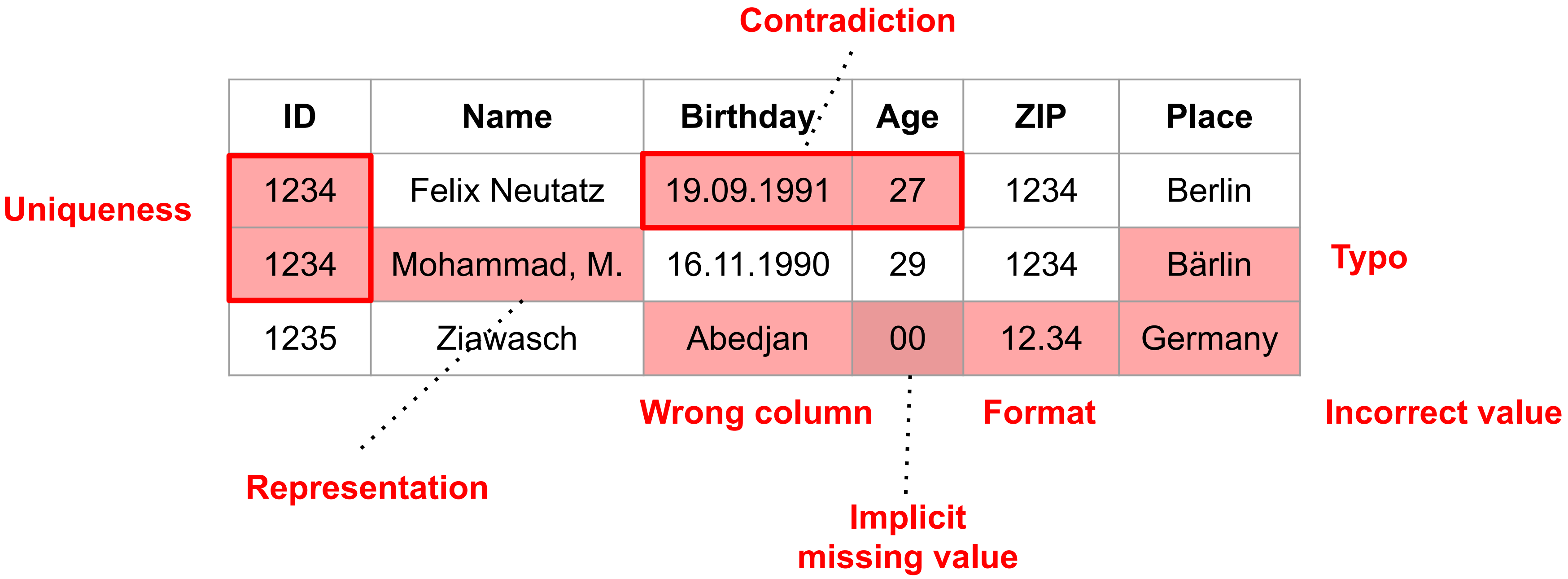
ED2: A Case for Active Learning in Error Detection

Felix Neutatz, Mohammad Mahdavi, Ziawasch Abedjan

November 06, 2019

The 28th ACM International Conference on Information and Knowledge Management

What Are Errors?



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

Typo

Wrong column

Format

Incorrect value

Representation

Implicit
missing value

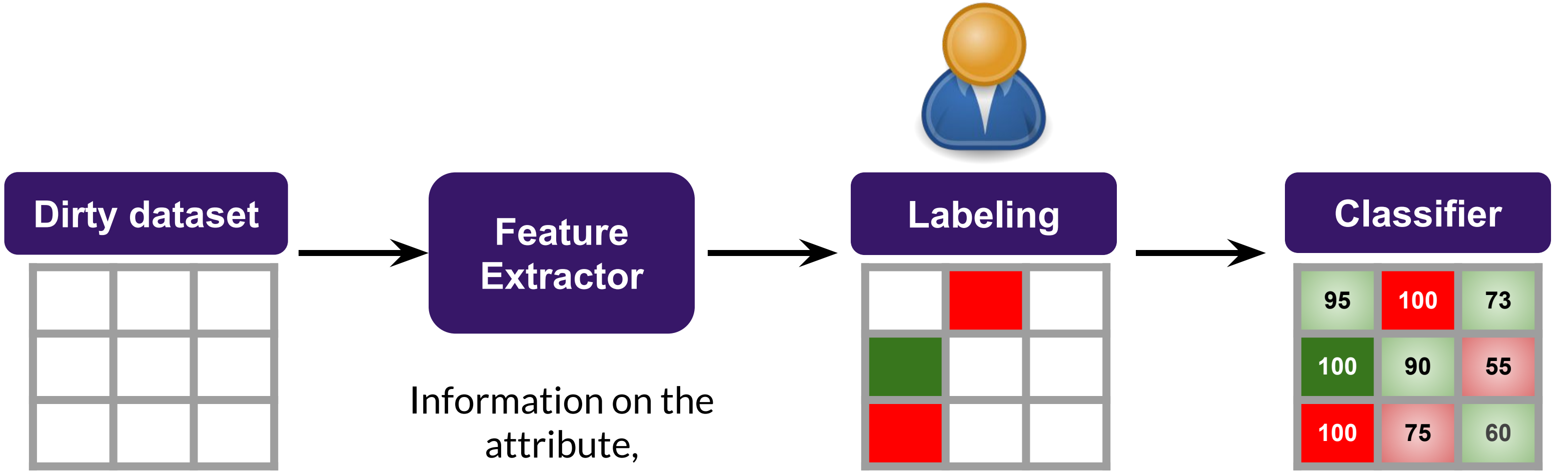
Can We Apply ML to Detect Errors?

Is this cell value correct?

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

Visengeriyeva, Larysa et al. 2018. Metadata-Driven Error Detection. SSDBM.
Heidari, Alireza et al. 2019. HoloDetect: Few-Shot Learning for Error Detection. SIGMOD.

Error Detection: Cell-Wise Classification



	🟥	
🟩		
🟥		

95	100	73
100	90	55
100	75	60

Information on the attribute,
tuple, and dataset level
for each data cell

Predictions
with certainty

Visengeriyeva, Larysa et al. 2018. Metadata-Driven Error Detection. SSDBM.
Heidari, Alireza et al. 2019. HoloDetect: Few-Shot Learning for Error Detection. SIGMOD.

How Many Labels Do We Need?

Methods	**Required Labels**
Visengeriyeva, Larysa et al. 2018. Metadata-Driven Error Detection. SSDBM.	1 %
Heidari, Alireza et al. 2019. HoloDetect: Few-Shot Learning for Error Detection. SIGMOD.	1 - 10 %

In the age of Big Data, 1% means a lot of labeling effort!

How Can We Apply Active Learning in 2D?

Which column should be labeled next?

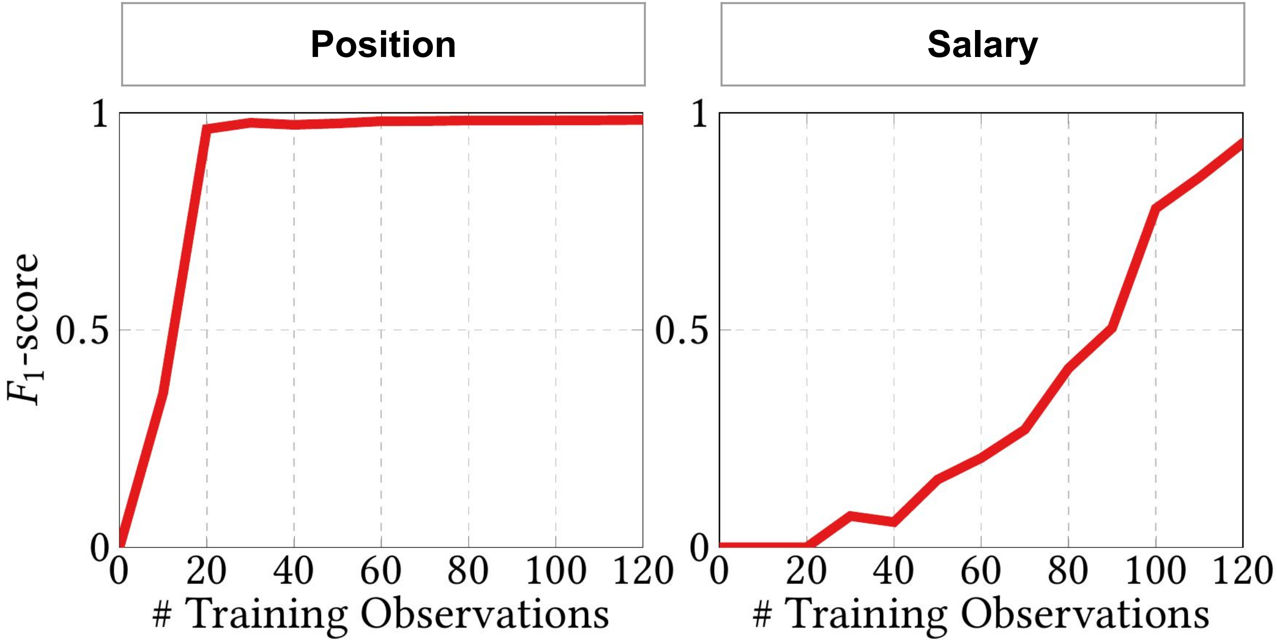
Position	Salary
senior_manager	10,000
senior accountant	5,000
junior engineer	4,000
senior accountant	11,000
senior_legal_counsel	6,000

Which data cells within a column should be labeled?

Syntax forbids '_'.

Senior staff earns more than 9,000.

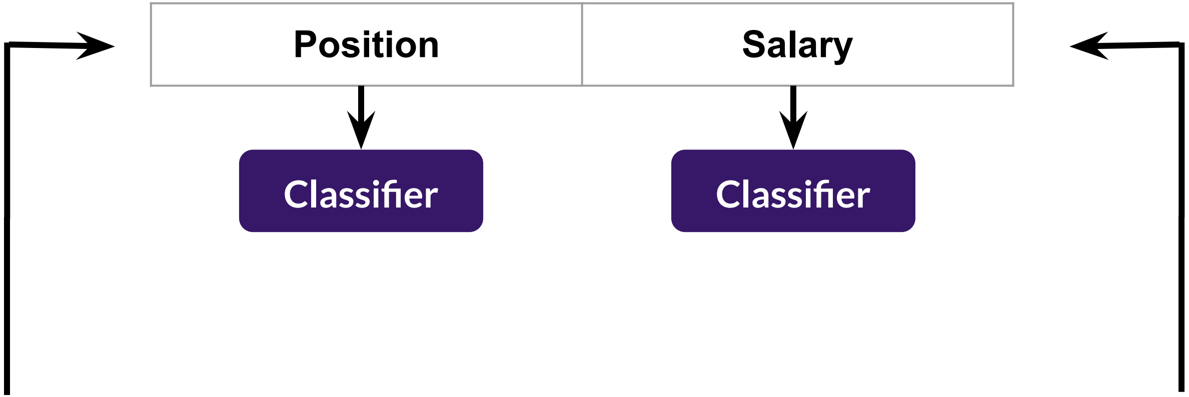
Why Does Column Selection Matter?



Syntax Error:
Position has '_'

Semantic Error:
Position has 'senior' &
Salary < 9,000

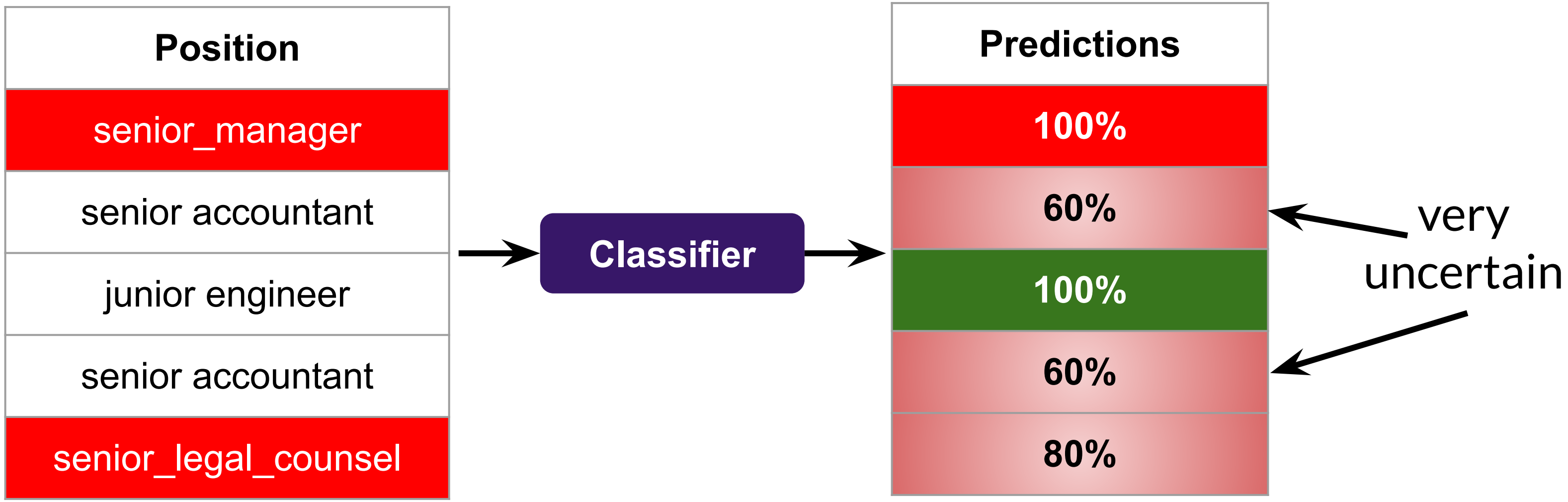
First Stage: Which Column Should Be Labeled?



$$\underset{\text{column}}{\arg\min} \underbrace{\sum_{v\in\text{column}} \left|P_{\text{column}}(y = \text{erroneous}|v) - 0.5\right|}_{\text{summed certainty across all values per column}}$$

Second Stage: Which Data Cells Should Be Labeled?

We let the user label the **k** most uncertain data cells with distinct values:

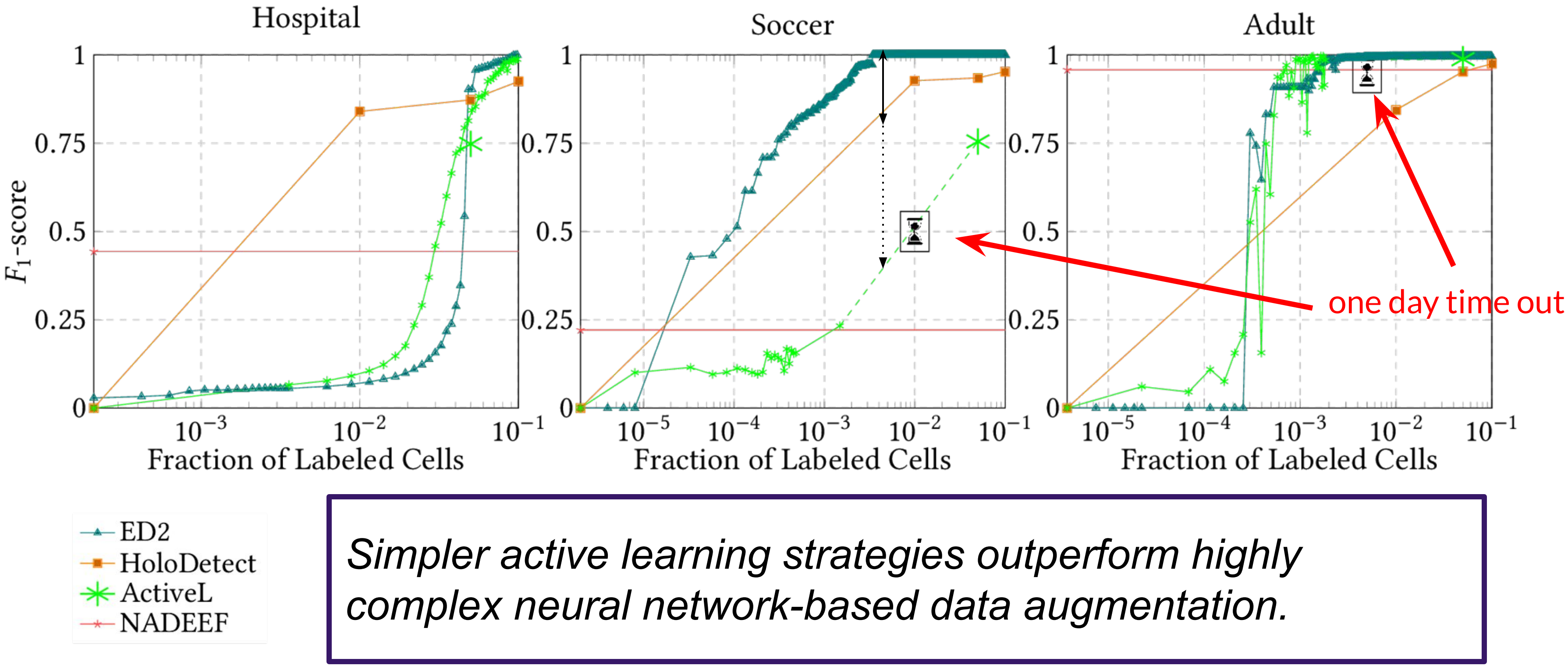


Position
senior_manager
senior accountant
junior engineer
senior accountant
senior_legal_counsel

Predictions
100%
60%
100%
60%
80%

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

Two-Stage Active Learning at Work



Simpler active learning strategies outperform highly complex neural network-based data augmentation.

Conclusion

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.

Source code is available here:

https://github.com/BigDaMa/ExampleDrivenErrorDetection