

# Współczesna teoria informacji

- ▶ Rozpoczyna się w I połowie XX wieku
- ▶ Podobnie jak wcześniej chciano przyspieszyć przekaz informacji tak aby był jasny sam w sobie. Wykorzystywano jednak coraz to nowsze osiągnięcia ludzkiego umysłu co zdecydowanie ułatwiało te dążenia. Współcześnie zatem przyspieszana jest przesyłka danych. Szczególnie po przez:
  - Rozważania na temat przepustowości sieci
  - Kompresję danych przesyłanych

# Przepustowość

- ▶ Aby zwiększyć przepustowość wbudowanych linii przesyłowych wykorzystywano cztery rodzaje napięcia: -3 V, -1 V, 1 V oraz 3 V.
- ▶ Starano się również wysyłać wiadomości jak najszybciej to możliwe, jednakże przy odpowiednio szybkim wprowadzaniu impulsów, na większym dystansie wysyłane fale zlewały się co często uniemożliwiało prawidłowe odczytanie wiadomości.
- ▶ Elementem mocno związanym z przepustowością jest liczba rodzaj symboli lub znaków użytych do przesyłania wiadomości. Im większe ich urozmaicenie tym więcej jesteśmy w stanie przekazać informacji w tym samym czasie.
- ▶ Odpowiada temu wzór  $S^N$ : gdzie S to liczba symboli, a N to ilość przesłanych symboli na sekundę.
- ▶ Np. w przekazie opisanym wyżej (różne napięcia elektryczne), przyjmując możliwość przesyłania 2 impulsów na sekundę możemy przekazać prawie 1000 znaków w ciągu minuty.

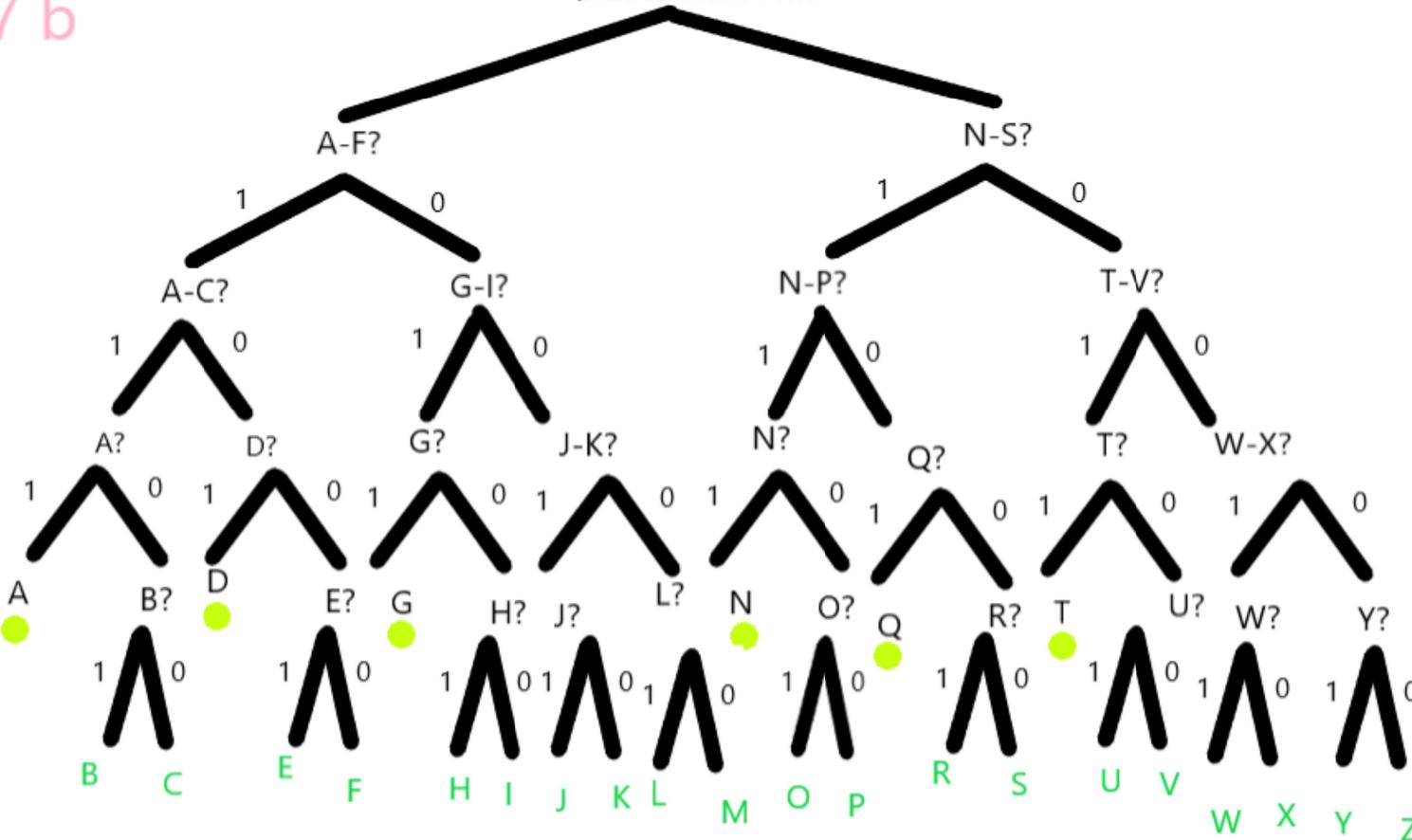
Przenosząc przepustowość na konkretne zagadnienie związane z informatyką...

Ile bitów potrzeba, aby określić dowolną literę wskazaną przez osobę?

Avg 4,7 b

A, B, C, D,  
E, F, G, H,  
I, J, K, L,  
M,  
N,  
O, P, Q, R,  
S, T, U, V  
W, X, Y, Z

Czy litera należy do  
przedziału: A-M?

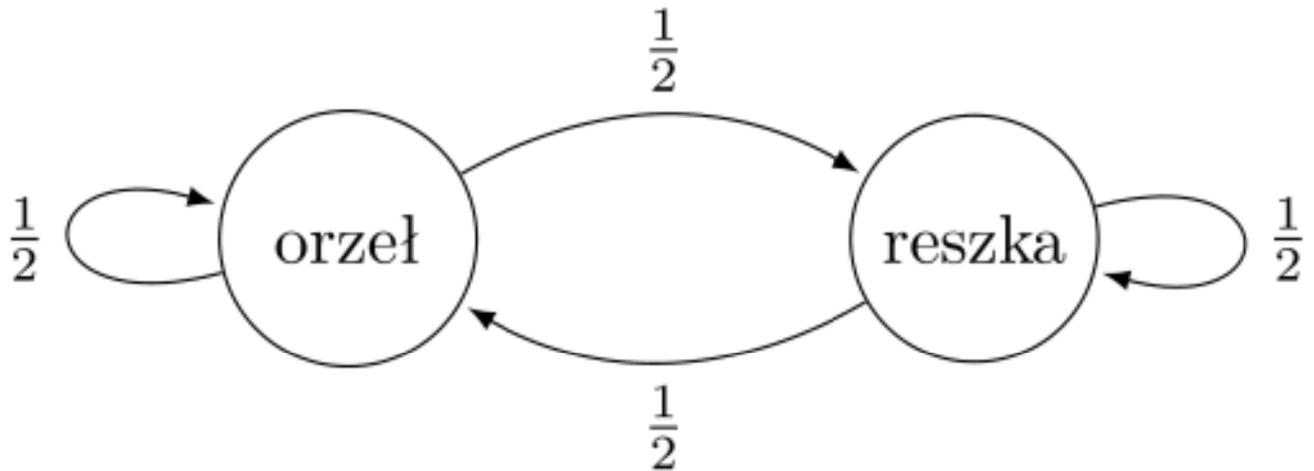


# Prawdopodobieństwo a kompresja danych

Bardzo pomocne w kompresji danych okazało się prawdopodobieństwo.

Andriej Markow w 1906 r. opracował podstawy, które później udoskonalone pozwoliły na utworzenie łańcucha Markowa. Jest to ciąg zdarzeń losowych, w których wynik poprzedniego losowania warunkuje wynik losowania następującego po nim.

Dzisiaj schemat ten wykorzystuje m. in. firma Google w sortowaniu wyszukiwanych stron, czwli w PageRank'u



# Kompresja danych

Cały czas dążono do kompresji danych. Jej częścią jest skracanie kodu równoznacznego z danym znakiem do minimum. Z pomocą prawdopodobieństwa i łańcucha Markowa udaje się to w znacznym stopniu. Wykorzystując zależność wysokości prawdopodobieństwa stwierdzono, że można do zapisu najczęściej występujących informacji wykorzystać mniej bitów, a do rzadziej występujących więcej bitów.

I tak na przykład:

Przyjmijmy, że w alfabetie od A do D prawdopodobieństwo kolejnych liter wynosi:  
A- 50% , D - 25 % oraz B i C po 12,5%.

Wykorzystując tą informację możemy skompresować dane potrzebne do ich przekazu w np.. w systemie binarnym dokładnie tak:

A- 1B- 0II C- 0I0 D- 00 \*

\* Musimy jednak uważać na powtarzające się znaki, gdyż kod 0II przy oznaczeniu litery D = 0, mógłby oznaczać jednocześnie= „B” = „DAA”, dlatego ważne jest tworzenie unikalnych kodów

# Entropia

Zależność tą opisał m.in. Claude Shannon. W 1948 r. w dziele „Matematyczna teoria informacji” opisał podwaliny pod pojęcie dzisiaj rozumiane jako teorię informacji. Opisując kompresję danych stworzył wzór na średnią liczbę bitów potrzebnych do zapisu danych. Nazywa się to Entropią. Shannon przebadał tą zależność również na języku angielskim rozpatrując poszczególne ciągi liter (jedno oraz kilku literowe), a także na ciągu pojedynczych oraz kilku słów.

Wzór na obliczanie entropii:

$$H(X) = \sum_{i=1}^n p(x_i) \log_r \frac{1}{p(x_i)} = - \sum_{i=1}^n p(x_i) \log_r p(x_i)$$

H(X) - entropia

p(x) - prawdopodobieństwo wystąpienia konkretnej zmiennej

r - liczba zmiennych (np. w systemie binarnym dwie)