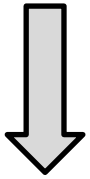
Cardinality Estimation Using Neural Networks

CASCON '15

Henry Liu	University of Waterloo, Waterloo, Canada
Mingbin Xu	York University, Toronto, Canada
Ziting Yu	University of Waterloo, Waterloo, Canada
Vincent Corvinelli	IBM Canada Ltd., Markham, Canada
Calisto Zuzarte	IBM Canada Ltd., Markham, Canada

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

Background

Terminologies in paper title:

Cardinality Estimation Using Neural Networks

Cardinality = # of total records * selectivity

name	age
J	3
A	5
P	3
X	4
C	5
I	4

Eg:

Select name from tb1 where age>3

# of total records: 6

Selectivity:0.67
(4/6)

Cardinality:4

Background

- Cardinality = # of total records * selectivity

How to estimate selectivity?

–a naïve approach: assume uniform distribution

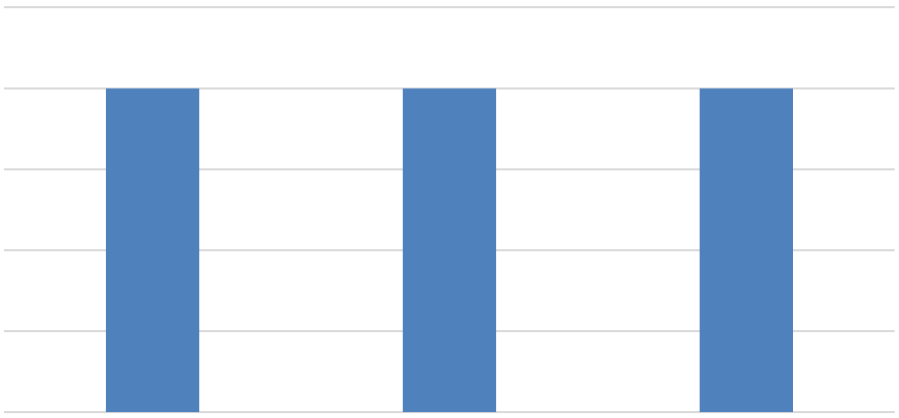
–a better approach: histogram

Background

Cardinality = # of total records * selectivity

- assume uniform distribution

age
3
5
3
4
5
4

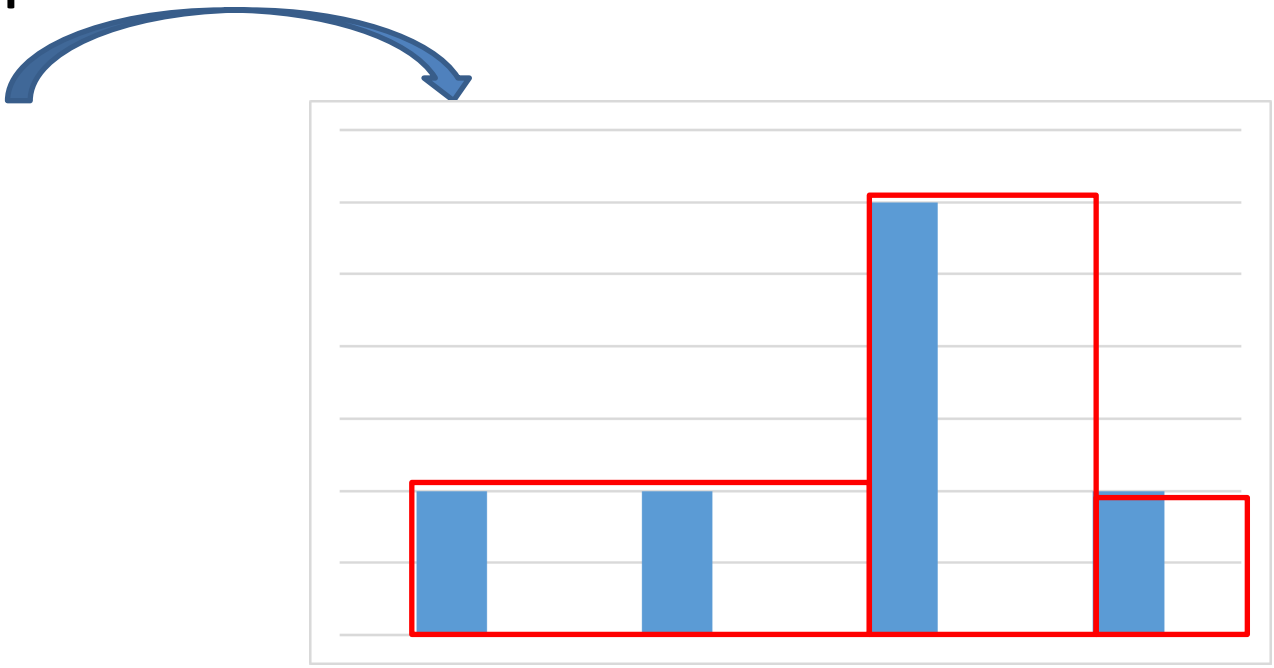


Background

Cardinality = # of total records * selectivity

- histogram

age
3
4
14
15
14
14



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
```

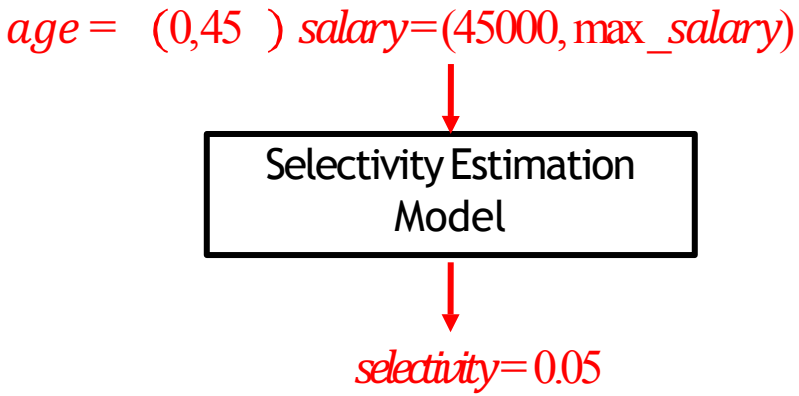
Multi-Column Selectivity Estimation

- 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 farless than that

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

Problem Formulation

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




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

Questions

- If columns are independent, can we do more better estimation of selectivity?
- If columns are dependent, can we improve the estimation of selectivity more?