



dati

Alberto Ferrari – Analisi dei Dati

parliamo di ...

- dati – informazioni
- dati – informazioni – conoscenza
- analogico – digitale
- rappresentazioni digitali



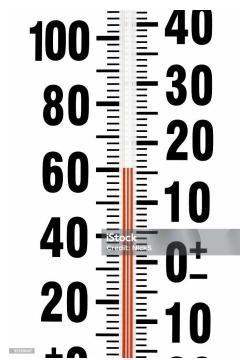
dato o informazione ?

50

Alberto Ferrari – Analisi dei Dati

contesto - interpretazione





60



Alberto Ferrari – Analisi dei Dati

dato - informazione

- un dato può essere interpretato in modi diversi dipendentemente dal contesto
- ... e rappresentare informazioni diverse



60





**pericolo: buche
sulla strada**

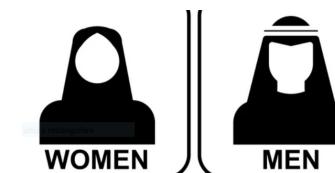


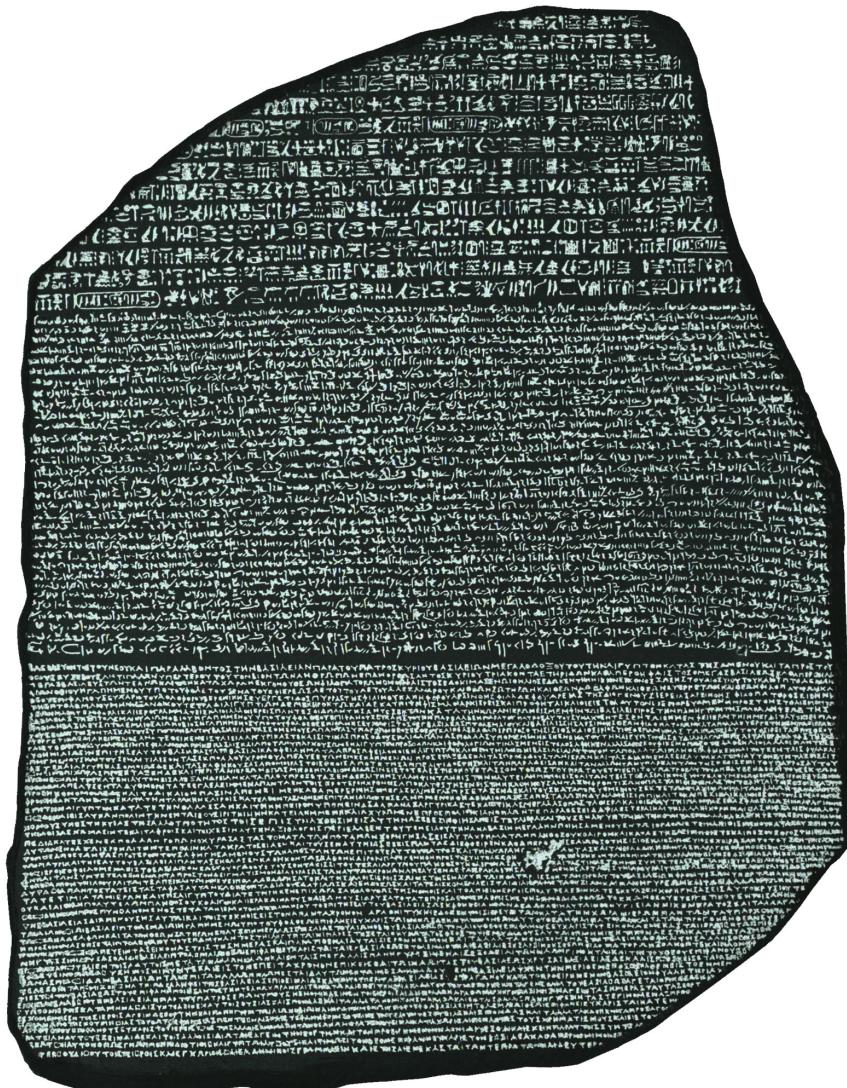
**danger: potholes
in the road**

κίνδυνος: λακκούβες στο δρόμο

informazione e rappresentazione (dato)

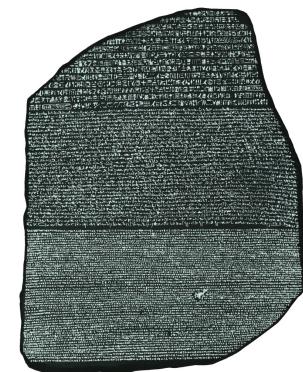
- una stessa informazione può essere rappresentata da dati diversi





la stele di Rosetta

- decifrare i geroglifici
- grazie a una pietra ritrovata dalle truppe di Napoleone, Jean-François Champollion poté decifrare la scrittura di una delle civiltà più importanti della storia
- la pietra contiene un decreto del faraone Tolomeo V scritto negli indecifrabili alfabeti **geroglifico** e **demotico** (derivato dal primo) e **greco antico**



Alberto Ferrari – Analisi dei Dati

dati – informazioni - conoscenza

- ***dati***

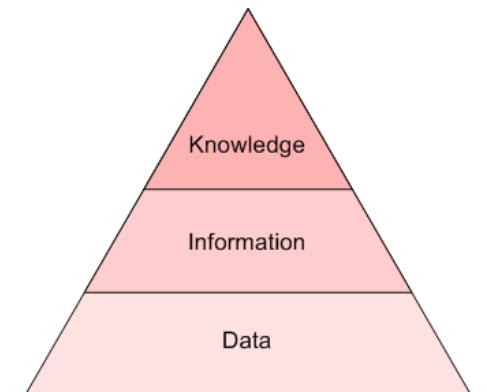
- registrazione di eventi che accadono, tutto ciò che può essere misurato o classificato può essere convertito in dati

- ***informazioni***

- studio e analisi dei dati per
 - capire la natura degli eventi
 - prendere decisioni
 - fare previsioni

- ***conoscenza***

- le informazioni vengono convertite in un insieme di regole per comprendere meglio alcuni meccanismi e fare previsioni sul evoluzione di alcuni eventi



analogico e digitale

- che cosa significa «rappresentazione analogica dei un dato» ?
- che cosa significa «rappresentazione digitale di un dato» ?



analogico - digitale

- la rappresentazione ***analogica*** di un'informazione si basa su un **insieme continuo di valori**
 - i dati da elaborare e trasmettere sono rappresentati da grandezze fisiche che assumono **infiniti valori**
- la rappresentazione ***digitale*** si basa su un **insieme discreto di valori**
 - i dati da elaborare e trasmettere non assumono tutti i valori di un intervallo e sono rappresentabili con simboli in **numero finito**

precisione

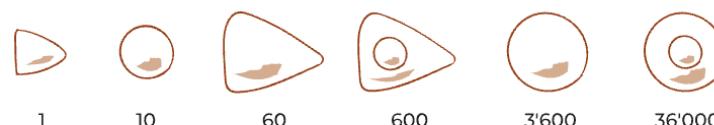
- è più precisa la rappresentazione analogica o quella digitale?



Alberto Ferrari – Analisi dei Dati

sistema numerico non posizionale

- **il valore di un simbolo non dipende dalla posizione** che occupa nel numero
 - ogni simbolo ha **sempre lo stesso valore**
 - si ripetono o si combinano i simboli
- **numeri romani:** $X = 10, L = 50, C = 100 \dots$
 - $XIV, X = 10, I = 1, V = 5$
 - il valore totale si ottiene con regole additive o sottrattive, ma non perché "I" è in una posizione particolare
- ogni simbolo rappresenta sempre la stessa quantità
- non esiste il concetto di zero
- è difficile rappresentare numeri grandi o fare calcoli complessi



Simboli numerici del sistema sumero.

sistema numerico posizionale

- il valore di una cifra dipende dalla posizione che occupa nel numero
- ogni cifra ha un **valore intrinseco** (0–9, ad esempio) e un **valore posizionale** (unità, decine, centinaia...).
- **decimale** (base 10): $343 = 3 \times 10^2 + 4 \times 10^1 + 3 \times 10^0$
- il valore di una cifra dipende dalla posizione
- include sempre lo zero
- è molto più efficiente per i calcoli

Indian numerals

0	1	2	3	4	5	6	7	8	9
०	१	२	३	४	५	६	७	८	९

digitale - binario

- il sistema binario è un sistema numerico posizionale in base 2
- utilizza solo due simboli (0 e 1)
- i numeri espressi nel sistema binario sono definiti "numeri binari"

Tabella di conversione bit byte kilobyte megabyte					
	bit	byte	Kilobyte	Megabyte	Gigabyte
bit	1				
byte	8	1			
Kilobyte	8,192	1,024	1		
Megabyte	8,388,608	1,048,576	1,024	1	
Gigabyte	8,589,934,592	1,073,741,824	1,048,576	1,024	1
Terabyte	8,796,093,022,208	1,099,511,627,776	1,073,741,824	1,048,576	1,024
Petabyte	9,007,199,254,740,990	1,125,899,906,842,620	1,099,511,627,776	1,073,741,824	1,048,576
Exabyte	9,223,372,036,854,780,000	1,152,921,504,606,850,000	1,125,899,906,842,620	1,099,511,627,776	1,073,741,824
Zettabyte	9,444,732,965,739,290,000,000	1,180,591,620,717,410,000,000	1,152,921,504,606,850,000	1,125,899,906,842,620	1,099,511,627,776

sistema binario

- **esempio: numero binario 1011_2**
- valore **decimale** \Rightarrow moltiplicare ogni cifra per la **potenza di 2** corrispondente alla sua posizione:
 - $1 \times 2^3 = 8$
 - $0 \times 2^2 = 0$
 - $1 \times 2^1 = 2$
 - $1 \times 2^0 = 1$
- sommare i risultati:

$$8 + 0 + 2 + 1 = 11$$

digit

- In informatica, termine equivalente all'italiano cifra.
- Si dice digitale un dispositivo che tratta informazioni rappresentate mediante cifre di un opportuno sistema di numerazione, in contrapposizione ad analogico che qualifica dispositivi che trattano informazioni espresse da una qualche grandezza fisica che sia funzione continua della entità da rappresentare.

Enciclopedia Treccani

bit

- contrazione di **bi**(nary) (digi)**t** «cifra binaria».
- 1. In teoria dell'informazione, l'unità di misura dell'informazione, corrispondente alla scelta tra due sole alternative possibili, ugualmente probabili (indicate, per es., con i simboli 0 e 1).
- 2. In informatica, sinonimo di **cifra binaria**: bit di informazione, di controllo, cifra (binaria) di informazione, di controllo

Enciclopedia Treccani

caratteri e testo

- necessaria **convenzione** per codifica numerica (**binaria**) dei caratteri
- codifica **ASCII** (American Standard **C**ode for **I**nformation **I**nterchange)
 - inizialmente 7 bit $\Rightarrow 2^7 = 128$ caratteri
- caratteri **alfanumerici**: *lettere maiuscole, minuscole, numeri, spazio*
- simboli e **punteggiatura**: @, #, ...
- caratteri di **controllo** (*non tutti visualizzabili*):
TAB, LF, CR, BELL ecc.

ascii table

Byte	Cod.	Char	Byte	Cod.	Char	Byte	Cod.	Char	Byte	Cod.	Char
00000000	0	Null	00100000	32	Spc	01000000	64	@	01100000	96	`
00000001	1	Start of heading	00100001	33	!	01000001	65	A	01100001	97	a
00000010	2	Start of text	00100010	34	"	01000010	66	B	01100010	98	b
00000011	3	End of text	00100011	35	#	01000011	67	C	01100011	99	c
00000100	4	End of transmit	00100100	36	\$	01000100	68	D	01100100	100	d
00000101	5	Enquiry	00100101	37	%	01000101	69	E	01100101	101	e
00000110	6	Acknowledge	00100110	38	&	01000110	70	F	01100110	102	f
00000111	7	Audible bell	00100111	39	,	01000111	71	G	01100111	103	g
00001000	8	Backspace	00101000	40	(01001000	72	H	01101000	104	h
00001001	9	Horizontal tab	00101001	41)	01001001	73	I	01101001	105	i
00001010	10	Line feed	00101010	42	*	01001010	74	J	01101010	106	j
00001011	11	Vertical tab	00101011	43	+	01001011	75	K	01101011	107	k
00001100	12	Form Feed	00101100	44	,	01001100	76	L	01101100	108	l
00001101	13	Carriage return	00101101	45	-	01001101	77	M	01101101	109	m
00001110	14	Shift out	00101110	46	.	01001110	78	N	01101110	110	n
00001111	15	Shift in	00101111	47	/	01001111	79	O	01101111	111	o
00010000	16	Data link escape	00110000	48	0	01010000	80	P	01110000	112	p
00010001	17	Device control 1	00110001	49	1	01010001	81	Q	01110001	113	q
00010010	18	Device control 2	00110010	50	2	01010010	82	R	01110010	114	r
00010011	19	Device control 3	00110011	51	3	01010011	83	S	01110011	115	s
00010100	20	Device control 4	00110100	52	4	01010100	84	T	01110100	116	t
00010101	21	Neg. acknowledge	00110101	53	5	01010101	85	U	01110101	117	u
00010110	22	Synchronous idle	00110110	54	6	01010110	86	V	01110110	118	v
00010111	23	End trans. block	00110111	55	7	01010111	87	W	01110111	119	w
00011000	24	Cancel	00111000	56	8	01011000	88	X	01111000	120	x
00011001	25	End of medium	00111001	57	9	01011001	89	Y	01111001	121	y
00011010	26	Substitution	00111010	58	:	01011010	90	Z	01111010	122	z
00011011	27	Escape	00111011	59	;	01011011	91	[01111011	123	{
00011100	28	File separator	00111100	60	<	01011100	92	\	01111100	124	
00011101	29	Group separator	00111101	61	=	01011101	93]	01111101	125	}
00011110	30	Record Separator	00111110	62	>	01011110	94	^	01111110	126	~
00011111	31	Unit separator	00111111	63	?	01011111	95	_	01111111	127	Del

tabella ascii estesa

- caratteri accentati + caratteri per grafici
 - code Page 437 per PC (DOS) in Nord America
 - possibile mischiare testo in inglese e francese (anche se in Francia CP850); ma non assieme greco (CP737), russo ecc.
 - ISO 8859, estensioni standard per ASCII ad 8 bit
 - ISO 8859-1 (o Latin1): Lingue dell'Europa Occidentale
 - ISO 8859-2: Lingue dell'Europa Orientale
 - ISO 8859-5: Alfabeto cirillico
 - ISO 8859-15: Latin1 con simbolo euro (€)

unicode

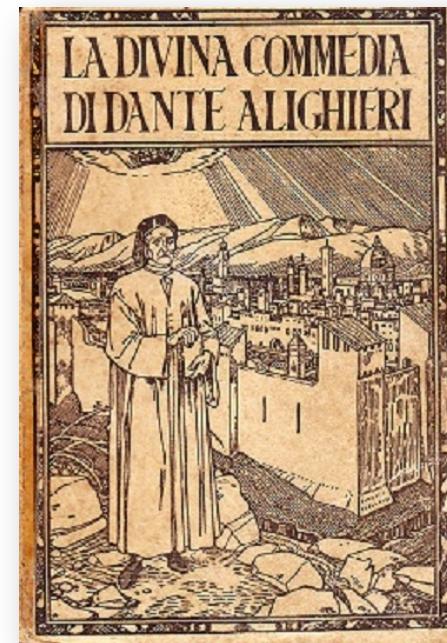
- *unicode* associa un preciso ***code-point (32 bit)*** a ciascun simbolo
 - possibile rappresentare miliardi di simboli
 - primi 256 code-point = Latin1
- attualmente più di 30 sistemi di scrittura
 - rappresentazione di geroglifici e caratteri cuneiformi
 - emoticon ed emoji 😊: ideogrammi per espressioni facciali, oggetti comuni, posti, eventi meteo e animali
 - proposta per Klingon (da Star Trek) ... rifiutata ☹

<https://unicode-table.com>

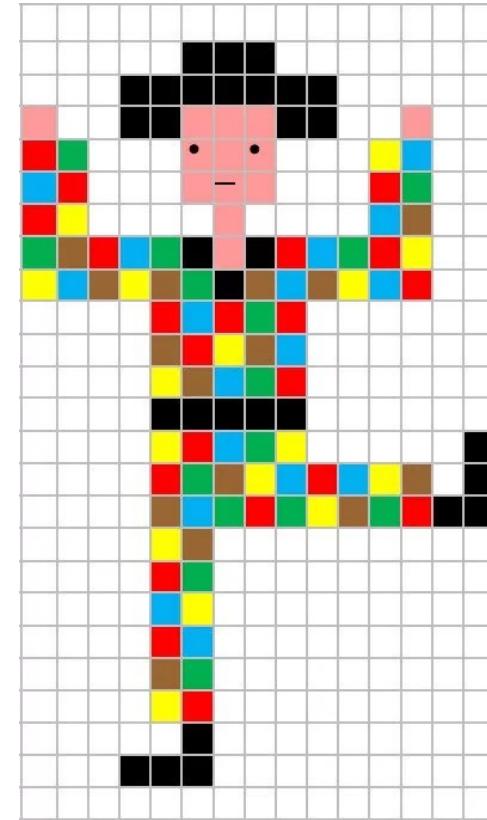


divina commedia

- *La Divina Commedia* di Dante Alighieri è composta da 671.447 caratteri
- 1 carattere = 1 byte
- **670 Kb** = 1 Divina Commedia \simeq 1 megabyte
- *universo digitale*
 - stima
 - attualmente **2.7 zettabyte**
1 zettabyte equivale a un triliardo di byte
 - previsione
 - entro il 2025 **180 zettabyte**



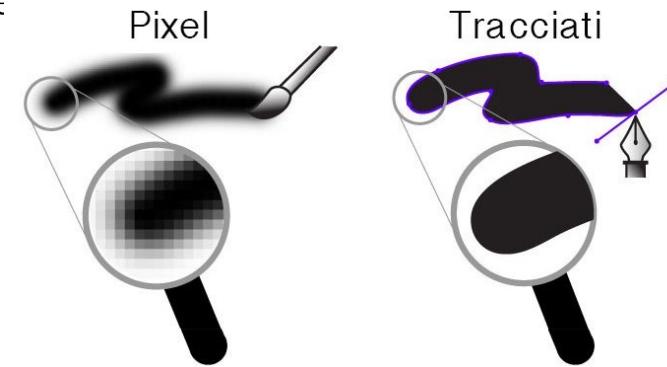
immagini digitali



Alberto Ferrari – Analisi dei Dati

immagini digitali

- **digitalizzazione**: procedimento per convertire un'immagine in una sequenza binaria
- tipologie di immagini digitali
 - **raster** ⇒ immagine suddivisa in una griglia di punti (**pixel**)
 - **vettoriali** ⇒ insieme di primitive geometriche
 - linee, poligoni



palette

- rappresentazione digitale di un'immagine
- la prima operazione è quella di definire una rappresentazione digitale per ogni **colore**
- stabilito il **numero di bit** da utilizzare si definisce l'insieme dei colori (tavolozza, **palette**) che saranno utilizzati per rappresentare l'immagine

colore	codice binario	valore decimale
	0000	0
Yellow	0001	1
Light Green	0010	2
Cyan	0011	3
Pink	0100	4
Blue	0101	5
Orange	0110	6
Lavender	0111	7
Teal	1000	8
Light Green	1001	9
Magenta	1010	10
Brown	1011	11
Olive Green	1100	12
Grey	1101	13
Light Blue	1110	14
Black	1111	15

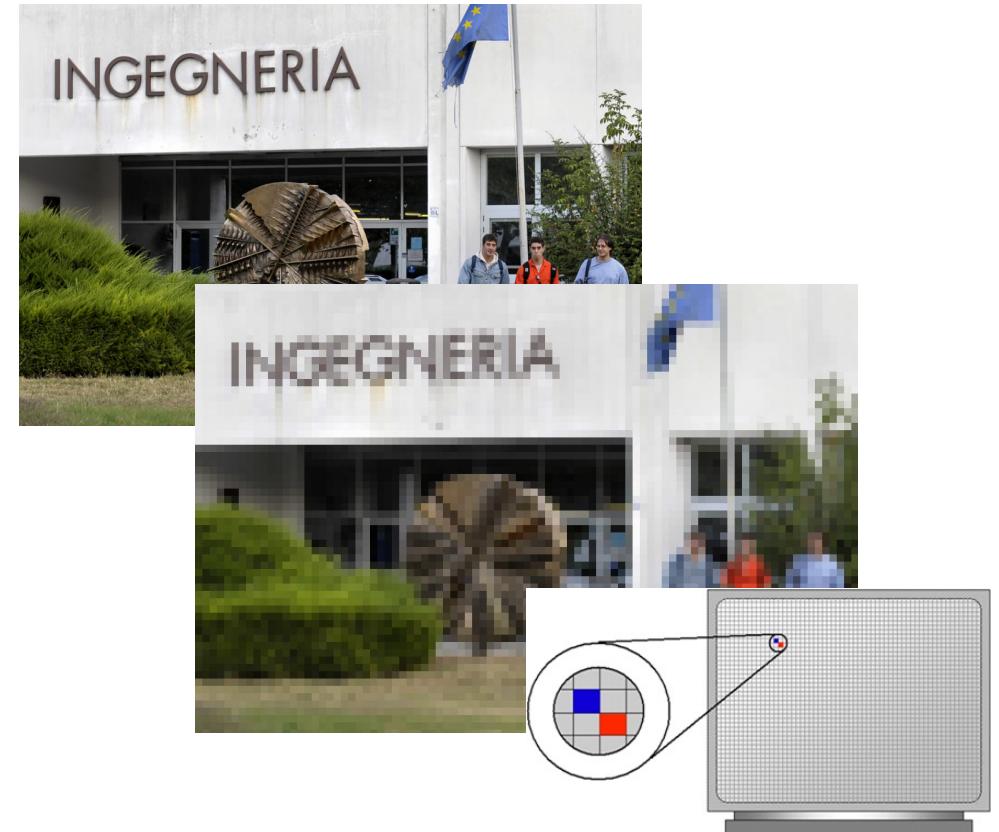
modelli di colore

- occhio sensibile a *variazioni luminosità*
 - *fotoricettori*: 6 mln di *coni*, 120 mln di *bastoncelli*
- **RGB**: rosso, verde, blu
 - 8 bit: 3 bit × R e G, 2 × B
 - 24 bit: 8 bit × R, G e B
 - 32 bit: canale alpha grado trasparenza/opacità
- **YUV**: luminosità, crominanza di R e B
 - sistema PAL, JPEG, MPEG
 - TV a colori (compatibilità B&W)
- **HSL (Hue Saturation Brightness)**: tonalità, saturazione e luminosità



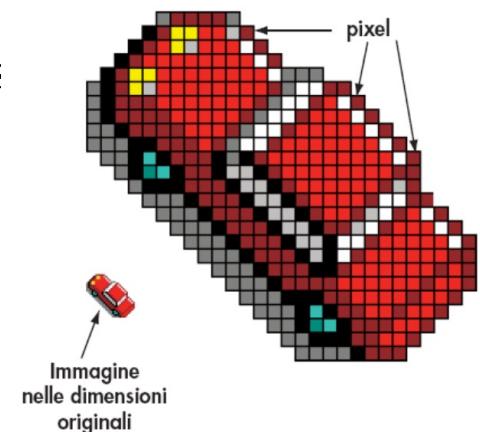
pixel

- immagine suddivisa in piccoli rettangoli
 - *elementi di base dell'immagine digitale*
 - ***pixel*** (*picture element*)
- per ogni pixel individuare un ***colore*** dominante
- l'immagine diventa una sorta di mosaico (*i tasselli del mosaico sono i pixel*)
- la tavolozza fornisce la sequenza di bit associata ad ogni pixel
- l'insieme di tutte le sequenze è la rappresentazione digitale dell'immagine



pixel

- il pixel è di un singolo colore
- il pixel non ha dimensione metrica
 - DPI (Dot Per Inch) (Punti per pollice)
 - DPI esprime la quantità di punti stampati o visualizzati su una linea lunga un pollice
- l'occhio umano non è in grado di percepire la suddivisione
 - su un monitor a 72 DPI
 - (le immagini con queste caratteristiche sono valide per il web)
 - su una stampa a 300 DPI
 - (600 DPI alta qualità)
- <https://gigapixelart.it/cristo-pantocrator-duomo-monreale/>



immagini – approssimazione e risoluzione

- aumentare il numero di pixel (*e ridurre quindi la loro dimensione*) migliora la **definizione** dell'immagine
- i monitor dei computer usano lo stesso procedimento per visualizzare le immagini
- la dimensione ridotta dei pixel e il numero elevato di colori fanno apparire al nostro occhio le immagini come se fossero formate da **linee continue** e infinite **sfumature di colore**
- **risoluzione** dell'immagine
 - **numero dei pixel**
(righe x colonne)
 - **profondità** di colore
(dimensione palette)



Alberto Ferrari – Analisi dei Dati

Sandro Botticelli - Primavera

il quadro



fotografia



Alberto Ferrari – Analisi dei Dati

Sandro Botticelli - Primavera

- originale – copia
- analogico – digitale
- più informazione – meno informazione
- <https://www.haltadefinizione.com/visualizzatore/opera/primavera-sandro-botticelli>

ingrandire ...

- Il microscopio più potente attualmente conosciuto è il microscopio a effetto tunnel (STM) e il microscopio a forza atomica (AFM). Questi strumenti possono risolvere dettagli a livello atomico, permettendo di osservare la struttura e le proprietà delle superfici e dei materiali su scala nanometrica.
- Il microscopio a effetto tunnel (STM) è principalmente considerato un microscopio digitale. Utilizza un ago affilato che scorre sulla superficie del campione a una distanza molto ridotta, e le variazioni di corrente elettrica tra l'ago e il campione vengono misurate e convertite in dati digitali. Questi dati vengono poi elaborati per creare immagini ad alta risoluzione della superficie a livello atomico. Quindi, non solo è digitale, ma offre anche funzionalità avanzate di imaging e analisi.

chatgpt.com

immagini - memoria

- il numero di bit necessario per rappresentare un'immagine è elevato
- es. risoluzione di 1920 x 1080 pixel e 24 bit colore:
 - risulta “scomposta” in $1920 \times 1080 \cong 2$ milioni pixel
 - per pixel colore a 24 bit (3 byte) $\cong 6$ Megabyte
- ***compressione***
 - per limitare l'occupazione di memoria si ricorre a rappresentazioni compresse
 - alcune tecniche di compressione mantengono inalterata la qualità dell'immagine, eliminando soltanto le informazioni ridondanti
 - altre riducono il numero di byte complessivi ma comportano anche perdita di qualità

le vostre fotografie

- quali sono le caratteristiche delle immagini
 - fotocamera / fotocamere del vostro smartphone
 - foto inviate tramite app di messaggistica (whatsapp, telegram ...)
 - foto condivise nei social media (facebook, instagram ...)

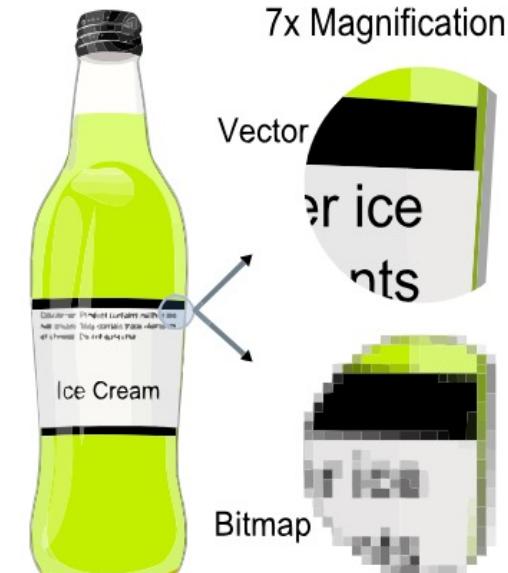


immagini raster - formati

- il formato delle immagini identifica il ***tipo*** di rappresentazione digitale
- ***BMP***: immagine (normalmente) non compressa
- ***TIFF, PNG***: ***comprimono*** l'immagine, per ridurne l'occupazione, senza deteriorarla (compressione ***lossless***)
- ***JPEG***: comprime (molto di più), ma deteriora l'immagine (compressione ***lossy***)

grafica vettoriale

- **immagine:** insieme di primitive geometriche
 - linee, poligoni..., colori, sfumature...
 - per ogni elemento vengono definite le coordinate dei punti di applicazione
-  qualità, a varie risoluzioni
-  compressione dati
-  gestione modifiche
-  non intuitiva
-  possibilmente onerosa



immagini vettoriali

- ***applicazioni***

- editoria (DTP), video-editing, architettura,
- grafica 3D (CAD)
- font vettoriali (*caratteri scalabili in dimensione senza perdere definizione*)

- ***formati***

- PS (PostScript), PDF (Portable Document Format), WMF (Windows MetaFile)
- DXF (AutoCAD), CDR (CorelDraw), SWF (Flash)
- SVG (Scalable Vector Graphics, utilizzato nel web)

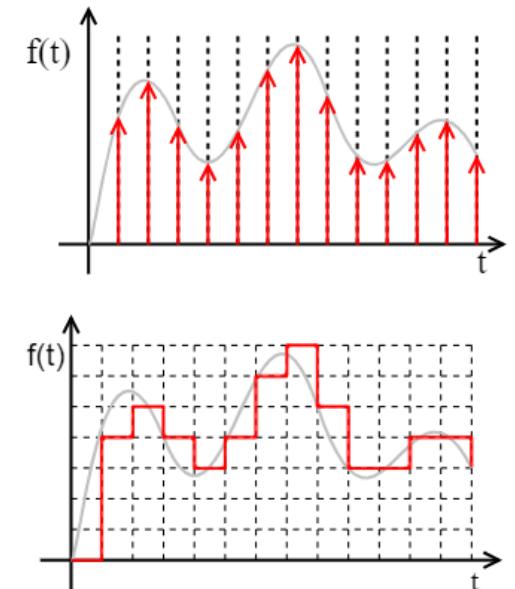
audio digitale



Alberto Ferrari – Analisi dei Dati

suono

- ***suono***
 - onde longitudinali, di ***compressione*** e ***rarefazione*** dell'aria
- grandezza analogica → ***discretizzazione***
- ***campionamento*** (*sampling*) nel tempo Hz [=] $\frac{1}{s}$
- ***quantizzazione*** (*quantizing*) nelle ampiezze bit
- qualità CD
 - 44 kHz, 16bit
 - spettro udibile: 20-20k Hz



Spotify

- Spotify utilizza una gamma di frequenze audio che va da circa 20 Hz a 20 kHz, che è l'intervallo udibile per la maggior parte degli esseri umani.
- Tuttavia, la qualità del suono e il bitrate possono variare a seconda delle impostazioni dell'utente e del piano di abbonamento.
- Spotify offre diverse opzioni di bitrate:
 - Oltre a 96 kbps per streaming su rete mobile.
 - 160 kbps per streaming di qualità standard.
 - 320 kbps per gli abbonati a Spotify Premium, che fornisce una qualità audio superiore.
- In termini di frequenze, tutte queste impostazioni possono riprodurre l'intero spettro udibile, ma la qualità del suono dipenderà anche dalla compressione audio e dalle attrezzature utilizzate per l'ascolto.



chatgpt.com

analogia immagini - suoni

- *analogia* fra il procedimento di digitalizzazione delle immagini e quello dei suoni:
 - scala dei valori sonori \Leftrightarrow tavolozza colori
 - frequenza di campionamento \Leftrightarrow numero pixel
- la scala dell'intensità sonora (numero di “*suoni differenti*”) e la frequenza di campionamento determinano la **qualità** del suono



suoni digitali

- come nel caso delle immagini la rappresentazione digitale dei suoni comporta un *elevato numero di byte*
- per *60 secondi* di audio
 - con rappresentazione a *8 bit* dell'intensità sonora e un campionamento a *8000 Hertz* sono necessari circa *660 Kbyte* (*qualità telefonica*)
 - con rappresentazione a *16 bit* dell'intensità sonora e un campionamento a *44 000 Hertz* i byte sono necessari circa *5 Mbyte* (*10Mb stereo*)

mp3

- analogamente alle immagini vengono usate **rappresentazioni compresse**
- la più nota è **MP3** (*Moving Picture Export Group Layer 3*)
 - l'**orecchio** umano è in grado di percepire solo suoni che stanno all'interno di un certo intervallo di frequenze
 - i suoni a frequenze superiori (**ultrasuoni**) o inferiori (**infrasuoni**) vengono eliminati dalla rappresentazione
 - questo, associato ad **altri procedimenti di compressione** permette di ridurre fino a oltre **12 volte** la quantità di dati digitali nella rappresentazione del suono senza un'apparente perdita di qualità



midi

- analogamente con quanto visto per le immagini vettoriali, nel caso di suoni prodotti da strumenti musicali, è possibile rappresentare, al posto del suono, la sequenza di **azioni** necessarie per **generarlo**
- si parla in questo caso di **suono sintetizzato**
- un esempio di questo tipo sono i suoni **MIDI** (*Musical Instrument Digital Interface*) nei quali vengono registrati gli eventi che generano un certo suono

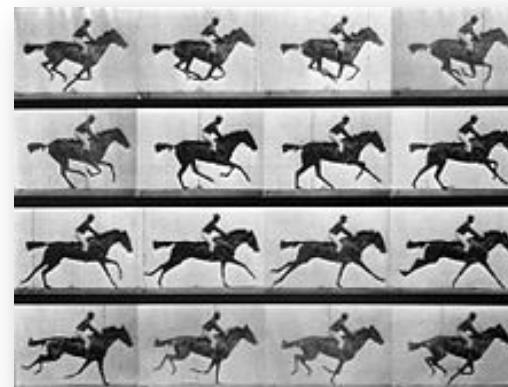


filmato digitale

Alberto Ferrari – Analisi dei Dati

filmati

- prendiamo come modello una pellicola cinematografica:
 - una sequenza di immagini statiche (fotogrammi)
 - una o più bande per il sonoro
- l'occhio umano non riesce a percepire come distinte due immagini separate da meno di un trentesimo di secondo



Alberto Ferrari – Analisi dei Dati

codifica filmati

- ogni singolo fotogramma viene digitalizzato utilizzando i procedimenti visti per la rappresentazione delle immagini
- la colonna sonora subisce lo stesso processo di conversione che abbiamo incontrato trattando i suoni digitali

memoria

- ***problema*** legato all'occupazione di memoria
 - (*soprattutto per trasmissione*)
- procedimenti di ***compressione*** per ridurre la dimensione
 - spesso solo una parte dell'immagine varia da un fotogramma al successivo
 - rappresentazione del fotogramma di partenza e poi solo della parte che in ogni fotogramma è differente dal precedente
- ***fattori*** che determinano la quantità di memoria:
 - ***lunghezza*** della sequenza
 - dimensione in ***pixel***
 - numero di ***colori***
 - numero di fotogrammi al secondo (***frame rate***)
 - qualità del ***sonoro***

formati video

- **MKV** (Matroska Video) è un formato video che memorizza diverse tracce audio e sottotitoli in un unico file. È molto diffuso per la memorizzazione di film.
- **MP4** è stato sviluppato dal Moving Picture Experts Group. Il formato supporta il codec video H.264 e molti altri. Il formato è supportato dalla maggior parte dei dispositivi moderni.

