# WeRateDogs 的推特档案的数据处理

## 1. 数据来源

数据集主要来源于 3 部分

1. 手头文件twitter_archive_enhanced.csv
2. 编程下载推特图像的预测数据，存储为image-predictions.tsv
3. 通过api获取的有关推特的附加数据，存储为tweet_json.txt

# 2. 数据评估

**质量问题**

***twitter-archive-enhanced 表***

1. retweeted_status_id、retweeted_status_user_id、retweeted_status_timestamp记录不为空的是转发的推特，不是我们要的数据；
2. text列含有“^RT @”的也表示转发数据，不是我们要的数据；
3. 缺失：有1976条数据没有狗的等级分类
4. 质量：timestamp：数据类型应为datetime，不是object
5. 质量：tweet_id 应为str类型，而不是int
6. 缺失：expanded_urls 有缺失
7. 质量：'rating_numerator' 有数值为几百，甚至上千的评分，可能有误识别
8. 质量：'rating_denominator' 分母固定为10，却只有2333个为10，其余的可能存在误识别的问题
9. 质量：'name' 有很多无意义的名字

***image-predictions 表***

1. 缺失（Maybe）：twitter-archive-enhanced 表中有2356条推特数据，image-predictions 表中只有2075条记录，有缺失
2. 质量：img_num 列应该只有1、2、3三种选择，出现了31个4，即表示最可能的是猜测4，猜测4并没有给出

***twett_json 表***

1. 质量：created_at 时间应为datetime，不是object
2. 缺失：twitter-archive-enhanced 表中有2356条推特数据，twett_json表中只有2351条记录，有缺失

**整洁度问题**

1. 狗的等级评定不应该是4列，而是一列“狗的等级”，数据类型为“分类”：doggo、floofer、pupper、puppo
2. 三张表的内容都是有关于一条推特的，应该合并

# 3. 数据处理

1. 筛选出不是转发的数据，我们只需要“WeRateDogs ”用户自己发的推特数据
2. 通过推特id合并三张数据表
3. 原本的狗狗等级为四个变量，转为一个变量dog_stage存储
4. 清除掉合并后重复的列
5. 将时间的数据类型改为datetime
6. 删除掉缺失严重的列（即，可能发的推文与狗狗无关）
7. 处理狗狗评分的分子、分母误读的问题（重新从text中读取，有些事人工读取）
8. 新增加了一列来展示评分：分子 / 分母
9. 删除了本次探索中不需要的列，重新对处理后的数据集进行排列列的顺序，

# 4 最终的数据集字段

In [2]:
```
import pandas as pd
df_twitter_clean = pd.read_csv('twitter_archive_master.csv')
df_twitter_clean.head()
```

Out[2]:

| | tweet_id | text | name | rating | rating_numerator | rating_denominator | dog_ |
|---|---|---|---|---|---|---|---|
| **0** | 892420643555336193 | This is Phineas. He's a mystical boy. Only eve... | Phineas | 13.0/10 | 13.0 | 10 | |
| **1** | 892177421306343426 | This is Tilly. She's just checking pup on you.... | Tilly | 13.0/10 | 13.0 | 10 | |
| **2** | 891815181378084864 | This is Archie. He is a rare Norwegian Pouncin... | Archie | 12.0/10 | 12.0 | 10 | |
| **3** | 891689557279858688 | This is Darla. She commenced a snooze mid meal... | Darla | 13.0/10 | 13.0 | 10 | |
| **4** | 891327558926688256 | This is Franklin. He would like you to stop ca... | Franklin | 12.0/10 | 12.0 | 10 | |

5 rows × 25 columns