International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 7 | ISSN : 2456-3307

A Comprehensive Study on Application and Future Trends in Data Science

Manohar Vemula[1], Shaik Balasaidulu[2]
[1]MCA,MCPD,CSM, Software Engineer, Hyderabad, Telangana, India
[2]MCA,Osmania University, Hyderabad, Telangana, India

ABSTRACT

Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information. Although the name Data Science seems to connect most strongly with areas such as databases and computer science, many different kinds of skills including non-mathematical skills are also needed here. Data Science is much more than simply analyzing data. There are many people who enjoy analyzing data who could happily spend all day looking at histograms and averages, but for those who prefer other activities, data science offers a range of roles and requires a range of skills. Data science includes data analysis as an important component of the skill set required for many jobs in the area, but is not the only skill. Data scientists play active roles in the design and implementation work of four related areas such as data architecture, data acquisition, data analysis and data archiving. In the present paper the authors will try to explore the different issues, implementation and challenges in area called Data science.

Keywords : Information, Data, Unstructured Data, Visualization, Management, Preservation

I. INTRODUCTION

Data Science is the extraction of learning from substantial volumes of information that are unorganized or unstructured, which is a continuation of the field of information mining and perceptive investigation, otherwise called information disclosure and information mining. "Unstructured data" can incorporate messages, features, photographs, online networking, and other client produced substance. Information science frequently obliges dealing with an awesome measure of data and composing calculations to concentrate bits of knowledge from this information. John Tukey's quote about data science and the necessity which aids its evolution: "The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data."[1][2] To quote Hal Varian, Google's Chief Economist[3], "The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it". The field of data science uses information planning, insights, and machine learning to research issues in different spaces, for example, advertising improvement, extortion discovery, setting open strategy, and so forth. Data science researchers utilize the capacity to discover and translate rich information sources; oversee a lot of information notwithstanding equipment, programming, and transfer speed imperatives; consolidate information sources; guarantee consistency of datasets; make representations to help in comprehension of information; construct scientific models utilizing the

information; and display and impart the information experiences/discoveries. The basic flow control in a data science process can be summed up in the following diagram

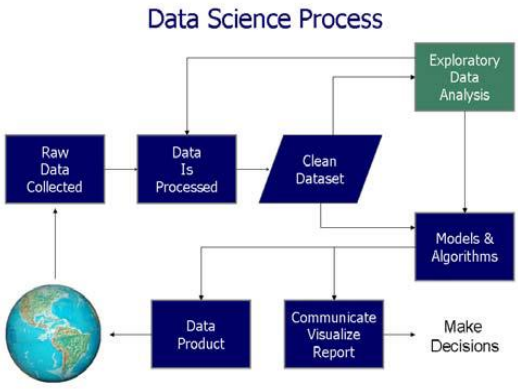

Figure 1. Steps Involved in a Data Science Process

II. BASIC STEPS OF DATA SCIENCE

The three segments included in data science are arranging, bundling and conveying information (the ABC of information). However bundling is an integral part of data wrangling, which includes collection and sorting of data. However 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: collection and preparation of the data, alternating between running the analysis and reflection to interpret the outputs, and finally dissemination of results in the form of written reports and/or executable code. The following are the basic steps involved in data science.

a) Data wrangling and munging

Collecting data from relevant areas and the process of manually converting or mapping data from one "raw" form into another format that allows for more convenient consumption and manipulation of the data with the help of semi-automated tools is referred to as data wrangling[4] or munging[5].Sorting out data includes the physical stockpiling and arrangement of information and joined best practices in information administration. It basically includes moving individuals and frameworks from current to new and from learner to master (start to finish). Propelling advances and abilities is the pith of development.

Bundling data is the next step that follows arranging data. Bundling data includes consistently controlling and joining the fundamental crude information into another representation and bundle. Bundling data is actually the opposite of sorting out data and includes moving individuals and frameworks from new to current and from master to apprentice (base to beat). This is the specialty of making things basic yet not less complex.

b) 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. The data is processed using various algorithms of statistics and machine learning to extract meaning and useful conclusions from the large volumes of data.

c) 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 technologyto appear like it was an integral part of the system.

III. SKILLS OF A DATA SCIENTIST

Data science as a field is the meeting point of a number of disciplines. A practitioner of data science possesses a combination of skills from across a variety of subjects. Basically, it is the combination of three major fields, as shown in the well-known Venn diagram(Fig. 5)[7]

1. Hacking skills – A data scientist must have the ability to extract and structure data. To do so, he/she must possess advanced programming abilities to manipulate data and apply algorithms.
2. Math 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 modelling.
3. Substantive Expertise – Since 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/her disposal, so that the best results can be obtained from the new data.

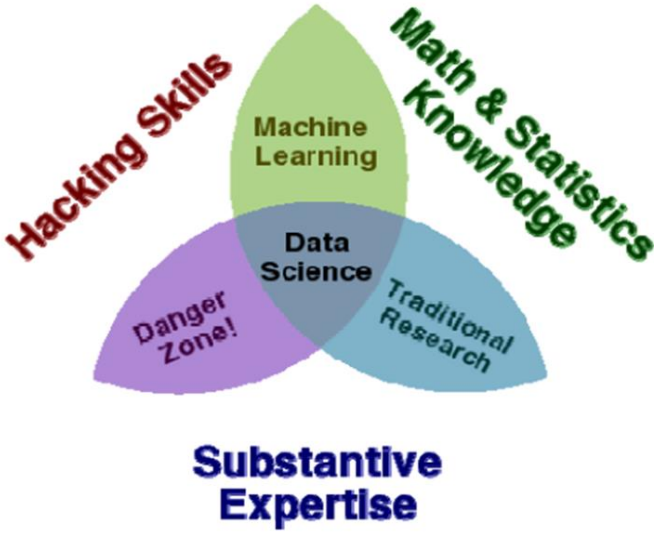

Figure 2 The Data Science Venn Diagram

IV. EVOLUTION OF DATA SCIENCE

In 1962, John W. Tukey wrote "The Future of Data Analysis"[8]. In this publication, Tukey coined the term "bit" which Claude Shannon used in his 1948 paper "A Mathematical Theory of Communications." In 1977, Tukey published Exploratory Data Analysis[9], arguing that more emphasis needed to be placed on using data to suggest hypotheses to test and that Exploratory Data Analysis and Confirmatory Data Analysis "can—and should—proceed side by side."

In 1974, Peter Naur published a "Concise Survey of Computer Methods"[10] which was a survey of contemporary data processing models that are used in a wide range of applications.Naur offered the following definition of data science: "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 International Association for Statistical Computing (IASC) is established as a Section of the ISI. The objective of the IASC was summed up as "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."

In 1989, Gregory Piatetsky-Shapiro organized and chaired the first Knowledge Discovery in Databases (KDD) workshop. In 1995, it became the annual ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD).[11]

Database marketing was talked about in the cover story of Business Week published in September 1994.Companies were collecting mountains of information, crunching it to predict how likely a customer is to buy a product, and using that knowledge to craft a marketing message precisely calibrated to get the desired customer response An earlier flush of enthusiasm prompted by the spread of checkout scanners in the 1980s ended in widespread disappointment: Many companies were too overwhelmed by the sheer quantity of data to do anything useful with the information. Still, many companies believed they have no choice but to brave the database-marketing frontier.In 1996 data science was included in the title of a conference for the first time.[12]

In 1996 Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth published "From Data Mining to Knowledge Discovery in Databases."[13] Between 1996-1999, several publications[14] [15] [16] demonstrate the rising relevance of data mining and knowledge discovery in databases. In 2001 William S.Cleveland[17] proposed a plan "to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called 'data science.'" He proposed to merge this new discipline in the context of computer science and data mining. In 2001 Leo Breiman[18] proposed two cultures in statistical modelling of data - stochastic data models and algorithmic models. The evolution and increasing popularity of data science is conclusive of its relevance in modern informatics. The applications of data science have been discussed in the following section

V. APPLICATIONS AND FUTURE SCOPE

Data science is a subject that arose primarily from necessity, in the context of real-world applications instead of as a research domain. Over the years, it has evolved from being used in the relatively narrow field of statistics and analytics to being a universal presence in all areas of science and industry. In this section, we look at some of the principal areas of applications and research where data science is currently used and is at the forefront of innovation.

1. Business Analytics –Collecting data about the past and present performance of a business can provide insight into the functioning of the business and help drive decision-making processes and build predictive models to forecast future performance. Some scientists have argued that data science is nothing more than a new word for business analytics[19], which was a meteorically rising field a few years ago, only to be replaced by the new buzzword data science. Whether or not the two fields can be considered to be mutually independent, there is no doubt that data science is in universal use in the field of business analytics.

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"[20]. Machine learning techniques are largely used to build predictive models in numerous fields.

3. Security – Data collected from user logs are used to detect fraud[21] using data science. Patterns detected in user activity can be used to isolate cases of fraud

and malicious insiders. Banks and other financial institutions chiefly use data mining and machine learning algorithms to prevent cases of fraud[22].

4. Computer Vision – Data from image and video analysis is used to implement computer vision[23], which is the science of making computers "see", using image data and learning algorithms to acquire and analyze images and take decisions accordingly. This is used in robotics, autonomous vehicles and human-computer interaction applications.

5. Natural Language Processing – Modern NLP techniques use huge amounts of textual data from corpora of documents to statistically model linguistic data, and use these models to achieve tasks like machine translation[24], parsing, natural language generation and sentiment analysis[25].

6. Bioinformatics – Bioinformatics[26] is a rapidly growing area where computers and data are used to understand biological data, such as genetics and genomics. These are used to better understand the basis of diseases, desirable genetic properties and other biological properties. As pointed out by Michael Walker – "Next-generation genomic technologies allow data scientists to drastically increase the amount of genomic data collected on large study populations. When combined with new informatics approaches that integrate many kinds of data with genomic data in disease research, we will better understand the genetic bases of drug response and disease."

7. Science and Research – Scientific experiments such as the well-known Large Hadron Collider project generate data from millions of sensors[27] and their data have to be analyzed to draw meaningful conclusions. Astronomical data from moderntelescopes[28] and climatic data stored by the NASA Center for Climate Simulation[29] are other examples of data science being used where the volume of data is so large that it tends towards the new field of Big Data.

8. Revenue Management - Real time revenue management is also very well aided by proficient data scientists. In the past, revenue management systems were hindered by a dearth of data points. In the retail industry or the gaming industry too datascience is used. As Jian Wang defines it:"Revenue management is a methodology to maximize an enterprise's total revenue by selling the right product to the right customer at the right price at the right time through the right channel."Now data scientists have the ability to tap into a constant flow of real-time pricing data and adjust their offers accordingly. It is now possible to estimate the most beneficial type of business to nurture at a given time and how much profit can be expected within a certain time span.

9. Government - Data science is also used in governmental directorates to prevent waste, fraud and abuse ,combat cyber-attacks and safeguard sensitive information, use business intelligence to make better financial decisions ,improve defense systems and protect soldiers on the ground. In recent times most governments have acknowledged the fact that data science models have great utility for a variety of missions.

The use of data science as a quantitative approach to turn information into something valuable has been trending since quite some time. The desire for "the statistician that can code" or "the programmer that knows stats" has arisen from the need to efficiently utilise data by bundling them according to relevance or importance and using the same for information mining.

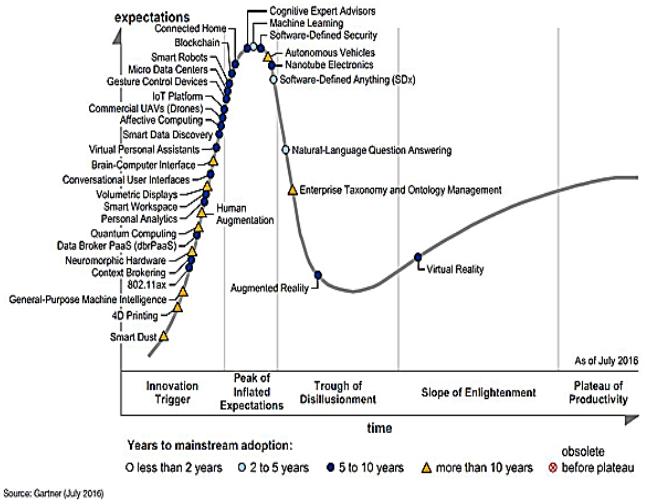

Figure 3 : Hype Cycle for Emerging technologies

The above chart is taken from "Hype Cycle for Emerging Technologies, 2013". We have crossed the peak of inflated expectation. And are now moving towards productivity.[30]This Hype Cycle brings together the most significant technologies from across Gartner's research areas. It provides insight into emerging technologies that have broad, cross-industry relevance, and are transformational and high-impact in potential.

The 2013 Emerging Technologies Hype Cycle highlights technologies that support all six of these areas including:

1. Augmenting humans with technology
2. Machines replacing humans
3. Humans and machines working alongside each other
4. The concept of Machines trying to develop a better understanding humans and the environment
5. Humans trying to understand machines as most mundane tasks are becoming automated.
6. Machines and humans becoming smarter and efficient.

As it can be seen data science is at the peak and it might be predicted that it can reach a plateau or inflation in few years.

For sure the future will be crowded with people trying to applying data science in all problems, kind of overusing it. But it can be sensed that we are going to see some real amazing applications of DS for a normal user apart from online applications (recommendations, ad targeting, etc). The skills needed for visualization, for client engagement, for engineering saleable algorithms, are all quite different. If we can perform everything perfectly at peak level it'd be great. However, if demand is robust enough companies will start accepting a diversification of roles and building teams with complementary skills rather than imagining that one person will cover all bases. Service Customization can be achieved by data science, one can achieve a person-level customization in almost any kind of services like healthcare, insurance, public services, banking, etc. We can utilize it to help approach making with the accessibility of most intricate topography level information on characteristic assets like water bodies, mineral stores, area sort/quality, and so on,man-made assets like streets, trains lines, air terminals, open workplaces/foundation, on citizens, their different properties, and their utilization example of items and administrations and even the government can make their approach making to a great degree modified, productive, shrewd, and receptive to changes. Knowledge creates knowledge to make future tasks easier. With more data the process of analysis and implementation becomes more efficient. Fields like 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 far fruitful process than many 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 intelligent.

VI. REFERENCES

1. Dhar, V. (2013). "Data science and prediction". Communications of the ACM 56.
2. Jeff Leek (2013-12-12). "The key word in 'Data Science' is not Data, it is Science". Simply Statistics.
3. Hal Varian on how the Web challenges managers. http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers
4. Parsons, MA, MJ Brodzik, and NJ Rutter. 2004. Data management for the cold land processes experiment: improving hydrological science.HYDROL PROCESS. 18:3637-653. http://www3.interscience.wiley.com/cgi-bin/jissue/109856902
5. Data Munging with Perl. DAVID CROSS. MANNING. Chapter 1 Page 4.
6. What is Data Science? http://www.datascientists.net/what-is-data-science
7. The Data Science Venn Diagram. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
8. Tukey, John W. The Future of Data Analysis. Ann. Math. Statist. 33 (1962), no.1,1--67. doi:10.1214/aoms/1177704711. http://projecteuclid.org/euclid.aoms/1177704711.
9. Tukey, John W. (1977). Exploratory Data Analysis. Pearson. ISBN 978-0201076165.
10. Peter Naur: Concise Survey of Computer Methods, 397 p. Studentlitteratur, Lund, Sweden, ISBN 91-44-07881-1, 1974
11. KDD-89: IJCAI-89 Workshop on Knowledge Discovery in Databases. August 20, 1989, Detroit MI, USA
12. Database Marketing Business Week. September 04, 1994
13. From Data Mining to Knowledge Discovery in Databases. Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth. AI Magazine Volume 17 Number 3 (1996)
14. "Statistics=Data Science?" C.F.Jeff Wu. University of Michigan, Ann Arbor. http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf
15. Data Mining and Knowledge Discovery. ISSN: 1384-5810 (print version). ISSN: 1573-756X (electronic version). Journal no. 10618
16. Mining Data for Nuggets of Knowledge Dec 10, 1999 Mining Data for Nuggets of Knowledge. http://knowledge.wharton.upenn.edu/article/mining-data-for-nuggets-of-knowledge/
17. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics William S. Cleveland Statistics Research, Bell Labs. http://www.stat.purdue.edu/~wsc/papers/datascience.pdf
18. Statistical Modelling: the two cultures Leo Breiman. Statistical Science. Vol. 16 No.3 (August 2001) 199-215.
19. Davenport, Thomas H. (January 1, 2006). "Competing on Analytics". Harvard Business Review
20. Glossary of Terms. Machine Learning - Special issue on applications of machine learning and the knowledge discovery process archive. Volume 30 Issue 2-3, Feb/March, 1998. Pages 271-274
21. Bolton, R. & Hand, D. (2002). Statistical Fraud Detection: A Review (With Discussion). Statistical Science 17(3): 235–255.
22. Neural data mining for credit card fraud detection. Brause, R.; Langsdorf, T.; Hepp, M. Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on. Publication Year: 1999, Pages: 103-106
23. ReinhardKlette (2014). Concise Computer Vision. Springer. ISBN 978-1-4471-6320-6.

24. Hutchins, W. John; Somers, Harold L. (1992). An Introduction to Machine Translation. London: Academic Press. ISBN 0-12-362830-X.
25. Akshi Kumar, Teeja Mary Sebastian. Sentiment Analysis on Twitter. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012. ISSN (Online): 1694-0814
26. Raul Isea. The Present-Day Meaning of the Word Bioinformatics, Global Journal of Advanced Research, 2015. Vol-2, Issue-1 PP. 70-73. ISSN: 2394-5788.
27. Brumfiel, Geoff (19 January 2011). "High-energy physics: Down the petabyte highway". Nature 469. pp. 282–83. doi:10.1038/469282a
28. Matthew Francis. Future telescope array drives development of exabyte processing. http://arstechnica.com/science/2012/04/future-telescope-array-drives-development-of-exabyte-processing/
29. "Supercomputing the Climate: NASA's Big Data Mission". CSC World. Computer Sciences Corporation. http://www.csc.com/cscworld/publications/81769/81773-supercomputing_the_climate_nasa_s_big_data_mission
30. Hype Cycle for Emerging Technologies, 2013.