Beyond the Numbers: The Essential Context of Data*

David James Dimalanta

2024-03-13

In an era dominated by the allure of big data and advanced analytics, the perception of data as a neutral arbiter of truth persists, underpinned by the belief that data, when allowed to 'speak for themselves,' can offer clear, unbiased insights into our complex world. This paper challenges this notion by arguing for the indispensability of context in data interpretation. By highlighting the nuances of data science—a field often misconstrued as purely objective—this paper advocates for a more ethical, informed, and context-aware approach to data analysis. The synthesis of these perspectives illuminates the urgent need for a data science practice that is critically engaged, ethically grounded, and contextually informed, thereby moving beyond mere numbers to uncover the rich, contextualized insights that data can offer.

Table of contents

---

*Code and data are available at:https://github.com/davidjamesdimalanta/Beyond_the_Numbers/tree/main

1 Introduction

The advent of big data and advanced analytics has brought with it a pervasive belief in the power of data to provide objective insights into our complex world. This belief hinges on the notion that data, if left to "speak for themselves," can offer clear, unbiased truths. Yet, this paper argues that such a view is not only simplistic but potentially dangerous. Drawing on the insights of scholars like Randy Au, Michael I. Jordan, and others, I, along with the help of my peer Harrison Huang, delve into the debate surrounding the autonomy of data in revealing truths, advocating for a balanced approach that recognizes the indispensable role of context in data interpretation.

2 The Illusion of Neutral Data and the Necessity of Context

Data's facade of neutrality is contested by the intricate process of data cleaning, which Randy Au (Au 2020) critiques for being superficially treated in mainstream discourse. Far from a mere preparatory step, Au argues that data cleaning is a critical phase of analysis, demanding significant decisions about data's validity and relevance. This perspective underscores the necessity for a deep understanding of data's origins and the subjective judgments influencing its preparation for analysis; thus challenging the notion of data neutrality and arguing against the idea of letting data speak for themselves without considering the context from which they emerge.

Similarly, Michael I. Jordan's (Jordan 2019) critique of the current enthusiasm surrounding artificial intelligence (AI) and machine learning (ML) echoes this sentiment. Jordan highlights a crucial misunderstanding of these technologies, emphasizing that, despite the hype, we are far from creating AI systems that can rival human intelligence. Instead, these systems should be seen as tools that augment human capabilities, necessitating human oversight for their application and interpretation. This argument underlines why data cannot merely be left to "speak for themselves," as data interpretation requires a nuanced understanding of the influence of societal biases, power dynamics, and cultural contexts on data collection and interpretation.

Expanding on the theme of context, D'Ignazio and Klein (D'Ignazio and Klein 2020) illustrate through various examples how data is never neutral, being produced and interpreted within a framework of unequal social relations and contexts. This argument challenges the myth of data's objectivity and highlights the potential for societal biases to be perpetuated when data is analyzed without a critical lens on its context. Their analysis, along with Jordan's emphasis on human-centric considerations in AI and data analysis, calls for a nuanced approach to data science—one that recognizes data as inherently situated and shaped by power dynamics and social relations.

3 The Persuasive Power of Visualizations

Kennedy et al. (Helen Kennedy and Allen 2016) highlight the persuasive power of data visualizations by examining how commonly accepted conventions—like two-dimensional viewpoints, clean layouts, geometric shapes, and the inclusion of data sources—impart a false sense of objectivity and transparency. For example, a two-dimensional map using color gradients to denote population density can subtly influence perceptions by emphasizing certain areas over others, depending on the chosen color scheme. Similarly, the inclusion of data sources, while seemingly promoting transparency, can also persuade viewers by implying a rigorous, unbiased selection of data, even when the data's collection and processing methodologies might themselves be biased.

Danks and London (Danks and London 2017) add depth to this analysis by addressing algorithmic biases within autonomous systems, illustrating how biases in training data, algorithmic focus, and deployment context can significantly influence system outputs. An autonomous vehicle trained primarily in urban environments, for instance, may perform suboptimally in rural areas, reflecting a bias embedded during the training phase. Yet, they argue, not all biases are inherently negative; some, like the preferential processing of emergency vehicle signals over standard traffic patterns, are ethically justified and necessary for societal well-being.

Overall, the work of Kennedy et al. and Danks et al. highlights how data visualizations carry persuasive power, shaping perceptions through conventions that imply objectivity. This analysis suggests that visual representations of data, while seemingly transparent, are influenced by choices that reflect and can reinforce societal narratives and biases.

4 Doing Good with Data

In contrast, the optimism about data visualization's potential to enhance public understanding and awareness is echoed in the work of Kennedy et al. (Helen Kennedy and Allen 2016). They reference the historical Isotype project by the Neuraths as a pioneering example of using visual data representation to educate the public. Today, interactive visualizations like those depicting the impacts of climate change leverage intuitive graphical elements and interactivity to make complex datasets understandable and engaging for the general audience. These tools aim to democratize data, allowing users not only to consume but also to explore and interrogate information, thereby fostering a more informed public discourse.

5 Towards a More Responsible and Informed Use of Data

Collectively, the insights presented by Au, Jordan, and D'Ignazio and Klein advocate for a context-aware approach to data analysis. This approach not only enriches our understanding but also ensures that our reliance on data advances knowledge and decision-making in a way

that is ethical and reflective of the multifaceted realities it seeks to represent. By acknowledging the complexities and inherent biases in data collection and interpretation processes, we can move beyond the numbers to uncover the rich, contextualized insights that data has the potential to offer.

These insights from the literature underscore the imperative for a context-aware approach to data analysis that appreciates the complexities and biases inherent in data collection, processing, and visualization. It calls for an ethical framework that considers not only the technical aspects of data science but also its societal impacts and the moral responsibilities of those who wield these powerful tools. By fostering a critical engagement with data and its representations, we can move toward a practice of data science that truly serves the public interest, offering not just numbers, but meaningful insights that reflect and respect the diversity and complexity of human experience.

6 Conclusion

In conclusion, it becomes evident that data, far from being a transparent lens onto reality, are deeply intertwined with the contexts of their creation and use. The insights from Kennedy et al. (Helen Kennedy and Allen 2016) and Danks and London (**Danks2017?**), alongside those of Au (Au 2020), Jordan (Jordan 2019), and D'Ignazio and Klein(D'Ignazio and Klein 2020), compel us to critically assess the processes through which data are generated, analyzed, and visualized. They highlight the urgent need for a data science practice that is both ethically grounded and context-aware, one that navigates the fine line between leveraging data for insights and recognizing the biases and power structures that data can perpetuate. Ultimately, the responsibility lies in our hands to ensure that our engagement with data is informed, critical, and just, moving beyond mere numbers to grasp the complex narratives they hold within.

References

Au, Randy. 2020. "Data Cleaning IS Analysis, Not Grunt Work." *Counting Stuff*. Counting Stuff. https://www.counting-stuff.com/data-cleaning-is-analysis-not-grunt-65949948516031001b3512e5/.

D'Ignazio, Catherine, and Lauren Klein. 2020. "6. The Numbers Don't Speak for Themselves." In *Data Feminism*.

Danks, David, and Alex John London. 2017. "Algorithmic Bias in Autonomous Systems." In *Ijcai*, 17:4691–97. 2017.

Helen Kennedy, Giorgia Aiello, Rosemary Lucy Hill, and William Allen. 2016. "The Work That Visualisation Conventions Do." *Information, Communication & Society* 19 (6): 715–35. https://doi.org/10.1080/1369118X.2016.1153126.

Jordan, Michael I. 2019. "Artificial Intelligence—The Revolution Hasn't Happened Yet." *Harvard Data Science Review* 1 (1).