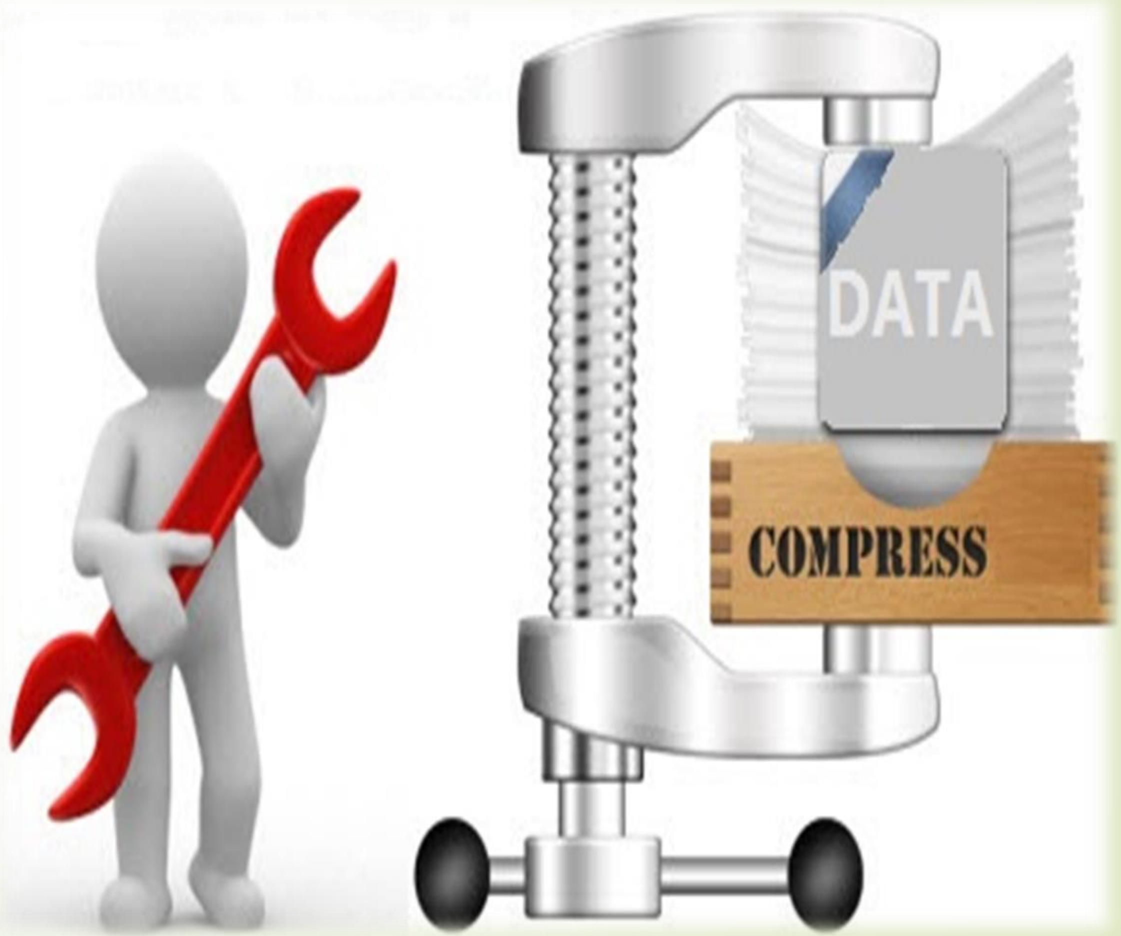
Data compression

Instructor

Ali Kadhum Al-Quraby

د. علي كاظم الغرابي



- In the last decade we have been witnessing a revolution—in the way we communicate
- The major contributors in this revolution are:
  - Internet;
  - The explosive development of mobile communications; and
  - The ever-increasing importance of video communication.
- Data compression is one of the enabling technologies for each of these aspects of the multimedia revolution.

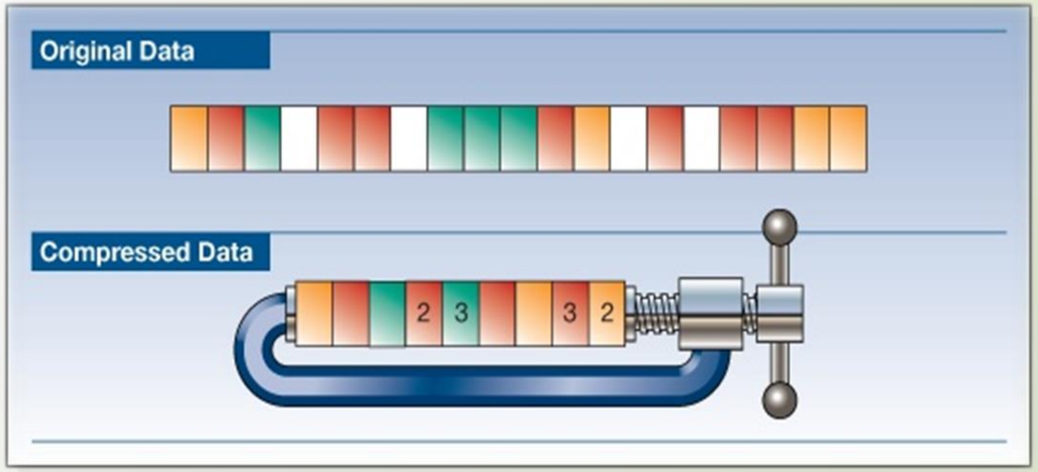
- ***So, what is data compression, and why do we need it?***

Data Compression

- Data Compression **is the process of converting an input data stream (the source stream or the original raw data) into another data stream (the output, the bit stream, or the compressed stream) that has a smaller size.**

- Data Compression **is the art or science of representing *information* in a compact form.**

- ***Compression reduces the size of a file:***
  - **To save space when storing it.**
  - **To save time when transmitting it.**
  - **Most files have lots of redundancy.**



Data Compression

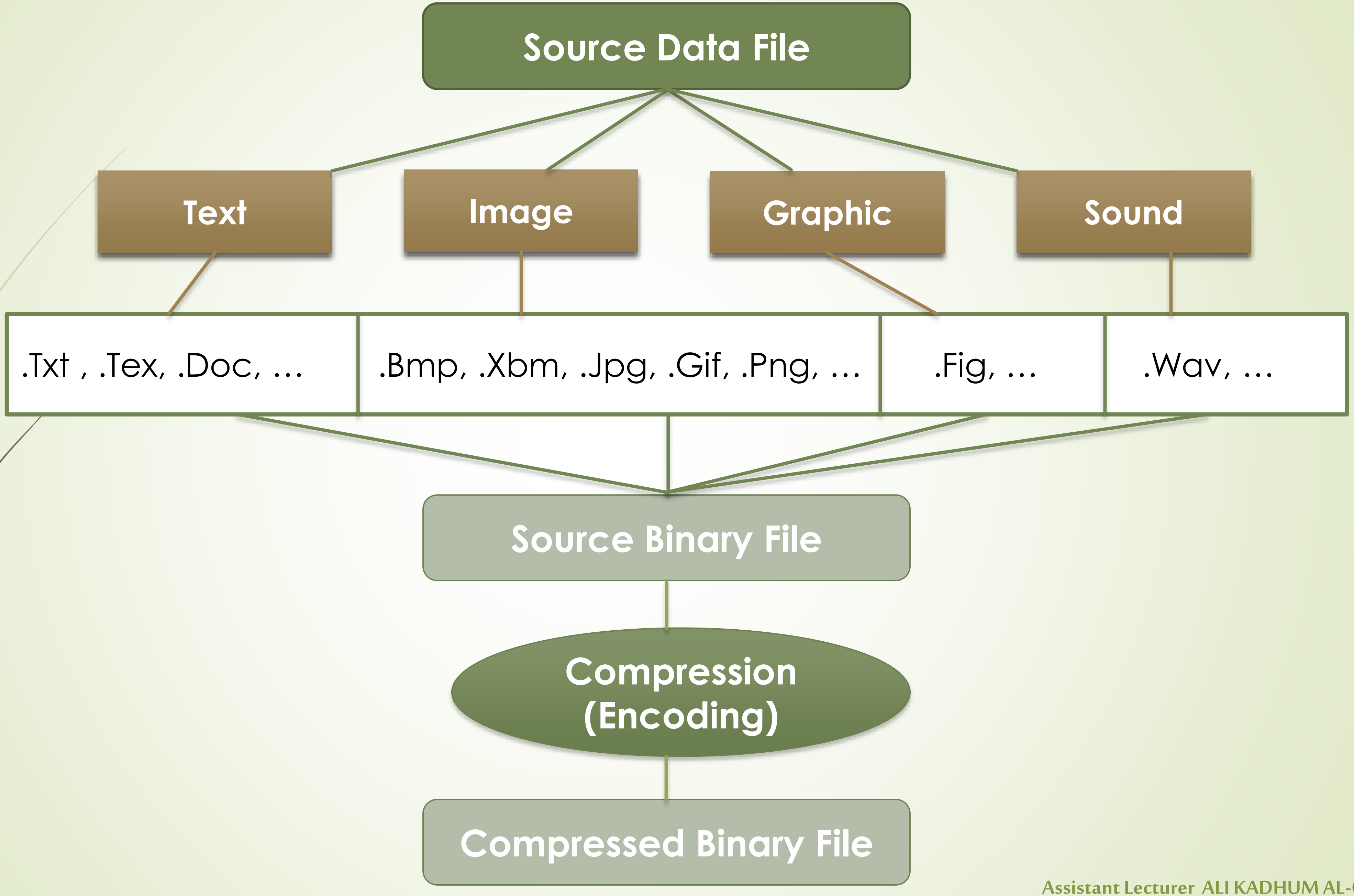
- Information is something that adds to people's knowledge; Information is not visible without some medium being a carrier.

- Data is the logical media often carried by some physical media such as a CD or a communication channel.

- Hence data can be viewed as a basic form of some factual information.

- The data before any compression process are called the **source data**, or the source for short.

- Examples of factual information may be classified broadly as **text**, **audio**, **image** and **video**.


Source Data File
Text
Image
Graphic
Sound
.Txt , .Tex, .Doc, …
.Bmp, .Xbm, .Jpg, .Gif, .Png, …
.Fig, …
.Wav, …
Source Binary File
Compression
(Encoding)
Compressed Binary File

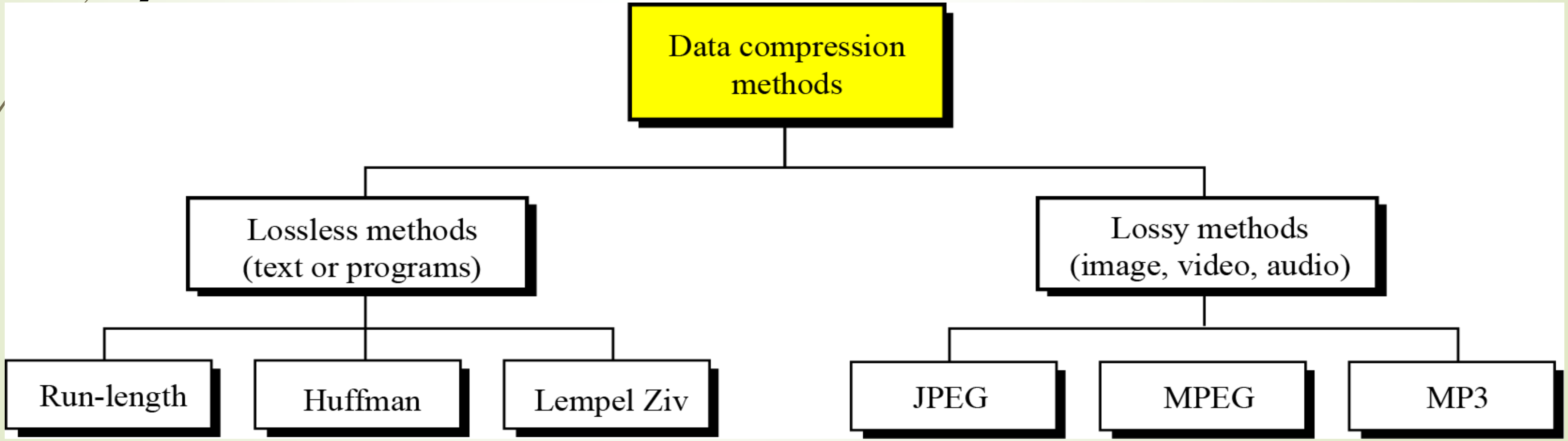
Encoding and Decoding

- ***There are two components in compression process:***
  - Encoding algorithm: **Generate a "compressed" representation C(*M*).**
  - Decoding algorithm: **Reconstruct original message or some approximation M′.**
  - Message: **Binary data M we want to compress.**

Data compression methods

- Data compression **is about storing and sending a smaller number of bits.**

- ***There're two major categories for methods to compress data:***
  - **lossless methods; and**
  - **lossy methods**



Lossless Compression Methods

- **In lossless methods, original data and the data after compression and decompression are *exactly the same.* Because, in these methods, the compression and decompression algorithms are *exact inverses of each other.***

- **Redundant data is removed in compression and added during decompression.**

- **Lossless methods are used when we can't afford to lose any data: legal and medical documents, computer programs.**

- **Consider the sentences "Do not send money" and "Do now send money."**

- **Compression ratio typically no better than 4:1 for lossless compression on many kinds of files.**

Lossless: M=M'

Lossless Compression Methods

- Statistical Techniques
  - **Huffman coding**
  - **Arithmetic coding**
  - **Golomb coding**

- Dictionary techniques
  - **LZW, LZ77**
  - **Sequitur**
  - **Burrows-Wheeler Method**

- Standards - **Morse code, Braille, Unix compress, gzip, zip, bzip, GIF, JBIG, Lossless JPEG**

Lossy Compression Methods

- Used for compressing
  - **Audio**
  - **Video**
  - **Still images, medical images, photographs**

- **Loss of information is acceptable because our eyes and ears cannot distinguish subtle changes, so lossy data is acceptable.**

- **These methods are cheaper, less time and space.**

- **Compression ratios of 10:1 often yield quite high fidelity results.**

Lossy Compression Methods

- Several methods:

  - **JPEG: compress pictures and graphics**
  - **MPEG: compress video**
  - **MP3: compress audio**
  - **Vector Quantization**
  - **Wavelets**
  - **Block transforms**

Measures of Performance

- **A compression algorithm can be evaluated in a number of different ways:**

  - ✓ **Measure the relative complexity of the algorithm,**
  - ✓ **The memory required to implement the algorithm,**
  - ✓ **How fast the algorithm performs on a given machine,**
  - ✓ **The amount of compression, and**
  - ✓ **How closely the reconstruction resembles the original.**

Measures of Performance

- Compression Ratio: **the ratio of the number of bits required to represent the data before compression to the number of bits required to represent the data after compression.**

$$Compression\ Ratio = \frac{size\ of\ the\ output\ stream}{size\ of\ the\ input\ stream}$$

- Note **Values greater than 1 imply an output stream bigger than the input stream (negative compression).**

- **The compression ratio can also be called bpb (bit per bit).**

Measures of Performance

- Saving percentage: **We can also represent the compression ratio as a percentage of the size of the original data.**

$$Saving\ Percentage = (1 - compression\ ratio) \times 100.$$

- Compression factor: **is the inverse of the compression ratio.**

$$Compression\ Factor = \frac{size\ of\ the\ input\ stream}{size\ of\ the\ output\ stream}$$

Note

In this case, values greater than 1 indicate compression and values less than 1 imply expansion.

Measures of Performance

- **Example:** **Suppose storing an image made up of a square array of 256×256 pixels requires 65536 bytes. the image is compressed into 16384 byte.**

$$Compression\ Ratio = \frac{16384}{65536} = \frac{1}{4}$$

$$Saving\ Percentage = \left(1 - \frac{1}{4}\right) \times 100 = 75\%$$

$$Compression\ Factor = \frac{65536}{16384} = 4$$

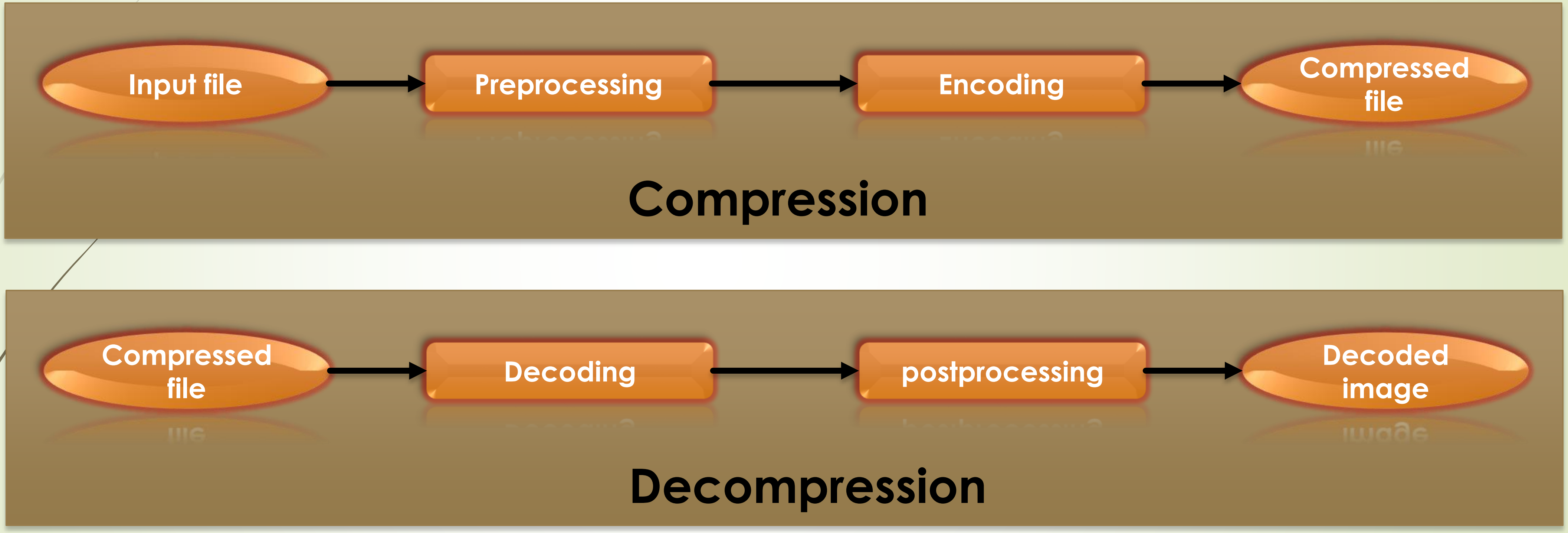
Compression System Model

- The ***Compression System Model*** **consists of two parts:**
  - ✓ The Compressor (Encoding), **and**
  - ✓ The Decompressor (Decoding).

- **The Compressor consists of:**
  - ✓ Preprocessing stage, **and**
  - ✓ Encoding stage.

- **The decompressor consists of a**
  - ✓ Decoding stage, **followed by**
  - ✓ Post processing stage.

Compression System Model

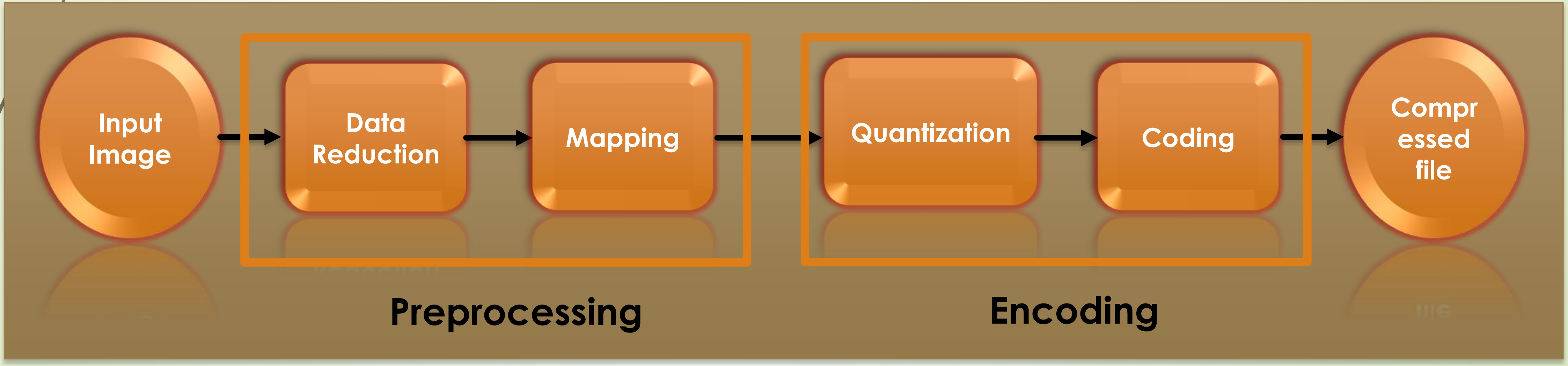


Compression System Model

The Compressor

❖ **The compressor can be further broken down into stages:**

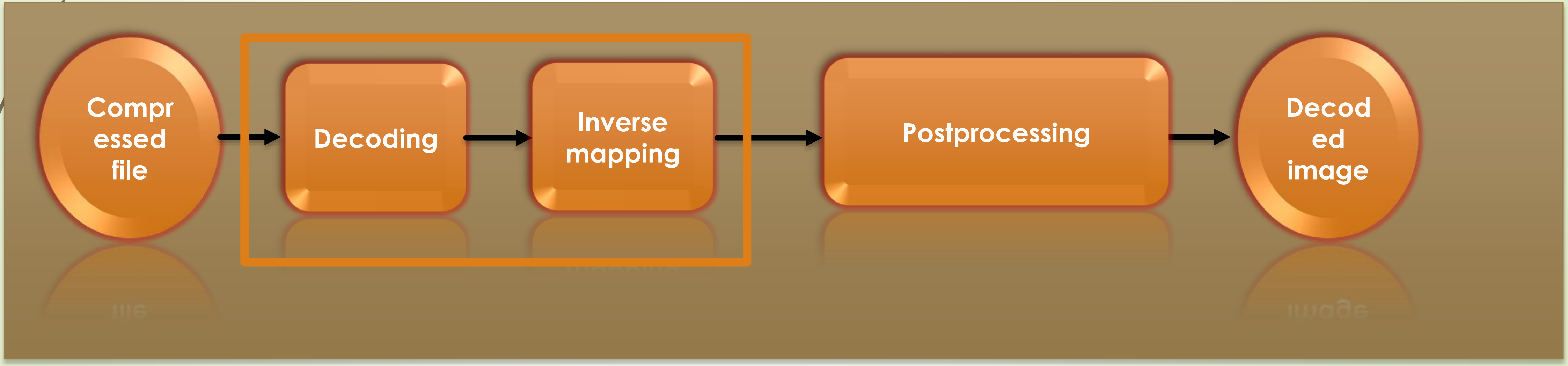
- Preprocessing:-
  - Data reduction.
  - Mapping process.
- Encoding process:-
  - Quantization stage
  - Coding



The Decompressor

❖ **The decompressor can be further broken down into stages:**

- Decoding process:-
  - Decoding stage.
  - Inverse mapping .
- Postprocessing



fundamentals

- The term data compression refers to the process of reducing the amount of data required to represent a given quantity of information
- Data $\neq$ Information
- Various amount of data can be used to represent the same information
- Data might contain elements that provide no relevant information : data redundancy
- Data redundancy is a central issue in image compression. It is not an abstract concept but mathematically quantifiable entity

Data redundancy

- Let n_1 and n_2 denote the number of information carrying units in two data sets that represent the same information
- The relative redundancy RD is define as :

$$R_D = 1 - \frac{1}{C_R}$$

- where C_R, commonly called the **compression ratio**, is

$$C_R = \frac{n_1}{n_2}$$

Data redundancy

- If $n_1 = n_2$, CR=1 and RD=0 no redundancy
- If $n_1 >> n_2$, $CR \rightarrow \infty$ and RD=1 high redundancy
- If $n_1 << n_2$, $CR \rightarrow 0$ and $RD \rightarrow \infty$ undesirable
- A compression ration of 10 (10:1) means that the first data set has 10 information carrying units (say, bits) for every 1 unit in the second (compressed) data set.
- So, RD=0.9, indicating that 90% of its data redundant.
- In Image compression , 3 basic redundancy can be identified
  - ✓ Coding Redundancy
  - ✓ Interpixel Redundancy
  - ✓ Psychovisual Redundancy

Fidelity Criteria

- The **key** in image compression algorithm is to determine the **minimal data** required retaining the necessary information.
- This is achieved by tacking advantage of the **redundancy** that exists in images.
- To determine exactly **what** information is important and to be able to **measure** image fidelity, we need to define an ***image fidelity criterion***.

Note

- ✓ The information required is application specific,
- ✓ With lossless schemes, there is no need for a fidelity criterion.

Fidelity Criteria

- Divided into two classes:

1. **Objective fidelity criteria:**
   - **This fidelity provides us with equations that can be used to measure the amount of error in the reconstructed (decompressed) image.**

- **Commonly used objective measures are:**
  - **The root-mean-square error (RMSE),**
  - **The root-mean-square signal-to-noise ratio (SNRRMS), and**
  - **The peak signal-to-noise ratio (SNRPEAK).**

- **The error between an *original*, uncompressed pixel value and the *reconstructed* (decompressed) pixel value is:**

$$Error(r,c) = g(r,c) - I(r,c)$$

$I(r,c)$: the original image. $g(r,c)$: the decompressed image. r, c: row & column

Fidelity Criteria

- The **total error** in an (N * N) decompressed image is:

$$Total_{Error} = \sum_{r=0}^{N-1} \sum_{c=0}^{N-1} [g(r,c) - I(r,c)]$$

- The root-mean-square error (**RMSE**),

$$RMSE = \sqrt{\frac{1}{N^2} \sum_{r=0}^{N-1} \sum_{c=0}^{N-1} [g(r,c) - I(r,c)]}$$

Notes

- ✓ The **smaller** the value of the error metrics, the **better** the compressed image represents the original images.
- ✓ Alternately, with the **signal-to-noise** (SNR) metrics, a **larger** number implies a **better** image.
- ✓ The **SNR** metrics consider the decompressed image g(r,c) to be "**signal**" and the error to be "**noise**".

Fidelity Criteria

- The root-mean-square signal-to-noise ratio (**SNRRMS**)

$$SNR_{RMS} = \sqrt{\frac{\sum_{r=0}^{N-1}\sum_{c=0}^{N-1}[g(r,c)]^2}{\sum_{r=0}^{N-1}\sum_{c=0}^{N-1}[g(r,c)-I(r,c)]^2}}$$

- The peak signal-to-noise ratio (**SNRPEAK**).

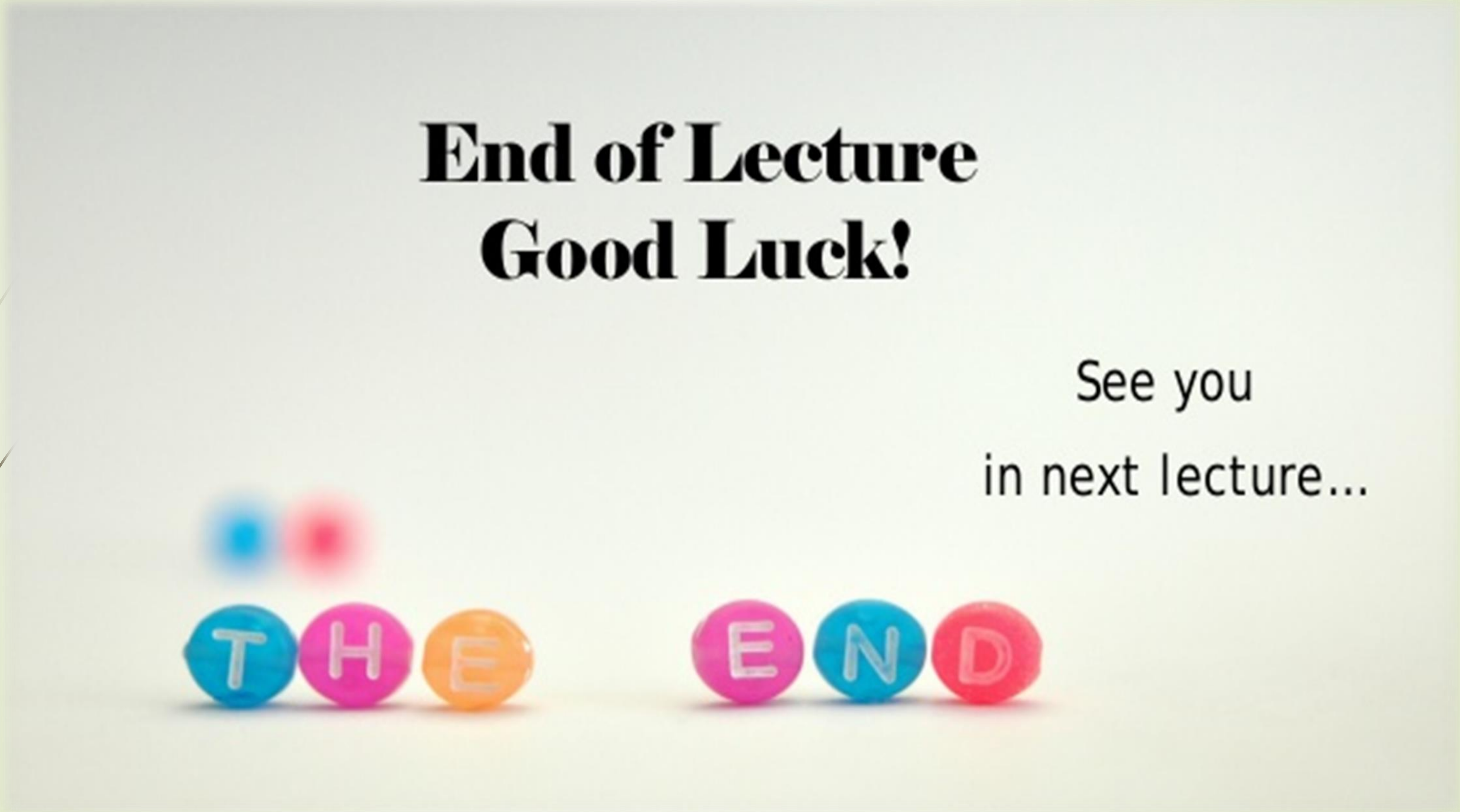
$$SNR_{PEAK} = 10\ \log_{10}\frac{(L-1)^2}{\frac{1}{N^2}\sum_{r=0}^{N-1}\sum_{c=0}^{N-1}[g(r,c)-I(r,c)]^2}$$

✓ **Where L : the number of gray levels (e.g., for 8 bits L =256).**

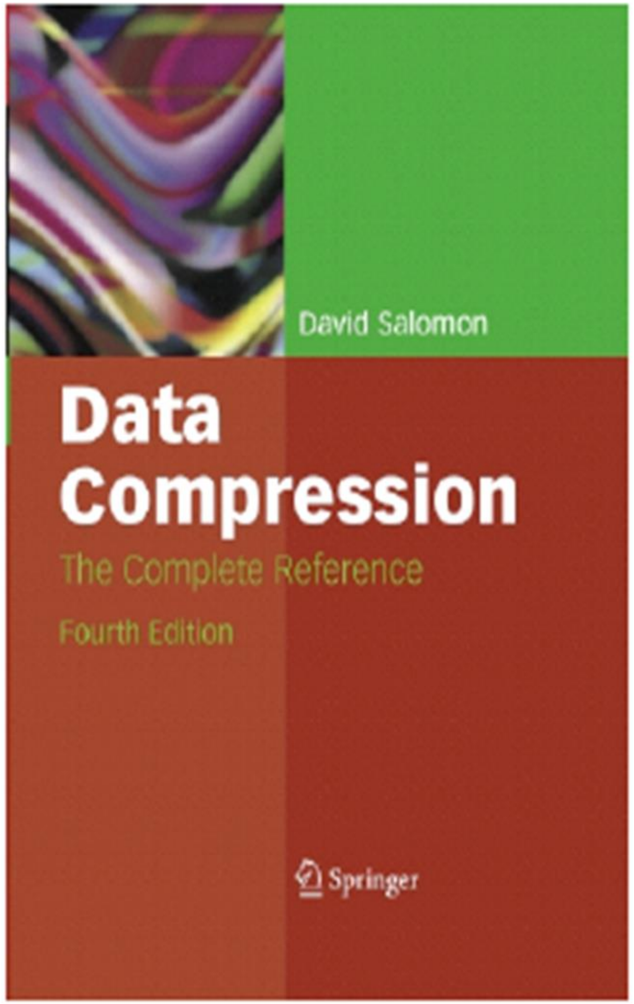
Fidelity Criteria

2. Subjective fidelity criteria (Viewed by Human):

these criteria require the definition of a qualitative scale to assess image quality. This scale can then be used by human test subjects to determine image fidelity. In order to provide unbiased results, evaluation with subjective measures requires careful selection of the test subjects and carefully designed evaluation experiments. The subjective measures are better method for comparison of compression algorithms, if the goal is to achieve high-quality images as defined by visual perception.

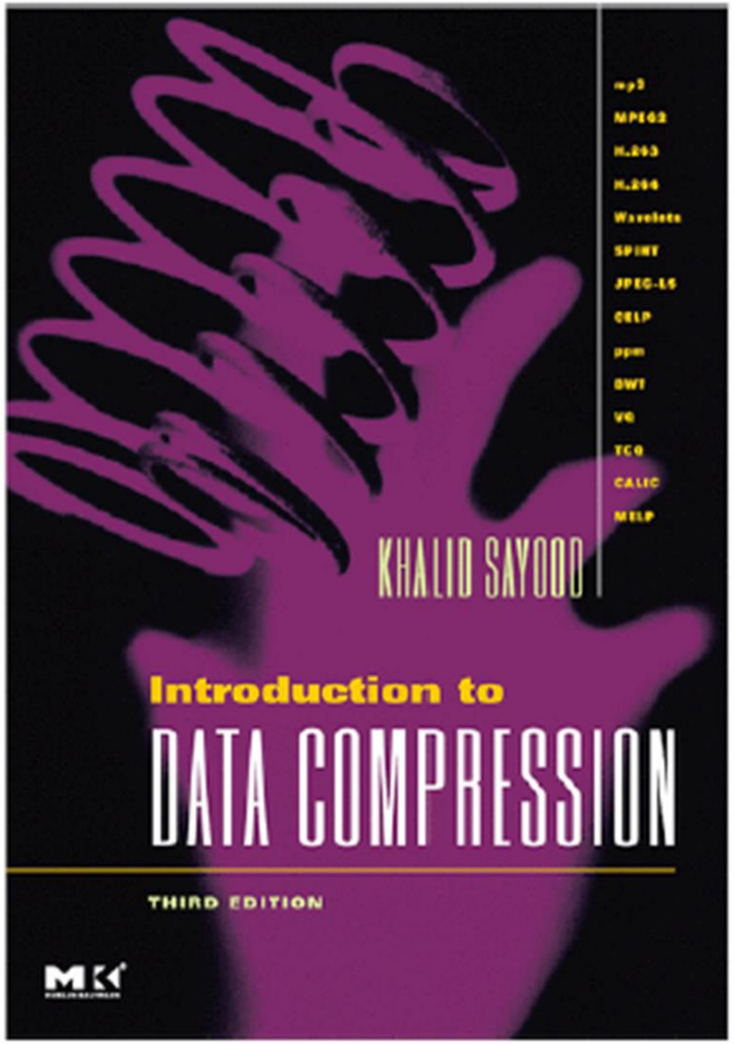

End of Lecture
Good Luck!
See you
in next lecture...
THE END

Data Compression: The Complete Reference (Fourth Edition)
David Salomon
Springer-Verlag London, 2007
(ISBN: 978-1846286025)
Comprehensive coverage of all compression algorithms and formats. Many more than covered in this course!



Introduction to Data Compression (3rd Edition)
Khalid Sayood
Morgan Kaufmann, 2005
(ISBN-13: 978-0126208627)

Excellent coverage of all compression algorithms and formats



Introduction to DATA COMPRESSION (4th Edition)
Khalid Sayood

Morgan Kaufmann, 2012
(ISBN 978-0-12-415796-5)

Excellent coverage of all compression algorithms and formats