Project Title: Reducing the Human Element within Forest Type Classification

Data set: http://archive.ics.uci.edu/ml/datasets/Covertype
~500K instances
Instance is composed of 54 (+1) features. These include reals (elevation, slope, distance to water), integers (hillshade index at different dates), and binaries (wilderness area, soil types)

Project Idea: Study previous work on this dataset to find areas for optimization, specifically with regards to reducing the reliance on (costly/expensive) labeled data.

Approach: First, I will read two academic papers written on classification approaches done on the same dataset. I will analyze their strengths, weaknesses and look for areas of improvement.

I will then run supervised classification techniques over the data: LDA and QDA. I will also run unsupervised K-means clustering, as I believe this will be the key towards completing my goal of the reduction of the human element. Ideally I am able to find a way to map clusters to labels. I will study the effects of increasing k and determine the optimal k value (elbow point) and how this affects the amount of human interaction involved.

Software:
I will use Sublime Text's Python interpreter as my IDE. I will use NumPy and the standard library to parse the data from its raw form. I will use Sci-Kit Learn's library for their LDA, QDA and K-means implementations, to avoid writing them from scratch. I will use Google's app suite for my graphs, presentation and final report.

References:

1. Blackard, Jock A., Dean, Denis J. "Comparative accuracies of artifical neural networks and discriminant analysis in predicting forest cover types from cartogrpahic variables". Computers and Electronics in Agriculture (1999).
2. Crain, Kevin, and Graham Davis. "Classifying Forest Type Using Cartographic Features." Stanford, Stanford, Dec. 2014, cs229.stanford.edu/proj2014/Kevin%20Crain,%20Graham%20Davis,%20Classifying%20Forest%20Cover%20Type%20using%20Cartographic%20Features.pdf.