Foundations of Information Processing

Data and data encoding

Considerations about data and its meaning

Consideration 1:

How to understand

data, information, and knowledge?

Consideration 2:

From where to get data?

Is it unambiguous?

Consideration 3:

How to code data?

What kind of representation notations to use?

Data and data encoding

“In God we trust, all others bring data.”

(Prof. William Edwards Deming, an American engineer, statistician, Yale University, PhD, sampling, Japanese concept of quality)

- Input of algorithms: data, sources of data, and data encoding.
- Concepts:
  - Data, information, knowledge (conclusions, experts).
  - ”To know is nothing, to imagine is everything”.
- Sources of data:
  - Observations/measurements.
  - Literature/stored data.
  - Experts.
- Data encoding:
  - Number theory: numbers and number systems.
  - Characters and notations.

Source (partly/modified from): J. Boberg, Johdatus tietojenkäsittelytieteeseen, Turun yliopisto, 2010 (in Finnish)

Concepts

- Data:
  - a regular presentation of the matter in a communicable or processable form.
- Information:
  - Conception or meaning produced by data for the human.
- Knowledge:
  - The result of the human's thinking based on data and information.
- In Information Processing:
  - Data is the presentation of the matter in a computer, for example, a string of characters which do not have any meaning before the interpretation of data.
  - Knowledge is the matter considered by the human, the interpretation of data.
  - 39 (just a number) => 39 °C (temperature of a human) => high fever (interpretation).

Definitions: data (singular datum)

Oxford English Dictionary, OED

- A thing given or granted; something known or assumed as fact, and made the basis of reasoning or calculation; an assumption or premise from which inferences are drawn.
- The *quantities*, *characters*, or *symbols* on which operations are performed by computers and other automatic equipment, and which may be *stored* or *transmitted* in the form of electrical signals, records on magnetic tape or punched cards, etc.
- Facts, esp. numerical facts, collected together for reference or information.

Definitions: information

OED

- Knowledge communicated concerning some particular fact, subject, or event; that of which one is apprised or told; intelligence, news.
- As a mathematically defined *quantity* divorced from any concept of news or meaning; spec. one which represents the degree of choice exercised in the *selection* or *formation* of one particular *symbol*, *message*, etc., out of a number of possible ones, and which is defined logarithmically in terms of the statistical *probabilities* of occurrence of the symbol or the elements of the message.
- Separated from, or without the implication of, reference to a person informed: that which inheres in one of two or more alternative *sequences*, *arrangements*, etc., that produce different *responses* in something, and which is capable of being *stored* in, *transmitted* by, and *communicated* to inanimate things.
- Contrasted with data: that which is obtained by the *processing* of data.

Definitions: knowledge

OED

- The fact of knowing a thing, state, etc., or (in general sense) a person; acquaintance; familiarity gained by experience.
- Acquaintance with a fact; perception, or certain information of, a fact or matter; state of being aware or informed; consciousness (of anything).
- Acquaintance with *facts*, range of *information*.
- Knowledge of a person, thing, or perception gained through *information* or *facts* about it rather than by direct experience.
- Acquaintance with a branch of learning, a language, or the like; theoretical or practical understanding of an art, science, industry, etc.
- The fact or condition of being instructed, or of having information acquired by *study* or *research*; acquaintance with ascertained *truths*, *facts*, or *principles*; information acquired by study; learning; erudition.
- The sum of what is known.

Summary about data

- Data, information, and knowledge is a concept that is ambiguous, and its comprehensive definition is challenging to make.
- Data alone do not reveal how useful or reliable is data.
- *Source criticism* and *understanding data*: the more important, the more consequences are based on its interpretation.
  - Especially nowadays when there is a lot of “information” available in real-time in many platforms of media.

    => a lot of disinformation/misinformation, unfortunately!
- Even observed or measured data are not objective
  - if understanding about the event of observation or measuring and its possible effects are incomplete.
  - For example, what radiation is dangerous?

Measurements as information source

A. Robinson, Mittaamisen historia, Multikustannus, 2008 (in Finnish).

- Importance of measuring:
  - ”Measuring is needed to succeed”
    - Is it so?
  - Are conclusions based on measurements valid?
  - Can we measure imagination and creativity?
  - Uniformity and precision.
    - Quantities and units: e.g., weight 92 kg.
    - The metric system (SI, International System of Units).
- What to measure?:
  - Nature: atoms, the Earth, the Universe.
  - Humans: the mind, the body, the society.

Measurements as information source: examples of quantities

- How many? => Number/Count.
- Weight, density.
- Height, distance, location.
- Area, volume.
- Angle.
- Currency, stock.
  - Value of currencies.
- Not so easy in all cases:
  - For example, time.
  - Why do we have sexagesimal as a number system?
  - Why do we have different calendar systems?
  - More about time in the next slides.

Time as a measured quantity

- Sexagesimal as a number system.
  - The base is 60.
  - 60 is a superior highly composite number: 2, 3, 4, 5, 6, 10, 12, 15, 20 and 30 where 2, 3, and 5 are prime numbers.
- The Babylonians in Mesopotamia around 4000 years ago (Iraq nowadays).
  - The circle was divided into 360 degrees where each degree into 60 minutes and each minute to 60 seconds.
- The Egyptians divided the day into hours and adopted 1 hour (h) = 60 minutes (min) from the Babylonians.
- The Babylonians were not able to measure time units less than one second (s or sec).
  - The decimal system appeared for fractions of the second only later.
- The units of time, length, area, and volume: divisible to each other.
- Basis coordinates: degree => 60 minutes => 60 seconds.

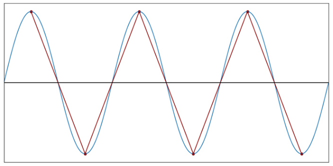
Time as a measured quantity (cont.)

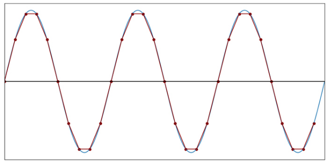
- The Sun and the Moon inspired:
  - Orbits, sunsets, sunrises, 24 h, 6 nighttime hours, 6 daylight hours.
- The Islamic calendar is a lunar calendar.
- The Julian calendar (Julius Caesar): every fourth year is a leap year (there is the leap day).
- The Gregorian calendar (Pope Gregorius XIII, Feb 24,1582):
  - The years divisible by 100 which are not divisible by 400 are not considered.
  - The period of 400 years includes 97 leap years, not 100.
  - When is the leap day? Now February 29th, earlier 24th (Caesar).
  - Finland started to apply the Gregorian calendar with the long delay in 1753 as a part of Sweden. Political and religious reasons?
  - What was the official time when Finland was a part of Russia which followed the Julian calendar? A handy solution:
    - The both dates were marked in official documents.

Measurements: properties

- Accuracy of the quantity.
  - The measured value vs. the real value.
- Precision.
  - Repeated measurements produce the same value.
- Error.
  - For example, systematic errors due to the reading of the measuring device or incorrect calibration.
- Uncertainty.
  - For example, due to measuring technology.
- Calibration and traceability.
  - “Exact” definitions: what is one meter or one kilogram?
  - “Natural” references.
    - For example, how to define an inch naturally?
    - 12th of a foot, or a width of a thumb, or something else?

Measurements: real results and their coding





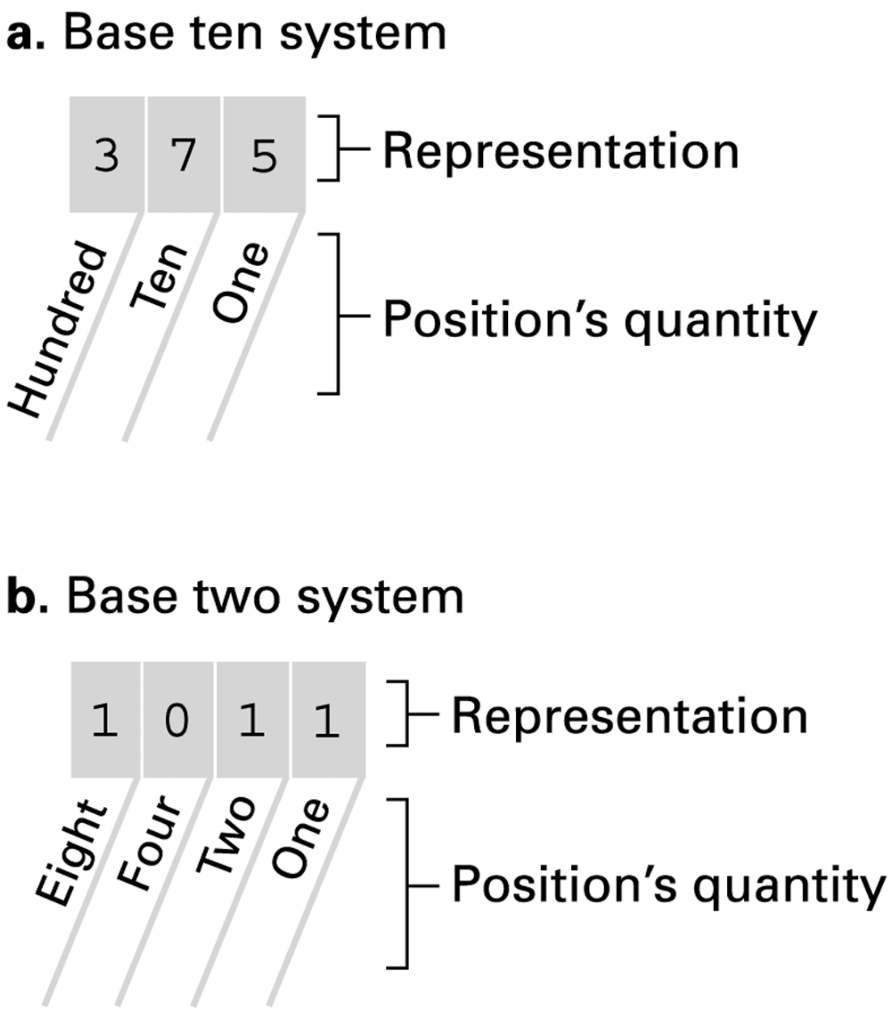
- Analog vs. digital signals, benefits of digitalization.
- Converting signals and preserving information:
  - Sampling frequency $F_s > 2 \cdot F_{max}$
    where F_{max} is the maximum frequency.
  - Nyquist–Shannon theorem.
  - Measure frequently enough, or information may be lost (too much).
  - Consider how a measuring environment and a measuring device may affect the phenomenon to be measured.
- Examples of real signals.
  - Mechanical vibration (speech, music), images, videos.
  - ECG (electrocardiograms).
  - Recognizing songs by a mobile.
  - Speech recognition.
  - Image and video processing.

Data encoding: number systems

- Number: a mathematical object which has the value.
- Digit: a representation of the number.
- Number systems (or numeral systems): the base-*k* system.
  - The base *k* is usually presented as a subscript in the number representation, e.g., 534_8 for the base-8 numbers.
  - The number can be presented using *k* digits.
  - Positional systems: the value of the digit depends on its value and its position related to other digits.
- Base-2, -8, -10, and -16 systems ⇒
  binary, octal, decimal, and hexadecimal numbers.
- Base-10 or decimal system:
  - The base k=10, digits = 0,1,…,8,9.
  - Why do we use the decimal system? 10 fingers??

Decimal and binary number systems

Brookshear, J.G. *Computer Science - An overview*, 7th ed. Addison Wesley, 2003



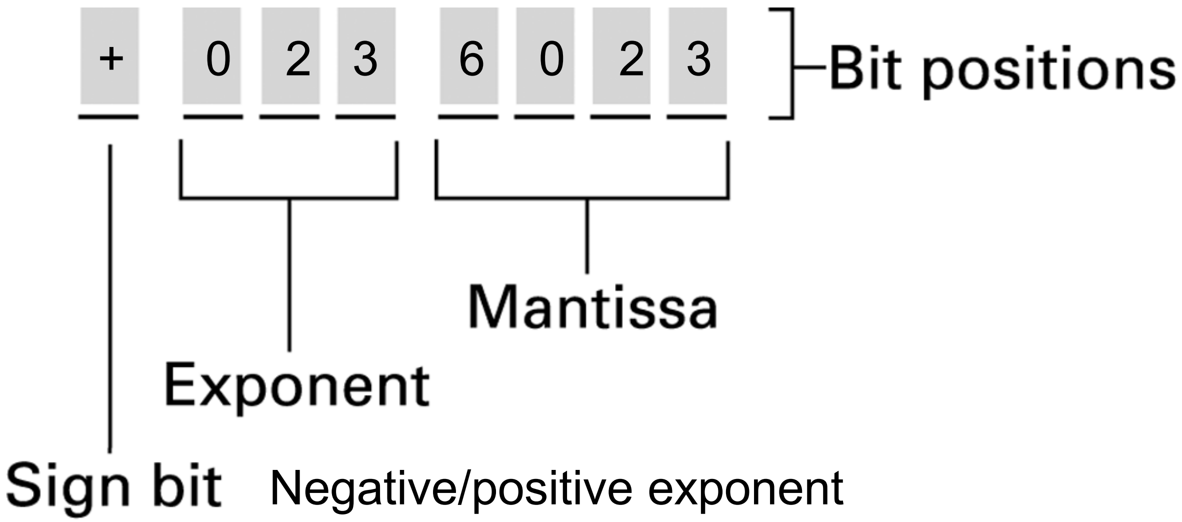
$$3 \cdot 10^2 + 7 \cdot 10^1 + 5 \cdot 10^0$$
$$300 + 70 + 5 = 375$$

$$1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0$$
$$8 + 0 + 2 + 1 = 11$$

Decimal system

Brookshear, J.G. Computer Science - An overview, 7th ed. Addison Wesley, 2003

- Two possible notations.
- Fixed-point representation:
  - The decimal point (also called a separator, either "." or ",") is fixed between the integer and the decimal:
    for example, 66,339443 or 0,0000489.
  - The range is limited in very large and small numbers.
- Floating-point representation: the decimal point floats as needed.
  - For example: $6{,}023 \cdot 10^{23}$



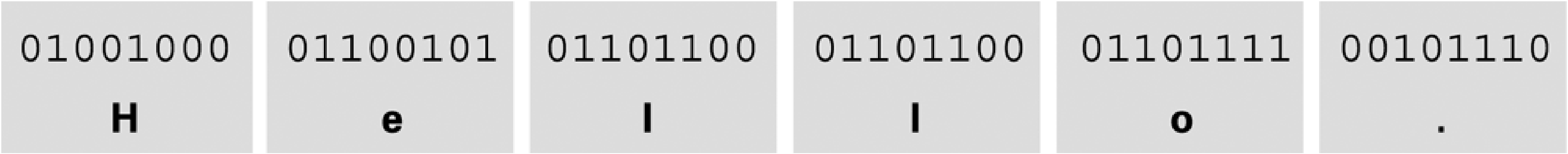
Negative/positive exponent

Representing text: characters and alphabets

- Characters and numbers in strings.
  - Different alphabets, special characters.
- 1 byte = 8 bits.
  - bit = binary digit, Claude Shannon, 1948.
- Each character is presented by own bit pattern (encoding).
- ASCII or ISO 8859:
  - Patterns of 7 or 8 bits = 1 byte/character.
  - For example, the Scandinavian characters are defined by the 8-bit patterns.
  - ASCII is not enough since there are many Chinese, Japanese, Hebrew, etc. characters.
- Unicode Transformation Format (UTF):
  - Up to 32-bit patterns => 1-4 bytes/character.
  - UTF-8 where 8-bit patterns are used is the ASCII encoding.

“Hello.” in ASCII

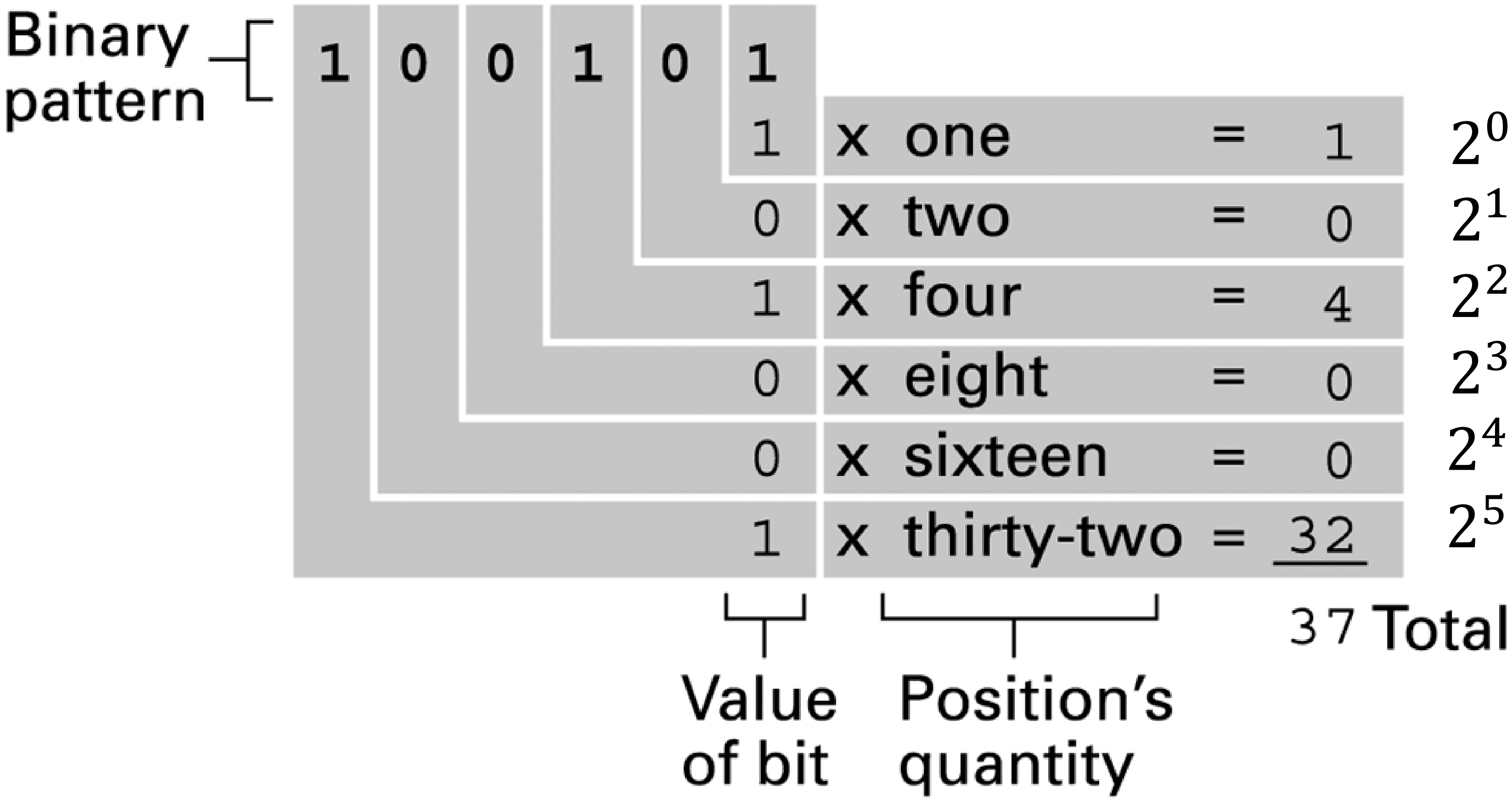
Brookshear, J.G. *Computer Science - An overview*, 7th ed. Addison Wesley, 2003



01001000	01100101	01101100	01101100	01101111	00101110
H	e	l	l	o	.

Number system conversions: Binary → Decimal

Brookshear, J.G. *Computer Science - An overview*, 7th ed. Addison Wesley, 2003



Binary pattern

1 0 0 1 0 1

1 x one = 1

0 x two = 0

1 x four = 4

0 x eight = 0

0 x sixteen = 0

1 x thirty-two = 32

37 Total

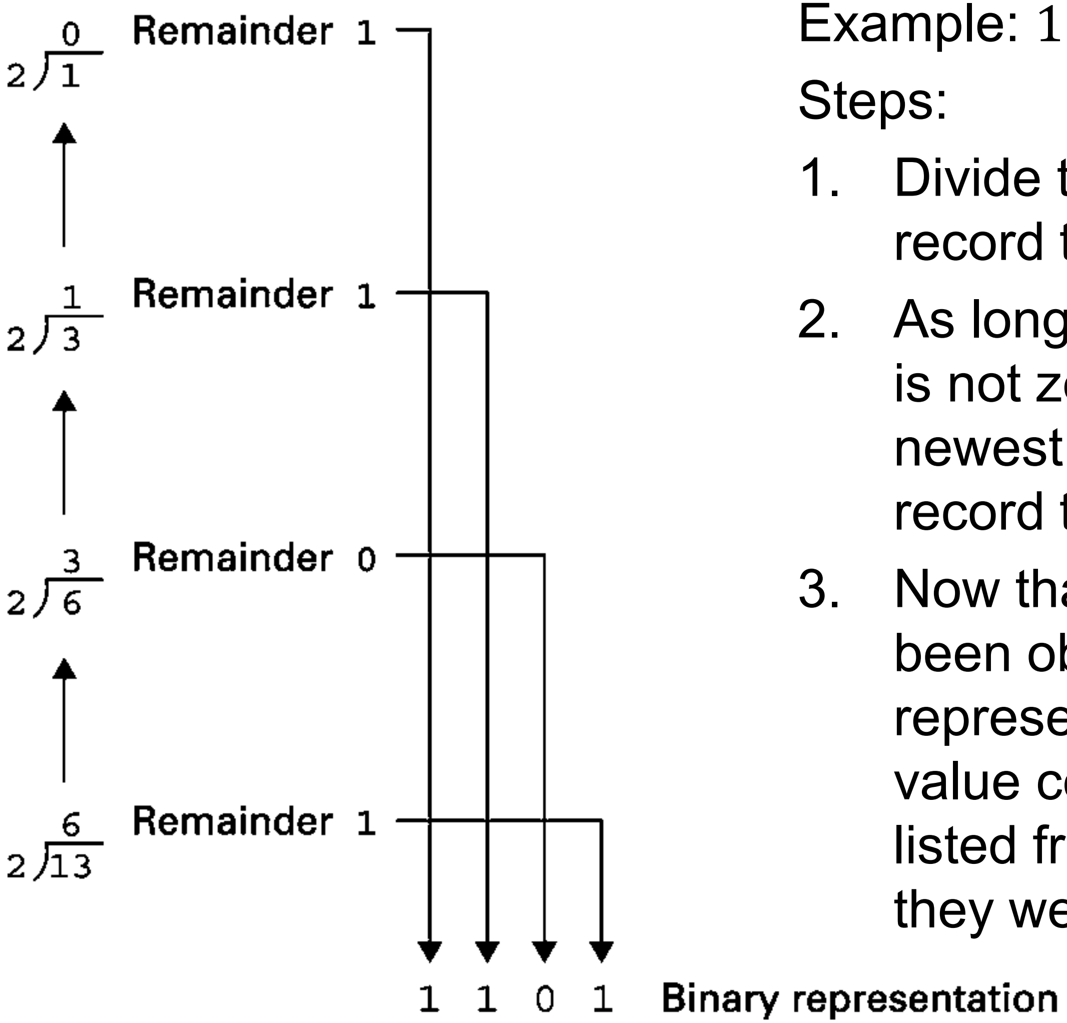
Value of bit

Position's quantity

$$2^0 = 1$$
$$2^1 = 2$$
$$2^2 = 4$$
$$2^3 = 8$$
$$2^4 = 16$$
$$2^5 = 32$$

Decimal → Binary

Brookshear, J.G. Computer Science - An overview, 7th ed. Addison Wesley, 2003



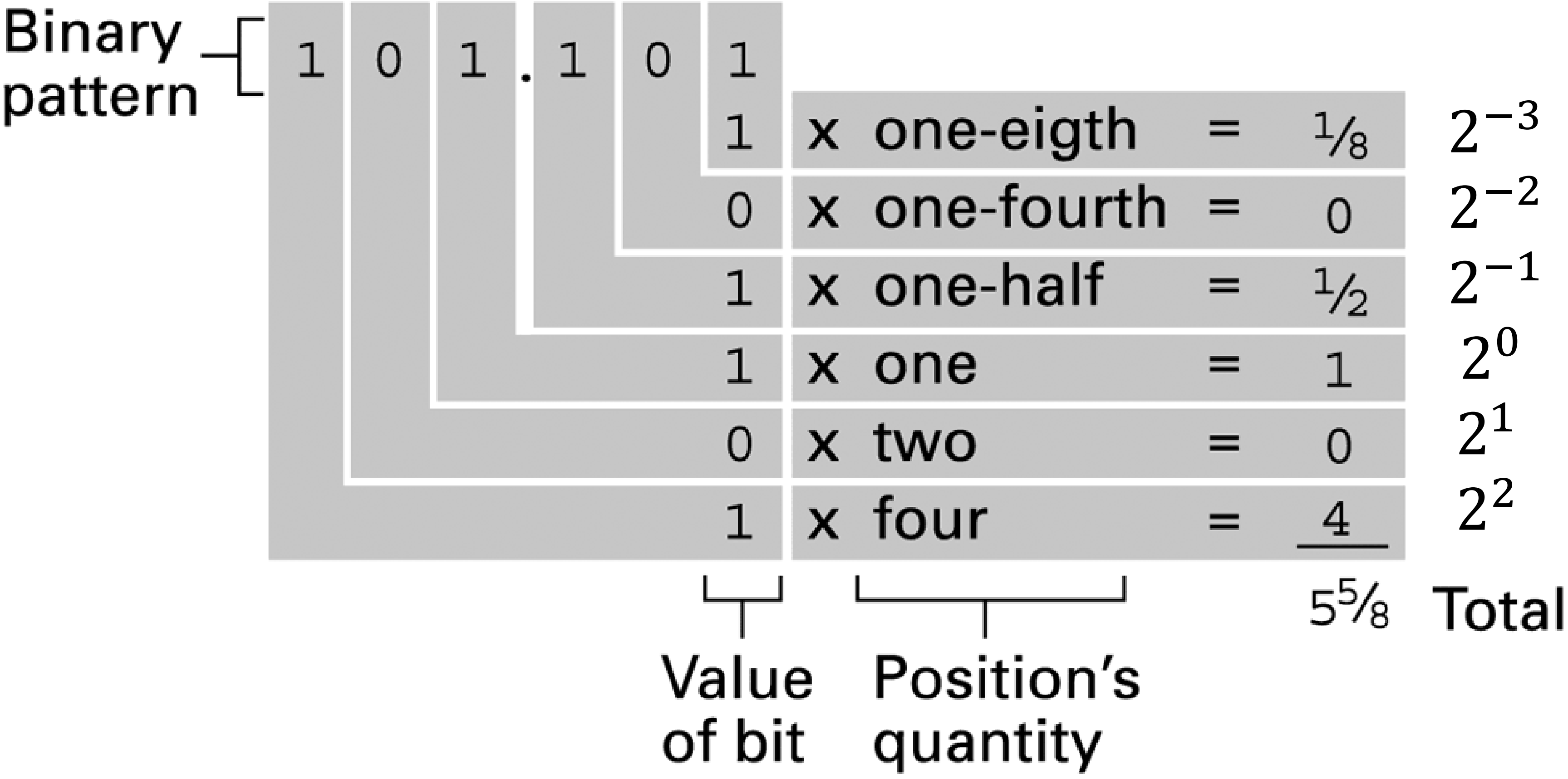
Example: 13_{10}

Steps:

1. Divide the value by two and record the reminder.
2. As long as the quotient obtained is not zero, continue to divide the newest quotient by two and record the reminder.
3. Now that a quotient of zero has been obtained, the binary representation of the original value consists of the reminders listed from right to left in the order they were recorded.

Fractions in the binary system

Brookshear, J.G. *Computer Science - An overview*, 7th ed. Addison Wesley, 2003



$2^{-3} = 1/8$

$2^{-2} = 1/4$

$2^{-1} = 1/2$

$2^{0} = 1$

$2^{1} = 2$

$2^{2} = 4$

Number system conversions in general

- First, convert the number of the base-b system to the decimal system as follows (a_i are the corresponding values to be converted):

$$a_{n-1}b^{n-1} + a_{n-2}b^{n-2} + \cdots + a_1 b + a_0 + a_{-1}b^{-1} + a_{-2}b^{-2} + \cdots + a_m b^{m-1}$$

- Divide the *integer* part of the decimal number by the base of the target system.
  - The reminders form the integer of the target system from right to left.
- Multiply the *decimal* of the decimal number by the base of the target system.
  - The integers form the decimals of the target system from left to right.
- Next, examples are shown.

Binary → Octal

- Convert 110101100.011010001_2 to the base-8 system (the octal system) via the decimal system.
- Conversion to the decimal system:

$$110101100.011010001_2$$

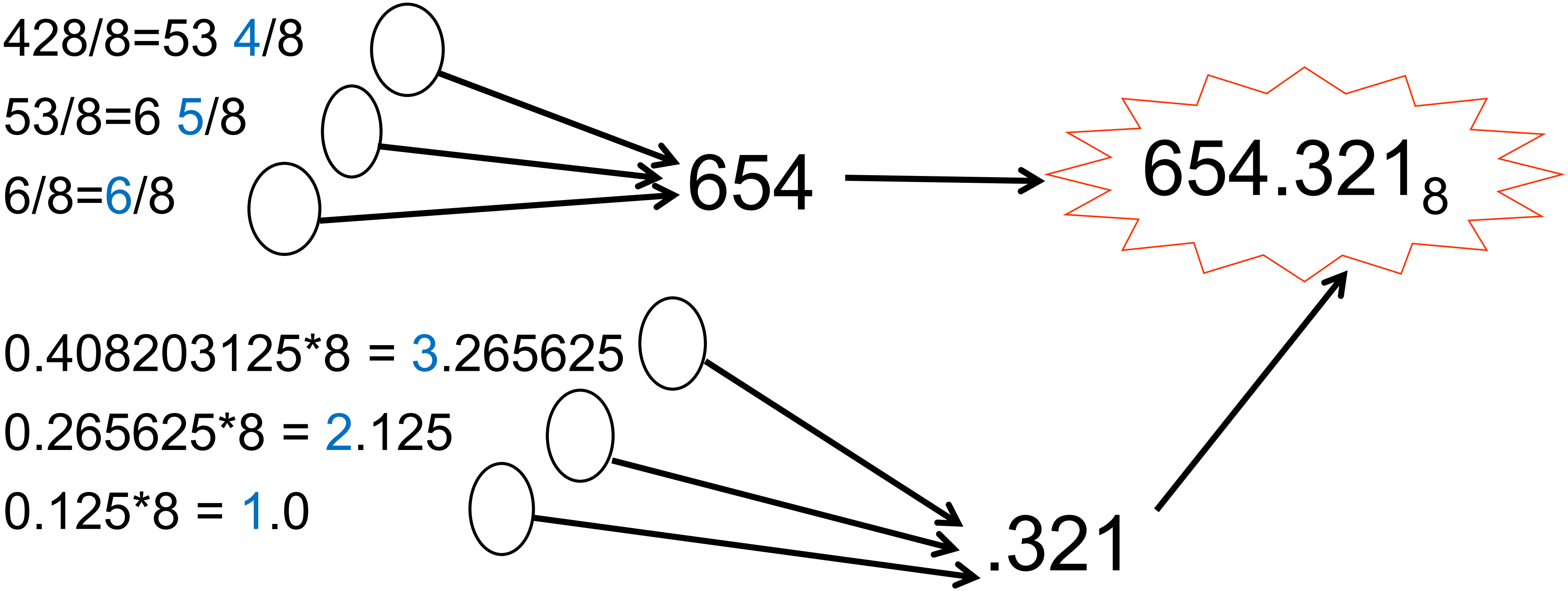
$$2^8 + 2^7 + 2^5 + 2^3 + 2^2 + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^5} + \frac{1}{2^9}$$

$$256 + 128 + 32 + 8 + 4 + \frac{1}{4} + \frac{1}{8} + \frac{1}{32} + \frac{1}{512}$$

$$428\frac{209}{512}$$

$$428.408203125_{10}$$

$428.408203125_{10} \rightarrow$ Octal



428/8=53 4/8

53/8=6 5/8

6/8=6/8

654

654.321_8

0.408203125*8 = 3.265625

0.265625*8 = 2.125

0.125*8 = 1.0

.321

Binary ↔ Hexadecimal

- 4 bits can represent $2^4 = 16$ numbers: 0000 0001 0010 0011 … 1111
- 2 → 16:
  - Arrange the binary value from right to left as 4-bit patterns.
  - Convert the patterns to the base-16 system:

    0,1,2,3,4,5,6,7,8,9, A, B, C, D, E, F

    0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15 (as decimal numbers)
  - Example: 10011101

    10011101 binary

    9 D hexadecimal

    (9 13 4-bit patterns as decimal)
- 16 → 2:
  - Convert the hexadecimal number digit by digit to the binary number.

    9 D hexadecimal

    10011101 binary

Binary addition

$$\begin{array}{r} 0 \\ +\,0 \\ \hline 0 \end{array} \qquad \begin{array}{r} 1 \\ +\,0 \\ \hline 1 \end{array} \qquad \begin{array}{r} 0 \\ +\,1 \\ \hline 1 \end{array} \qquad \begin{array}{r} 1 \\ +\,1 \\ \hline 10 \end{array}$$

Examples:

$$\begin{array}{r} 1\,1 \\ +\quad 0\,1 \\ \hline 1\,0\,0 \end{array} \qquad \begin{array}{r} 0\,1 \\ +\,1\,0 \\ \hline 1\,1 \end{array} \qquad \begin{array}{r} 1\,0\,0 \\ +\,0\,1\,1 \\ \hline 1\,1\,1 \end{array} \qquad \begin{array}{r} 1\,1 \\ +\quad 1\,1 \\ \hline 1\,1\,0 \end{array}$$

Integers: positive and negative signs

- Integers with no signs can be presented directly in the binary system.
- When computing with integers with signs, positive and negative numbers must be considered, depending on their signs.
- The sign bit (the first bit) and the absolute value (the rest).
  - Sign: 1 (negative) ja 0 (positive).
  - Problem: the number lines of positive binary patterns and negative binary patterns follow different directions.
  - Example: add -2 and +3.

$$\begin{array}{rcl} -2 & = & 1010 \\ \underline{+3} & \underline{=} & \underline{0011} \\ & & 1101 = -5 \end{array}$$

This is incorrect!

  - How to solve this problem?

Integers: one’s complement notation

- **One’s complement**: for negative numbers, make the complement of the corresponding bit pattern of the absolute value (0→1, 1→0).
- Example: -2 => patterns of length four => 1010 => 1101
- This way the number line of negative numbers becomes same as the number line of positive numbers.
- Does it work now?:

$$\begin{array}{rcl} -2 & = & 1101 \\ \underline{+3} & \underline{=} & \underline{0011} \\ & & 0000 = (+)0 \end{array}$$

- The result is wrong by one bit. The problem is that in this notation zero contains two representations (+0 and -0).
- The solution: let us have one representation of zero only (+0) so let us shift the bit patterns of negative numbers by one to remove -0.
    - **Two’s complement** notation.
- See examples about two’s complement on the next page.

Two’s complement notation

Brookshear, J.G. *Computer Science - An overview*, 7th ed. Addison Wesley, 2003

Two’s complement:

For negative numbers,

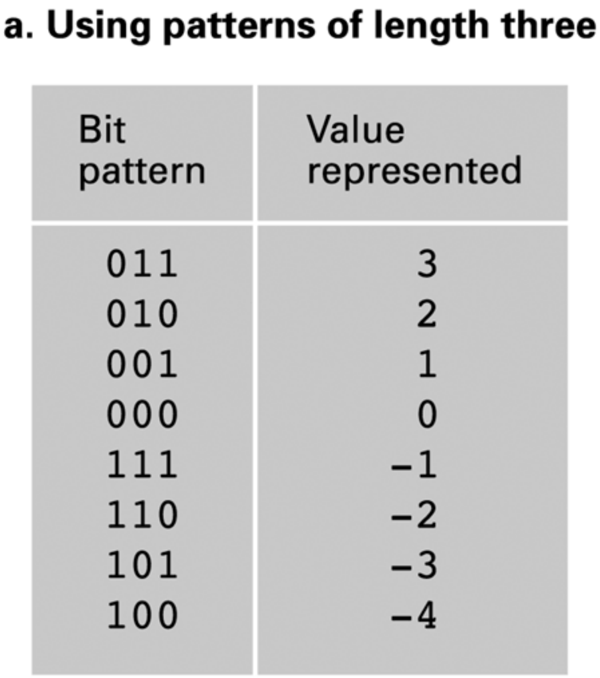
1. Make the complement of the corresponding bit pattern of the absolute value (0→1, 1→0).
2. Add one (+1).

For example: -1

1001 => 1110 + 1 = 1111

Our example:

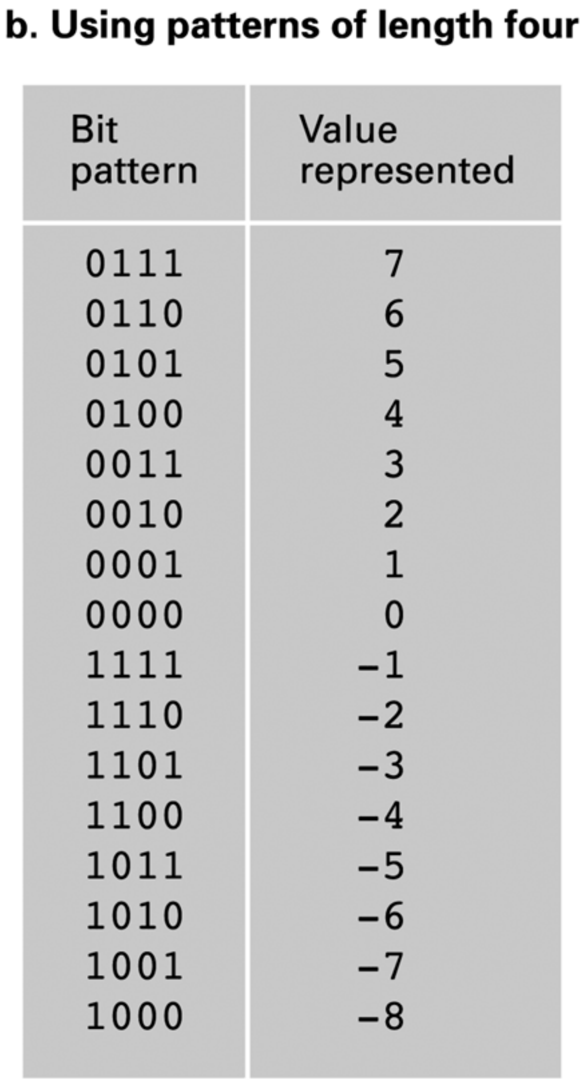
-2 = 1110
+3 = 0011
0001 = +1



a. Using patterns of length three

Bit pattern	Value represented
011	3
010	2
001	1
000	0
111	−1
110	−2
101	−3
100	−4

- The first bit is the sign bit.
- The rest of bits represent the value in the positive/negative representation.

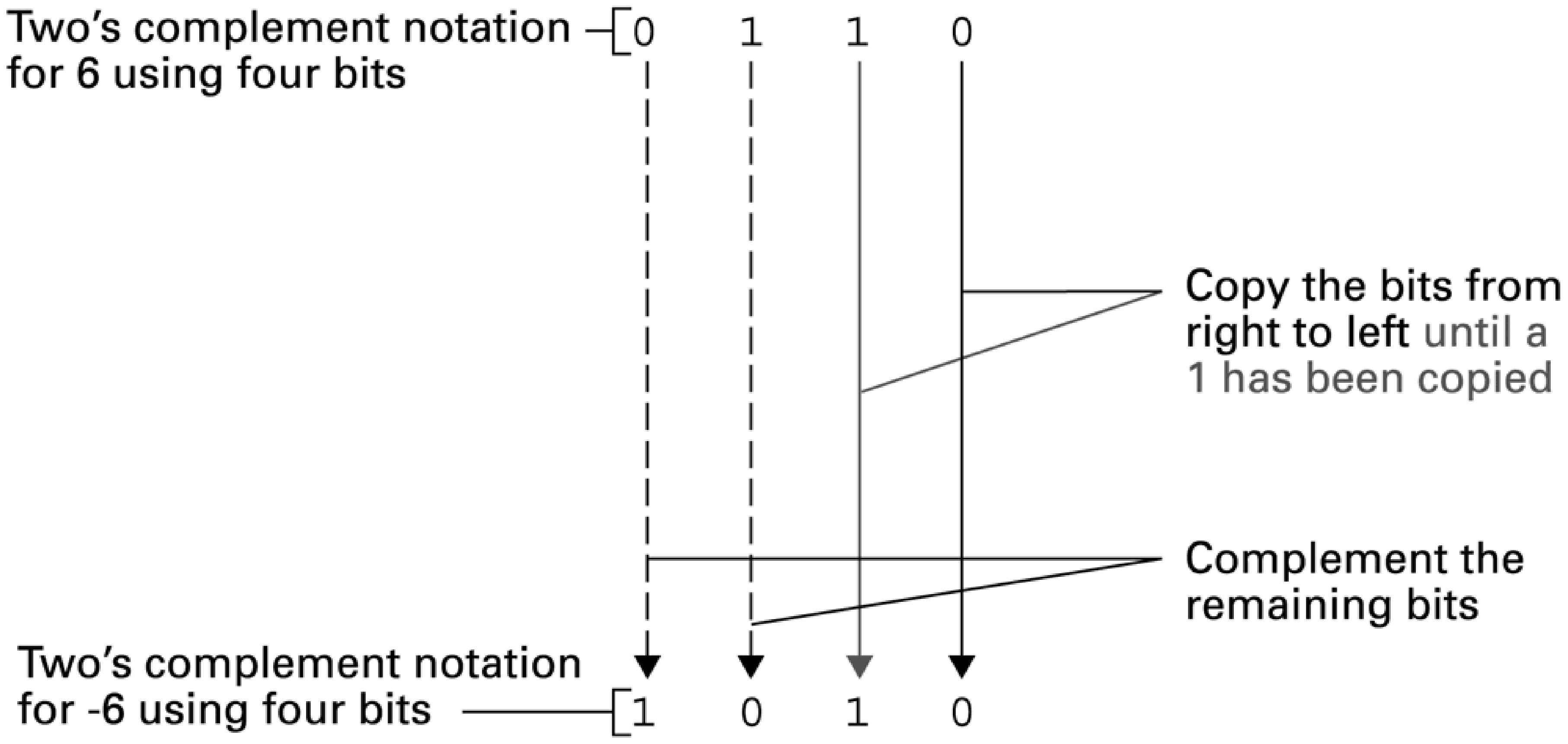


b. Using patterns of length four

Bit pattern	Value represented
0111	7
0110	6
0101	5
0100	4
0011	3
0010	2
0001	1
0000	0
1111	−1
1110	−2
1101	−3
1100	−4
1011	−5
1010	−6
1001	−7
1000	−8

Alternative way to form two’s complement: from +6_{10} to -6_{10}

Brookshear, J.G. *Computer Science - An overview*, 7th ed. Addison Wesley, 2003



Two’s complement notation for 6 using four bits

0 1 1 0

Copy the bits from right to left until a 1 has been copied

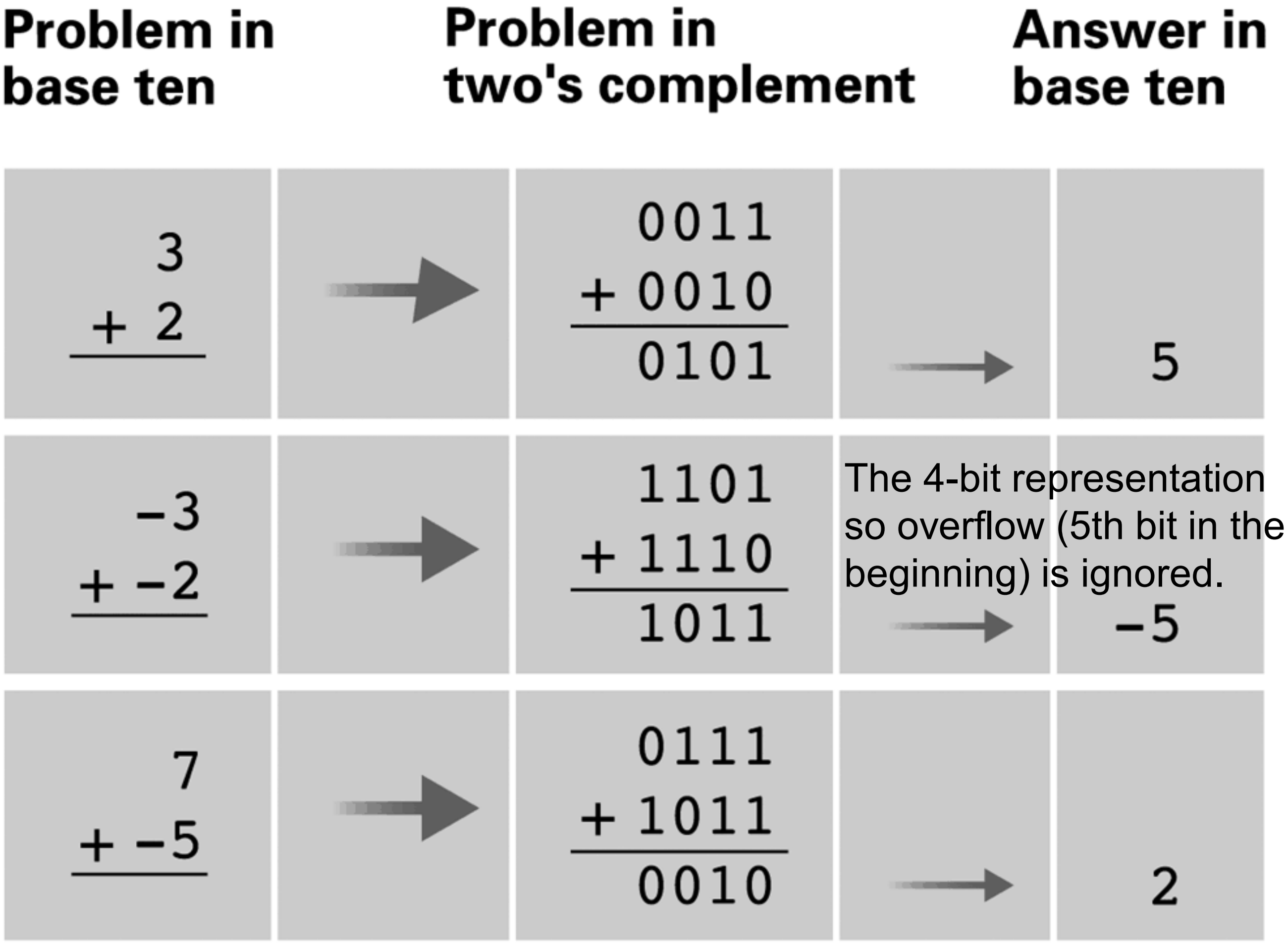
Complement the remaining bits

Two’s complement notation for -6 using four bits

1 0 1 0

Examples: addition problems converted to two’s complement notation (length four)

Brookshear, J.G. *Computer Science - An overview*, 7th ed. Addison Wesley, 2003



Problem in base ten		Problem in two's complement		Answer in base ten
3 + 2	→	0011 + 0010 0101	→	5
−3 + −2	→	1101 + 1110 1011	→	−5
7 + −5	→	0111 + 1011 0010	→	2

The 4-bit representation so overflow (5th bit in the beginning) is ignored.

+7	0111
+6	0110
+5	0101
+4	0100
+3	0011
+2	0010
+1	0001
0	0000
-1	1111
-2	1110
-3	1101
-4	1100
-5	1011

In case of subtraction, change it to addition: $(-5) - (-2) = -5 + 2$

$$\begin{array}{r} 1011 \\ +0010 \\ \hline 1101 \end{array}$$

Summary

- Data, information and knowledge are the input to algorithms.
- This input can be received from different sources, especially in technology usually from measurements.
- The results of measurements are presented as numbers and characters and are encoded for users according to the selected standard.
- Several number systems exist: the base of 2, 8, 10, 16.
- Binary integers are added using the two’s complement notation.