When Data Can Speak for Itself*

Russell Luchin

March 13, 2023

Data cleaning can be considered telling a story with data. When we engage with data we are not usually engaging with "raw" data, we're engaging with data that someone has processed and neatly plotted out to tell a story. With this, we can consider data cleaning the act of taking data and telling a story with it, rather than the data "speaking for itself". A data scientist becomes a story teller in this framework, and the data is simply the means to convey it.

As we've established data cleaning is storytelling, for data to speak for itself it must be in its raw, untouched form. This is possible, but rather unlikely. For example, if I look at the raw data for how long the runners of a marathon took to finish, all I'm seeing is a bunch of different times. Yes, if we abstract it we can argue that while looking at the data we are visualizing how it is plotted, and with the data is speaking for itself. This would just be impromptu data cleaning, however. Because of that the data isn't really speaking for itself, is it?

If we abstract further we can look to language. Language, based on the consensus of linguists (CITE), is a series of phonemes conveyed through phonetics (sounds) that we interpret into meaning. In a sense language is raw data speaking for itself. This argument is quite frankly ridiculous, however, and with enough abstraction we can create a framework where everything is data and everything is data speaking for itself. This is unproductive and does not tackle the question: to what extent should we let data speak for themselves?

I would argue that data, as we know it, is never speaking for itself. As Au Au (2020) discusses, data cleaning is discerning the "noise" from the useful signals. We are never engaging with the raw, unfiltered data in research; we're engaging with the cleaned version that can be used to tell a coherent story. Because the data we engage with is cleaned, are we ever engaging with the data alone? I would argue no, as the data was cleaned by a scientist, not magically

*Thanks to R Core Team or making the writing of this paper possible. No peer reviewer was available for this paper. Credit to Rohan Alexander for teaching the course which allowed me to form the arguments presented here. GitHub repo available at https://github.com/rluchin/tutorial9/

by itself. We're engaging with the work, and the story, of the scientist; to this end, the data cannot speak for themselves.

This takes me to my main point; I don't believe there is any way for data to "speak for itself" in an impactful way. It isn't a question of when data should speak for itself, but rather if it can at all. I argue that it cannot, as the majority of data science work is cleaning Au (2020), its not unreasonable to conclude that the data we engage with are stories told by the scientists cleaning it.

This isn't to say that data should be dismissed, not in the slightest. Rather, its a way to say that we should hold data scientists to a similar standard that we do journalists or authors. As discussed in D'Ignazio and Klein (2020), the majority of data scientists are caucasian men. They are also people, with beliefs and biases that can skew their perception of what is "good" practice in data. Like words, data can be easily manipulated to tell a story which may not be true in its entirety; such as when Purdue used data indicating OxyContin wasn't addictive to sell it to doctors (LaPerriere 2021). The doctors trusted the data to speak for itself, when in reality they were trusting the data scientist that prepared the data.

In conclusion, data can never speak for itself on a practical level. Sure, we can abstract the argument in ways that shows how raw, manipulated data tells its own story, but in reality what we consider to be "data" is actually cleaned data thats been prepared by a scientist. Data should never be trusted without scrutiny, and to let data "speak for itself" is to put blind faith into data scientists to never be wrong; to be blunt, it is a suboptimal practice which can lead to suboptimal outcomes (example: the opioid epidemic).

References

Au, Randy. 2020. "Data Cleaning IS Analysis, Not Grunt Work."

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism.* The MIT Press. https://data-feminism.mitpress.mit.edu/.

LaPerriere, Christy. 2021. "A HISTORY OF DANGEROUSLY MISLEADING DATA VISUALIZATION."