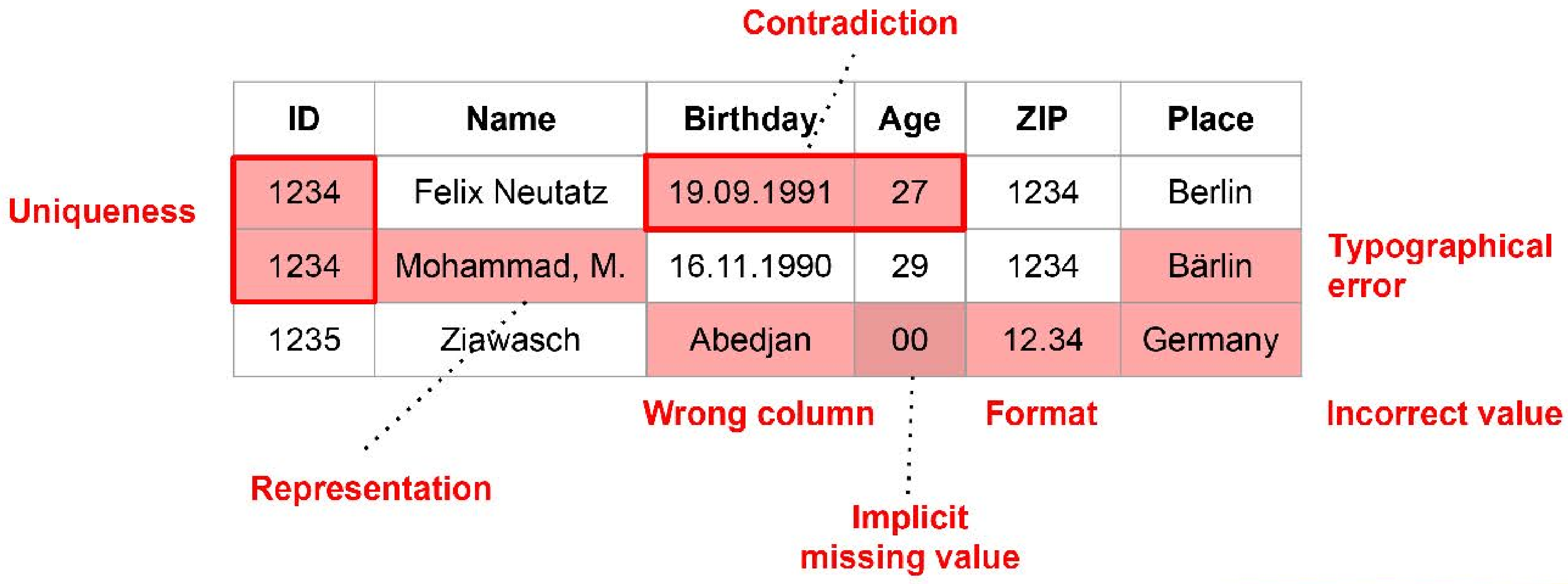
ED2: A Case for Active Learning in Error Detection

Felix Neutatz[1], Mohammad Mahdavi[2], and Ziawasch Abedjan[1,2]

[1] DFKI GmbH
felix.neutatz@dfki.de
[2] TU Berlin
{mahdavilahijani, abedjan}@tu-berlin.de

Motivation

Data scientists spend 80% of their time on data preparation. Error detection is a main part:



ID	Name	Birthday	Age	ZIP	Place
1234	Felix Neutatz	19.09.1991	27	1234	Berlin
1234	Mohammad, M.	16.11.1990	29	1234	Bärlin
1235	Ziawasch	Abedjan	00	12.34	Germany

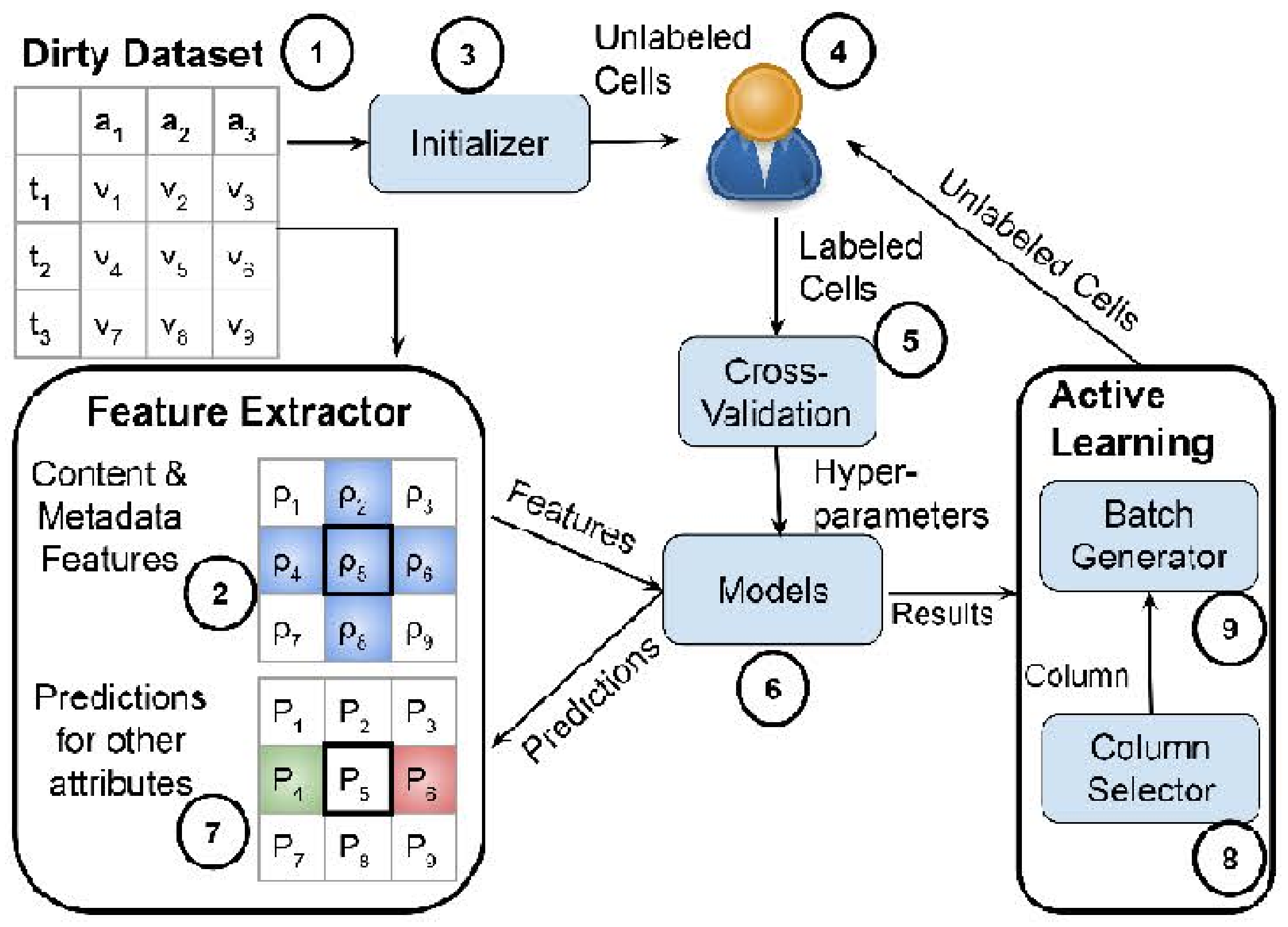
Contradiction
Uniqueness
Typographical error
Wrong column
Format
Incorrect value
Representation
Implicit missing value

Research Questions

- How can we reduce the user labeling effort for classification-based error detection methods [1,2]?
- How can we apply active learning for a two-dimensional relational table?

ED2 achieves state-of-the-art detection accuracy while two-stage active learning reduces the labeling effort by one order of magnitude for large datasets.

ED2 Architecture



1. Column Selection

$$\underset{\text{column}}{\arg\min} \sum_{v \in \text{column}} |P_{\text{column}}(y = \text{erroneous}|v) - 0.5|$$

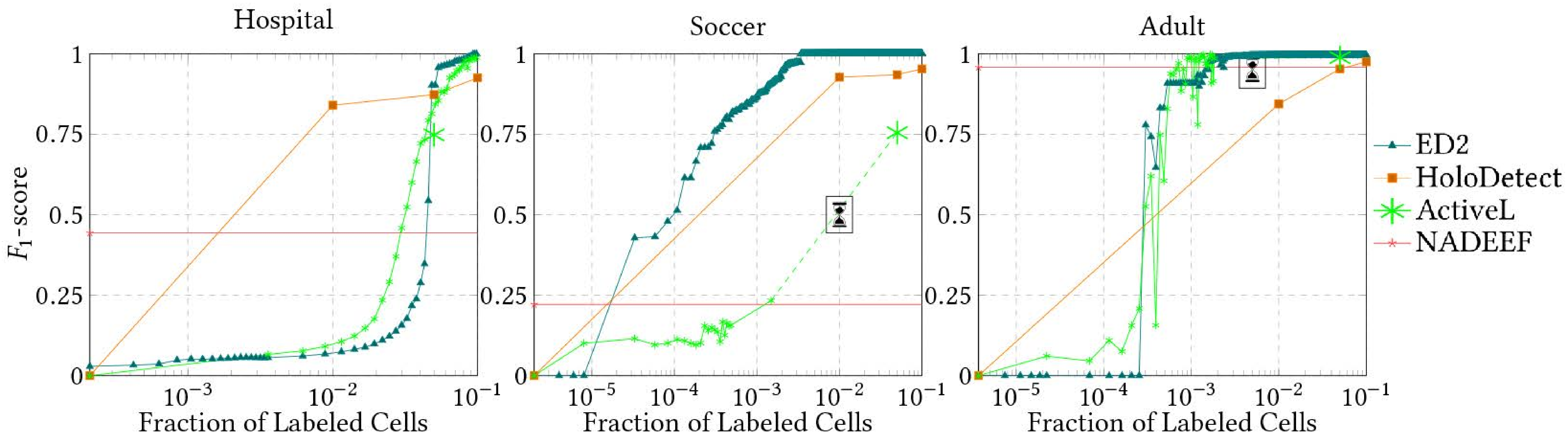
2. Batch Selection

$$\underset{V' \subset V, |V'| = k}{\arg\min} \sum_{v \in V'} |P(y = \text{erroneous}|v) - 0.5|$$

Experiments

Table 1: Experimental datasets.

Dataset	Columns	Rows	Errors
Hospital	19	1,000	2.65%
Soccer	10	200,000	1.56%
Adult	11	97,684	0.10%



[1] Visengeriyeva, L. et al. 2018. Metadata-Driven Error Detection. SSDBM.
[2] Heidari, A. et al. 2019. HoloDetect: Few-Shot Learning for Error Detection. SIGMOD.
[3] Dallachiesa, M. et al. 2013. NADEEF: a commodity data cleaning system. SIGMOD.

Open Source

Our system is available online:
https://github.com/BigDaMa/ExampleDrivenErrorDetection

Acknowledgement

This work was supported by the German Federal Ministry of Transport and Digital Infrastructure - project DAYSTREAM (19F2031A).

Federal Ministry of Transport and Digital Infrastructure