## Real Graph (bqe)

This project builds a machine learning model using a gradient boosting tree classifier to predict the likelihood of a Twitter user interacting with another user.

The algorithm works by first creating a labeled dataset of user interactions from a graph of Twitter users. This graph is represented in a BigQuery table where each row represents a directed edge between two users, along with various features such as the number of tweets, follows, favorites, and other metrics related to user behavior.

To create the labeled dataset, the algorithm first selects a set of candidate interactions by identifying all edges that were active during a certain time period. It then joins this candidate set with a set of labeled interactions that occurred one day after the candidate period. Positive interactions are labeled as "1" and negative interactions are labeled as "0". The resulting labeled dataset is then used to train a boosted tree classifier model.

The model is trained using the labeled dataset and various hyperparameters, including the maximum number of iterations and the subsample rate. The algorithm splits the labeled dataset into training and testing sets based on the source user's ID, using a custom data split method.

Once the model is trained, it can be used to generate a score estimating the probability of a user interacting with another user.

## Real Graph (scio)

This project aggregates the number of interactions between pairs of users on Twitter. On a daily basis, there are multiple dataflow jobs that perform this aggregation, which includes public engagements like favorites, retweets, follows, etc. as well as private engagements like profile views, tweet clicks, and whether or not a user has another user in their address book (given a user opt-in to share address book).

After the daily aggregation of interactions, there is a rollup job that aggregates yesterday's aggregation with today's interactions. The rollup job outputs several results, including the daily count of interactions per interaction types between a pair of users, the daily incoming interactions made on a user per interaction type, the rollup aggregation of interactions as a decayed sum between a pair of users, and the rollup aggregation of incoming interactions made on a user.

Finally, the rollup job outputs the ML predicted interaction score between the pair of users alongside the rollup aggregation of interactions as a decayed sum between them.