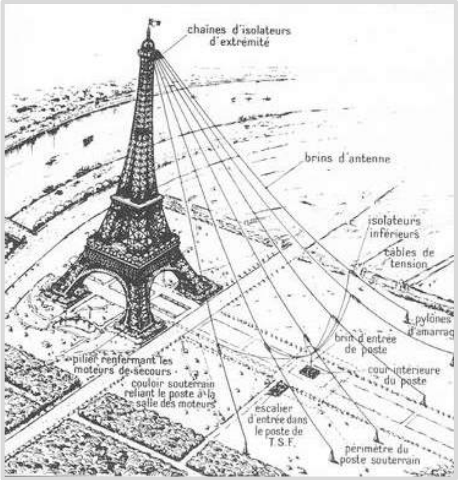
Second year of engineering degree

Source coding

Yves LOUËT

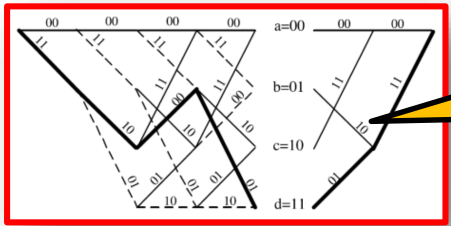
Full Professor with CentraleSupélec
IETR Lab. (UMR CNRS 6164)



The steps of information transmission

Analog source

… 0010101000 …

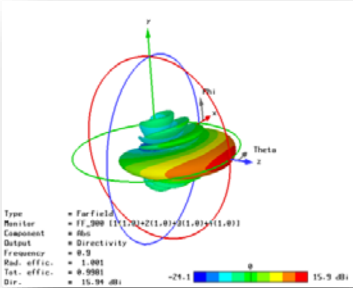


Information processing
Coding and modulation

Toward HW

RF front end

Transmission



Key word : **DIGITAL**

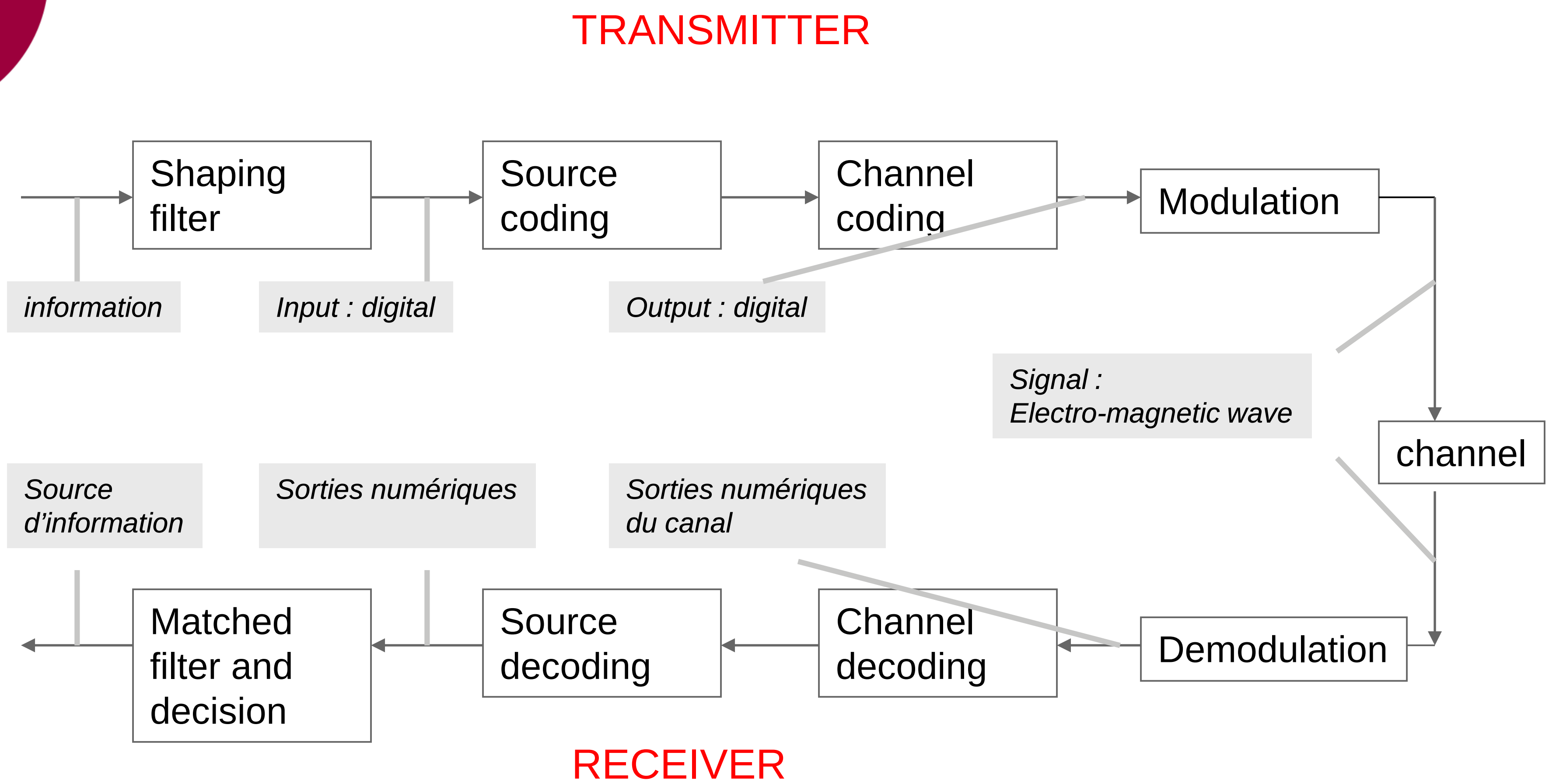
Objectives

Describe all the digital processings related to a transmission

- **source coding : compression**
- **channel coding : protection**
- **modulation**

- Establish the performance of digital transmissions

The basic transmission chain



4

Transmitter

Analog data (image, sound, …)

Sampling quantization

01001000110100001001110100 10 …

compresssion

Source coding (JPEG, MPEG, MP3, …)

111000 …

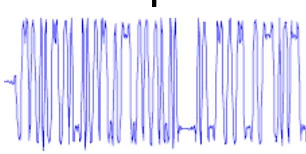
protection

Channel coding (RS, convolutif, …)

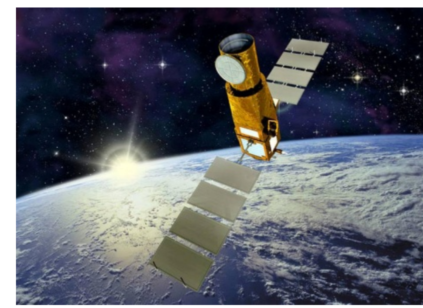
0101110001010 …

Filter and shaping

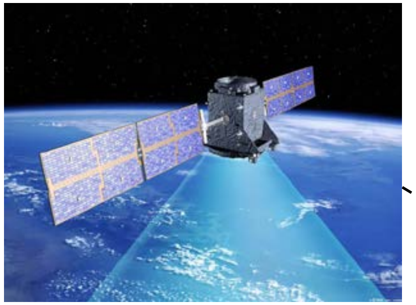
Modulation (to RF front end)



RF



Receiver



Convertor — *demodulation*

0101110001010 …

Error correction — *Channel decoding*

111000 …

decompression — *Source decoding*

01001000110100001001110100 10 …

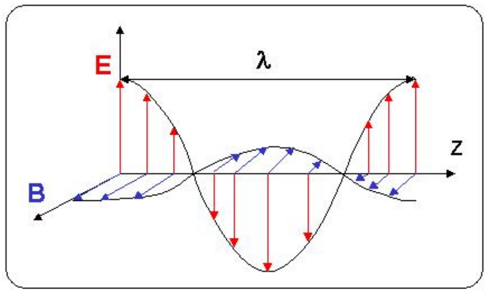
display

The origines of radio

Lodge, Popov (1895)

EM waves transmitter

Hertz (1887)



Maxwell (1864)

Branly (1890)

ME wave detector

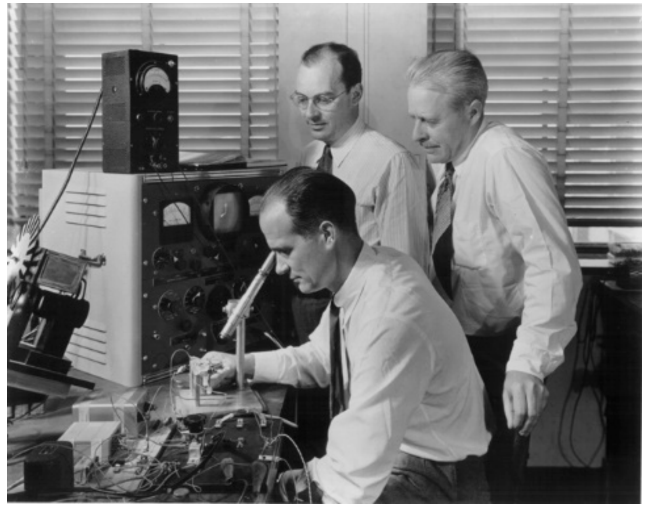
First radio transmission (in morse) : Marconi (1895), Popov (1896), …
First transmission between EU and USA (in morse) : Marconi (1901), Nobel Prize1909
First radio transmission (voice and musik) : Fessenden (1906)
The Eiffel tower used as a radio transmitter (and receiver) (Ferrié) : 1904

7

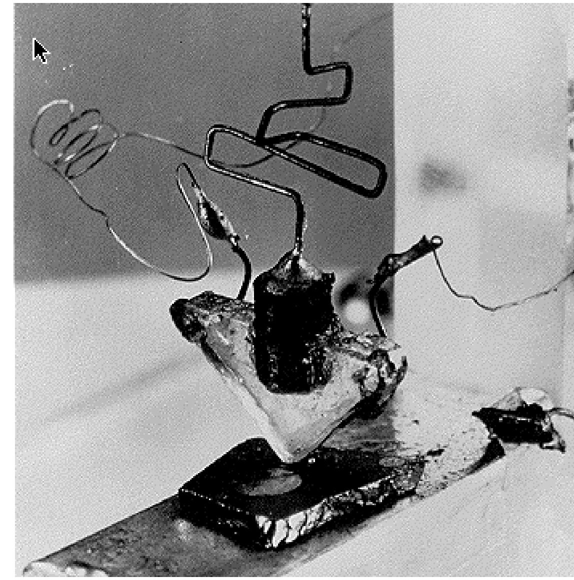
Digital communications were born in

1947-48

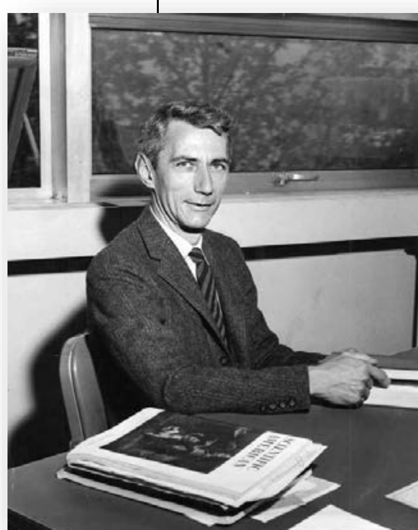
The transistor



Bardeen – Brattain - Shockley



Information theory

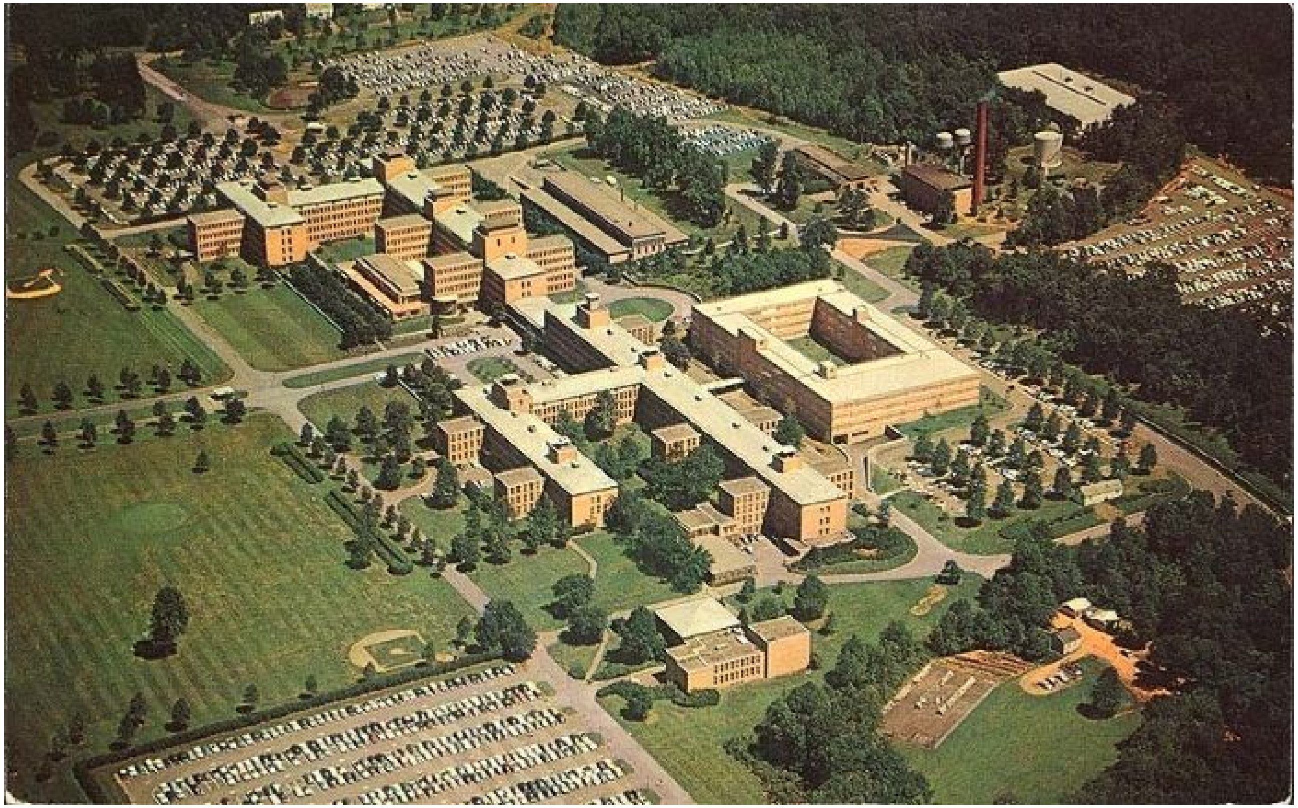


Reprinted with corrections from *The Bell System Technical Journal,*
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

A Mathematical Theory of Communication

By C. E. SHANNON

Bell labs (New Jersey)



- Shannon (th. de l'information)
- Smith (abaque)
- Carson (frequency modulation)
- Tuckey (Fast Fourier Transform)
- Ritchie (C langage)
- Nyquist (digital filtering)
- Bardain – Bradeen – Shockley (transistor)
- Moore law
- Bjarne Stroustrup (Langage C++)
- Boyle & Smith (LCD sensor)
- ….

The transistor

Information theory

+ antenna, networks, computers science, ...

Transmissions are everywhere

12

Key parameters

Carrier frequency

Bandwidth

Propagation conditions

Emitted power

Complexity

13

Information theory: compression and protection

- Association of **compression** (source coding) and **protection** (channel coding)

According to the information theory:

1. Redundancy is necessary to secure and recover the information
2. Over redundancy could be reduced to mitigate the data rate

=> Compression: reduce the over redundancy to a lower bound (defined as the entropy)

=> Once the data is reduced to its lowest rate (ie the entropy), this is time to protect it : this is the objective of channel coding

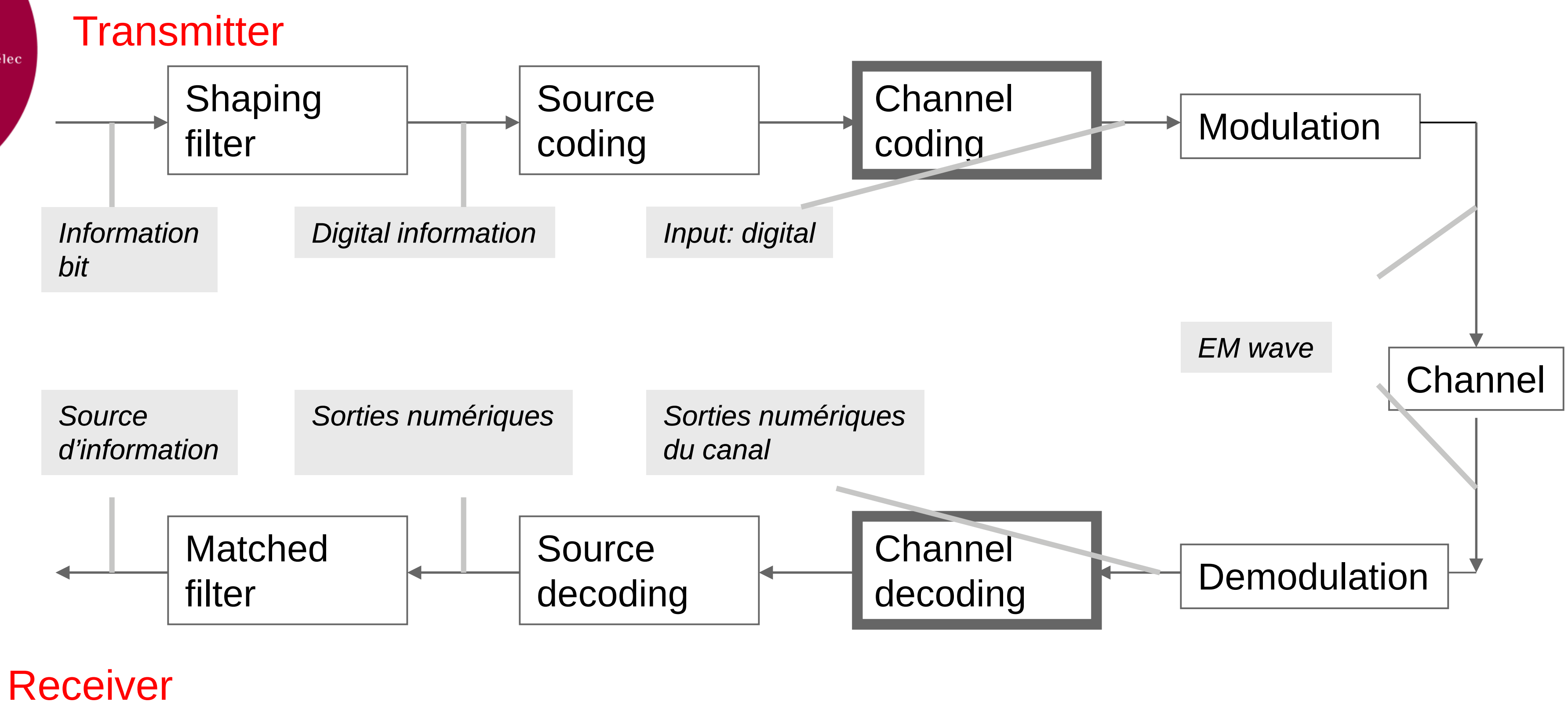
14

Digital communications

Some words on channel coding

Protect the information

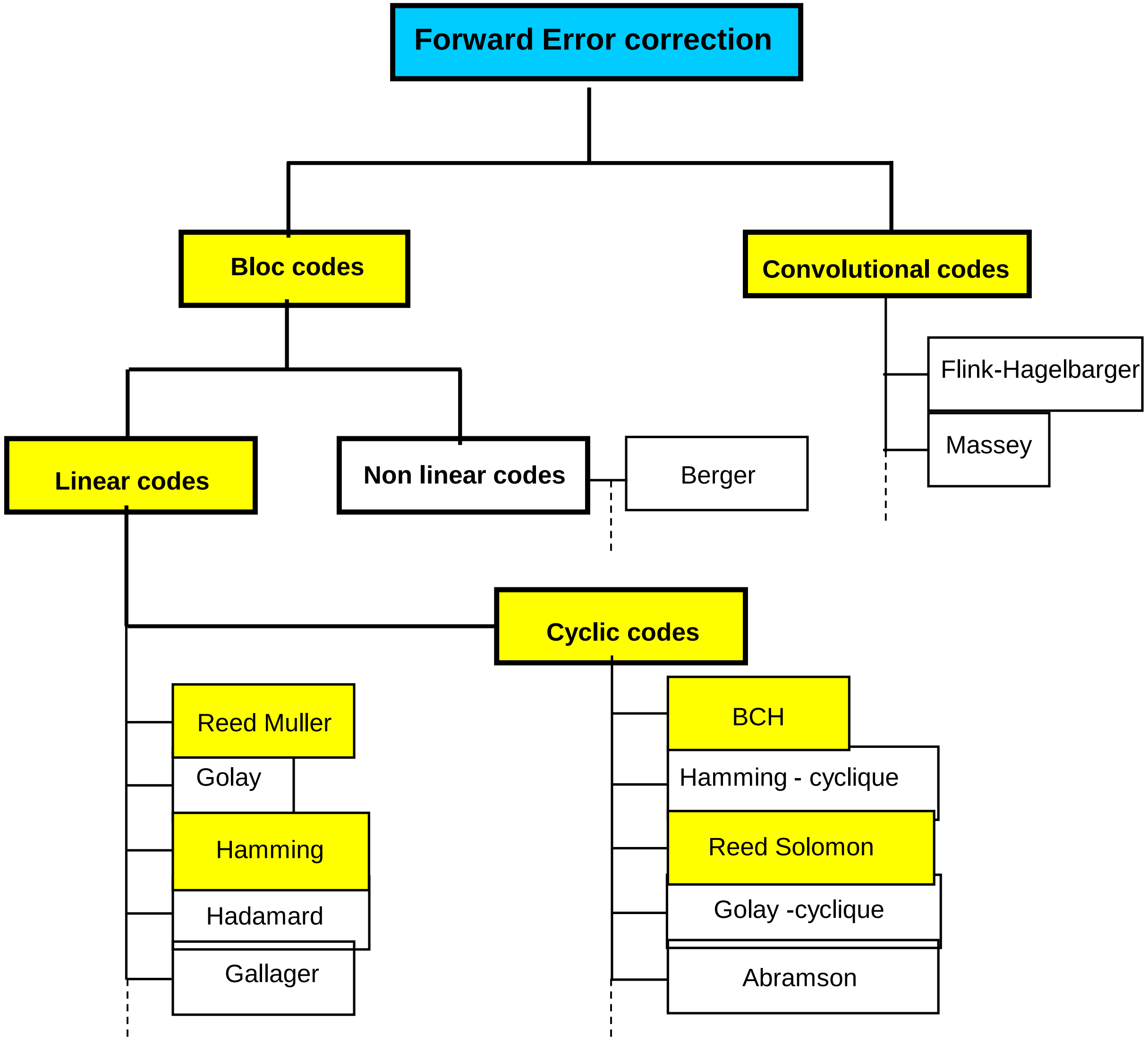
The communication chain



Channel coding

- Add redundancy to protect the information
- This adding is done according to a rule shared by the receiver and transmitter

Many types of codes



17

Principles of channel coding

k binary digits

First way

Systematic code: the redundancy is concatened with redundancy

k binary digits	r=n-k binary digits	Code word
Information	redundancy	

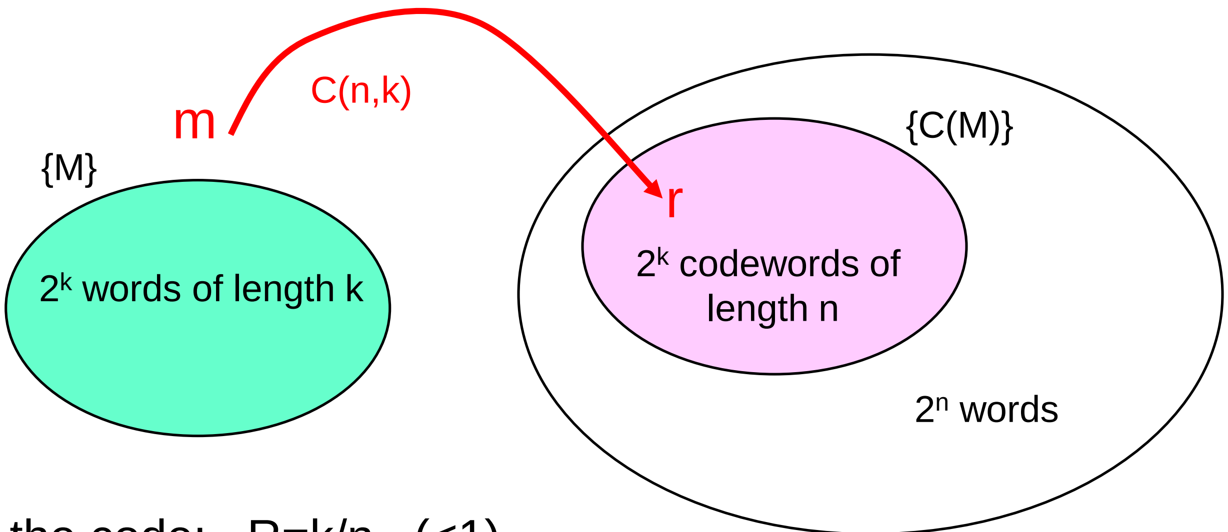
k binary digit — Code word

Second way

codage non systématique

n caractères

Principles of channel coding



Error detection if the received word does not belong to the set of codewords

Rate of the code: R=k/n (<1)

Dimension of the code : 2^k

Length of codewords: *n*

Distance of the code d_H : $d_H = \min_C w(C)$

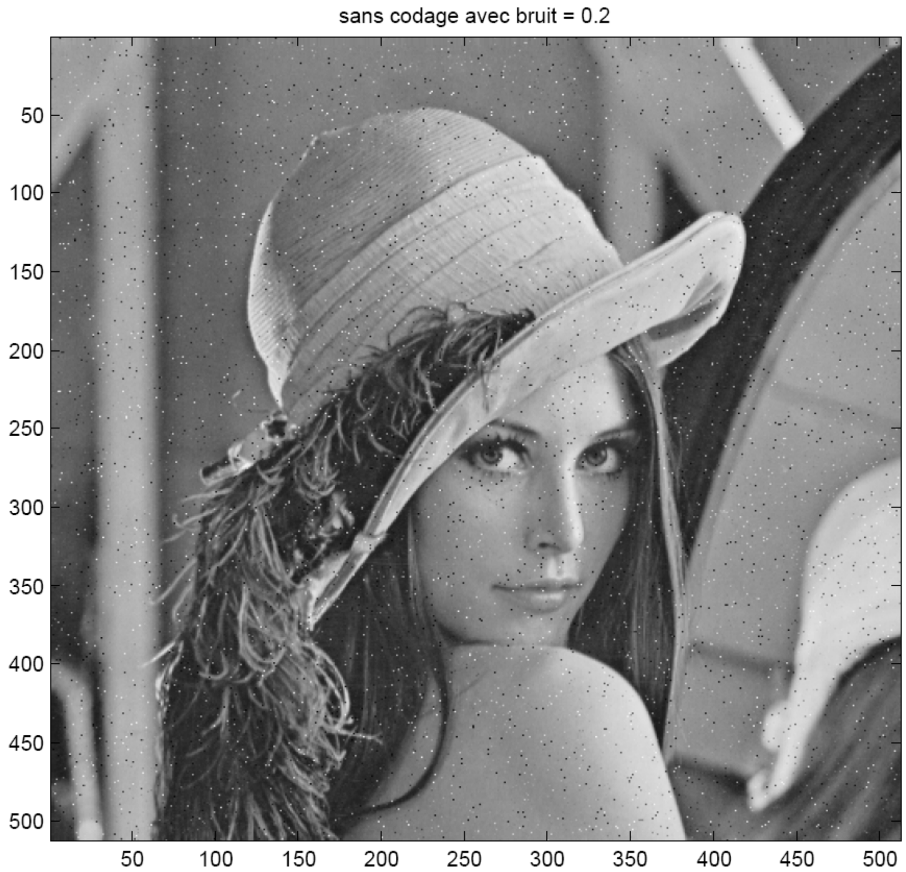
(*w* : Hamming weight)
(C : code word)

Number of corrected digits *e* :

$$e = \left[\frac{d_H - 1}{2} \right]$$

Powerfullness of error correction (1/2)

R=0.5

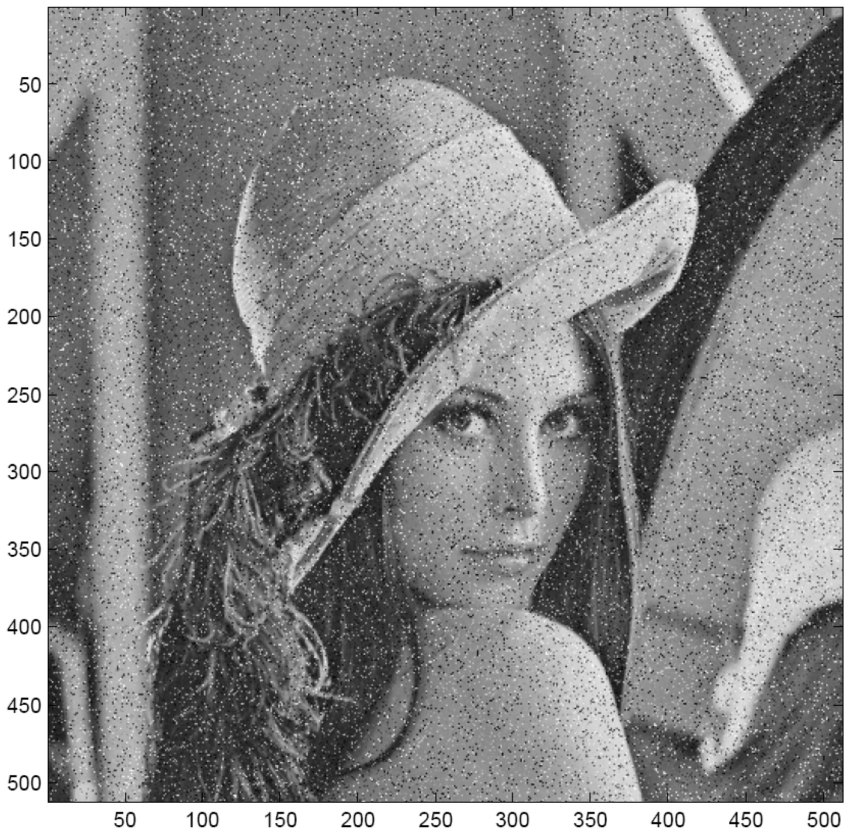


(R : rate of the code)

Powerfullness of error correction (2/2)

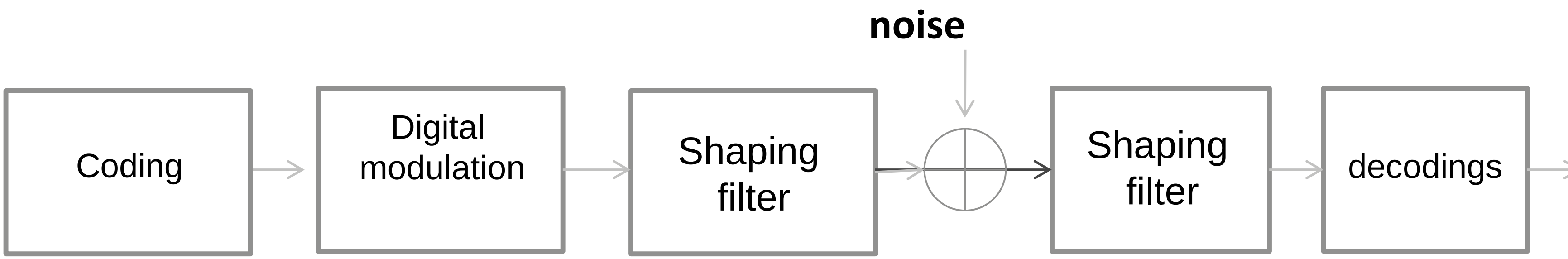
R=0.5

sans codage bruit=0.3



avec décodage BCH (127,64,t=10), R=0.3

Optimal scheme over a gaussian channel



Application: Digital Video Broadcasting (DVB)

- **Source coding**: MPEG-2
- **Error correction**: convolutional code + block code (Reed-Solomon)
- **Digital modulation**: QPSK (M=4)
- **Shaping filter:** Raised Root Cosine filter of coefficient β=0.35

Digital communications

Source coding

Compress the information

Source coding

An idea of data compression needs

Video

1 image: 700 pixels x 500 lines
25 images/s
1 pixel = 2 samples (brightness & chrominance)
1 sample : coded on 8 binary elements
⇒ Data rate of 140 Mbits/s
⇒ The compression standard is MPEG 2 (digital TV) yields a 4 Mbits/s data rate

Music (CD)

Sampling frequency: 44,1kHz
2 paths (stereo)
1 sample is coded on 16 digits
⇒ Data rate of 1,4 Mbits/s
⇒ MP3 compression algorithm provides a data rate of 128 kbits/s

GSM (2G)

Sampling frequency: 8kHz
1 sample is coded on 13 binary digits
⇒ Data rate of 104 kbits/s without compression
⇒ The usefull data rate is of 10 kbits/s

Source coding

Objectives

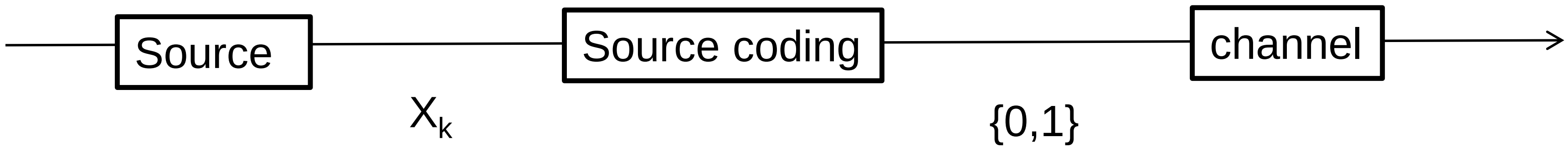
- At the transmitter, the aim is to change the information message provided by a source by the shortest one for material reasons (bandwidth, memories, storage, …)

- This process (long -> short messages) has to be traced-back at the Receiver to recover the original information

- To do so, all unecessary redundancies have to be removed
- At the maximum of the compression process, every quantum of data is absolutly necessary => necessay to protect them

Source coding

Fondamentals



The source provides **SYMBOLS** noted X_k
The source coding provides **CODEWORDS** noted C_k whose components are 0 or 1

The source codes should have the following properties:

- **Bijective application** : 1 symbole <-> 1 codeword
- **Uniquely decodable** : 1 sequence of symboles <-> 1 sequence of codewords

Source coding

Fondamentals

Ways to make a code uniquely decodable

1. use codewords of same length (OK but not the most efficient regarding the mean lenght)
2. use the same inter-word between codewords (not efficient)
3. by avoiding that a codeword has the same first digits than an other

These last condition (3) ensure that the code is a <u>prefix code</u> or that it is <u>instantaneous</u>

Example

Symbols	Code A	Code B	Code C
A	1	0	0
B	00	10	01
C	01	110	011
D	10	111	111

We want to transmit the message ‘BDC’

Code A : BDC -> 001001
At the receiver, there could be an ambiguity as 001001 can be viewed as ‘BABA’ (because ‘1’ is the beginning of ’10’)

Code C : BDC -> 01111011
At the receiver, one may decode as AD …. and 1011 has no associated symbols
$\Rightarrow$ the receiver has to start from scratch → the code is not instantaneous
(because ‘1’ and ‘01’ are the beginnings of ’111’ and ‘011’ respectively)

What about code B?

Extension of a code

Let be a memoryless source which delivers symbols X_k

Let us define the **extended** symbol $(X_1, X_2, \dots X_p)$ formed by the concatenation of p symbols of X

The set of all these extented symbols refered to the extension of order p of X. This set is noted X^p

Theorem 1

For a memoryless source X of entropy H(X) and p integer,

$$H(X^p) = pH(X)$$

To be shown by recurrence

Source coding

1 - Codes of fixed length

Definition 1: Mean length of a code

Let S be memoryless source of **n** symbols X_k
The probability distribution of the source is: $p_k = Pr[X_k]$
The symbols are coded into codewords of length n_k

Then the mean length **m** of the code is defined as:

$$m = \sum_{k=0}^{n-1} p_k n_k$$

Two types of codes:

fixed length
variable length

Source coding

2 – Efficiency of a fixed-length code

Property 1

X is a source of n symbols. It is possible to code it with a fixed-length Code of length m so that

$$log_2(n) \leqslant m \leqslant 1 + log_2(n)$$

The efficiency η of such a code is given by:

$$\eta = \frac{H(X)}{m}$$

As $H(X) \leq log_2(n)$

$$\eta \leqslant 1$$

η=1 ⇔ all symbols are of equal probability and **n** if a power of 2

Source coding

Example

X={0,1, 2, 3, 4, 5, 6, 7, 8, 9}
n_1=10

All the symbols are supposed to have same probability

Each symbol can be coded with **m**=4 bits

(because $\log_2 10 < 4 < 1+\log_2 10 \Leftrightarrow 2^3 < 4 < 2^4$)

$$\eta_1 = \frac{H(x)}{m} = \frac{log_2 10}{4} \approx 0,83$$

Now let us consider an extension of X of order p=2

Source coding

Example (cont'd)

X^2={00, 01, 02, …., 99}
n_2=100

The 100 symbols of X^2 can be coded on m_2=7 bits ($100<2^7$)

$$\eta_2 = \frac{H(x)}{m_2} = \frac{log_2 100}{7} \approx 0,95$$

And for p=3,

$$\eta_3 = \frac{H(x)}{m_3} = \frac{log_2 1000}{10} \approx 0,99$$

Source coding

Generalization

Theorem 2

Let X be a source of n symbols and let X^p its extension of order p.
It exists a code of fixed lenght m_p to code X^p which verifies

$$log_2(n) \leqslant \frac{m_p}{p} \leqslant \frac{1}{p} + log_2(n)$$ (to be proved)

Consequence 1

$$\lim_{p \to \infty} \frac{m_p}{p} = log_2(n)$$

If one defines the efficiency η_p of the coding of X^p

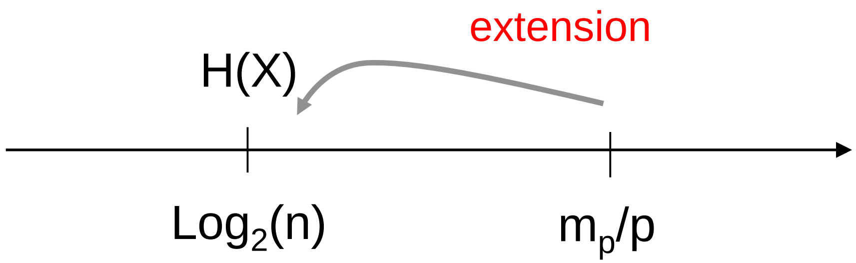
$$\eta_p = \frac{H(X^p)}{m_p}$$

$$\lim_{p \to \infty} \eta_p = \frac{H(X)}{log_2(n)}$$

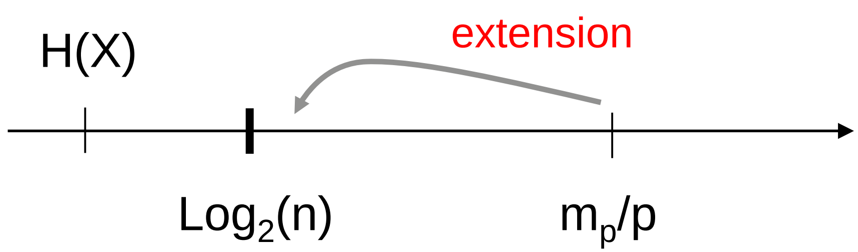
Generalization (cont'd)

Consequence 2

- If $H(X)<\log_2(n)$, the efficiency of the code will not tend towards 1
- If $H(X)=\log_2(n)$, the efficiency can be as closed as possible to 1



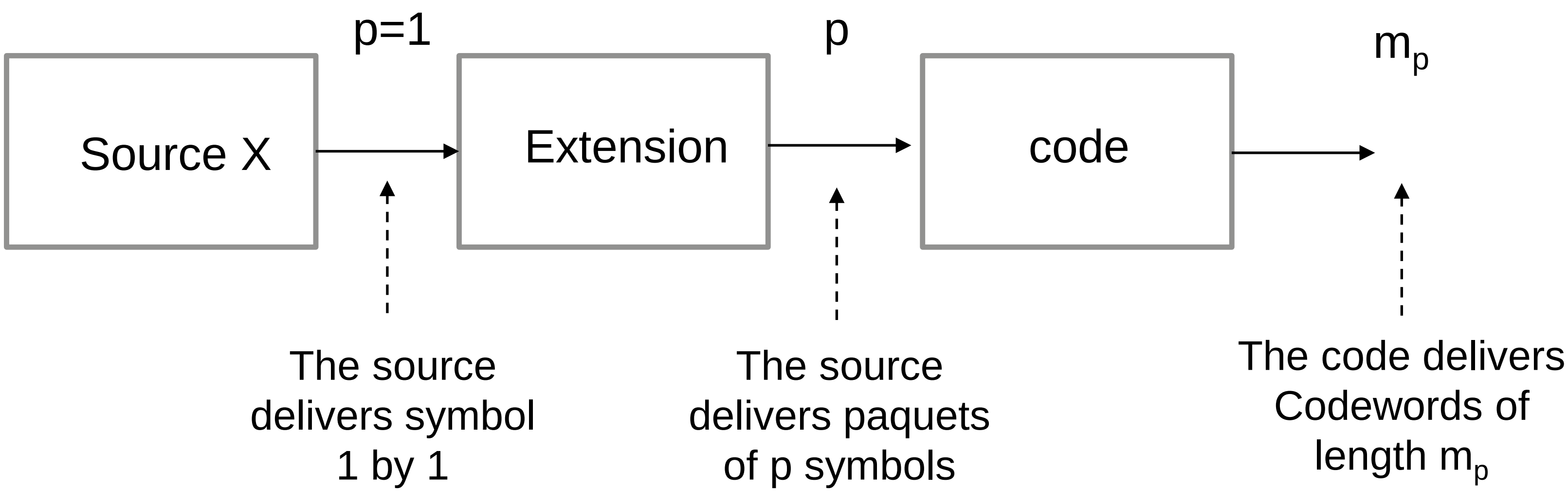
(source of equal probability)



(source of unequal probability)

Source coding

Interpretation of m_p/p



m_p binary digits ⇔ 1 symbol of X^p ⇔ p symbols of X

$\frac{m_p}{p}$: number of binary in a codeword for 1 symbol X
unity: bit/symbol of X

Source coding

Interpretation of m_p/p (cont'd)

Coming back to the example:

$$\frac{m_1}{1} = 4$$

$$\frac{m_2}{2} = 3,5$$

$$\frac{m_3}{3} = \frac{10}{3} = 3,33$$

$$\vdots$$

$$\frac{m_p}{p} \rightarrow log_2(10) = 3,32$$

The number of bit per
Transmitted symbol is decreasing
With p

This is the compression result

Source coding

Importance of the events probabilities of the source

Let us consider the following source

Elements of the source	Code I
A	00
B	01
C	10
D	11

The code is uniquely decodable, instantaneous
Considering that the 4 events have the same probability (1/4):

m_1=2 and p=1

Then $\mathbf{m_1/p = 2}$ bit/symbol of the source

and $H(X)=\log_2(4) = 2$

This is not possible to do better

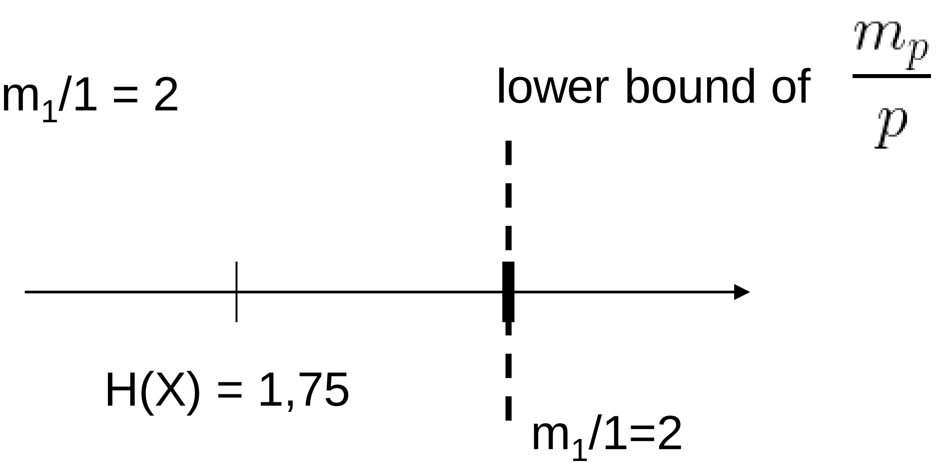
Source coding

But …. if the events don’t have the same probability

Elements of the source	Code I	Prob.
A	00	0,5
B	01	0,25
C	10	0,125
D	11	0,125

We still have m_1=2

But H(X)=1,75 < $\log_2(4) = m_1/1 = 2$



Source coding

As

$$\lim_{p\to\infty} \frac{m_p}{p} = log_2(n)$$ (=2 in the example)

this will not be possible for m_p/p to converge to H(X) while keeping the same number of digits per events

=> the length of the code has to be variable

Elements of the source	Code I	Prob.
A	0	0,5
B	10	0,25
C	110	0,125
D	1110	0,125

$$m = \sum_{k=0}^{n-1} p_k n_k$$

m=1,875 bit/symbol of the source

So m<2

Then this will be possible to converge toward the entropy (the lower bound)

Basic ideas of source coding

1. The codewords should have different length

2. This ensures the convergence toward the entropy so as to Minimize the length of the code

3. The fundamental rule is the following:

The events of highest probabilities will be coded with the lowest length

The events of lowest probabilities will be coded with the highest length

Source coding

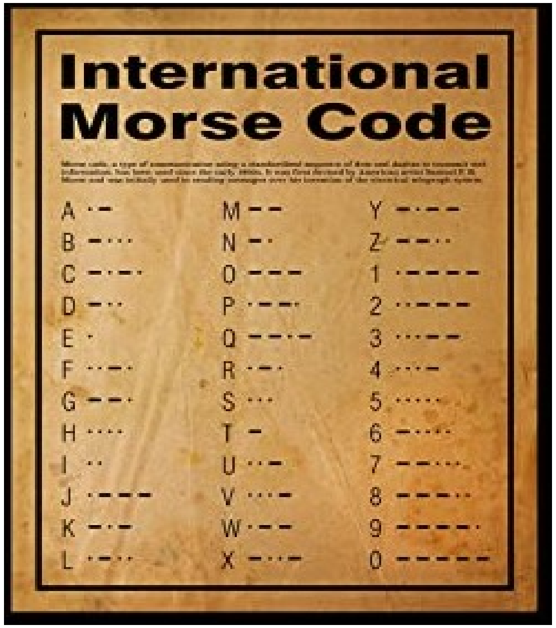
Basic ideas of source coding

Elements of the source	Code I	Prob.
A	0	0,5
B	10	0,25
C	110	0,125
D	1110	0,125

High probability
low number of bits

Low probability and
high number of bits

Idea of MORSE code (1832)

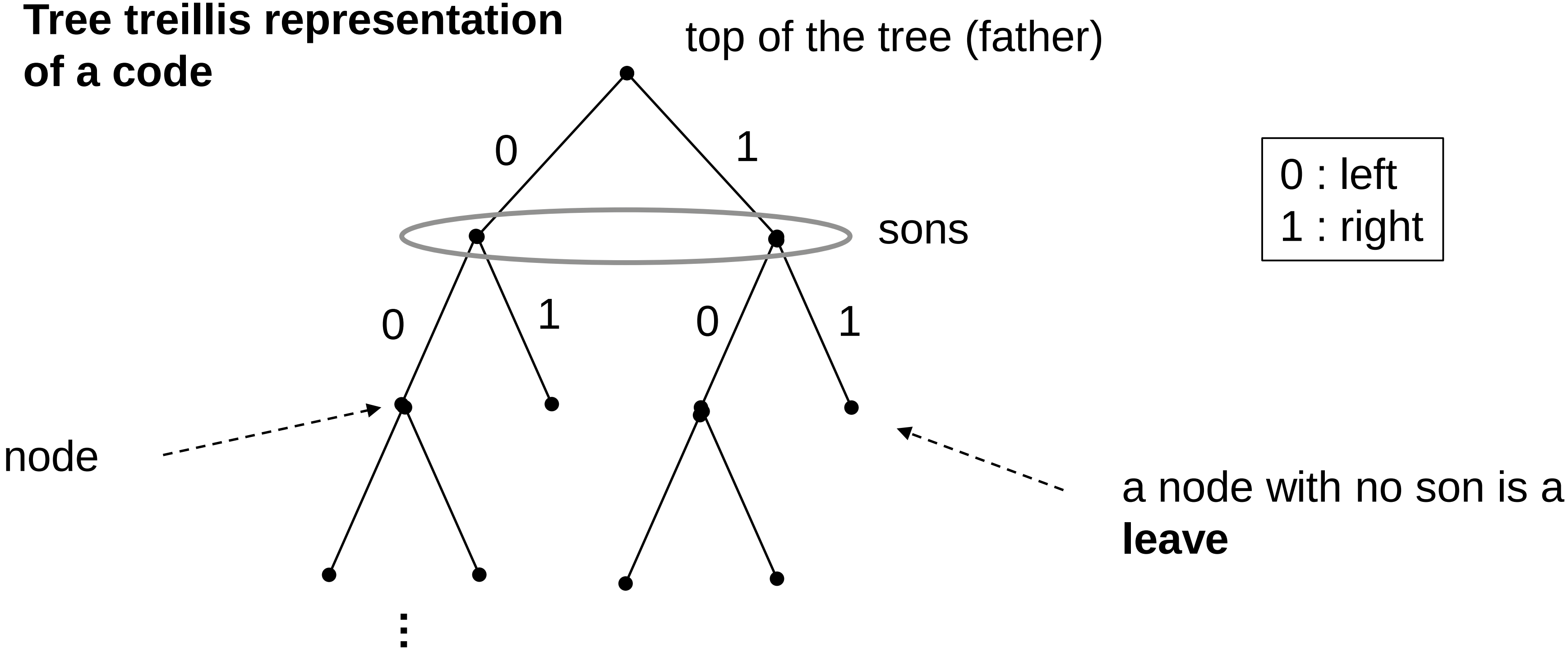


Source coding

Codes with variable codewords lengths

We focus on prefix codes (which are said instantaneous)

Tree treillis representation of a code

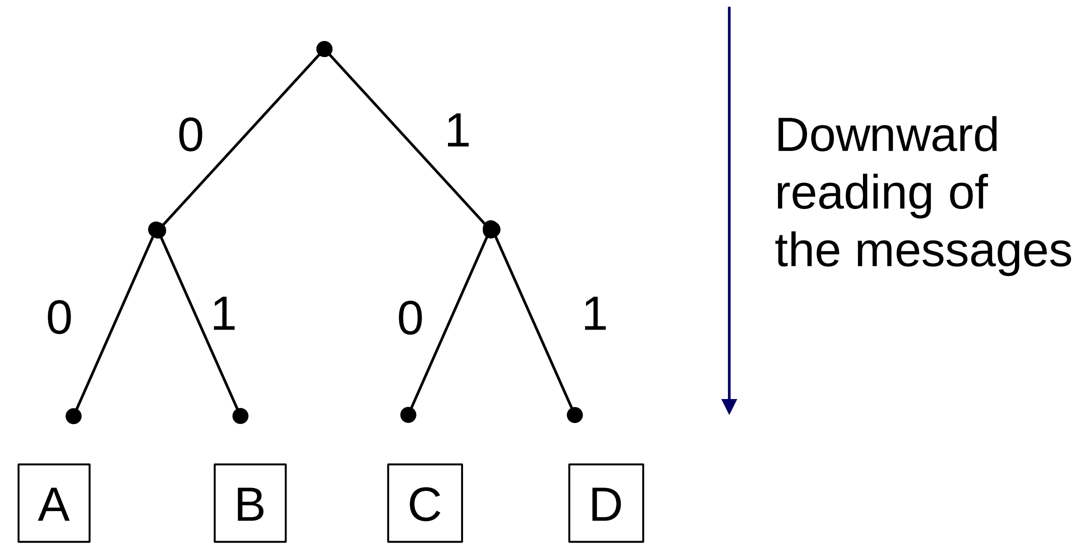


0 : left 1 : right

Source coding

Tree representation of a code

Elements of the source	Code I
A	00
B	01
C	10
D	11



A prefix code has a tree representation whose codewords are all leaves in the tree representation

What is the condition existence of prefix codes ?
This is the Kraft inequality

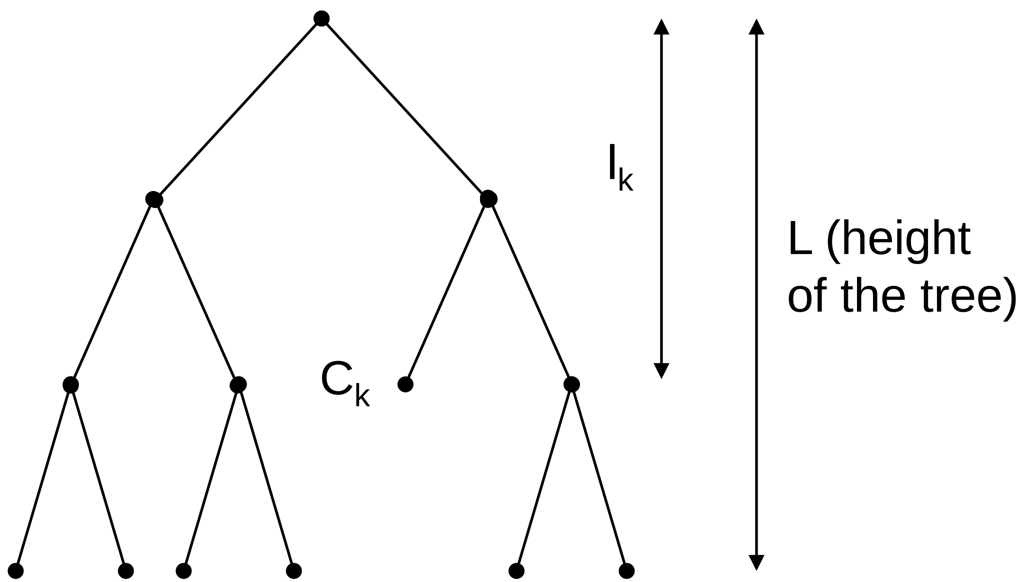
Source coding

Kraft inequality

This is possible to generate a prefix code whose codewords lengths are $l_1, l_2, \ldots, l_n$ (n is the number of codewords) if and only if

$$\sum_{i=1}^{n} 2^{-l_i} \leqslant 1$$

Proof

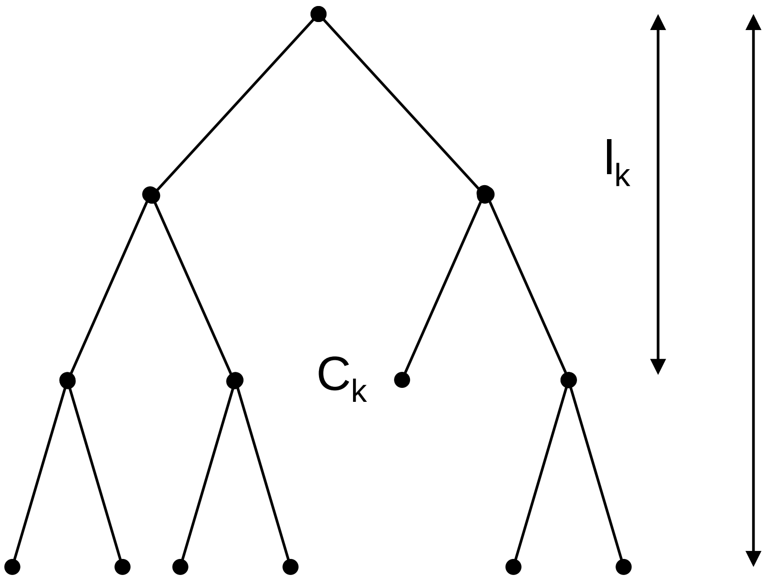


C_k is a codeword =>no son

2^{L-l_k} forbidden leaves

Source coding

Kraft inequality (cont'd)



Over n codewords, there are

$$\sum_{i=1}^{n} 2^{L-l_i} \leqslant 1$$

forbidden leaves

As $$\sum_{i=1}^{n} 2^{L-l_i} \leqslant 2^L$$

it gives $$\sum_{i=1}^{n} 2^{-l_i} \leqslant 1$$

Source coding

First Shannon theorem on source coding

For any source X of entropy H(X), it is possible to find a prefix source code of mean length m so that

$$H(X) \leqslant m \leqslant 1 + H(X)$$

<u>Proof</u>

- Part 1

$$\begin{aligned}\Delta = H(X) - m &= -\sum_i p_i log_2(p_i) - \sum_i p_i l_i \\ &= -\sum_i p_i log_2(p_i) - \sum_i p_i log_2(2^{l_i}) \\ &= \sum_i p_i log_2 \frac{2^{-l_i}}{p_i}\end{aligned}$$

Source coding

First Shannon theorem on source coding (cont'd)

$$\Delta = \sum_i p_i log_2(e) ln(\frac{2^{-l_i}}{p_i})$$

$$ln(X) \leqslant X - 1$$

=> $$\Delta \leqslant log_2(e) \left[\underbrace{\sum_i 2^{-l_i}}_{\leqslant 1} - \underbrace{\sum_i p_i}_{= 1} \right]$$

=> $$\boxed{H(X) \leqslant m}$$

Source coding

First Shannon theorem on source coding (cont'd)

$$H(X) \leqslant m$$

- there is equality when $p_i = 2^{-l_i}$

This is in perfect line with the fact that probability and length are closely related (high probability => short codeword ; low probability => long codeword)

In case of equality, $l_i = -log_2(p_i)$

But l_i has to be an integer and may be real in the general case where:

$$l_i - 1 \leqslant \underbrace{-log_2(p_i) \leqslant l_i}$$

because $\Delta = \sum_i p_i log_2(e) ln(\frac{2^{-l_i}}{p_i})$ is negative

Source coding

First Shannon theorem on source coding (cont'd)

Proof

- Part 2

=> $$l_i - 1 \leqslant -log_2(p_i) \leqslant l_i$$

=> $$p_i l_i - p_i \leqslant -p_i log_2(p_i)$$

By summation

$$m - 1 \leqslant H(x)$$

and

$$\boxed{m \leqslant H(x) + 1}$$

First Shannon theorem on source coding (cont'd)

$$H(X) \leqslant m \leqslant 1 + H(X)$$

The efficiency is still defined as

$$\eta = \frac{H(X)}{m}$$

The condition to get the maximul efficiency value (ie 1) is: $p_i = 2^{-l_i}$

This is a strong condition which is not usualy verified

The extension of the source of order *p* will help to converge to 1

Source coding

Second Shannon theorem on source coding

For any source X of entropy H(X) and of extension X^p, it is possible to find a prefix source code of mean length m_p so that

$$H(X) \leqslant \frac{m_p}{p} \leqslant H(X) + \frac{1}{p}$$

Proof: consider X^p and $H(X^p) = pH(X)$

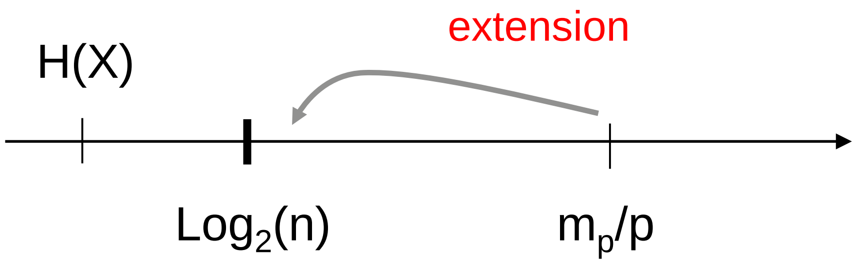
So it exists a source coding scheme for memoryless sources that make the efficiency the closest as possible to 1 by extending the orginal source

In these conditions, the mean length of the code converges to its lowest possible value: the entropy of the source

Source coding

Synthesis

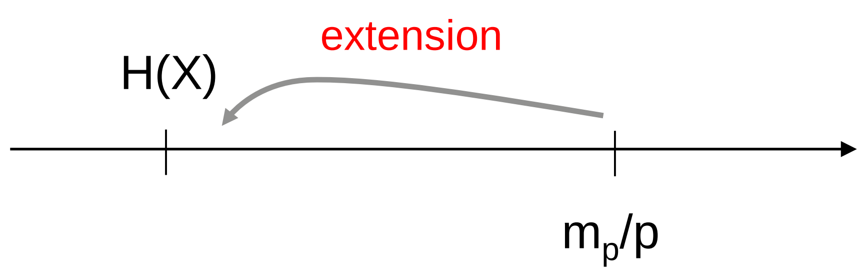
Codes of fixed lengths



(source of unequal probability)

Here the lowest value of the code (the entropy) cannot be reached unless the events have all same probabilities (strong condition)

Codes of variable lengths



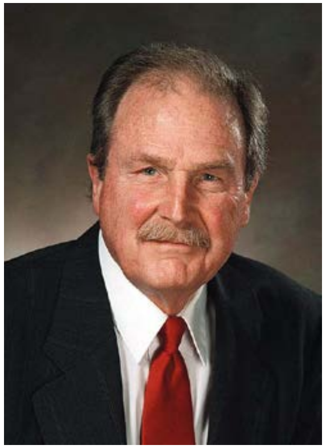
Here the lowest value of the code (the entropy) can be reached by extension whatever the source distribution

Source coding

Huffman coding method

This source coding method was suggested in 1952 by David Albert Huffman (1925-1999) during his PhD at MIT

It provides a code with variable codewords lenghts with a low mean length



Huffman coding method (cont'd)

The rules are the following:

1. From left to right, the events are ranked in the upward probability order

2. Gather the two events whose cumulative weight is the lowest.
These two events draw two branches of a tree to a father node whose weight is the lowest among all the possible couples

3. Repeat step 2 until a single event of probability one is obtained.
This event is the top father node of the tree

4. Process a backward coding of the code from the top father node to the initial events with the rule: 0 for a left-branch and 1 for a rigth branch

Huffman coding method (cont'd)
Example 1

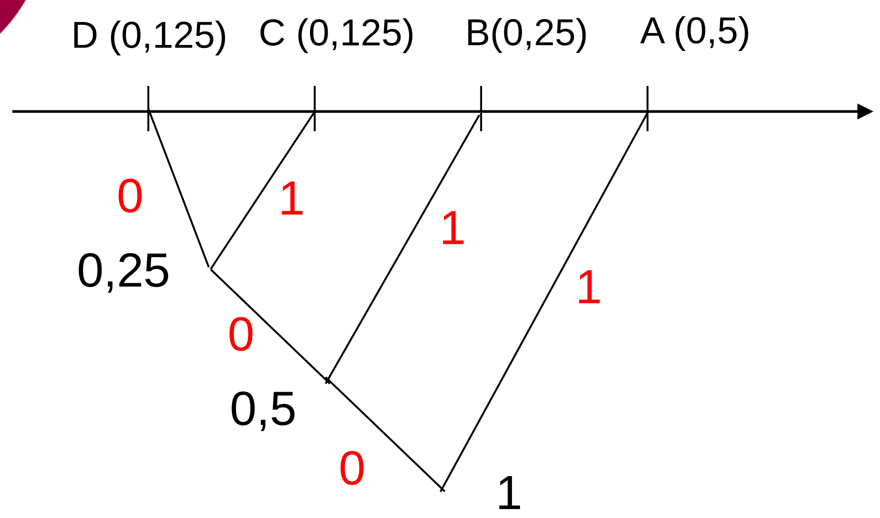
Events	**Prob.**
A	0,5
B	0,25
C	0,125
D	0,125

1. Suggest a source coding (from Huffman's algorithm)
2. What is the mean length of the code?
3. What is the efficiency of the code? Why?

56

Source coding

Huffman coding method (cont'd)
Example 1 (solution)



=>

Events	Codeword
A	1
B	01
C	001
D	000

Mean length:
m= 1x0,5 + 2x0,25 + 3x0,125 + 3x0,125 = 1,75 bit/symb.

Entropy of the source:
H(X)=1,75 bit/symb. (to be verified)

Here, m=H(X) what is exaclty the lower bound => not possible to decrease m
Why? Because in this example $pi=2^{-li}$

57

Source coding

Huffman coding method (cont'd)
Example 2

Consider a file to be transmitted with 35 characters (letters) with the following distribution:

Characters	#
A	5
B	7
C	1
D	14
E	6
F	2

1. What is the minimal number of bits necessary to transmit this file without any source coding strategy?
2. Propose a source coding strategy (following Huffman's algorithm)
3. What is the compression rate?
4. Is there a limit of the compression rate?

Yves LOUËT
Professeur de CentraleSupélec
Equipe SCEE - Laboratoire IETR

Avenue de la Boulaie, CS 47601
35576 CESSON SEVIGNE Cedex

https://research.centralesupelec.fr/yves.louet/
Yves.louet@centralesupelec.fr

59