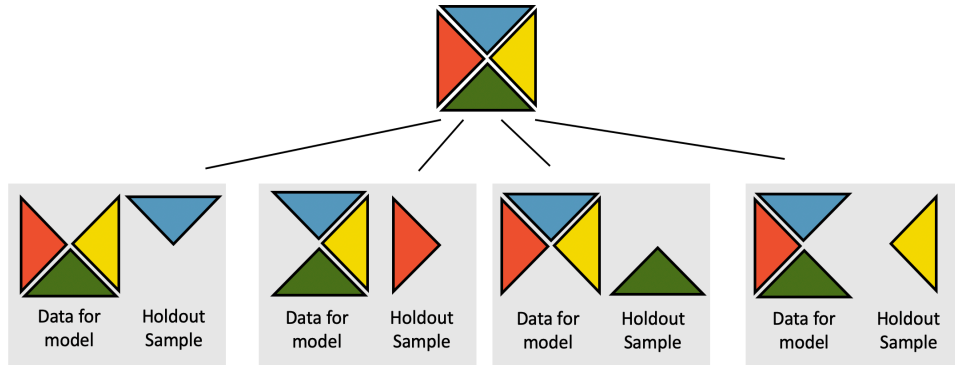


# the value of computational thinking in statistics education

Jo Hardin  
Pomona College

---



@jo\_hardin47

Github: hardin47

[jo.hardin@pomona.edu](mailto:jo.hardin@pomona.edu)

# Professional Guidelines

---

- GAISE - Guidelines for Assessment and Instruction in Statistics Education (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)
  - They 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 Curricula

Nolan & Temple Lang (2010)

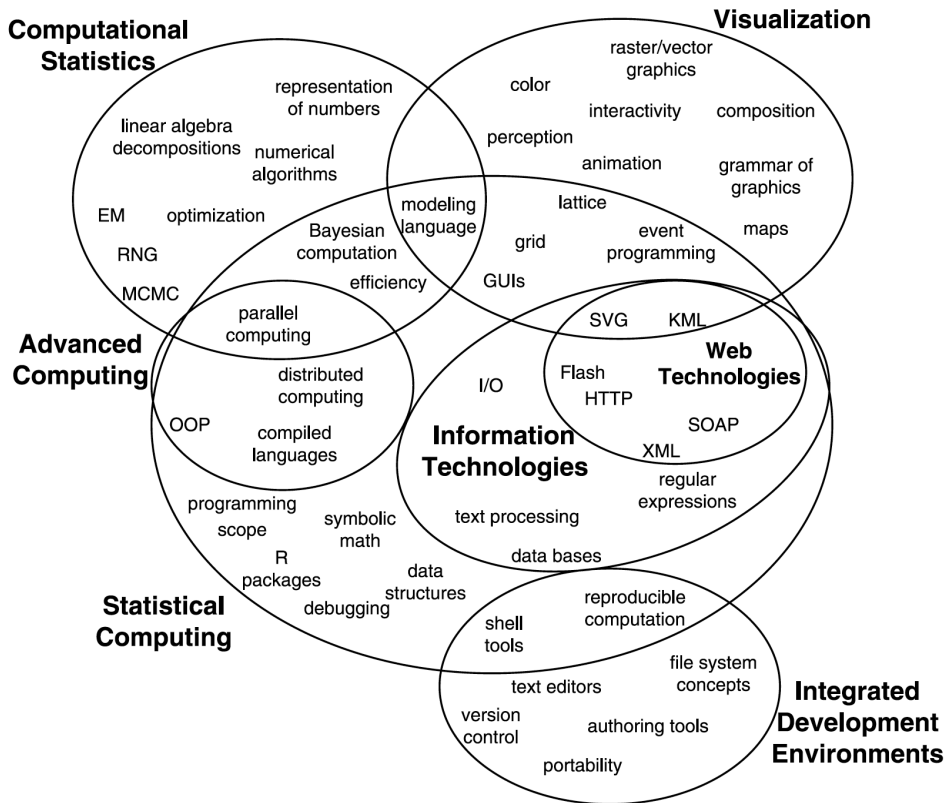
---

- 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 Curricula

Nolan & Temple Lang (2010)

---





What now / next?

---

- Special issue of Journal of Statistics Education on:

Computing in the Statistics and Data Science Curriculum  
(call in early 2019)

- Name of ASA Section (2019):  
Statistics Education ->  
Statistics and Data Science  
Education
- Name of JSE (2021):  
JSE ->  
Journal of Statistics and Data  
Science Education



## Journal of Statistics and Data Science Education

---

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

January 2021

editors: Jo Hardin & Nick Horton

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

# Creative teaching structures



## Easy-to-Use Cloud Computing for Teaching Data Science

Kim & Henke

	Tool	Function	Details
Step 1	Jupyter Notebooks	Document	Build teaching material.
Step 2	GitHub	Online Repository	Store notebooks online.
Step 3	Binder	Cloud Service	Deliver in the cloud.

Table 1: A summary of the tools and their uses for creating and delivering executable Jupyter notebooks.



## Teaching Statistical Concepts and Modern Data Analysis with a Computing-Integrated Learning Environment (ISLE)

Burckhardt, Nugent, & Genovese

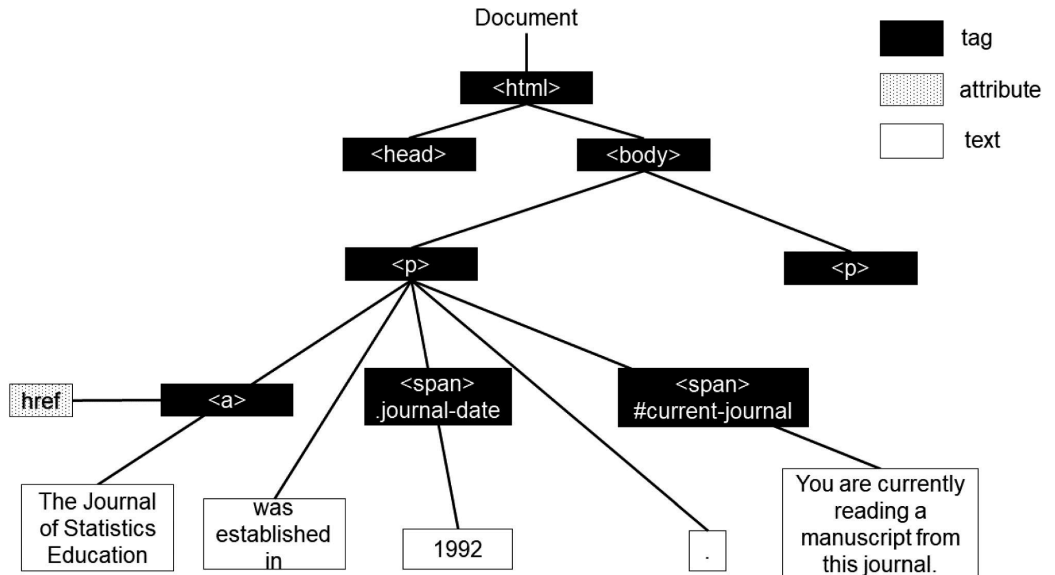
The screenshot displays a Jupyter Notebook interface. On the left, a code editor shows R code for an interactive lesson. The code includes a title, author, date, state, license, and a comment indicating it's an interactive lesson. It defines an RShell environment and sets up a mean function. A context menu is open over the code, showing options like 'Main', 'Display', 'Questions', 'Survey', 'Programmatic Components', and 'Learning Components'. On the right, the rendered notebook shows a yellow background with the text 'This is an interactive lesson.' and 'RShell'. Below this, there's a text input field containing the R code '1 mean( c(10, 5, 8, 2, 13) )' and a 'Submit' button. At the bottom right, there's a 'LaTeX' section with the text 'LaTeX equations:' and a partial mathematical expression  $f(x)dx$ .

# Novel and technical data science skills and habits



## Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities

Dogucu & Çentinkaya-Rundel

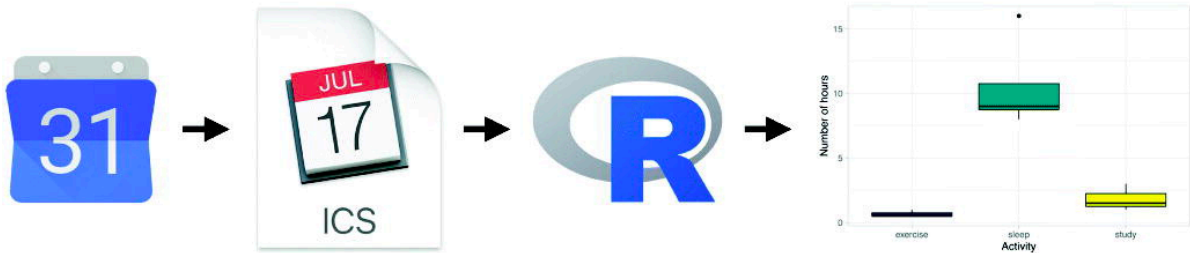


# Novel and technical data science skills and habits



Kim & Hardin

## "Playing the whole game": A data collection & analysis exercise with Google Calendar



1. Log activities in Google Calendar
2. Export to .ics file format
3. Import to R using `ical` package
4. Analyze

**Iterate as needed!**

# Novel and technical data science skills and habits



What is happening on Twitter?

A framework for student research projects with tweets

Boehm & Hanlon



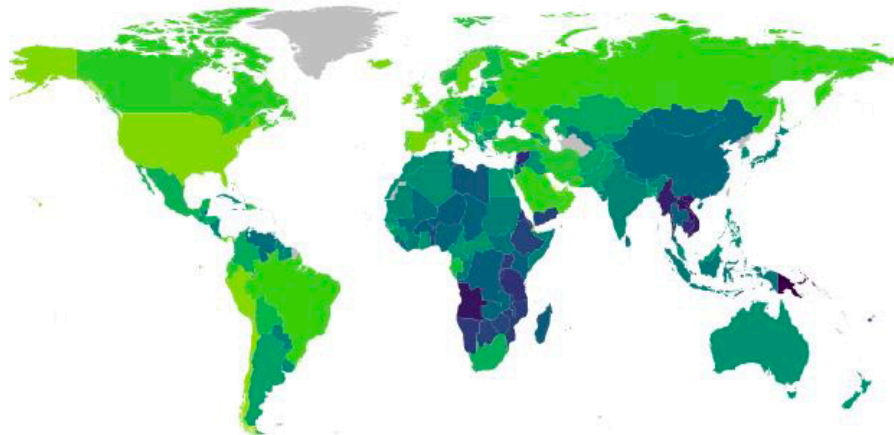
# Novel and technical data science skills and habits



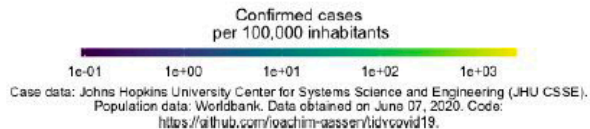
## Computational Skills for Multivariable Thinking in Introductory Statistics

Adams, Baller, Jonas, Joseph, & Cummiskey

Covid19: Confirmed cases (cumulative) as of June 06, 2020



“Proficiency in a statistical programming language facilitates the development of multivariable thinking by giving students tools to investigate complex data on their own.”



## Changing the way we think:

---



Computer science's contribution to biology goes beyond the ability to search through vast amounts of sequence data looking for patterns. The hope is that data structures and algorithms—our computational abstractions and methods—can represent the structure of proteins in ways that elucidate their function. Computational biology is changing the way biologists think.

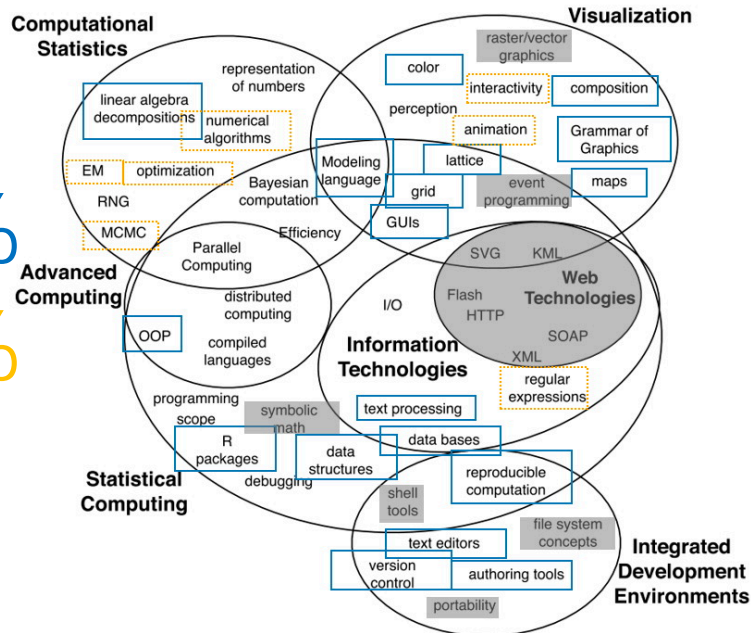


# Teaching computational thinking

covered > 75%

covered > 50%

not asked



The nature of doing computation in the classroom requires students to be familiar with concepts like debugging, code formatting, and reproducible programming. However, **are we truly developing students who understand** how R, Python, or any of the other computing languages used to teach data science “think”?



Data Science in 2020:  
Computing, Curricula, and  
Challenges for the Next 10  
Years

Schwab-McCoy, Baker, &  
Gasper

# Teaching computational thinking

---



## Teaching Creative and Practical Data Science at Scale

Donoghue, Voytek, & Ellis

Key skills for the budding data scientist include how to explore and debug both code and data issues, and how to decide on a path forward when what to do next is unclear. ... We seek to explicitly instruct students on the data-centric debugging strategies employed when analyzing data by running sessions on debugging and how to proceed if one's code is not working.



## Designing Data Science Workshops for Data-Intensive Environmental Science Research

Theobald, Hancock, Mannheim

The skills necessary for students to engage in [the data analysis] cycle may include general programming concepts such as looping, user-defined functions, or conditional statements.

# Teaching computational thinking in practice

---

- What does it all mean?
- How do we do this in the classroom?
- What are best practices?

# Tidy Data

country	year	cases	population
Afghanistan	1999	765	19987071
Afghanistan	2000	2666	20095360
Brazil	1999	31737	172006362
Brazil	2000	80488	174004898
China	1999	212258	1272015272
China	2000	216766	128042583

variables

country	year	cases	population
Afghanistan	1999	765	19987071
Afghanistan	2000	2666	20095360
Brazil	1999	31737	172006362
Brazil	2000	80488	174004898
China	1999	212258	1272015272
China	2000	216766	128042583

observations

country	year	cases	population
Afghanistan	1999	765	19987071
Afghanistan	2000	2666	20095360
Brazil	1999	31737	172006362
Brazil	2000	80488	174004898
China	1999	212258	1272015272
China	2000	216766	128042583

values

<https://towardsdatascience.com/what-is-tidy-data-d58bb9ad2458>

the tidyverse

The tidyverse is an opinionated collection of R packages... [sharing] an underlying design philosophy, grammar, and data structures.

<https://www.tidyverse.org/>

# Tidy Processing

$$f(x) = \log_{10}(\sqrt{x})$$

rounded, to 2 digits

```
> x <- c(1:10)
> x
[1] 1 2 3 4 5 6 7 8 9 10
>
> round(log(sqrt(x), base = 10), digits = 2)
[1] 0.00 0.15 0.24 0.30 0.35 0.39 0.42 0.45 0.48 0.50
>
> x %>% sqrt() %>% log(base = 10) %>% round(digits = 2)
[1] 0.00 0.15 0.24 0.30 0.35 0.39 0.42 0.45 0.48 0.50
```

not tidy



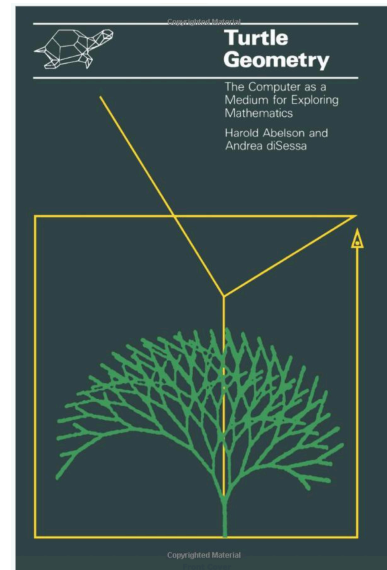
tidy



# Computing & Math

---

- **Turtle Geometry** is a college-level math text ... exploring mathematical properties visually via a simple programming language.



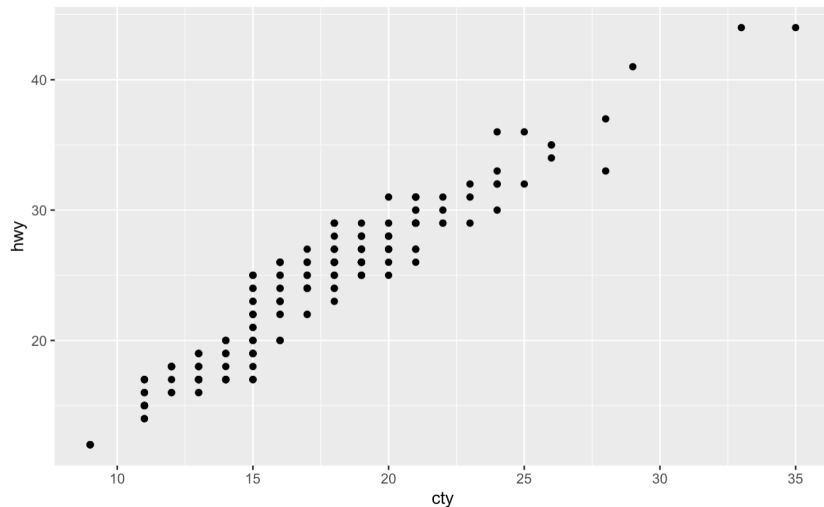
Programs must be written for people to read,  
and only incidentally for machines to execute.

Abelson & Sussman, "Structure and Interpretation of Computer Programs", preface to  
the first edition

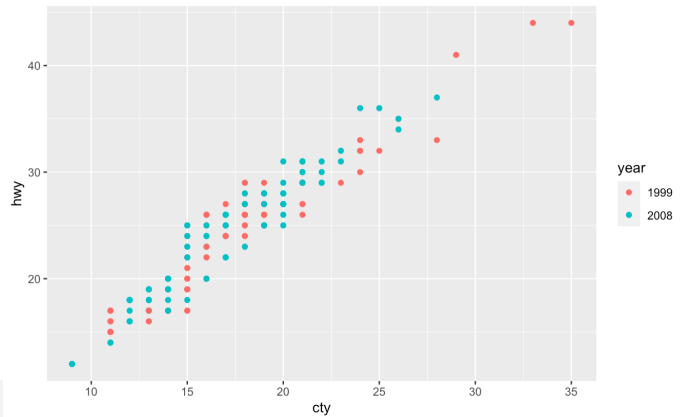
# Tidy Plots

```
60 ggplot(data = mpg)
```

```
63 ggplot(data = mpg) +  
64   geom_point(aes(x = cty, y = hwy))
```

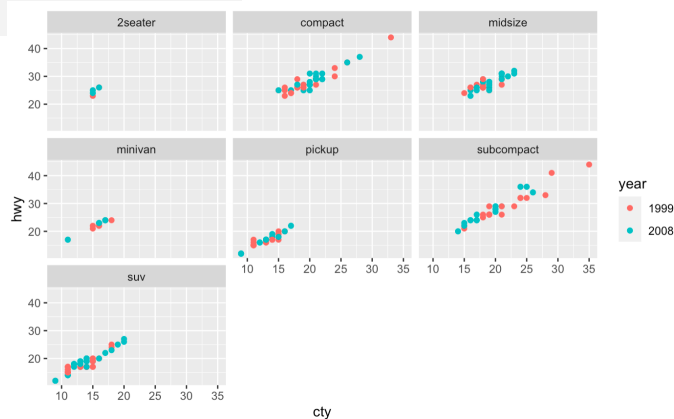


# Tidy Plots



```
67 ggplot(data = mpg) +  
68   geom_point(aes(x = cty, y = hwy, color = year))
```

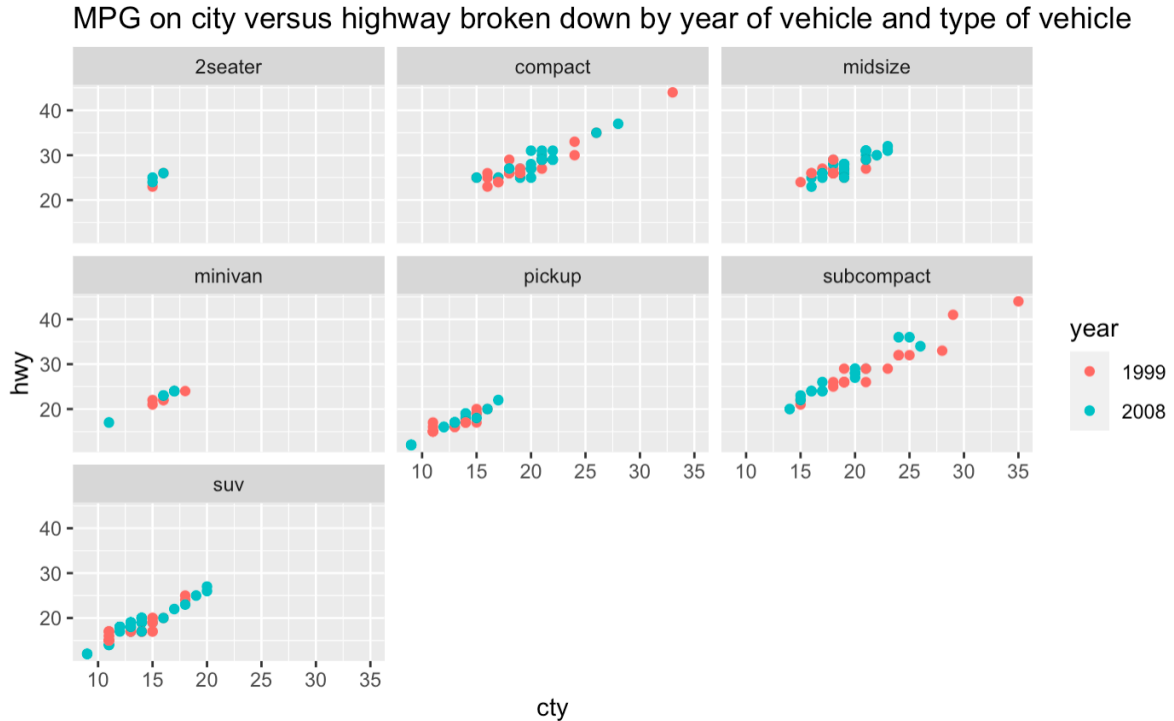
```
70 ggplot(data = mpg) + |  
71   geom_point(aes(x = cty, y = hwy, color = year)) +  
72   facet_wrap(~class)
```



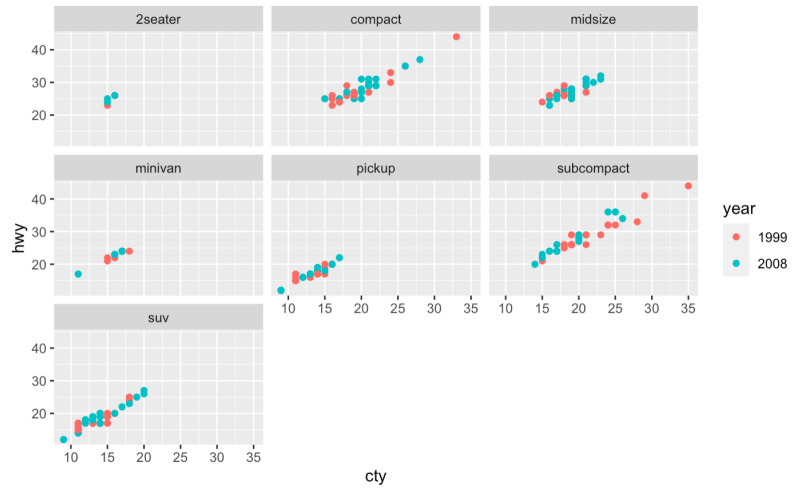
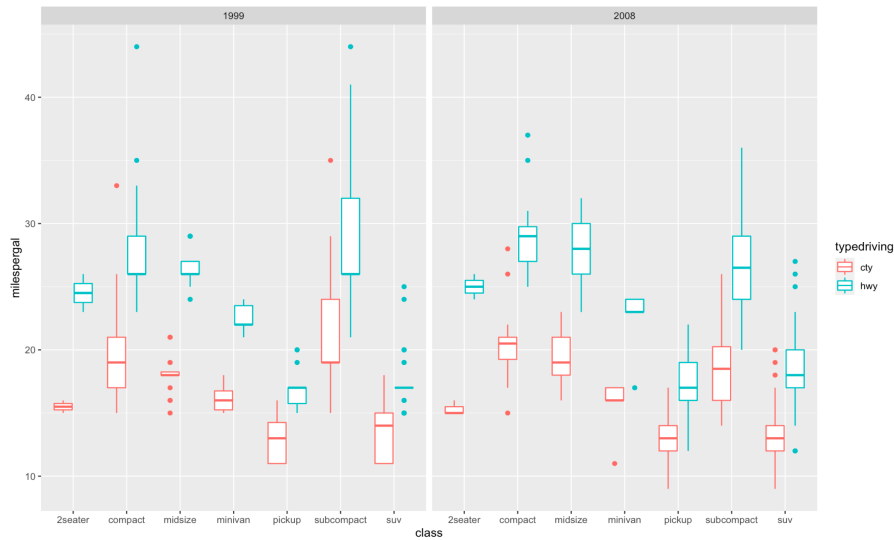


# Tidy Plots

```
74 ggplot(data = mpg) +  
75   geom_point(aes(x = cty, y = hwy, color = year)) +  
76   facet_wrap(~class) +  
77   ggtitle("MPG on city versus highway broken down by year of vehicle and type of vehicle")
```



Currently:  
we want a viz,  
the computer creates it



Instead:  
the computer creates,  
the viz elucidates  
→ what model?

# Randomization Tests

Our Experiment

$$\bar{x}_A - \bar{x}_B = 1.45$$

Data

69 <sup>A</sup> 87 <sup>B</sup>

60 <sup>A</sup> 71 <sup>B</sup>

99 <sup>A</sup> 77 <sup>B</sup>

81 <sup>A</sup> 79 <sup>B</sup>

↓ 65 <sup>B</sup>

$$\bar{x}_A = 77.25 \quad \bar{x}_B = 75.8$$

Simulation #1

$$\bar{x}_{A,stim1} - \bar{x}_{B,stim1} = -3.95$$

Shuffle

<sup>B</sup> <sup>B</sup>  
<sup>A</sup> <sup>B</sup>  
<sup>B</sup> <sup>A</sup>  
<sup>A</sup> <sup>A</sup>  
<sup>B</sup>

Reassign

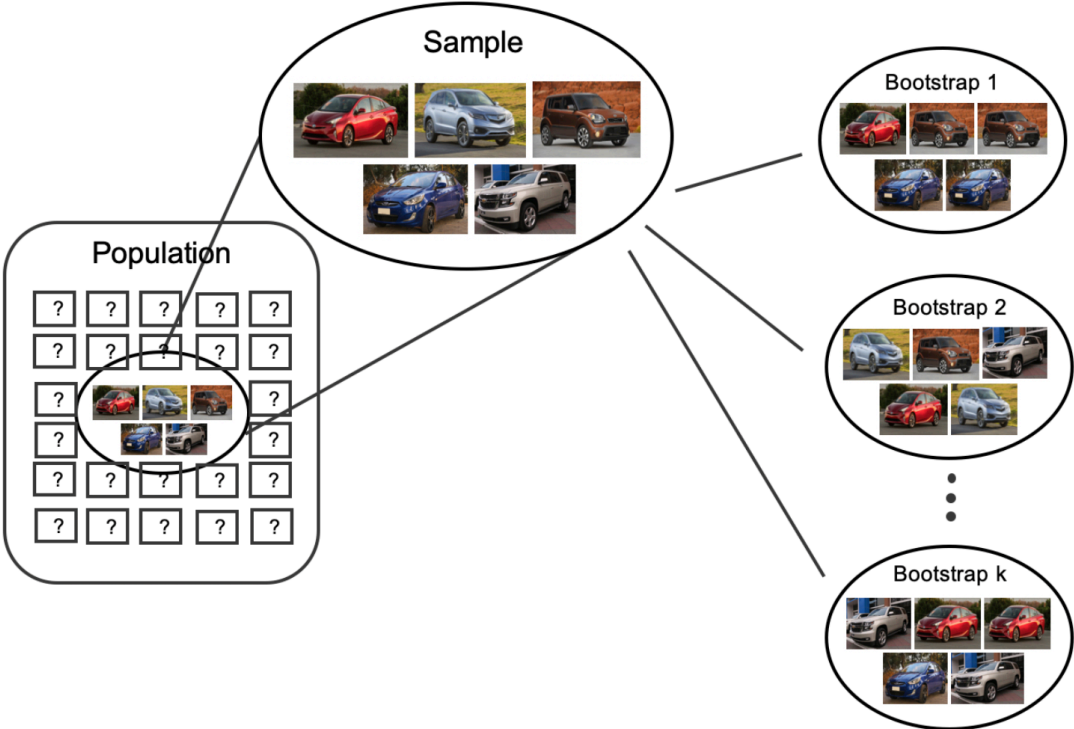
69 <sup>B</sup> 87 <sup>B</sup>  
60 <sup>A</sup> 71 <sup>B</sup>  
99 <sup>B</sup> 77 <sup>A</sup>  
81 <sup>A</sup> 79 <sup>A</sup>  
65 <sup>B</sup>

Sort

60 <sup>A</sup> 69 <sup>B</sup>  
77 <sup>A</sup> 87 <sup>B</sup>  
81 <sup>A</sup> 71 <sup>B</sup>  
79 <sup>A</sup> 99 <sup>B</sup>  
↓ 65 <sup>B</sup>  
 $\bar{x}_{A,stim1} = 74.25$   $\bar{x}_{B,stim1} = 78.2$

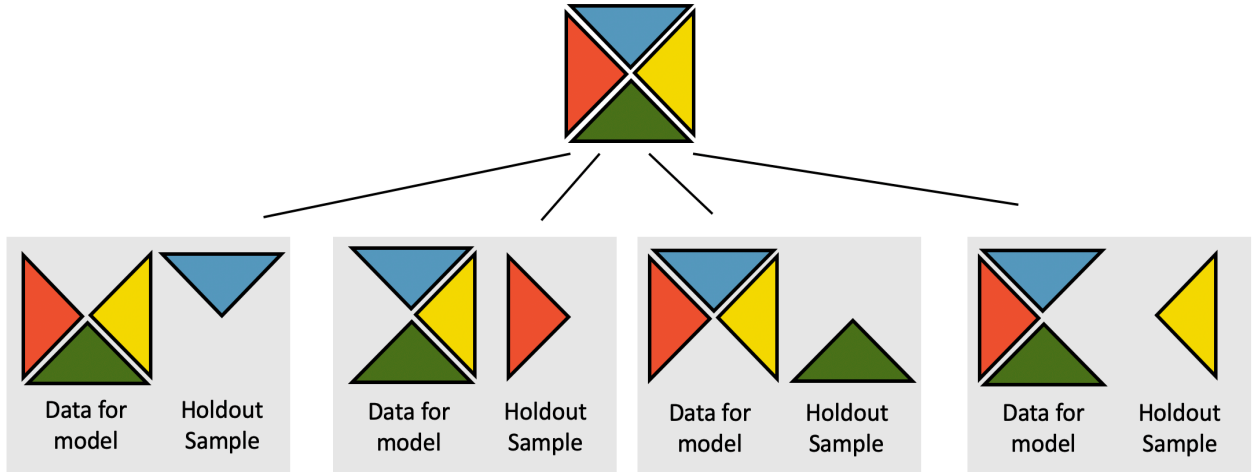
# Bootstrapping

---



# Cross Validation

---





### Data

The data we are using comes from the `ggplot2` package on fuel economy for 38 popular models of cars.

```
[r]
head(mpg)

#> A tibble: 6 x 11
#>   manufacturer m... di... y...   trans   ...   ...   fl
#>   <chr>        <chr> <dbl> <ctr> <chr>   <ct> <cin> <chr>
#> 1 audi         a4     1.8 1... 4 auto(5) f ...   p
#> 2 audi         a4     1.8 1... 4 manual(...) f ...   p
#> 3 audi         a4     2.0 2... 4 manual(...) f ...   p
#> 4 audi         a4     2.0 2... 4 auto(av) f ...   p
#> 5 audi         a4     2.8 1... 6 auto(5) f ...   p
#> 6 audi         a4     2.8 1... 6 manual(...) f ...   p

6 rows | 1-10 of 11 columns
```

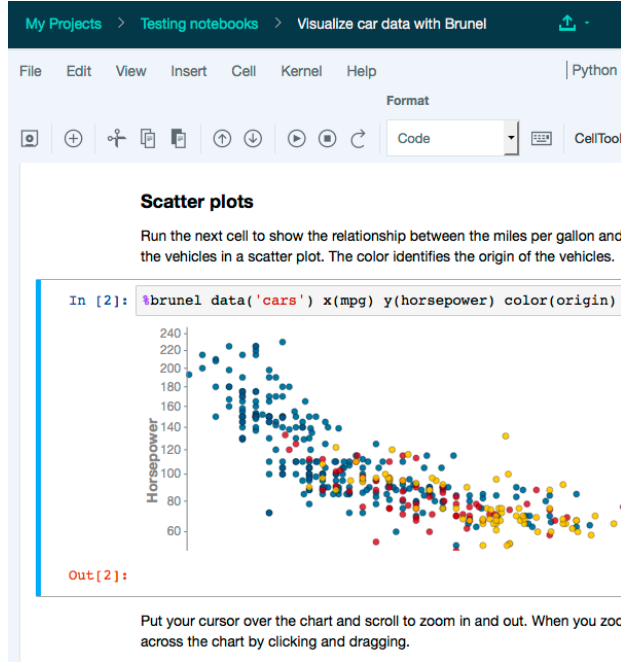
### Viz

Importantly, we'd like to visualize the relationship between some of the cars.

```
[r]
ggplot(data = mpg) +
  geom_point(aes(x = cty, y = hwy, color = year)) +
  facet_wrap(~class) +
  ggtitle("MPG on city versus highway broken down by year of vehicle and type of vehicle")

#> A ggplot object:
#>   • Aes:   cty, hwy, year
#>   • Data:   [data frame]
#>   • Layers:  geom_point
#>   • Facets:  ~class
```

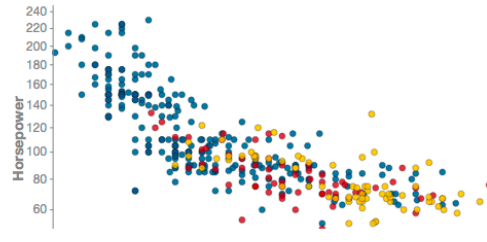
# R Markdown



### Scatter plots

Run the next cell to show the relationship between the miles per gallon and the vehicles in a scatter plot. The color identifies the origin of the vehicles.

```
In [2]: !brunel data('cars') x(mpg) y(horsepower) color(origin)
```



Out[2]:

Put your cursor over the chart and scroll to zoom in and out. When you zoom across the chart by clicking and dragging.

# Jupyter Notebook

technology & communication are *intimately* related



Image credit: Pomona College

Thank you

---

the value of computational thinking  
in statistics education  
Jo Hardin

[jo.hardin@pomona.edu](mailto:jo.hardin@pomona.edu)



[@jo\\_hardin47](https://twitter.com/jo_hardin47)



<https://github.com/hardin47>



<http://research.pomona.edu/johardin/>