horizontal line

**Amy Guan**

**Beth Emborsky**

**Shanker Nair**

Movie Database

**28th July 2020**

# OVERVIEW

Our project is to store movies, actors, producers, genres and any additional data in a relational database using Pandas to filter and clean the data prior to loading into SQL.

# DATA SOURCES

1. Kaggle IMDb datasets
   1. <https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset>
      1. IMDb movies.csv
      2. IMDb names.csv
      3. IMDb ratings.csv
2. *Guidebox API – maybe too much??*
   1. [*https://www.guidebox.com/*](https://www.guidebox.com/)
      1. *Gives current movies on streaming networks*

# TASK BREAKDOWN

Extracting data

1. Create python code to convert IMDB data CSVs into panda dataframe
2. *Create python code to pull Guidebox API data*

Transforming data

1. Complete ERD to show which tables to build
2. Form tables in PostGres w/ primary keys
3. Clean, merge, filter data to form matching PG tables

Loading datasets to Database

1. Load data from Panda dataframes into SQL