# Integrating Computational Thinking into Statistics Education

Jo Hardin, Victoria Woodard,
Nicholas J. Horton

# In the Chat

We would like to know the following about you!

- Name
- Affiliation
- Where do you personally see the biggest need to integrate computational thinking into Statistics Education? (Intro Stats, Upper level courses, Minor/Major curriculum development, etc…)

# Professional Guidelines

## GAISE - Guidelines for Assessment and Instruction in Statistics Education College Report (2016)

- It is important to view the use of technology **not just as a way to generate statistical output but as a way to explore conceptual ideas** and enhance student learning.
- Technology tools should also be used to help students visualize concepts and **develop an understanding of abstract ideas** by simulations.

## ASA Curriculum Guidelines for Undergraduate Programs in Statistical Science (2014)

- Students should be able to program in a higher-level language, **to think algorithmically**, to use simulation-based statistical techniques, and to undertake simulation studies.
- This capacity includes the ability to write functions and **use control flow** in a variety of languages.
- The capacity to undertake and interpret simulation studies as a way to **complement analytic understanding** and/or check results will be increasingly useful in the workplace.

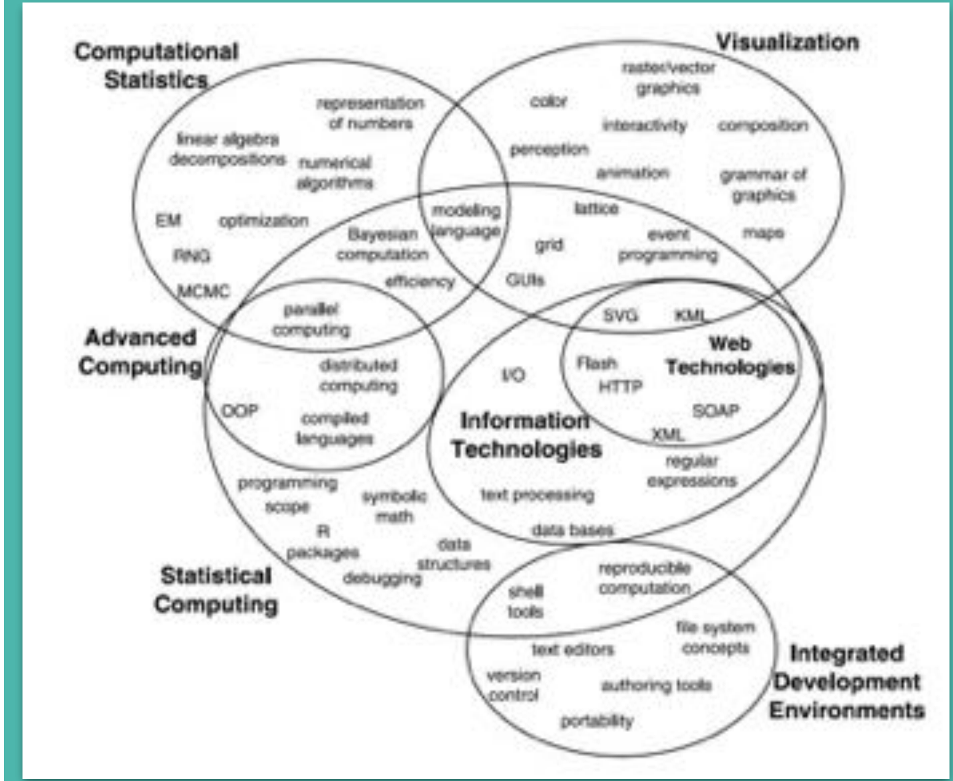
# Computing in the statistics curriculum

Nolan & Temple Lang (The American Statistician, 2010)

- Computational literacy and programming are as **fundamental to statistical practice** and research as mathematics.
- Our field needs to **define statistical computing more broadly** to include advancements in modern computing, beyond traditional numerical algorithms.
- Information technologies are increasingly important and should be added to the curriculum, as should the ability to **reason about computational resources**, work with large datasets, and perform computationally intensive tasks.

# Computing in the statistics curriculum

Nolan & Temple Lang (2010)



# Journal of Statistics and Data Science Education

Special Issue on Integrating computing in the statistics and data science curriculum

January 2021

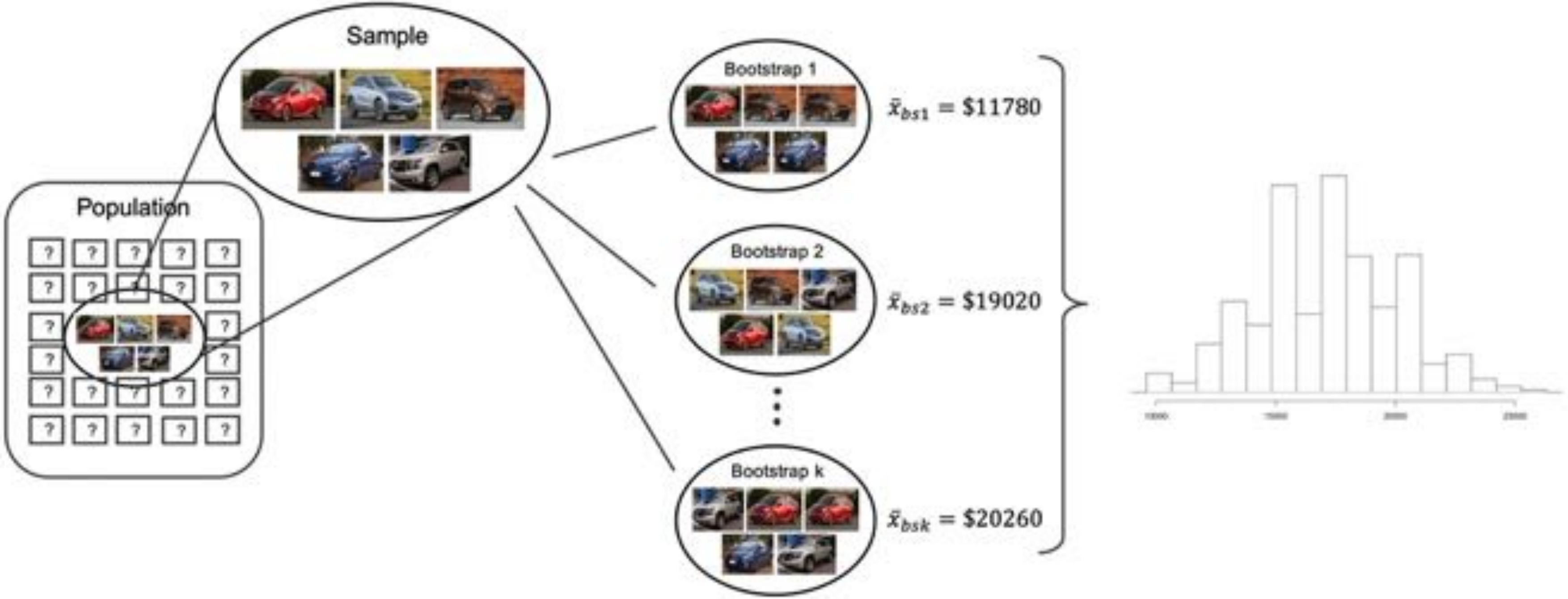
- Creative teaching structures
- Novel and technical data science skills and habits
- Teaching computational thinking

## Computational Foundations

- Basic abstractions,
- Algorithmic thinking,
- Programming concepts,
- Data structures, and
- Simulations.

National Academies of Sciences, Engineering & Medicine (2018) National Academies of Sciences, Engineering and Medicine. *Data science for undergraduates.* Washington, D.C.: The National Academies Press; 2018, https://nas.edu/envisioningds

# Bootstrapping



# Bootstrapping

Too soft:

```
```{r}
library(boot)
set.seed(47)
my.mean = function(x, indices){
  return(mean(x[indices]))}
boot(data = cars93$price, statistic = my.mean, R = 1000) %>%
boot.ci()
```
```

```
bootstrap variances needed for studentized intervalsBOOTSTRAP CONFIDENCE INTERVAL
CALCULATIONS
Based on 1000 bootstrap replicates

CALL : 
boot.ci(boot.out = .)

Intervals : 
Level      Normal              Basic         
95%   (16.97, 23.06 )   (16.75, 22.90 )  

Level     Percentile            BCa          
95%   (17.08, 23.23 )   (17.32, 23.65 )  
Calculations and Intervals on Original Scale
```

# Bootstrapping

## Too hard:

```r
```{r}
set.seed(47)
num.sim <- 1000
boot.means <- numeric(num.sim)
for(i in 1:num.sim){
  boot.means[i] <- mean(sample(cars93$price, replace = TRUE))
}
quantile(boot.means, c(0.025, 0.975))
```
```

```
     2.5%    97.5%
17.27731 23.28593
```

# Bootstrapping

## Just right:

```r
```{r}
library(infer)
set.seed(47)
cars93 %>%
  specify(response = price) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean") %>%
  get_ci(level = 0.95)
```
```

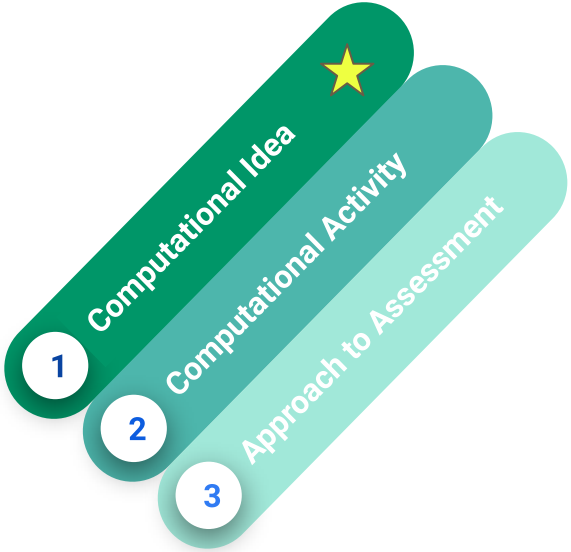
A tibble: 1 x 2

| lower_ci <dbl> | upper_ci <dbl> |
|---|---|
| 17.27731 | 23.28593 |

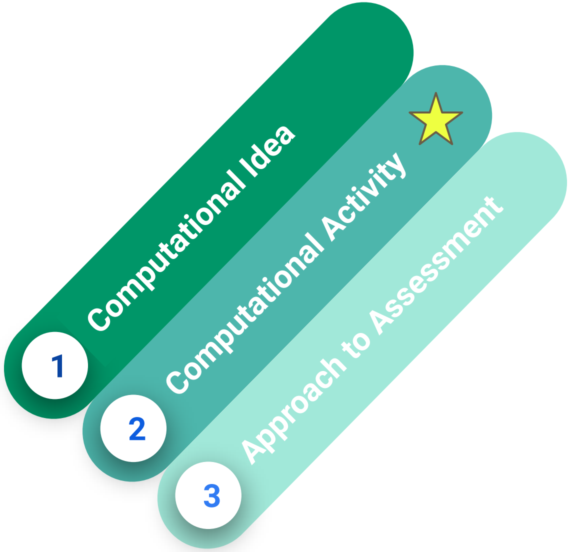
# axes of implementation

https://docs.google.com/presentation/d/1oLKZRYB4SRJAX43mrroUdHyS22MU2Ddh5GcDBjmOKiU/edit?usp=sharing

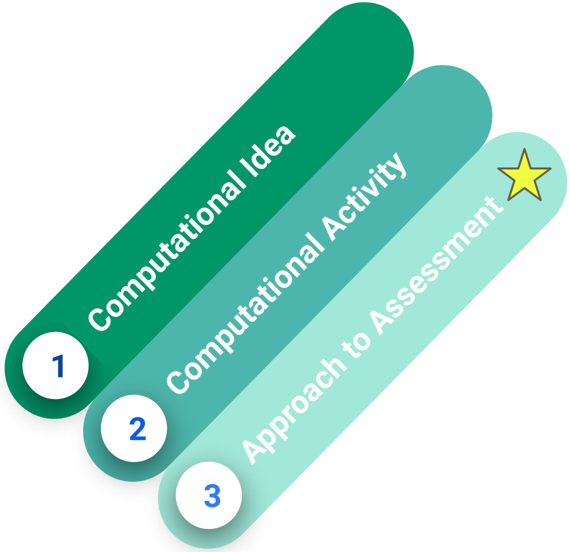
# 3 Examples of Activities for Computational Thinking




Data Viz on Day One




Playing the Whole Game




A Tip for ANOVA

# Data Viz on Day One

```
library(mosaic)
library(openintro)
my_loans <- loans_full_schema %>%
  filter(annual_income_joint > 10)
mplot(my_loans)
```



Data Viz on day one: Bringing big ideas into intro stats early and often (Wang, Rush, and Horton, TISE, 2017, https://escholarship.org/uc/item/84v3774z)

# Data Viz on Day One

## Computational Ideas

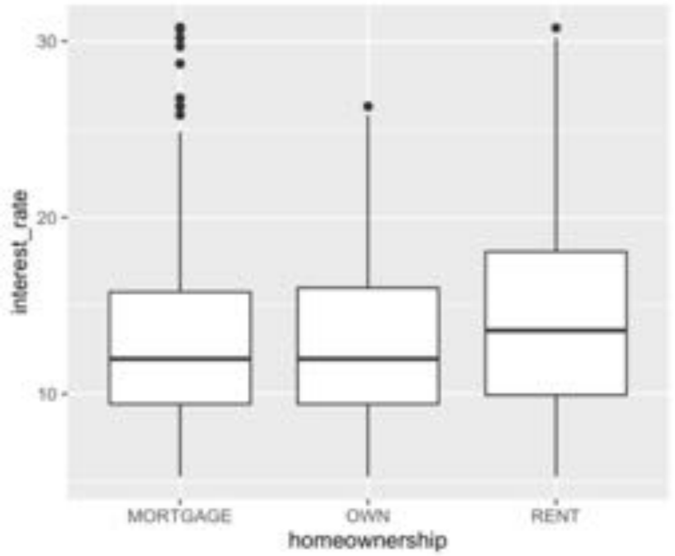
- Create a multivariate visualization
- Introduce a modeling language (formulas and the mosaic package)
- Utilize RMarkdown and reproducible analysis to share results and insights
- All on day one of class

```
> lm(interest_rate ~ homeownership, data = my_loans)

Call:
lm(formula = interest_rate ~ homeownership, data = my_loans)

Coefficients:
      (Intercept)   homeownershipOWN  homeownershipRENT
          12.8223            -0.2077             1.4894
```

```
> gf_boxplot(interest_rate ~ homeownership, data = my_loans)
```




```
> favstats(interest_rate ~ homeownership, data = my_loans)
  homeownership  min   Q1 median    Q3   max  mean    sd   n missing
1      MORTGAGE 5.31 9.43  11.99 15.77 30.79 12.82 5.323 950       0
2           OWN 5.31 9.43  11.99 16.01 26.30 12.61 4.897 183       0
3          RENT 5.31 9.93  13.59 18.06 30.75 14.31 5.555 362       0
```

# Data Viz on Day One


Clicking “Show Expression” yields the following output:

```
gf_jitter(
  interest_rate ~ annual_income_joint,
  data = my_loans,
  color = ~ homeownership
) %>%
  gf_lm() %>%
  gf_refine(scale_x_log10()) %>%
  gf_theme(legend.position = "top") %>%
  gf_labs(title = "", caption = "")
```


- In pairs or trios, students explore the personal loans dataset
- Using “mosaic::mplot()” they interactively generate multivariate plots
- Students copy the code into an RMarkdown file and add a brief interpretation
- Reports are shared via RPubs and discussed at the end of class

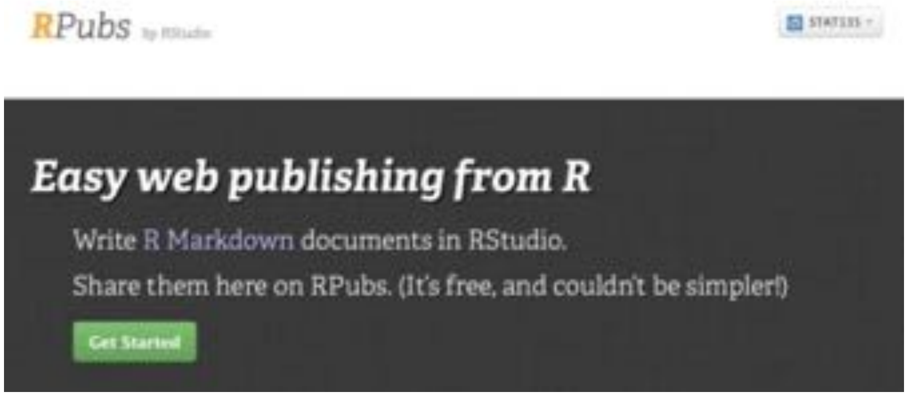
# Data Viz on Day One


"mplot()" also works with ggplot2!

```
ggplot(data = my_loans,
  aes(x = annual_income_joint, y = interest_rate)
) +
  geom_point() +
  aes(colour = homeownership) +
  scale_x_log10() +
  stat_smooth(method = lm) +
  theme(legend.position = "top") +
  labs(title = "")
```

# Data Viz on Day One

## Approach to Assessment




- Early in the semester (typically day one of intro stats)
- Serves to connect students with each other
- Brings big ideas into the course
- Full credit for submitting group report on RPubs
- Leverages RStudio Server Pro setup

Data Viz on day one: Bringing big ideas into intro stats early and often (Wang, Rush, and Horton, TISE, 2017, https://escholarship.org/uc/item/84v3774z)

# Novel and technical data science skills and habits

Kim & Hardin JSDSE 2021,
https://www.tandfonline.com/doi/full/10.1080/10691898.2020.1799728

# Playing the whole game

## Computational Idea

- Data collection
  1. Experience creating measurable data observations (e.g., how is “one day” measured, or what defines “studying”).
  2. Address data collection constraints due to limits in technological capacity and human behavior.
- Data ethics
  1. Practice the ethical and legal responsibilities of those collecting, storing, and analyzing data.
  2. Decide limits for personal privacy.
  3. Deliberate on the trade-offs between research results and privacy.
- “Playing the whole game”
  1. Tie together data collection, analysis, ethics, and communication components.
  2. Iterate between and within the components of the “whole game.”

# Playing the whole game

## Computational Activity

1. Collect data:
    1. Identify a question about how you use your time *that you feel comfortable sharing with your partner and the instructor*. Feel free to make up data.
    2. Start logging time in an electronic calendar app like Google Calendar, macOS Calendar, or Outlook. If you already use a calendar app, we suggest you create a new calendar dedicated to this activity. That way your pre-existing calendar's will be kept private.
2. Exchange data and analyze your partner's data:
    1. Export your calendar to an .ics file.
    2. Exchange your question and .ics data with your partner.
    3. Import your partner's calendar data into R
3. Write a 500 word joint reflection piece on this experience, keeping the Compromised Shoe Situation episode of the "Not So Standard Deviations" podcast in mind. I suggest you write it in Google Docs and then export to PDF. Address in particular:
    1. What difficulties in the data collection & analysis process did you encounter?
    2. As someone who provides data, what expectations do you have when you give your data?
    3. As someone who analyzes others' data, what ethical responsibilities do you have?

# Playing the whole game

## Approach to Assessment

- Class discussion
- Peer review
- Iteration on assignment
- HW assignment

# A Tip for ANOVA

Favorable weather has been shown to be associated with increased tipping. Will just the belief that future weather be favorable lead to higher tips? The researchers gave 60 index cards to a waitress at an Italian restaurant in New Jersey. Before delivering the bill to each customer, the waitress randomly selected a card and wrote on the bill the same message that was printed on the index card.

Twenty of the cards had the message "The weather is supposed to be really good tomorrow. I hope you enjoy the day!" Another 20 cards contained the message, "The weather is supposed to be not so good tomorrow. I hope you enjoy the day anyway!" The remaining 20 cards were blank, indicating that the waitress was not supposed to write any message.

The data for this problem can be found in R. Do the data support the hypothesis that there are differences among the tipping percentages for the three experimental conditions?

# A Tip for ANOVA

Computational Idea: Using data that is not in its “usual form”, can students transform data to be able to conduct ANOVA in R?

Computational Problem: ...Does the data support the hypothesis that there are differences among tipping percentages for the three experimental conditions?

```
good.report <- c(20.8, 18.7, 19.9, 20.6, 22.0, 23.4, 22.8, 24.9, 22.2, 20.3, 24.9, 22.3, 27.0, 20.4, 22.2, 24.0, 21.2, 22.1, 22.0, 22.7)
bad.report <- c(18.0, 19.0, 19.2, 18.8, 18.4, 19.0, 18.5, 16.1, 16.8, 14.0, 17.0, 13.6, 17.5, 19.9, 20.2, 18.8, 18.0, 23.2, 18.2, 19.4)
no.report <- c(19.9, 16.0, 15.0, 20.1, 19.3, 19.2, 18.0, 19.2, 21.2, 18.8, 18.5, 19.3, 19.3, 19.4, 10.8, 19.1, 19.7, 19.8, 21.3, 20.6)
```

# Identifying Computational Thinking

- ***Build*** one of many possible solutions to a problem to answer a statistical question (does not follow a prescribed set of procedures).
- ***Provide*** their own reasonable methods for analysis while utilizing the technology.
- ***Identify*** and understand information gathered that will be useful in solving the current problem.
- ***Recognize*** patterns in the output of an analysis and use that information to determine the next step of the process.
- ***Recognize*** patterns in the way things are coded and use that pattern to help solve the problem.
- ***Create*** a solution strategy and ***communicate*** that strategy to the computer software.
- ***Demonstrate*** critical or abstract ***thinking*** about a concept that the technology presents.
- ***Revise*** code that has been provided to them, to perform a new task.
- ***Use*** the technology to efficiently create graphs or summaries to make sense of the data, without overloading cognitive processes.
- ***Use*** the technology to perform statistical calculations and use the results to make appropriate statistical decisions without overloading cognitive processes.

Victoria Woodard & Hollylynne Lee (2021) How Students Use Statistical Computing in Problem Solving, Journal of Statistics and Data Science Education, 29:sup1, S145-S156, DOI: 10.1080/10691898.2020.1847007

# Identifying Computational Thinking

- ***Build*** one of many possible solutions to a problem to answer a statistical question (does not follow a prescribed set of procedures).
- ***Provide*** their own reasonable methods for analysis while utilizing the technology.
- ***Identify*** and understand information gathered that will be useful in solving the current problem.
- ***Recognize*** patterns in the output of an analysis and use that information to determine the next step of the process.
- ***Recognize*** patterns in the way things are coded and use that pattern to help solve the problem.
- ***Create*** a solution strategy and ***communicate*** that strategy to the computer software.
- ***Demonstrate*** critical or abstract ***thinking*** about a concept that the technology presents.
- ***Revise*** code that has been provided to them, to perform a new task.
- ***Use*** the technology to efficiently create graphs or summaries to make sense of the data, without overloading cognitive processes.
- ***Use*** the technology to perform statistical calculations and use the results to make appropriate statistical decisions without overloading cognitive processes.

Action!

Victoria Woodard & Hollylynne Lee (2021) How Students Use Statistical Computing in Problem Solving, Journal of Statistics and Data Science Education, 29:sup1, S145-S156, DOI: 10.1080/10691898.2020.1847007

# A Tip for ANOVA

Approach to Assessment: To truly identify if students are *thinking* computationally, we probably need to see them in action.

# A Tip for ANOVA

Approach to Assessment: To truly identify if students are *thinking* computationally, we probably need to see them in action.

## Assessment Options

| Assignment Type | Easy to Implement and Grade | Captures Individual Understanding | Identify Students' Computational Thinking |
|---|---|---|---|
| Traditional Homework | | | |
| Group Discussions | | | |
| Task Based Interview | | | |

# A Tip for ANOVA

Assessment of Computational Thinking: To truly identify if students are *thinking* computationally, we probably need to see them in action.

## Assessment Options

| Assignment Type | Easy to Implement and Grade | Captures Individual Understanding | Identify Students' Computational Thinking |
|---|---|---|---|
| Traditional Homework | ✓ | ✓ | ✗ |
| Group Discussions | ✓ | ✗ | ? |
| Task Based Interview | ✗ | ✓ | ✓ |

# A Tip for ANOVA

Assessment of Computational Thinking: To truly identify if students are *thinking* computationally, we probably need to see them in action.

Discussion Question: How do you think you can adequately assess an individual student's ability to think computationally, while balancing ease of implementation and grading?

# reflections on implementing computational thinking

We hope participants leave with:

- The sense that the computer provides a mechanism for students to experience the entire statistical analysis cycle and to deepen their understanding of fundamental statistical concepts such as variability, inference, and design.
- At least one idea for how to scaffold more computational thinking into the statistics classroom.
- At least one idea for how to assess computational thinking

https://docs.google.com/document/d/1TkwnMoQ5MNMYzh8I0WOslD-Ocd-vuuPTVTyePAQIz0w/edit?usp=sharing

# Thank you