# Wrangle Report

In this project, we are interested to extract and analyze data from the WeRateDogs twitter feed based comments, images, ratings, favorite and retweet counts.

But the original twitter archive did not provide all the information we needed to perform our analysis. Therefore we gather our data from another machine learning project that recognized dog breeds based on dog pictures and twitter API to obtain each tweet's favorite and retweet count.

The second step was to assess the data. Our new database was suffering from several quality and tidiness issues, including:

## Quality Issues

1. Missing data: There are 2356 different tweet id in total from the twitter archives, but
   - Data are missing from the image predictions files that contained information on 2075 tweet id. Those missing image rows needed to be excluded as we want tweet with images only
   - 59 rows are missing expanded_urls information
   - Favorite count and retweet count information are missing for 11 rows

2. Redundant information such as retweets or reply to a tweet that need to be excluded as we care only for original tweet

3. Rating extraction issue
   - The rating extraction program did not take into account decimals
   - The rating extraction program only considered the first "/" as a rating while 24/7, 7/11 are not ratings in the context

4. Dog Name extraction issue
   - Words such as "a", "an", "the", "this" were considered as dog name

5. Inconsistent expanded urls
   - Some extended urls contains two differents or duplicated urls
   - Some also links to gofundme.com, us.blastingnews.com, facebook, youtube, google links

6. Wrong data type
   - The data type for timestamps should be datetime64 type instead of object

7. Predictions name not related to dog breeds
   - For p1, p2, p3, there are some surprising entries such as umbrella, laptop, mailbox, snail...

8. Suspicious sources
   - Some sources seems to come from vine.co. Is it reliable ?

## Tidiness Issues

1. Dog stage column headers are value, not variable names
    - Dog stages can be consolidated from 4 columns to one

2. Data consolidation
    - The three dataset can be merged into one based from tweet_id to facilitate data wrangling and analysis

After assessing the quality and tidiness of the dataset, we organize the data cleaning.

- Whenever possible, we strive to clean the data programmatically using list, functions, panda library.
- For the data extraction issues, we used regex to identify regular expression in tweet text and extract information such as rating, dog names, dog stages

At the end, the dataset end up from 2356 rows to 1970 rows but ready for some data analysis.