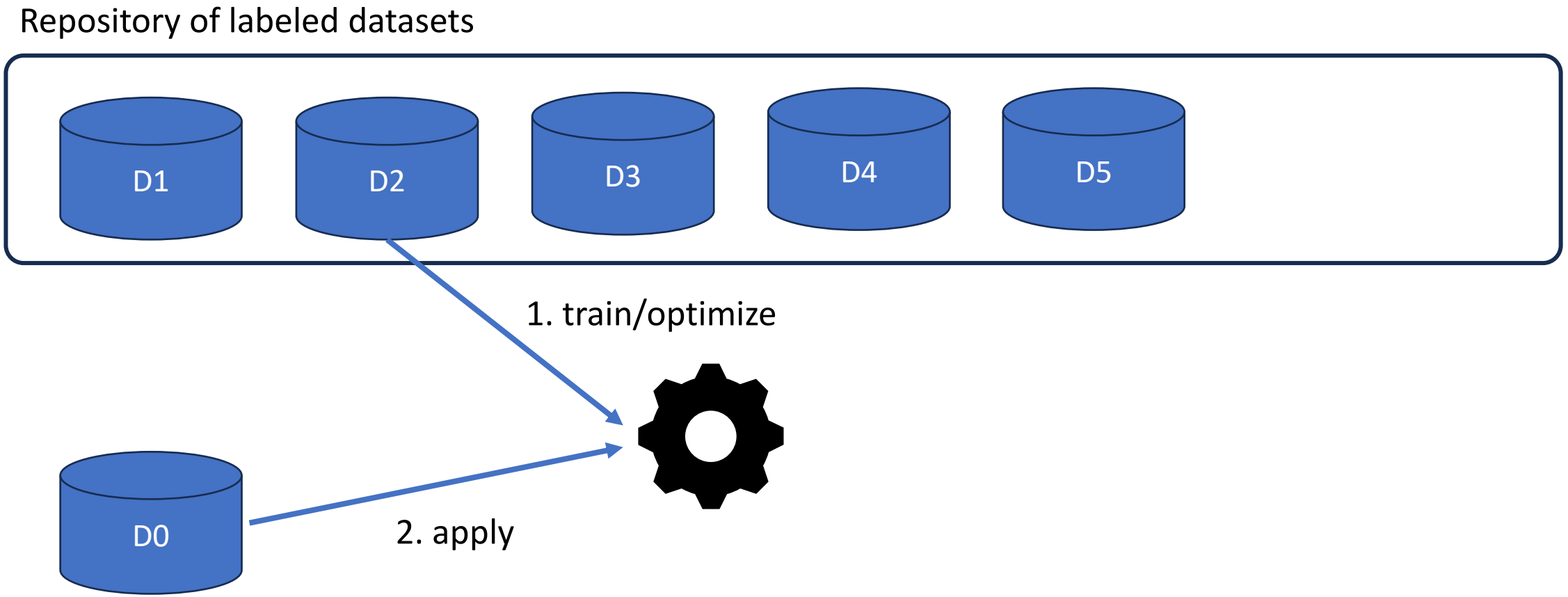
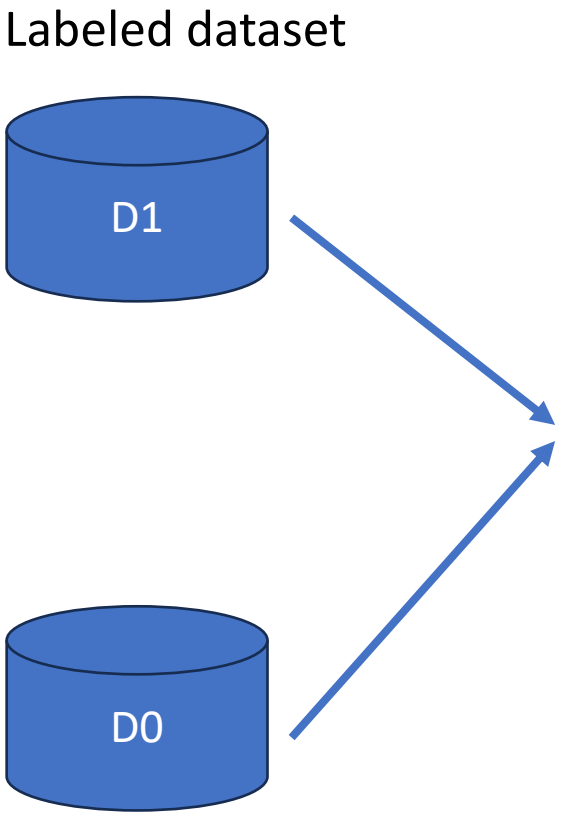
Finding Testdataset for Transfer Learning



Can we use any of the labeled dataset to configure a data matching approach for D0?

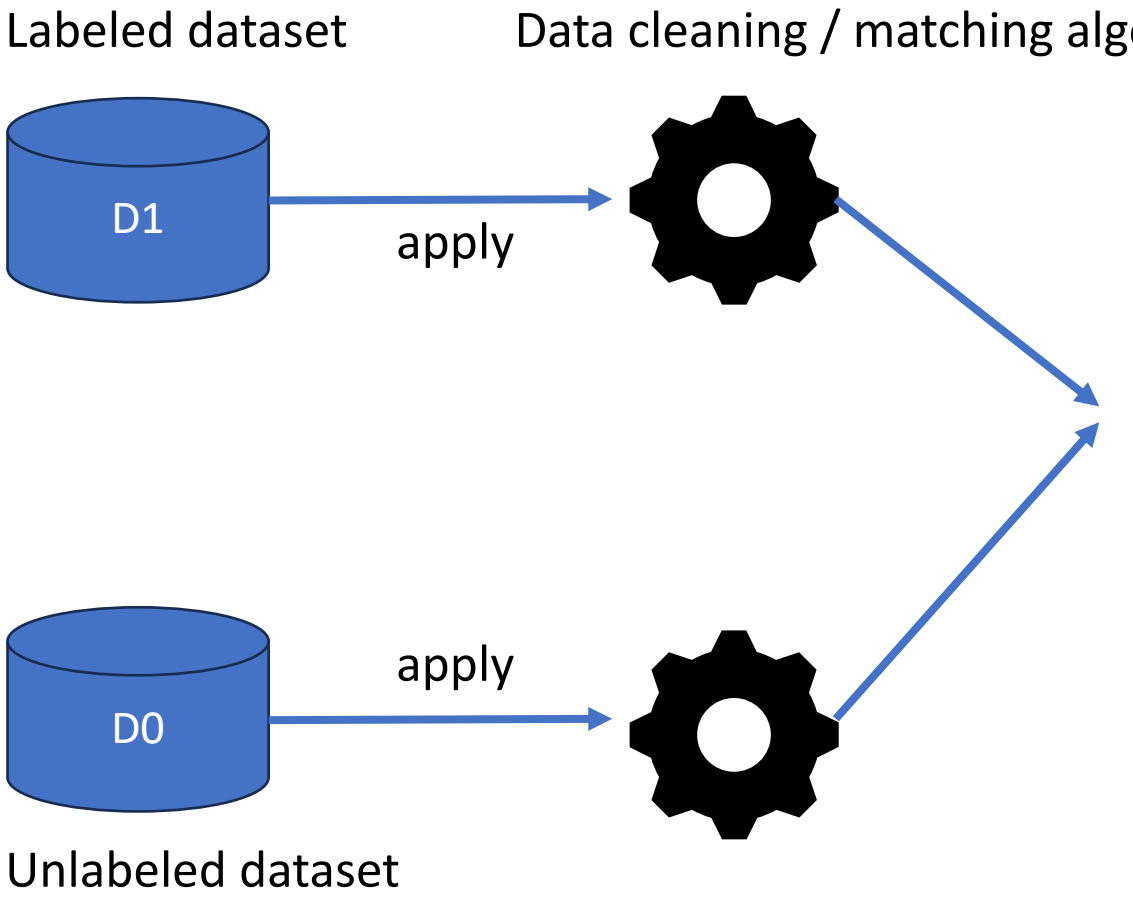
First Approach



Comparison of dataset characteristics, such as:

- Number attributes
- Number records
- Data types (numerical, categorical, text)
- Sparsity
- Integrity constraints

Second Approach

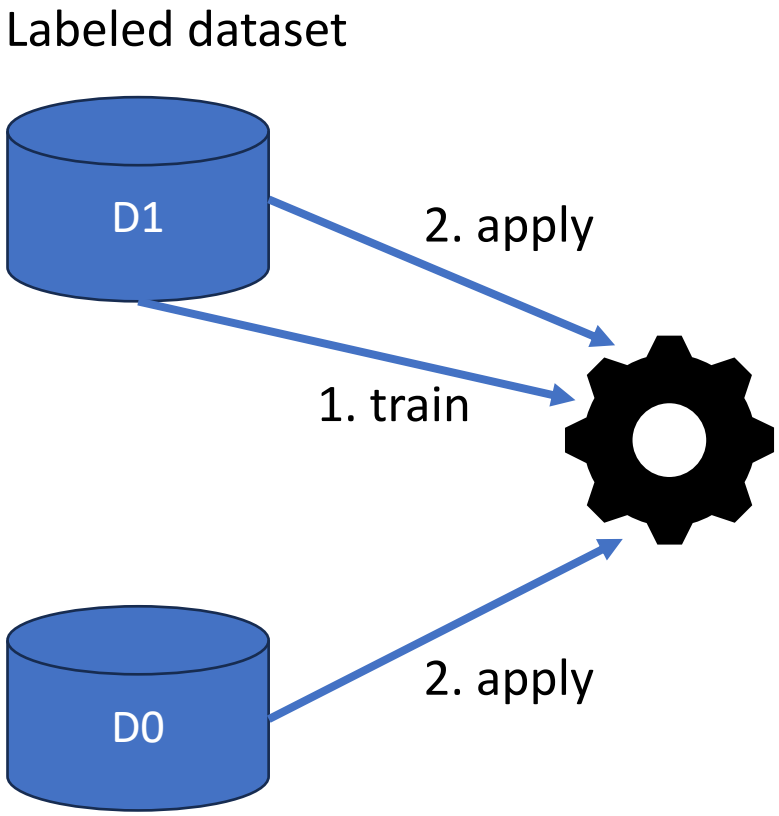


Comparison of data cleaning / matching results:

- Number of detected errors
- Types of detected errors
- Distribution of duplicate clusters
- Number of Matches misses for transitive closure
- Structure of duplicate pair graph

Maybe a similarity based classifier
where results are compared for
different threshold values
(e.g., comparison of two distributions)

Third Approach

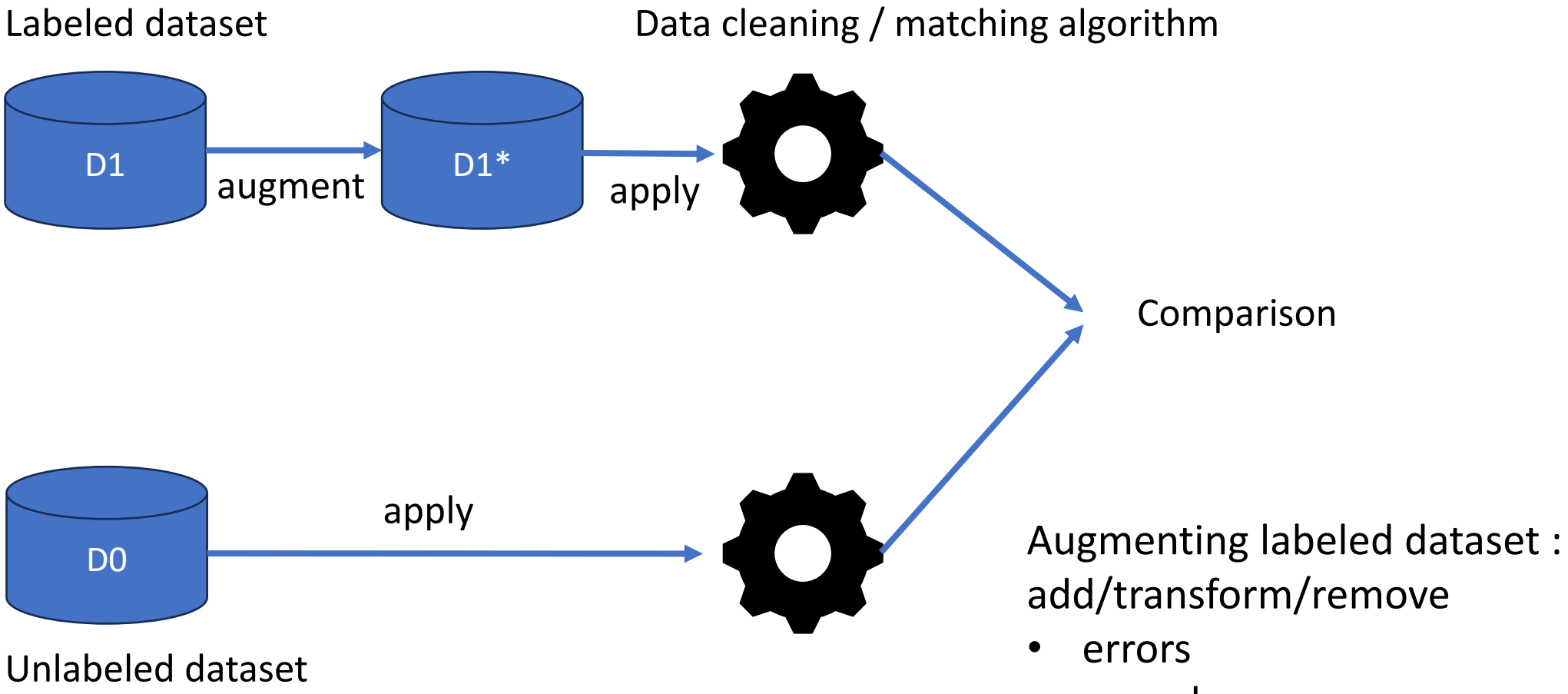


Comparison of matching results:

- Distribution of duplicate clusters
- Number of Matches misses for transitive closure
- Structure of duplicate pair graph

Maybe with different train/test splits of D1

Augmentation



Augmenting labeled dataset :
add/transform/remove

- errors
- records
- columns

to increase its suitability

First Steps

1. Building a suitable repository of 10-15 labeled datasets
2. Building a suitable repository of 3-5 data cleaning/matching algorithms
3. Evaluating suitability among these datasets
   (would be interesting to see how much the quality differs
   and depends on dataset used)