AI and Human in Data Analytics: Who leads this Dance?

Georgia Koutrika

Athena Research Center

Abstract

Integrating AI in data analysis methods promises unleashing the power of data and democratizing data access. It also raises several concerns. Perhaps, the most important one is the risk of less conscious decision-making or even pushing away domain and technical experts as automation is being applied over the whole data lifecycle.

1. AI in Data Analytics: The Opportunities

Machine learning has been used for extracting insights from data for a long time. For example, it is used for analyzing historical data to forecast future trends and behaviours, for extracting customer sentiment and preferences from customer reviews, or for anomaly detection in financial transactions to identify fraudulent activities. Recently, the introduction of deep learning methods to data analysis has enabled processing larger volumes of complex data at higher speeds and generating deeper insights. For instance, by using AI chatbots, businesses can allow average users to analyze large data sets and quickly extract key insights. For example, a sales person can ask questions such as: "Why did the sales decrease in March?". Generative AI can provide summary statistics and visualizations, offering an immediate and intuitive understanding of the data (e.g.,[1, 2]). If code is needed, AI can also help with code generation by translating to different programming languages and summarising code snippets (e.g., [3]).

AI used in data access and analysis promises data democratization, as more users can directly leverage stored data without programmers as middlemen, increased efficiency as many tasks from programming to data acquisition, computations and experimentations become faster or easier, and new discoveries that were not possible before or would be harder to reach.

2. AI in Data Analytics: The Risks

Example. At the time of the writing, a researcher asks ChatGPT for papers that cite the paper "Know What I don't Know: Handling Ambiguous and Unanswerable Questions for Text-to-SQL". The tool returns a list:

1. *Towards Unanswerable Question Detection in Text-to-SQL via Knowledge Graph and Semantic Reasoning (EMNLP 2023)*
2. *Handling Unanswerable Questions in Text-to-SQL with External Knowledge (ACL 2023)*
3. *Unanswerable Question Detection in Text-to-SQL with Natural Language Inference (arXiv 2023)*
4. *A Benchmark for Unanswerable Question Detection in Text-to-SQL (arXiv 2023)*

The papers are not only very relevant but are published in well-known venues. The researcher happily announces that there is work on this topic and the original paper is already well cited. Unfortunately, none of these papers exists. □

Inevitably, AI is transforming the way we work with and leverage data and that raises some strong concerns [4].

AI tools can lead to more reliance on pattern recognition without understanding the data. In the case of generative AI, people tend to treat it as a data retrieval tool, and hence trust its responses. However, as generative AI models learn the patterns and structure of their input training data, they can easily make up facts (hallucinations). Furthermore, algorithmic results can hide, or even, amplify bias in data.

On the other hand, when using deep learning models, answer verification, provenance and explainability are hard. Results can be irreproducible. Consider a traditional search engine, where a user submits a keyword search, say "data governance act" and the engine returns results. The user can easily check the result relevance and their credibility by simply checking the data (web pages). Using a tool like ChatGPT makes this (currently) impossible as there is no way to check how the answer is supported by the data. How can we trust the results? How can we base decisions on non-provable outcomes?

3. AI and the Human in Data Analytics

One of the biggest challenges we need to solve concerns the limits of AI, the nature of human intelligence and how best to regulate the interaction between the two.

DOLAP 2024: 26th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data, co-located with EDBT/ICDT 2024, March 25, 2024, Paestum, Italy

georgia@athenarc.gr (G. Koutrika)

This challenge encloses all applications of AI in our lives, and has several dimensions, including algorithmic, ethical, regulatory, and environmental. In this statement paper, we specifically focus on how AI in data access and analysis should lead to accelerated but not less-conscious decision-making. We are discussing two inter-connected directions researchers should work on.

Human-led Decision Making. Domain and technical experts are involved in all phases of the data lifecycle, from building the data pipelines and data analysis tools to using these tools for decision making. As AI is applied over the whole data lifecycle, it is natural to see the number of people involved decreasing. Nevertheless, humans should always be part of the process. Programmers should check that the code generated by AI is bug-free and performs what it is intended for. Domain experts should check that the algorithm or the data analysis tool performs well for a given domain. For example, can we build validation tools and benchmarks that help experts test the data analysis tool, and more specifically the AI algorithms working behind the scenes?

Humans should be also part of any decision-making process empowered by an AI algorithm or a tool. The tools should aid but not replace the humans. Humans, with their depth of understanding, ethical judgement and creativity, are irreplaceable. To actually enable conscious and informed decisions by humans, we need to build tools that enable and require human involvement.

Explainable by Design. Towards this direction, we need to build tools that can explain their answers to the user. In general, explainable AI implements specific techniques and methods to ensure that each decision made by the ML algorithm can be traced and explained. A data analysis tool supported by AI should be able to show how the answer is traced back to or supported by the original data. This entails work at both the algorithmic and the user interaction level.

At the algorithmic level, there are two ways to achieve explainability: (*in-processing*) algorithms with built-in explainability capabilities, and (*post-processing*) methods that trace the answer back to the data. At the interaction level, we can draw inspiration from the search engine paradigm, where results are connected to their source. Data analysis tools could follow a similar interaction paradigm. Results should be accompanied by evidence. In this way, checking the correctness of the answer will be inseparable part of the data analysis process.

In many ways, explainable AI is more critical than responsible or fair AI. Responsible AI looks at AI during the planning stages to make the AI algorithm responsible before the results are computed. Explainable AI looks at AI results after the results are computed. Without explainability, we cannot tell for sure whether the results are correct, biased, or generated by a responsible algorithm.

References

[1] Q. Wang, R. Castro Fernandez, Solo: Data discovery using natural language questions via a self-supervised approach, Proc. ACM Manag. Data 1 (2023). URL: https://doi.org/10.1145/3626756. doi:10.1145/3626756.

[2] Tableau GPT, https://tableau.com/products/tableau-ai (2023).

[3] W. Ahmad, S. Chakraborty, B. Ray, K.-W. Chang, Unified pre-training for program understanding and generation, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 2655–2668.

[4] R. V. Noorden, J. M. Perkel, AI and science: what 1,600 researchers think (2023).