RE for Data Cleaning with Machine Learning
CS 846

Presenter: Ishank Jain

UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS

OUTLINE

- Motivation
- Introduction
- Challenges
- Related Work
- Conclusion
- Questions ??

Sources

- ACM: SIGMOD
- VLDB
- CIDR: Conference on Innovative Data Systems Research
- STACS: Symposium on Theoretical Aspects of Computer Science

MOTIVATION

Databases can be **corrupted with various errors** such as missing (NULL, nan etc.), incorrect, or **inconsistent values**. An incorrect or **inconsistent data** can **lead** to false conclusions and **misdirected decisions**.

INTRODUCTION

The process of ensuring that **data adheres to desirable quality and integrity** is referred to as **data cleaning**, is **a major challenge** in most data-driven applications.

In this presentation, we will look at the requirements to perform data cleaning using machine learning techniques.

We will look at various tools such ActiveClean, BoostClean, Holoclean, and Tamr.

RELATED WORK

- Rule-based detection algorithms, such as **FDs, CFDs, and MDs**, and those have always been studied in isolation. Such techniques are usually applied in a **pipeline or interleaved**.
- Pattern enforcement and transformation tools such as OpenRefine. These tools discover patterns in the data, either syntactic or semantic, and use these to detect errors.
- Quantitative error detection algorithms that expose outliers, and glitches in the data.
- Record linkage and de-duplication algorithms for detecting duplicate data records, such as the Data Tamer system

REQUIRED CHARACTERISTICS

Scripting languages that are appropriate for skilled and unskilled programmers.

Systems will need to have automated algorithms with human help only when necessary.

New data sources must be integrated incrementally as they are uncovered.


CHALLENGES

CHALLENGES

Correctness

Dirty data identification

CHALLENGES

Synthetic data and errors: The lack of real data sets (along with ground truth) or a widely accepted benchmark makes it hard to judge the effectiveness

Human involvement: To verify detected errors, to specify cleaning rules, or to provide feedback that can be part of a machine learning algorithm

EXAMPLE APPLICATION

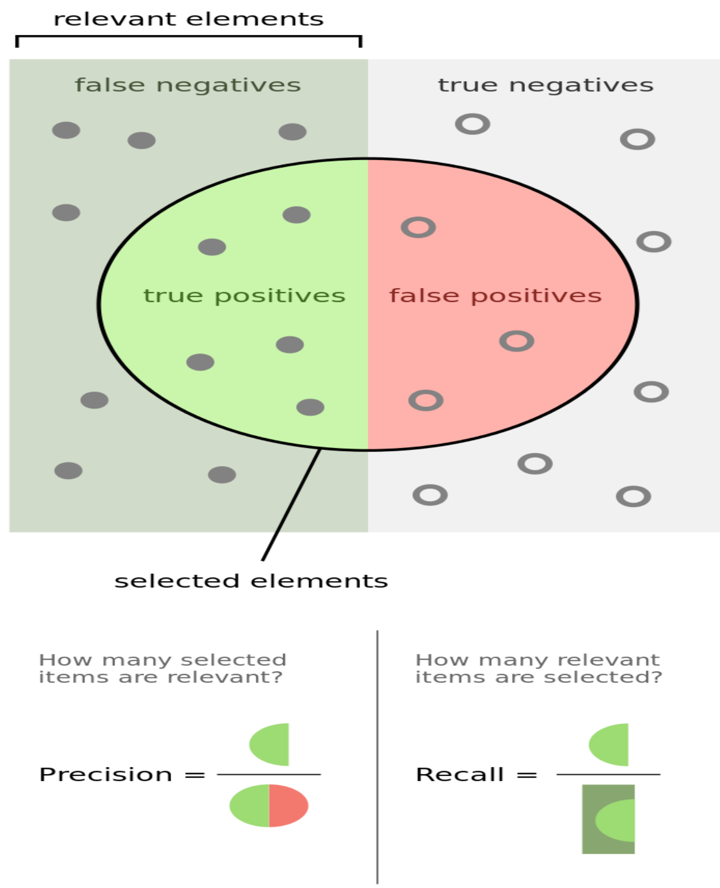
- **Health Services Application:** integrated database contains millions of records, and to consolidate claims data by medical provider. In effect, they want to de-dup their database, using a subset of the fields.
- **Web Aggregator:** integrates about URLs, collecting information on things to do" and events. Events include lectures, concerts, and live music at bars.
- **Hospital records:** medical records from different hospital branches needs to be integrated together.

REQUIREMENTS

- Datasets:
  - Training data
  - Clean data
  - Test data
- Rules and constraints to detect dirty cells.
- Machine learning architecture: this may include
  - Clustering algorithm to detect outliers, dirty cells. For instance ActiveClean, Tamr.
  - Neural network based algorithm which is trained on a feature graph model to generate potential domain, for instance, HoloClean.
  - Classification and boosting algorithm (SVM, Naïve Bais etc.) to assign the correct class label from the domain based on a loss minimization function or to detect duplicates, for instance, BoostClean and Tamr.

REQUIREMENTS

- Evaluation metrics:
  - Precision
  - Recall
  - Accuracy (sometimes)
  - F1 score (sometimes)



SETUP

- Input is a dirty training dataset which has training attributes and labels, where both the features X_{train} and labels Y_{train} may have errors, and test dataset (X_{test}, Y_{test}).
- Detection generator such as boolean expressions like FD's or outlier detection algorithm to find dirty data, duplicates, and missing data.
- Repair function which modifies the record's attributes based on domain to correct the dirty data.

SETUP: Detectors

- The ability for a data cleaning system to accurately identify data errors relies on the availability of a set of high-quality error detection rules.
- Different frameworks use different detector functions:
  1. Rules-based (for instance, Denial constraints in HoloClean),
  2. Use of classification algorithms to detect outliers like in BoostClean.

UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS

SETUP: Detectors

Rule-based data cleaning systems rely on data quality rules to detect errors. Data quality rules are often expressed using integrity constraints, such as functional dependencies or denial constraints.

$$c_1: \neg\big(G(g,f,n,r,c,a,s), G(g',f',n',r',c',a',s'), (c=c'), (s \neq s')\big)$$

$$c_2: \neg\big(G(g,f,n,r,c,a,s), G(g',f',n',r',c',a',s'), (r=r'), (c=\text{“}NYC\text{”}), (c' \neq \text{“}NYC\text{”}), (s' > s)\big)$$

LocalEmployeesSJ (L)

	LID	FN	LN	RNK	DO	Y	CT	MID	SAL
t_1	1	Paul	Smith	A	2	5	SJ	1	100
t_2	2	Mark	White	B	5	8	SJ	1	80

GlobalEmployees (G)

	GID	FN	LN	ROLE	CITY	AC	ST	SAL
t_3	102	Paul J.	Smith	V	SJ	639	CA	100
t_4	105	Anne	Nash	M	NYC	234	NY	110
t_5	211	Mark	White	E	SJ	639	CA	80
t_6	386	Mark	Lee	E	NYC	552	AZ	75

SETUP: Detectors

Use of classification algorithms to detect outliers like in BoostClean.

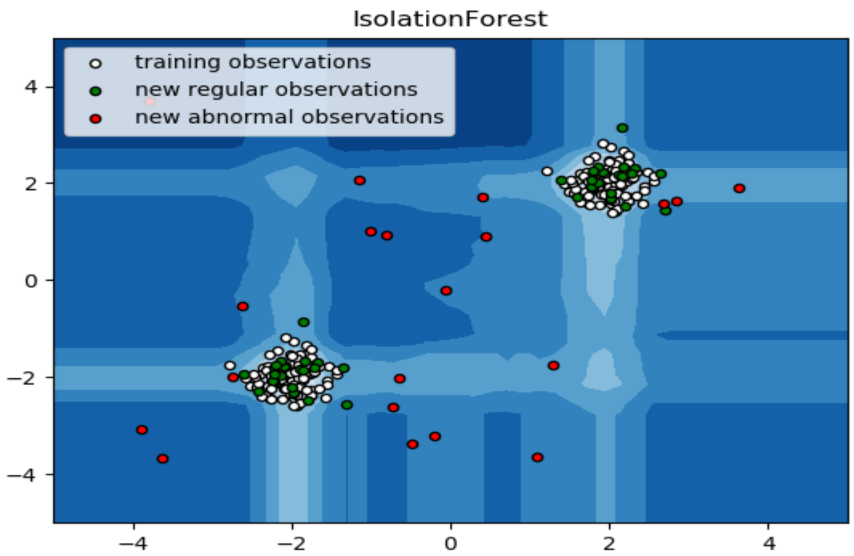
Isolation Forests. The Isolation Forest is inspired by the observation that outliers are more easily separable from the rest of the dataset than non-outliers.

The **length of the path to the leaf** node is a measure for the **outlierness** of the record—a shorter path more strongly suggests that the record is an outlier.

Isolation Forests have a **linear time complexity** and very small memory requirements. Isolation Forest provided the best trade-off between runtime and accuracy.

SETUP: Detectors

Random partitioning produces noticeable shorter paths for anomalies. Hence, when a forest of random trees collectively produce shorter path lengths for particular samples, they are highly likely to be anomalies.



SETUP: Detectors

Correlation clustering algorithm used in Tamr to detect duplicate tuples.

- The algorithm starts with all singleton clusters, and repeatedly merges randomly selected clusters that have a “connection strength" above a certain threshold.
- Tamr quantify the connection strength between two clusters as the number of edges across the two clusters over the total number of possible edges.

SETUP: Detectors

ActiveClean uses pointwise gradients to generalize the outlier filtering heuristics to select potentially dirty data even in complex models.

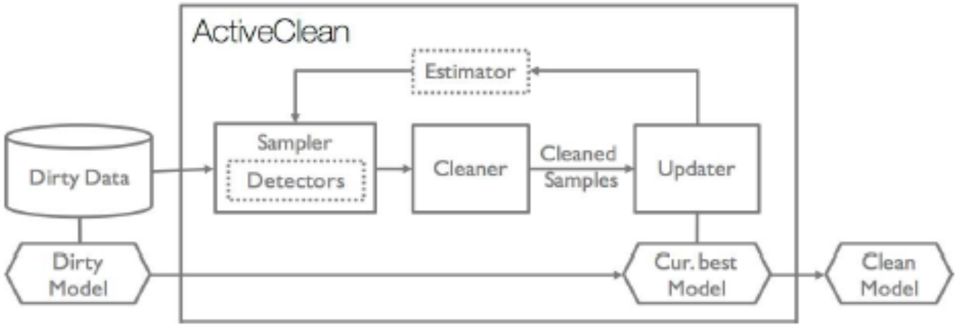
The cleaner (C) is as an oracle that maps a dirty example (x_i; y_i) to a clean example (x'_i; y'_i).

- Objective is a minimization problem that is solved with an algorithm called Stochastic Gradient Descent, which iteratively samples data, estimates a gradient, and updates the current best model.

$$\arg\min_{\theta \in \Theta} \sum_{i=1}^{N} \ell(C(x_i, y_i); \theta)$$

SETUP: Repair

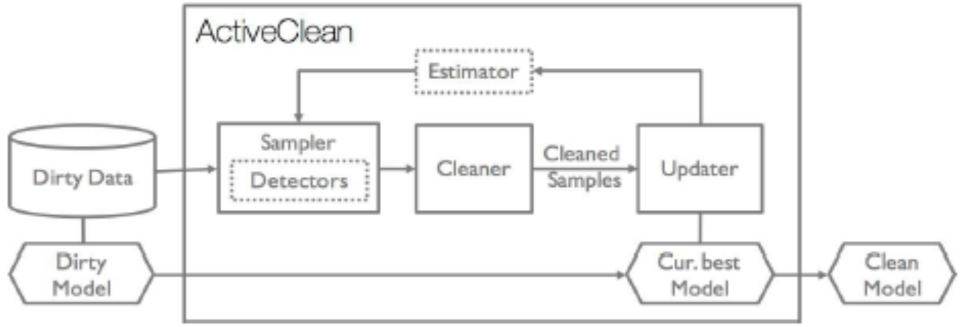
After the data sample is cleaned, ActiveClean updates the current best model, and re-runs the cross-validation to visualize changes in the model accuracy. At this point, ActiveClean begins a new iteration by drawing a new sampling of records to show the analyst.



SETUP: Repair

ActiveClean provides a Clean panel that gives the option to remove the dirty record, apply a custom cleaning operation (specified in Python), or pick from a pre-defined list of cleaning functions.

Custom cleaning operations are added to the library to help taxonomize different types of errors and reduce analyst cleaning effort.



SETUP: Repair

- BoostClean is pre-populated with a set of simple repair functions.
- Mean Imputation (data and prediction): Impute a cell in violation with the mean value of the attribute calculated over the training data excluding violated cells.
- Median Imputation (data and prediction): Impute a cell in violation with the median value of the attribute calculated over the training data excluding violated cells.

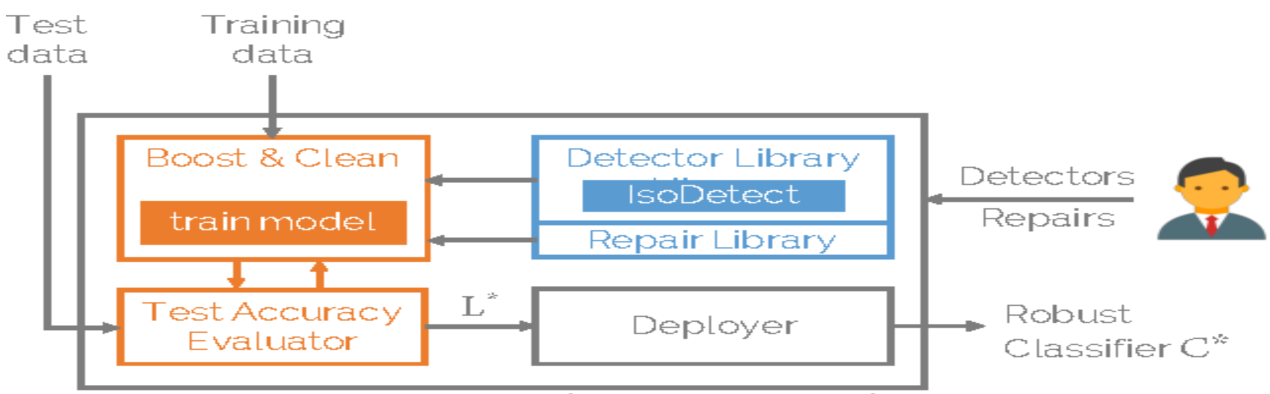


Figure 3: BoostClean system architecture.

UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS

SETUP: Repair

- Mode Imputation (data and prediction): Impute a cell in violation with the most frequent value of the attribute calculated over the training data excluding violated cells.
- Discard Record (data): Discard a dirty record from the training dataset.
- Default Prediction (prediction): Automatically predict the most popular label from the training data.

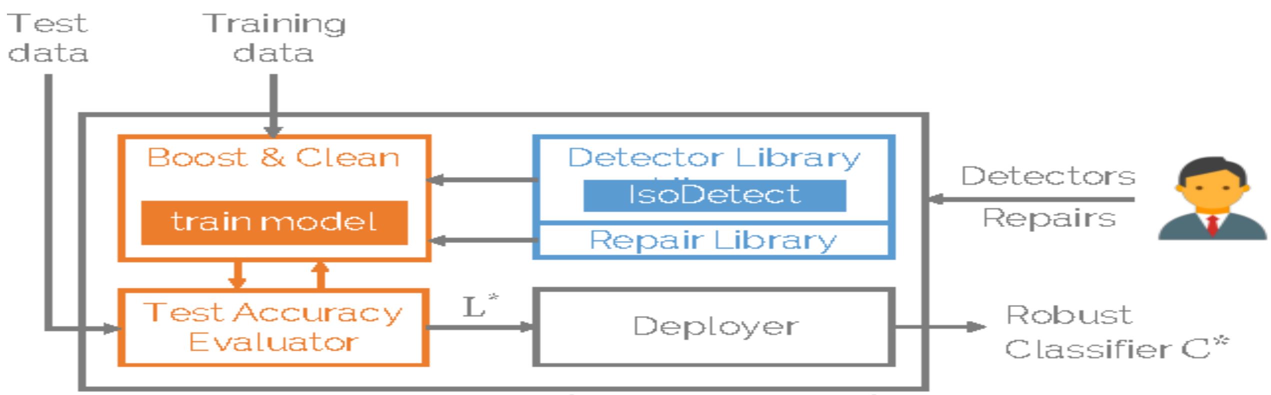
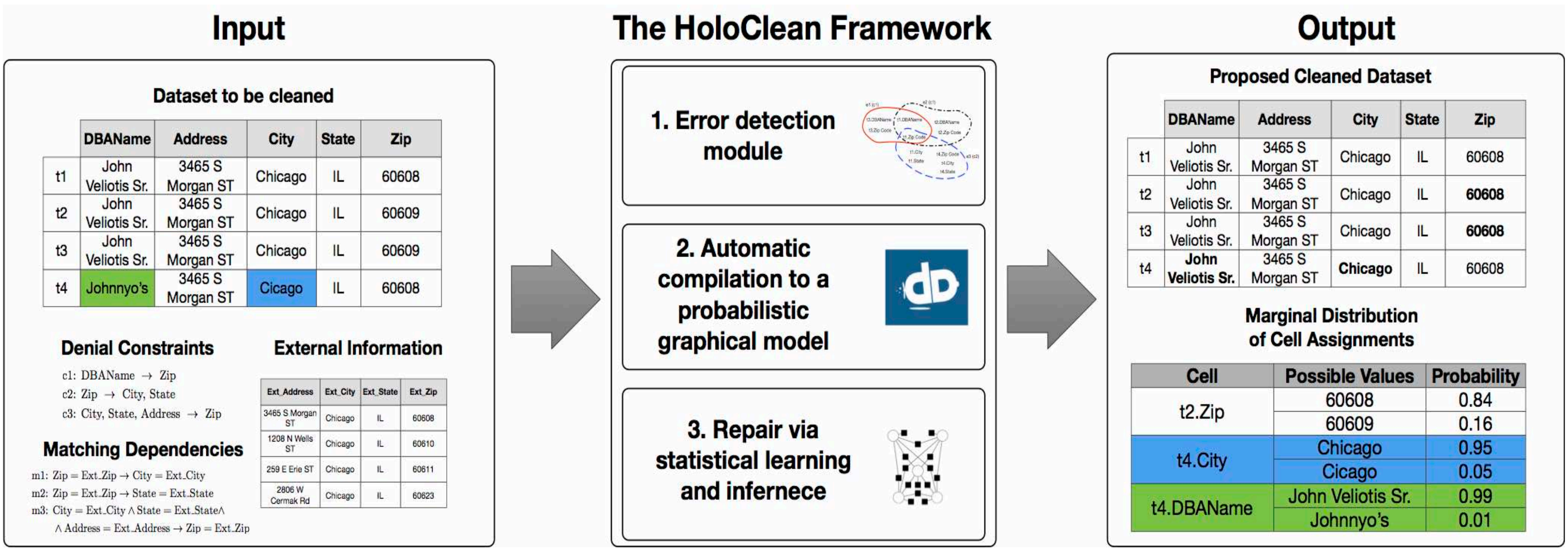


Figure 3: BoostClean system architecture.

SETUP: Repair



HoloClean Flow

Original Dataset

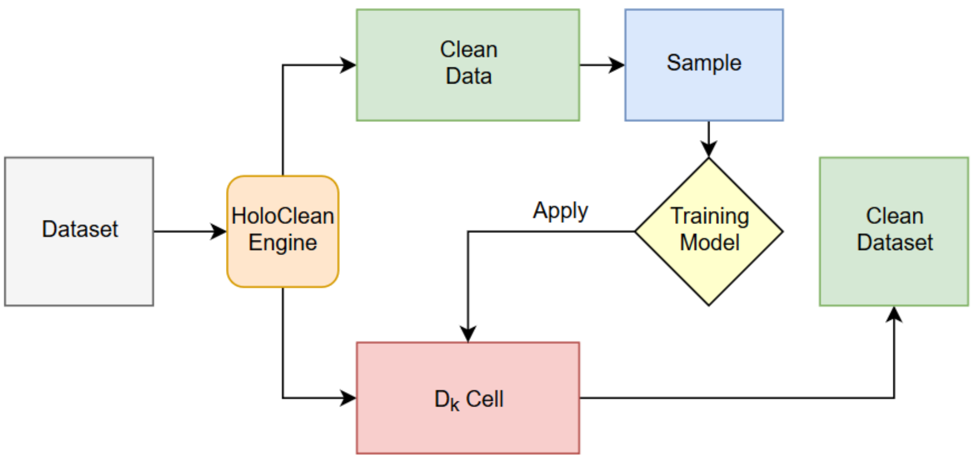
	A	B	C
t1	a1	b1	c1
t2	a1	b1	c2
t3	a2	b1	c3

Input Cell	Various features based on a cell's position				Label
t1.A = a1					1
t1.A = a2					0
t1.B = b1					1
t1.C = c1					1
t1.C = c2					0
t1.C =c3					0
t2.A = a1					1
…					…

Feature Matrix

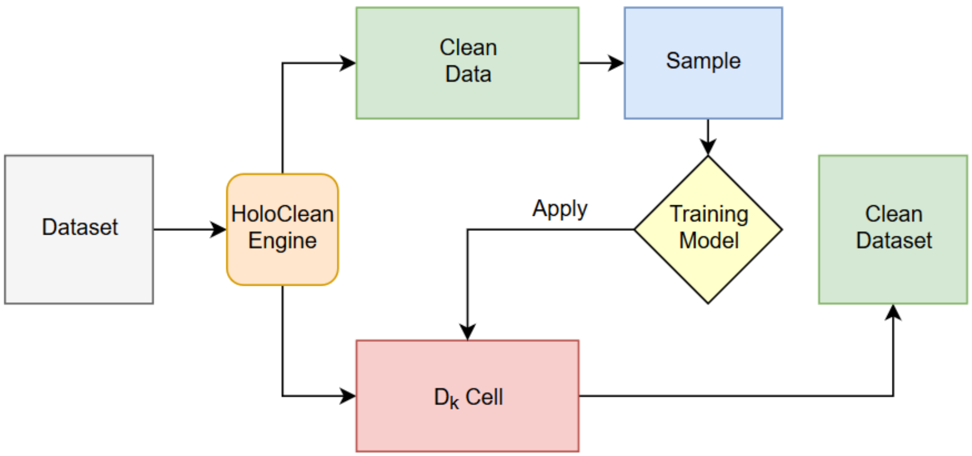
SETUP: Repair

- First HoloClean generates relations used to form the body of DDlog rules, and then uses those relations to generate inference DDlog rules that define HoloClean's probabilistic model. The output DDlog rules define a probabilistic program, which is then evaluated using the Deep-Dive framework.



SAMPLING BASED ON DIMENSIONAL MODEL

- Leverage the dimensional model of the dataset to sample meaningful representative cells.
- Leveraging the dimensional model's FDs, allows us to reduce the number of cells to be considered for dimensional columns at the most granular level.
- This allows us to implement clustered density sampling while leveraging the user's **domain knowledge** about dimensions and measures.



SAMPLING

Original Dataset

	A	B	C
t1	a1	b1	c1
t2	a1	b1	c2
t3	a2	b2	c3

Input Cell	Features depicting overall distribution				Label
t1.A = a1					1
t1.A = a2					0
t1.B = b1					1
t1.B = b2					1
t1.C = c1					0
t1.C = c2					0
t1.C =c3					1
t2.A = a1					0
t2.A =a2					1
t2.B = b1					1
t2.B = b2					0
t2.C = c2					1

Feature Matrix

EVALUATION

- The cleaned data test is matched to clean data that is prepared by a bunch of experts. The data is evaluated on:
  - Precision
  - Recall,
  - Accuracy (sometimes),
  - F1 score (sometimes).

EVALUATION

	Aggregator	Data Tamer
Total records	146690	
Pairs reported as duplicates	7668	180445
Common reported pairs	5437	
Total number of true duplicates (estimated)	182453	
Reported true duplicates (estimated)	7444	180445
Precision	97%	100%
Recall	4%	98.9%

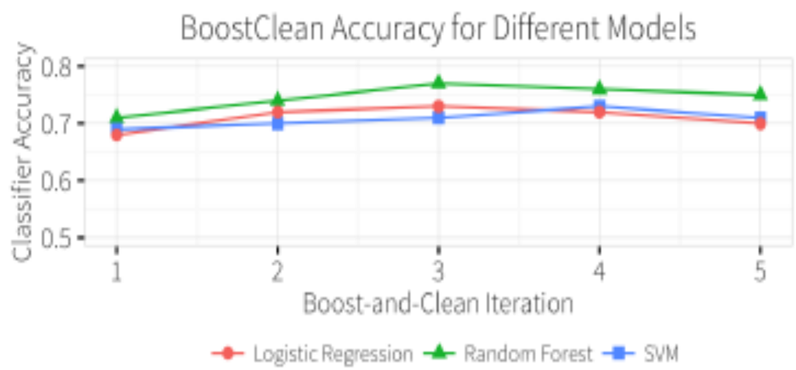
Figure 2: Quality results of entity consolidation for the web aggregator data

EVALUATION

- HoloClean Evaluation:
  - Evaluate on different datasets like hospital, flights, food, physicians.
  - On average the precision is 0.895,
  - On average the Recall is 0.765,
  - On average the F1 Score is 0.819.

EVALUATION

BoostClean achieves up to 81% accuracy and is competitive with hand-written rules, and the word embedding features significantly improve the detector accuracy.



CONCERN

Overfitting

This can lead to framework getting stuck at set of repairs which are incorrect and may require human intervention.

CONCERN

Cost

The cost related to human-interaction is not constant and may change depending on different datasets.

REFERENCES

- Krishnan, S., Franklin, M.J., Goldberg, K., Wang, J. and Wu, E., 2016, June. Activeclean: An interactive data cleaning framework for modern machine learning. In *Proceedings of the 2016 International Conference on Management of Data* (pp. 2117-2120). ACM.
- Yakout, M., Berti-Équille, L. and Elmagarmid, A.K., 2013, June. Don't be SCAREd: use SCalable Automatic REpairing with maximal likelihood and bounded changes. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (pp. 553-564). ACM.
- Stonebraker, M., Bruckner, D., Ilyas, I.F., Beskales, G., Cherniack, M., Zdonik, S.B., Pagan, A. and Xu, S., 2013, January. Data Curation at Scale: The Data Tamer System. In *CIDR*.
- Rekatsinas, T., Chu, X., Ilyas, I.F. and Ré, C., 2017. Holoclean: Holistic data repairs with probabilistic inference. *Proceedings of the VLDB Endowment*, *10*(11), pp.1190-1201.
- Krishnan, S., Franklin, M.J., Goldberg, K. and Wu, E., 2017. Boostclean: Automated error detection and repair for machine learning. *arXiv preprint arXiv:1711.01299*.

References

- Hao, Y.A.N. and Xing-chun, D., 2008. Optimal Cleaning Rule Selection Model Design Based on Machine Learning. In *2008 International Symposium on Knowledge Acquisition and Modeling*.
- Krishnan, S., Wang, J., Wu, E., Franklin, M.J. and Goldberg, K., 2016. ActiveClean: interactive data cleaning for statistical modeling. *Proceedings of the VLDB Endowment*, *9*(12), pp.948-959.
- C. Mathieu, O. Sankur, and W. Schudy. Online correlation clustering. In STACS, pages 573{584, 2010.

THANK YOU

QUESTIONS??