4. Data plan

We will use the business photos uploaded by Yelp users to predict the business attributes.

The data that we will use are:

train\_photos: photos of the training set

test\_photos: photos of the test set

train\_photo\_to\_biz\_ids.csv: maps the train\_photo id to business id

test\_photo\_to\_biz\_ids: maps the test\_photo id to business id

train: the main training dataset which includes the business id’s and their corresponding business attributes

sample\_submission: sample submission and the test dataset. It should include the business id’s and the corresponding predicted business attributes.

The business attributes that we will use are:

0: good\_for\_lunch

1: good\_for \_dinner

2: takes\_reservations

3: outdoor\_seating

4: restaurant\_is\_expensive

5: has\_alcohol

6: has\_table\_service

7: ambience\_is\_classy

8: good\_for\_kids

Since Yelp is a community driven website, we will also include the duplicated images in the dataset. The reasons for that are users may accidentally upload the same photo to the same business more than once or some chain businesses upload the same photo to different branches. Therefore, we will obtain above data from Kaggle’s Yelp Restaurant Photo Classification Competition.