Enliven Archive
Intensifying Thoughts

JOURNAL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

OPEN ACCESS

Research Article

A Study on Data Science Methodologies and Analytics

R Saradha*, Sudha Rajesh

Assistant Professor, BSA Crescent Institute of Science & Technology, Vandalur, Chennai, India.

***Corresponding Author:**

R. Saradha,
Assistant Professor, BSA Crescent Institute of Science & Technology, Vandalur, Chennai, India.
Tel: 9486123676
Email: saradaghari@gmail.com

Received: December 17, 2019
Accepted: January 19, 2020
Published: January 25, 2020

Abstract

As the world entered the era of big data, the need for its storage also grew. It was the main challenge and concern for the enterprise industries until 2010. The main focus was on building a framework and solutions to store data. Now when Hadoop and other frameworks have successfully solved the problem of storage, the focus has shifted to the processing of this data. Data Science is the secret sauce here. All the ideas which we see in Hollywood sci-fi movies can actually turn into reality by Data Science. Data Science is the future of Artificial Intelligence. Multiple research tracks will be championed by members of the data science community with the goal of enabling rigorous comparison of approaches through common tasks, datasets, metrics, and shared research challenges. This article examines foundational issues in data science including current challenges, basic research questions, and expected advances, as the basis for a new data science research program (DSRP) and associated data science evaluation (DSE) series, introduced by the National Institute of Standards and Technology (NIST) in the fall of 2015.

Keywords: Data Science Evaluation Series; Data Science Standards; Data Science Metrics; Data Science Measurements; Data Analytics.

Introduction

Data science is a way to try and discover hidden patterns in raw data. To achieve this goal, it makes use of several algorithms, Machine Learning (ML) principles, and scientific methods. The insights it retrieves from data lie in forms structured and unstructured. So in a way, this is like data mining. Data science encompasses all- data analysis, statistics, and machine learning. 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 [1, 2]. Data science is the same concept as data mining and big data: "use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems" [3].

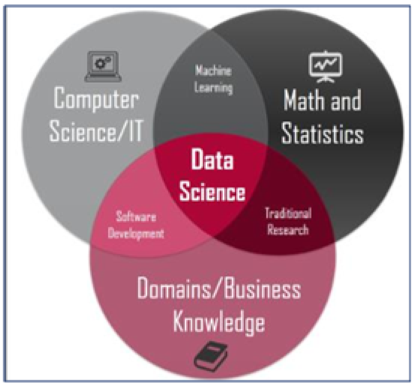
The modern definition of "data science" was first sketched during the second Japanese - French statistics symposium organized at the University of Montpellier II (France) in 1992 [4]. The attendees acknowledged the emergence of a new discipline with a specific focus on data from various origins, dimensions, types, and structures. They shaped the contour of this new science - based on established concepts and principles of statistics and data analysis with the extensive use of the increasing power of computer tools. The main advantage of enlisting data science in an organization is the empowerment and facilitation of decision-making. Organizations with data scientists can factor in quantifiable, data-based evidence into their business decisions. These data-driven decisions can ultimately lead to increased profitability and improved operational efficiency, business performance, and workflows. In customer - facing organizations, data science helps identify and refine target audiences. Data science can also assist recruitment: Internal processing of applications and data-driven aptitude tests and games can help an organization's human resources team make quicker and more accurate selections during the hiring process. The specific benefits of data science vary depending on the company's goal and the industry. Sales and marketing departments, for example, can mine customer data to improve conversion rates or create one-to-one marketing campaigns. Banking institutions are mining data to enhance fraud detection. Streaming services like Netflix mine data to determine what its users are interested in, and use that data to determine what TV shows or films to produce. Data-based algorithms are also used at Netflix to create personalized recommendations basedon a user's viewing history. Shipment companies like DHL, FedEx, and UPS use data science to find the best delivery routes and times, as well as the best modes of transport for their shipments.

Data science is still an emerging field within the enterprise be-cause the identification and analysis of vast amounts of unstructured data can prove too complex, expensive and time consuming for companies.

Classifications of Data Science Problems

This section examines several classes of problems for which techniques might be developed and evaluated across different domains and defines representative classes of problems accompanied by examples from the planned use case of traffic incident detection and prediction, although the problem classes are broader than this single use case. Different categories of algorithms and techniques in data science will be examined, with an eye toward building an assessment methodology for the DSE that covers each category.

Figure 1. Challenge in Data Science.



There are three fundamental methods in Data science. These methods are the basis for extracting useful knowledge from data, and also serve as a foundation for many well-known algorithms in data science. Detection, Regression, and prediction.

Detection

Detection aims to find data of interest in a given dataset. In the traffic domain, incidents are of interest, e.g., "traffic incident detection" is an important subproblem of the traffic use case. Yang et al., [5] analyze traffic flow in order to detect traffic incidents. Anomaly detection is the identification of system states that force additional pattern classes into a model. Relatedly, outlier detection is associated with identifying potentially erroneous data items that force changes in prediction models ("influential observations"). For example, through anomaly detection in health insurance claim records, potentially fraudulent claims can be identified. In the traffic case, an incident may be seen as an anomaly relative to data representing free- flowing traffic. Detection of incidents in traffic data with incident and non-incident data may also be seen as system state identification and estimation [6].

Cleaning

Cleaning refers to the elimination of errors, omissions, and inconsistencies in data or across datasets. In the traffic use case, cleaning might involve the identification and elimination of errors in raw traffic sensor data.

Regression

Regression refers to the process of finding functional relationships between variables. In the traffic pilot, the posed challenge might be to model the traffic flow rate as a function of other variables.

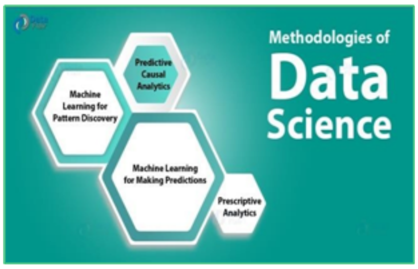
Prediction

Prediction refers to the estimation of a variable or multiple variables of interest at future times. In the pilot traffic flow prediction challenge, the participants are asked to predict traffic speed using covariates including flow volume, percentage occupancy, and training sets of past multivariate time series. Structured prediction refers to tasks where the outputs are structured objects, rather than numeric values [7, 8]. This is a desirable technique when one wishes to classify a variable in terms of a more complicated structure than producing discrete or real-number values. In the traffic domain, an example might be producing a complete road network where only some of the roads are observed. Knowledgebase construction Knowledge base construction refers to the construction of a database that has a predefined schema, based on any number of diverse inputs. Researchers have developed many tools and techniques for automated knowledge base construction (AKBC). In the traffic use case, a database of incidents and accidents could be constructed from news reports, time-stamped global positioning system (GPS) coordinates, historical traffic data, imagery, etc.

Other classes of problems Data science problems may involve ranking, clustering, and transcription (alternatively called "structured prediction" as defined above). Several of these are described by Bengio et al., [10]. Additional classes of problems rely on algorithms and techniques that apply to raw data at an earlier "preprocessing" stage. Given the broad scope of the classes of problems above, a number of different data processing algorithms and techniques may be employed for which an evaluation methodology is essential, for example, for benchmarking. The next section elaborates on the range of methodologies needed for measuring technology effectiveness within the new DSRP.

Strategies of Datascience

Figure 2. Methodologies of Data Science.


Methodologies of
Data
Science
Predictive Causal Analytics
Machine Learning for Pattern Discovery
Machine Learning for Making Predictions
Prescriptive Analytics

Machine Learning for Pattern Discovery

With this, clustering comes into play. This is an algorithm to use to discover patterns; an unsupervised model. When we don't have parameters on which to make predictions, clustering will let we find hidden patterns within adataset.

One such use-case is to use clustering in a telephone company to determine tower locations for optimum signal strength.

Machine Learning for Making Predictions

When we have the data we need to train our machine, we can use supervised learning to deal with transactional data. Making use of machine learning algorithms, we can build a model and determine what trends the future will observe.

Predictive Causal Analytics

Causal analytics lets us make predictions based on a cause. This will tell us how probable an event is to hold an occurrence in the future. One use-case will be to perform such analytics on payment histories of customers in a bank. This tells us how likely customers are to reimburse loans.

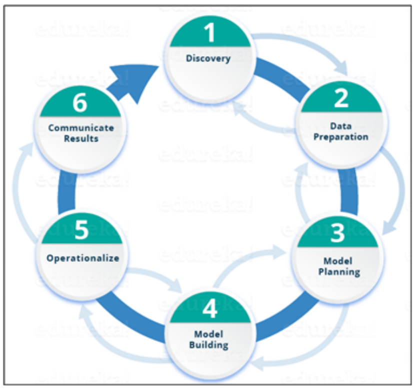
Prescriptive Analytics

Predictive analysis will prescribe the actions and the outcomes associated with those. This intelligence lets it take decisions and modify those using dynamic parameters. For a use-case, let us suggest the self-driving car by Google. With the algorithms in place, it can decide when to speed up or slow down, when to turn, and which road totake.

Life Cycle of DataScience

Here is a brief overview of the main phases of the Data Science Lifecycle:

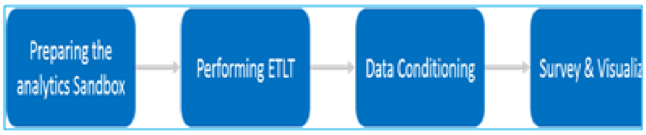
Figure 3. Life Cycle of Data Science.



Phase 1 - Discovery: Before we begin the project, it is important to understand the various specifications, requirements, Priorities and required budget [11]. We must possess the ability to ask the right questions. Here, we assess if we have the required resources present in terms of people, technology, time and data to support the project. In this phase, we also need to frame the business problem and formulate initial hypotheses (IH) to test.

Phase 2 - Data preparation: In this phase, we require analytical sandbox in which we can perform analytics for the entire duration of the project. We need to explore, preprocess and condition data prior to modeling. Further, we will perform ETLT (extract, transform, load and transform) to get data into thesandbox.

Figure 4. Statistical Analysis Flow.

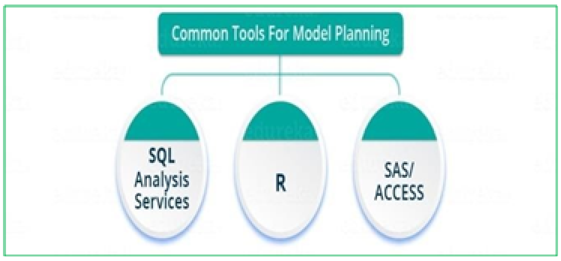


We can use R for data cleaning, transformation, and visualization. This will help us to spot the outliers and establish a relationship between the variables. Once we have cleaned and prepared the data, it's time to do exploratory analytics on it.

Phase 3 - Model planning: Here, we will determine the methods and techniques to draw the relationships between variables. These relationships will set the base for the algorithms which we will implement in the next phase. We will apply Exploratory Data Analytics (EDA) using various statistical formulas and visualization tools.

Let's have a look at various model planning tools.

Figure 5. Tools for Model Planning.



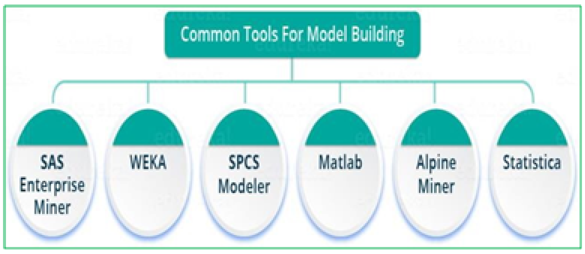
1. R has a complete set of modeling capabilities and provides a good environment for building interpretive models.
2. SQL Analysis services can perform in-database analytics using common data mining functions and basic predictive models.
3. SAS/ACCESS can be used to access data from Hadoop and is used for creating repeatable and reusable model flow diagrams.

Although, many tools are present in market R is the most commonly used tool.

Phase 4 - Model building: In this phase, we will develop datasets for training and testing purposes. We will consider whether our existing tools will suffice for running the models or it will need a more robust environment (like fast and parallel processing). We will analyze various learning techniques like classification, association, and clustering to build the model.

We can achieve model building through the following tools.

Figure 6. Model Building.



Phase 5 - operationalize: In this phase, we deliver final reports, briefings, code, and technical documents. In addition, sometimes a pilot project is also implemented in a real- time production environment. This will provide us a clear picture of the performance and other related constraints on a small scale before full deployment.

Phase 6 - Communicate results: Now it is important to evaluate if we have been able to achieve the goal that we had planned in the first phase. So, in the last phase, we identify all the key findings, communicate to the stakeholders and determine if the results of the project are a success or a failure based on the criteria developed in Phase 1.

Challenges of Data Science

Future Work

Big data analytics and data science are becoming the research focal point in industries and academia. Data science aims at researching big data and knowledge extraction from data. Applications of big data and data science include information science, uncertainty modeling, uncertain data analysis, machine learning, statistical learning, pattern recognition, data warehousing, and signal processing. Effective integration of technologies and analysis will result in predicting the future drift of events. The main focus of this section is to discuss open research issues in big data analytics. The research issues pertaining to big data analysis are classified into three broad categories namely the internet of things (IoT), cloud computing, bio-inspired computing, and quantum computing. However it is not limited to these issues. More research issues related to health care big data can be found in Husing Kuo et al., paper [12].

Knowledge acquisition from IoT data is the biggest challenge that big data professionals are facing. Therefore, it is essential to develop infrastructure to analyze the IoT data. An IoT device generates continuous streams of data and the researchers can develop tools to extract meaningful information from these data using machine learning techniques. Under-standing these streams of data generated from IoT devices and analyzing them to get meaningful information is a challenging issue and it leads to big data analytics.

Conclusion

In recent years data are generated at a dramatic pace. Analyzing this data is challenging for a general man. To this end in this paper, we survey the various research issues, challenges, and tools used to analyze these big data. From this survey, it is understood that every big data platform has its individual focus. Some of them are designed for batch processing whereas some are good at real-time analytics. Each big data platform also has specific functionality. Different techniques used for the analysis include statistical analysis, machine learning, data mining, intelligent analysis, cloud computing, quantum computing, and data stream processing. We believe that in future researchers will pay more attention to these techniques to solve problems of big data effectively and efficiently.

References

[1]. Dhar V. Data science and prediction. Communications of the ACM. 2013 Dec 1;56(12):64-73.

[2]. Leek J. The key word in'Data Science'is not Data, it is Science. Simply Statistics. 2013 Dec 12;12.

[3]. Smith M. The White House names Dr. DJ Patil as the first US chief data scientist. The White House Blog. 2015 Feb.

[4]. Bergen KJ, Chen T, Li Z. Preface to the focus section on machine learning in seismology. Seismological Research Letters. 2019 Mar;90(2A):477-80.

[5]. Yang S, Kalpakis K, Biem A. Detecting road traffic events by coupling multiple timeseries with a nonparametric bayesian method. IEEE Transactions on Intelligent Transportation Systems. 2014 Mar 11;15(5):1936-46.

[6]. Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM computing surveys (CSUR). 2009 Jul 30;41(3):1-58.

[7]. BakIr G, Hofmann T, Schölkopf B, Smola AJ, Taskar B, editors. Predicting structured data. MIT press; 2007.

[8]. Lafferty J, McCallum A, Pereira FC. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[9]. Dorr BJ, Greenberg CS, Fontana P, Przybocki M, Le Bras M, Ploehn C, Aulov O, Michel M, Golden EJ, Chang W. The NIST data science initiative. In2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA) 2015 Oct 19 (pp. 1-10). IEEE.

[10]. LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015 May;521(7553):436-44.

[11]. Hayashi C. What is data science? Fundamental concepts and a heuristic example. InData science, classification, and related methods 1998 (pp. 40-51). Springer, Tokyo.

[12]. Kuo MH, Sahama T, Kushniruk AW, Borycki EM, Grunwell DK. Health big data analytics: current perspectives, challenges and potential solutions. International Journal of Big Data Intelligence. 2014 Jan 1;1(1-2):114-26.