Letting data speak for itself*

Findings from a National Telephone Survey

Sehar Bajwa

March 12, 2024

---

1 Introduction

2 Research

2.1 Data Cleaning Paper

The first article 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 judgements on the data.

The implication from this reading is that if data were to speak for itself with all its inaccuracies and inconsistensies, 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 Data feminism paper

The next article delves into the principles of data feminism and pluralism, highlighting the significance of incorporating diverse perspectives to attain a comprehensive understanding.

*Code and data are available at: https://github.com/SEHB2012/Nonresponsebias

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

2.3 AI paper

The paper on Artificial Intelligence 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.