To what extent that data can truly speak for itself*

Kuiyao Qiao

Invalid Date

Table of contents

1 Introduction

Based on an article by (**jordan2019?**), (**dignazio2020?**), and (**au2020?**), the controversial topic of the extent to which humans should allow data to "speak for itself" will be explored in the article. In light of Michael I. Jordan's critique of the current state of AI, Randy Au's exploration of data cleansing as a critical analytical process, and other perspectives on the interpretation of data in a limited number of situations, data can stand on its own. Still, responsible data science requires careful consideration of context and the application of thoughtful human judgement, especially for data about people and social phenomena.

*Available at: https://github.com/QPP123/To-what-extent-that-data-can-truly-speak-for-itself/tree/main

2 Beyond Raw Data

The idea that data can directly reveal the truth without human interpretation has been convincingly criticized. (**dignazio2020?**) explains that "raw data is an oxymoron" and is always the product of human choices about measuring, categorizing, and recording it. What gets counted and how it gets counted reflects the priorities, blind spots, and sometimes even biases of the person doing the measuring.

(**au2020?**) makes a similar point: "Cleaning data is analysis, not grunt work." Decisions dealing with missing data, outlier handling, variable recoding, etc., are all analytical choices that can affect the conclusions drawn. So-called "data cleaning" is an integral part of the analytical process, not a rote preliminary step. Considering data cleansing as an analytical activity rather than a preparatory one underlines the subjective judgement embedded in making data analyzable.

Moreover, as (**dignazio2020?**) illustrates with the example of campus sexual assault data, it is not always apparent on the surface that data are influenced by structural forces such as sexism and racism. Considering the imbalance of power and incentives in the data collection environment is essential. Low numbers of violations may reflect a lack of institutional support for survivors and barriers to reporting rather than a lack of violations. The data itself does not tell the story without exploring these contextual factors.

3 Big data and data cleansing

The larger the dataset, the less need for theory or human insight has taken hold in some circles. (**anderson2008?**) makes the provocative point that large data volumes mean that "relevance is enough" and that the scientific method is obsolete. (**jordan2019?**) argues that this reveals a "culture of predictive pragmatism" spreading across AI and big data - a push for accurate predictions disconnected from understanding.

However, as (**jordan2019?**) points out, many headline-grabbing examples of significant data failures, such as Google Flu Trends, are due to ignoring critical context. The massive media coverage distracted Google's flu-tracking algorithm, leading more people to search for the topic. More data alone is only a panacea if it is grounded in domain knowledge.

However, data is collected, processed and analyzed in an environment with human bias. D'Ignazio and Klein (2020) further elaborate on this point, showing how social structures and power dynamics influence data interpretation, thus questioning the notion of data as a purely objective entity (**dignazio2020?**).

Randy O's perspective on data cleansing emphasizes the subjective nature of the process, whereby decisions about data inclusion, exclusion and transformation are made according to the understandings and biases of the cleanser. This process, while essential, complicates the

idea of letting data speak for itself by introducing layers of human judgment before data analysis even begins.

4 When does data speak for itself?

Au (**Au2020?**) suggests that the data-cleaning process is inherently analytical. This perspective calls for a reassessment of how researchers and practitioners deal with data, noting that decisions made during the data cleansing process can significantly influence the outcome of subsequent analyses. As a result, every step, from data collection to analysis, is riddled with human judgement, which detracts from the perception of the inherent objectivity of data.

This is not to say that data can never stand on its own. I think that in some cases, data can broadly "speak for itself," but there are still some caveats:

Statistics and analysis: Simple univariate summaries (means, medians, counts), correlations, data visualizations, etc., can convey meaningful patterns without much additional context. However, they still reflect upstream sampling and measurement choices.

Controls and Measurements: Data from well-designed instruments, laboratory experiments, or recordings may be superficially reliable, but there will still be errors that cannot be controlled. For example, astronomers working with telescope images can have reasonable confidence in low-level data, even if interpreting higher-level data requires more care. However, the potential for error and bias remains.

Predictive performance on retained data: In specific predictive tasks, such as spam filtering face recognition or even the latest AI, performance on unseen test data is a meaningful measure of validity without getting carried away with model details. But only if the assumptions this applies in the first place are that both the training and test data are unbiased samples.

So, while the data can be partially informative, human context and judgment are always required. And the more the data relates to human behaviour or social outcomes, the more critical other perspectives become.

5 Conclusion

Moving from collection to analysis is a profoundly human and subjective process. While sometimes data can stand independently, critical examination and careful interpretation of the processes by which it is generated is essential, especially for data about human beings. Developing a critical understanding of the limitations and biases in the interpretation of data will be necessary as technical capabilities advance.

6 References