Issues, Problems And Applied Research In Data Science

A. Victor Benevent Raj 1
Ananda College, Devakottai, India

J. John Kennedy 2
Assistant Professor,
Department of Computer Applications,
Ananda College, Devakottai, India

ABSTRACT

Data science, also identified as data-driven science, is an involving two or more academic, scientific, or artistic disciplines field about technical methods, processes, and systems to pull out information or insights from data in various forms, structured or unstructured, comparable to data mining.

Data science is about trade with large excellence of data for the rationale of extracting meaningful and rational results/conclusions/patterns. It's a recently rising field that encompasses a number of actions, such as data mining and data analysis. It employs techniques ranging from mathematics, statistics, and information technology, computer programming, data engineering, pattern recognition and learning, visualization, and towering presentation computing. This paper gives a clear plan about the diverse data science technologies used in Big data Analytics.

Data science is a "perception to unite statistics, data analysis and their connected methods" in order to "recognize and examine genuine phenomena" with data. It employs techniques and theories strained from many fields within the large areas of arithmetic, statistics, information science, and computer science, in particular from the sub domains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

Data Science is much more than purely analyzing data. There are a lot of people who take happiness in 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 variety of skills. Data science includes data analysis as an significant constituent of the talent set necessary for many jobs in the area, but is not the only skill. In this paper the author effort will rigorous on to travel around the diverse issues, accomplishment and challenges in Data science.

Keywords: *Data science, data, analytics, information, unstructured data, extraction, preservation, data visualization*

PROLOGUE

Data Science refers to a budding area of work anxious with the compilation, grounding, examination, apparition, administration, and protection of huge collections of in sequence. Even though the name Data Science seems to join most sturdily with areas such as databases and computer science, a lot of unlike kinds of skills counting nonmathematical skills are also desirable here.

Data Science is not only a artificial concept to unite statistics, data analysis and the aggravated methods, but also comprises its results. Data Science intends to examine and recognize genuine phenomena with "facts". In other words, the aspire of data science is to divulge the skin tone or the unseen structure of difficult accepted, human and social phenomena with data from a diverse tip of sight from the reputable or conventional hypothesis and technique.

In other words, the aspire of data science is to divulge the skin tone or the unseen structure of difficult accepted, human and social phenomena with data from a diverse tip of sight from the reputable or conventional hypothesis and technique. This point of outlook implies multi dimensional, dynamic and views.

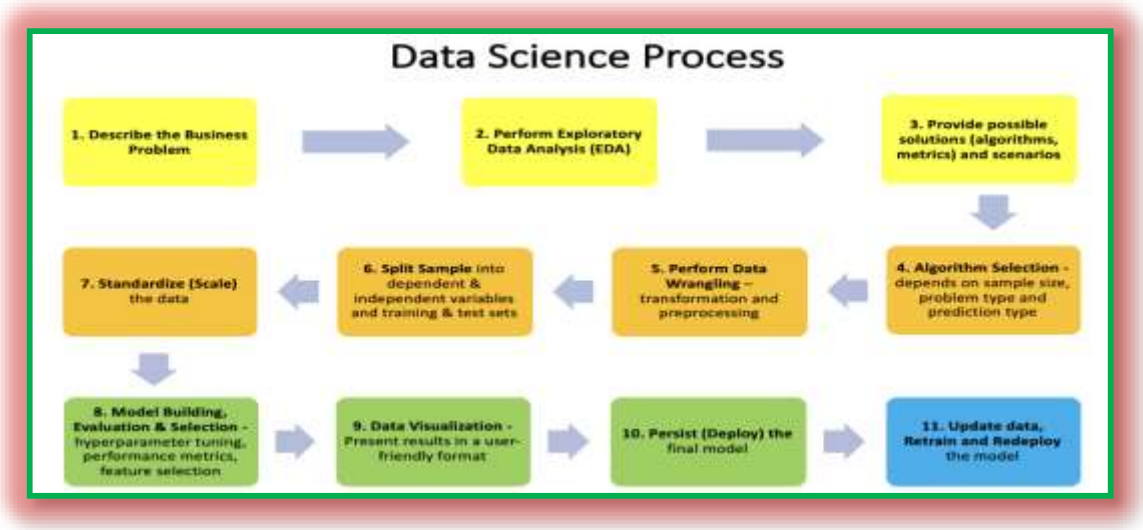
Change the way you think.Data research has

Three stages:

data creation, data collection, and data analysis. It is im portant to consider these three stages in an integrated co ncept based on the research concept explained below. A t these stages, a good stance should be taken to evaluate the established and practical methods of the product.

Data science is all about understanding from data; anal ytics is also about the need to “change the trend” and “understand the importance of business pressures” It is a change in the way information is sought, stored, and presented. Most often, it is used to describe an ext remely simple subject such as organizing, storing and analyzing data, as well as computer science, statistics, data warehousing and artificial intelligence. This is a new field, so there is no consensus on the management of what it has**.**

Data science is a combination of mathematics, statistics, operations, the details of the problem to be solved, the smart way to capture information that cannot be capture d at the moment, the ability to see things "another way", and of course important and important information. We need to clean, organize and organize documents.



I. CHALLENGES IN BIGDATA ANALYSIS

1. **Heterogeneity and Incompleteness** When people assimilate information; high levels of uniformity in heterogeneity are easily avoided. In fact, the lay ers and riches of natural language can provide dep th. However, analysis algorithms expect homogen eous data and cannot understand gradients. Theref ore, the data needs to be carefully designed as the first step (or preliminary step) in data analysis. For example, consider a patient undergoing various m edical procedures in a hospital. We can create reco rds for every treatment or test, records for the dura tion of the hospital stay, or records for every hospi tal interaction throughout the patient's life. But a computer works best when it can store many items of the same size and pattern. Quality representatio n, accessibility and analysis of semi-structured data require further study. Consider creating an ele ctronic medical record that includes fields for each patient's date of birth, occupation, and blood type.

2. **Scale Of course**, when it comes to big data, the fir st thing that comes to mind is its scale. After all, it has the word "big" in its name. Managing large an d rapidly growing data has been a challenge for de cades. In the past, this challenge has been alleviate d by speeding up processes based on Moore's Law and providing us with the resources we need to process more data. But now a significant chang e has occurred: data volume is expanding faster th an computing resources, and CPU speed is the sa me as before. First, there has been a huge change in technology in the last five years; Instead of dou bling the cycle every 18-24 months, the clock speed now remains the same and the clock speed does not change due to limited power are processe s. It constantly increases the number of cores. In the past, the big data processing system had to wor ry about the parallelism of nodes in the cluster. Now we need to address the parallelism in each other. Unfortunately, the data processing techniqu es used in the past to process data between nodes have not been directly applied to composites beca use the architecture looks very different; For exam ple, there are many hardware products. The processor cache and processor memory channels ar e shared between the cores in one of them.Additionally, the move to encapsulation of multiple sockets (10cores per socket) results in a different balance between cores. Finally, according to the prediction of Dark Silicon", future energy consumption may prevent us from consta ntly using all physical devices and computing systems. Energy management is necessary. These unprecedented changes require us to rethink how we create, build and operate information systems. The second major change underway is the move to cloud computing, which now involves many different tasks with different operational goals.

3. Time outside Libra is fast. The larger the data set to be processed, the longer the analysis will take. A design that controls size can result in a system that can process larger size files faster. But when people talk about speed in the context of big data, they often mean more than just speed. Instead, there iscompetition between purchasing, as discussed in the following Section, and the timing issues described later. In most cases, evaluation results are needed immediately.

For example, if a fraudulent credit card transaction is sus pected, it's a good idea to flag it before it goes through so the transaction is likely blocked altogether. Frankly, it is not possible to instantly analyze the user's entire purchas e history. Instead, we need part of the result first so we can use the new multiplication to make smaller numbers and get to a quick result. Given big data, it is often necessary to find content that meets the criteria specified in it.

This search will occur several times during the data analy sis process. It is not practical to analyze the entire data sets to find the right content. Instead, benchmarks are created in advance to ensure the right products are found quickly. The problem is that each standard is designed to support only certain types of standards. Since new analyz es are needed with large data sets, new models are deter mined and developed.

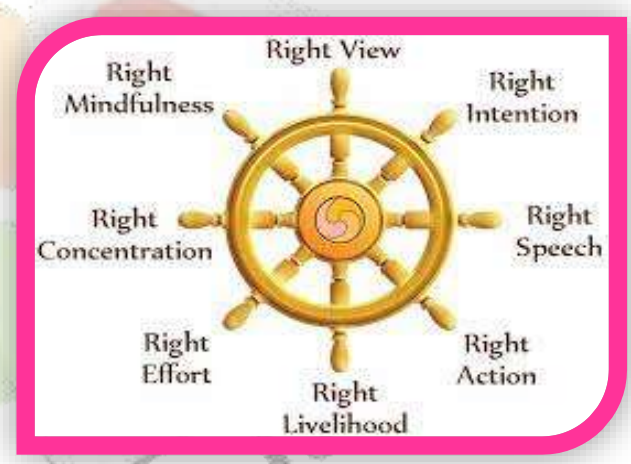
II. New benchmarks support such models. For example, consider a traffic management system that contains information about thousands of vehicles and local hotspots on the road. The system will need to predict traffic congestion points based on the user's chosen route and suggest an alternative route. To do this, it is necessary to examine various questions closely related to the path taken by moving objects. New standards are needed to support these questions. Building such a model can be particularly challenging when data volumes are growing rapidly and query response times are very short.

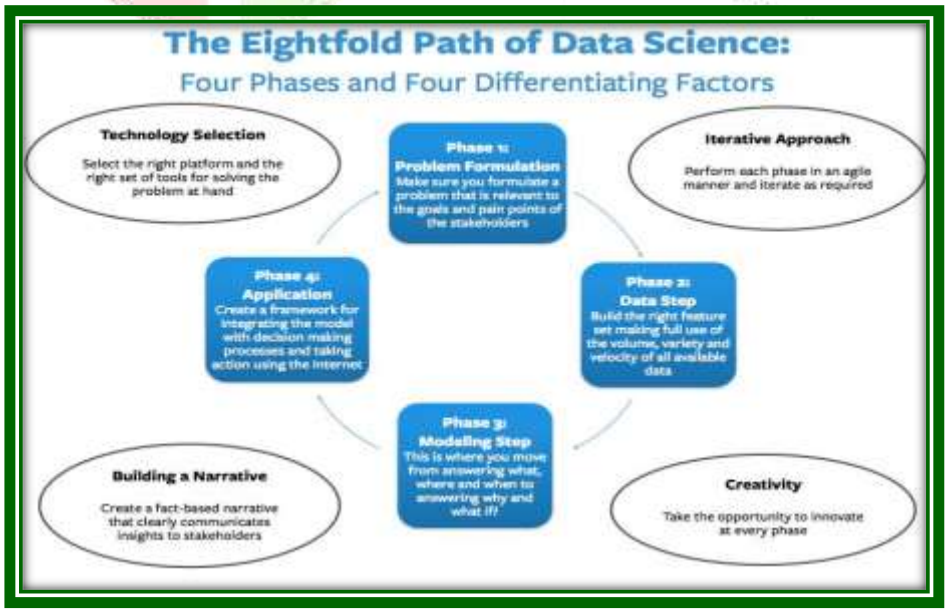
3. Privacy is another big problem and it is even worse in the context of big data. There are strict rules about what can and cannot be done with electronic health records. For other documents, regulations are particularly weak in the United States. However, the public is very concerned about the inappropriate use of personal information, especially by combining information from different sources. Self-regulation is actually both a skill and a health issue that must be addressed from both perspectives to realize the promise of big data.

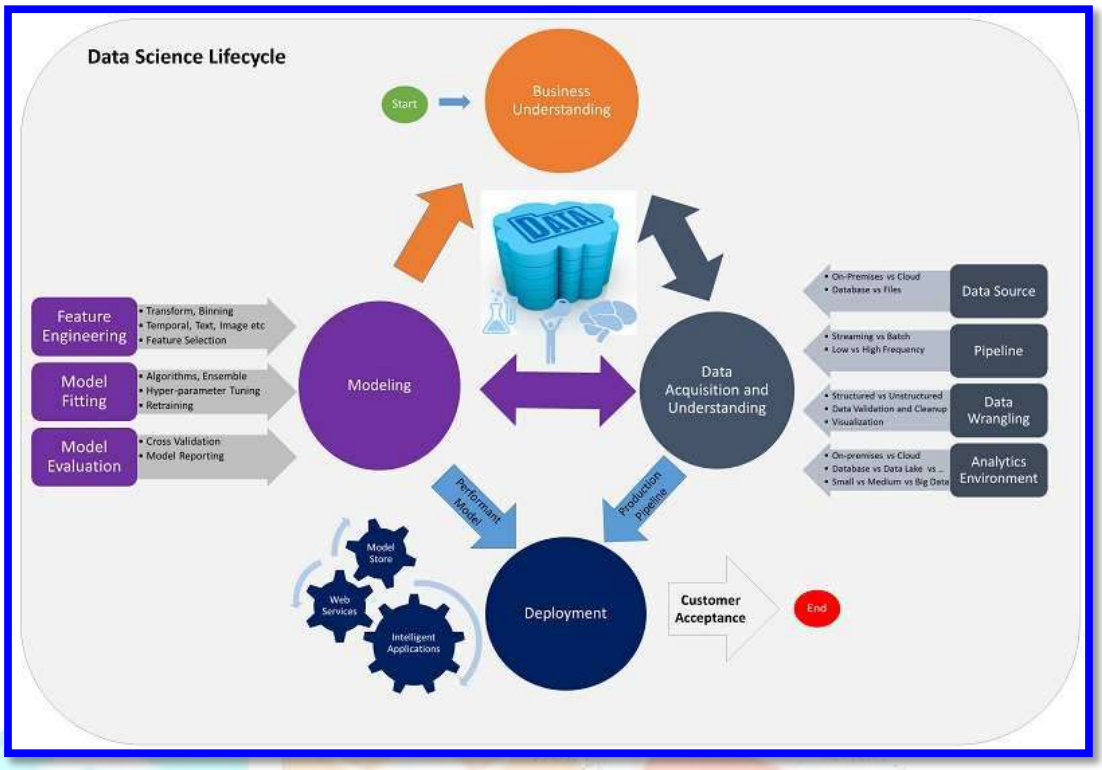
PHASES OF DATASCIENCE

The three foundations of data science are planning, packaging, and presenting data (data ABCs). But packaging is an important part of data organization that involves data collection and analysis. However, the difference between information science and other existing disciplines is that it still needs constant information about what, how, who and why. Data scientists must be aware of the benefits of transforming data science and have a clear vision of these benefits. Data scientists need to have a clear idea of

how to achieve this result within resources and time limits. Data scientists need a deep understanding of who will be involved in producing the results. The main steps in data research are: collecting and organizing the data, changing the focus and considering the interpretation of the results, and finally presenting the results in the form of written reports and/or success codes. Here are the simple steps to search data







III. TOOLSOFDATASCIENCE

1) Python

Python is a powerful, flexible language that is easy to learn, easy to use, and has a powerful library for data management and analysis. The syntax is simple, easy to understand for new programmers, and will be familiar to anyone with experience in Mat Labs, C/C++, Java, or Visual Basic. Over the years, Python has been used in computational science and many other fields such as finance, oil and gas, physics, and signal processing. It was used to develop the space shuttle's mission design, process images from the Hubble Space Telescope, and play a key role in collaborative physics experiments that led to the discovery of the Higgs boson, known as the "God particle." Python is one of the most popular programming languages
in the world, ranking higher than Perl, Ruby and JavaScript. Among today's languages, the power and productivity of Python-
based solutions are legendary. The future of Python depends on how many providers allow the use of the Python SDK and which Python modules expand the Python application portfolio.

The R Project for Statistical Computing

R is an excellent alternative to statistical software such as SPSS, SAS and Stata. It is compatible with Windows, Macintosh, UNIX and Linux platforms and provides an open source language and computing environment. The R environment provides software space from data management, computing to graphics.
Users can define new functions and manipulate R objects with the help of C code. So far, eight packages are available for users to use statistical methods. However, many routine statistics can be made with the help of CRAN series websites.
No license restricts anyone from contributing to code development or providing disease information.

R is an integrated set of software tools for data management, calculation and graphics. This includes:

- Effective data and storage.
- A set of operators for evaluating arrays (especially matrices).
- A large, comprehensive collection of intermediate tools for data analysis.
- Graphics equipment for analyzing and displaying data on the screen or in hard copy

- A well-developed, simple and effective programming language which includes conditionals, loops, user-defined
- Recursive functions and input and output facilities.

2) Hadoop

The name Hadoop has become synonymous with b ig data. It is an open source software framework fo r distributed storage of large amounts of data acros s clusters of computers. The relationship between d ata management and data analysis All this means y ou can scale your data up and down without having to worry about hardware failure. Hadoop provides massive storage, massive processing, and the abilit y to manage unlimited workloads or processes for any type of data. Hadoop is not suitable for beginn ers. You need to know Java to use its real power. It may be a commitment, but Hadoop is definitely w orth the effort since tons of other companies and te chnologies use it or integrate with it. However, Ha doop MapReduce is a batch-oriented system and is not suitable for interactive use; real-time operations such as streaming; and more calculations.

3) Visualization Tools

Data visualization is a modern branch of descri ptive statistics. It involves creating and examini ng visual representations of data; this means "d ata abstracted in some schematic form with attri butes or variables of data units." Some tools are software that use very different logic than usin g programs for data analysis. Think about the fi rst GUI that made computers popular with the p ublic, and suddenly everything becomes repetiti ve. A "pretty" paper is useless if it looks nice b ut doesn't tell you anything. But sometimes ma king information good and easy to understand c an also make it understandable to the average p erson. Tableau occupies a niche by allowing no n-programmers and brands to seamlessly extract data, perform fast searches, and create powerful plans (with interactive, interactive things, etc.) very quickly. D3: You should use D3.js becaus e it allows you to create the visual information you need. Graph/data visualization frameworks work well of decisions to make the framework easy to use. D3.js focuses on binding data to DOM elements. 3 stand for Data Driven information. We will explore the potential of D3.js. Dat a Wrappers: Data wrappers allow you to create charts a nd graphs in four steps. This tool cuts the time you need to create visualizations from hours to minutes. It's easy to use, all you have to do is upload your file, select a d ocument or map and print. Document wrappers are desi gned to fit your needs; The layout and visualization can be customized according to your style guide.

4) Paxata

- Paxata focuses on data cleaning and planning r ather than machine learning or modeling. The app is easy to use and its visual guidance make s it easy for users to assemble files to solve the problem, find and repair lost or damaged files. Information can be shared and reused with oth er groups. Suitable for people with limited pro gramming experience to manage data searches .

Here is the process provided by Pixata:

- Add new information from retrieve data from various sources.
- Any gap in the database can be detected during database search.
- Users can group or summarize data.
- Easily combine multiple data sets into a single
- response with Paxata's unique Smart Fusion tec hnology.
- Only one of them and together they will decide what is best.

IV. APPLICATIONS

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

Some people think that data science is just a new term for business analysis, that business analysis was a field that quickly emerged a few years ago and was later replaced by data new term science. Although the two fields can be considered independent of each other, there is no doubt that data science is frequently used in business marketing.

2. Prediction:

Large amounts of data collected and analyzed can be used to identify patterns in the data, which can then be used to develop predictive models. This is the basis of machine learning, where inductive algorithms and other so-called "learning" algorithms are used to find information. Machine learning techniques are frequently used to develop predictive models in many fields.

3. Security:

Information collected from user records is used to detect fraud using the search engine. Patterns detected in user activity can be used to isolate fraud and malicious internal events. Banks and other financial institutions often use data mining and machine learning algorithms to prevent fraud.

4. Computer Vision:

Information obtained from image and video analysis is used to use computer vision; it's the science of making computers "see," using image data and learning algorithms to capture and analyze images and determine compatibility. It is used in robots, autonomous vehicles and human-machine interaction applications.

5. Natural language processing:

Modern NLP technology uses large amounts of data, from data corpuses to model language data, and uses these models to perform tasks such as machine translation, parsing, natural language generation, and sentiment analysis.

6. Bioinformatics:

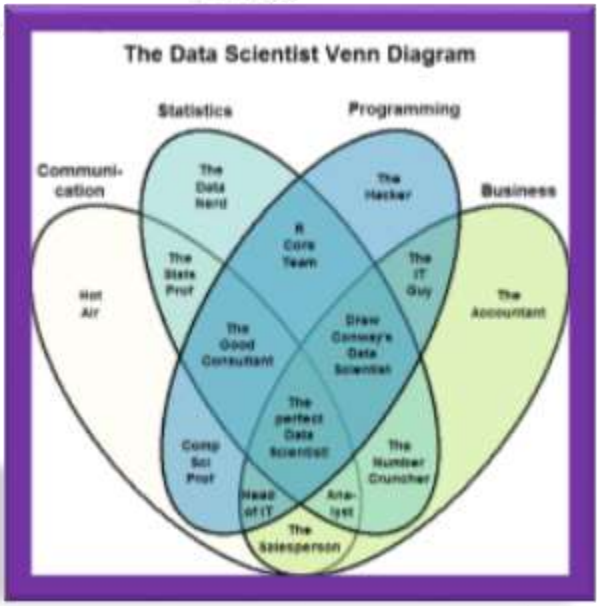
Bioinformatics is a rapidly growing field in which computers and data are used to understand biological data such as genetics and genomics. These are used to better understand the origin of the disease and the need for genetics and other biological materials. As Michael Walker points out, next-generation genomic technologies are helping data scientists augment genomic data collected from large populations. When genomic information is combined with new information that provides different information in disease research, we will better understand drug response and the genetics of the disease.

7. Scientific research: Scientific experiments, such as the famous Large Hard Collider project, generate data from millions of sensors that must be analyzed to draw conclusions. Astronomical data from today's radio stations and weather data stored at NASA's Climate Modeling Center are other examples of large-scale scientific data that tends to expand into new areas in file size.

8. Income Management: Instant income management also receives a lot of help from information experts. In the past, the revenue management process was hampered by a lack of data points. Data science is also used in the retail or gaming industry. Wang Jian puts it: Revenue management is the method of selling the right products to the right customers, at the right price, at the right time, with the right policy to ensure the total revenue of the enterprise. -Data scientists can now leverage real-time data streams and adjust their messages accordingly. It is now possible to predict the most profitable type of business that can be grown at the moment and how much profit will be made at a given time.

9. Government: Government agencies are also using data science to prevent waste, fraud and abuse, combat cyber attacks and protect sensitive data, make better financial decisions, use AI work to improve defense and protect ground forces. In recent years, most governments have realized the great utility of scientific models for various projects.

The Data Scientist Venn diagram



V. CONCLUSIONS

Better analysis of the vast amounts of data available th rough data science has the potential to drive advancem ents across many disciplines and increase the efficienc y and success of many businesses. However, before thi s potential can be realized, many of the challenges des cribed in this article need to be addressed. Challenges include not only the obvious issues of scale,but also all stages of the review process, from data collection to interpretation of results, such as inconsistencies, lack o f standards, errors, privacy,time,evidence and visibility . Because these challenges exist in many application d omains, it is not cost-effective to address them in just one domain. Additionally, these challenges require change and are not solved with the next product. If we want to realize the promised benefits of big data, we m ust support and encourage important research to solve these challenges.

The truth is that in the future there will be those who tr y to use scientific data for every problem, some of the m are overused. But in addition to online apps (recom mendations, ad targeting, etc.), we have a feeling we'll also see some great apps for DS aimed at average user s. Visualization, customer engagement, and engineerin g sales algorithms all require different skills. It would be nice if we could do everything with maximum perfe ction. But if demand is strong enough, companies will begin to take on different roles and create teams with expertise, rather than assuming that one person can co ver everything.

REFERENCES

1) Parsons, MA, MJ Brodzik, and NJ Rutter. 2004. Data management for the cold land processes experiment: improving hydrological science.HYDROLPROCESS.18:3637-653. http://www3.interscience.wiley.com/cgi-bin/jissue/109856902
2) Data Munging with Perl. DAVID CROSS. MANNING. Chapter 1 Page 4.
3) Blog post: Thoran Rodrigues in Big Data Analytics, titled “10 emerging technologies for Big Data”, December 4, 2012.
4) Proyag Pal1, Triparna Mukherjee, ”Challenges in Data Science: A Comprehensive Study on Application and Future Trends” published in international Journal of Advance Research in Computer Science and Management Studies , Volume 3, Issue 8, August 2015
5) What is Data Science? http://www.datascientists.net/what-is-data-science
6) The Data Science Venn Diagram. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
7) Tukey, John W. (1977). Exploratory Data Analysis. Pearson. ISBN 978-0201076165.
8) Peter Naur: Concise Survey of Computer Methods, 397 p. Student litteratur, Lund, Sweden, ISBN 91- 44-07881-1, 1974
9) Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyt,” From Data Mining to Knowledge Discovery in Databases. . AI Magazine Volume 17 Number 3 (1996)
10) “Research in Big Data and Analytics: An Overview” International Journal of Computer Applications(0975–8887)Volume108–No14, December2014
11) Douglas, Laney. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012
12) T. Giri Babu Dr. G. Anjan Babu, “A Survey on Data Science Technologies & Big Data Analytics” published in International Journal of Advanced Research in Computer Science and Software Engineering Volume 6, Issue 2, February 2016
13) Jeff Leek (2013-12-12). "The key word in 'Data Science' is not Data, it is Science". Simply Statistics.