Raha: A Configuration-Free Error Detection System

Mohammad Mahdavi[1], **Ziawasch Abedjan**[1], **Raul Castro Fernandez**[2], **Samuel Madden**[2], **Mourad Ouzzani**[3], **Michael Stonebraker**[2], and **Nan Tang**[3]

[1] TU Berlin
{mahdavilahijani, abedjan}@tu-berlin.de

[2] MIT
{raulcf, madden, stonebraker}@csail.mit.edu

[3] QCRI, HBKU
{mouzzani, ntang}@hbku.edu.qa

Motivation

- ❑ Error detection is the task of finding wrong values
  - E.g., the red values in the table
- ❑ Traditional algorithms need input rules or parameters
  - E.g., a rule violation detector needs $Kingdom \rightarrow Lord$
- ❑ ML-based approaches need a large training set
  - E.g., 1% of the dataset [6]

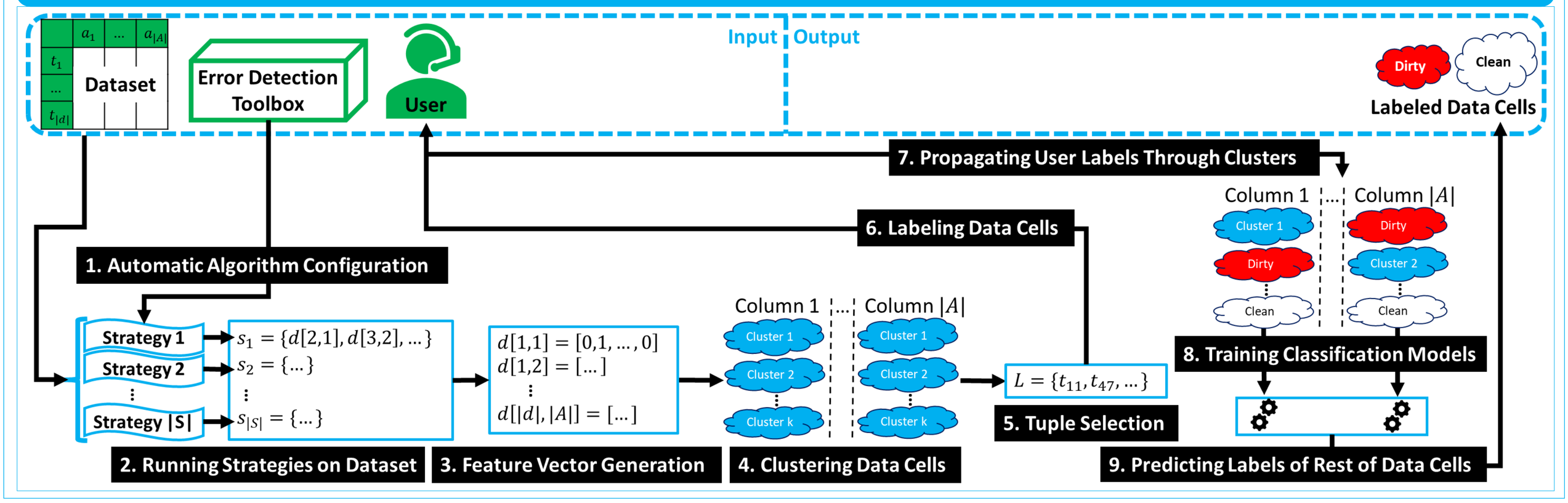
ID	Lord	Kingdom
1	Aragorn	Gondor
2	Gandalf	Ø
3	Saruman	Ø
4	Théoden	Shire

Research Question

- ❑ Given a dataset and a set of error detection algorithms, how can we accurately detect data errors without involving the user heavily in
  - Algorithm selection
    - The user should not select the promising algorithms
  - Algorithm configuration
    - The user should not provide any rules or parameters
  - Training data creation
    - The user may only label a few data values

Raha detects data errors with fewer than 20 labeled tuples due to its expressive feature representation and clustering-based sampling.

The Workflow of Raha



Experimental Overview

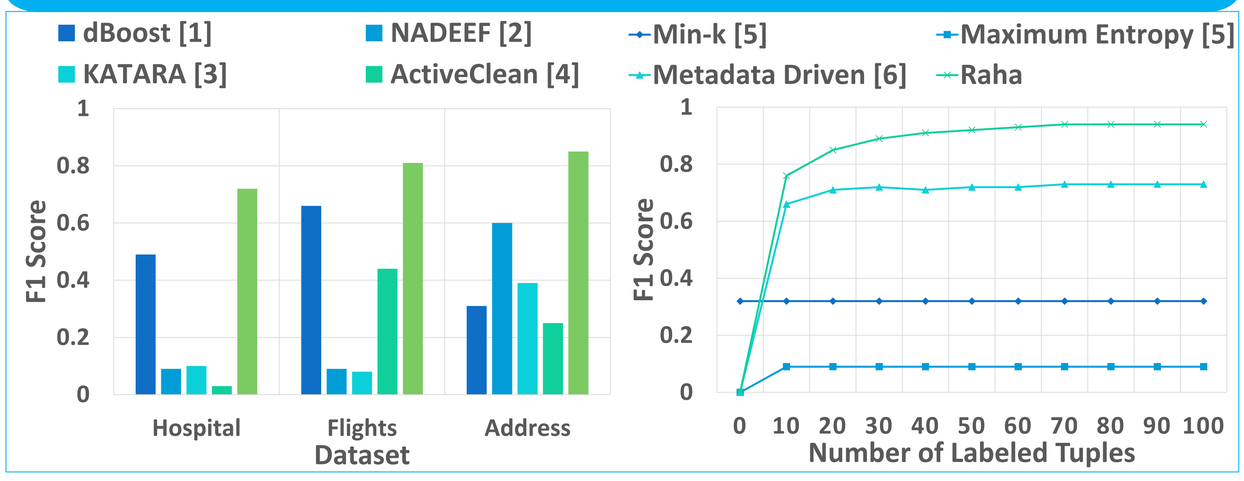
8 Datasets
- Hospital
- Flights
- Address
- Beers
- Rayyan
- Movies
- IT
- Tax

7 Baselines
- dBoost [1]
- NADEEF [2]
- KATARA [3]
- ActiveClean [4]
- Min-k [5]
- PBOS [5]
- MDED [6]

5 Evaluation Measures
- Precision
- Recall
- F_1 Score
- Runtime
- Labeled Tuples

10+ Experiments
- Performance
- Features
- Sampling
- Strategy Filtering
- User Labeling Error
- Scalability
- Classification Model

Raha Versus Baselines: Performance



References

[1] Clement Pit-Claudel et al. 2016. Outlier detection in heterogeneous datasets using automatic tuple expansion. Technical Report. CSAIL, MIT.
[2] Michele Dallachiesa et al. 2013. NADEEF: a commodity data cleaning system. In SIGMOD. 541–552.
[3] Xu Chu et al. 2015. Katara: A data cleaning system powered by knowledge bases and crowdsourcing. In SIGMOD. 1247–1261.
[4] Sanjay Krishnan et al. 2016. Activeclean: Interactive data cleaning for statistical modeling. PVLDB 9, 12, 948–959.
[5] Ziawasch Abedjan et al. 2016. Detecting data errors: Where are we and what needs to be done? PVLDB 9, 12, 993–1004.
[6] Larysa Visengeriyeva and Ziawasch Abedjan. 2018. Metadata-driven error detection. In SSDBM. 1–12.

Source Code

Our system is available online:
https://github.com/bigdama/raha

Acknowledgement

This project has been supported by the German Research Foundation (DFG) under grant agreement 387872445.

Deutsche Forschungsgemeinschaft