The Limits of Letting Data Speak for Themselves*

Zixuan Yang

February 13, 2024

1 Introduction

The advent of big data and powerful analytical algorithms has prompted some to suggest we should simply "let the data speak for themselves" to produce objective insights. However, a deeper look at data science reveals it inevitably involves human judgment and values. Data do not arise in a vacuum nor interpret themselves. Rather, responsible data analysis requires actively shaping data collection, cleaning, analysis and interpretation to serve society ethically. Several thinkers make compelling arguments against over-reliance on "letting data speak for themselves" without human guidance. Michael Jordan argues for developing principles to ensure data-driven systems enhance human life safely. Catherine D'Ignazio and Lauren Klein contend data science should empower marginalized people, not simply reinforce the status quo. Randy Au asserts data cleaning imposes judgments and assumptions upon data. Overall, while data provide key inputs and reveal patterns, diverse human oversight is essential to direct data science towards justice and avoid flawed, harmful outcomes. This essay will examine arguments against the notion that we should just "let the data speak for themselves."

In his article on data cleaning, (Au 2020) argues that the process of cleaning data inherently involves human judgment and values. The choices made in cleaning data allow chosen analysis algorithms to function and produce the desired results. By transforming the data in certain ways, whether removing anomalies or normalizing formats, the data cleaner imposes interpretations and assumptions on the data. Data cleaning is not a purely mechanical process, but reflects goals and biases.

(D'Ignazio and Klein 2020) similarly contend that data never simply "speak for themselves" in an objective manner. The choices made throughout data science, from collection and cleaning to analysis and interpretation, unavoidably reflect the perspectives and interests of those

*A GitHub Repository containing all data, Quartro file , and other files used in this investigation is located here: https://github.com/ZixuanYangFrank/miniessay6

involved. D'Ignazio and Klein advocate that the practice of data science should be guided by a goal of serving justice and empowering marginalized populations, not merely reinforcing dominant paradigms and the status quo.

(Jordan 2019) argues that society is currently building large-scale decision systems that integrate data, algorithms and physical entities. He contends principles are needed to thoughtfully design these societal systems in order to ensure they enhance human life in a safe, ethical manner. Leaving this responsibility to "data speaking for themselves" without human oversight and guidance invites major conceptual flaws and harms.

Jordan provides the example of the Large Hadron Collider to illustrate that even just "letting the data speak" involves significant human judgments, in determining how data will be filtered, selected and stored for analysis. These choices are guided by accumulated knowledge within physics about which data aspects are considered meaningful. Raw data alone does not produce scientific insights or truths.

Furthermore, analysis algorithms themselves cannot interpret raw data directly. For algorithms to function, the data must first be formatted, cleaned, structured and often reduced based on human goals, values, assumptions and knowledge. Data alone does not "speak" in a way algorithms can utilize without this human-driven preparation. Interpreting analytical results also requires human domain experts who deeply understand nuances in the data and complex interactions in the broader systems that generated it. Surface patterns in data alone do not speak the deeper truth.

Responsible institutions like the United States Census Bureau invest tremendous resources into meticulously documenting data collection methodologies and contexts. However, most real-world data lacks this level of documentation and context about its origins and biases. Therefore, data alone has limited interpretability without the involvement of knowledgeable humans.

Lastly, Jordan argues that for data insights to responsibly inform major decisions and policies impacting human lives, society must go beyond data to consider critical issues like uncertainty, causality, and long-term impacts using ethics and diverse human perspectives. Data alone is insufficient for guiding just and wise decision-making.

In summary, these three perspectives provide compelling arguments that while data can provide key inputs and reveal patterns, human judgment is absolutely essential to actively shape how data science is ethically applied to serve society. Expecting data alone to somehow objectively "speak for themselves" without diverse, thoughtful human guidance inevitably invites flawed, biased, and potentially harmful outcomes. Responsible data science requires human values, oversight and accountability.

2 Conclusion

In conclusion, while data can reveal valuable patterns and insights, it is dangerously naive to expect data alone to "speak for themselves" objectively without active human oversight. The arguments examined above comprehensively contend human judgment is absolutely essential to ethically shape data collection, cleaning, analysis and interpretation in service of society. If we passively leave this responsibility to "letting the data speak", we invite systemic flaws, biases, harms, and injustices into the powerful data systems influencing countless lives. Responsible data science requires proactive cultivation of a broad, diverse set of human voices and perspectives to guide the ethical application of these increasingly influential systems. We cannot neutrally let data speak for themselves -we must shape data science guided by human values. Diverse human minds must direct data systems to benefit humanity justly.

References

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

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

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