CountMin sketch

Ben Langmead

JOHNS HOPKINS
WHITING SCHOOL *of* ENGINEERING

Department of Computer Science

Please sign guestbook (www.langmead-lab.org/teaching-materials) to tell me briefly how you are using the slides. For original Keynote files, email me (ben.langmead@gmail.com).

Counting

Input is a "stream" of items $\{a_1, a_2, \ldots, a_m\}$, each from universe of size n.

Number of times a value x appears is its "count" or "frequency" f_x

Stream of zip-code digits: 2, 1, 2, 1, 8, 2, 6, 8, 2

$m = 9$

$n = |\{0,1,\ldots,9\}| = 10$

$f_1 = 2 \qquad f_2 = 4$

Aside on notation

Defining variables like n, m, N, M, and having to specify "distinct" versus not, can get tiresome

An alternative is to pick a variable for the input data stream, say $\mathbf{a}$

Then use double bars to express "moments"

$$\| \mathbf{a} \|_1 = \sum_{x \in distinct(\mathbf{a})} f_x \qquad \| \mathbf{a} \|_0 = \sum_{x \in distinct(\mathbf{a})} (f_x)^0$$

\# items in stream

\# *distinct* items in stream

Aside on notation

We can also consider higher moments like 2:

$$\| \mathbf{a} \|_2 = \sum_{x \in distinct(\mathbf{a})} (f_x)^2$$

Or, more generally, k:

$$\| \mathbf{a} \|_k = \sum_{x \in distinct(\mathbf{a})} (f_x)^k$$

Today we're concerned with f_x, not its powers

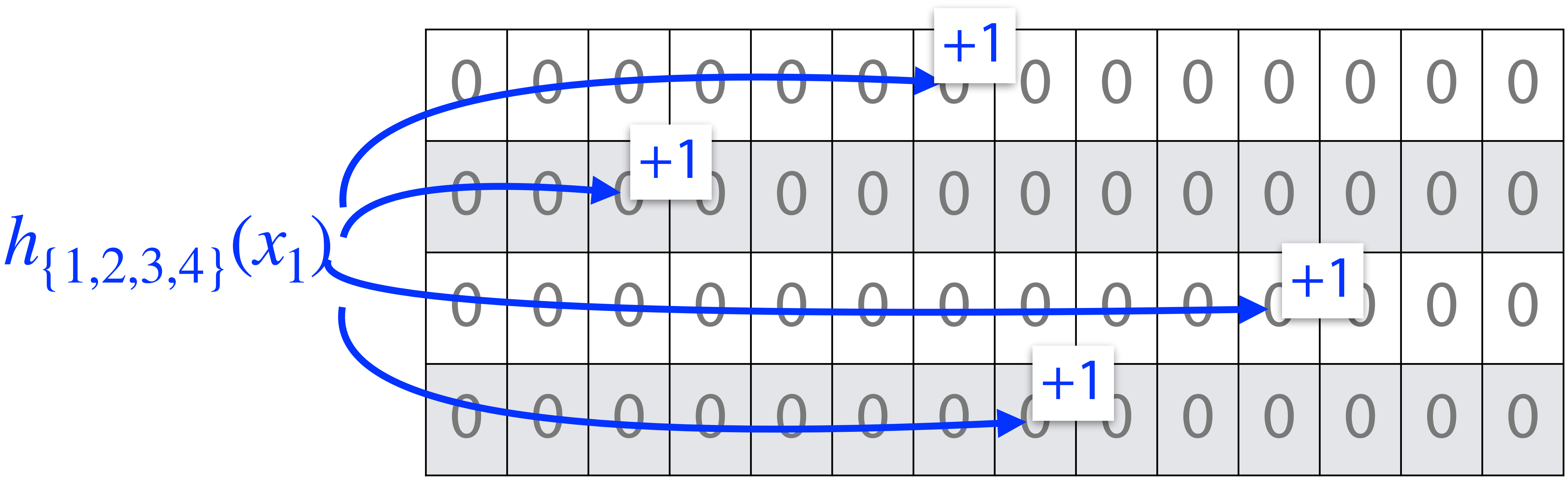
CountMin

Matrix of counters, all initially 0

0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0

CountMin

Insert:

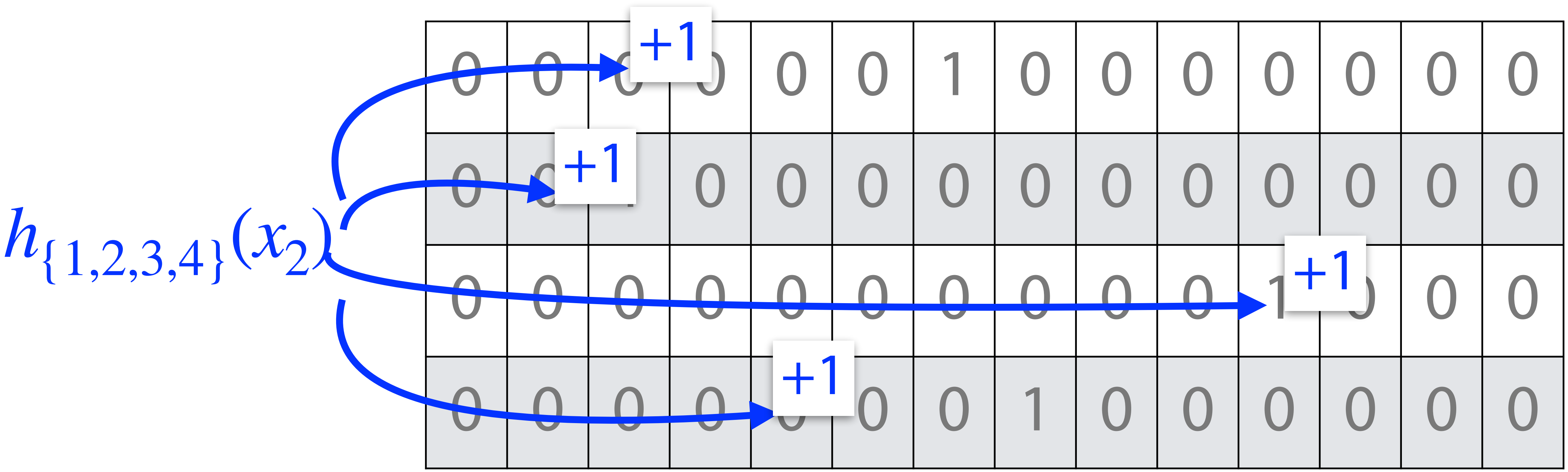


CountMin

0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0

CountMin

Insert:



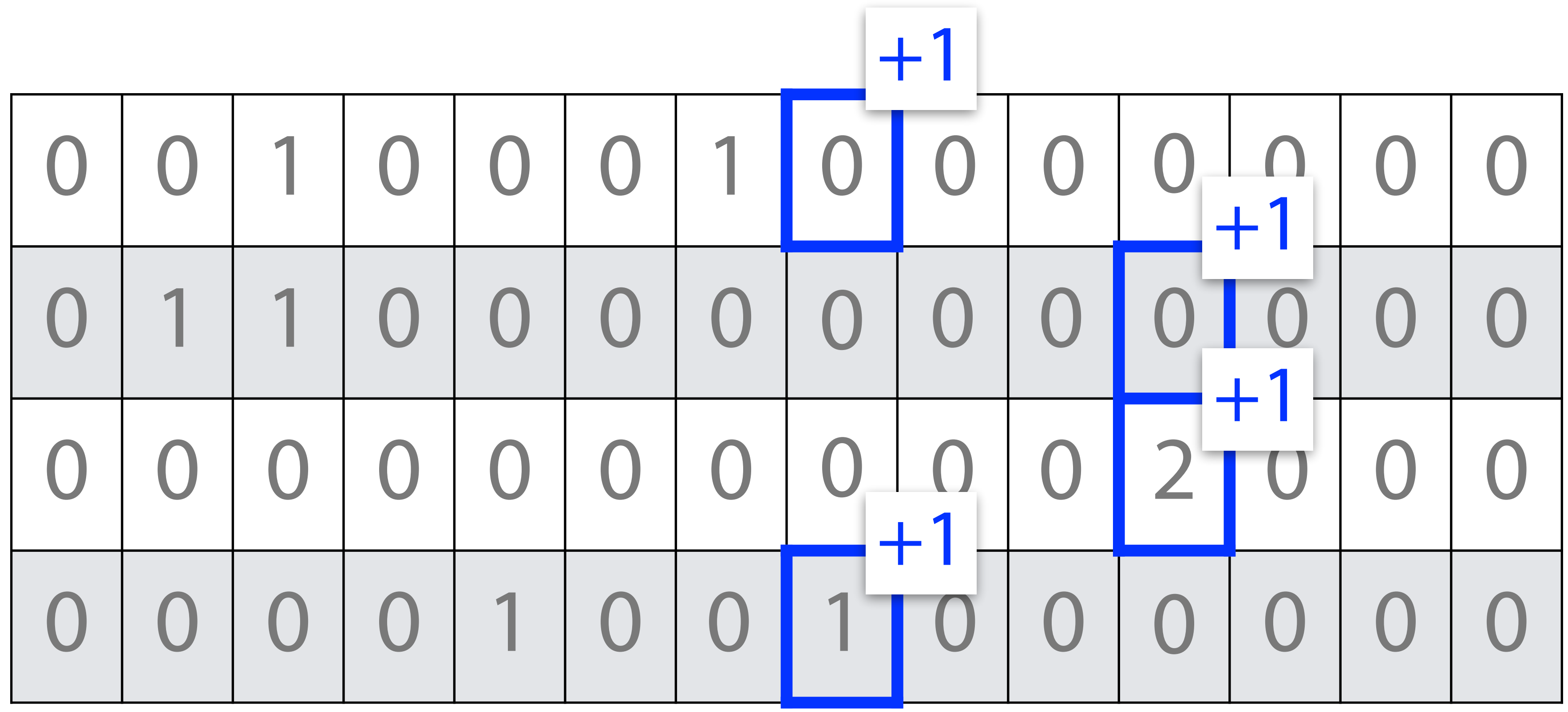
CountMin

0	0	1	0	0	0	1	0	0	0	0	0	0	0
0	1	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	2	0	0	0
0	0	0	0	1	0	0	1	0	0	0	0	0	0

CountMin

Insert:

$h_{\{1,2,3,4\}}(x_3)$



CountMin

0	0	1	0	0	0	1	1	0	0	0	0	0	0
0	1	1	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	3	0	0	0
0	0	0	0	1	0	0	2	0	0	0	0	0	0

CountMin

Point query:

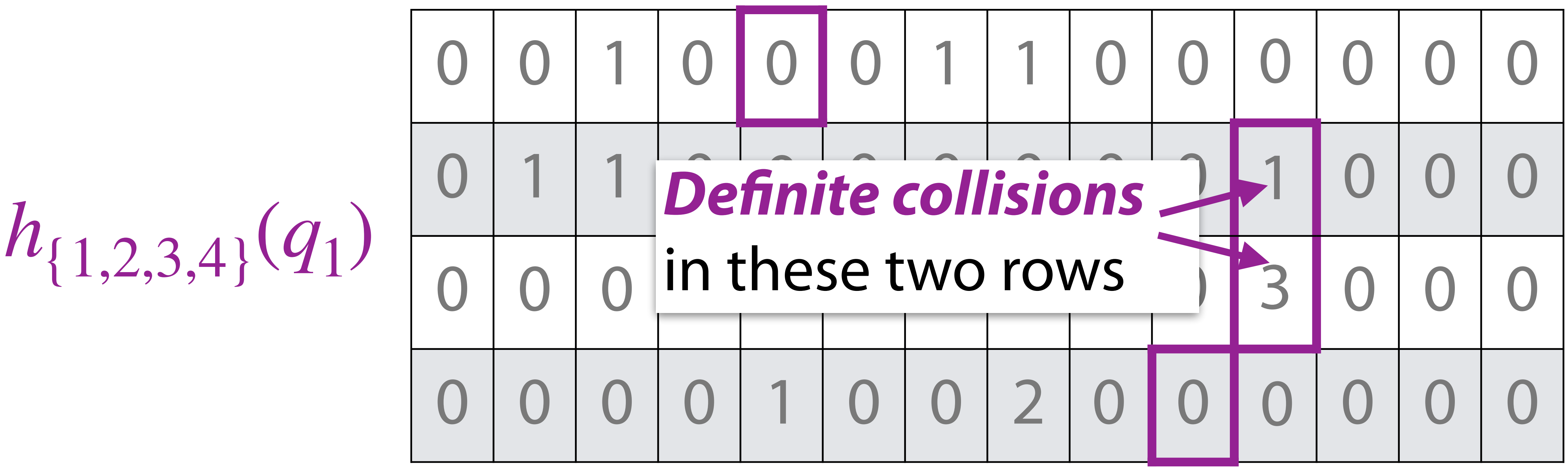
$h_{\{1,2,3,4\}}(q_1)$

0	0	1	0	0	0	1	1	0	0	0	0	0	0
0	1	1	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	3	0	0	0
0	0	0	0	1	0	0	2	0	0	0	0	0	0

What should the estimate $\tilde{f}_x$ be? 0

CountMin

Point query:



How much is due to collisions?

CountMin

Point query:

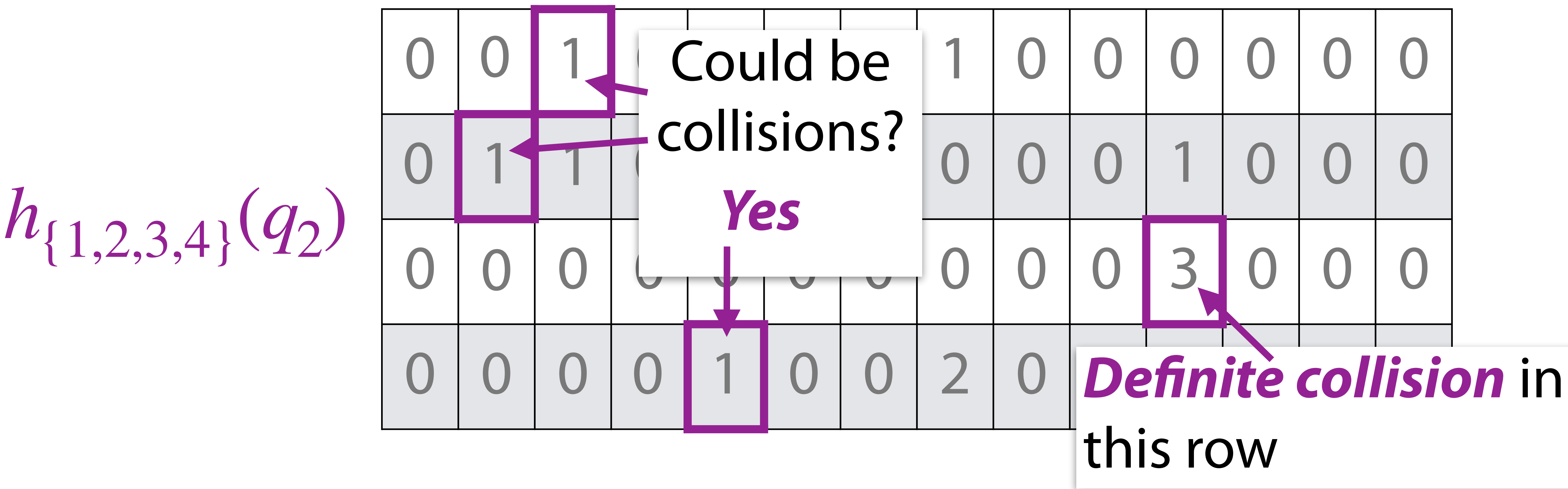
$h_{\{1,2,3,4\}}(q_2)$

0	0	**1**	0	0	0	1	1	0	0	0	0	0	0
0	**1**	1	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	**3**	0	0	0
0	0	0	0	**1**	0	0	2	0	0	0	0	0	0

What should the estimate $\tilde{f}_x$ be? 1

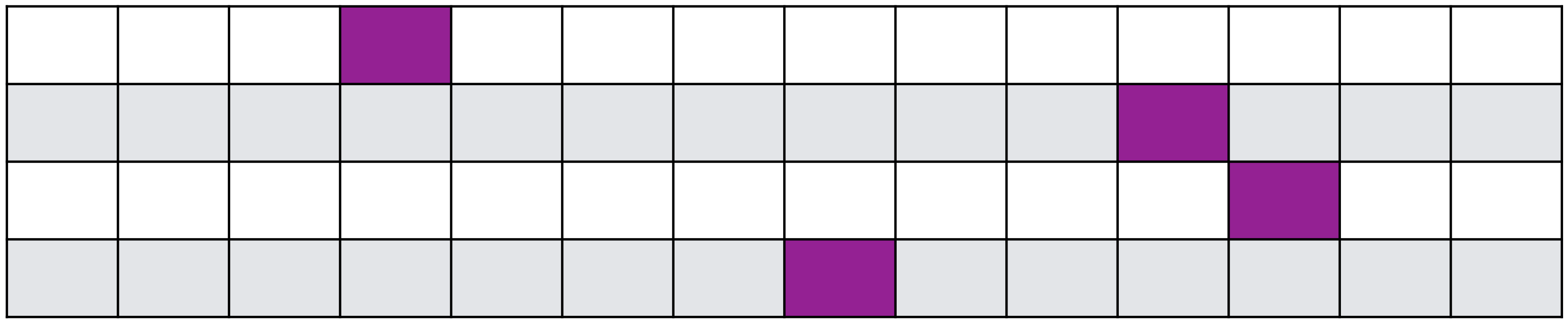
CountMin

Point query:



How much is due to collisions?

CountMin

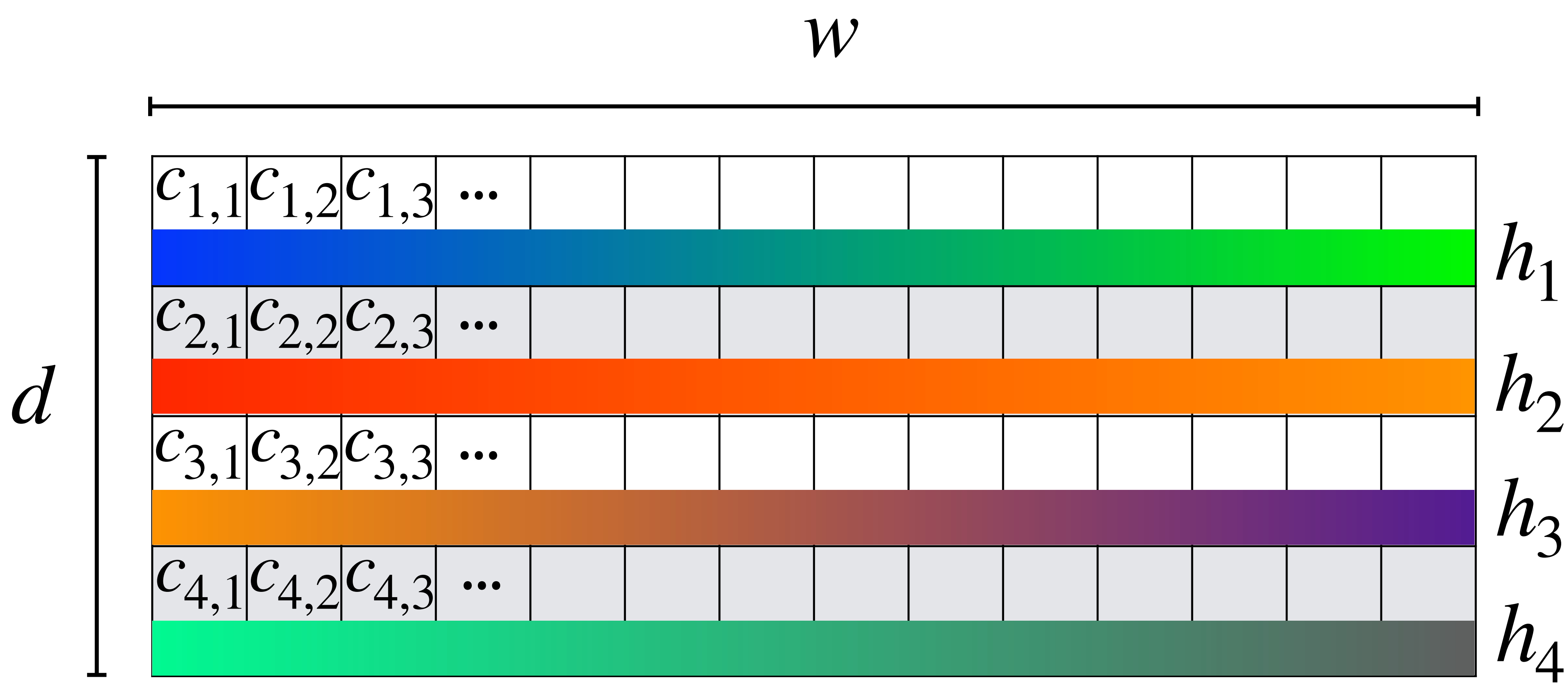


Query for item x returns ***minimum*** of the selected elements; call this estimate $\tilde{f}_x$

Collisions can make us overestimate f_x, but not underestimate; i.e. $\tilde{f}_x \geq f_x$

Can we argue $\tilde{f}_x$ is ***probably not too far*** from f_x?

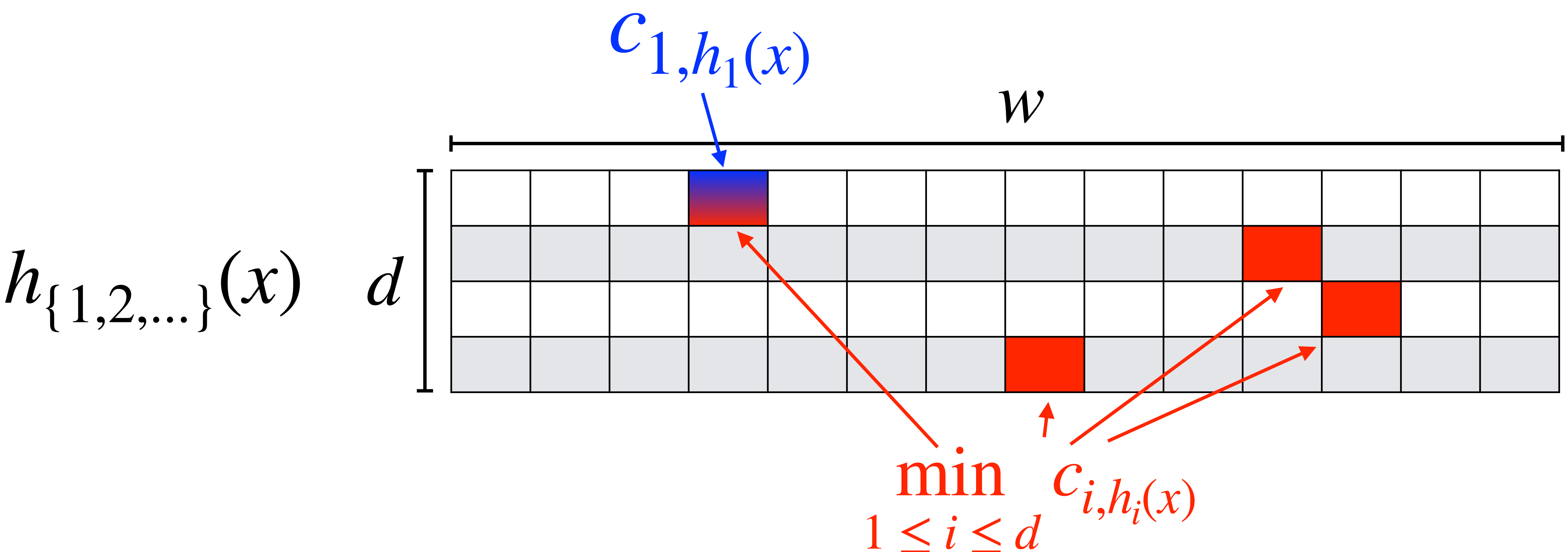
CountMin



We use functions $h_1, h_2, \ldots, h_d$ drawn from family H, each ranging over $\{1, 2, \ldots, w\}$

Let $c_{i,j}$ be the number of items x such that $h_i(x) = j$

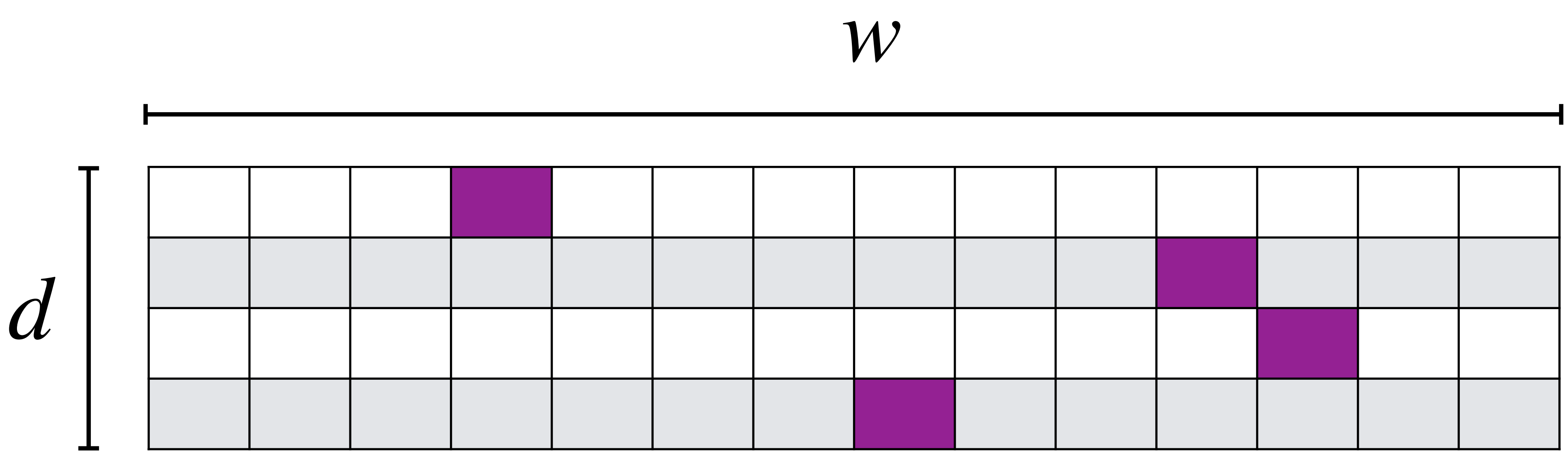
CountMin



What's the relationship between: f_x $c_{1,h_1(x)}$ $\min_{1 \le i \le d} c_{i,h_i(x)}$

$$f_x \le \min_{1 \le i \le d} c_{i,h_i(x)} \le c_{1,h_1(x)}$$

CountMin



Recall m = # of items in stream

Claim: if $w = 2/\epsilon$ and $d = \log_2 \delta^{-1}$, then

$$\Pr\left(\tilde{f}_x \le f_x + \epsilon m\right) \ge 1 - \delta$$

$\tilde{f}_x$ is ***probably not too far*** from f_x

CountMin

Pick item x and define r.v.s $\{Z_1, Z_2, \ldots, Z_d\}$ such that $Z_i = c_{i,h_i(x)} - f_x$

Z_i is the amount we over-counted in row i due to collisions

Argument modeled on Andrew McGregor's notes at https://people.cs.umass.edu/~mcgregor/711S18/vectors-3.pdf

CountMin

For $i \in \{1, 2, \ldots, d\}$, $y \in \{\text{distinct items}\} \setminus \{x\}$

$$X_{i,y} = \begin{cases} 1 & \text{if } h_i(y) = h_i(x) \\ 0 & \text{otherwise} \end{cases}$$

Recall: Z_i is the amount we over-counted in row i due to collisions

$$Z_i = \sum_{y \neq x} \left(f_y \cdot X_{i,y} \right)$$

CountMin

$$\mathbf{E}[Z_i] = \mathbf{E}\left[\sum_{x \neq y} f_y \cdot X_{i,y}\right]$$

$$= \sum_{x \neq y} f_y \cdot \mathbf{E}\left[X_{i,y}\right]$$ Linearity of expectation

$$= \sum_{x \neq y} f_y \cdot \Pr\left(h_i(y) = h_i(x)\right)$$ Expectation of indicator

What would we like to use next? **2-universality**

CountMin

Further assume that family H from which h_i's were drawn is 2-universal

$$\sum_{x \neq y} f_y \cdot \Pr\left(h_i(y) = h_i(x)\right) \leq \sum_{x \neq y} f_y \cdot \frac{1}{w}$$

2-universality

$$\leq \frac{m}{w}$$

Expected per-row excess $\mathbf{E}[Z_i]$ is at most m/w

CountMin

Z_i is a non-negative r.v., so:

$$\Pr\left(Z_i \geq a\right) \leq \frac{\mathbf{E}[Z_i]}{a}$$

Markov inequality

Let $b = \dfrac{a}{\mathbf{E}[Z_i]}$

$$\Pr\left(Z_i \geq b \cdot \mathbf{E}[Z_i]\right) \leq \frac{1}{b}$$

CountMin

$$\Pr\left(Z_i \geq b \cdot \mathbf{E}[Z_i]\right) \leq \frac{1}{b}$$

Combine with $\mathbf{E}[Z_i] \leq \dfrac{m}{w}$:

$$\Pr\left(Z_i \geq \frac{bm}{w}\right) \leq \Pr\left(Z_i \geq b \cdot \mathbf{E}[Z_i]\right) \leq \frac{1}{b}$$

Continue with these

CountMin

$$\Pr\left(Z_i \geq \frac{bm}{w}\right) \leq \frac{1}{b}$$

Let $b = w\epsilon$ (# columns times error tolerance):

$$\Pr\left(Z_i \geq \epsilon m\right) \leq \frac{1}{w\epsilon}$$

Let $w = 2/\epsilon$ (# columns from our Claim):

$$\Pr\left(Z_i \geq \epsilon m\right) \leq \frac{1}{2}$$

CountMin

When $w = 2/\epsilon$, probability that "bad thing" happens is at most 1/2

$$\Pr\left(Z_i \geq \epsilon m\right) = \Pr\left(f_x + Z_i \geq f_x + \epsilon m\right) \leq \frac{1}{2}$$

We want an upper bound of δ, δ being small

So: Repeat (across rows) and take minimum

CountMin

$$\Pr\left(Z_i \geq \epsilon m\right) \leq \frac{1}{2}$$

$$\Pr\left(\forall_{1 \leq i \leq d}\ Z_i \geq \epsilon m\right) \leq \left(\frac{1}{2}\right)^d$$

Independence of uniform & independently chosen hashes

Recall we set $d = \log_2 \delta^{-1}$

$$\left(\frac{1}{2}\right)^d = 2^{-\log_2 \delta^{-1}} = 2^{\log_2 \delta} = \delta$$

CountMin

$$\Pr\left(\forall_{1 \le i \le d}\ Z_i \ge \epsilon m\right) \le \delta$$

Probability of the bad thing is $\le \delta$

complement

$$\Pr\left(\exists_{1 \le i \le d}\ Z_i < \epsilon m\right) \ge 1 - \delta$$

Prob. of good thing is $\ge 1 - \delta$

To get the good thing, take the minimum

CountMin

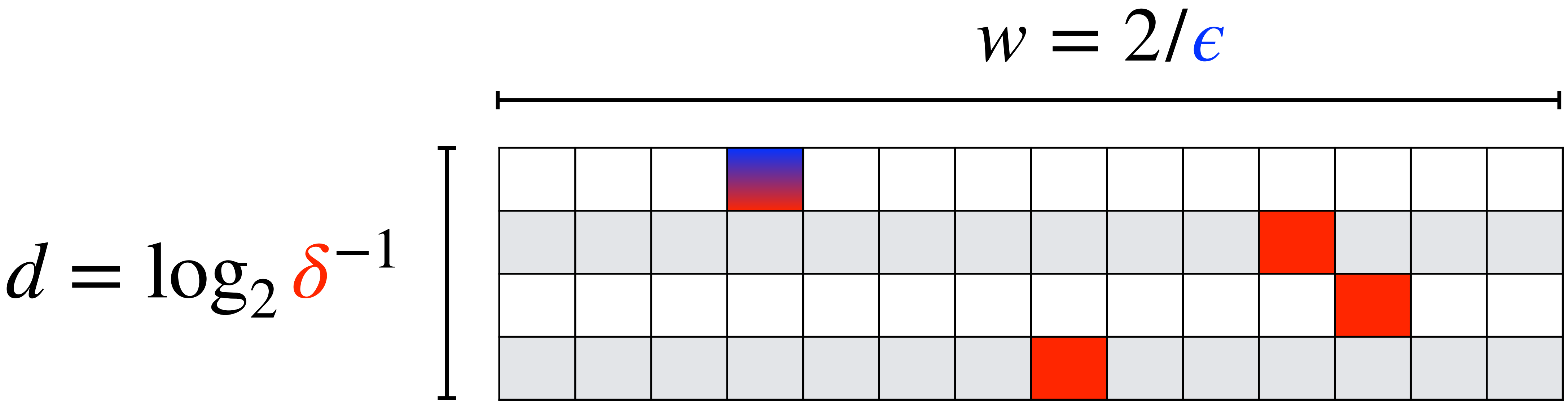
Claim is proved:

$$\tilde{f}_x = \min(c_{1,h_1(x)}, c_{2,h_2(x)}, \ldots, c_{d,h_d(x)})$$
$$= \min(f_x + Z_1, f_x + Z_2, \ldots, f_x + Z_d) \le f_x + \epsilon m$$

With probability $1 - \delta$ ✅

$\tilde{f}_x$ is ***probably*** ***not too far*** from f_x

CountMin



To achieve this, sketch must contain

$O(\epsilon^{-1} \log \delta^{-1})$ counters

CountMin

ϵ	δ	$\lceil (2/\epsilon) \rceil \cdot \lceil \log_2 \delta^{-1} \rceil$
10%	0.1	80
1%	0.01	1,400
0.1%	0.001	20,000
0.0001%	0.01	1,400,000

Remember that ϵ multiplies m, and a counter requires many (maybe 32 or 64) bits

CountMin

A common use of CountMin is to find ***heavy hitters***

Items with frequency over a threshold

$h_{\{1,2,3,4\}}(x)$

10	5	17	17	17	1	8	20 (+1)	18	14	8	5	20	13
15	4	4	19	20	0	8	17	15	19 (+1)	4	20	16	12
18	14	9	10	10	15	10	9	16	4	10 (+1)	18	20	10
13	3	6	18	8	19	1	15 (+1)	11	1	8	8	18	3

While adding items, maintain data structure containing items with point query result over the threshold

CountMin

$$\begin{bmatrix} & M & \end{bmatrix} \begin{bmatrix} \\ v \\ \\ \end{bmatrix} = \begin{bmatrix} Mv \end{bmatrix} \longrightarrow \text{answer}$$

Linear transformation interpretation

v is input data
M builds the sketch

Adapted from Andrew McGregor:
https://people.cs.umass.edu/~mcgregor/stocworkshop/mcgregor.pdf

CountMin

Result of applying h_1 to $x_1, x_2, \ldots, x_8$ — Row 1 of CountMin

$$\begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_{x_1} \\ f_{x_2} \\ f_{x_3} \\ f_{x_4} \\ f_{x_5} \\ f_{x_6} \\ f_{x_7} \\ f_{x_8} \end{bmatrix} = \begin{bmatrix} f_{x_3} + f_{x_4} \\ f_{x_1} + f_{x_5} + f_{x_7} \\ f_{x_2} + f_{x_6} + f_{x_8} \end{bmatrix}$$

Adapted from Andrew McGregor:
https://people.cs.umass.edu/~mcgregor/stocworkshop/mcgregor.pdf