

# The Tasting Game

A prototype multi-class predictor model for blind wine tasting simulation

**James Opacich**

Data Scientist

# 1. Problem Statement



FROM A TEMPERATE CLIMATE.





# Blind Tasting Simulator

## What Is Being Created?

- Prototype
- Multi-class Predictor
- Blind Wine Tasting Simulation

Grabby

# Blind Tasting Simulator

## Why is it Relevant?

- Mental Sharpness
- Physiological Health
- Fun
- Profit



# Blind Tasting Simulator

## Is There A Need?

- Demand Exists
- Board Games
- Tasting Trainers Abound
- There's a Void In The Market



The  
Who

# Blind Tasting Simulator



Consumers



Addressable Markets

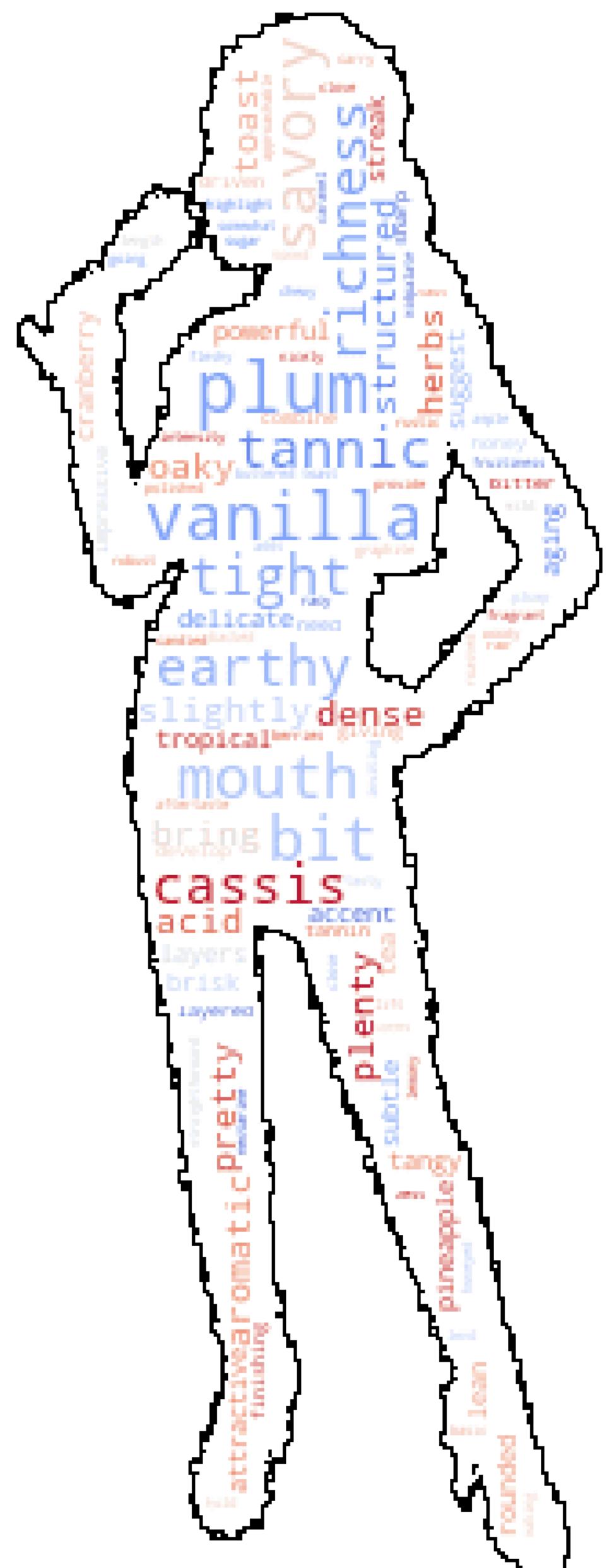
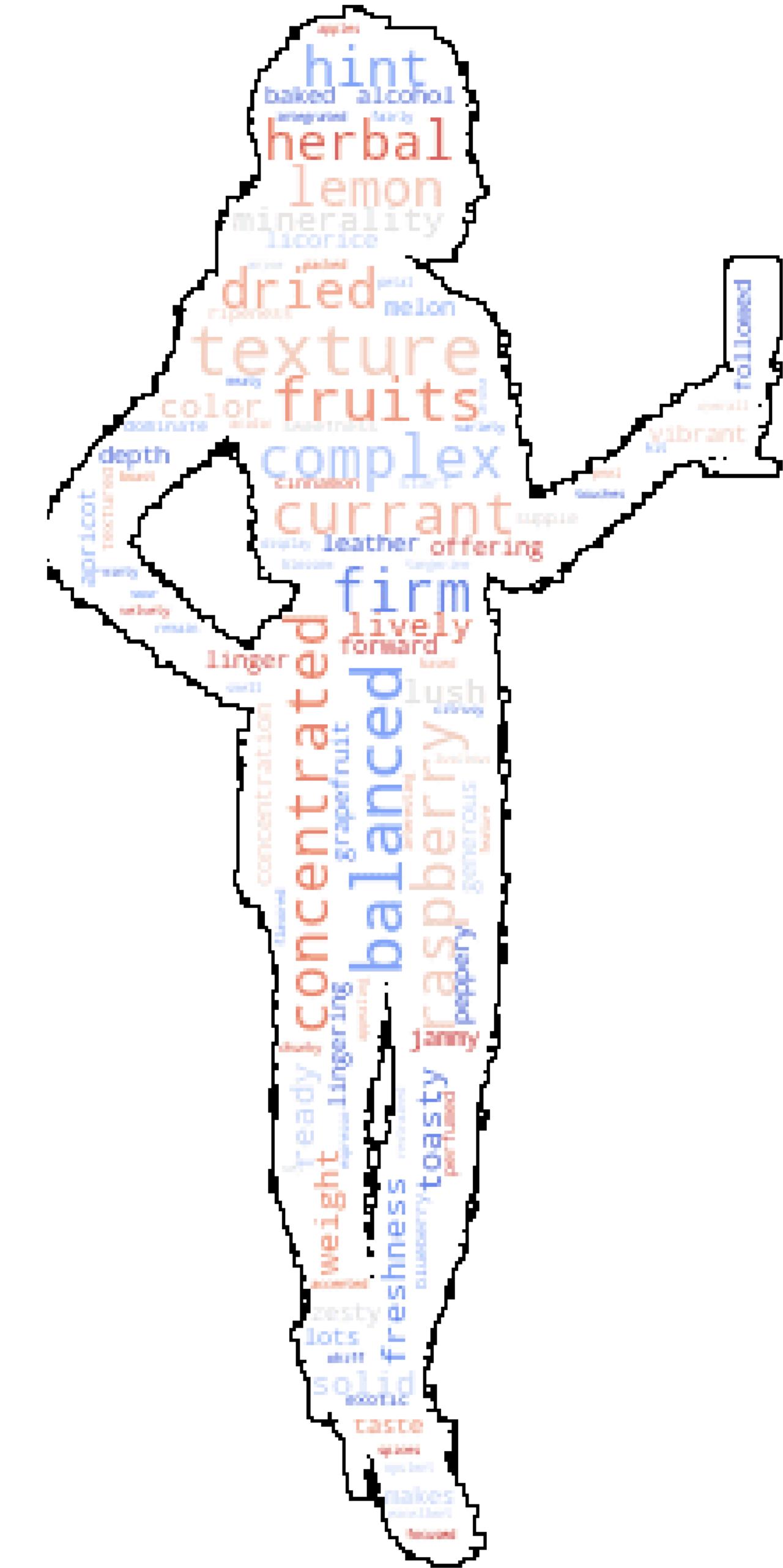


Distributors and Retailers

# Blind Tasting Simulator

## Measuring Success

- F1 Score
- Accuracy Score





# The Data

# Wine Enthusiast

# 130,000 Sommelier Tasting Notes (June - Dec 2017)

# Unique Counts

# 708 Unique Grapes

# 16,755 Wineries

# 1230 Regions

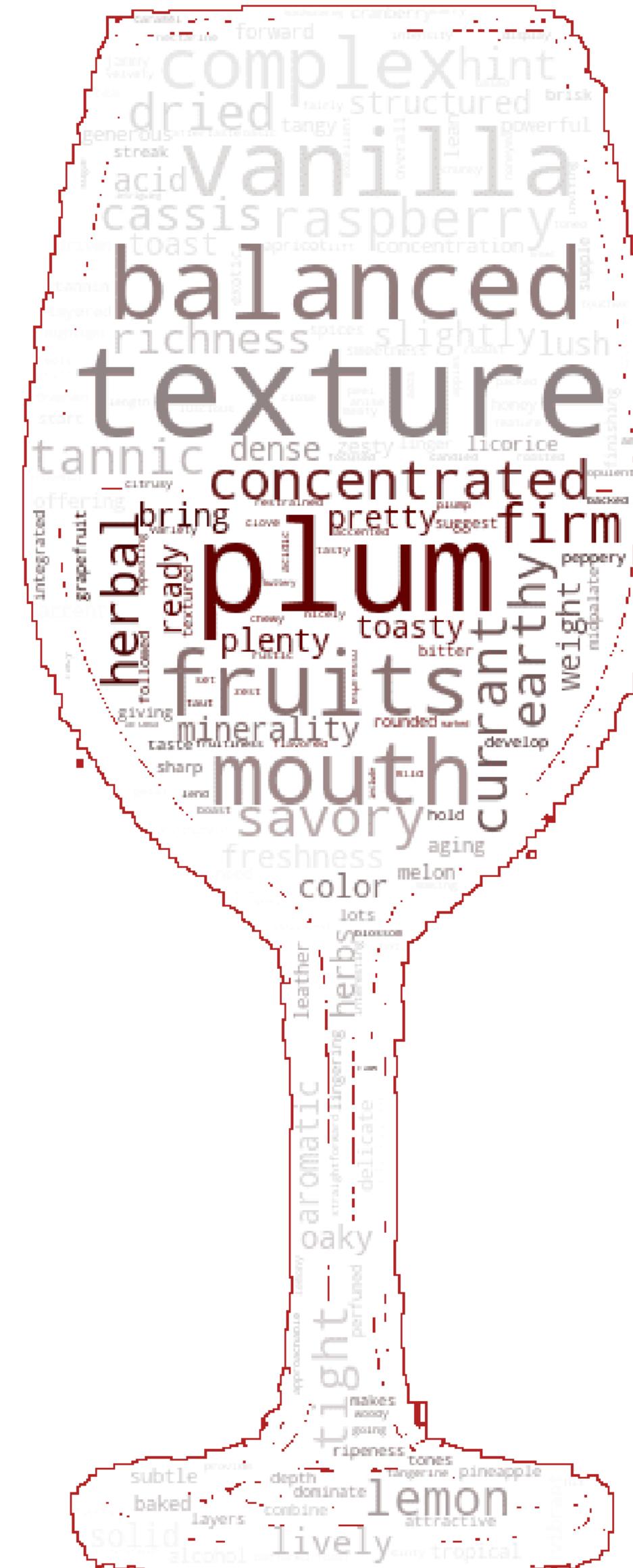
# 46 Countries

# Targets

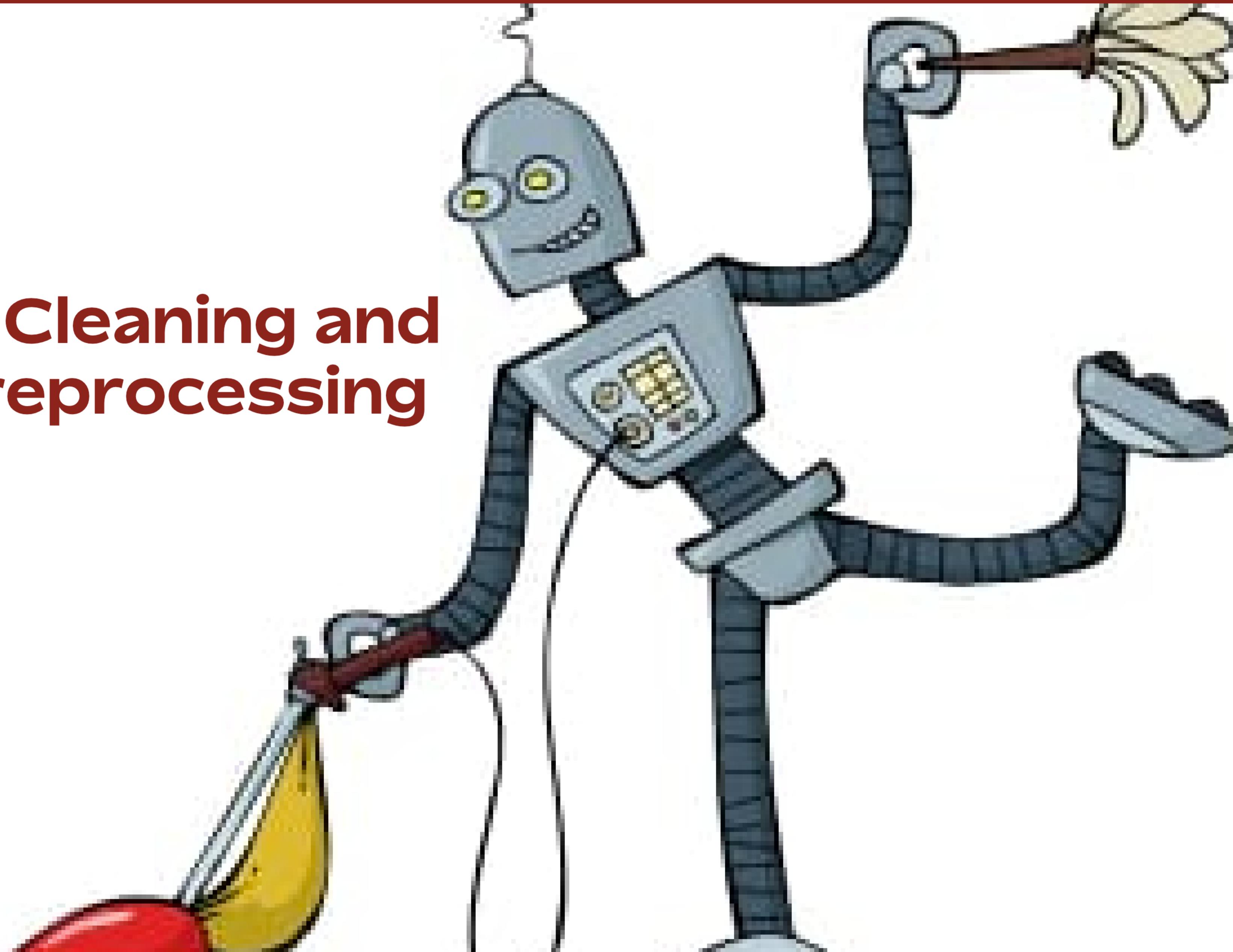
# 20 Classes

# 10 Red Varietals

# 10 White Varietals



## 2. Cleaning and Preprocessing



# Duplicates

Lost

11,614

Observations

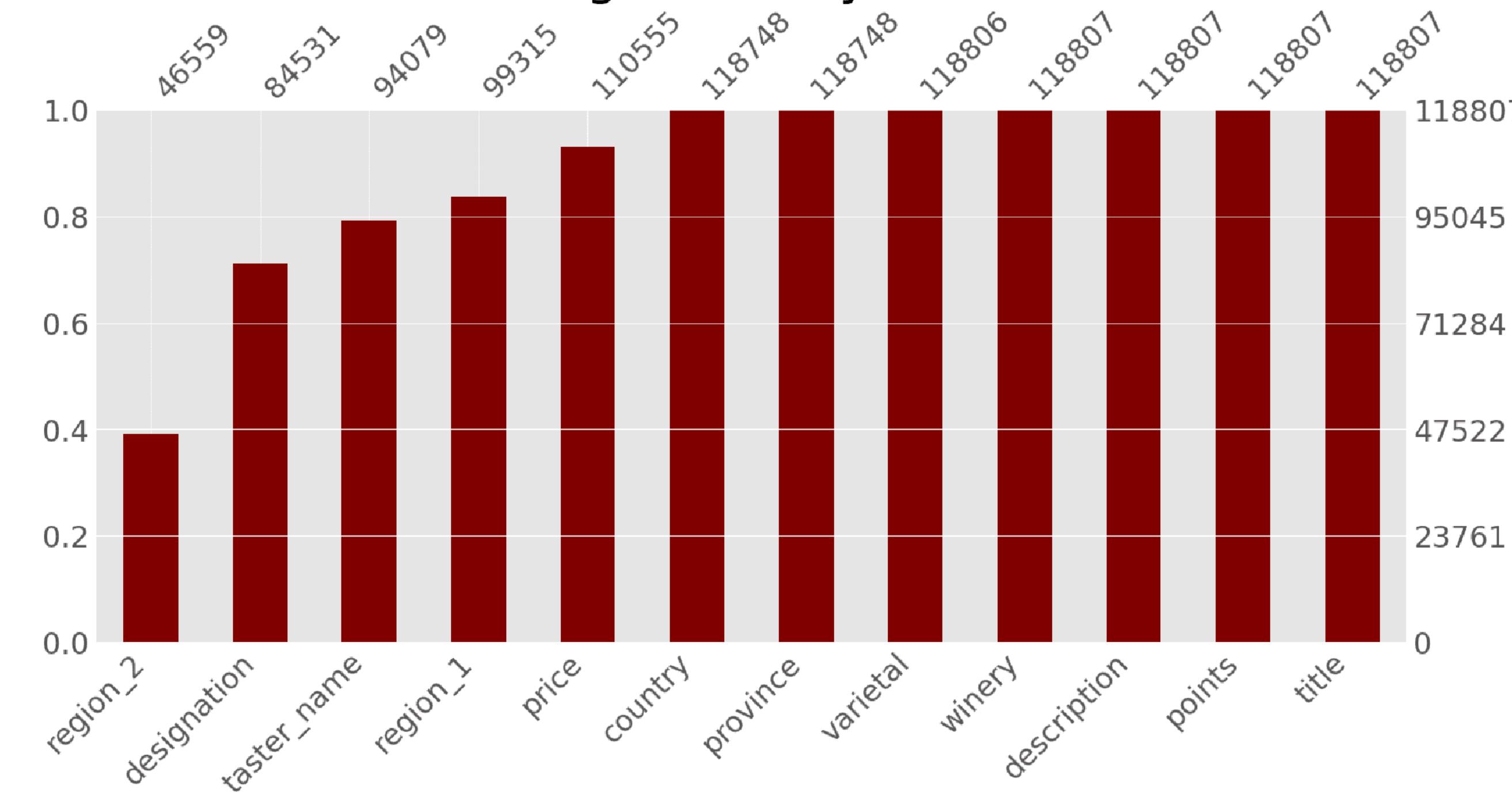
3%

of Total



# Missing Values

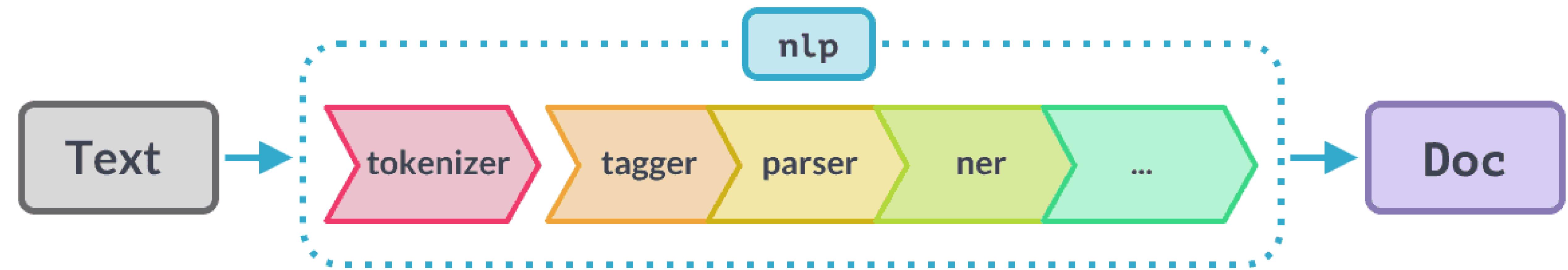
Missing Values by Feature



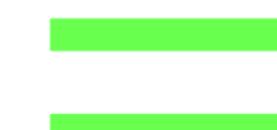
NLP

# Preprocessing

# spaCy



Word Embedding



Increased Model Performance!!!!

# Preprocessing

## Stop Word Stats

---

Length of Stop Word List: **22,403**

Total Stop Words Found: **> 2 Million**

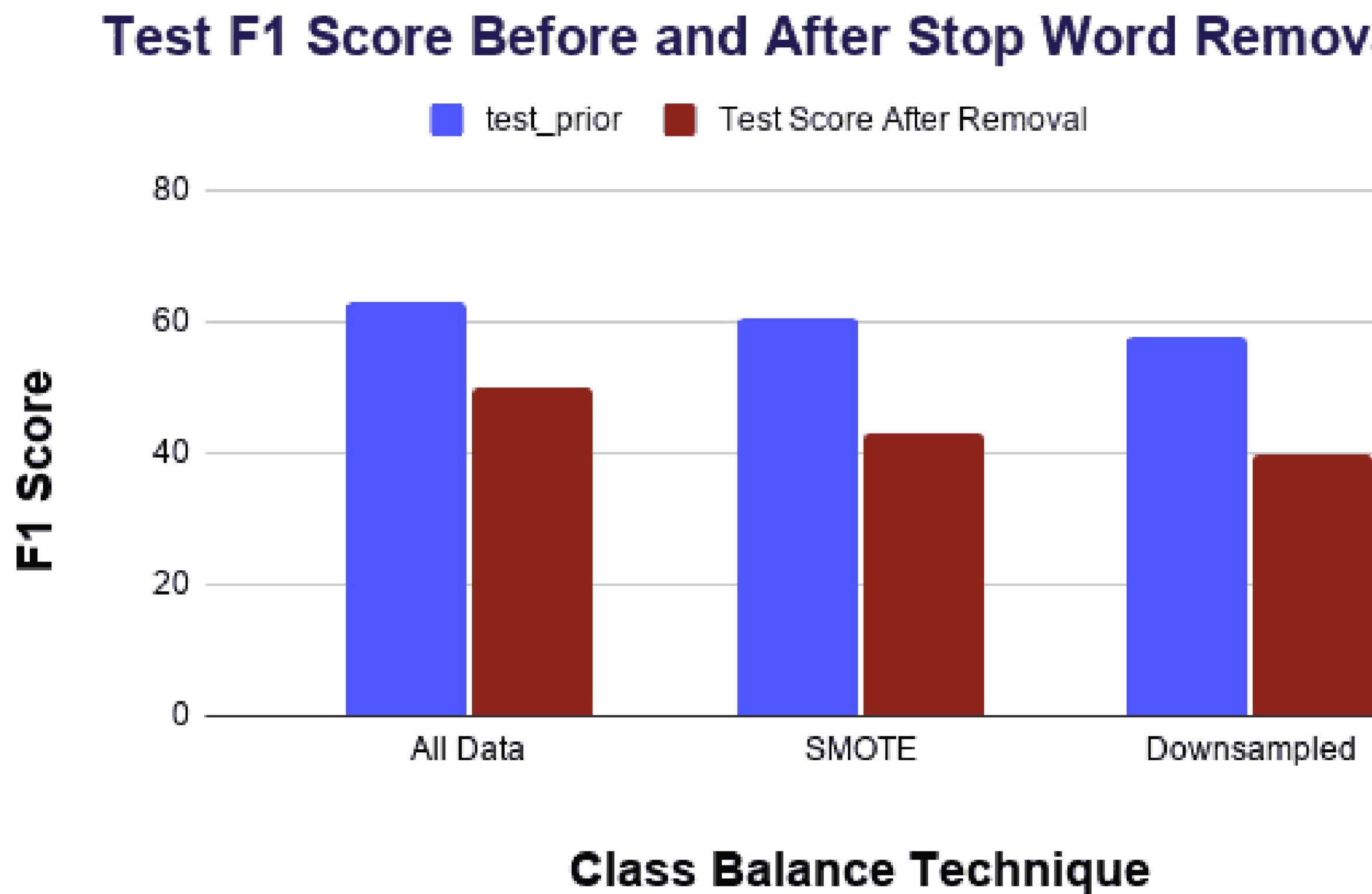
Percent of Total Words: **Over 80%**

# Stop Words

## Curated From

- Sklearn CVEC
- NLTK
- Spacy
- Domain Knowledge
- Scattertext Analysis

# Preprocessing



# Stop Words

## Significance

- More Truly Simulate Real World Environment
- Evaluate a More True Level of Accuracy in the Model



Word

# Vectorization

## TFIDF Vectorizer

- Performance: 37.24%  
Accuracy

## Count Vectorizer

- Performance: 37.39%  
Accuracy

On Preliminary Models



## Target Class Encoding

### Label Encoder

Labeled Target Classes in an ordinal-type fashion

Cabernet Sauvignon ..... → 1

Chardonnay ..... → 2

Pinot Noir ..... → 3

ETC.

# 3. Exploratory Data Analysis

Class

# Imbalance

Undersampling

Simple Random

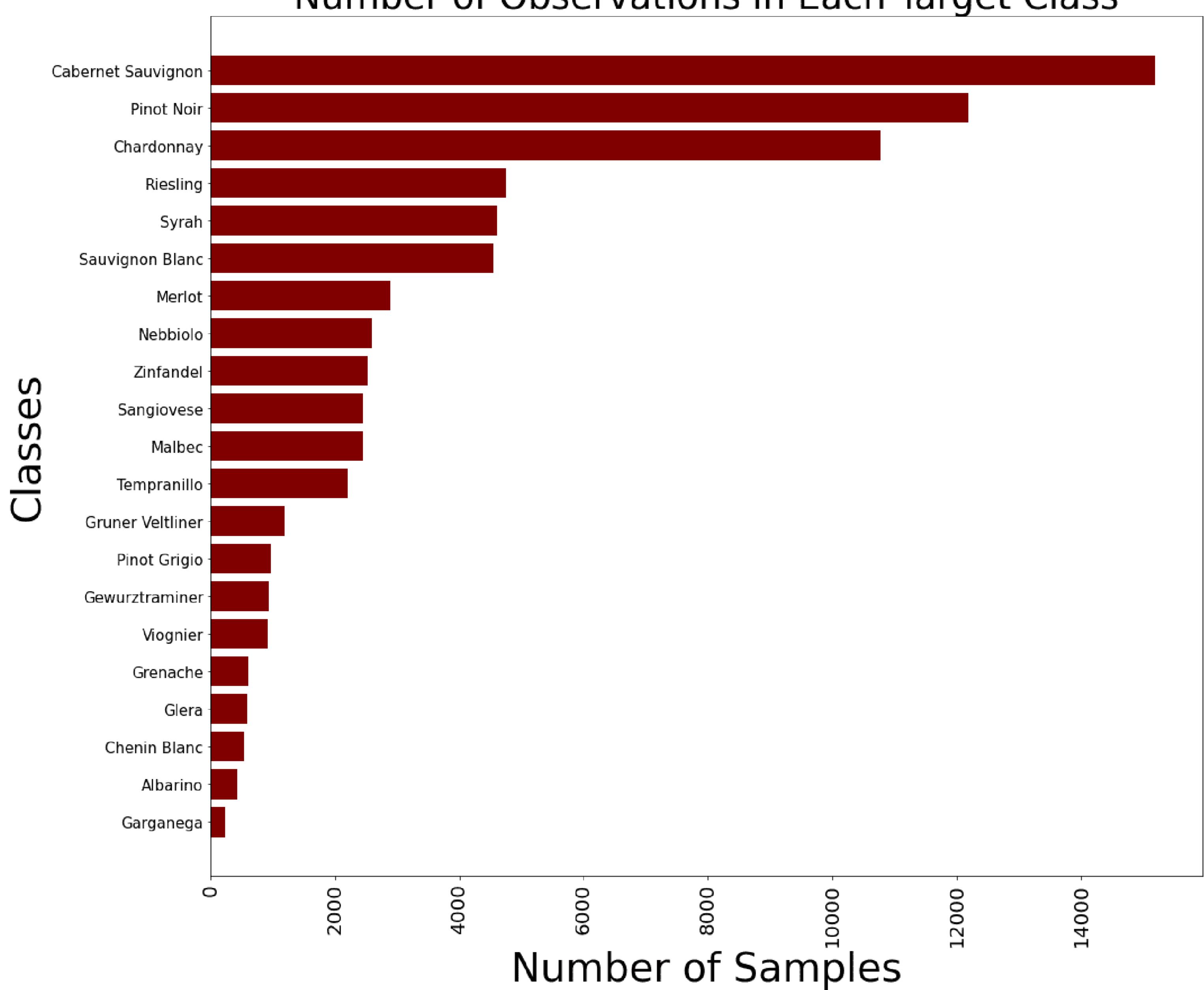
~400 Samples Each

Bootstrapping

SMOTE

Over 10,000 Samples Each.

Number of Observations in Each Target Class

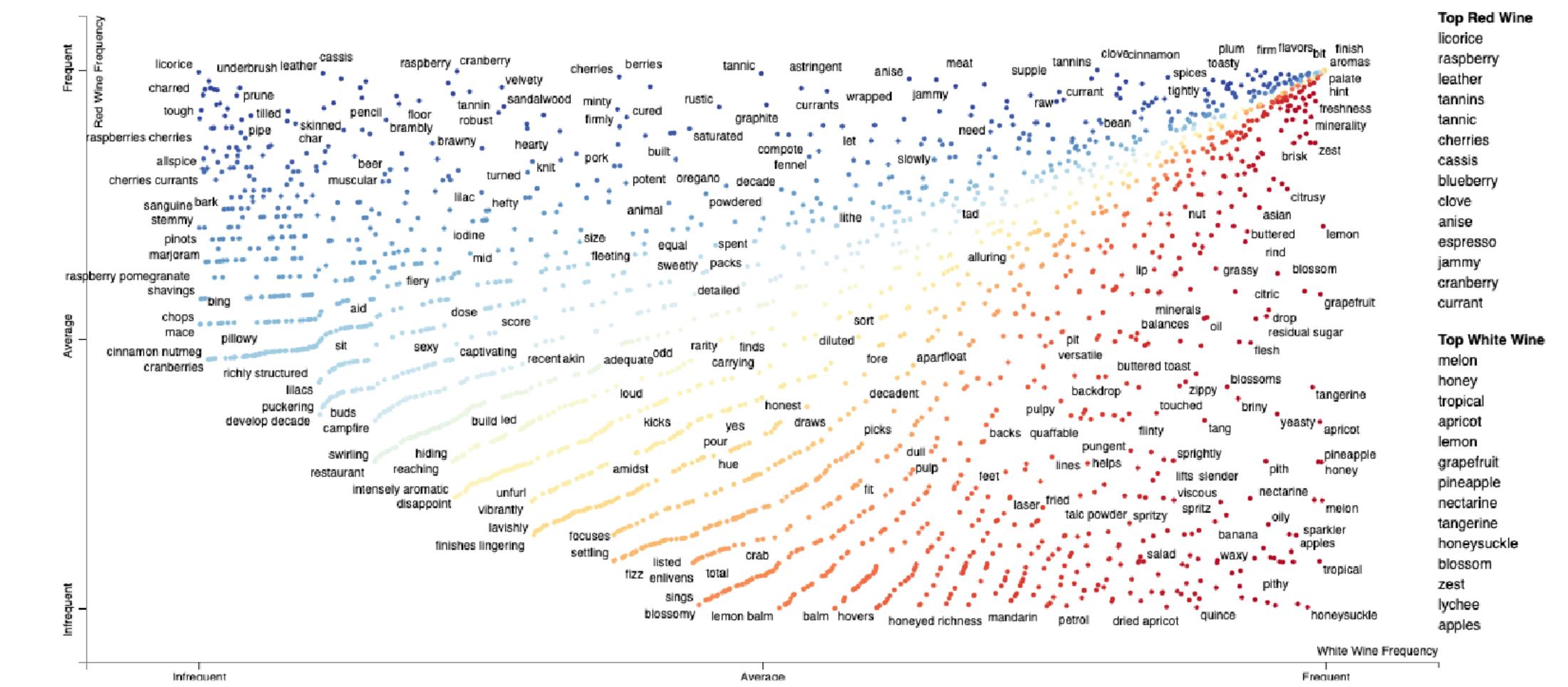


EDA

# Feature Analysis

- Term Frequencies
- Scaled F-Score
- Feature Recall
- Feature Precision

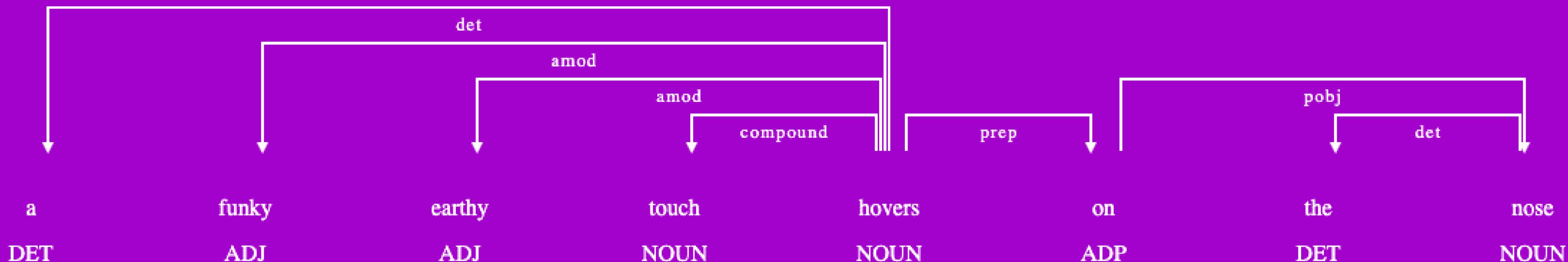
spaCy  
Scattertext



a visualization tool for comparing  
the importance scores of  
bag-of-words features to  
univariate metrics.

# Feature Analysis

## Dependency Tagger



A dependency visualizer that lets you see the way spaCy sets the parse and dependencies.



EDA

# Feature Analysis

spaCy  
DisplaCy

## (NER) Named Entity Recognizer and Tagger

this is surprisingly plump and accessible featuring scents of lime rind upfront then flavors of apple nectarine and citrus a medium bodied wine with a long

mouthwatering finish it's a tasty near term drinker but should develop well for up to years DATE

subdued hints of lily and jasmine rise from the glass the palate offers restrained notes of peach and mirabelle PERSON plum with a rounded off dry finish

focusing on blossom honey it's a gentle rounded and perfumed wine

as with the winery's riesling no vintage is indicated anywhere on the bottle listed at two percent PERCENT residual sugar this wine retains enough acid to

provide moderate lift to the spicy almost bitter pineapple and grapefruit flavors citrus rind phenolics lightly dappled with honeysuckle and other floral

highlights create a rather expansive wine with a substantial finish

once again this malbec merlot cabernet sauvignon blend from norton ORG excels the nose is a bit gray and candied but there's depth and a level of

intensity that bodes well for the future the palate is deep and concentrated with blackberry prune PERSON stewed plum graphite and coffee flavors in

Apple ORG is looking at buying U.K. GPE startup for \$1 billion MONEY

front of a mellow smooth finish drink now through

this wine tastes fresh and brisk like apricots and lemons yet a silky smooth texture adds depth and interest to make it more than just a one CARDINAL

note fruity style of wine it comes from an over achieving winery located one CARDINAL county inland from napa

---

A dependency visualizer that lets you see the way spaCy sets the parse and dependencies.

# 4. Modeling

I predict  
that it  
tastes like....  
Wine!



Model

# Overview

# Types of Models explored

## Multi-class Classification (21 Classes)

- Multinomial Naive Bayes
- Random Forest
- Logistic Regression
- KNN
- Voting Classifier
- CNN
- One Verse Rest



Model

# Evaluation

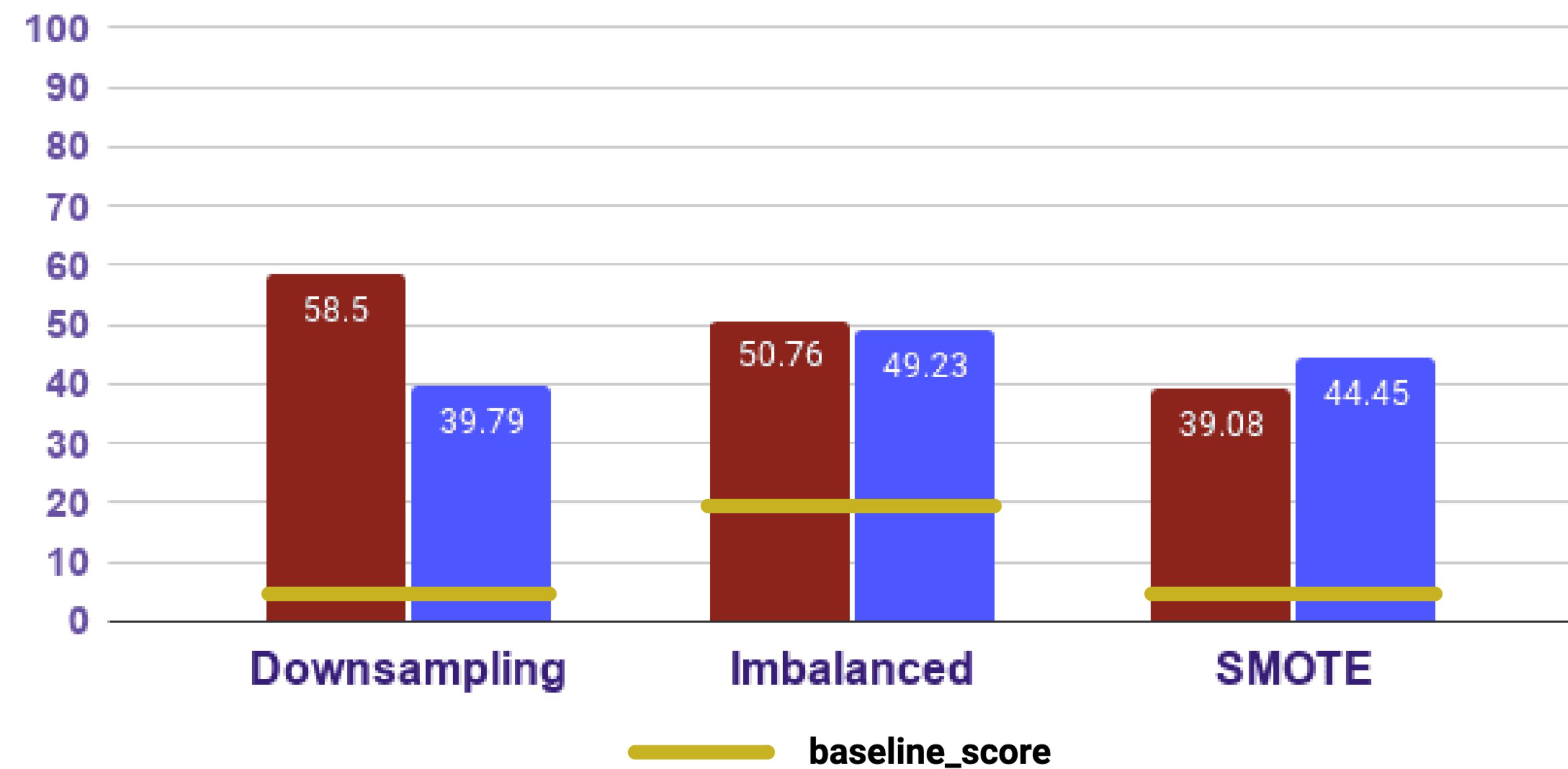
# Varietal Predictor

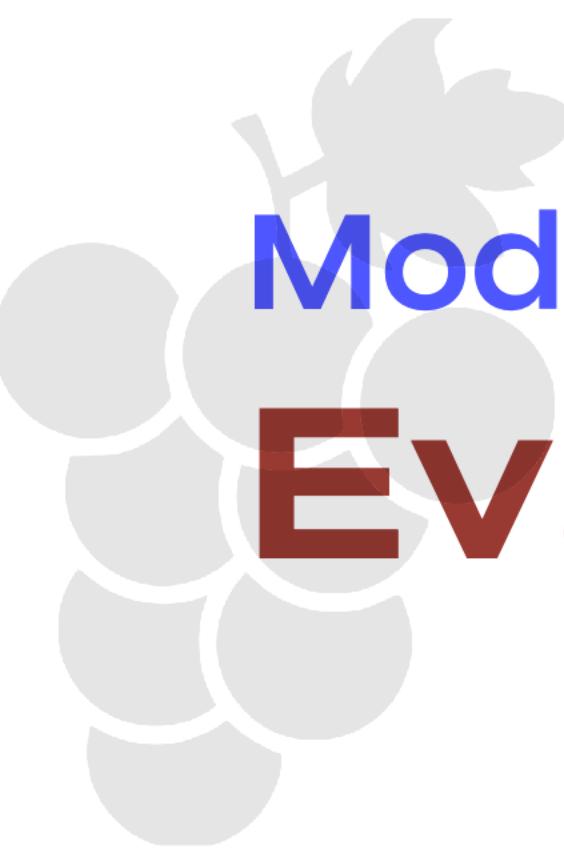
## Multinomial Bayes

### F1 Score For 20 Class Multinomial NB

Delineated by Class Imbalance Technique

■ train\_score ■ test\_score





Model

# Evaluation

# Varietal Predictor

## Convolutional Neural Net

↳ Model: "sequential\_5"

Layer (type)	Output Shape	Param #
=====		
embedding_5 (Embedding)	(None, 739, 32)	64000
conv1d_5 (Conv1D)	(None, 739, 64)	6208
max_pooling1d_5 (MaxPooling1D)	(None, 369, 64)	0
flatten_5 (Flatten)	(None, 23616)	0
dense_10 (Dense)	(None, 128)	3022976
dropout_5 (Dropout)	(None, 128)	0
dense_11 (Dense)	(None, 21)	2709
=====		
Total params: 3,095,893		
Trainable params: 3,095,893		
Non-trainable params: 0		

## Neural Net Structure

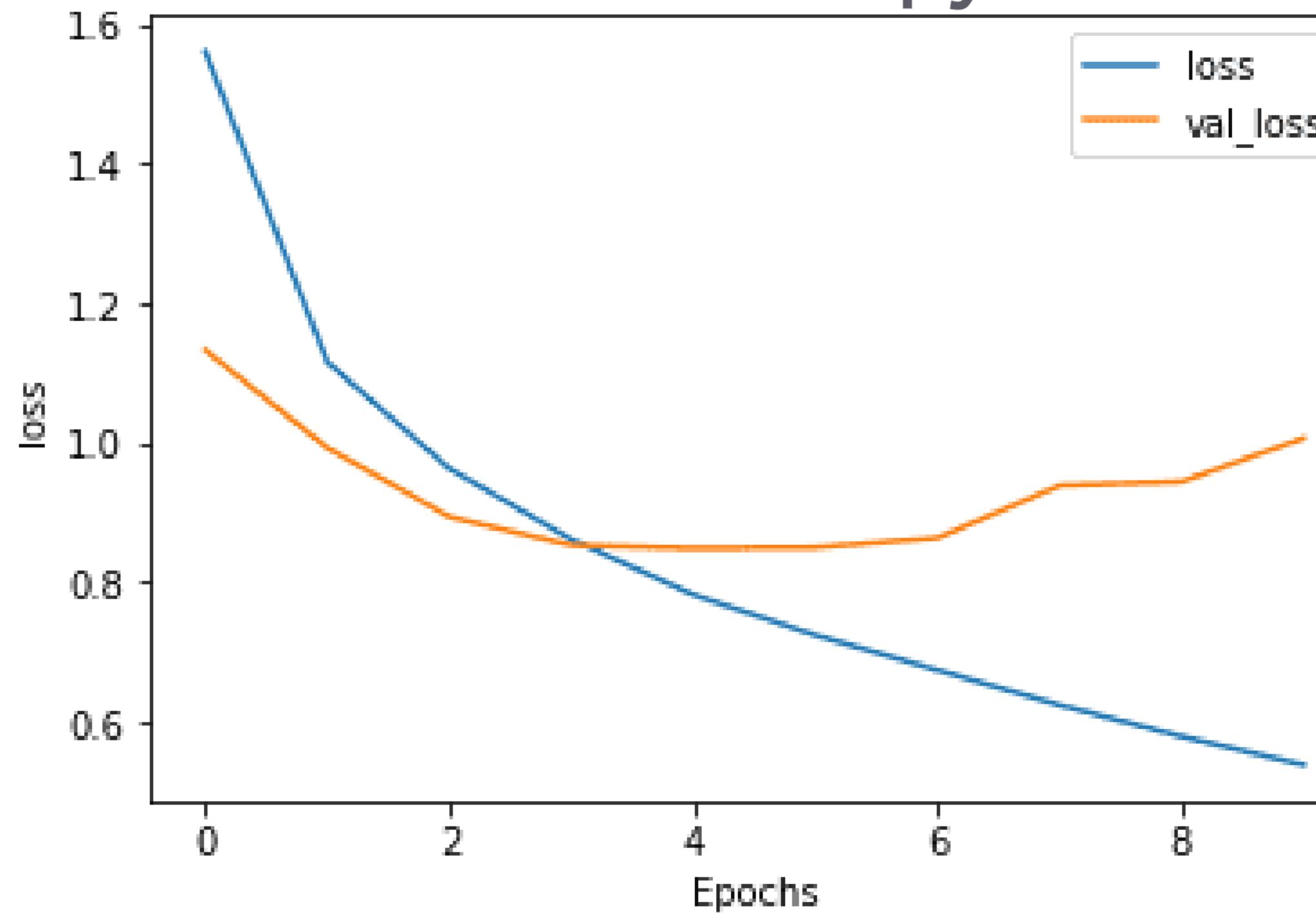
Model

# Evaluation

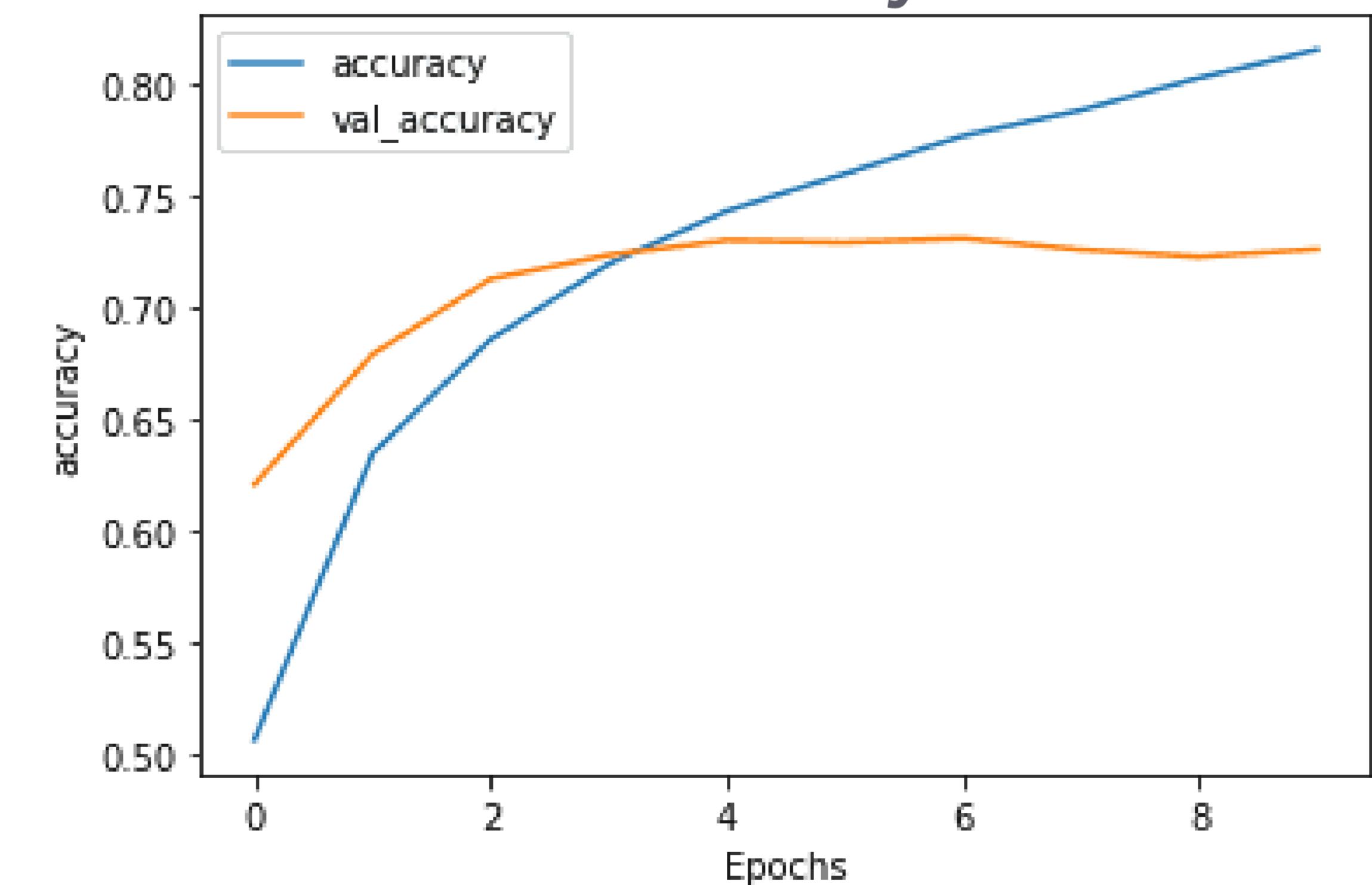
# Varietal Predictor

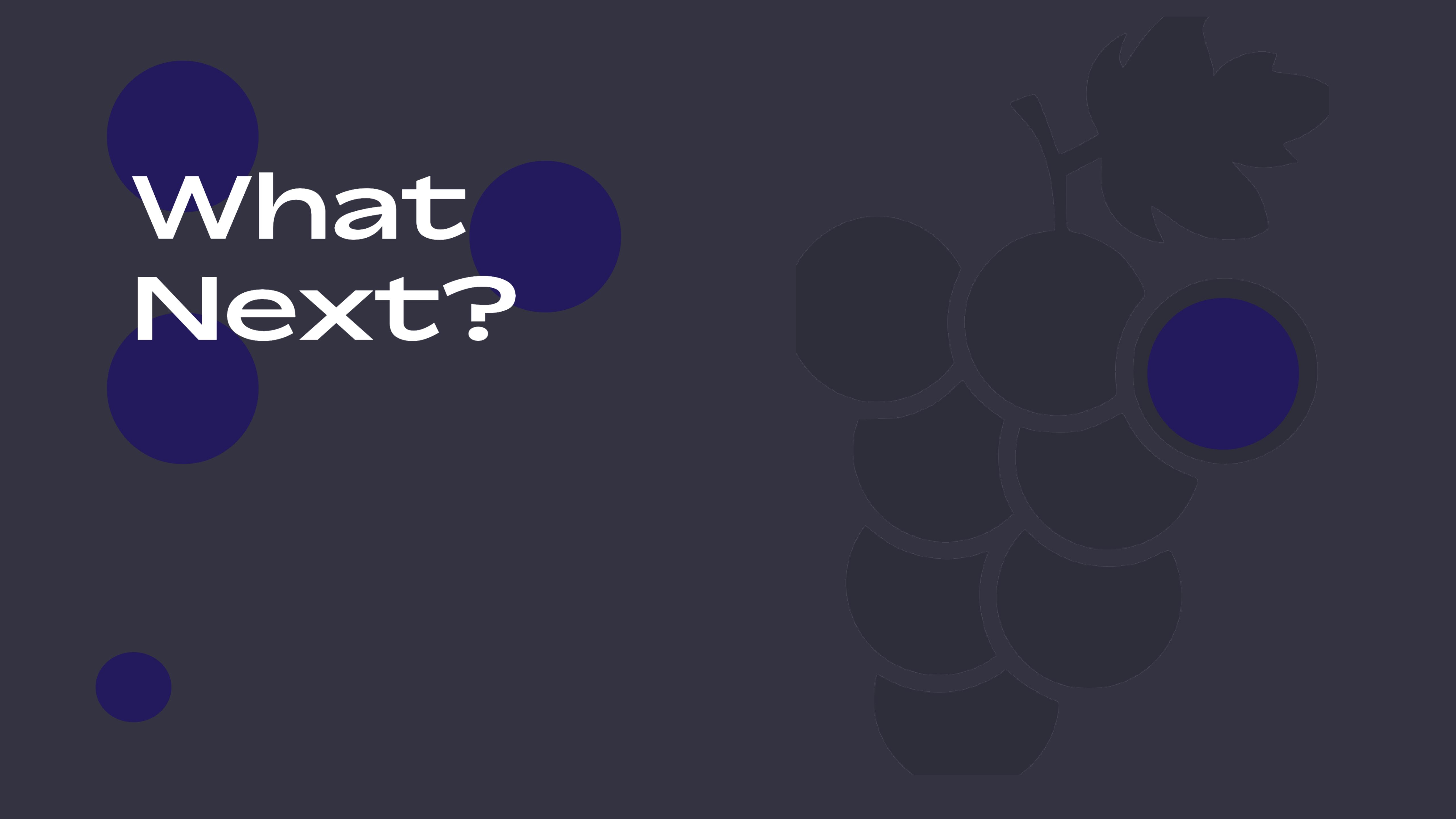
## Convolutional Neural Net

Sparse Categorical  
Crossentropy



Accuracy





what  
Next?

**FINAL**

# Overview

# Wish List

- Re-explore Methods for Feature Analysis
- RNN Classifier
- Binarized Labels
- Deployment



**FINAL**

# Overview

# Suggestions & Takeaways

- Preliminary scores are promising
- A better classifier can be built.
- More Money Needed
- Compute Power
- Hire a team
- Market Research

**Lets Move on to the Next Step!**





meredith

I love all of you!  
Thanks for an Experience!