Big Data - zagrożenia statystyczne

~ 1950-60 : badania zjawisk paranormalnych

10 000 uczestników:

odgadywanie ciągu 0/1 długości 10

- wynik I : ok 10 osób odgaduje poprawnie cały ciąg...
- wynik II : nikt nie odgaduje poprawnie

WYJAŚNIENIE : są osoby o nadnaturalnych

ALE : tracą je gdy dowiedzą się o nich.

$X_1, \dots, X_{10} \sim \mathrm{Ber}(\frac{1}{2})$ $\quad \mathbb{P}_0[X_i] = \begin{cases} \frac{1}{2} : X_i = 1 \\ \frac{1}{2} : X_i = 0 \end{cases}$

$$Y = X_1 + \dots + X_n$$

$$\Pr[Y = 10] = \left(\frac{1}{2}\right)^{10} \approx \frac{1}{1000}$$

$$Y_1, \dots, Y_{10000} \qquad S_i = [\![Y_i = 10]\!]$$

$$L = S_1 + \dots + S_{10000}$$

$$E[L] = 10000 \cdot \Pr[Y_i = 10] \approx \frac{10000}{1000} = 10$$

回

P. $\underbrace{0\,1\,0\,1\,0\,1\,0\,1\,\ldots\ldots\ldots}_{M \approx 10^7}$

jaka jest szansa, że $\frac{1}{2}$

Zadanie
zaś: to jest
ciąg losowych

Q. Ile ~~[illegible]~~ osób dziennie
w Polsce ramo widzi czarnego kota
a potem (w ciągu dnia) traci kobotę ?

MAP - REDUCE

① inverted index of word

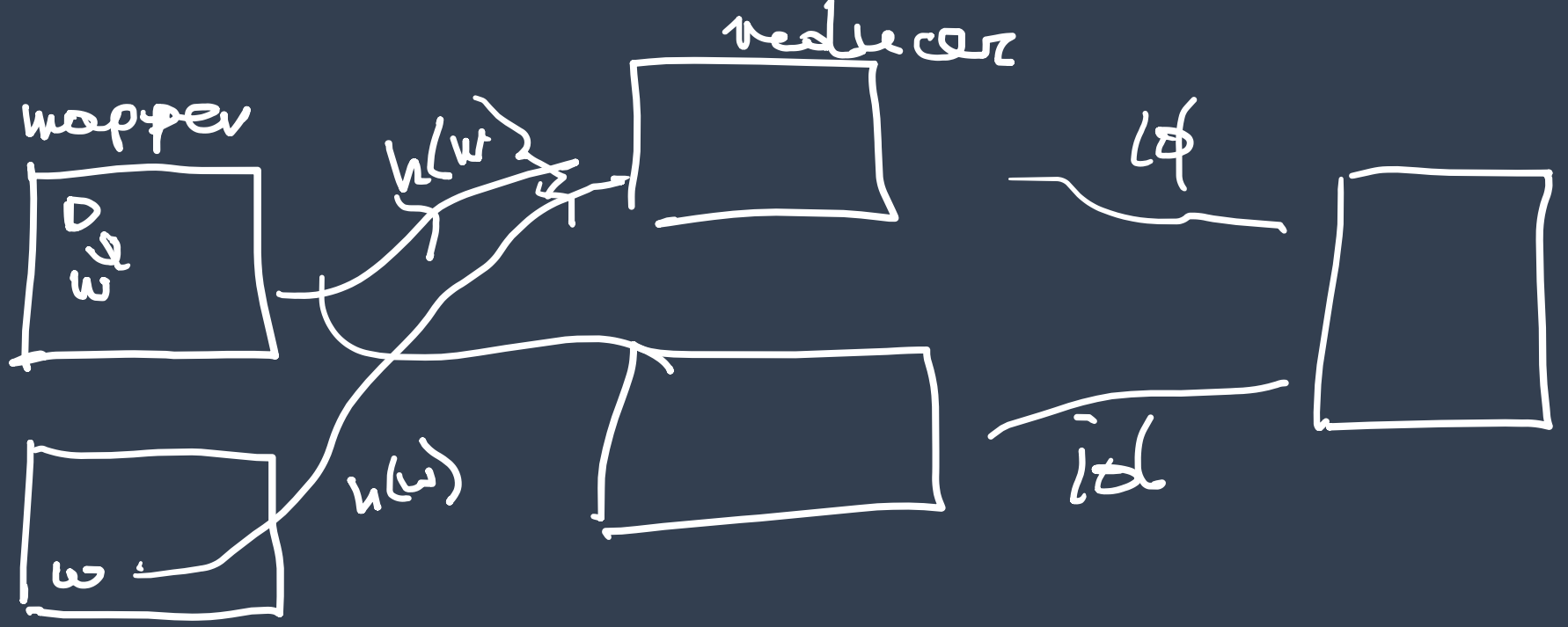
N <- duże

CEL: dla danego słowa w wyznacz listę dokumentów w których w występuje.

```
map(IdD, Doc) {
  forall w ∈ Doc do {
    emit(w, IdD);
  }
}
```

$[w_1, L_1], [w_2, [id^2_1, id^2_2, .. id^2_k]]$

```
reduce(w, L) {
  emit(w, L);
}
```



② Agregaty : $[x_1, \ldots, x_n] \rightsquigarrow$ suma, minimum, maksimum
(double)

? $x \rightsquigarrow$ klucz

$\varphi(x) =$ 10 ostatnich bitów $(x)_2$

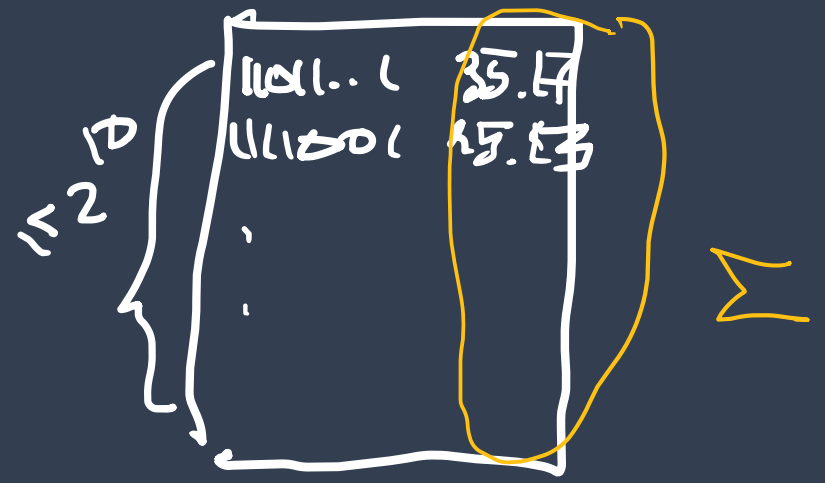
Suma

```
map(x) {
    emit(φ(x), x);
}
```

```
reduce(k, L) {
    emit(k, ΣL);
}
```

Redcer

$$["\text{słowo}", [y_1\ y_2 - \dots\ y_L]]$$



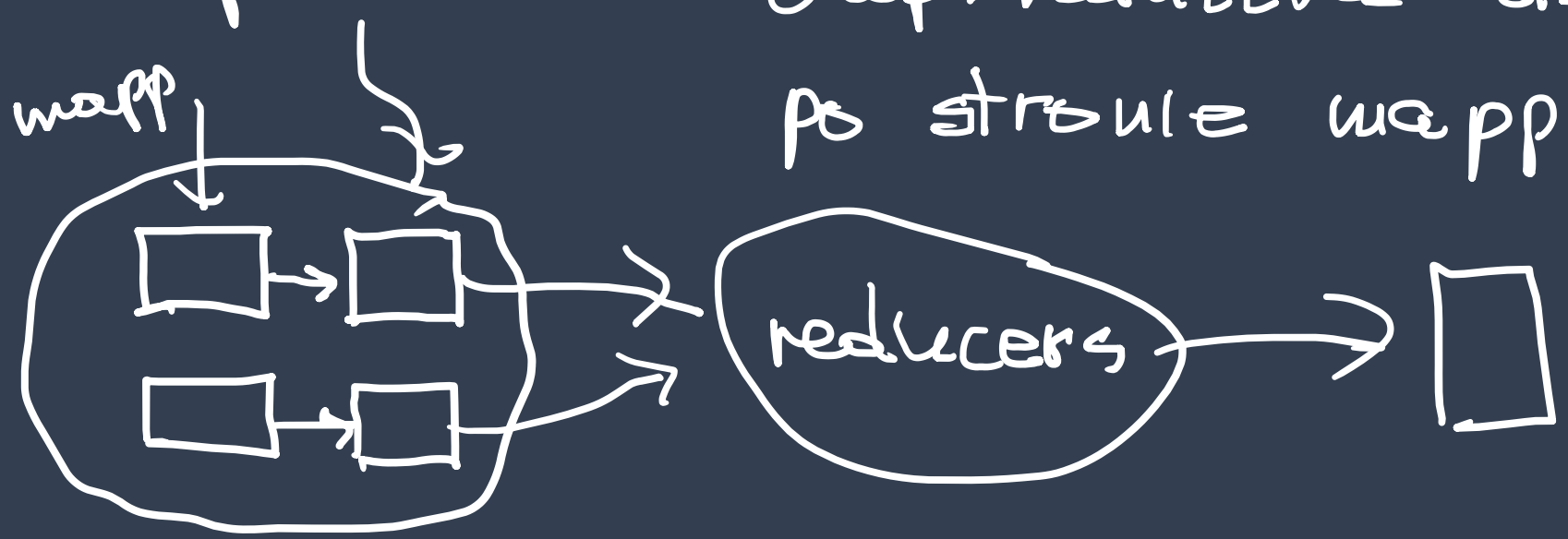
$$\sum_i \sum_j a_{ij} =$$

$$= (a_{11} + \dots + a_{1n_1}) + (a_{21} + \dots + a_{2n_2}) + \dots + (a_{m1} + \dots + a_{mn_m})$$

+ jest przemienne i łączne.

najczęściej frequent

Composer ≡ gdy reducerа działającej po stronie mappera



SUMY:

$$\begin{cases} \text{map}(x) \{ \text{emit}(\varphi(x), x) \} \\ \text{composer} (k, L) \{ \text{emit}(k, \Sigma L); \} \\ \text{reduce} (k, L) \{ \text{emit}(k, \Sigma L); \} \end{cases}$$

Przykłady łączny dla operacji: Σ, min, max

Uwaga: $*$ - łączne $\quad \ln(x_1 \cdot \ldots \cdot x_n) = \sum_{i=1}^{n} \ln(x_i)$

mnożenie:

map(x) {emit($\varphi(x)$, ln(x))}

Średnia : nie jest łączna

```
map(x) {emit(φ(x), x, 1); }
reducer (k, L){    // L = [(x_1,1),(x_2,1)] ... ,(x_n,1)]
   emit ( k, Σ π_1(L), length(L) )
}
```

$$\begin{bmatrix} k_1 & s_1 & L_1 \\ k_2 & s_2 & L_2 \\ \vdots & & \\ k_m & s_m & L_m \end{bmatrix} \longrightarrow \frac{s_1 + \ldots + s_m}{L_1 + \ldots + L_m}$$ średnia

ZADANIE : zrób to z składnikiem $(x, L_1) \oplus (x_2, L_2)$

Zad (P) $(x_1, \dots, x_n) \rightsquigarrow \frac{\sum x_i}{N}, \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$.

μ — var

$$\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 = \dots = \frac{1}{N}\sum x_i^2 - \left(\frac{\sum x_i}{N}\right)^2$$

zadanie

odpow: $\mathrm{Var}(X) = E[X^2] - (E[X])^2$.

```
map(x) { emit(φ(1), (1, x, x^2)); }
```

$1 \to N$, $x \to \sum x_i$, $x^2 \to \sum x_i^2$

SIŁA MAP-REDUCE

: LIPA

siła ≡ rozbicie zadania na
niezależne zadania

OPERACJE ZBIORÓW

dane: A, B ← zbiory CEL: A∪B, A∩B, A\B

x ∈ A : ("a", x)

x ∈ B : ("b", x)

⇝ mapper

```
map(k, x) {emit(x, k);}
```

suma:

```
reduce(x, L) {emit(x);}
```

iloczyn

```
reduce(x, L) { if |L| = 2 then emit(x); }
```

różnica (A\B)

```
reduce(x, L) { if (L = ["a"]) then emit(x); }
```

Operacje bazo-danowe :

Relacja : $R(a_1 .., a_n)$, $\theta(a_1 .. a_n) \leftarrow$ Booleen

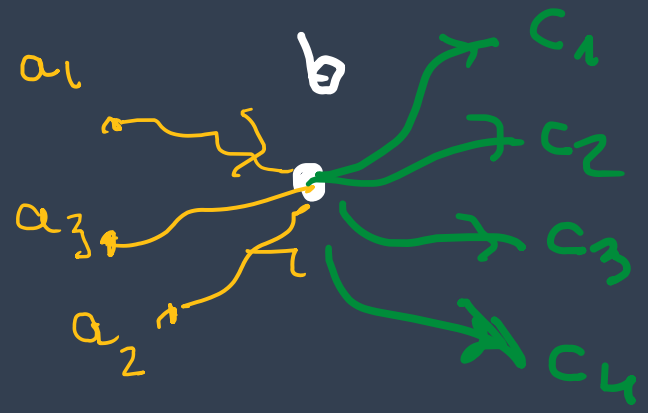
SELECT $a_1 .. a_n$ FROM R WHERE $\theta(a_1 .. a_n)$

map$(a_1 \ldots a_n)$ { if $\theta(\underbrace{a_1 .. a_n}_{\vec{a}})$ then emit $(\varphi(\vec{a}), \vec{a})$

$\varphi(a_1 .. a_n) = a_1 \| a_2 \| .. \| a_n$

JOIN : $R(a,b)$, $S(b,c)$ relație BD

SELECT R.a, S.c FROM R, S
WHERE R.b = S.b



done : (rel, x, y)
↑
R / S

```
map (r, x, y) {
   if (r = "R") {emit (y, ("R", x))
   else         {emit (x, ("S", y))
}
```

reduce (b, L) {

1) L_1 = posortuj L po 1 współrz.

2) rozbijamy

$L_1 = L_{11} \,||\, L_{12}$

↑ pierwsze R ↑ pierwsze S

3) forall $x \in \pi_2(L_{11})$ do
forall $y \in \pi_2(L_{12})$ do
emit (x, y);

}

$L = [(r_1, x_1), (r_2, x_2), \ldots, (r_m, x_m)]$

"R" lub "S"

$R \bowtie S$

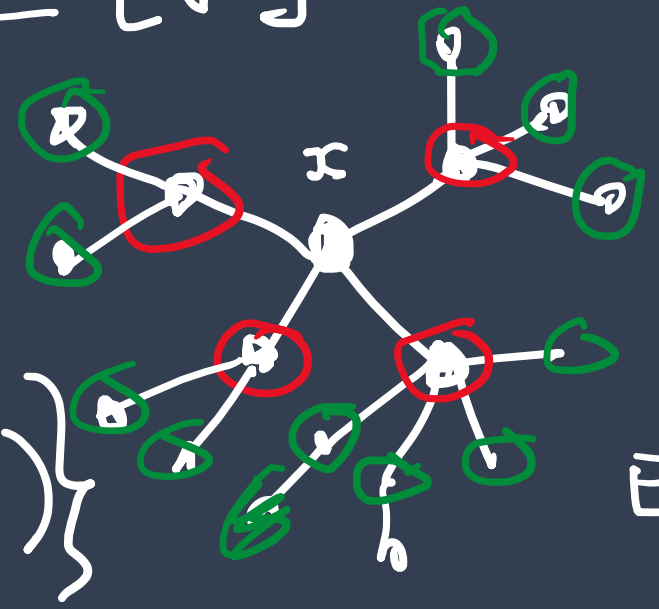
$L_1 = [\underbrace{(R, x_1), (R, x_2), \ldots, (R, x_k)}, (S, y_1), \ldots (S, y_e)]$

GRAFY : $G = (V, E)$ ← graf prosty

$E \subseteq [V]^2$

$N(x) = \{y : \{x,y\} \in E\}$

$N_2^+(x) = \{z : (\exists y)(\{x,y\} \in E \wedge \{y,z\} \in E)\}$



E

x	y_1
x	y_2
x	y_3
⋮	⋮

$E^* = E \cup flip(E)$

JOIN $E^* \circ E^*$.

(2) napisz to porządnie