**INFO 6105 (Data Science Engineering methods and tools) : Course Syllabus**

**Instructor: Ramkumar Hariharan, PhD**

*Course will be taught via more than 22 hrs of lively and, star-rated video lectures and other materials !*

***In addition to covering the material below, students are expected to work on a project. There are two options to choose a project — (1) From an instructor provided selection of real-world datasets, OR (2) come up with your own project idea and discuss it with the instructor. All discussions and communications during this online course, will be via email, or zoom.***

Course introduction — structure, operations, components, and content overview. Why and what is machine learning with some examples. The burning hot data science job market, and how this course will help you land a job in data science. Brief introduction to Python, and an overview of Python libraries commonly used by data scientists.

Supervised and unsupervised learning. Classification Vs Regression. Machine learning best practices — train-valid-test split. The K- nearest neighbor algorithm. Data pre-processing.Bias-variance trade off, and how to improve model performance. Measures of classifier performance. Decision trees.

Extending decision trees to Random Forests. Introduction to ensemble models and bagging. Using excel and Python to create a random forest classifier. More about Random Forests. Hyper-parameter tuning, and exploratory data analysis with random forests.

Introduction to artificial neural networks and deep learning — why are they popular, examples, their relation to linear algebra. Different neural network architectures. Basics of linear algebra. Neural networks as successive transformations of the input vector.

Components of neural networks — forward of activations, error calculation, back-propagation of the error gradients, weights updating. Activation functions - logistic, tanh, ReLU, and softmax. Loss functions — binary and categorical cross-entropy. Introduction to Keras, tensor flow and pytorch deep learning libraries. Stochastic gradient descent, learning rate, and the loss function landscape for deep neural networks. Saddle Points.

Convolutional neural network — theory and practice. Using excel to understand convolutions. Transfer learning with convolutional neural networks. Comparison and contrast with standard neural network architecture.

Unsupervised learning — Clustering and discovering structure in the data. Principal Component Analysis or PCA to visualize high dimensional datasets.