**MADALGO**
CENTER FOR MASSIVE DATA ALGORITHMICS

Zhewei Wei
Aarhus University

Danmarks Grundforskningsfond
Danish National Research Foundation

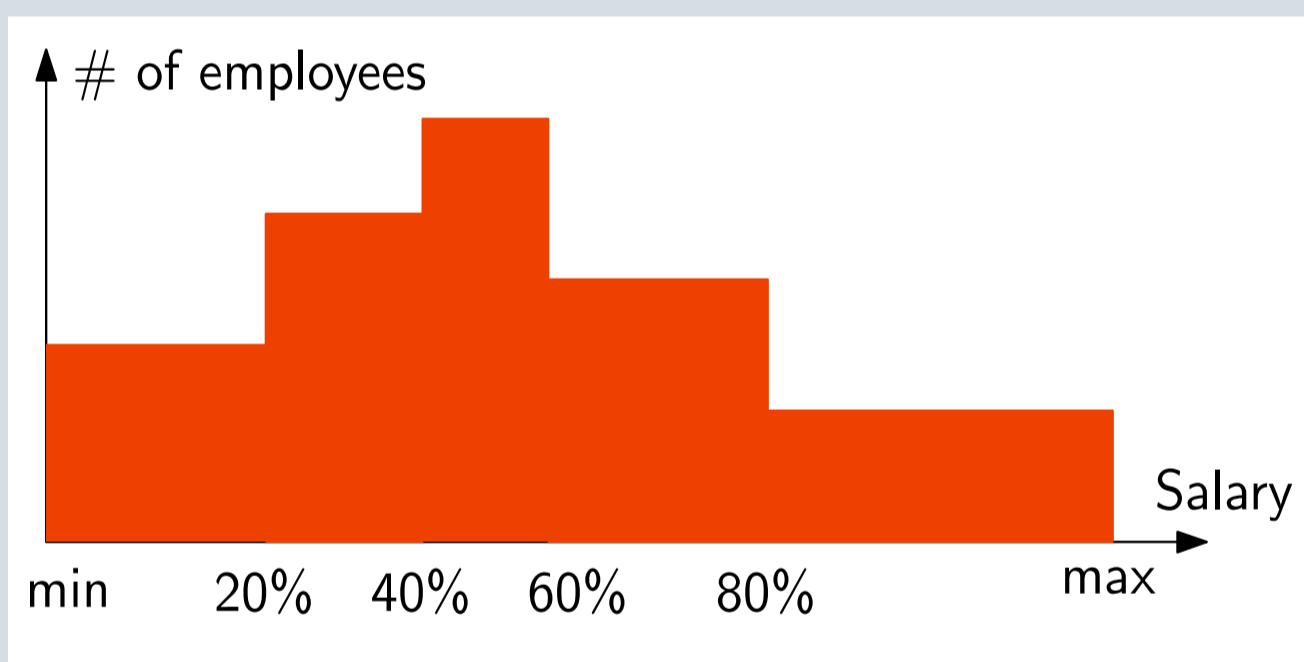# Indexing for Summary Queries: Theory and Practice

## Introduction

- Traditional database queries: Aggregation vs. Reporting

```
SELECT T.salary          SELECT AVG(T.salary)
FROM Table T             FROM Table T
WHERE 30 < T.age < 40    WHERE 30 < T.age < 40
```

$$\left.\begin{array}{l} \$32,000 \\ \$76,300 \\ \$54,400 \\ \ldots \\ \$68,000 \\ \$28,000 \end{array}\right\} 50,000 \text{ records} \qquad \$52,312$$

**Motivation**: Aggregation is fast, but reporting is more expressive [1].

Best of both world?

- (Q1) In a company's database: What is the distribution of salaries of all employees aged between 30 and 40?
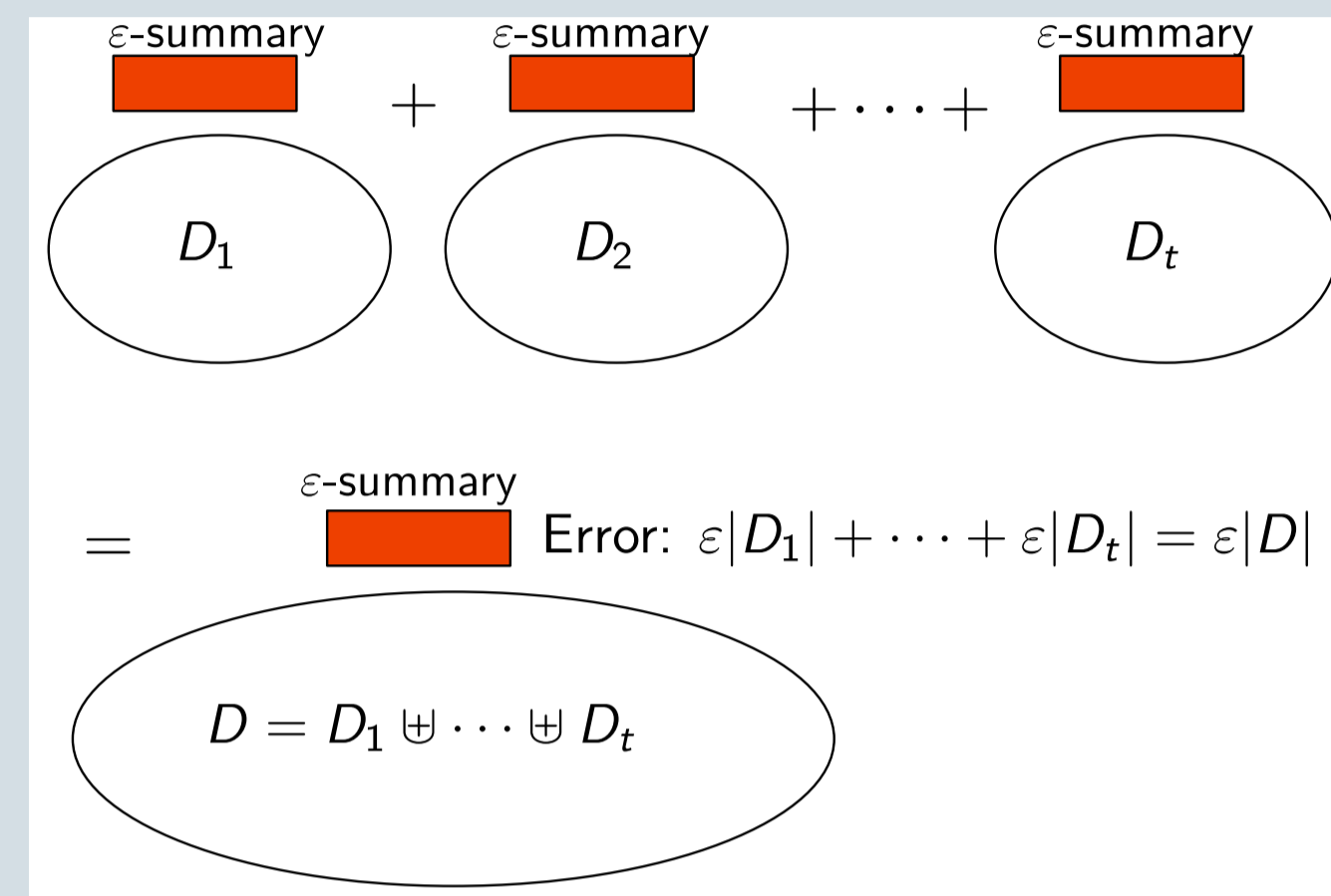


- (Q2) In a search engine's query logs: What are the most frequently queried keywords between March 11 and April 7, 2011?
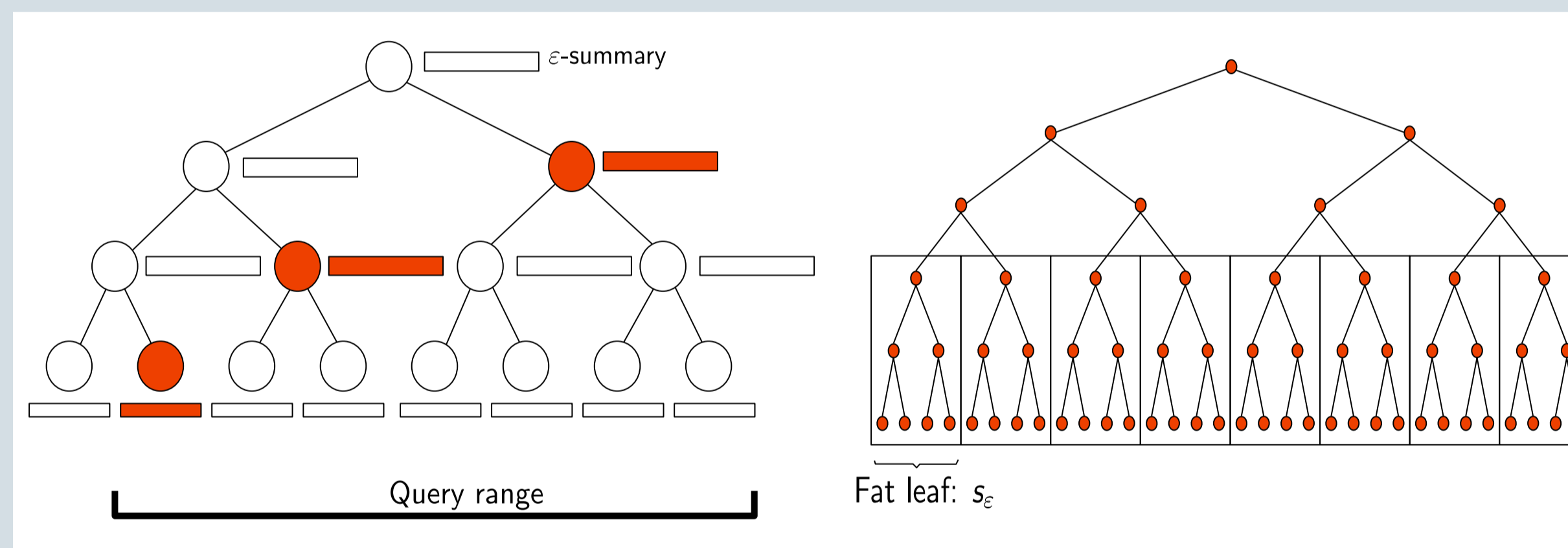
**Search Engine Log**

| Date | Keyword |
|------|---------|
| 2011.04.08 | Masters 2011 |
| 2011.04.08 | Libya |
| 2011.04.07 | Japan nuclear crisis |
| 2011.04.07 | Libya |
| ... | |
| 2011.03.11 | Japan earthquake |
| 2011.03.11 | Japan tsunami |
| 2011.03.10 | NCAA |
| ... | |

| Keyword | Frequency |
|---------|-----------|
| Libya | 19.3% |
| Japan nuclear crisis | 16.5% |
| Japan earthquake | 10.2% |
| ... | |

## Data Structure

- Decomposable summaries



$$\text{Error: } \varepsilon|D_1| + \cdots + \varepsilon|D_t| = \varepsilon|D|$$
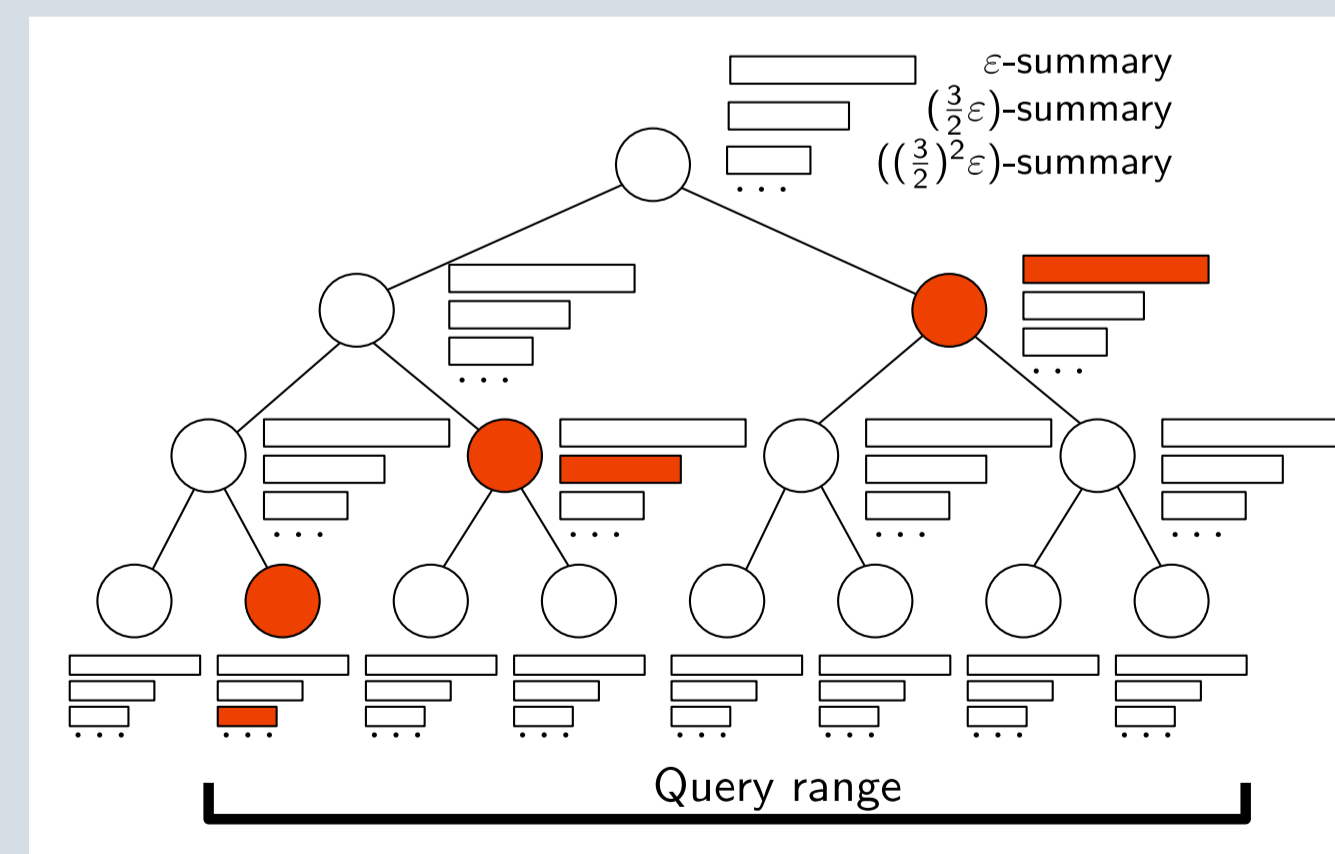
$$D = D_1 \uplus \cdots \uplus D_t$$

- Baseline structure



- Optimal structure



- Theoretical bounds:

Space: $O(N)$
Query cost: $O(\log N + s_\varepsilon)$ for internal memory
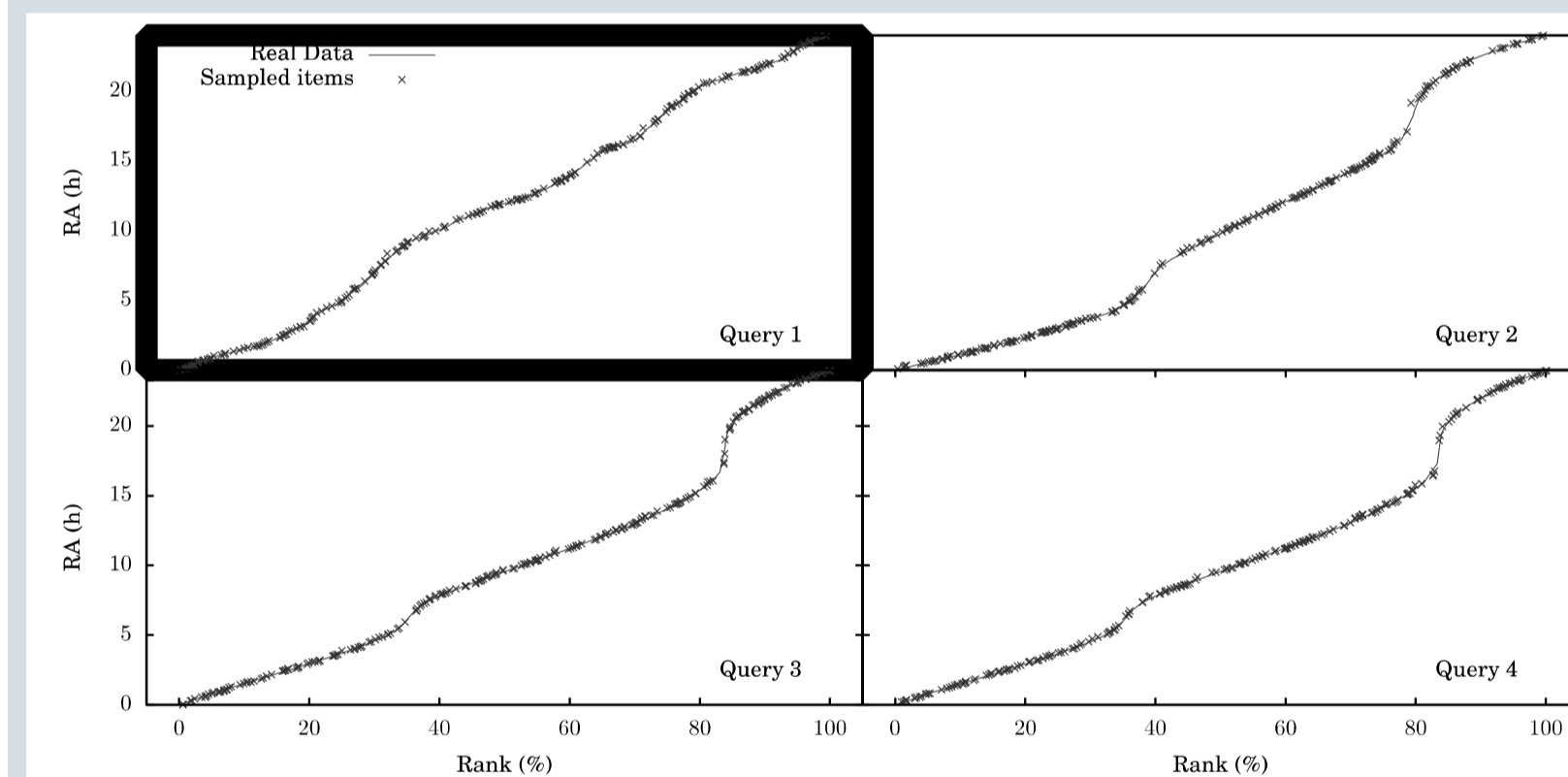$O(\log_B N + s_\varepsilon/B)$ I/Os for external memory

## Experiment

- Data set: MPCAT-OBS, observational sets for numbered and unnumbered minor planets and comets.

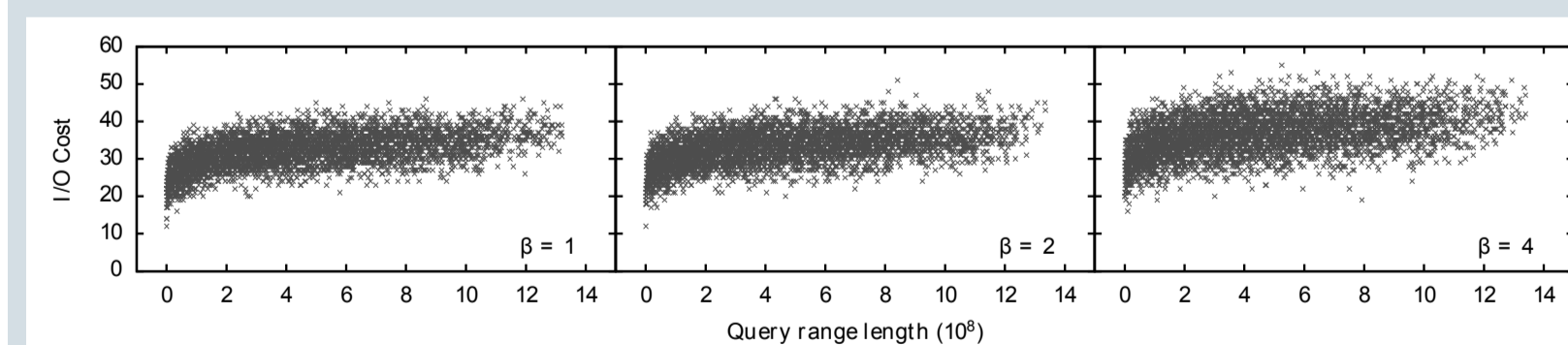- Date is the key attribute and observatory is the summary attribute.

- Query examples

| | Table II. Queries | | |
|---|---|---|---|
| # | Start Date | End Date | Length |
| Q1 | 1800.01.01 | 1900.01.01 | 9,993 |
| Q2 | 1900.01.01 | 2000.01.01 | 4,827,585 |
| Q3 | 2000.01.01 | 2100.01.01 | 82,850,545 |
| Q4 | 1800.01.01 | 2100.01.01 | 87,688,123 |

- Query results



- Query cost



## References

[1] P. K. Agarwal and J. Erickson. *Geometric range searching and its relatives*. In Advances in Discrete and Computational Geometry. American Mathematical Society, 1–56,1999.
[2] L. Wang, Z. Wei and K. Yi. *Indexing for summary queries: theory and practice*. Submitted to ACM Transactions on Database Systems (TODS).