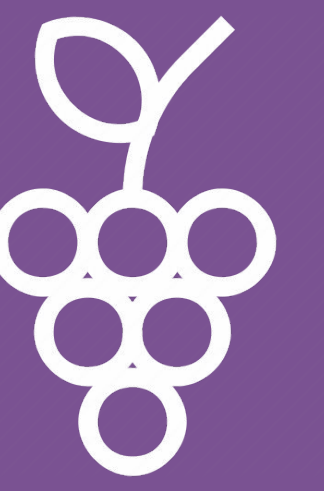# Your Cork | Personally tailored. Wine recommendations.

BIG DATA KNIGHTS

Chas Martin | Colin Murphy | Mike Nardella | Manjina Shrestha | Vivek Singh | Kishan Thapa

Introduction

Wine is a matter of personal preference and finding a wine that agrees with one's tastes can be both challenging and expensive.

Solution
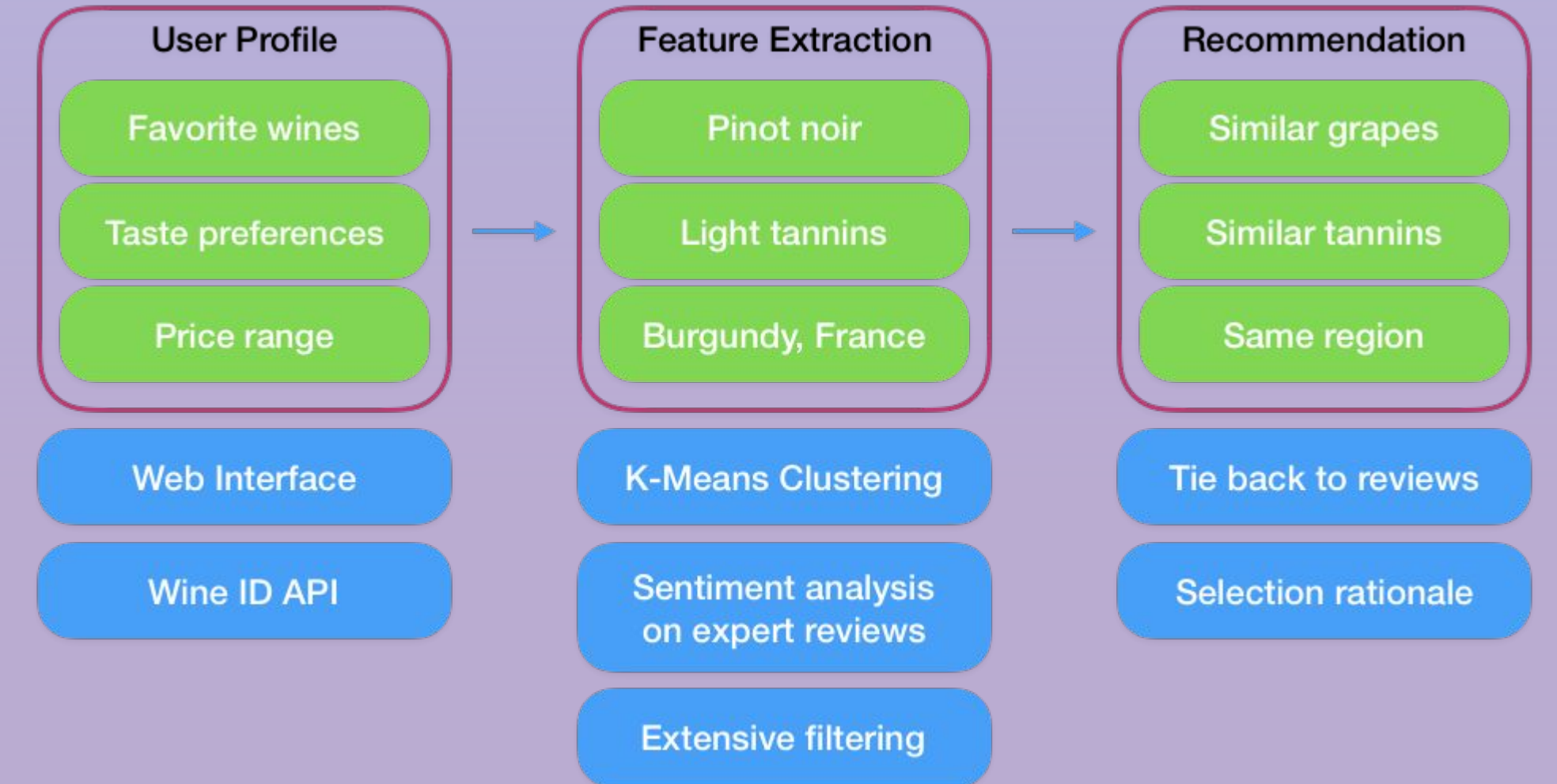- A recommendation system based on user's preferences

Benefits
- Reduces trial and error costs
- Simplifies wine choices
- Displays top 10 recommendations based on user's entered wine attributes
- Educates the user about wine regions
- Provides wine scores from sommeliers
- Enables users to explore neighboring wineries and wines with similar taste

Approach

Practical Application of Machine Learning
- Even a professional sommelier would be unable to keep up with a catalog of >100k wines
- Leverages the power of machine learning to extract key features from a comprehensive dataset
- Application responses are tailored to specific user preferences, including known wineries, price ranges, specific varieties, and free-form input
- Enables users to freely explore the world of wine by mapping regions to varieties and wineries
- Requires no prior knowledge of wine to interact

User Profile: Favorite wines, Taste preferences, Price range → Web Interface, Wine ID API

Feature Extraction: Pinot noir, Light tannins, Burgundy, France → K-Means Clustering, Sentiment analysis on expert reviews, Extensive filtering

Recommendation: Similar grapes, Similar tannins, Same region → Tie back to reviews, Selection rationale

Toolkit

| Regional Clustering | Global Exploration | General Filtering | Freeform Analysis |

Data

A globally scaled wine dataset was prepared with 15 features including expert reviews, their ratings, and winery details

- Data scraped from Wine Enthusiast
- Programming language used: Python
- Longitude and latitude queried on each winery with Google Maps API
- Size: ~80MB with 185k rows of wine and varieties
- Cleaned with OpenRefine (8GB configuration), Python, and Excel
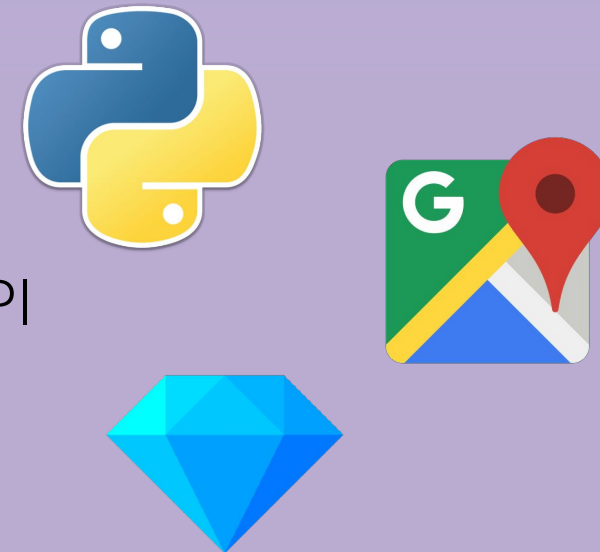
Methods

K-Means Clustering
- Wineries are classified geographically (on longitude and latitude) with the K-Means heuristic
- Research suggests grapes from regions with similar geographies produce wines with similar auromas
- Enables users to input a winery and explore others from the same area, labeled with type
- Simplifies finding / exploring wine with similar taste
- Also shows average price per winery and review score

Sentiment Analysis
- Classifies wine variety based on review column
- Allows a user to enter freeform taste preferences including food, color, or even another wine
- Leverages a variety of models including K Nearest Neighbors, Support Vector Machine, multivariable linear classifiers, decision trees, and random forest
- Hyperparameters are programmatically tuned and optimized for over 100,000 wines worldwide

Regional Highlights
- Hover over any country to find most popular wineries and varieties
- Perfect place to start if you're a wine novice
- Preparing dinner? Fixing something Italian? South American? Enjoy wine like the locals!
- Requires no prior knowledge or taste palate
- Learn here, explore further with clustering and sentiment analysis from the same user interface

Experiments and Results

Sentiment Analysis
- Experiment on corpus of wine review for sentiment analysis was done with the help of Support Vector Machine and Grid Search algorithm
- 185K reviews from wine experts was loaded using data frame and divided into 70-30 ratio of test and train set
- A sklearn pipeline model with sklearn countvectorizer having nltk stop words, & TFID transformer & SVM
- Grid search param was initialized with kernel [linear, rbf] c as [0.1,1,100] and n_gram with value of (1,2)
- Model was trained and experimented on 185k of reviews with accuracy of 0.62
- K nearest neighbor algorithm was run on predicted category in order to find top ten wines related to user taste along with Average price
- Histogram of predicted wine with average price and its reviewers score.

K-Means Clustering
- K value was tuned to produce cluster sizes balancing exploration and likeness
- The interactive cluster window is sized with padding to also show neighboring clusters
- Cartesian distance was selected as the quickest and most accurate method for determining which cluster a user-inputted winery falls

Conclusion and Future works

We developed a wine recommendation engine useful for both novices and wine experts. It will provide a personalized wine-related content with visualizations and wine suggestions using user inputs on regions of winery i.e., based on k-d tree and k-means clustering algorithm, and on flavor and characteristics i.e., based on sentiment analysis and knn algorithm. Future works would be fine-tuning the algorithms to increase the accuracy and hosting it on AWS.