Data Science: Concepts and Practice

Course slides

1. Introduction

Data Science – An Overview

AIM:

To familiarize students about data science.

Data Science – An Overview

Outcomes:

At the end of this module, you are expected to:

- Explain the concept of Data Science, its evolution and importance.

Data Science – An Overview

Content

- Introduction to Data Science
- Definition of Data Science
- Description of Data Science
- History and Development of Data Science

Data Science – An Overview

Introduction to Data Science

What is Data Science?

- Data Science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

- Data Science is the study of data.

- It involves developing methods of recording, storing, and analysing data to effectively extract useful information.

Data Science – An Overview

Definition of Data Science

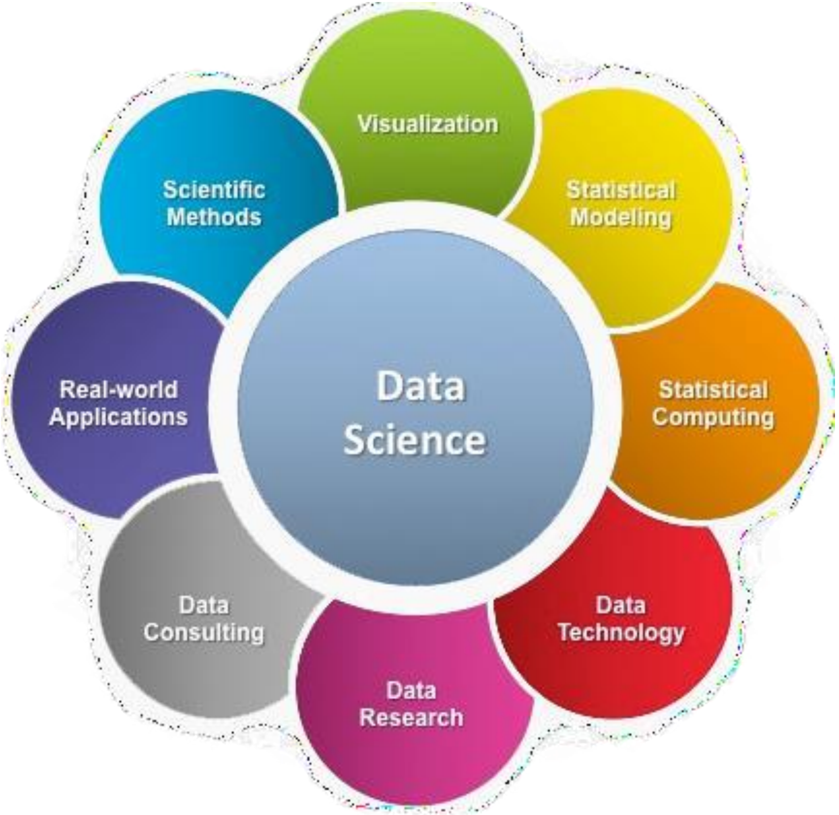
- Data science is a broad field that refers to the collective processes, theories, concepts, tools and technologies that enable the review, analysis, and extraction of valuable knowledge and information from raw data.

- It is geared towards helping individuals and organisations to make better decisions from stored, consumed and managed data.

- Data science was formerly known as datalogy.

Data Science – An Overview

Definition of Data Science



Data Science – An Overview

Description of Data Science

- Data science enables the use of theoretical, mathematical, computational and other practical methods to study and evaluate data.

- The key objective is to extract required or valuable information that may be used for multiple purposes, such as decision making, product development, trend analysis, and forecasting.

Data Science – An Overview

Description of Data Science

- Data science provides meaningful information based on large amount of complex data or big data.

- Data science, or data-driven science, combines different fields of work in statistics and computation to interpret data for decision-making purposes.

Data Science – An Overview

Description of Data Science

Understanding Data Science

- Data is drawn from different sectors, channels, and platforms including cell phones, social media, e-commerce sites, healthcare surveys, and internet searches.

- The increase in the amount of data available opened the doors to a new field of study based on big data - the massive data sets that contribute to the creation of better operational tools in all sectors.

Data Science – An Overview

Description of Data Science

(Continued) Understanding Data Science

- The continually increasing access to data is possible due to advancement in technology and collection techniques. Individuals buying patterns and behaviour can be monitored and predictions made based on the information gathered.
- The ever-increasing data is unstructured and requires parsing for effective decision making. This process is complex and time-consuming for companies - hence, the emergence of data science.

Data Science – An Overview

Description of Data Science

Data scientists are a new class of analytical data expert who has the technical skills to solve complex problems – and the interest to explore what kind of problems need to be solved. Data science techniques include data mining, big data analysis, data extraction, data retrieval and statistics.

Data science concepts and processes are derived from data engineering, statistics, programming, social engineering, data warehousing, machine learning and natural language processing, among others.

Data Science – An Overview

Goal of Data Science

- The goal of data science is to gain insights and knowledge from any type of data – both structured and unstructured.
- Data science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.

Data Science – An Overview

History and Development of Data Science

- The term data science has existed for the better part of the last 30 years and was originally used as a substitute for "computer science" in 1960.

- Approximately 15 years later, the term was used to define the survey of data processing methods used in different applications.

- In 2001, data science was introduced as an independent discipline.

- The Harvard Business Review published an article in 2012 describing the role of data scientist as the “sexiest job of the 21st century.”

Data Science – An Overview

History and Development of Data Science

- In 1974, Peter Naur authored the Concise Survey of Computer Methods, using the term “Data Science,” repeatedly. Naur presented his own convoluted definition of the new concept: “The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.”

- In 1977, The IASC, also known as the International Association for Statistical Computing was formed. The first phrase of their mission statement reads, “It is the mission of the IASC to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge.”

Data Science – An Overview

History and Development of Data Science

- In 1977, Tukey wrote a second paper, titled Exploratory Data Analysis, arguing the importance of using data in selecting "which" hypotheses to test, and that confirmatory data analysis and exploratory data analysis should work hand-in-hand.

- In 1989, the Knowledge Discovery in Databases, which would mature into the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, organised its first workshop.

Data Science – An Overview

History and Development of Data Science

In 1994, Business Week ran the cover story, Database Marketing, revealing the ominous news companies had started gathering large amount of personal information, with plans to start strange new marketing campaigns. The flood of data was, at best, confusing to company managers, who were trying to decide what to do with so much disconnected information.

Data Science – An Overview

History and Development of Data Science

In 1999, Jacob Zahavi pointed out the need for new tools to handle the massive amount of information available to businesses, in Mining Data for Nuggets of Knowledge.

He wrote, “Scalability is a huge issue in data mining. Conventional statistical methods work well with small data sets. Today’s databases, however, can involve millions of rows and scores of columns of data. Another technical challenge is developing models that can do a better job analysing data, detecting non-linear relationships and interaction between elements. Special data mining tools may have to be developed to address web-site decisions.

Data Science – An Overview

History and Development of Data Science

In 2001, William S. Cleveland laid out the plans for training Data Scientists to meet the needs of the future. He presented an action plan titled, Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics. It described how to increase the technical experience and range of data analysts and specified six areas of study for university departments. It promoted developing specific resources for research in each of the six areas. His plan also applies to government and corporate research.

In 2002, the International Council for Science: Committee on Data for Science and Technology began publishing the Data Science Journal, a publication focused on issues such as the description of data systems, their publication on the internet, applications and legal issues.

In 2006, Hadoop 0.1.0, an open-source, non-relational database, was released. Hadoop was based on Nutch, another open-source database.

Data Science – An Overview

History and Development of Data Science

In 2008, the title, “Data Scientist” became a buzzword, and eventually a part of the language. DJ Patil and Jeff Hammerbacher, of LinkedIn and Facebook, are given credit for initiating its use as a buzzword.

In 2009, the term NoSQL was reintroduced (a variation had been used since 1998) by Johan Oskarsson, when he organised a discussion on “open-source, non-relational databases”.

In 2011, job listings for Data Scientists increased by 15,000%. There was also an increase in seminars and conferences devoted specifically to Data Science and Big Data. Data Science had proven itself to be a source of profits and had become a part of corporate culture.

Data Science – An Overview

History and Development of Data Science

In 2011, James Dixon, CTO of Pentaho promoted the concept of Data Lakes, rather than Data Warehouses.

Dixon stated the difference between a Data Warehouse and a Data Lake is that the Data Warehouse pre-categorises the data at the point of entry, wasting time and energy, while a Data Lake accepts the information using a non-relational database (NoSQL) and does not categorise the data, but simply stores it.

Data Science – An Overview

History and Development of Data Science

In 2013, IBM shared statistics showing 90% of the data in the world had been created within the last two years.

In 2015, using Deep Learning techniques, Google's speech recognition, Google Voice, experienced a dramatic performance jump of 49 percent.

In 2015, Bloomberg's Jack Clark, wrote that it had been a landmark year for Artificial Intelligence (AI). Within Google, the total of software projects using AI increased from "sporadic usage" to more than 2,700 projects over the year.

Data Science – An Overview

History and Development of Data Science

In the past 10 years, Data Science has quietly grown to include businesses and organisations world-wide. It is now being used by governments, geneticists, engineers, and even astronomers.

During its evolution, Data Science's use of Big Data was not simply a "scaling up" of the data, but included shifting to new system for processing data and the ways data gets studied and analysed.

Data Science – An Overview

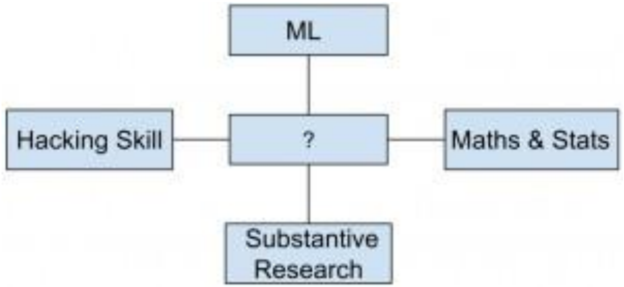
History and Development of Data Science

Data Science has become an important part of business and academic research. Technically, this includes machine translation, robotics, speech recognition, digital economy, and search engines.

In terms of research areas, Data Science has expanded to include the biological sciences, health care, medical informatics, the humanities, and social sciences. Data Science now influences economics, governments, and business and finance.

Self Assessment Question

1. Which one of the following would be more appropriate to be replaced with question mark in the following figure?



a) Data Analysis

b) Data Science

c) Descriptive Analytics

Answer: Data Science

Data Science – An Overview

Self Assessment Question

2. Which one of the following is performed by a Data Scientist?

a) Defines the question

b) Create reproducible code

c) Challenge results

d) All of the above

Answer: All of the above

Self Assessment Question

3. Which one of the following approach should be used to ask Data Analysis question?

    a) Find only one solution for particular problem

    b) Find out the question which is to be answered

    c) Find out answer from dataset without asking question

    d) None of the mentioned

Answer: Find out the question which is to be answered

Data Science – An Overview

Assignment

1. What is data science?

2. What is the goal of data science?

3. Summarise the evolution of data science.

Data Science – An Overview

Summary

- Data Science is a combination of various techniques to analyse the data.
- Statistics, computer science, machine learning etc are components of data science.
- Data science is the science of analysing data and finding some new patterns or hidden information which can be very fruitful for the business.

Data Science – An Overview

Document Links

Topics	URL	Notes
Data Science Definition	https://www.techopedia.com/definition/30202/data-science	This link explains the Introduction to Data Science
Understanding of Data Science	https://www.investopedia.com/terms/d/data-science.asp	This link explains the understanding of Data Science
History of Data Science	https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#351110b355cf	This link explains the history of Data Science
	https://www.dataversity.net/brief-history-data-science/	This link explains the history of Data Science

Data Science – An Overview

Video Links

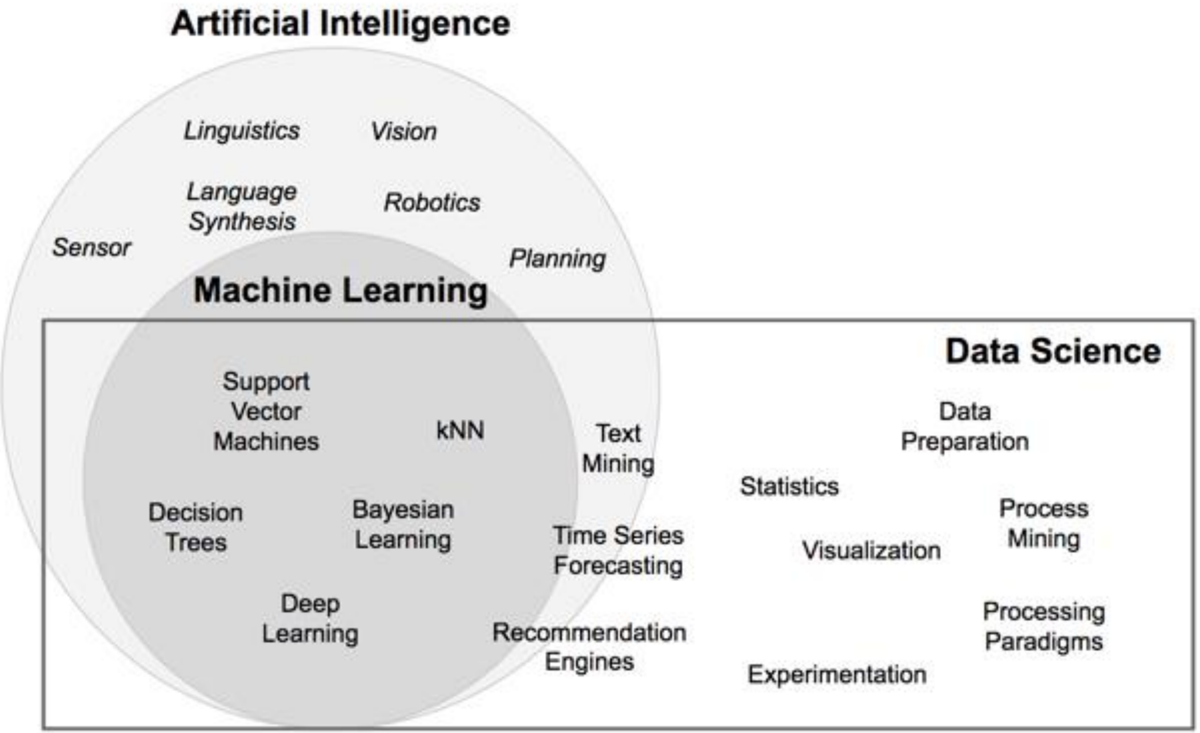
Topics	URL	Notes
Introduction to Data Science	https://www.youtube.com/watch?v=Oqb3WSPOyv8	This video explains introduction of data science
Foundation of Data Science	https://www.youtube.com/watch?v=2502R_mOhWc	How data science gets started
Data Science	https://www.youtube.com/watch?v=WEBUWYxaqLQ	What is features in data science

Data Science – An Overview

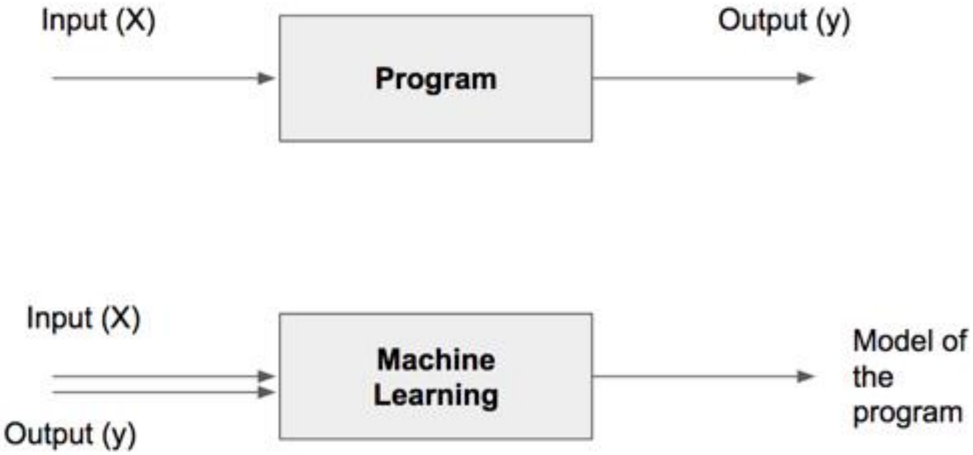
E-Book Links

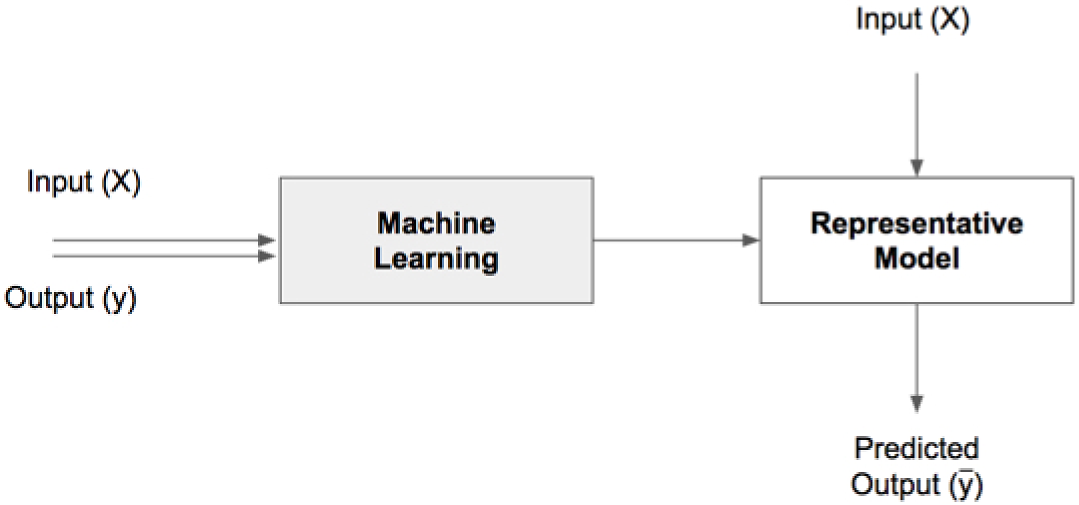
E-Book Name	Page Number	URL	Comments
An Introduction to Data Science by Jeffrey Satnton, Syracuse University - 3rd Edition	3 to 13	https://docs.google.com/file/d/0B6iefdnF22XQeVZDSkxjZ0Z5VUE/edit?pli=1	Explains introduction to data science
Introduction to Data Science by Sanjeev Ranjan Das	25 to 27	https://srdas.github.io/Papers/DSA_Book.pdf	Explains the steps involved in algorithmic problem solving.

What is Data Science



Models


Input (X)
Program
Output (y)
Input (X)
Output (y)
Machine Learning
Model of the program



Input (X)

Output (y)

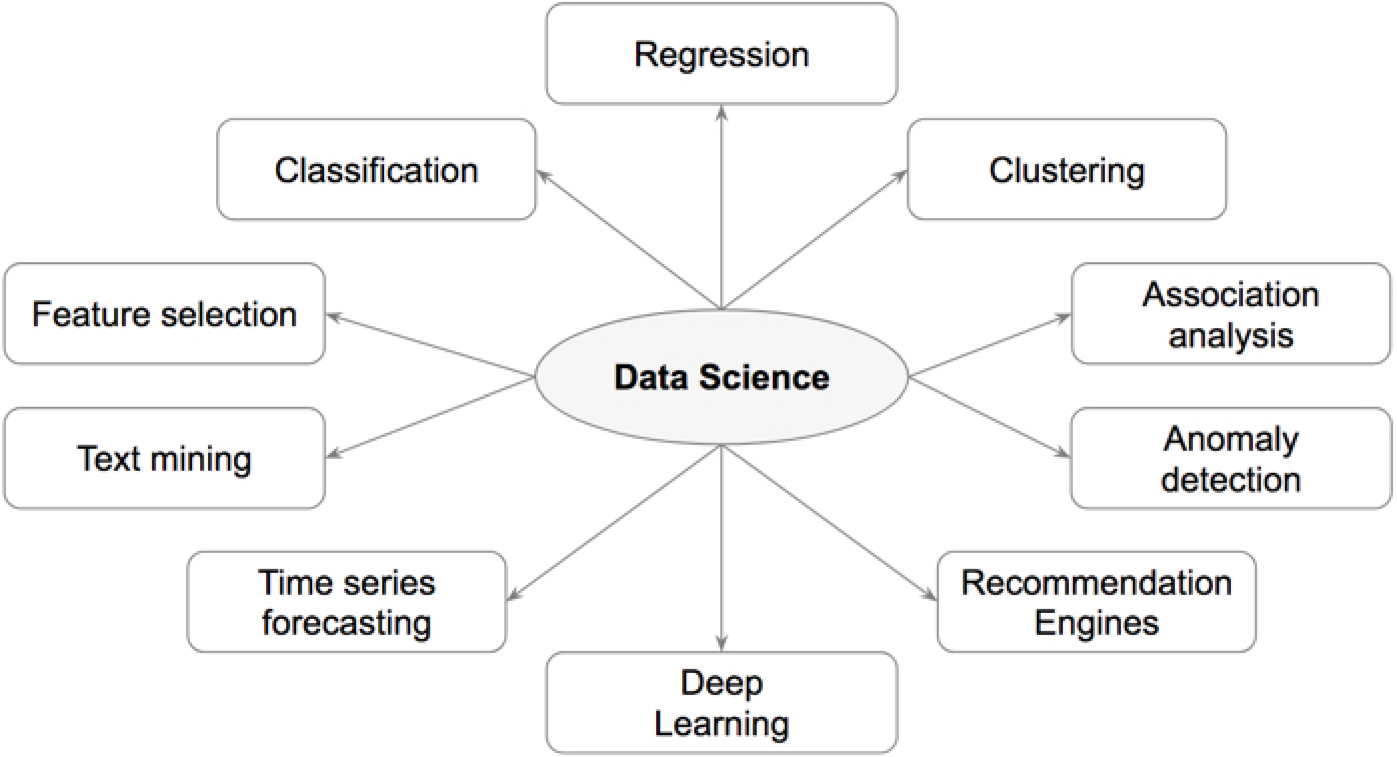
Machine Learning

Representative Model

Input (X)

Predicted Output ($\bar{y}$)

Types of Data Science



Data Science

Regression

Classification

Clustering

Feature selection

Association analysis

Text mining

Anomaly detection

Time series forecasting

Recommendation Engines

Deep Learning

Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of predefined classes. The prediction will be based on learning from known data set.	Decision Trees, Neural networks, Bayesian models, Induction rules, K nearest neighbors	Assigning voters into known buckets by political parties eg: soccer moms. Bucketing new customers into one of known customer groups.
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from known data set.	Linear regression, Logistic regression	Predicting unemployment rate for next year. Estimating insurance premium.
Anomaly detection	Predict if a data point is an outlier compared to other data points in the data set.	Distance based, Density based, LOF	Fraud transaction detection in credit cards. Network intrusion detection.
Time series	Predict if the value of the target variable for future time frame based on history values.	Exponential smoothing, ARIMA, regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the data set based on inherit properties within the data set.	K means, density based clustering - DBSCAN	Finding customer segments in a company based on transaction, web and customer call data.
Association analysis	Identify relationships within an itemset based on transaction data.	FP Growth, Apriori	Find cross selling opportunities for a retailor based on transaction purchase history.

Course outline

Process Basics

Data Science Process

Data Exploration

Model Evaluation

Core Algorithms

Classification

Decision Trees

Rule Induction

k-Nearest Neighbors

Naïve Bayesian

Artificial Neural Networks

Support Vector Machines

Ensemble Learners

Regression

Linear Regression

Logistic Regression

Association Analysis

Apriori

FP-Growth

Clustering

k-Means

DBSCAN

Self-Organizing Maps

Common Applications

Text Mining

Time Series Forecasting

Anomaly Detection

Feature Selection