# Project Proposal

**Team members:** Daniel Casares, Felipe Olivera

**Project title:** Detecting fraud on credit card transactions

**Description of the problem context:** Banks, merchants and credit card processors companies lose billions of dollars every year due to credit card fraud. Credit card data can be stolen by criminals using a variety of methods: bluetooth-enabled data skimming devices can be placed on card readers, the data might be stolen from a database by cyber-criminals, etc. Sometimes the criminal is simply the clerk at the checkout line at the grocery or in a restaurant, where the victim's card is swiped through a small device or surreptitiously jotted down.

Due to the increasing volume of transactions that financial institutions deal with everyday, the process of manually detecting fraud is very slow and expensive. Therefore, the objective is to create a data-based model that can detect fraud on new (previously unseen) transactions, willing to minimize the financial losses of the company related to credit card fraud. Real data from an anonymous latin american financial company will be used to build the model.

**Type of the problem:** Supervised learning; training data is labeled by class (fraud or no fraud) in advance and a model will be used to predict the class of new observations. Methods including decision trees and random forest can be investigated.

**Initial thoughts on techniques that might be used:** Preprocess the credit card transactions data to make it suitable for the model. Come up with new features to add valuable information to the model. Construct and validate the learning model using performance measures. To evaluate the model we would like to introduce a cost-based metric to minimize the losses of the company due to fraud. The objective is to construct a learning model that can detect fraud on new (previously unseen) credit card transactions.

**Description of the dataset:** We got the dataset while participating in a research project about aggregated features for credit card fraud detection, advised by Dr. Gustavo Vazquez, Director of the Computer Science Faculty at Universidad Catolica del Uruguay. (Remark: The results from this project only cover a small fraction of what we learned here in class). The data is not publicly available and is from a local company that wants to stay anonymous. The datasets contains 41,091,288 transactions made by credit cards between Jul-2014 and Jun-2015, where there are 12,632 frauds. The dataset is highly unbalanced, the positive class (frauds) account for 0.031% of all transactions. Each transaction contains the following variables:

- Time: Date and time of the transaction
- Tokenized PAN: Identification of the credit card
- Upscale: Indicates if the card holder is an upscale customer
- Entry mode: ie. Chip and PIN, magnetic stripe, …
- Amount: Amount of the transaction in US dollars
- Merchant code: Identification of the merchant type
- Merchant group: Merchant group identification
- Country: Country of transaction
- Bank: Issuer of the card