**Problem Statement**

Improve accuracy of self reported income for clients prospective clients of a credit card company.

*Proposal*

Predict individual salary based on demographic and employment data collected as part of the 1994/95 survey and leverage the predicted salary to compare against reported data.

**Client**

Client is a credit card company that wants to combine self reported income details by their clients against those predicted based on client demographics collected from various sources. They would use this analysis to understand the most critical data points needed for a reasonable accurate income prediction and then be able to combine self reported and predicted incomes for decisions like offer targeting and credit line assignment. The goal is to give higher lines to high earning individuals and minimize self reporting errors or system gaming

**Data Sourcing**

I will be using a 1994/95 census data set with demographic and employment variables.

This data set contains weighted census data extracted from the 1994 and 1995 [Current Population Surveys](http://www.census.gov/cps/methodology/techdocs.html) conducted by the U.S. Census Bureau. The data contains 41 demographic and employment related variables.

The instance weight indicates the number of people in the population that each record represents due to stratified sampling. To do real analysis and derive conclusions, this field must be used. This attribute should \*not\* be used in the classifiers.

<http://archive.ics.uci.edu/ml/machine-learning-databases/census-income-mld/census-income.data.html>

**Approach**

1. First step would be to explore the data and look at relationships between income and other variables
2. Do a first cut based on correlation and follow by variable clustering and random forrest approach to identify most critical variables and eliminate collinearity
3. Use regression and random forest as competing models to predict income

**Key Deliverables**

Key deliverables would include the following

1. Code for the analysis and model development
2. A doc/presentation with
   1. Detailed approach
   2. Reasons for variable selection
   3. Pros/cons of using a specific algorithm
   4. Regression / Forest algo. performance comparison