PCTO in Coding & Data Science

Modulo 1: Introduzione al coding

Parte A:
L’unità base: il dato

Cos’è secondo voi un dato?

Alcuni esempi di dati...

Età = 25
Colore capelli = marroni

I dati possono avere diversi **formati**, *numerico* o *carattere*.

Questi sono esempi di **dati "strutturati"**. Poi esistono i **dati "non strutturati".**

Una canzone, una foto.

Come rappresenteresti una canzone di Beyonce?

Wordcloud basata sui testi delle canzoni di Beyonce.

Fonte: https://www.databasic.io/en/wordcounter/

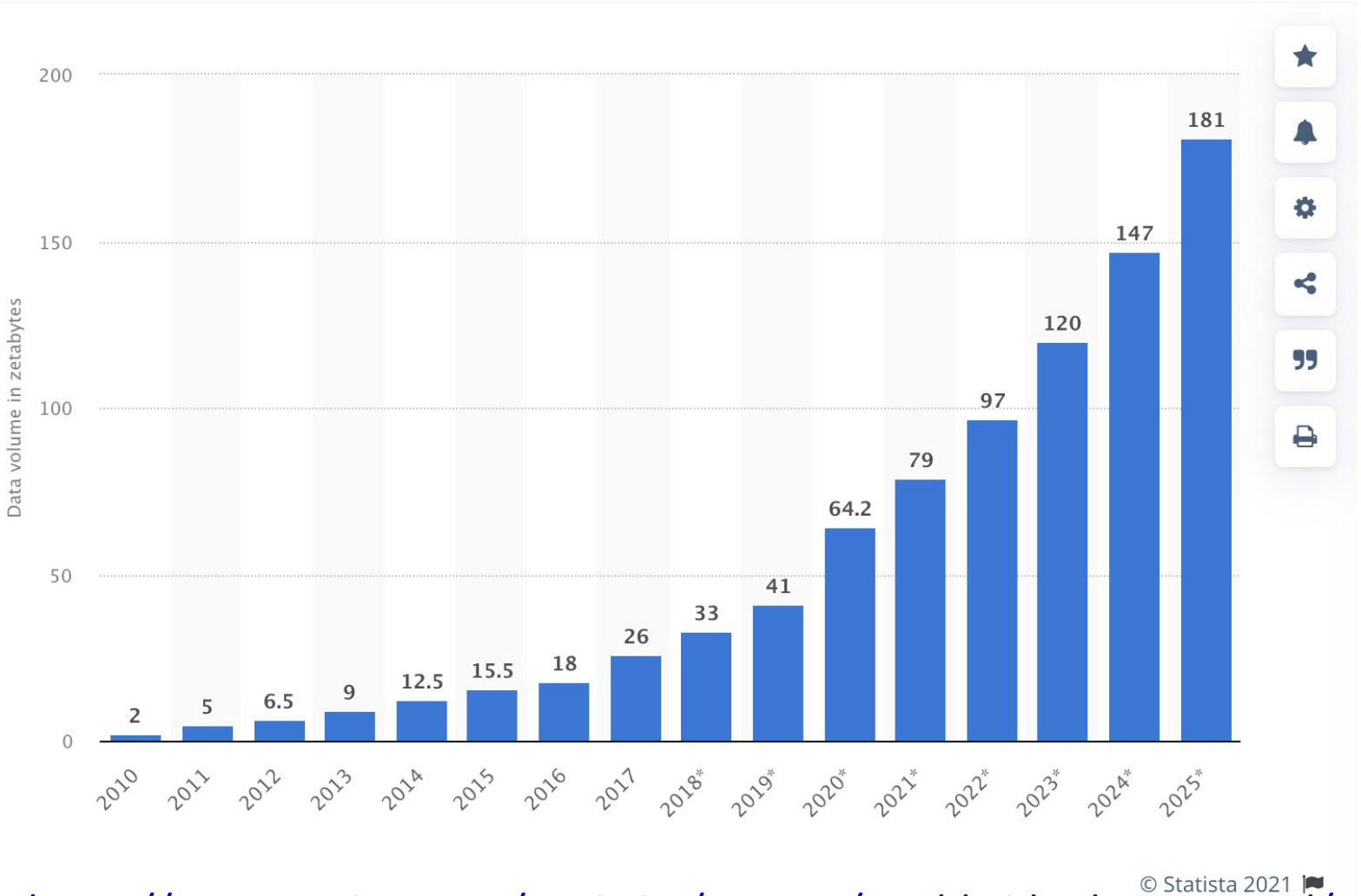
Quanti dati produciamo in un dato istante?

Produciamo dati ogni volta che...

- Accendiamo la luce
- Timbriamo il biglietto dell'autobus
- Siamo presenti all'appello in classe
- Facciamo un acquisto con la carta di credito
- Pubblichiamo una storia su Instagram
- Facciamo una ricerca su Google

  QUIZ: quante ricerche sono effettuate su Google in ogni dato istante?
- …

Nel 2020, ciascun essere umano ha creato in media 1.7 MB di dati AL SECONDO.



Fonte: https://www.statista.com/statistics/871513/worldwide-data-created/.
Uno zettabyte: 1 seguito da 21 zeri.

Nel 2018 1% dell’elettricità prodotta a livello globale era utilizzata per immagazzinare dati

Che cos'è la data science?

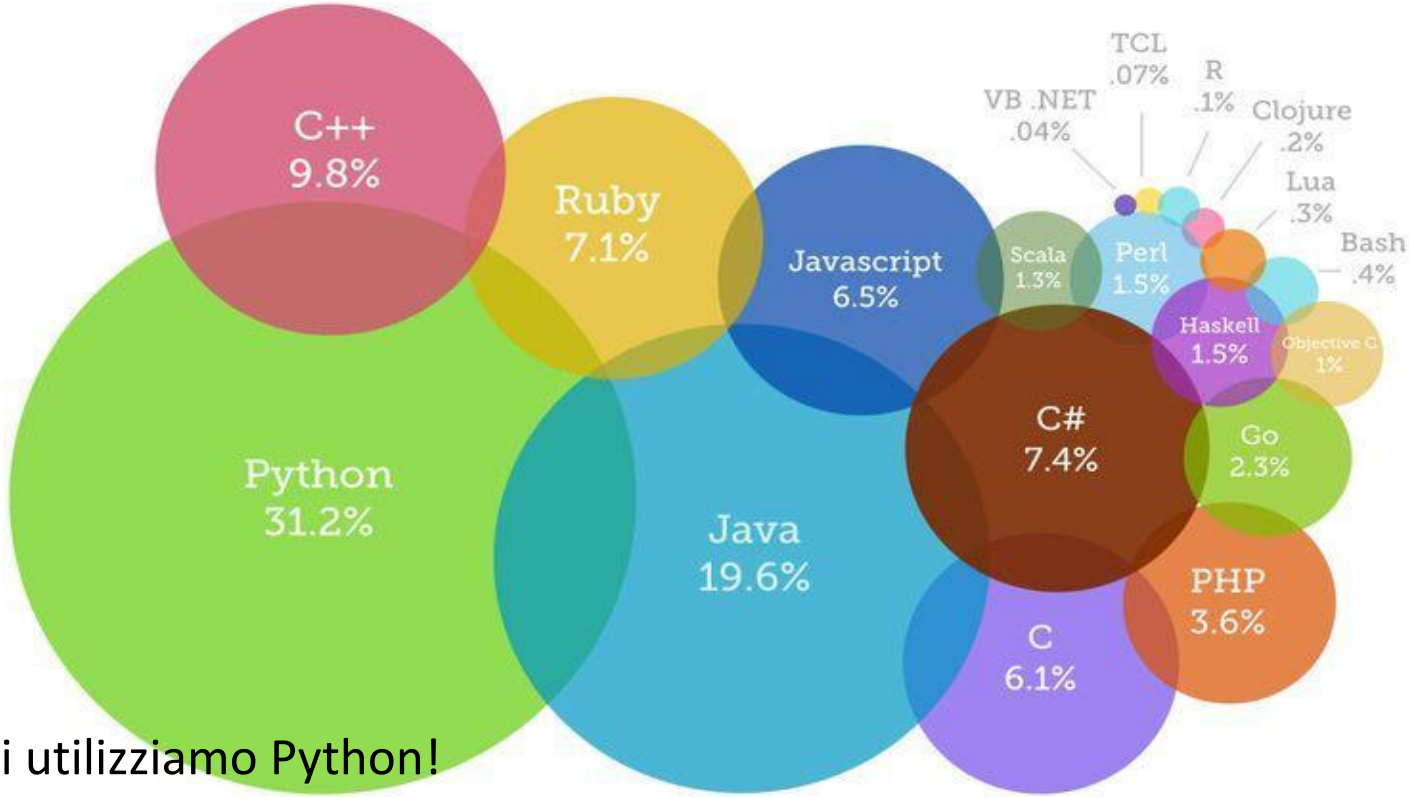
"... quella disciplina orientata all'estrazione e analisi di grandi volumi di dati ("big data") per mezzo di moderne tecniche/strumenti al fine di estrarne informazioni funzionali al miglioramento delle operazioni di una determinata organizzazione."

- Google search
- Pubblicità online
- Intelligent transport systems
- Veicoli a guida autonoma
- Ricerca medica

Nel 2020 quello del data scientist è risultato essere il terzo profilo più richiesto su Linkedin a livello globale!

Cos’è la programmazione?

“... il linguaggio che utilizziamo per comunicare con il computer.”



😎 noi utilizziamo Python!

Quali tipi di dati si possono avere in Python?

1. stringhe e numeri

```
nome = "John"          # Una stringa
```

assegnare a una variabile un valore

```
altezza = 170          # Un numero
```

2. liste

```
mylist = [ 'Annie', 160, 'John', 170 ]
```

3. dizionari: composti da chiavi (keys) e valori (values)

```
mydict = {'nome': 'John','altezza': 170}
```

Cosa ci facciamo con i dati: le funzioni

→ si applicano a degli argomenti

1. alcune funzioni pre-definite: print(), dir(), type()

```
print("Hello world")
>> Hello world
```

2. una funzione scritta da noi!

Dapprima la definiamo: definire serve per poter riutilizzare

```
def myage(x):
    print("La mia età è", x)
```

Dopo averla definita, la applichiamo:

```
myage(31)
>> La mia età è 31
```

Simili ma diversi rispetto alle funzioni: i metodi

→ si applicano a degli oggetti

Un esempio di metodo pre-definito:

```
mydict.keys()
>> dict_keys(['nome', 'altezza'])

mylist.append('Frank')
print(mylist)
>> ['Annie', 160, 'John', 170, 'Frank']
```

Simili ma diversi rispetto alle funzioni: i metodi

Un esempio di metodo scritto da noi: **l'area di un rettangolo**.
Ci serviamo delle **classi** (*"il sistema che definisce la relazione esistente tra le singole variabili e funzioni"*).

Dapprima le definizioni:

```
class Rettangolo():
        def __init__(self, latolungo, latocorto):     # attributi dell'oggetto
        self.lunghezza = latolungo                     # generico "self"
        self.larghezza  = latocorto
    def area(self):                                    # creazione del metodo
        return self.lunghezza*self.larghezza

myrectangle = Rettangolo(10, 5)                        # creiamo l'oggetto
print(myrectangle.area())                              # applichiamo il metodo!
>> 50
```

Simili ma diversi rispetto alle funzioni: i metodi

Si poteva fare più rapidamente con una funzione? Si, senza passare per lo step intermedio della creazione dell'oggetto "rettangolo", semplicemente *passando* alla funzione la lunghezza dei due lati:

```
def area(latolungo,latocorto):
    return latolungo * latocorto

print(area(10,5))
 >> 50
```

Parte B:
Un insieme di dati: il database

Cos’è il database?

“...un insieme organizzato di dati.”

Gli elementi costitutivi del database:

- le **osservazioni**
- le **variabili**
- gli **identificativi**
- gli **identificativi univoci**

Variabili e osservazioni

Variabili

Osservazioni

Mammals

Mammals (27 cases)

index	Mammal	Order	LifeSpan (years)	Height (meters)	Mass (kg)	Sleep (hours)	Speed (km/h)	Habitat
1	African ...	Probosc...	70	4	6400	3	40	land
2	Asian El...	Probosc...	70	3	5000	4	40	land
3	Big Bro...	Chiropt...	19	0.1	0.02	20	40	land
4	Bottlen...	Cetacea	25	3.5	635	5	37	water
5	Cheetah	Carnivo...	14	1.5	50	12	110	land
6	Chimpa...	Primate	40	1.5	68	10		land
7	Domest...	Carnivo...	16	0.8	4.5	12	50	land
8	Donkey	Perisso...	40	1.2	187	3	50	land

https://codap.concord.org/app/static/dg/en/cert/index.html

Identificativi e identificativi univoci

La combinazione di questi due dati identifica univocamente un'osservazione

	name	category	actbirthplace	actbirthyear	movietitle	movieyear	titletype	genres
0	Adrien Brody	actor	New York	1973	A Matador's Mistress	2008.0	movie	Biography,Drama,Romance
1	Adrien Brody	actor	New York	1973	American Heist	2014.0	movie	Action,Crime,Drama
2	Adrien Brody	actor	New York	1973	Backtrack	2015.0	movie	Drama,Fantasy,Mystery
3	Adrien Brody	actor	New York	1973	Blonde	2021.0	movie	Biography,Drama,Romance
4	Adrien Brody	actor	New York	1973	Bread and Roses	2000.0	movie	Drama
...	...	...	...	...	...	...	...	...
95	Alec Guinness	actor	England	1914	A Run for Your Money	1949.0	movie	Comedy
96	Alec Guinness	actor	England	1914	All at Sea	1957.0	movie	Comedy
97	Alec Guinness	actor	England	1914	Cromwell	1970.0	movie	Biography,Drama,History
98	Alec Guinness	actor	England	1914	Damn the Defiant!	1962.0	movie	Action,Drama,History
99	Alec Guinness	actor	England	1914	Hitler: The Last Ten Days	1973.0	movie	Biography,Drama,History

Al lavoro con il primo nostro database!

Obiettivi dell'esercitazione pratica: acquisire dimestichezza con *pandas*

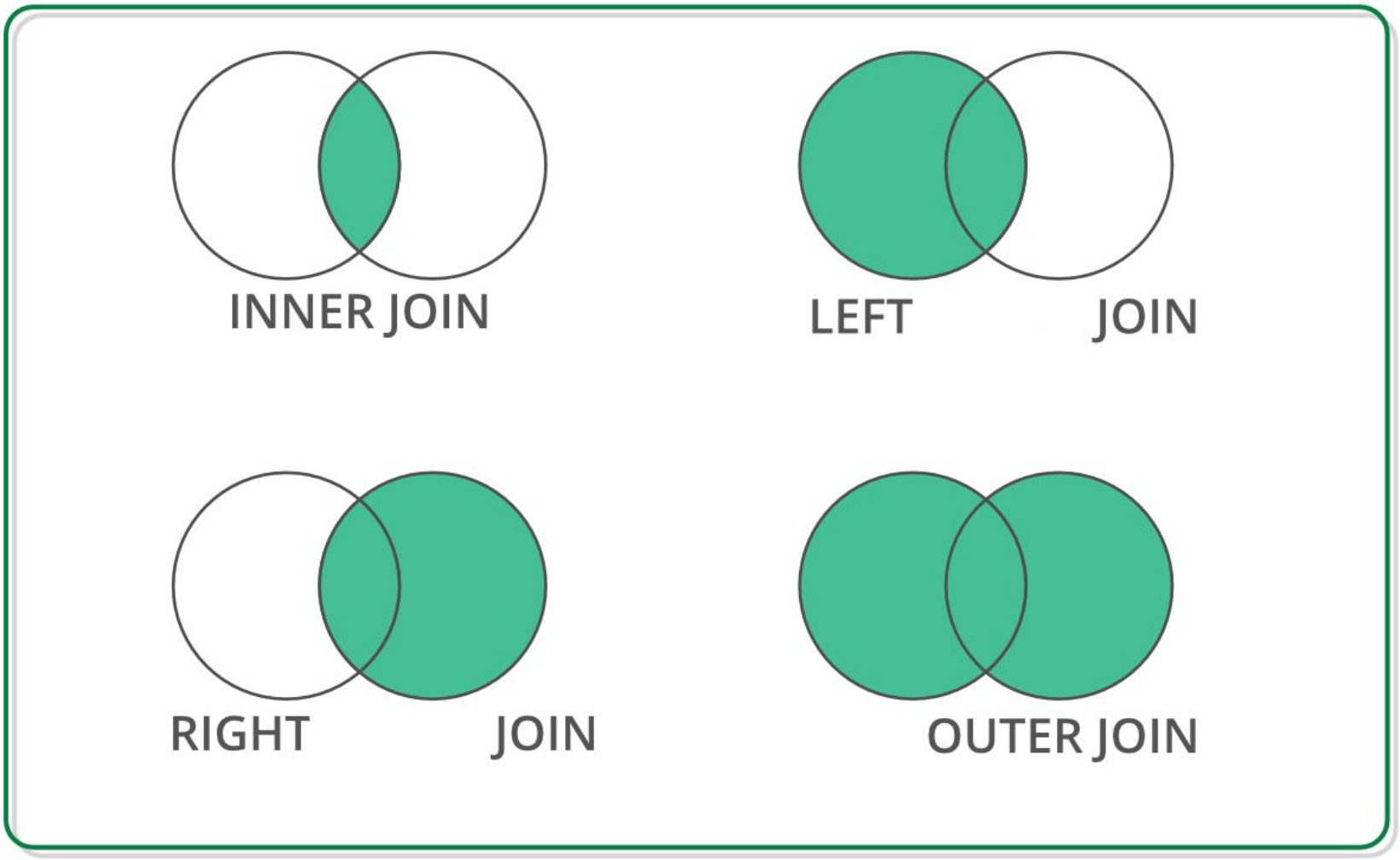
1. capire cosa vuol dire caricare librerie/pacchetti/moduli

2. riuscire a caricare il database

3. capire la struttura del database
   - numero di osservazioni e numero di variabili
   - tipi di variabili
   - identificativi e identificativi univoci
   - ispezione valori mancanti

4. semplice operazioni con un database
   - elimina una variabile, rinomina una variabile, elimina un gruppo di osservazioni, fai il merge con un altro database

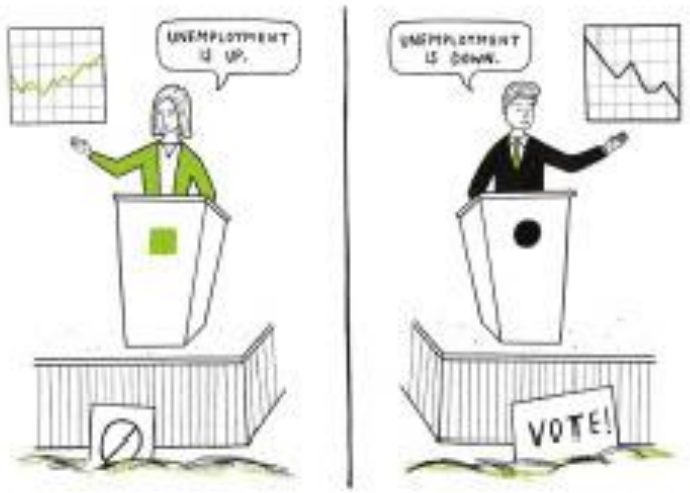
Come funziona il merge tra due database?



Rule n.1: non fidarsi mai completamente dei propri istinti

QUIZ: di quanto è aumentato il tasso di scolarizzazione primaria delle ragazze a livello mondiale dal 1970 ad oggi?

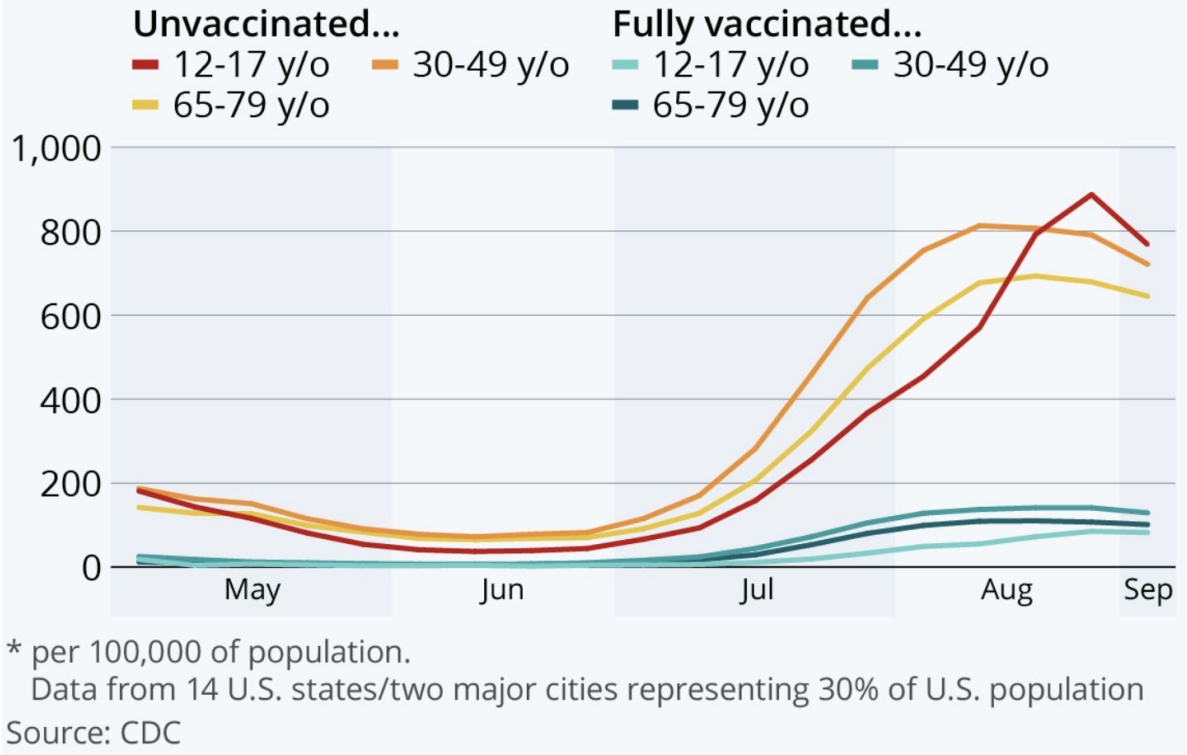
Rule n.2: guardare sempre con occhio critico ai dati presentati da altri



Il comportamento di chi riporta solamente i dati che gli fanno comodo….

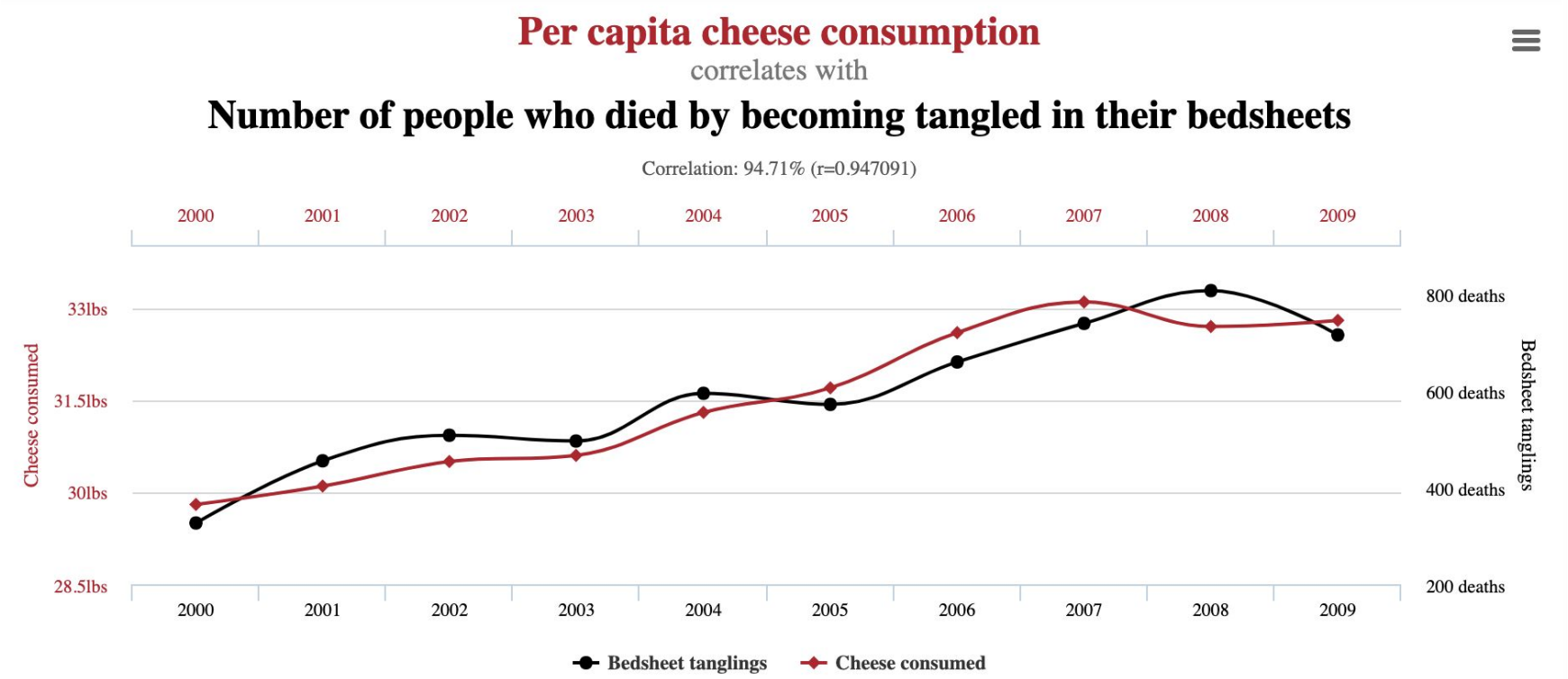
Un altro esempio di cherry picking

Quando i contagi hanno iniziato a calare quest'estate, negli Stati Uniti come altrove, in realtà calavano solamente per un sottoinsieme della popolazione...



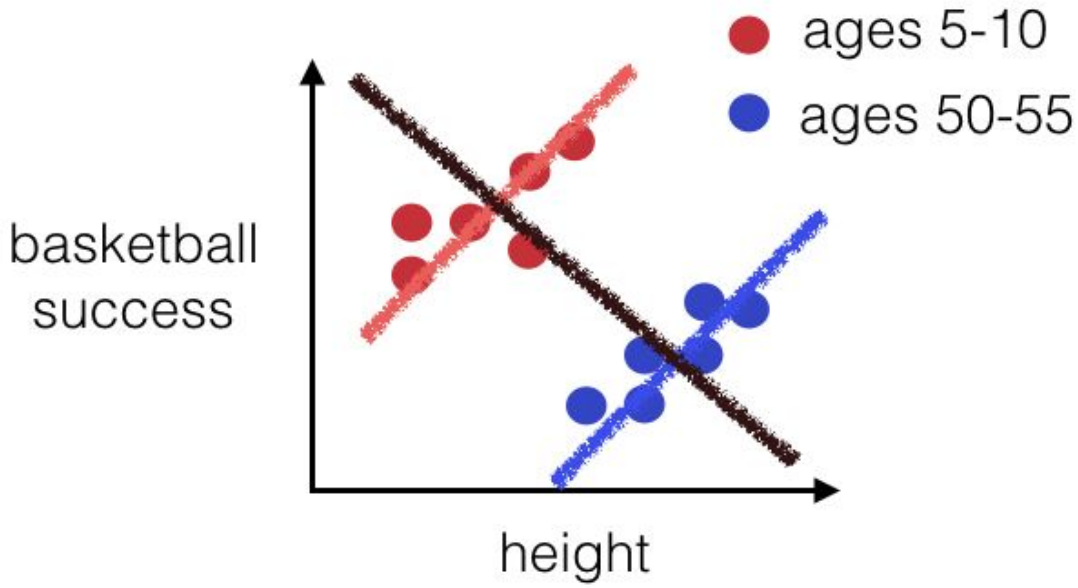
https://www.statista.com/chart/26159/covid-cases-us-age-group-vaccination-status/

Rule n.3: saper distinguere correlazione da causazione



https://www.tylervigen.com/spurious-correlations

Rule n.4: quello che è vero per dei sottogruppi, non è detto che sia vero per l'intera popolazione



Effetto Simpson: un certo risultato vale per un sottoinsieme della popolazione ma non a livello aggregato.

Bibliografia

AGGIUNGI RIFERIMENTI BIBLIOGRAFICI