**Project 2: ETL**

**By Emily Zhu, Surabhi Mukati, Stephen Kanaski, Hayden Rubin**

**Extract:**

* We extracted Netflix TV shows’ IMDB ratings by web scraping. After scraping, we input the content into a data frame. Below is a link: <https://www.imdb.com/search/title/?companies=co0144901&ref_=adv_prv>
* We also extracted a CSV file with a list of Netflix shows and movies from Kaggle. <https://www.kaggle.com/shivamb/netflix-shows>

**Transform**:

* The data cleaning and transformation that was done on the Netflix TV shows’ IMDB ratings file consisted of:
  + Dropping shows that didn’t have any ratings.
  + Using groupby() on the titles and using the average ratings as the rating for the shows that had multiple ratings for multiple episodes.
* The data cleaning and transformation that was done on the Netflix movies and shows CSV file consisted of:
  + Using loc to extract just the tv shows.
  + Dropping columns that we weren’t interested in, but including “type”, “title”, “country”, “release year”, and “category”.
  + Looking at the number of unique titles and dropping duplicates.

**Load**:

* We connected our Jupyter Notebook to a local PostGres database.
* Exported data frames into PostGres as tables.
* Final tables are: Netflixratings and Netflixshows.