The Limits of Data Interpretation*

Hari Lee Robledo

2024-03-12

Introduction

The question of how much we should let the data speak for themselves remains a topic of significant debate and contemplation. Drawing insights from the works of Jordan (Jordan 2019), D'Ignazio and Klein (D'Ignazio and Klein 2020) and Au (Au 2020), this paper explores the balance between allowing data autonomy and the necessity for interpretative guidance. By exploring the implications of unbridled data autonomy.

Discussion

In the article "Artificial Intelligence–The Revolution Hasn't Happened Yet" (Jordan 2019), Michael I. Jordan argues that the reliance on data analysis without attention to provenance is a critical issue that extends beyond individual medical care. He emphasizes the need for a deeper understanding of where the data originates, the inferences drawn from it, and the relevance of those inferences to the present situation.

The author proposes the development of a new interdisciplinary field that integrates data and humanities, emphasizing the importance of an engineering discipline with principles of analysis and design that goes beyond mere data-based decision-making. An example of progress in this realm is the creation of Intelligence Augmentation (IA) and Intelligent Infrastructure (II) to enhance human intelligence and creativity, as exemplified by search engines and natural language translation. This reading prompts reflection on our ability as humans to train critical thinking AI when we may still lack in this aspect. The creation of AI and II systems, as portrayed in the article, becomes a reflection of human nature, encouraging a re-evaluation of the extent to which we should rely solely on data without incorporating higher cognitive and social understanding.

*Code and data from this analysis are available at: https://github.com/hari-lr/The-Limits-of-Data-Interpretation. Thank you to _____ for your valuable insights and constructive feedback during the review process.

Similarly, in their book "Data Feminism: 6. The Numbers Don't Speak for Themselves" (D'Ignazio and Klein 2020) by Catherine D'Ignazio and Lauren Klein, the authors suggest that we should not allow the data to speak for themselves entirely. The central argument is that data is not a raw, neutral input; rather, it is already "cooked" and influenced by various social, political, and historical factors. The emergence of a "data creative" class underscores the importance of creatively mining and combining data to produce new insights. The reading emphasizes the need for a feminist strategy that involves interrogating the context, limitations, and validity of the data.

They exemplify the potential challenges and limitations of letting data speak for itself through the Google Flu Trends project. This initiative aimed to predict the prevalence of influenza by analyzing people's web searches for flu-related symptoms. The idea was that an increase in searches for flu symptoms could serve as an early indicator of a flu outbreak in a particular region. However, over time, the project faced significant challenges as the data searches were susceptible to external factors, such as media reporting about the flu. The reading highlights that relying solely on the data without critically examining its limitations leads to problems. It prompts data scientists to interrogate the context and validity of the data they work with rather than accepting it at face value.

Lastly, the article "Data Cleaning IS Analysis, Not Grunt Work" complements and extends the ideas presented in The Numbers Don't Speak for Themselves. Both readings emphasize the importance of critical thinking and human perspective in the data analysis process. Randy Au's article aligns with this perspective by highlighting that data cleaning involves making value judgments and interpretations, reinforcing the idea that data, in its raw form, cannot speak for itself. Moreover, Au's article indirectly emphasizes the human role in data cleaning, suggesting that, despite advancements in AI, certain aspects of data analysis, such as cleaning and interpretation, still require human expertise.

Conclusion

Based on these readings, I believe that the point at which we should let the numbers speak for themselves is when we want to make applications outside of context. It is important to understand that we cannot generalize data, as each dataset is unique and influenced by various factors. Specifically, when there is, as Jordan puts it, a lack of transparency regarding the provenance of the data (Jordan 2019). Similarly, D'Ignazio and Klein argue that raw data is not truly raw; it's already manipulated from its collection and measurement. Therefore, even with clear provenance, inherent biases in the data still exist. The way Randy approaches it, suggests that data manipulation in any form involves human input, especially critical and analytical thinking (Au 2020). So, considering his perspective, I reiterate that data shouldn't speak for themselves to generalize outcomes or trends.

Letting data speak for themselves is a good way to gain general knowledge and understanding of many topics and contexts, but biases in both the data and how they are presented pose

significant limitations. While these biases cannot be eliminated, they can be minimized by making the process of data collection, cleaning, analysis, and application transparent.

References

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