INFO I590 Data Science Onramp Advanced
(1-3 credit hours)

Data Science Onramp contains mini courses with the goal to build and enhance your data science skills which are oftentimes demanded or desired in data science related jobs. Each mini course will be counted as one credit hour. Each time you enroll, you can select 1-3 credit hours which means that you can select 1 or 2 or 3 mini courses. You are allowed to maximally enroll twice for this whole course.

Most of the mini courses are written in text format and few in video format. We provide Teaching Assistant (TA) support and office hours. If you encounter any problems, please feel free to reach our TAs either at their office hours or schedule an appointment which fits better to your schedule. You can learn your selected mini courses in sequence or parallel. But we HIGHLY recommend parallel learning because: 1) you can participate the online discussion with other classmates; 2) our TAs will have weekly office hours and monthly live demos based on weekly and monthly contents of the mini courses; and 3) you have a good reason to get your assignment done on time rather than rush to finish them before the end of the semester.

Each mini course has its own grading policy. In general, grading is based on assignments/projects, online discussions, and quizzes. If you select more than one mini course, the average of each mini course grade will be counted towards your final grade.

**Introduction to Spark**

Through this online course, we will introduce you what Apache Spark is, how it can be helpful, and where its power resides. The course is designed to be simple, to the point and instructive for the beginners. We will not be surprised to see many students who has already tried other online tutorials or coerces about Apache Spark, but very soon has found the concepts very confusing. However, here we understand this fact and it is number one priority to express all key concepts in a very straightforward language and try to avoid unnecessary and confusing fancy statements. Additionally, our preference has been to use real world examples to make sure that students actually can imagine how the skills will be helpful in a real-world setting. We want to provide you some hands-on experience by developing simple programs that can be easily deployed in many other situations only by being modified slightly. Moreover, we have tried to make the course easy to proceed by covering every basic concept and skill you need to develop Spark programs so you do not need to look for other resources frequently while taking the course.

**Course Structure:**
Introduction to Apache Spark

Apache spark components: Spark Core, Spark SQL, Spark Streaming, Spark MLLib

Installation of Apache Spark

Writing your first spark application

Resilient Distributed Datasets (RDD) in Spark

Data partitioning in Spark

Importing and exporting data into Spark

Accumulators and Broadcast variables

Spark interaction with R

Introduction to Spark SQL

Basics of Scala

Scala is a very fancy and new programming language. It is pretty popular especially in industry in the recent years. As a functional programming language, it is kind of similar to Java but with more flexibility. It can even run on JVM (Java virtual machine). This course was designed to get you familiar with Scala constructs and features. This course doesn't require any prerequisites but students should have a basic understanding of object-oriented programming. This course uses a data-centric approach to Scala. All content in this course is standard basics in Scala. If you can follow each session closely, you are guaranteed to get some useful knowledge about Scala at the end. And you are able to use Scala to solve some real-world problems.

Course Structure:

- Basic background of Scala
- Install Scala in your local environment
- Create a project in Scala IDE
- Scala REPL to run code in terminal
- OOP in Scala
- Write methods in Scala
- What is object in Scala
- Scala-particular basic concepts such as access modifiers and companion objects
- What are case object and case class
- Some synthetic methods
- Collections in Scala
- Sequences and sets in Scala
- Tuple and map in Scala
- Higher order functions in Scala

Introduction to Hadoop Framework

Unlike many of the online articles that you may have already seen, here we do not want to talk about how you can improve your resume by acquiring Hadoop MapReduce
knowledge and skills, nor do we want to emphasize the importance of Hadoop and MapReduce to the information technology industry, etc. We know that you already understand how important it is from different aspects, in fact that is probably why you are taking this course.

Our goal in this course is trying to teach you some practical skills so you can actually do something cool using Hadoop, like developing a program to rank some documents based on their relevance to a search query. We will start the course in the form of questions and answers, that is we assume that you have already faced with some questions when wanted to learn about Hadoop and MapReduce by yourself, but never found a clear answer for them. Then we will proceed by introducing different aspects of MapReduce and other systems designed on top of Hadoop. Throughout the course, we will make sure that you get hands on experience by developing simple programs to work on real-world data and scenarios. Moreover, we have tried to make the course easy to proceed by covering every basic concept and skill you need to develop Hadoop and MapReduce programs so you do not to look for other resources frequently while taking the course.

Course Structure:

- Basics of MapReduce
- Developing MapReduce programs in Java
- Installing Hadoop on your computer and running your first Hadoop program
- HDFS (Distributed File storage systems) and Yarn concepts
- MapReduce application development and configuration
- MapReduce Job architecture
- Inverted indexing technique for text retrieval
- Graph processing in Hadoop
- Analyzing stack exchange posts dataset using Hadoop
- Introduction to Apache HBase
- Writing MapReduce jobs on HBase
- Introduction to Apache Hive
- Analyzing Stack exchange dataset using Hive
- Final project-Implementing Pagerank algorithm using MapReduce

**Machine Learning with Spark**

Through this online course, we will introduce you how to do Machine Learning on large scale using Apache Spark. The course is designed to be simple, to the point and instructive for the beginners in Spark. We hope you enjoyed the "Introduction to Spark course" which is a prerequisite for the "Machine Learning with spark" course.

The "Machine Learning with spark" course starts with introduction to *Linear Algebra* and *Python in Spark* to brush-up your skills. The course discusses the MLlib which is Spark’s scalable machine learning library consisting of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, as well as underlying optimization primitives. The course ends
with topics like Text Mining, building a machine learning project pipeline and a final project. Our preference has been to use real world examples to make sure that students can imagine how the skills will be helpful in a real-world setting. We want to give you some hands-on experience by developing simple programs that can be easily deployed in many other situations only by being modified slightly. Moreover, we have tried to make the course easy to proceed by covering every basic concept and skill you need to develop your Machine Learning models in Spark.

Course Structure:

- Introduction to Linear Algebra
- Introduction of Python for spark
- Developing word count application of large data set using Spark
- Decision trees implementation in Spark
- Linear regression
- Logistic regression
- Unified view on Linear methods
- Unsupervised machine learning: Clustering
- Text analysis using Spark RDD
- Frequent patterns and occurrences in Spark
- Machine learning pipelines

Kaggle Cases

In this course we will focus on classic workflow of taking kaggle competitions. We will discuss three introductory Kaggle competitions. They are tasks about regression, binary classification and multiclass classification. We will get through all the necessary steps to complete these competitions, namely exploring and preprocessing data, constructing, tuning and evaluating models. Specifically, we will mainly demonstrate and discuss the relevant algorithms and techniques about missing value imputation, feature encoding and selection, linear regression, logistic regression, One-Vs-The-Rest, One-Vs-One, softmax regression, K-nearest neighbors, RBF regression, ridge and lasso regularization, K-fold cross validation and ensemble methods such as random forest and adaboost, etc. All the models and techniques learned in class to solve competitions will be implemented in Python, with the help of popular Python packages Jupyter notebook, scikit-learn and pandas.

Course Structure:

- Introduction to Kaggle
- Basic Knowledge Review_Part 1
- Basic Knowledge Review_Part 2
Deep Learning Principles

People who have some knowledge of machine learning and want to add deep learning to their arsenal are encouraged to take this class. While a machine learning class if not a hard prerequisite, knowing some general practical machine learning principles like regularization, validation sets, etc. will go a long way in helping you utilize the course to its maximum potential. But if you do not have a lot of machine learning experience, but are comfortable with coding in python and have some working knowledge of very basic linear algebra and high school calculus, you are welcome too. Many machine learning principles have been introduced from scratch but it is expected that you will learn the ones which haven't been dealt with in great detail. An introductory course like Machine Learning Principles will be very helpful before taking this class.

People who have never done any machine learning or aren't comfortable with programming in Python or aren't familiar with high school calculus and basic linear algebra shouldn't take this class. Finally, this is neither a completely theoretical course nor a hands-on recipe for implementing deep learning. If you want either of the two extremes, this course is not for you. It will try to strike a balance by first focusing on enough theory and then slowly build on more practical stuff.

You will learn about the following from this course:

- Feed-forward Neural Networks
- Deep Neural Networks
- Convolutional neural networks
- TensorFlow
- Keras

You will also develop a few interesting applications like handwritten digit recognition system in this course.
Course Structure:

- Machine Learning Primer
- Neurons – Introduction
- Neurons – Learning
- Neural Networks
- Neural Networks in Practice
- Deep Networks
- Practical issues in deep learning
- Convolutional Neural Networks
- Recurrent Neural Networks