2. Cardinality estimation
   - How much data?
   - How much data?

RMS Costs

1p H Join
1p T Join
Sort Merge Join
INL Join
2p H Join

\[
\text{# I/Os} = \begin{cases} 
0 & \text{if R or S is sorted} \\
0 + \frac{|R| \log |R| + |S| \log |S|}{\log |R|} & \text{otherwise}
\end{cases}
\]

\[
\text{either } |R| = 0 \text{ or } |S| = 0
\]

\[
2(|S| + |R|)
\]
1p HT

HashSet t
for s in S:
    add s to t
for r in R:
    find matches for r in t:
    emit (matcher r)

2p HT

for s in S:
    write s to on-disk bucket hash(s)
for r in R:
    write r to on-disk bucket hash(r)
for b in bucket:
    in mem join on all tuples in b
SML

```
INLS

for r in R:
  for s in S:
    index lookup(r): 
      emit(r,s)
```

look for # of records in R

cost depends on type of index

Hash: $O(1) \rightarrow 1$

Tree: $\log|S|$
\[ \text{Cost?} \downarrow \quad \text{INS} \]

- Index Scan

\[ \text{II, X: no change} \]

- RWS: sum of |R1, R5|
- RxS: product of |R1, R5|

\[ \text{O} \quad \text{R} \]

- \( A = \text{const} \)
- \( A < \text{const}, \text{const} < A < \text{const} \)
- \( A \neq \text{const} \)
- \( A_1 = A_2 \)
\[
A = \text{Const}
\]

Idea 1
\[
\begin{aligned}
&\text{SELECT } A, \text{ COUNT}(\ast) \\
&\text{FROM R} \\
&\text{GROUP BY A}
\end{aligned}
\]

\[\text{CTMI}\]

Idea 2
\[
\begin{aligned}
&\text{Store any MAX # of rows with a given } A \\
&\text{upper bound on } |O_{C=A} R|
\end{aligned}
\]

\[
\begin{aligned}
&\text{Avg = AVG(|O_{C=A} R|)} \\
&\text{Min = lower bound}
\end{aligned}
\]

Idea 3
\[
\begin{aligned}
&\text{Store min, max of } A
\end{aligned}
\]
SELECT Name
FROM People
WHERE Rank = 3
AND Age = 20

```
+-------+-----+
| Rank  | #   |
|-------+-----|
| 1     | 3   |
| 2     | 2   |
| 3     | 3   |
|       | #   |
| 19    | 1   |
| 20    | 4   |
| 21    | 2   |
| 22    | 1   |
```

Optimal size: 8

<table>
<thead>
<tr>
<th>Rank</th>
<th>Age</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>max 3</td>
<td>max 4</td>
<td></td>
</tr>
<tr>
<td>avg 2</td>
<td>avg 2</td>
<td></td>
</tr>
<tr>
<td>min 2</td>
<td>min 1</td>
<td></td>
</tr>
</tbody>
</table>

Count:

- max: 3
- avg: 2
- min: 1

Values:

- max: 3
- min: 19
- max: 22

Overall size: 8

Filter:

- O_Rank
- O_Age

Index on Rank

Index on Age

Optimal for 3, 20?
Using Statistics

Idea 2  → no more than 3 tuples in $\delta R$
         no more than 4 tuples in $\delta a R$

Idea 2.1 → avg of 2.6 tuples in $\delta R$
            2 tuples in $\delta a$

Idea 3  → avg of $\frac{8}{3}$ tuples in $\delta R$
            avg of $\frac{8}{4}$ tuples in $\delta a$