

# Approximating User-Defined Functions in BlinkDB

Sameer Agarwal, Ariel Kleiner, Henry Milner, Barzan Mozafari\*, Aurojit Panda, Purnamitra Sarkar, Ameet Talwalkar, Michael Jordan, Samuel Madden^, Ion Stoica (U Michigan-Ann Arbor\*, MIT^, UC Berkeley)

## BACKGROUND

```

SELECT avg(sessionTime)
FROM Table
WHERE city='Berkeley'
GROUP BY dt, os, isp
WITHIN 2 SECONDS

SELECT avg(sessionTime)
FROM Table
WHERE city='Berkeley'
GROUP BY dt, os, isp
ERROR 0.1 CONFIDENCE 95.0%
    
```

BlinkDB augments standard SQL-like queries to provide Response Time or Error/Confidence Interval Guarantees

**Offline Sampling Module:** BlinkDB maintains a set of *multi-dimensional* and *multi-granular* samples offline. These samples are either striped on hundreds of disks or cached in memory.

**Per-Query Sample Selection Module:** This module predicts per-query cost in BlinkDB and assigns an appropriately sized sample based on their error and/or response time requirements.

**Error Bars and Confidence Intervals:** All answers are then augmented with statistical measures of error bars and confidence intervals based on either statistical closed form measures (for AVG, SUM, COUNT, VARIANCE, PERCENTILES etc.) or Bag of Little Bootstraps technique (for arbitrary User-Defined Functions)

## MAINTAINS UNIFORM & BIASED SAMPLES

MILP optimization picks the set of columns to stratify on within a storage budget

ID	City	Ad Clicks
1	NYC	5
2	NYC	6
3	NYC	7
4	Berkeley	12
5	NYC	9
6	Berkeley	7

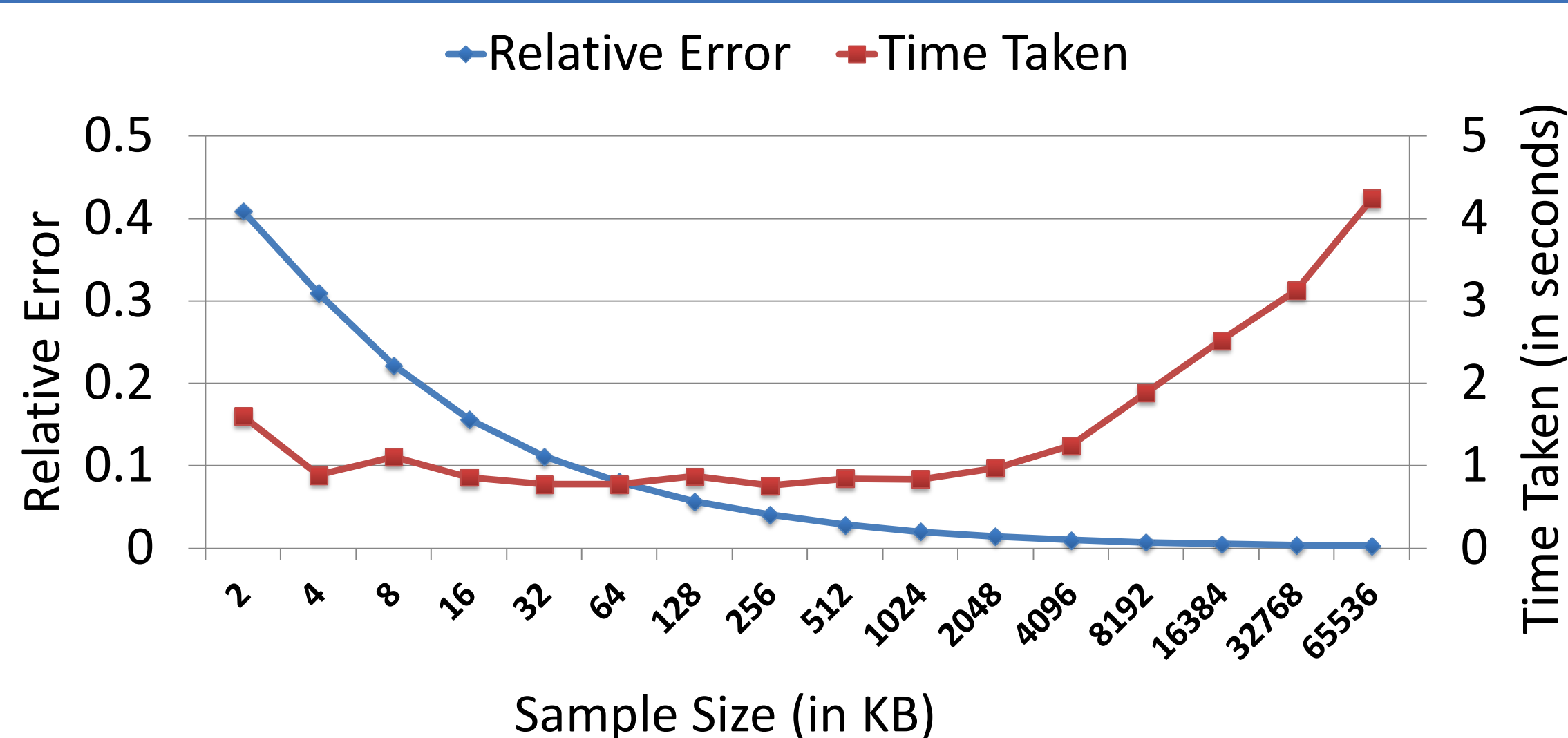
Uniform →

ID	City	Ad Clicks
1	NYC	5
3	NYC	7
5	NYC	9

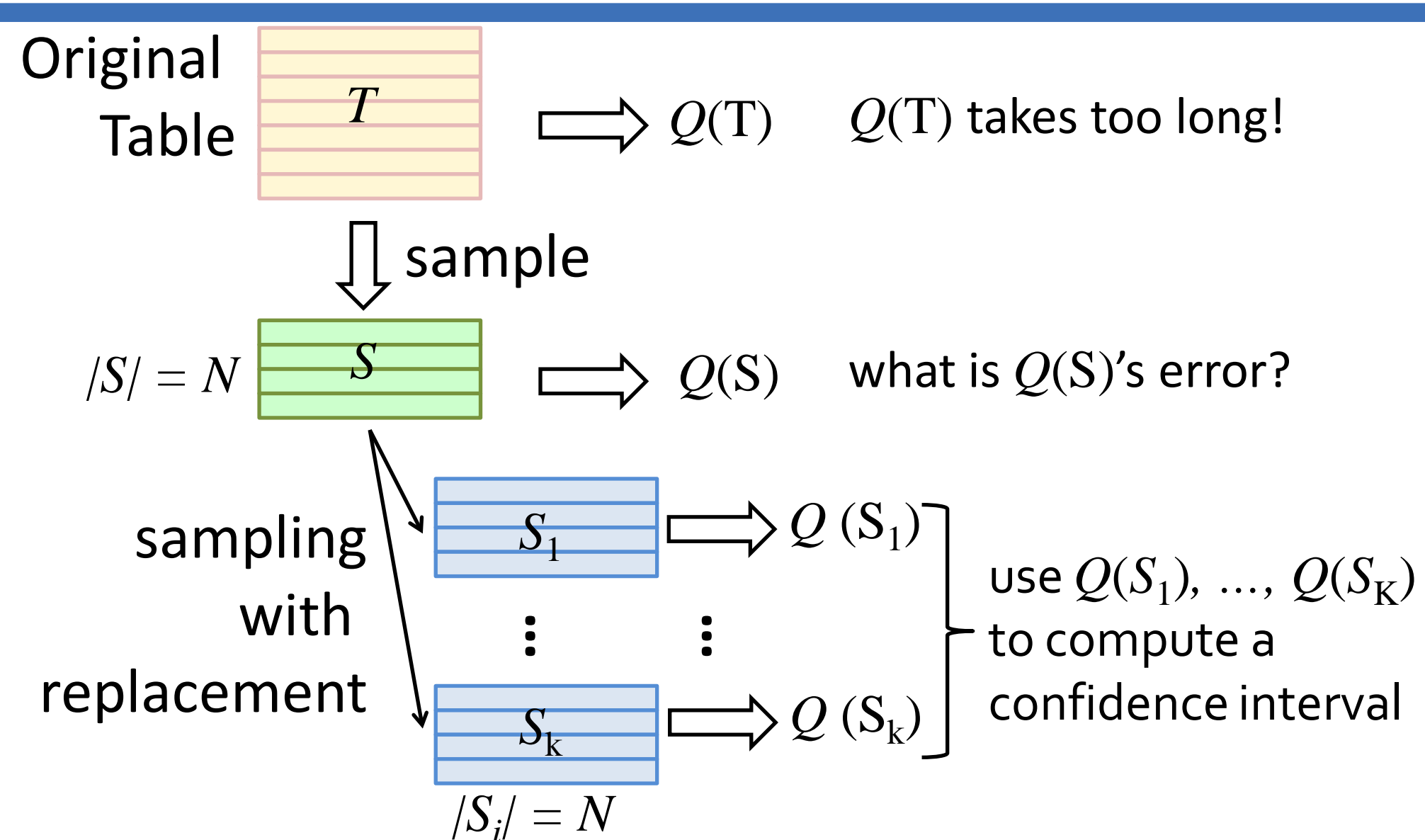
Biased/Stratified →

ID	City	Ad Clicks
1	NYC	5
3	NYC	7
6	Berkeley	7

## ERROR-LATENCY PROFILE

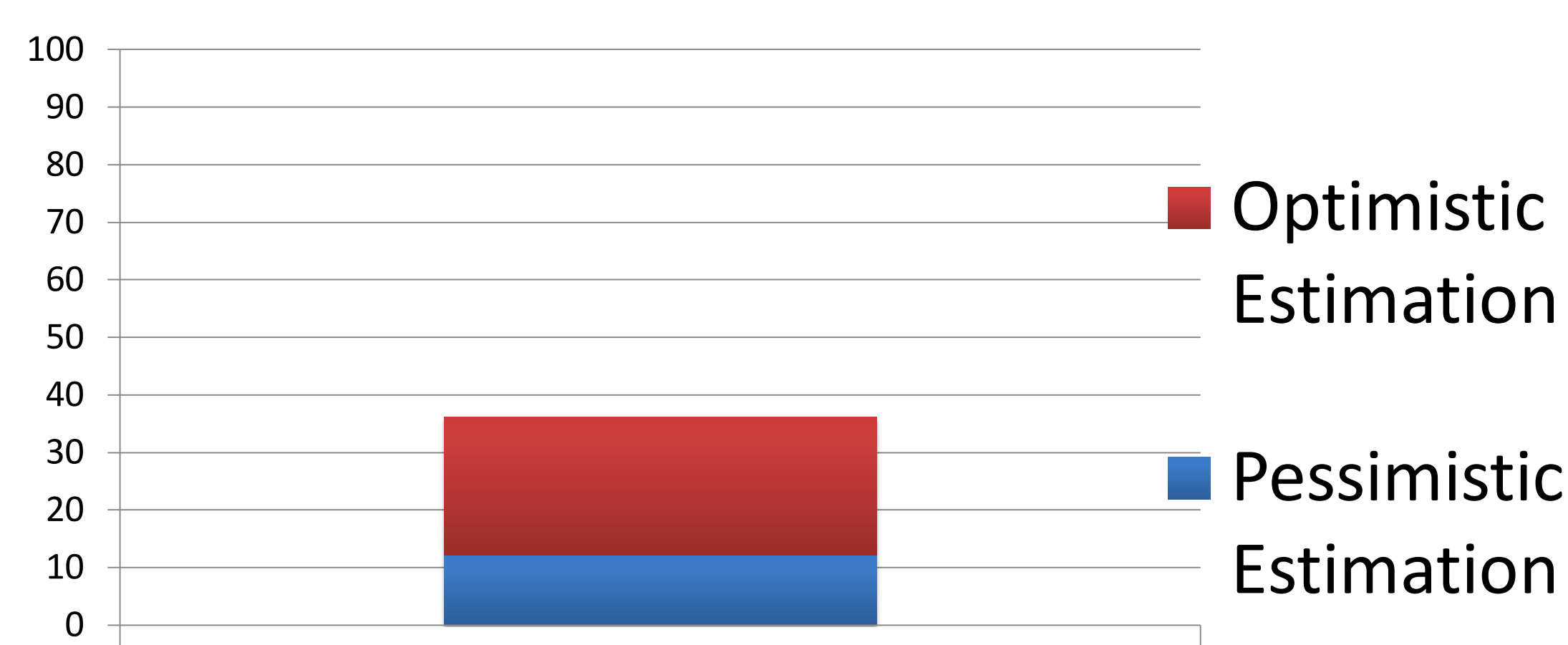
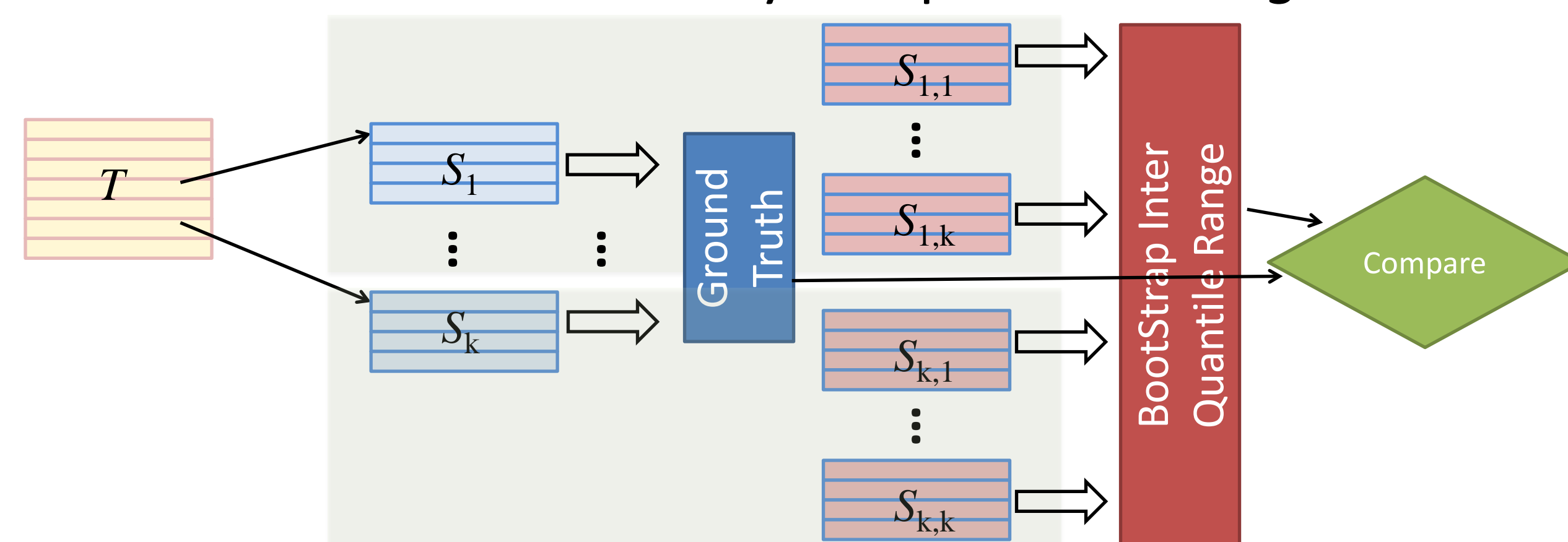


## ERROR QUANTIFICATION VIA BOOTSTRAP



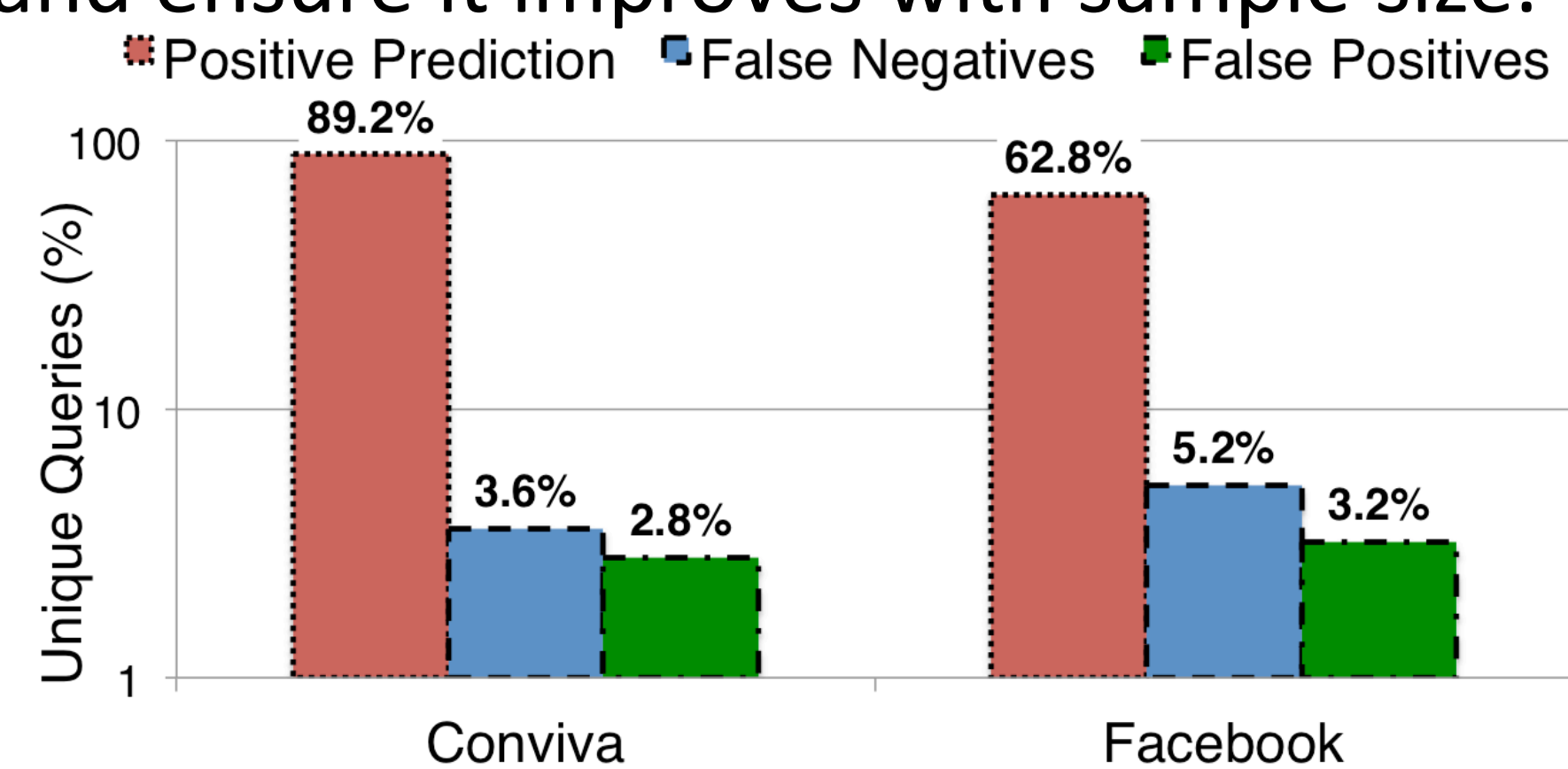
## RELIABILITY OF THE BOOTSTRAP

Bootstrap is not always reliable. Verification on a per-query/per data-distribution basis by comparison with *ground truth*.

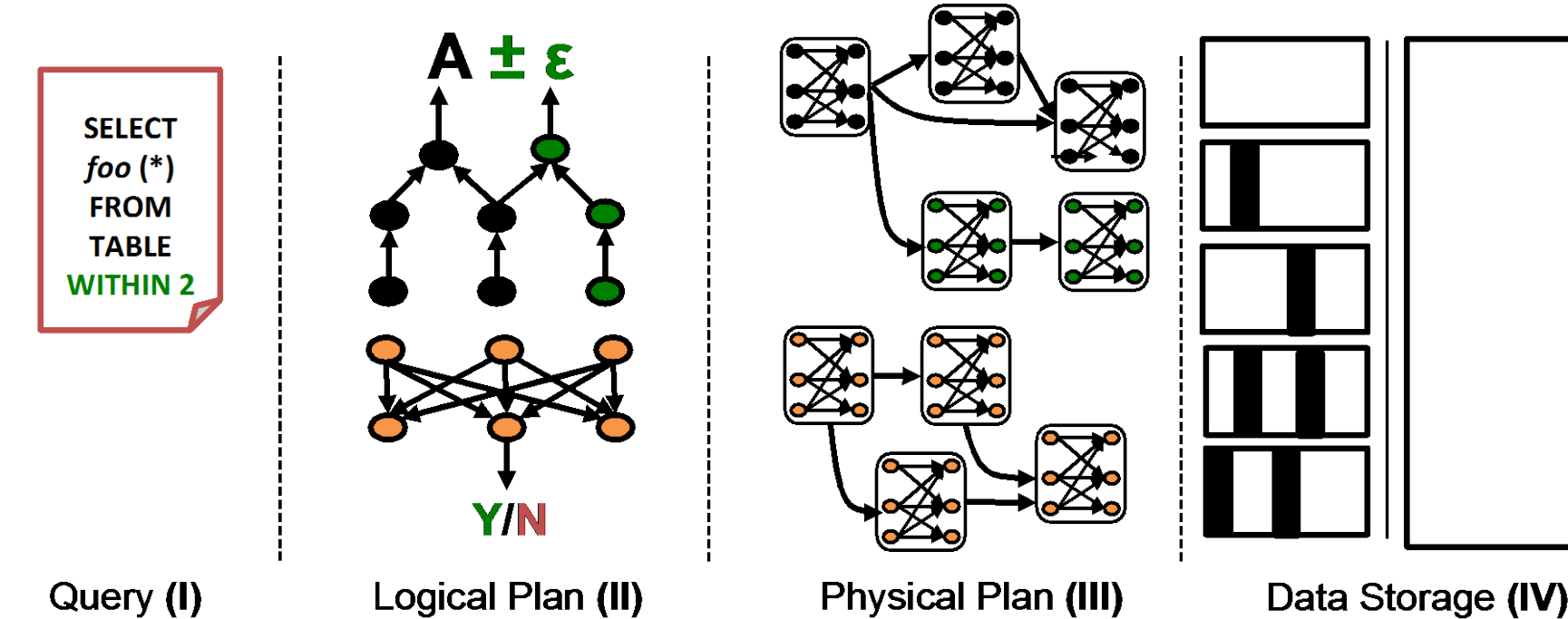


## RUNTIME DIAGNOSTICS

This verification is too slow. Good heuristic (Kleiner et al. 2012): Compare bootstrap with ground truth for *small samples* and ensure it improves with sample size.



## DESIGNED FOR INTERACTIVITY



- Filter Pushdown
- Task Consolidation
- Input Caching

## END-TO-END PERFORMANCE

