Andrew – Connecting the Game

Responding to Crawford's Claim About Data

When Kate Crawford claims that data cannot speak for itself, I feel that she is correct, but needs to be clearer about what this truly means. I think she means that data cannot tell its own story, and requires an outside, intelligent source to draw inferences, generalize, and provide proper analysis from a dataset.

To clarify, I believe that data cannot speak for itself because without someone to not only analyze data but also to collect data, the numbers would have no interpretation and they would just simply exist in our world. In addition, even though all datasets have a pattern which derives from our society, these patterns only present a risk from misinterpretation of their results which we can draw from analysis. I find that the following considerations provide a compelling argument to support my claim.

> Premise 1. Without someone to handle data, it seemingly exists in our world just waiting to be collected and analyzed.
>
> Premise 2. All datasets are naturally linked to human culture and society, and hidden biases present high risks.
>
> Premise 3. Depending on the person who collects and analyzes this data and the method they use, their interpretation might provide different results than someone else who collects and analyzes the same data.
>
> -----
>
> Conclusion 1. So, data is not objective. This means that a dataset can have different meanings to different people who collect and analyze it.
>
> -----
>
> Conclusion 2. Therefore, data cannot speak for itself and those who analyze it risk misunderstanding the generalizations they draw from it and, in turn, can misallocate resources.

Consequently, I feel that the philosophy of feminist epistemologists is heavily present in my argument. Data cannot control its biases and is often organized in ways that humans construct it. As mentioned in Premises 1 and 2, data "floats" throughout everyday life just waiting to be collected and given reason to exist. Much like how Catherine D'Ignazio and Catherine Klein say in their article that we must care about where our data comes from, it is important to note that data is a human construct. In the end, we choose where and how to collect datasets, whether it is through methods like experimentation, random sampling, or surveying, or even if we choose to use datasets that other people generated. Two more pertinent takeaways from D'Ignazio and Klein's article on data feminism are that "all knowledge is 'situated'…this means that context matters" and "when approaching new knowledge…it's essential to ask about the…conditions in which that knowledge was produced, as well as the identities of the humans who made that knowledge."[1] As data scientists, we should be constantly checking to make sure our results account for everyone who the data has involved, and that there is a history behind every number and statistic in a dataset. This history will reveal the hidden biases within data that it cannot control, such as why data has skewed characteristics, which are all intrinsically connected with society.

Nevertheless, someone might claim that Conclusion 1 poses a target of objection. The claim that data is not objective and that its results can be interpreted differently by different people can be legitimately challenged. According to the philosophy of anti-realists, specific scientific phenomena and patterns will occur because of reasons that humans cannot experience within our sensations. In this case, interpretation would not matter because there would only be

[1] D'Ignazio, Catherine, and Lauren Klein. "Chapter Five: The Numbers Don't Speak for Themselves." *MIT Press Open - Data Feminism*, January 15, 2019, 5.

one true reason embedded in our world as to why this result happens. Objectivity would not be a problem. A scientific example of this would be trying to explain why a ball falls to the ground when we throw it up in the air. No matter who you are and how you interpret this event, the reason the ball falls to the ground is because of gravity. From here, it would be recommended to not venture in to wondering why gravity occurs, and instead just accept that gravity is present in our world. Moreover, a data analytics example of this would be trying to explain if the relationship between two variables is significant. If two people used the same data for each and the same significance level, they could use general statistical calculations to determine if the explanatory variable is significant. These same two people would get the same result and should not have to worry about the impacts of different significance levels or what factors drive significance in data. This counterargument intends to show that experience is the only true form of knowledge and theorizing past our barrier of sensations, like asking why a ball falls to the ground or why significance occurs, is looking incorrectly into what makes science and data so important. Therefore, datasets can only have one meaning in the numbers they present and should not be interpreted differently by different people, since the reason for their results are the same.

In further response, this time to the paragraph above, I reject the objection. To take a more naturalistic standpoint, I want to ask: why should we not take into consideration that different significance levels might affect how we handle our data? If the relationship between two variables equates to a statistical measure of a probability-value of 0.02, we would reject our null hypothesis on a 0.05 significance level but accept our null hypothesis on a 0.01 significance level. In this scenario, we risk making decisions such as false negatives or false positives, which could harm the people impacted by our statistical claims. This is the problem with two people interpreting data differently. Since data itself cannot tell us to determine the significance of the relationship between variable X and variable Y using only a 0.01 threshold to get the output which will benefit society, it is extremely important for us data scientists to figure out what factors go into choosing significance levels. As for objectivity in data, we should want to be skeptical about how data is explained because different people will use different proxies. This technique of using proxies is very valuable in statistics, but it is necessary to ask where the proxy was derived from. Cathy O'Neil notes that "proxies vary in strength" and "not everything is measurable," so if someone makes a believable claim about a dataset using statistical measures on non-measurable data, people can easily be misled.[2] They will, in turn, not understand the truth behind the numbers, which the numbers themselves cannot warn us about.

All in all, I feel that Crawford's initial claim about data is an insightful take which opens a philosophical hotspot of discussion connected to the current state of data analytics. I have found that the most pressing issue surrounding data is context. All discussions about how to truly interpret data, how and where it was collected, what methods were used, and what to make of analytical results are all relevant. This big data era that we live in can cause us to easily forget where we are getting our data from and what it all means. It is crucial to take a step back when reading our data and ensure that the decisions we make on how to present it are ones that have positive outcomes.

[2] O'Neil, Cathy. "On Being a Data Skeptic," October 7, 2013, 5-6.