Efficient Skyline Computation over Low-Cardinality Domains

Michael Morse[1] Jignesh M. Patel[2] H.V. Jagadish[2]

[1]MITRE Corporation

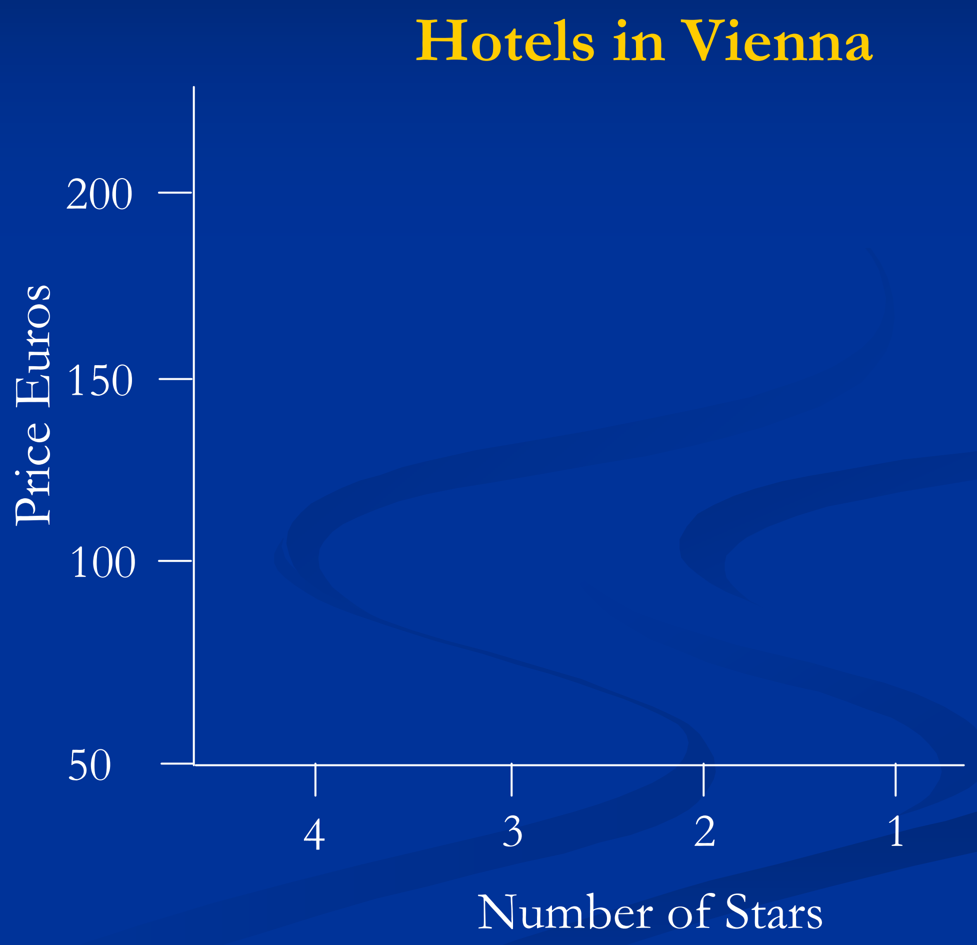
[2]University of Michigan

Overview

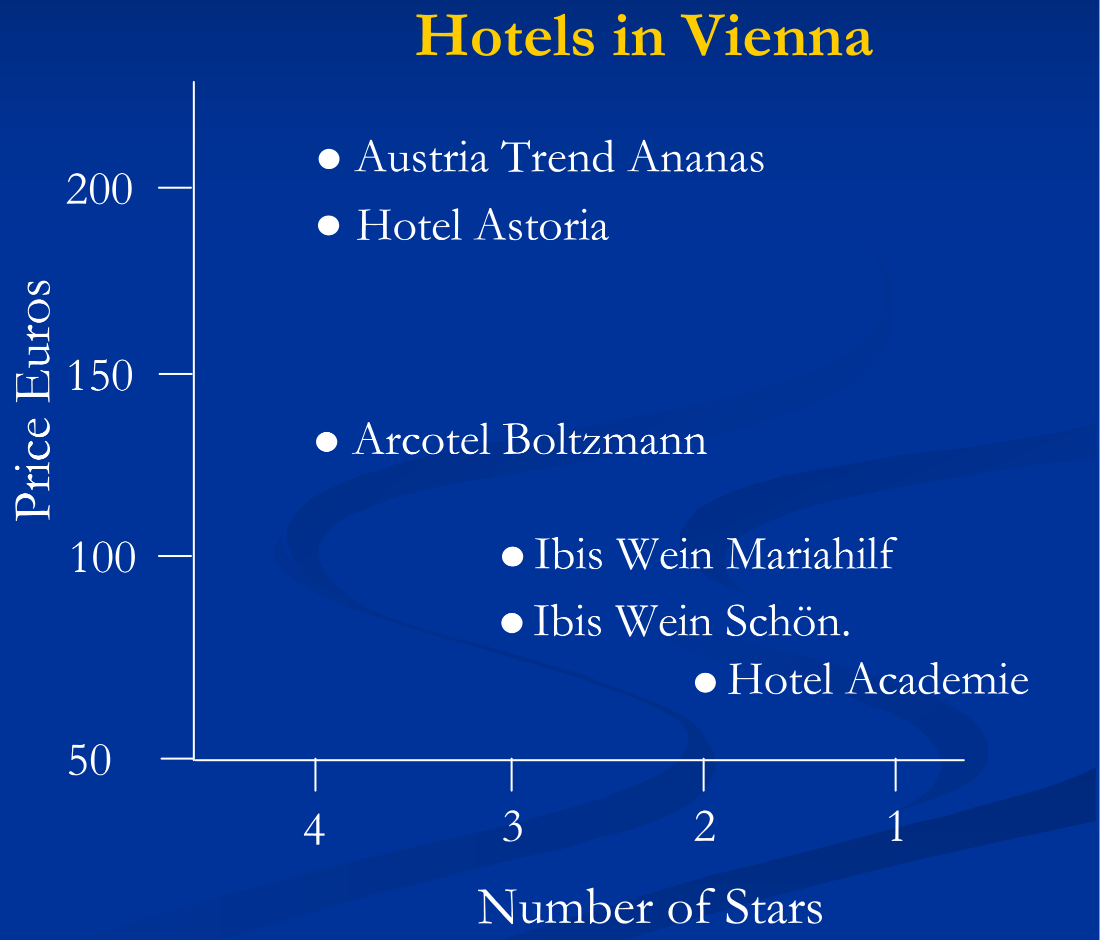
- Skyline Example and Definition.
- Discuss Low-Cardinality Attributes.
- Present the Lattice Skyline (LS) Algorithm.
- Discuss Experimental Results.
- Conclusions.

Traveling to VLDB

Traveling to VLDB

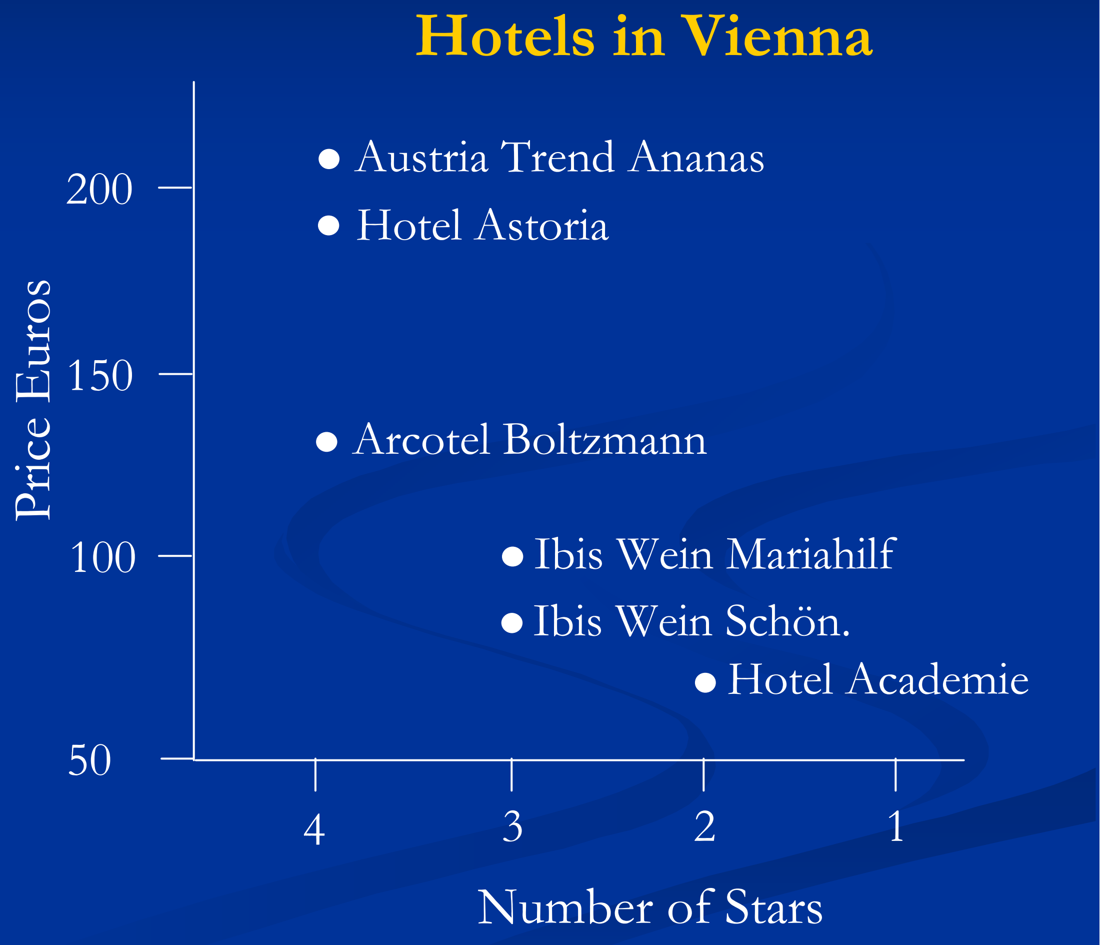


Traveling to VLDB



Traveling to VLDB

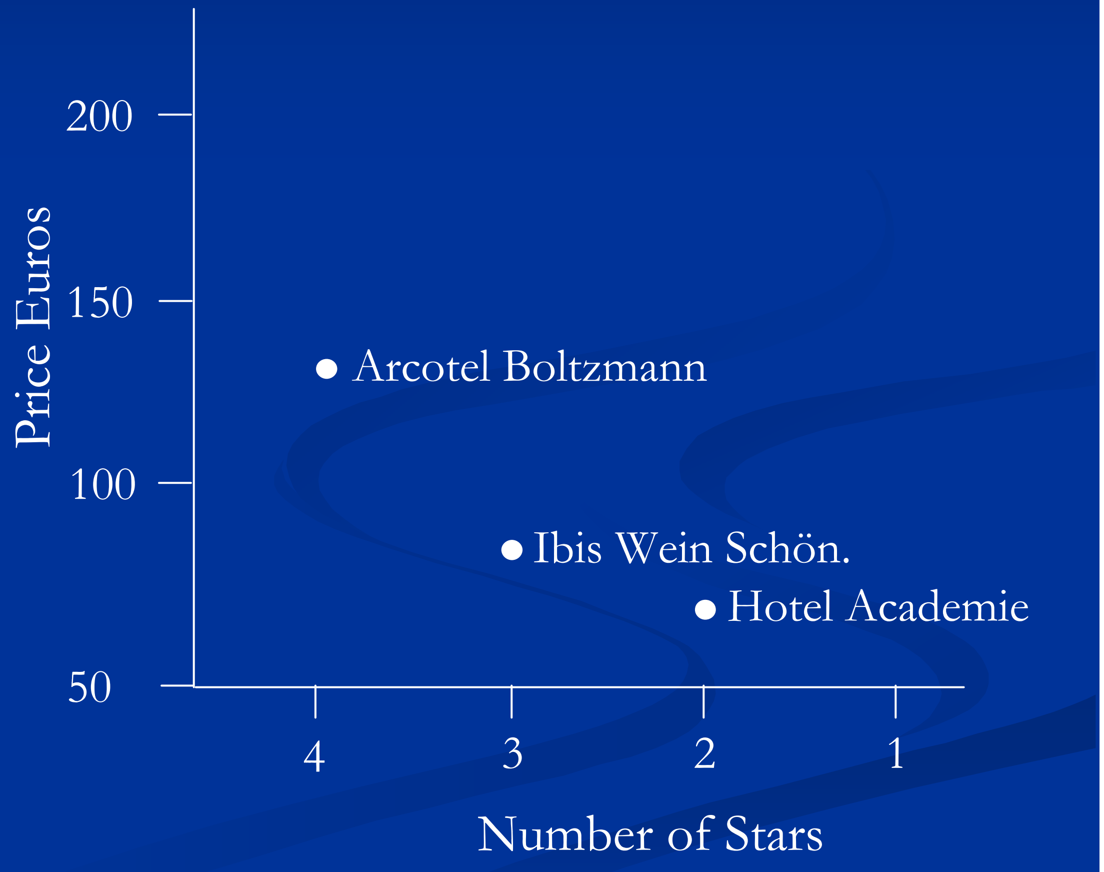
- Hotels that are more expensive than others and no higher rated are uninteresting.
  - e.g. The H. Astoria is more expensive than the Boltzmann,with the same rating.
- Such data points are said to be ‘dominated.’



Traveling to VLDB

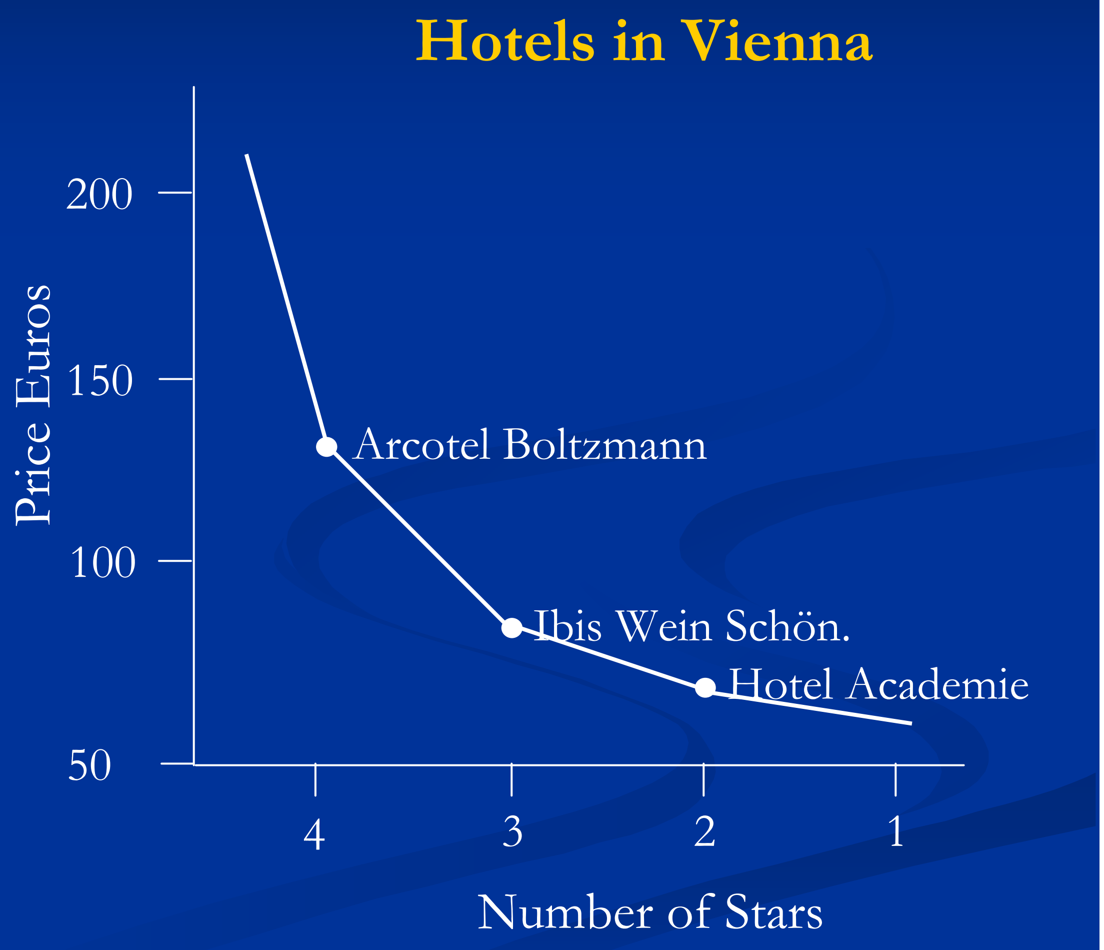
- Remove Dominated Hotels from consideration.

Hotels in Vienna



Traveling to VLDB

- Remove Dominated Hotels from consideration.
- We Obtain the Skyline for this Dataset.



Skyline Definition

- Skylines are an elegant summarization method for multidimensional datasets.
- Def: The skyline is the set of all points *p* in a dataset that are not dominated by some other point in that dataset.
- Equivalent to the Pareto Set or Maximal Vectors.

Overview

- Skyline Example and Definition.
- Discuss Low-Cardinality Attributes.
- Present the Lattice Skyline (LS) Algorithm.
- Discuss Experimental Results.
- Conclusions.

Attribute Domains

- Low-Cardinality Domain: the domain of possible values for attribute a_i is a small number.

Attribute Domains

- Low-Cardinality Domain: the domain of possible values for attribute a_i is a small number.
- We will consider datasets with d low-cardinality domains and optionally 1 unrestricted domain.

Attribute Domains

- Low-Cardinality Domain: the domain of possible values for attribute a_i is a small number.
- We will consider datasets with d low-cardinality domains and optionally 1 unrestricted domain.
- Example: We are interested in finding a highly rated hotel according to two different rating measures that is inexpensive.

Attribute Domains

- Low-Cardinality Domain: the domain of possible values for attribute a_i is a small number.
- We will consider datasets with d low-cardinality domains and optionally 1 unrestricted domain.
- Example: We are interested in finding a highly rated hotel according to two different rating measures that is inexpensive.

Hotel	Stars	Survey	Price
Slumber Well	⋆	Medium	120
Soporific Inn	⋆ ⋆	Low	65
Drowsy Hotel	⋆ ⋆	High	110
Celestial Sleep	⋆ ⋆ ⋆	Medium	101
Nap Motel	⋆ ⋆	Low	101

Related Algorithms

- Methods requiring indexing/preprocessing.
  - Nearest Neighbor [Kossman et al., VLDB 2002].
  - BBS [Papadias et al., SIGMOD 2003].
  - Bitmap, Index [Tan et al., VLDB 2001].
- Methods that require no preprocessing.
  - BNL [Borzsonyi et al., ICDE 2001].
  - SFS [Chomicki et al., ICDE 2003].
  - LESS [Godfrey et al., VLDB 2005].
- Many other related problems cited in the paper.
  - Probabilistic Skylines [Pei et al., VLDB 2007].
  - ZBtree [Lee et al., VLDB 2007].
  - Reverse Skylines [Dellis et al., VLDB 2007].

Related Algorithms

- Best Alternative: LESS

  [Godfrey et al. "Maximal Vector Computation in Large Datasets" VLDB 05]

  1. Preprocessing.
  2. Sorts data.
  3. Pairwise comparison of remaining tuples.

- Cost: between O(n) and O(n^2).
- One downside, can be sensitive to the dataset distribution and the tuple ordering.

Our Contribution

- We develop a new algorithm called the Lattice Skyline (LS) algorithm for skyline evaluation for datasets with low-cardinality domains.
- What we show in the experiments is that while LESS is more general, it is less efficient than LS.

Overview

- Skyline Example and Definition.
- Discuss Low-Cardinality Attributes.
- Present the Lattice Skyline (LS) Algorithm.
- Discuss Experimental Results.
- Conclusions.

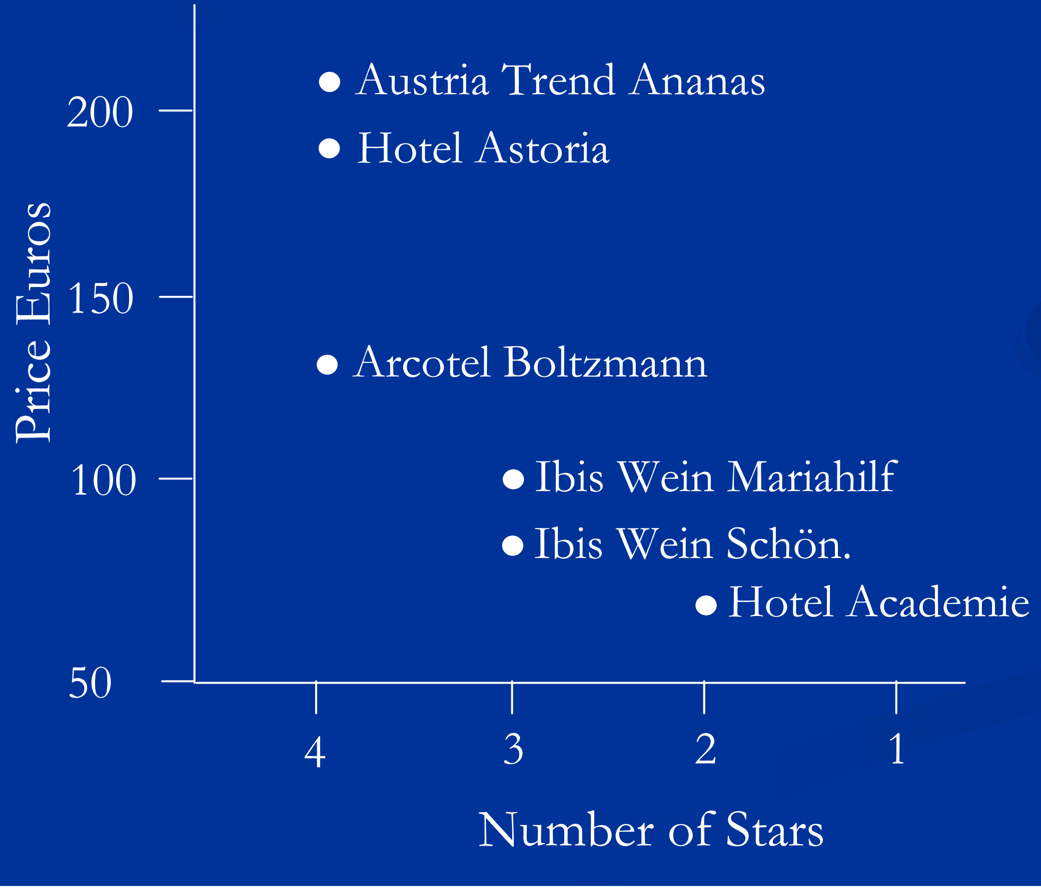
Partial Order Imposed by Skyline

- The skyline operator imposes a partial order on a dataset through the ‘dominance’ relationship ‘> ’.

Partial Order Imposed by Skyline

- The skyline operator imposes a partial order on a dataset through the 'dominance' relationship '> '.

Hotels in Vienna



Boltzmann > H. Astoria

Boltzmann ⊁ Mariahilf

Mariahilf ⊁ Boltzmann

Partial Order Imposed by Skyline

- The skyline operator imposes a partial order on a dataset through the ‘dominance’ relationship ‘> ’.
- This dataset and the skyline operator are not a lattice since there isn’t an upper or lower bound.

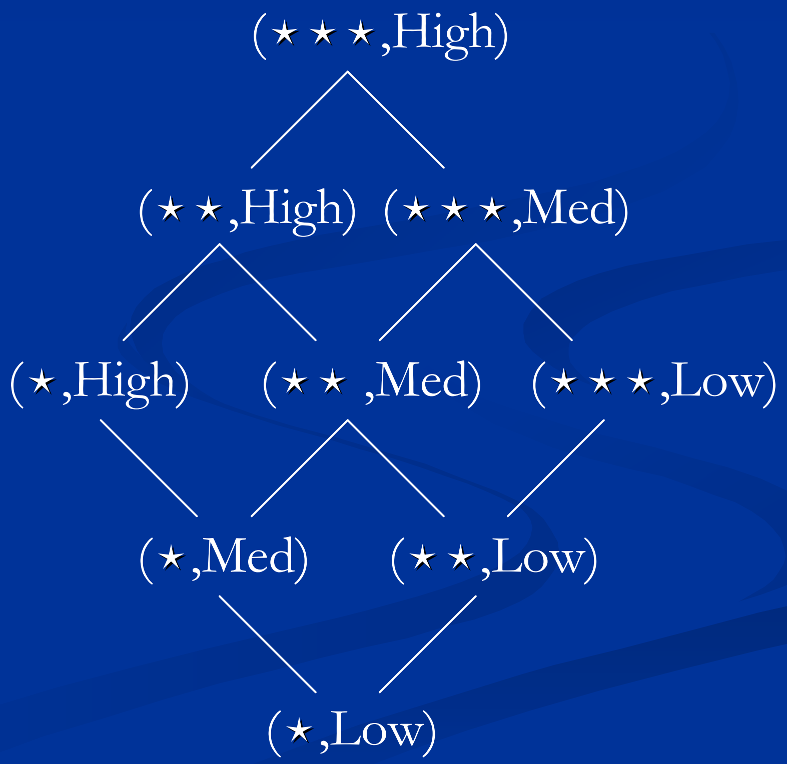
Partial Order Imposed by Skyline

- The skyline operator imposes a partial order on a dataset through the ‘dominance’ relationship ‘> ’.
- This dataset and the skyline operator are not a lattice since there isn’t an upper or lower bound.
- Dataspaces with attributes drawn from low-cardinality domains and the skyline operator are a lattice.

Lattice Structure

- If we consider the low-cardinality attribute space (Stars, Survey), we obtain a lattice:

Hotel	Position
Slumber Well	(⋆,Med)
Soporific Inn	(⋆⋆,Low)
Drowsy Hotel	(⋆⋆,High)
Celestial Sleep	(⋆⋆⋆,Med)
Nap Motel	(⋆⋆,Low)



(⋆⋆⋆,High)

(⋆⋆,High) (⋆⋆⋆,Med)

(⋆,High) (⋆⋆ ,Med) (⋆⋆⋆,Low)

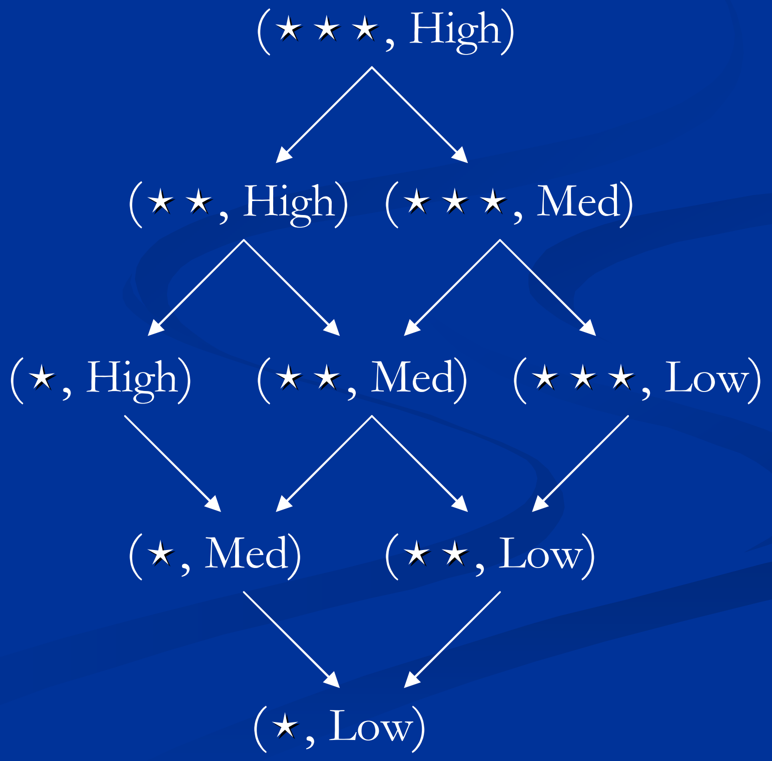
(⋆,Med) (⋆⋆,Low)

(⋆,Low)

Determining Dominance

- Elements that are reachable from others in the lattice-graph structure are dominated.

Hotel	Position
Slumber Well	(⋆,Med)
Soporific Inn	(⋆⋆,Low)
Drowsy Hotel	(⋆⋆,High)
Celestial Sleep	(⋆⋆⋆,Med)
Nap Motel	(⋆⋆,Low)



(⋆⋆⋆, High)

(⋆⋆, High) (⋆⋆⋆, Med)

(⋆, High) (⋆⋆, Med) (⋆⋆⋆, Low)

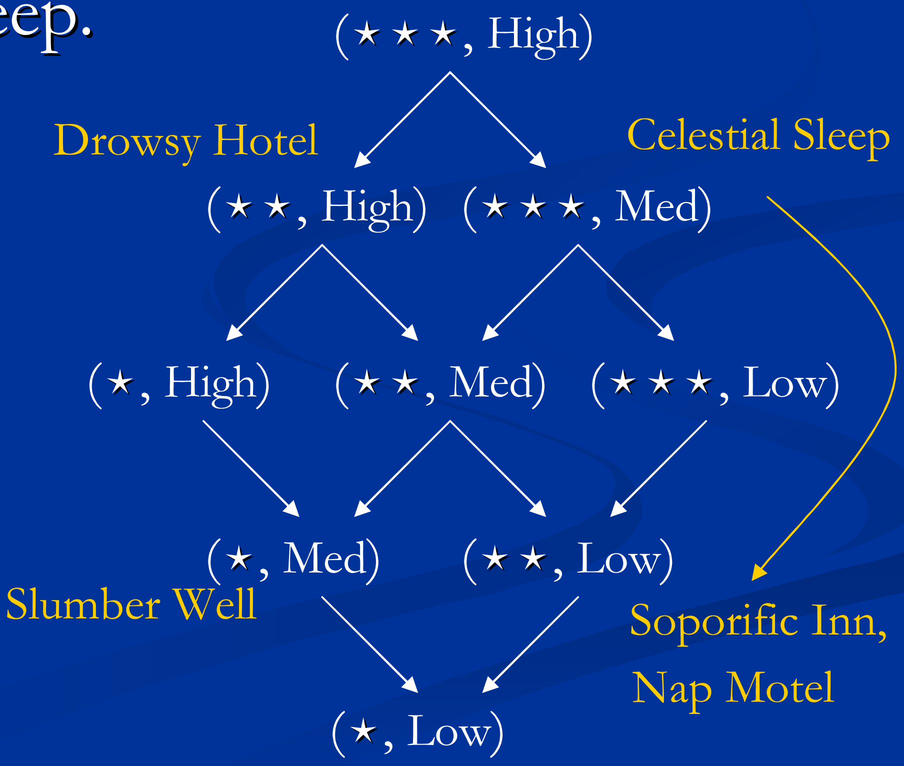
(⋆, Med) (⋆⋆, Low)

(⋆, Low)

Determining Dominance

- Elements that are reachable from others in the lattice-graph structure are dominated.
- Ex: (★★,Low) –Soporific Inn- is reachable from (★★★,Med) –Celestial Sleep.

Hotel	Position
Slumber Well	(★,Med)
Soporific Inn	(★★,Low)
Drowsy Hotel	(★★,High)
Celestial Sleep	(★★★,Med)
Nap Motel	(★★,Low)



(★★★, High)

Drowsy Hotel

Celestial Sleep

(★★, High) (★★★, Med)

(★, High) (★★, Med) (★★★, Low)

(★, Med) (★★, Low)

Slumber Well

Soporific Inn,

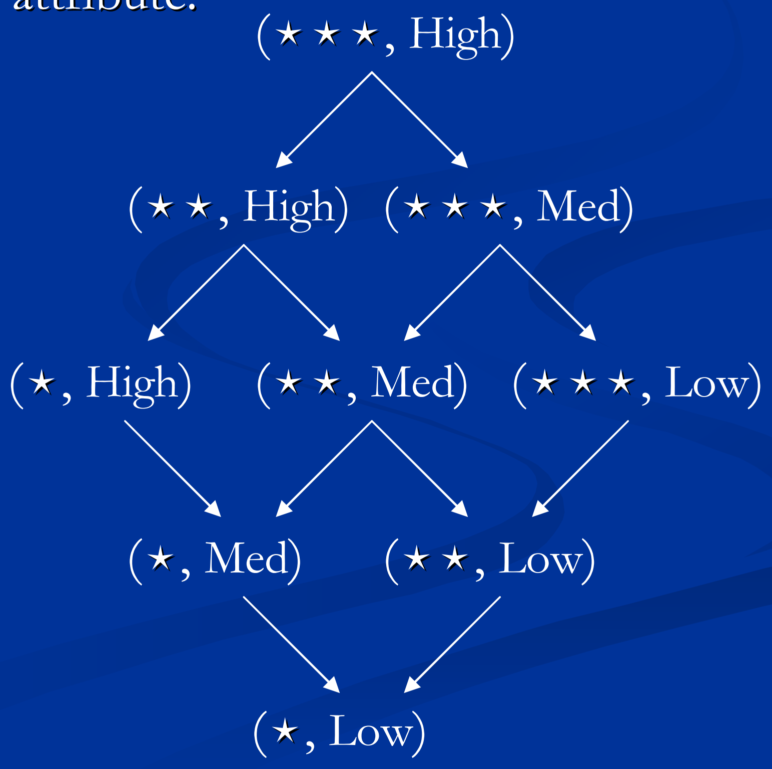
Nap Motel

(★, Low)

Lattice Skyline (LS) Algorithm

- For each lattice entry, maintain 2 pieces of information:
  1. Whether an element is present or not present in the data.
  2. The best value of the unrestricted attribute.

Hotel	Position
Slumber Well	(⋆,Med)
Soporific Inn	(⋆⋆,Low)
Drowsy Hotel	(⋆⋆,High)
Celestial Sleep	(⋆⋆⋆,Med)
Nap Motel	(⋆⋆,Low)



(⋆⋆⋆, High)

(⋆⋆, High) (⋆⋆⋆, Med)

(⋆, High) (⋆⋆, Med) (⋆⋆⋆, Low)

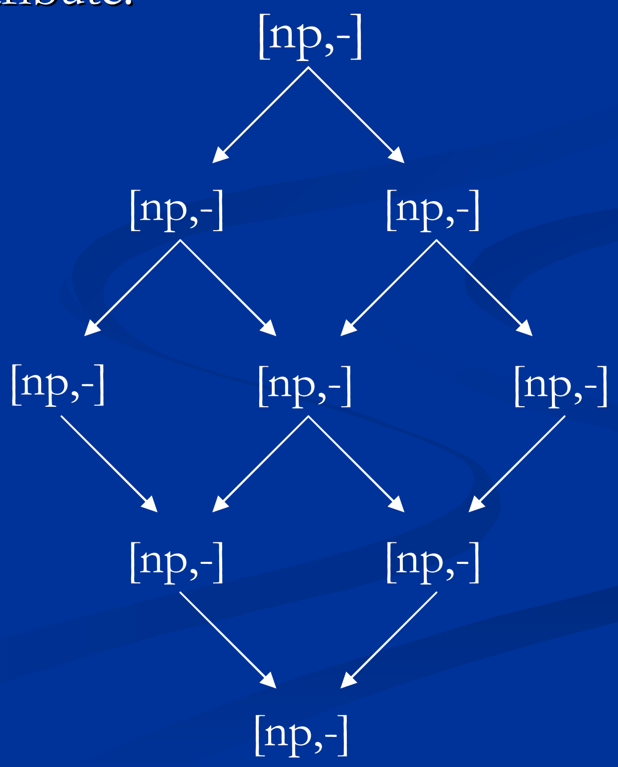
(⋆, Med) (⋆⋆, Low)

(⋆, Low)

Lattice Skyline (LS) Algorithm

- For each lattice entry, maintain 2 pieces of information:
    1. Whether an element is present or not present in the data.
    2. The best value of the unrestricted attribute.

Hotel	Position
Slumber Well	(⋆,Med)
Soporific Inn	(⋆⋆,Low)
Drowsy Hotel	(⋆⋆,High)
Celestial Sleep	(⋆⋆⋆,Med)
Nap Motel	(⋆⋆,Low)



[np,-]

[np,-] [np,-]

[np,-] [np,-] [np,-]

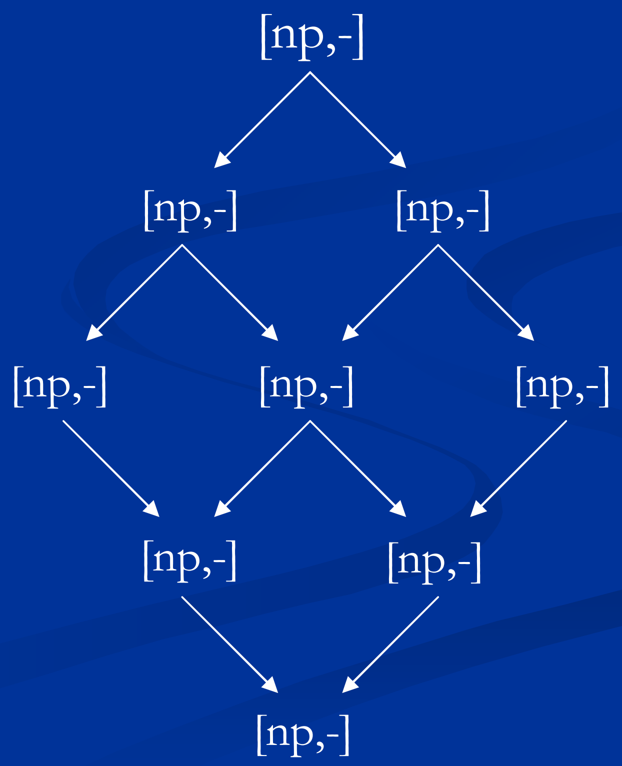
[np,-] [np,-]

[np,-]

Lattice Skyline (LS) Algorithm

- Iterate through the dataset.

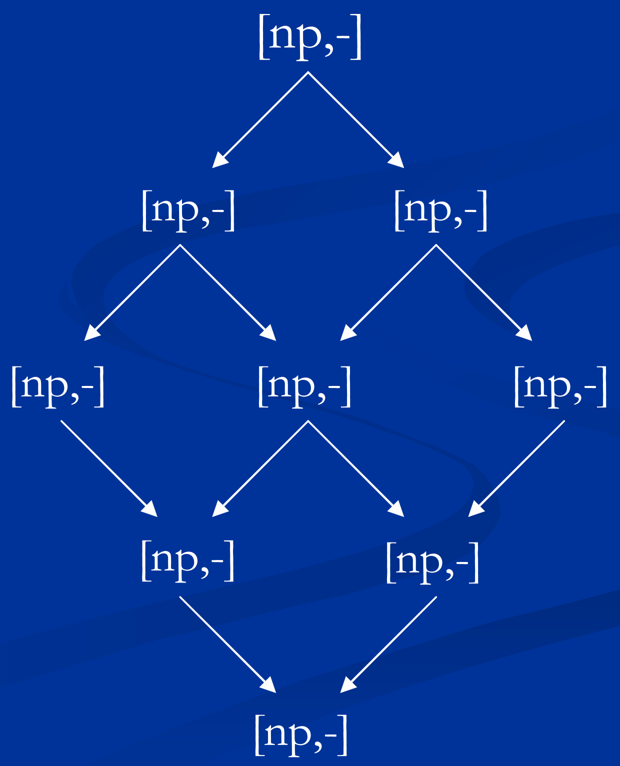
Hotel	Position	Price
Slumber Well	(★,Med)	120
Soporific Inn	(★★,Low)	65
Drowsy Hotel	(★★,High)	110
Celestial Sleep	(★★★,Med)	101
Nap Motel	(★★,Low)	101



Lattice Skyline (LS) Algorithm

- Iterate through the dataset.

Hotel	Position	Price
Slumber Well	(⋆,Med)	120
Soporific Inn	(⋆⋆,Low)	65
Drowsy Hotel	(⋆⋆,High)	110
Celestial Sleep	(⋆⋆⋆,Med)	101
Nap Motel	(⋆⋆,Low)	101



[np,-]

[np,-] [np,-]

[np,-] [np,-] [np,-]

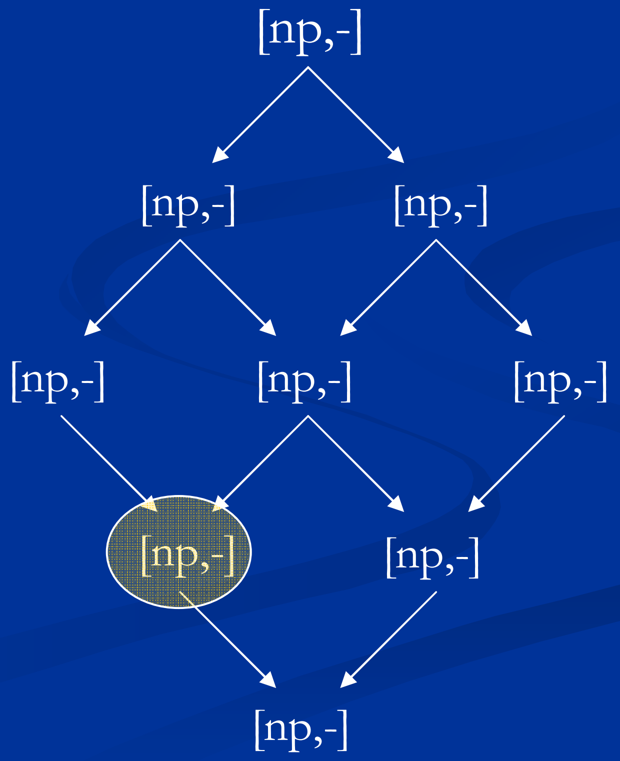
[np,-] [np,-]

[np,-]

Lattice Skyline (LS) Algorithm

- Iterate through the dataset.
- Modify the lattice position corresponding to the data point.

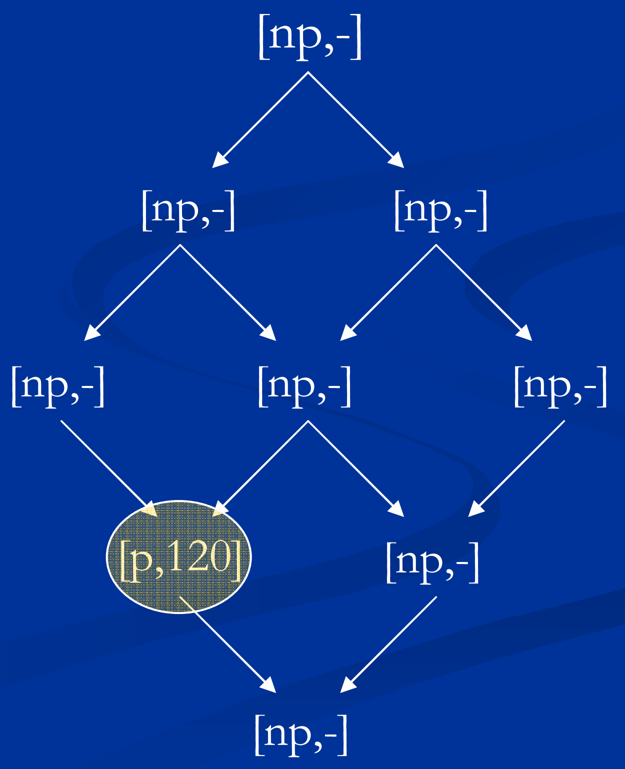
Hotel	Position	Price
Slumber Well	(★,Med)	120
Soporific Inn	(★★,Low)	65
Drowsy Hotel	(★★,High)	110
Celestial Sleep	(★★★,Med)	101
Nap Motel	(★★,Low)	101



Lattice Skyline (LS) Algorithm

- Iterate through the dataset.
- Modify the lattice position corresponding to the data point.

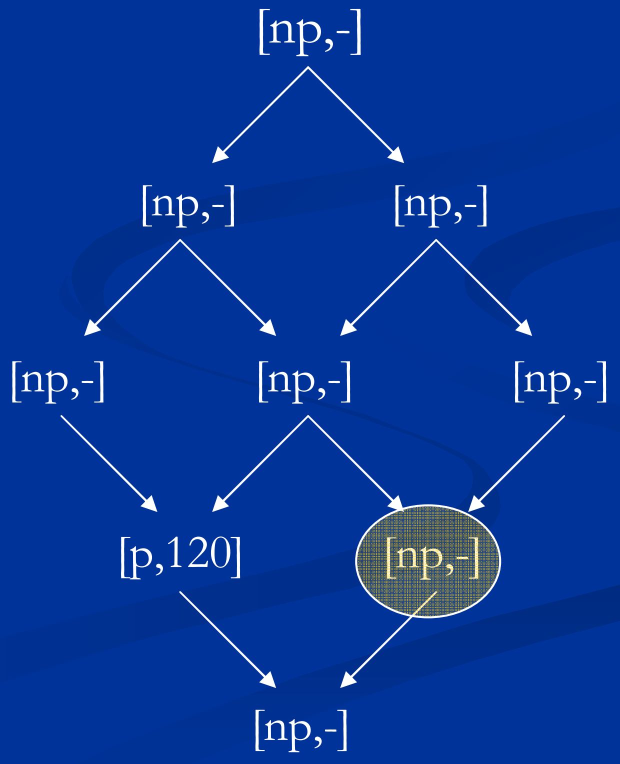
Hotel	Position	Price
Slumber Well	(⋆,Med)	120
Soporific Inn	(⋆⋆,Low)	65
Drowsy Hotel	(⋆⋆,High)	110
Celestial Sleep	(⋆⋆⋆,Med)	101
Nap Motel	(⋆⋆,Low)	101



Lattice Skyline (LS) Algorithm

- Iterate through the dataset.
- Modify the lattice position corresponding to the data point.

Hotel	Position	Price
Slumber Well	(★,Med)	120
Soporific Inn	(★★,Low)	65
Drowsy Hotel	(★★,High)	110
Celestial Sleep	(★★★,Med)	101
Nap Motel	(★★,Low)	101



[np,-]

[np,-] [np,-]

[np,-] [np,-] [np,-]

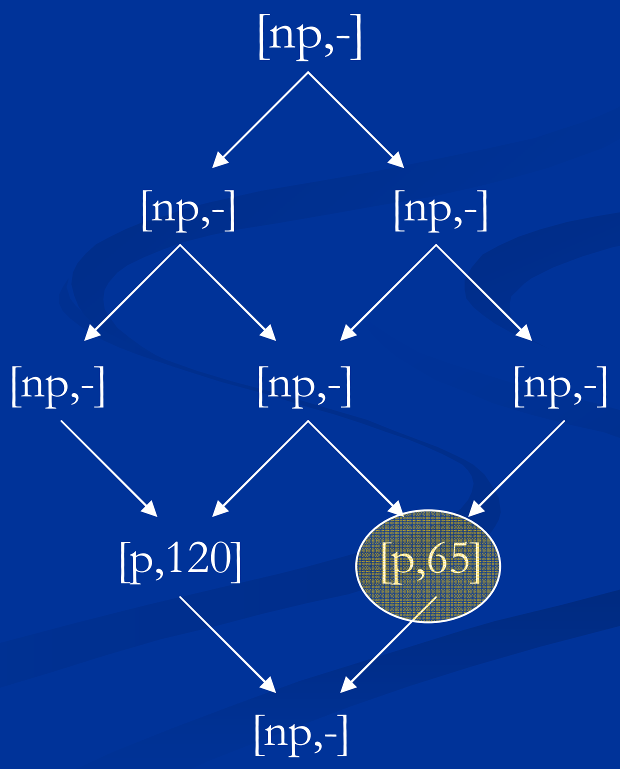
[p,120] [np,-]

[np,-]

Lattice Skyline (LS) Algorithm

- Iterate through the dataset.
- Modify the lattice position corresponding to the data point.

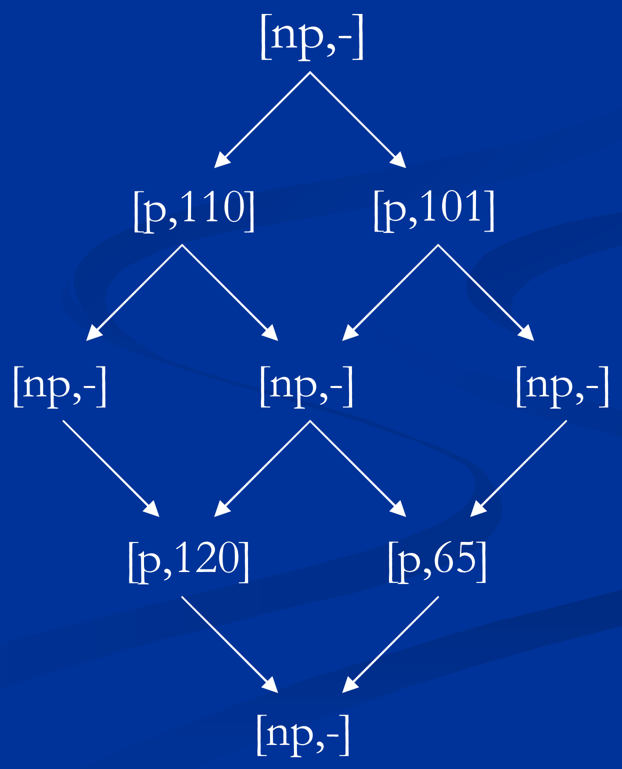
Hotel	Position	Price
Slumber Well	(⋆,Med)	120
Soporific Inn	(⋆⋆,Low)	65
Drowsy Hotel	(⋆⋆,High)	110
Celestial Sleep	(⋆⋆⋆,Med)	101
Nap Motel	(⋆⋆,Low)	101



Lattice Skyline (LS) Algorithm

- Iterate through the dataset.
- Modify the lattice position corresponding to the data point.

Hotel	Position	Price
Slumber Well	(★,Med)	120
Soporific Inn	(★★,Low)	65
Drowsy Hotel	(★★,High)	110
Celestial Sleep	(★★★,Med)	101
Nap Motel	(★★,Low)	101



[np,-]

[p,110] [p,101]

[np,-] [np,-] [np,-]

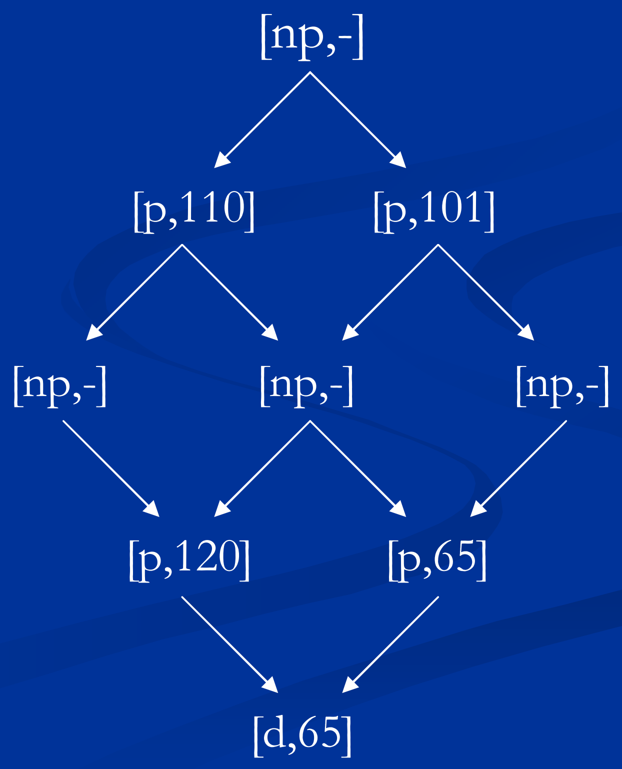
[p,120] [p,65]

[np,-]

Lattice Skyline (LS) Algorithm

- Compare each Lattice Element with Immediate Dominators in previous level.

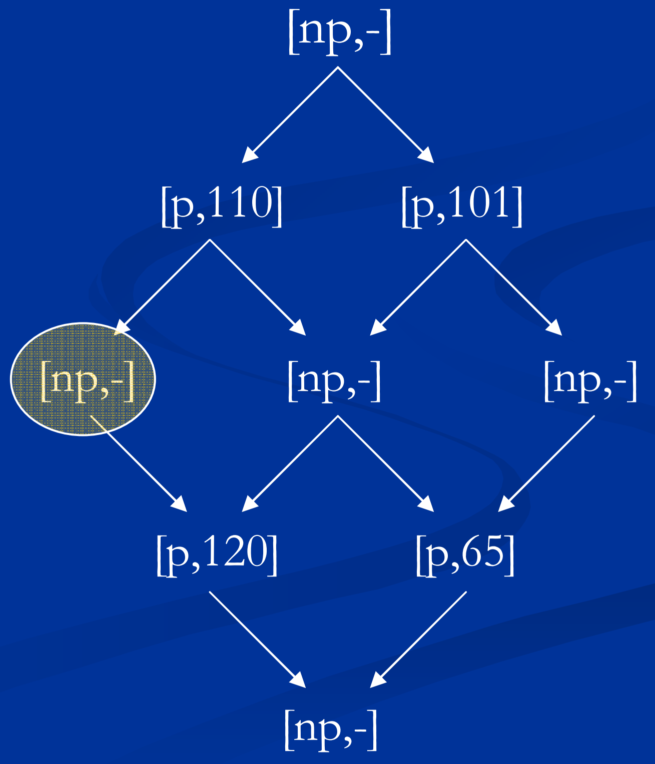
Hotel	Position	Price
Slumber Well	(⋆,Med)	120
Soporific Inn	(⋆⋆,Low)	65
Drowsy Hotel	(⋆⋆,High)	110
Celestial Sleep	(⋆⋆⋆,Med)	101
Nap Motel	(⋆⋆,Low)	101



Lattice Skyline (LS) Algorithm

- Compare each Lattice Element with Immediate Dominators in previous level.

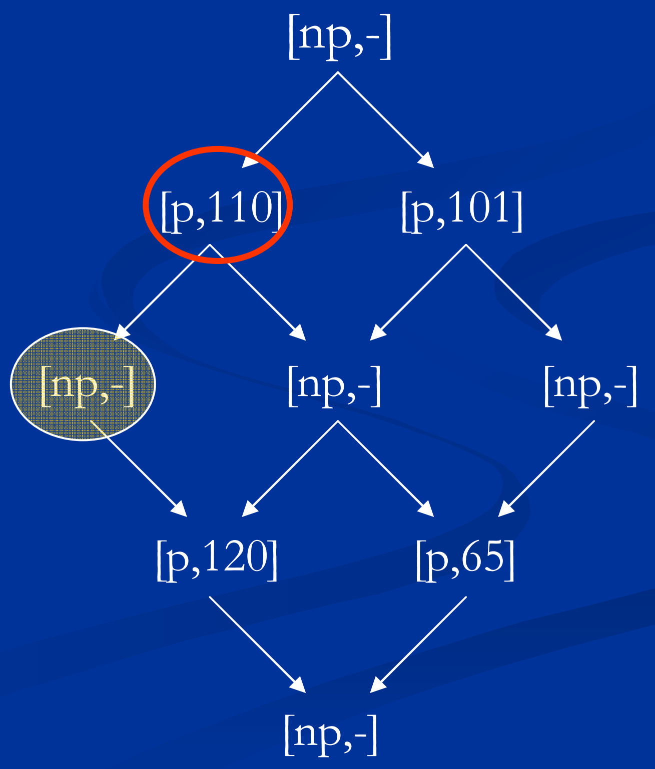
Hotel	Position	Price
Slumber Well	(⋆,Med)	120
Soporific Inn	(⋆⋆,Low)	65
Drowsy Hotel	(⋆⋆,High)	110
Celestial Sleep	(⋆⋆⋆,Med)	101
Nap Motel	(⋆⋆,Low)	101



Lattice Skyline (LS) Algorithm

- Compare each Lattice Element with Immediate Dominators in previous level.

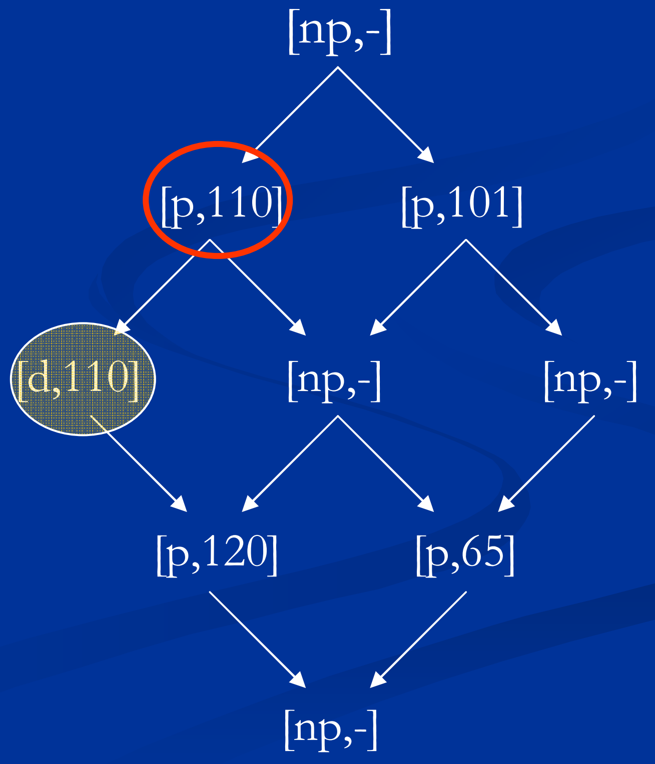
Hotel	Position	Price
Slumber Well	(⋆,Med)	120
Soporific Inn	(⋆⋆,Low)	65
Drowsy Hotel	(⋆⋆,High)	110
Celestial Sleep	(⋆⋆⋆,Med)	101
Nap Motel	(⋆⋆,Low)	101



Lattice Skyline (LS) Algorithm

- Compare each Lattice Element with Immediate Dominators in previous level.

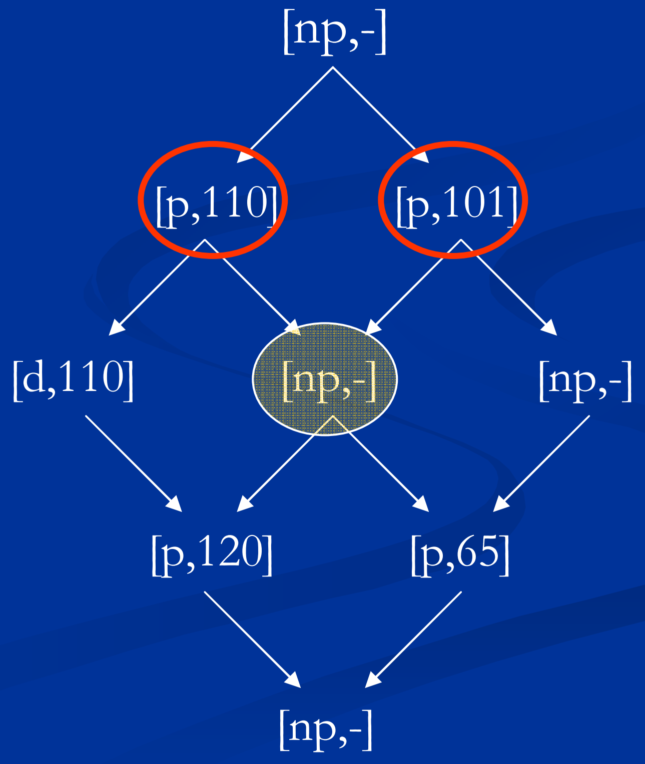
Hotel	Position	Price
Slumber Well	(⋆,Med)	120
Soporific Inn	(⋆⋆,Low)	65
Drowsy Hotel	(⋆⋆,High)	110
Celestial Sleep	(⋆⋆⋆,Med)	101
Nap Motel	(⋆⋆,Low)	101



Lattice Skyline (LS) Algorithm

- Compare each Lattice Element with Immediate Dominators in previous level.

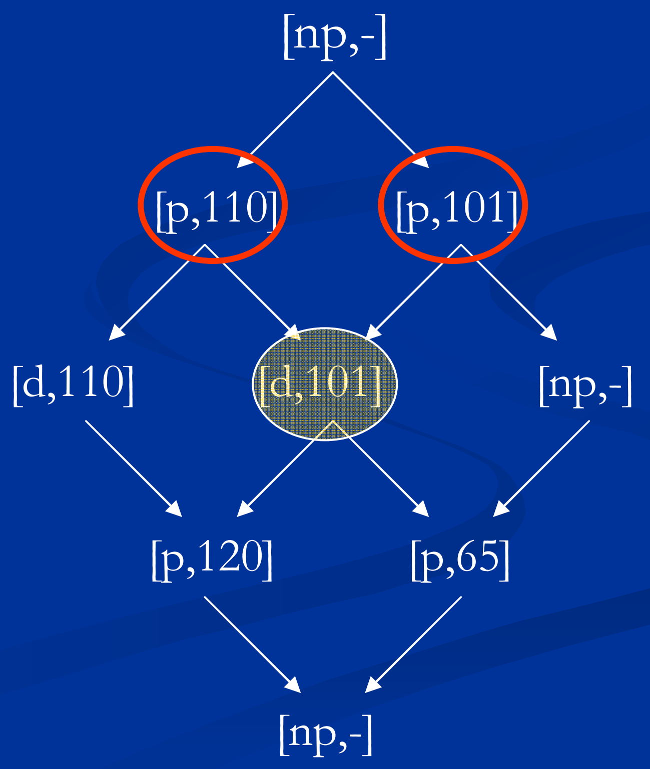
Hotel	Position	Price
Slumber Well	(⋆,Med)	120
Soporific Inn	(⋆⋆,Low)	65
Drowsy Hotel	(⋆⋆,High)	110
Celestial Sleep	(⋆⋆⋆,Med)	101
Nap Motel	(⋆⋆,Low)	101



Lattice Skyline (LS) Algorithm

- Compare each Lattice Element with Immediate Dominators in previous level.

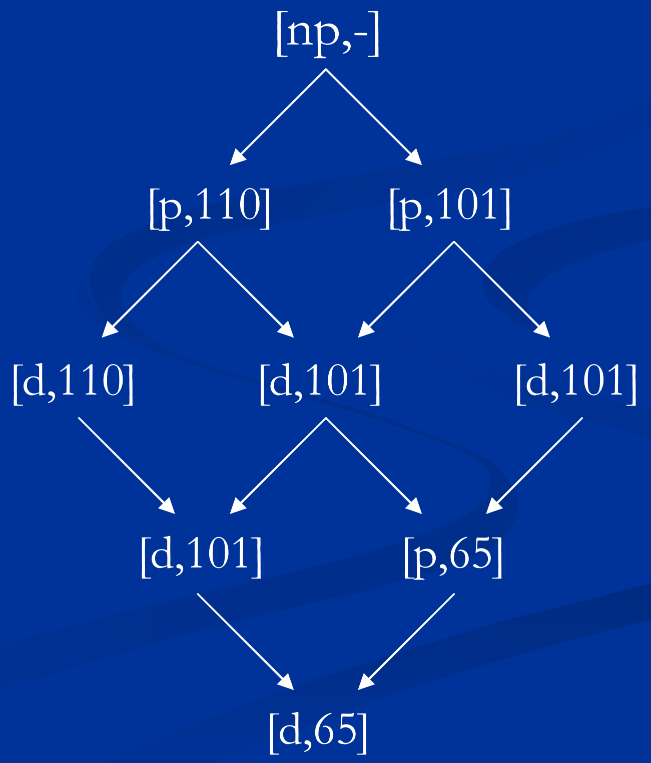
Hotel	Position	Price
Slumber Well	(⋆,Med)	120
Soporific Inn	(⋆⋆,Low)	65
Drowsy Hotel	(⋆⋆,High)	110
Celestial Sleep	(⋆⋆⋆,Med)	101
Nap Motel	(⋆⋆,Low)	101



Lattice Skyline (LS) Algorithm

- Compare each Lattice Element with Immediate Dominators in previous level.

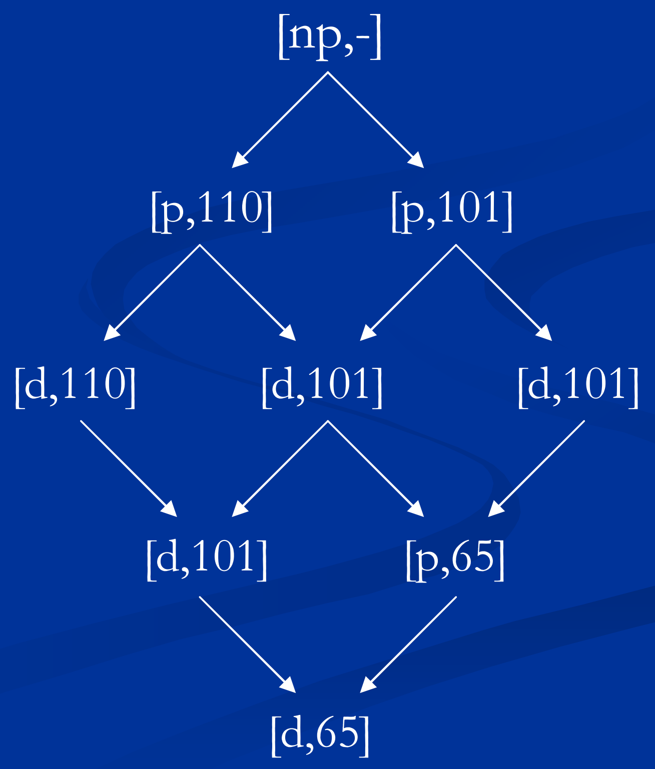
Hotel	Position	Price
Slumber Well	(⋆,Med)	120
Soporific Inn	(⋆⋆,Low)	65
Drowsy Hotel	(⋆⋆,High)	110
Celestial Sleep	(⋆⋆⋆,Med)	101
Nap Motel	(⋆⋆,Low)	101



Lattice Skyline (LS) Algorithm

- Compare each Lattice Element with Immediate Dominators in previous level.
- At this point, we know the skyline values present in the dataset.

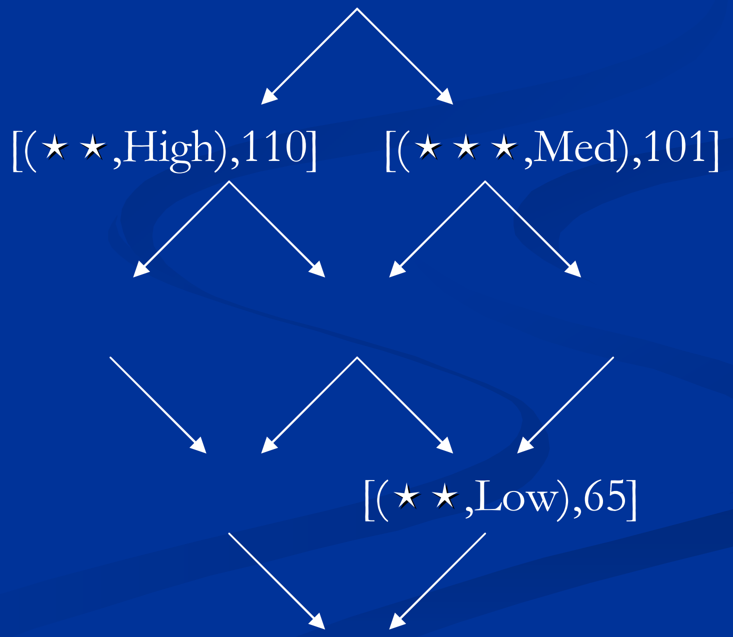
Hotel	Position	Price
Slumber Well	(⋆,Med)	120
Soporific Inn	(⋆⋆,Low)	65
Drowsy Hotel	(⋆⋆,High)	110
Celestial Sleep	(⋆⋆⋆,Med)	101
Nap Motel	(⋆⋆,Low)	101



Lattice Skyline (LS) Algorithm

- Iterate through the data.
- Output hotels matching the skyline values.

Hotel	Position	Price
Slumber Well	(⋆,Med)	120
Soporific Inn	(⋆⋆,Low)	65
Drowsy Hotel	(⋆⋆,High)	110
Celestial Sleep	(⋆⋆⋆,Med)	101
Nap Motel	(⋆⋆,Low)	101



Cost Analysis

- LS has 2 stages:

Complexity Analysis

- LS has 2 stages:
  - Iterating through the data and marking elements of the lattice [O(dn) cost].
    - d is the number of low cardinality dimensions
    - n is the number of tuples.

Complexity Analysis

- LS has 2 stages:
  - Iterating through the data and marking elements of the lattice [O(dn) cost].
    - d is the number of low cardinality dimensions
    - n is the number of tuples.
  - Finding skyline values in the lattice by examining the immediate dominators of each lattice position [O(dV) cost].
    - V is the domain cardinality product.
- This produces O(dn+dV) complexity.

Additional advantages

- The operation of LS does not vary with the input.
  1. Data ordering.
  2. Data distribution.
  - Additional advantage: Estimating running time is easy for an optimizer.

Overview

- Skyline Example and Definition.
- Discuss Low-Cardinality Attributes.
- Present the Lattice Skyline (LS) Algorithm.
- Discuss Experimental Results.
- Conclusions.

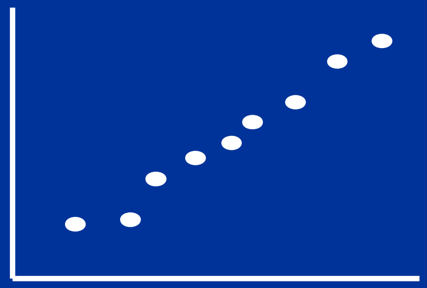
Experiments

- We tested LS against the best alternative technique LESS[1].
- We implemented LS and LESS with a 4KB page size and 500 buffer pool pages.
- 1.7 GHz Intel Xeon processor running Linux.
- Each tuple is a constant 100 bytes (includes some padding which models selection attributes such as a text attribute).
- We have run experiments on both synthetic and real datasets. Several of these results I will highlight here.

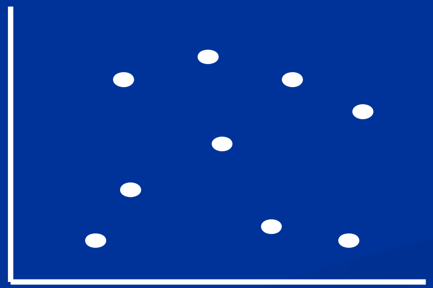
[1] [Godfrey et al. "Maximal Vector Computation in Large Datasets" VLDB 05]

Synthetic Datasets

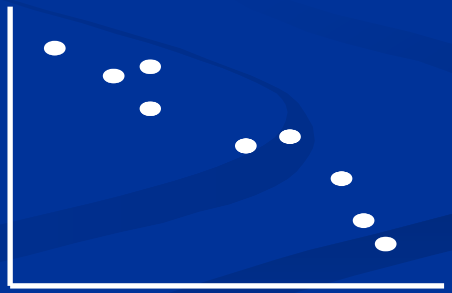
- Three synthetic datasets are commonly used in the evaluation of skyline techniques:
  - Correlated
  - Independent
  - Anti-correlated
- The anti-correlated dataset usually requires the most processing of the three.
- We vary the
  1. number of data tuples.
  2. Number of dimensions.
  3. Size of the low-cardinality domains.



Correlated



Independent

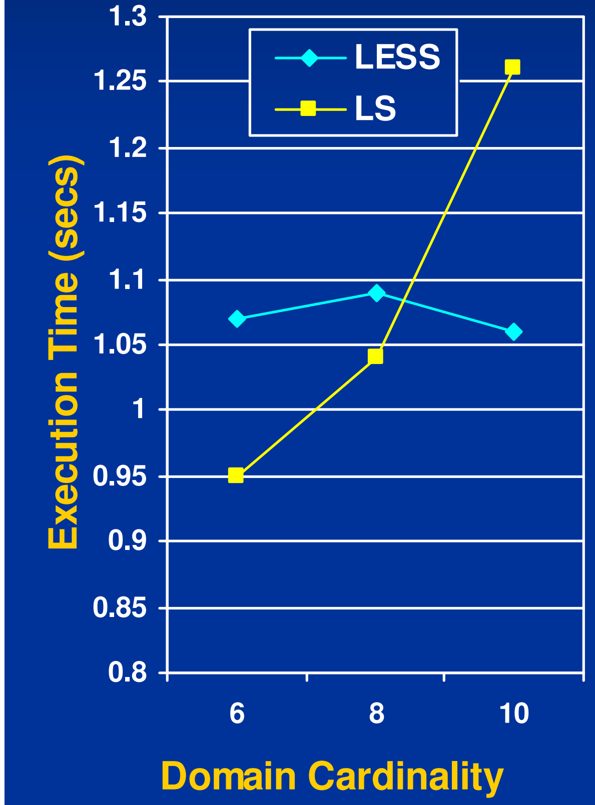


Anti-Correlated

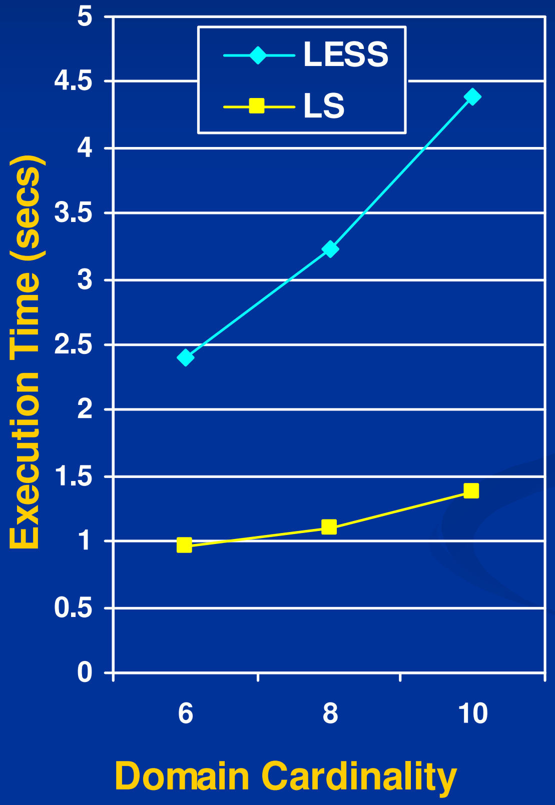
Real Dataset

- Zillow Housing Dataset: zillow.com lists information about real estate.
- We obtained a regional dataset with more than 160K entries with the below attributes.
- Low cardinality attributes include # of bedrooms, bathrooms, floors, and total rooms, and the garage capacity, with the estimated price as the unrestricted attribute.

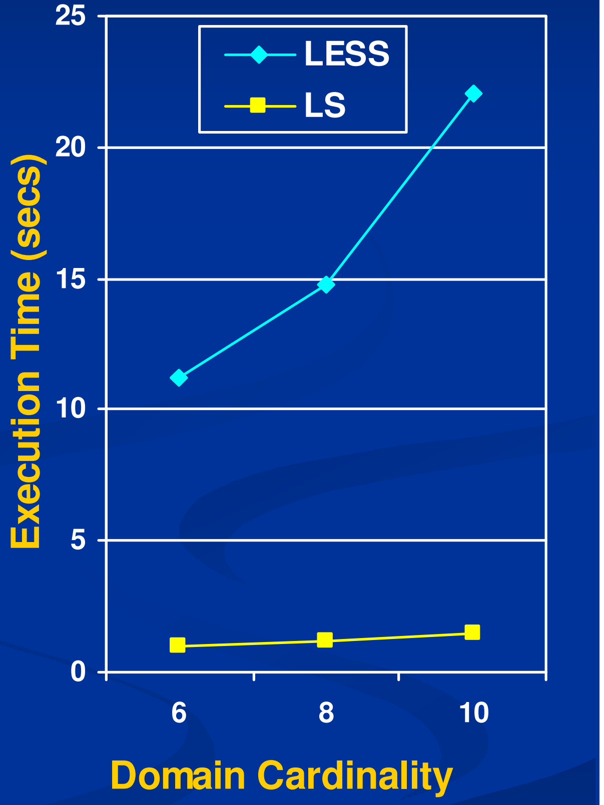
Results: Varying Domain Card.



Correlated



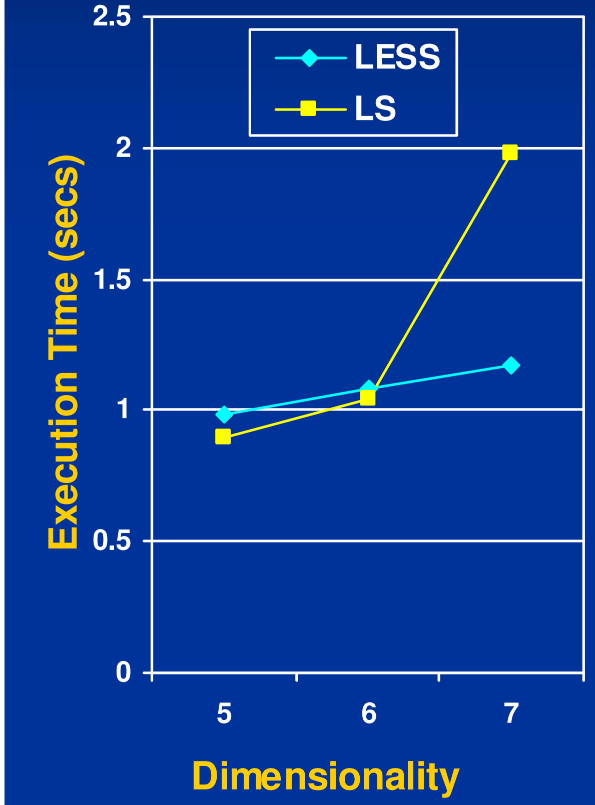
Independent



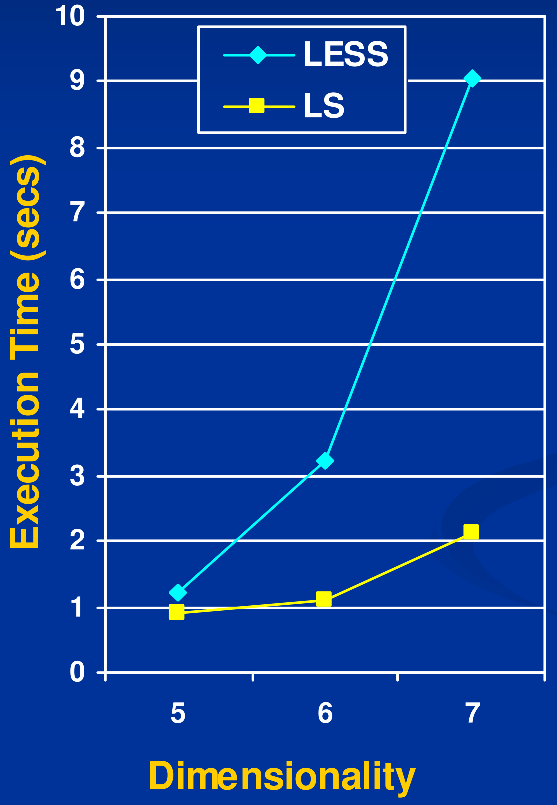
Anti-Correlated

Note: Graph Scales are not uniform.

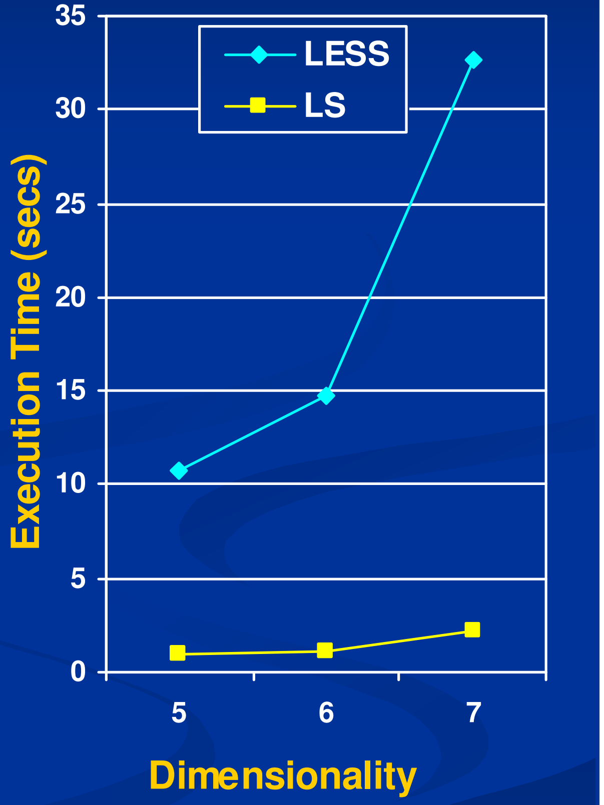
Results: Varying Dimensionality



Correlated



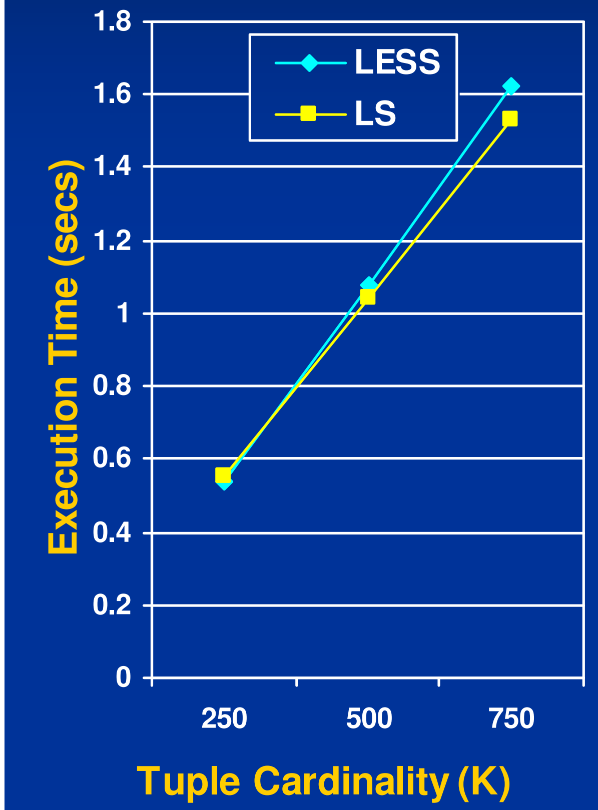
Independent



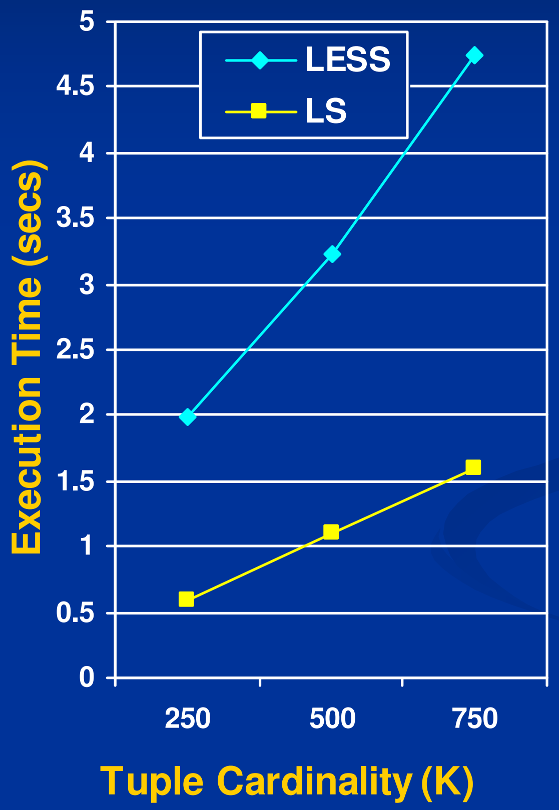
Anti-Correlated

Note: Graph Scales are not uniform.

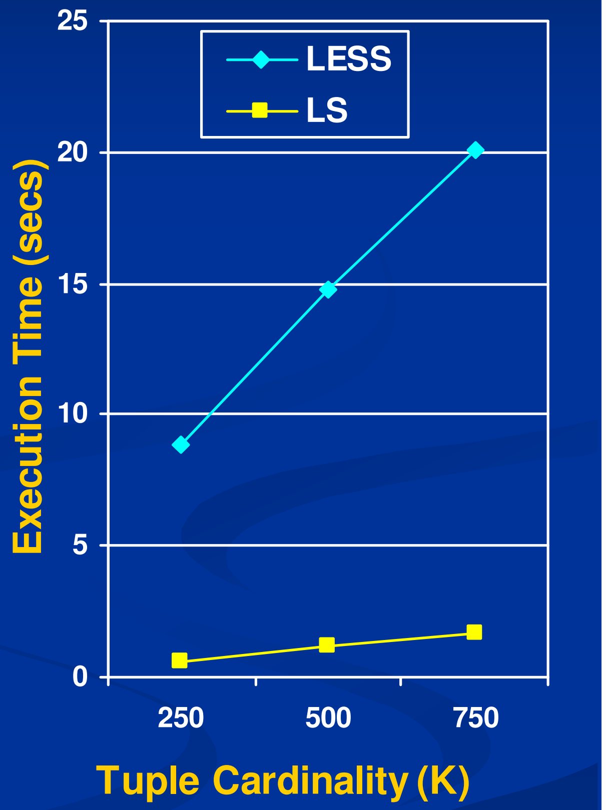
Results: Varying Tuple Card.



Correlated



Independent



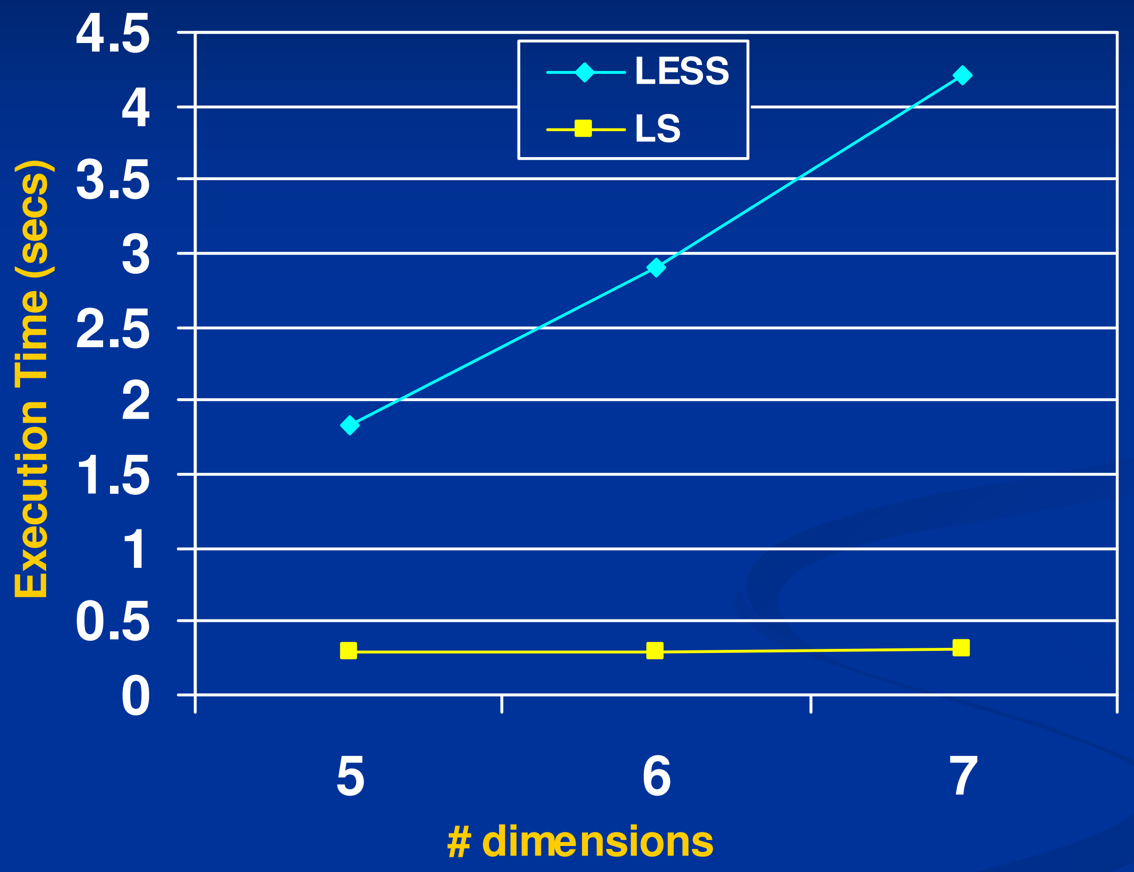
Anti-Correlated

Note: Graph Scales are not uniform.

Real Dataset

- Zillow Housing Dataset: zillow.com lists information about real estate.
- We obtained a regional dataset with more than 160K entries with the below attributes.
- Low cardinality attributes include # of bedrooms, bathrooms, floors, and total rooms, and the garage capacity, with the estimated price as the unrestricted attribute.
- Each tuple is a constant 100 bytes (usually includes some padding which models selection attributes such as a text attribute).

Results for Zillow Housing Dataset



Zillow Dataset

Overview

- Skyline Example and Definition.
- Discuss Low-Cardinality Attributes.
- Present the Lattice Skyline (LS) Algorithm.
- Discuss Experimental Results.
- Conclusions.

Conclusions

- We have proposed the Lattice Skyline Algorithm for skyline evaluation in the presence of datasets with low-cardinality attribute domains.
- The performance of the algorithm has been shown to be independent of dataset distribution and tuple ordering, both highly desirable properties for skyline evaluation.
- LS was shown to perform better than its nearest competitor, the LESS algorithm, in a number of synthetic and real dataset experiments.

Thank You!

Questions?

Back Up Slides

Real Dataset

- Zillow Housing Dataset: zillow.com lists information about real estate.
- We obtained a regional dataset with more than 160K entries with the below attributes.

Description	Values	Domain Cardinality
# of Bedrooms	Integer	7
# of Bathrooms	½ Increments	4
# of Floors	Integer	3
# of Rooms	Integer	10
Garage Capacity	Integer no. of cars	7
Asphalt Roof	Yes or No	2
Colonial Arch	Yes or No	2
Estimated Price	Dollar Value	Nearly 80K values