

情報 information

情報システムでは「データ」レベルの処理を示すことが多い。

- 価値判断や意思決定につながる状態にまである意味付けがなされた表現形態。

データを複合したもの、データを特定の形式で表現したもの

データ data

- それ単独ではある状態を示すが、判断や意思決定を表現するものではない。

身長、体重、血圧、血糖値など個別の数値

知識 knowledge

- 知性を伴う分析・評価によって、人間の判断の結果、生成されたもの。

医療情報

広義：患者自身の生体情報、医学知識、医療機関の情報など医療における情報の総称

狭義：患者から得られる情報、診療を行うための情報

_____：国際疾病分類コード

医療情報の特性



多面性・多層性



連続性



機密性

表現形態が多種類

- コード情報：薬品名、手術術式、病名、など
- テキスト情報：患者の主、読影レポートなど
- 数値情報：身長、体重、血液検査情報など
- 音情報：聴診器による心音、肺音など
- 波形情報：心電図、脳波、脈波など
- 画像情報：
 - ✓ 皮膚科写真などのカラー静止画像
 - ✓ X線検査などによる濃淡静止画像
 - ✓ 超音波エコー、冠動脈造影などの動画像
 - ✓ CT・MRIなどによる3次元画像
 - etc.

➤ 多面性・多層性

医療情報の内容の広がり深さの階層は非常に幅が広く多彩である

- 理学的情報：望診、視診、聴診、打診、触診を行い、経験に基づいて半定量的表現する
- 生理的信息：身長、体重などの形態的な計測値、血圧、脈拍数、心電図、脳波、など
- 臓器機能情報：肝機能、腎機能、脂質代謝、糖代謝などを生化学計測により検出
- 免疫的信息：細菌、ウィルス感染の有無などを検出
- 症候論的信息：患者の主訴、臨床症状、病状経過、生活史的背景、環境など



-
- 患者を専門職が観察・観測して問題を抽出、データを整理・統合処理を加え分析して、結果を判定し状態を判断（診断）する。診断結果に基づいて処置治療の必要性を決定、実行計画を立て実践（治療）する。成果を再び観察・観測して評価につなげる。このスパイラルが連続的に進行する。
 - これらの過程で下される診断や治療計画は、意思決定や思考を伴う高次の情報である。



-
- 誰もが自身の身体や病気に関わる情報を公開したくない。
 - 医療情報の中でも特に疾患情報は最も深遠な個人情報であるため、プライバシー保護、守秘性が確実に守られなければならない。
 - 他の情報に比して個人情報の保護、そのためのセキュリティ確保についての要求水準が高い。

医療情報の利用形態

・受益者(患者)に直接還元する_____

収集された種々の医療情報を、本来の目的(患者の直接利益)のために利用すること。

例えば、収集した検査データなどの処理により診断・治療を行う際、投薬結果に基づく検査データの変化から、患者の治療度を判定すること、など。

・公益のための_____

収集された種々の医療情報を、元来の目的意外のために利用すること。

例えば、血圧値や高コレステロール血症を示す患者数や割合などの統計データを算出する場合など。

→ 保健所、司法、警察への提供

病院経営管理のための利用、行政的利用、医学研究への利用、医学教育への利用、など。

↳ 診療報酬請求、病院の医療の質や効率の評価など

当該医療情報の発生源である患者などの了承が必要であることが多い。

医療情報の法的保存期間

・診療録

診療した医師は遅滞なく診療に関する事項を診療録に記録しなければならない。診療録は、5年間保存しなければならないとされている。その保存義務者は、勤務医に関しては病院または診療所の管理者であり、その他の場合はその医師である（医師法第24条、歯科医師法第23条）。

・処方せん

薬局で調剤済みとなった処方せんは、3年間保存しなければならないとされている（薬剤師法第27条）。

・その他の医療記録

「診療に関する諸記録は、過去2年間の病院日誌、各科診療日誌、処方せん、手術記録、看護記録、検査所見記録、エックス線写真、入院患者及び外来患者の数を明らかにする帳簿並びに入院診療計画書とする」（医療法施行規則第20条11号）。ただし、保健医療機関及び保健医療担当規則第9条では3年間の保存を義務付けている。

電子保存について

平成11年(1999年)年4月に、厚生労働省より「診療録などの電子媒体による保存について」が通達され、
_____ (真正性、見読性、保存性)が担保され、
プライバシー保護などが確保できれば、電子的な手段での保存が認められるようになった。

_____ : 間違いなく本物であること。診療において用いられた情報そのもの、診療の際に記録された情報そのものが保存されていること。

_____ : 必要に応じて表示および印刷ができること。セキュリティ用語の可用性(availability=システムが継続して稼働できる能力のこと)と同等。

_____ : 決められた期間、滞りなく保存できること。
民法第426条では民事訴訟までの期間を20年としているため、医療情報は最低5年間、できれば20年間の保管が望ましい。

コンピュータの5大機能

- _____: 装置 (CPU) → Control
- _____: 装置 (CPU) → Processing
- _____: 装置 (メモリー, HDD, CD-ROM等) → Memory
- _____: 装置 (キーボード, マウス等) → Input
- _____: 装置 (モニター, プリンター等) → Output

パソコンなどでは、制御装置と演算装置を一体化した中央処理装置 (CPU : Central Processing Unit) が使われている。

クライアント・サーバ型

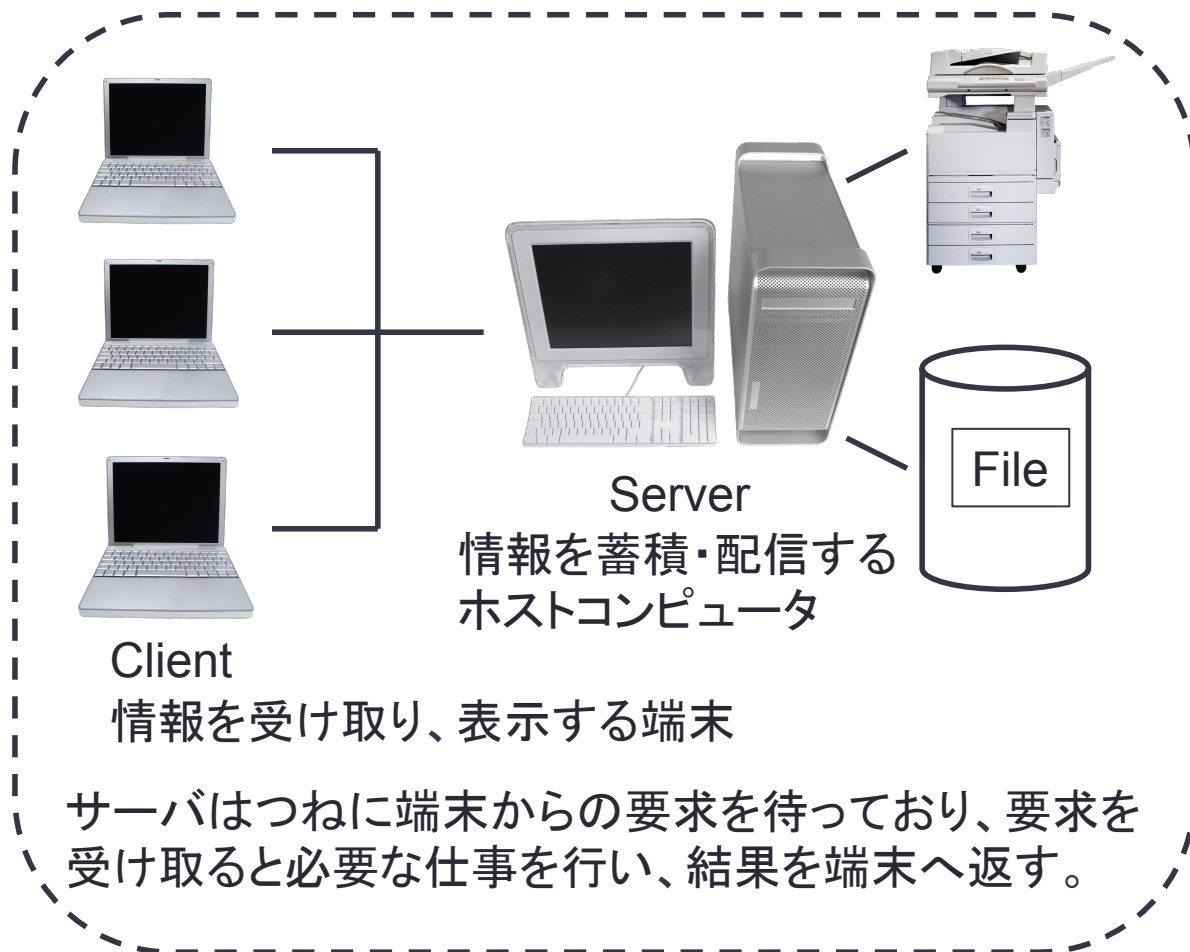
ネットワーク上のコンピュータのうち、サービスを受ける側()
とサービスを提供する側()とで明確に役割分担をしている運用形態。

・ _____方式

データのみをネットワークで送受信し、端末内でデータ処理を行って情報を表示する形態。

・ _____方式

全部の処理サーバ内で行い、結果の画面までサーバで作成し、その画面コピーを端末側に転送し表示させる形態。



サーバはつねに端末からの要求を待っており、要求を受け取ると必要な仕事を行い、結果を端末へ返す。

Local(局地的) Area(区域) Network(通信網)の略語であり、同一建物内や構内、敷地内など物理的に近い距離の限定された範囲をカバーするネットワークを指す。現在は、LANの接続形式としてEthernet(イーサネット)が標準的に用いられている。院内LAN、学内LAN、有線／無線LAN、など。

Wide(広い) Area(区域) Network(通信網)の略語であり、物理的に遠い距離にある拠点同士を結んだネットワークである。離れた場所に存在するLANを通信事業者が提供している公衆回線網や専用線などのWAN回線サービスを用いて相互に接続する。広域情報通信網とも呼ばれる。

Hardware & Software

	コンピュータ	人体なら
ハードウェア	パソコン本体, キーボード マウス, プリンター等	脳, 内臓, 骨
ソフトウェア	OS, アプリケーション, ミドルウェア, ファームウェア	血液の循環 神経の伝達



OS(オーエス)

O S

コンピューターのハードウェア、ソフトウェアを有効に利用するためにコンピューター全体の動きをまとめる基本的で重要な役割を持つソフトウェア

編集(E)	検索(S)	ヘルプ(H)
元に戻す(U)		Ctrl+Z
切り取り(T)		Ctrl+X
コピー(C)		Ctrl+C
貼り付け(P)		Ctrl+V
削除(L)		Del

ハードウェアとソフトウェアの制御・管理、ユーザビリティのためのユーザインタフェースの提供などを行う。

OSの種類	動作ハードウェア
_____	大型・専用コンピュータ
Windows	PC/AT互換機
Mac OS X	Apple社コンピュータ
_____	PC/AT互換機

アプリケーションソフトウェア

コンピュータを目的別に動かすためのソフトウェア

- ✓ 文書処理ソフトウェア (word processor, word processing software)
 - ✓ 表計算ソフトウェア (spreadsheet program)
 - ✓ プレゼンテーションソフトウェア (presentation software)
 - ✓ 画像処理プログラム (image processing program)
 - ✓ メーラー (mail software, mailer)
 - ✓ ウェブブラウザ (Web browser)
- など

医療情報処理システムでは、次のようなソフトウェアも利用される

- ・OMR (optical mark reader) → マークシートを読み取り、割り当てられている文字や数値として認識させるソフトウェア
- ・OCR (optical character reader) → 文字を含むスキャナ画像から、文字情報を自動で認識するソフトウェア
- ・バーコードリーダー (bar-code reader)
- ・QRコードリーダー (2次元バーコード)



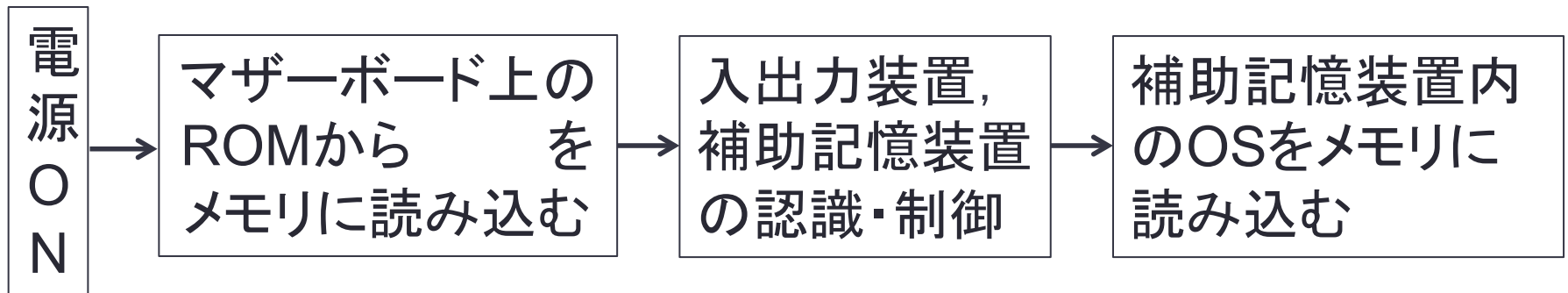
バーコード



QRコード

コンピュータのメモリは電源を入れた最初の状態では空であり、コンピュータを動かすには通常ハードディスクに保存されているオペレーティングシステムをメモリに読み込んでくる必要がある。この処理にはローダ(loader, boot loader)と呼ばれる読み込みのためのプログラムが必要である。ところが、ローダもプログラムであるから動くにはメモリに読み込む必要がある。

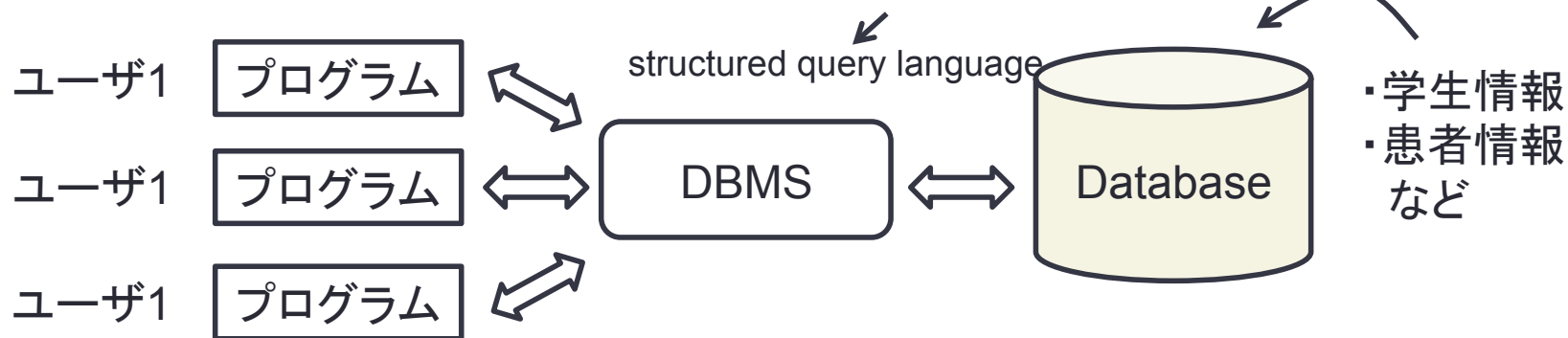
そこで、電源が消えても中身の消えないROMに、ローダを探してメモリに読み込むためのプログラム(IPL: Initial Program Loader)を記憶しておく方式がとられている。また、コンピュータの起動時に、通常、ディスプレイ、キーボード、マウス、ディスク装置が使える必要があるため、必要最小限の入出力制御プログラムもROMに記憶させてある。このようなROMに記録したソフトウェアのことを、ハードウェアとソフトウェアの中間に位置することからファームウェアと呼ぶ。IBM PC/AT互換機では、BIOS(Basic Input Output System)、MacintoshではOpen Firmwareがこれに相当する。



ミドルウェア middleware

オペレーティングシステムに組み入れるほどには基本的ではないが、ある分野においては複数の応用目的に共通に必要な機能がある。このようなレベルの共通機能を実装したソフトウェアをミドルウェア (middleware) と呼ぶ。

アプリケーションソフトウェアの多くは大量の情報を蓄積・管理し、検索等を行う。そのような共通の情報処理の要求に応じて、データベースの作成、維持、運用を効率的に行うために強力なデータ管理機能を提供するのがデータベース管理システム (DBMS: Database Management System) である。これはミドルウェアであり、専用言語としてSQLが用いられる。



データベース(database: DB)

データベースとは、一言でいえば、「たくさんのデータを蓄積して簡単に利用するための仕組み」である。

- 特定のテーマに沿ったデータを集めた管理し、容易に検索・抽出などの再利用をできるようにしたもの。
- 複数のアプリケーションやユーザによってデータが共有できるように整理したデータの集合体のこと。

医療情報は、「_____」、「_____」、「処方せん」、「予約」など、ほとんどデータベースに登録されている。

DBを構成するサーバや利用者端末の制御、また、データの収集、抽出、保存などの運用管理にはデータベース管理システム(database management system: _____)が用いられる。データベース言語として_____ (structured query language)が有名。

インタフェース(interface)

インタフェースとは、境界を意味する言葉である。例えば、人間がドアを開けるときの、ドアのノブをつかむ。その接したときにできる状態をインタフェースと言う。

- 二つのものの間に立って、情報のやり取りを仲介するもの。また、その規格。
 - コンピュータ本体と各種周辺装置やコンピュータ同士を接続し、電気信号の大きさを調整したり、データの形式を変換したりして、両者間のデータのやり取りを仲介する回路や装置。また、人間がコンピュータなどの装置を円滑に使用できるようにするための媒介装置、操作手順。
- Hardware interface → USB、Thunderbolt、IEEE1394など
 - Software interface → API、ネットワーク通信プロトコルなど
 - User interface → マウス、キーボード、GUI、CUI など

GUI:

コンピュータグラフィックスとポインティングデバイスを用いて、直感的な操作を提供するユーザインタフェース。

コンピュータへの命令をマウスを使用して、ウィンドウ、アイコン、ボタンなどで伝える。



例: cドライブをマウスクリック

CUI:

コンピュータへの命令を文字で伝える。

キーボードを用いて入力を行い、文字によって出力を行う様式のユーザインタフェース。

```
TRONLOG.TXT 1,000 01-08-27 12:57 TRONLOG.TXT
WINDOWS <DIR> 01-08-27 12:29 WINDOWS
NETLOG.TXT 6,132 01-08-27 12:42 NETLOG.TXT
AUTOEXEC.BAT 173 02-12-20 13:01 AUTOEXEC.BAT
DOSIME.SYS 155 01-08-27 12:31 DOSIME.SYS
PROGRAM*1 <DIR> 01-08-27 12:31 Program Files
MYDOCU*1 <DIR> 01-08-27 12:46 My Documents
WIN98 <DIR> 01-08-27 13:24 win98
KPCMS <DIR> 01-08-27 17:03 KPCMS
MYINST*1 <DIR> 02-10-28 9:36 My Installations
WINDOW*1 <DIR> 01-08-27 17:40 Windows Update Setup Files
SCANDISK.LOG 1,321 03-02-18 17:37 SCANDISK.LOG
DATA1 <DIR> 01-08-28 10:08 data1
TMP <DIR> 03-02-26 18:12 tmp
VCDROM <DIR> 02-04-30 11:32 VCDROM
ASLEEP_A.DAT 9,510 02-05-08 17:10 ASLEEP_A.DAT
ADOBEWEB.LOG 0 02-07-03 10:35 AdobeWeb.log
BRAINCT <DIR> 02-08-22 16:54 brainCT
PHP <DIR> 03-06-27 18:48 PHP
MYSQL <DIR> 03-06-27 19:11 mysql
MCAF.LOG 3,192 03-10-02 11:45 mcaf.log
10 個 140,049 バイトのファイルがあります。
13 ティレクトリ 1,143.09 メガバイトの空きがあります。
C:\>
```

例: `ls -l /c:¥`

そもそも情報とはなに？

「ある事柄に関して を得たり のより所としたりするために不可欠な、何らかの手段で伝達（入手）された種々の事項（の内容）」

（新明解国語辞典 第6版、三省堂より一部抜粋）

コンピュータで取り扱う情報の定義

・Claud Elwood Shannon（情報理論の発案者）

「 の中から選択できるもの」

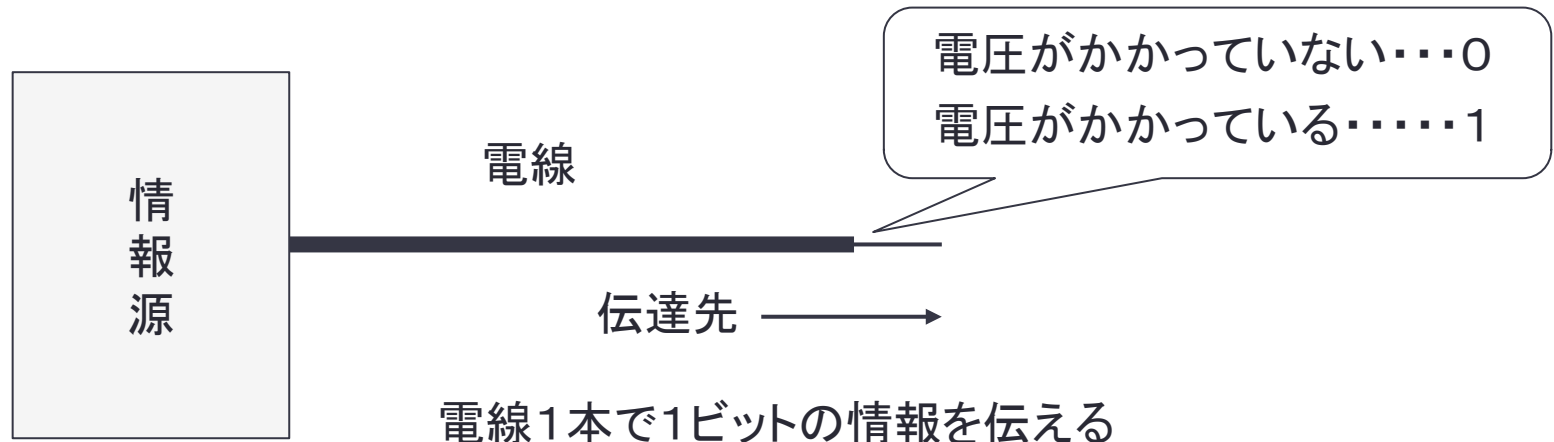
例：T子さんが結婚する（未婚→既婚への変化）

ここで問題です①

Q.「オバケのQ太郎」という漫画には、毎日三食とも必ずラーメンを食べている小池さんというキャラクターが登場します。小池さんが今日何を食べたかは、情報と呼べるでしょうか？

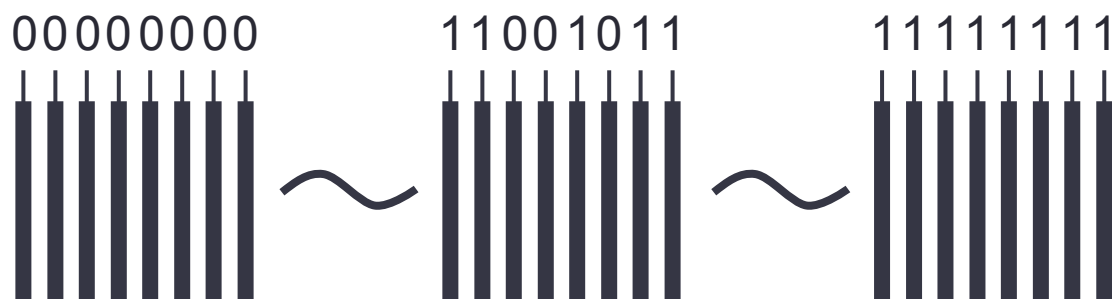
情報の最小単位

天気という情報は、晴れ、曇り、雨、雪の4通りに変化します。道路信号は、青、黄、赤の3通りに変化します。それでは、最も少ない変化は、何通りでしょう。YES/NO、男/女、前/後のような2通りの変化です。すなわち、2通りの変化が情報の最小単位であり、これを「ビット」と呼びます。ビット(bit)はbinary digit(2進数)を略した言葉です。



情報の基本単位

コンピュータは、8本の電線をセットで使った 8 ビットを、情報の基本単位としています。これを「バイト」と呼びます。8 ビット = 1 バイト (byte) です。



最近のコンピュータは
64ビットCPU (central
processing unit: 中央
処理装置) が搭載され
ている → 1800京通り

8本 = 1バイトで表せるパターンは256通り、すなわち256通りの変化を処理できる

1024 byte (B) = 1 KB (キロバイト)

→ 2^{10} byte

1024 KB = 1 MB (メガバイト)

→ 2^{20} byte

1024 MB = 1 GB (ギガバイト)

→ 2^{30} byte

1024 GB = 1 TB (テラバイト)

→ 2^{40} byte

n進数 → 10進数

$a_m a_{m-1} \cdots a_1 a_0$: n進数のある数

$$x_0 = n^0 \times a_0 + n^1 \times a_1 + \cdots + n^{m-1} \times a_{m-1} + n^m \times a_m$$

10進数表記: x_0

例1: 2進数 $(10101011)_2$
 $= (1 + 1 \times 2^1 + 1 \times 2^3 + 1 \times 2^5 + 1 \times 2^7)_{10}$
 $= (171)_{10}$

例2: 8進数 $(00012574)_8$
 $= (4 + 7 \times 8^1 + 5 \times 8^2 + 2 \times 8^3 + 1 \times 8^4)_{10}$
 $= (5500)_{10}$

10進数 → n進数

x_0 : 10進数表記のある数

$$x_1 = x_0 \div n, \quad a_0 = x_0 \bmod n$$

$$x_2 = x_1 \div n, \quad a_1 = x_1 \bmod n$$

$\vdots$
 $\uparrow$
 x を n で割った
余りを求める
という記号

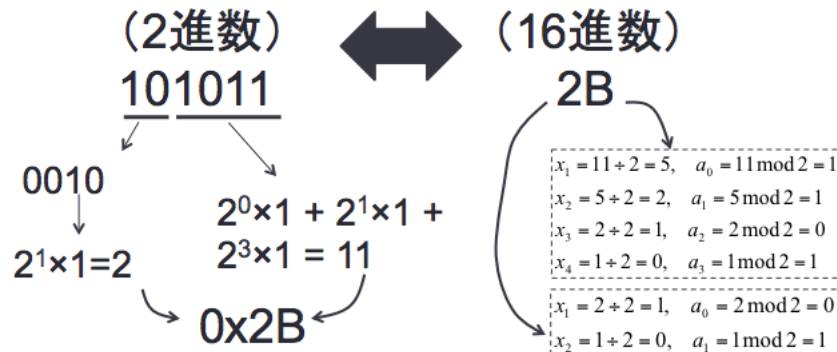
$$x_{m-1} = x_{m-2} \div n, \quad a_{m-1} = x_{m-2} \bmod n$$

$$x_m = x_{m-1} \div n = 0, \quad a_m = x_{m-1} \bmod n$$

n進数表記: $a_m a_{m-1} \cdots a_1 a_0$

2進数 ⇔ 16進数

基本: 16進数の1桁を2進数では4桁で表す



16進数は、数値の先頭に0xをつけて表記されることもある

2進数での負数の表現

- ・符号と絶対値を用いる方法
- ・1の補数による表現
- ・2の補数による表現

(先頭に置き、0を正、1を負とする)
符号ビット(1ビット)

絶対値(7ビット)

10進数	2進数
13	00001101
1	00000001
0	00000000 又 10000000
-1	10000001
-13	10001101

1の補数

ある2進数について、各桁の0/1を反転した表現を1の補数という

1の補数の例 →

元の2進数	1の補数
0101	1010
1100	0011
0000	1111
1111	0000

1の補数を用いて、先頭の桁が0のものが正数を、1のものが負数を表すとすれば、正負の数を表現することができる。負数の絶対値は、対応する正数の値とする。n桁の2進数では $+2^{n-1}-1 \sim -2^{n-1}-1$ までの数値を表現することができる。

10進数	1の補数による表現	0	0000 又は 1111
7	0111	-1	1110
6	0110	-2	1101
5	0101	-3	1100
4	0100	-4	1011
3	0011	-5	1010
2	0010	-6	1001
1	0001	-7	1000

小数の表現

例1

2進少数 $(0.101)_2$ を10進数へ変換する
 $(0.101)_2 = (1 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3})_{10}$
 $= (0.625)_{10}$

例2

16進少数 $(0.1f)_{16}$ を10進数へ変換する
 $(0.1f)_{16} = (1 \times 16^{-1} + 15 \times 16^{-2})_{10}$
 $= (0.12109375)_{10}$

・固定小数点方式

小数点の位置が固定されているものとして扱う方式



16ビットの値で、例えば、上位12ビットを整数、下位4ビットを小数とする、など
 $(000000111111 1110)_2 = (63.875)_{10}$
 $(100000111111 1110)_2 = (-63.875)_{10}$

MSB(最上位ビット)は符号ビット

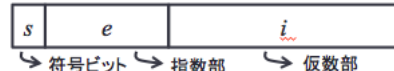
・浮動小数点方式

数値 a を次のような指数で表現する方式
コンピュータでは通常この方式を用いる

$$a = (-1)^s \times i \times n^e$$

例 -0.1101×2^3
 0.11111×2^{-6}

符号 仮数 指数 n (基数)は固定

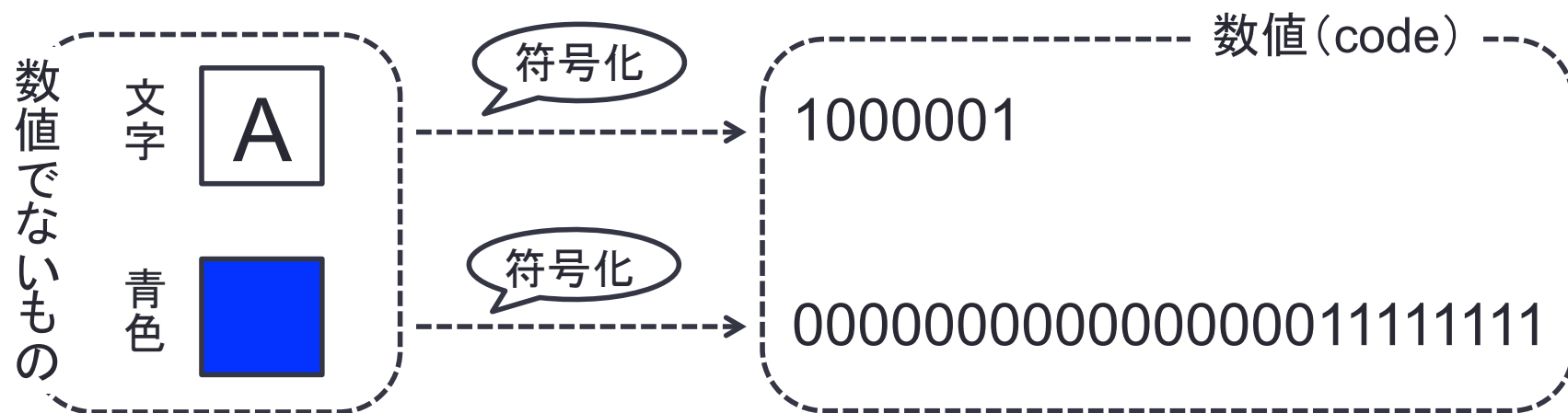


浮動小数点のビット表現

$-0.1101 \times 2^3 \rightarrow 1 0011 1101000000$
 $0.001011 \times 2^{-1} \rightarrow 0 1111 00101100000$

code and encoding

コンピュータの内部では、あらゆる「情報」を「数値」で表して「処理」しています。文字、色、画像、音声、動画のように、人間の感覚では数値でない情報も、何らかの数値に置き換えて処理しているのです。



本来数値でない情報を数値に置き換えたものを「**コード**(code、**符号**)」と呼びます。そして、どのような数値に置き換えるかを定める作業を「符号化(encoding)」と呼びます。文字を数値で表したものは「文字コード」、色を数値で表したもののなら「色コード」というわけです。

文字コード

- ASCII code
- Unicode (UTF-16など)
- JIS code (ISO-2022-JP)
- Shift-JIS code
- EUC code

情報の基本単位は1バイト=8ビットです。8ビットあれば256種類の文字を表せます。そこで、前半の128通りをASCIIコードと同じにして(MSB=0)、後半の128通りを各国独自の文字を割り当てる8ビットのコード体系がよく使われています。日本では、後半の128通りにカタカナや句読点などを割り当てています。しかし、8ビットの256通りでも漢字を表すのには足りません。そこで現在のコンピュータでは、16ビットで1文字を表す「Shift-JSコード」や「Unicode」というコード体系も使われています。16ビットあれば、65536通りの文字にコードを割り当てることができます。

表1.1 ASCIIコード

	010	011	100	101	110	111	← 上位ビット
0000	空白	0	@	P	`	p	
0001	!	1	A	Q	a	q	Aの文字コードは 100 0001
0010	"	2	B	R	b	r	
0011	#	3	C	S	c	s	
0100	\$	4	D	T	d	t	
0101	%	5	E	U	e	u	
0110	&	6	F	V	f	v	
0111	'	7	G	W	g	w	
1000	(	8	H	X	h	x	
1001	)	9	I	Y	i	y	
1010	*	:	J	Z	j	z	zの文字コードは 111 1010
1011	+	;	K	[	k	{	
1100	,	<	L	\	l		
1101	-	=	M	]	m	}	
1110	.	>	N	^	n	~	
1111	/	?	O	_	o	削除	削除を表すコードは 111 1111

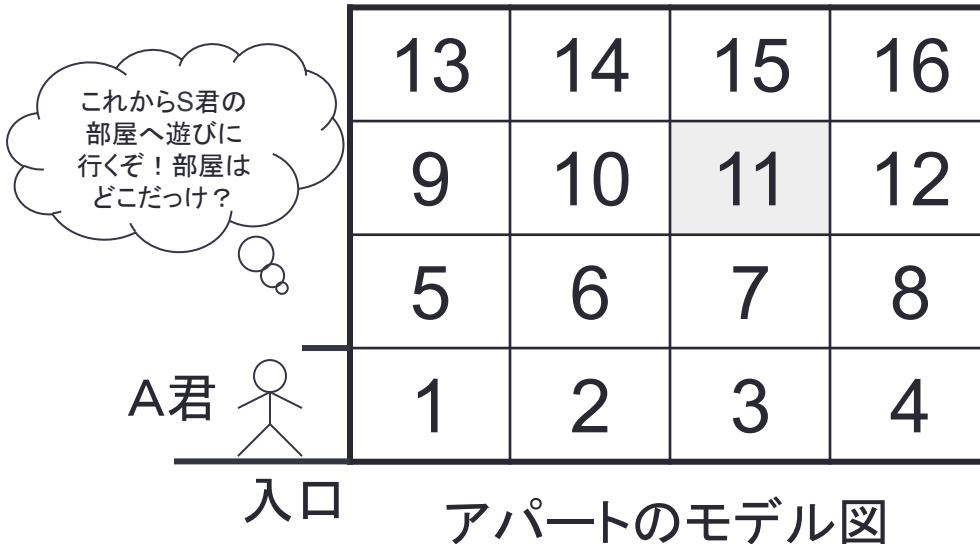
↑ 下位ビット

※\マークは、日本語環境では¥になります。

情報理論の基礎

- 自己情報量 (self information)
- 平均情報量 (average information)
- エントロピー (entropy)
- 最大エントロピー (maximum entropy)
- 冗長度 (redundancy)

情報の量とは、その情報を得た人にとっての内容の豊富さのことであると考えてよい。



A君は入口で次の情報を教えてもらった

情報： S君は11号室に住んでいる

この情報はA君にとって果たしてどれぐらいの情報の量であったのか。

Q. ケース1とケース2では、どちらが入口で得た情報の量が多い(豊富)と考えられるか？

Case 1: A君はS君が何階に住んでいるのか知らなかった.

Case 2: A君はS君が3階に住んでいるのを知っていた.

- 情報の量は、情報を得る前の_____（前例では部屋の数）に関係し、その数が増すにつれ情報の量も増すようであればならない（単調増加関数）。
- 情報の_____が成り立たなければならない。
 - 情報1：S君の部屋は16室の11号室である。
 - 情報2：S君の部屋は3階にある。
 - 情報3：S君の部屋は左から3番目の部屋である。

「情報1」の量＝「情報2」の量＋「情報3」の量



- 情報の量は、可能性の数の対数で定義するのが便利
- 対数の底を2にとると、最小の情報（二択）の量は_____となり、情報量の単位として都合がよい。

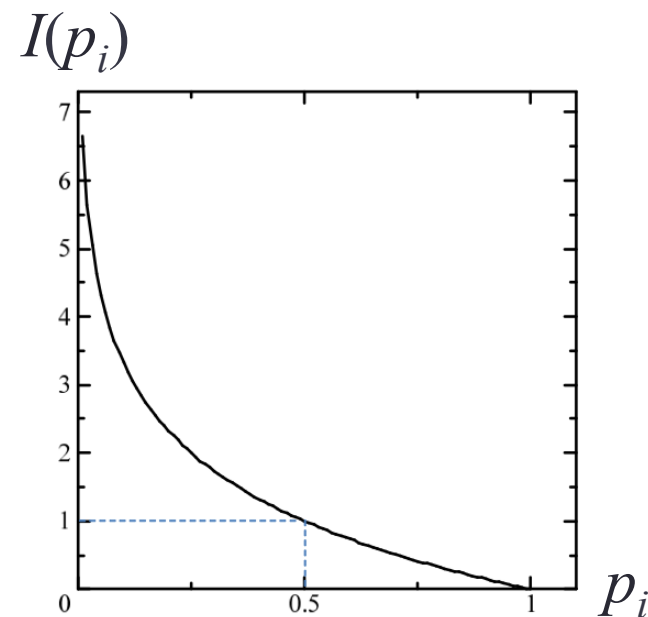
情報量の定義

p_i : ある出来事が発生する確率(事前確率)

$$\begin{aligned} I(p_i) &= \log_2 \frac{\text{事後確率}}{\text{事前確率}} \\ &= \log_2 \frac{1}{p_i} = -\log_2 p_i \end{aligned}$$

自己情報量 or 選択情報量

情報量の単位は_____



自己情報量のグラフ

$p_i = 1/2$ のとき、 $I(p_i) = 1$

$p_i = 1$ のとき、 $I(p_i) = 0$

自己情報量の例

[例1] 赤ん坊が生まれたとき、その男女比が1:1とする。男が生まれる事象をboy、女が生まれる事象をgirlとすると、それぞれの自己情報量は次のようになる。

$$I(\text{boy}) = -\log_2 \frac{1}{2} = 1 \text{ bit}$$

$$I(\text{girl}) = -\log_2 \frac{1}{2} = 1 \text{ bit}$$

[例2] ある試験では、合格する可能性が1/8である。この試験に合格した場合の自己情報量は

$$I(\text{合格}) = -\log_2 \frac{1}{8} = 3 \text{ bit}$$

となる。一方、不合格になったときの自己情報量は

$$I(\text{不合格}) = -\log_2 \frac{7}{8} = -\log_2 7 + \log_2 8 = -2.807 + 3 = 0.193 \text{ bit}$$

情報の加法性

情報1: S君の部屋は16室の11号室である.

情報2: S君の部屋は3階にある.

情報3: S君の部屋は左から3番目の部屋である.

「情報1」の量 = 「情報2」の量 + 「情報3」の量

[例] ジョーカーを除いた52枚のトランプを相手に引いて貰い、その内容を教えてもらうことを考える。

- ・引いたカードが、ハートのAであることを知ったときの情報量は

$$I(\text{ハートのA}) = -\log_2 \frac{1}{52} \approx 5.7 \quad \text{bit}$$

- ・引いたカードが、ハートであることのみを知ったときの情報量は

$$I(\text{ハート}) = -\log_2 \frac{1}{4} = 2 \quad \text{bit}$$

- ・引いたカードが、Aであることのみを知ったときの情報量は

$$I(A) = -\log_2 \frac{1}{13} \approx 3.7 \quad \text{bit}$$

平均情報量 (average information)

情報量の平均(期待値)について考えよう。いま、ある事象系を $\{a_1, a_2, \dots, a_N\}$ とする。これら N 個の事象は互いに排反で、その生起確率 p_i の総和は1とする(完全事象系)。このとき、情報量 $I(p_i)$ の平均、すなわち平均情報量 H は次のように示される。

$$\begin{aligned} H &= p_1(-\log_2 p_1) + p_2(-\log_2 p_2) + \dots + p_N(-\log_2 p_N) \\ &= -\sum_{i=1}^N p_i \log_2 p_i \quad \text{ただし、} \quad \sum_{i=1}^N p_i = 1.0 \end{aligned}$$

- ・平均情報量の取りうる値は $0 \leq H \leq \log_2 N$ bit
- ・事象系 $\{a_1, a_2, \dots, a_N\}$ において、一つの事象 a_i の生起確率が $p_i = 1$ で、その他の事象の生起確率がすべて0のとき $H = 0$ 、これは結果を聞く前から結果が既知なので情報としては価値がない。
- ・事象系 $\{a_1, a_2, \dots, a_N\}$ において、すべての事象の生起確率が _____ の場合は、平均情報量は最大の $H = \log_2 N$ bit となる。これはどれが起きるか全く予想できない状態。

平均情報量の例

[例1]ある家の本日の晩ご飯の生起確率が以下の通りだとする。

$$\begin{aligned}p(\text{カレー}) &= \frac{1}{4}, & p(\text{ハンバーグ}) &= \frac{1}{4}, & p(\text{焼き魚}) &= \frac{1}{8} \\p(\text{からあげ}) &= \frac{1}{8}, & p(\text{とんかつ}) &= \frac{1}{8}, & p(\text{さしみ}) &= \frac{1}{16} \\p(\text{ステーキ}) &= \frac{1}{16}, & p(\text{すき焼き}) &= 0\end{aligned}$$

$$\begin{aligned}H &= -\sum_{i=1}^8 p_i \log_2 p_i \\&= -\frac{1}{4} \log_2 \frac{1}{4} \times 2 - \frac{1}{8} \log_2 \frac{1}{8} \times 3 - \frac{1}{16} \log_2 \frac{1}{16} \times 2 - 0 \log_2 0 \\&= \frac{4}{4} + \frac{9}{8} + \frac{8}{16} + 0 = 1 + 1.125 + 0.5 = 2.625 \quad \text{bit}\end{aligned}$$

ただし、 $x \rightarrow 0$ のとき $x \log_2 x \rightarrow 0$

エントロピー (entropy)

熱力学における分子の

を表す尺度

熱力学における
エントロピー

$$H = -K \sum_k n_k \ln n_k$$

K はボルツマン定数, n_k は気体分子の k 番目のエネルギー状態にある確率

情報理論における
エントロピー

$$H = - \sum_k p_i \log_2 p_i$$

熱力学におけるエントロピーと平均情報量は定数倍, 対数の底を除いて一致する. そのため, 平均情報量を情報理論における
と呼ぶことにする.

最大エントロピー (maximum entropy)

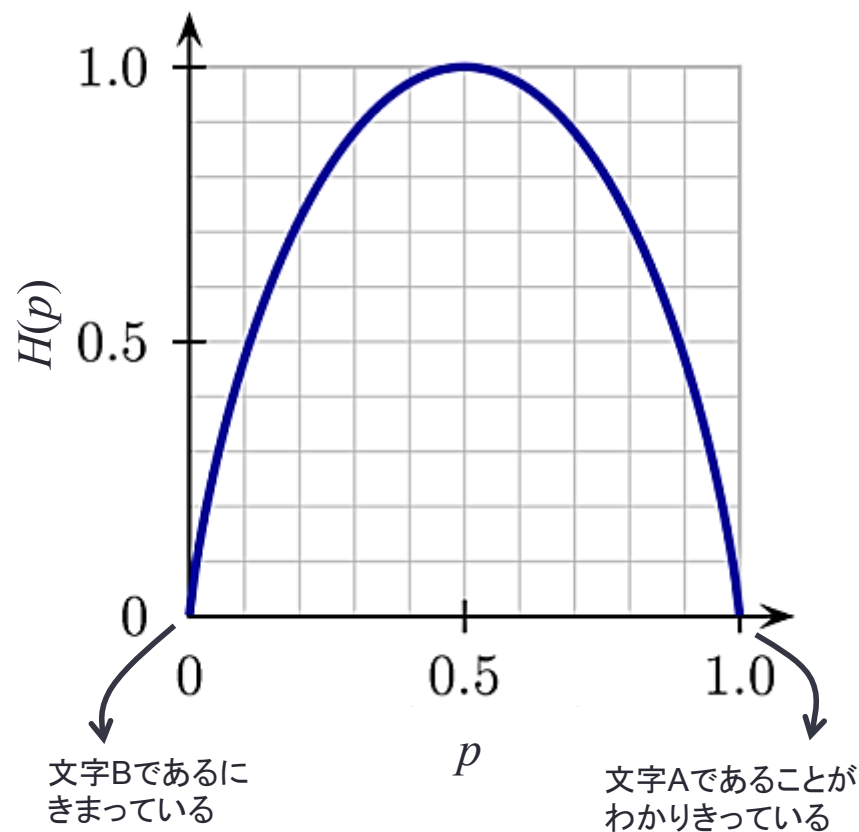
2種類の文字A, Bが, それぞれ確率 p , $1-p$ で生起する情報源のエントロピーは次のように表される.

$$H = -\{p \log p + (1-p) \log(1-p)\}$$

H を確率 p の関数として示すと右のようなグラフ(エントロピー関数)になる.

H が零になるのは, $p = 1$ か $p = 0$ のとき.

H が最大になるのは, $p = 0.5$ のとき.



すべての出来事が_____で発生すると仮定した場合のエントロピーを_____という.

最大エントロピーの例

以下の例はいずれも各事象の生起確率が等確率と仮定する.

- サイコロを一回振る時の最大エントロピー

$$H_{\max} = - \sum_{i=1}^6 \frac{1}{6} \log_2 \frac{1}{6} = \log_2 6 = 2.585 \quad \text{bit}$$

- 英数字 (A~Zと空白, 計27文字) の最大エントロピー

$$H_{\max} = - \sum_{i=1}^{27} \frac{1}{27} \log_2 \frac{1}{27} = \log_2 27 = 4.755 \quad \text{bit}$$

- 常用漢字 (1945文字) の最大エントロピー

$$H_{\max} = - \sum_{i=1}^{1945} \frac{1}{1945} \log_2 \frac{1}{1945} = \log_2 1945 = 10.925 \quad \text{bit}$$

冗長度 (redundancy)

最大エントロピー: $H_{\max}$ とエントロピー: H の違いは情報源に含まれる「無駄さ」である。与えられた情報源がどれだけ無駄なものを含むかの度合いを r という。これは情報の中で実際の情報以外のものの割合とも言える。

冗長度: r は次のように定義される

$$r = 1 - \frac{H}{H_{\max}} = 1 - \frac{\text{エントロピー}}{\text{最大エントロピー}}$$

冗長性があるおかげで

- ・データの圧縮が可能
- ・正確な情報通信が可能

情報そのものは減らすことなく、情報以外の冗長な部分を切り詰める

あえて冗長な部分を付加・利用して、情報の正確さを判定・担保する

例： 4つの文字 (A, B, C, D) からなるデータ

- ・ 出現率が均等である場合

$$p_A = p_B = p_C = p_D = 1/4$$

冗長度がなく、
最大エントロピー $H_{\max}$ に一致する

$$\text{平均情報量 } H = \log_2 4 = 2.0 \text{ bit}$$

- ・ 出現率に偏りがある場合

$$p_A = 1/2, p_B = 1/4, p_C = p_D = 1/8$$

冗長度があり、得られる
情報量は見かけより少ない

$$\begin{aligned} \text{平均情報量 } H &= \quad \times \log_2 2 + 1/4 \times \log_2 4 + 1/4 \times \log_2 8 \\ &= 0.5 + 0.5 + 0.75 \\ &= 1.75 \text{ bit} \end{aligned}$$

冗長度 $r =$ _____

情報量を、情報を表現できるビット数であると考えれば、出現率に偏りがある情報のほうが情報量が小さくなる。すなわち、少ないビット数で表現できることを意味する。これは、データ圧縮、特に可逆圧縮の基本原則である。可逆圧縮は元のデータに完全に復元できる。それに対して非可逆圧縮は元のデータに完全には復元できない。

安全性・信頼性を高めるために冗長性は必要不可欠である。

- ・列車や航空機での ブレーキ系統や制御系統は二重化
- ・コンピューターシステムにける, 同じシステムを二重に作っておくdual system
- ・信頼性工学における「繊維束モデル」, あるいは「鎖モデル」に対する「綱モデル」

要は, 多重化すること, スペアを持つこと, 冗長性を持つことの意味である。

コンピュータのデータ伝送時の「誤り検出訂正」ための方法として, 「パリティ・チェック」「ハミングコード」「CRC」などの方法があるが, そのいずれにおいても「冗長ビット」と云うものを付加する。

「パリティ・チェック」
(奇偶検査)

7桁ごとに数字を加算しこれにもう1桁を付加して8桁の合計が必ず奇数になるようにする

1011011 → の場合は合計が5で奇数であるから → 0 を付加して → 10110110にする

1011010 → の場合は合計が4で偶数であるから → 1 を付加して → 10110101にする

このようにすると8桁の合計が必ず奇数になる。だから、8桁区切りで合計して、奇数になっていなかったら、誤りがあることが分かるので、再伝送してもらう。この最後に加えた0または1の桁が「冗長ビット」である。