Lab 2 – Cardinality Estimation

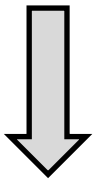
- Quick review query optimization and cardinality estimation
- Discuss how to estimate the cardinality for one column and multiple columns
- Try to estimate cardinality for multiple columns

Review: Query Optimization

- Main goal: generate efficient execution approaches for queries
- Among all the optimization goals, finding good join-orders for product plans is one of the most important goals in query optimization

Join-Orders Matter

```
SELECT * FROM a, b WHERE 5 < a.x < 9 AND 20 < b.y < 2000 AND a.z = b.z
```



nested-loop join

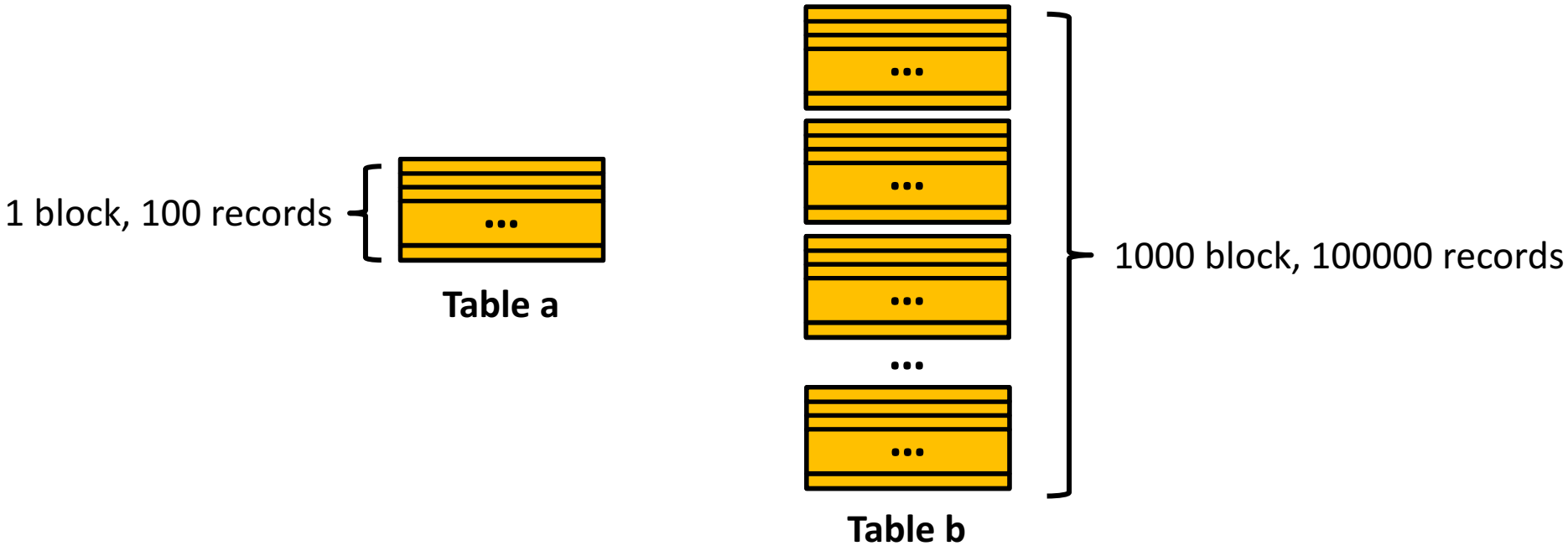
```
for outer_record in outer_table:
        for inner_record in inner_table:
                print(outer_record + inner record)
```

Option 1: a as outer table, b as inner table

Option 2: b as outer table, a as inner table

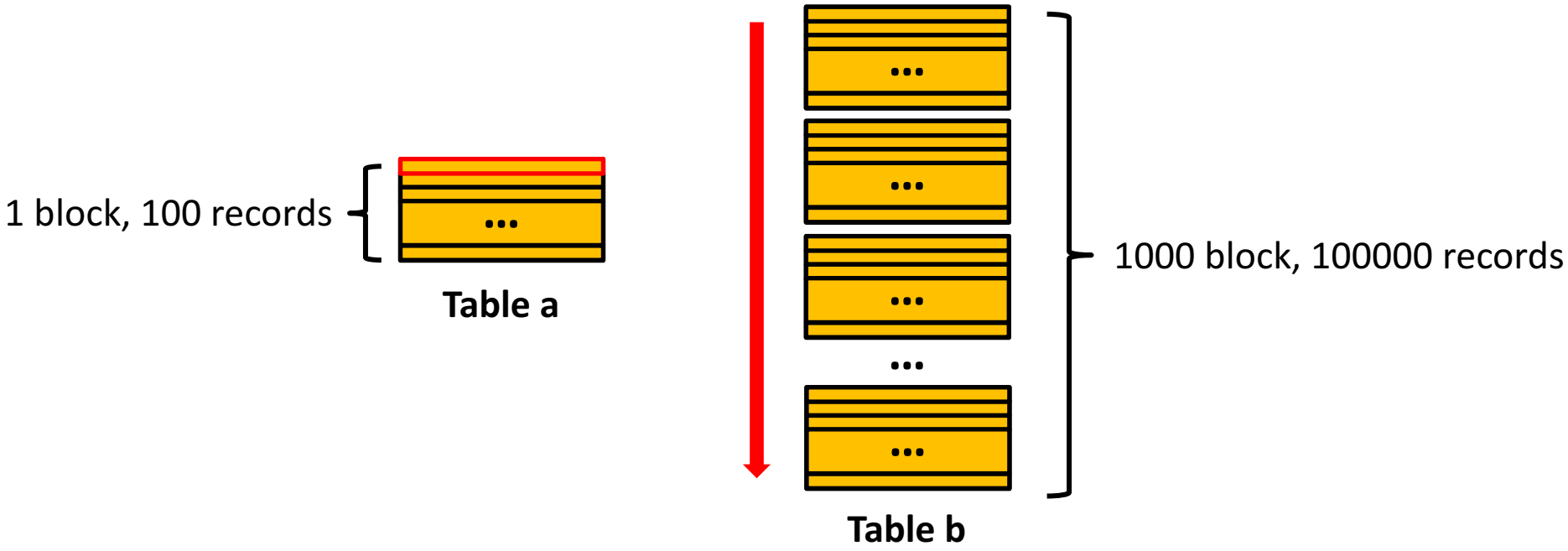
A as Outer Table, B as Inner Table

Assume the DBMS currently has 100 available buffers



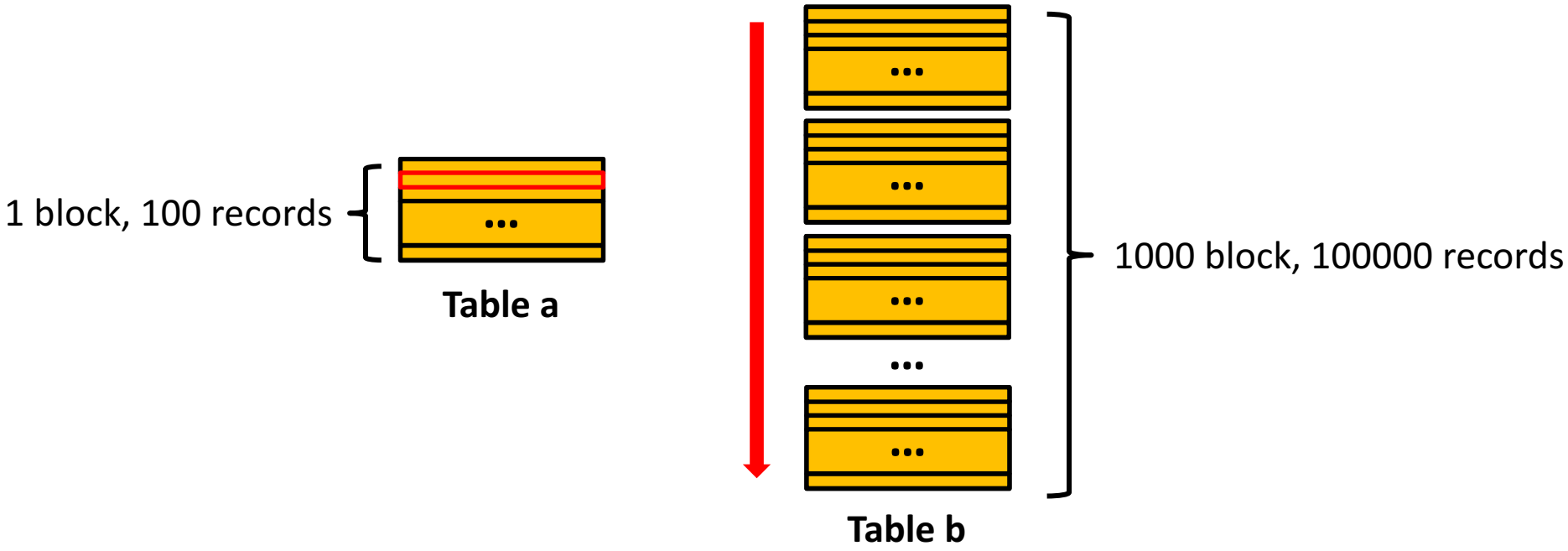
A as Outer Table, B as Inner Table

Assume the DBMS currently has 100 available buffers



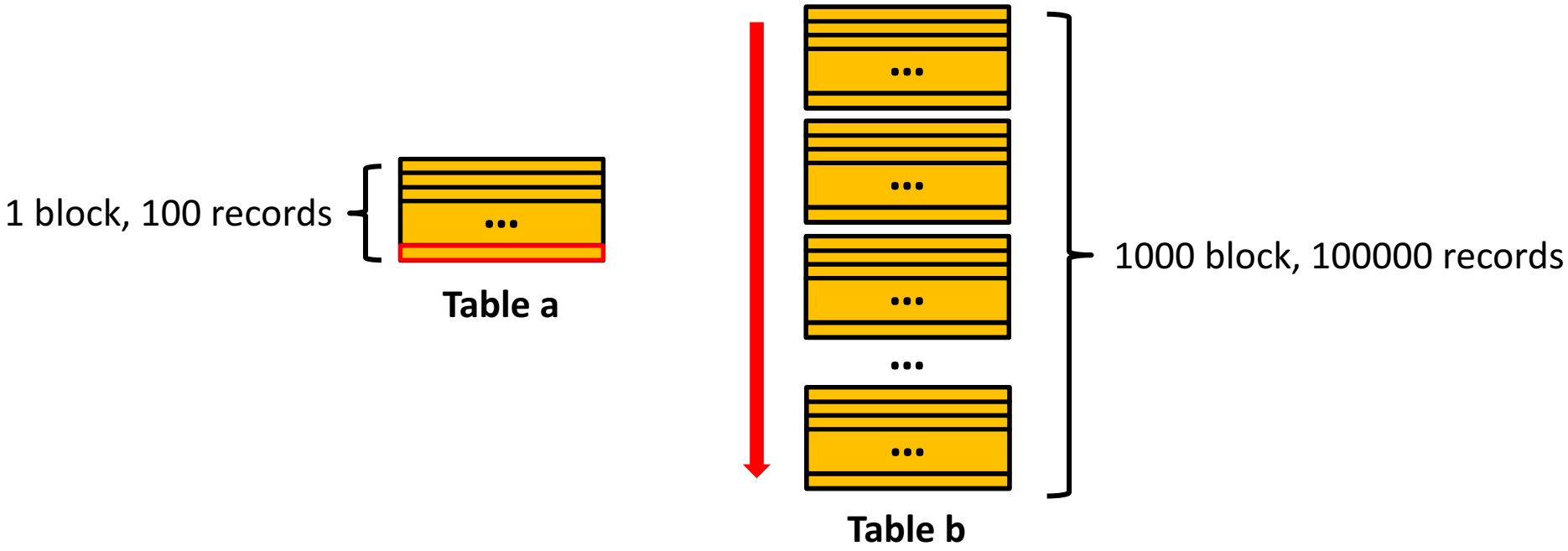
A as Outer Table, B as Inner Table

Assume the DBMS currently has 100 available buffers



A as Outer Table, B as Inner Table

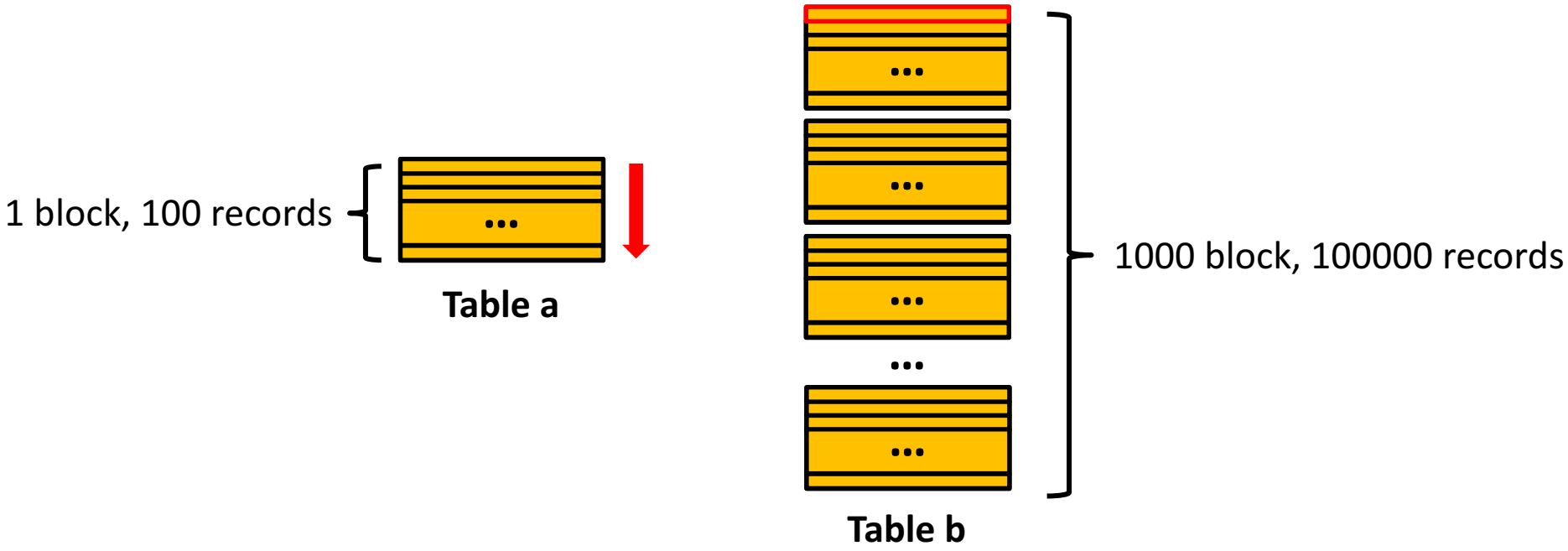
Assume the DBMS currently has 100 available buffers



Action	Cost (# of blocks accessed)
scan table a	1
scan table b	100 * 1000 = 100000

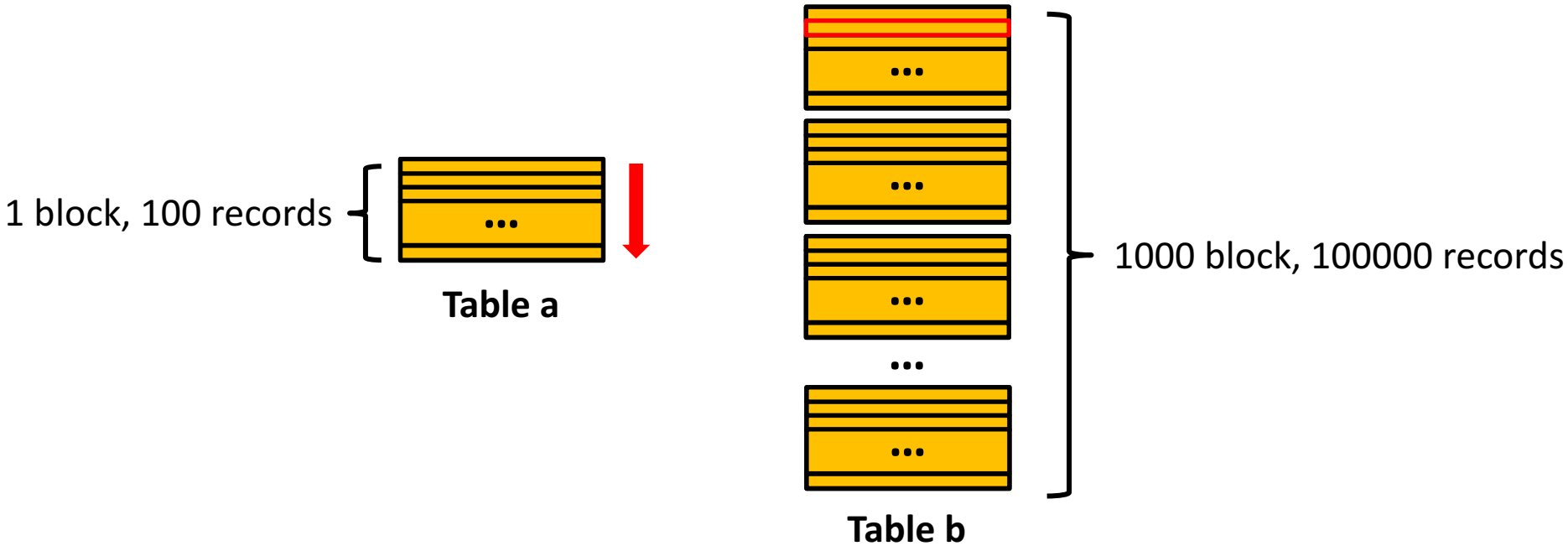
B as Outer Table, A as Inner Table

Assume the DBMS currently has 100 available buffers



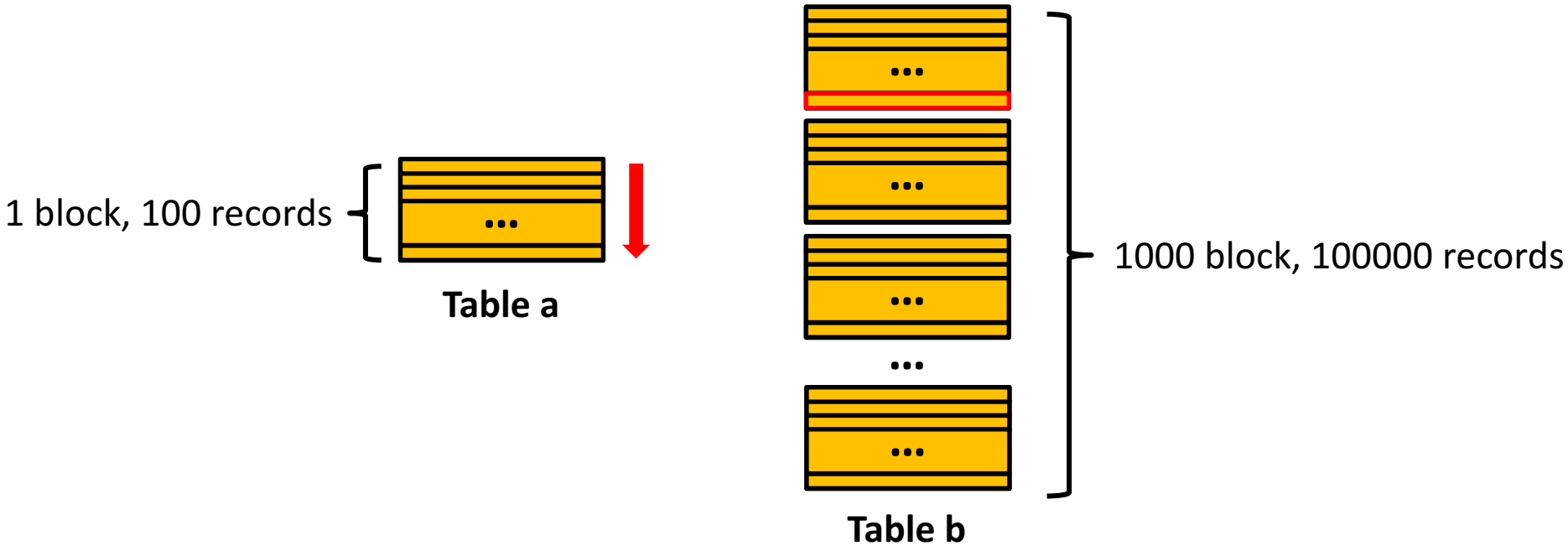
B as Outer Table, A as Inner Table

Assume the DBMS currently has 100 available buffers



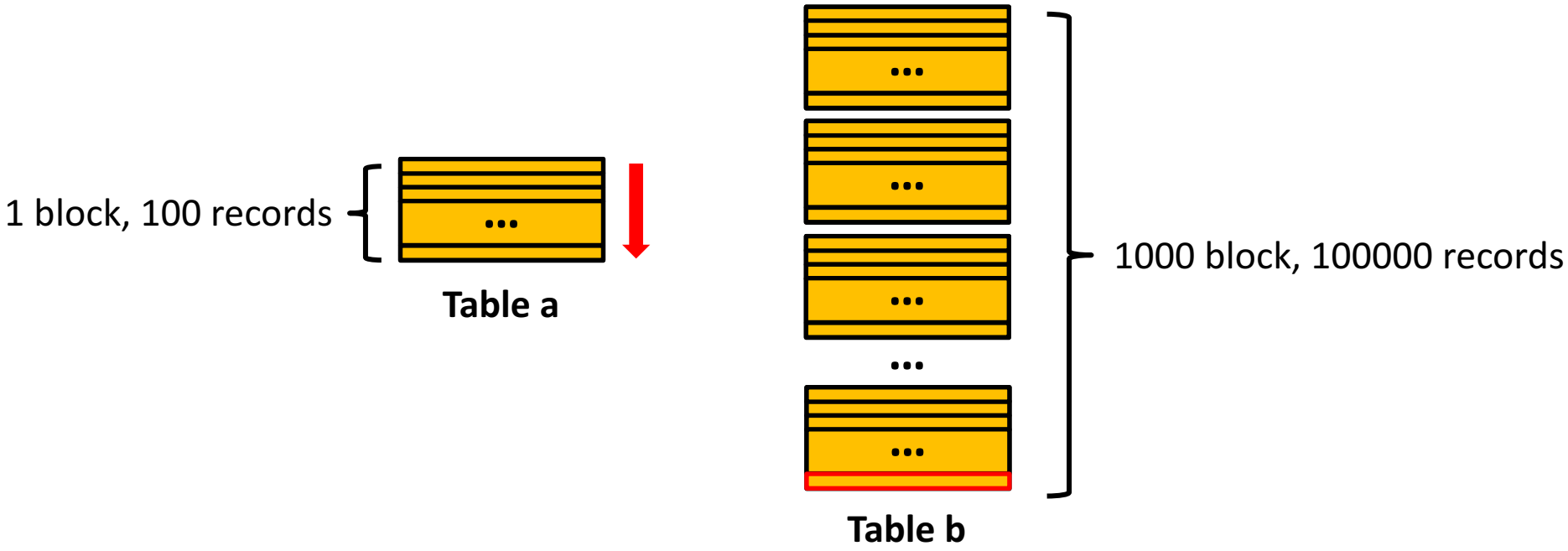
B as Outer Table, A as Inner Table

Assume the DBMS currently has 100 available buffers



B as Outer Table, A as Inner Table

Assume the DBMS currently has 100 available buffers



Action	Cost (# of blocks accessed)
scan table a	1
scan table b	1000

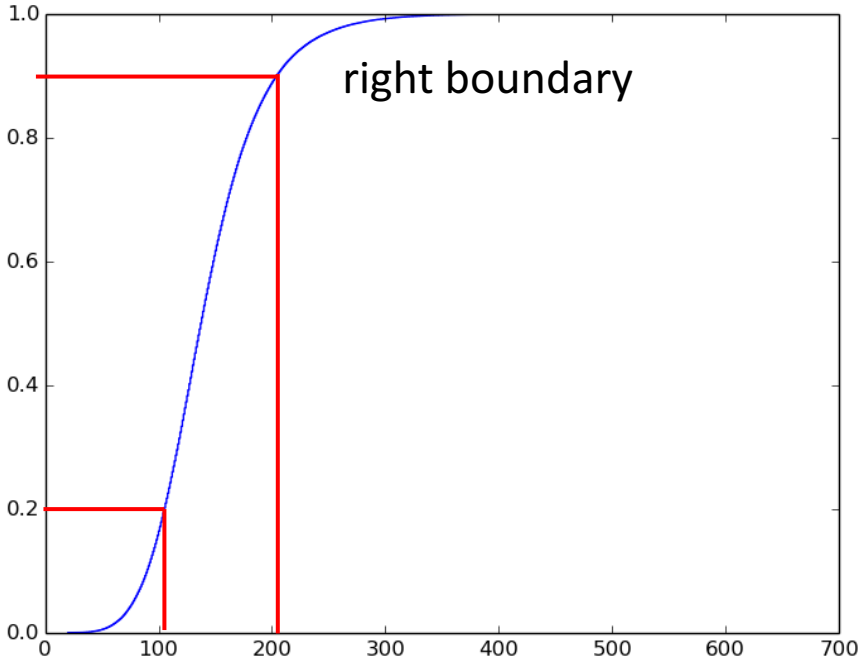
Cardinality Estimation

- How to estimate cardinality?
  - $cardinality = \#\ of\ total\ records\ * seletivity$
- How to estimate selectivity?
  - a naïve approach: assume uniform distribution
  - a better approach: histogram
  - how about machine learning?

```
SELECT * FROM a WHERE 110 < x < 210
```

One-Column Selectivity Estimation

- Learn CDF




```
SELECT * FROM a WHERE 110 < x < 210
```

Multi-Column Selectivity Estimation

- Assume columns are independent
  - estimate the selectivity of each column and obtains the product of them
- Correlation between columns

```
SELECT * FROM population WHERE age < 45 AND salary < 45000
```

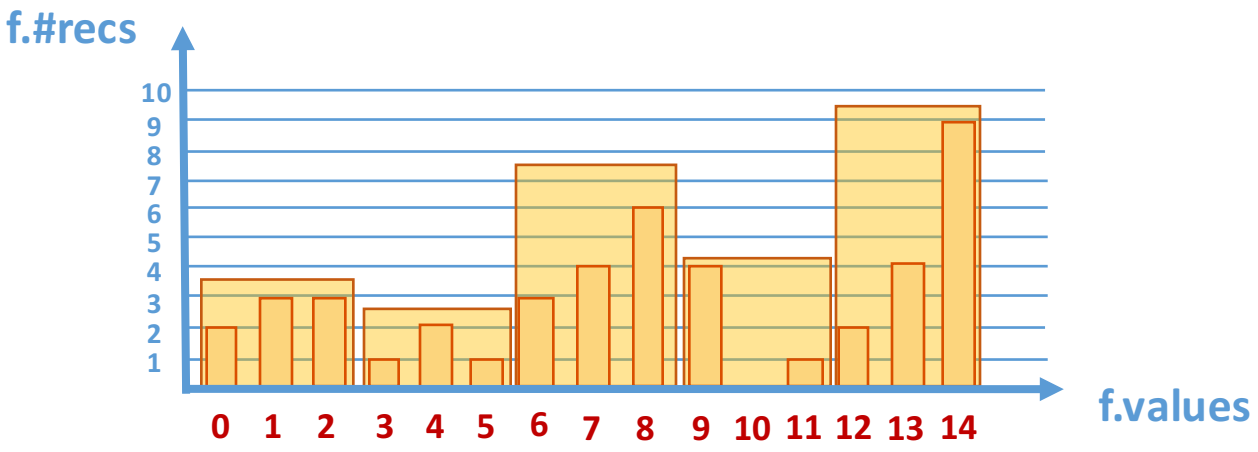
Multi-Column Selectivity Estimation

- Assume
  - half of the population is order than 45
  - half of the population has salary more than 45K
- Under the independent-column assumption
  - $sel = 0.5 * 0.5 = 0.25$
  - however, the actual selectivity may be far less than that

```
SELECT * FROM population WHERE age < 45 AND salary > 45000
```

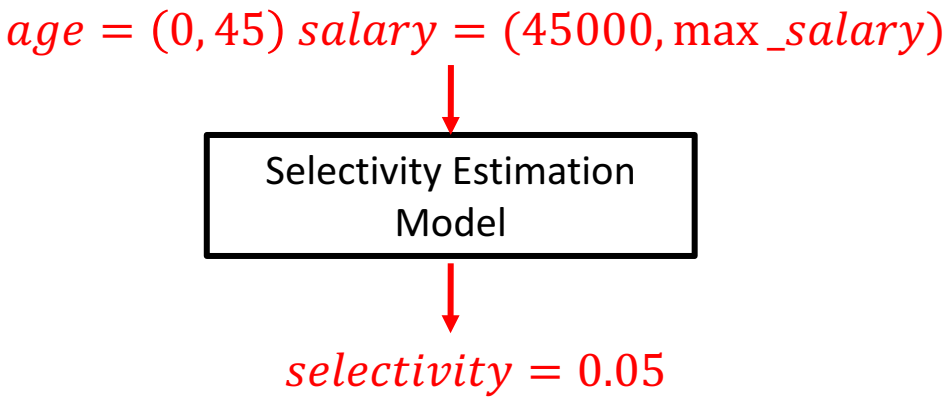
Multi-Column Selectivity Estimation

- How about high-dimensional histogram
  - high storage cost (exponential to dimensionality)
- Use machine learning model to estimate selectivity



Problem Formulation

- Input
  - *k* fields, *2k* boundaries
- Output
  - selectivity




```
SELECT * FROM population
WHERE 0 < age < 45 AND 45000 < salary < max_salary
```

https://github.com/yschang1206/dbai-labs