

Something old, something new, something borrowed, something **blue**

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



Dr. Chester Ismay
Data Science Curriculum Lead at DataCamp
GitHub: ismayc
Twitter: @old_man_chester

Slides available at <http://bit.ly/rstudioconf18>

Question

Slides available at <http://bit.ly/rstudioconf18>

Question

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

Question

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

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

Question

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

1. **Data science**: Data visualization, data wrangling, exploratory data analysis, data tidying
2. **Data modeling**: Explanation (causal inference) & prediction (machine learning), correlation

Question

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

1. **Data science**: Data visualization, data wrangling, exploratory data analysis, data tidying
2. **Data modeling**: Explanation (causal inference) & prediction (machine learning), correlation
3. **Statistical inference**: sampling distributions, standard errors, confidence intervals, hypothesis/A/B testing & p-values

General Framework



Emily Robinson

@robinson_es

Following

.@drob's teaching philosophy: have goals for what you want students to do and start them doing it as soon as possible, b/c:
- what you teach first matters
- you should show them why the material is useful [#rstudioconf](#)

Have goals for what you want your students to do,
and start them doing it as early as possible.

Slides available at <http://bit.ly/rstudioconf18>



“A Modern Dive into Data with R”

Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

“A Modern Dive into Data with R”

Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

“A Modern Dive into Data with R”

- Online textbook available at moderndive.com

Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

“A Modern Dive into Data with R”

- Online textbook available at moderndive.com
- Development version at moderndive.netlify.com

Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

“A Modern Dive into Data with R”

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

Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

“A Modern Dive into Data with R”

- Online textbook available at moderndive.com
- Development version at moderndive.netlify.com
- On GitHub at github.com/moderndive/
- Co-authored with [Albert Y. Kim](#)

Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

- Why is this needed?

Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

- Why is this needed?
- How can this help you and your colleagues?

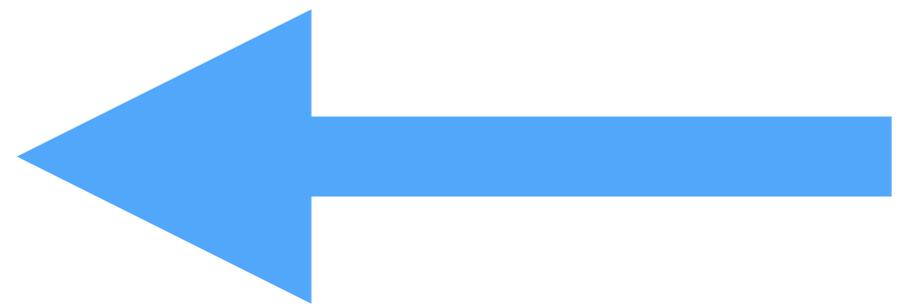
Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

- Why is this needed?
- How can this help you and your colleagues?
- How can you keep up-to-date on development?

Slides available at <http://bit.ly/rstudioconf18>



Analogy: Learning Long Division

Slides available at <http://bit.ly/rstudioconf18>

Analogy: Learning Long Division

Do this a few times:

A hand-drawn diagram of long division on lined paper. The divisor is 6, the dividend is 250, and the quotient is 41. A red arrow points from the digit 4 in the quotient to the digit 2 in the dividend. The remainder is 6.

041
6 / 250
- 24

10
- 6

wikiHow to Do Long Division

Slides available at <http://bit.ly/rstudioconf18>

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. The calculation shows: 6 goes into 25 four times (24), leaving a remainder of 10; then 6 goes into 10 one time (6), leaving a remainder of 4. A red arrow points from the digit 4 in the quotient to the digit 4 in the remainder 10.

041
6 / 250
-24

10
-6

4

wikiHow to Do Long Division

Then rely on this:



ggplot2 via the Grammar of Graphics

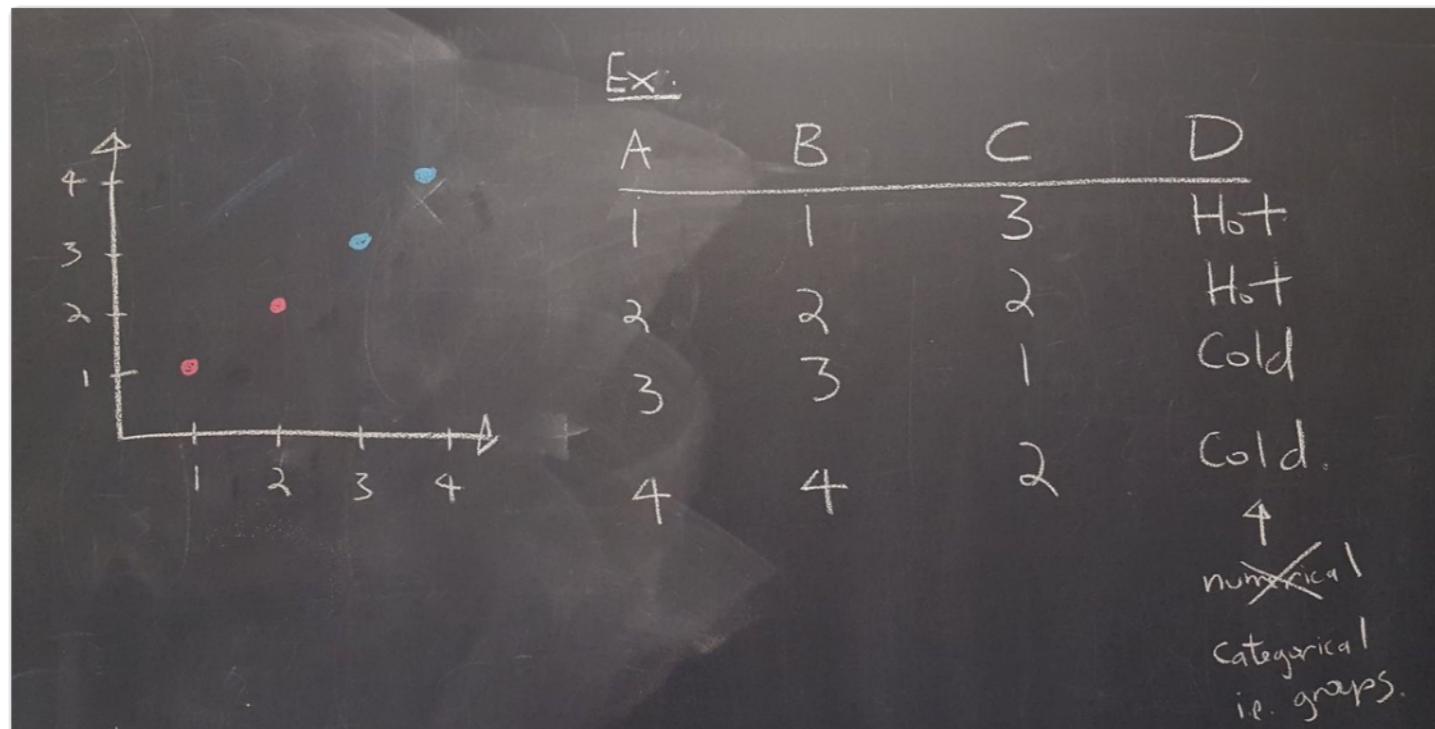


Albert Y. Kim

@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

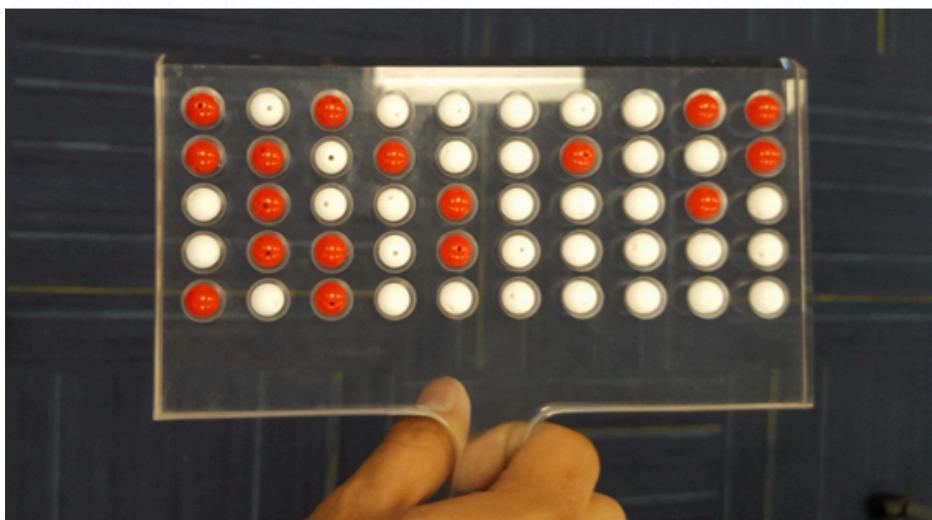
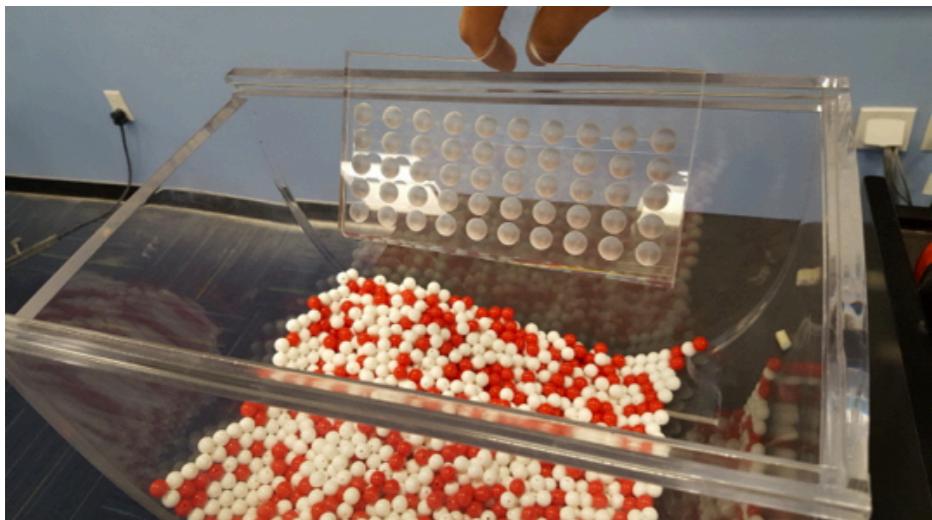


Slides available at <http://bit.ly/rstudioconf18>

Tactile simulation of
sampling to teach
sampling distributions

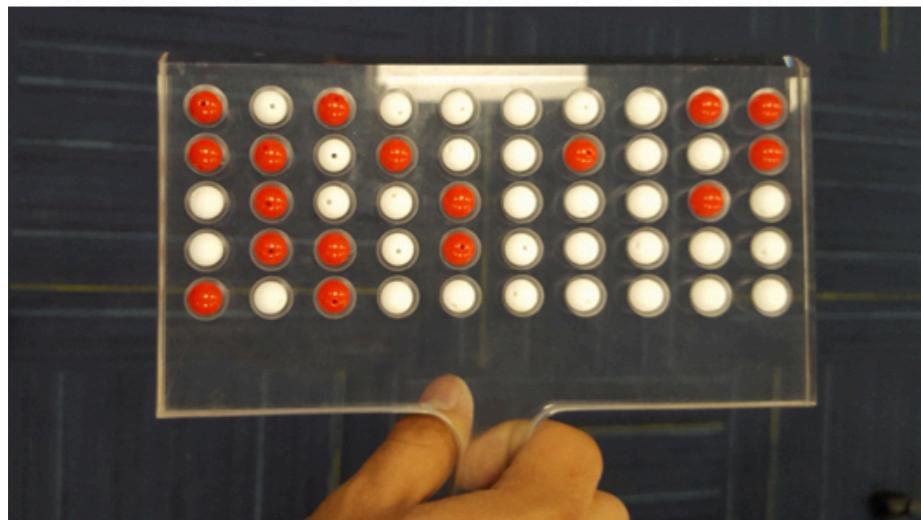
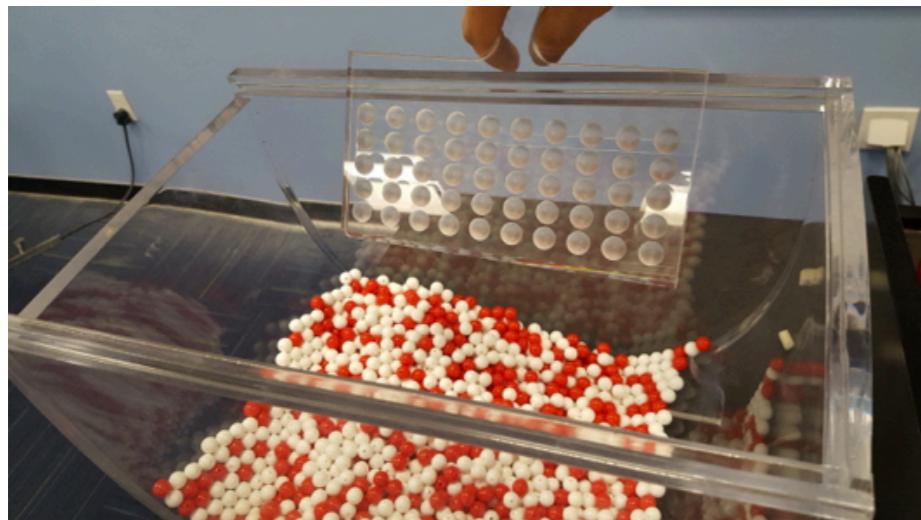
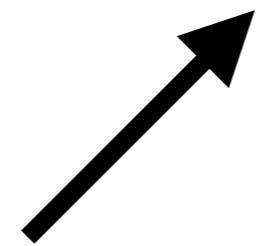
Slides available at <http://bit.ly/rstudioconf18>

Tactile simulation of sampling to teach sampling distributions



Slides available at <http://bit.ly/rstudioconf18>

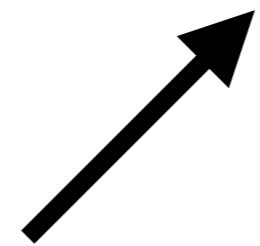
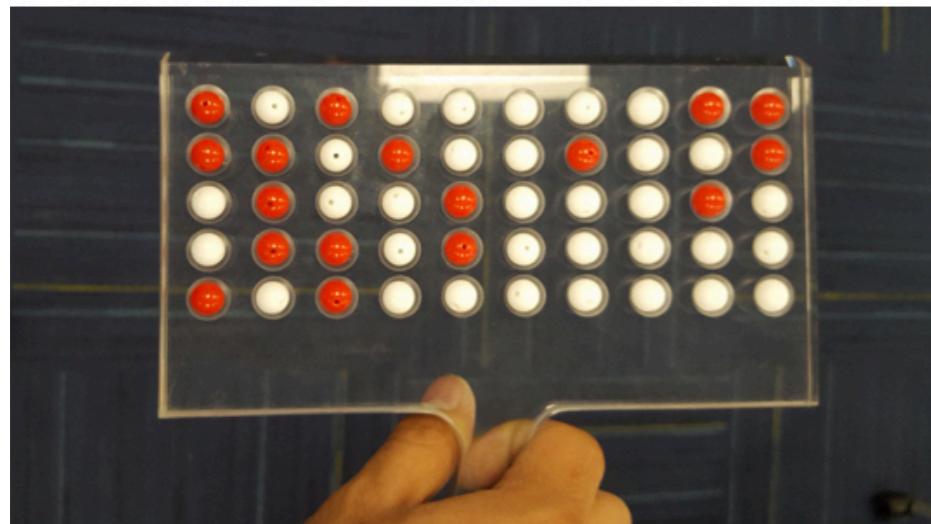
Tactile simulation of sampling to teach sampling distributions



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

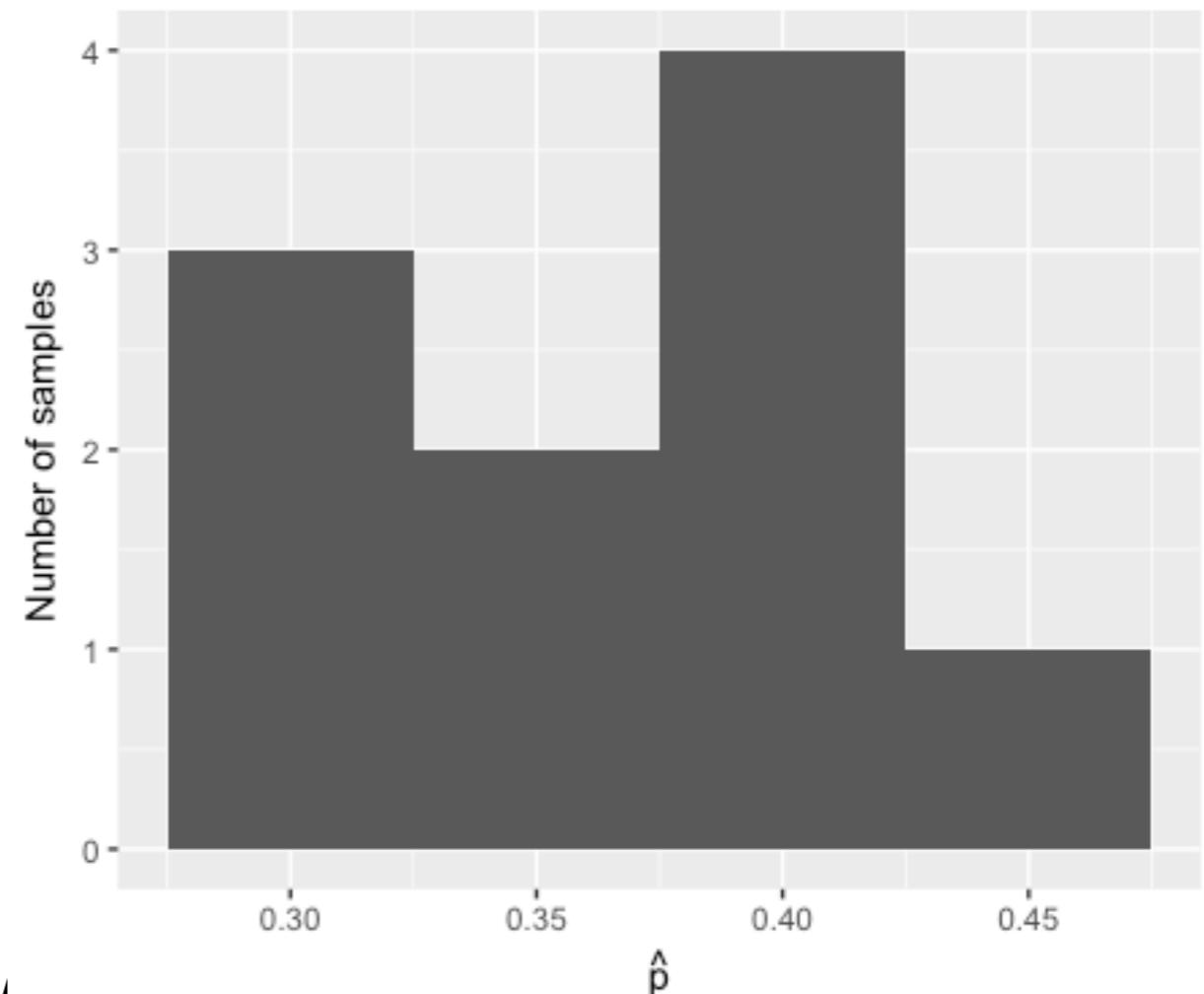
Slides available at <http://bit.ly/rstudioconf18>

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 $n = 50$



Slides available at <http://>

Computer simulation of sampling to teach sampling distributions

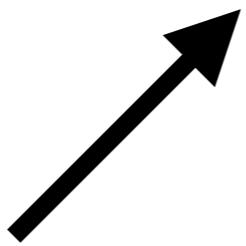
Slides available at <http://bit.ly/rstudioconf18>

Computer simulation of sampling to teach sampling distributions

```
> library(moderndive)
> bowl
# A tibble: 2,400 × 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)
```

Computer simulation of sampling to teach sampling distributions

```
> library(moderndive)
> bowl
# A tibble: 2,400 × 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)
```

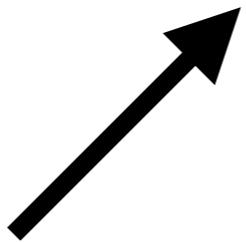


	replicate	red	n	prop_red
1	1	18	50	0.36
2	2	16	50	0.32
3	3	18	50	0.36
4	4	16	50	0.32
5	5	18	50	0.36
6	6	24	50	0.48
7	7	17	50	0.34
8	8	15	50	0.30
9	9	16	50	0.32
10	10	18	50	0.36

Showing 1 to 10 of 10,000 entries

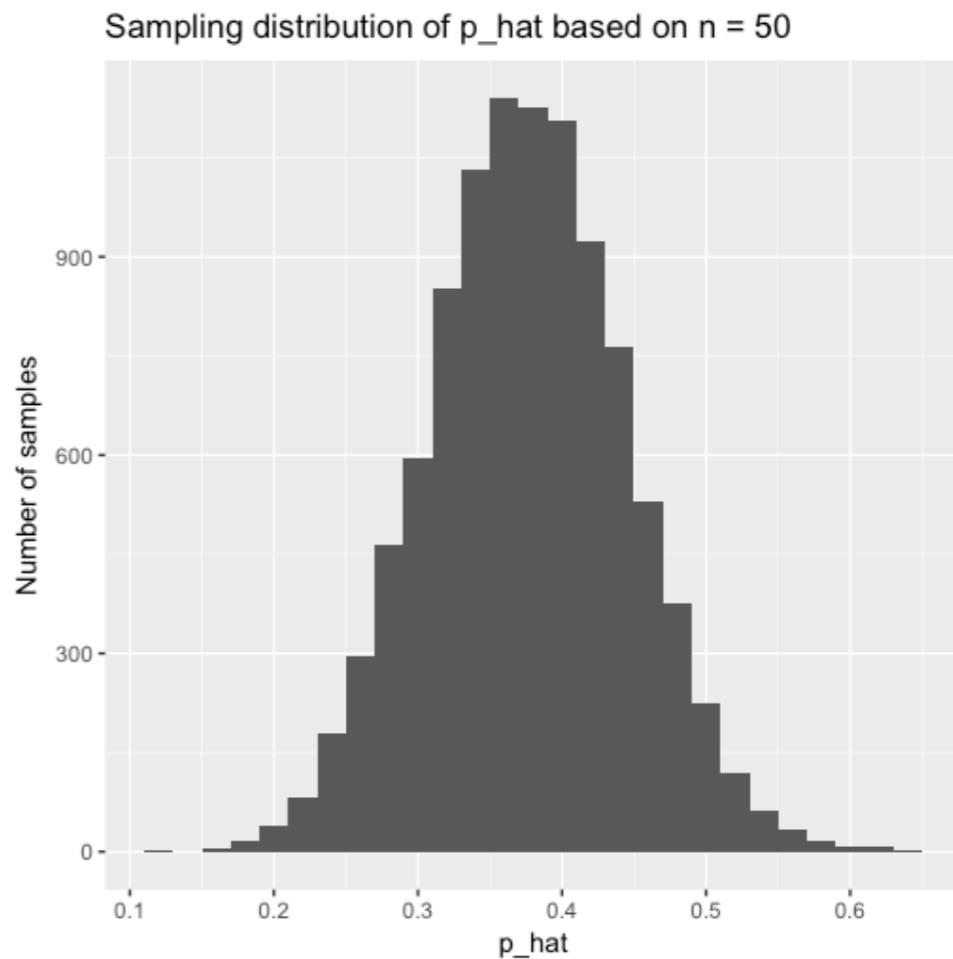
Computer simulation of sampling to teach sampling distributions

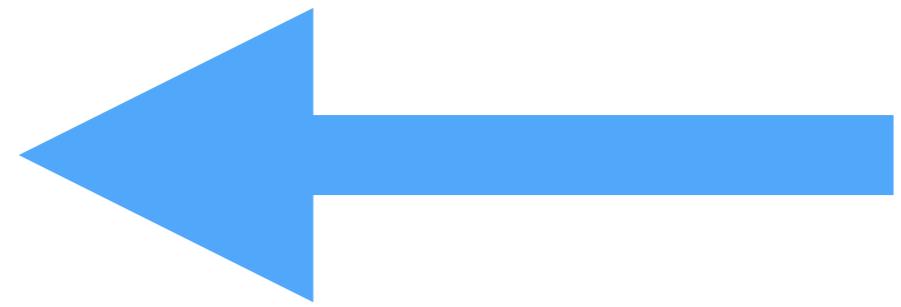
```
> library(moderndive)
> bowl
# A tibble: 2,400 × 2
  ball_ID color
  <int> <chr>
1      1 white
2      2 white
3      3 white
4      4 red
5      5 white
6      6 white
7      7 red
8      8 white
9      9 red
10     10 white
# ... with 2,390 more rows
> bowl %>%
  rep_sample_n(size = 50, reps =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





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 that students can get to the advanced methods later....

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 that students can get to the advanced methods later....

DataCamp

DataCamp: Immediate Feedback

Slides available at <http://bit.ly/rstudioconf18>

DataCamp: Immediate Feedback

- Students can practice failing, but with support.

Slides available at <http://bit.ly/rstudioconf18>

DataCamp: Immediate Feedback

- Students can practice failing, but with support.
- Difference with Coursera & Udacity?

Slides available at <http://bit.ly/rstudioconf18>

DataCamp: Immediate Feedback

- Students can practice failing, but with support.
- Difference with Coursera & Udacity?
- DataCamp will pick off low hanging fruit. Ex:

Slides available at <http://bit.ly/rstudioconf18>

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

DataCamp: Immediate Feedback

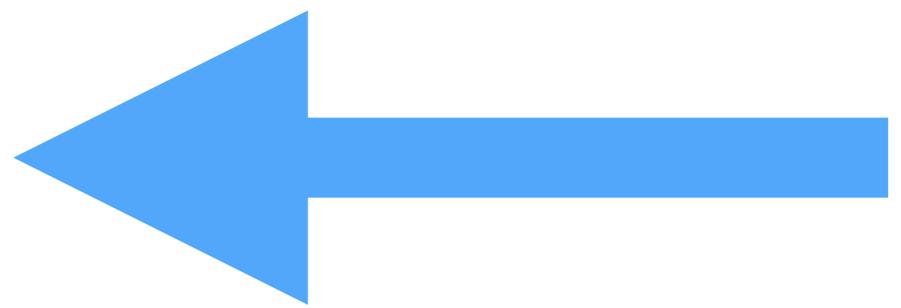
- Students can practice failing, but with support.
- Difference with Coursera & Udacity?
- DataCamp will pick off low hanging fruit. Ex:
 1. Matching parentheses
 2. Variable name misspellings

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

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 the “[Curse of knowledge](#)”



Leverage open source

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

Leverage open source

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

Bechdel test?

Leverage open source

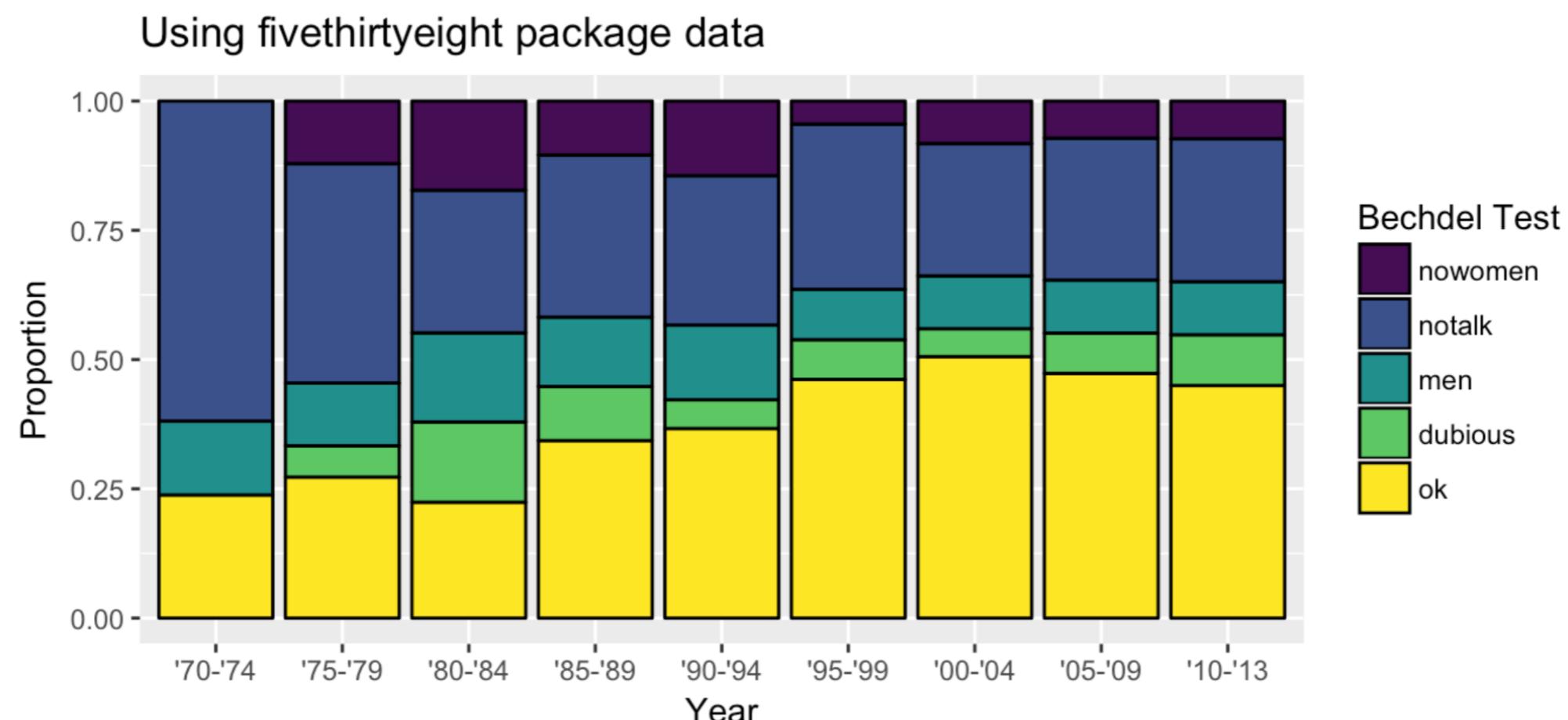
Open data, such as data in R packages like
`nycflights13`, `gapminder`, [fivethirtyeight](#)

Bechdel test? Original [538 article](#)

Leverage open source

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

Bechdel test? Original [538 article](#)



Slides available at <http://bit.ly/rstudioconf18>

New textbook authoring paradigm

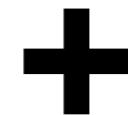
Slides available at <http://bit.ly/rstudioconf18>

New textbook authoring paradigm



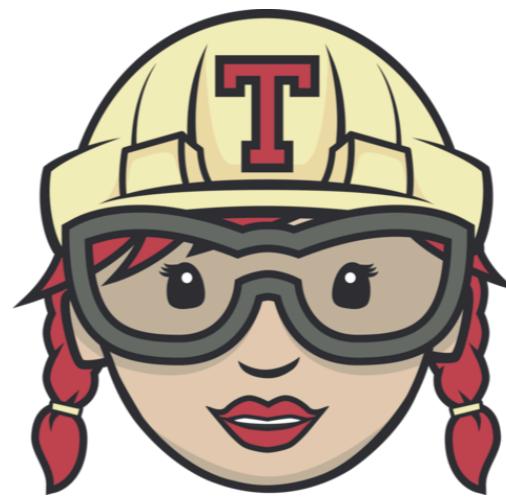
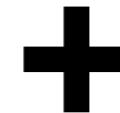
Slides available at <http://bit.ly/rstudioconf18>

New textbook authoring paradigm



Slides available at <http://bit.ly/rstudioconf18>

New textbook authoring paradigm

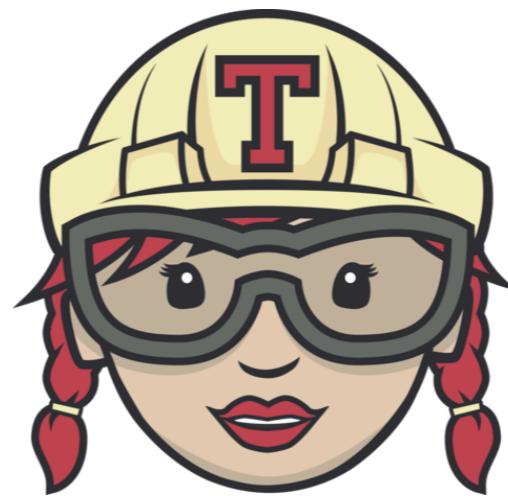


Slides available at <http://bit.ly/rstudioconf18>

New textbook authoring paradigm



+



Slides available at <http://bit.ly/rstudioconf18>

New textbook authoring paradigm

The screenshot shows the RStudio interface with the following details:

- Top Bar:** RStudio, File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Window, Help.
- Toolbar:** Go to file/function, Addins.
- File Path:** ~/Documents/moderndive/moderndive_book - master - RStudio
- Code Editor:** The file 03-visualization.Rmd is open. The code includes R code for creating a boxplot and learning checks. A note explains the use of the `factor()` function to convert a discrete value like `month` into a categorical variable.
- Build Tab:** Set to "All Formats" and "bookdown::gitbook".
- File Explorer:** Shows the directory structure of the book project, including files like style.css, Rproj files, and various chapters (01-12) and appendixes (91-99).
- Console:** Shows the command ~ /Documents/moderndive/moderndive_book/

Slides available at <http://bit.ly/rstudioconf18>

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 tidy

5 Data Wrangling via dplyr

II Data Modeling via moderndive

6 Basic Regression

ggsave("boxplot.png")

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

Figure 3.13: Month by temp boxplot

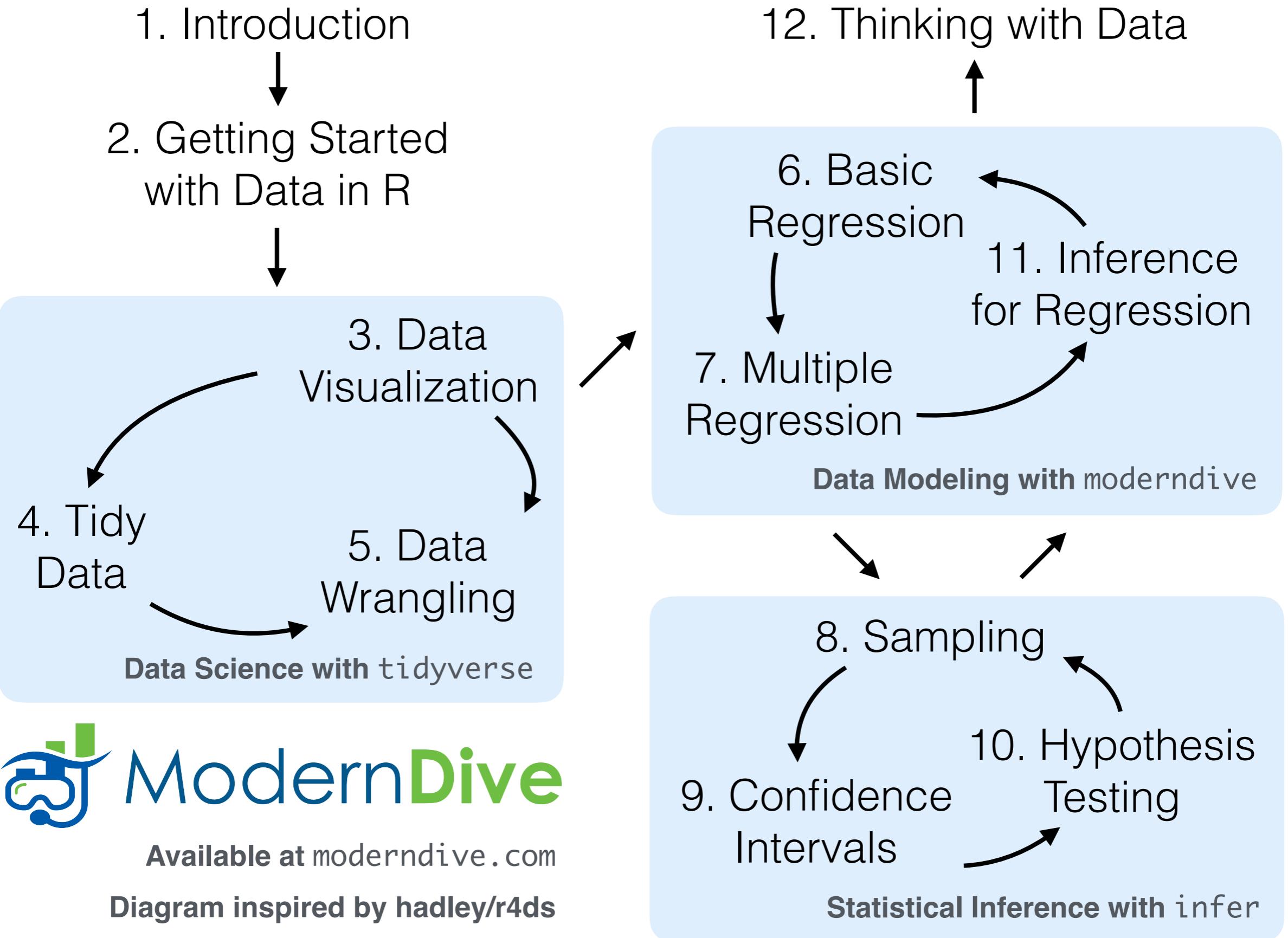
We have introduced a new function called `factor()` here. One of the things this function does is to convert a discrete value like `month` (1, 2, ..., 12) into a categorical variable. The “box” part of this plot represents the 25th percentile, the median (50th percentile), and the 75th percentile. The 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.

Slides available at <http://bit.ly/rstudioconf18>

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

[Reid Hoffman, founder of LinkedIn](#)

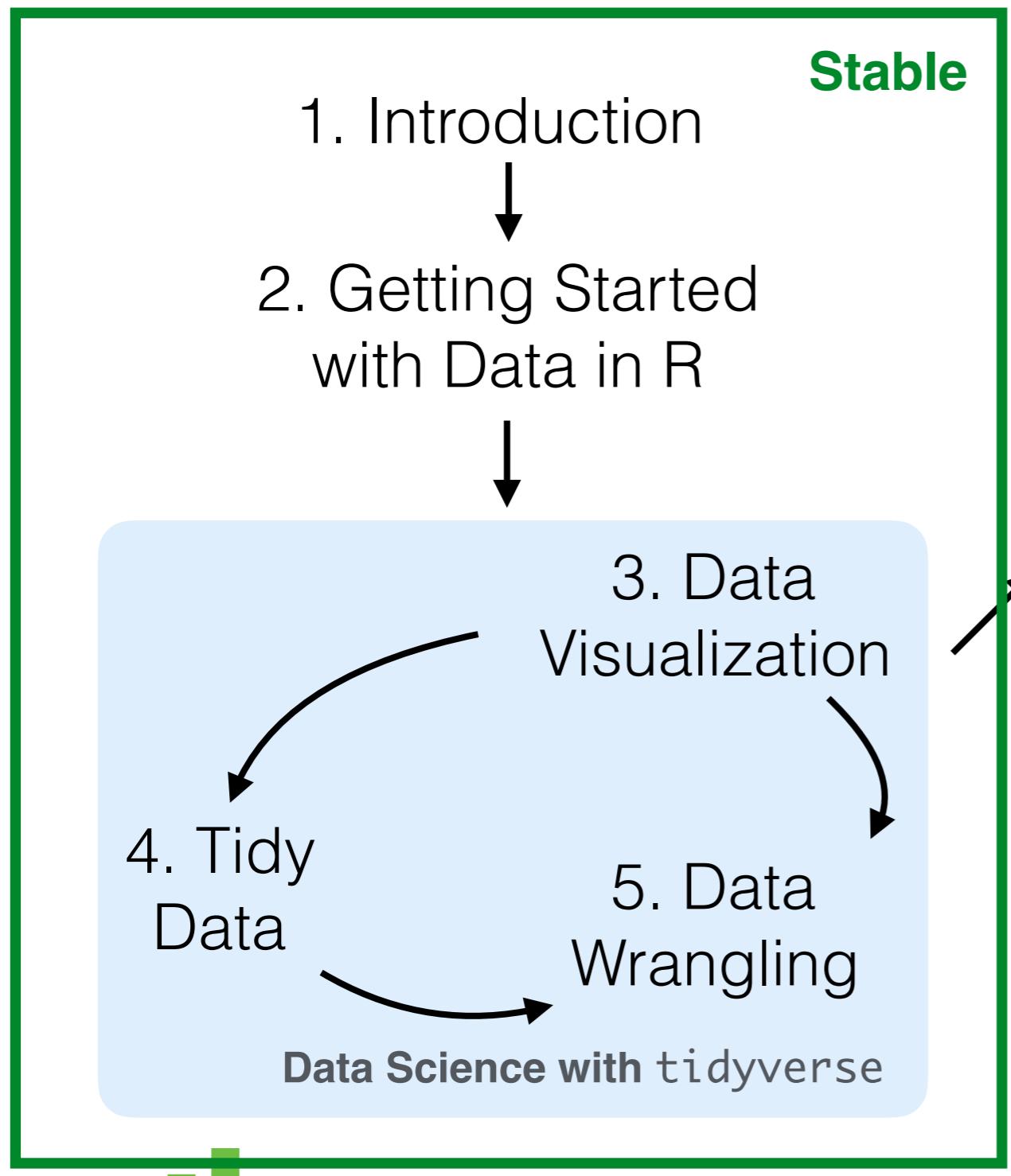
Slides available at <http://bit.ly/rstudioconf18>



Available at moderndive.com

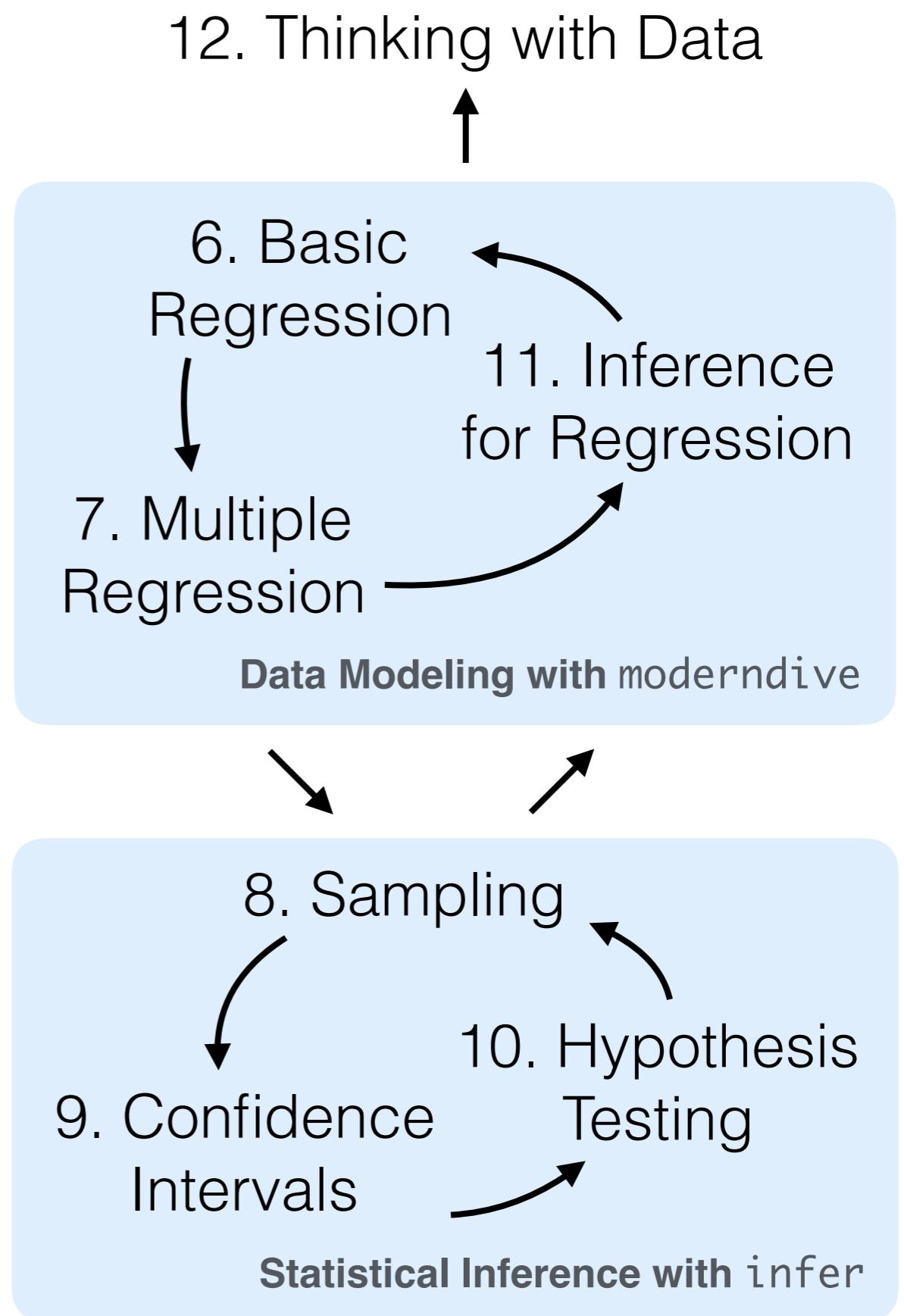
Diagram inspired by hadley/r4ds

Slides available at <http://bit.ly/rstudioconf18>

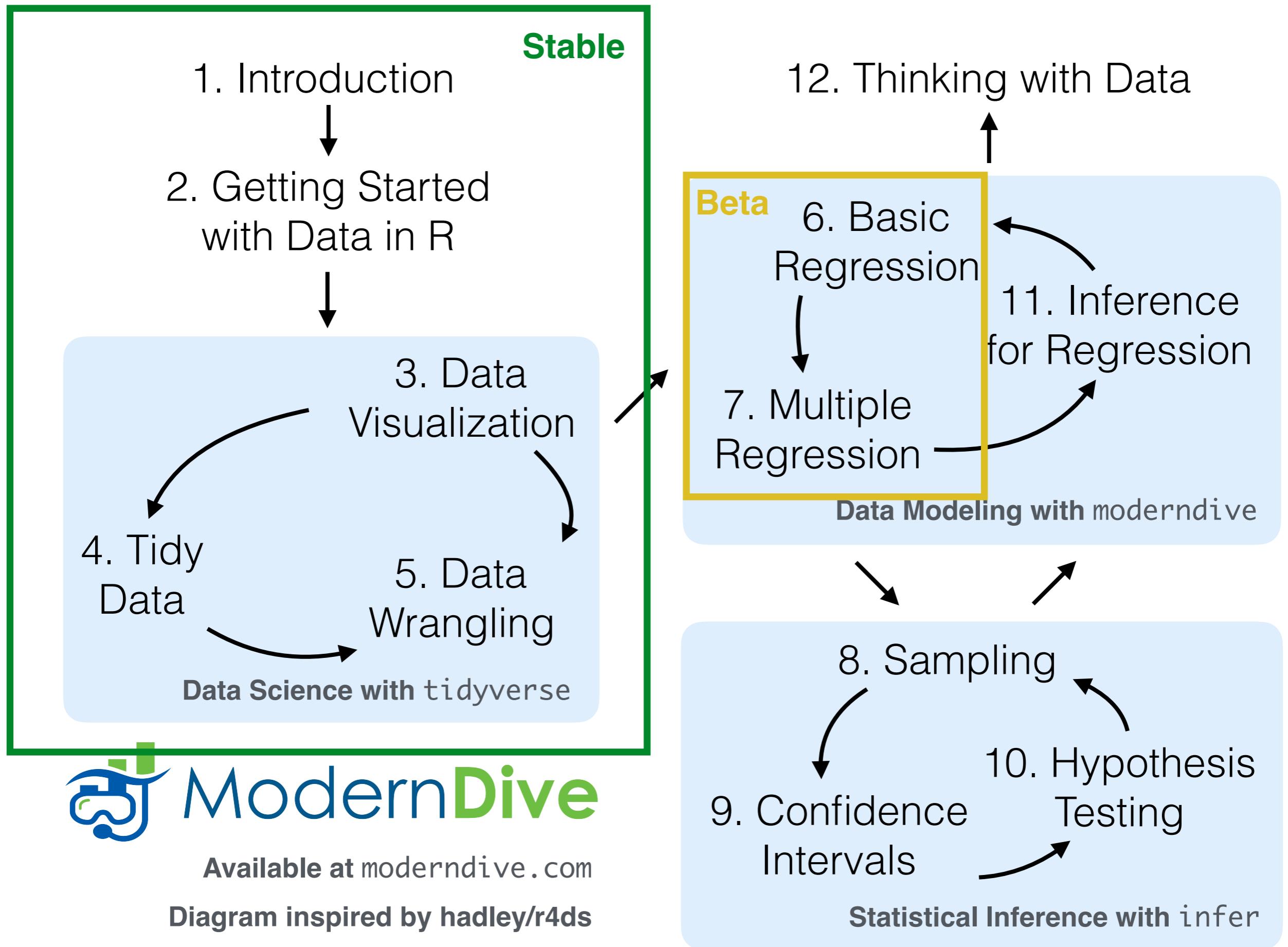


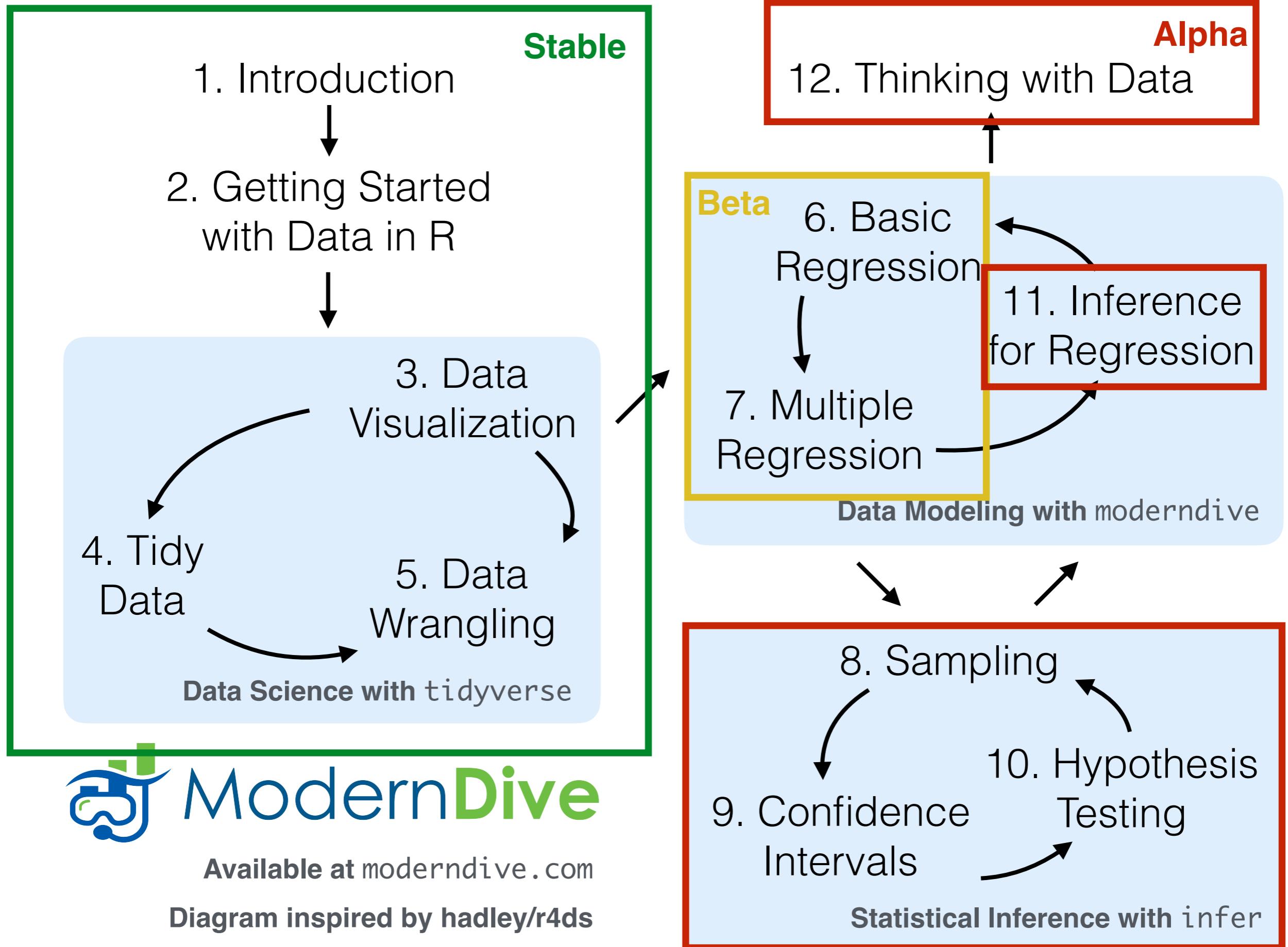
Available at moderndive.com

Diagram inspired by [hadley/r4ds](#)



Slides available at <http://bit.ly/rstudioconf18>





Available at moderndive.com

Diagram inspired by hadley/r4ds

Slides available at <http://bit.ly/rstudioconf18>

infer package for tidy statistical inference

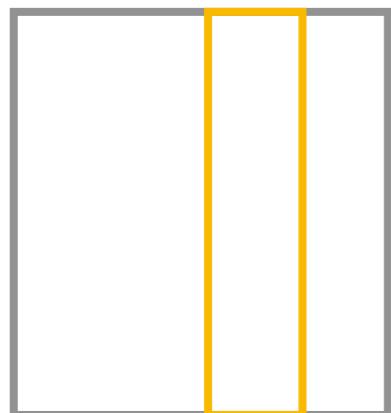
<http://infer.netlify.com/>

Slides available at <http://bit.ly/rstudioconf18>

infer package for tidy statistical inference

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

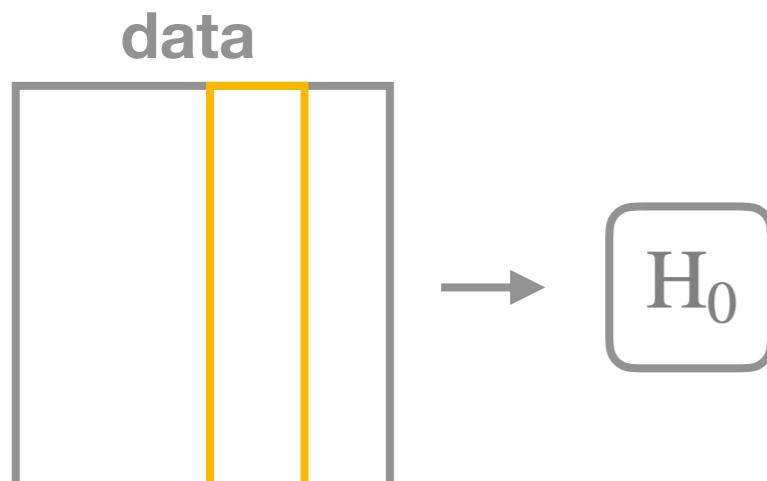
data



specify()

infer package for tidy statistical inference

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

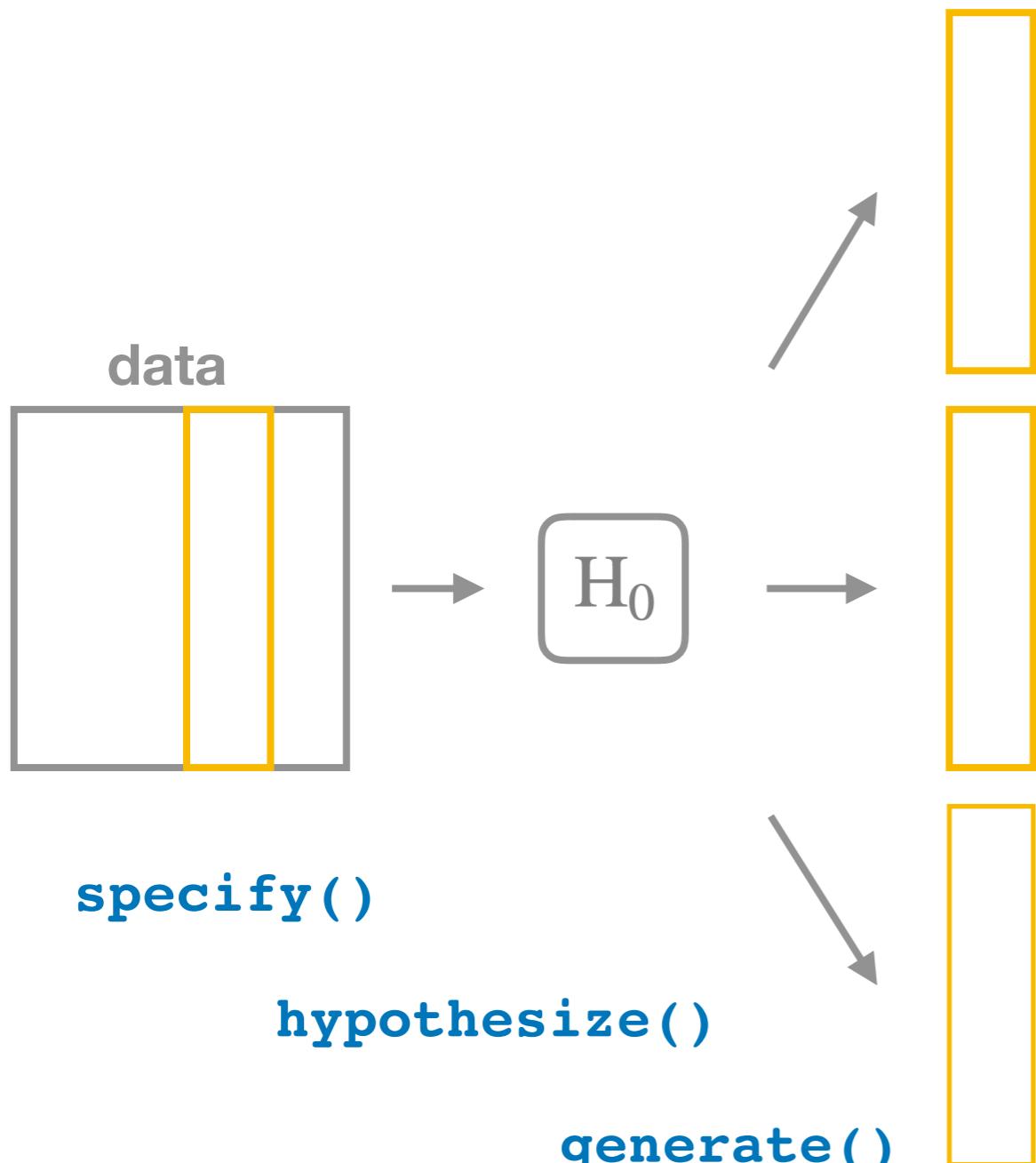


specify()

hypothesize()

infer package for tidy statistical inference

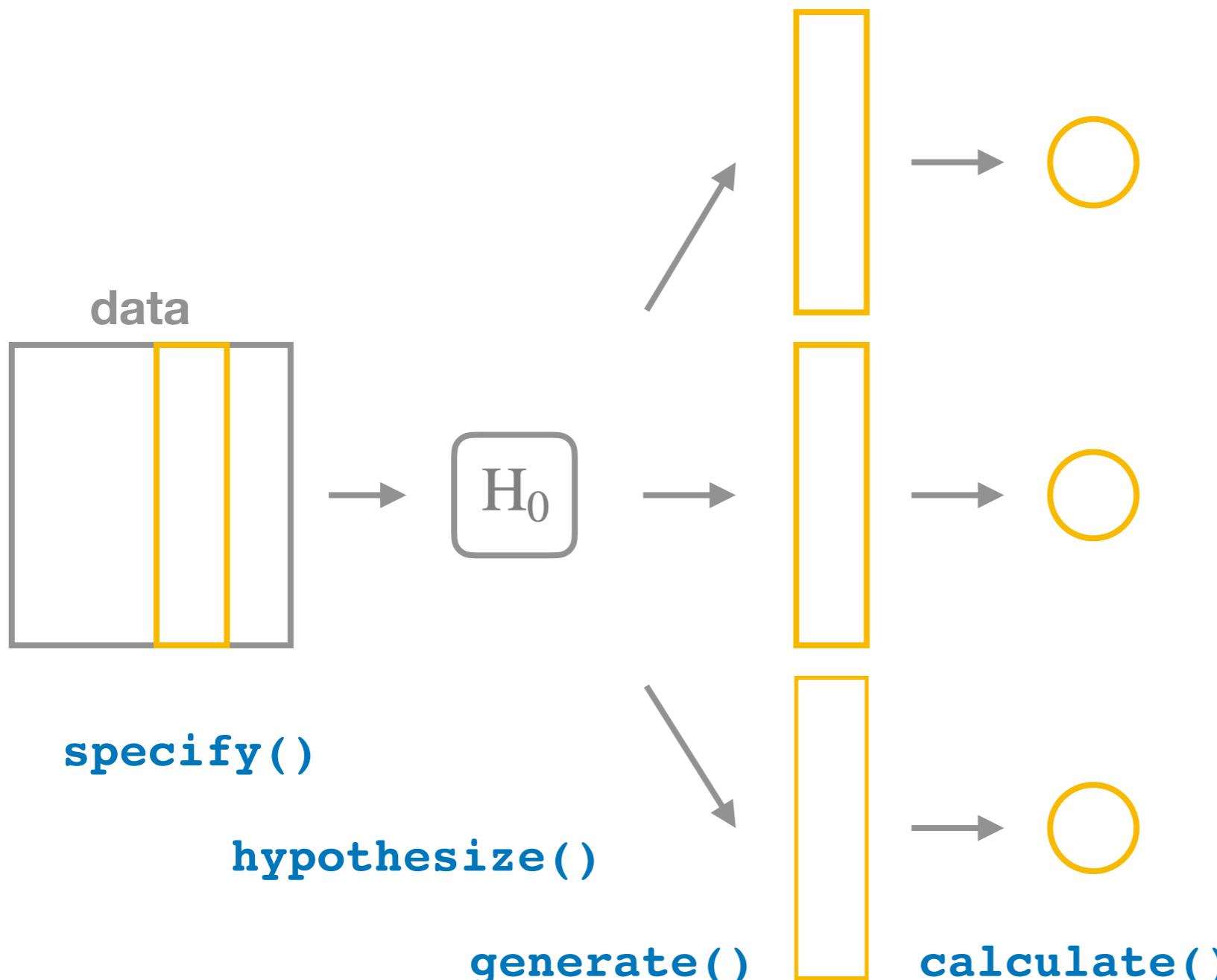
<http://infer.netlify.com/>



Slides available at <http://bit.ly/rstudioconf18>

infer package for tidy statistical inference

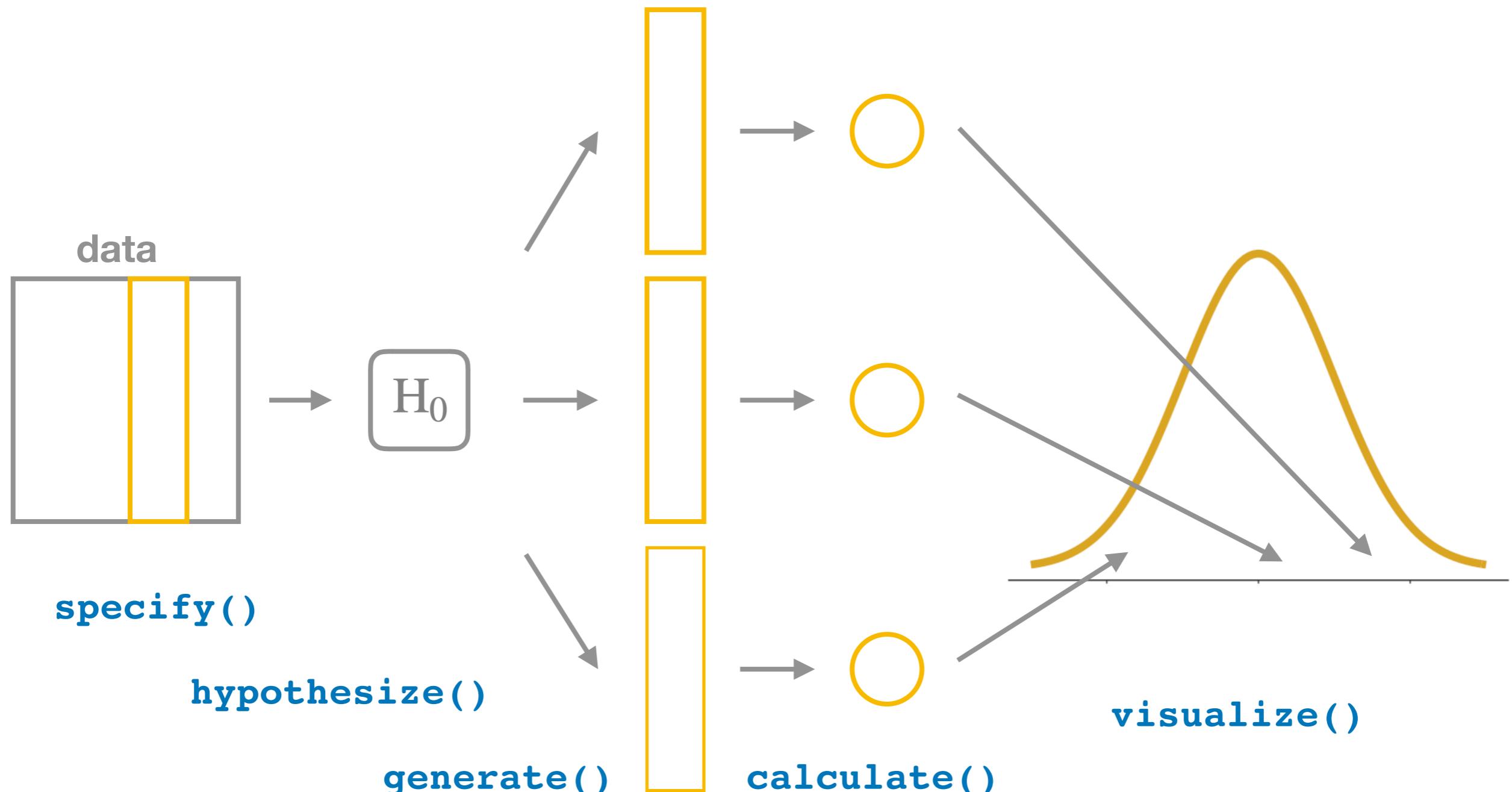
<http://infer.netlify.com/>



Slides available at <http://bit.ly/rstudioconf18>

infer package for tidy statistical inference

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



Slides available at <http://bit.ly/rstudioconf18>

infer package for tidy statistical inference

Slides available at <http://bit.ly/rstudioconf18>

infer package for tidy statistical inference



Lucy 🌻
@LucyStats

Following

Hearing Andrew Bray discuss his (& @old_man_chester @BaumerBen @minebocek) #rstats infer 📦. Building statistical inference in a tidy & transparent way, using the verbs:



specify()



hypothesize()



generate()



calculate()



visualize()

infer.netlify.com #rstudioconf

1:03 PM - 2 Feb 2018

23 Retweets 75 Likes



2



23



75



Slides available at <http://bit.ly/rstudioconf18>

“Thinking with Data”

Example student work

- Analysis of crime in [Chicago](#)
- How many [f**ks](#) does Tarantino give?
- Final projects: [Code and data](#)

New textbook authoring paradigm

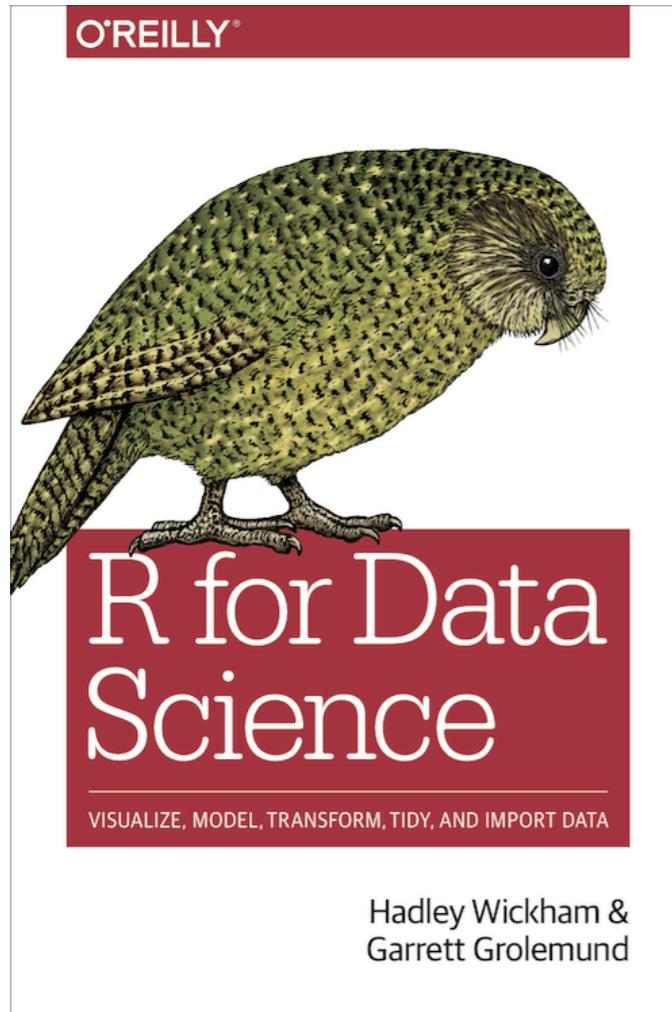
Slides available at <http://bit.ly/rstudioconf18>

New textbook authoring paradigm

“Versions, not editions”

Slides available at <http://bit.ly/rstudioconf18>

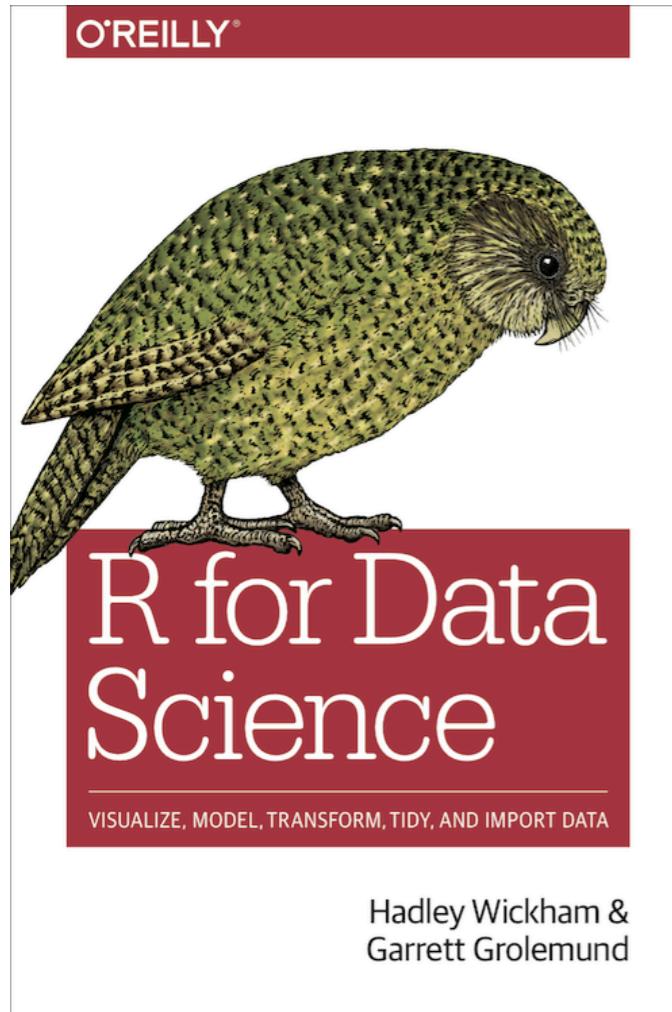
New textbook authoring paradigm



“Versions, not editions”

Slides available at <http://bit.ly/rstudioconf18>

New textbook authoring paradigm

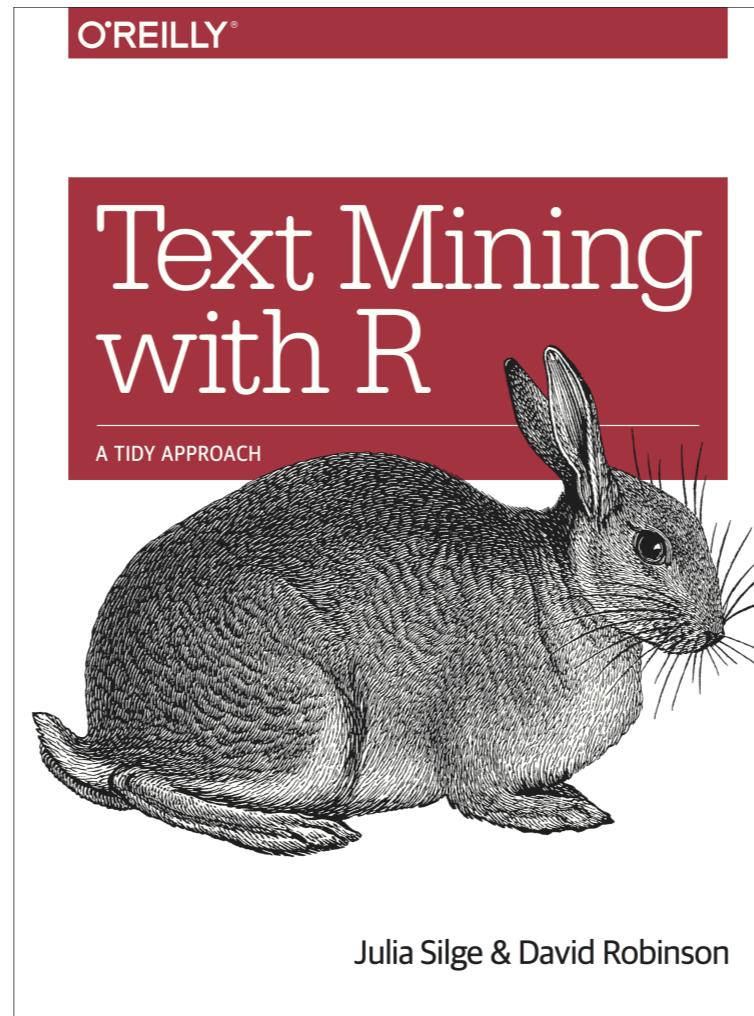
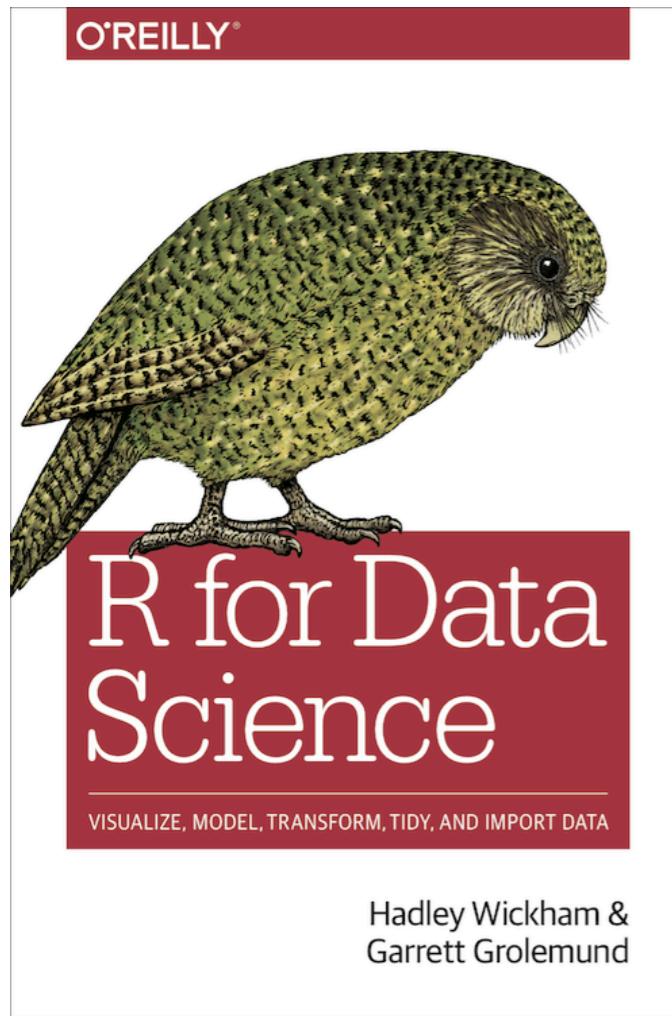


tidyverse

“Versions, not editions”

Slides available at <http://bit.ly/rstudioconf18>

New textbook authoring paradigm

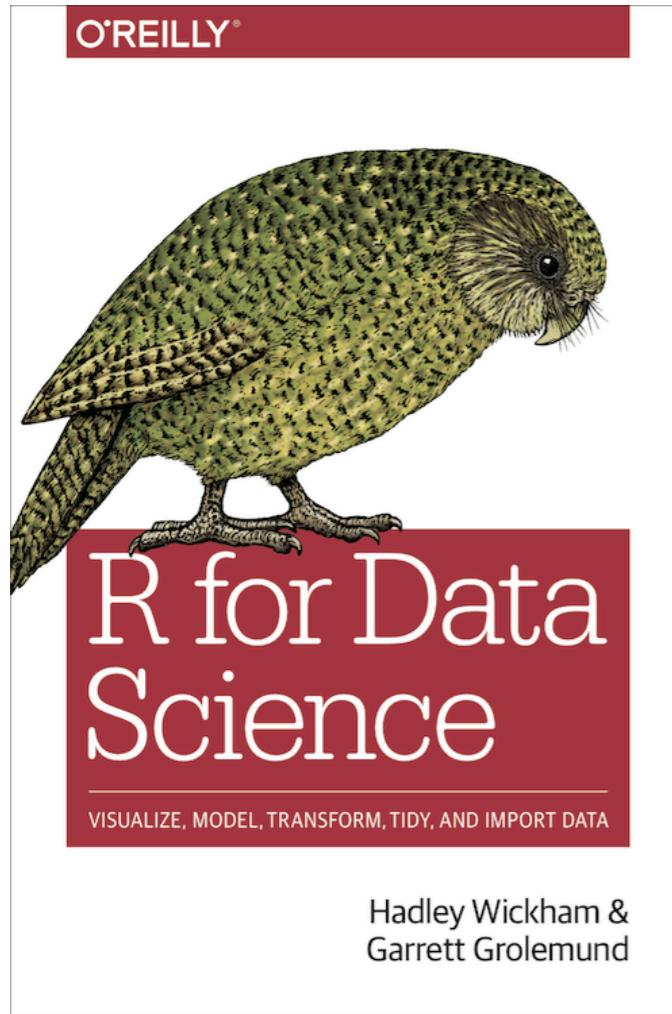


tidyverse

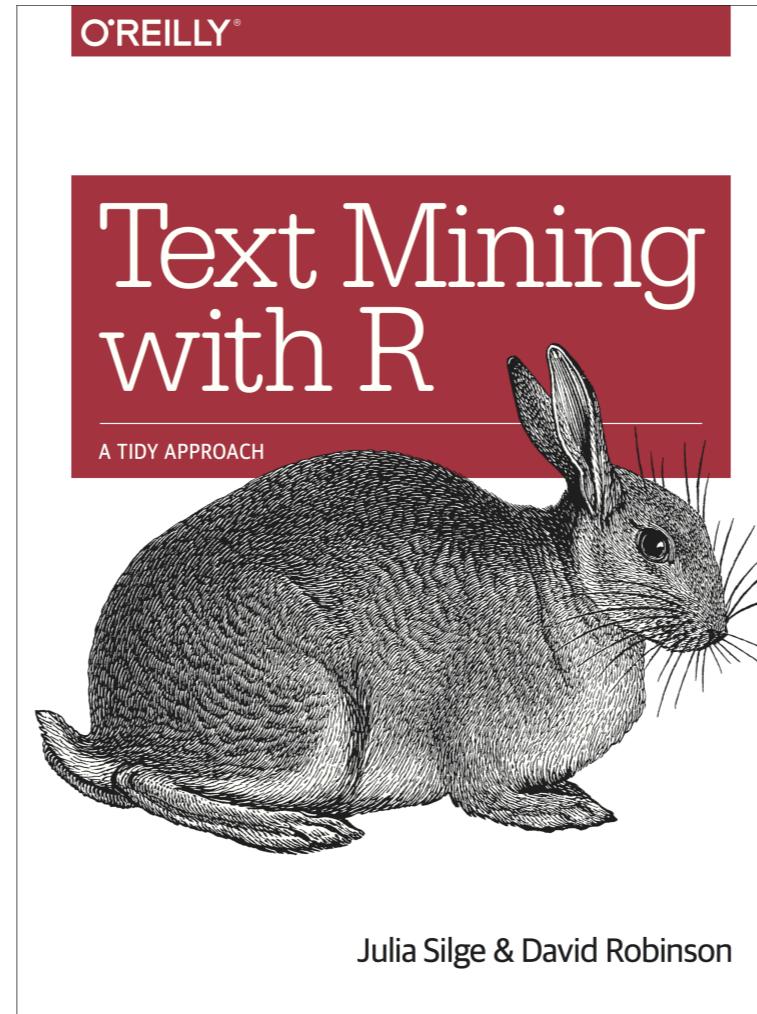
“Versions, not editions”

Slides available at <http://bit.ly/rstudioconf18>

New textbook authoring paradigm



tidyverse

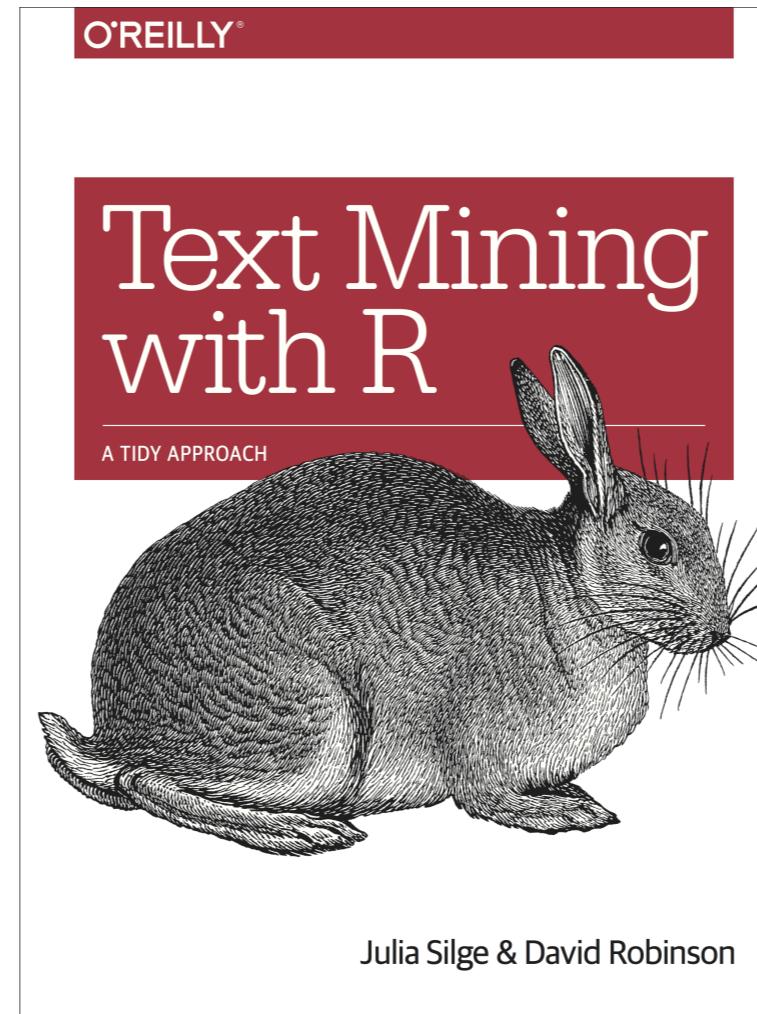
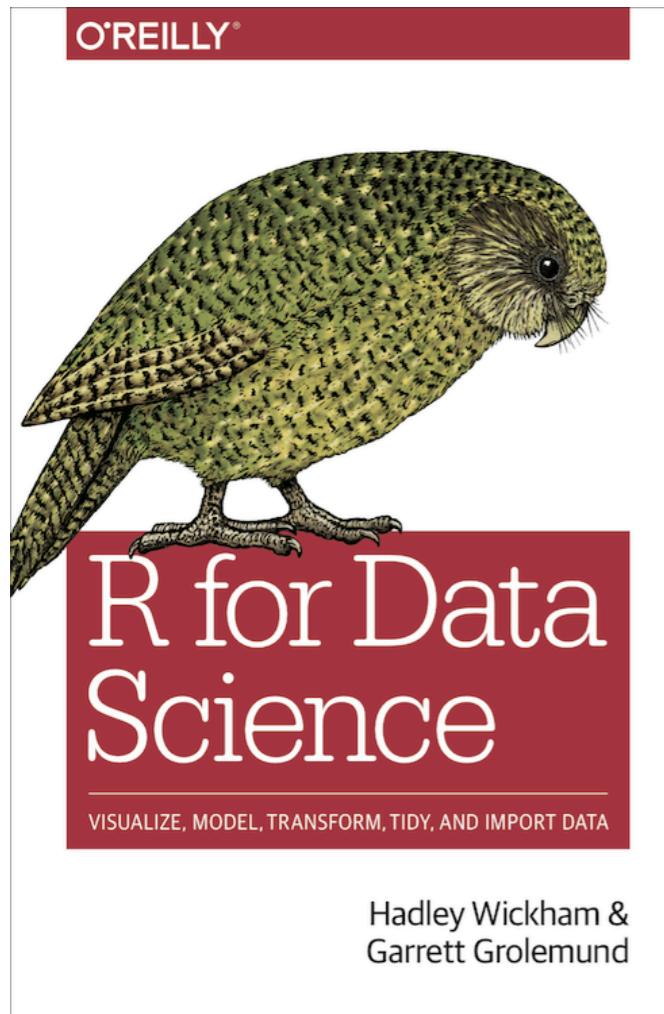


tidytext

“Versions, not editions”

Slides available at <http://bit.ly/rstudioconf18>

New textbook authoring paradigm



ModernDive



An Introduction to Statistical
and Data Sciences via



Chester Ismay
Albert Y. Kim

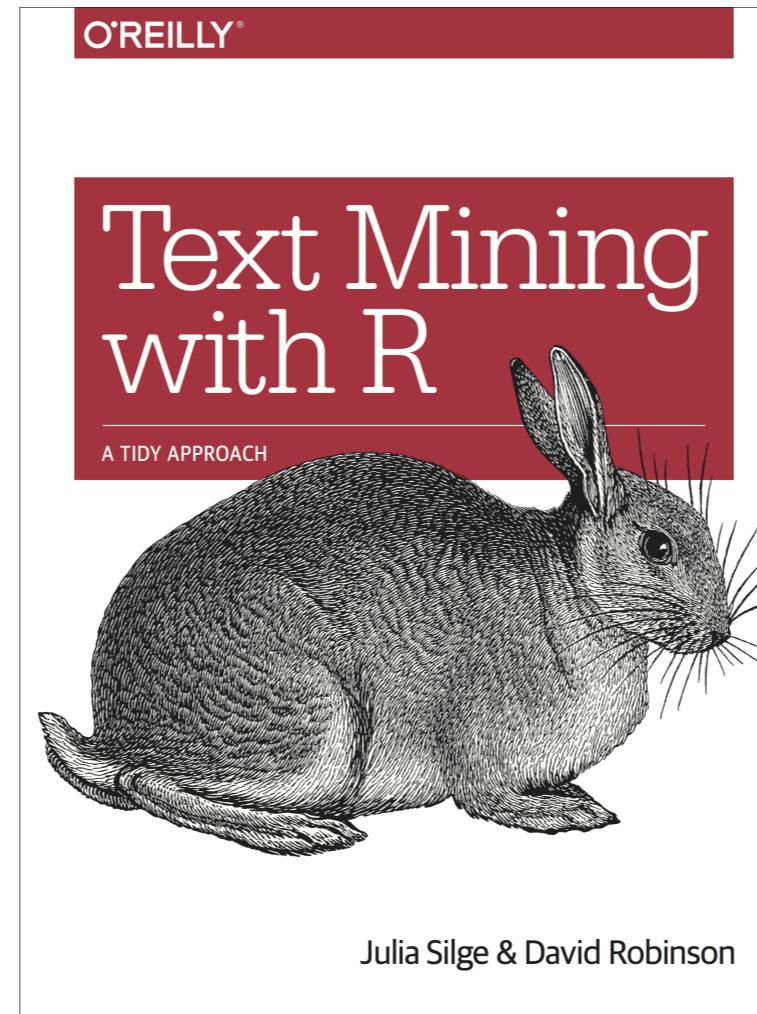
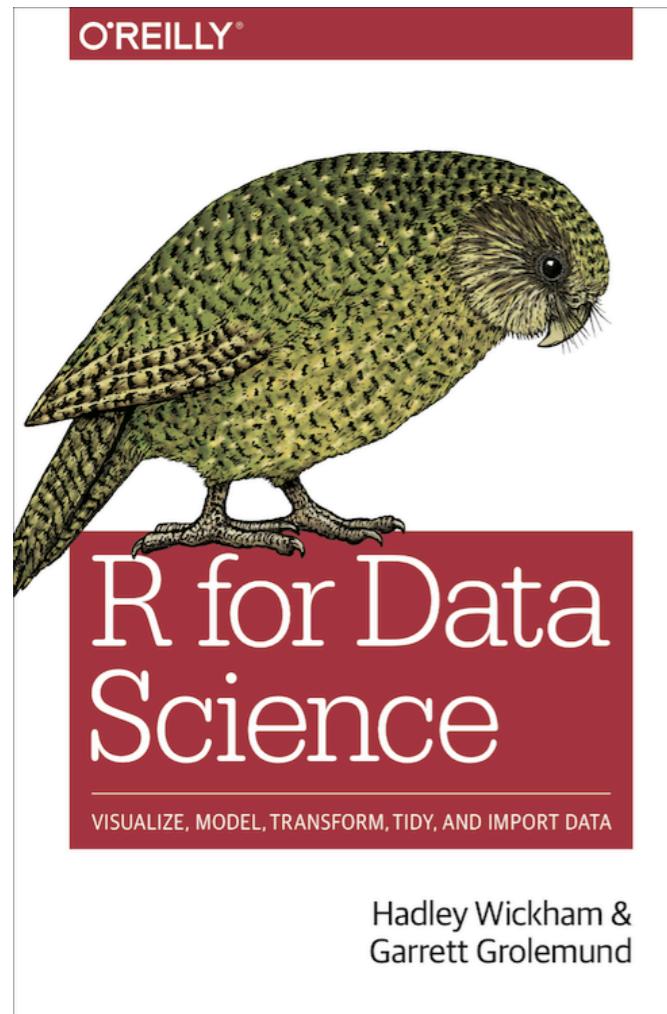
tidyverse

tidytext

“Versions, not editions”

Slides available at <http://bit.ly/rstudioconf18>

New textbook authoring paradigm



ModernDive



An Introduction to Statistical
and Data Sciences via



Chester Ismay
Albert Y. Kim

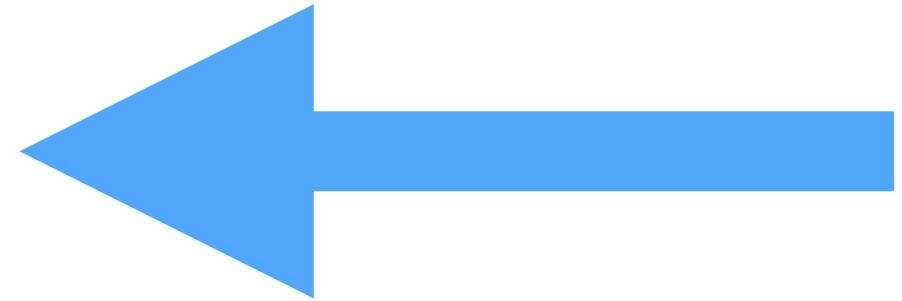
tidyverse

tidytext

moderndive
infer

“Versions, not editions”

Slides available at <http://bit.ly/rstudioconf18>



Slides available at <http://bit.ly/rstudioconf18>



Slides available at <http://bit.ly/rstudioconf18>



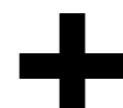
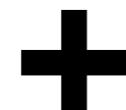
+



+



Slides available at <http://bit.ly/rstudioconf18>



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



Chester Ismay
DataCamp
Twitter: @old_man_chester
GitHub: ismayc

Slides available at <http://bit.ly/rstudioconf18>



Slides available at <http://bit.ly/rstudioconf18>



- Designed for the programming novice

Slides available at <http://bit.ly/rstudioconf18>



- Designed for the programming novice
- Fills the need for a mesh of data science and statistical concepts

Slides available at <http://bit.ly/rstudioconf18>



- Designed for the programming novice
- Fills the need for a mesh of data science and statistical concepts
- Integration of DataCamp courses for practice

Slides available at <http://bit.ly/rstudioconf18>



- Designed for the programming novice
- Fills the need for a mesh of data science and statistical concepts
- Integration of DataCamp courses for practice
- [Join our mailing list](#)

Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

“A Modern Dive into Data with R”

Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

“A Modern Dive into Data with R”

- Available at moderndive.com

Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

“A Modern Dive into Data with R”

- Available at moderndive.com
- Development version at moderndive.netlify.com

Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

“A Modern Dive into Data with R”

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

Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

“A Modern Dive into Data with R”

- Available at moderndive.com
- Development version at moderndive.netlify.com
- On GitHub at github.com/moderndive/
- Pull requests and feedback always welcomed!



An Introduction to Statistical and Data Sciences via R

“A Modern Dive into Data with R”

- Available at moderndive.com
- Development version at moderndive.netlify.com
- On GitHub at github.com/moderndive/
- Pull requests and feedback always welcomed!

v0.3.0 now released!

Slides available at <http://bit.ly/rstudioconf18>



An Introduction to Statistical and Data Sciences via R

“A Modern Dive into Data with R”

- Available at moderndive.com
- Development version at moderndive.netlify.com
- On GitHub at github.com/moderndive/
- Pull requests and feedback always welcomed!



v0.3.0 now released!



Slides available at <http://bit.ly/rstudioconf18>