# So where are we?

- Explosion of Stat & Data Science programs, courses, materials
- The People's Science
- We have no idea what the people are doing
- Or why they're doing it

- Human behavior is driving force in data analysis pipeline; "Many analysts, one data set" (Silberzahn, et al; April 2017)
- Adaptive instruction: how can we adapt if we don't know what we should be adapting for?
- Assessments largely focused on final product (reasonably so); what happened along the process?

Behavioral Data Science

# Carnegie Mellon University

- Private university in Pittsburgh, PA
- R1 research university designation
- $\approx$ 7000 undergrads, 7000 grads
- Six undergraduate colleges (admission is college-specific) College of Fine Arts, Dietrich College of Humanities & Social Sciences, College of Engineering, Mellon College of Science, School of Computer Science, Tepper School of Business
- Economics (joint in Tepper), English, History, Information Systems, Institute for Politics and Strategy, Modern Languages, Philosophy, Psychology, Social and Decision Science, Statistics & Data Science
- around 550 primary/additional majors; Statistics (Concentration: Open, Math, Neuroscience); Economics-Statistics, Statistics and Machine Learning

# Revamping Introductory Statistics/Data Science

In midst of general education re-design, assessing issues/needs

- Students don't know concepts, can't see big picture
- Get too tied to software steps, can't analyze later on their own

- More reasoning, more writing, more doing
- More interdisciplinary work
- More experiential learning and self-reflection

*Our goals:* emphasize concepts; tell stories with data; more student-driven inquiry; understand how students solve problems

# A quick comment about computing/programming

National Academy of Sciences recently finished a two year study on *Envisioning the Data Science Discipline: The Undergraduate Perspective*. Included an overview survey of data science courses/programs nationally.

- Largely dominated by master's programs
- First programs are hybrids from statistics, computer science, information systems; "low-hanging fruit"
- Most data science courses are strong advocates for early coding exposure; jump right in

We're not convinced this is the right order. Yes, students should code, but not at the expense of the other material. The cognitive load associated with programming syntax can drown out the concepts.

Let's start with teaching the pipeline, language-agnostic.

# Integrated Statistics Learning Environment (ISLE)

- No coding syntax. Coding *concepts*.
  Start analyzing case studies/research scenarios in week 1.
- No servers needed; browser-based, local scientific computing
- "Data Set Explorer": upload (formatted) data, variables
- Students can save graphs and work to editors that create reproducible websites/documents for a portfolio
- Interactive; collect and propagate student answers
- We collect information on clicks, decisions, times, text, anything
  How/why do students analyze data?
- Combining tools like Java Script with RMarkdown;
  built in modular form; can “mix-and-match”
- Development led by Philipp Burckhardt

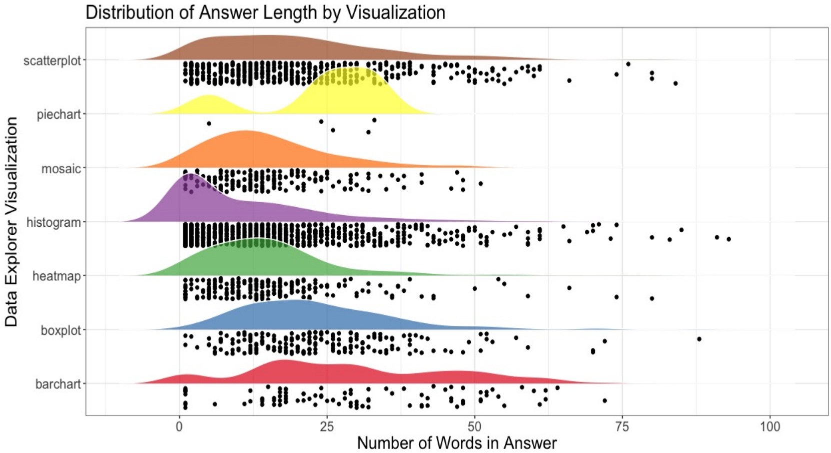
Hard to know best practices if you have no idea what they’re doing

# So what are we learning?

- IRB allows access to student action logs, etc after the semester is complete. Students can opt-out (so far they're not).
- Several rounds of smaller groups;
  live now in every intro stat class
- Examples from Fall 2017 ($n = 71$); Spring 2018 ($n = 130$)
  tens of thousands of actions, 11-12 labs, data analysis reports
- Everything tracked

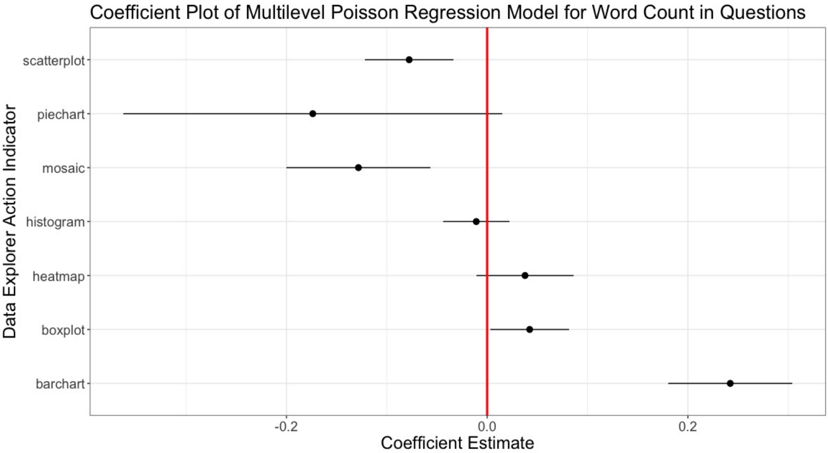
# Creating/Describing Graphs

Combine information about graphs they choose (parameters, etc) and how they describe them. Could do over time. Or use filters.



# Creating/Describing Graphs

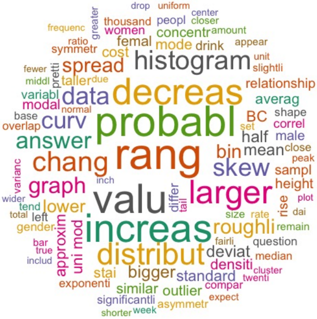
Relationships between visualization choice and answer length (Multi-level Poisson Regression)



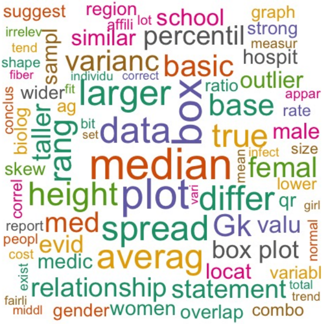
# Creating/Describing Graphs

Create word clouds using TF-IDF values from answers where students made histograms compared to boxplots

### Histograms

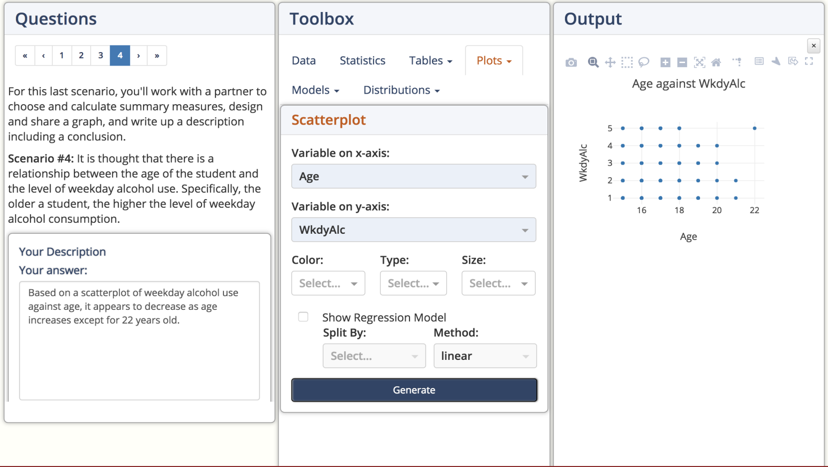


### Boxplots



# Open-ended Scenarios

Design graph for research question, critique current answer, rewrite



## Questions

For this last scenario, you'll work with a partner to choose and calculate summary measures, design and share a graph, and write up a description including a conclusion.

**Scenario #4:** It is thought that there is a relationship between the age of the student and the level of weekday alcohol use. Specifically, the older a student, the higher the level of weekday alcohol consumption.

Your Description

Your answer:

Based on a scatterplot of weekday alcohol use against age, it appears to decrease as age increases except for 22 years old.

## Toolbox

Scatterplot

Variable on x-axis: Age

Variable on y-axis: WkdyAlc

Color: Select...

Type: Select...

Size: Select...

□ Show Regression Model

Split By: Select...

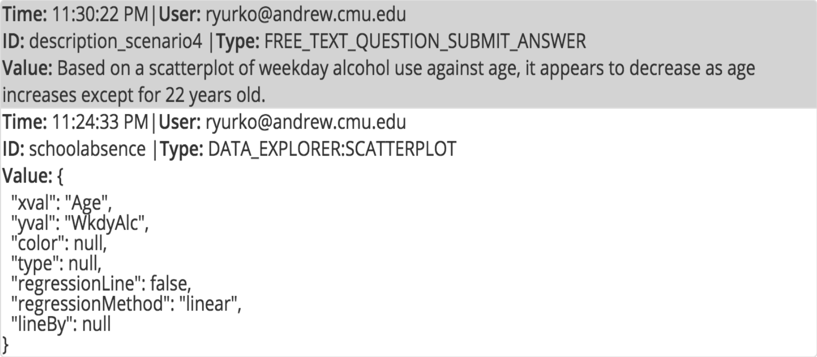
Method: linear

## Output

Age against WkdyAlc

# Open-ended Scenarios

Design graph for research question, critique current answer, rewrite



**Time:** 11:30:22 PM | **User:** ryurko@andrew.cmu.edu
**ID:** description_scenario4 | **Type:** FREE_TEXT_QUESTION_SUBMIT_ANSWER
**Value:** Based on a scatterplot of weekday alcohol use against age, it appears to decrease as age increases except for 22 years old.

**Time:** 11:24:33 PM | **User:** ryurko@andrew.cmu.edu
**ID:** schoolabsence | **Type:** DATA_EXPLORER:SCATTERPLOT
**Value:** {

```
 "xval": "Age",
 "yval": "WkdyAlc",
 "color": null,
 "type": null,
 "regressionLine": false,
 "regressionMethod": "linear",
 "lineBy": null
}
```

# Open-ended Scenarios

Lab session in week five of class uses a single dataset about school absences in Portugal, consists of four question scenarios:

- **Scenario 1**: Number of absences by location, urban or rural?
- **Scenario 2**: Older students more likely to miss school?
- **Scenario 3**: Academic performance by number of classes failed, differences between males and females?
- **Scenario 4**: Relationship between age and alcohol use?

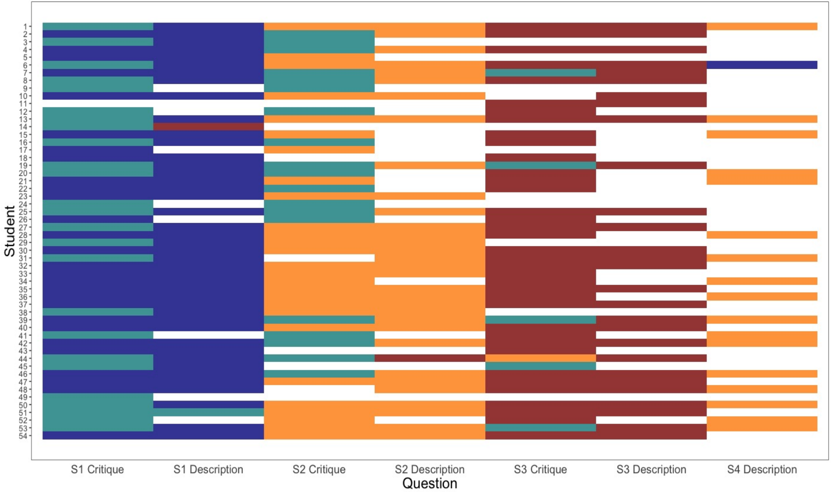
**Scenarios 1-3**: critique and write description with **explicit instructions** on what stats and graphs to edit/create

**Scenario 4**: only write description with **no guidance**

Refer to as: S1 Critique, S1 Description,..., S4 Description

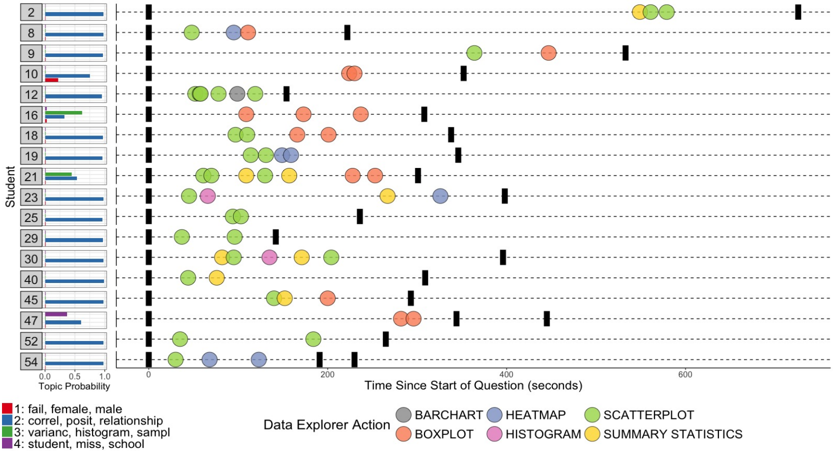
# Open-ended Scenarios

Cluster students by their TF-IDF values with spherical k-means



# Open-ended Scenarios

Topic models linking answers to timelines of their actions



# Other ISLE Features/Ongoing Research

- Voice activation, building slides/posters, random question generation; chat rooms; calculators
- Student/instructor progress dashboard (feedback)
- “Many Students, One Dataset” reproducibility studies
- Improving accessibility

- How/why do people do data science? Research data science?
- Students from different backgrounds might actually just be thinking about data differently (not incorrectly)
- Notions of reproducibility/replicability need to make room for “distributions of data analyses”; subjectivity of pipeline

# Access is not the same as equity

- Just building Data Science experiential learning/case study courses, programs, online materials, etc is not enough

“If you build it, they will come”
Sure, but will they actually play baseball when they get there?

- Non-STEM communities need accessible, understandable tools
- Need software/platforms that allow for customization without requiring comp background (for students, teachers)
- Give “ownership” to stakeholders

Upcoming Indian railway project: facilitating case-study based data analytics in classrooms with limited technological capability

Partnering with Problem Forward Data Science: training programs, Future of Work/Data Science ecosystem; jspm@problemforward.com

# The Behavioral Data Science Team/Upcoming

- Philipp Burckhardt
- Ron Yurko, Frank Kovacs
- Chris Genovese
- Ciaran Evans, Gordon Weinberg
- Yeuk Yu Lee, Robin Mejia, Wren Hemmel, Sarah Tanjung
- Alex Reinhart, Amanda Luby (soon Swarthmore), Jerzy Wieczorek (Colby), Josue Orrellana Arreaga, Peter Elliott, Kevin Lin, Justin Hyun, Christopher Peter Makris, Mikaela Meyer
- U.S. Conference on Teaching Statistics: Penn State, May 2019
- Carnegie Mellon Sports Analytics Camp `http://summer.stat.cmu.edu`
- Women in Data Science: Pittsburgh @CMU `http://www.stat.cmu.edu/wids`

rnugent@stat.cmu.edu, `http://www.stat.cmu.edu/~rnugent`; @CMU_Stats