Data Structures and Algorithms
Cardinality

CS 225
G Carl Evans

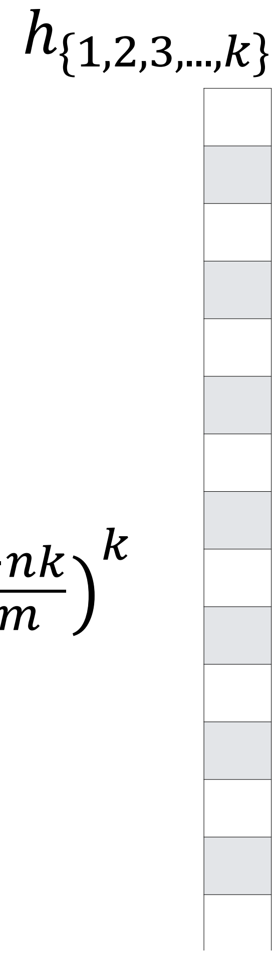
April 30, 2025

Department of Computer Science

From slides by Brad Solomon

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

Bloom Filter Use Cases

Which of the following problems can be solved with a bloom filter?

A) Find the closest matching item to a query of interest

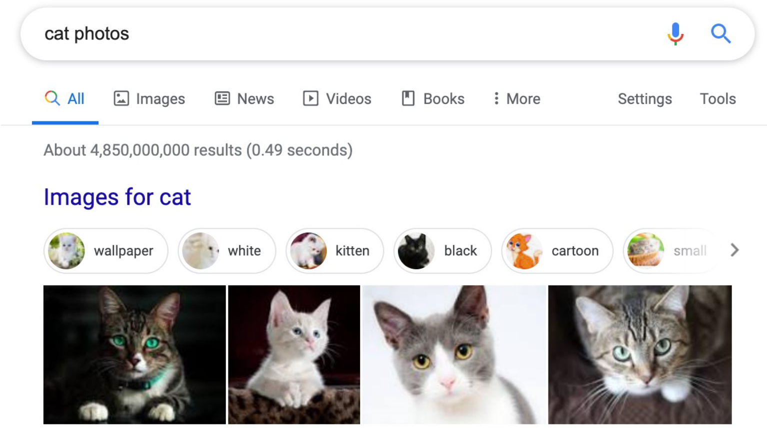
B) Check if a query exists in a dataset

C) Compare the similarity between two datasets

D) Count the number of unique items in a dataset

Cardinality

Sometimes its not possible or realistic to count all objects!



Estimate: 60 billion — 130 trillion

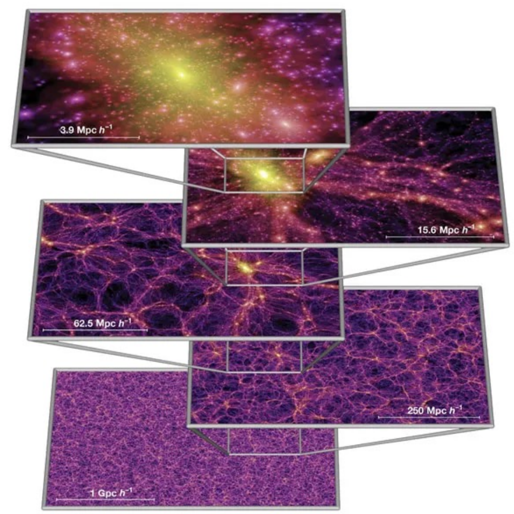


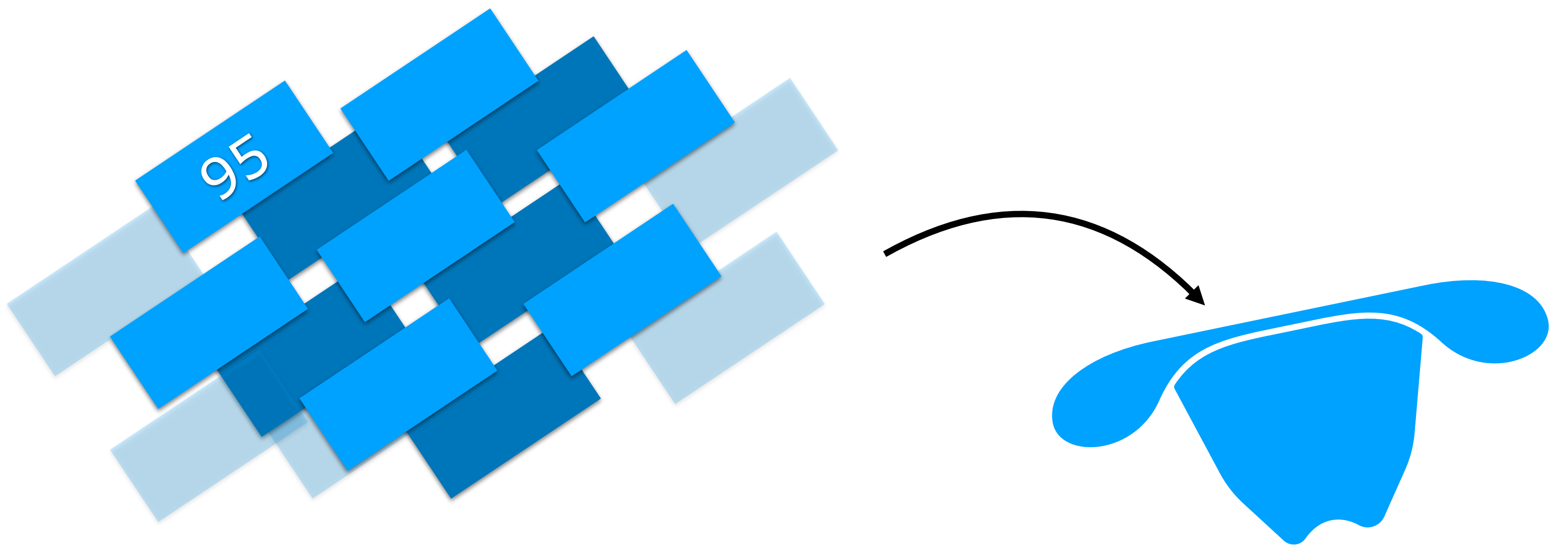
Image: https://doi.org/10.1038/nature03597

946
5581
8945
6145
8126
3887
8925
1246
8324
4549
9100
5598
8499
8970
3921
8575
4859
4960
42
6901
4336

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?

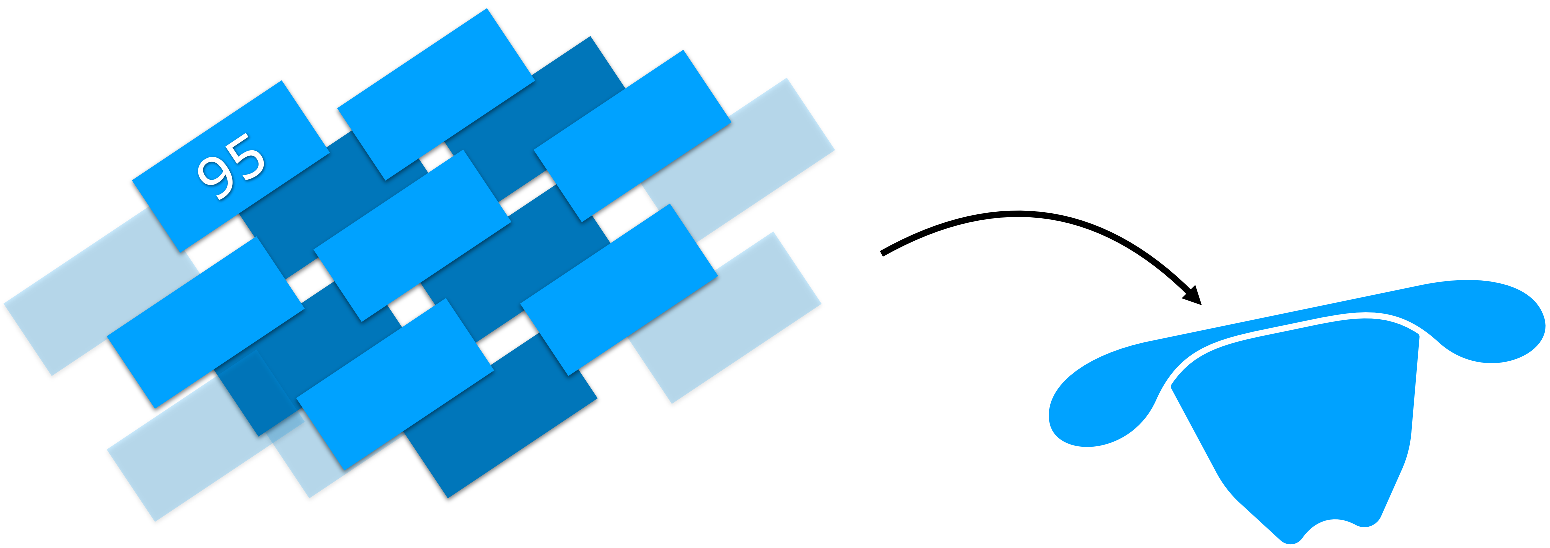


Analogy from Ben Langmead

Cardinality Estimation

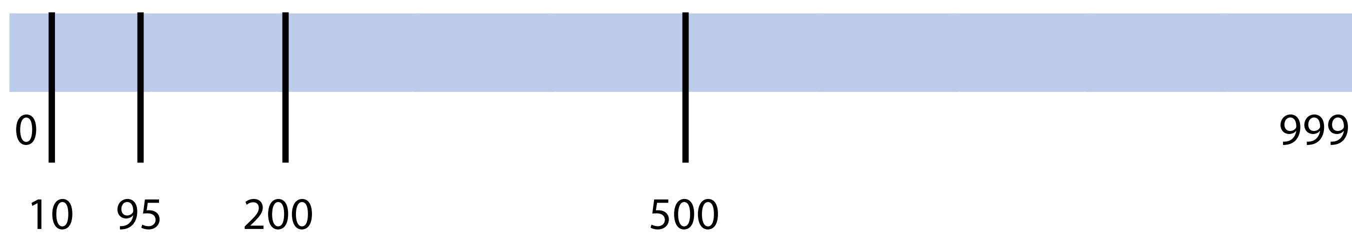
Imagine I fill a hat with **a random subset** of numbered cards **from 0 to 999**

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



Cardinality Estimation

Imagine we have multiple uniform random sets with different minima.



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

Cardinality Estimation

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

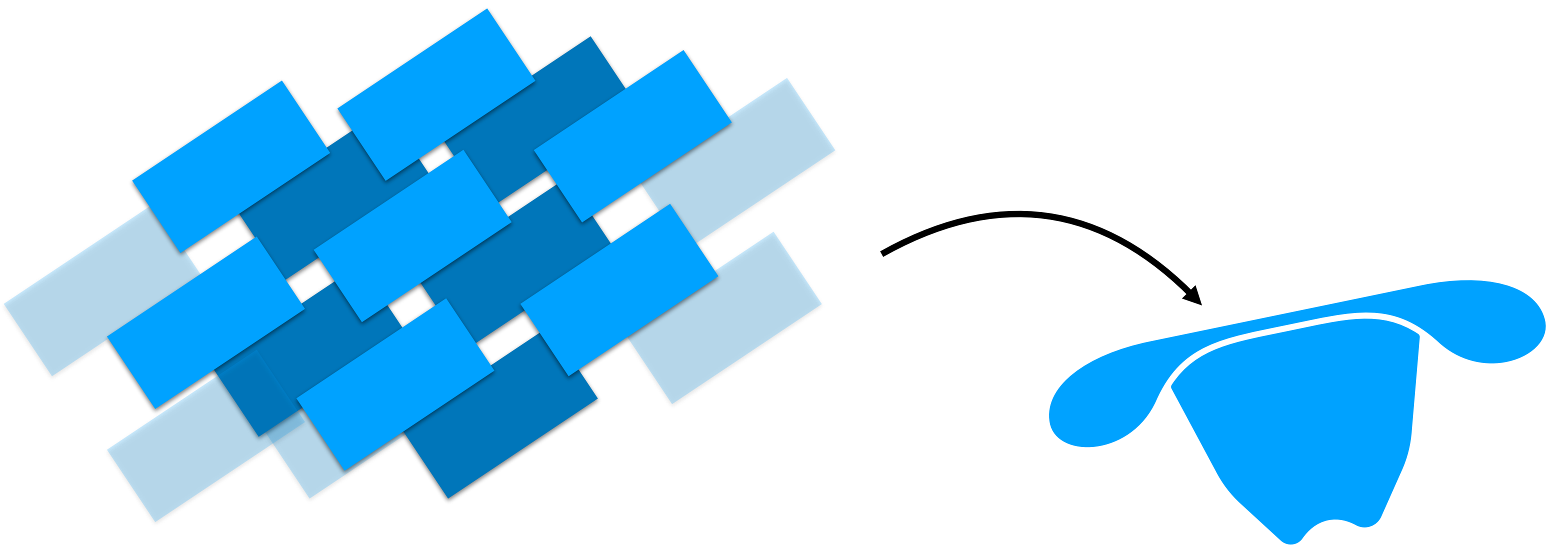
Conceptually: If we scatter N points randomly across the interval, we end up with $N + 1$ partitions, each about $1000/(N+1)$ long

Assuming our first 'partition' is about average:

$$\begin{aligned} 95 &\approx 1000/(N+1) \\ N+1 &\approx 10.5 \\ N &\approx 9.5 \end{aligned}$$

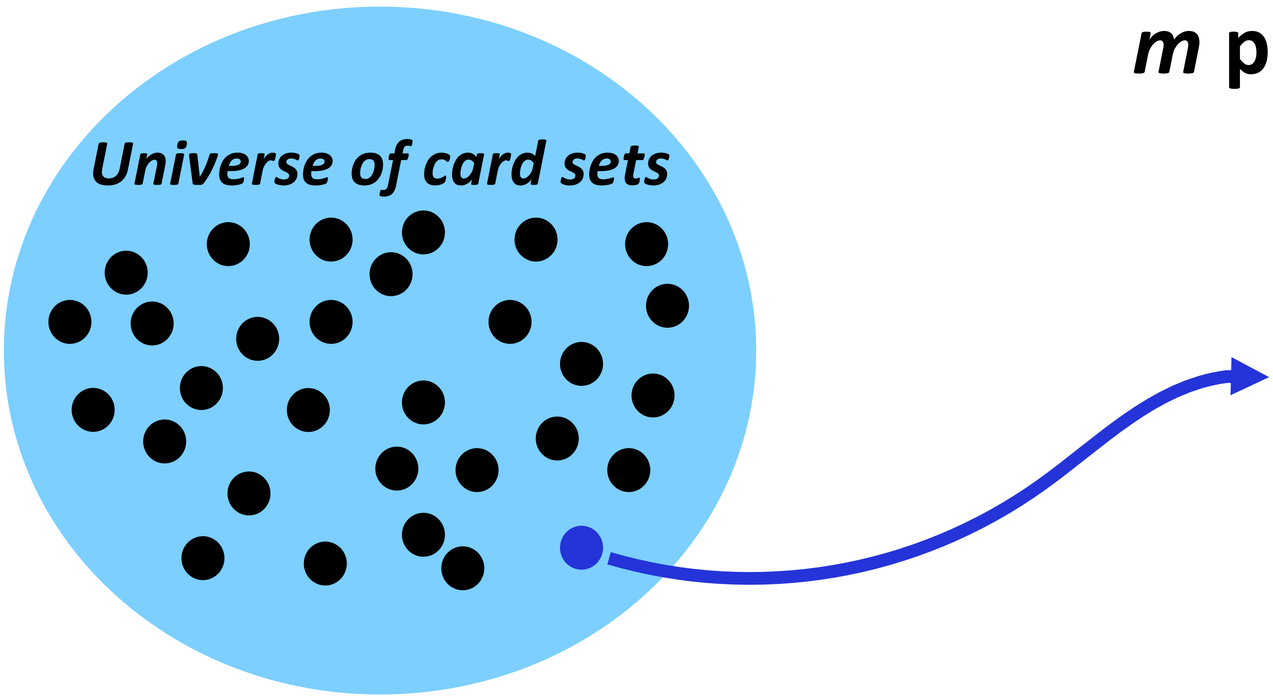
Cardinality Estimation

Why do we care about “the hat problem”?



Cardinality Estimation

Why do we care about "the hat problem"?



***m* possible minima**

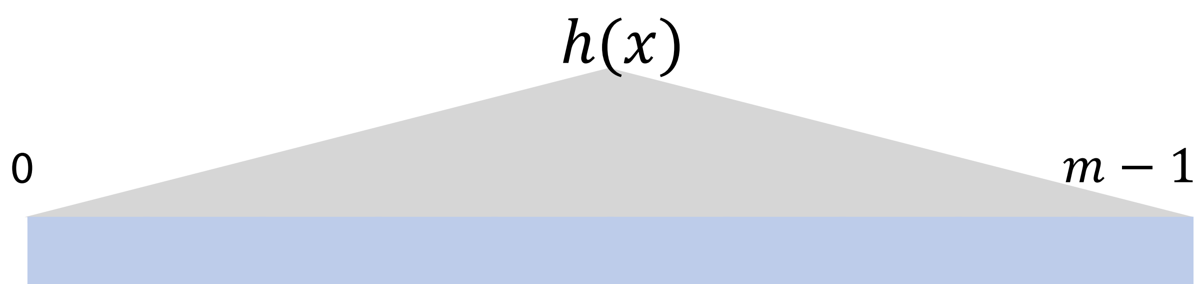
Key	Value

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!



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…

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

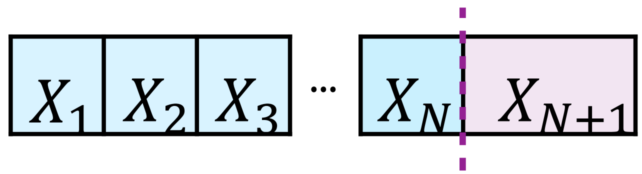
Cardinality Sketch

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

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

Cardinality Sketch

Consider an $N + 1$ draw:

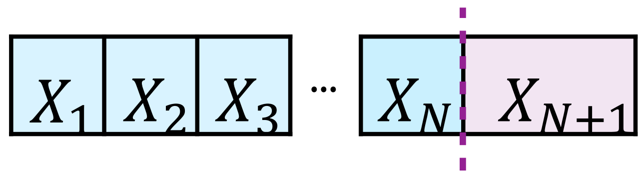


$$M = \min_{1 \le i \le N} X_i$$

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

Cardinality Sketch

Consider an $N + 1$ draw:



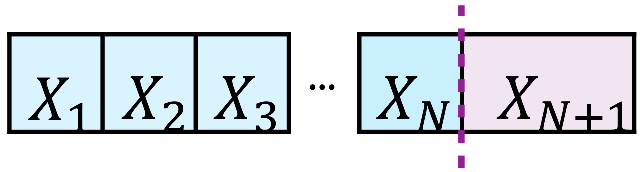
$$M = \min_{1 \le i \le N} X_i$$

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

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

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: X_1 | X_2 | X_3 ... X_N | X_{N+1}

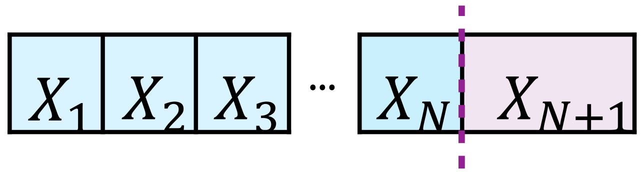
$$M = \min_{1 \le i \le N} X_i$$

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

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

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

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

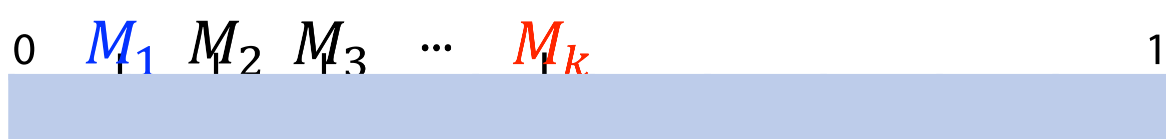
Cardinality Sketch

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

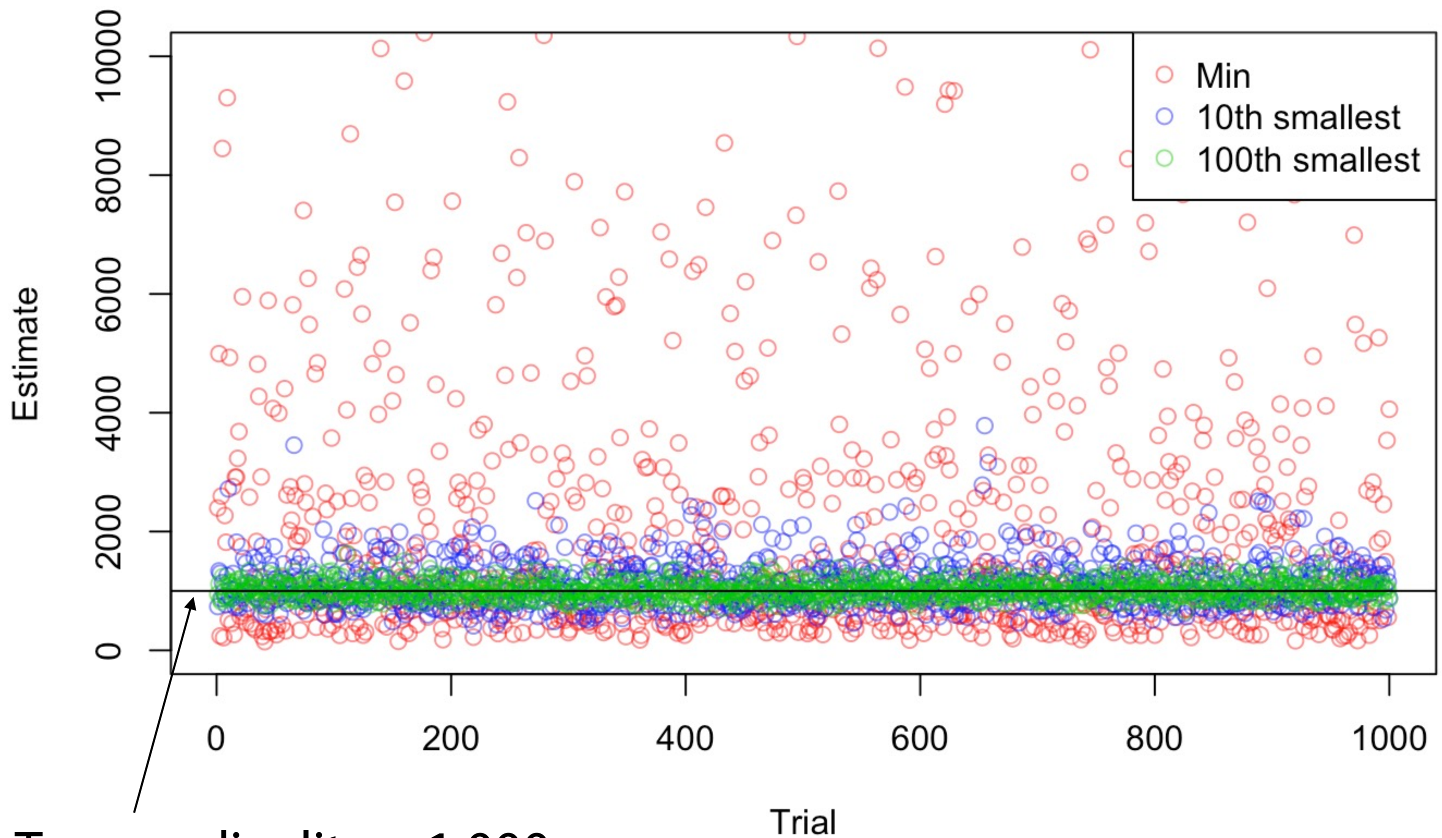
Cardinality Sketch

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

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



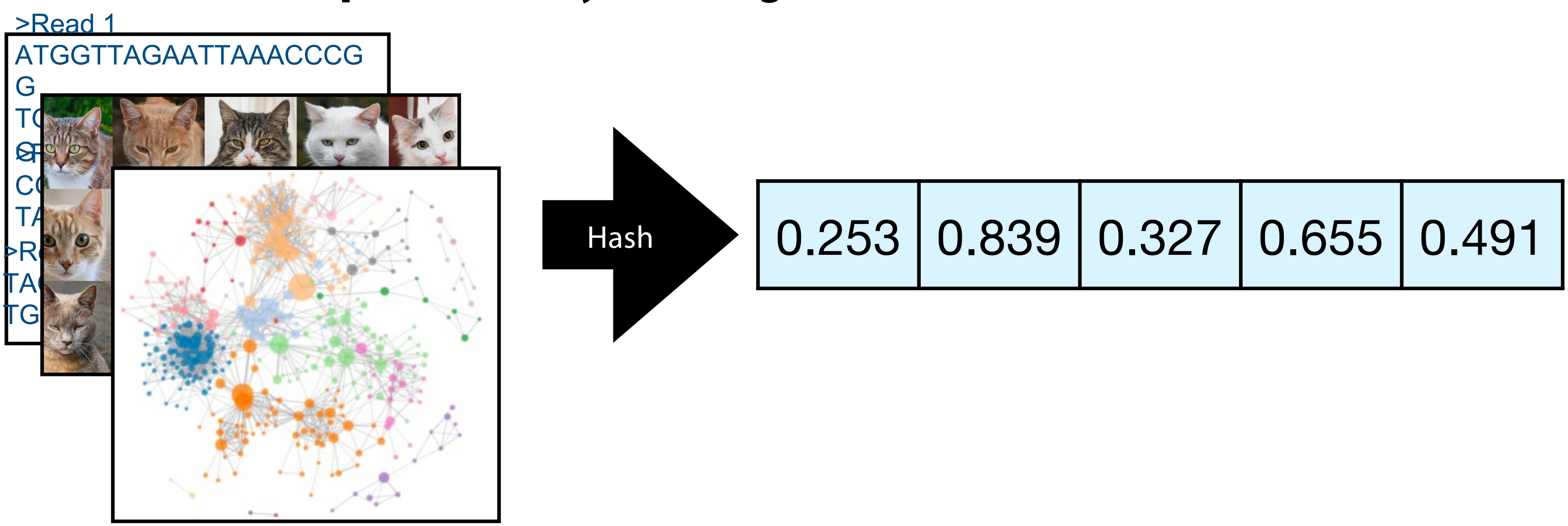
Cardinality Sketch



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**.



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

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

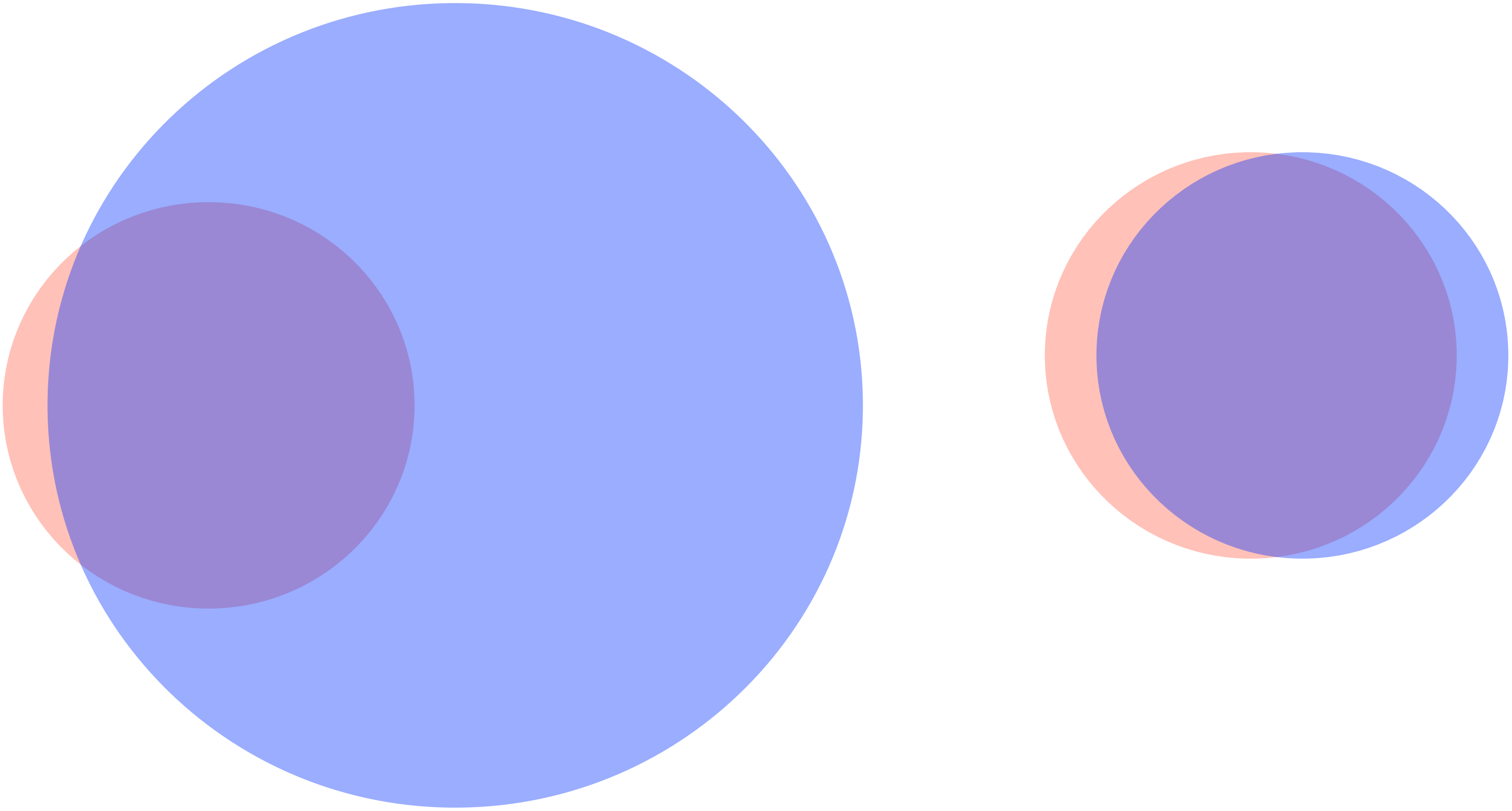
```
AGGCCACAGTGTATTATGACTG
|||||||||||  |||||||||
AGGCCACAGTGAGTTATGACTG

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

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

Set Similarity Review

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

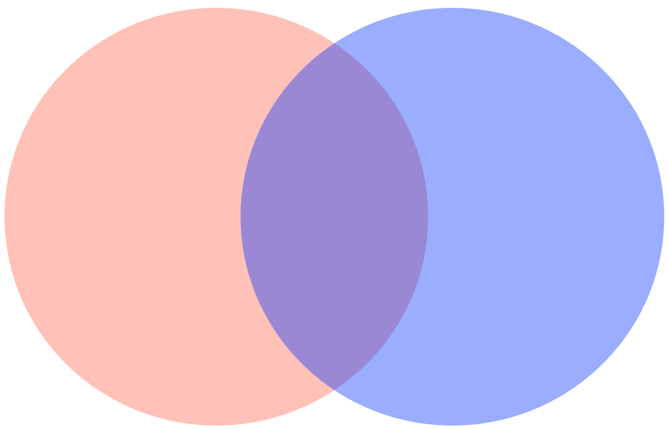
Set Similarity Review

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



Set Similarity Review

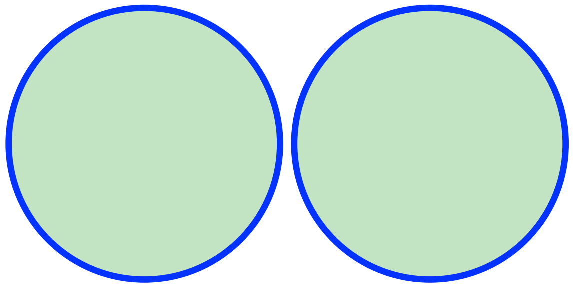
To measure **similarity** of A & B, we need both a measure of how similar the sets are but also the total size of both sets.



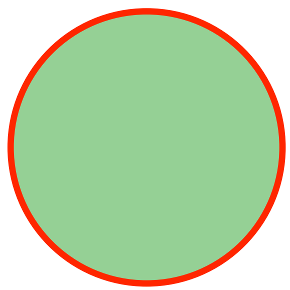
$$J = \frac{|A \cap B|}{|A \cup B|}$$

J is the ***Jaccard coefficient***

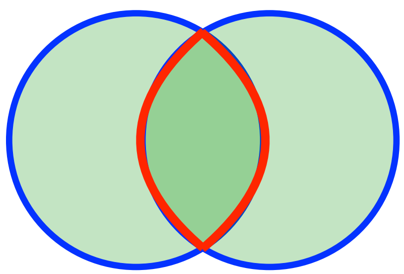
Set Similarity Review



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



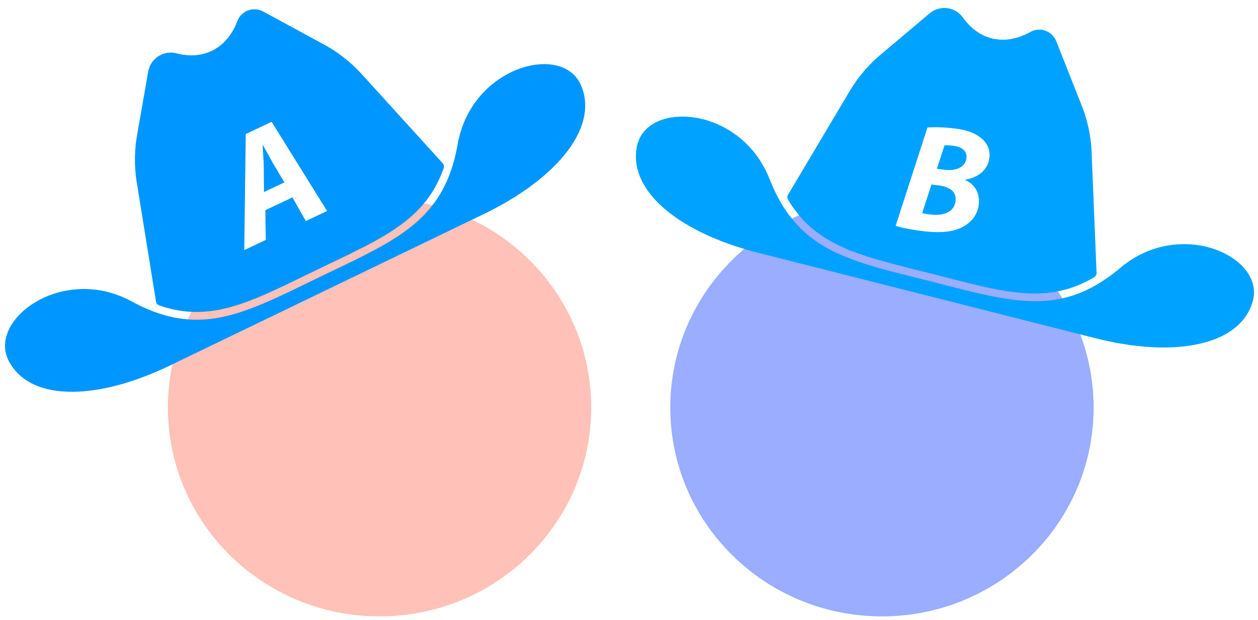
$$\frac{|A \cap B|}{|A \cup B|} = 1$$



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

Similarity Sketches

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



Similarity Sketches

Imagine we have two datasets represented by their kth minimum values

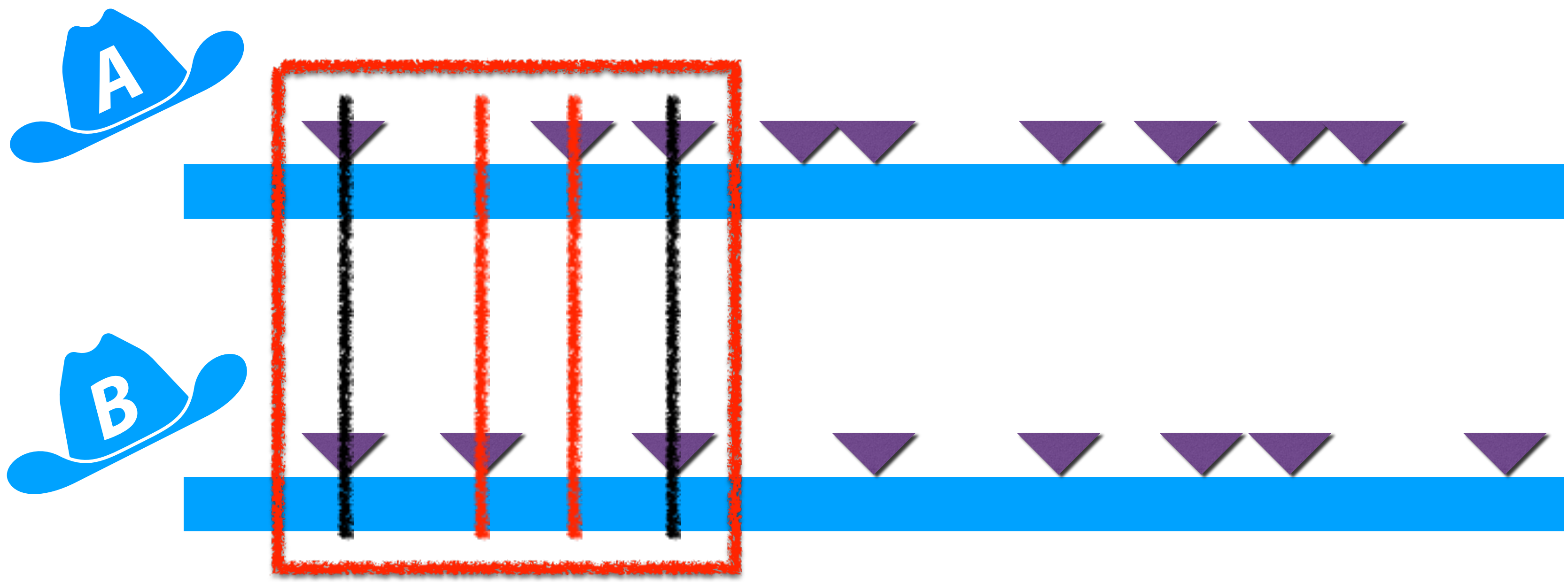


Image inspired by: Ondov B, Starrett G, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. **Mash Screen: high-throughput sequence containment estimation for genome discovery**. Genome Biol 20, 232 (2019)

Similarity Sketches

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

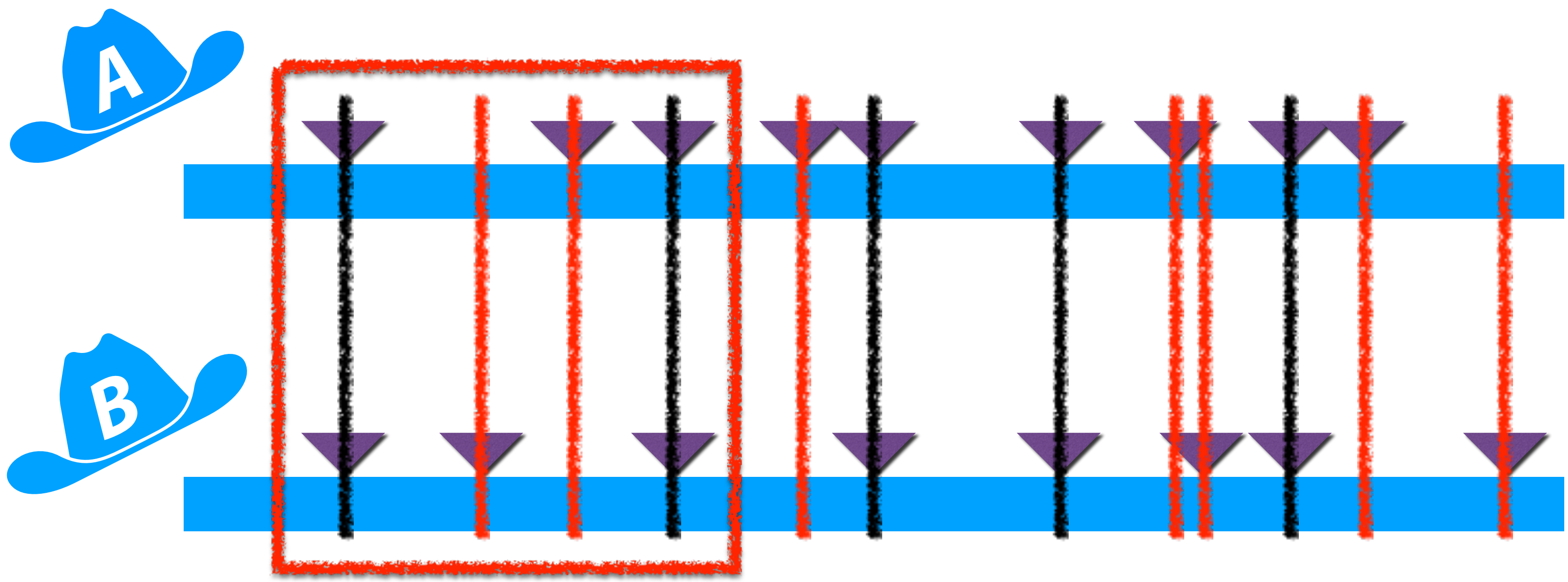


Image inspired by: Ondov B, Starrett G, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. **Mash Screen: high-throughput sequence containment estimation for genome discovery**. Genome Biol 20, 232 (2019)