Letting data speak for itself*

Unveiling the Voice of Data: Insights from Three Perspectives

Sehar Bajwa

March 13, 2024

> Oh you think we have good data on that? - attributed to Leslie Root by Rohan Alexander in his book 'Telling Stories with Data', and on Twitter, 2 August 2023. Spoiler alert - We do not.

---

1 Introduction

In the realm of statistical analysis, Jean-Paul Benzécri's notion to "let the data speak for itself" has reverberated through academic corridors since its articulation in his seminal work, *L'analyse des données*. The citation, originating from Benzécri's tome *L'Analyse des Correspondances* in 1973, encapsulates a fundamental principle in data analysis: the primacy of empirical evidence over theoretical assumptions.

Benzécri's proposition underscores a critical aspect of statistical modeling: the importance of ensuring that the structure of the model conforms to the inherent structure present within the data itself. Instead of imposing rigid preconceptions onto the data, Benzécri advocated for statistical models that extract insights directly from the data, allowing its inherent patterns and relationships to dictate the analytical framework.

Moreover, Benzécri's stance emphasizes the need for statistical methods to be robust and flexible, capable of accommodating the nuances and complexities inherent in real-world data. By allowing the data to guide the modeling process, analysts can avoid the pitfalls of overfitting and ensure that their conclusions are firmly grounded in empirical evidence.

In this article review, we delve into the multifaceted implications of Benzécri's dictum across various contexts in statistical analysis. Through exploration of diverse perspectives and case

*Code and data from this analysis are available at: https://github.com/SEHB2012/dataspeaks. Thank you to Hari Lee Robledo for your valuable insights and constructive feedback during the review process.

studies, we illuminate the significance of prioritizing empirical evidence and allowing the data to lead the way in statistical modeling and inference.

2 Research

2.1 Letting data speak for itself overlooks the inherent noise within raw data.

The first article(Au 2020) challenges the notion of letting data do its talking: it posits raw data is analogous to useful signals mixed in with noise. That is the fundamental premise of data cleaning: extracting these signals by preferentially transforming data so that the chosen analysis algorithm produces interpretable results, which is also the very act of data analysis. therefore, cleaning is but a subset of analysis: the minute one chooses to transform data implies decision-making and imposing value judgments on the data.

The implication from this reading is that if data were to speak for itself with all its inaccuracies and inconsistencies, it would spout gibberish. It also underscores the importance of analysts in actively shaping data and driving decisions at all stages of the process.

2.2 Letting data speak for itself risks reinforcing unjust power differentials and misaligned incentives, potentially leading to harmful outcomes.

The next article(D'Ignazio and Klein 2020) delves into the principles of data feminism and pluralism, highlighting the significance of incorporating diverse perspectives to attain a comprehensive understanding. It offers the Anti Eviction Mapping Project's (AEMP) Narratives of Displacement and Resistance map as a striking illustration of this approach. The map, featuring 5000 evictions depicted as red bubbles superimposed on a map of San Francisco, intentionally obscures the underlying map to underscore its message. This deliberate design choice challenges traditional information design norms, eschewing clarity and cleanliness in favor of emphasizing the crisis of gentrification.

In this instance, the data speaks for itself, unmistakably conveying the prevalence of evictions in the city. Moreover, the article acknowledges the historical context surrounding the concept of "cleaning" in data, recognizing its ties to eugenics and its potential to conceal diversity. This perspective aligns with the principles of pluralism, advocating for the incorporation of diverse viewpoints to achieve a more nuanced understanding of data and its implications.

Chapter 6 of the same publication makes notes that the premise of letting numbers speak for themselves posits that they are a raw input. However, data at the input stage is already full cooked - the result of a complex set of socio-political and historical circumstances.

2.3 Letting data speak for itself means prioritizing empirical evidence and rejecting preconceived notions.

The paper on Artificial Intelligence (Jordan 2019) recounts a troubling experience encountered by the author during his spouse's pregnancy. White spots detected around the fetus' heart were flagged as potential Down Syndrome markers, significantly increasing the risk of diagnosis. The recommended course of action was the risky procedure of amniocentesis, which carried a 1 in 300 fatality rate. However, leveraging his statistical expertise, the author delved deeper into the situation. He discovered that the new imaging machine responsible for the diagnosis produced higher quality images, raising the possibility that the observed calcium buildup spots were false positives.

Months later, the author was relieved to welcome the birth of a healthy baby. Yet, the episode continued to trouble him deeply. It underscored the crucial importance of letting the data speak for itself. Had the author not questioned the provenance of the data and accepted it at face value, it could have led to a perilous procedure and potentially a terminated pregnancy.

Moreover, the essay underscores the paramount significance of rigorous data analysis and the necessity for a systematic approach in designing large-scale systems that seamlessly integrate human and machine intelligence. Adhering to this principle enables decision-makers to steer clear of potential pitfalls and ensures that their actions are firmly grounded in evidence and sound reasoning derived from data.

In the broader context of AI and machine learning, allowing data to speak for itself entails prioritizing empirical evidence and insights garnered from thorough data analysis over preconceived notions or biases. This approach fosters transparency, accountability, and trust in AI systems, thereby fostering more ethical and responsible AI development and deployment.

3 Conclusion

In conclusion, the three articles explored in this paper collectively emphasize the significance of "letting data speak for itself" across diverse contexts. They underscore the critical role of analysts in shaping and interpreting data, advocating for the inclusion of diverse perspectives to amplify marginalized voices, and highlighting the consequences of blindly accepting data without critical examination. Together, they emphasize the importance of data literacy, critical thinking, and ethical considerations in navigating the complexities of the digital era. By empowering data to tell its story with clarity, accuracy, and compassion, we can harness its full potential to drive meaningful insights and positive societal change.

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.