Pre-Aggregation with Probability Distributions

Igor Timko (Free University of Bozen-Bolzano, Italy)
Curtis E. Dyreson (Washington State University, WA, USA)
Torben Bach Pedersen (Aalborg University, Denmark)

DOLAP 2006

Introduction

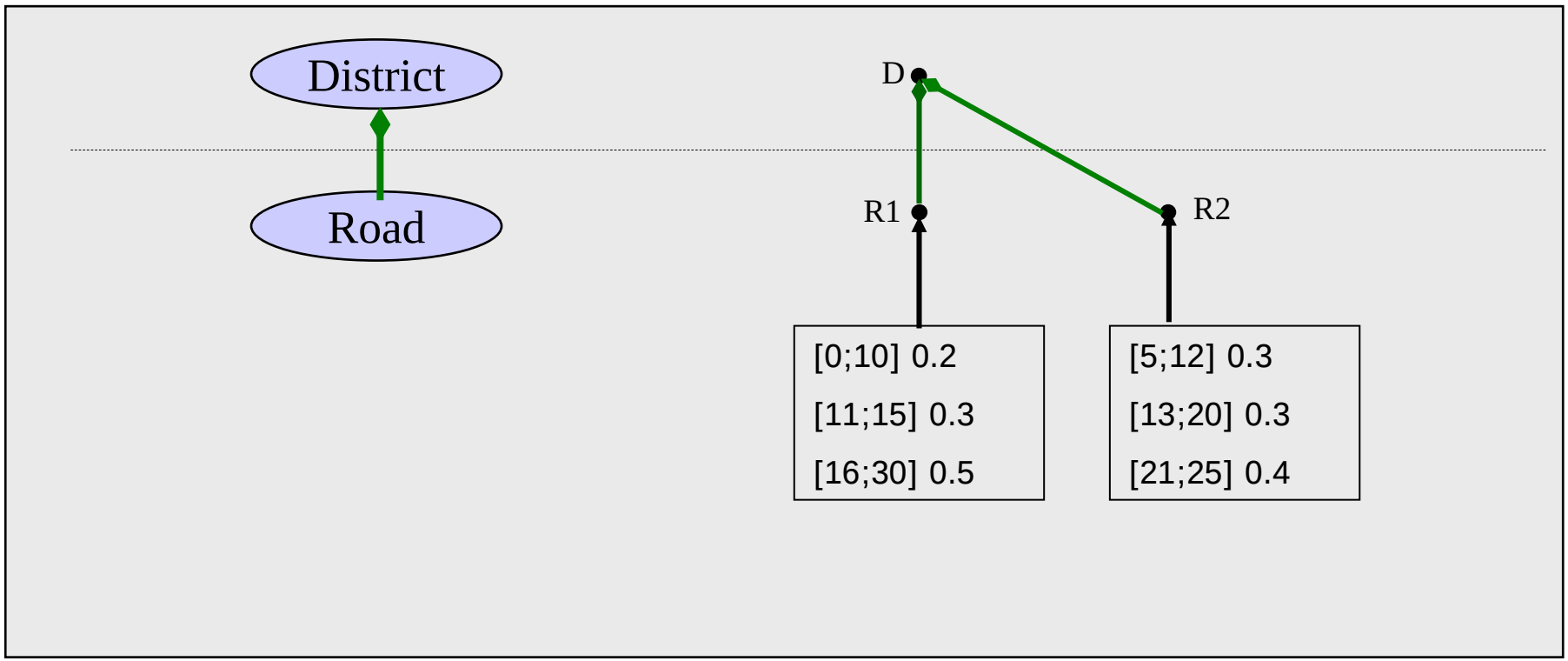
- LBS functionality (e.g., route finding) requires complex data analysis involving aggregation of probability distributions:
  - E.g., "Given probability distributions of COUNT of cars per city road in 5 minutes from now, what will the probability distribution of COUNT of cars be, per city district?"
- OLAP and DW enable complex analysis:
  - Multidimensional data model – high expressive power
  - Pre-computation of aggregate values (pre-aggregation) - fast aggregation
- We extend OLAP/DW to support aggregation and pre-aggregation of probability distributions:
  - Generalized measure - probability distribution
  - Approximate aggregation and pre-aggregation for probability distributions
  - Processing of queries over approximate probability distributions

Talk Outline

- Introduction
- Probability Distributions As Measures
- Aproximate Aggregation
- Pre-Aggregation
- Queries over Approximate Probability Distributions
- Conclusions and Future Work
- Related Work

Probability Distributions as Measures

- New type of aggregate values:
  - Aggregate value as probability distribution



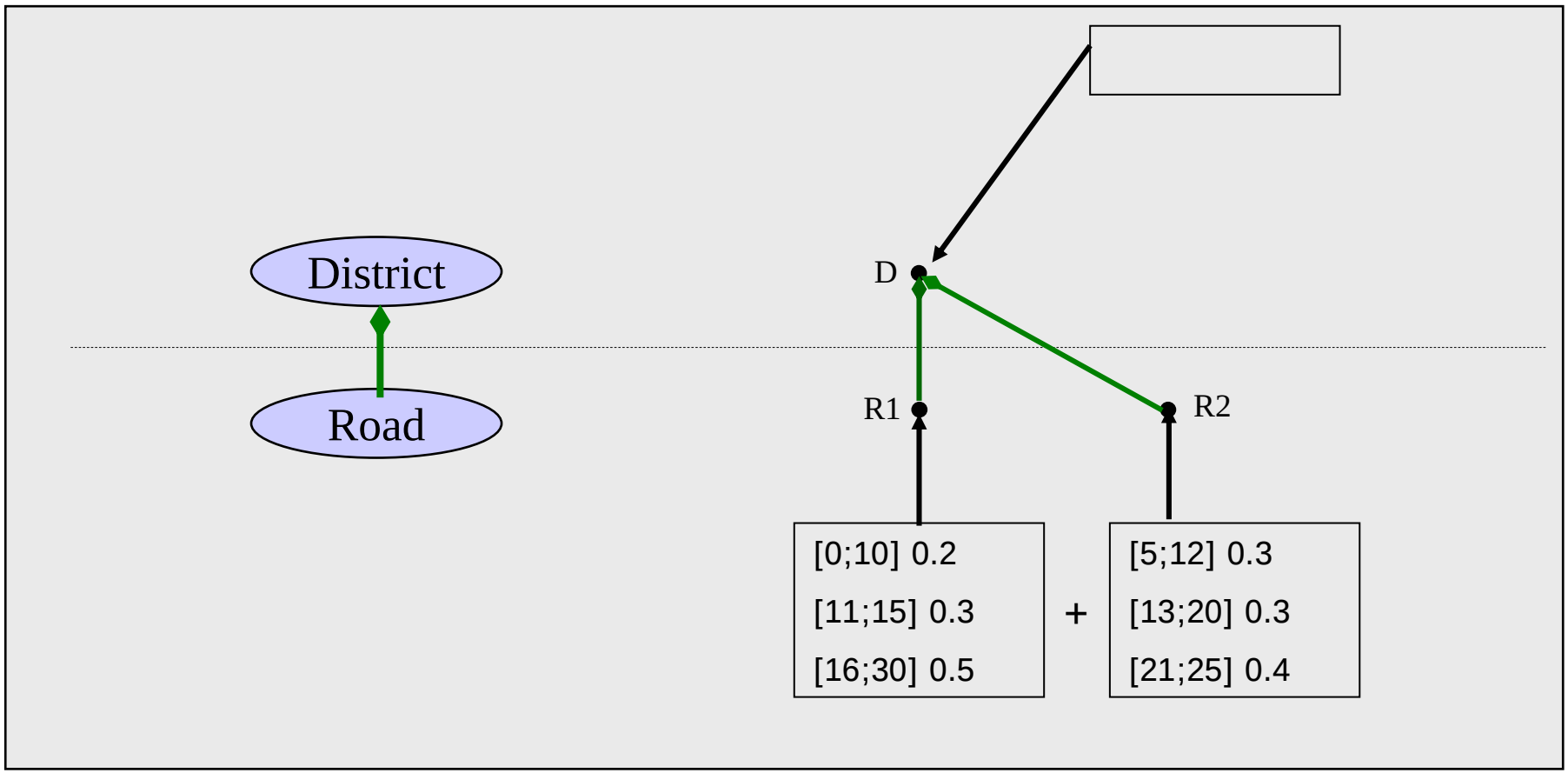
Talk Outline

- Introduction
- Probability Distributions As Measures
- Aproximate Aggregation
- Pre-Aggregation
- Queries over Approximate Probability Distributions
- Conclusions and Future Work
- Related Work

Aggregation

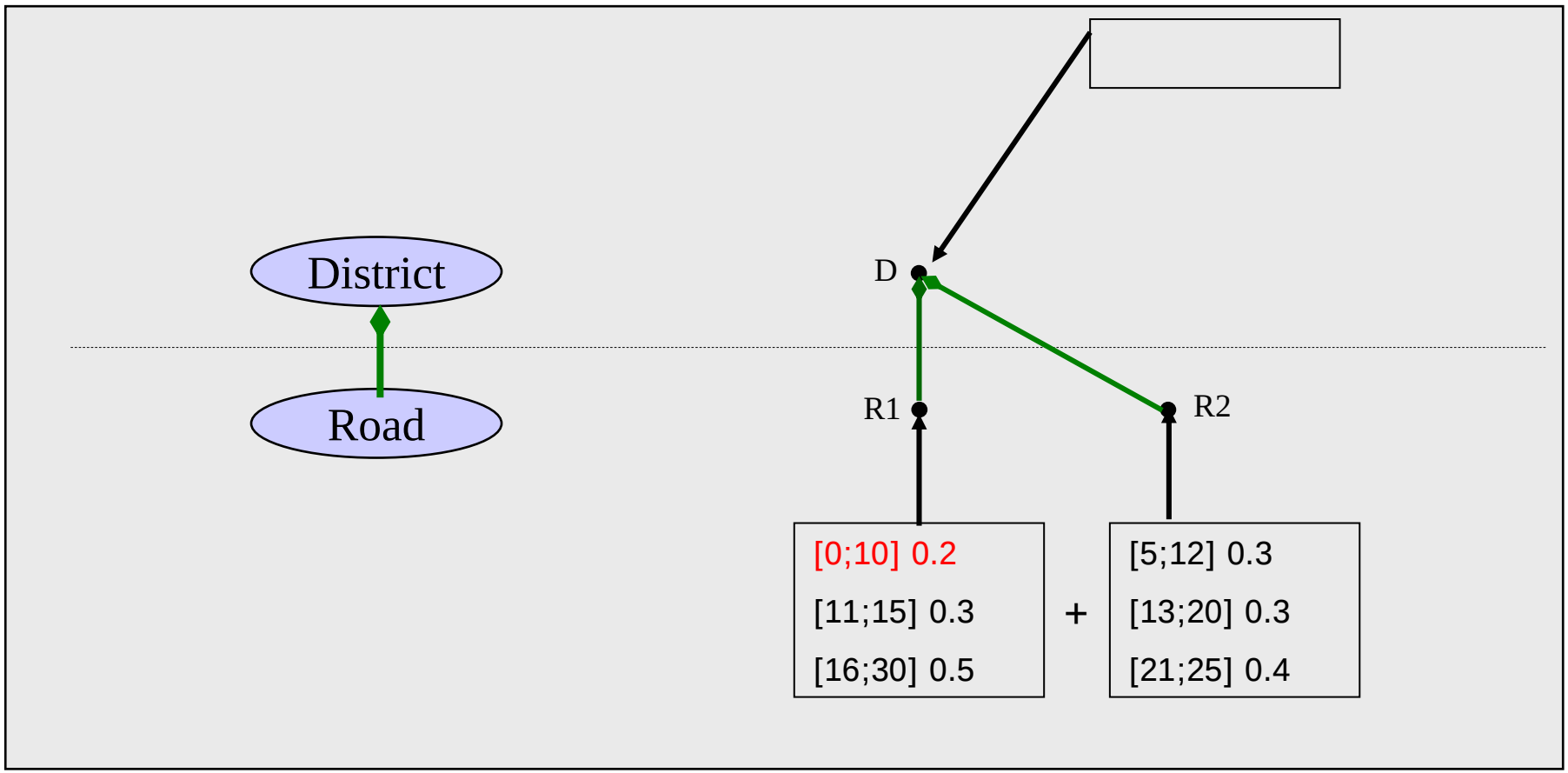
- Aggregation is based on summation of probability distributions:
  - Add each interval from A to each interval from B
  - Based on interval arithmetics:

    ([a;b], p) + ([x;y], q) = ([a+x, b+y], p * q)



Aggregation

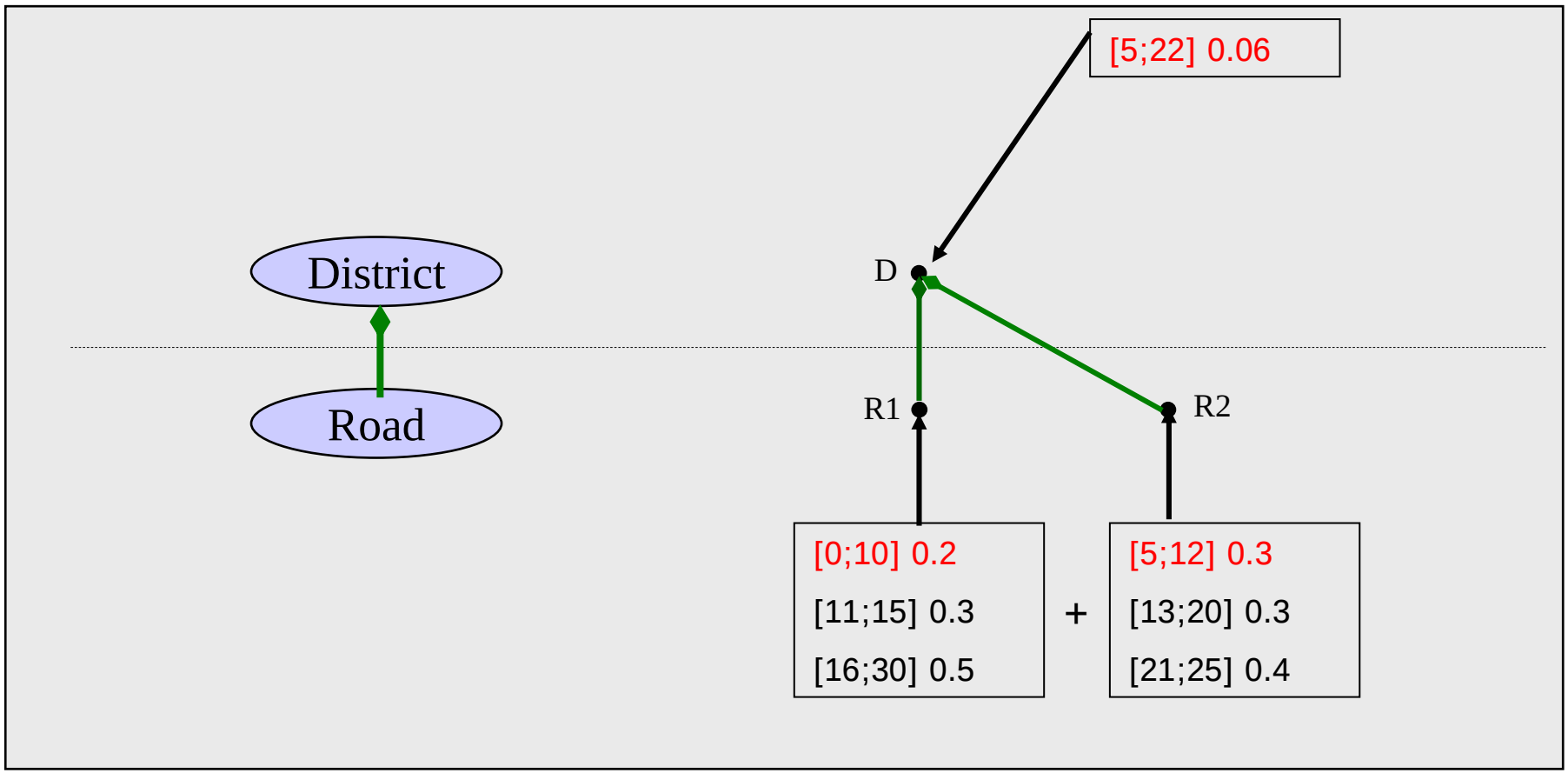
- Aggregation is based on summation of probability distributions:
  - Add each interval from A to each interval from B
  - Based on interval arithmetics:
    ([a;b], p) + ([x;y], q) = ([a+x, b+y], p * q)



Aggregation

- Aggregation is based on summation of probability distributions:
  - Add each interval from A to each interval from B
  - Based on interval arithmetics:

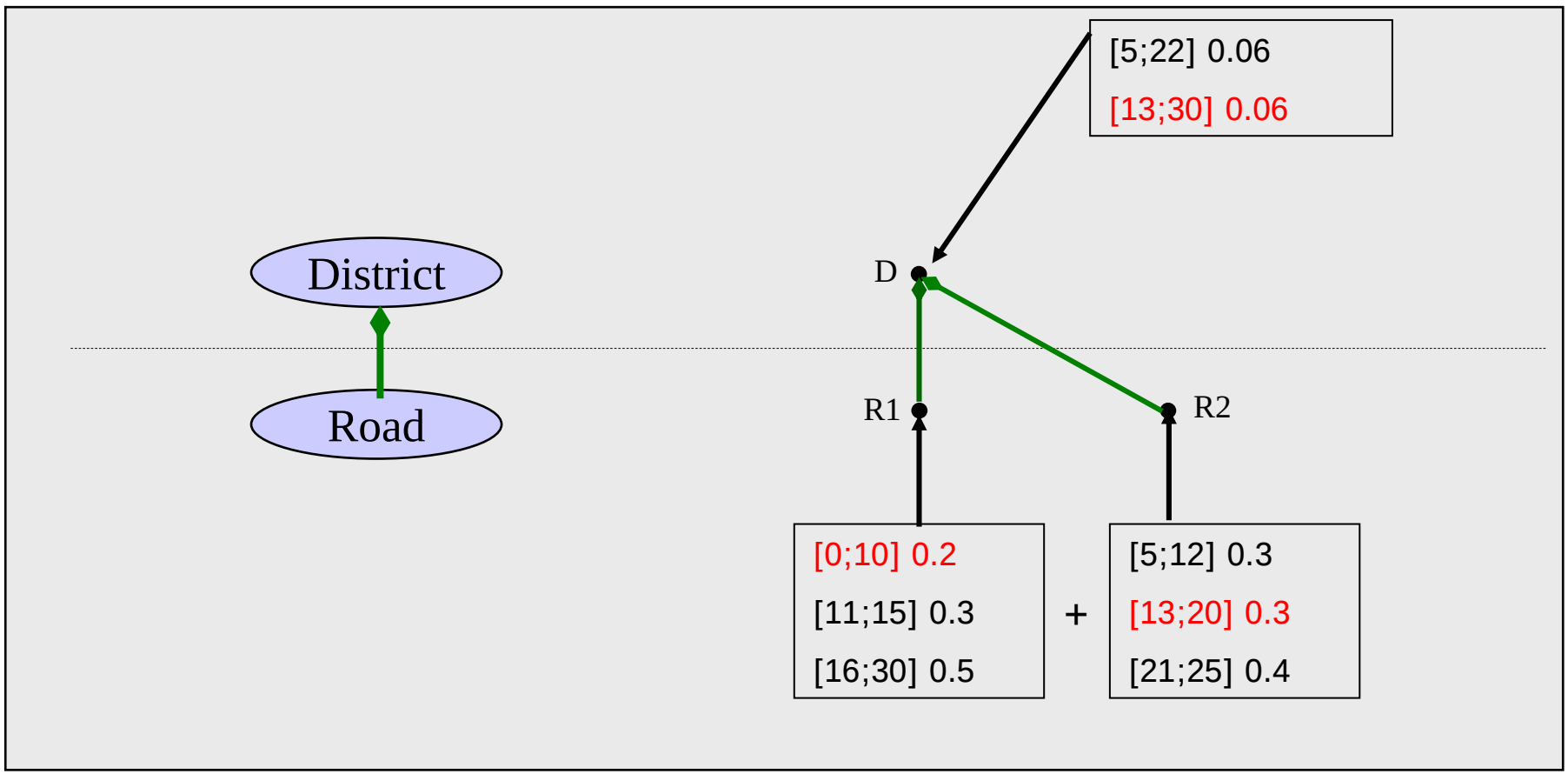
    ([a;b], p) + ([x;y], q) = ([a+x, b+y], p * q)



Aggregation

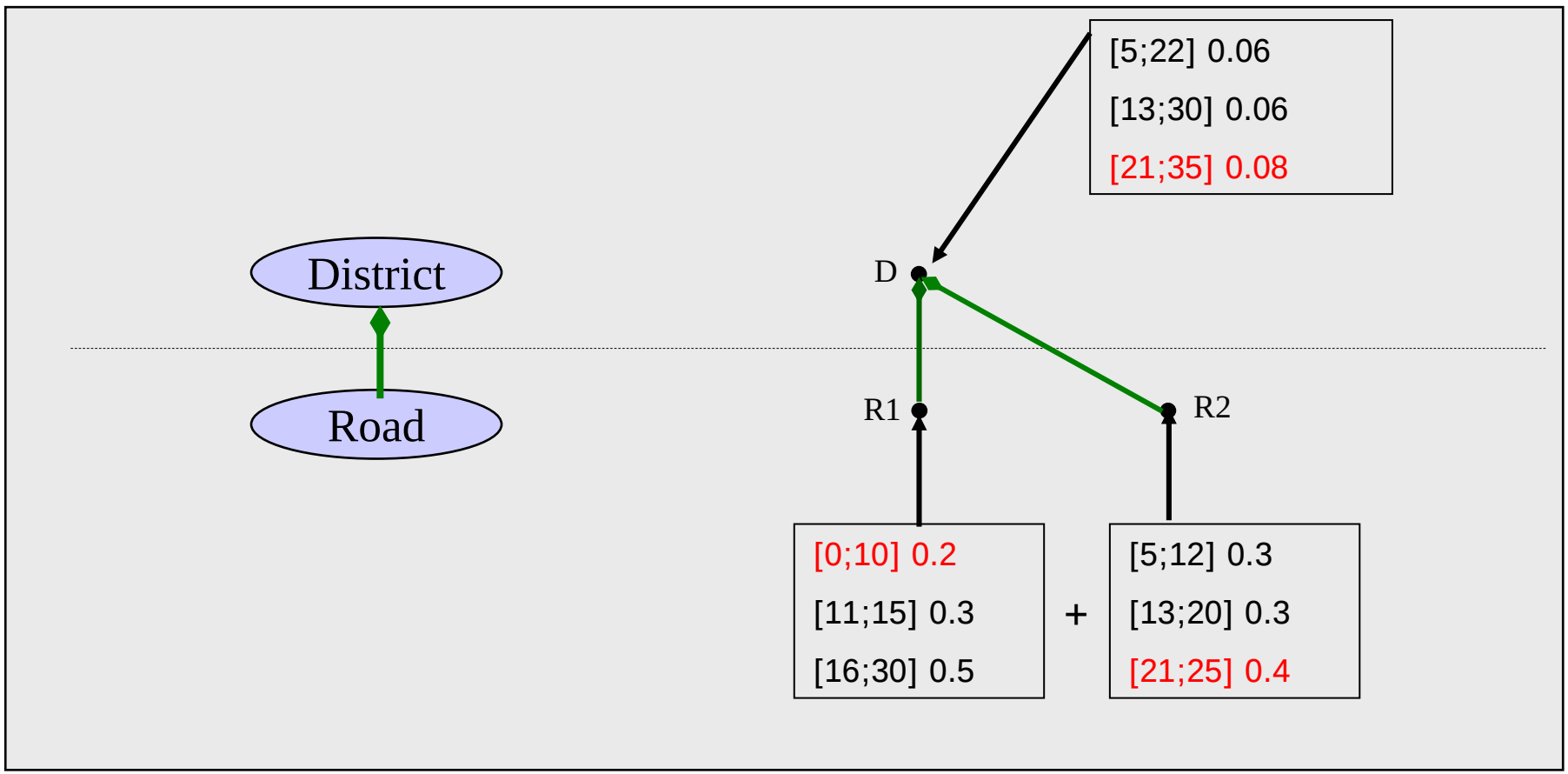
- Aggregation is based on summation of probability distributions:
  - Add each interval from A to each interval from B
  - Based on interval arithmetics:

    ([a;b], p) + ([x;y], q) = ([a+x, b+y], p * q)

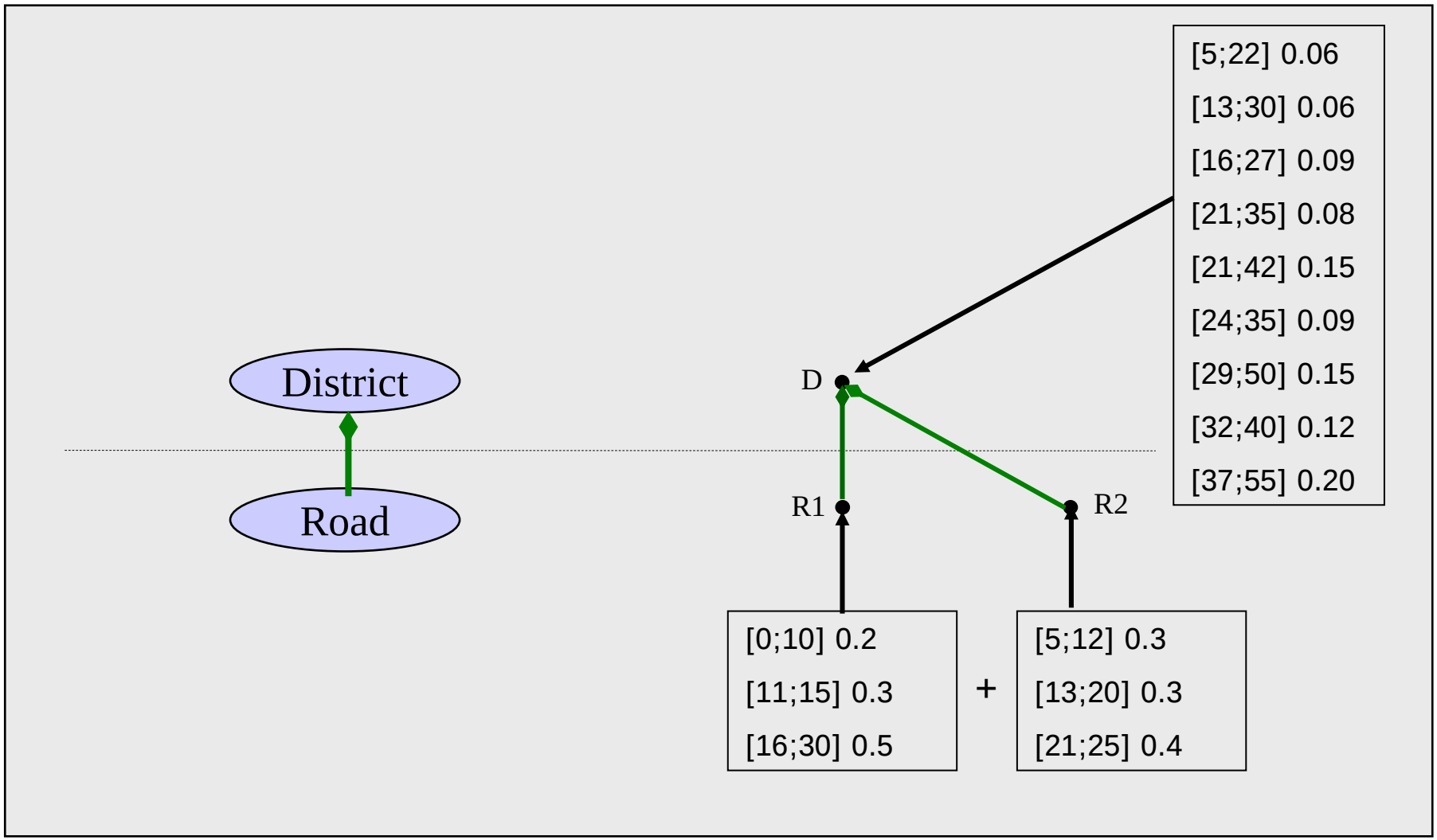


Aggregation

- Aggregation is based on summation of probability distributions:
  - Add each interval from A to each interval from B
  - Based on interval arithmetics:
    ([a;b], p) + ([x;y], q) = ([a+x, b+y], p * q)



Aggregation



Approximate Aggregation

- *Accurate* summation:
  - Sum M aggregate values, each with N intervals:
    - Need $N^2 + N^3 + \ldots + N^M = O(N^M)$ interval summations
    - Result contains N^M intervals
- *Accurate* summation:
  - Exponential time and space complexity
  - Precision of the result is too high
- Need to perform approximate summation!

Approximate Aggregation: Coalescion

- Idea of approximate summation:
  - Control number of intervals by coalescing intervals in each intermediate result
- Coalescion that preserves “shape” of the summation result:
  - Find “good” intervals, i.e., intervals with highest unit probability
  - Group intervals into groups of approx. equal total probabilities
  - Coalesce intervals in each group except good intervals
- Coalescion has linear time complexity

[5;22] 0.06

[13;30] 0.06

[16;27] 0.09

[21;35] 0.08

[21;42] 0.15

[24;35] 0.09

[29;50] 0.15

[32;40] 0.12

[37;55] 0.20

Approximate Aggregation: Coalescion

- Idea of approximate summation:
  - Control number of intervals by coalescing intervals in each intermediate result
- Coalescion that preserves “shape” of the summation result:
  - Find “good” intervals, i.e., intervals with highest unit probability
  - Group intervals into groups of approx. equal total probabilities
  - Coalesce intervals in each group except good intervals
- Coalescion has linear time complexity

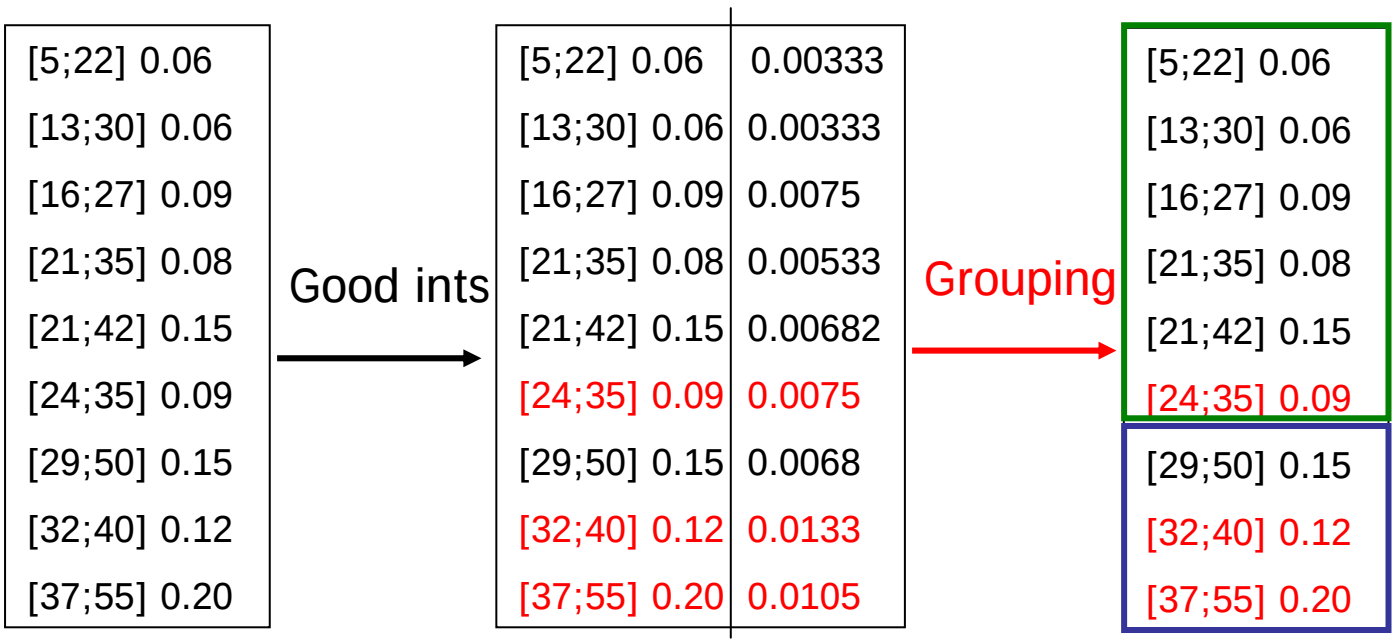
[5;22] 0.06
[13;30] 0.06
[16;27] 0.09
[21;35] 0.08
[21;42] 0.15
[24;35] 0.09
[29;50] 0.15
[32;40] 0.12
[37;55] 0.20

Good ints →

[5;22] 0.06	0.00333
[13;30] 0.06	0.00333
[16;27] 0.09	0.0075
[21;35] 0.08	0.00533
[21;42] 0.15	0.00682
[24;35] 0.09	0.0075
[29;50] 0.15	0.0068
[32;40] 0.12	0.0133
[37;55] 0.20	0.0105

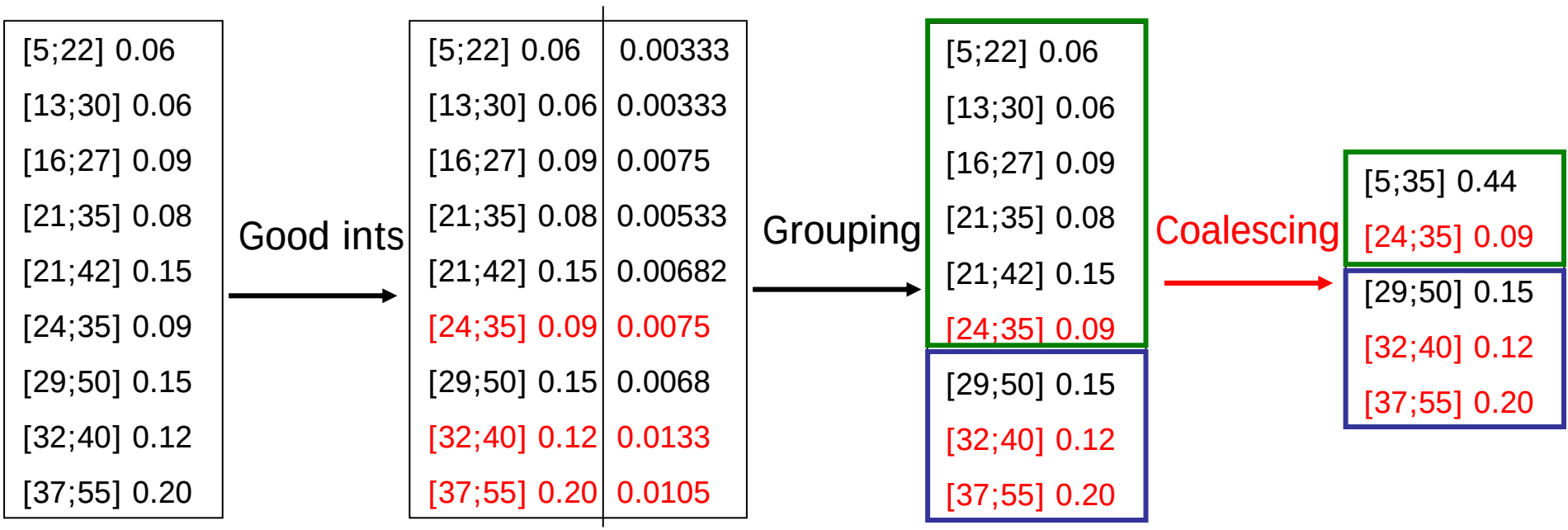
Approximate Aggregation: Coalescion

- Idea of approximate summation:
  - Control number of intervals by coalescing intervals in each intermediate result
- Coalescion that preserves "shape" of the summation result:
  - Find "good" intervals, i.e., intervals with highest unit probability
  - Group intervals into groups of approx. equal total probabilities
  - Coalesce intervals in each group except good intervals
- Coalescion has linear time complexity



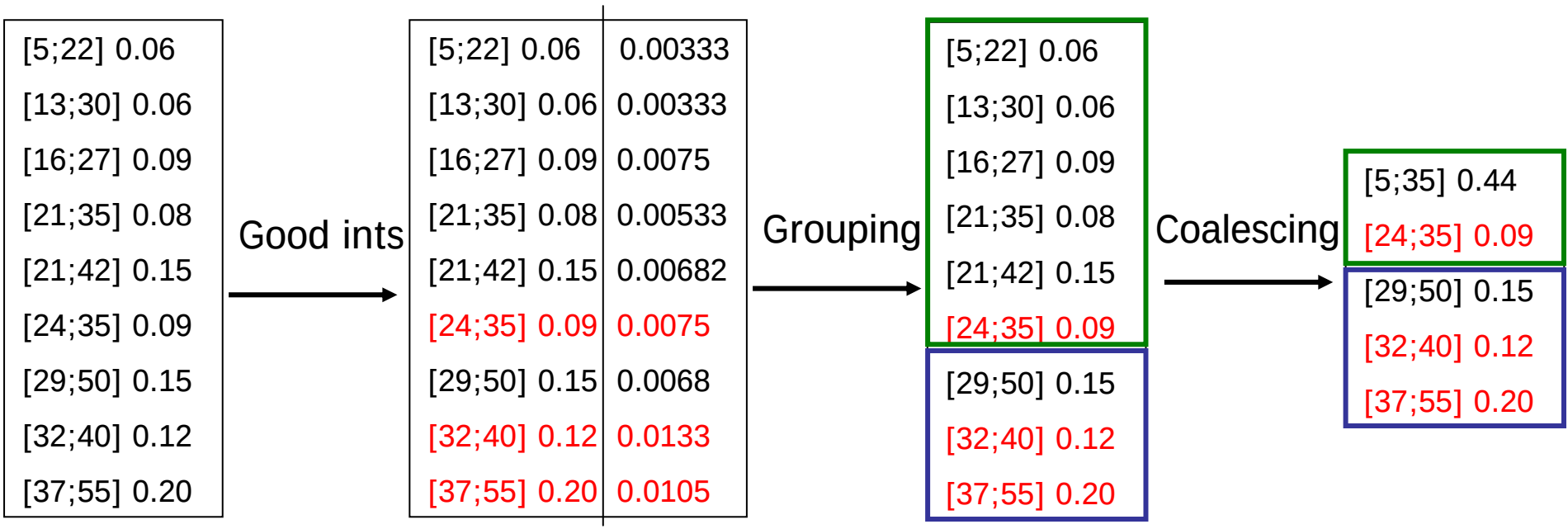
Approximate Aggregation: Coalescion

- Idea of approximate summation:
  - Control number of intervals by coalescing intervals in each intermediate result
- Coalescion that preserves “shape” of the summation result:
  - Find “good” intervals, i.e., intervals with highest unit probability
  - Group intervals into groups of approx. equal total probabilities
  - Coalesce intervals in each group except good intervals
- Coalescion has linear time complexity



Approximate Aggregation: Efficiency and Precision Control

- Maximum length of intermediate results
- Maximum number of “good” intervals
- “Good interval” threshold:
  - The threshold depends on average unit probability

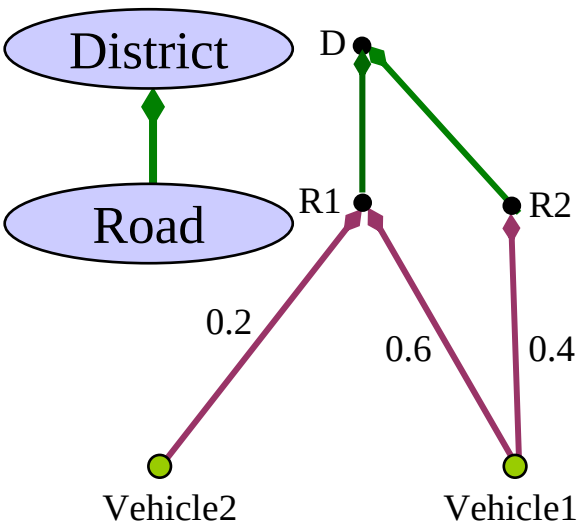


Talk Outline

- Introduction
- Probability Distributions As Measures
- Aproximate Aggregation
- Pre-Aggregation
- Queries over Approximate Probability Distributions
- Conclusions and Future Work
- Related Work

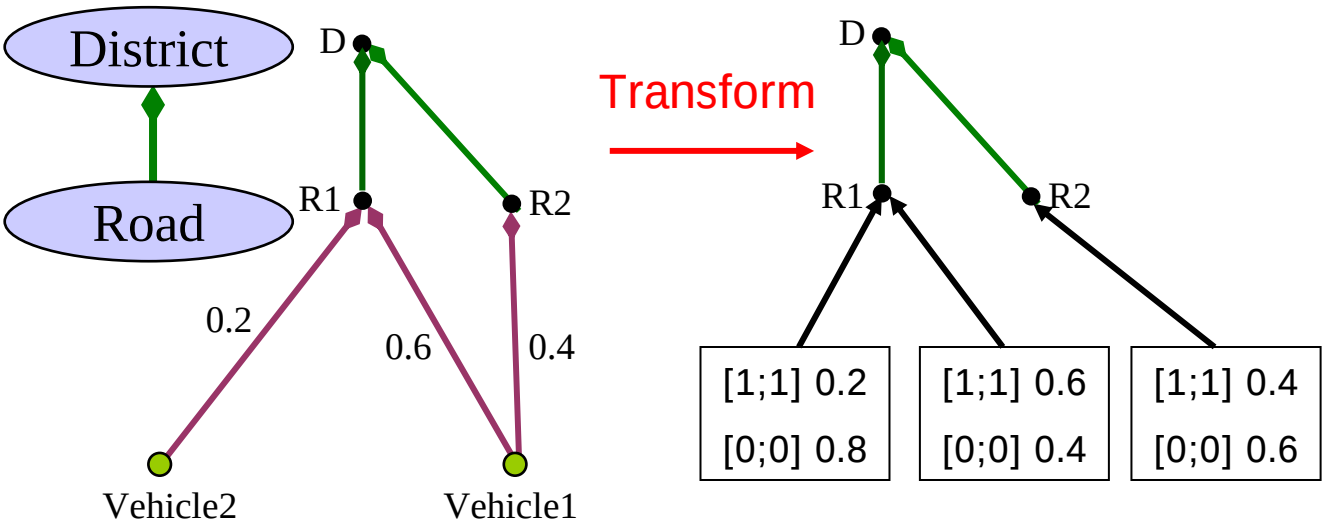
Pre-Aggregation

- Definition of Pre-Aggregation
  - Creating probability distributions out of fact data
- Adaptation of aggregation algorithm:
  - Each fact-dimension relationship is transformed into a probability distribution
  - The obtained distributions are summed



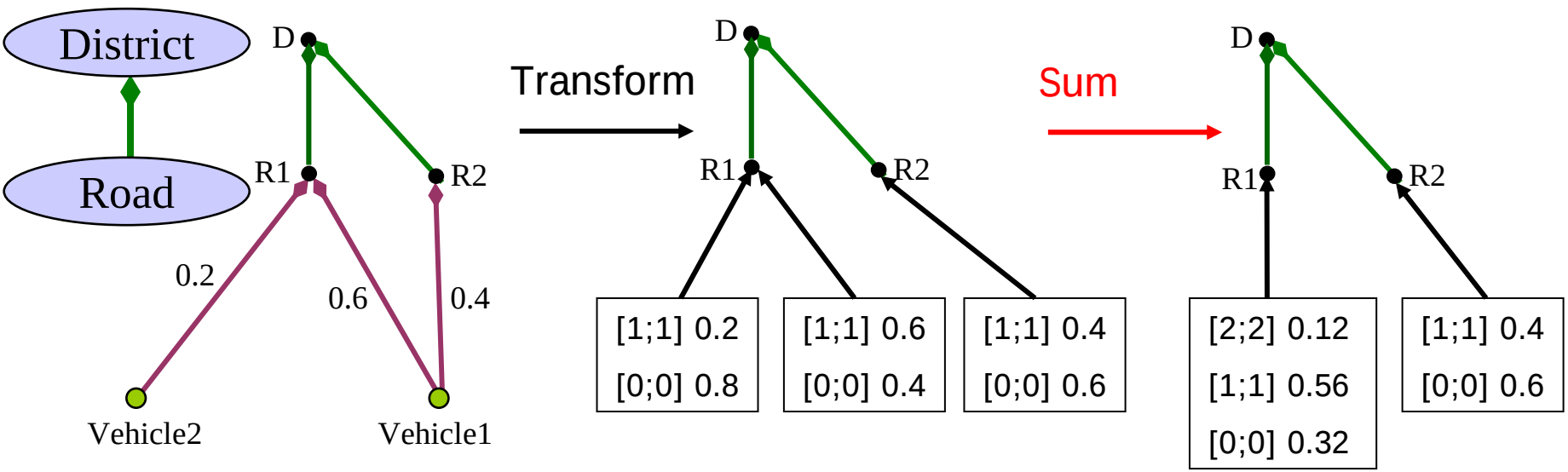
Pre-Aggregation

- Definition of Pre-Aggregation
  - Creating probability distributions out of fact data
- Adaptation of aggregation algorithm:
  - Each fact-dimension relationship is transformed into a probability distribution
  - The obtained distributions are summed



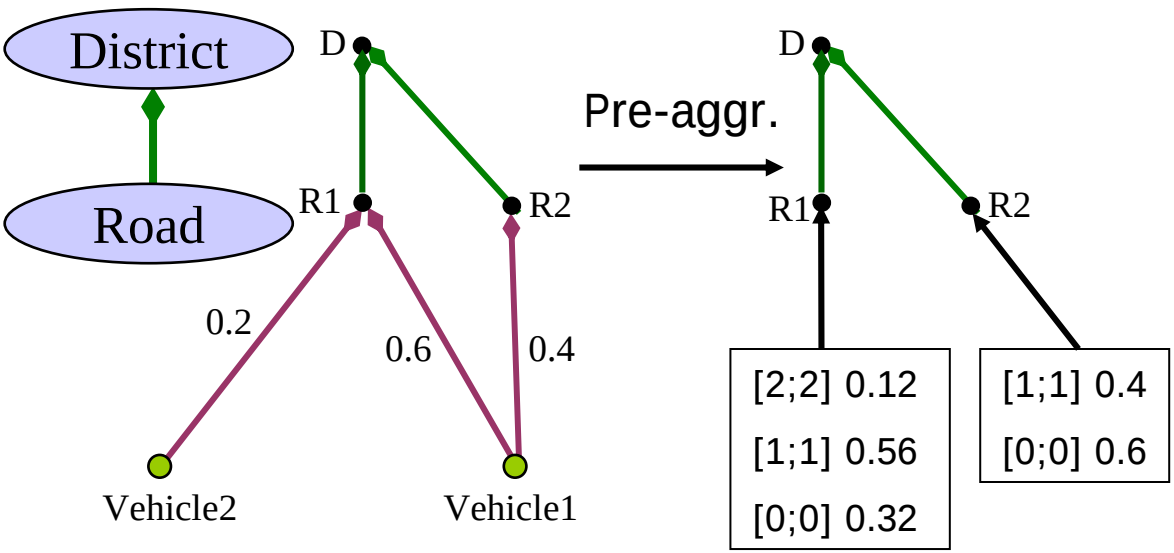
Pre-Aggregation

- Definition of Pre-Aggregation
  - Creating probability distributions out of fact data
- Adaptation of aggregation algorithm:
  - Each fact-dimension relationship is transformed into a probability distribution
  - The obtained distributions are summed



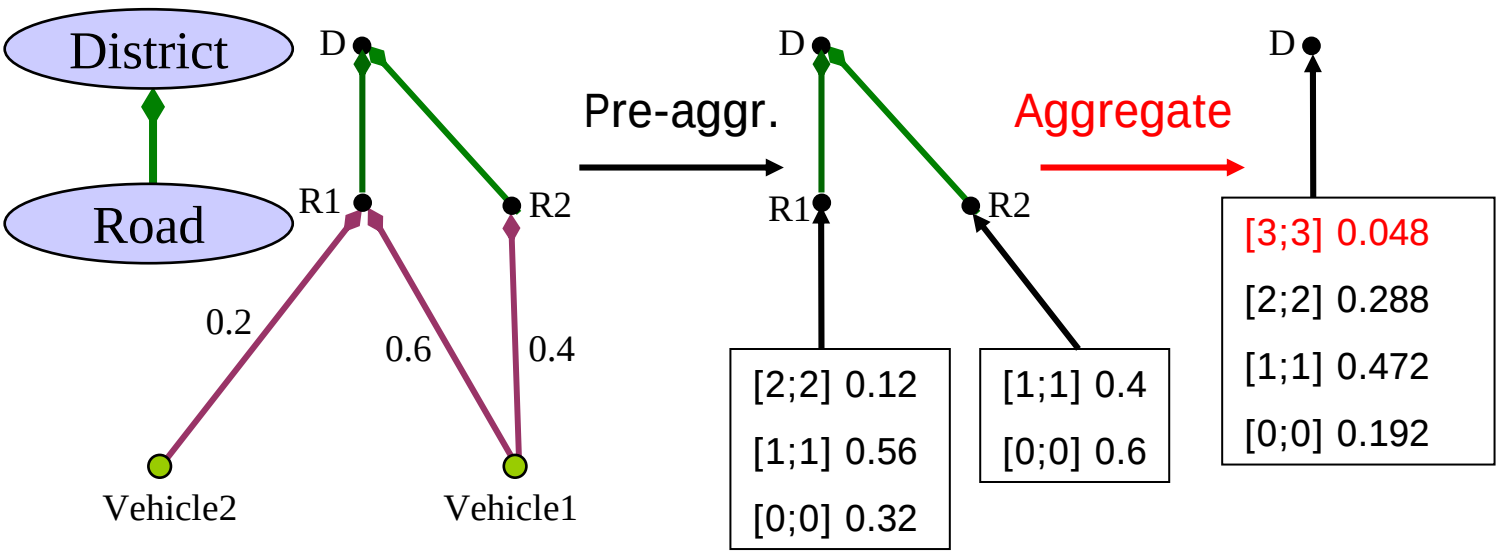
Pre-Aggregation: Summarizability

- Problem:
  - Vehicle attachments in pre-aggregated probability distributions are overcounted
- Solution:
  - If wrong values have small probabilities, filter them out



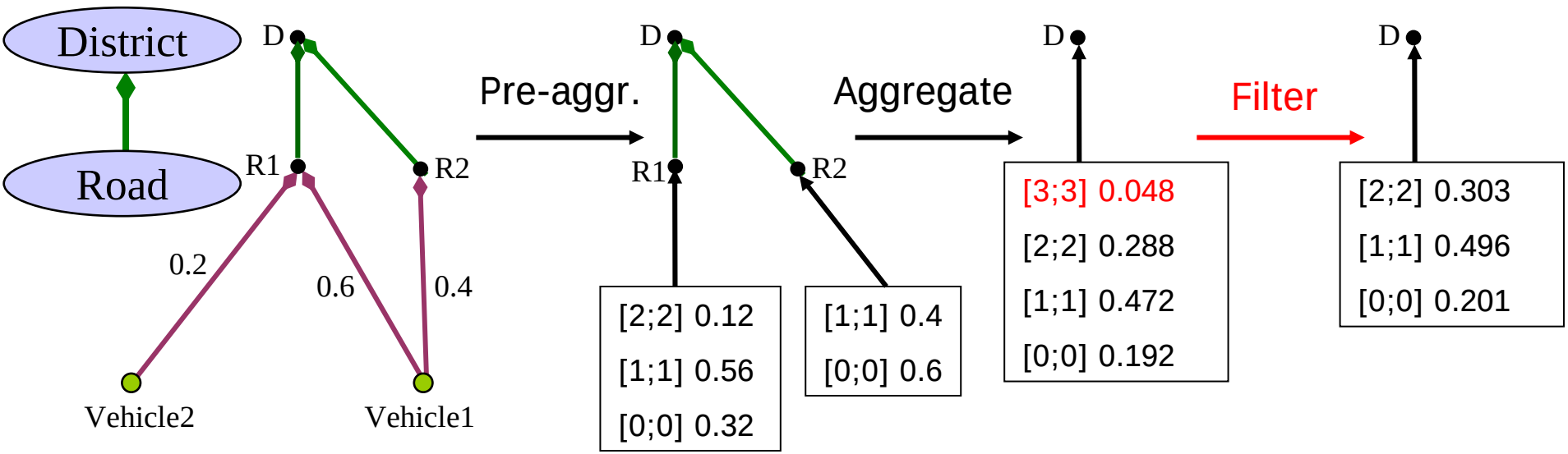
Pre-Aggregation: Summarizability

- Problem:
  - Vehicle attachments in pre-aggregated probability distributions are overcounted
- Solution:
  - If wrong values have small probabilities, filter them out



Pre-Aggregation: Summarizability

- Problem:
  - Vehicle attachments in pre-aggregated probability distributions are overcounted
- Solution:
  - If wrong values have small probabilities, filter them out

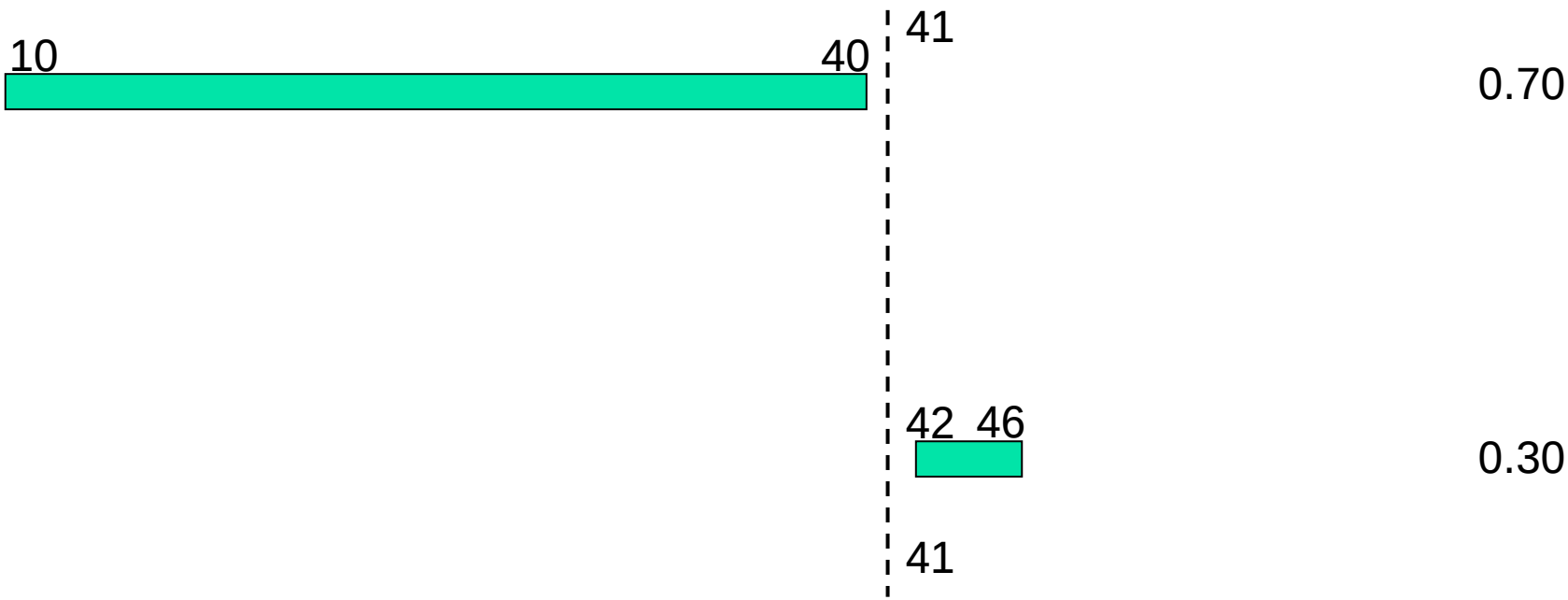


Talk Outline

- Introduction
- Probability Distributions As Measures
- Aproximate Aggregation
- Pre-Aggregation
- Queries over Approximate Probability Distributions
- Conclusions and Future Work
- Related Work

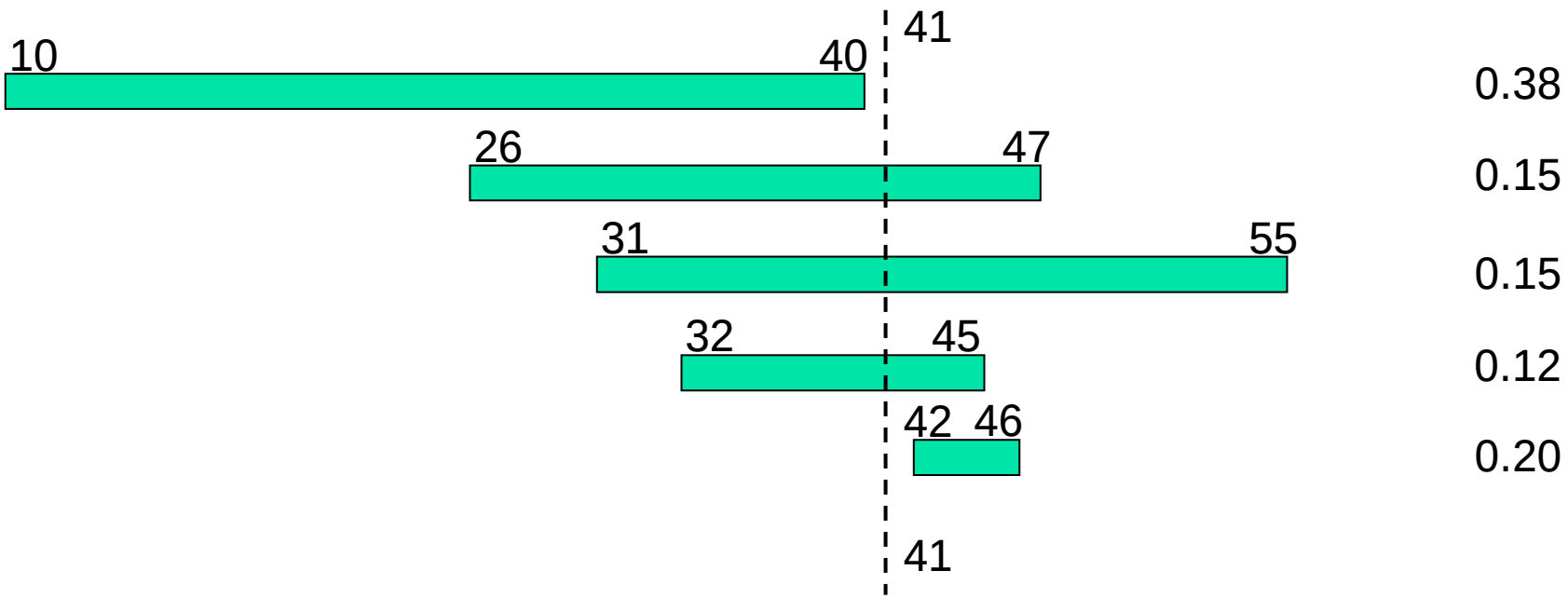
Query Processing: Probability Query

- Queries over an aggregate value (COUNT) for a dimension value, *s*
- Probability query:
  - E.g., “What is the probability that COUNT for *s* exceeds 41?”



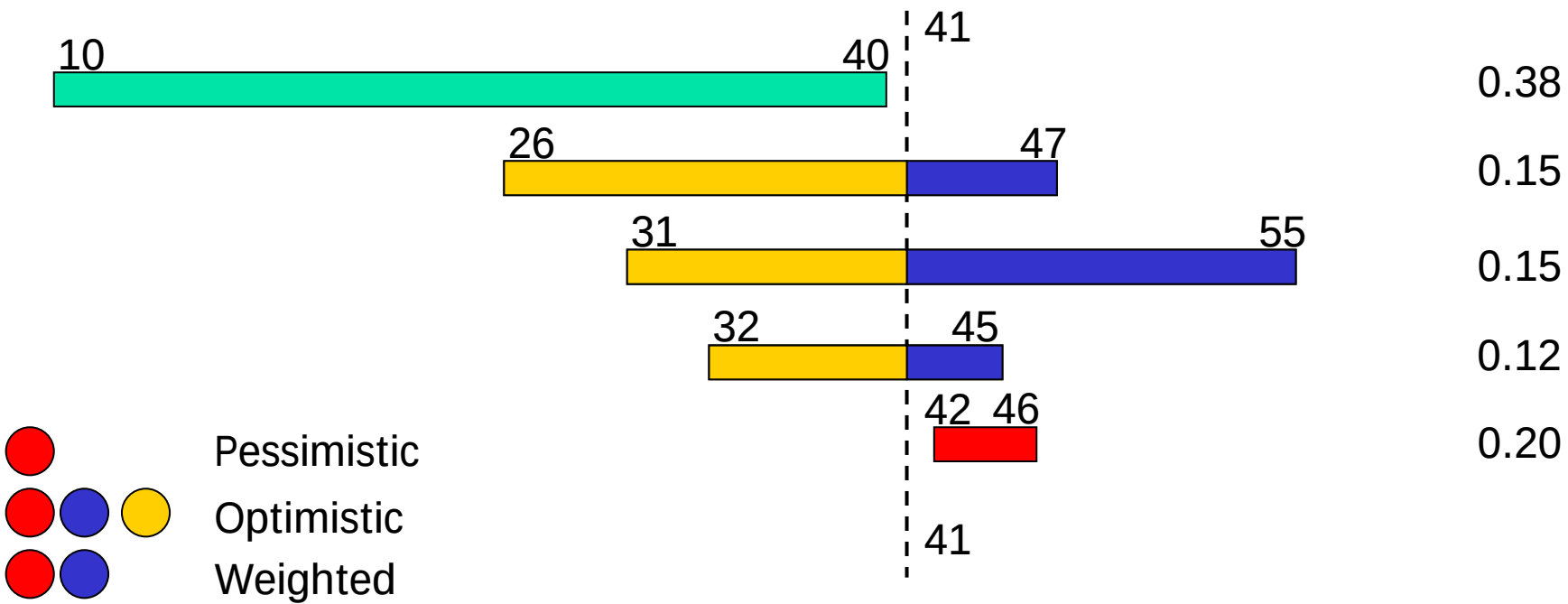
Query Processing: Probability Query

- Queries over an aggregate value (COUNT) for a dimension value, *s*
- Probability query:
  - E.g., “What is the probability that COUNT for *s* exceeds 41?”



Query Processing: Probability Query

- Queries over an aggregate value (COUNT) for a dimension value, *s*
- Probability query:
  - E.g., “What is the probability that COUNT for *s* exceeds 41?”



Conclusions

- OLAP/DW technology may enable complex analysis of LBS data that involve aggregation of probability distributions
- We are extending the current technology to support LBS data and queries
- Our contributions:
  - Generalized measure - probability distribution
  - Approximate aggregation and pre-aggregation for probability distributions
  - Processing of queries over approximate probability distributions

Future Work

- Integration of our methods and techniques into existing OLAP/DW systems
- Coalescion policies:
  - optimality (highest precision at lowest cost)
  - precision guarantees
- Support for dynamic content/future time queries:
  - Prediction of future aggregate values
  - Continuous evolution of aggregate values

Related Work

- OLAP Aggregation:
  - Approximate aggregation on certain data, but no uncertain data (e.g., Poosala and Ganti[SSDBM99])
  - Accurate aggregation of uncertain data, but no probability distributions (e.g., Moole[SoutheastCon03])
  - **Approximate aggregation of probability distributions, but no concrete representation of aggregate values and algorithms and no pre-aggregation (Burdick et al.[VLDB05])**
- Probabilistic Databases:
  - Uncertain data (e.g., Barbara et al.[TKDE92], Cavallo and Pitarelli[VLDB87], Dalvi and Suciu[VLDB04])
  - No approximate aggregation
- Spatio-temporal Databases:
  - Approximate aggregation of certain data (e.g., Tao et al.[ICDE2004])
  - No approximate aggregation of uncertain data
- Summation theory:
  - Accurate summation of uniformly sampled distributions (e.g., Regan et al.)
  - Approximate summation of infinitely many variables (e.g., Puckette)
  - Unrealistic assumptions
- Histograms:
  - Construction of optimal histograms (e.g., Jagadish et al.[VLDB98])
  - No summation of distributions

Questions?