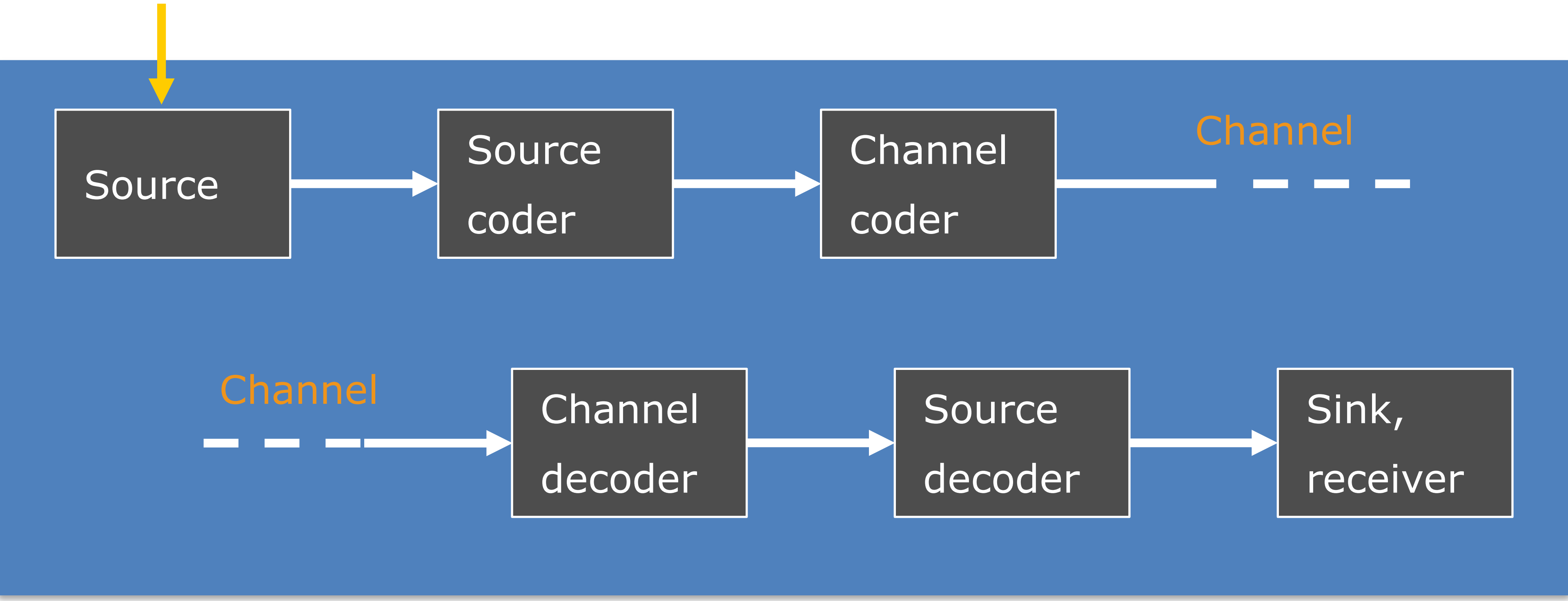
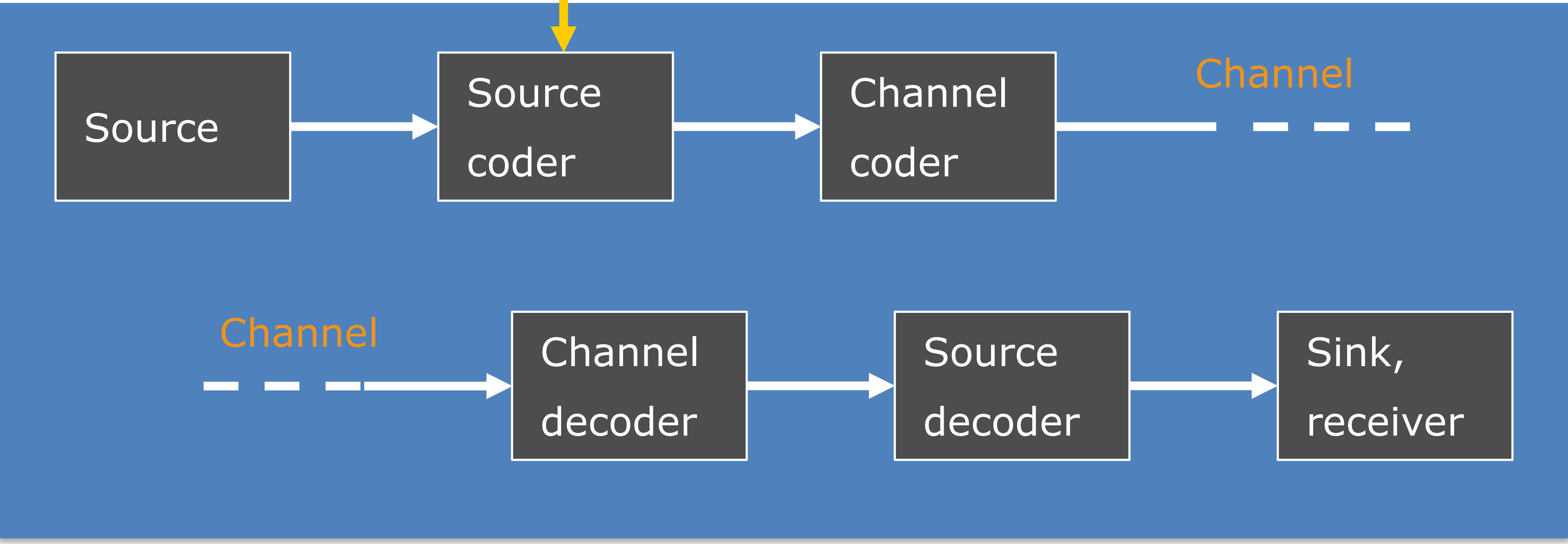
Is Entropy that Important?

YES!

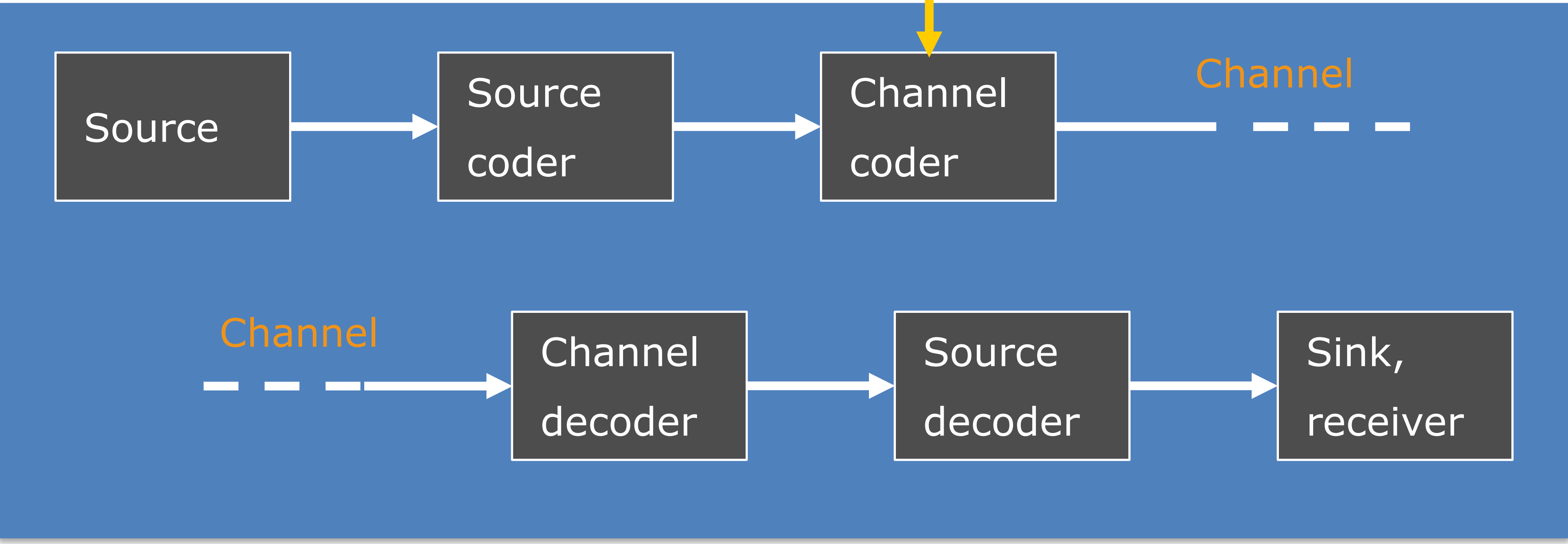
Any source of information

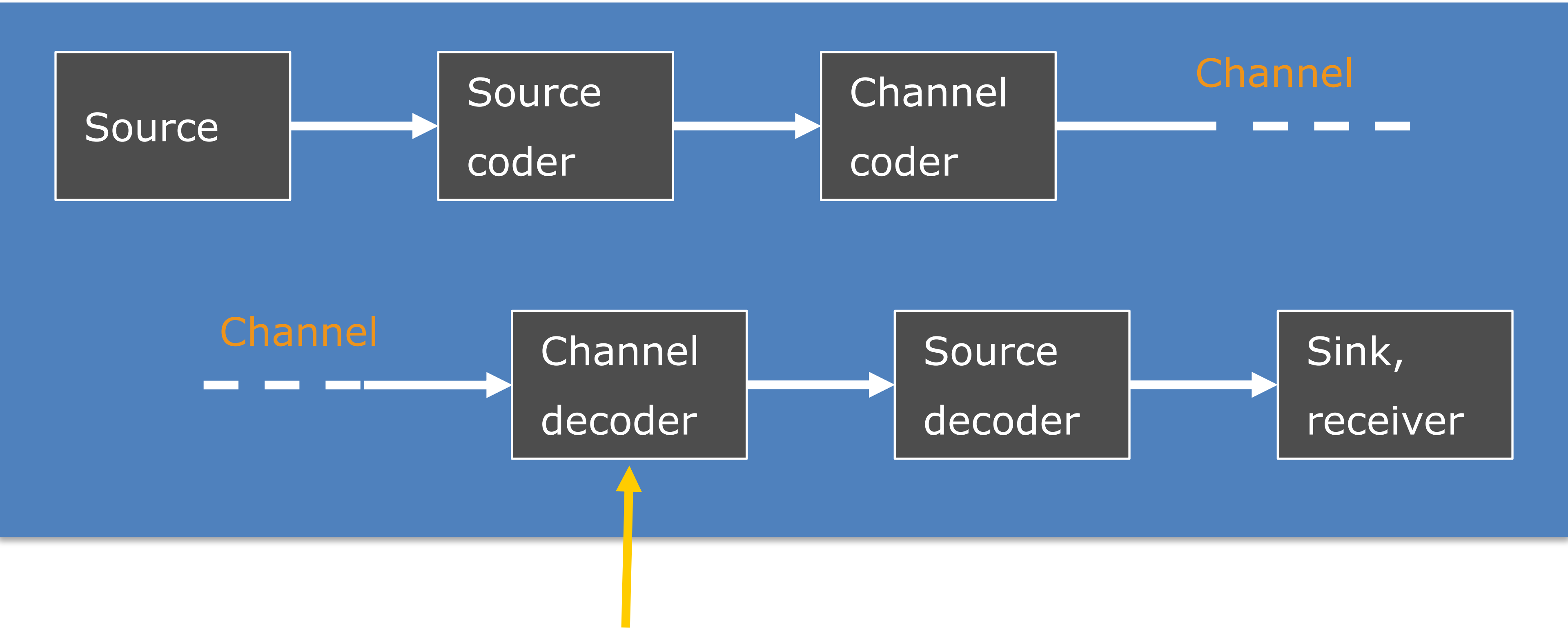


Change to an efficient representation,
i.e., data compression.

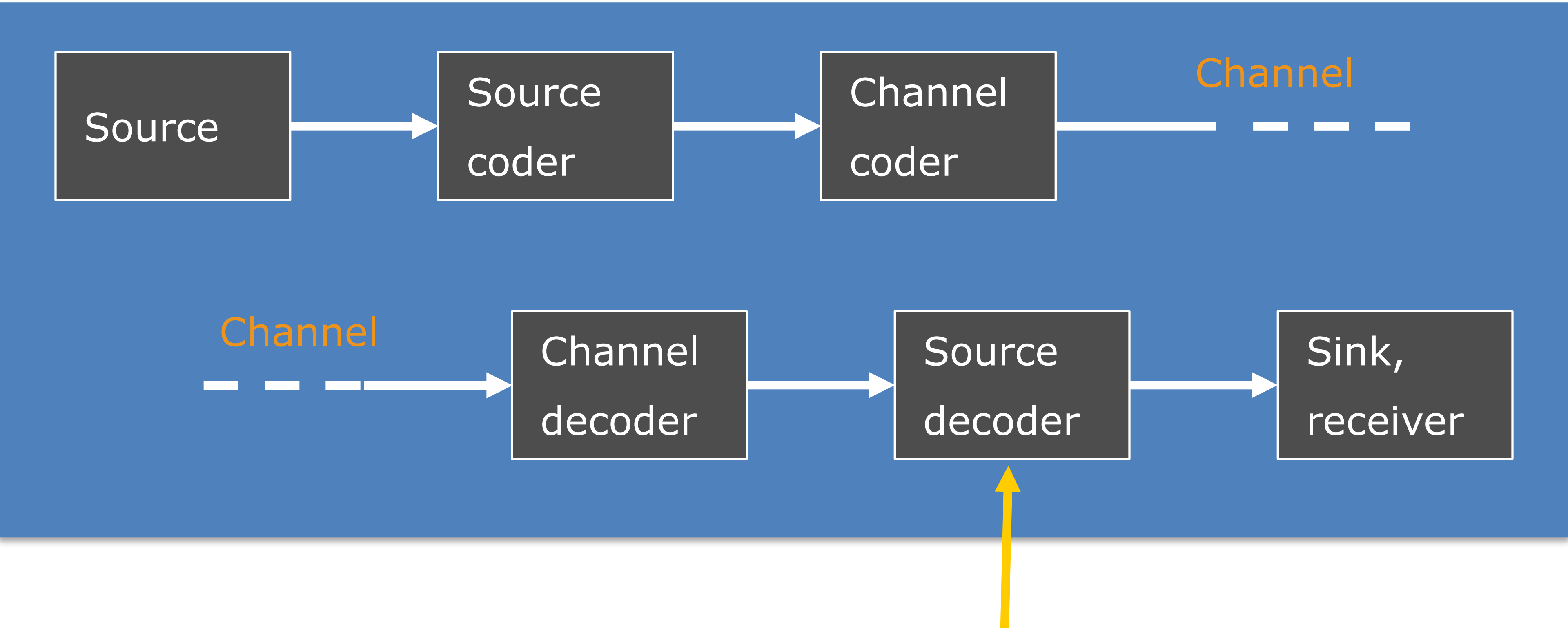


Change to an efficient representation for,
transmission, i.e., error control coding.

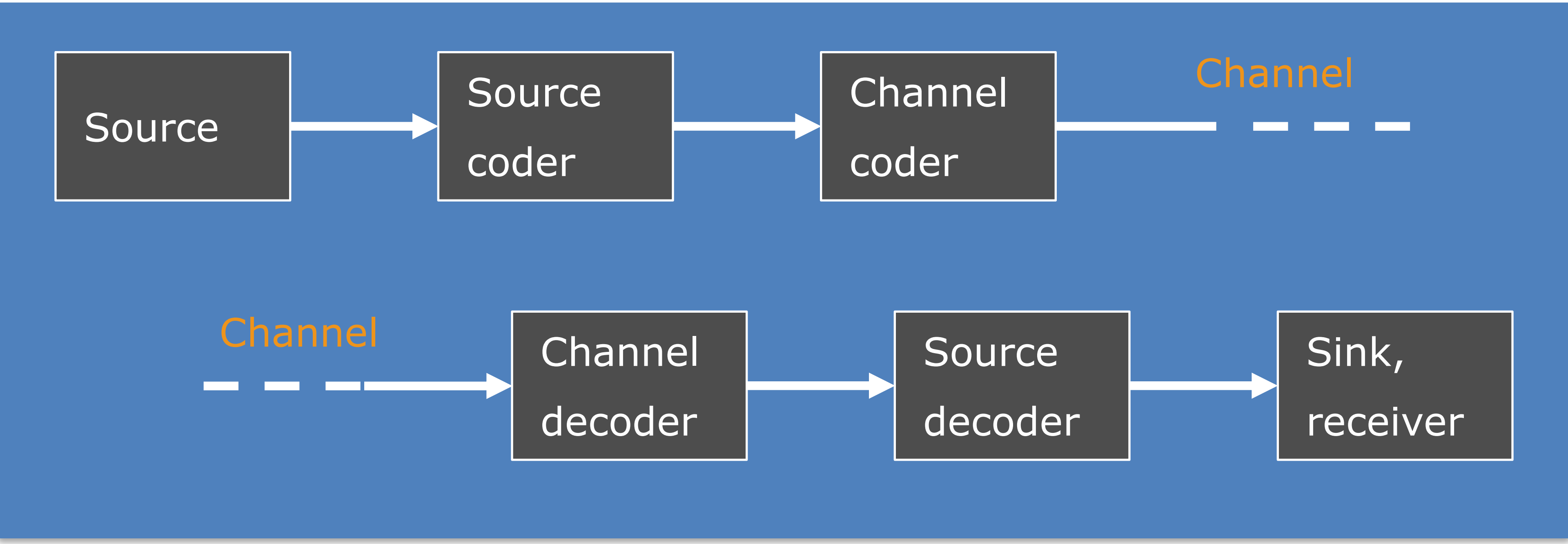




Recover from channel distortion.

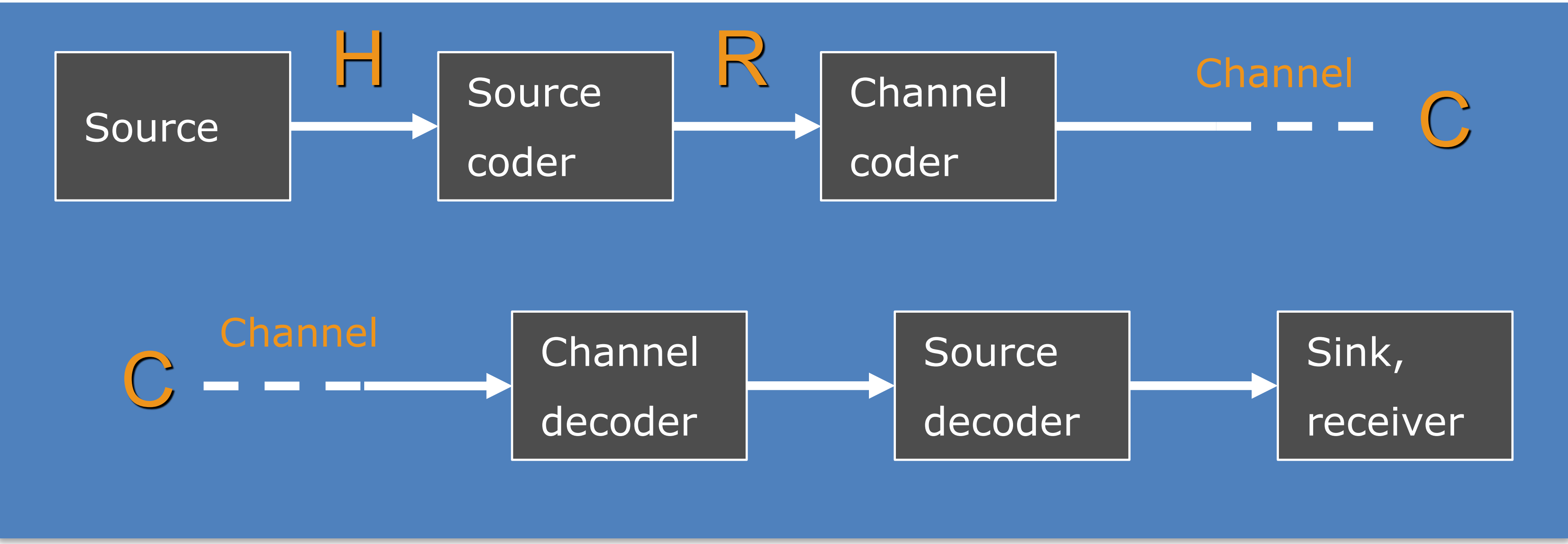


Reconstruct the source



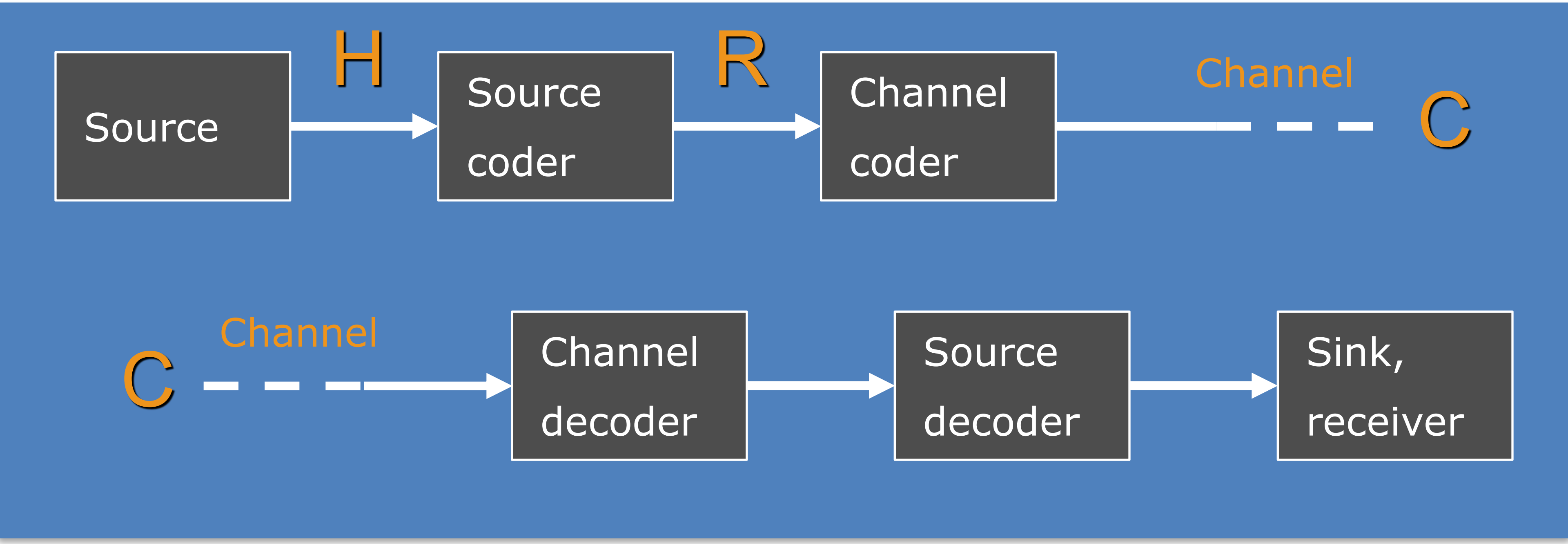
The channel is anything transmitting or storing information – a radio link, a cable, a disk, a CD, a piece of paper, …

Fundamental Entities



H: The information content of the source.
R: Rate from the source coder.
C: Channel capacity.

Fundamental Theorems



Shannon 1: Error-free transmission possible if $R \geq H$ and $C \geq R$.

Source coding theorem (simplified)

Channel coding theorem (simplified)

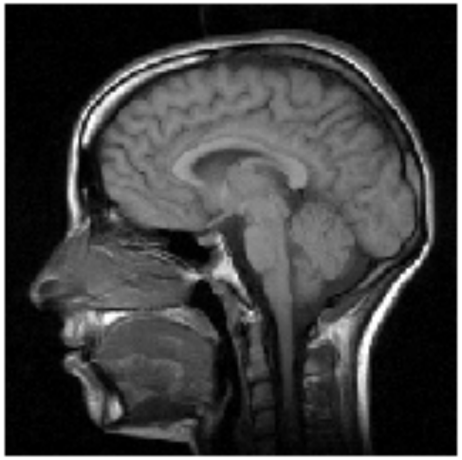
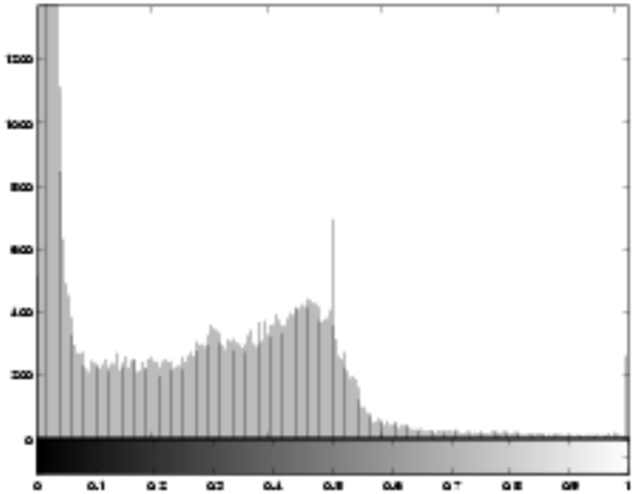
Key concept in data compression

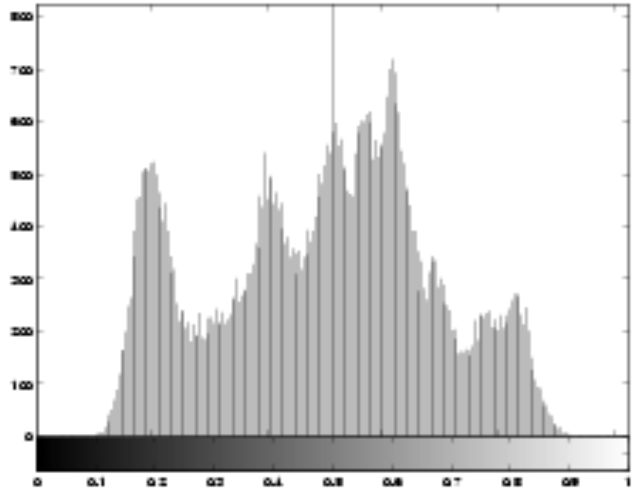
"Data compression can be achieved by assigning **short descriptions** to the **most frequent outcomes of the data source**, and necessarily longer descriptions to the less frequent outcomes." – Thomas & Cover

Why?

Coding Redundancy

- Some things are more common than others:



Short words are common

- Short representations for common things, and long representations for uncommon
- This concept is already familiar in many ways
- The 19 most common words in the English language are:

    the, of, are, I, and, you, a, can, to, he,

    her, that, in, was, is, has, it, him, his

Long words are uncommon

- Very long words are generally very uncommon:
  - Antidisestablishmentarianism
  - Floccipoccinihilipilification
  - Pneumonoultramicroscopicsilicovolcanicconiosis

Morse Code – Common Letters

Letter	Frequency per 1000 in English	International Morse Code
E	130	·
T	93	–
N	78	– ·
R	77	· – ·
I	74	· ·

Morse Code – Uncommon Letters

Letter	Frequency per 1000 English	International Morse Code
X	5	-..-
K	3	-.-
Q	3	--.-
J	2	.---
Z	1	--..

Source Coding

$$X \xrightarrow{\mathcal{X}:\{x_1,\, x_2,\, \dots\, x_j\}} \boxed{\text{Source coding}} \xrightarrow{\mathcal{D}:\{0,\, 1,\, 00,\, 01,\, 10,\, 11,\, 000,\, \dots\}} \mathcal{D}^*$$

- Let X be a random variable taking on values $x_1, x_2, \dots x_j$ from a finite alphabet $\mathcal{X}$
- Let $\mathcal{D}^*$ be the set of finite length strings of symbols from a D-ary alphabet
- For binary, $\mathcal{D}^* = \{0,1,00,01,10,11,000, \dots\}$
- A **source code** C for the random variable X is a mapping from $\mathcal{X}$ to $\mathcal{D}^*$
- $C(x)$ denotes the codeword corresponding to x
- $l(x)$ denotes the length of $C(x)$

Source Coding

EXAMPLE:

C(red) = 00, C(blue) = 11 is a source code for $\mathcal{X}$ = {red, blue} with alphabet $\mathcal{D} = \{0,1\}$

Expected Length

- The expected length of the code is:

$$L(X) = \sum_{x \in \mathcal{X}} p(x) l(x)$$

- ASCII is a **fixed-length** code (FLC)

a → 10000011 and A → 10000001

- **The same number of bits (7) is used to represent each symbol**

Variable Length Coding

- If we want to **reduce the number of bits required to represent different messages**, we should use different numbers of bits to represent different symbols

- Use fewer bits for things that occur more often

- This would be a ***variable length code*** **(VLC)**

Examples of Codes

Input letter	Prob.	Code 1	Code 2	Code 3	Code 4
A	1/2	0	0	1	0
B	1/4	0	1	01	01
C	1/8	1	00	001	011
D	1/8	10	11	000	0111
L(C)		1.125	1.25	1.75	1.875

Problems with these codes

- Code 1: Two input symbols have the same codeword
- Code 2: This problem is fixed. But suppose the decoder receives 00. What was the input? C or AA?
- Codes 3 and 4 look OK. Code 3 is shorter.
- Is that the best we can do?

Thank You