Data Structures and Algorithms

Cardinality and Similarity Sketches

CS 225

April 29, 2022

Brad Solomon

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Department of Computer Science

Learning Objectives

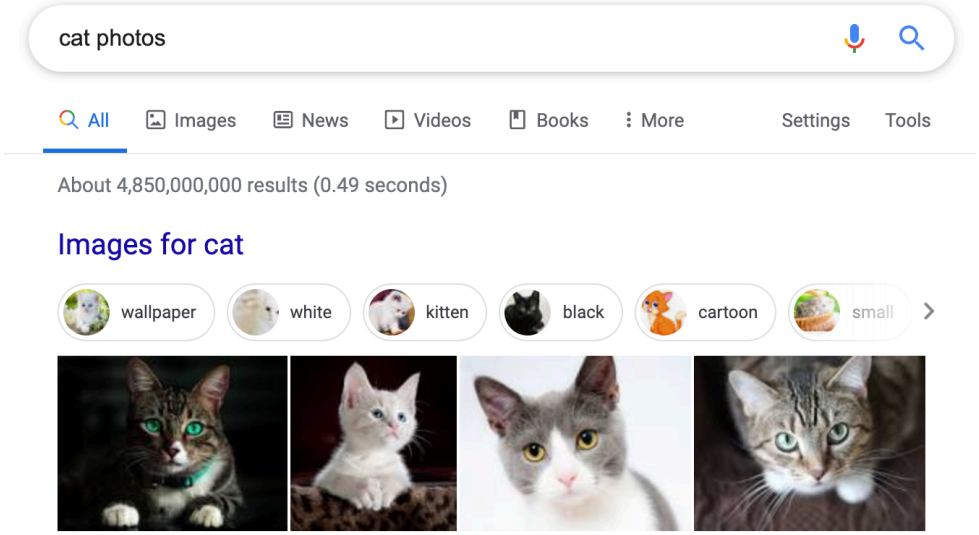
Introduce the concept of cardinality and cardinality estimation

See how hashing is an effective tool for approximation

Demonstrate the Minhash and HyperLogLog sketches

Cardinality

How many *distinct* (unique) values there are in a dataset



Google Index Estimate: >60 billion webpages

Google Universe Estimate (2013): >130 trillion webpages

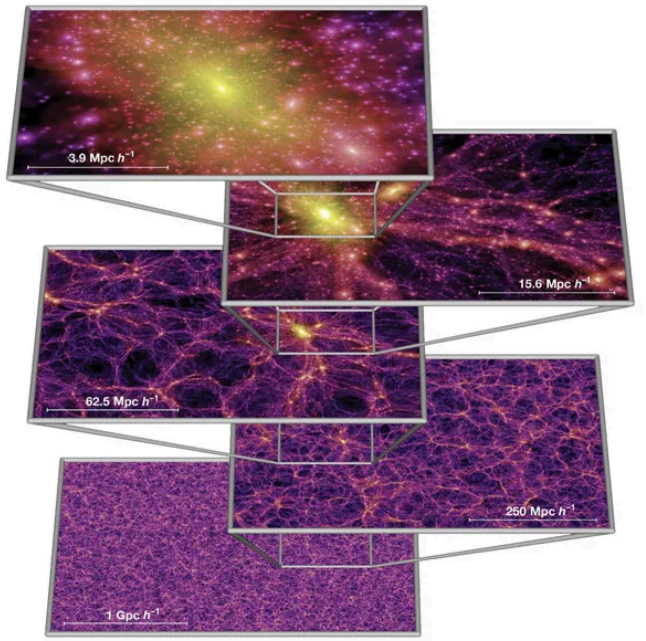
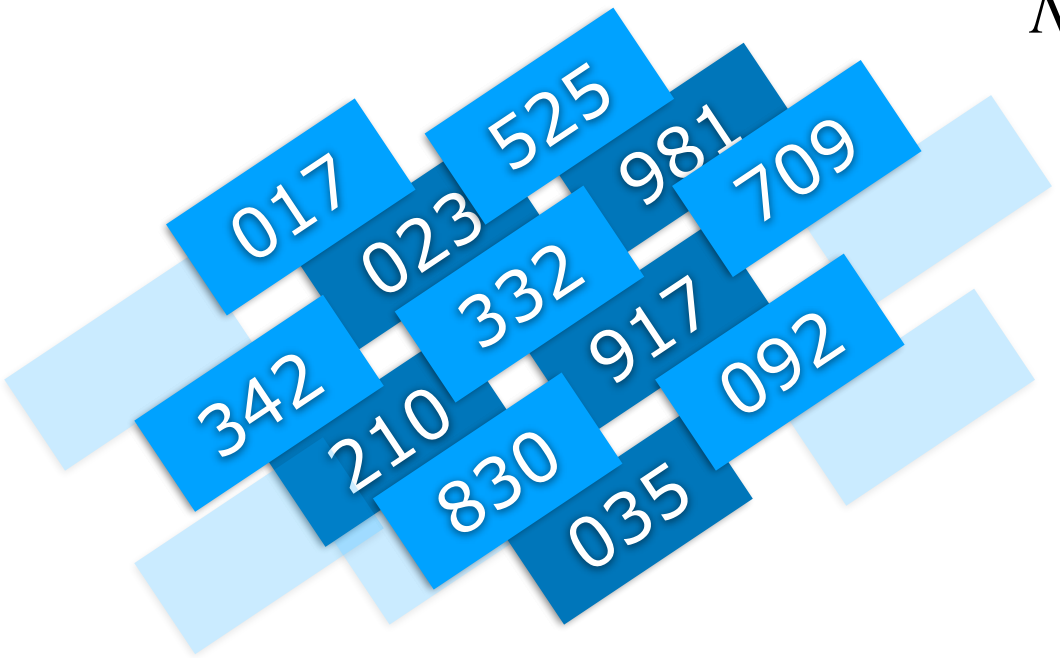


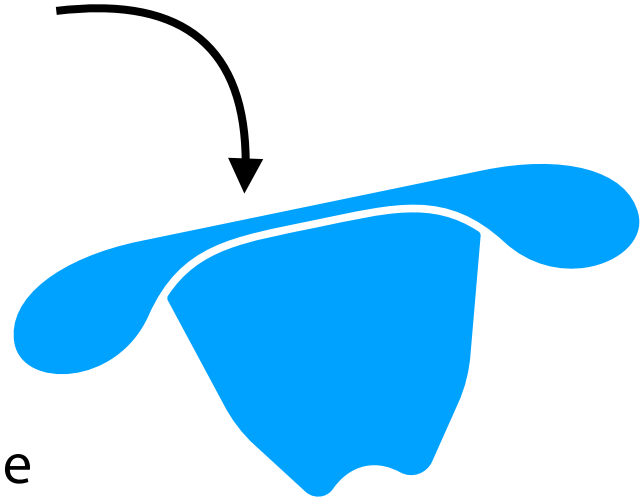
Image: https://doi.org/10.1038/nature03597

Cardinality

I take cards labeled 1--1,000 and choose a random subset of size N to hide in my hat



We want to estimate N



We can see **one representative** from the cards in the hat; which to pick?

Minimum, median, maximum? Something else?

Cardinality

What if **minimum** was 500? ...10? ... 4?

If minimum is 95, what's our estimate for N?

Cardinality

What if **minimum** was 500? ...10? ... 4?

If minimum is 95, what's our estimate for N?

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

Assuming our first 'partition' is about average:

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

$$N + 1 \approx 10.5$$

$$N \approx 9.5$$

Cardinality

Now imagine we have a SUHA hash (let h_{64} be a 64-bit hash)

$$\frac{h_{64}(x)}{2^{64}-1}$$

The randomness in the hash function turns any dataset-cardinality problem into the “hat problem”

Cardinality

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

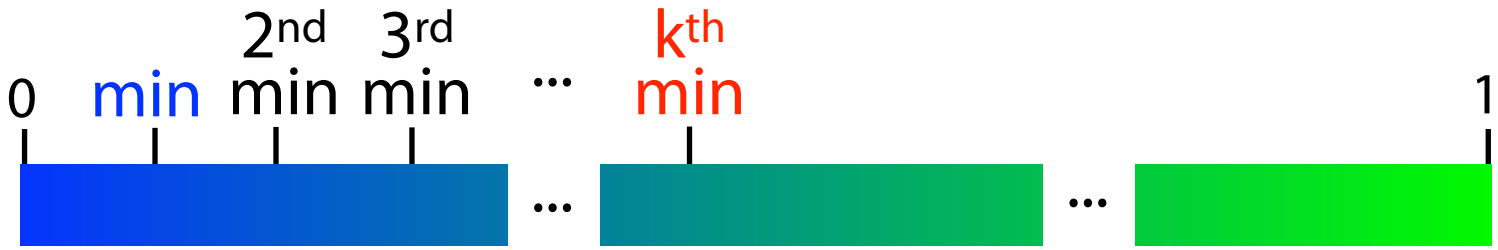
Claim: $\mathbf{E}[M] = \frac{1}{N+1}$

Cardinality

Attempt 1	0.455	0.220	0.951	0.236	0.979
Attempt 2	0.968	0.234	0.835	0.642	0.349
Attempt 3	0.774	0.484	0.309	0.526	0.143

Cardinality

Can the k^{th}-smallest hash value estimate the cardinality better than the minimum?



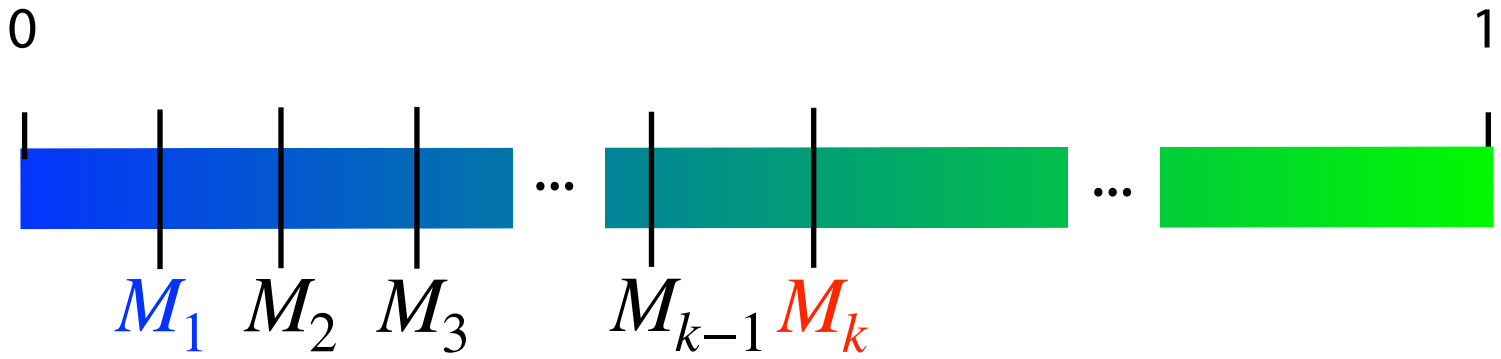
Cardinality

Can the k^{th}-smallest hash value estimate the cardinality better than the minimum?

$$\mathbf{E}[M_1] = \frac{1}{N+1} \qquad \mathbf{E}[M_k] = \frac{k}{N+1}$$

Cardinality

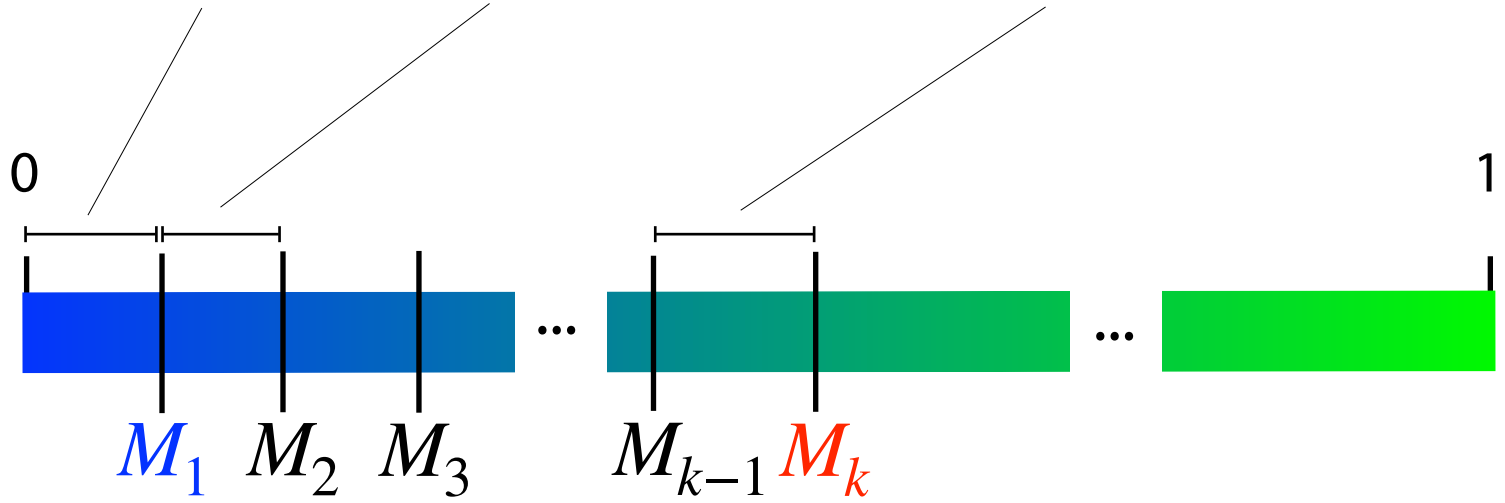
$$\frac{1}{N+1} = \frac{\mathbf{E}[M_k]}{k}$$



Cardinality

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

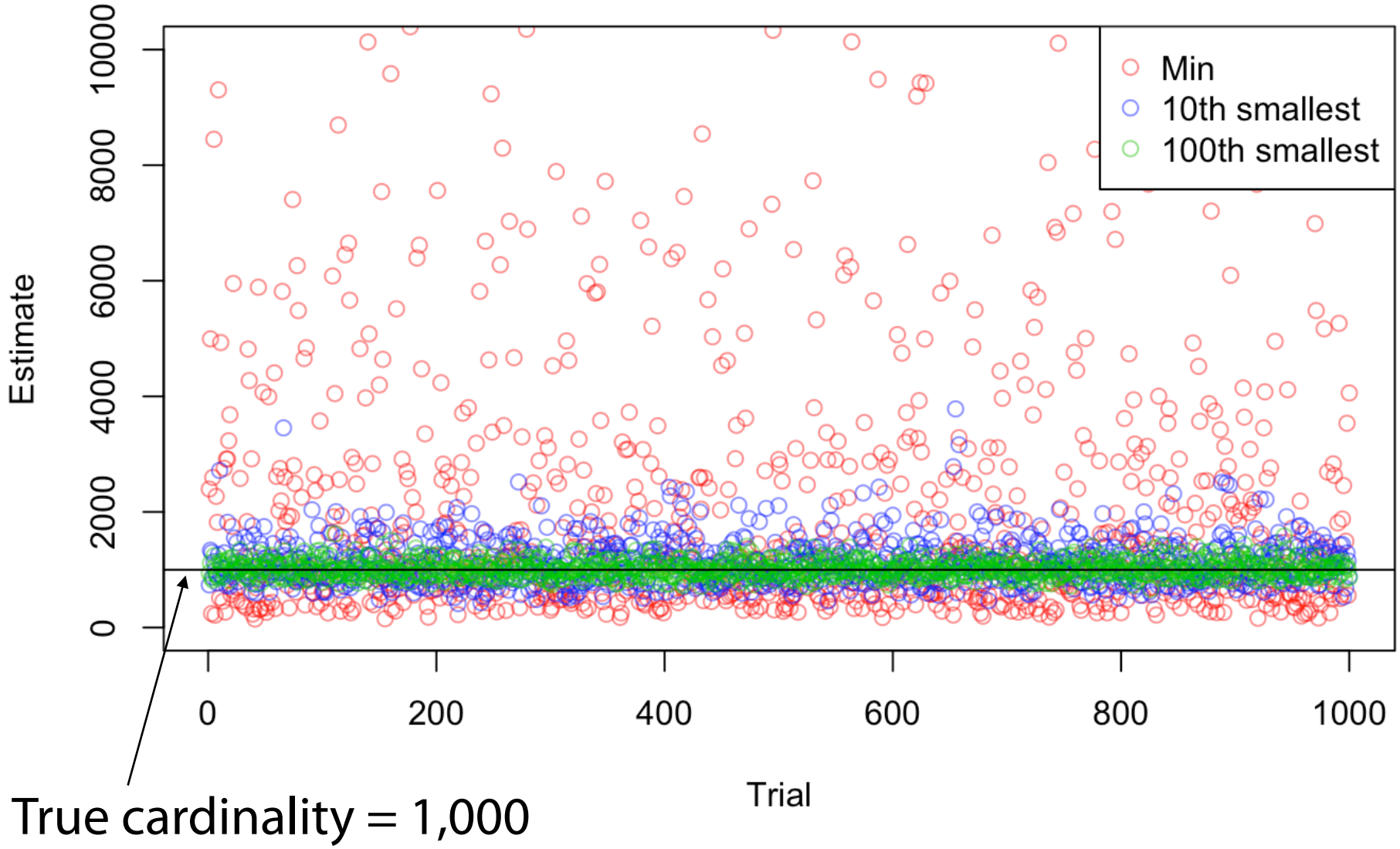
$$= \left[\mathbf{E}[M_1] + (\mathbf{E}[M_2] - \mathbf{E}[M_1]) + \ldots + (\mathbf{E}[M_k] - \mathbf{E}[M_{k-1}])\right] \cdot \frac{1}{k}$$



k^{th} minimum value (KMV)

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

Cardinality



Cardinality

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

Hash

0.135	0.220	0.236	0.36	0.41

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 Operations

$A = \{1, 2, 3, 4\}$ $B = \{3, 4, 5, 6, 7\}$

Union $A \cup B$

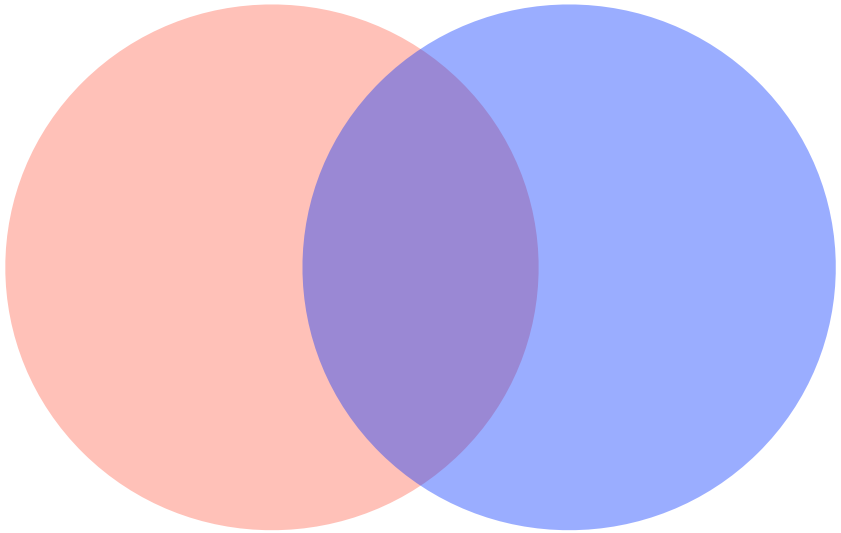
Intersection $A \cap B$

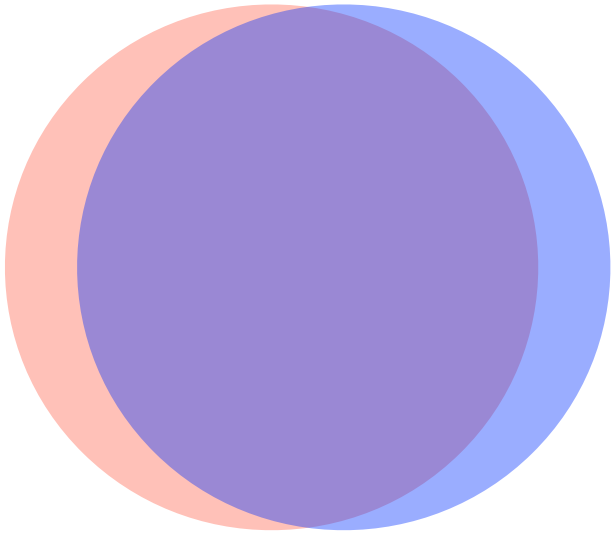
Difference A / B

Symmetric difference $A \triangle B$

Set Similarity

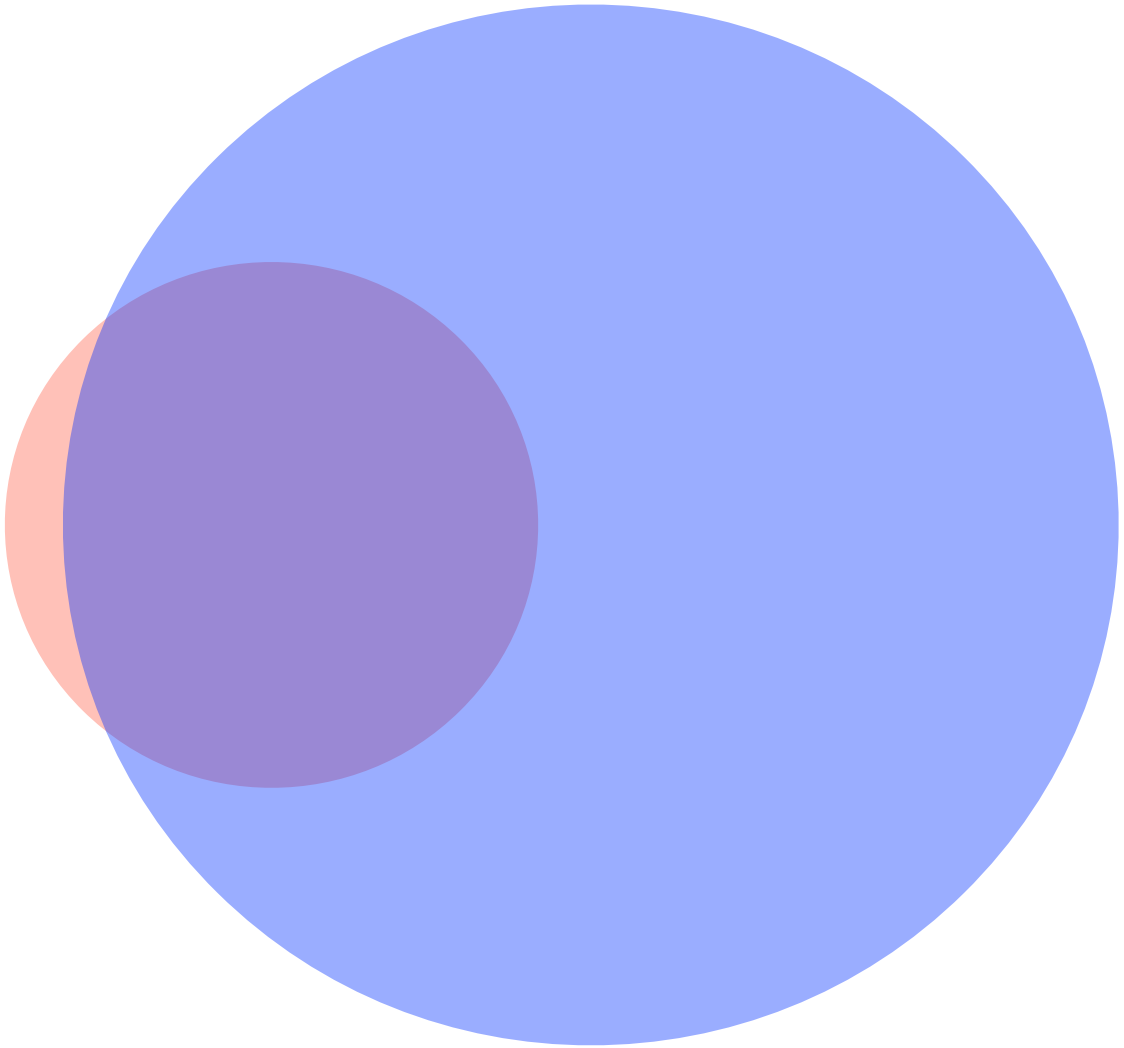
How can we describe how ***similar*** two sets are?

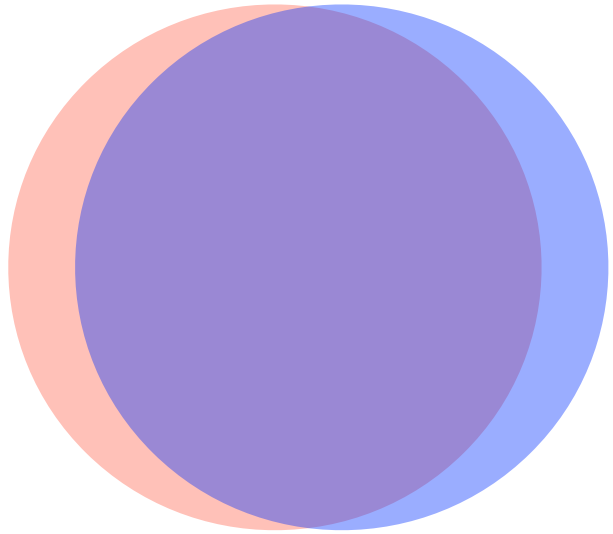




Set Similarity

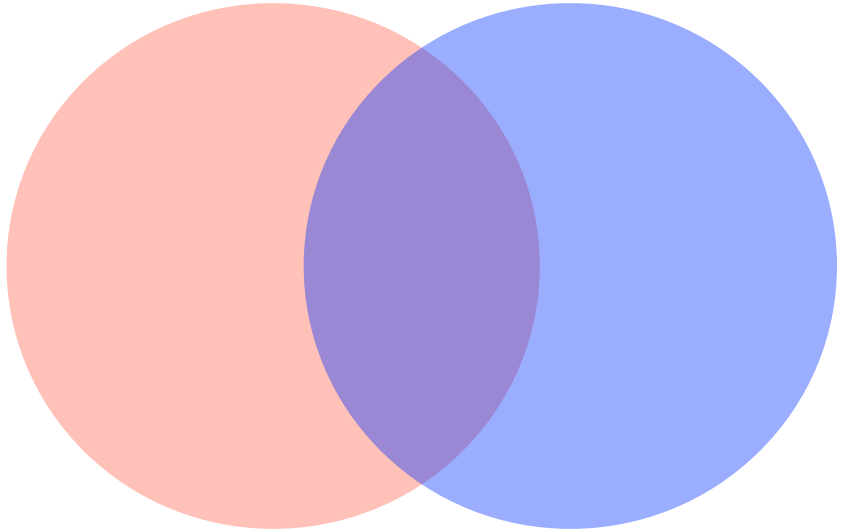
How can we describe how ***similar*** two sets are?





Set Similarity

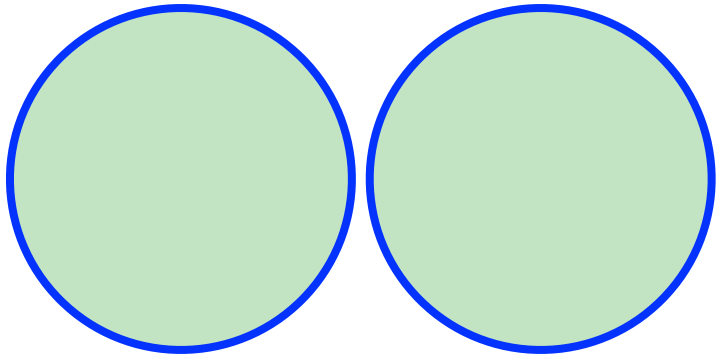
To measure **similarity** of A & B, we need both a measure of how similar the sets are but also the total size of both sets.



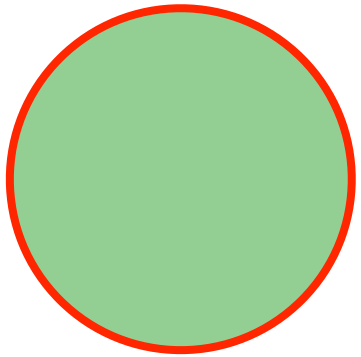
$$J = \frac{|A \cap B|}{|A \cup B|}$$

J is the ***Jaccard coefficient***

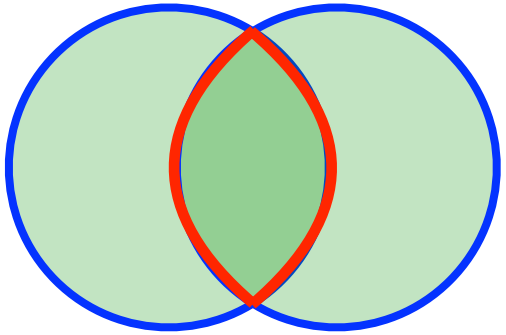
Set Similarity



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



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



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

Set Similarity

$$A = \{1, 2, 3, 4\} \qquad B = \{3, 4, 5, 6, 7\}$$

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

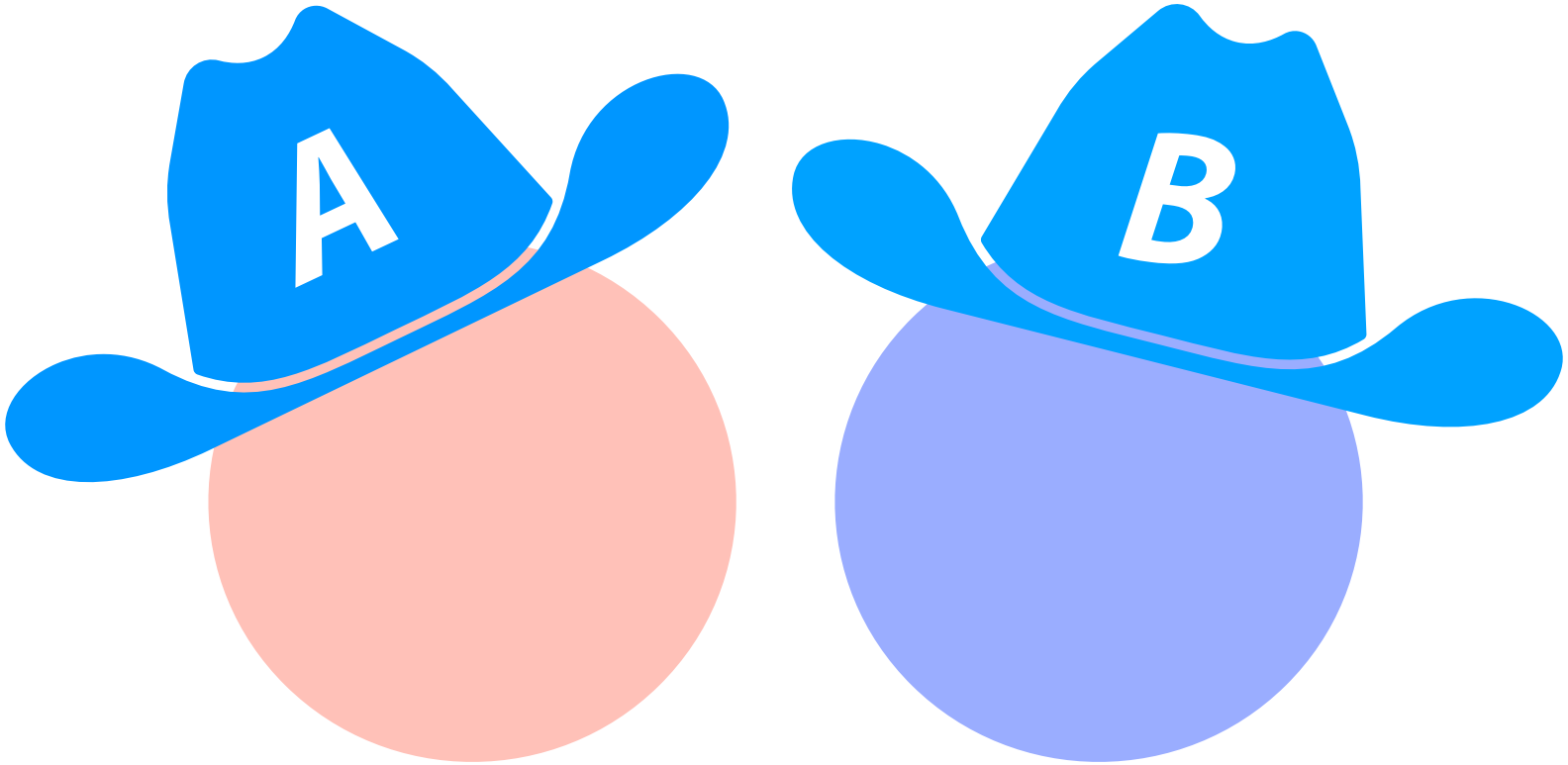
Set Similarity

$$A = \{1, 2, 3, 4\} \qquad B = \{3, 4, 5, 6, 7\}$$

$$J = \frac{|A \cap B|}{|A \cup B|} = \frac{|\{3,4\}|}{|\{1,2,3,4,5,6,7\}|} = \frac{2}{7}$$

Similarity Sketches

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



Similarity Sketches

Imagine we 'sketched' two datasets by hashing all objects…

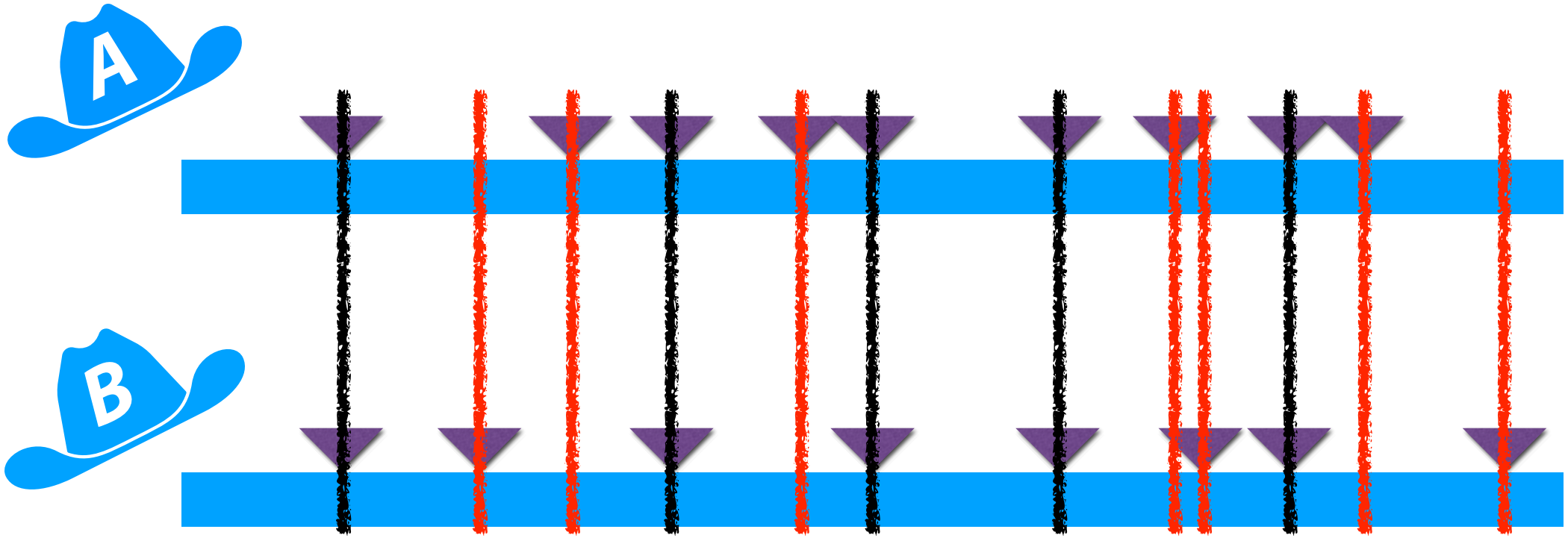


Image inspired by: Ondov B, Starrett G, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. **Mash Screen: high-throughput sequence containment estimation for genome discovery**. *Genome Biol* 20, 232 (2019)

Similarity Sketches

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

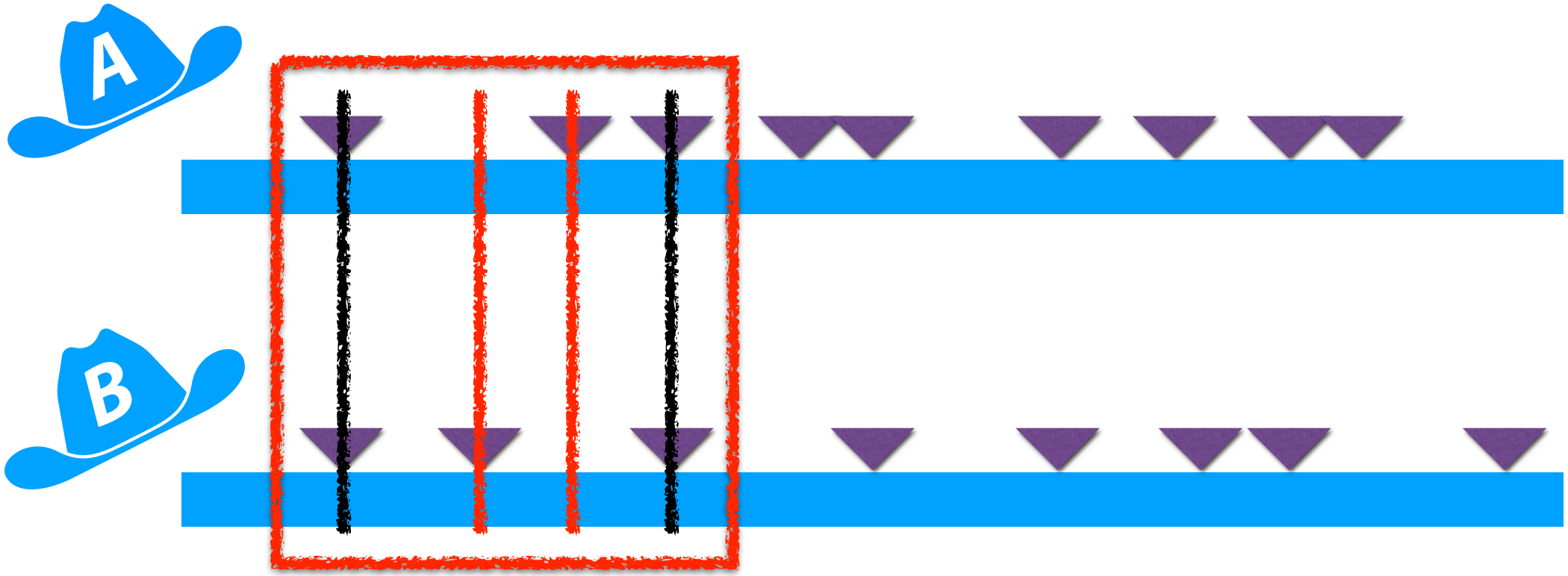


Image inspired by: Ondov B, Starrett G, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. **Mash Screen: high-throughput sequence containment estimation for genome discovery**. *Genome Biol* 20, 232 (2019)

Similarity Sketches

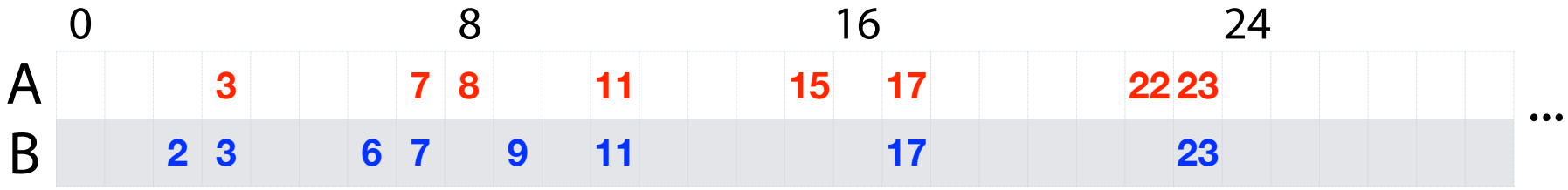
Say we find the 8 minimum hashes (bottom-8) for items in set A and B

Sketch A

3	15
7	17
8	22
11	23

Sketch B

2	9
3	11
6	17
7	23



Similarity Sketches

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

Sketch A

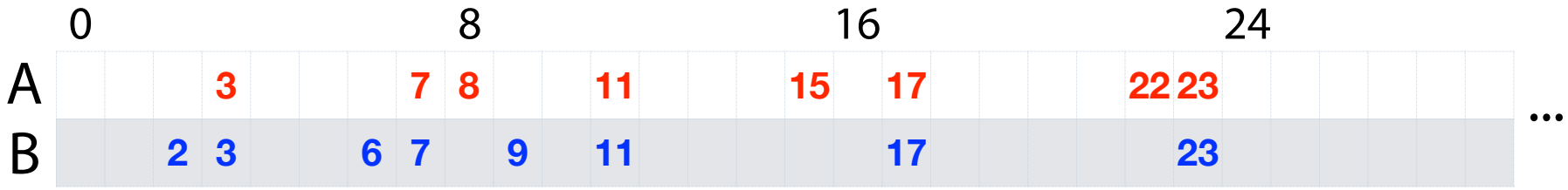
3	15
7	17
8	22
11	23

$\cup$

Sketch B

2	9
3	11
6	17
7	23

$=$



Similarity Sketches

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

Sketch A

3	15
7	17
8	22
11	23

$\cup$

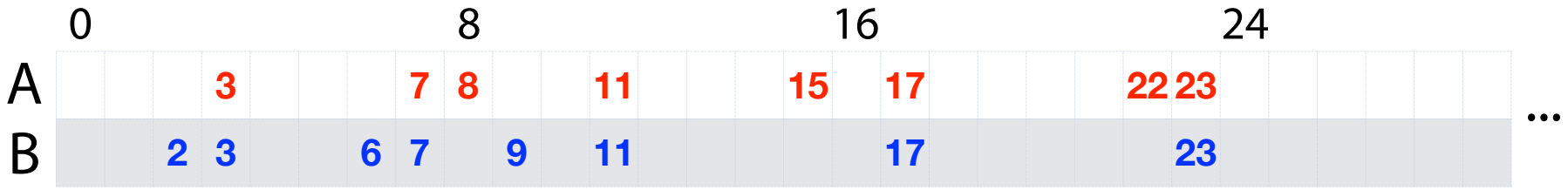
Sketch B

2	9
3	11
6	17
7	23

$=$

$|A \cup B|$

2	8
3	9
6	11
7	15



Similarity Sketches

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

Sketch A

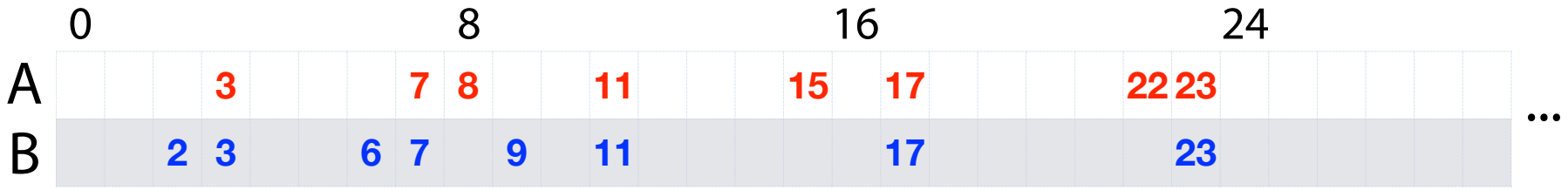
3	15
7	17
8	22
11	23

$\cap$

Sketch B

2	9
3	11
6	17
7	23

$=$



Similarity Sketches

Claim: Can approximate the intersection of our sketches as our datasets!

Sketch A

3	15
7	17
8	22
11	23

Sketch B

2	9
3	11
6	17
7	23

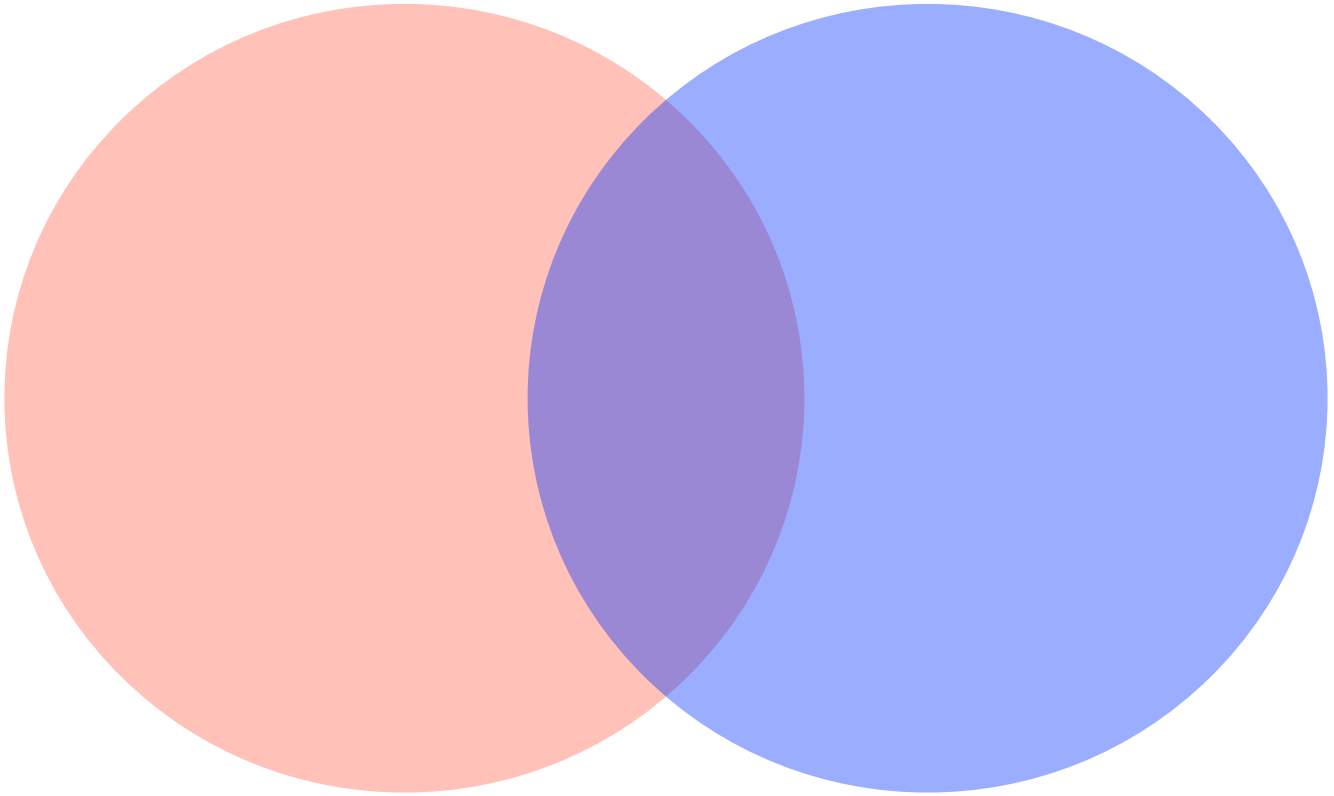
$|A \cup B|$

2	8
3	9
6	11
7	15

$$|A \cap B| \approx |S(A \cup B) \cap S(A) \cap S(B)|$$

Inclusion-Exclusion Principle

$$|A \cap B| =$$



Similarity Sketches

Claim: Can approximate the intersection of our sketches as our datasets!

Sketch A

3	15
7	17
8	22
11	23

Sketch B

2	9
3	11
6	17
7	23

$|A \cup B|$

2	8
3	9
6	11
7	15

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

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

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

$$\approx 0.29$$

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

Similarity Sketches

Claim: Can approximate the intersection of our sketches as our datasets!

Sketch A

3	15
7	17
8	22
11	23

Sketch B

2	9
3	11
6	17
7	23

$|A \cup B|$

2	8
3	9
6	11
7	15

All computation here is simple

- Hash functions
- Bottom k (heap / sorted list)
- k^{th} minimum value (lookup)
- Get union sketch (merge heaps / lists)
- Calculate Jaccard (during merge)

1) Sequence decomposed into **kmers**

S_1 : CATGGACCGACCAG GCAGTACCGATCGT : S_2

CAT GAC GAC GTA CGA CGT

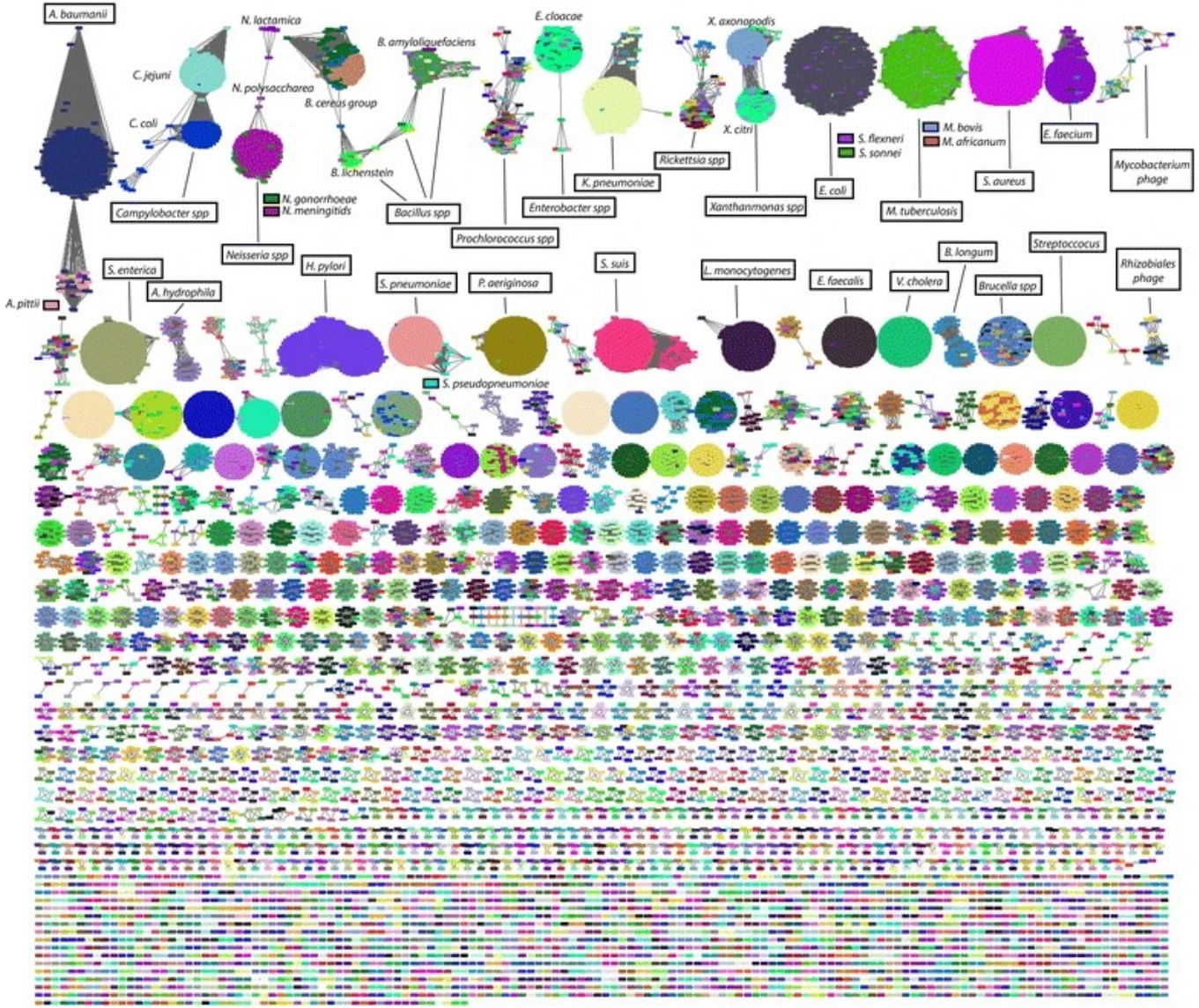
ATG ACC ACC AGT CCG TCG

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*

Sketching Summary

If my dataset is too large to handle, I can still answer many questions:

Does my object exist in a set?

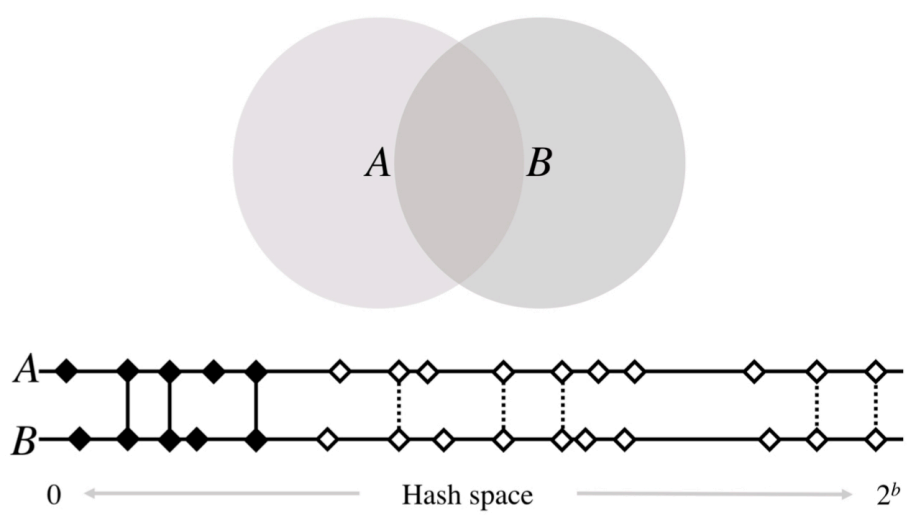
How often is a particular value repeated?

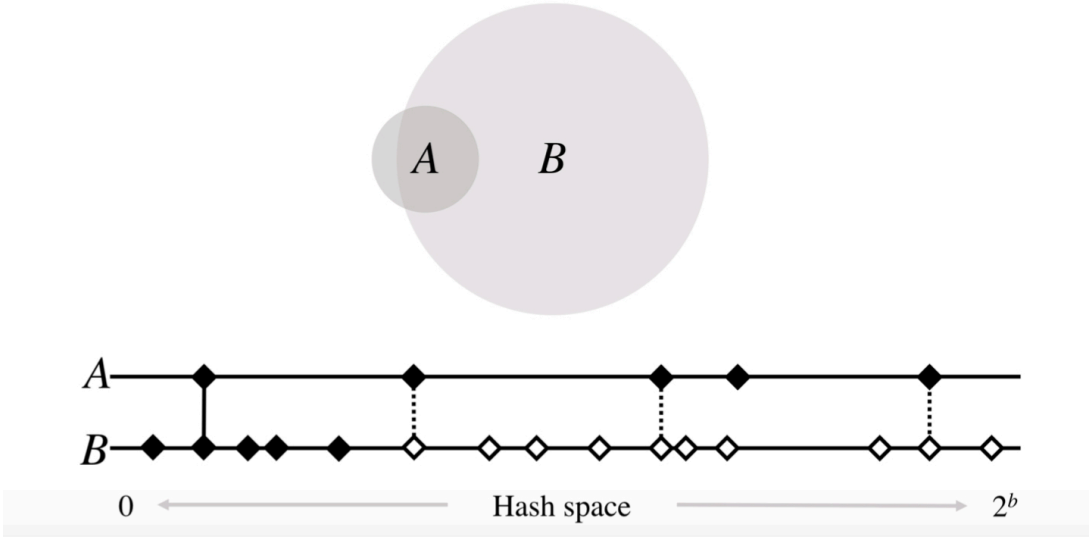
How many unique objects do I have?

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

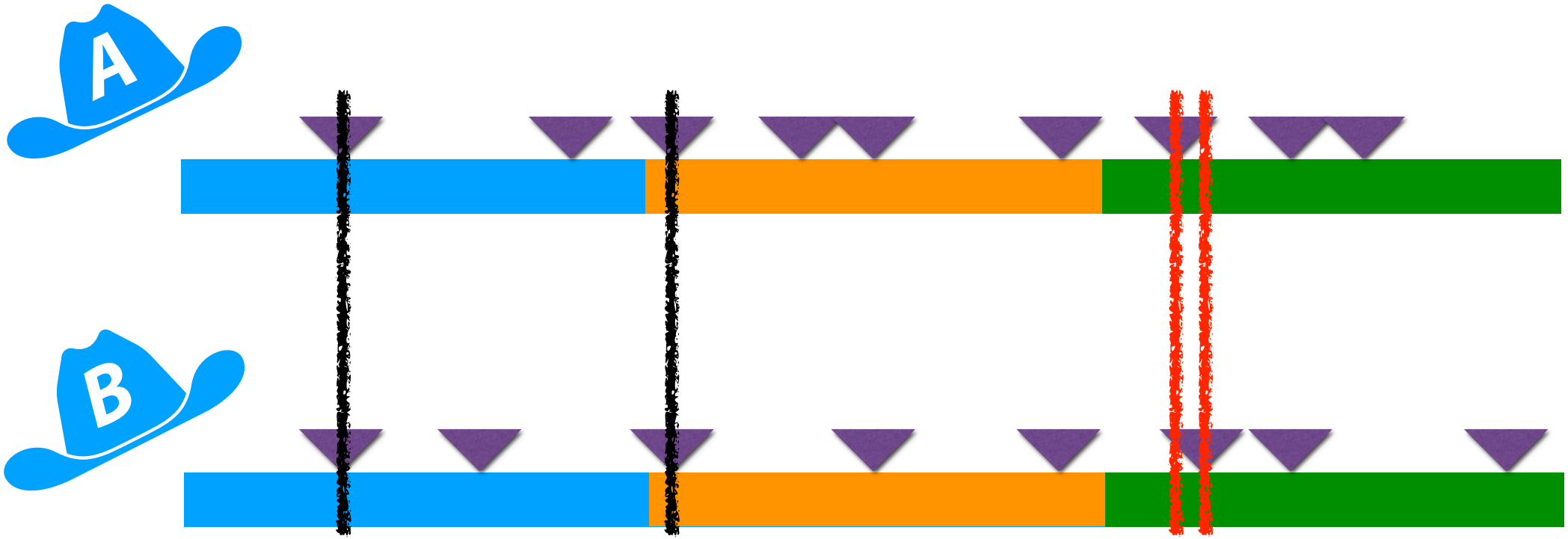




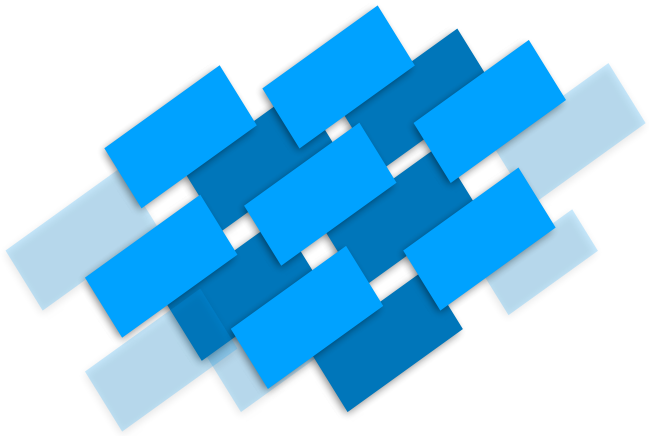
Ondov, Brian D., Gabriel J. Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B. Buck, and Adam M. Phillippy. **Mash Screen: High-throughput sequence containment estimation for genome discovery**. *Genome biology* 20.1 (2019): 1-13.

K-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

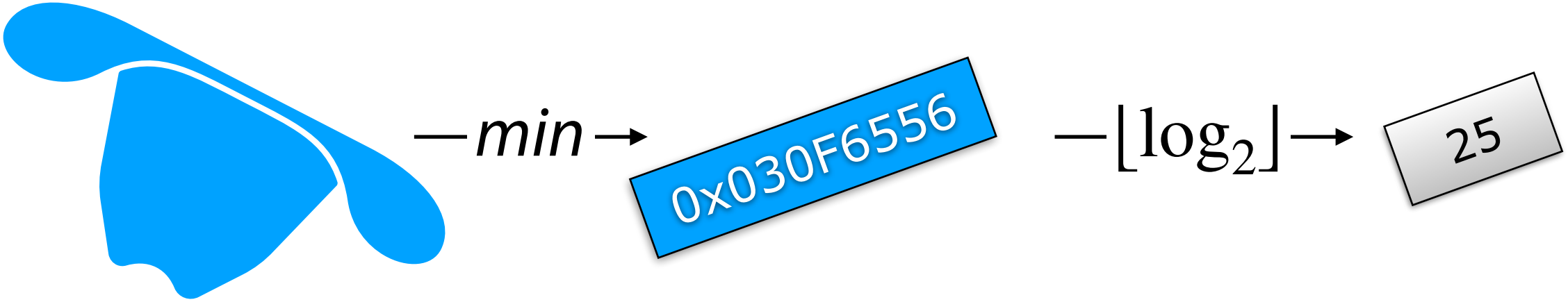
11010110

11

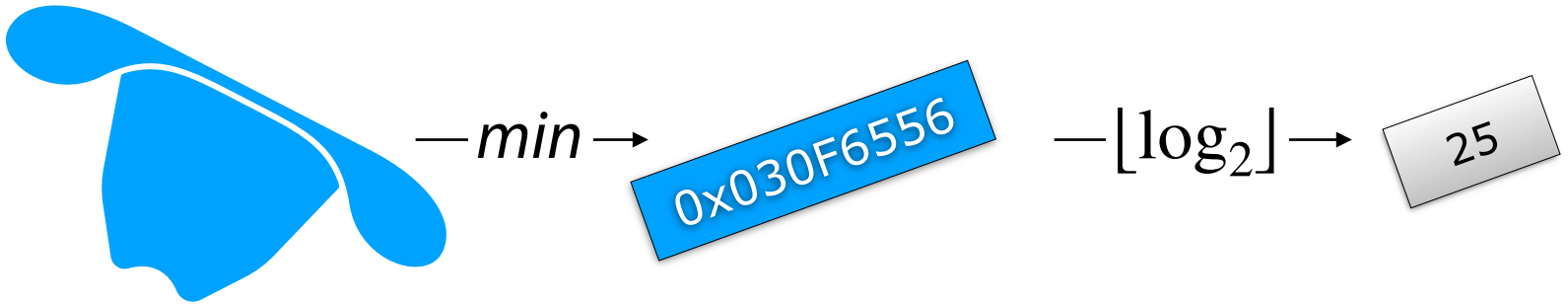
01101011

HyperLogLog

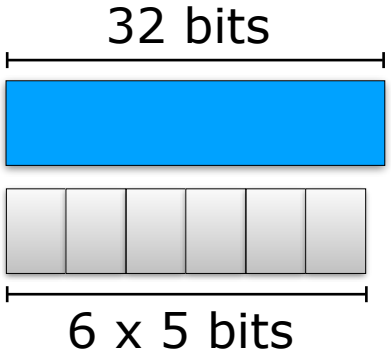
Instead of *minimum*, say we use *log-minimum*



HyperLogLog

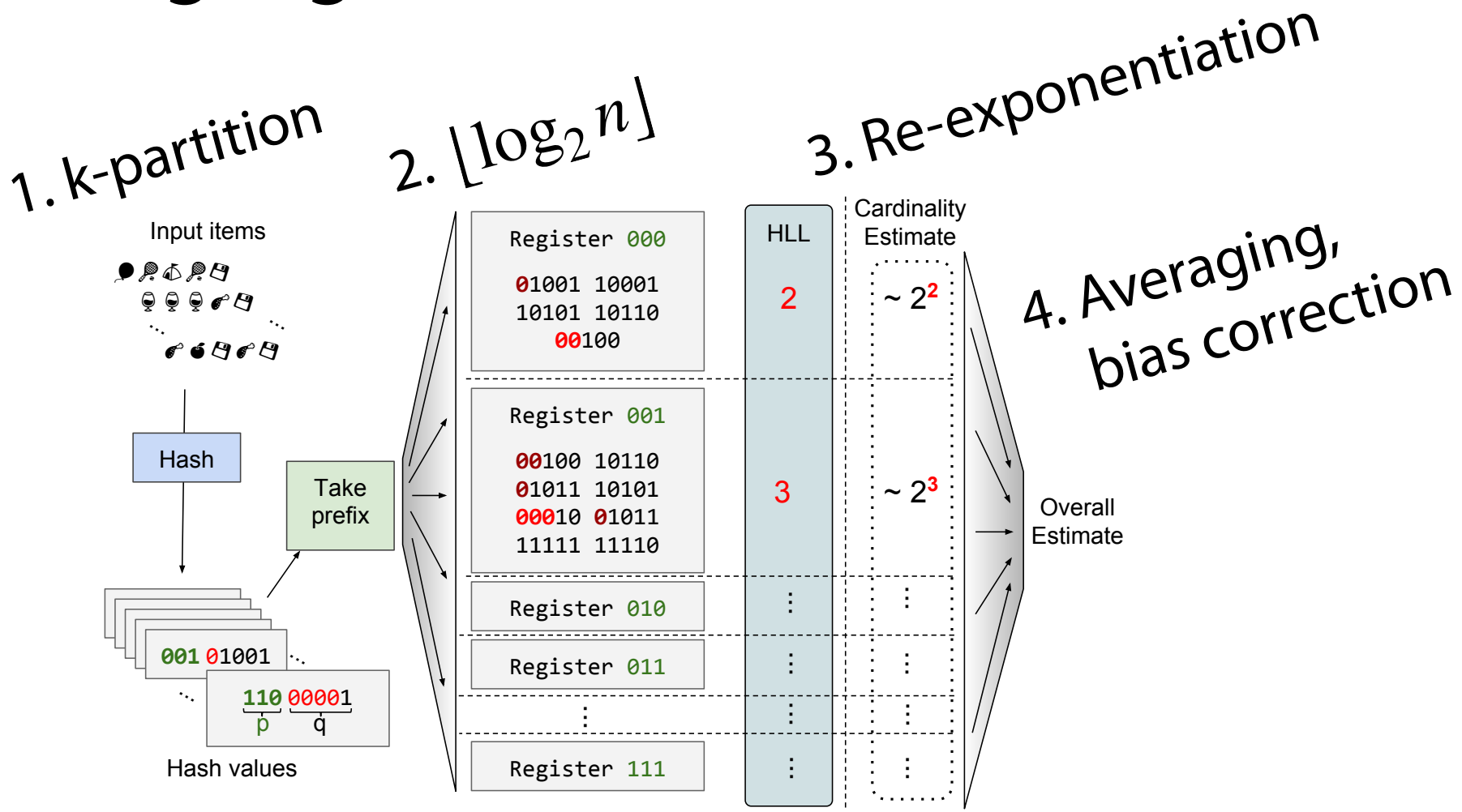


Pro: Representatives take $\log \log U$ rather than $\log U$ bits



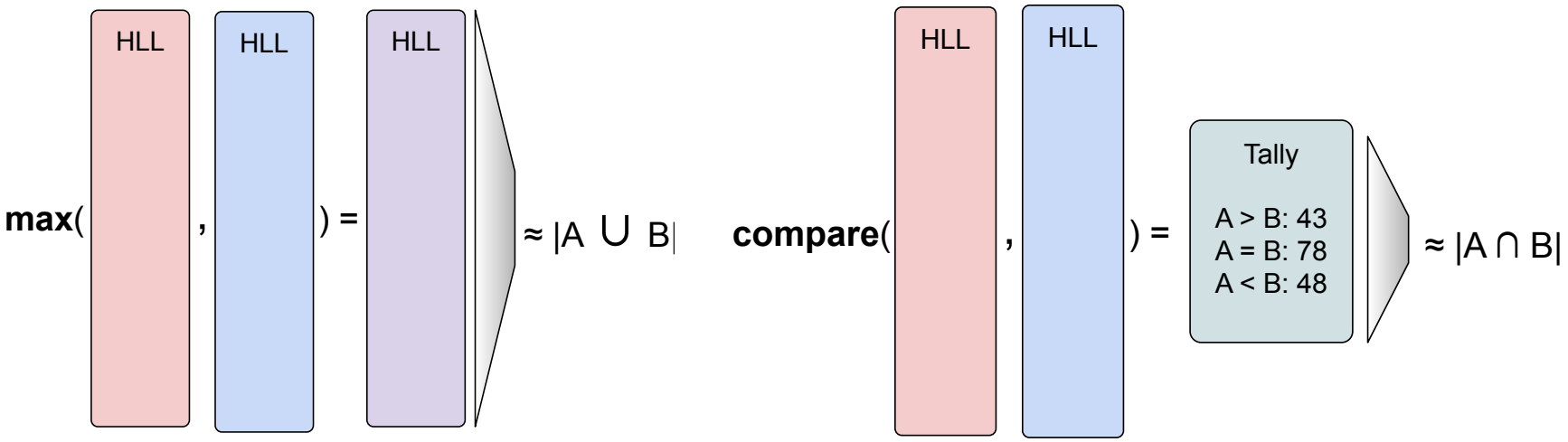
Con: Estimate is of $\lfloor \log_2 n \rfloor$; can re-exponentiate later, but with added variance & bias

HyperLogLog



Baker DN, Langmead B. **Dashing: fast and accurate genomic distances with HyperLogLog**. In press, *Genome Biology*.

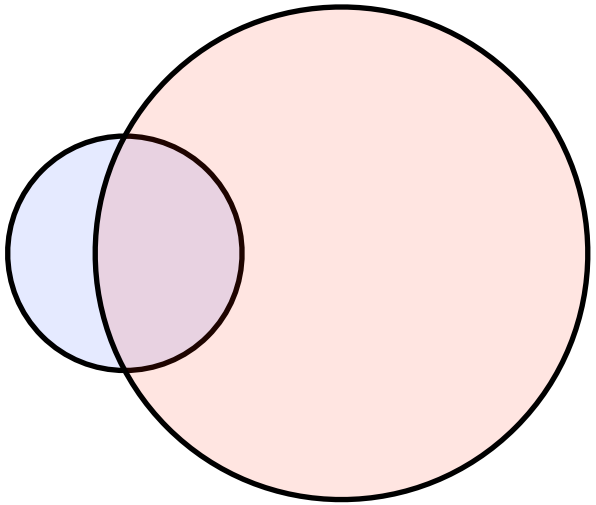
HyperLogLog

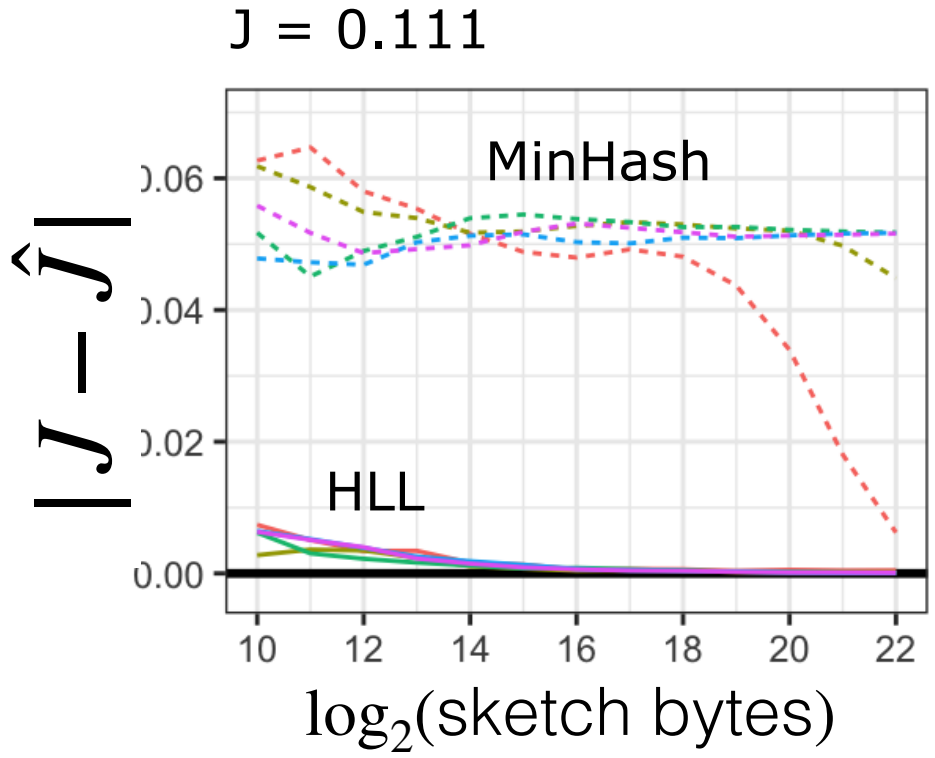


Union *and intersection* cardinalities can be estimated *directly*. No need for $|A \cap B| \approx |S(A \cup B) \cap S(A) \cap S(B)|$.

HyperLogLog

HLL handles lopsided sets better than bottom-k MinHash [1,2]

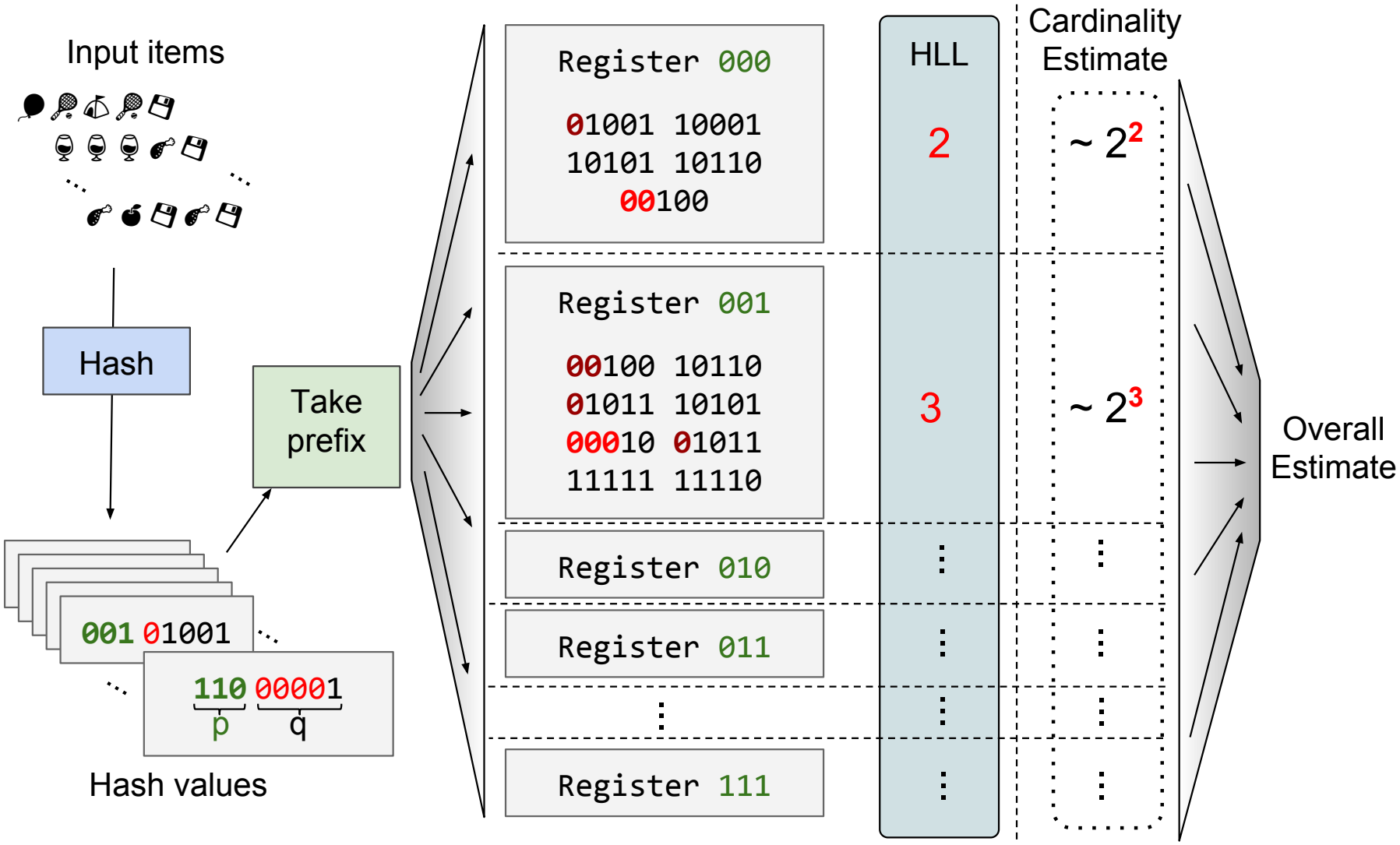




1. Koslicki, David, and Hooman Zabeti. **Improving MinHash via the containment index with applications to metagenomic analysis**. *Applied Mathematics and Computation* 354 (2019): 206-215.

2. 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)

HyperLogLog



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