CSE 344 Section 4
### RA Reference Table

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>Projection</td>
<td>$\pi$</td>
</tr>
<tr>
<td>Join</td>
<td>$\bowtie$</td>
</tr>
<tr>
<td>Group By</td>
<td>$\gamma$</td>
</tr>
<tr>
<td>Set Difference</td>
<td>$-$</td>
</tr>
<tr>
<td>Duplicate Elimination</td>
<td>$\delta$</td>
</tr>
</tbody>
</table>
Typical Plan for a Query (1/2)

\[ \sigma_{\text{having condition}} \]
\[ \gamma_{\text{fields, sum/count/min/max(fields)}} \]
\[ \pi_{\text{fields}} \]
\[ \sigma_{\text{where condition}} \]
join condition

... ... 

SELECT fields
FROM R, S, ...
WHERE condition
GROUP BY fields
HAVING condition
Cost Parameters

- Cost = I/O + CPU + Network BW
  - We will focus on I/O

- Parameters:
  - $B(R) = \#$ of blocks (i.e., pages) for relation $R$
  - $T(R) = \#$ of tuples in relation $R$
  - $V(R, a) = \#$ of distinct values of attribute $a$
    - When $a$ is a key, $V(R,a) = T(R)$
    - When $a$ is not a key, $V(R,a)$ can be anything $< T(R)$

- Where do these values come from?
  - DBMS collects statistics about data on disk
Selectivity Factors for Conditions

- **A = c**  
  
  \[ \sigma_{A=c}(R) = \frac{1}{V(R,A)} \]

- **A < c**  
  
  \[ \sigma_{A<c}(R) = \frac{c - \text{Low}(R,A)}{\text{High}(R,A) - \text{Low}(R,A)} \]

- **c1 < A < c2**  
  
  \[ \sigma_{c1<A<c2}(R) = \frac{c2 - c1}{\text{High}(R,A) - \text{Low}(R,A)} \]
Cost of Reading Data From Disk

- Sequential scan for relation R costs $B(R)$

- Index-based selection
  - Estimate selectivity factor $X$ (see previous slide)
  - Clustered index: $X \times B(R)$
  - Unclustered index $X \times T(R)$

Note: we ignore I/O cost for index pages