Query Processing and Advanced Queries

Advanced Queries (3): Skyline Query

Motivation

Suppose we want to look for a vacation package

3 packages

We want to have a cheaper package.

We want to have a higher hotel-class.

Package ID	Price	Hotel-class
a	1600	4
b	2400	1
c	3000	5

Package a "**dominates**" package b

We know that
1. Package a has a cheaper price
2. Package a has a higher hotel-class

skyline

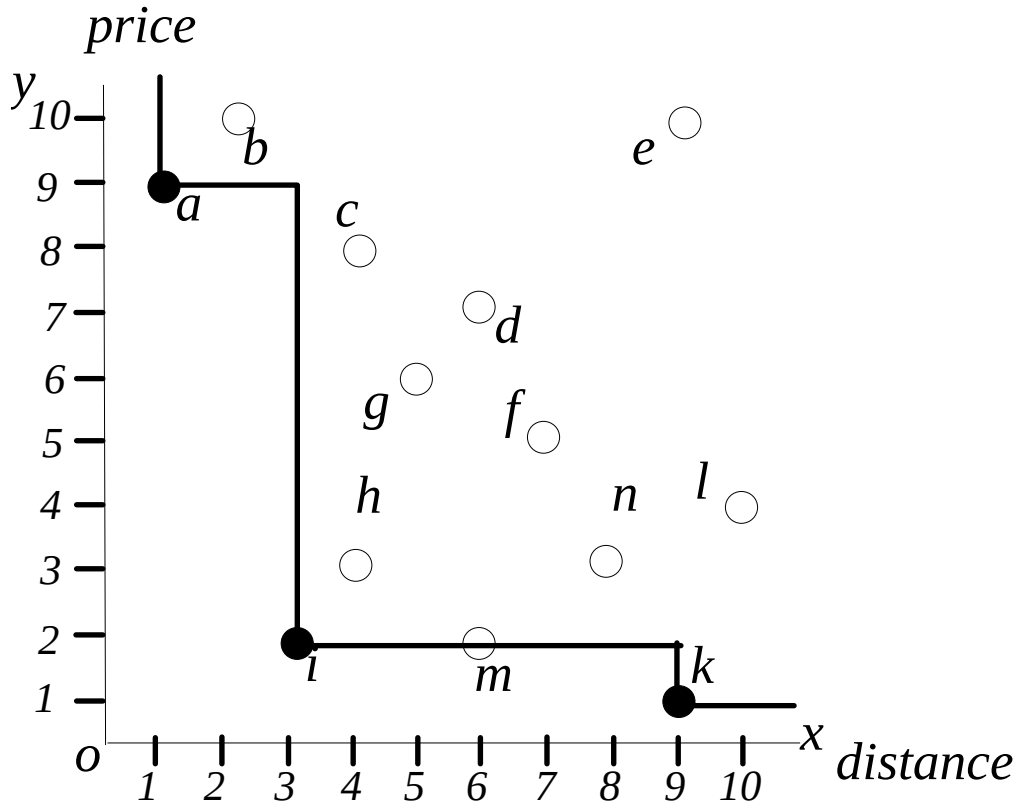
We want to find a set of packages which are NOT dominated by any other pacakges

All of the "best" possible choices.
i.e., {a, c}

Skyline Operator

- A new operator (like “ORDER BY”) in database systems.
- Skyline Points
  - A set of data points that are not dominated by any other data points.

Skyline Queries



- Retrieve points **not dominated** by any other point:
  - A point p *dominates* another point q if p is as good or better as q in all dimensions and better in at least one dimension.

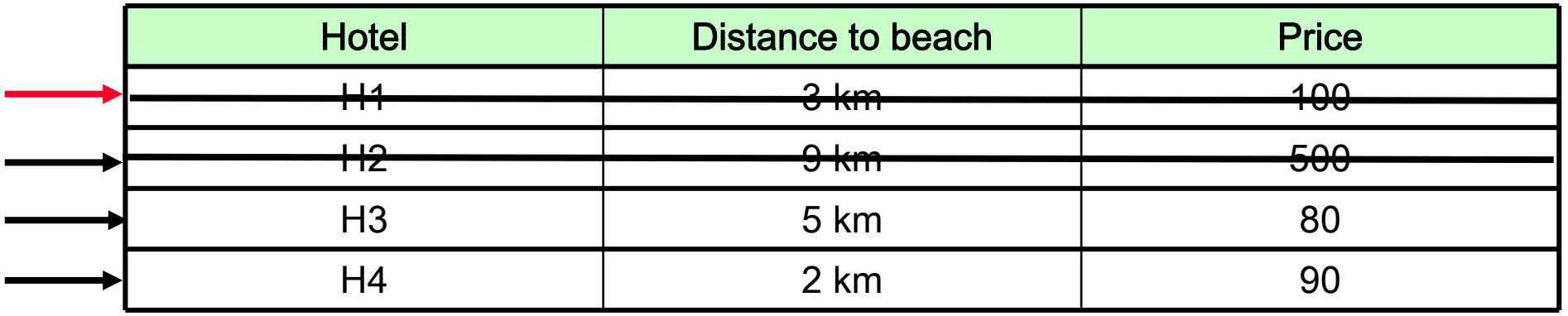
Skyline of Manhattan

- Which buildings can we see?
  - *Higher or nearer*

    (a building dominates another building if it is higher, closer to the river, and has the same *x* position)

A Naïve Algorithm

- For each point p
  - Check any other points in the dataset;
  - If p is dominated by at least one point q, p cannot be in the output;
  - If p dominates one point q, q cannot be in the output;
- Return all the skyline points



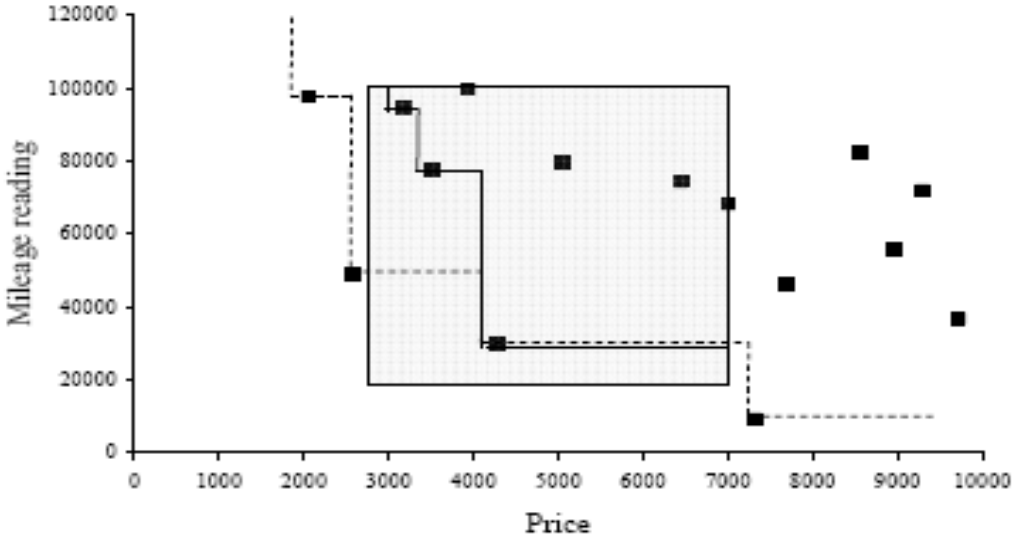
Hotel	Distance to beach	Price
~~H1~~	~~3 km~~	~~100~~
~~H2~~	~~9 km~~	~~500~~
H3	5 km	80
H4	2 km	90

Pruning using R-tree

- To use an R-tree to compute the Skyline of cheap hotels near the beach, we exploit the following fact:
  - Given a hotel *h*, we need not search in any branches of the R-tree which are guaranteed to contain only hotels that are dominated by *h*.
- For example, if we know that there is a hotel that costs $30 and is located 1.0 miles from the beach, then we need not consider any branches in the R-tree which include, for instance, hotels in the price range of ($40, $60) and distance range of (2.0 miles, 3.5 miles).
- As a consequence, the idea is to traverse the R-tree in a depth first way and *prune* branches of the R-tree with every new hotel found.

Extensions

- Constrained Skyline (car database):
  - A user may only be interested in records within the price range from 3 thousand to 7 thousand dollars and with mileage reading between 20K and 100K.



- The traditional skyline (dashed line) fails to return interesting points.

Extensions (cont.)

- Subspace Skyline:
  - A car database could contain many other attributes of the cars:
    - horsepower, age, fuel consumption, etc…
  - A customer that is sensitive on the price and the mileage reading (2-dimensional subspace) would like to pose a skyline query on those attributes, rather than on the whole data space.

- While the dimensionality of the corresponding data space might be rather high, skyline queries generally refer to a *low dimensional* subspace.
- The *constrained subspace skyline queries* form the generalization of all meaningful skyline queries over a given dataset.