Active Learning for ML Enhanced Database Systems

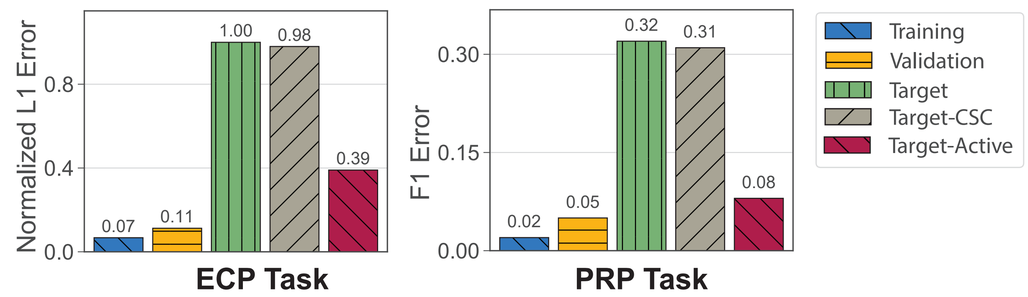
Lin Ma, Bailu Ding, Sudipto Das, Adith Swaminathan

Background

- ML enhanced databases are emerging
  - › E.g., academic contributions on ML enhanced query optimizer, index advisor, or query latency predictor
- ML enhanced databases face challenges at deployments: the ML model's prediction error increases significantly
- Key reason:
  - › Data distribution shift due to differences on table sizes, table data distributions, and operator combinations on production workloads

Motivation Experiment

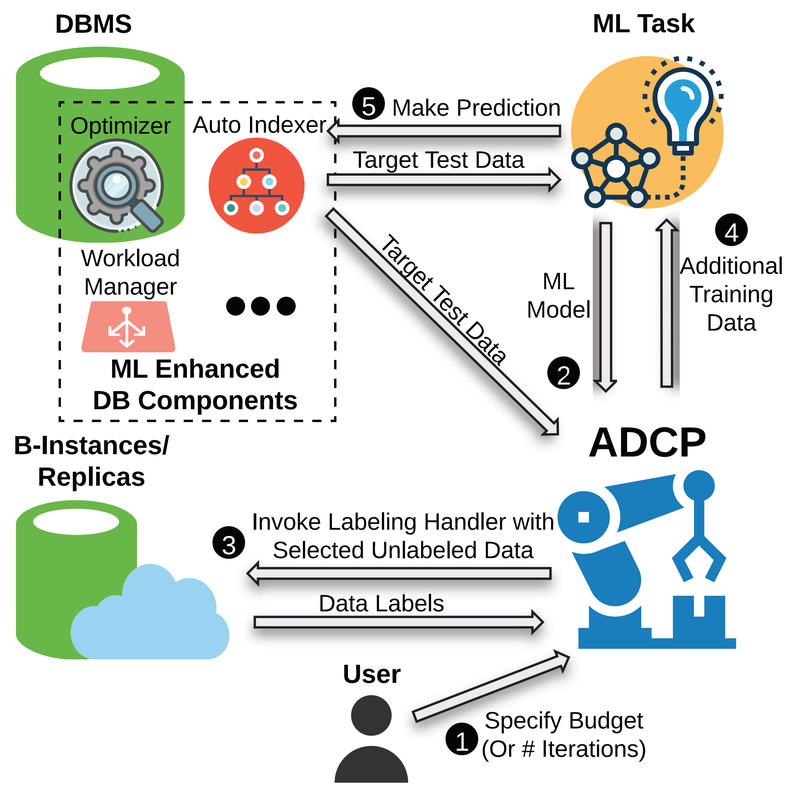
- ECP Task: given a query plan, predict its execution cost
- PRP Task: given two query plans, predict which is cheaper
- Applications: ML enhanced query optimizer or index advisor



Solution

- Actively collect additional training data for individual databases during deployments
  - › Using B-instances or replicas to execute additional queries without impacting normal business operation
  - › Focus on "target test data" for specific prod. workload
    - E.g., the optimizer's plan space for a set of queries
- A general platform supporting various ML tasks for DBs

Active Data Collection Platform

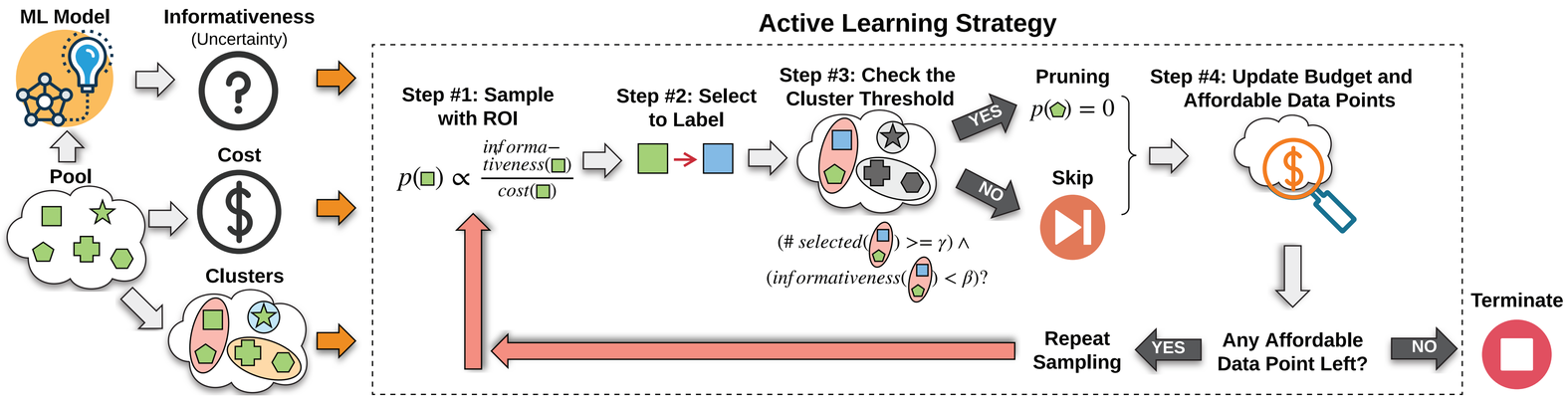


Active Learning

- Challenge: target test data can be large
  - › E.g., optimizer enumerates hundreds of thousands of plans for complex queries
- Active learining: select best training data from a pool of test data to improve ML model
  - › Typically define an informativeness score (e.g., uncertainty), then acquire the most informative labels

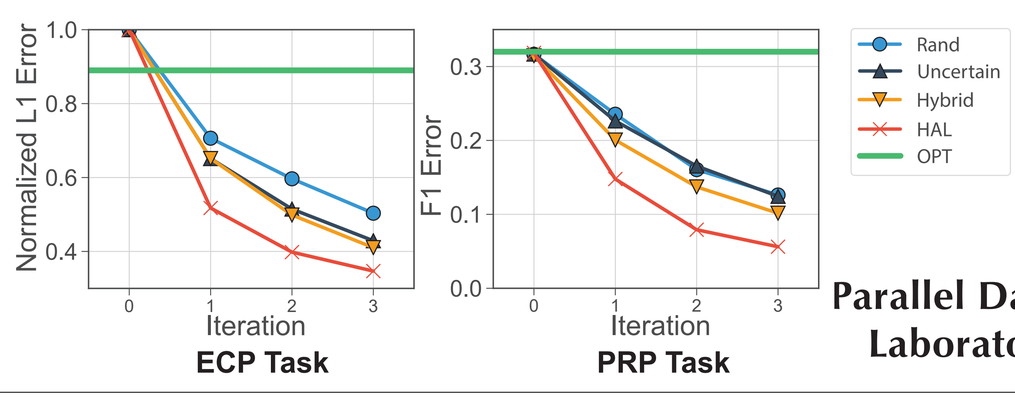
Holistic Active Learner

- Challenge: canonical active learning has not focused on several holistic challenges arising from DB deployments
  - › Noisy uncertainty, labeling cost differences, and batch-labeling before retraining
- HAL: a new active learning strategy that is robust, cost-sensitive, and batch-friendly
  - › Techniques: biased sampling, cost weighting, and redundancy rejection



Results

- Reduces up to 75% test error on new databases at deployments by executing ~100 queries



Carnegie Mellon University

Parallel Data Laboratory

PDLSVD20