* Which topic did you choose to apply the data science methodology to? **(2 marks)**
  + I want to proceed with "Credit Cards"
* Next, you will play the role of the client and the data scientist.

Using the topic that you selected, complete the Business Understanding stage by coming up with a problem that you would like to solve and phrasing it in the form of a question that you will use data to answer. **(3 marks)**

You are required to:

Describe the problem, related to the topic you selected.

Phrase the problem as a question to be answered using data.

For example, using the food recipes use case discussed in the labs, the question that we defined was, "Can we automatically determine the cuisine of a given dish based on its ingredients?".

* + **Problem:** Credit card fraud stands as a major problem for worldwide financial institutions. Annual lost due to these accounts to billions of dollars. We can observe this from many financial reports. Such as (Bhattacharyya et al., 2011) 10th annual online fraud report by CyberSource shows that estimated loss due to online fraud is $4 billion for 2008 which is 11% increase than $3.6 billion loss in 2007and in 2006, fraud in the United Kingdom alone was estimated to be £535 million in 2007 and now costing around 13.9 billion a year (Mahdi et al., 2010). From 2006 to 2008, the UK alone has lost £427.0 million to £609.90 million due to credit and debit card fraud (Woolsey & Schilz, 2011). Although there is some decrease in such losses after implementation of detection and prevention systems by government and bank, card-not-present fraud losses are increasing at a higher rate due to online transactions. The worst thing is it is still increasing un-protective and un-detective way.

**Question:**Flagging or finding the probability of a transaction being a fraud.

* Briefly explain how you would complete each of the following stages for the problem that you described in the Business Understanding stage, so that you are ultimately able to answer the question that you came up with. **(5 marks)**:

Analytic Approach

Data Requirements

Data Collection

Data Understanding and Preparation

Modeling and Evaluation

You can always refer to the labs as a reference with describing how you would complete each stage for your problem.

* + Logistic Regression can be used to classify the transactions are fraudulent or not based on the probability of the transaction being a fraud 2. Data Requirements can range from user data to individual transaction details. It includes the city, state, and Country of user's residence, user interests, spending patterns, current occupation, earnings, transaction details, date of transaction, time of the transaction, the vendor of the transaction, amount, location, and device used for the transaction. 3. Data can be collected from different sources. Like user data, spending pattern and interests can be obtained during the application process of the card by giving out detailed forms for collecting personal information. bank account statements, social media profiles, if available and having personal discussions with the user before giving out a credit card. Also, a dump of all the transactions that have been flagged as safe or fraud and all the user and transaction related data for the same transactions. 4. A basic understanding can be generated from the data by clustering the spending and social media analysis to identify the interests and spending patterns of the user. Using the same, coupled with the other available user and transaction data, we can establish our data set by cleaning and structuring the available data. Missing and non-uniform information should be ignored as of now and can be incorporated in further iterations of the model or as and when we run campaigns to collect the missing or additional data. 5. Now split the data set into 70-30 for training and testing the model. Develop a logistic regression model using the training data set to determine the probability of a transaction is a fraud. And later test the same model on the testing data set. Now, the model can be evaluated by using it on the live data set to predict if the transaction is fraud or not.

There will be several iterations and introduction or removal of features for fraud detection before we implement the model for good.