

情報のデジタル化

広島大学 AI・データイノベーション教育研究センター

村上 祐子

目標

情報をデジタル表現する過程について説明できるようになる

この教材で紹介すること

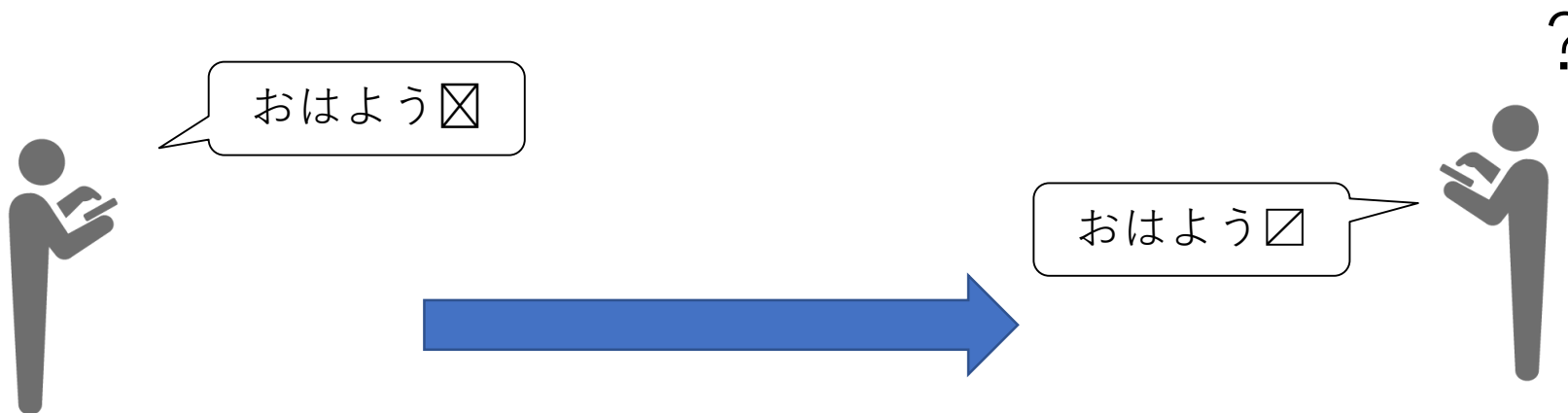
- デジタル表現とは何か
- 文字をデジタル表現にする過程

キーワード

情報の単位、2進法、文字コード、文字符号化集合、文字符号化方式

こんなことはありませんか？

SNSで文字や絵文字が正しく表示されない



コンピュータやスマートフォンなどは文字をどのように認識しているのか

アナログとデジタル

アナログ

- 連続した量で表現される
 - 人間が聞く音
 - 人間が見る物（文字）
- 多くの情報を表現できる
 - 筆跡の違い
 - 音楽のニュアンスの差
- コピーするためには物理的な移動が必要
 - 授業ノートを手書きで移す

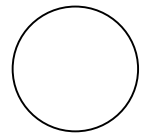
デジタル

- 離散的な量で表現される
 - コンピュータの処理・表現は0と1の2通りのみで行われる
- 物理的な移動を伴わなくてもコピーできる
 - 授業資料を電子メールでやり取り
- コピーの精度が高い
 - チラシを数百枚印刷しても同じ絵柄

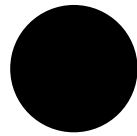
情報の単位

情報の大きさを表す単位

bit(ビット) 2通りの情報を表現できる



か



か

オン

か

オフ

か

白

か

黒

か

0

か

1

か

コンピュータは2進法のデジタルデータを処理している

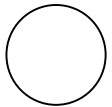
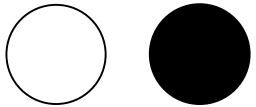

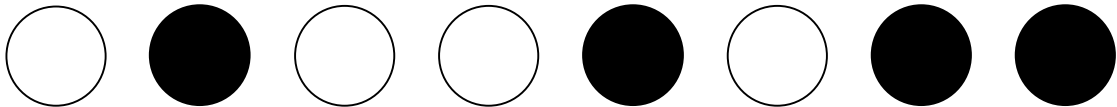


命令 : 01110101001100110101010...



応答 : 101010101010101010101...

情報の単位

		2進法	表現方法
1bit		0	2通り
2bit		01	$2 \times 2 = 4$ 通り
3bit		011	$2 \times 2 \times 2 = 8$ 通り
8bit		01001011	$2 \times 2 \times 2 \times \dots = 256$ 通り

1byte(バイト)とする

デジタルの表現

- 10進法
通貨などで一般生活になじみ深い
0~9の10通りで量を表す
- 2進法
0と1の2通りで量を表す
- 16進法
0~9 + A~Fの16通りで量を表す

10進法	2進法	16進法
0	0	0
1	1	1
2	10	2
3	11	3
4	100	4
5	101	5
6	110	6
7	111	7
8	1000	8
9	1001	9
10	1010	A
11	1011	B
12	1100	C
13	1101	D
14	1110	E
15	1111	F
16	10000	10

情報の単位

より膨大な情報量を表現する単位

1KB(キロバイト) = 1,024byte

1MB(メガバイト) = 1,024KB = 1,024 × 1,024 byte

1GB(ギガバイト) = 1,024MB = 1,024 × 1,024 × 1,024 byte

1TB(テラバイト) = 1,024GB = 1,024 × 1,024 × 1,024 × 1,024 byte



1kg=1,000gのk(キロ)などとは単位の変換が違うことに注意

例題

1. 32bitと5byteではどちらの情報量が大きい？
2. コンピュータの処理は2進法で行われますが、その処理は16進法で表現されることが多いです。日ごろの生活で慣れ親しんでいる10進法ではなく、16進法で表されるのはなぜでしょうか？



日常的に10進法を使うなら10進法で表現すればいいのに

解説①

1. 32bitと5byteではどちらの情報量が大きい？

情報の単位を揃えます。

1byte=8bitなので、

5byte=5×8bit=40bit

これは32bitよりも大きいので、

32bitと5byteでは5byteのほうが大きい

解説②

2. コンピュータの処理は2進法で行われますが、その処理は16進法で表現されることが多いです。日ごろの生活で慣れ親しんでいる10進法ではなく、16進法で表されるのはなぜでしょうか？

- 2進法の4桁を16進法1桁の数に置き換えできて便利だから
- 2進法を10進法で表現しようとするとう桁上がりになずれが生じる

0000 0010 1010 1111
↓ ↓ ↓ ↓
0 2 A F

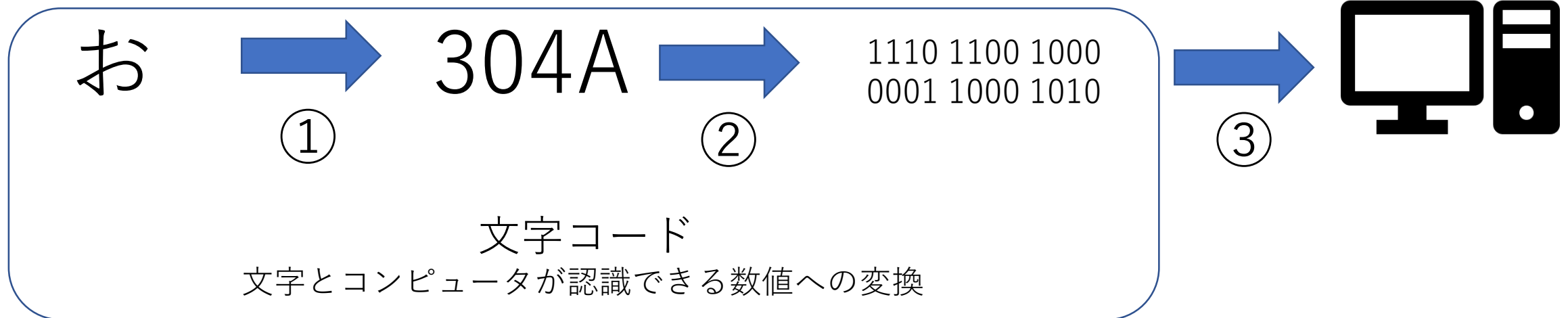
10進法	2進法
9	1001
10	1010
11	1011

「2進法4桁の数字を10進法で**1桁**か**2桁**で表す」
というルールは
はっきりしていなくてやりにくい

文字がコンピュータに認識されるまで

SNSやメールに入力した文字は

- ① 文字に割り当てた番号へ変換
- ② 手順①で変換した番号をコンピュータが理解可能な数字(符号)へ変換
- ③ 手順②で変換された符号をコンピュータが理解する



文字符号化集合

世界中のあらゆる言語、記号などを数字に割り当てたもの

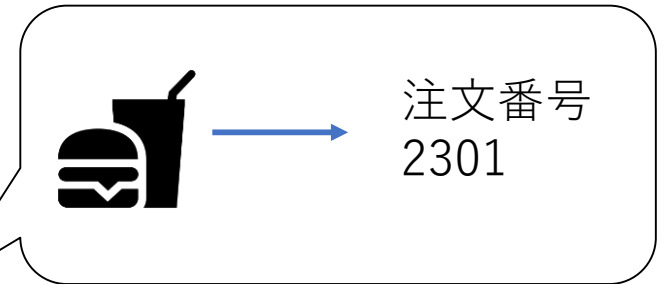
例：Unicodeで定義されている文字符号化集合

文字	文字符号化集合
お	304A
M	004D
°C	2103

16進法で表現



カタログのようなもの



Wikipedia Unicode表, <https://ja.wikipedia.org/wiki/Category:Unicode%E8%A1%A8>

文字符号化方式

符号化文字集合で文字に対応付けられた番号をコンピュータが理解できるデータ列に変換する方法

例：UnicodeをUTF-8に変換する

- ① 下の表からUnicodeの符号範囲に対するUTF-8のビット列表現を調べる

Unicodeの符号範囲	UTF-8のビット列 (2進法)
0000-007F	0xxx xxxx
0080-07FF	110x xxxx 10xx xxxx
0800-FFFF	1110 xxxx 10xx xxxx 10xx xxxx
10000-10FFFF	1111 0xxx 10xx xxxx 10xx xxxx 10xx xxxx

- ② Unicodeの符号(16進法)を2進法に変換する
- ③ ①の表のxxxに②で変換した数値をあてはめる

「お」の文字がコンピュータに認識されるまで

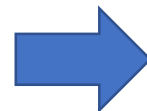
「お」の文字はUnicodeで「304A」と表される

- ① 下の表からUnicodeの符号範囲に対するUTF-8のビット列表現は
1110 **xxxx** 10**xx** **xxxx** 10**xx** **xxxx**

Unicodeの符号範囲	UTF-8のビット列 (2進法)
0000-007F	0 xxx xxxx
0080-07FF	110 x xxxx 10 xx xxxx
0800-FFFF	1110 xxxx 10 xx xxxx 10 xx xxxx
10000-10FFFF	1111 0 xxx 10 xx xxxx 10 xx xxxx 10 xx xxxx

- ② 304A(16進法)は2進法に変換すると0011 0000 0100 1010.
③ ①の表のxxxに②で変換した数値をあてはめると

1110 **0011** 10**00** **0001** 10**00** **1010**



「お」の情報量は3byte

問題

文字コードの規格Unicodeを用いて、「おはよう」という文字の情報量を考えます。

1. 「は」、「よ」、「う」の文字符号化集合での表現を調べてみましょう。
2. 問題1.で調べた文字符号化集合をUTF-8のビット列表現に変換してみましょう。
3. 「おはよう」という文字列の情報量は全部でいくらになるのでしょうか

文字	文字符号化集合	UTF-8のビット列 (2進法)
お	304A	1110 1100 1000 0001 1000 1010
は		
よ		
う		