CS 124 — Lecture 9

Birthday paradox

$Pr[\text{2 people do not share the same bday}] = 1 - \frac{365}{365^2} = 1 - \frac{1}{365}$

person by person

$Pr[\text{no 2 people share the same bday}) = \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{n}{365}\right)$

$p(2 \text{ no bd})\, p(3 \text{ no bd} \mid 2 \text{ no bd})\, p(4 \text{ no bd} \mid 3 \text{ no bd}) \cdots p(n \text{ no bd} \mid n-1 \text{ no bd})$

$1 - \left(\frac{365}{365^2} + \frac{364}{365^2}\right) = 1 - \frac{2}{365}$

$[365 \cdot 364 \cdots (365 - n)]$

$n = 60 \sim 0.5\%$

group

$Pr[\ldots] = \frac{\binom{365}{n} n!}{365^n}$ } orderings of [redacted] days, not the same, across n slots / all orderings of n days across n slots

$n > 60$
$Pr[\_] \sim 0.5\%$

Generalized (balls, bins)

m "people", m "days"

$Pr[\text{no collision}] = \prod_{j=1}^{m-1} \left(1 - \frac{j}{m}\right) \approx \prod_{j=1}^{m-1} e^{-j/m} =$

Approx: $1 - \frac{k}{m} \approx e^{-k/m}$, small k

$= \exp\left[-\sum_{j=1}^{m-1} \frac{j}{m}\right] = \exp\left[-\frac{1}{m} \cdot \frac{m-1}{2} \cdot m\right] \approx e^{-m^2/2m}$

Q: What is the value of m for which $Pr[\text{collision}] = 1/2$

$\ln\left(e^{-m^2/2 \cdot 365}\right) = \ln\left(\frac{1}{2}\right)$

$-\ln\left(\frac{1}{2}\right) = \ln(1) - \ln\left(\frac{1}{2}\right) = \ln\left(\frac{1}{1/2}\right) = \ln(2)$

$\frac{+m^2}{2 \cdot 365} = \ln(2)$

[redacted]

$n \approx 22.49$

**Q: How many bins ("days") are empty?**

$\Pr[\text{a bin is empty}] = \left(1 - \frac{1}{m}\right)^n \approx e^{-1/m^n} =$

↑ hit the bin

$E\left[\sum_{i=1}^{m} X_i\right] = \sum_{i=1}^{m} E(X_i)$

linearity of expectation over indicator r.v.s

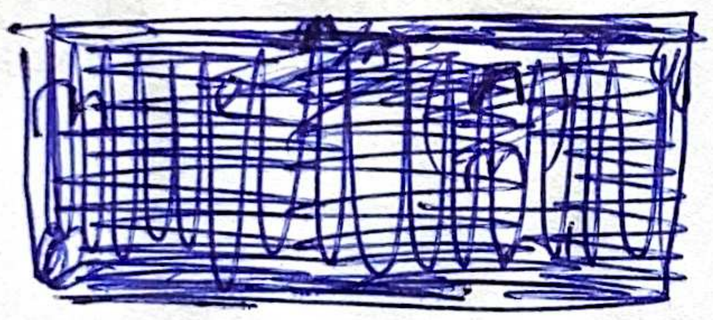
$\#\text{ empty bins} = m \cdot e^{-n/m} = e^{\frac{-n}{m}}$

**Q: prob. a bin has r balls**

$\Pr[\text{given bin has r balls}] = \binom{n}{r}\left(\frac{1}{m}\right)^r\left(\frac{m-1}{m}\right)^{n-r}$

$\approx e^{-\frac{n}{m}} \frac{\left(\frac{n}{m}\right)^r}{r!}$

**Q: # of bins with r balls in each**

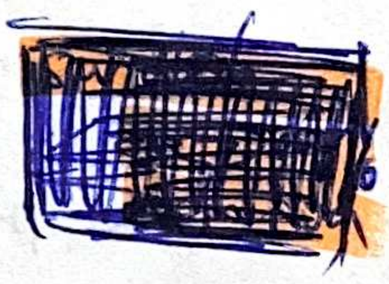


$m \cdot e^{-n/m} \frac{\left(\frac{n}{m}\right)^r}{r!}$

**Hash function**: maps a larger set to a smaller set

(Birthday is a hash function)

→ eventually get collisions



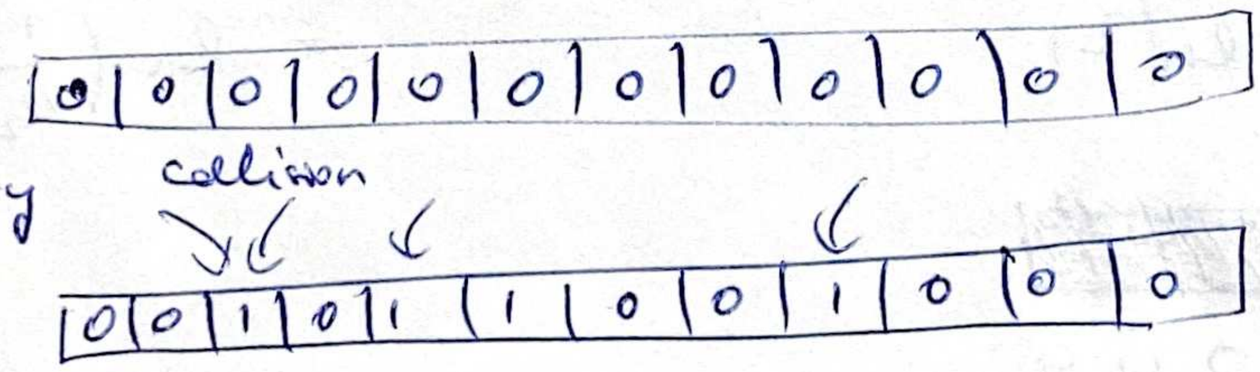
**Defn**: $f: \{0 \ldots n-1\} \longrightarrow \{0 \ldots m-1\}$

(usually $n << m$)

mapping is deterministic

each elt in $\{0 \ldots m-1\}$ is equally likely mapped to any elt in $\{0 \ldots m-1\}$

Password ex



trade-off btwn. space and false positive rate

fraction $e^{-n/m}$ bins will be empty (on average)

$$P[\text{reject strong pwd}] = 1 - e^{-n/m}$$ ← FP rate

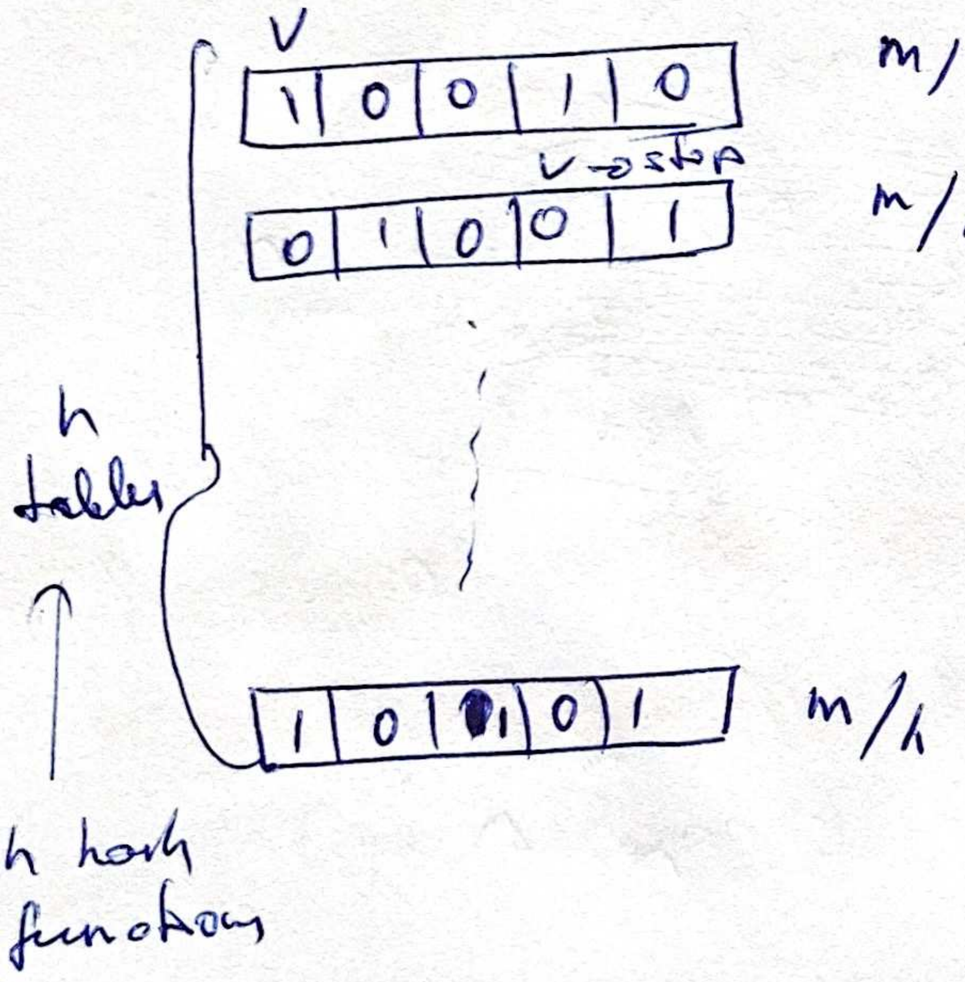
$p = e^{-n/m}$ a strong pwd hits any bin

equally likely. with $1-p$ the bin that was hit is non-empty

**Bloom filters**

smart way to deal with space vs. FP rate trade-off

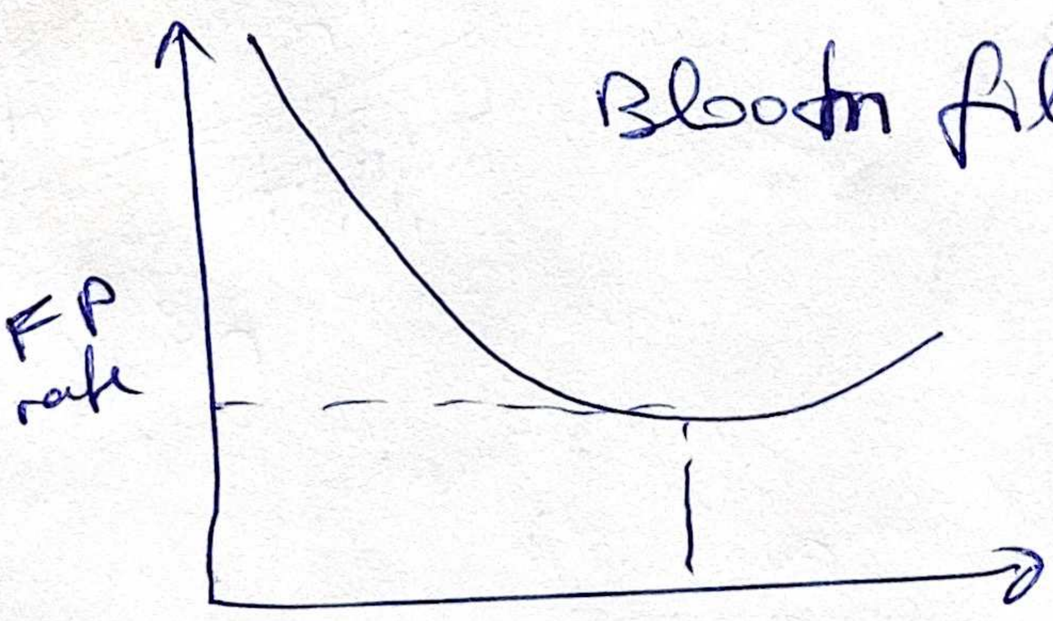
$m$ bits, but $h$ tables, each of size $\frac{m}{h}$




Pr [reject a strong pwd] $= \left(1 - e^{-n/(m/h)}\right)^h = \left(1 - e^{-nh/m}\right)^h$

optimal $h$ ?

make smaller ↓ h ↑ make larger




$$h_{opt} = (\ln 2)\frac{m}{n}$$

e.g. 100,000 weak pwds

~ 7 chars long

700,000 bytes → 350,000 bytes compressed

Bloom filter: 5 tables, 160k bits

FP = 2%