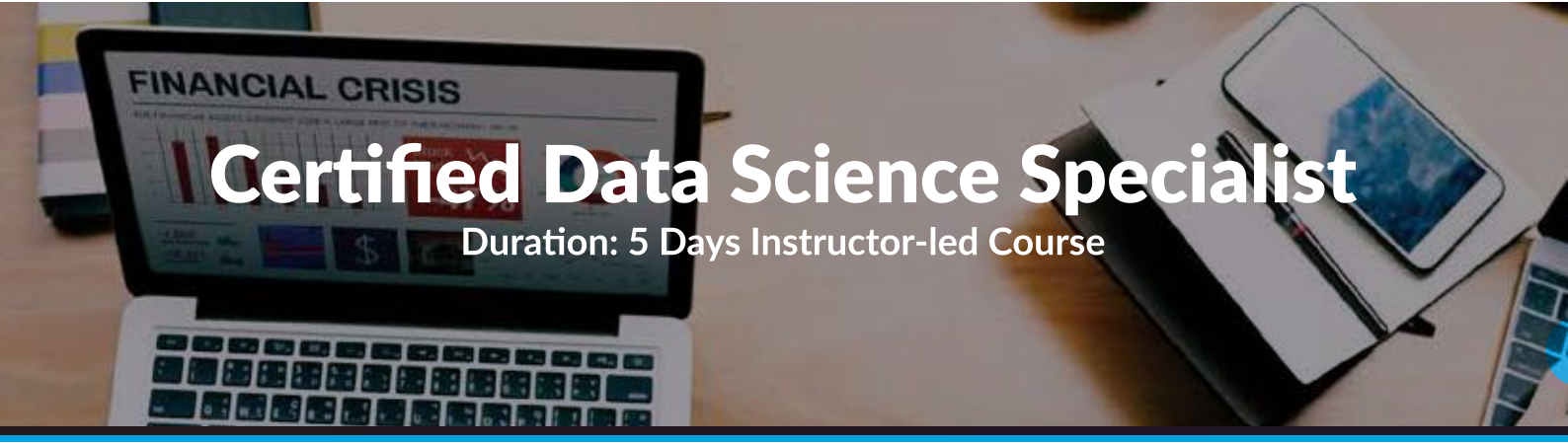
CYBERPORT ACADEMY
Business Unit Of MSC Cyberport Sdn. Bhd.

DATA SCIENCE

BIG DATA IS THE FOUNDATION OF ALL CURRENT MEGATRENDS, FROM SOCIAL TO MOBILE TO THE CLOUD TO GAMING

The Future belongs to people who can't transform data into strategic Business decisions and value-driven products. Data science is a field of Big Data Which seeks to provide meaningful information from large amounts of complex data using various tools, alogrithms and machine learning principles

Worldwide revenues for big data and buisness analytics will grow from $130.1 billion in 2016 to more than $203 billion in 2020, at a compound annual growth rate (CAGR) of 11.7%. **Source: IDC**



Certified Data Science Specialist

Duration: 5 Days Instructor-led Course

Course Overview

Our lives are flooded by large amounts of information, but not all of them are useful data. Therefore it is essential for us to learn how to apply data science to every aspect of our daily life from personal finances, reading and lifestyle habits, to making informed business decisions. In this course you will learn how to leverage on data to ease life,or unlock new economic value for a business.

This course is a hands-on guided course for you to learn the concepts, tools, and techniques that you need to begin learning data science. We will cover the key topics from data science to big data, and the processes of gathering, cleaning and handling data. This course has a good balance of theory and practical applications, and key concepts are taught using case study references.

Upon completion, participants will be able to perform basic data handling tasks, collect and analyze data, and present them using industry standard tools.

Prerequisites

All participants should have basic understanding of data,relations, and basic knowledge of mathematics.

Who Should Attend

This workshop is intended for individuals who are interested in learning data science,or who want to begin their career as a data scientist.

Exam Format

The CDSS Certification Exam duration is 2 hours, consisting of 50 Multiple Choice Questions, with a Passing Score of 70%. You will receive a professional CDSS Certification upon passing the exam.

Learning Outcomes

Upon completion of this course, you will be able to:

- Identify the appropriate model for different data types
- Create your own data process and analysis workflow
- Define and explain the key concepts and models relevant to data science.
- Differentiate key data ETL process, from cleaning, processing to visualization.
- Implement algorithms to extract information from dataset.
- Apply best practices in data science, and become familiar with standard tools.

Course Outline

Day 1

Introduction to Data Science

- What is Data?
- Types of Data
- What is Data Science?
- Knowledge Check
- Lab Activity

Data Science Workflow

- Data Gathering
- Data Preparation & Cleansing
- Data Analysis - Descriptive, Predictive, and Prescriptive
- Data Visualization and Model Deployment
- Knowledge Check

Life of a data scientist

- What is a Data Scientist?
- Data Scientist Roles
- What does a Data Scientist Look Like?
- T-Shaped Skillset
- Data Scientist Roadmap
- Data Scientist Education Framework
- Thinking like a Data Scientist
- Knowns and Unknowns
- Demand and Opportunity
- Labor Market
- Applications of Data Science
- Data Science Principles
- Data-Driven Organization
- Developing Data Products
- Knowledge Check

Data Gathering

- Obtain data from online repositories
- Import data from local file formats (json, xml)
- Import data using Web API
- Scrape website for data
- Knowledge check

Day 2

Data Science Prerequisites

- Probability and Statistics
- Linear Algebra
- Calculus
- Combinatorics

Beginning Databases

- Types of Databases
- Relational Databases
- NoSQL
- Hybrid database
- Knowledge Check Lab activity

Structured Query Language (SQL)

- Performing CRUD (Create, Retrieve,Update, Delete)
- Designing a Real world database
- Normalizing a table
- Knowledge Check Lab Activity

Introduction to Python

- Basics of Python language
- Functions and packages
- Python lists
- Functional programming in Python
- Numpy and Scipy
- iPython
- Knowledge check
- Lab Activity
- Lab: Exploring data using Python

Course Outline

Day 3

Data Preparation and Cleansing

- Extract, Transform and Load (ETL) - Pentaho, Talend, etc
- Data Cleansing with OpenRefine
- Aggregation, Filtering, Sorting, Joining
- Knowledge Check Lab Activity

Introduction to R

- Packages for data import, wrangling, and visualization
- Conditionals and Control Flow
- Loops and Functions
- Knowledge check
- Lab activity
- Lab: Exploring data using R

Exploratory Data Analysis (Descriptive)

- What is EDA?
- Goals of EDA
- The role of graphics
- Handling outliers
- Dimension reduction

Data Quality

- Raw vs Tidy Data
- Key Features of Data Quality
- Maintenance of Data Quality
- Data Profiling
- Data Completeness and Consistency

Day 4

Machine Learning (Predictive)

- Bayes Theorem
- Information Theory
- NLP
- Statistical Algorithms
- Stochastic Algorithms

Introduction to Text Mining

- What is Text Mining?
- Natural Language Processing
- Pre-processing text data
- Extracting features from documents
- Using BeautifulSoup
- Measuring document similarity
- Knowledge check Lab activity

Supervised, Unsupervised, and Semi-supervised Learning

- What is prediction?
- Sampling, training set, testing set.
- Constructing a decision tree
- Knowledge check Lab Activity

Course Outline

Day 5

Data Visualization

- Choosing the right visualization
- Plotting data using Python libraries
- Plotting data using R
- Using Jupyter Notebook to validate scripts
- Knowledge check
- Lab activity

Big Data Landscape

- What is small data?
- What is big data?
- Big data analytics vs Data Science
- Key elements in Big Data (3Vs)
- Extracting values from big data
- Challenges in Big data

Data Analysis Presentation

- Using Markdown language
- Convert your data into slides
- Data presentation techniques
- The pitfall of data analysis
- Knowledge check
- Lab activity
- Group presentation Lab: Mini Project

Big data Tools and Applications

- Introducing Hadoop Ecosystem
- Cloudera vs Hortonworks
- Real world big data applications
- Knowledge check
- Group discussion

What's Next?

- Preview of Data Science Specialist
- Showing advanced data analysis techniques
- Demo: Interactive visualizations