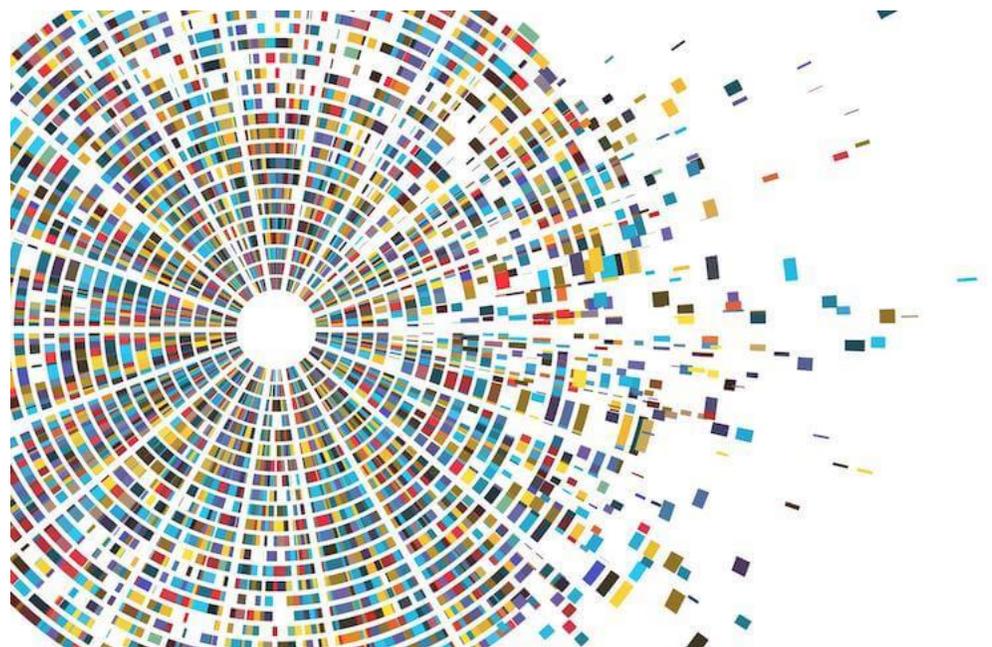


# *data analysis*



## ***Big data – le 3 V del modello del 2001 (Doug Laney)***

- ***Volume***

- l'ingente massa di informazioni generate nell'attività quotidiana che non è possibile raccogliere con tecnologie tradizionali
- il volume dei dati è in continua crescita ed è difficile identificare un valore limite al di sopra del quale si può parlare di Big Data

- ***Velocità***

- i dati vengono generati e acquisiti sempre più rapidamente
- necessità non solo di raccogliere ma anche di analizzare i dati tempo reale per poter prendere decisioni tempestive

- ***Varietà***

- le differenti tipologie di dati provenienti da un numero crescente di fonti eterogenee
- aziendali, sensori, social network, open data

## *Big data – da 3 V a 5 V*

- *Veridicità*

- i dati devono essere affidabili, raccontare il vero
- la qualità e l'integrità delle informazioni è un pilastro imprescindibile per permettere analisi utili e affidabili

- *Variabilità*

- la mutevolezza del significato dei dati è un aspetto da tenere in considerazione nel momento in cui vengono interpretati

## *Big data – la sesta V*

- *Valore*

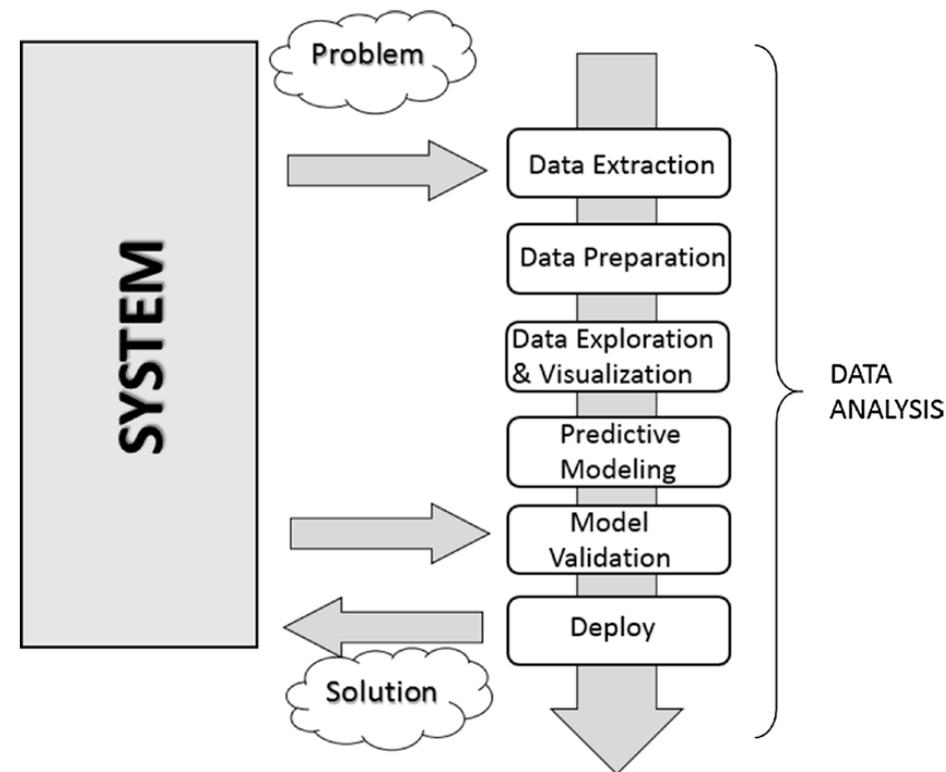
- i dati sono diventati una fonte inestimabile di valore
- limitarsi a raccogliere i dati non garantisce di avere informazioni e soprattutto di estrarre conoscenza
- sono necessari strumenti di Analytics (Big Data Analytics)

## *data analysis – data analytics*

- *data analysis*
  - processo di estrazione delle informazioni dai dati grezzi
- *data analytics*
  - disciplina generale per la gestione completa dei dati e comprende:
    - analisi
    - raccolta dei dati
    - organizzazione
    - archiviazione
    - strumenti e tecniche utilizzate
- è compito del *data analyst* raccogliere, analizzare e tradurre i dati in informazioni accessibili e identificare modelli per aiutare le organizzazioni a prendere decisioni 'aziendali' migliori

## *processo di data analysis*

- *definizione del problema*
- *estrazione* / recupero dati
- *preparazione* dei dati
  - pulizia dei dati
  - trasformazione dei dati
- *esplorazione e visualizzazione* dei dati
- *modellazione* predittiva
- *validazione* / test del modello
- *rilascio (deployment)*
  - visualizzazione e interpretazione dei risultati
  - distribuzione della soluzione



## ***dati – informazioni - conoscenza***

- ***dati***

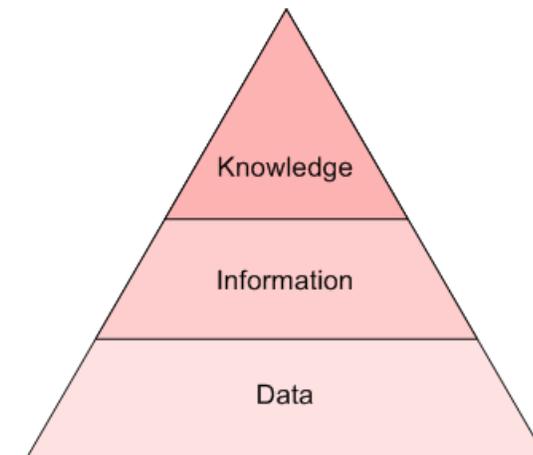
- registrazione di eventi che accadono, tutto ciò che può essere misurato o classificato può essere convertiti in dati

- ***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



# *i dati*



# *dati*

- dati *quantitativi*
  - possono essere descritti tramite *numeri*
  - su di essi è possibile eseguire *operazioni matematiche*
  - possono essere *discreti* o *continui*
- dati *qualitativi*
  - non possono essere descritti tramite numeri
  - in genere, vengono descritti usando delle categorie e un linguaggio «naturale»



## *dati qualitativi*

- ***ordinali***

- esprimono un ordine o una gerarchia e possono essere collocati su una scala
- es. classificazione in base alla massa corporea (sottopeso, normopeso, sovrappeso, obesi)
- esprimono una scala (possiamo associare un valore da 1 a 4)

- ***nominali***

- non esprimono un ordine o una scala
- es. classificazione in base al colore dei capelli (biondo, castano, moro, ecc.)
- possiamo associare un valore (es. biondo = 1, castano = 2, moro = 3, ecc.)
- qualsiasi calcolo basato su tali numeri sarà privo di senso.

- ***dicotomici***

- variabili che possono assumere soltanto due valori
- vero o falso, presente o assente, sì o no
- è possibile associare a ciascun valore un numero (es. vero=1, falso=0)

## *dati quantitativi*

- *discreti*

- il dato può assumere un insieme finito o numerabile di valori
- es. numero di tiri effettuati da una squadra in una partita di calcio, numero di punti realizzati da un giocatore di volley in un set, ecc.

- *continui*

- il dato può assumere tutti i valori compresi in un determinato intervallo, senza soluzione di continuità tra un valore e l'altro
- es. percentuale di realizzazione di tiri da tre punti di un giocatore di basket in una stagione, percentuale di ace in relazione al numero di servizi di un tennista in un torneo

## *struttura dei dati*

- ***dati strutturati***

- rispondono a criteri stabiliti da un modello dati predefinito
- es. le tabelle (ogni riga corrisponde a un record e ogni colonna è un attributo) e nelle celle sono presenti le informazioni
- sono facili da gestire e da analizzare

- ***dati semi-strutturati***

- non seguono un modello tabellare preciso ma mantengono comunque una loro struttura che ne agevola l'analisi
- uso di tag e marcatori per stabilite delle gerarchie tra i campi e tra i record
- es. i file XML, JSON e HTML

- ***dati non strutturati***

- manca un modello predefinito
- es. file di testo, pdf, immagini, e-mail
- sono più difficili da gestire

# dati strutturati - esempi

REGIONE	POSITIVI SARS-CoV2				DIMESSI GUARITI	Deceduti	Casi totali	Incremento casi totali (rispetto al giorno precedente)	Casi identificati dal sospetto diagnostico	Casi identificati da attività di screening	CASI TOTALI	Totale casi testati	Totale tamponi effettuati	INCREMENTO TAMPONI
	Ricoverati con sintomi	Terapia intensiva	Isolamento domiciliare	Totale attualmente positivi										
Lombardia	5.563	570	111.233	117.366	107.463	18.118	242.947	9.934	190.116	52.831	242.947	1.960.088	3.172.359	46.401
Piemonte	3.871	268	43.777	47.916	36.993	4.549	89.458	4.878	39.162	50.296	89.458	703.606	1.117.280	21.288
Campania	1.677	180	60.339	62.196	15.017	796	78.009	4.508	75.357	2.652	78.009	739.433	1.075.201	23.897
Veneto	1.184	169	42.584	43.937	25.612	2.543	72.092	3.297	25.565	46.527	72.092	937.280	2.417.660	20.005
Emilia-Romagna	1.673	177	31.880	33.730	28.559	4.752	67.041	1.953	46.964	20.077	67.041	897.938	1.695.309	20.847
Lazio	2.511	234	43.939	46.684	12.653	1.370	60.707	2.699	19.176	41.531	60.707	1.296.386	1.591.303	28.744
Toscana	1.303	209	37.928	39.440	16.738	1.502	57.680	2.592	46.541	11.139	57.680	780.994	1.180.350	15.743
Liguria	1.315	78	10.299	11.692	20.881	1.879	34.452	1.127	24.320	10.132	34.452	244.325	475.685	5.772
Sicilia	1.157	159	18.197	19.513	8.684	628	28.825	1.423	17.968	10.857	28.825	514.907	737.769	9.525
Puglia	787	122	14.720	15.629	7.450	802	23.881	946	6.712	17.169	23.881	423.213	594.560	7.728
Marche	444	60	8.736	9.240	7.379	1.037	17.656	697	17.400	256	17.656	199.235	338.947	3.842
Umbria	326	53	8.506	8.885	4.532	174	13.591	767	3.841	9.750	13.591	185.144	320.893	4.855
Abruzzo	468	42	8.071	8.581	4.340	584	13.505	395	7.396	6.109	13.505	190.210	308.505	3.067
Friuli Venezia Giulia	260	44	6.451	6.755	6.162	435	13.352	542	11.675	1.677	13.352	238.032	567.637	6.552
P.A. Bolzano	356	35	6.655	7.046	4.196	330	11.572	713	11.572	0	11.572	127.570	249.844	2.961
Sardegna	392	46	7.248	7.686	3.477	249	11.412	359	4.384	7.028	11.412	242.502	286.076	3.699
P.A. Trento	226	17	2.190	2.433	7.561	459	10.453	261	5.864	4.589	10.453	117.055	304.662	3.463
Calabria	212	15	4.254	4.481	2.101	132	6.714	264	1.243	5.471	6.714	289.214	292.222	2.861
Valle d'Aosta	154	13	2.001	2.168	1.600	195	3.963	129	3.481	482	3.963	27.094	44.359	719
Basilicata	100	16	2.219	2.335	778	59	3.172	249	1.124	2.048	3.172	112.020	112.980	1.461
Molise	26	8	1.371	1.405	749	45	2.199	76	2.149	50	2.199	63.905	68.295	815
<b>TOTALE</b>	<b>24.005</b>	<b>2.515</b>	<b>472.598</b>	<b>499.118</b>	<b>322.925</b>	<b>40.838</b>	<b>862.681</b>	<b>37.809</b>	<b>562.010</b>	<b>300.671</b>	<b>862.681</b>	<b>10.290.151</b>	<b>16.951.896</b>	<b>234.245</b>

```
{
  "utenti": [
    {
      "nome": "Mario Rossi",
      "foto": {
        "file": "mrossi.jpg"
      }
    },
    {
      "nome": "Laura Verdi",
      "foto": {
        "file": "lverdi.jpg"
      }
    }
  ]
}
```

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <anagrafica>
- <cliente>
  <nome>Massimo</nome>
  <cognome>Furia</cognome>
  <citta>Varese</citta>
  <ragionesociale>LogicImage srl</ragionesociale>
  <piva>012345678011</piva>
- <fatturato>
  <annuo>1000000</annuo>
  <ultimoanno>2007</ultimoanno>
  <mediamensile>10000</mediamensile>
  <passivita>3000</passivita>
  <ammortamenti>1000</ammortamenti>
  <utile>6000</utile>
</fatturato>
  <codicecliente>0450222</codicecliente>
</cliente>
```

## *metadati*

- metadato (dal greco antico μετά, «oltre, dopo, per mezzo» e dal latino datum, «informazione»)
- è un dato che descrive una qualche proprietà di un altro dato
- i metadati forniscono informazioni sui dati stessi
- es. l'autore di un file di testo, il timestamp di un'immagine, la risoluzione di una foto, ecc.

## *dati non o semi organizzati*

- ***non organizzati*** (non strutturati)
  - in formato libero
    - testo o audio grezzo o segnali che devono essere analizzati meglio per poter essere organizzati
  - rappresentano **80% - 90%** dei dati mondiali
  - necessario utilizzo di tecniche di ***preprocessing*** per dare una struttura ad almeno una parte dei dati e fornirli alla successiva analisi
- ***semi organizzati*** (semi strutturati)
  - presentano una parte dotata di struttura e una parte non strutturata
  - esempi
    - un documento Word, o PDF, possiede una serie di metadati che sono molto ben strutturati (titolo , autore, numero di parole ...) mentre il corpo del documento è costituito da testo
    - le immagini all'interno del file presentano una serie di metadati che descrivono lo scatto (risoluzione, impostazioni fotocamera, coordinate gps ...)

## *esempio preprocessing su un testo «libero»*

- esempio tweet  
*«This Wednesday morn, are you early to rise? Then look East. The Crescent Moon joins Venus & Saturn. Afloat in the dawn skies»*
- possibile eseguire:
  - conteggio di parole/frasi
  - presenza di determinati caratteri speciali (*emoticon ...*)
  - lunghezza del testo
  - individuazione degli argomenti (*sentiment analysis ...*)

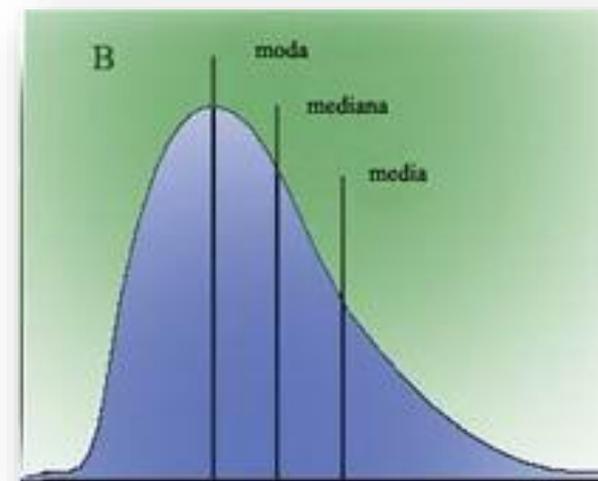
	this	wednesday	morn	are	this wednesday	?	Lunghezza relativa	Argomento
Conteggio parole	1	1	1	1	1	1	4.03	astronomia

## *livelli dei dati*

- livello *nominale*
  - dati qualitativi *classificati* (raggruppati) in categorie qualitative
    - nazionalità, specie, colore
  - esiste la sola relazione (identità, appartenenza)
  - statistica: moda
- livello *ordinale*
  - esiste un ordine di valutazione ma non si hanno le differenze relative fra le osservazioni (non è possibile sommarle o sottrarle)
  - statistica: mediana, moda
- livello degli *intervalli*
  - si possono ordinare e confrontare i dati ma anche sottrarli e sommarli
  - statistica: media, mediana, moda, deviazione standard
- livello dei *rapporti*
  - ha il vantaggio di avere un'origine reale (misure in cui zero significa quantità nulla)
  - es: l'altezza, la distanza, la velocità, l'età, il peso, il reddito

## *statistica cenni*

- *media aritmetica*
- *mediana*
  - valore che si trovano nel mezzo della distribuzione (ordinata)
- *moda*
  - è il valore che compare più frequentemente
- *deviazione standard* (scarto quadratico medio)
  - differenza al quadrato fra i punti e la media
  - permette di enfatizzare i valori che sono eccezionalmente lontani dalla media



## *media aritmetica*

- si sommano tutti i valori presenti nell'insieme dei dati
- si conta il numero di valori presenti nell'insieme di dati
- si divide la somma per il numero dei valori
- insieme di dati : 2, 3, -5, 3, 7, 2, 2, 4
- media aritmetica?

$$\text{Somma} = x_1 + x_2 + x_3 + \dots + x_n$$

$$\text{Media} = \frac{\text{Somma}}{n}$$

## *moda*

- la **moda** è il valore che appare più frequentemente in un insieme di dati
- un dataset può avere una moda, più di una moda (se ci sono più valori con la stessa frequenza massima) o nessuna moda (se tutti i valori si presentano con la stessa frequenza)
- insieme di dati : 2, 3, -5, 3, 7, 2, 2, 4
- moda : ?

## *mediana*

- la **mediana** è il valore centrale di un insieme di dati ordinati
- per calcolare la mediana, si dispone l'insieme di dati in ordine crescente (o decrescente) e si identifica il valore che si trova nel mezzo
  - se ci sono un numero dispari di osservazioni, la mediana è il valore centrale
  - se ci sono un numero pari di osservazioni, la mediana è la media dei due valori centrali
- insieme di dati : 2, 3, -5, 3, 7, 2, 2, 4
- mediana : ?

## *deviazione standard*

<https://mathcracker.com/it/calcolatrice-deviazione-standard-media>

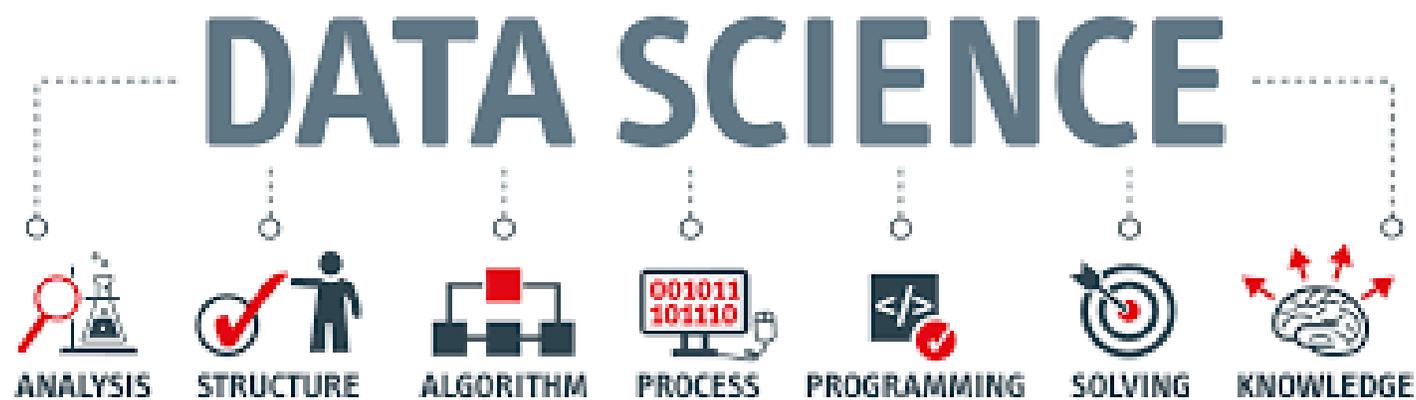
- la **deviazione standard** indica quanto i valori di un insieme di dati si discostano dalla media e fornisce un'indicazione della dispersione o della variabilità dei dati
  - una deviazione standard bassa indica che i dati sono vicini alla media
  - una deviazione standard alta indica che i dati sono più sparsi e distanti dalla media
- come di calcola:
  - si calcola la media di tutti i valori
  - per ogni valore si calcola la differenza rispetto alla media
    - $\text{media} - \text{valore}$
  - si eleva al quadrato il dato ottenuto
  - la **varianza** è la somma di tutti questi quadrati fratto il numero dei valori presenti nell'insieme dei dati
  - la **deviazione standard** è la radice quadrata della varianza



## *nominale o ordinale?*

- l'origine dei semi nella vostra tazzina di caffè
- la posizione ottenuta dai partecipanti a una gara podistica
- il metallo usato per la medaglia ricevuta dopo aver partecipato alla suddetta gara
- il numero telefonico di un cliente
- quante tazzine di caffè bevete in una giornata

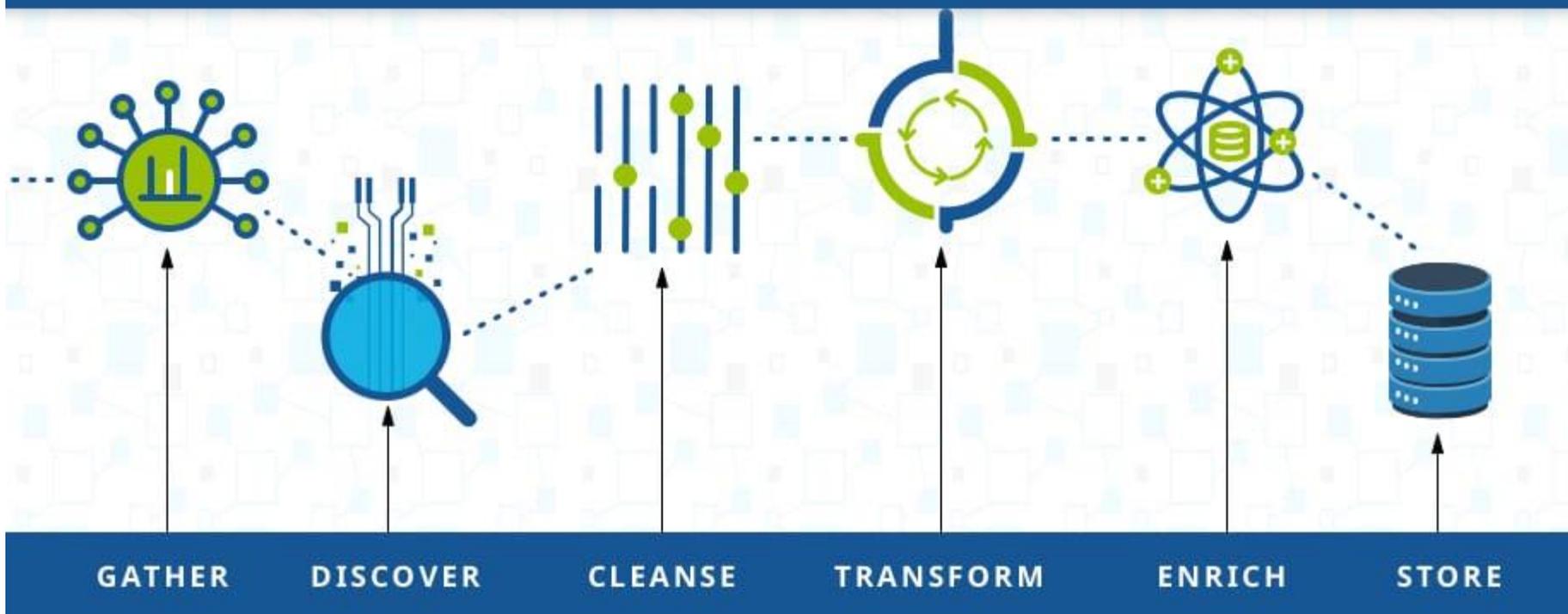




## *le cinque fasi della data science*

- *individuare un obiettivo*
- *ottenere i dati*
  - dove si possono trovare i dati?
  - all'interno dell'azienda? pubblici o privati? costano?
- *esplorare i dati*
  - riconoscere i diversi tipi di dati, trasformare i dati per migliorare la qualità dell'intero dataset e per prepararlo per la fase di modellazione
- *creare un modello per i dati*
  - di modelli statistici e di machine learning
- *comunicare e presentare i risultati*
  - presentare i risultati in forma chiara e comprensibile è più difficile di quanto si possa immaginare

# DATA PREPARATION



## *esplorazione dei dati (data preparation)*

- *importante*
  - alla preparazione dei dati è da imputare la maggior parte del tempo speso per un progetto di analisi predittiva
- *pulizia dei dati*
- *feature engineering*
  - creazione delle variabili di input a partire dai dati
- *gestione dei valori mancanti*
- *individuazione dei valori anomali (outliers)*
- *normalizzazione (rescaling)*

## *normalizzazione*

- gli algoritmi basati sul calcolo della distanza tra punti nello spazio multidimensionale beneficiano particolarmente della *normalizzazione* dei valori ad un intervallo  
 (es  $[-1, +1]$ ,  $[0, 1]$ )
- per non dare un *peso eccessivo* ai valori più grandi
- di solito si usa per normalizzare:
  - valori di *minimo*, *massimo*
  - *deviazione standard*
    - $u$  media  $\sigma$  deviazione standard

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

$$z = \frac{x_i - u}{\sigma}$$

## *altre operazioni sui dati*

- *variabili categoriche*

- raggruppamento (riduzione del numero di livelli)
- ordinal encoding (codifica con valori numerici)
- one hot encoding (creazione di una colonna [val 0,1] per ogni categoria)

- *variabili continue*

- discretizzazione (creazione di fasce di valori)

# *comunicare i dati*



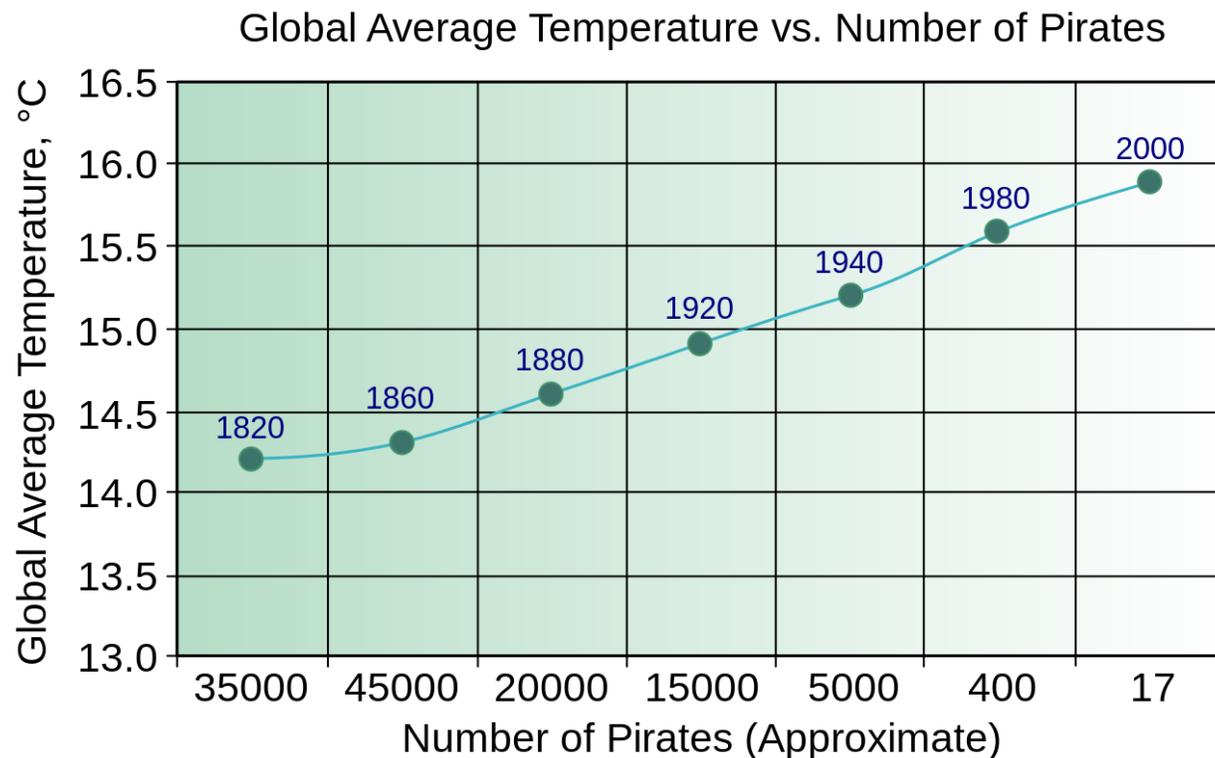
## *comunicare i dati*

- *scopo*
  - presentare i *risultati* ottenuti in modo *coerente* e *comprensibile*
- *obiettivo*
  - chiunque sia in grado di *comprendere* e *utilizzare* i nostri risultati
- le *rappresentazioni grafiche* favoriscono la sintesi
  - identificare i *metodi di presentazione efficaci* (e quelli inefficaci)
  - esistono grafici che hanno lo scopo di «*ingannare*» il pubblico
  - importante distinguere *causalità* da *correlazione*

## *correlazione vs causalità*

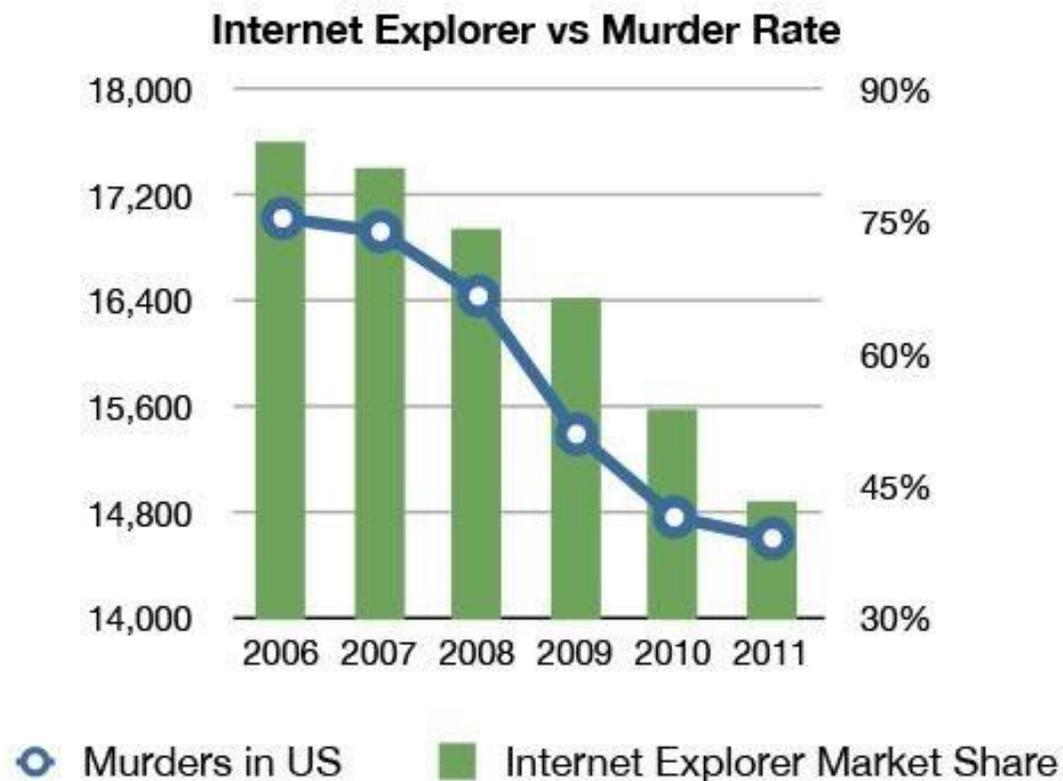
- *“la correlazione non è necessariamente indice di causalità”*
- *correlazione*
  - *relazione* tra due variabili tale che a ciascun valore della prima corrisponda un valore della seconda, seguendo una certa regolarità
- *causalità*
  - principio interpretativo della realtà, che si fonda sul rapporto di *causa ed effetto*

## *correlazione o causalità?*



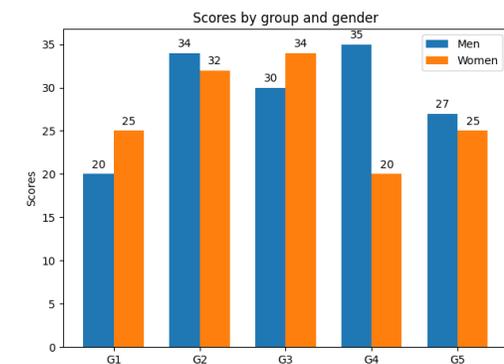
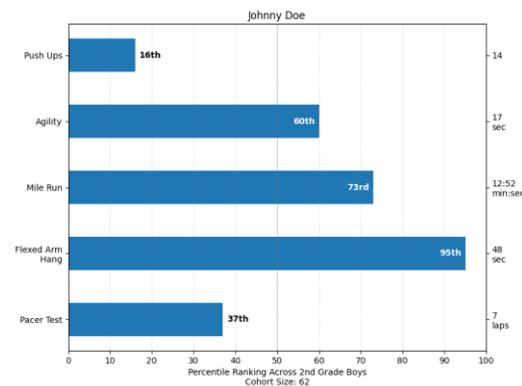
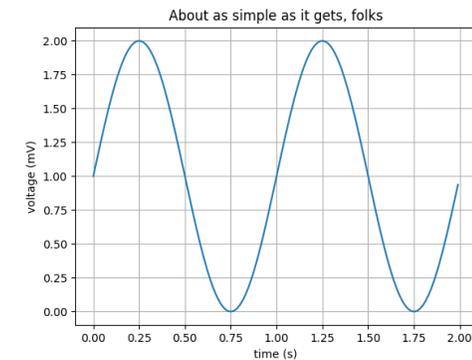
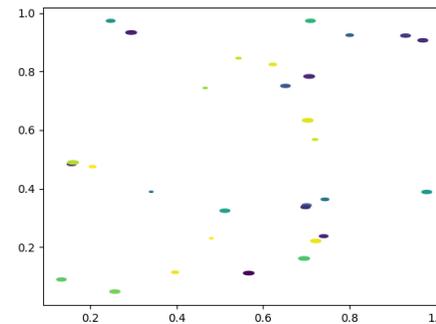
You can see that as the *number of pirates* in the world has *decreased* over the past 130 years, *global warming* has gotten steadily *worse*. In fact, this makes it entirely clear that *if you truly want to stop global warming, the most impactful thing to do is -- become a pirate.*

## *vendite di Microsoft Internet Explorer e numero di omicidi*



## *tipi di grafici*

- *grafici a dispersione*
- *grafici a linee*
- *diagrammi a barre*
- *istogrammi*



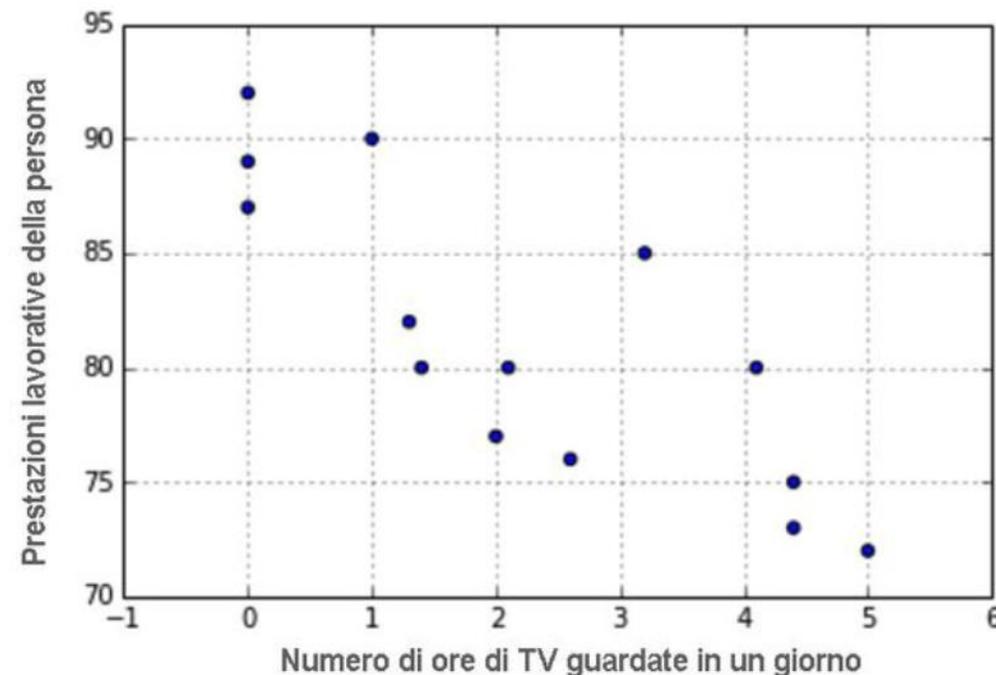
## *grafico a dispersione (scatter plot)*

- i due *assi* sono *quantitativi*
- ogni punto rappresenta un'osservazione
- si evidenziano le *relazioni* esistenti fra *due variabili*
  - se possibile si individua una *correlazione*

*il grafico sembra mostrare una relazione: «il numero di ore giornaliere in cui guardiamo la TV determina le nostre prestazioni lavorative»*

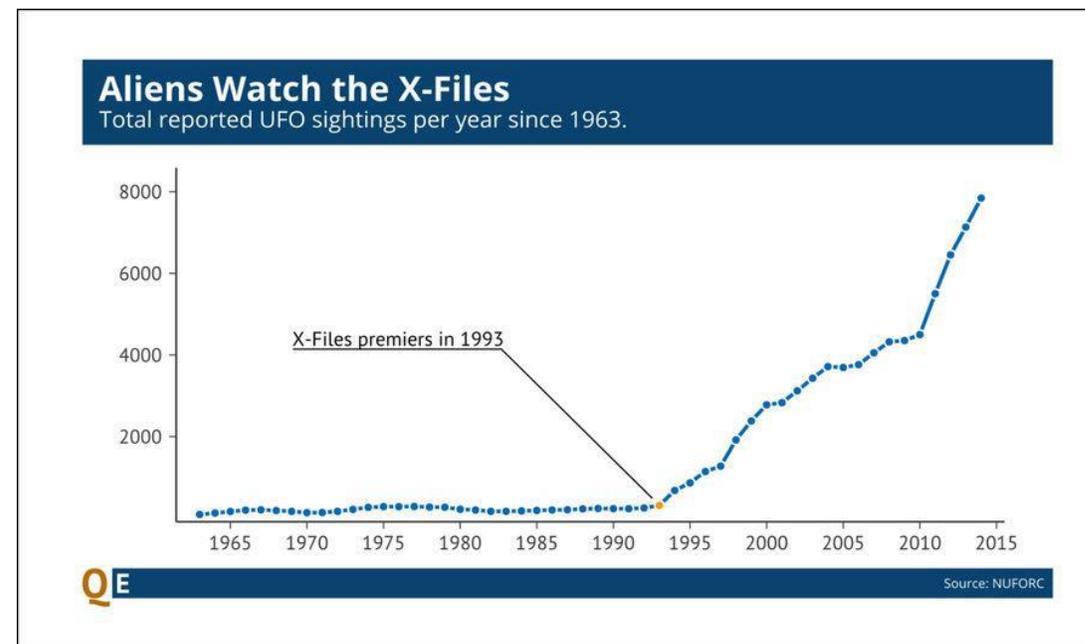
*ma potrebbero non esserci elementi di causalità*

*il grafico può solo aiutare a rilevare una correlazione o un'associazione, ma non una causalità*



## *grafico a linee (line chart)*

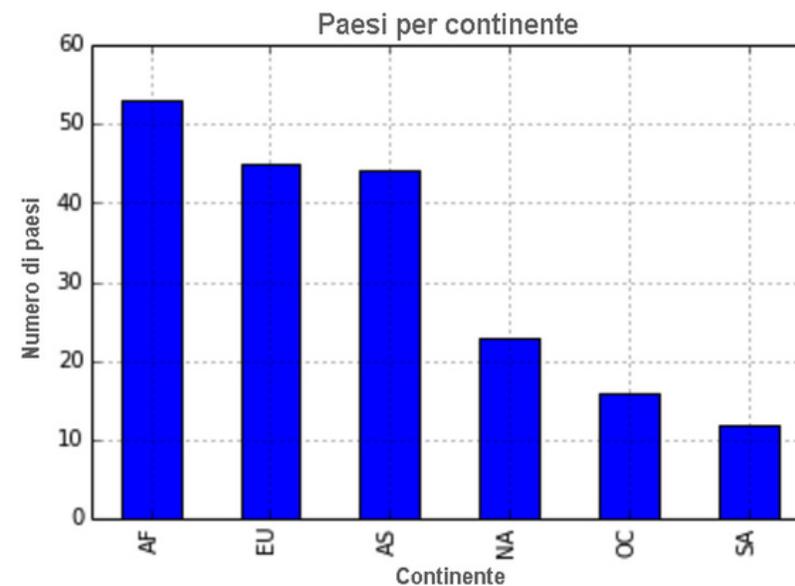
- normalmente l'asse x rappresenta una linea del **tempo** e l'asse y è di tipo quantitativo
- è costituito da una serie di punti (rilevazioni) uniti da linee rette



*Sembra evidente che, subito dopo il 1993 (anno di uscita della prima stagione di X Files) il numero di avvistamenti di UFO si è impennato*

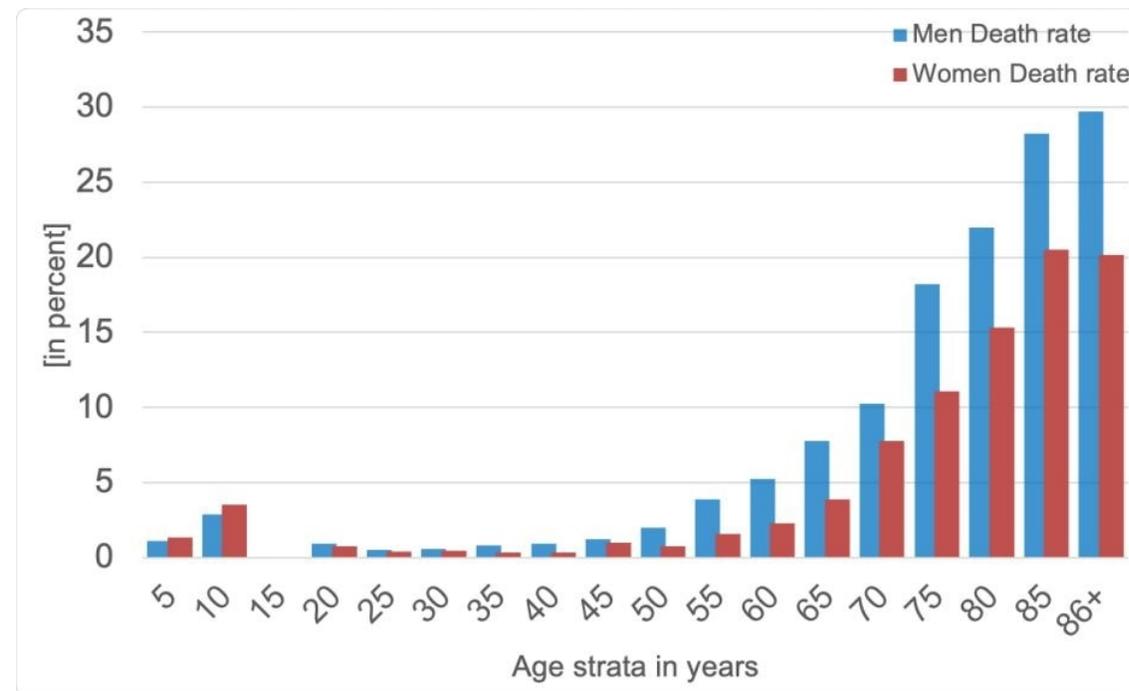
## *diagramma a barre (bar chart)*

- permette di ***confrontare*** le valori appartenenti a vari ***gruppi***
- sull'asse x si trova una variabile ***categorica***
- sull'asse y una variabile ***quantitativa***



## *istogramma (histogram)*

- rappresentazione della distribuzione di frequenza di un'unica variabile quantitativa
- l'asse y è quantitativo
- l'asse x è categorico
  - ogni categoria comprende un determinato intervallo di valori



## *grafici e dati online*

- Organizzazione Mondiale della Sanità (grafici)
  - <https://worldhealthorg.shinyapps.io/covid/>



- Presidenza del Consiglio dei Ministri (protezione civile)
  - <https://github.com/pcm-dpc/COVID-19>

