

### Data Acquisition

- IRS 990 forms on AWS
- NCCS BMF files

### Dataset Preprocessing

- Confidence A
- Link datasets

### Text Preprocessing

- Tokenizing / stop words
- Check spelling

## 1. Data Preprocessing

Naive Bayes /  
Random Forest

Neural Network

### Bag-of-Words Approach

- Stemming / Lemmatizing
- Word Count
- TF-IDF

### Word Embedding Approach

- GloVe 6B, 100 dimensions.

## 2. Word Representation and Feature Extraction

### Grid Search

- Satisficing decision table
- Optimizing decision table

### Stochastic Search

- Hyperparameters for hidden layers

← CNN

## 3. Training and Intermediate Decision Making

### 4. Training Model Finalist

- 90% training, 10% testing.