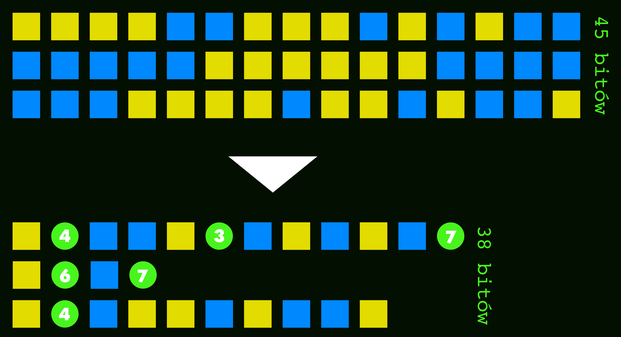
KOMPRESJA DANYCH

Dane to ciągi bitów, czyli zero-jedynkowych wartości oznaczających obecność lub brak sygnału elektrycznego w układzie elektronicznym przetwarzającym dane. Ilość miejsca zajmowanego przez bity jest bardzo cenna - dlatego powstał cały szereg technik pozwalających na ich kompresję. Jest to nic innego, jak zmniejszenie ilości zajmowanego miejsca przez takie zakodowanie informacji, które umożliwia wyrażenie jej za pomocą mniejszej liczby bitów. Obecnie większość szeroko wykorzystywanych plików graficznych czy wideo stosuje różne rodzaje kompresji pozwalającej wyrazić złożoną informację w sposób "zużywający" znacznie mniej miejsca do ich przechowania.



Przykładowe dane zawierają wiele powtarzających się sekwencji bitów - możemy je częściowo zastąpić zliczeniami tych bitów (np. gdy występuje obok siebie 3 lub więcej bitów).

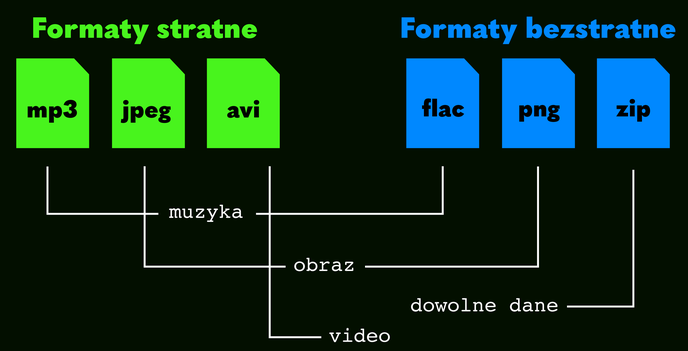
1 0 = 1 bit

● = 3 bity (w naszym przypadku maksymalne "zliczenie" wynosi 7 - do zapisu takiej wartości w systemie dwójkowym potrzebujemy 3 bitów (dziesiętne 7 = dwójkowe 111)

Redukcja informacji

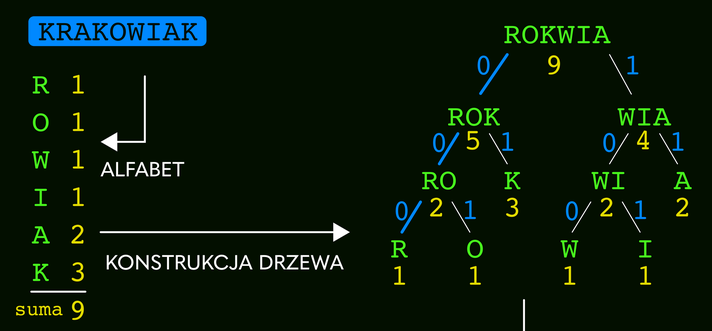
Kompresja danych to redukcja ilości informacji potrzebnej do przekazania danej treści. Może być **stratna** (tracimy pewne szczegóły informacji ale jej treść pozostaje czytelna) oraz **bezstratna** (możemy odtworzyć 100% oryginalnej informacji dekompresując dane).

Wiele popularnych plików przechowujących muzykę, wideo lub grafikę stosuje różne typy kompresji.



Drzewo Huffmana

Dawniej skompresowanie ciągu dowolnych znaków można było osiągnąć za pomocą tzw. drzewa Huffmana. Jego zastosowanie polega na skonstruowaniu grafu, w którym litery/znaki grupowane są wg ich częstości w kodowanym zdaniu. Graf wykorzystje się następnie do zastąpienia każdego znaku ciągiem bitów (zer i jedynek). Spróbujmy zakodować słowo **KRAKOWIAK**.



1
Częstość (wagi) znaków i słów alfabetu

0/1
Wartości binarne każdego rozwidlenia drzewa

Odnalezienie każdego znaku na drzewie Huffmana wraz z kodowaniem binarnym każdej gałęzi pozwala na zakodowanie oryginalnych znaków za pomocą bitów. W tym przypadku wszystkie znaki zakodowane są za pomocą maksymalnie 3 bitów.

R 000
O 001
W 100
I 101
A 11
K 01

Oryginalne słowo zapisane w kodowaniu Huffmana wygląda tak jak poniżej. Oryginalny tekst składał się ze znaków literowych ASCII, których zakodowanie **wymaga 8 bitów/znak, co daje sumaryczną zawartość informacji 9×8 = 72 bity**. Po zastosowaniu drzewa Huffmana ciąg znaków ma **22 bity, czyli ponad 3 razy mniej**.

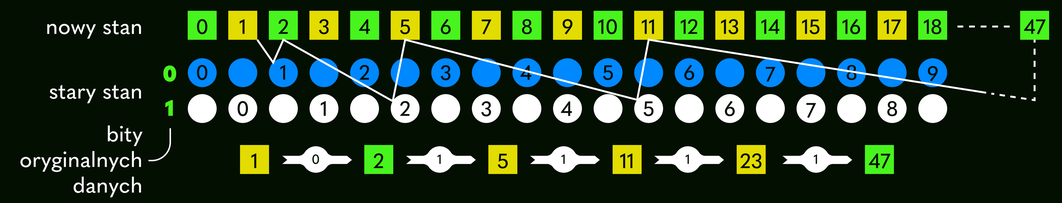
KRAKOWIAK 72 bity

K R A K O W I A K
01 000 11 01 001 100 101 11 01 22 bity (3.27x mniej)

Kodowanie asymetryczne

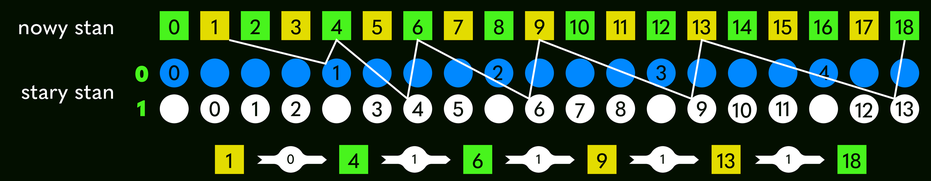
Zadanie: zakodowanie danych 01111 w postaci liczby naturalnej.

Przepis: rozpoczynając od 1, dla każdego bitu kodowanej informacji (01111) dodaj do rosnącej liczby odpowiednią wartość z tablicy kodowej wg reguły: **nowy stan = 2×stary stan + bit**.



"Standardowy" układ dwójkowy zakłada **równy udział zer i jedynek** w zapisywanej informacji. Stąd w powyższej tabeli liczby naturalne używane do kodowania (●●) rozmieszczone są w równych odstępach, takich samych dla kodowania zera (górny rząd●) i jedynki (dolny rząd●).

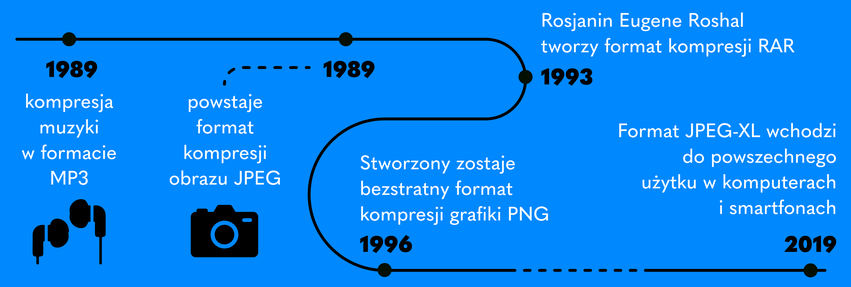
Ale - w naszym przypadku informacja zawiera nadmiar jedynek (01111). Zastosujmy więc "**asymetryczny system dwójkowy**" gdzie liczby kodujące jedynki występują gęściej (poniżej na każdą liczbę kodującą dla bitu "zero" przypadają trzy liczby kodujące dla bitu "jeden"). To tak, jakby **bity przyjmowały wartości ułamkowe**. Matematyczny przepis przyjmuje postać: **nowy stan = stary stan/prawdopodobieństwo zaobserwowania danego bitu**.



Kodowanie standardowym układem dwójkowym **wymaga dłuższej tablicy liczb, i zwraca wartość 47**. Ścieżka kodowania (poniżej) jest znacznie dłuższa, mniej efektywna i wymaga do działania większej ilości pamięci.

Kodowanie systemem asymetrycznym **zwraca wartość 18 (zajmującą mniej miejsca), jest szybsze (wymaga mniejszej tablicy liczb)**. Jest więc ono **mniej kosztowne**, bo **bierze pod uwagę "skład" oryginalnej informacji**..

Wykorzystanie kodowania asymetrycznego (ANS) do kompresji to polski wynalazek! **W latach 2006-2013 dr Jarosław Duda z Wydziału Matematyki i Informatyki UJ** wprowadził ten rodzaj algorytmu, który dzięki zastosowaniu "ułamkowych" bitów pozwala uzyskać lepszą kompresję oraz szybszą dekompresję danych. Wprowadzany właśnie nowy **format JPEG-XL, oparty na kompresji ANS,** lada chwila wejdzie do powszechnego użycia zastępując format JPEG i nawet trzykrotnie polepszając sprawność kompresji plików graficznych.



#NAUKOSKOP

Dla Cafe Nauka, Uniwersytet Jagielloński. Projekt graficzny: Drobne Kreski / szdrobniak.pl • Źródła: D. Salomon "Concise introduction to data compression"; http://th.if.uj.edu.pl/~dudaj/