# Project Proposal: Wine Reviews

**INLS 613.001.S18 Text Mining**

Patric Carman Jordan Chasteen Andrew Garcia

19 February 2018

## Problem description

*A description of the problem area you wish to investigate*

Representing an enigmatic global economic structure, the world wine market is driven by a large number of feature variables, each with varying weight. A certain wine's price is determined by its feature variables which can range from obvious things like region, vintage (production year), grape quality to not so obvious things like soil nutrients, vine height and even varietal of barrel used for fermenting. Due to the high number and specific variety of interesting features, the problem of predicting a wine's price given its production variables is a problem that is almost a perfect fit for machine learning. Using machine learning we hope to gain insight on how the price of any given wine is determined, trained and tested on a database of 130k wines, listed with all sorts of variables such as origin, description, designation, points, provenance, regions, title, variety and winery.

## Literature review

*A list of 6-8 papers you plan to survey as part of your literature review. For each paper, provide a brief justification for why you selected it (3-5 sentences).*

1. *Predicting the Quality and Prices of Bordeaux Wine*
   Author: Orley Ashenfelter.
   JSTOR permalink
   This paper provides a more economics driven analysis of wine prices, and pertains to predicting prices of a specific type of wine – namely the Bordeaux varietal. This paper provides some much needed information on the economics of wine, and give us a bit of background in the area of study we chose – an area in which we are not particularly experts.

2. *Statistical analysis of the price and subjective quality ratings on Australian wines*
   Author: Peter M. Visscher
   arXiv permalink
   This paper from a researcher from Queensland focuses on analyzing price and subjective quality ratings on Australian wines. This is a similar research area to that of our project, and will be useful moving forward through our project. It provides detailed statistics on relationships between price and rating, variations in price and variations in ratings. If we determine at any point in our research that we need to adjust for these variables, we will have adjustment coefficients readily available for a sample subset.
3. TODO Patric
4. TODO Patric
5. TODO Andrew
6. TODO Andrew

## Experiment description

*A description of the purpose of your experiments. What are you testing?*

According to The Wine Institute, 62% of Americans aged 21+ regularly make wine purchases. Additionally, the Wine Genome study performed by Constellation Brands found that one-in-five wine purchasers makes their wine decision purely based on price, and on no other factors. These facts together imply that close to 30 million (27.5 million) Americans are blindly picking up a wine off the shelf based on price, with no regard for quality or popularity. What's even more surprising is that the entities who set the price of wine are those who produce and profit from it. This creates a positive feedback loop, where by setting higher prices, yet keeping quality consistent, you will maintain a large swath of customers, only encouraging higher prices on lower quality wines. It makes the wine pricing process seem to have the same randomness as a blind dart throw. We selected this problem in order to give method to this current madness in wine pricing.

### TODO What are you testing

## Dataset

*A description of the dataset you will use to conduct your experiments. If you are using an existing dataset (highly recommended), provide a short description and a hyperlink so that I*

*can review it. If you plan to collect our own dataset, provide a short description of how you will collect it (e.g., By scraping text from the web? Will you annotate it manually?).*

TODO Andrew

## Associated risks

*A description of any risks associated with your proposed plan.*

As with any machine learning experiment, the hypothesis that there is a significant correlation between features and class could be wrong. In this instance the risk is that wine price is not dependent on the feature variables we possess in our dataset. However, even if we encounter this problem, we will still have garnered a valuable piece of information about wine pricing models – that it is more arbitrary than we initially thought.