Approximate Frequency:

Finding frequent items deterministically

soap, ipad

* Stream: $\sigma = \langle a_1, \ldots, a_j, a_m \rangle$, $a_i \in \{1, \ldots, n\}$

j: j^{th} token

m: length of stream

1, 2, 2, 2, 1, 10, 9, 1, 2, 3, 4, 5, . . .

* $f = (f_1, f_2, \ldots, f_n)$, $f_1 + \ldots + f_n = m$

Majority

$$\exists\, j : f_j > \frac{m}{2}$$

Frequent

$$\{ j : f_j > \frac{m}{k} \}$$

any one pass $\Rightarrow$ Majority requires at $\min\{m, n\}$ space

Misra-Gries Frequency Estimation

* Process σ
* Provide $\hat{f}_j$, $j \in \{1, \ldots, n\}$

Init : $A \leftarrow$ empty (Balanced Binary Search tree)

↑ approximate freer estimator

Process (token j)

if $j \in keys(A)$

$A[j] \leftarrow A[j] + 1$

else if $|keys(A)| \leq k-1$ then

$A[j] \leftarrow 1$

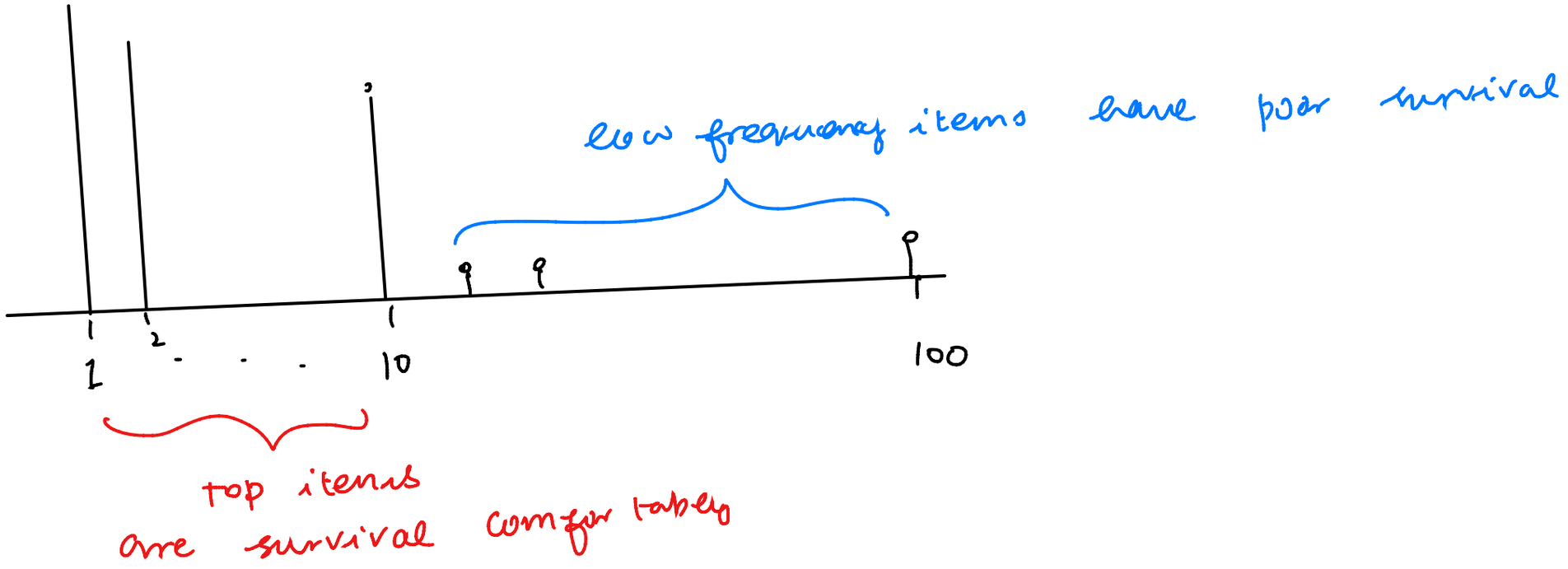
else (if $|keys(A)| = k$)

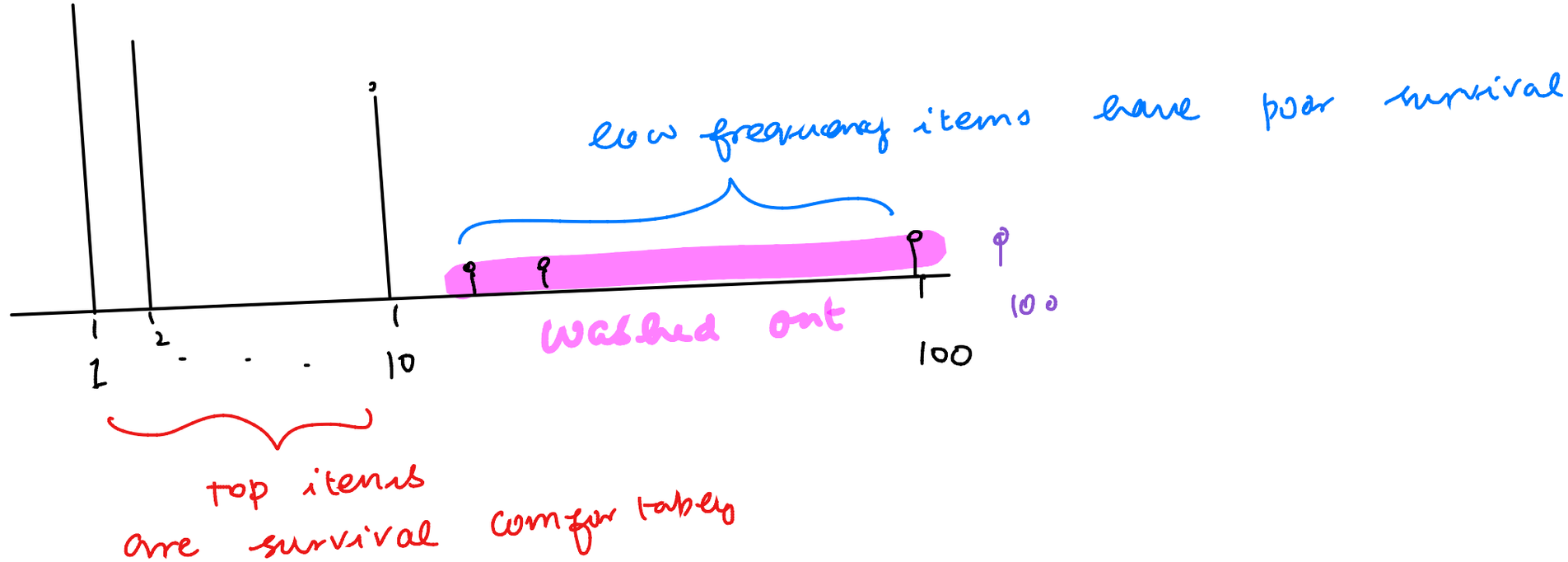
for each $\ell \in key(A)$

$A[\ell] \leftarrow A[\ell] - 1$ → decrement step

if $A[\ell] = 0$, remove it from A

top 10, $n = 10^6$, $k = 100$





$n = 10^6$, top items are soap, pen, mobile, laptop, pencil

$a_1, \ldots, a_m$

Soap Bag_1 = { 2, 30, 31, 52 }

Pen = { 1, 20, 25 }

mobile = { 3, 18, }

laptop = { 5 }

pencil = { 4, . . . }

Bag_k = { }

A[j]

on the decrement step I think that the first items in the bag drop out

* Items come into bag once, if they leave they do not come back into the bag

* When decrement happens, k items leave at once, one from each bag.

$\Rightarrow$ # decrements is at max $\frac{m}{k}$

$$f_j - \frac{m}{k} \leq \hat{f}_j \leq f_j$$

Typical scenario, $m = 10^6$ (↑ search queries), $k = 100$, top 10 are frequent

$$f \sim \frac{m}{10} \sim 10^5, \quad \frac{m}{k} = \frac{10^6}{100} = 10^4$$

$$10^5 - 10^4 \leq \hat{f}_j \leq 10^5$$

$f_i > \frac{m}{k} \Rightarrow \hat{f}_i > 0$

But not the other way around $\Rightarrow$ one can output in frequent items

Say, $k=2$, $m=5$, $S = \{a, b, c\}$

a, a, a, a, b

$\hat{f}_1 = 4$

$\hat{f}_2 = 1 \quad \not\Rightarrow$ item b is occuring more than $\frac{5}{2}$

a, a, b, c, a

$\hat{f}_1 = 1, 2, 2, 1, 2$

$\hat{f}_2 = -, -, 1, -, -$