**Identify Fraud from Enron Email**

Enron dataset building a « person of interest » identifier

Established in 1985, Enron became one of the largest companies in the world in 2000. In December 2001 it filed for bankruptcy, obliterating thousands of jobs and $60 billion in market value.

Investigation and trial lasted four years.

Arthur Andersen audit company was involved in the scandal. In 2002, the firm voluntarily surrendered its licenses to practice as [Certified Public Accountants](https://en.wikipedia.org/wiki/Certified_Public_Accountant) in the United States after it was [found guilty of crimes](https://en.wikipedia.org/wiki/Enron_scandal) in the firm's [auditing](https://en.wikipedia.org/wiki/Audit) of [Enron](https://en.wikipedia.org/wiki/Enron).

Enron scandal was a major milestone for modern ethics and compliance rules :

In reaction to this major corporate and accounting scandal together with others in the same period, as Worldcom, a law was voted, called Sarbanes–Oxley Act , also known as the "Public Company Accounting Reform and Investor Protection Act". The sections of the bill cover responsibilities of a public corporation's board of directors, add criminal penalties for certain misconduct, and require the [Securities and Exchange Commission](https://en.wikipedia.org/wiki/United_States_Securities_and_Exchange_Commission) to create regulations to define how public corporations are to comply with the law.

The investigation into Enron's collapse was conducted by the Enron Task Force, a team of federal prosecutors within the Justice Department's Criminal Division, and agents from the FBI and the Internal Revenue Service Criminal Investigations Division. The Enron Task Force also coordinated with the Securities and Exchange Commission. The Enron Task Force was part of the President's Corporate Fraud Task Force, created in July 2002 to investigate allegations of fraud and corruption at U.S. corporations.

Enron Task force madesignificant amount of information public record, including tens of thousands of emails and detailed financial data for Enron employees. Dataset is made of this information.

In addition to Enron dataset, a list of Persons Of Interest (POI) was created manually. Is a POI either :someone who was indicted, or who settled without admitting guilt or who testified in exchange for immunity.

Goal of this project is to build a person of interest identifier based on financial and email data made public as a result of the Enron scandal. For this, as the amount of data available is hudge, machine learning tools are a must.

# Dataset

Aim of this project is to define features that allow to identify person of interest.

This is meant to identify features that could help to identify future fraudsters, dishonnest employees, managers or consultants.

## Dataset source and python modules used.

*We will use Enron email dataset, in its 2015 version downloaded from link below :*

[*https://www.cs.cmu.edu/~./enron/enron\_mail\_20150507.tar.gz*](https://www.cs.cmu.edu/~./enron/enron_mail_20150507.tar.gz)

*In order to be in line with udacity « machine learning » explanations, we will use python 2.7 version together with skitlearn version 0.18 :*

*>>> import sklearn*

*>>> sklearn.\_\_version\_\_*

*'0.18'*

*However, both those versions are now decommissioned and code might have to be adapted to python 3 and sklearn 0.24.*

Email and finance data are combined into a single dataset, that we will explore in in this project.

## Enron dataset : Persons Of Interest (Poi)

Among 146 records in the dataset, 18 relate to persons of interest : (‘poi’)

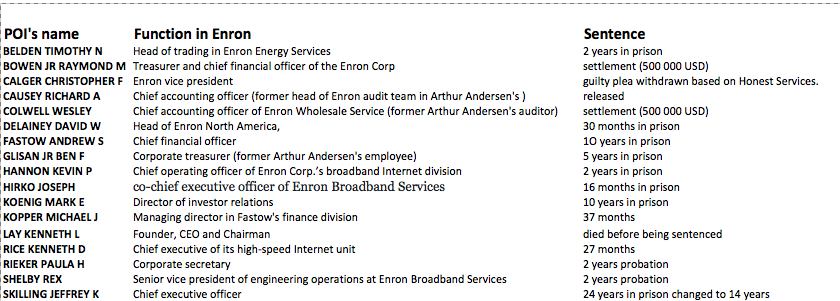
Enron had 29000 employees in 2002, so 146 records of which 18 of POI class could seem not a large sample. However, 18 persons of interest seems a significant proportion in the sample : 12%. In addition, each record correspond to a person and gathers a lot of emails for each.

Considering the size of present sample, any conclusion would have to be validated with a larger sample.

About POI in our dataset :

Here is information about their former function in Enron and how they were sentenced.

We can see that all POIs of the dataset seem not to have the same level in implication in fraud.



## Enron dataset : features and labels

The dataset, contains 146 records .consisting of  20 features  and ‘poi’label

There are 14 features related to monetary information (« MONEY ») and 6 about mail content (« MAIL »)

Features\_list : <class 'pandas.core.frame.DataFrame'>

Index: 146 entries, ALLEN PHILLIP K to YEAP SOON

Data columns (total 21 columns): My feature/label classification

bonus MONEY

deferral\_payments MONEY

deferred\_income MONEY

director\_fees MONEY

email\_address MAIL

exercised\_stock\_options MONEY

expenses MONEY

from\_messages MAIL

from\_poi\_to\_this\_person MAIL

from\_this\_person\_to\_poi MAIL

loan\_advances MONEY

long\_term\_incentive MONEY

other MONEY

poi LABEL

restricted\_stock MONEY

restricted\_stock\_deferred MONEY

salary MONEY

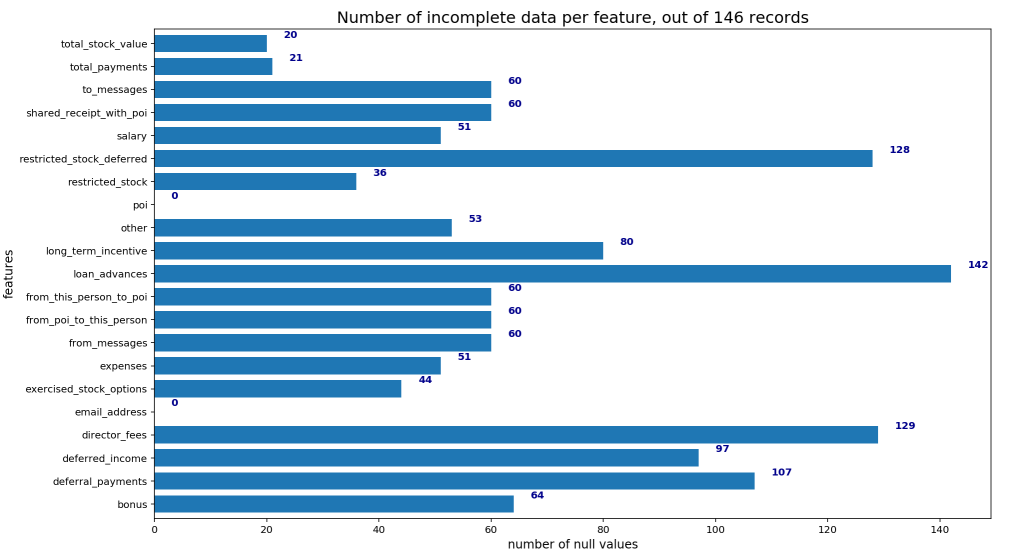
shared\_receipt\_with\_poi MAIL

to\_messages MAIL

total\_payments MONEY

total\_stock\_value MONEY

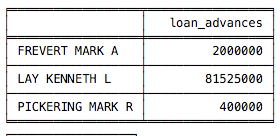
## Dataset : removal of non significant records



We can see that three features have a lot of null values :

* restricted\_stock\_deffered ;
* loan advances ;
* directors fees ;

In particular, let’s have a look at non zero loan advances records :



Number of non zero values shows that Mark Frevert and Mark Pickering, Enron executives who were not involved in Enron scandal,received loans that were high but much lower than that of Kenneth Lay (81M USD !!). The fact that only three records are significant allows us to remove the « loan\_advances » column from dataset. This feature seems not to be very significant for our analysis : This can not help to identify other PoI.

Other two features are kept : also there are not a large amount of information, those two information might be of some interest.

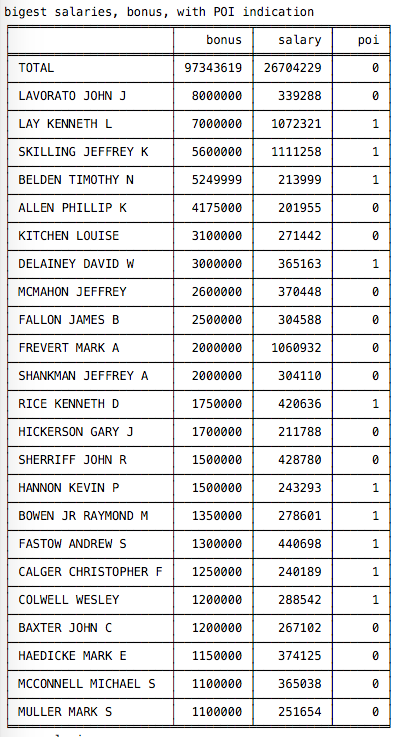
We will then remove rows with no significant data :

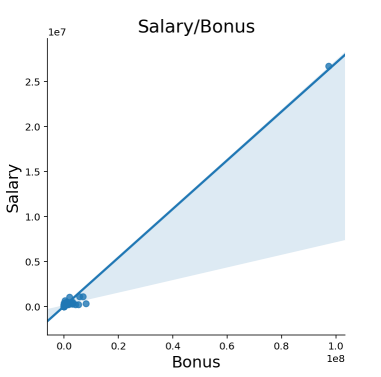
Let’s identify rows where more than 19 features are NAN and remove them from the dataset.

After removal, dataset contains 141 records and 20 features.

From this point of the analysis, all NaN values in money features are replaced by 0.

## Outliers identification and removal





A scattered plot together with list of bonus and salaries per person reveals outlier to be obviously removed from dataset : Total

After removal, remaining outliers values appear to be that of most severly sentenced POI.

But.. Among some of the top ranking bonus, we can find non POI. Internet search about John Lavorato leads to an article in the British newspaper « the Guardian », that shows why John Lavorato and Louise Kitchen received such bonuses and were not indicted :

*Quote :« The doubts were raised about Mr Skilling's testimony on the same day as CNN revealed that some 500 Enron staff had received windfalls ranging from $1,000 to $5m. The payments were made to retain staff as the firm faced collapse. To get the cash, the staff agreed to stay for 90 days.*

*The highest payment of $5m went to John Lavorato, who ran Enron's energy trading business, while Louise Kitchen, the division's British-born chief operating officer, pocketed $2m. Both have taken up new jobs with UBS Warburg, the investment bank that now owns the division. » (…)*

*Workers laid off by Enron have, by contrast, been paid the minimum severance of $4,500 before tax and many are struggling to find work. » end of Quote*

*We can see that both John Lavorato and Louise Kitchen ratio bonus/salary is much higher than that of Mark Frevert* who was chief executive of Enron Europe from 1996 until 2000 and then appointed chairman of Enron in 2001. All three are not POI.

This ratio bonus/salary, seems to us an interesting feature to create. As those tremendous bonuses were granted a few months before bankrupt, this ratio marks desperate attempt to hide desastrous situation of the company. This sort of indicator could be an alert for SEC in future comparable situations.

Another interesting key indicator is the mail flow from and to POI. John Lavorato and Louise Kitchen were granted bonuses by well informed people about fraud, so, they were in contact with POI .

# Features

Exploration of dataset enabled to remove outliers, gave hints for new features to create and others featrures to supppress.

## Additional features creation

We create three additional features :

* % of mails from POI
* % of mails to POI
* salary/bonus ratio