magnitia
IT's next generation

# Data Science with Python

## Introduction

- What is data science?
- Where do we apply data science?
- Why is data science a "science"?
- What makes up data science?
- What is the difference between data science, machine learning and AI?
- Data Science vs Data Analytics vs Big Data
- Why do we need to understand the difference?
- The business of data science
- Additional details
  - Questions that a data scientist must ask
  - Is your data ready for data science?
  - Ask the right questions
  - Pre-requisites and the process

## Data Science from scratch

- Tools required (pic)
- Algorithms that can be used

## Some basic python

- Numpy and scipy
  - Data objects
  - Math
  - Comparison Operators
  - Condition Statements
  - Loops
  - Lists
  - Tuples
  - Sets
  - Dictionaries
  - Functions
  - Array
  - Selecting Data
  - Slicing
  - Iterating
  - Manipulations
  - Stacking
  - Splitting Arrays

- Pandas
  - Pandas overview
  - Series and Data Frame
  - Manipulation
- Scikit learn

## Visualizing data – bar plots, line charts, scatterplots, histograms – Matplotlib or seaborn

## Linear Algebra

- Vectors
- Matrices

## Statistics

- Central Tendencies, dispersion, Mean, Median and Mode, Data Variability, Standard Deviation, Z-Score, Outliers
- Correlation & Causation
- sampling

## Probability

- Dependence & Independence
- Conditional Probability – CDF's and PDF's
- Bayes Theorem
- Random Variables
- Continuous distributions, Normal Distributions
- Central Limit Theorem
- Skewness & Curtosis

## Hypotheses and Inferences

- Hypothesis testing: Null hypothesis, p-values
- Confidence intervals
- Type Errors
- A/B testing, T-Tests, ANOVA test
- Bayesian Inferences

## Gradient Descent

- Estimation
- Usage
- Stochastic Gradient Descent

## Getting Data

- Reading files
- De-limited files

- Scrapping the web and using APIs
- Authentication of connection, APIs

## Data Munging/cleaning

- Data exploration: 1-d, 2-d & multi-d data
  - Grouping
  - Aggregation
  - Treating Missing Values
  - Removing Duplicates
  - Cleaning data
  - Munging data
  - Manipulation of data
  - Scaling data
  - Dimensionality reduction
  - Transforming data

## What is "Machine Learning"?

- Modelling
- Overfitting of data
- Underfitting of data
- Correctness of data
- Bias and variance – trade off
- Feature extraction and selection

## Algorithm (s)

- K-Nearest Neighbours
- Naïve Bayes
- Linear Regression
- Multi Regression
- Logistic Regression and SVM's
- Decision Trees
  - Entropy
  - Random Forests

## Other directions:

- Gradient Boosting
- xg-boost
- PCA
- Deep Learning and GANs
- Semi-supervised learning
- Active learning

## Neural Nets

- Perceptron
- Feed-forward NNs
- Back-propagation
- CNN
- RNN

## Clustering

## NLP – Natural Language Processing

- Lexical processing, syntactic processing, semantic processing
- Word clouds
- N-gram models
- Grammars
- Gibbs Sampling
- Topic modelling

## Recommender Systems

- Trending Recommendations
- User based Collaborative filtering
- Item based collaborative filtering

## DB and SQL Overview

- Table creation and insert
- Update
- Delete
- Select
- Group by
- Order by
- Join
- Sub-queries
- Indices
- Query Optimization
- No-SQL

## MapReduce Overview

- Why MapR
- General applications

## Tableau Overview

- Overview of Tableau
- Connect to Data and make visualizations
- Features of Tableau

## Azure ML

- What is Azure ML?
- What can I do with Azure ML?
- The Azure ML process and provisions
- Azure ML components and API's

## What is Time Series?

- Trend, Seasonality, cyclical and random

- White Noise
- Auto Regressive Model (AR)
- Moving Average Model (MA)
- ARMA Model
- Stationarity of Time Series
- ARIMA Model – Prediction Concepts
- ARIMA Model Hands on with Python
- Case Study Assignment on ARIMA

**Plan & Version Control Tools :** Jira, Git

**IDE**: Pycharm

**Distribution:** Anaconda (The Anaconda distribution would allow us to implicitly install python 3x, R along with scikit-learn, numpy, scipy and pandas. Anaconda includes jupyter and spyder as part of the package).

**Platforms :** Windows, Linux