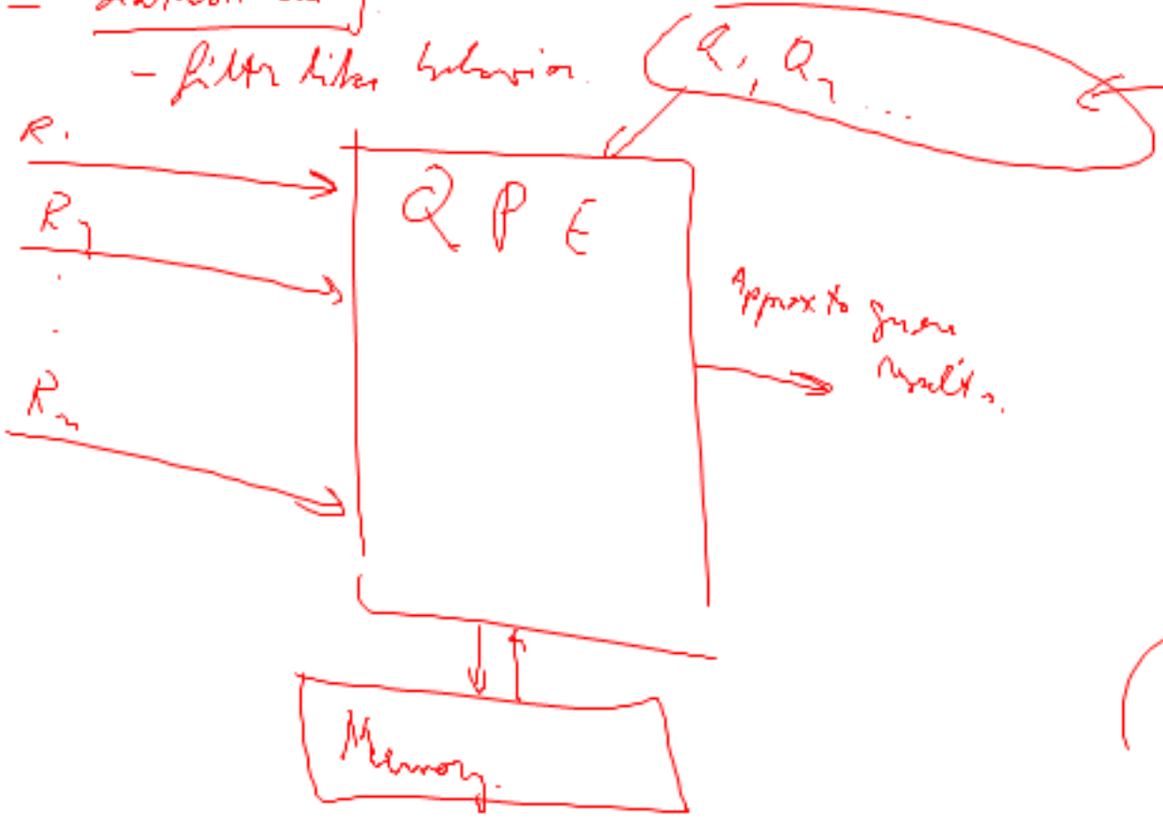
Lecture 14

Sketches

- look at all the data
- use small amount of memory
- datastreaming
  - filter like behavior



$Q_1, Q_2 \ldots$ Queries given

Approx to given queries

- no assumption
- not making predictions

Given $f(D)$ want to compute

$f(D)$ could require a lot of space

Compute $\tilde{f}(D)$

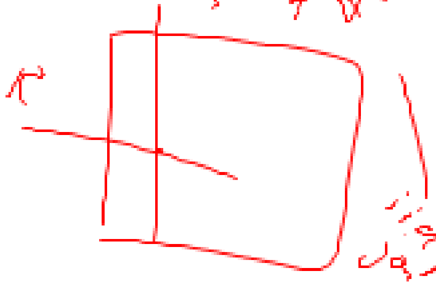
$\tilde{f}(D)$ approx $f(D)$

$\tilde{f}(D)$ using small memory

History of Sampling

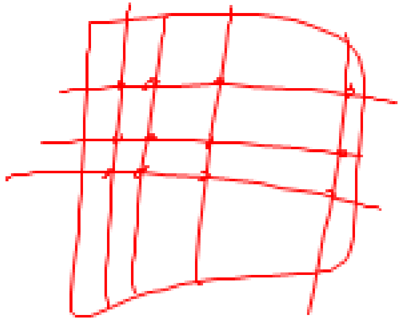
- survey sampling.
  Pearson, Fisher 1930
- in DB
  - Frank Olken 1981
    selectivity estimation
    queries with joins

$|R \bowtie S| = \sigma_{R.a=S.a}(R \times S)$



i.i.d. sample from $R \times S$

Before (Ripple Join)
Get R', S' samples
consider $|R' \bowtie S'|$ ...



$|R'| \cdot |S'|$

↓

W, analyzed them

$V_q \sim \frac{1}{|R'||S'|}$

sample $R \times S$

hard to analyze

$V_q \sim \frac{1}{|R'|}$

$|R'| = |S'|$

Linearity of Sketching

Flajolet - Martin alg for counting distinct elements

- space O(# of distinct elements)
- can you count in less space?
  - need approx

Let $h(x)$ be a univ. hash fn.

$h_s(x)$

- for every item x on the stream
  - compute $h(x)$
  - compute the ~~number of trai~~ position of the first 1 in the binary rep. of $h(x)$

10 110110 010

q

maintain max such number

Claim:

2^{max} approx the number of distinct values