Data Distribution Tailoring Revisited:
Cost-Efficient Integration of Representative Data

MOTIVATION

WANT:
- Cheap & representative data.
- Model charging for difficult subgroups.

REALITY:
- Data collection is expensive.
- Existing datasets are biased.

SOLUTION:
1. Combine existing data from data repositories.
2. Enforce group count requirements.

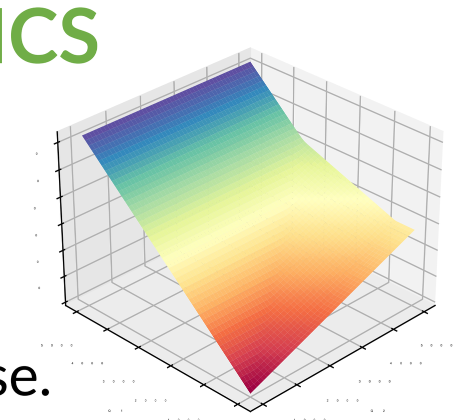
PROBLEM DEFINITION

GIVEN:
- Data sources with sampling costs.
- Groups with minimum count requirements.

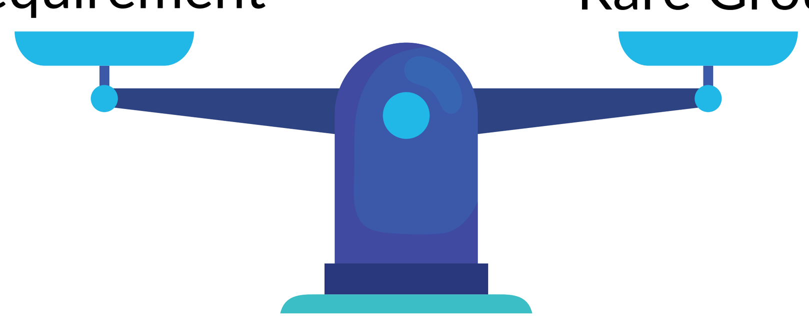
GOAL: Minimize expected total query cost.
CONSTRAINT: Satisfy minimum count requirement.
QUERY MODEL: Uniformly random sampling.

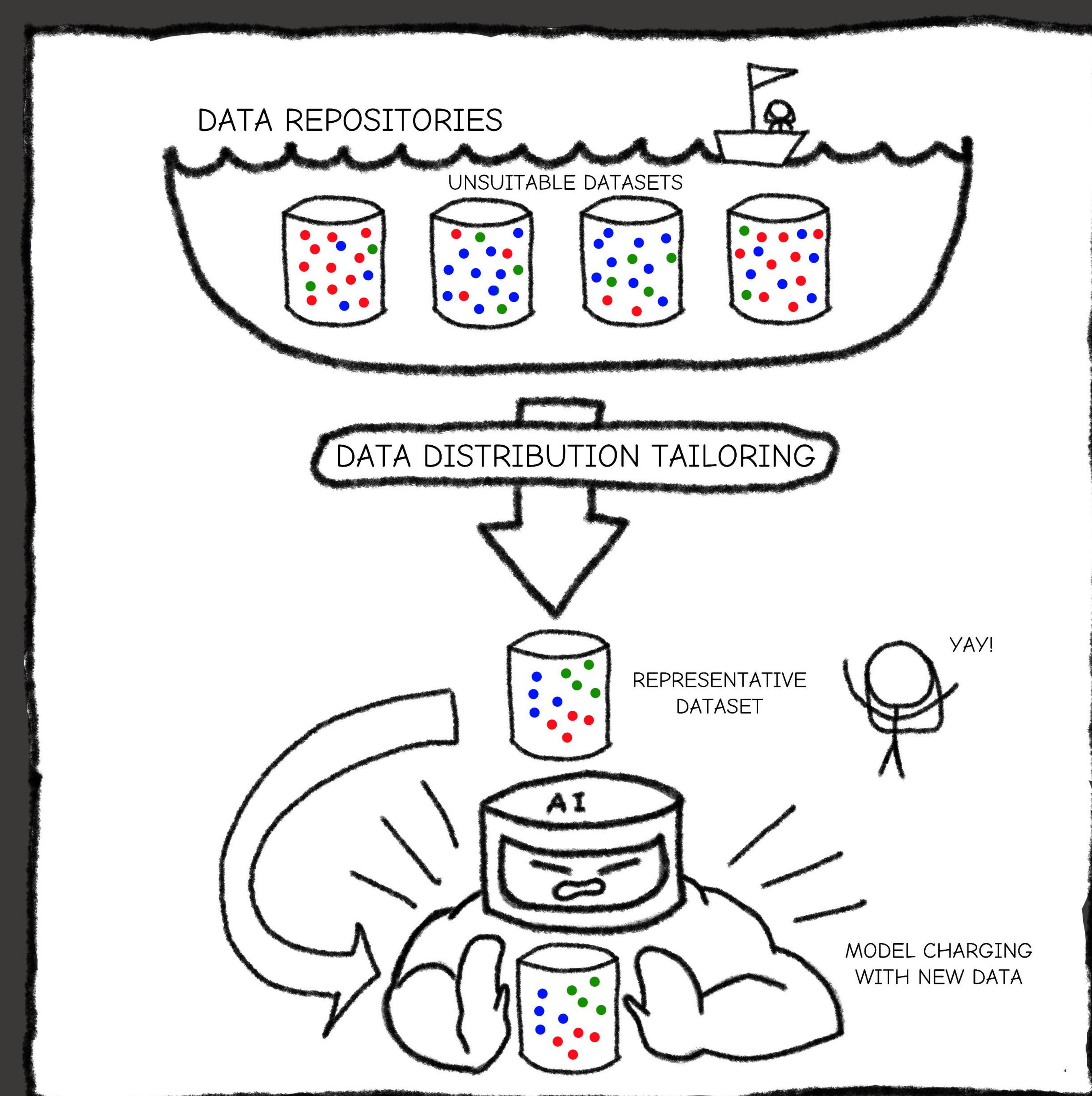
KNOWN STATISTICS

DYNAMIC PROGRAMMING
- Optimal but slow $O(Q^n)$ time.



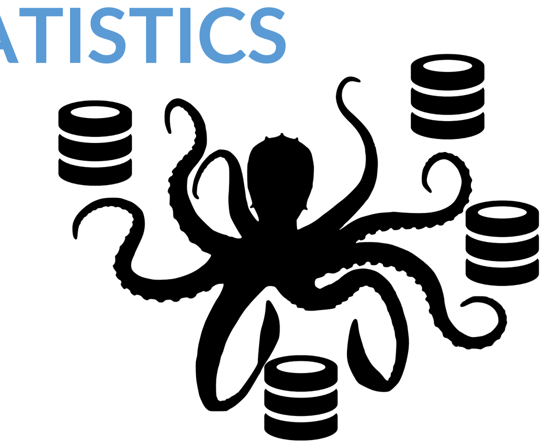
HEURISTIC ALGORITHM **RatioColl**
- Only $\sqrt{Q}$ overhead cost in special case.
- Linear upper bound in general.





UNKNOWN STATISTICS

ε-GREEDY BANDIT
- Same heuristic as **RatioColl**.
- No priors needed.
- Sublinear regret bound.



ALGORITHMS

RatioColl

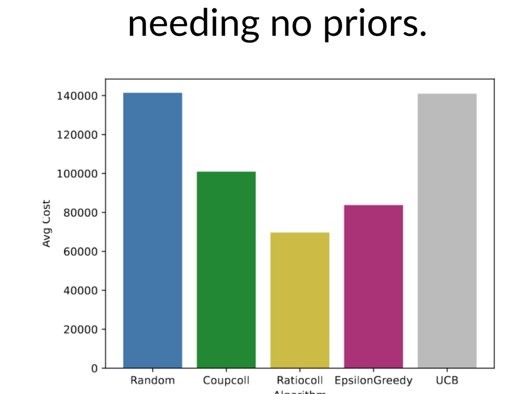
$O \leftarrow \emptyset$
while query is not satisfied:
  $G^* = \text{argmax}_{G_j, Q_j > 0}\left(Q_j \cdot \min_{i\in[n]}\left(\frac{C_i}{P_{i,j}}\right)\right)$ // Choose priority group
  $D^* = \text{argmin}_{D_i}\left(\frac{C_i}{P_{i,*}}\right)$ // Maximize priority group per cost
  $s \leftarrow \text{Query}(D^*)$
  $O \leftarrow O \cup \{s\}$
return O

EpsilonGreedy

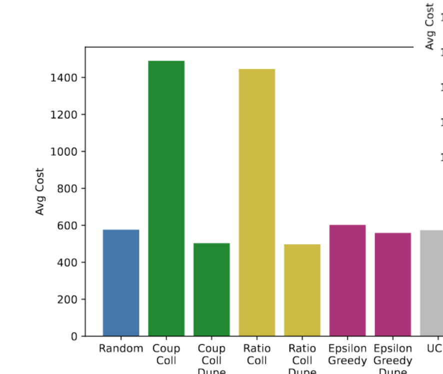
$O \leftarrow \emptyset$
while query is not satisfied:
  if $t \leq n$ **then** $D^* \leftarrow D_t$ // Initialization
  else:
    with probability $p = \sqrt[3]{\ln t / t}$: // Exploration round
      $D^* \leftarrow$ random data source
    else: // Greedy exploitation round
      let $\overline{P_{i,j}} \leftarrow$ estimate of $P_{i,j}$ based on sample mean
      $R(G_j) \leftarrow \left(Q_j \cdot \min_{i\in[n]}\left(\frac{C_i}{\overline{P_{i,j}}}\right)\right)$ for each G_j // Reward of group
      $D^* \leftarrow \text{argmax}_{D_i}\left(\frac{1}{C_i}\sum_{j\in[m]} \overline{P_{i,j}} \cdot R(G_j)\right)$ // Reward of data source
  $s \leftarrow \text{Query}(D^*)$
  $O \leftarrow O \cup \{s\}$
  update trackers for probability estimation
return O

SELECTED RESULTS

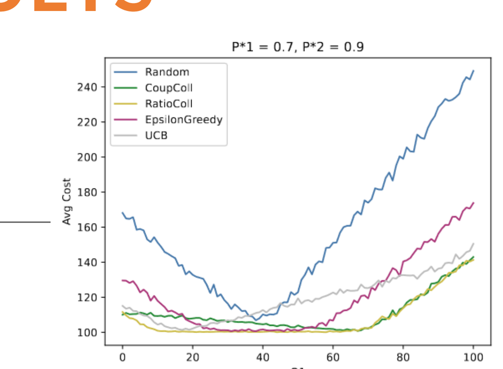
- RatioColl consistently out-performs SOTA.
- EpsilonGreedy competitive with SOTA despite needing no priors.



Flights Dataset



COMPAS



Synthetic Benchmark

Jiwon Chang[1], Bohan Cui[1], Fatemeh Nargesian[1],
Abolfazl Asudeh[2], H.V. Jagadish[3]
[1]University of Rochester [2]University of Illinois Chicago [3]University of Michigan