1. 数据收集
   1. WeRateDogs推特档案数据直接导入
   2. 通过URL编程下载推特图像的预测数据，即根据神经网络，对出现在每个推特中狗的品种（或其他物体、动物等）进行预测的结果。
   3. 通过Tweepy库下载额外的推特数据，至少要包含转发数（retweet\_count）和喜欢数（favorite\_count）
2. 数据评估

通过目测评估和编程评估的方式对数据进行质量及整洁度的评估

* 1. 质量：
     1. twitter\_archive 表格
* tweet\_id是整数格式
* in\_reply\_to\_status\_id, in\_reply\_to\_user\_id, retweeted\_status\_id, retweeted\_status\_user\_id, retweeted\_status\_timestamp 列有太多缺失值
* 只需要含有图片的原始评级 (不包括转发)
* timestamp是字符串格式不是时间格式，并且都有+0000
* source列太多多余的信息，如http连接地址
* expanded\_urls列中有缺失值或者一行中内容重复
* rating\_numerator和rating\_denominator的最大值最小值异常
* name有很多None或者a/an
* doggo, floofer, pupper, puppo中有很多None
* 不需要2017年8月1日后的数据
  + 1. *image\_predictions* 表格
* tweet\_id和 img\_num是整数格式
* jpg\_url列有重复项
  1. 整洁度：
* 狗狗阶段分为了四列 (doggo, floofer, pupper, puppo)
* 三张表格可以整合

1. 数据清理

3.1 数值缺失问题

*#### 质量 quality*

*\*twitter\_archive table\**

*1. 'tweet\_id' is an int not a string*

*2. The columns 'in\_reply\_to\_status\_id', 'in\_reply\_to\_user\_id', 'retweeted\_status\_id', 'retweeted\_status\_user\_id', 'retweeted\_status\_timestamp' have too much missing data*

*3. The dataset doesn't need retweet contents*

*4. 'timestamp' is a string not a datatype*

*5. 'source' includes useless infomation such like http link*

*6. 'expanded\_urls' has null value or duplicated urls in a row*

*7. Min and max numbers of the 'rating\_numerator' and 'rating\_denominator' are not correct*

*8. Delete the data after 8/1/2017*

*\*image\_predictions table\**

*1. 'tweet\_id' and 'img\_num' is an int not a string*

*2. The duplicates in 'jpg\_url' column*

*#### 整洁度 tidiness*

*\* Dog stage is in four different columns (doggo, floofer, pupper, puppo)*

*\* Three tables can be merged*