Interactive Data Exploration Using Semantic Windows
Alexander Kalinin, Ugur Cetintemel, Stan Zdonik
Brown University

Overview
- New data exploration framework – Semantic Windows
- Users express exploration goals in terms of “windows of interest”.
- Windows of interest are expressed as a number of conditions on shape (e.g., 2x3) and content (e.g., \( \text{avg(brightness)} > 0.8 \))
- The primary focus is on incremental results.

- User-friendly SQL extensions for more efficient exploration
- On-line algorithm providing interactive results
- Adaptable prefetching technique to deal with “online vs. total results” trade-off
- Different data back-ends: PostgreSQL, SciDB

Searching for a 2D region
- Region of shape 6 by 5
- Average brightness > 0.8

Searching for a time interval
- Time interval of 1 to 3 years
- Average stock price > 50

Window-based search
Looking for: SIZE < 5, SUM > 10

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

Online results: 5 4 6 5 And so on...

Algorithm Overview
Priority Queue (utility-ordered)

```
1 0.98 0.80 0.79
3 2 1 4
```

```
1 0.98 0.85 0.80 0.79
3 4 2 1 3 1 4
```

```
0.98 0.85 0.80 0.80 0.79
2 1 3 1 2
```

```
0.98 0.85 0.80 0.80 0.79
3 4
```

Utility = f(benefit, cost)

Distance to goals
AVG > 200, Value = 100?
8 = 1 – \( \frac{200 – 100}{500} \) = 0.8

Cost of disk read
Number of cells adjusted for skew – cached cells

Online Results Times

Optimizations
- Prefetching. Deals with the “online vs. total results” trade-off. Increases locality to reduce the number of disk seeks, but delays online results.
- Pruning. Allows to get rid of parts of the graph that cannot contain results. Based on user-specified criteria.
- Sampling. Allows to estimate utilities without reading the exact data from disk. Guides the search process.

Ongoing Work
- “Low-resolution” picture, showing results across all search space at the same pace. Show negative results along with positive!
- Approximate answering. Outputting windows with condition values and confidence intervals.
- Parallel search.