# WRANGLE REPORT

- Sanjay (26/5/2020)

Note: The report is also present as an internal document in the wrangle_act.ipynb using the markdown cells.

## A) Gather :

We gather from three types of resources in three different formats as follows.

- df1:
  * Twitter data
  * Format - CSV file
  * Source - Manual Download of the dataset
- df2:
  * Image predictions
  * Format - TSV file
  * Source - Requesting the resource URL
- df3:
  * Twiiter additional data
  * Format - json file
  * Source - Requesting from twitter API

## B) ASSESS :

- We assess the data gathered to observe and find a minimum of 8 cleanliness(dirty data) issues and 2 tidiness(messy data) issues.
- The assessment was conducted visually to give starting insights and the above code assesses the issues programmatically to make observations.
- Note: There might be many assessments that could have been made, but here are the top issues that need to be cleaned in the future cleaning steps to yield better results.

- **Cleanliness :**

  1) High null values in columns like 'in_reply_to_status_id', 'in_reply_to_user_id','retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'.

  2) 'expanded_urls' have null values. (Image source URL is missing)

  3) Dog stages have null values represemted with a string "None".

  4) Datatype of 'timestamp' and 'retweeted_status_timestamp' is object instead of datetime64.

  5) Datatype of 'tweet_id' is integer instead of an object(string) in df1,df2.

6) Retweets are present which are supposed to be excluded.

7) Source is has html attribute like the tags <a, href> complicating things.

8) Unstardadized Rating scores. For instance having denominators as zero is absurdly wrong.

- **Tidiness :**
  1) Having 3 data frames df1,df2, and df3 instead of a single dataset

  2) Dog stages are represented in four columns when they can better be represented in one 'dog stage' column.

  3) Having multiple non-necessary columns

## C)Clean :

1. We clean the issues assessed in the gathered data to make a master dataset usable for visualization.
2. High null values in columns like 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id','retweeted_status_user_id,'retweeted_status_timestamp' is treated as follows :
    i. Drop 'in_reply_to_status_id' and 'in_reply_to_user_id' columns(Since they're not required for the analysis phase).
    ii. Fill in the null values of 'retweeted_status_id', retweeted_status_user_id & 'retweeted_status_timestamp' with zeros .
    iii. Note: Anyways retweeted columns are going to be dropped later once we exclude the retweets.
3. 'expanded_urls' have null values that need to be dropped as the Image source URL is missing and is not feasible to find.
4. Replace the string 'None' of 'doggo', 'floofer', 'pupper', 'puppo' (dog stages) with empty strings.
5. Convert the datatype of 'timestamp' and 'retweeted_status_timestamp' to datetime64 instead of the object.
6. Convert the datatype of 'tweet_id' to an object(string) instead of an integer because 'tweet_id' is a nominal data
7. Exclude Retweets to only have original tweet threads.
8. Remove the HTML tags from the source.
9. Standardize ratings at a 10 point scale.
10. Represent the four columns 'doggo', 'floofer', 'pupper', 'puppo' in one 'dog_stage' column.
11. Make a single master dataset from the 3 dataframes df1c,df2c, and df3c.

12. Remove the non-necessary columns to make it precisely usable master-dataset