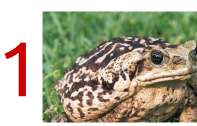
Top-k Query Processing in Probabilistic Databases with Non-Materialized Views

Maximilian Dylla
Iris Miliaraki
Martin Theobald

Tuple-Independent Probabilistic Database

Encounter		p
Kangaroo	Land	0.8
Shark	Water	0.25
Crocodile	Land	0.1
Crocodile	Water	0.2
Cane Toad	Land	0.95

Fatality	p
Crocodile	0.7
Kangaroo	0.05
Shark	0.2
Cane Toad	0.01



Queries

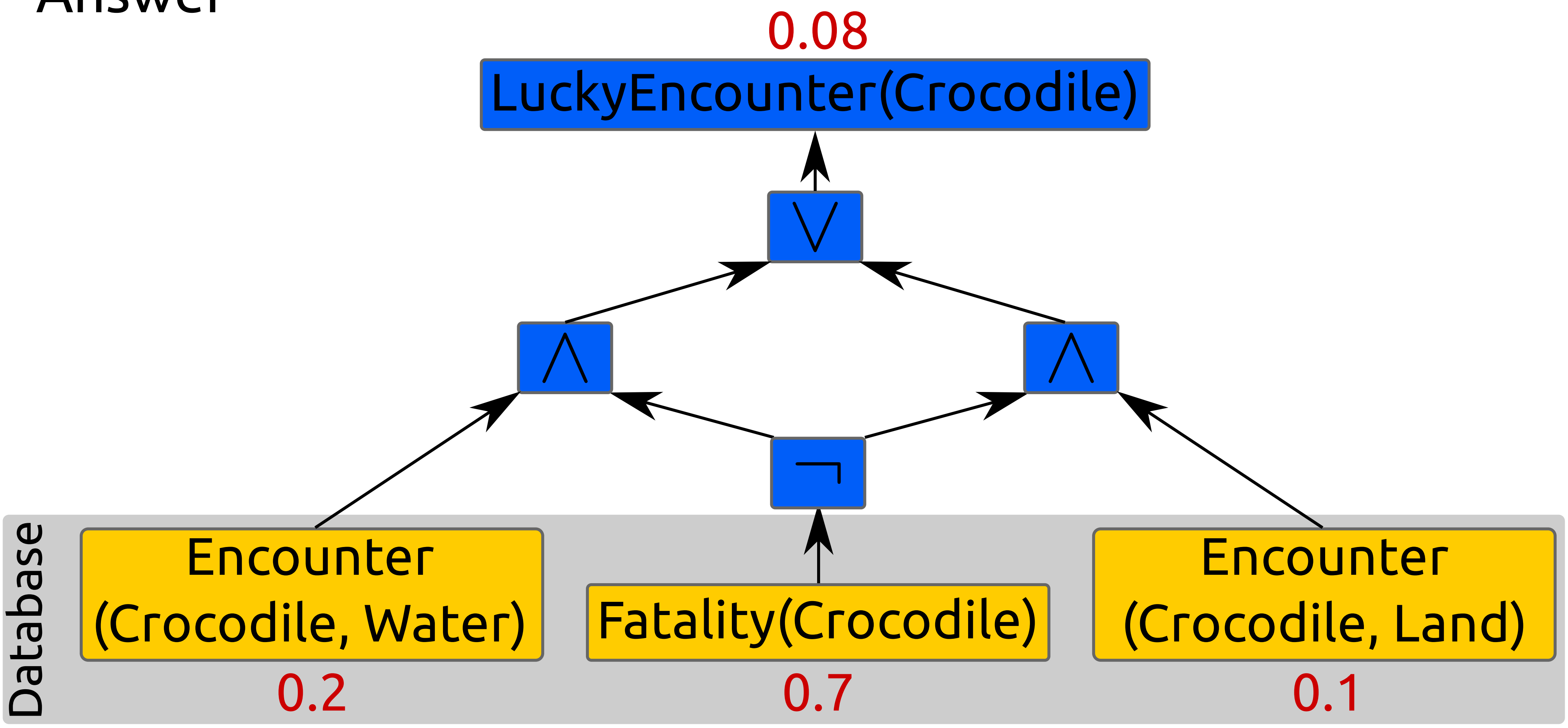
Query

$$LuckyEncounter(?A)$$

Deduction rule

$$LuckyEncounter(?A) \leftarrow \exists ?L\ \ Encounter(?A, ?L) \wedge \neg Fatality(?A)$$

Answer

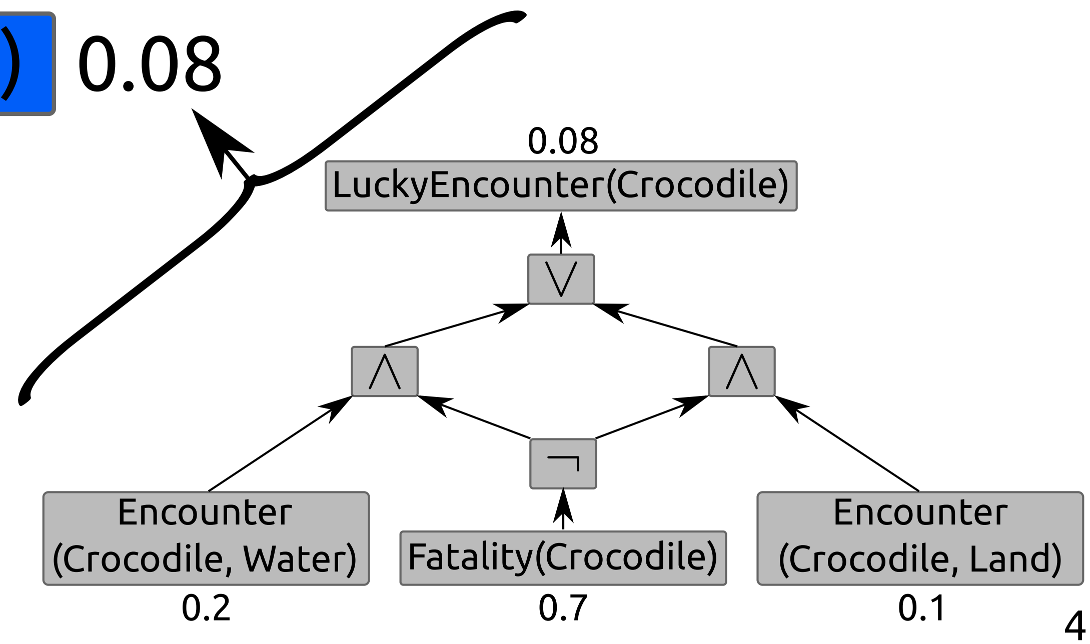


Top-k Answers

Query $LuckyEncounter(?A), \color{red}{k = 2}$

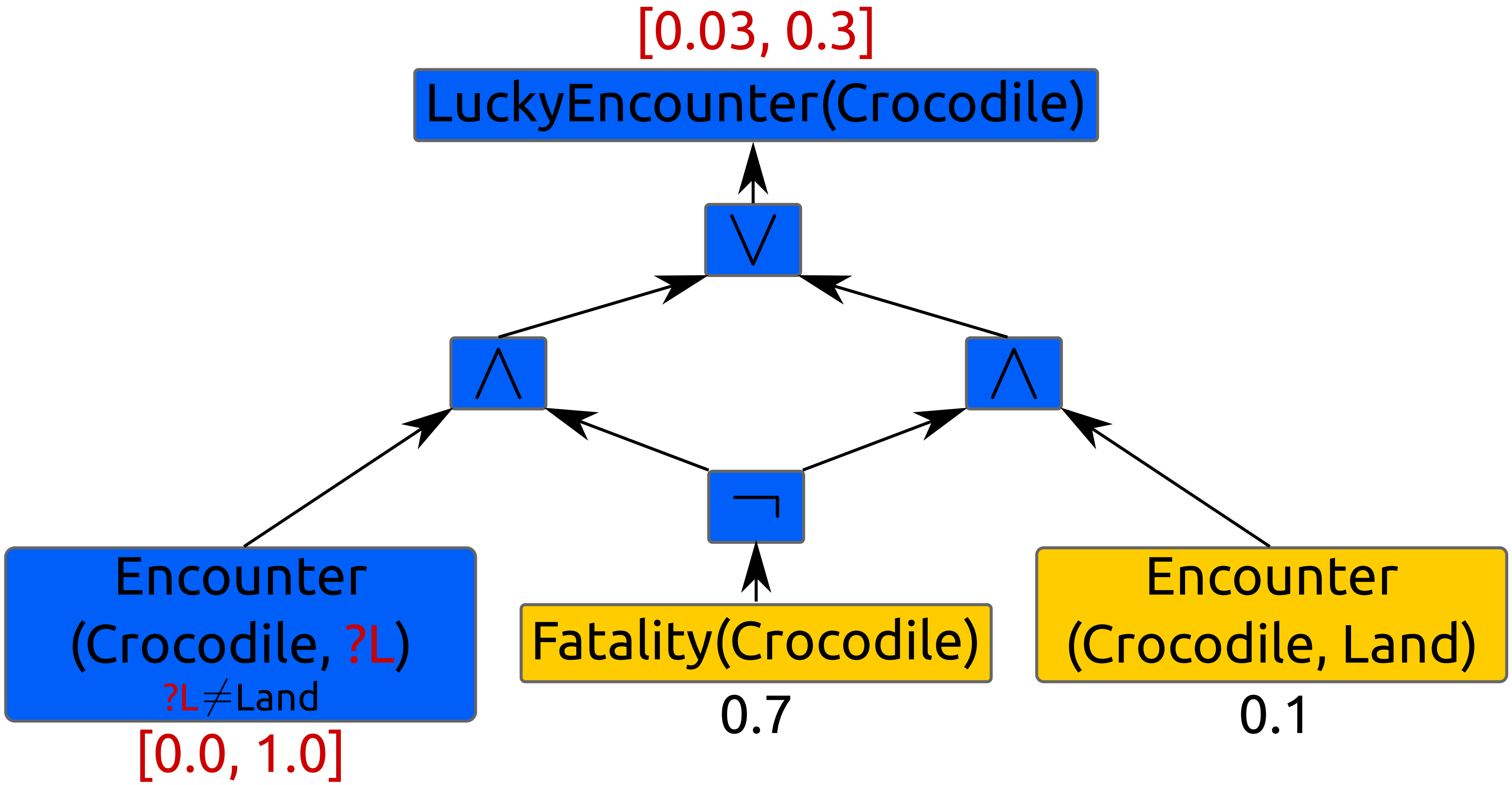
Top-2

Answers	p
LuckyEncounter(Cane Toad)	0.94
LuckyEncounter(Kangaroo)	0.76
	$\geq$
LuckyEncounter(Shark)	0.2
LuckyEncounter(Crocodile)	0.08



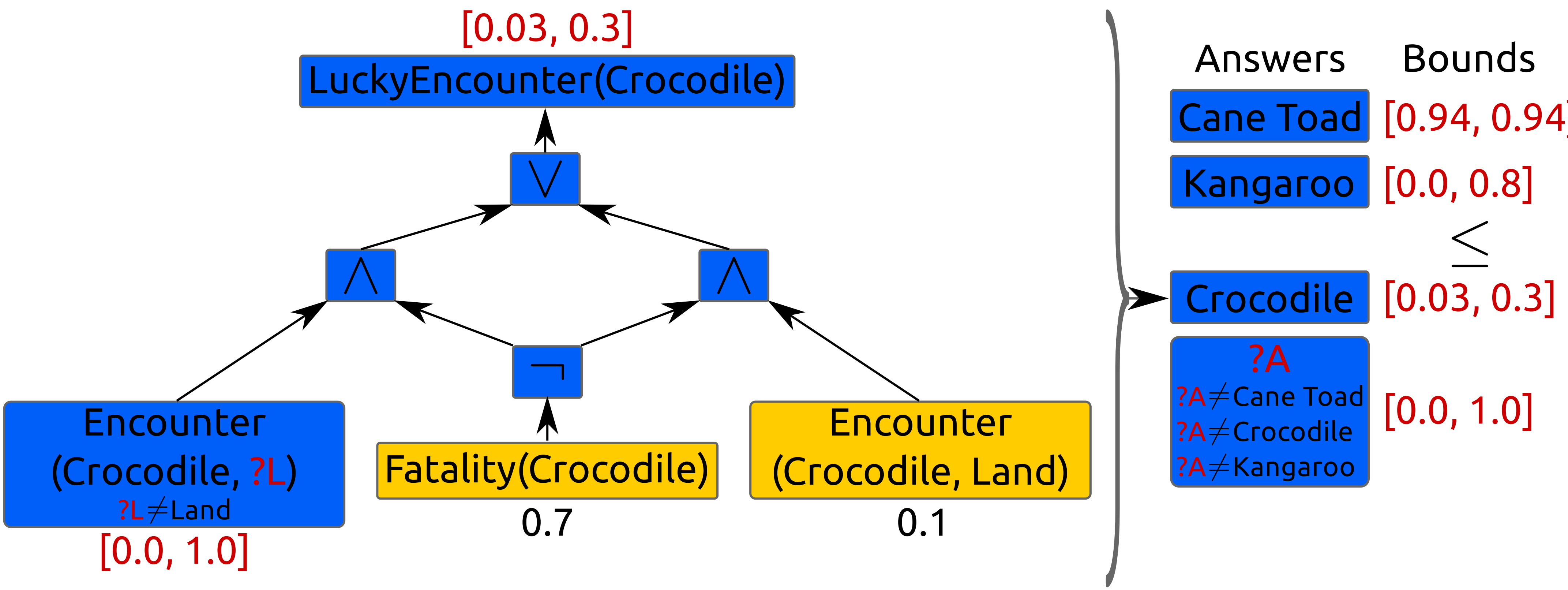
First-Order Lineage

$$LuckyEncounter(?A) \leftarrow \exists ?L\ Encounter(?A, ?L) \land \neg Fatality(?A)$$



First-Order Lineage and Top-k

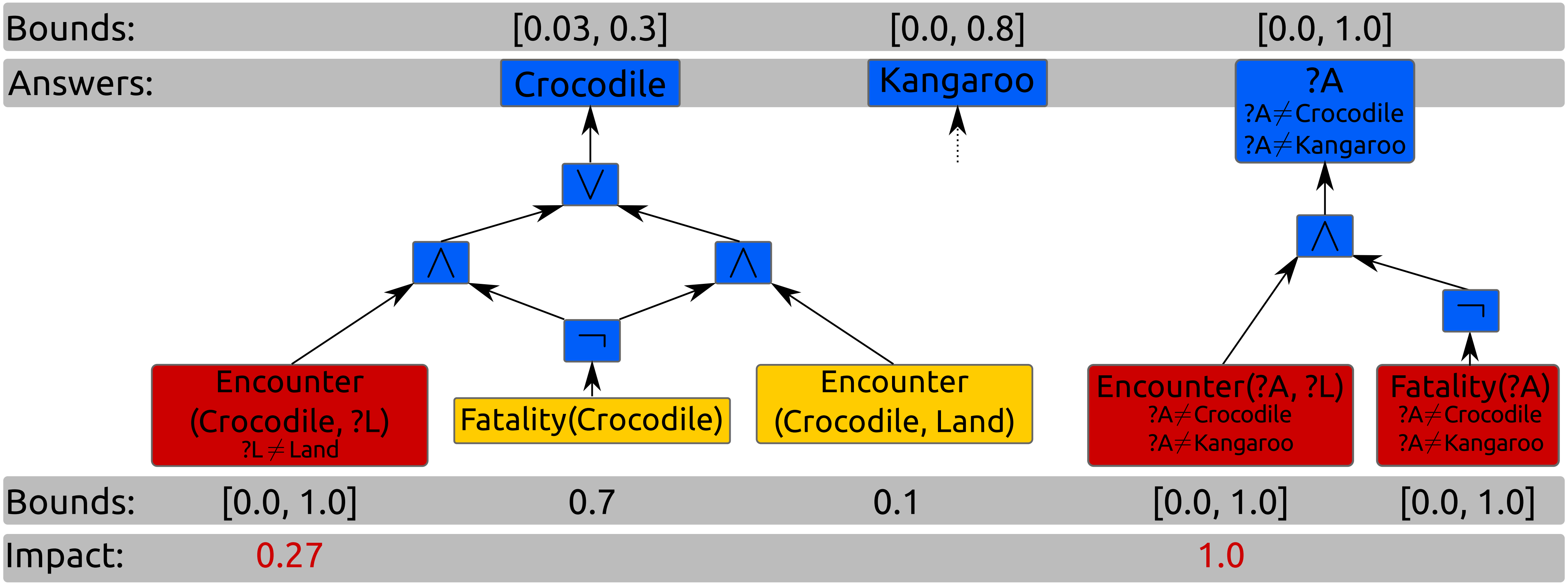
$$LuckyEncounter(?A) \leftarrow \exists ?L \;\; Encounter(?A, ?L) \wedge \neg Fatality(?A)$$



Scheduling

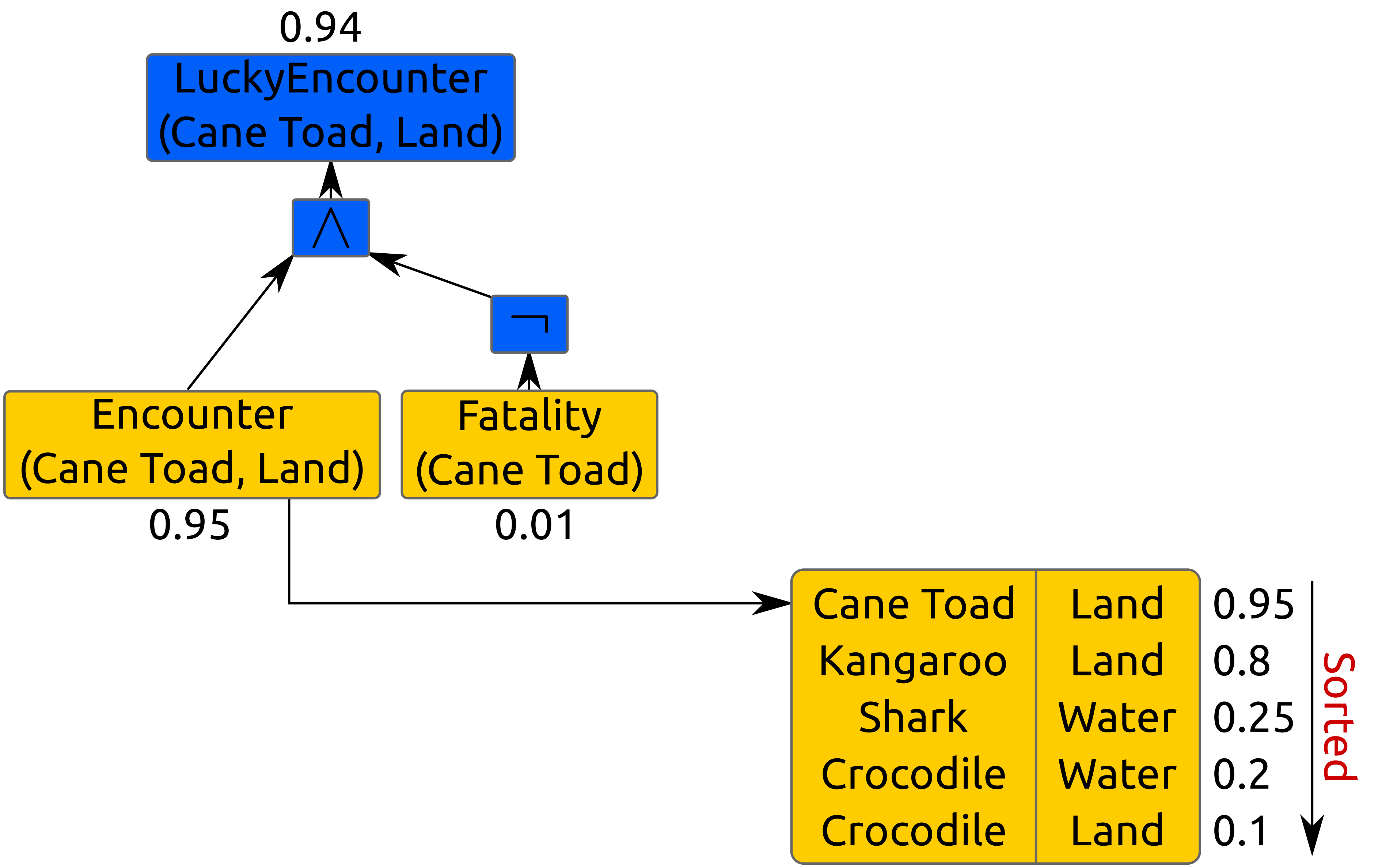
Impact on answer's bounds:

$$\frac{d}{dg} = P\left(\phi_{[g \to true]}\right) - P\left(\phi_{[g \to false]}\right)$$



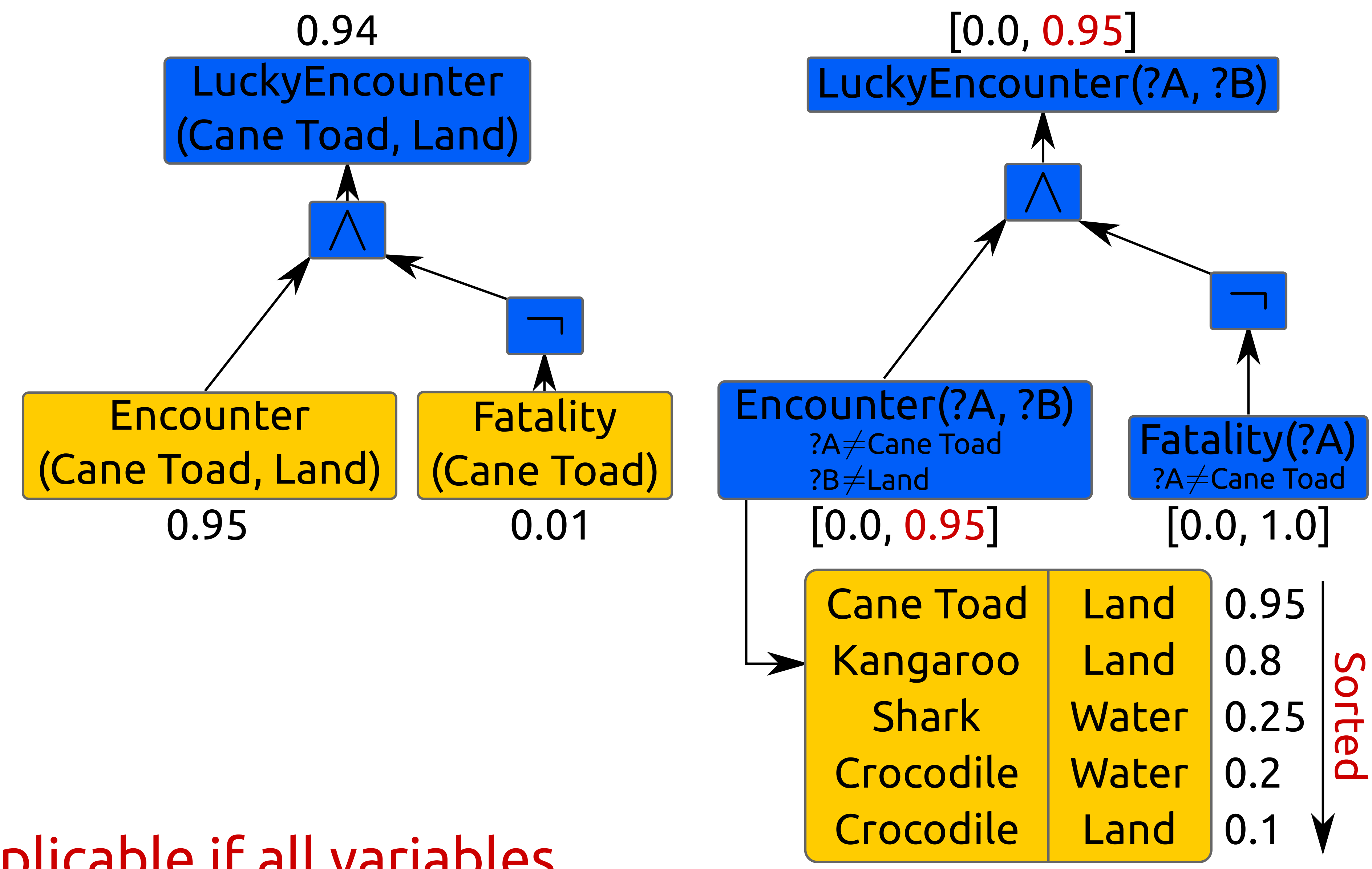
Sorted Input Lists

Rule: $LuckyEncounter(?A, ?B) \leftarrow Encounter(?A, ?B) \wedge \neg Fatality(?A)$



Sorted Input Lists

Rule: $LuckyEncounter(?A, ?B) \leftarrow Encounter(?A, ?B) \wedge \neg Fatality(?A)$



Applicable if all variables are query variables!

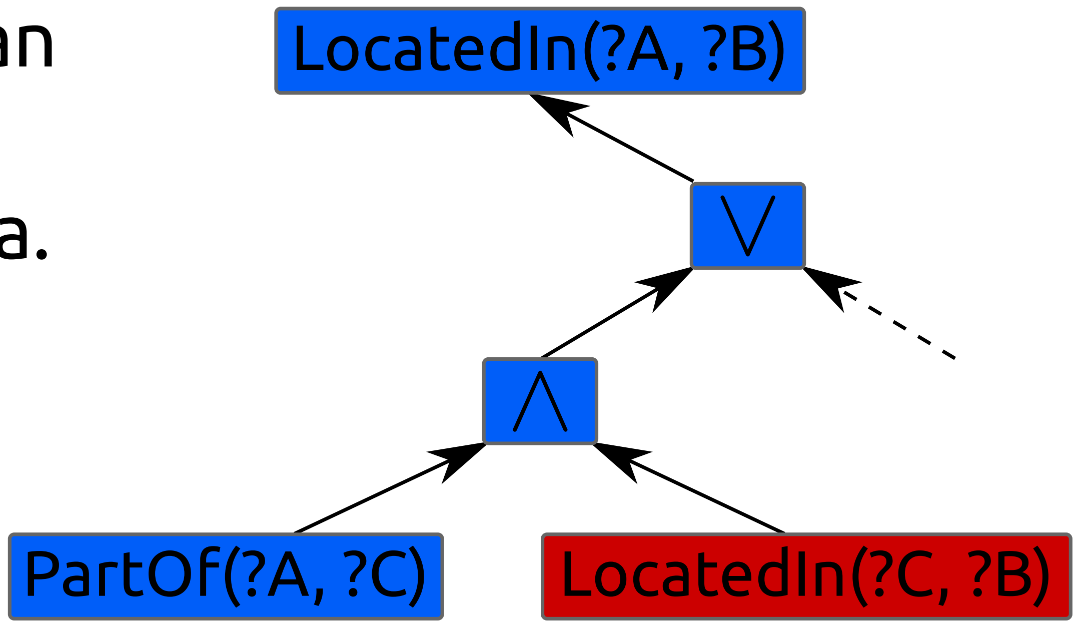
Recursion

Rule: $LocatedIn(?A, ?B) \leftarrow \exists ?C \;\; PartOf(?A, ?C) \wedge LocatedIn(?C, ?B)$

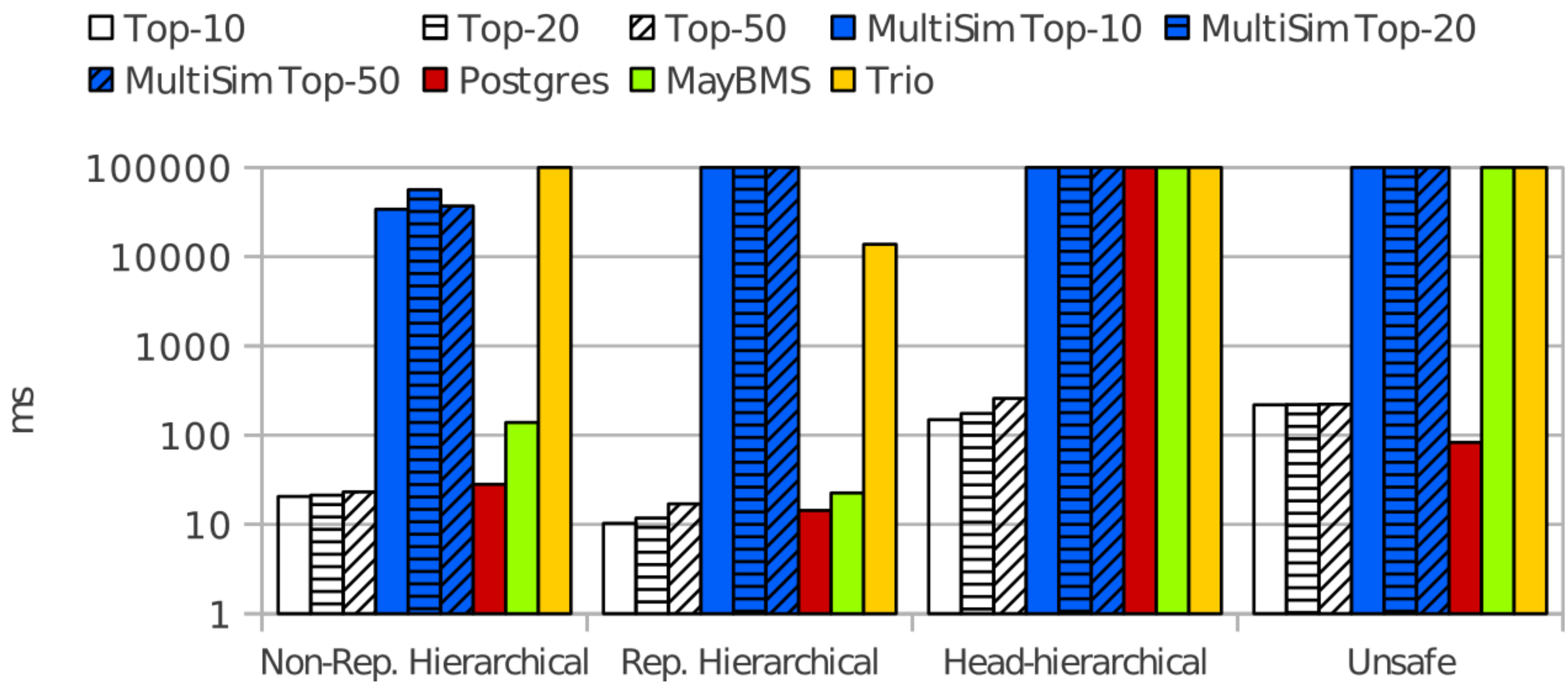
Block cycles during grounding

Theorem:

Expanding a cycle more than once does not alter the validity of a lineage formula.



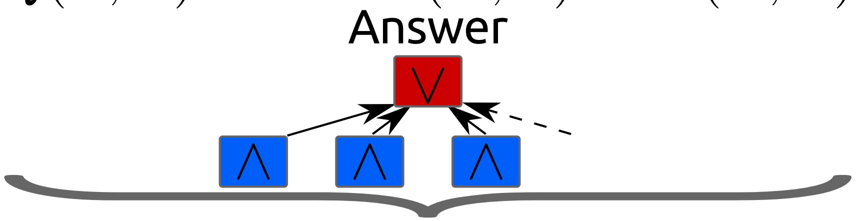
Experiments: Query Classes

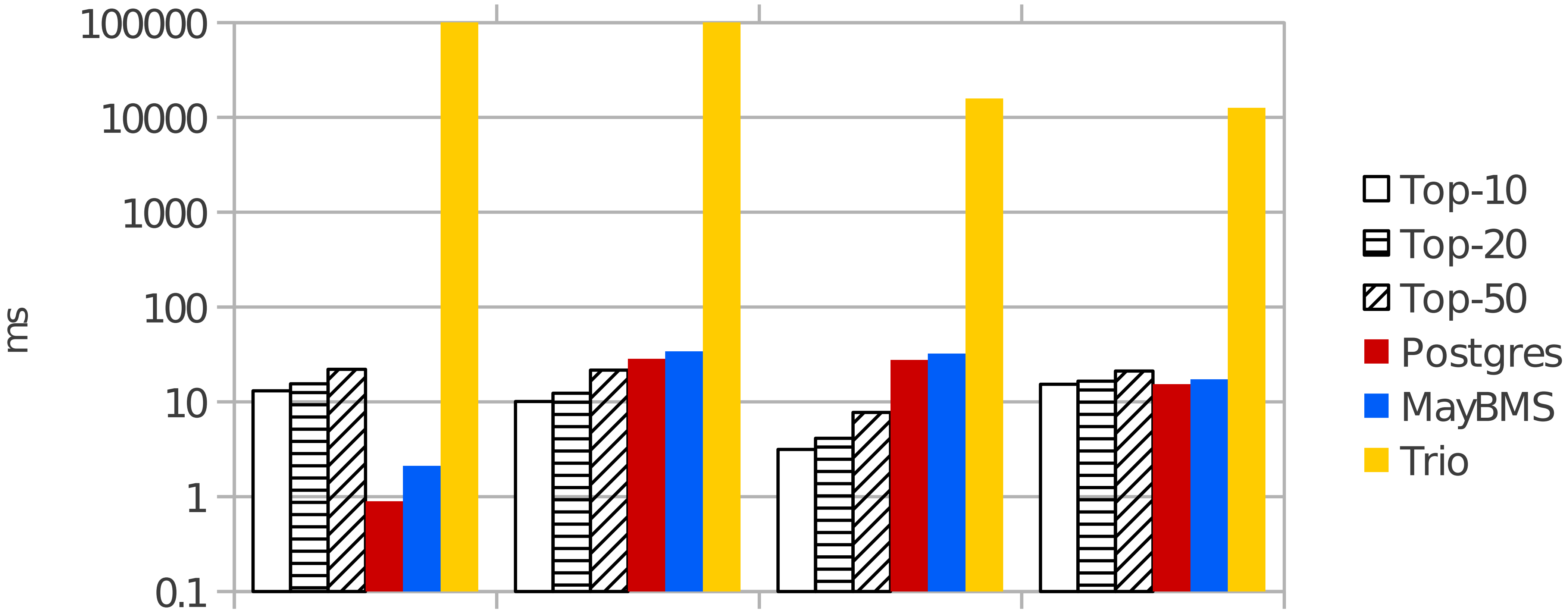


Data: Imdb, 26M tuples, uniformly sampled confidences.
Queries: Each query pattern instantiated by 1000 constants.

Experiments: Performance Factors

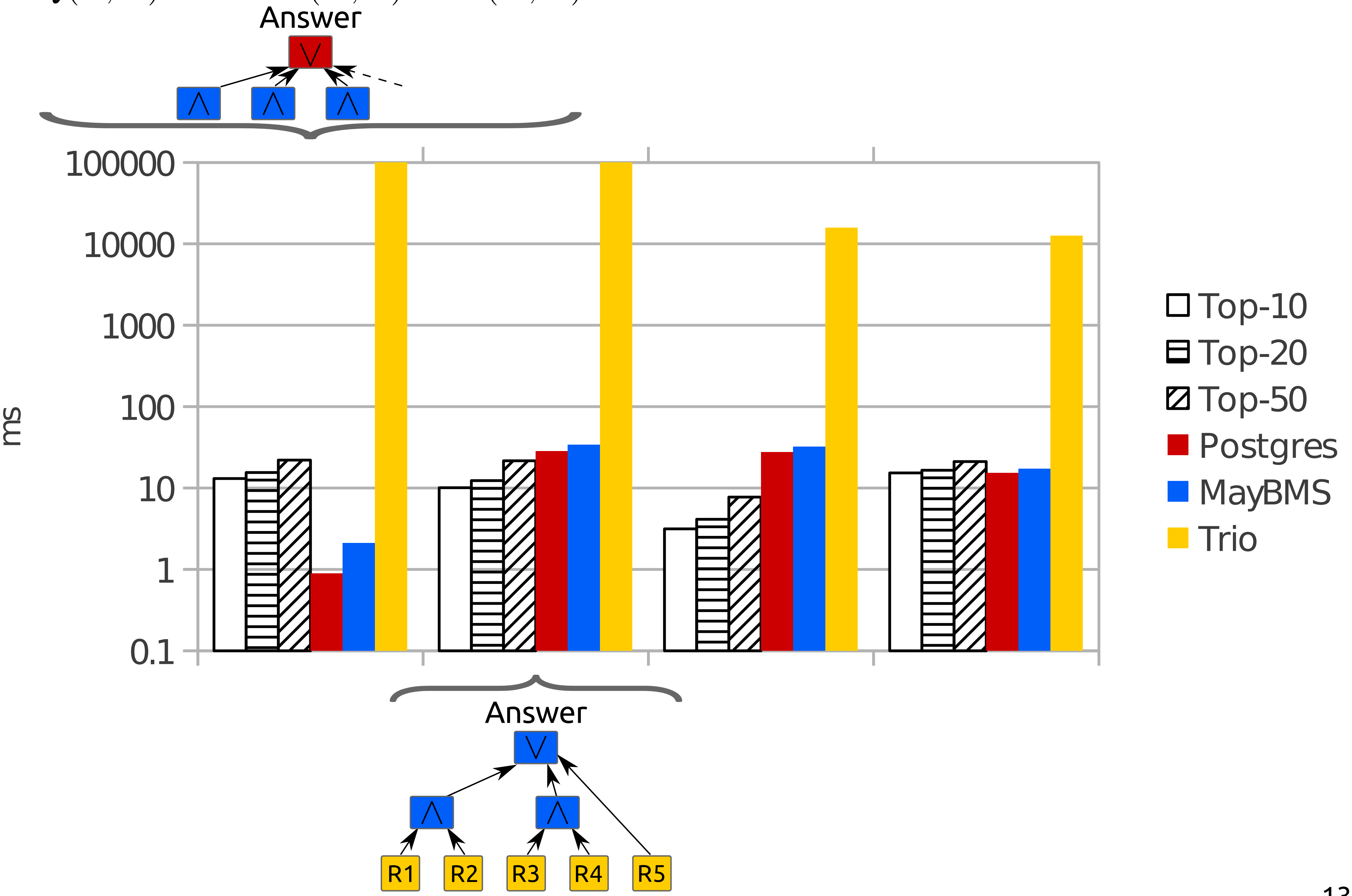
$Q(A,B) \leftarrow \exists X\, R1(X,A) \wedge R2(X,B)$





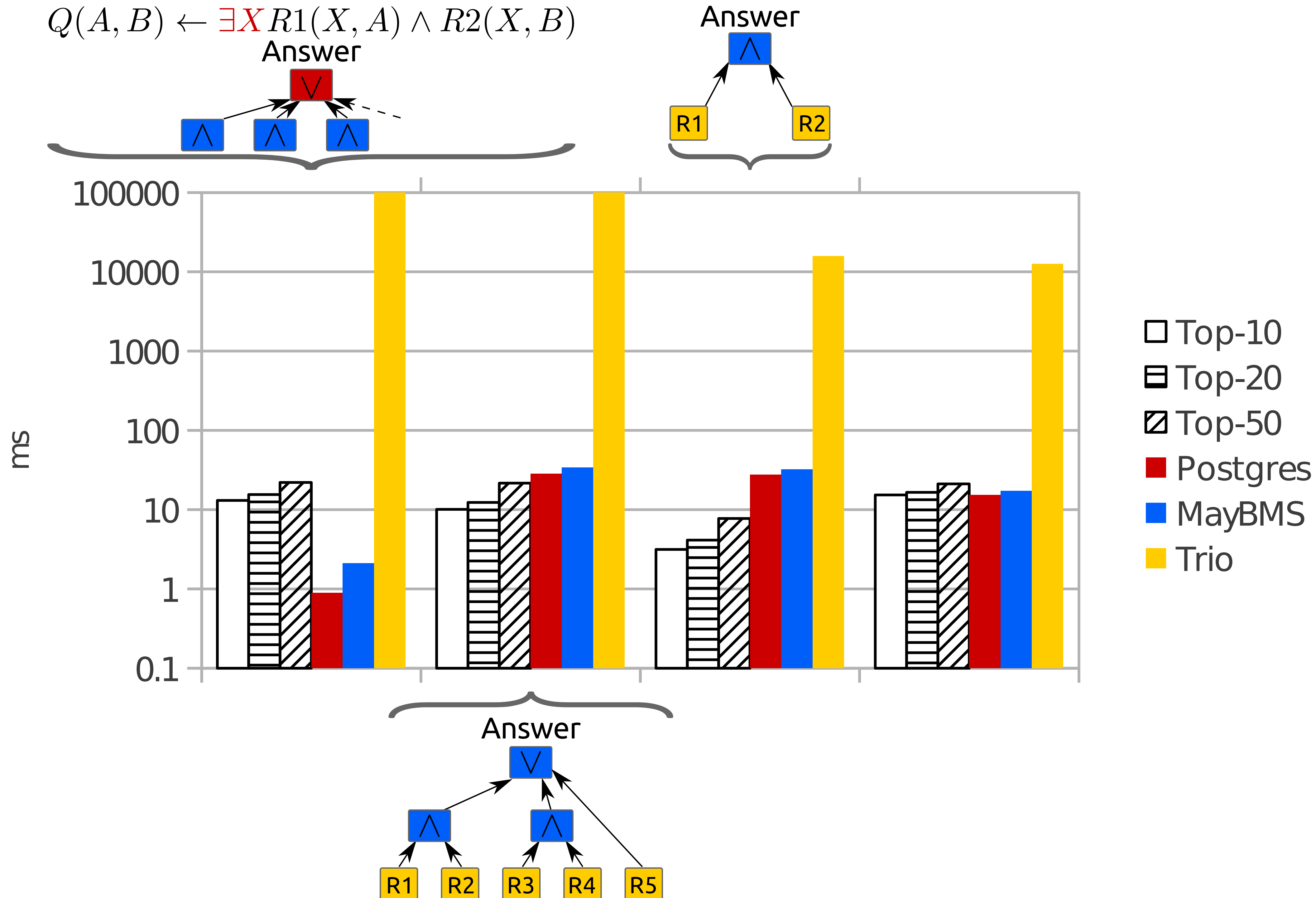
Experiments: Performance Factors

$$Q(A,B) \leftarrow \exists X\, R1(X,A) \wedge R2(X,B)$$



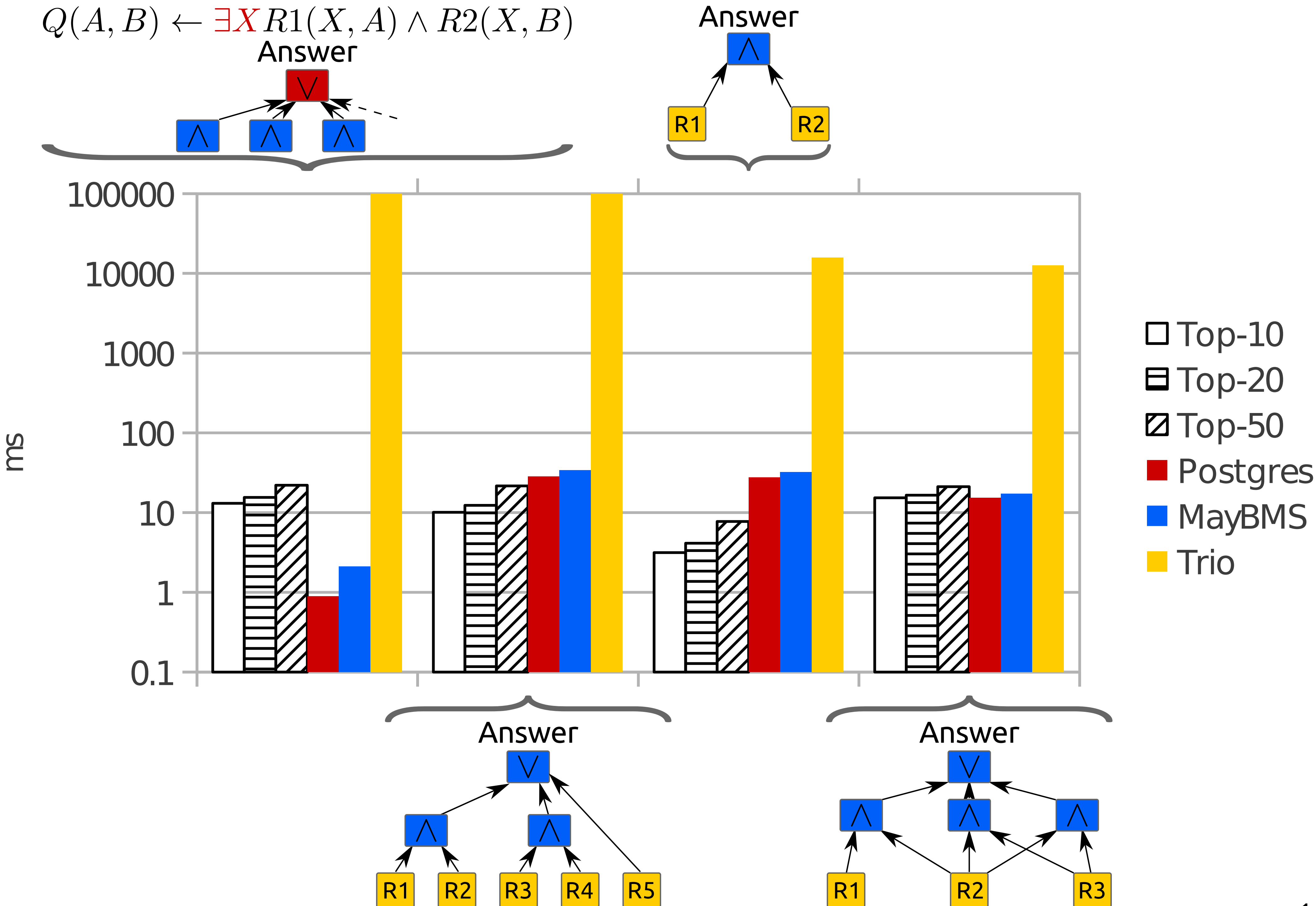
Experiments: Performance Factors

$Q(A,B) \leftarrow \exists X\, R1(X,A) \wedge R2(X,B)$



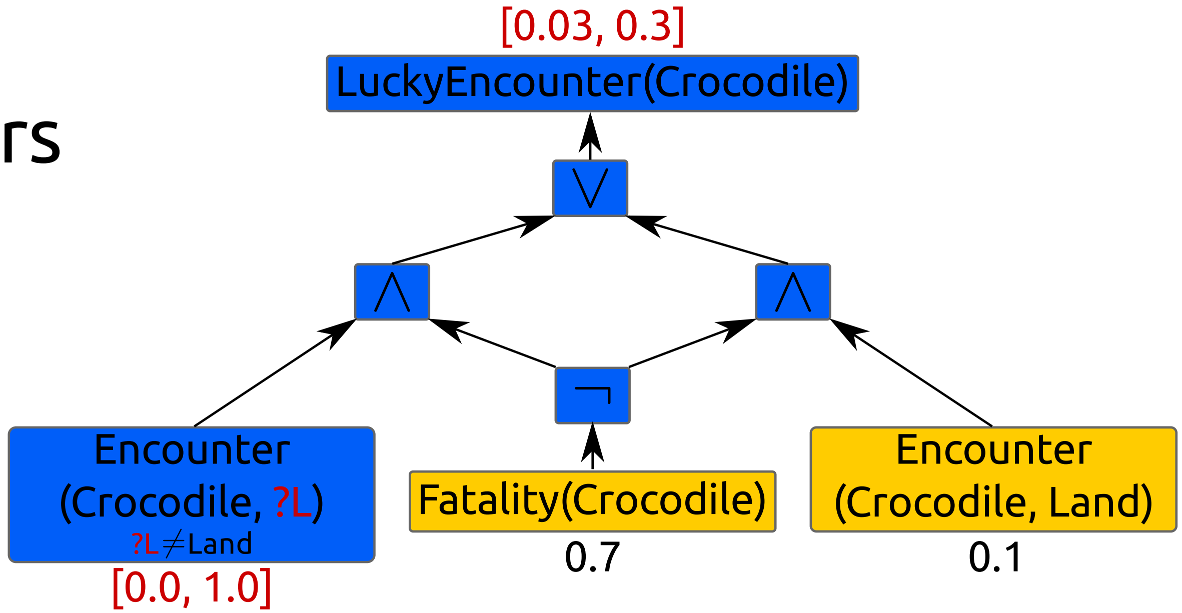
Experiments: Performance Factors

$$Q(A,B) \leftarrow \exists X\, R1(X,A) \wedge R2(X,B)$$



Summary

First-order lineage representing sets of answers + bounds on probabilities.



Integration of data and confidence computations.

Support for all select-project-join queries.