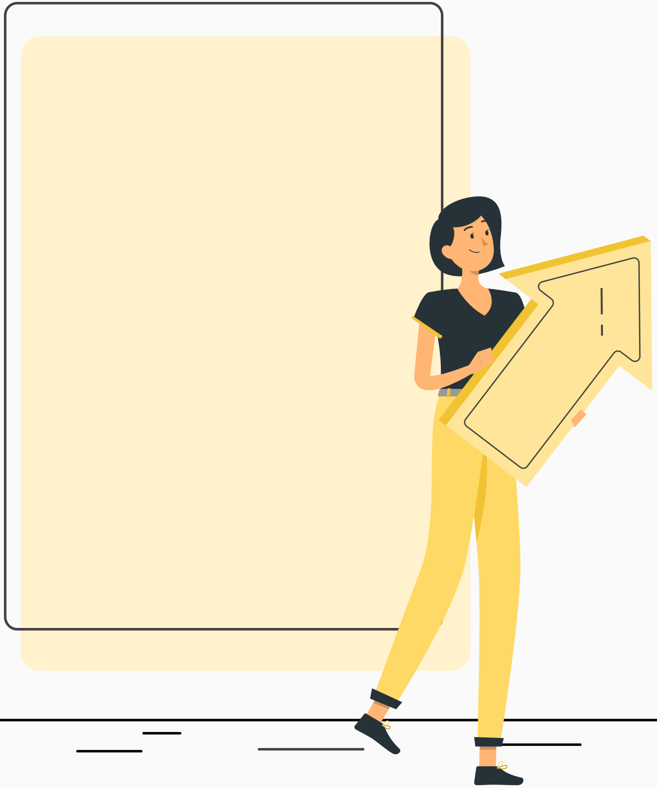
APPLIED DATA SCIENCE II

Week 5: REEEEEESSAAMMPPLLIINNGG

Kyle Scot Shank
WI-22



6:00 - 6:30

HW REVIEW

Let’s walk through it!

6:30-7:30

TOPICS + CODE!

Let’s pump up our power with some **resampling methods**!

7:30-7:45

SNACK BREAK!

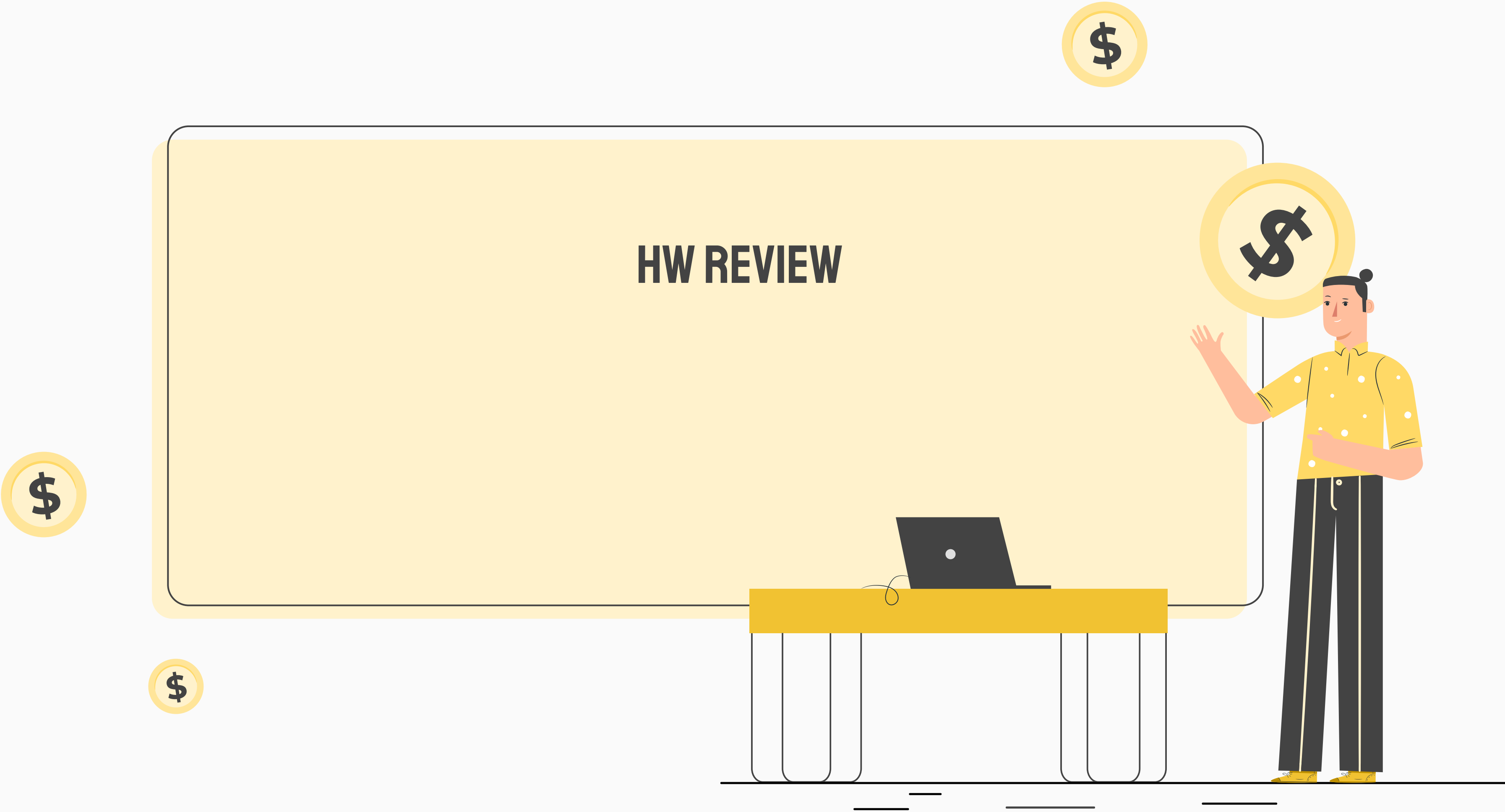
Time for some munchies

7:45 - 9:00

HANDS-ON CODE LAB

Work through stuff together

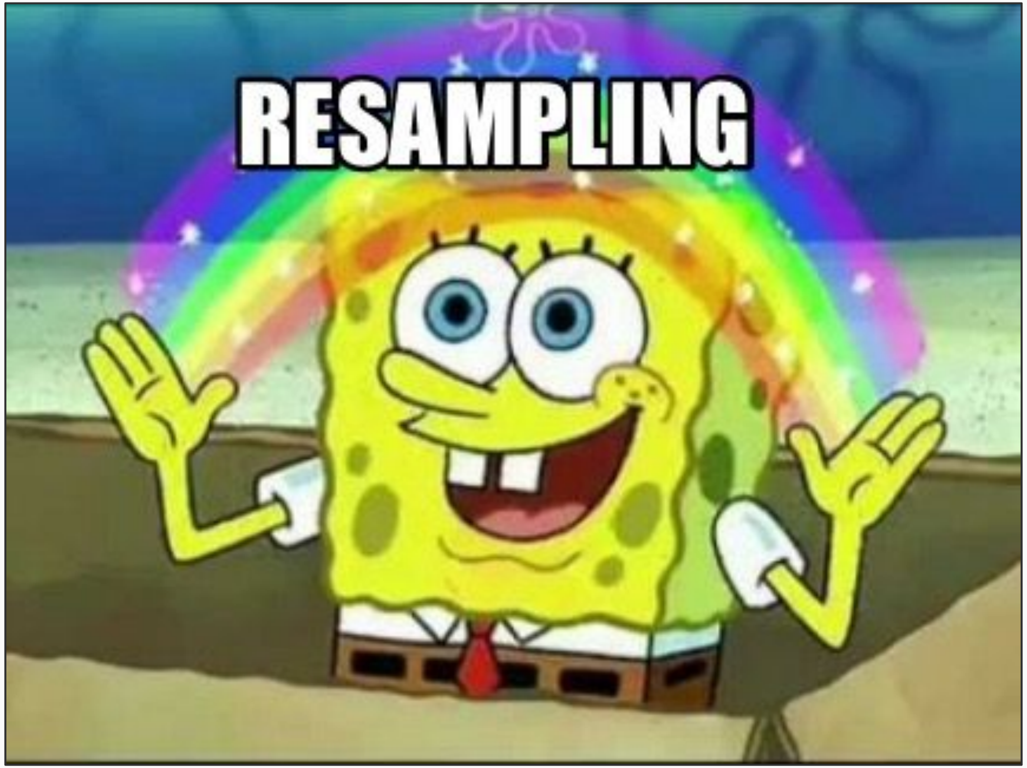
HW REVIEW



TOPIC OVERVIEW

RESAMPLING METHODS

WHAT IS THE POINT OF RESAMPLING?



RESAMPLING METHODS

Formal definition:

Resampling is a methodology of economically using a data sample to improve the accuracy and quantify the uncertainty of a population parameter.

Text

Human language definition:

“Ain’t nothing wrong with being cheap and thrifty!”

RESAMPLING METHODS

There is a ton of literature on statistical resampling and all the miraculous things it can do.

This literature is very boring.

We're going to be focused on (primarily) how we can use some of these resampling methods to make our predictive models even more accurate!

***So we're going to focus on two:** k-fold cross-validation **and** the bootstrap.*

These methods refit a model of interest to samples formed from the training set, in order to obtain additional information about the fitted model. For example, they provide estimates of test-set prediction error, and the standard deviation and bias of our parameter estimates

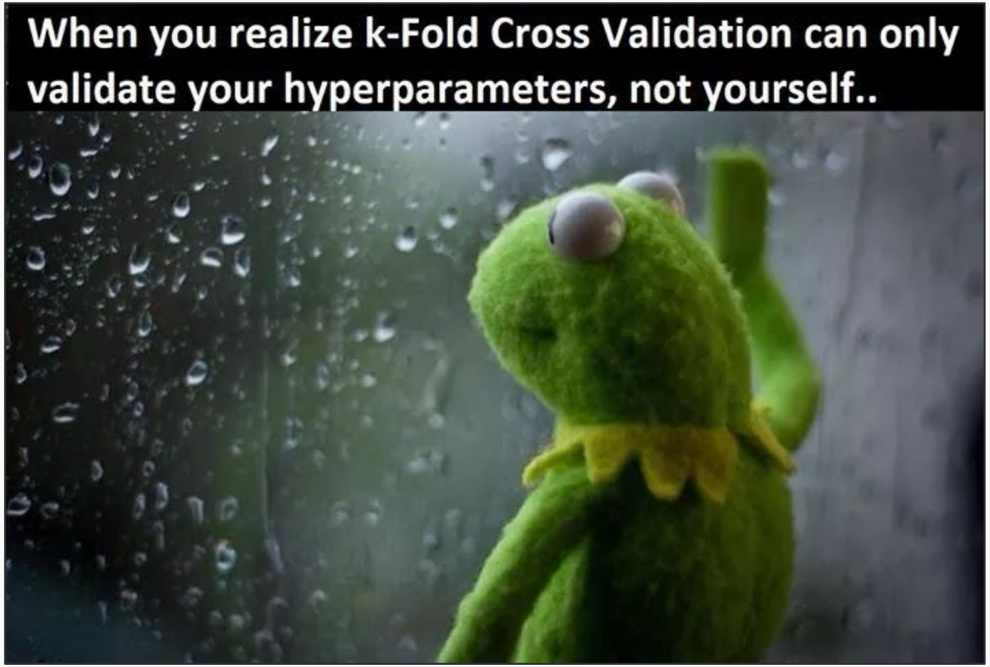
K-FOLD CROSS-VALIDATION

Recall the distinction between the test error and the training error:

> *The test error is the average error that results from using a model to predict the response on a new observation, one that was not used in training the method. The training error is just using the model to understand the differences between the predicted and actual values you used to generate your model.*

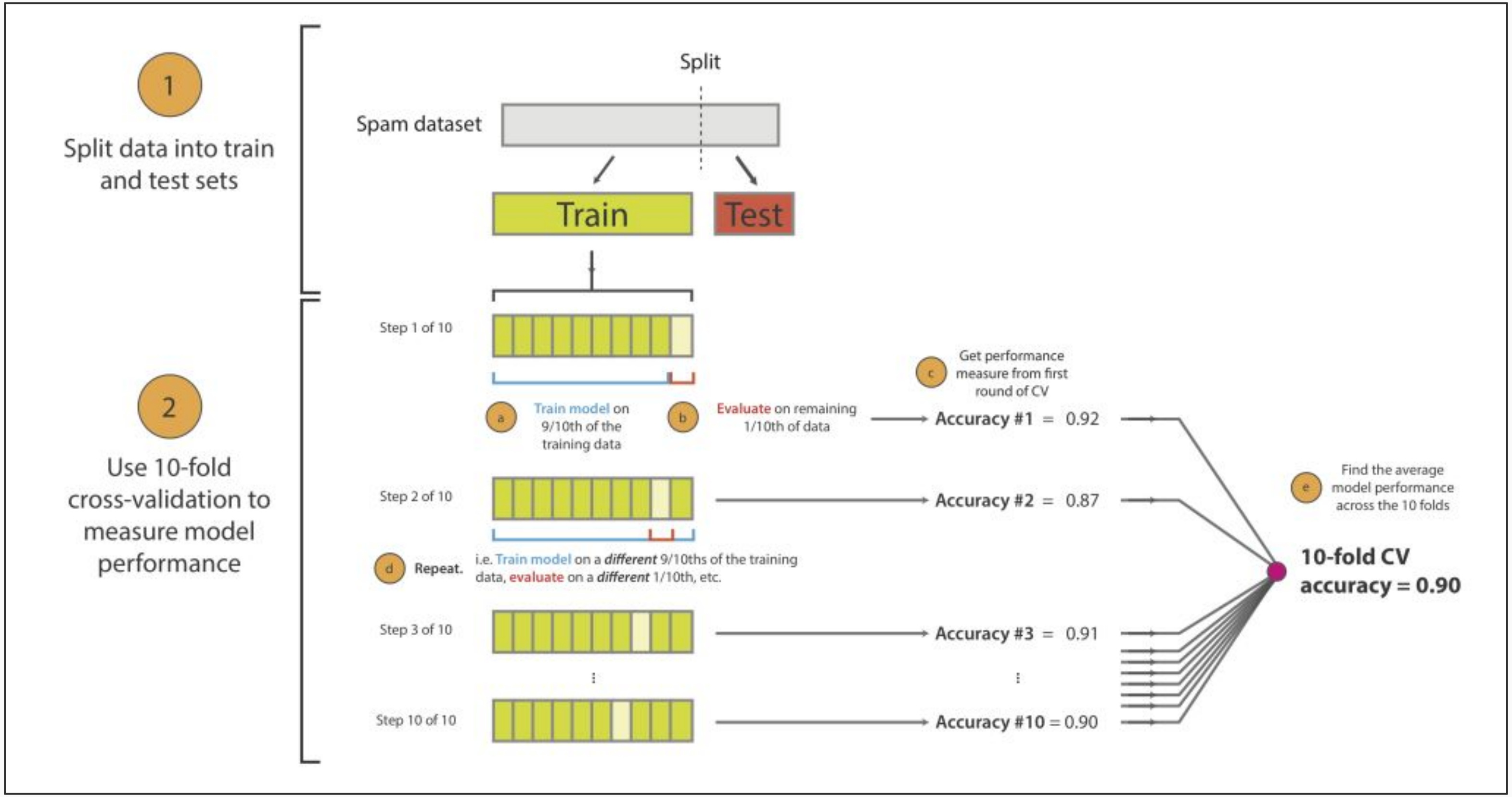
The problem we have is that training error rate often is quite different from the test error rate, and in particular the former can dramatically underestimate the latter.

*Enter **k-fold cross-validation**!*


When you realize k-Fold Cross Validation can only validate your hyperparameters, not yourself..

K-FOLD CROSS-VALIDATION

How it works:


1
Split data into train and test sets
Split
Spam dataset
Train
Test
2
Use 10-fold cross-validation to measure model performance
Step 1 of 10
a
Train model on 9/10th of the training data
b
Evaluate on remaining 1/10th of data
c
Get performance measure from first round of CV
Accuracy #1 = 0.92
Step 2 of 10
Accuracy #2 = 0.87
d
Repeat.
i.e. Train model on a *different* 9/10ths of the training data, evaluate on a *different* 1/10th, etc.
Step 3 of 10
Accuracy #3 = 0.91
Step 10 of 10
Accuracy #10 = 0.90
e
Find the average model performance across the 10 folds
10-fold CV accuracy = 0.90

K-FOLD CROSS VALIDATION

- *Let’s do this together!*

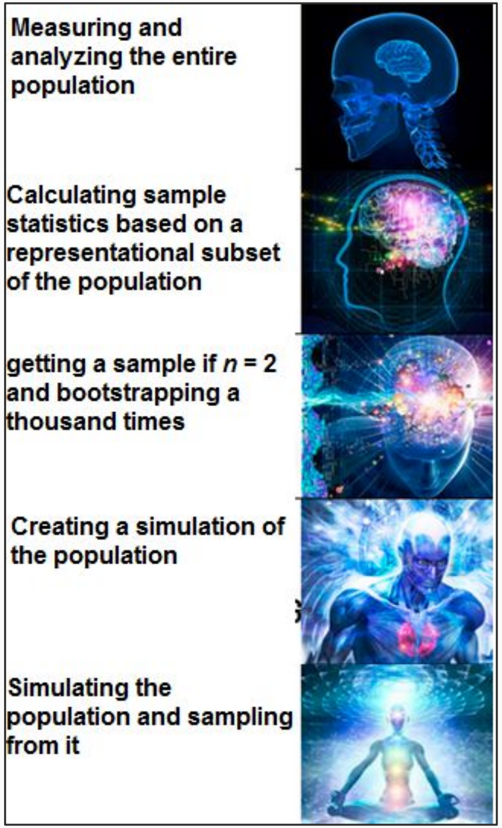
Open up R!

THE BOOTSTRAP

The bootstrap is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or predictive model.

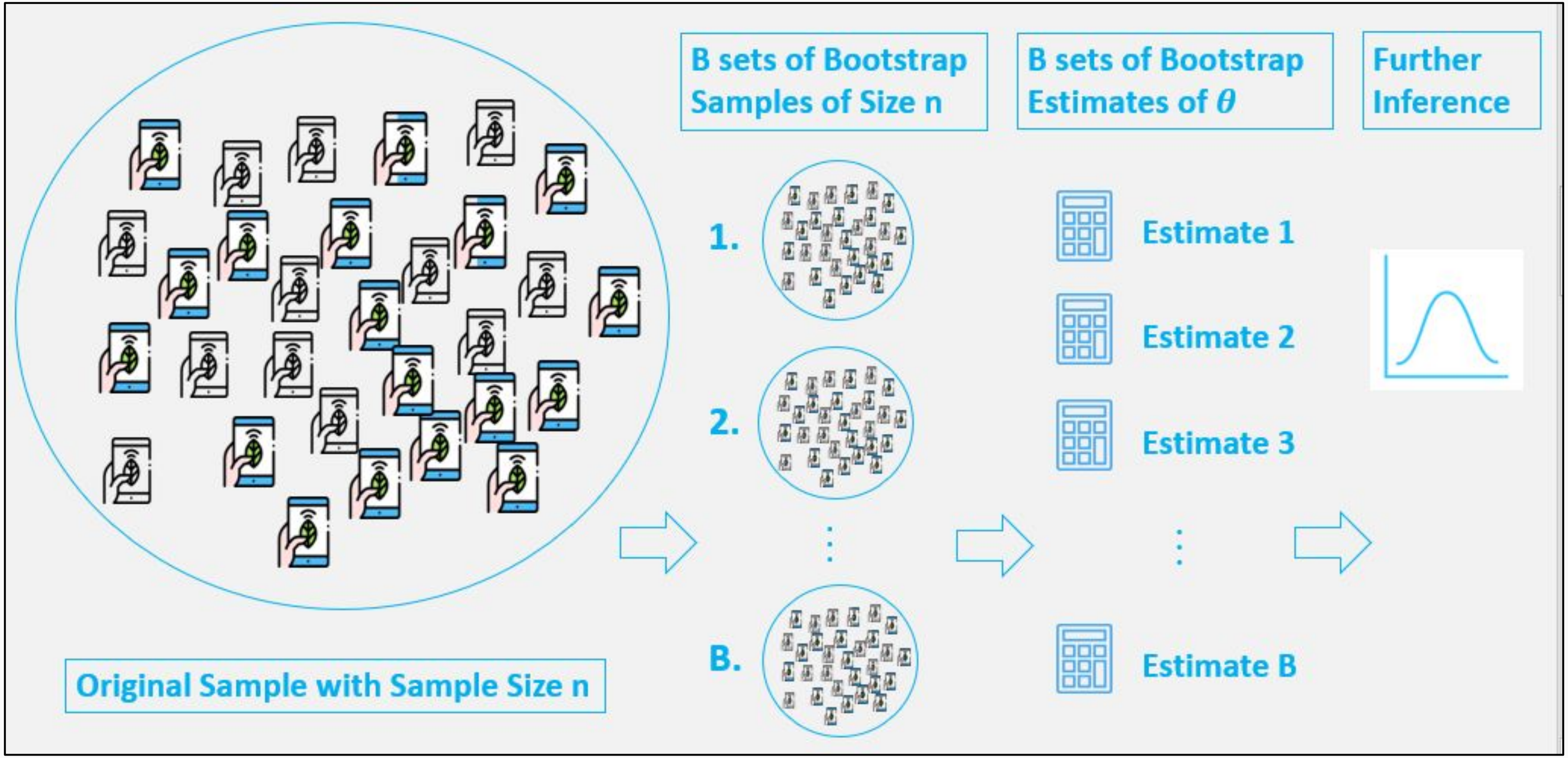
For example, it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.

*For our purposes: bootstrapping techniques let us derive **significantly** more rigorous estimates of the predictive power of our given model by allowing us to leverage resampling techniques.*


Measuring and analyzing the entire population
Calculating sample statistics based on a representational subset of the population
getting a sample if *n* = 2 and bootstrapping a thousand times
Creating a simulation of the population
Simulating the population and sampling from it

THE BOOTSTRAP

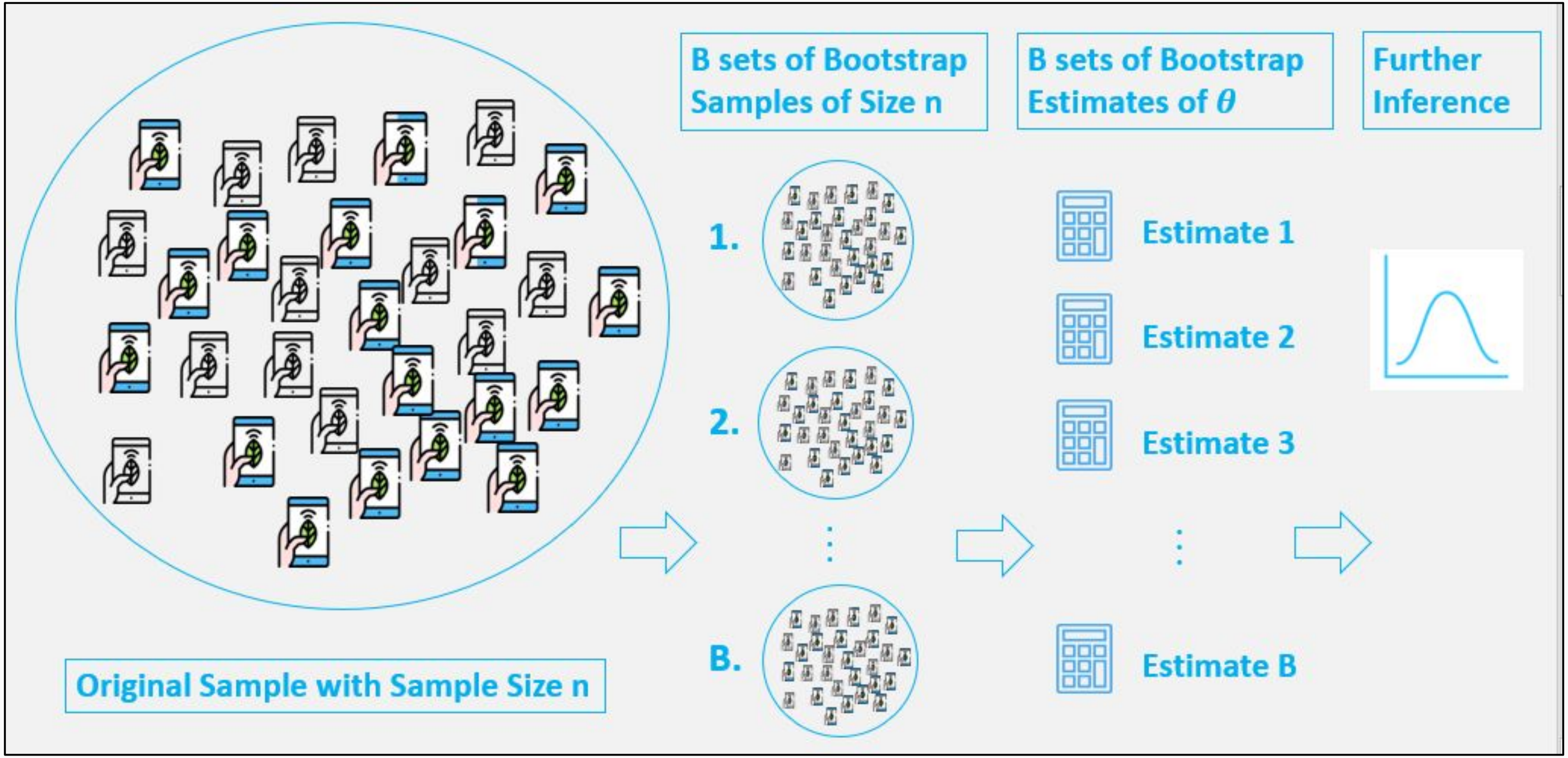
How it works:



B sets of Bootstrap Samples of Size n
B sets of Bootstrap Estimates of θ
Further Inference
1.
2.
B.
Estimate 1
Estimate 2
Estimate 3
Estimate B
Original Sample with Sample Size n

THE BOOTSTRAP

How it works:



B sets of Bootstrap Samples of Size n
B sets of Bootstrap Estimates of θ
Further Inference
1.
2.
B.
Estimate 1
Estimate 2
Estimate 3
Estimate B
Original Sample with Sample Size n

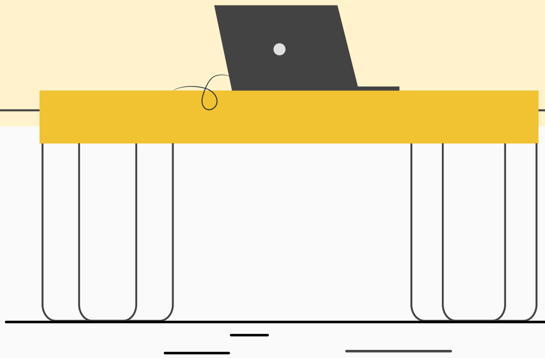
K-Nearest Neighbors

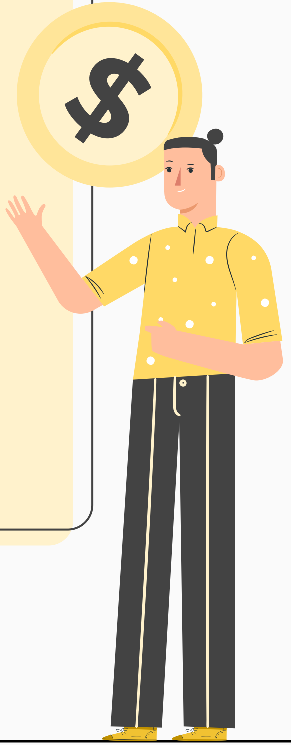
- ***Let's do this together!***

Open up R!

SNACK BREAK!

COME BACK IN 15!





CODE LAB!

OPEN UP RSTUDIO