

The logo for ICT Foundation is a light gray rectangular shape with a white archway in the center. The text "ICT Foundation" is written in white inside the archway.

ICT Foundation

# Data Compression

# Data Compression

---

- The process of encoding information using fewer bits than an unencoded representation would use.
- Lossless compression
  - Can reproduce the original information completely.
  - Examples: PNG, GIF (Only for less than 256 color images), ZIP
- Lossy compression
  - Cannot reproduce the original information.
  - Example: JPEG, MP3, MPEG (compressing data by thinning out the difficult parts of the human perceives, therefore thinned data can not be reproduced).

# Compression principles (1)

## Run-length Coding

---

- Representing bit string by a number of repeated sequence and the sequence string.
  - Example: 00011000000011111110000011111000 (32 bits)
  - Repeat counts from the right:  $0 \times 3$ ,  $1 \times 2$ ,  $0 \times 7$ ,  $1 \times 7$ ,  $0 \times 5$ ,  $1 \times 5$ ,  $0 \times 3$
  - Repeat counts is less than 7, so they can be represented using 3 bits: 011 010 111 111 101 101 011 (21 bits)
  - Note that if 0 and 1 appear alternately, the value of each part (0 or 1) can be skipped.
  - If one value is repeated more than 8 times, it can be represented by a different bit value (for example, 0 is repeated 12 times, so representing 0 as:  $0 \times 7$ ,  $1 \times 0$ ,  $0 \times 5$ , then encoding: 111 000 101)

# Compression principles (2)

## Huffman Coding

---

- Assign short bit string for big probability of occurrence of one element.
  - To encode 4 types of weather in one week (sunny, rainy, snowy, cloudy), two bits are needed.
    - Weather log in one week: sunny, sunny, sunny, cloudy, cloudy, cloudy, rainy.
    - 2-bit encoding string is 00 00 00 11 11 11 10 (14 bits)
    - The probability of occurrence sunny > cloudy > rainy > snowy  
 → Encode sunny: 1 cloudy: 01 rainy: 001 snowy: 000 → string 1 1 1 01 01 01 001 (12 bits)

Weather	2-bit encoding	Variable-length encoding by the probability of occurrence
sunny	00	1
rainy	01	001
snowy	10	000
cloudy	11	01