Set 0: Data Literacy

Skill 0.1: Investigate the importance of data literacy
Skill 0.2: Provide examples of "Garbage in, garbage out"
Skill 0.3: Identify bias in data
Skill 0.4: Explain the benefits of statistics
Skill 0.5: Explore the importance of data visualizations

Skill 0.1: Investigate the importance of data literacy

Skill 0.1 Concepts

Skill 0.1 Exercise 1

It's no secret that data is an incredibly powerful tool. With all that's at stake, it's also not a surprise that understanding a data-driven conclusion can feel overwhelming sometimes – both as an audience member and as an analyst. No matter which side we find ourselves on, data literacy is about how well we read, interpret, and communicate with data.

Data literacy also helps us to produce readable work for other people. As we'll see, even when good data is there, the inability to tell a clear story can have dire consequences.

Let's dive in with some case studies about data literacy triumphs and failures!

Skill 0.2: Provide examples of "Garbage in, garbage out"

Skill 0.2 Concepts

Garbage in, garbage out is a data-world phrase that means "our data-driven conclusions are only as strong, robust, and well-supported as the data behind them."

Example 1

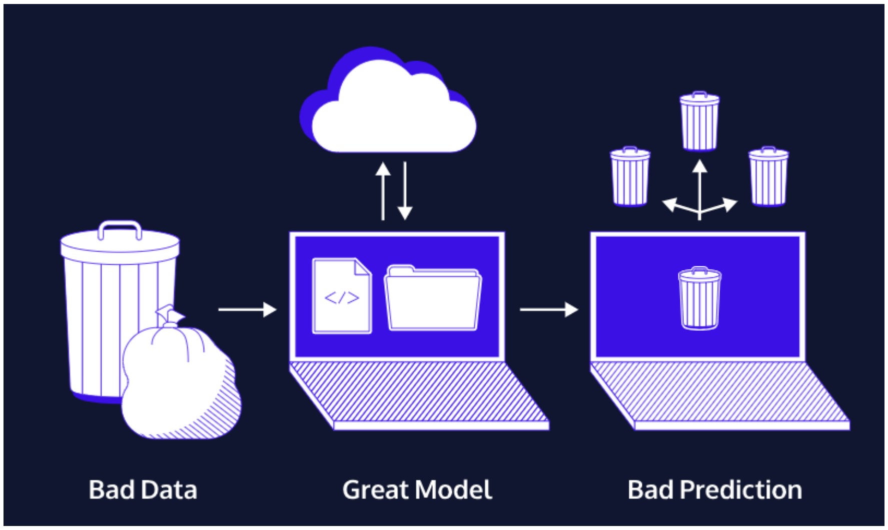
Garbage out: Heart disease is the leading cause of death in women.
Garbage in: As of 2021, women account for only 38% of participants in relevant research studies.

Example 2

Garbage out: 80% of the applicants recommended by a chatbot for a computer science position were white men
Garbage in: The model was trained on historical data of previous applicants that were hired, most of which were white men

Example 3

Garbage out: Don't get vaccinated
Garbage in: vaccines cause autism



The image above is an example of what “garbage in, garbage out” means when it comes to “feeding” a model. Even if we have an excellent mathematical model of a situation, it can only make predictions as good as the data that goes into it. Garbage data will make garbage predictions, no matter how good the model is.

How does data literacy factor in? Part of understanding and communicating with data means asking the right questions so that we end up with useful, relevant data. We can already answer lots of questions about heart attacks, but we won’t learn the ins and outs of women’s heart attacks by studying mostly men.

Part of practicing good data literacy means asking…

Do we have sufficient data to answer the question at hand?
Can my data answer my exact question?

Skill 0.2 Exercise 1

Skill 0.3: Identify bias in data

Skill 0.3 Concepts

One question the data on heart attacks might prompt is “why did the trials have only 38% female participation?”

In part, for historical reasons: in the 1950s, pregnant women in Europe and Canada were prescribed a drug called thalidomide for morning sickness. This drug resulted in severe birth defects and was taken off the market. As a result, in 1977 the US Food and Drug Administration (FDA) recommended excluding from early-stage clinical trials all women who could become pregnant. While intended to protect women, the recommendation put them at risk in a different way, limiting our knowledge of the effects of drugs on women’s bodies.

The FDA reversed these recommendations in the 1990s, and today government-funded clinical trials must include women and other minorities. Yet, the trials don’t need to include minority groups at representative levels, and the majority of drug trials in the US aren’t government-funded anyway.

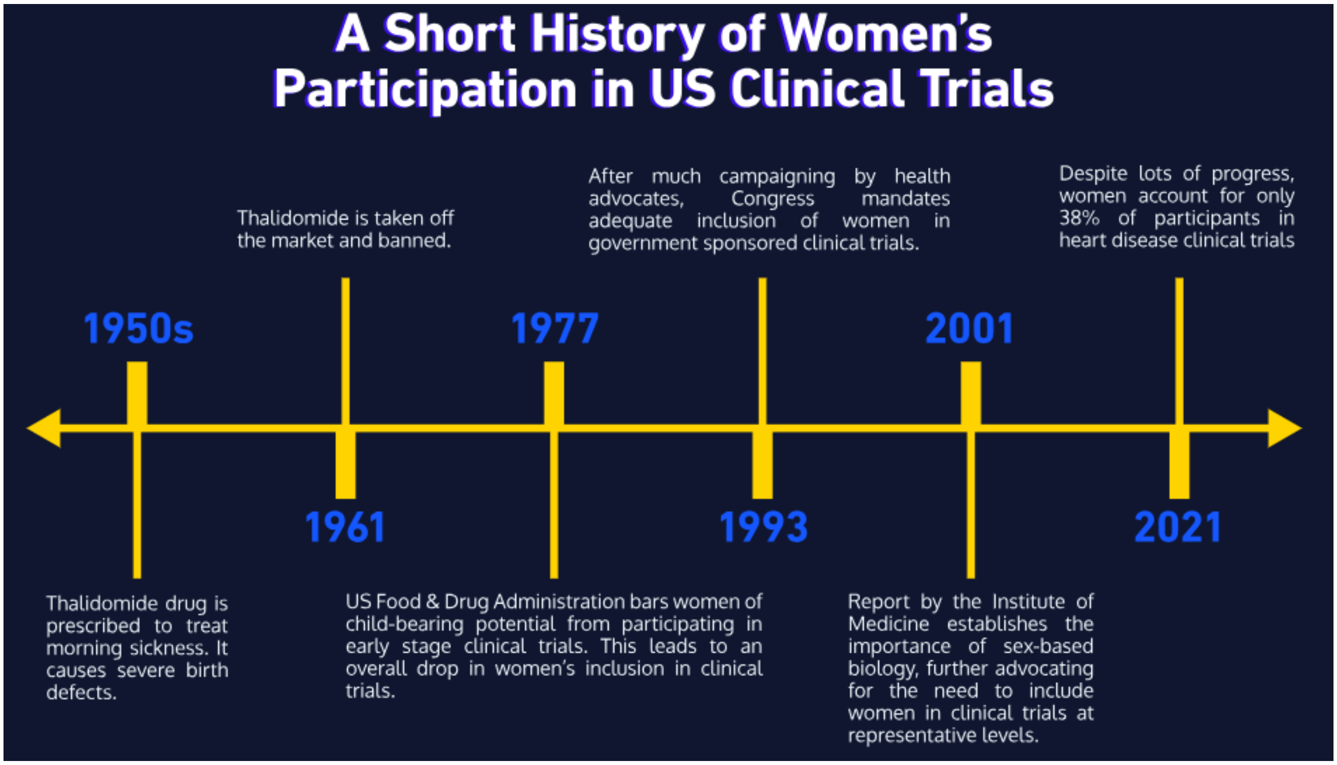
In this case, participation might also be impacted by media representations. In typical TV or movie heart attacks, we almost always see a man clutching at his arm or chest. Not only do women have heart attacks too (we wouldn’t know it from watching TV), they rarely experience chest pain as a symptom.

(In fact, in the top 20 “heart attack” movies* on IMDB, only two heart attacks happen to women: one is fake, and the other is a disguised murder. So… zero real heart attacks in women in a list of top 20 “heart attack” movies!)

It might seem like a stretch from data literacy to TV heart attacks, but sound science means examining bias and controlling variables wherever possible.

Part of practicing good data literacy means asking…

- Who participated in the data?
- Who is left out?
- Who made the data?



A Short History of Women’s Participation in US Clinical Trials

1950s: Thalidomide drug is prescribed to treat morning sickness. It causes severe birth defects.

1961: Thalidomide is taken off the market and banned.

1977: US Food & Drug Administration bars women of child-bearing potential from participating in early stage clinical trials. This leads to an overall drop in women’s inclusion in clinical trials.

1993: After much campaigning by health advocates, Congress mandates adequate inclusion of women in government sponsored clinical trials.

2001: Report by the Institute of Medicine establishes the importance of sex-based biology, further advocating for the need to include women in clinical trials at representative levels.

2021: Despite lots of progress, women account for only 38% of participants in heart disease clinical trials

Skill 0.3 Exercise 1

Skill 0.4: Explain the importance of statistics

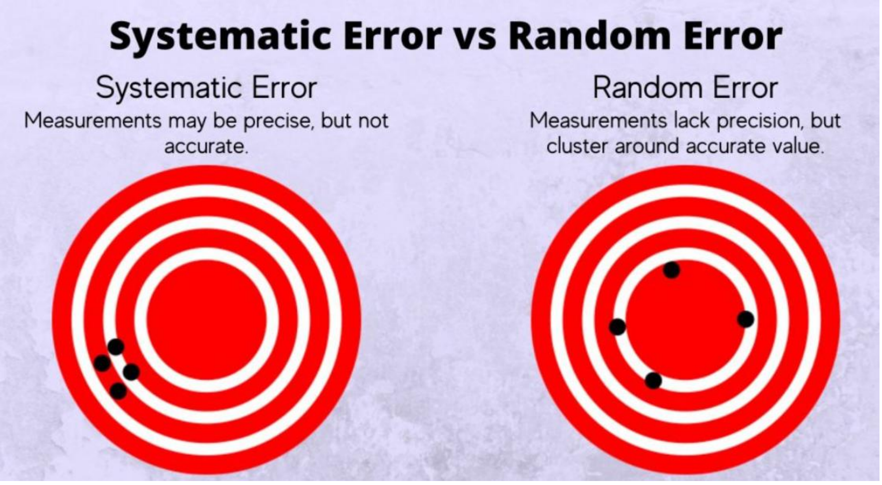
Skill 0.4 Concepts

Big, amorphous injustices like hiring discrimination are hard to prove in court. Hiring discrimination is a pattern of biased behavior towards candidates. This bias results in qualified candidates not getting hired because of their traits.

Throughout the 1900s, companies in the US were able to justify hiring on a case-by-case basis. After all, it’s legal to hire or not hire candidates based in part on soft qualities such as “fit” and “office culture.” But if these qualities are a mask for factors like a candidate’s race, gender, or religion, the company has broken anti-discrimination laws.

Usually, a lawyer would have to show the many individual cases that proved a company was discriminatory. Instead, lawyer Elaine W. Shoben shifted the burden of proof to companies. How was she able to do this with data literacy? She used the power of statistics! **Statistics helps us test the likelihood of an event happening by random chance versus systematically.**

Systematic versus random data can be visualized below,


Systematic Error vs Random Error
Systematic Error
Measurements may be precise, but not accurate.
Random Error
Measurements lack precision, but cluster around accurate value.

So how did Elaine Shoben show that discrimination was at play in hiring decisions using statistics? Below, breaks down the argument,

Step	Logic	Example
Step 1	Could the hiring results have happened by random chance? Or is that statistically impossible?	In the last 5 years, StarComm Corporation had 1,000 candidates and hired 200 people. Of the 1,000 candidates, 400 were women (40%). Of the 200 people hired, only 20 were women (10%).
Step 2	If the hiring results haven't happened by chance, they must have happened by "purposeful exclusion"	Statisticians determined that the probability of getting these hiring results by chance is essentially zero. Lawyers can then conclude that the low number of women hired isn't accidental, but purposeful in some way.
Step 3	If the employer is aware of this "purposeful exclusion", they show "reckless disregard" for the rights of individual candidates not to be discriminated against.	StarComm Corporation is now aware that their hiring practice discriminates against women. So lawyers can argue that SCC violated the rights of women candidates to have a fair shot (without discrimination) in the hiring process.
Step 4	The burden of proof shifts to the employer to prove why hiring requirements are valid and necessary.	The burden is now on StarComm Corporation to get its hiring process into legal shape OR to prove why its hiring process has to be the way it is. It's no longer the job of individual women candidates to prove they are up against an unfair process.

The above conclusion illustrates statistics at work! That's definitely a bit of legal jargon, but how cool is it to use statistics to reveal a systematic pattern of discrimination, rather than trying to piece together a case from individual experiences. That's really what stats is all about.

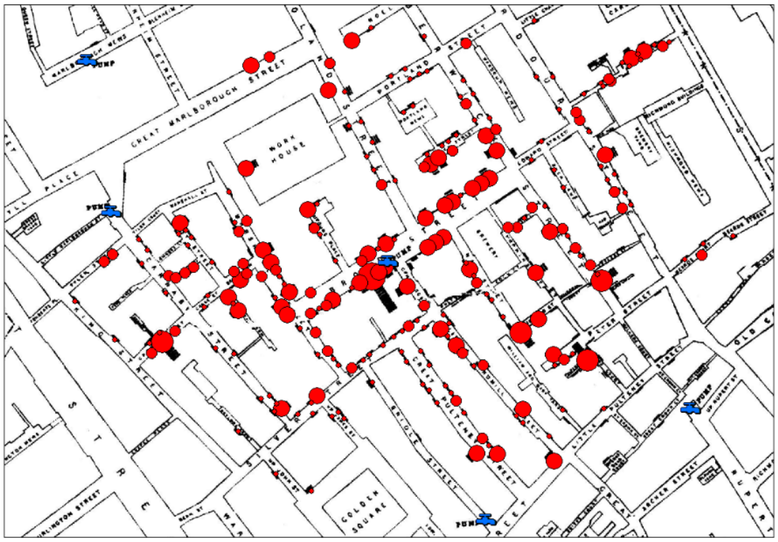
Skill 0.4 Exercise 1

Skill 0.5: Investigate the importance of data visualizations

Skill 0.5 Concepts

We’ve looked at examples of data quality and bias along with how statistics can be used to answer big legal questions. Visualizations are another powerful tool for interacting with data. They help us understand data-driven arguments and are a powerful tool for communication.

A famous data visualization created by John Snow in 1854, enabled Dr. Snow to link a cholera outbreak in London to a contaminated water pump. Up until this time, people believed that cholera was caused by vapors rising from the burial grounds of plague victims from two centuries earlier. (A good try, but cholera is actually a waterborne disease caused by bacteria found in sewage. It causes severe dehydration and has a fatality rate of over 50% when untreated.)



Dr. John Snow’s breakthrough started with how he visualized his data: he organized cholera death records by location rather than by time, which was more common. He made a map, and discovered that the deaths centered around a water pump on Broad Street.

From there, Dr. Snow used death records that seemed to contradict his theory to strengthen his explanation. For instance, a woman who died of cholera in a completely different neighborhood had just visited her aunt’s house near Broad Street and drunk water from the pump.

Dr. Snow also found that a workhouse and a brewery near the pump both had few or no cholera deaths. Upon investigation, he learned that the workhouse had its own water supply, and that the brewers not only had access to a well at the brewery, but that they drank only malt liquor and never visited the Broad Street pump.

Snow advised that the handle be taken off the Broad Street pump to prevent people from drinking the contaminated water. The handle was removed, and this action coincided with the end of that outbreak. The number of deaths was already trailing off (more than 75% of residents had left the area to avoid “choleric vapors”), but this public health intervention prevented the disease from recurring as people returned, and the epidemic ended.

Skill 0.5 Exercise 1