**How do I Complete this Project?**

This project is connected to the [Intro to Machine Learning](https://www.udacity.com/course/viewer#!/c-ud120-nd) course, but depending on your background knowledge of machine learning, you may not need to take the whole thing to complete this project.

A note before you begin: the mini-projects in the Intro to Machine Learning class were mostly designed to have lots of data points, give intuitive results, and otherwise behave nicely. This project is significantly tougher in that we're now using the real data, which can be messy and does not have as many data points as we usually hope for when doing machine learning. Don't get discouraged--imperfect data is something you need to be used to as a data analyst! If you encounter something you haven't seen before, take a step back and think about a smart way around. You can do it!

**Project Overview**

In 2000, Enron was one of the largest companies in the United States. By 2002, it had collapsed into bankruptcy due to widespread corporate fraud. In the resulting Federal investigation, a significant amount of typically confidential information entered into the public record, including tens of thousands of emails and detailed financial data for top executives. In this project, you will play detective, and put your new skills to use by building a person of interest identifier based on financial and email data made public as a result of the Enron scandal. To assist you in your detective work, we've combined this data with a hand-generated list of persons of interest in the fraud case, which means individuals who were indicted, reached a settlement or plea deal with the government, or testified in exchange for prosecution immunity.

**Resources Needed**

You should have python and sklearn running on your computer, as well as the starter code (both python scripts and the Enron dataset) that you downloaded as part of the first mini-project in the Intro to Machine Learning course. You can get the starter code on git: git clone https://github.com/udacity/ud120-projects.git

The starter code can be found in the final\_project directory of the codebase that you downloaded for use with the mini-projects. Some relevant files:   
  
poi\_id.py : Starter code for the POI identifier, you will write your analysis here. You will also submit a version of this file for your evaluator to verify your algorithm and results.   
  
final\_project\_dataset.pkl : The dataset for the project, more details below.   
  
tester.py : When you turn in your analysis for evaluation by Udacity, you will submit the algorithm, dataset and list of features that you use (these are created automatically in poi\_id.py). The evaluator will then use this code to test your result, to make sure we see performance that’s similar to what you report. You don’t need to do anything with this code, but we provide it for transparency and for your reference.   
  
emails\_by\_address : this directory contains many text files, each of which contains all the messages to or from a particular email address. It is for your reference, if you want to create more advanced features based on the details of the emails dataset. You do not need to process the e-mail corpus in order to complete the project.

**Steps to Success**

We will provide you with starter code that reads in the data, takes your features of choice, then puts them into a numpy array, which is the input form that most sklearn functions assume. Your job is to engineer the features, pick and tune an algorithm, and to test and evaluate your identifier. Several of the mini-projects were designed with this final project in mind, so be on the lookout for ways to use the work you’ve already done.

As preprocessing to this project, we've combined the Enron email and financial data into a dictionary, where each key-value pair in the dictionary corresponds to one person. The dictionary key is the person's name, and the value is another dictionary, which contains the names of all the features and their values for that person. The features in the data fall into three major types, namely financial features, email features and POI labels.

**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', 'poi', 'shared\_receipt\_with\_poi'] (units are generally number of emails messages; notable exception is ‘email\_address’, which is a text string)

**POI label**: [‘poi’] (boolean, represented as integer)

You are encouraged to make, transform or rescale new features from the starter features. If you do this, you should store the new feature to my\_dataset, and if you use the new feature in the final algorithm, you should also add the feature name to my\_feature\_list, so your evaluator can access it during testing. For a concrete example of a new feature that you could add to the dataset, refer to the lesson on Feature Selection.

In addition, we advise that you keep notes as you work through the project. As part of your project submission, you will compose answers to a series of questions (given on the next page) to understand your approach towards different aspects of the analysis. Your thought process is, in many ways, more important than your final project and we will by trying to probe your thought process in these questions.

## General Submission and Evaluation Overview

Your submission will contain several files: the code/classifier you create and some written documentation of your work. We will evaluate your project according to the rubric [here](https://docs.google.com/a/knowlabs.com/document/d/17-JwNQH1aRxtqMkJ6zpCL_68kh5F6uSbDXcJS26vZWY/pub?embedded=true); only projects that satisfy **all** "meets expectations" items will pass. **Please self-evaluate before you submit!** If you don't think your project meets all the criteria, the project evaluator likely won't either.

## Submission

Ready to submit your project? Go back to your Udacity Home, click on the project, and follow the instructions to submit!

* You can either send us a GitHub link of the files or upload a compressed directory (zip file).
* Inside the zip folder include a text file with a list of Web sites, books, forums, blog posts, GitHub repositories etc that you referred to or used in this submission (Add N/A if you did not use such resources).

It can take us up to two weeks to grade the project so keep checking back for updates.

If you are having any problems submitting your project or wish to check on the status of your submission, please email us at dataanalyst-project@udacity.com.

### Items to include in submission:

### Code/Classifier

When making your classifier, you will create three pickle files (my\_dataset.pkl, my\_classifier.pkl,my\_feature\_list.pkl). The project evaluator will test these using the tester.py script. You are encouraged to use this script before submitting to gauge if your performance is good enough. You should also include your modified poi\_id.py file in case of any issues with running your code or to verify what is reported in your question responses (see next paragraph). Notably, we should be able to run poi\_id.py to generate the three pickle files that reflect your final algorithm, without needing to modify the script in any way. If you have intermediate code that you would like to provide as supplemental materials, it is encouraged for you to save them in files separate from poi\_id.py. If you do so, be sure to provide a readme file that explains what each file is for.

### Documentation of Your Work

Document the work you've done by answering (in about a paragraph each) the questions found [here](https://docs.google.com/document/d/1NDgi1PrNJP7WTbfSUuRUnz8yzs5nGVTSzpO7oeNTEWA/pub?embedded=true). You can write your answers in a PDF, text/markdown file, HTML, or similar format. The responses in your documentation should allow a reviewer to understand and follow the steps you took in your project and to verify your understanding of the methods you have performed.

### Text File Listing Your References

A list of Web sites, books, forums, blog posts, github repositories etc. that you referred to or used in this submission (add N/A if you did not use such resources). Please carefully read the following statement and include it in your document “I hereby confirm that this submission is my work. I have cited above the origins of any parts of the submission that were taken from Websites, books, forums, blog posts, github repositories, etc.

### Good Luck!