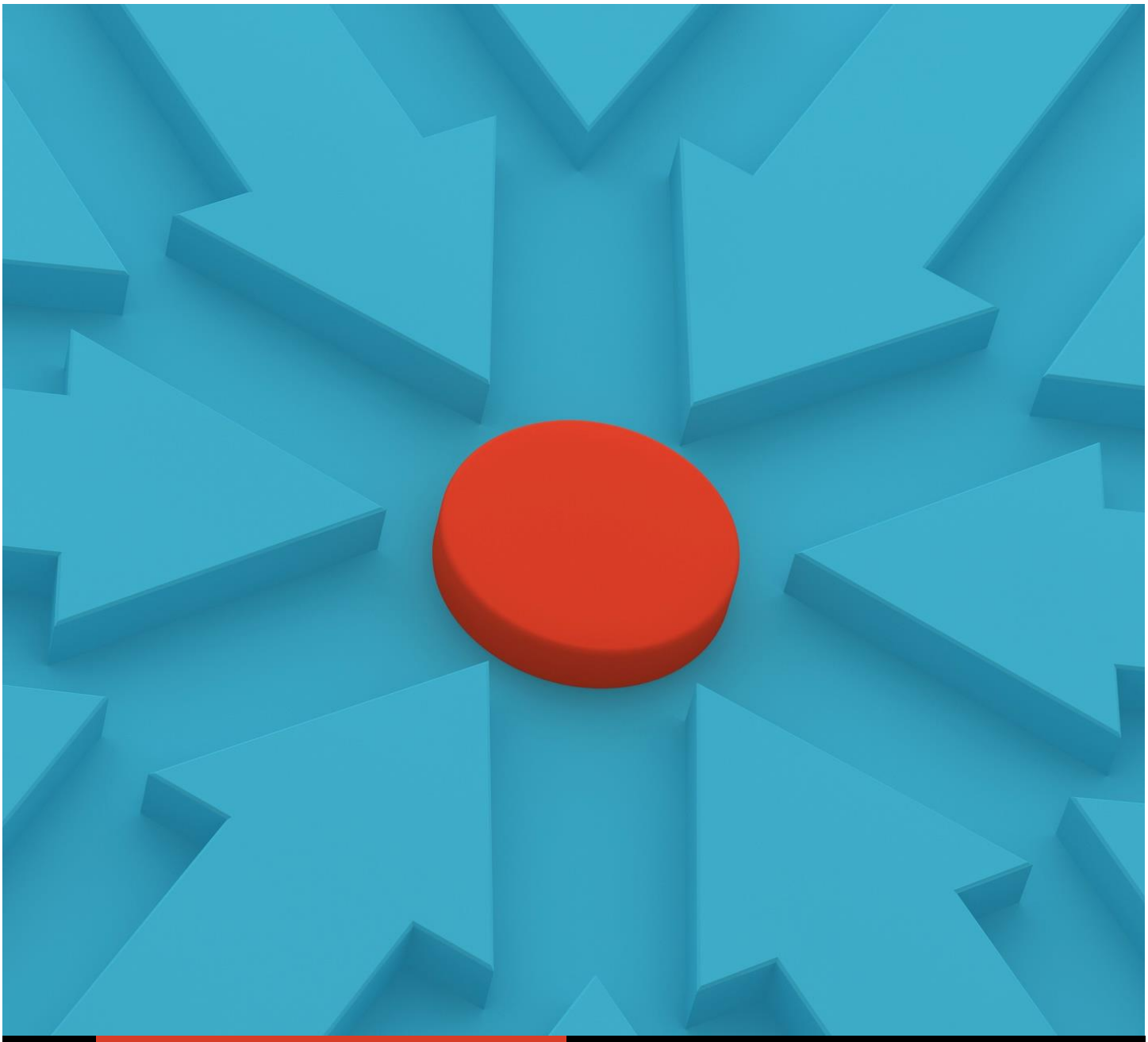
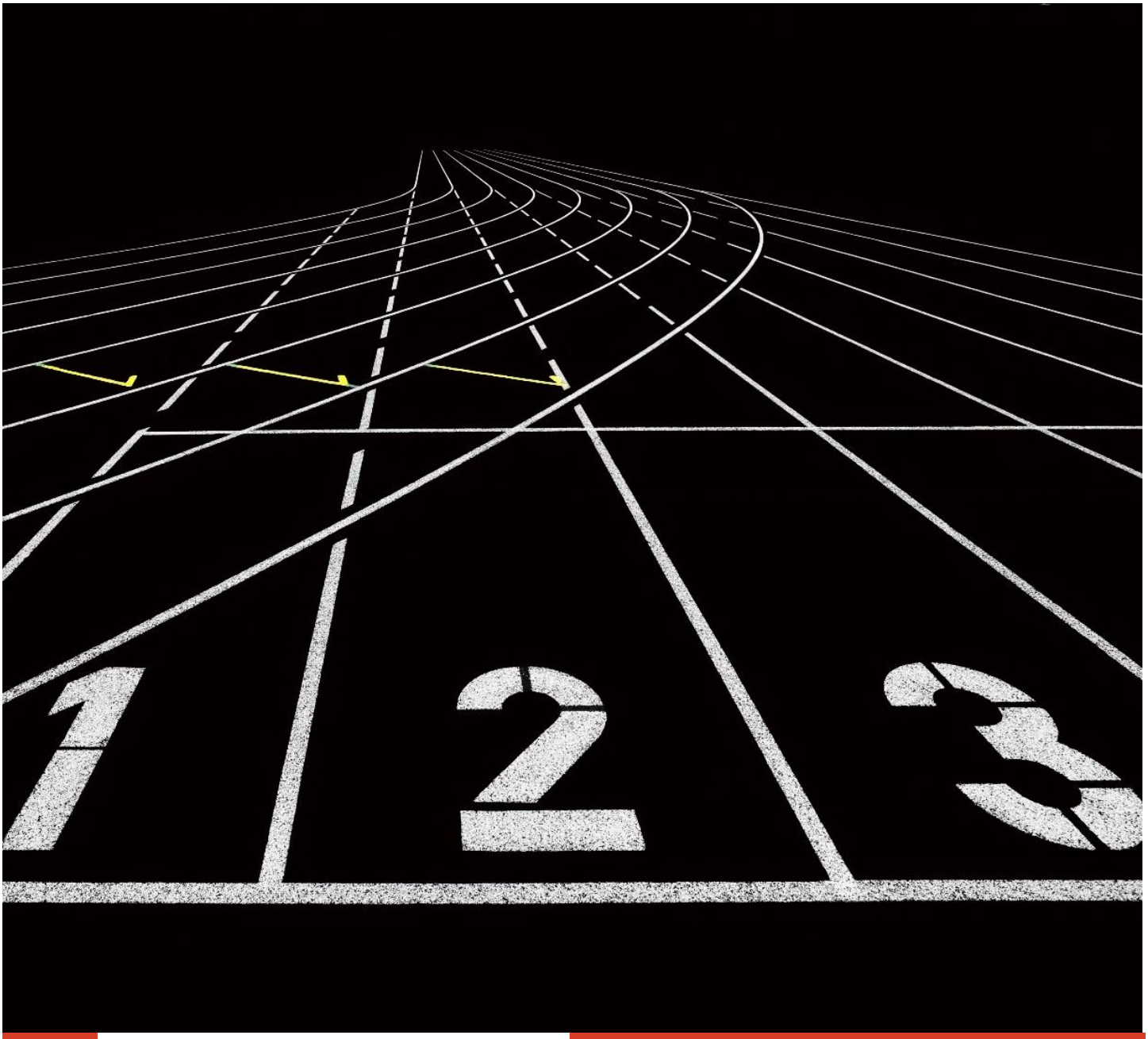


# WARANGLE REPORT

**UDACITY PROJECT**

Belalnasser1999@gmail.com




# OVERVIEW

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

# MY TASKS IN THIS PROJECT ARE AS FOLLOWS: -

### DATA WRANGLING CONSISTS OF:

- Gathering data
- Assessing data
- Cleaning data

### Gathering data

Gather each of the three pieces of data as described below in a Jupyter Notebook titled `wrangle_act.ipynb:`

1. The WeRateDogs Twitter archive. I am giving this file to you, so imagine it as a file on hand. Download this file manually by clicking the following link: `twitter_archive_enhanced.csv`
2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (`image_predictions.tsv`) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the followingURL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called `tweet_json.txt` file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

# Data Assessing

- ## Quality issues

### TWITTER_ARCHIVE_ENHANCED-DATA:

- **Drop retweets by filtering the NaN of retweeted_status_user_id**
- **Many NaN values in in_reply_to_status_id, in_reply_to_user_id,retweeted_status_timestamp, r etweeted_status_id and retweeted_status_user_id columns**
- **The Source column data is invalid and needs cleaning "HTML tags"**
- **Correct Invalid rating_numerator values and drop exetrem values.**
- **Correct numerators with decimals**

- **Correct Invalid rating_denominator values and drop exetrem values.**
    - **Manually (assessed by individual print text).**
    - **programarlly(Tweets with denominator not equal to 10 are usually multiple dogs).**
- **timestamp column format is string and needs to be splited onto 3 columns**
- **convert floar numerator_rating to int64**

## IMAGE_PREDICTION:

- **Removing retweets will make retweeted_status column useless.**
- **Drop jpg_url (duplicated)**
- **Create 1 column for image prediction and 1 column for confidence level**

## TWEET_JSON:

- **Drop retweets**

- ## Tidiness issues
- **Change tweet_id to type int64 in order to merge with the other 2 tables**
- **Merge doggo, floofer, pupper and puppo into a single column called dog stage.**
- **Merging three data sets into one**

# Data cleaning

- **PROGRAMATICALLY CLEANED QUALITY ISSUES**
    - **Drop columns that contain a lot of NAN values and will not be used in analysis and visualization. (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_timestamp, retweeted_status_id and retweeted_status_user_id columns).**
    - **Split timestamp column into three columns (Year, Month, Day).**
    - **Clean HTML tags from source columns.**
    - **Melt doggo, floofer, pupper and puppo into a single column called dog stage.**
    - **Removing retweets then drop retweeted_status column.**
    - **Convert numerator_rating and denominator_rating data type to float.**
    - **Convert tweet_id to int64**
    - **Removing retweets will make retweeted_status column useless.**
    - **Drop jpg_url (duplicated)**
    - **Create 1 column for image prediction and 1 column for confidence level**

- **MANUALLY CLEANING**

**IN order to extract the correct rating in the tweet I checked every tweet text with extreme values and started to correcting them manually, then drop tweets**

  - **Correct Invalid rating_numerator values and drop exetrem values.**
    - **Correct numerators with decimals.**
  - **Correct Invalid rating_denominator values and drop exetrem values.**
    - **Manually (assessed by individual print text).**
    - **(Tweets with denominator not equal to 10 are usually multiple dogs).**