# Predict Fraud from Financial Text by Machine Learning

According to the survey by Ernst & Young (2010), the fraudulent activity has increased in the post financial crisis year. Therefore, auditors, regulators, and investors are responsible for detecting fraud by choosing both cost-effective and high-quality tools. Studies have shown that publicly listed companies' textual disclosures contain significant information about economic activities and intangibles related to unethical corporate behaviors. There are some scholars conducting several methods to detect the fraud by the language used in the management discussion and analysis. Thus, this project is going to explore the feasibility and accuracy of detecting fraud via several machine learning algorithms. Machine learning method can be a cost-effective fraud-detection tool for professionals in the financial industry. By comparing that effectiveness of our method to alternative fraud detection approaches across our data, we want to figure out the best predicted method to perform the fraud.

**Data Source:**

The text data is from the firm's annual 10-K filing which is public online, and we will mainly focus on the discussion part, management discussion and analysis (MD&A).



**UNITED STATES**
**SECURITIES AND EXCHANGE COMMISSION**
Washington, D.C. 20549

**FORM 10-K**

☒ ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the fiscal year ended September 30, 2017
or
☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the transition period from _____ to _____
Commission File Number: 001-36743

**Apple Inc.**
(Exact name of Registrant as specified in its charter)

Example: 10-K filing of Apple Inc.

Example: MD&A of Apple Inc.

Since 1982, the U.S. Securities and Exchange Commission has issued Accounting and Auditing Enforcement Releases (AAERs) during or at the end of an investigation against a company, an auditor, or an officer for alleged accounting and/or auditing misconduct. Based on the AAERs, we can obtain our label: 1 – fraud; 0 – not fraud.

Specifically, our project is a classification problem and we need to build a classifier to tell labels.

**Method**

1. Data Clean: such as merging the information from AAERs and MD&A; removing punctuations and other basic work for next step

2. Data exploratory Analysis: create word cloud; plot hist gram of the number of sentences in each company's AAERs

3. Word2Vect: convert the text information into numerical information

4. Modeling: Apply the dataset to several machine learning method, such as SVM, Neural Network, Random Forrest etc.

5. Comparison of the machine learning algorithms by decision matrix