Want to store a set $S$ of items from a massive universe of possible items (e.g., images, text documents, IP addresses).

- Allow small probability $\delta > 0$ of false positives. I.e., for any $x$,

$$ \Pr(query(x) = 1 \text{ and } x \not\in S) \leq \delta. $$
Want to store a set $S$ of items from a massive universe of possible items (e.g., images, text documents, IP addresses).

**Goal:** support \( \text{insert}(x) \) to add \( x \) to the set and \( \text{query}(x) \) to check if \( x \) is in the set. Both in \( O(1) \) time.

- Allow small probability \( \delta > 0 \) of false positives. I.e., for any \( x \),

\[
\Pr(\text{query}(x) = 1 \text{ and } x \notin S) \leq \delta.
\]
Approximately Maintaining a Set

Want to store a set $S$ of items from a massive universe of possible items (e.g., images, text documents, IP addresses).

**Goal:** support $insert(x)$ to add $x$ to the set and $query(x)$ to check if $x$ is in the set. Both in $O(1)$ time.

- Allow small probability $\delta > 0$ of false positives. I.e., for any $x$,

$$\Pr(query(x) = 1 \text{ and } x \notin S) \leq \delta.$$
Want to store a set $S$ of items from a massive universe of possible items (e.g., images, text documents, IP addresses).

**Goal:** support $\text{insert}(x)$ to add $x$ to the set and $\text{query}(x)$ to check if $x$ is in the set. Both in $O(1)$ time.

- Allow small probability $\delta > 0$ of false positives. I.e., for any $x$,

\[
\Pr(\text{query}(x) = 1 \text{ and } x \notin S) \leq \delta.
\]
Want to store a set $S$ of items from a massive universe of possible items (e.g., images, text documents, IP addresses).

**Goal:** support *insert*($x$) to add $x$ to the set and *query*($x$) to check if $x$ is in the set. Both in $O(1)$ time.

- Allow small probability $\delta > 0$ of false positives. I.e., for any $x$,

$$\Pr(\text{query}(x) = 1 \text{ and } x \notin S) \leq \delta.$$ 

**Solution:** Bloom filters (repeated random hashing). Will use much less space than a hash table.
BLOOM FILTERS

Chose $k$ independent random hash functions $h_1, \ldots, h_k$ mapping the universe of elements $U \rightarrow [m]$. No false negatives. False positives more likely with more insertions.
BLOOM FILTERS

Chose \( k \) independent random hash functions \( h_1, \ldots, h_k \) mapping the universe of elements \( U \rightarrow [m] \).

- Maintain an array \( A \) containing \( m \) bits, all initially 0.
Chose $k$ independent random hash functions $h_1, \ldots, h_k$ mapping the universe of elements $U \rightarrow [m]$.

- Maintain an array $A$ containing $m$ bits, all initially 0.
- $\text{insert}(x)$: set all bits $A[h_1(x)] = \ldots = A[h_k(x)] := 1$. 

No false negatives. False positives more likely with more insertions.
BLOOM FILTERS

Chose \( k \) independent random hash functions \( h_1, \ldots, h_k \) mapping the universe of elements \( U \rightarrow [m] \).

- Maintain an array \( A \) containing \( m \) bits, all initially 0.
- \( insert(x) \): set all bits \( A[h_1(x)] = \ldots = A[h_k(x)] := 1 \).
- \( query(x) \): return 1 only if \( A[h_1(x)] = \ldots = A[h_k(x)] = 1 \).

No false negatives. False positives more likely with more insertions.
BLOOM FILTERS

Chose $k$ independent random hash functions $h_1, \ldots, h_k$ mapping the universe of elements $U \rightarrow [m]$.

• Maintain an array $A$ containing $m$ bits, all initially 0.
• $\text{insert}(x)$: set all bits $A[h_1(x)] = \ldots = A[h_k(x)] := 1$.
• $\text{query}(x)$: return 1 only if $A[h_1(x)] = \ldots = A[h_k(x)] = 1$.

m bit array $A$

| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

No false negatives. False positives more likely with more insertions.
BLOOM FILTERS

Chose $k$ independent random hash functions $h_1, \ldots, h_k$ mapping the universe of elements $U \rightarrow [m]$.

- Maintain an array $A$ containing $m$ bits, all initially 0.
- $insert(x)$: set all bits $A[h_1(x)] = \ldots = A[h_k(x)] := 1$.
- $query(x)$: return 1 only if $A[h_1(x)] = \ldots = A[h_k(x)] = 1$.

\[
\begin{array}{cccccccccccc}
\text{Insertions} \\
\hline
\hline
\text{m bit array } A \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}
\]

Queries:
BLOOM FILTERS

Chose $k$ independent random hash functions $h_1, \ldots, h_k$ mapping the universe of elements $U \rightarrow [m]$.

- Maintain an array $A$ containing $m$ bits, all initially 0.
- $\text{insert}(x)$: set all bits $A[h_1(x)] = \ldots = A[h_k(x)] := 1$.
- $\text{query}(x)$: return 1 only if $A[h_1(x)] = \ldots = A[h_k(x)] = 1$. 

\[ 
\begin{array}{cccccccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array} 
\]
BLOOM FILTERS

Chose $k$ independent random hash functions $h_1, \ldots, h_k$ mapping the universe of elements $U \to [m]$.

- Maintain an array $A$ containing $m$ bits, all initially 0.
- $\text{insert}(x)$: set all bits $A[h_1(x)] = \ldots = A[h_k(x)] := 1$.
- $\text{query}(x)$: return 1 only if $A[h_1(x)] = \ldots = A[h_k(x)] = 1$. 

No false negatives. False positives more likely with more insertions.
Chose $k$ independent random hash functions $h_1, \ldots, h_k$ mapping the universe of elements $U \rightarrow [m]$. 

- Maintain an array $A$ containing $m$ bits, all initially 0.
- $\text{insert}(x)$: set all bits $A[h_1(x)] = \ldots = A[h_k(x)] := 1$.
- $\text{query}(x)$: return 1 only if $A[h_1(x)] = \ldots = A[h_k(x)] = 1$. 

No false negatives. False positives more likely with more insertions.
Chose $k$ independent random hash functions $h_1, \ldots, h_k$ mapping the universe of elements $U \rightarrow [m]$.

- Maintain an array $A$ containing $m$ bits, all initially 0.
- $insert(x)$: set all bits $A[h_1(x)] = \ldots = A[h_k(x)] := 1$.
- $query(x)$: return 1 only if $A[h_1(x)] = \ldots = A[h_k(x)] = 1$. 

Insertions: $X$ 

$h_1(x)$ $h_2(x)$ $h_3(x)$ 

m bit array $A$ 

| 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |

Queries:
Chose $k$ independent random hash functions $h_1, \ldots, h_k$ mapping the universe of elements $U \rightarrow [m]$.

- Maintain an array $A$ containing $m$ bits, all initially 0.
- $insert(x)$: set all bits $A[h_1(x)] = \ldots = A[h_k(x)] := 1$.
- $query(x)$: return 1 only if $A[h_1(x)] = \ldots = A[h_k(x)] = 1$. 

![Bloom Filter Diagram]

No false negatives. False positives more likely with more insertions.
Chooses \( k \) independent random hash functions \( h_1, \ldots, h_k \) mapping the universe of elements \( U \to [m] \).

- Maintain an array \( A \) containing \( m \) bits, all initially 0.
- \( \text{insert}(x) \): set all bits \( A[h_1(x)] = \ldots = A[h_k(x)] := 1 \).
- \( \text{query}(x) \): return 1 only if \( A[h_1(x)] = \ldots = A[h_k(x)] = 1 \).
Chose $k$ independent random hash functions $h_1, \ldots, h_k$ mapping the universe of elements $U \rightarrow [m]$.

- Maintain an array $A$ containing $m$ bits, all initially 0.
- $\text{insert}(x)$: set all bits $A[h_1(x)] = \ldots = A[h_k(x)] := 1$.
- $\text{query}(x)$: return 1 only if $A[h_1(x)] = \ldots = A[h_k(x)] = 1$.

No false negatives. False positives more likely with more insertions.
BLOOM FILTERS

Chose $k$ independent random hash functions $h_1, \ldots, h_k$ mapping the universe of elements $U \rightarrow [m]$.

- Maintain an array $A$ containing $m$ bits, all initially 0.
- $insert(x)$: set all bits $A[h_1(x)] = \ldots = A[h_k(x)] := 1$.
- $query(x)$: return 1 only if $A[h_1(x)] = \ldots = A[h_k(x)] = 1$. 

![Diagram of Bloom Filter](image)
BLOOM FILTERS

Chose $k$ independent random hash functions $h_1, \ldots, h_k$ mapping the universe of elements $U \rightarrow [m]$.

- Maintain an array $A$ containing $m$ bits, all initially 0.
- $\text{insert}(x)$: set all bits $A[h_1(x)] = \ldots = A[h_k(x)] := 1$.
- $\text{query}(x)$: return 1 only if $A[h_1(x)] = \ldots = A[h_k(x)] = 1$.

No false negatives. False positives more likely with more insertions.
Chose $k$ independent random hash functions $h_1, \ldots, h_k$ mapping the universe of elements $U \rightarrow [m]$.

- Maintain an array $A$ containing $m$ bits, all initially 0.
- $\text{insert}(x)$: set all bits $A[h_1(x)] = \ldots = A[h_k(x)] := 1$.
- $\text{query}(x)$: return 1 only if $A[h_1(x)] = \ldots = A[h_k(x)] = 1$. 

No false negatives. False positives more likely with more insertions.
Chose \( k \) independent random hash functions \( h_1, \ldots, h_k \) mapping the universe of elements \( U \rightarrow [m] \).

- Maintain an array \( A \) containing \( m \) bits, all initially 0.
- \( \text{insert}(x) \): set all bits \( A[h_1(x)] = \ldots = A[h_k(x)] := 1 \).
- \( \text{query}(x) \): return 1 only if \( A[h_1(x)] = \ldots = A[h_k(x)] = 1 \).
Bloom Filters

Chose \( k \) independent random hash functions \( h_1, \ldots, h_k \) mapping the universe of elements \( U \to [m] \).

- Maintain an array \( A \) containing \( m \) bits, all initially 0.
- \( \text{insert}(x) \): set all bits \( A[h_1(x)] = \ldots = A[h_k(x)] := 1. \)
- \( \text{query}(x) \): return 1 only if \( A[h_1(x)] = \ldots = A[h_k(x)] = 1. \)

No false negatives. False positives more likely with more insertions.
For a bloom filter with $m$ bits and $k$ hash functions, the insertion and query time is $O(k)$. 

How does the false positive rate $\delta$ depend on $m$, $k$, and the number of items inserted?

Step 1: What is the probability that after inserting $n$ elements, the $i$th bit of the array $A$ is still 0?

$n \times k$ total hashes must not hit bit $i$.

$$Pr(A[i] = 0) = Pr(h_1(x_1) \neq i \cap \ldots \cap h_k(x_k) \neq i \cap \ldots \cap h_k(x_2) \neq i \cap \ldots) = Pr(h_1(x_1) \neq i) \times \ldots \times Pr(h_k(x_1) \neq i) \times \ldots$$

$k \cdot n$ events each occurring with probability $1 - \frac{1}{m} = \left(1 - \frac{1}{m}\right)^k n^3$
For a bloom filter with $m$ bits and $k$ hash functions, the insertion and query time is $O(k)$. How does the false positive rate $\delta$ depend on $m$, $k$, and the number of items inserted?
For a bloom filter with $m$ bits and $k$ hash functions, the insertion and query time is $O(k)$. How does the false positive rate $\delta$ depend on $m$, $k$, and the number of items inserted?

**Step 1:** What is the probability that after inserting $n$ elements, the $i^{th}$ bit of the array $A$ is still 0?
For a bloom filter with $m$ bits and $k$ hash functions, the insertion and query time is $O(k)$. How does the false positive rate $\delta$ depend on $m$, $k$, and the number of items inserted?

**Step 1:** What is the probability that after inserting $n$ elements, the $i^{th}$ bit of the array $A$ is still 0? $n \times k$ total hashes must not hit bit $i$.

$$
\Pr(A[i] = 0) = \Pr\left(h_1(x_1) \neq i \cap \ldots \cap h_k(x_k) \neq i \cap \ldots \cap h_1(x_2) \neq i \ldots \cap h_k(x_2) \neq i \cap \ldots \right)
$$
For a bloom filter with $m$ bits and $k$ hash functions, the insertion and query time is $O(k)$. How does the false positive rate $\delta$ depend on $m$, $k$, and the number of items inserted?

**Step 1**: What is the probability that after inserting $n$ elements, the $i^{th}$ bit of the array $A$ is still 0? $n \times k$ total hashes must not hit bit $i$.

$$\Pr(A[i] = 0) = \Pr \left( h_1(x_1) \neq i \cap \ldots \cap h_k(x_k) \neq i \cap h_1(x_2) \neq i \ldots \cap h_k(x_2) \neq i \cap \ldots \right)$$

$$= \Pr \left( h_1(x_1) \neq i \right) \times \ldots \times \Pr \left( h_k(x_1) \neq i \right) \times \Pr \left( h_1(x_2) \neq i \right) \ldots$$

$k \cdot n$ events each occurring with probability $1 - \frac{1}{m}$
For a bloom filter with $m$ bits and $k$ hash functions, the insertion and query time is $O(k)$. How does the false positive rate $\delta$ depend on $m$, $k$, and the number of items inserted?

**Step 1:** What is the probability that after inserting $n$ elements, the $i^{th}$ bit of the array $A$ is still 0? $n \times k$ total hashes must not hit bit $i$.

$$\Pr(A[i] = 0) = \Pr(h_1(x_1) \neq i \cap \ldots \cap h_k(x_k) \neq i \cap \ldots \cap h_1(x_2) \neq i \cap \ldots \cap h_k(x_2) \neq i \cap \ldots)$$

$$= \Pr(h_1(x_1) \neq i) \times \ldots \times \Pr(h_k(x_1) \neq i) \times \Pr(h_1(x_2) \neq i) \ldots$$

$$= \left(1 - \frac{1}{m}\right)^{kn}$$
How does the false positive rate $\delta$ depend on $m$, $k$, and the number of items inserted?

What is the probability that after inserting $n$ elements, the $i^{th}$ bit of the array $A$ is still 0?

$$\Pr(A[i] = 0) = \left(1 - \frac{1}{m}\right)^{kn}$$

$n$: total number items in filter, $m$: number of bits in filter, $k$: number of random hash functions, $h_1, \ldots h_k$: hash functions, $A$: bit array, $\delta$: false positive rate.
How does the false positive rate $\delta$ depend on $m$, $k$, and the number of items inserted?

What is the probability that after inserting $n$ elements, the $i^{th}$ bit of the array $A$ is still 0?

$$\Pr(A[i] = 0) = \left(1 - \frac{1}{m}\right)^{kn} \approx e^{-\frac{kn}{m}}$$
How does the false positive rate $\delta$ depend on $m$, $k$, and the number of items inserted?

What is the probability that after inserting $n$ elements, the $i^{th}$ bit of the array $A$ is still 0?

$$\Pr(A[i] = 0) = \left(1 - \frac{1}{m}\right)^{kn} \approx e^{-\frac{kn}{m}}$$

Let $T$ be the number of zeros in the array after $n$ inserts. Then,

$$E[T] = m \left(1 - \frac{1}{m}\right)^{kn} \approx me^{-\frac{kn}{m}}$$

$n$: total number items in filter, $m$: number of bits in filter, $k$: number of random hash functions, $h_1, \ldots, h_k$: hash functions, $A$: bit array, $\delta$: false positive rate.
If \( T \) is the number of 0 entries, for a non-inserted element \( w \):

\[
\Pr(A[h_1(w)] = \ldots = A[h_k(w)] = 1)
= \Pr(A[h_1(w)] = 1) \times \ldots \times \Pr(A[h_k(w)] = 1)
= (1 - T/m) \times \ldots \times (1 - T/m)
= (1 - T/m)^k
\]
If $T$ is the number of 0 entries, for a non-inserted element $w$:

$$\Pr(A[h_1(w)] = \ldots = A[h_k(w)] = 1) = \Pr(A[h_1(w)] = 1) \times \ldots \times \Pr(A[h_k(w)] = 1) = (1 - T/m)^k$$

- How small is $T/m$? Note that $\frac{T}{m} \geq \frac{m-nk}{m} \approx e^{-kn/m}$ when $kn \ll m$. More generally, it can be shown that $T/m = \Omega \left(e^{-kn/m}\right)$ via Theorem 2 of:

  cglab.ca/~morin/publications/ds/bloom-submitted.pdf
**False Positive Rate**: with $m$ bits of storage, $k$ hash functions, and $n$ items inserted $\delta \approx \left(1 - e^{\frac{-kn}{m}}\right)^k$. 
False Positive Rate: with \( m \) bits of storage, \( k \) hash functions, and \( n \) items inserted \( \delta \approx \left(1 - e^{-kn/m}\right)^k \).
False Positive Rate: with $m$ bits of storage, $k$ hash functions, and $n$ items inserted $\delta \approx \left(1 - e^{-kn/m}\right)^k$. 
**False Positive Rate:** with $m$ bits of storage, $k$ hash functions, and $n$ items inserted $\delta \approx \left(1 - e^{-kn/m}\right)^k$. 

![Graph showing the false positive rate vs. number of hash functions]
**False Positive Rate:** with $m$ bits of storage, $k$ hash functions, and $n$ items inserted $\delta \approx \left(1 - e^{-\frac{kn}{m}}\right)^k$.

- Can differentiate to show optimal number of hashes is $k = \ln 2 \cdot \frac{m}{n}$.  

![Graph showing false positive rate vs number of hash functions](image)
False Positive Rate: with $m$ bits of storage, $k$ hash functions, and $n$ items inserted $\delta \approx \left(1 - e^{-kn/m}\right)^k$.

- Can differentiate to show optimal number of hashes is $k = \ln 2 \cdot \frac{m}{n}$.
- Balances between filling up the array with too many hashes and having enough hashes so that even when the array is pretty full, a new item is unlikely to have all its bits set (yield a false positive)
Questions on Bloom Filters?
Stream Processing: Have a massive dataset $X$ with $n$ items $x_1, x_2, \ldots, x_n$ that arrive in a continuous stream. Not nearly enough space to store all the items (in a single location).

- Still want to analyze and learn from this data.
Stream Processing: Have a massive dataset $X$ with $n$ items $x_1, x_2, \ldots, x_n$ that arrive in a continuous stream. Not nearly enough space to store all the items (in a single location).

• Still want to analyze and learn from this data.
• Typically must compress the data on the fly, storing a data structure from which you can still learn useful information.
**Stream Processing:** Have a massive dataset \( X \) with \( n \) items \( x_1, x_2, \ldots, x_n \) that arrive in a continuous stream. Not nearly enough space to store all the items (in a single location).

- Still want to analyze and learn from this data.
- Typically must compress the data on the fly, storing a data structure from which you can still learn useful information.
- Often the compression is randomized. E.g., bloom filters.
**Stream Processing:** Have a massive dataset $X$ with $n$ items $x_1, x_2, \ldots, x_n$ that arrive in a continuous stream. Not nearly enough space to store all the items (in a single location).

- Still want to analyze and learn from this data.
- Typically must compress the data on the fly, storing a data structure from which you can still learn useful information.
- Often the compression is randomized. E.g., bloom filters.
- Compared to traditional algorithm design, which focuses on minimizing *runtime*, the big question here is how much *space* is needed to answer queries of interest.
• **Sensor data:** images from telescopes (15 terabytes per night from the Large Synoptic Survey Telescope), readings from seismometer arrays monitoring and predicting earthquake activity, traffic cameras and travel time sensors (Smart Cities), electrical grid monitoring.
• **Sensor data:** images from telescopes (15 terabytes per night from the Large Synoptic Survey Telescope), readings from seismometer arrays monitoring and predicting earthquake activity, traffic cameras and travel time sensors (Smart Cities), electrical grid monitoring.
• **Sensor data:** images from telescopes (15 terabytes per night from the Large Synoptic Survey Telescope), readings from seismometer arrays monitoring and predicting earthquake activity, traffic cameras and travel time sensors (Smart Cities), electrical grid monitoring.

• **Internet Traffic:** 500 million Tweets per day, 5.6 billion Google searches, billions of ad-clicks and other logs from instrumented webpages, IPs routed by network switches, ...
• **Sensor data:** images from telescopes (15 terabytes per night from the Large Synoptic Survey Telescope), readings from seismometer arrays monitoring and predicting earthquake activity, traffic cameras and travel time sensors (Smart Cities), electrical grid monitoring.

• **Internet Traffic:** 500 million Tweets per day, 5.6 billion Google searches, billions of ad-clicks and other logs from instrumented webpages, IPs routed by network switches, ...

• **Datasets in Machine Learning:** When training e.g. a neural network on a large dataset (ImageNet with 14 million images), the data is typically processed in a stream due to storage limitations.
SOME EXAMPLES

• **Sensor data:** images from telescopes (15 terabytes per night from the Large Synoptic Survey Telescope), readings from seismometer arrays monitoring and predicting earthquake activity, traffic cameras and travel time sensors (Smart Cities), electrical grid monitoring.

• **Internet Traffic:** 500 million Tweets per day, 5.6 billion Google searches, billions of ad-clicks and other logs from instrumented webpages, IPs routed by network switches, ...

• **Datasets in Machine Learning:** When training e.g. a neural network on a large dataset (ImageNet with 14 million images), the data is typically processed in a stream due to storage limitations.
Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, output the number of distinct elements in the stream.
Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, output the number of distinct elements in the stream. E.g.,

$$1, 5, 7, 5, 2, 1 \rightarrow 4 \text{ distinct elements}$$
Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements in the stream. E.g.,

$$1, 5, 7, 5, 2, 1 \rightarrow 4 \text{ distinct elements}$$
Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements in the stream. E.g.,

$$1, 5, 7, 5, 2, 1 \rightarrow 4 \text{ distinct elements}$$

Applications:

- Distinct IP addresses clicking on an ad or visiting a site.
- Distinct values in a database column (for estimating sizes of joins and group bys).
- Number of distinct search engine queries.
- Counting distinct motifs in large DNA sequences.
Distinct Elements (Count-Distinct) Problem: Given a stream \( x_1, \ldots, x_n \), estimate the number of distinct elements in the stream. E.g.,

\[
1, 5, 7, 5, 2, 1 \rightarrow 4 \text{ distinct elements}
\]

Applications:

- Distinct IP addresses clicking on an ad or visiting a site.
- Distinct values in a database column (for estimating sizes of joins and group bys).
- Number of distinct search engine queries.
- Counting distinct motifs in large DNA sequences.

Google Sawzall, Facebook Presto, Apache Drill, Twitter Algebird
DISTINCT ELEMENTS IDEAS
**Distinct Elements (Count-Distinct) Problem:** Given a stream \( x_1, \ldots, x_n \), estimate the number of distinct elements.
Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Min-Hashing for Distinct Elements (variant of Flajolet-Martin):

- Let $h : U \rightarrow [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, h(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$
Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Min-Hashing for Distinct Elements (variant of Flajolet-Martin):

- Let $h : U \rightarrow [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, h(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$
Distinct Elements (Count-Distinct) Problem: Given a stream \( x_1, \ldots, x_n \), estimate the number of distinct elements.

Min-Hashing for Distinct Elements (variant of Flajolet-Martin):

- Let \( h : U \rightarrow [0, 1] \) be a random hash function (with a real valued output)
- \( s := 1 \)
- For \( i = 1, \ldots, n \)
  - \( s := \min(s, h(x_i)) \)
- Return \( \tilde{d} = \frac{1}{s} - 1 \)
Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Min-Hashing for Distinct Elements (variant of Flajolet-Martin):

- Let $h : U \rightarrow [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, h(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$
Distinct Elements (Count-Distinct) Problem: Given a stream \(x_1, \ldots, x_n\), estimate the number of distinct elements.

Min-Hashing for Distinct Elements (variant of Flajolet-Martin):

- Let \(h : U \rightarrow [0, 1] \) be a random hash function (with a real valued output)
- \(s := 1\)
- For \(i = 1, \ldots, n\)
  - \(s := \min(s, h(x_i))\)
- Return \(\tilde{d} = \frac{1}{s} - 1\)
Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Min-Hashing for Distinct Elements (variant of Flajolet-Martin):

- Let $h : U \to [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, h(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$
Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Min-Hashing for Distinct Elements (variant of Flajolet-Martin):

- Let $h : U \rightarrow [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, h(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$
Distinct Elements (Count-Distinct) Problem: Given a stream \( x_1, \ldots, x_n \), estimate the number of distinct elements.

Min-Hashing for Distinct Elements (variant of Flajolet-Martin):

- Let \( h : U \rightarrow [0, 1] \) be a random hash function (with a real valued output)
- \( s := 1 \)
- For \( i = 1, \ldots, n \)
  - \( s := \min(s, h(x_i)) \)
- Return \( \tilde{d} = \frac{1}{s} - 1 \)
Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Min-Hashing for Distinct Elements (variant of Flajolet-Martin):

1. Let $h : U \rightarrow [0, 1]$ be a random hash function (with a real valued output)
2. $s := 1$
3. For $i = 1, \ldots, n$
   - $s := \min(s, h(x_i))$
4. Return $\tilde{d} = \frac{1}{s} - 1$
Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Min-Hashing for Distinct Elements (variant of Flajolet-Martin):

- Let $h : U \rightarrow [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, h(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$
Min-Hashing for Distinct Elements:

- Let $h : U \rightarrow [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, h(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$
Min-Hashing for Distinct Elements:

- Let $h : U \rightarrow [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, h(x_i))$
  - Return $\tilde{d} = \frac{1}{s} - 1$

- After all items are processed, $s$ is the minimum of $d$ points chosen uniformly at random on $[0, 1]$. Where $d = \#$ distinct elements.
Min-Hashing for Distinct Elements:

- Let $h: U \rightarrow [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, h(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$

After all items are processed, $s$ is the minimum of $d$ points chosen uniformly at random on $[0, 1]$. Where $d = \#$ distinct elements.

Intuition: The larger $d$ is, the smaller we expect $s$ to be.
Min-Hashing for Distinct Elements:

- Let \( h : U \rightarrow [0, 1] \) be a random hash function (with a real valued output)
- \( s := 1 \)
- For \( i = 1, \ldots, n \)
  - \( s := \min(s, h(x_i)) \)
- Return \( \tilde{d} = \frac{1}{s} - 1 \)

After all items are processed, \( s \) is the minimum of \( d \) points chosen uniformly at random on \([0, 1]\). Where \( d = \# \) distinct elements.

Intuition: The larger \( d \) is, the smaller we expect \( s \) to be.

Same idea as Flajolet-Martin algorithm and HyperLogLog, except they use discrete hash functions.
\[ s \text{ is the minimum of } d \text{ points chosen uniformly at random on } [0, 1]. \]

Where \( d = \# \text{ distinct elements.} \)
s is the minimum of \( d \) points chosen uniformly at random on \([0, 1]\). Where \( d = \# \) distinct elements.

\[
\mathbb{E}[s] = \frac{1}{d + 1} + \int_0^\infty \text{Pr}(s > x) \, dx
\]

\( \hat{d} \) is an estimate of this expectation. Does this mean \( \mathbb{E}[\hat{d}] = d \)?

No, but:

- Approximation is robust: if \(|s - \mathbb{E}[s]| \leq \epsilon \cdot \mathbb{E}[s] \) for any \( \epsilon \in (0, \frac{1}{2}) \) and a small constant \( c \leq 4 \):
  \[
  (1 - c \epsilon) d \leq \hat{d} \leq (1 + c \epsilon) d
  \]
**PERFORMANCE IN EXPECTATION**

$s$ is the minimum of $d$ points chosen uniformly at random on $[0, 1]$. Where $d = \#$ distinct elements.

\[
\mathbb{E}[s] = \frac{1}{d + 1} \quad \text{(using } \mathbb{E}(s) = \int_0^\infty \Pr(s > x)dx + \text{calculus})}
\]
s is the minimum of d points chosen uniformly at random on [0, 1]. Where d = # distinct elements.

\[ E[s] = \frac{1}{d + 1} \] (using \( E(s) = \int_0^\infty \Pr(s > x)dx + \text{calculus} \))

- So estimate of \( \hat{d} = \frac{1}{s} - 1 \) output by the algorithm is correct if s exactly equals its expectation.
s is the minimum of d points chosen uniformly at random on [0, 1]. Where d = # distinct elements.

\[ \mathbb{E}[s] = \frac{1}{d + 1} \] (using \[ \mathbb{E}(s) = \int_{0}^{\infty} \Pr(s > x) dx \] + calculus)

- So estimate of \( \hat{d} = \frac{1}{s} - 1 \) output by the algorithm is correct if s exactly equals its expectation. Does this mean \( \mathbb{E}[\hat{d}] = d \)?
$s$ is the minimum of $d$ points chosen uniformly at random on $[0, 1]$. Where $d = \#$ distinct elements.

\[
\mathbb{E}[s] = \frac{1}{d + 1} \quad \text{(using } \mathbb{E}(s) = \int_0^\infty \Pr(s > x)dx \text{)} + \text{calculus)}
\]

- So estimate of $\hat{d} = \frac{1}{s} - 1$ output by the algorithm is correct if $s$ exactly equals its expectation. Does this mean $\mathbb{E}[\hat{d}] = d$? No, but:
s is the minimum of d points chosen uniformly at random on [0, 1]. Where \( d = \# \text{ distinct elements} \).

\[
\mathbb{E}[s] = \frac{1}{d + 1} \quad (\text{using } \mathbb{E}(s) = \int_0^{\infty} \Pr(s > x) dx + \text{calculus})
\]

- So estimate of \( \hat{d} = \frac{1}{s} - 1 \) output by the algorithm is correct if s exactly equals its expectation. **Does this mean \( \mathbb{E}[\hat{d}] = d \)?** No, but:

- **Approximation is robust**: if \( |s - \mathbb{E}[s]| \leq \epsilon \cdot \mathbb{E}[s] \) for any \( \epsilon \in (0, 1/2) \) and a small constant \( c \leq 4 \):

\[
(1 - c\epsilon)d \leq \hat{d} \leq (1 + c\epsilon)d
\]
So question is how well $s$ concentrates around its mean.

$$\mathbb{E}[s] = \frac{1}{d + 1}.$$  

---

$s$: minimum of $d$ distinct hashes chosen randomly over $[0, 1]$, computed by hashing algorithm. $\hat{d} = \frac{1}{s} - 1$: estimate of $\#$ distinct elements $d$. 
So question is how well $s$ concentrates around its mean.

$$\mathbb{E}[s] = \frac{1}{d+1} \text{ and } \text{Var}[s] \leq \frac{1}{(d+1)^2} \text{ (also via calculus).}$$

$s$: minimum of $d$ distinct hashes chosen randomly over $[0,1]$, computed by hashing algorithm. $\hat{d} = \frac{1}{s} - 1$: estimate of # distinct elements $d$. 
So question is how well \( s \) concentrates around its mean.

\[
E[s] = \frac{1}{d+1} \quad \text{and} \quad Var[s] \leq \frac{1}{(d+1)^2} \quad (\text{also via calculus}).
\]

**Chebyshev’s Inequality:**

\[
Pr \left[ |s - E[s]| \geq \epsilon E[s] \right] \leq \frac{Var[s]}{(\epsilon E[s])^2}.
\]

---

s: minimum of \( d \) distinct hashes chosen randomly over \([0, 1]\), computed by hashing algorithm. \( \hat{d} = \frac{1}{s} - 1 \): estimate of \# distinct elements \( d \).
So question is how well \( s \) concentrates around its mean.

\[
\mathbb{E}[s] = \frac{1}{d+1} \quad \text{and} \quad \text{Var}[s] \leq \frac{1}{(d+1)^2} \quad (\text{also via calculus}).
\]

**Chebyshev’s Inequality:**

\[
\Pr[|s - \mathbb{E}[s]| \geq \epsilon \mathbb{E}[s]] \leq \frac{\text{Var}[s]}{(\epsilon \mathbb{E}[s])^2} = \frac{1}{\epsilon^2}.
\]

\( s \): minimum of \( d \) distinct hashes chosen randomly over \([0, 1]\), computed by hashing algorithm. \( \hat{d} = \frac{1}{s} - 1 \): estimate of \# distinct elements \( d \).
So question is how well \( s \) concentrates around its mean.

\[
E[s] = \frac{1}{d + 1} \quad \text{and} \quad \text{Var}[s] \leq \frac{1}{(d + 1)^2} \quad (also \ via \ calculus).
\]

**Chebyshev's Inequality:**

\[
Pr \left[ |s - E[s]| \geq \epsilon E[s] \right] \leq \frac{\text{Var}[s]}{(\epsilon E[s])^2} = \frac{1}{\epsilon^2}.
\]

Bound is vacuous for any \( \epsilon < 1 \).

---

\( s \): minimum of \( d \) distinct hashes chosen randomly over \([0, 1]\), computed by hashing algorithm. \( \hat{d} = \frac{1}{s} - 1 \): estimate of \# distinct elements \( d \).
So question is how well $s$ concentrates around its mean. 

$$E[s] = \frac{1}{d + 1} \text{ and } \text{Var}[s] \leq \frac{1}{(d + 1)^2} \text{ (also via calculus).}$$

**Chebyshev’s Inequality:**

$$\Pr [ |s - E[s]| \geq \epsilon E[s] ] \leq \frac{\text{Var}[s]}{(\epsilon E[s])^2} = \frac{1}{\epsilon^2}.$$

Bound is vacuous for any $\epsilon < 1$. **How can we improve accuracy?**

$s$: minimum of $d$ distinct hashes chosen randomly over $[0, 1]$, computed by hashing algorithm. $\hat{d} = \frac{1}{s} - 1$: estimate of $\#$ distinct elements $d$. 
Leverage the law of large numbers: improve accuracy via repeated independent trials.
Leverage the law of large numbers: improve accuracy via repeated independent trials.

**Hashing for Distinct Elements (Improved):**

- Let $h : U \rightarrow [0, 1]$ be a random hash function
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, h(x_i))$
- Return $\hat{d} = \frac{1}{s} - 1$
Leverage the law of large numbers: improve accuracy via repeated independent trials.

Hashing for Distinct Elements (Improved):

- Let $h_1, h_2, \ldots, h_k : U \rightarrow [0, 1]$ be random hash functions
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, h(x_i))$
- Return $\hat{d} = \frac{1}{s} - 1$
Leverage the law of large numbers: improve accuracy via repeated independent trials.

Hashing for Distinct Elements (Improved):

- Let \( h_1, h_2, \ldots, h_k : U \rightarrow [0, 1] \) be random hash functions
- \( s_1, s_2, \ldots, s_k := 1 \)
- For \( i = 1, \ldots, n \)
  - \( s := \min(s, h(x_i)) \)
- Return \( \hat{d} = \frac{1}{s} - 1 \)
Leverage the law of large numbers: improve accuracy via repeated independent trials.

**Hashing for Distinct Elements (Improved):**

- Let $h_1, h_2, \ldots, h_k : U \rightarrow [0, 1]$ be random hash functions
- $s_1, s_2, \ldots, s_k := 1$
- For $i = 1, \ldots, n$
  - For $j=1,\ldots,k$, $s_j := \min(s_j, h_j(x_i))$
- Return $\hat{d} = \frac{1}{s} - 1$
Leverage the law of large numbers: improve accuracy via repeated independent trials.

**Hashing for Distinct Elements (Improved):**

- Let $h_1, h_2, \ldots, h_k : U \rightarrow [0, 1]$ be random hash functions
- $s_1, s_2, \ldots, s_k := 1$
- For $i = 1, \ldots, n$
  - For $j=1,\ldots,k$, $s_j := \min(s_j, h_j(x_i))$
- $s := \frac{1}{k} \sum_{j=1}^{k} s_j$
- Return $\hat{d} = \frac{1}{s} - 1$
Leverage the law of large numbers: improve accuracy via repeated independent trials.

**Hashing for Distinct Elements (Improved):**

- Let $h_1, h_2, \ldots, h_k : U \rightarrow [0, 1]$ be random hash functions
- $s_1, s_2, \ldots, s_k := 1$
- For $i = 1, \ldots, n$
  - For $j=1,\ldots,k$, $s_j := \min(s_j, h_j(x_i))$
- $s := \frac{1}{k} \sum_{j=1}^{k} s_j$
- Return $\hat{d} = \frac{1}{s} - 1$
\[ s = \frac{1}{k} \sum_{j=1}^{k} s_j. \] Have already shown that for \( j = 1, \ldots, k \):

\[ \mathbb{E}[s_j] = \frac{1}{d + 1} \]

\[ \text{Var}[s_j] \leq \frac{1}{(d + 1)^2} \]

**s\_j**: minimum of \( d \) distinct hashes chosen randomly over \([0, 1]\). \( s = \frac{1}{k} \sum_{j=1}^{k} s_j \).

\( \hat{d} = \frac{1}{s} - 1 \): estimate of \# distinct elements \( d \).
\[ s = \frac{1}{k} \sum_{j=1}^{k} s_j. \] Have already shown that for \( j = 1, \ldots, k \):

\[
\mathbb{E}[s_j] = \frac{1}{d + 1} \implies \mathbb{E}[s]
\]

\[
\text{Var}[s_j] \leq \frac{1}{(d + 1)^2}
\]

\( s_j \): minimum of \( d \) distinct hashes chosen randomly over \([0, 1]\). \( s = \frac{1}{k} \sum_{j=1}^{k} s_j \).

\( \hat{d} = \frac{1}{s} - 1 \): estimate of \# distinct elements \( d \).
\( s = \frac{1}{k} \sum_{j=1}^{k} s_j \). Have already shown that for \( j = 1, \ldots, k \):

\[
\mathbb{E}[s_j] = \frac{1}{d+1} \implies \mathbb{E}[s] = \frac{1}{d+1} \quad \text{(linearity of expectation)}
\]

\[
\text{Var}[s_j] \leq \frac{1}{(d+1)^2}
\]

- \( s_j \): minimum of \( d \) distinct hashes chosen randomly over \( [0, 1] \).
- \( s = \frac{1}{k} \sum_{j=1}^{k} s_j \).
- \( \hat{d} = \frac{1}{s} - 1 \): estimate of \# distinct elements \( d \).
$$s = \frac{1}{k} \sum_{j=1}^{k} s_j. \text{ Have already shown that for } j = 1, \ldots, k:\n$$

$$\mathbb{E}[s_j] = \frac{1}{d+1} \implies \mathbb{E}[s] = \frac{1}{d+1} \text{ (linearity of expectation)}$$

$$\text{Var}[s_j] \leq \frac{1}{(d+1)^2} \implies \text{Var}[s]$$

$s_j$: minimum of $d$ distinct hashes chosen randomly over $[0, 1]$. $s = \frac{1}{k} \sum_{j=1}^{k} s_j$.

$\hat{d} = \frac{1}{s} - 1$: estimate of \# distinct elements $d$. 
\[ s = \frac{1}{k} \sum_{j=1}^{k} s_j. \] Have already shown that for \( j = 1, \ldots, k \):

\[ \mathbb{E}[s_j] = \frac{1}{d + 1} \implies \mathbb{E}[s] = \frac{1}{d + 1} \] (linearity of expectation)

\[ \text{Var}[s_j] \leq \frac{1}{(d + 1)^2} \implies \text{Var}[s] \leq \frac{1}{k \cdot (d + 1)^2} \] (linearity of variance)

\( s_j \): minimum of \( d \) distinct hashes chosen randomly over \([0, 1]\). \( s = \frac{1}{k} \sum_{j=1}^{k} s_j \).

\( \hat{d} = \frac{1}{s} - 1 \): estimate of \# distinct elements \( d \).
\[ s = \frac{1}{k} \sum_{j=1}^{k} s_j. \] Have already shown that for \( j = 1, \ldots, k \):

\[ \mathbb{E}[s_j] = \frac{1}{d+1} \implies \mathbb{E}[s] = \frac{1}{d+1} \quad \text{(linearity of expectation)} \]

\[ \text{Var}[s_j] \leq \frac{1}{(d+1)^2} \implies \text{Var}[s] \leq \frac{1}{k \cdot (d+1)^2} \quad \text{(linearity of variance)} \]

**Chebyshev Inequality:**

\[ \Pr [ |s - \mathbb{E}[s]| \geq \epsilon \mathbb{E}[s] ] \leq \frac{\text{Var}[s]}{(\epsilon \mathbb{E}[s])^2} \]

\( s_j \): minimum of \( d \) distinct hashes chosen randomly over \([0, 1]\). \( s = \frac{1}{k} \sum_{j=1}^{k} s_j \).

\( \hat{d} = \frac{1}{s} - 1 \): estimate of \# distinct elements \( d \).
\[ s = \frac{1}{k} \sum_{j=1}^{k} s_j. \] Have already shown that for \( j = 1, \ldots, k: \]

\[ \mathbb{E}[s_j] = \frac{1}{d+1} \implies \mathbb{E}[s] = \frac{1}{d+1} \] (linearity of expectation)

\[ \text{Var}[s_j] \leq \frac{1}{(d+1)^2} \implies \text{Var}[s] \leq \frac{1}{k \cdot (d+1)^2} \] (linearity of variance)

**Chebyshev Inequality:**

\[
\Pr \left[ |s - \mathbb{E}[s]| \geq \epsilon \mathbb{E}[s] \right] \leq \frac{\text{Var}[s]}{(\epsilon \mathbb{E}[s])^2} = \frac{\mathbb{E}[s]^2/k}{\epsilon^2 \mathbb{E}[s]^2} = \frac{1}{k \cdot \epsilon^2}
\]

**s_j:** minimum of \( d \) distinct hashes chosen randomly over \([0, 1]\). \( s = \frac{1}{k} \sum_{j=1}^{k} s_j \).

\( \hat{d} = \frac{1}{s} - 1: \) estimate of \# distinct elements \( d \).
\( s = \frac{1}{k} \sum_{j=1}^{k} s_j \). Have already shown that for \( j = 1, \ldots, k \):

\[
\mathbb{E}[s_j] = \frac{1}{d+1} \implies \mathbb{E}[s] = \frac{1}{d+1} \quad \text{(linearity of expectation)}
\]

\[
\text{Var}[s_j] \leq \frac{1}{(d+1)^2} \implies \text{Var}[s] \leq \frac{1}{k \cdot (d+1)^2} \quad \text{(linearity of variance)}
\]

**Chebyshev Inequality:**

\[
\Pr \left[ \left| d - \hat{d} \right| \geq 4\epsilon \cdot d \right] \leq \frac{\text{Var}[s]}{(\epsilon \mathbb{E}[s])^2} = \frac{\mathbb{E}[s]^2/k}{\epsilon^2 \mathbb{E}[s]^2} = \frac{1}{k \cdot \epsilon^2}
\]

---

\( s_j \): minimum of \( d \) distinct hashes chosen randomly over \([0, 1]\). \( s = \frac{1}{k} \sum_{j=1}^{k} s_j \).

\( \hat{d} = \frac{1}{s} - 1 \): estimate of \# distinct elements \( d \).
\( s = \frac{1}{k} \sum_{j=1}^{k} s_j \). Have already shown that for \( j = 1, \ldots, k \):

\[
\mathbb{E}[s_j] = \frac{1}{d+1} \implies \mathbb{E}[s] = \frac{1}{d+1} \quad \text{(linearity of expectation)}
\]

\[
\text{Var}[s_j] \leq \frac{1}{(d+1)^2} \implies \text{Var}[s] \leq \frac{1}{k \cdot (d+1)^2} \quad \text{(linearity of variance)}
\]

**Chebyshev Inequality:**

\[
\Pr \left[ \left| d - \hat{d} \right| \geq 4\epsilon \cdot d \right] \leq \frac{\text{Var}[s]}{(\epsilon \mathbb{E}[s])^2} = \frac{\mathbb{E}[s]^2/k}{\epsilon^2 \mathbb{E}[s]^2} = \frac{1}{k \cdot \epsilon^2}
\]

How should we set \( k \) if we want \( 4\epsilon \cdot d \) error with probability \( \geq 1 - \delta \)?

\( s_j \): minimum of \( d \) distinct hashes chosen randomly over \([0, 1]\). \( s = \frac{1}{k} \sum_{j=1}^{k} s_j \).

\( \hat{d} = \frac{1}{s} - 1 \): estimate of \# distinct elements \( d \).
\( s = \frac{1}{k} \sum_{j=1}^{k} s_j \). Have already shown that for \( j = 1, \ldots, k \):

\[
\mathbb{E}[s_j] = \frac{1}{d+1} \implies \mathbb{E}[s] = \frac{1}{d+1} \quad \text{(linearity of expectation)}
\]

\[
\text{Var}[s_j] \leq \frac{1}{(d+1)^2} \implies \text{Var}[s] \leq \frac{1}{k \cdot (d+1)^2} \quad \text{(linearity of variance)}
\]

**Chebyshev Inequality:**

\[
\Pr \left[ \left| d - \hat{d} \right| \geq 4\epsilon \cdot d \right] \leq \frac{\text{Var}[s]}{(\epsilon \mathbb{E}[s])^2} = \frac{\mathbb{E}[s]^2/k}{\epsilon^2 \mathbb{E}[s]^2} = \frac{1}{k \cdot \epsilon^2}
\]

How should we set \( k \) if we want \( 4\epsilon \cdot d \) error with probability \( \geq 1 - \delta \)?

\[
k = \frac{1}{\epsilon^2 \cdot \delta}.
\]

\( s_j \): minimum of \( d \) distinct hashes chosen randomly over \([0,1]\). \( s = \frac{1}{k} \sum_{j=1}^{k} s_j \).

\( \hat{d} = \frac{1}{s} - 1 \): estimate of \# distinct elements \( d \).
\( s = \frac{1}{k} \sum_{j=1}^{k} s_j \). Have already shown that for \( j = 1, \ldots, k \):

\[
E[s_j] = \frac{1}{d+1} \implies E[s] = \frac{1}{d+1} \quad \text{(linearity of expectation)}
\]

\[
\text{Var}[s_j] \leq \frac{1}{(d+1)^2} \implies \text{Var}[s] \leq \frac{1}{k \cdot (d+1)^2} \quad \text{(linearity of variance)}
\]

**Chebyshev Inequality:**

\[
\Pr \left[ \left| d - \hat{d} \right| \geq 4\epsilon \cdot d \right] \leq \frac{\text{Var}[s]}{(\epsilon E[s])^2} = \frac{E[s]^2/k}{\epsilon^2 E[s]^2} = \frac{1}{k \cdot \epsilon^2} = \frac{\epsilon^2 \cdot \delta}{\epsilon^2} = \delta.
\]

How should we set \( k \) if we want \( 4\epsilon \cdot d \) error with probability \( \geq 1 - \delta \)?

\[
k = \frac{1}{\epsilon^2 \cdot \delta}.
\]

\( s_j \): minimum of \( d \) distinct hashes chosen randomly over \([0, 1] \). \( s = \frac{1}{k} \sum_{j=1}^{k} s_j \).

\( \hat{d} = \frac{1}{s} - 1 \): estimate of \# distinct elements \( d \).
SPACE COMPLEXITY

Hashing for Distinct Elements:

- Let \( h_1, h_2, \ldots, h_k : U \rightarrow [0, 1] \) be random hash functions
- \( s_1, s_2, \ldots, s_k := 1 \)
- For \( i = 1, \ldots, n \)
  - For \( j = 1, \ldots, k \), \( s_j := \min(s_j, h_j(x_i)) \)
- \( s := \frac{1}{k} \sum_{j=1}^{k} s_j \)
- Return \( \hat{d} = \frac{1}{s} - 1 \)

- Setting \( k = \frac{1}{\epsilon^2 \delta} \), algorithm returns \( \hat{d} \) with \( |d - \hat{d}| \leq 4\epsilon \cdot d \) with probability at least \( 1 - \delta \).
Space Complexity

Hashing for Distinct Elements:

- Let $h_1, h_2, \ldots, h_k : U \rightarrow [0, 1]$ be random hash functions
- $s_1, s_2, \ldots, s_k := 1$
- For $i = 1, \ldots, n$
  - For $j=1, \ldots, k$, $s_j := \min(s_j, h_j(x_i))$
- $s := \frac{1}{k} \sum_{j=1}^{k} s_j$
- Return $\hat{d} = \frac{1}{s} - 1$

- Setting $k = \frac{1}{\epsilon^2 \delta}$, algorithm returns $\hat{d}$ with $|d - \hat{d}| \leq 4\epsilon \cdot d$ with probability at least $1 - \delta$.
- Space complexity is $k = \frac{1}{\epsilon^2 \delta}$ real numbers $s_1, \ldots, s_k$. 
Hashing for Distinct Elements:

- Let $h_1, h_2, \ldots, h_k : U \rightarrow [0, 1]$ be random hash functions.
- $s_1, s_2, \ldots, s_k := 1$
- For $i = 1, \ldots, n$
  - For $j=1, \ldots, k$, $s_j := \min(s_j, h_j(x_i))$
- $s := \frac{1}{k} \sum_{j=1}^{k} s_j$
- Return $\hat{d} = \frac{1}{s} - 1$

- Setting $k = \frac{1}{\epsilon^2 \delta}$, algorithm returns $\hat{d}$ with $|d - \hat{d}| \leq 4\epsilon \cdot d$ with probability at least $1 - \delta$.
- Space complexity is $k = \frac{1}{\epsilon^2 \delta}$ real numbers $s_1, \ldots, s_k$.
- $\delta = 5\%$ failure rate gives a factor 20 overhead in space complexity.
How can we improve our dependence on the failure rate $\delta$?
How can we improve our dependence on the failure rate $\delta$?

**The median trick:** Run $t = O(\log 1/\delta)$ trials each with failure probability $\delta' = 1/4$ – each using $k = \frac{1}{\delta' \epsilon^2} = \frac{4}{\epsilon^2}$ hash functions.
How can we improve our dependence on the failure rate $\delta$?

**The median trick:** Run $t = O(\log 1/\delta)$ trials each with failure probability $\delta' = 1/4$ – each using $k = \frac{1}{\delta' \epsilon^2} = \frac{4}{\epsilon^2}$ hash functions.

- Letting $\hat{d}_1, \ldots, \hat{d}_t$ be the outcomes of the $t$ trials, return $\hat{d} = \text{median}(\hat{d}_1, \ldots, \hat{d}_t)$. 
How can we improve our dependence on the failure rate $\delta$?

**The median trick:** Run $t = O(\log 1/\delta)$ trials each with failure probability $\delta' = 1/4$ – each using $k = \frac{1}{\delta' \epsilon^2} = \frac{4}{\epsilon^2}$ hash functions.

- Letting $\hat{d}_1, \ldots, \hat{d}_t$ be the outcomes of the $t$ trials, return $\hat{d} = \text{median}(\hat{d}_1, \ldots, \hat{d}_t)$. 

![Diagram showing the median trick with a line segment from $(1 - 4\epsilon)d$ to $(1 + 4\epsilon)d$ and a median point $\hat{d}$]
How can we improve our dependence on the failure rate $\delta$?

**The median trick:** Run $t = O(\log 1/\delta)$ trials each with failure probability $\delta' = 1/4$ — each using $k = \frac{1}{\delta' \epsilon^2} = \frac{4}{\epsilon^2}$ hash functions.

- Letting $\hat{d}_1, \ldots, \hat{d}_t$ be the outcomes of the $t$ trials, return $\hat{d} = \text{median}(\hat{d}_1, \ldots, \hat{d}_t)$.

- If $> 1/2$ of trials fall in $[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$, then the median will.
\begin{itemize}
\item $\hat{d}_1, \ldots, \hat{d}_t$ are the outcomes of the $t$ trials, each falling in $[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$ with probability at least 3/4. Let $\hat{d} = median(\hat{d}_1, \ldots, \hat{d}_t)$.

What is the probability that the median $\hat{d}$ falls in $[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$?
\end{itemize}
• \( \hat{d}_1, \ldots, \hat{d}_t \) are the outcomes of the \( t \) trials, each falling in 

\[
[(1 - 4\epsilon)d, (1 + 4\epsilon)d]
\]

with probability at least 3/4. Let \( \hat{d} = \text{median}(\hat{d}_1, \ldots, \hat{d}_t) \).

What is the probability that the median \( \hat{d} \) falls in \( [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \)?

• Let \( X \) be the \# of trials falling in \( [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \).
• $\hat{d}_1, \ldots, \hat{d}_t$ are the outcomes of the $t$ trials, each falling in 

$$[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$$

with probability at least $3/4$. Let $\hat{d} = median(\hat{d}_1, \ldots, \hat{d}_t)$.

What is the probability that the median $\hat{d}$ falls in $[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$?

• Let $X$ be the number of trials falling in $[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$. 

$$Pr\left(\hat{d} \notin [(1 - 4\epsilon)d, (1 + 4\epsilon)d]\right) \leq Pr\left(X < \frac{1}{2} \cdot t\right)$$
THE MEDIAN TRICK

• \( \hat{d}_1, \ldots, \hat{d}_t \) are the outcomes of the \( t \) trials, each falling in
  \[
  [(1 - 4\epsilon)d, (1 + 4\epsilon)d]
  \]
  with probability at least \( \frac{3}{4} \). Let \( \hat{d} = \text{median}(\hat{d}_1, \ldots, \hat{d}_t) \).

What is the probability that the median \( \hat{d} \) falls in \( [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \)?

• Let \( X \) be the \# of trials falling in \( [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \). \( \mathbb{E}[X] \geq \).

\[
\Pr \left( \hat{d} \notin [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \right) \leq \Pr \left( X < \frac{1}{2} \cdot t \right)
\]
THE MEDIAN TRICK

• $\hat{d}_1, \ldots, \hat{d}_t$ are the outcomes of the $t$ trials, each falling in

$$[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$$

with probability at least $3/4$. Let $\hat{d} = \text{median}(\hat{d}_1, \ldots, \hat{d}_t)$.

What is the probability that the median $\hat{d}$ falls in $[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$?

• Let $X$ be the # of trials falling in $[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$. $\mathbb{E}[X] \geq \frac{3}{4} \cdot t$.

$$\Pr\left(\hat{d} \not\in [(1 - 4\epsilon)d, (1 + 4\epsilon)d]\right) \leq \Pr\left(X < \frac{1}{2} \cdot t\right)$$
THE MEDIAN TRICK

• \( \hat{d}_1, \ldots, \hat{d}_t \) are the outcomes of the \( t \) trials, each falling in
  
  \[ [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \]

  with probability at least \( 3/4 \). Let \( \hat{d} = \text{median}(\hat{d}_1, \ldots, \hat{d}_t) \).

  What is the probability that the median \( \hat{d} \) falls in \( [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \)?

• Let \( X \) be the \# of trials falling in \( [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \). \( \mathbb{E}[X] \geq \frac{3}{4} \cdot t \).

\[
\Pr \left( \hat{d} \notin [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \right) \leq \Pr \left( X < \frac{1}{2} \cdot t \right)
\]
• \( \hat{d}_1, \ldots, \hat{d}_t \) are the outcomes of the \( t \) trials, each falling in 
\[ [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \]
with probability at least \( 3/4 \). Let \( \hat{d} = \text{median}(\hat{d}_1, \ldots, \hat{d}_t) \).

What is the probability that the median \( \hat{d} \) falls in \( [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \) ?

• Let \( X \) be the \# of trials falling in \( [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \). \( \mathbb{E}[X] \geq \frac{3}{4} \cdot t \).

\[
\Pr \left( \hat{d} \notin [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \right) \leq \Pr \left( X < \frac{1}{2} \cdot t \right) \leq \Pr \left( |X - \mathbb{E}[X]| \geq \frac{1}{4} t \right)
\]
THE MEDIAN TRICK

• $\hat{d}_1, \ldots, \hat{d}_t$ are the outcomes of the $t$ trials, each falling in $[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$ with probability at least $3/4$. Let $\hat{d} = \text{median}(\hat{d}_1, \ldots, \hat{d}_t)$.

What is the probability that the median $\hat{d}$ falls in $[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$?

• Let $X$ be the # of trials falling in $[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$. $\mathbb{E}[X] \geq \frac{3}{4} \cdot t$.

$$\Pr \left( \hat{d} \notin [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \right) \leq \Pr \left( X < \frac{1}{2} \cdot t \right) \leq \Pr \left( |X - \mathbb{E}[X]| \geq \frac{1}{4} t \right)$$

Apply Chernoff bound:
THE MEDIAN TRICK

• $\hat{d}_1, \ldots, \hat{d}_t$ are the outcomes of the $t$ trials, each falling in 

$$[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$$

with probability at least $3/4$. Let $\hat{d} = \text{median}(\hat{d}_1, \ldots, \hat{d}_t)$.

What is the probability that the median $\hat{d}$ falls in $[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$?

• Let $X$ be the # of trials falling in $[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$. $\mathbb{E}[X] \geq \frac{3}{4} \cdot t$.

$$\Pr \left( \hat{d} \notin [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \right) \leq \Pr \left( X < \frac{1}{2} \cdot t \right) \leq \Pr \left( |X - \mathbb{E}[X]| \geq \frac{1}{4} t \right)$$

Apply Chernoff bound:

$$\Pr \left( |X - \mathbb{E}[X]| \geq \frac{1}{3} \mathbb{E}[X] \right) \leq 2 \exp \left( -\frac{\frac{1}{3} \cdot \frac{3}{4} t}{2 + 1/3} \right) = O \left( e^{-\Theta(t)} \right).$$
THE MEDIAN TRICK

• $\hat{d}_1, \ldots, \hat{d}_t$ are the outcomes of the $t$ trials, each falling in 
  $$[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$$
  with probability at least $3/4$. Let $\hat{d} = \text{median}(\hat{d}_1, \ldots, \hat{d}_t)$.

What is the probability that the median $\hat{d}$ falls in $[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$?

• Let $X$ be the $\#$ of trials falling in $[(1 - 4\epsilon)d, (1 + 4\epsilon)d]$. $\mathbb{E}[X] \geq \frac{3}{4} \cdot t$.

$$\Pr\left(\hat{d} \notin [(1 - 4\epsilon)d, (1 + 4\epsilon)d]\right) \leq \Pr\left(X < \frac{1}{2} \cdot t\right) \leq \Pr\left(|X - \mathbb{E}[X]| \geq \frac{1}{4}t\right)$$

Apply Chernoff bound:

$$\Pr\left(|X - \mathbb{E}[X]| \geq \frac{1}{3} \mathbb{E}[X]\right) \leq 2 \exp\left(-\frac{\left(\frac{1}{3} \cdot \frac{3}{4} t\right)^2}{2 + 1/3}\right) = O\left(e^{-O(t)}\right).$$

• Setting $t = O(\log(1/\delta))$ gives failure probability $e^{-\log(1/\delta)} = \delta$. 

20
**Upshot:** The median of $t = O(\log(1/\delta))$ independent runs of the hashing algorithm for distinct elements returns $\hat{d} \in [(1 - 4\epsilon)d, (1 + 4\epsilon)d]$ with probability at least $1 - \delta$. 
Upshot: The median of $t = O(\log(1/\delta))$ independent runs of the hashing algorithm for distinct elements returns $\hat{d} \in [(1 - 4\epsilon)d, (1 + 4\epsilon)d]$ with probability at least $1 - \delta$.

Total Space Complexity: $t$ trials, each using $k = \frac{1}{\epsilon^2 \delta'}$ hash functions, for $\delta' = 1/4$. Space is $\frac{4t}{\epsilon^2} = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ real numbers (the minimum value of each hash function).
Upshot: The median of \( t = O(\log(1/\delta)) \) independent runs of the hashing algorithm for distinct elements returns \( \hat{d} \in [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \) with probability at least \( 1 - \delta \).

Total Space Complexity: \( t \) trials, each using \( k = \frac{1}{\epsilon^2 \delta'} \) hash functions, for \( \delta' = 1/4 \). Space is \( \frac{4t}{\epsilon^2} = O \left( \frac{\log(1/\delta)}{\epsilon^2} \right) \) real numbers (the minimum value of each hash function).

No dependence on the number of distinct elements \( d \) or the number of items in the stream \( n \)! Both of these numbers are typically very large.
**Upshot:** The median of \( t = O(\log(1/\delta)) \) independent runs of the hashing algorithm for distinct elements returns \( \hat{d} \in [(1 - 4\epsilon)d, (1 + 4\epsilon)d] \) with probability at least 1 - \( \delta \).

**Total Space Complexity:** \( t \) trials, each using \( k = \frac{1}{\epsilon^2 \delta'} \) hash functions, for \( \delta' = 1/4 \). Space is \( \frac{4t}{\epsilon^2} = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right) \) real numbers (the minimum value of each hash function).

No dependence on the number of distinct elements \( d \) or the number of items in the stream \( n! \). Both of these numbers are typically very large.

**A note on the median:** The median is often used as a robust alternative to the mean, when there are outliers (e.g., heavy tailed distributions, corrupted data).