The proliferation of data hosted on data lakes present a great asset for AI practitioners. However finding relevant datasets remains time-consuming. This is because Data lakes:

- Are schema-agnostic: Data Portals have limited or no metadata associated with the datasets (Fig 1A)
- Offer basic data discovery search operations such as keyword search, filtering, or sorting. (Fig 1B)

**Problem & Motivation**

**Profiler:** Creates profiles for the different columns in the tables in the data lake.

**Knowledge Graph Builder:** Leverages the profiles to create a schema modeled as a KG.

**Linker:** Merges a given KG with the existing one by leveraging Graph Neural Networks.

**APIs** Allow the user to interact with the KG to perform data discovery and data enrichment operations.

1- Load data from data lakes.
2- Profile the raw data and store them on a document DB.
3- Build Knowledge Graph by using the profiles and inferring relationships based on the profiles.
4- Host the Knowledge Graph on a RDF-star store.
5- Find relevant data using the APIs on top of the KG.

A - Load a new Knowledge Graph to merge with the existing one.
B - Merge the two KGs and host the result on the RDF-star store.

**Knowledge Graph**

The KG is the schema for the data loaded by the Profiler. The KG has a hierarchical granularity: dataset, table, and column.

Edges are annotated: «lac:A lac:contentSimilarity lac:B» lac:certainty 0.88