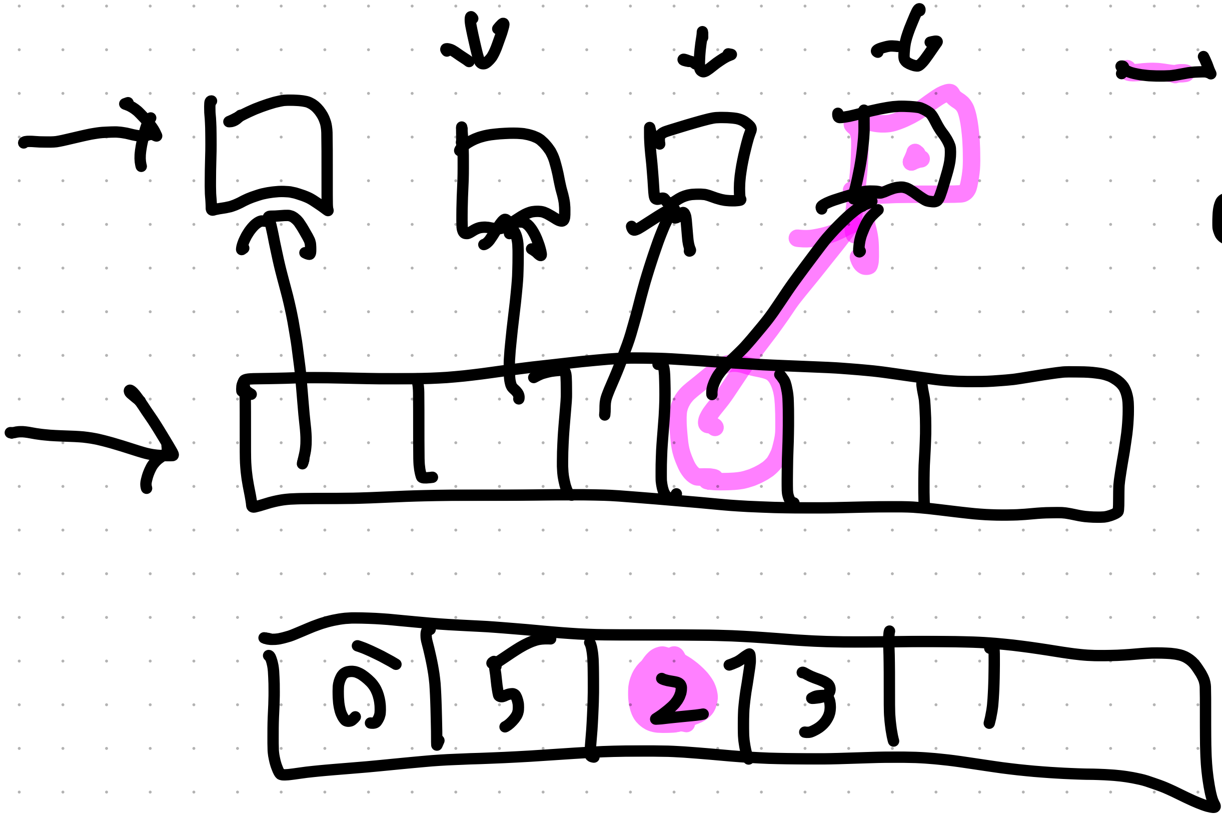
Big Data Algorithms

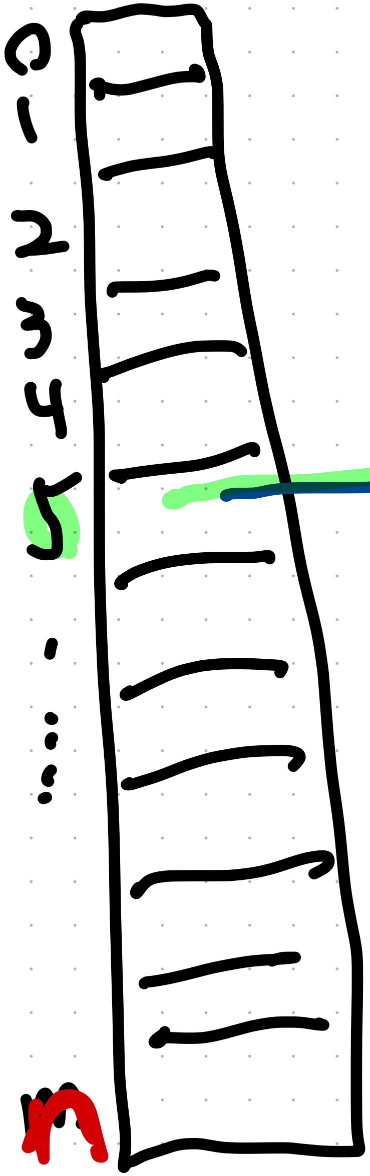
Week 3

start at 8:05

— Better analysis of hashing



→ Sublinear algorithms / streaming



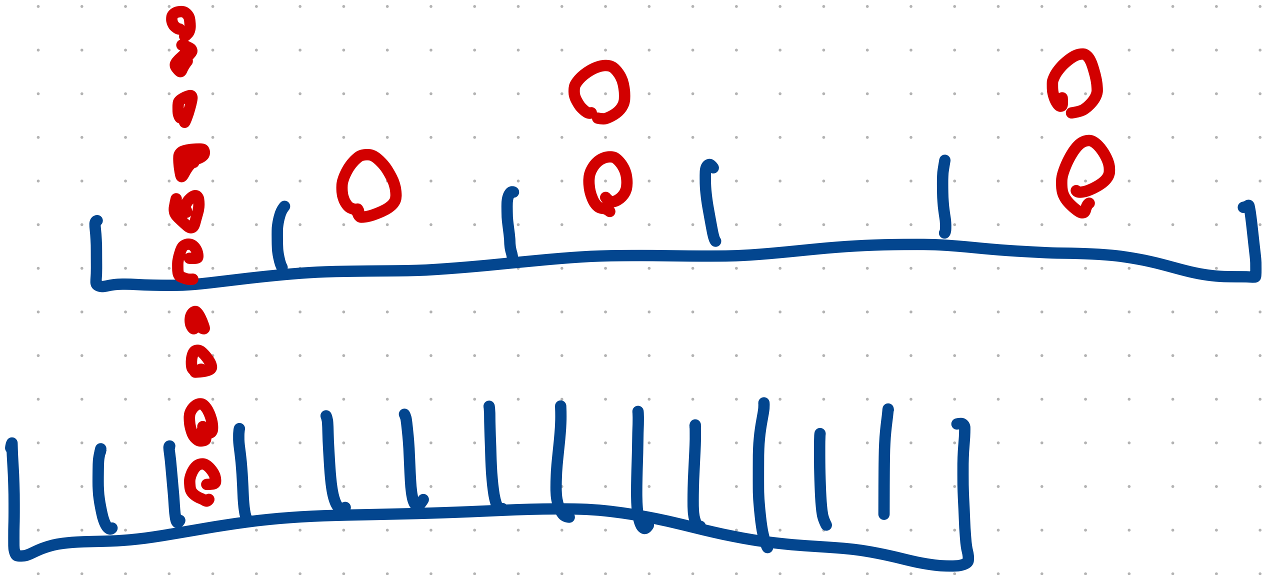
$\text{hash}(x) \% m$

$\text{hash}(10) \% m = 5$

Insert n items

"Balls in Bins"

Given n balls, throw them randomly into n bins



N Balls N bins

1. What is the average size bin?
2. What is the expected size of a bin?

- 3. Can you say: Bin i has size at most x with prob $\leq p$
- 4. Can you say: All bins have size at most x with prob $\leq p$

Average

b_i = size of bin i

$$\text{Average Bin size} = \frac{1}{N} \underbrace{\sum_{i=1}^{N} b_i}_{N} = \frac{1}{N} \cdot N = 1$$

Expected Value

$$E[X] = \sum_{i} i \, Pr[X = i]$$

random variable

$$E[b_i] = \sum_{j=0}^{N} j \, \underbrace{Pr[b_i = j]}_{\text{what is this}}$$

$$Pr[b_i = 0] = \left(1 - \frac{1}{N}\right)^N \underset{\text{fact from last time}}{\approx} \frac{1}{e}$$

$$Pr[b_i = j] = \binom{n}{j} \left(1 - \frac{1}{N}\right)^{N-j} \left(\frac{1}{N}\right)^{j}$$

Linearity of Expectation

$$E[X+Y] = E[X] + E[Y]$$

$E[X \cdot Y]$ Not true that always $= E[X \cdot Y]$

$$b_{i,j} = \begin{cases} 1 & \text{if } j^{th} \text{ ball is in } b_i \\ 0 & \text{otherwise} \end{cases}$$

"Indicator random variable"

$$E[b_i] = E\left[\sum_{j=1}^{N} b_{i,j}\right]$$

$$= \sum_{j=1}^{N} E[b_{i,j}]$$

$$= \sum_{j=1}^{N} \frac{1}{N} = N \cdot \frac{1}{N} = 1$$

$$E[b_{i,j}] = \sum_{k} k \Pr[b_{i,j} = k]$$

$$= \Pr[b_{i,j} = 1]$$

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

Markov's Inequality

100 students

Average score is 4

Claim: At most 50 students have a grade of 8 or higher

Assumption: Scores are not negative

$$\Pr[X > a] \leq \frac{E[X]}{a}$$

wanted: bin i has size ⎵ with pr ⎵

$$\Pr[b_i > X] \leq \frac{E[b_i]}{X} = \frac{1}{X}$$

bin i has size >100 with probability $\leq \frac{1}{100}$

≤ 100 99%

Chernoff Bounds

They bound $\sum_{i=1}^{k} X_i$ where the X_i are **independent**

rand var

Defn. X, Y are ind if $\Pr[X=i \text{ and } Y=j] = P[X=i] \cdot P[Y=j]$

$b_{i,j}$ = Ball j is in bin i

$\underbrace{b_{i,j}}$ and $b_{i,j'}$ $j' \neq j$ $\underbrace{\text{Ind}}$

$b_{i,j}$ and $b_{i',j}$ $i' \neq i$

$\left(\frac{1}{w}\right)^N$ b_i b_j $i \neq j$

Break until

9:10

Chernoff Bounds

They bound $\sum_{i=1}^{k} X_i$ where the X_i are independent

$$\Pr[X \leq (1-\delta)E[X]] \leq \frac{1}{e^{\delta^2 E[X]/2}} \qquad 0 \leq \delta \leq 1$$

$$\Pr[X \geq (1+\delta)E[X]] \leq \frac{1}{e^{\delta^2 E[X]/(2+\delta)}} \qquad 0 \leq \delta$$

$$\Pr[b_i \geq (1+\delta)\underbrace{E[b_i]}_{1}] \leq e^{-\delta^2 \underbrace{E[b_i]}_{1} /(2+\delta)}$$

$$\Pr[b_i \geq (1+\delta)] \leq e^{-\delta^2/(2+\delta)}$$

$$\Pr\Big[b_i \geq (\overbrace{1+\delta}^{100})^{1}\Big] \leq e^{-\delta^2/(2+\delta)}$$

what is $\Pr[b_i \geq 100] \leq e^{-99^2/(2+99)}$

$\delta = 99$

$$= e^{-\frac{9801}{11}}$$

$$= 0.000000000000000000000000000000000000000000718$$

With High Probability

Event e happens "with polynomially high probability" means it happens with prob $1 - O\left(\frac{1}{n^c}\right)$ for same c

Intuition: Chernoff $e^{-\delta}$ $\nearrow$ $e^{-c \ln N} = N^{-c} = \frac{1}{N^c}$

Set $\delta = c \ln N$

What is the chance $b_i \geq 1 + \alpha \ln n$?

$$\Pr[b_i \geq (1+\delta)] \leq e^{-\delta^2/2+\delta}$$

$$\Pr[b_i \geq 1+\alpha \ln n] \leq e^{-(\alpha \ln n)/(2+\alpha \ln n)}$$

$$\approx \frac{1}{n^{\alpha-1}}$$

$$\Pr[b_i \geq 1+3 \ln n] \leq \frac{1}{n^2}$$

"bin i is $\leq 1+3\ln n$ w.h.p."

"bin i is $O(\log n)$ w.h.p."

Union Bound

$$P[X=i \text{ and } Y=j] \leq P[X=i] + P[Y=j]$$

$$P\left[\begin{array}{c}\text{4 coins tossed} \\ \text{all heads}\end{array}\right] \leq 4 \cdot P[\text{one heads}]$$

$$= 4 \cdot \frac{1}{2}$$

$$= 2$$

$$Pr[b_i \geq 1 + 3\ln n] \leq \frac{1}{n^2}$$

$$Pr[\underset{i}{\overset{\text{For}}{\text{All}}} \text{ in } 1..N \;\; b_i \geq 1 + 3\ln N] \leq n \cdot \frac{1}{N^2} = \frac{1}{N}$$

Approximate Median Finding

$\overset{10000}{\cancel{1}}$, 3, 99, 32, 78, 2, 4, 6, 8

Average: 25.8 → 1136

Median: 6 → 8

ϵ - Approx Median

x is an ϵ-approx median if it is between the

$\left(\frac{1-\epsilon}{2}\right)N$ th and $\left(\frac{1+\epsilon}{2}\right)N^{th}$ largest item

Stupid alg: Pick a random item x

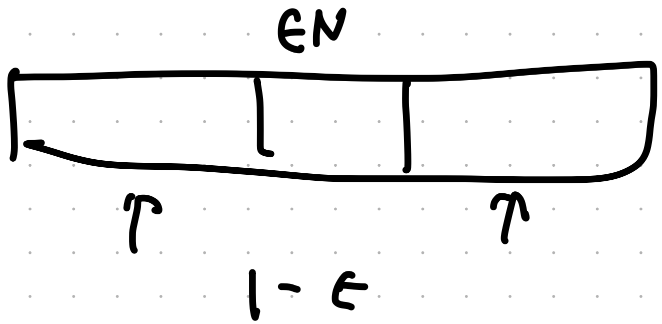
$\rightarrow$ x is a 50%-median with prob 50%

Better algorithm: Pick K random elements
return the median of these elements

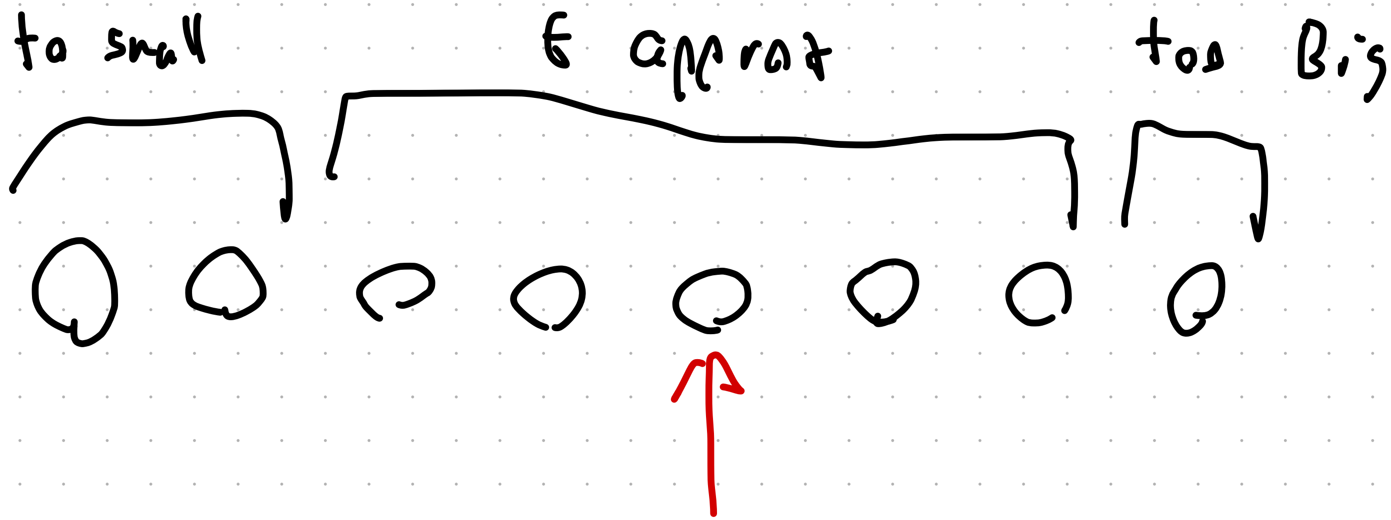
X_i = ith sample is not a ϵ-approx median

Bad sample

$$E[X_i] = 1-\epsilon$$



X = Total number of Bad samples



$K=2$

$K=1$