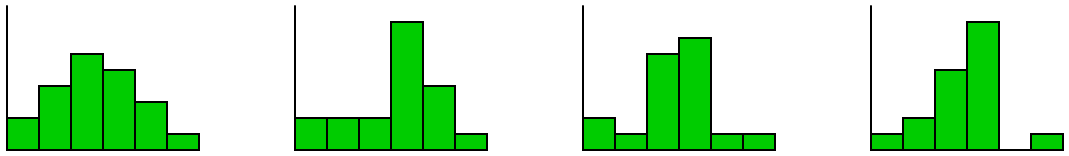
Histograms and Wavelets on Probabilistic Data



Graham Cormode

AT&T Labs-Research

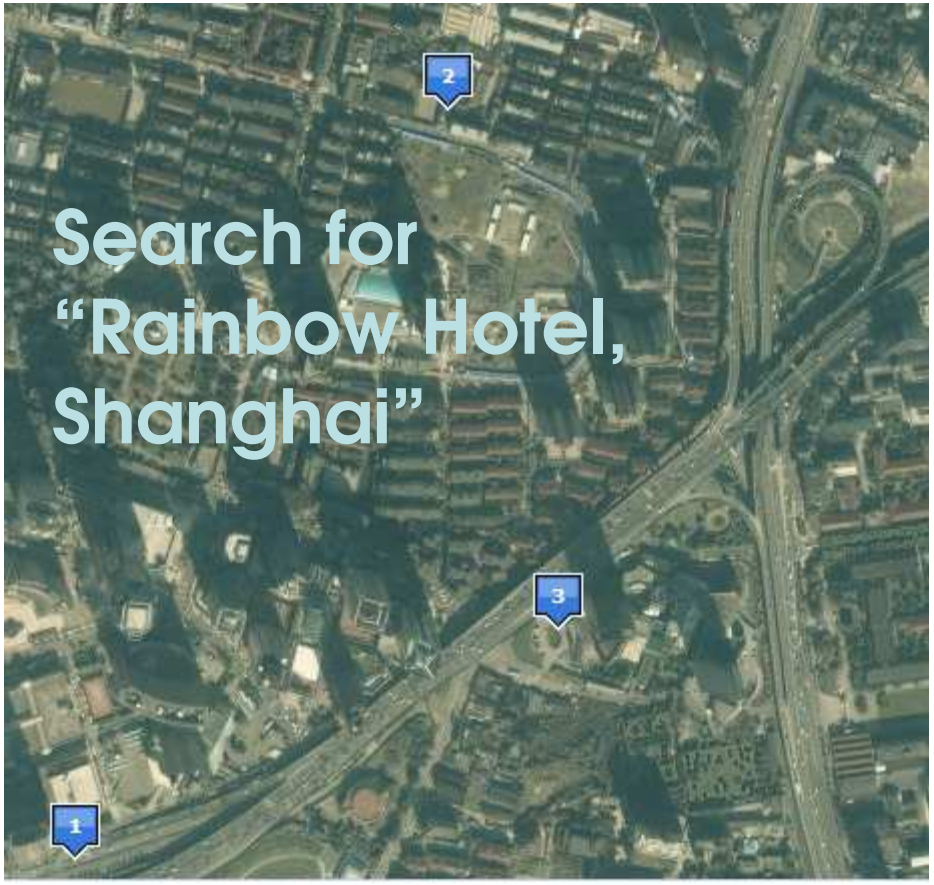
Minos Garofalakis

Technical University of Crete

Sources of Probabilistic Data

- Increasingly data is *uncertain* and *imprecise*
  - Data collected from sensors has errors and imprecisions
  - Record linkage has confidence of matches
  - Learning yields probabilistic rules
- Recent efforts to build uncertainty into the DBMS
  - Mystiq, Trio and MayBMS projects
  - Model uncertainty and correlations within tuples
  - Aim to allow general purpose queries over uncertain data

Uncertain Data



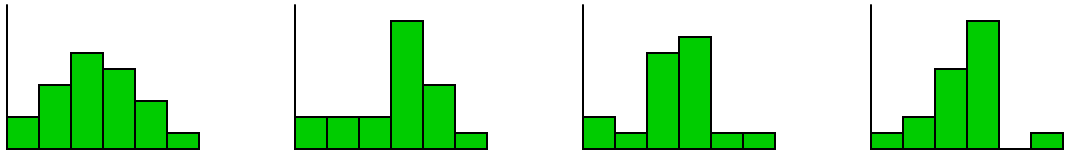
Query: How close is Galaxy Hotel to Rainbow Hotel?

Probabilistic Data Reduction

- Probabilistic data can be difficult to work with
  - Even simple queries can be #P complete [Dalvi, Suciu '04]
  - Want to avoid materializing all possible worlds
- Seek compact representations of probabilistic data
  - Data synopses which capture key properties
  - Can perform expensive operations on compact summaries
  - Histograms and wavelets used in traditional systems
- Challenge: how to build optimal synopses?

Models of Data

- Model defines a distribution over possible worlds, W
- Limit correlations to keep models compact
- Value pdf model:
  - Each item has independent distribution of frequencies



- Tuple pdf model:
  - Each tuple has a distribution of possible values
  - Can interpret as non-independent Value pdf model

Histograms for Probabilistic Data

- A histogram partitions a domain into buckets
  - All values within a bucket behave similarly
  - Can be represented by a single value
- Apply same idea to probabilistic data
  - Partition domain to minimize *expected* error
  - Key problem is finding cost of a given bucket
  - Use dynamic programming to find overall cost

Sum Squared Error Histograms

- Given a bucket b=(s,e), choose representative value r
  - Pick c to minimize expected squared error
  - Frequency of item i in world W is $g_i(W)$
  - (Expected) cost = $\sum_{i=s}^{e} \sum_{\text{worlds } W} \Pr[W] \cdot (g_i(W) - r)^2$
- Cost minimized by r = mean value in the bucket
  - Mean given by $r = \sum_{i=s}^{e} \sum_{\text{worlds } W} \Pr[W] \cdot g_i(W)$
  - Generalizes the deterministic case
  - Cost of a bucket is sum of expected sum of squares, less scaled expected square of sums
- How to compute the cost efficiently on demand?

Sum Squared Error Histograms

- Use the fact that $E[X^2] = Var[X] + E[X]^2$ to simplify
  - Apply independence and summation of variance
  - Rewrite cost of a bucket in terms of sums of values per item
  - Keep prefix sums of these values to find sum of any range
- With precomputation, find bucket cost in $O(1)$ time
  - Find optimal B-bucket histogram in time $O(Bn^2)$ via DP
- Holds for both tuple and value pdf models
  - Linearity of expectation handles dependencies for tuple pdf

Sum Squared Relative Error

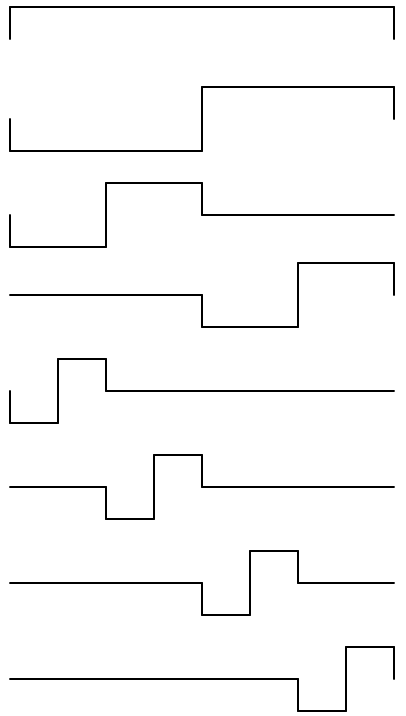
- Cost of bucket is $E_W[\sum_{i=s}^{e} (g_i(W) - r)^2/\max(c^2, g_i^2(W))]$
  - For representative value r, and constant c
- Expand out the quadratic numerator
  - Observe that denominator is fixed in any world W
  - Differentiate to find optimal value of r
- Gives bucket cost in terms of three expectations:
  - $\sum_i 1/\max(c^2, g_i^2)$; $\sum_i g_i/\max(c^2, g_i^2)$ and $\sum_i g_i^2/\max(c^2, g_i^2)$
  - Use prefix sums to find bucket cost in constant time
- Find optimal B-bucket SSRE histogram in time $O(Bn^2)$

Sum of Absolute Error

- Cost of bucket is $E_W[\sum_{i=s}^{e} | g_i(W) - r |]$
  - Break into sum of values above r, and those below
  - Minimize when r is some value with non-zero probability
  - As r varies, cost decreases to a minimum, then increases
- Can precompute prefix sums for different r values
  - Ternary search to find best choice of r for a bucket
  - Takes $O(\log |V|)$ steps over $|V|$ different values
- Find opt B-bucket SAE histogram in $O(n^2(B+\log |V|))$

Wavelets for Probabilistic Data

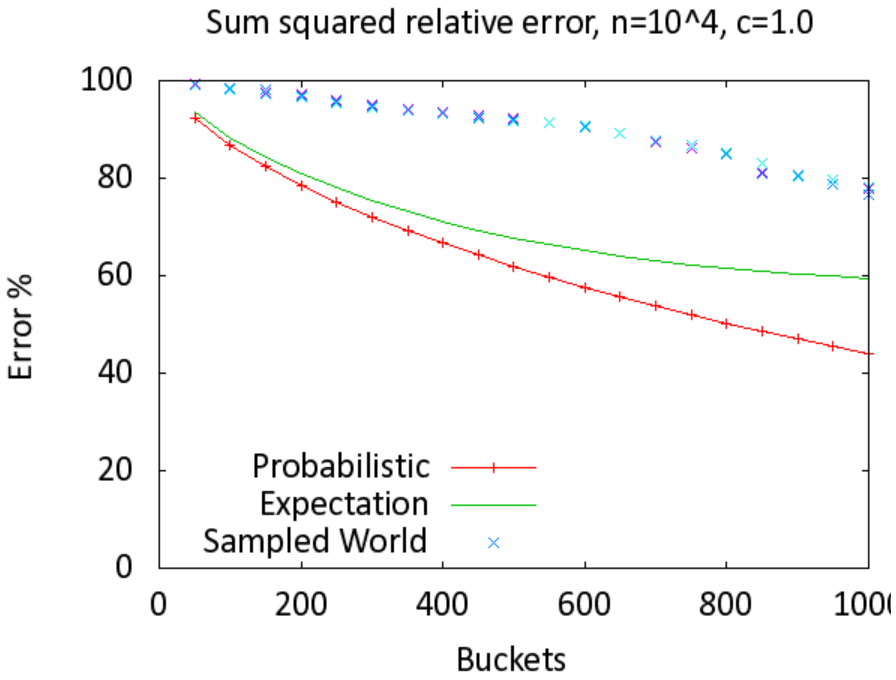
- Express data via B Haar basis functions
  - Seek to minimize expected squared error
- Use linearity of wavelet transform
  - Optimal to take expected coefficient values
  - Error due to dropping i'th coefficient = square of expected value
  - Best to pick B largest expected coefficients
- More complex under other error metrics
  - Perform DP over tree structure and coefficient values

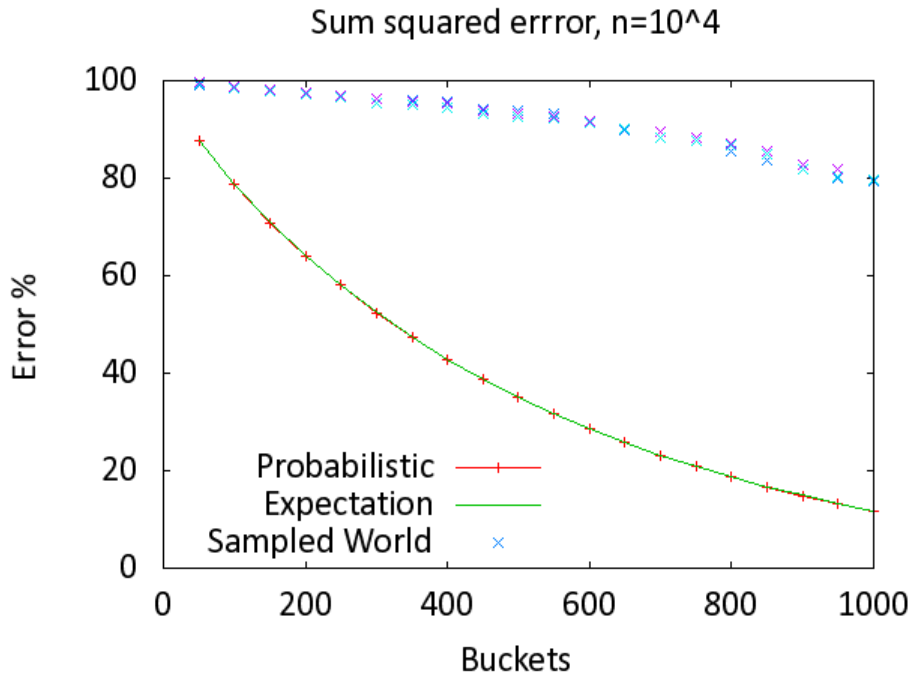


Experimental Study

- Evaluated on two probabilistic data sets
  - Real data from Mystiq Project (10K items)
  - Synthetic data from MayBMS generator (30K items)
- Compare to naïve methods to generate summaries:
  - Build wavelets/histograms over sampled possible world
  - Build wavelets/histograms over expected values
- Plot fraction of cost of 1 bucket – cost of n buckets

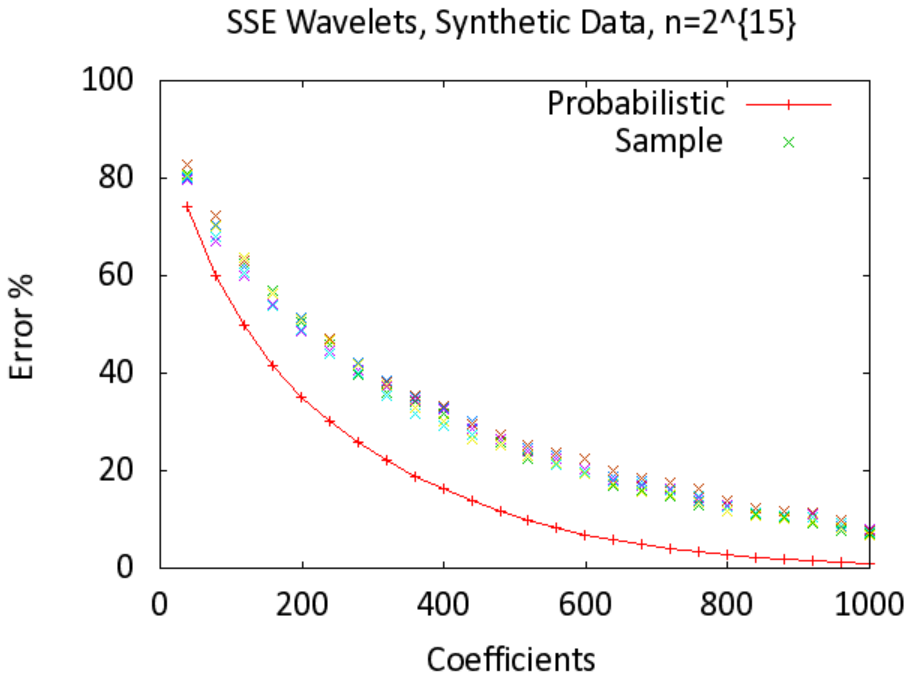
Sum Squared Error Histograms

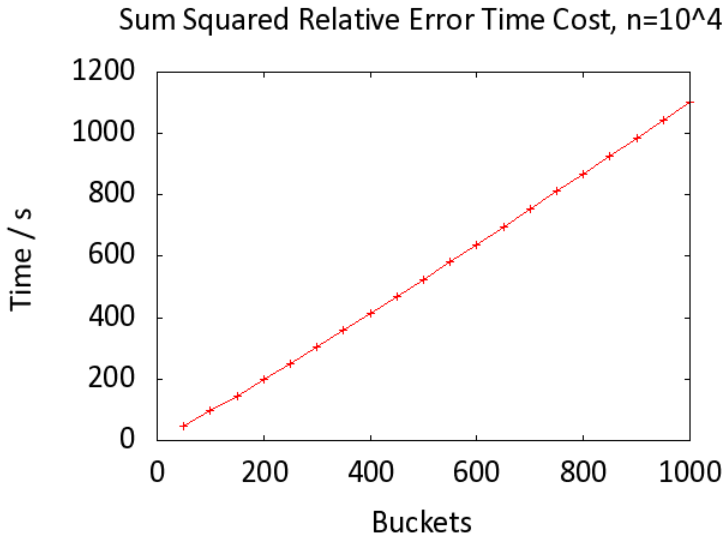


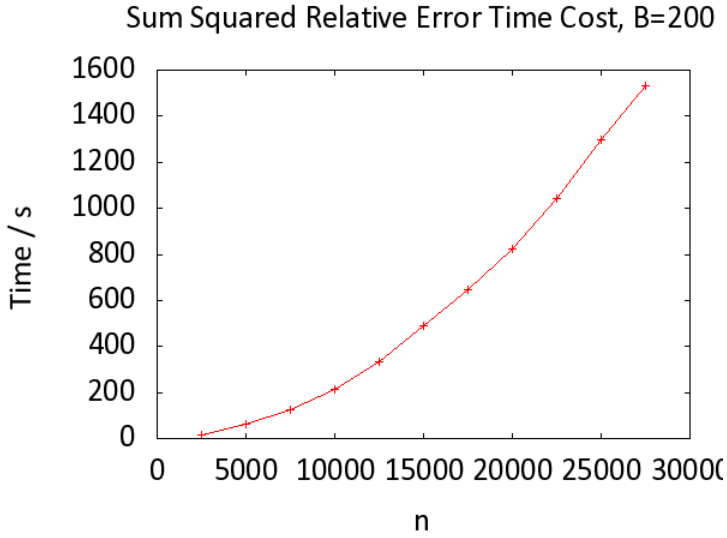


- Clear benefit for relative error over naïve methods
- Histograms on expected values almost as good for SSE

Time and Wavelets







- Time cost is linear in B, quadratic in n
  - Same cost for histogram of sample
- Expected coefficients shows clear benefit over sampling possible worlds

Concluding Remarks

- Can build synopses for probabilistic data

Advantages:

- Histograms and wavelets are familiar objects
- Leverage existing methods for processing summaries

Disadvantages:

- Dynamic programming can be slow (quadratic cost)
  - Can approximate using standard techniques
- Representation loses probabilistic semantics
  - Look for summaries that are more like pdfs?