To What Extent Should We Let the Data Speak for Themselves?*

Bolin Shen

February 11, 2024

As a student majoring statistic, I need to learn how to analyze data. To what extent should we let the data speak for themselves is a pivotal question. To answer the question, I will draw insights from Jordan (2019), D'Ignazio and Klein (2020, chap. 6), and Au (2020).

Jordan (2019) emphasizes the need for broadly understanding artificial intelligence (AI). He is against the prevailing focus on human-imitative AI, urging for attention to Intelligence Augmentation (IA) and Intelligent Infrastructure (II). Jordan reveals that letting data speak for itself is insufficient through the case of his spouse. Instead, he advocates for an approach that considers human values, social sciences, and humanities. The concept of letting data speak for itself is at odds with Jordan's vision, as he emphasizes the importance of combining principles and perspectives beyond raw data. D'Ignazio and Klein (2020), in Chapter 6 of "Data Feminism," provide a critical analysis on letting data speak for itself. They highlight cases where relying solely on data, without considering context, can lead to misleading interpretations. The concept of "Big Dick Data" emphasizes their criticism of projects that prioritize size over contextual understanding. They stress the necessity of context in data analysis, challenging the idea that numbers can speak for themselves (D'Ignazio and Klein 2020). They assert the contextual nature of knowledge and emphasize the importance of theory and context in meaningful data interpretation. Randy Au (2020) challenges the perception of data cleaning as grunt work. While not explicitly addressing the issue of letting data speak for itself, Au's point is consistent with the idea that data should be actively shaped and analyzed. He sees data cleansing as an integral part of data analytics, highlighting its role in building transformations. Au's stance means that without proper cleaning and shaping, data cannot effectively speak for itself. This paper encourages practitioners to consider data cleansing as a critical and analytical aspect of their work, emphasizing the need for data to be actively engaged in meaningful analysis.

Although the idea of letting data speak for itself might seem objective and scientific as discussed in these works, it does have limitations. To start with, data in its raw form lacks the ability to

*this analysis is available at: https://github.com/Brian031205/Data-Speak-for-Themselves

convey meaning and insights without proper context, interpretation, and ethical considerations. In terms of Context Matters, "Big Dick Data" was criticized to show the importance of context in data analysis (D'Ignazio and Klein 2020). Context is not just an additional layer but a crucial factor that shapes the meaning of the data. Without context, data might easily be misinterpreted, as demonstrated by the flawed reporting of kidnapping rates in the case of the Nigerian schoolgirls (D'Ignazio and Klein 2020). Without context, the data analysis might be at risk of having bias and harmful stereotypes. Secondly, in terms of Human Values and principles, emphasizing the need for principles and diverse perspectives is of crucial importance in the development of AI systems. This implies that data, to be meaningful and beneficial, must be satisfied with human values and ethical considerations (Jorden 2019). Letting data speak for itself neglects the importance of incorporating these broader principles into the analysis. Lastly, regarding Data Cleaning as Analytical Work, the perception that data cleaning is menial work is being challenged and, by extension, it is supported that data analysis is an active, iterative process (Au 2020). If data were to speak for itself, there would be no need for the process of cleaning and shaping it. Au's argument reinforces the notion that meaningful analysis involves careful decisions at every step, from data collection to interpretation.

Given these views, it's clear that raw data cannot provide a complete representation of reality. Thus, letting the data speak for itself is an incomplete and potentially misleading approach. Jordan (2019) advocates a framework of principles, D 'Ignazio and Klein emphasize (2020) points out the importance of context, and Au (2020) emphasizes the importance of proactive data shaping. Overall, these views advocate a more thoughtful and engaged analysis. It's clear that analysts play a vital role in turning data into actionable insights. This role involves understanding context, applying a principled framework, and actively shaping data through processes such as cleansing. In the age of big data, it is necessary to go beyond the simple concept of letting the data speak for itself. Instead, we should take a more critical and principled approach to data analysis. By doing so, we can achieve meaningful insights from data while avoiding biases and misunderstanding.

Reference:

Jordan, Michael. 2019. "Artificial Intelligence–The Revolution Hasn't Happened Yet." Harvard Data Science Review 1 (1). https://doi.org/10.1162/99608f92.f06c6e61.

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

Au, Randy. 2020. "Data Cleaning IS Analysis, Not Grunt Work," September. https://counting.substack.com/p/data-cleaning-is-analysis-not-grunt.

(I extend thanks to my classmates, Heng Ma, as he provided useful suggestions for this mini essay.)