

# Democratizing Data (Science): Empowering and Expanding Opportunities for Both Students and Educators

Rebecca Nugent

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

USCOTS 2021

# First up, how are you?

[https://www.youtube.com/watch?v=Zy\\_y9yOrgxk](https://www.youtube.com/watch?v=Zy_y9yOrgxk)

# Gratitude, admiration, and respect

- ▶ You are all amazing. For real.
- ▶ This past 15 months has been utterly exhausting.  
And yet you're still here, pushing forward stat & data science education.

Possibly in your pajamas and drinking an adult beverage.  
But you're here.

- ▶ Everyone is under-resourced and under-paid.  
Once we recover, let's tackle that as a group. Invest in people first.
- ▶ Huge support and appreciation to everyone creating virtual/hybrid content and working to provide quality materials and educational experiences for the 100000s of students nationally taking our classes.

Special shout-out to Kelly McConville and Allan Rossman and the entire CAUSE/USCOTS team

# Data Science, A View

A process or workflow; solving real problems with “extracting value from data”



J. Wing (2019), Harvard Data Science Review

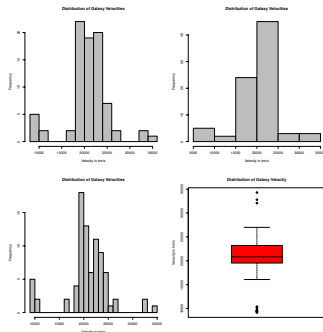
- ▶ *Management* includes security, elements of data engineering
- ▶ *Interpretation* includes communication

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

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.



# The Science of Data Science

A process or workflow; solving real problems by “extracting value from 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 (where hopefully the data analysis doesn't lead to being trapped in a cave forever):



With apologies to Edward Packard

# 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.

# Reproducibility vs Replicability

What does this mean for (intro) students?

▶ **Reproducibility:**

- ▶ Do the same steps I did last time, get the same thing
- ▶ Or I did the same steps as my friend/used the same code, we got the same thing; are we now in trouble for cheating?
- ▶ I definitely copied my friends' answers/code and am now claiming reproducibility....

▶ **Replicability:**

- ▶ My friends and I have different random samples of the same data set/distribution; slightly different but similar results
- ▶ My friends 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...

*Ron Wasserstein, ASA*

# Reproducibility vs Replicability

Common issues that we face teaching these concepts:

▶ **Reproducibility:**

- ▶ How do students keep track of their work?
- ▶ Writing (commented) code is one way; what happens with pre-coding classes?
- ▶ More importantly, how do you keep track of decisions?

▶ **Replicability:**

- ▶ Students commonly work on one piece at a time (t-test, graph interpretation, etc)
- ▶ See variation in simulation-based/sampling distributions activities, not in real data analysis decisions
- ▶ Students rarely all work on/present the same project

# Open-Ended Data Analysis

Issues/questions we commonly hear:

- ▶ How do we know this is right?
- ▶ My friend's graph looks like this; why does mine look different?
- ▶ My friend's p-value is this; why is mine different?
- ▶ How do I decide what to do?
- ▶ Seriously, is this right?

(Intro) students don't have enough exposure to the entire data science/statistical analysis pipeline.

Have little to no intuition for variation introduced by decision-making; how do we reconcile differences across analyses?

# Data Science Experiential Learning

## Partnering with Industry

- ▶ Recent job market has strongly pulled students toward wanting real-world engagement
- ▶ Impact of experiential learning can be large; learning to work with outside clients, etc
- ▶ Match teams of students to projects with clients (PhD, MSP, UG)
- ▶ Students have faculty supervision and PhD assistance
- ▶ Educational Project Agreement (semester, year, etc); IP, NDAs, legal
- ▶ Promotes synthesis and just-in-time learning
- ▶ Students can start this as rising sophomores

# Data Science Experiential Learning

- ▶ Forecasting/nowcasting the flu using disparate data sources
- ▶ Modeling the impact of state exec orders on COVID-19 cases/deaths
- ▶ Build algorithms/models to optimize job opportunities for gig workers with less stable income
- ▶ Disruption of global supply chains; determine effect of social and political events on deliveries; predict need to re-route earlier
- ▶ Characterizing and forecasting the needs of grocery retail (particularly in the face of shortages)
- ▶ Building predictive models to characterize injury/incident rates at construction sites; zero harm initiatives
- ▶ Using scanned receipts to describe consumer behavior; improve marketing, item availability
- ▶ “Tinder for Brands” - based on opinion surveys, match celebrities to brand endorsements
- ▶ Build models to predict expected win probability play-by-play:

`nflscrapR`

# Reducing Incidents and Injuries

Global engineering, procurement, construction company specializing in infrastructure development (water, power, telecom, oil and gas)

- ▶ Projects all over the world; span a few months to several years
- ▶ Employees include subcontractors, local unions
- ▶ Have a strong commitment to reducing/eliminating injuries and incidents
- ▶ Ranges from first aid to property damage to theft to environ. events

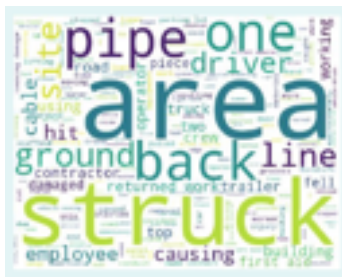
Characterize and Predict Injury Types and Rates

- ▶ Use a third-party data warehouse to log incidents
- ▶ Including typical info needed for Occupational Safety and Health Administration requirements/standards
- ▶ Additional info about projects, employees, injuries including written description of what occurred
- ▶ Mandatory reporting; varying levels of compliance



# Reducing Incidents and Injuries

- ▶ Build tools to allow silos of data warehouse to “talk to each other”
- ▶ Interactive interface that summarizes groups of similar projects and models most common injuries
- ▶ Allows for early identification of potential problems; can deploy educational programs, prevention tips prior to project start
- ▶ Increases employee health outcomes and satisfaction; decreases company financial losses, improves project completion



# Modeling Pandemics

Carnegie Mellon DELPHI group

*Developing the Theory and Practice of Epidemiological Forecasting*

- ▶ Joint with Statistics & Data Science, Machine Learning, Comp Biology
- ▶ Part of Pittsburgh-based MIDAS National Center of Excellence (epidemiologists, virologists, infectious disease experts, legal experts,..)
- ▶ <https://delphi.cmu.edu/>
  
- ▶ National competition for predicting flu rates: “nowcasting”
- ▶ DELPHI won four years in a row; now a CDC Center for Excellence
- ▶ Incorporating non-traditional data sources and information

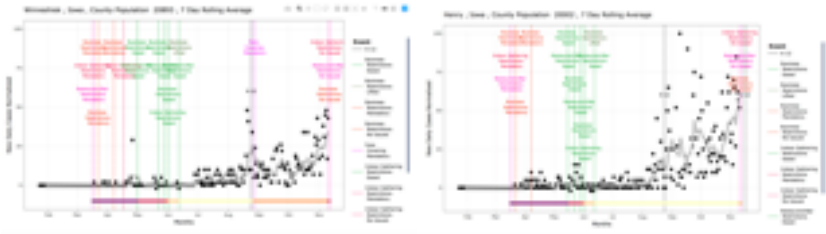
And now more recently, we've turned to COVID-19:

<https://covidcast.cmu.edu/>

# Modeling Pandemics

PHIGHT COVID (collab with Pitt Microbiology & Molecular Genetics )

<https://phightcovid.org>



## Takeaways/Next Steps

- ▶ Data Science Pipeline: integrated set of tasks that all impact each other
- ▶ Goal should be “better” data and more robust, simpler AI/ML models
- ▶ What impact does human behavior have on our data? variables?  
On the choice of model? On the model performance?

If you are not talking to content experts early and often, you will be starting over. Not data-driven decision-making - it's data-informed decision making.

- ▶ Experiential Learning programs are valuable and really hard to build
- ▶ Want to build a network/repositories of shared materials and projects that support broader range of educational institutions (e.g., ISLE)
- ▶ Piloting project partnerships now
- ▶ Please contact me if you're interested (rnugent@stat.cmu.edu)

# 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

# Where we were/Where we're going

Interviews w/ different depts (and industries) about issues/needs

- ▶ Students don't know the concepts
- ▶ Get tied to the specific software syntax or steps (Minitab)
- ▶ Can't see the big picture
- ▶ Classes aren't really for them

*Our goals:*

- ▶ Modernize courses; support different styles of learning (incl. remote)
- ▶ Emphasize concepts; tell stories with data
- ▶ More student-driven inquiry, case studies
- ▶ More adaptive material
- ▶ learn how different students interact with data

# Integrated Statistics Learning Environment (ISLE)



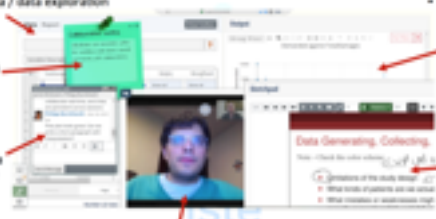
<http://www.stat.cmu.edu/isle>

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

# Integrated Statistics Learning Environment (ISLE)

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

- ▶ browser-based: multiple operating systems and devices can be downloaded for offline access
- ▶ Conformance to most of the World Wide Web Consortium's Web Content Accessibility Guidelines (W3C WCAG 2.4)
- ▶ Work on improving usability when using screen readers
- ▶ integrated video & audio chatting through Jitsi meet

# Integrated Statistics Learning Environment (ISLE)

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

- ▶ Hundreds of students at Carnegie Mellon, undergraduate and graduate
- ▶ Thousands of students in beta at other universities, community colleges, liberal arts, etc; more each semester
- ▶ Not just Statistics/Data Science; have other STEM, Forensic Science, Criminal Justice, English, Philosophy, Prof Dev Skills, Data Literacy..
- ▶ Flipped classroom, remote learning, choose your own adventure
- ▶ Retraining/upskilling/ExecEd: health care, finance, manufacturing, etc
- ▶ Interactive industry dashboards; telling “Stories with Data”
- ▶ Interactive journal article content

Roles	Phases in the Data Life Cycle:										Data Science Capability Level	
	Question	Define	Coordinate	Generate	Collect	Curate and Manage	Analyze	Visualize	Discern and Interpret	Assess		
Data engineer		implementation specialist										
Data scientist	assets refine	science and theory					deeper capabilities				acumen	
Data analyst	assets refine						deeper capabilities					
Data users	assets refine											
Domain experts	pose										literacy	
Leaders, decision-makers, & managers	pose											

- ▶ NASEM, *Empowering the Defense Acquisition Workforce to Improve Mission Outcomes using Data Science*, May 2021
- ▶ Focus on government (DoD Acquisitions) but included data science industry teams
- ▶ Vast majority of workers will fall into non-technical data roles but will need data science literacy
- ▶ Seeing a big increase in industry reaching out for upskilling/re-training to think/talk/write about data

# Integrated Statistics Learning Environment (ISLE)

<https://isle.stat.cmu.edu/USCOTS2021/USCOTSSandbox>

# So what are we learning/researching?

- ▶ IRB allows access to action logs, etc after the semester is complete. Students can opt-out (so far they're mostly 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
- ▶ Optimizing lesson/lecture structure wrt student engagement (*early*)
- ▶ Adaptive review repositories with random question generation (*early*)
- ▶ Examples from Fall 2017 Intro Stat ( $n = 71$ ); Spring 2018 ( $n = 130$ ) tens of thousands of actions, 11-12 labs, data analysis reports



# 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

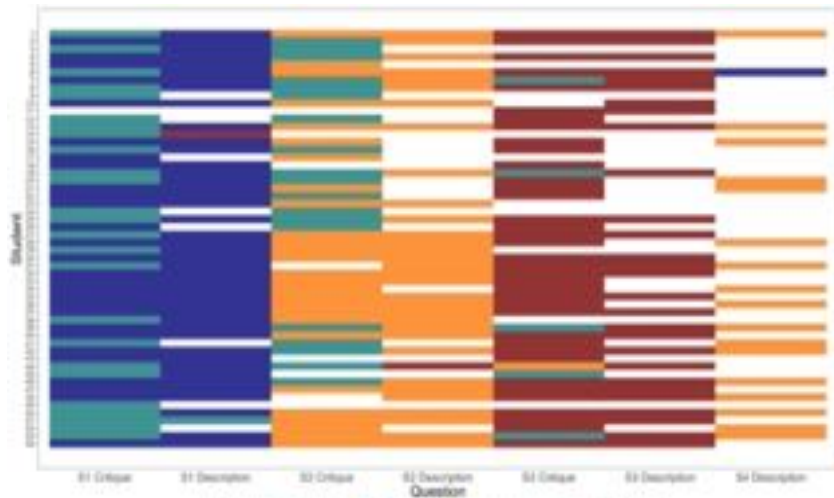
**Scenario 4:** only write description with **no guidance**

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



# Open-ended Scenarios

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



# Incorporating Timelines

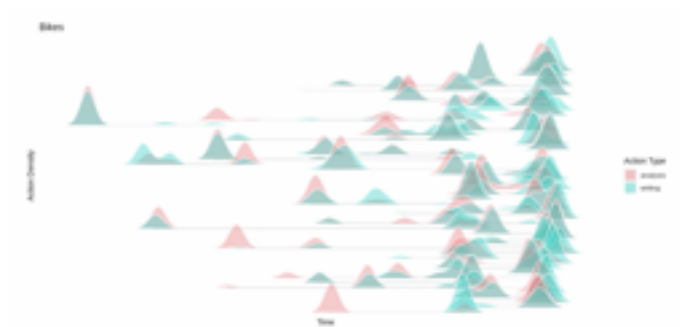
## Understanding and Analyzing Analysis Pathways



Sunburst chart of any three subsequent action types in students' data analyses. The light-red shaded pieces correspond to histograms, purple for summary statistics, and light-green ones for two-sample z-tests

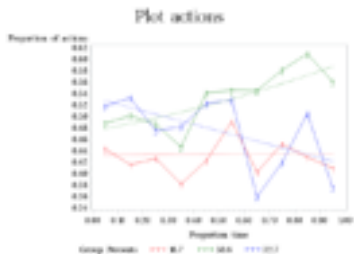
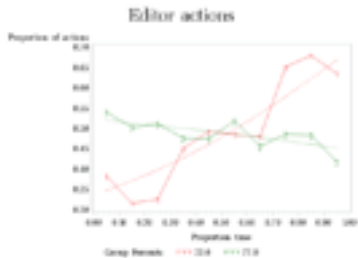
# Incorporating Timelines

Analyzing (and clustering) how people build and write data analysis reports



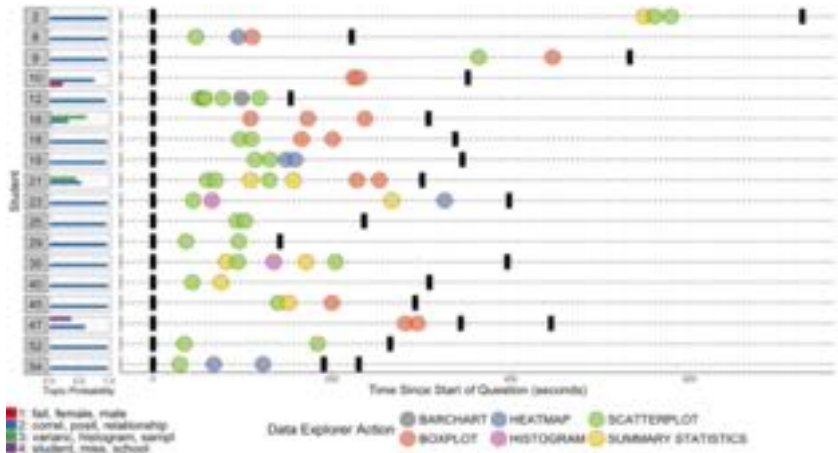
# Incorporating Timelines

Analyzing (and clustering) how people build and write data analysis reports



# Incorporating Timelines

Topic models linking answers to timelines of their actions



# Data Science, a Team Sport

We want to:

- ▶ teach our students/practitioners how to do (good) data analysis
- ▶ teach them how to work together

Now we have different sets of decisions from different people with different backgrounds and somehow they're supposed to agree on one final product.

So what happens? How do we do it?

- ▶ Have them sit and work in groups during lab/class  
*Some of them talk to each other; others stare at their laptop/paper; others wait until class is over*
- ▶ Have them work together outside of class  
*Completely impossible to find time that all 4-5 students can meet  
Students end up just dividing up work; no cohesion  
1-2 people do all the work and the others do the intro/conclusion*
- ▶ Have them use "shareable" tools: Google, Slack, etc  
*Easy to ignore groupmates but sort of works. Assumes that students are okay with downloading and running tools. Can get by without speaking to each other.*

# Data Science, a Team Sport

The truth is: we don't know what they're doing. And we're not doing a very good job of teaching them how to optimize working together.

We want to help them learn about data by using data.

<https://isle.stat.cmu.edu/USCOTS2021/USCOTSSandbox/>

In ISLE:

- ▶ can be assigned to groups
- ▶ can chat with each other and the instructor/TA
- ▶ can view each other over video

On the back end:

- ▶ track all the actions, time stamps, you name it
- ▶ can analyze when and how people worked
- ▶ can tie to performance but also tailor instruction

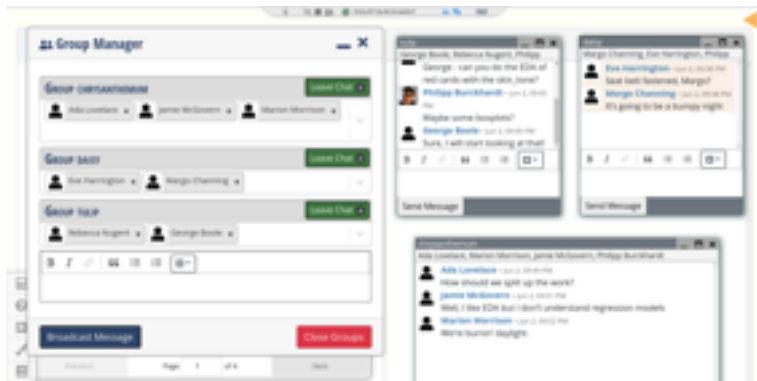
# Data Science, a Team Sport





# Data Science, a Team Sport

*The Instructor View:*



# Data Science, a Team Sport

## *Challenges:*

- ▶ How can we reduce the mental tax of having to switch between different contexts (chat, report writing, data analysis)?
- ▶ How can one avoid having some students not participating, leaving their groupmates alone in finishing their reports?
- ▶ When work independently on different sections, how can one synthesize the findings?

## *Current Work:*

- ▶ Improved collaborative editing experience including a searchable history of analysis actions and text edits
- ▶ Full group projects which may be completed asynchronously over a period of time
- ▶ Giving the students feedback as they work
- ▶ Compare different kinds of group assignments

## Takeaways so far

- ▶ Only building Data Science courses/programs is missed opportunity
- ▶ How/why do people do data science? Research data science?
- ▶ Notions of reproducibility/replicability need to make room for “distributions of data analyses”; subjectivity of pipeline
- ▶ Non-STEM communities need accessible tools
- ▶ People from different backgrounds might actually just be thinking about data differently (not incorrectly)
- ▶ Need software/platforms that allow for customization without requiring comp background (for students, teachers)
- ▶ Not just going to build traditional courses; need more flexible content delivery
- ▶ More interaction with data analysis pipeline (start to end)
- ▶ Give “ownership” to stakeholders; data science is for everyone

# Data Science-Related Outreach

- ▶ “Intro to Data Science” Workshop for CMU graduate students; expanding to other non-CMU populations
- ▶ *Carnegie Mellon Sports Analytics Center*: [stat.cmu.edu/cmsac](http://stat.cmu.edu/cmsac)  
Faculty/PhD research, conferences, workshops, heavy outreach component, **summer research programs**

Hub for SCORE, a national network in partnership with ESPN among others focusing on building a broad pipeline to STEM through sports

*Birds of a Feather 1-06*: Tues 6/29 11:15am EDT, Michael Schuckers

- ▶ *Women in Data Science Pittsburgh @ CMU*: [stat.cmu.edu/wids](http://stat.cmu.edu/wids)  
networking and research events across multiple campuses; strong engagement within Data Science/AI industry community, local schools, etc; Women in Statistics mentoring programs

# Looking Forward

- ▶ Incorporating Behavioral Sciences into Data Science, Machine Learning, AI, the next big buzz word, etc
- ▶ No one asks us about how to code; everyone asks about how to communicate, work together, ask the right questions about data
- ▶ Cool, new tools are fun but will only get us so far. Humans are the problem, but also the solution.
- ▶ More focus on people and how they interact with the tools, support different learning paths and perspectives
- ▶ Promote sharing of projects, materials; take advantage of remote teaching lessons learned
- ▶ Invest in people, their careers and their lives

*The Behavioral Data Science Team:* Rebecca Nugent, Philipp Burckhardt, Jamie McGovern, Chris Genovese, Ciaran Evans, Gordon Weinberg, David Brown, Mike Laudenschach, Ron Yurko, Frank Kovacs, Wren Hemmel, Sarah Tanjung, Xiaoyi Yang

Have a wonderful USCOTS!

