Wrangle Report :

The data around us in the world is in good quality and in strong structure, thus will affect the insights gatherd from the data. So before any steps of making a dcesion or prediction for example everyone need to sure that it's data in good qulaity and tidness to make observaions and excellent visualization.

In this project I will do the most important steps in the wrangling phase which are Gathering, Assessing and Cleaning:

- Gathering:

In this step, The analyist should collect data from anywhere and with different extennsions of files like what the anlysisng did, then save them or load it to your computer. Some times you need to collect the data using programmatic way which will make it more efficient like what anlaysing did with Image_prediction data:

In [ ]:
```
## Download and save Images data:
url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'

response = requests.get(url)


with open(url.split('/')[-1], mode = 'wb') as outfile:
    outfile.write(response.content)

imagesdf = pd.read_csv('image-predictions.tsv' , sep='\t')
#** DONE

## loading archive file :
archivedf = pd.read_csv('twitter-archive-enhanced.csv')
#** DONE

## load download json file :
tweetdf = pd.read_json('tweet-json.txt' , lines=True)
#** DONE
```

The gathering now is done with different ways:

Note : There is an efficient way i read about it in the guidlines which using Twitter API, but the analysis did not use it.

- Assessing:

After gathering data, analyst is going to focus on two aspects quality and tidness of data.

Here is some problems found in qulaity and tidness in each data

---

Tweet dataframe:

Quality issues :

- Missing values in some cloumns
- many columns that are not needed such as 'qouted_status'
- Id type is int64 and of course we need the identifire to be str so I will change it to be object.
- `display_text_range` has a range with fixed start value equals to 0 so need to fix it to be an integer.

Tidness :

- Id column should named 'tweet_id' as the others data have.

Archive dataframe:

Quality issues :

- Missing values in some cloumns
- `rating_numerator` and `rating_denominator` should merge to be `rating_value` as one columns after fix them wrong values and the type of the rating should be float64 as it division.
- Id type is int64 and of course we need the identifire to be str so I will change it to be object.
- `timestamp` column should be converted to date if we need to deal with it.

Tidness :

- the stages of dogs are seperated in many columns and it should be merged in one column.

---

images dataframe:

Quality issues :

- Id type is int64 and of course we need the identifire to be str so I will change it to be object.
- there are non dogs prediction like some fruits, this is not related to the data, it should be deleted.

Tidness :

- some unneed columns, which should be dropped.

- Cleaning:

The final step is cleaning the data above and merge them in one datafram if it can be, this step depends very much on assessing step.

the analysis will fix each quality and tidness issues wroted in assessing step, and find some more features to add or droping som unneeded data.

cleaning depends very much on what you want to do, what questions you ask for? what information you need to get at the end.

Some questions I ask before do cleaning after mixed up the datasets are :

- What is the most kind of dogs is prefered ?
- Who have got the highest favorites and retweets?
- What is the top frequent predictions of dogs have occured?
- What is the most name repeated?

At the end I answerd taht questions and wrote the observasions of what I found.