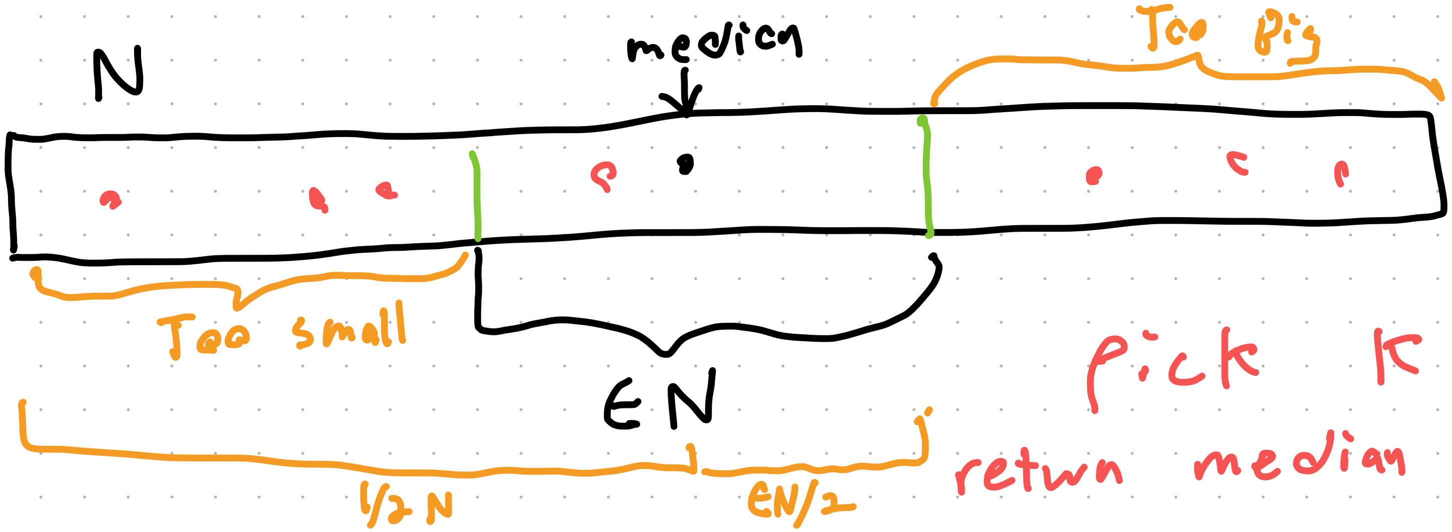
Lecture 4

Start at 8:05



Pick K items
return median of K

What is the chance a random item is too big?

X_i: 1 if ith is too big
0 otherwise

$\frac{1-\epsilon}{2}$

want: Bound $Pr[\sum X_i \geq \frac{1}{2}K]$ ← Failure

$$E[X_i] = \frac{1-\epsilon}{2}$$

$$\Pr\left[\sum x_i \geq \frac{1}{2}k\right]$$

$$E[X_i] = \frac{1-\epsilon}{2} \qquad E\left[\sum_{i=1}^{k} x_i\right] = k\left(\frac{1-\epsilon}{2}\right)$$

Chernoff: $$\Pr\{X \geq \underbrace{(1+\delta)}\underbrace{E[X]}_{k\left(\frac{1-\epsilon}{2}\right)}\} \leq e^{-\frac{\delta^2 \overbrace{E[X]}^{k\frac{1-\epsilon}{2}}}{2+\delta}}$$

$$\frac{1}{2}k = (1+\delta)\, k \cdot \frac{1-\epsilon}{2}$$

$$\delta = \frac{1}{1-\epsilon} - 1$$

$$\leq e^{-\frac{\left(\frac{1}{1-\epsilon}-1\right)^2 k \frac{1-\epsilon}{2}}{2+\delta \quad 3}}$$

$$\leq e^{-k\epsilon^2/6}$$

$$\Pr\left[\sum x_i \geq \frac{1}{2}k\right] \leq e^{-k\epsilon^2/6}$$

$$\Pr[\text{m of m is too Big}] \leq e^{-k\varepsilon^2/6}$$

$$\Pr[\text{m of m is too small}] \leq e^{-k\varepsilon^2/6}$$

$$\Pr[\underbrace{\text{m of m is too Big or Small}}_{\text{Error}}] \leq 2e^{-k\varepsilon^2/6}$$

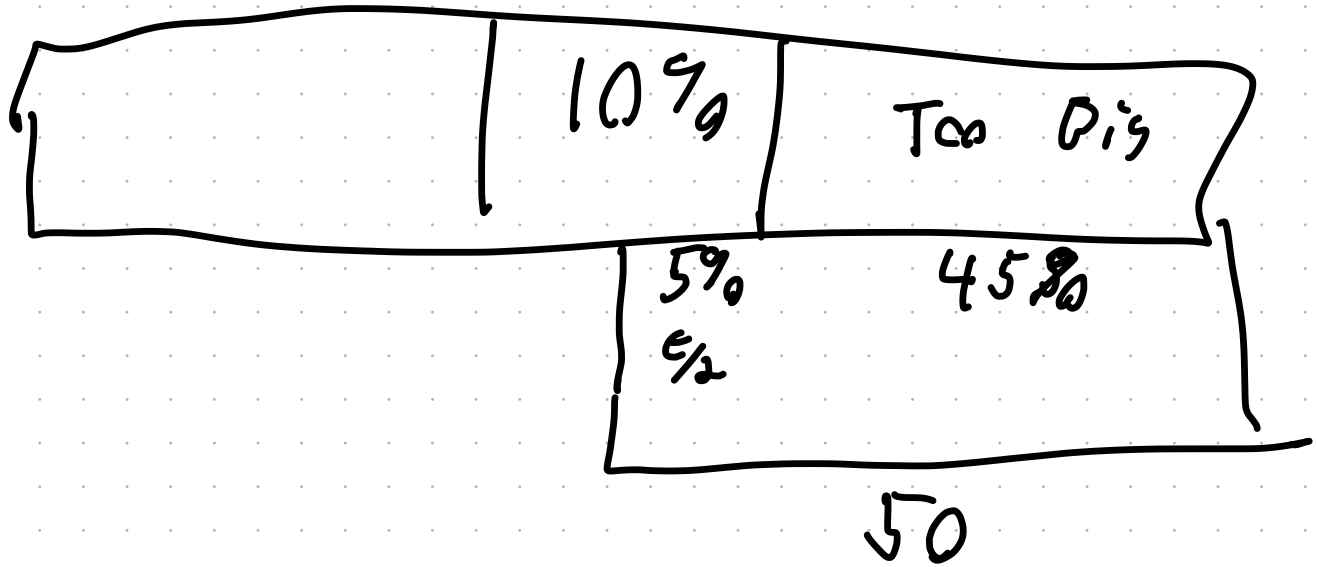
10% median
0.01% failure rate

$\epsilon = 0.1$
$\underline{\delta} = 0.0001$

$\underline{\delta} = 2e^{-k\varepsilon^2/6}$

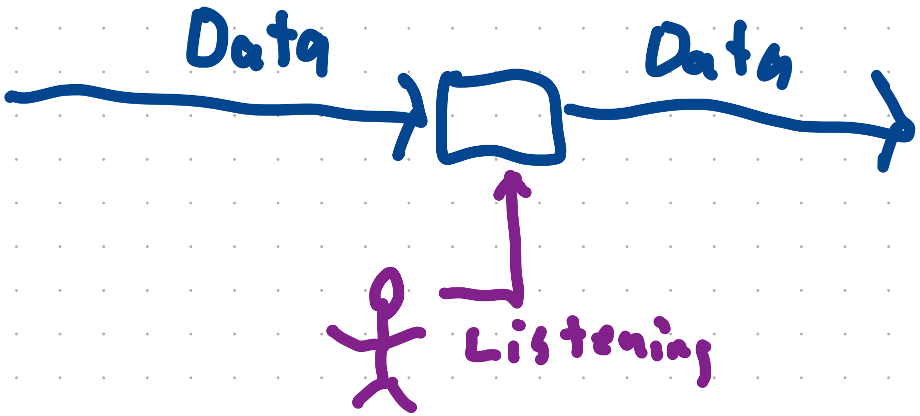
$k = \frac{6}{\varepsilon^2} \ln \frac{2}{\delta}$

$k = 5942$



$$\frac{1}{2} - \frac{\varepsilon}{2} = \frac{1-\varepsilon}{2}$$

Streaming Algoritms



- Approx Median
- Distinct Elements
- Frequencies
- Heavy Hitters

Approx median: Generate K random samples, take median.

$d_1 \quad d_2 \quad d_3 \quad d_4 \quad \cdots \quad d_K \cdots\cdots \quad d_N$

$\uparrow$ Pick as sample

$\uparrow$ Pick $1/2$

$\uparrow$ $1/3$

$\uparrow$ $1/4$

$\uparrow$

K K+1 K+2

$$\frac{1}{\cancel{K}} \cdot \frac{\cancel{K}}{\cancel{K+1}} \cdot \frac{\cancel{K+1}}{\cancel{K+2}} \cdots\cdots \frac{\cancel{N-1}}{N}$$

$$= \frac{1}{N}$$

Break until

9:07

Distinct Elements

$\xrightarrow{\text{Elements}}$ How many different?

$h(x)$ = number in same range.

$h(x) = h(x)$

$h(x) \neq h(y)$ (usually)

$h(x)$ = "looks like a random number"

$\Pr[h(x) = \text{even}] = \frac{1}{2}$

$\Pr[h(x) \text{ is div by } 4] = \frac{1}{4}$

$\Pr[h(x) \text{ is div by } 8] = \frac{1}{8}$

$\vdots$

$zeros(x) = i$ means 2^i is the largest power of 2 that x is divisible by.

$zeros(48) = 4$

$$\begin{array}{l} 48 \\ 24 \\ 12 \\ 6 \\ 3 \text{ odd} \end{array}$$

110000

$Pr[zeros(h(x)) = i] = \frac{1}{2^i}$

$\frac{1}{2} \quad \frac{1}{4} \quad \frac{1}{8} \quad \frac{1}{16} \ldots$

Stream $x_1, x_2 \ldots x_k$

$$\boxed{\text{return } 2^{\max(zeros(h(x_i))) + \frac{1}{2}}}$$

Analysis of distinct elements

$Z_j(x_i) = 1$ when $z(x_i) \geq j$

Let $x_1, x_2, \ldots x_d$ Be the distinct element

$d = \#$ of distinct elements

$Y_j = \sum Z_j(x_i)$ # of items that have $\geq j$ 0's

$z(x_i) = \#$ of zeros in $2(\text{hash}(x_i))$

$Z = \max_i z(x_i)$ max # of zeros seen

$\hat{d} = 2^{Z+\frac{1}{2}}$ This is the output, We hope $\hat{d}$ is close to d

$\Pr[\hat{d} \geq 3d]$ What is the chance $\hat{d}$ is too big, at least a factor of 3

$\Pr[2^{Z+\frac{1}{2}} \geq 2^{\log 3d}]$

$\Pr[Z+\frac{1}{2} \geq \log 3d]$

$\Pr[Y_{\boxed{\log 3d - \frac{1}{2}}} \geq 1]$

$$\Pr[X > K] < \frac{E[X]}{K}$$

$$\frac{d}{2^{\log 3d - \frac{1}{2}}} = \frac{d}{3d \frac{1}{\sqrt{2}}} = \frac{\sqrt{2}}{3}$$

$$= 47\%$$

$$E[Y_j] = \sum_{i=1}^{d} E[Z_j(x_i)]$$

$$= d \cdot \frac{1}{2^j}$$

$\Pr[\hat{d} \geq 3d] \leq 47\%$

$\Pr[\hat{d} \leq \frac{1}{3}d] \leq 47\%$ (will be in the notes)

$\Pr[\frac{1}{3}d \leq \hat{d} \leq 3d] \leq 6\%$

The alg gives a factor-3 approximation with a success rate of 6%

$\epsilon = 6\%$

$$\frac{1}{\epsilon^2} \log \frac{1}{\delta}$$