STREAMING

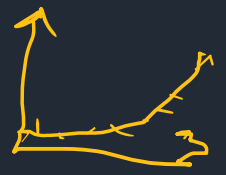
Sampling


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

otrzymujemy zmienną losową o rozkł. jedn.

(sample with replacements)

MOŻLIWE ZMIANY:



```
if (random() < 2/n) { ... }
if (random() < 1/√n) { ... }
```

SLIDING WINDOW

N = liczba ostatnich obserwacji

$mg(X) \in \{1, \ldots, N\}$

rozkł. jednost.

± • 2003 : P. Indyk, Motwani, ...

± • 2012 : Braverman ← prosty

± • 2021 : D. Bojko, M.K, JCL

COUNTING

STREAM : $(a_i)_{i=1\dots N}$

ALPHABET : $|\{a_i : i \in \{1..N\}\}| = 2^{20}$

(P) aabaaacaaaa bb : $|\{a, b, c\}| = 3$

MALWARE : sort: aaaaa bbb c

$h : \Sigma^* \longrightarrow \{0,1\}^n$

$|\{h(a_i) : i = 1..n\}| \leq |\{a_i : i..\}|$

LINEAR COUNTING

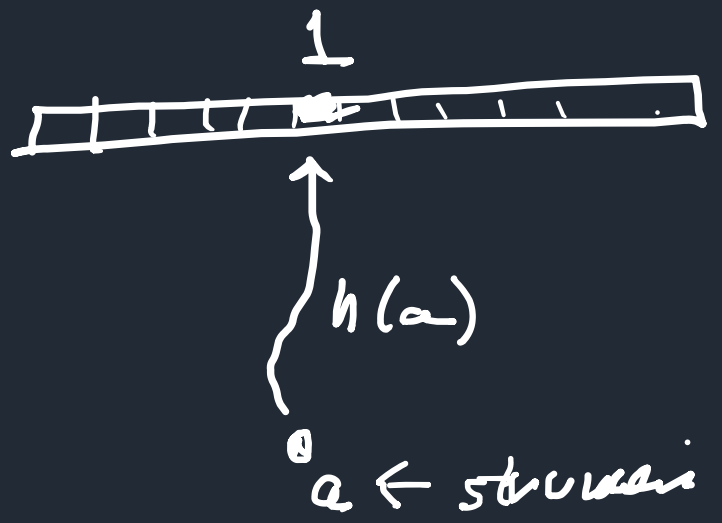
Mamy hash function $h: \Sigma^* \longrightarrow \{1 \ldots, N\}$

Mamy tablicę bitów $X[1 \ldots N]$

INIT : $X \leftarrow (0, 0 \ldots 0)$.

```
On Get(a) {
    X[h(a)] = 1
}
```

Przechowujemy n elementów.



$a \neq b$: $h(a)$ może $h(b)$

$P[u_i = 0] =$ $\{a_1, \dots, a_m\}$ - pewien wzór

$$= \left(\frac{N-1}{N}\right)^m$$

$$L = \sum_{i=1}^{N} [\![u_i = 0]\!]$$

$$E[L] = \sum_{i=1}^{N} E([\![u_i = 0]\!]) = \sum_{i=1}^{N} Pr(u_i = 0) = N\left(1 - \frac{1}{N}\right)^m$$

$m \approx N$

$\approx \frac{N}{e}$

$u \leftarrow$ liczba pustych serii

$$u = N\left(1 - \frac{1}{N}\right)^m \quad ; \quad \frac{u}{N} = \left(1 - \frac{1}{N}\right)^m$$

$$\ln\left(\frac{u}{N}\right) = m \ln\left(1 - \frac{1}{N}\right)$$

$$m = \frac{\ln\left(\frac{u}{N}\right)}{\ln\left(1 - \frac{1}{N}\right)}$$

$$\simeq -N \ln\left(\frac{u}{N}\right) =$$

$$= N \ln \frac{N}{u} \quad \text{0}$$

$$\hat{m} = N \ln \frac{N}{u}$$

Estymator liczby m.

$$\ln \frac{1}{1-x} = x + \frac{1}{2}x^2 + \frac{1}{3}x^3 + \dots$$

$$= \sum_{k \geq 1} \frac{1}{k} x^k$$

$$\ln \frac{1}{1 - \frac{1}{N}} \simeq \frac{1}{N}$$

$$\| \quad \text{1}$$

$$-\ln\left(1 - \frac{1}{N}\right)$$

PROBLEM: $u = 0$

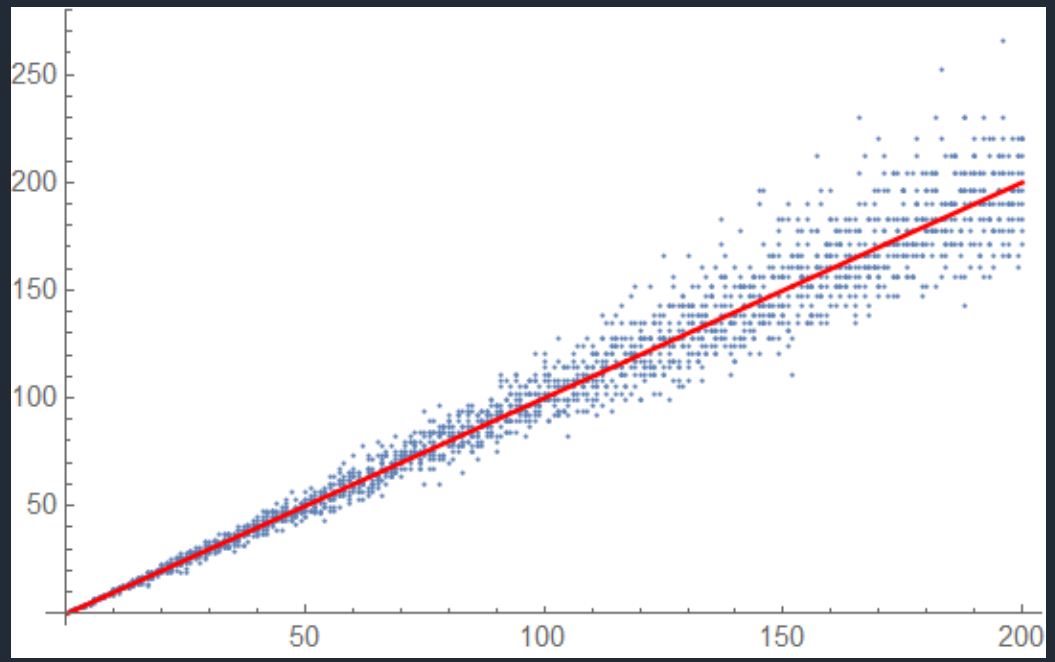
? $u < N \longrightarrow$ NIE

CCP: zupełniacie $\simeq$ $m = N \ln N$

2. Policzcie • $\mathrm{var}(L)$ $(= E[L^2] - (E[L])^2)$

• $\dfrac{\mathrm{var}(L)}{(E[L])^2}$

$$P[|L - E[L]| \geq \alpha E[L]] \leq \frac{\mathrm{var}(L)}{\alpha^2 (E[L])^2}.$$



simulations
$N = 100$;
$k = 1, 2, \ldots, 200$
(10 simulations for each k)

Ograniczenie LIN. COUNTING: $m \ll N \ln N$

LICZNIK PROBABILISTYCZNY

- Licznik; $\begin{cases} \text{init}: n = 0; \\ \text{onGet}(x)\ \{ n{+}{+} \} \end{cases}$

ile bitów potrzebuję do zapisania $n \in \{1, \dots, N\}$

$$L \approx \log_2 N.$$

$a_0 + a_1 2 + \dots + a_k 2^k \leq$

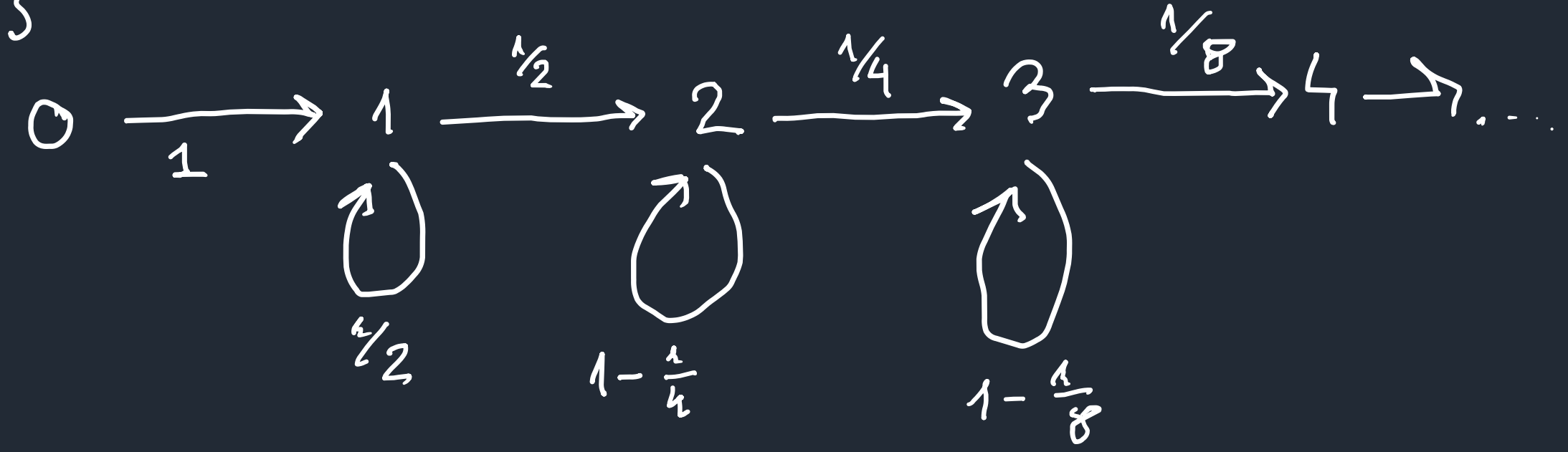
$\leq 1 + \dots + 2^k = 2^{k+1} - 1 \leq N$

- $\approx$ 1955; MORRIS IMB

IDEA: nie muszę liczyć dokładnie!

Licznik Morrisa:

```
init: c = 0
onGet(x) {
    if (random() < 1/2^c) { c++; }
}
```



$$E[2^{C_n}] = ?$$

$$E[2^{C_0}] = E[2^0] = 1$$

$$E[2^{C_1}] = E[2^1] = 2,$$

$n \geq 1$

$$E[2^{C_{n+1}}] = \sum_k 2^k P[C_{n+1} = k] =$$

$$= \sum_k 2^k \left(P[C_{n+1} = k \mid C_n = k] \cdot P[C_n = k] + P[C_{n+1} = k \mid C_n = k-1] \cdot P[C_n = k-1] \right) =$$

$$= \sum_k 2^k \left(\left(1 - \frac{1}{2^k}\right) P[C_n = k] + \frac{1}{2^{k-1}} \cdot P[C_n = k-1] \right) =$$

$$= \underbrace{\sum_k 2^k P[C_n = k]}_{E[2^{C_n}]} - \underbrace{\sum_k P[C_n = k]}_{1} + 2 \underbrace{\sum_k P[C_n = k-1]}_{1}.$$

C_n = wartość licznika po przeczytaniu n-elem.

$$E\left[2^{C_{n+1}}\right] = E\left[2^{C_n}\right] + 1$$

$$E\left[2^{C_0}\right] = 1$$

$$E\left[2^{C_n}\right] = n + 1$$

C - wartość łącznika : $\hat{n} = 2^{C} - 1$

$$\hat{n} = 2^c - 1$$

$$2^c = n + 1$$

$$c \cong \log_2(n+1)$$

Ile potrzebujemy bitów na c?

$$\log_2(\log_2(n+1))$$