## Problem Statement

All online travel agencies are scrambling to meet the Artificial Intelligence driven personalization standard set by Amazon and Netflix. In addition, the world of online travel has become a highly competitive space where brands try to capture our attention (and wallet) with recommending, comparing, matching, and sharing. For this assignment, we aim to create the optimal hotel recommendations for Expedia’s users that are searching for a hotel to book. For this assignment, you need to predict which “hotel cluster” the user is likely to book, given his (or her) search details. In doing so, you should be able to demonstrate your ability to use four different algorithms (of your choice). The data set can be found at Kaggle: Expedia Hotel Recommendations. To get you started, I would suggest you use train.csv which captured the logs of user behavior and destinations.csv which contains information related to hotel reviews made by users. You are also required to write a one page summary of your approach in getting to your prediction methods. I expect you to use a combination of R and Python in your answer.

### Data Data for this assignment has been collected from the following Kaggle competition: <https://www.kaggle.com/c/expedia-hotel-recommendations/data>

**train.csv** - the training set

**test.csv** - the test set

**destinations.csv** - hotel search latent attributes

**Assignment 5.4 - Summary**

The algorithms were chosen based on the understanding of the data from the EDA and preliminary analysis. Here are the basis of the selection.

**1. SVM or Support Vector Machine** SVM is capable of doing both classification and regression. Since we are trying to clasify users to hotel clusters, I have chosen SVM. As the clusters are multi level (100), this is using non-linear SVM. Non-linear SVM means that the boundary that the algorithm calculates doesn't have to be a straight line. The benefit is that we can capture much more complex relationships between the datapoints without having to perform difficult transformations on our own. The downside is that the training time is much longer as it's much more computationally intensive.

Using SVM, I have achieved the highest cross validation score.

**2. Naive Bayes classifier** I wanted to try Naive Bayes algorithm because it is simple and if the conditional independence assumption actually holds, a it will converge quicker than discriminative models like logistic regression, hence it will need less training data.

But it has the worst performance of the four models. Therefore, this classifier is not recommended for the problem at hand.

**3. Logistic Regression** Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). In this case, we can think of the problem as, whether a hotel fall in a specific cluster (yes/no) based on the chosen features.

Logistic Regression was close to the performance of SVM but slightly worse.

**4. K-Nearest Neighbor classifier** KNN is a non-parametric, lazy learning algorithm. In the currentcontext, the idea is to teach the model which users (with other similar characteristics) chose which hotel cluster, and predict future cluster assignment based on that learning.

KNN performed very similar to Logistic Regression for the model in question.