**Introduzione alla Data Science**
**Progetto**

Per ogni passaggio, commentare opportunamente e fornire giustificazioni delle scelte operate.

## Dataset

Ti sono stati assegnati i dataset

- netflix_titles_1.csv
- netflix_titles_2.csv
- netflix_credits.csv
- amazon_titles_1.csv
- amazon_titles_2.csv
- amazon_credits.csv

## 1. Integrazione

Per ciascun gruppo di dati hai 2 tabelle ***titles***. Procedi con i seguenti passi:

- Fai join tra queste 2 tabelle per ottenere una tabella integrata
- Crea una nuova tabella <piattaforma>_titles_combinata.csv ottenuta a partire da <piattaforma>_titles_1.csv ed aggiungendo le colonne *date_added* e *country* che trovi in <piattaforma>_titles_2.csv
- Fai join tra la tabella ottenuta e la tabella credits
- Procedi con la pulizia dei dati: elimina colonne inutili o ripetute, righe non significative, gestisci la presenza di eventuali valori nulli o mancanti, ecc...

Dopo aver preparato i due dataset, puoi decidere di concatenarli per ottenerne uno unico, oppure mantenerli separati.

## 2. Trasformazione

- Sostituisci la colonna *date_added* con 2 colonne *year_added* e *month_added*
- Sostituisci la colonna *genres* con una colonna *genres_number* che contiene il numero di generi associati a quel dato

Raggruppa le righe che contengono informazioni su uno stesso dato (questa situazione si crea quando facciamo join tra tabelle titles e tabelle credits), e crea

- una colonna *cast_number* che conta il numero di attori in un film/tv show
- una colonna *director* che riporta il nome del regista

## 3. Esplorazione

Visualizzare con boxplot la distribuzione del numero di attori nelle due categorie FILM e TV SHOW separando i campioni delle due piattaforme streaming

Rappresentare la distribuzione della durata di FILM e TV SHOW.

4. **Test statistici**

   Verificare con un test statistico se la distribuzione del numero di attori in TV SHOW e FILM è diversa tenendo separate le piattaforme, e poi riunendo tutto per un unico test statistico

   Verificare con un test statistico se la distribuzione della durata dei FILM e TV SHOW sulla prima piattaforma che stai analizzando è diversa da quella della seconda, tenendo prima separati i dati dalle due piattaforme e poi riunendo tutto per un unico test statistico

5. **OLAP**

   Costruire una rappresentazione OLAP che conteggi i dati nelle due piattaforme raggruppando per

   - Anno di caricamento sulla piattaforma
   - Tipologia (TV SHOW o FILM)
   - Paese di produzione

   Proporre e discutere 2 visualizzazioni

6. **Metodi predittivi**

   Costruire un descrittore composto da

   - type
   - imdb_score
   - tmdbs_score
   - tmdbs_popularity
   - runtime

   Applicare K-Means per individuare eventuali gruppi coerenti tra i dati. Calcolare l'indice di Silhouette per individuare il numero migliore di cluster

   Applicare PCA per poter visualizzare i dati, colorando ogni campione secondo l'appartenenza ad uno dei cluster.

   Verificare cosa succede tenendo i dati separati per piattaforma.