S&M
Split and Merge Compression Algorithm

By

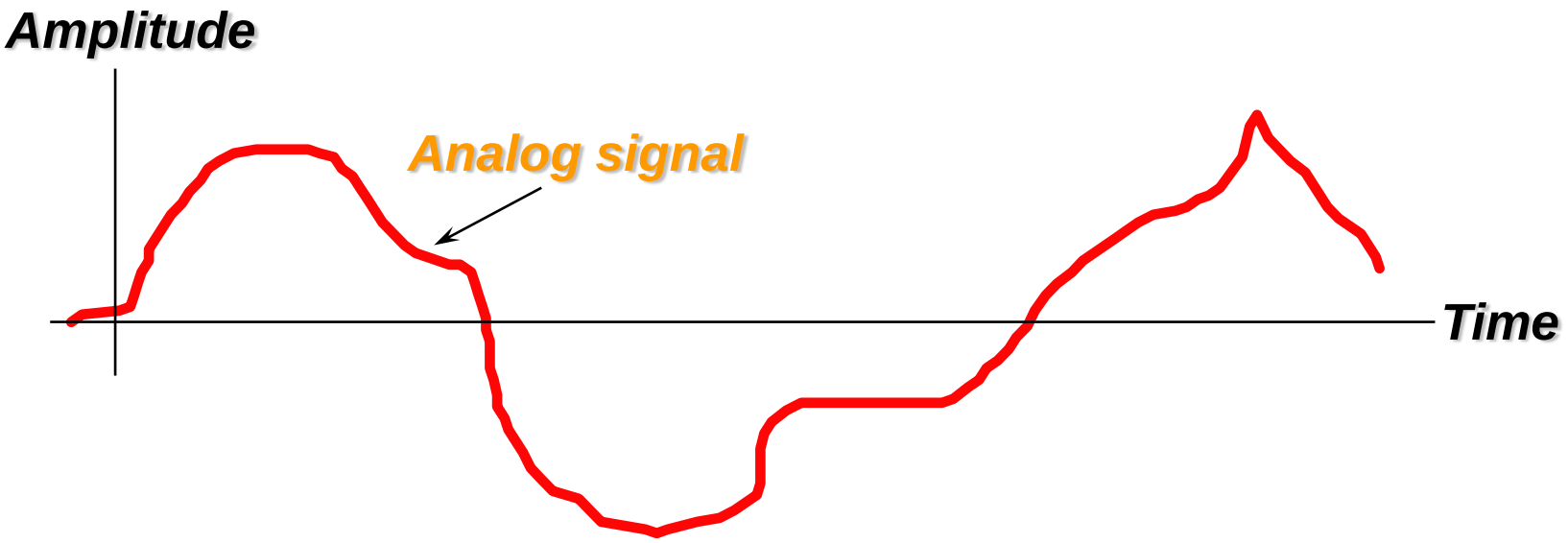
Abdullah Hashim

Compression Fundamentals

Information

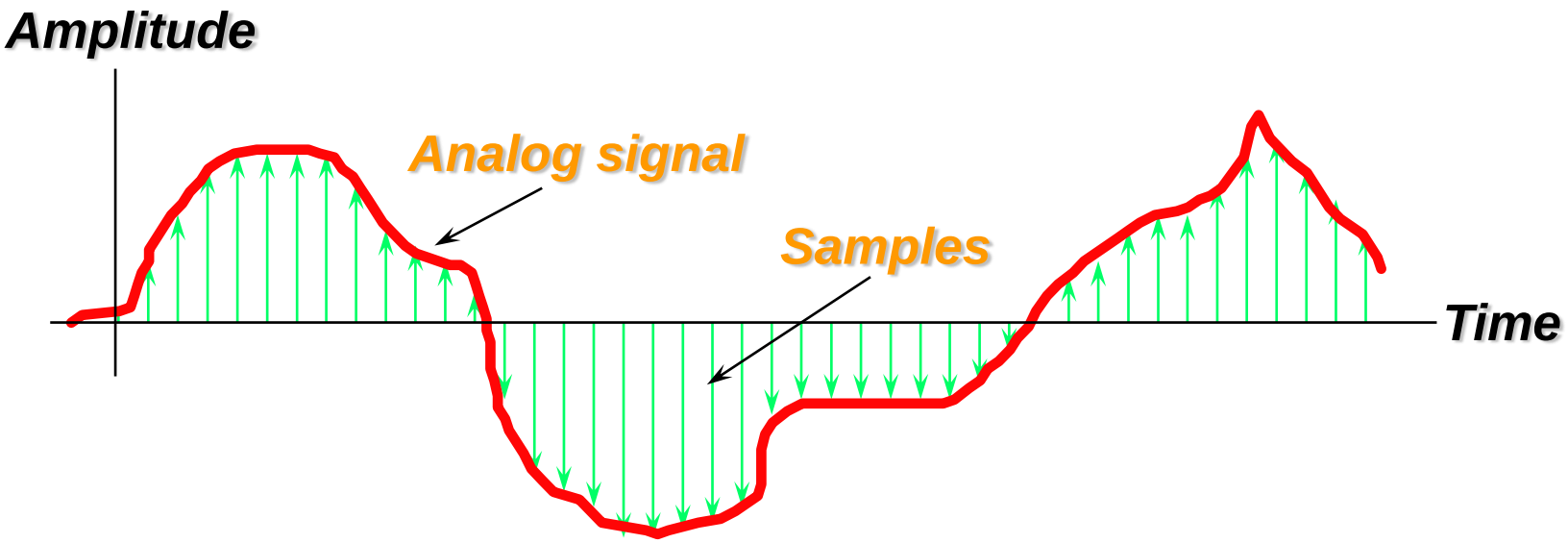
Any source which generates a variable quantity is called an information source.

Information



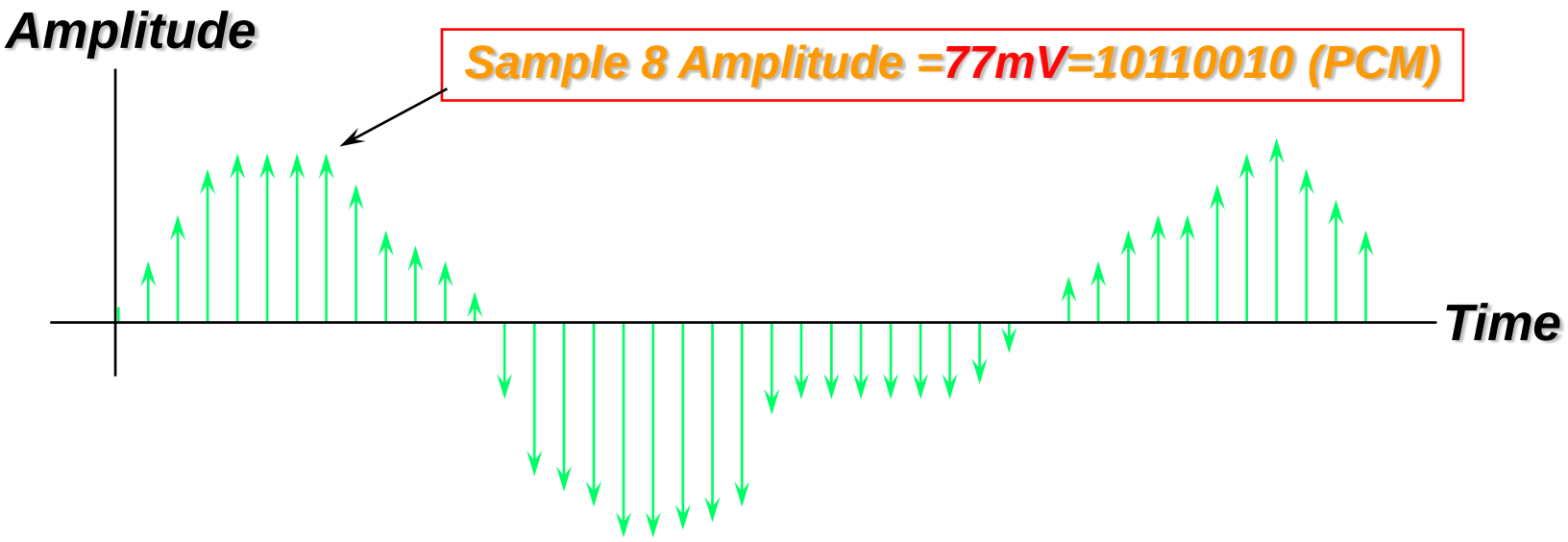
Any source which generates a variable quantity is called an information source.

Information



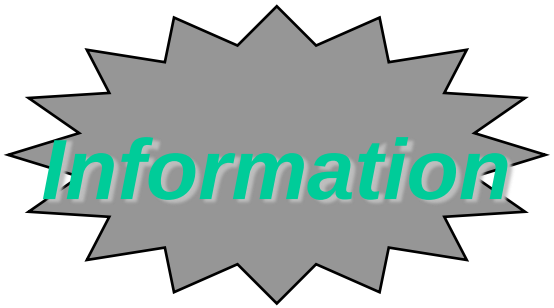
- **Samples rate => twice the maximum frequency of the Analog signal**
- **The amplitude of the sample may be quantised to 256 discrete levels (8 bits).**

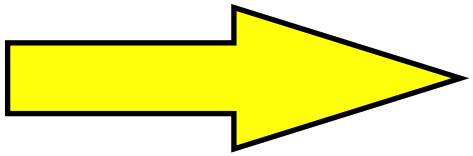
Information



Sample 8 Amplitude - Sample 9 amplitude =13mV=1011 (ADPCM)

Information



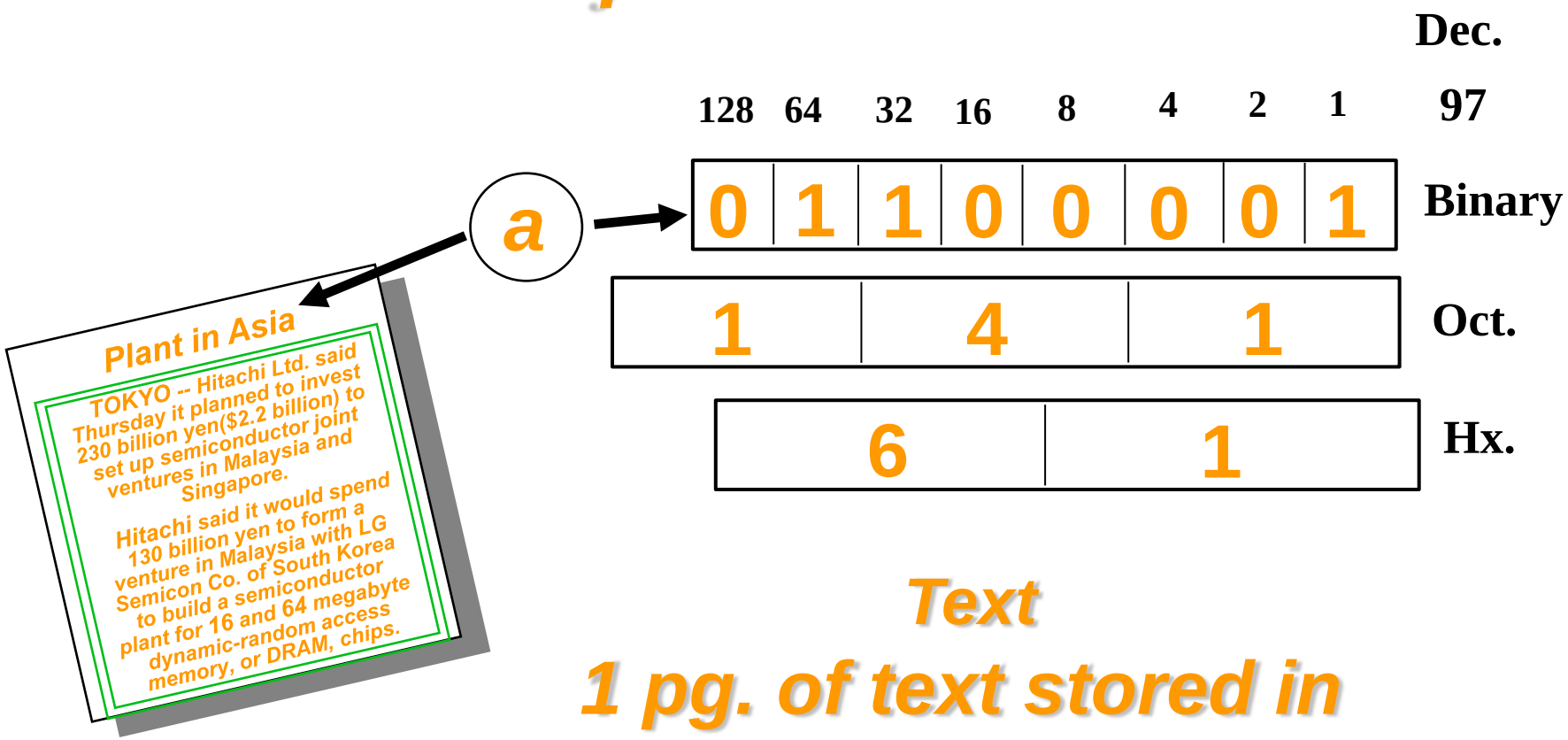


Stream of Bits

10110010 ---- 11001111 --

Eight Bits = One Byte = ASCII Character

Information



Text

1 pg. of text stored in 5 KByte

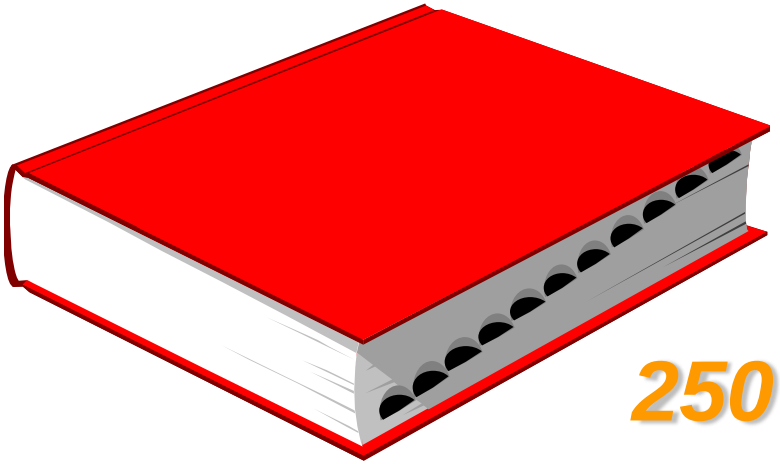
Any source which generates a variable quantity is called an information source.

Information

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

Information

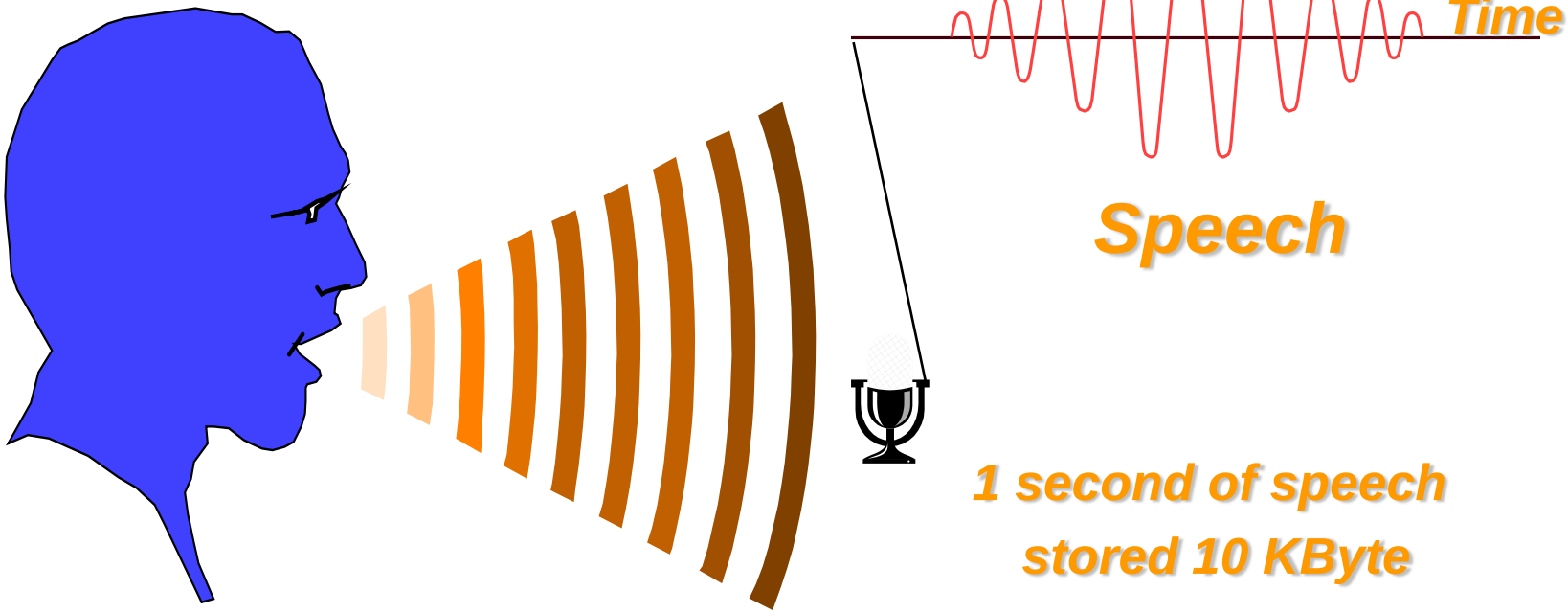


Text

250 pg. book stored in 1MByte

Any source which generates a variable quantity is called an information source.

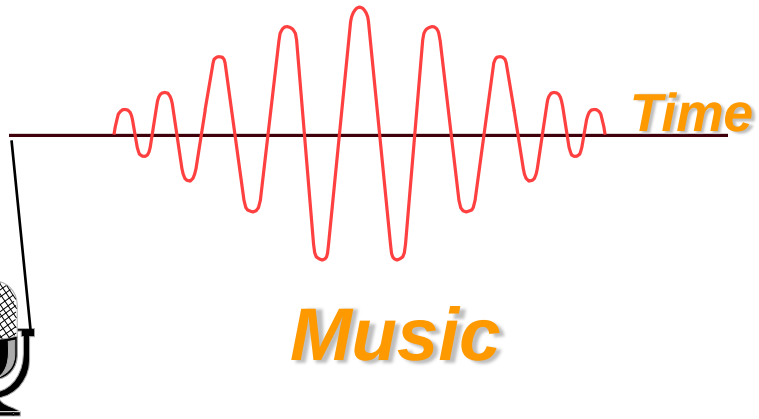
Information



1 second of speech
stored 10 KByte

Any source which generates a variable quantity is called an information source.

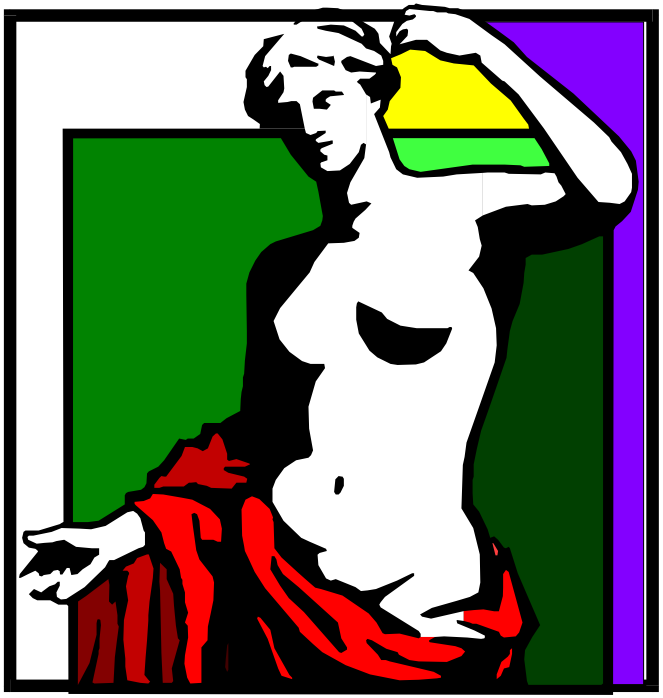
Information



1 second of music
stored 50 KByte

Any source which generates a variable quantity is called an information source.

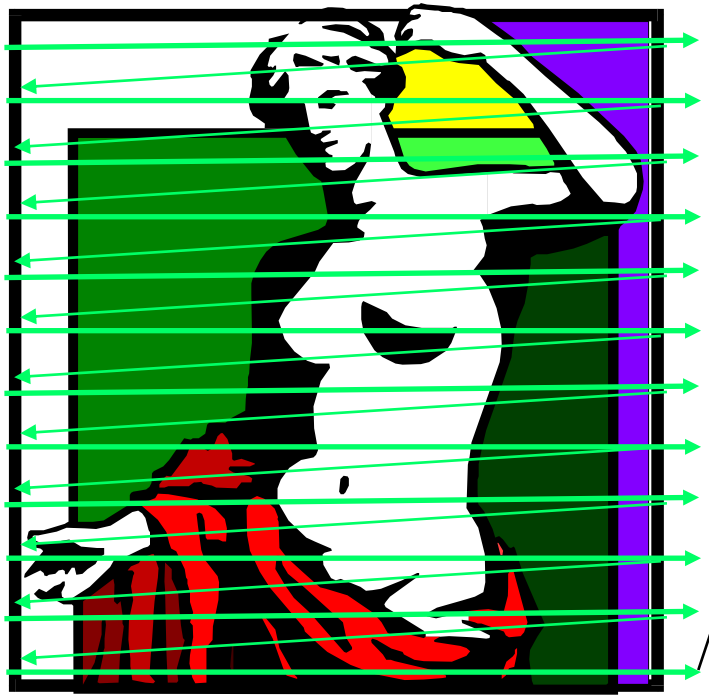
Information

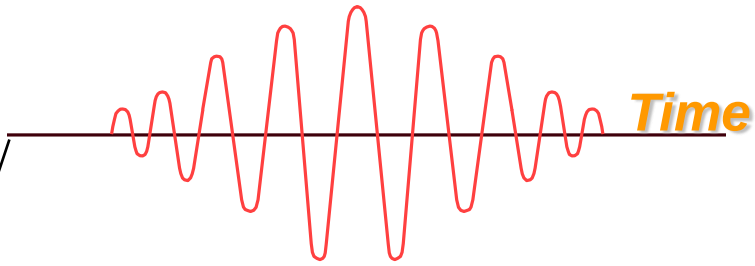


Image

Any source which generates a variable quantity is called an information source.

Information



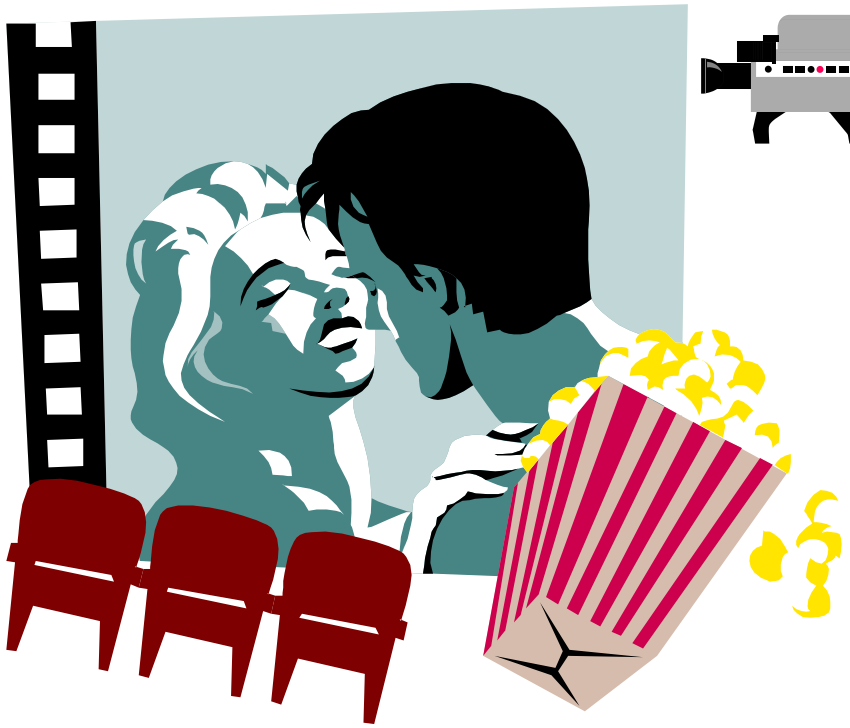


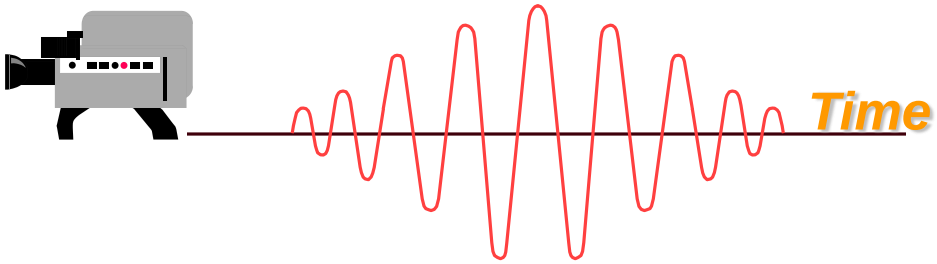
Image

256 X256 pixel image
stored in 64 KByte

Any source which generates a variable quantity is called an information source.

Information



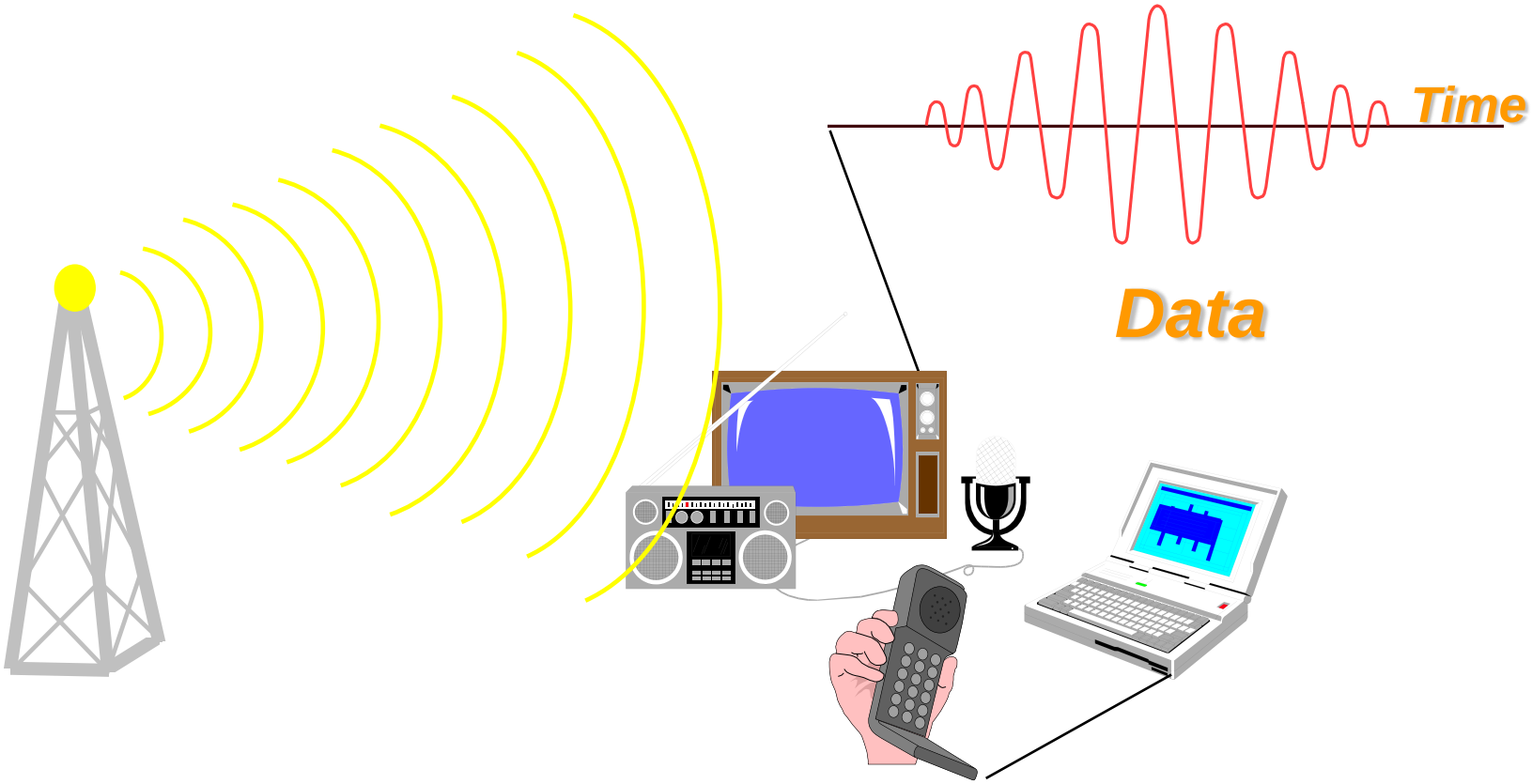


Video

1 second of video stored 7.68 MByte

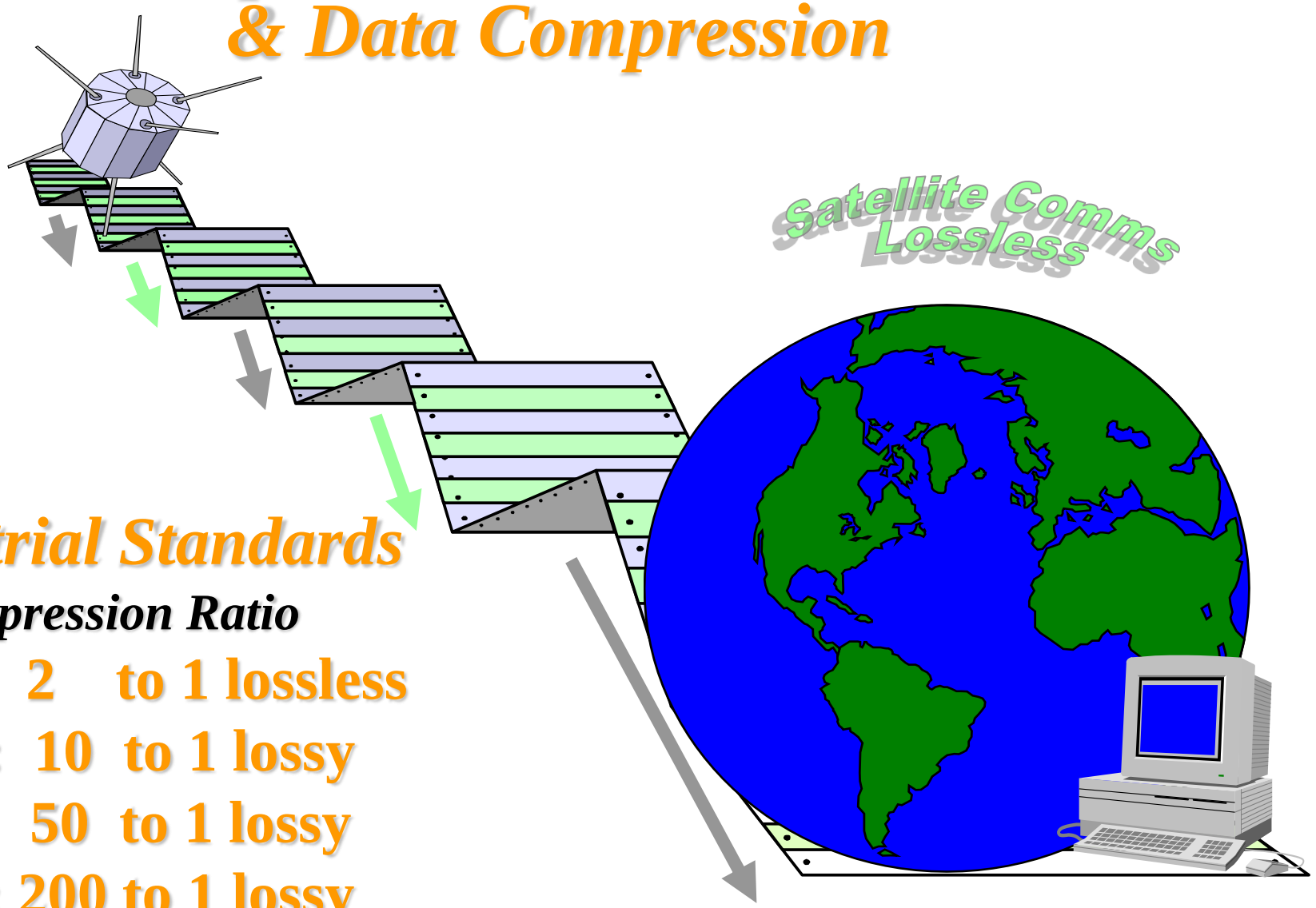
Any source which generates a variable quantity is called an information source.

Information



Any source which generates a variable quantity is called an information source.

IT Age
Computers, Communications & Data Compression



Industrial Standards

Compression Ratio

Text : 2 to 1 lossless

Sound: 10 to 1 lossy

Image: 50 to 1 lossy

Video : 200 to 1 lossy

Compression Fundamentals
-Information Theory-

Hartley (1928). (*Transmission of Information*)

Shannon (1948). (*Information Theory*)

Information content (I) in bits of a symbol (α_i) with a probability (p_i) is given by the expression:

$$I_i = \log_2(1 / p_i) = - \log_2(p_i) \quad \text{bits}$$

The measure of the average information per symbol of N symbols source (s_i) is called the entropy (H) of the source is given by the following expression:

$$H = \sum_{i=0}^{N-1} (p_i I_i) \quad \text{bits per symbol}$$

Compression Fundamentals

Examples of Data Source Entropies

Sample Type	Sample Entropy	Comments
English Text	**4.03 bits per symbol**	**Shannon (1951)**
Portuguese Text	**3.92 bits per symbol**	**Manfrino (1969)**
C++ Code	**5.29 bits per symbol**	**Measured**
Executable Code	**5.80 bits per symbol**	**Measured**

Source is called memoryless if

i^{th+1} event is independent of the i^{th} event.

Compression Fundamentals

Variable Length Coding

When the average codeword length $L_{average}(\alpha_i)$ of source of (N) symbols equal to the source entropy the code is said to be ***optimal code***:

$$L_{average}(\alpha_i) = \sum_{i=0}^{N-1}[p_i\, L(\alpha_i)] \Rightarrow \sum_{i=0}^{N-1}(p_i\, I_i) = H$$

Compression Ratio $R = \log_2(N) / L_{average}(a_i)$

Prefix code consists of comma less unique codewords, i.e. no codeword may be a prefix of any other codeword

Note: *R* decreases to a great extent with slight changes of probability set p_i from that of the assumed values.

To ensure robustness against probability set p_i variation, codewords length are bounded to a given value say (L), where $L < (N - 1)$, N-1 is the longest possible length of codewords.

Compression Fundamentals

Entropy of a binary source with two symbol α and θ

Note:

when: $p_\alpha = 0$ **and**

$p_\theta = 1$

the entropy is minimum

$H_{min} \rightarrow 0$

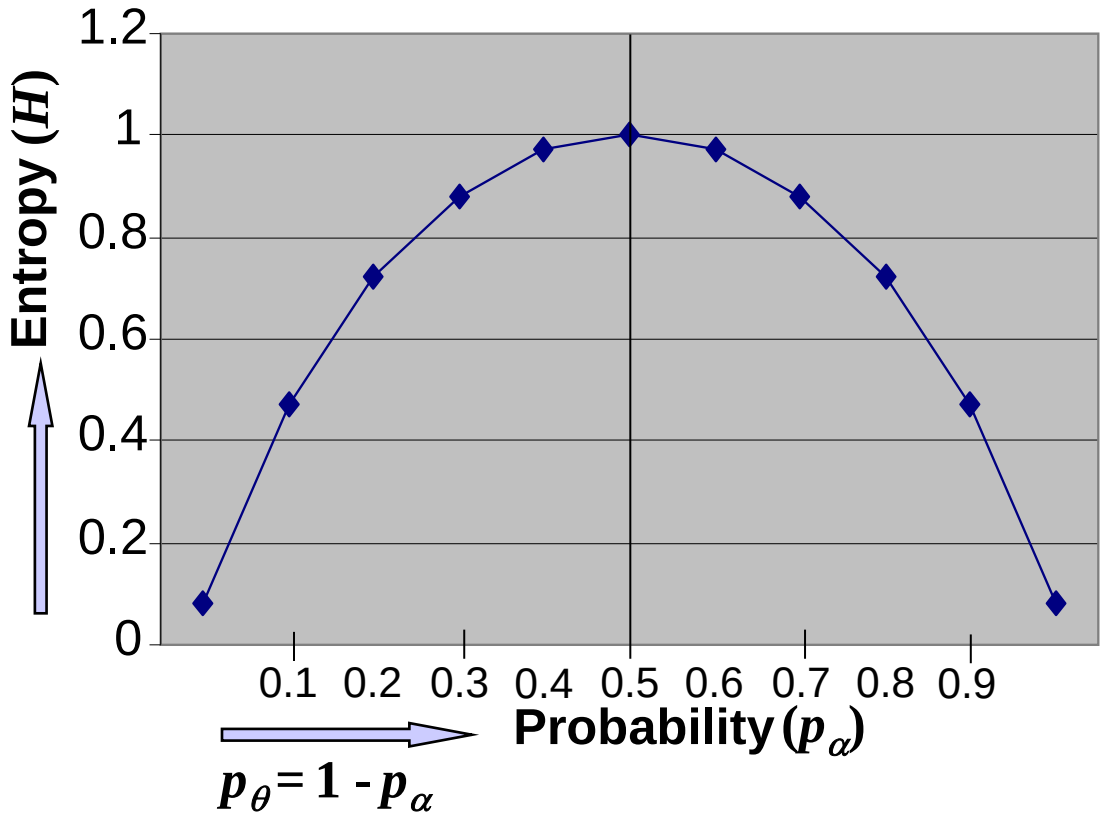
when: $p_\alpha = p_\theta = 1/N = 1/2$

the entropy is maximum

$H_{max} = log_2(N)$

where N **is the number of symbols in the source**

$$0 < H \leq log_2(N)$$



Note:

Source with N **equiprobable symbols is called a random source, and its entropy is equal to its symbol codeword length of** $[log_2(N)]$**. Random source is coded optimally with equal length binary codewords.**

Compression Fundamentals
Shannon-Fano Code

i) Arrange the symbols in order of decreasing probability, s_0 that symbol has the highest and s_{N-1} the lowest probability.

ii) Determine the cumulative probability P(<=j) for each symbol s_j the sum of probabilities P_k for values of k from 0 to j.

iii) The j-th codeword is given by the expansion as a binary number of p(>=j), the expansion being carried out to L_j places, where L_j is given by:-

$$I(<=j) <= L_j < 1 + I(<=j)$$

Compression Fundamentals

Practical implementation of Huffman code

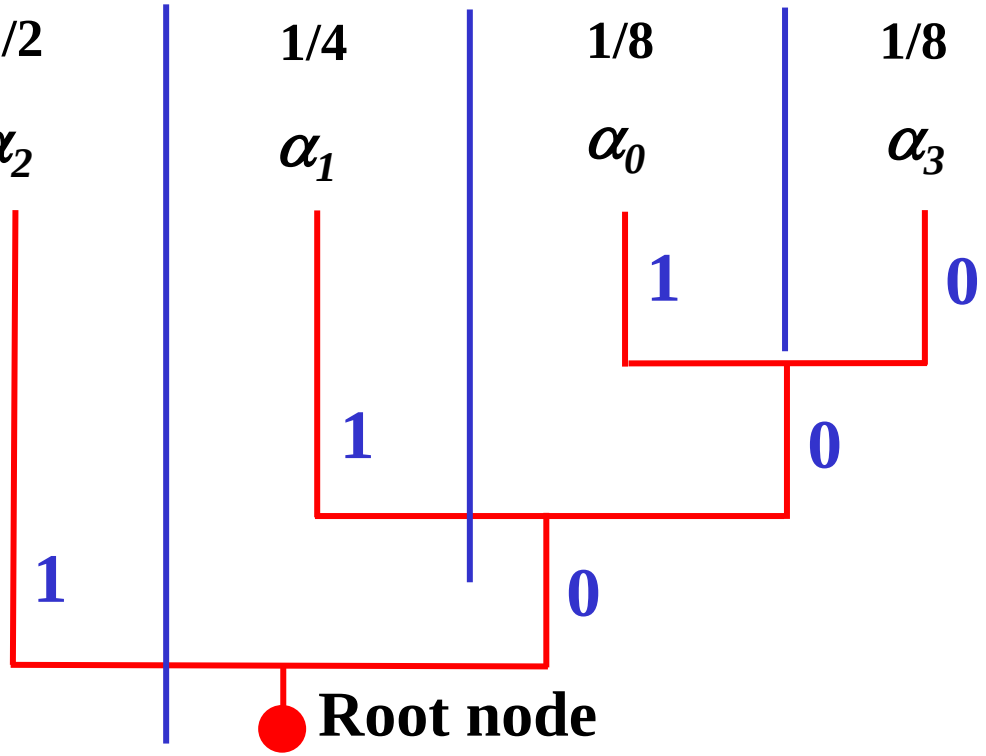
Consider a message of N symbols, α_0, α_1, α_2 and α_3, N=4, symbols probabilities are: $p(\alpha_2) = 1/2$, $p(\alpha_1) = 1/4$, $p(\alpha_0) = 1/8$, $p(\alpha_3) = 1/8$.

Probabilities 1/2 1/4 1/8 1/8

Symbols α_2 α_1 α_0 α_3

Coding

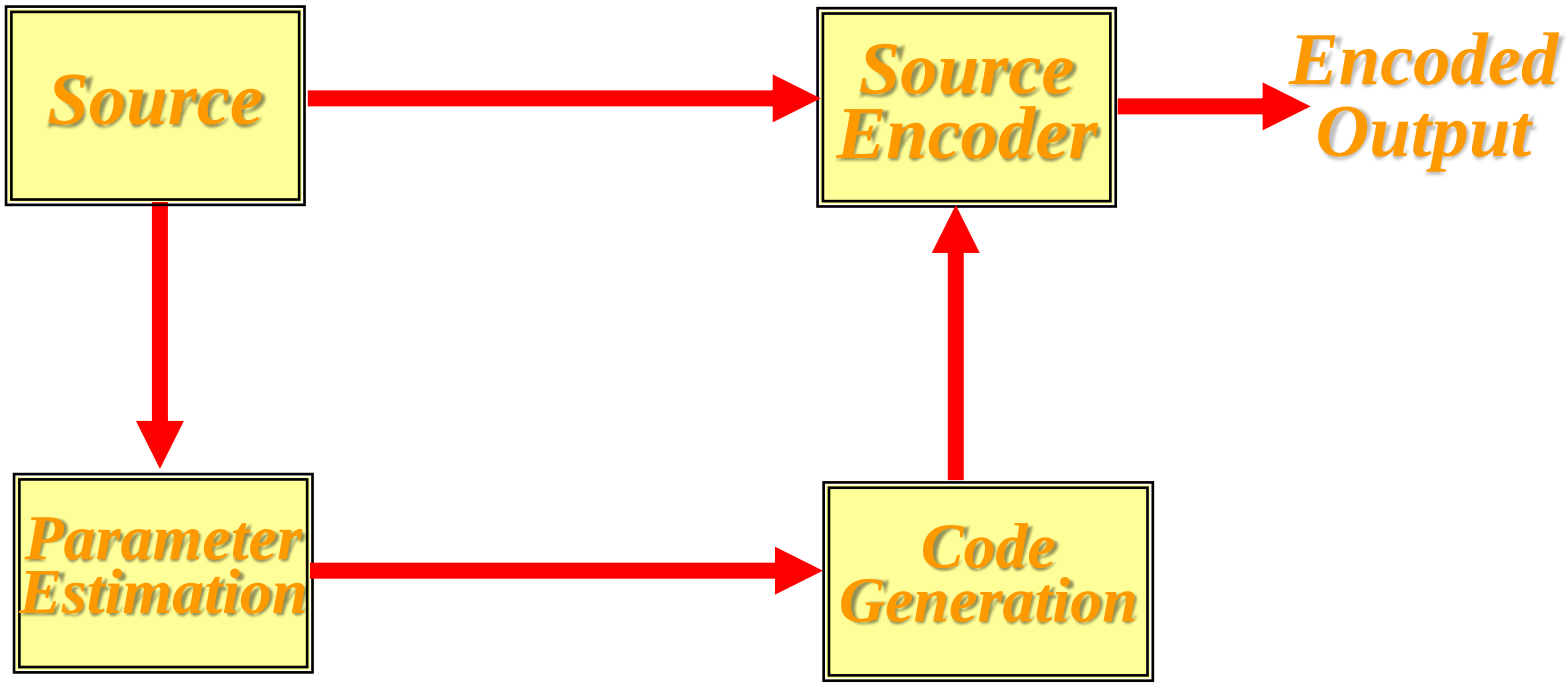
$\alpha_0 \rightarrow 001$, $\alpha_1 \rightarrow 01$,
$\alpha_2 \rightarrow 1$, and $\alpha_3 \rightarrow 000$



The longest possible length of codeword is: (N-1) = 3

Compression Ratio R = $\log_2(N)$ / $L_{average}(\alpha_i)$ = 2.0 / 1.75 = 1.1428

Compression Fundamentals
Adaptive Variable Length Coding



Compression Fundamentals

Source With Memory

- Real information sources exhibit local dependence between message symbols.
- Conditional Probability is given by:

$$P_{i/j} = P_i \cdot P_{j/i}$$

$$I_{i/j} = I_i - I_{j/i}$$

$$H_{i/j} = H_i - H_{j/i}$$

$$H_{1st\ order} \geq H_{2nd\ order} \geq \ldots H_{ith\ order} \geq H_{ith+1\ order} \geq \ldots$$

for very large value of i; $H_{ith\ order} \rightarrow 0$

Compression Fundamentals

Source With Memory

Pair (α_i, α_j)	$P_{j/i}$	I_j (bits)	$I_{j/i}$ (bits)
(e,)	**0.0341**	**2.77**	**1.77**
(,t)	**0.0264**	**3.76**	**2.48**
(t,h)	**0.0239**	**4.55**	**1.63**
(h,e)	**0.0223**	**3.11**	**1.94**
(s,)	**0.0197**	**2.77**	**1.57**

$$H_{1st\ order} \geq H_{2nd\ order} \geq \ldots H_{ith\ order} \geq H_{ith+1\ order} \geq \ldots$$

for very large value of i; $H_{ith\ order} \rightarrow 0$

Compression Fundamentals

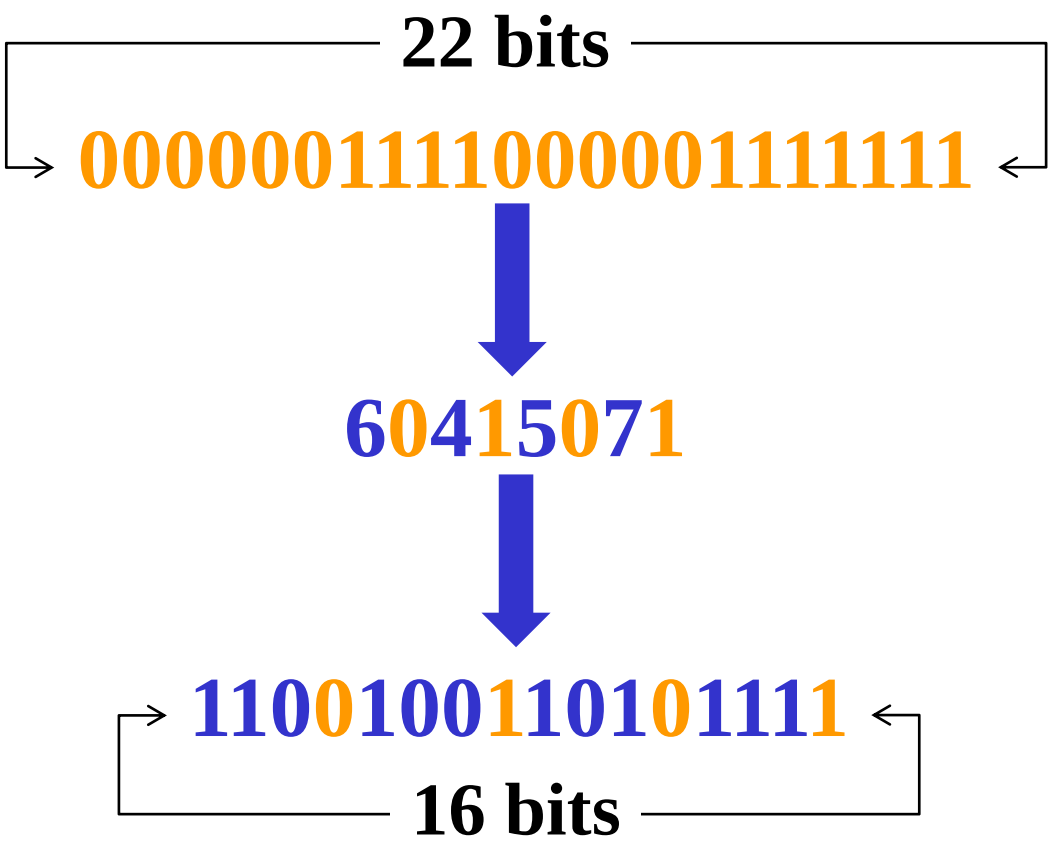
Sample entropy per symbol

Sample type	1st order H_i	Joint $H_{i,j}$	Conditional $H_{j/i}$
Arabic	**4.21**	**3.99**	**3.77**
English	**4.03**	**3.67**	**3.32**
TV Signal	**4.39**	**3.15**	**1.91**
Average	**4.21**	**3.61**	**3.00**

$$H_{1st\ order} \geq H_{2nd\ order} \geq \ldots H_{ith\ order} \geq H_{ith+1\ order} \geq \ldots$$

for very large value of i; $H_{ith\ order} \rightarrow 0$

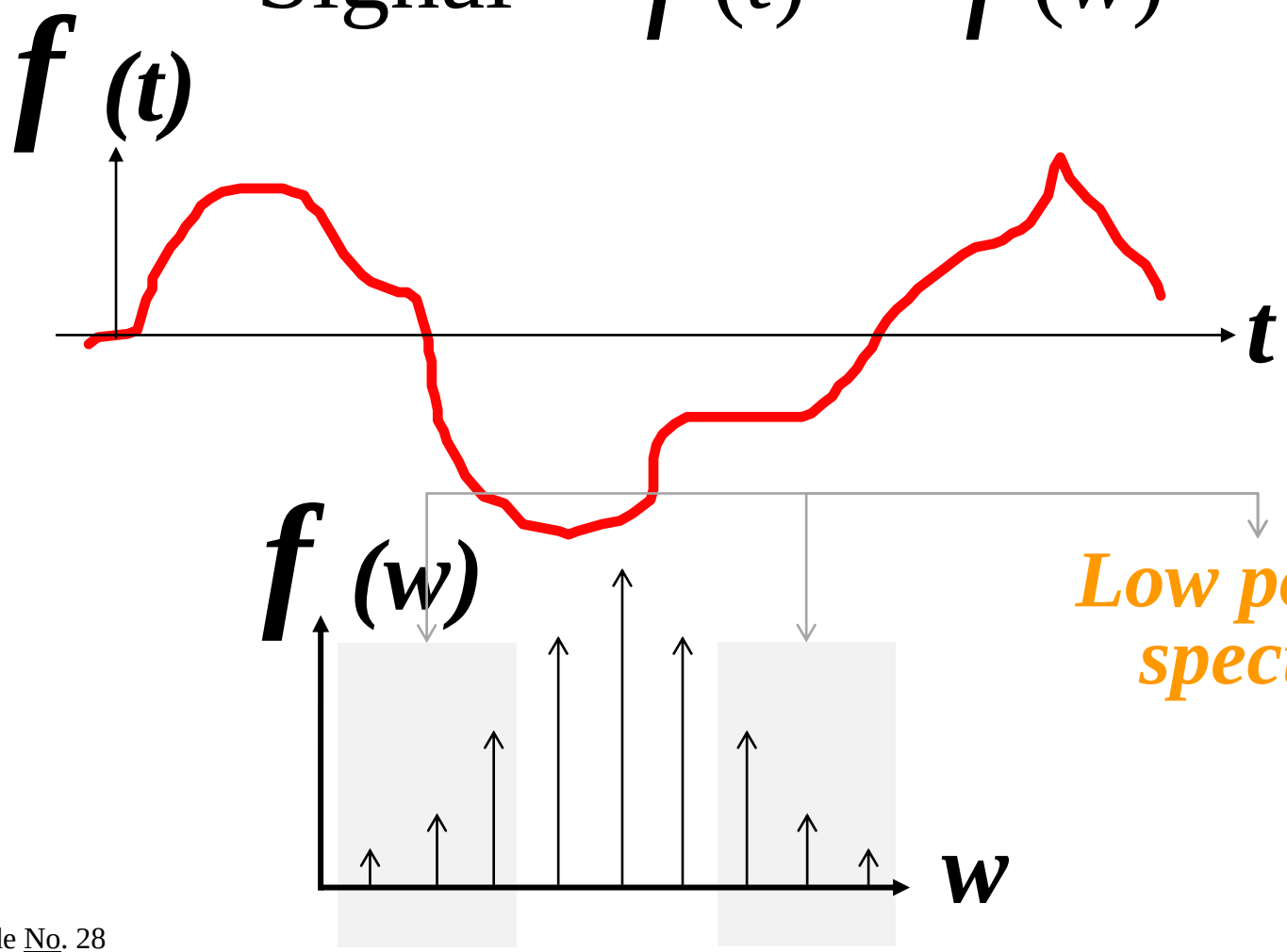
Compression Fundamentals
Structure – Generic Compression



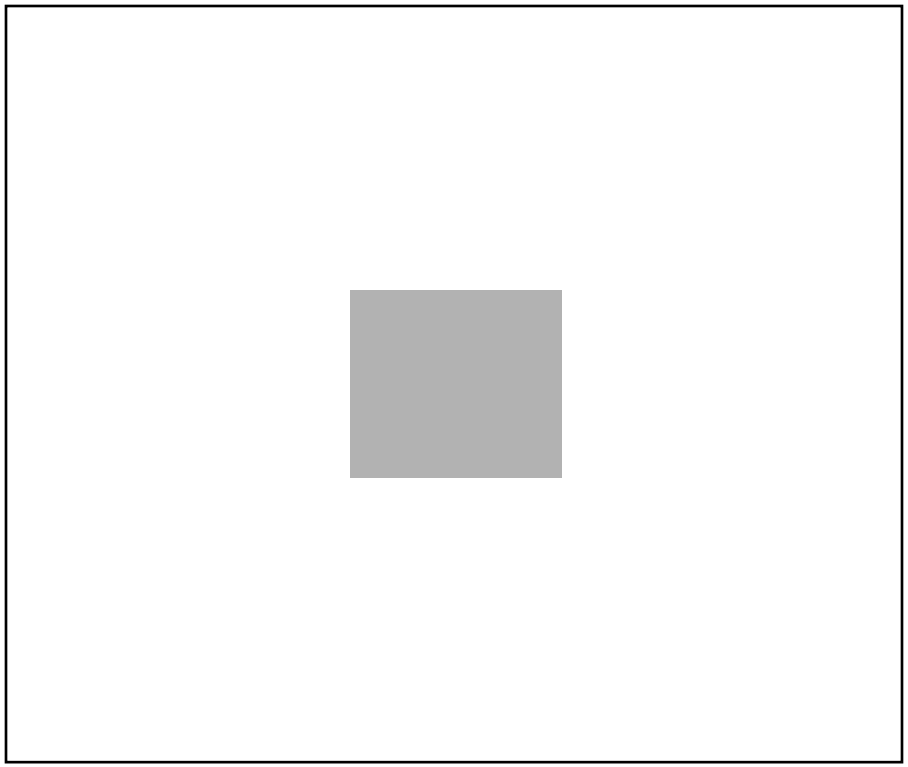
$$\textit{Compression Ratio} = \frac{22}{16} = 1.375$$

Compression Fundamentals
Transform Coding

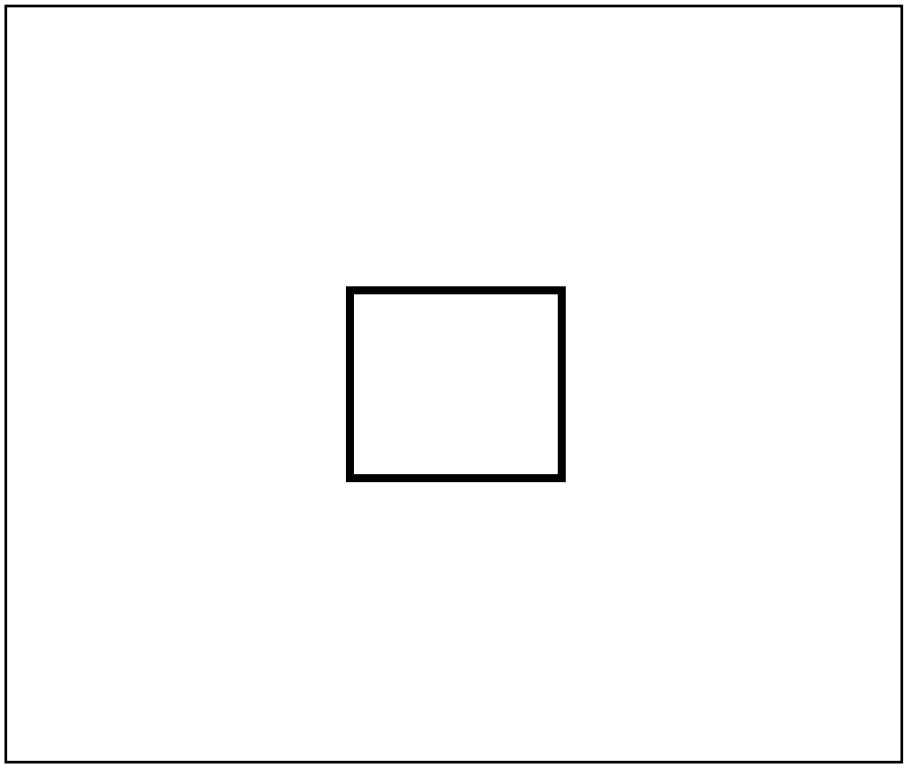
Signal = $f(t) = f(w)$



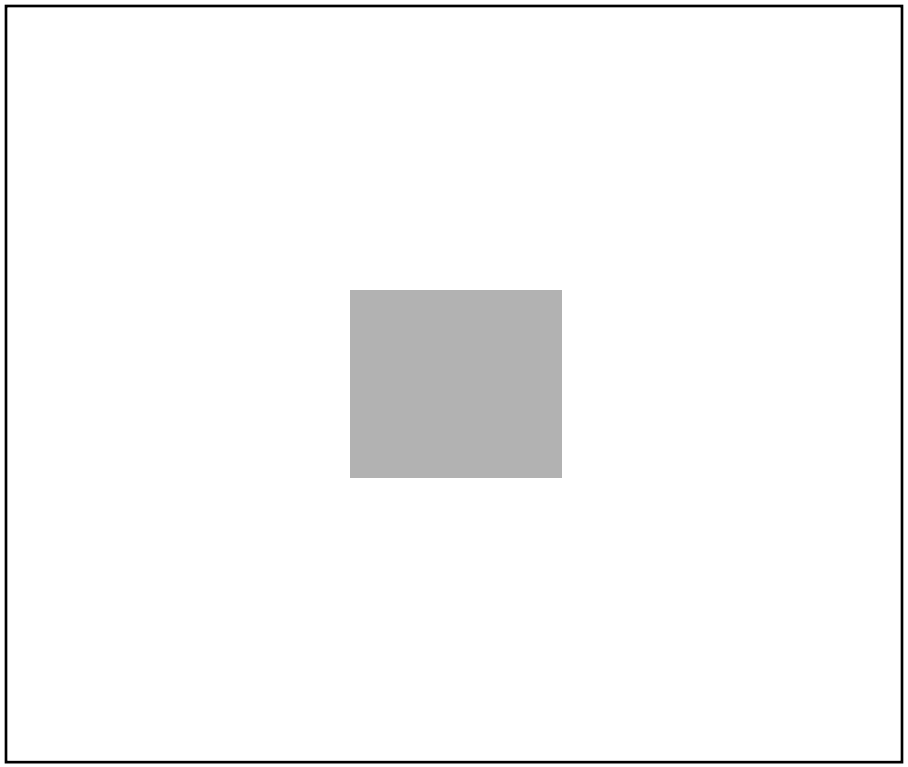
Compression Fundamentals
3rd Generation Compression



Compression Fundamentals
3rd Generation Compression



Compression Fundamentals
3rd Generation Compression



Compression Fundamentals
Dictionary Coding

1. **Matching:** The input sequence of source symbols is matched with a longest word stored in the dictionary.
2. **Coding:** The words are coded by equal length binary codeword.
3. **Pruning:** If the dictionary reached to its maximum size, prune the oldest word added to the dictionary.
4. **New Dictionary Word:** The last unmatched character of the input sting is appended to the matched word to form a new dictionary word**.**
5. The process repeated to EOF.

Dictionary

Word No.	The Word	Word Codeword
0	NUL	00 0000 0000
1	SOH	00 0000 0001
2	STX	00 0000 0010
...		
...		
48	0	00 0011 0001
49	1	00 0011 0001
...		
...		
97	a	00 0011 0001
98	b	00 0011 0010
..		
255	is	00 1111 1111
256	the	001000 0000
...		
...		
1023	There	11 1111 1111

Dictionary size is 1024 words