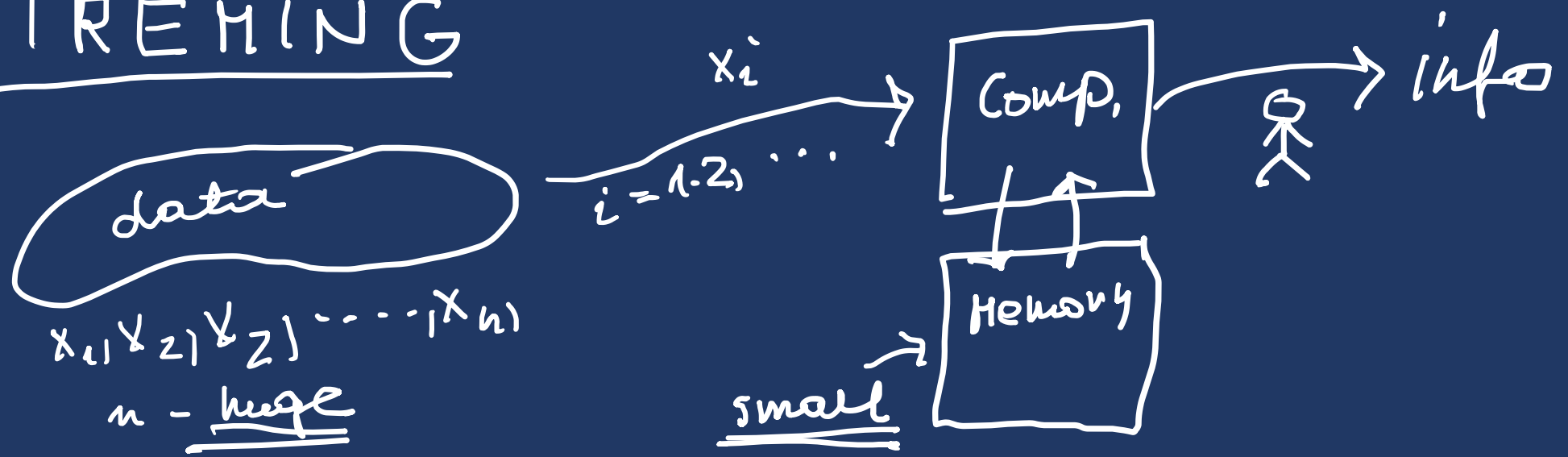
STREMING



Q: 1) n ?

2) $\left(\sum_{i \le n} x_i / n\right)$

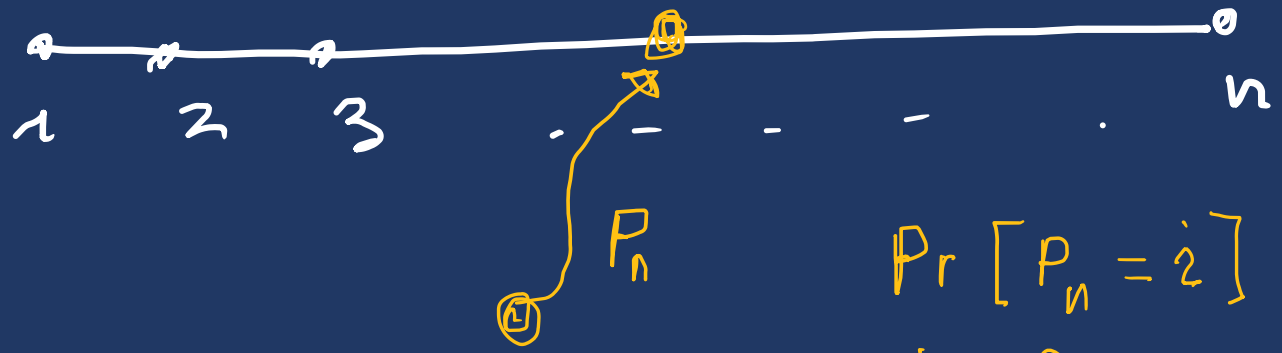
3) $\left(\sum_{i \le n} x_i^2 / n\right)$

} trivial

SAMPLING :

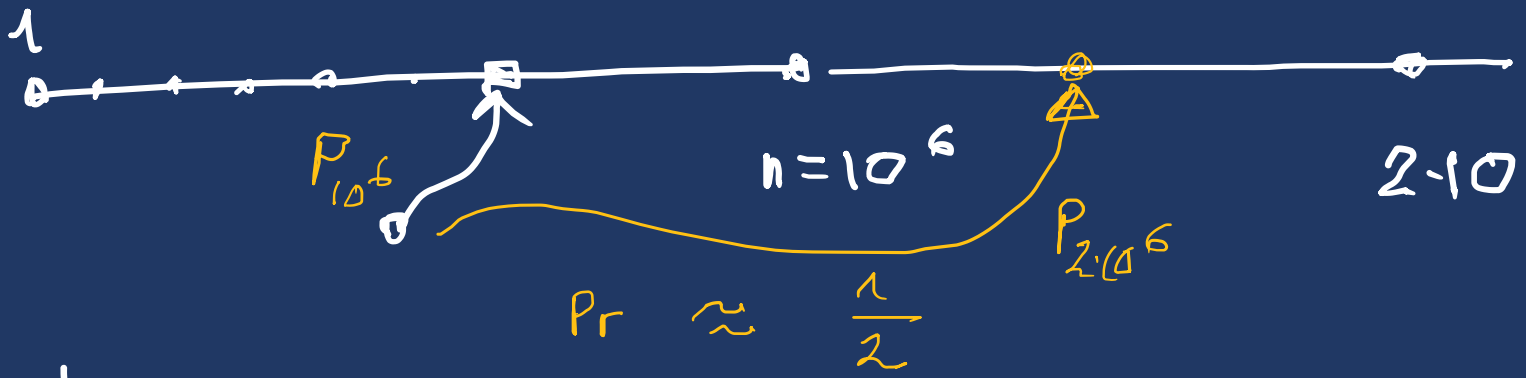
$k \simeq 1000$

sample : uniformly distributed



$$Pr[P_n = i] = \frac{1}{n}$$

$i \in \{1 \dots n\}$



Vitter's R-algorithm.

```
onInit () {
   n = 0 ;   // numbe of elem.
   p = 0 ;   // pointer
data = nil ;  // obs. data by pointer
}
```

```
onGet ( ) {
  return ((p, data));
}
```

```
onRead (x) {
   n++ ;
   if (random (0,1) < 1/n) {
      p = n ;
   } data = x ;
}
```

P_n = the value of pointer after reading n-th element.

- $P_n \in \{1, \dots, n\}$
- GOAL: $(\forall n)(\forall i \in \{1, \dots, n\})\left(\Pr[P_n = i] = \frac{1}{n}\right)$ (*)

1) $P_1 = 1$: ok.
2) assume that (*) holds at n.

C = "there was a change at $n+1$"

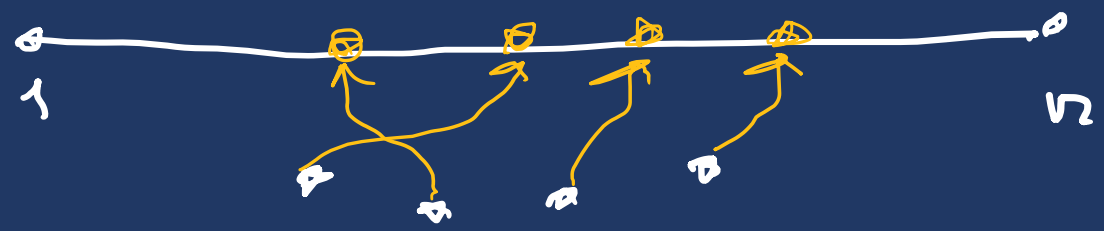
$$P[P_{n+1} = i] = P[P_{n+1} = i \mid C] \cdot \Pr[C] + \Pr[P_{n+1} = i \mid \neg C] \cdot P[\neg C]$$

- $i \in \{1 \dots n\}$: $P[P_{n+1} = i] = 0 \cdot \frac{1}{n+1} + P[P_n = i] \cdot \left(1 - \frac{1}{n+1}\right) =$
$$= \frac{1}{n} \cdot \frac{n+1-1}{n+1} = \frac{1}{n+1}.$$
- $i = n+1$: $P[P_{n+1} = i] = 1 \cdot \frac{1}{n+1} + 0 \cdot \left(1 - \frac{1}{n+1}\right) = \frac{1}{n+1}$

∎

- Build a sample of size k use independently $X_1, \dots, X_k$ from R-alg.

$P[A_1 \cup \dots \cup A_k] \le P(A_1) + \dots + P(A_k)$



$$P\Big[\bigwedge_{1 \le i < j \le k} (X_i \neq X_j)\Big] = 1 - P\Big[\bigvee_{1 \le i < j \le k} (X_i = X_j)\Big] \ge$$

$$1 - \sum_{1 \le i < j \le k} P[X_i = X_j] = 1 - \binom{k}{2} \frac{1}{n} = 1 - \frac{k(k-1)}{2} \cdot \frac{1}{n} \simeq$$

$$\simeq 1 - \frac{k^2}{2n} .$$

- if $n \gg \frac{k^2}{2}$ then w.h.p $X_i \neq X_j$ for all $1 \le i < j \le k$

Birthday paradox $\simeq \sqrt{n}$.

- ~~Different versions~~

we keep in memory array $[X_1, \dots, X_k] = X$

$$V_i = (p_i, data_i)$$

$$(\forall n)\ (\forall A \subseteq \{1..n\})(|A| = k \rightarrow P(\{p_1, \dots, p_k\} = A) = \frac{1}{\binom{n}{k}})$$

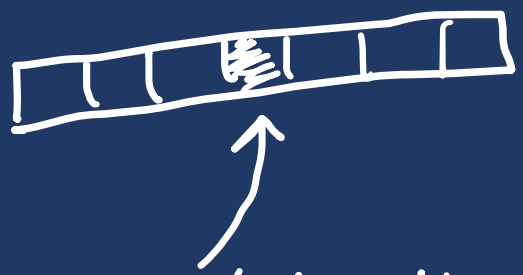
- read $X_1 \dots X_k$; put

$$[(1, x_1), \dots, (k, x_k)]$$

- if $n > k$:

```
if (random(0,1) < k/n) {
    i = random {1, ..., k}
    X[i] = (n, x)
}
```

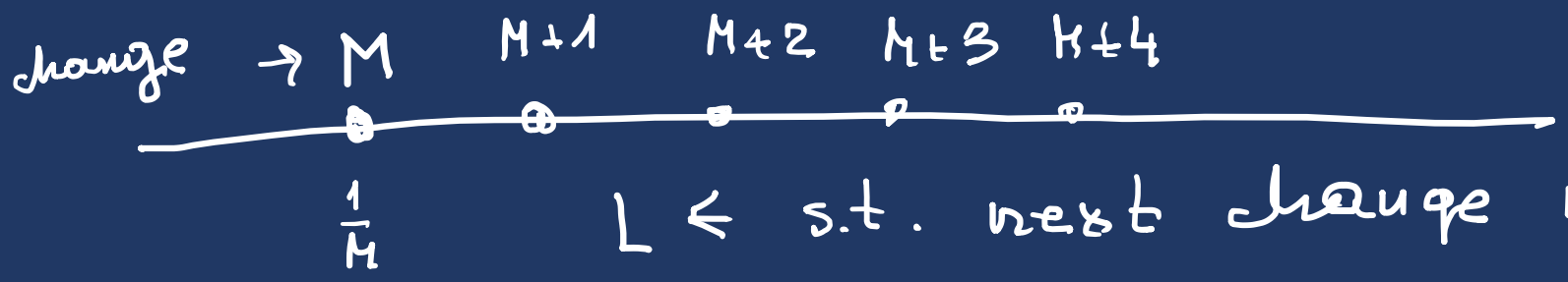
THIS IS CORRECT



modification with prob $\frac{k}{n}$.

EXERCISE

$\binom{n}{k}$ $\binom{n+1}{k}$



$\frac{1}{M}$ $L \leftarrow$ s.t. next change is at $M+L$

$$P[L \geq 1] = 1$$

$$P[L \geq 2] = 1 - \frac{1}{M+1} = \frac{M}{M+1}$$

$$P[L > 2] = P[L \geq 3] = \left(1 - \frac{1}{M+1}\right) \cdot \left(1 - \frac{1}{M+2}\right) = \frac{M}{M+1} \cdot \frac{M+1}{M+2}$$

$$P[L > K] = \frac{M}{M+1} \cdot \frac{M+1}{M+2} \cdot \ldots \cdot \frac{M+(K-1)}{M+K} = \frac{M}{M+K}$$

$$P[L \leq K] = 1 - \frac{M}{M+K} = \frac{K}{M+K} \xrightarrow[K \to \infty]{} 1$$

$$P[L \le k] = \frac{k}{M+k}$$

$$p_i = P[L = i] = \frac{i}{M+i} - \frac{(i-1)}{k+(i-1)}, \quad i = 1, 2, 3, \dots$$

$$\sum_{i=1}^{\infty} p_i = 1$$

$$P[u \in I_4] = p_4$$

$$P[L \le k-1] = p_1 + p_2 + \dots + p_{k-1} < u \le p_1 + \dots + p_k = P[L \le k]$$

u = rand(0,1); find k s.t. (*); return k.

find min k : $u \le \frac{k}{M+k}$

$$u \leq \frac{k}{M+k}$$

$$M u + k \cdot u \leq k$$

$$k - k \cdot u \geq M \cdot u$$

minimal $\rightarrow$ $k \geq \frac{M \cdot u}{1-u}$

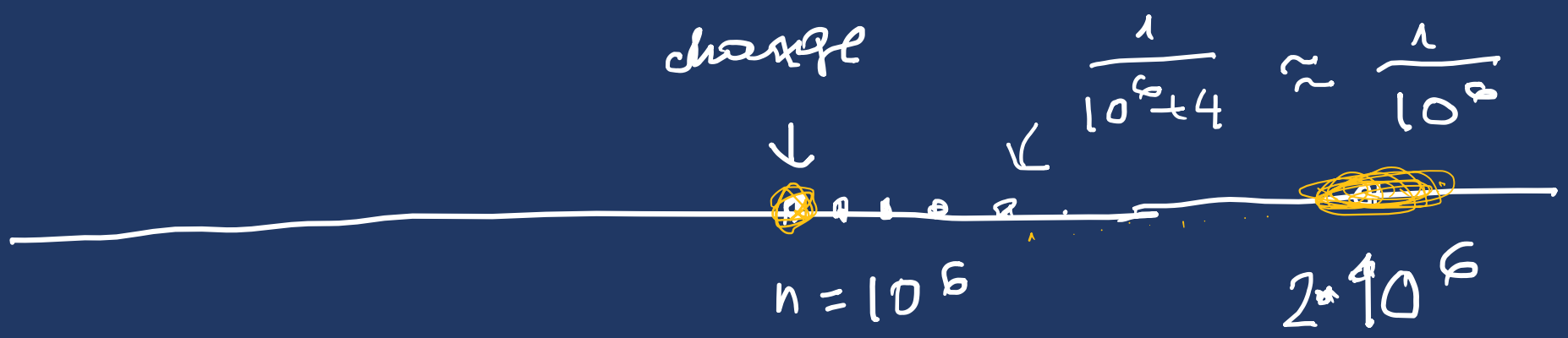
$$k = \left\lceil \frac{M \cdot u}{1-u} \right\rceil$$

```
{ u = rand(0,1)
  return ceiling((M·u)/
         (1-u)) ;
```

```
init : n=0; p=0; data = self; c=0; L=1;

onRead (x) {
    n++; c++;
    if (c==L) {
        p = n;
        data = x;
        u = random(0,1);
        L = ceiling((n*u)/(1-u));
        c = 0;
    }
```

good version of R-algorithm



$L \sim Geo\left(\frac{1}{10^6}\right)$

next change $\approx 2 \cdot 10^6$