**Description**

This course introduces the students to Data Science and Analytics Technologies, Data Analytics at scale, and Data-driven Science systems in order to extract insights from data assuming the students have no prior data science experience. These scientific processes will include various phases and techniques such as Data Preparation, Model Building, and Prediction, Clustering, Association, Regression (Linear and Logistic), Classification, Decision Trees, Textual Data Analysis and Data Presentation. The basic concepts will be covered with examples which can be tried on R or Python by using RStudio and/or Jupyter Notebooks (aka IPython Notebooks). These miniaturized examples of real-world problems are designed in such way that the student will gain a clear understanding and get firm foundation of the methods covered in the course. In addition, the course gives an introduction to R Statistical Language, Apache (Databricks) Spark, and Anaconda Analytics platforms.

**Objective**

After completing this course, the students will have a clear understanding and of data mining and predictive modeling concepts and tools. The student will be able to identify the goals of an analytical project and define and outline the detailed steps for implementation. After mastering the examples given in the course, the student will be in a very good position to apply his/her knowledge to do real-world projects in analytics. Having this introductory exposure to modern big data tools and architecture, such as Hadoop and Spark; students will know when these tools are necessary and will be poised to quickly train up and utilize them in a big data project. Both R and Python have become the de facto languages for Data Science projects. R was built with statistics and data analysis in mind. Although Python is a general purpose language many libraries have been added through packages, primarily for Machine Learning applications. Should you choose R or Python for Data Science? Thanks to recent advances made in both languages, you really can't go wrong with either. I personally favor Python, especially its interactive mode (Jypyter Notebooks, previously known as IPython Notebooks).

**Course Outline**

**Week 01**

* Introduction to Data Science -- Examples of Big Data Analytics (ML & DL)
* Data collection, Data warehousing, data modeling
* Sign-up with Anaconda-Cloud. Download Anaconda 4.2.x (Python 3.x version)

**Week 02, 03 and 04**

* Statistics for data analysis
* Introduction to statistical concepts
* Mean, Median, Mode, Ratio
* Probability Theory, Probability Distributions
* Hypothesis testing, regression analysis, A/B testing
* Statistical modeling and fitting
* Causation vs. correlation

**Week 05**

* Introduction to MS Excel
* Basic Excel functionality

**Week 06 and 07**

* Introduction to Relational Database
* SQL querying, syntax,
* Select, Where, Insert, Delete, Joins
* Create DB, Drop DB, Create Table, Drop Table
* Primary Key, Foreign Key
* Views, Injection, Hosting

**Week 08 and 09**

* Fundamental of python as OOP
* Introduction to Pypl

**Week 10 and 11**

* Data Analytics Lifecycle, Data Science Workflows
* Introduction to MapReduce (MR1) and Hadoop
* Introduction to Data Types, Time Series Data, Longitudinal Data, Unstructured Text Data
* Data Exploration in Python (Data Wrangling): NumPy, SciPy, Pandas
* Introduction to Python’s Visualization: Matplotlib, Plotly, Seaborn, Bokeh

**Week 12**

* Introduction to R (RStudio IDE)
* Semiology of Graphics (James Bertin) -- R Graphics for Data Display (ggplot2 library)
* Introduction to Anaconda Open Data Science platform

**Week 13**

* Data Analytics at Scale / Streaming Analytics
* Introduction to Apache Spark (MR2), Scala (Functional) Programming Language
* Introduction to Spark’s MLlib and Sparklyr (dplyr backend)

**Week 14**

* Machine Learning, Supervised Learning Techniques
* Naïve Bayes Algorithm
* Introduction to scikit-learn Python library

**Week 15**

• Project one

**Week 16**

* Introduction to Linear Regression (LR) & Prediction using LR
* Machine Learning Model Validation, Regularization, Overfitting, H-Parameter Tuning

**Week 17**

* Introduction to Logistic Regression & Scoring with Logistic Regression
* Risk Models / Response Models
* Introduction to Support Vector Machine (SVM) and Singular Value Decomposition
* (SVD)

**Week 18**

* Advanced Analytical Theory and Methods: Classification
* Introduction to Decision Trees Modeling (CART, Boosted Trees)
* Decision Tree Algorithms, Bagging Predictors, the XGBoost Algorithm



**Week 19**

* Unsupervised Learning -- Cluster Analysis (k-Means)
* The Curse of Dimensionality
* Principal Component Analysis (PCA) : Eigenvectors, Eigenvalues and Dimension Reduction

**Week 20**

* Natural Language Processing (NLP)
* Introduction to TF-IDF (Term Frequency - Inverse Document Frequency) statistical model
* Introduction to Word2vec and GloVe (Python) libraries/models

**Week 21**

* Introduction to Deep Learning
* Biological Neural Networks vs Artificial Neural Networks
* Neural Networks – Perceptron Learning
* Introduction to TensorFlow, Keras (Python)

**Week 22**

* Computer Vision – Convolutional Neural Networks (CNNs)
* Recurrent Neural Networks (RNNs) and LSTNs (Long Short-Term Networks)
* Introduction to TensorFlow, Keras (Python)

**Week 23**

• Final Project

**Week 24**

* Job preparation, Resume building, job searching, interview technique