**我要如何完成此项目？**

在开始之前，你应该注意，此迷你项目需要大量数据点才能给出直观的结果，并且良好地运行起来。 此项目更为棘手的原因在于，我们使用了真实的数据，这些数据可以是杂乱无章的，而且在进行机器学习时不具有我们所希望的大量数据点。 不要失去信心——作为数据分析师，你只需要习惯不完美的数据！如果你遇到之前没有见过的事物，请退后一步想想聪明的解决之道。要相信自己！

**项目概述**

安然曾是 2000 年美国最大的公司之一。2002 年，由于其存在大量的企业欺诈行为，这个昔日的大集团土崩瓦解。 在随后联邦进行的调查过程中，大量有代表性的保密信息进入了公众的视线，包括成千上万涉及高管的邮件和详细的财务数据。 你将在此项目中扮演侦探，运用你的新技能，根据安然丑闻中公开的财务和邮件数据来构建相关人士识别符。 为了协助你进行侦查工作，我们已将数据与手动整理出来的欺诈案涉案人员列表进行了合并， 这意味着被起诉的人员要么达成和解，要么向政府签署认罪协议，再或者出庭作证以获得免受起诉的豁免权。

**需要的资源**

你的计算机上应有 python 和 sklearn，以及你随“机器学习入门”课程的首个迷你项目一并下载的初始代码（python 脚本和安然数据集）。 你可以从 git 上获取初始代码：

git clone https://github.com/udacity/ud120-projects.git

你可以在下载下来用于迷你项目的代码库 final\_project 目录中找到初始代码。相关文件如下所示：

poi\_id.py：用于 POI 识别符的初始代码，你将在此处撰写你的分析报告。你也将提交此文件的副本，用于评估人员检验你的算法和结果。

final\_project\_dataset.pkl：项目数据集，详情如下。

tester.py：在你提交供优达学城评估的分析报告时，你将随附算法、数据集和你使用的特征列表（这些是在 poi\_id.py 中自动创建的）。 评估人员将在此后使用这一代码来测试你的结果，以确保性能与你在报告中所述类似。你无需处理这一代码，我们只是将它呈现出来供你参考。

emails\_by\_address：该目录包含许多文本文件，每个文件又包含特定邮箱的往来邮件。 你可以进行参考，并且可以根据邮件数据集的详细信息创建更多的高级特征。你无需处理电子邮件语料库来完成项目。

**迈向成功**

我们将给予你可读入数据的初始代码，将你选择的特征放入 numpy 数组中，该数组是大多数 sklearn 函数假定的输入表单。 你要做的就是设计特征，选择并调整算法，用以测试和评估识别符。 我们在设计数个迷你项目之初就想到了这个最终的项目，因此请记得借助你已完成的工作成果。

在预处理此项目时，我们已将安然邮件和财务数据与字典结合在一起，字典中的每对键值对应一个人。 字典键是人名，值是另一个字典（包含此人的所有特征名和对应的值）。 数据中的特征分为三大类，即财务特征、邮件特征和 POI 标签。

**财务特征**: ['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'] (单位均是美元）

**邮件特征**: ['to\_messages', 'email\_address', 'from\_poi\_to\_this\_person', 'from\_messages', 'from\_this\_person\_to\_poi', 'shared\_receipt\_with\_poi'] (单位通常是电子邮件的数量，明显的例外是 ‘email\_address’，这是一个字符串）

**POI 标签**: [‘poi’] (boolean，整数)

我们鼓励你在启动器功能中制作，转换或重新调整新功能。如果这样做，你应该把新功能存储到my\_dataset，如果你想在最终算法中使用新功能，你还应该将功能名称添加到 my\_feature\_list，以便于你的评估者可以在测试期间访问它。关于如何在数据集中添加具体的新要素的例子，可以参考“特征选择”这一课。

此外，我们还建议你可以在完成项目过程中做一些记号。你可以写出系列问题的答案（在下一页），将这个作为提交的项目的一部分，以便于评估者了解到你对于不同方面分析的方法。你的思维过程在很大程度上比你的最终项目更重要，我们将通过你在这些问题的解答中了解你的思维过程。