**Project 2 Proposal - ETL of Youtube Videos**

**Group Members:**

* Ludan Zhang
* Yuan Zhong
* John Stauffer
* Bryan Carro

**ETL Goal:**

* We will use jupyter notebook and pgAdmin4 to Extract, Transfer, and Load two datasets into our PostgreSQL database.
* The first dataset we are working with is a csv file that contains top trending youtube videos in the United States. We will remove unnecessary columns such as “tags” associated with the video, and dislikes because that feature has been reduced on youtube. Therefore, this table will contain information like views, likes, and published time.
* The second dataset we are working with is a json file that contains the video categories from Youtube’s API, which is where the csv is sourced from. The actual category id and category name are contained within a nested dictionary. We need to extract that information and organize it so that it can easily be join with the videos table.

**Sources:**

* [Videos](https://www.kaggle.com/datasnaek/youtube-new/version/115?select=USvideos.csv) (csv)
* [Categories](https://www.kaggle.com/datasnaek/youtube-new/version/115?select=US_category_id.json) (json)

**Future Areas of Analysis Made Possible by Database:**

This database will allow analysts to investigate a variety of different questions related to videos and a number of different variables:

* Differences between different video categories in count of trending videos
* Differences between categories in the correlation between likes and views
* Average views of trending videos
* Relationship between published month and average view