**Project Proposal**

**Group Member:**

He huang

Ye Liu

Zhu wang

Congying Xia

Lichao Sun

Fan Zhu

**Dataset:**

# Our project is about a competition named “What’s Cooking?” on Kaggle, [www.kaggle.com](http://www.kaggle.com). The dataset is provided by Yummy, [www.yummly.com](http://www.yummly.com).

# Food is an important part of most cultures. Some of our strongest geographic and cultural associations are tied to a region's local foods. In the dataset, we have a train file and a test file. The train file include the recipe id, the type of cuisine, and the list of ingredients of each recipe (of variable length). The data is stored in JSON format. In the test file, the format of a recipe is the same as the train file, only the cuisine type is removed, as it is the target variable we are going to predict.

# In the train file, there are 39774 recipes and 20 kinds of cuisines in total. In the test file, there are 9944 recipes that we need to predict.

# You can find the competition from this website: [www.kaggle.com/c/whats-cooking](http://www.kaggle.com/c/whats-cooking)

# And you can download the dataset from this website: [www.kaggle.com/c/whats-cooking/data](http://www.kaggle.com/c/whats-cooking/data)

**Machine Learning Task:**

Our aim with this project is to use recipe ingredients to categorize the cuisine.

Problem definition: Multiclass Classification

1) Features: all V ingredients, and each recipe is represented as a vector of length V.

2) Classes: all 20 kinds of cuisines

**Machine Learning Techniques:**

ML Algorithms:

1. SVM (multiple classes version)

2. Random Forest

3. Naive Bayes

4. Class-baed Association Rule Mining

5. Neural Network

Bagging & Boosting

Visualization:

1.reduce V-dimensional vectors to 2- or 3- dimensional ones

2. Methods:

1) PCA

2) t-SNE

Contribution:

1. Using many traditional ML techniques

2. Compare the performance of each used technique

3. Visualizing the recipes in 2-D/3-D space to show the closeness of recipes.