Randomized Algorithms

南京大学

尹一通

Balls-into-bins model

throw *m* balls into *n* bins
uniformly and independently

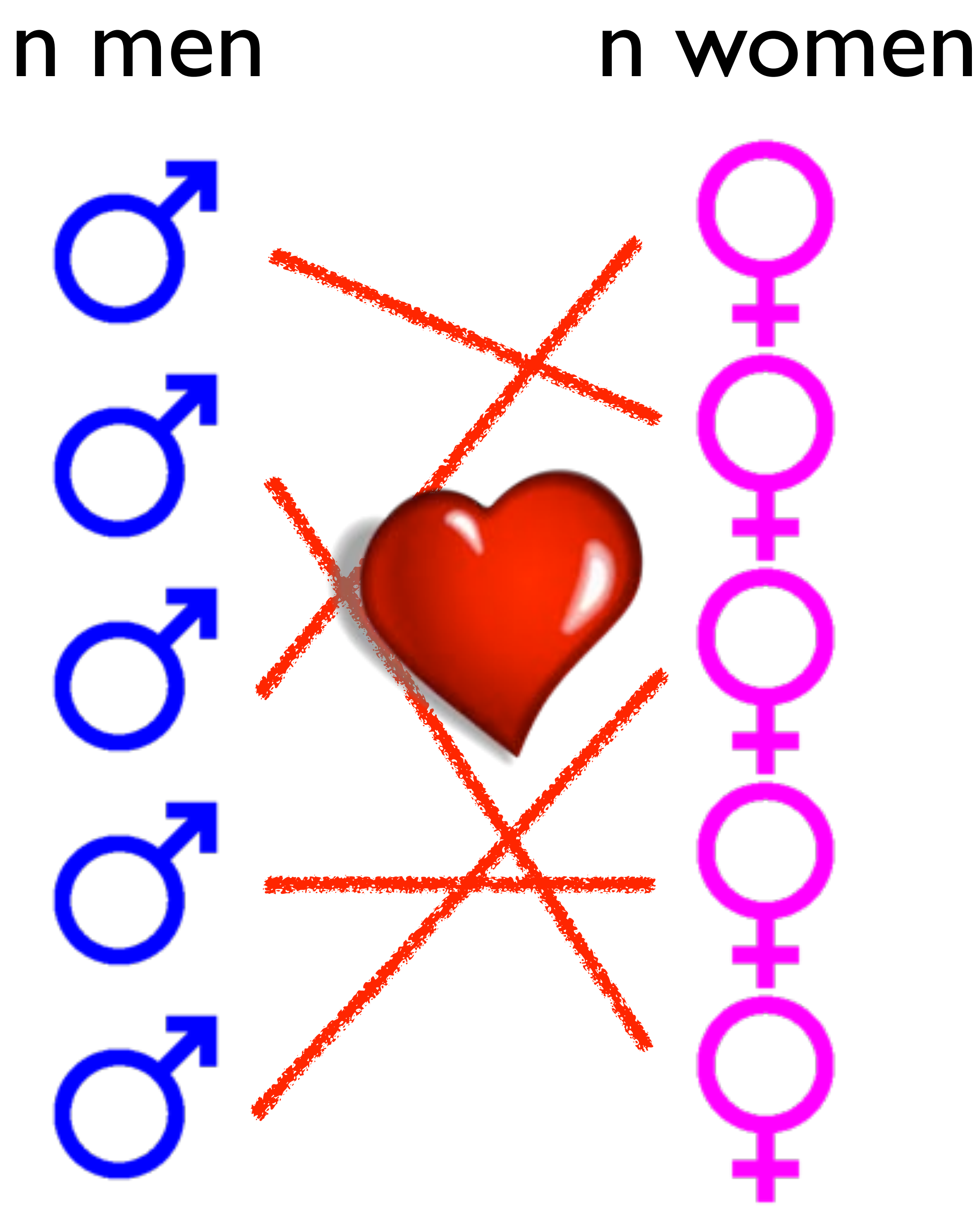
uniform random function
$$f : [m] \to [n]$$

1-1	birthday problem
on-to	coupon collector
pre-images	occupancy problem

- The threshold for being 1-1 is $m = \Theta(\sqrt{n})$.
- The threshold for being on-to is $m = n \ln n + O(n)$.
- The maximum load is

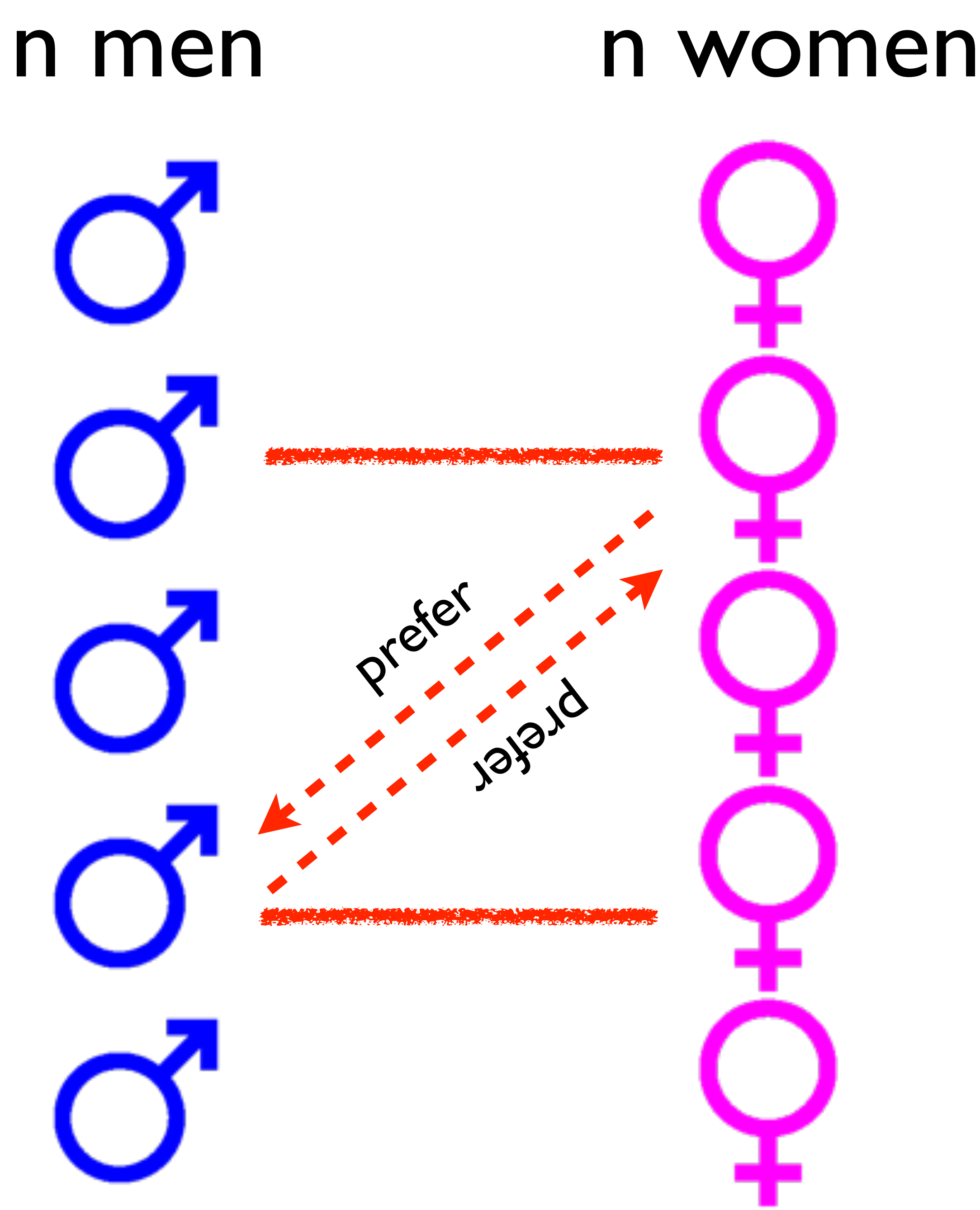
$$\begin{cases} O(\frac{\ln n}{\ln \ln n}) & \text{for } m = \Theta(n), \\ O(\frac{m}{n}) & \text{for } m = \Omega(n \ln n). \end{cases}$$

Stable Marriage



- each man has a preference order of the *n* women;
- each woman has a preference order of the *n* men;
- solution: *n* couples
- Marriage is stable!

Stable Marriage



unstable:

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

stability: local optimum

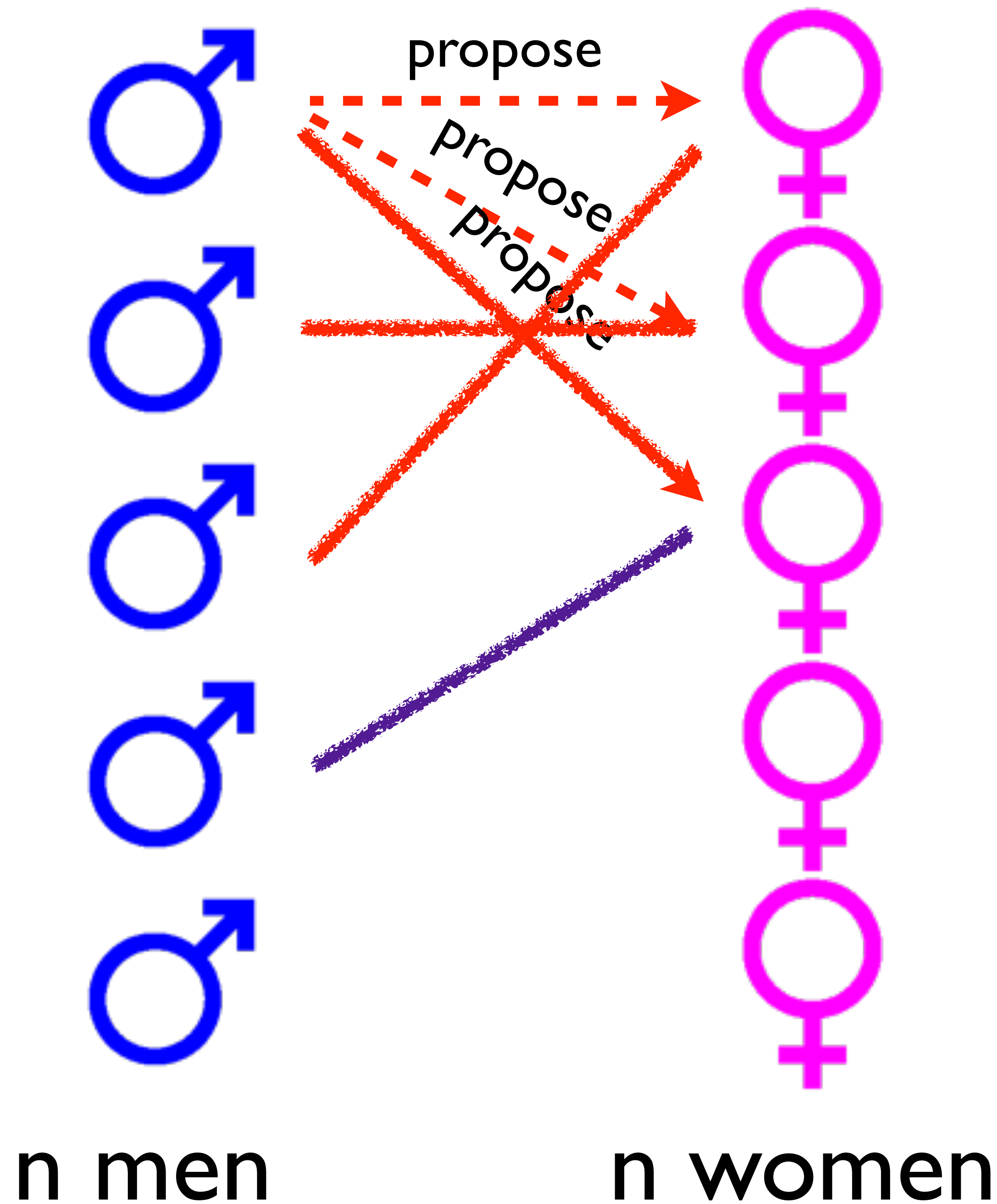
fixed point

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

- 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

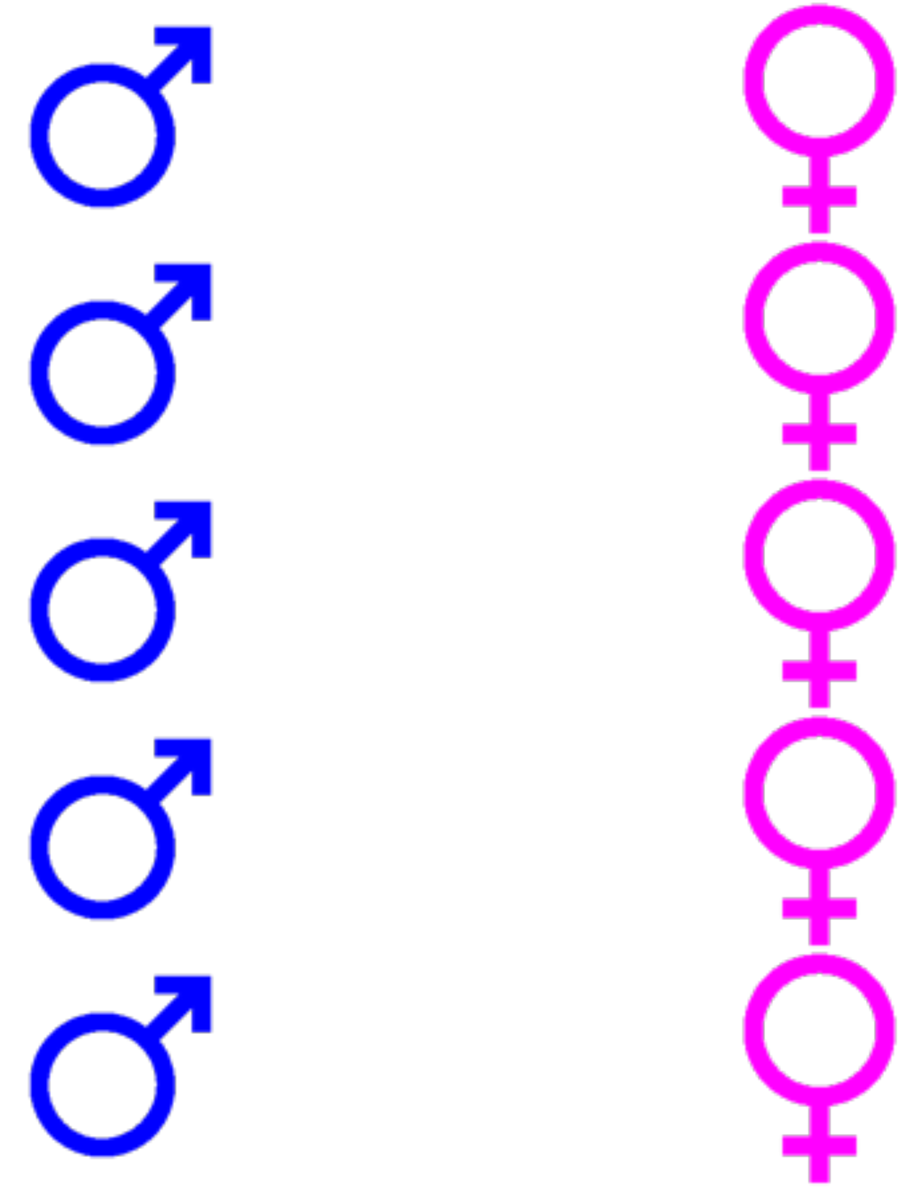
Woman:
upon received a proposal: accept if she's single or married to a less preferable man (divorce!)

Average-case

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

everyone has an ordered list.
proposing, rejected, accepted,
running off with another man ...

Looks very complicated!



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

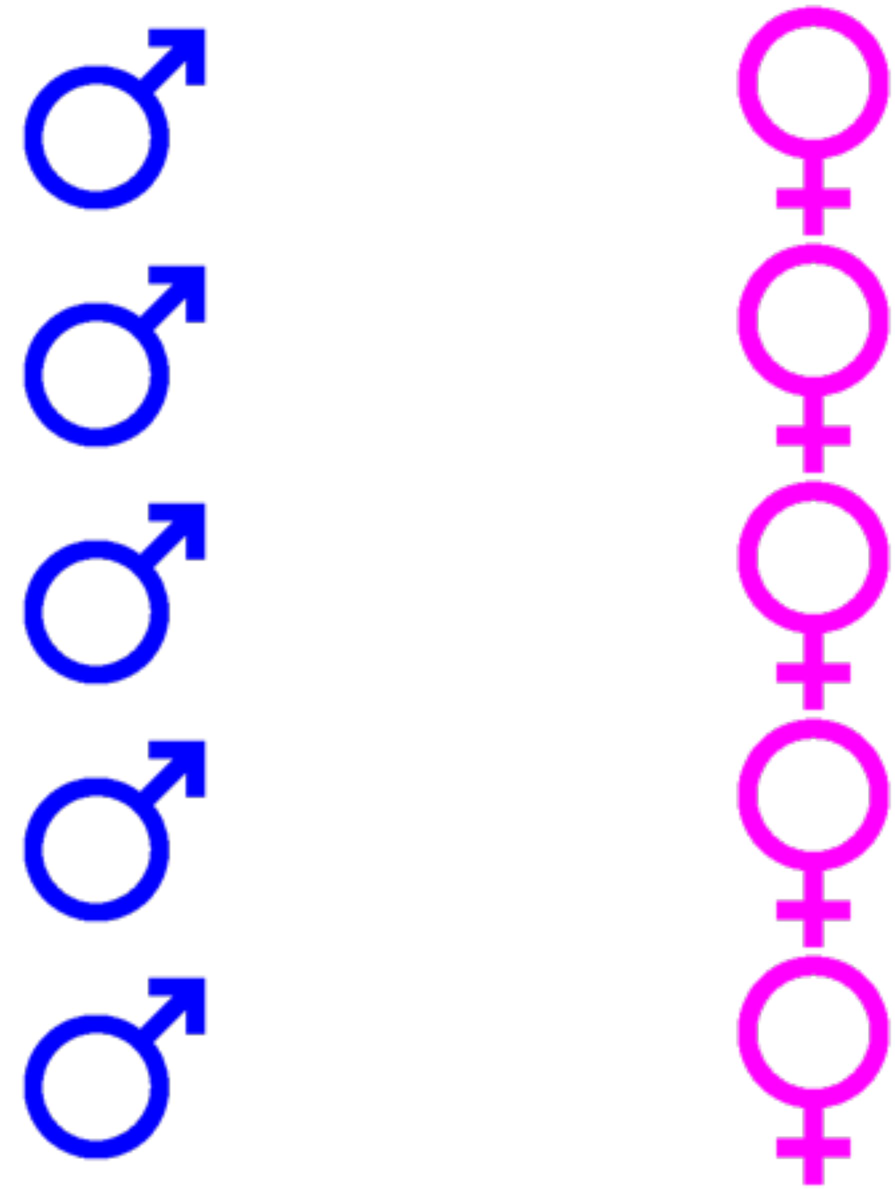
⇕

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

men propose

women change minds



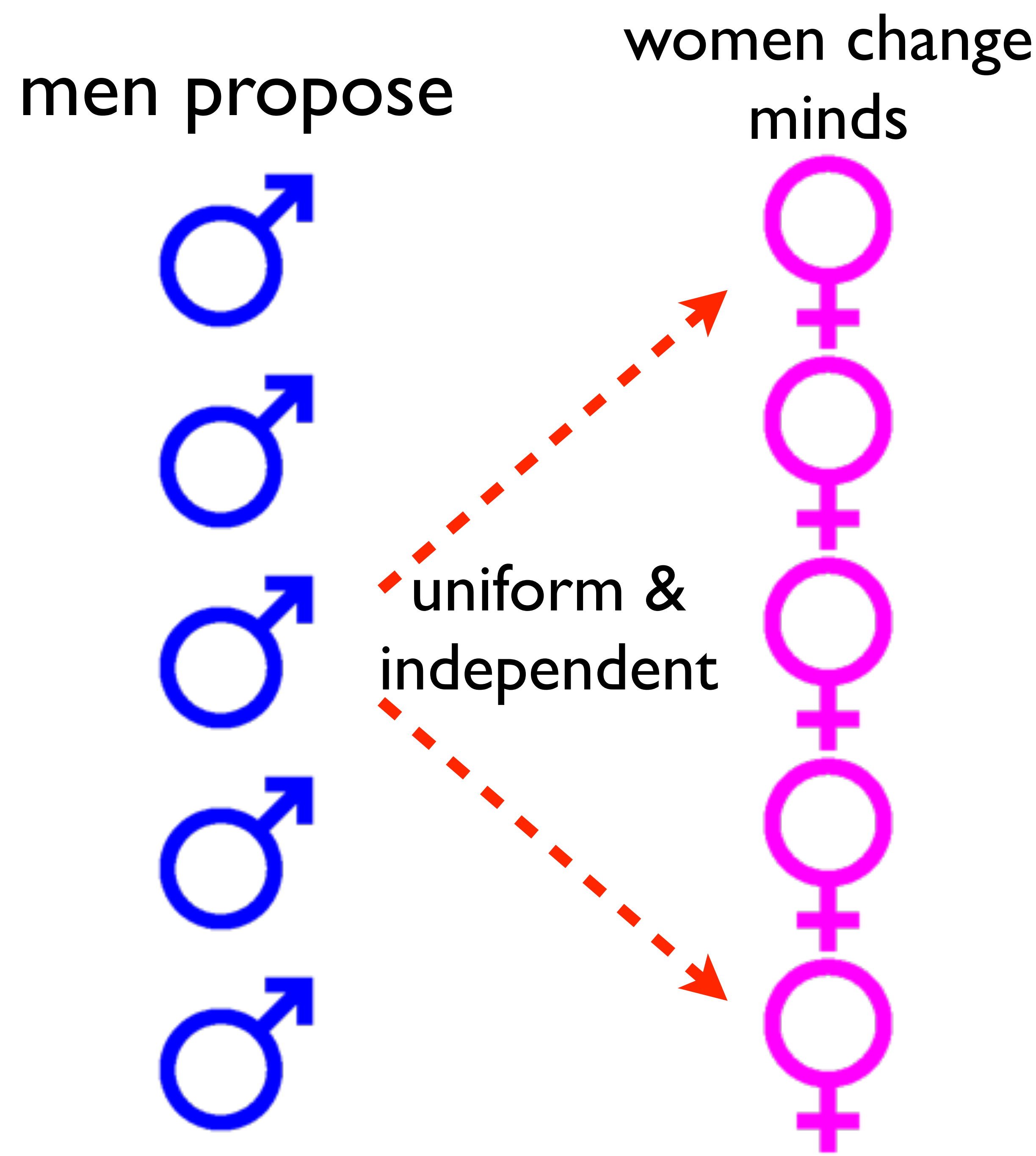
Principle of Deferred Decisions

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

$\leq$

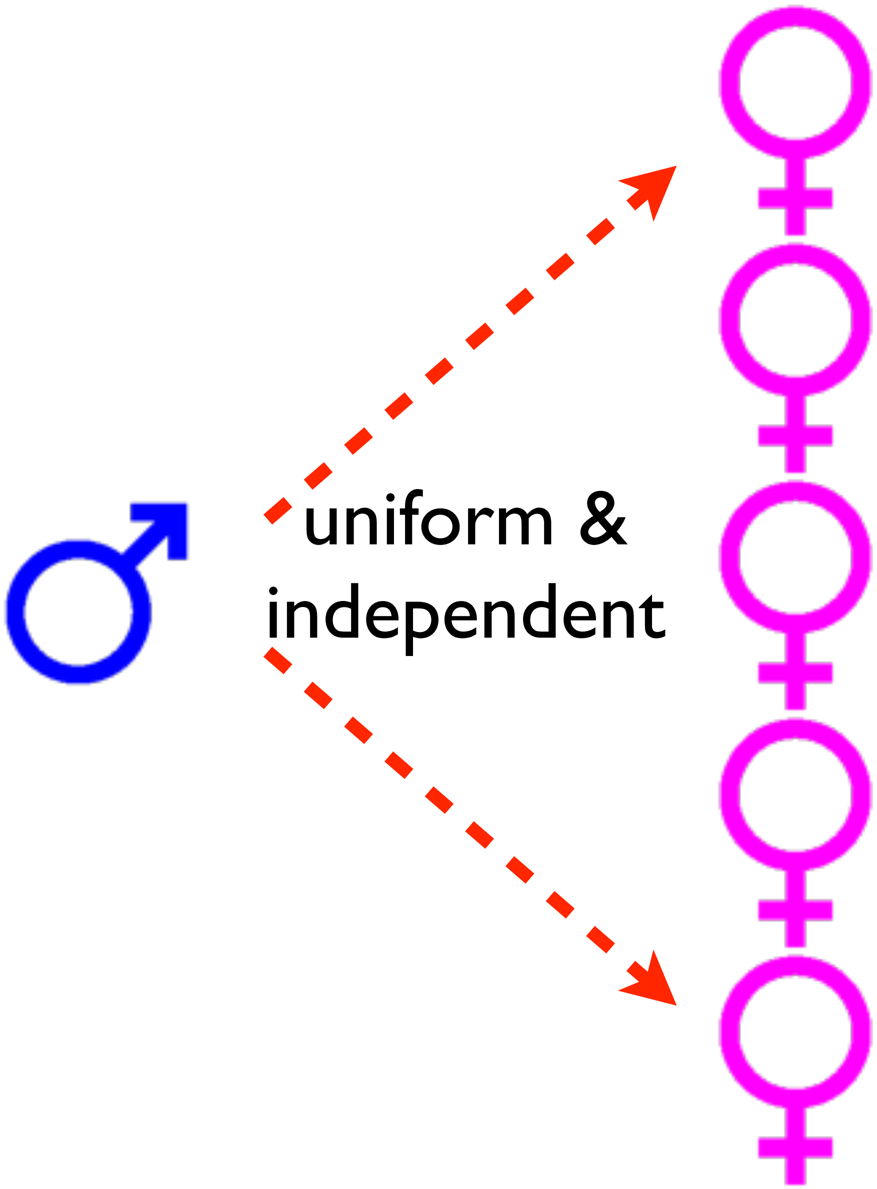
at each time, proposing to
a uniformly & independently
random woman

the man forgot who had
rejected him (!)



Principle of Deferred Decisions

- uniformly and independently proposing to n women
- Alg stops once all women got proposed.
- Coupon collector!
- Expected $O(n \ln n)$ proposals.

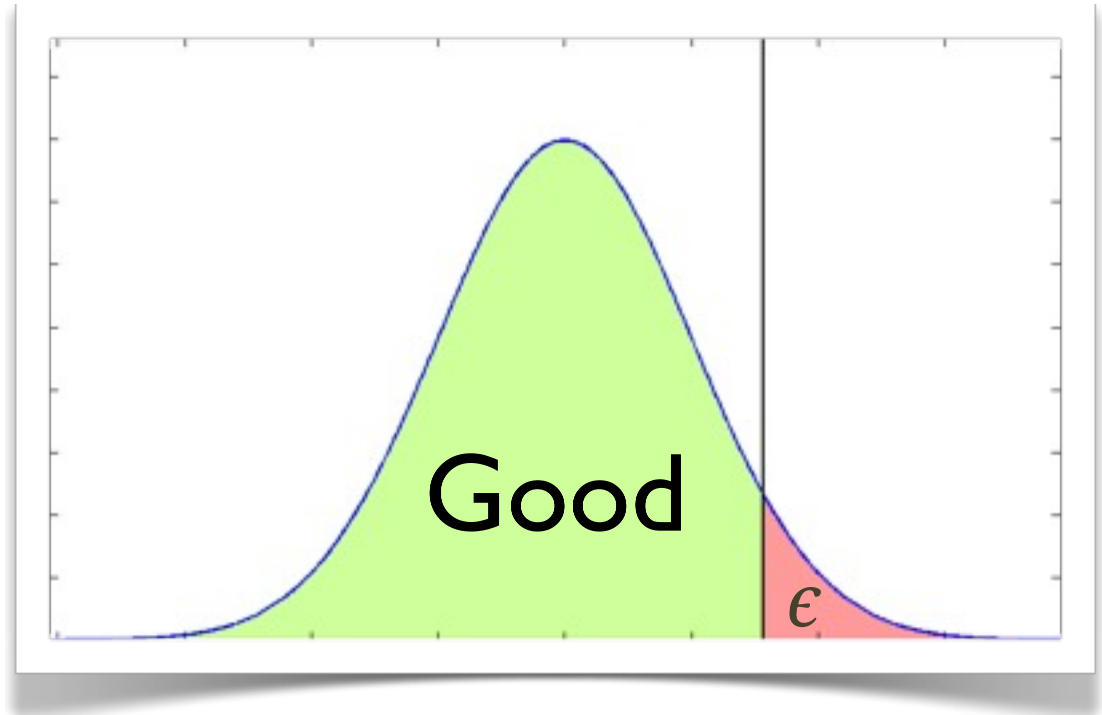


Tail Inequalities

Tail bound:

$$\Pr[X > t] < \epsilon.$$

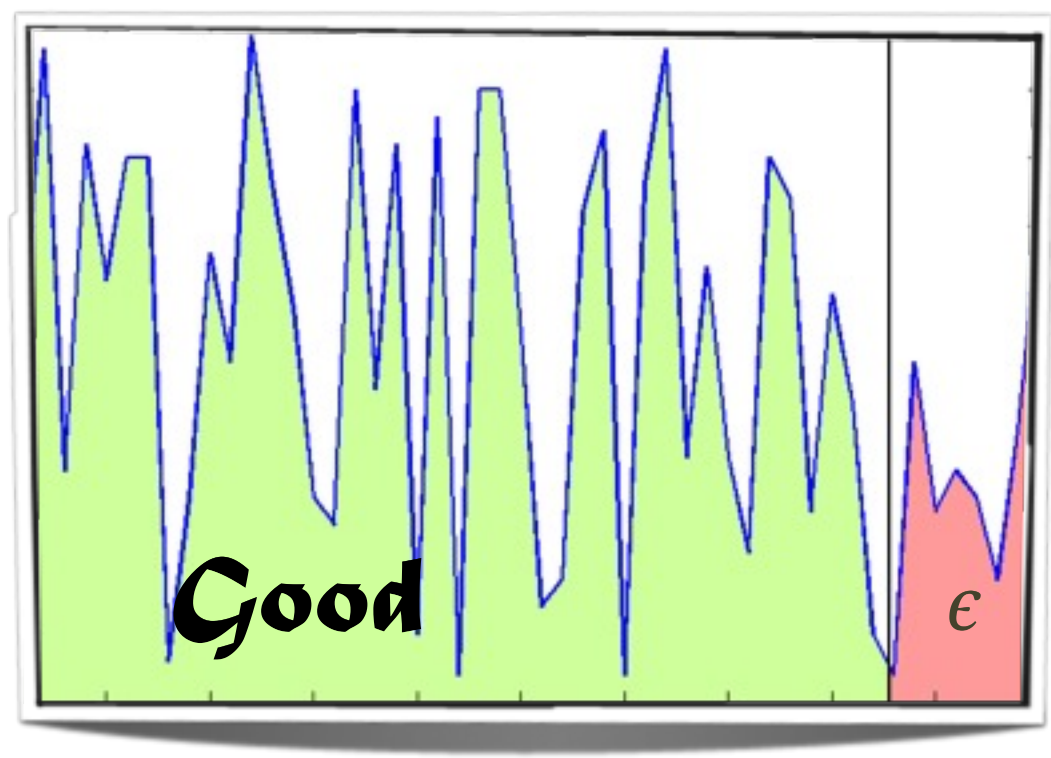
Pretty:



Thresholding:

- The running time of a Las Vegas Alg.
- Some cost (e.g. max load).
- The probability of extreme case.

Ugly:



Tail bound:

$$\Pr[X > t] < \epsilon.$$

Take 1: Counting

- calculation
- smartness

tail bounds for dummies?

n-ball-to-n-bin:

$$\begin{aligned}
&\Pr[\text{load of the first bin} \geq t] \\
&\leq \binom{n}{t}\left(\frac{1}{n}\right)^t \\
&= \frac{n!}{t!(n-t)!n^t} \\
&= \frac{1}{t!} \cdot \frac{n(n-1)(n-2)\cdots(n-t+1)}{n^t} \\
&= \frac{1}{t!} \cdot \prod_{i=0}^{t-1}\left(1-\frac{i}{n}\right) \\
&\leq \frac{1}{t!} \\
&\leq \left(\frac{e}{t}\right)^t
\end{aligned}$$

Tail bound:

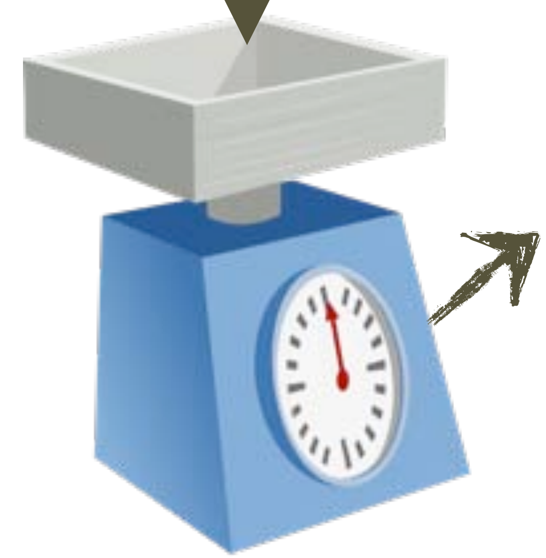
$$\Pr[X > t] < \epsilon.$$

Take II: Characterizing

Relate tail to some measurable characters of X

Reduce the tail bound to the analysis of the characters.

X follows distribution $\mathscr{D}$



character I

$$\Pr[\, X > t \,] < f\,(t,\, I\,)$$

Markov's Inequality

Markov's Inequality:

For *nonnegative* X, for any $t > 0$,

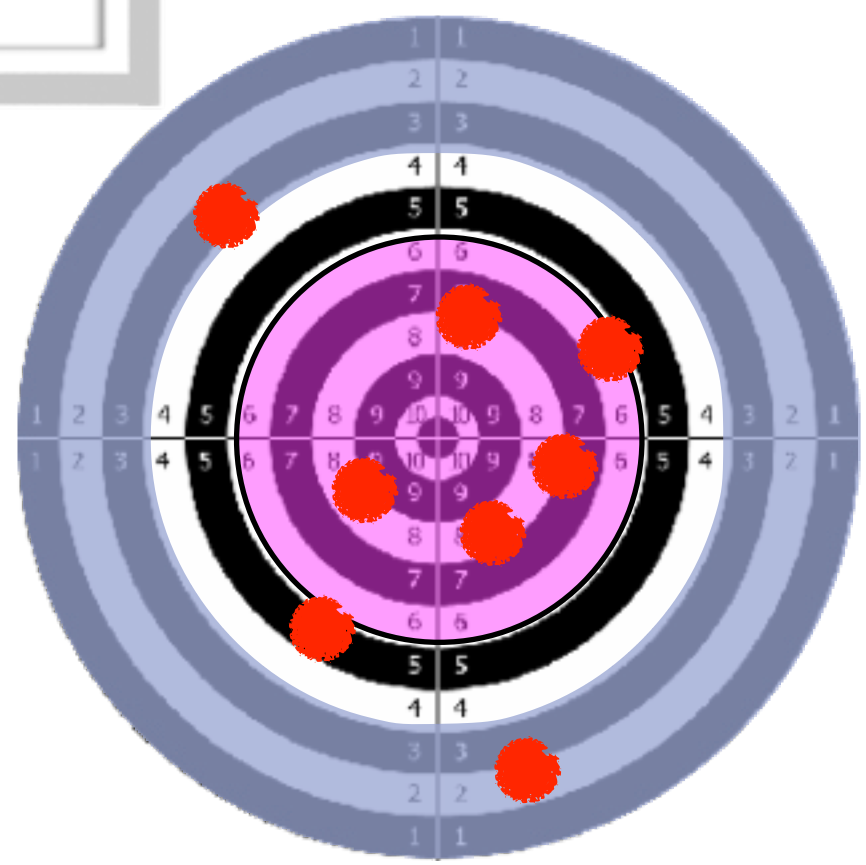
$$\Pr[X \geq t] \leq \frac{\mathbf{E}[X]}{t}.$$

Proof:

Let $Y = \begin{cases} 1 & \text{if } X \geq t, \\ 0 & \text{otherwise.} \end{cases} \Rightarrow Y \leq \left\lfloor \frac{X}{t} \right\rfloor \leq \frac{X}{t},$

$$\Pr[X \geq t] = \mathbf{E}[Y] \leq \mathbf{E}\left[\frac{X}{t}\right] = \frac{\mathbf{E}[X]}{t}.$$

QED



tight if we only know the expectation of X

Las Vegas to Monte Carlo

- Las Vegas: running time is random, always correct.
- A: Las Vegas Alg with worst-case expected running time $T(n)$.
- Monte Carlo: running time is fixed, correct with chance.
- B: Monte Carlo Alg ...

```
B(x):
    run A(x) for 2T(n) steps;
    if A(x) returned
        return A(x);
    else return 1;
```

one-sided error!

$$\Pr[\text{error}]$$

$$\leq \Pr[T(A(x)) > 2T(n)]$$

$$\leq \frac{\mathbf{E}[T(A(x))]}{2T(n)} \leq \frac{1}{2}$$

ZPP ⊆ **RP**

A Generalization of Markov's Inequality

Theorem:

For any X, for $h: X \mapsto \mathbb{R}^+$, for any $t > 0$,

$$\Pr[h(X) \geq t] \leq \frac{\mathbf{E}[h(X)]}{t}.$$

Chebyshev, Chernoff, ...

Chebyshev's Inequality

Chebyshev's Inequality:

For any $t > 0$,

$$\Pr\left[|X - \mathbf{E}[X]| \geq t\right] \leq \frac{\mathbf{Var}[X]}{t^2}.$$

Variance

Definition (variance):

The variance of a random variable X is

$$\mathbf{Var}[X] = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] = \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2.$$

The standard deviation of random variable X is

$$\delta[X] = \sqrt{\mathbf{Var}[X]}$$

Covariance

Definition (covariance):

The covariance of X and Y is

$$\mathbf{Cov}(X,Y) = \mathbf{E}\left[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\right].$$

Theorem:

$$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}(X,Y);$$

$$\mathbf{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{Var}[X_i] + \sum_{i \neq j} \mathbf{Cov}(X_i, X_j).$$

Covariance

Theorem:

For independent X and Y, $\mathbf{E}[X \cdot Y] = \mathbf{E}[X] \cdot \mathbf{E}[Y]$.

Theorem:

For independent X and Y, $\mathbf{Cov}(X, Y) = 0$.

Proof:

$$\begin{aligned} \mathbf{Cov}(X, Y) &= \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] \\ &= \mathbf{E}[X - \mathbf{E}[X]]\,\mathbf{E}[Y - \mathbf{E}[Y]] \\ &= 0. \end{aligned}$$

Variance of sum

Theorem:

For independent X and Y, $\mathbf{Cov}(X,Y) = 0$.

Theorem:

For **pairwise** independent $X_1, X_2, \ldots, X_n$,

$$\mathbf{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{Var}[X_i].$$

Variance of Binomial Distribution

- Binomial distribution: number of successes in n i.i.d. Bernoulli trials.
- X follows binomial distribution with parameter n and p

$$X = \sum_{i=1}^{n} X_i \qquad X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

$$\mathbf{Var}[X_i] = \mathbf{E}[X_i^2] - \mathbf{E}[X_i]^2 = p - p^2 = p(1-p)$$

$$\mathbf{Var}[X] = \sum_{i=1}^{n} \mathbf{Var}[X_i] = p(1-p)n$$

(independence)

Chebyshev's Inequality

Chebyshev's Inequality:

For any $t > 0$,

$$\Pr\left[|X - \mathbf{E}[X]| \geq t\right] \leq \frac{\mathbf{Var}[X]}{t^2}.$$

Proof:

Apply Markov's inequality to $(X - \mathbf{E}[X])^2$

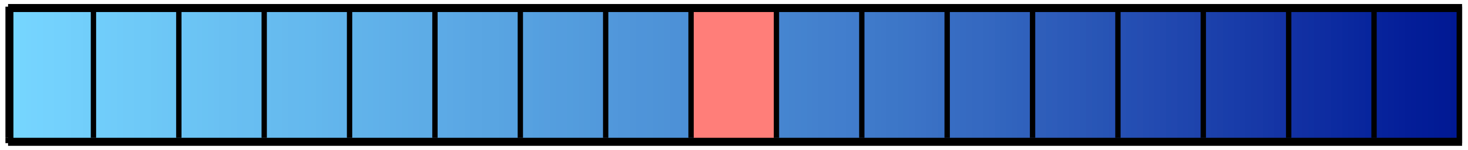
$$\Pr\left[(X - \mathbf{E}[X])^2 \geq t^2\right] \leq \frac{\mathbf{E}\left[(X - \mathbf{E}[X])^2\right]}{t^2}$$

QED

Selection Problem

Input: a set of n elements

Output: median



straightforward alg:

sorting, $\Omega(n \log n)$ time

sophisticated deterministic alg:

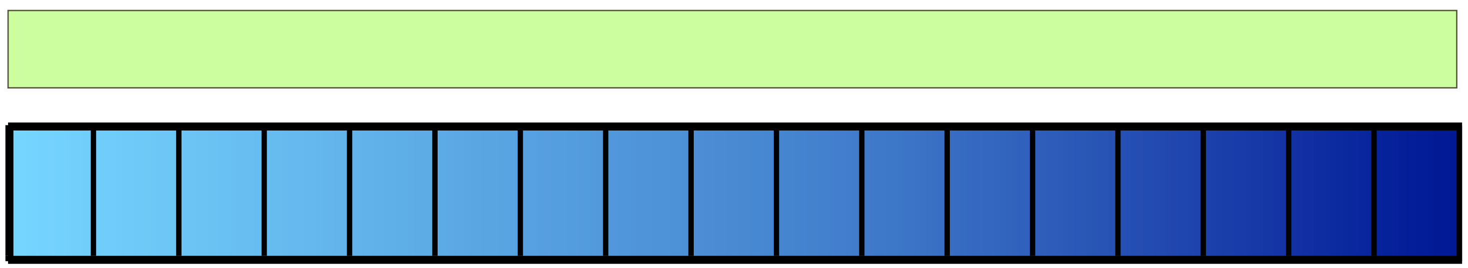
median of medians, $\Theta(n)$ time

simple randomized alg:

LazySelect, $\Theta(n)$ time, find the median whp

Selection by Sampling

distribution:

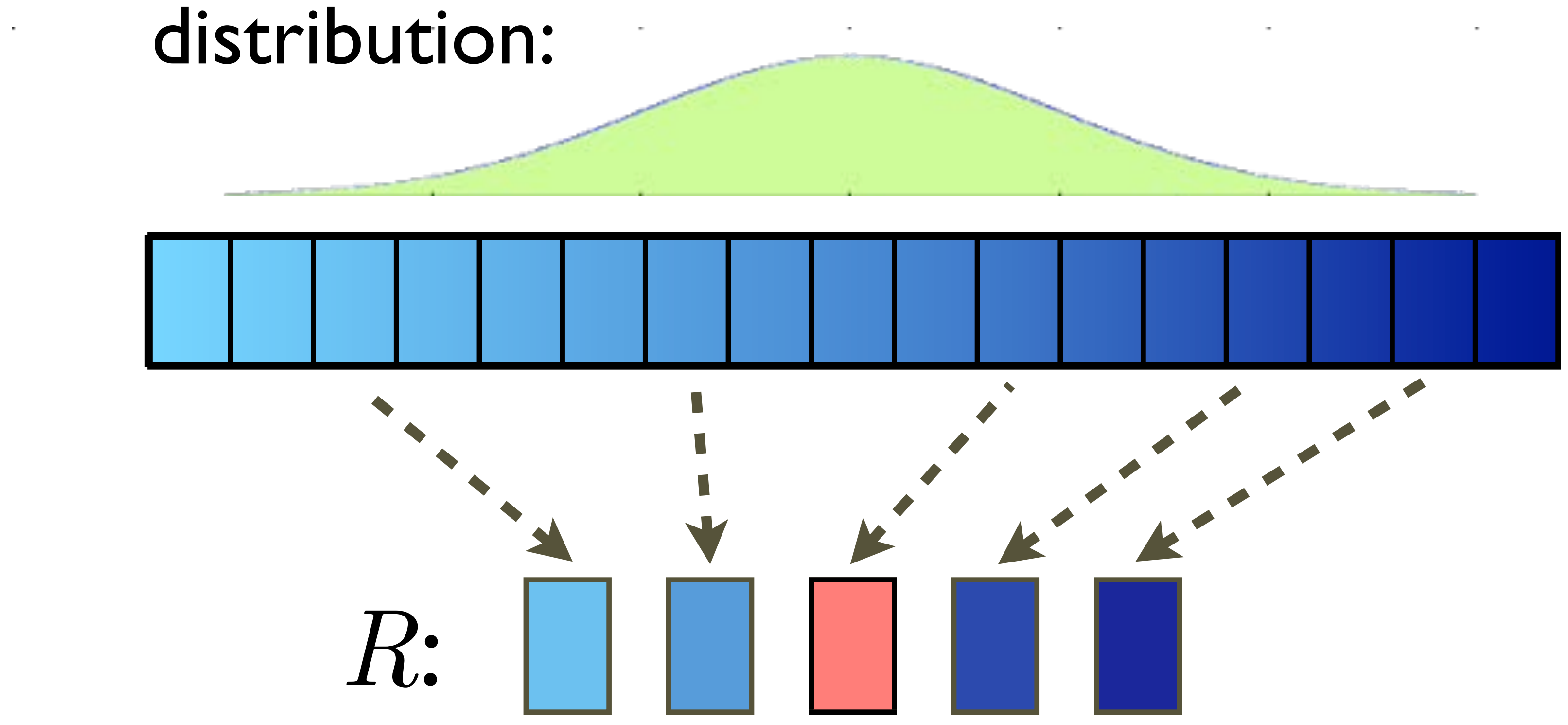


Naive sampling:

uniformly choose an random element

make a wish it is the median

Selection by Sampling



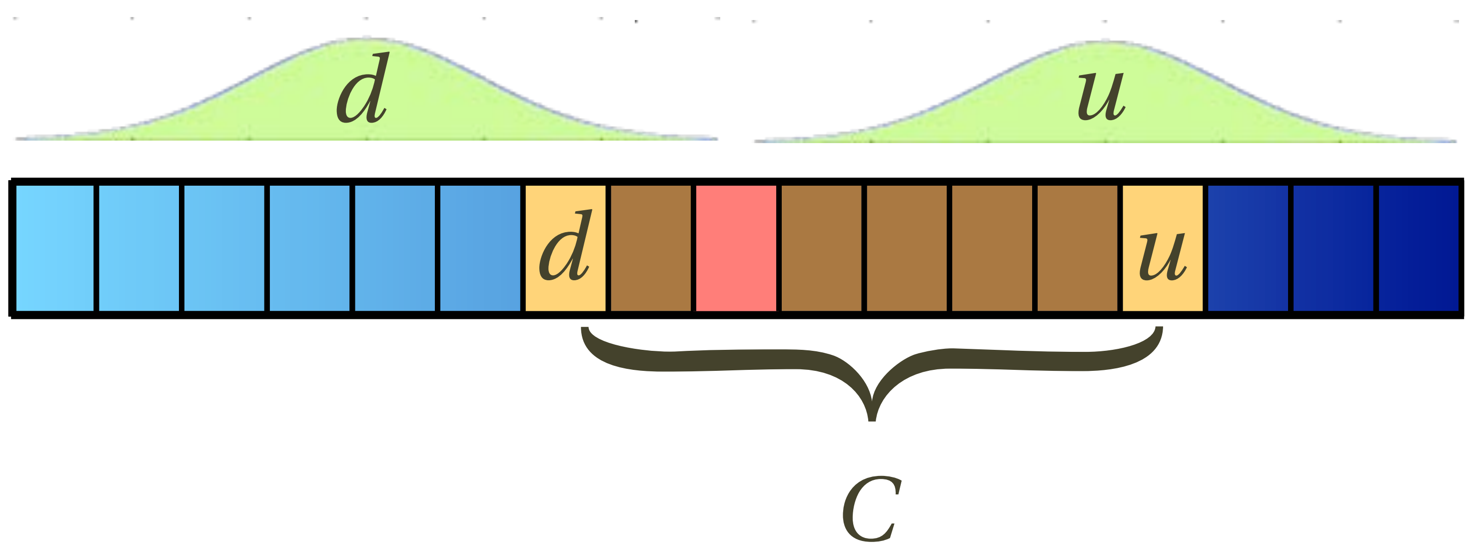
sample a small set R,

selection in R by sorting

roughly concentrated, but not good enough

Selection by Sampling

distributions:

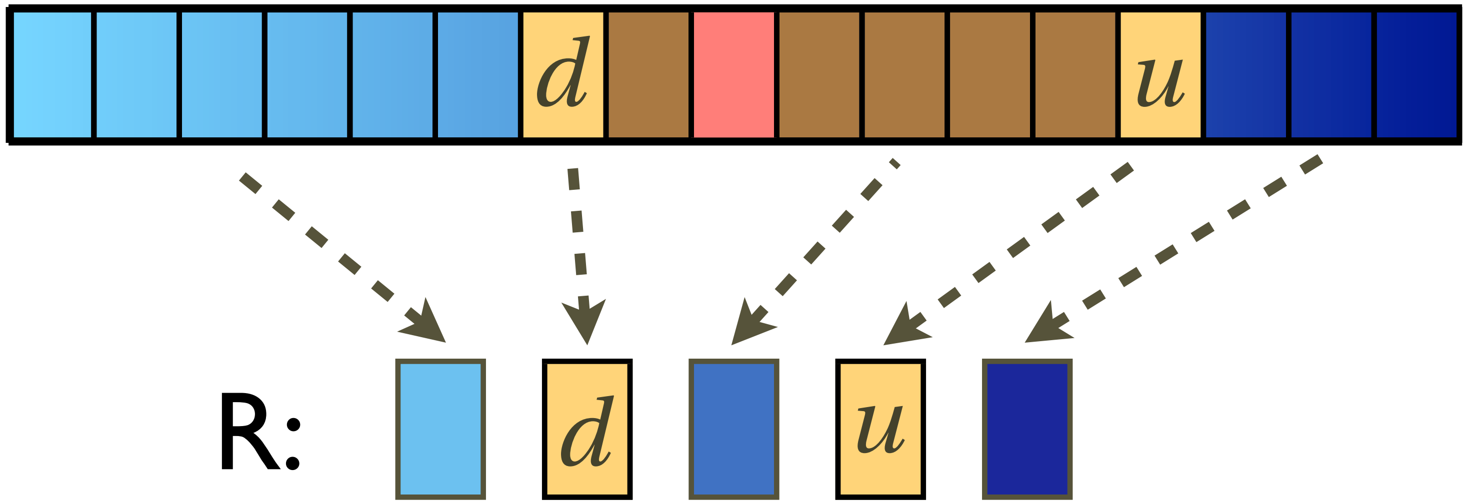


Find such d and u that:

- Let $C = \{x \in S \mid d \leq x \leq u\}$.
- The median is in C.
- C is not too large (sort C is linear time).

LazySelect

(Floyd & Rivest)



Size of R: r

Offset for d and u from the median of R: k

Bad events: median is not between d and u;

too many elements between d and u.

(inefficient to sort)

$O(r \log r)$ 1. Uniformly and independently sample r elements from S to form R; and sort R.

$O(1)$ 2. Let d be the $(\frac{r}{2} - k)$th element in R.

$O(1)$ 3. Let u be the $(\frac{r}{2} + k)$th element in R.

$O(n)$ 4. If any of the following occurs

$$|\{x \in S \mid x < d\}| > \frac{n}{2};$$

$$|\{x \in S \mid x > u\}| > \frac{n}{2};$$

$$|\{x \in S \mid d \le x \le u\}| > s;$$

then FAIL. Pr[FAIL] < ?

$O(s \log s)$ 5. Find the median of S by sorting $\{x \in S \mid d \le x \le u\}$.

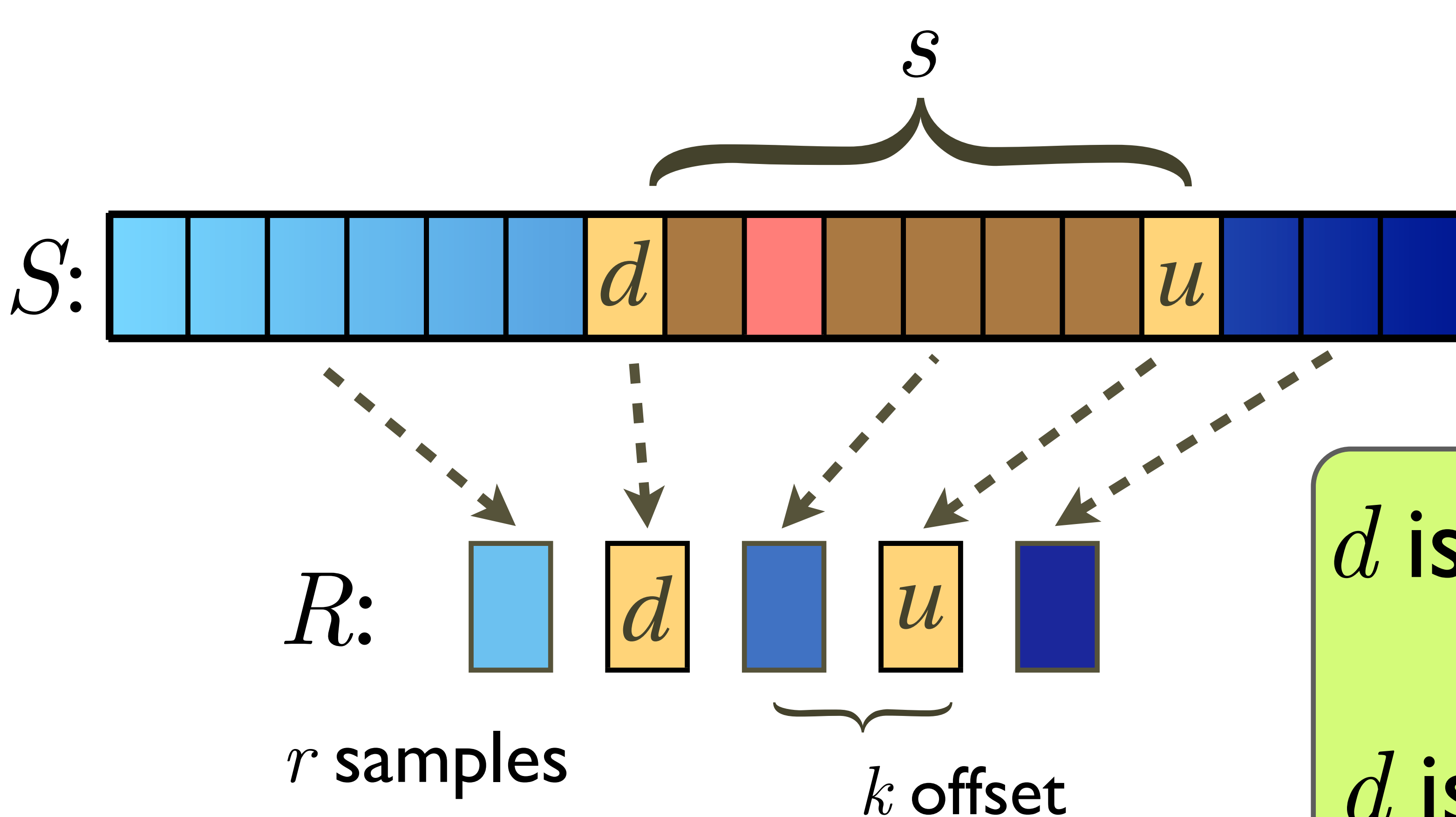
Bad events:

1. $|\{x \in S \mid x < d\}| > \frac{n}{2}$;
2. $|\{x \in S \mid x > u\}| > \frac{n}{2}$;
3. $|\{x \in S \mid d \le x \le u\}| > s$;

or $|\{x \in S \mid x < d\}| < \frac{n}{2} - \frac{s}{2}$;

$|\{x \in S \mid x > u\}| < \frac{n}{2} - \frac{s}{2}$;

Symmetry!



Bad events for d:

d is too large:

$|\{x \in S \mid x < d\}| > \frac{n}{2}$

d is too small:

$|\{x \in S \mid x < d\}| < \frac{n}{2} - \frac{s}{2}$

Bad events for d:

d is too large:

$$|\{x \in S \mid x < d\}| > \frac{n}{2}$$

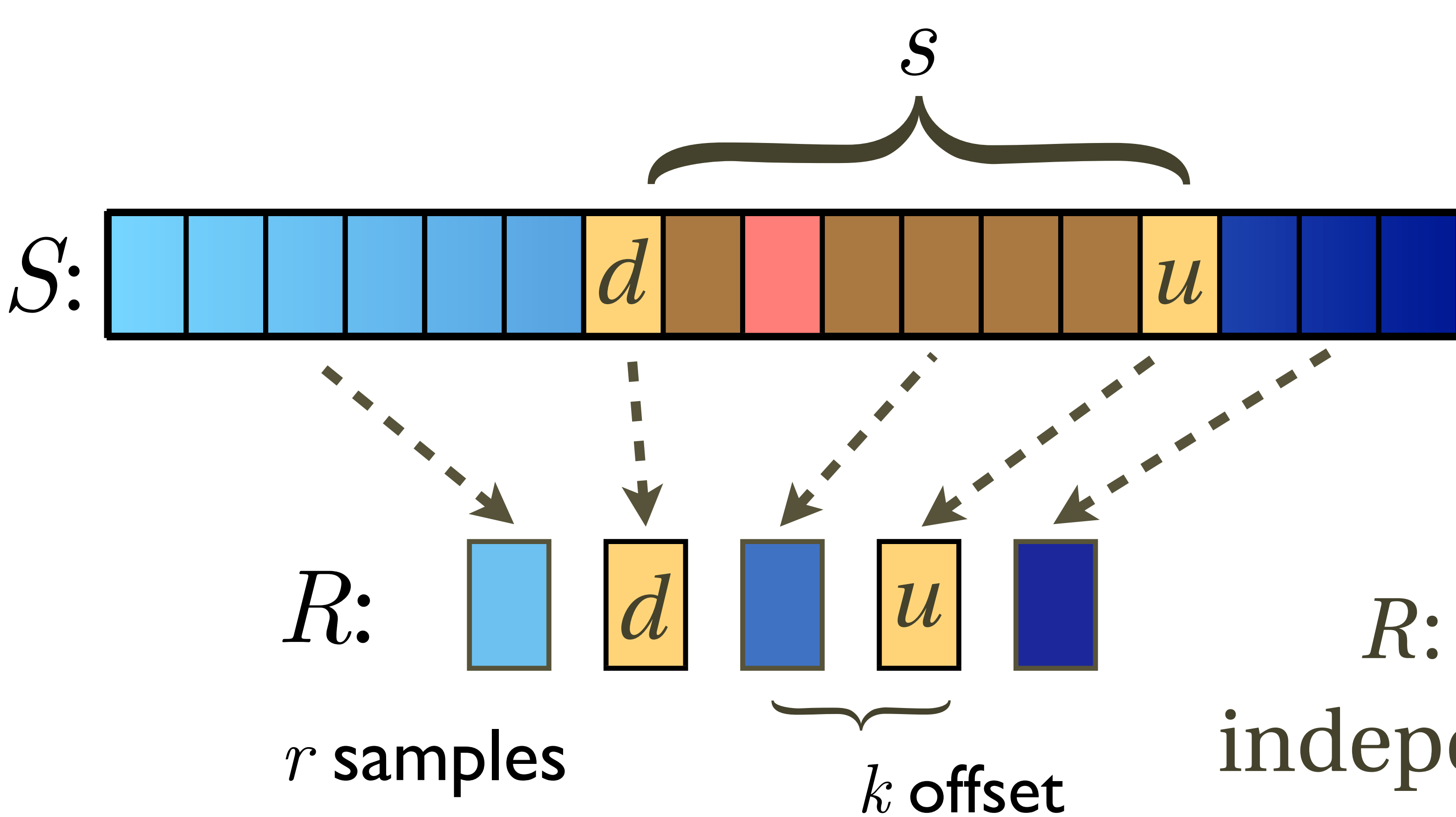
d is too small:

$$|\{x \in S \mid x < d\}| < \frac{n}{2} - \frac{s}{2}$$

Bad events for R:

the sample of rank $\frac{r}{2} - k$ is ranked $> \frac{n}{2}$ in S.

the sample of rank $\frac{r}{2} - k$ is ranked $\leq \frac{n}{2} - \frac{s}{2}$ in S.



R: r uniform and independent samples from S

Bad events for d:

d is too large:

$$|\{x \in S \mid x < d\}| > \frac{n}{2}$$

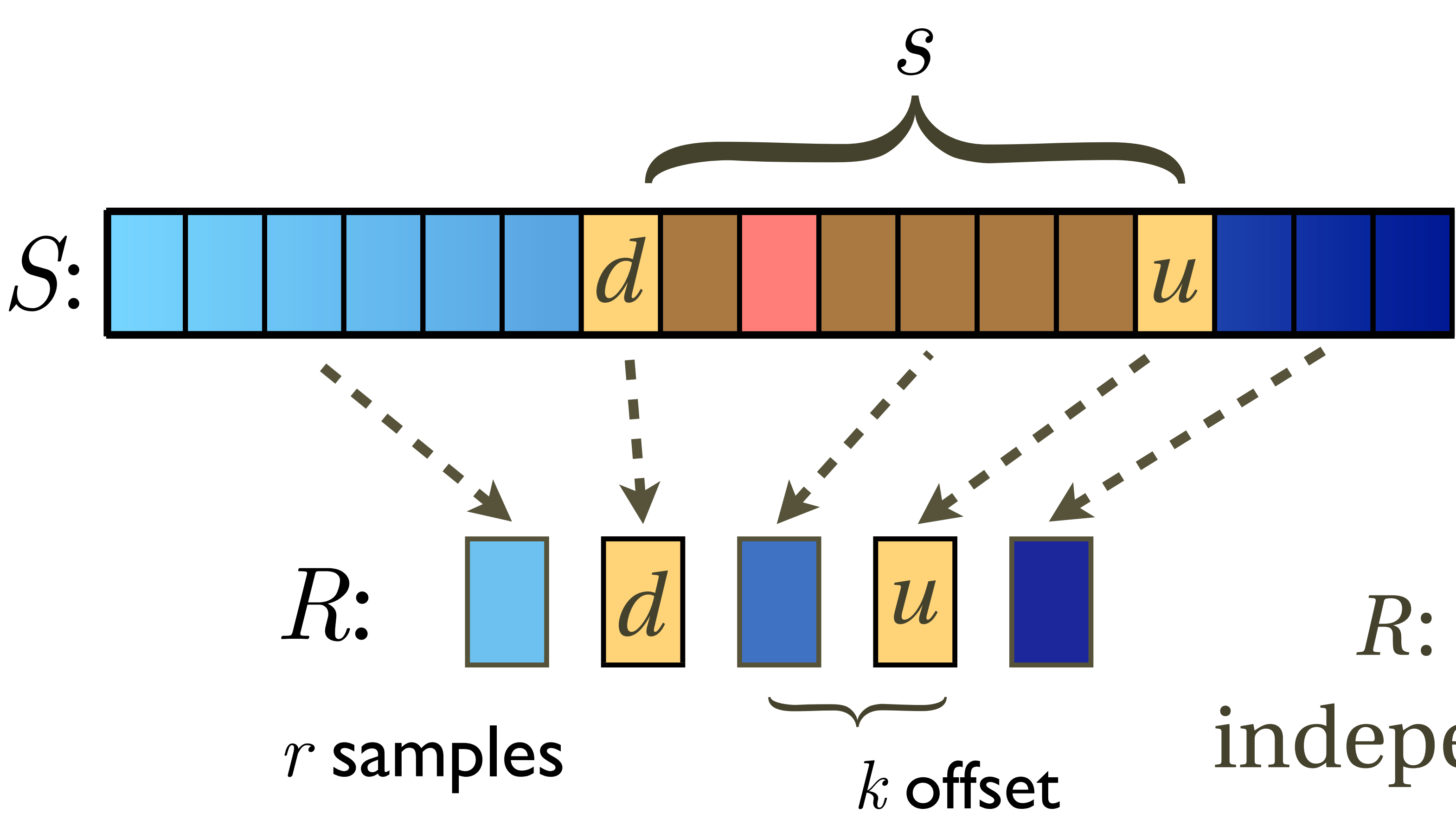
d is too small:

$$|\{x \in S \mid x < d\}| < \frac{n}{2} - \frac{s}{2}$$

Bad events for R:

$< \frac{r}{2} - k$ samples are among the smallest half in S.

$\geq \frac{r}{2} - k$ samples are among the $\frac{n}{2} - \frac{s}{2}$ smallest in S.



R: r uniform and independent samples from S

R: r uniform and independent samples from S

$$X_i = \begin{cases} 1 & i\text{th sample ranks} \le n/2, \\ 0 & \text{otherwise.} \end{cases}$$

Bad events for R:

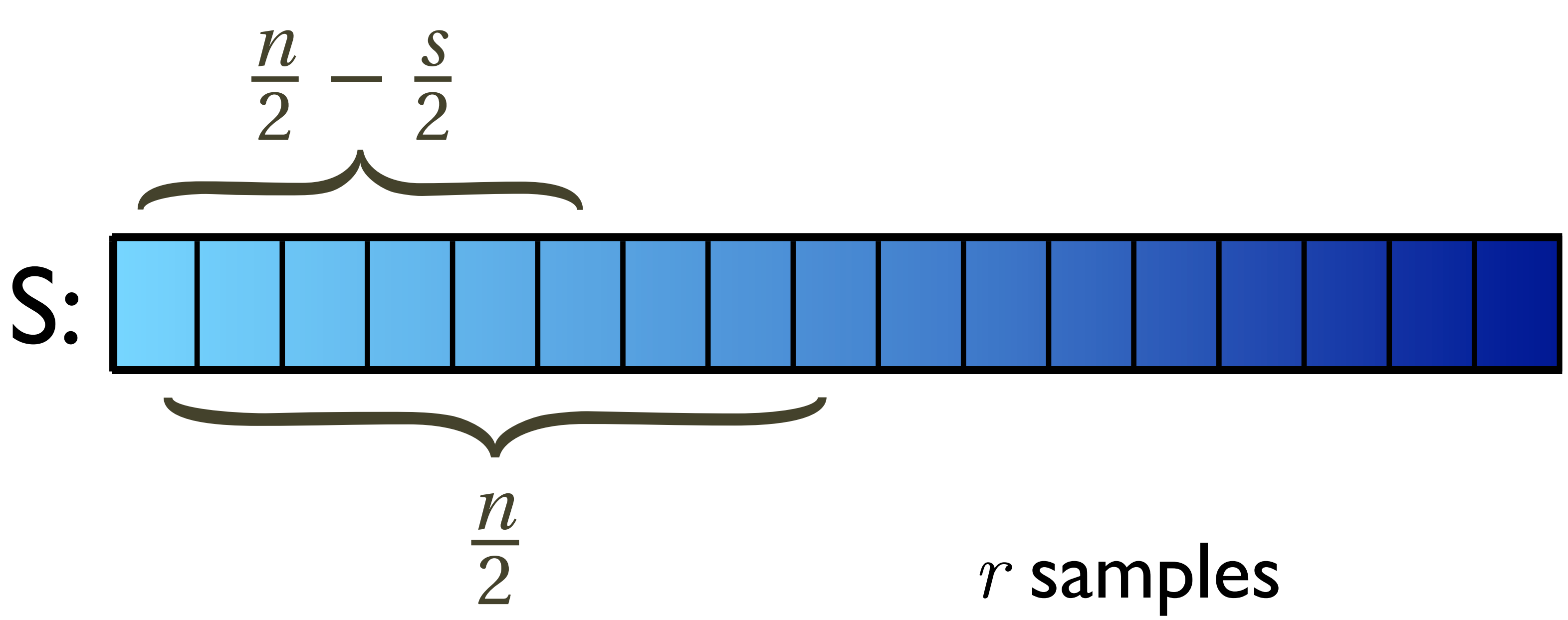
$\mathscr{E}_1$: $< \frac{r}{2} - k$ samples are among the smallest half in S.

$\mathscr{E}_2$: $\ge \frac{r}{2} - k$ samples are among the $\frac{n}{2} - \frac{s}{2}$ smallest in S.

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

$$Y_i = \begin{cases} 1 & i\text{th sample ranks} \le \frac{n}{2} - \frac{s}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

$$Y = \sum_{i=1}^{r} Y_i$$



r samples

R: r uniform and independent samples from S

$$X_i = \begin{cases} 1 & \text{with prob } \frac{1}{2} \\ 0 & \text{with prob } \frac{1}{2} \end{cases}$$

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

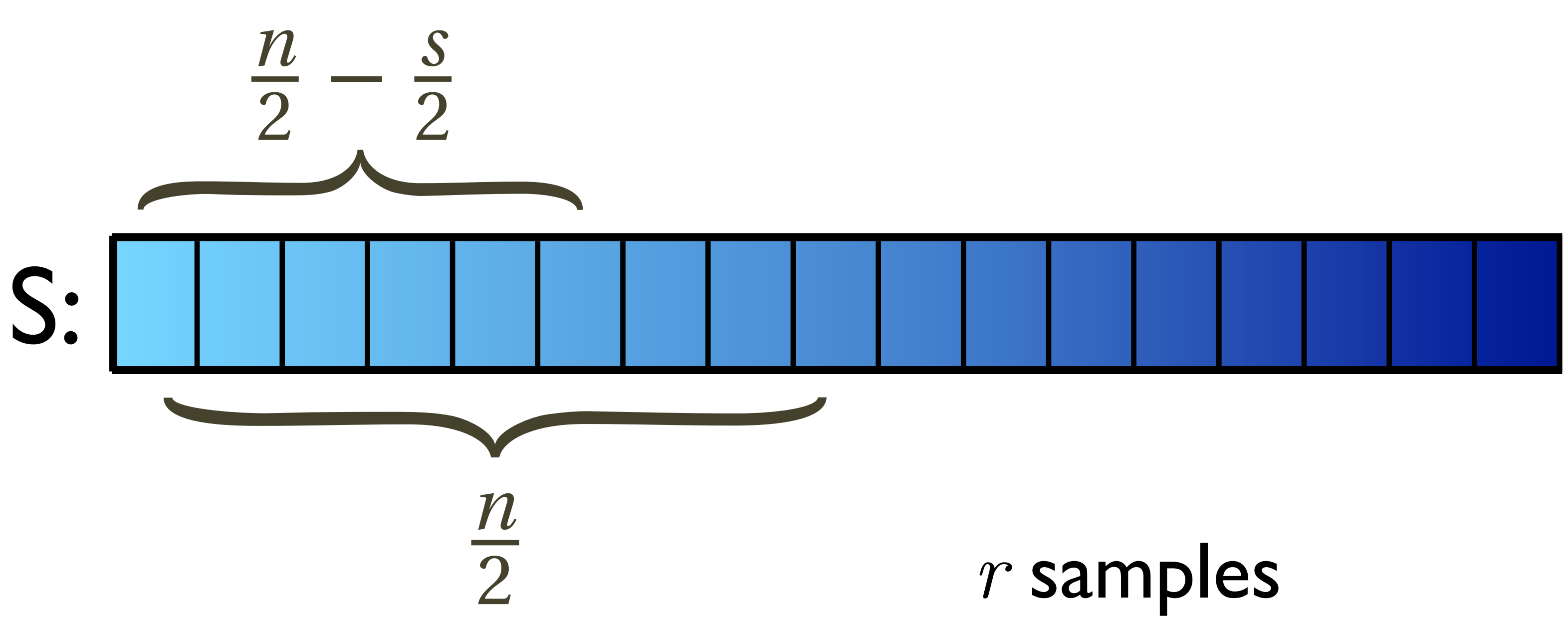
Bad events for R:

$\mathcal{E}_1$: $< \frac{r}{2} - k$ samples are among the smallest half in S.

$\mathcal{E}_2$: $\geq \frac{r}{2} - k$ samples are among the $\frac{n}{2} - \frac{s}{2}$ smallest in S.

$$Y_i = \begin{cases} 1 & \text{with prob } \frac{1}{2} - \frac{s}{2n} \\ 0 & \text{with prob } \frac{1}{2} + \frac{s}{2n} \end{cases}$$

$$Y = \sum_{i=1}^{r} Y_i$$



r samples

Bad events:

$$\mathcal{E}_1 : \quad X < \frac{r}{2} - k$$

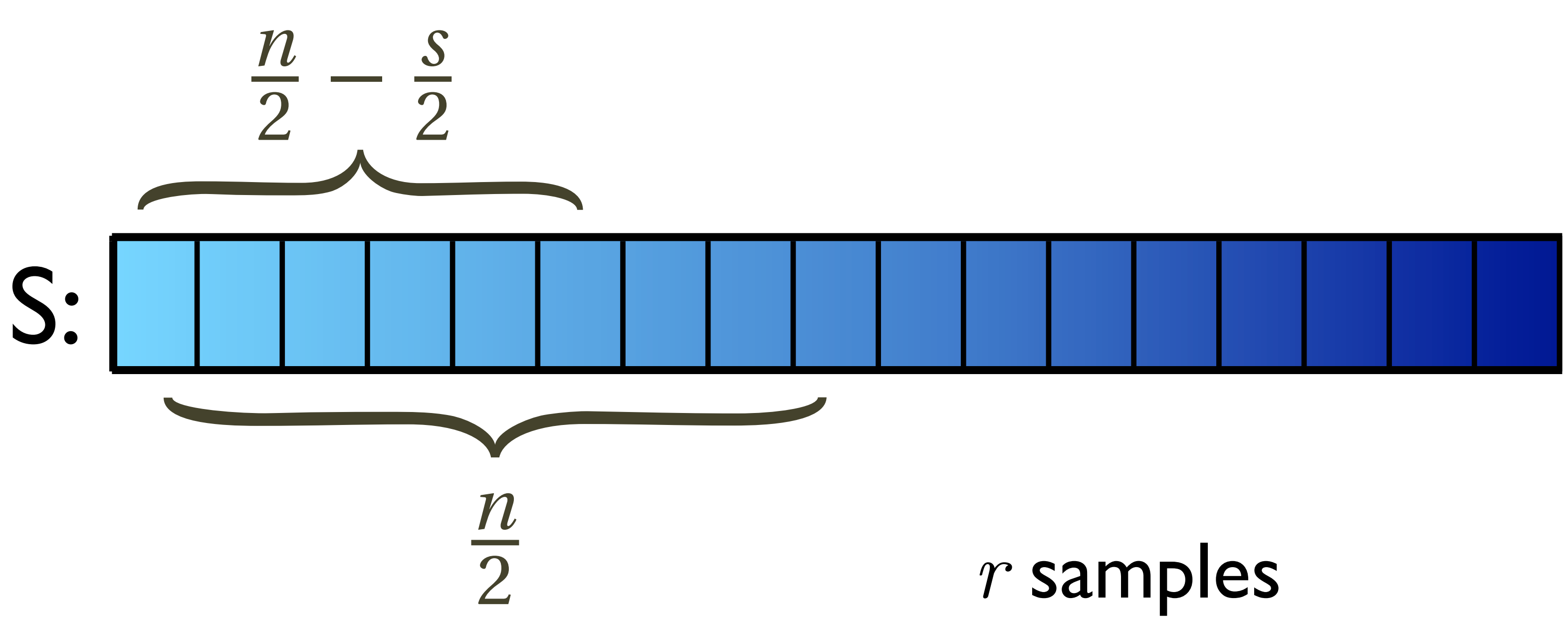
$$\mathcal{E}_2 : \quad Y \geq \frac{r}{2} - k$$

$$X_i = \begin{cases} 1 & \text{with prob } \frac{1}{2} \\ 0 & \text{with prob } \frac{1}{2} \end{cases}$$

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

$$Y_i = \begin{cases} 1 & \text{with prob } \frac{1}{2} - \frac{s}{2n} \\ 0 & \text{with prob } \frac{1}{2} + \frac{s}{2n} \end{cases}$$

$$Y = \sum_{i=1}^{r} Y_i$$



r samples

Bad events:

$$\mathcal{E}_1 : \quad X < \frac{r}{2} - k$$

$$\mathcal{E}_2 : \quad Y \geq \frac{r}{2} - k$$

$$X_i = \begin{cases} 1 & \text{with prob } \frac{1}{2} \\ 0 & \text{with prob } \frac{1}{2} \end{cases}$$

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

$$Y_i = \begin{cases} 1 & \text{with prob } \frac{1}{2} - \frac{s}{2n} \\ 0 & \text{with prob } \frac{1}{2} + \frac{s}{2n} \end{cases}$$

$$Y = \sum_{i=1}^{r} Y_i$$

X and Y are binomial!

$$\mathbf{E}[X] = \frac{r}{2} \qquad \mathbf{Var}[X] = \frac{r}{4}$$

$$\mathbf{E}[Y] = \frac{r}{2} - \frac{sr}{2n} \qquad \mathbf{Var}[Y] = \frac{r}{4} - \frac{s^2 r}{4n^2}$$

Bad events:

$\mathcal{E}_1$: $$X < \frac{r}{2} - k$$

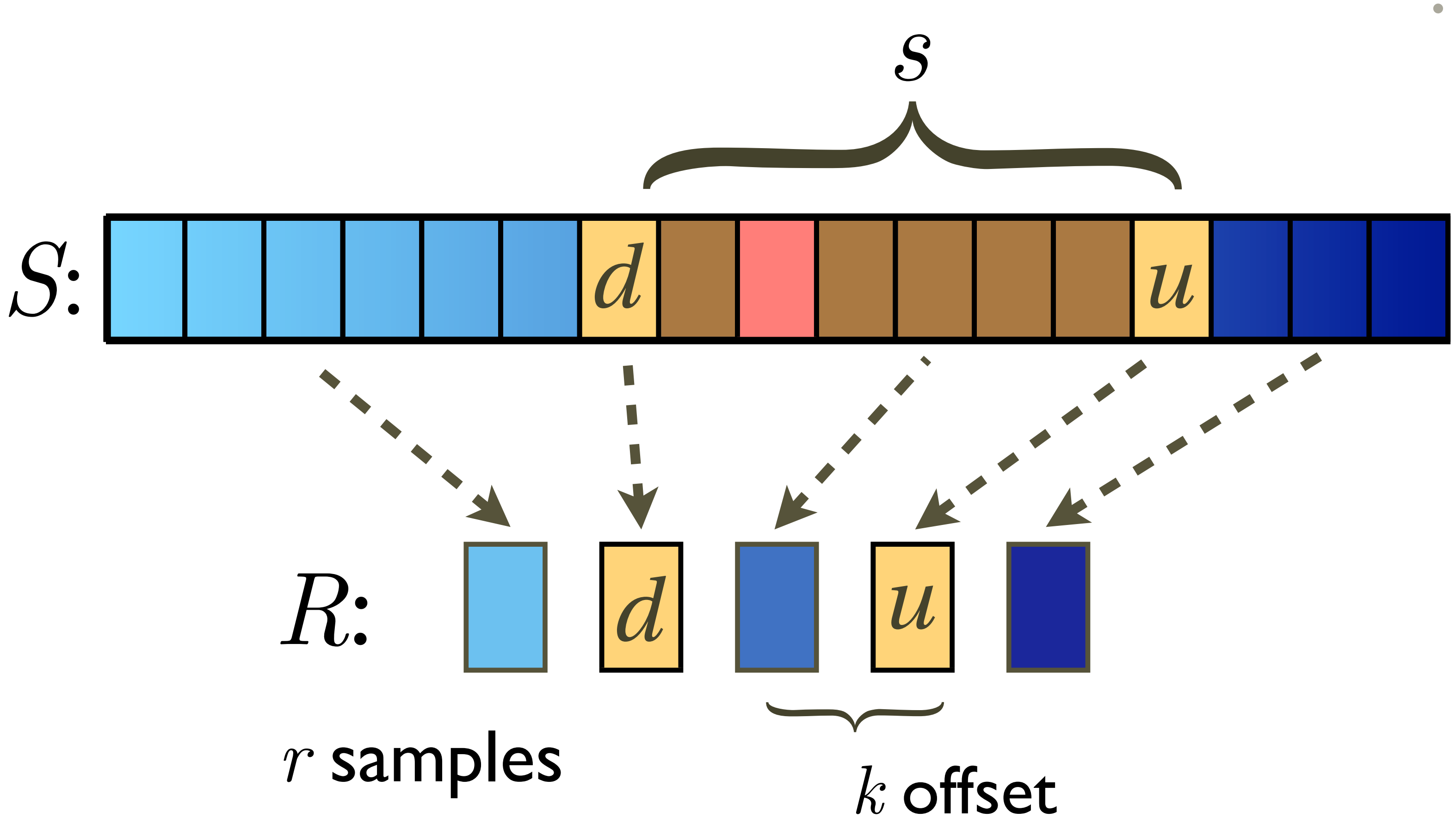
$\mathcal{E}_2$: $$Y \geq \frac{r}{2} - k$$

$$\mathbf{E}[X] = \frac{r}{2}$$

$$\mathbf{Var}[X] = \frac{r}{4}$$

$$\mathbf{E}[Y] = \frac{r}{2} - \frac{sr}{2n}$$

$$\mathbf{Var}[Y] = \frac{r}{4} - \frac{s^2 r}{4n^2}$$



$$r = n^{3/4}$$

$$k = n^{1/2}$$

$$s = 4n^{3/4}$$

Bad events:

$\mathcal{E}_1$: $X < \frac{1}{2}n^{3/4} - \sqrt{n}$

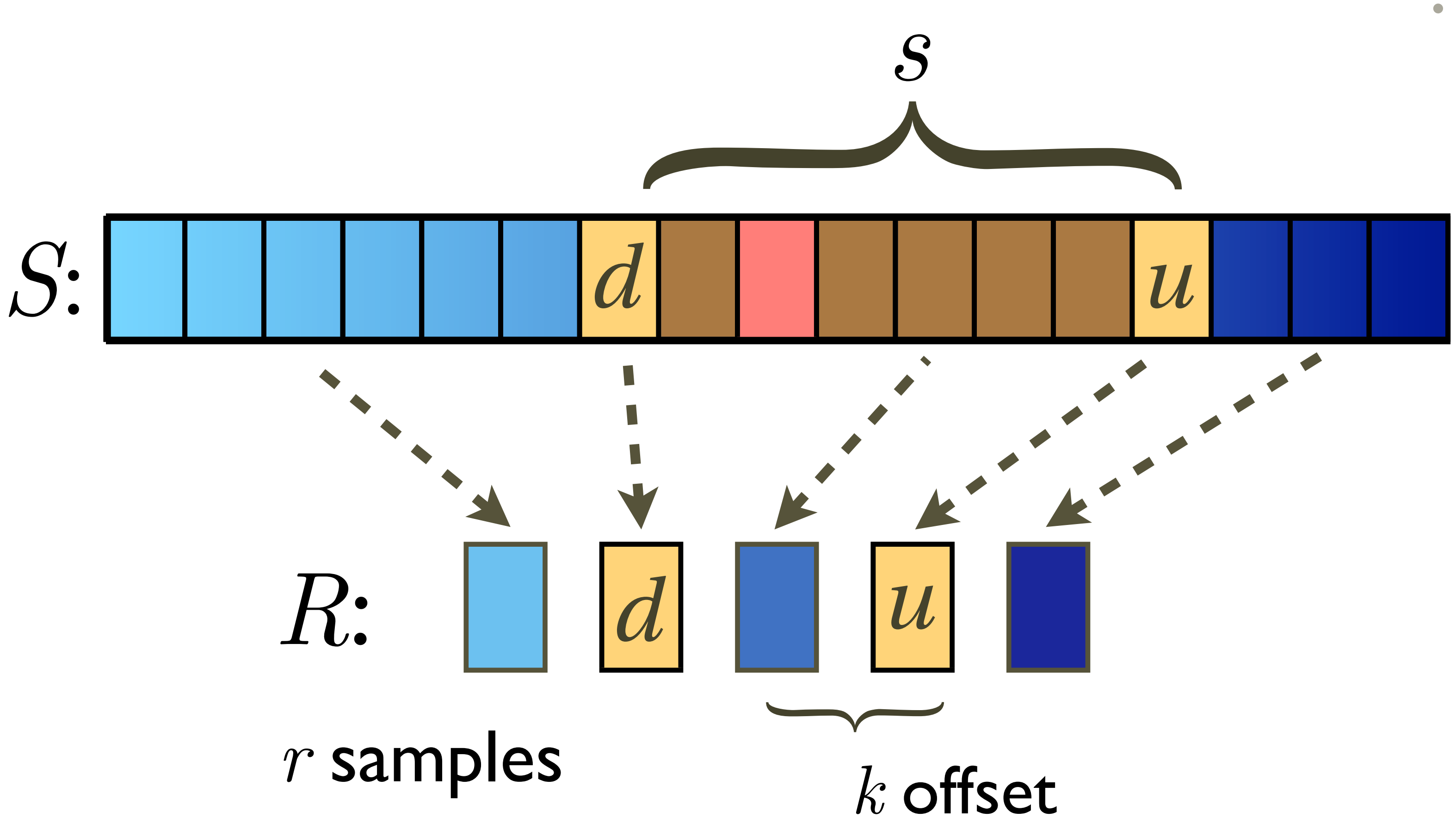
$\mathcal{E}_2$: $Y \geq \frac{1}{2}n^{3/4} - \sqrt{n}$

$$\mathbf{E}[X] = \frac{1}{2}n^{3/4}$$

$$\mathbf{Var}[X] = \frac{1}{4}n^{3/4}$$

$$\mathbf{E}[Y] = \frac{1}{2}n^{3/4} - 2\sqrt{n}$$

$$\mathbf{Var}[Y] < \frac{1}{4}n^{3/4}$$



$r = n^{3/4}$

$k = n^{1/2}$

$s = 4n^{3/4}$

Bad events:

$\mathcal{E}_1 : \quad X < \frac{1}{2}n^{3/4} - \sqrt{n}$

$\mathcal{E}_2 : \quad Y \geq \frac{1}{2}n^{3/4} - \sqrt{n}$

$$\mathbf{E}[X] = \frac{1}{2}n^{3/4}$$

$$\mathbf{Var}[X] = \frac{1}{4}n^{3/4}$$

$$\mathbf{E}[Y] = \frac{1}{2}n^{3/4} - 2\sqrt{n}$$

$$\mathbf{Var}[Y] < \frac{1}{4}n^{3/4}$$

$$\Pr[\mathcal{E}_1] = \Pr\left[X < \frac{1}{2}n^{3/4} - \sqrt{n}\right] \leq \Pr\left[|X - \mathbf{E}[X]| > \sqrt{n}\right] \leq \frac{\mathbf{Var}[X]}{n} \leq \frac{1}{4}n^{-1/4}$$

$$\Pr[\mathcal{E}_2] = \Pr\left[Y \geq \frac{1}{2}n^{3/4} - \sqrt{n}\right] \leq \Pr\left[|Y - \mathbf{E}[Y]| \geq \sqrt{n}\right] \leq \frac{\mathbf{Var}[Y]}{n} \leq \frac{1}{4}n^{-1/4}$$

Bad events:

$$\mathscr{E}_1 : \quad X < \frac{1}{2}n^{3/4} - \sqrt{n}$$

$$\mathscr{E}_2 : \quad Y \geq \frac{1}{2}n^{3/4} - \sqrt{n}$$

$$\Pr[\mathscr{E}_1] \leq \frac{1}{4}n^{-1/4}$$

$$\Pr[\mathscr{E}_2] \leq \frac{1}{4}n^{-1/4}$$

union bound:

$$\Pr[d \text{ is bad}] \leq \Pr[\mathscr{E}_1 \vee \mathscr{E}_2] \leq \Pr[\mathscr{E}_1] + \Pr[\mathscr{E}_2] \leq \frac{1}{2}n^{-1/4}$$

symmetry:

$$\Pr[u \text{ is bad}] \leq \frac{1}{2}n^{-1/4}$$

union bound:

$$\Pr[\text{FAIL}] \leq n^{-1/4}$$

1. Uniformly and independently sample $n^{3/4}$ elements from S to form R; and sort R.

2. Let d be the $(\frac{1}{2}n^{3/4} - \sqrt{n})$th element in R.

3. Let u be the $(\frac{1}{2}n^{3/4} + \sqrt{n})$th element in R.

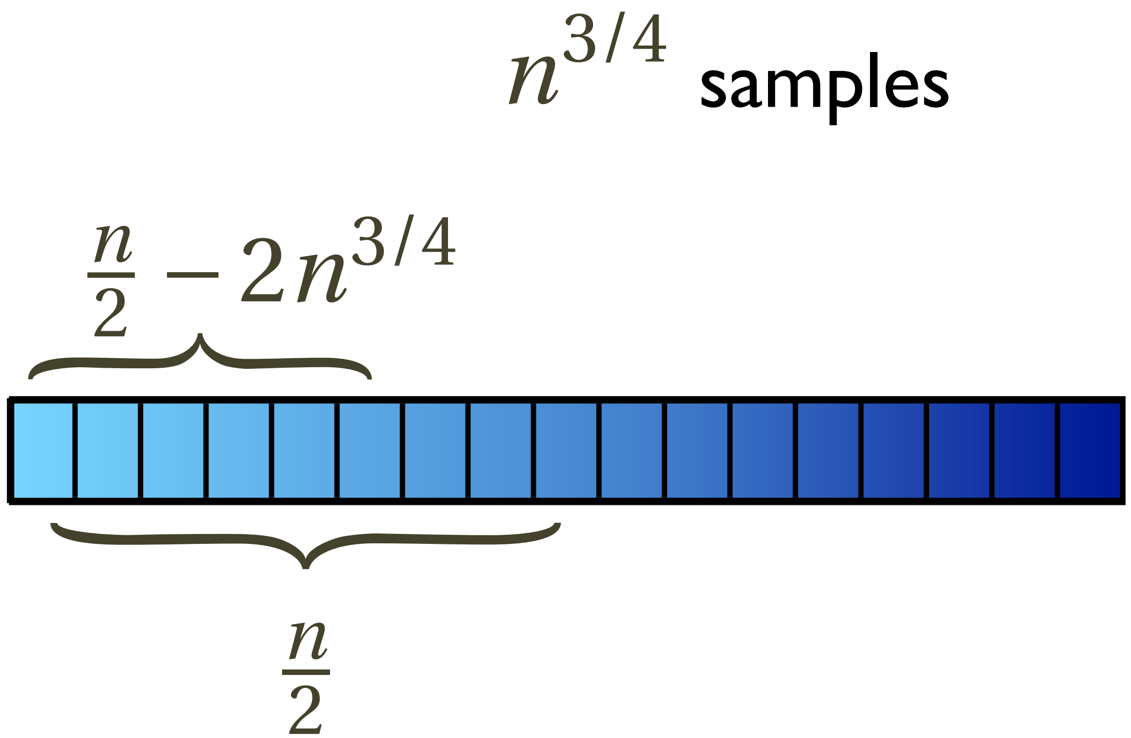
4. If any of the following occurs

   $|\{x \in S \mid x < d\}| > \frac{n}{2}$;

   $|\{x \in S \mid x > u\}| > \frac{n}{2}$;

   $|\{x \in S \mid d \leq x \leq u\}| > s$;

   then FAIL.



5. Find the median of S by sorting C.