Review on Data Science and Prediction

Abdulwahab Alazeb   
*dept. Computer Science  
University of Arkansas*  
Fayetteville AR, USA *Najran University, Najran, KSA*  
afalazeb@{uark.edu, *nu.edu.sa}*

Mohammed Alshehri  
*dept. Computer Science  
University of Arkansas*  
Fayetteville AR, USA *Najran University, Najran, KSA*  
msalsheh@uark.edu

Sultan Almakdi  
*College of Computer Science and Information Systems  
Najran University, Najran, KSA* smalmukadi@nu.edu.sa

*Abstract*—The exponential growth of data over the years has brought into focus the increasingly-important task of managing all of it. This need has been driven by advancements in technology that have created large increases in data volume. Due to the societal importance of many newly-developed forms of technology, management of data is no longer just a matter of storage, but security and availability as well. In the business world, the management of data affects the delivery of services and overall productivity. This paper seeks to explain the different aspects of data science related to modern needs and future significance. Data science is explored in detail with emphasis on the background, history, and concepts of data management. Various sources from existing literature and academia are thoroughly discussed regarding the growing importance of data management and techniques involved. An evaluation of these techniques is also presented.

Keywords— Data Science, Big Data, IoT, Data Prediction.

# Introduction

Data science is study of the extraction of valuable information and its transformation into usable forms. This could also be viewed as the practice of developing valuable insights from existing information. Whatever the perspective, data science has been very effective in combating the challenges faced in processing huge data sets commonly referred to as Big Data. The term Big Data is referred to the structured, semi-structured, and unstructured data produced by large-scale enterprises. In recent times, a veritable explosion of data has occurred due to smart devices, social media, and the web.

In order to handle the many aspects of data science, data scientists and experts are needed. They handle the statistics, software engineering, and visualization of Big Data. Data scientists often focus on areas of finance, fraud, security, medicine, and marketing. They largely depend on signal processing, machine learning, natural-language processing, statistics, and text retrieval for data analyses and result interpretations. Additionally, data experts interact with parallel processing systems, map-reduction computing, and petabyte-sized SQL databases.

The significance of data science became evident many years ago due to the increasing demand for data. Nearly every aspect of human life became interconnected with large collections of information, creating a parallel necessity for people who could manage and track the collection and handling of this information. For business purposes, these people primarily extract valuable information to solve business problems. Particular business interest in data science has been in the health care sector in regards to providing valuable medical services. From the viewpoint of engineering, the incredible scale of Big Data has proven that many database models are outdated in regards to the discovery of knowledge within such databases. These outdated models were created for data summarization and quick access and not for knowledge discovery [3]. They are optimized for user queries and not discovery of various complex patterns that are not well-formed in large volumes of data.

Database discovery utilizes inquiries into patterns satisfying certain criteria while data querying simply locates information based upon the specific query. Database discovery is chiefly concerned with finding robust and interesting patterns [2]. To address this issue, firstly educators must first rethink their ways in passing knowledge of skills and program design. Secondly, decision-making skills must be revamped regarding the management of Big Data. Thirdly, scientists must discover improved ways of exploiting big data and optimizing computational power for scientific enquiries. Fourthly, scientific enquiries and processes must be updated to incorporate new ways of thinking into traditional methods of conducting scientific enquiries. Finally, scientists must find ways of augmenting their viewpoints regarding innovation and discovery.

As was previously mentioned, data science is important in the health care sector, especially in regards to clinical trials. Gathering clinical data and processing it in a timely manner greatly aids in providing answers to many medical problems. Data science is very helpful in clinical trials as it helps illustrate the safety and efficacy of existing and upcoming compounds. Clinical data scientists develop statistical analyses of clinical trials in order to aid in the development of health care applications. Clinical data scientists have become integral to clinical trials in planning the collection of data and transforming such data into useful results that can be summarized and communicated to professionals and consumers.

The need for effective data science has become even more critical due to the fact that data has become spread across wide networks, often in a heterogeneous and unstructured fashion. Much of the unstructured data could be in the form of text, images, videos, and sounds. Research has shown that the volume of data has grown exponentially, which has led researchers to the estimation that there would exist more than 125 billion connected IoT devices by 2030 [1] that can produce in excess of 79 zettabytes of data daily [20, 21]. The analysis of all of this data requires integration, interpretation, making sense of it all combined with the tools of computer science, software engineering, linguistics, econometrics, sociology, and other disciplines in order to achieve proper data management [2]. Markup languages and tags are needed to provide computers with the ability to interpret data automatically and assist in crucial decision-making. Early markup languages, such as HTML, were designed to focus on the presentation of information in a form that human beings could understand. Today, markup languages are being oriented to computer understanding to aid in the automated processing of information from different resources so that extracted information can be used to make automated decisions. Often it is necessary to be able to determine if extracted information contains a workable solution for the future based-on a determination of reliability from past data. This results in the predication of trends of the near future [2].

# Background: Data Science

The concept of data science was introduced in 1960 by Peter Naur who used the term at that time instead of computer science, indicating that computer science was originally known as data science. In 1974, Naur published the book “Concise Survey of Computer Methods”. In his survey Naur used the term “data” frequently while surveying contemporary data processing methods. Some members from the International Federation of Classification Societies organized a conference in 1996 in Tokyo in which the term “data science” was first given the proper recognition. The conference name was “Data science, classification, and related methods”. In 1998, C.F. Jeff Wu gave a lecture entitled “Statistics Data Science” in which he emphasized the inclusion of statistics in science. He showed the world the importance of statistics in terms of data science by describing the treatment of statistical data as data collection, data modeling and analysis, problem solving, and decision-making.

William Cleveland furthered existing research and presented data science as an autonomous discipline, proposing the extension of the field of statistics in combination with technology of computing devices in his article about Data Science in 2001. Cleveland claimed that data science should be incorporated with other fields of science in order to collaboratively form “Data Science”. The fields of focus were multidisciplinary studies, and schemes for data, data computation, education, evaluation and assessment methods, and theoretical framework [3]. In 2002, the International Council for Science Committee on Data for Science and Technology (CODATA) began the Data Science Journal focused mainly on issues that appear in research regarding data science, such as details of data systems, the publication of such details over the Internet, applications of data systems, and legal issues concerning data systems.

Early in 2003, Columbia University launched its “Journal of Data Science” with the primary goal of presenting data scientists worldwide a platform upon which to share ideas and views and collaborate with other scientist for research purposes. One of the founding purposes of the journal was to collect research on the application of statistical methods on quantitative research. Two years later, the Board of National Science issued "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century" in which they identified themselves as data scientists, but stated that they can be pointed to by other titles such as computer scientist, software programmers and database, and field experts along with expert annotators, librarians, and archivists. Those with these titles will be important in the successful management of digital data collections. The Board identified these people as those who are able to conduct creative inquiry and analysis [3].

In order to understand the meaning of “Big Data” in the field of data science, we should begin at the start of the 21st century when the term was first introduced and spread with many organizations embracing it. Many of these organizations were based online or startup companies. The ability to run business related to Big Data and simultaneously research the field offered a symbiotic seed from which to begin. Giants firms like Google, eBay, LinkedIn, and Facebook actually had the “Big Data” concept in mind because they started operating on the basis of analyzing the data gathered through their websites [3]. These were not traditional firms from their beginnings. They did not need to integrate Big Data technologies with traditional IT infrastructures because they never had traditional IT infrastructures. Big Data can by itself stand alone, allowing many firms to primarily focus on analytics and architectures.

## Big Data Integration and Prediction

Big Data is primarily concerned with analytics, making it inherently merged with the field of analytics. Many researchers have the idea that data scientists need to work their way through all the quantitative tools that are present [16]. One of the scientific methods used in the field of data science is the analytics of data, which shows the actual value of a business and hence how to help the business flourish. Analytics used in the past are not concise enough to give businesses revolutionary perspectives. Today the proper application of data is far more powerful than the previous analytics methods could accommodate. Fortunately, the developing methods of data science have allowed more precise collection of data and thus more powerful analytics. This results in more intelligent predictions made from Big Data resulting in better decision-making. Big Data technologies can help us target more effective interventions and improve analyses, predications, and measurements in areas that were not possible a few years ago. It seems likely that Big Data will continue spreading worldwide with developing tools, research, and philosophies that will change ideas and make comfortable facing evolving challenges in data management. Smart industries have already begun using new Big Data technologies in their efforts to change the world in many perspectives and grant people the ability to make predictions that were no possible until recently. This does not mean that Big Data is without challenges. Companies seeking to adopt the philosophies of Big Data need to do so at the organizational level, which can be quite challenging [16].

In 2012 the world possessed almost 2.5 exabytes of previously-generated data and that amount has grown steadily ever since. Statistics have shown that the global quantity of data on the Internet is now increasing more rapidly each second than the entire quantity of data present on the Internet twenty years ago [16]. Whereas previously organizations would have been working with megabytes or gigabytes of data, they are now working with petabytes of data. Due to parallel and distributed processing, this large quantity of data can sometimes be manipulated as a single entity rather than as a large workload. It is not just Internet data that has seen strong application of data science. Research shows that data gathered by Walmart every hour from consumer deals and businesses could be more than 2.5 petabytes. This quantity of data is far too large to handle manually without unreasonable coordination and expense, which is yet another driver of applying the latest data science technologies.

Perhaps what has shocked many regarding the current developments in data science is the speed of data creation, which is effectively not measurable. Speed is often considered to be more relevant than size when it comes to data because it is the agility of approaches and not the scale that lead to sizeable advantages over competitors in the business world. Real-time information is needed nowadays to make many businesses successful. To look at an example of how Big Data can affect business, Alex Pentland with his crowd at the Massachusetts Institute of Technology media lab used location data from an array of mobile phones to calculate how many people were present in Macy’s parking lots on Black Friday. Right at the beginning of the Christmas shopping season in the United States, his group had estimated the sales of all stores before Macy’s itself had recorded anything. This is how Wall Street and Main Street get ideas of trends followed by checking out the data present in the market. [3] It is not just the massive quantity of data that is of focus, but the numerous varieties and multifaceted nature of the data. All of those zeroes and ones are now messages, updates, images, sounds, videos, and more. The refresh and pictures which are posted on social networking sites, are a tremendous source of Big Data. Another source is GPS signals from smart phones. Often forms of Big Data are inconsistent, heterogeneous, and complex in nature.

The many forms and types of Big Data that exist have added considerable complexity to the tasks of traditional tools. Many data resources are newer than the concept of Big Data itself. For example, Facebook, launched in 2004, and Twitter, launched in 2006, were founded well after the conception of Big Data. Due to the enormous size of Big Data and the lack of existence of past records for reference, structured infrastructures are not able to hold and process the quantity of information. Ironically, the reduction in cost of computing devices has led to an even wider gap in the generation of data versus its processing. Traditional methods are simply unable to keep pace. When it comes to the question of whether Big Data intelligence can help improve business performance, the answer is obvious and proven by data [16]. The research journal article by Andrew et al. entitled “*Big Data: The Management Revolution”* outlines an experiment organized by a team at the MIT Center for Digital Business who worked in collaboration with McKinsey’s business office and their partners from Wharton and a Ph.D. student from the Massachusetts Institute of Technology. Their main agenda was to test the hypothesis that “*data-driven companies would perform better*”. They conducted interviews with company executives in approximately 330 public North American companies regarding their organizational structures and technology management practices. They gathered performance data from the companies’ annual reports and from independent resources. The group discovered that not everyone was interested in having a data-driven organizational policy. From the data they collected from the companies, the group also realized that one aspect stood out clearly: companies that implemented data-driven policies performed better financially and operationally. These companies typically achieved 5% higher production and 6% more profit than their non-data-driven competitors [16]. Their research showed that when data is used intelligently, it will be reflected in the area of application. They took a further step to discover how managers of different companies far from the Silicon Valley are responding to trends in Big Data. The companies that were not fully-versed technologically had to increase their business or expand their sales.

Other researchers have also developed different ideas of using Big Data to predict different aspects of research. An example of how big data is able to help predict changes in prices of different objects involved simply collecting search data that is readily available. Researchers Eric et al. were able to customize the free web search data to forecast how housing prices would change in metropolitan areas across the United States. Although they were not experts of the housing market, after their research they were able to deduce from virtually real-time search data short-term predictions regarding the housing market. They presented their research with so much precision that their results were more accurate than the official results of the National Association of Realtors, which had a very complex model for prediction. The researchers simply relied on relatively slow-changing historical data [15].

There are many cases that show how Big Data driven analytic approaches can produce excellent results [16]. Some scholars observed that they were able to use data from Google flu trends to predict the flow of patients in flu linked to urgent clinic visits by at lease days before the Center for Disease Control sends alerts . Likewise, looking at news articles from the last five years related to Big Data, it can be seen that Twitter updates were very accurate in tracking the spread of cholera in Haiti after the 2010 earthquake according to the official reports that were presented afterwards, two weeks prior to their release. There is no shortage of cases where data science and its related predictions were as-accurate or more accurate than the official reports issued later. There seems to be no limit to the variety of predictions that can be made using data science [4].

## Big Data Challenges

A number of different research efforts, such as Fulgoni, have suggested that today organizations are facing huge explosions of data. Per the research conducted by Enterprise Strategy, the respondents of Group 1 stated that they were facing an 11% to 30% increase in volume of data with 28% of the respondents facing a 30% growth in data volume [4]. With such a large amount of data being crated and replicated, IT departments are struggling to find new ways of managing, storing, and safeguarding such huge volume of data. The modern day trend is to leverage Big Data to gain competitive advantages and assist organizations in achieving their goals. New types of information are being generated, such as seismic exploration outcomes, pharmaceutical trial data, and website comments. Such information is being accumulated and sifted for answers and insights [5]. For instance, the Swiss Institute of Bioinformatics generates huge volume of data from its genome-sequencing process. One experiment generates 743,000 files for every run with each run every 3.5 days generating about 2TB [6]. A study by Hawkins et.al asserts that organizations are converting much of their content into digital format. Data in the form of video has become a significant communication, training, and marketing tool. High definition and special effects also enhance the volume of data generated and much more space is required for storing a 3-D video as it requires two cameras for shooting the same footage [6]. Various gaps have been identified in IT architectures regarding supporting advanced technologies.

Today, governments, non-profit organizations, and businesses have realized the significance of data as their IT departments are facing rising challenges. Per study, growth in data is the most critical challenge in association with data center hardware infrastructure for big enterprises [7]. There are various challenges associated with developing strategies that support the dependence of an organization on data. There are integrity concerns, availability requirements, complexity sprawl, and budget constraints. Today the economy is quite uncertain and organizations want to make optimum utilization of their resources and not make huge investments in new technology. IT organizations have been forced to make their technological infrastructures more complex due to strained technological talent. Thus, they are seeking new ways for empowering end users with user-friendly tools . Organizations need to ensure easy access to data for their employees irrespective of the place where it is stored. They believe that their data will be unaltered and intact, requiring IT departments to provide such outcomes [6].

## Big Data Today

Today, the IT world is largely focused on Big Data and the multiple meanings associated with it. Big Data has many new forms and variations, including collections of small volumes of data like traffic camera feeds, underwater ocean-floor photographs, and comments from social networking sites. Uncertain data stores of particular information types also grow rapidly into Big Data. Describes that today organizations need huge storage capacities to manage big files [2]. Big Data like any other information technology brings about cost reductions and improvements in service offerings, new products, and time needed for performing computing tasks. The concepts and technologies supporting Big Data permit organizations to reach their objectives, not only financially, but also regarding the processes associated with Big Data [19].

As explained by Biesdorf et.al organizations that pursue technologies associated with Big Data strongly believe that structured data can most economically be delivered via Big Data technologies such as Hadoop clusters. For example, cost comparisons of one company estimated that storing one TB would cost about $37,000 annually for an outdated relational database and $5000 for the database appliance, but only $2000 for a Hadoop cluster. However, Data security methods have not yet been developed completely for Hadoop clusters [19]. Once its primary objectives are accomplished, an organization might examine ways to improve its economic efficiency. One such case is that of GroupM, the popular media-purchasing subsidiary of WPP, an advertising company. It is the biggest worldwide buyer of media and deploys Big Data tools [3]. The key problem GroupM faces is that it has 120 offices globally and every office has its own approaches and technologies. If the company permitted each of its office to deploy its own tools for Big Data, it would cost about $1 million for every site. GroupM deployed centralized services for Big Data from its office in New York. With 25 markets worldwide, it expects that it will cost only about a third of the amount that it currently spends with a decentralized approach [2].

## Developing Big Data-Based Goods and Services

The most innovative motivation for a company regarding Big Data is in its use for the development of new goods and services. Most companies like this are usually online firms that need to implement data-based goods and services [9]. For example, LinkedIn has been using data scientists and Big Data for improving multiple characteristics of their prodacts such as Jobs You May Be Interested In, People You May Know, Who’s Viewed My Profile, Groups You May Like, etc. Such offerings have helped in bringing millions of consumers to LinkedIn [2]. Google is another major contender developing goods and services using Big Data. It mainly employs a lot of algorithms and mechanisms for Big Data. The company is also developing new and innovative goods and services comprised of Big Data algorithms for its ad-placement and search functions for Google Apps, Google Plus, and Gmail. Its self-driving car has been referred to by the company as an application of Big Data. A few of its product developments failed and a few others paid-off, but in reality there is hardly any more prolific offerings creator than Google. GE is another significant creator of new offerings using Big Data. It mainly emphasizes optimization of maintenance intervals and service contracts for industrial goods [15].

A few other companies, such as T-Mobile, Sprint, and Verizon Wireless, are already selling services depending on the location and data from mobile devices. For example, the Precision Market Insight offering by Verizon evaluates the effectiveness of external media locations, venue audiences, and retail store locations. Netflix Prize is a popular offering developed by Netflix by its data science group for optimizing the movie recommendations given by the company to its customers [14]. Per the study by Schultz, Big Data is being used for creating proprietary content. It is being used by ‘Kaplan’, a testing firm, to advise its customers about test-preparation strategies and effective learning [14]. The Big Data offerings of the company focus directly on goods, services, and end users. An organization that is serious about generation of its goods and services using Big Data must develop a wide-ranging initiative or program to do so. It must be comprised of a set of technologies, people, and tools well-suited to manipulating Big Data.

# Methods

Many methods of data science have been used over the years since data science became a field. Scientists have worked tirelessly to achieve various beneficial results from different technologies using data science techniques and approaches to Big Data. One such method is related to the “Web Usage Mining” technique to gather data based on user access patterns of web servers. Many companies gather a massive amounts of data everyday, often produced repeatedly by web servers and gathered in the logs of the server. In the paper “*Improved User Navigation Pattern Prediction Technique from Web Log Data*” written by *V. Sujatha and Punitha Valli,* an ideology of “*Prediction of User navigation patterns using Clustering and Classification (PUCC)”* was considered based on this issues [17]. They worked in stages where the first stage of PUCC focused on separating the potential users from the web log data and the second stage of PUCC used a clustering process in order to make groups of different users based on the characteristics of similar interests. In the final stage, PUCC predicted the future navigation patterns of users. The research showed that using this approach can improve the quality of business in huge web sites by predicting users’ next requests and being prepared for them.

Web mining is generally alienated according to the data type. Three types of web mining are available in the industry: the content of web mining, the usage web mining, and the structure of web mining. There are many analysis tools available on the market to mine the data, however the efficiency of these tools is still in question as Big Data is very huge. Clustering and classification are two very important areas of machine learning research that are helping data scientists to work more accurately. Clustering separates Big Data of web logs into smaller groups depending on certain predefined criteria like similar interest, race, location, etc. Classification helps data scientists understand the categories of users presented in web logs [17]. The general process of web mining is comprised of four stages: resource discovery, information pre-processing, generalization, and pattern analysis. Resource discovery is used to retrieve information from the Big Data available in web log files and other available documents. Information pre-processing converts the discovered information into meaningful data. The process of revealing the general repetitions of users across various websites is called the generalization. Pattern analysis is a technique to validate the mined patterns. The problem of analyzing user navigation patterns comes under the domain of web usage mining [17].

Identifying user navigation patterns through web log data has been very popular among data scientists as a research topic. Baraglia al et. introduced a Web Usage Mining system named “SUGGEST” that was able to provide useful information in order to make web user navigation easier and optimize web server performance at the same time. Liu and Keselj proposed an automatic classification of web user navigation patterns in 2007 that classified user navigation patterns and predicted future navigation patterns of users. Jespersen et al in 2002 proposed a hybrid approach that was able to analyze the clicking sequences of users who visited a particular site. Jalali et al in 2007 and 2008 proposed an automated system for discovering user navigation patterns through a graph portioning model. In the model there was an undirected graph built based on the connections between each pair of web pages and weights were assigned to all links of the graph [17]. The architecture proposed in the paper was a generalized PUCC (*Prediction of User navigation patterns using Clustering and Classification*) system. The main component of the PUCC system is web registry data because data is the actual fundamental component that gives knowledge regarding navigation patterns. The web data registry stores all successful hits that reach the internet. The word “hit” is expressed as an inquiry that would assist a user to display any valid media or the HTML. The data in the registry file of the Web can be obtained from any client-side or proxy server or from an organization’s database server upon request. The registry consists of data in form of an IP address of the computer making the request, user ID, date and time of the request, a status field that shows whether the request was successful or not, size of the file transferred, referring URL of the page that contains the link that generated the request, and the name and version of the browser that was used for this transaction.

A web registry file is created automatically and maintained by a server, another important aspect of the proposed PUCC server. The web registry file is used to save every transaction that reaches the web server including each view of an HTML document, image, or other object. There are two log file patterns available. One is known as “Common Log File” (CLF) and the other is known as “Extended Log File” (ELF). Whichever format is used, the log file is key to locating information regarding navigation patterns of different segments of the full web traffic. Log files can save information related to single users, multiple users, single-site browsing, and multi-site access patterns. Log files are ordinary text files with uniform data that show the requests made to a server. The first stage of PUCC is the pre-processing of web log data in which the unstructured, heterogeneous data is converted into a form suitable for mining. The formatted data from this stage will not be understandable by researchers, but it will help in completing the next stage of the process. The primary purpose of the pre-processing step is to prepare the data so that users and sessions can be identified in the next stage [17]. The next stage focuses on identifying potentials users from a large pool of users. Researchers have used a variety of different methods and procedures over time to perform this identification. Suneetha et al. employed the decision tree classification algorithm for this purpose along with a set of decision rules. Their algorithm was an improvement over previous works, but neglected entries made by the robots, which affected their results [17]. The authors tried to resolve this issue by specifically identifying the entries created by robots in the log file and deleting them before separating users into potential and non-potential categories. They were able to identify the robotic entries by referencing the “robot.txt” file that is always created at a website’s root directory. Having removed the robotic entries, the authors calculated browsing speed by dividing the number of pages viewed by a user by the length of the session time of that user. The decision rule that the authors used to identify users was to check if the session time exceeded thirty minutes, the number of pages accessed exceeded five, and the HTML method was POST. If all of these conditions were met, a user was classified as potential. Otherwise, the user was classified as non-potential. All of this was accomplished by reducing the log file to allow for effective cluster portioning and prediction. In regards to cluster portioning, the authors used a graph portioning for grouping users with similar navigation patterns. When Jalali *et al* used this method in 2005, they utilized a weighted unidirectional graph to connecte each web page to its pair set. The authors of the aforementioned paper assigned the likes weight to every link on the graph. Connectivity time was used to measure the degree of visits between two pages during a session. The new technology introduced by this work was the prediction engine utilized within the architecture to classify user navigation pattern and predict users’ future requests. Predictions were accomplished utilizing the “Longest Common Subsequence” algorithm to find the longest subsequence common to all sequences in a set of sequences. This is accomplished by initial realizing that if two sequences both end with the same item, their longest common subsequence can be establish by eliminating the last item and then discovering the longest common subsequence of the shortened sequence. Once the shortest such sequence has been attained, the longest common subsequence is found by finding the longest common subsequence of the two shortened sequences with the last character of the first removed or the last character of the second removed, whichever is longer.

Other methods of data processing for data mining purposes exist in regards to Big Data. In the paper entitled “*Predicting emerging technologies with the aid of text-based data mining: the micro approach*” by Dr. Smalheiser, it is proposed to do something with the flood of text data worldwide. Text data mining is used to create new technologies and new uses for existing technologies. It is focused on merging different domains, fields, and technologies to obtain fresh and more beneficial results from the existing technologies [18]. The focus of the paper was in predicting genetic engineering technologies that might impact viral warfare in the future by investigating a mixture of conventional Medline searches and utilizing a bundle of well developed methods well-known as Arrowsmith. The paper clearly showed that genetic packaging technologies, such as DEAE-dextran, cationic liposomes and cyclodextrins, can influence the effects of infections caused by different viruses delivered via an aerosol route. Text data mining is classified into two categories: macro and micro. Macro techniques perform analysis by crunching large sets of data in order to recognize trends that occur on large scale to classify and organize the information. Micro techniques utilize complementary information that connects to specified fields of inquiry. In a paper by the later, micro technologies were shown to be able to be implemented in order to aid in making policy decisions regarding technical innovations. The authors divided their work into steps. The first step was to define the problem based on two subsets of the literature: viral warfare and genetic engineering. These subsets were deemed likely to be implicitly related to each other in a complementary way. The second step was to retrieve existing literature regarding ways that could alter the aerosol stability of viruses [18].

## Data Science: Technical Approaches and skills

A study by Mozafari et al highlighted that today machine learning skills have become an integral part of data science because companies direct the data flood and are making efforts to develop automatic decision systems emphasizing predictive accuracy. Thus, even the most basic machine learning course is necessary for today’s competitive markets. Knowledge of text mining and text processing have become increasingly valuable because of the recent text explosion and unstructured pool of data in social networks, healthcare systems, and other areas. It is important to have deep understanding regarding tag languages, such as XML and its derivatives, because content today is frequently labeled hence data will be interpreted robotically [13]. Data scientists must learn more about machine learning along with basic skills that can be classified as statistics, particularly Bayesian statistics that are comprised of knowledge regarding multivariate analysis, hypothesis testing, distributions, and probability. This knowledge can be gained in up to three course sessions. Econometrics and multivariate analysis are usually overlapped as econometrics focuses on alignment of statistical models with economic data [12].

A study by Null in 2013 stated that data science requires skills in the field of computer science in focusing on ways in which computers represent and manipulate data internally. This requires a number of classes based on OS, Algorithms, and Data Structures [11]. System skills along with scripting languages are the basic building blocks needed for managing datasets of considerable size. It should be kept in mind that standard database systems based on relational data models have various limitations in managing large datasets [10].

The high demand now is cloud computing and how efficiently and effectively handle the massive datasets, which needs professional skills for the specialist. To this effect, it is necessary to have knowledge regarding causation and correlation, which is the core of any modeling exercise that involves data. Observational data is usually limited to correlations. However, huge data can sometimes represent randomized trials and also represent the likelihood of calculating conditional probabilities that enable the discovery of a causal structure [11]. Causal models are highly desirable in domains where reasonable confidence is associated with the wholeness of a formulated stable model or where the causal model and its observed data are stable. Data scientists must be aware of the differences between causality, correlation, and the talent necessary to assess models that are practical, desirable, and feasible in various settings. Another set of skills necessary are less standardized and more elusive and might be classified as a craft. This set of skills requires data scientists to be capable of formulating problems such that effective solutions are derived [8].

Some problems are sometimes comprised of unbalanced target classes that denote dependent variables. Such cases are interesting to predict. These problems are challenging to model as require predictions that might be incorrect if the model is not good at differentiating between the classes. Such problems are well-known to experienced data scientists who must formulate them such that systems are capable of making accurate predictions [8]. This indicates that problem-devising skills will represent a key requirement for data scientists in the years to come. The phrase ‘computational thinking’, coined by Papert and further explained by Wing, is indicative of the necessary skills described. Modern day universities are training students to build problem devising skills with electives structured about the core suitable to particular disciplines.

As described by HSinchun, the revolution of data science is posing critical challenges on organizations. Other than identification and cultivation of suitable skill sets, it is necessary for managers to shift their considerations toward decision-making based on current data as an augmentation or replacement for past practices and intuitions. W. Edwards Demming was an American statistician of the 20th century who was quoted as saying that ‘In God we trust, everyone else please bring data’. This quote describes the new orientation of thinking that has shifted from intuition-based decisions to fact-based decisions. The decision-making today is relay on the Big Data where computers are acting as better decision-makers than people. Here, the term better refers to scalability, accuracy, and cost. It has already taken place in the data-intensive world of trade where most investment decisions are made by computers within just a few milliseconds of new data becoming accessible [7]. This is applicable in the case of online sales as multiple auctions take place in just a fraction of a second every day along with routing package delivery, controlling air traffic, and various types of planning tasks requiring accuracy, speed, and scale.

## Big Data Analytics and Insights

In 2013 Hoffmann highlighted that the evolution of Big Data in the last few years has increased the significance of data scientists. These scientists are savvy and inquisitive professionals capable of synthesizing large volumes of information, making them practically sensible at deciphering the meaning of data such that they can aid organizations in bringing about change [8]. WPI is one university that is well-prepared to train graduates to meet the challenges of this evolving field. It has a long history of carrying out research studies in data management, business analytics, statistics, and cutting-edge technologies. It is also now offering its first-ever Data Science Program in Massachusetts that is interdisciplinary across various departments based on important data science skills [6]. WPI students and faculty in areas ranging from data mining to actual analysis are developing efficient methods for accessing and using Big Data. This type of education aids in creating solutions to problems arising in different domains creating a significant social impact.

## Knowledge Discovery

A study by Dhar describes that today there are large volumes of data available and there is no need to make peace with wrong models. Predictions made by machine learning algorithms unassisted by humans have proven significant for many businesses [2]. Businesses like Google symbolize the victory of AI over the growth of top-down model. The language translator of Google cannot comprehend language and its algorithms cannot determine the webpage contents.

IBM’s Watson failed to comprehend projected questions and cannot use causal knowledge for determining the answers to these questions. In the same way, there are many unknown companies capable of predicting the odds of an individual responding to a particular displayed advertisement devoid of a solid theory [2]. An article by Davenport and Patil in 2012 stirred a vigorous debate among academic circles. Now the question is regarding the existence of predictive and scientific models not based on any theory. It has been observed that patterns evolve long before reasons for their existence are visible. These patterns seem to echo including specialists mainly in healthcare, marketing, etc. Now the question is about developing a predictive model using available data that is updated periodically when the problem in not stationary and the model is just an approximation [3]. When observations seem to indicate natural experiments, the data may express causality. For example, when a new drug is recommended for a sick individual, it will indicate biasness in the data.

Davenport and Patil observed that viewpoint is highly relevant in the fields of earth, social, and health sciences in the Big Data era. Because these fields lack significant theories, huge volumes of data serve as the basis for developing theories and understanding changes [3]. On the other hand, social sciences and physics contradict the predictive powers of theories. From the viewpoint of physics, a theory seems complete when the association between particular variables is capable of explaining the phenomenon without any exceptions. This type of a model is capable of making perfect predictions committing with some measurement errors not resulting from unintended consequences or omitted variables. The model is also capable making predictions in the presence of input changes. It is not enough to have a model that is 95% sure about the results with the remaining 5% subject to chance. While engineering is largely dependent on science, social sciences featured incomplete models that are partial guesstimates of truth and are usually based on simplistic assumptions about human behavior [4].

# Evaluation

Now we discuss the ways in which Big Data can operate with appropriate domains on robust grounds. It is possible to assume complete models for practical purposes. This brings about the likeliness of extracting causal models from huge volume of data. Until prediction errors become relevant, data serves as the right basis for developing a theory. For instance, a research scientist in the health-care industry remarked recently about an observed pattern for coronary failure that was triggered by a critical infection [5]. The main assumption of the scientist was that infection was the main cause for loosened plaque and inflamed arteries resulting in coronary failure. Other explanations could also be given, but if this observed pattern can be predicted, it can be enquired-into further and published. Now the question is if the Popperian test for predictive accuracy can be considered for future data favoring precise predictive models as the chief elements of future theory [5]. The earlier trend was to develop a causal model which was tested based on data. Whatever the case, it is important to discuss the accuracy of predictive models and the sources of their errors.

It has been believed that prediction errors result from three sources. The first of these sources is from misspecification of a particular method. This can happen for instance if a linear method is assumed for a non-linear phenomenon, resulting in errors due to the biasness imposed by the linear method. The other basis of commons faults comes from samples utilized for estimating parameters. Small samples increase the biasness of method estimates [3]. The third source of errors is in randomness despite perfect specification of the model. Big Data helps data scientists minimize errors from the first two sources. Big Data helps in considering models that make limited assumptions regarding the functionality of logistical or linear regressions because considerable data is available for testing the models and computing error. The second source of error can be eliminated by Big Data because the sample acts as a proxy for an entire population.

Observational data has a theoretical limitation evident in such examples regardless of the size of the data. Data is usually considered passive in that it represents actual happenings rather than things that might have taken place in varying circumstances. With respect to the field of healthcare, this is evident when observing the utilization of a health-care system in a passive manner, understanding its retrospective, and extracting its predictive patterns. It is fortunate when data is derived from the most accurate experiments [8]. However, alternate happenings are still not taken into account for situations in which alternative treatments would have been recommended to a particular patient. This results in failing to achieve a well-controlled and randomized experiment that would have otherwise established controls and determined the different impacts of treatments of matched pairs.

Big Data has one of the greatest application in the field of politics as demonstrated by the huge investments of Democratic National Committee (DNC) in analytics and data before President Barack Obama won the 2012. The DNC developed predictive models in the campaign based on the outcomes from big experiments, which helped in manipulating attitudes. This campaign helped in predicting the pattern in which every eligible voter would be voting and also ways of turning an individual into the type of person desired [2].

# Conclusion

Hypothesis-driven approaches and research for development of theories have been serving data scientists quite well. However, huge volumes of data are evolving today that thwart traditional approaches for identification of structures. These traditional approaches have failed to make the most of observations taking place in uncontrolled circumstances. Controlled experiments in the field of health care have helped in identifying various causes underlying diseases, but failed to reflect health complexities. Some estimates assert that clinical trials ignore 80% of the conditions in which a drug must be recommended. One such case is when a patient is taking multiple medications [5].

Big Data increases the possibility of uncovering causal models for data generation in situations where randomized trials can be designed. It has been shown that for diabetes related healthcare that Big Data increases the feasibility of a machine to enquire-about and answer questions that might not be considered by humans. The chief skill involved is to formulate predictive models for making active business decisions. Earth, ecological, and social sciences are the main data-starved areas where data brings about opportunities to develop theories and discover knowledge. These areas in the past did not have data of today’s scale and variety.

The landscape of data science in many areas is emerging and requires integrative skills on the part of data scientists. Data science is defined as the practice of developing valuable insights from existing data. This practice has emerged as a way of combating the challenges faced while processing huge data sets commonly referred to as Big Data. Academic programs in the fields of business management, engineering, and computer science provide valuable skills regarding the usage of Big Data, but the integration of many disciplines is necessary to produce effective data scientists [5]. Paypal has been highly successful in capturing and dominating consumer-to-consumer payments as it is capable of predicting the distribution of loss for every transaction and take suitable steps. This ability to treat transactions based on data contradicts traditional practices of handling transactions from a risk viewpoint. The most recent expedition of Google in the form of Knowledge Graph aims at understanding the entities equivalent to the strings processed by it constantly [5]. Google aims to understand the things behind the strings. The responsible administrations are struggling to capitalize on the phenomenal development of Big Data. In contrast, data has simplified the testing of established intuitions, allowing for cheap experiments and accurate decision-making. This requires a fundamental shift in an organization’s culture, as is evident by those organizations that have adapted to the evolving world of Big Data for efficient decision-making

##### References

1. Alazeb, A., & Panda, B. (2019, July). Ensuring data integrity in fog computing based health-care systems. In International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage (pp. 63-77). Springer.
2. Dhar, V.: Data Science and Prediction. Communications Of The ACM, Vol. 56 No. 12. (2013) 64-73.
3. Davenport, T. H., & Patil, D. J.: Data Scientist: The Sexiest Job Of the 21st Century. Harvard Business Review. Vol. 90 (10). Boston 70-76.
4. Fulgoni, G.: Big Data: Friend or Foe of Digital Advertising? Five Ways Marketers Should Use Digital Big Data to Their Advantage. Journal Of Advertising Research. Vol. 53(4), Oxfordshire, The United Kingdom (2013) 372-376.
5. Gertner, J., McCorvey, J. J., & Goodman, J.: Nate Silver Cover Story. Fast Company. Vol. 176. Harlan (2013) 70-154.
6. Hawkins, P., Rinard, M., Aiken, A., Sagiv, M., & Fisher, K.: An Introduction to Data Representation Synthesis. Communications Of The ACM, Vol. 55(12). New York (2012) 91-99.
7. Hsinchun, C., Chiang, R. L., & Storey, V. C.: Business Intelligence And Analytics: From Big Data To Big Impact. MIS Quarterly, Vol. 36(4). (2012) 1165-1188.
8. Hoffmann, L.: Looking Back at Big Data. Communications Of The ACM, Vol. 56(4). New York (2013) 21-23.
9. Jacobs, A.: The Pathologies of Big Data. Communications Of The ACM, Vol. 52(8). New York (2009) 36-44.
10. Meijer, E.: All Your Database Are Belong to Us. Communications Of The ACM, 55(9). (2012) 54-60.
11. Null, C.: How small businesses can mine big data. PC World, Vol. 31(11). San Francisco (2013) 67-70.
12. Nunan, D., & Di Domenico, M.: Market Research And The Ethics Of Big Data. International Journal Of Market Research, Vol. 55(4). Washington, DC(2013) 2-13.
13. Mozafari, b., kai, z., d'antoni, l., & zaniolo, C.: High-Performance Complex Event Processing over Hierarchical Data. ACM Transactions On Database Systems, Vol. 38(4), New York (2013) 1-39.
14. Schultz, B.: Big Data In Big Companies. Baylor Business Review, Vol. 32(1). Waco, Texas (2013) 20-21.
15. Xindong Wu; Xingquan Zhu; Gong-Qing Wu; Wei Ding.: Data mining with big data. Knowledge and Data Engineering, IEEE Transactions on , vol.26, no.1. (2014) 97- 107.
16. McAfee Andrew; Brynjolfsson Erik.: Big Data: The Management Revolution. Harvard Business Review, Vol. 90 Issue 10. Boston (2012) 60-68. 8p. 3
17. V. Sujatha, and Punithavalli.: Improved User Navigation Pattern Prediction Technique from Web Log Data. Procedia Engineering, Vol. 30. Philadelphia (2012) 92–99
18. Smalheiser, N. R.: Predicting emerging technologies with the aid of text-based data mining: the micro approach. Technovation, Vol. 21. Chicago (2001) 689–693
19. Biesdorf, S., Court, D., & Willmott, P.: Big data: What's your plan?. Mckinsey Quarterly, Vol. 2. (2013) 40-51.
20. Alazeb, A., & Panda, B. (2019, December). Maintaining Data Integrity in Fog Computing Based Critical Infrastructure Systems. In 2019 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 40-47). IEEE.
21. Panda, B., & Alazeb, A. (2020, August). Securing Database Integrity in Intelligent Government Systems that Employ Fog Computing Technology. In *2020 International Conference on Computing and Data Science (CDS)* (pp. 202-207). IEEE.