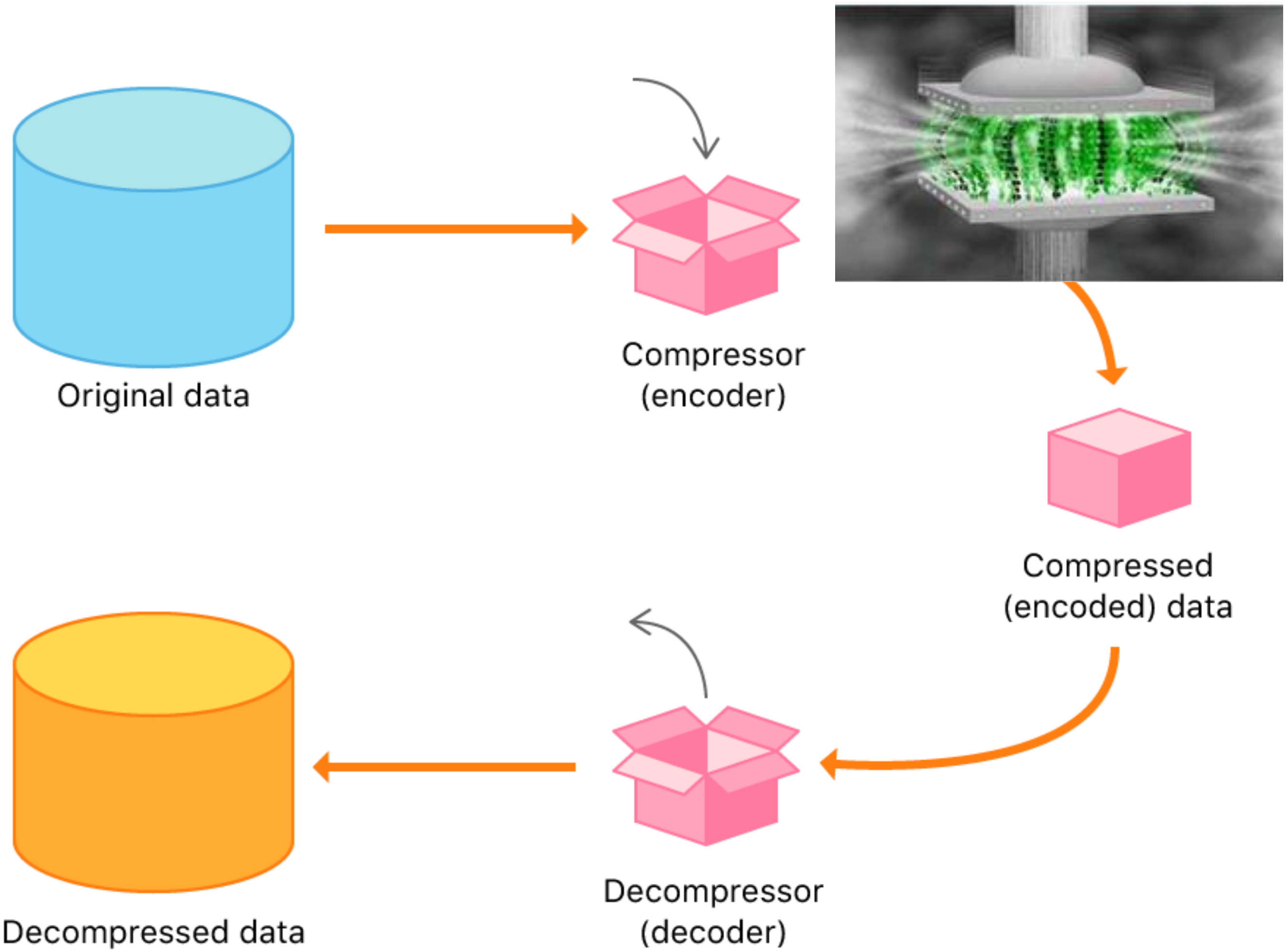
Source Coding

What is Data Compression?



What is data compression?

- Data compression is the **representation** of an **information source** (e.g., data file, speech signal, video signal) as accurately as possible using the **fewest number of bits**

- Compressed data can only be understood if the decoding method is known by the receiver

Source Coding

Lossless
(text, programs)

- **Lossless compression:** the compressed image/video can be decompressed to be identical to the original

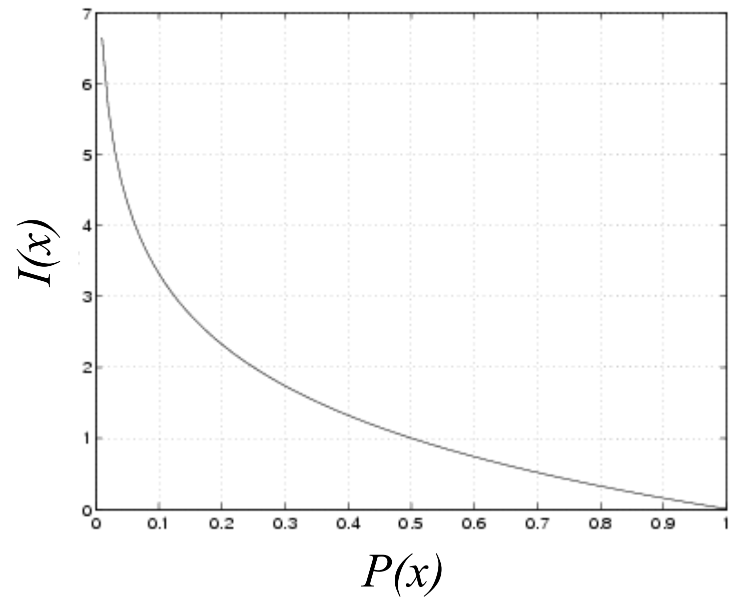
Lossy
(images, videos)

- With **lossy compression**, the image is degraded, mathematically not identical to original. Data compression introduces a **distortion** of the source.

Source Information

- Let X be a discrete random variable taking on values x_1, x_2, … x_j from a finite alphabet A

- Probabilities of occurrence are $p(x_1), p(x_2), \ldots p(x_j)$



- The **information** associated with symbol s_i is defined to be

$$I(x_i) = log_2 \frac{1}{p(x_i)}$$

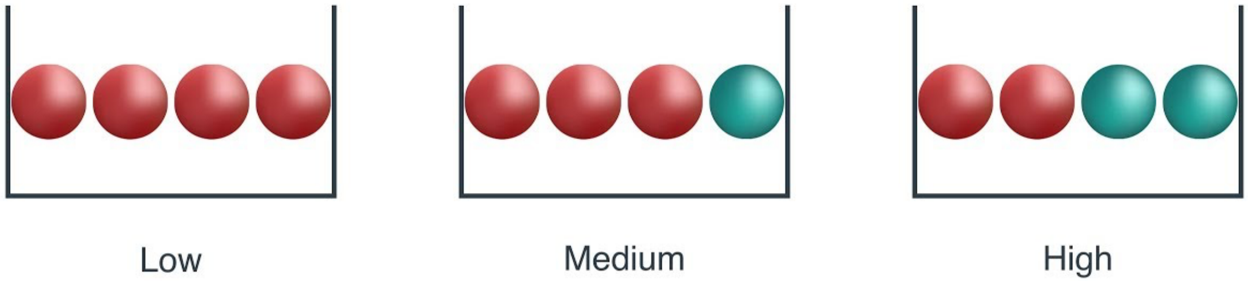
Entropy

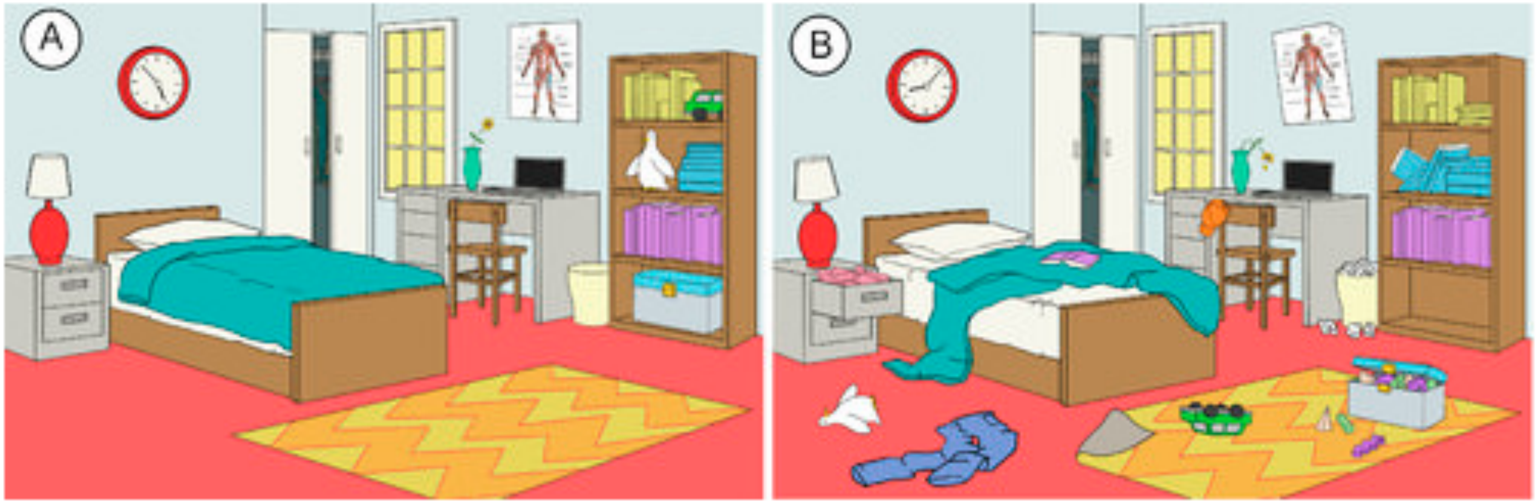
- (First-order) Entropy is a measure of the uncertainty of a random variable.
- **Entropy** is the average information per symbol

$$H(X) = \sum_{i=1}^{J} p(x_i) log_2 \frac{1}{p(x_i)}$$

- Entropy is expressed in bits
- Entropy of a fair coin toss is 1 bit
- Average length of the shortest description of the random variable

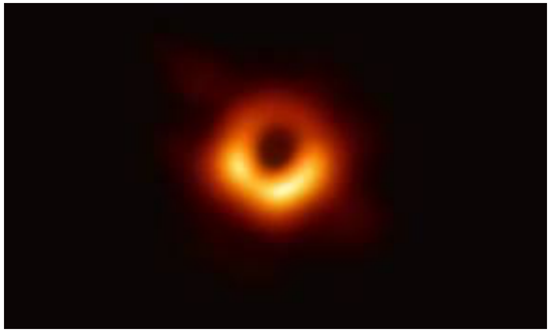
Entropy





Source Information

- Sunrise Time – not that much informative



- Black Hole Pic – highly informative

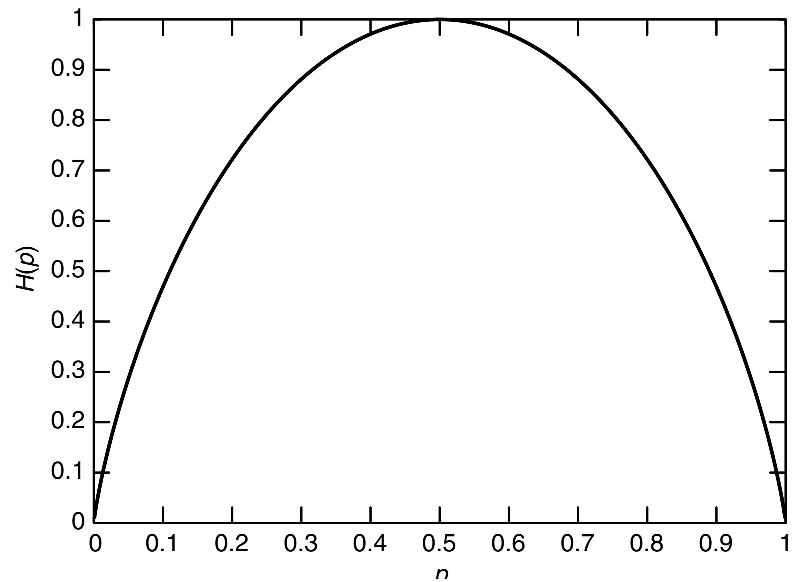
Example 1

Let $$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p \end{cases}$$

Then $$H(X) = -p\, log_2 p - (1-p) log_2 (1-p) \overset{\text{def}}{=} H(p)$$

Entropy of a fair coin toss is 1 bit.

Or in other words p=0.5 → H(p)=1



Example 2

Let

$$X = \begin{cases} a & \text{with probability} \frac{1}{2}, \\ b & \text{with probability} \frac{1}{4}, \\ c & \text{with probability} \frac{1}{8}, \\ d & \text{with probability} \frac{1}{8}. \end{cases}$$

Then, the entropy of X is

$$H(X) = -\frac{1}{2}\, log_2 \frac{1}{2} - \frac{1}{4}\, log_2 \frac{1}{4} - \frac{1}{8}\, log_2 \frac{1}{8} - \frac{1}{8}\, log_2 \frac{1}{8} = \frac{7}{4} \text{ bits}$$

What is the minimum number of questions you need to determine X?

“Is X = a?”

“Is X = b

“Is X = c?”

The resulting expected number of binary questions required is *1.75*.

The minimum expected number of binary questions required to determine X lies between $H(X)$ and $H(X) + 1$.

Thank You