**Joseph Deal**

Oklahoma

NLP Wine Descriptions

**October 18, 2020**

# Overview

A Wine Review database from Kaggle.com will be utilized for a Natural Language Processing project. The database contains wine reviews from Wine Enthusiast magazine and provides professional descriptions of wines from around the globe. The emphasis of using NLP on the descriptions is to see if there are enough unique descriptors to be utilized to evaluate other features found in the database. The ability to use the words as predictors for type of grape or wine, the quality or ranking of the wine and predicting the cost of the wine. The dataset of 150,000 descriptions would provide the basis for evaluating the descriptions from professional wine critics (sommeliers).. There are 632 different wine varieties. Ratings are generally between 80 and 100, while pricing varies from 4 USD to 3,300 USD.

# Goals

1. **Predict Type of Grapes or Wine:** Utilizing machine learning techniques including Natural Language Processing (NLP) to evaluate the type of grape used in each wine.
2. **Estimate the Ratings:** Utilizing learned machine learning techniques including NLP to predict the rating of a wine based on a professional wine-taster or Sommelier’s description of the product.
3. **Estimate the Pricing:** Utilizing NLP on the Sommeliers’ description of the wine evaluated to initiate various other machine learning techniques to provide an estimate of the pricing of the wine.

# Milestones

## Collect and Clean Data

Collecting the data is downloading the csv file from Kaggle.com. Cleaning will ascertain the extreme values looking for missing or mischaracterized data. Look for ways to remediate missing and/or mischaracterized data. Focusing on just a few variables, such as description, price, and ratings, the balance of the variables will seem marginally impactful. These marginal variables might include taster-name, province, designation, which is the winery name, country, region1, and region2.

One focus of cleaning data would be the spelling found in the description variable. Spelling might be an issue in the winery names, country, provinces and regions as well. If there are missing data for these variables might be solved with a little internet research.

## Explore the Data

Utilizing natural language processing of the description variable should provide a treasure trove of information relative to the quality and pricing of the various wines produces. The peripheral variables should show some varying degrees of correlation to those values as well. Exploring those relationships will be key in producing a predictive tool connecting the description with various analytics such as pricing and ratings.

## Destination

The end goal is to evaluate whether the professional wine critics’ comments can actually relate to pricing and ratings over the global wine industry. The correlation between these critiques has merit, then the individual wines could be adjusted accordingly or utilized in a way that could be promoted to increase sales volume or sales pricing while maintaining market pricing models in the industry. There may also be regional pricing components and premiums applicable in certain circumstances.