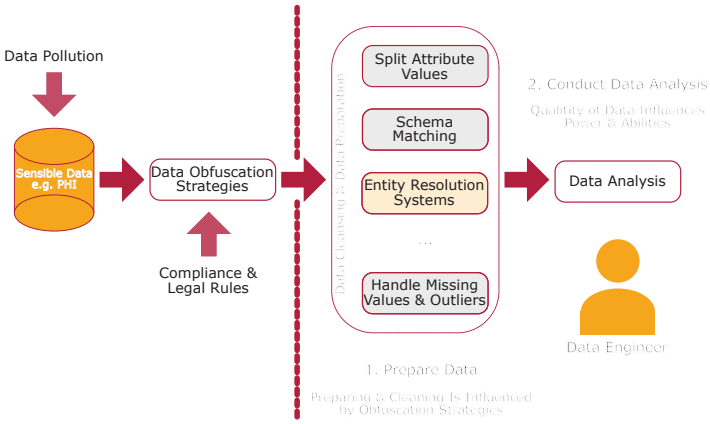
Analyzing the Impact of Data Obfuscation Strategies on Entity Resolution Systems

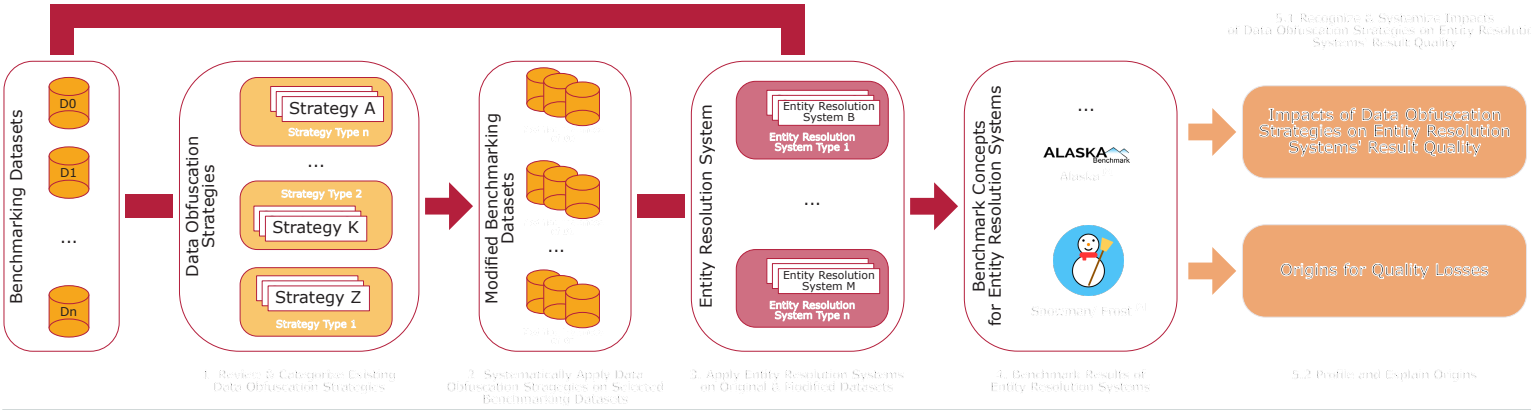
Data Obfuscation methods are used for de-identifying data while preserving original data formats and data types. **Entity Resolution** systems search for entries in the dataset that reference the same real-world object, so-called duplicates.

Data Obfuscation methods and *Entity Resolution* systems are common parts of real-world data integration pipelines. Recent research neglected the impact of *Data Obfuscation* methods on the quality of *Entity Resolution* systems' results. Hence, I propose a research project to close this gap.

Problem. Legal and compliance rules require the application of *Data Obfuscation* strategies before sensitive datasets are handed to data engineers. For instance, only *Data Obfuscation* strategies allow sharing Protected Health Information (PHI) with data engineers. In practice, *duplicates* represent one main cause of dirty data which can be cleaned with *Entity Resolution* systems. Recent studies analyzed the impact of *Data Obfuscation* strategies on *Entity Recognition* systems[1][3] and *Collaborative Filter* systems[4]. However, data engineers cannot estimate if and how previously applied *Data Obfuscation* strategies worse *Entity Resolution* systems' result qualities due to missing studies.



Analyze the impact of actual Data Obfuscation strategies on Entity Resolution systems' result quality and trace origins of quality losses.



References.

1) Catelli R., Garbigulo F., Casola V., De Pietro G., Fujita H., Esposito M. (2020) Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set. Applied Soft Computing. vol 94. ELSEVIER
2) Crescenzi V., De Angelis A., Firmani D., Mazzei M., Merialdo P., Piai F., Srivastava D. (2021) Alaska: A Flexible Benchmark for Data Integration Tasks. Computing Research Repository
3) Berg H., Henriksson A., Dalianis H. (2020) The impact of De-identification on Downstream Named Entity Recognition in Clinical Text. Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis. Association for Computational Linguistics
4) Berkovsky S., Kuflik T., Ricci F. (2012) The impact of data obfuscation on the accuracy of collaborative filtering. Expert Systems with Applications. vol 39. ELSEVIER
5) Markl V., Traub J., Kaoudi Z., Quiané-Ruz J.-A. (2021) Agora: Bringing Together Datasets, Algorithms, Models and More in a Unified Ecosystem [Vision]. ACM SIGMOD. vol 49. ACM DL
6) Naumann F., Gremmelspacher R., Panse F., Laskowski L., Graf M., Sold F., Papsdorf F. (2021) Frost: Benchmarking and Exploring Data Matching Results
7) Saeedi A., Peukert E., Rahm E. (2020) Incremental Multi-source Entity Resolution for Knowledge Graph Completion. The Semantic Web. ESWC 2020. Lecture Notes in Computer Science, vol 12123. Springer

Lecture Series on Database Research.

- Prof. Markl envisions a unified asset ecosystem, called AGORA[5]. First, its market aggregator may use *Entity Resolution* systems for finding equivalences among data assets. Furthermore, data asset constraints may allow a broader sharing of sensitive data if Data Obfuscation strategies were applied.
- Prof. Rahm emphasizes the relevance of Entity Resolution systems in data integration pipelines as well as for enabling automatic creation and refinement of large-scale knowledge graphs[7].

Florian Papsdorf

Lecture Series on Database Research, Master Program IT-Systems Engineering
Hasso Plattner Institute, Potsdam, Germany

E-Mail: firstname.lastname@student.hpi.de

HPI Hasso Plattner Institut