# Bachelor Thesis Topic: Visualisation and Interactive Exploration of Data Changes in Data Engineering Workflows

## 1 Introduction

Data Engineering workflows are usually scripts that combine different data preprocessing algorithms. In a Data engineering workflow, the input dataset sets (tidy datasets) are stepwise adapted and changed. Tidy datasets are datasets in which "'each variable is a column, each observation is a row, and each type of observational unit is a table"' ([Wic]) or shortly relational-like datasets.

Aim is to check data quality characteristics, select data, to improve the data quality and to transform data into a target structure.

## 2 Main Task of this Bachelor Thesis

Often the Data Engineering workflows are python scripts using pre-implemented data preprocessing algorithms.

```python
import pandas as pd
from sklearn.impute import SimpleImputer

# read the data from a csv file into a pandas Dataframe
raw_data = pd.read_csv('students.csv')
# delete entries with the same matriculation number (deduplication)
data = data.drop_duplicates(subset='matriculation_number')
# impute missing values for column "age" by setting it to the mean
imputer = SimpleImputer(strategy='mean')
data['age'] = imputer.fit_transform(data['age'])
# next step would be to pass the processed data to a ML model
...
```

Listing 1: Simple data pipeline implemented in Python

Listing 1 gives a sample script for a data engineering script. This notation is readable for computer scientists, but more complicated for scientists from other fields like biology, chemistry or economy. From the scripts, it is not clear how the datasets are changed. Figure 1 sketches a visualisation of the data engineering workflow.

To sum it up: A question yet to be answered is to explain the data changes which a data engineering workflow produces.

In this Bachelor thesis, a first step in this direction shall be made. A program that visualises data changes shall be developed with the aim to visualise the data changes and to offer an interactive method to explore these changes. It can be assumed that:

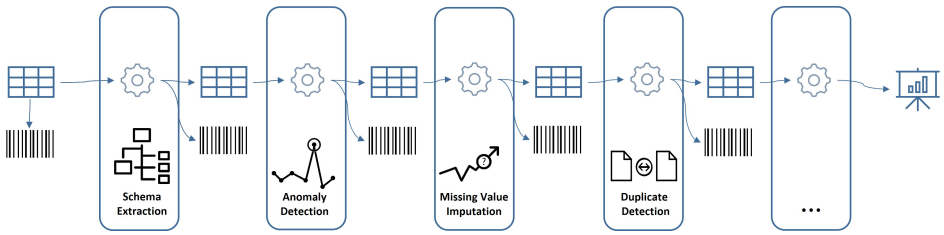


Figure 1: Visualisation of a Data Engineering Pipeline.

- a (JSON) file describing the DE workflow (DE algorithms, input and output datasets of each algorithm, and the sequence of the DE algorithms) is available and
- the diff files which are providing all differences between all datasets are also available (see Figure 2)

In both cases, these interfaces can be extended or changed if necessary.

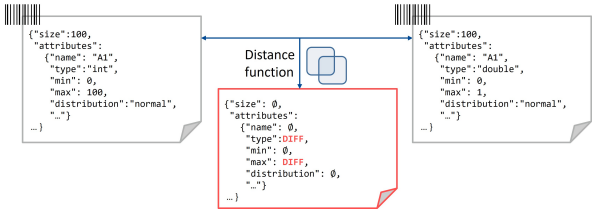


Figure 2: Diff File representing the Differences between two Data Sets.

Main Aim of the Bachelor Thesis is a program that supports a human-in-the-loop-approach in which users can *visualise the characteristics of datasets* and *interactively explore the differences between two datasets.*

## 3 Subtasks

In this Bachelor Thesis, the following subtasks shall be solved:

- requirements definition
- selection of the suitable libraries for visualisation and interactive approaches
- developing a graphical presentation to visualise the DE workflow
- visualising the data in a DE workflow
  1. visualisation of the different data characteristics and
  2. visualisation of the differences between two datasets
  3. interactive method for exploring data characteristics
  4. interactive method for exploring the differences between two datasets.

## 4 Application fields

An in-depth analysis of a dataset is needed for different data engineering tasks. It can be used for

1. selecting datasets for certain analytical tasks. Each of these selection processes starts with "understanding" the dataset, evaluating size, completeness and data quality.

An in-depth comparison of two datasets is needed for different data engineering tasks. It is used for

1. understanding the data changes in one data engineering workflow
2. furthermore, it is prerequisite for comparing two different data engineering workflows (e.g. in case of workflow evolution or workflow adaptation)

In the first case, the data sets are compared before and after each data processing algorithm. For example, in an outlier removal algorithm, the input data set (before the algorithm is executed) and the output data set (after the algorithm is executed) are compared. The results between the two data sets is represented in the diff files (see Figure 2) and describes the effect of the algorithm, even if the data processing algorithm itself is treated as a black box. The interactive graphical user interface which shall be developed in the thesis shall enable users to see and understand the workflow.

In the second application, two different data engineering algorithms can be compared based on the data diff files. This can be used in cases of exchanged algorithms and the effects shall be compared.

# 5 Technical Environment

- datasets in JSON syntax
- implementation in python and python libraries for visualisation (or other languages are possible, too)

## 5.1 Literature

- Meike Klettke: Between Data Lakes and Research Data Management – Data Engineering Tasks for the Next Decade, Fresh Thinking Talk, BTW 2023
- S Grafberger, J Stoyanovich, S Schelter: Lightweight inspection of data preprocessing in native machine learning pipelines, Conference on Innovative Data Systems Research (CIDR), 2021
- Puplications of the HILDA (Workshop on Human-In-the-Loop Data Analytics) Workshops 2020 and 2022, 2023: https://hilda.io/2020/ and https://hilda.io/2022/, and (soon available) https://hilda.io/2023/

# References

[Wic] H. Wickham. Tidy Data. *Journal of Statistical Software, 59(10)*, page 1–23. https://doi.org/10.18637/jss.v059.i10.