Title: Human-guided Labeling for high-quality Data Collection

Themes: databases, data cleaning, human-in-the-loop

Project research oriented

Advisors
- Andrea Mauri (Lyon 1 Univ., andrea.mauri@univ-lyon1.fr)
- Angela Bonifati (Lyon 1 Univ., angela.bonifati@univ-lyon1.fr)

Context

Data collection is the first and the most crucial step in any data-intensive application. It is the foundation of any data analysis pipeline, both the ones based on machine learning technology and the traditional data querying and integration approaches. Existing data quality annotations and repair methods are either at dataset level - or very high level - and while they provide valuable information to the people who desire to use the datasets, they do not provide enough information about the quality of the data itself nor contribute to the data cleaning process- or they often limited on a subset of inconsistencies (e.g., like denial constraint), or works on specific data structures (e.g., relational or graph), and - while they are very efficient in term of execution - they cannot solve inconsistencies where external knowledge is needed.

Assignment

This project will investigate how to include a human perspective in the data collection and preprocessing phase. It will develop new human-in-the-loop algorithms that employ human expertise and domain knowledge to either collect data or support the pre-processing step to obtain data of higher quality.
In particular, the project will focus on the definition of metrics and algorithms to evaluate the quality of data during the data collection process.
The assumption is that humans are particularly knowledgeable of the data when they are its **owner** because they know how the data was collected (e.g, with a wearable device) and in which context (e.g., while running, sleeping) and may be aware of conditions that could have altered the quality of the data (e.g., the data is missing because the person didn't wear the smartwatch, or there is an outlier because the bracelet was loose, etc..).
However, humans are slow, costly to engage with (both in terms of time and money), and have limited cognitive capacity. For this reason, the developed algorithms should integrate human intelligence and automatic approaches in a smart, efficient, and effective way.

Tasks:
- Investigate existing data quality metrics (e.g. inconsistency degrees [2,3]) that are suitable for this kind of approach (or come up with a new one). Specifically which kind of human annotations can contribute to which data quality metric.
- Inspired by the data donation concept [5] and existing existing human-guided approaches [1,4], develop a human-in-the-loop algorithm to annotate the data and calculate these metrics.

- Evaluate your proposed solution through internal indicators on the quality of the metrics (e.g., consistency, labelers inter-agreement in case of human annotators, etc..) and effectiveness on data repair tasks.

Expected abilities:

- Very good programming skills
- Very good communication skills
- Familiarity with data management techniques
- Familiarity with machine learning methods appreciated

Opportunities: The selected students (the topic can be pursued by a team of students) will have the opportunity to work with top-class researchers in the area of data management and to be involved in writing a research article documenting the results.
The internship also offers the possibility to continue for a PhD scholarship.

References

1. Andrea Mauri and Alessandro Bozzon "Towards a human in the loop approach to preserve privacy in images." *11th Italian Information Retrieval Workshop, IIR 2021*. 2021. https://pure.tudelft.nl/ws/files/103169092/paper6.pdf

2. Parisi, Francesco, and John Grant. "On Database Inconsistency Measures." 29th Italian Symposium on Advanced Database Systems, SEBD 2021. Vol. 2994. CEUR-WS, 2021. http://ceur-ws.org/Vol-2994/paper20.pdf

3. Ousmane Issa, Angela Bonifati, and Farouk Toumani. 2020. Evaluating top-k queries with inconsistency degrees. Proc. VLDB Endow. 13, 12 (August 2020), 2146–2158. https://doi.org/10.14778/3407790.3407815

4. Alessandro Bozzon, Marco Brambilla, Stefano Ceri, Andrea Mauri, Riccardo Volonterio. "Pattern-based specification of crowdsourcing applications." *International Conference on Web Engineering*. Springer, Cham, 2014. https://link.springer.com/chapter/10.1007/978-3-319-08245-5_13

5. Gomez Ortega, A., Bourgeois, J., & Kortuem, G. (2022, October). Reconstructing Intimate Contexts through Data Donation: A Case Study in Menstrual Tracking Technologies. In the Nordic Human-Computer Interaction Conference (pp. 1-12). https://dl.acm.org/doi/10.1145/3546155.3546646