Data Structures and Algorithms

Cardinality & MinHash Sketch (hopefully)

CS 225

Brad Solomon

May 1, 2026

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Department of Computer Science

Scalable Comparison of AI vs Manual Grading

CS 225 is exploring LLM-based grading tools for free-response questions

This is a research question — **your data will not be used without your explicit informed consent!**

2 points of extra credit will be awarded for filling out the survey regardless of whether you consent or decline to participate!

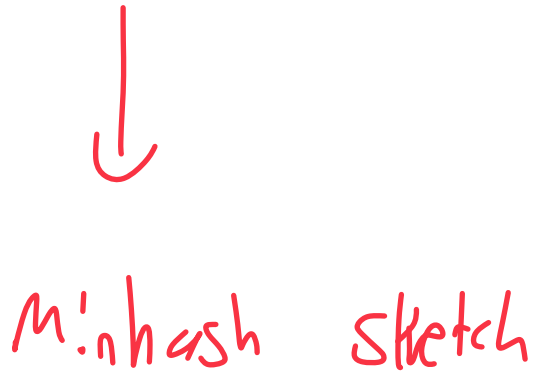
For details, please refer to the Prairielearn question available now.

The goal is to use AI for busywork tasks not replacing course staff!

Learning Objectives

Review bloom filters and identify the 'weakness' of BFs

Introduce the concept of cardinality and cardinality estimation



Bloom Filters

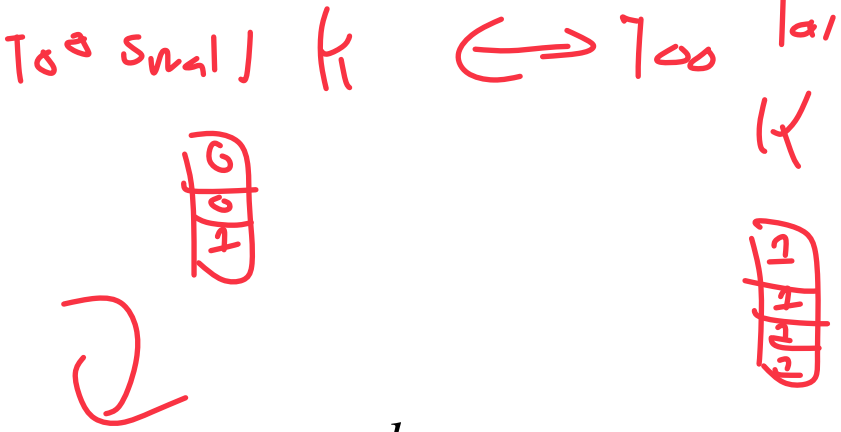
A probabilistic data structure storing a set of values

$h_{\{1,2,3,\ldots,k\}}$

Has three key properties:

- k, number of hash functions
- n, expected number of insertions
- m, filter size in bits

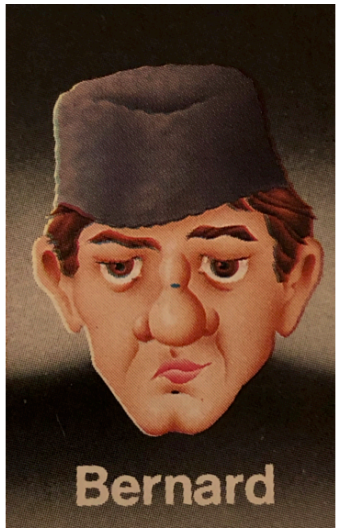
Too small k ⟷ Too large k

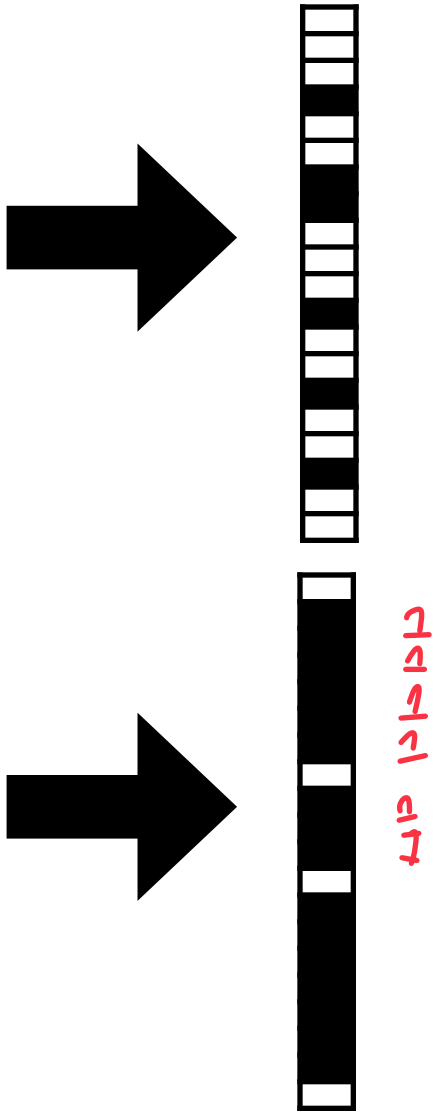


Expected false positive rate: $\left(1 - \left(1 - \frac{1}{m}\right)^{nk}\right)^{k} \approx \left(1 - e^{\frac{-nk}{m}}\right)^{k}$

Optimal accuracy when: $k^* = \ln 2 \cdot \frac{m}{n}$

The hidden problem with (most) sketches…





Seems saturated

Cover leads to many new strings

Cardinality

Cardinality is a measure of how many unique items are in a set

2
4
9
3
7
9
7
8
5
6

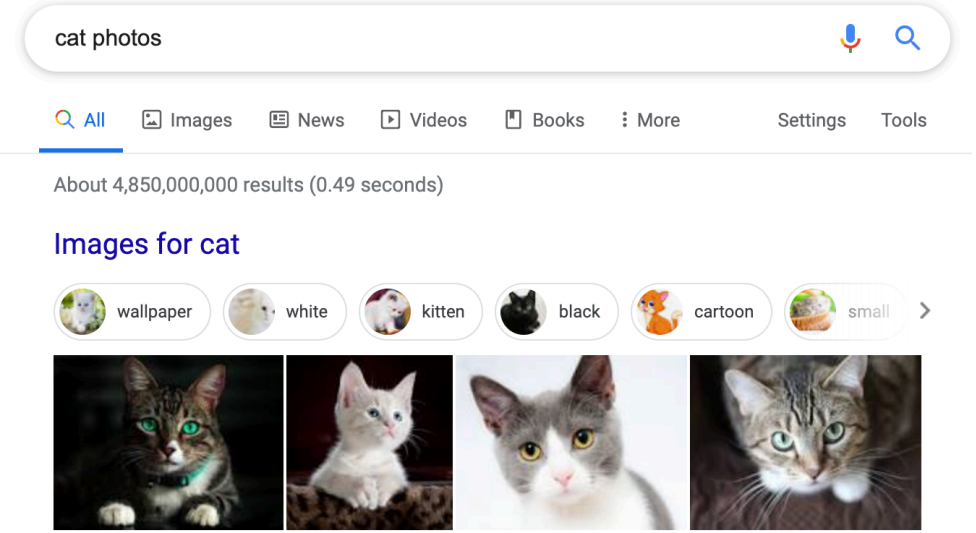
for each item

add item to set

size(set)

Cardinality

Sometimes its not possible or realistic to count all objects!



Estimate: 60 billion — 130 trillion

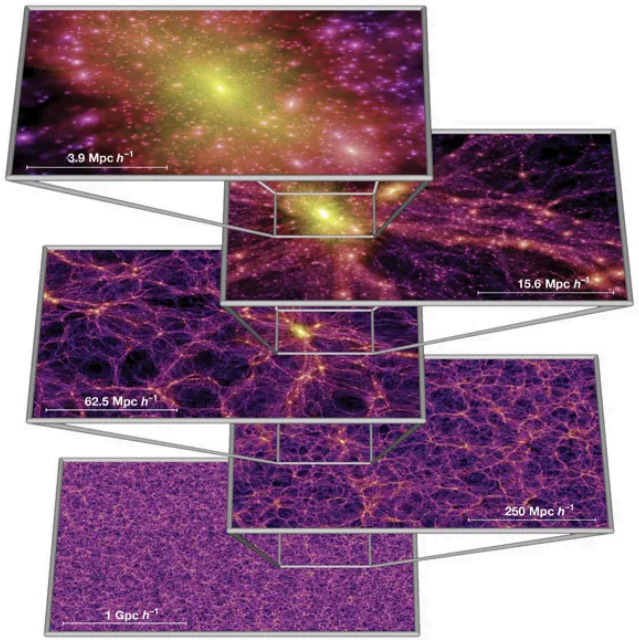


Image: https://doi.org/10.1038/nature03597

If cant compute exactly → estimate!

data stream

5581
8945
6145
8126
3887
8925
1246
8324
4549
9100
5598
8499
8970
3921
8575
4859
4960
42
6901
4336
9228
3317
399
6925
2660
2314

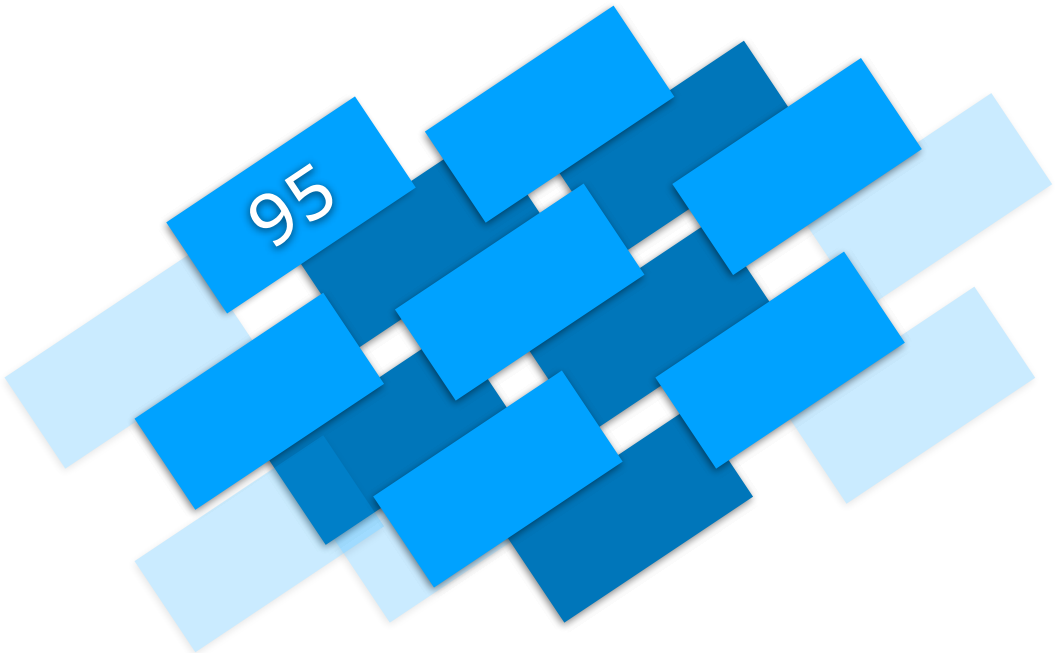
Cardinality Estimation

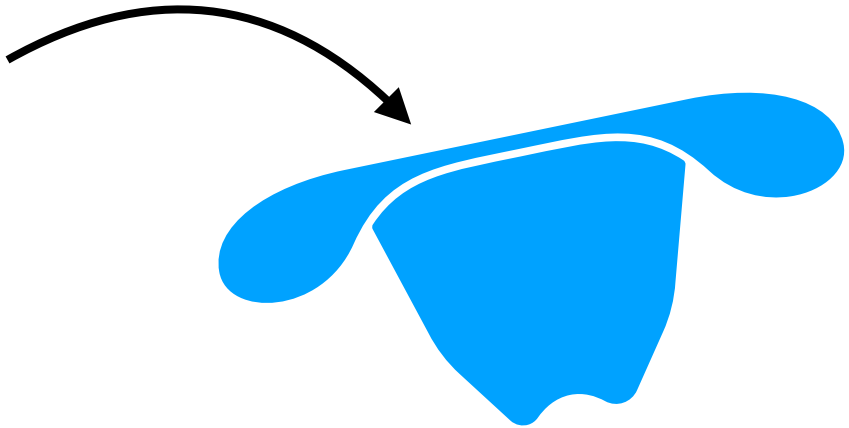
Imagine I fill a hat with numbered cards and draw one card out at random.

If I told you the value of the card was 95, what have we learned?

↳ 95 is in Set

↳ Not much else





Analogy from Ben Langmead

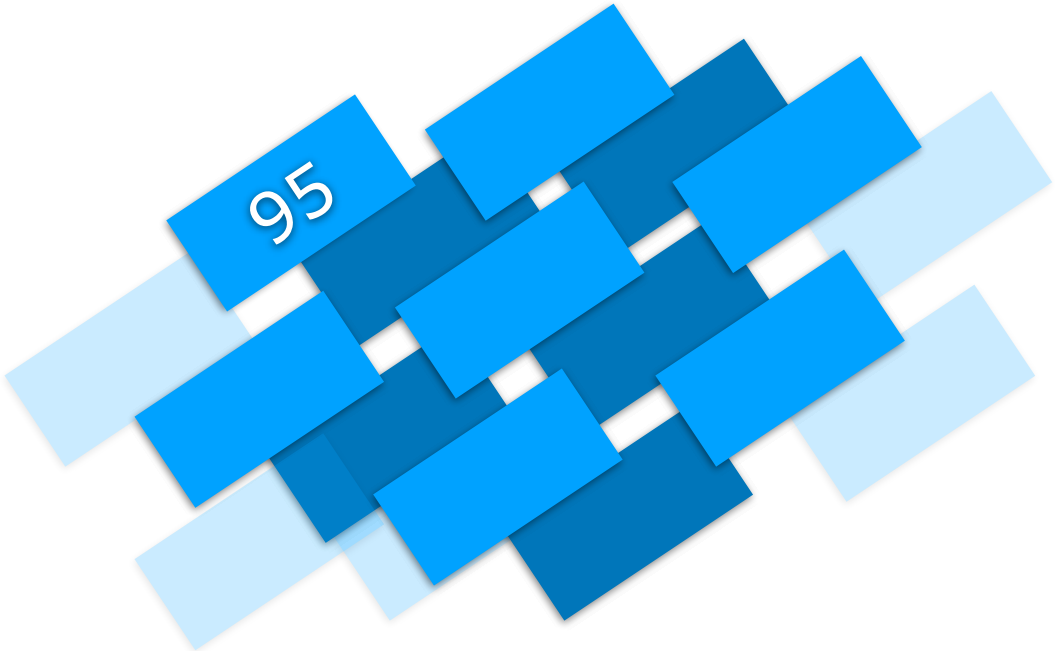
Cardinality Estimation

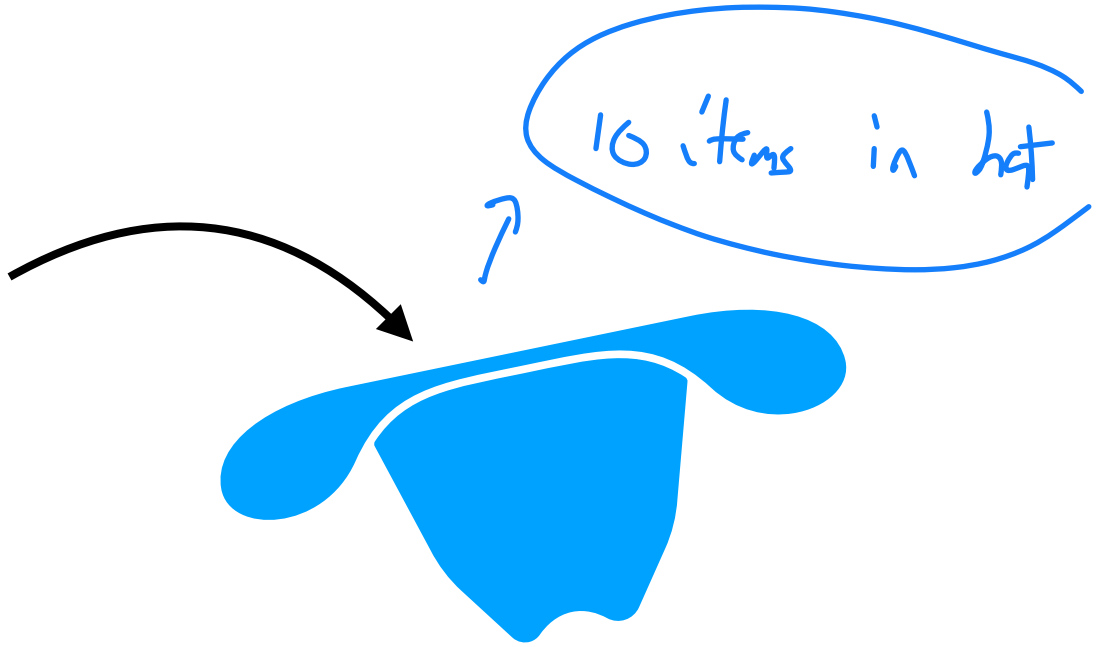
universe

Imagine I fill a hat with **a random subset** of numbered cards **from 0 to 999**

If I told you that the **minimum** value was 95, what have we learned?

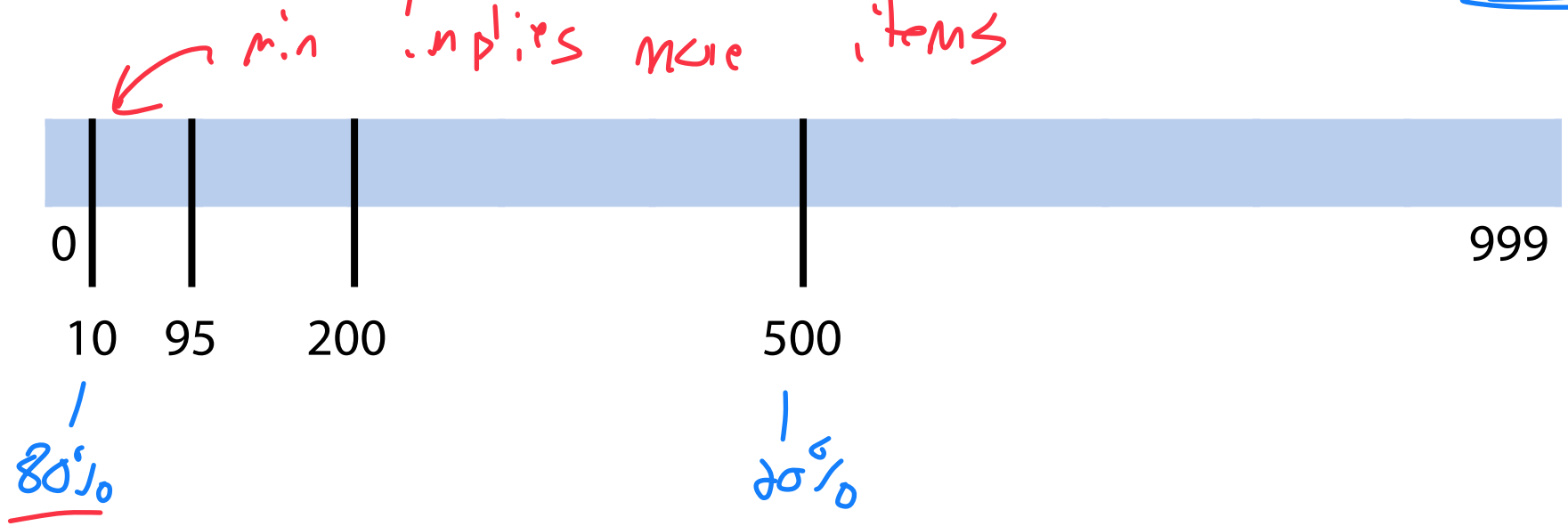
↳ we know 1 - 904 items in hat





Cardinality Estimation

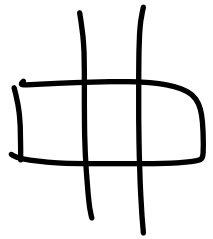
Imagine we have multiple uniform random sets with different minima.



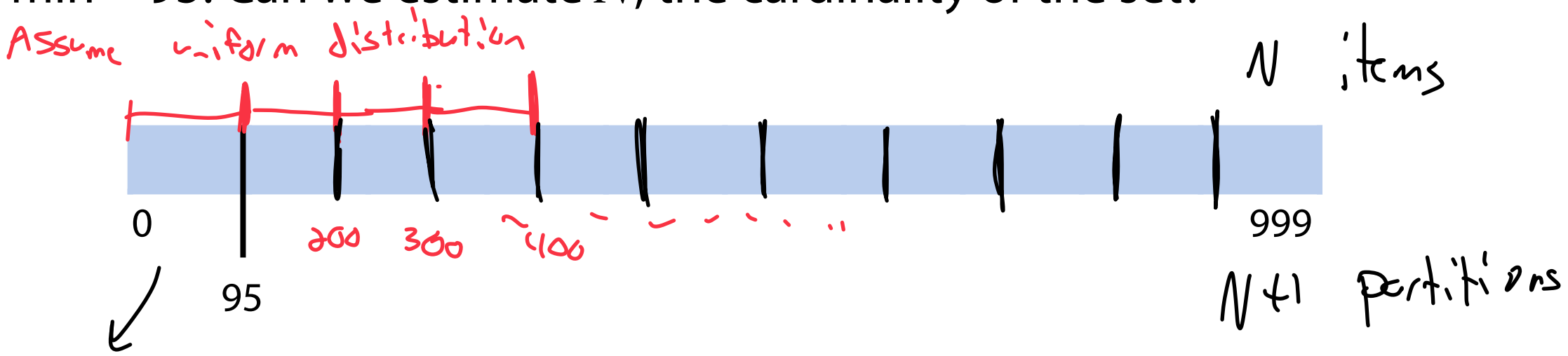
N=5

N=5

Cardinality Estimation



Let min = 95. Can we estimate N, the cardinality of the set?



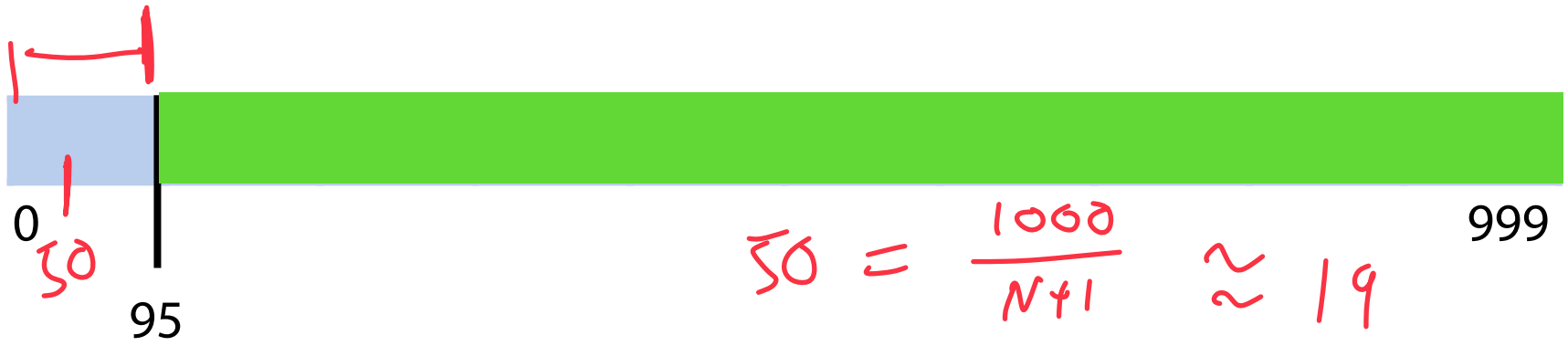
$$95 \cong \frac{1000}{N+1}$$

Cardinality Estimation

Let min = 95. Can we estimate N, the cardinality of the set?

Claim: $95 \approx \frac{1000}{(N+1)}$

Cardinality Estimation



Let min = 95. Can we estimate N, the cardinality of the set?

Conceptually: If we scatter N points randomly across the interval, we end up with $N + 1$ partitions, each about $1000/(N + 1)$ long

Assuming our first 'partition' is about average:

$$95 \approx 1000/(N+1)$$

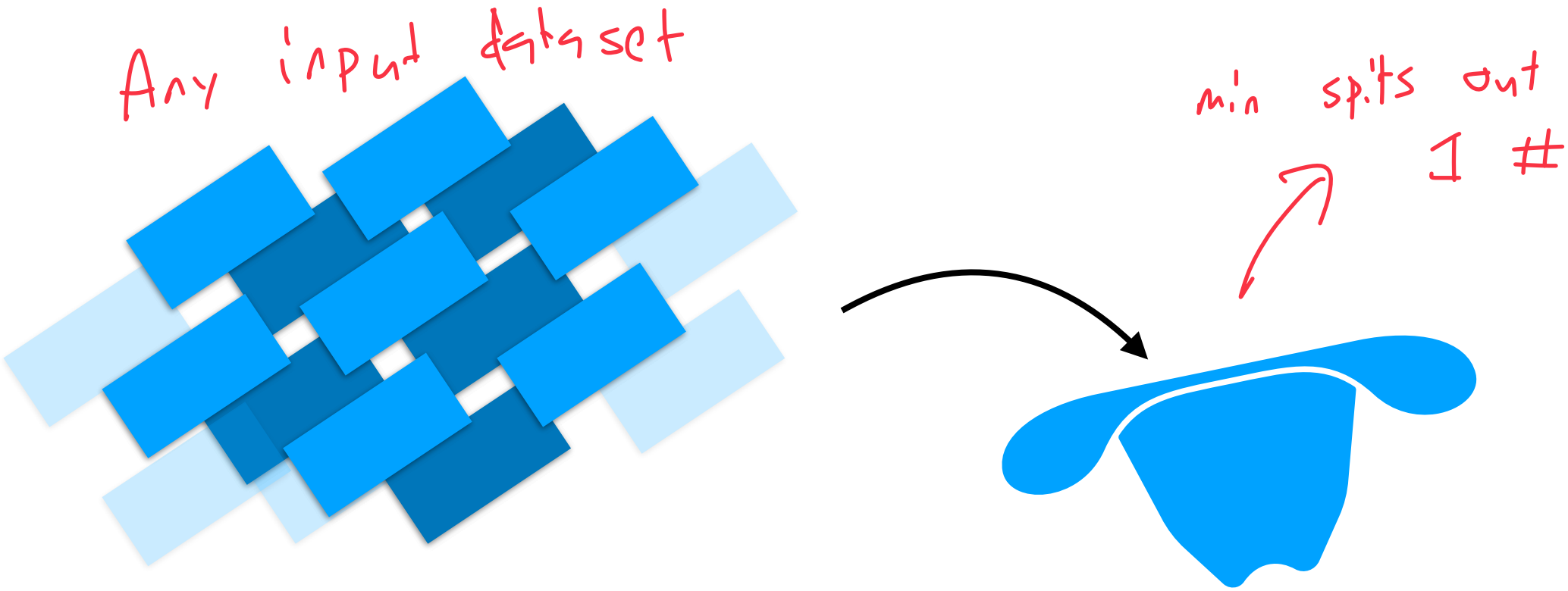
$$N+1 \approx 10.5$$

$$N \approx 9.5$$

→ 10 items in hat

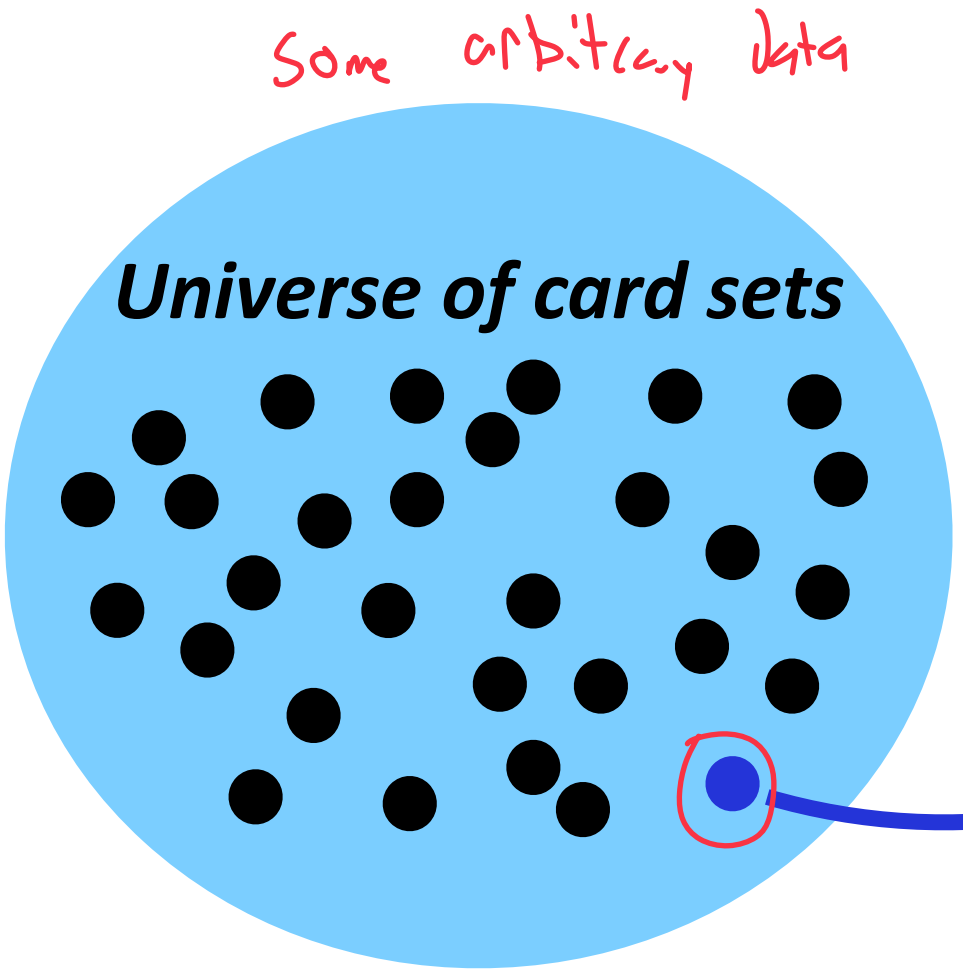
Cardinality Estimation

Why do we care about “the hat problem”?



Cardinality Estimation

Why do we care about "the hat problem"?



m possible minima

Key	Value
int address	

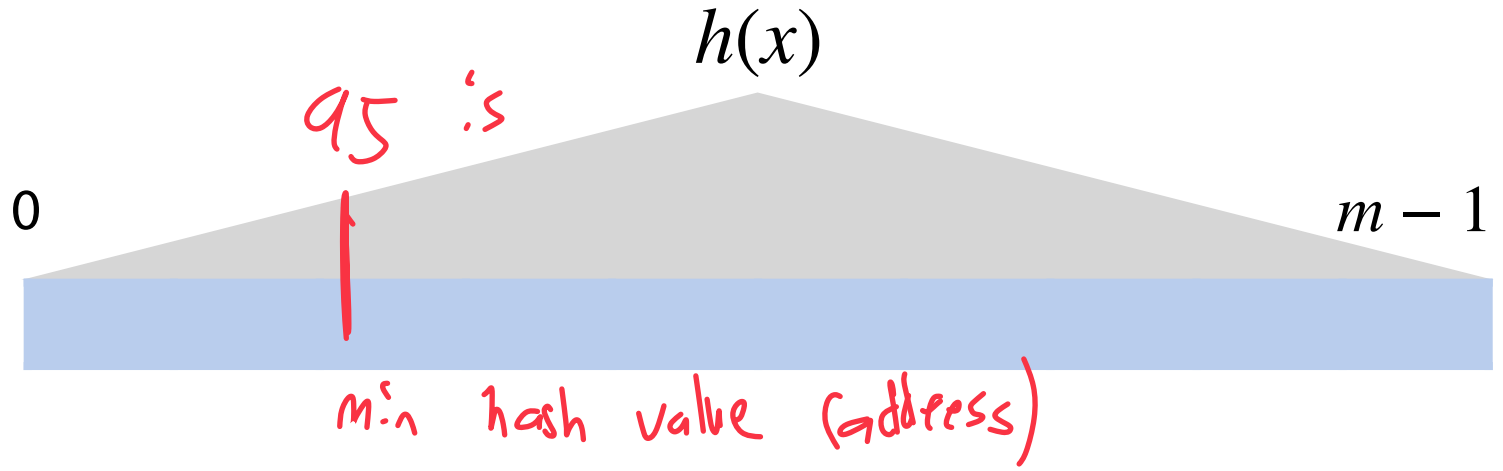
Cardinality Estimation

Imagine we have a SUHA hash h over a range m.

Inserting a new key is equivalent to adding a card to our hat!

Tracking only the minimum value is a **sketch** that estimates the cardinality!

lossy estimate of



Cardinality Estimation

Imagine we have a SUHA hash h over a range m.

Inserting a new key is equivalent to adding a card to our hat!

Tracking only the minimum value is a **sketch** that estimates the cardinality!

To make the math work out, lets normalize our hash…

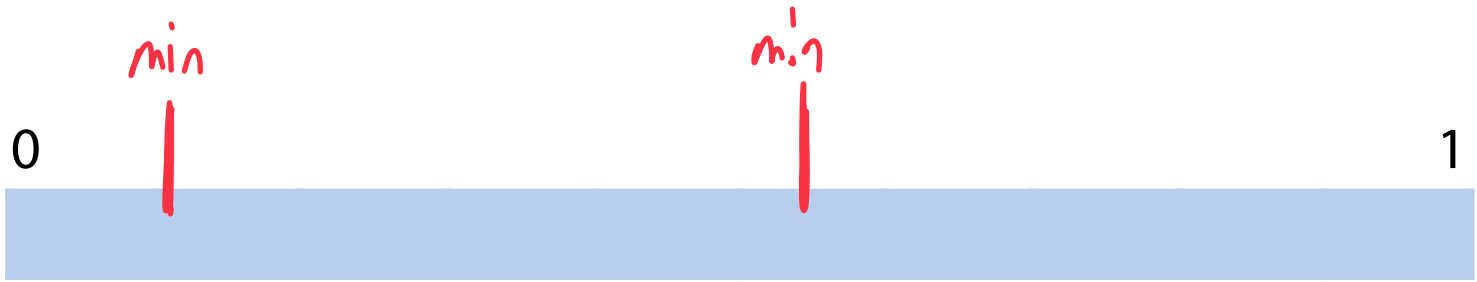
$$h'(x) = h(x) \ / \ (m - 1)$$

Cardinality Sketch

N the # of items

Let $M = min(X_1, X_2, \ldots, X_N)$ where each $X_i \in [0, 1]$ is an uniform independent random variable

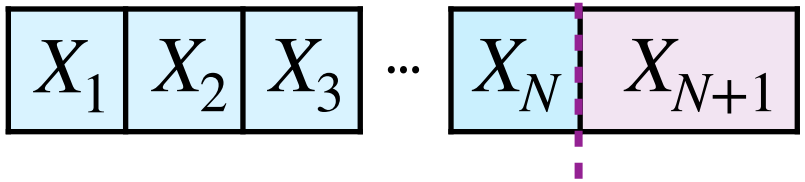
Claim: $\mathbf{E}[M] = \frac{1}{N+1}$



Cardinality Sketch

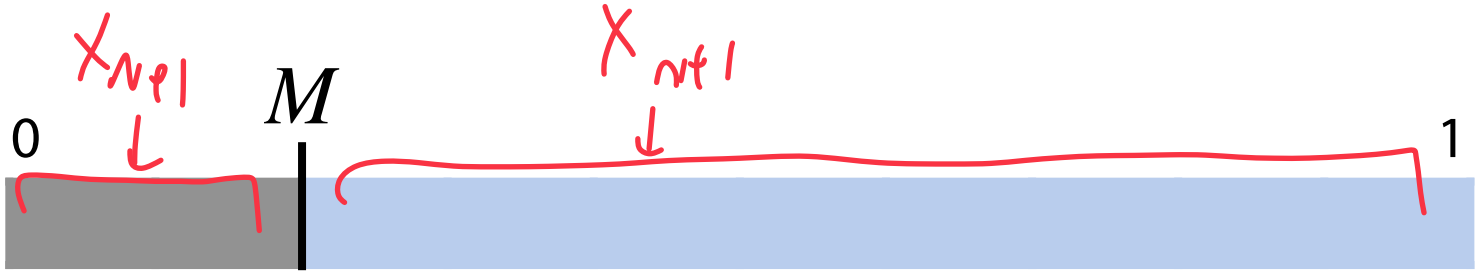
* tip for future

Consider an $N+1$ draw:



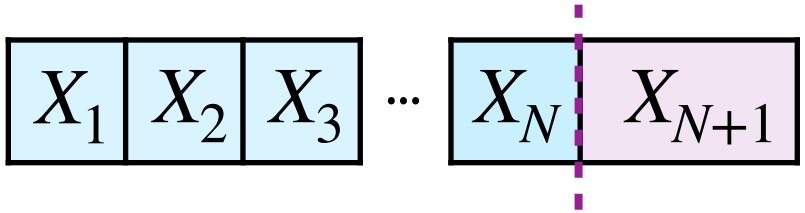
$$M = \min_{1 \le i \le N} X_i$$

X_{N+1} can end up in one of two ranges:



Cardinality Sketch

Consider an $N+1$ draw:

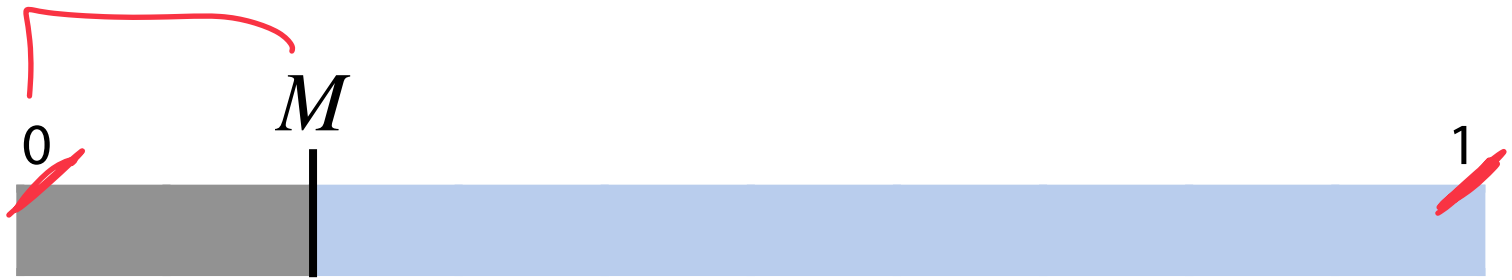


$$M = \min_{1 \le i \le N} X_i$$

X_{N+1} can end up in one of two ranges:

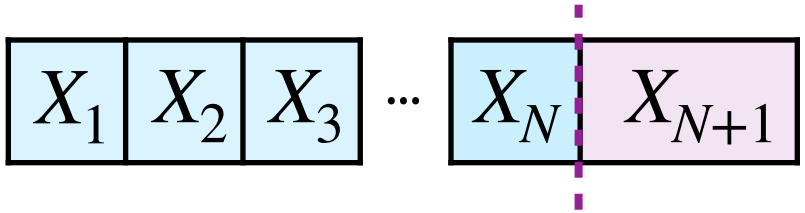
X_{N+1} will be the new minimum with probability M

uniform iid variable



Cardinality Sketch

Consider an $N + 1$ draw:



$$M = \min_{1 \le i \le N} X_i$$

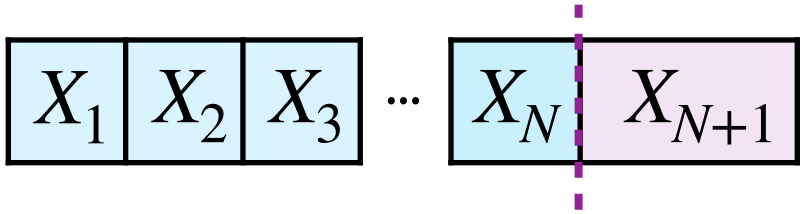
X_{N+1} can end up in one of two ranges:

X_{N+1} will be the new minimum with probability M

X_{N+1} will not change minimum with probability $1 - M$

Cardinality Sketch

Consider an $N+1$ draw:

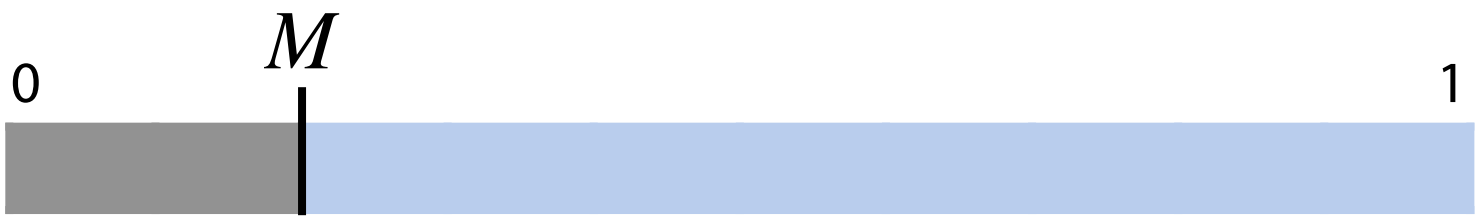


$$M = \min_{1 \le i \le N} X_i$$

X_{N+1} will be the new minimum with probability M

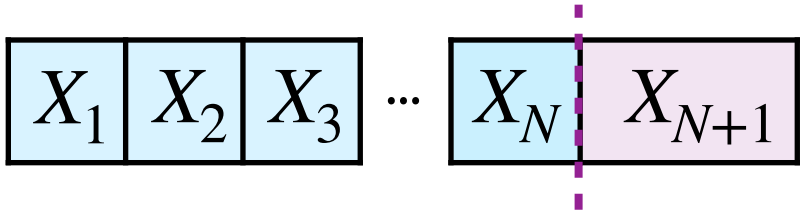
By definition of SUHA, X_{N+1} has a $\frac{1}{N+1}$ chance of being smallest item

All $N+1$ items equally likely to be min



Cardinality Sketch

Consider an $N+1$ draw:

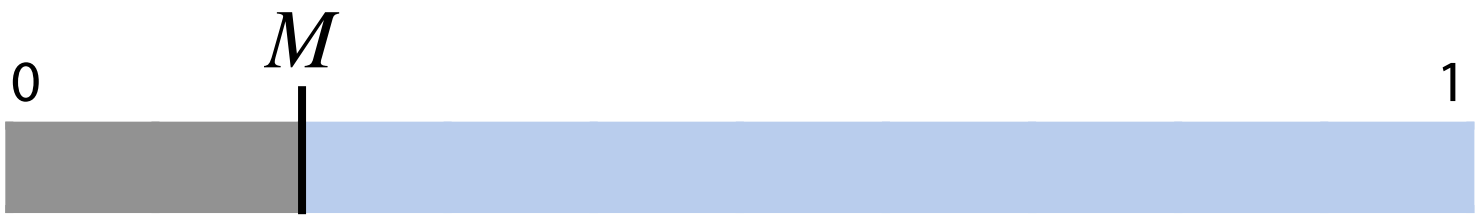


$$M = \min_{1 \le i \le N} X_i$$

X_{N+1} will be the new minimum with probability M

By definition of SUHA, X_{N+1} has a $\frac{1}{N+1}$ chance of being smallest item

Thus, $\mathbf{E}[M] = \frac{1}{N+1}$



Cardinality Sketch

Claim: $\mathbf{E}[M] = \frac{1}{N+1}$ $\Rightarrow$ $N \approx \frac{1}{M} - 1$

5 items!

Attempt 1

0.962	0.328	0.771	0.952	0.923

$N = 2.05$ ← underestimate

Attempt 2

0.253	0.839	0.327	0.655	0.491

$N = 2.95$ ← underestimate

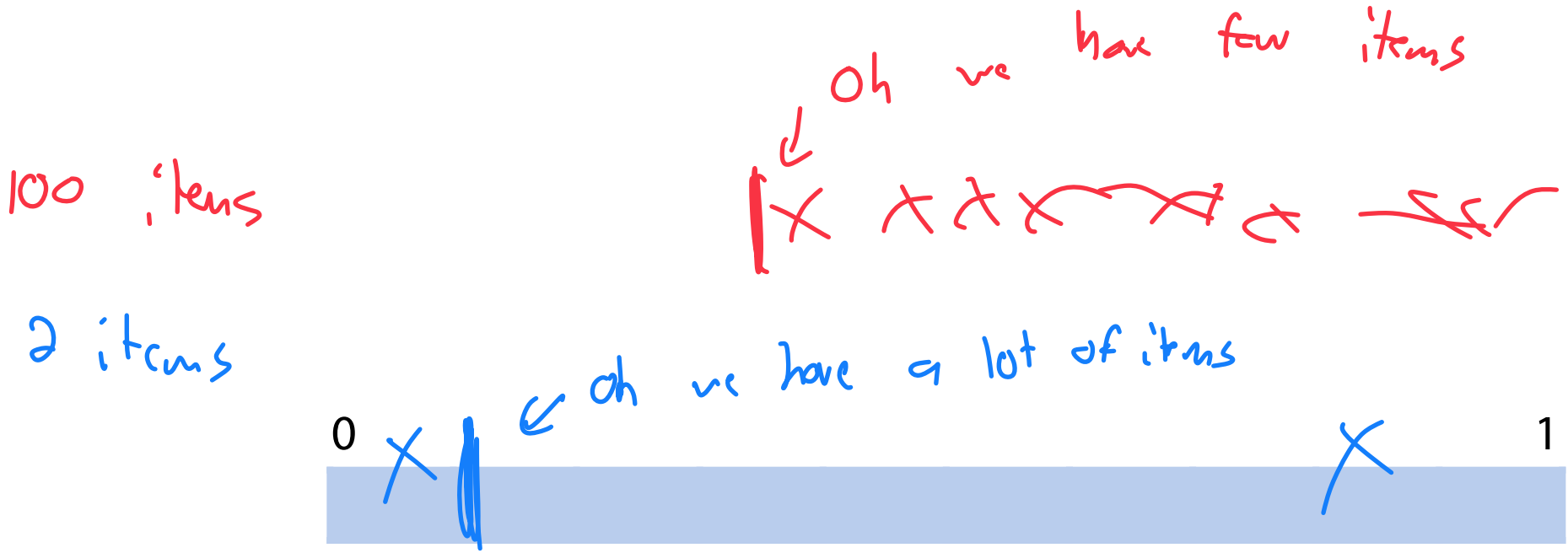
Attempt 3

0.134	0.580	0.364	0.743	0.931

$N = 6.5$ ← overestimate

Cardinality Sketch

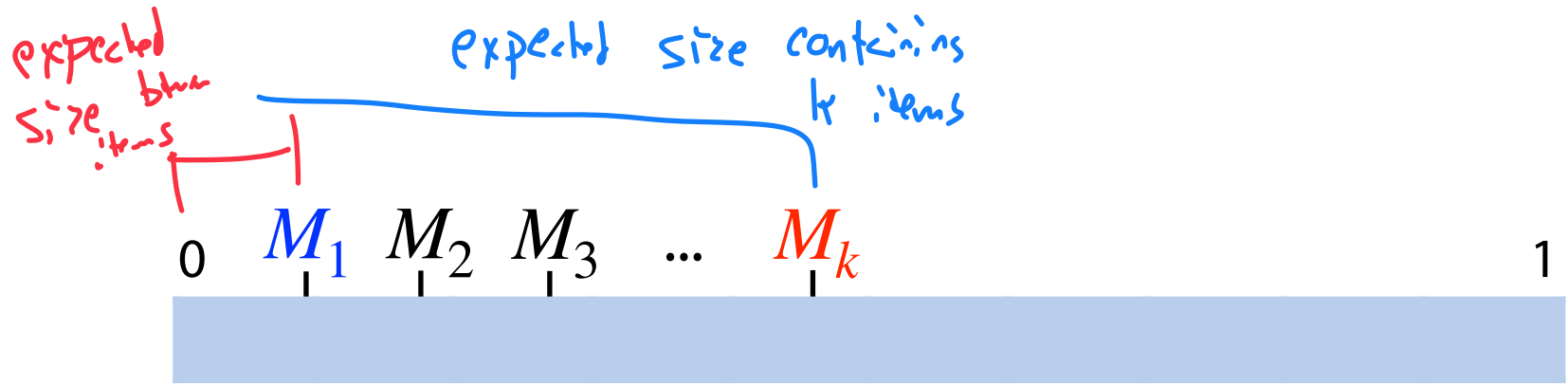
The minimum hash is a valid sketch of a dataset but can we do better?



Cardinality Sketch

Claim: Taking the k^{th}-smallest hash value is a better sketch!

Claim: $\mathbf{E}[M_k] = \dfrac{k}{N+1}$



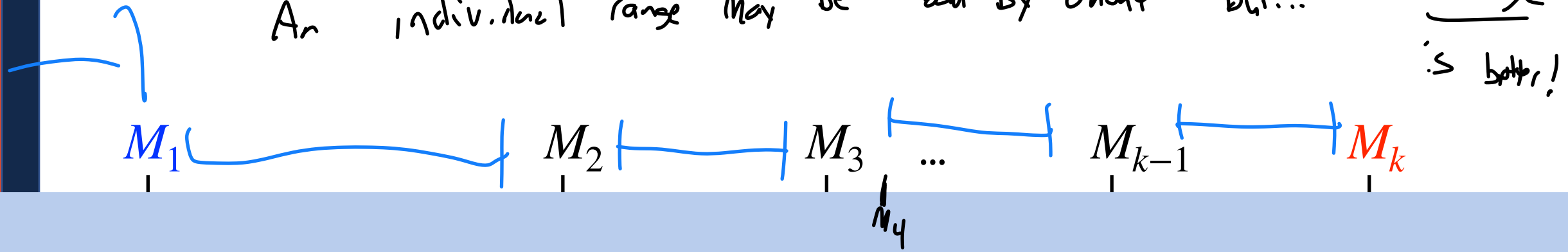
Cardinality Sketch

Claim: Taking the k^{th}-smallest hash value is a better sketch!

Claim: $$\frac{\mathbf{E}[M_k]}{k} = \frac{1}{N+1}$$

$$= \left[\mathbf{E}[M_1] + (\mathbf{E}[M_2] - \mathbf{E}[M_1]) + \ldots + (\mathbf{E}[M_k] - \mathbf{E}[M_{k-1}])\right] \cdot \frac{1}{k}$$

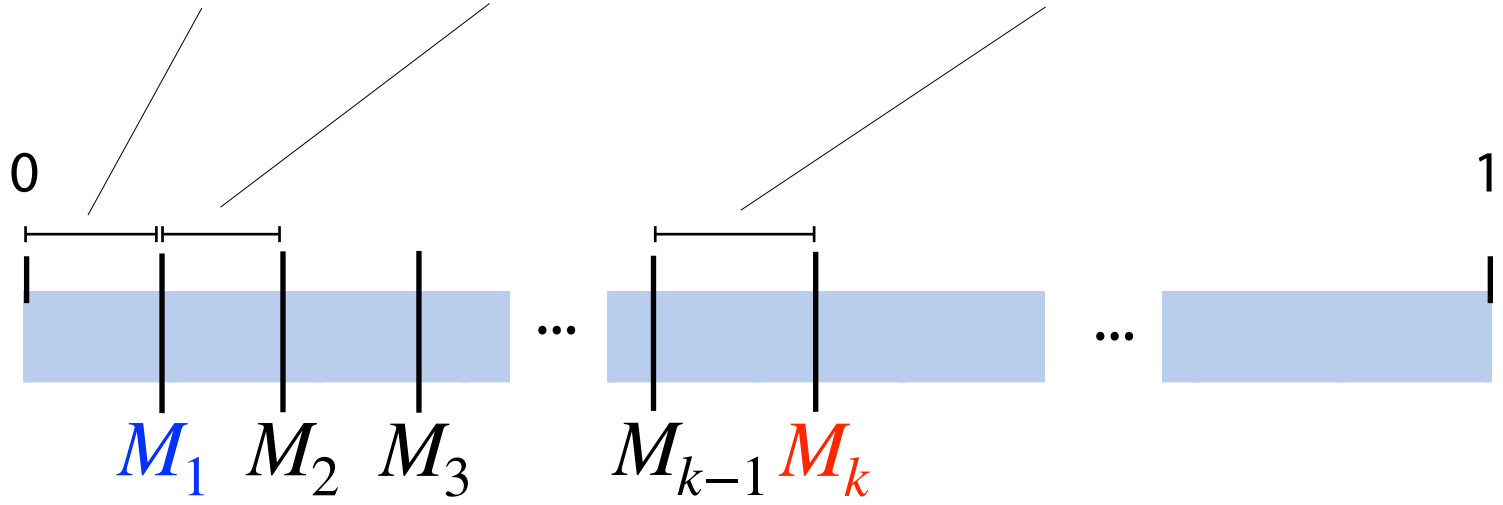
An individual range may be bad by chance but... the average is better!



Cardinality Sketch

$$\frac{1}{N+1} = \frac{\mathbf{E}[M_k]}{k}$$

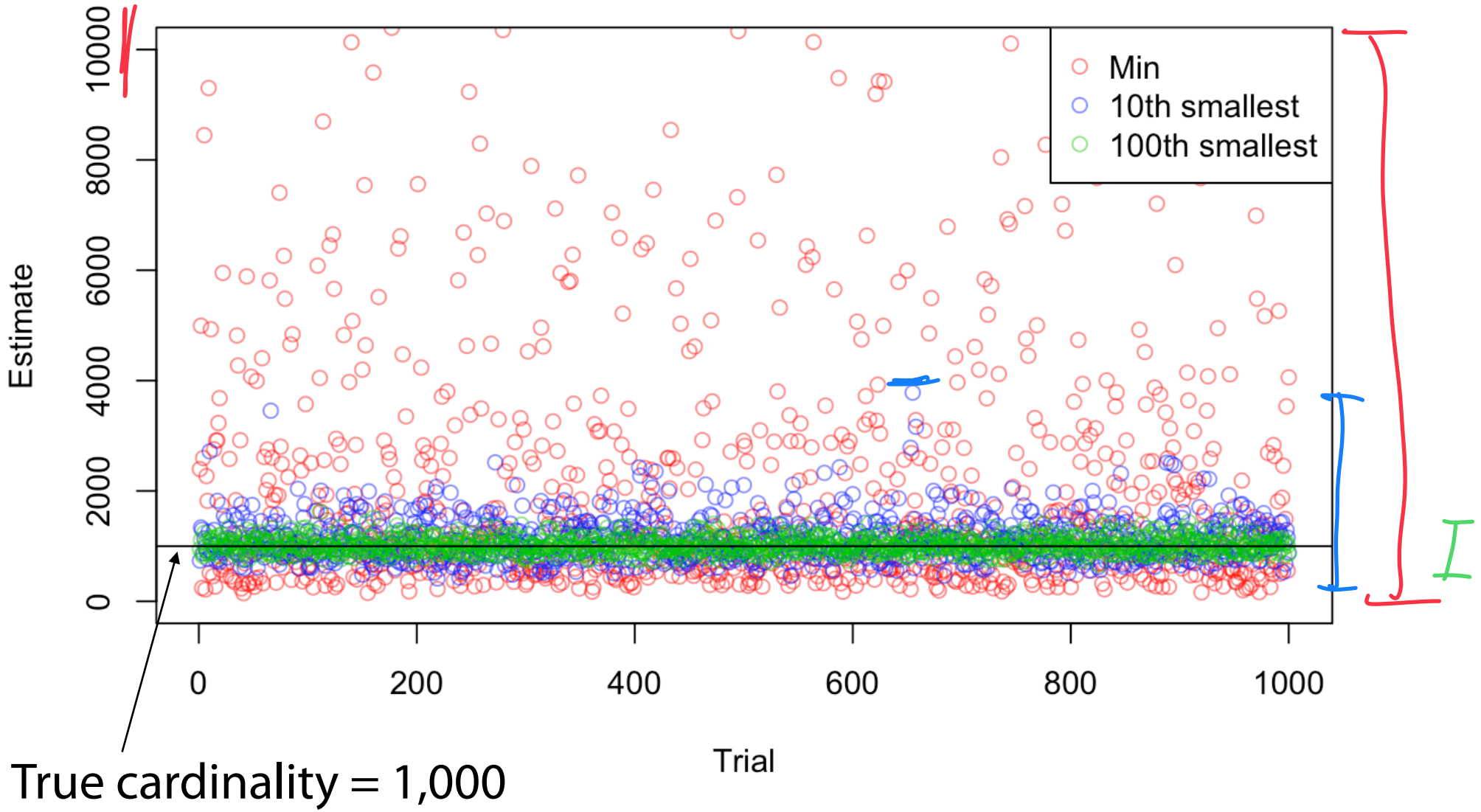
$$= \left[\mathbf{E}[M_1] + (\mathbf{E}[M_2] - \mathbf{E}[M_1]) + \ldots + (\mathbf{E}[M_k] - \mathbf{E}[M_{k-1}])\right] \cdot \frac{1}{k}$$



k^{th} minimum value (KMV)

Averages k estimates for $\frac{1}{N+1}$

Cardinality Sketch



Cardinality Sketch

Given any dataset and a SUHA hash function, we can **estimate the number of unique items** by tracking the **k-th minimum hash value**.

Hash

0.253	0.839	0.327	0.655	0.491

$\frac{3}{.491} - 1 \simeq 5.1$

$\frac{1}{.253} - 1 \simeq 3$

$\frac{2}{.327} - 1 \simeq 5.1$

To use the k-th min, we have to track k minima. **Can we use ALL minima?**

Tradeoff: higher k better estimate (more storage cost / slower runtime)

Applied Cardinalities

complex
slow

Cardinalities

$|A|$

$|B|$

$|A \cup B|$

$|A \cap B|$

Set similarities

$$O = \frac{|A \cap B|}{min(|A|, |B|)}$$

$$J = \frac{|A \cap B|}{|A \cup B|}$$

Real-world Meaning

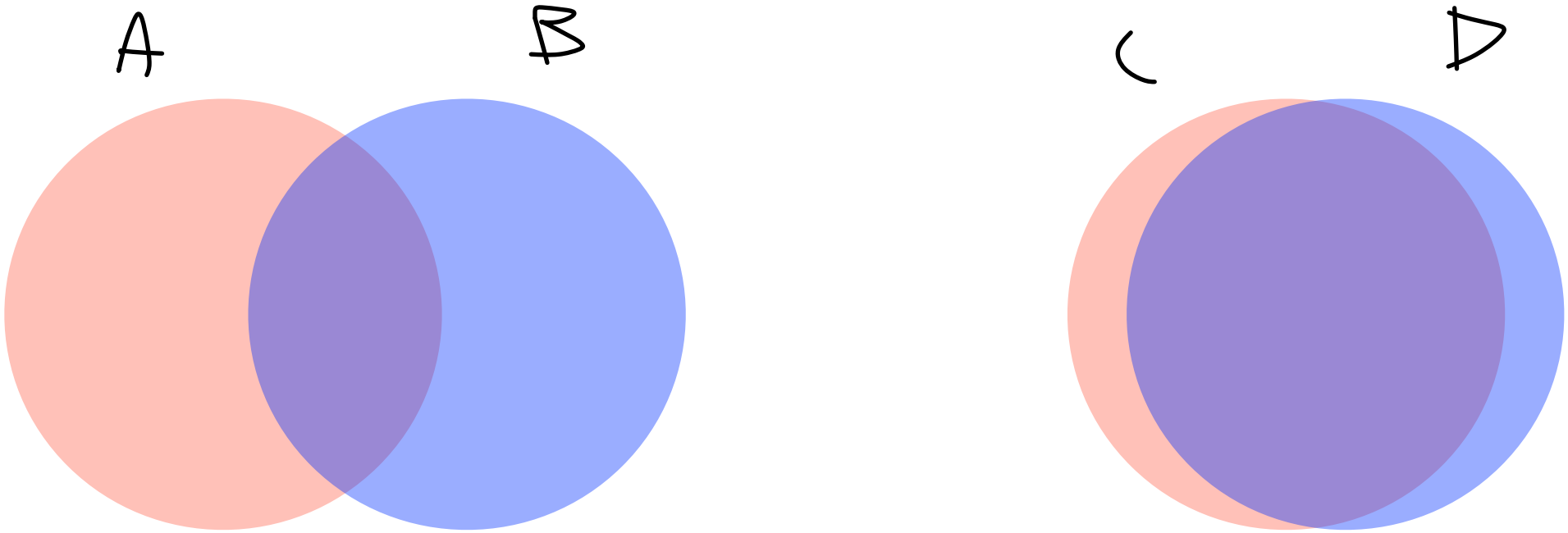
```
AGGCCACAGTGTATTATGACTG
|||||||||||  |||||||||
AGGCCACAGTGAGTTATGACTG

AAAAAAAAAAAGATGT-AAGTA
|||||||||||||||| |||||
AAAAAAAAAAAGATGTAAAGTA

GAGG--TCAGATTCACAGCCAC
||||  ||||||||||||||||
GAGGGGTCAGATTCACAGCCAC
```

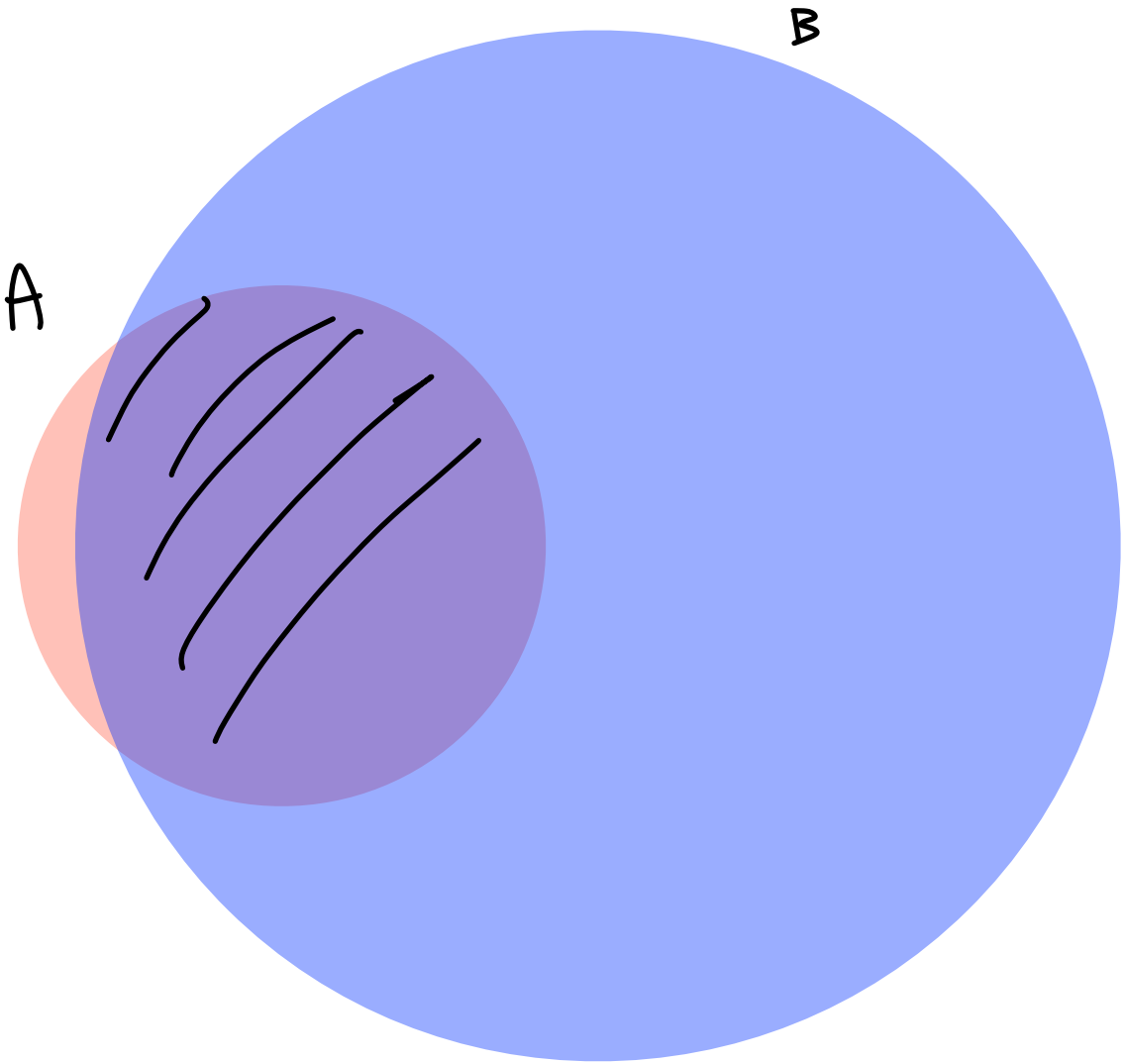
Set Similarity Review

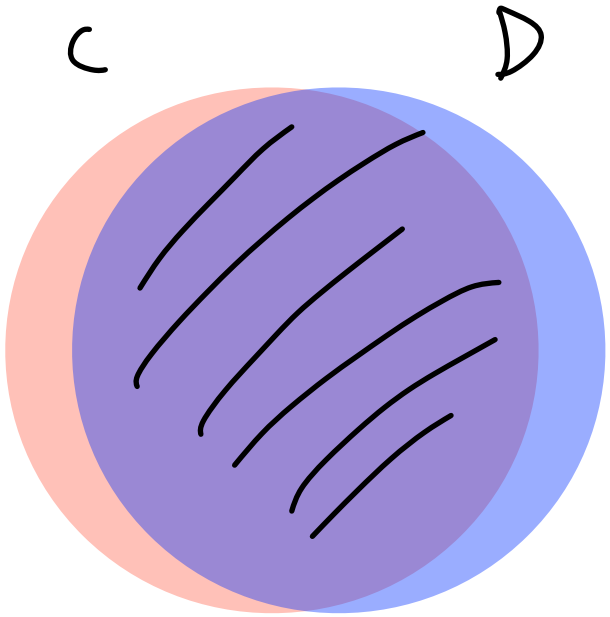
How can we describe how ***similar*** two sets are?



Set Similarity Review

How can we describe how ***similar*** two sets are?

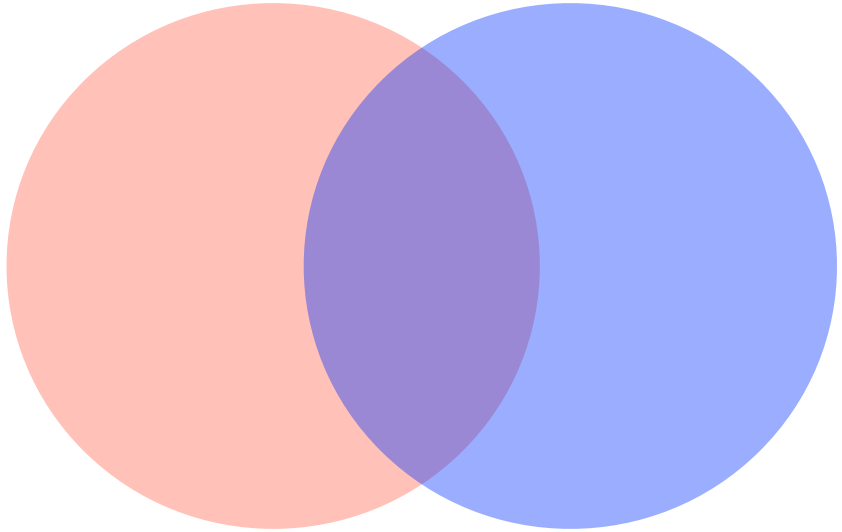




The intersection is the "same"

Set Similarity Review

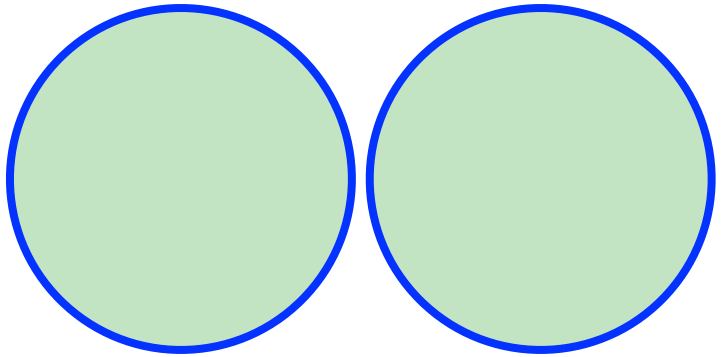
To measure **similarity** of A & B, we need both a measure of how similar the sets are but also the total size of both sets.



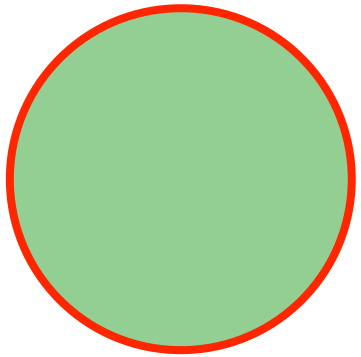
$$J = \frac{|A \cap B|}{|A \cup B|}$$

J is the ***Jaccard coefficient***

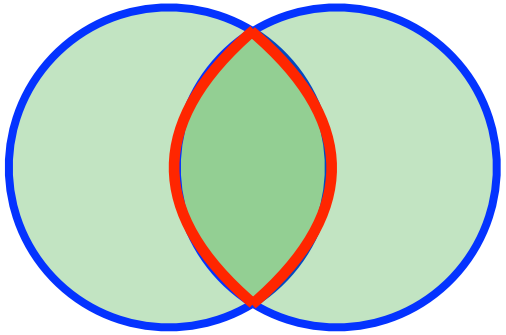
Set Similarity Review



$$\frac{|A \cap B|}{|A \cup B|} = 0$$



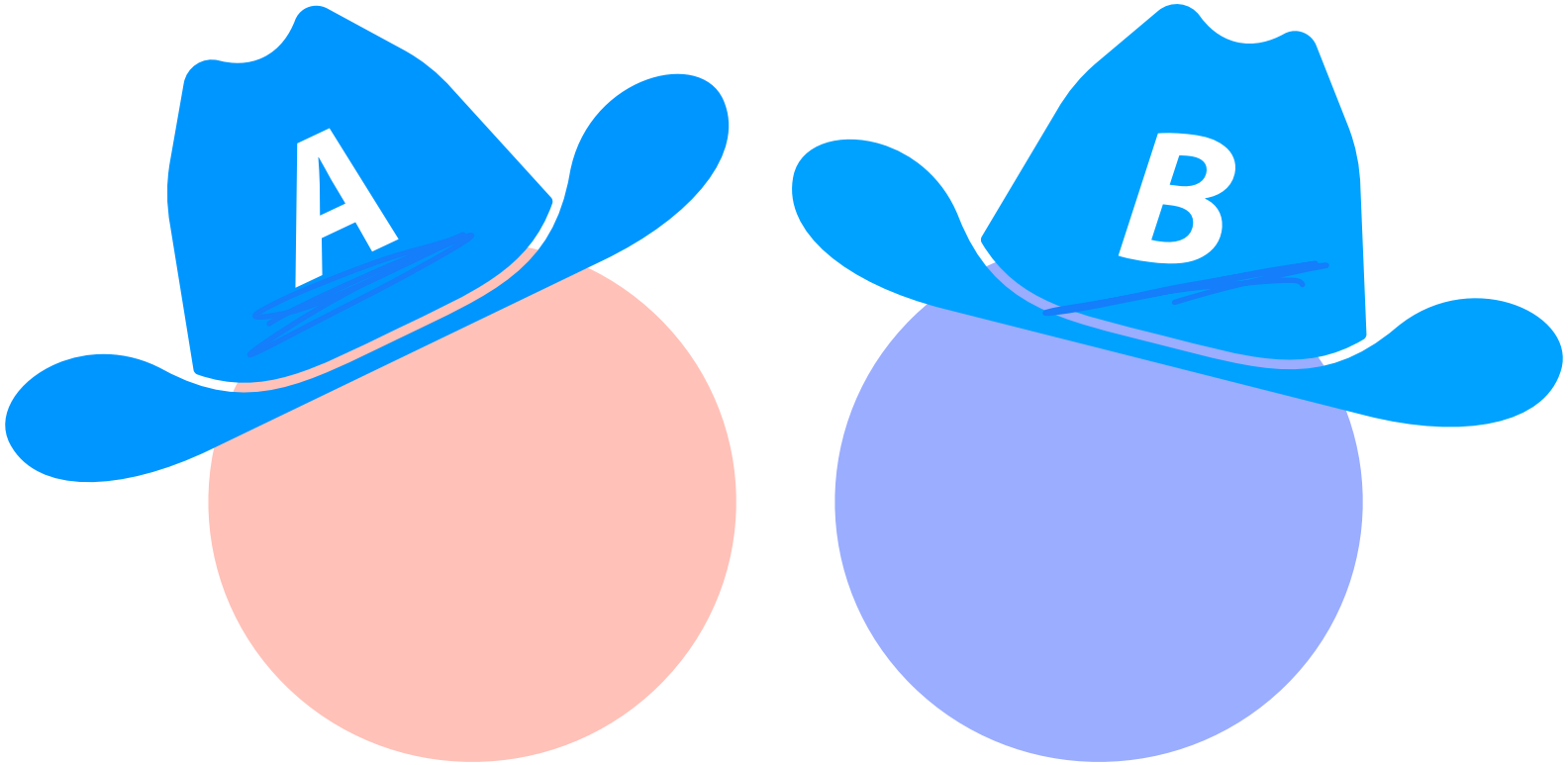
$$\frac{|A \cap B|}{|A \cup B|} = 1$$



$$0 < \frac{|A \cap B|}{|A \cup B|} < 1$$

Similarity Sketches

But what do we do when we only have a sketch?



Similarity Sketches

Imagine we have two datasets represented by their kth minimum values

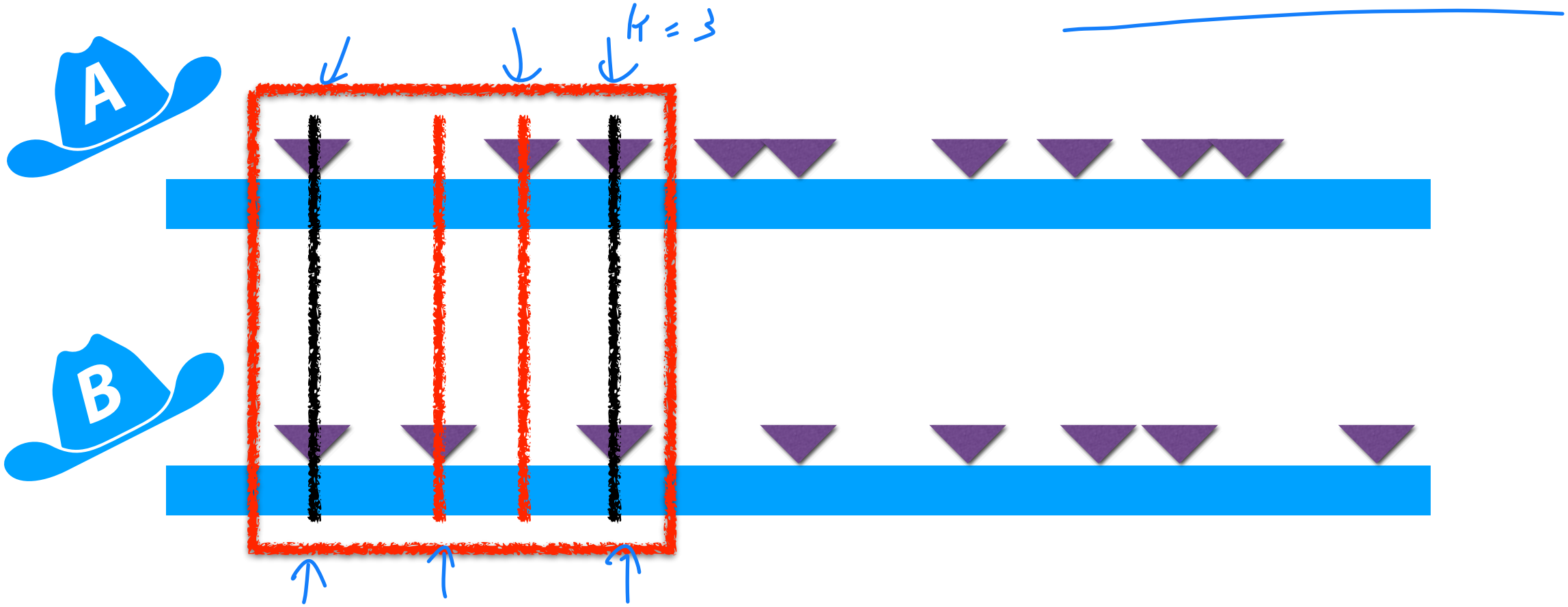


Image inspired by: Ondov B, Starrett G, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. **Mash Screen: high-throughput sequence containment estimation for genome discovery**. *Genome Biol* 20, 232 (2019)

Similarity Sketches

Claim: Under SUHA, set similarity can be estimated by sketch similarity!

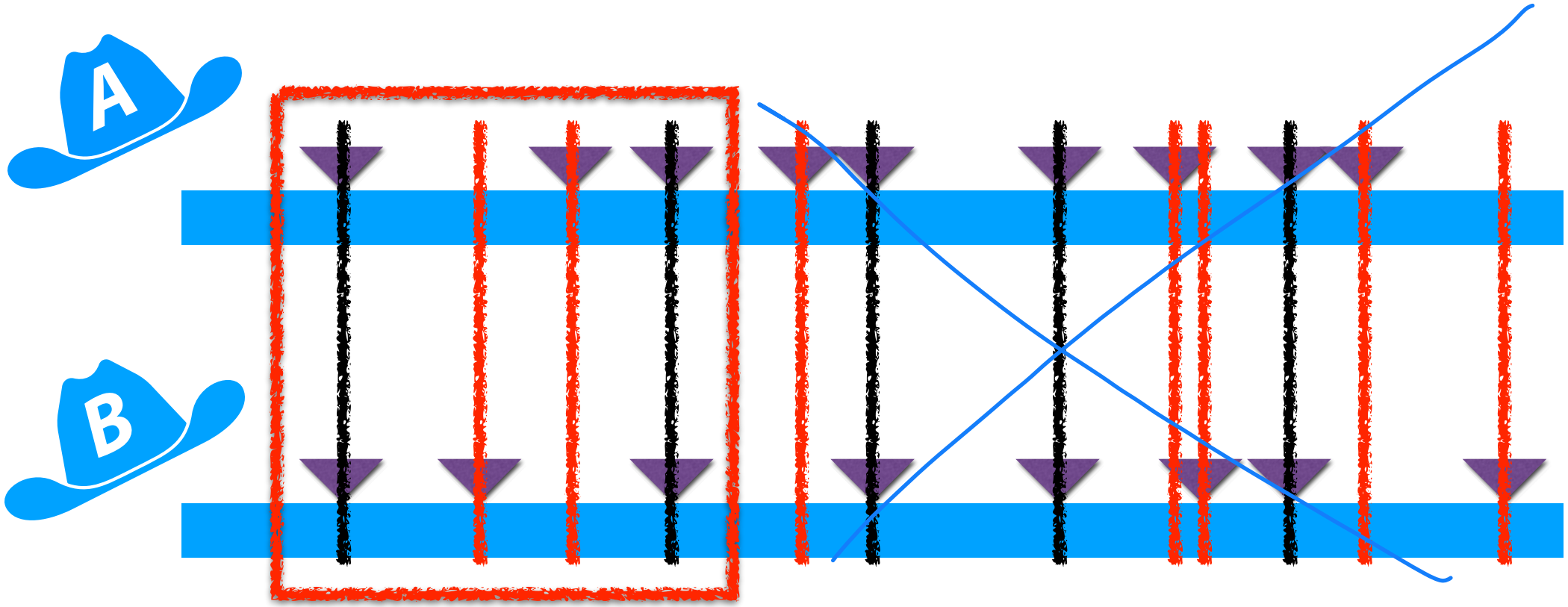
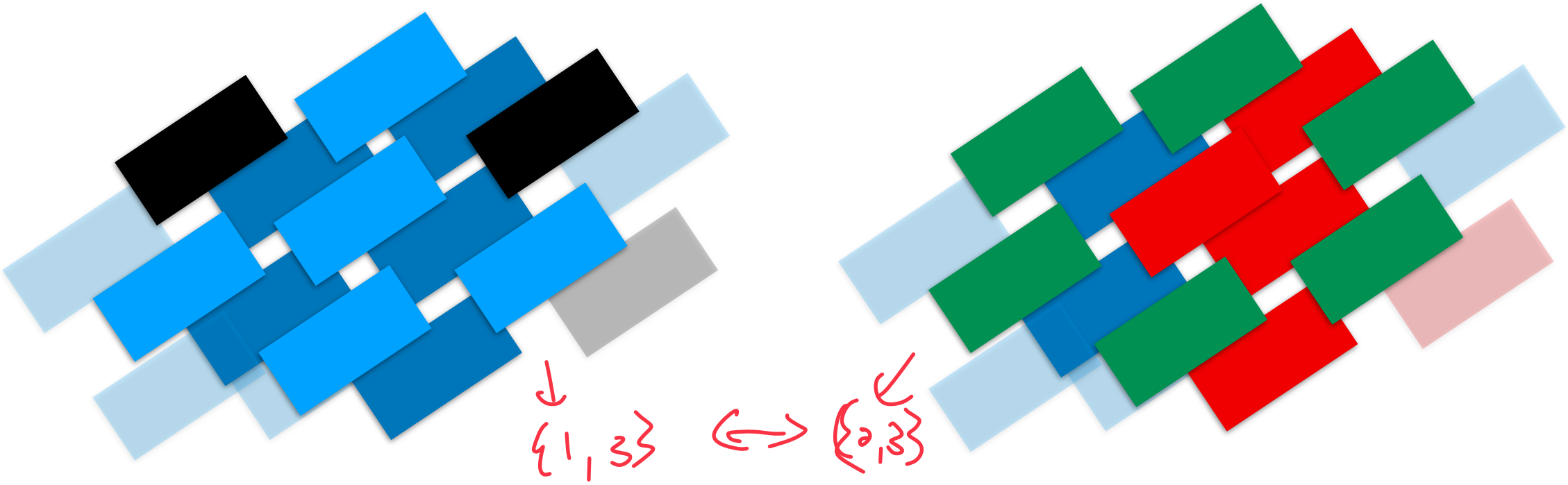


Image inspired by: Ondov B, Starrett G, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. **Mash Screen: high-throughput sequence containment estimation for genome discovery**. *Genome Biol* 20, 232 (2019)

MinHash Sketch

The **k-th minimum value sketch** is built by tracking k minima but only uses one value (the k-th minima) to get **cardinality!**

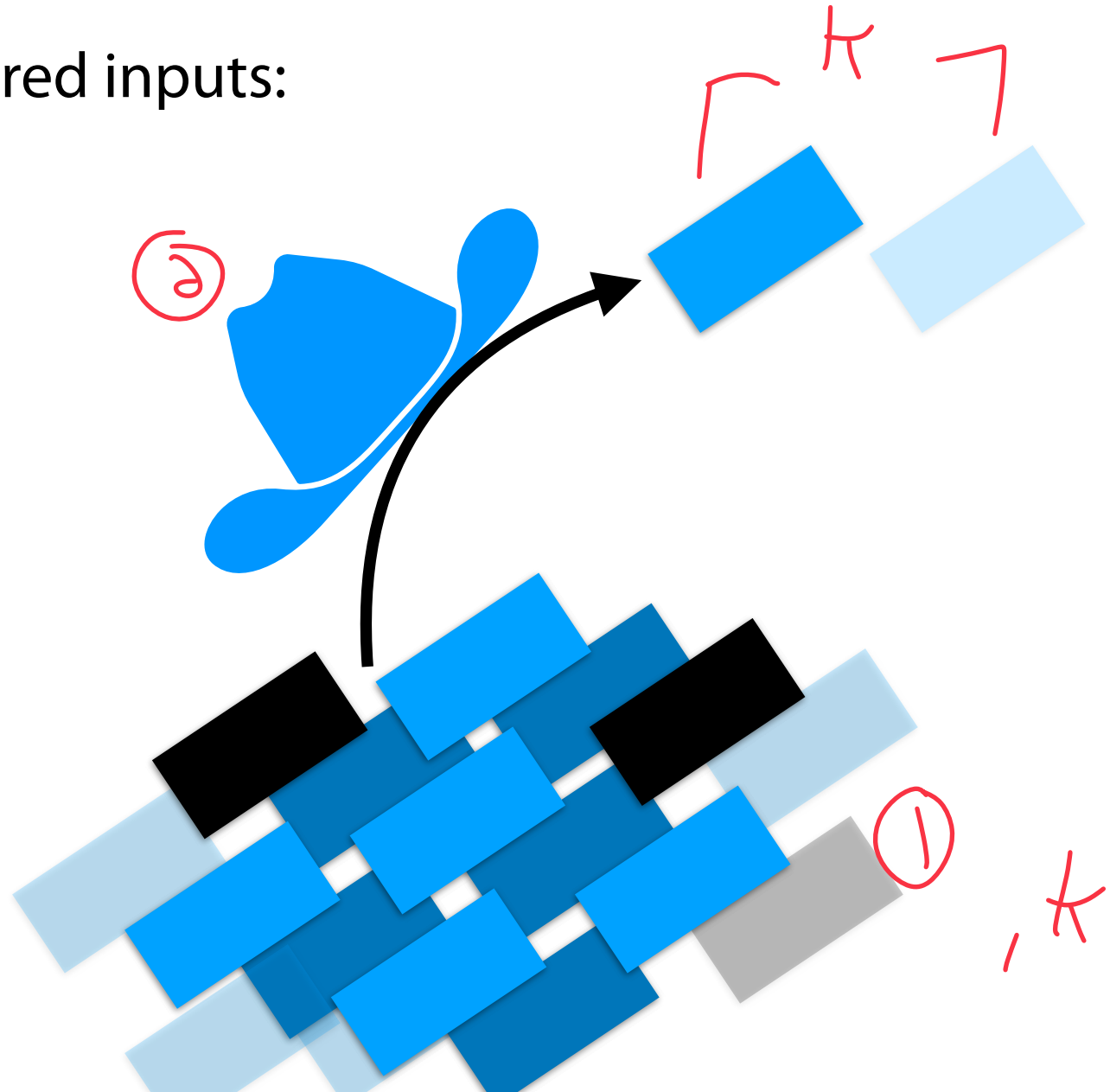
We can extend this approach into a full **MinHash sketch** that can also estimate **set similarities.**



MinHash Construction

A MinHash sketch has three required inputs:

1. Dataset
2. Hash Function
3. Some size k



MinHash Construction

S = { 16, 8, 4, 13, 15}

h(x) = x % 7

k = 3

Algorithm is trivial:

1. Hash each item
2. Keep the k-minimum values in memory
   (Ignore collisions / duplicates)

16 → 2

8 → 1

4 → 4

13 → 6 X

15 % 7 → 1

store hash values