Data Structures and Algorithms
Cardinality Sketches

CS 225
Brad Solomon

Dec 2, 2022

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Department of Computer Science

Reminder: Final Exam Scheduling

You can sign up now for the final exam

There are no extensions or make-ups for the final exam

Learning Objectives

Introduce the concept of cardinality and cardinality estimation

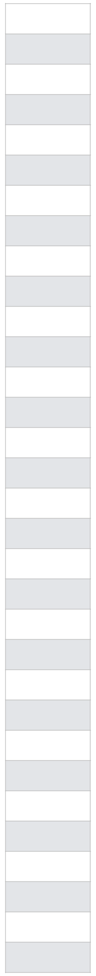
Demonstrate the relationship between cardinality and similarity

Introduce the MinHash Sketch for set similarity detection

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

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

Count-Min Sketch

A probabilistic data structure storing a set of values

Has **four** key properties:

- k, number of hash functions
- n, expected number of insertions
- m, filter size in ***registers***
- b**, number of bits per register**

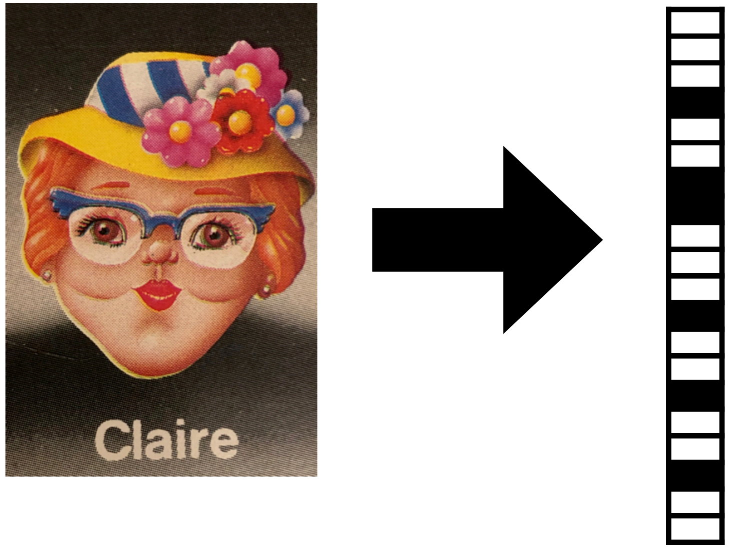
$h_{\{1,2,3,\ldots,k\}}$

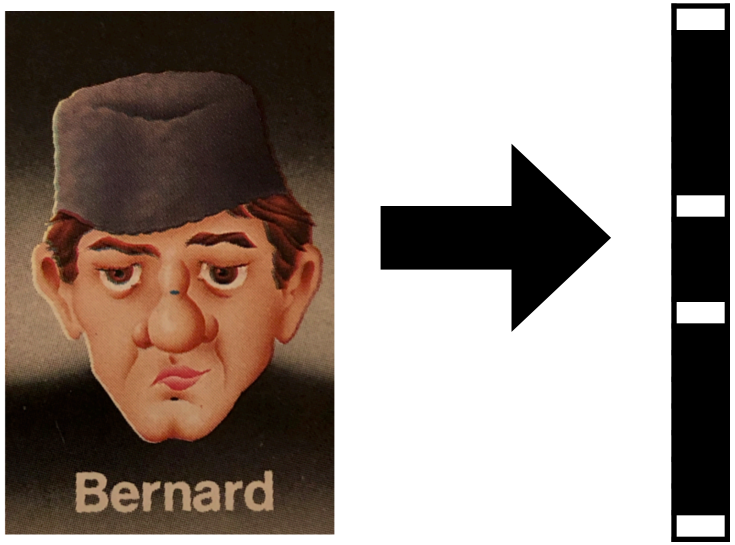
0	3	1	0
0	0	2	3
2	0	1	0
3	1	0	1
0	0	2	2

Minimal increase reduces overcounting by identifying collisions.

(Count returned by sketch) $\geq$ (True count of the query)

The hidden problem with sketches…





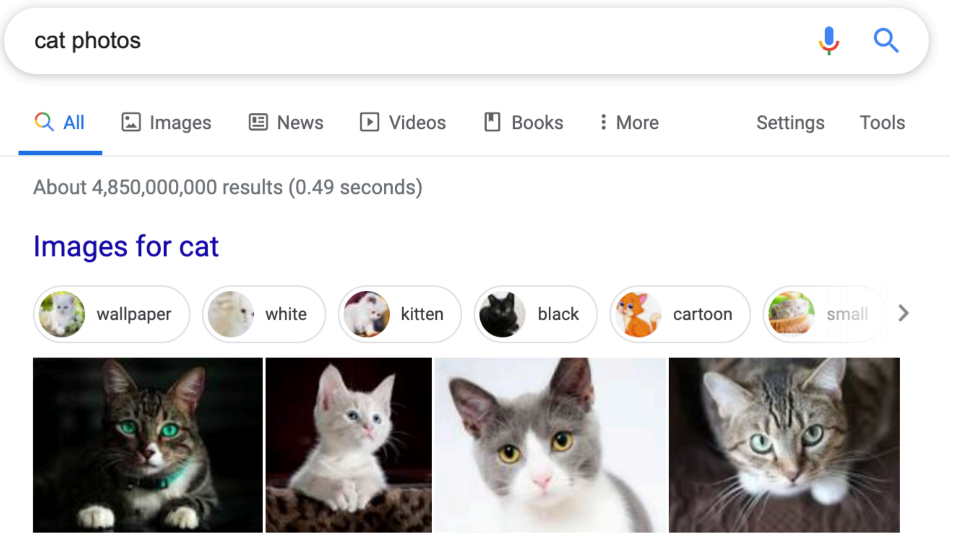
Cardinality

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

2
4
9
3
7
9
7
8
5
6

Cardinality

Sometimes its not possible or realistic to count all objects!



Estimate: 60 billion — 130 trillion

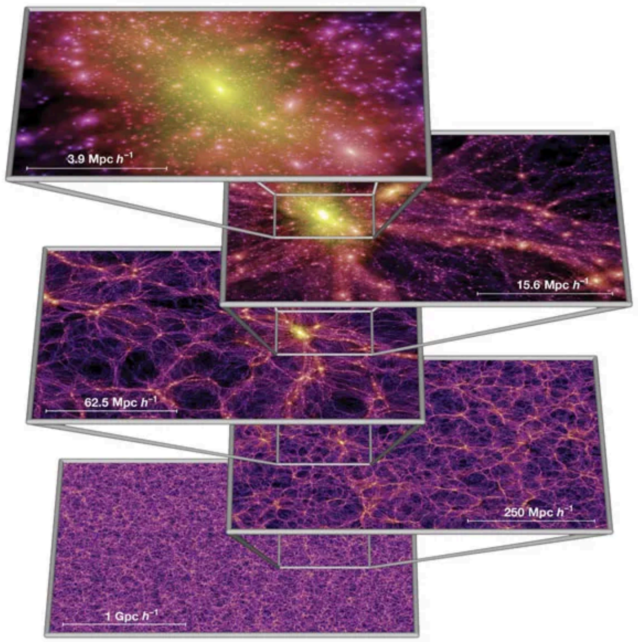


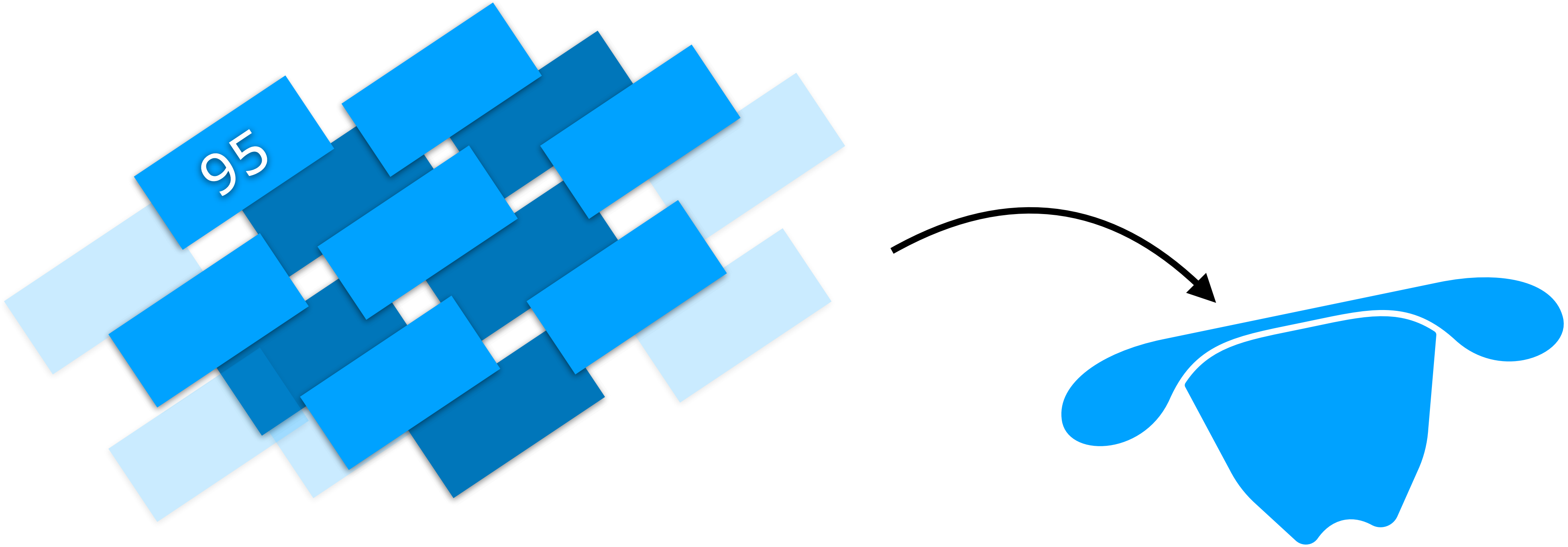
Image: https://doi.org/10.1038/nature03597

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.

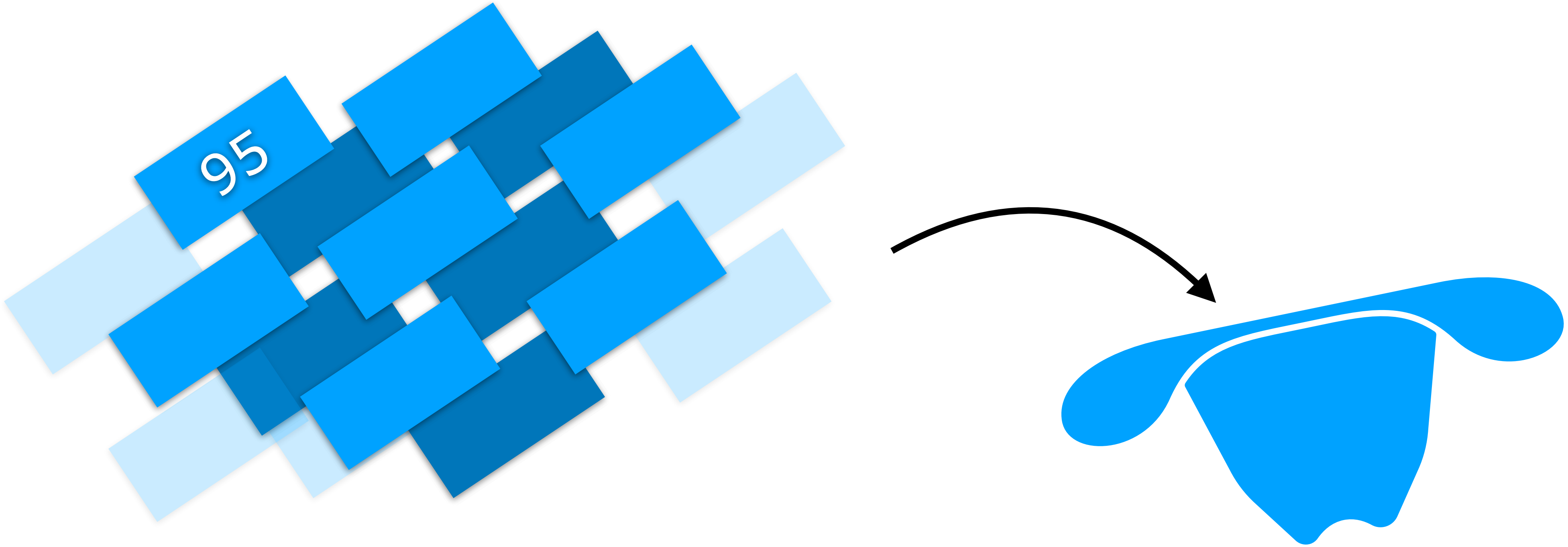
If I told you the value of the card was 95, what have we learned?



Cardinality Estimation

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?



Cardinality Estimation

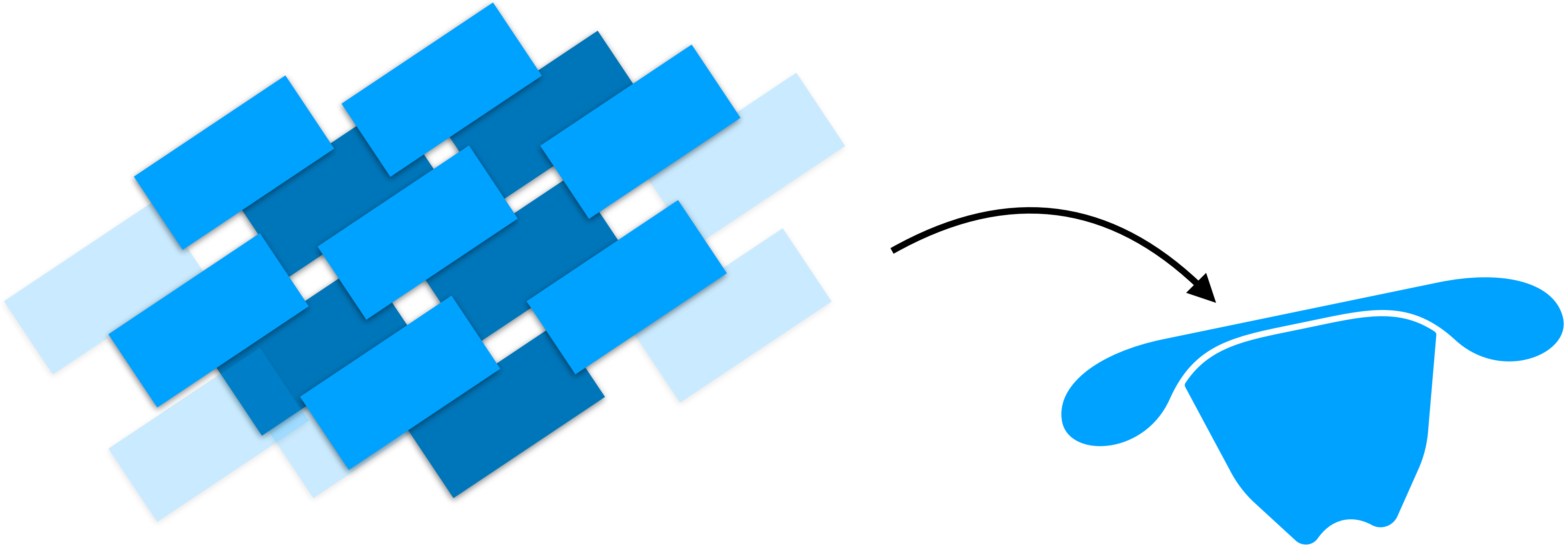
Imagine we have multiple sets (multiple minimums).

Cardinality Estimation

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

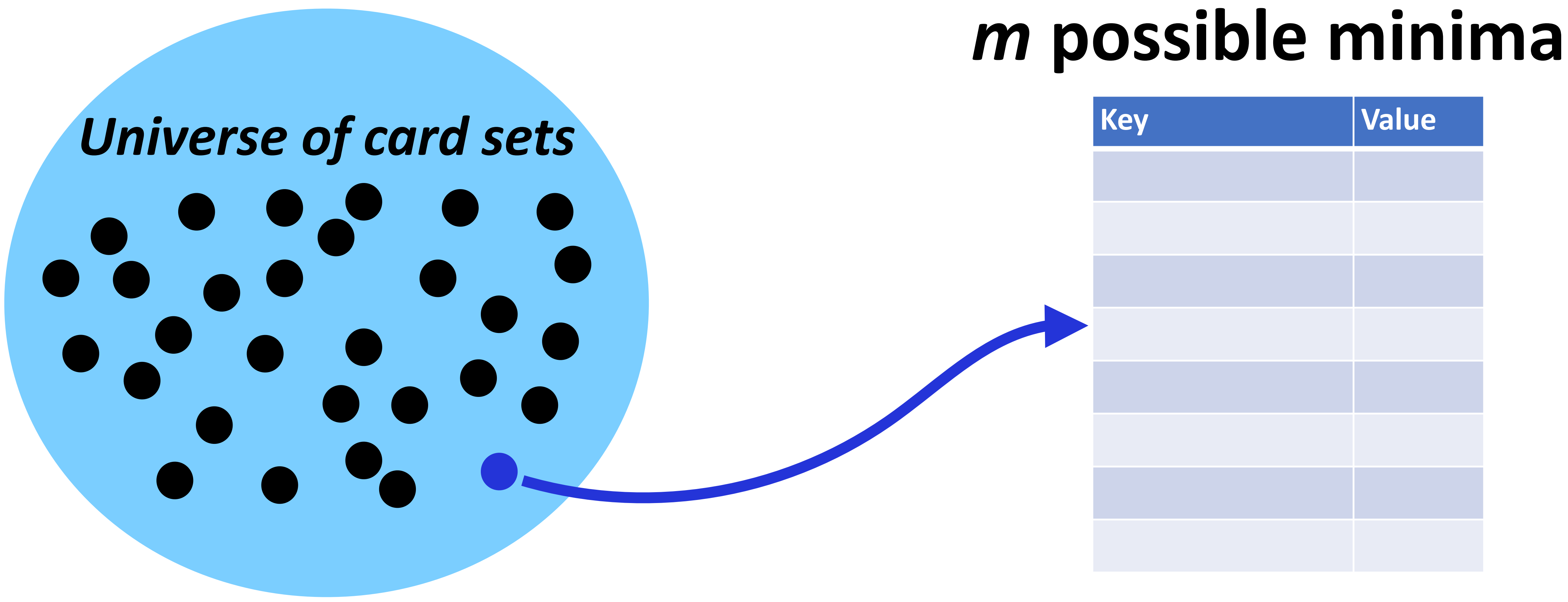
Cardinality Estimation

Why do we care about “the hat problem”?



Cardinality Estimation

Why do we care about "the hat problem"?

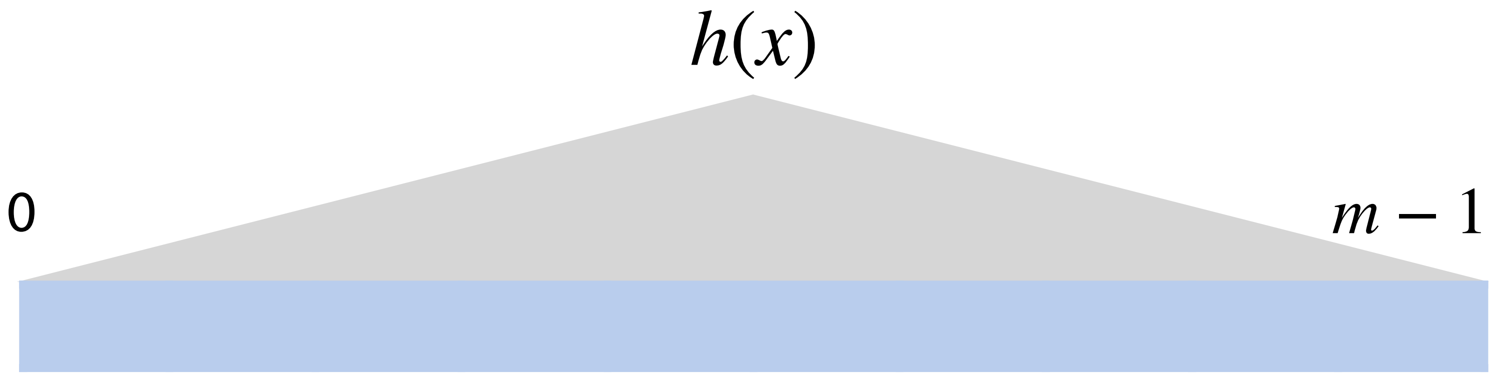


Cardinality Estimation

Now imagine we have a SUHA hash h over a range m.

Here a hash insert is equivalent to adding a card to our hat!

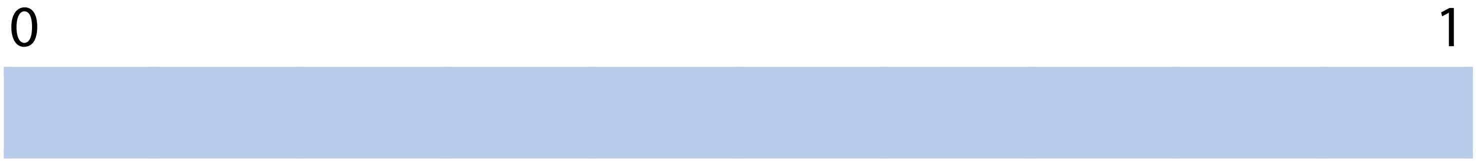
Now storing only the minimum hash value is a **sketch!**



Cardinality Sketch

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

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



Cardinality Sketch

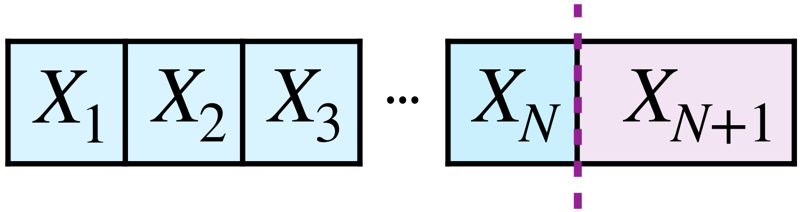
Claim: $\mathbf{E}[M] = \dfrac{1}{N+1}$

$$N \approx \frac{1}{M} - 1$$

Attempt 1	0.962	0.328	0.771	0.952	0.923
Attempt 2	0.253	0.839	0.327	0.655	0.491
Attempt 3	0.134	0.580	0.364	0.743	0.931

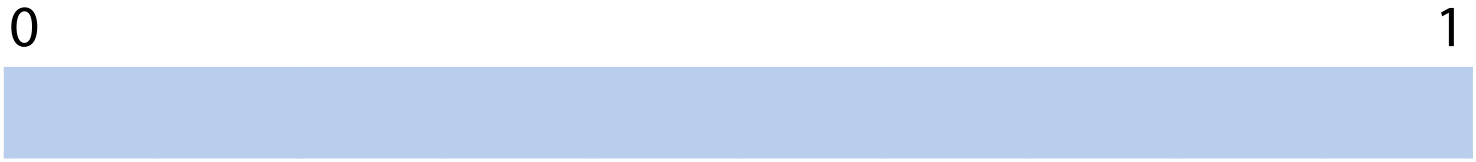
Cardinality Sketch

Consider an $N + 1$ draw:



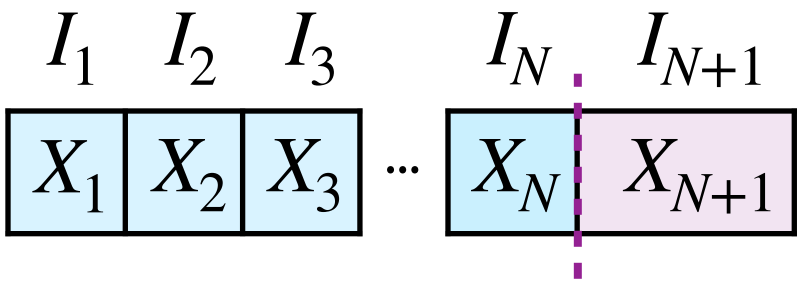
$$M = \min_{1 \le i \le N} X_i$$

Claim: $\mathbf{E}[M] = Pr\left(X_{N+1} < \min_{1 \le i \le N} X_i\right)$



Cardinality Sketch

Consider an $N+1$ draw:

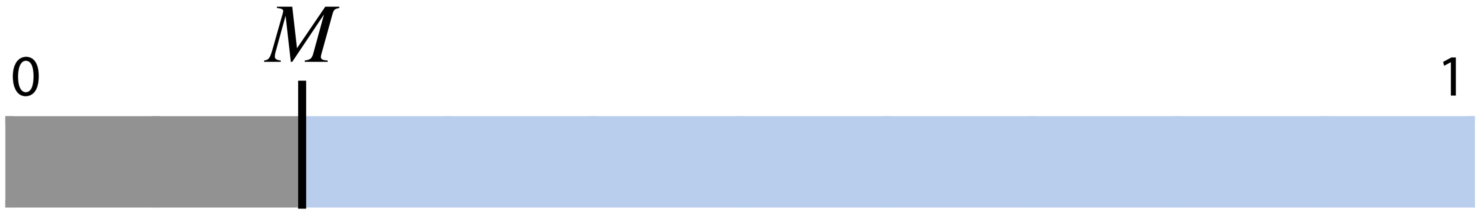


$$M = \min_{1 \le i \le N} X_i$$

Claim: $\mathbf{E}[M] = Pr\left(X_{N+1} < \min_{1 \le i \le N} X_i\right)$

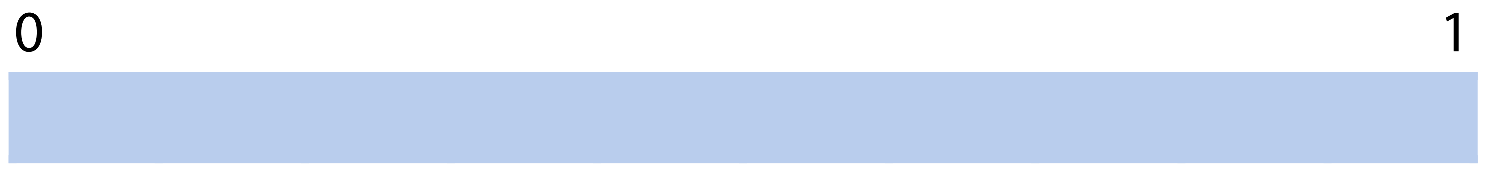
$$I_i = \begin{cases} 1 & \text{if } X_i < \min_{j \neq i} X_j \\ 0 & \text{otherwise} \end{cases}$$

$$\Pr\left(X_{N+1} < \min_{1 \le i \le N} X_i\right) = \mathbf{E}[I_{N+1}] = \frac{1}{N+1} = \mathbf{E}[M]$$



Cardinality Sketch

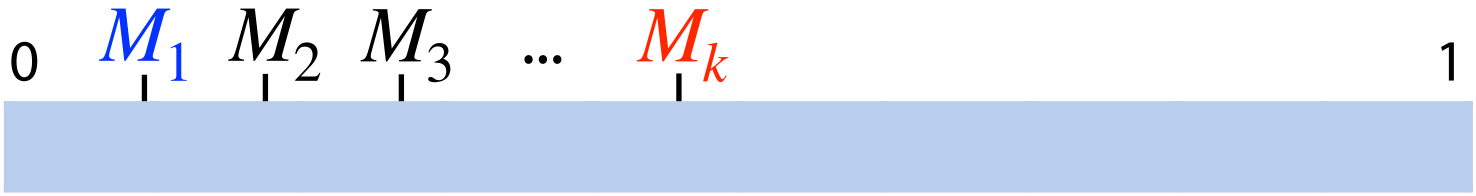
The minimum hash is a valid sketch of a dataset but can we do better?



Cardinality Sketch

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

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



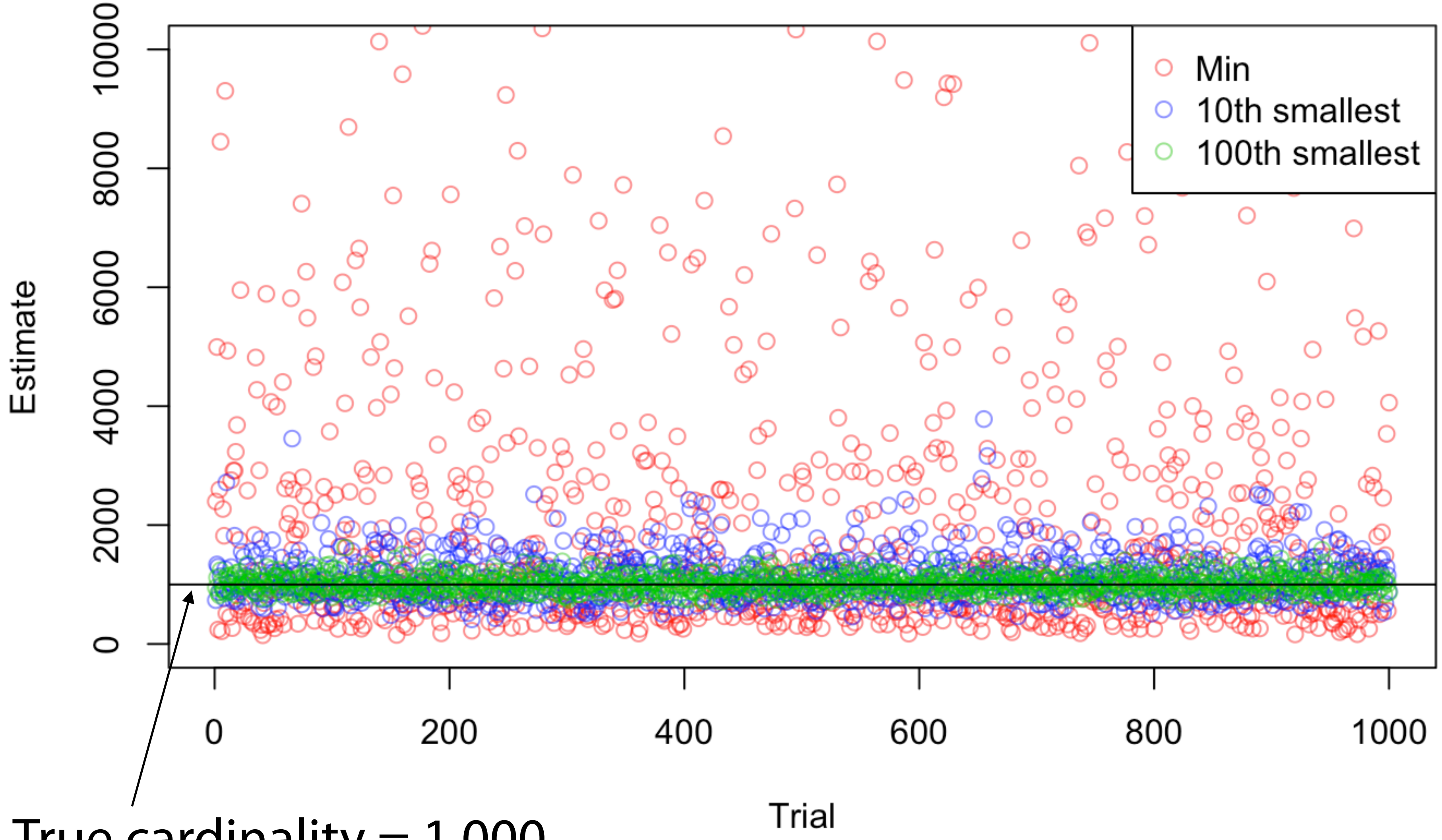
Cardinality Sketch

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

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

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

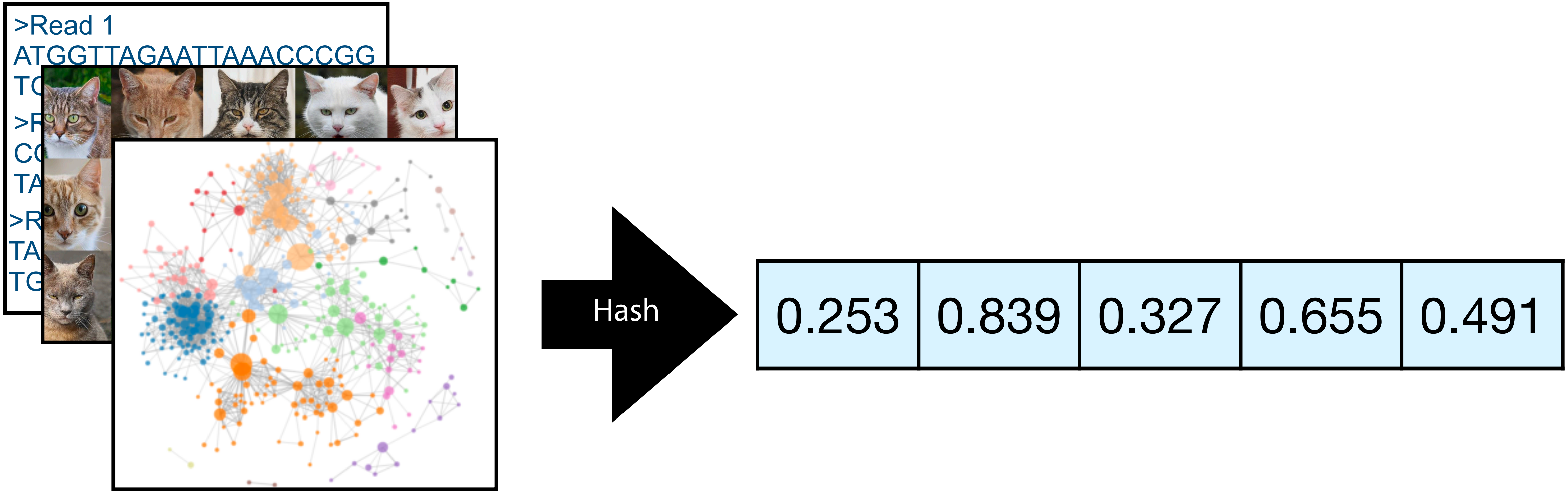
Cardinality



True cardinality = 1,000

Cardinality

Given any dataset and a SUHA hash function, we can estimate the number of unique items by tracking the minimum hash values.



Applied Cardinalities

Cardinalities

$$|A|$$

$$|B|$$

$$|A \cup B|$$

$$|A \cap B|$$

Set similarities

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

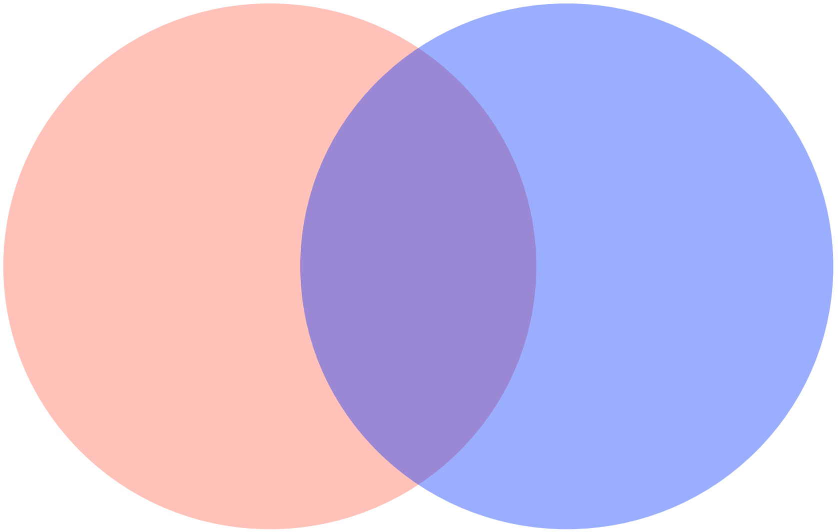
$$J = \frac{|A \cap B|}{|A \cup B|}$$

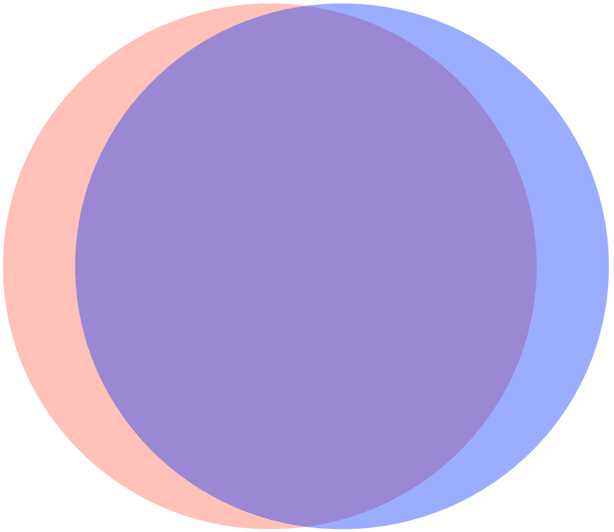
Real-world Meaning

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

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

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

Set Similarity

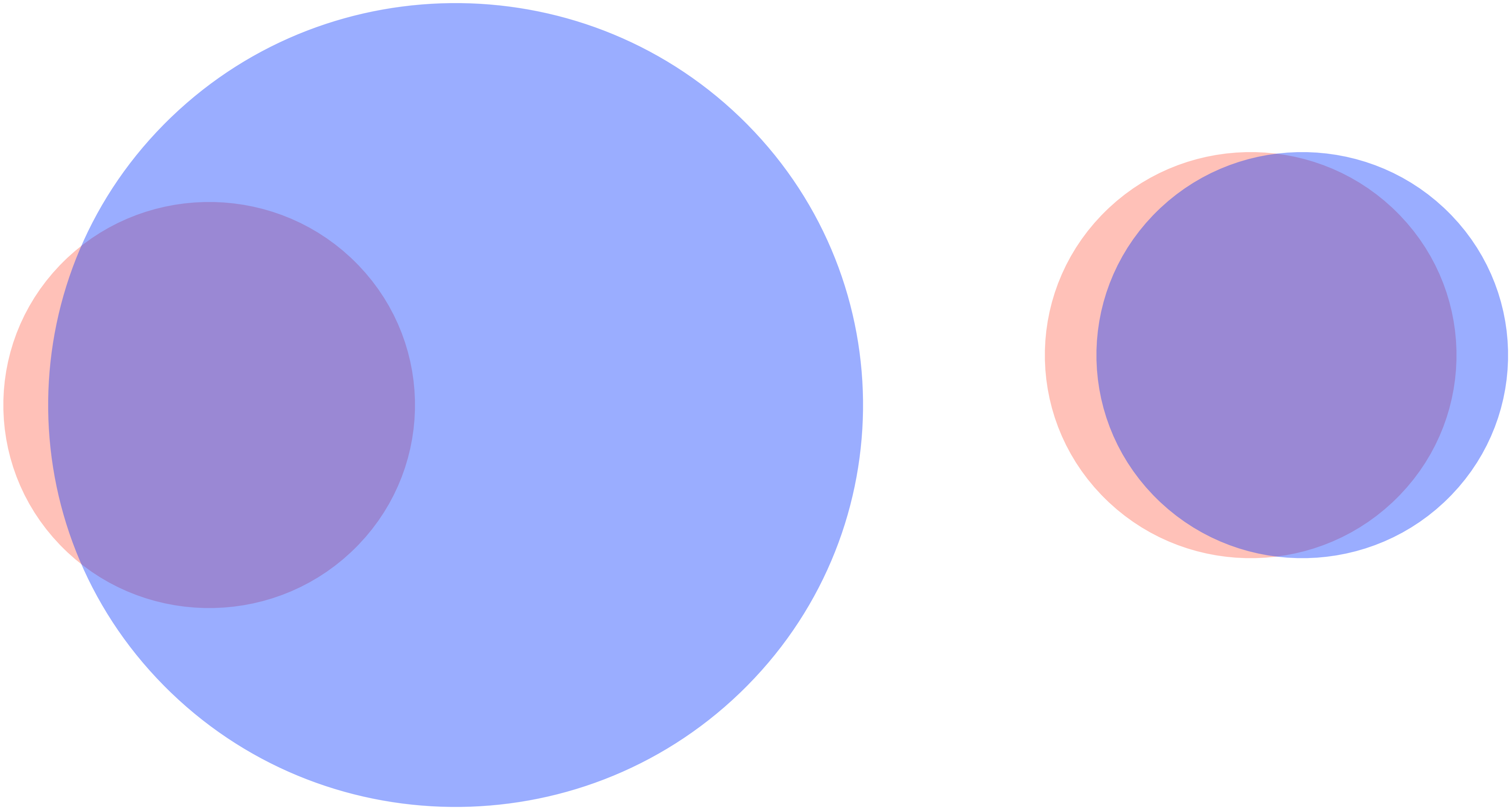
How can we describe how ***similar*** two sets are?





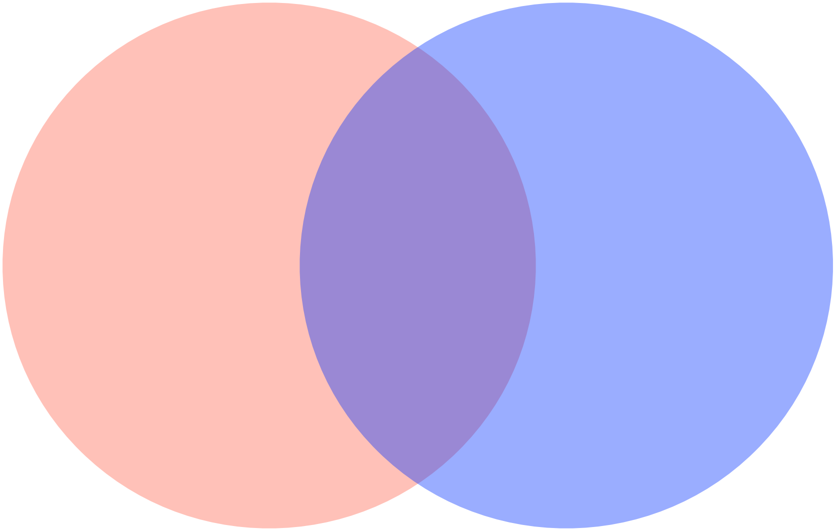
Set Similarity

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



Set Similarity

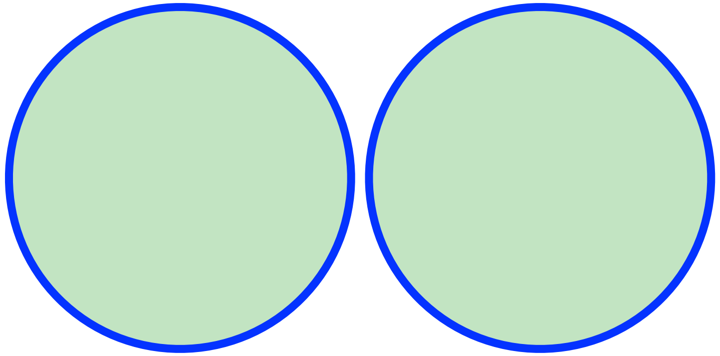
To measure **similarity** of A & B, we need both a measure of how similar the sets are but also the total size of both sets.



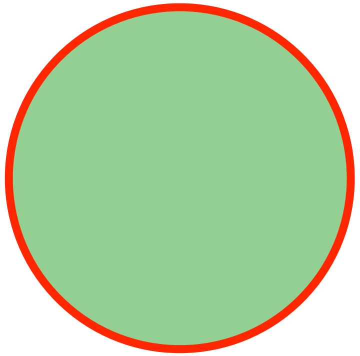
$$J = \frac{|A \cap B|}{|A \cup B|}$$

J is the ***Jaccard coefficient***

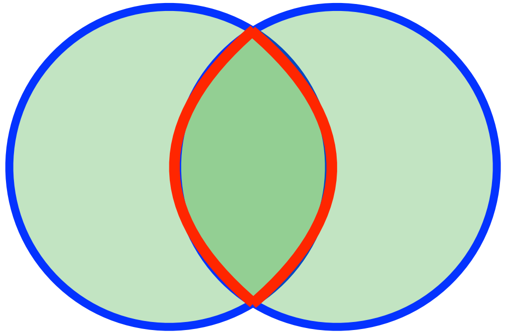
Set Similarity



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



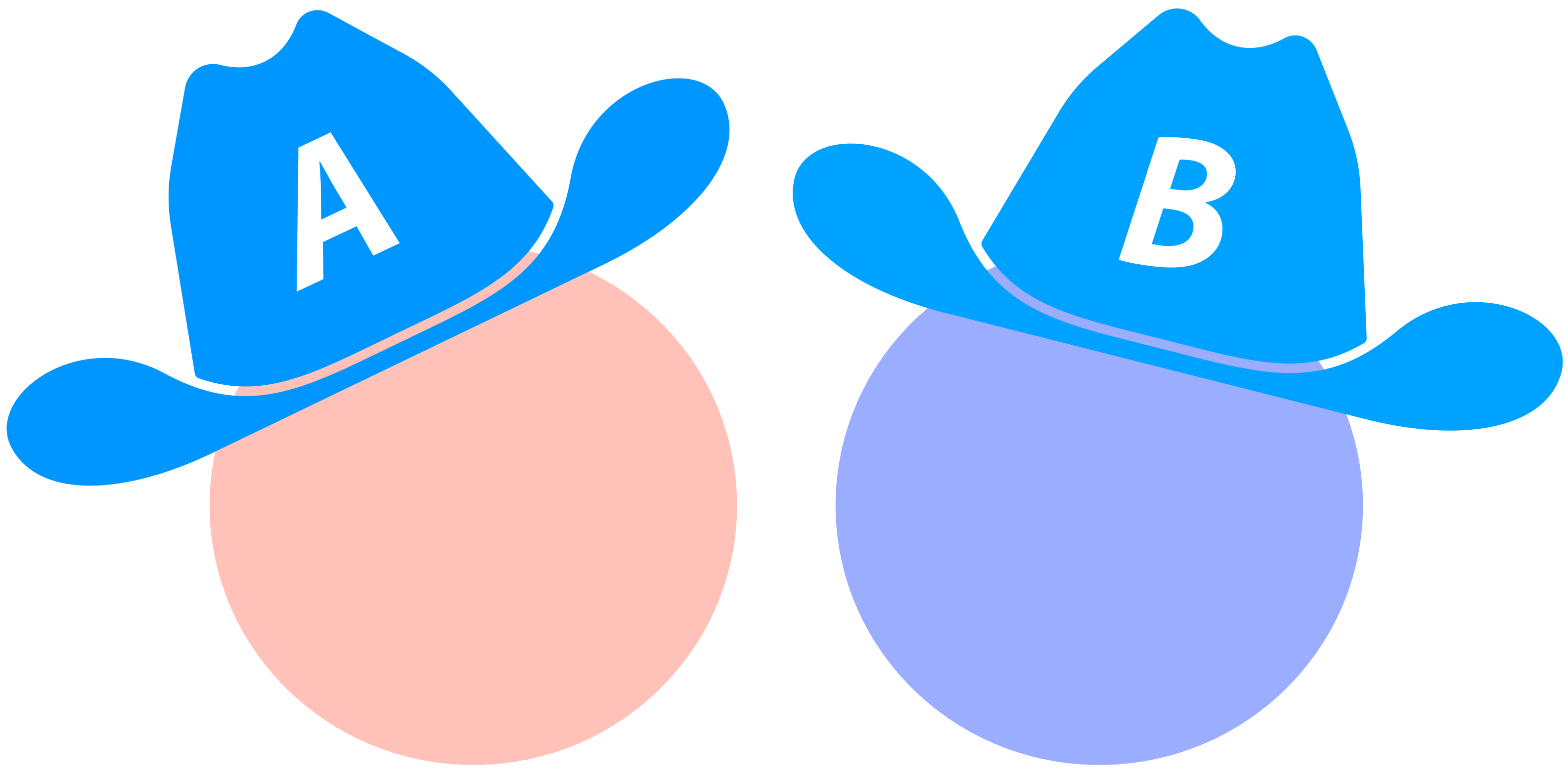
$$\frac{|A \cap B|}{|A \cup B|} = 1$$



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

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

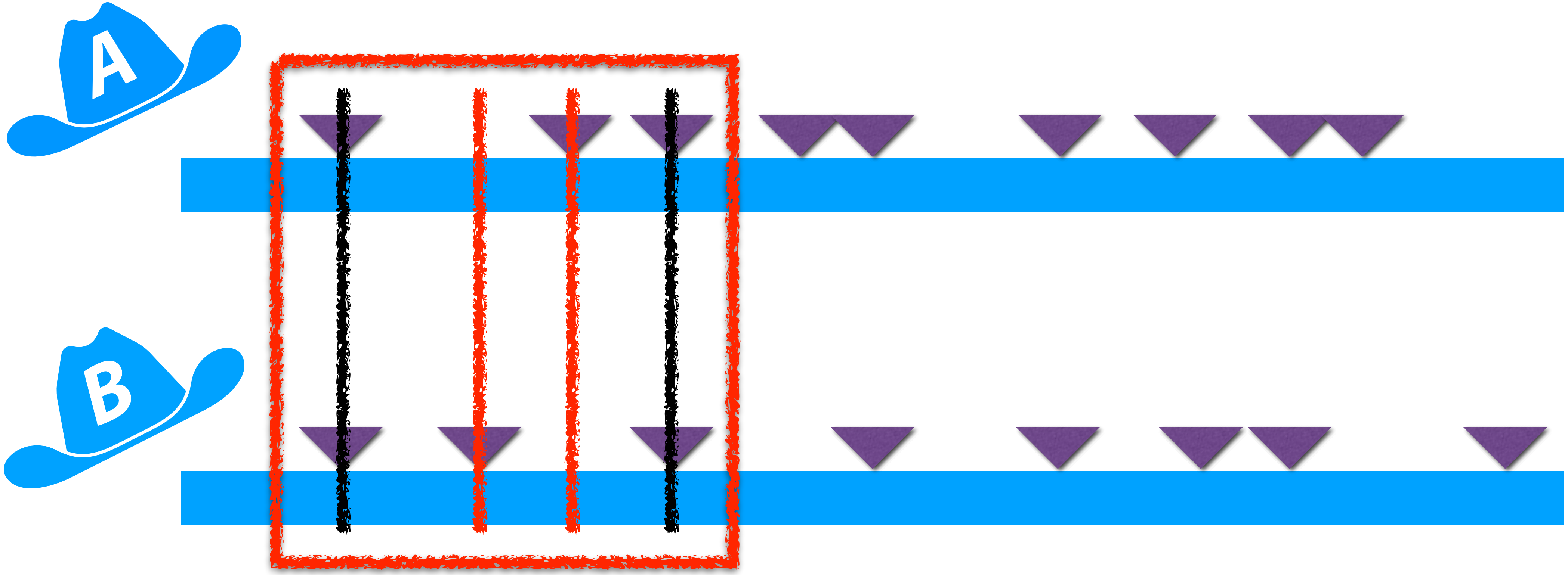


Image inspired by: Ondov B, Starrett G, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. **Mash Screen: high-throughput sequence containment estimation for genome discovery**. *Genome Biol* 20, 232 (2019)

Similarity Sketches

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

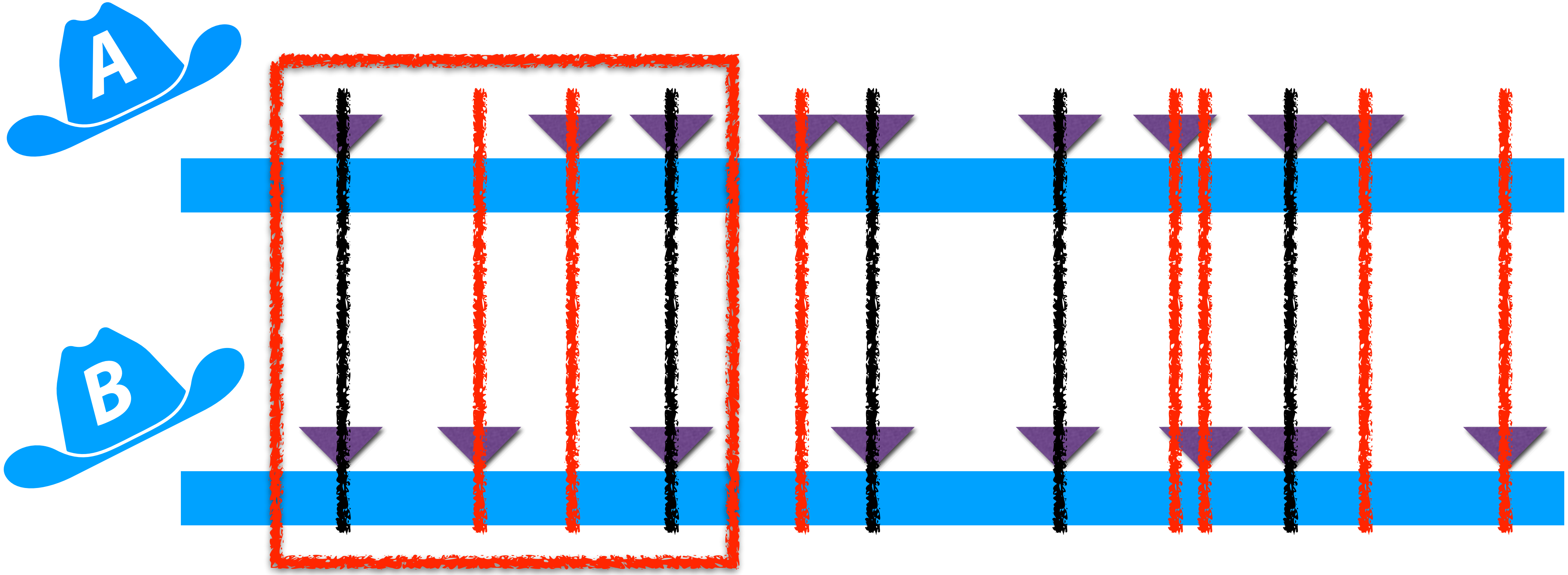


Image inspired by: Ondov B, Starrett G, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. **Mash Screen: high-throughput sequence containment estimation for genome discovery**. *Genome Biol* 20, 232 (2019)

The MinHash Sketch

Let sets A and B be two arbitrary sets of at least 8 elements

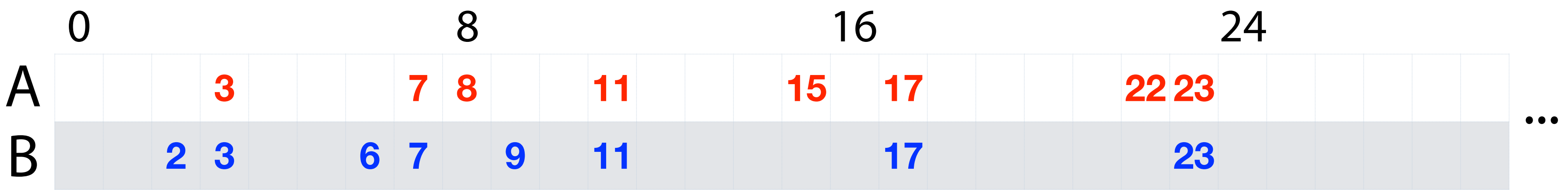
The eight minimum hash values for sets A and B is a **MinHash Sketch**

Sketch A

3	15
7	17
8	22
11	23

Sketch B

2	9
3	11
6	17
7	23



The MinHash Sketch

To get similarity, we want to estimate $|A \cup B|$ and $|A \cap B|$…

Sketch A

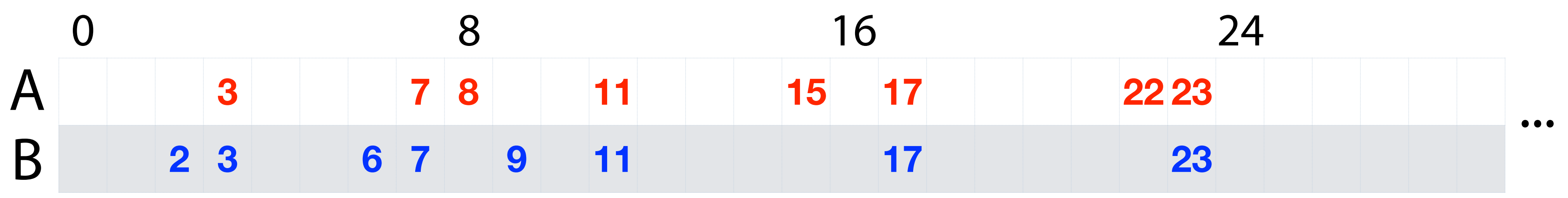
3	15
7	17
8	22
11	23

$\cup$

Sketch B

2	9
3	11
6	17
7	23

$=$



The MinHash Sketch

To get similarity, we want to estimate $|A \cup B|$ and $|A \cap B|$…

Sketch A

3	15
7	17
8	22
11	23

$\cup$

Sketch B

2	9
3	11
6	17
7	23

$=$

Sketch of $|A \cup B|$

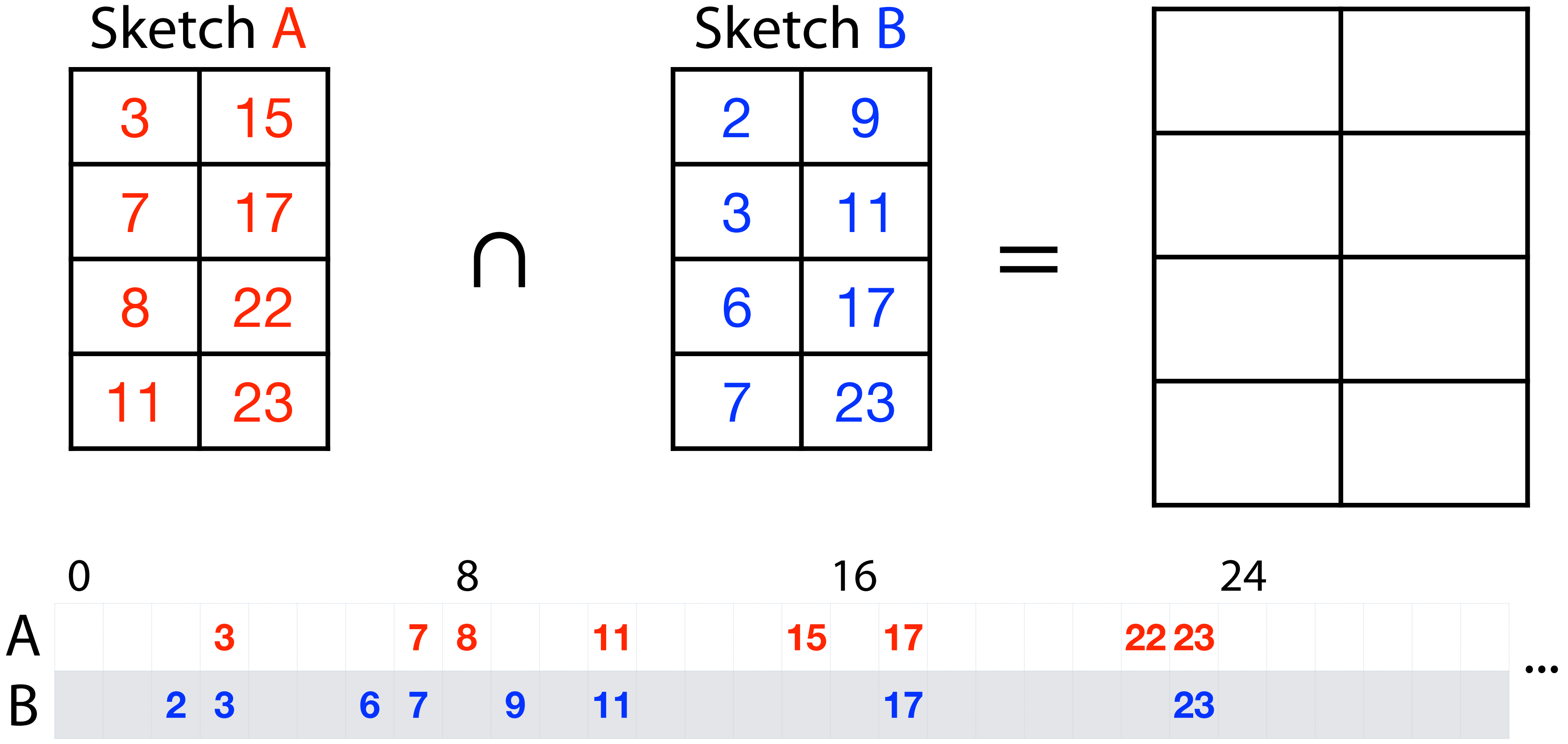
2	8
3	9
6	11
7	15

	0								8								16								24					
A				3				7	8			11				15		17					22	23						
B			2	3			6	7		9		11						17						23						

…

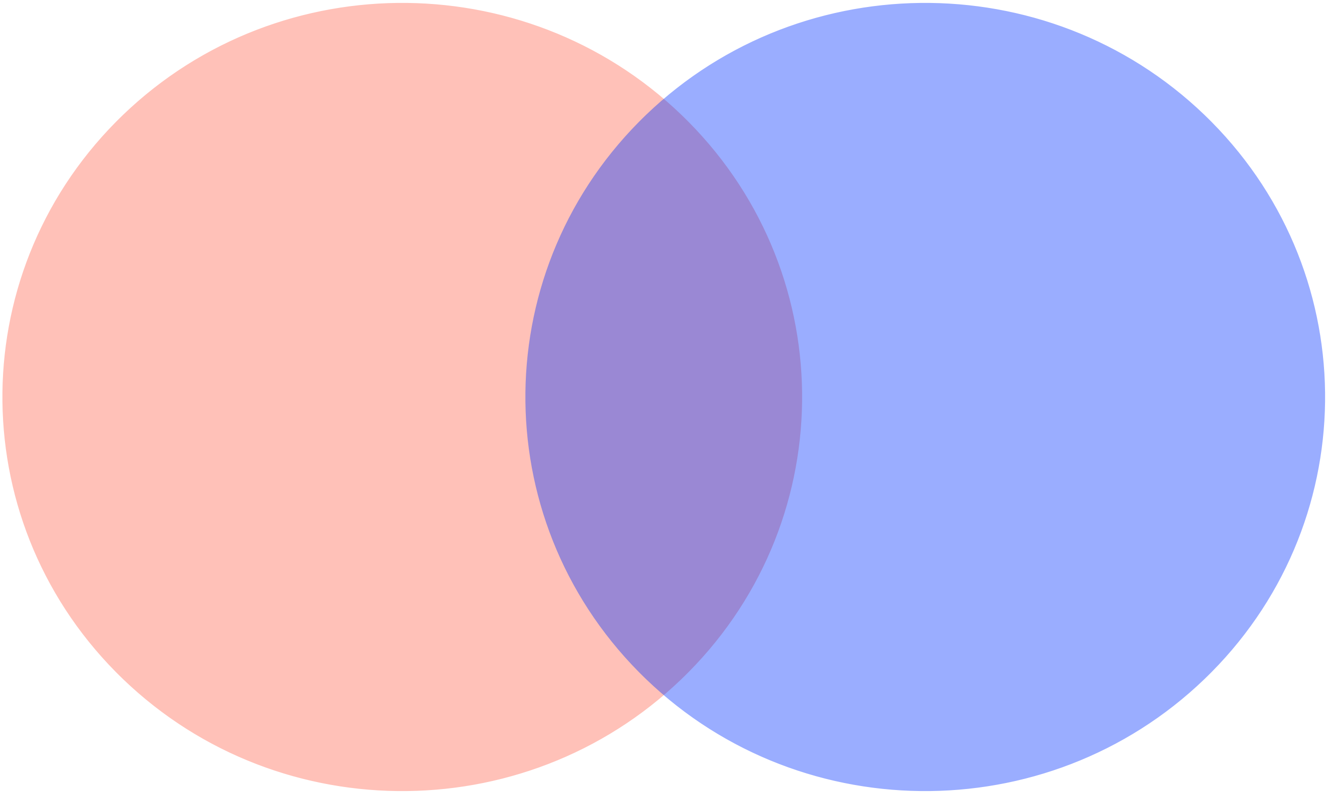
The MinHash Sketch

To get similarity, we want to estimate $|A \cup B|$ and $|A \cap B|$…



Inclusion-Exclusion Principle

$$|A \cap B| =$$



The MinHash Sketch

Using **inclusion-exclusion** principle and KMV, we can estimate similarity!

Sketch A

3	15
7	17
8	22
11	23

Sketch B

2	9
3	11
6	17
7	23

Sketch of $|A \cup B|$

2	8
3	9
6	11
7	15

kth minimum value (KMV) with $k = 8$, assuming hash range is integers in [0, 100):

$$= \frac{800/23 - 1 + 800/23 - 1 - 800/15 - 1}{800/15 - 1}$$

$$= \frac{34.782 + 34.782 - 53.333 - 1}{53.333 - 1}$$

$$\approx 0.29$$

$$\frac{|A| + |B| - |A \cup B|}{|A \cup B|}$$

The MinHash Sketch

Claim: Cardinality of the intersection can also be estimated directly!

Sketch A

3	15
7	17
8	22
11	23

Sketch B

2	9
3	11
6	17
7	23

Sketch of $|A \cup B|$

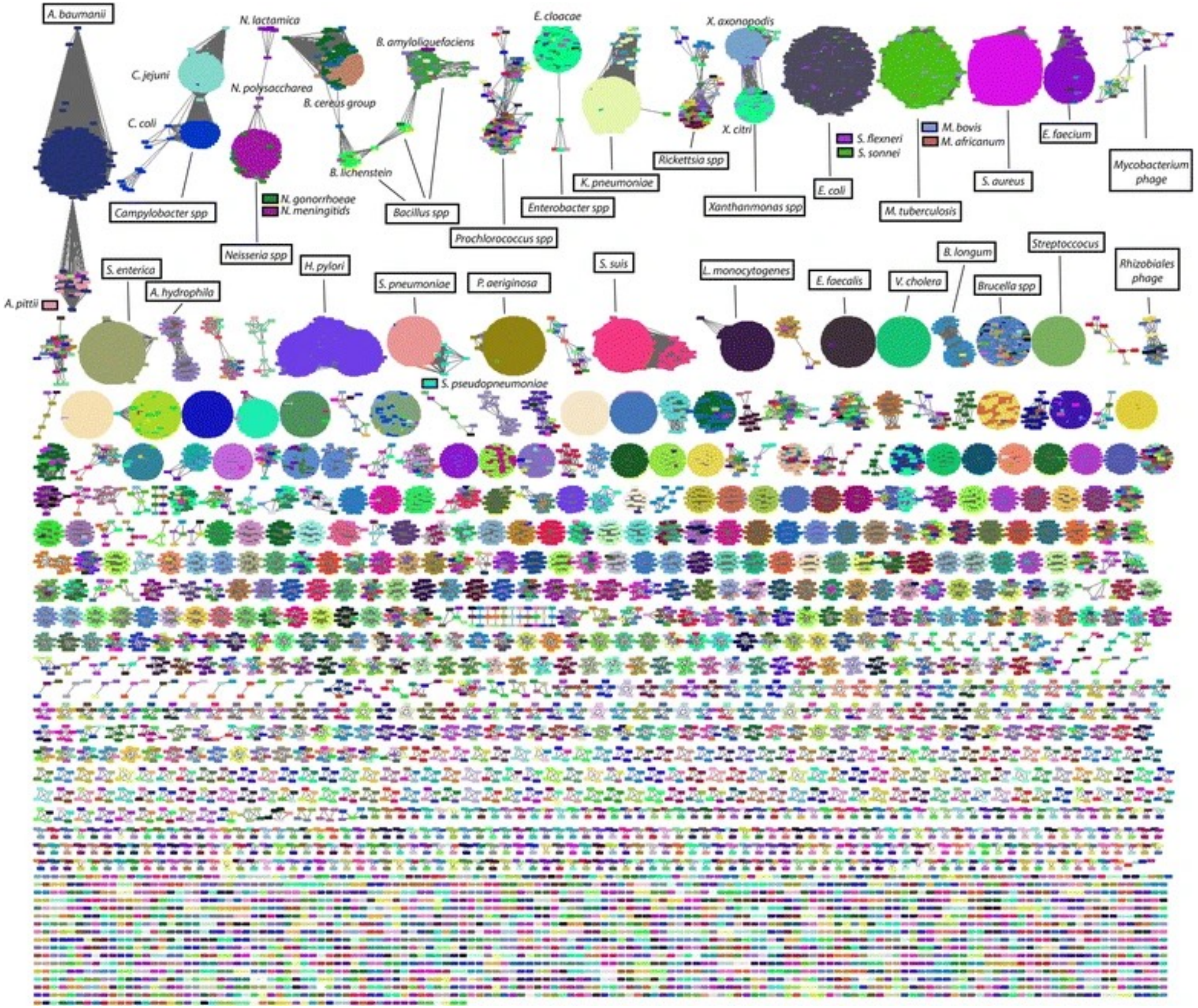
2	8
3	9
6	11
7	15

1) Sequence decomposed into **kmers**

S_1: CATGGACCGACCAG	GCAGTACCGATCGT : S_2
CAT GAC GAC	GTA CGA CGT
ATG ACC ACC	AGT CCG TCG
TGG CCG CCA	CAG ACC ATC
GGA CGA CAG	GCA TAC GAT

Assembling large genomes with single-molecule sequencing and locality-sensitive hashing
Berlin et al (2015) *Nature Biotechnology*

MinHash in practice



Mash: fast genome and metagenome distance estimation using MinHash
Ondov et al (2016) *Genome Biology*

Reviewing probabilistic data sketches

What sketch would I use for the following:

Does a *specific* object exist in my data?

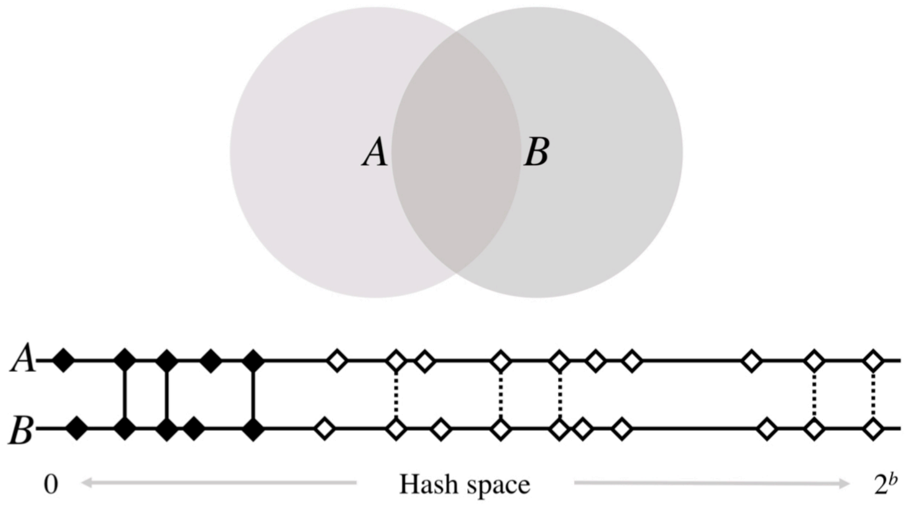
How often is a *specific* object repeated in my data?

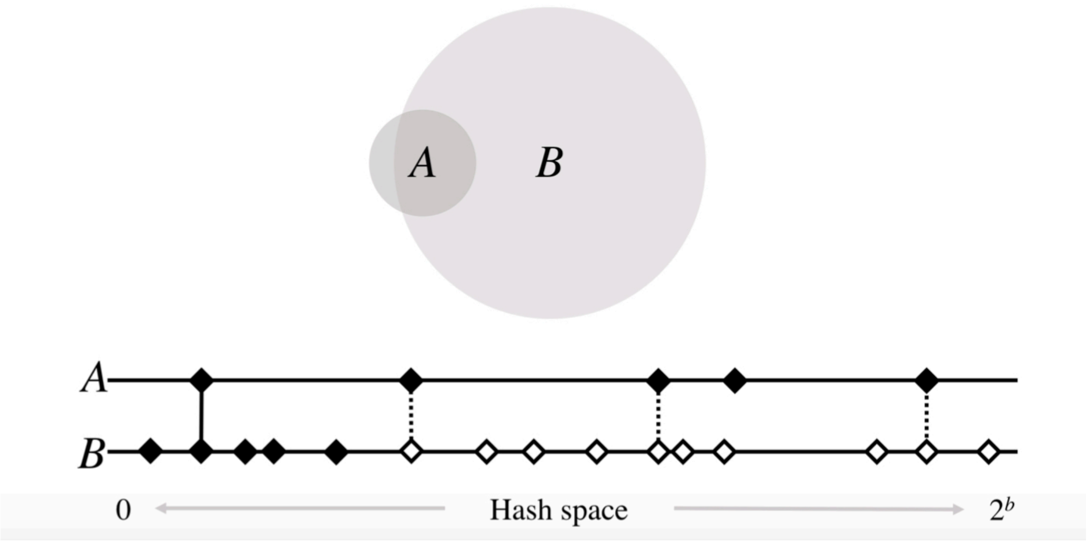
How many unique objects do I have in my set?

How similar are two datasets?

Bonus Slides (Taking it one step further…)

Bottom-k minhash has low accuracy if the cardinality of sets are skewed

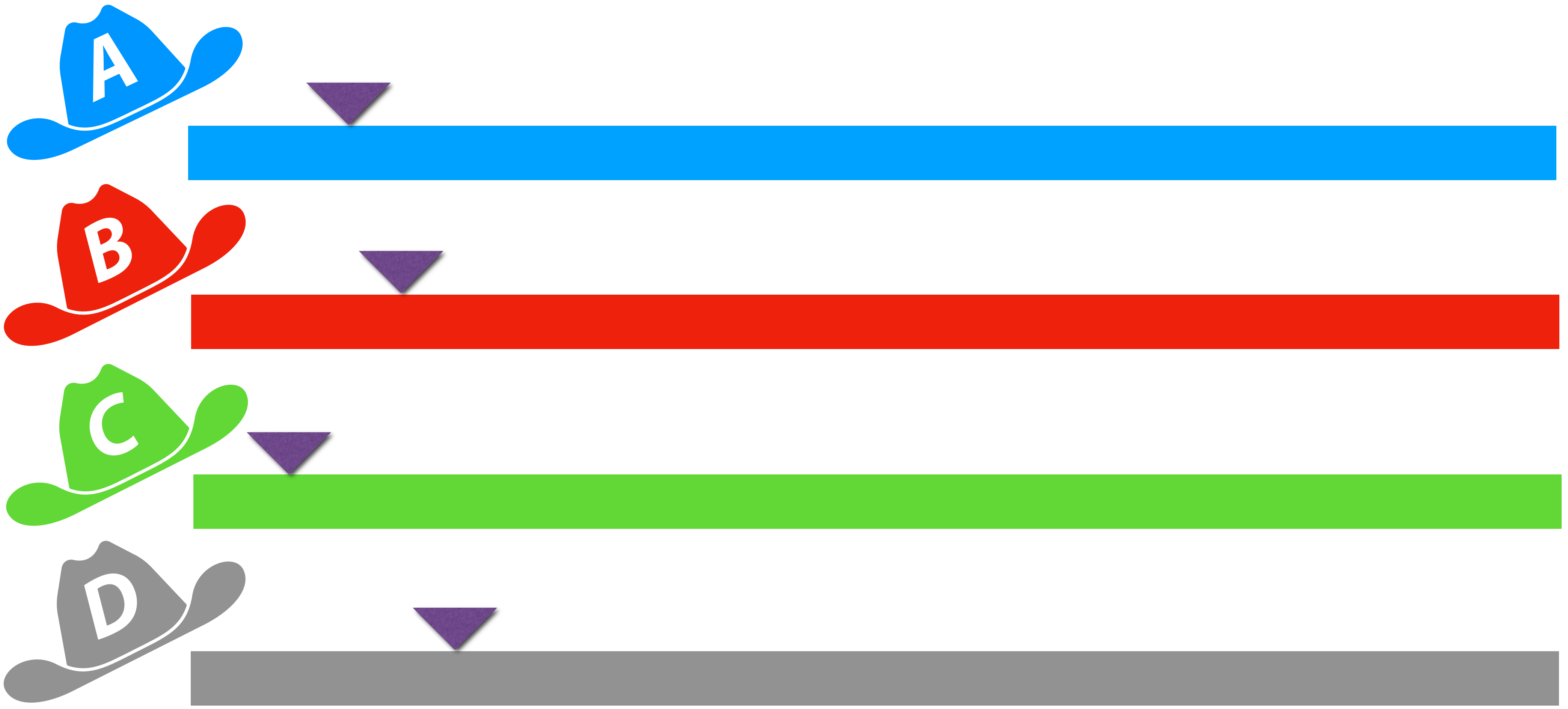




Ondov, Brian D., Gabriel J. Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B. Buck, and Adam M. Phillippy. **Mash Screen: High-throughput sequence containment estimation for genome discovery**. *Genome biology* 20.1 (2019): 1-13.

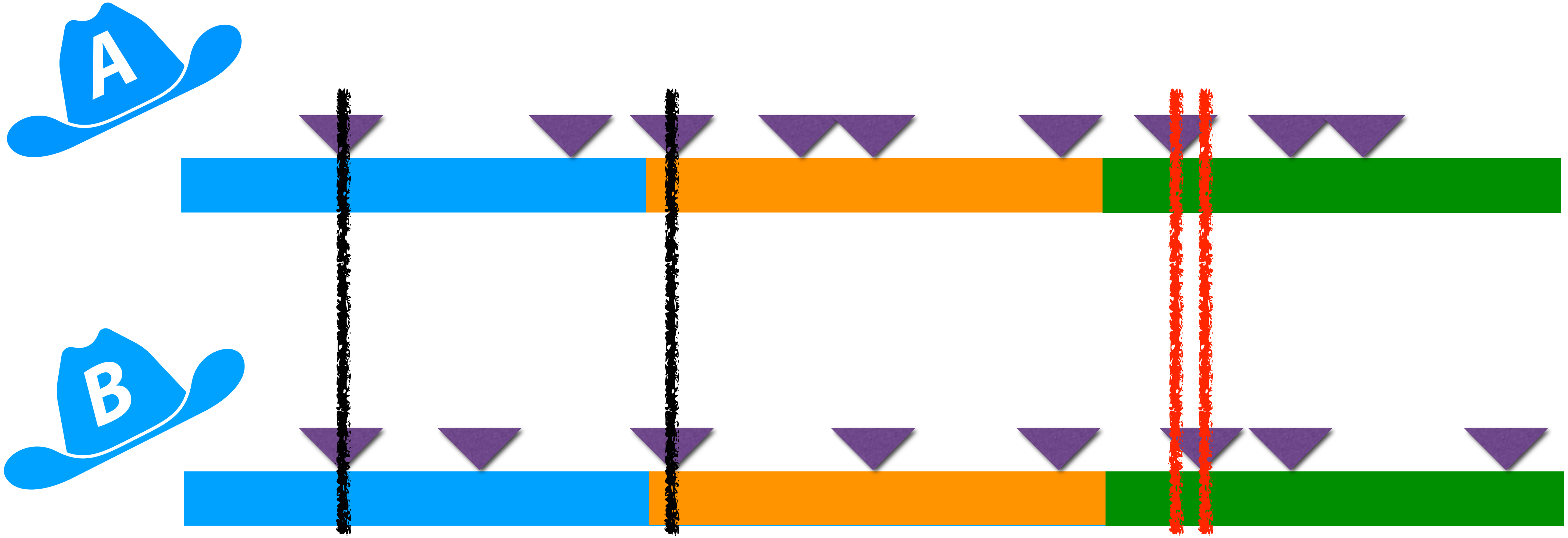
K-Hash Minhash

What if instead we used k different hashes and took the min each time?

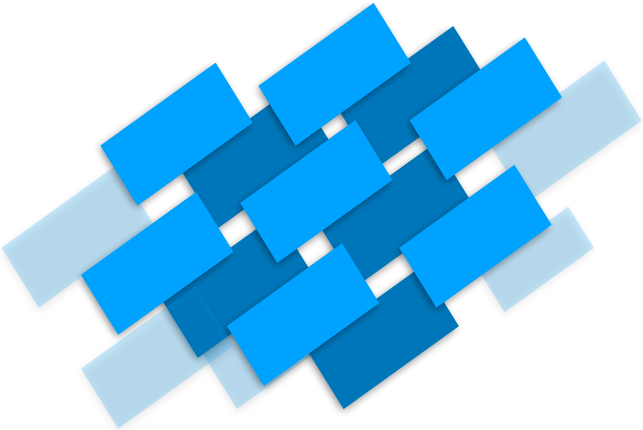


K-Partition Minhash

What if we instead took the minimum of k-partitions?



K-Partition Minhash



Hash

1010110101
0001111010
1101101011
1011010110
0101100000
0010001101

Partition

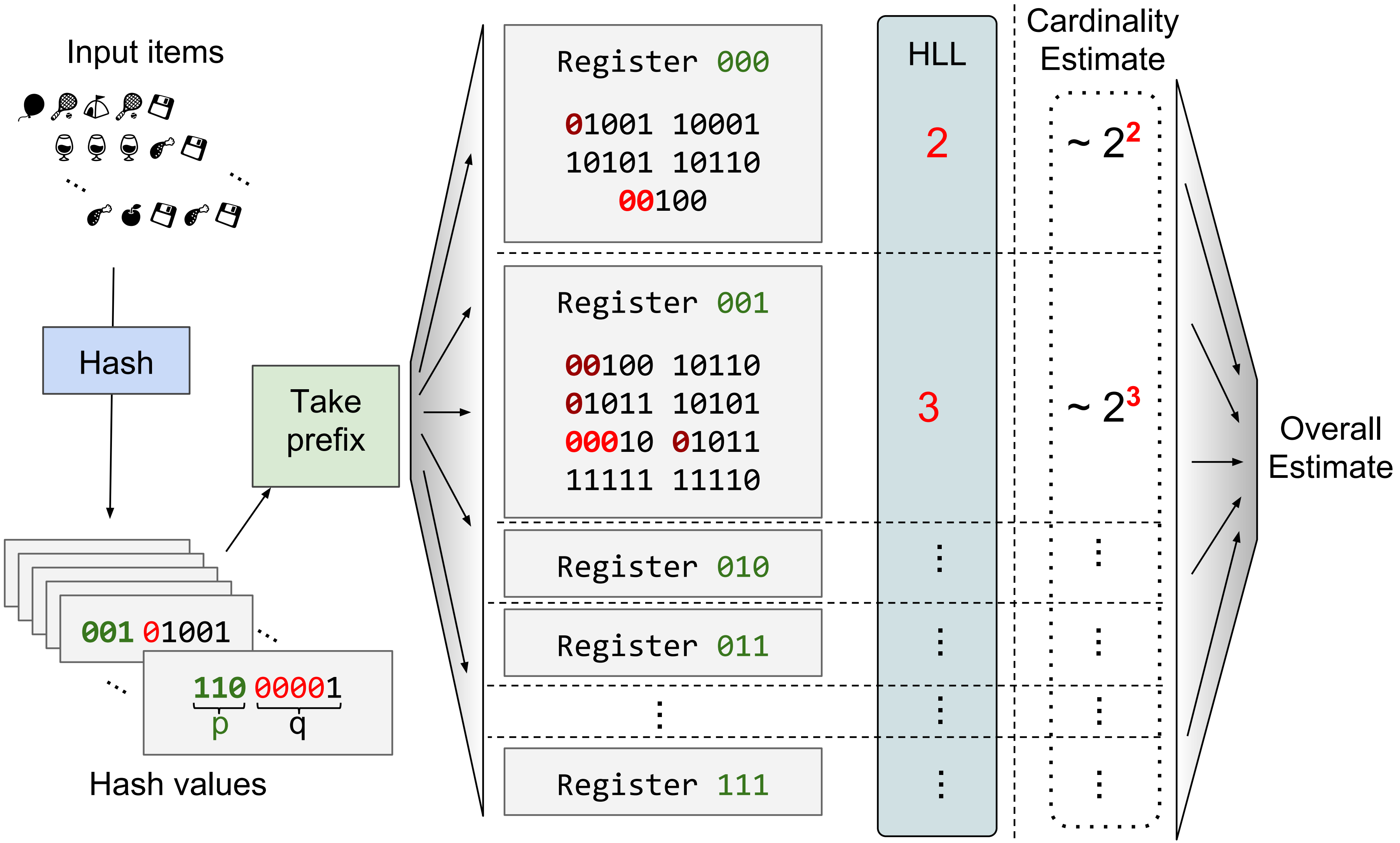
00
01111010
10001101

01
01100000

10
10110101
11010110

11
01101011

HyperLogLog



Baker, Daniel et al. "Dashing: fast and accurate genomic distances with HyperLogLog." *Genome biology* 20.1 (2019): 1-12.