DATA STORAGE - DATA COMPRESSION

- Data compression: using compression algorithm to make the data file smaller
- Benefits of data compression
  - Files are smaller
  - Take less data storage
  - Fast transfer
  - Download/upload faster
  - Meet email attachment size restriction
- Drawbacks
  - Compress and decompress takes time
  - Require compatible software to compress and decompress
  - Affect the quality of file
  - It removes some data permanently (lossy)

DATA STORAGE - DATA COMPRESSION

- Lossless compression
  - Using compression algorithm to replace repeating data with smaller data
  - Can restore back to original file, no data is lost
  - Example: zip, rar
  - Run length encoding (RLE)
    - It replaces repeating data with 2 bytes
    - The first byte is the number of occurrences
    - The second byte is the item of information
    - However, it is not good when the text has not many repeating words
    - example

cccmmmmmsssssdddcccccc

would be represented by

3c5m5s3d6c

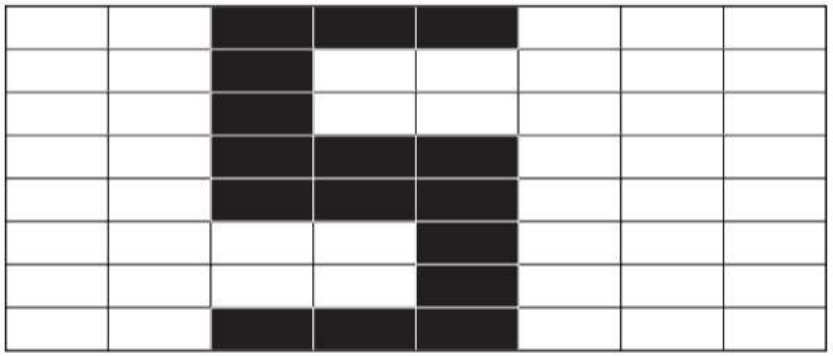

Figure 3.10 A bitmap diagram of the number 5

To demonstrate run-length encoding, the letters 'w' and 'b' are used to represent the white and black pixels instead of 1s and 0s

Code	RLE version	Size of coded version
wwbbbwww	2w3b3w	6
wwbwwwww	2w1b5w	6
wwbwwwww	2w1b5w	6
wwbbbwww	2w3b3w	6
wwbbbwww	2w3b3w	6
wwwwbwww	4w1b3w	6
wwwwbwww	4w1b3w	6
wwbbbwww	2w3b3w	6
64 bytes		**48 bytes**

The file size for this character has been reduced from 64 to 48 bytes. It could be reduced even further if the number 1 is omitted where there is a run of only one. The algorithm could be easily adapted.

DATA STORAGE - DATA COMPRESSION

- Lossless compression
  - It creates a table to store information of the compression
  - The repeated words are identified and indexed and put into the table
  - The position of each word is also stored in the table

red, green, yellow, green, purple, blue, red, purple, blue, yellow, grey, black, pink, red,

Word	red	green	yellow	green	purple	blue	red	purple	blue	yellow	grey	black	pink	red
Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Token	0000	0001	0010		0011	0100					0101	0110	0111	

Word	Token	Position
red	0000	1/7/14
green	0001	2/4
yellow	0010	3/10
purple	0011	5/8
blue	0100	6/9
grey	0101	11
black	0110	12
pink	0111	13

DATA STORAGE - DATA COMPRESSION

- Lossy compression
  - It removes unnecessary data from the file
  - It cannot decompress back to original file
  - Thus, it cannot be used with text file or program coding because if some texts are removed, it will not readable
  - Often use with sound and bit-map image
  - Bit-map image
    - If the image uses 24 bits depth, it can represent 224 or 16 million colors
    - Human eyes cannot see different in the detail
    - The compression algorithm will remove some bits of the bit-depth
    - The quality of image still good
    - Example file format: jpg, gif

Digitally compressing an image makes little difference to our eyes but huge savings in file size

DATA STORAGE - DATA COMPRESSION

- Lossy compression
  - Audio files
    - Uncompressed audio file format, example WAV
    - It takes 30 MB for three minutes of sound
    - But MP3 which lossy compression file format need only 3 MB for three minutes of sound
    - It removes frequency that human cannot hear
    - When two sounds are played together the softer sound will be removed because we hear only the louder sound

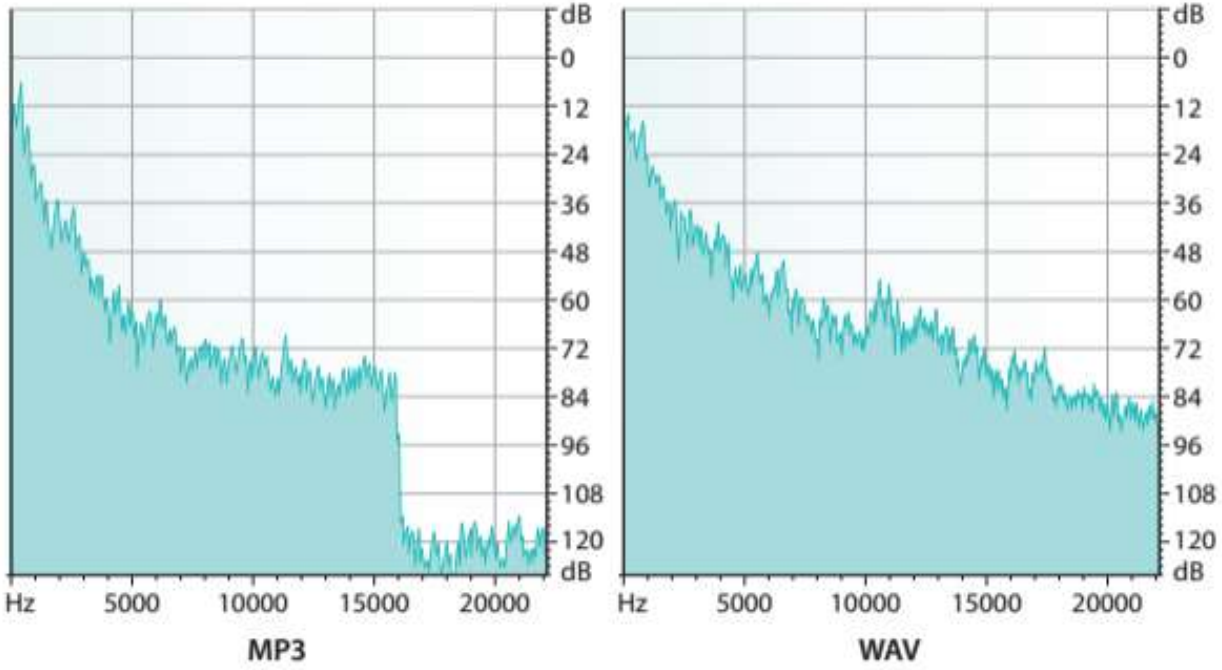

Figure 3.13 The difference in ranges between an MP3 and a WAV sound sample can clearly be seen

DATA STORAGE - DATA COMPRESSION

- Sarah compresses photographs to store them on the USB flash memory drive. It is important that the compression does not affect the quality of the photographs in any way. State which type of compression is the most suitable. Justify your choice

DATA STORAGE - DATA COMPRESSION

- Smith has finished a coding which take space around 100 MB. He would like to upload the code to his server, but he also want to upload faster so he decides to use compression, but he doesn't know to use lossy or lossless compression. Which one you would recommend and why?