

MODELOWANIE RZECZYWISTOŚCI

Daniel Wójcik

Instytut Biologii Doświadczalnej PAN
Szkoła Wyższa Psychologii Społecznej

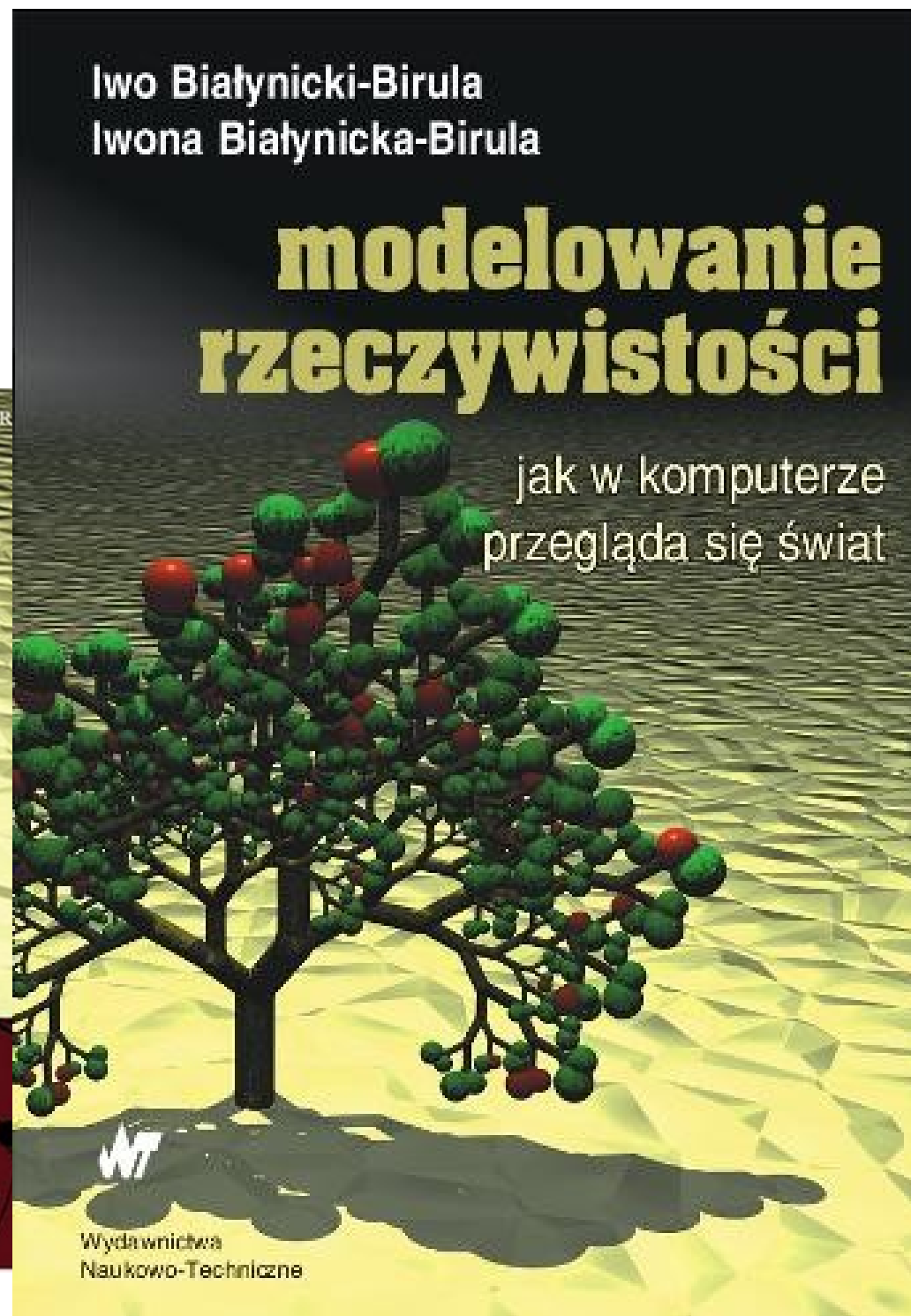
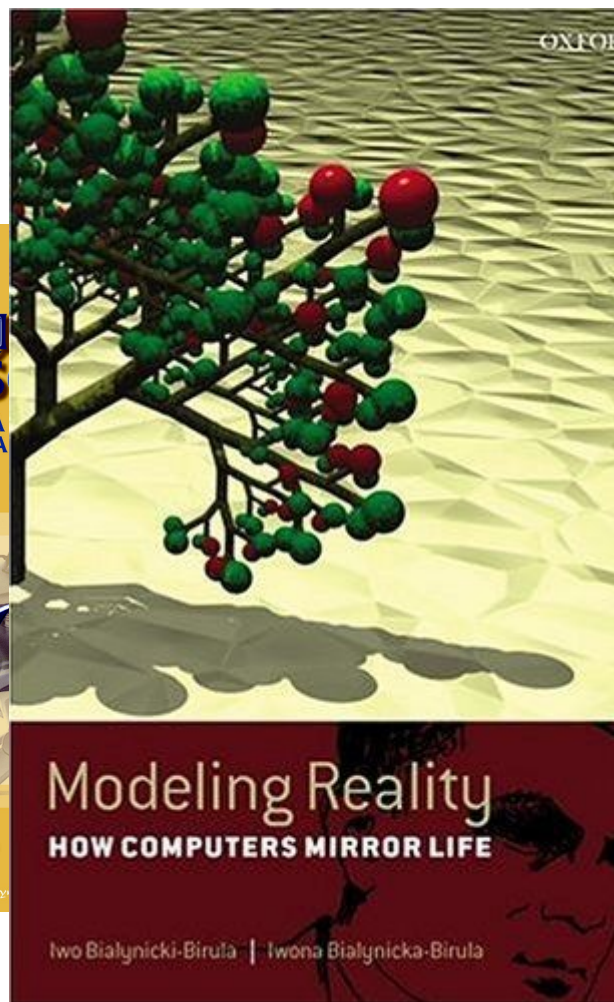
d.wojcik@nencki.gov.pl
dwojcik@swps.edu.pl

tel. 022 5892 424

<http://www.neuroinf.pl/Members/danek/swps/>

Podręcznik

Iwo Białynicki-Birula
Iwona Białynicka-Birula



Informacja i niepewność

- Matematyczna teoria informacji zajmuje się pojemnością kanału transmisji informacji, zupełnie abstrahuje od znaczenia, wartości i sensu przekazywanej informacji.
- W teorii informacji możemy porównywać zawartość informacji w zdaniach
 - Szła dziewczeczka do laseczka
 - Masa zakrzywia czasoprzestrzeńnie zważając na ich treść

Informacja i niepewność

- Informacja i niepewność to dwie strony tego samego medalu: zdobywając informację usuwamy niepewność i na odwrót, tracąc informację powiększamy niepewność.
- Im większa niepewność co do poszukiwanego wyniku, tym więcej informacji zdobywamy poznając ten wynik.
- Informacja jak woda

Bit

- Miarą informacji jest **bit** – skrót od binary digit. Jest to miara informacji otrzymanej w odpowiedzi na **elementarne pytanie**, to jest pytanie na które odpowiedź może brzmieć tylko „tak” lub „nie”.
- Większe jednostki to **bajt, kilobajt, megabajt, gigabajt, terabajt, itd.**
- **UWAGA:**
kilometr to $1000=10^3$ metrów
kilobajt to $1024 = 2^{10}$ bajtów

Nośniki informacji

- Informacja może mieć różne postacie: dźwięku, obrazu, tekstu, filmu. My skupimy się na tekście.
- Tekst jest uporządkowanym ciągiem znaków z pewnego alfabetu. Ten sam tekst można zapisać w różnych alfabetach, podobnie jak liczby można zapisać w różnych systemach.

Nośniki informacji

- Im krótszy alfabet, tym więcej musimy napisać, żeby przekazać tą samą informację.
- Żeby kodować informację potrzebujemy przynajmniej 2 symboli.
- Będziemy używać alfabetu dwójkowego (binarnego). Wtedy każdy znak niesie ze sobą jeden bit informacji.

Własności informacji

- Przyjmijmy, że informacja jest zapisana w alfabecie binarnym (0,1).
- **Słowem binarnym** jest ciąg zer i jedynek o długości N . Liczba N mierzy objętość nośnika informacji. Informacja zawarta w słowie jest proporcjonalna do N .
- Informacja, jaka może być zawarta w danym ciągu znaków jest proporcjonalna do długości tego ciągu. (Informacja jest wielkością **ekstensywną**)

Miara informacji

- Postulujemy zatem, żeby miarą informacji była długość słowa binarnego

Informacja $H = \text{Długość_słowa_binarnego}$

- Istnieje 2^N słów binarnych o długości N znaków.
Zatem

$\text{Długość_słowa} = N = \log_2 (\text{Liczba_słów}) = \log_2 (2^N)$

Miara informacji cd

- Jeżeli prawdopodobieństwo wystąpienia każdego słowa jest takie samo, to wynosi ono

$$p = \frac{1}{\text{Liczba_słów}}$$

- Ponieważ $\log(1/x) = -\log x$
więc informacja zawarta
w pojedynczym słowie wynosi

$$H = -\log_2(p)$$

Wzór Shannona

- Claude Shannon uogólnił tą intuicję do przypadków, kiedy różne słowa występują z różnymi prawdopodobieństwami
- Według niego, zawartość informacyjna przekazu złożonego z n znaków wyrażona jest przez prawdopodobieństwa p_i występowania tych znaków

$$H = - \sum_{i=1}^n p_i \log_2(p_i)$$

Wzór Shannona

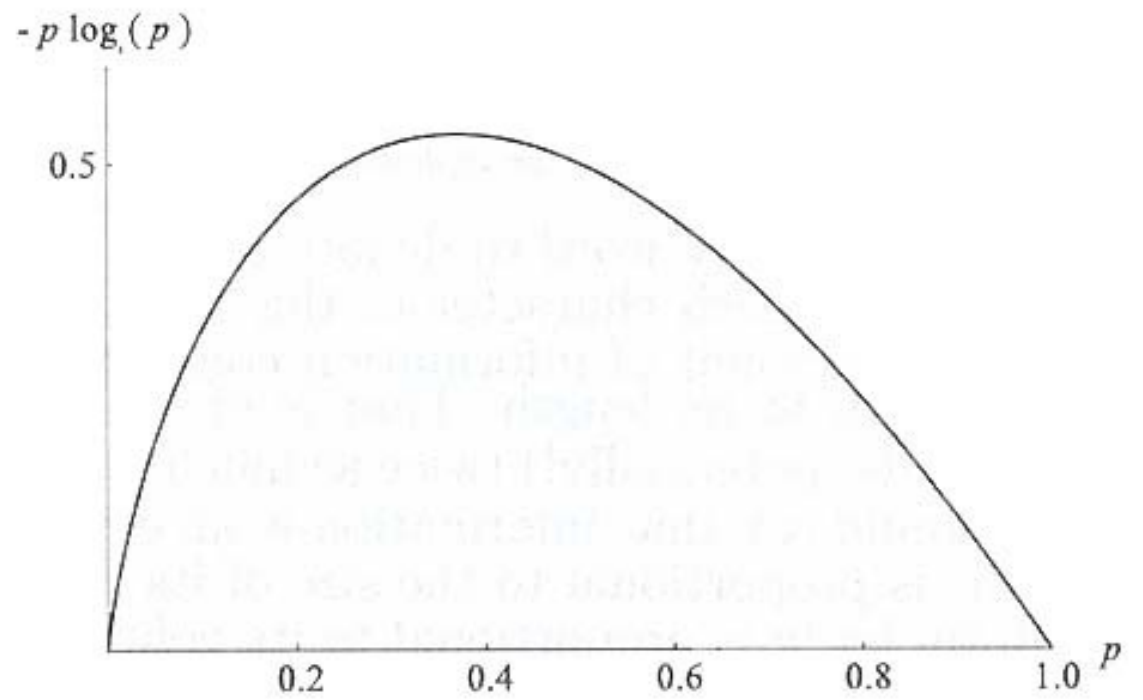
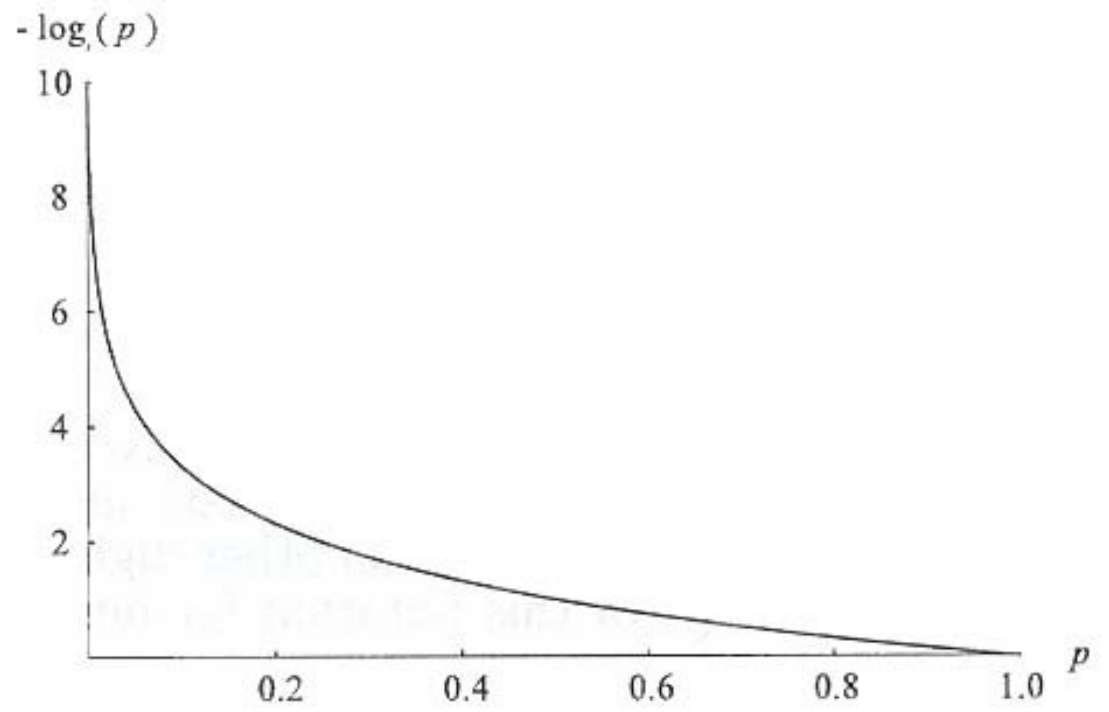
- Wielkość H

$$H = - \sum_{i=1}^n p_i \log_2(p_i)$$

oznacza informację mierzona w bitach.
Nazywa się ją **entropią informacyjną**.

- H jest nieujemna, bo logarytm z liczby od 0 do 1 jest ujemny

$$-P \log_2(P)$$

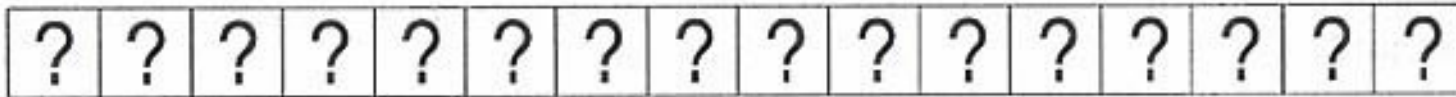


Gra w 20 pytań

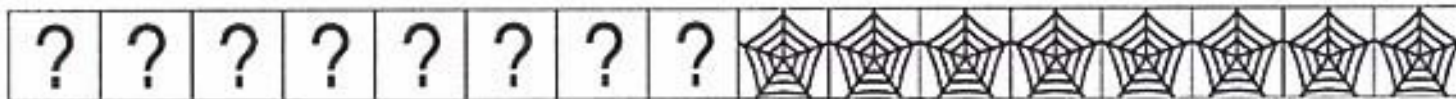
- Pokażemy, że entropia równa jest liczbie elementarnych pytań potrzebnych do odgadnięcia słowa.
- Rozważmy uproszczoną sytuację, kiedy jest 2^N równoprawdopodobnych słów o długości N . Ponumerujmy je wszystkimi liczbami naturalnymi od 1 do 2^N .
- Mamy 2^N zakrytych komórek. W jednej z nich jest „skarbu”. Znalezienie „skarbu” jest tym samym co odgadnięcie słowa.

Optymalna strategia

Initially we know nothing:



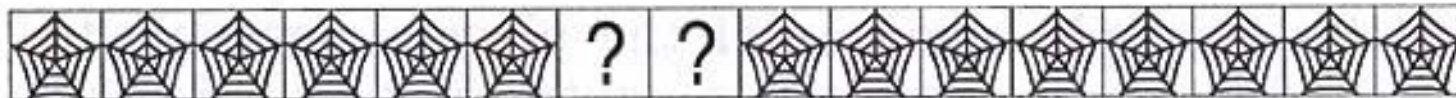
After the first question we know:



After the second question we know:



After the third question we know:



After four questions we know where the treasure is:



20 pytań cd

- Liczba pytań potrzebna do uzyskania pełnej informacji równa jest początkowej niepewności
- Na ogół prawdopodobieństwa wystąpienia różnych słów nie są takie same:
 - Kiedy grając w szubienicę mamy odgadnąć słowo postaci **KU_A** nie wiemy, czy jest to **KUFA**, **KULA**, **KUMA**, **KUŃA**, **KUPA** czy **KURA**.
 - Kiedy mamy słowo postaci **ŚW_T**, to nie ma problemu.

Rozkład 1

- Rozważmy skarb ukryty w jednej z 4 komórek, z prawdopodobieństwami $p_1 = 0.5$, $p_2 = 0.25$, $p_3 = 0.125$, $p_4 = 0.125$
- Pierwotna strategia daje średnio 2 pytania do osiągnięcia sukcesu
- Lepsza strategia:
 - Czy skarb jest w pierwszej komórce?
 - Czy skarb jest w drugiej komórce?
 - Czy skarb jest w trzeciej komórce?
- Średnio prowadzi ona do $1*0.5+2*0.25+3*0.25=7/4 < 2$ pytań zatem jest lepsza od bisekcji.

Rozkład 2

- Rozważmy rozkład 6 komórkowy:

$$p_1 = 1/3, \quad p_2 = 1/5, \quad p_3 = 1/5,$$
$$p_4 = 2/15, \quad p_5 = 1/15, \quad p_6 = 1/15$$

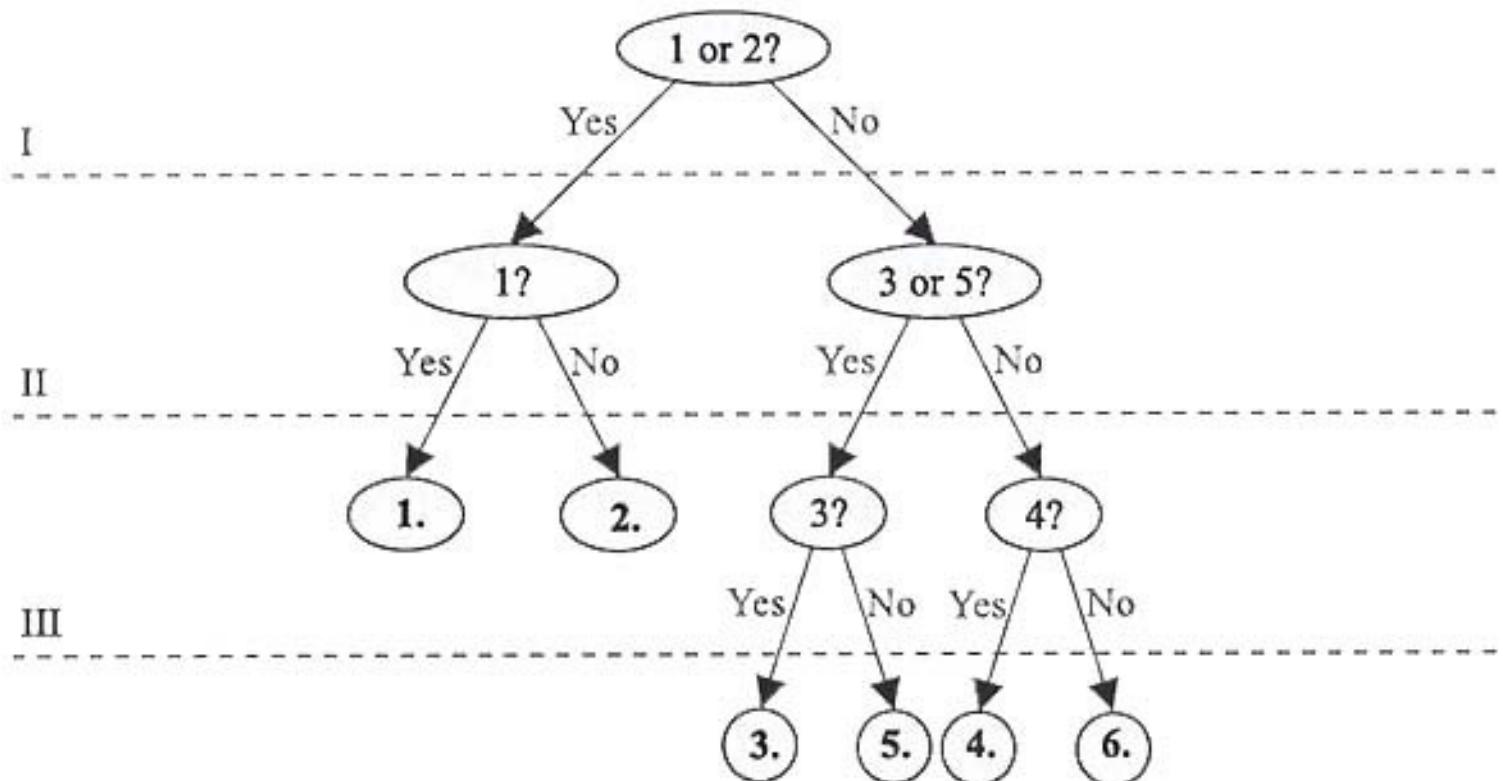
Rozkład 2

- Rozważmy rozkład 6 komórkowy:

$$p_1 = 1/3, \quad p_2 = 1/5, \quad p_3 = 1/5,$$
$$p_4 = 2/15, \quad p_5 = 1/15, \quad p_6 = 1/15$$

Równomierny
podział:

średnia liczba
pytań wynosi
 $37/15$



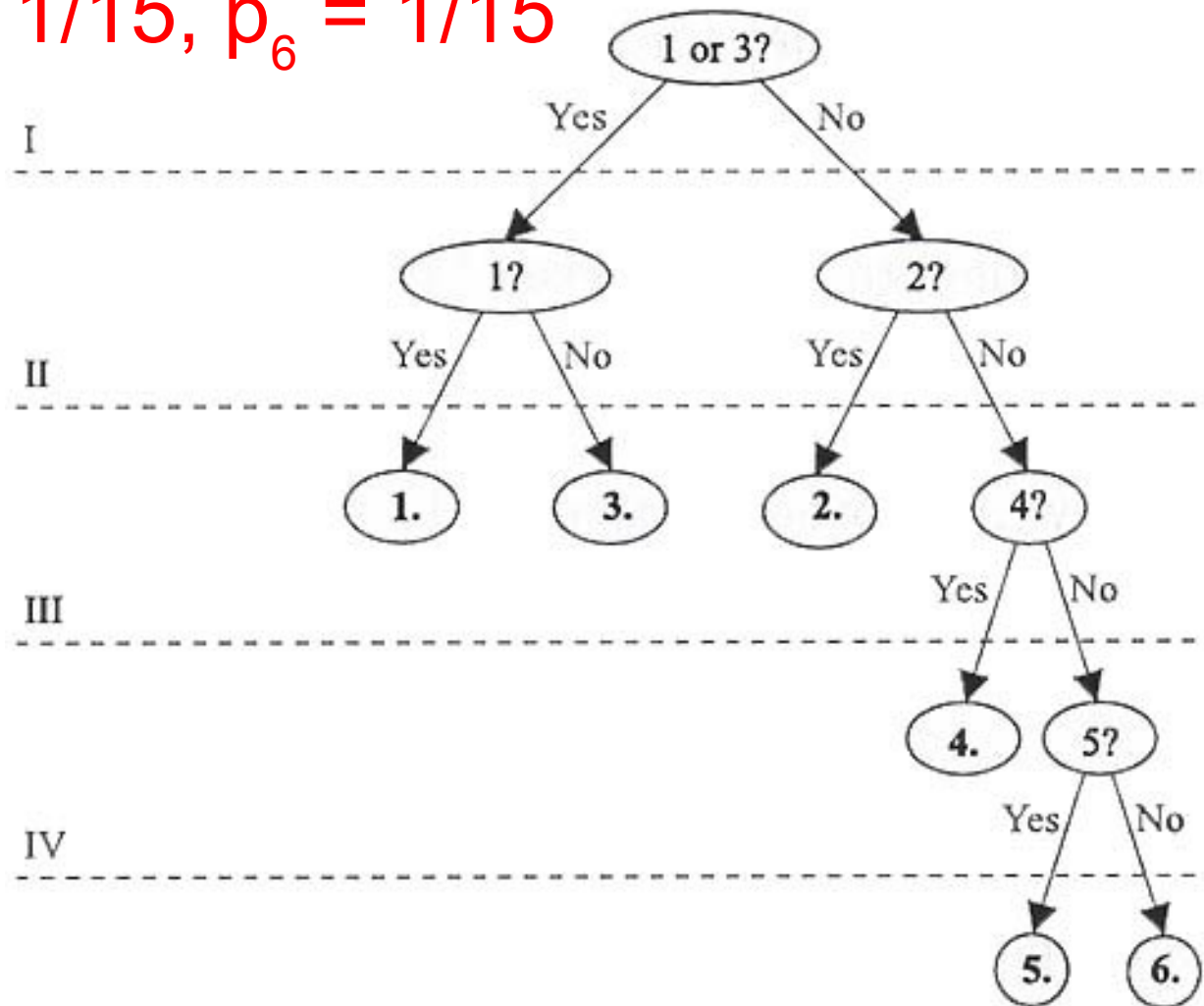
Rozkład 2 – druga strategia

- Rozważmy rozkład 6 komórkowy:

$$p_1 = 1/3, \quad p_2 = 1/5, \quad p_3 = 1/5,$$
$$p_4 = 2/15, \quad p_5 = 1/15, \quad p_6 = 1/15$$

Średnia liczba
pytań wynosi tu
 $36/15$

Czy można to
zmniejszyć?



Optymalna strategia – algorytm Huffmana

- Z początkowego rozkładu $p^0_1, p^0_2, \dots, p^0_n$ wybieramy dwa najmniej prawdopodobne zdarzenia p^0_i oraz p^0_j .
- Łączymy je w jedno o prawdopodobieństwie p^1_k , mamy nowy rozkład $p^1_1, p^1_2, \dots, p^1_{n-1}$
- Powtarzamy procedurę $n-1$ razy
- W ten sposób, od dołu, powstaje optymalne **drzewo pytań**.

Przykład działania strategii

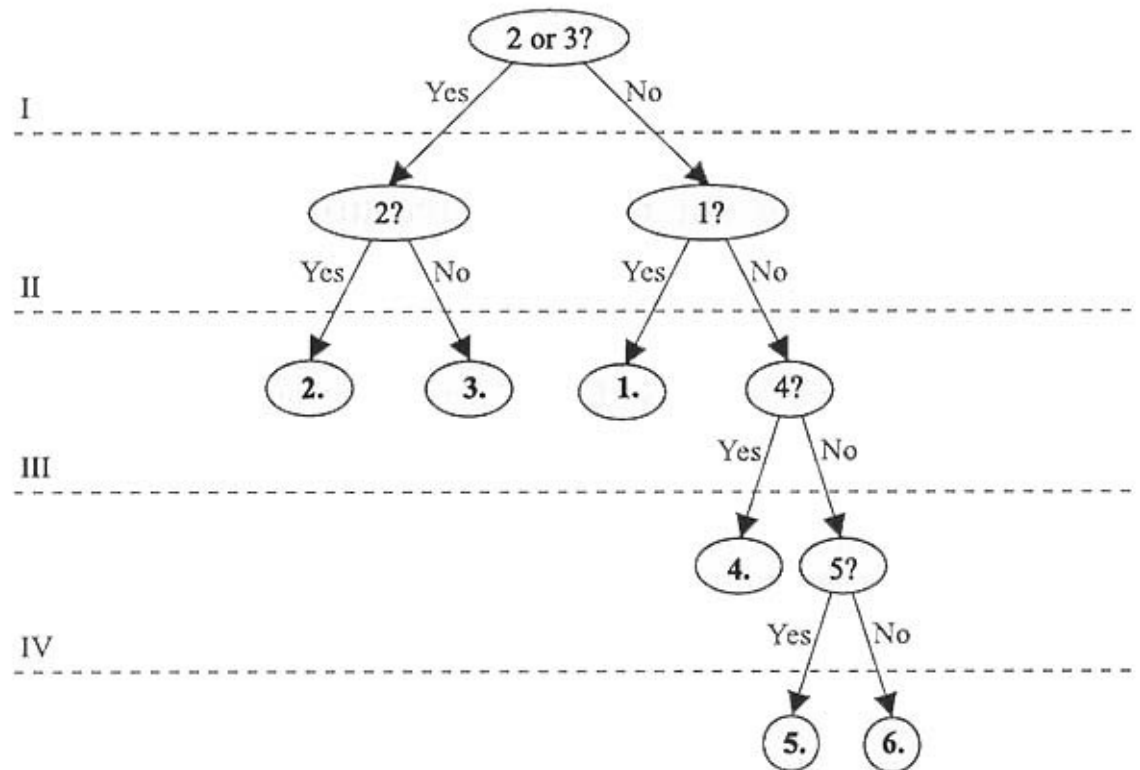
- Zaczynamy od rozkładu $p^0_1=1/3$, $p^0_2=1/5$,
 $p^0_3=1/5$, $p^0_4=2/15$, $p^0_5=1/15$, $p^0_6=1/15$
- Łączymy p^0_5 i p^0_6 w p^1_5 . Dostajemy:
 $p^1_1=1/3$, $p^1_2=1/5$, $p^1_3=1/5$, $p^1_4=2/15$, $p^1_5=2/15$
- Łączymy p^1_4 i p^1_5 w p^2_4 . Dostajemy:
 $p^2_1=1/3$, $p^2_2=1/5$, $p^2_3=1/5$, $p^2_4=4/15$
- Łączymy p^2_2 i p^2_3 w p^3_3 . Dostajemy:
 $p^3_1=1/3$, $p^3_2=4/15$, $p^3_3=6/15$
- Łączymy p^3_1 i p^3_2 w p^4_2 . Dostajemy:
 $p^4_1=9/15$, $p^4_2=6/15$

Rozkład 2 – trzecia strategia

- Rozważmy rozkład 6 komórkowy:

$$p_1 = 1/3, \quad p_2 = 1/5, \quad p_3 = 1/5,$$
$$p_4 = 2/15, \quad p_5 = 1/15, \quad p_6 = 1/15$$

Średnia liczba
pytań wynosi tu
również $36/15$

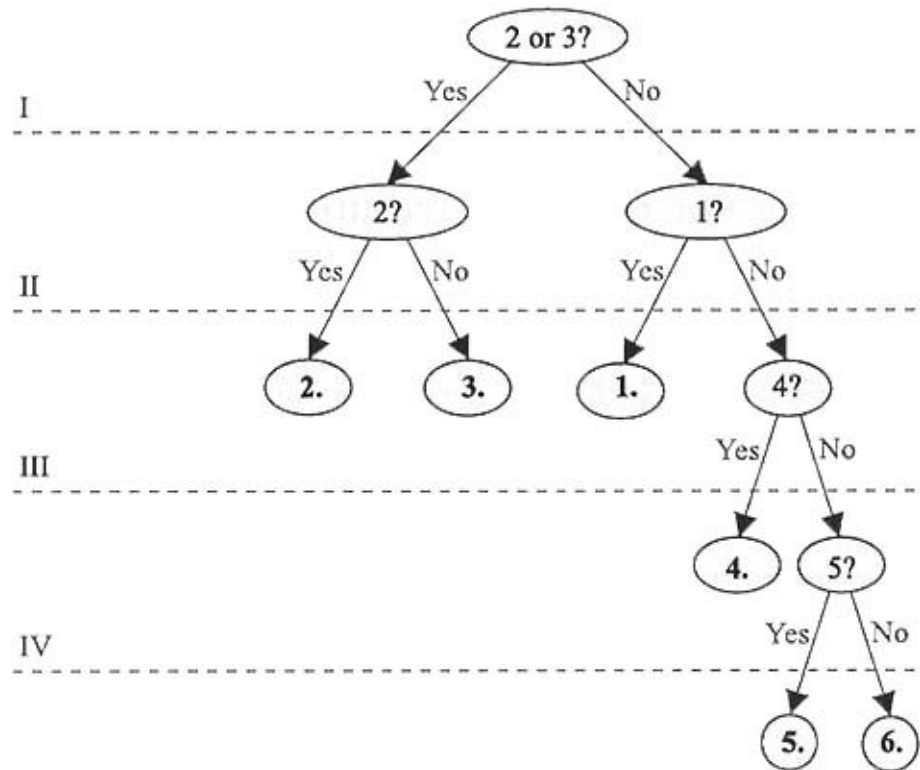
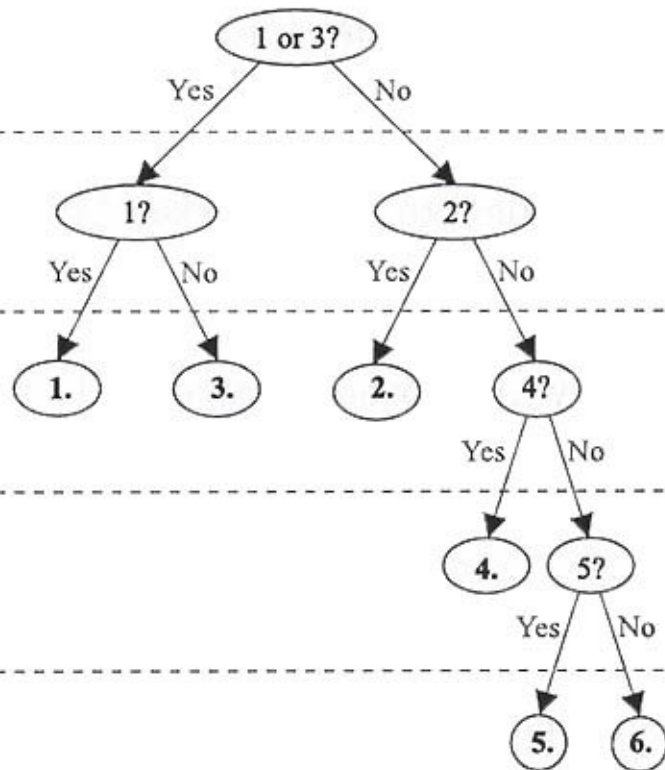


Porównanie dwóch ostatnich strategii

- Rozważmy rozkład 6 komórkowy:

$$p_1 = 1/3, \quad p_2 = 1/5, \quad p_3 = 1/5,$$

$$p_4 = 2/15, \quad p_5 = 1/15, \quad p_6 = 1/15$$



Średnia liczba pytań wynosi dla obu strategii $36/15$

Średnia informacja

- Nasze rozważania pokazują, że entropia Shannona mierzy średnią informację obliczoną w przypadku, gdy znane są wszystkie prawdopodobieństwa elementarne.
- Ogólne twierdzenie o bezszumowym kodowaniu możemy sformułować tak:

Nie istnieje strategia o średnio mniejszej liczbie pytań niż entropia Shannona

Doświadczenia Hymana

- R. Hyman pokazał, że czas reakcji na bodźce o określonej zawartości informacji jest proporcjonalny do entropii Shannona.
- Program **Hyman**