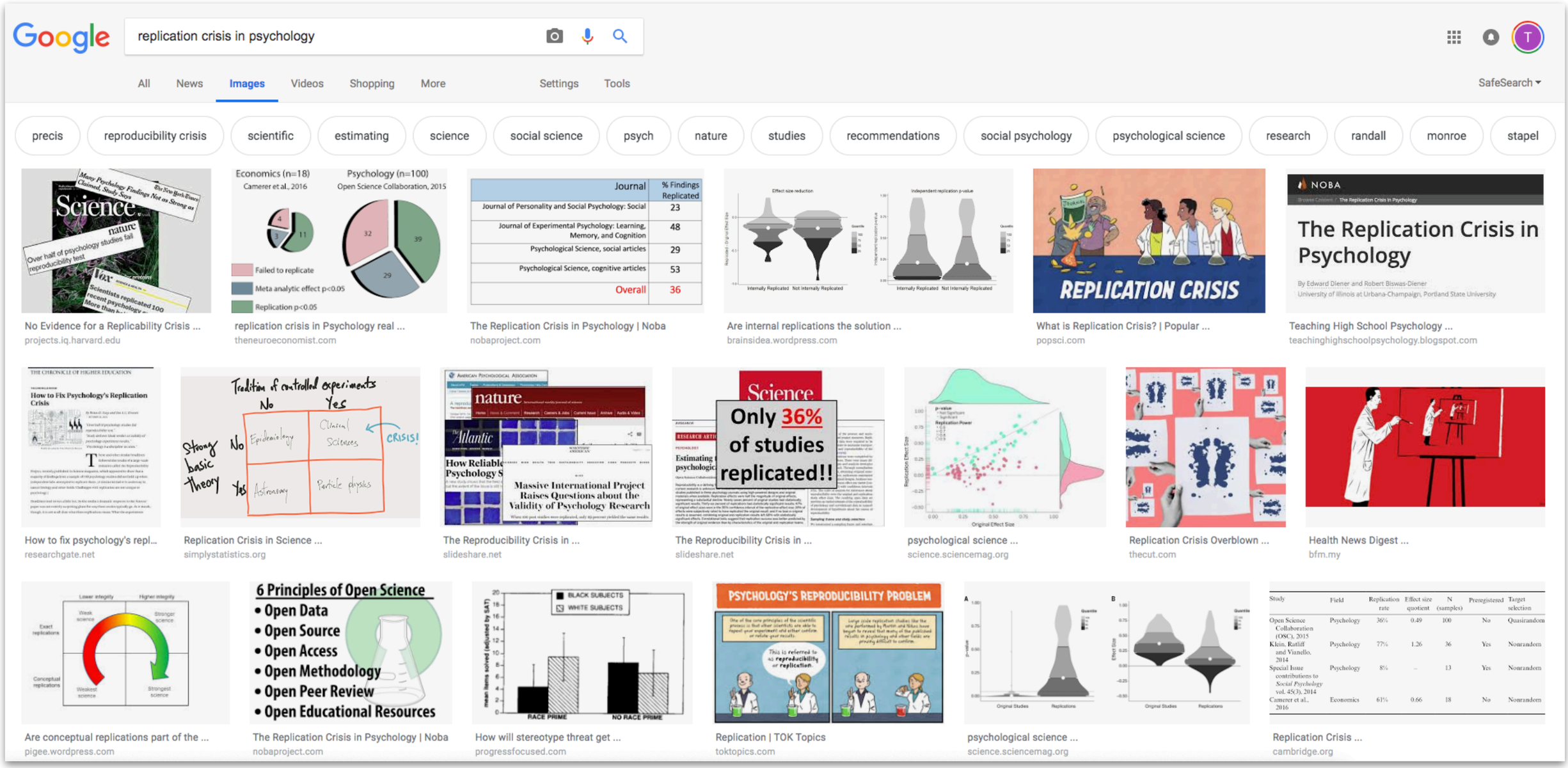


Psychology grad student survey



Report results

Replication crisis



No Evidence for a Replicability Crisis ...
projects.iq.harvard.edu

replication crisis in Psychology real ...
theneuroeconomist.com

The Replication Crisis in Psychology | Noba
nobaproject.com

Are internal replications the solution ...
brainsidea.wordpress.com

What is Replication Crisis? | Popular ...
popsci.com

Teaching High School Psychology ...
teachinghighschoolpsychology.blogspot.com

How to fix psychology's repl...
researchgate.net

Replication Crisis in Science ...
simplystatistics.org

The Reproducibility Crisis in ...
slideshare.net

The Reproducibility Crisis in ...
slideshare.net

psychological science ...
science.sciencemag.org

Replication Crisis Overblown ...
thecut.com

Health News Digest ...
bfm.my

Are conceptual replications part of the ...
pigee.wordpress.com

The Replication Crisis in Psychology | Noba
nobaproject.com

How will stereotype threat get ...
progressfocused.com

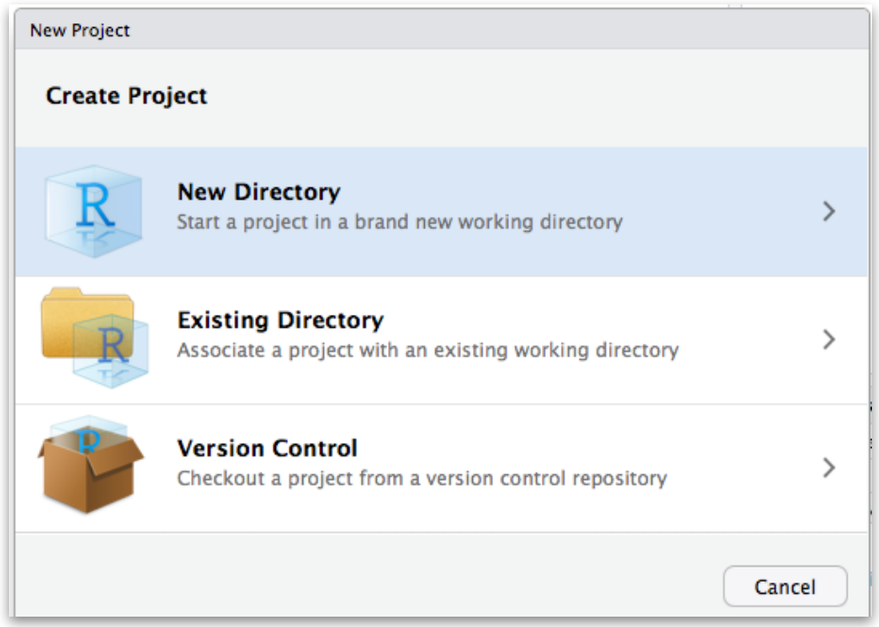
Replication | TOK Topics
toktopics.com

psychological science ...
science.sciencemag.org

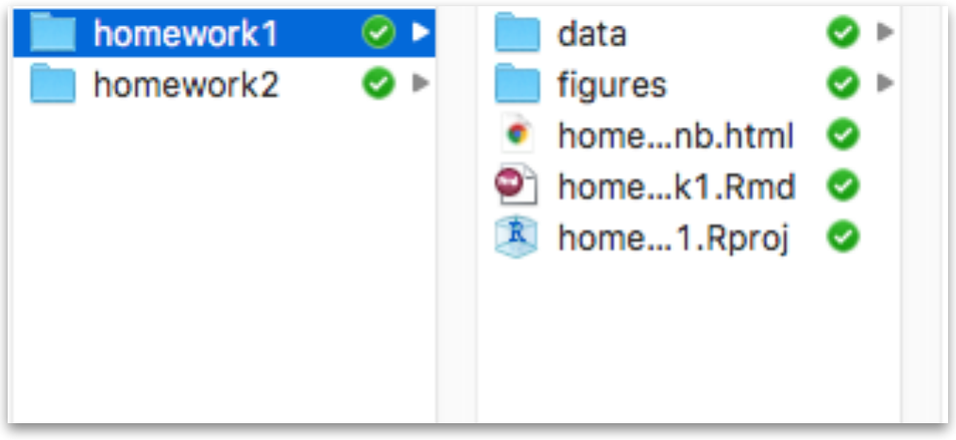
Replication Crisis ...
cambridge.org

Project organization

- Use **R Projects**
  - makes sure that the working directory is set correctly
  - allows you to work on multiple projects at the same time



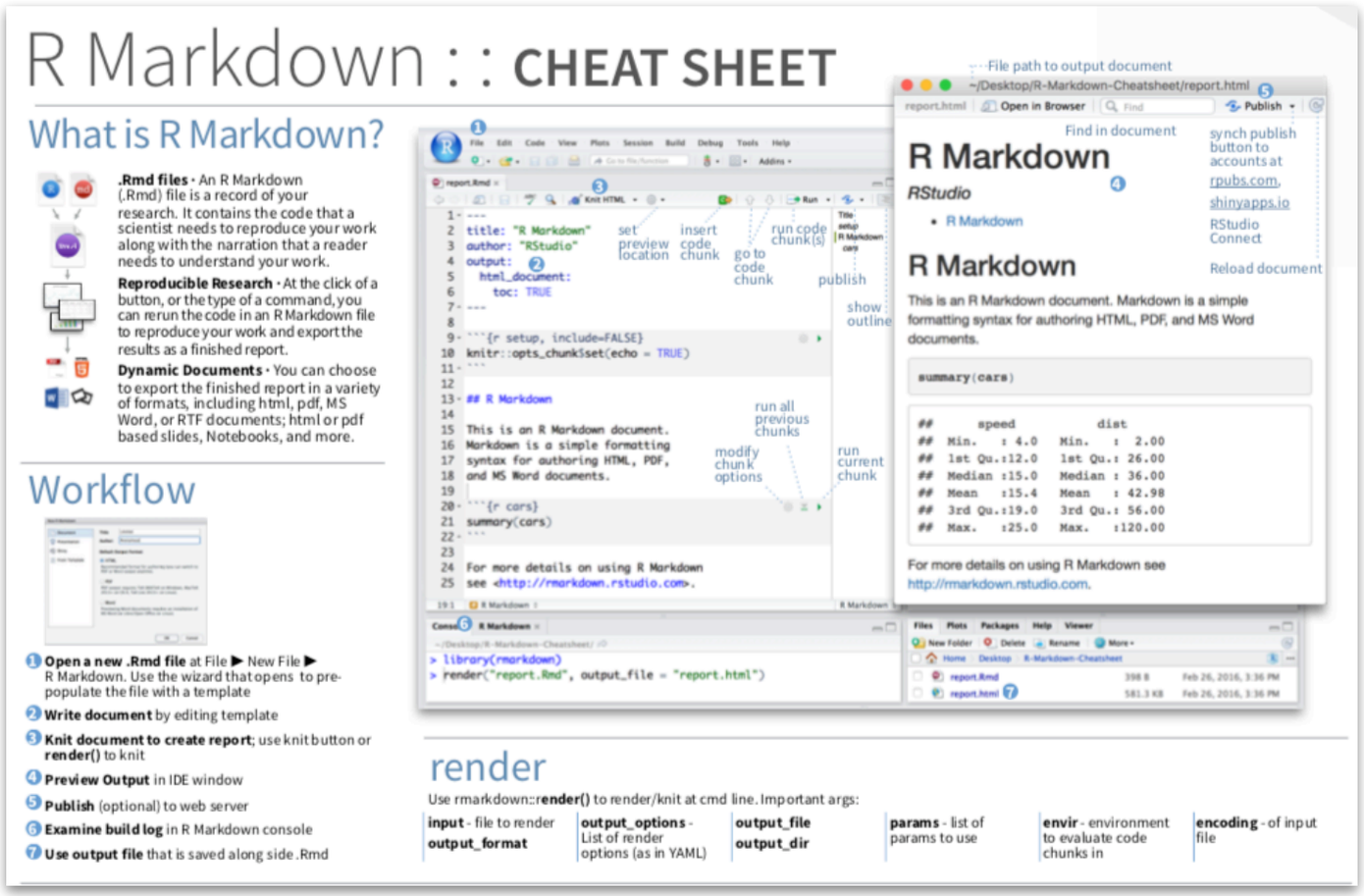
- **Organize** the folder structure
- Use **relative paths**



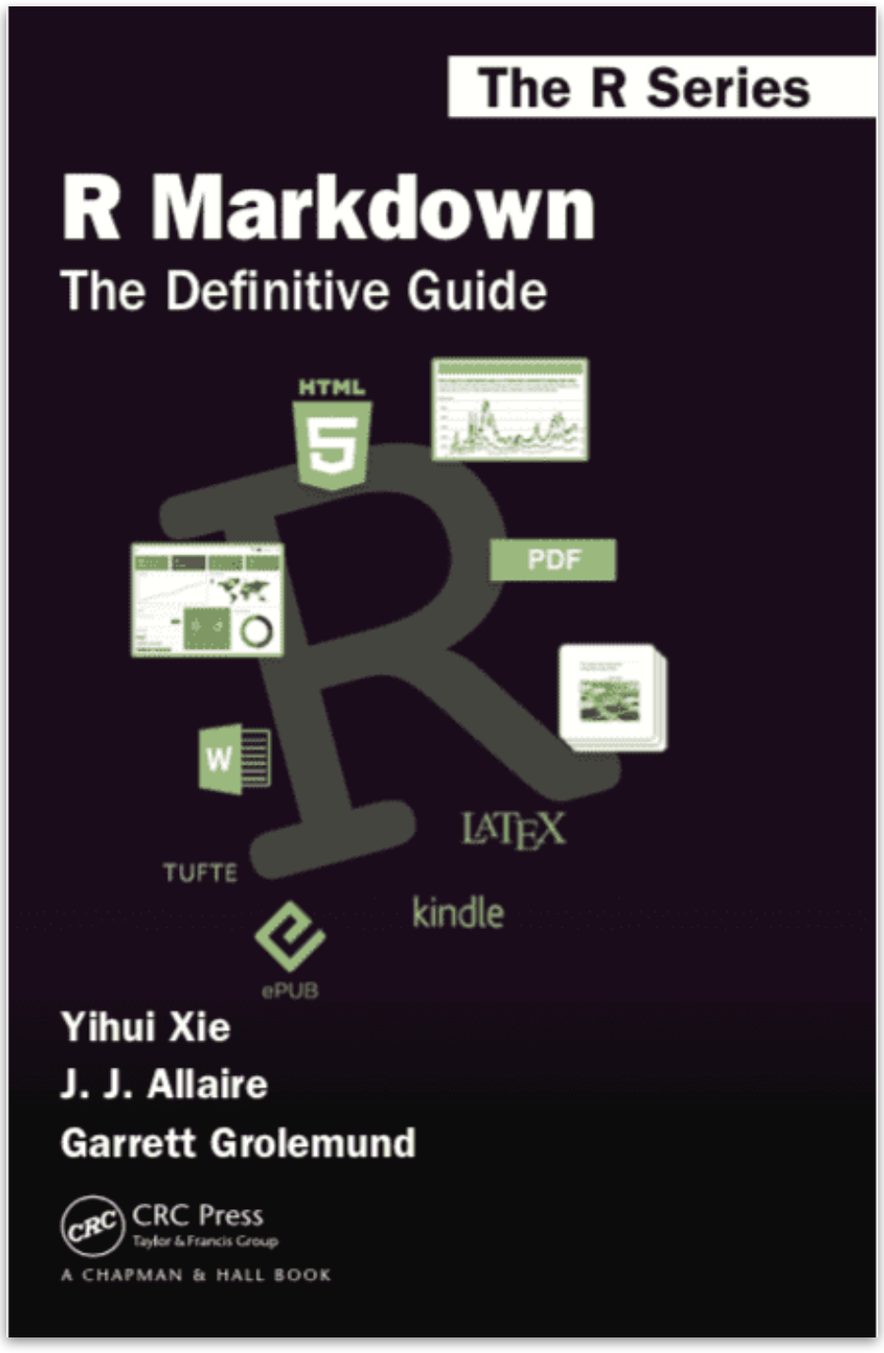
✓
```
1 read_csv("data/dataset.csv")
```

✗
```
1 read_csv("~/tobi/Documents/work/homework/data/
dataset.csv")
```

R Markdown

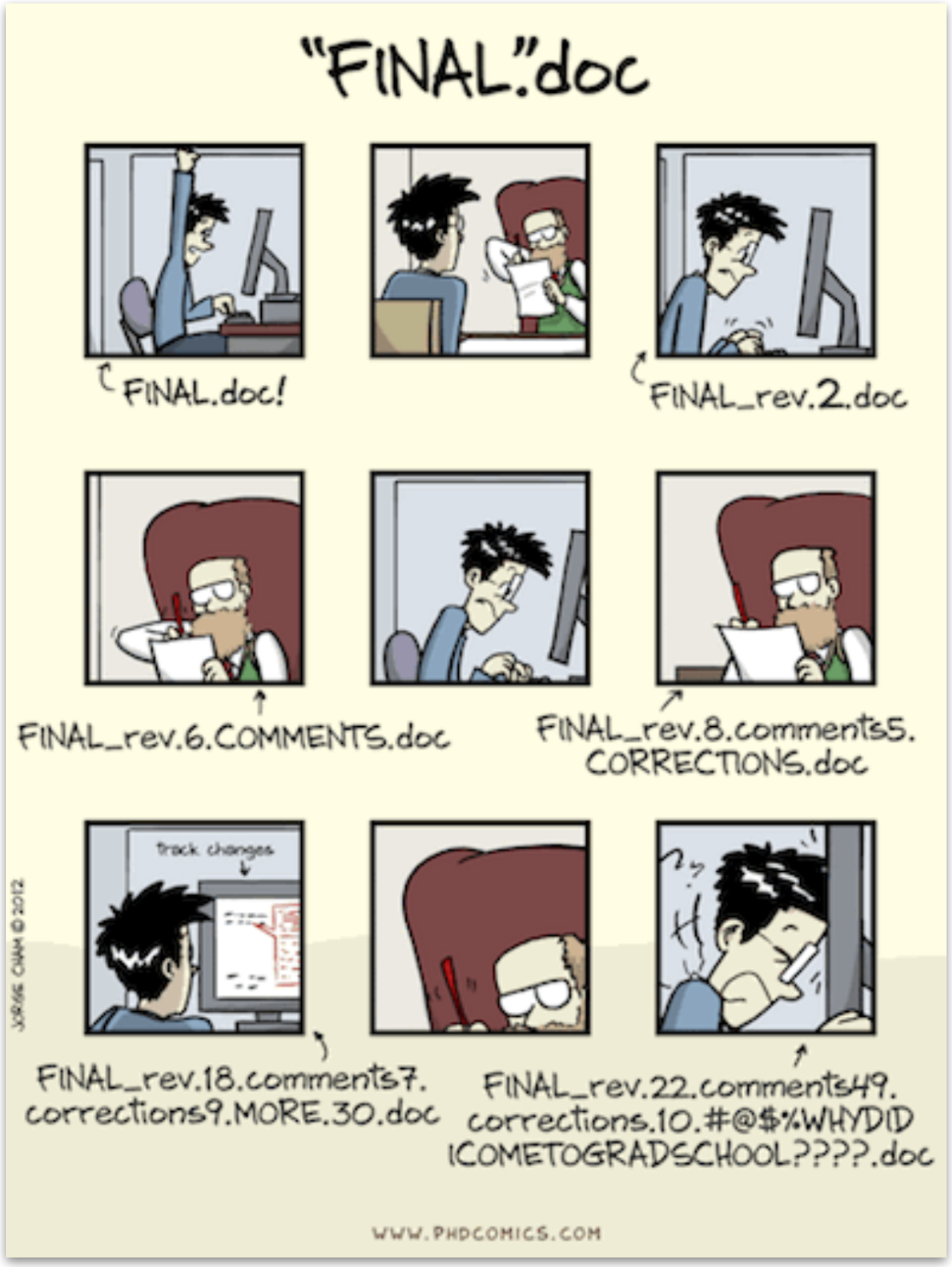


https://github.com/rstudio/cheatsheets/raw/master/rmarkdown-2.0.pdf



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)

How will we learn?

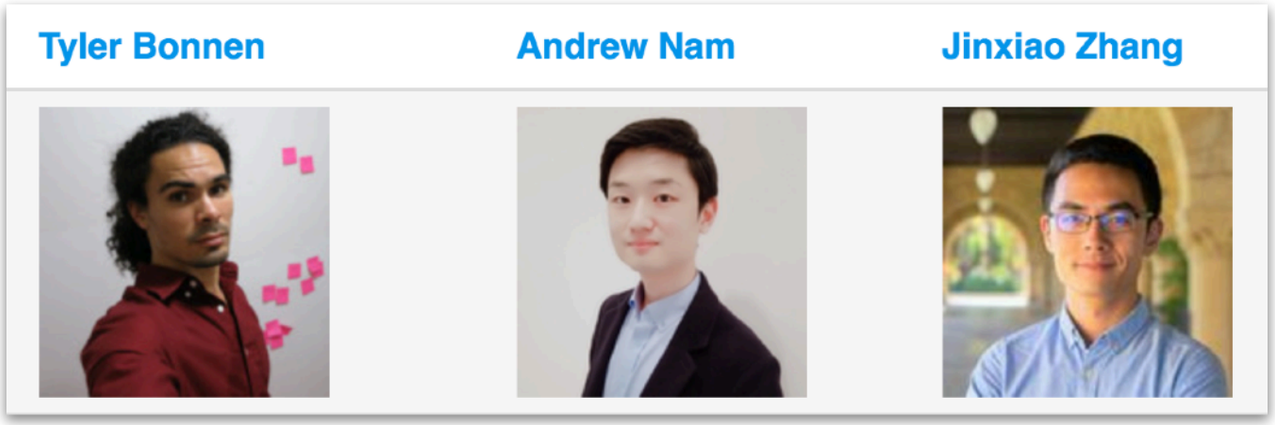
How will we learn?

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

How will we learn?

- **Lectures**
  - Monday, Wednesday, Friday @ 1:30pm in Gates B12 (here!)
  - get familiar with R and RStudio
    - visualization
    - data manipulation
    - simulation
  - introduce and discuss statistical methods

How will we learn?



- **Sections**

  - Section times are **Tuesday** and **Thursday** **4:30-5:20pm** in **160-322**.

Overview

Day	Date	Topic
Monday	January 6th	Introduction
Wednesday	January 8th	Visualization I
Friday	January 10th	Visualization II
Monday	January 13th	Data wrangling I
Tuesday	January 14th	Homework section: Visualization
Wednesday	January 15th	Data wrangling II
Thursday	January 16th	Application section: Visualization & Data wrangling
Friday	January 17th	Probability
Monday	January 20th	No class (Martin Luther King Jr. Day)

Homework section

Application section

How will we learn?

- **Homework sections (on Tuesdays)**
  - TAs facilitate working on the weekly homework
  - you'll be able to work together and ask the TAs questions
  - come here instead of office hours when you have questions about the homework

How will we learn?

- **Application sections (on Thursdays)**
  - TAs will help you apply what we've discussed in lectures to cool datasets
  - will build on lectures but also go beyond
  - highly recommended to attend!

How will we learn?

- **Homework**
  - one assignment per week (7 in total)
  - you can work in groups (indicate who you worked with on your submission)
  - download and submit via Canvas
  - you'll submit an RMarkdown file rendered as a pdf
  - homework will be available online after class on Friday, and is due **Thursday 8pm** the week after

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 (2/7) and is due on Thursday (2/13) at 8pm
    - you can use any resources you like
    - will be submitted to Canvas as a RMarkdown file

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

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. <u>short written project proposal</u>

     - due February 18th
     - one page maximum
     - we'll provide an RMarkdown template
     - and some examples from last year

How will we learn?

- **Final project**

  2. <u>oral presentation</u>

     - on exam day March 19th
     - short slideshow of the project
     - groups present together
     - presentation times scales with group size
     - TAs and I will grade independently and then pool our grades

How will we learn?

- **Final project**

  3. written report

     - due March 20th
     - ~ 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
     - we'll make some examples available from last year

How will we learn?

late submissions count as 0

- **Grading**
  - Homework: 40% (the homework with the lowest score doesn't count)
  - Midterm: 20%
  - Final project: 40%
    - Proposal: 5%
    - Presentation: 10%
    - Report: 25%
  - Extra credit
    - Piazza: 2%
    - Datacamp: 2%

How will we learn?

- **Office hours**

	Tobi Gerstenberg	Tyler Bonnen	Andrew Nam	Jinxiao Zhang
				
Role:	Instructor	Teaching assistant	Teaching assistant	Teaching assistant
Email:	gerstenberg@stanford.edu	bonnen@stanford.edu	ajhnam@stanford.edu	jinx.zhang@stanford.edu
Office:	420-302	420-406	420-326	420-425
Office hours:	Monday 4:00-5:00pm	Wednesday 3:00-4:00pm	Friday 12:30-1:30pm	Tuesday 3:00-4:00pm

Tools we will use in class

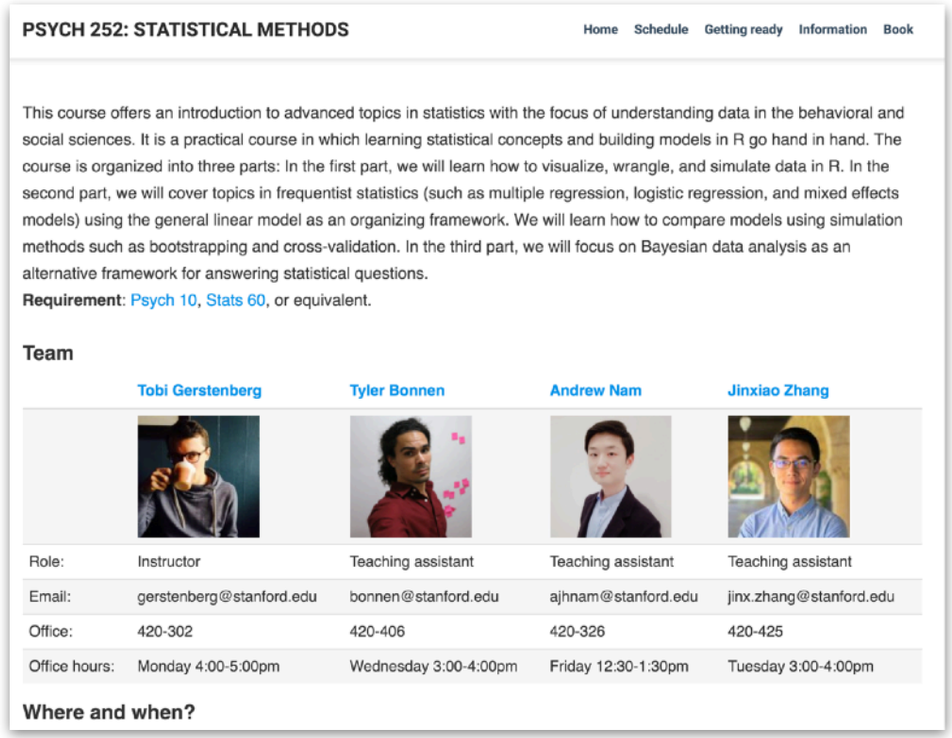
- Canvas
- Course homepage
- PollEverywhere
- Datacamp
- Piazza
- Free online books

Tools we will use in class

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

Tools we will use in class

- **Course homepage**
  - general information
  - schedule with links to materials (will keep being updated)
  - **https://psych252.github.io/**
  - you can access it via canvas



PSYCH 252: STATISTICAL METHODS

Home Schedule Getting ready Information Book

This course offers an introduction to advanced topics in statistics with the focus of understanding data in the behavioral and social sciences. It is a practical course in which learning statistical concepts and building models in R go hand in hand. The course is organized into three parts: In the first part, we will learn how to visualize, wrangle, and simulate data in R. In the second part, we will cover topics in frequentist statistics (such as multiple regression, logistic regression, and mixed effects models) using the general linear model as an organizing framework. We will learn how to compare models using simulation methods such as bootstrapping and cross-validation. In the third part, we will focus on Bayesian data analysis as an alternative framework for answering statistical questions.

Requirement: Psych 10, Stats 60, or equivalent.

Team

	Tobi Gerstenberg	Tyler Bonnen	Andrew Nam	Jinxiao Zhang
Role:	Instructor	Teaching assistant	Teaching assistant	Teaching assistant
Email:	gerstenberg@stanford.edu	bonnen@stanford.edu	ajhnam@stanford.edu	jinx.zhang@stanford.edu
Office:	420-302	420-406	420-326	420-425
Office hours:	Monday 4:00-5:00pm	Wednesday 3:00-4:00pm	Friday 12:30-1:30pm	Tuesday 3:00-4:00pm

Where and when?

Tools we will use in class

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

How are you feeling today?

Demonstrate that data are in fact anonymous

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/psych252datacamp20**
  - get extra credit for taking courses

Schedule

Week	Day	Date	Topic	Content	Reading	Resources	Datacamp
1	M	7-Jan	Introduction	• Course introduction		• Cheatsheet R Studio • Cheatsheet R Markdown 1 • Cheatsheet R Markdown 2 • R Markdown for class reports	• Introduction to R • RStudio IDE 1 • RStudio IDE 2 • RMarkdown
	W	9-Jan	Visualization I	• Best practices • Introduction to RStudio • Introduction to `library(ggplot2)` • Reporting results using Rmarkdown	• Data visualization (#1) • Data visualization (#3)	• Cheatsheet ggplot2	• ggplot part 1 • ggplot part 2 • Reporting

Tools we will use in class

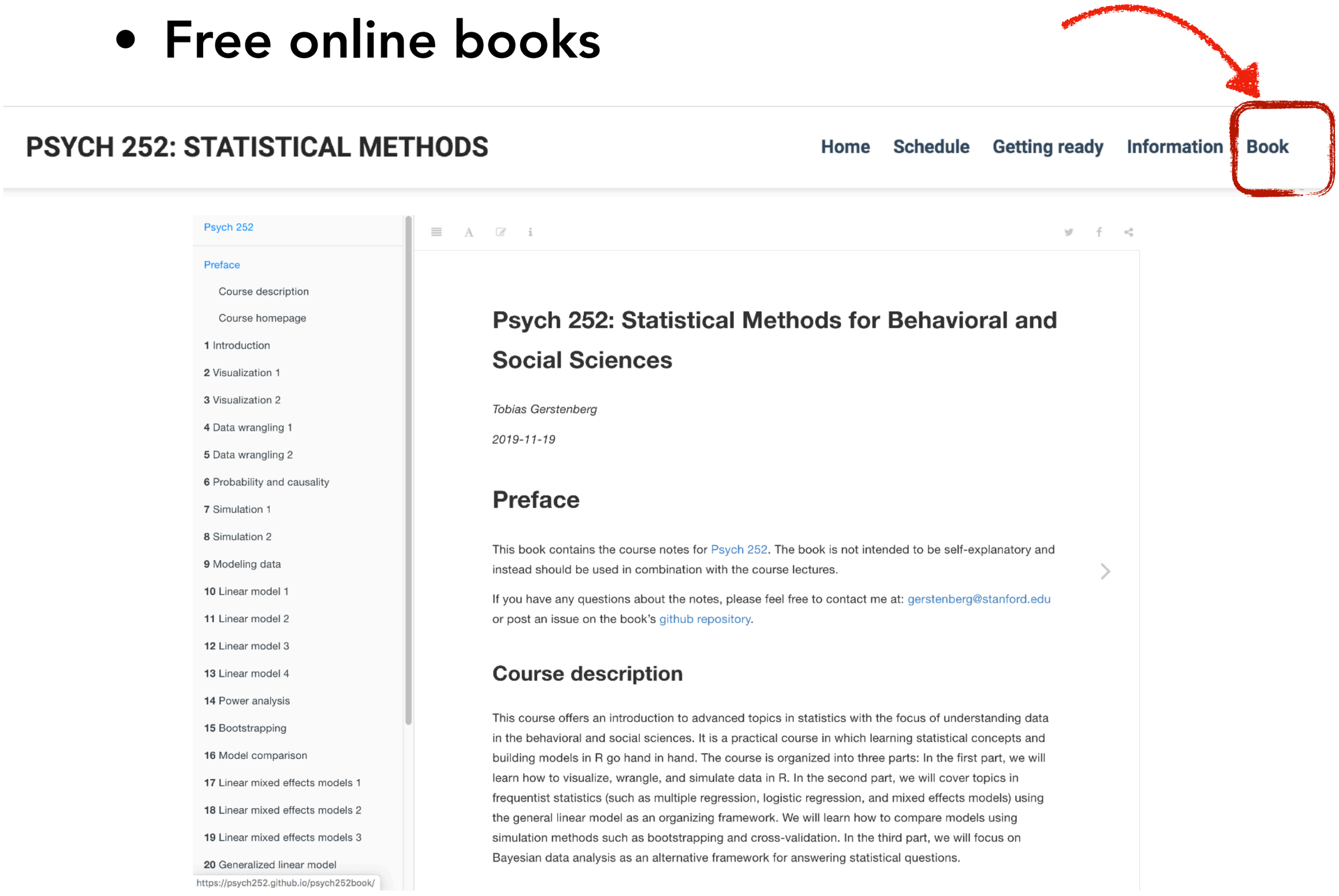
- **Piazza**
  - post your questions and answer your colleagues' questions (for extra credit, yay!)
  - form groups for the final project
  - sign up here: **https://tinyurl.com/psych252piazza20**
  - access code**: sunnystats**

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 I didn't find a single book that fits the bill
  - **but**: many great, free books online
  - **and**: I will point out suggested readings as we go along
  - **also**: I'll update an online book based on the 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

Piazza

I have a question about the homework.

Canvas

I would like to submit my homework on time.

Course website

I don't remember the schedule.

Datacamp

I would like to learn more about a topic.

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**

Vision for this class

In "A Vision for Stanford", university president Marc Tessier-Lavigne states that Stanford wants to be

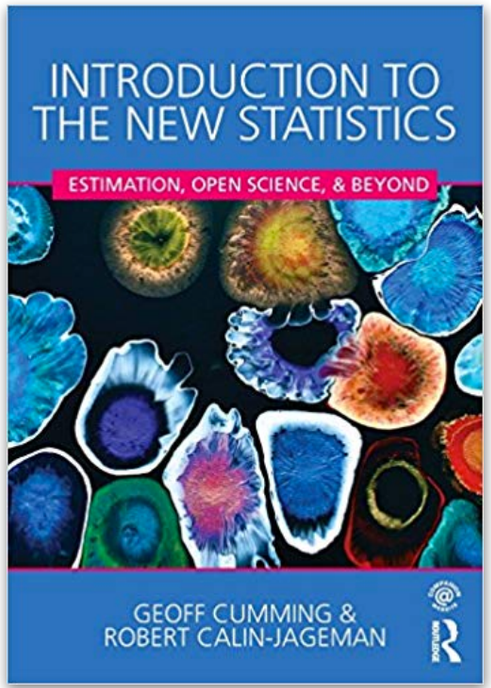
"an inspired, inclusive and collaborative community of diverse scholars, students and staff, where all are supported and empowered to thrive."

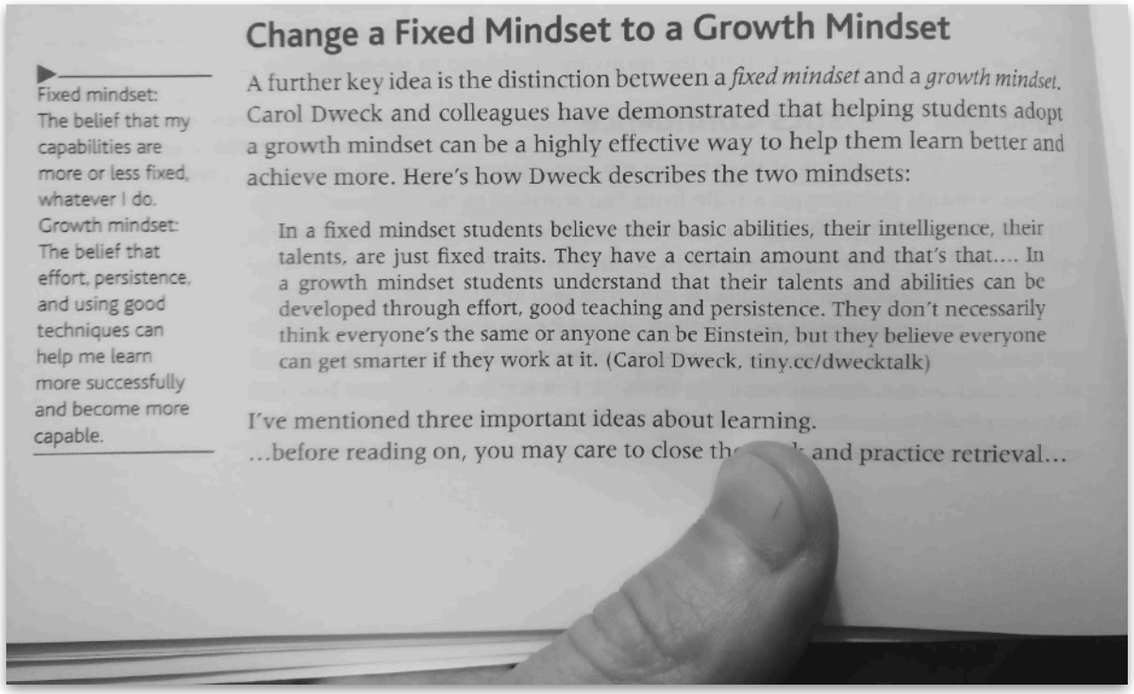
Let's try our best together in this class to make this happen!

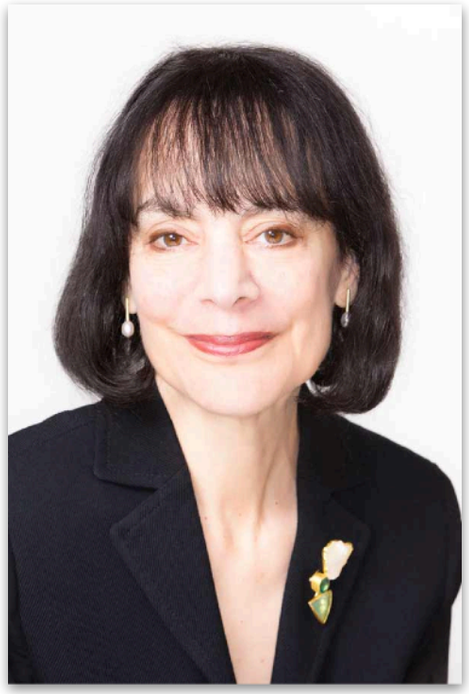
Fear of statistics

- I don't want to make a mistake ...
- What if my analysis script has bugs?
- Best response: openness!
- We will use **github** for the final projects.

Fear of statistics







my 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

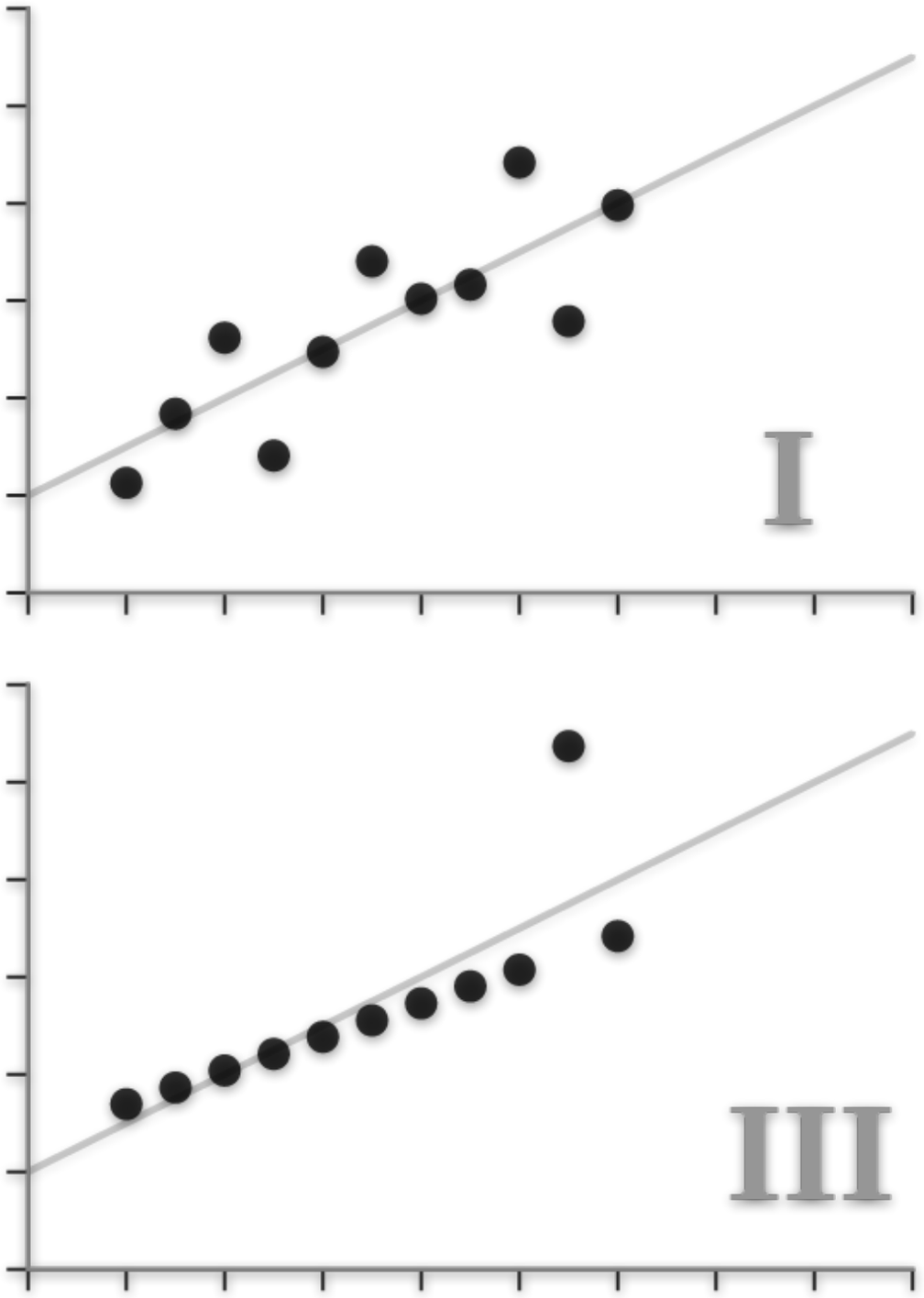
Visualization

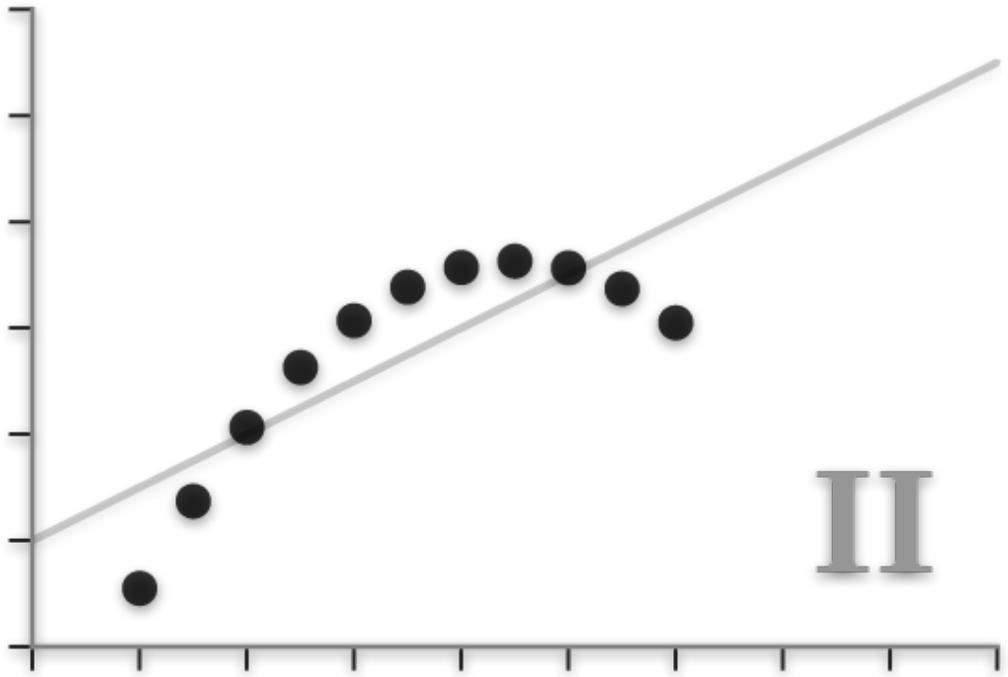
some inspirational quotes ...

> The greatest value of a picture is when it forces us to notice what we never expected to see. — John Tukey

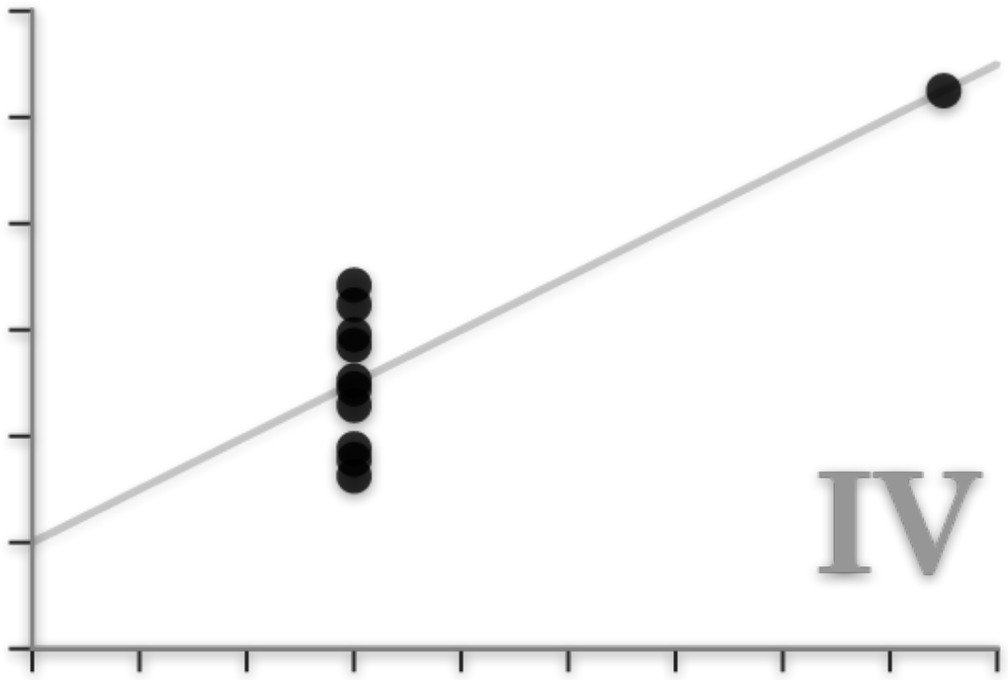
> There is no single statistical tool that is as powerful as a well–chosen graph. — Chambers et al. (1983)

> …make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding. — Anscombe (1973)

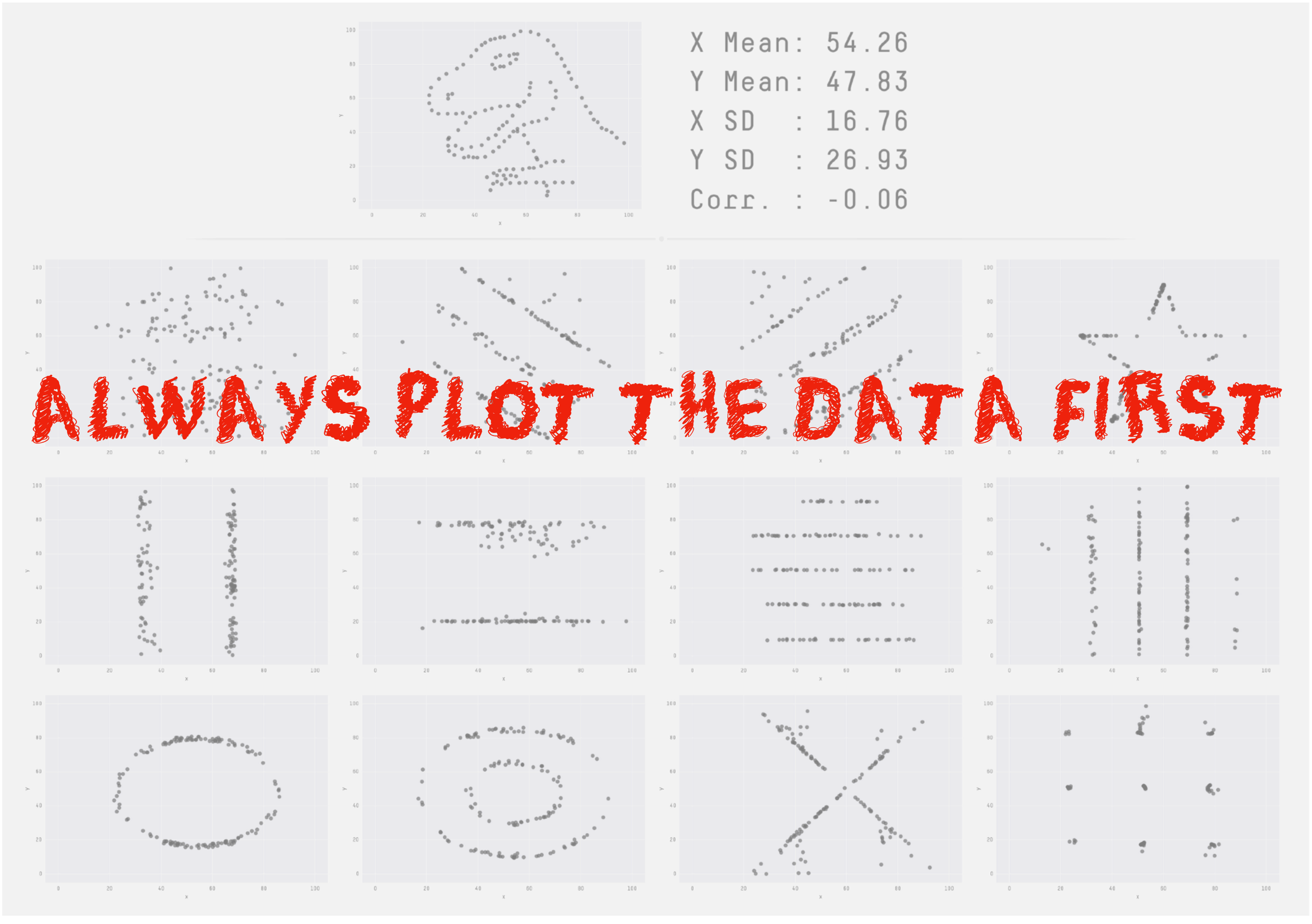

I


II

III

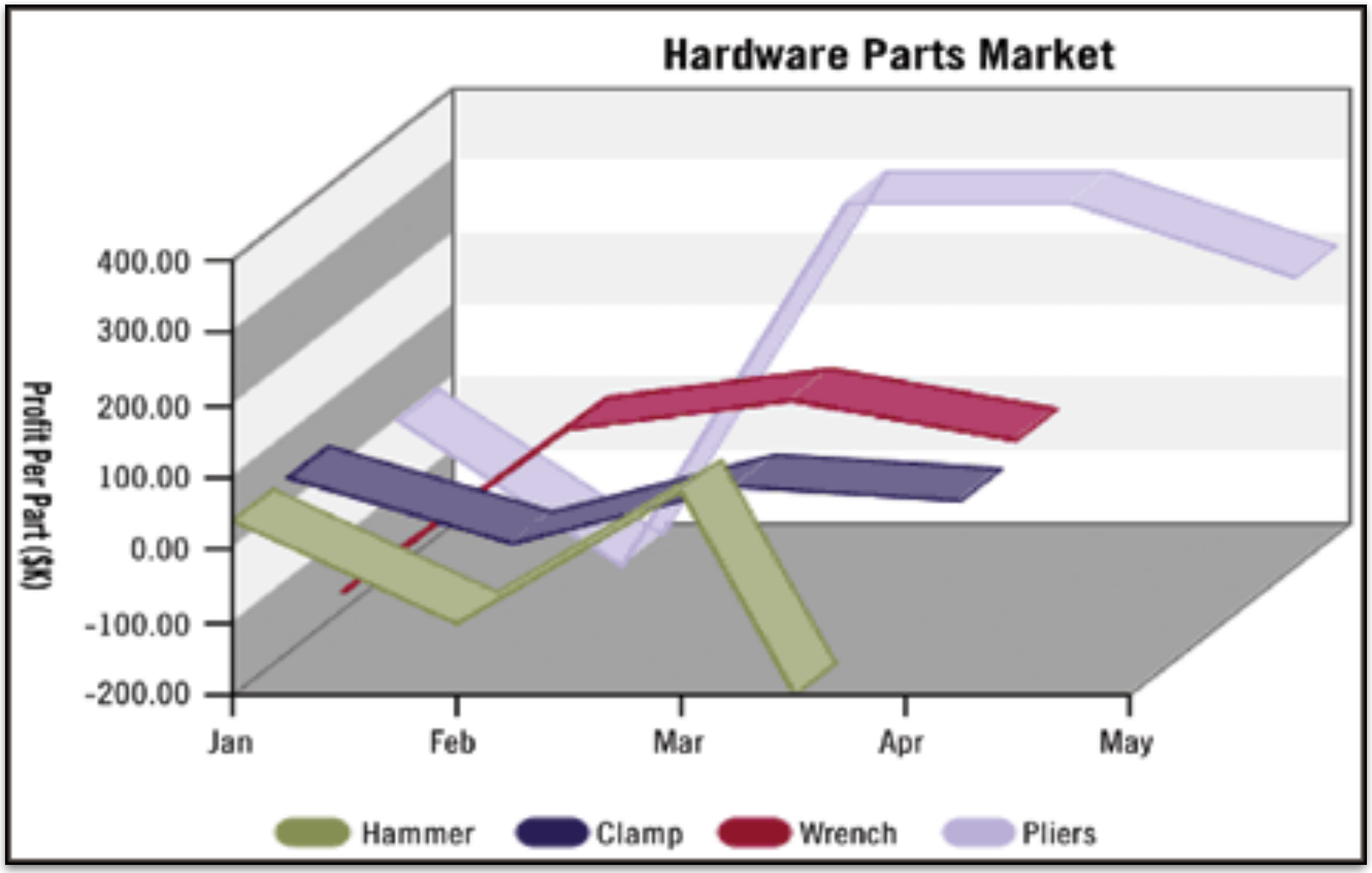

IV

From Anscombe to T-Rex

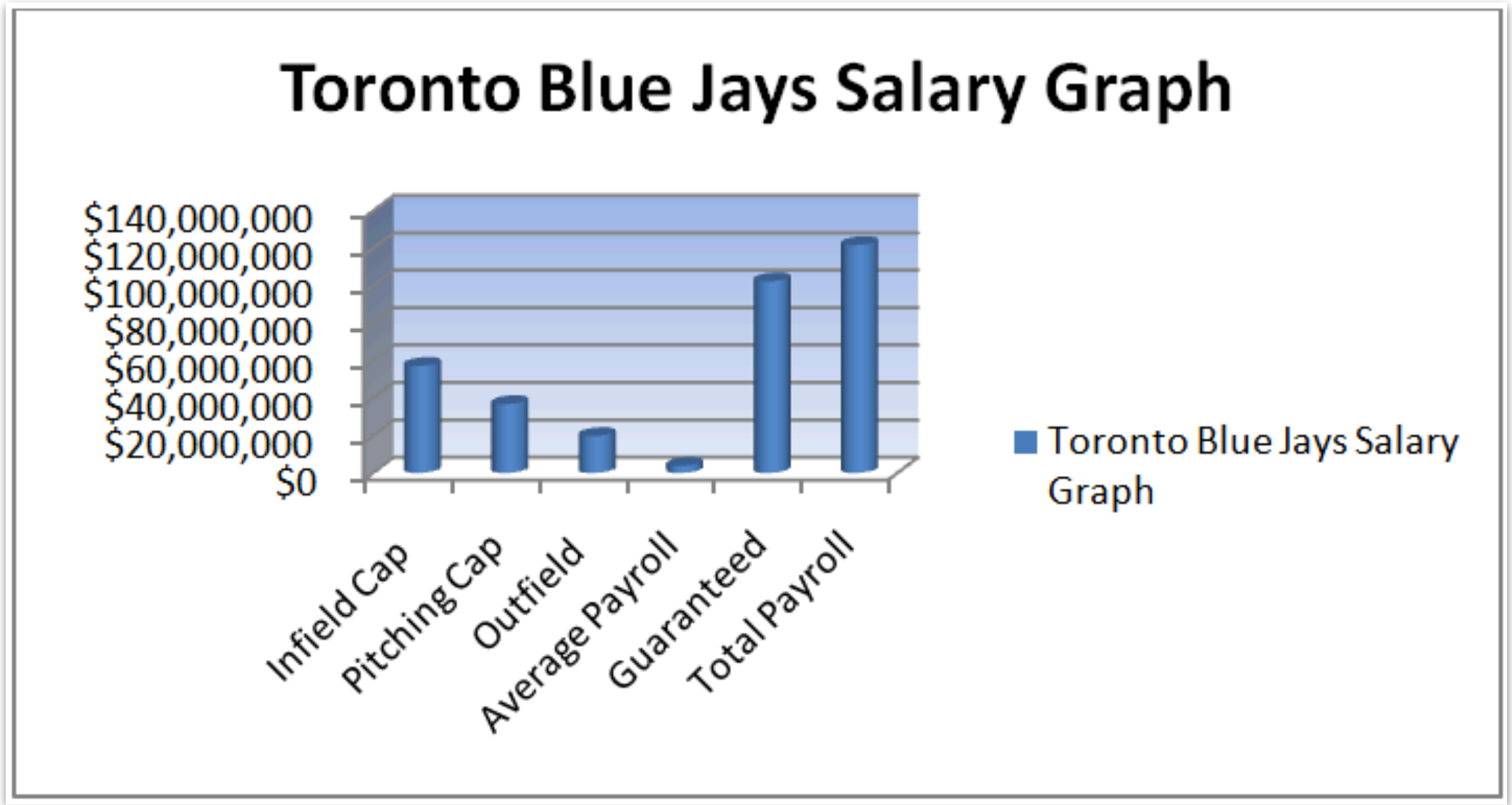


Plots that can be improved ...

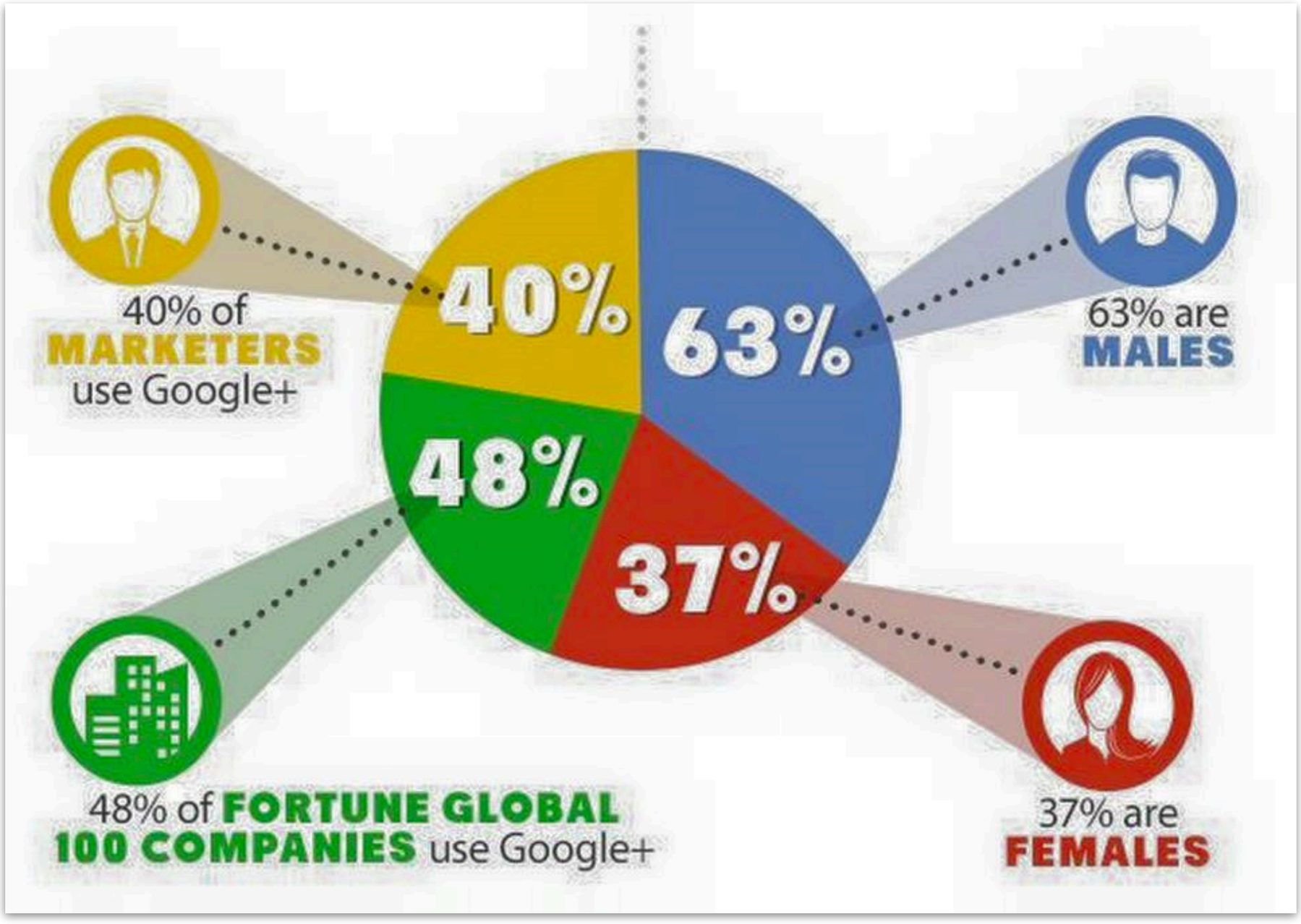
Plots that can be improved ...



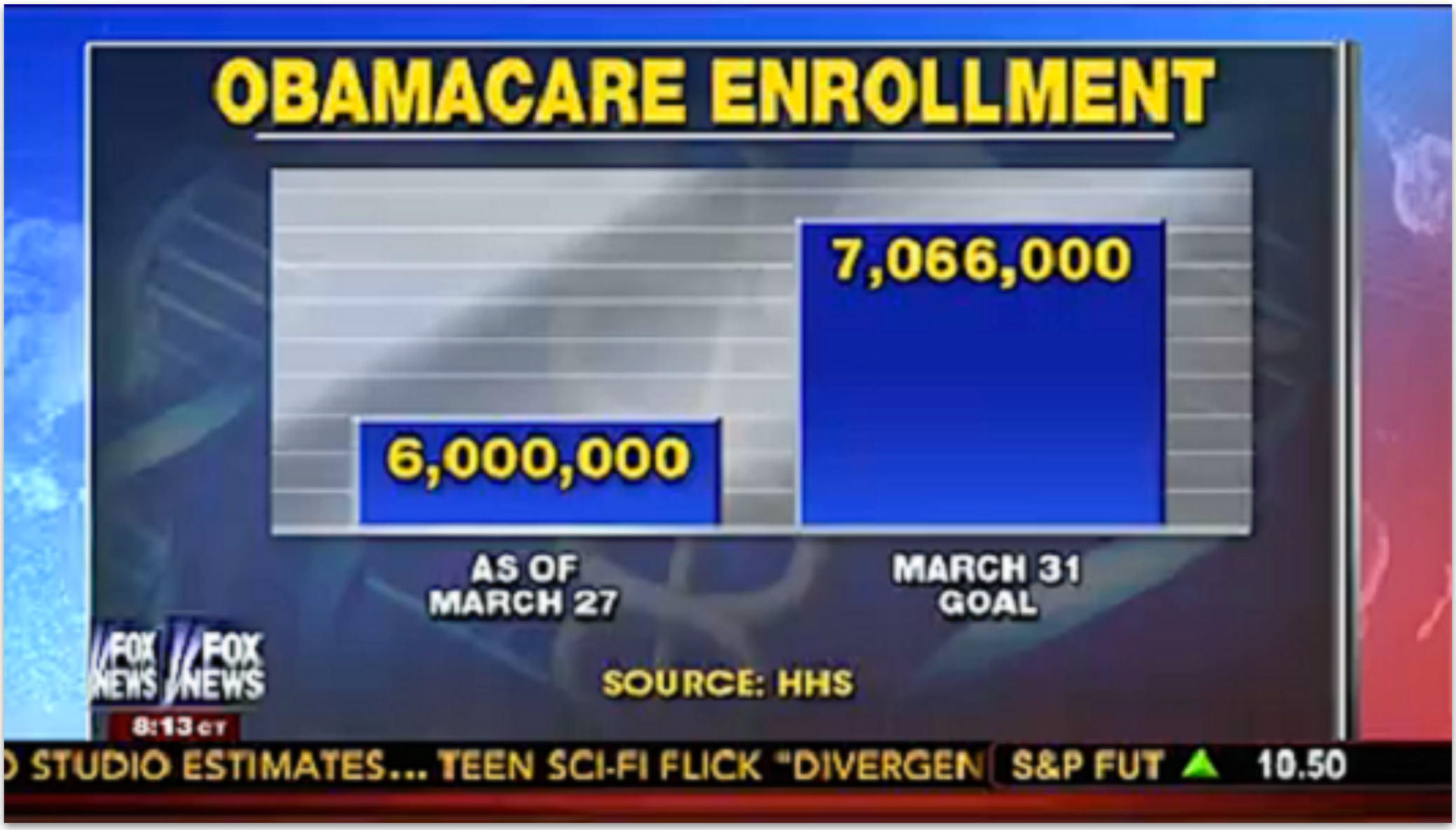
Plots that can be improved ...



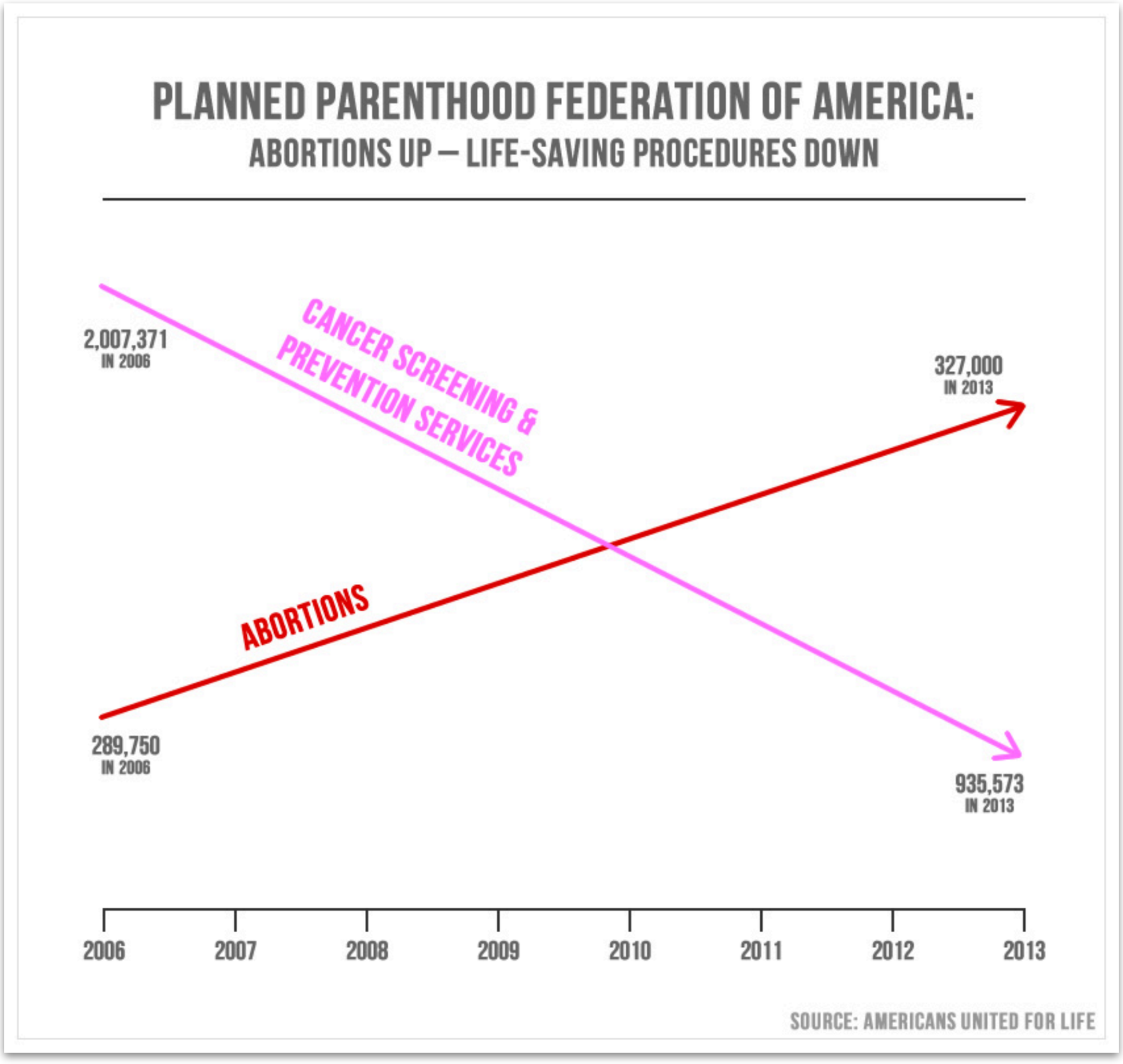
Plots that can be improved ...



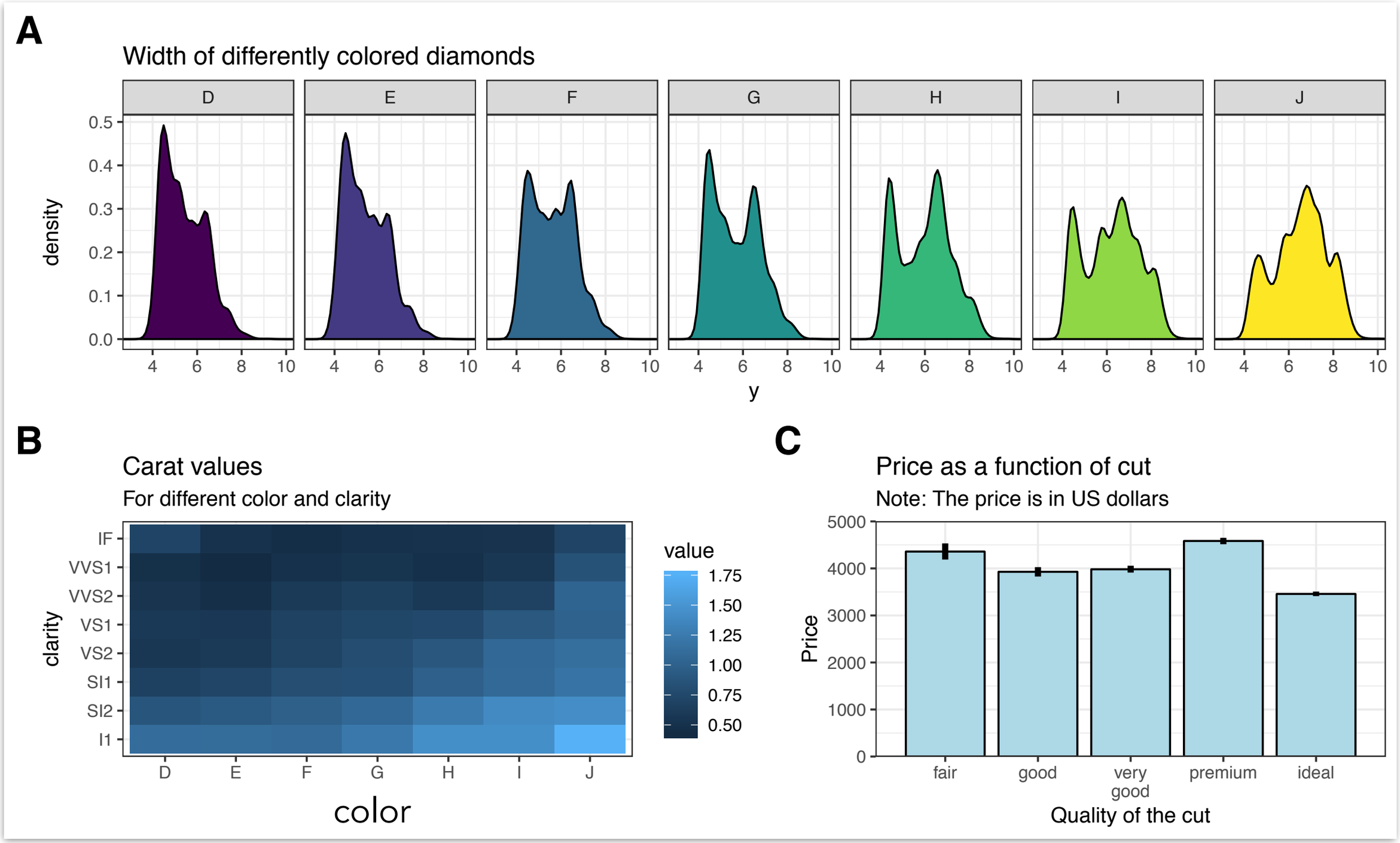
Plots that can be improved ...



Plots that can be improved ...



A nicer plot


A
Width of differently colored diamonds
D
E
F
G
H
I
J
density
0.5
0.4
0.3
0.2
0.1
0.0
4
6
8
10
y
B
Carat values
For different color and clarity
clarity
IF
VVS1
VVS2
VS1
VS2
SI1
SI2
I1
D
E
F
G
H
I
J
color
value
1.75
1.50
1.25
1.00
0.75
0.50
C
Price as a function of cut
Note: The price is in US dollars
Price
5000
4000
3000
2000
1000
0
fair
good
very good
premium
ideal
Quality of the cut

Feedback

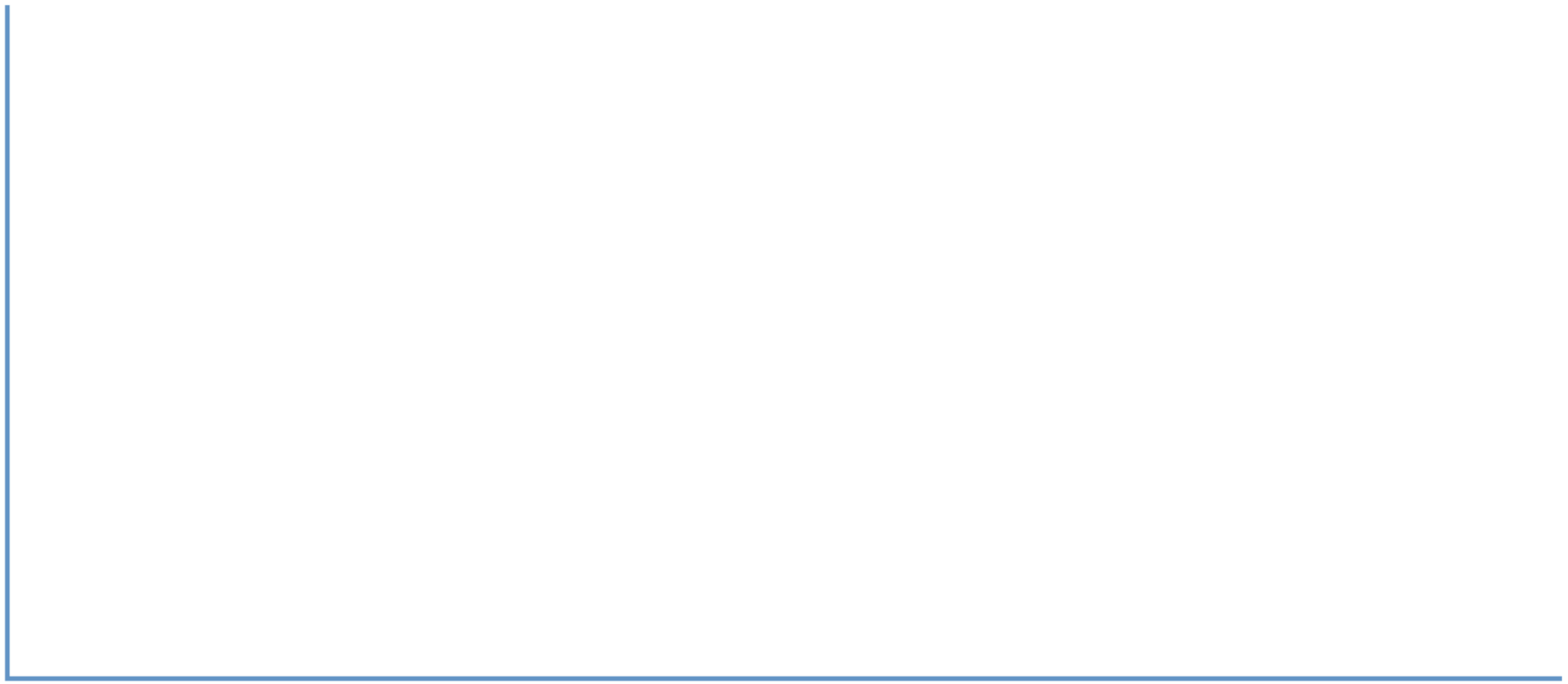
Feedback

- ask questions in class and/or come to us after class
- post questions on Piazza (you can post anonymously)
- send us an email (but use Piazza for questions about homework or class!)
- use anonymous feedback form (link at the bottom of the course website) https://tinyurl.com/psych252feedback
- come to office hours

Feedback


Dan Quintana
@dsquintana
stats lecture
Your audience when giving ~~a talk:~~
stats lecture
What you see when giving ~~a talk:~~
2:37 AM · Nov 24, 2019 · Twitter for iPhone
162 Retweets
1.7K Likes

How was the pace of today's class?



much too slow

a little too slow

just right

a little too fast

much too fast

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
Ben Baumer
Benoit Monin
Datacamp
David Lagnado
Ewart Thomas
Henrik Singmann
Julian Jara-Ettinger
Justin Gardner
Kevin Smith
Maarten Speekenbrink
Matthew Kay
Matthew Salganik
Mika Braginsky
Mike Frank
Mine Çetinkaya-Rundel
Patrick Mair
Peter Cushner Mohanty
Richard McElreath
Russ Poldrack
Stephen Dewitt
Tom Hardwicke
Tristan Mahr

Thanks!

see you on **Wednesday**

bring a laptop

with **R** and **RStudio** installed and up to date

post on Piazza if you experience any problems