



Cover Page



UNDERSTANDING THE DATA SCIENCE DOMAIN WITH ITS COMPONENTS AND DATA SOURCES

¹Kardile Vilas Vasantrao and ²Patil Rahul Ashok

^{1&2}Computer Science Department

¹Tuljaram Chaturchand College, Baramati and ²K.T.H.M College, Nashik
Pune, Maharashtra, India

Abstract: This era Data science is most popular and attracting topic of information technology. Information technology is a fusion of computer science, management and technology. Data Science is glimmer of such fusion that can brings several changes and challenges. These changes and challenges are produce a path of improvement with several perceptual and misperception aspects. ‘In such environment, minor mistake in appreciation create lot of unsolvable problem. With the motivation of “good knowledge is provide the best opportunity for developos to good foundation of landmark to useful application”. An objective of this study is to understand the data science, its aspects and abstracts the pitfalls.

Keywords: Data Science Concept, v3’s, Need of Data Science, Components, Toolbox, Application, Data and Its sources, Challenges of Data Science.

Introduction

Data science is branch of computer science that helps the user to gain insight from a large amount of available data using a variety of scientific method, algorithms and process on the basis of hidden patterns. Numerous Authors define data science in various prospectus but in here abstract it as it as per Wikipedia it is define as “Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and non-structured data, and applies knowledge and actionable insights from data across a wide range of application domains.” [7]

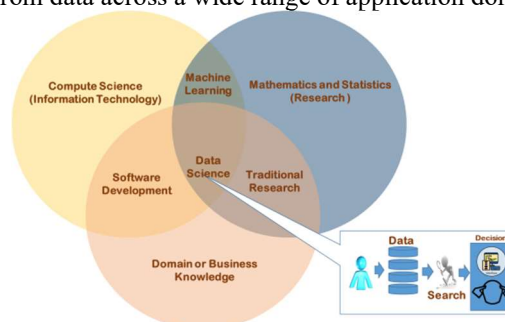


Figure 1: Domicile of Data Science

The term ‘Data Science’ is fusion of various disciplines that allows users to extract knowledge from structured or unstructured data. It helps the user to solve a business problem through a research project and then use it in a practical way. So, it needs to be understood very correctly, because to avoid violating the basic structure of the underlying concept, it takes a fallacious logical path that can lead to confusion through misunderstanding. In this consideration in Section–I the conceptual framework of data science is explain, theoretical prospects of data science process framework is explained in II-section, III-section deliberated with data science’s Components and Data sources. IV section focuses relevance and its application, In V-Section considered the summary of this study.

Section –I Conceptual framework of Data science: Today, we can see the use of data science in various areas of the world for the purpose of creating new knowledge. It is a knowledge revolution in every sector. It became most attracting phenomena globally by every sector with the aim of “significantly increase performance on the basis of by strategic manner performing utilization of data in scientific way”.

The conceptual framework of data science is based on the 3 V’s: Volume (the volume of data contains the data that has need to consider for their importance), Velocity (The speed of real time data generation online or offline is called as Velocity of data), Variety (The various type of data has been created manually or machine online or offline instance). [2]



In data science data storage and its source, data analysis, visualize, Data manipulate, and data operate with security privilege are very important activity perform on the basis of following task.

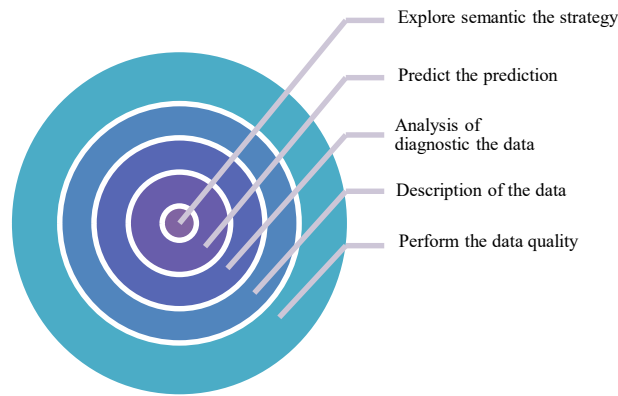


Figure 2: Conceptual framework of Data Science

Section –II Theoretical prospects of data science process framework: In this consideration available literature most prominently focus that, “the data science is a practice that generates the most strategic actionable data from data”. Simply we can understand it as it is cyclic process that starts with data and end with data. It became more understandable when we focus phases. It have 7 phases[1,3,5] (Business Understanding (Preliminary Investigation, Data Mining (Discovery), Data cleaning (Data Preparation), Exploration (Model Planning), Feature Engineering (Model Building), Prediction Modeling (Operationalize), Data Visualization (communicate Result)): that indicated is lifecycle, as showing following figure.



Figure 3: The life-cycle of data science.



Cover Page



Business Understanding (Preliminary Investigation): This is initial phase user's basic requirement, it's 'Problem' are understand and discuss the proposed solution. This phase fact finding phase and utilizes techniques like (Interview, Questionnaire, Record review and Observation) for understand the problem and try to propose solution in favor of organization.

Data Mining (Discovery): This is second phase, where data scientist identify basic requirement, priorities, budget, no of stakeholder, technology, time in hand, data availability and end goal. At end of the phase Data scientist can frame-out proposed problem and its solution at the first hypothesis level.

Data cleaning (Data Preparation): This is third phase; rough data obtained from previous phase may in scattered position in various sources. So, there is need to inline the needed data and clean the unwanted data. In this phase data is clean by data cleaning process like (data reduction process and Data integration process). This phase end with data manipulates and transformation activities that produce input data in further process.

Exploration (Model Planning): Data exploration is done in fourth phase, in this phase data analysis and defines method and techniques to establish the relation between input variables very carefully. In this phase data scientist utilize "Exploratory Data Analytics" (EDA) for graphical representation technique like graph (Histogram, Scatter plots... etc.).

Feature Engineering (Model Building): This is fifth phase; in this phase feature engineering process is applied with machine learning. In this phase data scientist create data set for design and testing purpose on the basis of association, Classification and Clustering data to develop the model. Here data scientists outline domain knowledge and deep learning of data is required to model's executing algorithm to improve the features predication.

Prediction Modeling (Operationalize): This is the sixth phase that executes the project and predicts the future event and action on the basis of outcome of project. The final reports of the project with technical document are delivering and define along with discussions in this phase. In this phase, Data scientist provides a clear overview of complete projects execution and its performance before the full deployment.

Data Visualization (communicate Result): This is last phase, in this phase data visualization process is proceed with define the information about the outcome of projects. Here outcome of the project verified and validate as per its goal with intelligent perception of user.

Section- III: Components of data science and its Data sources

Data Science Components: The main components of Data Science are given below:

Statistics: The important component of data science is 'Statistics'. It assists for collect and analyze numerical information and produce meaningful outcome from it.

Domain Expertise: The data science is interdisciplinary field that has various knowledgeable or skill expertise. Domain expertise binds all these expertise together.

Data engineering: An Important segment of data science is Data engineering. It contains acquiring, storing, retrieving and transformation of the data and metadata (data about data).

Visualization: Data visualization is nothing but representing huge number data in visual context that will assist the user to understand the significance of data.

Advanced computing: Advanced computing is nothing but designing, writing, debugging, and maintaining the source code of computer programs.

Mathematics : It contains the Study of quantity, structure, scope and changes. Good knowledge of mathematics is essential to become good data scientist.

Machine learning: Backbone of data science is Machine learning. It provides training to a machine. It is nothing but machine activation as per human knowledge and logic. Several machine learning algorithms to solve the problems.

Data Scientist's Toolbox: Data scientist utilizes various tools on various circumstances apart from that there are some tools mostly utilized for solving the data science project as follows:

Table 1: Data science Components or Tools and its Technology	
Components or Tools	Technology
Data Storage (Warehouse)	SQL, Hadoop, ETL, Informatica, Talend, AWS, Redshift, MongoDB, Hbase, Cassandra----- etc.
Data Analysis tools	JAVA, C++, MS-Excel, SAS Statistical tools, Python, R and R Studio, MATLAB, Jupyter, Rapidminer, ---, etc.
Machine Learning	JAVA, Ruby, Perl, C++, ML Studio, Spark, Azure, Mahout., etc
Data Visualizations	R, Tableau, Cognos, Jupyter, -----, etc.



Cover Page



Types of Data: To utilize the above components successfully there is need to understand the data its type, its data sources. Generally, above compotes process its operation on diffident types of data like (Structured data, Semi-structured data, and unstructured data).

Structured: The Structured data is defined as data or record that stored in fixed field within file. The data stored in predefine and searchable format of DBMS/RDBMS or warehouse or data lake system is called structured data. Generally, structured data contain number, text and generated by machine and manually. For example (no, name..., etc.)

Semi-structured: The Semi-structured data is define as data or record that have self-describing structure. Such data does not fit into the formal structure of DBMS/RDBMS but, it serves the system with different element and allows to search. For example (Smartphones photo, CSV and XML file,.....,etc)

Unstructured Data: The unstructured data is defined as data or record that is not in predefined structured format or not having any kind data model. Such kind of data is not stored in its native format. For example rich media, text, social media activity, surveillance image....etc. In the consideration of structured and semi-structured data, unstructured data is much larger. It is more than 80% proposed enterprise data and it is continuously grows. Such data are not take into account are missing out on a valuable business intelligence of organization.

Data sources: For the data science data scientist are able to uses various data sources like (Open Data, Social Media Data, Multimodal Data, Standard datasets,...,etc)

Data Formats (Integers, Floats, Dense Numerical Arrays, Text, Compressed, Archived Data)

Various data files (Text Files, CSV Files, JSON Files, XML Files, HTML Files, Tar Files, G-Zip Files, .Zip Files, Image Files (Rasterized, Vectored).

Section -IV: Relevance of data science and its application: In today's fast-growing world, it is basic need that, to achieve enhancement, there is need to take a proper decision making by intelligence strategy (right tools, technologies, and algorithm). Data Science is grown up as an intelligence tool that discloses the need-base insights to make efficient strategy and policy for enhancement [6]. It is proven with various examples that; it assists the user to choose enhancing strategy in relevant sector with various applications like (Internet Search. Recommendation Systems, Image & Speech Recognition, Gaming world, Online Price Comparison, Fraud and Risk Detection, Healthcare, Airline route planning, Augmented Reality, Transport, etc)

Threats (Risk and Challenges) : In this study there is need to focus on the some threats, that will assist to us for better understand. The available published literature specifies, some points that may be consider as some risks and challenges. Apart from this, in this study few of them are specified as follows.

- There are various data sources are utilizes in application so, data ownership, data standardization and authentication issues and policy violation issues may be raises.
- Data integration, missing and noisy data handling and its reliability, consistency, timeliness, incompleteness, and scalability of data problems may be raises.

Opportunities: Data is a very valuable segment in today's competitive, swift and dynamic environment, we have a various type of data and its sources, but it is indolent stored position. So it is necessary to find valuable insights to make smart data based decisions. The available published literature significantly explores that various applications are used today in various industry domains like as marketing, healthcare, finance, banking policy work and more.

Section –V Summary of this study: With the consideration of available published publication consider for this study, one can easily understand “Data science is glimmer of various segments fusion like (Statistics, Visualization, Deep Learning, Machine Learning, Deep learning). It has capability to produces extracting visions from data by the utilizing of various scientific methods, algorithms, and processes with 6-7 phases in cyclic formation. It provides most attracting profile with its own challenges like (error handling, lack of standardization, policy and legal issues, heterogeneity, inconsistency, timeliness, incompleteness, and scalability of data). Available published publications highlight the need to provide solutions for risks and challenges. However, regardless of their risks and challenges, data science has many current and future opportunities that have the potential to contribute to economic growth of nations, data science sector is bound with good opportunities for organizations, industries and nations. This is very initial attempt for understand the Data Science, this study appreciates, there is need to practically execute application for understand it more exclusively.



Cover Page



Reference

1. Joel Grus” Data Science from Scratch” 2nd edition., O’Reilly
2. Lillian Pierson”Data Science For Dummies”, 1st edition, John Wiley & Sons
3. Martin KleppmannDesigning Data-Intensive Applications”,1st edition O’Reilly Media
4. Viktor Mayer-Schönberger “Big Data” Reprint edition,Harper Business
5. Cole Nussbaumer Knaflie “Story telling with Data”1st edition, Wiley
6. V. Dhar “Data science and prediction. Commun”. ACM 56(12), 64–73 (2013). doi:10.1145/2500499
7. Wikipedia<https://www.wikipedia.org>