Letting Data Speak: Between Insight and Oversight*

Wanling Ma

March 13, 2024

Introduction

In today's world, where decisions are increasingly driven by data, the idea that data can simply "speak for themselves" is widely debated. This paper examines critical viewpoints from Jordan (2019), who questions the merge of AI with human oversight, D'Ignazio and Klein (2020), who challenge the neutrality of big data, and Au (2020), who redefines data cleaning as an analytical process. Together, these perspectives argue for a deeper, more nuanced understanding of data analysis, highlighting the essential role of human judgment in interpreting data.

Theoretical Perspective

Jordan (2019) provides a compelling critique of the contemporary understanding of Artificial Intelligence (AI), advocating for a shift away from the fascination with AI that mimics human intelligence. He stresses that the real challenge and opportunity lie in creating systems that enhance human decision-making rather than replace it. Jordan argues for a blend of computer science and statistics to construct inference-and-decision-making systems that safely and effectively augment human capabilities. This perspective underscores the significance of human oversight in the era of data and computing, advocating for an engineering discipline that respects human nuances and societal needs.

D'Ignazio and Klein (2020) delve into the inherent biases within big data projects, critiquing the often-unquestioned assumption of data's objectivity. They introduce the term "Big Dick Data" to describe projects that embody a patriarchal view, aiming to dominate and control through data collection and analysis, ignoring the critical importance of context and the situated nature of all knowledge. Their work calls for a feminist approach to data science, one

*Code and data are available at: https://github.com/WanlingMa/Letting-Data-Speak-Between-Insight-and-Oversight

that recognizes data as never neutral but always shaped by human values, biases, and socio-political contexts. This approach emphasizes the need for awareness and correction of biases in data analysis to ensure fairness and equity.

Case studies

Jordan's experience with an ultrasound predicting Down syndrome risk for his unborn child exemplifies the intricacies of data interpretation. He discovered that the increased resolution of newer ultrasound machines led to false positives in detecting markers for Down syndrome (Jordan 2019). This case underscores the need for understanding the provenance of data—where it comes from, how it's collected, and its limitations—emphasizing that data without context can lead to misguided decisions.

D'Ignazio and Klein's analysis of the reporting on the Boko Haram kidnapping incident by FiveThirtyEight showcases the pitfalls of uncritical reliance on big data. The outlet had used the Global Database of Events, Language, and Tone (GDELT) and inaccurately reported the number of kidnappings due to a misunderstanding of the data's nature (D'Ignazio and Klein 2020). This incident illustrates the dangers of ignoring the context in which data is generated and analyzed, highlighting the need for a nuanced approach that considers the human and social dimensions of data.

Conclusion

This examination of perspectives from Jordan, D'Ignazio, Klein, and Au challenges the notion that data inherently possesses objectivity, emphasizing instead the critical role of human oversight in data interpretation. The key takeaway is that data does not exist in a vacuum; it reflects the biases, judgments, and contexts of those who collect, clean, and analyze it. Jordan's narrative on medical diagnostics, D'Ignazio and Klein's analysis of big data misinterpretations, and Au's insights into data cleaning collectively underline the importance of scrutinizing data beyond its surface value. For data science to advance responsibly, it is imperative to acknowledge the subjective lens through which data is viewed and analyzed. Balancing the perceived objectivity of data with the subjectivity of its interpretation requires a nuanced approach, one that embraces complexity and seeks to understand the broader implications of data-driven conclusions. As the field evolves, fostering critical thinking and ethical practices among data scientists will be essential in navigating the intricate interplay between data and its interpretation.

References

Au, Kevin. 2020. "On Data Cleaning."

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism.* Cambridge, MA: MIT Press.

Jordan, Michael I. 2019. "Artificial Intelligence—the Revolution Hasn't Happened Yet." *Harvard Data Science Review.*