Data Structures and Algorithms
MinHash Sketch

CS 225
Brad Solomon

April 24, 2024

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Department of Computer Science

Learning Objectives

Finish discussing bloom filters

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

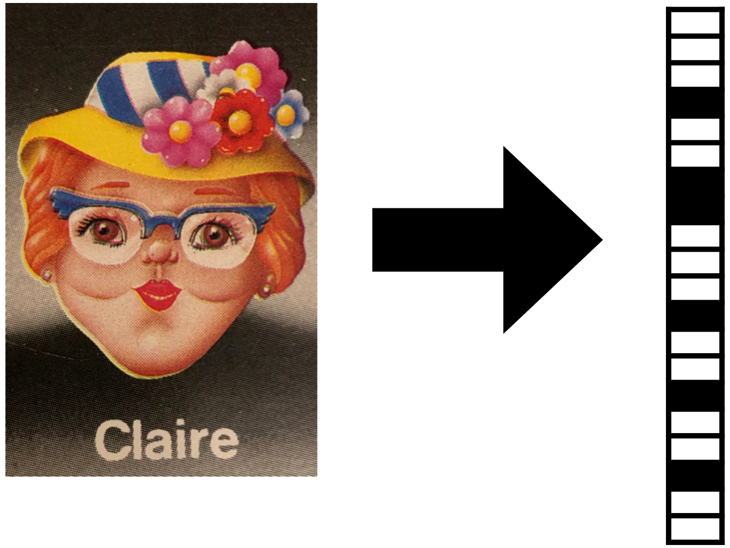
Has three key properties:

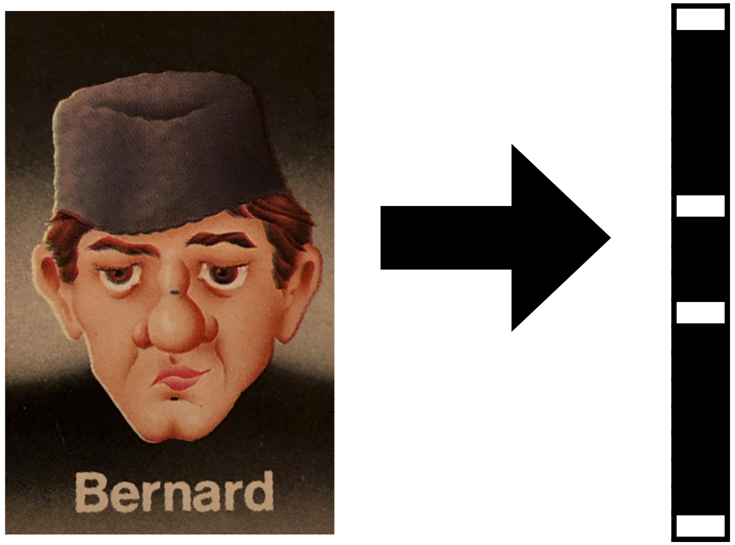
- k, number of hash functions
- n, expected number of insertions
- m, filter size in bits

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

The hidden problem with (most) 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

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

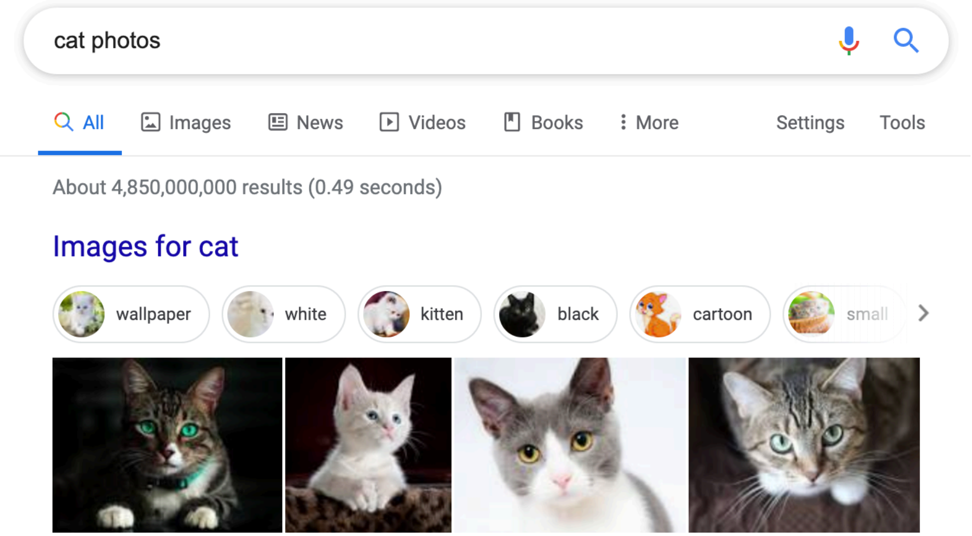
```
AGGCCACAGTGTATTATGACTG
|||||||||||  |||||||||
AGGCCACAGTGAGTTATGACTG

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

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

Cardinality

Sometimes its not possible or realistic to count all objects!


cat photos

About 4,850,000,000 results (0.49 seconds)

Images for cat

Estimate: 60 billion — 130 trillion

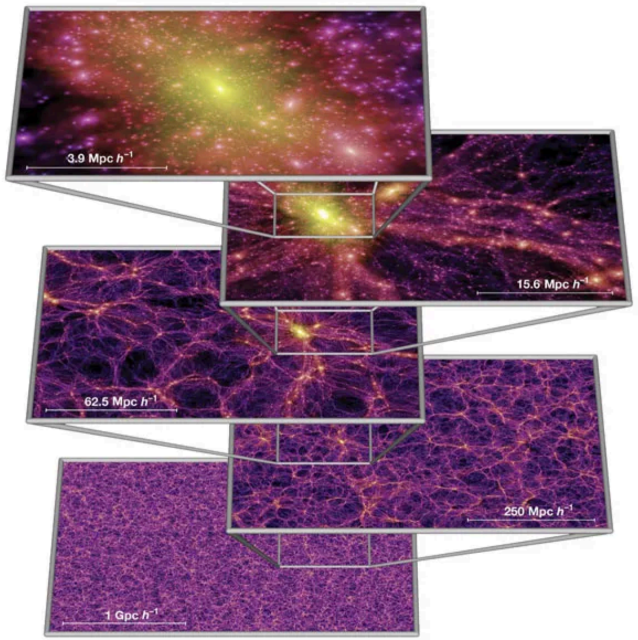


Image: https://doi.org/10.1038/nature03597

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

Applied Cardinalities

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

Given any two values, we can optimize the third

$n = 100$ items $k = 3$ hashes $m =$

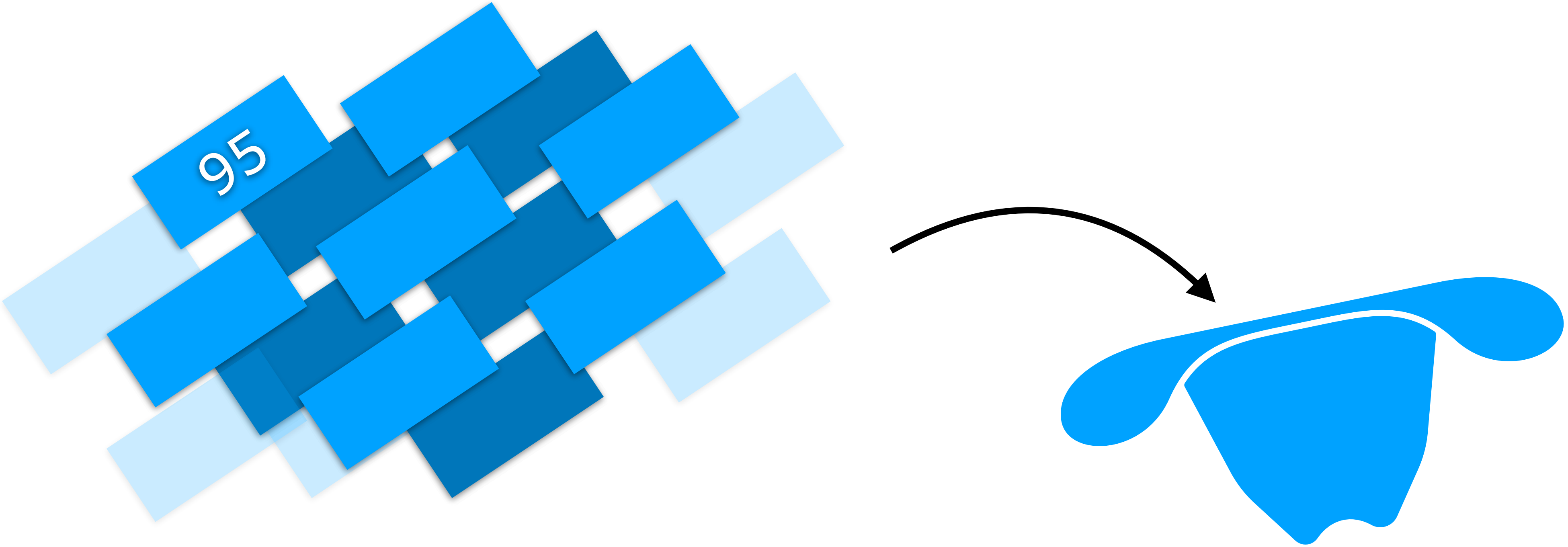
$m = 100$ bits $n = 20$ items $k =$

$m = 100$ bits $k = 2$ items $n =$

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?

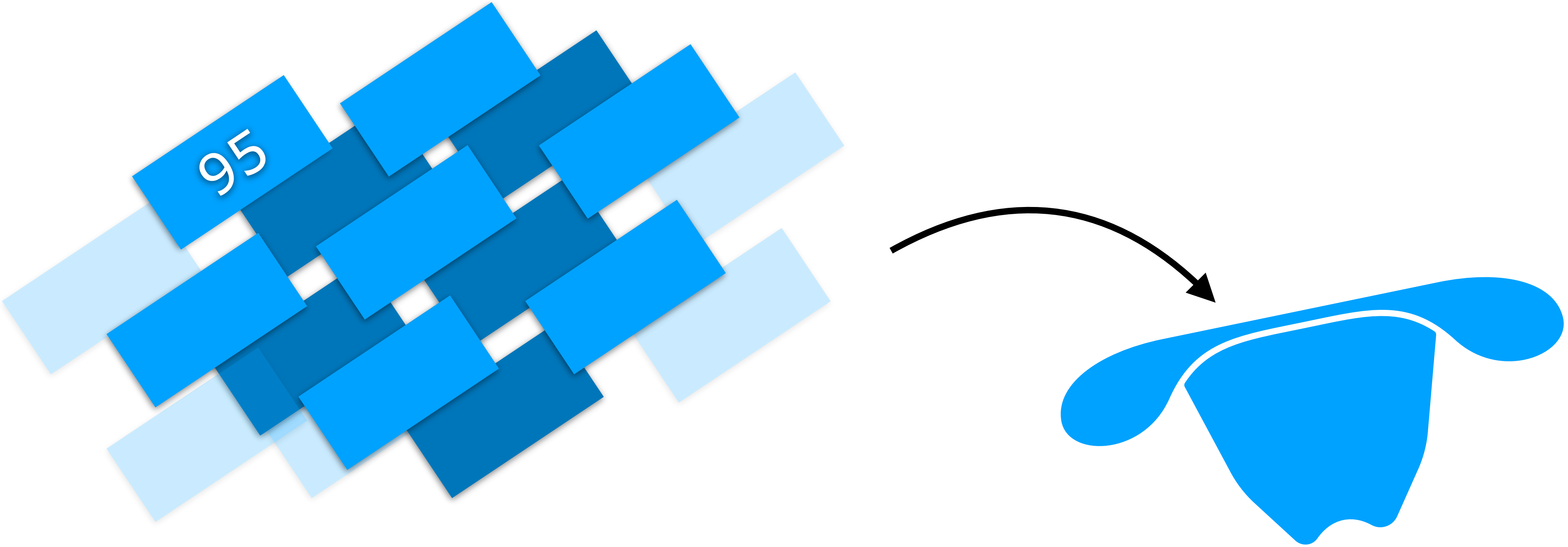


Analogy from Ben Langmead

Cardinality Estimation

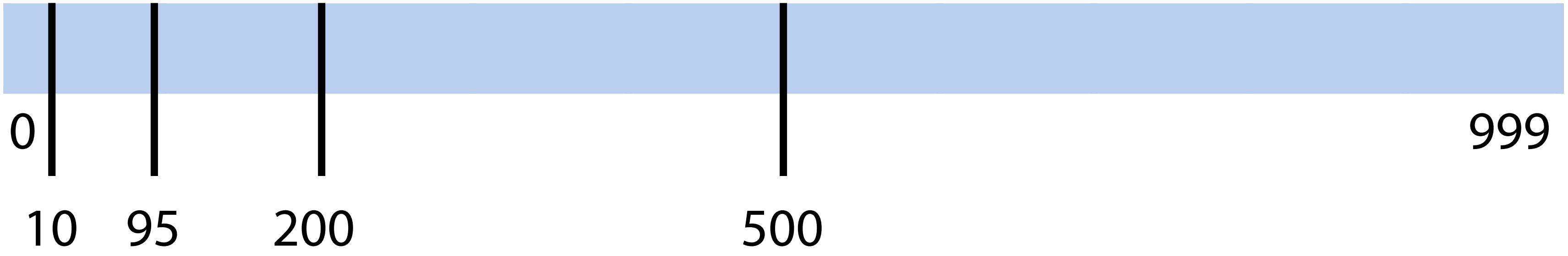
Imagine I fill a hat with **a random subset** of numbered cards **from 0 to 999**

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?

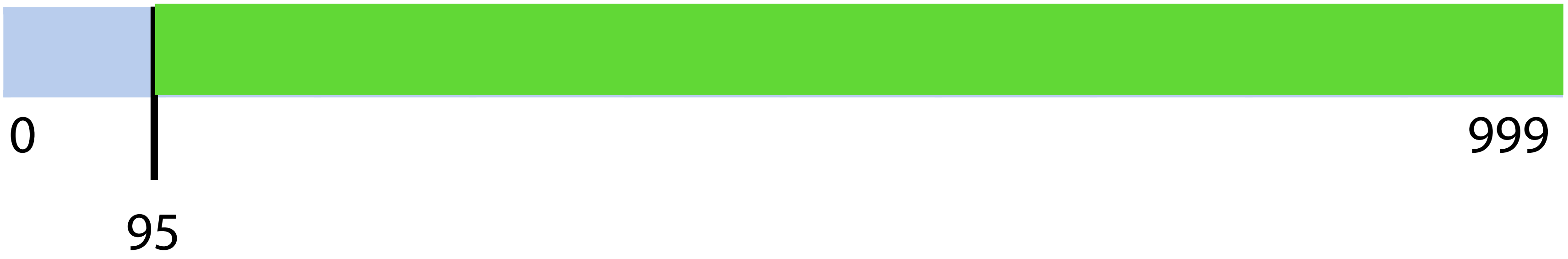
Cardinality Estimation

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

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

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:

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

$$N + 1 \approx 10.5$$

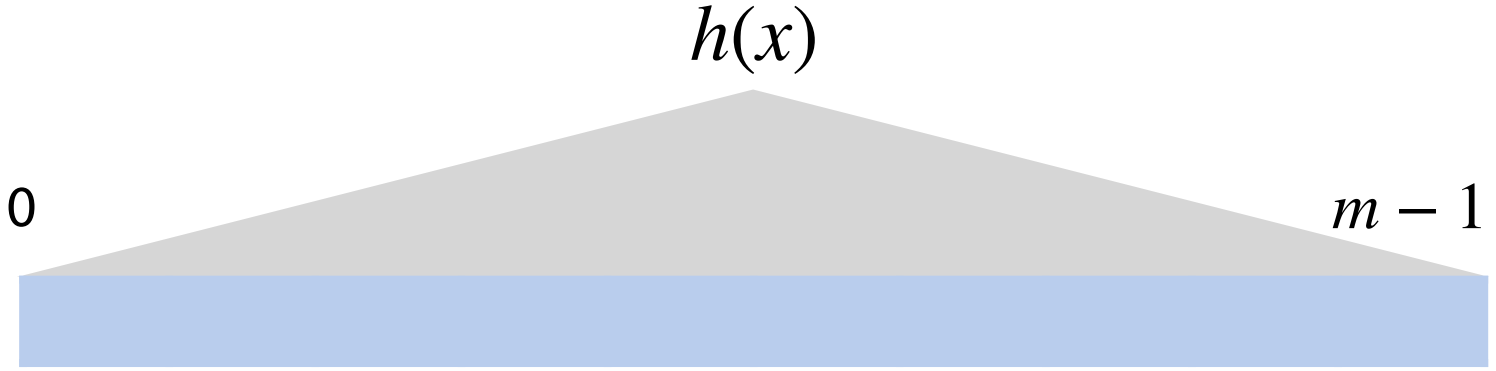
$$N \approx 9.5$$

Cardinality Estimation

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

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

Tracking only the minimum value can let us estimate the cardinality!



Cardinality Estimation

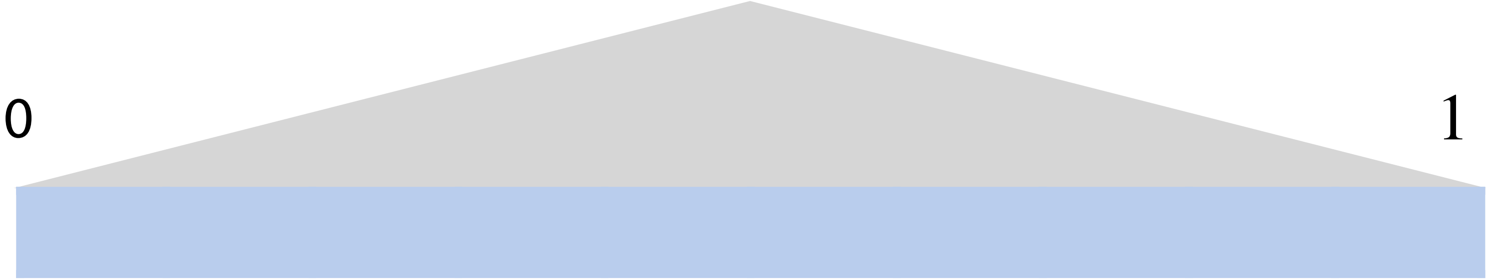
Imagine we have a SUHA hash h over a range m.

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

Tracking only the minimum value can let us estimate the cardinality!

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

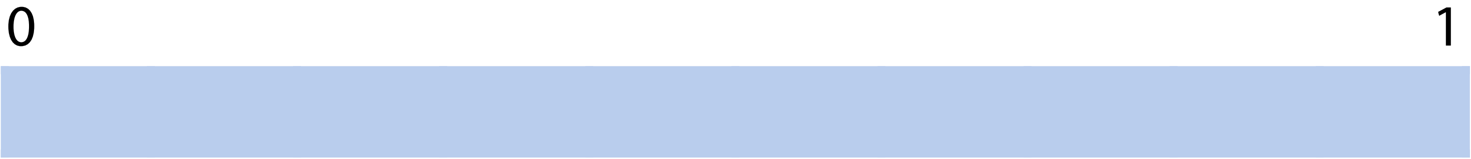
$$h'(x) = h(x)\ /\ (m - 1)$$



Cardinality Sketch

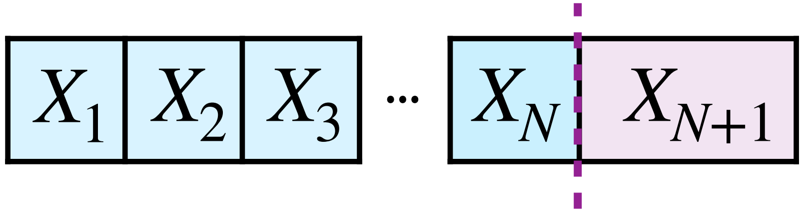
Let $M = min(X_1, X_2, \ldots, X_N)$ where each $X_i \in [0, 1]$ is an uniform independent random variable

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



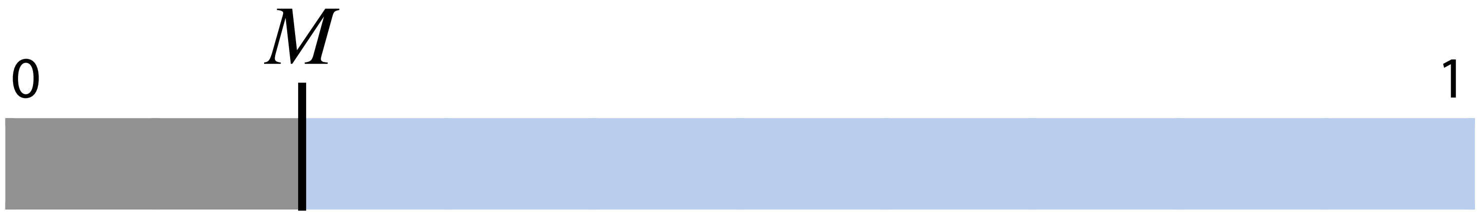
Cardinality Sketch

Consider an $N + 1$ draw:



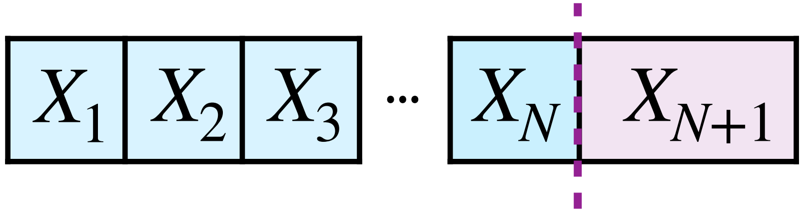
$$M = \min_{1 \leq i \leq N} X_i$$

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



Cardinality Sketch

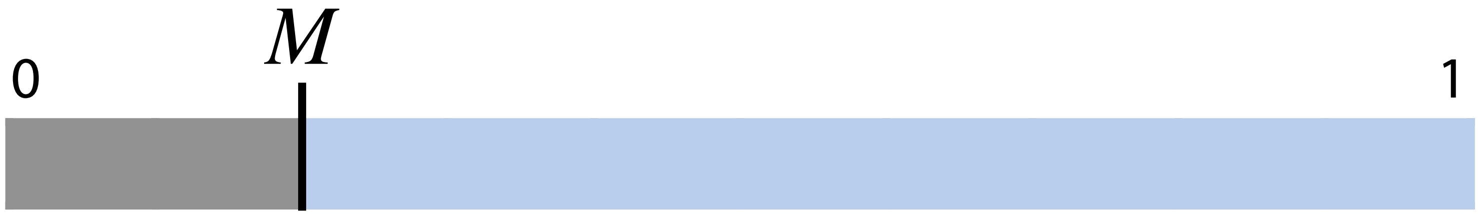
Consider an $N+1$ draw:



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

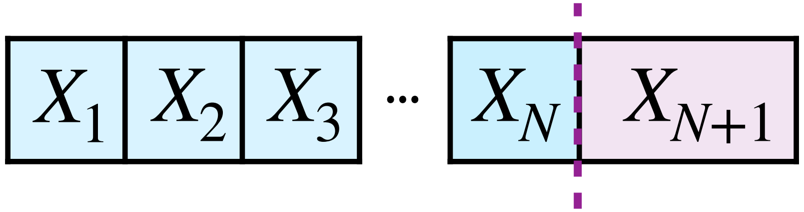
X_{N+1} can end up in one of two ranges:

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



Cardinality Sketch

Consider an $N + 1$ draw:

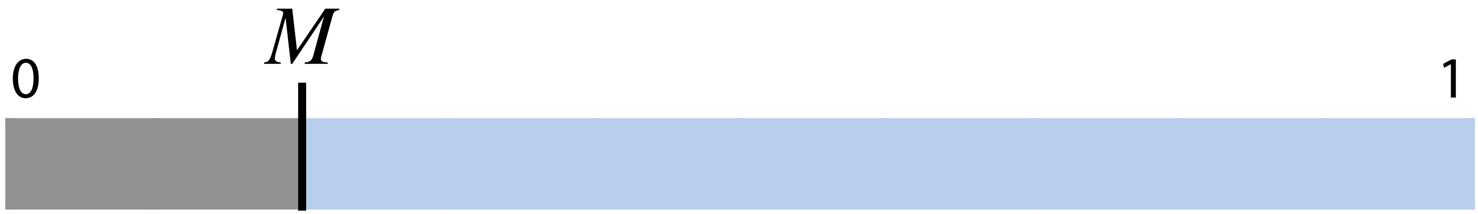


$$M = \min_{1 \le i \le N} X_i$$

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

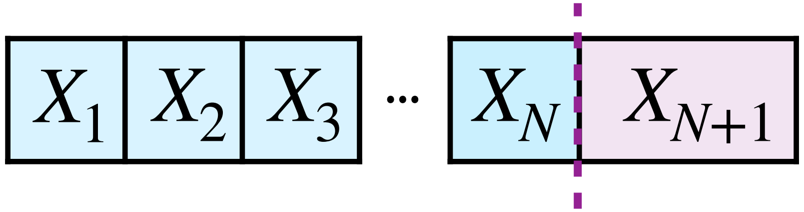
X_{N+1} will be the new minimum with probability M

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



Cardinality Sketch

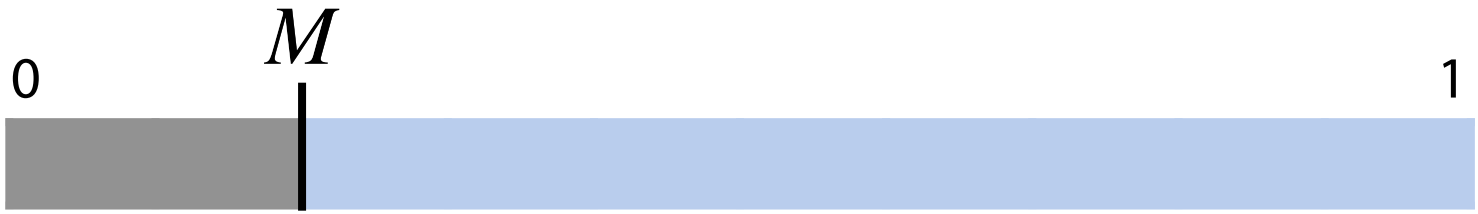
Consider an $N + 1$ draw:



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

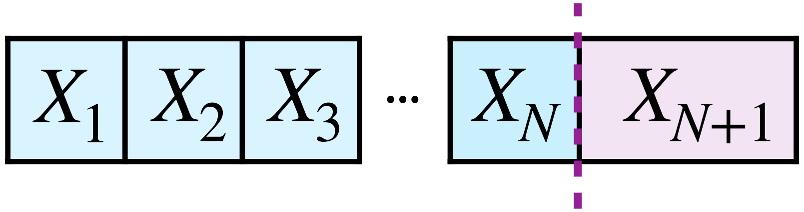
X_{N+1} **will be the new minimum with probability** M

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



Cardinality Sketch

Consider an $N + 1$ draw:

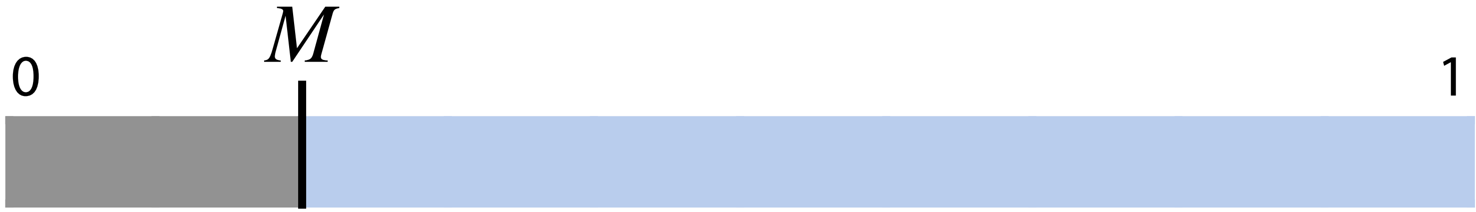


$$M = \min_{1 \le i \le N} X_i$$

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

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

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



Cardinality Sketch

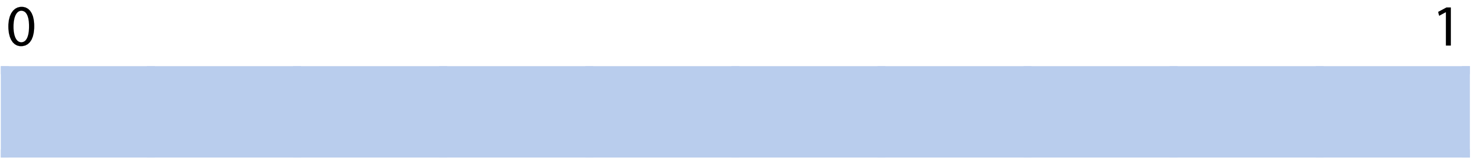
Claim: $\mathbf{E}[M] = \dfrac{1}{N+1}$

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

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

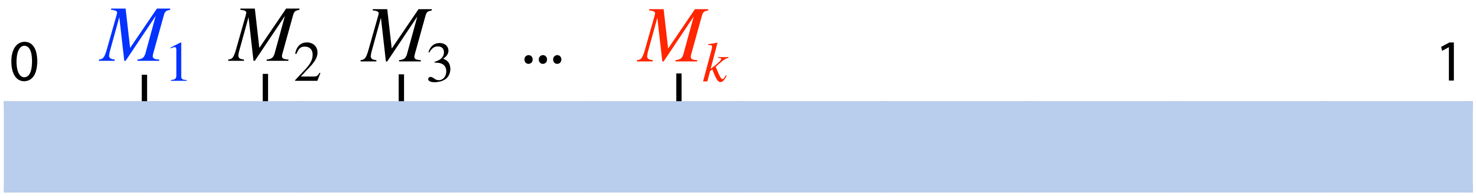
The minimum hash is a valid sketch of a dataset but can we do better?



Cardinality Sketch

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

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

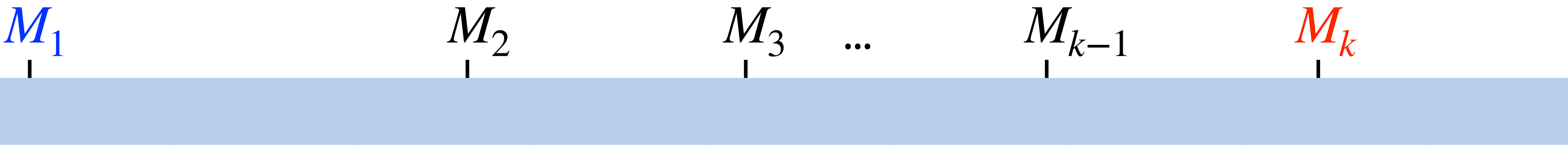


Cardinality Sketch

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

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

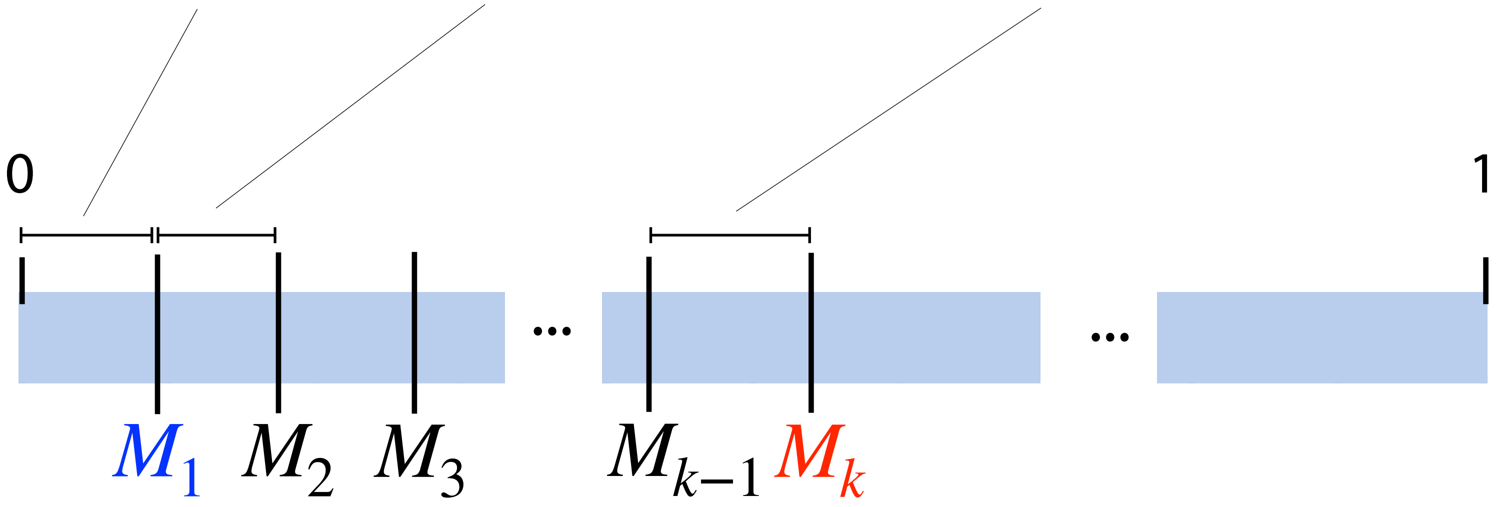
$$= \left[\mathbf{E}[M_1] + (\mathbf{E}[M_2] - \mathbf{E}[M_1]) + \ldots + (\mathbf{E}[M_k] - \mathbf{E}[M_{k-1}])\right] \cdot \frac{1}{k}$$



Cardinality Sketch

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

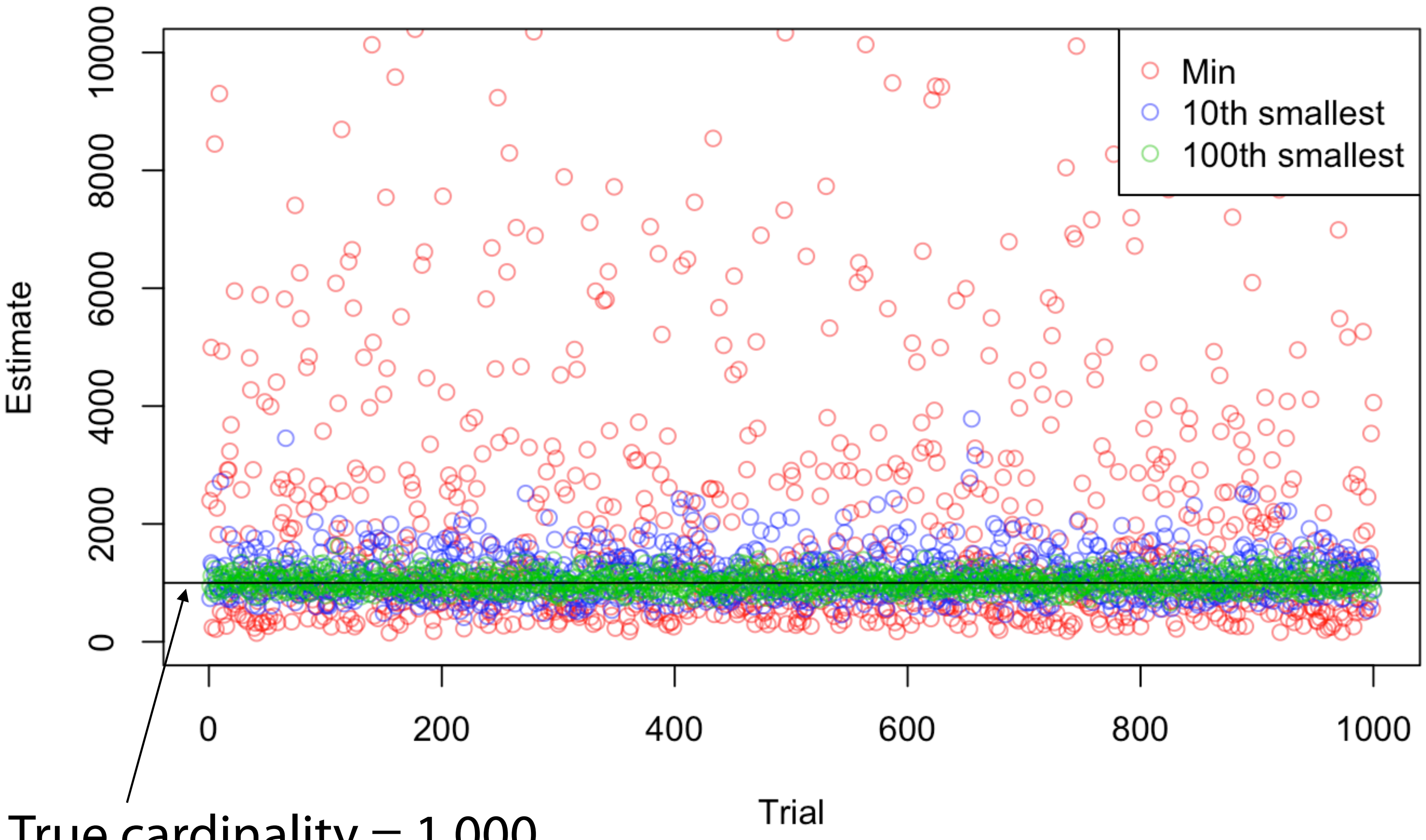
$$= \left[\underbrace{\mathbf{E}[M_1]}_{} + \underbrace{(\mathbf{E}[M_2] - \mathbf{E}[M_1])}_{} + \ldots + \underbrace{(\mathbf{E}[M_k] - \mathbf{E}[M_{k-1}])}_{}\right] \cdot \frac{1}{k}$$



k^{th} minimum value (KMV)

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

Cardinality Sketch



True cardinality = 1,000

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**.

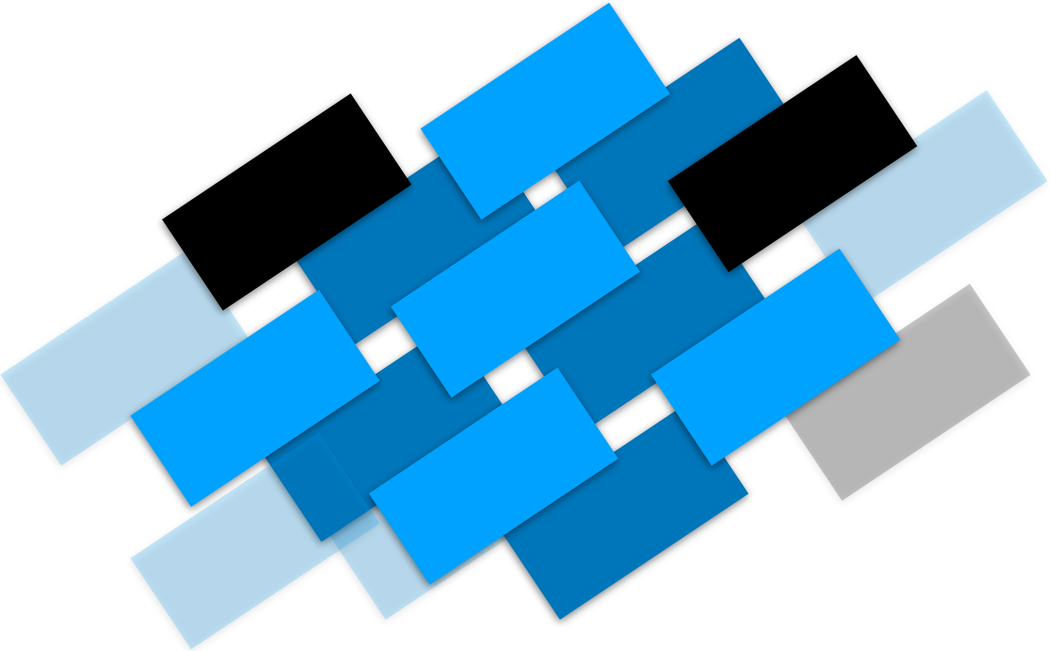
Hash

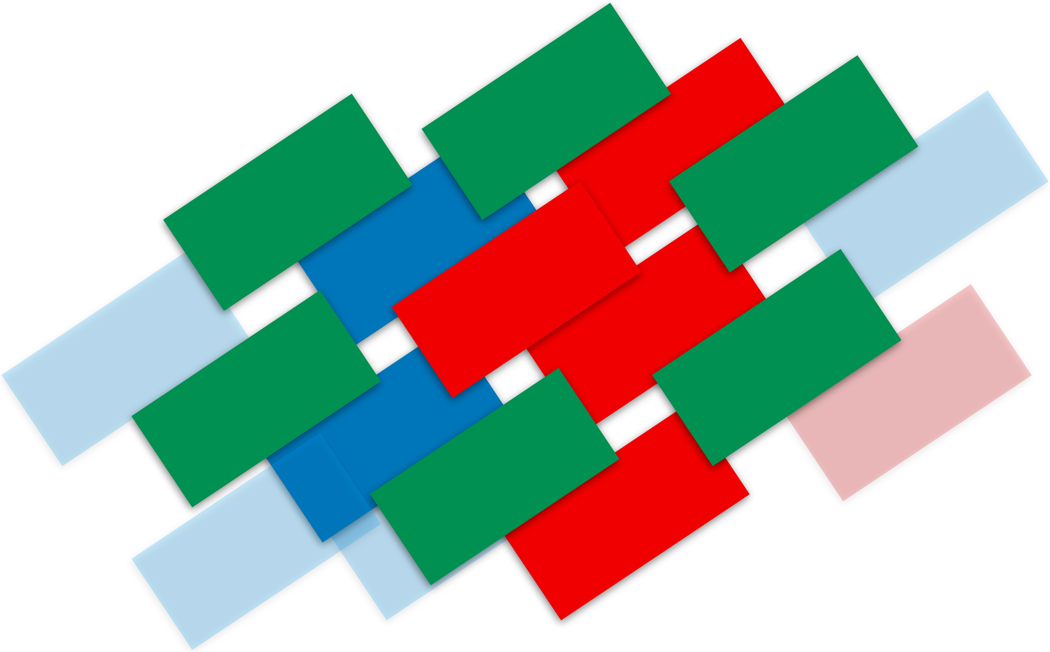
0.253	0.839	0.327	0.655	0.491

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.**



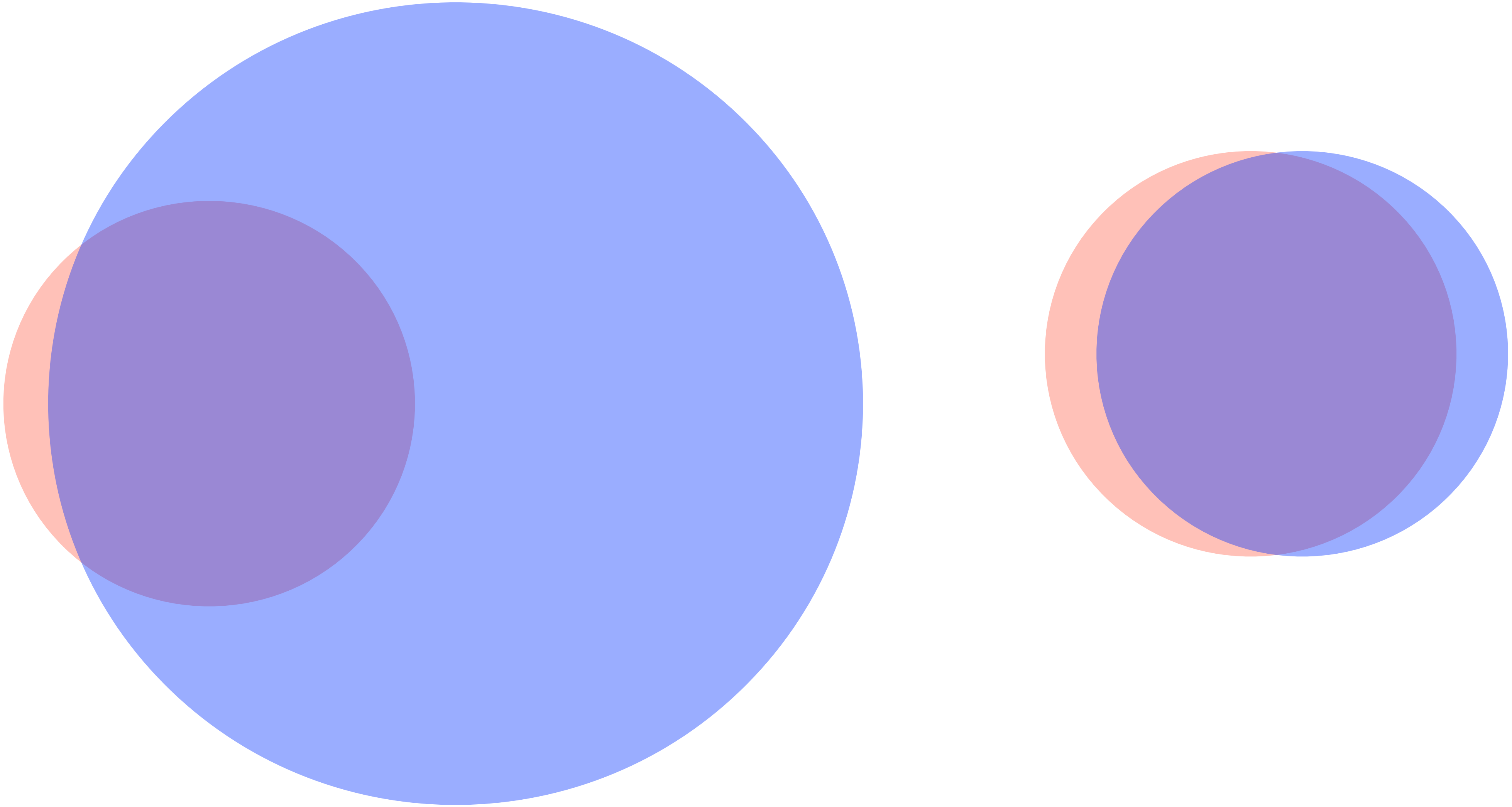


Set Similarity Review

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

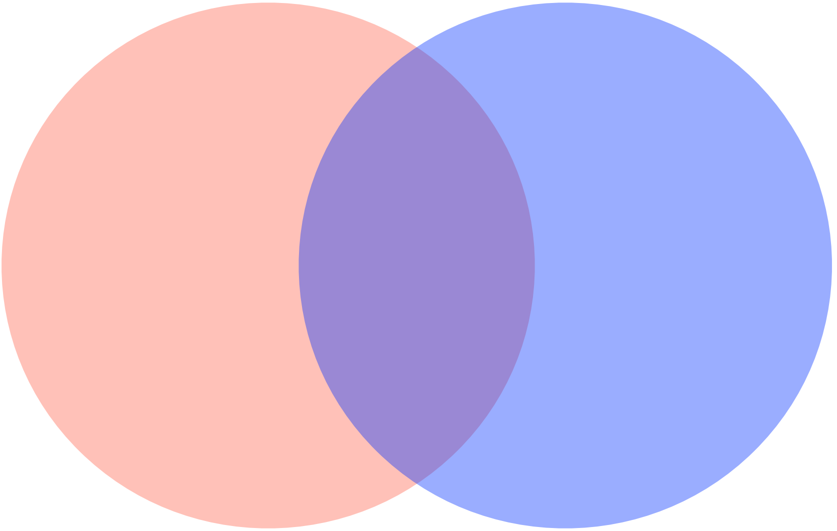
Set Similarity Review

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



Set Similarity Review

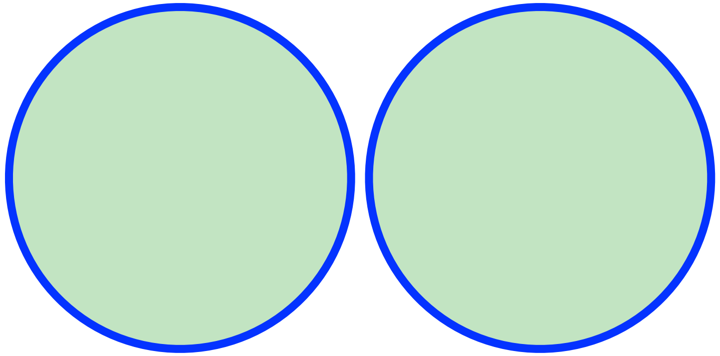
To measure **similarity** of A & B, we need both a measure of how similar the sets are but also the total size of both sets.



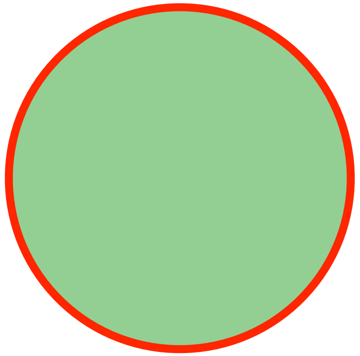
$$J = \frac{|A \cap B|}{|A \cup B|}$$

J is the ***Jaccard coefficient***

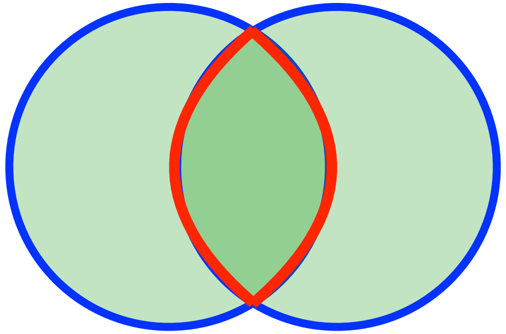
Set Similarity Review



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



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



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

Similarity Sketches

Imagine we have two datasets represented by their kth minimum values

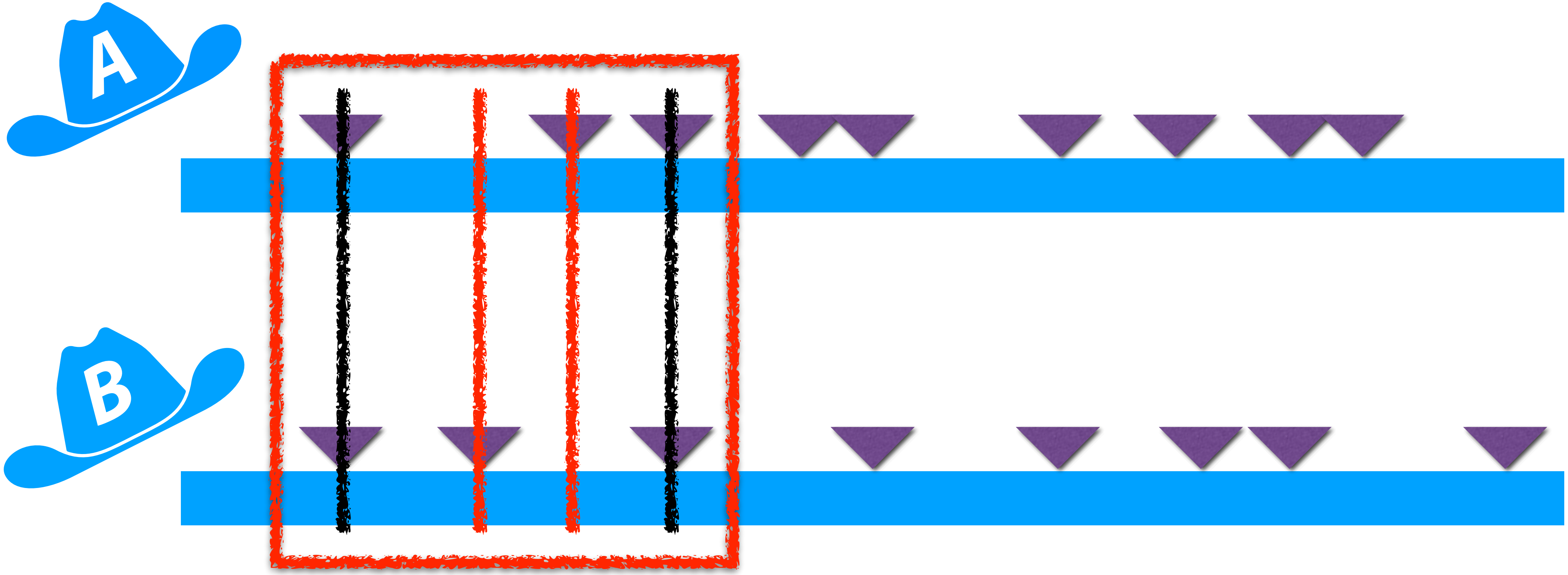


Image inspired by: Ondov B, Starrett G, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. **Mash Screen: high-throughput sequence containment estimation for genome discovery**. *Genome Biol* 20, 232 (2019)

Similarity Sketches

Claim: Under SUHA, comparing all k minima can estimate set similarity!

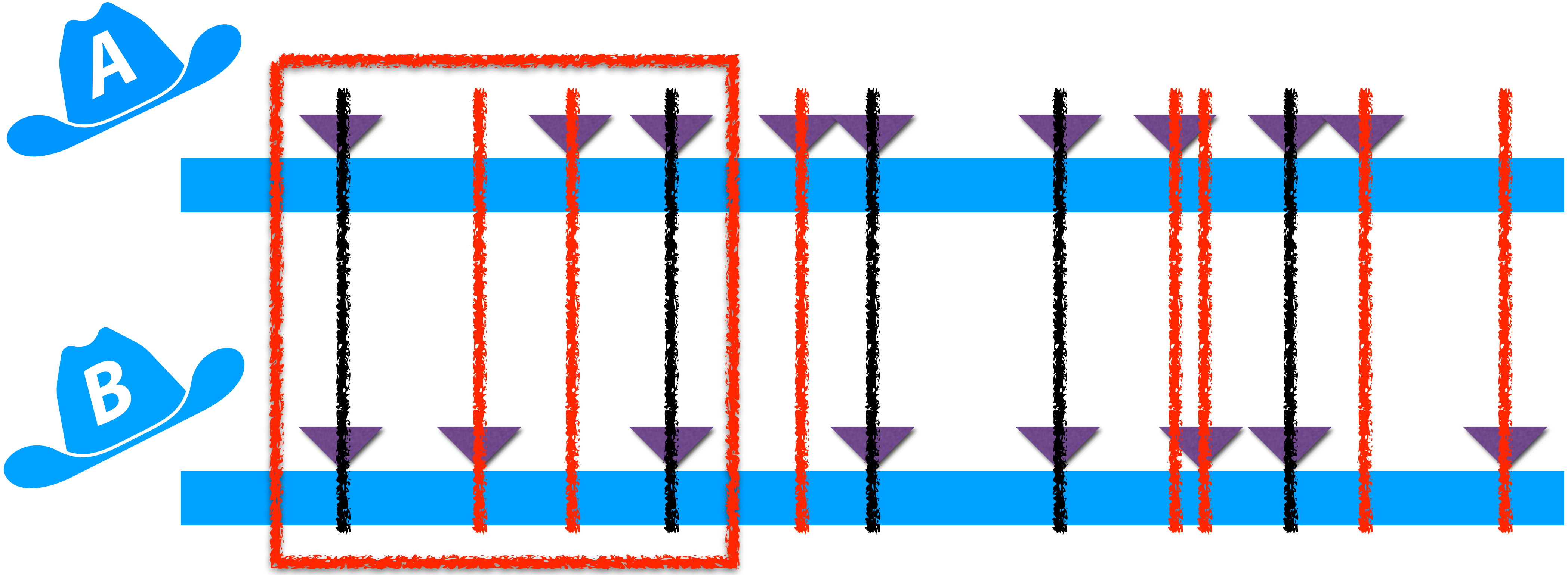


Image inspired by: Ondov B, Starrett G, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. **Mash Screen: high-throughput sequence containment estimation for genome discovery**. *Genome Biol* 20, 232 (2019)

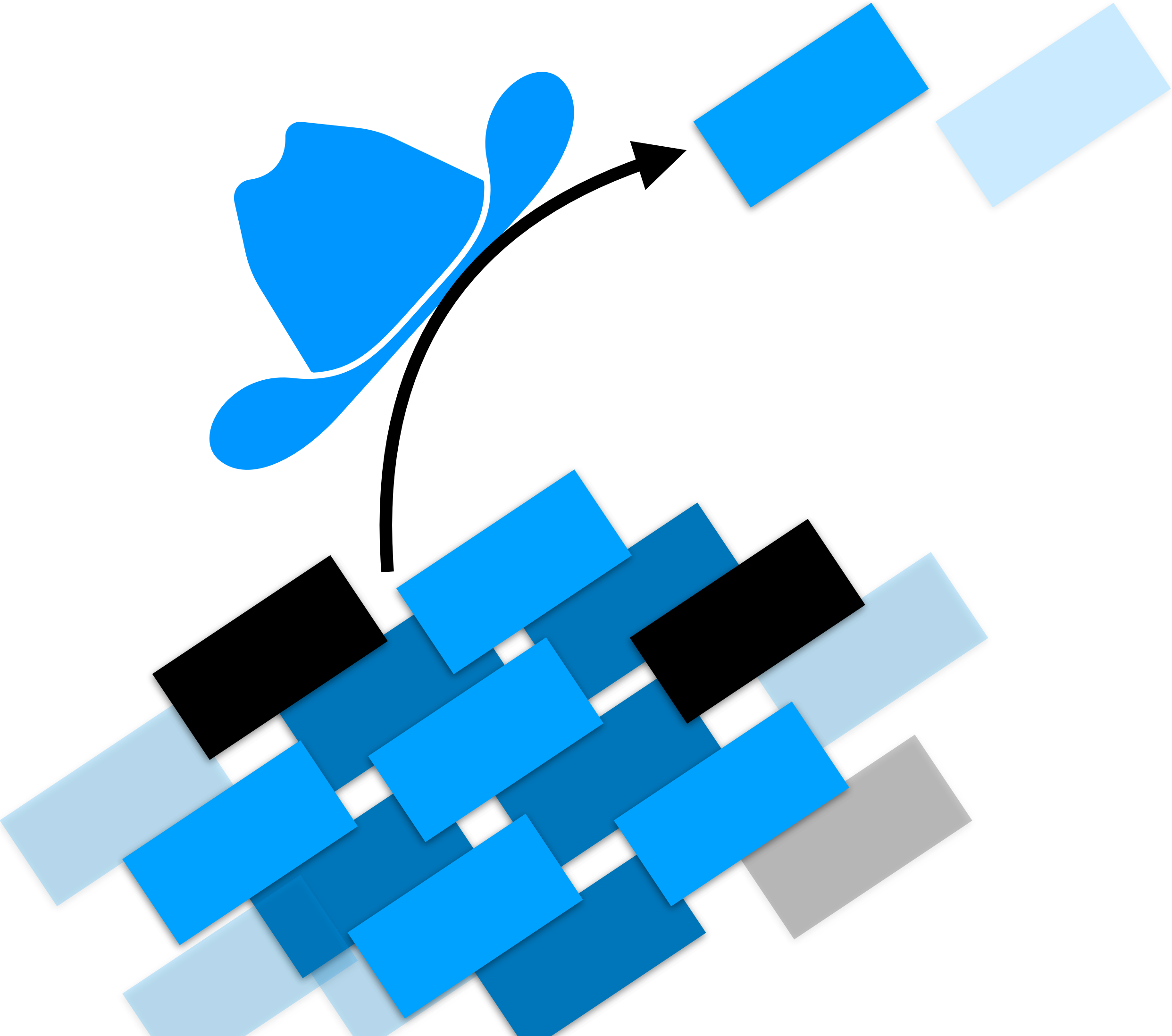
MinHash Construction

A MinHash sketch has three required inputs:

1.

2.

3.



MinHash Construction

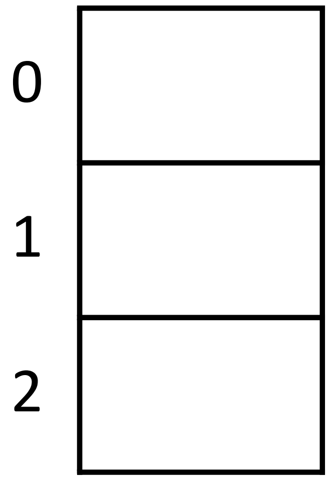
S = { 16, 8, 4, 13, 15}

h(x) = x % 7

k = 3

Algorithm is trivial:

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



MinHash Jaccard Estimation

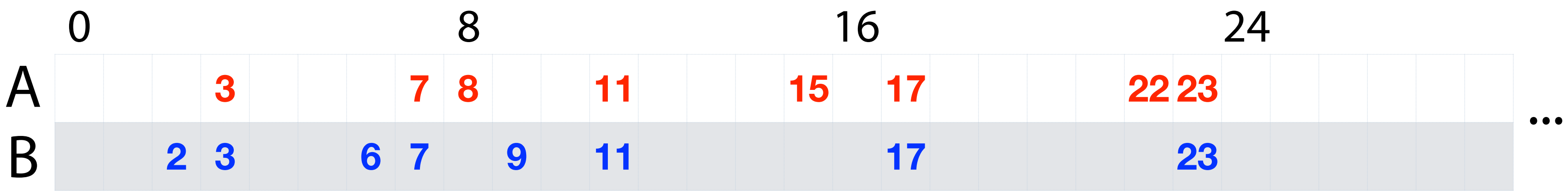
Given sets A and B sampled uniformly from [0, 100], store the bottom-8 **MinHash:**

Sketch A

3	15
7	17
8	22
11	23

Sketch B

2	9
3	11
6	17
7	23



MinHash Jaccard Estimation

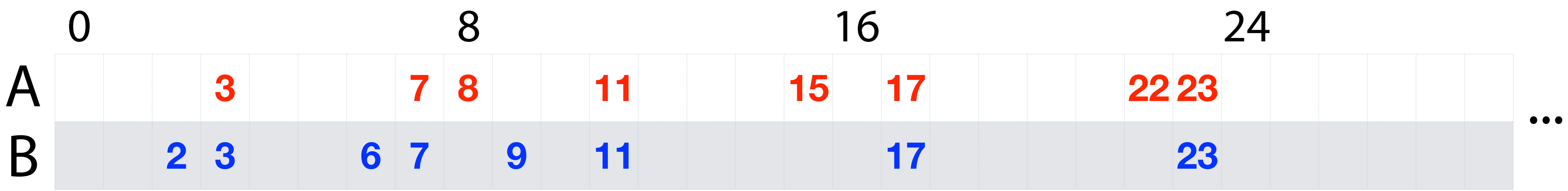
We want to estimate the Jaccard Coefficient: $\frac{|A \cup B|}{|A \cap B|}$

Sketch A

3	15
7	17
8	22
11	23

Sketch B

2	9
3	11
6	17
7	23



MinHash Jaccard Estimation

What do we know about $A \cup B$?

Sketch A

3	15
7	17
8	22
11	23

Sketch B

2	9
3	11
6	17
7	23

$A \cup B$

…

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

MinHash Jaccard Estimation

We dont *know* $A \cup B$, but we can make a sketch!

Sketch A

3	15
7	17
8	22
11	23

$\cup$

Sketch B

2	9
3	11
6	17
7	23

$=$

Sketch $A \cup B$

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

…

MinHash Jaccard Estimation

Estimate $|A \cup B|$ (the cardinality of the union) from sketch:

Sketch $A \cup B$

2	8
3	9
6	11
7	15

Our sets sampled from [0, 100].

MinHash Jaccard Estimation

Can we build a 8-Minhash of $A \cap B$?

Sketch A

3	15
7	17
8	22
11	23

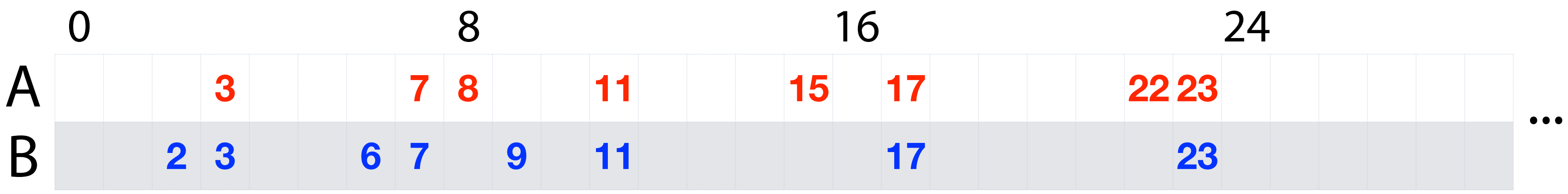
$\cap$

Sketch B

2	9
3	11
6	17
7	23

$=$

Sketch $A \cap B$



MinHash Jaccard Estimation

Not guaranteed to be able to get a full sketch of the intersection!

Sketch A

3	15
7	17
8	22
11	23

$\cap$

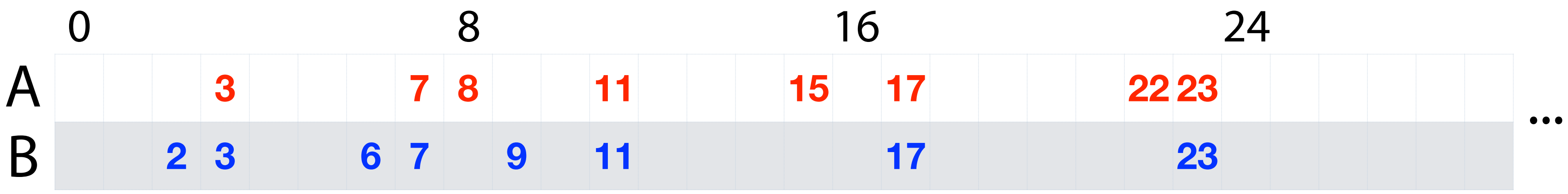
Sketch B

2	9
3	11
6	17
7	23

$=$

Sketch $A \cap B$

3	23
7	
11	
17	



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

…

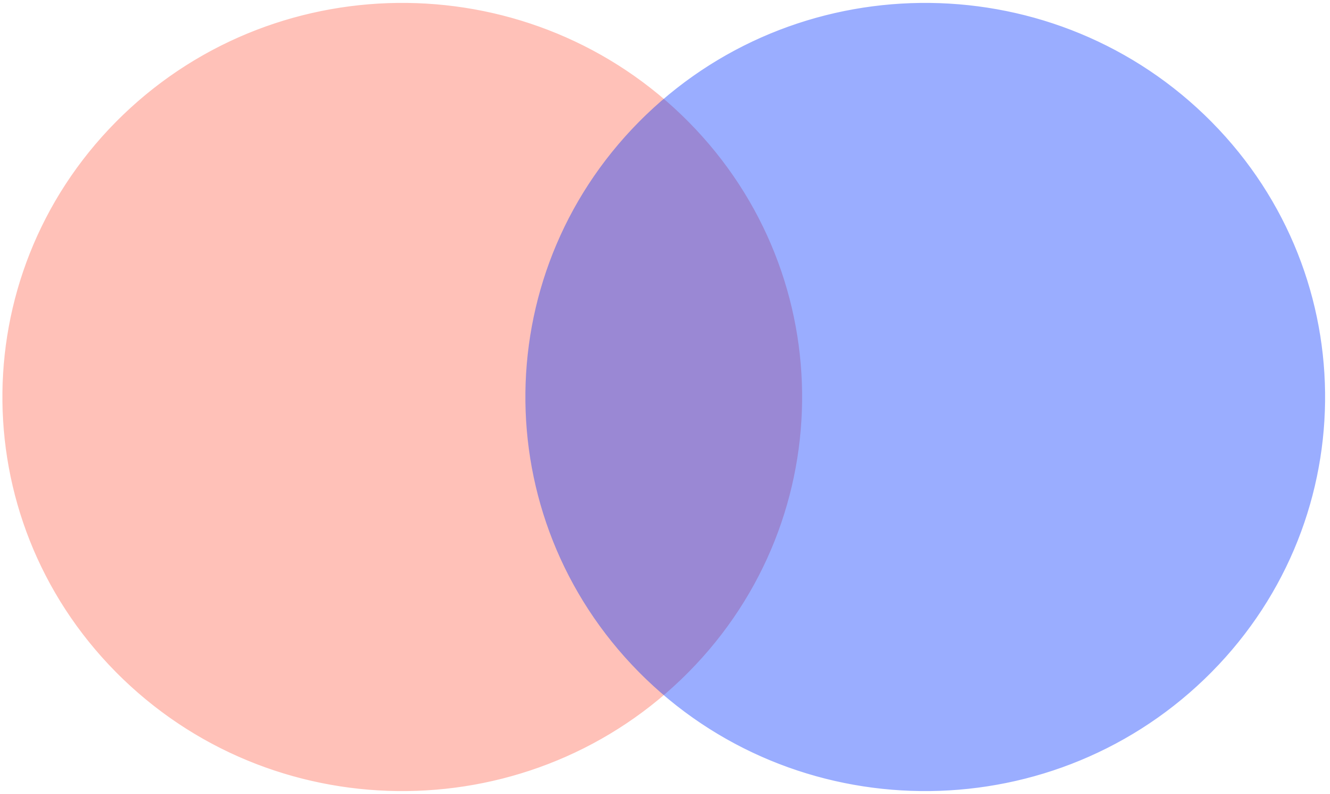
MinHash Jaccard Estimation

Using MinHash sketches, we can estimate $|A|$, $|B|$, and $|A \cup B|$

Is this enough to estimate the Jaccard?

Inclusion-Exclusion Principle

$$|A \cap B| =$$



MinHash Indirect Jaccard Estimation

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

$k = 8$ MinHash sketches

Our sets sampled from [0, 100]

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

$$= \frac{(800/23-1)+(800/23-1)-\left(800/15-1\right)}{800/15-1}$$

$$= \frac{34.782+34.782-53.333-1}{53.333-1} \approx 0.29$$

MinHash Direct Jaccard Estimate

We can also estimate cardinality directly using our sketches!

Sketch A

3	15
7	17
8	22
11	23

Sketch B

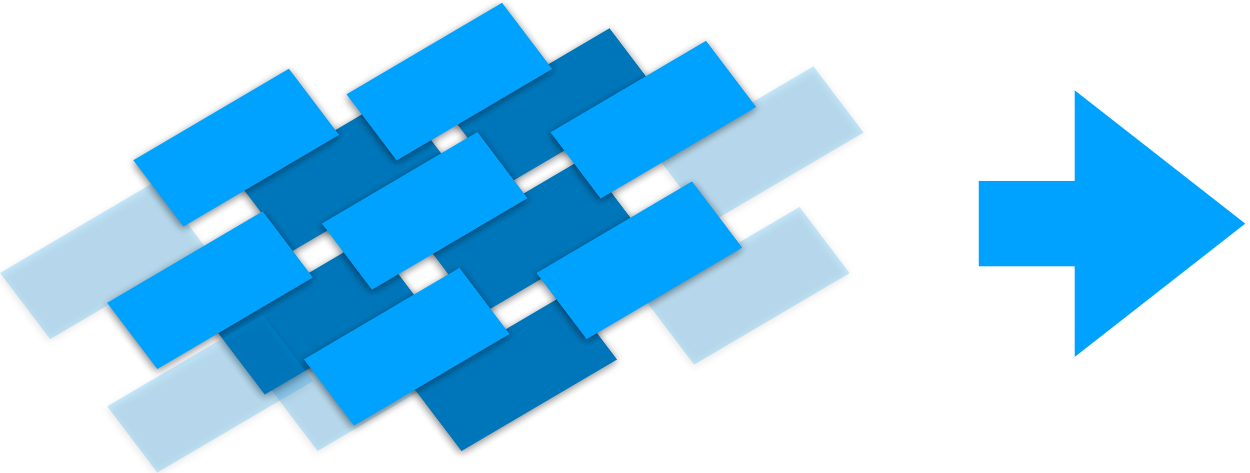
2	9
3	11
6	17
7	23

Intersection

Union

MinHash Sketch

We can convert any hashable dataset into a **MinHash sketch**



{1 , 3 , 5}

We lose our original dataset, but we can still estimate two things:

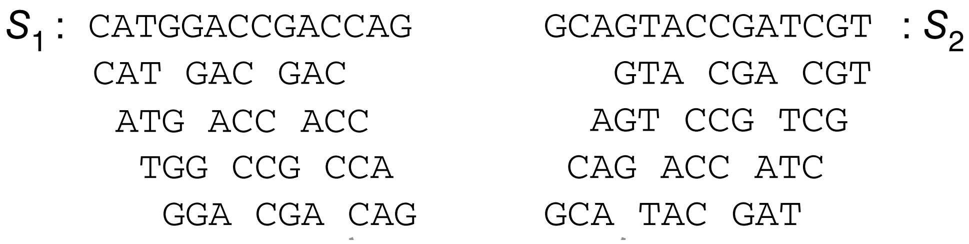
1.

2.

Alternative MinHash Sketch Approaches

Rather than use one single hashes and take bottom-k, we can also use k hashes — **if you have access to that many independent hashes**!

1) Sequence decomposed into **kmers**



S_1 : CATGGACCGACCAG

CAT GAC GAC

ATG ACC ACC

TGG CCG CCA

GGA CGA CAG

GCAGTACCGATCGT : S_2

GTA CGA CGT

AGT CCG TCG

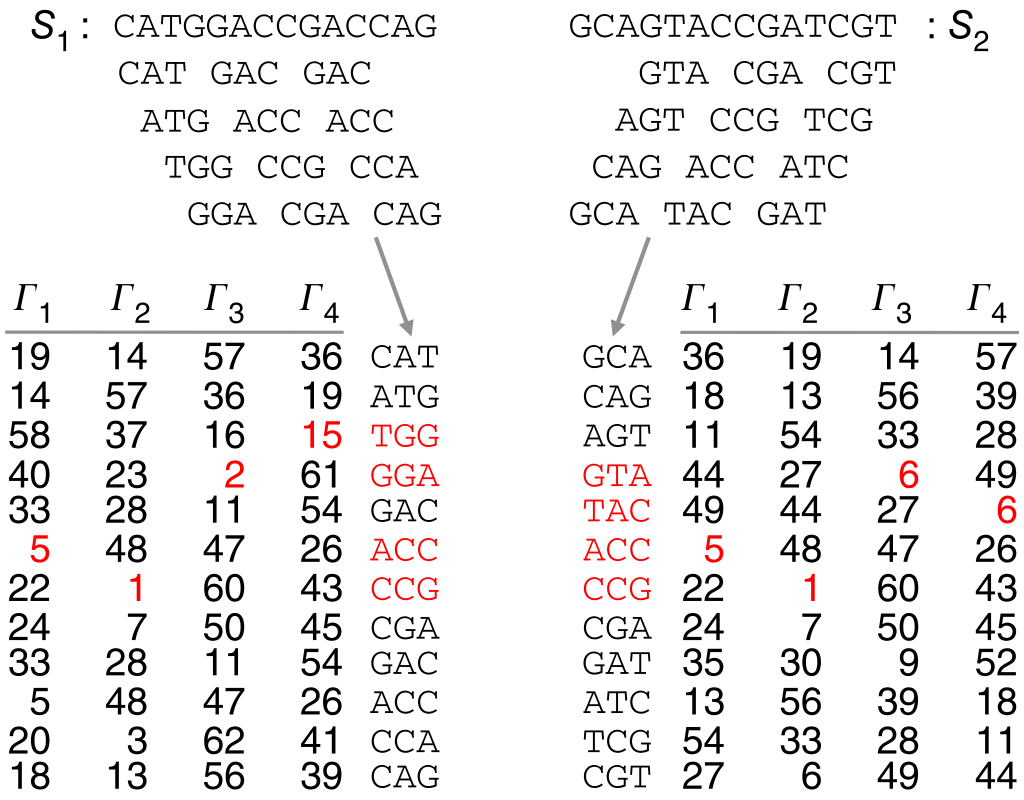
CAG ACC ATC

GCA TAC GAT

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

1) Sequence decomposed into **kmers**

2) Multiple hash functions (**Γ**) map kmers to values.



S_1: CATGGACCGACCAG
CAT GAC GAC
ATG ACC ACC
TGG CCG CCA
GGA CGA CAG

GCAGTACCGATCGT : S_2
GTA CGA CGT
AGT CCG TCG
CAG ACC ATC
GCA TAC GAT

Γ_1	Γ_2	Γ_3	Γ_4	
19	14	57	36	CAT
14	57	36	19	ATG
58	37	16	15	TGG
40	23	2	61	GGA
33	28	11	54	GAC
5	48	47	26	ACC
22	1	60	43	CCG
24	7	50	45	CGA
33	28	11	54	GAC
5	48	47	26	ACC
20	3	62	41	CCA
18	13	56	39	CAG

	Γ_1	Γ_2	Γ_3	Γ_4
GCA	36	19	14	57
CAG	18	13	56	39
AGT	11	54	33	28
GTA	44	27	6	49
TAC	49	44	27	6
ACC	5	48	47	26
CCG	22	1	60	43
CGA	24	7	50	45
GAT	35	30	9	52
ATC	13	56	39	18
TCG	54	33	28	11
CGT	27	6	49	44

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

1) Sequence decomposed into **kmers**

S_1 : CATGGACCGACCAG

CAT GAC GAC
ATG ACC ACC
TGG CCG CCA
GGA CGA CAG

GCAGTACCGATCGT : S_2

GTA CGA CGT
AGT CCG TCG
CAG ACC ATC
GCA TAC GAT

2) Multiple hash functions (**Γ**) map kmers to values.

Γ_1	Γ_2	Γ_3	Γ_4	
19	14	57	36	CAT
14	57	36	19	ATG
58	37	16	15	TGG
40	23	2	61	GGA
33	28	11	54	GAC
5	48	47	26	ACC
22	1	60	43	CCG
24	7	50	45	CGA
33	28	11	54	GAC
5	48	47	26	ACC
20	3	62	41	CCA
18	13	56	39	CAG

	Γ_1	Γ_2	Γ_3	Γ_4
GCA	36	19	14	57
CAG	18	13	56	39
AGT	11	54	33	28
GTA	44	27	6	49
TAC	49	44	27	6
ACC	5	48	47	26
CCG	22	1	60	43
CGA	24	7	50	45
GAT	35	30	9	52
ATC	13	56	39	18
TCG	54	33	28	11
CGT	27	6	49	44

3) The smallest values for each hash function is chosen

[5, 1, 2, 15]
Sketch (S_1)

[5, 1, 6, 6]
Sketch (S_2)

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

1) Sequence decomposed into **kmers**

2) Multiple hash functions (**Γ**) map kmers to values.

3) The smallest values for each hash function is chosen

4) The Jaccard similarity can be estimated by the overlap in the **Min**imum **Hash**es **(MinHash)**

S_1: CATGGACCGACCAG

CAT GAC GAC
ATG ACC ACC
TGG CCG CCA
GGA CGA CAG

GCAGTACCGATCGT : S_2

GTA CGA CGT
AGT CCG TCG
CAG ACC ATC
GCA TAC GAT

Γ_1	Γ_2	Γ_3	Γ_4	
19	14	57	36	CAT
14	57	36	19	ATG
58	37	16	15	TGG
40	23	2	61	GGA
33	28	11	54	GAC
5	48	47	26	ACC
22	1	60	43	CCG
24	7	50	45	CGA
33	28	11	54	GAC
5	48	47	26	ACC
20	3	62	41	CCA
18	13	56	39	CAG

	Γ_1	Γ_2	Γ_3	Γ_4
GCA	36	19	14	57
CAG	18	13	56	39
AGT	11	54	33	28
GTA	44	27	6	49
TAC	49	44	27	6
ACC	5	48	47	26
CCG	22	1	60	43
CGA	24	7	50	45
GAT	35	30	9	52
ATC	13	56	39	18
TCG	54	33	28	11
CGT	27	6	49	44

[5, 1, 2, 15]
Sketch (S_1)

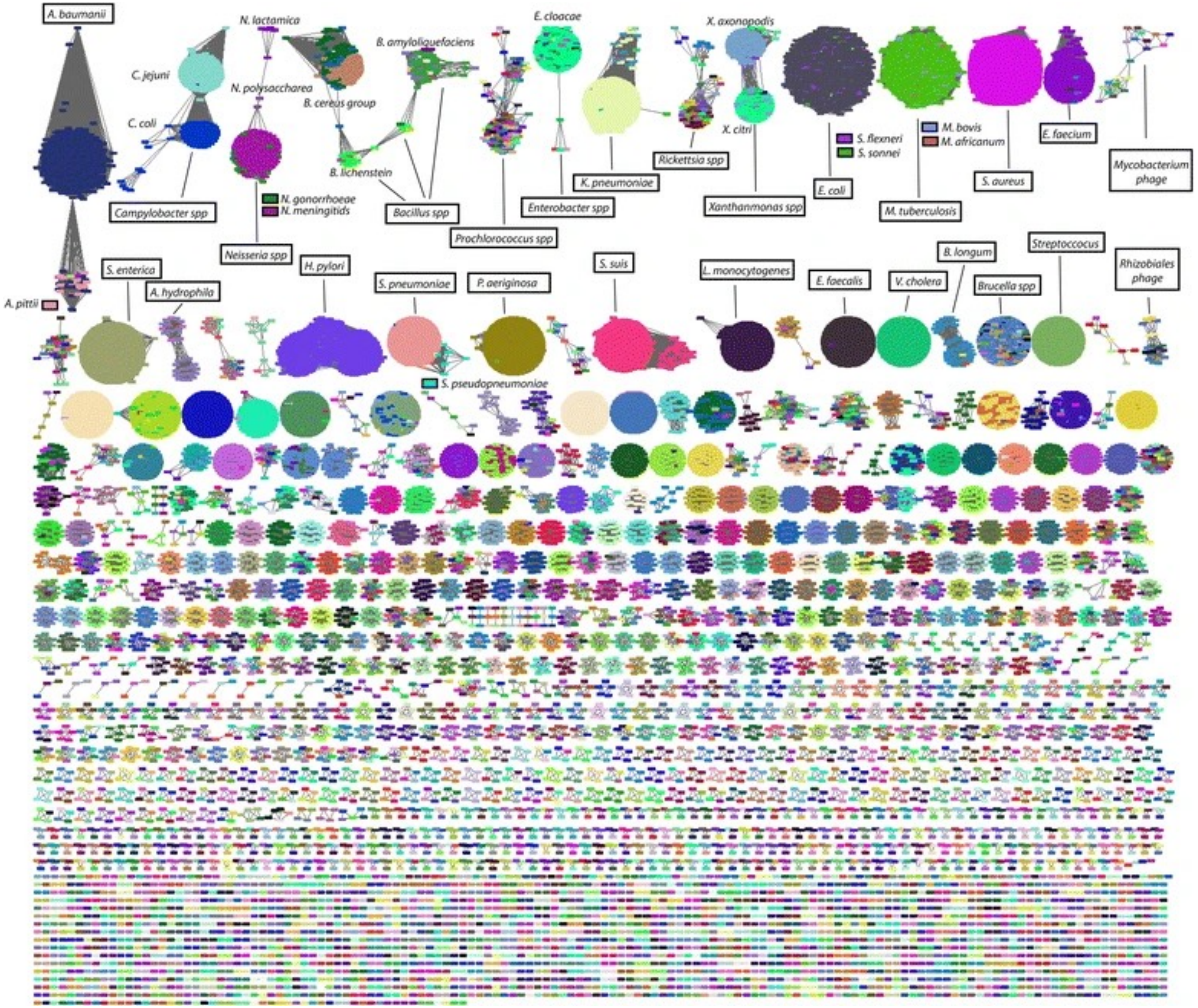
[5, 1, 6, 6]
Sketch (S_2)

J (S_1, S_2) ≈ 2/4 = 0.5

S_1: CATGGACCGACCAG
| ||||| |
S_2: GCAGTACCGATCGT

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

MinHash in practice



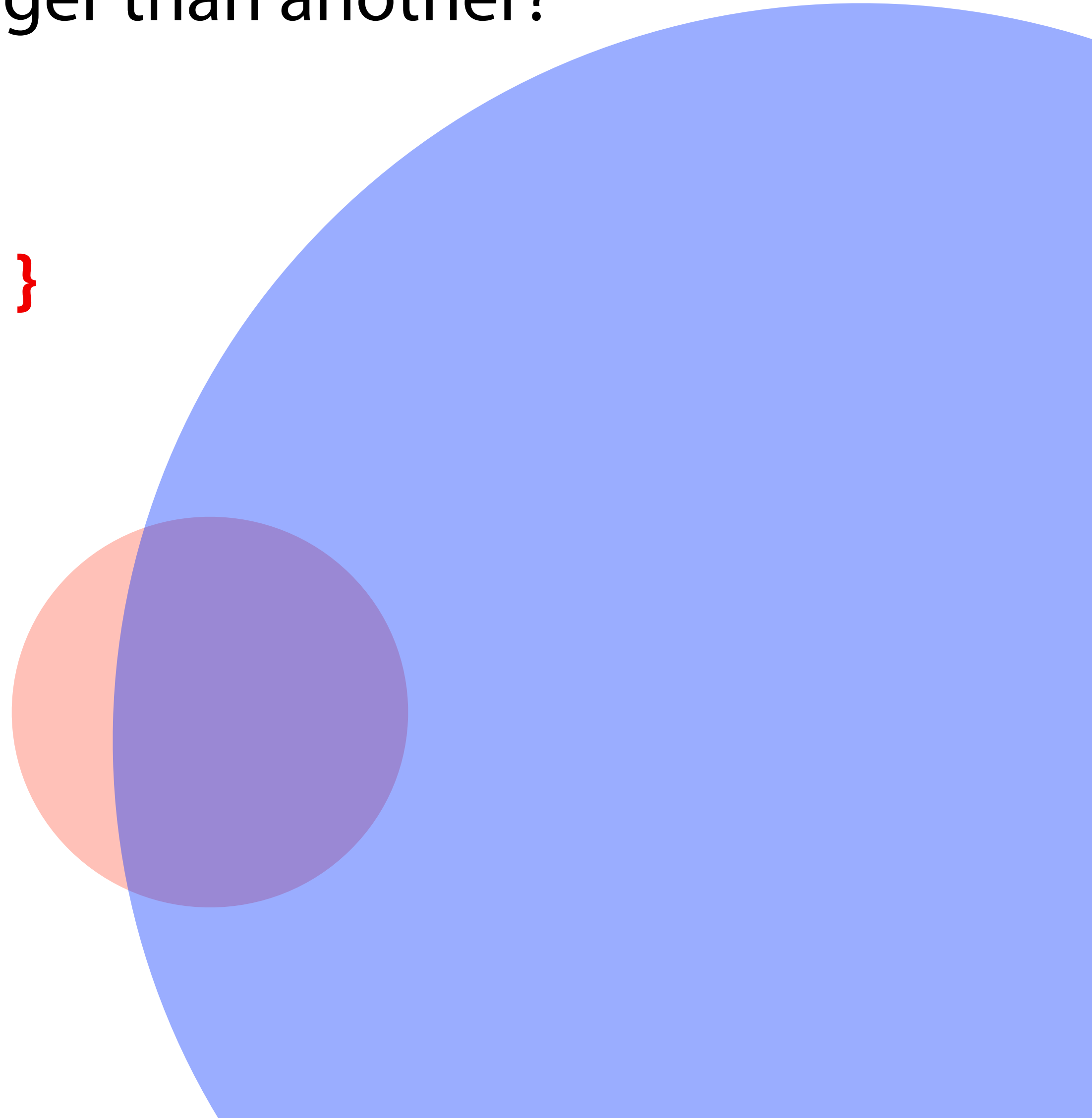
Mash: fast genome and metagenome distance estimation using MinHash
Ondov et al (2016) *Genome Biology*

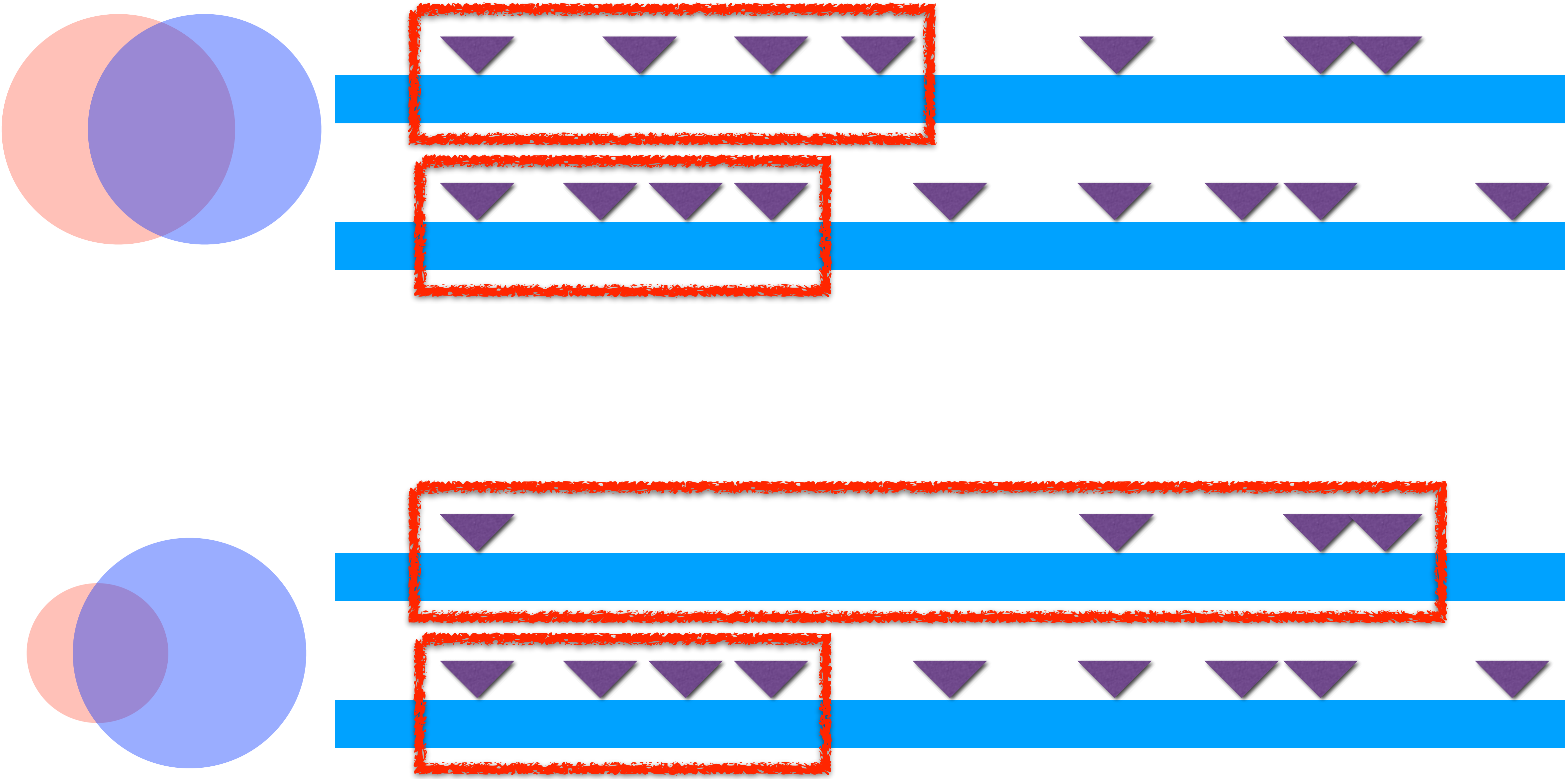
Alternative MinHash Sketch Approaches

What if I have a dataset which is **much** larger than another?

S_1 = { 1, 3, 40, 59, 82, 101 }

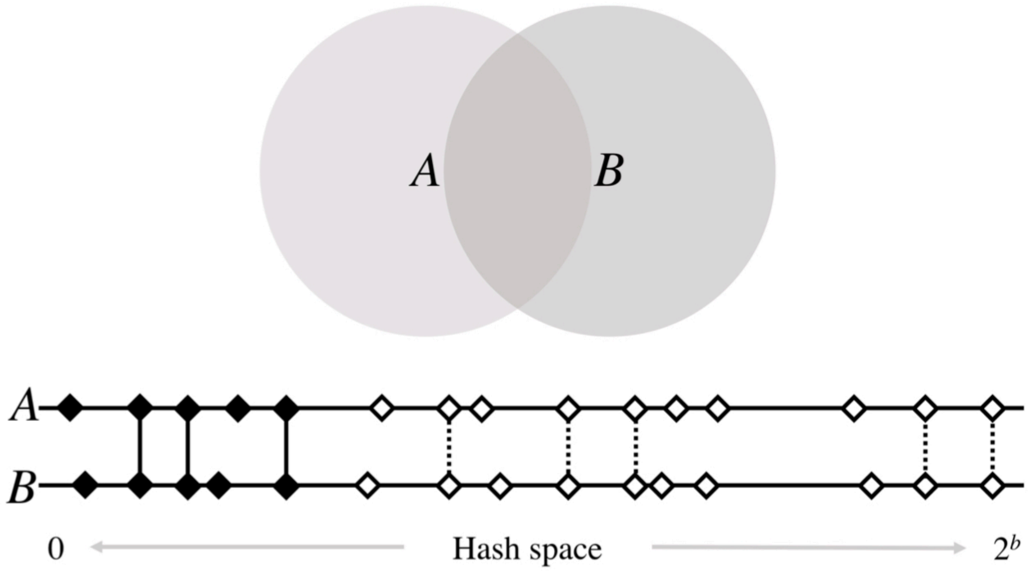
S_2 = { 1, 2, 3, 4, 5, 6, 7, … 59, 82, 101, … }

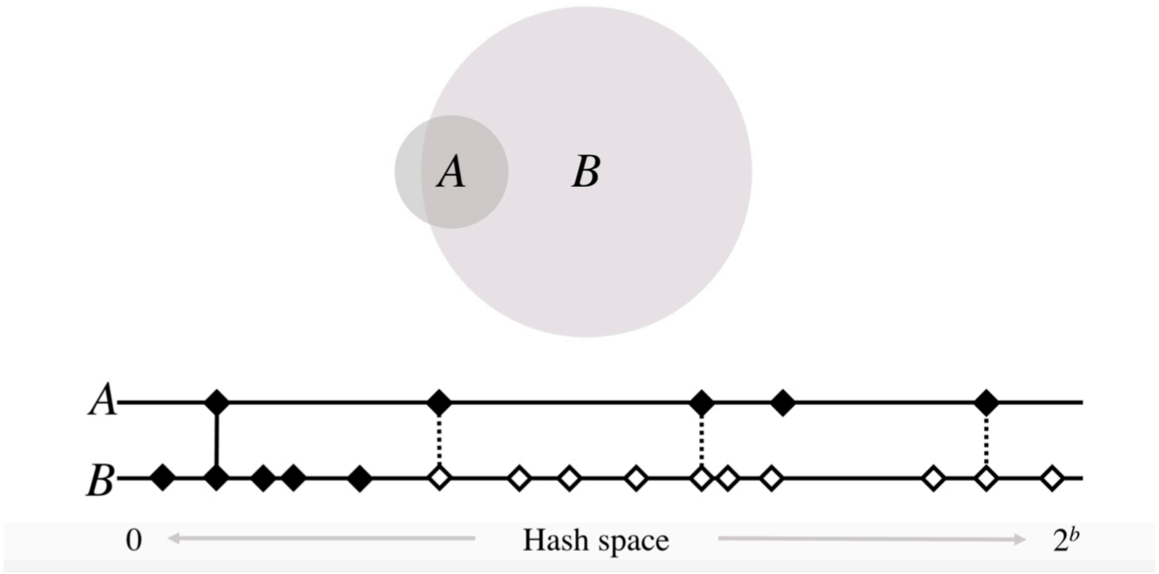




Alternative MinHash sketches

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

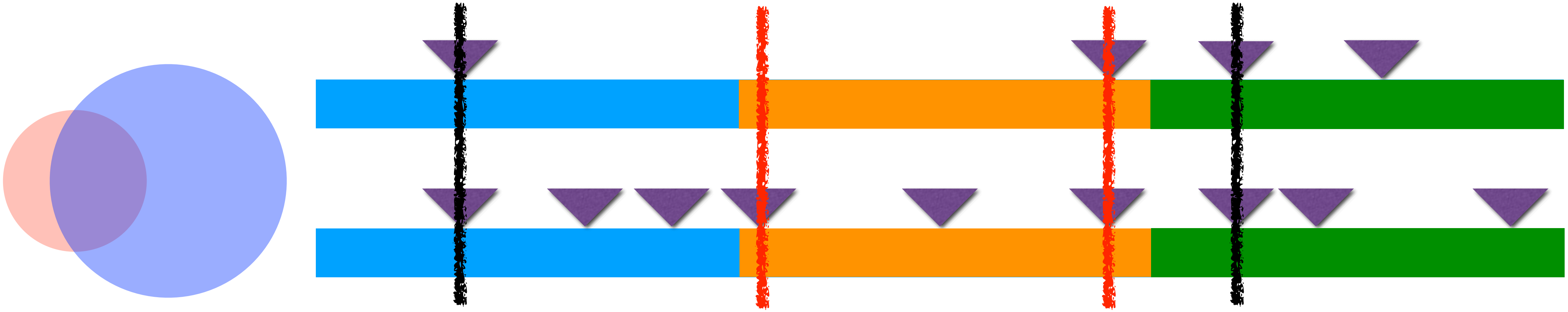




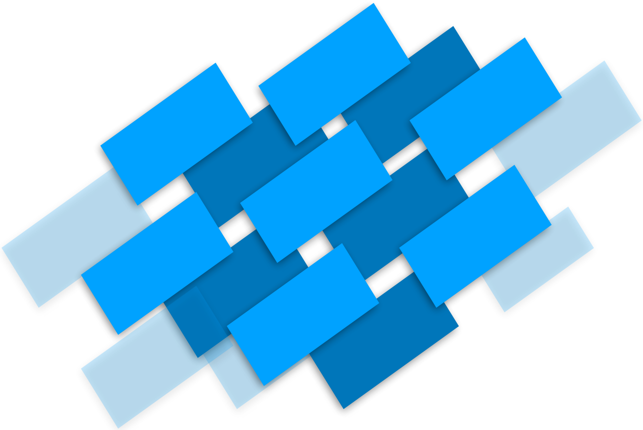
Ondov, Brian D., Gabriel J. Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B. Buck, and Adam M. Phillippy. **Mash Screen: High-throughput sequence containment estimation for genome discovery**. *Genome biology* 20.1 (2019): 1-13.

Alternative MinHash Sketch Approaches

If there is a large cardinality difference, **use k-partitions!**



K-Partition Minhash



Hash

1010110101
0001111010
1101101011
1011010110
0101100000
0010001101

Partition

00
01111010
10001101

01
01100000

10
10110101
11010110

11
01101011

Probabilistic Data Structures

Probabilistic data structures trade accuracy for efficiency

Most can maintain surprisingly good accuracy

“Cheat” Big O limitations on conventional data analysis