**Group 4 Project Proposal: Youtube Video Views by Country**

* **The sources of data that you will extract from:**
  + We will be pulling Youtube data for US, CA, and GB (United States, Canada, and Great Britian), cleaning up the original CSV to narrow views for top performing videos, and merging together a relative .json which includes identifiers for video categories so we can also narrow in on top performing video categories
    - US:
      * <https://www.kaggle.com/datasnaek/youtube-new?select=USvideos.csv>
      * <https://www.kaggle.com/datasnaek/youtube-new?select=US_category_id.json>
    - CA:
      * <https://www.kaggle.com/datasnaek/youtube-new?select=CAvideos.csv>
      * <https://www.kaggle.com/datasnaek/youtube-new?select=CA_category_id.json>
    - GB:
      * <https://www.kaggle.com/datasnaek/youtube-new?select=GBvideos.csv>
      * https://www.kaggle.com/datasnaek/youtube-new?select=GB\_category\_id.json
* **The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).**
  + Remove unnecessary columns from each individual CSV (unique columns such as “description” and “tags” etc.)
  + Convert the time stamp to UTC for when viewed
  + Join together the CSV with the JSON on the category id primary key
  + Filter by views from top to bottom
  + Merge all tables together to create a “world view” master table
  + Sort based off of country
* **The type of final production database to load the data into (relational or non-relational).**
  + Postgres – non-relational; the tables only relate on category id
* **The final tables or collections that will be used in the production database.**
  + Tables for each country joined together with category id descriptions
  + Tables for sorted by top views in each country
  + Master table with all countries and videos