Trust but Verify: Archaeologist

Sahil Bhatia

Based on: Sample and Seek

1. **AQP** trades off **accuracy** for **speed** in data analysis
    a. Generates approximate answers to queries using sampling or statistical techniques
2. Issues with existing AQP:
    a. Precision metrics
        i. Confidence Intervals : focus on estimating parameters within each group or subset but may not reveal the overall distribution errors
    b. Unbounded Errors
        i. CI Width $\propto$ std(Sm) / $\sqrt{}$m.

Sample and Seek

Support AQP with a user-specified error bound

Sample and Seek

1. Distributional Precision
    a. Measures overall precision over all the groups
2. **L2 distance** between normalized distributions of the **approximate answer** and the **exact one**.
3. System guarantees to produce approximations bounded by user-specified error ε.

Sample and Seek

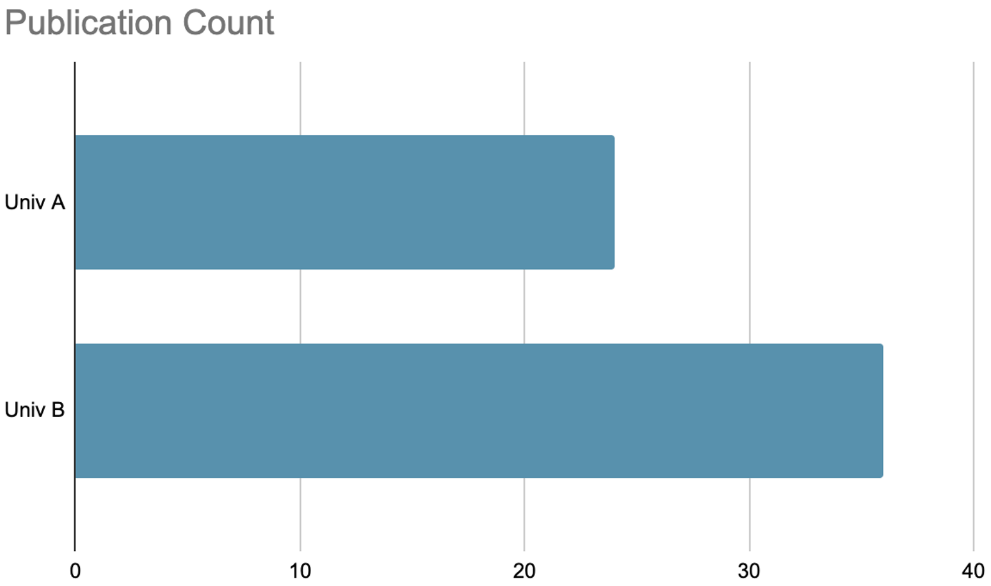
1. Measure-based Sampling
    a. Rows with higher measure attribute values are more likely to be included in the sample.
    b. A($25), B($20), C($10)
2. Indexes for selective predicates
    a. measure -augmented inverted index - maps values to their respective rews
    b. low- frequency group index - identify rows that belong to low-frequency groups and store them sequentially on disk

Influenced: ProReveal: Progressive Visual Analytics With Safeguards

1. Progressive Visual Analysis allows access to partial result in middle of computations
   a. Infeasible to compute precise results
2. Not all systems guarantees how **long** they have to wait for **exact results**
3. Users can take a decision using partial results
   a. Account for worst case

ProReveal

1. Represent intermediate knowledge as as guards (logical formulas).



pubCount(Univ A) < pubCount(Univ B)

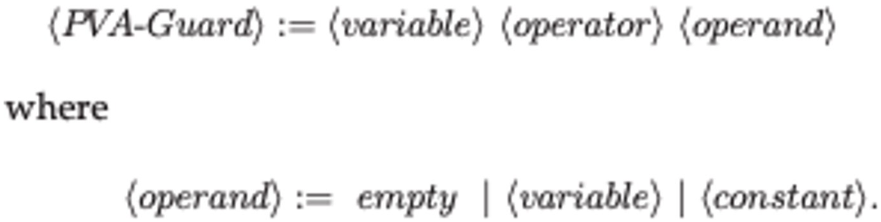
ProReveal

1. Represent intermediate knowledge as as guards (logical formulas).
2. System continuously gives feedback on validity
    a. If invalidated systems notifies user

ProReveal

1. Moritz et al.’s research on optimistic visualization [9] is one of the studies that motivated this research.
2. Differences:
   a. Precise results is not obtainable
   b. Continuous feedback on validity of guard
   c. Intermediate knowledge can be represented structurally

ProReveal



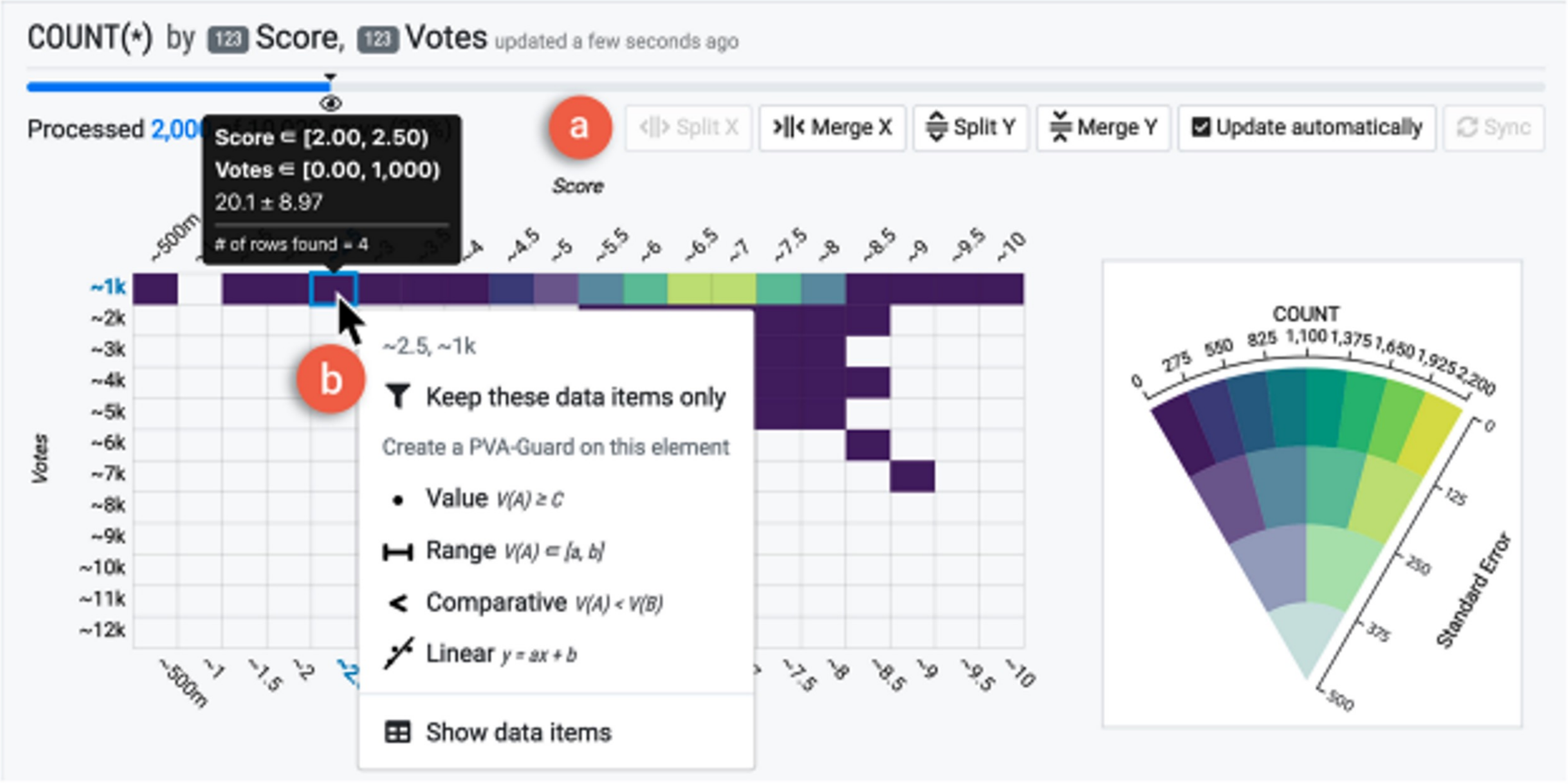
$$\langle \textit{PVA-Guard} \rangle := \langle \textit{variable} \rangle \ \langle \textit{operator} \rangle \ \langle \textit{operand} \rangle$$

where

$$\langle \textit{operand} \rangle := \ \textit{empty} \ \mid \langle \textit{variable} \rangle \mid \langle \textit{constant} \rangle.$$

1. Variables can be single value (cell of heatmap) or even a distribution of values

ProReveal



COUNT(*) by 123 Score, 123 Votes updated a few seconds ago

Processed 2,00

Score ∈ [2.00, 2.50)
Votes ∈ [0.00, 1,000)
20.1 ± 8.97
# of rows found = 4

~2.5, ~1k