Rsample Sampler

Resampling Methods with R

R-Ladies San Francisco 2018-08-15 *Fanny Chow* *@frannystats*

Setup! 🍽️✋🏽

- Download 
  - R (https://cran.r-project.org)
  - RStudio (https://www.rstudio.com/products/rstudio/download/#download)
- Within RStudio: Click Tools > Install Packages… 📦
  - tidyverse
  - tidymodels

If you need help, please raise your hand✋🏽 !

File > New File > RScript

Then load the packages by typing into your RScript:

```
library(tidyverse)
library(tidymodels)
```

Let's run this code!

With your cursor on the text:

- hit 'command enter' on Mac
- 'cntrl enter' on Windows.

>Whoami

Introduced to STEM 

"This is cool but I'm terrified." 

"You can do it!" 


"You should major in statistics." 


---

startups

differential privacy

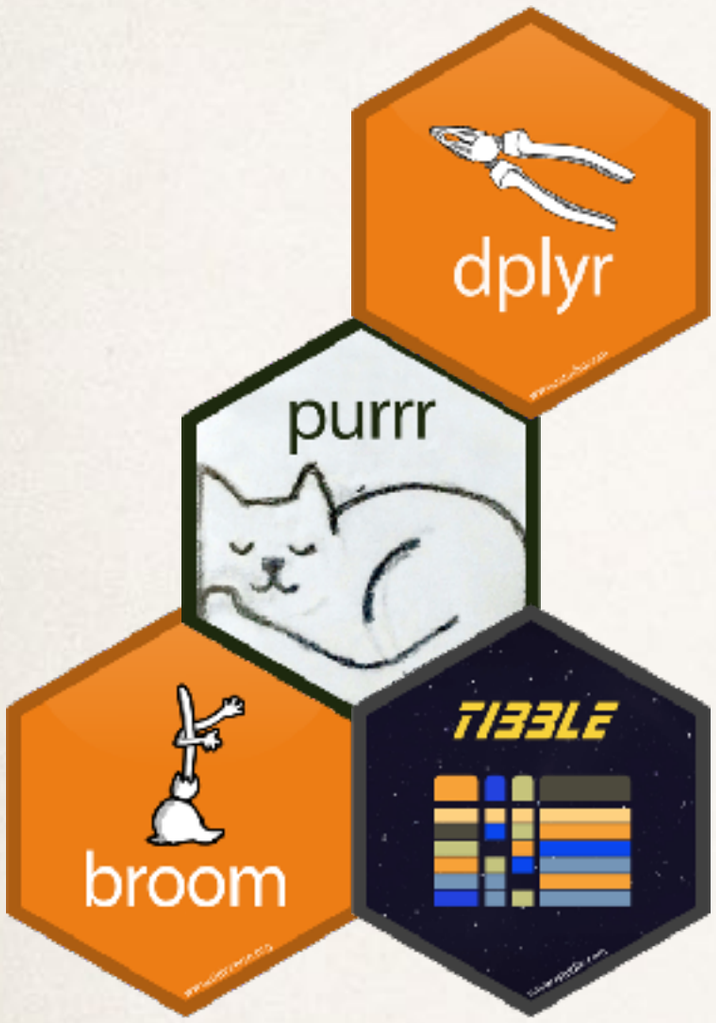
web security

build & deploy models

Intern 


Introducing

- Consistent model interfaces
- High and low-level APIs with practical defaults
- Suite of modular tools
  - functional programming (purrr)
  - Chaining %>%
  - Tidy data (2nd, 3rd Normal Form)



https://github.com/tidymodels

https://tidymodels.github.io/rsample

Functions & data structures specifically for resampling data

Infrastructure to assess and validate model performance

Easy to keep track of how your data is split

Recall: what is sampling?

We can't measure every person, place, or thing.

So we randomly collect a sample that's reasonably representative of the population we want to make inferences or predictions about.

- U.S. Census (American Community Survey)
- A/B test measuring click-through-rate on a sample of site visitors
- Risk of heart disease by race & ethnicity

What's resampling?

Resampling is taking repeated samples from our data set.

"You randomly sample your random sample!"

Sampling

Resampling

Focus

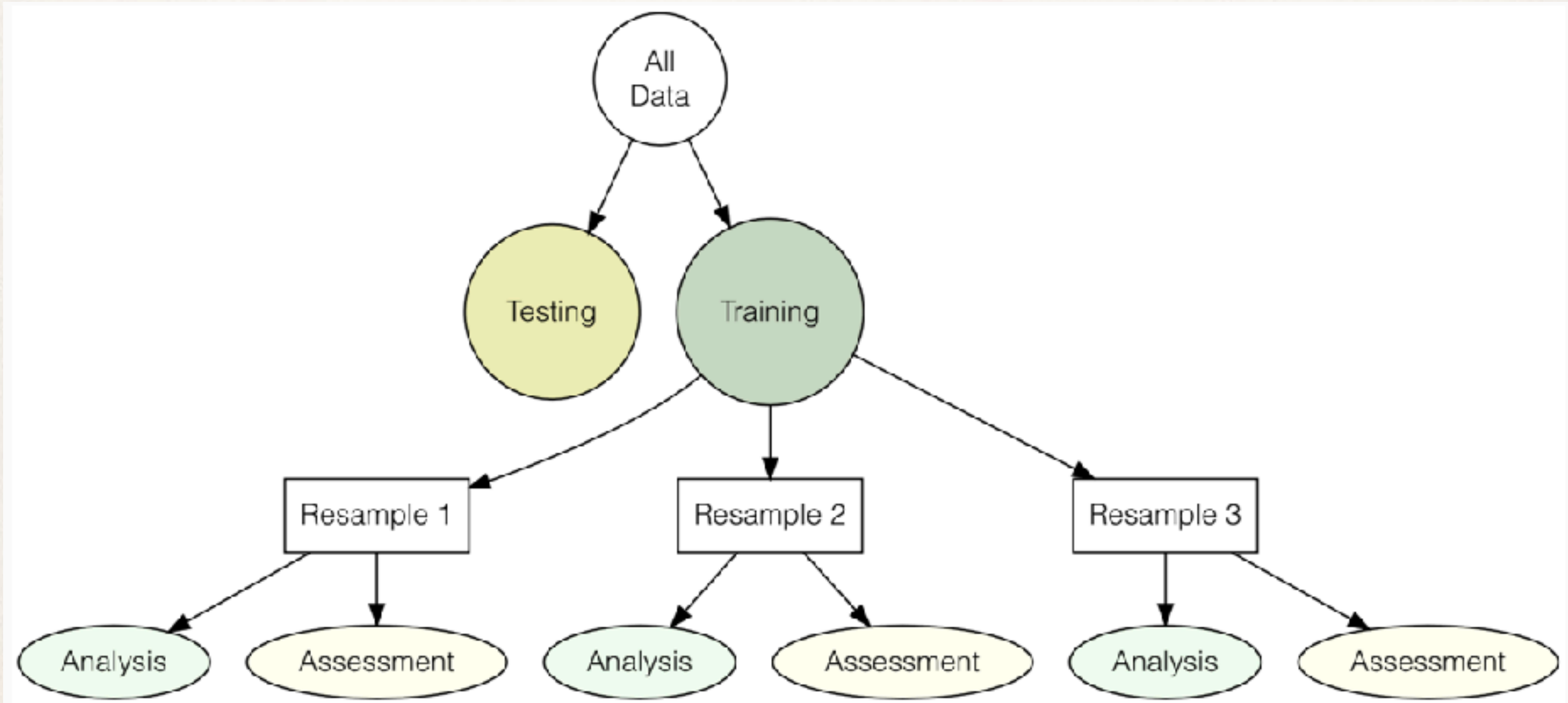
Estimate a statistic of interest

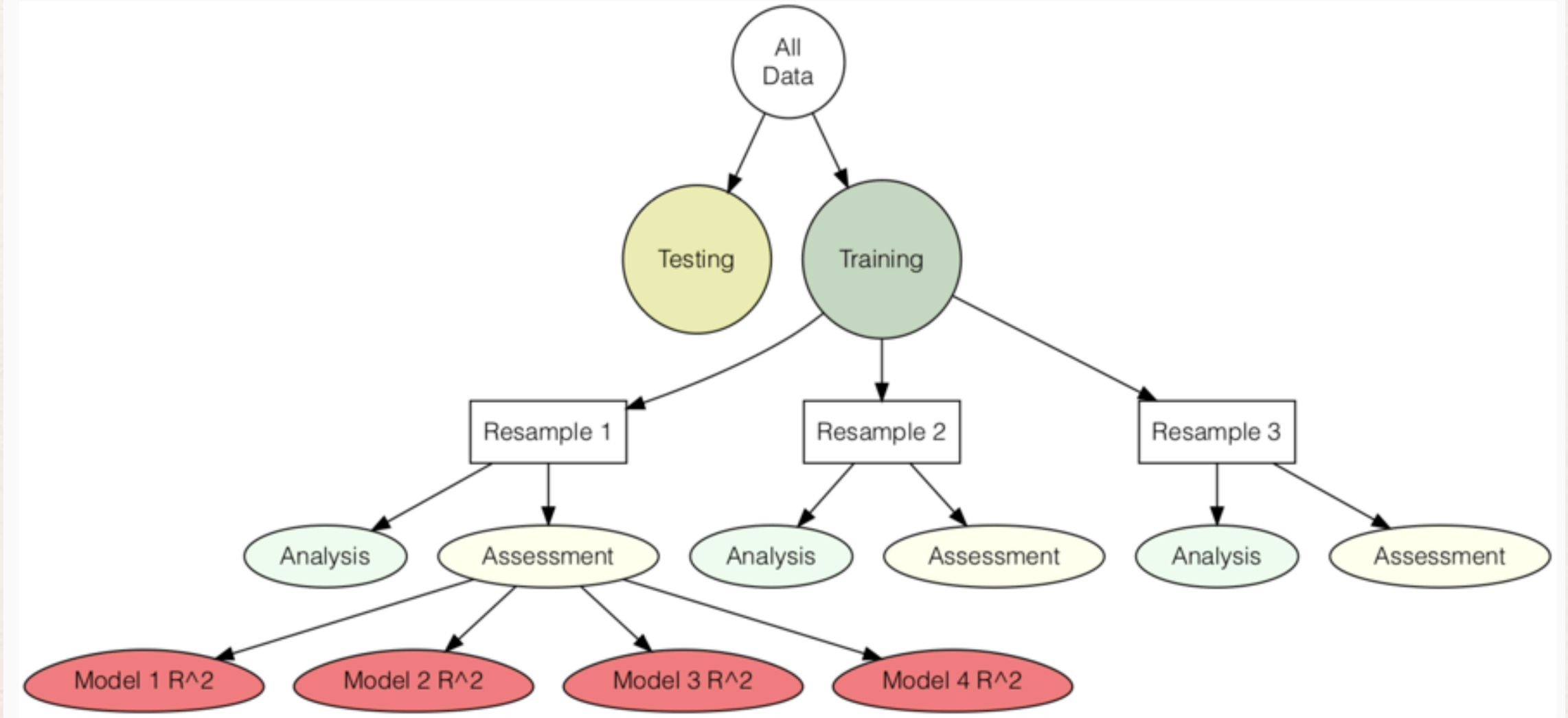
mean, median, regression coefficients

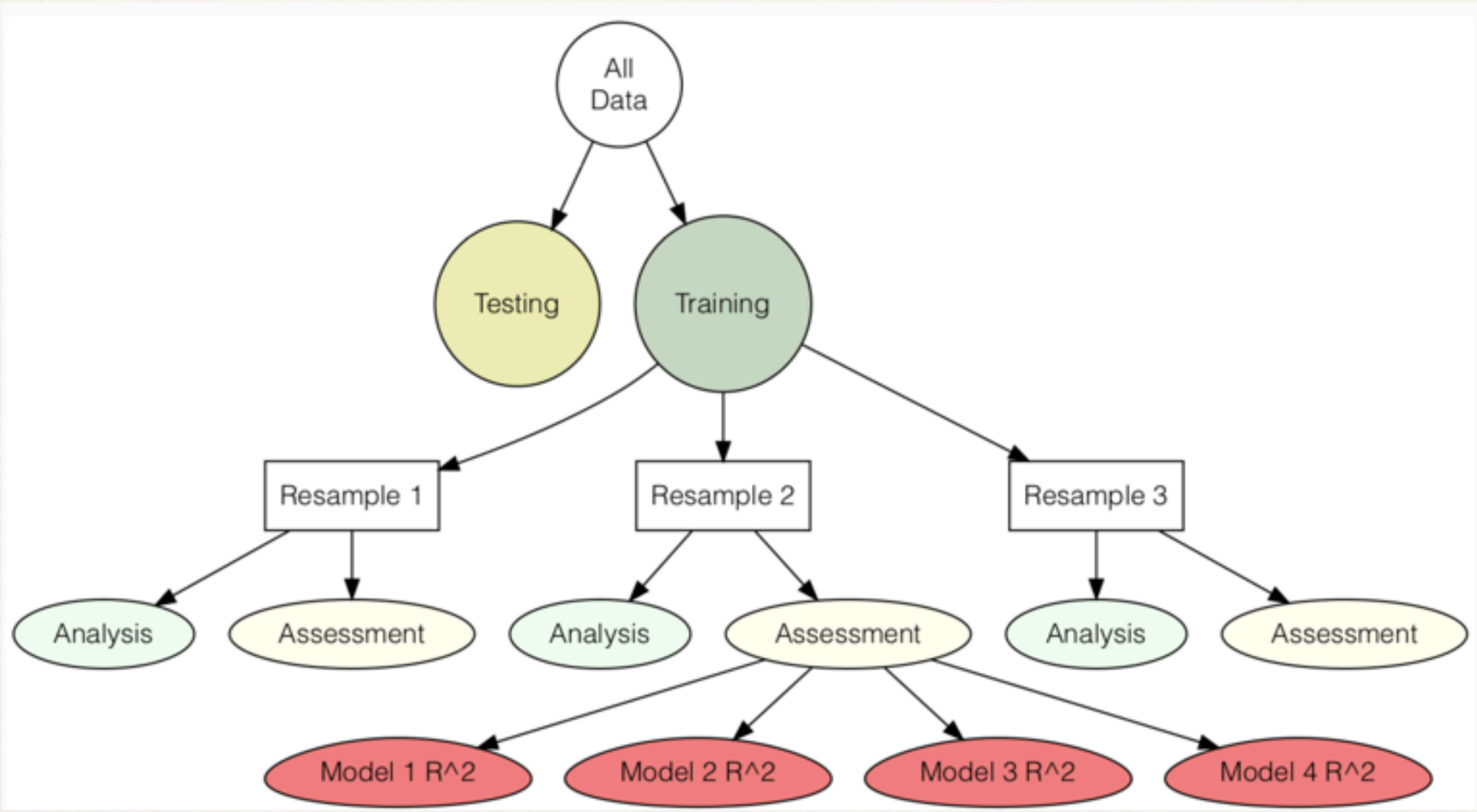
Gauge the variability or uncertainty in our statistic of interest

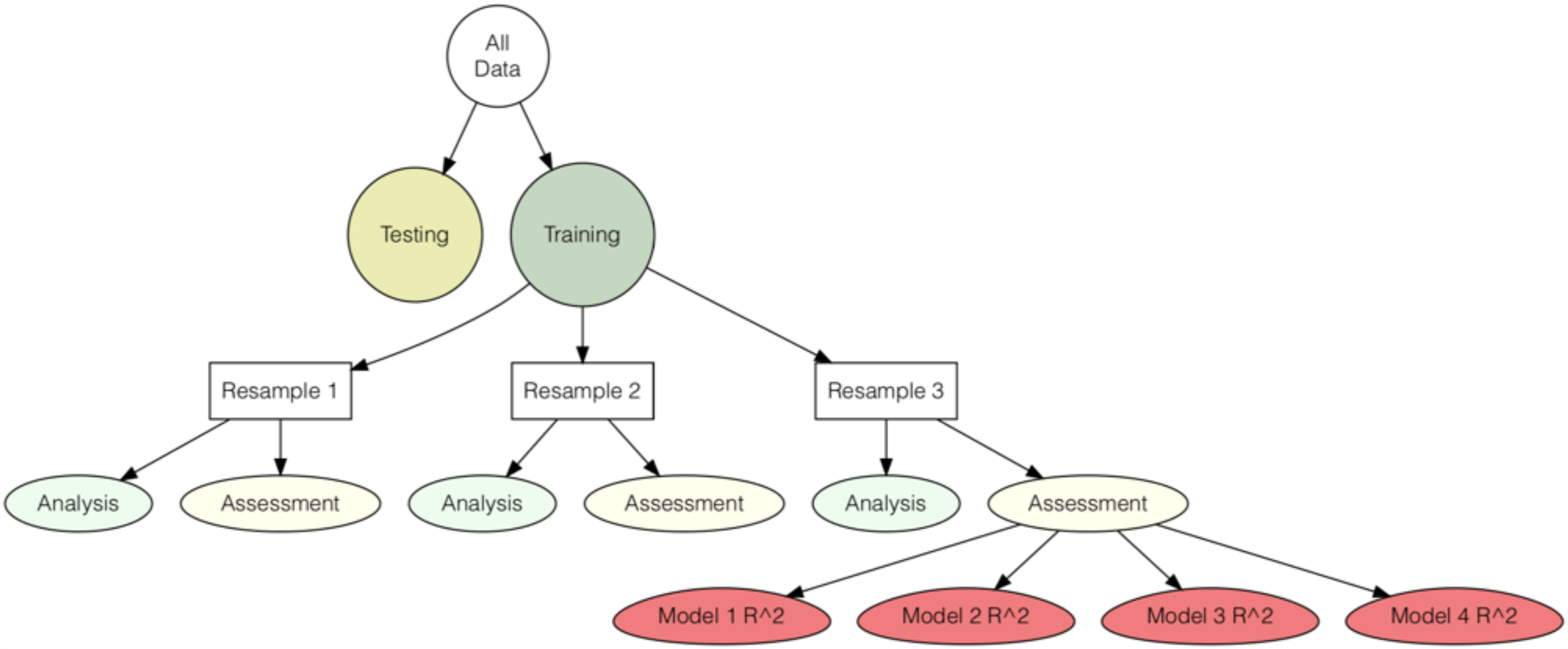
standard error, standard deviation, confidence intervals

What does resampling look like?


All
Data
Testing
Training
Resample 1
Resample 2
Resample 3
Analysis
Assessment
Analysis
Assessment
Analysis
Assessment


All Data
Testing
Training
Resample 1
Resample 2
Resample 3
Analysis
Assessment
Analysis
Assessment
Analysis
Assessment
Model 1 R^2
Model 2 R^2
Model 3 R^2
Model 4 R^2


All Data
Testing
Training
Resample 1
Resample 2
Resample 3
Analysis
Assessment
Analysis
Assessment
Analysis
Assessment
Model 1 R^2
Model 2 R^2
Model 3 R^2
Model 4 R^2


All Data
Testing
Training
Resample 1
Resample 2
Resample 3
Analysis
Assessment
Analysis
Assessment
Analysis
Assessment
Model 1 R^2
Model 2 R^2
Model 3 R^2
Model 4 R^2

Initial Split into Test & Train

We've only got so much data so budget wisely! We take a random portion of our data to **train** (*fit our model)* on. And we set aside the rest to test or (*validate*) our models on.

Luckily the `initial_split` function in rsample makes it easy!

```
names_df <- tribble(
   ~name, ~fav_num,
      "Dana", 4,
     "Irene", 8,
      "Mara", 2,
      "Ming", 5,
    "Chaita", 10,
       "Jen", 7,
     "Becky", 4,
     "Jenny", 9,
      "Mine", 2,
     "Emily", 6
           )

boxplot(names_df$fav_num)
View(names_df)
nrow(names_df)
str(names_df)
summary(names_df)

 first_split <- initial_split(names_df)
  train_names <- analysis(first_split)
 test_names <- assessment(first_split)
```

Start a new script and try the same thing on a different data set.
This time let's use the attrition data set within the rsample package.

Let's load the data by typing:

```
data("attrition")
```

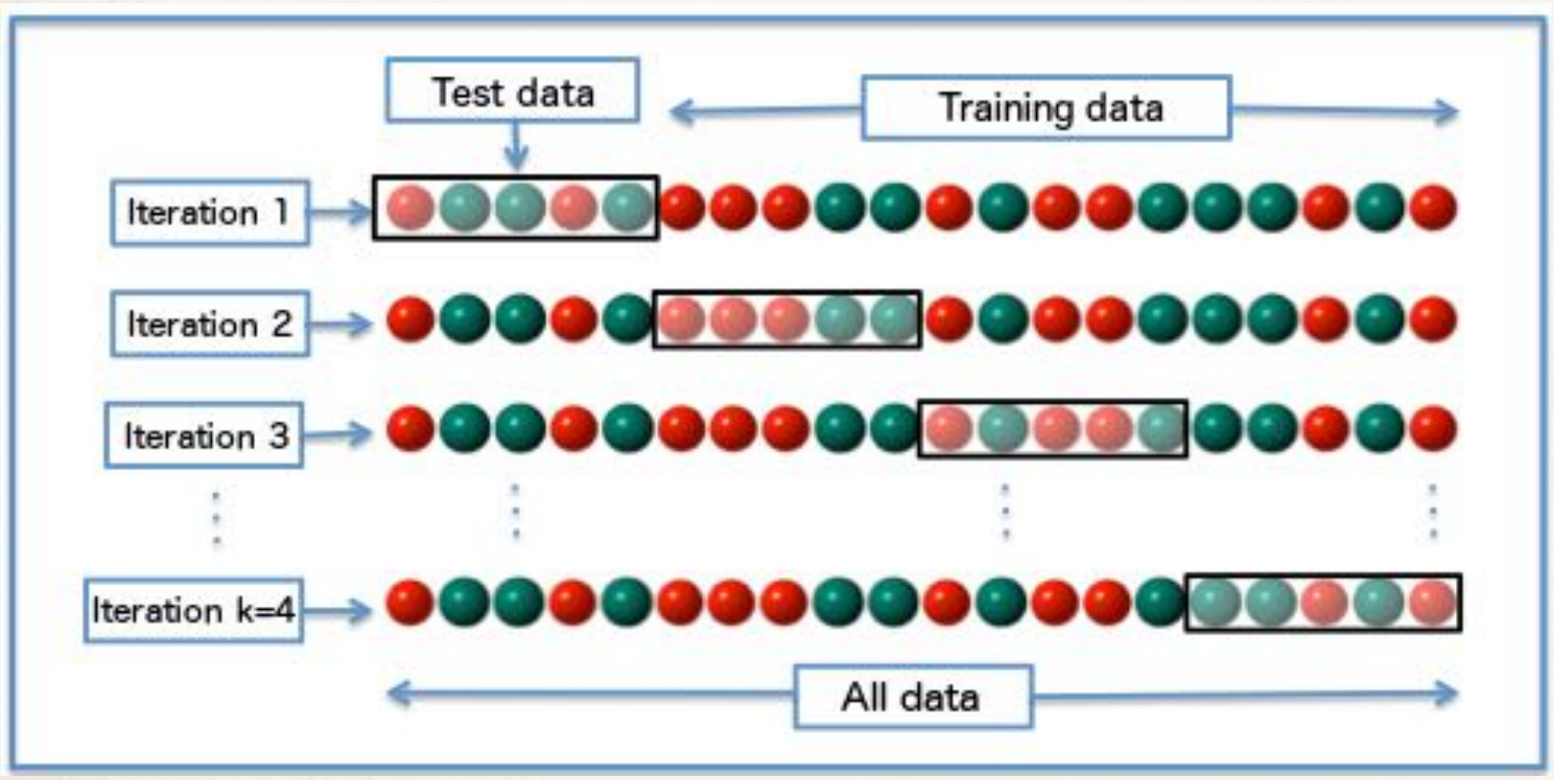
Today's Resampling Sampler

1. Cross-validation

2. Bootstrapping

1. Cross Validation

V-Fold Cross Validation



Wikipedia

Used to test how the results of a statistical analysis will generalize to a new situation

Let's try vfold_cv()

```
vfold_cv(data, v = 10, repeats = 1, strata = NULL, ...)
```

Let's try vfold_cv()

- We will get a rset object
  - Which contains many resamples

- Each resample is an rsplit object

It's easier to understand when we see it in action. Let's try.

- Great! We have many tidy resamples!

- We can use functional programming functions from the purr package 😺to fit a model on each resample.

- For the sake of information overload, let's save that for next time 😀.

2. Bootstrapping

Bootstrapping

- resampling with replacement
- (all values in the sample have an equal probability of being included, including multiple times, so a value could have a duplicate)
- Can help you calculate statistics with less strict mathematical assumptions

Ex: throw 10 paper slips in a hat, pick name from a hat, write down name, throw paper back in, repeat 10x

- Bootstrapping is usually used to help us make sense of the distribution of a statistic of interest.

- But we can also use bootstrapping to perform inference.

Bootstraps function

```
bootstraps(data, times = 25, strata = NULL, apparent =
FALSE, ...)
```

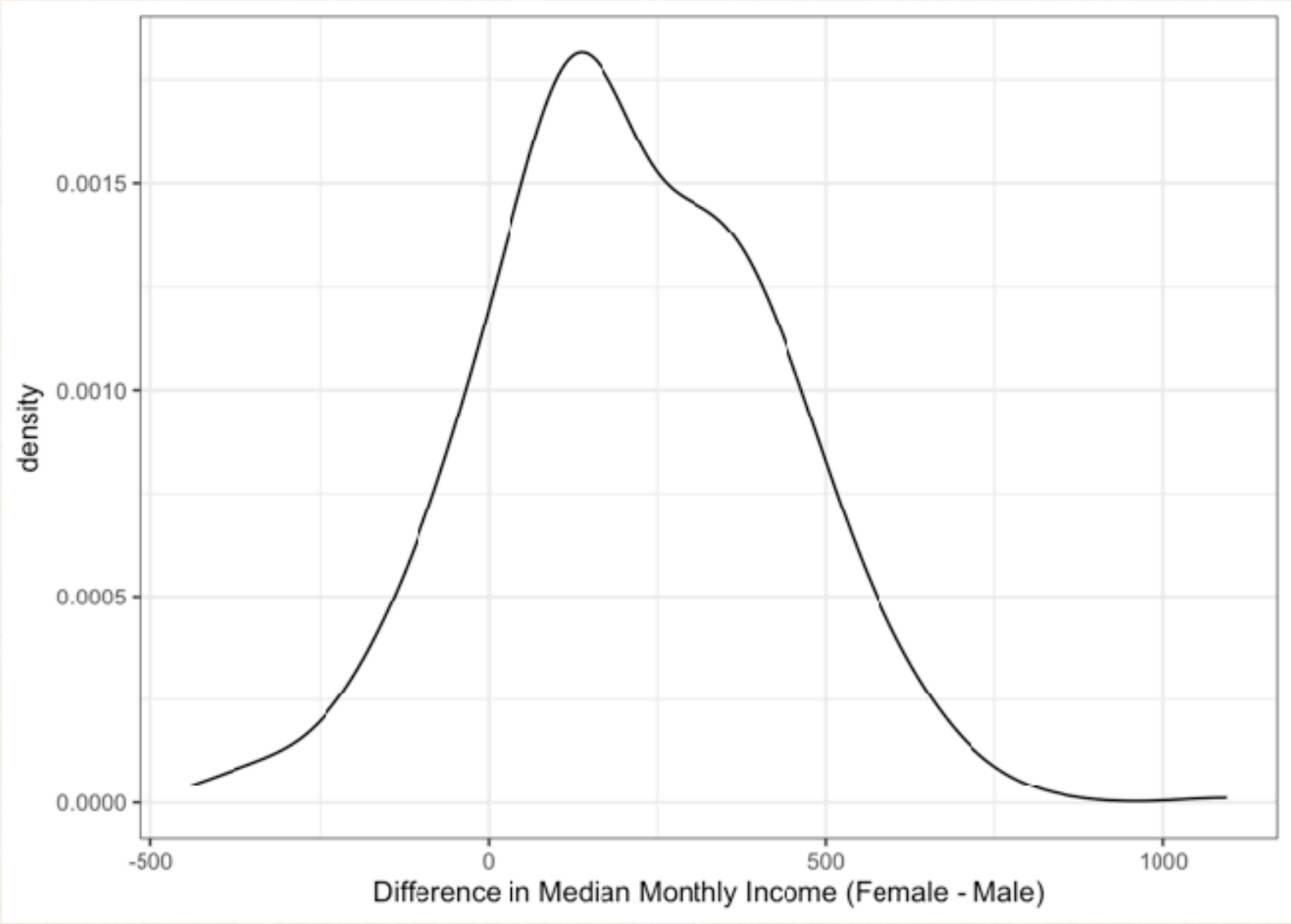
Let's say we wanted to see if there's a marked difference between median income for men and women.

An example from Max's rsample vignette

95% Confidence Interval (Percentile Method)

```
quantile(bt_resamples$wage_diff,
         probs = c(0.025, 0.500, 0.975))
#>  2.5%   50% 97.5%
#>  -207   190   618
```

Is there a difference between median monthly income between groups?



Summer Project

- Bootstrap confidence intervals
  - Percentile
  - Student-t
  - Bias-Corrected Accelerated (BCA)

Live demo

(still in dev mode, sit back and relax)

www.github.com/fbchow/rsample

Thanks! Questions?