Warper

Efficiently Adapting Learned Cardinality Estimators to Data and Workload Drifts

Beibin Li[1,2], Yao Lu[2], Srikanth Kandula[2]

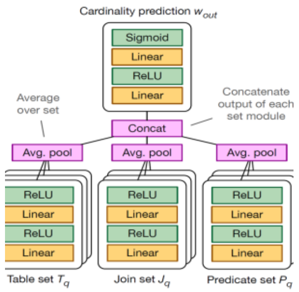
[1] University of Washington, Seattle, WA

[2] Microsoft Research, Redmond, WA

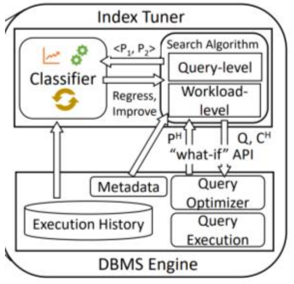
2022

Machine Learning (ML) for Database Systems

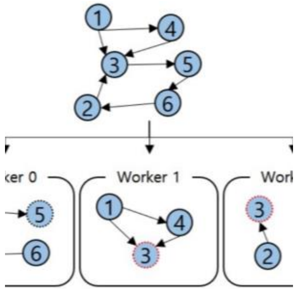
Research



Cardinality Estimation

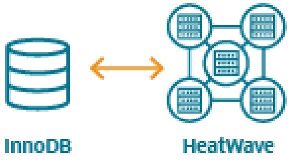


Index Tuning



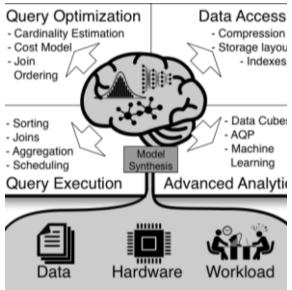
Graph Paritioning

Practice



Oracle's HeatWave

Google's AlloyDB



MIT's SageDB

Challenge in AI Applications:

Training and Testing Distribution Shifts (Drifts)

Cardinality Estimation and Its Machine Learning Process

SELECT * FROM t1 JOIN t2 ON t1.id == t2.id
WHERE $t1.col_1 \geq L_1$ AND $t1.col_1 \leq U_1$ AND
$t2.col_1 \geq L'_1$ AND ...

Machine Learning Model
(black box)

$$f(x) = y$$

Query

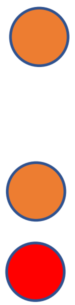
Ground Truth
(Value)

Data Shifts

(e.g., insert, delete, update)

f(weight <=3 and price <= 3)

Weight	Price
1.0	1.0
1.0	1.5
3.0	4.2



New D

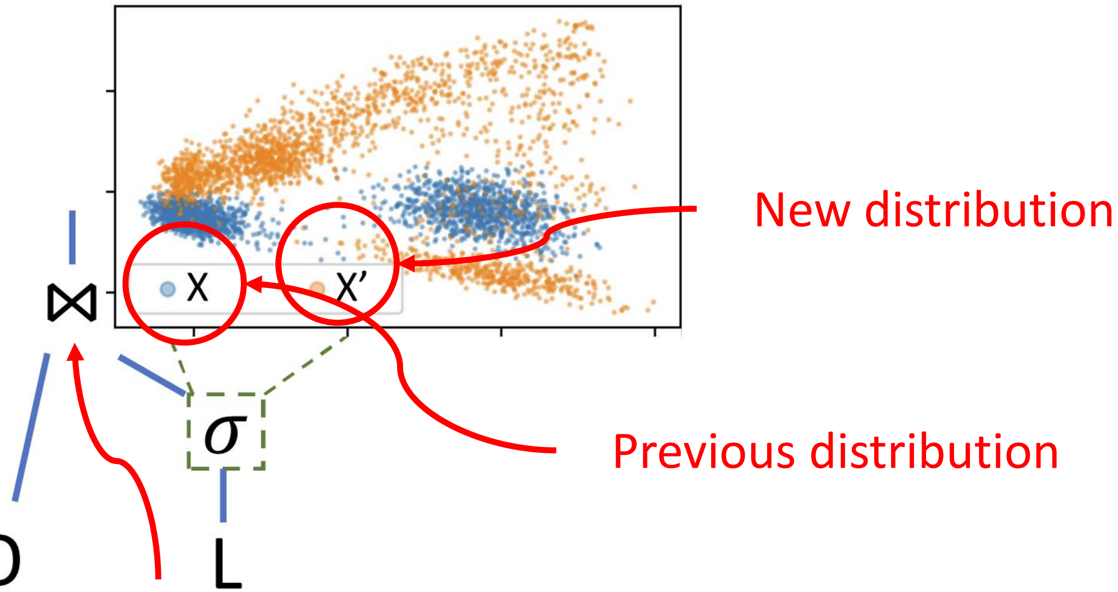
Workload Shifts

P(x): distribution of input data features

Weight Bound	Price Bound
< 1.0	< 1.0
< 2.1	< 2.0
< 4.0	< 3.8

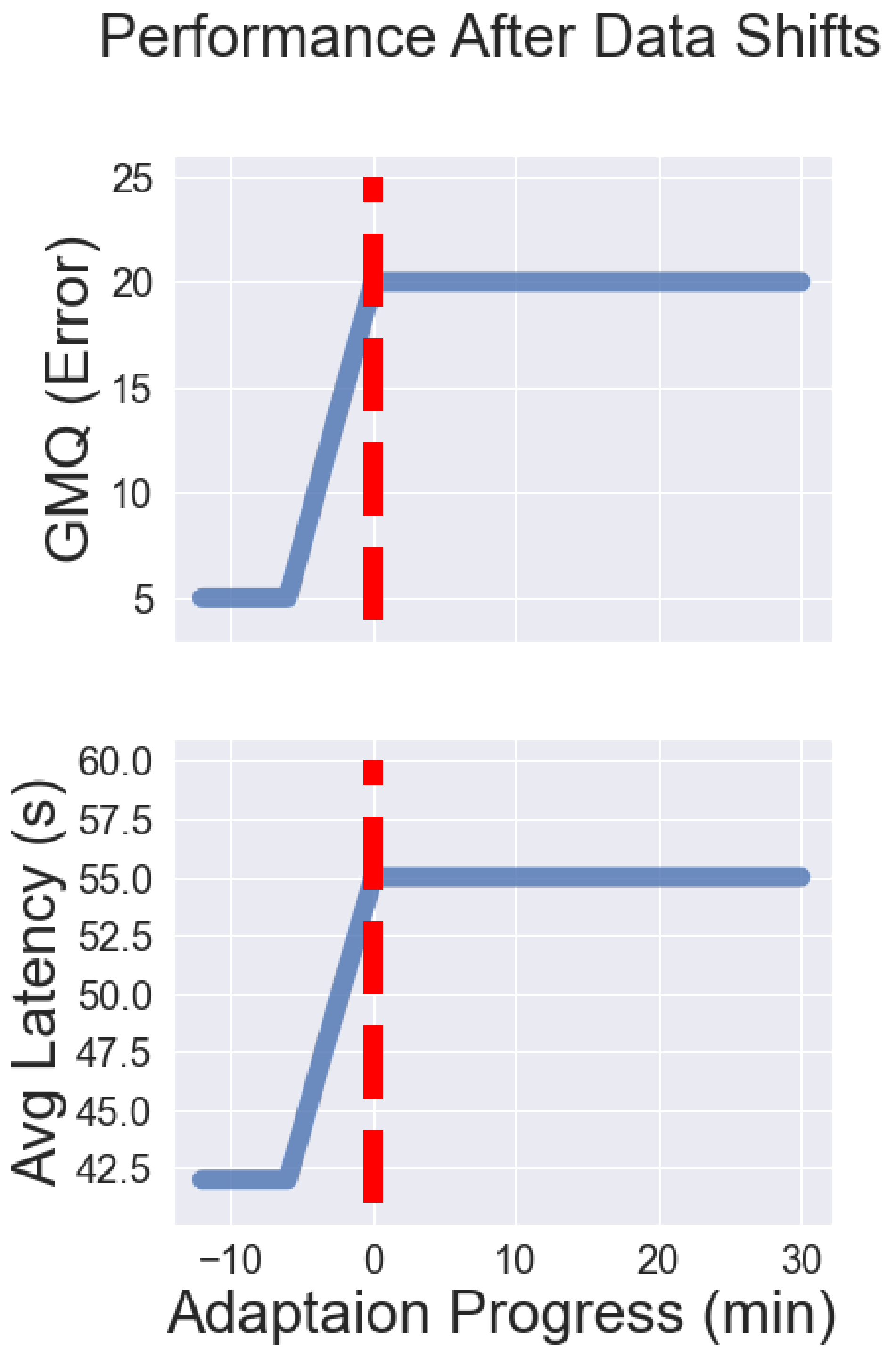
Shifts Lead to Regression

Q-Error: $q = \max(\frac{gt}{pred}, \frac{pred}{gt})$



Join the "**O**rder" and "**L**ineitem" tables

Example in TPC-H dataset



State-of-the-art for Adaption in CE

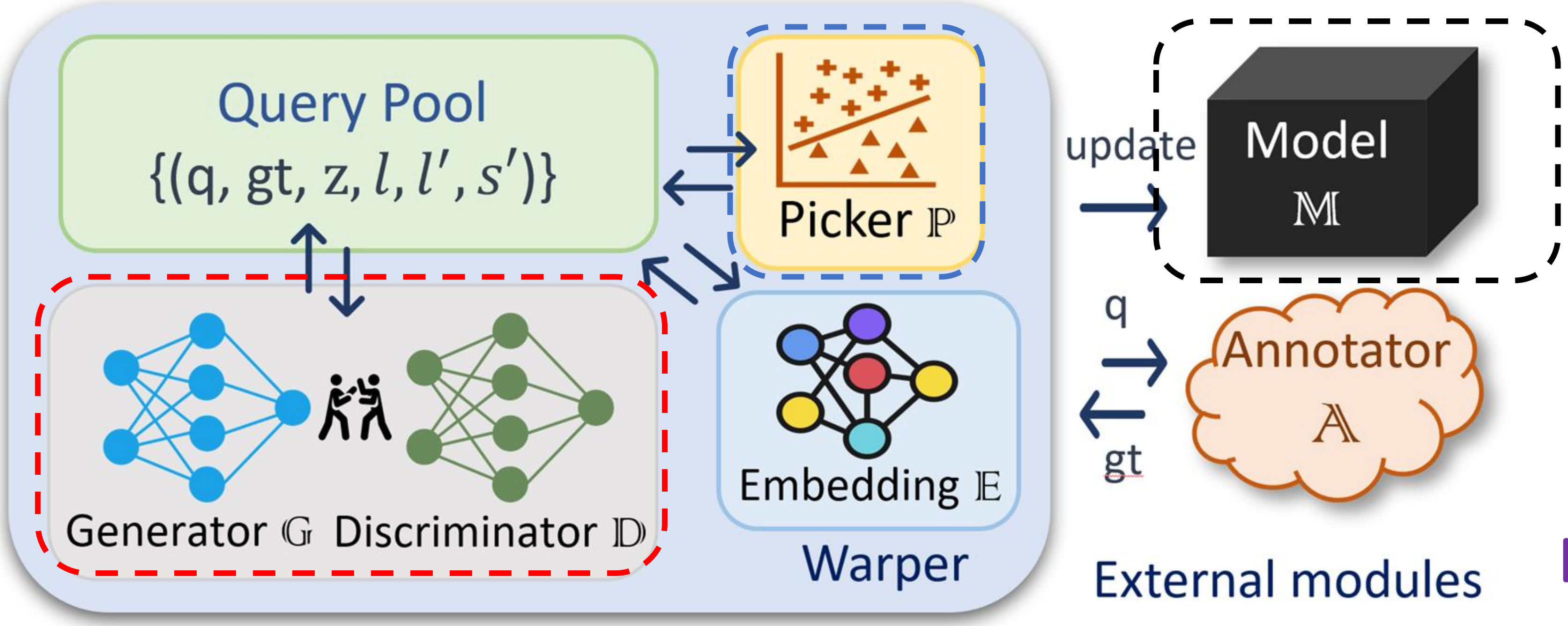
Scheme	How
LM[1] Naru[3]	Re-train
MSCN[2]	Completely re-train or fine-tune model
DeepDB[4]	Partial re-train

- Other DB tasks (e.g., [5]) use apriori training
  - Ad-hoc
  - Domain insights
  - Overly general.

[1] Dutt, Anshuman, et al. "Selectivity estimation for range predicates using lightweight models." *Proceedings of the VLDB Endowment* 12.9 (2019): 1044-1057.
[2] Kipf, Andreas, et al. "Learned cardinalities: Estimating correlated joins with deep learning." *CIDR* (2019).
[3] Yang, Zongheng, et al. "Deep unsupervised cardinality estimation." *Proceedings of the VLDB Endowment, 13.3 (2019)*
[4] Hilprecht, Benjamin, et al. "DeepDB: learn from data, not from queries!." *Proceedings of the VLDB Endowment* 13.7 (2020): 992-1005.
[5] Ma, Lin, et al. "Active Learning for ML Enhanced Database Systems." *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2020.

Goals

- Agnostic to ML Models
- Small Computation Overhead
- Quick Adaptation



Examples

Drift Case 1: Workload Distribution Drift

- **Table stays the same**
- **Workload changed**

Previous Distribution

Many queries

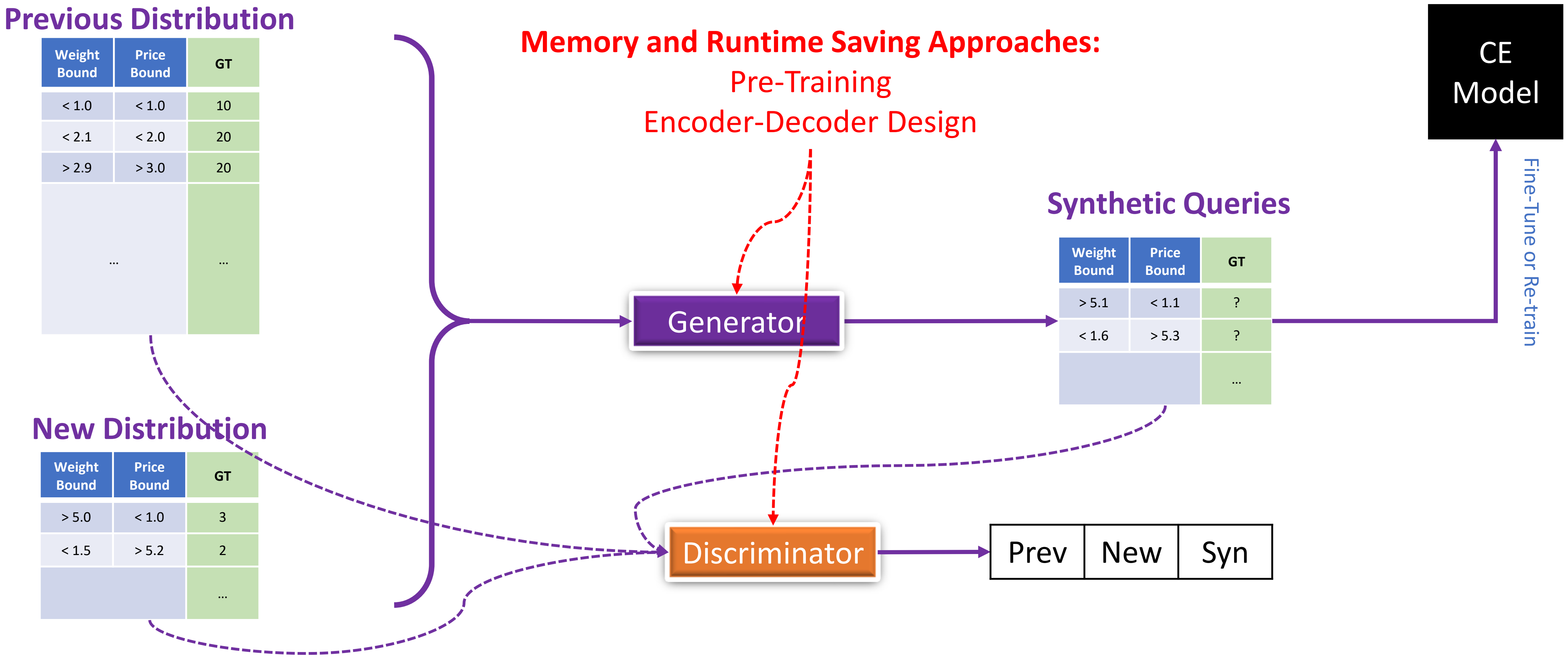
Weight Bound	Price Bound	Card. GT
< 1.0	< 1.0	10
< 2.1	< 2.0	20
> 2.9	> 3.0	20
…		…

New Distribution

A few queries

Weight Bound	Price Bound	Card. GT
> 5.0	< 1.0	3
< 1.5	> 5.2	2
		…

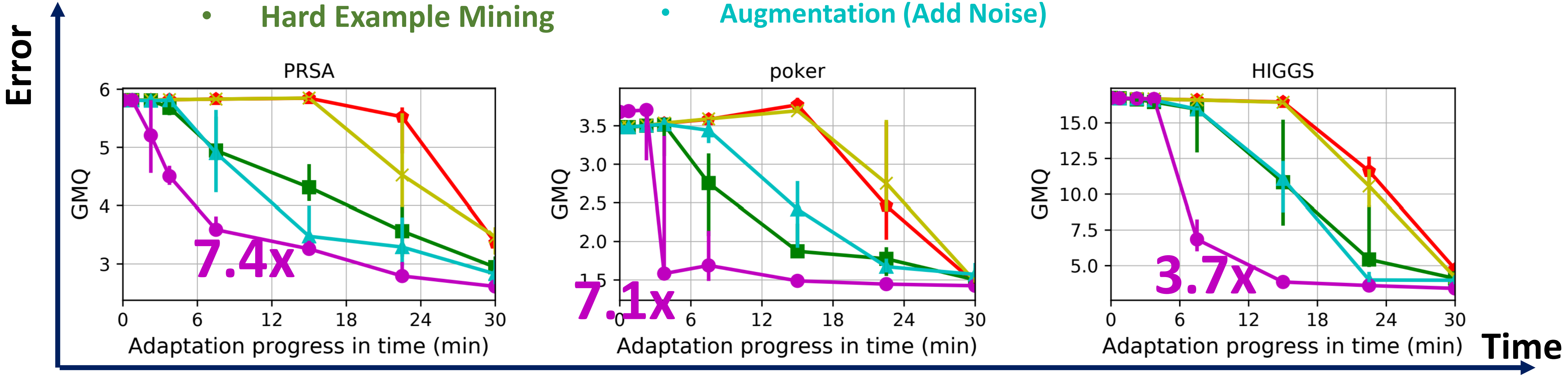
Drift Case 1: ***Warper* with GAN** (Generative Adversarial Network)



Drift Case 1: Experiments

- Black Box Models: LM (VLDB 19'), MSCN (CIDR 19')
- Three datasets
- 12 new queries arrive per minute.

- Fine-Tune
- Hard Example Mining
- Mixture
- Augmentation (Add Noise)
- Warper (Ours)



Drift Case 1: Summary

- Workload Drift
- Scenario: **number of new** query predicates **is small**
- Goal: **adapt** the black-box ML model **quickly**
- Solution: **synthesize additional queries** for the model

- However, what if the number of new queries is large?

Next Drift Case

Drift Case 2: Too Many to Label

- **Table stays the same**
- **Workload changed**

Previous Distribution

Many Queries

Weight Bound	Price Bound	Card. GT
< 1.0	< 1.0	10
< 2.1	< 2.0	20
> 2.9	> 3.0	20
…		…

New Distribution

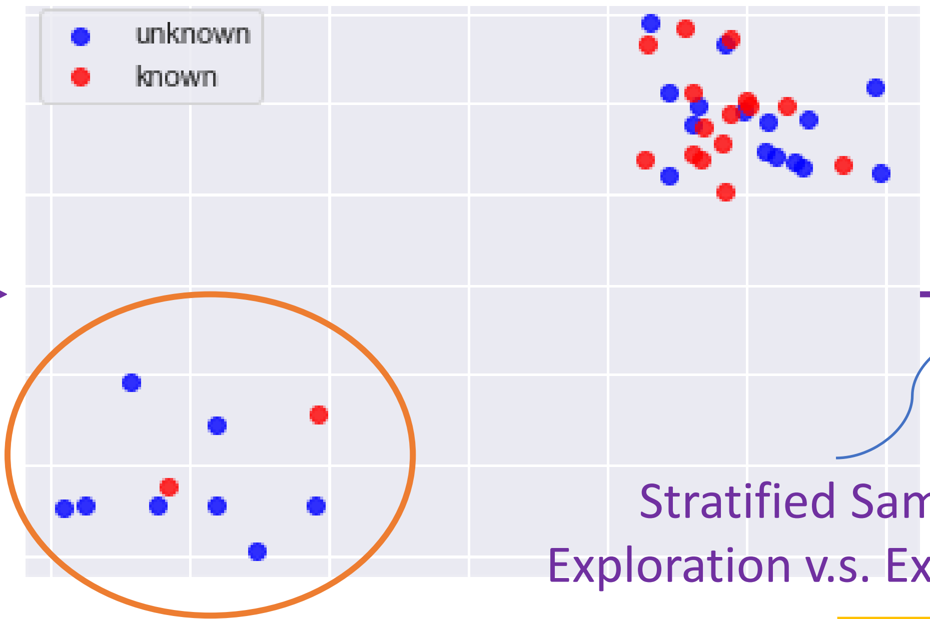
Many Queries

Weight Bound	Price Bound	Card. GT
< 1.2	< 1.1	?
< 2.1	< 2.0	?
> 0.9	> 4.2	?
…		…

Drift Case 2: ***Warper*** with Picker

Weight Bound	Price Bound	Card. GT	Source
< 1.0	< 1.0	10	**Prev**
< 2.1	< 2.0	20	**Prev**
> 2.9	> 3.0	20	**Prev**
…		…	
< 1.2	< 1.1	?	**New**
< 2.1	< 2.0	?	**New**
> 0.9	> 4.2	?	**New**
…		…	

Encoder-Decoder Design
(reuse components in GAN)



Stratified Sampling
Exploration v.s. Exploitation

Weight Bound	Price Bound	Card GT.
> 5.0	< 1.0	3
< 1.5	> 5.2	4
		…

Fine-Tune or Re-train

CE Model

Speedup		
PRSA	Poker	HIGGS
1.1x	1.4x	1.2x

Drift Case 2: Summary

- Workload Drift
- Scenario: **number of new** queries **is large**
- Solution: **select novel queries** with higher priority

A few queries
CPU is in idle

Lots of queries
CPU cannot label all

- Both Drift Case 1 and Drift Case 2 are workload drift

Data Drift Case

Drift Case 3: Data Shifted

Data Table

Weight	Price
1.0	1.0
5.0	4.5
1.0	1.5
...	

Data Shift

Weight	Price
1.0	1.1
~~5.0~~	~~4.5~~
1.0	1.5
5.0	6.9
...	

Queries

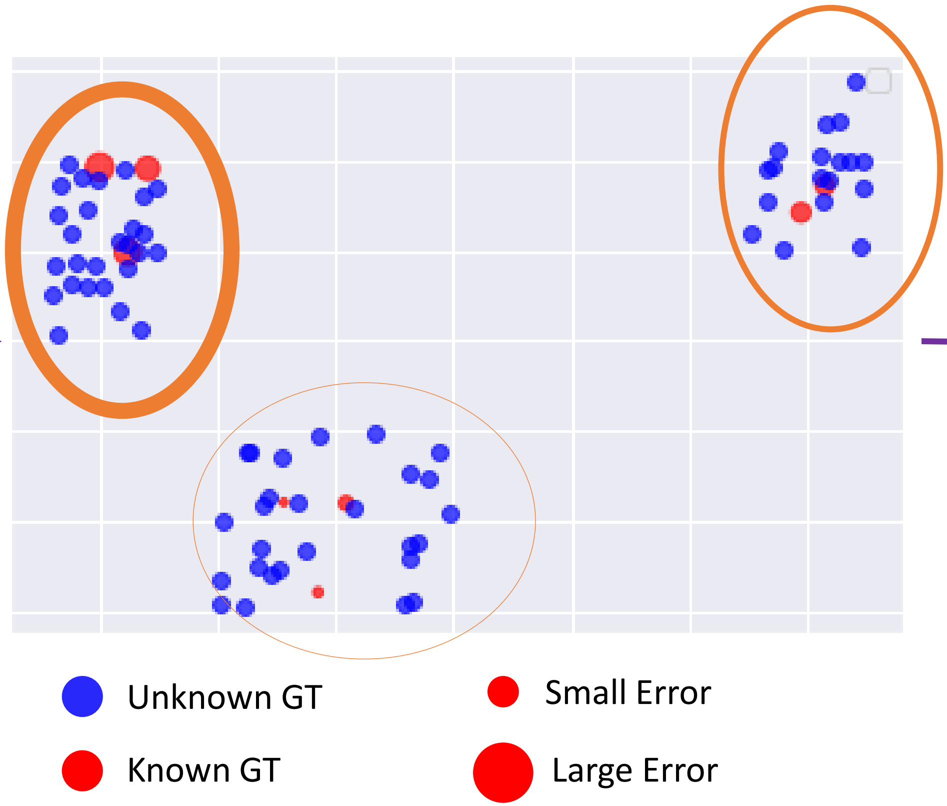
Weight Bound	Price Bound	Card. GT
< 1.0	< 1.0	10
< 2.1	< 2.0	20
...		...

Re-calculate Ground Truth

Weight Bound	Price Bound	Card. GT
< 1.0	< 1.0	?
< 2.1	< 2.0	?
...		?

Drift Case 3: Solution (for Data Shift)

Weight Bound	Price Bound	Card. GT	Error
< 1.0	< 1.0	10	5.3
< 2.1	< 2.0	?	
…		?	
		12	1.2
		?	
		?	
		…	



Weight Bound	Price Bound	Card. GT
> 5.0	< 1.0	3
< 1.5	> 5.2	4
		…

CE Model

Speedup		
PRSA	Poker	HIGGS
3.0x	1.3x	1.5x

Summary of These Drift Cases

Workload Shift

- Drift Case 1: Too Few New Queries
- Drift Case 2: Too Many New Queries
- Drift Case 3: Data Shift

Synthesis with GAN

Picker

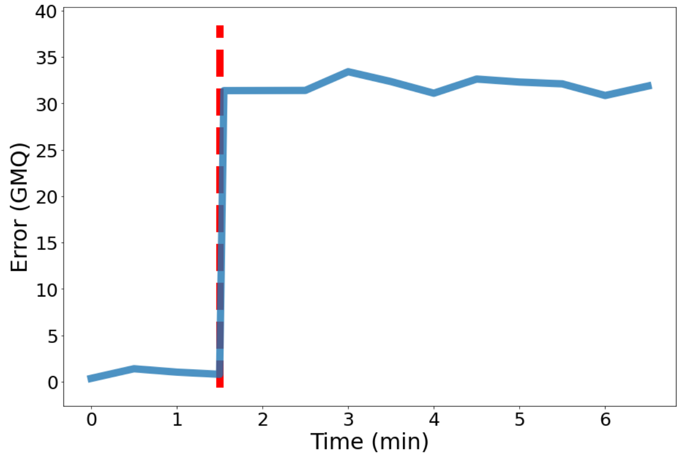
- Workload and Data shifts can happen at the same time.
- Other different drift examples (scenarios), and End-to-End experiment with continuous drifts are shown in the paper.

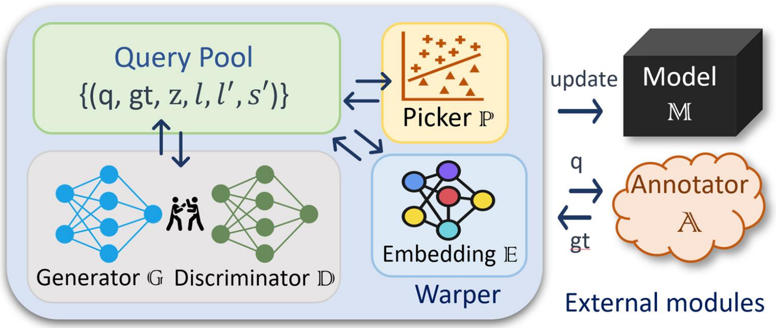
Related Work

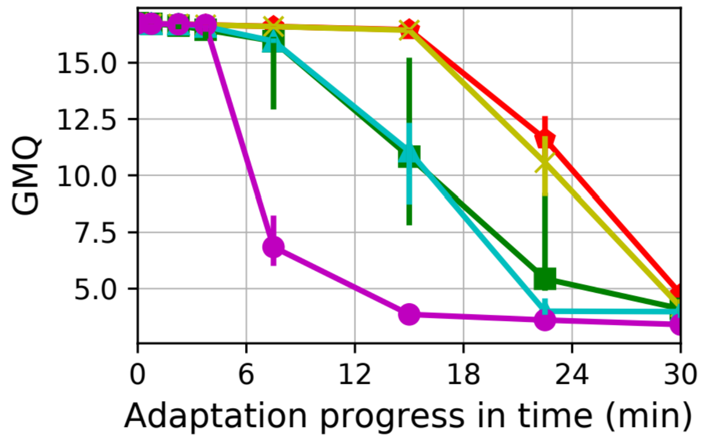
- Active Learning
  - HAL (SIGMOD 19’), ADCP (SIGMOD 20’), Wilds (PMLR 21’), …
- Generative Adversarial Network (GAN)
  - Deep Learning: GAN (NeurIPS 14’), InfoGAN (NeurIPS, 16’), CycleGAN (ICCV 17’), StyleGAN2 (ECCV 20’), TransGAN (NeurIPS 21’)
  - DB: Relative Data Synthesis (VLDB 20’)

Conclusion

- ML for System also Suffers from Data and Workload Shifts
- Create ***Warper*** to Adapt for Cardinality Estimation
  - Low Computation Overhead
  - 3x – 6x Faster Adaptation in Slow.
  - 1-2x Faster Adaptation in Fast.
- Future: Examine More Real Workloads in End-to-End Setting







Thank You!