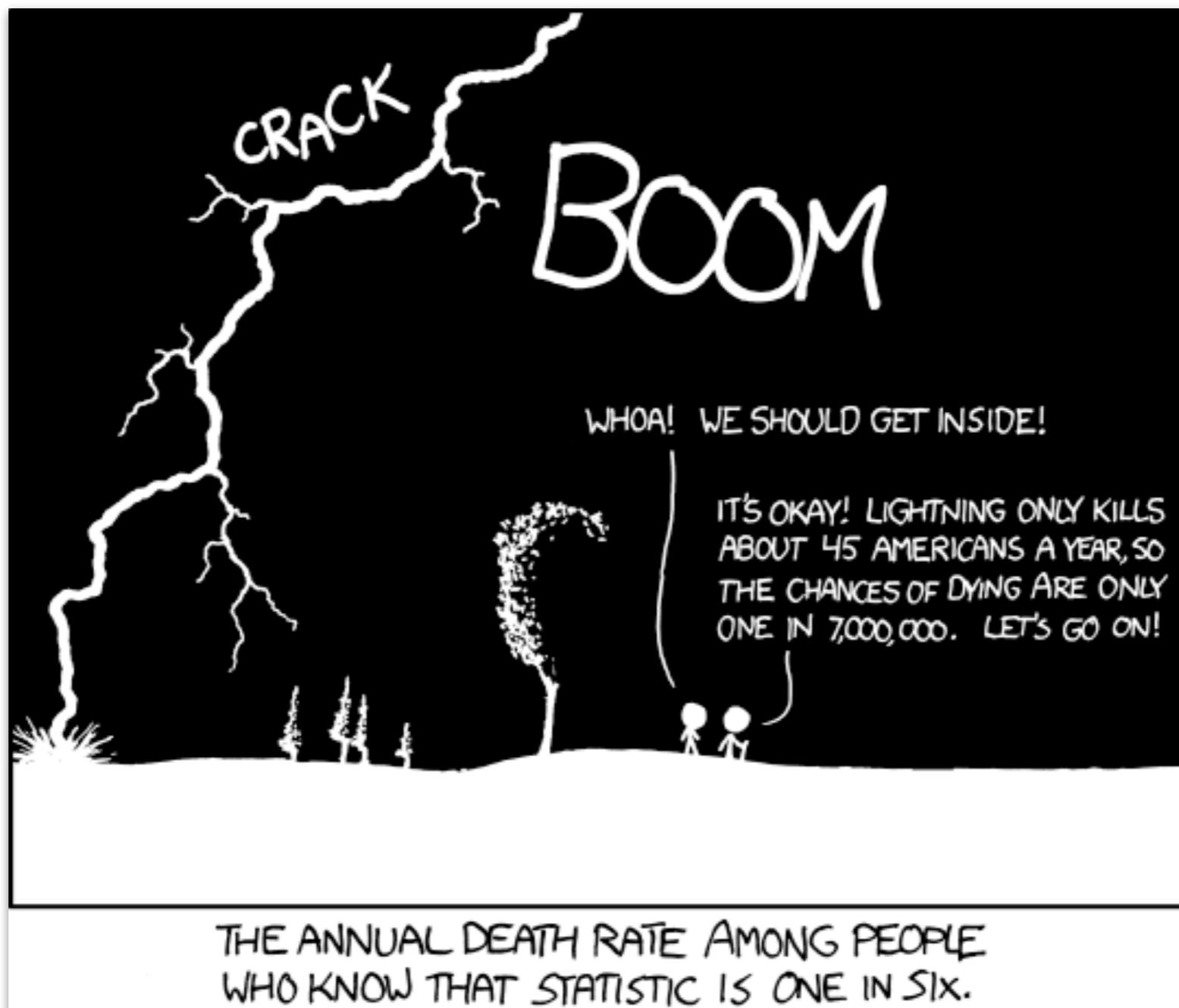


Introduction



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

The screenshot shows a Google Forms survey titled "Psych 252: Introductory survey". The interface includes a green header bar with the title, a toolbar with various icons, and a sidebar on the right with additional tools. The main content area contains the survey introduction, a required email address field, a question about graduate school year, and a question about department.

Psych 252: Introductory survey

This survey is meant to obtain some background information about the students in Psych 252, as well as to obtain some different types of data that we will use to analyze in the class. Please answer all questions as honestly as possible.

Some of the data collected in this survey may be made available to the class for use in analysis exercises. However, the data will be made anonymous before anything is released by removing any potentially identifying information, so that it will not be possible to determine which data came from which person in the group.

Please use your Stanford email address in order to get credit for completing the survey.

Email address*

Valid email address

This form is collecting email addresses. [Change settings](#)

What year of graduate school are you in?*

1. Undergraduate
2. 1-2
3. 3-4
4. 5+

What department are you in?*

Psychology

Education

(announcement soon via Canvas)

Tobias Gerstenberg (he/him/his)

Tobi Gerstenberg

- 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?

feel free
to call me



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

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)

[Tyler Bonnen](#)

- 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



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

Andrew Nam (he/him/his)

[Andrew Nam](#)

- 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.

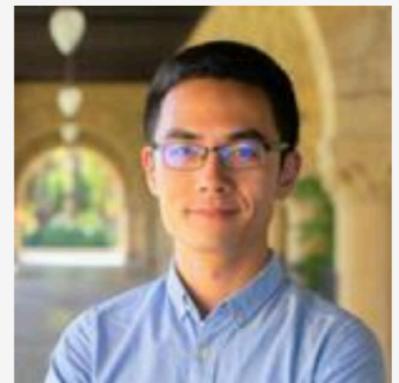


Role:	Teaching assistant
Email:	ajhnam@stanford.edu
Office:	420-326
Office hours:	Friday 12:30-1:30pm

Jinxiao Zhang (he/him/his)

[Jinxiao Zhang](#)

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



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

- Formulate hypotheses as statistical models
- Bayesian statistics

3. Report results

- Reproducible research

**all the time
(& Week 10)**

Learning goals

You will learn how to **use R** to ...

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

Use R!

Why ?

Visualization

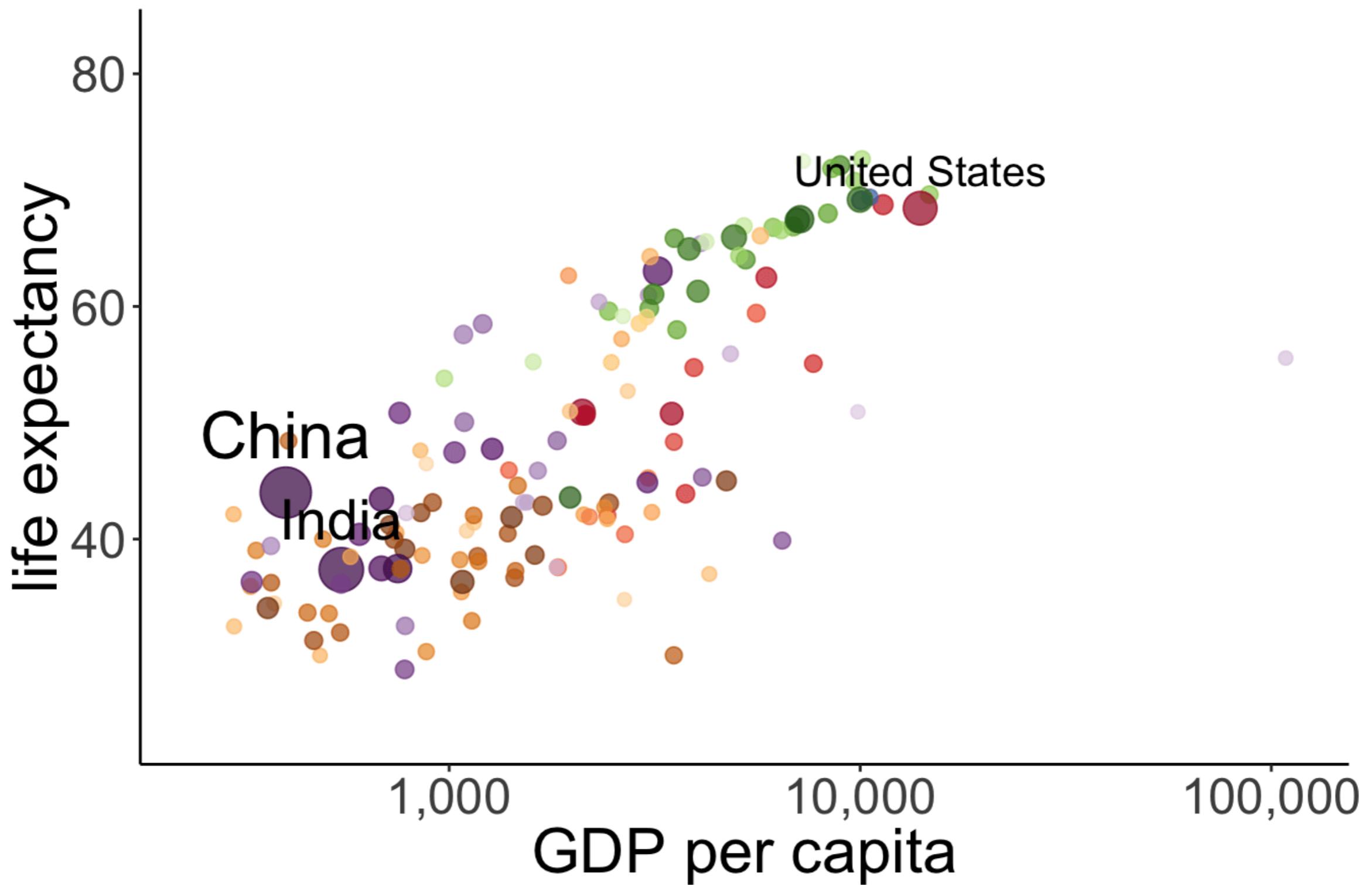
Data manipulation

Statistical modeling

Communicating results

Why R ? Visualization

Year: 1952



Why R ? Visualization

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

Why R ? Data manipulation

Data Transformation with dplyr :: CHEAT SHEET



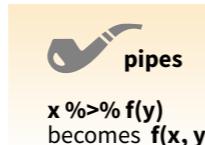
dplyr functions work with pipes and expect **tidy data**. In tidy data:



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

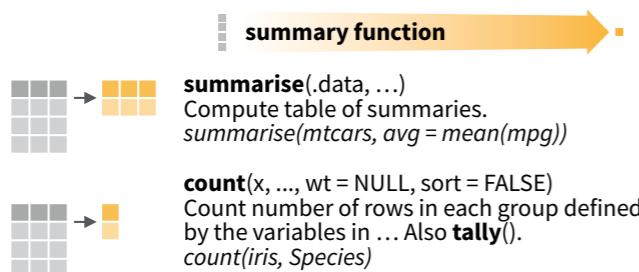


Each **observation**, or **case**, is in its own **row**



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).

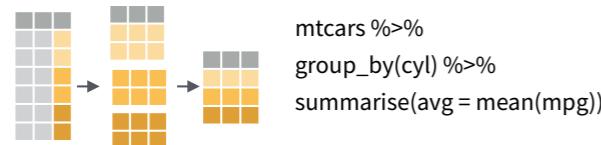


VARIATIONS

- 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.



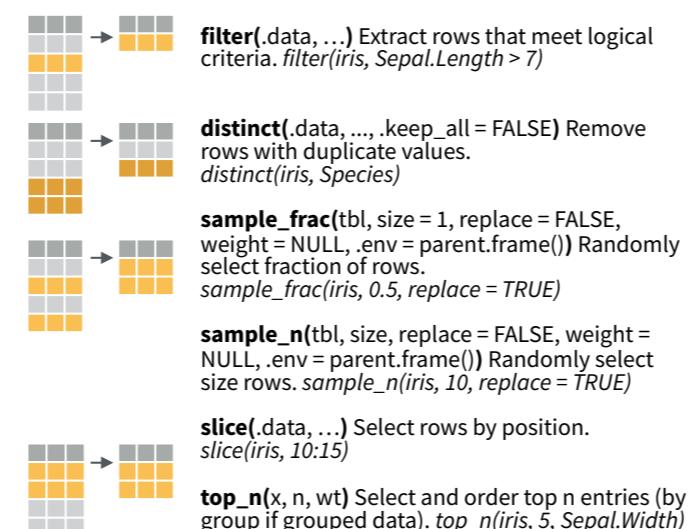
group_by(.data, ..., add = FALSE)
Returns copy of table grouped by ...
`g_iris <- group_by(iris, Species)`

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.



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))`

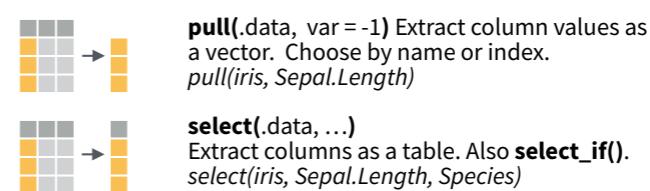
ADD CASES

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.

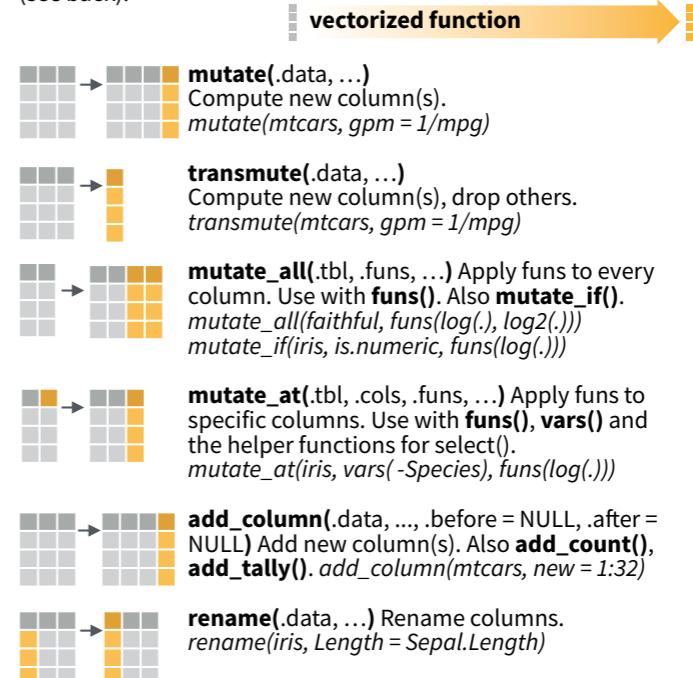


Use these helpers with **select ()**,
e.g. `select(iris, starts_with("Sepal"))`

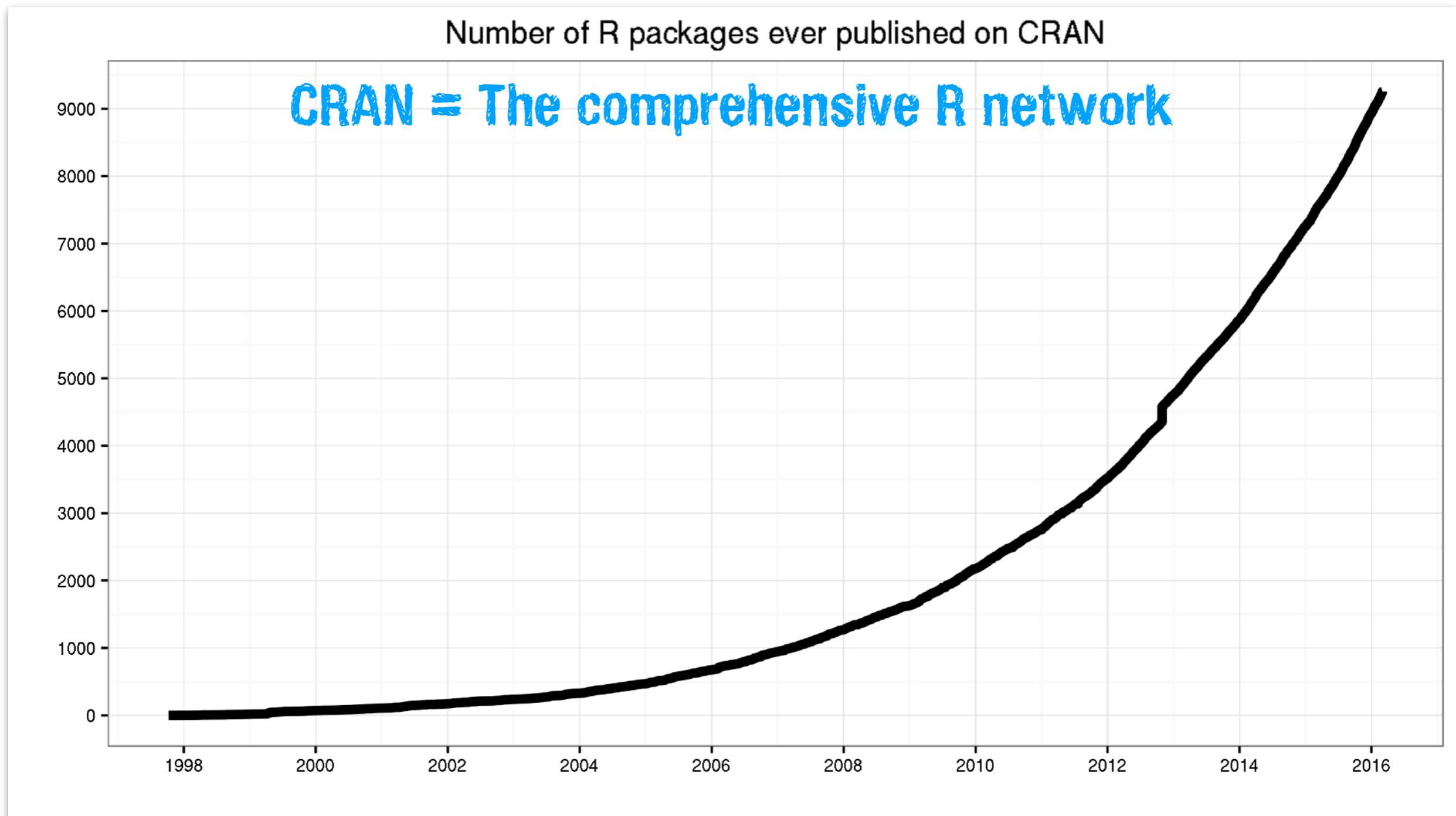
contains(match)	num_range(prefix, range)	:, e.g. <code>mpg:cyl</code>
ends_with(match)	one_of(...)	-, e.g. <code>-Species</code>
matches(match)	starts_with(match)	

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).



Why R ? Statistical modeling



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

Why R? Communicating results

```
01-intro.Rmd x
ABC 🔎 Knit ▾

1 # Introduction {#intro}
2
3 You can label chapter and section titles using `#{label}` after them,
do not manually label them, there will be automatic labels anyway, e.
4
5 Figures and tables with captions will be placed in `figure` and `tabl
6
7
8 ```{r nice-fig, fig.cap='Here is a nice figure!', out.width='80%', fi
9 par(mar = c(4, 4, .1, .1))
10 plot(pressure, type = 'b', pch = 19)
11 ```
12
13 Reference a figure by its code chunk label with the `fig:` prefix, e.
can reference tables generated from `knitr::kable()`, e.g., see Table
14
15 ```{r nice-tab, tidy=FALSE}
16 knitr::kable(
17   head(iris, 20), caption = 'Here is a nice table!',
18   booktabs = TRUE
19 )
20 ```
21
22 You can write citations, too. For example, we are using the **bookdown**
was built on top of R Markdown and **knitr** [@xie2015].
23
24 ```{r stats-help, fig.cap='Stats cheatsheet',fig.align='center',echo=
25 knitr::include_graphics('figures/cheatsheets/stats-help.jpg')
26 ```
27
28 See figure \@ref(fig:stats-help)
29
30
31 ```{r klippy, echo=FALSE, include=TRUE}
32 klippy::klippy()
33 ...
34
```

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.

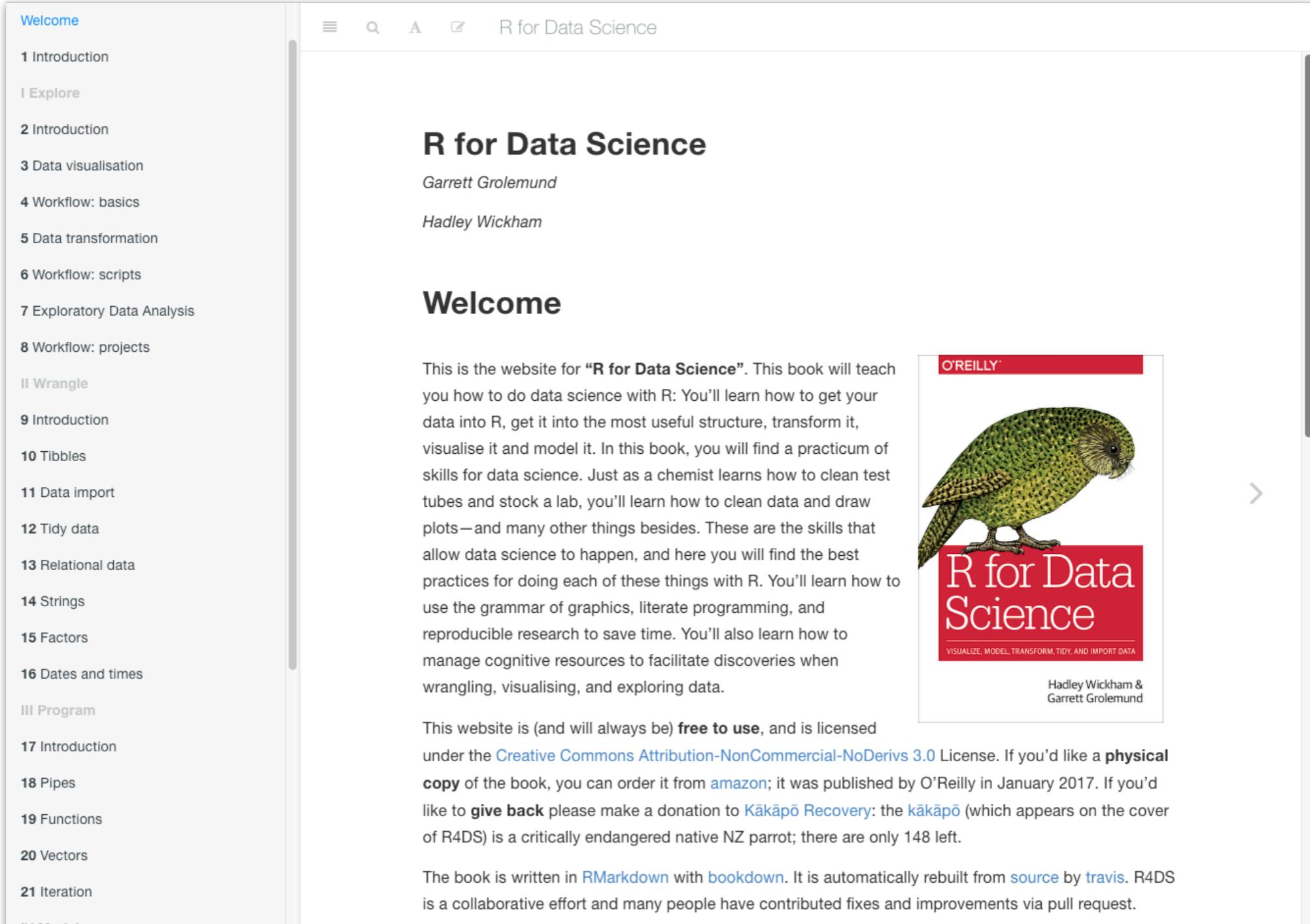
```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

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.

R Markdown

Why R ? Communicating results



The screenshot shows the homepage of the R for Data Science website. The left sidebar contains a table of contents for the book, organized into four main sections: I Explore, II Wrangle, III Program, and IV Model. The main content area features the book's title "R for Data Science" in large bold letters, followed by the authors' names, Garrett Grolemund and Hadley Wickham. Below this is a "Welcome" section with a detailed description of what the book teaches. To the right of the welcome text is an image of the book cover, which features a green parrot (Kākāpō) and the title "R for Data Science". At the bottom of the page, there is a note about the website's license and a link to the book's source code.

Welcome

R for Data Science

Garrett Grolemund
Hadley Wickham

Welcome

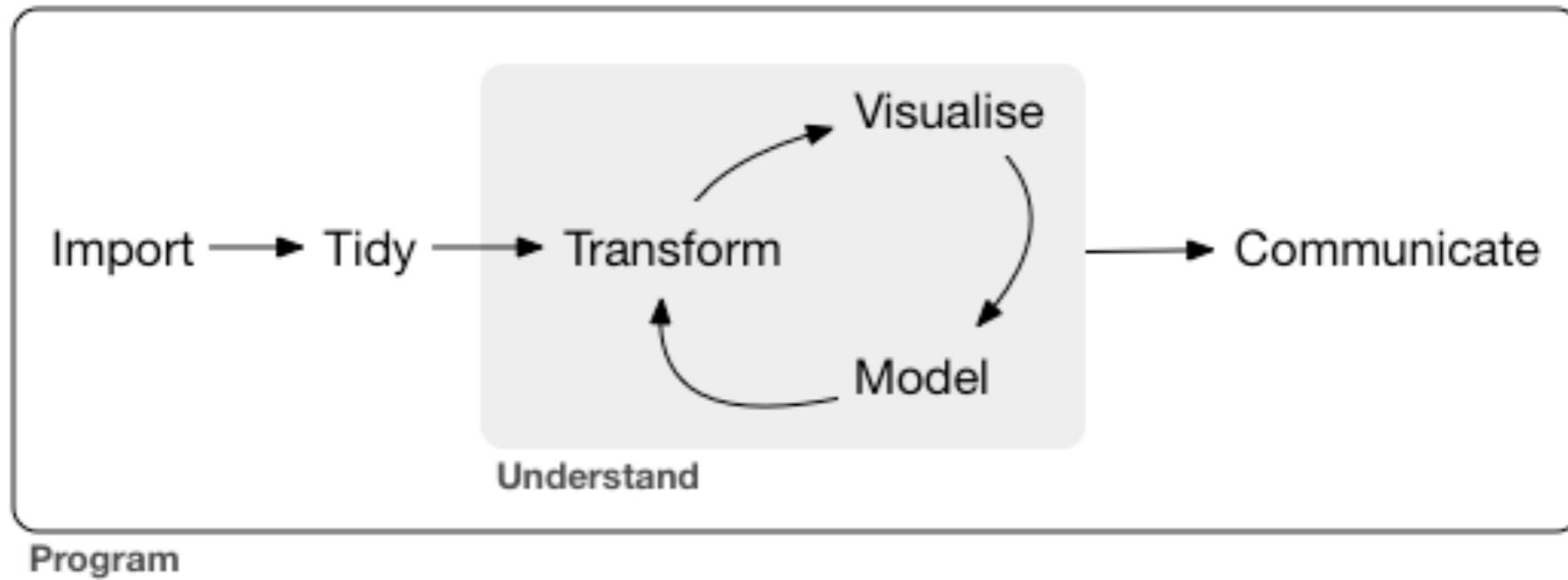
This is the website for “**R for Data Science**”. This book will teach you how to do data science with R: You’ll learn how to get your data into R, get it into the most useful structure, transform it, visualise it and model it. In this book, you will find a practicum of skills for data science. Just as a chemist learns how to clean test tubes and stock a lab, you’ll learn how to clean data and draw plots—and many other things besides. These are the skills that allow data science to happen, and here you will find the best practices for doing each of these things with R. You’ll learn how to use the grammar of graphics, literate programming, and reproducible research to save time. You’ll also learn how to manage cognitive resources to facilitate discoveries when wrangling, visualising, and exploring data.

This website is (and will always be) **free to use**, and is licensed under the [Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License](#). If you’d like a **physical copy** of the book, you can order it from [amazon](#); it was published by O’Reilly in January 2017. If you’d like to **give back** please make a donation to [Kākāpō Recovery](#): the kākāpō (which appears on the cover of R4DS) is a critically endangered native NZ parrot; there are only 148 left.

The book is written in [RMarkdown](#) with [bookdown](#). It is automatically rebuilt from [source](#) by [travis](#). R4DS is a collaborative effort and many people have contributed fixes and improvements via pull request.

<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

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 R? Also ...

instead of

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

(google search "complicated statistical equation")

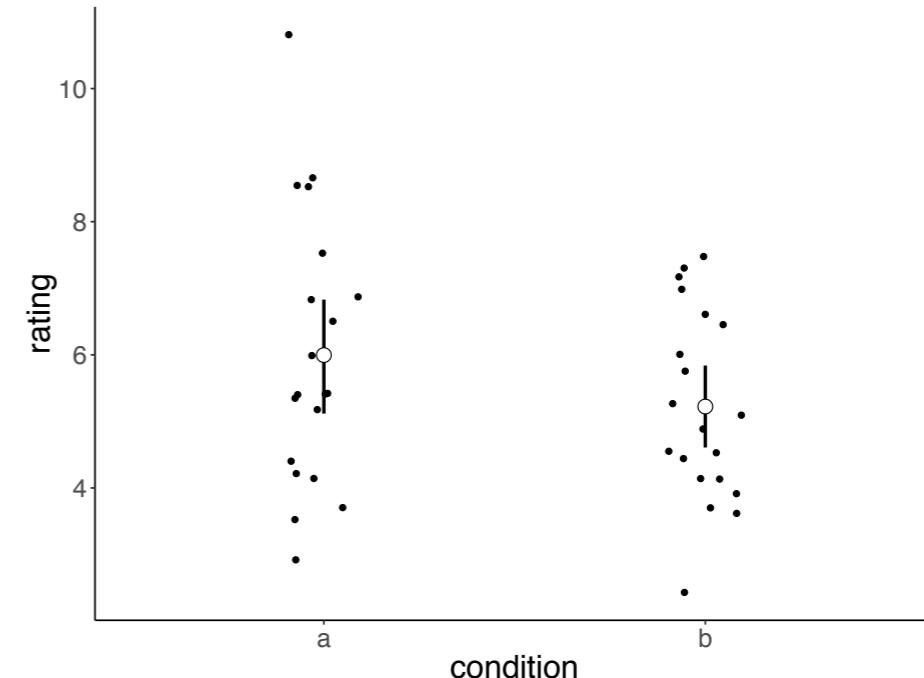
df	Level of Significance for Directional Test (t_{crit})					
	$\alpha=0.10$	$\alpha=0.05$	$\alpha=0.025$	$\alpha=0.01$	$\alpha=0.005$	$\alpha=0.0005$
Level of Significance for Non-directional Test (t_{crit})						
1	3.0780	6.3140	12.7100	31.8200	63.6600	636.6000
2	1.8860	2.9200	4.3030	6.9650	9.9250	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

```

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  -

```



But what about python ™ ?

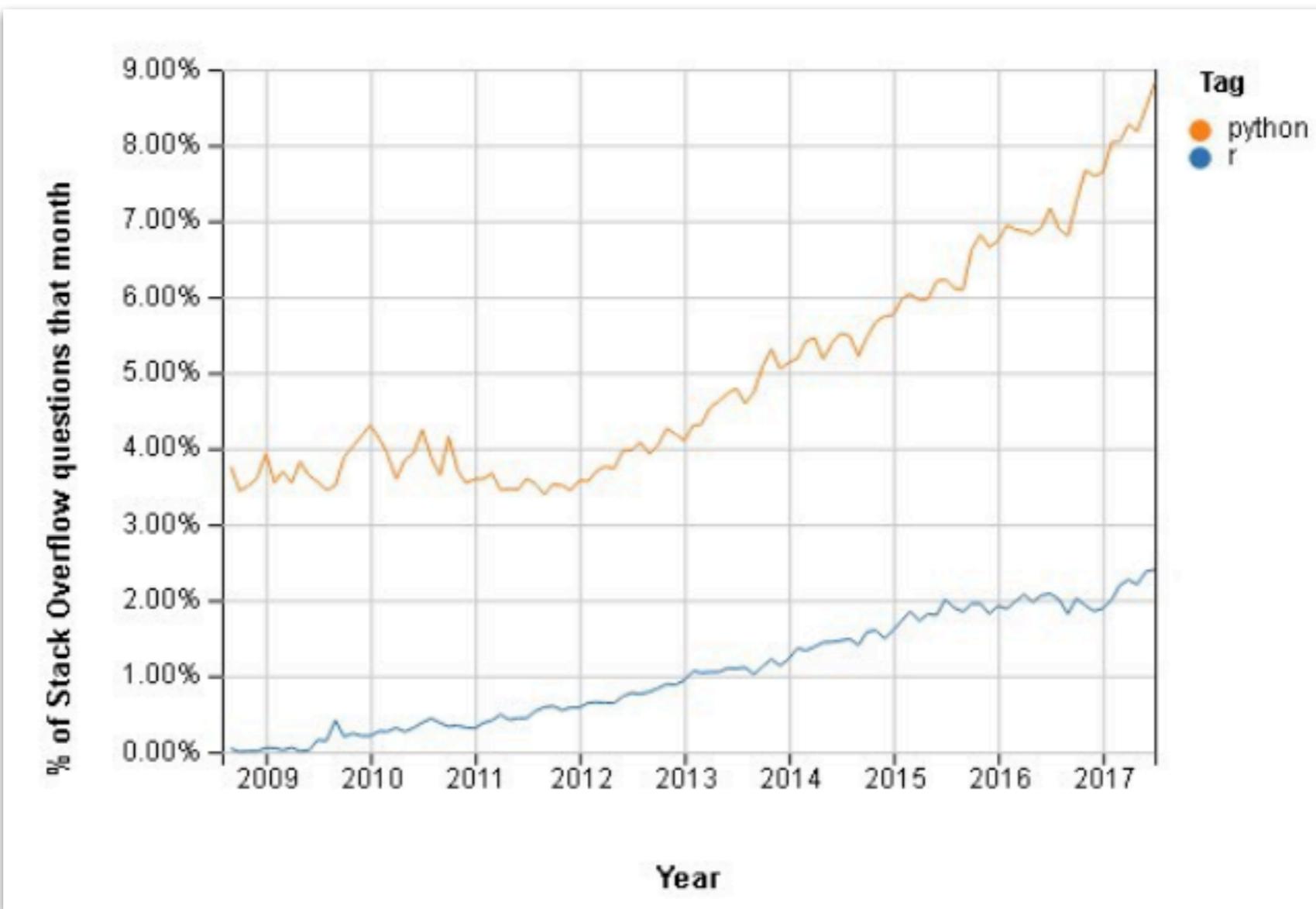
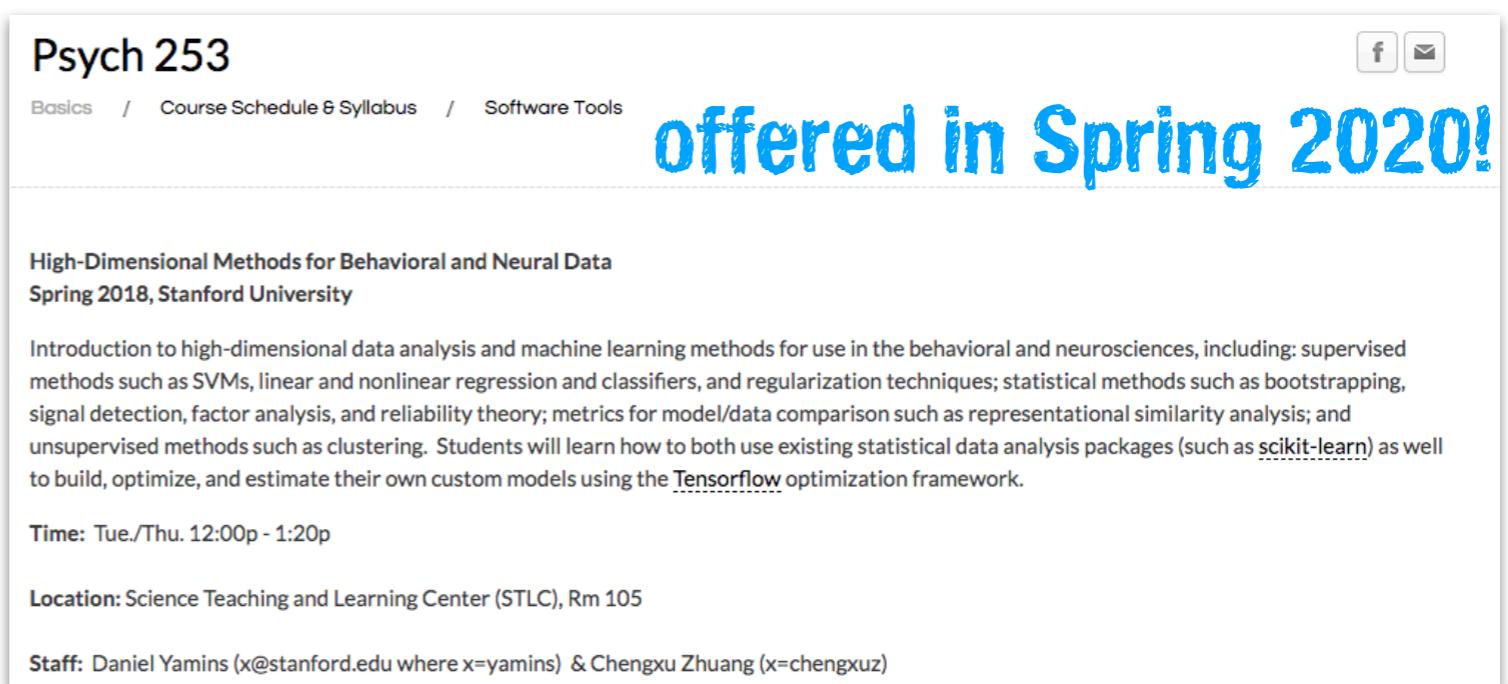


Image Source: <https://dzone.com/articles/r-or-python-data-scientists-delight>

But what about python™ ?

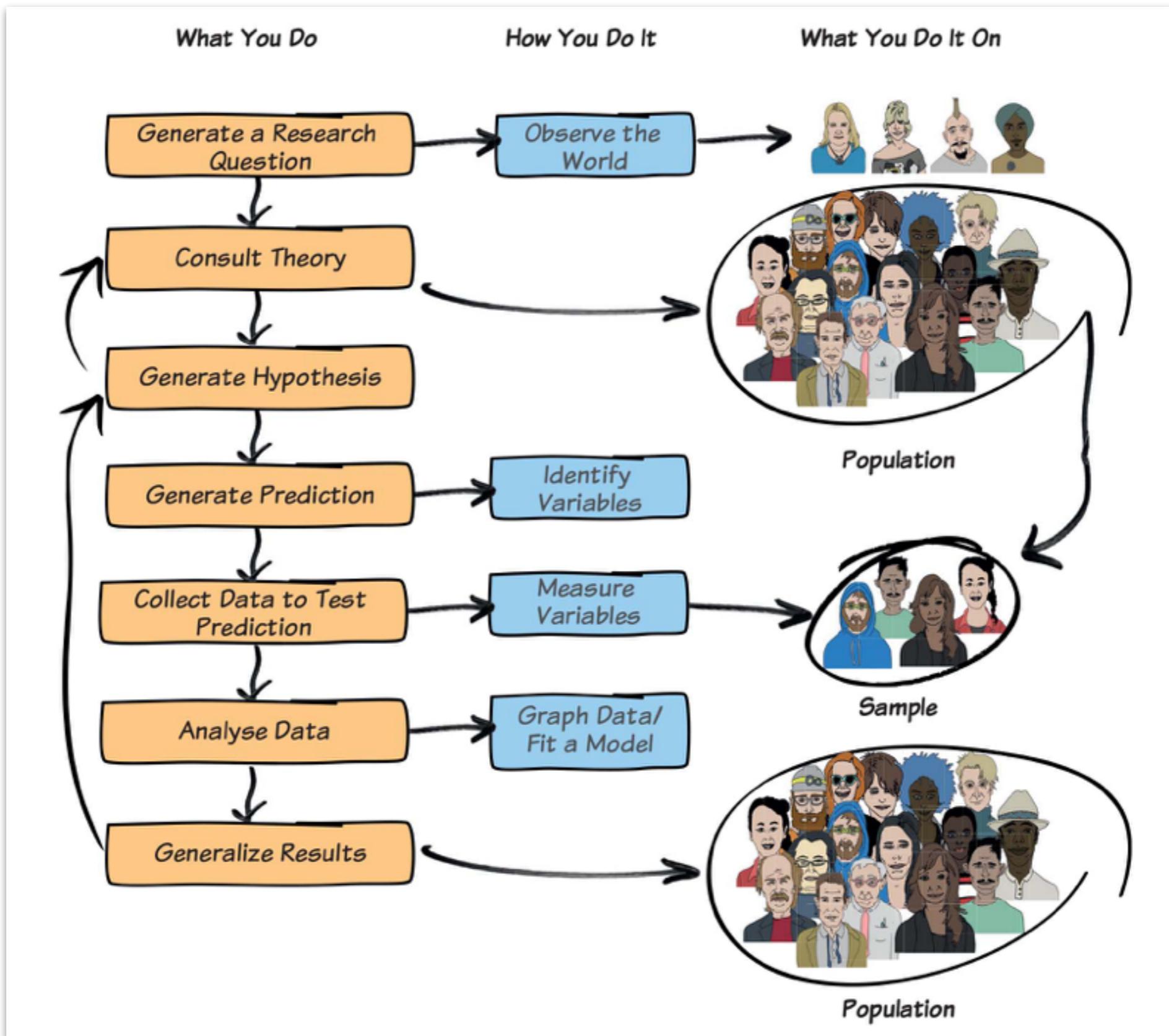
- You should learn python, too! :)
- Python is the more general purpose programming language.
- Python is great for machine learning (e.g. neural networks and deep learning).
- You can learn about all of these things in **Psych 253!**



The screenshot shows the homepage of the Psych 253 course website. At the top, there's a navigation bar with links for "Basics", "Course Schedule & Syllabus", and "Software Tools". To the right of the navigation are social media icons for Facebook and Email. A large blue banner across the top right reads "offered in Spring 2020!". Below the banner, the course title "Psych 253" is displayed, along with the subtitle "High-Dimensional Methods for Behavioral and Neural Data" and the text "Spring 2018, Stanford University". The main content area describes the course's focus on high-dimensional data analysis and machine learning methods for behavioral and neurosciences, mentioning supervised and unsupervised learning, regularization techniques, and clustering. It also notes the use of scikit-learn and Tensorflow. At the bottom, details about the course schedule ("Time: Tue./Thu. 12:00p - 1:20p"), location ("Location: Science Teaching and Learning Center (STLC), Rm 105"), and staff ("Staff: Daniel Yamins (x@stanford.edu where x=yamins) & Chengxu Zhuang (x=chengxuz)") are provided.

Build models

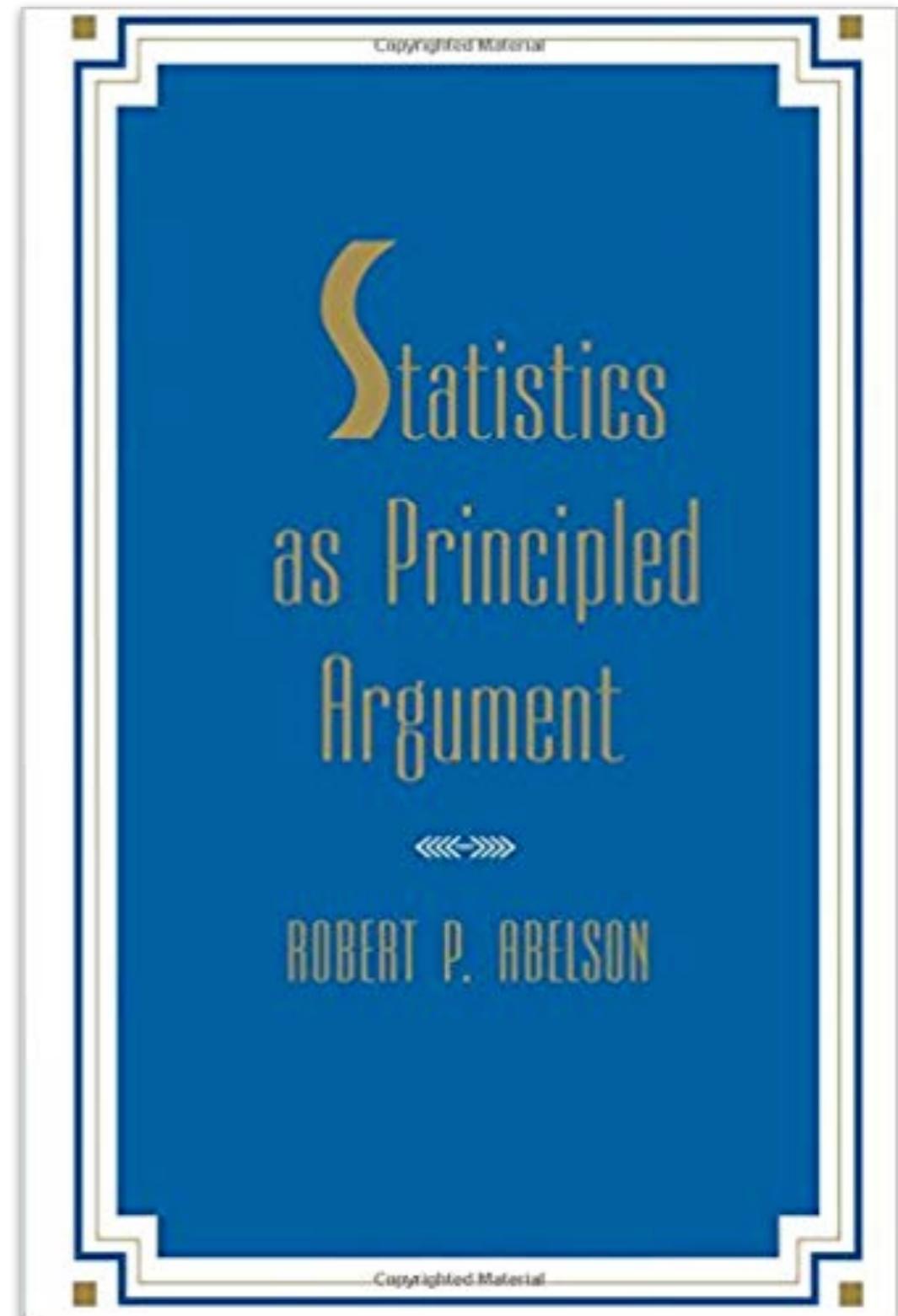
The research process



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, ...

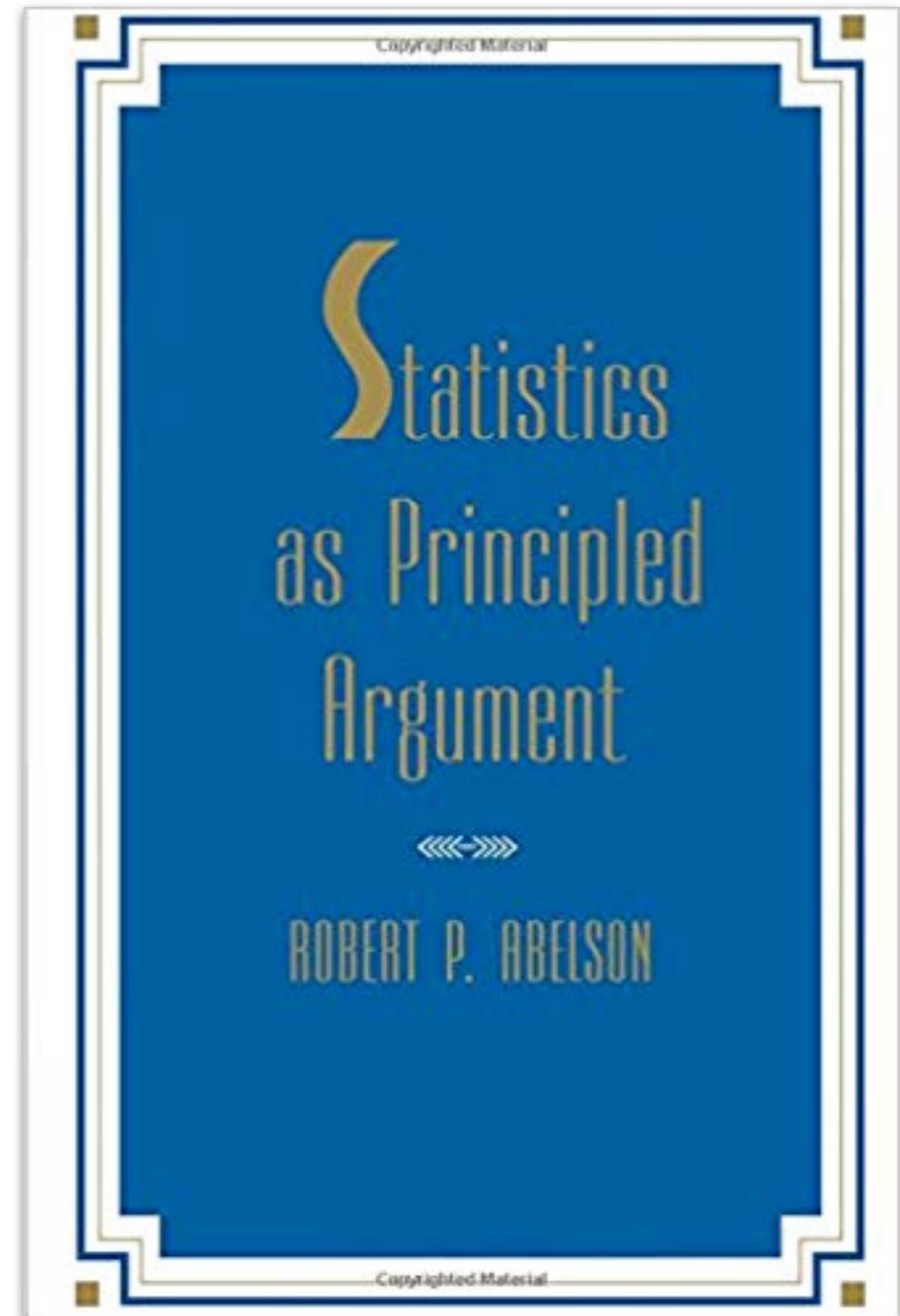


Abelson, R. P. (1995). *Statistics as principled argument*. New York: Psychology Press.

Statistics as argument

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

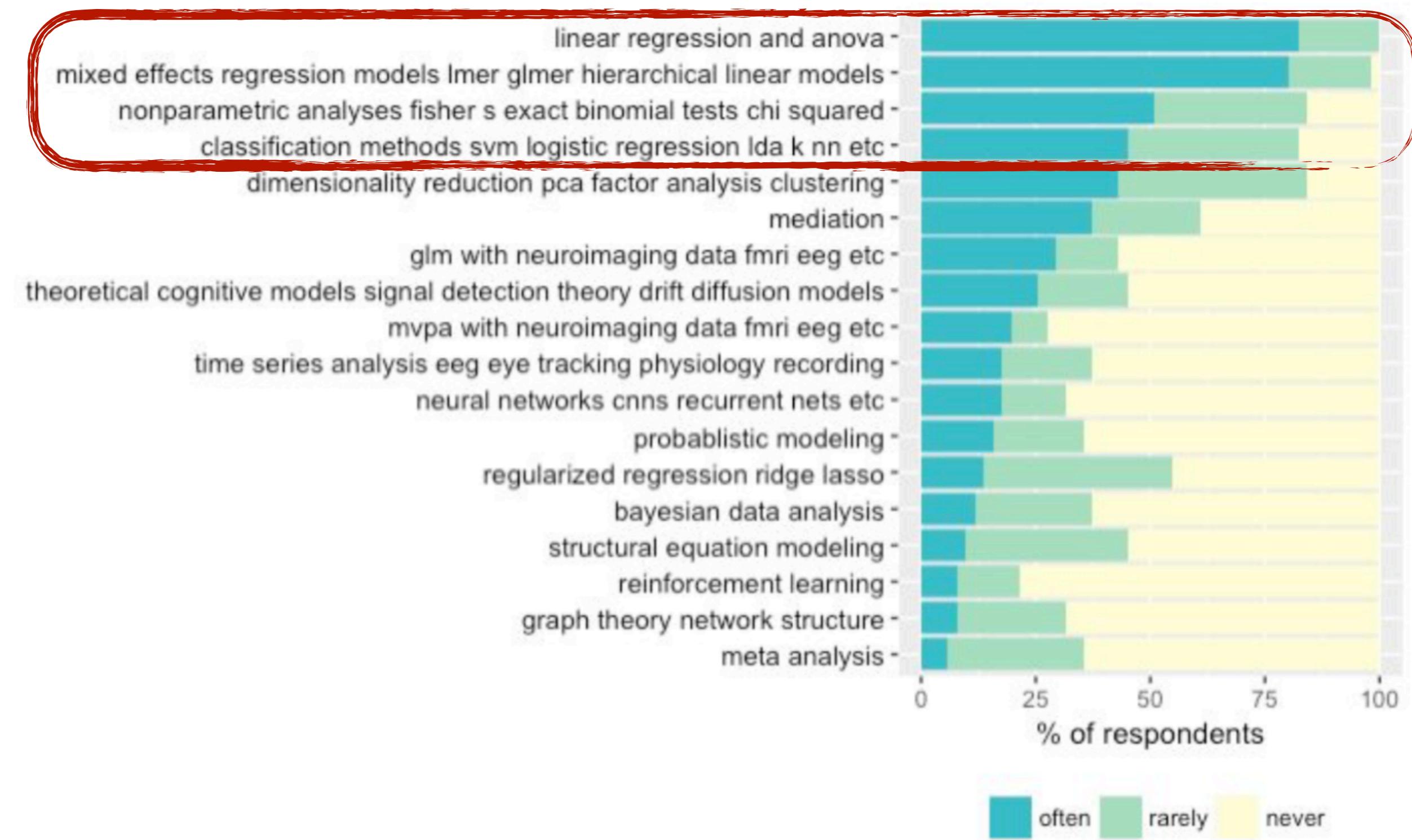
- **Magnitude**: Larger effects are more persuasive
- **Articulation**: Understandability of results
- **Generality**: How much does the effect depend on contextual factors
- **Interestingness**: Change what people believe about an important issue
- **Credibility**: Believability of a research claim,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



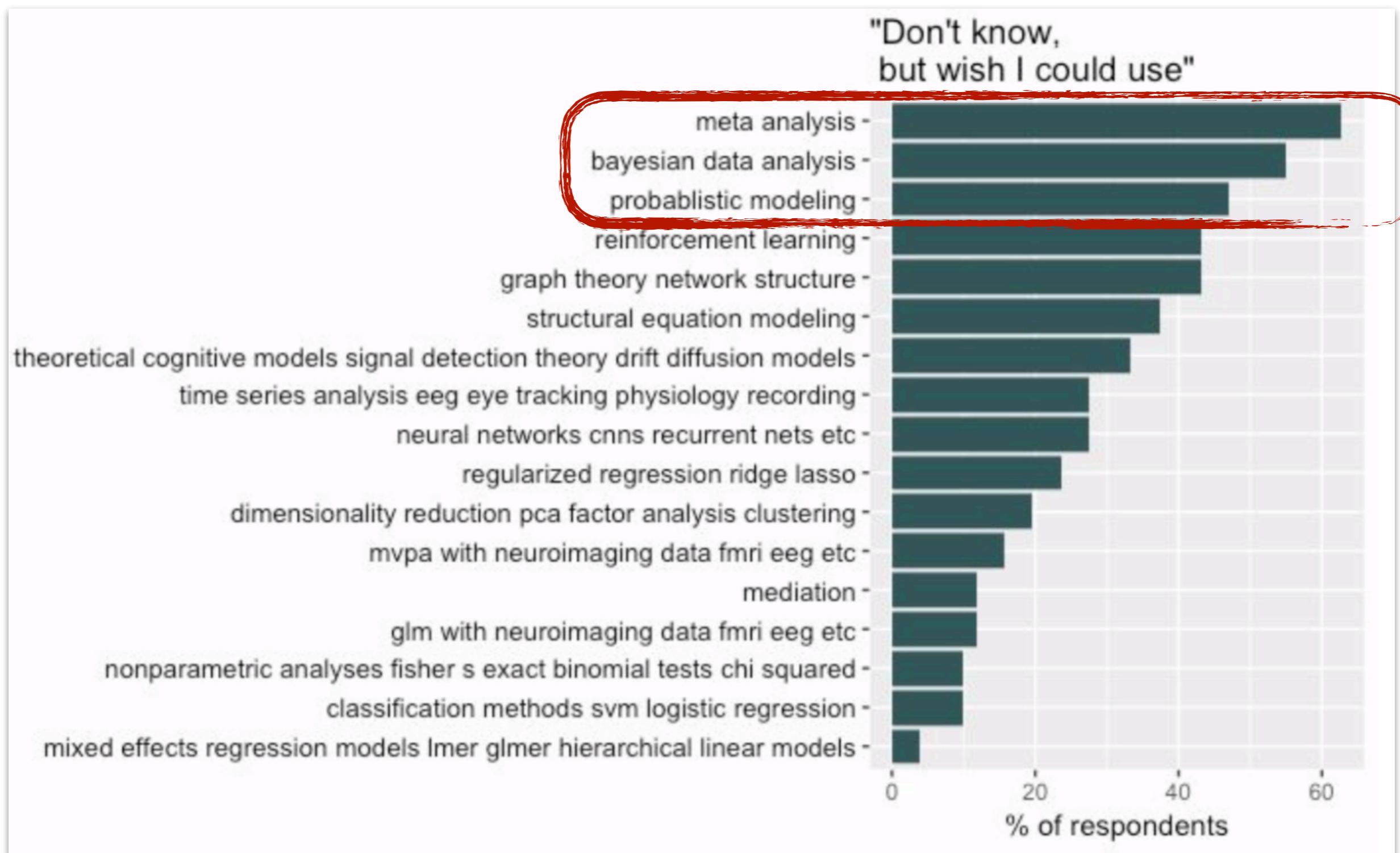
Abelson, R. P. (1995). *Statistics as principled argument*. New York: Psychology Press.

Psychology grad student survey

“How often do you **use** the following skills?”



Psychology grad student survey



Report results

Replication crisis

Google replication crisis in psychology

All News Images Videos Shopping More Settings Tools SafeSearch ▾

precis reproducibility crisis scientific estimating science social science psych nature studies recommendations social psychology psychological science research randall monroe stapel

Economics (n=18) Psychology (n=100)
Camerer et al., 2016 Open Science Collaboration, 2015

Failed to replicate Meta analytic effect p<0.05 Replication p<0.05

Journal % Findings Replicated
Journal of Personality and Social Psychology: Social 23
Journal of Experimental Psychology: Learning, Memory, and Cognition 48
Psychological Science, social articles 29
Psychological Science, cognitive articles 53
Overall 36

Effect size reduction Independent replication p-value

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

THE CHRONICLE OF HIGHER EDUCATION How to Fix Psychology's Replication Crisis

Tradition of controlled experiments No Yes
No Epidemiology Clinical Sciences CRISIS!
Yes Strong basic theory Astrology Particle physics

nature How Reliable Psychology S Massive International Project Raises Questions about the Validity of Psychology Research

Science Only 36% of studies replicated!!

Replication Effect Size P-value Non-Significant Replication Power

psychological science ... science.sciencemag.org

Replication Crisis Overblown ... thecut.com

Health News Digest ... bfm.my

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

6 Principles of Open Science

- Open Data
- Open Source
- Open Access
- Open Methodology
- Open Peer Review
- Open Educational Resources

Lower integrity Higher integrity
Weak science Strongest science
Exact replications Conceptual replications

mean items solved (adjusted by SAT)

RACE PRIME NO RACE PRIME

■ BLACK SUBJECTS □ WHITE SUBJECTS

PSYCHOLOGY'S REPRODUCIBILITY PROBLEM

LARGE SCALE REPPLICATION STUDIES like the one performed by Rutter and Niles have begun to reveal that many of the published results in psychology are not replicable, making it difficult to confirm.

A B

Study Field Replication rate Effect size quotient N (samples) Preregistered Target selection

Open Science Collaboration (OSCI), 2015	Psychology	36%	0.49	100	No	Quasirandom
Klein, Ratliff, and Vianello, 2014	Psychology	77%	1.26	36	Yes	Nonrandom
Special Issue contributions to <i>Social Psychology</i> vol. 45(3), 2014	Psychology	8%	—	13	Yes	Nonrandom
Camerer et al., 2016	Economics	61%	0.66	18	No	Nonrandom

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

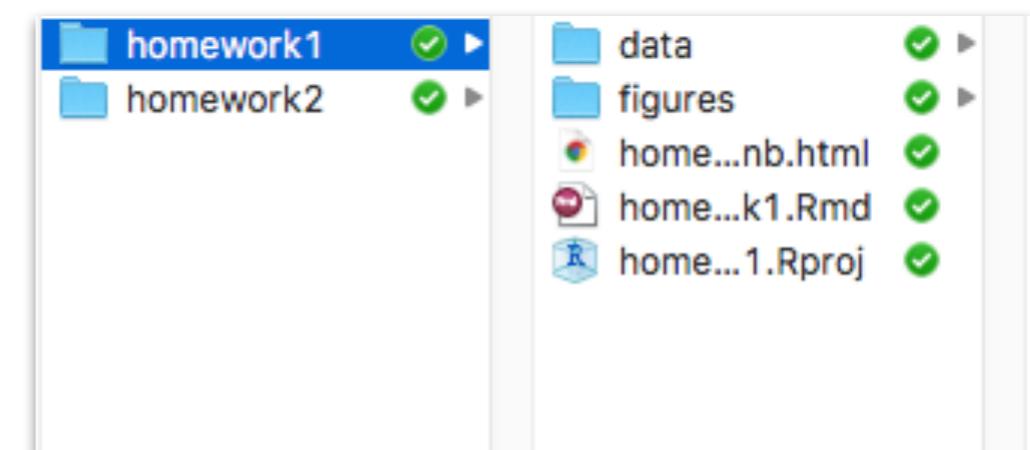
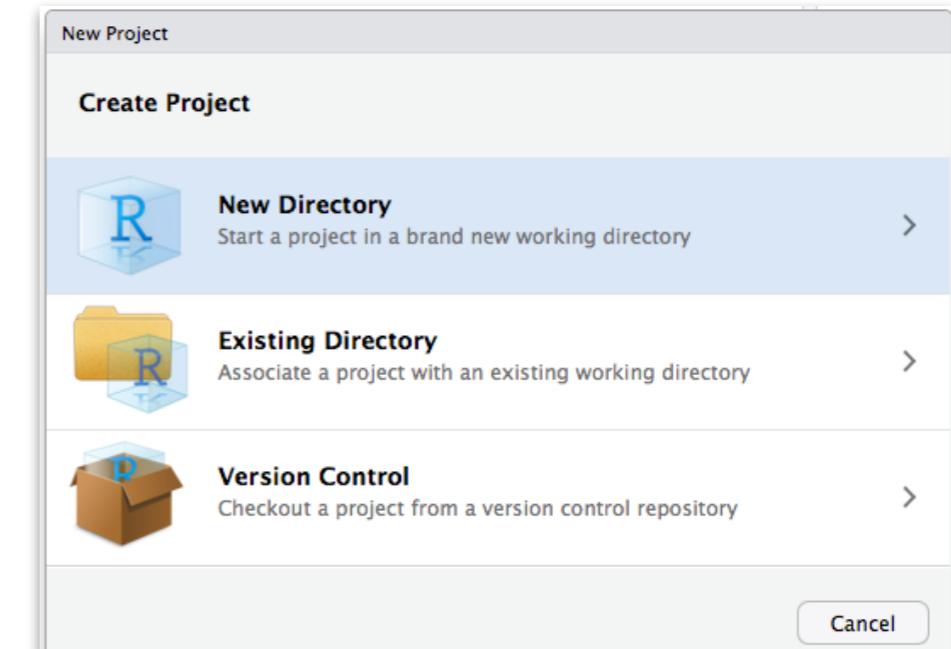
- **Organize** the folder structure
- **Use relative paths**



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



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



R Markdown

R Markdown :: CHEAT SHEET

What is R Markdown?

.Rmd files - An R Markdown (.Rmd) file is a record of your research. It contains the code that a scientist needs to reproduce your work along with the narration that a reader needs to understand your work.

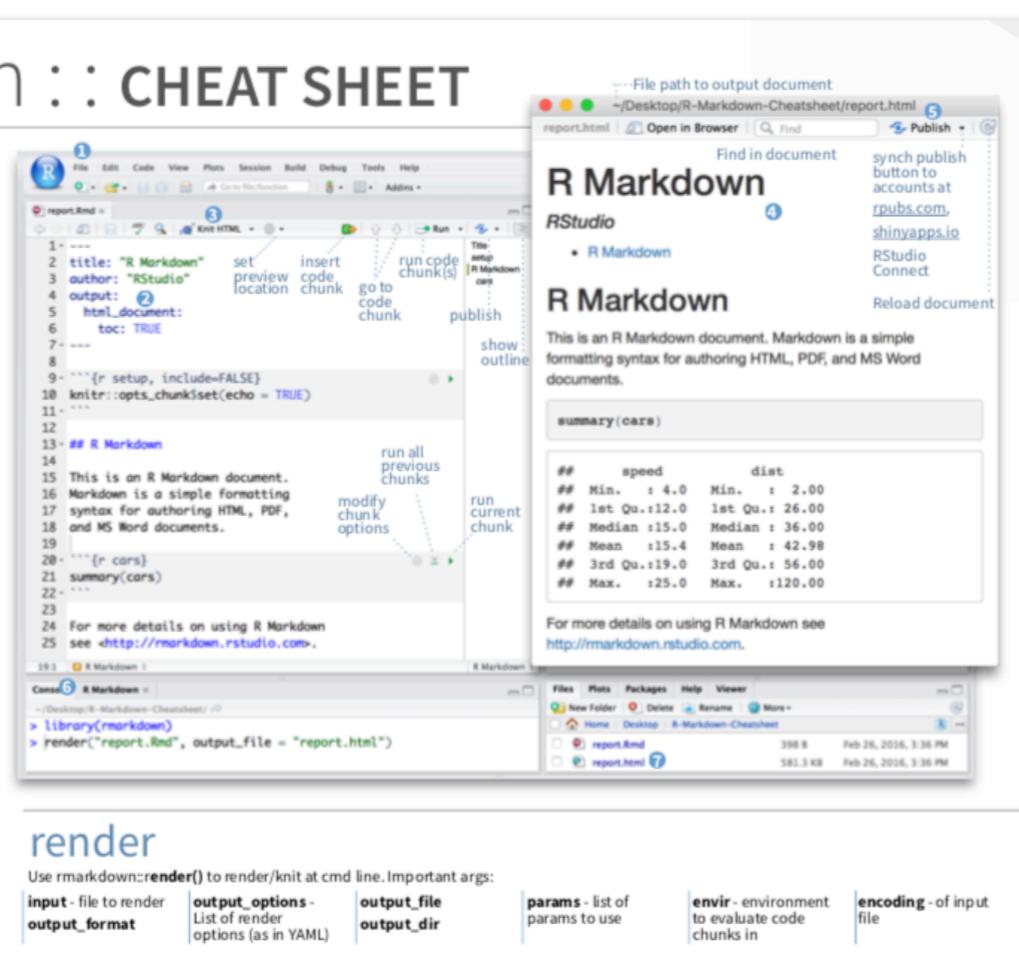
Reproducible Research - At the click of a button, or the type of a command, you can rerun the code in an R Markdown file to reproduce your work and export the results as a finished report.

Dynamic Documents - You can choose to export the finished report in a variety of formats, including html, pdf, MS Word, or RTF documents; html or pdf based slides, Notebooks, and more.

Workflow



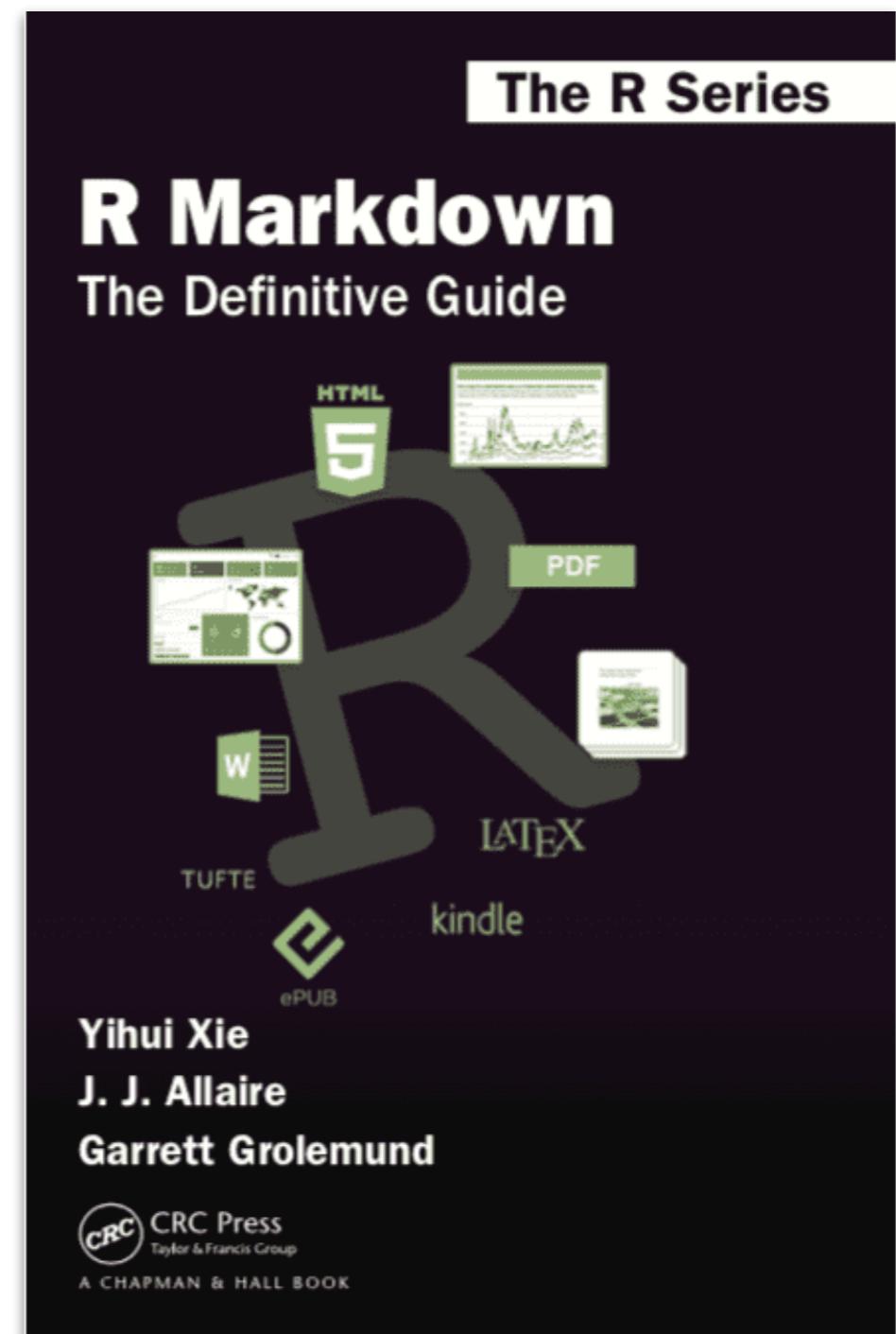
- ① Open a new .Rmd file at File ▶ New File ▶ R Markdown. Use the wizard that opens to pre-populate the file with a template
- ② Write document by editing template
- ③ Knit document to create report; use knit button or render() to knit
- ④ Preview Output in IDE window
- ⑤ Publish (optional) to web server
- ⑥ Examine build log in R Markdown console
- ⑦ Use output file that is saved along side.Rmd



The screenshot shows the RStudio interface with the following components:

- File Path:** ~/Desktop/R-Markdown-Cheatsheet/report.html
- Toolbar:** File, Edit, Code, View, Plots, Session, Build, Debug, Tools, Help
- Code Editor:** A .Rmd file titled "report.Rmd" containing R code and comments.
- IDE Window:** Displays the rendered output of the R code, showing the summary of the "cars" dataset.
- Console:** Shows the command `library(rmarkdown)` and `render("report.Rmd", output_file = "report.html")` being run.
- File Explorer:** Shows the file structure with "report.Rmd" and "report.html".
- Help:** A sidebar with links for "synch publish button to accounts at rpubs.com, shinyapps.io" and "RStudio Connect".

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



The R Series

R Markdown

The Definitive Guide

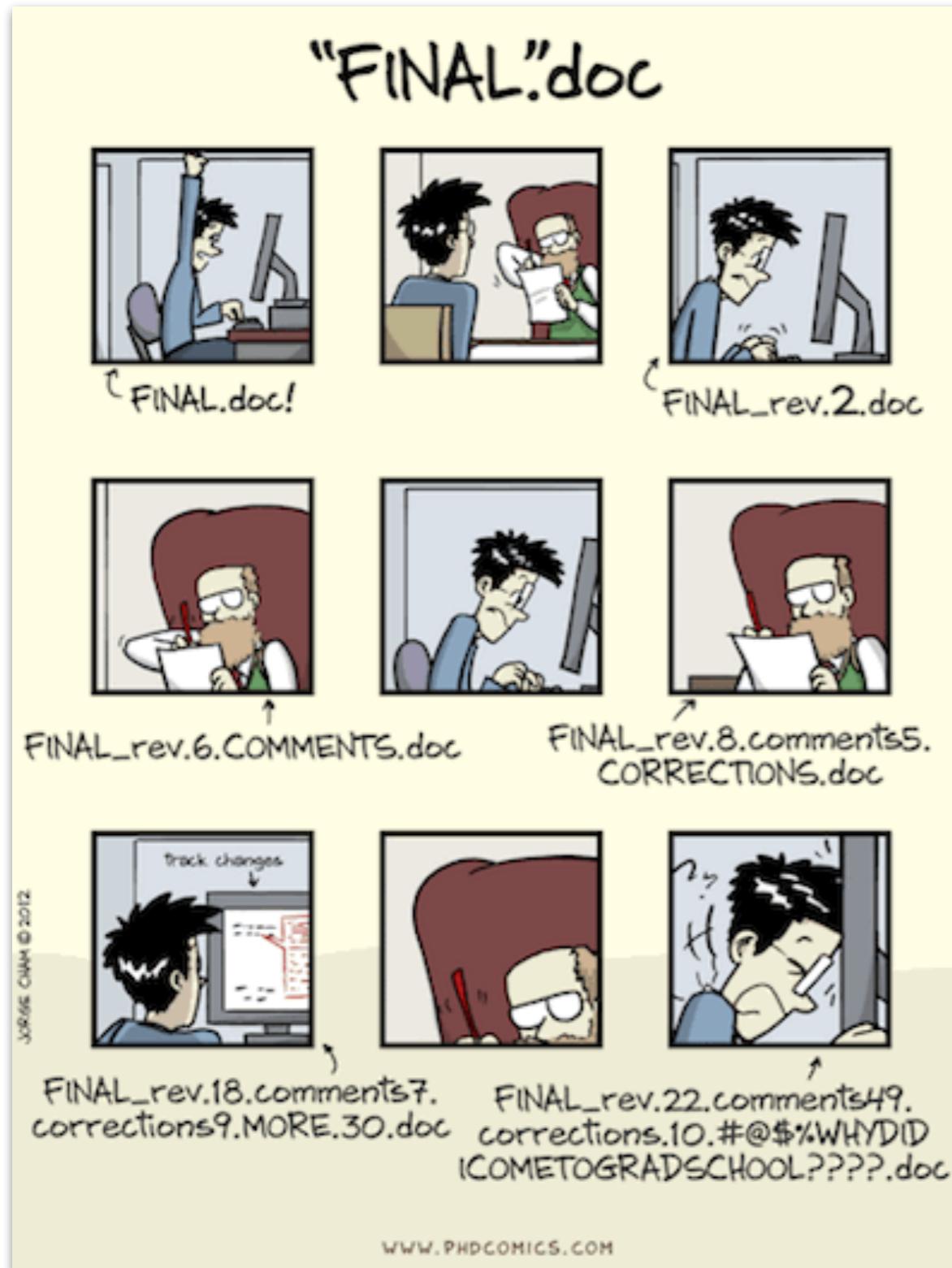
Yihui Xie
J. J. Allaire
Garrett Grolemund

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

HTML 5
PDF
W
TUFTE
ePUB
LATEX
kindle

<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?

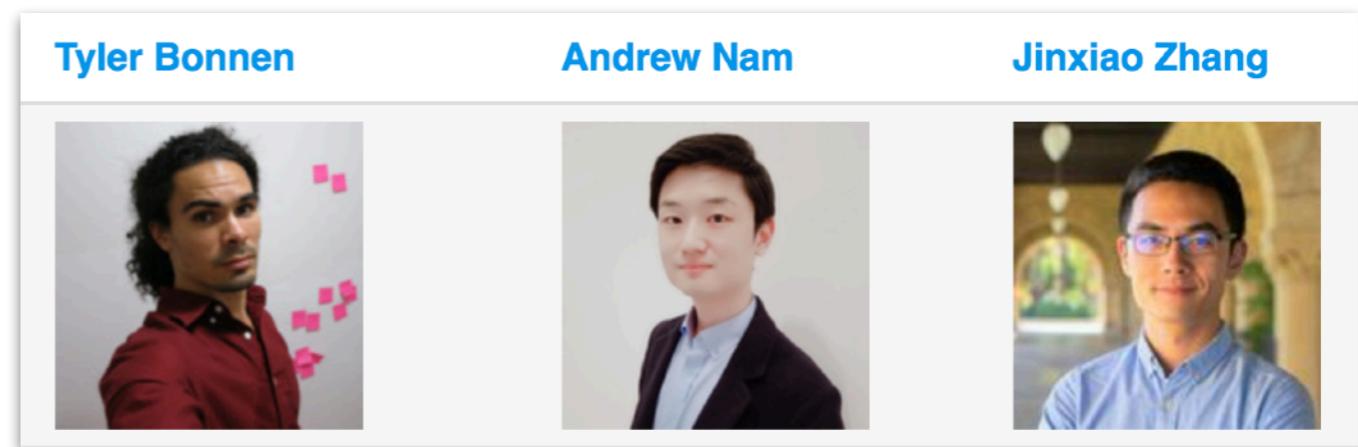
- Lectures
- Sections
- Homework
- Midterm exam
- Final project
- Grading
- Office hours

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. short written project proposal

- 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. oral presentation

- 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?

- **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%

late submissions
count as 0

How will we learn?

- **Office hours**

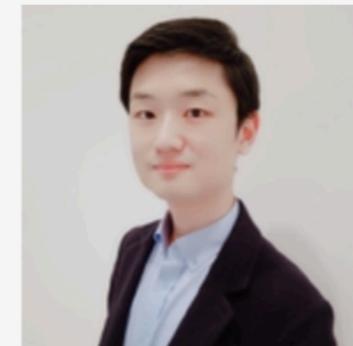
Tobi Gerstenberg



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	<ul style="list-style-type: none">• Course introduction
				<ul style="list-style-type: none">• Cheatsheet R Studio• Cheatsheet R Markdown 1• Cheatsheet R Markdown 2• R Markdown for class reports
				<ul style="list-style-type: none">• Introduction to R• RStudio IDE 1• RStudio IDE 2• RMarkdown
				<ul style="list-style-type: none">• ggplot part 1• ggplot part 2• Reporting
	W	9-Jan	Visualization I	<ul style="list-style-type: none">• Best practices• Introduction to RStudio• Introduction to <code>library(ggplot2)</code>• Reporting results using Rmarkdown
				<ul style="list-style-type: none">• Data visualization (#1)• Data visualization (#3)• Cheatsheet ggplot2



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

PSYCH 252: STATISTICAL METHODS

Home Schedule Getting ready Information **Book**



Psych 252

Preface

- Course description
- Course homepage
- 1 Introduction
- 2 Visualization 1
- 3 Visualization 2
- 4 Data wrangling 1
- 5 Data wrangling 2
- 6 Probability and causality
- 7 Simulation 1
- 8 Simulation 2
- 9 Modeling data
- 10 Linear model 1
- 11 Linear model 2
- 12 Linear model 3
- 13 Linear model 4
- 14 Power analysis
- 15 Bootstrapping
- 16 Model comparison
- 17 Linear mixed effects models 1
- 18 Linear mixed effects models 2
- 19 Linear mixed effects models 3
- 20 Generalized linear model

<https://psych252.github.io/psych252book/>

Psych 252: Statistical Methods for Behavioral and Social Sciences

Tobias Gerstenberg
2019-11-19

Preface

This book contains the course notes for [Psych 252](#). The book is not intended to be self-explanatory and instead should be used in combination with the course lectures.

If you have any questions about the notes, please feel free to contact me at: gerstenberg@stanford.edu or post an issue on the book's [github repository](#).

Course description

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.

<https://github.com/psych252/psych252book>

Tools we will use in class

Piazza

I have a question about the homework.

Datacamp

I would like to learn more about a topic.

PollEverywhere

I have an idea how to make this better!

Canvas

I would like to submit my homework on time.



Course website

I don't remember the schedule.

Office hours

I'd like to discuss something in person.

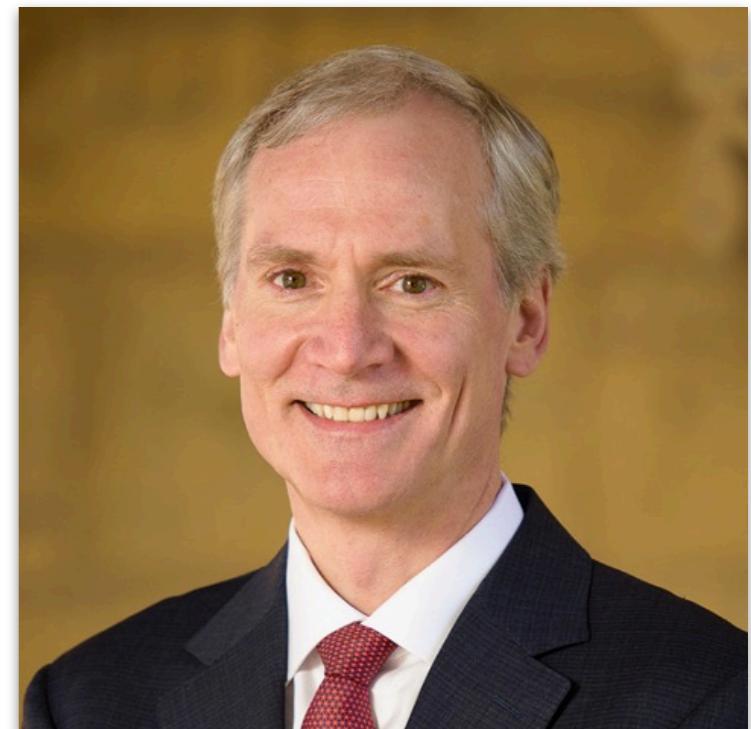
Anonymous feedback form

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



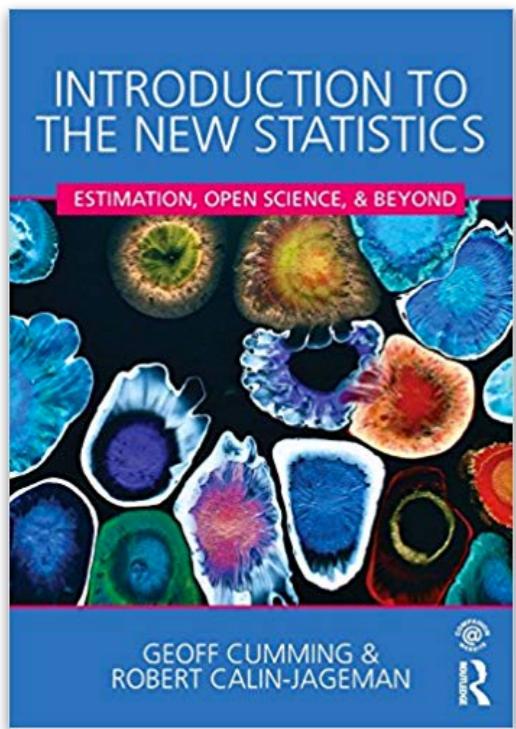
“an inspired, inclusive and collaborative community of diverse scholars, students and staff, where all are supported and empowered to thrive.”

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



Change a Fixed Mindset to a Growth Mindset

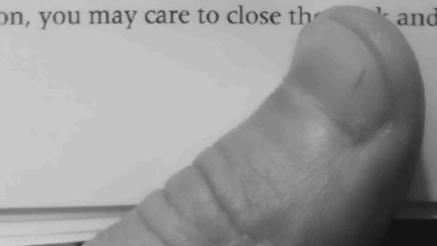
A further key idea is the distinction between a *fixed mindset* and a *growth mindset*. Carol Dweck and colleagues have demonstrated that helping students adopt a growth mindset can be a highly effective way to help them learn better and achieve more. Here's how Dweck describes the two mindsets:

Fixed mindset:
The belief that my capabilities are more or less fixed, whatever I do.

Growth mindset:
The belief that effort, persistence, and using good techniques can help me learn more successfully and become more capable.

In a fixed mindset students believe their basic abilities, their intelligence, their talents, are just fixed traits. They have a certain amount and that's that.... In a growth mindset students understand that their talents and abilities can be developed through effort, good teaching and persistence. They don't necessarily think everyone's the same or anyone can be Einstein, but they believe everyone can get smarter if they work at it. (Carol Dweck, tiny.cc/dwecktalk)

I've mentioned three important ideas about learning.
...before reading on, you may care to close the book and practice retrieval...



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

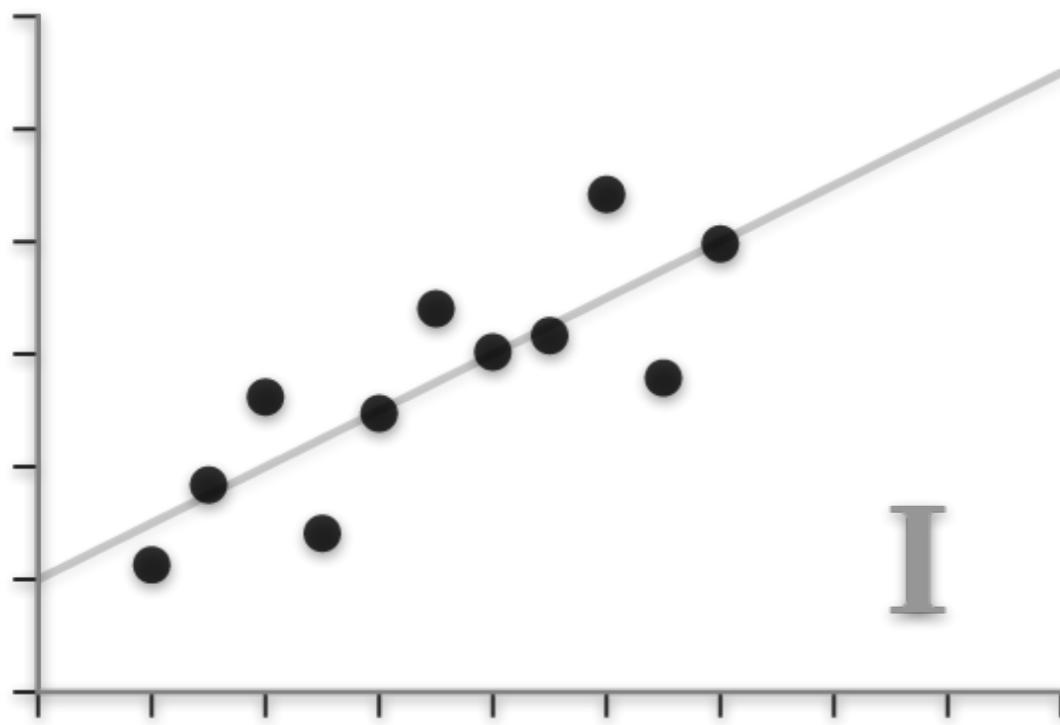
Visualization

some inspirational quotes ...

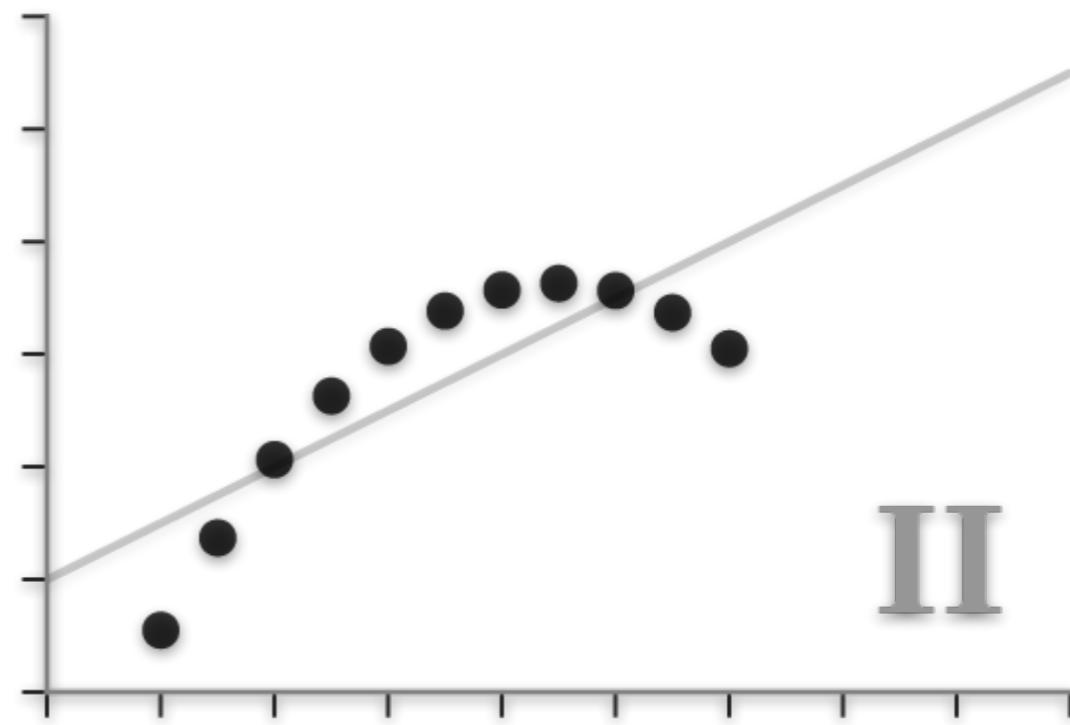
The greatest value of a picture is when it forces us to notice what we never expected to see. — John Tukey

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

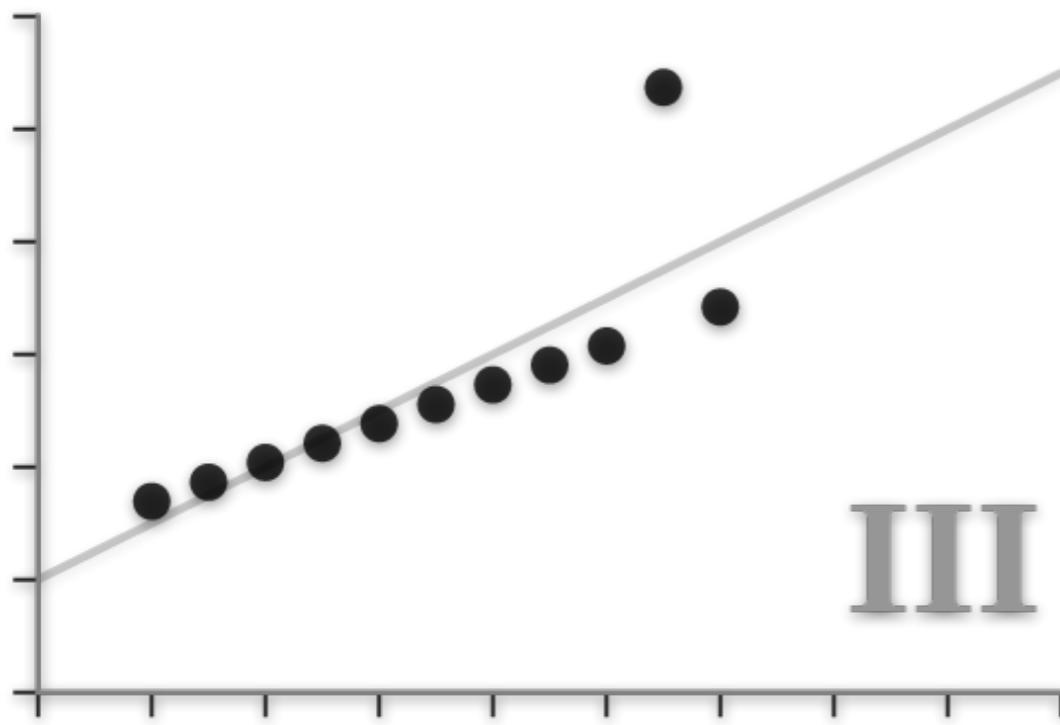
...make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding. — Anscombe (1973)



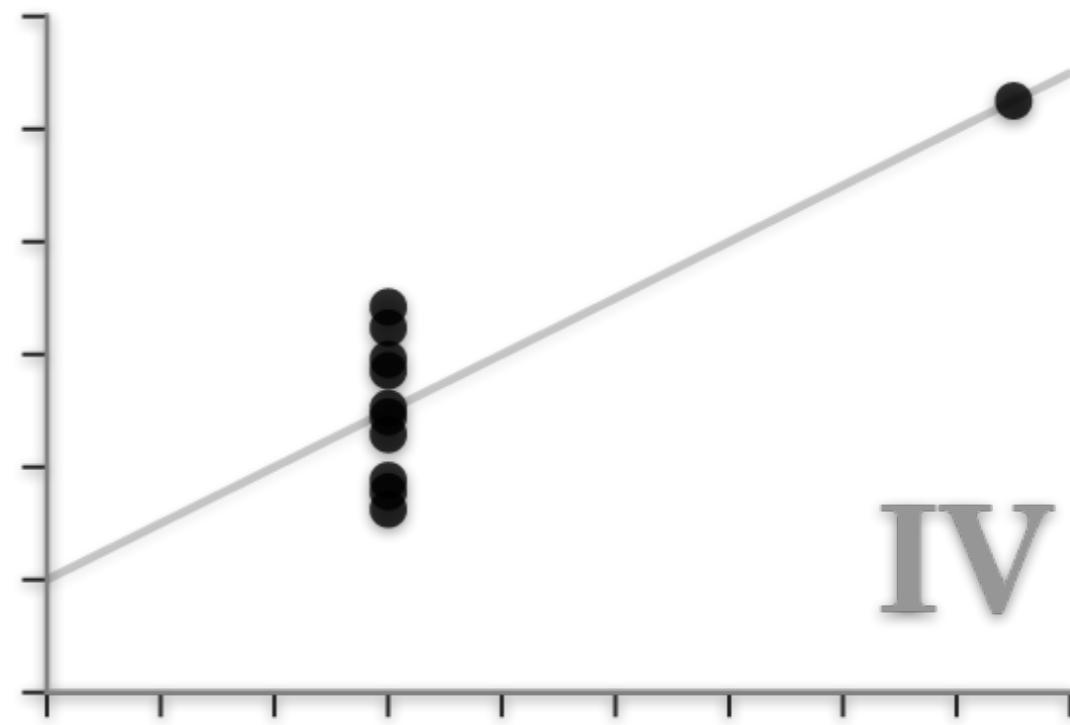
I



II



III

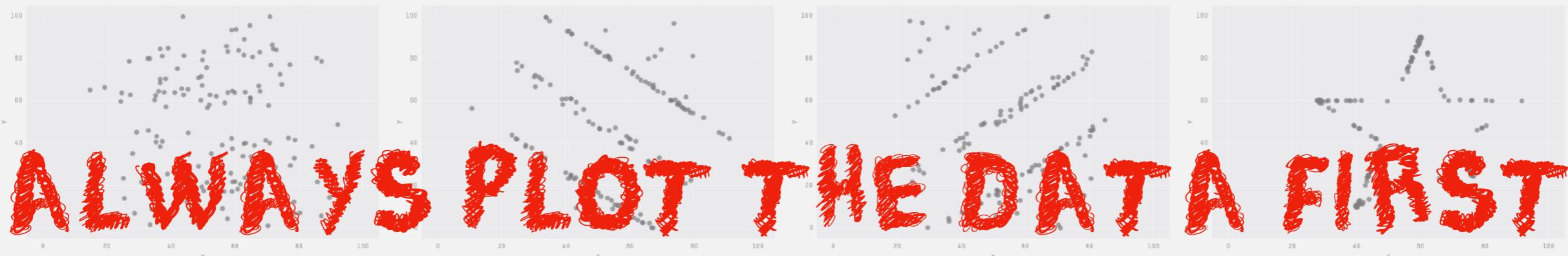


IV

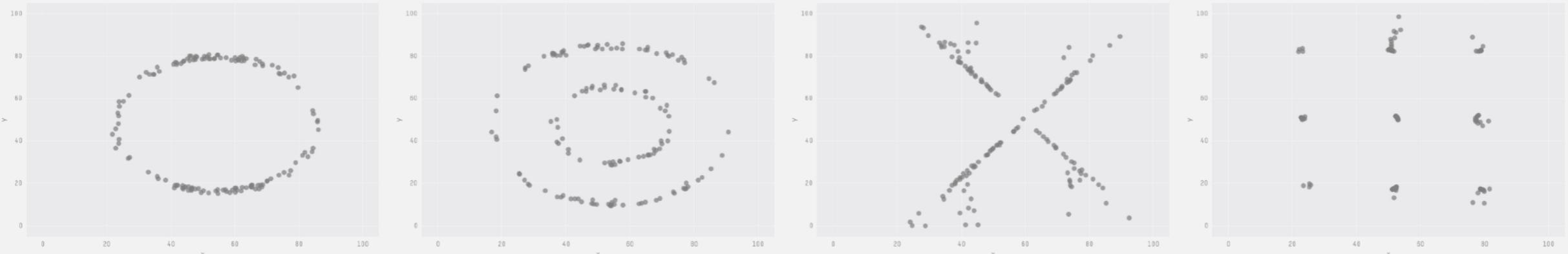
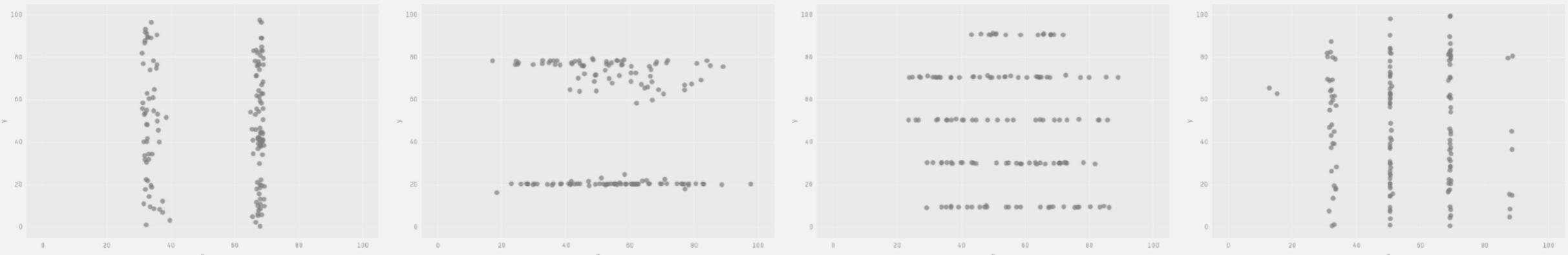
From Anscombe to T-Rex



X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06



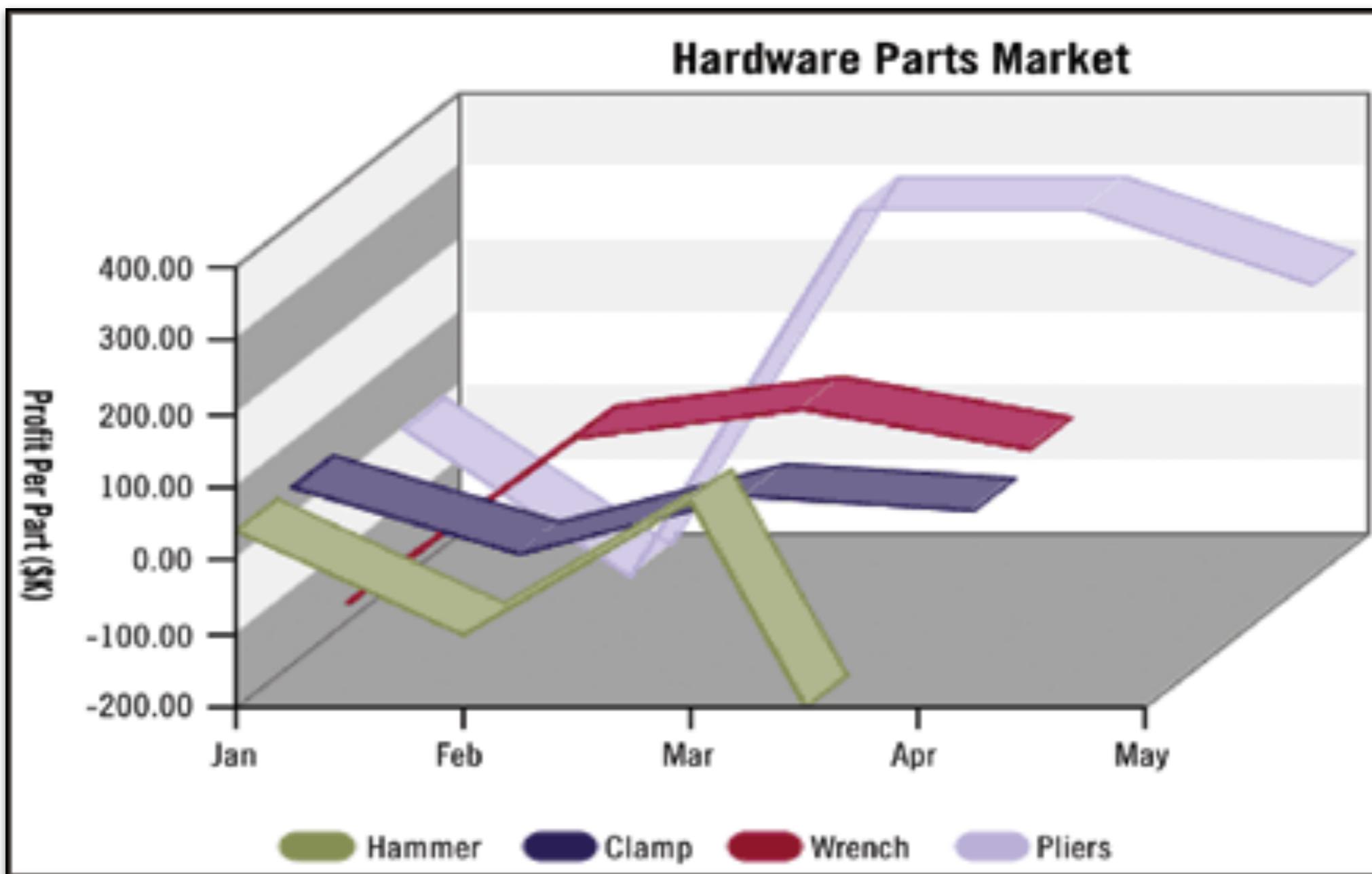
ALWAYS PLOT THE DATA FIRST



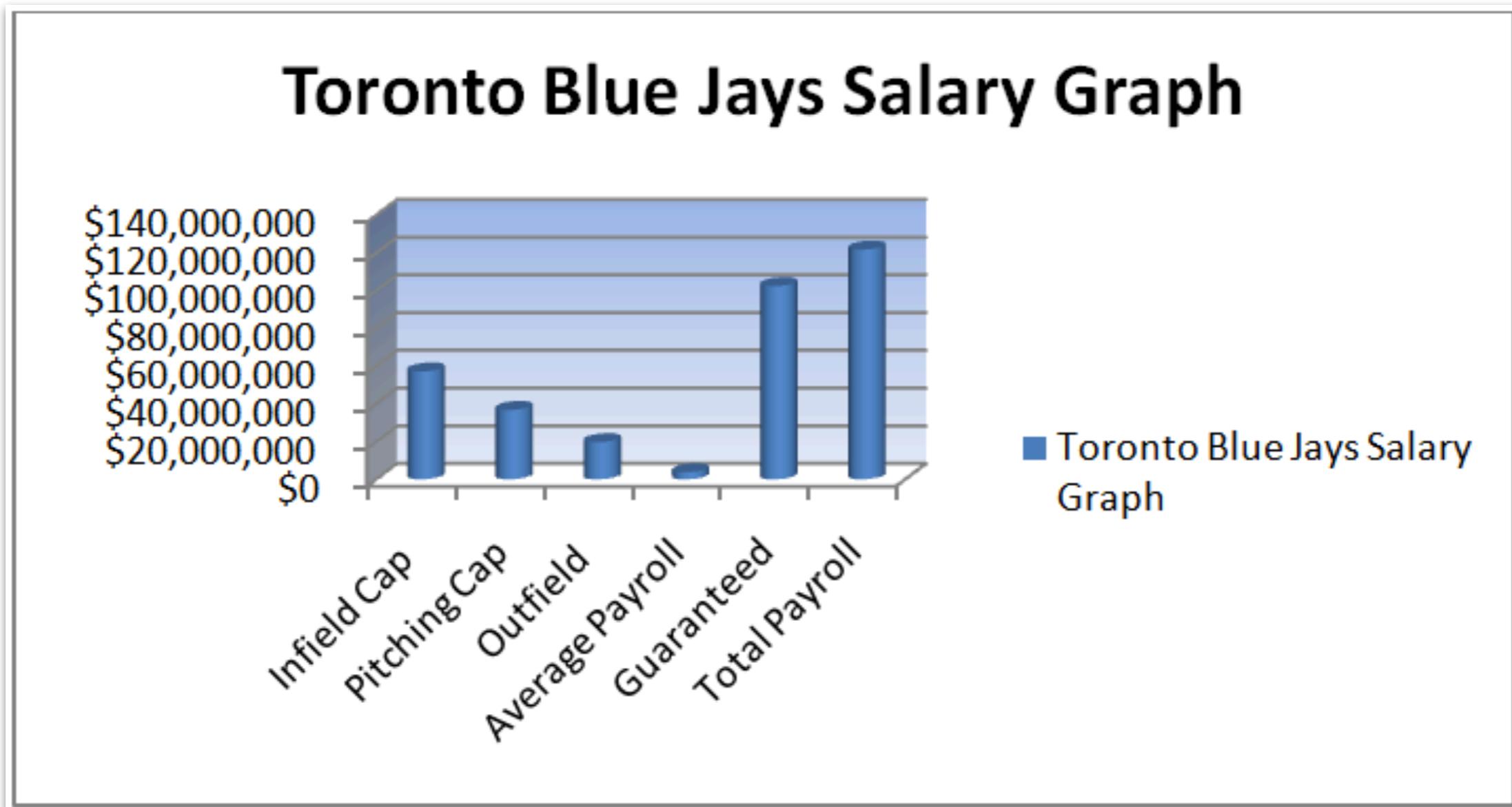
Plots that can be improved ...



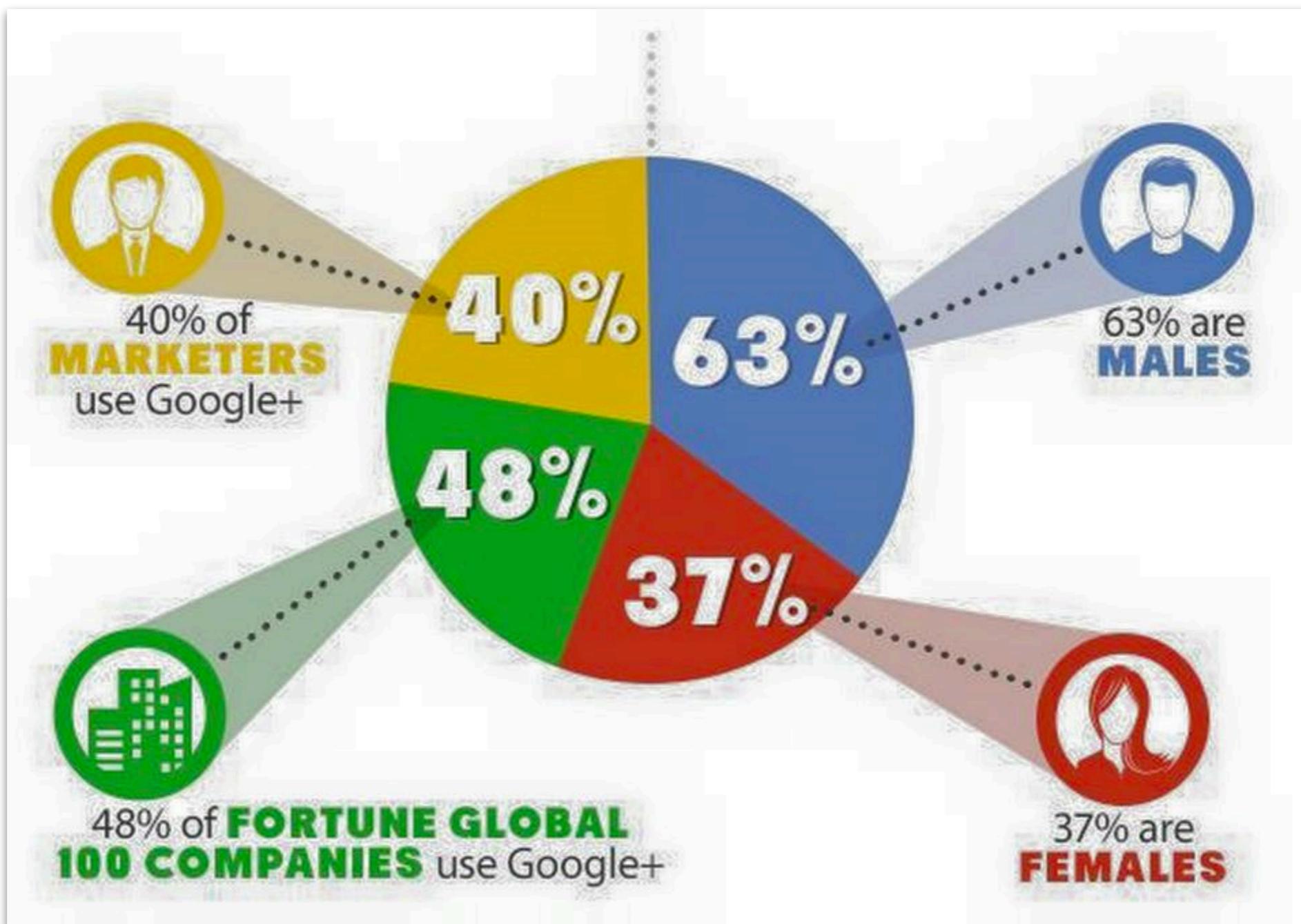
Plots that can be improved ...



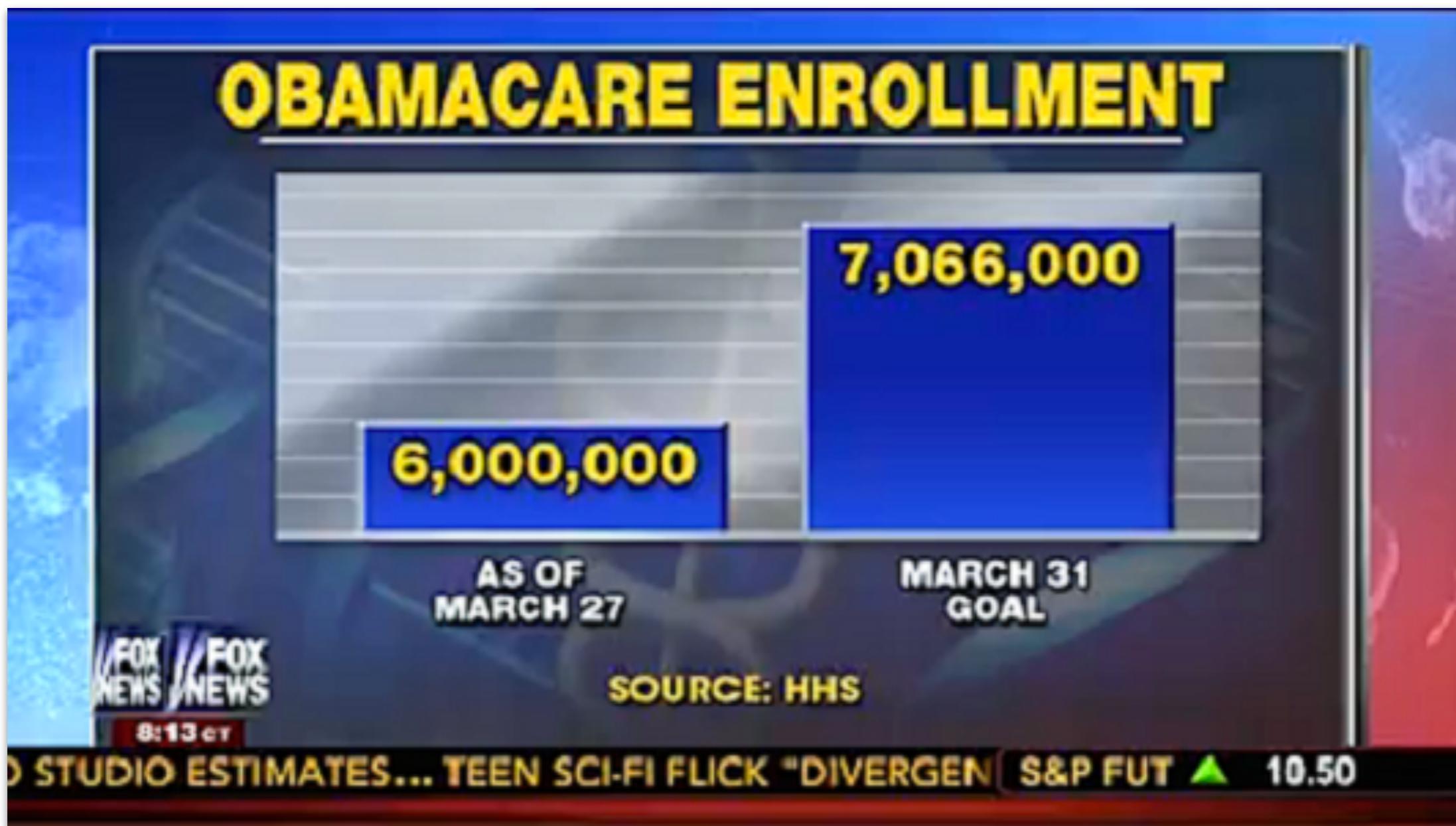
Plots that can be improved ...



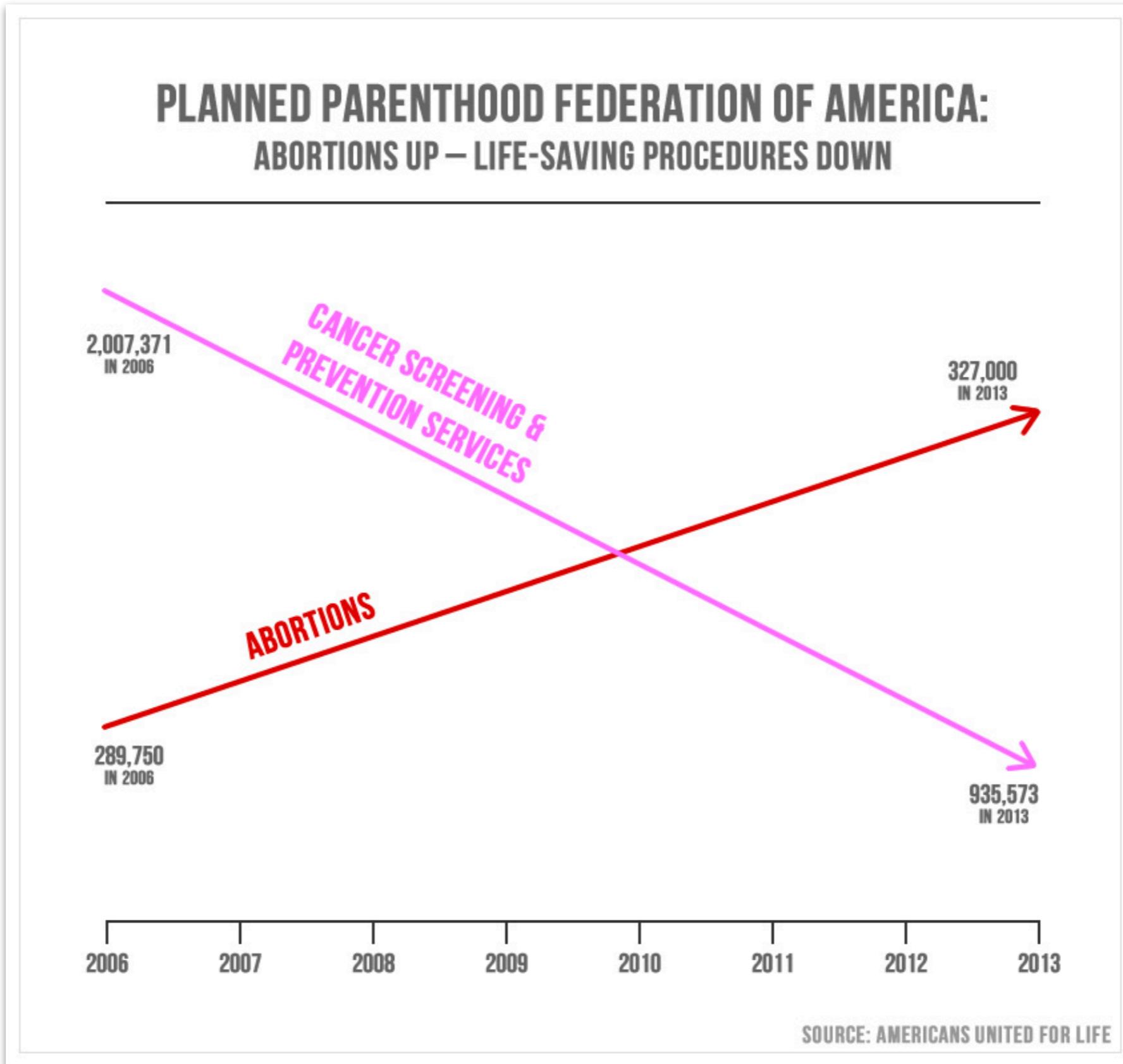
Plots that can be improved ...



Plots that can be improved ...



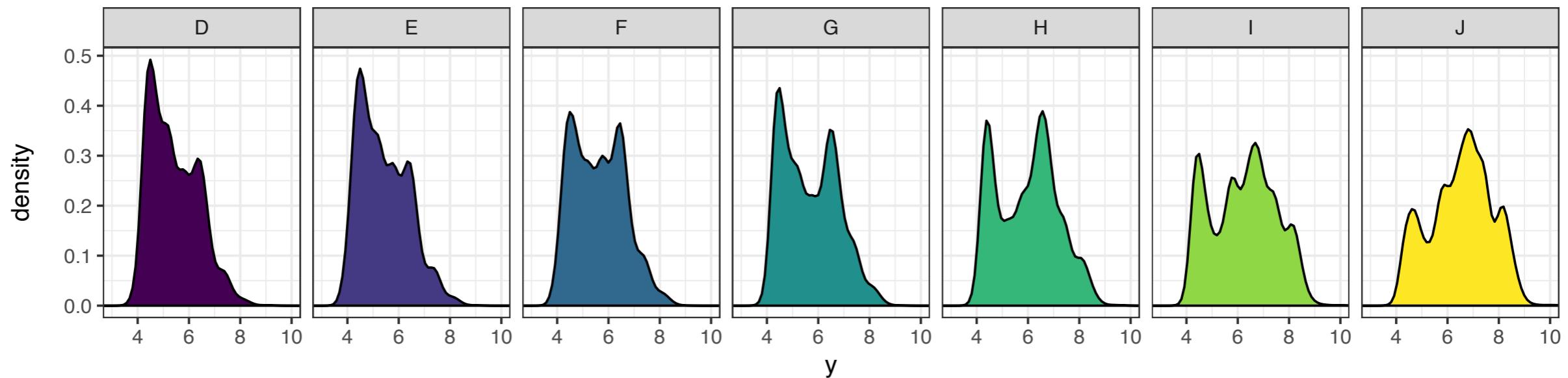
Plots that can be improved ...



A nicer plot

A

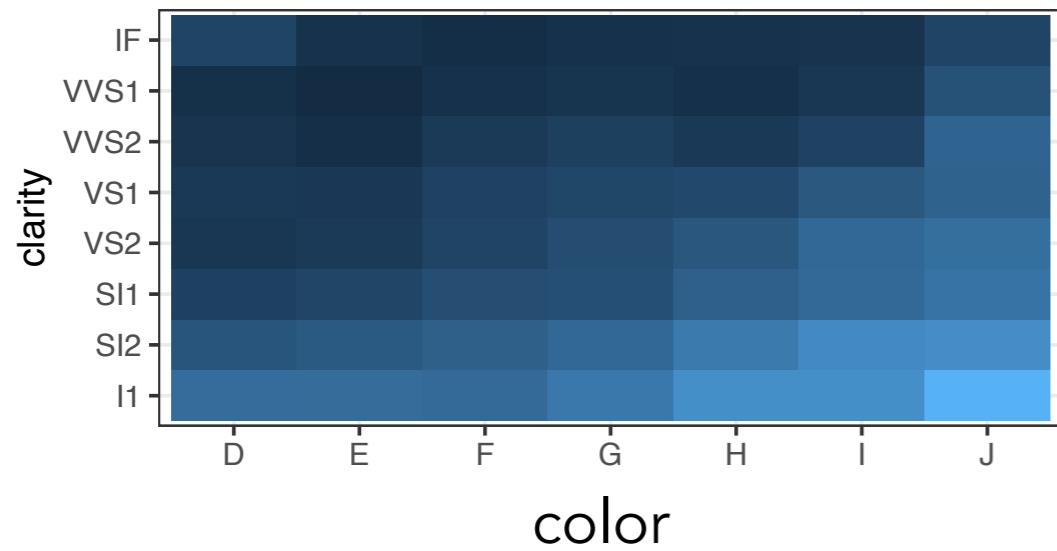
Width of differently colored diamonds



B

Carat values

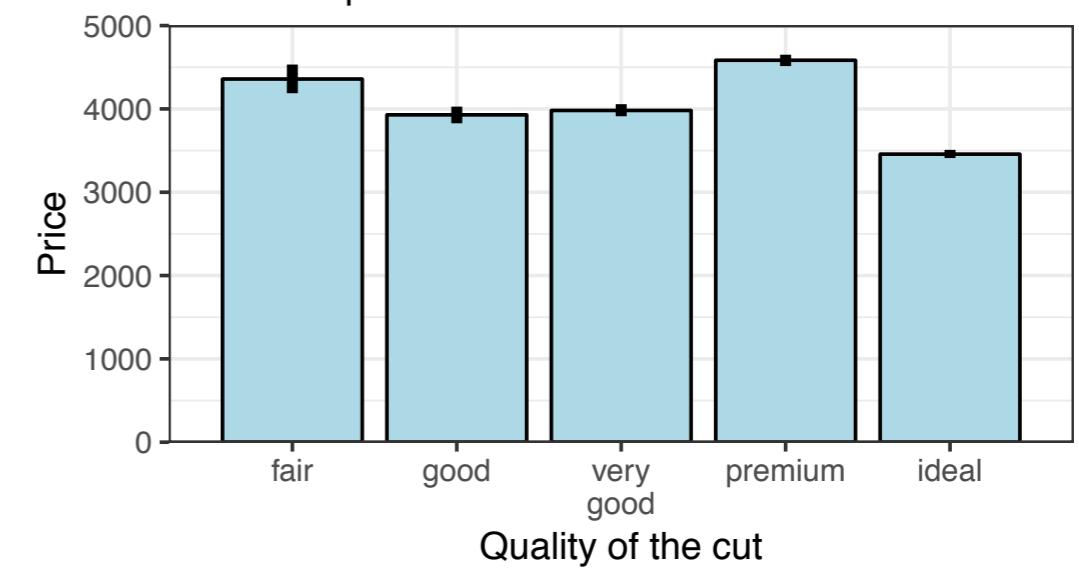
For different color and clarity



C

Price as a function of cut

Note: The price is in US dollars



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~~:



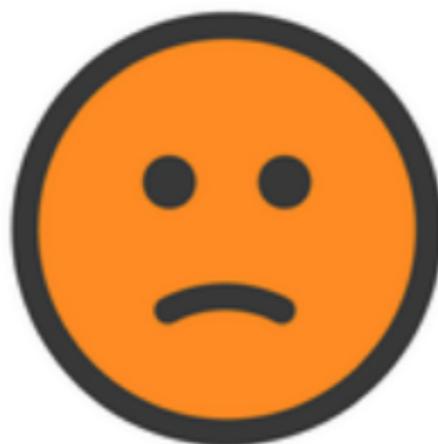
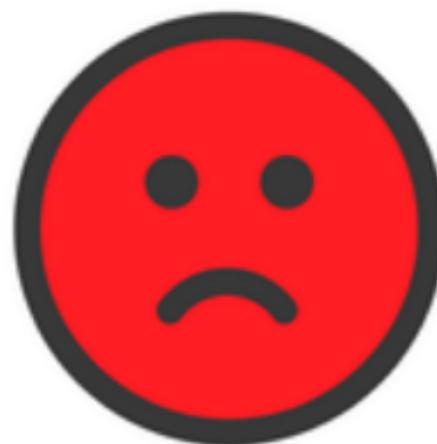
2:37 AM · Nov 24, 2019 · Twitter for iPhone

162 Retweets **1.7K Likes**

How was the pace of today's class?

much a little just a little much
too too right too too
slow slow fast 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

