Statistical inference via data science: A "tidy" approach

Albert Y. Kim

Joint Math Meetings
Denver CO, USA
January 18, 2020

Slides available at twitter.com/rudeboybert

Statistical inference **via** data science…

What is the tidyverse?

R packages for data science

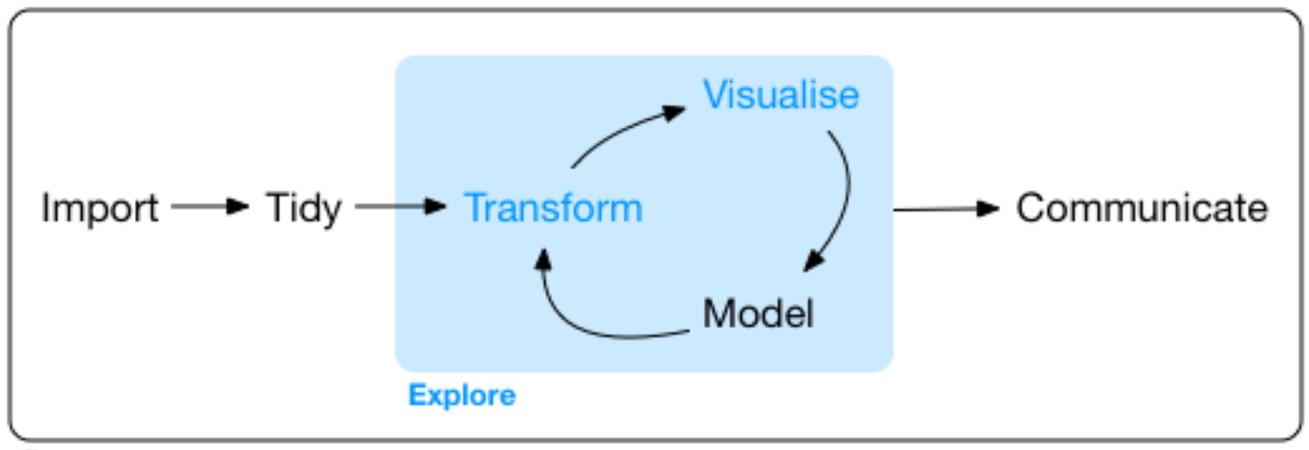
The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

From: tidyverse.org

Why use the tidyverse?

1. It encourages students to “play the whole game”
2. It’s transferable
3. It bridges the gap between tools for *learning* statistics & tools for *doing* statistics

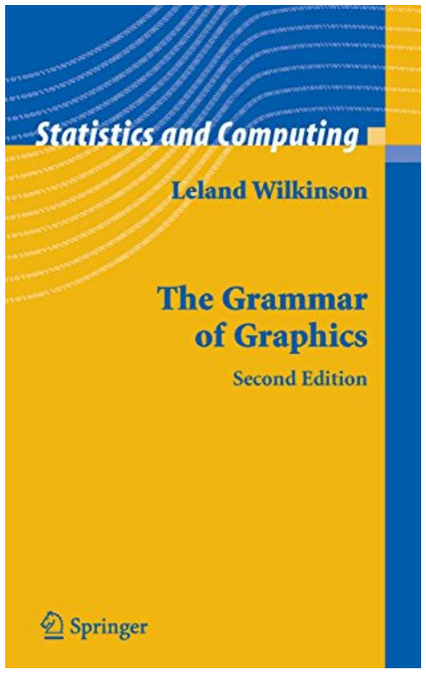
1. It encourages students to “play the whole game”



- Exploratory data analysis (EDA)
- “To (data) wrangle or not to wrangle?”
- IMO to do no data wrangling betrays true nature of the work

From: YouTube, r4ds (2017), Perkins (2009)

2.a) It transfers: Data visualization



Salesforce closes $15.7B Tableau deal

Ron Miller @ron_miller / 7:44 am MDT • August 1, 2019

Comment

From: Wilkinson (2005), ggplot2 package, TechCrunch

2.b) It transfers: Data wrangling

Normal forms & database normalization

From: Codd (1970)

3. It bridges the gap between tools for *learning* statistics & tools for *doing* statistics

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....*

tidyverse principle #4: Design for humans

GitHub

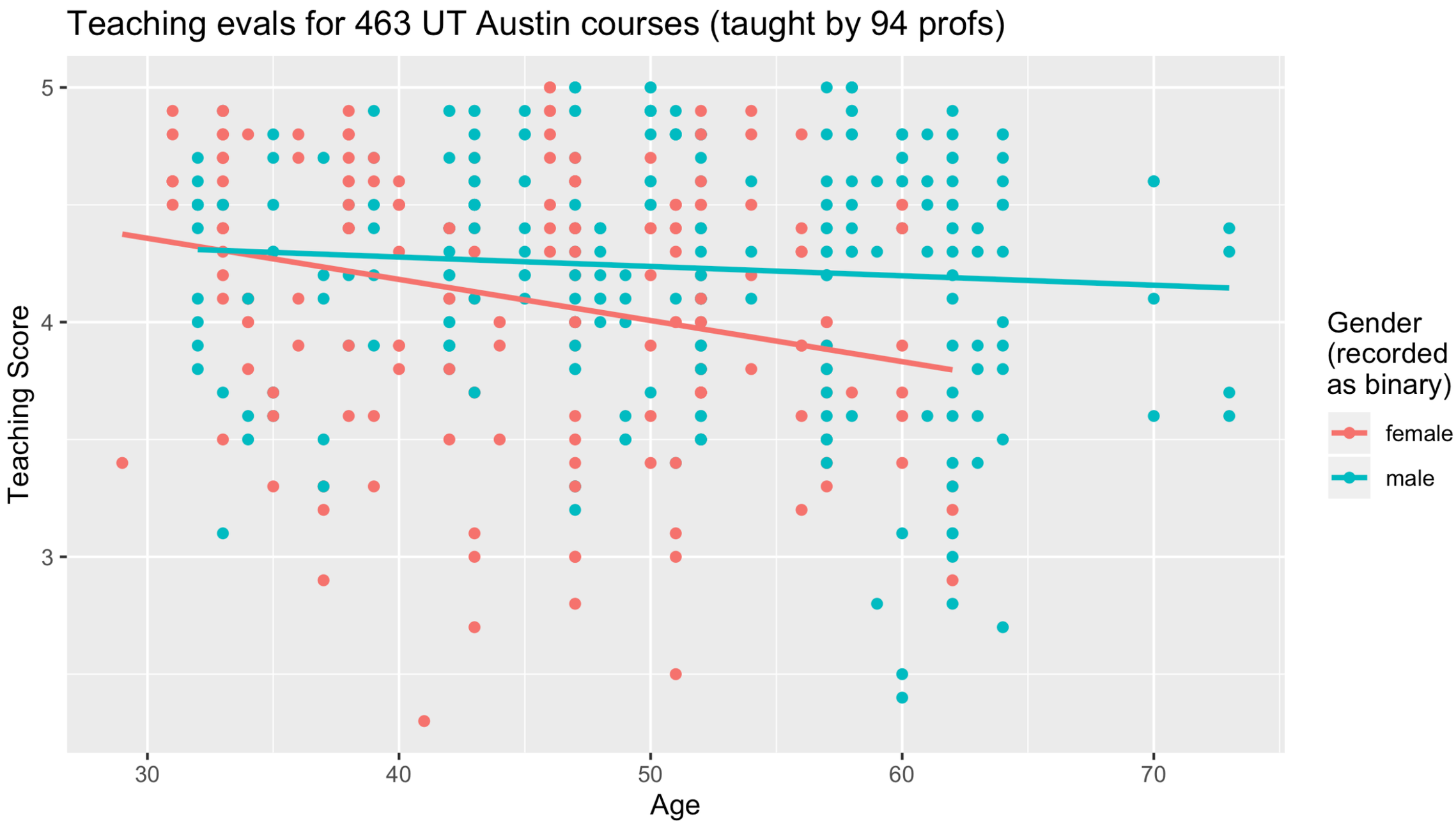
stackoverflow

From: McNamara (2015), Robinson blogpost, tidy tools manifesto

Using the tidyverse in intro stats for:

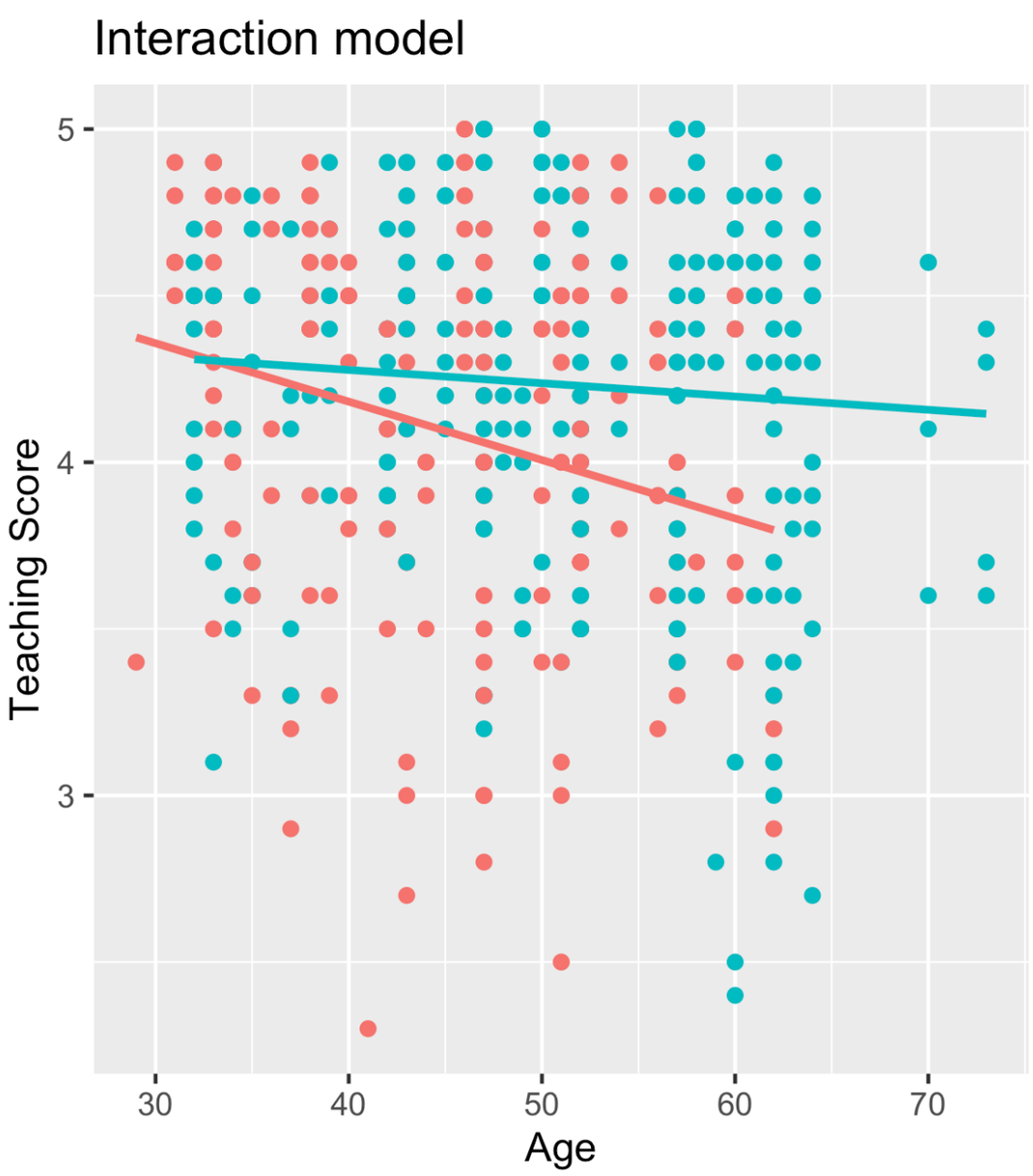
1. Statistical modeling
2. Statistical inference

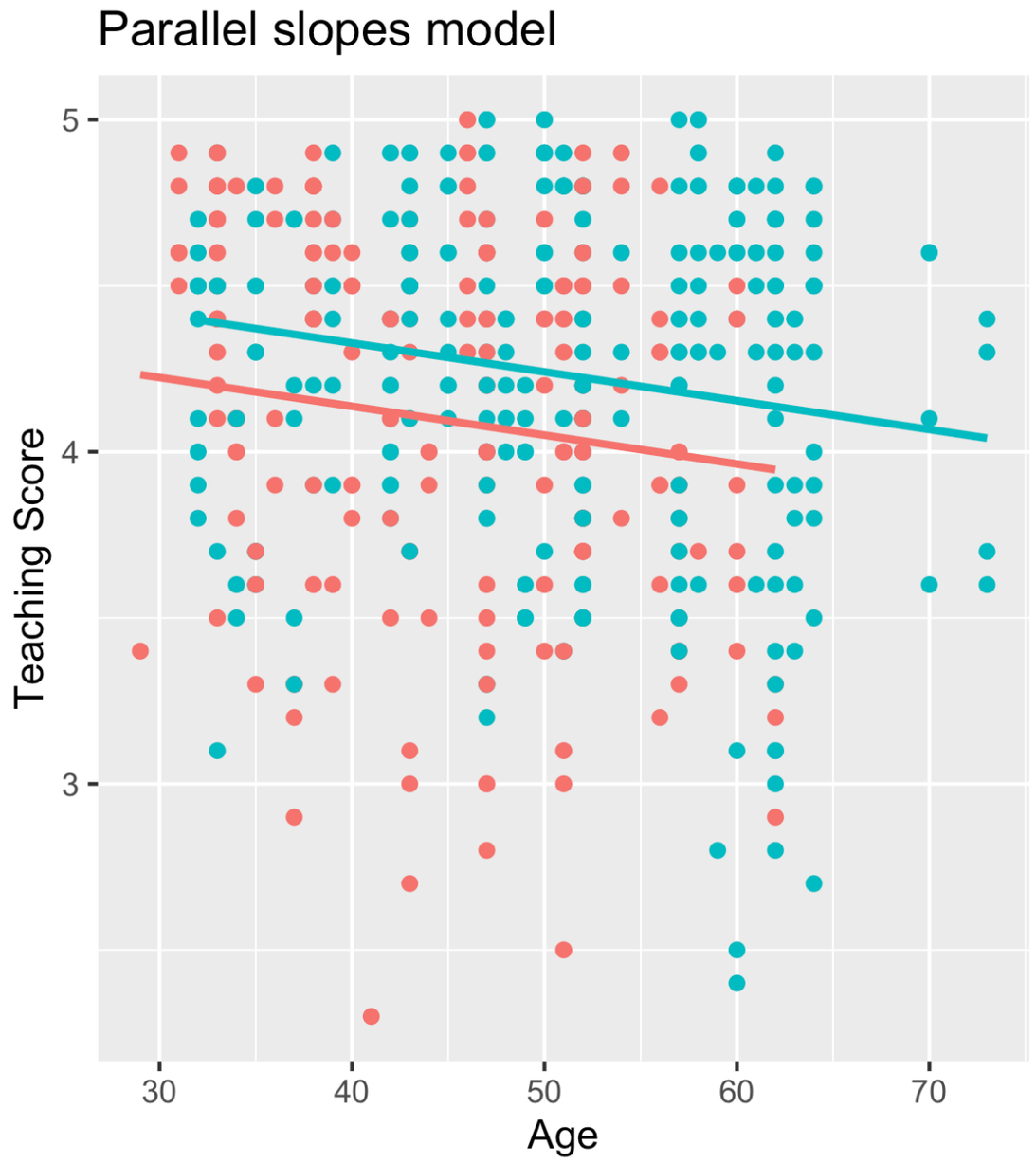
EDA to Motivate Statistical Modeling



From: [Chance Magazine]

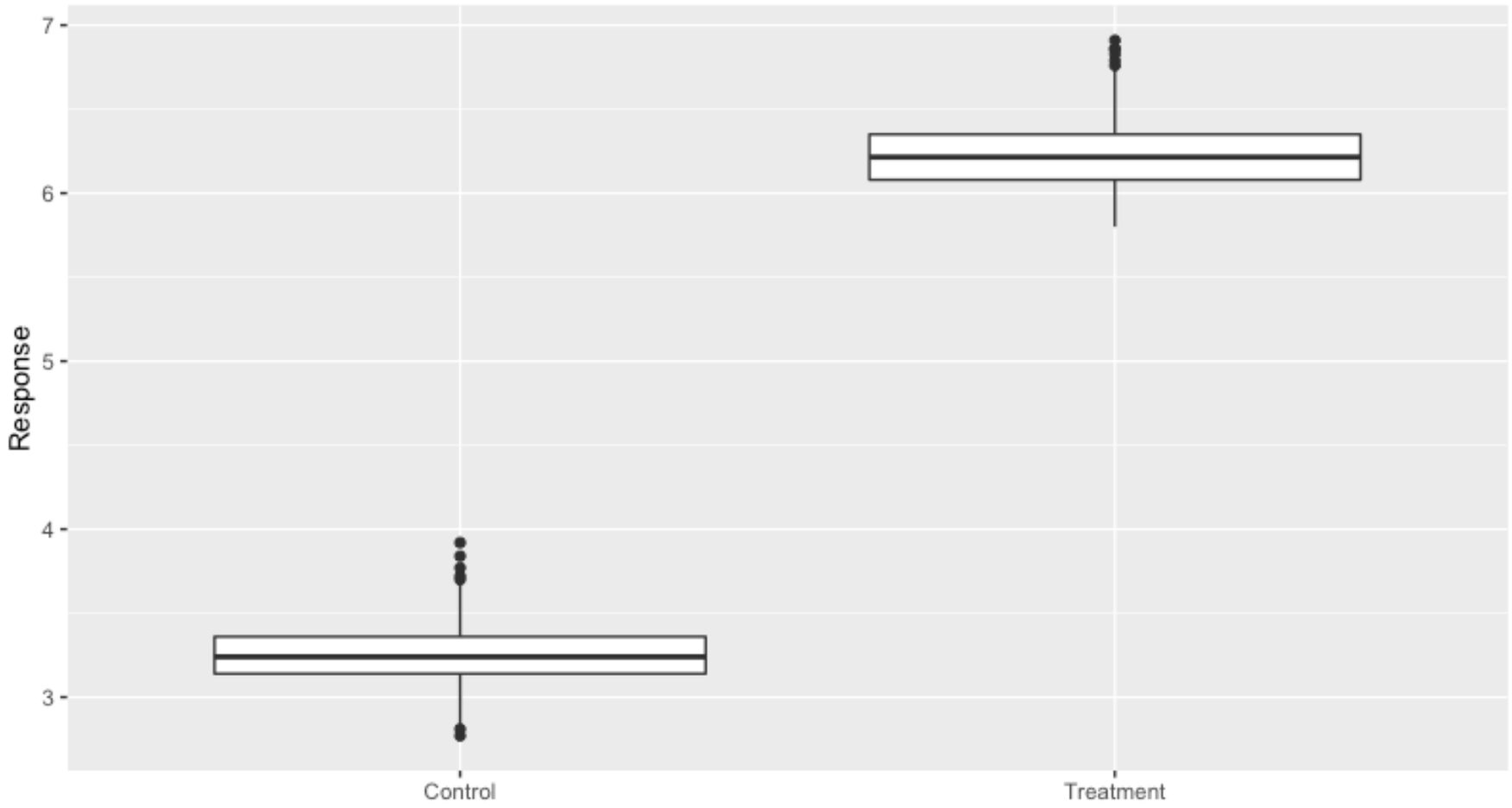
EDA to Motivate Model Selection





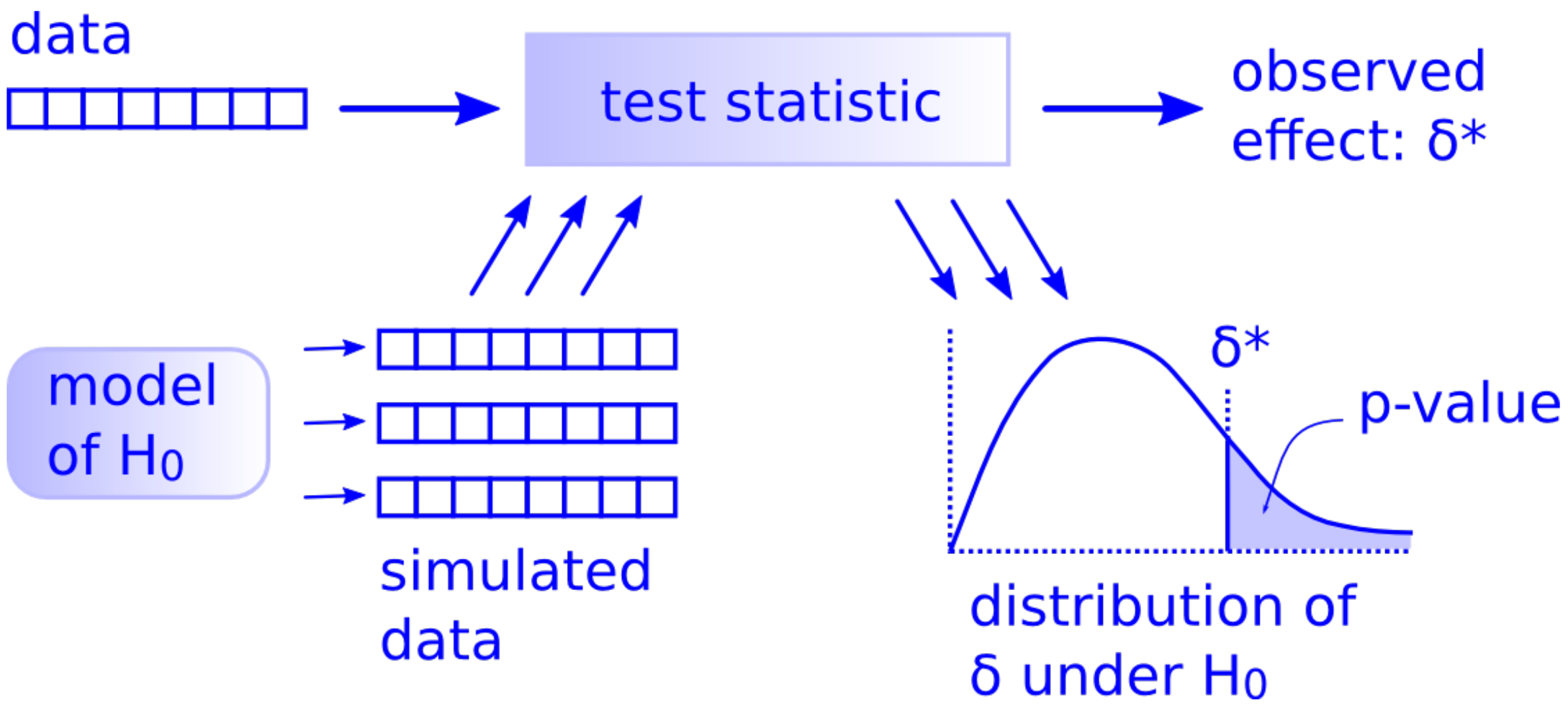
EDA to Motivate Statistical Inference

A “you don’t need no PhD in Statistics” moment:



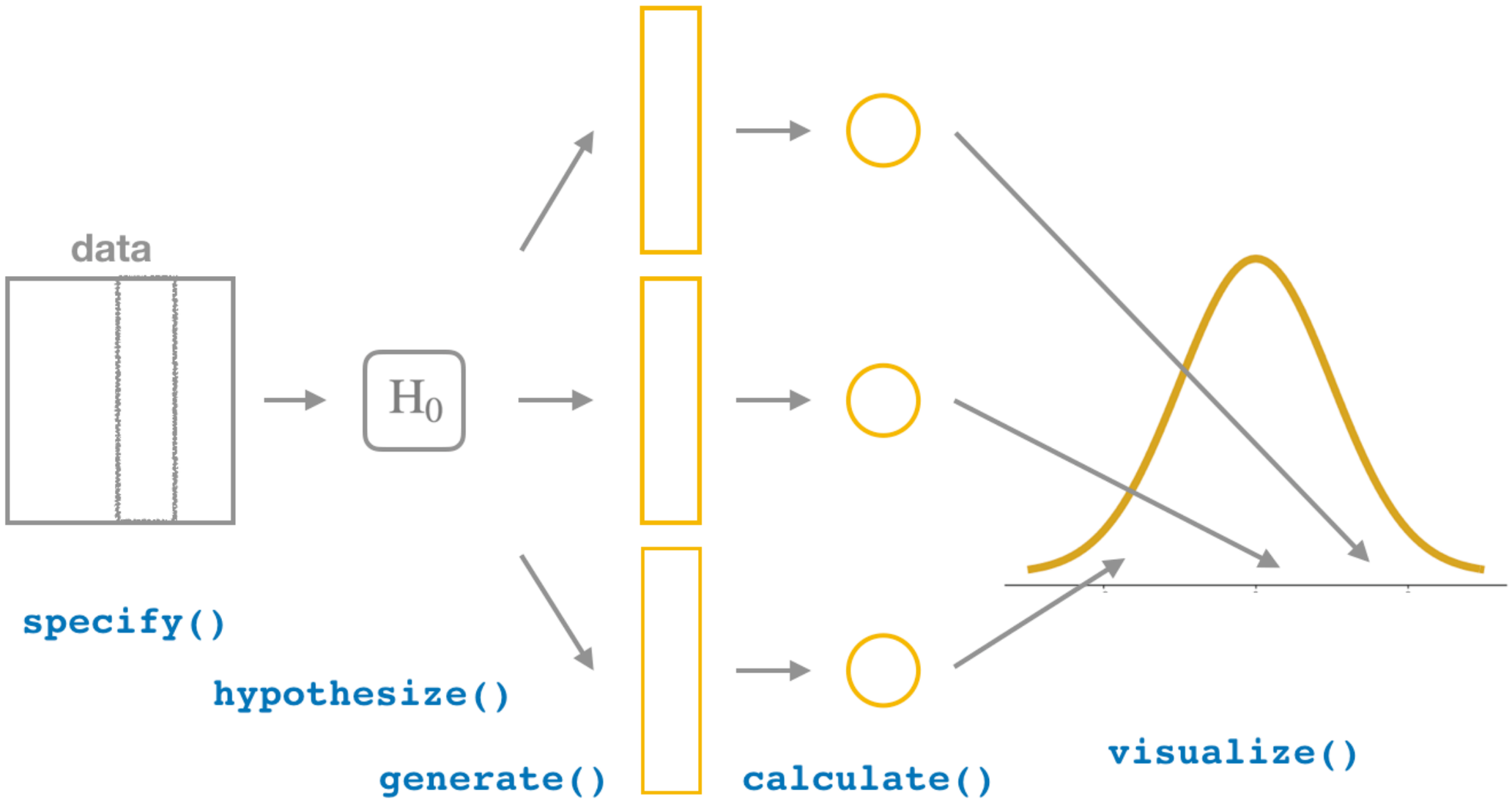
vs just reporting “the p-value is 0”

“There is only one test”



From: [Downey blogpost]

infer package for “tidy” statistical inference



From: [Bray, Ismay, Chasnovski, Baumer, and Cetinkaya-Rundel]

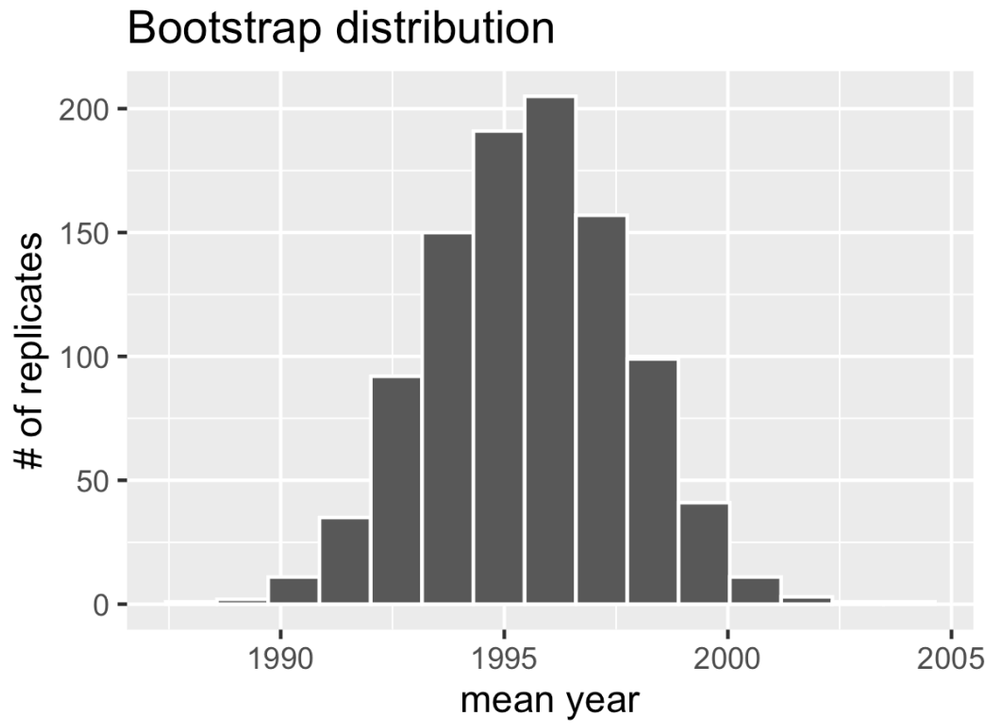
Ex: Inferring the mean year of all 🇺🇸 pennies

```
> library(moderndive)
> pennies_sample
# A tibble: 50 x 2
      ID  year
   <int> <dbl>
 1     1  2002
 2     2  1986
 3     3  2017
 4     4  1988
 5     5  2008
 6     6  1983
 7     7  2008
 8     8  1996
 9     9  2004
10    10  2000
# … with 40 more rows
```

Using bootstrap resampling with replacement:

```
library(tidyverse)
library(infer)

pennies_sample %>%
  specify(response = year) %>%
  generate(reps = 1000) %>%
  calculate(stat = "mean")
```



How to make room for the tidyverse

- Drop probability theory
- IMO: De-emphasize χ^2 tests & ANOVA
- Lean on “There is only one test” framework
- Drop asymptotic theory in favor of simulation based inference

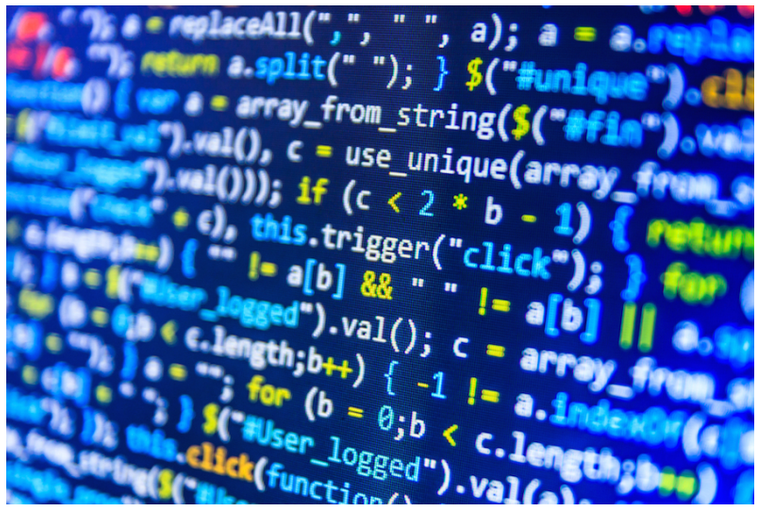
Guiding Paper

"Mere Renovation is Too Little Too Late: We Need to Rethink Our Undergraduate Curriculum from the Ground Up" by Cobb (2015)

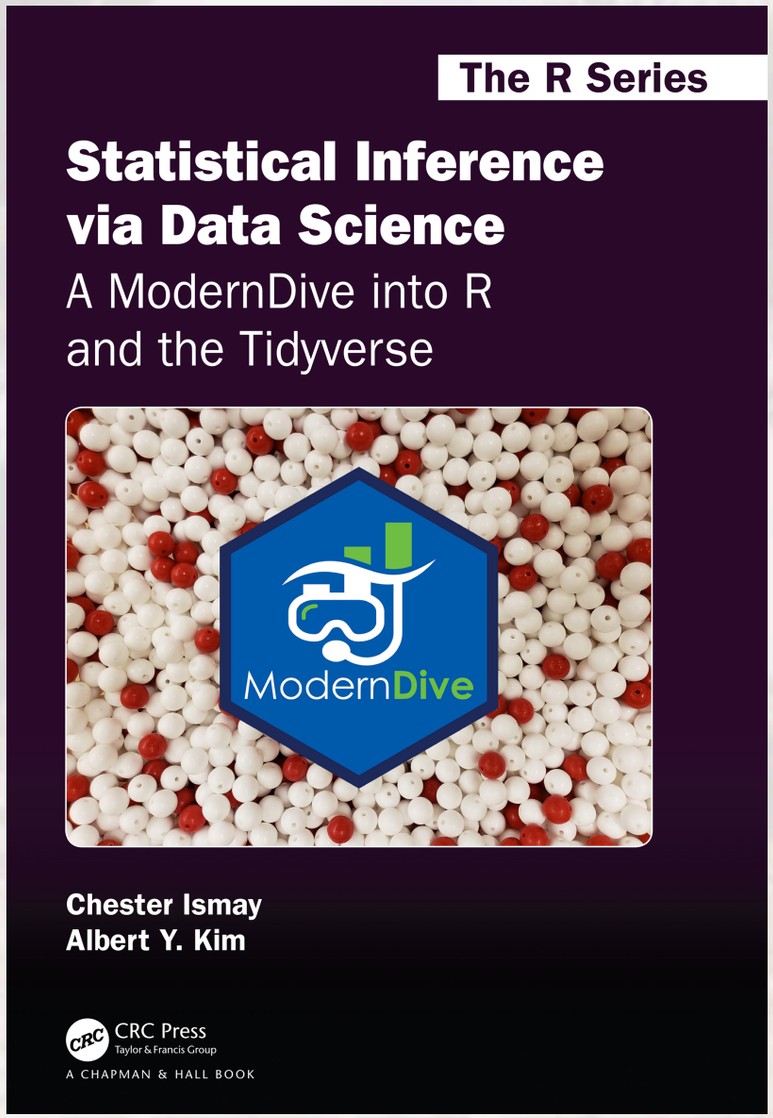
- Make fundamental concepts accessible
- Minimize prerequisites to research
- Replace "mathematics" with "computation" as the *engine of statistics*

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

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



For more info check out:



- Available free online at moderndive.com
- Print copies now on sale at Taylor & Francis booth & CRC Press website: Use discount code ASA18
- Slides available at twitter.com/rudeboybert

`infer`-ring the mean year of all 🇺🇸 pennies using bootstrap resampling with replacement

Collect a sample of 50 pennies (is sampling representative?) →

```
> library(moderndive)
> pennies_sample
# A tibble: 50 x 2
      ID  year
   <int> <dbl>
 1     1  2002
 2     2  1986
 3     3  2017
 4     4  1988
 5     5  2008
 6     6  1983
 7     7  2008
 8     8  1996
 9     9  2004
10    10  2000
# … with 40 more rows
```

Then generate 1000 resamples with replacement of size 50, compute mean for each. →

```
library(tidyverse)
library(infer)

pennies_sample %>%
  specify(response = year) %>%
  generate(reps = 1000) %>%
  calculate(stat = "mean")
```

Then visualize. In what range do “most” values lie? →