INTRODUCTORY DATA SCIENCE AND ANALYTICS

OVERVIEW

This workshop will aim to provide an introduction to the hot field of data science. According to a McKinsey study, by 2018, the US alone can face a shortage of ~150K people with deep analytical skills. This workshop will give a brief overview of the field and its future and explain core algorithms and techniques used to understand big data. By the end of this workshop, attendees can expect to take away techniques that will help them to look at datasets like a data scientist and glean insights from real-life data that will help answer strategic business questions.

SYLLABUS:

1. **Data Science 101**
   What is all the buzz about data science? Organizations these days are overwhelmed with the data they collect and are able to obtain from varied sources. This lecture will give newcomers, a quick overview of the essentials of this field. There are 4 key aspects of a typical data science project: data wrangling (cleaning and curating to get a sample data set), exploratory data analysis (understanding what is in your data), modeling & predicting using the data, communication: creating visualizations and dashboards to communicate results from the models. We will go over these 4 aspects and help attendees to feel comfortable transitioning from one phase to another.

2. **Data Munging with Python:**
   We will start our journey into the field of data science with the basics on how to use Python to clean messy and incomplete datasets. This is the crucial first step before any kind of analysis can be done. We will focus on some best practices and libraries in python that can be used to get our data into a usable format.

3. **Exploratory Data Analysis and Visualization:**
   The next key aspect in becoming a good data scientist is to understand your data from different angles. Are there correlation effects, what are the general trends the data is suggesting, how are variables related to each other? – These are some of the questions we will try to answer so we can build accurate models that reflect reality. The lab part of this lecture will introduce students to some basic packages to visualize their data quickly.

4. **Introduction to Machine Learning**
   This session will be an overview of the current state of Machine Learning in industry with case studies from some of the leading tech companies. We will learn some basic algorithms in machine learning and also understand how they were used in industrial situations to deliver business insights. Lectures will be followed by hands-on workshops
where students will learn how to apply the algorithms they learnt to actual data sets from industry and government.

5. **Communication of results:**
   A key skill to be a successful data scientist is to be able to explain the outcomes of a model to a non-technical audience. The value of the model we build is only as good as how well we explain it to the business stakeholders in an organization. This lecture and lab series will go over some best practices in how to communicate the results of the models we built. We will build interactive dashboards that will convey the key insights we have developed.

**PROJECT TOPICS:**

1. Data science in AdTech: Multi Touch Attribution- assigning the right credit to an advertisement in driving a conversion
2. Social Media Analytics: Finding your spheres of influence
3. Data science in HealthCare: Predicting prominent risk factors for cardio-vascular diseases
4. Finance Data Science: Predicting defaulting customers – credit risk

**PRE-REQUISITES:**

1. **Basic statistics and probability:** Undergrad level of proficiency.
2. **Introductory Python:** Comfortable in the python and ipython environment. How to install and import packages
3. **Introductory Git:** Git is a version control software. We will commit the code we write to a common repository so it is available for everyone in the workshop to see. Install git and understand the basics.

**HARDWARE & SOFTWARE REQUIREMENTS:**

1. Python 2.x or above
2. Git
3. Access to AWS s3 account (optional)

**KEY LEARNING OUTCOMES FROM WORKSHOP:**

1. Master the 4 key aspects of data science: (1) Data Wrangling (2) Exploratory Data Analysis (3) Predictive Modeling (4) Communication of results
2. Develop comfortable proficiency to be able to tackle some open data challenges in Kaggle or other open datasets
3. Apply project based learnings to a real life dataset and deliver key business insights based on the models we build.