**Data Science Course**

**Project Description**

**Fannie Tseng**

**November 15, 2015**

**Project Title:** Using Machine Learning to Predict Completion and Learning Outcomes in Massive Open Online Courses (MOOCs)

**Dataset Description:**

The file is a person-course level file from all students who registered for any of the 16 MOOCs offered by HarvardX and MITX during the program’s first year (2013-14). This de-identified dataset has 641,139 person-course observations. The dataset includes administrative data (documenting the students’ interactions with the course materials) and user-provided data. The administrative data includes the following variables:

* Course name
* UserID
* 0/1 indicator for whether the student registered for the class
* 0/1 indicator for whether the student viewed the course materials
* 0/1 indicator for whether the student completed at least 50% of the lessons
* 0/1 indicator for whether the student earned a certificate
* Interaction with course materials: # of videos watched, # of times posted to forums, first and last dates interacting with course materials
* Course grade

The user-provided data includes the following variables:

* Year of birth
* Gender
* Country of residence
* Highest degree attained

**Analysis Plan:**

I would like to use machine learning to analyze the extent to which the very basic user-provided data (age, gender, country, degree) and course characteristics (subject, level of difficulty, as measured by prerequisites) can predict how the student will interact with the course materials (i.e. which of the outcomes the student will achieve).

**Research Questions:**

* To what extent does highest degree attained help predict MOOC outcomes?
* Do MOOCs really democratize learning? Are students from all backgrounds accessing and learning from the materials?
* Is it possible to identify students who “give up” because the course materials are too difficult?