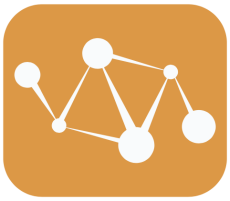

**DataScienceWorks**
**Training Institute**

# All of Data Science

Available from **Feb 2020**

| | | | |
|---|---|---|---|
| **Instructor:** | Dr. Nayyar Zaidi | **Time/Place:** | TBA |
| | | **Duration:** | 4-6 Days |
| | | **Price:** | Ask for the quote |
| | | **Contact:** | info@datascience-works.com |

## Description:

This is one of the most comprehensive yet succinct training course that provides you with a complete picture of Data Science and Machine Learning. The course is combination of theory and practice, and is up-to-date with the latest research and trends in Data Science. For example, it covers topics such as Representation Learning (new emerging field with-in Machine Learning), Deep Learning, Feature Engineering (secret sauce behind all practical and effective algorithms), Data Engineering and Machine Learning Engineering (marriage of Big Data world of databases with Machine Learning world of Analytics and models).

Things have changed and are changing very quickly in the world of Data Science and Machine Learning. E.g., last few years have seen the rise of deep learning and representation learning. And, then there has been the rise of Big Data and Big Data processing platforms such as Hadoop and Spark. On the other hand, we are witnessing a tremendous progress in typical Artificial Intelligence tasks due to emergence and success of deep Reinforcement Learning. The goal of this training is to bring you to speed and get you acquainted with foundations, advancements as well as future trends in the field of Data Science.

This course is designed to provide a full view of Data Science landscape. If you are a seasoned data scientists or machine learning practitioner, this course will serve as an excellent refresher course to bring you up to date with current research and trends in the field. If you are an engineer, computer scientist, business analyst, database specialist, or related professional - this course provides you with an excellent start in the world of data science.

The course builds from the fundamentals and provides contents in sequential manner. It starts by introducing mathematical and programming fundamentals. It later covers traditional Machine Learning topics and then dive into Deep Learning. Finally, it introduces Machine Learning Engineering and concludes with advance topics in Machine Learning.

## Training Learning Outcome (TLO):

After the training, you are expected to:

- Have an excellent understanding of various topics in Data Science.
- Be able to comprehend, communicate, disseminate complex Data Science concepts.
- Have a basic to expert understanding of inner workings of various Data Science algorithms.
- Be able to scope a Data Science project.

## Target Audience:

The training is for any aspiring or seasoned data scientists, and is perfect for:

- Computer Scientists and I.T Professionals,
- Engineers (Electrical, Mechanical, Industrial, etc),
- First year Ph.D. students in any field looking to break in Data Science,
- Post-doc fellows and Early Career Researchers in any field.

## Duration:

The course is expected to be delivered in 5 days (9-5pm), but the duration can be adjusted based on audience experience and background.

## Outline:

The following outline is tentative, and can be customized based on audience demand. The electives can be chosen and replaced from Table 2.

| | Day 1 | Day 2 | Day 3 | Day 4 | Day 5 |
|---|---|---|---|---|---|
| **Session 1** | Introduction | Model Selection | Unsupervised Machine Learning | Big Data Engineering | Large Scale ML Platforms |
| **Session 2** | Maths and Programming Fundamentals | Feature Engineering | Deep Learning Lab | Machine Learning Engineering I | Graph Analytics |
| **Session 3** | Supervised Machine Learning | Supervised Machine Learning Lab | Representation Learning | Machine Learning Engineering II | Reinforcement Learning |
| **Session 4** | Python & Data Wrangling Lab | Deep Learning I | Deep Learning II | PySpark Lab | Further Topics in Data Science |

Table 1: 5 Days – Training Outline.

| | | | | | |
|---|---|---|---|---|---|
| **Electives** | Data Generation | Data Visualization | Data Science for Managers | Recommender Systems | Audience Presentations |

Table 2: Other Electives.

Let us delve deep into the details (outline of topics covered) of each session in the following. (E) denotes an elective session.

---

**Introduction**

- Machine learning, Artificial Intelligence, Statistics, Data Mining and More
- Machine learning applications

- Introduction to Data Science and Big Data
- Ingredients of Machine Learning – Data, Model and Process
- Training your first practical model

---

**Programming Fundamentals (E)**

- Programming Languages
- Data Structures
- Algorithms and Complexity Analysis
- Databases
  - SQL
  - Towards Data Warehouses
- Introduction to Python
  - Lists, Dictionaries, Sets, Tuples
  - Numpy
  - Pandas
  - Scipy, Matplotlib, etc.

---

**Mathematics Fundamentals (E)**

- Linear Algebra
- Probability and Statistics
- Calculus
- Linear Algebra Re-visted
  - Gradients, Hessians
  - Local/Global Minima/Maxima of a function
- Optimization

---

**Supervised Machine Learning**

- Regression
  - Linear Regression, Polynomial Regression
- Classification
  - Logistic Regression
  - LDA/QDA
  - Naive Bayes, Decision Trees
  - Nearest Neighbour Methods
  - Generative vs. Discriminative Learning
- Forecasting - Time Series Analysis
  - Seasonality and Trends
  - Moving Averages
  - Holt-Winters Method Family
  - ARIMA

---

**Model Selection**

- Bias and Variance Analysis
  - Achieving Low-variance
  - Regularization
  - Feature Selection
- Achieving Low-bias
  - Feature Construction

  - Kernels and Kernel trick
  - Feature Engineering
- Evaluating and Comparing Models
  - Cross-valiation
  - Lift Charts, ROC, RPC, other metrics
  - Statistical Tests, Null-Hypothesis, Friedman Statistics, etc.

---

**Feature Engineering**

- Ensemble-based Methods
  - Bagging, Stacking
  - Random Forest
  - Gradient Boosting
  - XGBoost
- Extreme Non-linear Models
  - Generalized Linear Models
  - Factorization Machines
  - Support Vector Machines
  - Artificial Neural Networks
- Need for Deep Learning

---

**Artificial Neural Networks**

- Foundations of Deep Learning
  - Backpropagation
  - Gradient Computation
  - Optimization
  - Gradient Vanishing/Exploding
  - Model Architecture
  - Batch Normalization
  - Error Surfaces of ANN
  - Miscellaneous Issues
- Why Deep Learning Now?

---

**Unsupervised Machine Learning**

- Clustering
  - K-means, DB-Scan, Hierarchical
- Density Estimation
- Bayesian Networks
- EM Algorithms for Clustering and Gaussian Mixture Models
- Curse of Dimensionality
- Similarity Measurements
  - Exact vs. Approximate Measures
- Local Sensitive Hashing
  - Data Pre-processing
  - Dimensionality Reduction
  - Data Standardization
  - Data Munging
  - Feature Hashing
- Overview of anomaly detection
- Association rules and discovery

  - APriori Algorithm

---

**Representation Learning**

- Introduction
  - Principal Component Analysis (PCA)
  - Singular-value-Decomposition (SVD)
  - First look at Embeddings
- Unsupervised ANN
  - Restricted Boltzmann Machines (RBM)
  - Greedy Layer-wise Training
  - Contrastive Divergence
- Auto-Encoders (AE)
  - Encoder/Decoder Architecture
  - Stacked Auto-Encoders
  - Sparse Auto-Encoders
- Embeddings
  - Word2Vec (Skipgram, CBOW), Node2Vec, etc.
- Transfer Learning
- Tensor Decomposition

---

**Deep Learning I**

- Introduction
  - Convolution
  - Feature Maps, Max Pooling
- CNN Architectures
  - LeNet-5, AlexNet, VGG-16, ResNet
  - Transfer Learning Revisited
- Object Detection
  - Harnessing the value of Covolution for Object Detection
- Face Learning
  - Face Verification
    - Triplet Loss
    - Siamese Network
  - FaceNet and DeepFace
- CNN for non-images
- 1-d, 3-d and 4-d Convolutions

---

**Deep Learning II**

- Introduction
  - Various Architectures
  - Applications
- RNN Embeddings
- LSTM, GRU
- Attention-based models

---

**Data Generation (E)**

- Introduction
  - Importance/Rejection Sampling

  - Gibbs Sampling and MCMC
- Pixel CNN and Pixel RNN
- Variational Auto-Encoders (VAE)
- Generative Adversarial Networks (GAN)
  - Conditional GAN
  - Applications
- Adversarial Learning
  - Defence Methods
  - Attack Methods

---

**Big Data Engineering**

- Databases - Management, Query, Indexing, Control
- Relational Algebra
- Brief overview of Business Intelligence
  - Data Cubes, Marts
  - Data Warehouses
- Introduction to Big Data
  - SQL vs. NoSQL
- Parallel Query Processing
- Google Big Table, GFS, HDFS, Hadoop
- Hadoop Echosystem
- NoSQL Databases
  - Hbase
  - MongoDB

---

**Large Scale ML Platforms**

- Introduction Apache Spark
- Spark Internals
  - Diving Deep into Spark
  - Spark runtime and application architecture
  - Spark RDDs
  - Spark as distributed file systems
- Spark SQL architecture
  - Spark SQL
- Data Processing with Spark
- Introduction to Yarn
- Introduction to Streaming Architecture
  - Kafka

---

**Machine Learning Engineering I**

- Introduction
- Cloud Architecture
  - Cloud Computing
  - Available Resources: AWS, Google Cloud, etc.
- Enviroments
  - Docker
  - Kubernetes
- Model Training
  - TF Serving

  - Training Models on GPUs
  - Training Models across Multiple Devices
- Model Deployment

---

**Machine Learning Engineering II (E)**

- Introduction
- Building Models
- Bad Data
- Evaluating Models
- Error Analysis
- Handling Distribution Shift
- Analysis in Production

---

**Reinforcement Learning (E)**

- Introduction
- Markov Decision Process and RL
- Introduction to Dynamic Programming
- Q-Learning
  - Deep Q-Learning
  - Double-deep Q-Learning
- Policy-gradient Methods
  - REINFORCE
  - Variance-control Methods
- Actor-Critic Models

---

**Recommender Systems (E)**

- Recommender Systems – An Introduction
  - Evaluating Recommendations
  - Data Structure of Recommender Systems
- Content-based Filtering
  - Neighbourhood-based
  - Similarity-based
  - Model-based
- Collaborative Filtering
  - Baseline Estimates
  - Neighbourhood-based
  - Model-based
- Singular Value Decomposition (SVD)
  - Matrix Factorization
  - Constrained Matrix Factorization (Non-negative Matrix Factorization)
  - SVD++
  - Collaborative Filtering meets Content-based Filtering
- Advertising on the Web
  - Ad Placement
  - Multi-arm Bandits
  - Contextual Multi-arm Bandits

---

**Further Topics in Machine Learning (E)**

- Bayesian Machine Learning
- Text Mining
  - Representing Texts (TF-IDFs)
  - Name Entity Recognition
  - Topic Models (pLSA)
- Learning to Rank
- Causality
  - Randomized Clinical Trials
  - AB Testing from Software Engineering Perspective

---

**Graph Analytics (E)**

- Introduction to Graph
  - DFS, BFS and other basic graph algorithms
- Graph Embeddings
  - Node2Vec
  - LINE
  - Graph Factorization
  - Graph Convolution Network
- Typical graph analytics problems and their solutions

---

**Data Science for Managers (E)**

- Running and Managing a Data Science Project
- Explainable AI
- Ethical ML Models
- Data Privacy Issues

---

**Data Visualization (E)**

- Introduction
- Data Types and Visual Mappings
- Data Combinations and Dimensions
- Effective Visual Encoding
- On Visual Perception
- Evaluating Visualizations

---

**Audience Presentations Analytics (E)**

- 3-5 minutes presentation by member of the audience on their previous data science project
- Discussion on how the lessons learned from this training can help improve that project

---

## About the Instructor:

Dr. Nayyar Zaidi is the lead Data Scientist at DataScienceWorks and a Senior Lecturer of Computer Science at Deakin University. He received the B.S. degree in computer science and engineering from the University of Engineering and Technology, Lahore, in 2005, and the Ph.D. degree in Artificial Intelligence from Monash University, Melbourne, Australia, in 2011. He worked as a Research Fellow, a Lecturer, and a Research Fellow, from 2011 to 2013, from 2013 to 2014, and from 2014 to 2017, respectively, at the Faculty of Information Technology, Monash University. From 2017 to 2019, he worked as Research Scientist at Credit AI (Trusting

Social) Melbourne Lab. His research interests include feature engineering, data generation explainable models and ethical AI. He is also interested in practical (applied) data science, machine learning engineering, and data science training.

## Frequently Asked Questions:

### Why should I attend this course?

Good question! Let us ask you some counter questions:

- 1) Are you interested in exploring the breadth and depth of Data Science?
- 2) Are you curious about the inner workings of Machine Learning algorithms?
- 3) Interested in Data Science but do not know where to start? Or overwhelmed by so much information?
- 4) Would you like to bring yourself to what is state of the art in analytics?
- 5) Would you like to become a pure Data Scientist?
- 6) Would you like to become a full-stack Machine Learning Engineer? If answer to any of the above questions is yes, then you should attend this course.

### How is this course different from others?

This course covers the same amount of material which generally 3 to 4 under-graduate or post-graduate units contain. This course is intentionally designed to be exhaustive to provide the depth and breadth of data science and machine learning spectrum. In our opinion, not many courses outside academia has this much information delivered in short amount of time.

### Do I need a deep Math background?

Not deep. But you should be familiar with some basic Statistics, Linear Algebra and Calculus. Note, there is also an introductory lecture on Mathematical Fundamentals.

### Do I need to bring my laptop?

Yes, there are lab components in most of the lectures as well as dedicated labs. It will be beneficial if you bring your laptop to run code and exercises on your computer.

### Do I need to know computer programming?

Yes. Some computer programming experience is required if you are interested in implementing some of the ideas in the workshop.

### I am not a Python programmer?

There is a module that covers introduction to Python.

## About this Document: