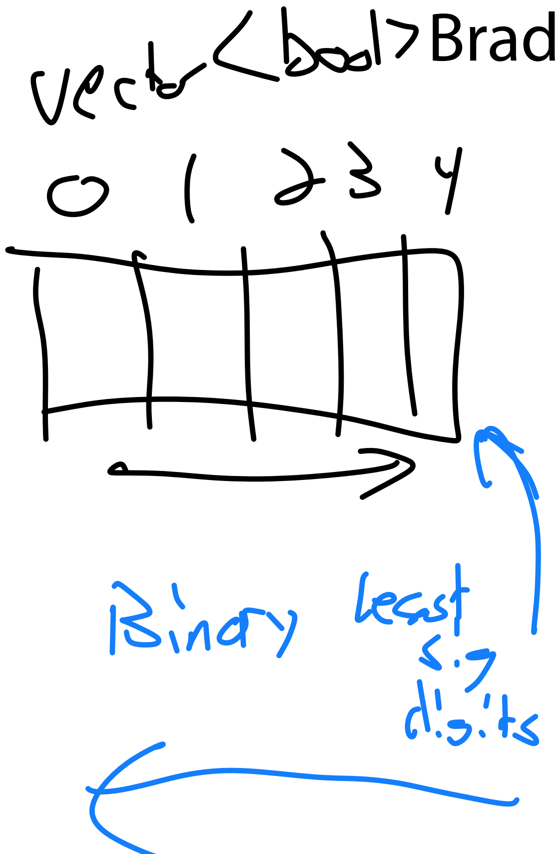
Data Structures and Algorithms
MinHash Sketch

CS 225
Brad Solomon

November 8, 2023



UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN
Department of Computer Science

Extra Credit Project — Next Steps

~20% acceptance rate on extra credit projects

↳ # of last minute submissions very high

If you were not approved, its just means you will not receive extra credit

Mentors will be notifying you sometime this week

Be sure to submit a weekly development log! Schedule a check-in meeting!

Learning Objectives

Review the concept of cardinality and cardinality estimation

# of unique items

↳ hard to get exact count

- Memory issues
- Data Scale

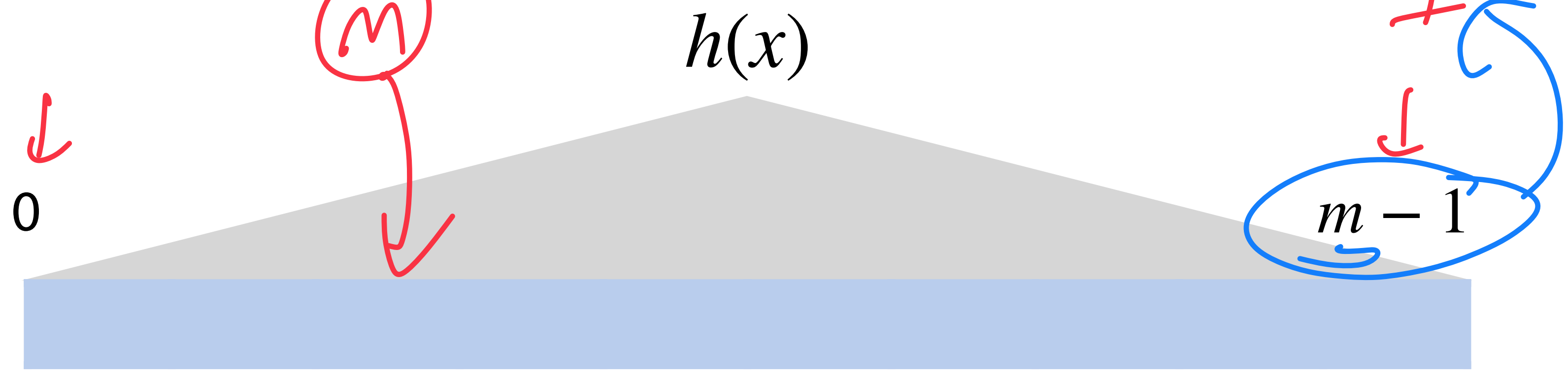
Improve our cardinality estimation approach

Demonstrate the relationship between cardinality and similarity

Introduce the MinHash Sketch for set similarity detection

Cardinality Estimation

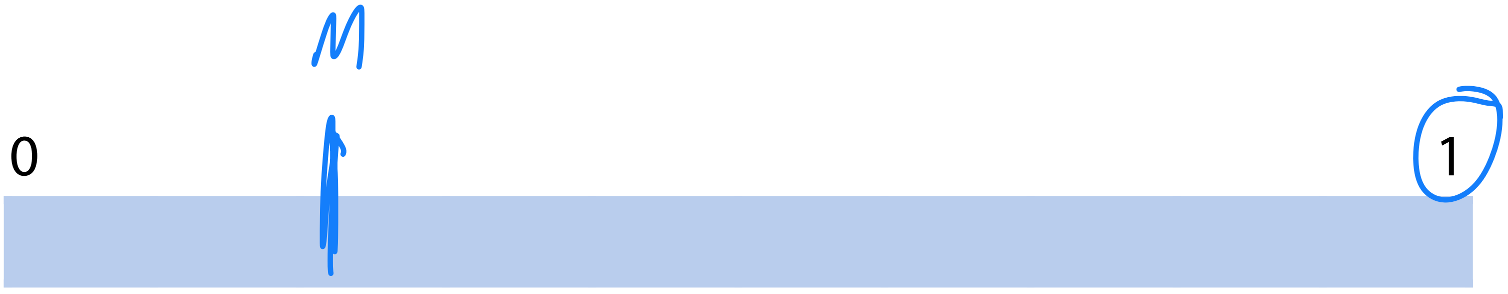
Given a SUHA hash h over a range m, we can estimate cardinality:



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



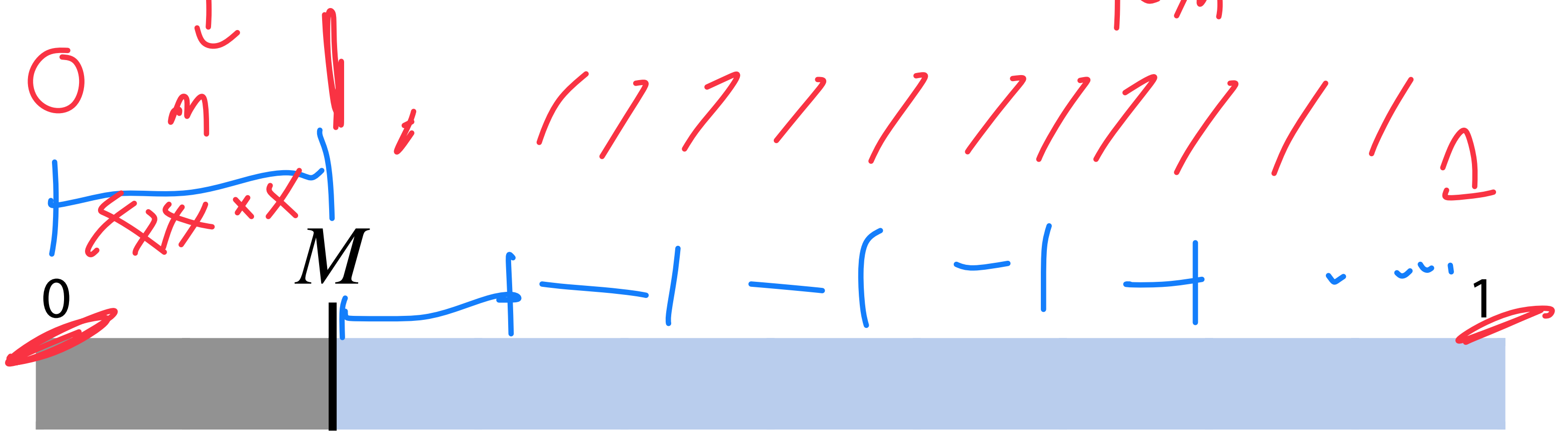
Cardinality Sketch

$\mathbf{E}[M]$ defines the range from 0 to the min value $\left(M = \min_{1 \le i \le N} X_i\right)$

Consider an $N + 1$ draw: X_1 | X_2 | X_3 ... X_N | X_{N+1}

Not real — hypothetical

$\text{Prob}(X_{N+1} \text{ is min}) = M$



Cardinality Sketch

uniform indep & every X_i is equally likely to be min

Consider an $N+1$ draw: X_1 | X_2 | X_3 ... X_N | X_{N+1}

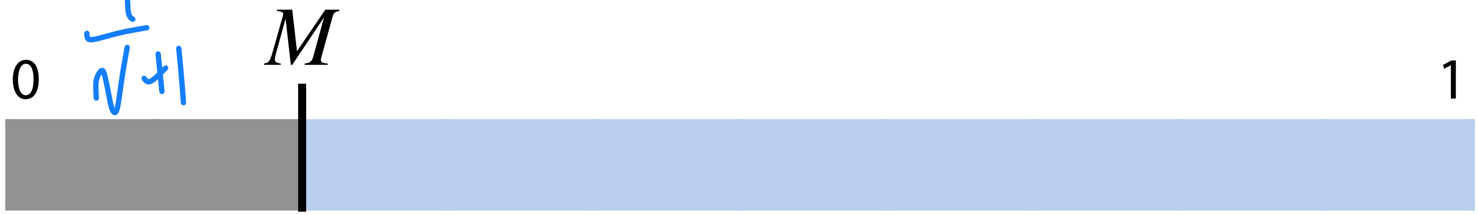
$\frac{1}{N+1}$

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

Define an **indicator:**

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

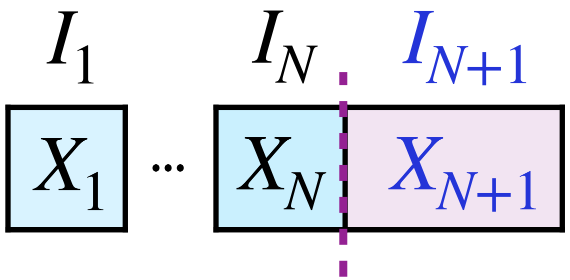
$$\mathbf{E}[I_i] = \sum_v \text{Prob} \cdot v = \Pr(X_i < M) \cdot 1 + 0 \cdot \sum \Pr(X_i \text{ not} < M)$$



Cardinality Sketch

Hypothetical Draw

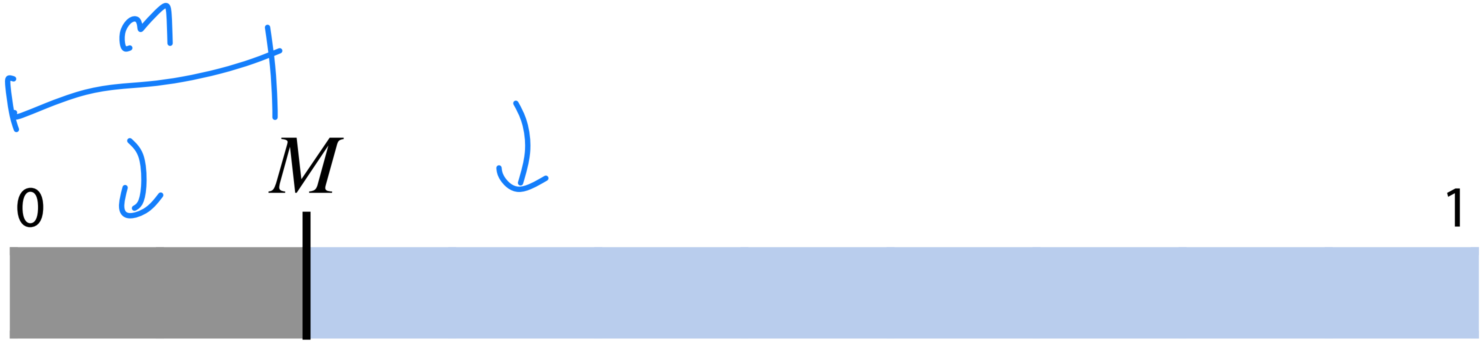
Claim: $\mathbf{E}[M] = \mathbf{E}[I_{N+1}]$ $= \frac{1}{N+1}$



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

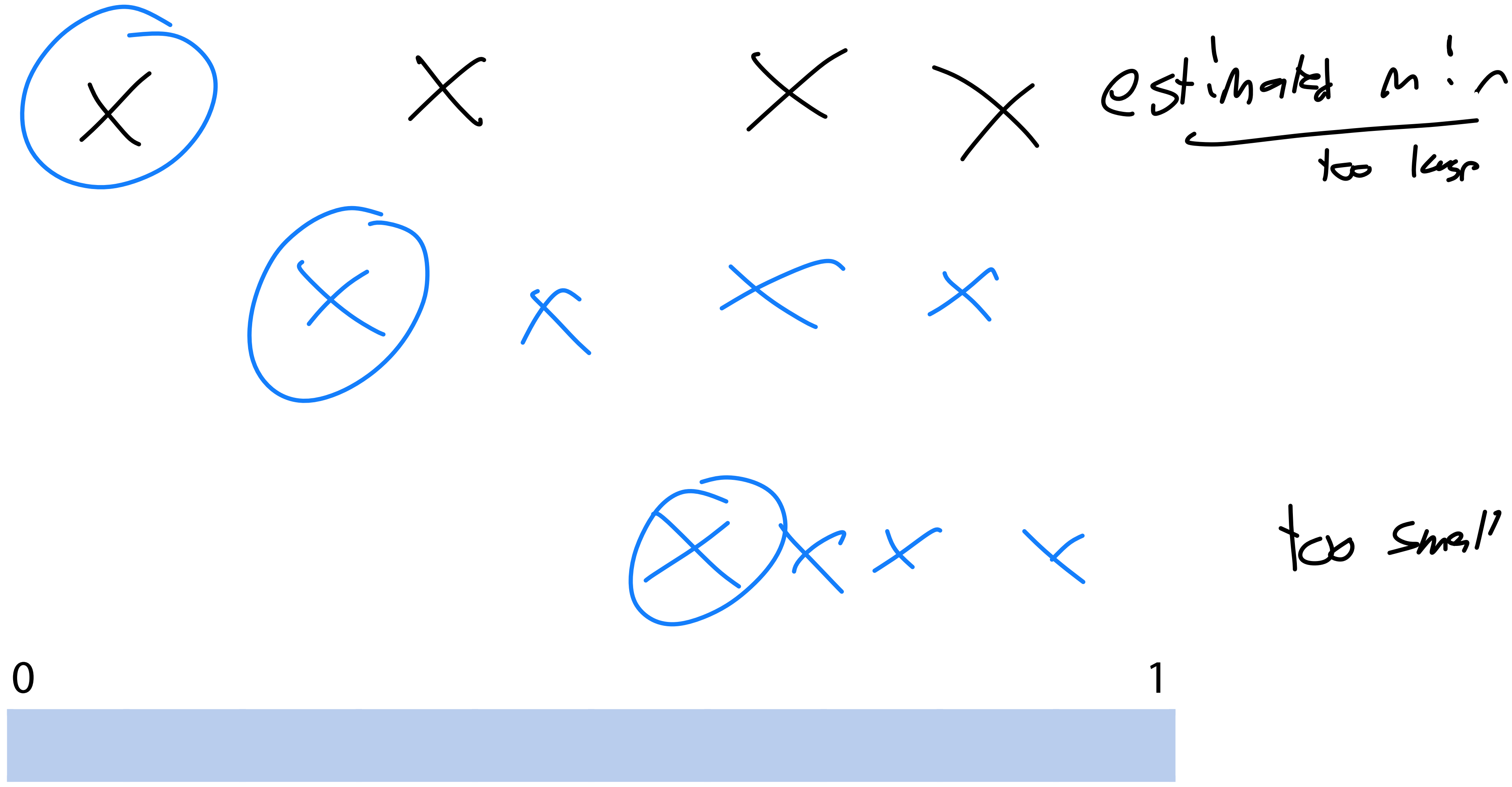
By definition, $\mathbf{E}[I_{N+1}] = \Pr(X_{N+1} < M) = \dfrac{1}{N+1}$

← cardinality (# unique)



Cardinality Sketch

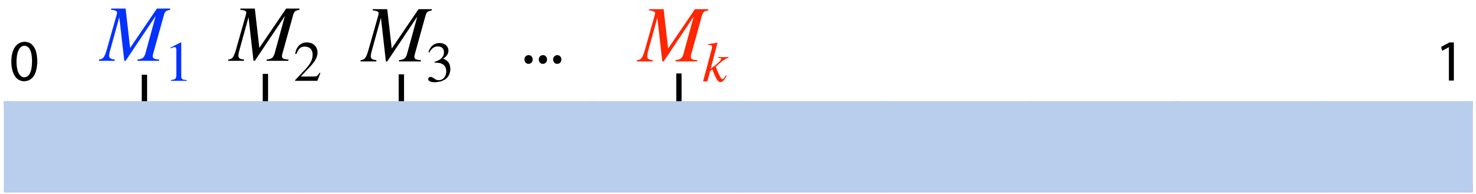
The minimum hash is a valid sketch of a dataset but can we do better?



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!

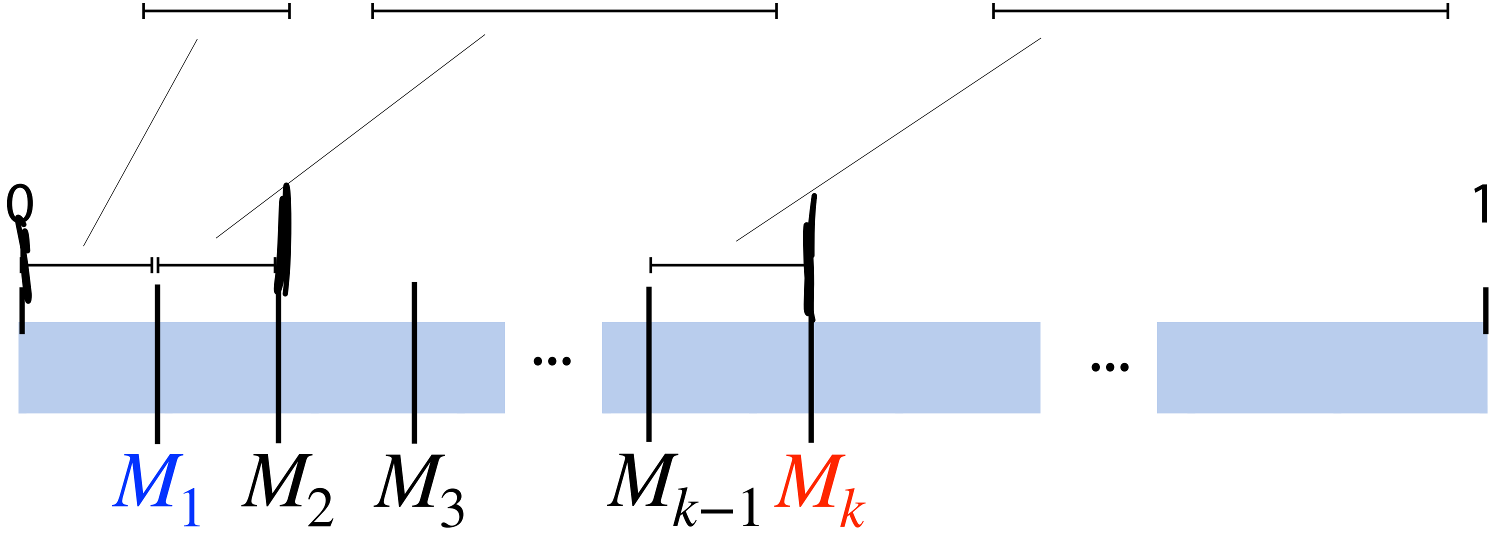
Claim: $\dfrac{\mathbf{E}[M_k]}{k} = \dfrac{1}{N+1}$

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

Cardinality Sketch

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

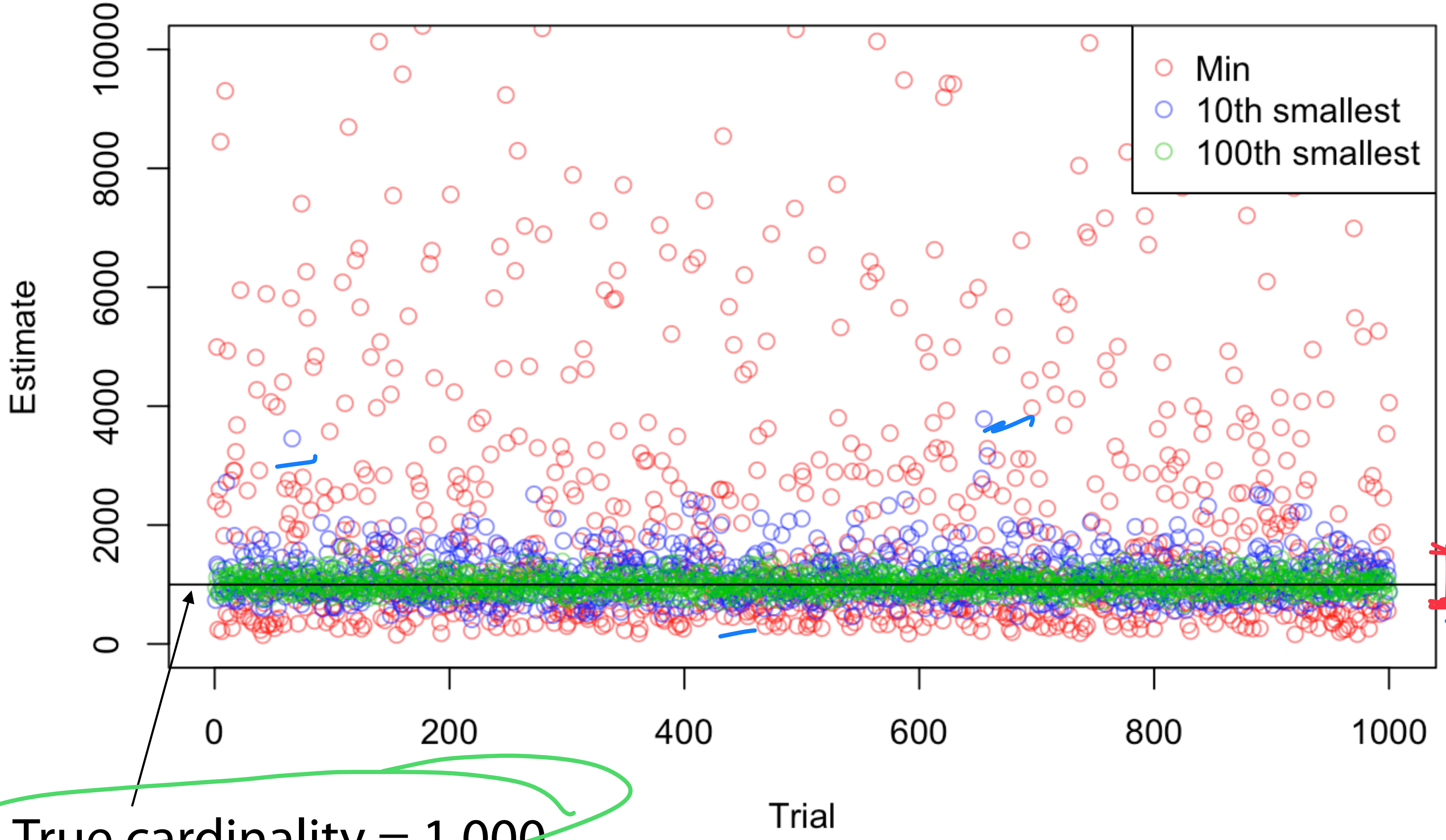
$$= \Big[\mathbf{E}[M_1] + (\mathbf{E}[M_2] - \mathbf{E}[M_1]) + \ldots + (\mathbf{E}[M_k] - \mathbf{E}[M_{k-1}])\Big] \cdot \frac{1}{k}$$



k^{th} minimum value (KMV)

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

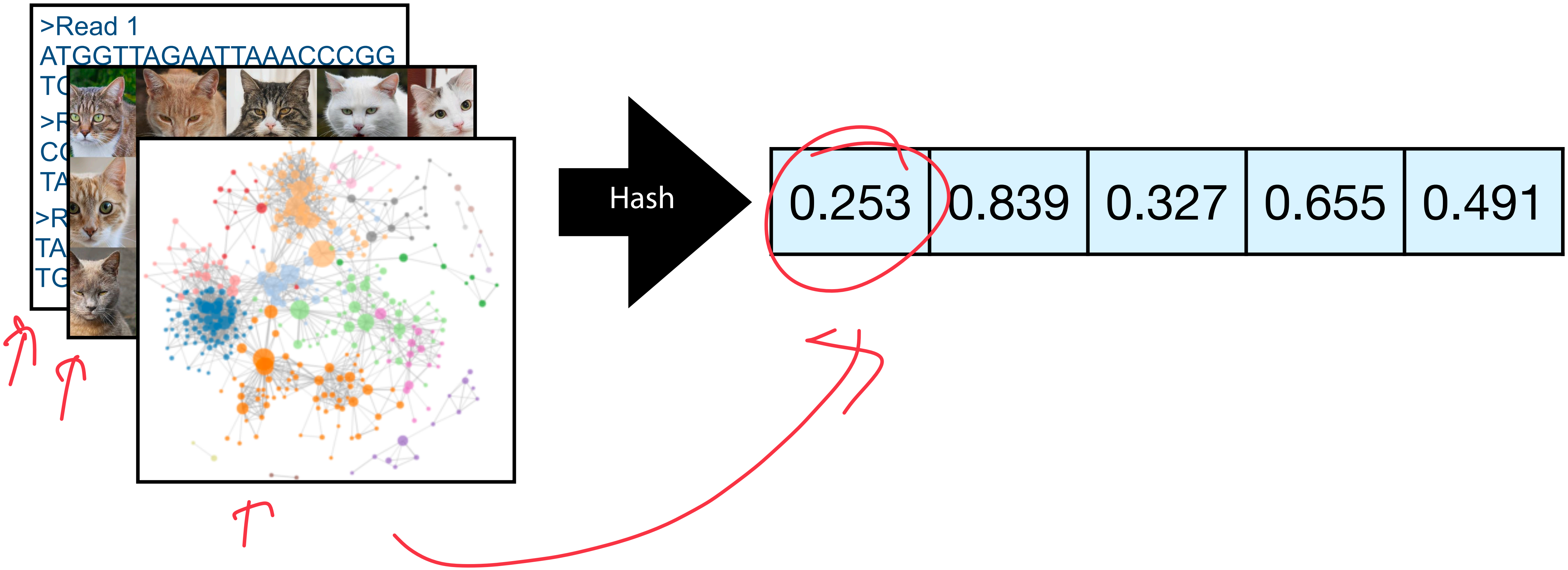
Cardinality Sketch



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**.



Applied Cardinalities

Sketch

Cardinalities → Set similarities → Real-world Meaning

$$|A|$$

$$|B|$$

$$|A \cup B|$$

$$|A \cap B|$$

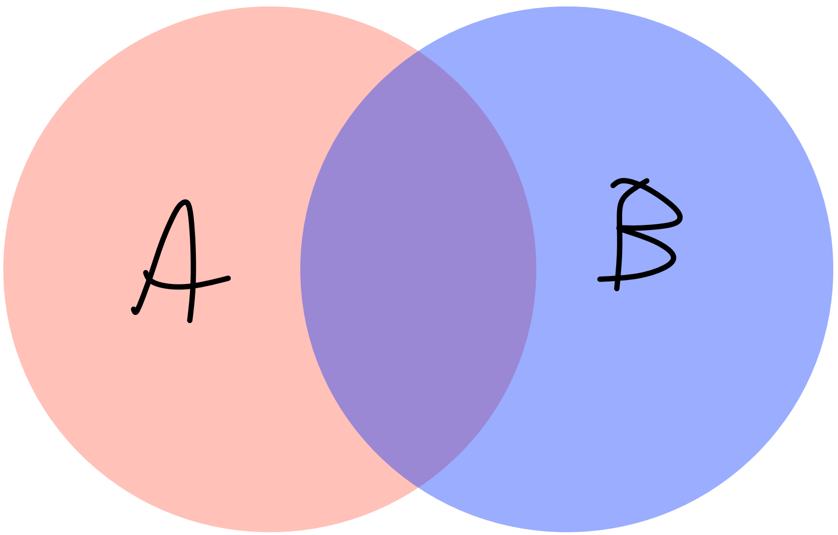
$$O = \frac{|A \cap B|}{min(|A|, |B|)}$$

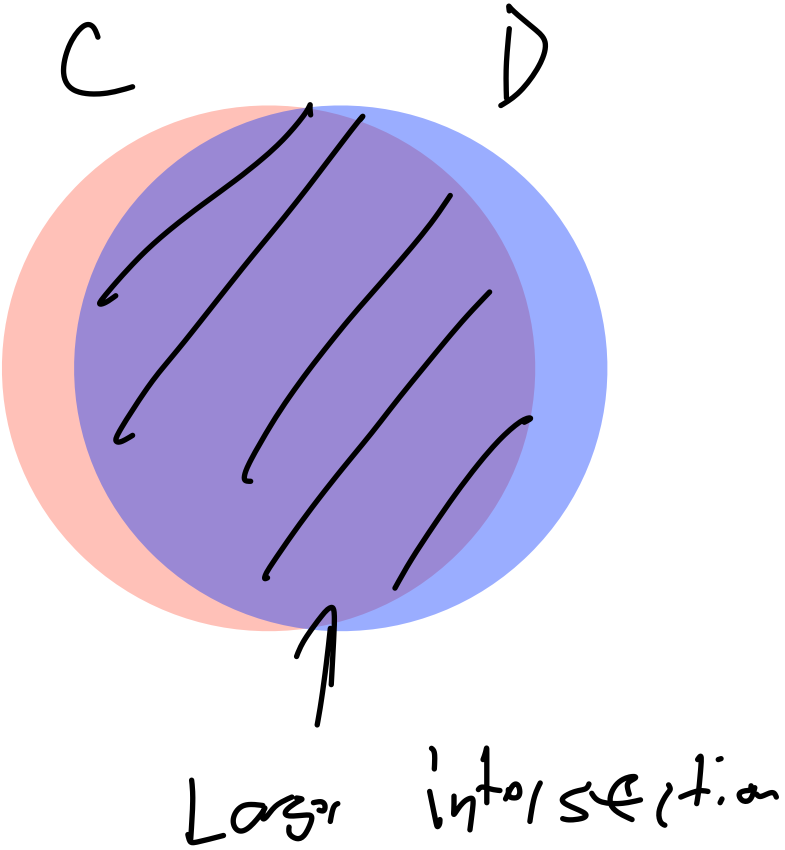
$$J = \frac{|A \cap B|}{|A \cup B|}$$

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

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

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

Set Similarity Review

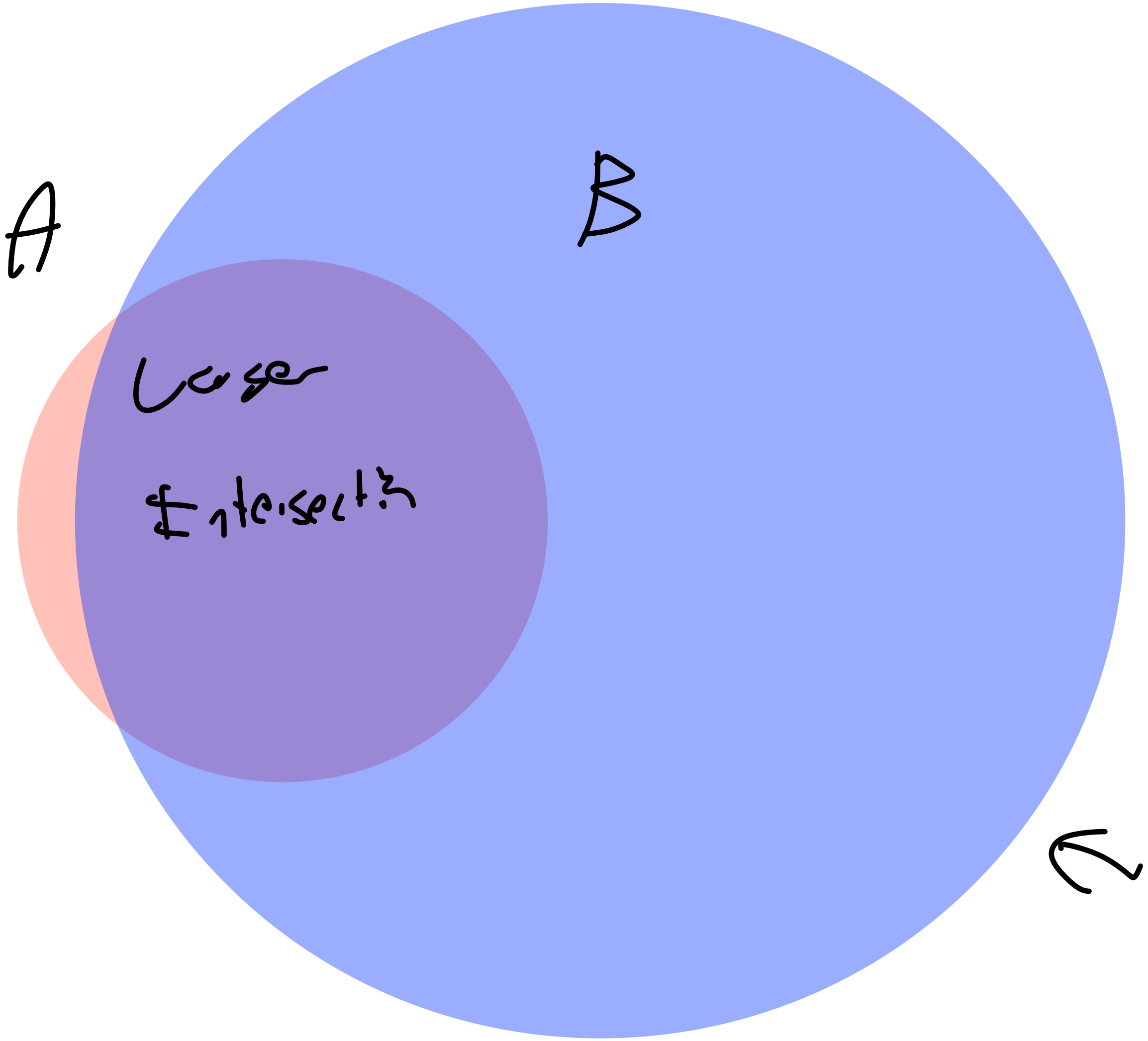
How can we describe how ***similar*** two sets are?

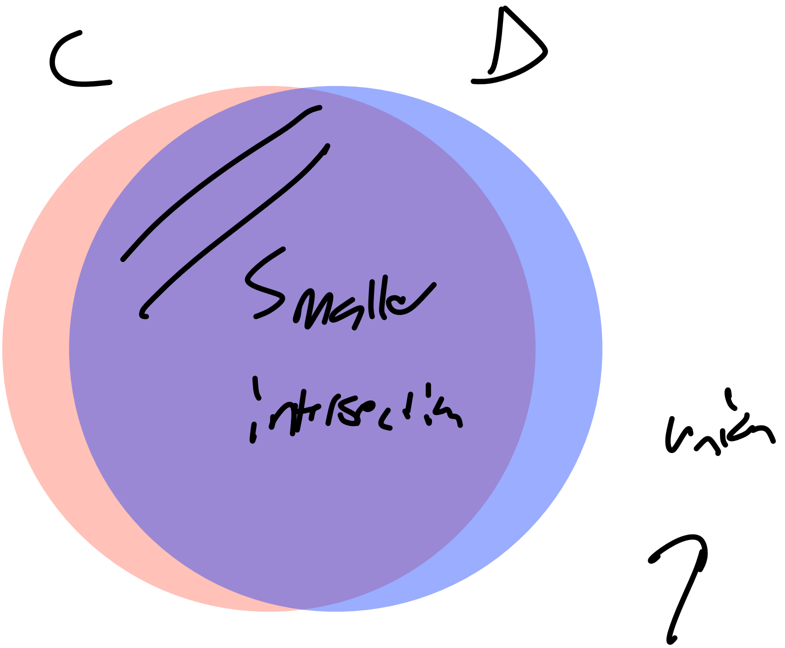




Set Similarity Review

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

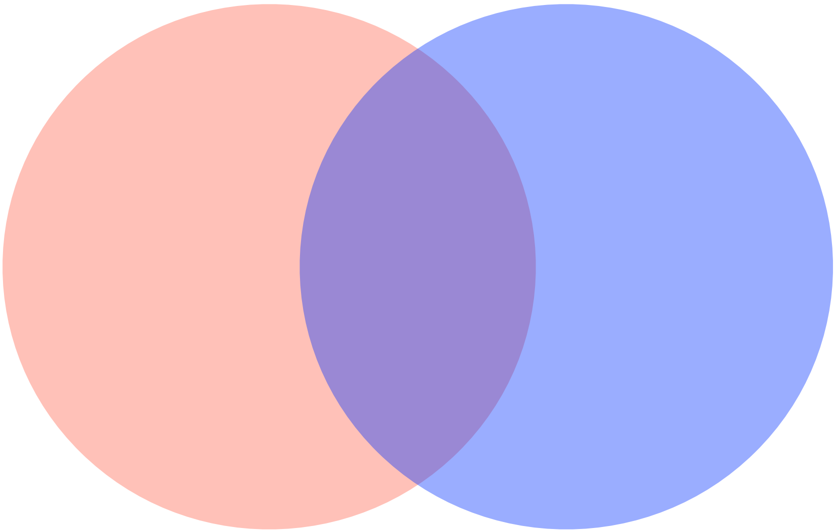




take into account the total size

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.



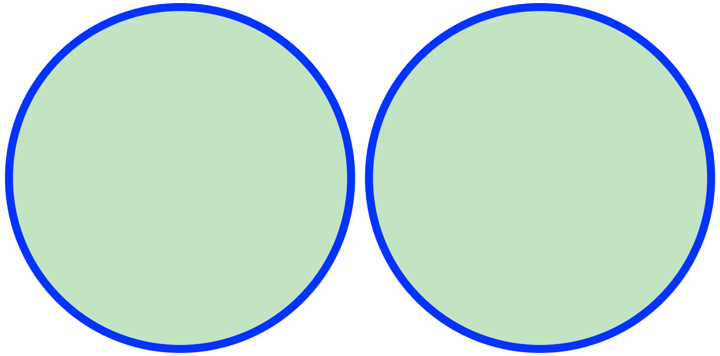
$$J = \frac{|A \cap B|}{|A \cup B|}$$

intersection

union

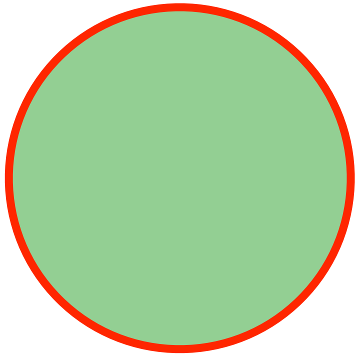
J is the ***Jaccard coefficient***

Set Similarity Review

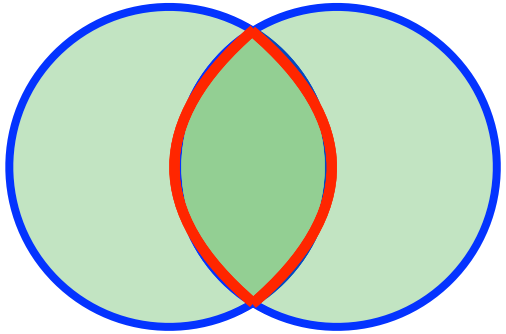


$$\frac{|A \cap B|}{|A \cup B|} = 0$$

Non zero



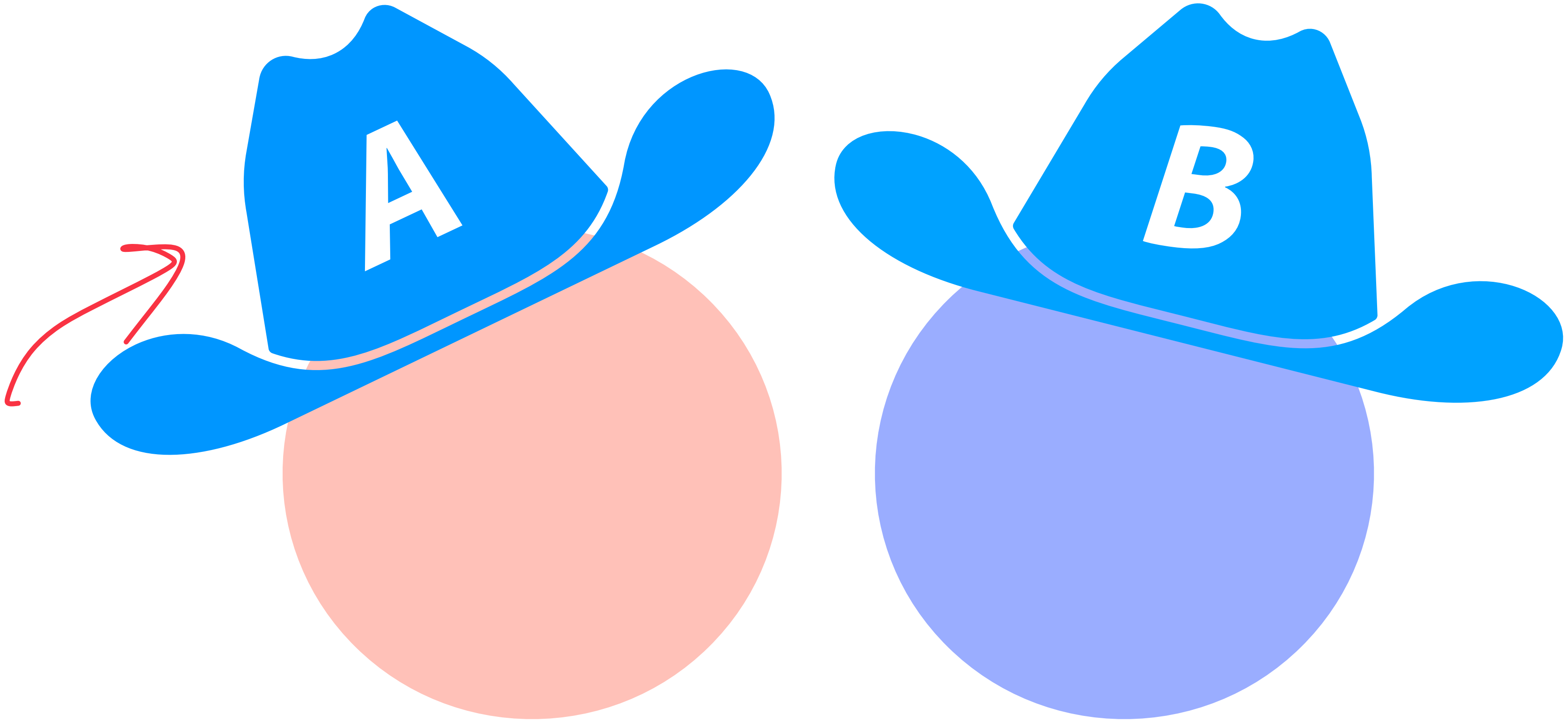
$$\frac{|A \cap B|}{|A \cup B|} = 1$$



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

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

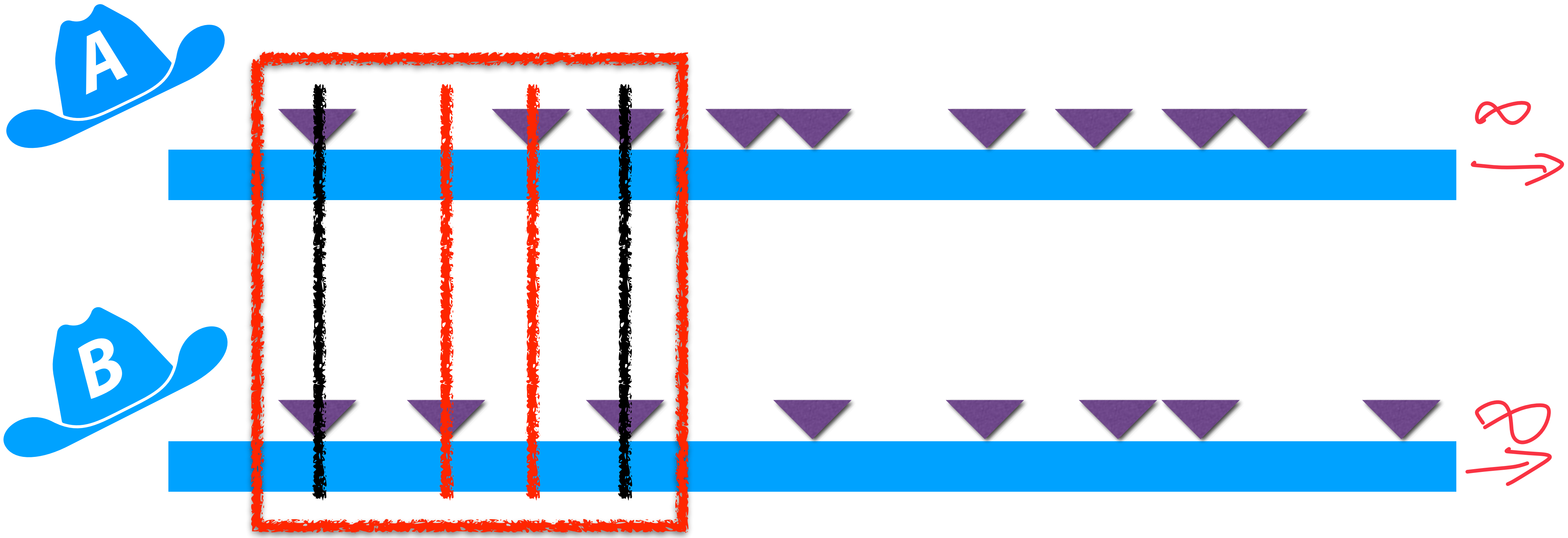


Image inspired by: Ondov B, Starrett G, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. **Mash Screen: high-throughput sequence containment estimation for genome discovery**. *Genome Biol* 20, 232 (2019)

Similarity Sketches

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

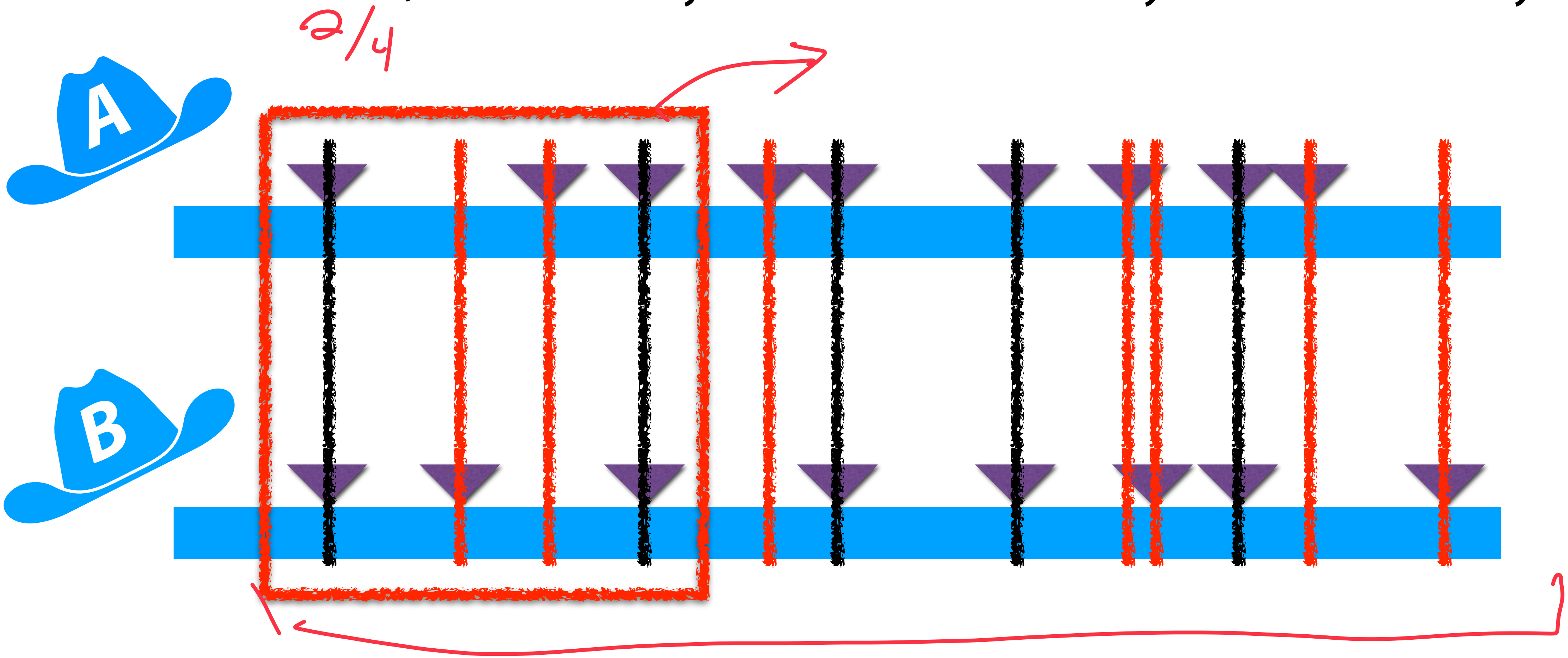
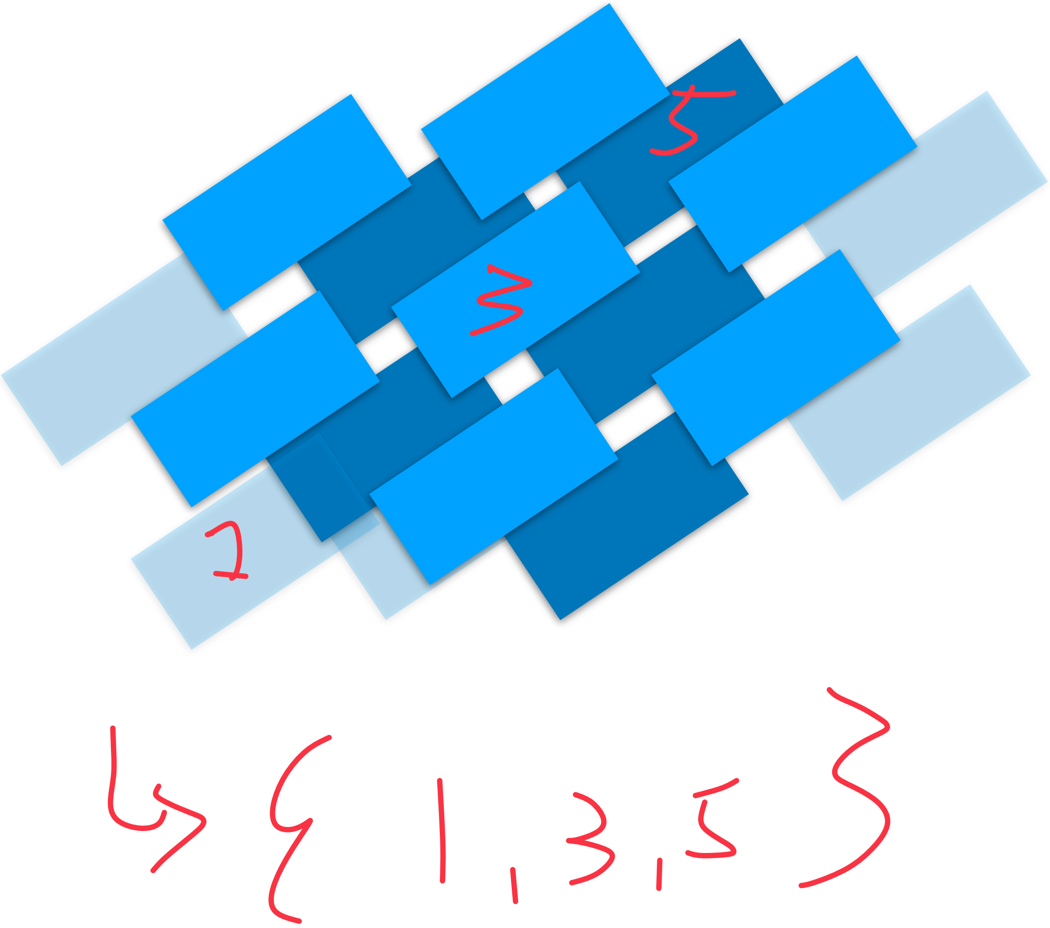
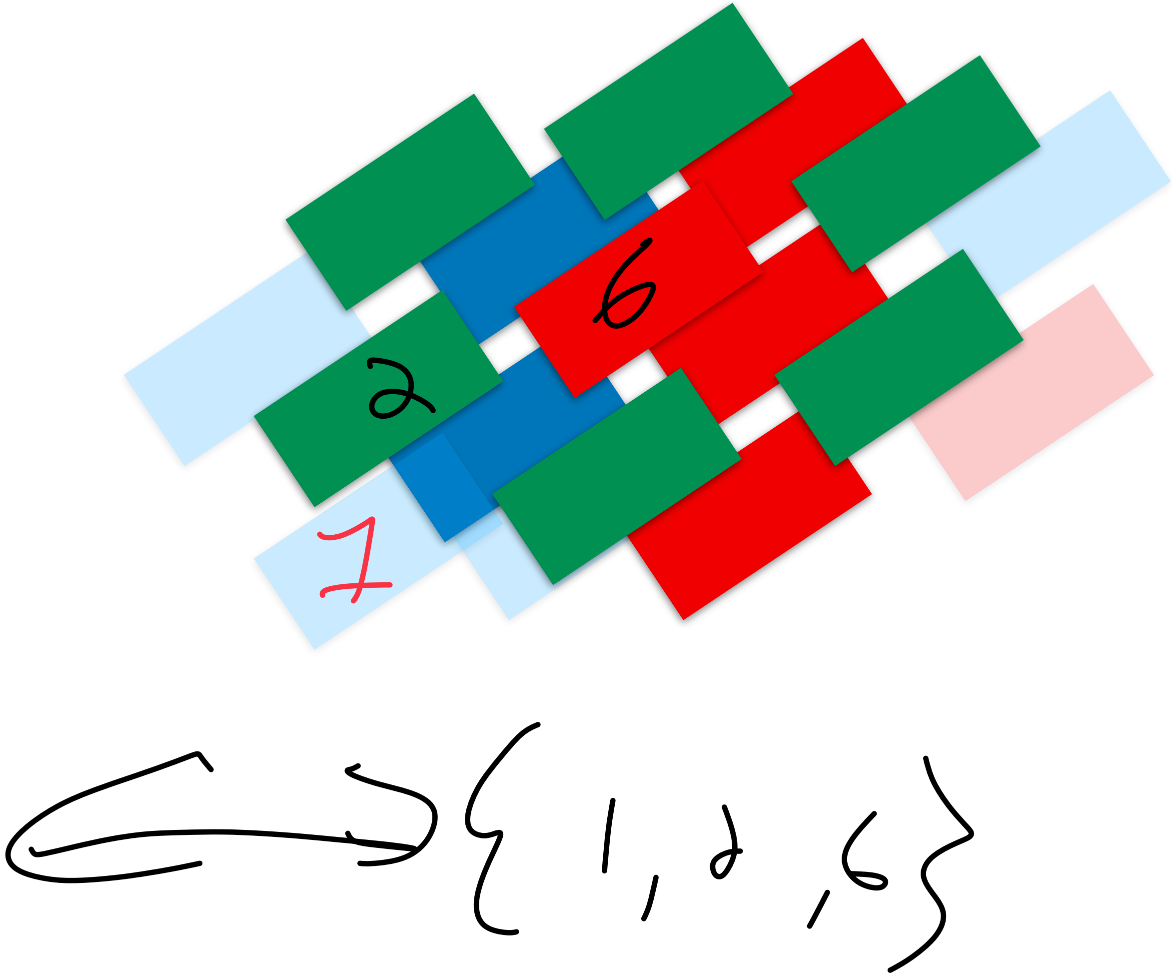


Image inspired by: Ondov B, Starrett G, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. **Mash Screen: high-throughput sequence containment estimation for genome discovery**. *Genome Biol* 20, 232 (2019)

Minhash Sketch

An approximation for a full dataset capable of **estimating set similarity**





Minhash Sketch 'ADT' (Use Cases)

Constructor → K min values

Cardinality Estimation — Seen this

Set Similarity Estimation

MinHash Construction

A MinHash sketch has three required inputs: (parameters)

1. A dataset (PNG, list, text file, ...) → int?
2. An integer K (the # of min elements)
3. One or more hash functions

MinHash Construction

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

h(x) = x % 7

k = 3

Assume h(x) is SHA1 min values

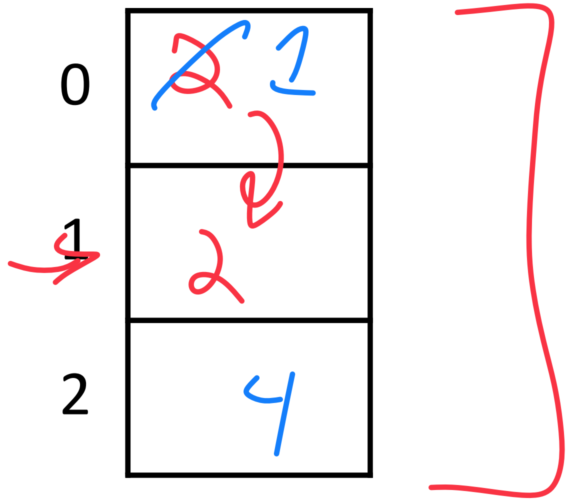
8 % 7 = 1

16 % 7 = 2

1) Hash each item

15 % 7 = 1

k min values in order



Minhash sketch is a set (No duplicates)

MinHash Cardinality Estimation → # of unique items

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

h(x) = x % 7 → {0, 1, 2, 3, 4, 5, 6}   (m-1)

k = 3

0	1
1	2
2	4

1) Normalize hash values

$M_1 = 1/6 = \frac{1}{N+1} \to N = 6-1 = 5 \approx 5$

$M_2 = 2/6 = \frac{2}{N+1}$

$M_3 = 4/6 = \frac{3}{N+1}$

* Normalization & How to calc K min values

MinHash Jaccard Estimation

Let's assume we have sets A and B sampled uniformly from [0, 100).

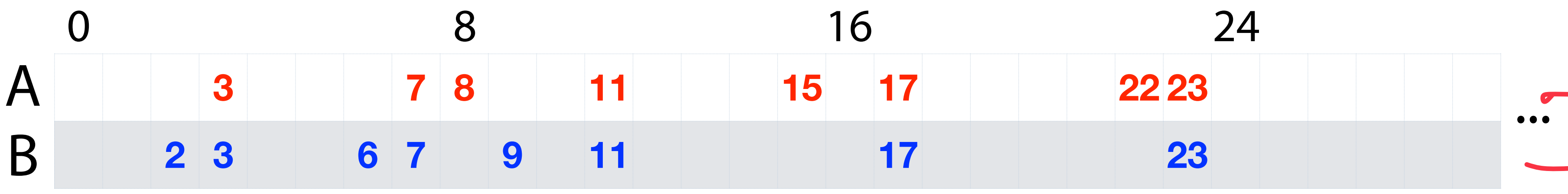
Instead of storing A & B, we store the bottom-8 **MinHash**

Sketch A

3	15
7	17
8	22
11	23

Sketch B

2	9
3	11
6	17
7	23



MinHash Jaccard Estimation

$$\frac{A \cap B}{A \cup B}$$

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

Sketch A

3	15
7	17
8	22
11	23

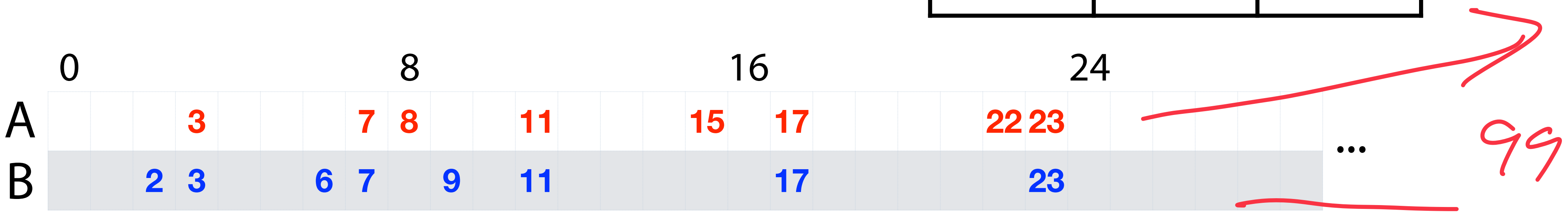
Sketch B

2	9
3	11
6	17
7	23

$|A \cup B|$

2	8	17
3	9	22
6	11	23
7	15	

…



MinHash Jaccard Estimation

8 min items

We dont *know* $|A \cup B|$, but we can estimate it!

Sketch of $|A \cup B|$ → ???

Sketch A

3	15
7	17
8	22
11	23

$\cup$

Sketch B

2	9
3	11
6	17
7	23

$=$

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

…

MinHash Jaccard Estimation

We can estimate the cardinality of $|A \cup B|$ from this sketch.

Sketch of $|A \cup B|$

2	8
3	9
6	11
7	15

KMV

Our sets sampled from [0, 100).

est size of $|A \cup B|$

$$15/99 = \frac{8}{N+1}$$

MinHash Jaccard Estimation

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

min 8 values in intersection

Sketch A

3	15
7	17
8	22
11	23

$\cap$

Sketch B

2	9
3	11
6	17
7	23

$=$

Sketch of $|A \cap B|$

3	23
7	X
11	X
17	X

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 are 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 of $|A \cap B|$

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

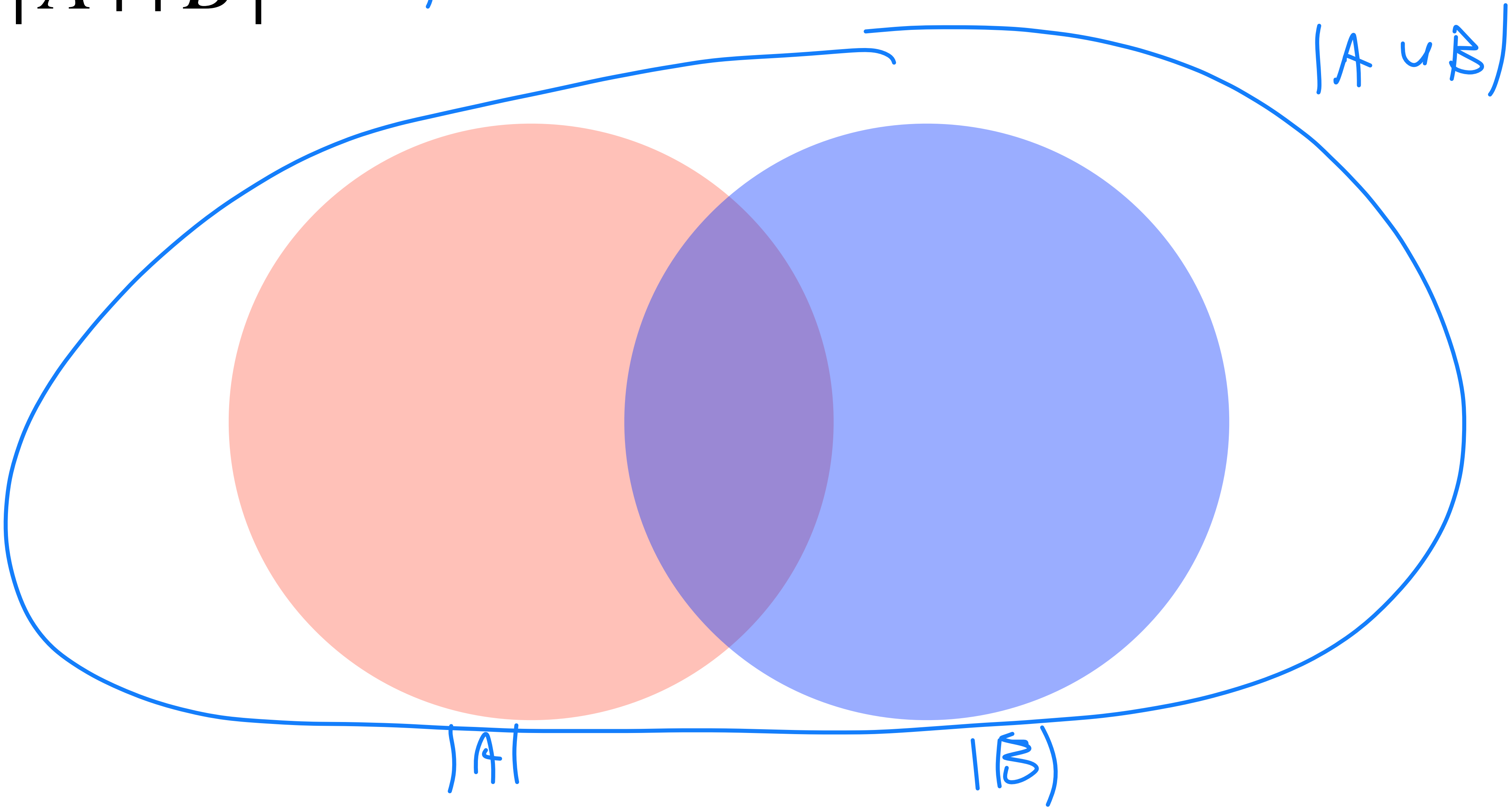
Using MinHash sketches, we can estimate $|A|$, $|B|$, and $|A \cup B|$

Not intersection

Is this enough to estimate the Jaccard?

Inclusion-Exclusion Principle

$$|A \cap B| = |A| + |B| - |A \cup B|$$



MinHash Jaccard Estimation

Math double check later!

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

$k = 8$ 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

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

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

The MinHash Sketch

We can also estimate cardinality directly using our sketches!

Sketch A

3	15
7	17
8	22
11	23

Sketch B

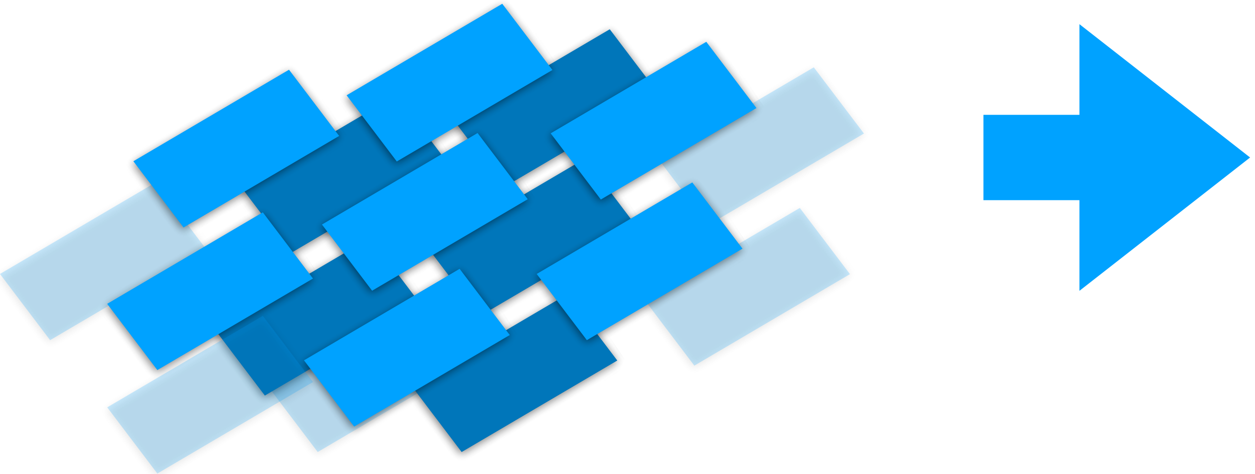
2	9
3	11
6	17
7	23

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

The **easiest** version of MinHash uses k hashes. How might this work?

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)

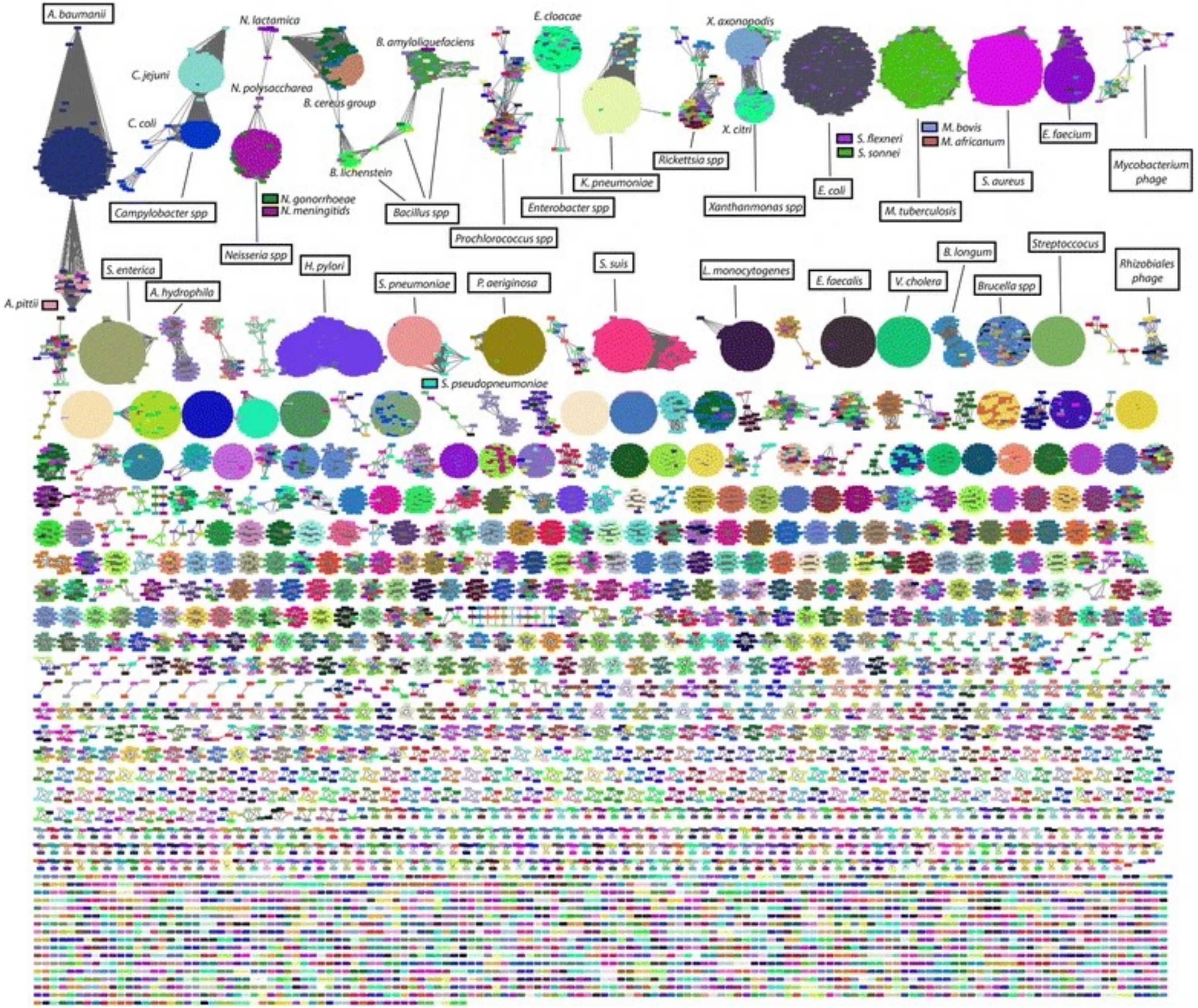
[5, 1, 6, 6]
Sketch (S_2)

$J(S_1, S_2) \approx 2/4 = 0.5$

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



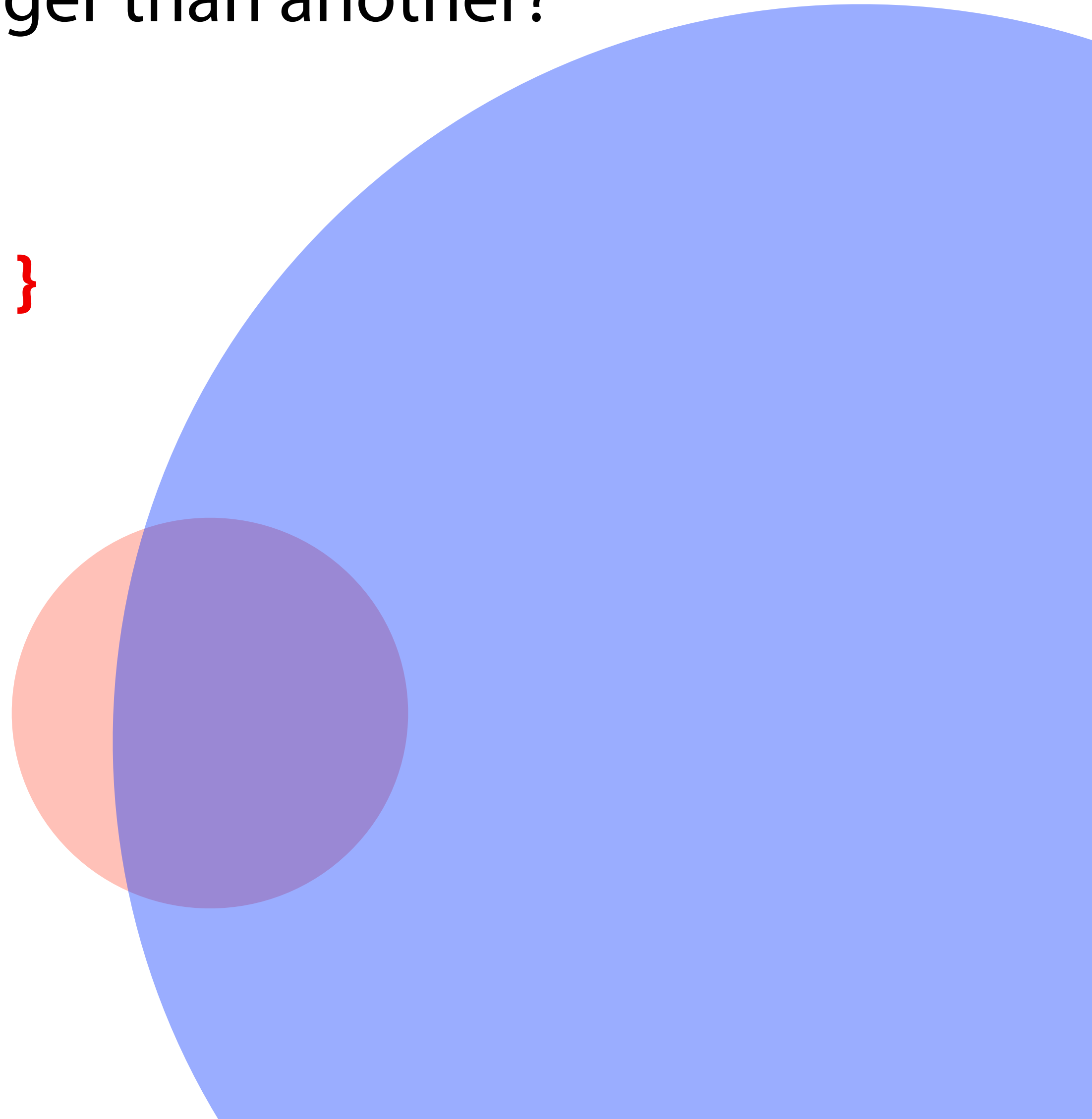
Mash: fast genome and metagenome distance estimation using MinHash
Ondov et al (2016) *Genome Biology*

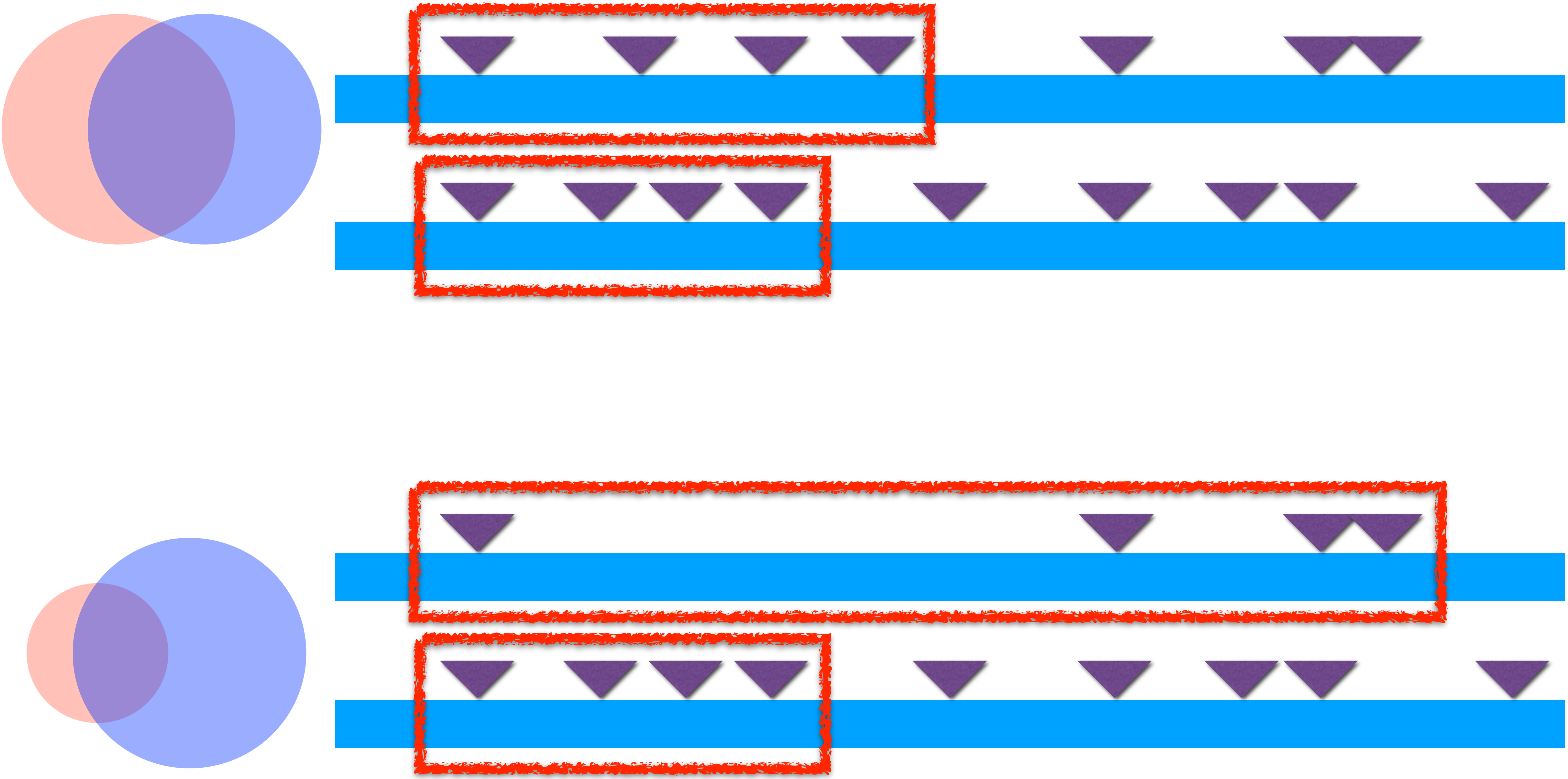
Alternative MinHash Sketch Approaches

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

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

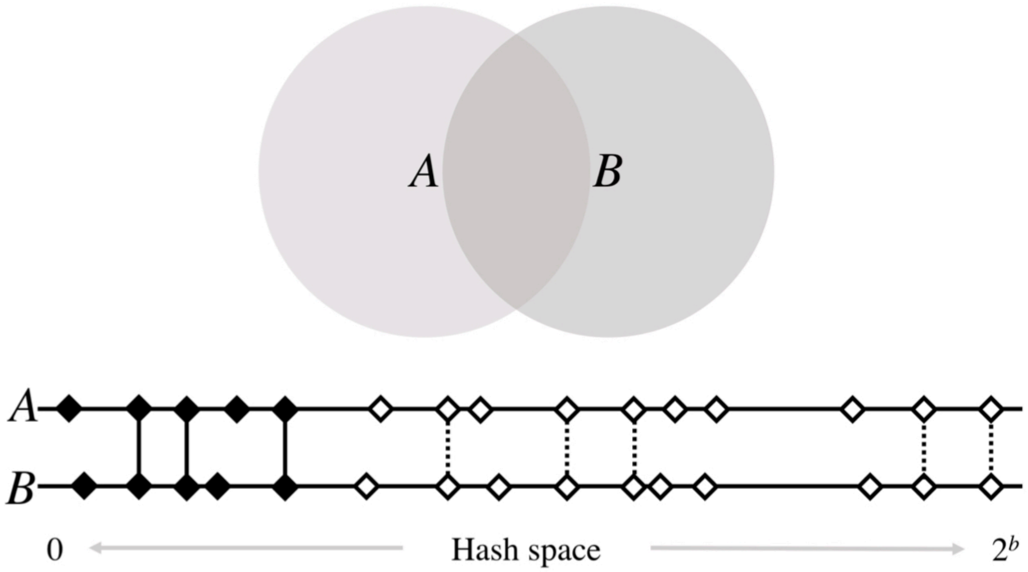
S_2 = { 1, 2, 3, 4, 5, 6, 7, … 59, 82, 101, … }

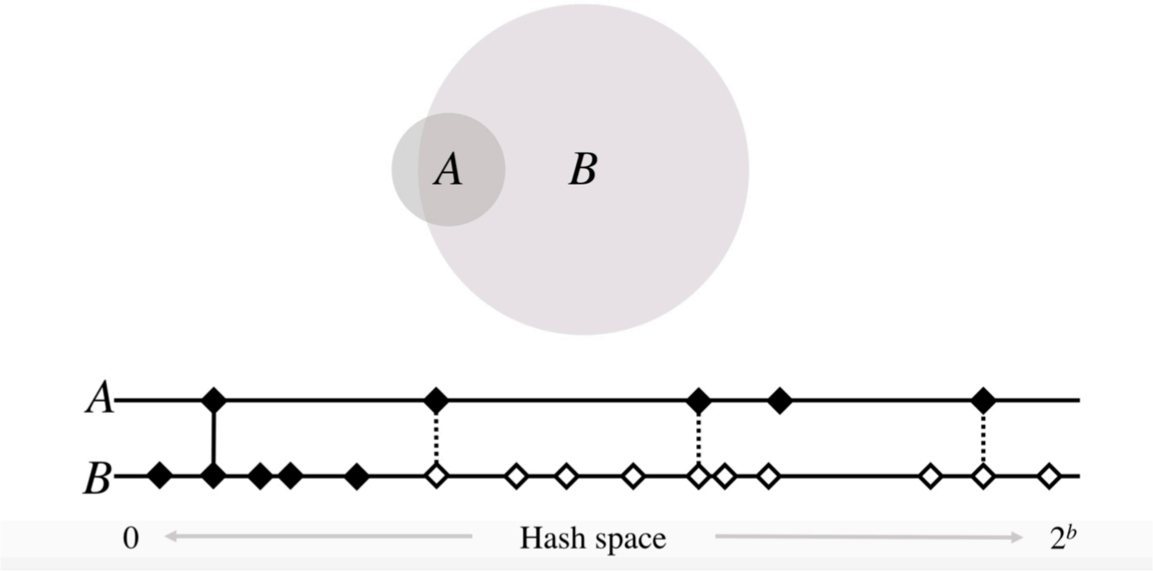




Alternative MinHash sketches

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

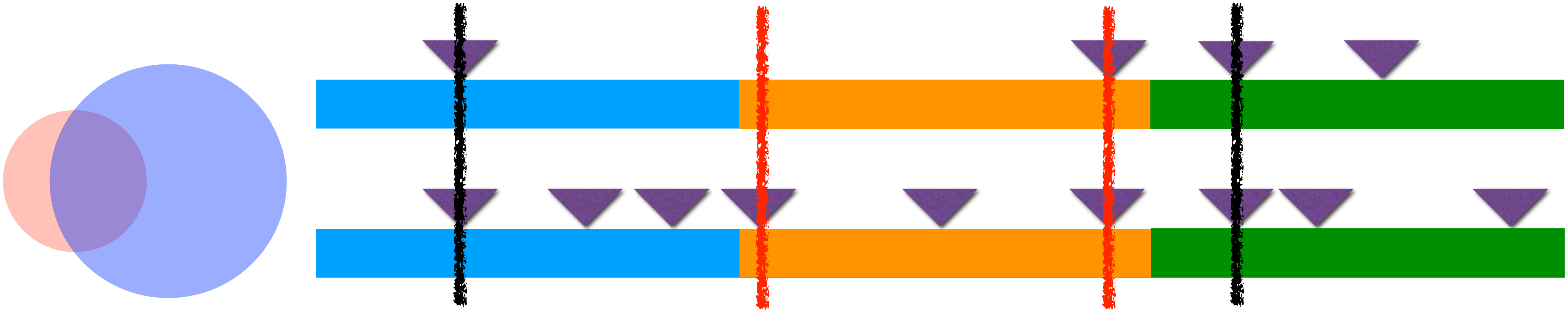




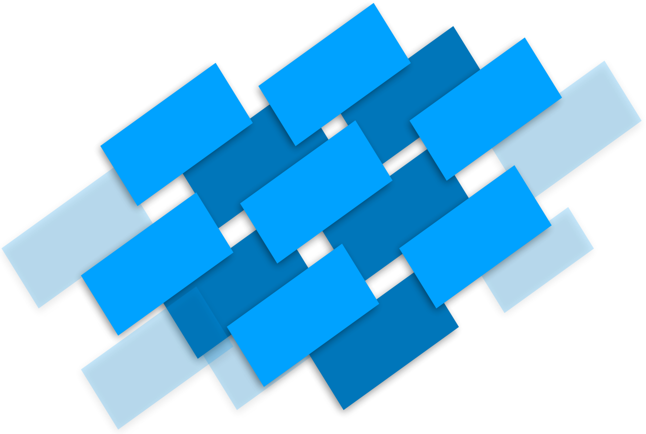
Ondov, Brian D., Gabriel J. Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B. Buck, and Adam M. Phillippy. **Mash Screen: High-throughput sequence containment estimation for genome discovery**. *Genome biology* 20.1 (2019): 1-13.

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

K-Partition Minhash

Hint: What bitwise operator will allow me to do this?

```
1010110101    →  Partition  →    10
0001111010                       00
```

What information do I need to do this in general?

MP_Sketching: A MinHash experiment

Using legitimate hashes, write MinHash sketch three ways:

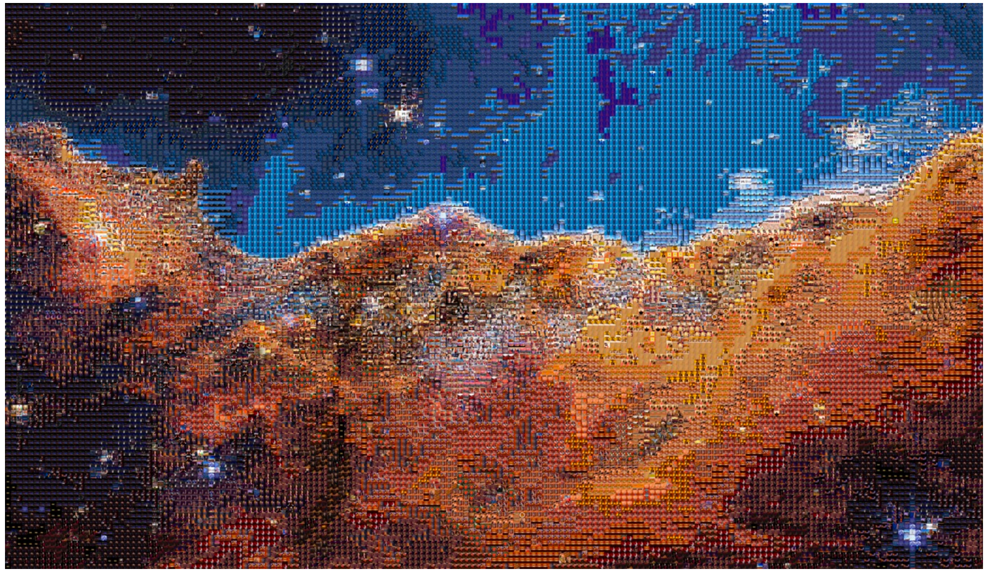
```
std::vector<uint64_t> khash_minhash(std::vector<int> inList, std::vector<hashFunction> hv);
```

```
std::vector<uint64_t> kminhash(std::vector<int> inList, unsigned k, hashFunction h);
```

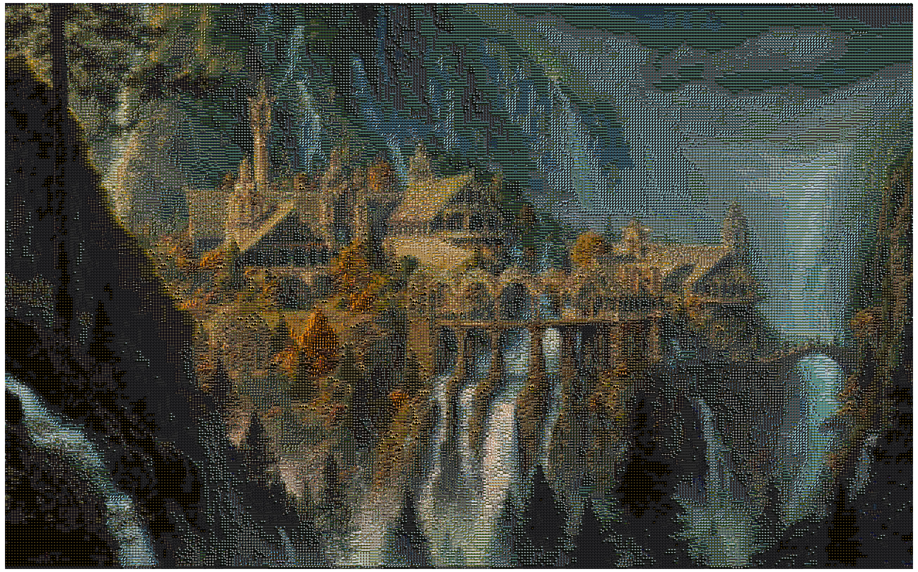
```
std::vector<uint64_t> kpartition_minhash(std::vector<int> inList, int part_bits, hashFunction h);
```

MP_Sketching: A MinHash experiment

Use MinHash sketches to estimate PNG similarity



Mosaics (Discord: Bose)



Mosaics (Discord: LightningStorm)

MP_Sketching: A MinHash experiment

Build a weighted graph of every possible pairwise comparison!