## EDA: 欄位的資料類型介紹及處理

* 我們在取得資料後的第一步就是要先了解資料
  1. 將資料做EDA (4步驟)
  2. 如果某feature的資料類型是字串的話，須將他們轉換為數值資料類型(可以參考<https://www.itread01.com/content/1546669104.html>)
     1. Label encoding : 該feature的字串之間是"有順序"的關係，那麼就可以把他們依照排序給上1,2,3,4….的代號。
     2. One Hot encoding : 該feature的字串之間是"平等"的關係，例如:國家。就把他們用dummies來做處理。
* 在pandas DataFrame中最常見的資料類型有三種
  1. Float64
  2. Int64
  3. Object :只要資料中包含”字串”，就會被電腦歸類為此種類
* 遇到的問題:

# 設定 data\_path, 並讀取 app\_train

#記得路徑形式的字串前面要加上"R"，而且屁股的地方記得不能有"\(斜線)"

dir\_data = R'C:\Users\gerog\Desktop\files of HC\Get from compagn\Common Data'

f\_app\_train = os.path.join(dir\_data, 'application\_train.csv')

app\_train = pd.read\_csv(f\_app\_train)

DF.select\_dtypes(include=["object"]) 會回傳符合條件的部分DF