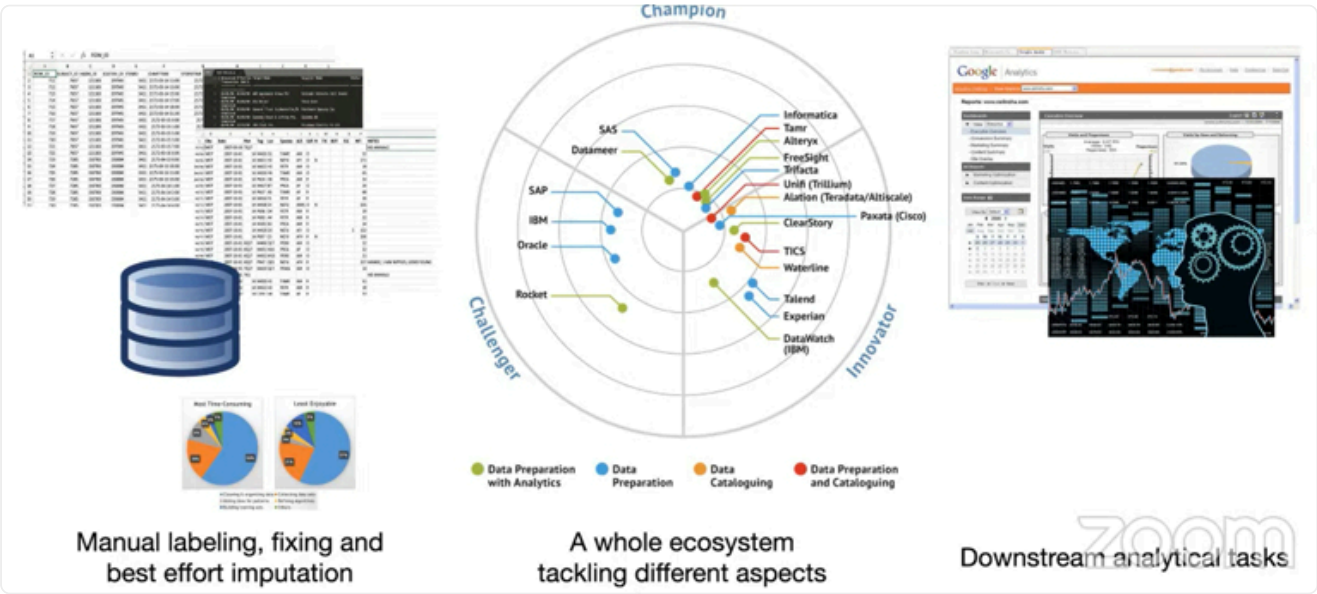
GROUP > SEMINAR & TALK > STANFORD MLSYS SEMINAR

Episode 18

Structure is all you need

- Software 2.0 for data quality management
- Theodoros Rekatsinas | UW-Madison

The Notorious data quality problem



Data quality management task

Error detection tasks

- Tuple (sample) validation

- Cell-value validation

Data repairs

- Missing data imputation
- Data repairs (value replacement)

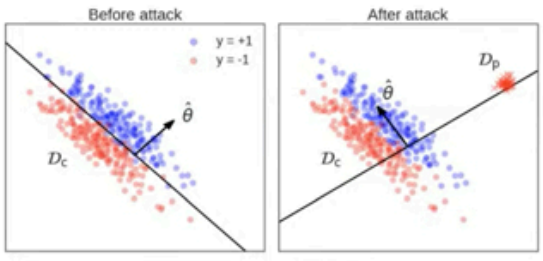
Address	City	State	Zip	Property Tax Rate
640 W Washington Ave	Madison	WI	53703	1.76
652 W Washington Ave	Madison	WI	53703	1.76
2707 ~~W~~ E Washington Ave	Madison	WI	53704	1.05
641 W Washington Ave	adison, WI	53703	53703	1.65

Invalid tuple Erroneous value Data repair Value imputation

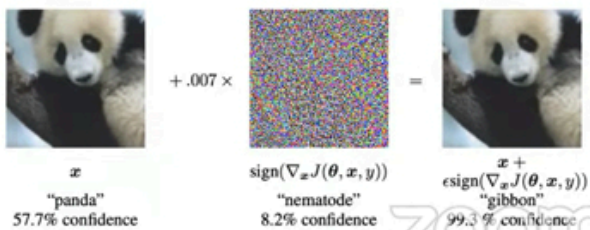
- Push to ML model?

ML models are sensitive to low-quality data

- In the training stage, corrupted data can bias the ML models
- Data poisoning techniques destroy a model by adding a small fraction of adversarially crafted points



- In the inference stage, errors in the inference queries can result in wrong predictions
- Adversarial attacks add noise to flip the prediction



[Koh et al. 2018, Goodfellow et al. 2015]

Goal: Streamline data quality management

- **Many projects:** mainly HoloClean (Inductiv) and recently Picket and Marius
- **Question:** Can we automate data quality management tasks?
- **Focus:** data validation tasks (i.e., data cleaning, out-of-distribution detection)
- **Automation opportunity:** adopt probabilistic semantics and rely on self-supervised ML approaches to learn how clean data looks like and use that to drive automation

Reasoning about **structured context** is key for automated data validation

- Marius: graphs, heteregeneous-structure data

Example: data validation for mean estimation

The simplest analytical model: compute the mean **X** of the following sample

X1	X2	X3	X4	X5
2.39	1.34	1.05	0.01	0.35
NaN	-1.3	1.92	NaN	1.14
0.19	NaN	1.76	1.90	NaN
0.99	-2.16	3.15	1.72	-3.15

Coordinate-wise mean: **X** = (1.19, -0.706, 1.97, 1.21, -0.553)

Mean-after filtering: **X** = (1.69, -0.41, 2.1, 0.865, -1.4)

- Discrepancies between two estimate

X1	X2	X3	X4	X5
2.39	1.34	1.05	0.01	0.35
0.62	-1.3	1.92	NaN	1.14
0.19	-1.57	1.76	1.90	NaN
0.99	-2.16	3.15	1.72	-3.15

Improved estimate **X** = (1.048, -0.923, 1.97, 1.21, -0.553); why?

- Filling some of the values, and leaving the others

X1	X2	X3	X4	X5
2.39	1.34	1.05	0.01	0.35
0.62	-1.3	1.92	NaN	1.14
0.19	-1.57	1.76	1.90	NaN
0.99	-2.16	3.15	1.72	-3.15

Improved estimate **X** = (1.048, -0.923, 1.97, 1.21, -0.553); why?

Exploit the fact that X1 = X2 + X3, and X4, X5 are independent

- If know this dependency in advanced, then we could make better estimate

Structure-aware data cleaning is necessary

- Two-step meta-algorithm for robust mean estimation:
  - (1) Recover: use the dependencies across coordinates of the data (i.e., the structure) to recover the values of corrupted samples
  - (2) Estimate: After fixing corruptions, perform standard statistical estimation
- Information theoretically optimal error-estimation in the presence of structure:

Structure	Entire-sample corruption (epsilon-samples corrupted)	Coordinate-level Corruption (alpha entries corrupted)
No Structure	$\Theta(\epsilon)$	$\Theta(\alpha n)$
Linear Structure (x = Az)	$\Theta(\epsilon)$	$\Theta(\alpha n / m_A)$

n: number of coordinates

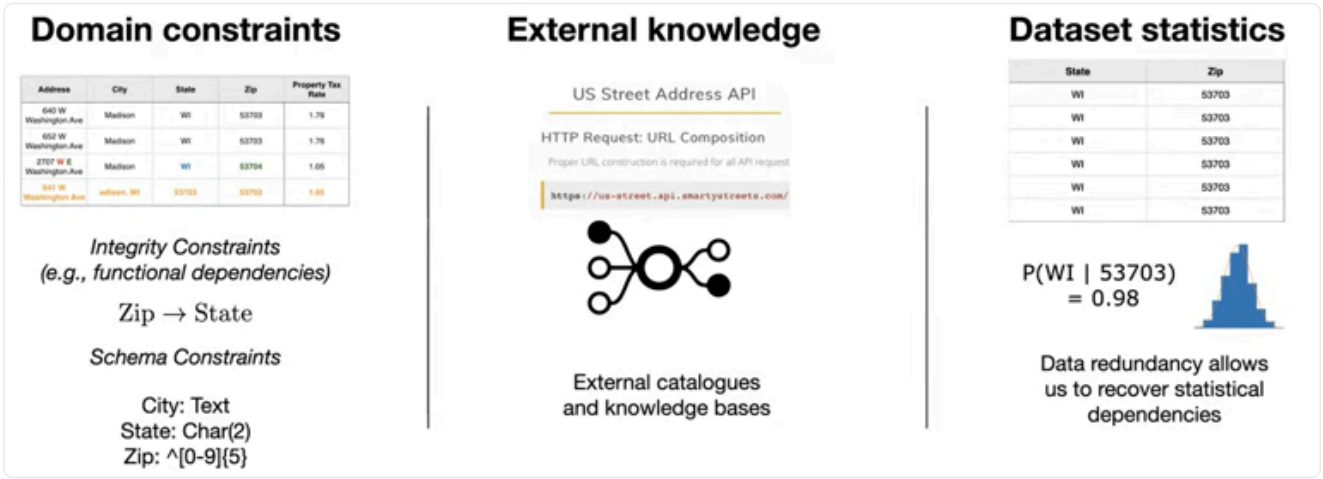
ϵ: % of corrupted samples

α: % of corrupted entries

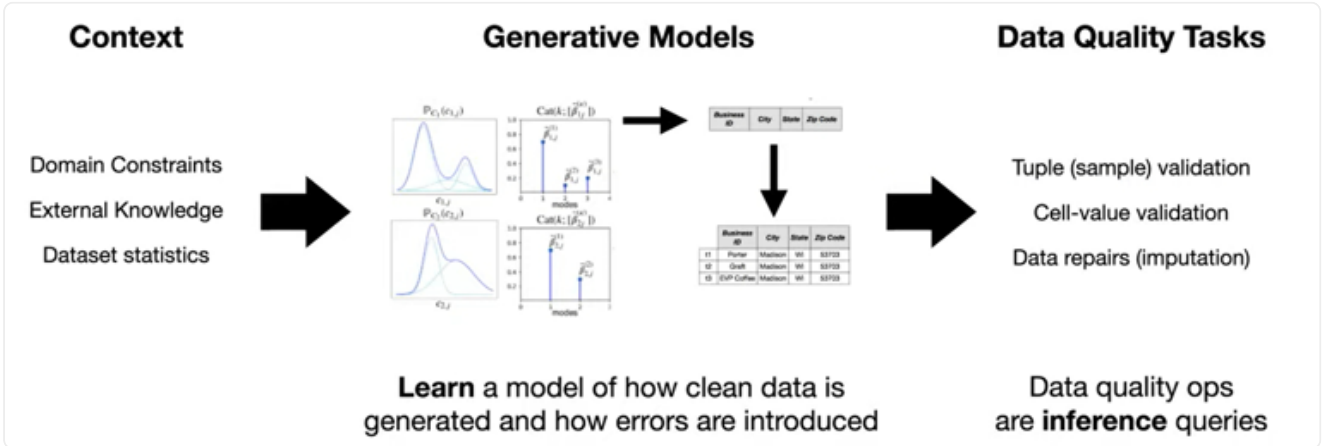
m_A: number of corruptions to reduce the row space by one

Ref: Structure-aware robust mean estimation [Liu et al., 2020]
https://arxiv.org/abs/2002.04137

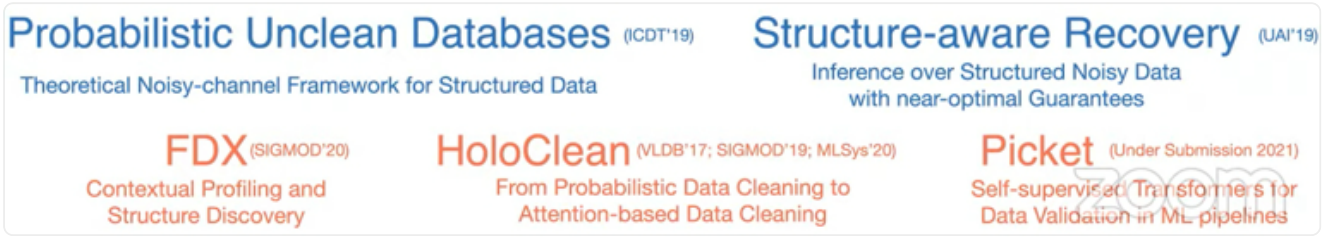
Heterogeneous types of structure



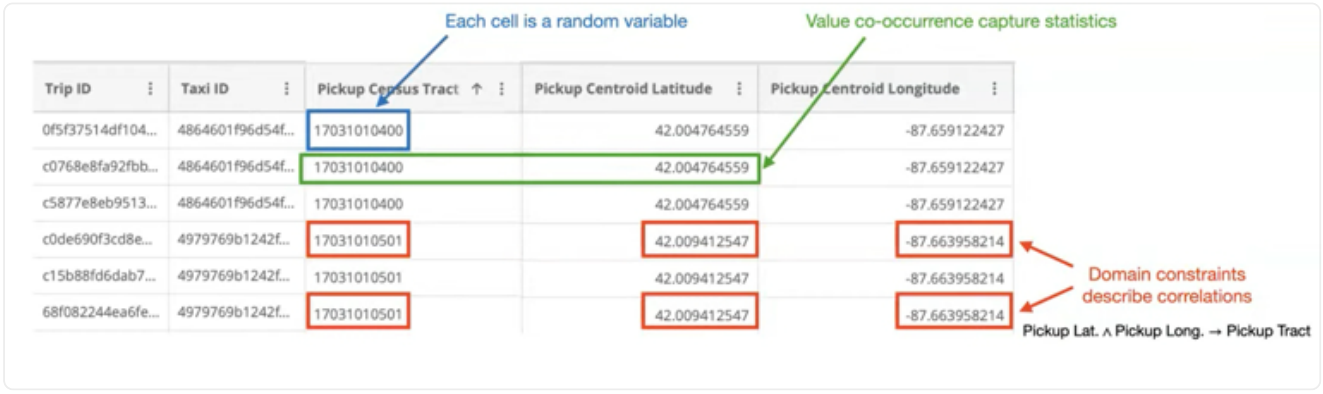
Contextual ML for automated data quality

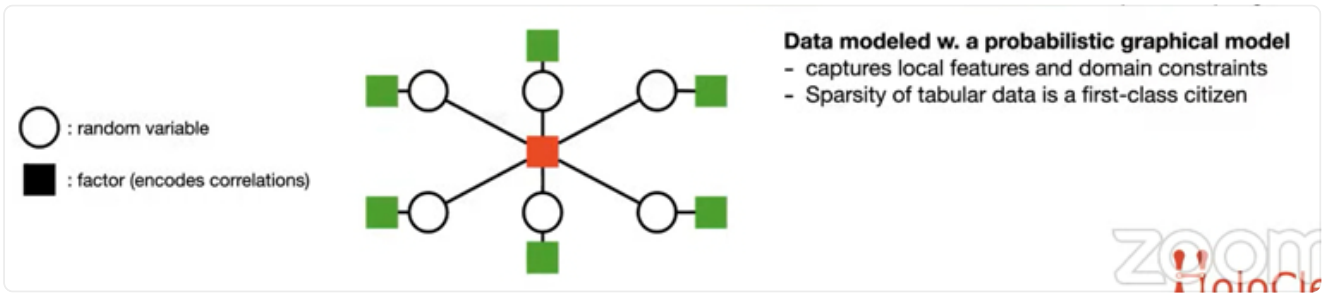


- Unsupervised manner?
  - Can I run this inference queries fast

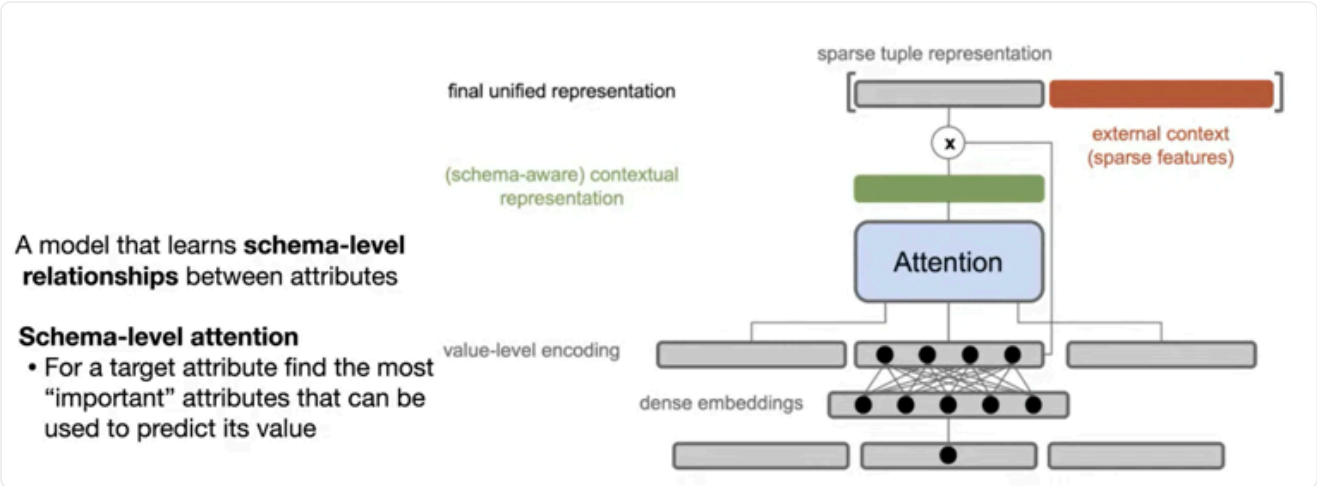


HoloClean: Probabilistic Data Repairs

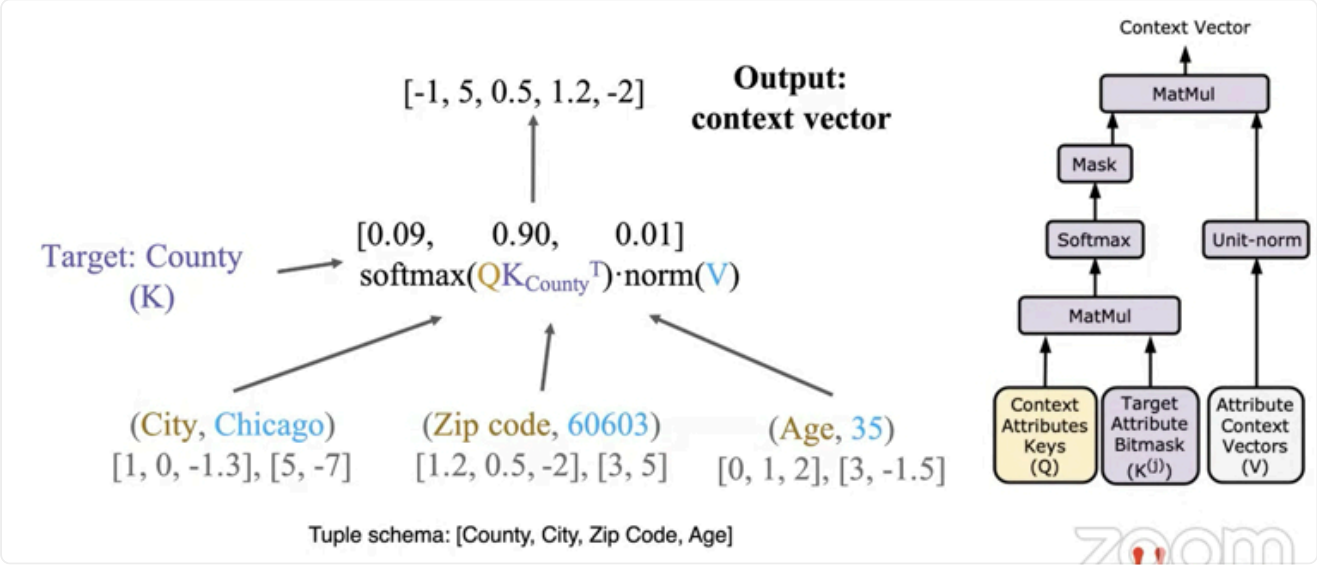




("Take two")

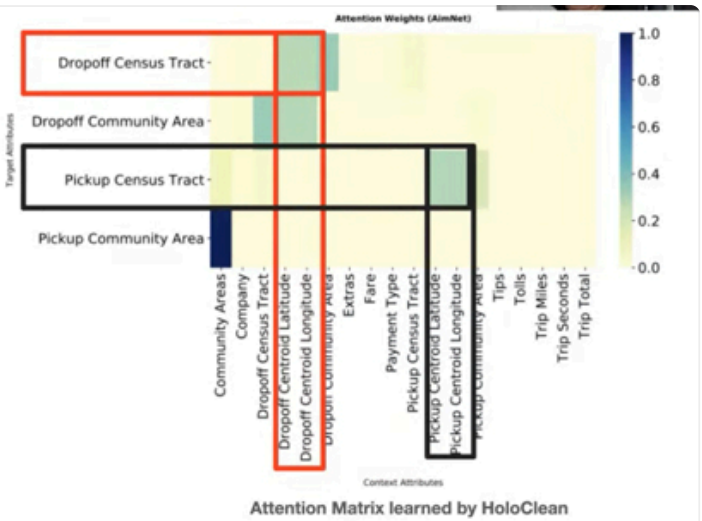


Schema-level Attention



Why Attention?

Attention can capture dependencies like the ones in graphical model; analogue to kernels that can used to learn graphical models



Attention Matrix learned by HoloClean

Refs:
- Learning Graphical Models with Mercer Kernels, Bach and Jordan, 2002;
- Transformer Dissection: A Unified Understanding of Transformer's Attention via the Lens of Kernel, Tsai et al., 2019

Naturally-occurring missing data

Chicago taxi data
- Benchmark in Google's TFX data validation
- Pickup/drop-off info, fare, company
- Naturally-occurring missing values w/ ground truth
- Systematic bias between companies

	Company	Pickup Census Tract	Pickup Centroid Latitude	Pickup Centroid Longitude
2366	Chicago Medallion Leasing INC	nan	41.975171	-87.687516
78445	Dispatch Taxi Affiliation	nan	41.975171	-87.687516
57109	Taxi Affiliation Services	17031040401	41.972036	-87.686100

All within "17031040401" census tract

- 40+ F1 points improvements over next-best method
- 10x faster

AimNet (MLSys'20)	HCQ	XGB	MIDAS	GAIN	MF	MICE
HoloClean with Attention	HoloClean with quantization	XGBoost	Denoising Autoencoder	GAN	Random Forest	Linear regression

Accuracy on discrete attributes for the Chicago data set

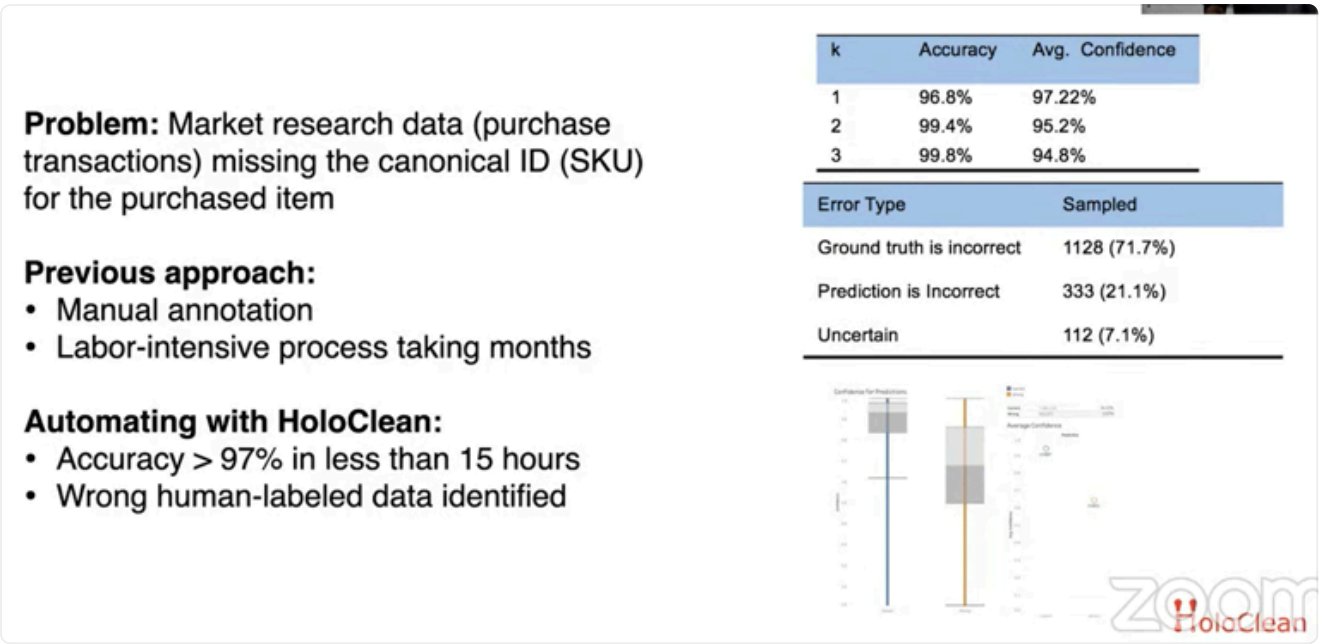
AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
0.73 ± 0.01	0.07 ± 0.0	0.27 ± 0.0	0.09 ± 0.01	0.01 ± 0.01	0.3 ± 0.0	—

Run time (minutes) for the Chicago data set

AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
53	124	5350	176	186	7439	—

24

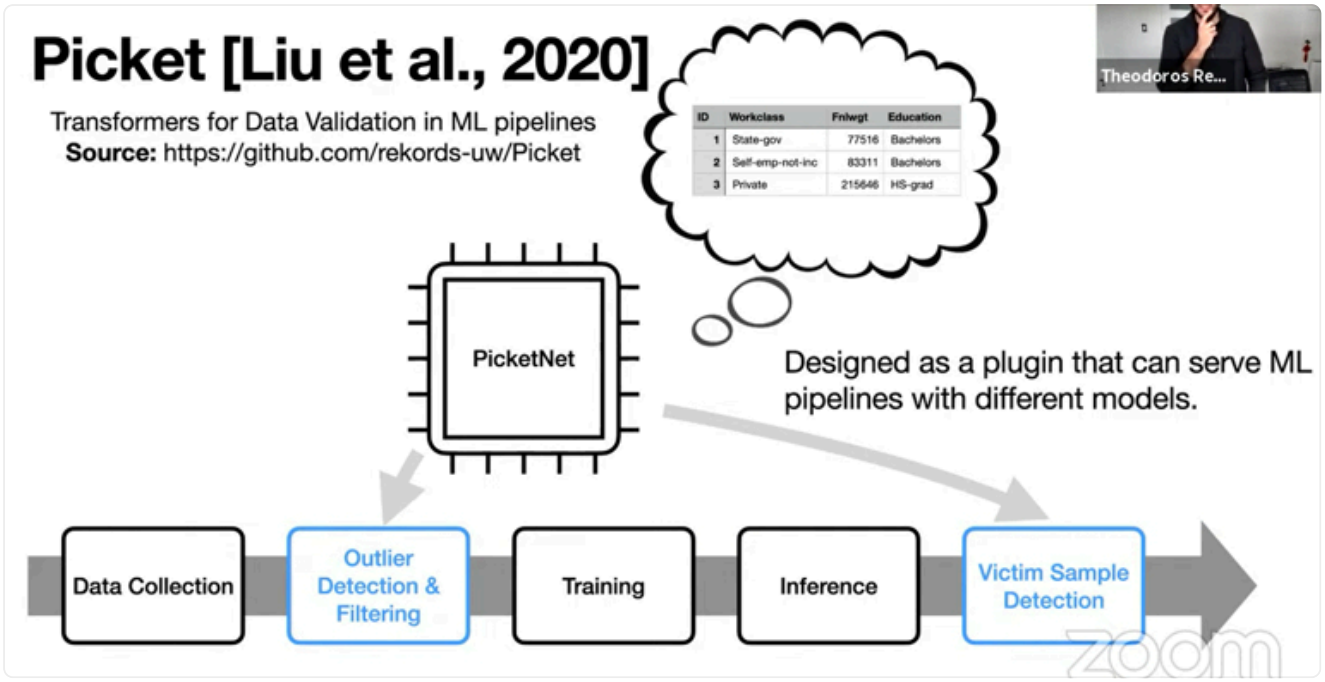
Use case: Data Categorization

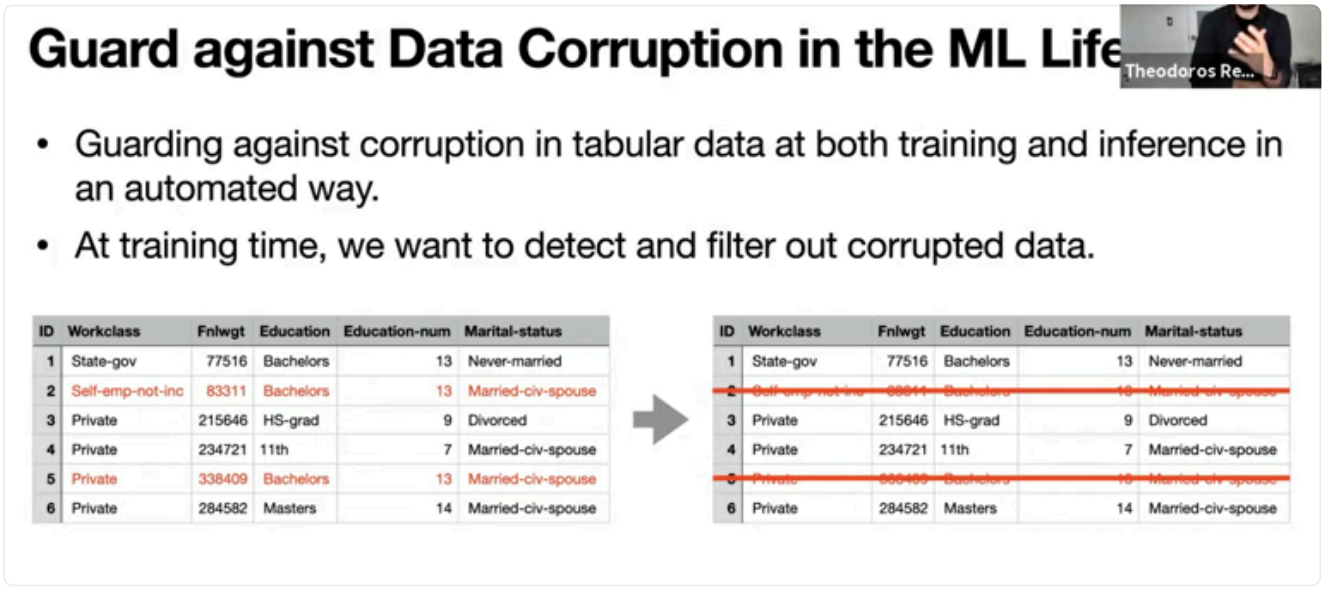


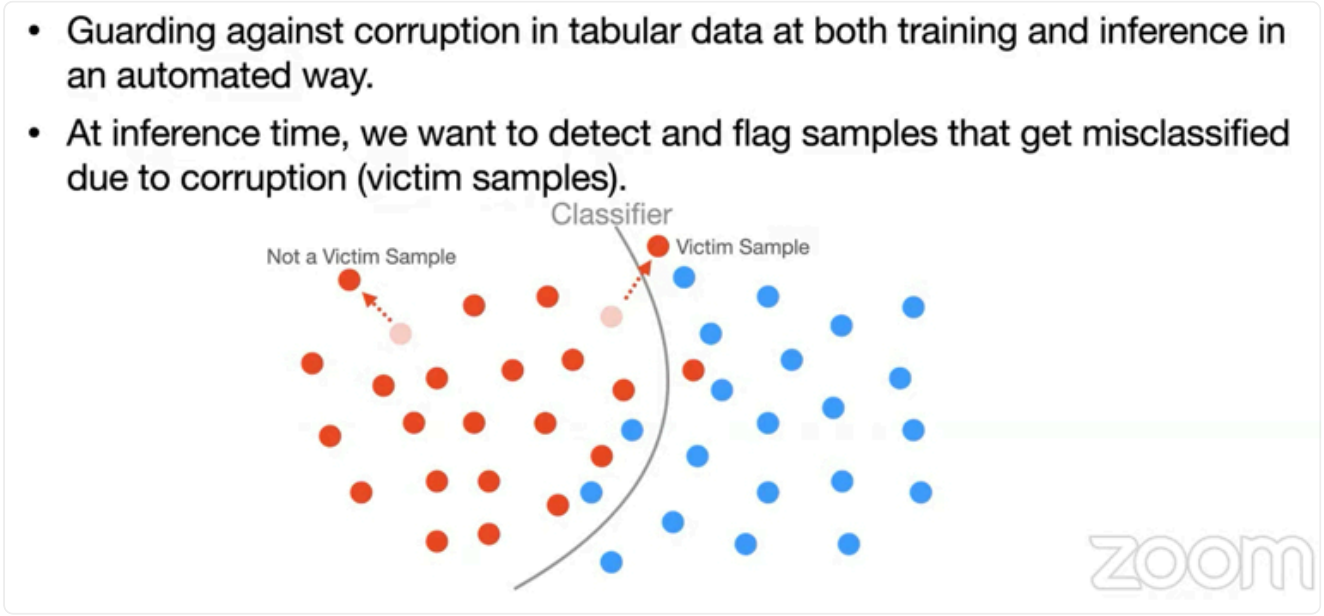
Other use cases

- Error detection in demographic data used for policy decision
- KIP tracking
- Imputation of numerical data for industrial machinery monitoring

Picket: self-supervised transformers for data validation in ML pipelines

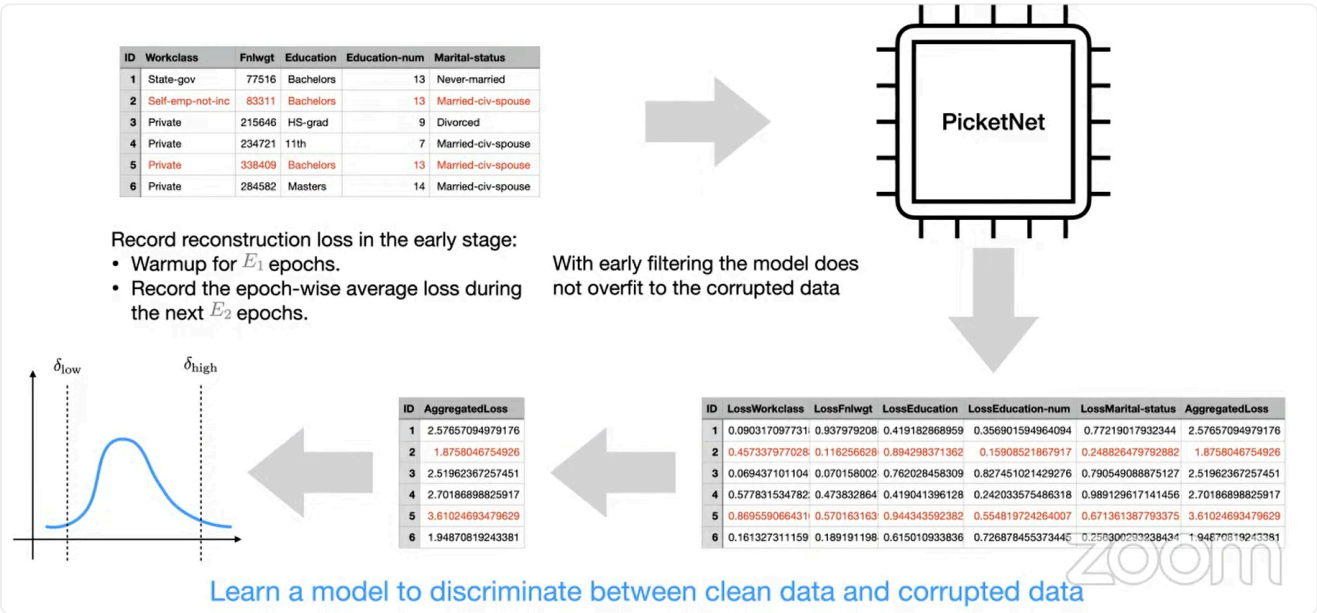

Picket [Liu et al., 2020]
Transformers for Data Validation in ML pipelines
Source: https://github.com/rekords-uw/Picket
Theodoros Re...
ID Workclass Fnlwgt Education
1 State-gov 77516 Bachelors
2 Self-emp-not-inc 83311 Bachelors
3 Private 215646 HS-grad
PicketNet
Designed as a plugin that can serve ML pipelines with different models.
Data Collection
Outlier Detection & Filtering
Training
Inference
Victim Sample Detection
zoom


Guard against Data Corruption in the ML Life
Theodoros Re...
• Guarding against corruption in tabular data at both training and inference in an automated way.
• At training time, we want to detect and filter out corrupted data.
ID Workclass Fnlwgt Education Education-num Marital-status
1 State-gov 77516 Bachelors 13 Never-married
2 Self-emp-not-inc 83311 Bachelors 13 Married-civ-spouse
3 Private 215646 HS-grad 9 Divorced
4 Private 234721 11th 7 Married-civ-spouse
5 Private 338409 Bachelors 13 Married-civ-spouse
6 Private 284582 Masters 14 Married-civ-spouse

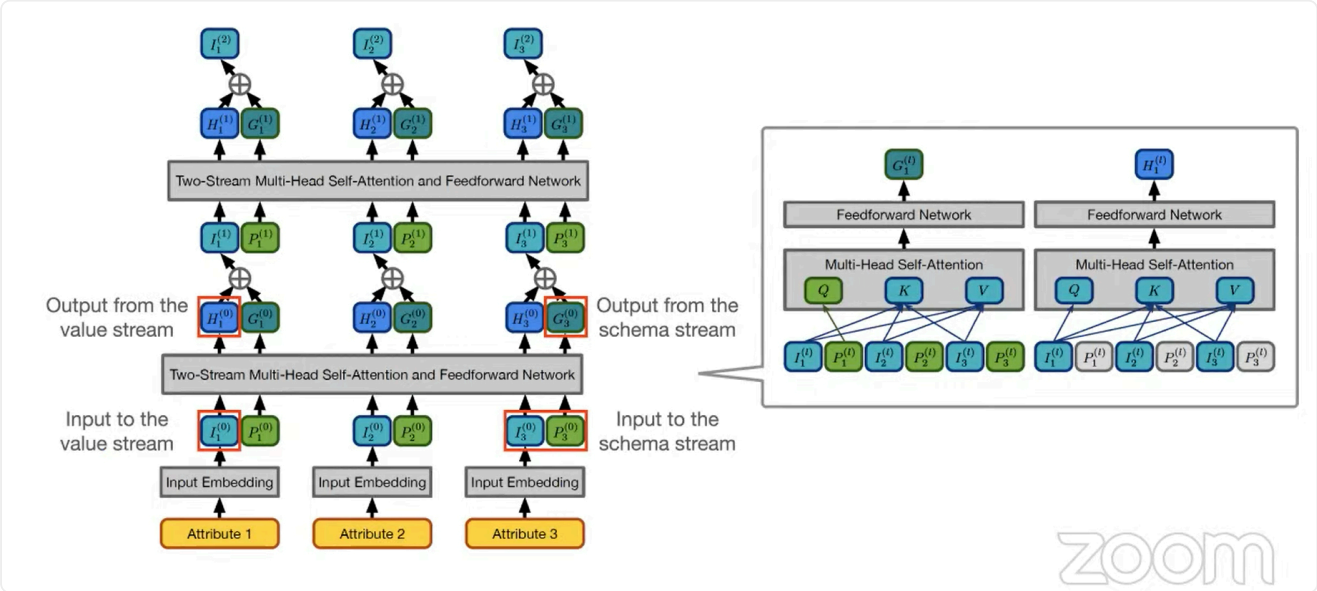

• Guarding against corruption in tabular data at both training and inference in an automated way.
• At inference time, we want to detect and flag samples that get misclassified due to corruption (victim samples).
Classifier
Not a Victim Sample
Victim Sample
zoom

Loss-based Outlier Detection and Filtering

1. Go back to the idea of learning a model to capture the clean data, and use this model in decisions
    a. PicketNet: transformer
    b. Outlier detection problem



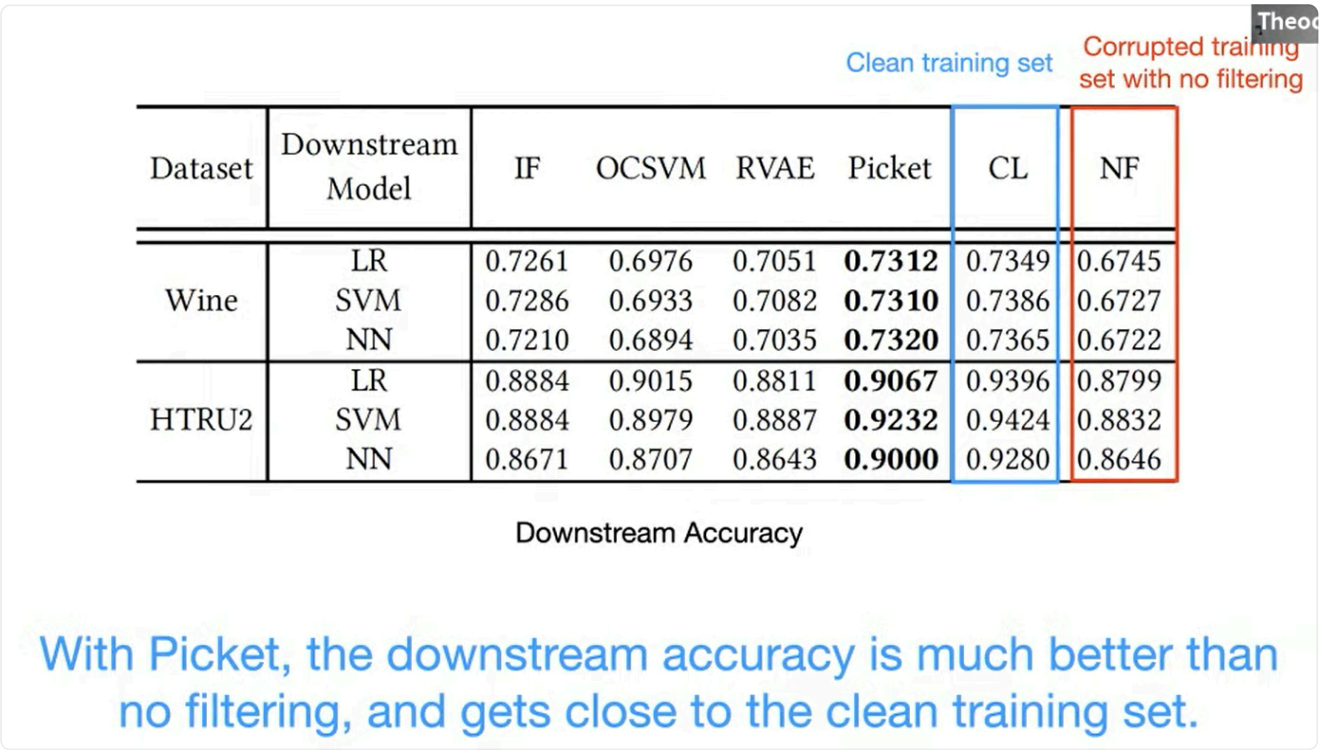
PicketNet: two-stream transformer for tabula data



- Benefits:
  - Value stream: flexibility
  - Schema stream: regularization

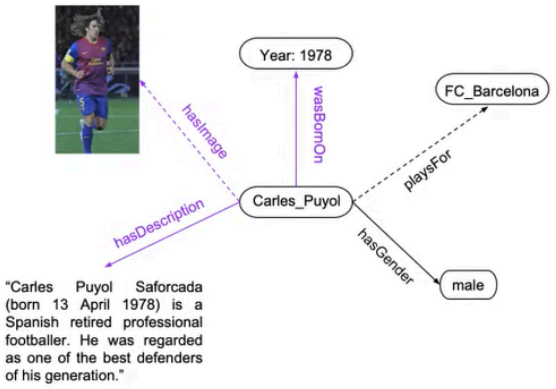
Experimental Highlights: Poisoning Attacks

- Aim to destroy



Dataset	Downstream Model	IF	OCSVM	RVAE	Picket	CL	NF
Wine	LR	0.7261	0.6976	0.7051	**0.7312**	0.7349	0.6745
	SVM	0.7286	0.6933	0.7082	**0.7310**	0.7386	0.6727
	NN	0.7210	0.6894	0.7035	**0.7320**	0.7365	0.6722
HTRU2	LR	0.8884	0.9015	0.8811	**0.9067**	0.9396	0.8799
	SVM	0.8884	0.8979	0.8887	**0.9232**	0.9424	0.8832
	NN	0.8671	0.8707	0.8643	**0.9000**	0.9280	0.8646

Contexual ML for automated data quality ops



Source: Embedding Multimodal Relational Data for Knowledge Base Completion, Pezeshkpour et al., ACL 2018

Extend the ideas behind HoloClean and Picket to heterogeneous, open-domain data (e.g., Knowledge Graphs)

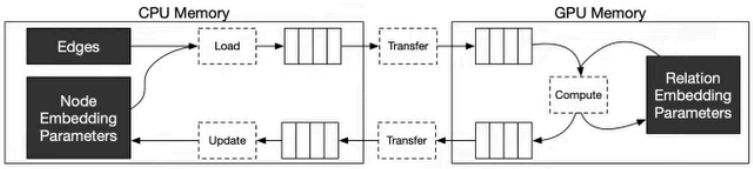
Graph embeddings and contextual models over 100s of millions of entities (nodes) and 10s of billions of facts (edges)

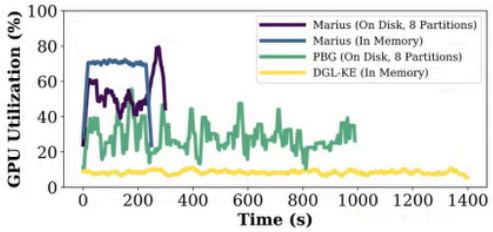
Challenge: scalable graph learning with optimized resource utilization

Scalable no-code graph learning

marius

Learning Massive Graph Embeddings on a Single Machine, OSDI'2021
Find more at **marius-project.org**





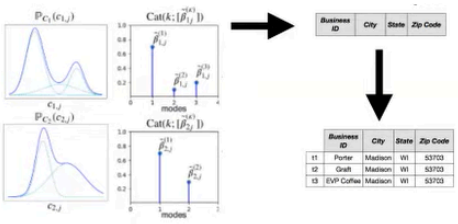
Pipelining and a novel data replacement policy allow Marius to maximize resource utilization of the entire memory hierarchy (including disk, CPU, and GPU memory)

Achieves graph learning over billion edge graphs **in a single machine**

Context

Domain Constraints
External Knowledge
Dataset statistics

Generative Models



Data Quality Tasks

Tuple (sample) validation
Cell-value validation
Data repairs (imputation)

Learn a model of how clean data is generated and how errors are introduced

Data quality ops are **inference** queries

Thank you!
@thodrek

- Systematic variation of the data?
  - Types of noise: real-world data, no assumption about the noise. Types of noise can be random, or systematic error (integrating things across different sources)
    - Random
    - Systematic: repeated instances of the noise, if we condition, then it's not random
    - Adversarial noise: are aware of the downstream task, and go and attack that system
- Holoclean
  - Attention: handle this gracefully, pick up the type of strong bias
- Picket:
  - Worst-possible case (Adversarial)
  - Not overfitting to examples
- Distinguish between out-of-distribution or systematic change
  - Solution: two streams (scheme, value)
    - Value: robust towards the case. Kernel structure that operates on this level.
  - Profiling mechanism
- Hetereogeneous data types
  - Tabula data: higher level constraints
    - Encode them as functional dependencies in DB
    - or pick them up through attention mechanism
    - Back in the day: user specification
  - Structure learning over the data
    - Exactly the attention matrix, in a faster and cheaper way
  - Hetereogenous

        - Same mechanism can potentially hold for a graph
        - Running structural learning type of profiling
            - Identify homogenous area as a pre-processing step
            - And preprocessing ...
            - Filtering away and keep
        - Doing this heavily in Holoclean
- Monitor data and see if something is going on in the data pipeline
    - Reliable data
    - How that setting is different? Or some of the goals change?
        - Reconstruction
            - Signal and context
            - With high confidence, then it should be outlier
        - Information about the likelihood

Goal:

- applying the rules at scale (ETL, ...)
- Start-up and companies: specific problem
    - Identify duplicates in records
    - Infer rules to prepare and standardize
    - AI --> platform for error detection and fixing
- Nobody is targeting
    - Automating this
    - Position: reasoning about noisy structured data

Challenges and what that issue looks like:

- What is the model doing? Aspects

    - Attention: interpretable (know the semantics of the attributes, put more weights or less)
    - Allow people not immediately accept them. Have confidence over the prediction of the model.
        - Accept the one that makes sense
- Also, allow users to introduce business /external features that would allow you to introduce

Holoclean

- Accept logic rules, convert them to features
- Support matching functions

Which one should I trust?

- Ensemble (weighted vote)
- In real cases, people believe their rules...

Previous
Episode 17

Next
Index

Last updated 3 years ago