LLM-aided data exploration and storytelling

The position is a 3 years fully funded PhD, starting September or October 2025.

Keywords: data stories, data narration, automatic data exploration, automatic data visualization, large language models, interactivity, human-machine interaction, Human-in-the-Loop data analysis

1 Context

In this data-driven era, it is of paramount importance to be able to access and exploit relevant data efficiently. Indeed, people and organizations need to collect, clean, transform, integrate, store, explore, analyze and summarize large amounts of data in order to **gain insights** and support any data-driven decision-making. **Data narration**, is about telling stories supported by insights extracted from data analysis, and rendered using interactive visualizations [2]. Indeed, data narration covers the whole cycle of data analysis, from data collection, wrangling and exploration, to insight reporting, visualization and storytelling [15].

The debate around AI, especially Machine Learning (ML) and Large Language Models (LLM), and their supposed capacity at automating decision making, is very intense these days. People may even wonder whether data narration be fully automated and eventually leave decisions to an Artificial Intelligence (AI). In the database (DB) community, and more particularly in the Business Intelligence (BI) community, there is long tradition of having the decision maker at the center of the data analysis process. At the inverse of automated application of algorithms, BI has been, since its inception, all about facilitating the task of interactive exploration of a dataspace. One could even say that BI is the ancestor of the Human-in-the-Loop Data Analysis phenomenon [6].

On the other hand, the automation of data narration is drawing attention nowadays [4, 13], which poses great challenges as background knowledge and human judgment are the keys to success [4]. Many recent works address the automatic discovery of insights [5, 9, 11] and their usage for enhancing data exploration [1, 3] and automating the overall data narration process (e.g. [20, 18, 19]). Even if such works envision full automation, they only deal with very specific scenarios where data exploration is reduced to the search of statistical findings.

LLM have the potential to automate many data narration tasks. Some first proposals essentially use LLM for matching users' goal [10, 17, 8]. So far, these proposals achieve best results when combining LLM with other narration techniques.

Since the early works demonstrating the potential for LLM as a key part of data narration, LLMs research has taken a few large leaps. LLMs are now capable of splitting complex reasoning tasks into many steps even relying on external tools and knowledge [7]. LLM personalisation has also been the subject of many studies with many techniques now covering several aspects of personalisation (tone and style, relevance, accuracy) [21].

2 Thesis topic

This PhD topic follows up the data narration model and process proposed by the PhD thesis of Faten El Outa [14], and ambitions to automatize much of the tedious parts of the process, while letting the end user in the loop.

This need, introduced in [16] is a first step to a broader vision of interactive data narration that opens major research questions, including:

1. How to leverage the narrator, auditor or both user's intentions to guide the data narration process? Letting the user in command means a tight connection among the automatic processes and the user preferences and feedback. The LLM as a comprehensive user interface should be flexible and its prompt well adapted to the user inquiry and preferences.

2. How to ensure that data stories are personalized? The data story should be adapted to the auditors' profiles (decision markers, data enthusiasts, ...). In particular, the substance of the story should be adapted to the user's knowledge. Also, the discourse (structure and presentation) should change. For example, the choice of structures and ordering should reflect diversity among the auditors, and insight-to-visual mappings should be personalized.

3. How to declaratively enhance the interplay between data exploration and narration tasks? The state of the art shows that data exploration is too much dissociated from narration, a full automatic approaches generally neglecting one of them. Such a declarative enhancement could take the form of an Explore-Narrate-Explore paradigm, letting the user intervene after the presentation of the story at each iteration, which reduces the overall time-to-message.

4. What is the quality of a data story? This transverse aspect of data story should be addressed to correctly assess proposals made to answer any of the previous points. Benchmarking data stories and the underlying data narration process is very challenging because of the complexity of the process. Many human-, system- and data-oriented measures exist (see a

survey in [12]), but their combination and the proposal of new holistic measures is still to come.

It is expected that this PhD contributes to answer one of the first 3 questions, while keeping an eye on the last one. The applicants are expected to indicate and motivate, in their applications, which of the questions they prefer to address.

3 Advising and application

Employers: University of Tours and University of Orléans (France)

Labs: Laboratoire d'Informatique Fondamentale et Appliquée de Tours (LIFAT) and Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)

Locations: Blois and Orléans (France)

Supervisors:

- Patrick Marcel, Full Professor, University of Orléans, patrick.marcel@univ-orleans.fr,
- Verónika Peralta, Associate Professor HDR, University of Tours, veronika.peralta@univ-tours.fr,
- Alexandre Chanson, Associate Professor, University of Tours, chanson@univ-tours.fr,
- Valentin Nyzam, Associate Professor, University of Tours, nyzam@univ-tours.fr

Requirements: Applicants are expected to hold a Master's degree in Computer Science, be skilled in databases, machine learning, programming and be fluent in English. Applicants must demonstrate proficiency in one of the following topics: data exploration, data narration, natural language processing. Applications failing to match these expectations to will be automatically rejected.

Application: Applicants will email, **before May 12, 2025 (firm deadline)**, the following documents to the supervisors: CV, transcripts of the Master's program, Master's thesis dissertation, cover letter, reference letters.

Shortlisted applicants will be contacted for a video interview that will include a discussion of the scientific literature relevant to the topic.

References

[1] Ori Bar El, Tova Milo, and Amit Somech. Automatically generating data exploration sessions using deep reinforcement learning. In *SIGMOD'2020*, Portland, USA, 2020.

[2] Sheelagh Carpendale, Nicholas Diakopoulos, Nathalie Henry Riche, and Christophe Hurter. Data-driven storytelling (dagstuhl seminar 16061). *Dagstuhl Reports*, 6(2):1–27, 2016.

[3] Alexandre Chanson, Nicolas Labroche, Patrick Marcel, Stefano Rizzi, and Vincent T'kindt. Automatic generation of comparison notebooks for interactive data exploration. In *EDBT'2022*, Edinburgh, UK, 2022.

[4] Tijl De Bie, Luc De Raedt, José Hernández-Orallo, Holger H. Hoos, Padhraic Smyth, and Christopher K. I. Williams. Automating data science. *Commun. ACM*, 65(3):76–87, 2022.

[5] Rui Ding, Shi Han, Yong Xu, Haidong Zhang, and Dongmei Zhang. Quickinsights: Quick and automatic discovery of insights from multi-dimensional data. In *SIGMOD'2019*, Amsterdam, The Netherlands, 2019.

[6] AnHai Doan. Human-in-the-loop data analysis: A personal perspective. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics, HILDA@SIGMOD 2018, Houston, TX, USA, June 10, 2018*, pages 1:1–1:6, 2018.

[7] Silin Gao, Jane Dwivedi-Yu, Ping Yu, Xiaoqing Ellen Tan, Ramakanth Pasunuru, Olga Golovneva, Koustuv Sinha, Asli Celikyilmaz, Antoine Bosselut, and Tianlu Wang. Efficient tool use with chain-of-abstraction reasoning. *ArXiv*, abs/2401.17464, 2024.

[8] Tavor Lipman, Tova Milo, Amit Somech, Tomer Wolfson, and Oz Zafar. LINX: A language driven generative system for goal-oriented automated data exploration. In Alkis Simitsis, Bettina Kemme, Anna Queralt, Oscar Romero, and Petar Jovanovic, editors, *Proceedings 28th International Conference on Extending Database Technology, EDBT 2025, Barcelona, Spain, March 25-28, 2025*, pages 270–283. OpenProceedings.org, 2025.

[9] Pingchuan Ma, Rui Ding, Shi Han, and Dongmei Zhang. Metainsight: Automatic discovery of structured knowledge for exploratory data analysis. In *SIGMOD'21*, pages Xi'an, China, 2021.

[10] Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, and Dongmei Zhang. Insightpilot: An llm-empowered automated data exploration system. In *EMNLP'2023 (System Demonstrations)*, Singapore, 2023.

[11] Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, and Dongmei Zhang. Xinsight: explainable data analysis through the lens of causality. *Proc. ACM Manag. Data*, 1(2):156:1–156:27, 2023.

[12] Patrick Marcel, Veronika Peralta, and Sihem Amer-Yahia. Data narration for the people: Challenges and opportunities. In *EDBT'2023, tutorial*, Ioannina, Greece, 2023.

[13] Tova Milo and Amit Somech. Automating exploratory data analysis via machine learning: An overview. In *SIGMOD'2020*, Portland, USA, 2020.

[14] Faten El Outa. *A framework for crafting data narratives.* PhD thesis, University of Tours, France, 2023.

[15] Faten El Outa, Patrick Marcel, Veronika Peralta, and Panos Vassiliadis. Highlighting the importance of intentional aspects in data narrative crafting processes. *Inf. Syst. Frontiers*, 25, 2023.

[16] Verónika Peralta. *Leveraging users' behavior, intentions and interests for enhancing Exploratory Data Analysis and Data Narration.* University of Tours, France, 2024.

[17] Gerard Pons Recasens, Miona Dimic, and Besim Bilalli. Capturing analytical intents from text. In *ADBIS'2024 (Short Papers)*, Bayonne, France, 2024.

[18] Danqing Shi, Xinyue Xu, Fuling Sun, Yang Shi, and Nan Cao. Calliope: Automatic visual data story generation from a spreadsheet. *IEEE Trans. Vis. Comput. Graph.*, 27(2):453–463, 2021.

[19] Mengdi Sun, Ligan Cai, Weiwei Cui, Yanqiu Wu, Yang Shi, and Nan Cao. Erato: Cooperative data story editing via fact interpolation. *IEEE Trans. Vis. Comput. Graph.*, 29(1):983–993, 2023.

[20] Yun Wang, Zhida Sun, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang. Datashot: Automatic generation of fact sheets from tabular data. *IEEE Trans. Vis. Comput. Graph.*, 26(1):895–905, 2020.

[21] Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen K. Ahmed, and Yu Wang. Personalization of large language models: A survey. *CoRR*, abs/2411.00027, 2024.