**Description**

The [competition is to predict repeat buyers](http://www.kaggle.com/c/acquire-valued-shoppers-challenge/) (those who redeem a coupon and purchase that product afterwards). For this we have the labelled data (did become repeat buyer, did not become repeat buyer) for about 160,000 shoppers (the trainHistory set).

Our task is to predict the labels for about 150,000 other shoppers (the testHisotory set). For this we can use a file called transactions.csv. It’s a huge file (unzipped about 22GB) containing nearly 350 million rows. The total amount spend in the transaction data is nearing 1.5 billion.

1. Data Reduction: chose transactions from only one chain (chain\_id=4) and keep data only from 1 year.
   * reduce data from 350M to 3.3M
2. Combine trainHistory and testHistory: In test dataset, assign 'repeatrips' to 0 and 'repeater' to 'f'
   * combined data has 0.3M rows and 8 features
3. Use only customers in combined data which have at least one transaction in the reduced transaction dataset
   * combine\_reduced data has 4133 rows and 8 features
4. Inner join 'combine\_reduced' with 'offers.csv' on feature='offer'
   * add offer's category/company/brand
   * new datasets has 4133 rows and 13 features
5. Feature Extraction: for each customer id, calculate total amount/quantity the customer pays for under offer's category/company/brand as well as total payment and total trips within one year
   * create X\_features: 4133 rowas and 22 features
6. Encoding Features: company/category/brand
7. Create X\_train,X\_test, Y\_train, Y\_test
   * test data is 'repeater' for each customer
8. Fit the model using Random Forest
   * AUC score is 1 under training data
   * can get probability that a customer repeat purchased for test data