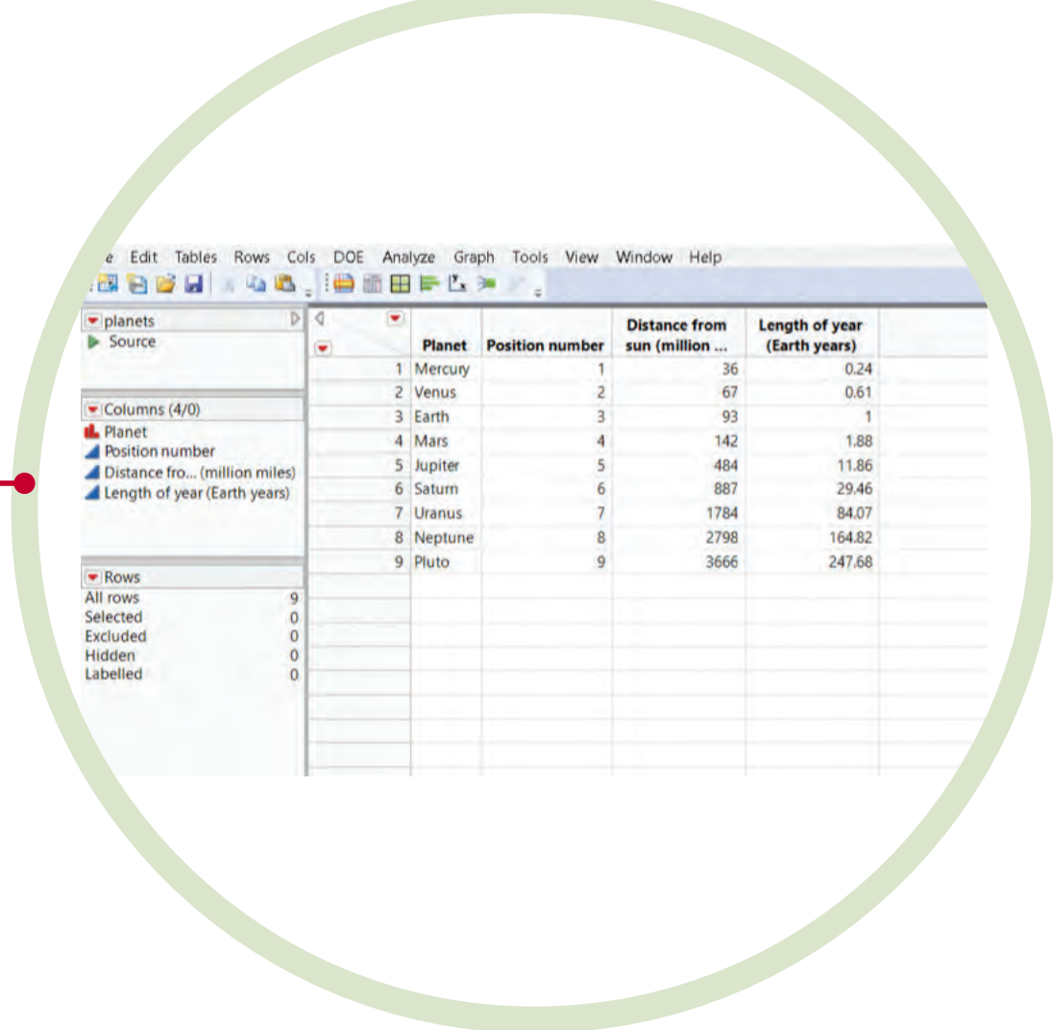
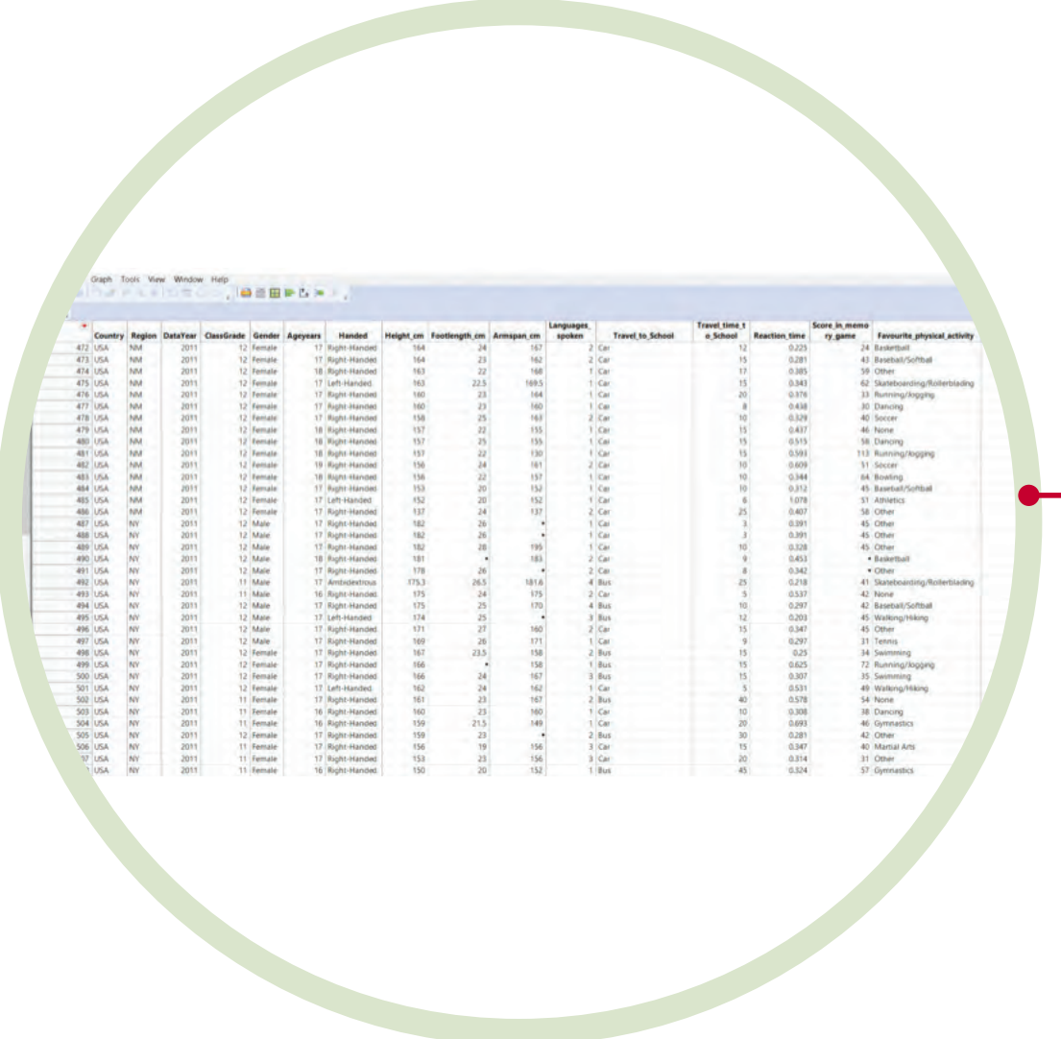
THE EVOLUTION OF AP® STATISTICS: **HOW BIG DATA AND MACHINE LEARNING ARE CHANGING THE COURSE**

2008–2019 TIMELINE



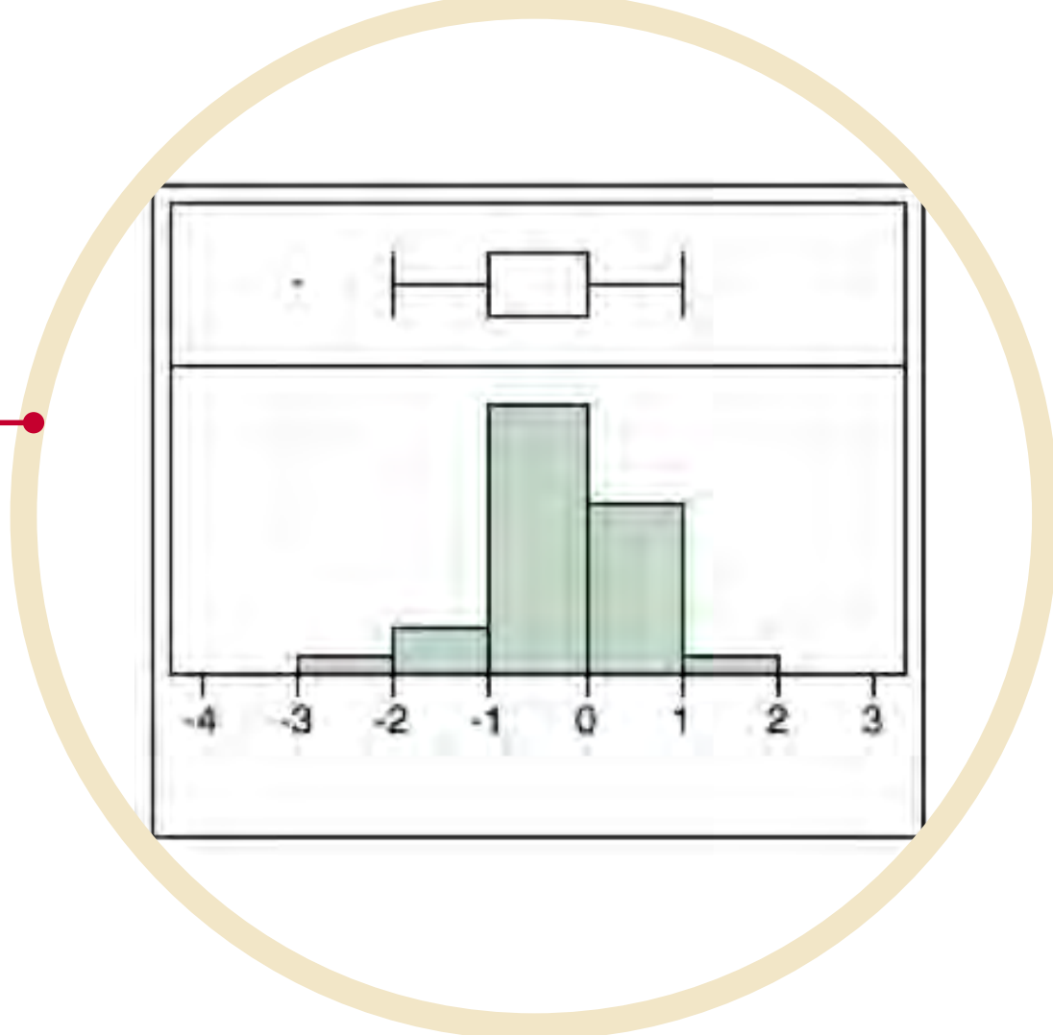
2008

Typical Data File
(9 observations, 4 variables)



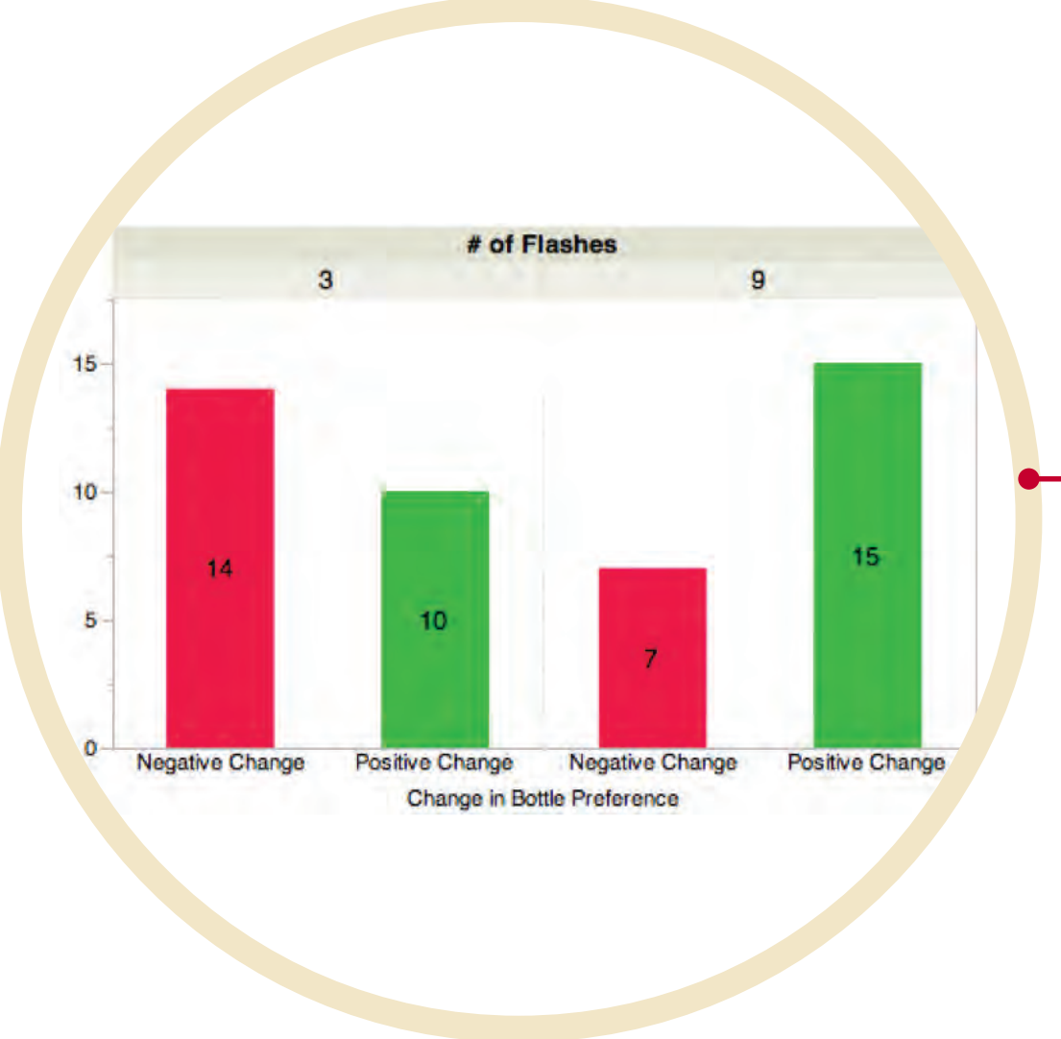
2012

Subset of Census at School Data
(838 observations , 60 variables)



2012

Difference between a Student's Stress Level Normally and when she is Sleep Deprived

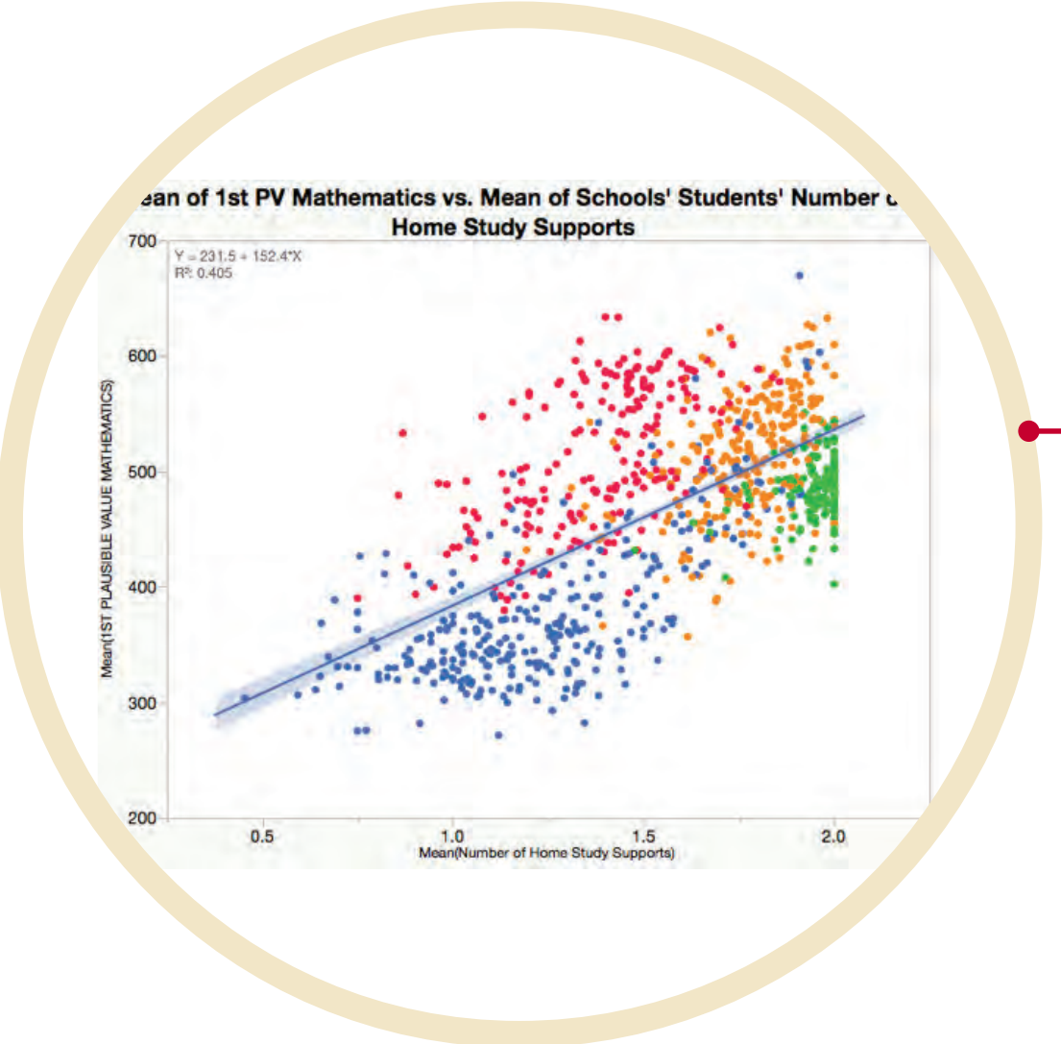


2015

Change in Opinion of Water Bottles by Treatment

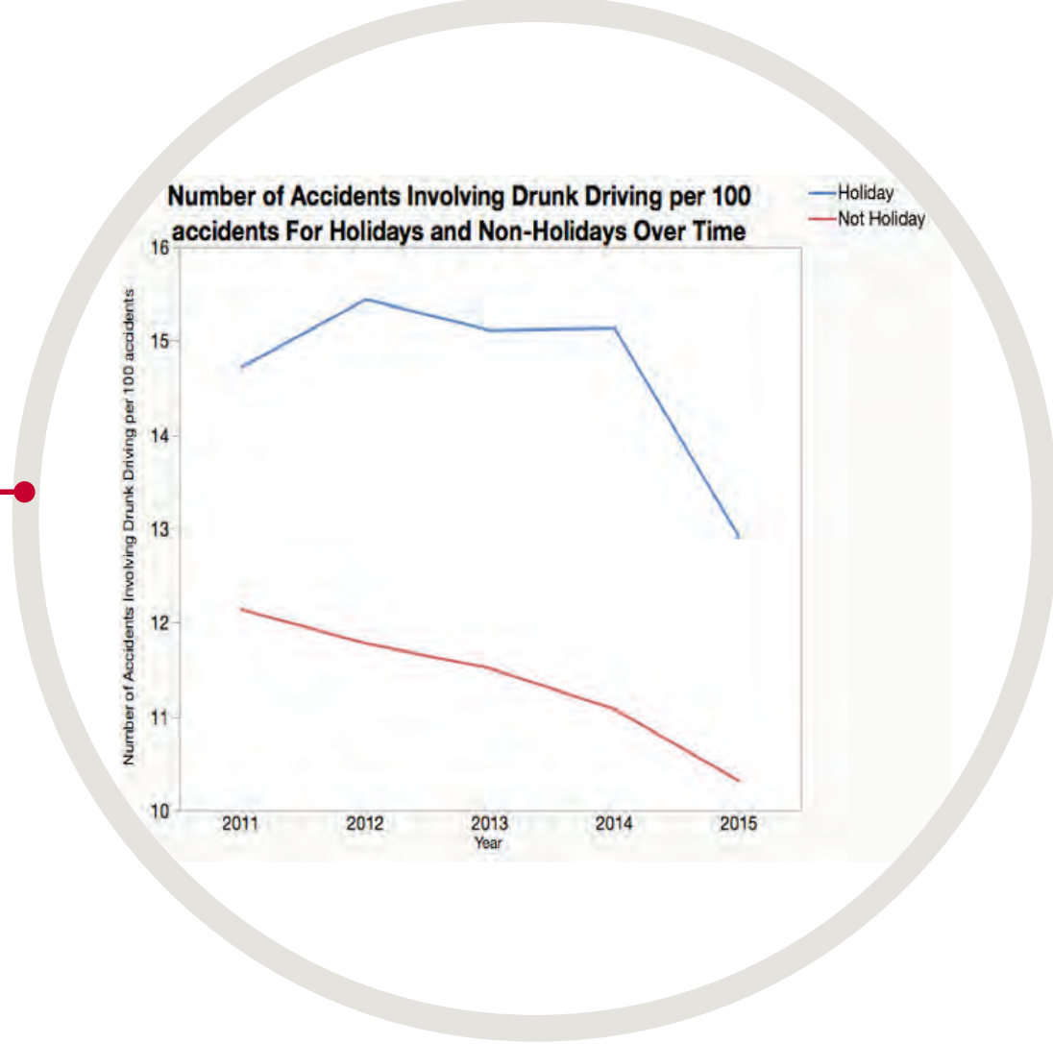
2017

Snapshot of Fatality Analysis Data
(378,000 observations, 24 variables)



2018

Mean of 1st PV Mathematics vs Mean of Schools' Students' number of Home Study Supports



2018

Fatality Analysis: Number of Accidents Involving Drunk Driving per 100 accidents For Holidays and Non-Holidays Over Time

CATALYSTS FOR CHANGE | CURRICULUM CONTENT | DATA & TECHNOLOGY | PROJECT

Workshops (Darren Starnes and Floyd Bullard) Emphasized the important role of simulations to anchor key statistical concepts	2008	
	2009	Introduced simulations into curriculum across curricular elements
Industry Initiatives for Math and Science Education (Summer Internship) Developed unit to introduce students to JMP – Penn State dataset 227 observations, 8 variables	2010	Introduced Year-long statistics project – Literature review and analysis activity in the fall; data gathering and data analysis in the spring
	2011	Introduced students to JMP for statistical analysis; Added this skill without explicitly teaching it; Ability varied widely from student to student; Good idea, poor execution
Sabbatical Research on Statistical Literacy big data, importance of statistical literacy, introduced JMP more formally to students	2012	Conducted a review of statistical software; R was widely used but had too steep a learning curve; Excel did not support the range of statistical analysis activities; JMP chosen as the best tool based on functionality, ease of use, and cost
Statistics Integration Grant Outgrowth of sabbatical research, impact on student preparation for AP stats; increased knowledge of descriptive statistics in grades 6-10	2013	Revised teaching of Statistical inference; used a holistic approach and encouraged students to look at the similarities across both tests for proportions and tests for means; Student understanding of the tests improved by focusing on common elements and difference between tests
Introduced Census at School data to teach basic data analysis skills (838 observations, 60 variables) Most projects were student surveys or experiments; a couple of student did analysis projects with outside organizations	2014	Less time on descriptive statistics – students were better prepared as a result of greater statistical literacy
Increasing availability of public data; Students (with support) can begin to combine data from different sources	2015	Provided additional scaffolding for project; reviewed drafts of most major elements before final submission; Not sustainable with larger student population
Research module on machine learning Realized the central role of linear and logistic regression in machine learning	2017	Introduced FARS (Fatality Analysis and Reporting System) data for student analysis; Data on 5 years of traffic fatalities in the US; all students required to work with this dataset (378,000 observations, 24 variables); Students to determine own research question and use JMP to analyze; insufficient teacher support for JMP
Elective course on Artificial Intelligence Many classifiers require knowledge of statistical techniques All projects were data analysis projects; the datasets the students started with had 1000s to over 1,000,000 observations and 20 to several hundred variables; students learned to create meaningful subsets for analysis and focus on the variables that mattered	2019	Added data analysis unit that focused on teaching students how to use JMP for data analysis; unit culminates with analysis of FARS data

FUTURE DIRECTIONS

CURRICULUM CONTENT

- Big data and machine learning will influence future course content
- Supervised Learning classifiers are key – linear regression (multivariate) is common for prediction problems and logistic regression is common for classification problems; students should understand the basics of both of these approaches
- Computers do the mechanics of all statistics tests; Should the future direction focus more on the value humans can add to the tests:
  - » Understand your data
  - » Constructing hypotheses
  - » Checking conditions
  - » Interpreting results
- Inference with large datasets (>10,000 observations) requires more critical analysis of results; introduce effect size calculation?

TECHNOLOGY AND DATA

- Continue work with large datasets (50,000+ observations, many variables)
- Focus on understanding the data
- Emphasize working with data subsets and validating results on different subsets
- Work with data in the cloud?

PROJECTS

- Projects promote authentic learning and provide critical analysis skills
- Most future projects will be data analysis projects
- Having a project stakeholder who cares about results is valuable; need a sustainable model for doing this
- Students still want to take data at face value without considering how the data was produced or generated; how do we change this attitude?
- Projects are time intensive to review; need to consider models that are scalable

Kyle Barriger

Castilleja