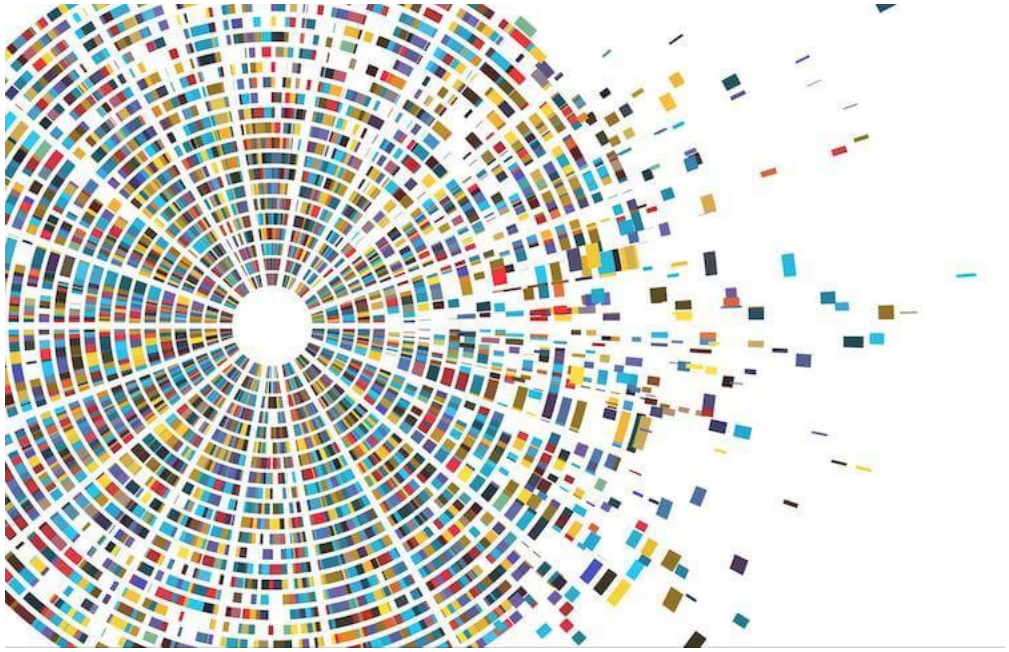
data analysis

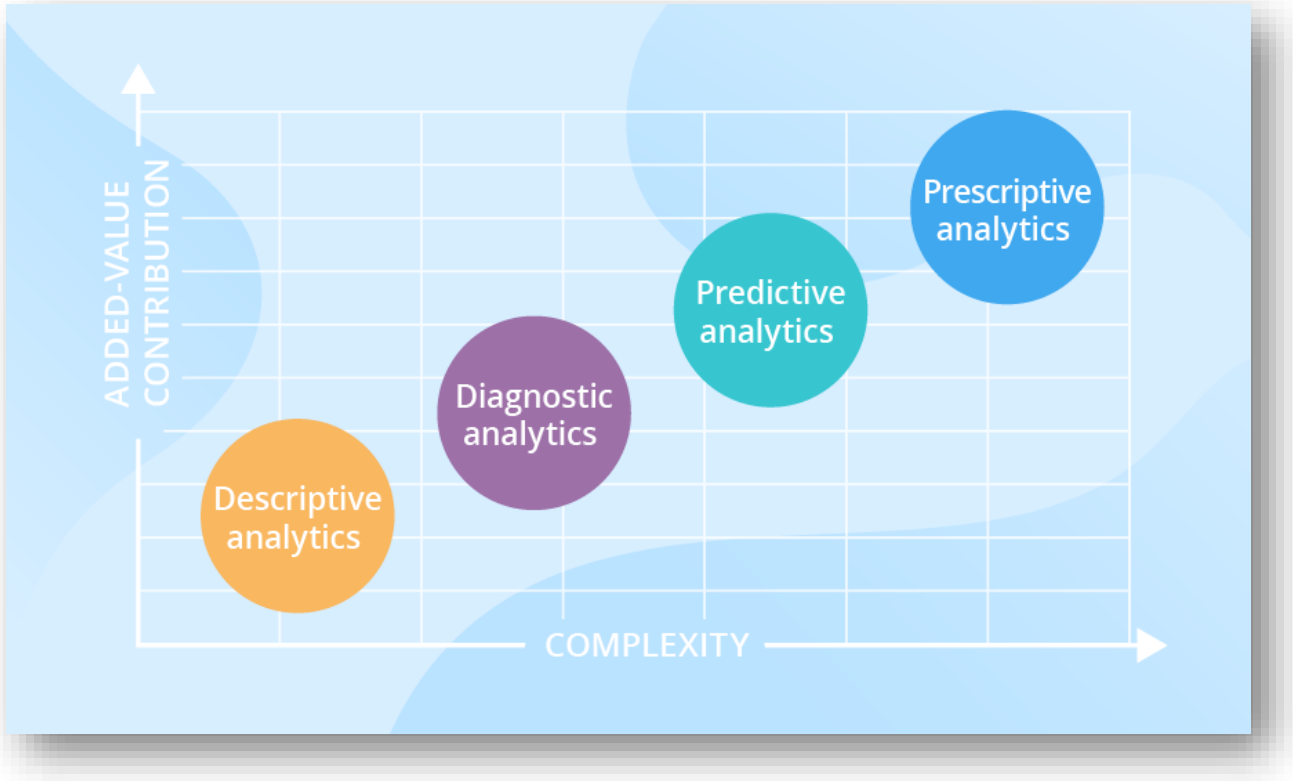


data analysis – data analytics

- ***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

analytics

- ***descriptive*** analytics
  - diagnostic analytics
- ***predictive*** analytics
- ***prescriptive*** analytics
- ***automated*** analytics



descriptive - diagnostic analytics

- ***what happened?***
  - insieme di strumenti orientati a descrivere la ***situazione attuale e passata*** dei processi aziendali e/o aree funzionali
  - strumenti che permettono di ***accedere ai dati*** secondo viste logiche flessibili e di ***visualizzare*** in modo sintetico e grafico i principali indicatori di prestazione
- ***why did it appen?***
  - ***data mining*** / correlazione

data mining
individuazione di informazioni di varia natura (non risapute a priori) tramite estrapolazione mirata da una singola banca dati o da diverse banche dati (incrociando i dati di queste)

predictive analytics

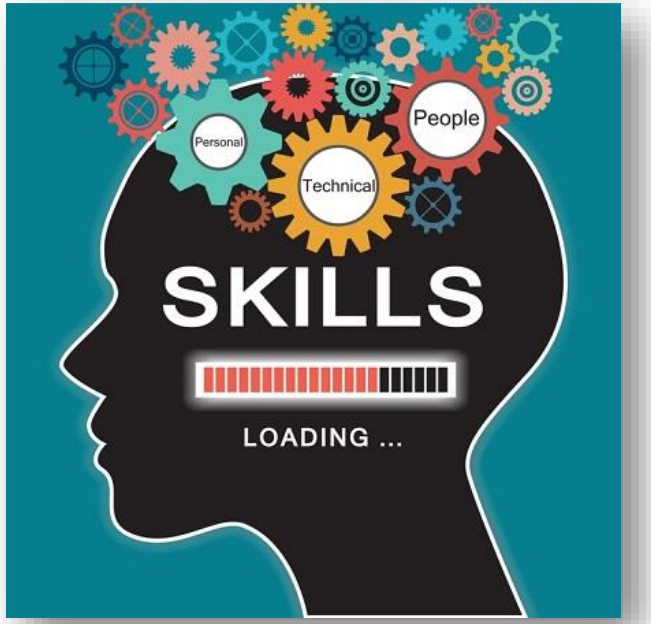
- ***what will appen?***
  - strumenti avanzati che effettuano l'analisi dei dati per rispondere a domande relative a cosa potrebbe accadere nel ***futuro***
  - tecniche matematiche
    - regressione
    - forecasting
      (interpretare le performance e le esigenze aziendali in divenire)
    - modelli predittivi

prescriptive / automatic analytics

- ***how can we make it appen?***
  - applicazioni ***big data*** avanzate
  - insieme all'analisi dei dati propongono al decision maker soluzioni operative/strategiche sulla base delle analisi svolte
- ***automated analytics***
  - capaci di implementare ***autonomamente*** l'azione proposta secondo il risultato delle analisi svolte

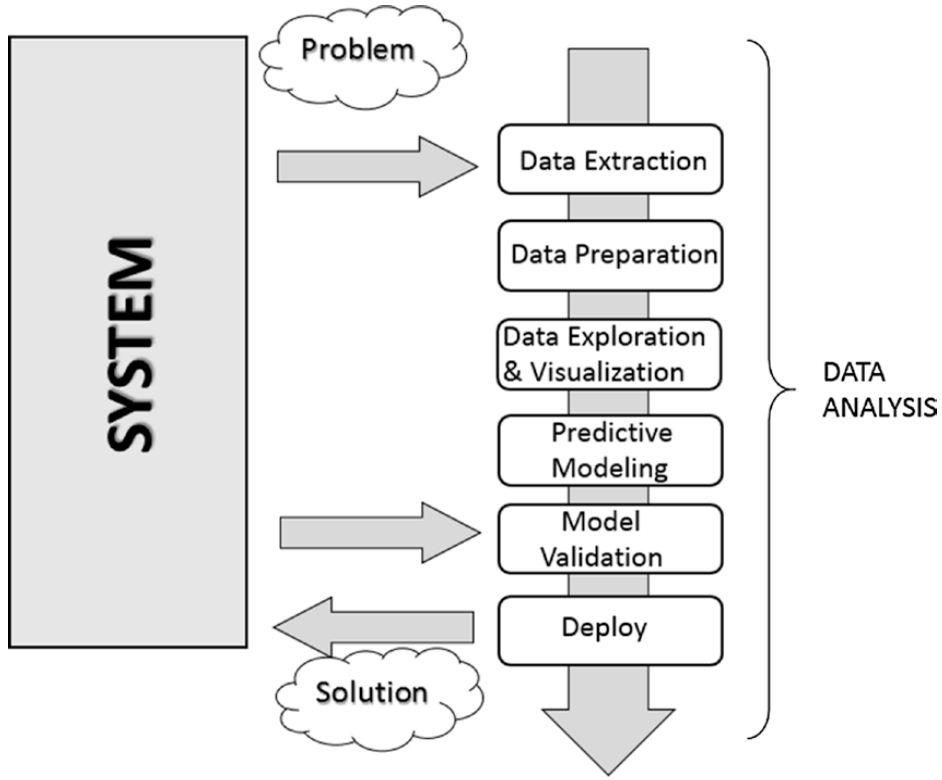
skill (competenze)

- ***statistica e matematica***
  - regressione
  - overfitting
  - supervised vs. unsupervised learning
- ***software development***
  - R, ***python***, MATLAB
  - database SQL, NoSQL
- ***business experience***
  - conoscenza dei processi aziendali



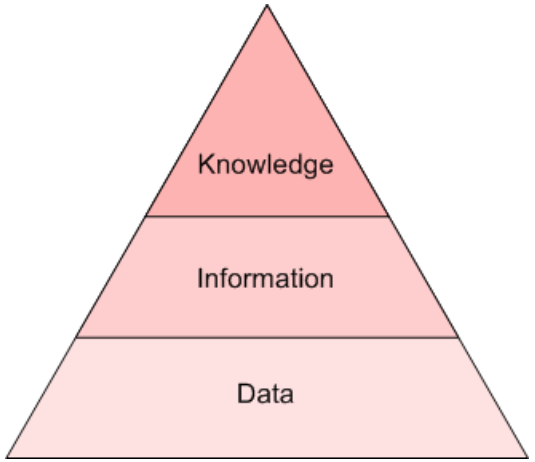
processo di data analysis

- ***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



consolidamento dei dati

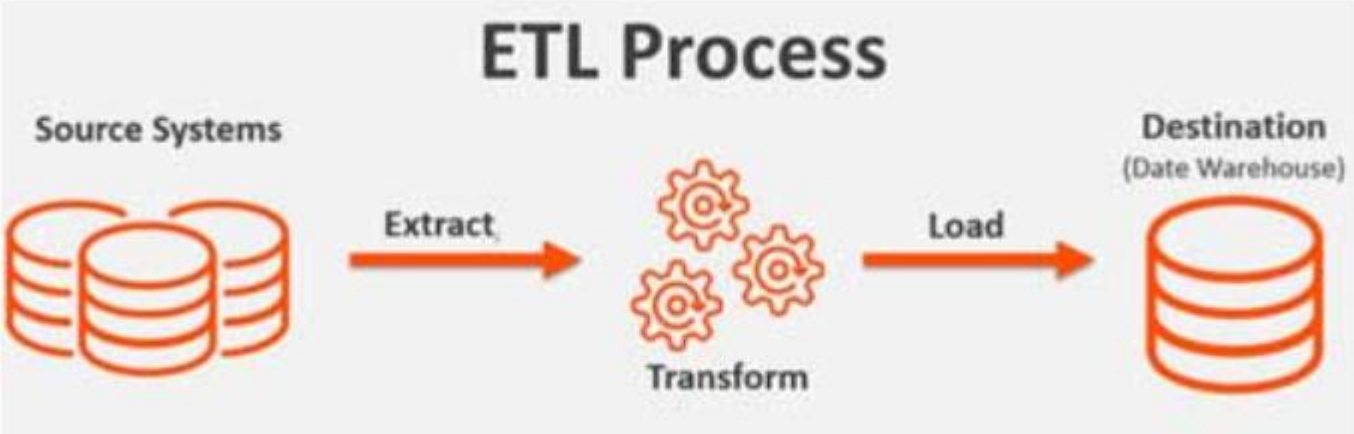
- consolidamento dei dati
  - estrarre i dati
  - renderli ***coerenti***
  - ***ripulirli*** il più possibile
- i dati vengono spesso archiviati
  - in diversi formati
  - sono spesso incoerenti tra le fonti
  - potrebbero essere ***sporchi***
    con valori internamente incoerenti o mancanti

inconsistent data

- i dati potrebbero essere ***incoerenti***
  - due mappe potrebbero avere nomi diversi per indicare la stessa parte fisica
  - rappresentare True e False, un sistema può usare 1 e 0 mentre un altro sistema può usare T e F
- i dati memorizzati in paesi diversi probabilmente memorizzeranno le vendite nella loro valuta locale
  - queste vendite devono essere convertite in una valuta comune

Extraction, Transformation, and Loading (ETL)

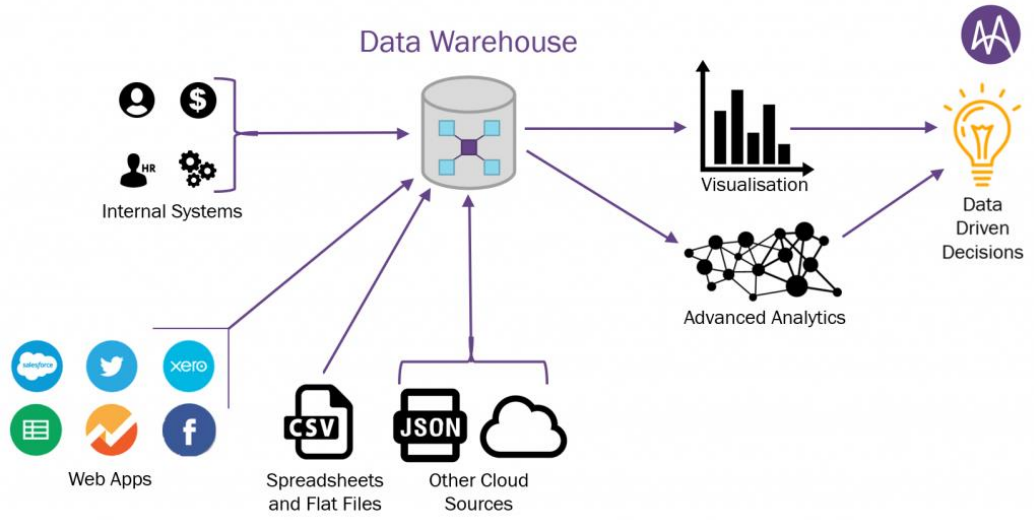
- la fase di ***estrazione*** estrae i dati dalle varie fonti
- i dati vengono quindi ***trasformati*** e resi coerenti
- infine i dati vengono ***caricati*** in un repository di dati
  - factory di dati (***data warehouse***)
- il processo ETL spesso consuma l'80% del tempo di sviluppo totale



data warehouse

*"collezione o aggregazione di **dati** grezzi **strutturati** e **non strutturati** utili ad analisi e reporting prima tramite strumenti appositi di **trasformazione** dei dati di tipo **ETL**, provenienti da fonti **interne** (DBMS) ed **esterne** al sistema informativo aziendale, e poi a strumenti di analisi di tipo OLAP (On-Line Analytical Processing), (business intelligence, data mining, big data analytics ecc…), tipicamente ad uso strategico aziendale nei processi decisionali di business."*

Wikipedia



i dati

Alberto Ferrari – Big Data

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»

struttura dei dati

- dati ***organizzati/ strutturati***
  - ordinati in una struttura a righe e colonne
    - ogni riga rappresenta un'unica osservazione
    - le colonne rappresentano le caratteristiche di tale osservazione
  - formato XML o JSON
    - assieme ai dati contengono i metadati che definiscono i nomi dei campi e la loro struttura.

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
TOTALE	24.005	2.515	472.598	499.118	322.925	40.638	862.681	37.809	562.010	300.671	862.681	10.290.151	16.951.896	234.245

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

```xml
<?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>
```

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

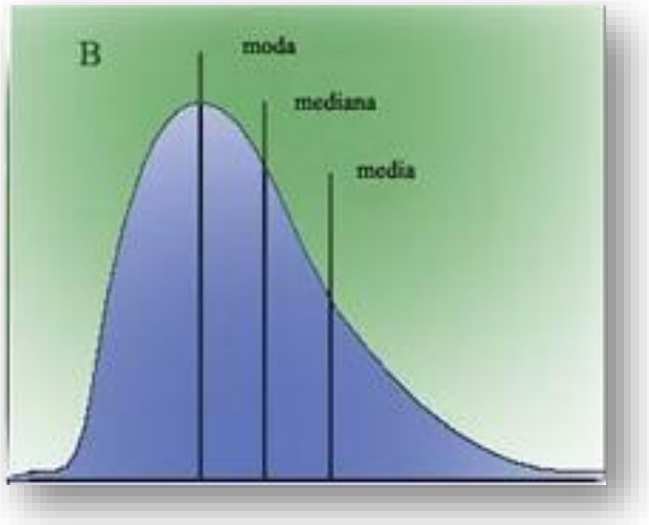
livelli dei dati

- 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

$$M_a = \frac{1}{n}\sum_{i=1}^{n} x_i .$$



esempio livello ordinale

- ***analisi di un sondaggio*** (apprezzamento con valori da 1 a 5)
- risultati di un sondaggio:

```
5, 4, 3, 4, 5, 3, 2, 5, 3, 2, 1, 4, 5, 3, 4, 4, 5, 4, 2, 1, 4, 5, 4, 3, 2, 4, 4, 5, 4, 3, 2, 1
```

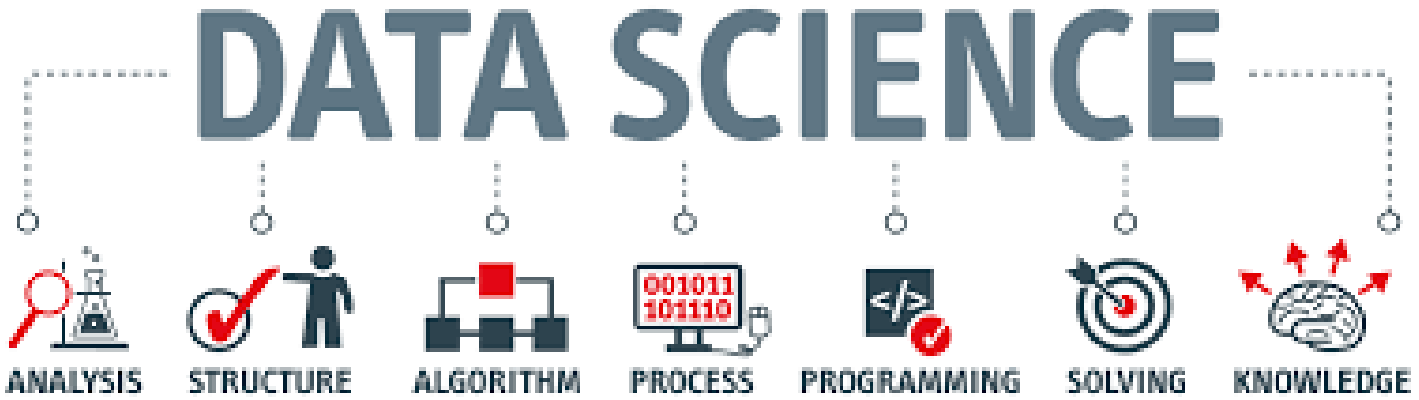
- la ***mediana*** (*il valore centrale fra i dati numerici*) è, normalmente, un modo appropriato per definire il centro dei dati
- si può anche usare la ***moda*** ma non la media perché a questo livello la somma o la sottrazione non hanno senso e neanche la divisione.
- risultati ordinati:

```
1 , 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5
```

- mediana: **`4`**
- moda **`4`**
- media: **`3.4375`** meno adatta a rappresentare il risultato

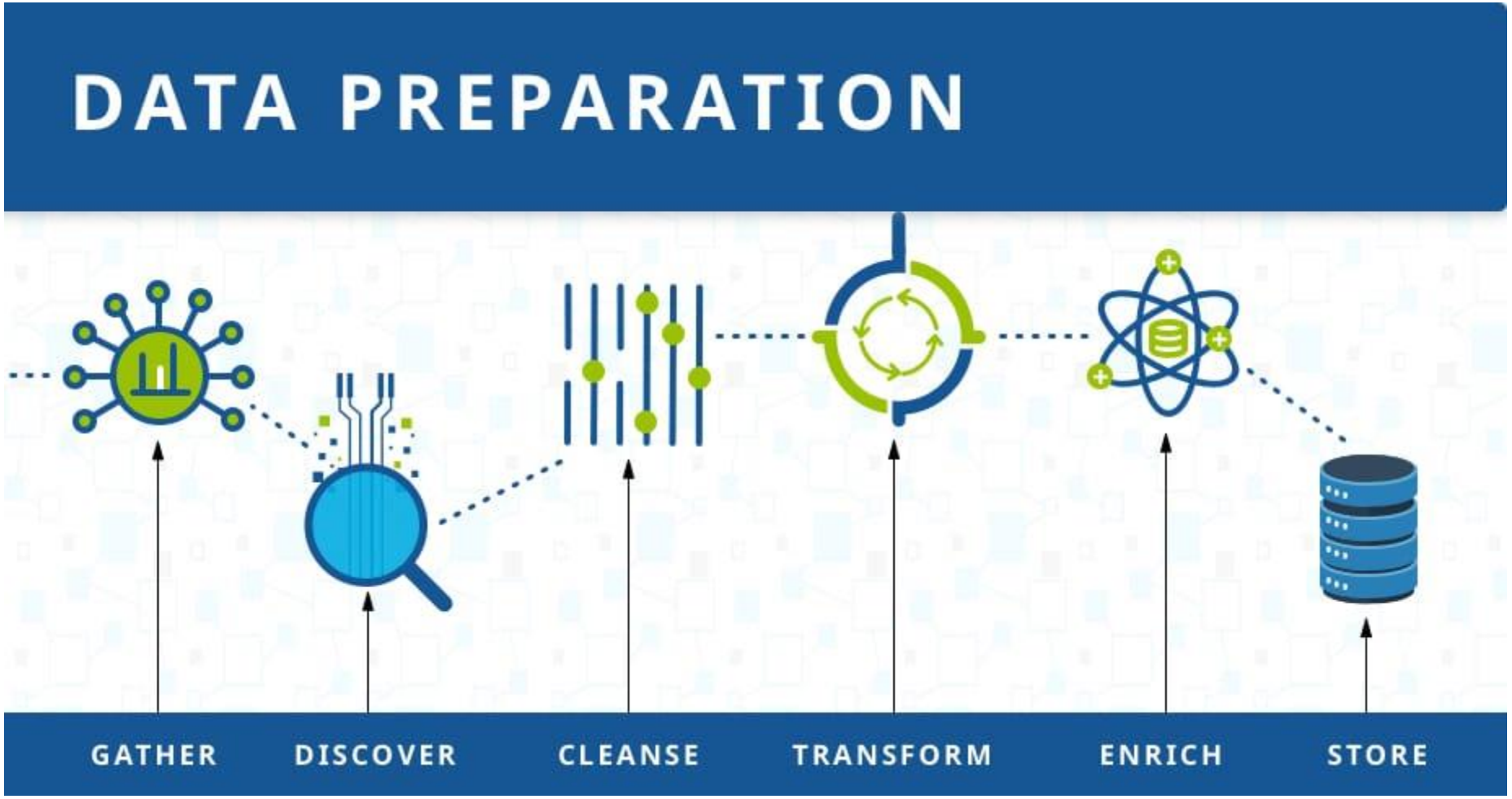
nominale o ordinale?

- 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


DATA SCIENCE
001011
101110
ANALYSIS
STRUCTURE
ALGORITHM
PROCESS
PROGRAMMING
SOLVING
KNOWLEDGE

i cinque passi 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 presentare i risultati***
  - presentare i risultati in forma chiara e comprensibile è più difficile di quanto si possa immaginare


DATA PREPARATION
GATHER
DISCOVER
CLEANSE
TRANSFORM
ENRICH
STORE

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 σ 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

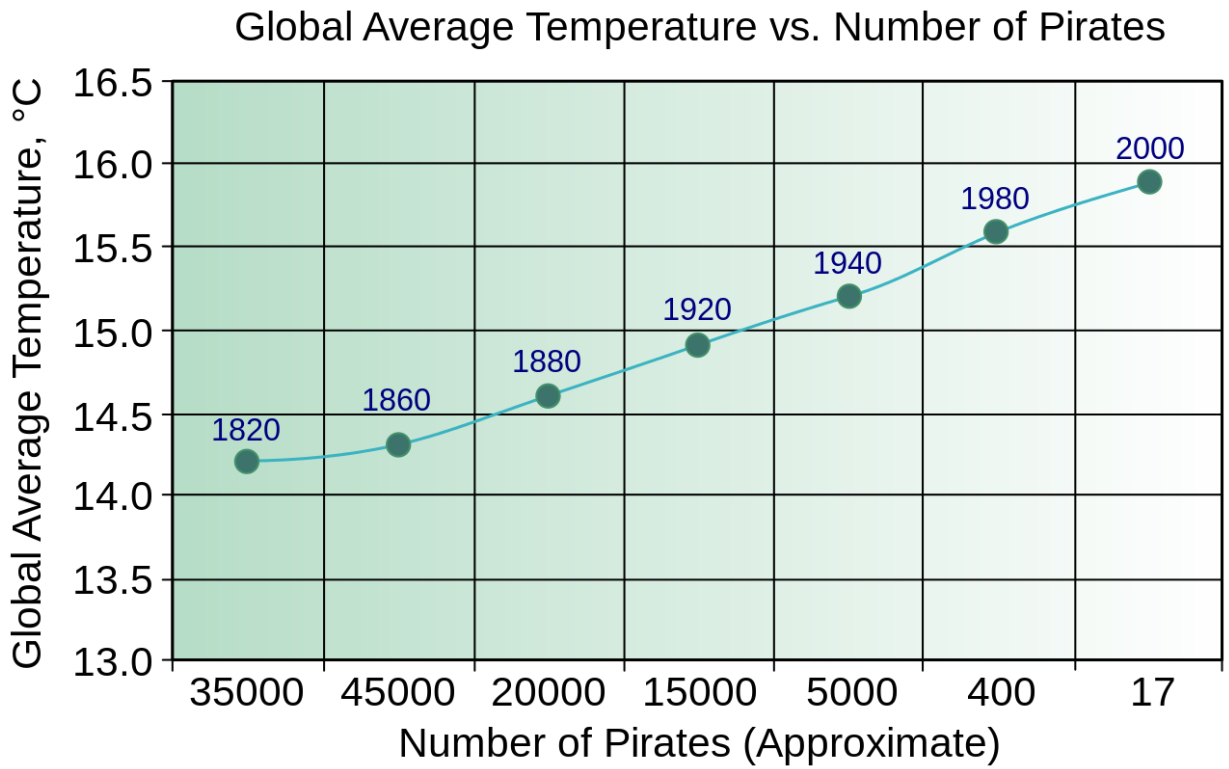
comunicare i dati

- ***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

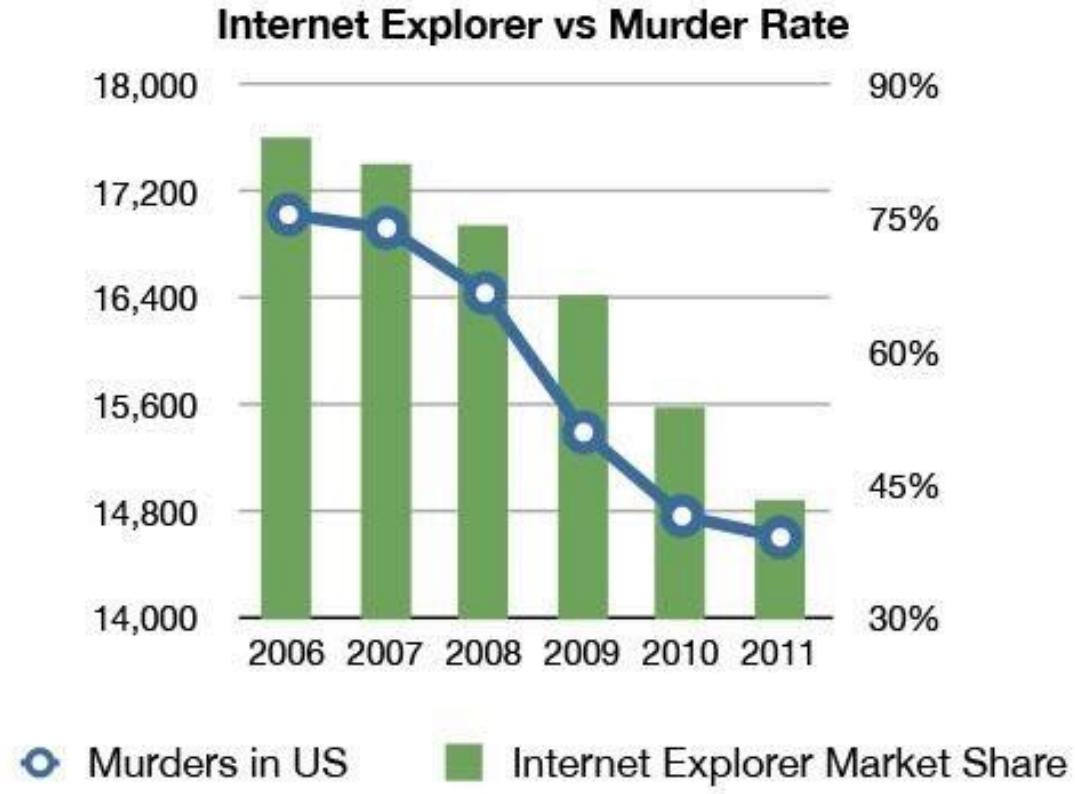
- ***"la correlazione non è necessariamente indice di causalità"***
- ***correlazione***
  - ***relazione*** tra due variabili tale che a ciascun valore della prima corrisponda un valore della seconda, seguendo una certa regolarità
- ***causalità***
  - principio interpretativo della realtà, che si fonda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

correlazione o causalità?



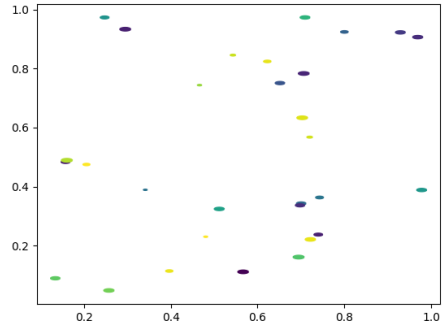
You can see that as the ***number of pirates*** in the world has ***decreased*** over the past 130 years, ***global warming*** has gotten steadily ***worse***.
In fact, this makes it entirely clear that ***if you truly want to stop global warming, the most impactful thing to do is -- become a pirate.***

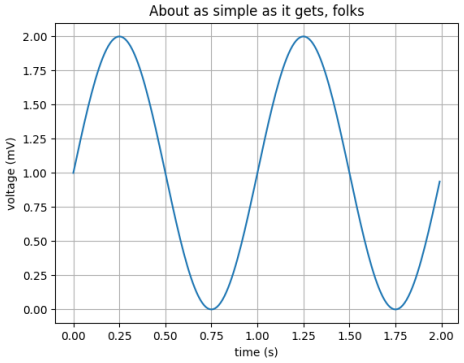
vendite di Microsoft Internet Explorer e numero di omicidi

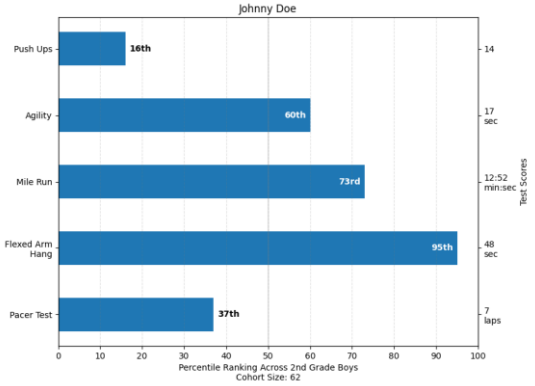


tipi di grafici

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







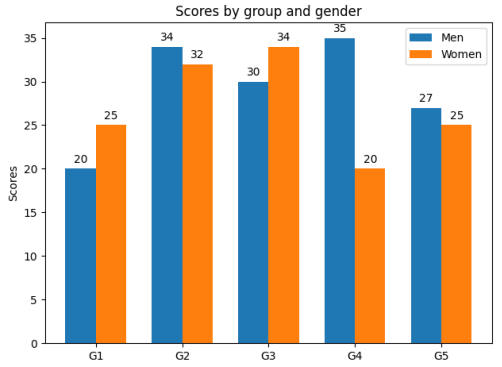
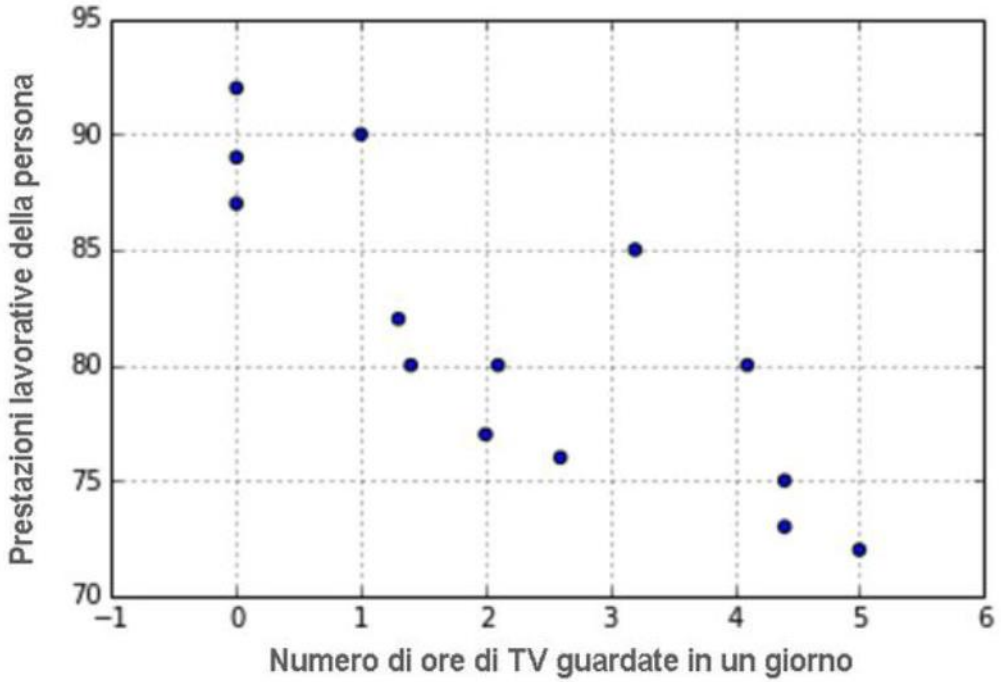


grafico a dispersione (scatter plot)

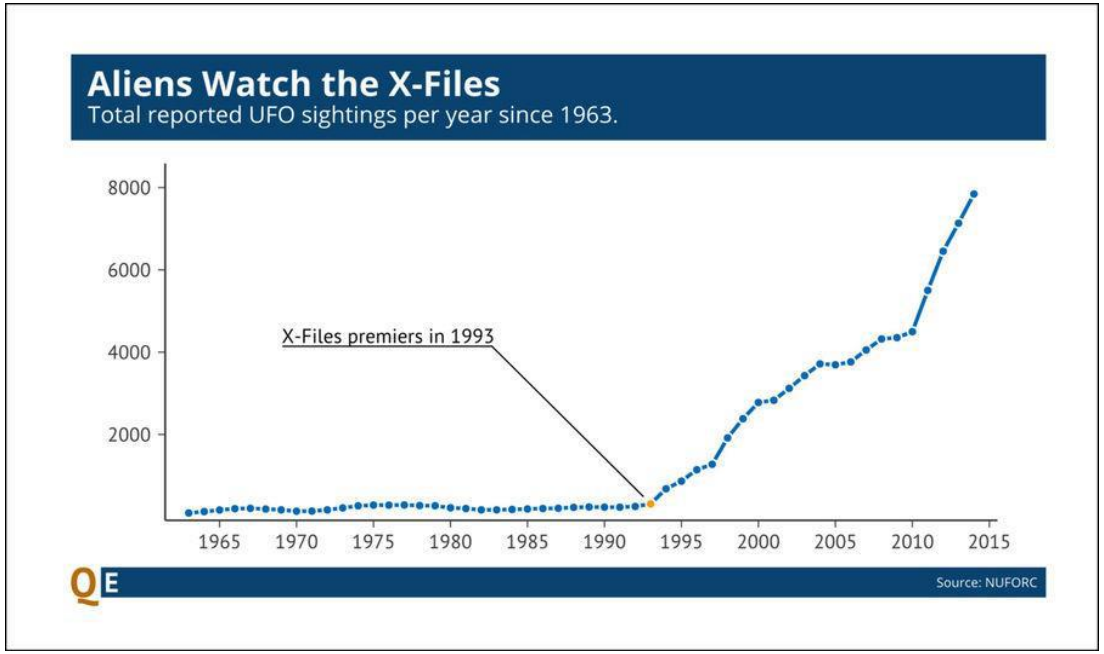
- i due ***assi*** sono ***quantitativi***
- ogni punto rappresenta un'osservazione
- si evidenziano le ***relazioni*** esistenti fra ***due variabili***
  - se possibile si individua una ***correlazione***



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)

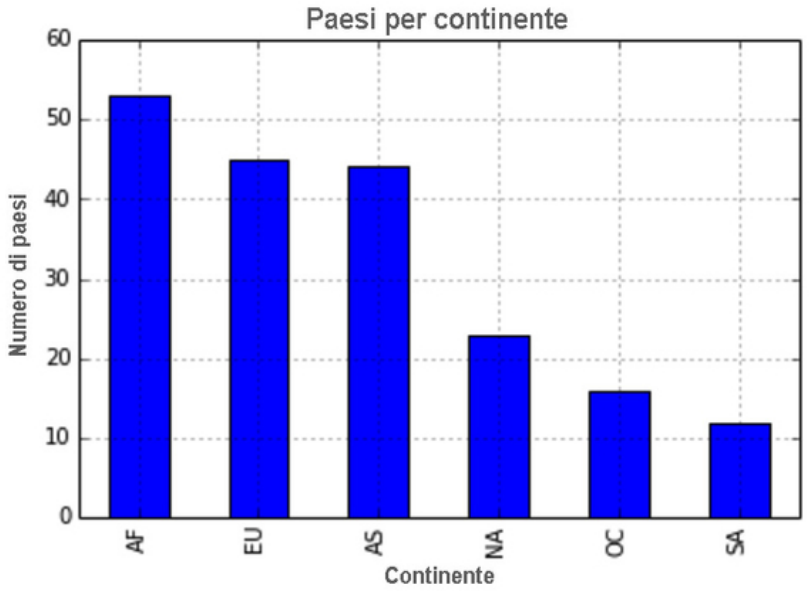
- normalmente l'asse x rappresenta una linea del ***tempo*** e l'asse y è di tipo quantitativo
- è costituito da una serie di punti (rilevazioni) uniti da linee rette



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

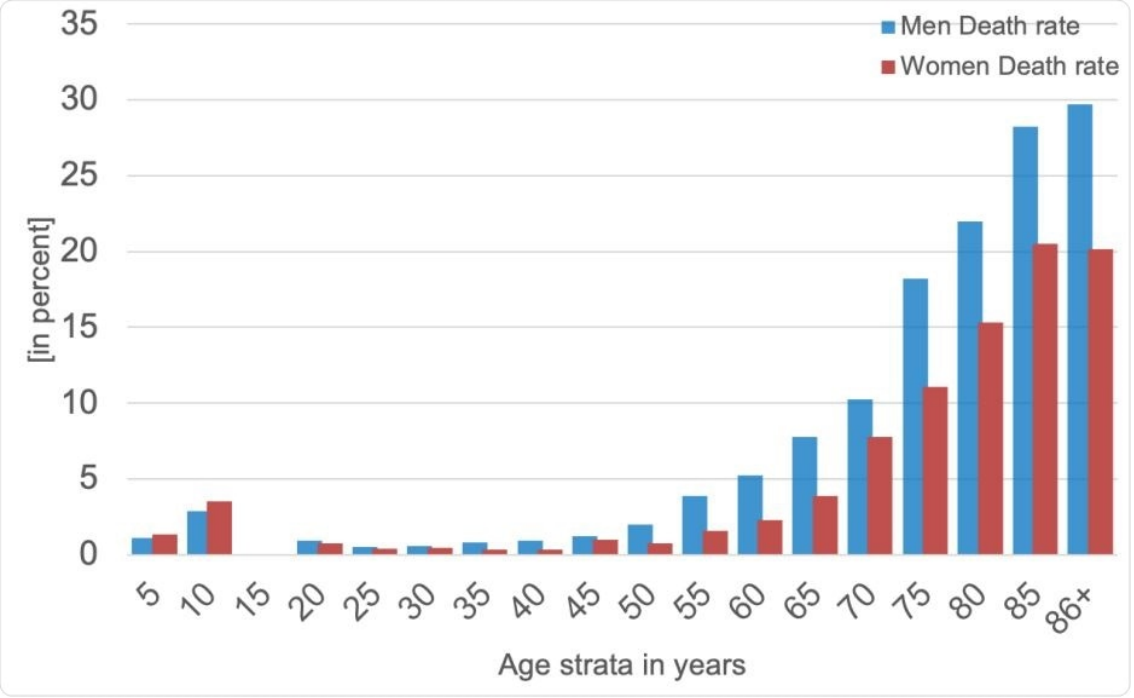
diagramma a barre (bar chart)

- 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