**Executive Summary**

*By: Thomas Bacas - Data Scientist*

**The Problem**

Reddit, the open forum of the internet. Where individuals post their beliefs and ideas. Many people including myself see the an internet forum like reddit the true place for people's opinions, where no one speaker is hindered by threat of personal attack.

A place like reddit is home to endless amounts of information, with the proper tools and ingenuity you can draw valuable information from the open forum.

**Tools and Code**

The primary method for retrieving the post information was querying the reddit API. By utilizing for loops and one can quickly return 25k+ post including the comments, title, post contents, users and more. These posts were collected into a pandas DataFrame for easy evaluation and exploration.

After doing this, each post was countervectorized a process called tokenizing is used to break sentences into individual words (or characters in some cases) so the computer can evaluate their importance to classification. Without this step, the computer would not be able to properly run the functions below.

Once our important words were selected from each post they were used as features for our Random Forest Regressor, an aggregated decision tree model which develops branches based on information gain from each stage. With this information the model is able to predict the subreddit origin with 93% accuracy (r2 Score).

The developed model has the capability to return which words are most important for the classification process. In the example of this project - 'Risk' was the determining word with almost 75% importance. That is not surprising coming from r/Finance where much of the regular discussion is about financial risks.

**What Next**

The next stages of this project would be to utilize similar techniques as above to determine trends within posts. Which posts return the most karma, have the most comments, which titles attract the most attention, etc. These processes are currently in development and will be using similar code as above.