**Research Question and Significance?**

Significance could be:

1. Develop a tutorial/guide to rapidly learn ML techniques and how you could apply them into one of the WOS projects.
2. How to use ML techniques to classify AFM images for research purposes.

**Background**

Machine learning (ML) applications have become an integral part of our daily lives. Large companies such as Google, Amazon, and Apple have developed tools that utilize ML techniques to solve complex problems. For example, the speech recognition technology behind Siri and Amazon’s Alexa are great examples of ML in everyday life. In addition, Amazon is beta testing a new kind of store that uses ML named AmazonGo which will require no checkout. Customers can grab the items they need and simply walk out of the store. Although machine learning has fairly recently become a popular interest the term has been around for more than fifty years. In 1959, the computer scientist Arthur Samuel defined machine learning as “*The field of study that gives computers the ability to learn without being explicitly programmed.” (ref 1*) Samuel’s research was important because the central principles of machine learning were verified by his early experiments. A more recent formal definition came in 1997 by Tom Mitchell in which he stated “*A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.” (ref2)* Mitchell’s definition gives us a more detailed description of what it means for a computer to learn. Depending on the complexity of the problem ML can be separated into two areas: Supervised and Unsupervised learning.

In supervised learning, a data set is provided and we have a notion of what the correct output should appear like and we are aware that there is a relationship between the input and the output. Supervised learning investigations can be divided into regression and classification problems. In a regression problem, results are being predicted from a continuous output. In essence, input variables are being mapped to a continuous function. In contrast, for a regression problem the input variables are being mapped into discrete categories. In the past, supervised learning problems have included topics such as housing prices and breast cancer prediction.

In unsupervised learning, a data set is given and some structure is found within the given data. The difference here is that we do not have a notion of what the results should look like. This approach can derive structure from the data where the effect of the variables is unknown. The structure is derived by clustering the data based on relationships among the variables.

A supervised learning problem can often be thought of as a regression problem. An important aspect in linear regression is to generate a good output function traditionally called a hypothesis. A hypothesis function will map input variables to predict some output variable. A general form of the hypothesis function is shown below in Figure 1.

**Figure 1.**

The theta ( components are the parameters for the function that will dictate the shape of it. For every input variable (x) we are trying to map the predicted value. The goal is to generate a straight line through the training data which is represented by .

The accuracy of the hypothesis function can be measured by using a cost function. The cost function takes an average difference of every result from the hypothesis, input variables (x), and the true output values (y). The goal is to get the smallest difference between the predicted value and the true value. The smaller the difference the more confident we can be in our hypothesis function.

A research problem that could benefit from machine learning is the task of classifying atomic force microscopy (AFM) images. AFM is a high resolution imaging technique which allows researchers to gain insight into the external properties of materials (ref). This type of microscopy can be applied to synthetic polymers such as high and low density polyethylene or it could also be applied to biological units such as DNA (ref). It is not uncommon for experts in the field of AFM to debate on the interpretation of the images (ref?). There can also be uncertainty in what the images are actually showing, such as a residual complex that might be a result of the scanning probe (ref). By carefully training a ML algorithm to classify AFM images perhaps some insight can be gained from a non-human interpreter.

Machine learning has become a large field with a variety of disciplines for solving complex problems. The ability to process images and to get meaningful predictions from those models will be an area of interest for this project. There are several online resources for implementing machine learning techniques such as the classification of images. Google’s Tensorflow and SciKit are good examples of some open source software libraries that are freely available to use. Also, there is an 11-week online machine learning course by Dr. Andrew Ng at Stanford University.

**Project Design/Methods**

The approach that this project will take will be to survey the Coursera Machine Learning course by Dr. Ng at Stanford. The ML course is 11 weeks long and will cover basic topics in machine learning and will also cover some more advanced topics such as neural networks. The course will also use the scientific programming language Octave. Octave is comparable to MatLab and is free to use. Currently, we have completed the first week’s lecture material and we are currently in the second week. A general course outline is shown below:

**Course Outline**

Week 1 – Estimated time: 3h 37m

* Introduction
* Linear Regression with One Variable
* Linear Algebra Review

Week 2 – Estimated time: 7h 28m

* Linear Regression with Multiple Valuables
* Octave/Matlab Tutorial

Week 3 – Estimated Time: 6h 2m

* Logistic Regression
* Regularization

Week 4 – Estimated time: 4h 42m

* Neural Networks: Representation

Week 5 – Estimated time: 5h 8m

* Neural Networks Learning

Week 6 – Estimated Time: 6h 7m

* Advice for Applying Machine Learning
* Machine Learning System Design

Week 7 – Estimated time: 4h 57m

* Support Vector Machines

Week 8 – Estimated time: 5h 26m

* Unsupervised Learning
* Dimensionality Reduction

Week 9 – Estimated time: 6h 9m

* Anomaly Detection
* Recommender Systems

Week 10 – Estimated time: 1h 23m

* Large Scale Machine Learning

Week 11 – Estimated time: 1h 16m

* Application Example: Photo OCR

This course can act as a foundation for our capstone to build upon. In addition, the programming language that is primary used in most machine learning tutorials is Python. There are several online tutorials and free books available. It would be prudent to learn some basic Python conditionals, control flow and function syntax. One free resource available is the CodeAcademy Python course. (ref) In addition, there is an online book called Scipy Lecture. (ref) This online book was written to supply beginners with the information that is needed to use Python for science. SciPy has a chapter that goes through the basics of Python all the way through object-oriented programming in Python. A general outline of the chapter is listed below:

1.2 The Python Language

* First steps
* Basic types
* Control Flow
* Defining functions
* Reusing code: scripts and modules
* Input and Output
* Standard Library
* Exception handling in Python
* Object-oriented programming (OOP)