Librerie per il data preprocessing

# Introduzione

Con pochi dati nel dataset dobbiamo usare modelli semplici, per evitare l’overfitting; con tanti dati possiamo usare modelli complessi

I dati all’interno del dataset sono **indipendenti e identicamente distribuiti (i.i.d.)**, cioè gli esempi del dataset sono presi dalla stessa distribuzione ma estratti in maniera indipendente gli uni dagli altri. Tutti i modelli che vedremo faranno l’assunzione di avere i dati i.i.d. 🡪 **learning proposizionale** (dati formati da un insieme di feature categoriche, numeriche o ordinali)

PS. In alcuni casi i dati possono non essere i.i.d., ad esempio, prendendo un caso in cui vogliamo spostare degli agenti all’interno di un mondo di blocchi, ovvero, dati dei blocchi tipo lego vogliamo costruire qualcosa, se un esempio è “ho spostato il blocco 0 nella posizione 3 e ho avuto successo” e un altro “ho spostato il blocco 2 nella posizione 3 e ho avuto successo”, non possiamo pensare che queste due azioni siano indipendenti; quindi, dovremo usare un altro approccio (che non vedremo), perché se occupo 3 con il blocco 0 non posso anche occuparlo con il blocco 2.

Come librerie vedremo

* **numpy** (np)
* **pandas** (pn)
* **matplotlib.pyplot** (plt) 🡪 libreria di data visualization molto usata (basso livello)
* **seaborn** (sns) 🡪 libreria di data visualization che costruisce sopra plt (alto livello, con meno righe di codice permette di ottenere dei bei risultati)