Start at

8:05

Streaming

Given a data item, what is its frequency in the stream?

f_x = # of times x has appeared

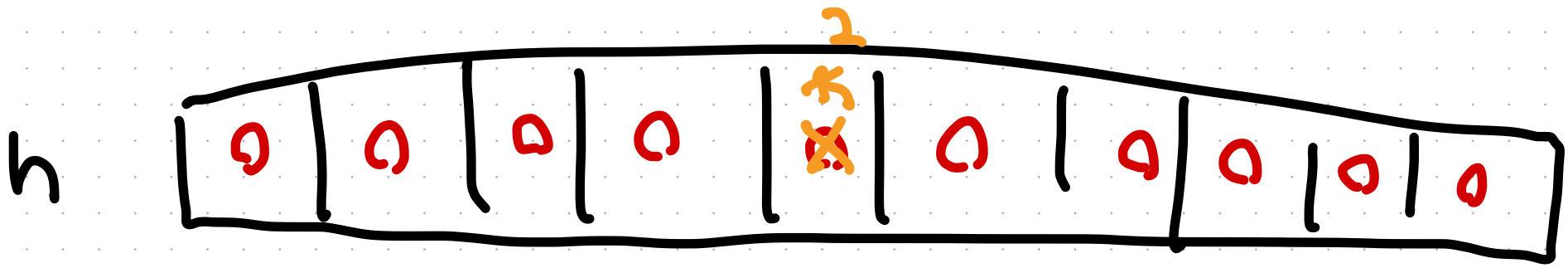
m = total amount of data

$m = \sum f_x$

Compute $\hat{f}_x$ such that $f_x \leq \hat{f}_x \leq f_x + m\epsilon$ with probability $\geq 1 - \delta$

Dad Way:

Create a hash table of size w

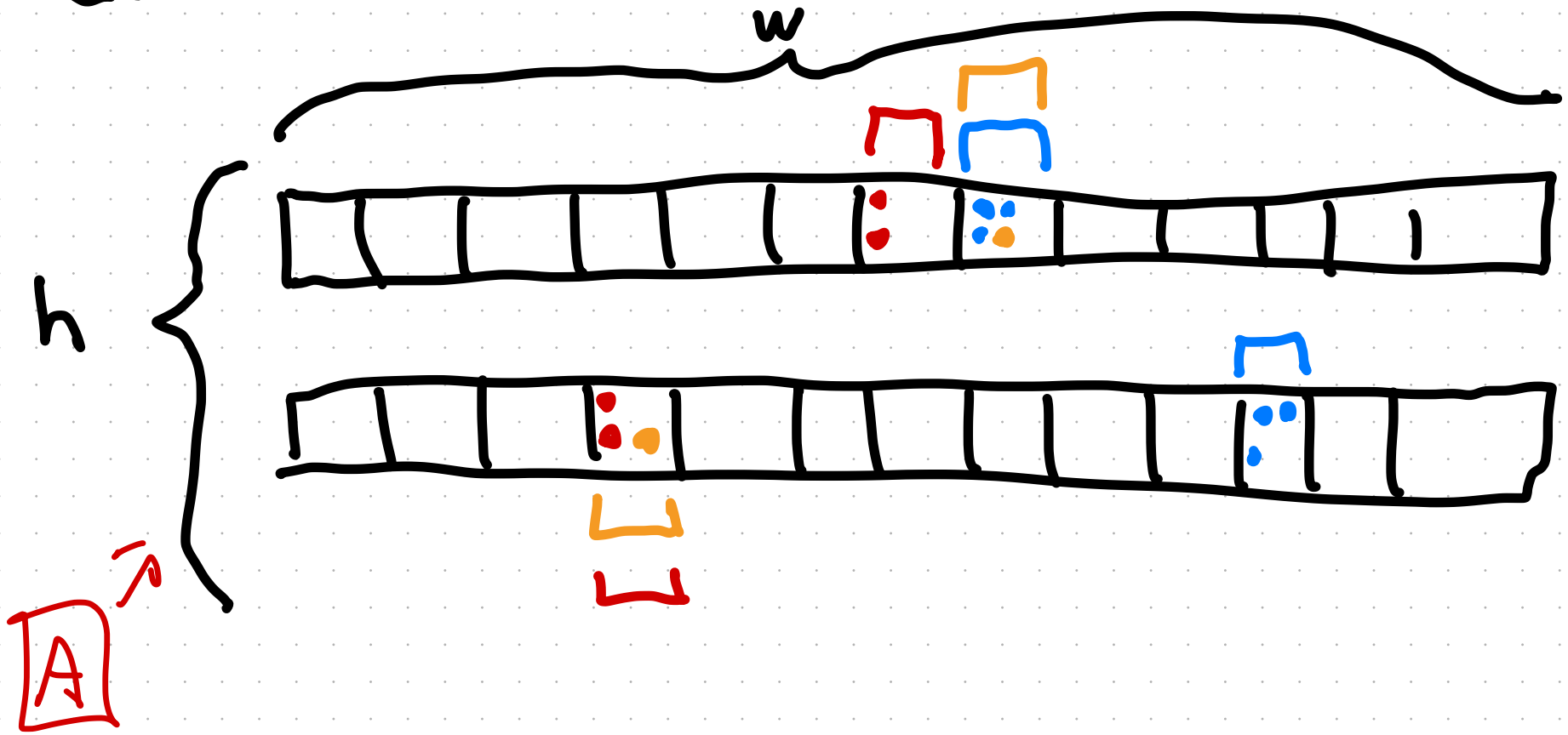


Add("x")
Add("x")
Query("xyz")

Add(x): h[hash(x) % w] += 1

Query(x): return h[hash(x) % w]

Count-min sketch:



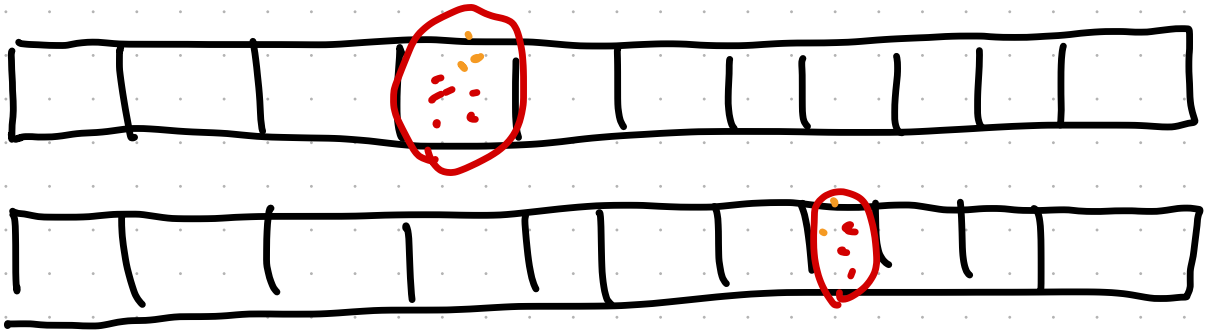
Add(x): $A[h][\text{hash}_h(x) \% w] += 1$ for all h

Query(x): $\min_h A[h][\text{hash}_h(x) \% w]$

h = # of tables = $\log \frac{1}{\delta}$ $\quad$ δ = failure prob

w = size of table = $\frac{2}{\epsilon}$ $\quad$ ϵ = accuracy

m = total ammount of stuff inserted



Let $X_{x,i} = A[i][hash_i(x)] - f_x$

$E[X_{x,i}] = \frac{m}{w} \rightarrow Pr[X_{x,i} \geq \frac{2m}{w}] \leq \frac{1}{2}$ ($\frac{2m}{w} = m\epsilon$)

$$Pr\left[\bigwedge_{i=1}^{h} [X_{x,i} \geq m\epsilon]\right] = \left(\frac{1}{2}\right)^h = \delta$$

Problem

Several texts

Ask questions about similarity

- Give me all pairs of texts that are similar
- Given a new text, is it similar to existing ones?

This ␣ is ␣ a ␣ cat.

"Edit distance"

There ␣ as ␣ a ␣ gnat.

"Jaquard distance" between x and y

$$\frac{\text{Fragments in both } x \text{ and } y}{\text{Fragents in } x \text{ and/or } y}$$

Compute fragments:

A scan, if the document has n characters, generate n fragments

Jaquard (x, y)

Takes time $O(|x|+|y|)$

Suppose you have m documents of size n

have a new document of size n

To find the one with min distance