Data Science for Everyone with ISLE

Leveraging Web Technologies to Increase Data Acumen

Rebecca Nugent

Stephen E. and Joyce Fienberg Professor of Statistics & Data Science
Carnegie Mellon Statistics & Data Science

isle

Interacting with Data Science

Can be as “small” as participating in a survey

WWIB

Who Wore This Cat-Lady Look Best: Molly Sims or Beth Stern?

usmagazine.com

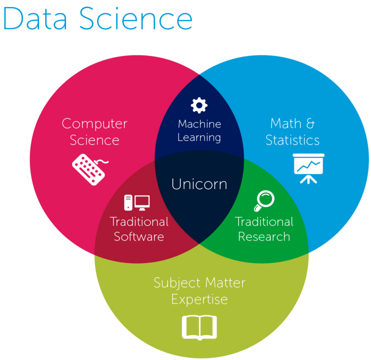
Interacting with Data Science

Or as “large” as living in fully simulated environment

The Matrix

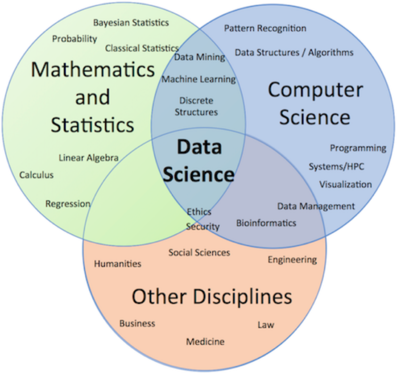
Early Definitions

Focused on overlapping sets of skills from different disciplines; static



Early Definitions

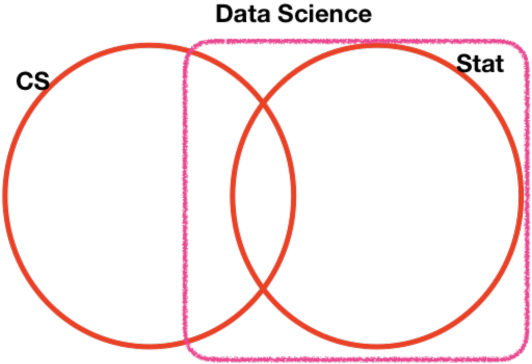
Venn Diagram viewpoint created competing ownership claims



ACM Task Force on Data Science White Paper Draft

Early Definitions

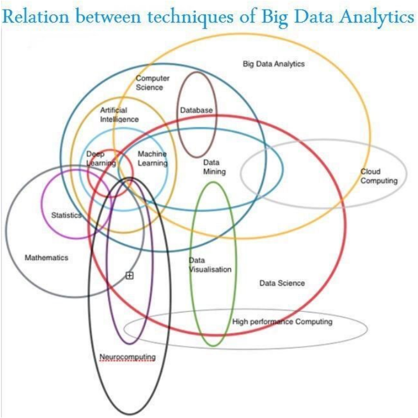
Including suggestions that Data Science is just a re-branding of Statistics with techniques for “Big Data” sets



Sent by Rob Gould, UCLA

Early Definitions

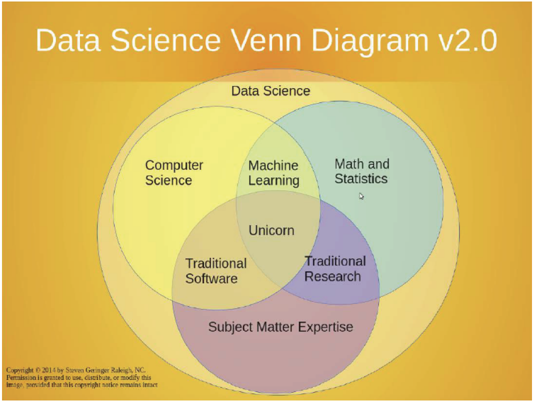
Conversation got a little out of control....



Mara Averick, RStudio

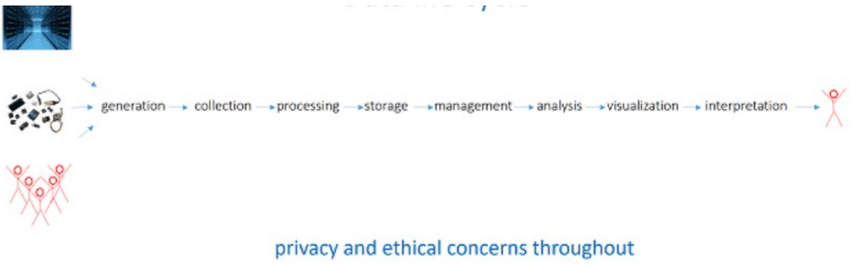
Early Definitions

Initially landed on perception of all-encompassing; curriculum/program development struggled with how to train students and professionals



Data Science, A View

Thought of as an process or workflow; solving real problems with real data



J. Wing (2019), Harvard Data Science Review

- *Management* includes security, elements of data engineering
- *Interpretation* includes communication

In practice, move roughly from left to right but with loops and iterations; experts often focus on specific pieces; project managers oversee pipeline

The Science of Data Science

Huge emphasis on having reproducible and/or replicable results;
made far more complicated by the pipeline nature of the problems

- **Reproducibility**: ability to implement the same experiment/code/procedures with the same data to obtain the exact same results
- **Replicability**: obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data (NAS)

Most can agree on need to carefully document all code, analyses, algorithms;
slightly smaller group would add requirements to public post/disseminate
all work, code, data sets, etc.

The Science of Data Science

What does this mean for students and practitioners?

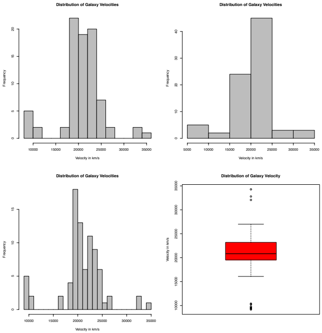
- **Reproducibility**:
  - Do the same steps I did last time, get the same thing
  - Oh god, can't find my notes, have no idea how I got this result
  - I copied my friends' answers/code but claiming reproducibility....
- **Replicability**:
  - My friends and I have different random samples of the same data set/distribution; slightly different but similar results
  - My colleagues and I collected different data sets in a similar way (survey, etc); have same/different results for same question

And p-values? Really like swiping right on Tinder.
Not so much a lifelong commitment but more just a sign of interest...

The Science of Data Science

While much of data science relies on extracting signal/structure using machine learning algorithms, much is based on human subjective decisions.

Velocities of 82 galaxies; multimodality - voids and superclusters (Roeder, JASA, 1990)



The Science of Data Science

Many analysts, one dataset *(Silberzahn, et al 2018)*

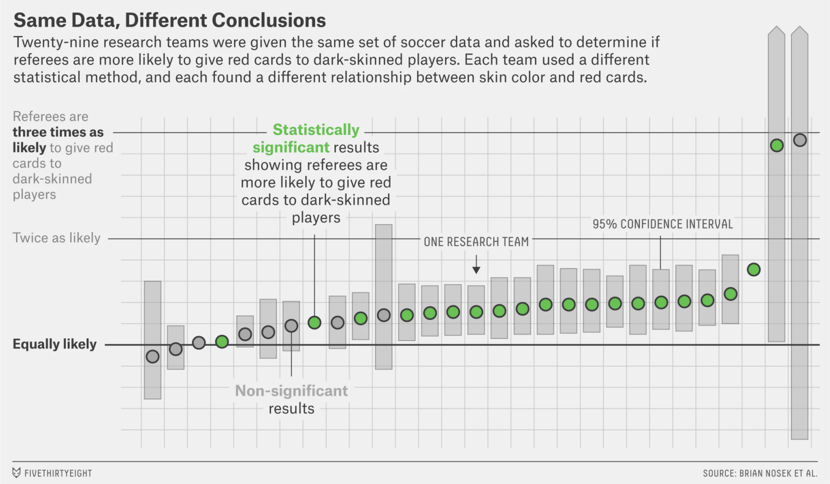
29 teams of analysts, same dataset, same question:

> *Are soccer referees more likely to give red cards to players with dark skin than to players with light skin?*

Analysis stages:

- Teams worked independently
- Peer-review, exchanged information and analysis
- Revisions and submit final conclusions

The Science of Data Science



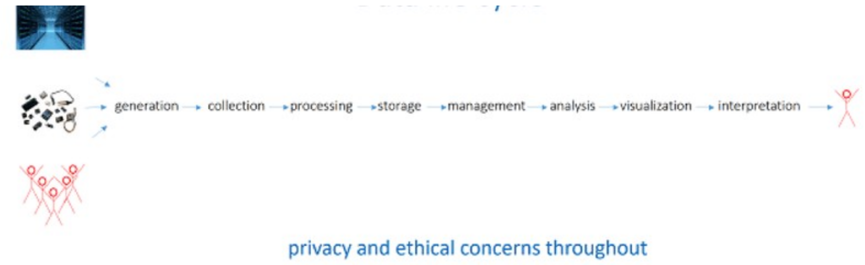
Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.

SOURCE: BRIAN NOSEK ET AL.

The Science of Data Science

Thought of as an process or workflow; solving real problems with real data



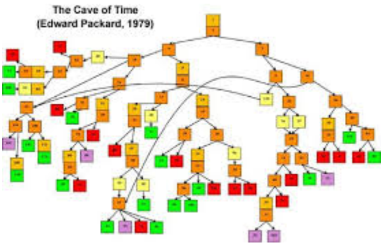
J. Wing (2019), Harvard Data Science Review

- *Management* includes security, elements of data engineering
- *Interpretation* includes communication

In practice, move roughly from left to right but with loops and iterations; experts often focus on specific pieces; project managers oversee pipeline

The Science of Data Science

The Ultimate Choose Your Own Adventure Book
(hopefully data science doesn't lead to being trapped in a cave forever)



With apologies to Edward Packard

The Science of Data Science

- Explosion of Stat & Data Science programs, courses, materials, tools
- 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
- How can we incorporate human decision-making into a data science interface/pipeline?

Behavioral Data Science

Some current actions/questions:

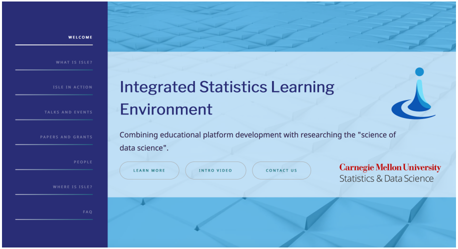
- *Think-Alouds*: recording what you're thinking while doing your work
- *Crowd-Sourcing*: have groups work independently on same problem; how do you reconcile differences in data analysis variations?
- *Data Analysis Population*: Is our one data analysis is "different"?

Carnegie Mellon University

- Private university in Pittsburgh, PA; R1 research university designation
- $\approx$ 7000 undergrads, 7000 grads
- Seven colleges: College of Fine Arts, Dietrich College of Humanities & Social Sciences, College of Engineering, Heinz College of Information Systems and Public Policy, 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
- $\approx$ 550 primary/additional majors; Statistics (Concentration: Open, Math, Neuroscience); Economics-Statistics, Statistics and Machine Learning
- Almost all of our course sizes (UG through PhD) are in the hundreds

Commonly hear that learning software (early) gets in way of learning content

Integrated Statistics Learning Environment (ISLE)



Integrated Statistics Learning Environment

Combining educational platform development with researching the "science of data science".

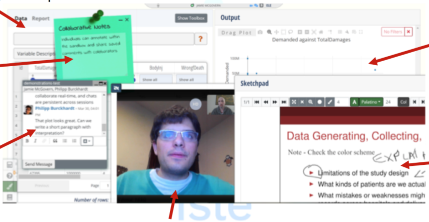
http://www.stat.cmu.edu/isle

- Labs; Surveys; Widgets
- Sketch Pads/Lecture Slides; Group Collaboration
- Data Explorer; Reports; Presentations
- Peer to Peer Sharing; Chat Rooms
- Data Provenance; Reproducibility
- Action Logs; Grading/Annotations

Integrated Statistics Learning Environment (ISLE)

Enabling Technology: Capstone Collaboration in the ISLE Sandbox

- Interact with data / data exploration
- Share comments on data sets
- Communicate via chat with simple file sharing (images/pdfs)



Collaborative Notes

individuals can annotate within the sandbox and share saved comments with collaborators

demonstrations-test

Jamie McGovern, Philipp Burckhardt

collaborate real-time, and chats are persistent across sessions

Philipp Burckhardt • Mar 30, 04:01 PM

That plot looks great. Can we write a short paragraph with interpretation?

Data Generating, Collecting,

Note - Check the color scheme

EXPLAIN

- Limitations of the study design
- What kinds of patients are we actual
- What mistakes or weaknesses migh

- Collaborate on group reports with templates
- Work via shared notes/sketchpad functionality
- Communicate via real-time video

Integrated Statistics Learning Environment (ISLE)

http://www.stat.cmu.edu/isle

- browser-based: multiple operating systems and devices
- moving computational load from server to client via JavaScript, stdlib (`https://stdlib.io`)
- web technologies typically not built with computing needs in mind; slowly changing
- continuous real-time connection between users and server through web sockets (socket.io)
- integrated video & audio chatting through Jitsi meet
- recomposable components (React.js) for e-learning that can be combined/customized in an accompanying editor (Electron application)

Integrated Statistics Learning Environment (ISLE)

http://www.stat.cmu.edu/isle

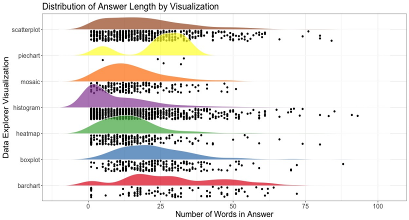
- Hundreds of students at Carnegie Mellon, undergraduate and graduate
- In beta at other universities
- Statistics/Data Science through English/Humanities classes
- Analyze how different fields write
- Flipped classroom, remote learning, choose your own adventure
- Retraining/upskilling/ExecEd: health care, finance, manufacturing, etc
- Interactive journal article content
- UN pilot initiative to improve statistics/data science education in developing countries

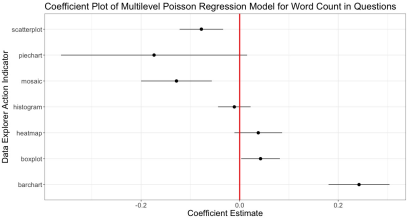
So what are we learning/researching?

- IRB allows access to action logs, etc after the course is complete. Students can opt-out (so far they're not).
- Everything tracked. Everything.
- Writing and structuring arguments about data
- How to optimize a data science team; group collaboration
- Populations and variance of data analyses ("*Many Students, One Dataset*")
- Data literacy; longitudinal impact related to access and equity
- Examples from Fall 2017 Intro Stat ($n = 71$); Spring 2018 ($n = 130$) tens of thousands of actions, 11-12 labs, data analysis reports

Creating/Describing Graphs

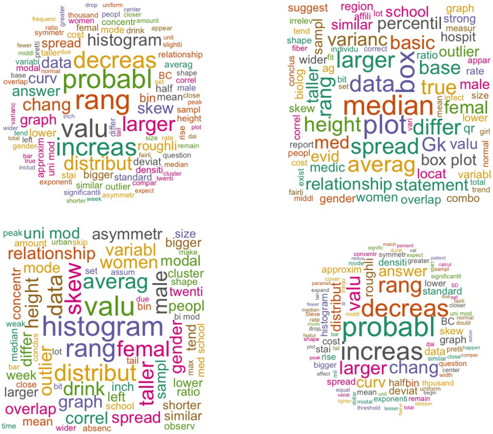
Combine information about graphs they choose (parameters, etc) and how they describe them. Could do over time. Or use filters.





Creating/Describing Graphs

Comparison word clouds via answer TF-IDF values (graph type; over time)



Open-ended Scenarios

Studying school absences in Portugal:

- **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?

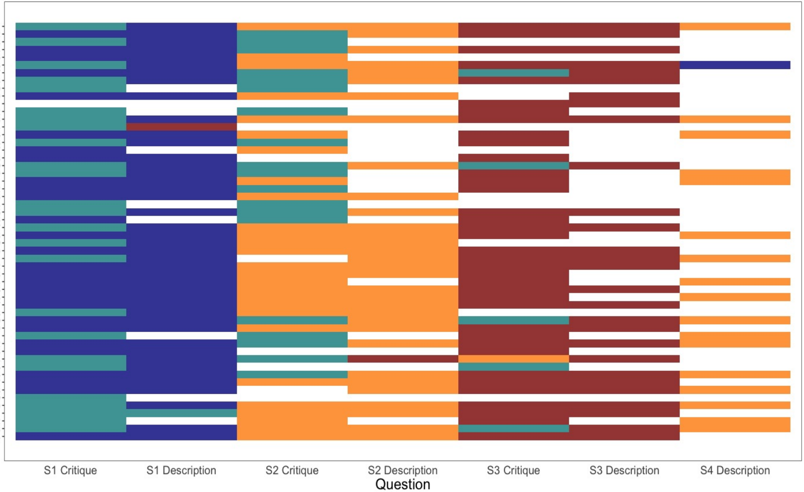
Scenarios 1-3: critique and write description with **explicit instructions** on what stats and graphs to edit/create

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

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

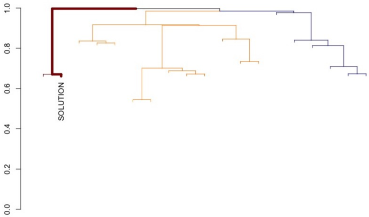
Open-ended Scenarios

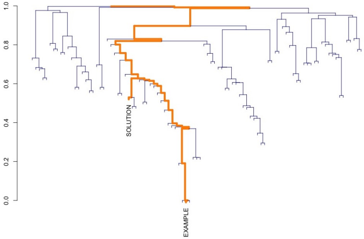
Cluster students by their TF-IDF values with spherical k-means



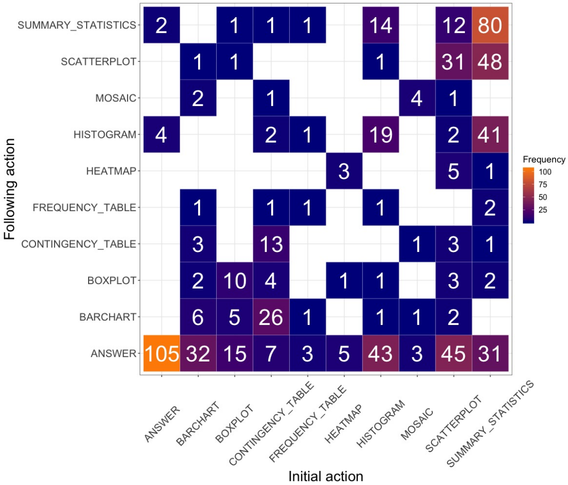
Open-ended Scenarios

How different are the answers from the solution? auto-grading/copy-paste





Transition Matrix for Data Analysis Actions

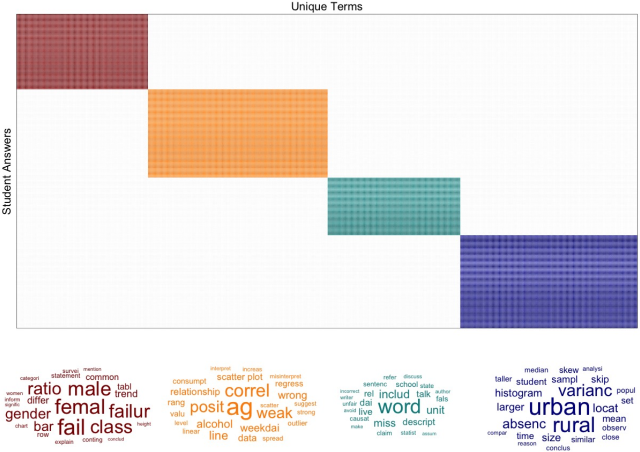


Following action \ Initial action	ANSWER	BARCHART	BOXPLOT	CONTINGENCY_TABLE	FREQUENCY_TABLE	HEATMAP	HISTOGRAM	MOSAIC	SCATTERPLOT	SUMMARY_STATISTICS
SUMMARY_STATISTICS	2		1	1	1		14		12	80
SCATTERPLOT		1	1				1		31	48
MOSAIC		2		1				4	1	
HISTOGRAM	4			2	1		19		2	41
HEATMAP						3			5	1
FREQUENCY_TABLE		1		1	1		1			2
CONTINGENCY_TABLE		3		13				1	3	1
BOXPLOT		2	10	4		1	1		3	2
BARCHART		6	5	26	1		1	1	2	
ANSWER	105	32	15	7	3	5	43	3	45	31

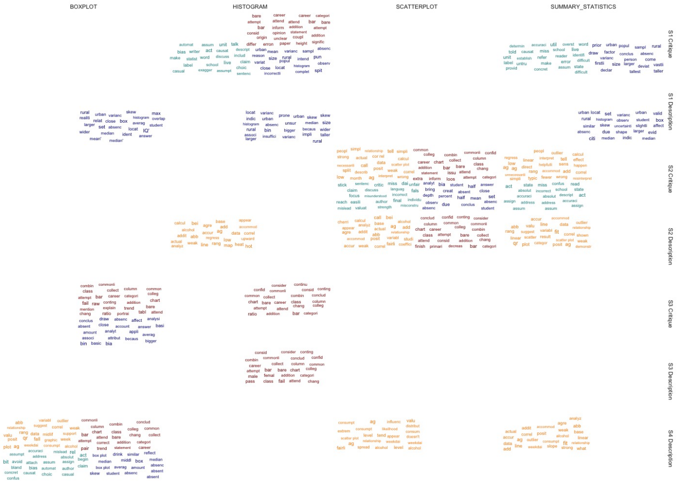
Looking for Clusters of Users and their Words

Directional Co-Clustering using von-Mises Fisher Mixture Models

(Banerjee, 2005; Raftery, Dean 2006)

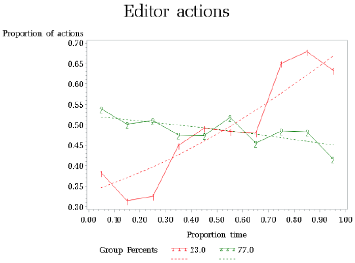


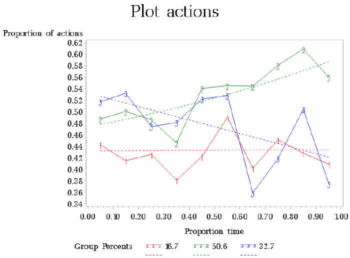
Looking for Clusters of Users and their Words



Incorporating Timelines

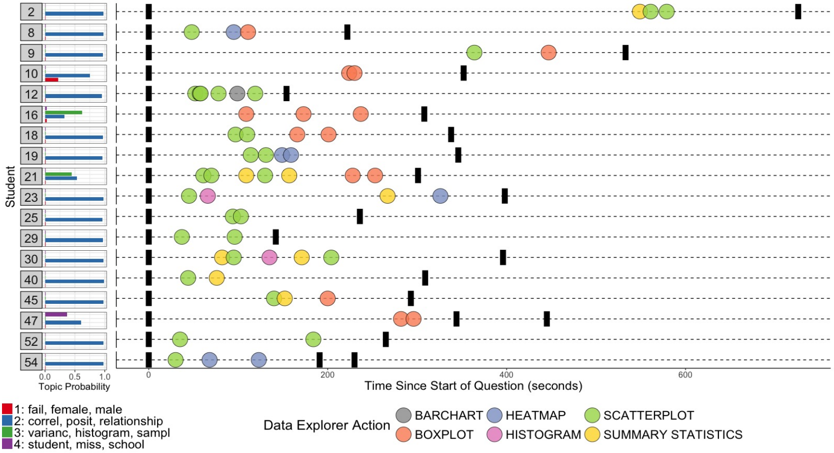
Analyzing how people write data analysis reports





Incorporating Timelines

Topic models linking answers to timelines of their actions



Takeaways so far

- Focusing only on building Data Science tools is missed opportunity
- How/why do people do data science? Research data science?
- Not just for tech folks; non-STEM communities need accessible tools
- People from different backgrounds often just thinking about data differently (not incorrectly)
- Need software/platforms that allow for customization without requiring comp background (for students, teachers)
- More interaction with data analysis pipeline (start to end)
- Give “ownership” to stakeholders
- Notions of reproducibility/replicability need to make room for “distributions of data analyses”; subjectivity of pipeline

Looking Forward

- New frontier is the Nexus of Humanities and Technology
- Incorporating Behavioral Sciences into Data Science, Machine Learning, AI, the next big buzz word, etc
- Need to improve our understanding of variation in decisions, actions and their downstream impact; secondary, tertiary effects on society
- Cool, new tools are fun but will only get us so far.
  Humans are the problem, but also the solution.

The Behavioral Data Science Team

- Rebecca Nugent, Philipp Burckhardt
- Ron Yurko, Frank Kovacs, Ciaran Evans, Gordon Weinberg,
  Chris Genovese, Wren Hemmel, Sarah Tanjung, Jamie McGovern

Feel free to contact Rebecca Nugent or isle@stat.cmu.edu to learn more!