Active Learning for ML-Enhanced Database Systems

Lin Ma, Bailu Ding, Sudipto Das, Adith Swaminathan

Carnegie Mellon University

Microsoft Research

lin.ma@cs.cmu.edu

Emerging ML-Enhanced Databases

Many academic contributions

Query Run-time Prediction

Query Optimization

Index Recommendation

Autonomous Administration

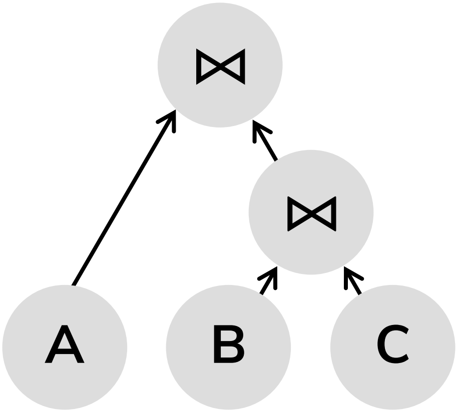
Challenge at deployments

ML-Enhanced Database Example

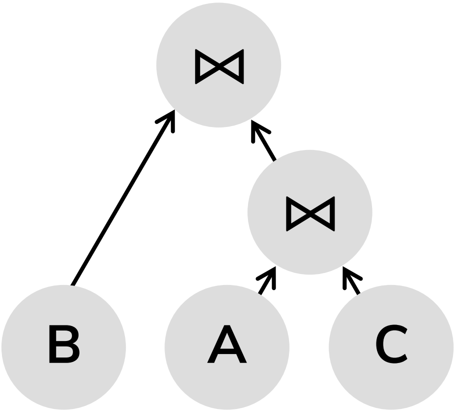
ML Model:

- [Ding, B., et al. SIGMOD 2019]

Model Input



P1

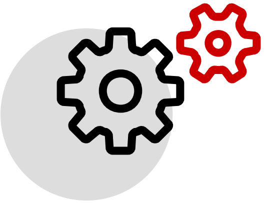


P2

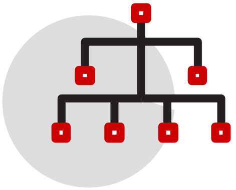
Model Output

P1 is cheaper than P2

Applications

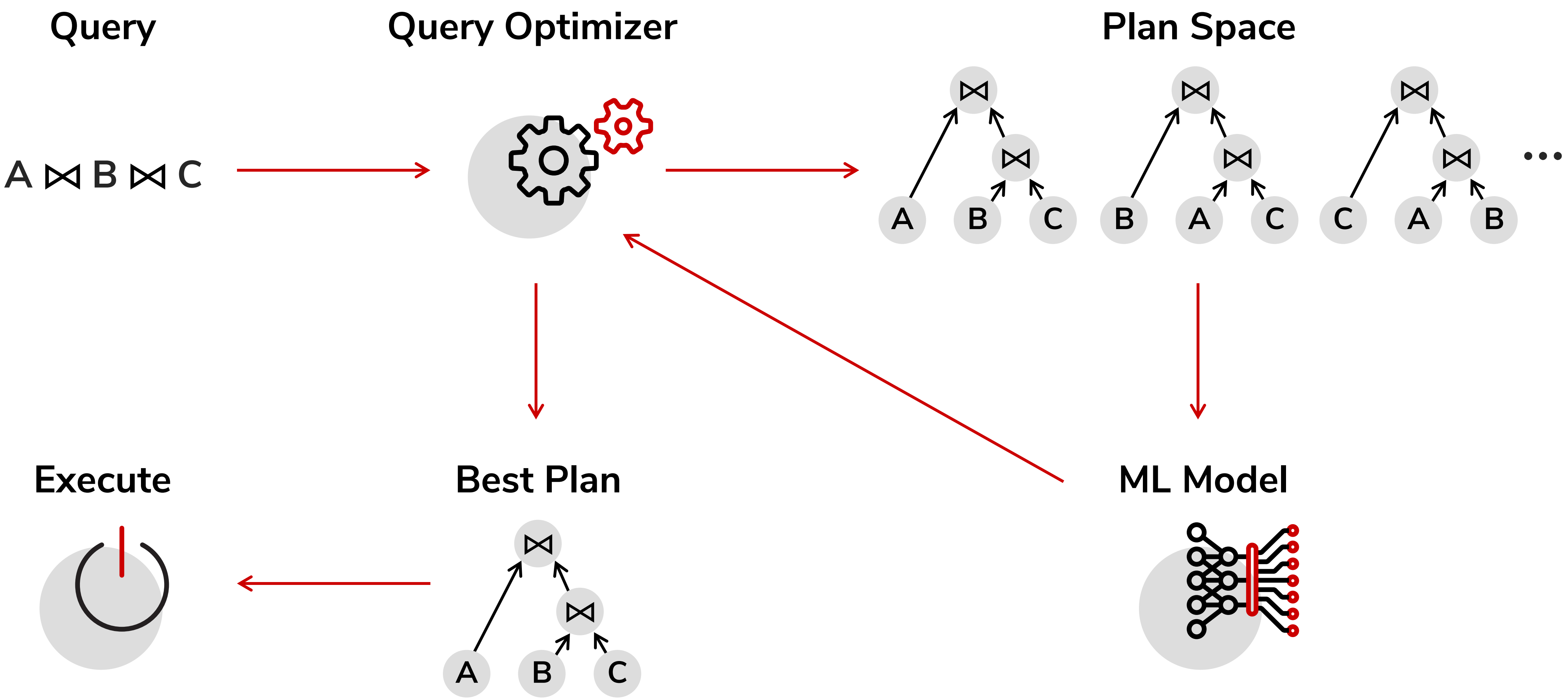


Query Optimizer



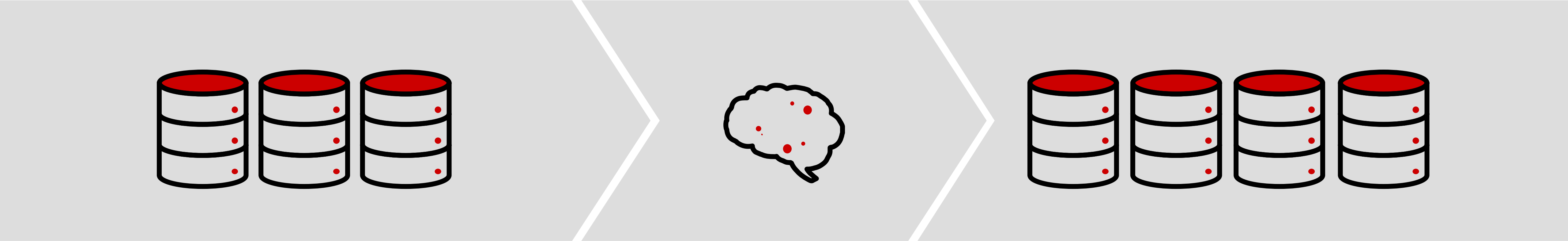
Index Advisor

ML-Enhanced Database Example



Simulated Model Training and Deployment

Collect Data | **Train** | **Deploy (simulated)**



Standard Benchmarks
and Available Workloads

Training Error: 2%
Validation Error: 5%

Error: 32%

What’s wrong?

Challenge: Data Distribution Shift

ML assumes same training-test distribution

Test data distribution varies heavily in production databases

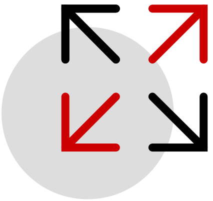


Table Sizes

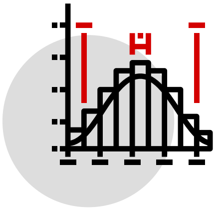
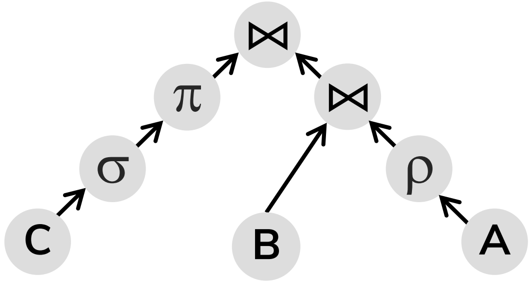


Table Data Distributions



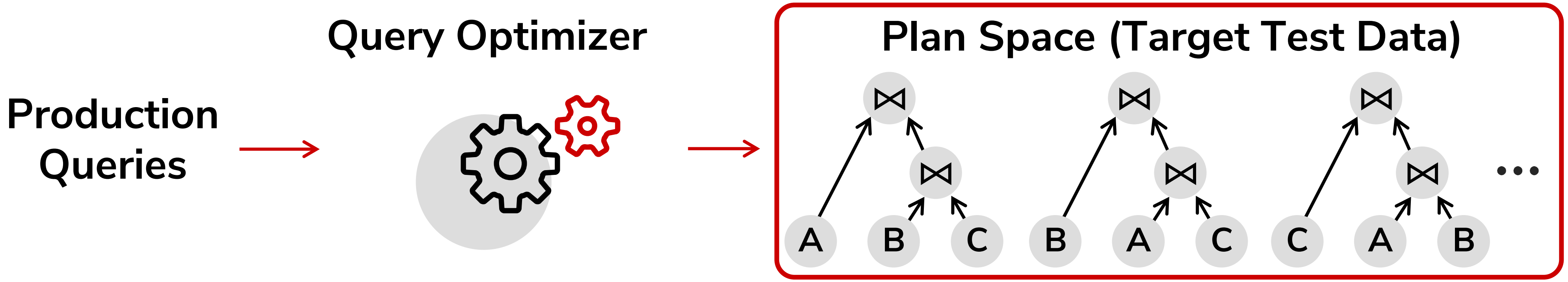
Operator Combinations

Key barrier to productionize ML for databases

Solution: Collect More Data in Deployments

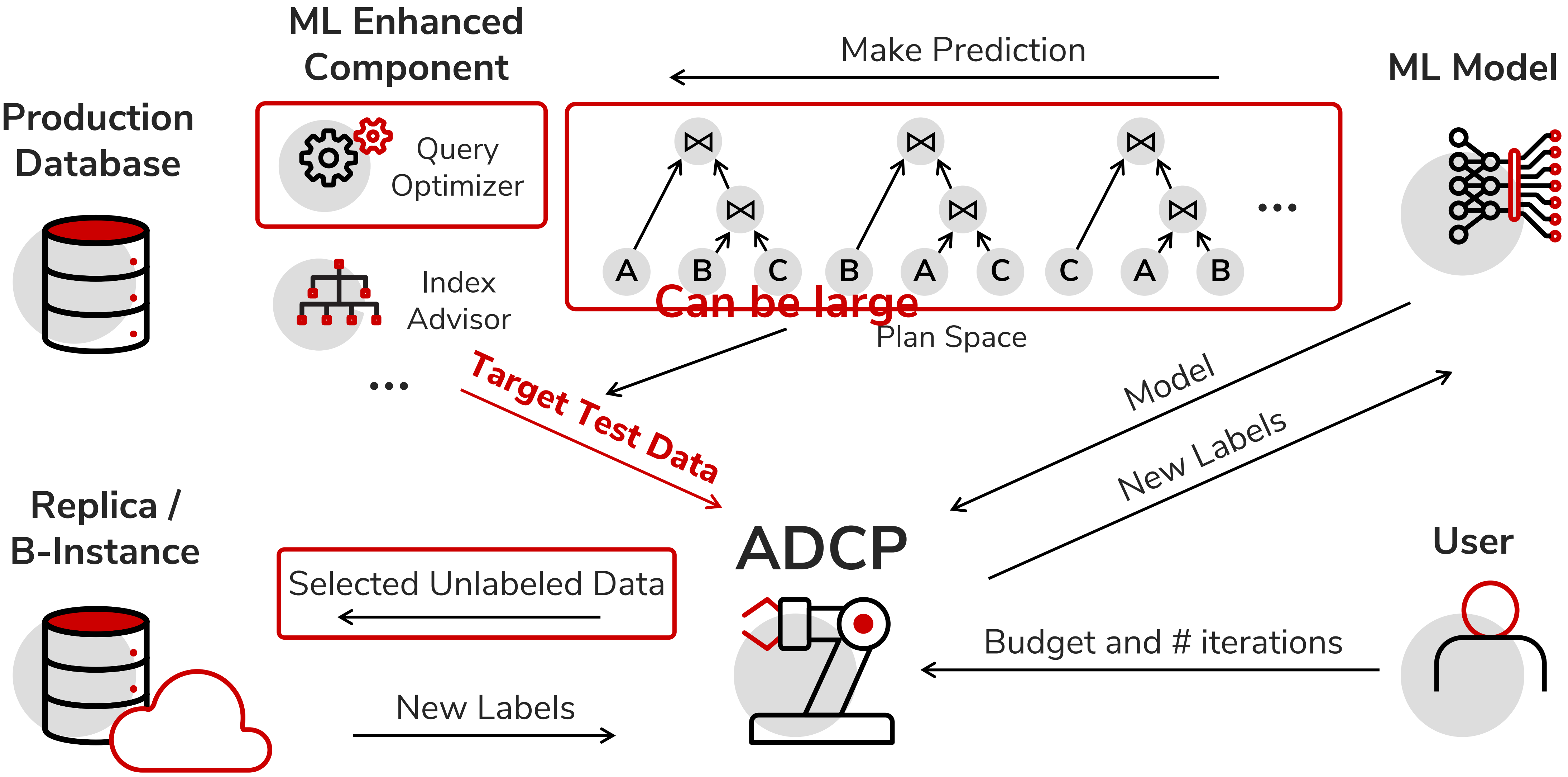
Insight: *actively collect data for individual database deployments*

- Acquire labels from replicas (b-instances) without impacting the normal operation
- The "target test data" is often derivable for a specific workload



Reduces 75% error by executing ~100 queries

Active Data Collection Platform



Active Learning

AL strategy selects the best training data from a pool of unlabeled data

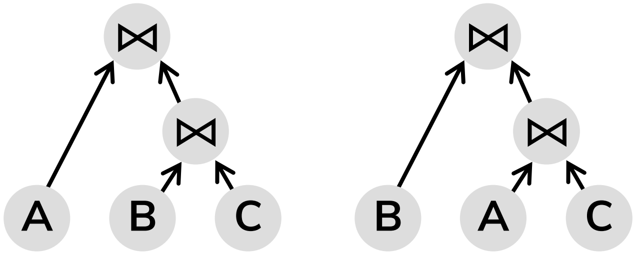
- Long and successful history in database crowdsourcing

Typical AL:

w(x)	↓	↓	↓	↓					
	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
	x1	x2	x3	x4	x5	x6	x7	x8	x9

Most common w(x): uncertainty

Model Input



P1 **P2**

Model Output

P1 is cheaper than P2: **70%**
P2 is cheaper than P1: **30%**

Uncertainty: 30%

Holistic AL Challenges

Robust

Noisy uncertainty signal under significant distribution shift

Cost-sensitive

Drastically different labeling costs, especially with index creations

Batch-friendly

Expensive model retraining

Holistic AL Challenges

AL Strategy	Robust	Cost-sensitive	Batch-friendly
Uncertainty			
Cost		Y	
Hybrid	Y		
RBMAL			Y
ROUND		Y	Y
SIMILAR		Y	Y

Fertile area of future research

Holistic Active Learner (HAL) for ADCP

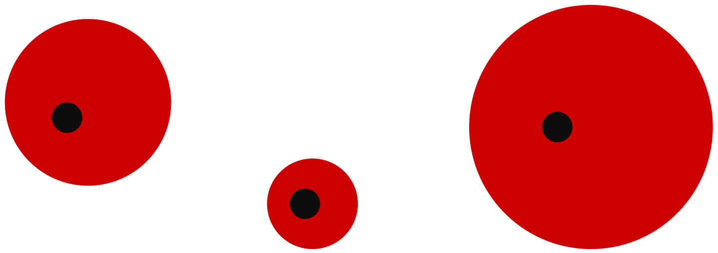
Biased sampling: robust

	↓	↓		↓		↓			
w(x)	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
	x1	x2	x3	x4	x5	x6	x7	x8	x9

Cost weighting: cost-sensitive

- Per “cost unit” uncertainty

Redundancy rejection: batch friendly



Evaluation

14 workloads include industrial standard benchmarks (e.g., TPC-DS) and customer workloads

- Hold out each workload as the target production database, and round robin
- 30K plans, 1M plan pairs

Multiple AL iterations with evenly split budget for each iteration

- Total budget of 150x average estimate plan cost

Different ML tasks, budget sizes, models, features, cost types, or no cost estimation

Baselines

Optimizer

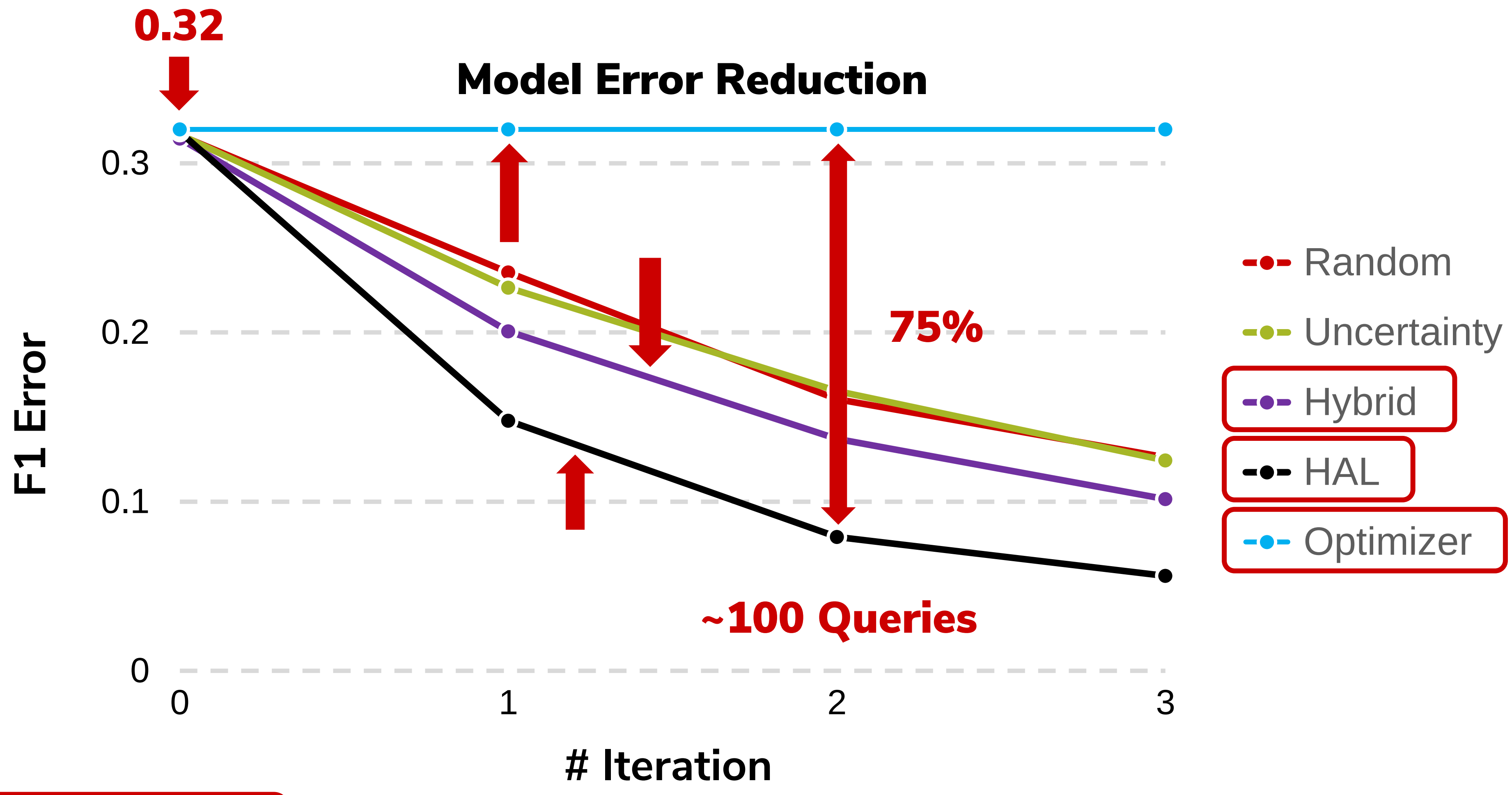
Random

Uncertainty

Hybrid

- Random + Uncertainty
- [Hass, D., et al. VLDB 2015]

Results



Budget: 50x average query cost per iteration

Takeaway

Addressing the training/deployment distribution shift is crucial for ML-enhanced databases

A practical solution to actively collect training data during deployment using replicas and HAL

Fertile area of future research

- Better address the holistic AL challenges
- Better use the training data during deployments

lin.ma@cs.cmu.edu