Stats can be fun, I swear!

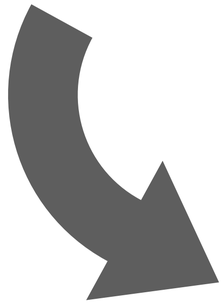
Cornell University - PSYCH1130

Ashley Mullan - September 24, 2024

Follow along with me!

https://bit.ly/ash-talks

(load the slides!)



Coming Soon to Theaters

1. **Invited Talk**: Misclassification Challenges in Quantifying the Relationship Between Diabetes and Local Access to Healthy Food
   Department of Mathematics, University of Scranton - March 2025

2. **Invited Panel**: Graduate Admissions Advice
   Electronic Undergraduate Student Research Conference
   November 2024

3. **Invited Talk**: For Better or For Worse: The First Kiss Effect on TV Ratings
   International Day of Women in Statistics and Data Science
   October 2024

4. **Guest Lecture**: Stats can be fun, I swear!
   PSYCH1130, Cornell University - September 2024
   Slides - App

Today’s TL;DR

1. How’d I end up on your screen today?
2. Here are some of my hot takes about statistics.
3. Don’t believe me? Try it yourself!
4. Ask me things!

How’d I end up on your screen today?

Try new things, you'll meet people!

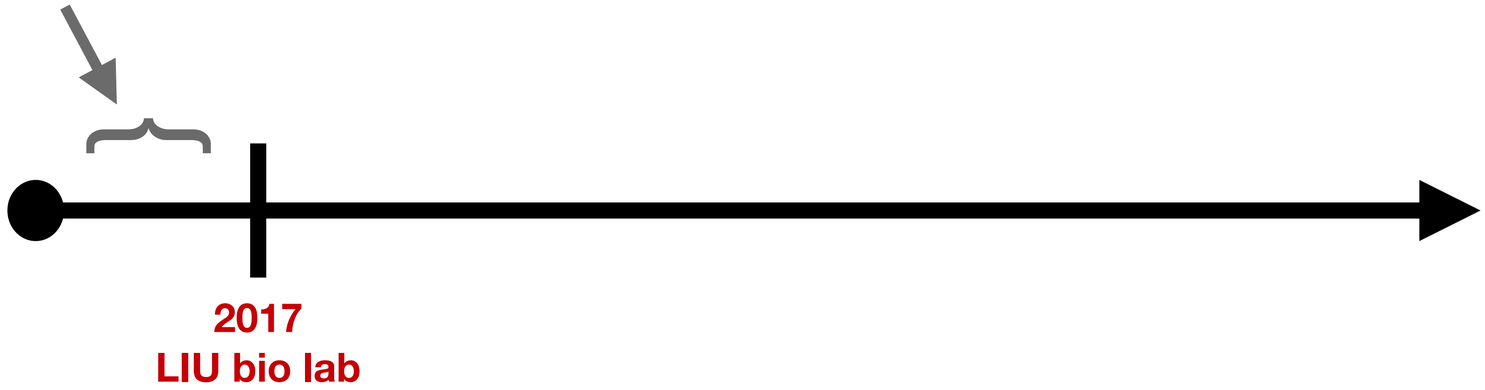
(the origin story)

(not even close to scale)

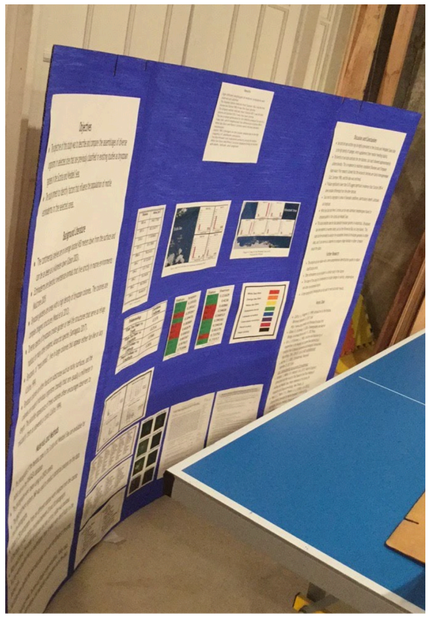
Try new things, you'll meet people!

(the origin story)

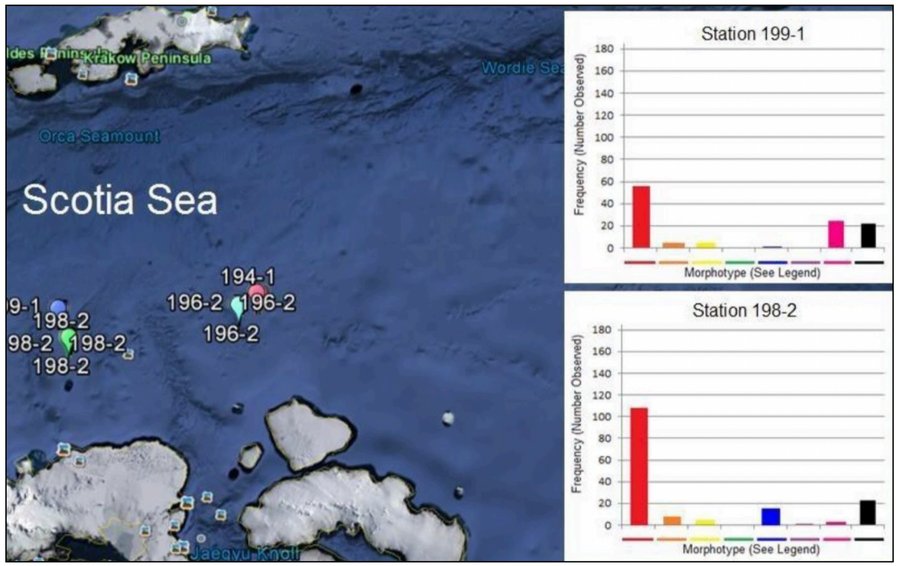
(not even close to scale)



2017
LIU bio lab



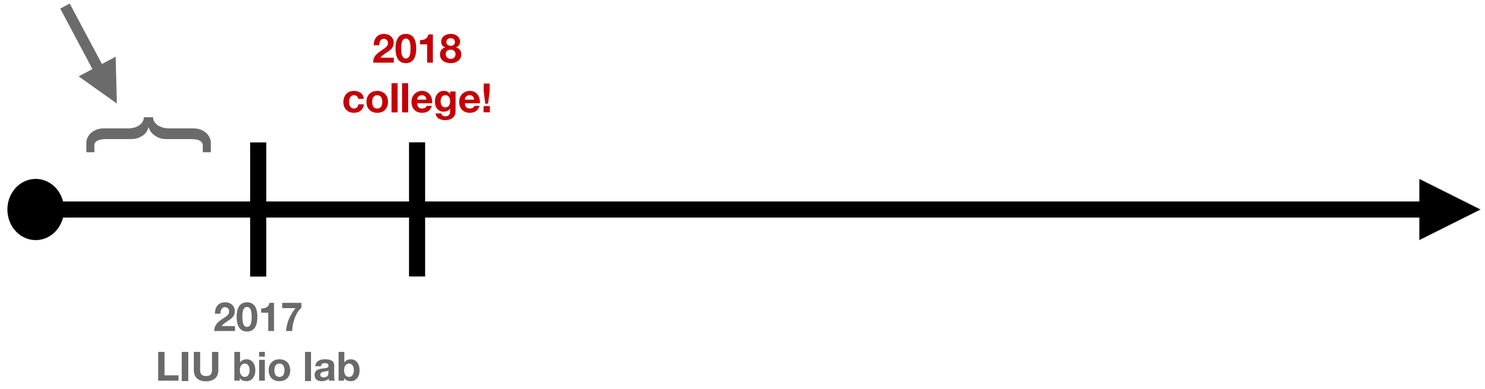
biology era



Try new things, you'll meet people!

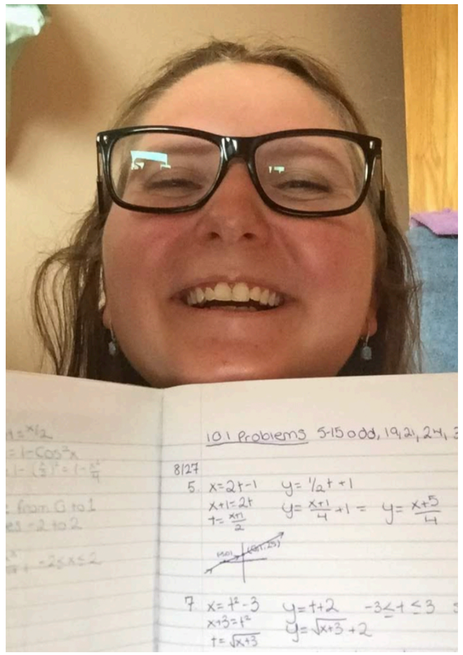
(the origin story)

(not even close to scale)



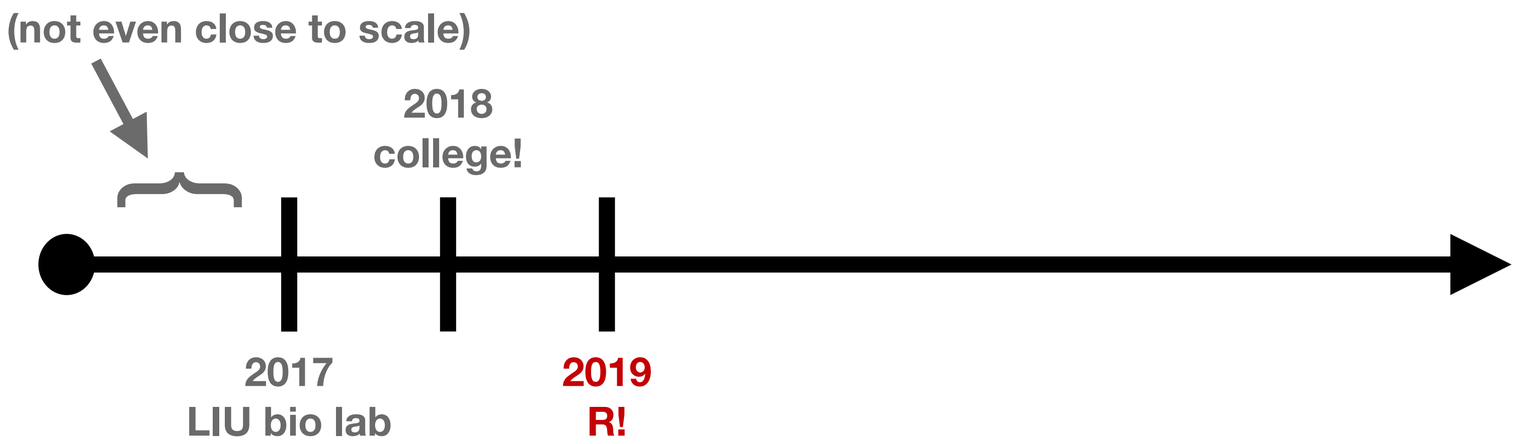
math era

THE UNIVERSITY OF SCRANTON
A JESUIT UNIVERSITY



Try new things, you'll meet people!

(the origin story)

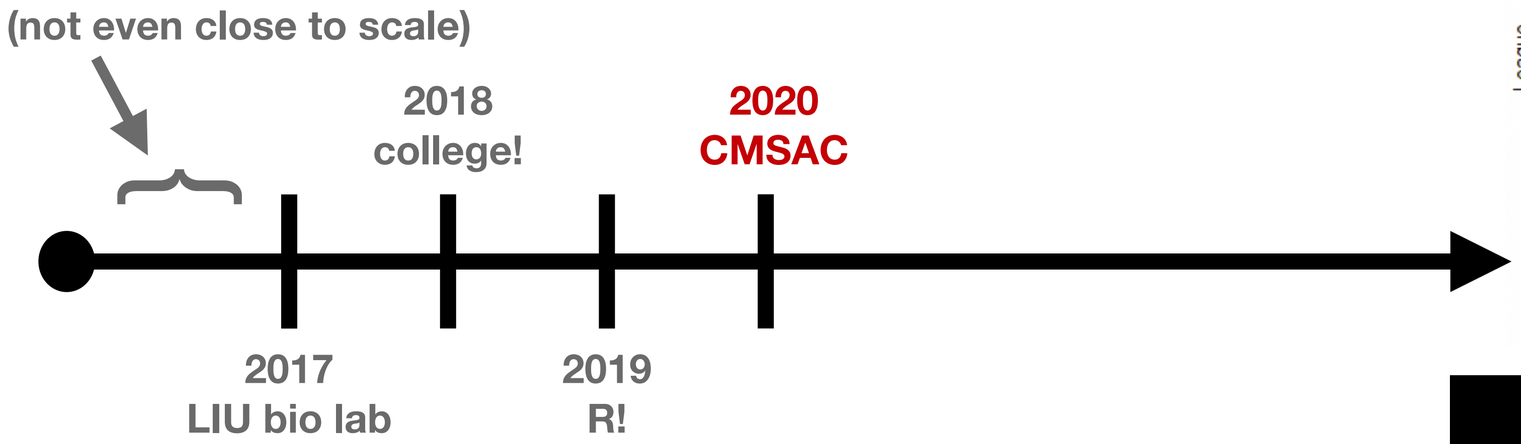


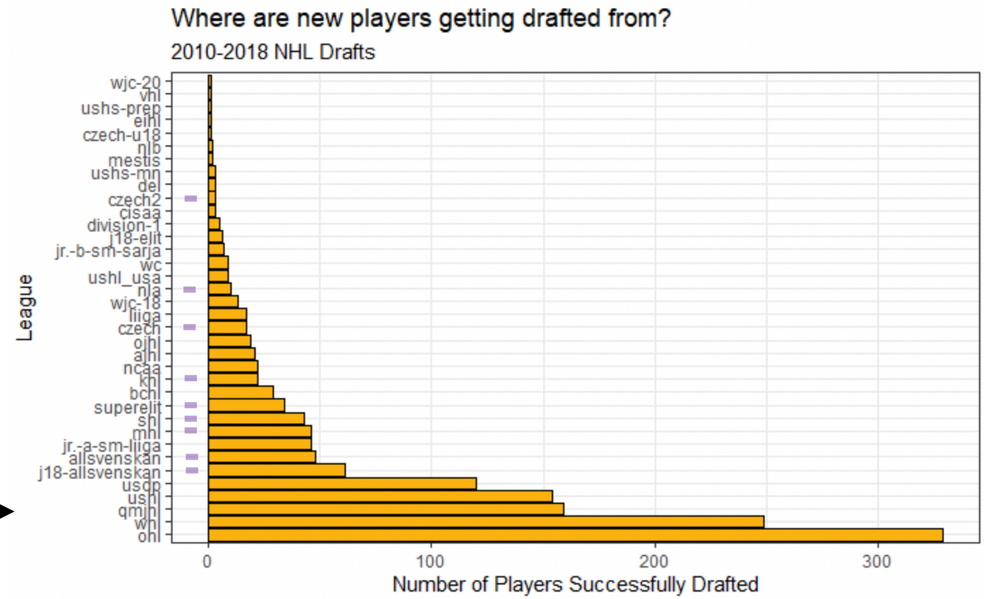
panic era

Try new things, you'll meet people!

(the origin story)

sports era



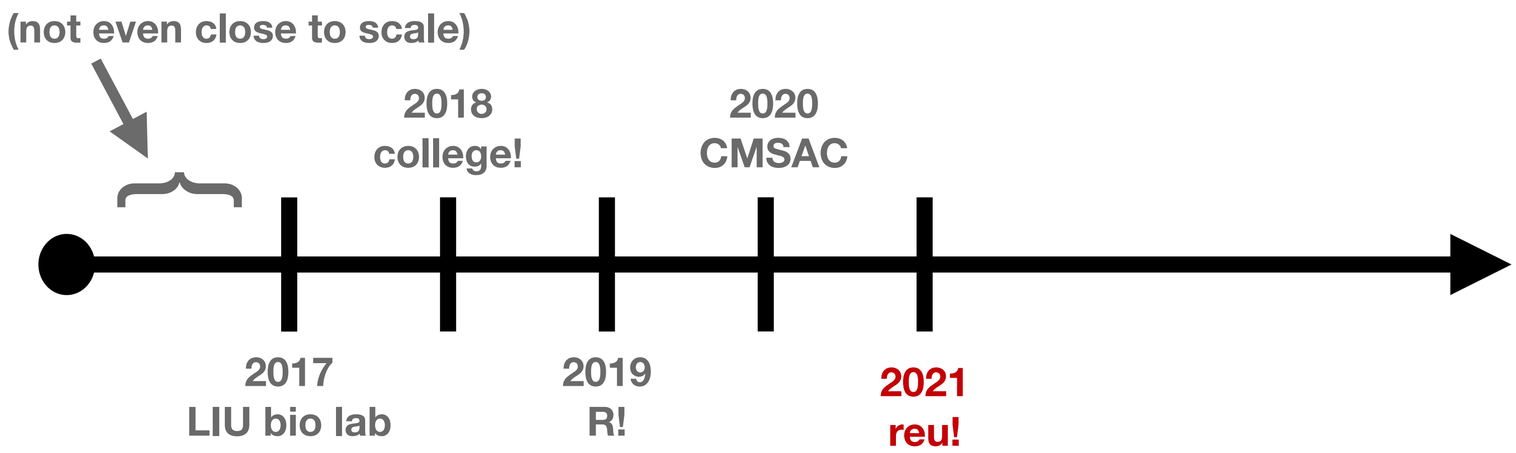


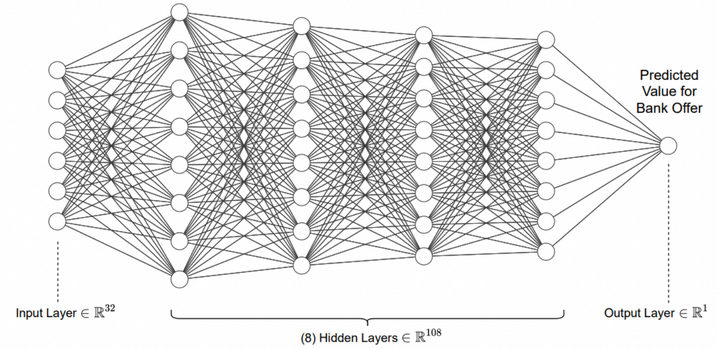
A Puck Above the Rest:
Exploring the Effects of New Data on 2020 NHL Draft Decisions

Ashley Mullan and Lucy Ward
October 25, 2020
Advisors: S. Ventura, N. Citrone, R. Yurko

Try new things, you'll meet people!

(the origin story)

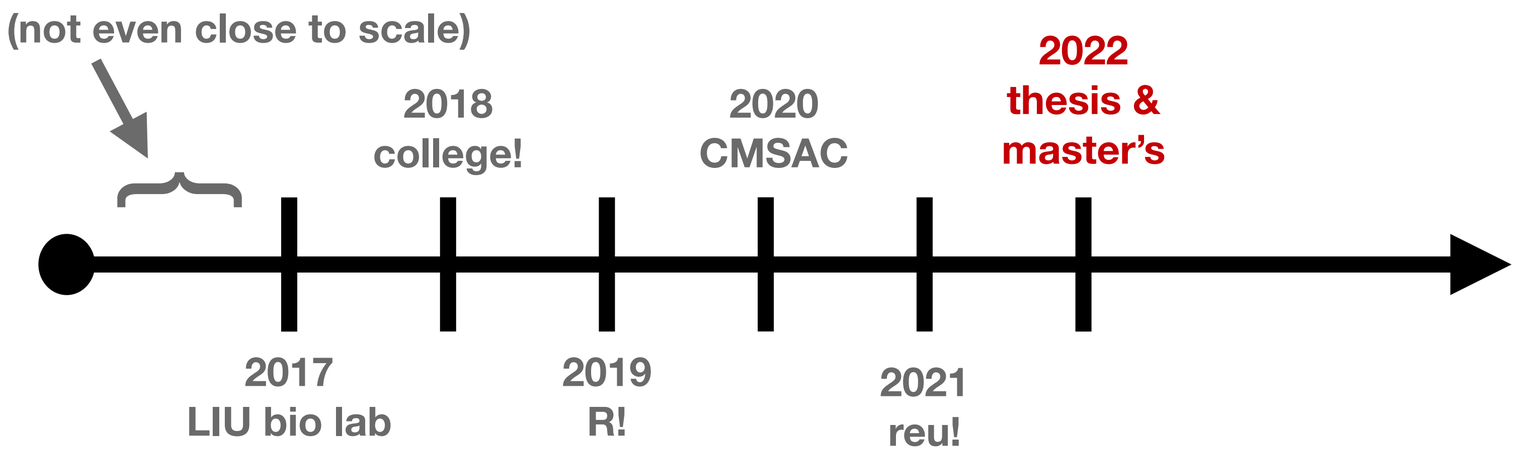




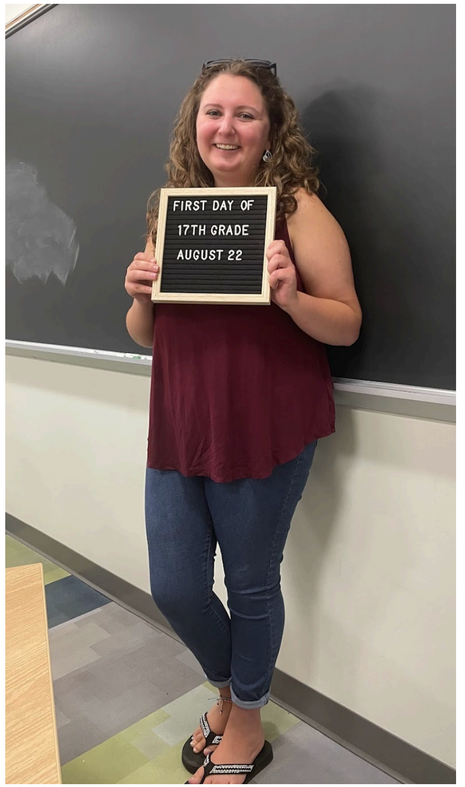
game show era

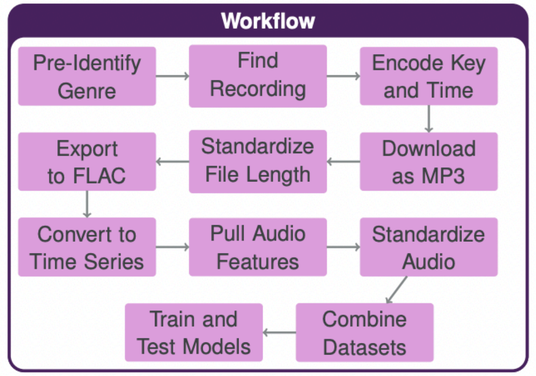
Try new things, you'll meet people!

(the origin story)



first grad school era

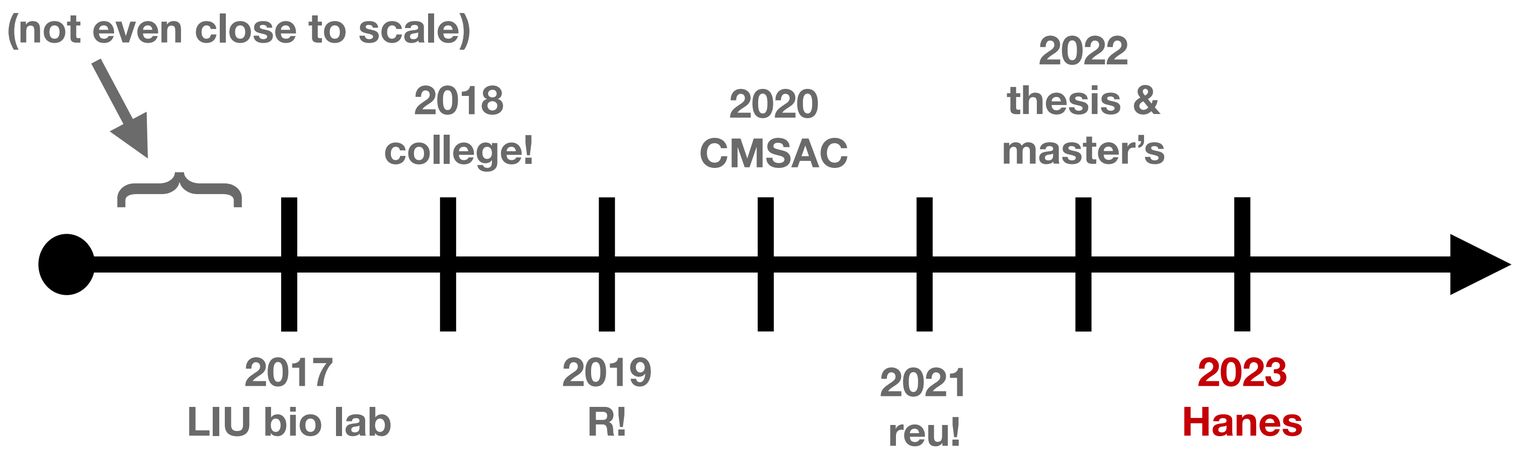




spotify era

Try new things, you'll meet people!

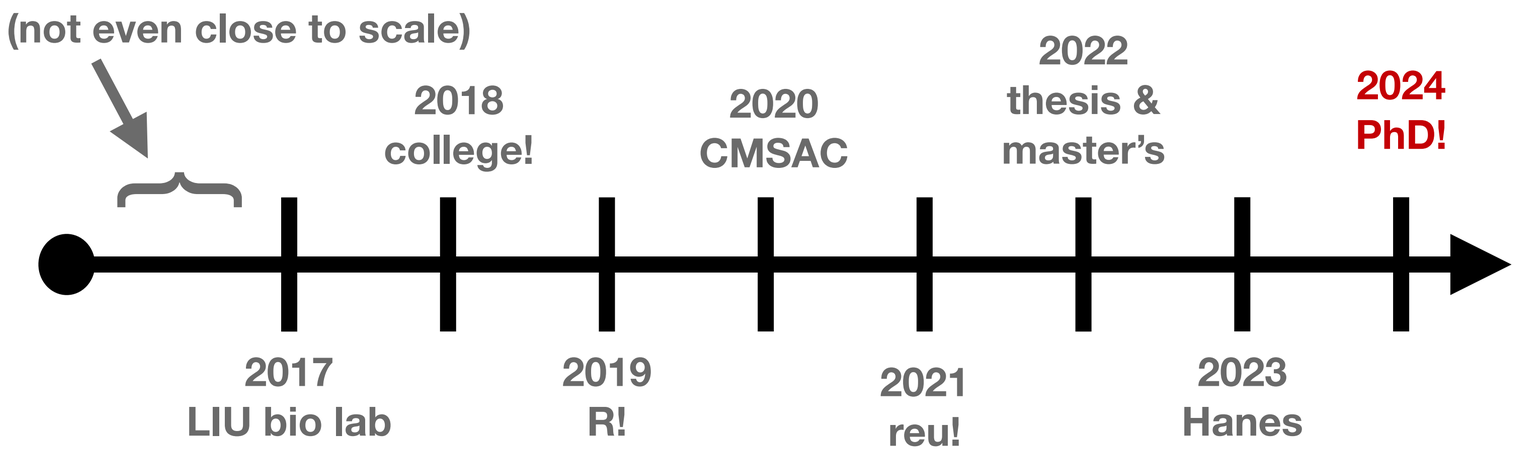
(the origin story)



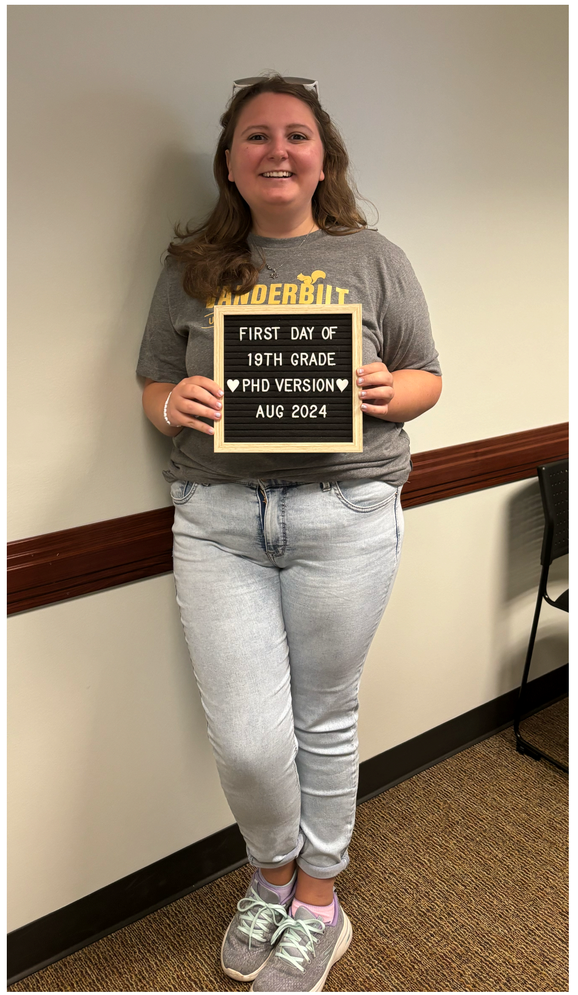
corporate era

Try new things, you'll meet people!

(the origin story)



grad school era
(round two)



Here are some of my hot takes about statistics.

Note: This might be a little overdramatic, but bear with me.

“Anyone who says ~on average~ might be trying to sneak something past you.”

-Me, literally any time I read something resembling statistics in a Buzzfeed article.

What do I mean by that?

(but first, a few definitions)

A **measure of central tendency** is a way to describe your data. Here are some examples!

- The **mean** is the average value in your data. Add up all of your observations and divide by your sample size to find it.
- The **median** is the 50th percentile of your data. Exactly half of the points are less than it, and half are greater than it.
- The **mode** is the value that appears the **most** in your data.

Let’s practice. Grab your pencils! ✏️

Suppose we ask five students how many pencils they have in their backpacks. We get the following data, and we want to compute the mean, median, and mode.

3, 0, 1, 3, 8

That’s me, I’m a walking Staples!

That’s lame, come to class prepared!

Let’s practice. Grab your pencils! ✏️

Suppose we ask five students how many pencils they have in their backpacks. We get the following data, and we want to compute the mean, median, and mode.

0, 1, 3, 3, 8

Let’s practice. Grab your pencils! ✏️

Suppose we ask five students how many pencils they have in their backpacks. We get the following data, and we want to compute the mean, median, and mode.

Step 2: The Mean
Sum up your data!
Then, divide by 5
for the 5 students.

$$\frac{0+1+3+3+8}{5} = \frac{15}{5} = 3$$

Let’s practice. Grab your pencils! ✏️

Suppose we ask five students how many pencils they have in their backpacks. We get the following data, and we want to compute the mean, median, and mode.

0, 1, 3, 3, 8

Let’s practice. Grab your pencils! ✏️

Suppose we ask five students how many pencils they have in their backpacks. We get the following data, and we want to compute the mean, median, and mode.

Step 3: The Median
Work from the edges
(this is why we sorted!)
to find your median.

~~0~~, 1, 3, 3, ~~8~~

Let’s practice. Grab your pencils! ✏️

Suppose we ask five students how many pencils they have in their backpacks. We get the following data, and we want to compute the mean, median, and mode.

Step 3: The Median
Work from the edges
(this is why we sorted!)
to find your median.

~~0~~, ~~1~~, 3, ~~3~~, ~~8~~

Let’s practice. Grab your pencils! ✏️

Suppose we ask five students how many pencils they have in their backpacks. We get the following data, and we want to compute the mean, median, and mode.

Step 3: The Median
Work from the edges
(this is why we sorted!)
to find your median.

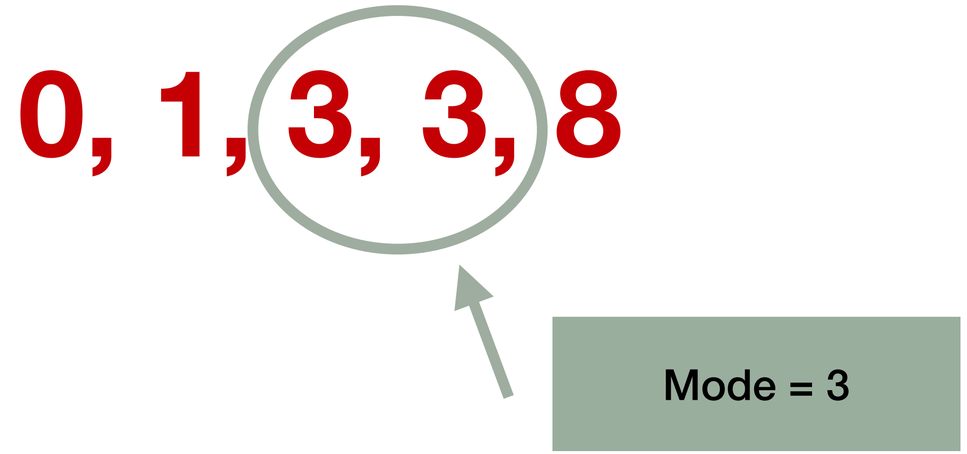
~~0~~, ~~1~~, 3, ~~3~~, ~~8~~

Found it!

Let’s practice. Grab your pencils! ✏️

Suppose we ask five students how many pencils they have in their backpacks. We get the following data, and we want to compute the mean, median, and **mode**.

Step 4: The Mode
Look for any duplicates
to find your mode!

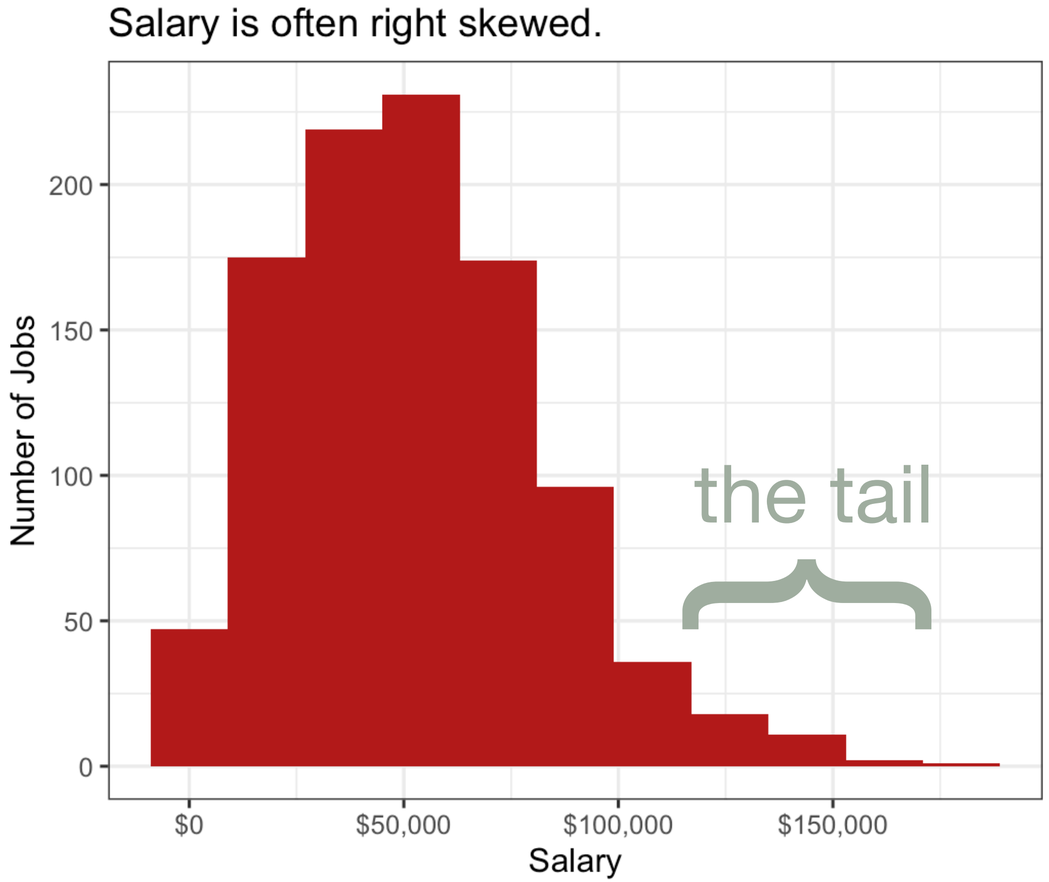


Huh, was that a coincidence? 🤔

- Did you notice that the mean, median, and mode all matched?
- We just showed that different measures of central tendency **can match**, but there are some cases where they don't. That's where the hot take comes in!
- If the mean and median are **similar**, you're looking at **symmetric** data. That means that you don't have way too many "big" data points or way too many "small" data points.
- If the mean and median are **drastically different**, you may want to take a closer look at your data. You could be dealing with **skew** or an **outlier**!

But Ashley, what are those?

- A **skewed** dataset has a tail. There are more extreme values in one direction than you'd expect!
- An **outlier** is a point that doesn't look like the others. It may be much bigger or much smaller!
- We'll see an example of an outlier later and explore how it affects the reporting decisions you make.



How am I supposed to know which measure to use? 🤯

Don’t worry!

We’ll play around with this in Part 3.

(Being a good statistician is all about asking these kinds of questions, great work!)

“Although our intellect always longs for clarity and certainty, our nature often finds uncertainty fascinating.”

-Carl von Clausewitz

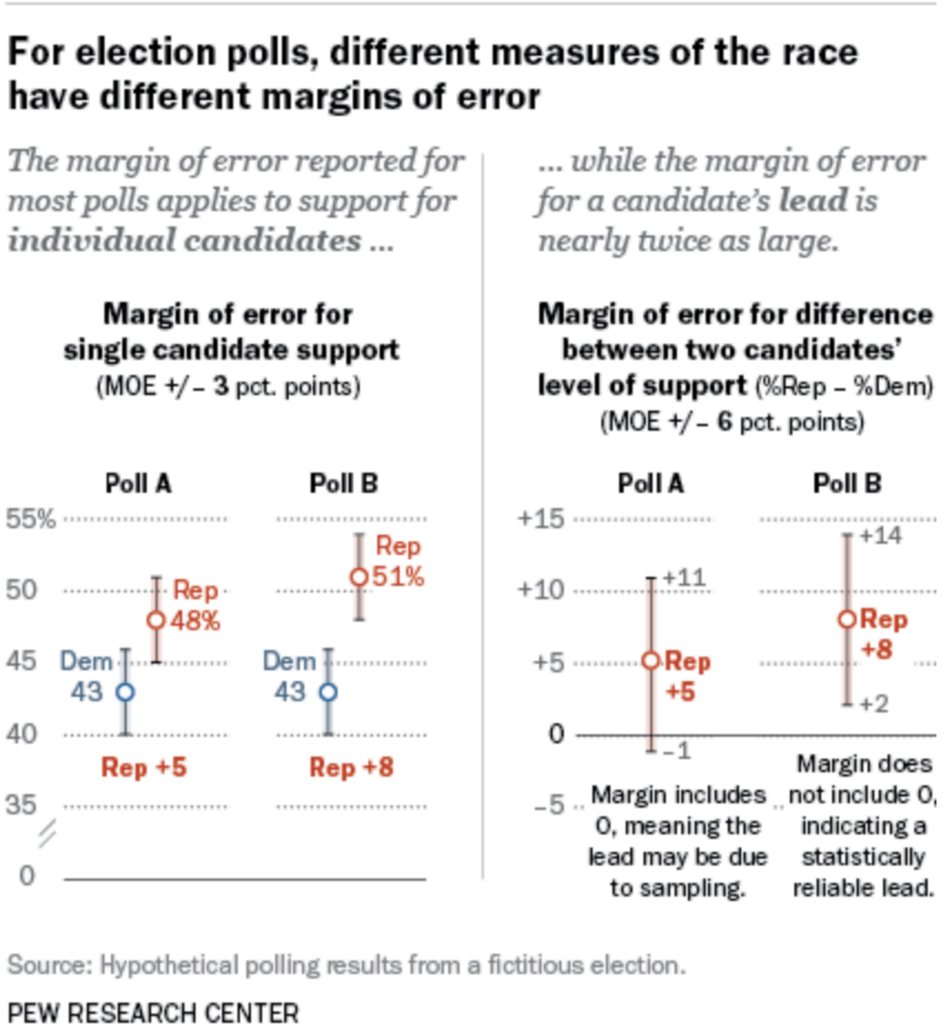
Carl von Clausewitz, painted by Wilhelm Wach.
Thanks, Wikipedia!

Prussian generals do statistics? 💣

- Obviously, we want our statistical models to be correct!
- When we try to **predict** unknown quantities or make **inferences** about a larger group (a **population**) from a smaller group (a **sample**), we will never be 100% right all the time. 😢
- Statisticians have many tools to quantify their **uncertainty**. In other words, we often give both an **estimate** for a value and **error bounds** to state how far away from our estimate we think might be a reasonable guess.

This can be pretty helpful IRL…

Political polls often report a margin of error, since it's really hard to either poll every single American or get a fair sample!



https://www.pewresearch.org/short-reads/2316/09/08/understanding-the-margin-of-error-in-election-polls/

It might even help with some of your upcoming work*!

This is how statistical results often appear in a journal article.

*You might see this article sometime next week; I won't tell James if you won't! (but who do you think sent me the article lol)

Excerpt from Zhdanova et. al (1998)

Endogenous Melatonin Levels and the Fate of Exogenous Melatonin: Age Effects

dose of melatonin significantly increased circulating melatonin levels after 30 min ($p < .001$). Serum and salivary melatonin measurements showed that mean group AUC (± *SD*) for the younger group was 441.9 ± 121.07 pg/ml h (serum) and 136.4 ± 28.11 pg/ml h (saliva) (Figure 4).

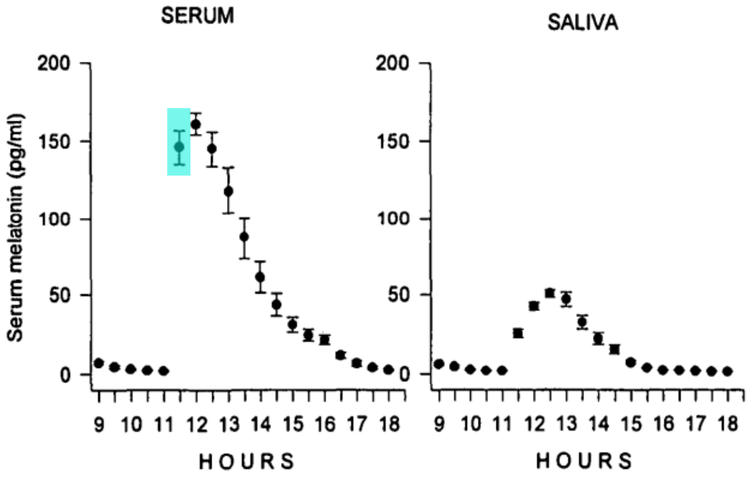


Figure 4. Profiles of serum and saliva melatonin concentrations in the group of younger volunteers (mean ± *SEM; n* = 10). Samples collected prior to and after the ingestion of a 0.3 mg dose of melatonin at 11.00 h.

How can I quantify uncertainty?

- Uncertainty can often come from **variation** either **across** samples or **within** your sample!
- One way to measure variation is with a **standard deviation**. This measures how far (on average) you expect to be from the sample mean!
- Let's go back to the ✏️ example. We had a mean of 3 ✏️ in a student's 🎒.
- I'll skip over the math, but the **standard deviation** ends up being about **3.08**.
- This means that we **expect** to be within about 0 and 6 pencils for most students! Checks out, only the walking Staples exceeded those bounds.

Don’t believe me? Try it yourself!

Hot takes deserve case studies.

1. *How would you describe my reading speed?*

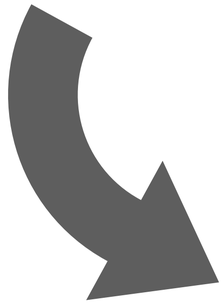
   I provide a list of books I've read this year and how long it took me to finish them. Something in that dataset is a little funny, can you find it?

2. *Which section of statistics would you rather take?*

   Professor A's students have the higher average, but the test scores are all over the place. Professor B's students have a slightly lower average grade, but you know pretty well where the scores fall and think it might be less risky.

https://bit.ly/ash-talks

(except this time, load the app)



Coming Soon to Theaters

1. **Invited Talk**: Misclassification Challenges in Quantifying the Relationship Between Diabetes and Local Access to Healthy Food
   Department of Mathematics, University of Scranton - March 2025

2. **Invited Panel**: Graduate Admissions Advice
   Electronic Undergraduate Student Research Conference
   November 2024

3. **Invited Talk**: For Better or For Worse: The First Kiss Effect on TV Ratings
   International Day of Women in Statistics and Data Science
   October 2024

4. **Guest Lecture**: Stats can be fun, I swear!
   PSYCH1130, Cornell University - September 2024
   Slides - App

Let’s think about those case studies!

(books first)

- Before removing the outliers, where does the mean lie in relation to the median?
- Do the outliers affect the scale of the plot at all?
- How extreme do you think the outliers are?
- Does the median or the mean better represent the center of the data?
- After removing the outliers, where does the mean lie in relation to the median?
- Does this difference really matter? Why or why not?
- What does this tell you about reporting measures of center?

Let’s think about those case studies!

(now the grades)

- Do you have a section preference if you only consider the mean and median? Why?
- Do you have a section preference if you only consider the standard deviation? Why?
- Do these preferences change when you consider everything together?
- BONUS: Are there any other factors that you might want to consider?

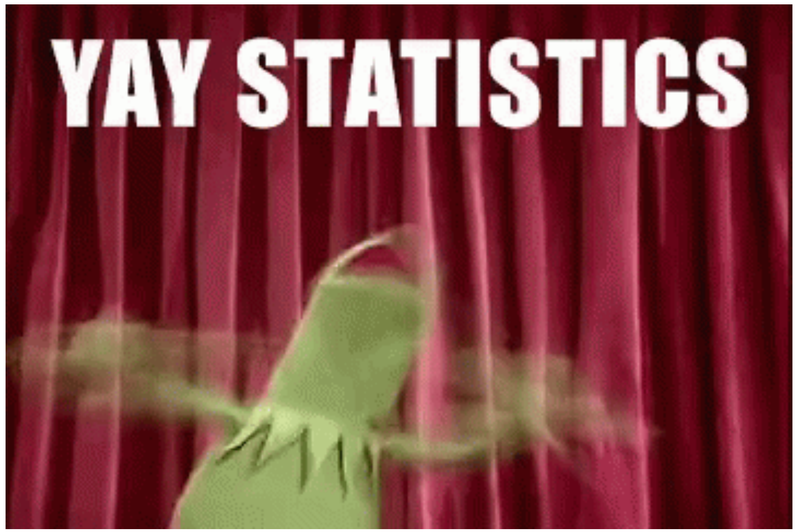
Ask me things!

Takeaways

1. Always check for outliers! The mean and the median work together to give you the most reliable information about your data.

2. Always report some kind of error margin or measure of data variation!

3. Document the data decisions (like how you treat your outliers) when you write up your homework assignments, scientific articles, and business reports so future users can understand and reproduce your work!



Thank you! 😊

If you’re interested in more fun stats takes, check out my website!

ashleymullan.github.io