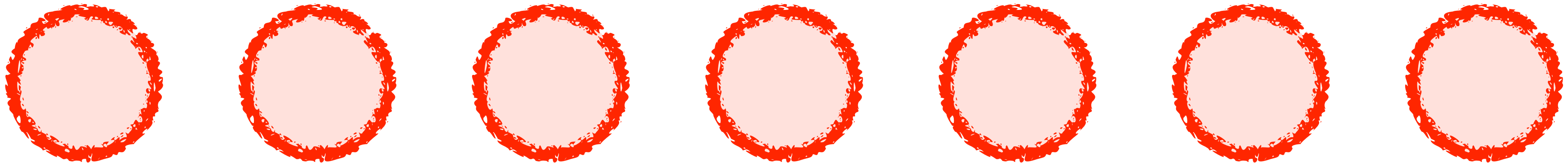
Advanced Algorithms

Balls into Bins

尹一通 Nanjing University, 202 Fall

Balls into Bins

n balls



uniform & independent

m bins

random function $f : [n] \to [m]$

birthday, coupon collector, occupancy, ...

Random Function

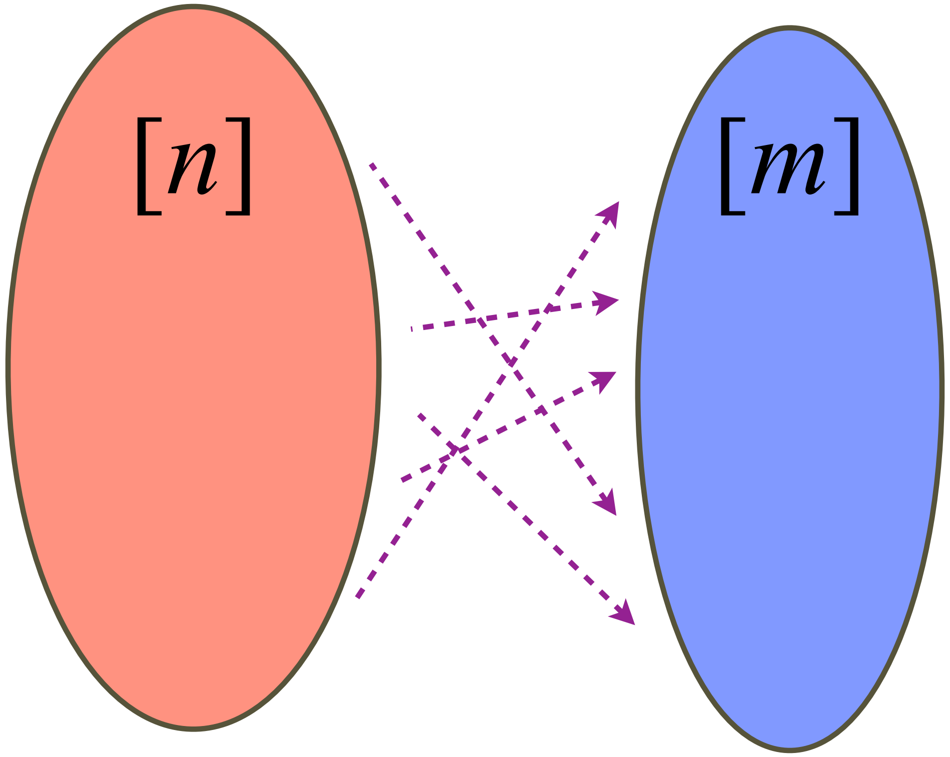
- n balls into m bins:

$$\Pr[\text{assignment}] = \frac{1}{m} \cdots \frac{1}{m} = \frac{1}{m^n}$$

- uniform random function:

$$\Pr[f] = \frac{1}{\left|[n] \to [m]\right|} = \frac{1}{m^n}$$

1-1	birthday
on-to	coupon collector
pre-image size	occupancy



uniform random function

$$f : [n] \to [m]$$

Birthday Paradox

Paradox:

(i) a statement that leads to a contradiction;

(ii) a situation which defies intuition.

In a class of m>57 students, with >99% probability, there are two students with the same birthday.

Assumption: birthdays are uniformly & independently distributed.

n balls are thrown into m bins:

event $\mathscr{E}$: each bin receives ≤ 1 balls

Birthday Paradox

n balls are thrown into m bins:

event $\mathscr{E}$: each bin receives ≤ 1 balls

$$\Pr[\mathscr{E}] = \frac{\left|[n] \xrightarrow{1-1} [m]\right|}{\left|[n] \to [m]\right|} = \frac{m(m-1)\cdots(m-n+1)}{m^n}$$

$$= \prod_{i=0}^{n-1}\left(1 - \frac{i}{m}\right)$$

Birthday Paradox

n balls are thrown into m bins:
event $\mathscr{E}$: each bin receives ≤ 1 balls

Suppose that balls are thrown one-by-one:

$$\Pr[\mathscr{E}] = \Pr[\text{all } n \text{ balls are thrown into ditinct bins}]$$

chain rule

$$= \prod_{i=1}^{n} \Pr[\text{the } i\text{th ball is thrown into an empty bin} \mid \text{first } i-1 \text{ balls are thrown into ditinct bins}]$$

$$= \prod_{i=1}^{n}\left(1 - \frac{i-1}{m}\right) = \prod_{i=0}^{n-1}\left(1 - \frac{i}{m}\right)$$

Birthday Paradox

n balls are thrown into m bins:

event $\mathscr{E}$: each bin receives ≤ 1 balls

(Taylor: $1 - x \approx \mathrm{e}^{-x}$ for $x = o(1)$)

$$\Pr[\mathscr{E}] = \prod_{i=0}^{n-1}\left(1 - \frac{i}{m}\right) \approx \prod_{i=0}^{n-1} \mathrm{e}^{-\frac{i}{m}} \approx \mathrm{e}^{-n^2/2m}$$

Formally: $\mathrm{e}^{-(1+o(1))n^2/2m} \leq \prod_{i=0}^{n-1}\left(1 - \frac{i}{m}\right) \leq \mathrm{e}^{-(1-o(1))n^2/2m}$

(assuming $n \ll m$)

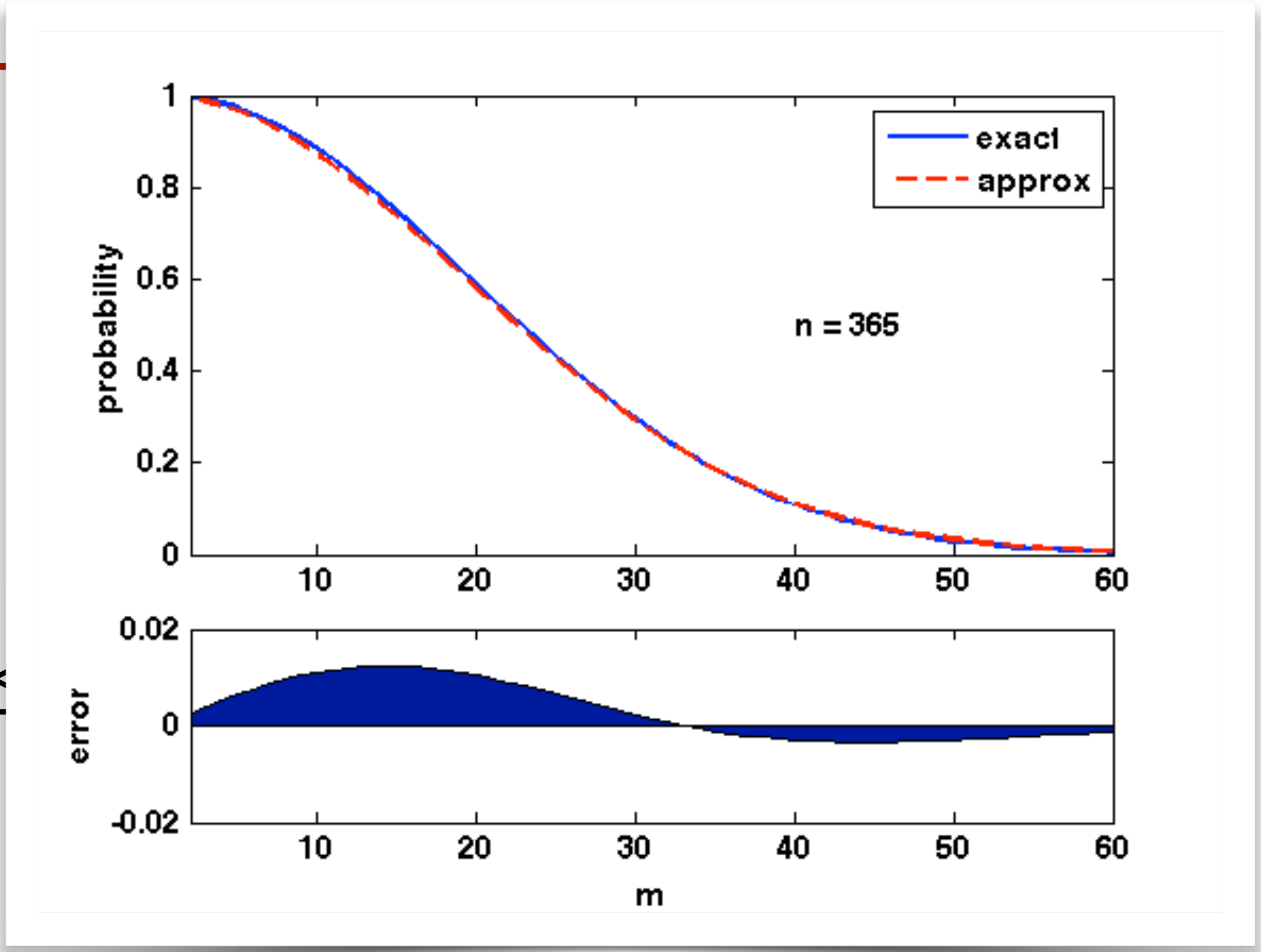
when $n = \sqrt{2m \ln \frac{1}{p}} \implies \Pr[\mathscr{E}] = (1 \pm o(1))p$

Birthday Paradox

n balls are thrown into m bins:

event $\mathscr{E}$: each bin receives ≤ 1 balls

$$\Pr[\mathscr{E}] = \prod_{i=0}^{n-1}\left(1-\frac{i}{m}\right)$$



Formally: $e^{-(1+o(1))n^2/2m}$

(assuming $n \ll m$)

when $n = \sqrt{2m\ln\frac{1}{p}} \implies \Pr[\mathscr{E}] = (1 \pm o(1))p$

Data Structure for Set

Data: a set S of n items $x_1, x_2, \ldots, x_n \in U = [N]$

Query: an item $x \in U$

Determine whether $x \in S$.

- **Space cost:** size of data structure (in bits)
  - entropy of a set: $\mathrm{O}(n \log N)$ bits (when $N \gg n$)
- **Time cost:** time to answer a query (in memory accesses)
- Balanced tree: $\mathrm{O}(n \log N)$ space, $\mathrm{O}(\log n)$ time
- Perfect hashing: $\mathrm{O}(n \log N)$ space, $\mathrm{O}(1)$ time

Perfect Hashing

$S = \{a, b, c, d, e, f\} \subseteq [N]$ of size n

uniform random $\boxed{h}$ $[N] \to [m]$

no collision

Pr[perfect] $\approx \mathrm{e}^{-n^2/2m} > 1/2$

$m = n^2$

Birthday

Table T:

e	b		d		f		c	a	

SUHA: Simple Uniform Hash Assumption

```
Query(x):
    retrieve hash function $h$;
    check whether $T[h(x)] = x$;
```

Universal Hashing

Universal Hash Family (Carter and Wegman 1979):

A family $\mathscr{H}$ of hash functions in $U \to [m]$ is **k-universal** if for any distinct $x_1, \ldots, x_k \in U$,

$$\Pr_{h \in \mathscr{H}} \left[h(x_1) = \cdots = h(x_k) \right] \le \frac{1}{m^{k-1}}.$$

Moreover, $\mathscr{H}$ is **strongly k-universal** (k-wise independent) if for any distinct $x_1, \ldots, x_k \in U$ and any $y_1, \ldots, y_k \in [m]$,

$$\Pr_{h \in \mathscr{H}} \left[\bigwedge_{i=1}^{k} h(x_i) = y_i \right] = \frac{1}{m^k}.$$

k-Universal Hash Family

hash functions $h : U \to [m]$

- **Linear congruential hashing**:
  - Represent $U \subseteq \mathbb{Z}_p$ for sufficiently large prime p
  - $h_{a,b}(x) = ((ax + b) \bmod p) \bmod m$
  - $\mathscr{H} = \left\{ h_{a,b} \mid a \in \mathbb{Z}_p \backslash \{0\}, b \in \mathbb{Z}_p \right\}$

Theorem:

The linear congruential family $\mathscr{H}$ is 2-wise independent.

- **Degree-k polynomial in finite field with random coefficients**
- Hashing between binary fields: $GF(2^w) \to GF(2^l)$

$$h_{a,b}(x) = \texttt{(a*x+b)>>(w-l)}$$

Birthday Paradox (pairwise independence)

n balls are thrown into m bins: by 2-universal hashing
event $\mathscr{E}$: each bin receives ≤ 1 balls

- Location of n balls: $X_1, X_2, \ldots, X_n \in [m]$
- Total # of collisions:

$$Y = \sum_{i<j} I[X_i = X_j]$$

- Linearity of expectation:

$$\mathbb{E}[Y] = \sum_{i<j} \Pr[X_i = X_j] \leq \binom{n}{2}\frac{1}{m}$$

2-universal

when
$n \leq \sqrt{2m\epsilon}$

- Markov's inequality: $\Pr[\neg\mathscr{E}] = \Pr[Y \geq 1] \leq \mathbb{E}[Y] \leq \epsilon$

Perfect Hashing

$S = \{a, b, c, d, e, f\} \subseteq [N]$ of size n

2-universal $\boxed{h}$ $[N] \to [m]$

Pr[*imperfect*] $= \dfrac{n(n-1)}{2m}$

m

Table T:

e	b		d		f		c	a	

For 2-universal family $\mathscr{H}$ from $[N]$ to $[m]$, if $m > \binom{n}{2}$, for any $S \subseteq [N]$ of size n, there is an $h \in \mathscr{H}$ that cause no collisions over S.

Query(x):

```
    retrieve hash function $h$;
    check whether $T[h(x)] = x$;
```

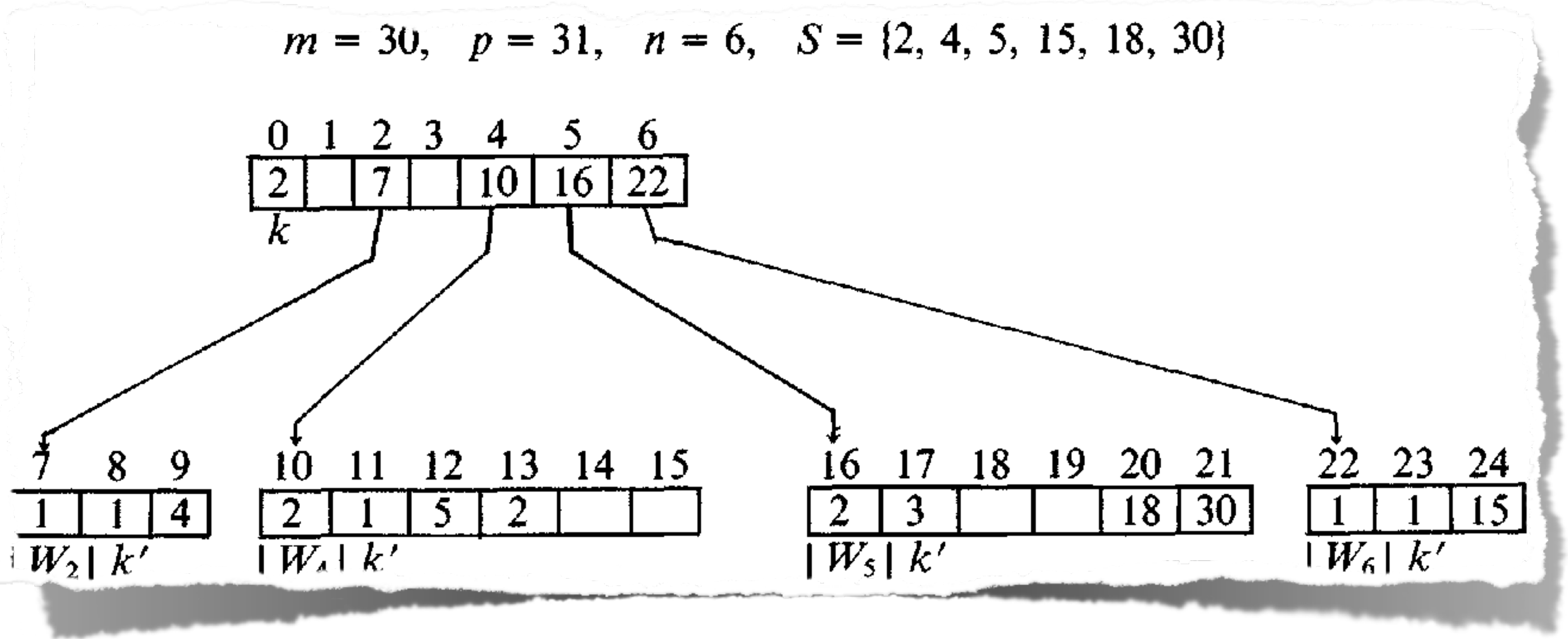
FKS Perfect Hashing

(Fredman, Komlós, Szemerédi, 1984)

Data: a set S of n items $x_1, x_2, \ldots, x_n \in U = [N]$

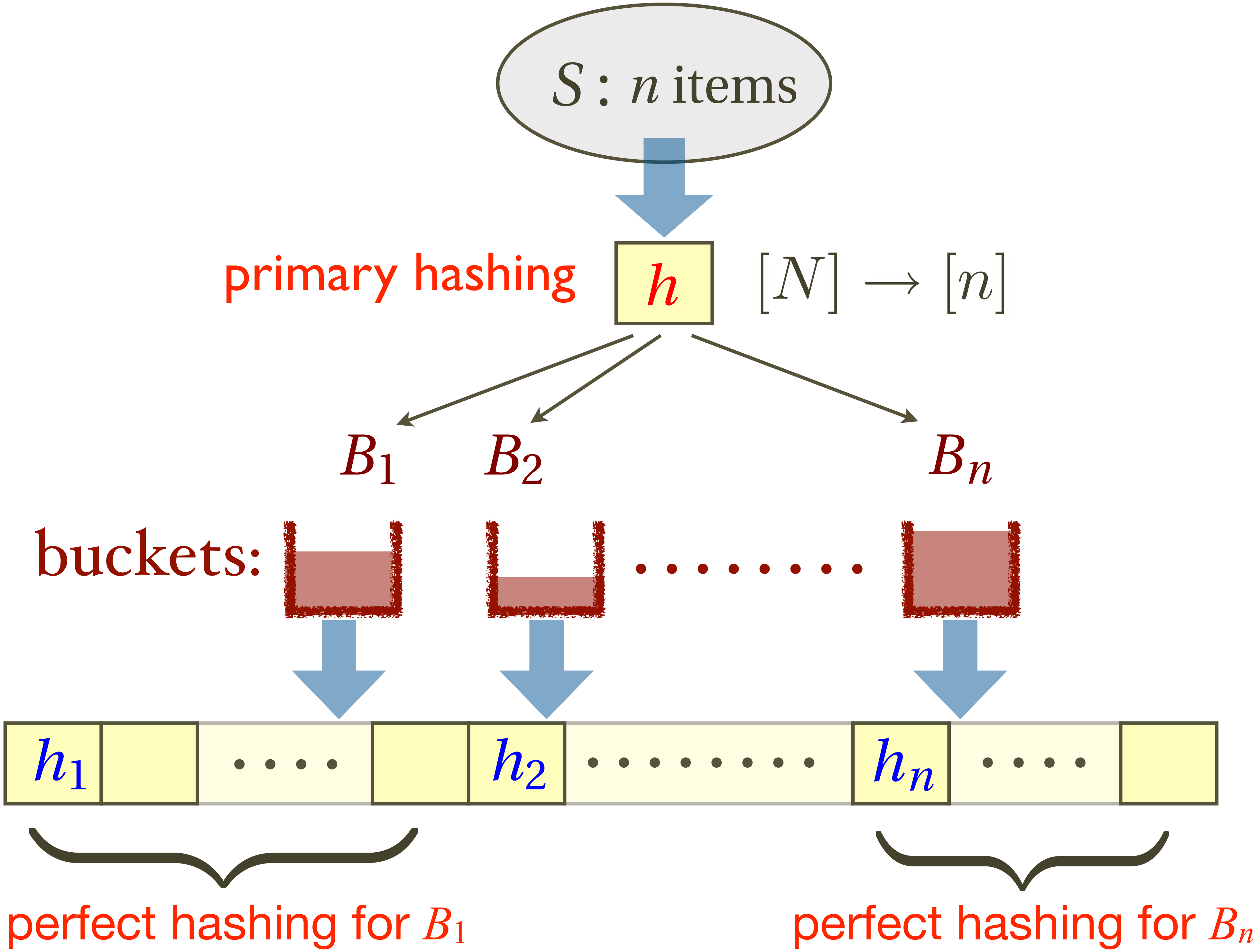
Query: an item $x \in U$

Determine whether $x \in S$.



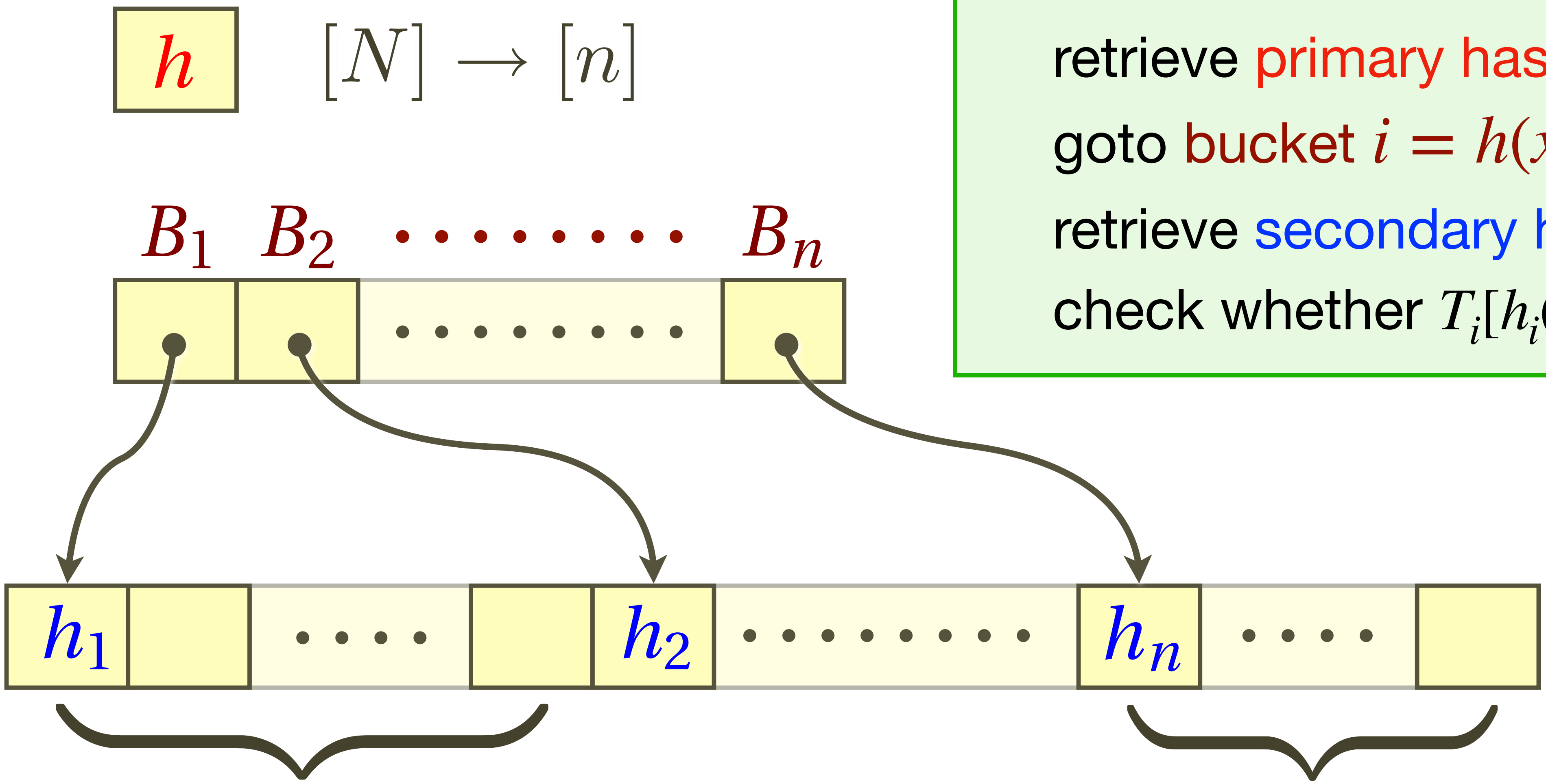
- **Space cost:** $O(n)$ words (each of $O(\log N)$ bits)
- **Time cost:** $O(1)$ for each query in the worst case

FKS Perfect Hashing



FKS Perfect Hashing

Set $S \subseteq [N]$ of size n



Query(x):

```
    retrieve primary hash $h$;
    goto bucket $i = h(x)$;
    retrieve secondary hash $h_i$;
    check whether $T_i[h_i(x)] = x$;
```

perfect hashing for B_1
using space $|B_1|^2$

perfect hashing for B_n
using space $|B_n|^2$

- $\exists\, h_1, \ldots, h_n$ from 2-universal family s.t. h_i is perfect for B_i for all i

Collision Number

n balls are thrown into m bins by 2-universal hashing

- Location of n bins: $X_1, X_2, \ldots, X_n \in [m]$

Collision #: $Y = \sum_{i<j} I[X_i = X_j]$

- Linearity of expectation:

$$\mathbb{E}[Y] = \sum_{i<j} \Pr[X_i = X_j] \leq \binom{n}{2} \frac{1}{m}$$

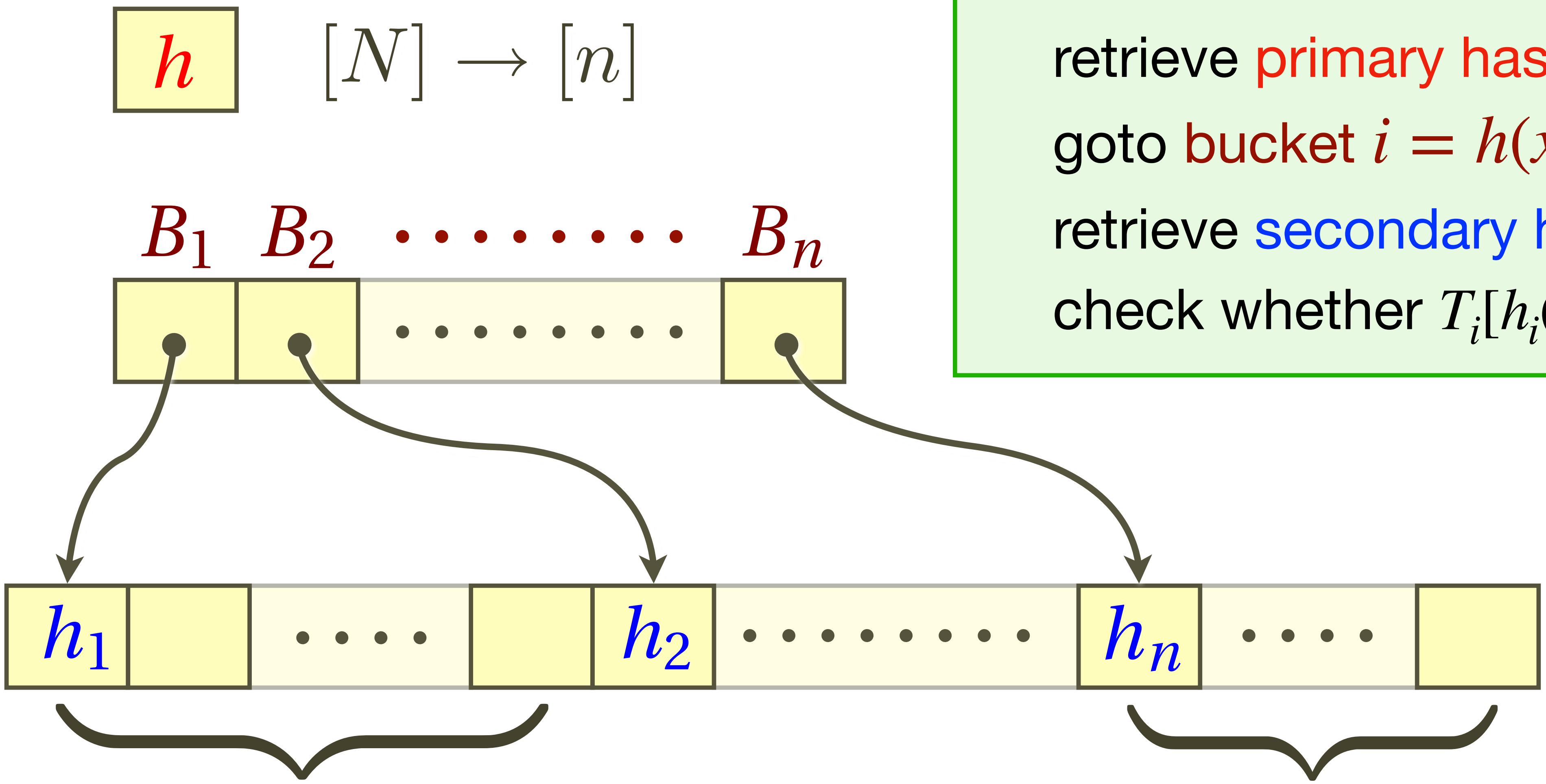
2-universal

- Size of the i-th bin: $|B_i|$

$$Y = \sum_{i=1}^{n} \binom{|B_i|}{2} = \frac{1}{2} \sum_{i=1}^{n} |B_i|(|B_i| - 1) \implies \mathbb{E}\left[\sum_{i=1}^{n} |B_i|^2\right] = \frac{n(n-1)}{m} + n$$

FKS Perfect Hashing

Set $S \subseteq [N]$ of size n



perfect hashing for B_1 using space $|B_1|^2$

perfect hashing for B_n using space $|B_n|^2$

Query(x):

retrieve primary hash h;

goto bucket $i = h(x)$;

retrieve secondary hash h_i;

check whether $T_i[h_i(x)] = x$;

- $\exists h$ from a 2-universal family s.t. the total space cost is O(n)

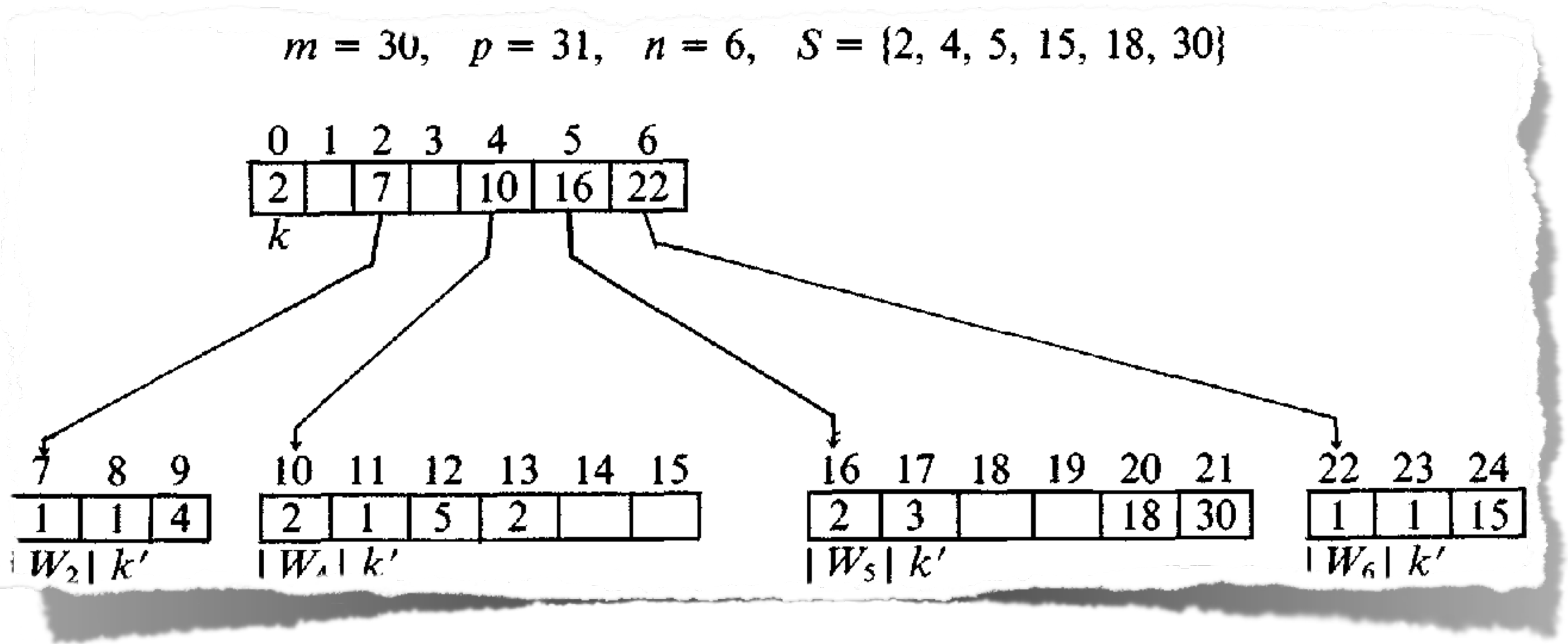
FKS Perfect Hashing

(Fredman, Komlós, Szemerédi, 1984)

Data: a set S of n items $x_1, x_2, \ldots, x_n \in U = [N]$

Query: an item $x \in U$

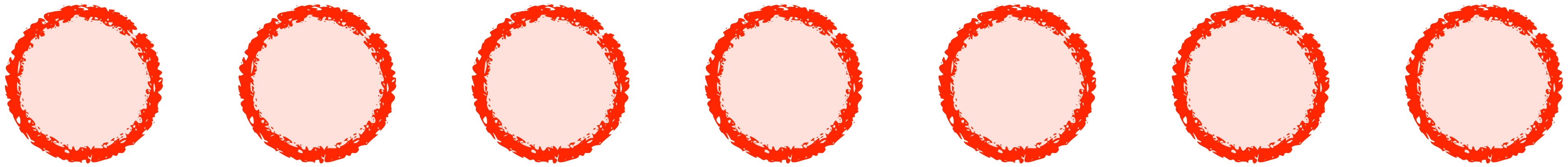
Determine whether $x \in S$.



- $O(n \log N)$ space, $O(1)$ time in the worst case
- Dynamic version: [Dietzfelbinger, Karlin, Mehlhorn, Heide, Rohnert, Tarjan, 1984]

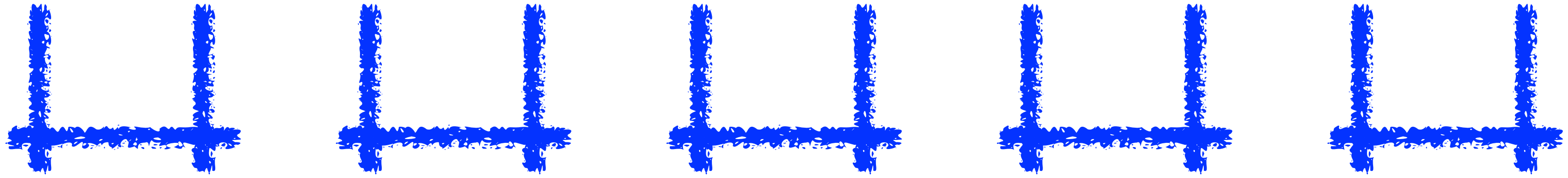
Balls into Bins

(Coupon Collector)



uniform & independent

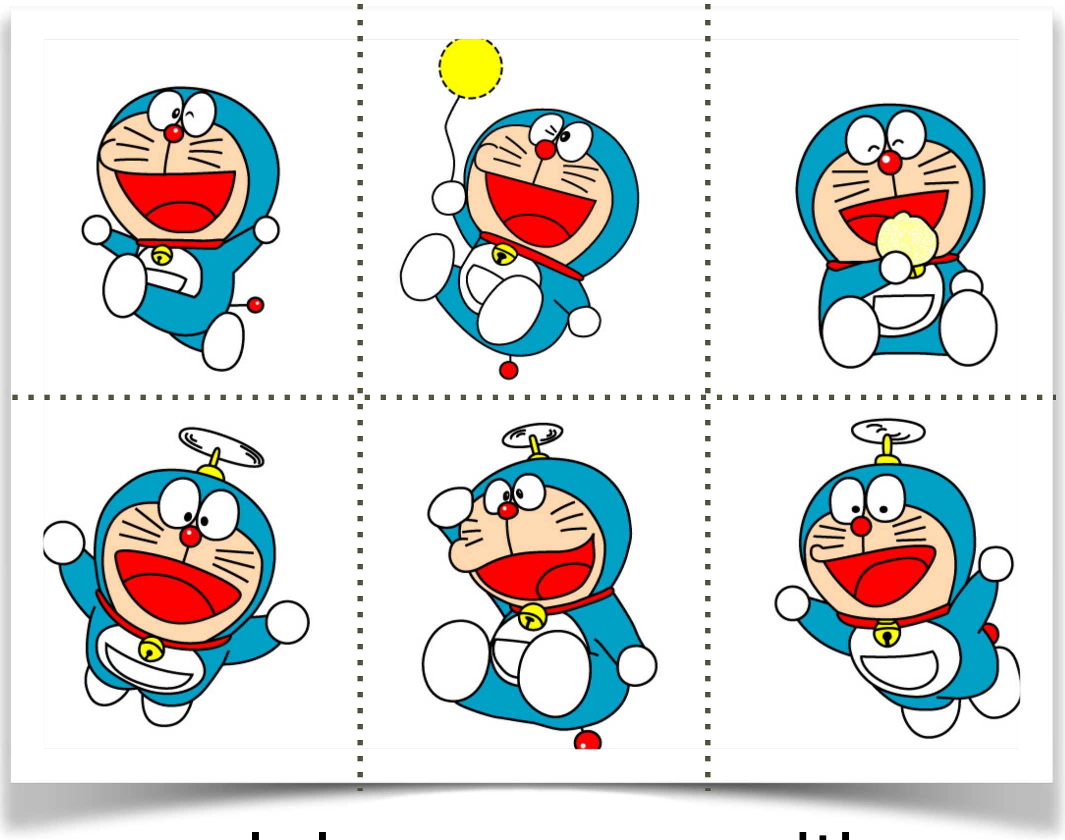
n bins



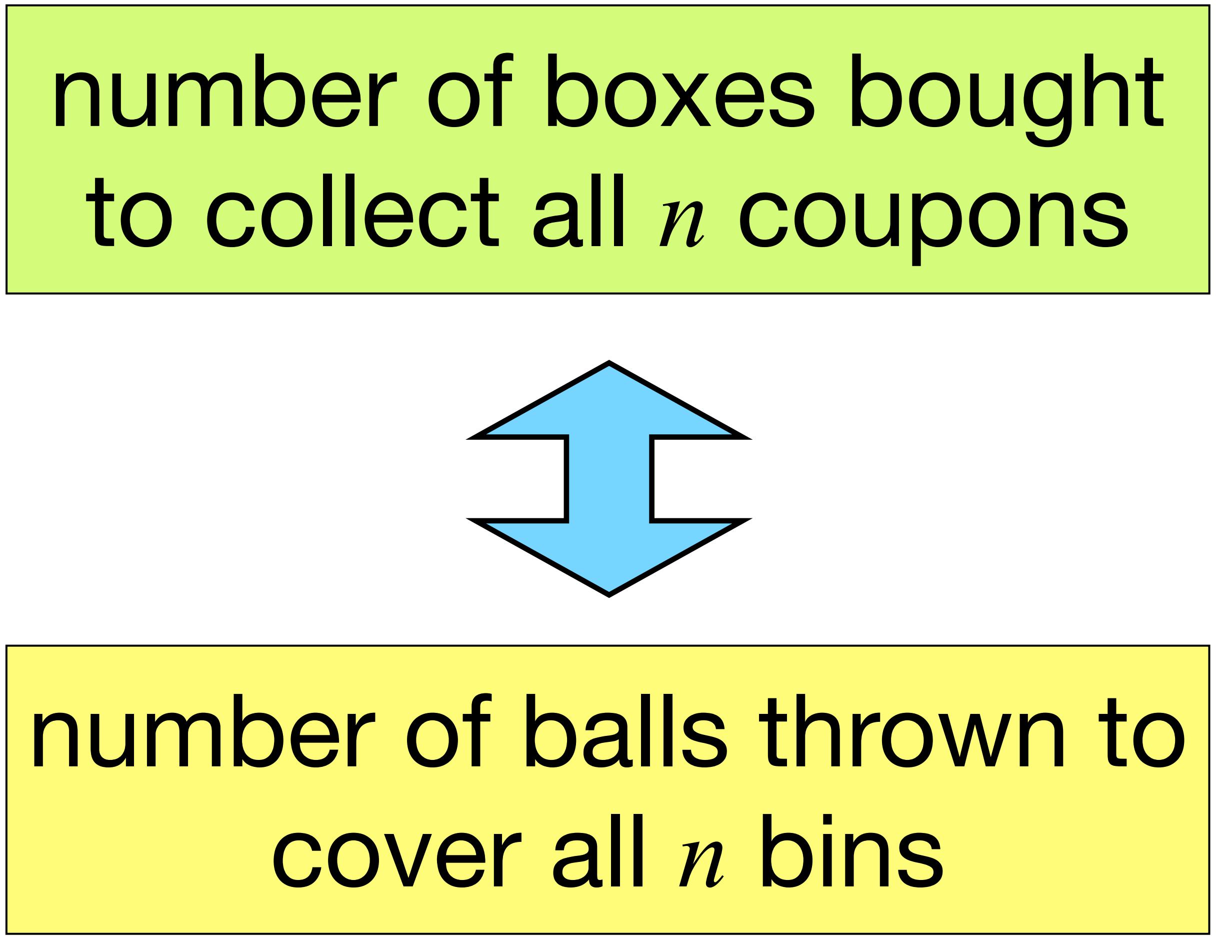
surjection (cover all bins)

Coupon Collector

coupons in cookie box



each box comes with a uniformly random coupon



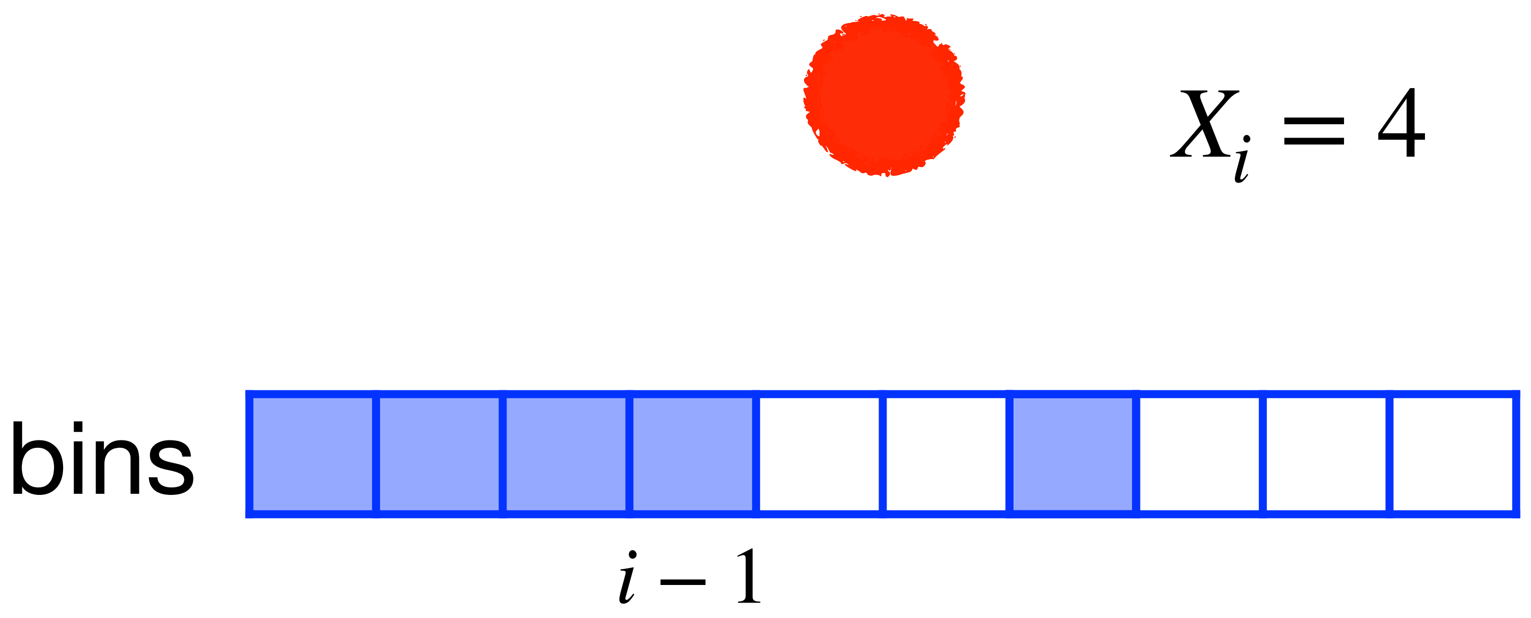
number of boxes bought to collect all n coupons

number of balls thrown to cover all n bins

Coupon Collector

$X:$ number of balls thrown to make all the n bins nonempty

$$X = \sum_{i=1}^{n} X_i$$



$X_i = 4$

X_i is geometric!

with $p_i = 1 - \dfrac{i-1}{n}$

$$\mathbb{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}$$

Coupon Collector

$X:$ number of balls thrown to make all the n bins nonempty

$X_i:$ number of balls thrown while there are exactly $(i\text{-}1)$ nonempty bins

$$X = \sum_{i=1}^{n} X_i$$

$$\mathbb{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}$$

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$

linearity of expectations

$$= \sum_{i=1}^{n} \frac{n}{n-i+1}$$

$$= n\sum_{i=1}^{n} \frac{1}{i}$$

Harmonic number

$$= nH(n)$$

expected $n \ln n + O(n)$ balls

Coupon Collector

X : number of balls thrown to make all the n bins nonempty

Theorem: For $c > 0$,

$$\Pr[\,X \geq n \ln n + cn\,] \leq \mathrm{e}^{-c}$$

Proof: For one bin, it misses all balls with probability

$$\left(1 - \frac{1}{n}\right)^{n \ln n + cn} = \left(1 - \frac{1}{n}\right)^{n(\ln n + c)}$$

$$< \mathrm{e}^{-(\ln n + c)}$$

$$< \frac{1}{n\mathrm{e}^{c}}$$

Coupon Collector

X : number of balls thrown to make all the n bins nonempty

Theorem: For $c > 0$,

$$\Pr[\,X \geq n\ln n + cn\,] \leq \mathrm{e}^{-c}$$

Proof: For one bin, it misses all balls with probability

$$< \frac{1}{n\mathrm{e}^{c}}$$

union bound!

$$\Pr[\,\exists \text{ a bin misses all balls}\,] \leq n\Pr[\,\text{first bin misses all bins}\,]$$

$$< \mathrm{e}^{-c}$$

Coupon Collector

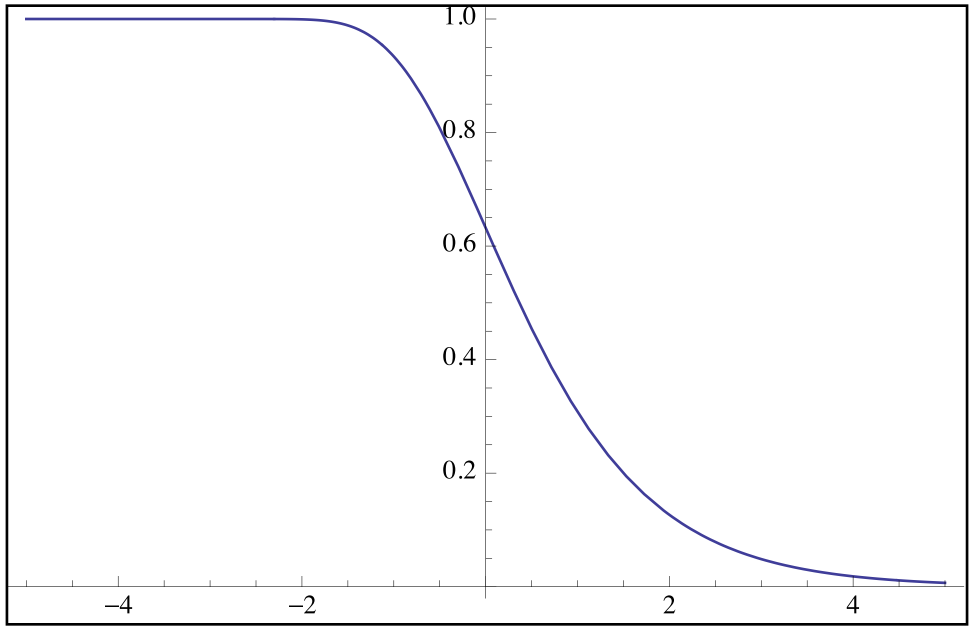
X : number of balls thrown to make all the n bins nonempty

Theorem: For $c > 0$,

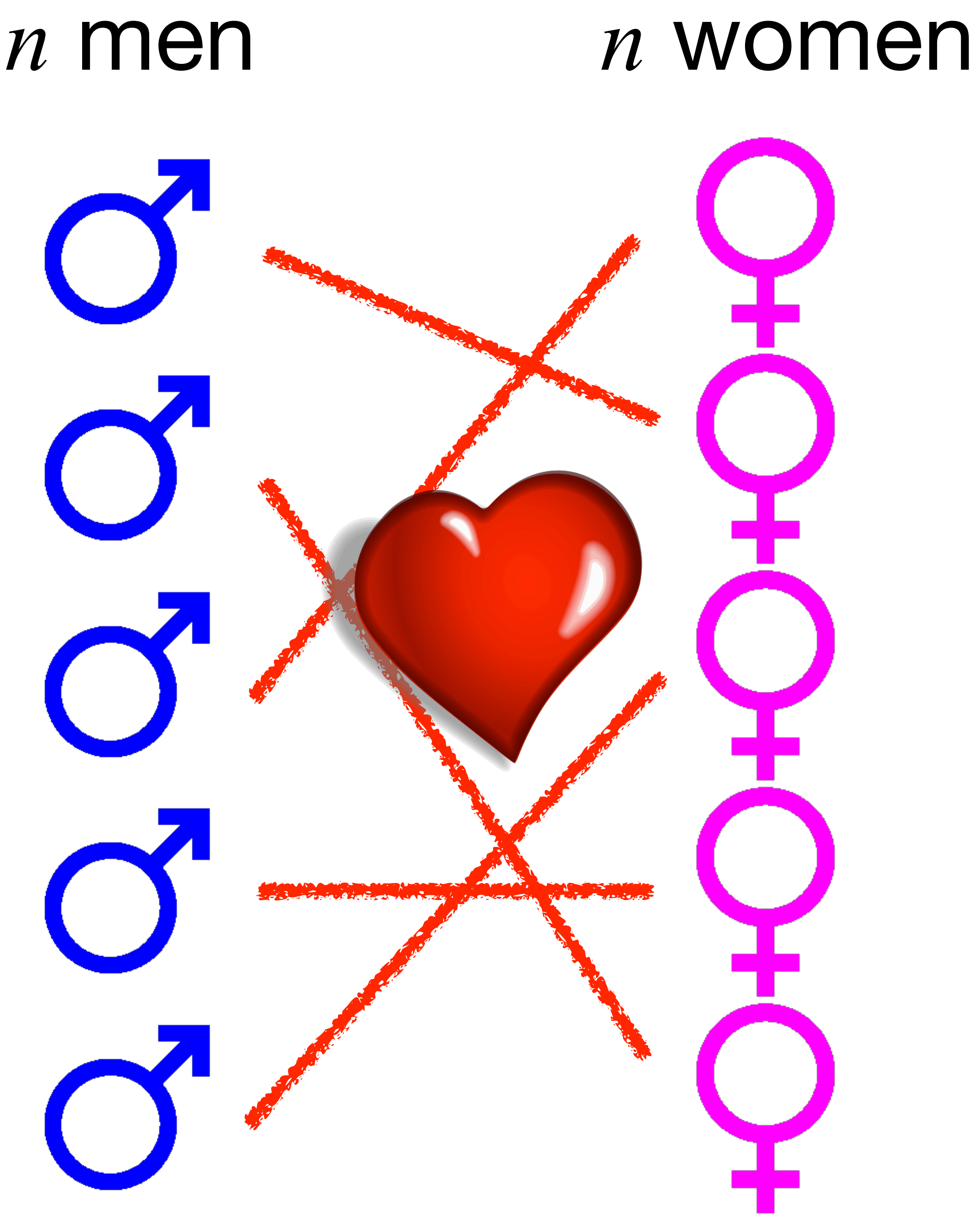
$$\Pr[\, X \geq n \ln n + cn \,] \leq \mathrm{e}^{-c}$$

a sharp threshold:

$$\lim_{n\to\infty} \Pr[X \geq n \ln n + cn] = 1 - \mathrm{e}^{-\mathrm{e}^{-c}}$$

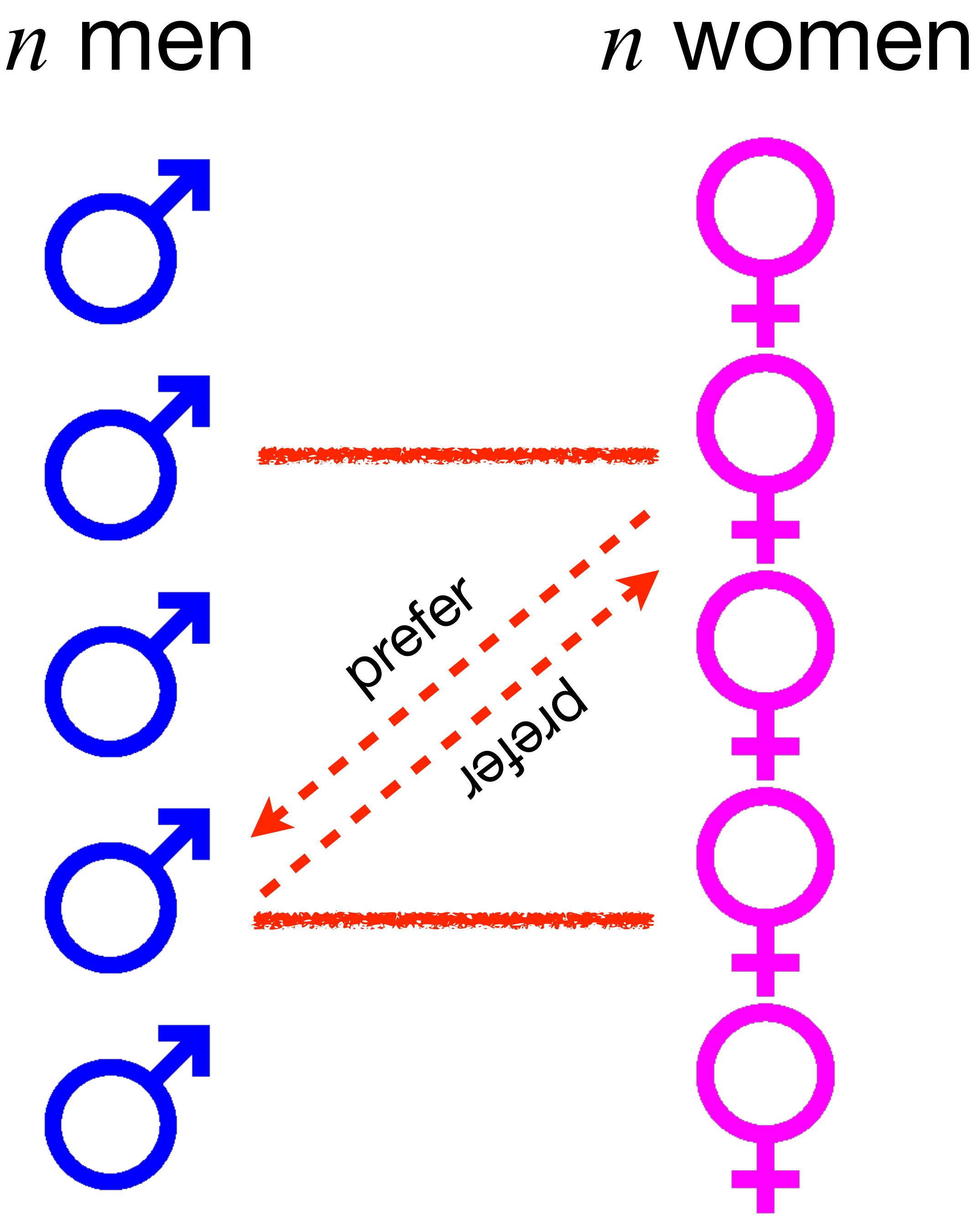


Stable Matching



- each man has a preference order of the n women;
- each woman has a preference order of the n men;
- solution: n couples
- Marriages are stable!

Stable Matching



n men

n women

prefer

prefer

unstable (blocking pair):

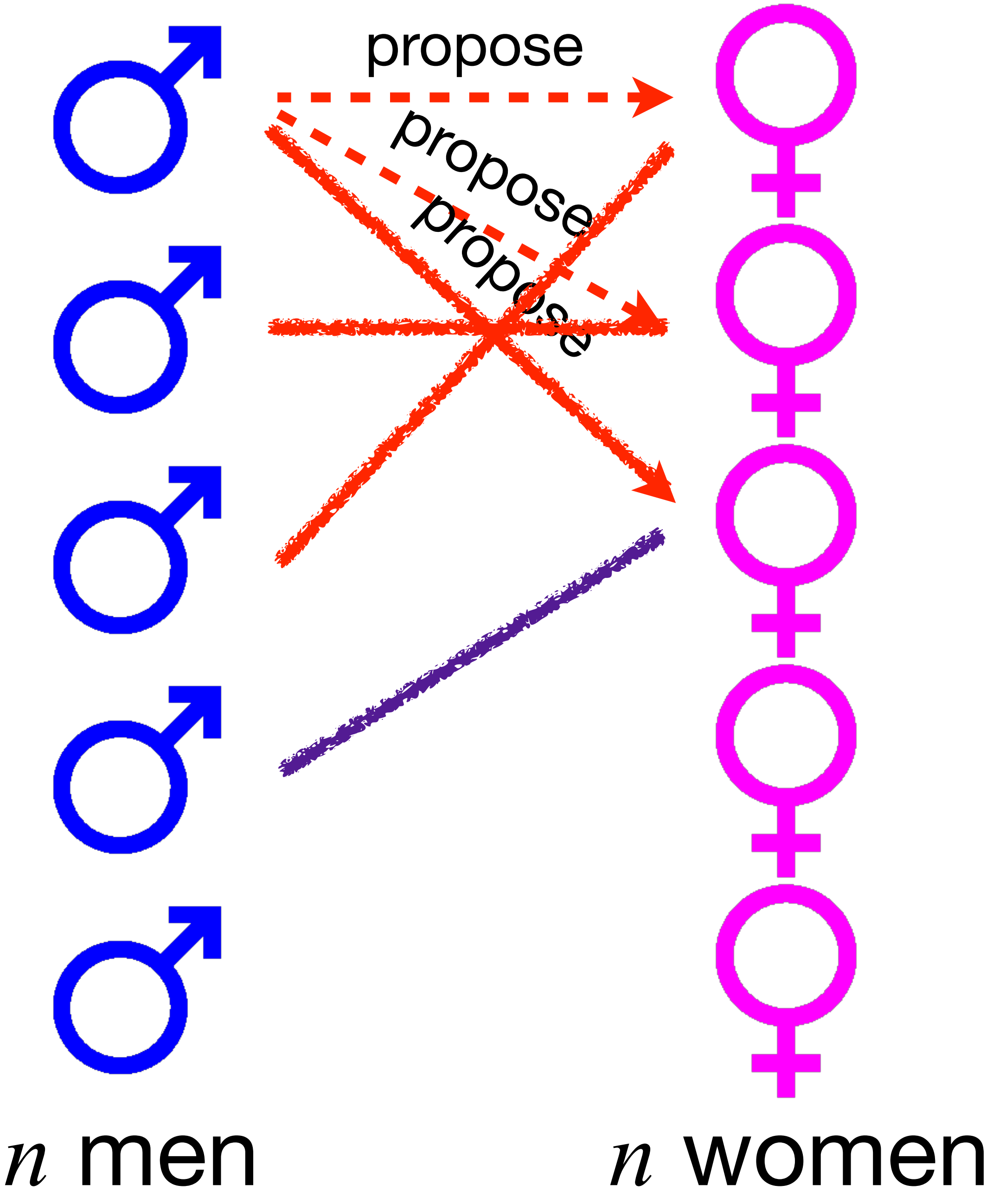
a man and a woman, who prefer each other to their current partners

stable: no blocking pairs

local optimum
fixed point
equilibrium
deadlock

Proposal Algorithm

(Gale-Shapley 1962)



Single man:

propose to the most preferable women who has not rejected him

Woman:

upon received a proposal: accept if she's single or married to a less preferable man (divorce!)

Proposal Algorithm

(Gale-Shapley 1962)

- woman: once got married always married

(will only switch to better men!)

- man: will only get worse ...

- once all women are married, the algorithm terminates, and the marriages are stable

- total number of proposals:

$$\leq n^2$$

Single man:

propose to the most preferable women who has not rejected him

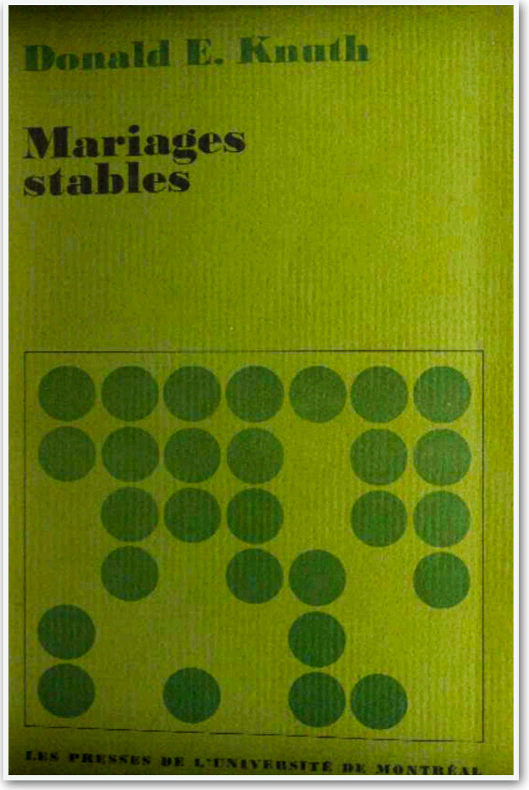
Woman:

upon received a proposal: accept if she's single or married to a less preferable man (divorce!)

Average-Case Performance

(Knuth 1976)

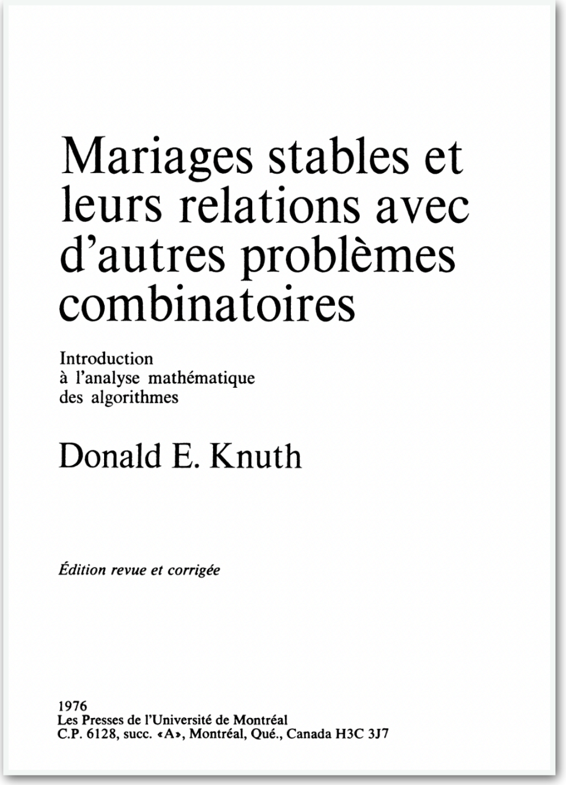
- Every man/woman has a uniform random permutation as preference list
- Expected total number of proposals?



Donald E. Knuth

Mariages stables

LES PRESSES DE L'UNIVERSITÉ DE MONTRÉAL



Mariages stables et leurs relations avec d'autres problèmes combinatoires

Introduction
à l'analyse mathématique
des algorithmes

Donald E. Knuth

Édition revue et corrigée

1976
Les Presses de l'Université de Montréal
C.P. 6128, succ. «A», Montréal, Qué., Canada H3C 3J7

Single man:

propose to the most preferable women who has not rejected him

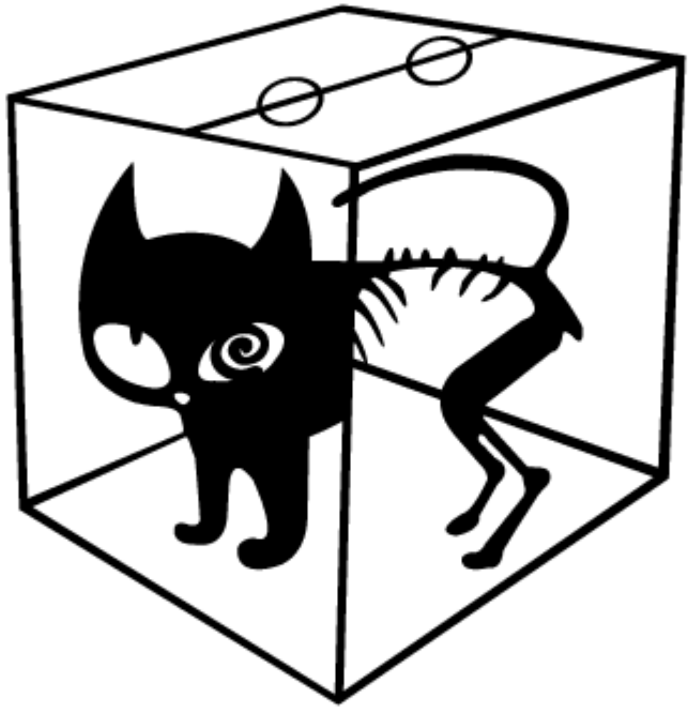
Woman:

upon received a proposal: accept if she's single or married to a less preferable man (divorce!)

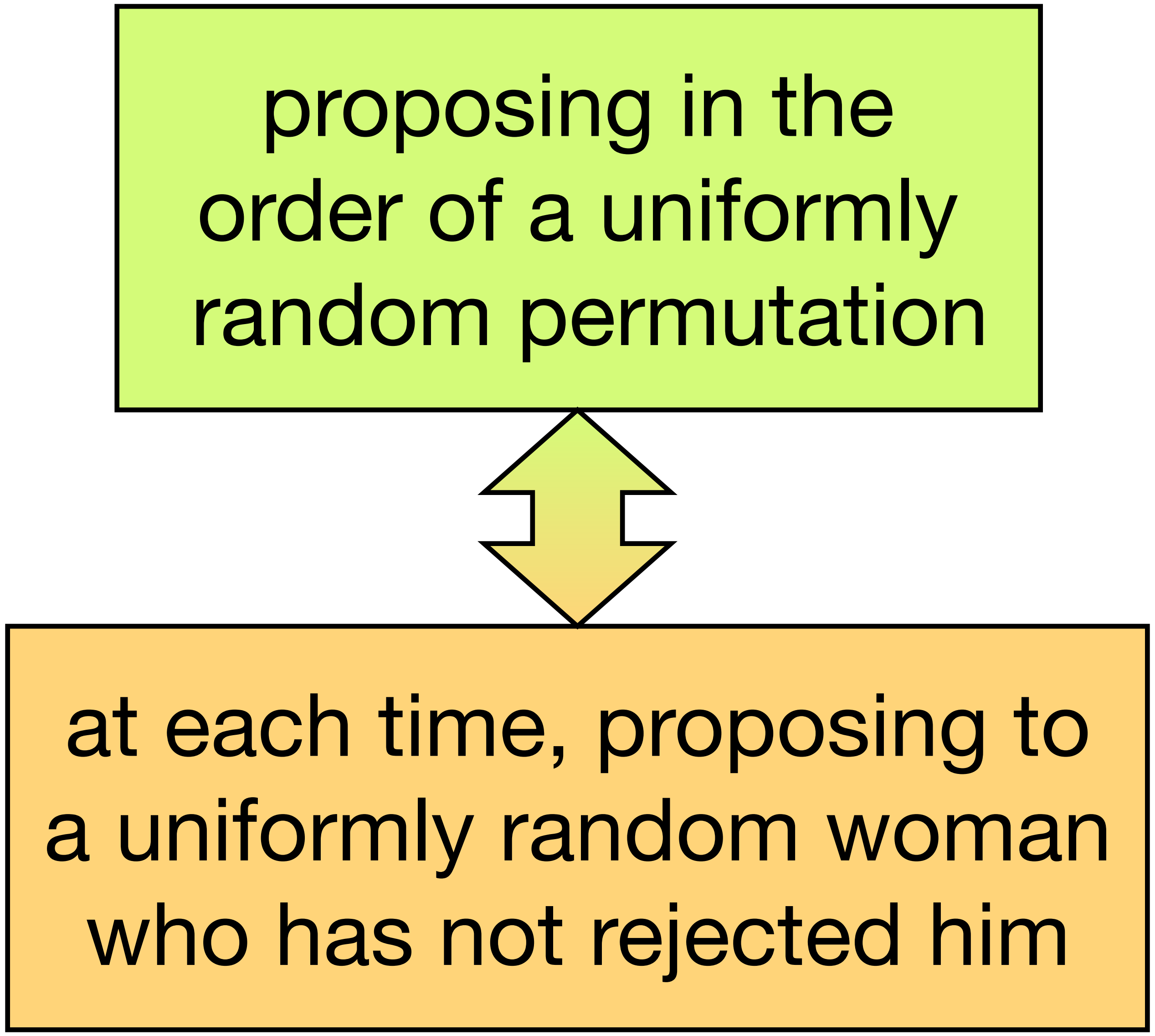
Principle of Deferred Decisions

Principle of deferred decision

The decision of random choice in the random input is deferred to the running time of the algorithm.



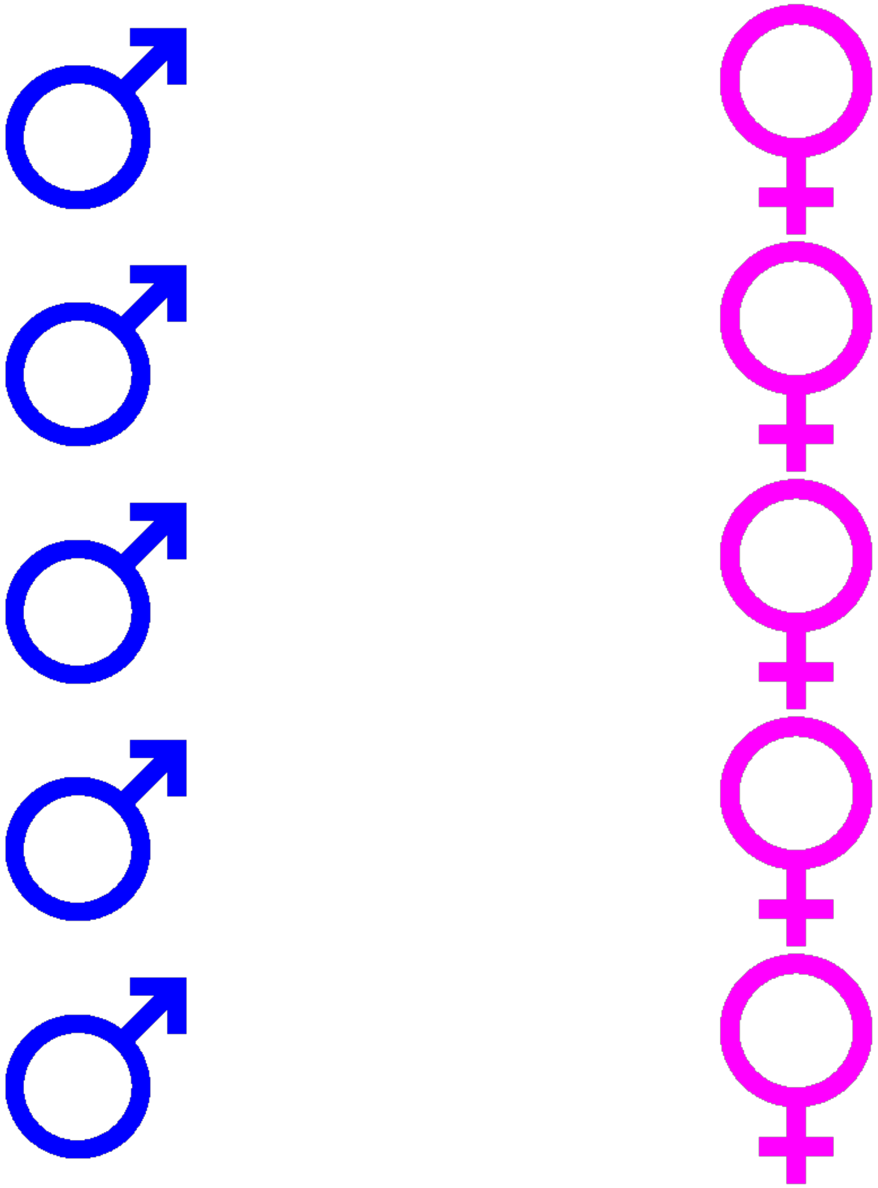
Principle of Deferred Decisions



proposing in the order of a uniformly random permutation

at each time, proposing to a uniformly random woman who has not rejected him

decisions of the inputs are deferred to the time when Alg accesses them



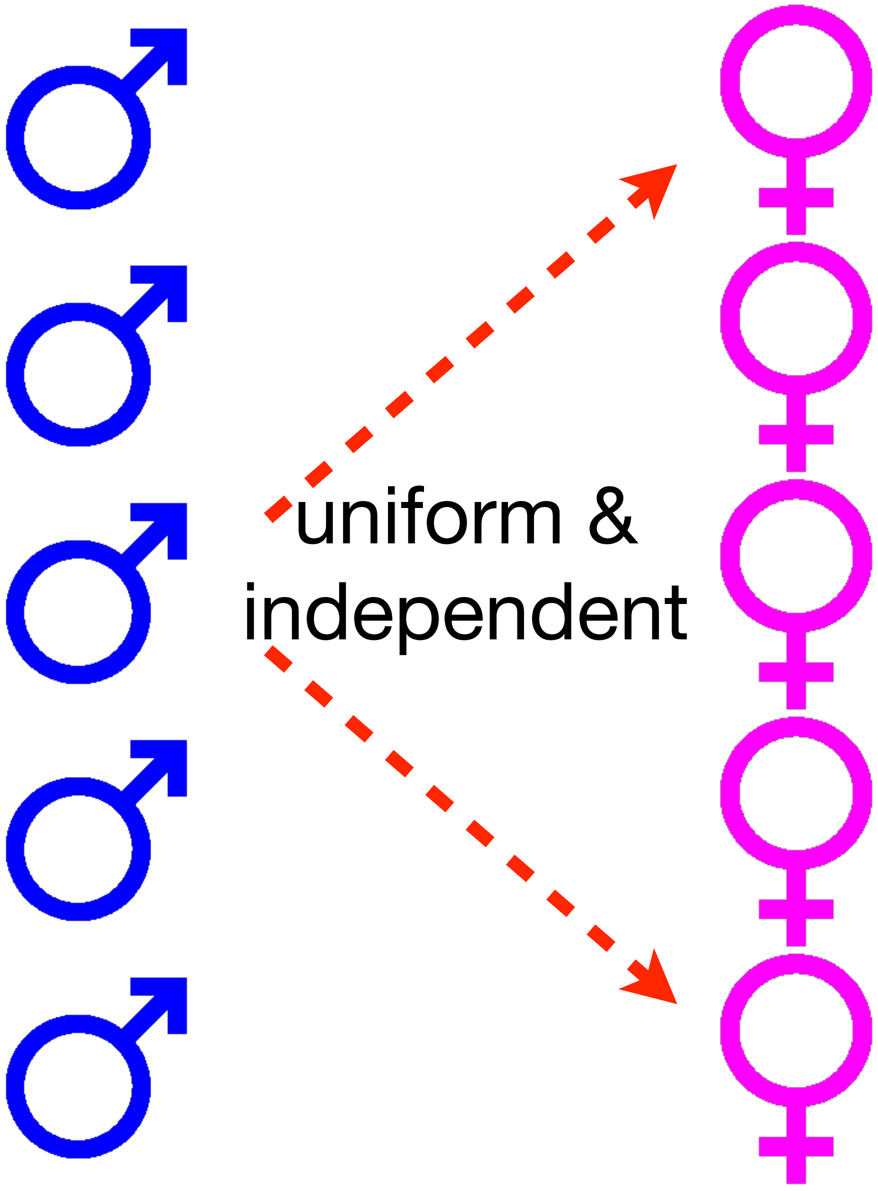
Stochastic Domination

at each time, proposing to a uniformly random woman who has not rejected him

$\leq$

at each time, proposing to a uniformly & independently random woman

the man forgot who had rejected him (!)



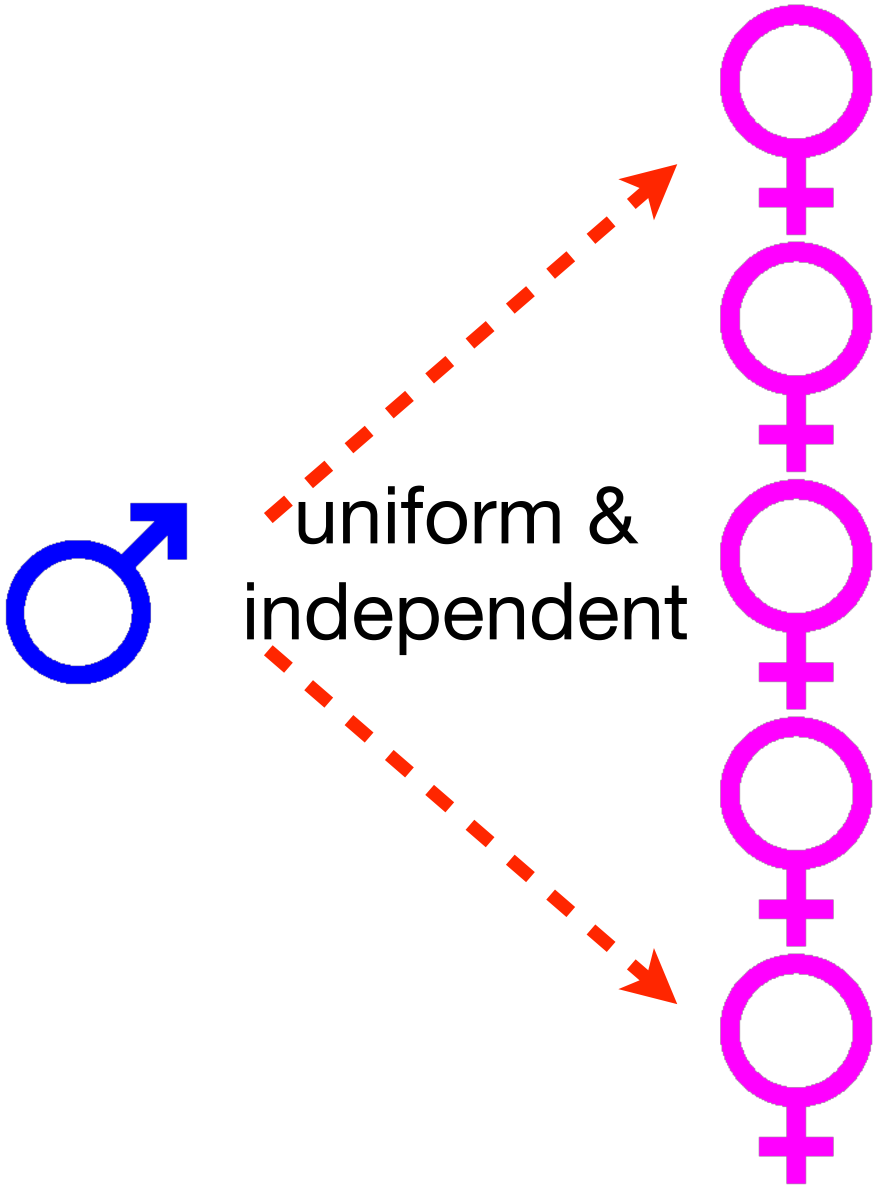
Average-Case Performance

- uniformly and independently proposing to n women

- Alg stops once all women got proposed.

- Coupon collector!

- Expected $n \ln n + O(n)$ proposals.


uniform &
independent