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:

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

Question

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

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

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

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

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

Question

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

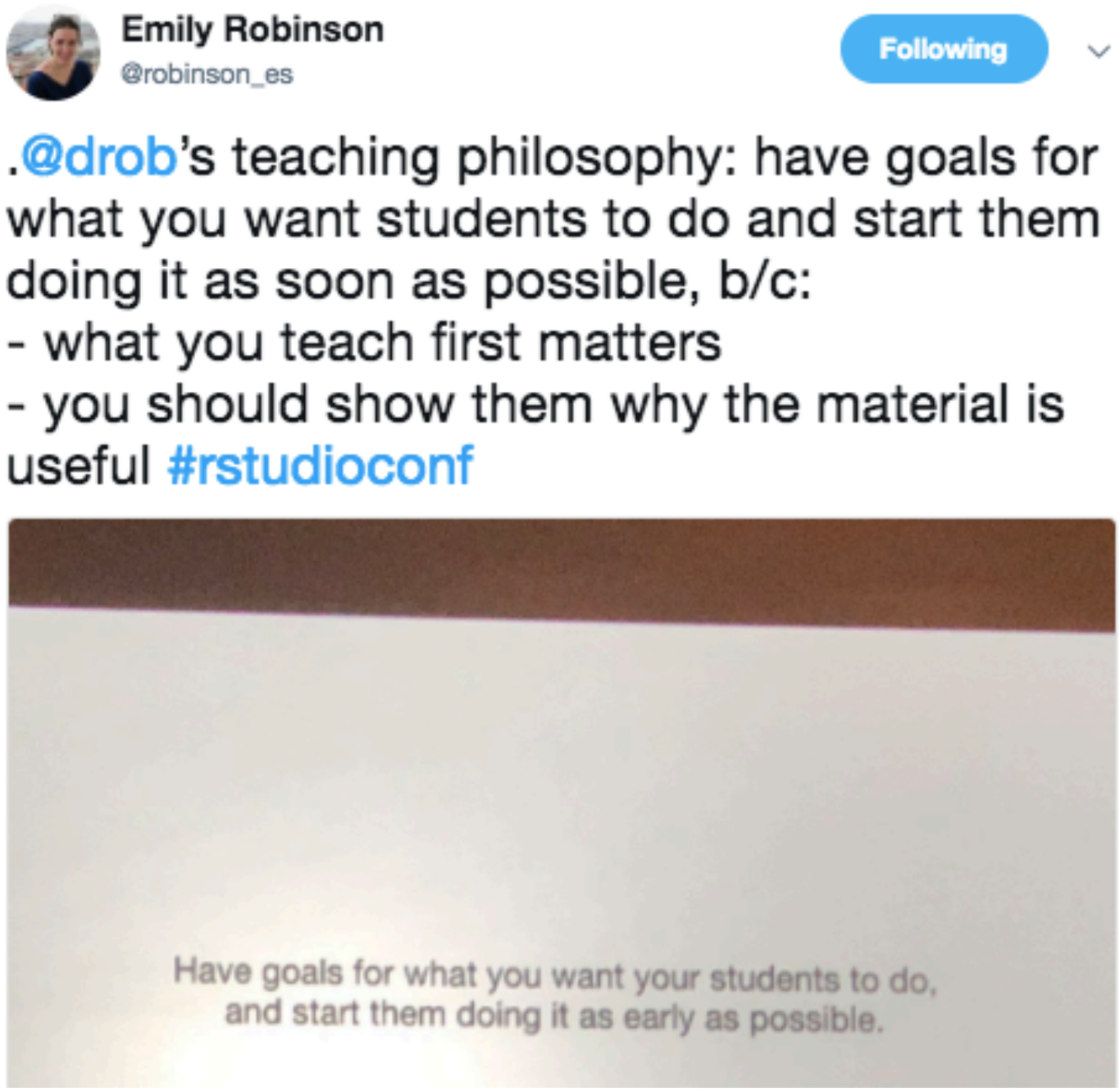
1. **Data science**: Data visualization, data wrangling, exploratory data analysis, data tidying

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

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

General Framework



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

"A Modern Dive into Data with R"

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

ModernDive

An Introduction to Statistical and Data Sciences via R

"A Modern Dive into Data with R"

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

ModernDive

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

ModernDive

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

ModernDive

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

ModernDive

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

ModernDive

An Introduction to Statistical and Data Sciences via R

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

ModernDive

An Introduction to Statistical and Data Sciences via R

- Why is this needed?

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

ModernDive

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

ModernDive

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

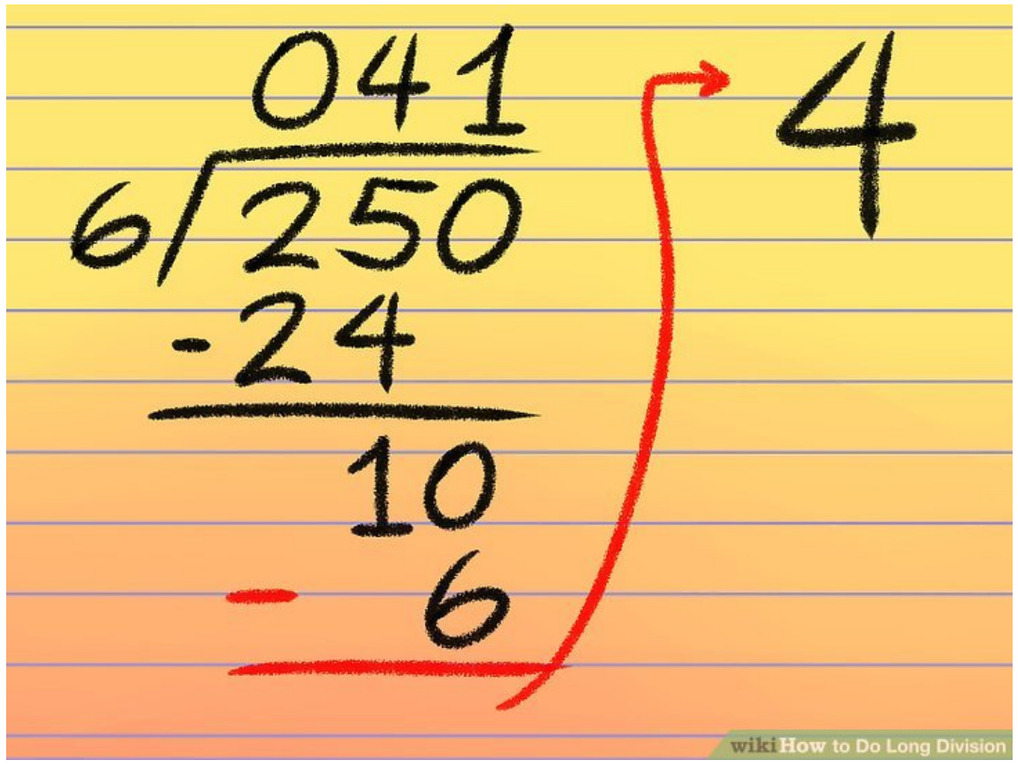

Something
Old
New
Borrowed
Blue

Analogy: Learning Long Division

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

Analogy: Learning Long Division

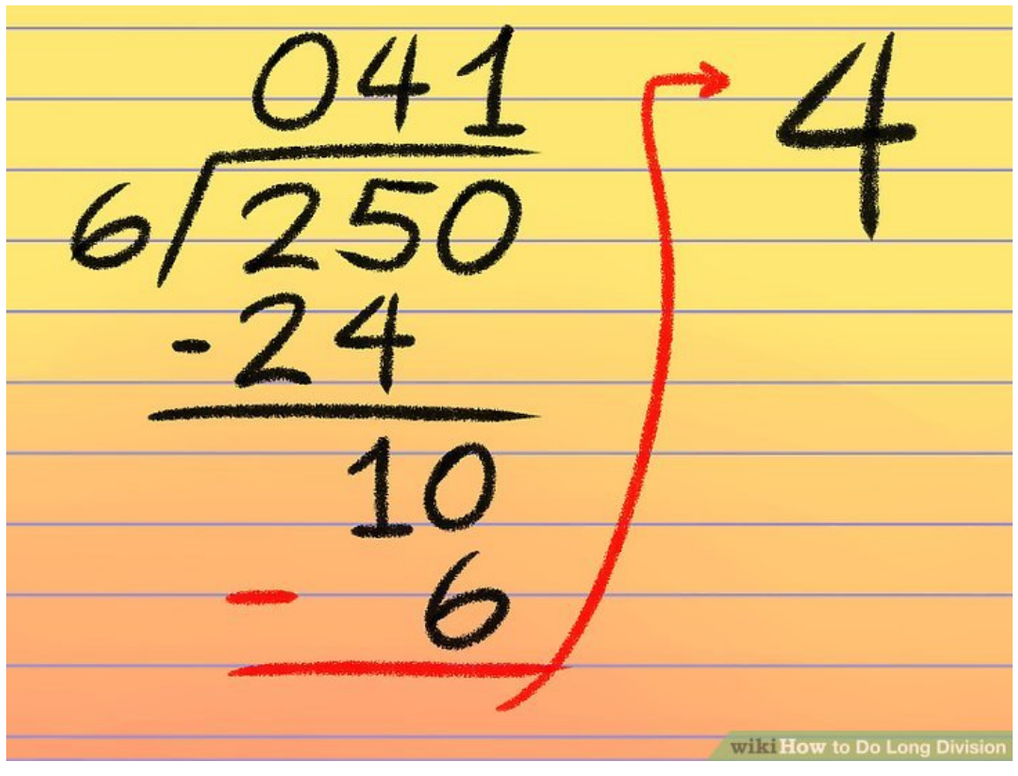
Do this a few times:



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

Analogy: Learning Long Division

Do this a few times:



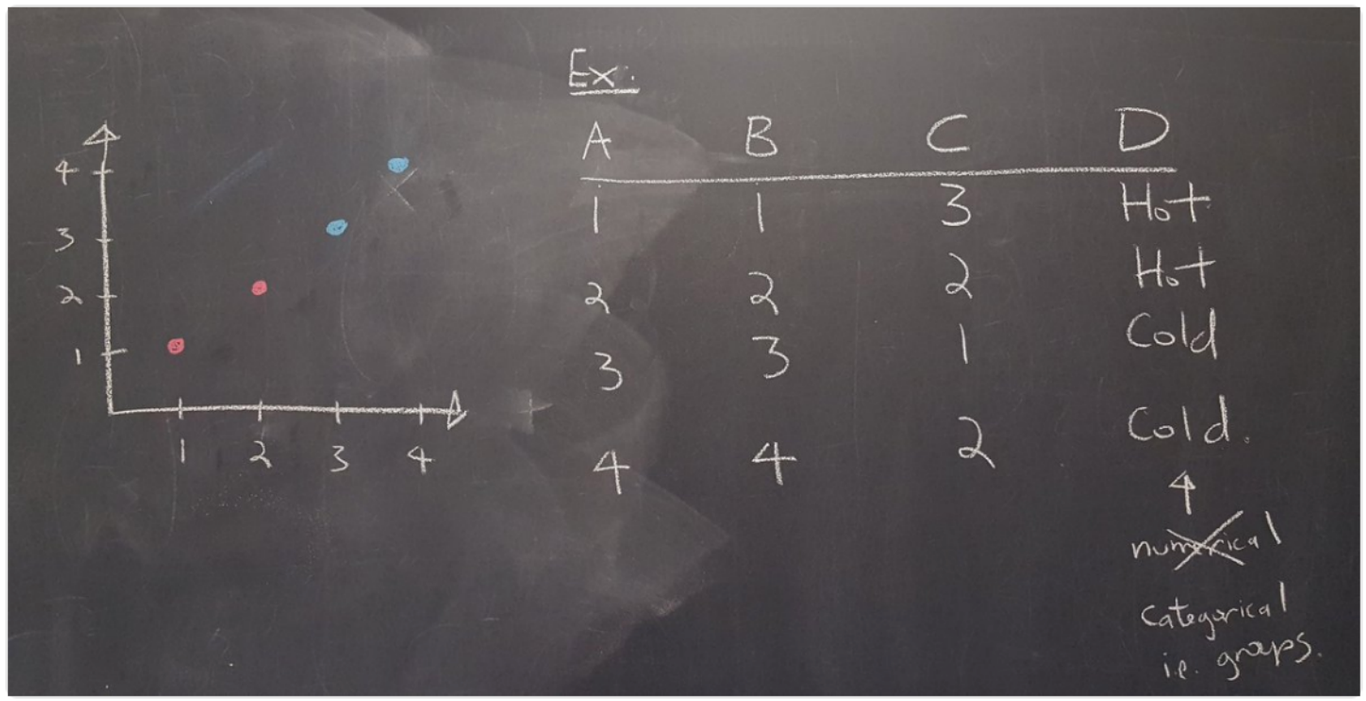
Then rely on this:

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

`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

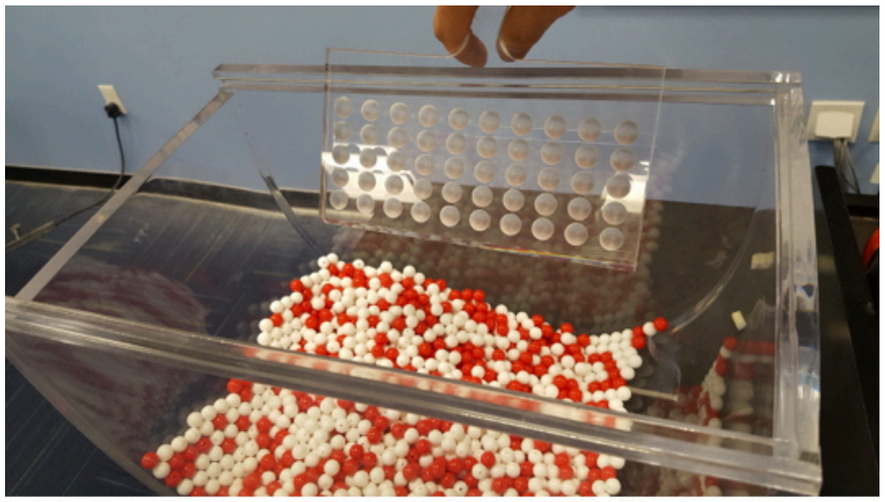
3 5 29

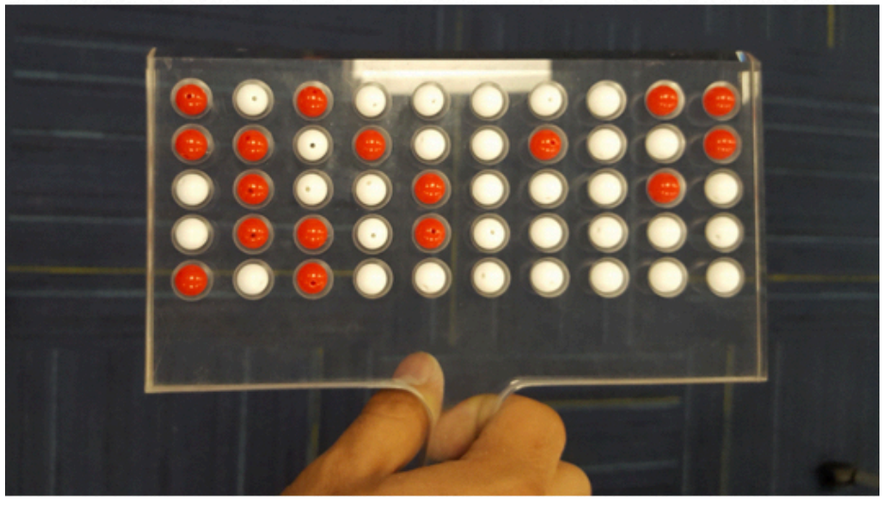
Slides available at http://bit.ly/rstudioconf18

Tactile simulation of sampling to teach sampling distributions

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

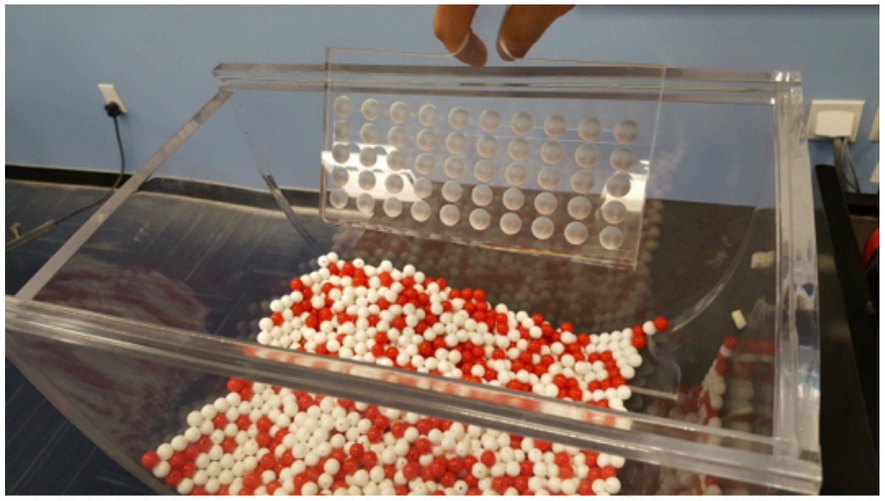
Tactile simulation of sampling to teach sampling distributions

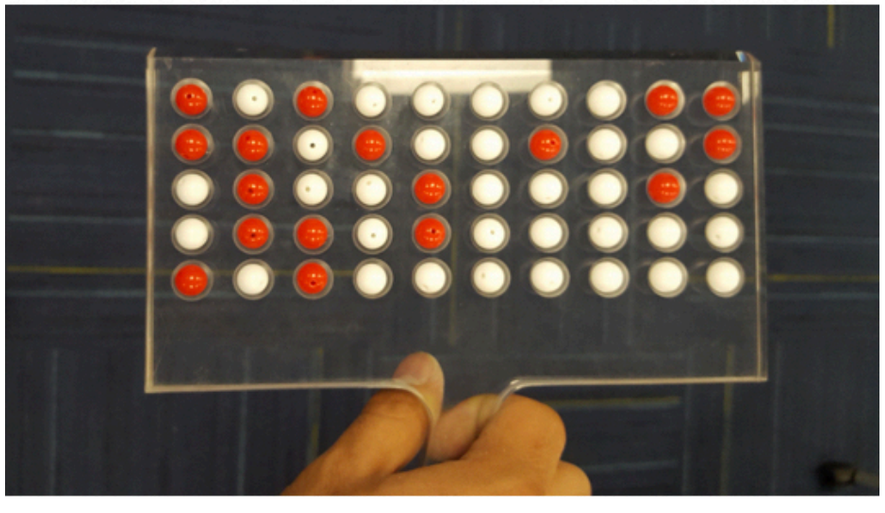




Slides available at http://bit.ly/rstudioconf18

Tactile simulation of sampling to teach sampling distributions

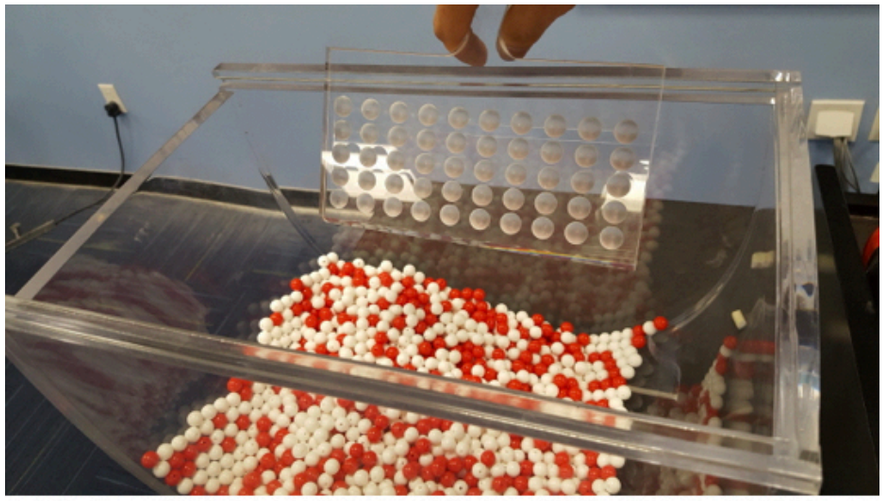


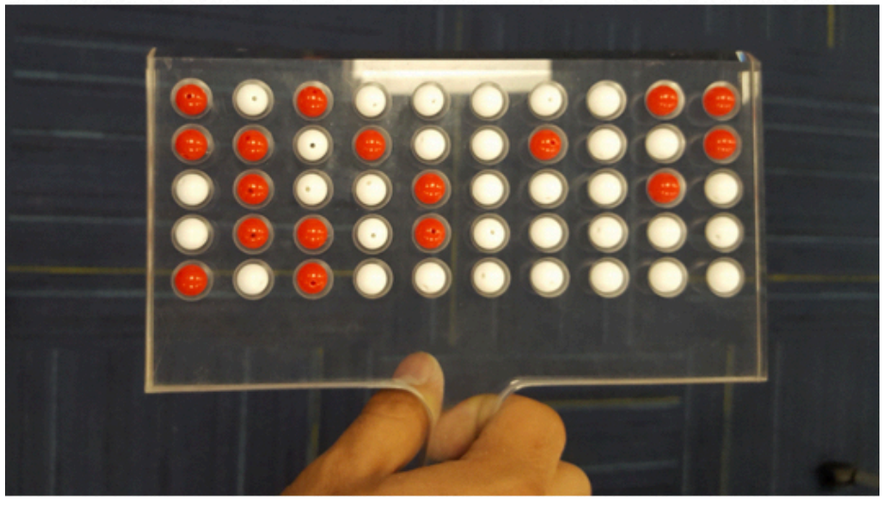


	group	red	n	prop_red
1	Kathleen and Max	18	50	0.36
2	Sean, Jack, and CJ	18	50	0.36
3	X and Judy	22	50	0.44
4	James and Jacob	21	50	0.42
5	Hannah and Siya	16	50	0.32
6	Niko, Sophie, and Caitlin	14	50	0.28
7	Niko, Sophie, and Caitlin	19	50	0.38
8	Aleja and Ray	20	50	0.40
9	Yaw and Drew	16	50	0.32
10	Yaw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

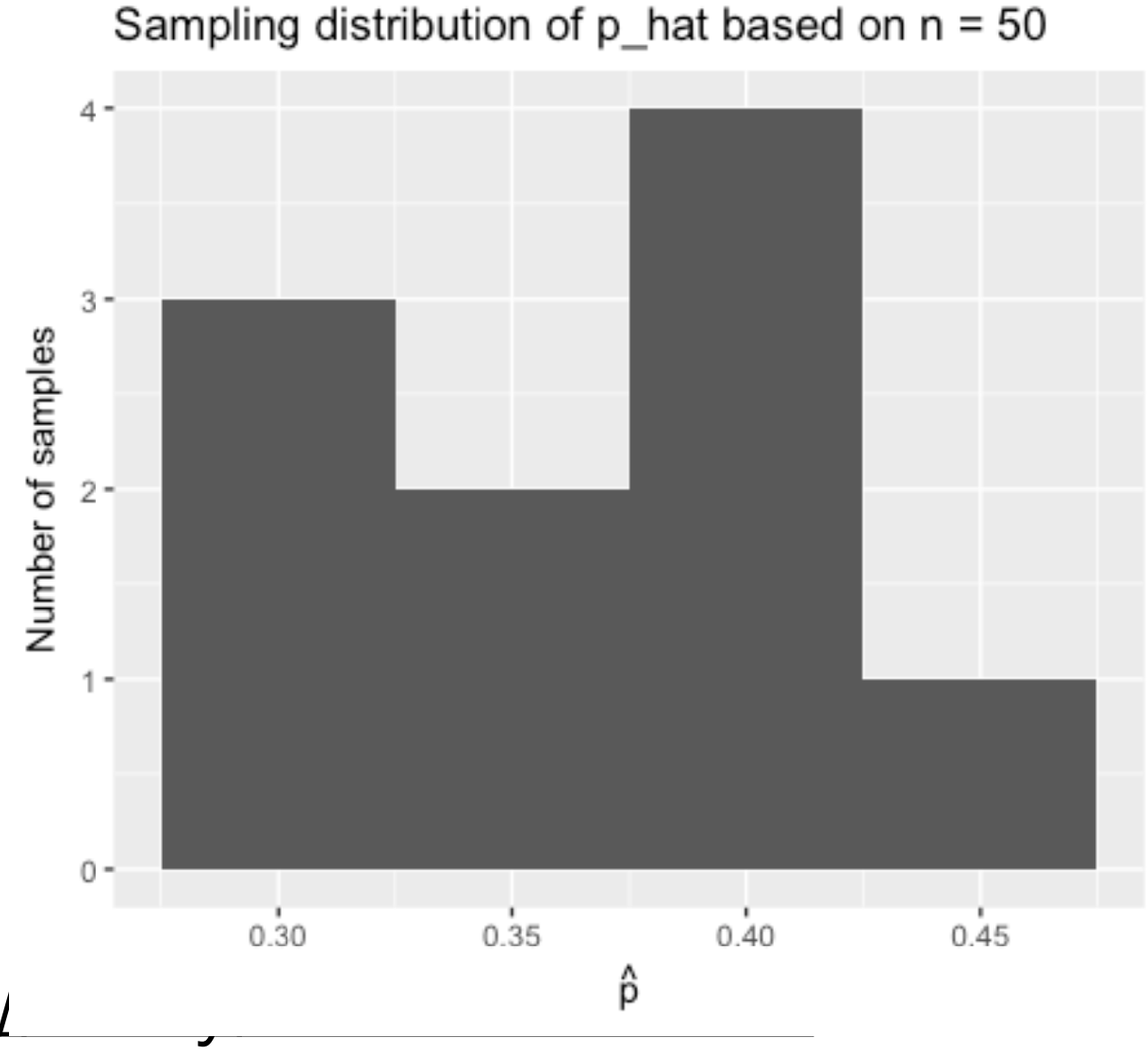
Slides available at http://bit.ly/rstudioconf18

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



Slides available at http:/

Computer simulation of sampling to teach sampling distributions

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

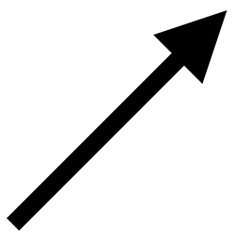
Computer simulation of sampling to teach sampling distributions

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

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

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

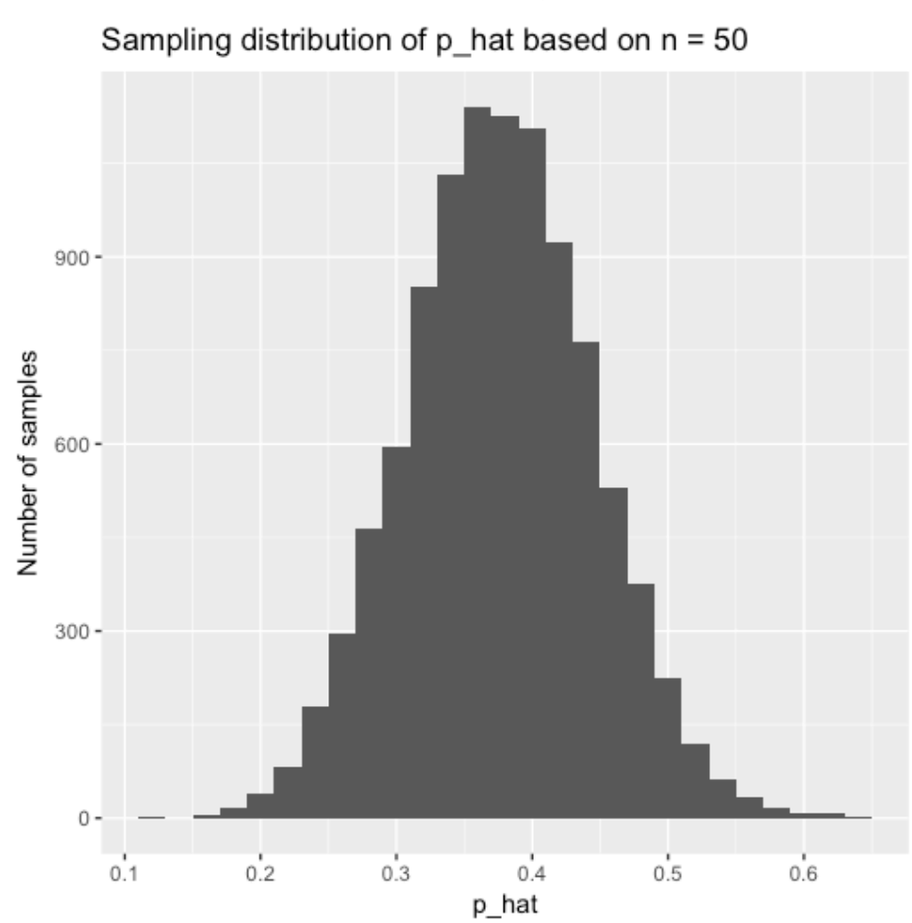
Slides available at http://bit.ly/rstudioconf18

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



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


Something
Old
New
Borrowed
Blue

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....*

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

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

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

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

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
  2. Variable name misspellings

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
  2. Variable name misspellings
  3. Linearity of programs

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
    2. Variable name misspellings
    3. Linearity of programs
- Examples of the "Curse of knowledge"

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


Something
Old
New
Borrowed
Blue

Leverage open source

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

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

Leverage open source

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

Bechdel test?

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

Leverage open source

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

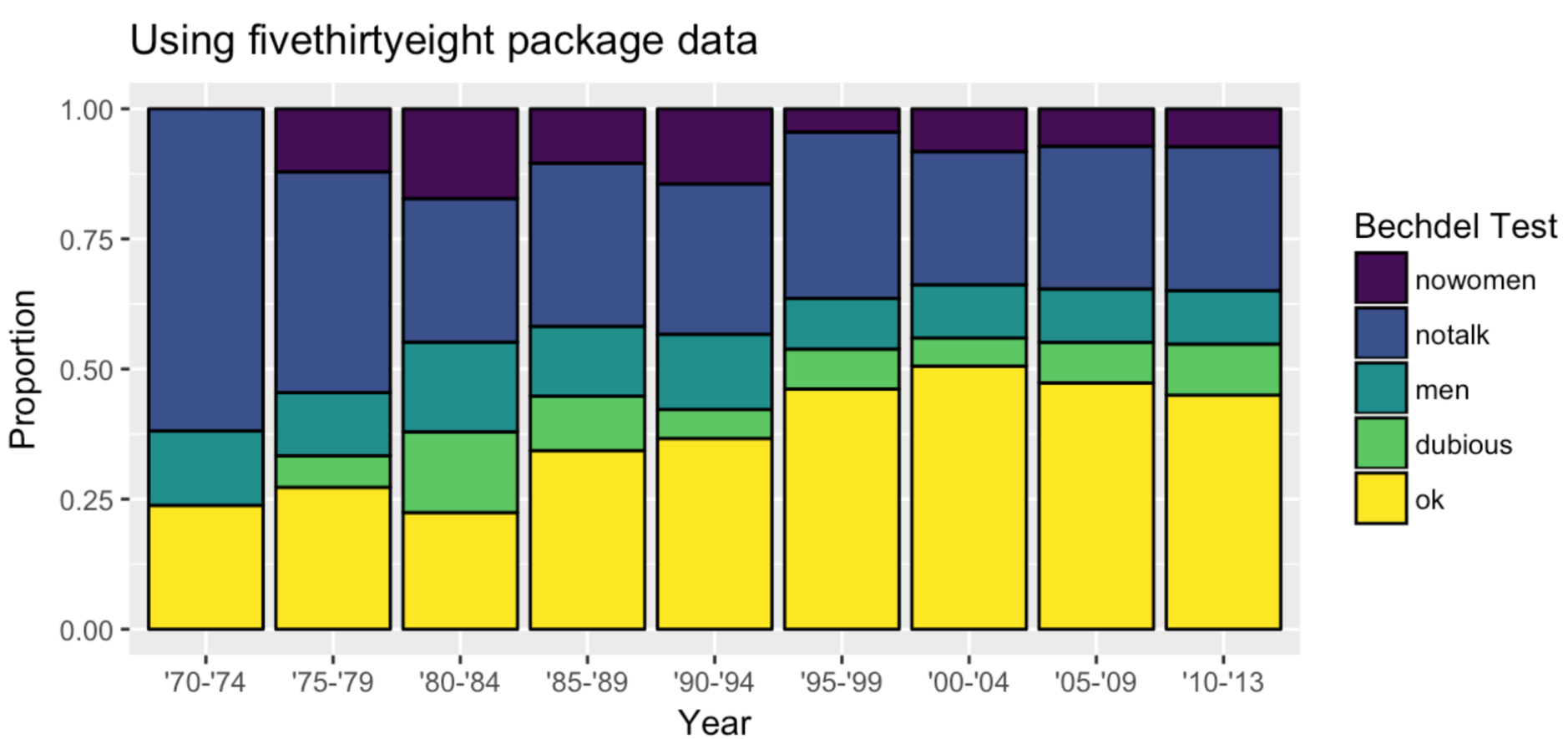
Bechdel test? Original 538 article

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

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

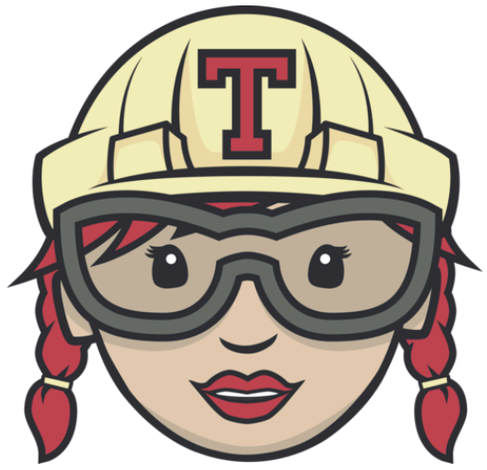
New textbook authoring paradigm

+

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

New textbook authoring paradigm

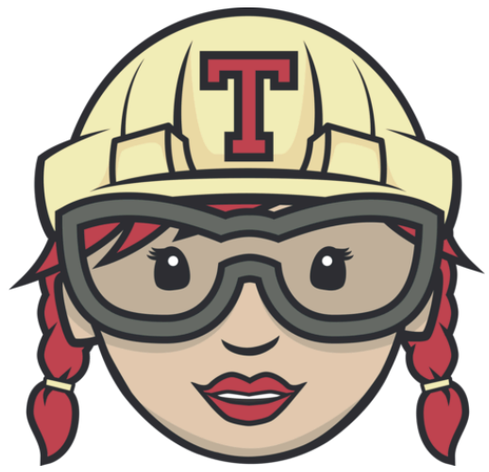
+



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

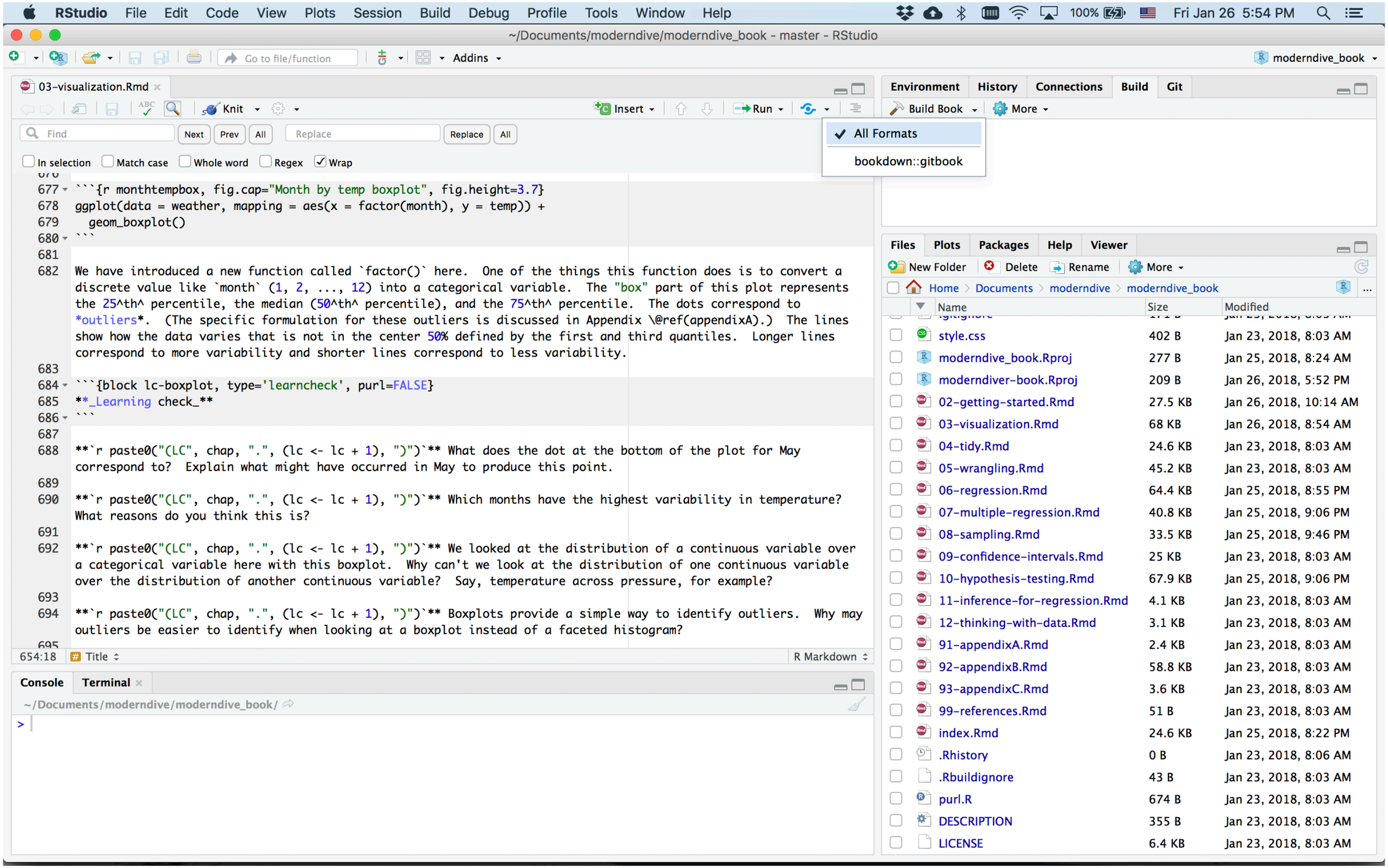
New textbook authoring paradigm

+



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

New textbook authoring paradigm



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

New textbook authoring paradigm


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

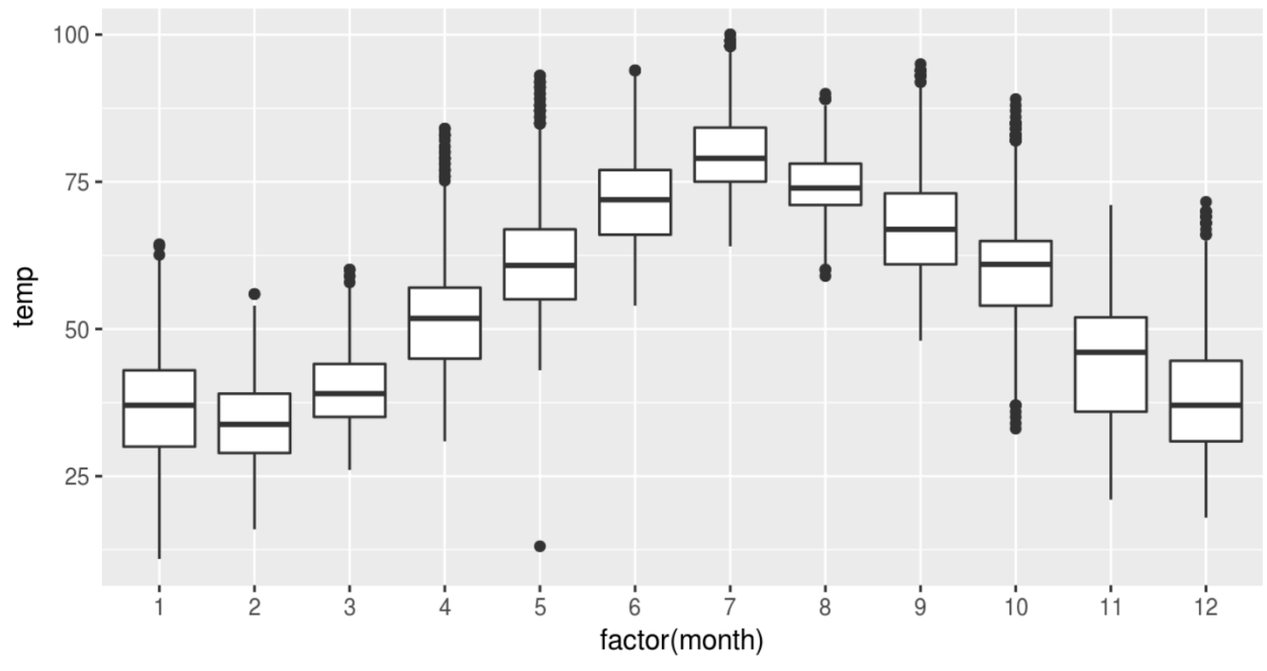


Figure 3.13: Month by temp boxplot

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

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

12. Thinking with Data

Data Modeling with `moderndive`

8. Sampling

9. Confidence Intervals

10. Hypothesis Testing

Statistical Inference with `infer`

ModernDive

Available at `moderndive.com`

Diagram inspired by hadley/r4ds

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

Stable

1. Introduction

2. Getting Started with Data in R

3. Data Visualization

4. Tidy Data

5. Data Wrangling

Data Science with `tidyverse`

12. Thinking with Data

6. Basic Regression

7. Multiple Regression

11. Inference for Regression

Data Modeling with `moderndive`

8. Sampling

9. Confidence Intervals

10. Hypothesis Testing

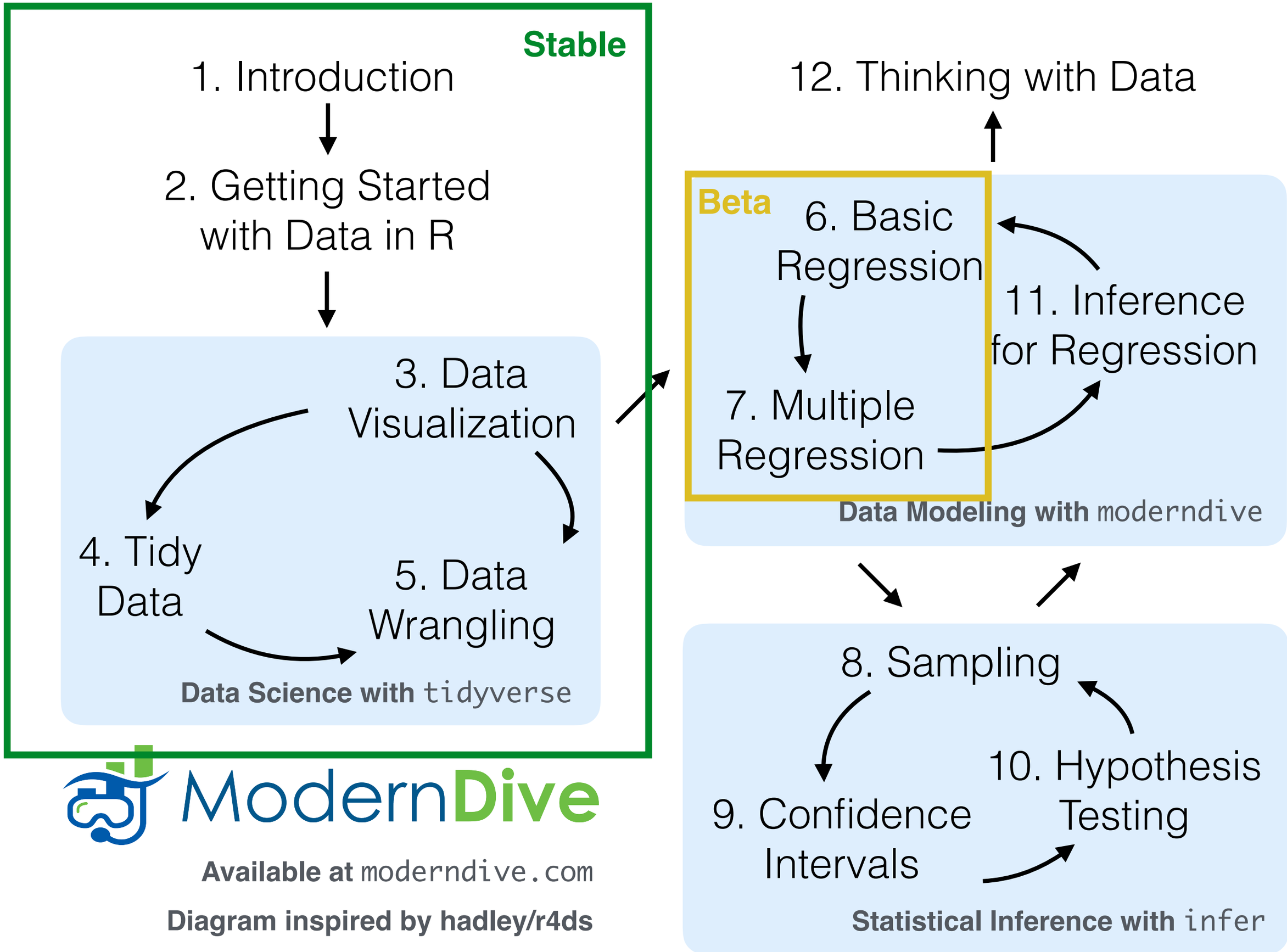
Statistical Inference with `infer`

ModernDive

Available at `moderndive.com`

Diagram inspired by hadley/r4ds

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



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

Stable

1. Introduction

2. Getting Started with Data in R

3. Data Visualization

4. Tidy Data

5. Data Wrangling

Data Science with `tidyverse`

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`

ModernDive

Available at `moderndive.com`

Diagram inspired by hadley/r4ds

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

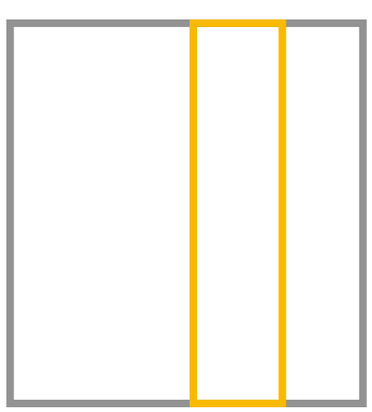
`infer` package for tidy statistical inference

http://infer.netlify.com/

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

`infer` package for tidy statistical inference

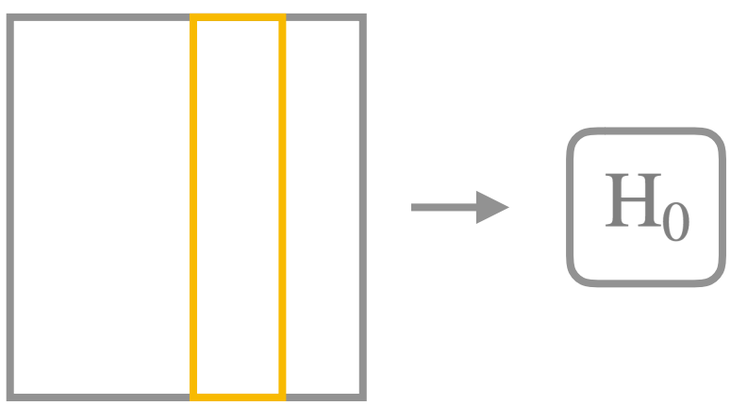
http://infer.netlify.com/



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

`infer` package for tidy statistical inference

http://infer.netlify.com/



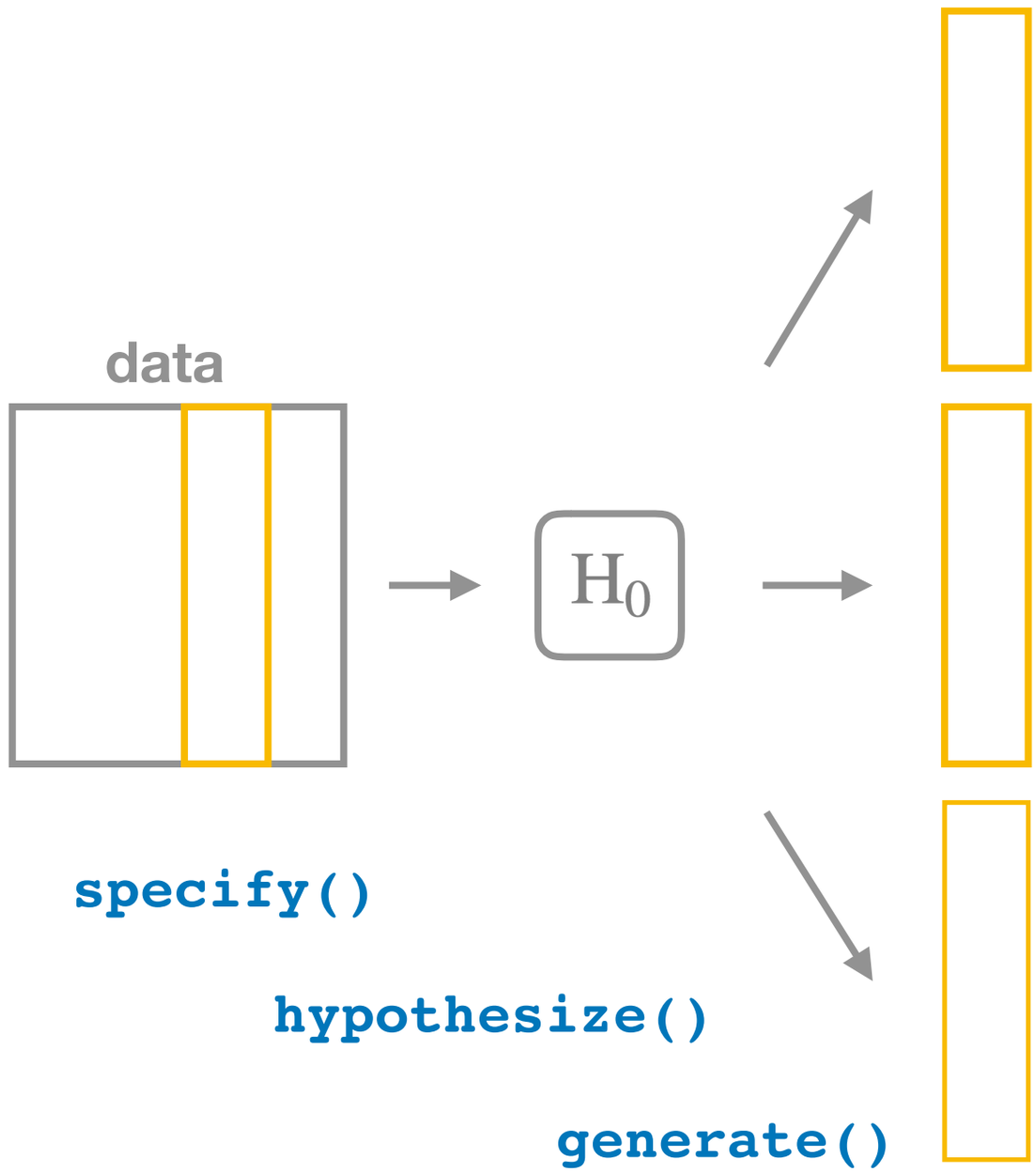

```
specify()

    hypothesize()
```

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

`infer` package for tidy statistical inference

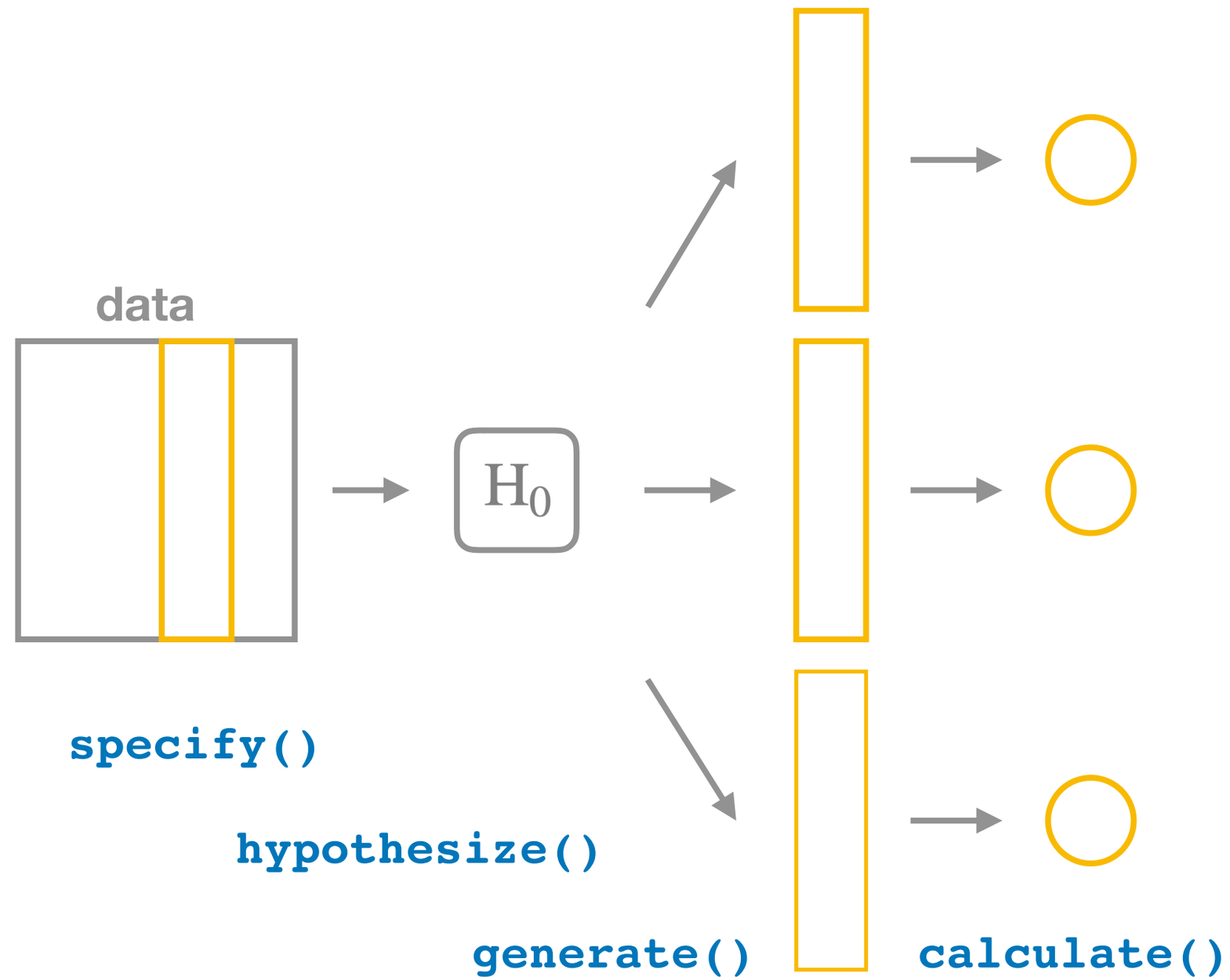
http://infer.netlify.com/



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

`infer` package for tidy statistical inference

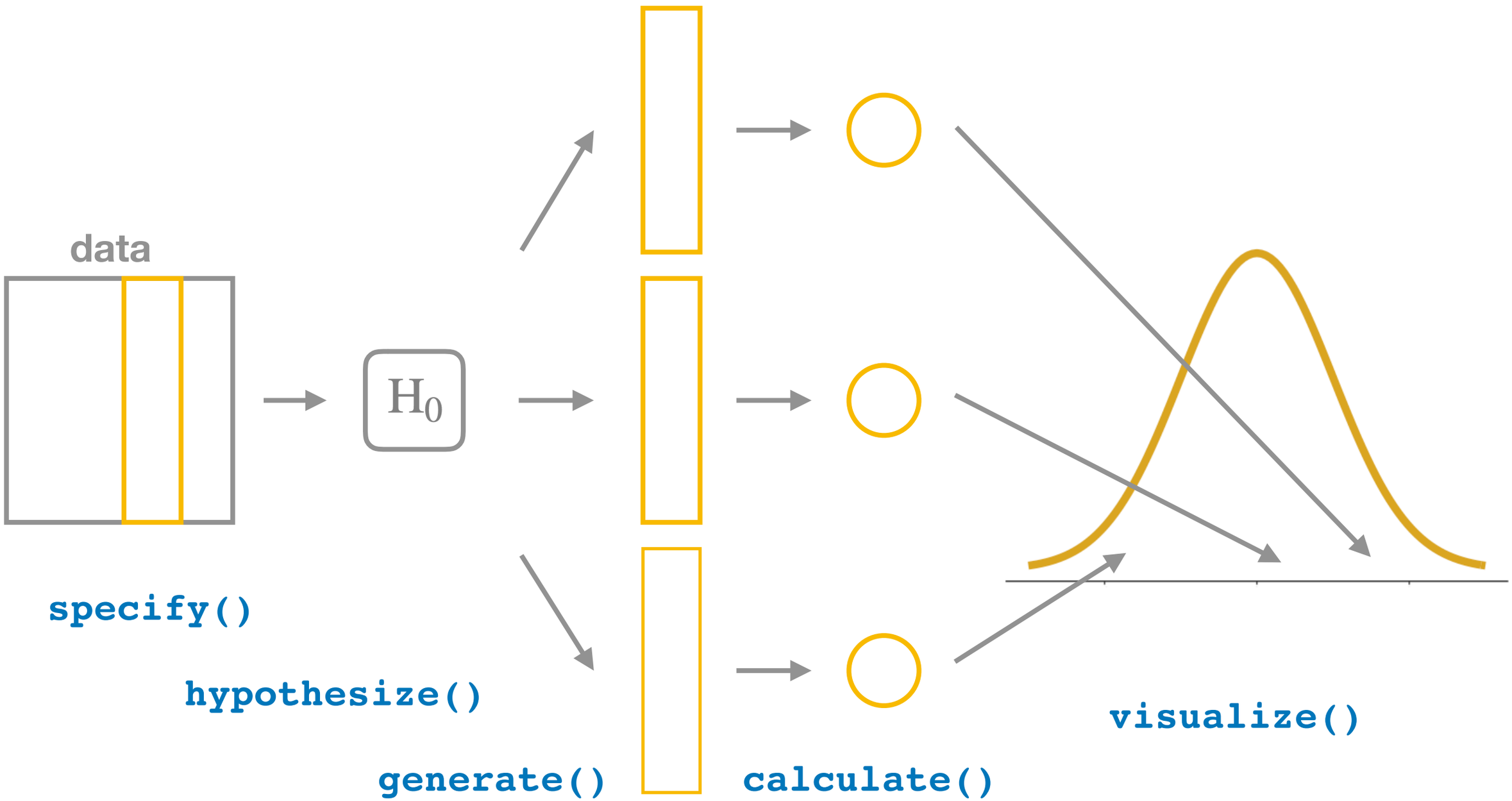
http://infer.netlify.com/



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

`infer` package for tidy statistical inference

http://infer.netlify.com/

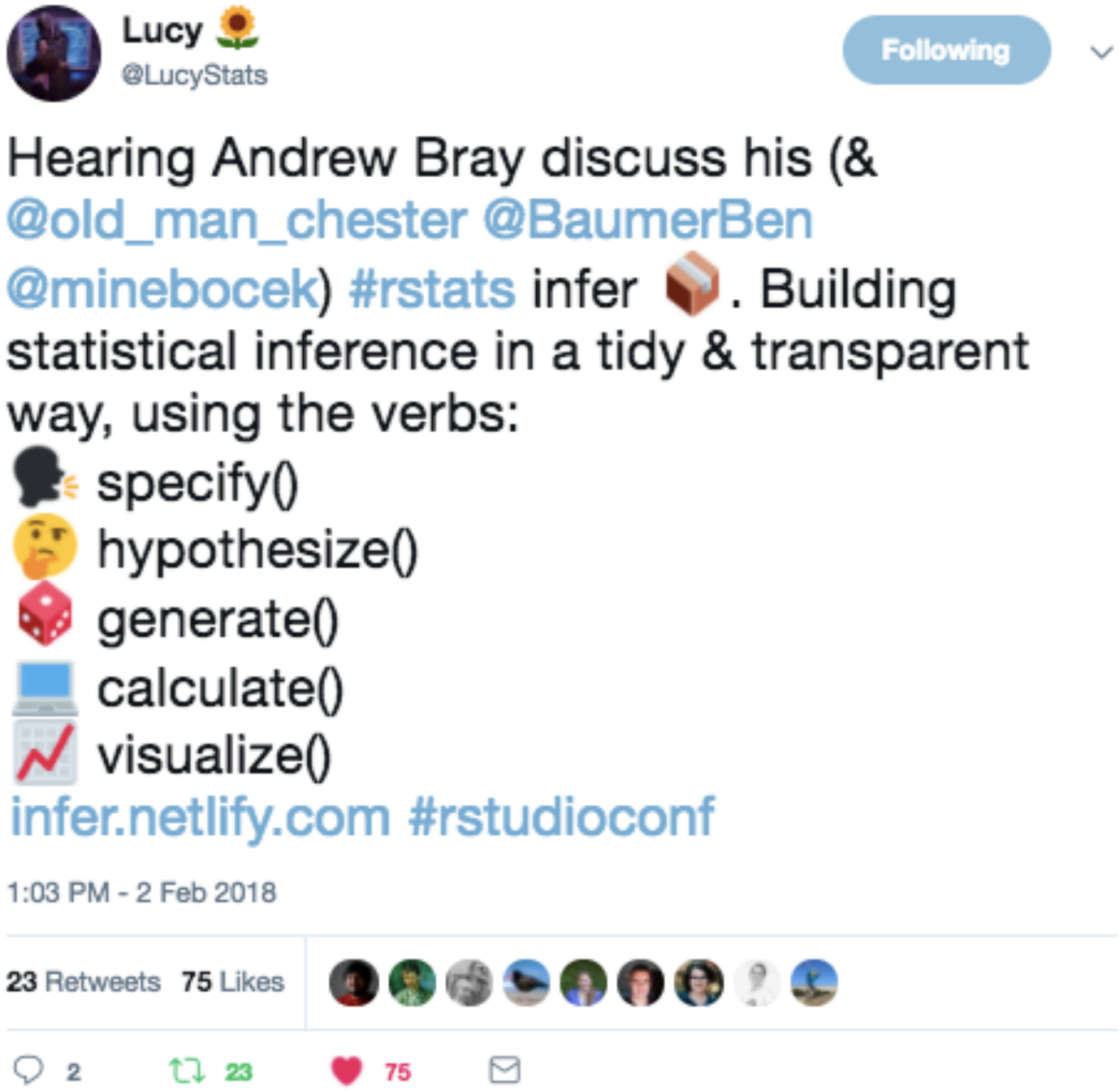


Slides available at http://bit.ly/rstudioconf18

`infer` package for tidy statistical inference

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

`infer` package for tidy statistical inference



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

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

New textbook authoring paradigm

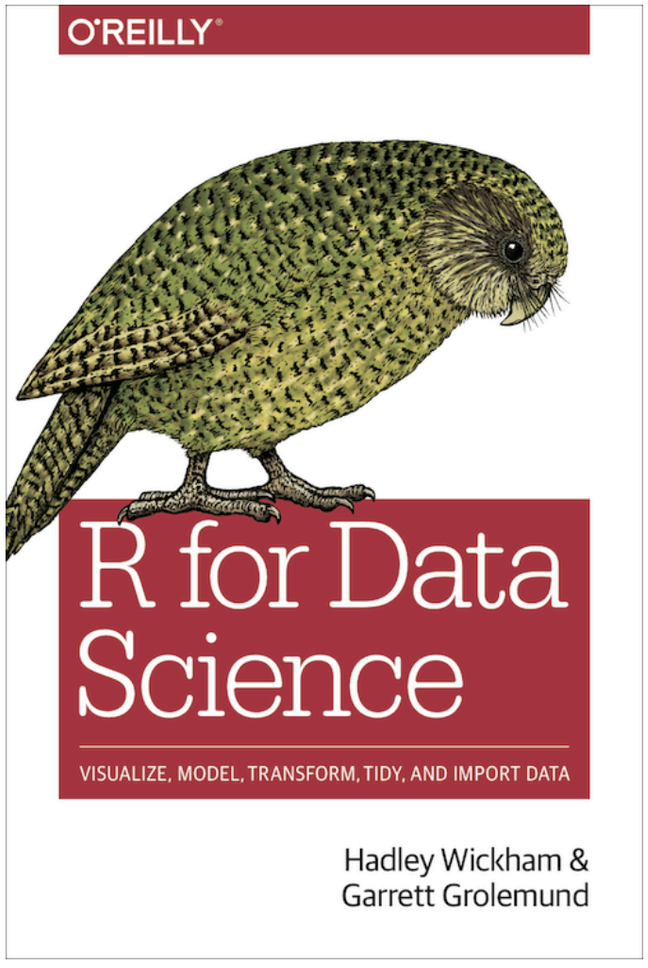
Slides available at http://bit.ly/rstudioconf18

New textbook authoring paradigm

"Versions, not editions"

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

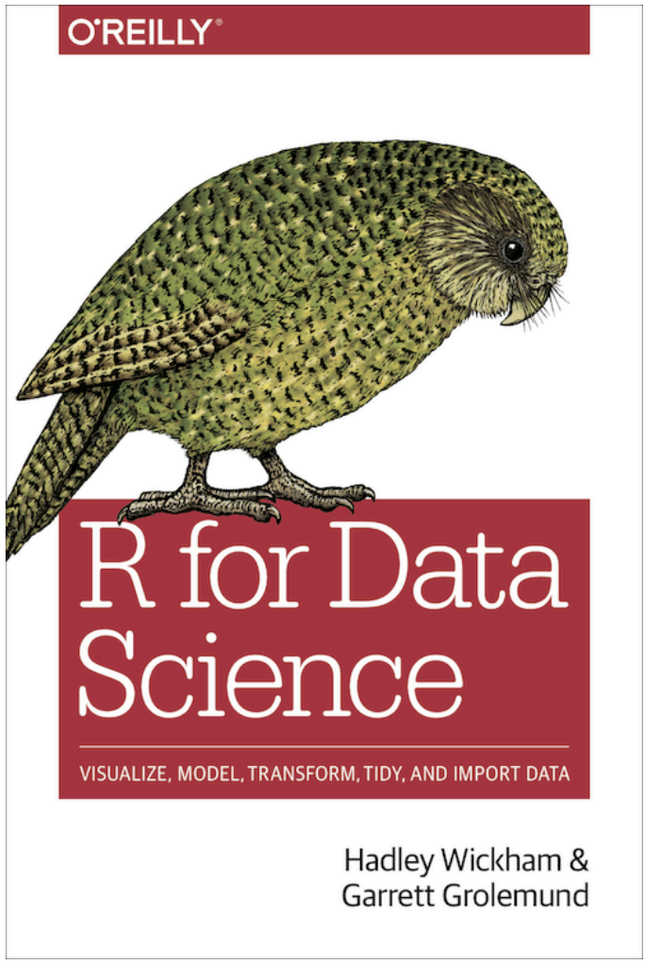
New textbook authoring paradigm



"Versions, not editions"

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

New textbook authoring paradigm

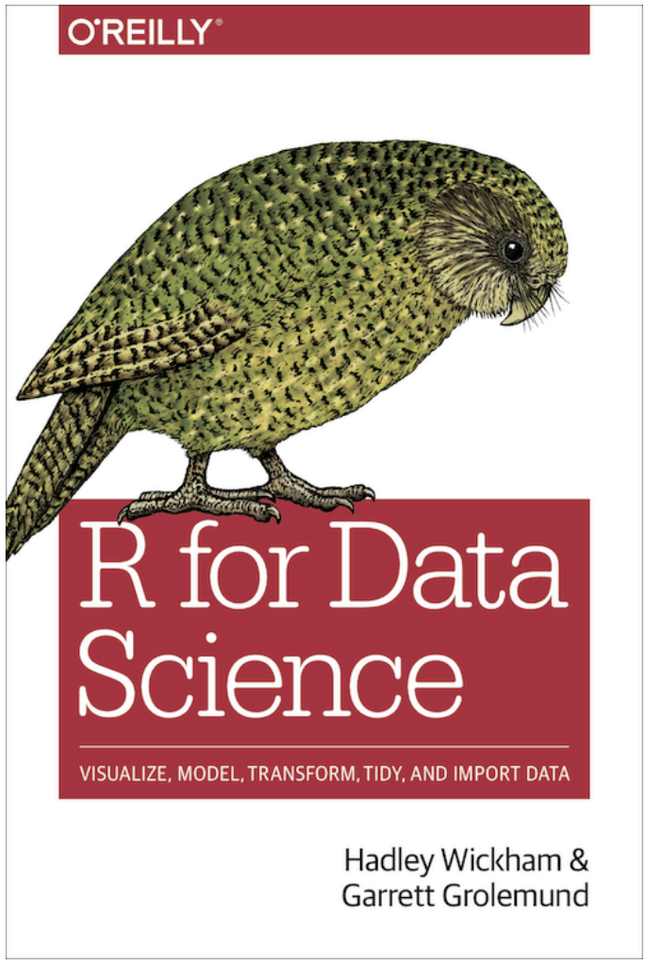


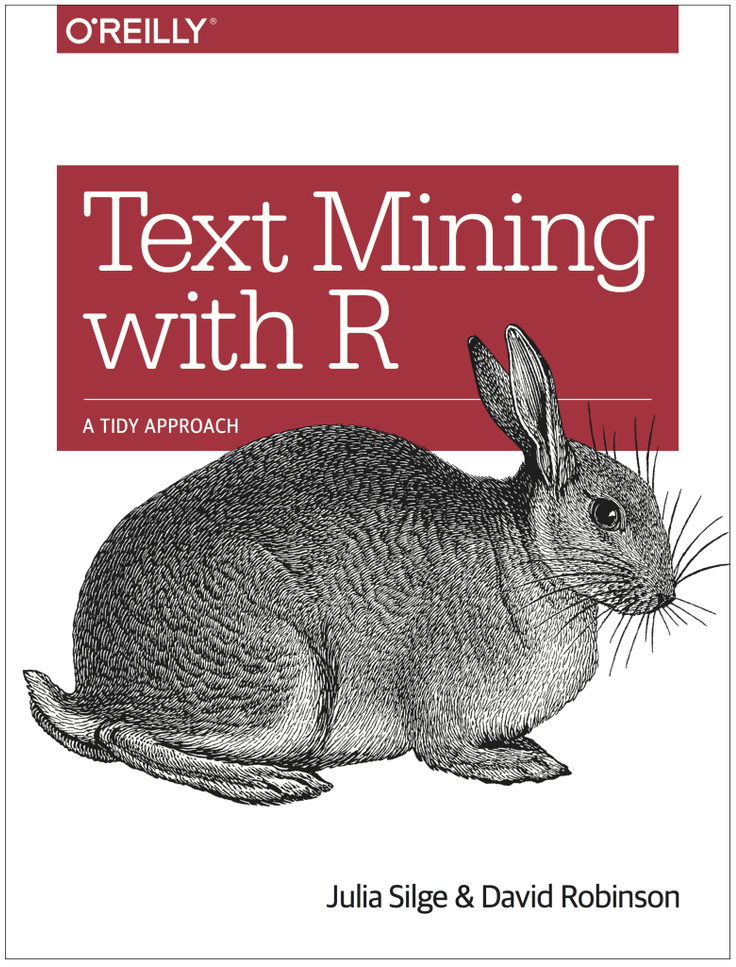
`tidyverse`

"Versions, not editions"

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

New textbook authoring paradigm



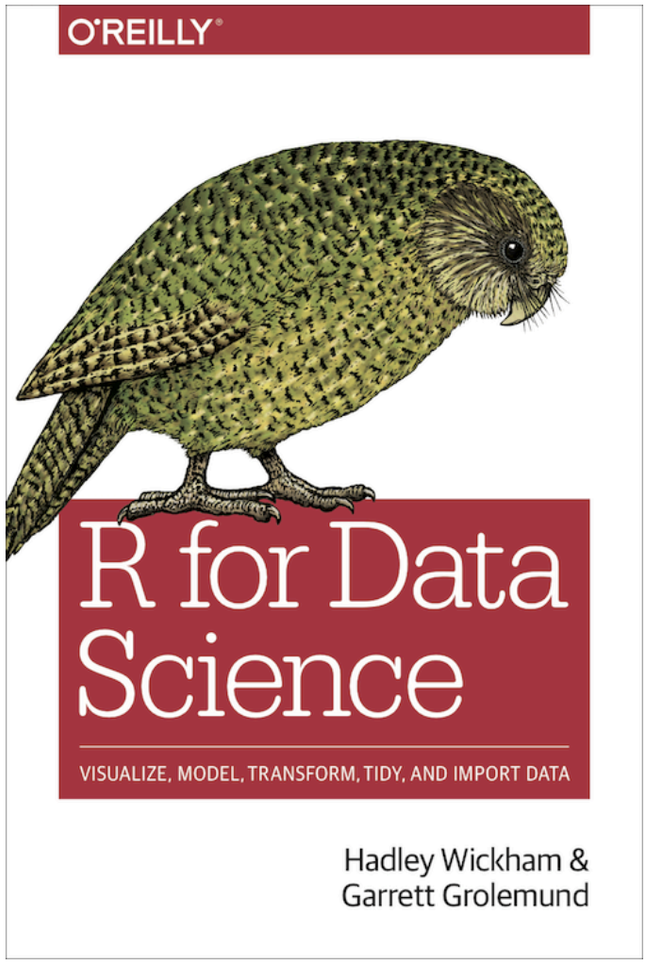


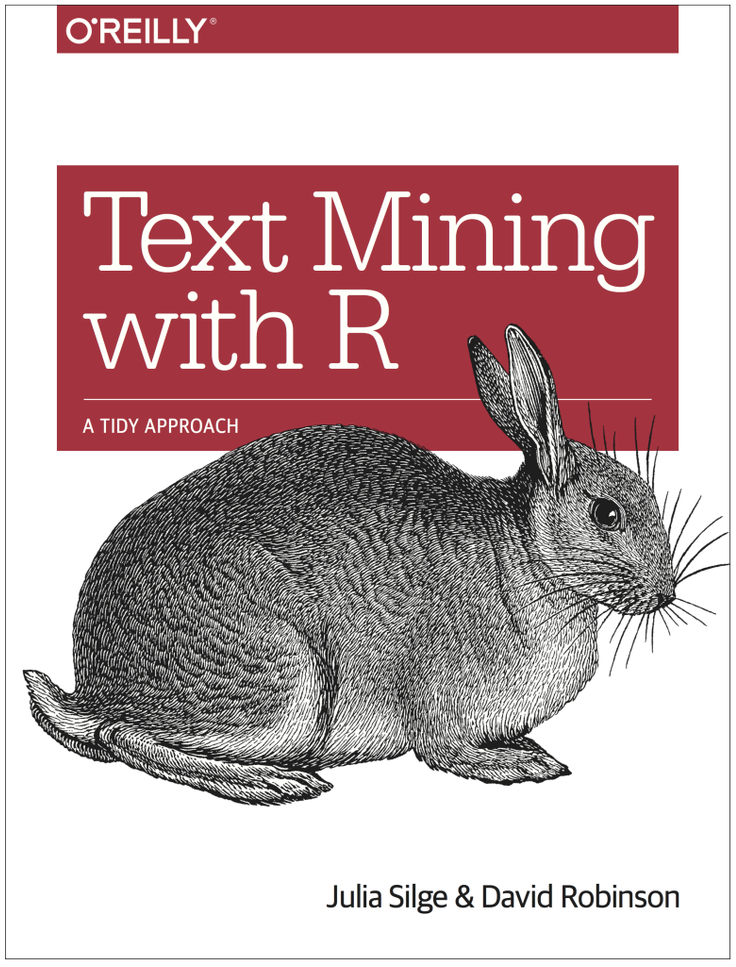
`tidyverse`

"Versions, not editions"

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

New textbook authoring paradigm



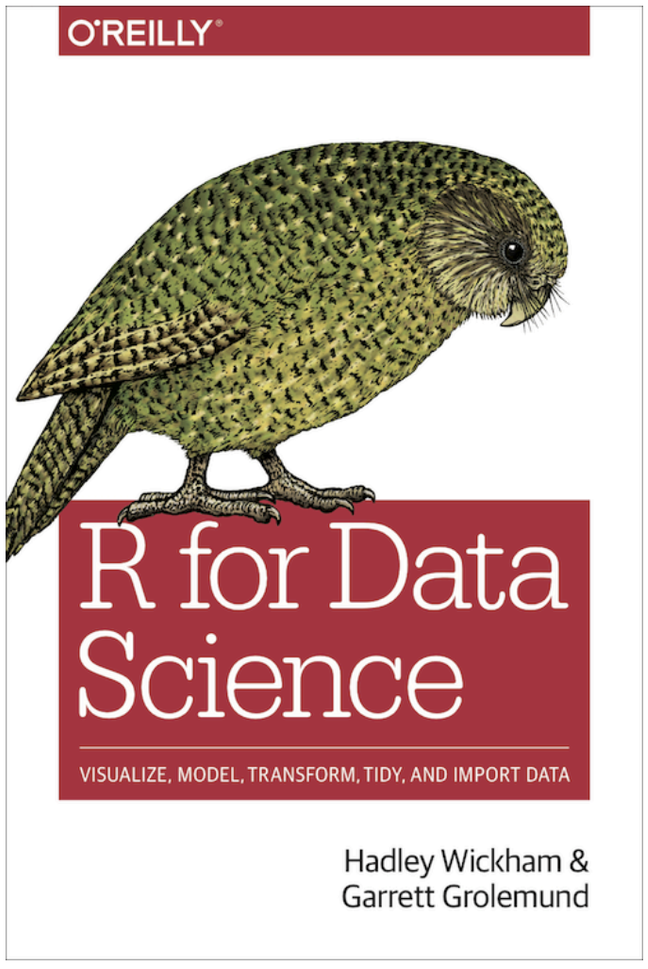


`tidyverse` `tidytext`

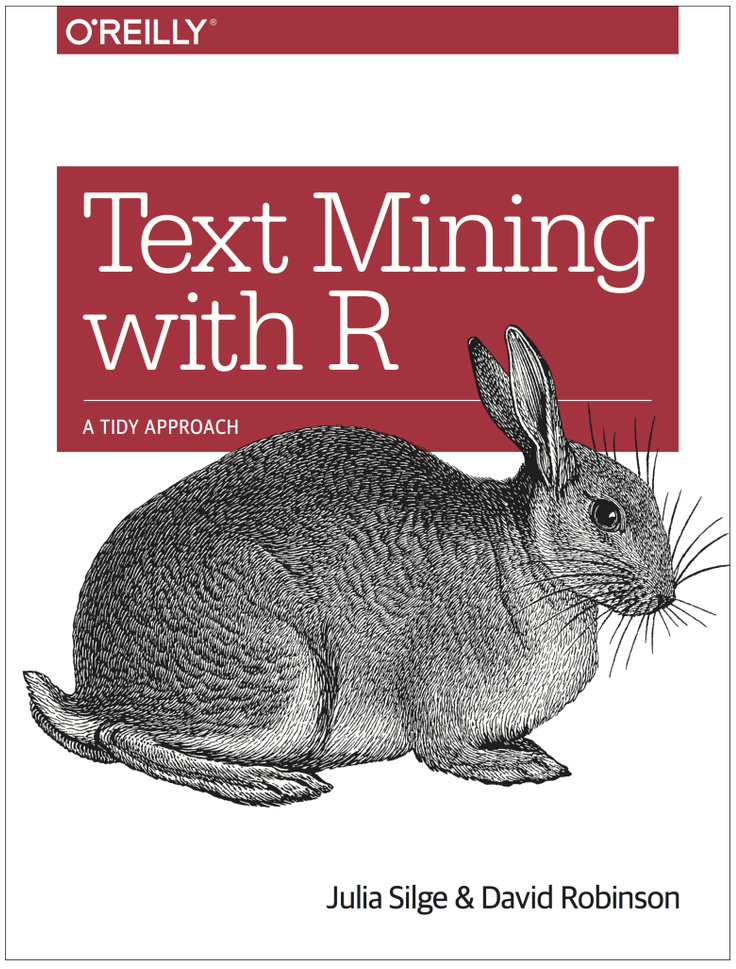
"Versions, not editions"

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

New textbook authoring paradigm



`tidyverse`



`tidytext`

ModernDive

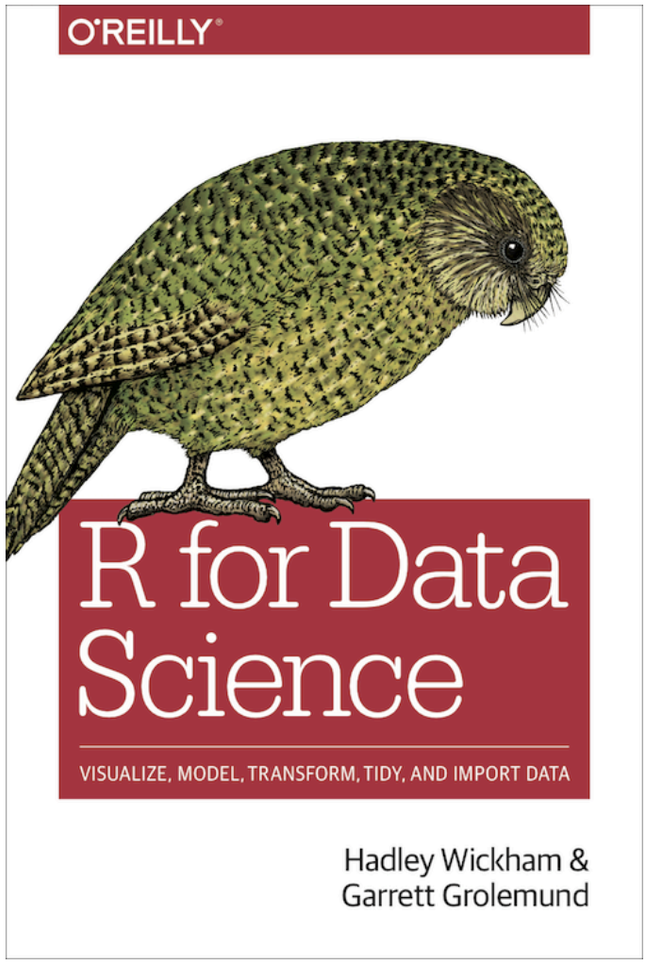
An Introduction to Statistical and Data Sciences via R

Chester Ismay
Albert Y. Kim

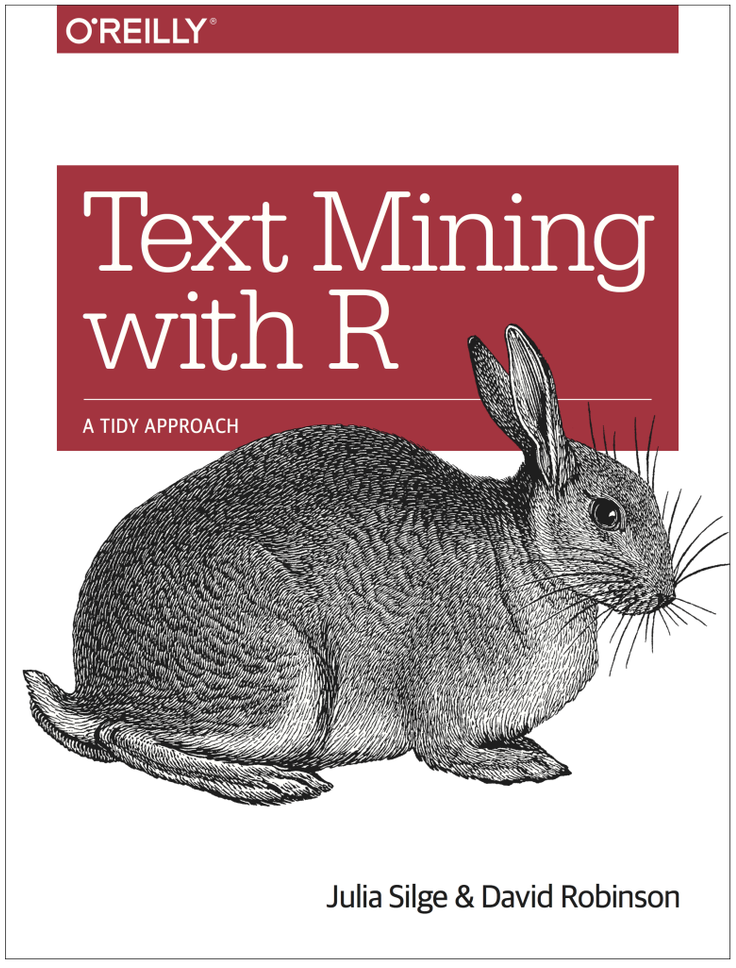
"Versions, not editions"

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

New textbook authoring paradigm



`tidyverse`



`tidytext`

ModernDive

An Introduction to Statistical and Data Sciences via R

Chester Ismay
Albert Y. Kim

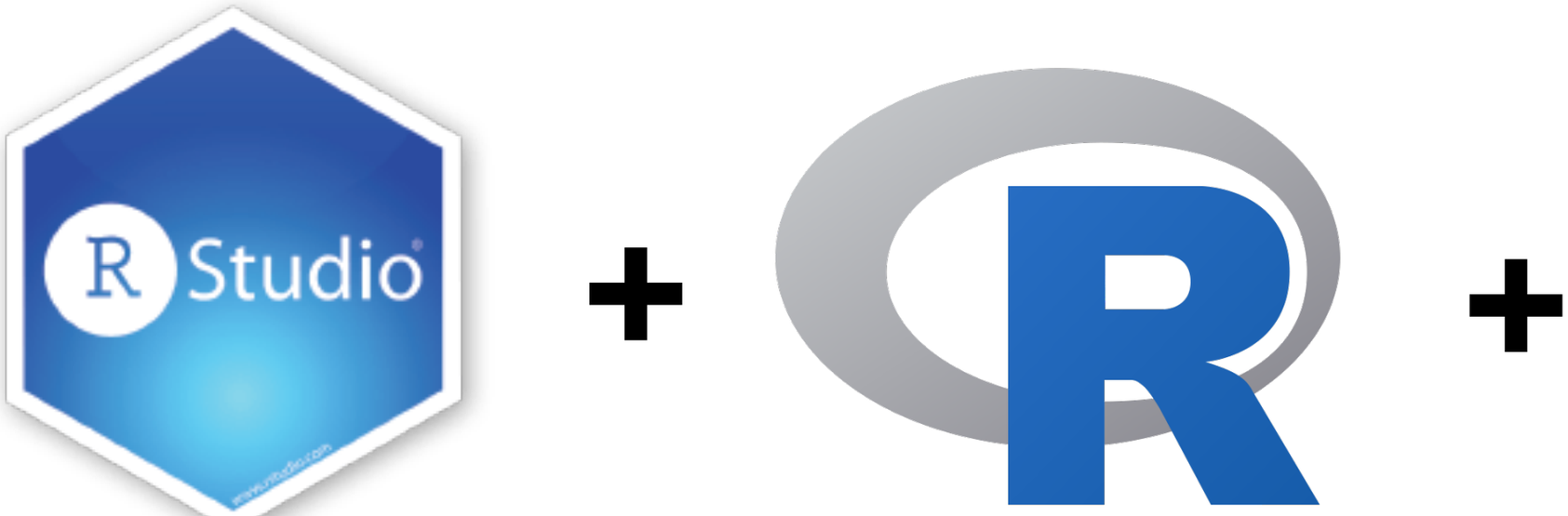
`moderndive`
`infer`

"Versions, not editions"

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


Something
Old
New
Borrowed
Blue

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

ModernDive

- Designed for the programming novice

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

ModernDive

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

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

ModernDive

- 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

ModernDive

- 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

ModernDive

An Introduction to Statistical and Data Sciences via R

"A Modern Dive into Data with R"

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

ModernDive

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

ModernDive

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

ModernDive

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

ModernDive

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!

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

ModernDive

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

ModernDive

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