**ENRON Final project**

**By Eric Rabaute**

# Introduction

This analysis is part of Udacity training.

The objective is to make a data analysis based on Machine learning methods and tools.

# MachinE learning objective

The questions we want to answer in this study are:

* How can we identify with a good accurate result the person of interest related to Enron Fraud? (ie.: wich available features are the most valuable?)
* How can we optimize this Person of Interest Identifier thanks to Machine learning methods and tools?

# Features selection

The first task is to choose among all available features the one that will helps us

**financial features**: ['salary', 'deferral\_payments', 'total\_payments', 'loan\_advances', 'bonus', 'restricted\_stock\_deferred', 'deferred\_income', 'total\_stock\_value', 'expenses', 'exercised\_stock\_options', 'other', 'long\_term\_incentive', 'restricted\_stock', 'director\_fees'] (all units are in US dollars)

**email features**: ['to\_messages', 'email\_address', 'from\_poi\_to\_this\_person', 'from\_messages', 'from\_this\_person\_to\_poi', 'shared\_receipt\_with\_poi'] (units are generally number of emails messages; notable exception is ‘email\_address’, which is a text string)

Number of persons on the sample:146

Number of poi on the sample:18 (0.12%)

First observations on data show a lot of features missed (ie. NaN value). I proposed to make a quantitative check on this value for all features first. Indeed, if a feature is miss for more than 50% absent of all data set, it could be not interesting to go further with it.

Hereafter the count of NaN for all features (max to min order) for a total of **146 persons**:

(1) loan\_advances:142

(2) director\_fees:129

(3) restricted\_stock\_deferred:128

(4) deferral\_payments:107

(5) deferred\_income:97

(6) long\_term\_incentive:80

(7) bonus:64

(8) to\_messages:60

(9) shared\_receipt\_with\_poi:60

(10) from\_poi\_to\_this\_person:60

(11) from\_messages:60

(12) from\_this\_person\_to\_poi:60

(13) other:53

(14) salary:51

(15) expenses:51

(16) exercised\_stock\_options:44

(17) restricted\_stock:36

(18) email\_address:35

(19) total\_payments:21

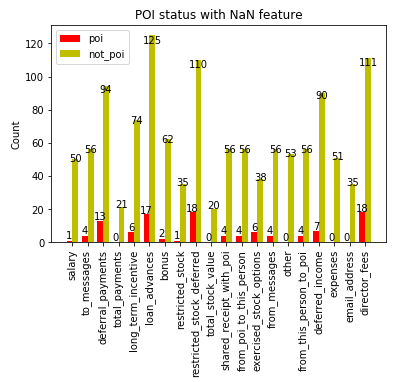
(20) total\_stock\_value:20

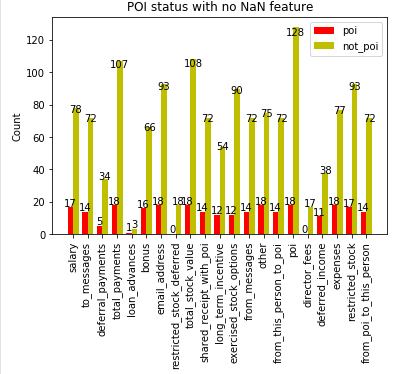
It is interesting to see that some features are more than 97% missing (loan advances). It could be explained by the fact the information is difficult to get that is not a very valuable information or this feature absence means something about the person (ex.: director fees means the person is a high ranking person and so likely a POI).

I propose to move forward this question by correlating the number of NaN and the POI status (ie. Have we got a higher POI status rate if NaN value is more or less present for one feature).

Hereafter two figures:

* The first one is number of POI and not POI persons according features equals to NaN value
* The second is the same with persons with value different of NaN





**Analysis**: At a first sight, no clear output is raised except that theory about ‘director fees’ is false, indeed, there is no ‘poi’ with director fees different from NaN. This is strange since we could expect to have POI inside director committee.

I propose to go a last step further with the ratio between poi and not poi for none NaN value.

(1) loan\_advances: poi=0.25% (n=4)

(2) deferred\_income: poi=0.22% (n=49)

(3) bonus: poi=0.2% (n=82)

(4) other: poi=0.19% (n=93)

(5) expenses: poi=0.19% (n=95)

(6) salary: poi=0.18% (n=95)

(7) long\_term\_incentive: poi=0.18% (n=66)

(8) to\_messages: poi=0.16% (n=86)

(9) shared\_receipt\_with\_poi: poi=0.16% (n=86)

(10) from\_poi\_to\_this\_person: poi=0.16% (n=86)

(11) from\_messages: poi=0.16% (n=86)

(12) from\_this\_person\_to\_poi: poi=0.16% (n=86)

(13) email\_address: poi=0.16% (n=111)

(14) restricted\_stock: poi=0.15% (n=110)

(15) total\_payments: poi=0.14% (n=125)

(16) total\_stock\_value: poi=0.14% (n=126)

(17) deferral\_payments: poi=0.13% (n=39)

(18) exercised\_stock\_options: poi=0.12% (n=102)

(19) poi: poi=0.12% (n=146)

(20) restricted\_stock\_deferred: poi=0.0% (n=18)

(21) director\_fees: poi=0.0% (n=17)

**Analysis**: On first observations, we noticed that the ratio between poi and not poi in the sample is 12% (18/146). We can assume the features we would like to select have the less NaN value to provide most information as possible and the poi sample mean different from sample mean.

I propose to exclude in a first step the features where NaN values are more than 50% (ie. 73), ie:

(1) loan\_advances: poi=0.25% (n=4)

(2) deferred\_income: poi=0.22% (n=49)

(7) long\_term\_incentive: poi=0.18% (n=66)

(17) deferral\_payments: poi=0.13% (n=39)

(20) restricted\_stock\_deferred: poi=0.0% (n=18)

(21) director\_fees: poi=0.0% (n=17)

(3) bonus: poi=0.2% (n=82)

(4) other: poi=0.19% (n=93)

(5) expenses: poi=0.19% (n=95)

(6) salary: poi=0.18% (n=95)

(8) to\_messages: poi=0.16% (n=86)

(9) shared\_receipt\_with\_poi: poi=0.16% (n=86)

(10) from\_poi\_to\_this\_person: poi=0.16% (n=86)

(11) from\_messages: poi=0.16% (n=86)

(12) from\_this\_person\_to\_poi: poi=0.16% (n=86)

(13) email\_address: poi=0.16% (n=111)

(14) restricted\_stock: poi=0.15% (n=110)

(15) total\_payments: poi=0.14% (n=125)

(16) total\_stock\_value: poi=0.14% (n=126)

(18) exercised\_stock\_options: poi=0.12% (n=102)

(19) poi: poi=0.12% (n=146)

