

**Lecture 10**  
**10/15/09**

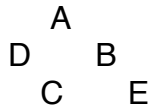
Transactions

Lab 2 due  
 Project proposals due  
 PS2 due next Tuesday  
 Q1 next Thursday

**Finish discussion of Query Optimization**  
**Selinger -- Slides**

how are things different in the real world?

- real optimizers consider bushy plans (why?)



- selectivity estimation is much more complicated than selinger says and is very important.

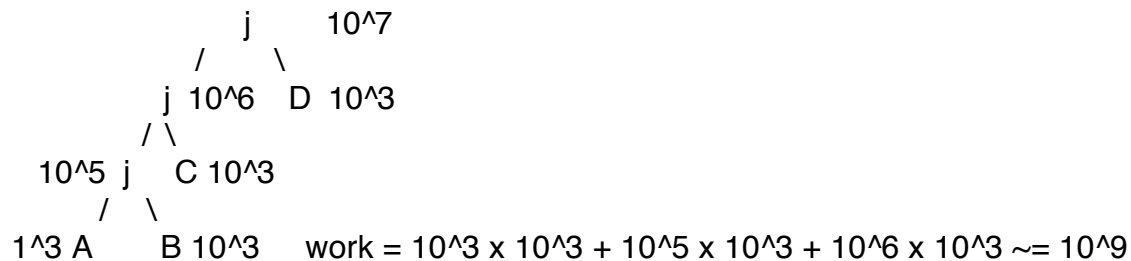
how does selinger estimate the size of a join?

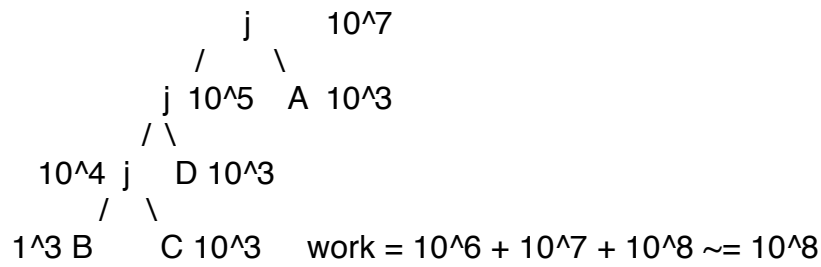
- selinger just uses rough heuristics for equality and range predicates.

- what can go wrong?

consider ABCD A.f1 = B.f2, B.f3 = C.f4, C.f5 = D.f6  
 suppose sel (A join B) = .1  
 everything else is .01

|A| = |B| = |C| = |D| = 1000 ; NL join



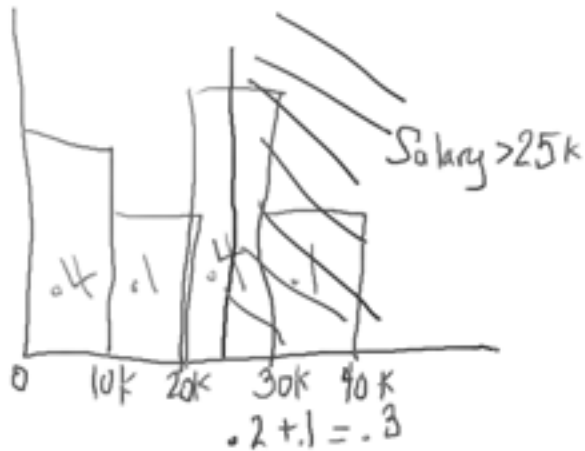


If I don't leave A join B until last, I do 10x more work

Naive selinger estimates all joins have same selectivity

- how can we do a better job?
- (multi-d) histograms, sampling, etc.

example: 1d hist



example: 2d hist (40,80 -> 30,60)

