Data Gaps

3 min

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."

For example: we have a lot of data on heart attacks, but there's room for improvement when it comes to data quality. Heart disease is the leading cause of death in women, but as of 2021, women account for only 38% of participants in relevant research studies.

There are key differences between men's and women's heart attacks that impact how they're treated, but our data doesn't yet adequately outline those differences. This leads ultimately to worse outcomes in treatment and a higher post-heart attack mortality rate for women.

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?

Instructions

The image to the right 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.

Imagine having a model that can accurately predict the weather in Rio de Janeiro 90% of the time, but then using data from only United States weather stations as the input values. Or using temperature data that was collected only at noon each day. Or using wind speed data collected only by estimating with a licked finger held up in the breeze. No matter how good the model is, it will need precise and accurate data from Brazil to give a helpful output.

Bad Data   Great Model   Bad Prediction