# Data Driven Science: Application and Future

**Mritunjay[1] Anoop Kumar[2] Amit Kumar[3]**
[1,2,3]Research Scholar
[1,2,3]Department of Information & Technology Engineering
[1,2,3]IIMT College of Engineering, Greater Noida, India

*Abstract*— The volume of data that is generated each day is rising rapidly. There is a need to analyze this data efficiently and produce results quickly. Data science offers a formal methodology for processing and analyzing data. It involves a work-flow with multiple stages, such as, data collection, data wrangling, statistical analysis and machine learning. In this paper, we look at data analytics systems that support the data science work-flow. Diving in at a granular level to mine and understand complex behaviors, trends, and inferences. It's about surfacing hidden insight that can help enable companies to make smarter business decisions.

***Keywords:*** Data Driven Science, Hacking, Data Science

## I. INTRODUCTION

Data-driven science is the extract learning of large information that is instructed or unorganized, which is known as information mining. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.[4] It employs techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, and information science .Peter Naur quoted the term data logy for data science which existed for over thirty years and was first used as a substitute for comport science by him. In 1974,Peter Naur published" Concise Survey of Computer Methods" which freely used the term data science in its survey of the contemporary data processing methods and that are used in a wide range of applications. The field of data science uses information planning, insights, and machine learning to research issues in different places for example, advertising, setting open strategy. Data science researchers may utilize the capacity to discover and translate rich information sources, programming equipment's, transfer speed, data set consistency, and comprehensive of information, scientific models and more important discoveries. William S. Cleveland introduced data driven science as an independent discipline, which extending the field of statistics to incorporate and advances in computing with data in his article.

## II. LITERATURE

### *A. Steps in Data Science:*

The three segments included in data science are as follows:
1) Arranging
2) Bundling and
3) Conveying information

However bundling is an integral part of data wrangling, which includes collection and sorting of data. What isolates data science from other existing disciplines is that they additionally need to have a nonstop consciousness of What, How, Who and Why. A data science researcher needs to realize what will be the yield of the data science transform and have an unmistakable vision of this yield. A data science researcher needs to have a plainly characterized arrangement on in what manner this yield will be accomplished inside of the limitations of accessible assets and time. A data scientist needs to profoundly comprehend who the individuals are that will be included in making the yield.

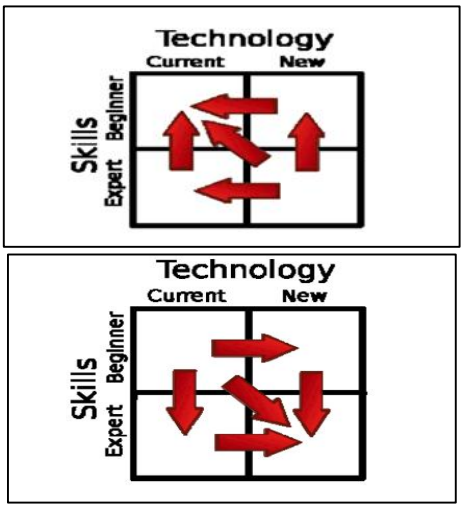
The steps of Data Science are mainly:
1) Collection
2) Preparation of the data

### *B. The following are the Basic Steps Involved in Data Science:*

#### *1) Arranging Data and Munging:*

Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. This process typically includes manually converting/mapping data from one raw form into another format to allow for more convenient consumption and organization of the data. Raw data can be unstructured and messy, with information coming from disparate data sources, mismatched or missing records, and a slew of other tricky issues. By Data Munging we mean the process of taking raw data, understanding it, cleaning it and preparing it for analysis or modeling. It is by no means the glamorous part of data science however if done well it plays a more important role in getting to powerful models and insights than what algorithm you use. Work of cleaning up data so that it is polished and ready for downstream usage.

The key steps to data wrangling:
- Data Acquisition: Identify and obtain access to the data within your sources
- Data Cleansing: Redesign the data into a usable/functional format and correct/remove any bad data.




#### *2) Data Analysis:*

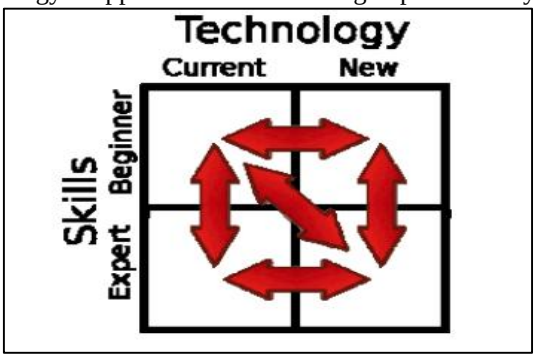
Analysis or investigation of data is a procedure of assessing, changing, and demonstrating information with the objective of finding helpful data, recommending conclusions, and supporting decision-making.

Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data.

*3) Convey Data:*

Conveying data includes methods to transform the mathematical or statistical conclusions drawn from the data into a form that can be easily understood and interpreted by those in need of it. Conveying data is empowering the development starting with one perspective then onto the next, empowering a beginner to turn into an expert, current technology to appear to be new and allowing the modeled information to be seen by apprentices and making new technology to appear like it was an integral part of the system.




a) Hacking

Wearer not talking about hacking as in breaking into computers. We're referring to the tech programmer subculture meaning of hacking i.e., creativity and ingenuity in using technical skills to build things and find clever solutions to problems.

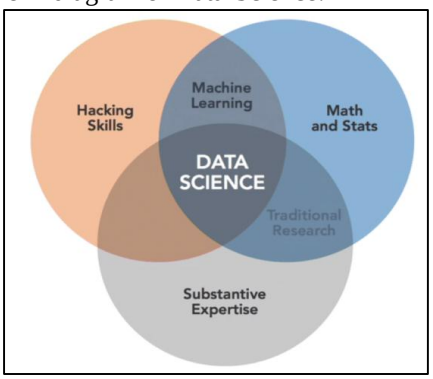
b) Math's and Statistics Knowledge:

To extract meaning from large volumes of data, a data scientist must have knowledge of at least some basic level of mathematics and statistics, since most data science techniques involve statistical computation and modeling.

c) Expertise:

The fundamental aim of data science is to build knowledge, it must build upon previous knowledge bases and discoveries. This requires that the data scientist must have a large amount of experience at his disposal, so that the best results can be obtained from the new data.

d) Venn diagram of Data-Science:




e) Applications:

Using data science, companies have become intelligent enough to push & sell products as per customer's purchasing power & interest. Here's how they are ruling our hearts and minds:

1) Internet search: Search engines like Yahoo, Bing, Ask, AOL, Duckduckgo etc. All these search engines including Google which is widely used search engine make use of data science algorithms to deliver the best result for our searched query in fraction of seconds.
2) Prediction: Large amounts of data collected and analyzed can be used to identify patterns in data, which can in turn be used to build predictive models. This is the basis of the field of machine learning, where knowledge is discovered using induction algorithms and on other algorithms that are said to learn. Machine learning techniques are largely used to build predictive models in numerous fields.
3) Security: Security mostly refers to protection from hostile forces, but it has a wide range of other senses. Data collected from user log are used to detect fraud using data science.

f) Computer vision:

Computer Vision or Machine Vision's main role is to determine whether or not data in an image consists of some specific object or activity. Algorithms are used to automatically analyze images and extract information

4) Natural language processing: It is broadly defined as the automatic manipulation of natural language, like speech and text, by software. Natural Language Processing (NLP) is the art and science which helps us extract information from text and use it in our computations and algorithms. Given then increase in content on internet and social media, it is one of the must have still for all data scientists out there.
5) Future Trends: In near future, the demand for data science at all levels will keep increasing while the shortage of data scientists will remain for a while. This means that it will be relatively easy to get entry-level jobs.

But in the far future, the entry-level jobs might be "threatened" by automated machine learning and data analytics tools like automatic science tools will be smart and powerful enough to replace data scientists form all tasks. In this case, we certainly need more hard work to get data science jobs. We need to be able to work on tasks that "machines" are not good at, such as formulating good hypothesis / questions, engineering features for machine learning models, fine tuning models for specific use cases, and etc. Material Sciences, Drug discovery, Quantum mechanics, Neuroscience, Nanotechnology, and many more have greatly benefited from change in method in which studies are done, data analytics have proved to be a useful process than other .The surge in huge information, examination and intellectual figuring methodologies will give choice backing and computerization to people, and mindfulness and knowledge to machines. These advancements can be utilized to make both people and things more astute.

## III. Conclusion

Practice of data science can be explained as the combination of exploration and engineering analytics. This presents the problem that needs to be likely solved and is restated to data mining tasks. The problem is firstly decomposed in to subtasks that are solved using the existing tools and the rests

are assumed that the tools are not enough to solve the problem so in the cases we have to mine the data and conduct evaluation to see and if it also doesn't work we apply the completely different method. In that case we may discover the knowledge or discover something unexpected that leads to important success. Neither the exploration nor is the analytical engineering omitted while drafting the solution of a problem. The hiding of captcha, building an entire STO (send time optimization), improving user experience, maintain the security of users IP address etc. will be included in the major sourcing points of data mining and this thesis is all about it. This method and research will definitely having the good future scope with security, improved user experience along with hassle free working.

## REFERENCES

[1] https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article
[2] https://www.analyticsvidhya.com/blog/2015/09/applications-data-science/
[3] Data Mining and Knowledge Discovery. ISSN: 1384-5810 (print version). ISSN: 1573-756X (electronic version). Journal no. 10618
[4] Mining Data for Nuggets of Knowledge Dec 10, 1999 Mining Data for Nuggets of Knowledge. http://knowledge.wharton.upenn.edu/article/mining-datafor-nuggets-of-knowledge/
[5] Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics William S. Cleveland Statistics Research, Bell Labs.
[6] Statistical Modelling: the two cultures Leo Breiman. Statistical Science. Vol. 16 No.3 (August 2001) 199-215.