

**Principles of Digital Communications**  
**Prof. Shabbir N. Merchant**  
**Department of Electrical Engineering**  
**Indian Institute of Technology, Bombay**

**Lecture – 01**  
**Introduction to Information Theory**

Welcome back, with this module we will begin our study of Information Theory. The purpose of a communication system is to transmit signals generated by a source of information over a communication channel, but what do you mean by the term information? To address this important issue we need to understand the fundamentals of information theory. Let me take 1 illustrative example to highlight the importance of the role of information theory in a communication system.

So, let us say we have 2 cities A and B, the weather conditions in both the cities change very rapidly during the day. And, it is require to transmit every half an hour or show, the status of the weather condition in both the cities. So, let us see how do we do this?.

(Refer Slide Time: 01:25)

Example:

<p><b>A → C</b></p> <p>sunny - 1/4 - 00</p> <p>cloudy - 1/4 - 01</p> <p>rainy - 1/4 - 10</p> <p>foggy - 1/4 - 11</p> $L_{av} = \frac{2 \times 1}{4} + \frac{2 \times 1}{4} + \frac{2 \times 1}{4} + \frac{2 \times 1}{4}$ $= 2 \text{ bits/message}$	<p>---</p>	<p><b>B → C</b></p> <p>sunny - 1/2 - 0</p> <p>cloudy - 1/8 - 1110</p> <p>rainy - 1/8 - 110</p> <p>smoggy - 1/4 - 10</p> $L_{av} = \frac{1 \times 1}{2} + \frac{4 \times 1}{8} + \frac{3 \times 1}{8} + \frac{2 \times 1}{4}$ $= \frac{16}{8} \text{ bits/message}$
--	------------	--

I have city A and there are 4 weather conditions; sunny, cloudy, rainy, and foggy. And for city B there are 4 weather conditions; sunny, cloudy, rainy, smoggy. The only difference being foggy and smoggy, from the communication perspective both these cities generate 4 messages.

So, let us take the transmission of the weather condition from city A to headquarter weather station located in city C and similarly we will do for transmission of messages from B to C. We, will use binary digits for the transmission I will use the word binit for binary digits. Let me also assume there the probability of occurrence of each of this message is given as one-fourth, one-fourth, one-fourth and one-fourth.

Now, using the binit, I want to label these messages. So, 1 way of doing it would be as follows I give this as 0 0 0 1 1 0 and 1 1, this labels which I have assigned for each of this messages are also known as code words and the set of the code words corresponding to each of the messages is known as a code. Now, for cost effective communication it is desirable to have this binit per message as low as possible.

So, what would be the average length of this labels or code words, that would be given as follows an average is equal to we have 2 binit for the first matches and the probability is one-fourth plus 2 multiplied by one-fourth plus 2 multiplied by one-fourth plus 2 multiplied by one-fourth and this gives us 2 binit per message.

Now, let us take the case of message transmission from city b to city C. And let me assume that the probability of occurrence of this weather is as follows sunny is half cloudy is one-eighth rainy is one-eighth and smoggy is one-fourth. Now, in principle I could have use the same labeling as I did in the case for transmission from A to C and I could have assigned 0 0 0 1 1 0 1 1, but if I do that way it is easy for you to see that the average length, which I will get will turn out to be again 2 binit per message.

But, in this case I can use a simple common sense logic and do the labeling as follows the the logic is that the messages which occur very frequently, I will choose those labels which have smaller length. So, for sunny I will assign a 0 for smoggy I will assign as 1 0 and rainy I will assign 1 1 0 and this I will assign as 1 1 0. Now, it is also important to note that this labeling, which I have done this mapping from the message to the labels is unique in the sense given any of the labels out here I can say what was the message transmitted?

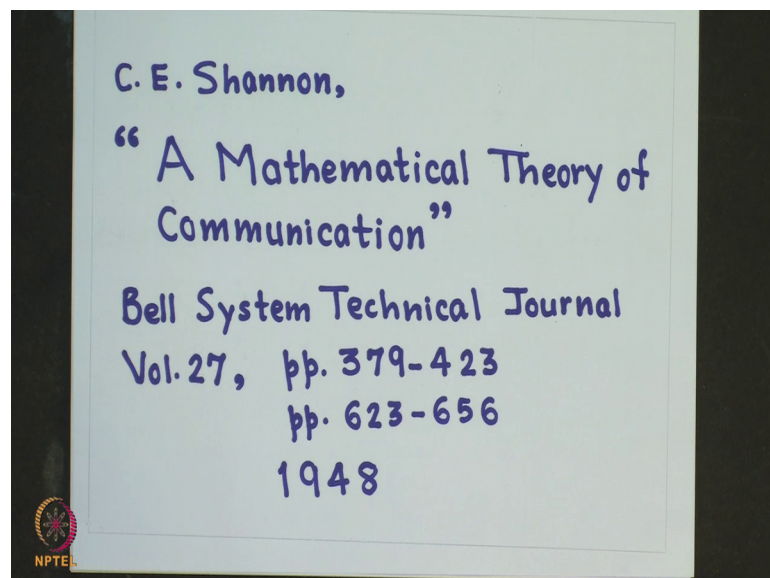
Now, if we calculate the average length in this case it will turn out to be 1 multiplied by half plus 4 multiplied by 1 eighth plus 3 multiplied by 1 eighth and 2 multiplied by one-fourth and you can show that this turns out to be 15 by 8 binit per message. So, what I have shown that using this kind of a clever scheme of labeling, I could save about 1 by

eighth binit compare to this kind of labeling which I did for A to C. So, this is a roughly about 6 percent of saving.

So, the question that comes to my minds are following is it possible for me to use a better labeling scheme and bring down this value still lower, because I desire that for cost effective communication. And, if I can do this what is the lower bound or limit to, which I can reach and even if I can calculate the theoretical lower bound for that how do I achieve that in a practical scenarios. Now, in this case is possible to show that is the better scheme than this and we can get further savings, but we will take up this problem later on as we learn information theory.

So, there are 2 key aspects in performance evaluation of a digital communication system. These are first the efficiency with which information from a given source can be represented and the second the rate at which information can be transmitted reliably over a channel. The fundamental limits on this key aspects of system performance have the roots in information theory information theory was founded by 1 of the scientist Shannon. And, the first white paper in information theory was published way back in 1948.

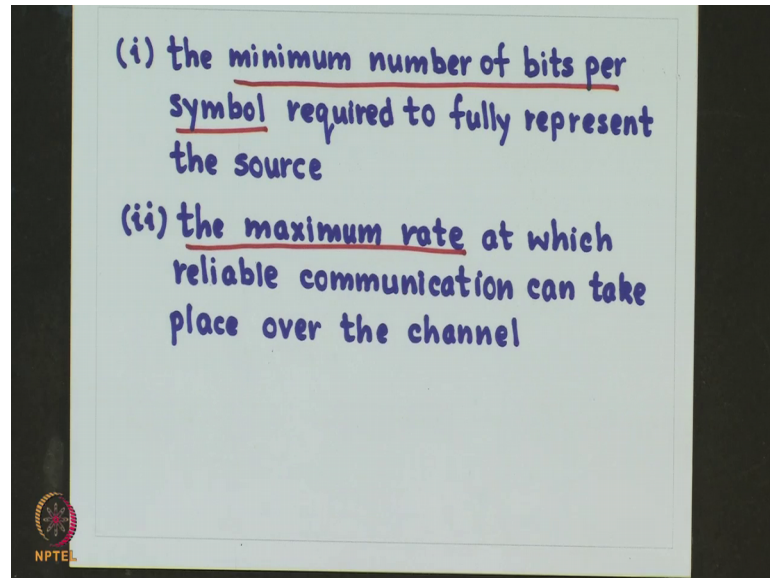
(Refer Slide Time: 08:20)



And, the title of the paper was “A Mathematical Theory of Communication” and it was published in 2 parts in bell system technical journal, volume 27 in the year 1948.

So, given a information source and noisy channel information theory provides the limits on the following 1 the minimum number of bits per symbol required to fully represent the source.

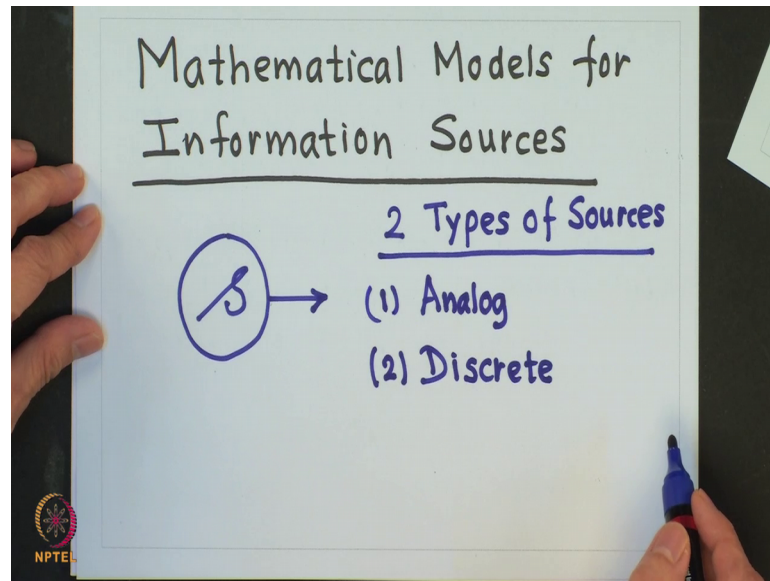
(Refer Slide Time: 08:50)



And, the second is the maximum rate at which reliable communication can take place over the channel. In the context of communication information theory deals with mathematical modeling and analysis of a communication system rather than physical sources and physical channels.

Let us look at the mathematical models for information source.

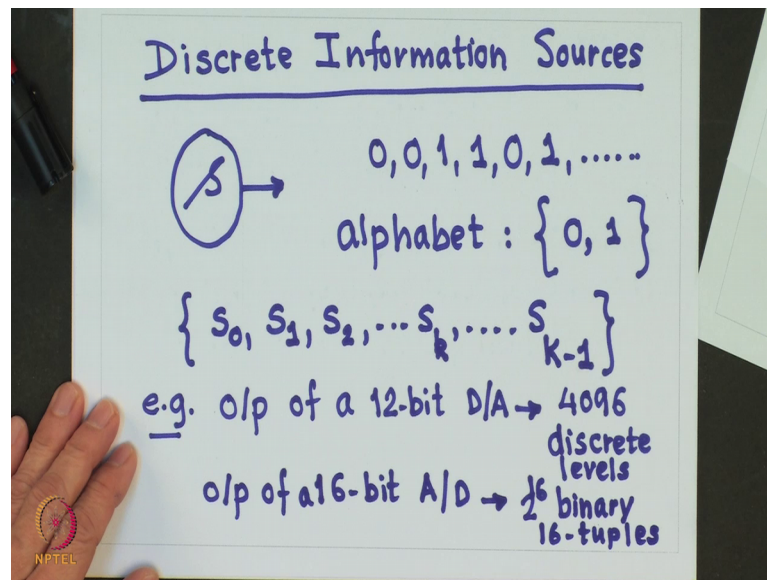
(Refer Slide Time: 09:30)



So, any information source let me depicted here by capital S produces an output, that is random that is the source output is characterized in statistical terms. If the source output was known or deterministic I need not transmit and there are 2 types of sources; 1 is analog source an example of that would be the audio and speech signals. And the second is the discrete source an example of that would be the output from computer and storage devices.

We will study both these sources analog and discrete and postulate mathematical models for each type of this source. So, let us start with discrete information sources.

(Refer Slide Time: 11:20)

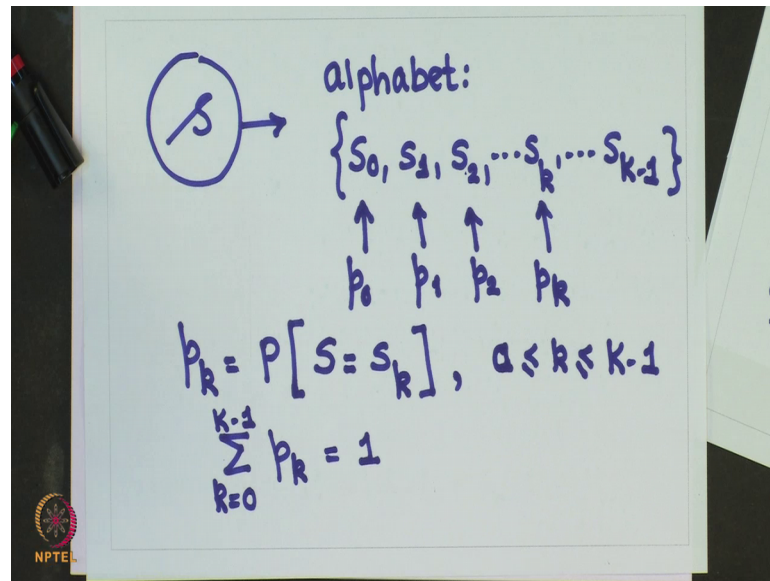


So, the simplest type of discrete source generates or emits a sequence of symbols or letters selected from a finite source alphabet for example, a binary sequence of the form 0 0 1 1 0 1 the alphabet for this is the set 0 and 1. So, these are the 2 symbols or the letters of this alphabet more generally a discrete information source would have an alphabet of  $k$  possible letters or symbols for example,  $S_0 S_1 S_2$ .

. So, this alphabet corresponds to another discrete information show, where there are  $k$  letters or symbols. So, when this alphabet contains a finite number of symbols in this case  $k$  symbols the source is said to be a finite discrete source. An example would be output of a 12 bit D to A converter. So, this outputs 1 of 4 0 9 6 discrete levels, another example would be output of a 16 bit A to D converter.

In this output is 1 of 2 raised to 16 binary 16 tuples. So, to construct a mathematical model for a finite discrete source we assume that each letter in the alphabet like what I have shown you here;  $S_0 S_1 S_2$  up to  $S_{K-1}$  has a given probability  $p_k$  of occurrence.

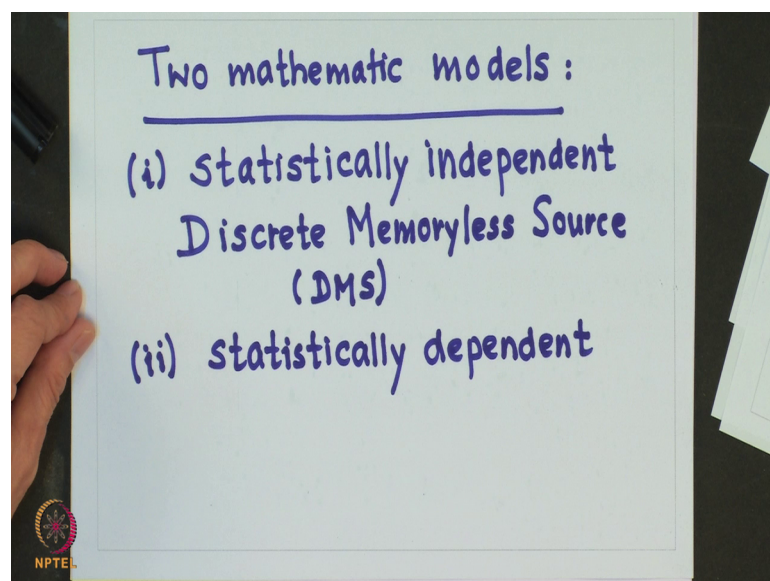
(Refer Slide Time: 15:05)



That is  $p_k$  is equal to probability of the output of the source or that probability of the event is equal to  $S_k$  for  $k$  equal to 0 to capital  $K$  minus 1, and ; obviously, the summation of  $p_k$  overall capital  $K$  should be equal to 1.

So, a finite discrete source is defined by the list of symbols or letters this list is known as alphabet, and the probability assigned to these symbols or letters there are two mathematical models of discrete source

(Refer Slide Time: 16:54)

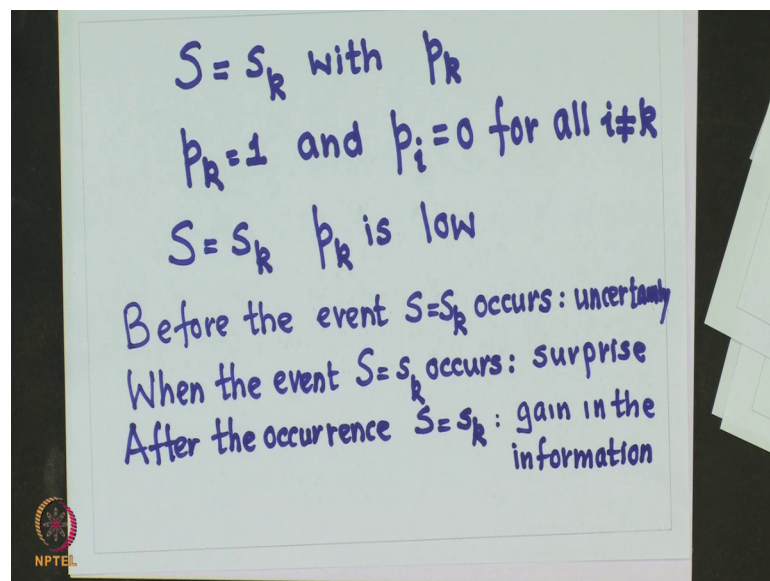


One the output sequence from the source is statistically independent. In such a case we say that the source is a memoryless source. So, we have what is known as discrete memoryless source and we could have another discrete source, where the output is statistically dependent the example of this could be English text.

So, given a symbol or letter  $q$  in English text most probably the next symbol or letter is going to be  $u$ . So, the occurrence of  $u$  basically is influenced by the occurrence of  $q$  or if I have text English text as  $th$ , then the next occurrence of the letter is more likely to be  $e$   $a$   $i$   $o$   $u$   $r$  what I want to convey is that there is a statistical dependency among the symbols or letters, which are being generated or emitted from the source during a particular signaling interval. So, the next question is can we measure how much information is generated by a source which has been defined mathematically as we have done now.

And, we will see that the idea of information is closely related to that of uncertainty or surprise.

(Refer Slide Time: 19:55)



So, let us consider the event  $S$  is equal to  $S_k$  with the probability of its occurrence  $p_k$ . Now, if  $p_k$  is equal to 1 and  $p_i$  is equal to 0 for all  $i$  not equal to  $k$ , then whenever this event occurs there is no surprise and there is no information gain, but from the same source if the probabilities of these symbols or letters were unequal. Then whenever  $s$  is equal to  $S_k$  occurs and let us assume that,  $p_k$  is low then in this case there will be more surprise or more information. When this event occurs compared to the occurrence of any

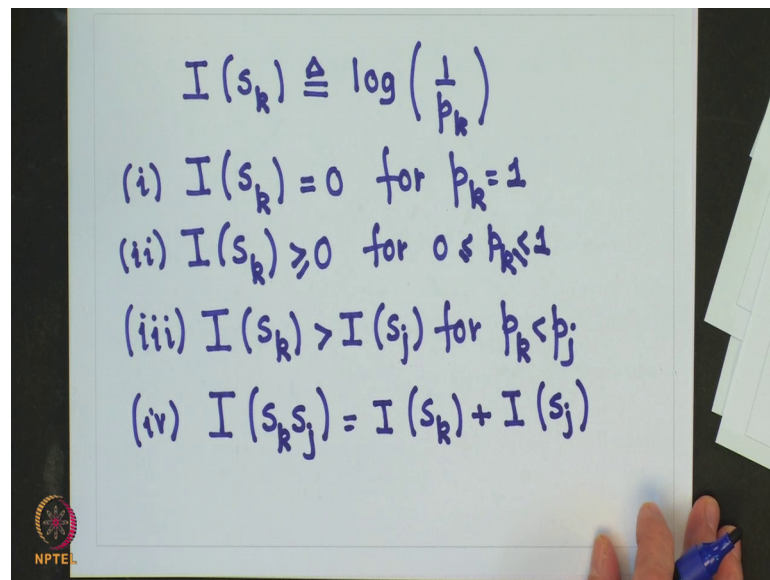


other event where the probability of that event is higher than the probability of this event  $S$  is equal to  $S_k$ .

So, what it means that the concept of uncertainty surprise information are all related. So, before the event  $S$  is equal to  $S_k$  occurs there is amount of uncertainty. When the event  $S$  is equal to  $S_k$  occurs there is a amount of surprise. And, after the occurrence of this event  $S$  is equal to  $S_k$ , this uncertainty which we had before the occurrence of event gets resolved or we can say there is gain in the information.

So, the amount of information is related to the inverse of the probability of occurrence of that event. So, 1 measure for defining information could be as follows.

(Refer Slide Time: 23:24)



The image shows a whiteboard with handwritten mathematical definitions for information  $I(s_k)$ . The definitions are:

$$I(s_k) \triangleq \log\left(\frac{1}{p_k}\right)$$

- (i)  $I(s_k) = 0$  for  $p_k = 1$
- (ii)  $I(s_k) \geq 0$  for  $0 < p_k \leq 1$
- (iii)  $I(s_k) > I(s_j)$  for  $p_k < p_j$
- (iv)  $I(s_k s_j) = I(s_k) + I(s_j)$

A hand holding a blue marker is visible at the bottom right of the whiteboard. In the bottom left corner, there is a small logo for NPTEL.

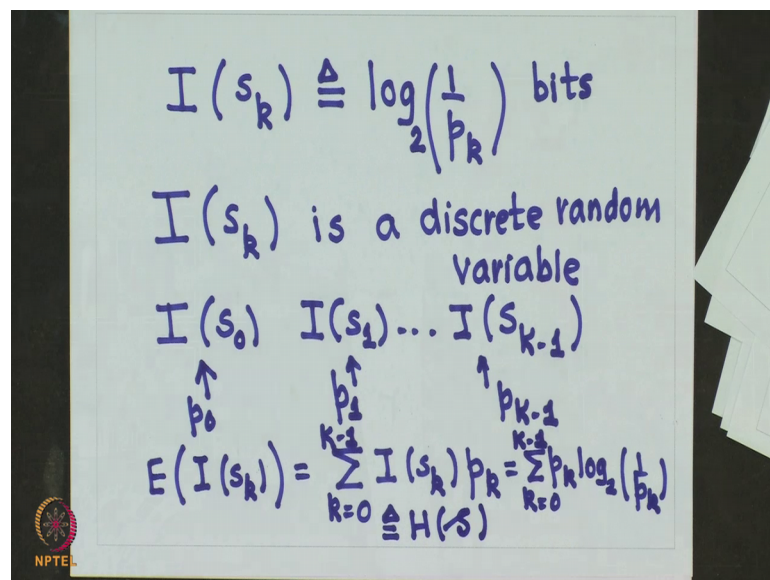
. So, whenever the event  $S$  is equal to  $S_k$  occurs I could say the amount of information gained after observing this event  $S$  is equal to a  $S_k$  is by definition equal to logarithmic of 1 by  $p_k$ . Now, this definition of low battening 1 by  $p_k$  has important properties that are intuitively satisfying.

Let us take the first property information I get when the event  $S$  is equal to  $S_k$  will be equal to 0 for  $p_k$  equal to 1, because in this case we are certain of the outcome of an event even before it occurs and therefore, no information is gained and this metric of the measure and this measure of the information also conveys the same. The second is

basically this is greater than equal to 0 for 0 for  $p_k$  greater than 0 and less than equal to 1.

So, what it means that it provides some or no information, but no loss of information. The third is  $I(S_k)$  is greater than  $I(S_j)$  for  $p_k$  less than  $p_j$ . So, what it means that less probable event the more information be gain when it occurs. And finally, because of this measure if you have 2 independent events  $S_k$  and  $S_j$ , then the joined information which I get from this is equal to the information which I get from the occurrence of event  $s$  is equal to  $S_k$  and from the occurrence of the event  $S$  is equal to  $S_j$  which again intuitively is satisfied.

(Refer Slide Time: 26:12)



We have this measure for information whenever event  $s$  is equal to  $S_k$  occurs as by definition  $\log$  of  $1$  by  $p_k$ , what is the base of this logarithm? It is more natural to use to the base 2 and when you use to the base 2 then the units of measurement of information is given in terms of bits. So, whenever event  $s$  is equal to  $S_k$  occurs I get the information, which is given by this quantity what is the significance of the base 2 is as follows if I have a source, which emits only 2 messages or 2 symbols or 2 letters for example, tossing of a coin the outputs are head or tail.

And, if you assume that probabilities of occurrence of both of this are equal, then whenever head or tail occurs the information I get from it is equal to 1 bit. Now, we know that if I have an output which is binary like tossing of a coin, then to physically

represent this output in terms of binary data I require 1 binary digit. So, this measure of information correlates with our physical understanding of a process fine. Now, the next question is given this definition for information measured for the occurrence of a particular event, how do I generalize to the case where I want to calculate the average information, which is being generated and emitted from the source.

Now, it is very easy to do that if you realize that  $I(S_k)$  is a discrete random variable. So, in this case we will have this  $I(S_k)$  taking different values as follows is  $0, 1, \dots, K-1$  up to is capital  $K$  minus 1. And each of this will occur with probabilities given by  $p_0, p_1, \dots, p_{K-1}$  and  $p_{K-1}$ .

So, we can find out the average of this value as follows this would be equal to  $I(S_k)$  multiplied by the probability of this random variable, which is equal to  $p_k$ . So, if I do this I get this as and so, this is the average information in terms of bits per source letter or symbol and by definition this is known as the entropy of a discrete memoryless source.

So, in this module we have studied what is a discrete information source, how to measure information, and the definition of entropy for a discrete memoryless source.

Thank you.