Accelerating Query Explanations Using Fine-Grained Provenance

Alex Yao, Columbia University

Columbia University Data Science Institute

Background & Motivation

User Complaint: Too High!

Explanation Engine

Counterfactual Deletions:
```
DELETE state="PA"
DELETE state="PA" AND age=30
DELETE name="John Doe"
…
```

Query Re-execution

Top Score

Delete John Doe from Table A
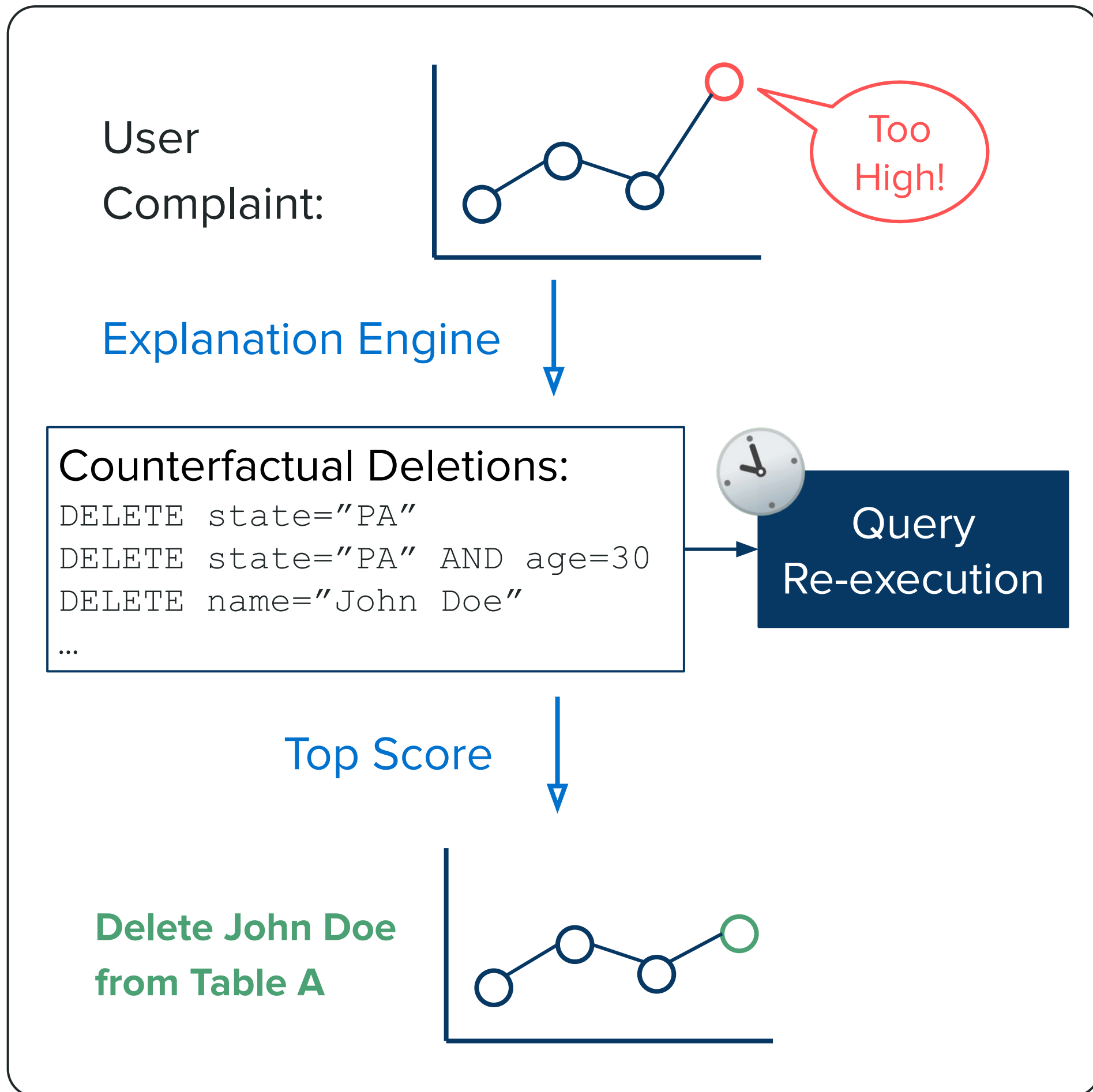
Problem Statement

Explanation engines are bottlenecked by **Query Re-execution** 🕐, which naively needs to run the original query once for each proposed deletion.
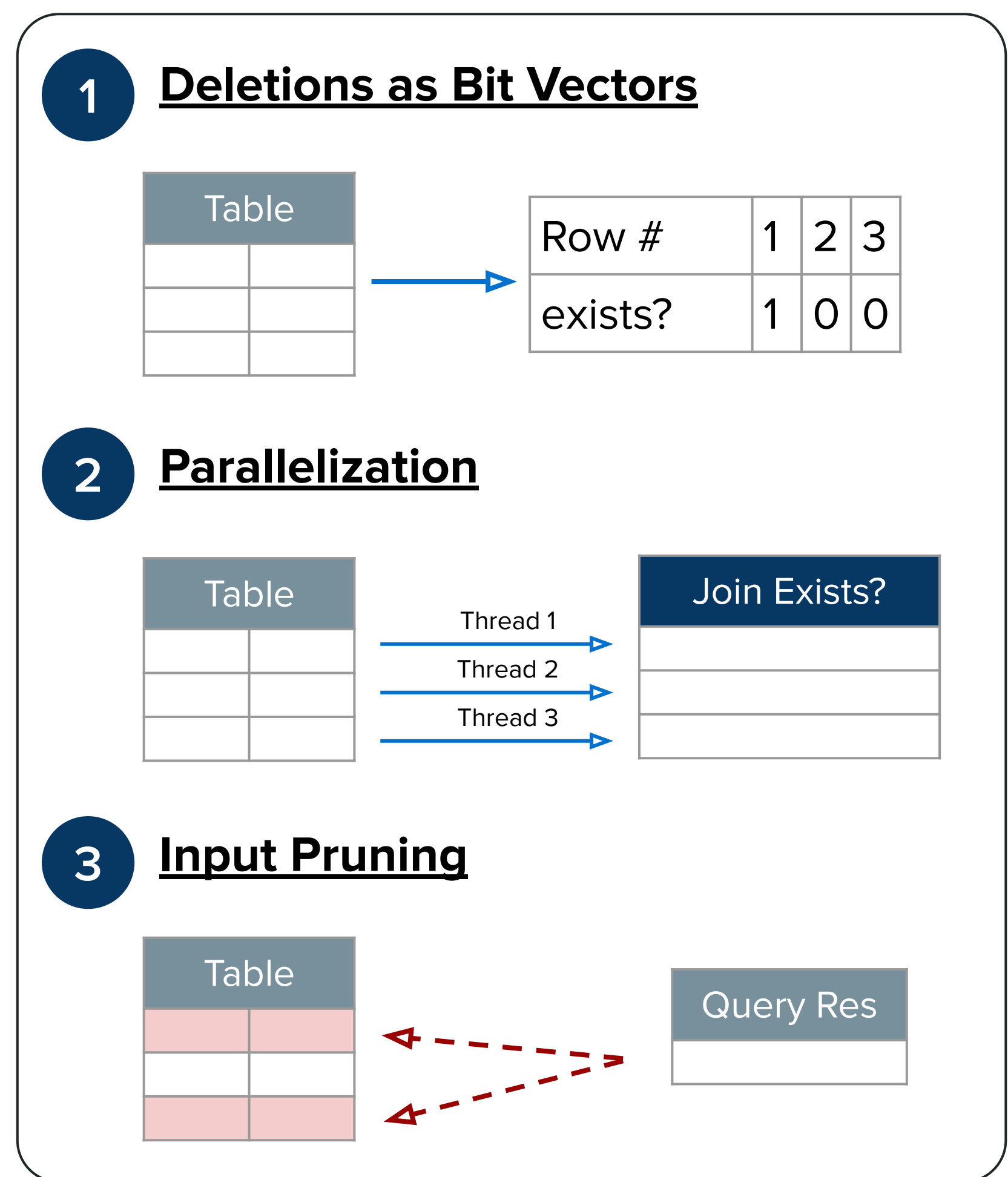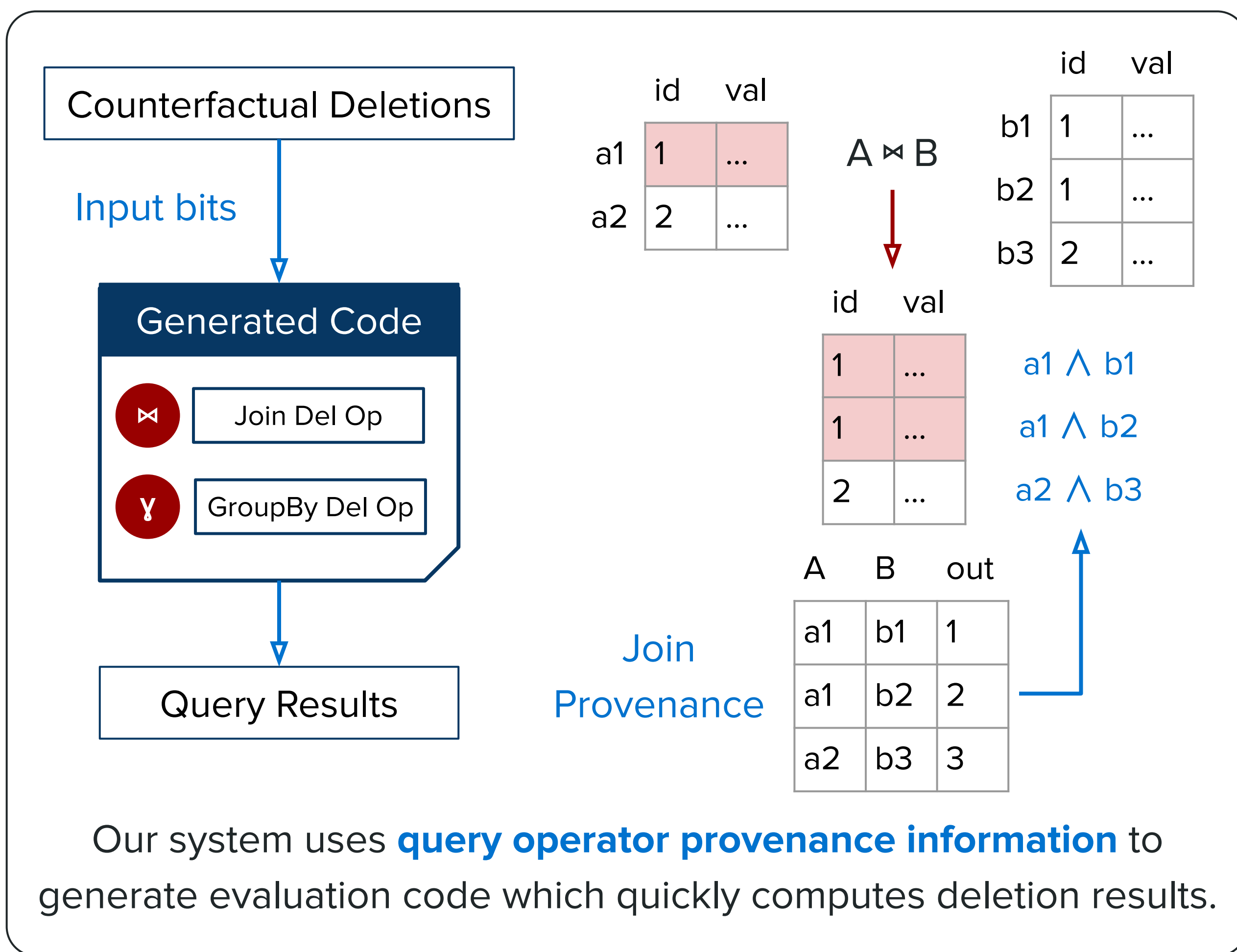
Existing Approach 1: IVM

Poor performance over large deletions and complex queries

$(A - \nabla A) \bowtie (B - \nabla B) =$

$(A \bowtie B) - (\nabla A \bowtie B + A \bowtie \nabla B - \nabla A \bowtie \nabla B)$

Existing Approach 2: Provenance Circuits

Conceptually great, but impractical, inefficient evaluation logic

T, F, T → ∧, ∧ → ∨ → o_1

Our Approach

Counterfactual Deletions

Input bits

Generated Code
- ⋈ Join Del Op
- γ GroupBy Del Op

Query Results

id	val
a1	1	...
a2	2	...

$A \bowtie B$

id	val
b1	1	...
b2	1	...
b3	2	...

id	val
1	...
1	...
2	...

$a1 \wedge b1$
$a1 \wedge b2$
$a2 \wedge b3$

Join Provenance

A	B	out
a1	b1	1
a1	b2	2
a2	b3	3

Our system uses **query operator provenance information** to generate evaluation code which quickly computes deletion results.

1 Deletions as Bit Vectors

Table

Row #	1	2	3
exists?	1	0	0

2 Parallelization

Table → Thread 1, Thread 2, Thread 3 → Join Exists?

3 Input Pruning

Table → Query Res

Experiments & Results

TPC-H at SF 1, randomized independent deletions

Small Deletions (0.05%) — >20x

Medium Deletions (10%) — >600x

Latency in ms (log) — lower is better

Legend: IVM, Ours

Query: 5, 9, 12

Takeaway: Our performance is invariant to deletion size and consistently faster than IVM systems.

Performance in Perspective

Avg Throughput: 810,000 deletions/s

Speedup (1 – 20)

Legend: Vectorization, 8 Threads, Both

Capable of handling over **1 million deletions a second** on analytical queries over multiple tables

Beats IVM and existing provenance systems by **600x** and **10,000x**