Something old, something new, something borrowed, something blue

Ways to teach data science (and learn it too!)

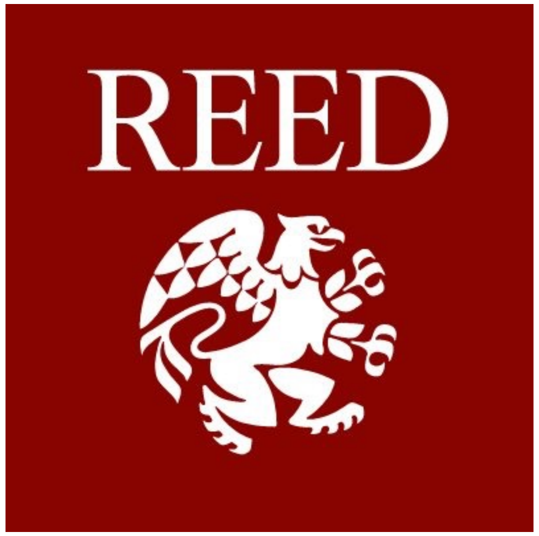
Albert Y. Kim
Amherst College (Smith College July 2018)

Slides available at twitter.com/rudeboybert

Background

W
WASHINGTON

Google


REED

1800
Middlebury

AMHERST
MAMMOTHS

SMITH COLLEGE

Focus of Today

- Talk is nominally is about how I teach intro statistics and data science courses
- However can apply to a broader target demographic
- R-centric, but many of these ideas are language agnostic

Amherst College STAT135

- Course webpage
- Heterogeneous group: Backgrounds and socio-economics status
- Majors: Math, Stats, Econ, Bio, Neuroscience, Psych, Poli Sci, Environmental Studies
- All had high school algebra, most had no coding experience

Question

How can we introduce ***data and computation*** novices to:

1. **Data science**: Data visualization, data wrangling, exploratory data analysis

2. **Data modeling**: Explanation (causal inference) & prediction (machine learning), correlation

3. **Statistical inference**: elementary probability theory, sampling distributions, standard errors, confidence intervals, hypothesis/AB testing & p-values

ModernDive

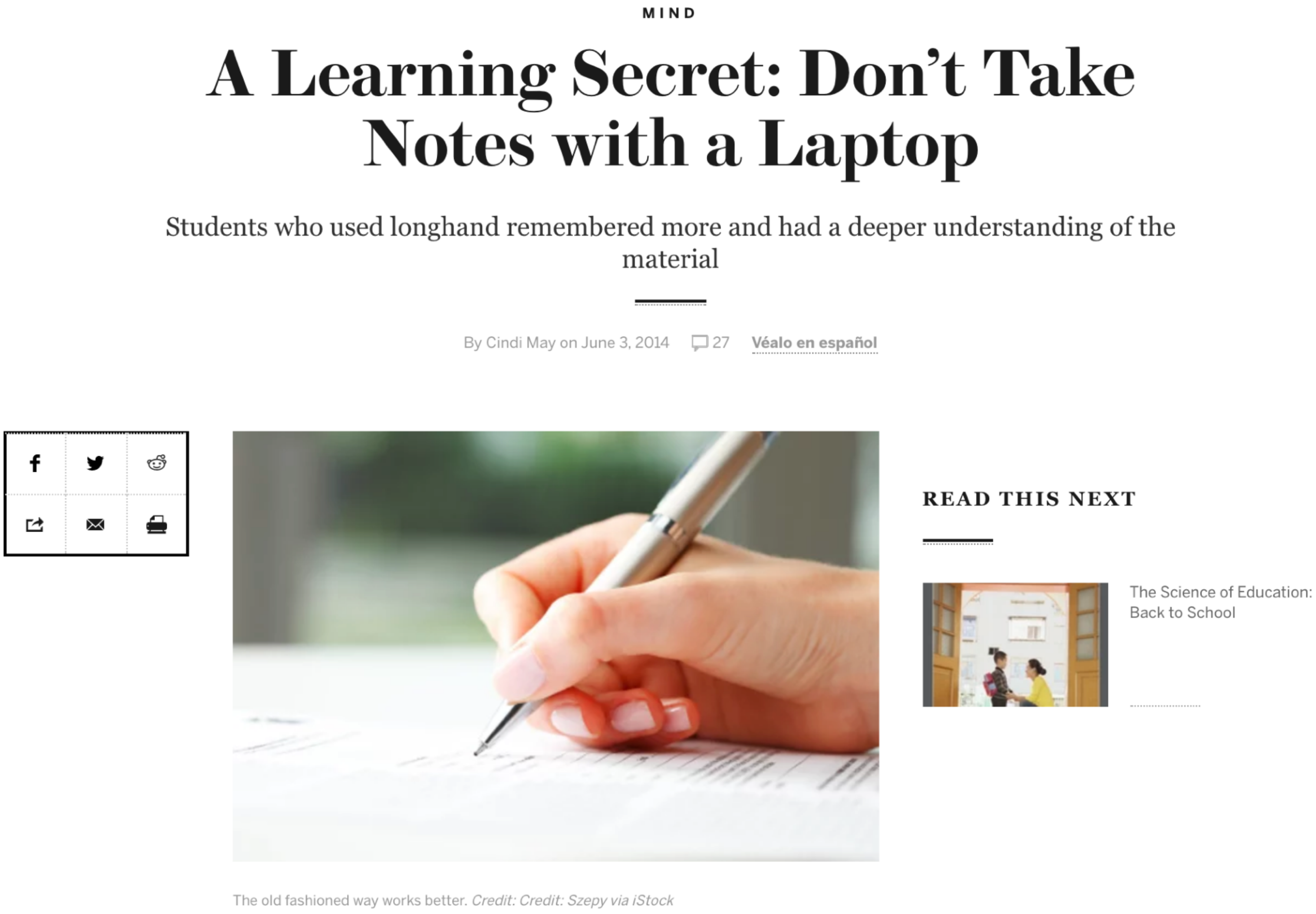
An Introduction to Statistical and Data Sciences via R

- Online textbook available at moderndive.com
- Development version at moderndive.netlify.com
- On GitHub at github.com/moderndive/

Something
Old
New
Borrowed
Blue

Technology in the classroom?

The debate continues…



MIND

A Learning Secret: Don't Take Notes with a Laptop

Students who used longhand remembered more and had a deeper understanding of the material

By Cindi May on June 3, 2014 27 Véalo en español

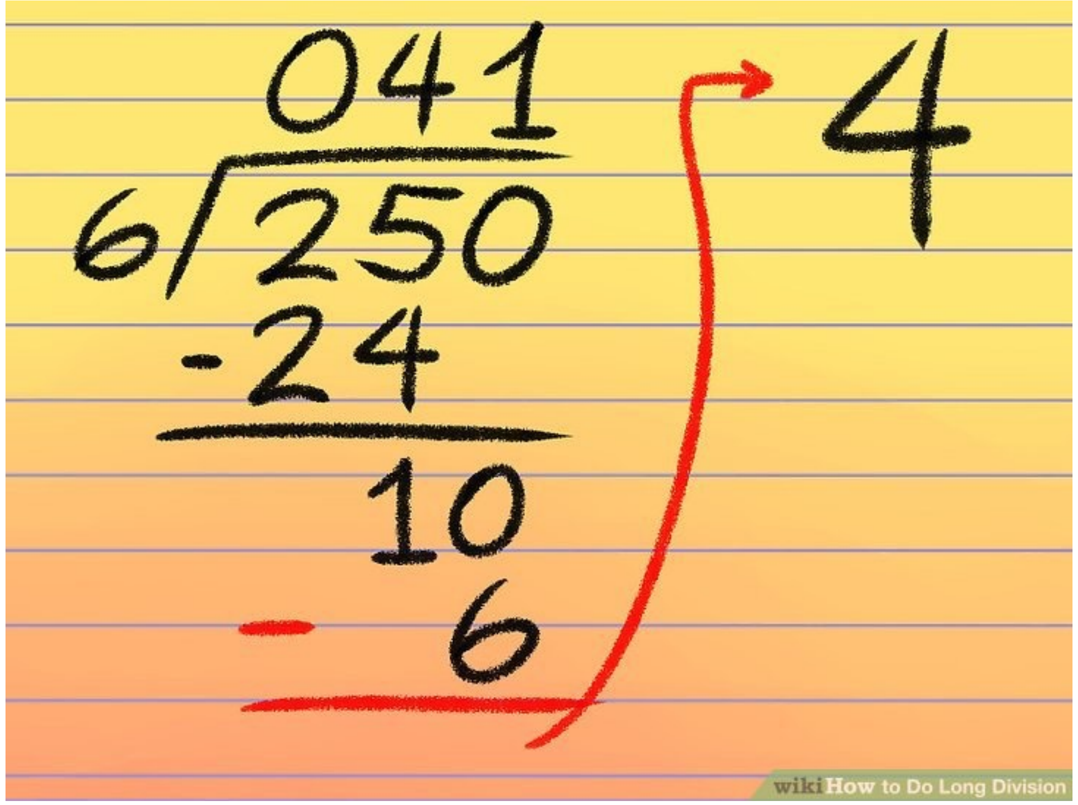
READ THIS NEXT

The Science of Education: Back to School

The old fashioned way works better. *Credit: Credit: Szepy via iStock*

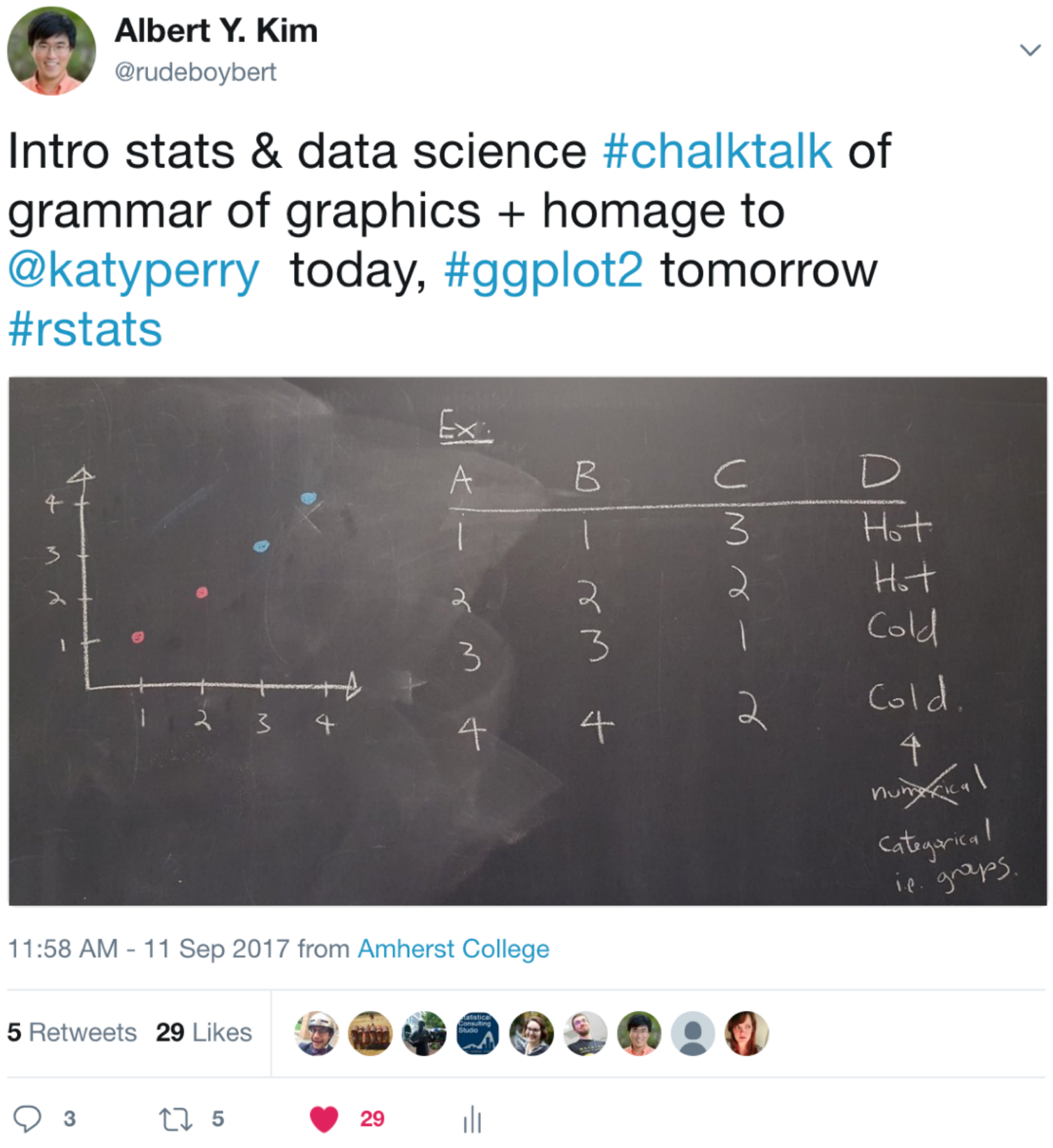
Analogy: Learning Long Division

Do this a few times:



Then rely on this:

`ggplot2` via the Grammar of Graphics



Albert Y. Kim
@rudeboybert

Intro stats & data science #chalktalk of grammar of graphics + homage to @katyperry today, #ggplot2 tomorrow #rstats

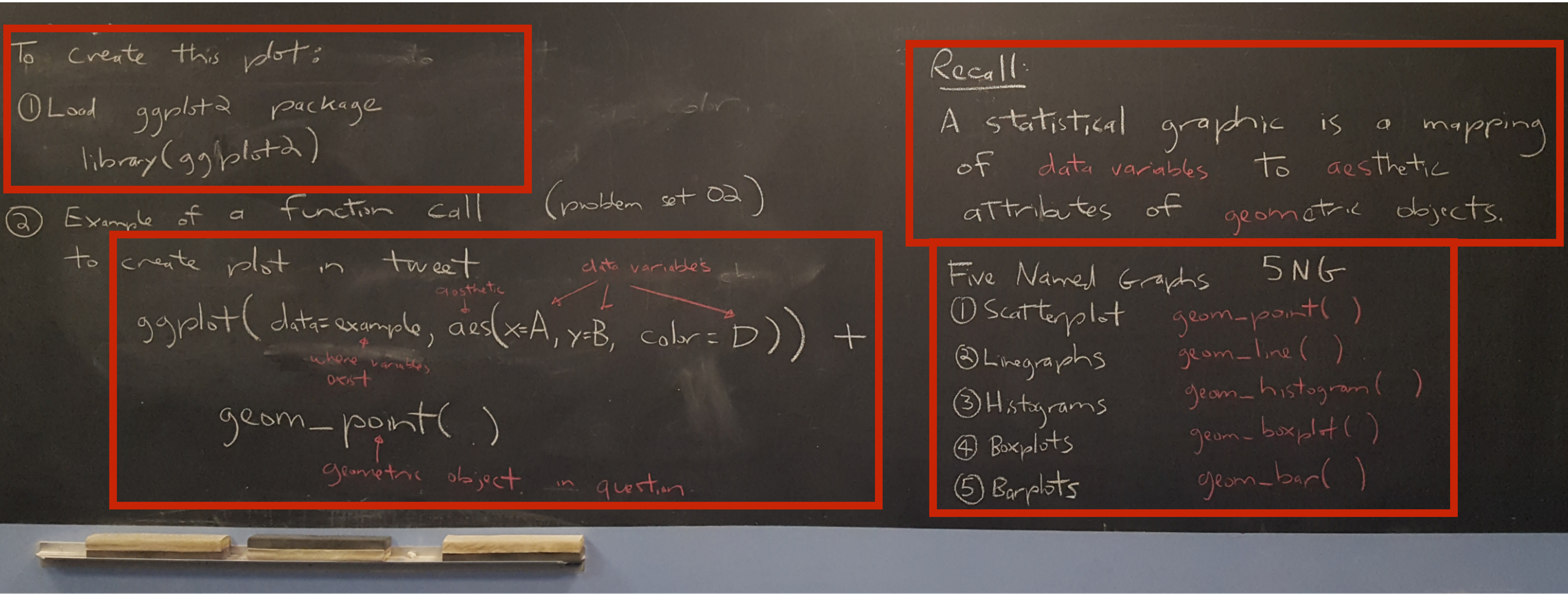
Ex:

A	B	C	D
1	1	3	Hot
2	2	2	Hot
3	3	1	Cold
4	4	2	Cold.

11:58 AM - 11 Sep 2017 from Amherst College

5 Retweets **29** Likes

`ggplot2` via the Grammar of Graphics



To create this plot:

① Load ggplot2 package

```
library(ggplot2)
```

② Example of a function call (problem set 02) to create plot in tweet

aesthetic

data variables

```
ggplot( data=example, aes(x=A, y=B, color=D)) +

    geom_point( )
```

where variables exist

geometric object in question

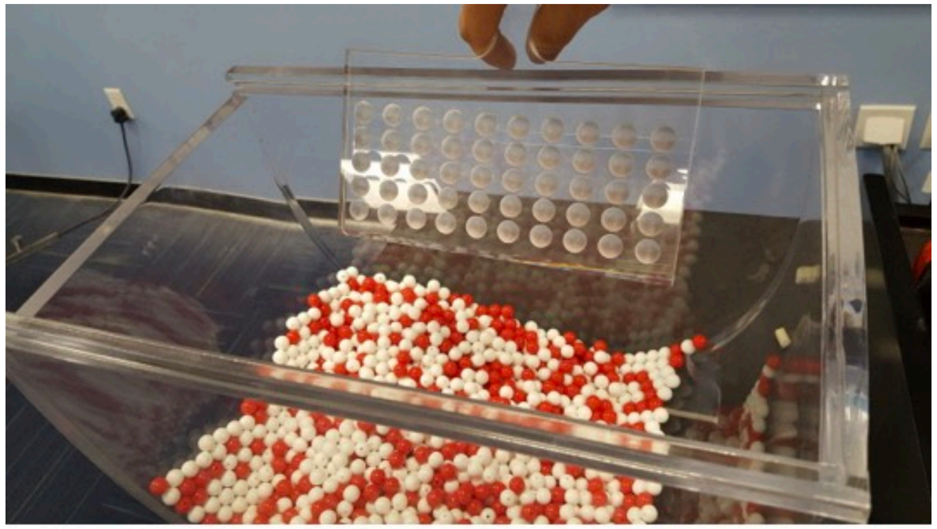
Recall:

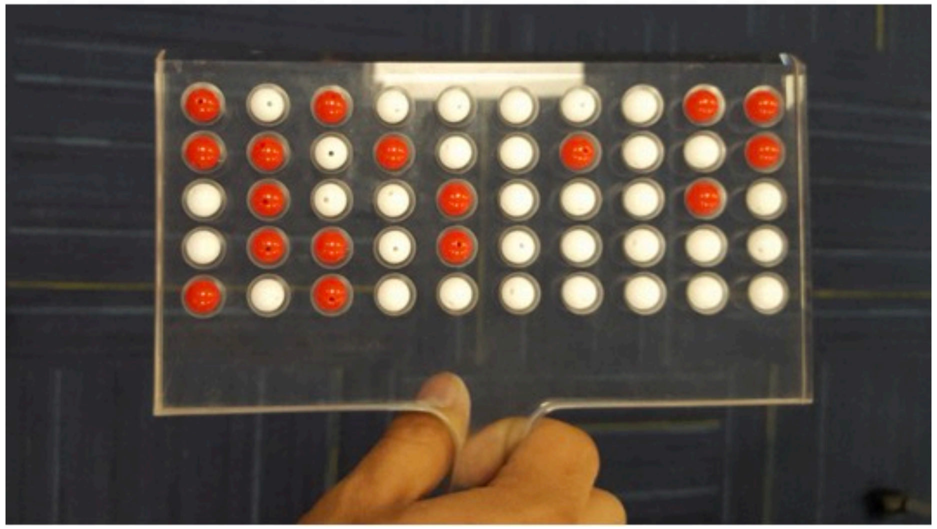
A statistical graphic is a mapping of data variables to aesthetic attributes of geometric objects.

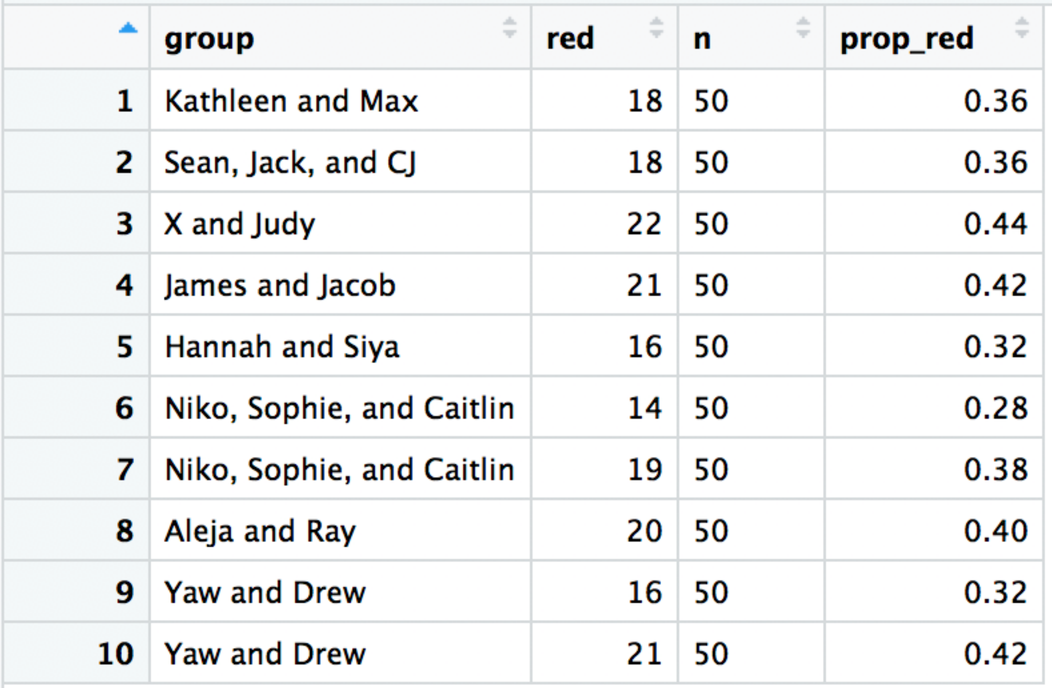
Five Named Graphs 5NG

① Scatterplot	geom_point( )
② Linegraphs	geom_line( )
③ Histograms	geom_histogram( )
④ Boxplots	geom_boxplot( )
⑤ Barplots	geom_bar( )

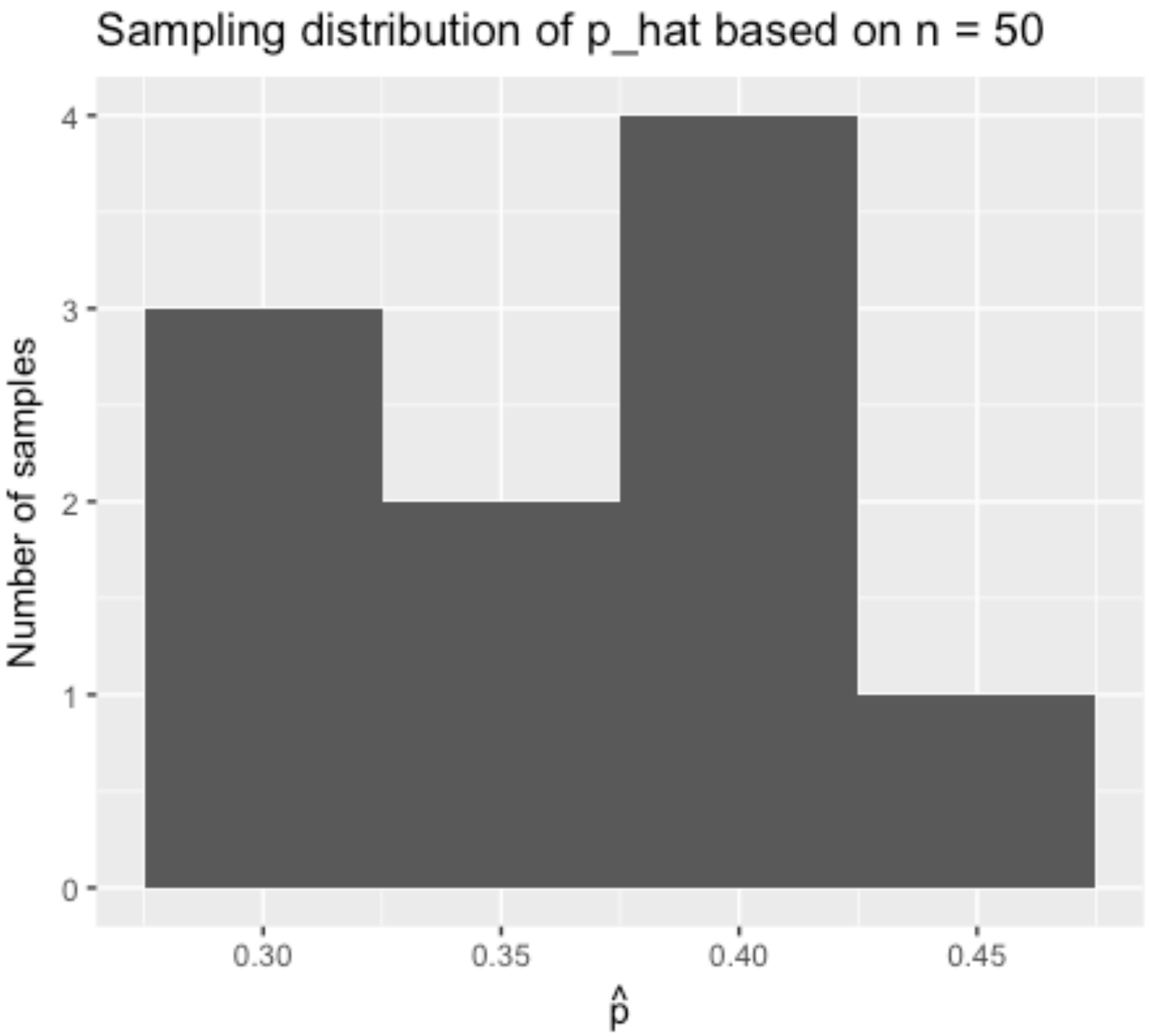
Tactile simulation of sampling to teach sampling distributions



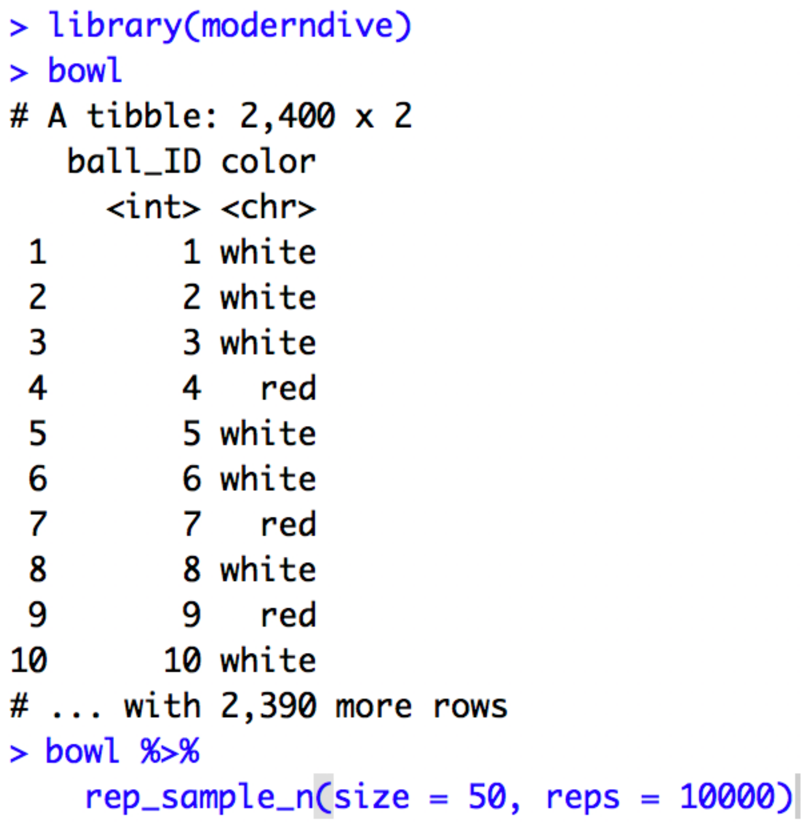




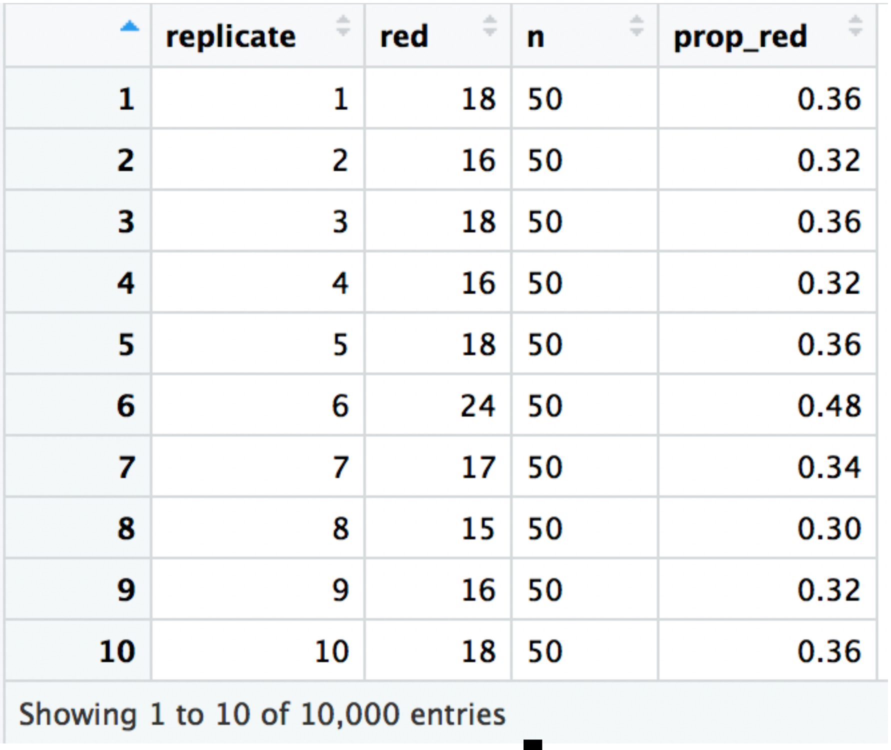
	group	red	n	prop_red
1	Kathleen and Max	18	50	0.36
2	Sean, Jack, and CJ	18	50	0.36
3	X and Judy	22	50	0.44
4	James and Jacob	21	50	0.42
5	Hannah and Siya	16	50	0.32
6	Niko, Sophie, and Caitlin	14	50	0.28
7	Niko, Sophie, and Caitlin	19	50	0.38
8	Aleja and Ray	20	50	0.40
9	Yaw and Drew	16	50	0.32
10	Yaw and Drew	21	50	0.42



Computer simulation of sampling to teach sampling distributions

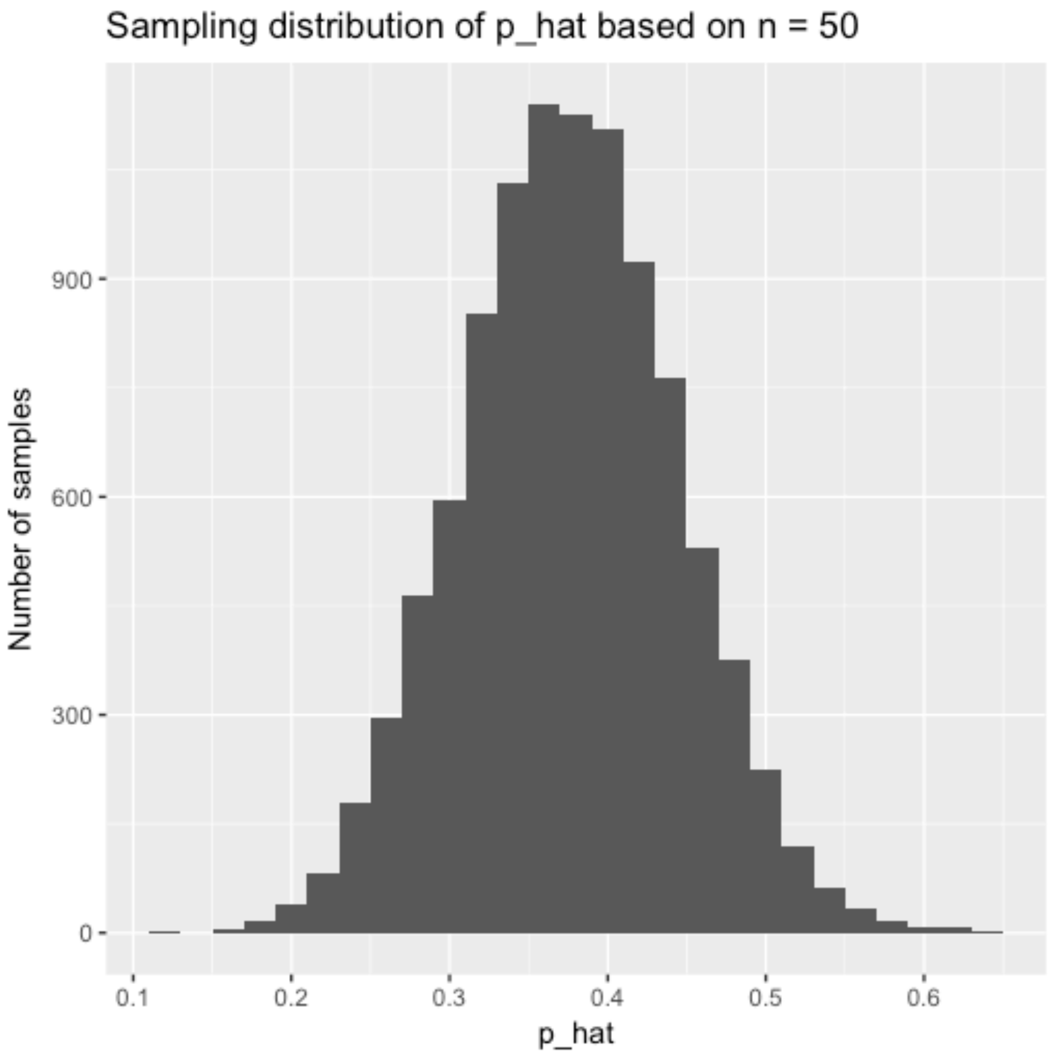


```
> library(moderndive)
> bowl
# A tibble: 2,400 x 2
   ball_ID color
     <int> <chr>
 1       1 white
 2       2 white
 3       3 white
 4       4   red
 5       5 white
 6       6 white
 7       7   red
 8       8 white
 9       9   red
10      10 white
# ... with 2,390 more rows
> bowl %>%
    rep_sample_n(size = 50, reps = 10000)
```



	replicate	red	n	prop_red
1	1	18	50	0.36
2	2	16	50	0.32
3	3	18	50	0.36
4	4	16	50	0.32
5	5	18	50	0.36
6	6	24	50	0.48
7	7	17	50	0.34
8	8	15	50	0.30
9	9	16	50	0.32
10	10	18	50	0.36

Showing 1 to 10 of 10,000 entries



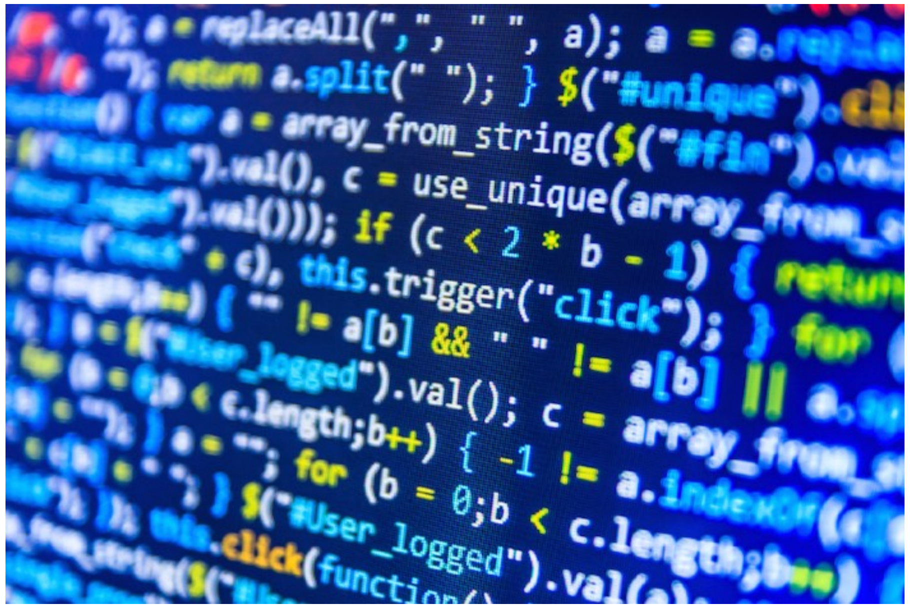
Something
Old
New
Borrowed
Blue

Coding

Cobb (2015) argued there are two possible computational engines for statistics:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_{\overline{X}_1 - \overline{X}_2}} = \frac{\overline{X}_1 - \overline{X}_2}{S_{\overline{X}_1 - \overline{X}_2}}$$

$$S_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\left[\frac{1}{N_1} + \frac{1}{N_2}\right]}$$



```
"); a = replaceAll(",", " ", a); a = a.repl
""); return a.split(" "); } $("#unique").cl
function() { var a = array_from_string($("#fin").va
("#test_val").val(), c = use_unique(array_from_s
("#user_logged").val())); if (c < 2 * b - 1) { return
("check" + c), this.trigger("click"); } for (
c.length;b++) { "" != a[b] && " " != a[b] || a.sp
b = $("#User_logged").val(); c = array_from_st
for (b = 0;b < c.length;b++) { -1 != a.indexOf(
a = ""; for (b = 0;b < c.length;b++)
} $("#User_logged").val(a)
this.click(function(
```

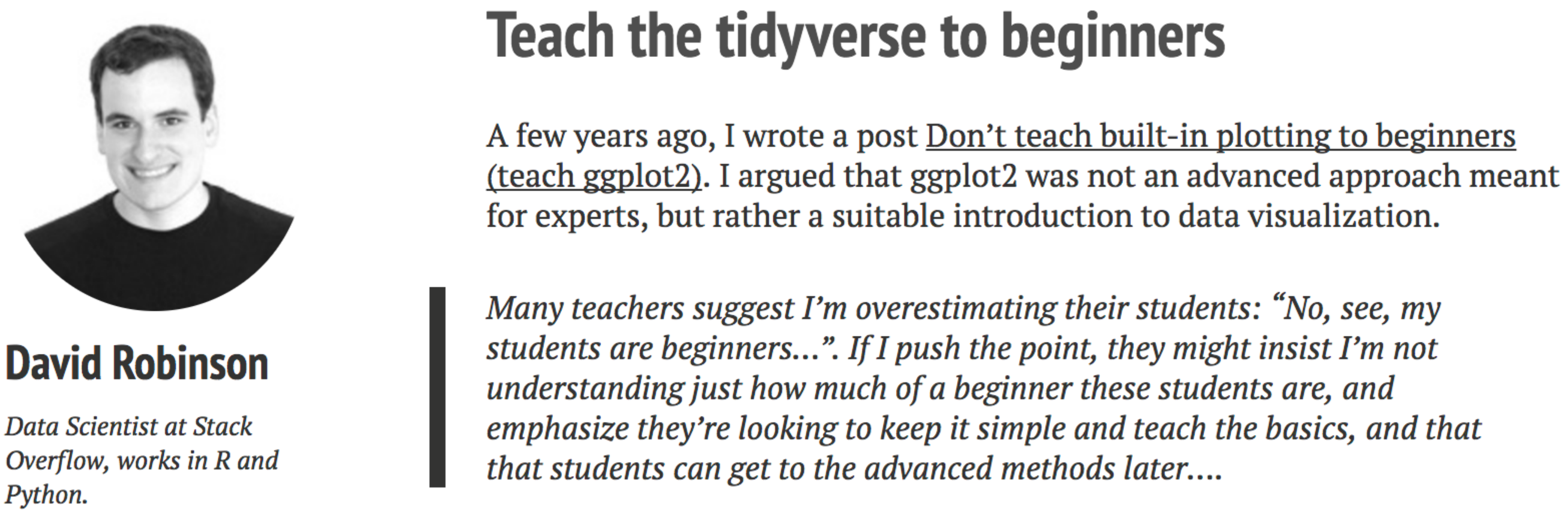
Teaching/Learning Code

- Learn how a practitioner would learn: the "Copy/paste/tweak approach"
- Borrow elements of "flipped classroom": how to use time we're all in the same room together?

Teaching Coding: The Battle is Psychological

- “Don’t code from scratch, take the copy/paste/tweak approach!”
- “Computers are stupid!”
- “Learning to code is similar to learning a language”

New Tools Specific for Data Science



David Robinson

Data Scientist at Stack Overflow, works in R and Python.

Teach the tidyverse to beginners

A few years ago, I wrote a post Don't teach built-in plotting to beginners (teach ggplot2). I argued that ggplot2 was not an advanced approach meant for experts, but rather a suitable introduction to data visualization.

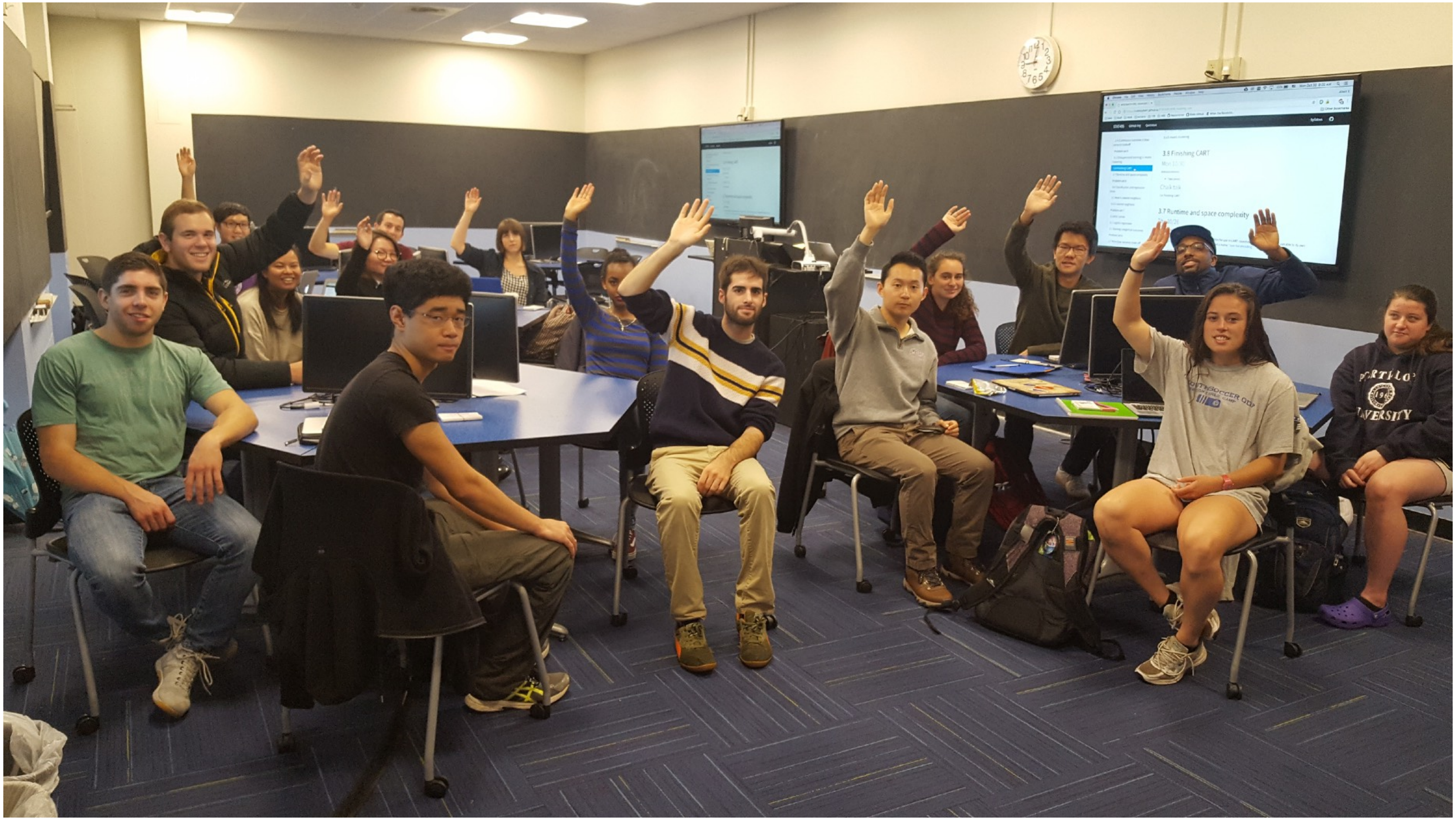
> *Many teachers suggest I'm overestimating their students: "No, see, my students are beginners…". If I push the point, they might insist I'm not understanding just how much of a beginner these students are, and emphasize they're looking to keep it simple and teach the basics, and that students can get to the advanced methods later….*

DataCamp

DataCamp: Immediate Feedback

- Students can practice failing, but with support.
- Difference with Coursera & Udacity?
- DataCamp will pick off low hanging fruit. Ex:
    1. Matching parentheses
    2. Variable name misspellings
    3. Linearity of programs
- Examples of “Curse of knowledge”

Without DataCamp: # of Questions on Coding



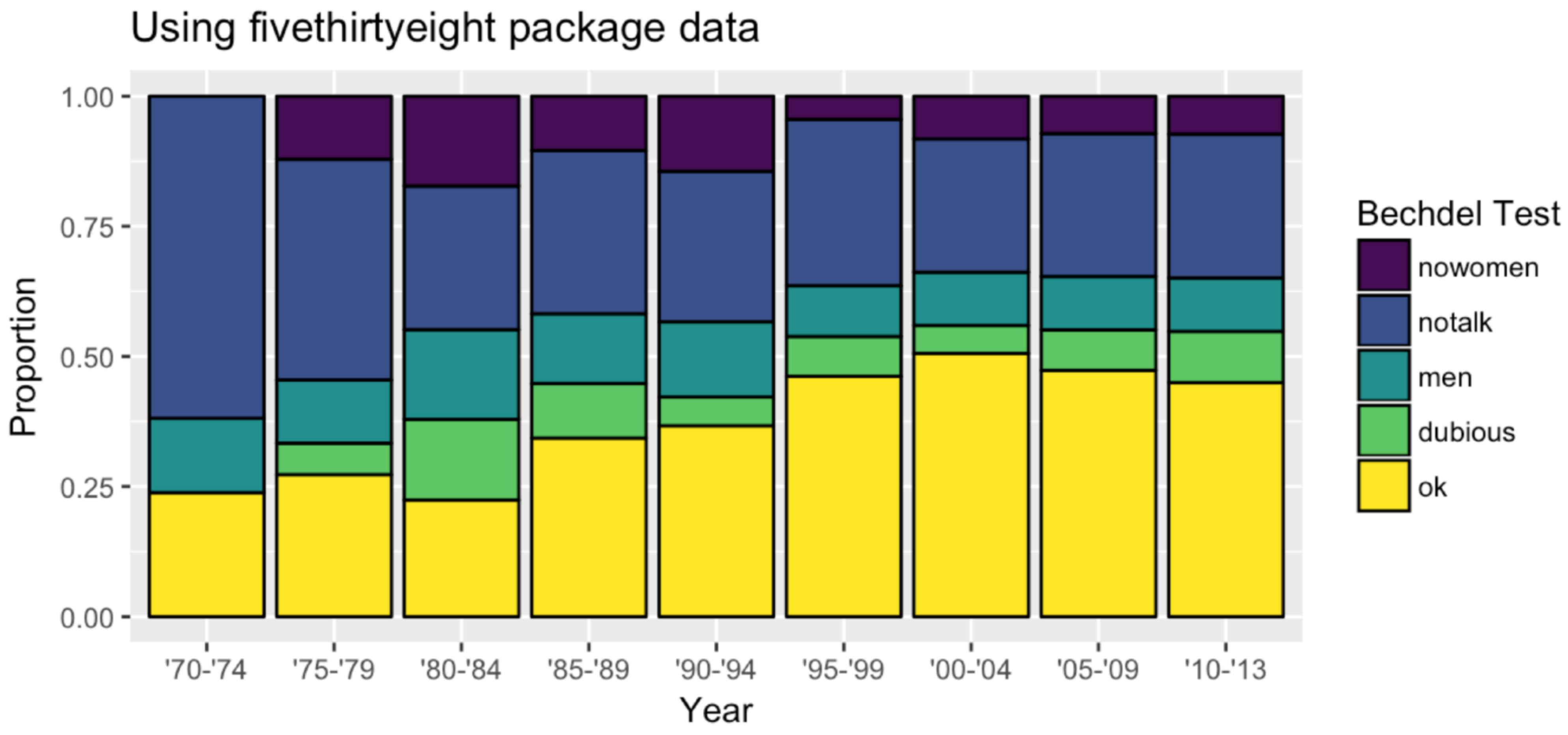
With DataCamp: # of Questions on Coding

Something
Old
New
Borrowed
Blue

Leverage open source

Open data, such as data in R packages like nycflights13, gapminder, fivethirtyeight

Bechdel test?   Original 538 article



Leverage open source

New textbook authoring paradigm

+

The R Series

bookdown

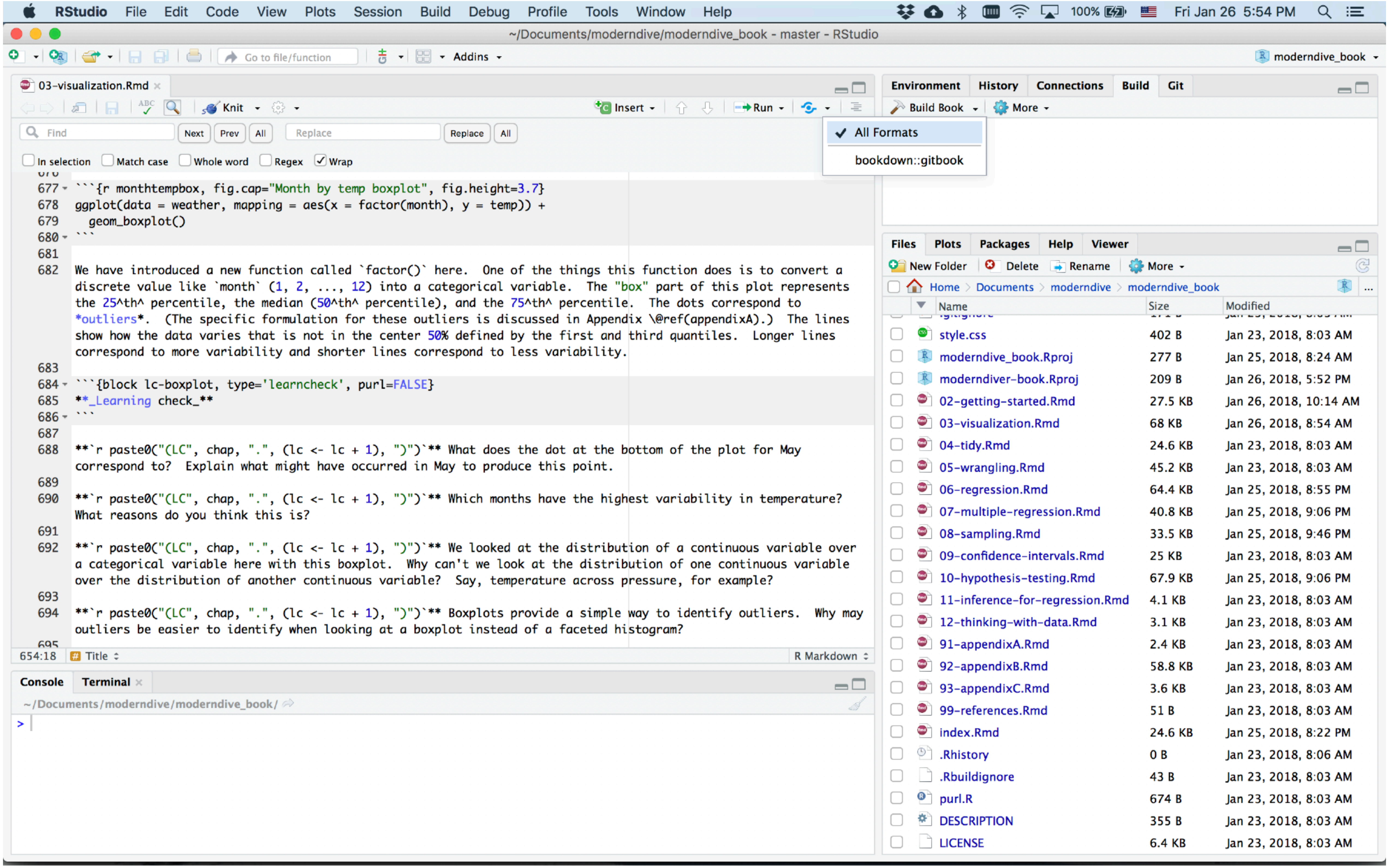
Authoring Books and Technical Documents with R Markdown

Yihui Xie

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

+

New textbook authoring paradigm



□ In selection □ Match case □ Whole word □ Regex ☑ Wrap

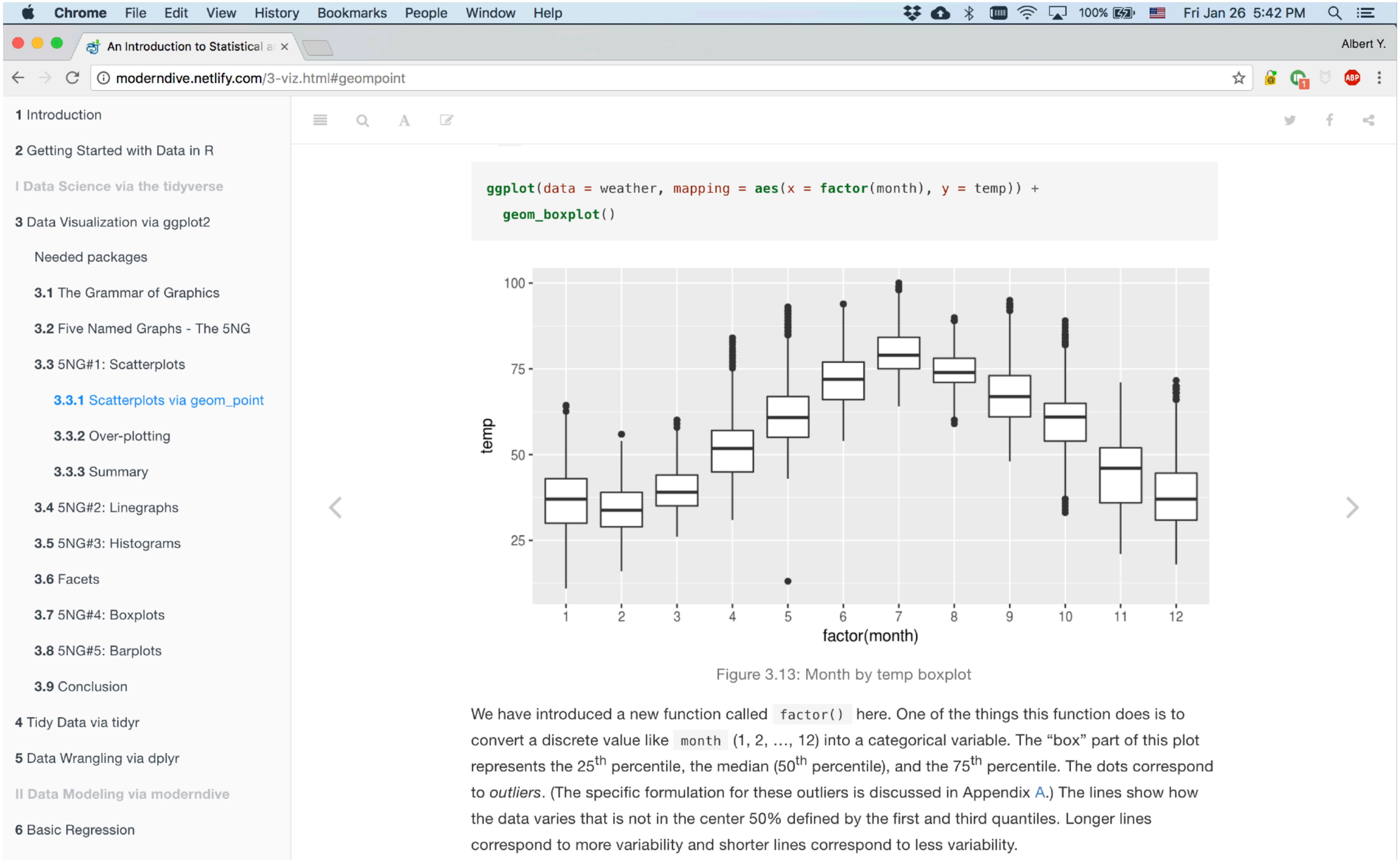
✓ All Formats
bookdown::gitbook

~~~markdown
```{r monthtempbox, fig.cap="Month by temp boxplot", fig.height=3.7}
ggplot(data = weather, mapping = aes(x = factor(month), y = temp)) +
  geom_boxplot()
```

We have introduced a new function called `factor()` here.  One of the things this function does is to convert a discrete value like `month` (1, 2, ..., 12) into a categorical variable.  The "box" part of this plot represents the 25^th^ percentile, the median (50^th^ percentile), and the 75^th^ percentile.  The dots correspond to *outliers*.  (The specific formulation for these outliers is discussed in Appendix \@ref(appendixA).)  The lines show how the data varies that is not in the center 50% defined by the first and third quantiles.  Longer lines correspond to more variability and shorter lines correspond to less variability.

```{block lc-boxplot, type='learncheck', purl=FALSE}
**_Learning check_**
```

**`r paste0("LC", chap, ".", (lc <- lc + 1), ")")`** What does the dot at the bottom of the plot for May correspond to?  Explain what might have occurred in May to produce this point.

**`r paste0("LC", chap, ".", (lc <- lc + 1), ")")`** Which months have the highest variability in temperature?  What reasons do you think this is?

**`r paste0("LC", chap, ".", (lc <- lc + 1), ")")`** We looked at the distribution of a continuous variable over a categorical variable here with this boxplot.  Why can't we look at the distribution of one continuous variable over the distribution of another continuous variable?  Say, temperature across pressure, for example?

**`r paste0("LC", chap, ".", (lc <- lc + 1), ")")`** Boxplots provide a simple way to identify outliers.  Why may outliers be easier to identify when looking at a boxplot instead of a faceted histogram?
~~~

```
~/Documents/moderndive/moderndive_book/
>
```

Name	Size	Modified
style.css	402 B	Jan 23, 2018, 8:03 AM
moderndive_book.Rproj	277 B	Jan 25, 2018, 8:24 AM
moderndiver-book.Rproj	209 B	Jan 26, 2018, 5:52 PM
02-getting-started.Rmd	27.5 KB	Jan 26, 2018, 10:14 AM
03-visualization.Rmd	68 KB	Jan 26, 2018, 8:54 AM
04-tidy.Rmd	24.6 KB	Jan 23, 2018, 8:03 AM
05-wrangling.Rmd	45.2 KB	Jan 23, 2018, 8:03 AM
06-regression.Rmd	64.4 KB	Jan 25, 2018, 8:55 PM
07-multiple-regression.Rmd	40.8 KB	Jan 25, 2018, 9:06 PM
08-sampling.Rmd	33.5 KB	Jan 25, 2018, 9:46 PM
09-confidence-intervals.Rmd	25 KB	Jan 23, 2018, 8:03 AM
10-hypothesis-testing.Rmd	67.9 KB	Jan 25, 2018, 9:06 PM
11-inference-for-regression.Rmd	4.1 KB	Jan 23, 2018, 8:03 AM
12-thinking-with-data.Rmd	3.1 KB	Jan 23, 2018, 8:03 AM
91-appendixA.Rmd	2.4 KB	Jan 23, 2018, 8:03 AM
92-appendixB.Rmd	58.8 KB	Jan 23, 2018, 8:03 AM
93-appendixC.Rmd	3.6 KB	Jan 23, 2018, 8:03 AM
99-references.Rmd	51 B	Jan 23, 2018, 8:03 AM
index.Rmd	24.6 KB	Jan 25, 2018, 8:22 PM
.Rhistory	0 B	Jan 23, 2018, 8:06 AM
.Rbuildignore	43 B	Jan 23, 2018, 8:03 AM
purl.R	674 B	Jan 23, 2018, 8:03 AM
DESCRIPTION	355 B	Jan 23, 2018, 8:03 AM
LICENSE	6.4 KB	Jan 23, 2018, 8:03 AM

New textbook authoring paradigm



Chrome File Edit View History Bookmarks People Window Help
100%
Fri Jan 26 5:42 PM
An Introduction to Statistical a ×
Albert Y.
moderndive.netlify.com/3-viz.html#geompoint
100
75
50
25
temp
1 2 3 4 5 6 7 8 9 10 11 12
factor(month)

```r
ggplot(data = weather, mapping = aes(x = factor(month), y = temp)) +
  geom_boxplot()
```

Figure 3.13: Month by temp boxplot

We have introduced a new function called `factor()` here. One of the things this function does is to convert a discrete value like `month` (1, 2, …, 12) into a categorical variable. The “box” part of this plot represents the 25th percentile, the median (50th percentile), and the 75th percentile. The dots correspond to *outliers*. (The specific formulation for these outliers is discussed in Appendix A.) The lines show how the data varies that is not in the center 50% defined by the first and third quantiles. Longer lines correspond to more variability and shorter lines correspond to less variability.

New textbook authoring paradigm

+

The R Series

bookdown

Authoring Books and Technical Documents with R Markdown

Yihui Xie

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

www.rstudio.com

+

"Versions, not editions"

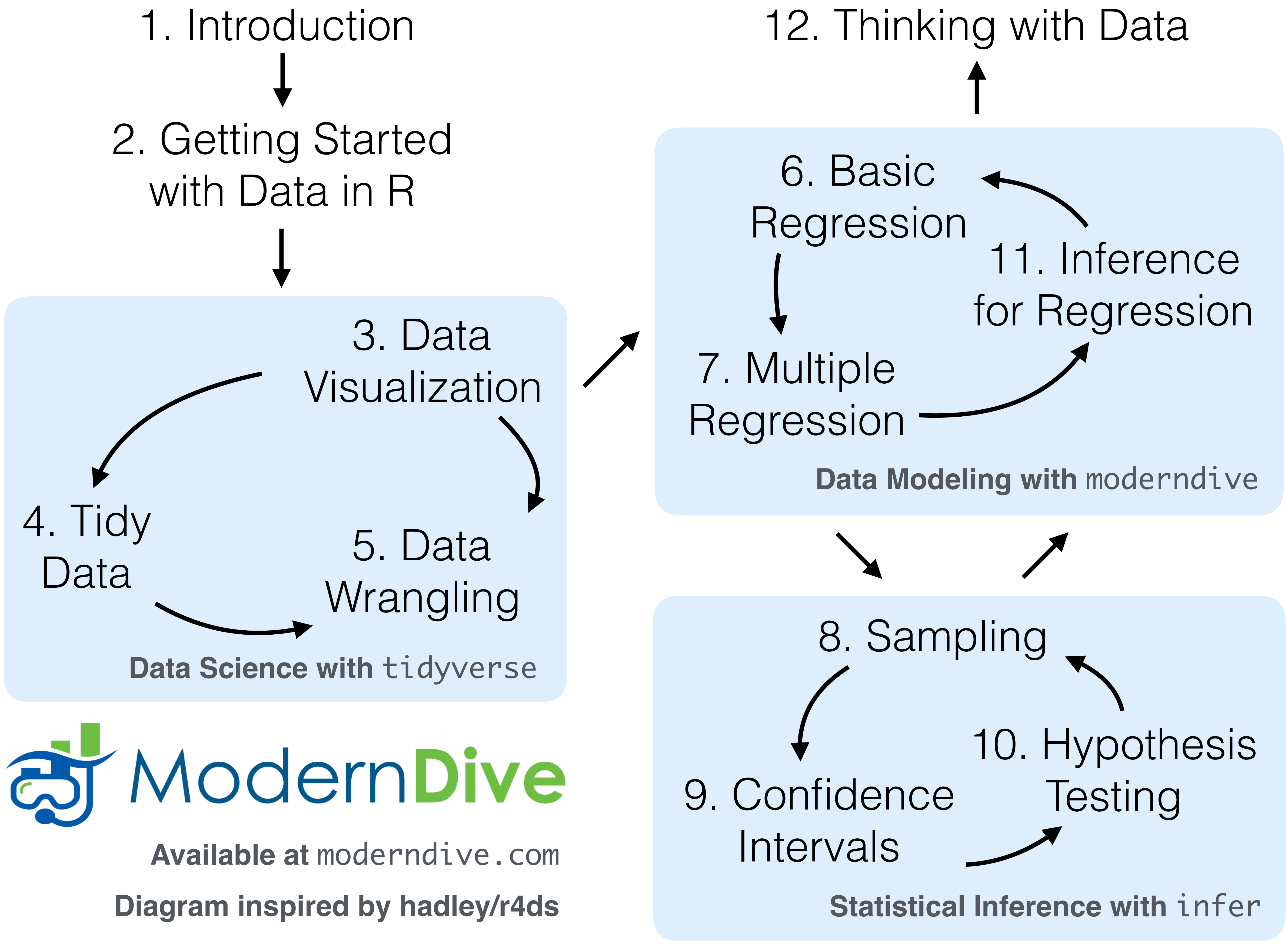
On GitHub at github.com/moderndive/

ModernDive

An Introduction to Statistical and Data Sciences via R

- Available at moderndive.com
- Development version at moderndive.netlify.com
- On GitHub at github.com/moderndive/

v0.3.0 to be released next week! What's new?



1. Introduction

2. Getting Started with Data in R

3. Data Visualization

4. Tidy Data

5. Data Wrangling

Data Science with tidyverse

6. Basic Regression

7. Multiple Regression

11. Inference for Regression

Data Modeling with moderndive

12. Thinking with Data

8. Sampling

9. Confidence Intervals

10. Hypothesis Testing

Statistical Inference with infer

ModernDive

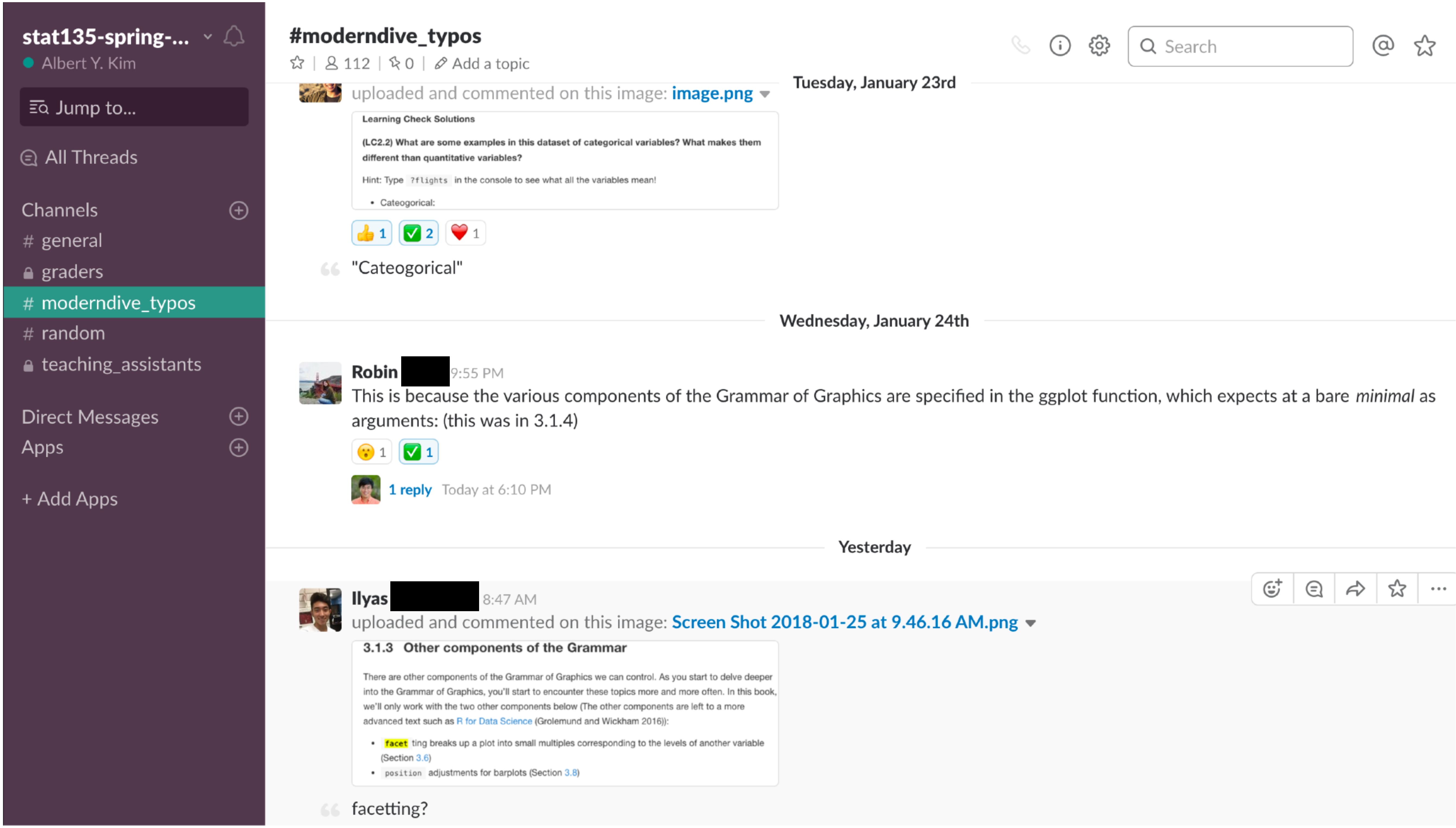
Available at moderndive.com

Diagram inspired by hadley/r4ds

"If You're Not Embarrassed By The First Version Of Your Product, You’ve Launched Too Late"

Reid Hoffman, founder of LinkedIn

Crowdsourcing Typos



Tuesday, January 23rd

uploaded and commented on this image: image.png

Learning Check Solutions

(LC2.2) What are some examples in this dataset of categorical variables? What makes them different than quantitative variables?

Hint: Type `?flights` in the console to see what all the variables mean!

- Cateogorical:

👍 1 ✅ 2 ❤️ 1

"Cateogorical"

Wednesday, January 24th

Robin [redacted] 9:55 PM

This is because the various components of the Grammar of Graphics are specified in the ggplot function, which expects at a bare *minimal* as arguments: (this was in 3.1.4)

😯 1 ✅ 1

1 reply Today at 6:10 PM

Yesterday

Ilyas [redacted] 8:47 AM

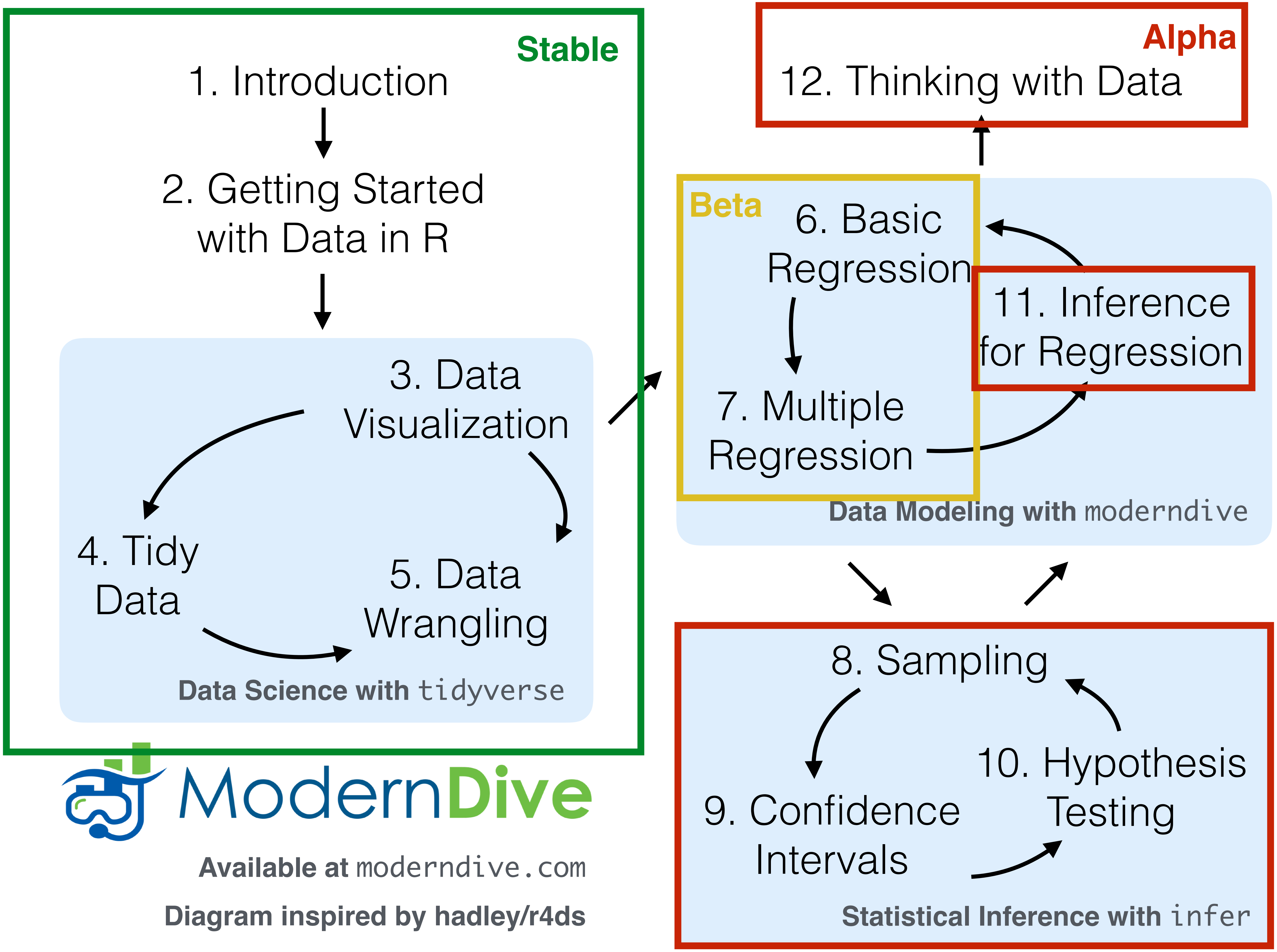
uploaded and commented on this image: Screen Shot 2018-01-25 at 9.46.16 AM.png

3.1.3 Other components of the Grammar

There are other components of the Grammar of Graphics we can control. As you start to delve deeper into the Grammar of Graphics, you'll start to encounter these topics more and more often. In this book, we'll only work with the two other components below (The other components are left to a more advanced text such as R for Data Science (Grolemund and Wickham 2016)):

- `facet` ting breaks up a plot into small multiples corresponding to the levels of another variable (Section 3.6)
- `position` adjustments for barplots (Section 3.8)

facetting?



Stable

1. Introduction

2. Getting Started with Data in R

3. Data Visualization

4. Tidy Data

5. Data Wrangling

Data Science with `tidyverse`

ModernDive

Available at `moderndive.com`

Diagram inspired by hadley/r4ds

Alpha

12. Thinking with Data

Beta

6. Basic Regression

7. Multiple Regression

11. Inference for Regression

Data Modeling with `moderndive`

8. Sampling

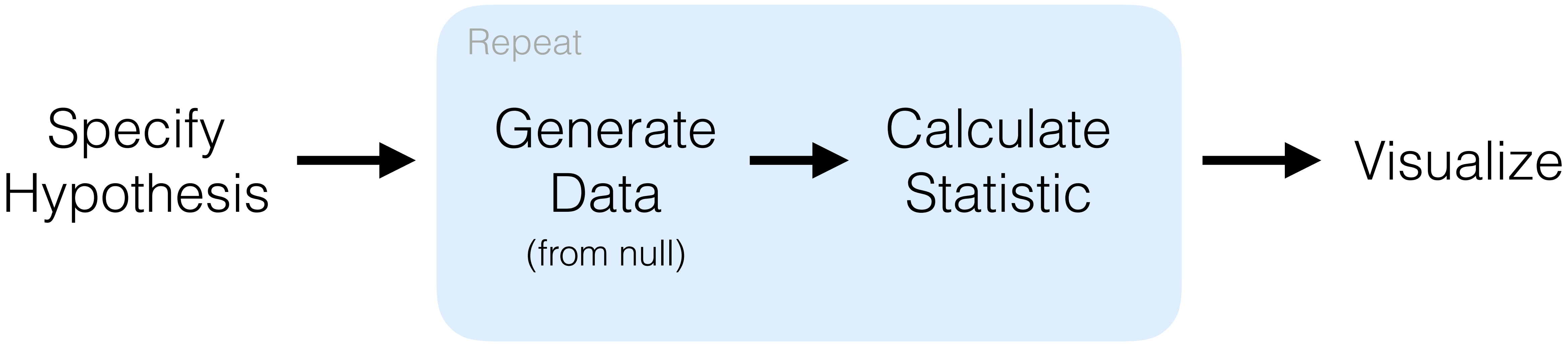
9. Confidence Intervals

10. Hypothesis Testing

Statistical Inference with `infer`

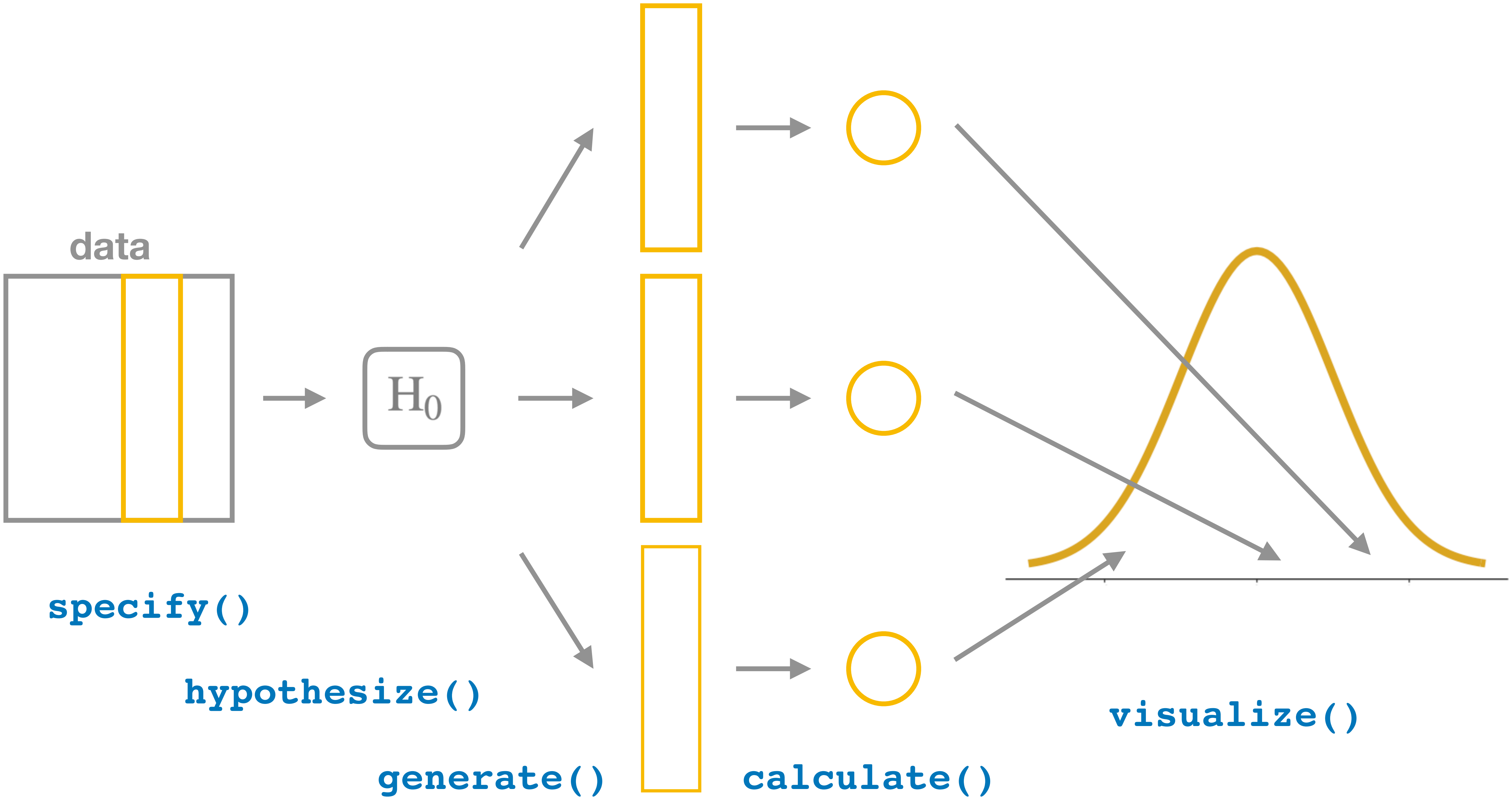
`infer` package for tidy statistical inference

http://infer.netlify.com/



```
hypothesize(null)  %>%  generate(reps)  %>%  calculate(stat)  %>%  visualize()
```

Hypothesis Testing



data

H_0

specify()

hypothesize()

generate()

calculate()

visualize()

“Thinking with Data”

Example student work

- Analysis of crime in Chicago
- How many f**ks does Tarantino Give?
- Final projects: Code and data

Something
Old
New
Borrowed
Blue

+

+

Albert Y. Kim
Amherst College
Twitter: @rudeboybert
GitHub: rudeboybert

Chester Ismay
DataCamp
Twitter: @old_man_chester
GitHub: ismayc

ModernDive

An Introduction to Statistical and Data Sciences via R

- Available at moderndive.com
- Development version at moderndive.netlify.com
- On GitHub at github.com/moderndive/

v0.3.0 to be released next week! What's new?

Slides available at twitter.com/rudeboybert