

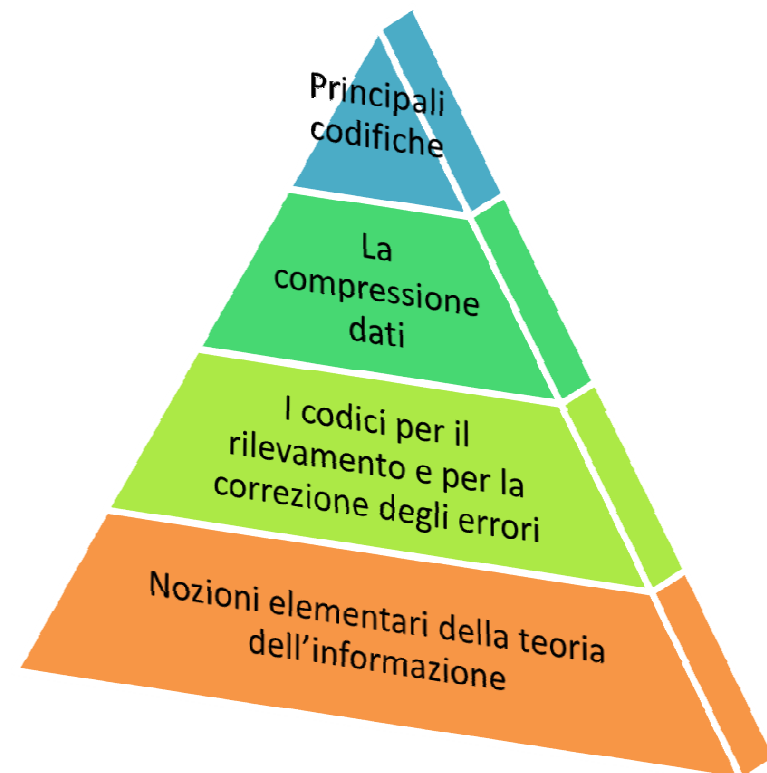


Architettura degli elaboratori

Fondamenti di Teoria
dell'Informazione

ARGOMENTI DELLA LEZIONE

- ☐ Nozioni elementari della teoria dell'informazione
- ☐ I codici per il rilevamento e per la correzione degli errori
- ☐ La compressione dati
- ☐ Principali codifiche





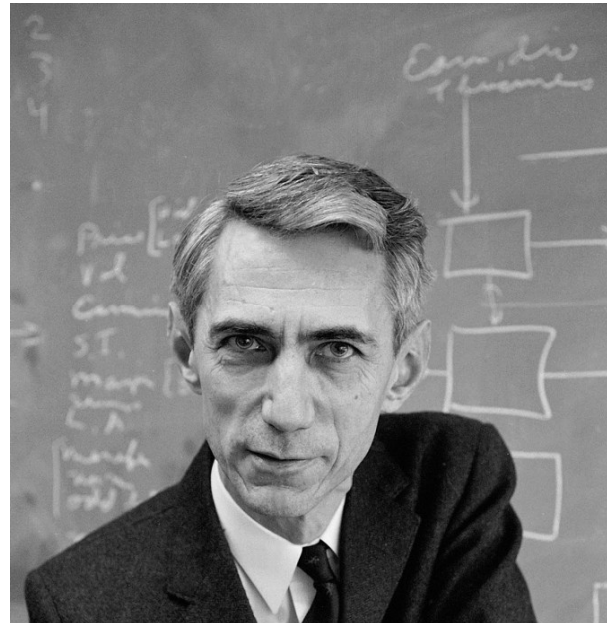
Architettura degli Elaboratori

Principi

TEORIA DELL'INFORMAZIONE

Teoria dell'informazione

- ❑ *A Symbolic Analysis of Relay and Switching Circuits* (1937)
 - ❑ Si dimostra che le reti digitali possono essere utilizzate per risolvere espressioni booleane e viceversa
- ❑ *A Mathematical Theory of Communication* (1948)
 - ❑ Definizione delle componenti fondamentali delle comunicazioni digitali



CLAUDE SHANNON

TEORIA DELL'INFORMAZIONE

Teoria dell'informazione

- ❑ La teoria enunciata da Shannon pose l'attenzione su come riprodurre in un determinato punto, in modo esatto (o con una buona approssimazione), un **messaggio** emesso da un luogo differente e rappresentato con dei segnali (o simboli) e da qui propose i componenti essenziali di un **sistema di comunicazione**
- ❑ Shannon introdusse anche la possibilità che il canale fosse affetto da rumore; cioè si verifica un'alterazione che cambia, durante la trasmissione, il segnale originariamente inviato e quindi modifica il messaggio ricevuto dal destinatario



TEORIA DELL'INFORMAZIONE

Teoria dell'informazione

- ❑ Shannon non prese in considerazione il significato dei messaggi, ma nel definire la **sorgente** evidenziò che questa può essere definita come **un insieme di messaggi possibili**
- ❑ Lo scienziato statunitense comprese che in un sistema efficiente è importante garantire solamente che siano i singoli segnali a giungere a destinazione in maniera corretta
- ❑ Ad esempio usando come mezzo di comunicazione il telegrafo quello che bisogna rispettare è la successione di simboli, linee e punti, che devono giungere a destinazione (di cui, in ricezione, non si sa niente); mentre è irrilevante approfondire l'equivalenza tra i gruppi dei segni grafici e le lettere dell'alfabeto, compito che può essere demandata ai telegrafisti o a delle tabelle di conversione

Sistema di comunicazione
Acquisizione corretta dei dati

-... .. --- --. -. .- / -.. .-.-. .-.. .-.. / .-- .-. - ---

BISOGNA DECIDERE PRESTO

Sistema di comunicazione
Acquisizione non corretta dei dati

-... .. --- --. -. .- / -.. .-.-. .-.. .-.. / .-- .-. - ---

BISOGNA DECEDERE PRESTO



TEORIA DELL'INFORMAZIONE

Teoria dell'informazione

- ❑ L'elemento che differenzia le sorgenti (e che può creare una cattiva interpretazione al destinatario) è la diversa prevedibilità dei messaggi inviabili: il livello d'incertezza è più alto quanto più numerosi sono i messaggi che possono provenire da quella sorgente (supposto che tutti i messaggi abbiano uguale probabilità di essere trasmessi)
- ❑ Shannon pertanto definì **l'informazione come la misura della complessità di una sorgente** o, in maniera equivalente, la misura dell'imprevedibilità di un messaggio
- ❑ Inoltre affermò che la prevedibilità di un messaggio è inversamente proporzionata al numero di messaggi erogabili dalla sorgente; e, viceversa, una sorgente è tanto più imprevedibile quanto più numerosi ed egualmente probabili sono i messaggi

Esempio

Una sorgente con solo due segnali e 32 messaggi richiedono da 0, rappresentato con $(00000)_2$, a 31, con $(11111)_2$.

Con il sistema decimale per rappresentare 32 messaggi bastano due cifre, ma i simboli potenzialmente trasmettibili sono dieci, infatti si usa l'alfabeto decimale $\{0, \dots, 9\}$, e quindi occorrono altrettanti segnali fisici diversi

TEORIA DELL'INFORMAZIONE

Definizione di bit

- ❑ Nella teoria dell'informazione la quantità minima di informazione che serve a discernere tra due eventi equiprobabili è detta **bit**
- ❑ In informatica il bit è una cifra binaria, ovvero uno dei due simboli (*zero*, 0, e *uno*, 1) del sistema numerico binario

La base ottima, cioè quella che consente di rappresentare numeri con poche cifre (m_b minimo) e con un alfabeto composto da un numero ristretto di simboli (b minimo) si individua minimizzando b .

Considerando un numero N , allora sarà una funzione legata alla variabile b . Per trovare il minimo, effettueremo la derivazione della funzione ed individueremo con b_{min} il valore per cui si annulla

Per semplificare i calcoli utilizzeremo logaritmi naturali anziché quelli in base b .

Pertanto avremo che $N \leq b^{m_b}$ e $\log_b(N) = m_b$ per cui $m_b = \log_e(N) / \log_e(b)$, e quindi:

$$\frac{d(m_b b)}{db} = \frac{d((\log_e(N) / \log_e(b))b)}{db} = \frac{\log_e(N)}{\log_e(b)} - \frac{\log_e(N)b}{(\log_e(b))^2 b} = \frac{\log_e(N)(\log_e(b) - 1)}{(\log_e(b))^2}$$

Che si annulla per $b_{min}=e$



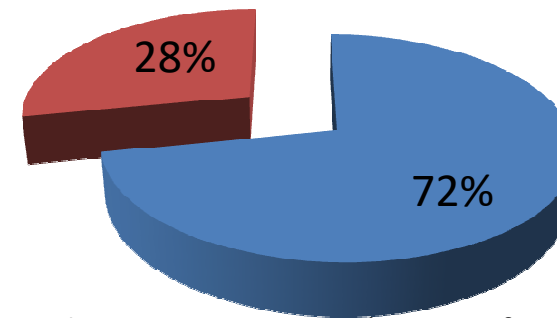
Architettura degli elaboratori

*Codici per il rilevamento e per
la correzione di errori*

TEORIA DELL'INFORMAZIONE

Errori

- ❑ Un sistema di telecomunicazione reale, come detto, è soggetto a **errori** dovuti alla presenza di rumore sul canale. Analogamente i supporti di conservazione dei dati sono realizzati con materiali soggetti a degrado o a malfunzionamenti tecnici che comportano la perdita d'informazione
- ❑ L'alterazione dei dati può essere mitigata con delle **informazioni ridondanti**, cioè usando più bit di quelli necessari per rappresentare il messaggio (e quindi con un incremento del tempo richiesto per la trasmissione e uno spazio di occupazione più ampio nel caso di archiviazione), e ricorrendo a tecniche e codici per il rilevamento e la correzione degli errori



- Errori dovuti a programmi (virus, software mal progettato), componenti fisici (danneggiamento hardware, perdita di dati dai supporti o in trasmissione), personale (negligenza, dolo)
- Sicurezza Tecnologica inadeguata



TEORIA DELL'INFORMAZIONE

Rilevamento: il bit di parità

- ☐ Un semplice codice per il rilevamento degli errori è il **bit di parità**
- ☐ Questa tecnica interviene su una parola di lunghezza determinata (*word*) da cui si deriva la parola di codice (*codeword*) che ha un bit supplementare accodato il cui valore è 1 se il numero di 1 presenti nella parola è dispari e 0 altrimenti

Bit di parità

Word	Numero di 1	Bit di parità	Codeword
10101100	4	0	101011000
11101111	7	1	111011111



TEORIA DELL'INFORMAZIONE

Rilevamento: il bit di parità

- ❑ Questo metodo individua solamente parole con un numero di errori dispari e non rileva la posizione in cui si è verificato l'errore

Bit di parità

Codeword	Numero di errori	Veridicità del valore	Bit di parità
101011000	0	Corretto	Corretto
111011000	1	Errato	Errato
001010000	2	Errato	Corretto
100000000	3	Errato	Errato
000000000	4	Errato	Corretto



TEORIA DELL'INFORMAZIONE

Rilevamento e correzione: codice lineare

- ❑ Una strategia di rilevamento e di correzione degli errori è il **codice lineare**, in cui una parola di correzione (linear word) è ottenuta dalla somma (or esclusivo) tra due generiche parole (word 1 e word 2)
- ❑ Il rilevamento si effettua in ricezione svolgendo l'or esclusivo delle due parole e confrontando il risultato con la parola di correzione
- ❑ Ad esempio se occorre un errore al quarto bit meno significativo della prima word (111101), si ricostruisce l'informazione corrotta eseguendo l'or esclusivo delle altre parole: (010010 XOR 100111)=110101

Codice Lineare

Word1	Word2	Linear Word
110101	010010	100111



Architettura degli elaboratori

Compressione dati



TEORIA DELL'INFORMAZIONE

Compressione dati digitali

- ❑ La teoria della compressione dei dati riguarda la trasmissione delle informazioni binarie e la conservazione dei documenti elettronici
- ❑ Nel primo caso si cerca di ottimizzare il canale di comunicazione: una riduzione del numero di dati permette il trasferimento di più informazioni in un intervallo temporale prestabilito (di solito si valuta in KB per secondi)
- ❑ Riguardo alla conservazione si sfrutta al meglio lo spazio di memorizzazione offerto dai supporti digitali (nel 2015 sono state stimate circa 2.5 miliardi di caselle di posta elettronica attive e il numero di bit costituenti i documenti elettronici prodotti negli ultimi anni ha superato l'Exabyte)

Produzione digitale

147GB per persona al giorno

90% dei contenuti digitali esistenti creati negli ultimi 5 anni

463EXABYTE creati nel 2025

95000000 foto e video condivisi su Instagram ogni giorno

306.4miliardi di email ogni giorno

500milioni di *twitter* ogni giorno

TEORIA DELL'INFORMAZIONE

Compressione dati digitali

- ❑ In letteratura sono riportati diversi algoritmi utilizzati per la compressione dati e si suddividono in due categorie:
- ❑ **senza perdita di informazione (*lossless*)**, in cui, cioè, il messaggio ricostruito dopo il processo di compressione risulta essere uguale a quello originale. Si ricorre a questa tipologia di algoritmi per rappresentare testo, immagini mediche o militari; dove l'integrità dell'informazione è un aspetto determinante.
- ❑ **con perdita di informazione (*lossy*)** nei quali a fronte di una maggiore riduzione dei dati si verifica una alterazione della fedeltà del dato originario. Ciò è quanto avviene su internet per immagini (es.: *jpeg*), filmati (es.: *mp4*) o suoni (es.: *mp3*)

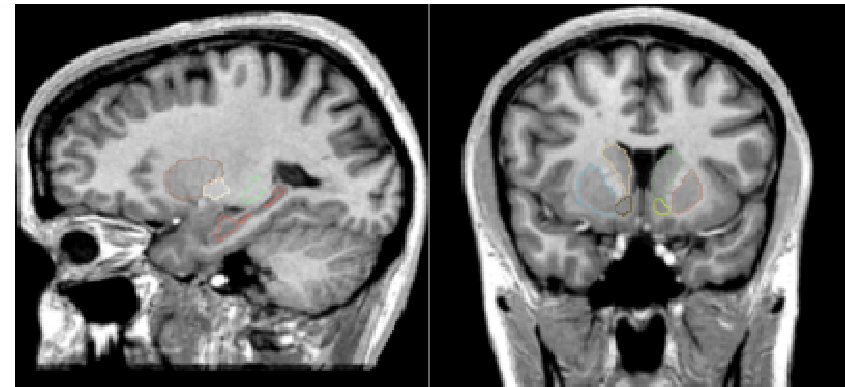


Immagine senza perdita di dati



Immagine originale

Immagine
compressione con
10% perdita di dati

Immagine
compressione con
40% perdita di dati

TEORIA DELL'INFORMAZIONE

Compressione dati digitali

- ❑ Esistono numerose tecniche per ottenere una riduzione dell'informazione, ma principalmente due sono i rami in cui è possibile classificarle: il **modello probabilistico**, in cui si opera sui dati senza alcuna considerazione della tipologia d'informazione processata, ma si considera la frequenza con cui questi si ripetono; e quelle che invece operano **sul concetto di contesto**, esaminando il modo e la posizione in cui l'informazione si ripete nel messaggio (algoritmo a dizionario, LZ)

A B BC BCA BA BCAA BCAAB

Output	Indice	Stringa codificata
(0,A)	1	A
(0,B)	2	B
(2,C)	3	BC
(3,A)	4	BCA
(2,A)	5	BA
(4,A)	6	BCAA
(6,B)	7	BCAAB



TEORIA DELL'INFORMAZIONE

Compressione dati digitali: RLE

- ❑ Un semplice algoritmo di codifica senza perdita di dati basato sulla tecnica di contesto è il **Run Length Encoding** (RLE) che ricerca una serie consecutiva di elementi uguali (run) e li codifica riportando solo il primo seguito da un contatore che indica quante volte è ripetuto..

ALGORTIMO

```
for (int i = 0; str[i]; i++)  
{  
    count = 1;  
    while (str[i] == str[i + 1])  
    {  
        count++;  
        i++;  
    }  
    encoding += str[i]+to_string(count);  
}
```



TEORIA DELL'INFORMAZIONE

Compressione dati digitali: Huffman

- ❑ L'**algoritmo di Huffman**, invece, è un procedimento di compressione dati senza perdita di informazione di tipo entropico. Il metodo valuta la probabilità di ciascuna parola in un flusso informativo; in seguito si costruiscono dei codici univoci e a lunghezza variabile da associare a ogni parola seguendo la logica che a parole frequenti si accomuna un codice con pochi simboli e viceversa.

ALGORITMO

1. Selezionare dall'insieme dei valori della distribuzione di probabilità due parole con probabilità minima assegnando ai rispettivi codici i simboli 0:: e 1:: e segnare marcate le due distribuzioni di probabilità;
2. Sostituire ai due valori la loro somma, ottenendo così una nuova distribuzione di probabilità;
3. Iterare il procedimento dal punto 1 fino a quando tutte le distribuzioni di probabilità sono marcate.



TEORIA DELL'INFORMAZIONE

Compressione dati digitali: Huffman

MESSAGGIO (parole di tre bit)

101 101 101 000 011 100 000 000 010 100 100 001 001 011 101 101 000 101

DISTRIBUZIONE PROBABILITA'

$P(000) = 4/18 = 0.22$

$P(001) = 2/18 = 0.11$

$P(010) = 1/18 = 0.06$

$P(011) = 2/18 = 0.11$

$P(100) = 3/18 = 0.17$

$P(101) = 6/18 = 0.33$

TEORIA DELL'INFORMAZIONE

Compressione dati digitali: Huffman

$P(000) = 4/18 = 0.22$
 $P(001) = 2/18 = 0.11$
 $P(010) = 1/18 = 0.06$
 $P(011) = 2/18 = 0.11$
 $P(100) = 3/18 = 0.17$
 $P(101) = 6/18 = 0.33$

$P(A) = 0.17$

$P(B) = 0.28$

$P(C) = 0.39$

$P(D) = 0.61$

$P(Z) = 1$

Codifica Huffman:

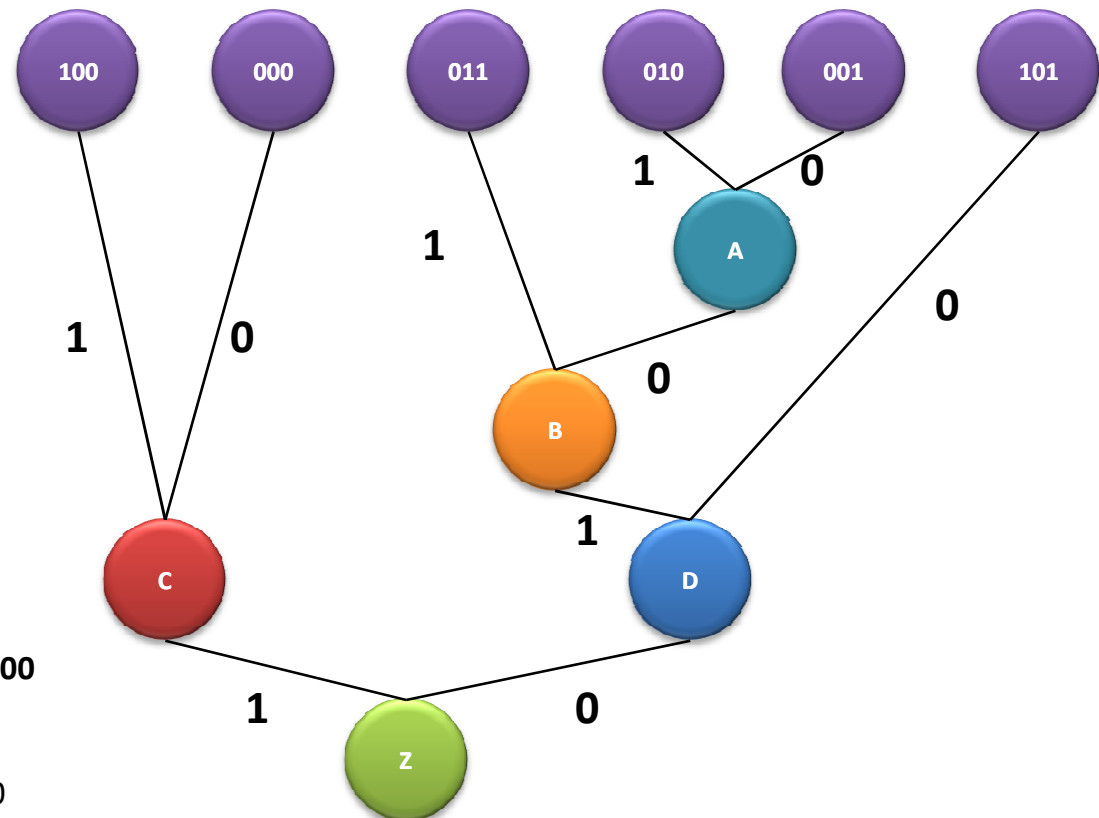
000=10, 001=0100, 010=0101, 011=011, 100=11, 101=00

Lunghezza media messaggi: 2.83

Codifica stringa:

0000000101111101001011110100010001100001000

Tasso di compressione: 18.52%





Architettura degli elaboratori

Principali codifiche



TEORIA DELL'INFORMAZIONE

Impiego del bit nei sistemi digitali

- ☐ Un messaggio può essere identificato da una sequenza (stringa) di bit detta **parola** (*word*)
- ☐ Ad ogni parola può essere associato un significato in base al codice prescelto
 - ☐ Carattere
 - ☐ Punto di colore
 - ☐ Numero
 - ☐ Un suono

TEORIA DELL'INFORMAZIONE

Formato testuale

- ❑ Nel 1963 un gruppo di lavoro per la standardizzazione, American National Standard Institute, propose per rappresentare i **caratteri** nelle comunicazioni fra telescriventi (poi adottato anche in campo informatico) lo **American Standard Code for Information Interchange** (ASCII). Si trattava di una codifica a 7 bit (a cui si aggiunse il bit di parità, per evidenziare eventuali errori durante la trasmissione) in grado di rappresentare 128 caratteri tra cui le cifre decimali, le lettere maiuscole e minuscole, i simboli aritmetici, i segni di punteggiatura, e 33 simboli di controllo

Decimal - Binary - Octal - Hex – ASCII
Conversion Chart

Decimal	Binary	Octal	Hex	ASCII	Decimal	Binary	Octal	Hex	ASCII	Decimal	Binary	Octal	Hex	ASCII	Decimal	Binary	Octal	Hex	ASCII
0	00000000	000	00	NUL	32	00100000	040	20	SP	64	01000000	100	40	@	96	01100000	140	60	`
1	00000001	001	01	SOH	33	00100001	041	21	!	65	01000001	101	41	A	97	01100001	141	61	a
2	00000010	002	02	STX	34	00100010	042	22	"	66	01000010	102	42	B	98	01100010	142	62	b
3	00000011	003	03	ETX	35	00100011	043	23	#	67	01000011	103	43	C	99	01100011	143	63	c
4	00000100	004	04	EOT	36	00100100	044	24	\$	68	01000100	104	44	D	100	01100100	144	64	d
5	00000101	005	05	ENQ	37	00100101	045	25	%	69	01000101	105	45	E	101	01100101	145	65	e
6	00000110	006	06	ACK	38	00100110	046	26	&	70	01000110	106	46	F	102	01100110	146	66	f
7	00000111	007	07	BEL	39	00100111	047	27	'	71	01000111	107	47	G	103	01100111	147	67	g
8	00001000	010	08	BS	40	00101000	050	28	(72	01001000	110	48	H	104	01101000	150	68	h
9	00001001	011	09	HT	41	00101001	051	29)	73	01001001	111	49	I	105	01101001	151	69	i
10	00001010	012	0A	LF	42	00101010	052	2A	*	74	01001010	112	4A	J	106	01101010	152	6A	j
11	00001011	013	0B	VT	43	00101011	053	2B	+	75	01001011	113	4B	K	107	01101011	153	6B	k
12	00001100	014	0C	FF	44	00101100	054	2C	,	76	01001100	114	4C	L	108	01101100	154	6C	l
13	00001101	015	0D	CR	45	00101101	055	2D	-	77	01001101	115	4D	M	109	01101101	155	6D	m
14	00001110	016	0E	SO	46	00101110	056	2E	.	78	01001110	116	4E	N	110	01101110	156	6E	n
15	00001111	017	0F	SI	47	00101111	057	2F	/	79	01001111	117	4F	O	111	01101111	157	6F	o
16	00010000	020	10	DLE	48	00110000	060	30	0	80	01100000	120	50	P	112	01110000	160	70	p
17	00010001	021	11	DC1	49	00110001	061	31	1	81	01100001	121	51	Q	113	01110001	161	71	q
18	00010010	022	12	DC2	50	00110010	062	32	2	82	01100010	122	52	R	114	01110010	162	72	r
19	00010011	023	13	DC3	51	00110011	063	33	3	83	01100011	123	53	S	115	01110011	163	73	s
20	00010100	024	14	DC4	52	00110100	064	34	4	84	01010100	124	54	T	116	01110100	164	74	t
21	00010101	025	15	NAK	53	00110101	065	35	5	85	01010101	125	55	U	117	01110101	165	75	u
22	00010110	026	16	SYN	54	00110110	066	36	6	86	01010110	126	56	V	118	01110110	166	76	v
23	00010111	027	17	ETB	55	00110111	067	37	7	87	01010111	127	57	W	119	01110111	167	77	w
24	00011000	030	18	CAN	56	00111000	070	38	8	88	01011000	130	58	X	120	01111000	170	78	x
25	00011001	031	19	EM	57	00111001	071	39	9	89	01011001	131	59	Y	121	01111001	171	79	y
26	00011010	032	1A	SUB	58	00111010	072	3A	:	90	01011010	132	5A	Z	122	01111010	172	7A	z
27	00011011	033	1B	ESC	59	00111011	073	3B	;	91	01011011	133	5B	[123	01111011	173	7B	{
28	00011100	034	1C	FS	60	00111100	074	3C	<	92	01011100	134	5C	\	124	01111100	174	7C	
29	00011101	035	1D	GS	61	00111101	075	3D	=	93	01011101	135	5D]	125	01111101	175	7D	}
30	00011110	036	1E	RS	62	00111110	076	3E	>	94	01011110	136	5E	^	126	01111110	176	7E	~
31	00011111	037	1F	US	63	00111111	077	3F	?	95	01011111	137	5F	_	127	01111111	177	7F	DEL

TEORIA DELL'INFORMAZIONE

Codifica immagine

- ❑ Un'altra codifica è quella dei **punti di colore** grazie ai quali è possibile definire una immagine digitale. Il primo a derivare da una fotografia analogica un surrogato digitale fu Russell Kirsch, un ingegnere informatico statunitense che nel 1957 lavorava al National Bureau of Standards. Kirsch digitalizzò con una risoluzione colore bitonale (bianco e nero) una fotografia del figlio Walden, con un rudimentale sistema di acquisizione fotoelettrico (*scanner*).
- ❑ Una **immagine digitale** è una matrice di punti di informazione luminosa detti *pixel* (contrazione di *picture element*). L'informazione luminosa è un numero intero non negativo (rappresentato in binario per essere gestito dagli elaboratori elettronici) e definisce la risoluzione radiometrica (o **profondità del colore**).

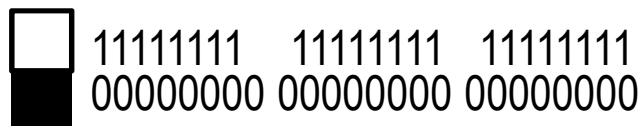
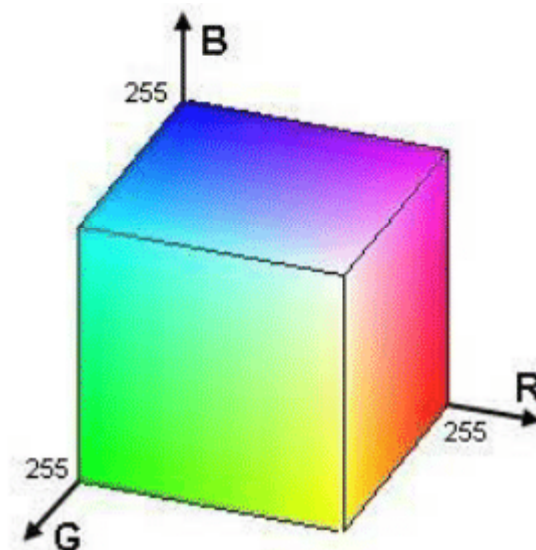


00010000 10110000 01010000

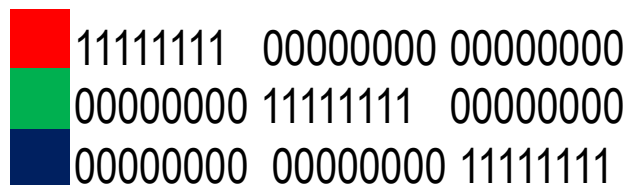
TEORIA DELL'INFORMAZIONE

Formato immagine

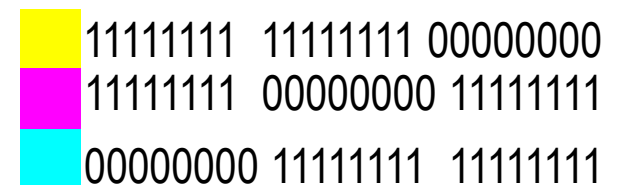
- ❑ Un **modello colore** specifica i toni cromatici con modalità standardizzate, che fanno riferimento ad un sistema di coordinate tridimensionali, o meglio ad un suo sotto-spazio, nel quale ogni colore è rappresentato da un punto
- ❑ Nel modello RGB, ad esempio, ogni tinta è definita mediante le intensità dei tre colori primari che lo compongono: il canale Rosso (R), il canale Verde (G) e il canale Blu (B). Il modello è basato su un sistema di coordinate cartesiane, in cui il sotto-spazio di riferimento è un cubo di lato unitario (tutti i valori si intendono normalizzati in modo da ricadere nell'intervallo $[0,1]$ o in binario per pixel da 24 bit da $[0..255]$)



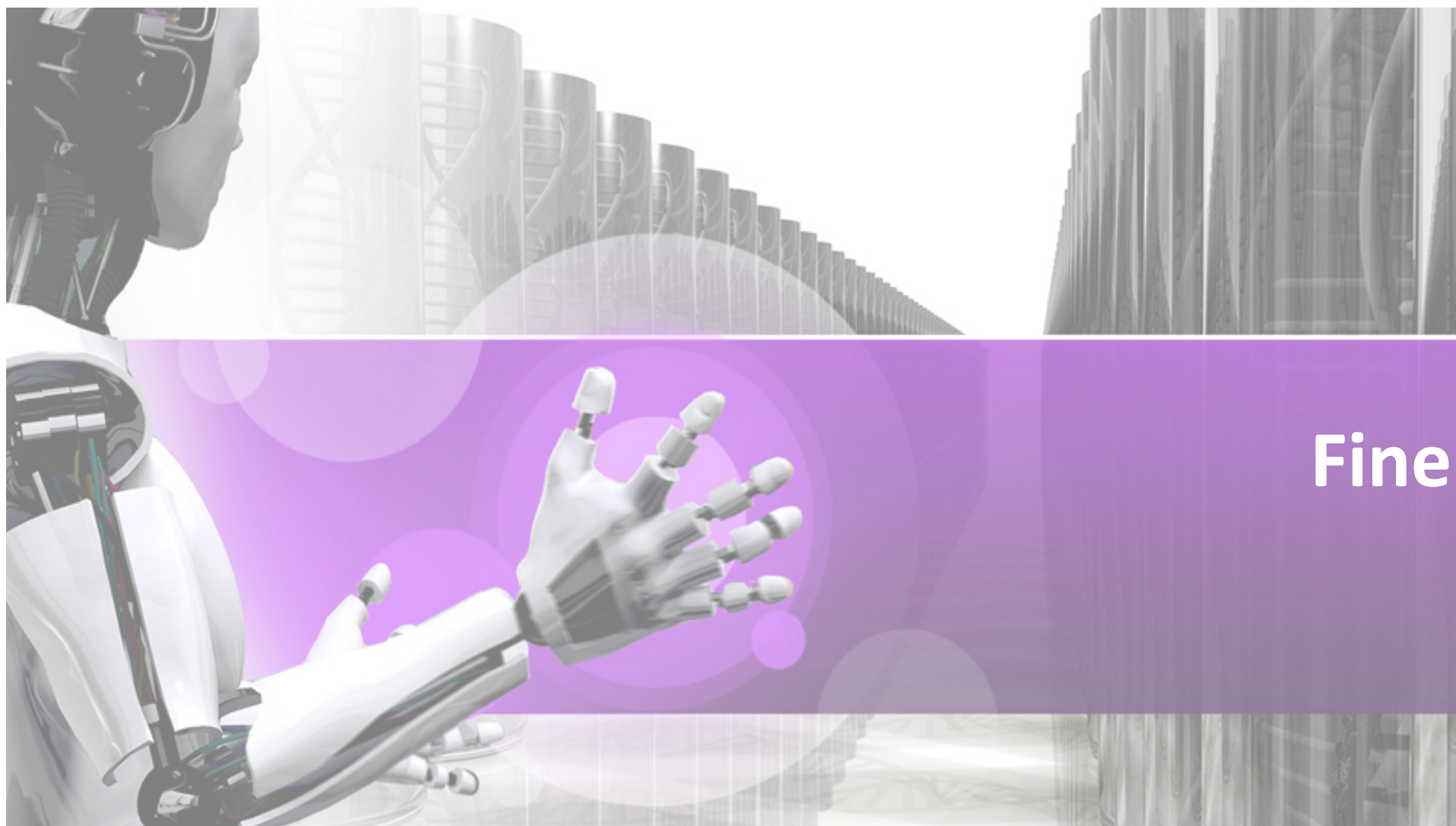
11111111 11111111 11111111
00000000 00000000 00000000



11111111 00000000 00000000
00000000 11111111 00000000
00000000 00000000 11111111



11111111 11111111 00000000
11111111 00000000 11111111
00000000 11111111 11111111



Fine