

PhD topic: data stories for interactive intentional analytics

The position is a 3 years fully funded PhD.

Employer: University of Tours (France) in partnership with University of Ioannina (Greece)

Lab: Laboratoire d'Informatique Fondamentale et Appliquée de Tours (LI-FAT)

Location: Blois (France) with 3-12 months stay in Ioannina (Greece). The recruited student can apply to an Erasmus grant for covering the stay in Greece.

Supervisors:

- Patrick Marcel, Associate Professor HDR, University of Tours, patrick.marcel@univ-tours.fr,
- Verónica Peralta, Associate Professor, University of Tours, veronika.peralta@univ-tours.fr,
- Panos Vassiliadis, Professor, University of Ioannina, panos.vassiliadis@cs.uoi.gr

Keywords: data analytics, data exploration, data stories, data quality, data mining, intentional OLAP

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.

Application: Applicants will email, **before May 13th, 2019 (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 Skype interview that will include a discussion of the scientific literature relevant to the topic.

1 Context

Can data analysis be fully automated and eventually an Artificial Intelligence (AI) makes the decision? The debate around AI, especially Machine Learning (ML), and their supposed capacity at automating decision making, is very intense these days. In the database (DB) community, and more particularly in the data warehousing (DW) 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 [4], DW has been, since its inception, all about facilitating the task of interactive exploration of a dataspace, and not let e.g., an algorithm automatically mine this space for patterns. One could even say that DW is the ancestor of the Human-In-the-Loop Data Analysis phenomenon [2].

This PhD topic follows up from the reinvention of OLAP described in [10, 11], and ambitions to automatize further interactive data analysis, while letting the end user in command. This reinvention of OLAP introduces an analytics model redefining what a query is, with respect to both what users ask the system, what the answer entails, and how this answer is computed. An implementation is currently being done.

The work introduced in [11] is a first step to a broader vision of interactive data analysis [7], that opens major research questions, including:

- How to facilitate the understanding of data? This demands to precisely define what are the answers to complex sequences of high level intentions, and package them into coherent data stories accessible to even non expert users.
- How to optimize the overall interactive analysis? Performance of query processing has traditionally been the focus of the DB community, but it has overlooked key indicators pertaining to the quality of the overall interaction with the data.

2 Expected contributions

This PhD will focus on the topics listed above, and more precisely:

- In [11], it is proposed that answers to intentional operators are no more traditional set of tuples, but dashboards including data, charts, informative summaries of KPI performance, as well as concise representations of knowledge hidden in the data. The first challenge of this PhD work will be to define how to structure such dashboards in a context where the interactive data analysis is a sequence of possibly complex queries, each being a composition of intentions, in a personalized way [8].
- Facilitating data understanding also demands to automatically put query results in a shape that is easy to grasp and that facilitates data storytelling. The second challenge of the PhD lies in identifying the most ap-

appropriate graphical representation of query answers, and to automatically craft narratives, commenting on the highlights presented, etc. [5, 6].

- Finally, executing complex intentional statements requires an optimization phase to decide which logical operators and model mining algorithms to execute [7]. This optimization can be thought in terms of performance, but also in terms of information content delivered and in terms of the quality of the user’s experience [1], measuring the number of insights [3, 9], controlling false discoveries [12], etc.

References

- [1] Mahfoud Djedaini, Krista Drushku, Nicolas Labroche, Patrick Marcel, Verónica Peralta, and Willeme Verdeau. Automatic assessment of interactive OLAP explorations. *Information Systems*, 82:148–163, 2019.
- [2] 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.
- [3] Philipp Eichmann, Emanuel Zraggen, Zheguang Zhao, Carsten Binnig, and Tim Kraska. Towards a benchmark for interactive data exploration. *IEEE Data Eng. Bull.*, 39(4):50–61, 2016.
- [4] Matthias Feurer, Aaron Klein, Katharina Eggenberger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2962–2970, 2015.
- [5] Dimitrios Gkesoulis, Panos Vassiliadis, and Petros Manousis. Cinecubes: Aiding data workers gain insights from OLAP queries. *Inf. Syst.*, 53:60–86, 2015.
- [6] Jessica Hullman, Steven M. Drucker, Nathalie Henry Riche, Bongshin Lee, Danyel Fisher, and Eytan Adar. A deeper understanding of sequence in narrative visualization. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2406–2415, 2013.
- [7] Patrick Marcel, Nicolas Labroche, and Panos Vassiliadis. Towards a benefit-based optimizer for interactive data analysis. In *Proceedings of the 21st International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data, co-located with EDBT/ICDT Joint Conference, DOLAP@EDBT/ICDT 2019, Lisbon, Portugal, March 26, 2019.*, 2019.

- [8] Tova Milo and Amit Somech. Next-step suggestions for modern interactive data analysis platforms. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 576–585, 2018.
- [9] Amit Somech, Tova Milo, and Chai Ozeri. Predicting ”what is interesting” by mining interactive-data-analysis session logs. In *Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019*, pages 456–467, 2019.
- [10] Panos Vassiliadis and Patrick Marcel. The road to highlights is paved with good intentions: Envisioning a paradigm shift in OLAP modeling. In *Proceedings of the 20th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data co-located with 10th EDBT/ICDT Joint Conference (EDBT/ICDT 2018), Vienna, Austria, March 26-29, 2018.*, 2018.
- [11] Panos Vassiliadis, Patrick Marcel, and Stefano Rizzi. Beyond roll-up’s and drill-down’s: An intentional analytics model to reinvent OLAP. *Accepted to Information Systems*, <https://doi.org/10.1016/j.is.2019.03.011>, 2019.
- [12] Zheguang Zhao, Lorenzo De Stefani, Emanuel Zraggen, Carsten Binnig, Eli Upfal, and Tim Kraska. Controlling false discoveries during interactive data exploration. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, pages 527–540, 2017.