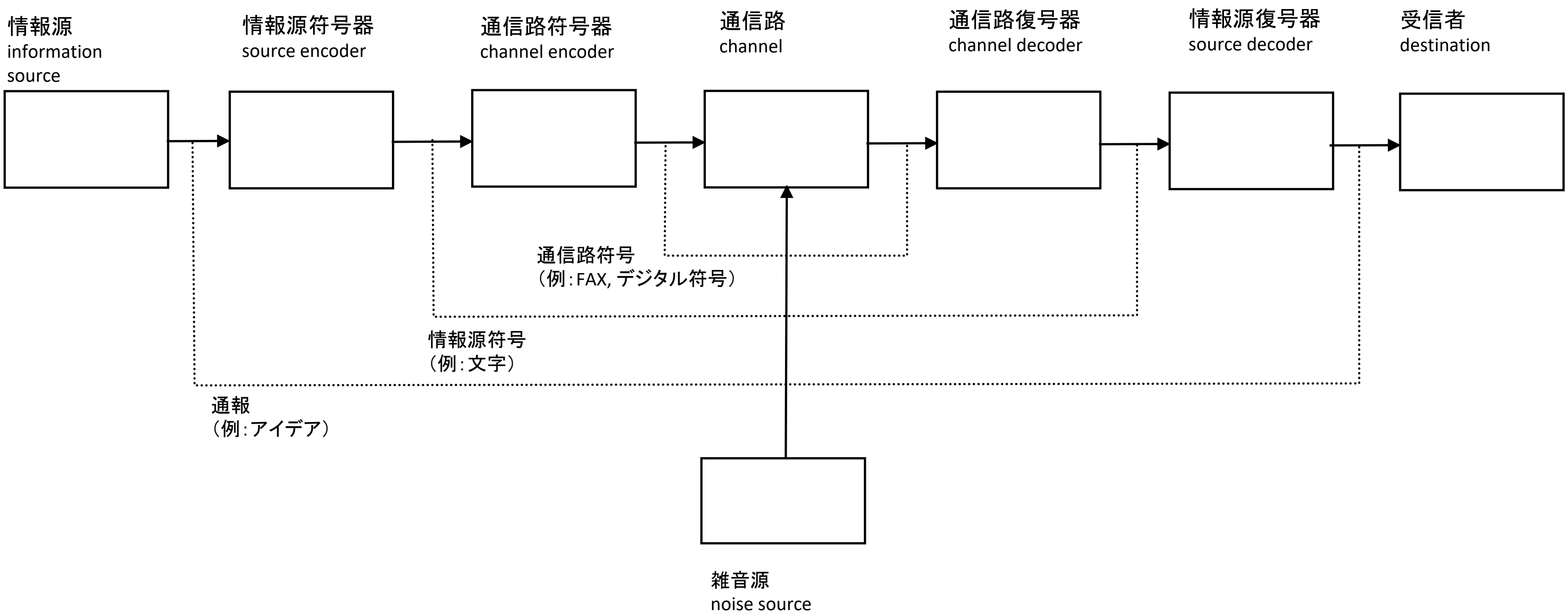
# 情報源符号化とその限界

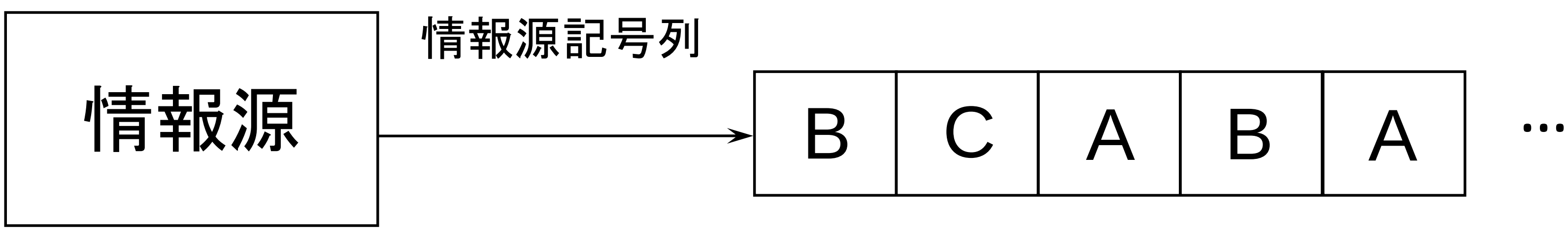
# 本科目の構成


情報源
information source
情報源符号器
source encoder
通信路符号器
channel encoder
通信路
channel
通信路復号器
channel decoder
情報源復号器
source decoder
受信者
destination
通信路符号
（例：FAX, デジタル符号）
情報源符号
（例：文字）
通報
（例：アイデア）
雑音源
noise source

# 情報源のモデル化

情報源：確率モデルを用いてモデル化する.

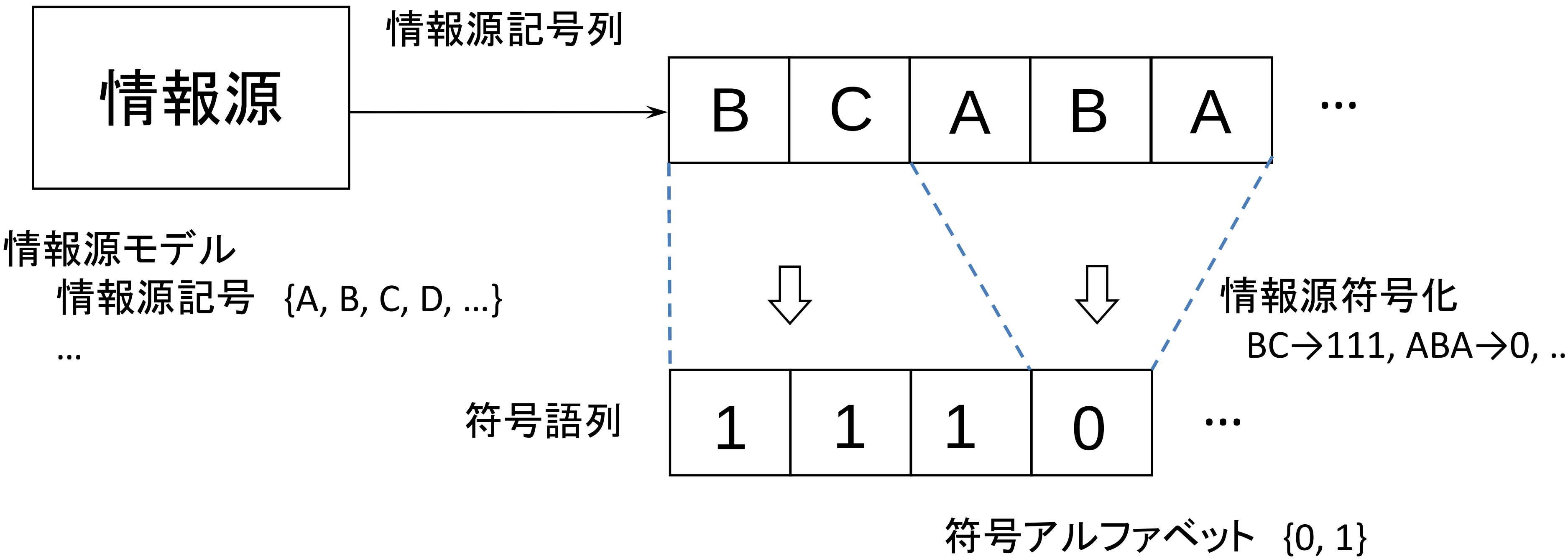
(例)　人，生命，…




情報源モデル
　情報源記号　{A, B, C, D, …}
　情報源記号の生成を説明する確率モデル

# 情報源符号化

情報源から発せられる情報源記号(列)を符号語列に変換する.



# 平均符号長

情報源

情報源符号化

A→00, B→01, C→10, D→11

情報源モデル

情報源記号 {A, B, C, D}

p(A)=0.25, p(B)=0.25, p(C)=0.25, p(D)=0.25

平均符号長＝2

符号アルファベット {0, 1}

# 平均符号長

情報源

情報源符号化

A→00, B→01, C→10, D→11

情報源モデル

情報源記号 {A, B, C, D}

p(A)=0.6, p(B)=0.2 p(C)=0.1, p(D)=0.1

平均符号長＝2

符号アルファベット {0, 1}

# 平均符号長

情報源

情報源符号化

A→0, B→01, C→10, D→11

情報源モデル

情報源記号　{A, B, C, D}

p(A)=0.6, p(B)=0.2 p(C)=0.1, p(D)=0.1

平均符号長＝1×0.6+2×0.2+2×0.1+2×0.1＝1.4

符号アルファベット　{0, 1}

# 平均符号長さえ短ければいいのか？

情報源

情報源符号化

A→0, B→01, C→10, D→11

情報源モデル

情報源記号 {A, B, C, D}

p(A)=0.6, p(B)=0.2 p(C)=0.1, p(D)=0.1

ADA? BC?

0110

平均符号長＝1×0.6+2×0.2+2×0.1+2×0.1＝1.4

符号アルファベット {0, 1}

# 符号化に課せられる条件

- 一意復号可能性

- 瞬時性

# 一意復号可能な符号

情報源

情報源符号化

A→0, B→01, C→011, D→111

情報源モデル

情報源記号　{A, B, C, D}

p(A)=0.6, p(B)=0.2 p(C)=0.1, p(D)=0.1

平均符号長＝1×0.6+2×0.2+3×0.1+3×0.1＝1.6

符号アルファベット　{0, 1}

# 瞬時符号

情報源

情報源モデル
　情報源記号　{A, B, C, D}
　p(A)=0.6, p(B)=0.2 p(C)=0.1, p(D)=0.1

一意復号可能だが非瞬時
　A→0, B→01, C→011, D→111

瞬時符号
　A→0, B→10, C→110, D→111

平均符号長＝1×0.6+2×0.2+3×0.1+3×0.1＝1.6

符号アルファベット　{0, 1}

# ここまでのまとめ

- Shannon-Fanoモデル
- いろいろな情報源
- 情報源のモデル化
- 情報源符号化
- 平均符号長
- 情報源符号の持つべき性質 ― 一意復号可能性，瞬時性

# 残った疑問

- 平均符号長はどこまで短くできるのか？
- 一意復号可能性の判定法？
- 瞬時符号の判定法？
- 平均符号長最短の符号(コンパクト符号)構成法？

…

# 符号の分類

- 符号語の長さによる分類
  - 等長符号
  - 非等長符号
- 復号の可能性による分類
  - 可逆符号
    - 瞬時符号
    - 非瞬時符号
  - 非可逆符号

# 情報源符号化のタイプ

（例） 各情報源記号ごとに符号語を割り当てる場合

| 情報源記号 | 符号 | | | | | | |
|---|---|---|---|---|---|---|---|
| | C1 | C2 | C3 | C4 | C5 | C6 | C7 |
| A | 000 | 0 | 00 | 0 | 0 | 00 | 00 |
| B | 001 | 10 | 01 | 10 | 01 | 10 | 01 |
| C | 010 | 110 | 10 | 110 | 011 | 01 | 10 |
| D | 011 | 1110 | 110 | 1110 | 0111 | 011 | 111 |
| E | 100 | 11110 | 1110 | 11110 | 01111 | 0111 | 1110 |
| F | 101 | 11111 | 1111 | 111111 | 11111 | 1111 | 1111 |
| 等長／非等長 | 等長 | 非等長 | 非等長 | 非等長 | 非等長 | 非等長 | 非等長 |
| 瞬時符号 | ○ | ○ | ○ | ○ | × | × | × |
| 一意復号可能 | ○ | ○ | ○ | ○ | ○ | ○ | × |

# 瞬時性の判定

情報源

情報源符号化

符号アルファベット　{0, 1}

符号1：一意復号可能，非瞬時

A→0, B→01, C→011, D→111

符号2：瞬時

A→0, B→10, C→110, D→111

情報源モデル

情報源記号　{A, B, C, D}

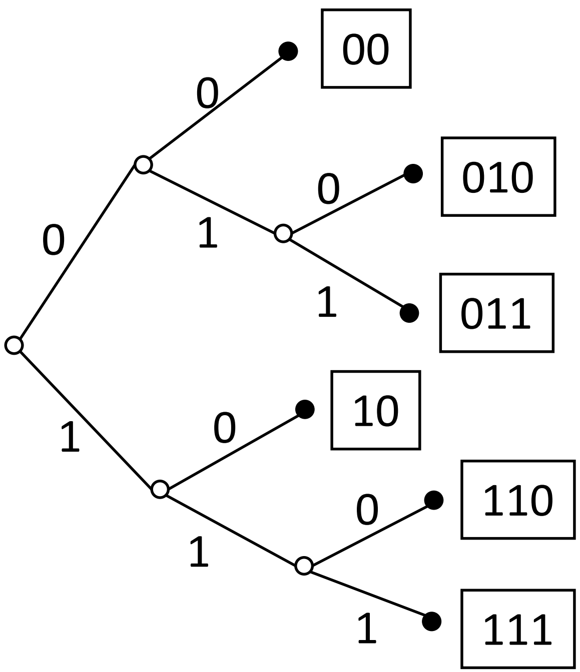
p(A)=0.6, p(B)=0.2 p(C)=0.1, p(D)=0.1

平均符号長＝1×0.6+2×0.2+3×0.1+3×0.1＝1.6

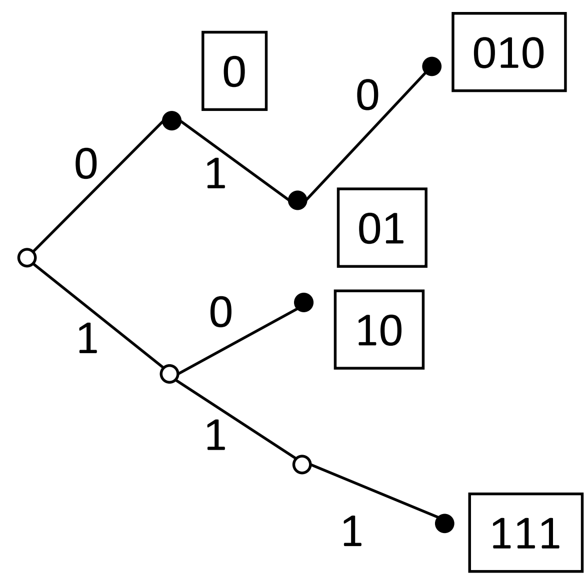
# 瞬時性の判定

符号の木：符号化で使われる符号語の集合を構造的に表現したもの

$C_1=\{00,010,011,10,110,111\}$




$C_2=\{0,01,010,10,111\}$



# 瞬時性の判定

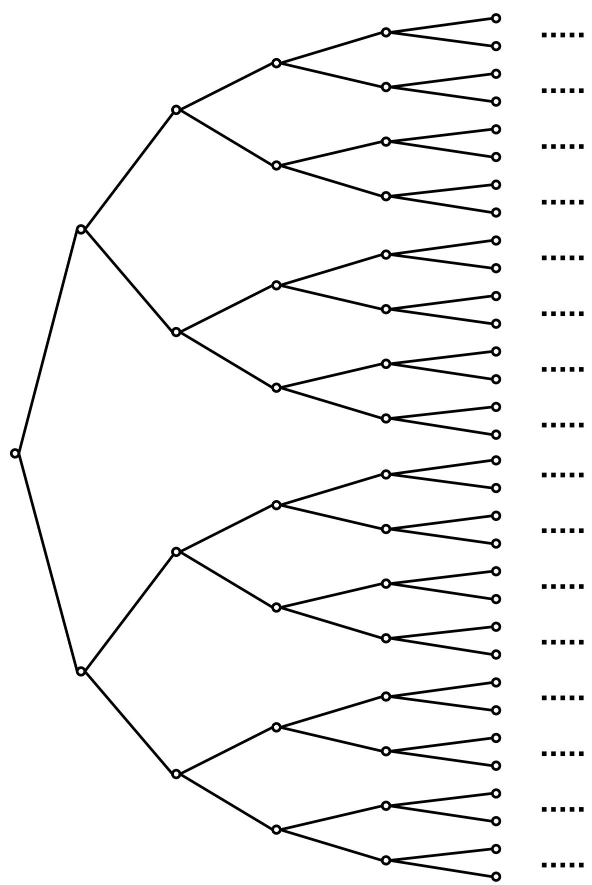
- 符号の木で$w_i$が$w_j$の接頭
  ⇔
  $w_i$が$w_j$の上流にある

- 符号の木$t$が接頭条件を満足する
  ⇔
  $t$のどの符号語も他の符号語の接頭になっていない.

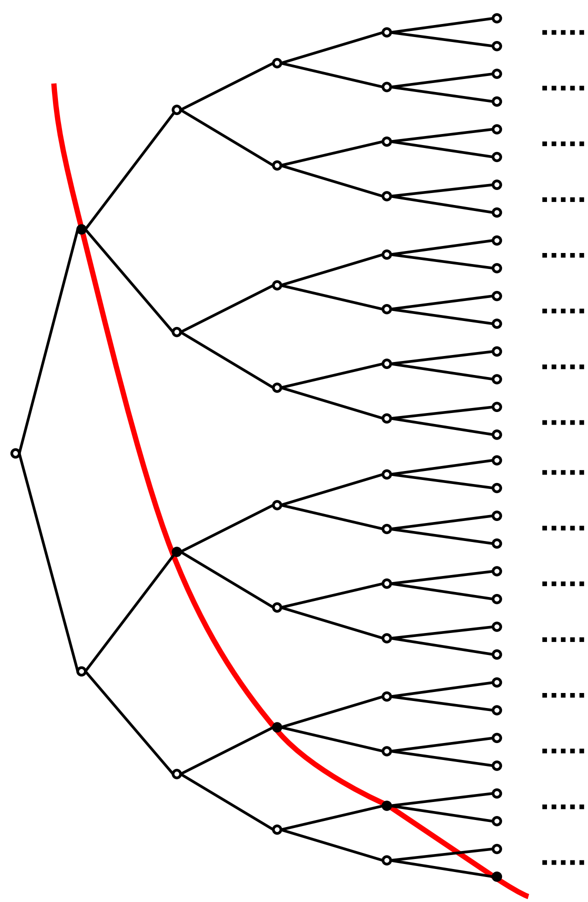
- 符号化$C$が瞬時符号である
  ⇔
  $C$に対する符号の木が接頭条件を満足している.

# 瞬時性の判定

例：　情報源記号$\{A_1, A_2, A_3, A_4, A_5\}$に対する2元瞬時符号



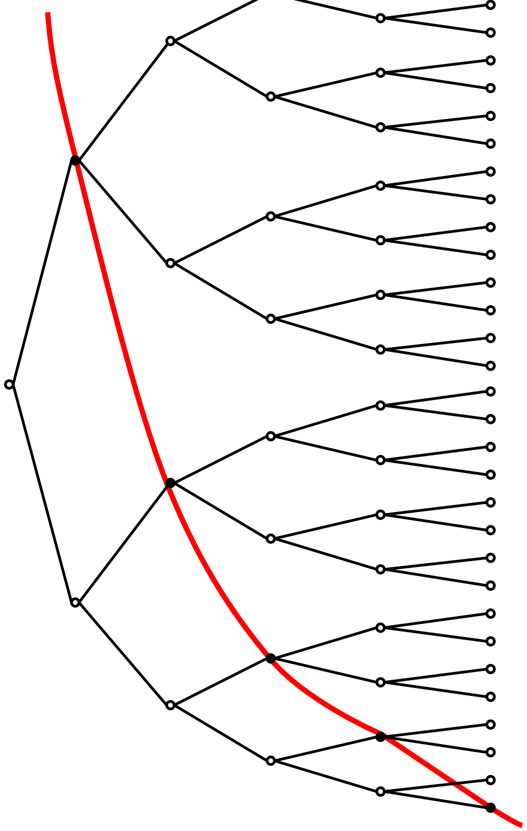

接頭条件を満たすように枝刈りする



# 瞬時性の判定

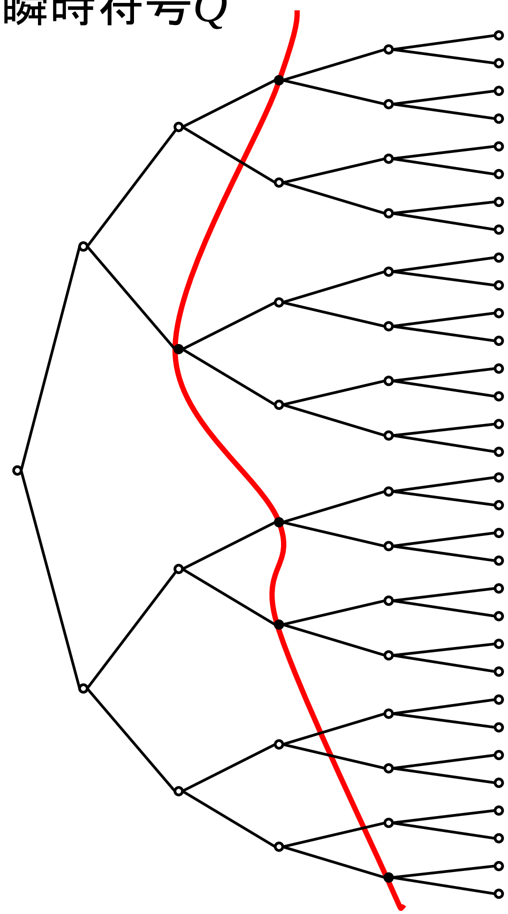
例： 情報源記号$\{A_1, A_2, A_3, A_4, A_5\}$に対する2元瞬時符号

瞬時符号$P$



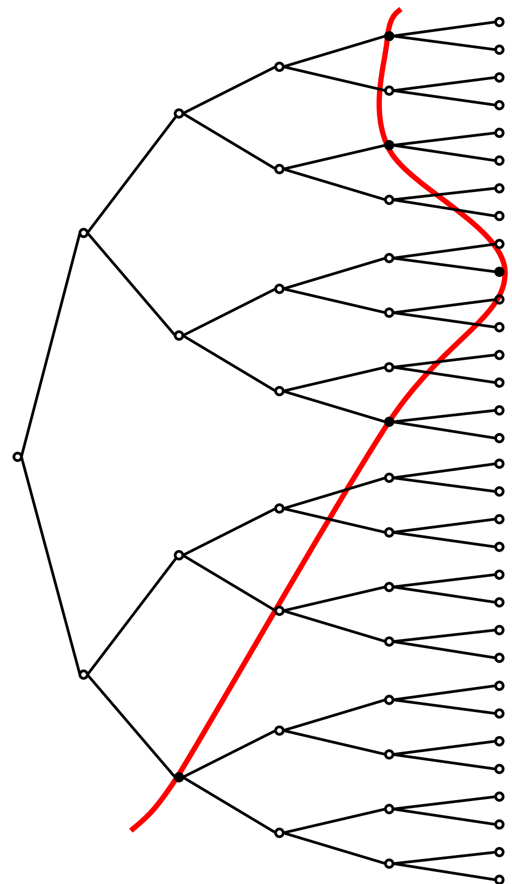
符号長バッグ：{1, 2, 3, 4, 5}

瞬時符号$Q$



符号長バッグ：{3, 2, 3, 3, 4}

瞬時符号$R$



符号長バッグ：{4, 4, 5, 4, 2}

…

瞬時符号構成可能な符号語バッグの条件？

# 瞬時符号構成可能性

クラフトの不等式

長さが$l_1, l_2, \cdots, l_M$の$M$個の符号語$c_1, c_2, \cdots, c_M$をもつ$q$元瞬時符号を構成できる.

$$\Leftrightarrow \quad q^{-l_1} + q^{-l_2} + \cdots + q^{-l_M} \leq 1$$

例題 : 長さ1,2,3,3の4個の符号語をもつ2元瞬時符号は構成可能か？

解答 : Yes!

$$2^{-1} + 2^{-2} + 2^{-3} + 2^{-3} = \frac{4+2+1+1}{8} = 1 \leq 1$$

# 瞬時符号構成可能性

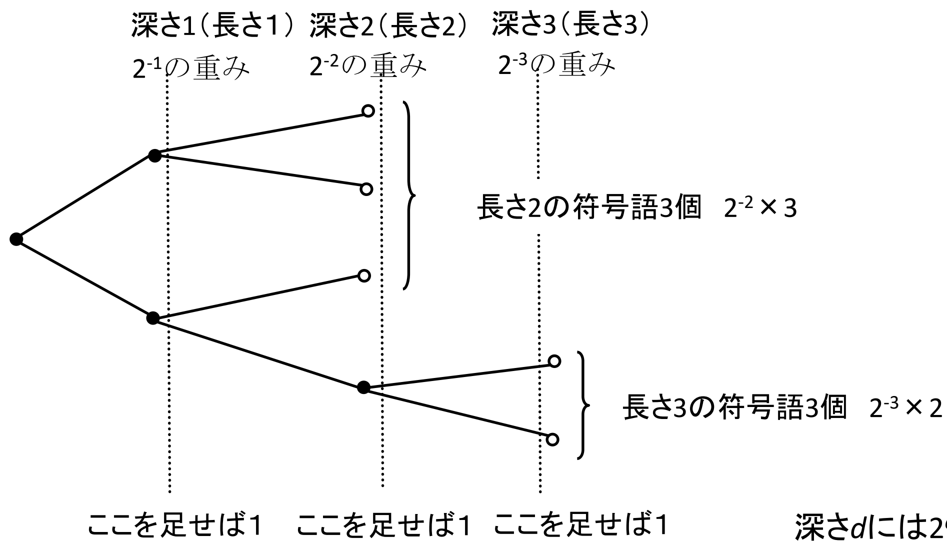
## クラフトの不等式

長さが$l_1, l_2, \cdots, l_M$の$M$個の符号語$c_1, c_2, \cdots, c_M$をもつ$q$元瞬時符号を構成できる.

$$\Leftrightarrow \quad q^{-l_1} + q^{-l_2} + \cdots + q^{-l_M} \leq 1$$

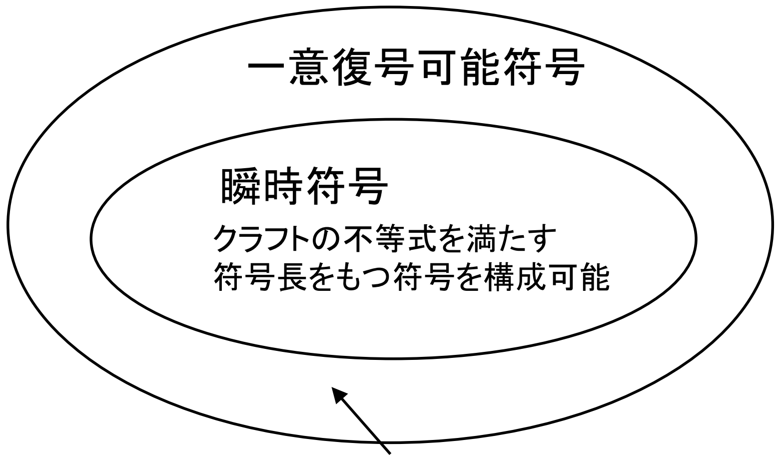
## 直観的証明：リソースという考え方を使う

2元符号の場合




深さ$d$には$2^d$個の符号語を割り当てられる.

# 一意復号可能な符号の構成可能性




クラフトの不等式を満たさない符号長をもつ
一意復元可能な符号を構成可能？ ⇨ NO!

マクミランの不等式

$l_1, l_2, \cdots, l_M$のM個の符号語$c_1, c_2, \cdots, c_M$をもつ$q$元**一意復号可能**な符号を構成できる.

$$\Leftrightarrow \quad q^{-l_1} + q^{-l_2} + \cdots + q^{-l_M} \leq 1$$

# マクミランの不等式の証明

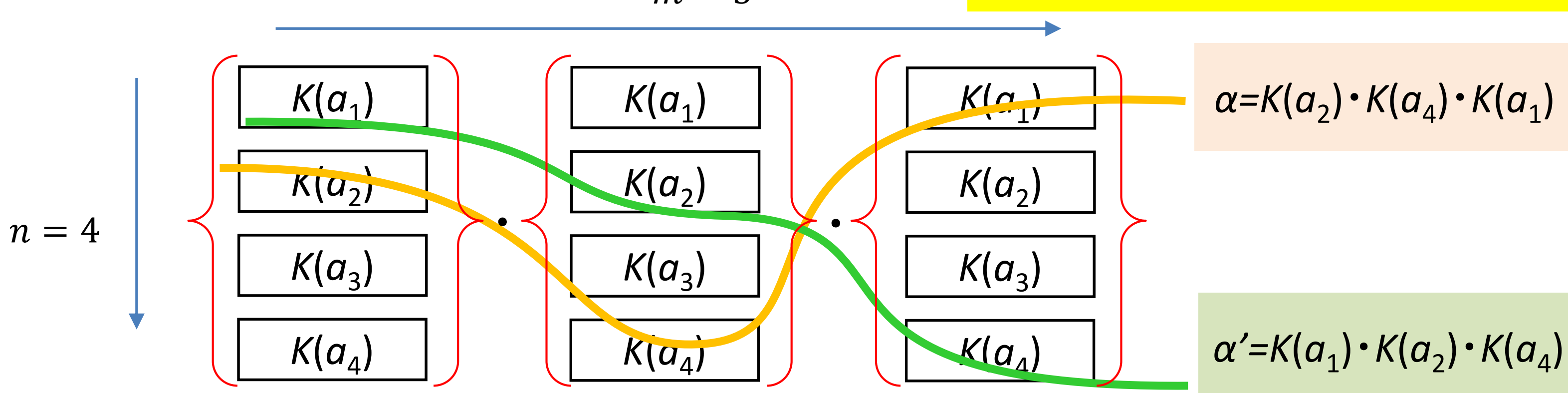
- $a_1, a_2, \cdots, a_n$を情報源記号，$K(a_1), K(a_2), \cdots, K(a_n)$を符号語，$d_1, d_2, \cdots, d_n$を各符号語の長さとする.
- $c = 2^{-d_1} + 2^{-d_2} + \cdots + 2^{-d_n}$という量に着目し，$c > 1$ならば，一意に復号可能でないことを示す
.
- $\mathrm{c}^m$という量について考えてみよう.
$\mathrm{c}^m$は $\{c_1, c_2, \cdots, c_n\}$ のなかの符号語を$m$個つないでできるすべての系列$\alpha$について，
$2^{-|\alpha|}$ を計算し，加えたものに等しい.

（例） $n = 4, m = 3$

$$
\begin{aligned}
c^3 &= \left(2^{-d_1} + 2^{-d_2} + 2^{-d_3} + 2^{-d_4}\right) \times \left(2^{-d_1} + 2^{-d_2} + 2^{-d_3} + 2^{-d_4}\right) \times \left(2^{-d_1} + 2^{-d_2} + 2^{-d_3} + 2^{-d_4}\right) \\
&= 2^{-(d_1+d_1+d_1)} + 2^{-(d_1+d_1+d_2)} + 2^{-(d_1+d_1+d_3)} + \cdots + 2^{-(d_4+d_4+d_3)} + 2^{-(d_4+d_4+d_4)} \\
&= \sum_{3\cdot \min_{1\le i\le 4} d_i \le l \le 3\cdot \max_{1\le i\le 4} d_i} N_l \cdot 2^{-l}
\end{aligned}
$$

一般には，$\mathrm{c}^m = \displaystyle\sum_{m\cdot \min_{1\le i\le n} d_i \le l \le m\cdot \max_{1\le i\le m} d_i} N_l \cdot 2^{-d_l}$

$m = 3$



# マクミランの不等式の証明

- 所与の符号が一意復号可能であるためには，長さ$l$になる系列の個数$N_l$は，長さ$l$の可能な2元符号語の個数$2^l$を超えてはならない．つまり，$N_l \leq 2^l$でなければならない．

- 従って，$c^m = \sum_{m\cdot \min_{1\leq i\leq n} d_i \leq l \leq m\cdot \max_{1\leq i\leq n} d_i} N_l \cdot 2^{-l} \leq \sum_{m\cdot \min_{1\leq i\leq n} d_i \leq l \leq m\cdot \max_{1\leq i\leq n} d_i} 2^l \cdot 2^{-l} \leq m \cdot \max_{1\leq i\leq n} d_i$

- 前項で得られた$N_l \leq 2^l$は，（固定された$l_1, l_2, \cdots, l_n$に対して）任意の$m$について成立しなければならない．

- ここで

$$\frac{c^m}{m} = \frac{\left((c-1)+1\right)^m}{m} = \frac{1 + m(c-1) + \frac{m(m-1)(c-1)^2}{2} + \cdots}{m} > \frac{(m-1)(c-1)^2}{2}$$

であるので，$c > 1$であればこの条件を満足できない．

- 従って，$c \leq 1$でなければならない．

# まとめ

- Shannon-Fanoモデル
- いろいろな情報源
- 情報源のモデル化
- 情報源符号化
- 平均符号長
- 情報源符号の持つべき性質 ― 一意復号可能性，瞬時性
- 瞬時性の判定 ― 接頭条件
- 瞬時符号構成可能性 ― クラフトの不等式
- 一意復号可能な符号の構成可能性 ― マクミランの不等式
- 残された課題：平均符号長はどこまで短くできるのか？一意復号可能性の判定法？平均符号長最短の符号の構成法？