Towards Automated Data Cleaning Workflows

Mohammad Mahdavi, Felix Neutatz, Larysa Visengeriyeva, Ziawasch Abedjan

October 02, 2019

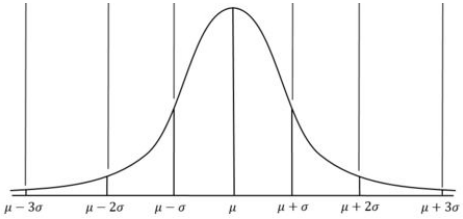
Learning. Knowledge. Data. Analytics. 2019

How Data Science Starts...

ID	Name	Birthday	Age	ZIP	Place
1234	Franz Müller	19.09.1991	27	1234	Berlin
1235	Jane Schmidt	-	9999	1234	Bärlin
1234	Lara	Weber	126	12.34	Germany

Credit_Approval
YES
NO
YES

Data Cleaning



Outlier

Contradiction

Uniqueness

ID	Name	Birthday	Age	ZIP	Place	Credit_Approval
1234	Franz Müller	19.09.1991	27	1234	Berlin	YES
1235	Jane Schmidt	-	9999	1234	Bärlin	NO
1234	Lara	Weber	265	12.34	Germany	YES

Format

Misfielded Value

Incorrect value

Typographical error

Missing value

People work on data cleaning for a very long time...

Brodie, M.L., **1980**. Data quality in information systems. *Information & Management*.

Rahm, E. and Do, H.H., **2000**. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*.

... but in 2019, people still write custom cleaning scripts!

Data Cleaning

ID	Name	Birthday	Age	ZIP	Place
1234	Franz Müller	19.09.1991	27	1234	Berlin
1235	Jane Schmidt	-	9999	1234	Bärlin
1234	Lara	Weber	265	12.34	Germany

Credit_Approval
YES
NO
YES

Missing value

Data Cleaning

ID	Name	Birthday	Age	ZIP	Place
1234	Franz Müller	19.09.1991	27	1234	Berlin
1235	Jane Schmidt	-	9999	1234	Bärlin
1234	Lara	Weber	265	12.34	Germany

Credit_Approval
YES
NO
YES

Missing value

Data Cleaning

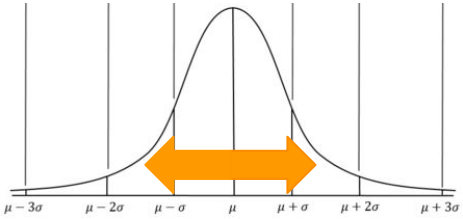
ID	Name	Birthday	Age	ZIP	Place
1234	Franz Müller	19.09.1991	27	1234	Berlin
1235	Jane Schmidt	**05.08.1991**	**28**	1234	Bärlin
1234	Lara	Weber	265	12.34	Germany

Credit_Approval
YES
NO
YES

Knowledge Base-Driven Correction

Rule-Driven Correction

Data Cleaning



Outlier

ID	Name	Birthday	Age	ZIP	Place
1234	Franz Müller	19.09.1991	27	1234	Berlin
1235	Jane Schmidt	**05.08.1991**	**28**	1234	Bärlin
1234	Lara	Weber	265	12.34	Germany

Credit_Approval
YES
NO
YES

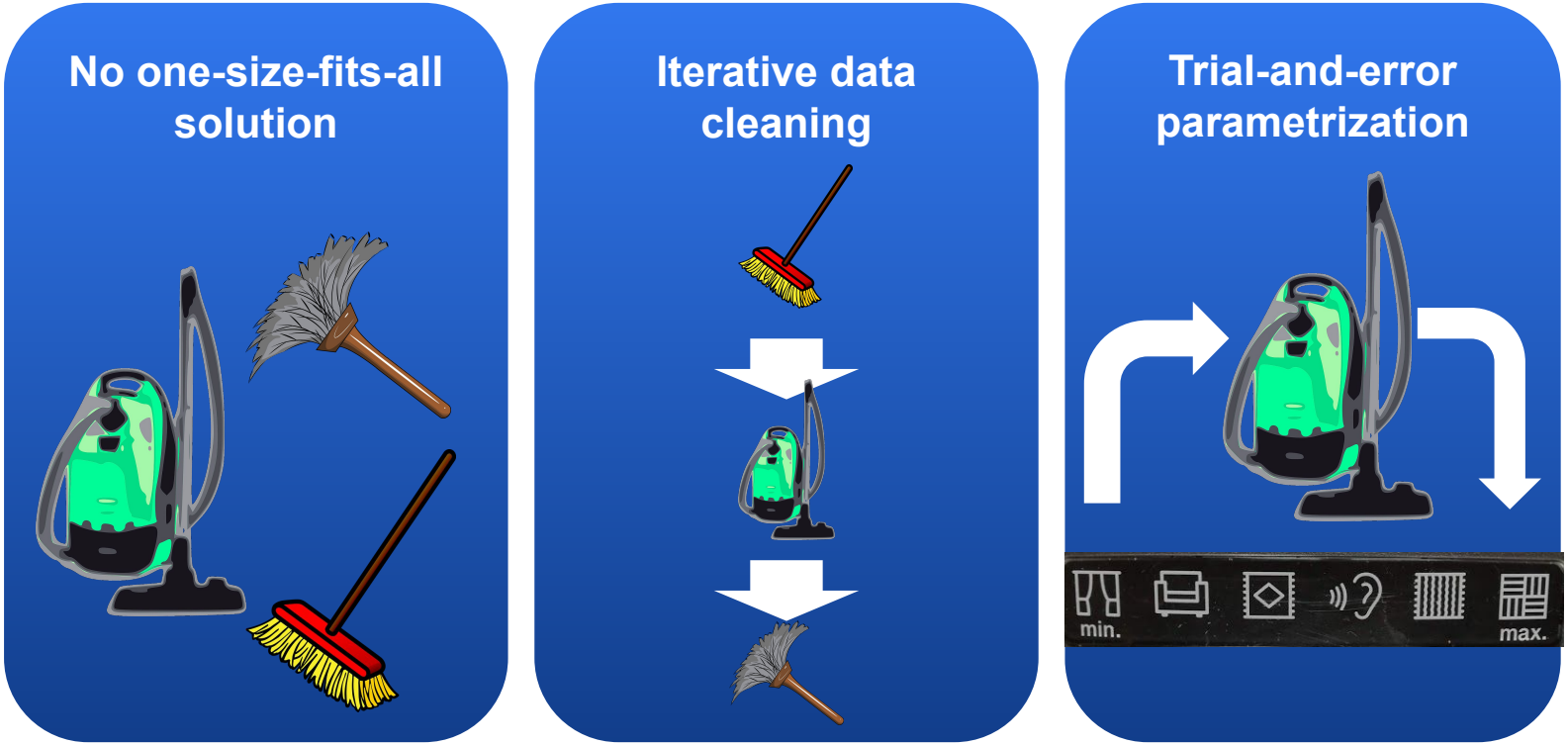
Data Cleaning

ID	Name	Birthday	Age	ZIP	Place
1234	Franz Müller	19.09.1991	27	1234	Berlin
1235	Jane Schmidt	**05.08.1991**	**28**	1234	Bärlin
1234	Lara	Weber	**29**	12.34	Germany

Credit_Approval
YES
NO
YES

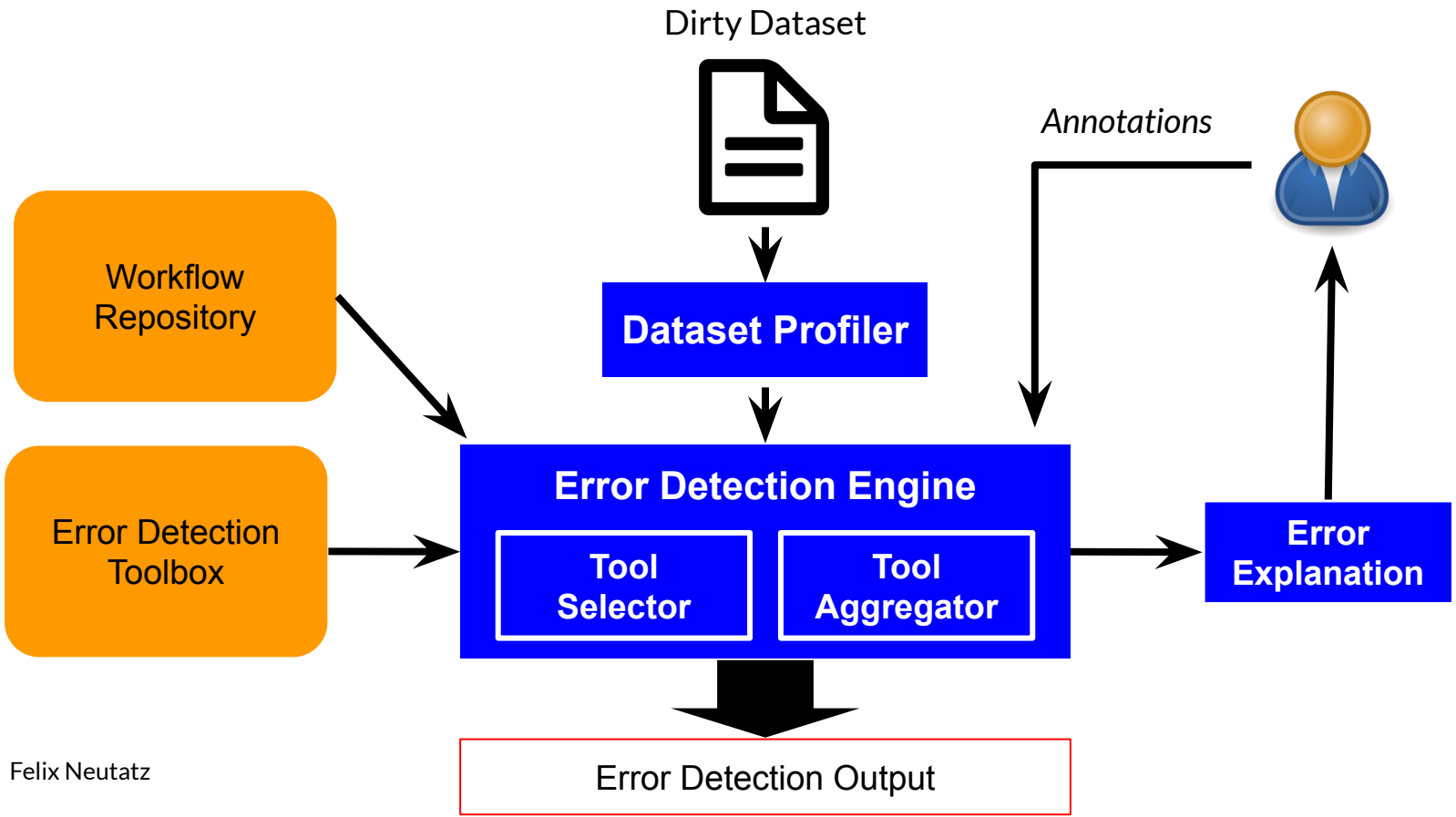
Imputation / Regression

Data Cleaning: The Usage Gaps

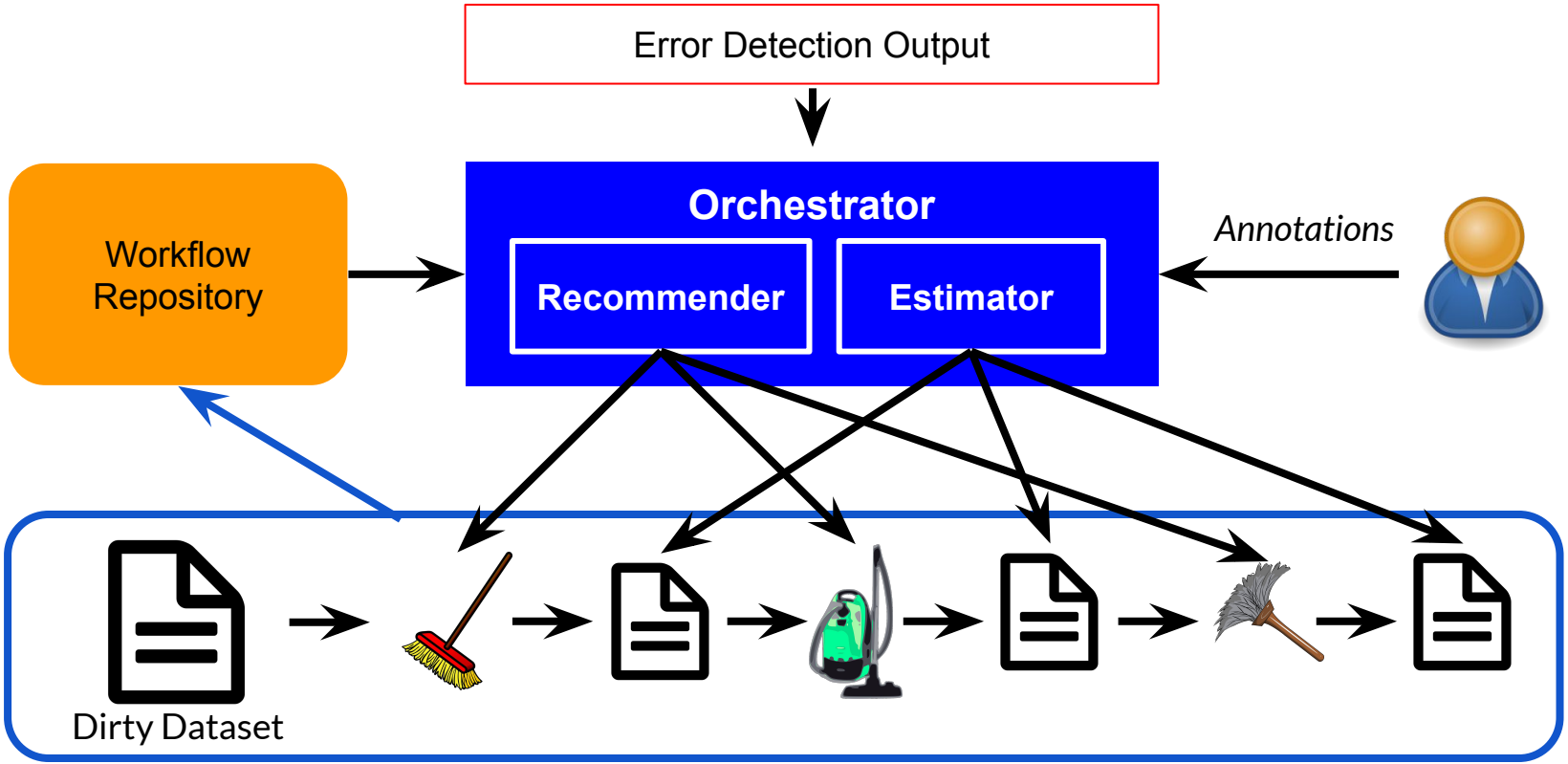


We need a workflow orchestrator that learns from previous tasks to propose promising data cleaning workflows for a new dataset.

Vision: Cleaning Workflow Orchestrator



Vision: Cleaning Workflow Orchestrator



Cleaning Workflow

Current Status

MDED, a system that learns to aggregate error detection strategies via metadata

Larysa Visengeriyeva et al.

SSDBM`18

http://bit.ly/systems-aggregation

REDS, a system that estimates the performance of error detection strategies via metadata

Mohammad Mahdavi et al.

SSDBM`19

https://github.com/bigdama/reds

ED2, an active learning-driven error detection system

Felix Neutatz et al.

CIKM`19

http://bit.ly/2mjyiTO

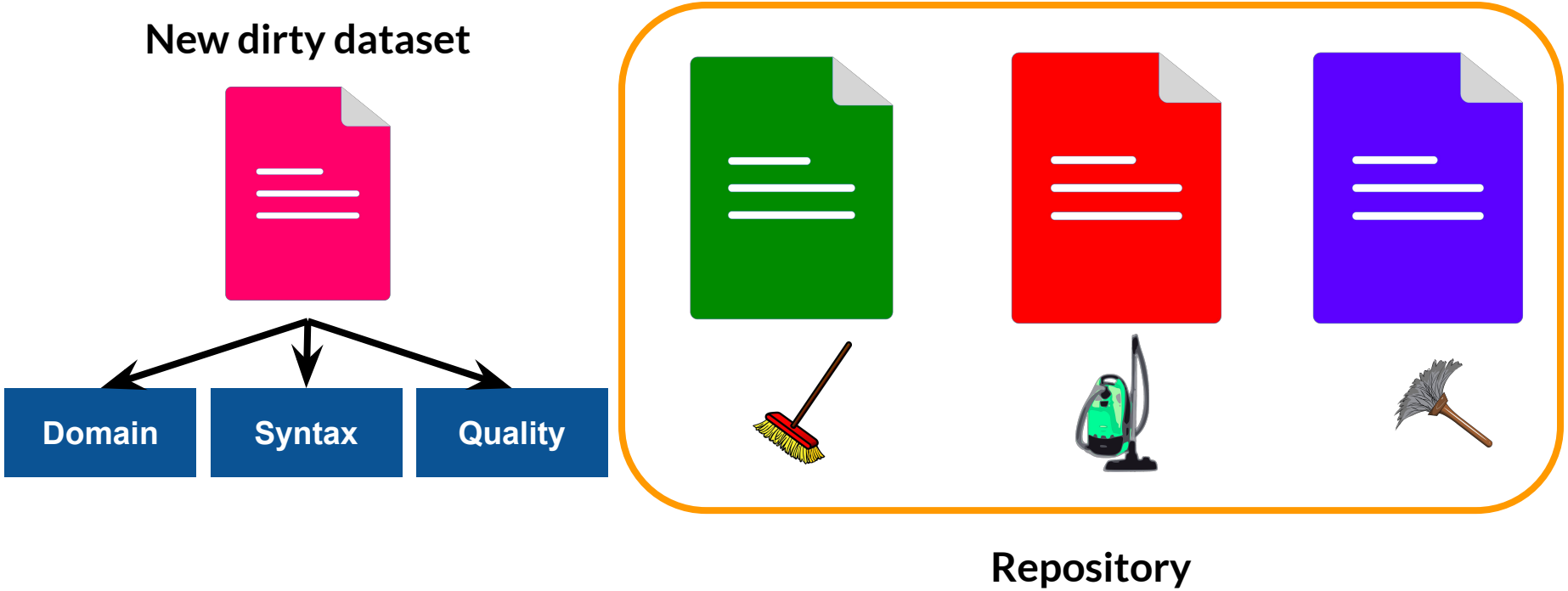
Raha, a configuration-free error detection system to detect data errors holistically

Mohammad Mahdavi et al.

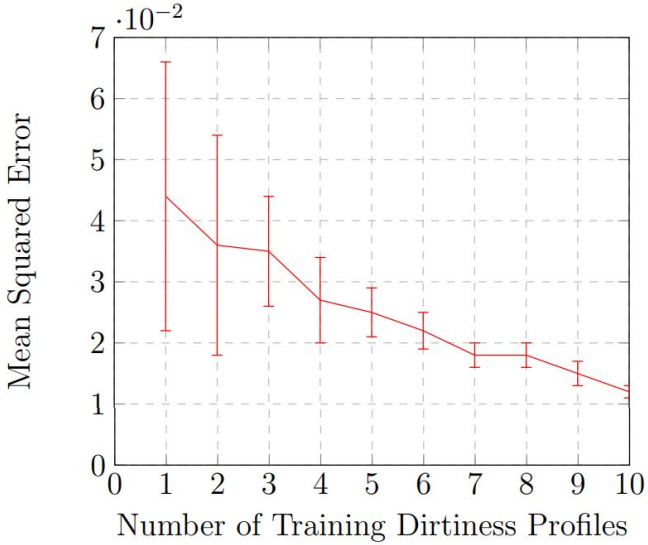
SIGMOD`19

https://github.com/BigDaMa/raha

Configuration-Free Tool Selection



Configuration-Free Tool Selection

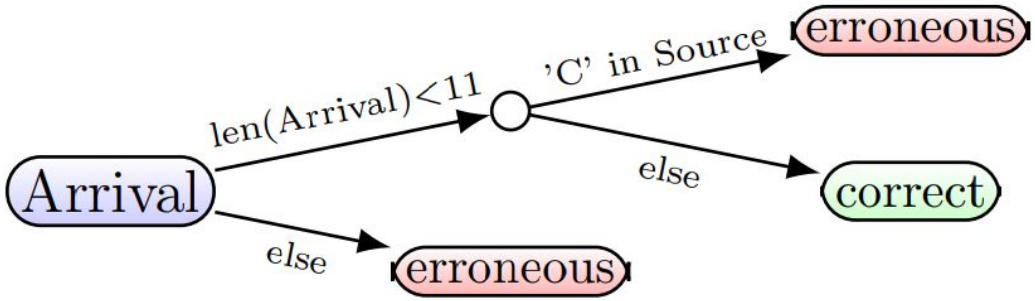


The more datasets in our repository, the higher is the accuracy of estimating the performance of cleaning strategies on a new dataset.

Error Explanation

Source	Flight	Departure	Arrival
A	LH-100	10:00 a.m.	**2:00 p.m. Dec 02**
B	LH-100	10:00 a.m.	2:00 p.m.
C	LH-100	10:00 a.m.	**3:00 p.m.**

Given some user labels and a fine-grained feature represenation, we can learn the error patterns:



How Do We Get Labels?

Two-stage Active Learning

City	ZIP
BERLIN	10247
BERLIN	10249
BÄRLIN	10247
Augsburg	86I99

Clustering-based Label Propagation

Future Directions

- Recommending sequences of multiple repair methods or even more complex control flow.
- Understanding metadata:
  Mapping metadata to data quality issues.
- Learning to transform data values:
  Machine translation of dirty data to clean data.
  Especially out-of-domain errors are hard to correct.

Takeaways

- Gaps between theory and practice:
  - No one-size-fits-all solution
  - Iterative data cleaning
  - Trial-and-error parametrization
- We need a workflow orchestrator that learns to propose promising data cleaning workflows.
- Towards this vision, we developed a cleaning strategy selector, a cleaning strategy aggregator, and an error explanation system.

Towards Automated Data Cleaning Workflows

MDED, a system that learns to aggregate error detection strategies via metadata

Larysa Visengeriyeva et al.

SSDBM`18

http://bit.ly/systems-aggregation

REDS, a system that estimates the performance of error detection strategies via metadata

Mohammad Mahdavi et al.

SSDBM`19

https://github.com/bigdama/reds

ED2, an active learning-driven error detection system

Felix Neutatz et al.

CIKM`19

http://bit.ly/2mjyiTO

Raha, a configuration-free error detection system to detect data errors holistically

Mohammad Mahdavi et al.

SIGMOD`19

https://github.com/BigDaMa/raha