

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

- About me: Washington Stats, Google, Colleges x 4
- 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
- Most with high school math, most without coding experience

Question

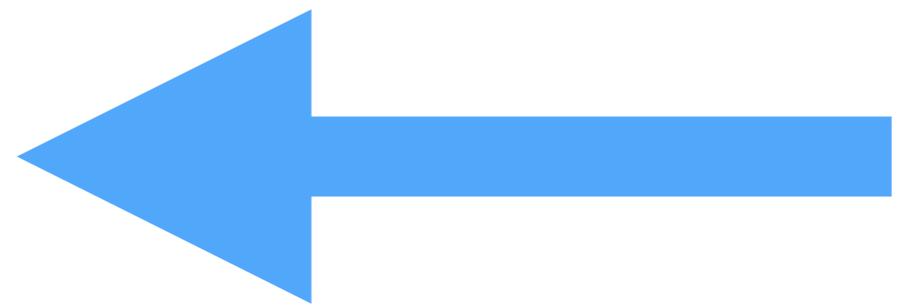
How can we introduce ***data and computation*** novices to:

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



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/



Technology in the classroom?



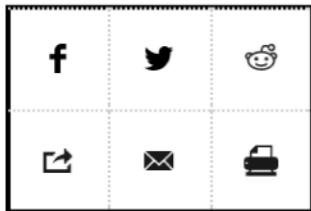
The debate continues...

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](#)



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

READ THIS NEXT



The Science of Education:
Back to School

Analogy: Learning Long Division

Do this a few times:

A handwritten long division problem on yellow lined paper. The divisor is 6, the dividend is 250, and the quotient is 41. A red bracket groups the digits 25 and 0, and a red arrow points from the digit 4 in the quotient to this bracket, indicating the current step of dividing 25 by 6.

041
6 / 250
- 24

10
- 6

wikiHow to Do Long Division

Then rely on this:



ggplot2 via the Grammar of Graphics

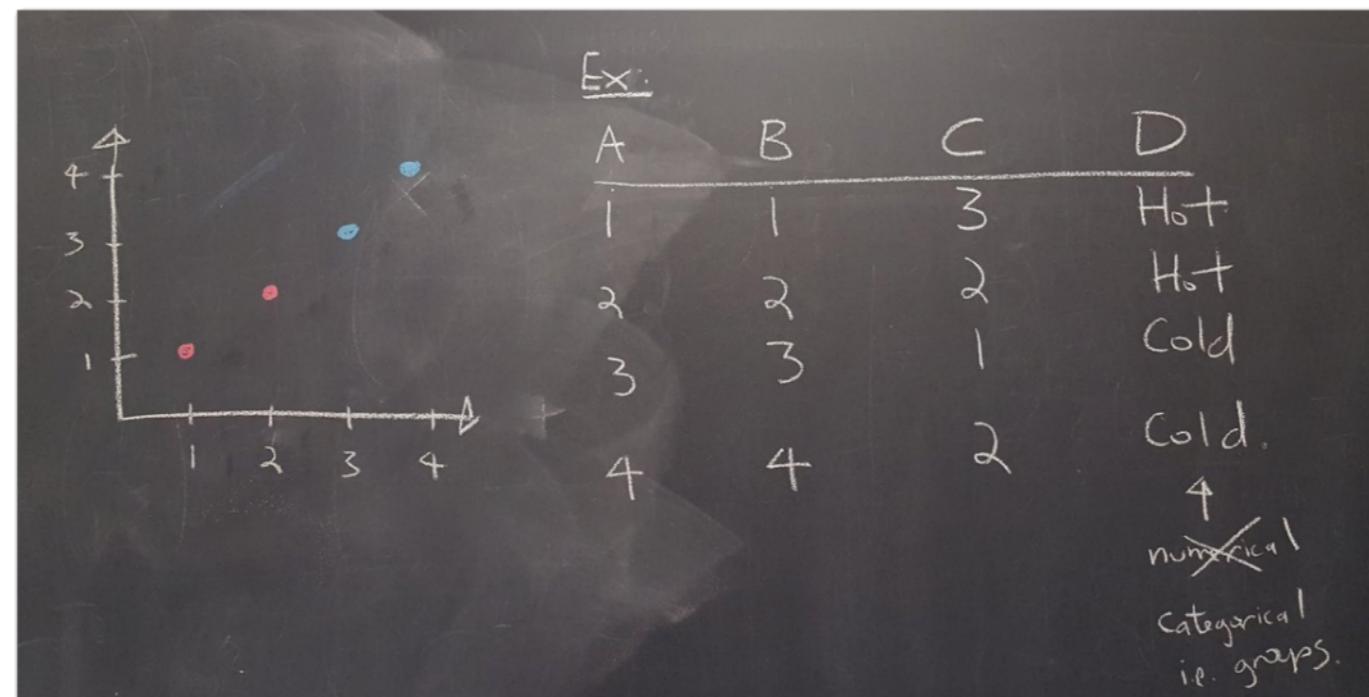


Albert Y. Kim

@rudeboybert



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



11:58 AM - 11 Sep 2017 from Amherst College

5 Retweets 29 Likes



3



5



29



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

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

Annotations:
- 'aes' is labeled 'aesthetic'
- 'x=A' and 'y=B' are labeled 'data variable's'
- 'color=D' is labeled 'color'
- 'geom_point()' is labeled 'geometric object in question'
- 'example' is labeled 'where variables exist'

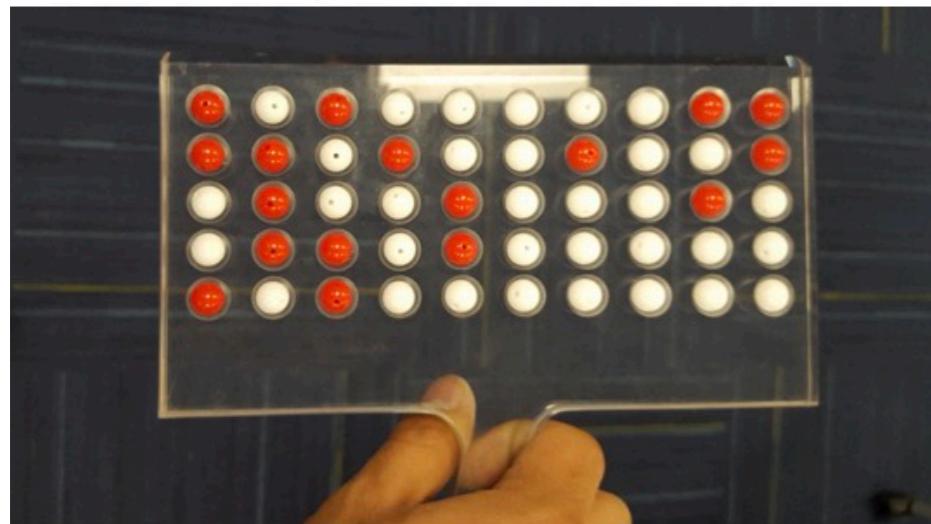
Recall:

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

Five Named Graphs 5NG

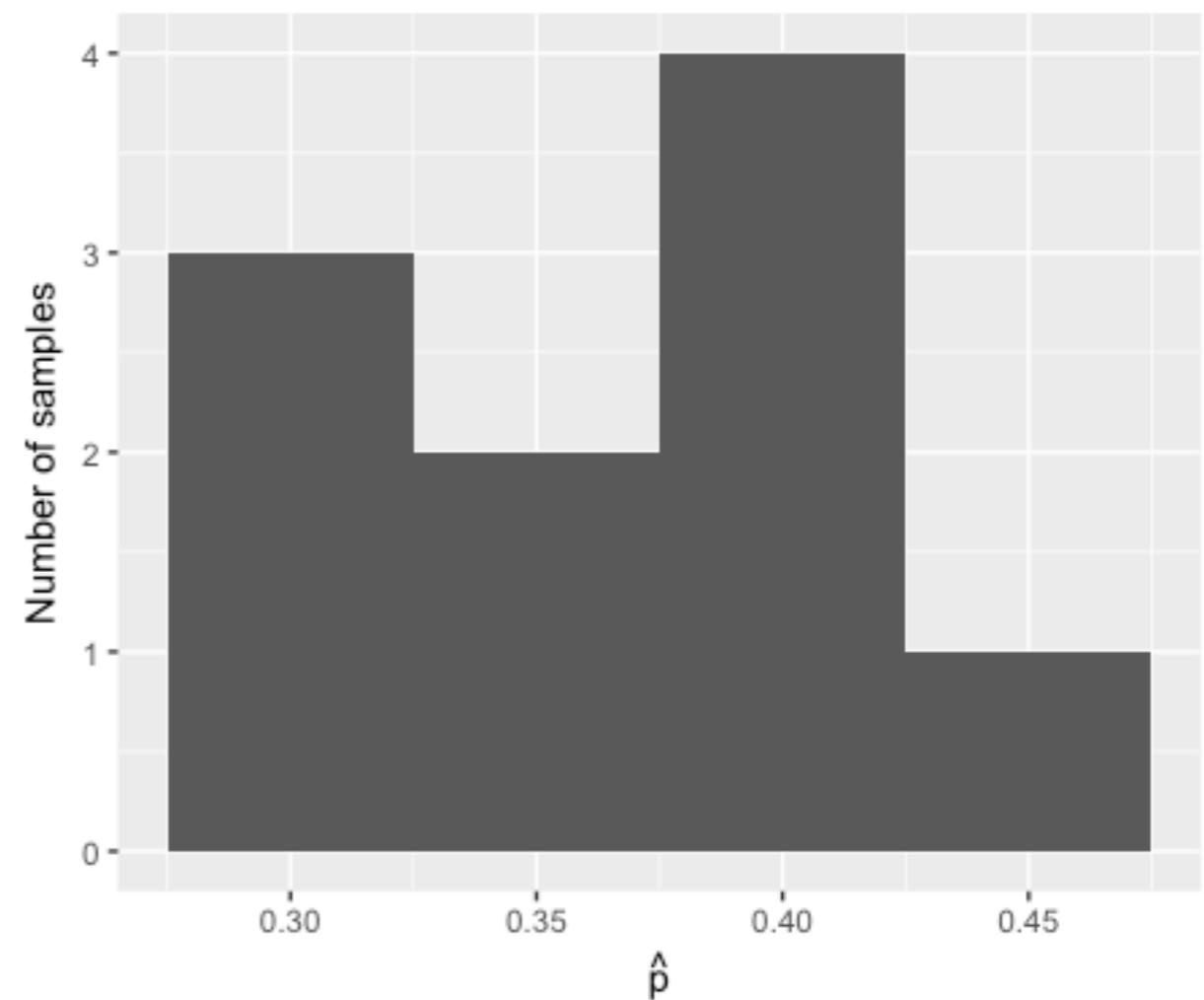
- ① Scatterplots geom_point()
- ② Line graphs geom_line()
- ③ Histograms geom_histogram()
- ④ Boxplots geom_boxplot()
- ⑤ Barplots geom_bar()

Tactile simulation of sampling to teach sampling distributions



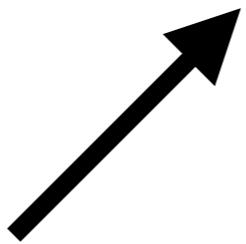
group	red	n	prop_red
1 Kathleen and Max	18	50	0.36
2 Sean, Jack, and CJ	18	50	0.36
3 X and Judy	22	50	0.44
4 James and Jacob	21	50	0.42
5 Hannah and Siya	16	50	0.32
6 Niko, Sophie, and Caitlin	14	50	0.28
7 Niko, Sophie, and Caitlin	19	50	0.38
8 Aleja and Ray	20	50	0.40
9 Yaw and Drew	16	50	0.32
10 Yaw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

Sampling distribution of \hat{p} based on $n = 50$



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)
```

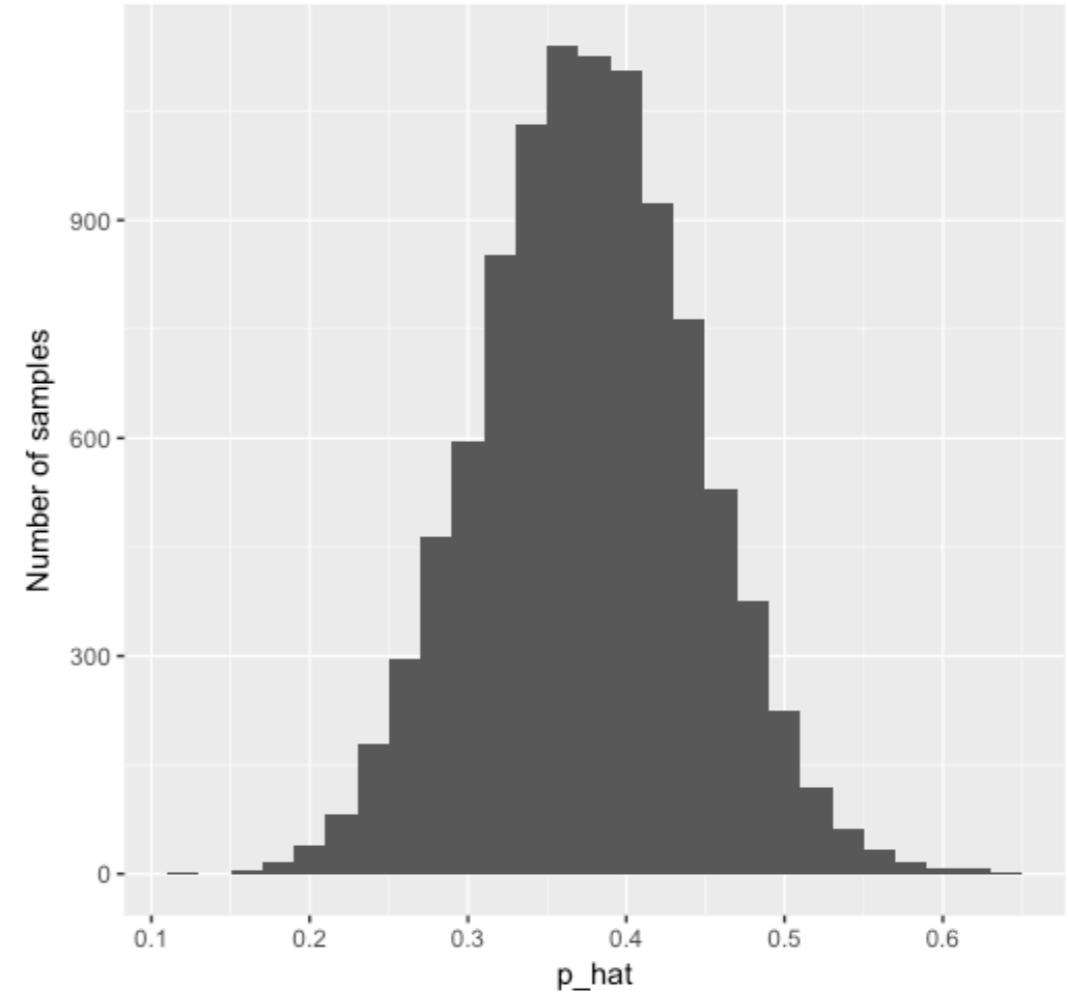


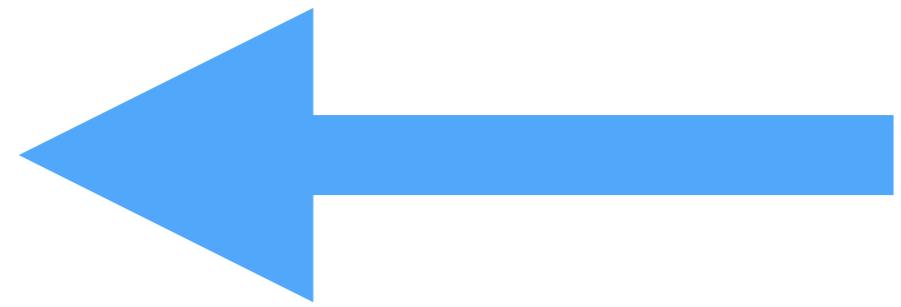
replicate	red	n	prop_red
1	1	18	0.36
2	2	16	0.32
3	3	18	0.36
4	4	16	0.32
5	5	18	0.36
6	6	24	0.48
7	7	17	0.34
8	8	15	0.30
9	9	16	0.32
10	10	18	0.36

Showing 1 to 10 of 10,000 entries



Sampling distribution of p_{hat} based on $n = 50$





Coding

[Cobb \(2015\)](#) argued there are two possible computational engines for statistics:

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

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



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

DataCamp 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 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

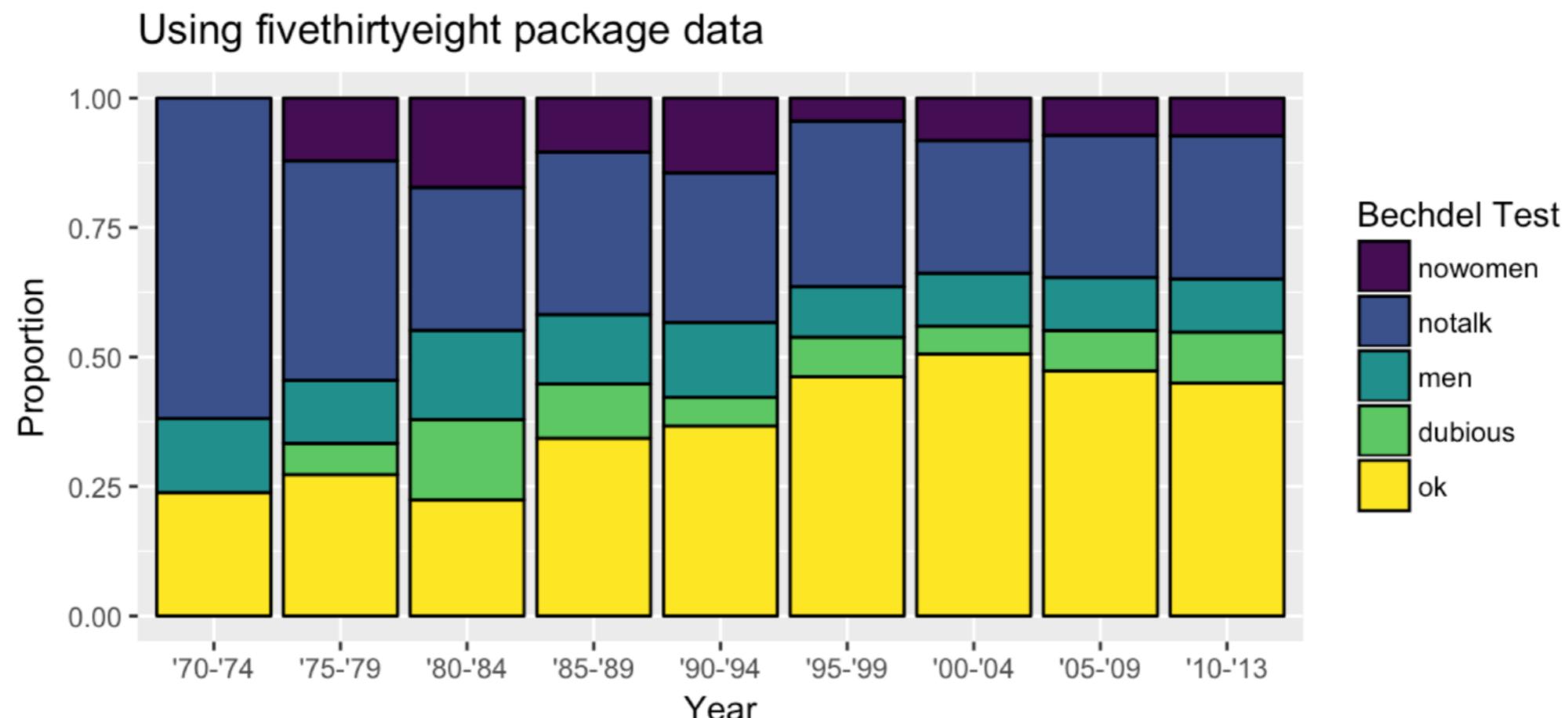




Leverage open source

Open data, such as data in R packages like nycflights13, gapminder, [fivethirtyeight](#)

Bechdel test? Original [538 article](#)



Leverage open source



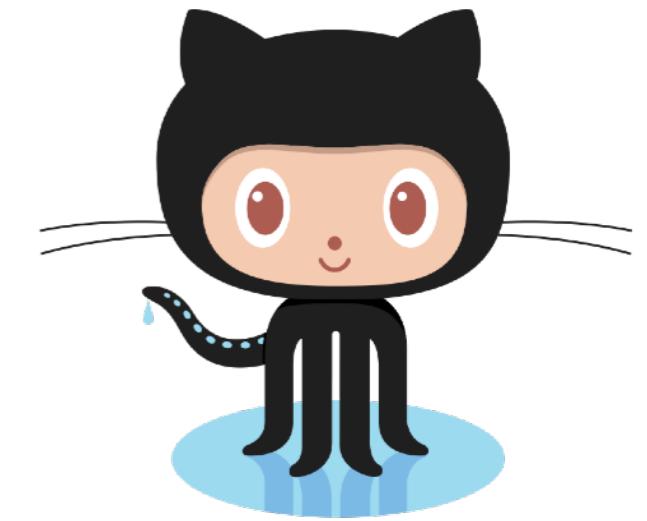
New textbook authoring paradigm



+



+



New textbook authoring paradigm

The screenshot shows the RStudio interface with the following components:

- Top Bar:** RStudio, File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Window, Help.
- Toolbar:** Go to file/function, Addins.
- Code Editor:** The active tab is "03-visualization.Rmd". It contains R code and explanatory text. The explanatory text discusses the `geom_boxplot()` function, its parameters, and how it represents data distribution. It also includes several `r` blocks for learning checks and questions about temperature variability.
- Build Tab:** Shows "All Formats" selected, with "bookdown::gitbook" highlighted.
- File Explorer:** Shows the directory structure of the "moderndive_book" project, listing files like "01-introduction.Rmd", "style.css", and "DESCRIPTION".
- Console:** Displays the command `~/Documents/moderndive/moderndive_book/` followed by a prompt `>`.

New textbook authoring paradigm

Chrome File Edit View History Bookmarks People Window Help

An Introduction to Statistical ... Albert Y.

moderndive.netlify.com/3-viz.html#geompoint

1 Introduction

2 Getting Started with Data in R

I Data Science via the tidyverse

3 Data Visualization via ggplot2

Needed packages

3.1 The Grammar of Graphics

3.2 Five Named Graphs - The 5NG

3.3 5NG#1: Scatterplots

3.3.1 Scatterplots via geom_point

3.3.2 Over-plotting

3.3.3 Summary

3.4 5NG#2: Linegraphs

3.5 5NG#3: Histograms

3.6 Facets

3.7 5NG#4: Boxplots

3.8 5NG#5: Barplots

3.9 Conclusion

4 Tidy Data via tidyr

5 Data Wrangling via dplyr

II Data Modeling via moderndive

6 Basic Regression

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



+



+



“Versions, not editions”

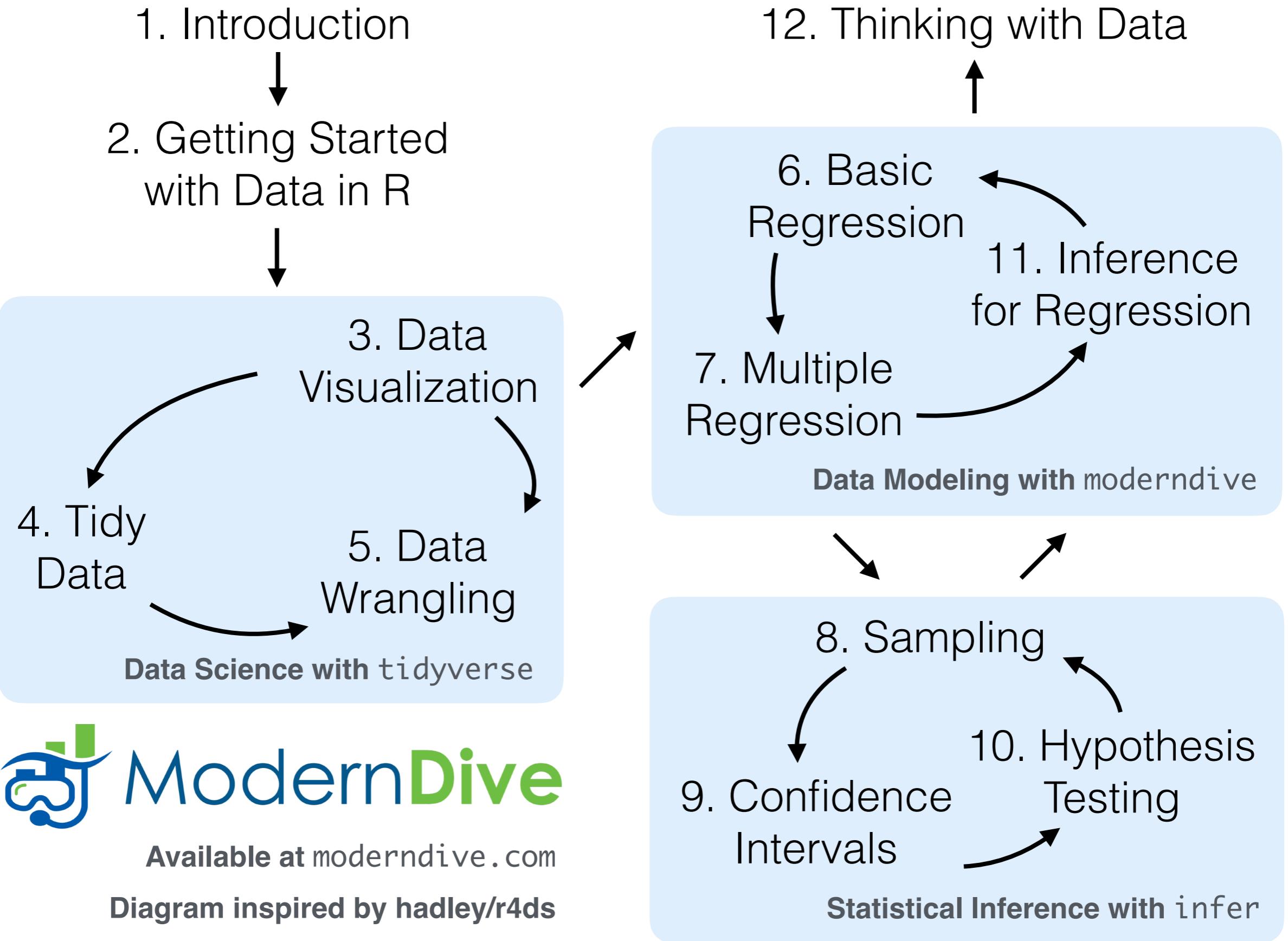
On GitHub at github.com/moderndive/



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?



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

[Reid Hoffman, founder of LinkedIn](#)

Crowdsourcing Typos

stat135-spring-... Albert Y. Kim

Jump to... All Threads Channels # general # graders # moderndive_typos # random # teaching_assistants Direct Messages Apps + Add Apps

#moderndive_typos 112 0 Add a topic

uploaded and commented on this image: [image.png](#)

Tuesday, January 23rd

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!

• Categorical:
 1 2 1

“Categorical”

Wednesday, January 24th

Robin 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

Illyas 1 reply Today at 6:10 PM

Yesterday

Illyas 8:47 AM
uploaded and commented on this image: [Screen Shot 2018-01-25 at 9.46.16 AM.png](#)

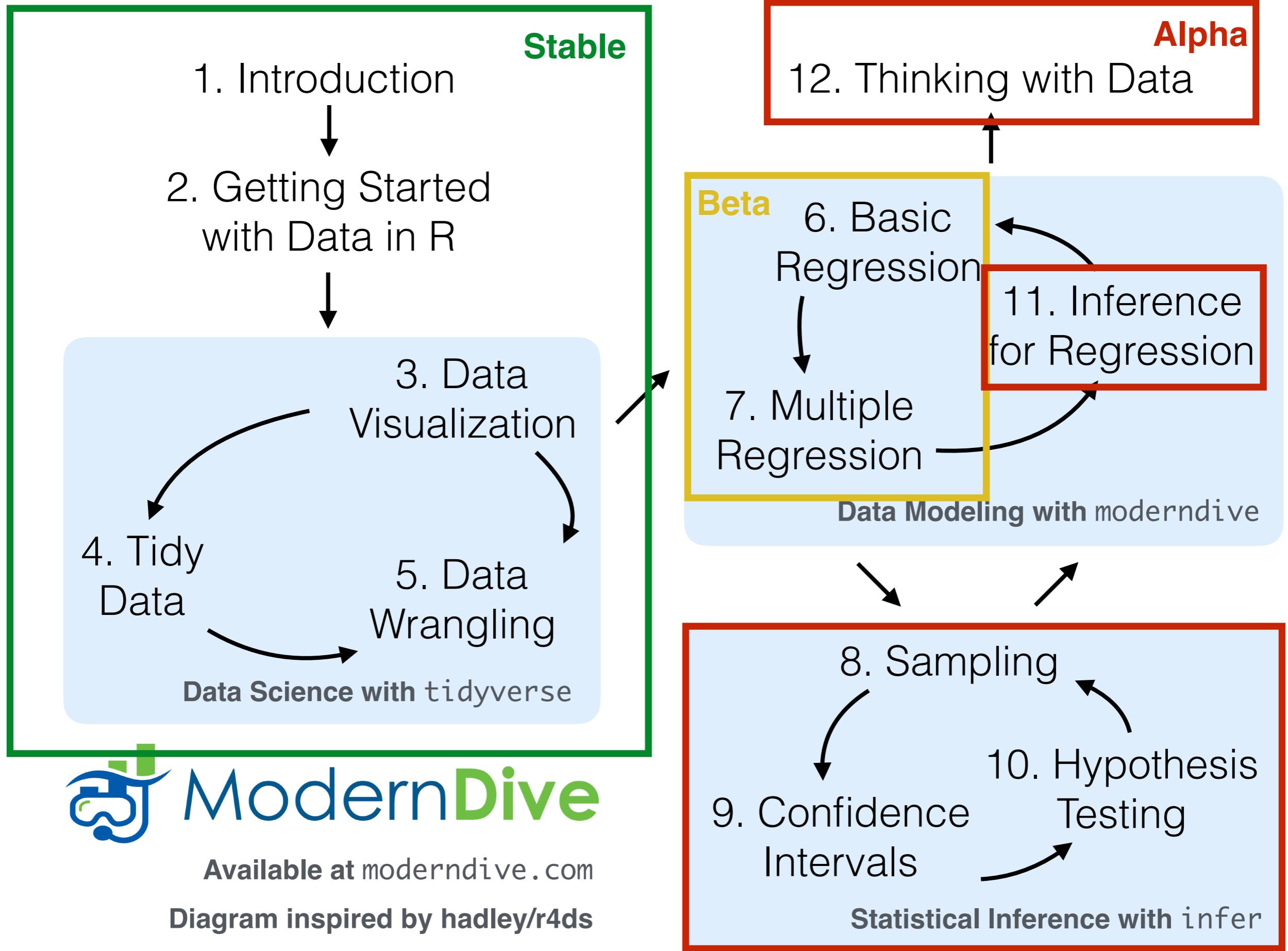
3.1.3 Other components of the Grammar

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](#) (Gromelund 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)

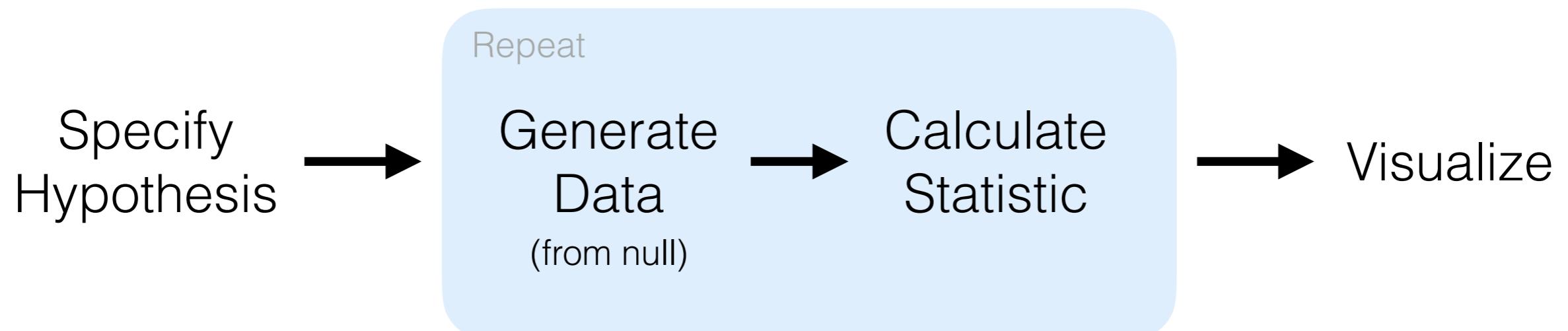
“facetting?”

Search @ ☆



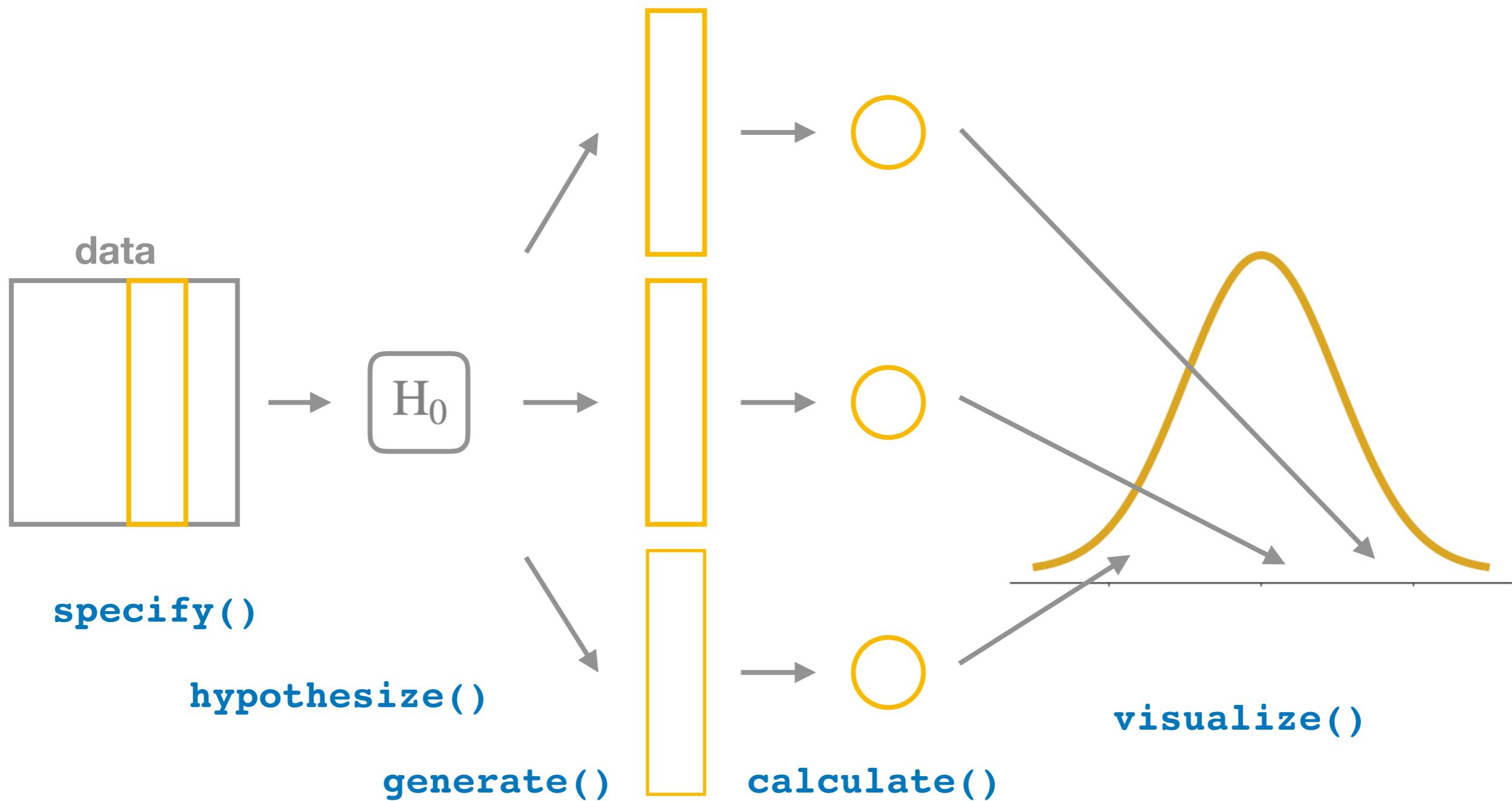
infer package for tidy statistical inference

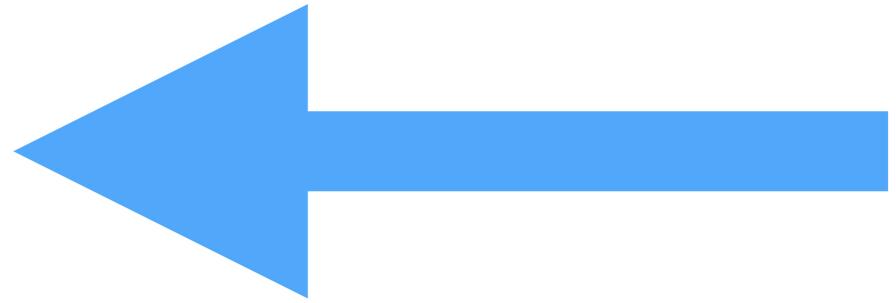
<http://infer.netlify.com/>



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

Hypothesis test







+



+



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



Chester Ismay
DataCamp
Twitter: @old_man_chester
GitHub: ismayc



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