Today

Streaming.
Today

Streaming.
Frequency Moments.
Streaming.

Input:
Streaming.

Input:

\[ x_1, \ldots, x_n. \]
Streaming.

Input:

\(x_1, x_2,\)
Streaming.

Input:

\[ x_1, x_2, x_3, \ldots, x_n, \ldots \]

Actually, no. \((\log c_n)\) space. Model LARGE data small space. Extreme mismatch.
Streaming.

Input:

\(x_1, x_2, x_3, \ldots,\)
Streaming.

Input:

\[ x_1, x_2, x_3, \ldots, x_n. \]
Streaming.

Input:

$x_1, x_2, x_3, \ldots, x_n$.

One at a time.
Streaming.

Input:

\( x_1, x_2, x_3, \ldots, x_n \).

One at a time.

Pikachu,
Streaming.

Input:

$x_1, x_2, x_3, \ldots, x_n$.

One at a time.

Pikachu, Squirtle,
Input:

\(x_1, x_2, x_3, \ldots, x_n\).

One at a time.

Pikachu, Squirtle, Mew,
Streaming.

Input:
\[ x_1, x_2, x_3, \ldots, x_n. \]

One at a time.
Pikachu, Squirtle, Mew, Pikachu,
Streaming.

Input:
\[ x_1, x_2, x_3, \ldots, x_n. \]

One at a time.

Pikachu, Squirtle, Mew, Pikachu, …
Streaming.

Input:

\(x_1, x_2, x_3, \ldots, x_n\).

One at a time.

Pikachu, Squirtle, Mew, Pikachu, ...

Got to get ’em all!
Input:

\[ x_1, x_2, x_3, \ldots, x_n. \]

One at a time.

Pikachu, Squirtle, Mew, Pikachu, …

Got to get ’em all!

Actually, no.
Streaming.

Input:
\[ x_1, x_2, x_3, \ldots, x_n. \]
One at a time.
Pikachu, Squirtle, Mew, Pikachu, \ldots
Got to get ’em all!
Actually, no. \( O(\log^c n) \) space.
Streaming.

Input:

$x_1, x_2, x_3, \ldots, x_n$.

One at a time.

Pikachu, Squirtle, Mew, Pikachu, …

Got to get 'em all!

Actually, no. $O(\log^c n)$ space.

Model LARGE data
Streaming.

Input:

$x_1, x_2, x_3, \ldots, x_n$.

One at a time.

Pikachu, Squirtle, Mew, Pikachu, …

Got to get ’em all!

Actually, no. $O(\log^c n)$ space.

Model LARGE data small space.
Streaming.

Input:
$x_1, x_2, x_3, \ldots, x_n$.

One at a time.
Pikachu, Squirtle, Mew, Pikachu, …

Got to get ’em all!

Actually, no. $O(\log^c n)$ space.

Model LARGE data small space.

Extreme mismatch.
What to compute.

Data.

Moments!

\[ F_k = \sum_i m_k_i \]

- number of items of type \( i \).

E.g., number of Pikachus, Squirtles, ...

\[ F_0 \] Number of distinct elements.

How to compute?

\[ F_1 \] Length of stream.

Easy to compute!

\[ F_2 \] How to compute?
What to compute.

Data.
Moments!

\[ F_k = \sum_i m_k \cdot m_i \]

- number of items of type \( i \).

E.g., number of Pikachus, Squirtles, ...

\[ F_0 \]: Number of distinct elements.

\[ F_1 \]: Length of stream. Easy to compute!

\[ F_2 \]: How to compute?
What to compute.

Data.

Moments!

$F_k$
What to compute.

Data.

Moments!

\[ F_k = \sum_i m_i^k \]
What to compute.

Data.

Moments!

\[ F_k = \sum_i m_i^k \]

\( m_i \) - number of items of type \( i \).
What to compute.

Data.

Moments!

\[ F_k = \sum_i m_i^k \]

\( m_i \) - number of items of type \( i \).

E.g., number of Pikachus, Squirtles, …
What to compute.

Data.

Moments!

\[ F_k = \sum_i m_i^k \]

- \( m_i \) - number of items of type \( i \).

E.g., number of Pikachus, Squirtles, ...  

\( F_0 \):
What to compute.

Data.

Moments!

\[ F_k = \sum_i m_i^k \]

\( m_i \) - number of items of type \( i \).

E.g., number of Pikachus, Squirtles, …

\( F_0 \): Number of distinct elements.
What to compute.

Data.

Moments!

\[ F_k = \sum_i m_i^k \]

\( m_i \) - number of items of type \( i \).

E.g., number of Pikachus, Squirtles, …

\( F_0 \): Number of distinct elements.

How to compute?
What to compute.

Data.

Moments!

\[ F_k = \sum_i m_i^k \]

\( m_i \) - number of items of type \( i \).

E.g., number of Pikachus, Squirtles, …

\( F_0 \): Number of distinct elements.

  How to compute?

\( F_1 \):
What to compute.

Data.

Moments!

\[ F_k = \sum_i m_i^k \]

\( m_i \) - number of items of type \( i \).

E.g., number of Pikachus, Squirtles, ... \[ F_0 \]: Number of distinct elements.

How to compute?

\[ F_1 \]: Length of stream.
What to compute.

Data.

Moments!

\[ F_k = \sum_i m_i^k \]

- \( m_i \) - number of items of type \( i \).

E.g., number of Pikachus, Squirtles, …

\( F_0 \): Number of distinct elements.

- How to compute?

\( F_1 \): Length of stream.

- Easy to compute!
What to compute.

Data.

Moments!

\[ F_k = \sum_i m_i^k \]

\( m_i \) - number of items of type \( i \).

E.g., number of Pikachus, Squirtles, …

\( F_0 \): Number of distinct elements.

How to compute?

\( F_1 \): Length of stream.

Easy to compute!
What to compute.

Data.

Moments!

\[ F_k = \sum_i m_i^k \]

\( m_i \) - number of items of type \( i \).

E.g., number of Pikachus, Squirtles, ... 

\( F_0 \): Number of distinct elements.

How to compute?

\( F_1 \): Length of stream.

Easy to compute!

\( F_2 \): How to compute?
Number Distinct Elements

Claim: takes $\Omega(n)$ space for exact number of distinct items!
Number Distinct Elements

Claim: takes $\Omega(n)$ space for exact number of distinct items!

Pikachu, Squirtle, Mew, Squirtle, Pikachu, Squirtle

How many distinct elements?

Answer: 3.

$\Omega(n)$ time.

Algorithm A takes stream $S$ maintains number of distinct elements.

Is $x \in S$?

Add $x$, see if number of distinct elements change.

Must know subset of $\left[ n \right]$ (at most $n$ types) $2^n$ possibilities $\rightarrow$ requires $\Omega(n)$ bits!
Number Distinct Elements

Claim: takes $\Omega(n)$ space for exact number of distinct items!
Pikachu, Squirtle, Mew, Squirtle, Pikachu, Squirtle
How many distinct elements?
Number Distinct Elements

Claim: takes $\Omega(n)$ space for exact number of distinct items!

Pikachu, Squirtle, Mew, Squirtle, Pikachu, Squirtle

How many distinct elements?

Answer: 3.
Claim: takes \( \Omega(n) \) space for exact number of distinct items!

Pikachu, Squirtle, Mew, Squirtle, Pikachu, Squirtle

How many distinct elements?

Answer: 3.

See!
Claim: takes $\Omega(n)$ space for exact number of distinct items!
Pikachu, Squirtle, Mew, Squirtle, Pikachu, Squirtle
How many distinct elements?
Answer: 3.
See! $\Omega(n)$ time.
Number Distinct Elements

Claim: takes $\Omega(n)$ space for exact number of distinct items!

Pikachu, Squirtle, Mew, Squirtle, Pikachu, Squirtle

How many distinct elements?

Answer: 3.

See! $\Omega(n)$ time.

Algorithm A takes stream $S$
   maintains number of distinct elements.
Number Distinct Elements

Claim: takes $\Omega(n)$ space for exact number of distinct items!
Pikachu, Squirtle, Mew, Squirtle, Pikachu, Squirtle
How many distinct elements?
Answer: 3.
See! $\Omega(n)$ time.
Algorithm A takes stream $S$
   maintains number of distinct elements.
Is $x \in S$?
Number Distinct Elements

Claim: takes $\Omega(n)$ space for exact number of distinct items!

Pikachu, Squirtle, Mew, Squirtle, Pikachu, Squirtle

How many distinct elements?

Answer: 3.

See! $\Omega(n)$ time.

Algorithm A takes stream $S$
   maintains number of distinct elements.
   Is $x \in S$?
   Add $x$, see if number of distinct elements change.
Number Distinct Elements

Claim: takes $\Omega(n)$ space for exact number of distinct items!

Pikachu, Squirtle, Mew, Squirtle, Pikachu, Squirtle

How many distinct elements?

Answer: 3.

See! $\Omega(n)$ time.

Algorithm A takes stream $S$
  maintains number of distinct elements.
  Is $x \in S$?
  Add $x$, see if number of distinct elements change.
  Must know subset of $[n]$
Number Distinct Elements

Claim: takes $\Omega(n)$ space for exact number of distinct items!
Pikachu, Squirtle, Mew, Squirtle, Pikachu, Squirtle
How many distinct elements?
Answer: 3.
See! $\Omega(n)$ time.

Algorithm A takes stream $S$
  maintains number of distinct elements.
  Is $x \in S$?
  Add $x$, see if number of distinct elements change.
  Must know subset of $[n]$
  (at most $n$ types)
Number Distinct Elements

Claim: takes $\Omega(n)$ space for exact number of distinct items!

Pikachu, Squirtle, Mew, Squirtle, Pikachu, Squirtle

How many distinct elements?

Answer: 3.

See! $\Omega(n)$ time.

Algorithm A takes stream $S$
  maintains number of distinct elements.
  Is $x \in S$?
  Add $x$, see if number of distinct elements change.
  Must know subset of $[n]$
  (at most $n$ types)

$2^n$ possibilities
Number Distinct Elements

Claim: takes $\Omega(n)$ space for exact number of distinct items!

Pikachu, Squirtle, Mew, Squirtle, Pikachu, Squirtle

How many distinct elements?

Answer: 3.

See! $\Omega(n)$ time.

Algorithm A takes stream $S$
  maintains number of distinct elements.
  Is $x \in S$?
  Add $x$, see if number of distinct elements change.
  Must know subset of $[n]$
  (at most $n$ types)

$2^n$ possibilities $\rightarrow$ requires $\Omega(n)$ bits!
Toy problem

Alg: Number of distinct elements

"no"

"yes"

Don't care if in between.

Randomized Algorithm:

1. Choose random hash function. $h: [n] \rightarrow [B]$, where $B = k$.
2. If any $h(x) = 0$, say "yes", else "no".

$$\Pr[A(x) = \text{No} | N \leq k] = \left(1 - \frac{1}{B}\right)^N$$

$$\Pr[A(x) = \text{No} | N > 2k] = \left(1 - \frac{1}{B}\right)^{N \leq 2k}$$

Constant gap (roughly $1/e - 1/e^2$).

Many trials, in parallel gives good result.

More later.

Number of bits for random hash function? $k \cdot \log k$ bits to specify!
Toy problem

Alg: Number of distinct elements
≤ k
Toy problem

Alg: Number of distinct elements
\[ \leq k \] Output: “no”
Toy problem

Alg: Number of distinct elements

$\leq k$ Output: “no”

$\geq 2k$
Toy problem

Alg: Number of distinct elements

\( \leq k \) Output: “no”
\( \geq 2k \) Output: “yes”
Toy problem

Alg: Number of distinct elements
  $\leq k$ Output: “no”
  $\geq 2k$ Output: “yes”

Don’t care if in between.
Toy problem

Alg: Number of distinct elements
\[ \leq k \] Output: “no”
\[ \geq 2k \] Output: “yes”

Don’t care if in between.

Randomized Algorithm:
Toy problem

Alg: Number of distinct elements

\( \leq k \) Output: “no”

\( \geq 2k \) Output: “yes”

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.
Toy problem

Alg: Number of distinct elements

$\leq k$ Output: “no”
$\geq 2k$ Output: “yes”

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.
$h : [n] \rightarrow [B]$, where $B = k$. 
Toy problem

Alg: Number of distinct elements

≤ k Output: “no”
≥ 2k Output: “yes”

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.

\[ h : [n] \rightarrow [B], \text{ where } B = k. \]

(2) If any \( h(x_i) = 0 \), say “yes”, else “no”.

\[ \Pr[A(x) = \text{No} | N \leq k] = (1 - \frac{1}{2^{B}})^N \]
\[ \Pr[A(x) = \text{No} | N > 2k] = (1 - \frac{1}{2^{B}})^{N \leq 2k} \]

Constant gap (roughly \( \frac{1}{e} - \frac{1}{e^2} \)).

Many trials, in parallel gives good result.

..more later.

Number of bits for random hash function?

\( k\log k \) bits to specify!
Toy problem

Alg: Number of distinct elements
\[ \leq k \] Output: “no”
\[ \geq 2k \] Output: “yes”

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.
\[ h : [n] \to [B] \], where \( B = k \).
(2) If any \( h(x_i) = 0 \), say “yes”, else “no”.

Pr \([A(x) = \text{No} | N \leq k]\) = \((1 - 1/B)^N\)
Pr \([A(x) = \text{No} | N > 2k]\) = \((1 - 1/B)^{2k}\)

Constant gap (roughly 1/e - 1/e^2).

Many trials, in parallel gives good result.

Number of bits for random hash function?
\[ k \cdot \log k \] bits to specify!
Toy problem

Alg: Number of distinct elements
\( \leq k \) Output: “no”
\( \geq 2k \) Output: “yes”

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.
\( h : [n] \rightarrow [B] \), where \( B = k \).
(2) If any \( h(x_i) = 0 \), say “yes”, else “no”.

\[
\Pr[A(x) = \text{No} \mid N \leq k] = \left( 1 - \frac{1}{B} \right)^N \geq \left( 1 - \frac{1}{B} \right)^k
\]
Toy problem

Alg: Number of distinct elements

$\leq k$ Output: “no”
$\geq 2k$ Output: “yes”

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.

$h : [n] \rightarrow [B]$, where $B = k$.

(2) If any $h(x_i) = 0$, say “yes”, else “no”.

$$\Pr[A(x) = No \mid N \leq k] = \left(1 - \frac{1}{B}\right)^N \geq \left(1 - \frac{1}{B}\right)^k$$

$$\Pr[A(x) = No \mid N > 2k] = \left(1 - \frac{1}{B}\right)^N \leq \left(1 - \frac{1}{B}\right)^{2k}$$
Toy problem

Alg: Number of distinct elements

\( \leq k \) Output: “no”

\( \geq 2k \) Output: “yes”

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.

\( h : [n] \rightarrow [B], \text{ where } B = k. \)

(2) If any \( h(x_i) = 0 \), say “yes”, else “no”.

\[
\Pr[A(x) = No \mid N \leq k] = \left(1 - \frac{1}{B}\right)^N 
\geq \left(1 - \frac{1}{B}\right)^k 
\]

\[
\Pr[A(x) = No \mid N > 2k] = \left(1 - \frac{1}{B}\right)^N 
\leq \left(1 - \frac{1}{B}\right)^{2k} 
\]

Constant gap
Toy problem

Alg: Number of distinct elements

\[ \leq k \text{ Output: “no”} \]
\[ \geq 2k \text{ Output: “yes”} \]

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.
\[ h : [n] \rightarrow [B], \text{ where } B = k. \]
(2) If any \( h(x_i) = 0 \), say “yes”, else “no”.

\[
\Pr[A(x) = No \mid N \leq k] = \left(1 - \frac{1}{B}\right)^N \geq \left(1 - \frac{1}{B}\right)^k
\]
\[
\Pr[A(x) = No \mid N > 2k] = \left(1 - \frac{1}{B}\right)^N \leq \left(1 - \frac{1}{B}\right)^{2k}
\]

Constant gap (roughly \(1/e - 1/e^2\)).
Toy problem

Alg: Number of distinct elements
≤ k Output: “no”
≥ 2k Output: “yes”

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.
    \( h : [n] \rightarrow [B] \), where \( B = k \).
(2) If any \( h(x_i) = 0 \), say “yes”, else “no”.

\[
\Pr[A(x) = No \mid N \leq k] = \left(1 - \frac{1}{B}\right)^N \geq \left(1 - \frac{1}{B}\right)^k
\]

\[
\Pr[A(x) = No \mid N > 2k] = \left(1 - \frac{1}{B}\right)^N \leq \left(1 - \frac{1}{B}\right)^{2k}
\]

Constant gap (roughly \( 1/e - 1/e^2 \)).

Many trials,
Toy problem

Alg: Number of distinct elements
≤ k Output: “no”
≥ 2k Output: “yes”

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.
    \( h : [n] \rightarrow [B] \), where \( B = k \).
(2) If any \( h(x_i) = 0 \), say “yes”, else “no”.

\[
\Pr[A(x) = No \mid N \leq k] = \left(1 - \frac{1}{B}\right)^N \geq \left(1 - \frac{1}{B}\right)^k
\]

\[
\Pr[A(x) = No \mid N > 2k] = \left(1 - \frac{1}{B}\right)^N \leq \left(1 - \frac{1}{B}\right)^{2k}
\]

Constant gap (roughly \( 1/e - 1/e^2 \)).

Many trials, in parallel.
Toy problem

Alg: Number of distinct elements

≤ k Output: “no”
≥ 2k Output: “yes”

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.
   \( h: [n] \rightarrow [B] \), where \( B = k \).
(2) If any \( h(x_i) = 0 \), say “yes”, else “no”.

\[
\Pr[A(x) = \text{No} \mid N \leq k] = \left(1 - \frac{1}{B}\right)^N \geq \left(1 - \frac{1}{B}\right)^k
\]

\[
\Pr[A(x) = \text{No} \mid N > 2k] = \left(1 - \frac{1}{B}\right)^N \leq \left(1 - \frac{1}{B}\right)^{2k}
\]

Constant gap (roughly \( 1/e - 1/e^2 \)).
Many trials, in parallel gives good result.
Toy problem

Alg: Number of distinct elements

\( \leq k \) Output: “no”
\( \geq 2k \) Output: “yes”

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.
\( h : [n] \rightarrow [B], \text{ where } B = k. \)
(2) If any \( h(x_i) = 0 \), say “yes”, else “no”.

\[
\Pr[A(x) = \text{No} \mid N \leq k] = \left(1 - \frac{1}{B}\right)^N \geq \left(1 - \frac{1}{B}\right)^k
\]

\[
\Pr[A(x) = \text{No} \mid N > 2k] = \left(1 - \frac{1}{B}\right)^N \leq \left(1 - \frac{1}{B}\right)^{2k}
\]

Constant gap (roughly \( 1/e - 1/e^2 \)).
Many trials, in parallel gives good result. ..more later.
Toy problem

Alg: Number of distinct elements

≤ k Output: “no”

≥ 2k Output: “yes”

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.

\( h : [n] \rightarrow [B] \), where \( B = k \).

(2) If any \( h(x_i) = 0 \), say “yes”, else “no”.

\[
\Pr[A(x) = No \mid N \leq k] = \left(1 - \frac{1}{B}\right)^N \geq \left(1 - \frac{1}{B}\right)^k
\]

\[
\Pr[A(x) = No \mid N > 2k] = \left(1 - \frac{1}{B}\right)^N \leq \left(1 - \frac{1}{B}\right)^{2k}
\]

Constant gap (roughly \( 1/e - 1/e^2 \)).

Many trials, in parallel gives good result. ..more later.

Number of bits for random hash function?
Toy problem

Alg: Number of distinct elements
≤ k Output: “no”
≥ 2k Output: “yes”

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.
   \( h : [n] \to [B] \), where \( B = k \).
(2) If any \( h(x_i) = 0 \), say “yes”, else “no”.

\[
\Pr[A(x) = No \mid N \leq k] = \left(1 - \frac{1}{B}\right)^N \geq \left(1 - \frac{1}{B}\right)^k
\]

\[
\Pr[A(x) = No \mid N > 2k] = \left(1 - \frac{1}{B}\right)^N \leq \left(1 - \frac{1}{B}\right)^{2k}
\]

Constant gap (roughly \(1/e - 1/e^2\)).
Many trials, in parallel gives good result. ..more later.

Number of bits for random hash function?
\( k^n \) hash functions.
Toy problem

Alg: Number of distinct elements

$\leq k$ Output: “no”

$\geq 2k$ Output: “yes”

Don’t care if in between.

Randomized Algorithm:
(1) Choose random hash function.

$h : [n] \rightarrow [B]$, where $B = k$.

(2) If any $h(x_i) = 0$, say “yes”, else “no”.

$$
\Pr[A(x) = No \mid N \leq k] = \left(1 - \frac{1}{B}\right)^N \geq \left(1 - \frac{1}{B}\right)^k
$$

$$
\Pr[A(x) = No \mid N > 2k] = \left(1 - \frac{1}{B}\right)^N \leq \left(1 - \frac{1}{B}\right)^{2k}
$$

Constant gap (roughly $1/e - 1/e^2$).

Many trials, in parallel gives good result. ..more later.

Number of bits for random hash function?

$k^n$ hash functions. $n \log k$ bits to specify!
2-wise independent hash functions

The family $\mathcal{H} : [n] \rightarrow [p]$
2-wise independent hash functions

The family \( \mathcal{H} : [n] \to [p] \)

\[ h_{a,b}(x) = ax + b \mod p, \text{ prime } p \geq n, \ a, b \in \{0, \ldots, p - 1\} \]
2-wise independent hash functions

The family $H : [n] \rightarrow [p]$

$h_{a,b}(x) = ax + b \mod p$, prime $p \geq n$, $a, b \in \{0, \ldots, p - 1\}$

is 2-wise independent:
2-wise independent hash functions

The family \( \mathcal{H} : [n] \to [p] \)

\[ h_{a,b}(x) = ax + b \mod p, \text{ prime } p \geq n, \ a, b \in \{0, \ldots, p - 1\} \]
is 2-wise independent:

\[
\Pr_{a,b}[h(x) = c \land h(y) = d] = \frac{1}{p^2} \quad \forall x \neq y
\]
2-wise independent hash functions

The family \( \mathcal{H} : [n] \rightarrow [p] \)

\( h_{a,b}(x) = ax + b \mod p \), prime \( p \geq n \), \( a, b \in \{0, \ldots, p - 1\} \)

is 2-wise independent:

\[
\Pr_{a,b}[h(x) = c \land h(y) = d] = \frac{1}{p^2} \quad \forall x \neq y
\]

Proof: If \( h(x) = c \) and \( h(y) = d \) then

\[
\Pr_{a,b}[h(x) = c \land h(y) = d]\]
2-wise independent hash functions

The family $\mathcal{H} : [n] \rightarrow [p]$

$$h_{a,b}(x) = ax + b \mod p, \text{ prime } p \geq n, \ a, b \in \{0, \ldots, p - 1\}$$

is 2-wise independent:

$$\Pr_{a,b}[h(x) = c \land h(y) = d] = \frac{1}{p^2} \quad \forall x \neq y$$

**Proof:** If $h(x) = c$ and $h(y) = d$ then

$$ax + b = c \pmod{p} \quad ay + b = d \pmod{p}$$
2-wise independent hash functions

The family \( \mathcal{H} : [n] \to [p] \)

\( h_{a,b}(x) = ax + b \mod p, \text{ prime } p \geq n, a,b \in \{0,\ldots,p-1\} \)

is 2-wise independent:

\[
\Pr_{a,b}[h(x) = c \land h(y) = d] = \frac{1}{p^2} \quad \forall x \neq y
\]

Proof: If \( h(x) = c \) and \( h(y) = d \) then

\[
ax + b = c \pmod{p} \quad ay + b = d \pmod{p}
\]

has unique solution for \( a,b \) since \( (x-y) \neq 0 \).
2-wise independent hash functions

The family \( \mathcal{H} : [n] \rightarrow [p] \)
\[ h_{a,b}(x) = ax + b \mod p, \text{ prime } p \geq n, \ a, b \in \{0, \ldots, p - 1\} \]
is 2-wise independent:

\[
\Pr_{a,b}[h(x) = c \land h(y) = d] = \frac{1}{p^2} \quad \forall x \neq y
\]

**Proof:** If \( h(x) = c \) and \( h(y) = d \) then

\[
ax + b = c \pmod{p} \quad ay + b = d \pmod{p}
\]

has unique solution for \( a, b \) since \((x - y) \neq 0\).

\( \rightarrow \) One \( h_{a,b} \) out of \( p^2 \) functions has \( h(x) = c \) and \( h(y) = d \). \( \square \)
2-wise independent hash functions

The family $\mathcal{H} : [n] \rightarrow [p]$

$h_{a,b}(x) = ax + b \mod p$, prime $p \geq n$, $a, b \in \{0, \ldots, p-1\}$

is 2-wise independent:

$$\Pr_{a,b}[h(x) = c \land h(y) = d] = \frac{1}{p^2} \quad \forall x \neq y$$

Proof: If $h(x) = c$ and $h(y) = d$ then

$$ax + b = c \pmod{p} \quad ay + b = d \pmod{p}$$

has unique solution for $a, b$ since $(x - y) \neq 0$.

$\rightarrow$ One $h_{a,b}$ out of $p^2$ functions has $h(x) = c$ and $h(y) = d$. \hfill \Box$

Nonprime $|B| < p.$
2-wise independent hash functions

The family \( \mathcal{H} : [n] \rightarrow [p] \)
\[
h_{a,b}(x) = ax + b \mod p, \text{ prime } p \geq n, \ a, b \in \{0, \ldots, p - 1\}
\]
is 2-wise independent:

\[
\Pr_{a,b}[h(x) = c \land h(y) = d] = \frac{1}{p^2} \quad \forall x \neq y
\]

**Proof:** If \( h(x) = c \) and \( h(y) = d \) then

\[
ax + b = c \pmod{p} \quad ay + b = d \pmod{p}
\]

has unique solution for \( a, b \) since \( (x - y) \neq 0 \).

\( \rightarrow \) One \( h_{a,b} \) out of \( p^2 \) functions has \( h(x) = c \) and \( h(y) = d \). \( \Box \)

Nonprime \( |B| < p \).

\( \mathcal{H} : [n] \rightarrow |B|, \)
2-wise independent hash functions

The family \( \mathcal{H} : [n] \to [p] \)
\[
h_{a,b}(x) = ax + b \mod p, \text{ prime } p \geq n, a, b \in \{0, \ldots, p - 1\}
\]
is 2-wise independent:
\[
\Pr_{a,b}[h(x) = c \land h(y) = d] = \frac{1}{p^2} \quad \forall x \neq y
\]

**Proof:** If \( h(x) = c \) and \( h(y) = d \) then
\[
ax + b = c \pmod{p} \quad ay + b = d \pmod{p}
\]
has unique solution for \( a, b \) since \( (x - y) \neq 0 \).

\( \rightarrow \) One \( h_{a,b} \) out of \( p^2 \) functions has \( h(x) = c \) and \( h(y) = d \).

Nonprime \( |B| < p \).

\( \mathcal{H} : [n] \to |B|, h_{a,b} = (ax + b) \pmod{p} \pmod{|B|} \)
2-wise independent hash functions

The family $\mathcal{H} : [n] \to [p]$

$$h_{a,b}(x) = ax + b \mod p,$$ prime $p \geq n$, $a, b \in \{0, \ldots, p - 1\}$
is 2-wise independent:

$$\Pr_{a,b}[h(x) = c \land h(y) = d] = \frac{1}{p^2} \quad \forall x \neq y$$

**Proof:** If $h(x) = c$ and $h(y) = d$ then

$$ax + b = c \pmod{p} \quad ay + b = d \pmod{p}$$

has unique solution for $a, b$ since $(x - y) \neq 0$.

$\rightarrow$ One $h_{a,b}$ out of $p^2$ functions has $h(x) = c$ and $h(y) = d$.

Nonprime $|B| < p$.

$\mathcal{H} : [n] \to |B|$, $h_{a,b} = (ax + b) \pmod{p} \pmod{|B|}$

Approximately 2-wise independent.

$$\Pr[\text{collision at } c \text{ and } d] \approx \frac{1}{|B|^2} \left(1 \pm \frac{k}{p}\right)^2$$

Assume $p \gg 1$, so basically assume perfectly independent.
2-wise independent hash functions

The family $\mathcal{H} : [n] \rightarrow [p]$,

$$h_{a,b}(x) = ax + b \mod p, \text{prime } p \geq n, \ a, b \in \{0, \ldots, p-1\}$$

is 2-wise independent:

$$\Pr_{a,b}[h(x) = c \land h(y) = d] = \frac{1}{p^2} \quad \forall x \neq y$$

**Proof:** If $h(x) = c$ and $h(y) = d$ then

$$ax + b = c \pmod{p} \quad ay + b = d \pmod{p}$$

has unique solution for $a, b$ since $(x - y) \neq 0$.

→ One $h_{a,b}$ out of $p^2$ functions has $h(x) = c$ and $h(y) = d$.

Nonprime $|B| < p$.

$\mathcal{H} : [n] \rightarrow |B|, \ h_{a,b} = (ax + b) \pmod{p} \pmod{|B|}$

Approximately 2-wise independent.

$$\Pr[\text{collision at } c \text{ and } d] \approx \frac{1}{|B|^2} \left(1 \pm \frac{k}{p}\right)^2 \text{ Assume } p >> 1, \text{ so basically assume perfectly independent.}$$

$(k$-wise independent hash family.)
2-wise independent hash functions

The family $\mathcal{H} : [n] \to [p]$

$h_{a,b}(x) = ax + b \mod p$, prime $p \geq n$, $a, b \in \{0, \ldots, p-1\}$
is 2-wise independent:

$$\Pr_{a,b}[h(x) = c \land h(y) = d] = \frac{1}{p^2} \quad \forall x \neq y$$

**Proof:** If $h(x) = c$ and $h(y) = d$ then

$$ax + b = c \pmod{p} \quad ay + b = d \pmod{p}$$

has unique solution for $a, b$ since $(x - y) \neq 0$.
→ One $h_{a,b}$ out of $p^2$ functions has $h(x) = c$ and $h(y) = d$. □

Nonprime $|B| < p$.

$\mathcal{H} : [n] \to |B|$, $h_{a,b} = (ax + b) \pmod{p} \pmod{|B|}$

Approximately 2-wise independent.

$$\Pr[\text{collision at } c \text{ and } d] \approx \frac{1}{|B|^2} \left(1 \pm \frac{k}{p}\right)^2 \text{ Assume } p \gg 1, \text{ so}$$
basically assume perfectly independent.

($k$-wise independent hash family. degree $k$ polynomials.)
Distinct elements with 2-wise hash functions.

\( N \) distinct items.

Toy Alg:

1. Random hash \( h \) from \( H \): \( n \rightarrow 4k \).
2. If \( h(x_i) = 0 \), say "yes", else say "no".

Union Bound:

\[
\Pr[A \cup B] \leq \Pr[A] + \Pr[B]
\]

\[
\Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i \Pr[A_i]
\]

\[
\Pr["yes" \mid N < k] \leq \sum_j \Pr[h(j) = 0] \leq k(1 - \frac{1}{4k})
\]

Inclusion/Exclusion:

\[
\Pr[A \cup B] \geq \Pr[A] + \Pr[B] - \Pr[A \cap B]
\]

\[
\Pr[\bigcup A_i] \geq \sum_i \Pr[A_i] - \sum_i, j \Pr[A_i \cap A_j]
\]

\[
\Pr["yes" \mid N \geq 2k] \geq \frac{2k}{2k} - \frac{2k}{(2k-1)} = \frac{3}{8}
\]

See this as one of two coins. Either heads with prob \( \leq \frac{1}{4} \) or \( \frac{3}{8} \). Gap of \( \frac{1}{8} \).

Flip coin (in parallel) to pump up the probability!
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:

1. Random hash $h$ from $H$: $\{n\} \rightarrow \{4^k\}$.
2. If $h(x_i) = 0$, say "yes," else say "no"

Union Bound:

$\Pr[A \cup B] \leq \Pr[A] + \Pr[B]$

$\Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i \Pr[A_i]$

$\Pr["yes" | N < k] \leq k(\frac{1}{4^k}) \leq 1^4$

Inclusion/Exclusion:

$\Pr[A \cup B] \geq \Pr[A] + \Pr[B] - \Pr[A \cap B]$

$\Pr[\bigcup A_i] \geq \sum_i \Pr[A_i] - \sum_{i,j} \Pr[A_i \cap A_j]$

$\Pr["yes" | N \geq 2k] \geq 2^k - 2^k(\frac{1}{2^k}) = (\frac{3}{8})$

See this as one of two coins.

Either heads with prob $\leq \frac{1}{4}$.

Either heads with prob $\leq \frac{3}{8}$.

Gap of $\frac{1}{8}$.

Flip coin (in parallel) to pump up the volume!
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:
(1) Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$. 

See this as one of two coins.
Either heads with prob $\leq \frac{1}{4}$
Either heads with prob $\leq \frac{3}{8}$
Gap of $\frac{1}{8}$.

Flip coin (in parallel) to pump up the volume!
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:
(1) Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$.
(2) If $h(x_i) = 0$, say “yes”, else say “no”
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:

1. Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$.
2. If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound:
Distinct elements with 2-wise hash functions.

$N$ distinct items.
Toy Alg:
(1) Random hash $h$ from $\mathcal{H} : [n] \to [4k]$.
(2) If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]$
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:

(1) Random hash $h$ from $H : [n] \rightarrow [4k]$.

(2) If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]$

$Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]$
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:
1. Random hash $h$ from $\mathcal{H} : [n] \to [4k]$.
2. If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]
Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]
Pr[\text{"yes"}|N < k] \leq $
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:
(1) Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$.
(2) If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]$

$Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]$

$Pr[\text{“yes”}|N < k] \leq \sum_j Pr[h(j) = 0]$
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:

1. Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$.
2. If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]$  
$Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]$  
$Pr["yes" | N < k] \leq \sum_j Pr[h(j) = 0] \leq k(\frac{1}{4k})$
Distinct elements with 2-wise hash functions.

\( N \) distinct items.

Toy Alg:
1. Random hash \( h \) from \( \mathcal{H} : [n] \rightarrow [4k] \).
2. If \( h(x_i) = 0 \), say “yes”, else say “no”

Union Bound:
\[
Pr[A \cup B] \leq Pr[A] + Pr[B]
\]
\[
Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]
\]
\[
Pr[“yes”|N < k] \leq \sum_j Pr[h(j) = 0] \leq k(\frac{1}{4k}) \leq \frac{1}{4}
\]
Distinct elements with 2-wise hash functions.

N distinct items.

Toy Alg:
1. Random hash \( h \) from \( \mathcal{H} : [n] \rightarrow [4k] \).
2. If \( h(x_i) = 0 \), say “yes”, else say “no”

Union Bound: 
\[
Pr[A \cup B] \leq Pr[A] + Pr[B]
\]
\[
Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]
\]
\[
Pr[“yes”|N < k] \leq \sum_j Pr[h(j) = 0] \leq k \left( \frac{1}{4k} \right) \leq \frac{1}{4}
\]

Inclusion/Exclusion:
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:
1. Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$.
2. If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]

Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]

Pr[“yes”|N < k] \leq \sum_j Pr[h(j)= 0] \leq k\left(\frac{1}{4k}\right) \leq \frac{1}{4}

Inclusion/Exclusion: $Pr[A \cup B] \geq Pr[A] + Pr[B] - Pr[A \cap B]$
Distinct elements with 2-wise hash functions.

\(N\) distinct items.

Toy Alg:
1. Random hash \(h\) from \(\mathcal{H} : [n] \rightarrow [4k]\).
2. If \(h(x_i) = 0\), say “yes”, else say “no”

Union Bound: \(Pr[A \cup B] \leq Pr[A] + Pr[B]\)
\[Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]\]

\[Pr[“yes”|N < k] \leq \sum_j Pr[h(j) = 0] \leq k\left(\frac{1}{4k}\right) \leq \frac{1}{4}\]

Inclusion/Exclusion: \(Pr[A \cup B] \geq Pr[A] + Pr[B] - Pr[A \cap B]\)
\[Pr[\bigcup A_i] \geq \sum_i Pr[A_i] - \sum_{i,j} Pr[A_i \cap A_j]\]
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:
1. Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$.
2. If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]$

$Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]$

$Pr["yes" | N < k] \leq \sum_j Pr[h(j) = 0] \leq k(\frac{1}{4k}) \leq \frac{1}{4}$

Inclusion/Exclusion: $Pr[A \cup B] \geq Pr[A] + Pr[B] - Pr[A \cap B]$

$Pr[\cup A_i] \geq \sum_i Pr[A_i] - \sum_{i,j} Pr[A_i \cap A_j]$

$Pr["yes" | N \geq 2k] \geq \frac{2k}{B}$
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:
1. Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$.
2. If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]$

$Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]$

$Pr["yes" | N < k] \leq \sum_j Pr[h(j) = 0] \leq k\left(\frac{1}{4k}\right) \leq \frac{1}{4}$

Inclusion/Exclusion: $Pr[A \cup B] \geq Pr[A] + Pr[B] - Pr[A \cap B]$

$Pr[\cup A_i] \geq \sum_i Pr[A_i] - \sum_{i,j} Pr[A_i \cap A_j]$

$Pr["yes" | N \geq 2k] \geq \frac{2k}{B} - \frac{2k(2k-1)}{2B}$
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:
(1) Random hash $h$ from $\mathcal{H}: [n] \rightarrow [4k]$.
(2) If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]$  
$Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]$  
$Pr[\text{“yes”} | N < k] \leq \sum_j Pr[h(j) = 0] \leq k(\frac{1}{4k}) \leq \frac{1}{4}$

Inclusion/Exclusion: $Pr[A \cup B] \geq Pr[A] + Pr[B] - Pr[A \cap B]$  
$Pr[\cup A_i] \geq \sum_i Pr[A_i] - \sum_{i,j} Pr[A_i \cap A_j]$  
$Pr[\text{“yes”} | N \geq 2k] \geq \frac{2k}{B} - \frac{2k(2k-1)}{2B} \geq \frac{2k}{B} (1 - \frac{k}{B})$
Distinct elements with 2-wise hash functions.

N distinct items.

Toy Alg:
1. Random hash $h$ from $\mathcal{H}: [n] \rightarrow [4k]$.
2. If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]$  
$Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]$

$$Pr["yes"|N < k] \leq \sum_j Pr[h(j) = 0] \leq k\left(\frac{1}{4k}\right) \leq \frac{1}{4}$$

Inclusion/Exclusion: $Pr[A \cup B] \geq Pr[A] + Pr[B] - Pr[A \cap B]$  
$Pr[\cup A_i] \geq \sum_i Pr[A_i] - \sum_{i,j} Pr[A_i \cap A_j]$

$$Pr["yes"|N \geq 2k] \geq \frac{2k}{B} - \frac{2k(2k-1)}{2B} \geq \frac{2k}{B} \left(1 - \frac{k}{B}\right) = \left(\frac{3}{8}\right)$$
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:
(1) Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$.
(2) If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]

Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]

Pr[“yes”|N < k] \leq \sum_j Pr[h(j) = 0] \leq k\left(\frac{1}{4k}\right) \leq \frac{1}{4}$

Inclusion/Exclusion: $Pr[A \cup B] \geq Pr[A] + Pr[B] - Pr[A \cap B]

Pr[\bigcup A_i] \geq \sum_i Pr[A_i] - \sum_{i,j} Pr[A_i \cap A_j]

Pr[“yes”|N \geq 2k] \geq \frac{2k}{B} - \frac{2k(2k-1)}{2B} \geq \frac{2k}{B} \left(1 - \frac{k}{B}\right) = \left(\frac{3}{8}\right)$

See this as one of two coins.
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:

1. Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$.
2. If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: 

\[
Pr[A \cup B] \leq Pr[A] + Pr[B] \\
Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]
\]

\[
Pr[“yes”|N < k] \leq \sum_j Pr[h(j) = 0] \leq k\left(\frac{1}{4k}\right) \leq \frac{1}{4}
\]

Inclusion/Exclusion: 

\[
Pr[A \cup B] \geq Pr[A] + Pr[B] - Pr[A \cap B] \\
Pr[\cup A_i] \geq \sum_i Pr[A_i] - \sum_{i,j} Pr[A_i \cap A_j]
\]

\[
Pr[“yes”|N \geq 2k] \geq \frac{2k}{B} - \frac{2k(2k-1)}{2B} \geq \frac{2k}{B} \left(1 - \frac{k}{B}\right) = \left(\frac{3}{8}\right)
\]

See this as one of two coins.

Either heads with prob $\leq \frac{1}{4}$
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:
(1) Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$.
(2) If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]$

$Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]$

$Pr[\text{“yes”} | N < k] \leq \sum_j Pr[h(j) = 0] \leq k\left(\frac{1}{4k}\right) \leq \frac{1}{4}$

Inclusion/Exclusion: $Pr[A \cup B] \geq Pr[A] + Pr[B] - Pr[A \cap B]$

$Pr[\cup A_i] \geq \sum_i Pr[A_i] - \sum_{i,j} Pr[A_i \cap A_j]$

$Pr[\text{“yes”} | N \geq 2k] \geq \frac{2k}{B} - \frac{2k.(2k-1)}{2B} \geq \frac{2k}{B} \left(1 - \frac{k}{B}\right) = \left(\frac{3}{8}\right)$

See this as one of two coins.

Either heads with prob $\leq \frac{1}{4}$
Either heads with prob $\leq \frac{3}{8}$
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:
(1) Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$.
(2) If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]$
$Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]$

$Pr["yes" | N < k] \leq \sum_j Pr[h(j) = 0] \leq k(\frac{1}{4k}) \leq \frac{1}{4}$

Inclusion/Exclusion: $Pr[A \cup B] \geq Pr[A] + Pr[B] - Pr[A \cap B]$
$Pr[\bigcup A_i] \geq \sum_i Pr[A_i] - \sum_{i,j} Pr[A_i \cap A_j]$

$Pr["yes" | N \geq 2k] \geq \frac{2k}{B} - \frac{2k(2k-1)}{2B} \geq \frac{2k}{B} (1 - \frac{k}{B}) = (\frac{3}{8})$

See this as one of two coins.

Either heads with prob $\leq \frac{1}{4}$
Either heads with prob $\leq \frac{3}{8}$
Gap of $\frac{1}{8}$. 
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:

1. Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$.
2. If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound:

$Pr[A \cup B] \leq Pr[A] + Pr[B]$  
$Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]$  

$Pr[\text{“yes”} | N < k] \leq \sum_j Pr[h(j) = 0] \leq k \left( \frac{1}{4k} \right) \leq \frac{1}{4}$

Inclusion/Exclusion:

$Pr[A \cup B] \geq Pr[A] + Pr[B] - Pr[A \cap B]$  
$Pr[\cup A_i] \geq \sum_i Pr[A_i] - \sum_{i,j} Pr[A_i \cap A_j]$  

$Pr[\text{“yes”} | N \geq 2k] \geq \frac{2k}{B} - \frac{2k(2k - 1)}{2B} \geq \frac{2k}{B} \left( 1 - \frac{k}{B} \right) = \left( \frac{3}{8} \right)$

See this as one of two coins.

Either heads with prob $\leq \frac{1}{4}$
Either heads with prob $\leq \frac{3}{8}$
Gap of $\frac{1}{8}$.

Flip coin (in parallel)
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:
(1) Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$.
(2) If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]$
$Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]$

$Pr[\text{"yes" } | N < k] \leq \sum_j Pr[h(j) = 0] \leq k(\frac{1}{4k}) \leq \frac{1}{4}$

Inclusion/Exclusion: $Pr[A \cup B] \geq Pr[A] + Pr[B] - Pr[A \cap B]$
$Pr[\cup A_i] \geq \sum_i Pr[A_i] - \sum_{i,j} Pr[A_i \cap A_j]$

$Pr[\text{"yes" } | N \geq 2k] \geq \frac{2k}{B} - \frac{2k \cdot (2k - 1)}{2B} \geq \frac{2k}{B} \left(1 - \frac{k}{B}\right) = \left(\frac{3}{8}\right)$

See this as one of two coins.

Either heads with prob $\leq \frac{1}{4}$
Either heads with prob $\leq \frac{3}{8}$
Gap of $\frac{1}{8}$.

Flip coin (in parallel) to pump up the volume!
Distinct elements with 2-wise hash functions.

$N$ distinct items.

Toy Alg:
1. Random hash $h$ from $\mathcal{H} : [n] \rightarrow [4k]$.
2. If $h(x_i) = 0$, say “yes”, else say “no”

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]$

$Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]

Pr[“yes” | N < k] \leq \sum_j Pr[h(j) = 0] \leq k(\frac{1}{4k}) \leq \frac{1}{4}$

Inclusion/Exclusion: $Pr[A \cup B] \geq Pr[A] + Pr[B] - Pr[A \cap B]$

$Pr[\cup A_i] \geq \sum_i Pr[A_i] - \sum_{i,j} Pr[A_i \cap A_j]$}

$$Pr[“yes” | N \geq 2k] \geq \frac{2k}{B} - \frac{2k.(2k-1)}{2B} \geq \frac{2k}{B} (1 - \frac{k}{B}) = \left(\frac{3}{8}\right)$$

See this as one of two coins.

Either heads with prob $\leq \frac{1}{4}$

Either heads with prob $\leq \frac{3}{8}$

Gap of $\frac{1}{8}$.

Flip coin (in parallel) to pump up the volume! probability!
It gets better.

**Simpl. Chernoff:** Number of heads $\hat{b}$ in $k = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$ flips of bias $b$ coin satisfies $bk(1 - \varepsilon) \leq \hat{b} \leq bk(1 + \varepsilon)$ with probability $1 - \delta$. 
It gets better.

**Simpl. Chernoff:** Number of heads $\hat{b}$ in $k = O(\frac{\log(1/\delta)}{\epsilon^2})$ flips of bias $b$ coin satisfies $bk(1 - \epsilon) \leq \hat{b} \leq bk(1 + \epsilon)$ with probability $1 - \delta$.

Alg:
It gets better.

**Simpl. Chernoff:** Number of heads $\hat{b}$ in $k = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$ flips of bias $b$ coin satisfies $bk(1 - \varepsilon) \leq \hat{b} \leq bk(1 + \varepsilon)$ with probability $1 - \delta$.

**Alg:**
- “yes” with probability at most $1/4$ when $N < k$.
- “yes” with probability at least $3/8$ when $N > 2k$. 

Run $\Theta\left(\log\frac{1}{\delta}\right)$ independent copies of Alg.

Output “yes” if more than $\frac{5}{16}$ yes’s.

Use claim with $\varepsilon = \frac{1}{3}$.

$\rightarrow$ Correct with probability $\geq 1 - \delta$. 

Run $\log n$ times to get within factor of two.

Choose $|B| = \theta(k\varepsilon)$ in Alg.

“yes” with probability at most $\tau$ when $N < k$.

“yes” with probability at least $\left(1 + \varepsilon\right)\tau$ when $N > (1 + \varepsilon)k$.

Run $\log \frac{1}{\delta\varepsilon^2}$ times to pump up the probability.

Run $\log \frac{1}{1 + \varepsilon} n$ times to get within factor of $1 + \varepsilon$.

$O\left(\log n \log \frac{1}{\delta\varepsilon^2}\right)$ space, $(1 \pm \varepsilon)$ estimate, w/prob $1 - \delta$. 
It gets better.

**Simpl. Chernoff:** Number of heads $\hat{b}$ in $k = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$ flips of bias $b$ coin satisfies $bk(1 - \varepsilon) \leq \hat{b} \leq bk(1 + \varepsilon)$ with probability $1 - \delta$.

**Alg:**
- "yes" with probability at most $1/4$ when $N < k$.
- "yes" with probability at least $3/8$ when $N > 2k$.

Run $\Theta(\log \frac{1}{\delta})$ independent copies of Alg.
It gets better.

**Simpl. Chernoff:** Number of heads \( \hat{b} \) in \( k = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right) \) flips of bias \( b \) coin satisfies \( bk(1 - \varepsilon) \leq \hat{b} \leq bk(1 + \varepsilon) \) with probability \( 1 - \delta \).

**Alg:**

“yes” with probability at most \( 1/4 \) when \( N < k \).

“yes” with probability at least \( 3/8 \) when \( N > 2k \).

Run \( \Theta(\log \frac{1}{\delta}) \) independent copies of Alg.

Output “yes” if more than \( \frac{5}{16} \) yes’s.
It gets better.

**Simpl. Chernoff:** Number of heads $\hat{b}$ in $k = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$ flips of bias $b$ coin satisfies $bk(1 - \varepsilon) \leq \hat{b} \leq bk(1 + \varepsilon)$ with probability $1 - \delta$.

Alg:
- “yes” with probability at most $1/4$ when $N < k$.
- “yes” with probability at least $3/8$ when $N > 2k$.

Run $\Theta(\log \frac{1}{\delta})$ independent copies of Alg.
- Output “yes” if more than $\frac{5}{16}$ yes’s.

Use claim with $\varepsilon = \frac{1}{3}$. 
It gets better.

**Simpl. Chernoff:** Number of heads $\hat{b}$ in $k = O(\frac{\log(1/\delta)}{\varepsilon^2})$ flips of bias $b$ coin satisfies $bk(1 - \varepsilon) \leq \hat{b} \leq bk(1 + \varepsilon)$ with probability $1 - \delta$.

Alg:

“yes” with probability at most $1/4$ when $N < k$.

“yes” with probability at least $3/8$ when $N > 2k$.

Run $\Theta(\log \frac{1}{\delta})$ independent copies of Alg.

Output “yes” if more than $\frac{5}{16}$ yes’s.

Use claim with $\varepsilon = \frac{1}{3}$.

$\rightarrow$ Correct with probability $\geq 1 - \delta$. 
It gets better.

**Simpl. Chernoff:** Number of heads $\hat{b}$ in $k = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ flips of bias $b$ coin satisfies $bk(1 - \epsilon) \leq \hat{b} \leq bk(1 + \epsilon)$ with probability $1 - \delta$.

**Alg:**

“yes” with probability at most $1/4$ when $N < k$.

“yes” with probability at least $3/8$ when $N > 2k$.

Run $\Theta(\log \frac{1}{\delta})$ independent copies of Alg.

Output “yes” if more than $\frac{5}{16}$ yes’s.

Use claim with $\epsilon = \frac{1}{3}$.

→ Correct with probability $\geq 1 - \delta$.

Run log $n$ times to get within factor of two.
It gets better.

**Simpl. Chernoff:** Number of heads \( \hat{b} \) in \( k = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right) \) flips of bias \( b \) coin satisfies \( bk(1 - \varepsilon) \leq \hat{b} \leq bk(1 + \varepsilon) \) with probability \( 1 - \delta \).

Alg:

- “yes” with probability at most \( 1/4 \) when \( N < k \).
- “yes” with probability at least \( 3/8 \) when \( N > 2k \).

Run \( \Theta(\log \frac{1}{\delta}) \) independent copies of Alg.

Output “yes” if more than \( \frac{5}{16} \) yes’s.

Use claim with \( \varepsilon = \frac{1}{3} \).

\( \rightarrow \) Correct with probability \( \geq 1 - \delta \).

Run \( \log n \) times to get within factor of two.

Factor of \( (1 + \varepsilon) \)?
It gets better.

**Simpl. Chernoff:** Number of heads $\hat{b}$ in $k = O(\frac{\log(1/\delta)}{\epsilon^2})$ flips of bias $b$ coin satisfies $bk(1 - \epsilon) \leq \hat{b} \leq bk(1 + \epsilon)$ with probability $1 - \delta$.

Alg:
- “yes” with probability at most $1/4$ when $N < k$.
- “yes” with probability at least $3/8$ when $N > 2k$.

Run $\Theta(\log \frac{1}{\delta})$ independent copies of Alg.
- Output “yes” if more than $\frac{5}{16}$ yes’s.

Use claim with $\epsilon = \frac{1}{3}$.
→ Correct with probability $\geq 1 - \delta$.
- Run $\log n$ times to get within factor of two.

Factor of $(1 + \epsilon)$? Choose $|B| = \theta\left(\frac{k}{\epsilon}\right)$ in Alg.
It gets better.

**Simpl. Chernoff:** Number of heads $\hat{b}$ in $k = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$ flips of bias $b$ coin satisfies $bk(1 - \varepsilon) \leq \hat{b} \leq bk(1 + \varepsilon)$ with probability $1 - \delta$.

**Alg:**

“yes” with probability at most $1/4$ when $N < k$.
“yes” with probability at least $3/8$ when $N > 2k$.

Run $\Theta(\log \frac{1}{\delta})$ independent copies of Alg.
Output “yes” if more than $\frac{5}{16}$ yes’s.

Use claim with $\varepsilon = \frac{1}{3}$.
→ Correct with probability $\geq 1 - \delta$.
Run $\log n$ times to get within factor of two.

Factor of $(1 + \varepsilon)$? Choose $|B| = \theta\left(\frac{k}{\varepsilon}\right)$ in Alg.
“yes” with probability at most $\tau$ when $N < k$. 
It gets better.

**Simpl. Chernoff:** Number of heads $\hat{b}$ in $k = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ flips of bias $b$ coin satisfies $bk(1 - \epsilon) \leq \hat{b} \leq bk(1 + \epsilon)$ with probability $1 - \delta$.

Alg:

“yes” with probability at most $1/4$ when $N < k$.

“yes” with probability at least $3/8$ when $N > 2k$.

Run $\Theta(\log \frac{1}{\delta})$ independent copies of Alg.

Output “yes” if more than $\frac{5}{16}$ yes’s.

Use claim with $\epsilon = \frac{1}{3}$.

→ Correct with probability $\geq 1 - \delta$.

Run $\log n$ times to get within factor of two.

Factor of $(1 + \epsilon)$? Choose $|B| = \theta\left(\frac{k}{\epsilon}\right)$ in Alg.

“yes” with probability at most $\tau$ when $N < k$.

“yes” with probability at least $(1 + \epsilon)\tau$ when $N > (1 + \epsilon)k$. 
It gets better.

**Simpl. Chernoff:** Number of heads $\hat{b}$ in $k = O(\log(1/\delta)\epsilon^2)$ flips of bias $b$ coin satisfies $bk(1 - \epsilon) \leq \hat{b} \leq bk(1 + \epsilon)$ with probability $1 - \delta$.

Alg:

“yes” with probability at most $1/4$ when $N < k$.
“yes” with probability at least $3/8$ when $N > 2k$.

Run $\Theta(\log \frac{1}{\delta})$ independent copies of Alg.

Output “yes” if more than $\frac{5}{16}$ yes’s.

Use claim with $\epsilon = \frac{1}{3}$.

$\rightarrow$ Correct with probability $\geq 1 - \delta$.

Run $\log n$ times to get within factor of two.

Factor of $(1 + \epsilon)$? Choose $|B| = \theta(\frac{k}{\epsilon})$ in Alg.

“yes” with probability at most $\tau$ when $N < k$.
“yes” with probability at least $(1 + \epsilon)\tau$ when $N > (1 + \epsilon)k$.

Run $\frac{\log \frac{1}{\delta}}{\epsilon^2}$ times to pump up the probability.
It gets better.

**Simpl. Chernoff:** Number of heads $\hat{b}$ in $k = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ flips of bias $b$ coin satisfies $bk(1 - \epsilon) \leq \hat{b} \leq bk(1 + \epsilon)$ with probability $1 - \delta$.

**Alg:**

“yes” with probability at most $1/4$ when $N < k$.

“yes” with probability at least $3/8$ when $N > 2k$.

Run $\Theta(\log \frac{1}{\delta})$ independent copies of Alg.

Output “yes” if more than $\frac{5}{16}$ yes’s.

Use claim with $\epsilon = \frac{1}{3}$.

→ Correct with probability $\geq 1 - \delta$.

Run $\log n$ times to get within factor of two.

Factor of $(1 + \epsilon)$? Choose $|B| = \theta\left(\frac{k}{\epsilon}\right)$ in Alg.

“yes” with probability **at most** $\tau$ when $N < k$.

“yes” with probability **at least** $(1 + \epsilon)\tau$ when $N > (1 + \epsilon)k$.

Run $\frac{\log \frac{1}{\delta}}{\epsilon^2}$ times to pump up the probability.

Run $\log_{1+\epsilon} n$ times to get within factor of $1 + \epsilon$. 


It gets better.

**Simpl. Chernoff:** Number of heads $\hat{b}$ in $k = O(\frac{\log(1/\delta)}{\varepsilon^2})$ flips of bias $b$ coin satisfies $bk(1 - \varepsilon) \leq \hat{b} \leq bk(1 + \varepsilon)$ with probability $1 - \delta$.

Alg:
- “yes” with probability at most $1/4$ when $N < k$.
- “yes” with probability at least $3/8$ when $N > 2k$.

Run $\Theta(\log \frac{1}{\delta})$ independent copies of Alg.
Output “yes” if more than $\frac{5}{16}$ yes’s.

Use claim with $\varepsilon = \frac{1}{3}$.
$\rightarrow$ Correct with probability $\geq 1 - \delta$.
- Run $\log n$ times to get within factor of two.

Factor of $(1 + \varepsilon)$? Choose $|B| = \theta\left(\frac{k}{\varepsilon}\right)$ in Alg.
- “yes” with probability at most $\tau$ when $N < k$.
- “yes” with probability at least $(1 + \varepsilon)\tau$ when $N > (1 + \varepsilon)k$.

Run $\frac{\log \frac{1}{\delta}}{\varepsilon^2}$ times to pump up the probability.
Run $\log_{1+\varepsilon} n$ times to get within factor of $1 + \varepsilon$.

$O(\log n \log_{1+\varepsilon} n \frac{\log \frac{1}{\delta}}{\varepsilon^2})$ space,
It gets better.

**Simpl. Chernoff:** Number of heads $\hat{b}$ in $k = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ flips of bias $b$ coin satisfies $bk(1 - \epsilon) \leq \hat{b} \leq bk(1 + \epsilon)$ with probability $1 - \delta$.

Alg:

“yes” with probability at most $1/4$ when $N < k$.

“yes” with probability at least $3/8$ when $N > 2k$.

Run $\Theta(\log \frac{1}{\delta})$ independent copies of Alg.

Output “yes” if more than $\frac{5}{16}$ yes’s.

Use claim with $\epsilon = \frac{1}{3}$.

→ Correct with probability $\geq 1 - \delta$.

Run $\log n$ times to get within factor of two.

Factor of $(1 + \epsilon)$? Choose $|B| = \theta\left(\frac{k}{\epsilon}\right)$ in Alg.

“yes” with probability at most $\tau$ when $N < k$.

“yes” with probability at least $(1 + \epsilon)\tau$ when $N > (1 + \epsilon)k$.

Run $\frac{\log \frac{1}{\delta}}{\epsilon^2}$ times to pump up the probability.

Run $\log_{1+\epsilon} n$ times to get within factor of $1 + \epsilon$.

$O(\log n \log_{1+\epsilon} n \frac{\log \frac{1}{\delta}}{\epsilon^2})$ space, $(1 \pm \epsilon)$ estimate,
It gets better.

**Simpl. Chernoff:** Number of heads \( \hat{b} \) in \( k = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right) \) flips of bias \( b \) coin satisfies \( bk(1 - \varepsilon) \leq \hat{b} \leq bk(1 + \varepsilon) \) with probability \( 1 - \delta \).

**Alg:**
- “yes” with probability at most \( 1/4 \) when \( N < k \).
- “yes” with probability at least \( 3/8 \) when \( N > 2k \).

Run \( \Theta(\log \frac{1}{\delta}) \) independent copies of Alg.
- Output “yes” if more than \( \frac{5}{16} \) yes’s.

Use claim with \( \varepsilon = \frac{1}{3} \).
- Correct with probability \( \geq 1 - \delta \).
  - Run \( \log n \) times to get within factor of two.

Factor of \( (1 + \varepsilon) \)? Choose \( |B| = \theta\left(\frac{k}{\varepsilon}\right) \) in Alg.
- “yes” with probability at most \( \tau \) when \( N < k \).
- “yes” with probability at least \( (1 + \varepsilon) \tau \) when \( N > (1 + \varepsilon)k \).

Run \( \frac{\log \frac{1}{\delta}}{\varepsilon^2} \) times to pump up the probability.
Run \( \log_{1+\varepsilon} n \) times to get within factor of \( 1 + \varepsilon \).

\( O(\log n \log_{1+\varepsilon} n \frac{\log \frac{1}{\delta}}{\varepsilon^2}) \) space, \( (1 \pm \varepsilon) \) estimate, w/prob \( 1 - \delta \).
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$. 
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H}: [n] \rightarrow \pm 1$. 

Estimating $F_2$

Second Moment: $F_2 = \sum_j m^2_j$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \to \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H}: [n] \rightarrow \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \to \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$$h(j) = Y_j$$
$$Z = \sum_i h(x_i)$$
$$E[Z^2] = \sum_j E[Y_j^2] m_j^2 + 3 \sum_{i, j} E[Y_i] E[Y_j] m_i^2 m_j^2 = \sum_i m_i^4 + 6 \sum_{i, j} m_i^2 m_j^2 \leq 2F_2^2$$
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \to \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$h(j) = Y_j$
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \rightarrow \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$h(j) = Y_j$

$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \rightarrow \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$h(j) = Y_j$
$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$

$E[Z^2]$
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \rightarrow \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$h(j) = Y_j$

$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$

$E[Z^2] = \sum_j E[Y_j^2] m_j^2$
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H}: [n] \to \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$h(j) = Y_j$

$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$

$E[Z^2] = \sum_j E[Y_j^2] m_j^2 + \sum_{i,j} E[Y_i] E[Y_j] m_i m_j$
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
1. Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \rightarrow \pm 1$.
2. Output $Z^2 = (\sum_i h(x_i))^2$


$$h(j) = Y_j$$
$$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$$

$$E[Z^2] = \sum_j E[Y_j^2] m_j^2 + \sum_{i,j} E[Y_i] E[Y_j] m_i m_j = \sum_i m_i^2 = F_2$$
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \rightarrow \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$h(j) = Y_j$
$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$

$E[Z^2] = \sum_j E[Y_j^2] m_j^2 + \sum_{i,j} E[Y_i] E[Y_j] m_i m_j = \sum_i m_i^2 = F_2$

Show good probability of success?
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H}: [n] \rightarrow \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$h(j) = Y_j$
$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$

$E[Z^2] = \sum_j E[Y_j^2] m_j^2 + \sum_{i,j} E[Y_i] E[Y_j] m_i m_j = \sum_i m_i^2 = F_2$

Show good probability of success? Calculate variance.
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$. 

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \rightarrow \pm 1$.  
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$h(j) = Y_j$

$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$

$E[Z^2] = \sum_j E[Y_j^2] m_j^2 + \sum_{i,j} E[Y_i] E[Y_j] m_i m_j = \sum_i m_i^2 = F_2$

Show good probability of success? Calculate variance.

$Var(X) = E[X^2] - (E[X])^2$
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \rightarrow \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$$h(j) = Y_j$$
$$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$$

$$E[Z^2] = \sum_j E[Y_j^2] m_j^2 + \sum_{i,j} E[Y_i] E[Y_j] m_i m_j = \sum_i m_i^2 = F_2$$

Show good probability of success? Calculate variance.

$$Var(X) = E[X^2] - (E[X])^2$$
$$E[Z^4]$$
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \rightarrow \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$$h(j) = Y_j$$
$$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$$

$$E[Z^2] = \sum_j E[Y_j^2] m_j^2 + \sum_{i,j} E[Y_i] E[Y_j] m_i m_j = \sum_i m_i^2 = F_2$$

Show good probability of success? Calculate variance.

$$Var(X) = E[X^2] - (E[X])^2$$
$$E[Z^4] = \sum_i E[Y_i^4 m_i^4] + 3 \sum_{i,j} E[Y_i^2 Y_j^2 m_i^2 m_j^2]$$
Estimating $F_2$

Second Moment: $F_2 = \sum j \ m^2_j$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \rightarrow \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$h(\cdot) = Y_j$
$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$

$E[Z^2] = \sum_j E[Y_j^2] m_j^2 + \sum_{i,j} E[Y_i] E[Y_j] m_i m_j = \sum_i m_i^2 = F_2$

Show good probability of success? Calculate variance.

$Var(X) = E[X^2] - (E[X])^2$

$E[Z^4] = \sum_i E[Y_i^4] m_i^4 + 3 \sum_{i,j} E[Y_i^2 Y_j^2 m_i^2 m_j^2] = \sum_i m_i^4 + 6 \sum_{i,j} m_i^2 m_j^2$
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \to \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$h(j) = Y_j$
$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$

$E[Z^2] = \sum_j E[Y_j^2] m_j^2 + \sum_{i,j} E[Y_i] E[Y_j] m_i m_j = \sum_i m_i^2 = F_2$

Show good probability of success? Calculate variance.

$Var(X) = E[X^2] - (E[X])^2$

$E[Z^4] = \sum_i E[Y_i^4] m_i^4 + 3 \sum_{i,j} E[Y_i^2 Y_j^2 m_i^2 m_j^2] = \sum_i m_i^4 + 6 \sum_{i,j} m_i^2 m_j^2$

$Var(Z^2)$
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \to \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$h(j) = Y_j$
$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$

$E[Z^2] = \sum_j E[Y_j^2] m_j^2 + \sum_{i,j} E[Y_i] E[Y_j] m_i m_j = \sum_i m_i^2 = F_2$

Show good probability of success? Calculate variance.

$\text{Var}(X) = E[X^2] - (E[X])^2$

$E[Z^4] = \sum_i E[Y_i^4 m_i^4] + 3 \sum_{i,j} E[Y_i^2 Y_j^2 m_i^2 m_j^2] = \sum_i m_i^4 + 6 \sum_{i,j} m_i^2 m_j^2$

Estimating $F_2$

Second Moment: $F_2 = \sum j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \to \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


\[ h(j) = Y_j \]
\[ Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j \]

\[ E[Z^2] = \sum j E[Y_j^2] m_j^2 + \sum_{i,j} E[Y_i] E[Y_j] m_i m_j = \sum i m_i^2 = F_2 \]

Show good probability of success? Calculate variance.

\[ \text{Var}(X) = E[X^2] - (E[X])^2 \]

\[ E[Z^4] = \sum_i E[Y_i^4 m_i^4] + 3 \sum_{i,j} E[Y_i^2 Y_j^2 m_i^2 m_j^2] = \sum_i m_i^4 + 6 \sum_{i,j} m_i^2 m_j^2 \]

\[ \text{Var}(Z^2) = E[Z^4] - E[Z^2]^2 = 2 \sum m_i^2 m_j^2 \]
Estimating $F_2$

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:
(1) Random $h$ from 4-wise ind. family $\mathcal{H} : [n] \to \pm 1$.
(2) Output $Z^2 = (\sum_i h(x_i))^2$


$h(j) = Y_j$
$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$

$E[Z^2] = \sum_j E[Y_j^2] m_j^2 + \sum_{i,j} E[Y_i] E[Y_j] m_i m_j = \sum_i m_i^2 = F_2$

Show good probability of success? Calculate variance.

$\text{Var}(X) = E[X^2] - (E[X])^2$

$E[Z^4] = \sum_i E[Y_i^4 m_i^4] + 3 \sum_{i,j} E[Y_i^2 Y_j^2 m_i^2 m_j^2] = \sum_i m_i^4 + 6 \sum_{i,j} m_i^2 m_j^2$

$\text{Var}(Z^2) = E[Z^4] - E[Z^2]^2 = 2 \sum m_i^2 m_j^2 \leq 2F_2^2$
Core Alg: analysis cont.

\[ E[Z^2] = F_2. \]
Core Alg: analysis cont.

\[ E[Z^2] = F_2. \]

\[ \text{Var}(Z^2) \]
Core Alg: analysis cont.

\[ E[Z^2] = F_2. \]
\[ \text{Var}(Z^2) = E[Z^4] - E[Z^2]^2 \]
Core Alg: analysis cont.

\[ E[Z^2] = F_2. \]

\[ \text{Var}(Z^2) = E[Z^4] - E[Z^2]^2 = 2 \sum m_i^2 m_j^2 \]
Core Alg: analysis cont.

\[ E[Z^2] = F_2. \]

\[ \text{Var}(Z^2) = E[Z^4] - E[Z^2]^2 = 2 \sum m_i^2 m_j^2 \leq 2F_2^2 \]
Core Alg: analysis cont.

\[ E[Z^2] = F_2. \]

\[ \text{Var}(Z^2) = E[Z^4] - E[Z^2]^2 = 2 \sum m_i^2 m_j^2 \leq 2F_2^2 \]

Close to expectation?
Core Alg: analysis cont.

\[ E[Z^2] = F_2. \]

\[ \text{Var}(Z^2) = E[Z^4] - E[Z^2]^2 = 2 \sum m_i^2 m_j^2 \leq 2F_2^2 \]

Close to expectation? \[ |Z^2 - \mu| \leq \varepsilon F_2 ? \]
Core Alg: analysis cont.

\[
\]

\[
\text{Var}(Z^2) = E[Z^4] - E[Z^2]^2 = 2 \sum m_i^2 m_j^2 \leq 2F_2^2
\]

Close to expectation? \(|Z^2 - \mu| \leq \varepsilon F_2\) ?

Chebyshev: \(\Pr[|X - \mu| > \Delta] \leq \frac{\text{Var}(X)}{\Delta^2}\)
Core Alg: analysis cont.

\[ E[Z^2] = F_2. \]

\[ \text{Var}(Z^2) = E[Z^4] - E[Z^2]^2 = 2 \sum m_i^2 m_j^2 \leq 2F_2^2 \]

Close to expectation? \( |Z^2 - \mu| \leq \varepsilon F_2 \)?

Chebyshev: \( Pr[|X - \mu| > \Delta] \leq \frac{\text{Var}(X)}{\Delta^2} \)

For \( Z^2 \), \( Pr[|Z^2 - \mu| > \varepsilon F_2] \leq \frac{2F_2^2}{\varepsilon^2 F_2^2} = \frac{2}{\varepsilon^2} \)
\[ E[Z^2] = F_2. \]

\[ \text{Var}(Z^2) = E[Z^4] - E[Z^2]^2 = 2 \sum m_i^2 m_j^2 \leq 2F_2^2 \]

Close to expectation? \( |Z^2 - \mu| \leq \varepsilon F_2 \) ?

Chebyshev: \( Pr[|X - \mu| > \Delta] \leq \frac{\text{Var}(X)}{\Delta^2} \)

For \( Z^2 \), \( Pr[|Z^2 - \mu| > \varepsilon F_2] \leq \frac{2F_2^2}{\varepsilon^2 F_2^2} = \frac{2}{\varepsilon^2} \)

Uh oh.
Core Alg: analysis cont.

\[ E[Z^2] = F_2. \]

\[ \text{Var}(Z^2) = E[Z^4] - E[Z^2]^2 = 2 \sum m_i^2 m_j^2 \leq 2F_2^2 \]

Close to expectation? \(|Z^2 - \mu| \leq \varepsilon F_2?\)

Chebyshev: \( \text{Pr}[|X - \mu| > \Delta] \leq \frac{\text{Var}(X)}{\Delta^2} \)

For \( Z^2 \), \( \text{Pr}[|Z^2 - \mu| > \varepsilon F_2] \leq \frac{2F_2^2}{\varepsilon^2 F_2^2} = \frac{2}{\varepsilon^2} \)

Uh oh. Bigger than one for \( \varepsilon \leq 2! \)
Independent trials.

Run Core Alg $k$ times.
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$. 

$$E[Z^2] = \text{Var}(Z) \leq 2F^2.$$ 

Output average. 

$$Y = \frac{1}{k} \sum_i Z^2_i \quad E[Y] = \frac{1}{k} \sum_i E[Z^2_i] = F^2 \quad \text{Var}(cX) = c^2 \text{Var}(X) \quad \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y); \text{ independent } X \text{ and } Y \quad \text{Var}(Y) = \frac{1}{k^2} \sum_i \text{Var}(Z^2_i) = \frac{2}{k}F^2,$$

and Chebyshev 

$$\Pr[|Y-\mu| \geq \epsilon F^2] \leq \delta.$$ 

Space: $O(\log n \epsilon^{-2} \delta^{-2})$. Could get $O(\log n \log \frac{1}{\delta} \epsilon^{-2})$ using a Central Limit Theorem.
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$.

$\mathbb{E}[Z_i^2] = F_2$
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$.

$$(E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2).$$
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$.

$$(E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2.)$$

Output average.
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$.

$(E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2).$ 

Output average. $Y = \frac{1}{k} \sum_i Z_i^2$
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$.

$(E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2.)$

Output average. $Y = \frac{1}{k} \sum_i Z_i^2$

$E[Y] =$
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$.

$$(E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2.)$$

Output average. $Y = \frac{1}{k} \sum Z_i^2$

$E[Y] = \frac{1}{k} \sum E[Z_i^2]$
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$.

$(E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2 )$

Output average. $Y = \frac{1}{k} \sum_i Z_i^2$

$E[Y] = \frac{1}{k} \sum E[Z_i^2] = F_2$
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$.

$$(E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2.)$$

Output average. $Y = \frac{1}{k} \sum_i Z_i^2$

$$E[Y] = \frac{1}{k} \sum E[Z_i^2] = F_2$$

$$Var(cX) = c^2 Var(X)$$
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$.

$$(E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2).$$

Output average. $Y = \frac{1}{k} \sum_i Z_i^2$

$E[Y] = \frac{1}{k} \sum E[Z_i^2] = F_2$

$Var(cX) = c^2 \ Var(X)$

$Var(X + Y) = Var(X) + Var(Y)$; independent $X$ and $Y$
Independent trials.

Run Core Alg \( k \) times. \( Z_1, \ldots, Z_k \).

\[
(E[Z_i^2] = F_2 \quad \text{Var}(Z_i^2) \leq 2F_2^2 .)
\]

Output average. \( Y = \frac{1}{k} \sum_i Z_i^2 \)

\[
E[Y] = \frac{1}{k} \sum E[Z_i^2] = F_2
\]

\[
\text{Var}(cX) = c^2 \text{Var}(X)
\]

\[
\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y); \text{ independent } X \text{ and } Y
\]

\[
\text{Var}(Y)
\]
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$.

$$(E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2.)$$

Output average. $Y = \frac{1}{k} \sum_i Z_i^2$

$E[Y] = \frac{1}{k} \sum E[Z_i^2] = F_2$

$Var(cX) = c^2 Var(X)$

$Var(X + Y) = Var(X) + Var(Y)$; independent $X$ and $Y$

$Var(Y) = \frac{1}{k^2} \sum_i Var(Z_i^2)$
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$. 

$$(E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2.)$$

Output average. $Y = \frac{1}{k} \sum_i Z_i^2$

$E[Y] = \frac{1}{k} \sum E[Z_i^2] = F_2$

$Var(cX) = c^2 Var(X)$

$Var(X + Y) = Var(X) + Var(Y); \text{ independent } X \text{ and } Y$

$Var(Y) = \frac{1}{k^2} \sum_i Var(Z_i^2) = \frac{2F_2^2}{k}$
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$.

$$E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2.$$ 

Output average. $Y = \frac{1}{k} \sum_i Z_i^2$

$$E[Y] = \frac{1}{k} \sum E[Z_i^2] = F_2$$

$$Var(cX) = c^2 Var(X)$$

$$Var(X + Y) = Var(X) + Var(Y); \text{ independent } X \text{ and } Y$$

$$Var(Y) = \frac{1}{k^2} \sum_i Var(Z_i^2) = \frac{2F_2^2}{k}$$

$k = \frac{2}{\delta \varepsilon^2}$ and Chebyshev
Independent trials.

Run Core Alg \( k \) times. \( Z_1, \ldots, Z_k \).

\[
(E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2.)
\]

Output average. \( Y = \frac{1}{k} \sum_i Z_i^2 \)

\[
E[Y] = \frac{1}{k} \sum E[Z_i^2] = F_2
\]

\[
Var(cX) = c^2 Var(X)
\]

\[
Var(X + Y) = Var(X) + Var(Y); \text{ independent } X \text{ and } Y
\]

\[
Var(Y) = \frac{1}{k^2} \sum_i Var(Z_i^2) = \frac{2F_2^2}{k}
\]

\[
k = \frac{2}{\delta \varepsilon^2} \text{ and Chebyshev}
\]

\[
Pr[|Y - \mu| \geq \varepsilon F_2] \leq \delta
\]
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$.

$$(E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2.)$$

Output average. $Y = \frac{1}{k} \sum_i Z_i^2$

$$E[Y] = \frac{1}{k} \sum E[Z_i^2] = F_2$$

$$\Var(cX) = c^2 \Var(X)$$

$$\Var(X + Y) = \Var(X) + \Var(Y); \text{ independent } X \text{ and } Y$$

$$\Var(Y) = \frac{1}{k^2} \sum \Var(Z_i^2) = \frac{2F_2^2}{k}$$

$$k = \frac{2}{\delta \varepsilon^2} \text{ and Chebyshev}$$

$$\Pr[|Y - \mu| \geq \varepsilon F_2] \leq \delta$$

Space:
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$.

$(E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2.)$

Output average. $Y = \frac{1}{k} \sum_i Z_i^2$

$E[Y] = \frac{1}{k} \sum E[Z_i^2] = F_2$

$Var(cX) = c^2 Var(X)$

$Var(X + Y) = Var(X) + Var(Y)$; independent $X$ and $Y$

$Var(Y) = \frac{1}{k^2} \sum_i Var(Z_i^2) = \frac{2F_2^2}{k}$

$k = \frac{2}{\delta \varepsilon^2}$ and Chebyshev

$Pr[|Y - \mu| \geq \varepsilon F_2] \leq \delta$

Space: $O(\frac{\log n}{\varepsilon^2 \delta})$. 
Independent trials.

Run Core Alg $k$ times. $Z_1, \ldots, Z_k$.

$\left( E[Z_i^2] = F_2 \ Var(Z_i^2) \leq 2F_2^2 \right)$

Output average. $Y = \frac{1}{k} \sum_i Z_i^2$

$E[Y] = \frac{1}{k} \sum E[Z_i^2] = F_2$

$Var(cX) = c^2 Var(X)$

$Var(X + Y) = Var(X) + Var(Y)$; independent $X$ and $Y$

$Var(Y) = \frac{1}{k^2} \sum_i Var(Z_i^2) = \frac{2F_2^2}{k}$

$k = \frac{2}{\delta \varepsilon^2}$ and Chebyshev

$Pr[|Y - \mu| \geq \varepsilon F_2] \leq \delta$

Space: $O\left(\frac{\log n}{\varepsilon^2 \delta}\right)$.

Could get $O\left(\frac{\log n \log \frac{1}{\delta}}{\varepsilon^2}\right)$ using a Central Limit Theorem.
See you on Tuesday.