RESEARCH ARTICLE OPEN ACCESS

# Mining Big Data to Predicting Future

Amit K. Tyagi[1], R. Priya[2],A. Rajeswari[3]
Department of Computer Science and Engineering,
Pondicherry Engineering College, Puducherry-605014, INDIA

**Abstract**
Due to technological advances, vast data sets (e.g. big data) are increasing now days. Big Data a new term; is used to identify the collected datasets. But due to their large size and complexity, we cannot manage with our current methodologies or data mining software tools to extract those datasets. Such datasets provide us with unparalleled opportunities for modelling and predicting of future with new challenges. So as an awareness of this and weaknesses as well as the possibilities of these large data sets, are necessary to forecast the future. Today's we have an overwhelming growth of data in terms of volume, velocity and variety on web. Moreover this, from a security and privacy views, both area have an unpredictable growth. So Big Data challenge is becoming one of the most exciting opportunities for researchers in upcoming years.
Hence this paper discuss about this topic in a broad overview like; its current status; controversy; and challenges to forecast the future. This paper defines at some of these problems, using illustrations with applications from various areas. Finally this paper discuss secure management and privacy of big data as one of essential issues.

***General Terms*—Big data, data mining, Large datasets; Internet of things.**

## I. Introduction

Recent years have witnessed a dramatic increase in our ability to collect data from various sensors, devices in multiple formats, from independent or connected applications [1, 2 and 3]. This data has outpaced our capability to process, analyze, store and understand these datasets [2]. For e.g. Internet data. Imagine that, the web pages indexed by Google were around 1 million in 1998, but quickly it reached 1 billion in 2000 and 1 trillion in 2008 and so on. So this rapid expansion is accelerated by the dramatic increase in acceptance of social networking applications, such as Facebook, Twitter etc., that allow users to create contents freely and amplify the already huge Web volume and give an overview to predict the future of users. Based on this, these companies read mind of their users like to providing ads. Furthermore, with mobile phones becoming the sensory gateway to get real-time data on people from different aspects, the vast amount of data that mobile carrier can potentially process to improve our daily life has significantly outpaced our past CDR (Call Data Record)-based processing for billing purposes only [1, 2 and 3].

It can be foreseen that Internet of things (IoT) applications will raise the scale of data to an unprecedented level. People and devices (from home coffee machines to cars, to buses, railway stations and airports) are all loosely connected [1, 2, 3]. Trillions of such connected components generates a huge data ocean; by that valuable information can be discover to help people and to improve the quality of their life and to make this world a better place for them for e.g. after we get up in every morning, in order to optimize our commute time to work and complete the optimization before we arrive at once, the system needs to process information from traffic, weather, construction, police activities to our calendar schedules.

After that we perform our deep optimization schedule under the tight time constraints.

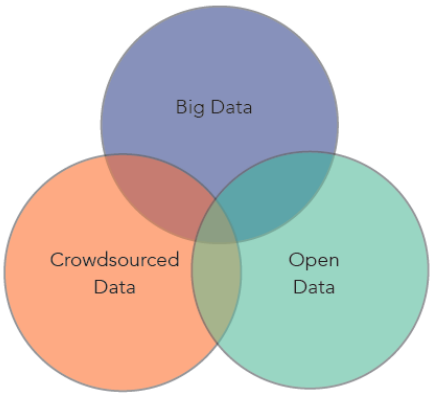


**Fig. 1. Relationship between Big Data, Crowdsourced Data, and Open Data**

Figure 1 shows the relation among various types of data existed in this real world. With the rapid increment in sensor technologies, mobile, cloud and high speed computing, we have already witnessed the emergence of the Big Data era [4]. While the immense

volume of data being produced at exponential rates by plethora of increasingly heterogeneous computing devices and sensors, that constitutes the Internet of Things (IoT), show huge potential for the better understanding of various phenomena and events through predictive data analytics[4]. The same data can be significantly misused and exploited for harvesting privacy sensitive information, or affecting the veracity or integrity of results of data analytics that will be used by people and enterprises for decision making purposes or predicting future. The diversity of data sources introduces immense variety in the media types such as text, images, videos and variations within these; further, establishing reliability and trustworthiness, as well as completeness of data from different sources become very difficult [4].

These exacerbate the problem of ensuring overall quality of data and information related to an individual or an enterprise throughout its lifecycle. The richness of multimedia data generates unique privacy risks as they can be correlated to reveal very sensitive information [4]. The velocity with which multimedia data flows through the networks and devices enabled by rapid advances in mobile and cloud computing, and networking technologies, which adds another level of challenge w.r.t to securely processing potentially inaccurate, unfiltered data in motion. The increasing volume, velocity, and variety of data and the increasing challenge with regards to establishing veracity of such data, present an unprecedented level of security and privacy challenges [4]. In particular, the threat landscape has been seen an immense growth resulting in a significant increase in number of threats witnessed in short periods of time [4]. This rapid growth in the threat spectrum have also resulted in many sophisticated hacker tools and cyber criminals that never existed before. Seen along with such increase in threats are also many sophisticated hacker tools, which if coupled with emerging big data analytics tools, will enable cyber criminals to acquire computing resources to create large scale security incidents that never existed before [4].

Today's we have been witnessing an overwhelming growth of data in terms of volume, velocity and variety, interestingly So from a security and privacy standpoint view, the threat landscape; security and privacy risks have also seen an unprecedented growth. Privacy is a pillar of democracy, we must remain alert to the possibility that it might be compromised by the rise of new technologies, and put in place all necessary safeguards [5, 6]. It is discussed as a main issue in this paper.

But in all these applications, we are facing various significant challenges in leveraging the vast amount of data, including challenges in (1) system capabilities (2) algorithmic design (3) business models (4) security (5) privacy. As an example, Big Data is having in the data mining community, which is being considered as one of the most exciting opportunities in the years to come [3]. Moreover this, Secure Management of Big Data with today's threat spectrum is also a biggest challenging problem. This paper shows that significant research effort is needed to build a generic architectural framework towards addressing these security and privacy challenges in a holistic manner as future work.

Finally this paper introduce Big Data mining and its applications in Section 2. We summarize the papers presented in this issue in Section 3, and discuss about Big Data problems in Section 4. Section 5 discuss about Big Data controversy. We point the importance of open-source software tools in Section 6 and give some challenges and forecast to the future in Section 7. Finally Section 8 conclude this paper.

## II. Mining Big Data

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the NextWave of InfraStress" [7,8,9]. Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya [3, 7]. However, the first academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold [2, 3, 7, and 10].

The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day [7]. Usama Fayyad in his invited talk at the KDD (Knowledge Discovery Database) BigMine'12 Workshop presented amazing data numbers about internet usage that is: each day Google has more than 1 billion queries per day, Twitter has more than 250 million tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of zettabytes, and it is increasing around 45% every year. A new large source of data is going to be generated from mobile devices; big companies as Google, Apple, Facebook, Yahoo, and Twitter are starting to look carefully to this data to find useful patterns to improve user experience [3, 7, 8, 10, 11, and 12]. Based on this, these companies read mind of their users like to providing ads.

Big Data's Key characteristics are Huge with heterogeneous and diverse data sources, Decentralized control, Complex data and knowledge associations. It (big data) can be two types like; structured and unstructured [7]. Firstly ***structured data*** are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smartphones, and global positioning system (GPS) devices [7, 13]. Structured data also include things like sales figures, account balances, and transaction data. And secondly ***unstructured data*** include more complex information,

such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites [13]. These data cannot easily be separated into categories or analyzed numerically.

Big Data mining is the capability of extracting useful information from these large datasets or streams of data that is due to its volume, variability, and velocity, it was not possible before to do it [3, 14, 15, 16, 17, and 18]. We need new algorithms and new tools to deal with all of this collected data. Doug Laney [2, 8] was the first one in talking about 3 V's in Big Data management (*refer figure 2*):

- **Variety:** there are many different types of data, exists as text, sensor data, audio, video, graph, and more.
- **Volume:** there is more data available than ever before, its size continues increasing day by day, but not the percent of data that our tools can analyses
- **Velocity:** data is arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time.

*Nowadays, there are two more V's:*

- **Value:** business value that gives organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach.
- **Variability:** there are changes in the structure of the data and how users want to interpret that data.

Gartner summarizes this in their definition of Big Data in 2012 as high volume, velocity and variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making [2, 3, 10, 11, 19, and 20]. There are various applications exists about Big Data for e.g.:

- **Health:** mining DNA of each person, to discover, monitor and improve health aspects of every one.
- **Business:** customer personalization, churn detection.
- **Smart cities:** cities focused on sustainable economic development and high quality of life, with wise management of natural resources.
- **Technology:** reducing process time from hours to seconds.

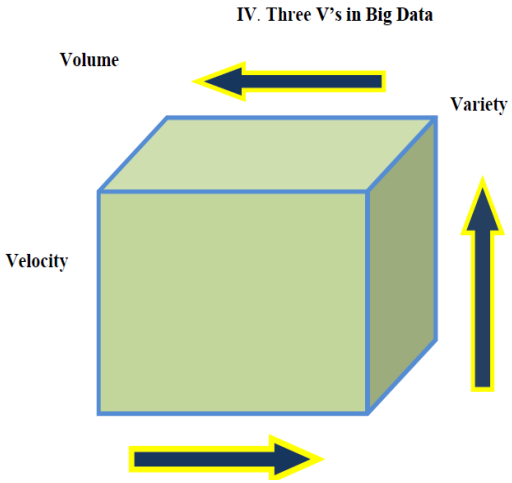


**Fig. 2.3 V" s in Big Data Management**

These applications will allow people to have better services, better customer experiences, and also be healthier, as personal data will permit to prevent and detect illness much earlier than before [2, 3].

### 2.1 Global Pulse: "Big Data as a Big Asset for Future Development"

To show the usefulness of Big Data mining, Global Pulse is using Big Data to improve life in developing countries [3]. Global Pulse is a United Nations initiative, launched in 2009, that functions as an innovative lab i.e. Based on mining Big Data for developing countries [21]. They work with a strategy that consists of 1) researching innovative methods and techniques for analyzing real-time digital data to detect early emerging vulnerabilities 2) assembling free and open source technology toolkit for analyzing real-time data and sharing hypotheses; and 3) establishing an integrated, global network of Pulse Labs, to pilot the approach at country level [3]. Global Pulse describe the main opportunities of Big Data that offers to developing countries in their White paper "Big Data for Development: Challenges & Opportunities":

- **Real-time awareness:** design programs and policies with a more fine-grained representation of reality [21].
- **Early warning:** develop and provide fast response in time of crisis, detecting anomalies in the usage of digital media [21].
- **Real-time feedback:** check what policies and programs fails, monitoring it in real time, and using this feedback make the needed changes. The Big Data mining revolution is not restricted to the industrialized world, as mobiles are spreading in developing countries as well. It is estimated that there are over five billion mobile phones, and that 80% are located in developing countries [21].

Big data for development is about turning imperfect, complex, often unstructured data into actionable information [22]. Big Data for Development sources generally share some or all of these features:

(1) **Digitally Generated-** i.e. the data are created digitally (as opposed to being digitised manually), and can be stored using a series of ones and zeros, and thus can be manipulated by computers [23, 24, 25, 26].

(2) **Passively Produced**- a by product of our daily lives or interaction with digital services [23,24,25].

(3) **Automatically Collected-** there is a system in place that extracts and stores the relevant data as it is generated [23,24]

(4) **Geographically or Temporally Trackable –** for e.g. mobile phone location data or call duration time [23,25]

(5) **Continuously Analysed**– information is relevant to human well - being and development and can be analyses in real time [27, 28].

## 2.2 Forecasting

To forecasting, this paper start with an example; like steering a car by looking through the rear view mirror i.e. of course, one would never steer a car like that. To steer a car, one looks ahead, noting that one is approaching a bend in the road, that there is another vehicle bearing down on one, and that there is a cyclist just ahead on the near side. That is, in steering a car, one sees that certain things lie ahead, which will have to be taken into account. The presumption in this is that one cannot see what lies ahead, but, instead, has to try to predict it based on an analysis of past data [29].

In such a retrospective analysis, one examines configurations of incidents from the past, seeking arrangements which are similar to those of the present, so that one can extrapolate from these past incidents through the present to the future. Sophisticated extrapolations also take into account the uncertainties involved, giving distributions or confidence intervals for likely future values. The fact is, however, that in steering a car one is making exactly the same kind of retrospective analysis. One observes the car ahead, and, based on one's previous experience with approaching vehicles, assumes that the vehicle will continue to proceed, in a relatively uniform manner, on the correct side of the road [29].

The key here about the user and car, is that one's predictions, one's forecasts, are based on assumptions of continuity with one's past experience. Sometimes in some similar situations in past, almost all had been followed by a particular event, then one would have considerable confidence that the same thing would happen the next time: the sun rising tomorrow is the good example [29]. The trick in all situations is quantifying the degree of continuity, and in a sense, that is what all forecasting is about. The desire to forecast is universal i.e. one of those things we all wish we could do is know the future.

Forecasting has several aspects. One is defining the degree of similarity between the present and the past. Another is determining the range and variability of events which followed these similar past events, and a third is deciding whether one understands enough about the underlying process to adopt a particular model form [29]. *Forecasting also has its limitations.* Firstly, there are chaotic limitations. These are fundamental in the sense that they tell us that no matter how much we know about the past and the present, and no matter how accurately we know it, there comes a point in the future at which the minuscule inaccuracies in our present knowledge will have been amplified to [3, 29] render our forecasts wide of the mark. This is nicely illustrated by weather forecasting, where, thanks to vast computational resources and huge statistical models, we can now forecast reasonably accurately perhaps five days ahead, but where extending this significantly further requires dramatically more extensive data and greater computer power. Secondly, there are stochastic limitations. These are the sudden, unpredicted and unpredictable jolts to the system which are often caused by external agencies, or perhaps by inadequacies in the model [29]. A nice recent example of that is the current global financial crisis. I have been asked: could we have seen it coming? The short answer, of course, is that we did: there are many economic forecasters, and at least as many forecasts. Some of these were sufficiently confident of the danger to act on it (some hedge funds did very well out of it). If we combine data mining with forecasting we can always find someone who (on looking back) gave the right forecast [29]. This is the basis for the surefire way of making money as a stock market tipster by making a series of multiple different forecasts, and eventually selecting just those potential customers to whom you gave a series which happened to turn out to be correct. It also illustrates the difficulties of making inferences in data mining, when huge data sets and numbers of data configurations are involved [29].

Hence this section discuss about this forecasting in detail. Now section will dealt with the contributed points related to this paper.

# III. Related Main Issues

Four contributions in Big Data Mining that together shows very significant state of the art research and provides a broad overview of the field which is used to predict the future:

- **Big Graph Mining:** Algorithms and discoveries by U Kang and Christos Faloutsos (Carnegie Mellon Universitiy). This paper presents an overview of mining big graphs, focusing in the use of the Pegasus tool, showing some findings in the Web Graph and Twitter social network [30]. The paper gives

inspirational future research directions for big graph mining [3, 30].

- **Scaling Big Data Mining Infrastructure:** The Twitter Experience by Jimmy Lin and Dmitriy Ryaboy (Twitter, Inc.). This paper presents insights about Big Data mining infrastructures, and the experience of doing analytics at Twitter [8, 30]. It shows that due to the current state of the data mining tools, it is not straightforward to perform analytics. Most of the time is consumed in preparatory work to the application of data mining methods, and turning preliminary models into robust solutions [3, 30].
- **Mining Large Streams of User Data for Personalized Recommendations by Xavier Amatriain (Netix).**
- **Mining Heterogeneous Information Networks:** A Structural Analysis Approach by Yizhou Sun (North- eastern University) and Jiawei Han (University of Illinois at Urbana-Champaign). This paper define that mining heterogeneous information networks is a new and promising research frontier in Big Data mining research [31]. It considers interconnected, multi-typed data, including the typical relational database data, as heterogeneous information networks [30]. These semi-structured heterogeneous information network models leverage the rich semantics of typed nodes and links in a network and can uncover surprisingly rich knowledge from interconnected data [3, 30, and 31].

Hence this paper presents some lessons learned with the Netix Prize. It discusses recent important problems and future research directions. Now section 4 contains an interesting discussion about problems arises in big data mining, which will have a most important role to discover future.

## IV. Problem Arises

There is undoubtedly enthusiasm about the emerging data revolution, and the possibilities of making use of Big Data for measuring and monitoring progress in societies. Researchers are taking note of this alternative data source, but with some degree of caution, as bigger data need not always mean better data. There is some problems in big data mining, as the latter were not tailor made for statistical purposes and its use provides the risk of yielding figures that are far from reality. Big data is largely unstructured, unfiltered data exhaust from digital products, such as electronic and online transactions, social media, sensors (GPS, climate sensors), and consequently, analytics can be poor, unlike traditional data sources utilized for official statistics that are well-structured with good analytics, but with a fairly high cost (for data collection), and typically infrequent conduct with time lags that stakeholders find unreasonable in an age of fast data [32, 33].

The combination of large data sets and observational data mean that data mining exercises are often at risk of drawing misleading conclusions. This section describe just four of these problems. These problems are certainly not things alone have detected. Indeed, within the statistics community, they are others problems also which are well understood. However, by the central philosophy of data mining, throw sufficient computer power at a large enough data set and interesting things will be revealed [29]. It meant that they have often been overlooked in data mining exercises. Unfortunately, the solution is to temper the enthusiasm, and to recognize that rather more complex models are necessary. Statisticians generally do not build complicated models simply for fun, but for good reasons [29]. From above discussion, we can say "Is it Big Data or Big Value or Big Problem"? So four problems arise in mining big data are:

**Problem 1. Selectivity bias.**

**Problem 2. Out of date data.**

**Problem 3. Empirical rather than iconic models.**

**Problem 4. Measuring performance.**

Hence this section dealt with problems arises in big data mining. Now further section will deal with controversy points arises in big data mining.

## V. Controversy about Big Data Mining

Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data [7, 35]. As Big Data is a new hot topic with mining to predict the future. There have been a lot of controversy also about it i.e. "Is it Big Data or Big Value or Big Problem"? This paper summarize all as:

- There is no need to distinguish Big Data analytics from data analytics, as data will continue growing, and it will never be small again [35].
- Big Data may be a hype to sell Hadoop based computing systems. Hadoop is not always the best tool. It seems that data management system sellers try to sell systems based in Hadoop, and MapReduce may be not always the best programming platform, for example for medium-size companies [3, 37].
- In real time analytics, data may be changing. In that case, what it is important is not the size of the data, it is its recency [3, 37].
- Claims to accuracy are misleading.
- Bigger data are not always better data. It depends if the data is noisy or not, and if it is representative of what we are looking for [3, 37].

For example, sometimes twitter users are assumed to be a representative of the global population, when this is not always the case [17].

- Ethical concerns about accessibility. The main issue is if it is ethical that people can be analyzed without knowing it [37].
- Limited access to Big Data creates new digital divides. There may be a digital divide between people or organizations being able to analyze Big Data or not. Also organizations with access to Big Data will be able to extract knowledge that without this Big Data is not possible to get. We may create a division between Big Data rich and poor organizations [37] to forecasting.

Hence, this section discuss controversy points in big data mining. Section 6, contain information about important tools to mining big data to discover the future events.

## VI. Tools: Open Source Revolution

The Big Data phenomenon is intrinsically related to the open source software revolution. Most of the large companies like as Facebook, Yahoo!, Twitter, and LinkedIn etc. benefit and contribute working on open source projects. Big Data infrastructure deals with Hadoop, and other related software as:

- **Apache Hadoop [3]:** this software is used for data-intensive distributed applications; based in the MapReduce programming model and a distributed file system called Hadoop Distributed File system (HDFS). Hadoop allows writing applications that rapidly process large amounts of data in parallel on large clusters of compute nodes. A MapReduce job divides the input dataset into independent subsets that are processed by map tasks in parallel [38]. This step of mapping is then followed by a step of reducing tasks. These reduced tasks use the output of the maps to obtain the final result of the job.
- **Apache S4 [2, 3, and 16]:** it provides a platform for processing continuous data streams. S4 is designed specifically for managing data streams. S4 apps are designed combining streams and processing elements in real time [38].
- **Apache Hadoop related projects [3, 16]:** Apache Pig, Apache Hive, Apache HBase, Apache ZooKeeper, Apache Cassandra, Cascading, Scribe and many others [38].
- **Storm [1, 3]:** this software is used for streaming data-intensive distributed applications, similar to S4, and developed by Nathan Marz at Twitter.

*In Big Data Mining, there are many open source initiatives. The most popular are the following:*

- **Apache Mahout [4, 7]:** Scalable machine learning and data mining open source software based mainly in Hadoop [38]. It has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining.
- **MOA [3]:** Stream data mining open source software to perform data mining in real time. It has implementations of classification, regression, clustering and frequent item set mining and frequent graph mining. It started as a project of the Machine Learning group of University of Waikato, New Zealand, famous for the WEKA software.The streams framework provides an environment for defining and running stream processes using simple XML based definitions and is able to use MOA, Android and Storm. SAMOA is a new upcoming software project for distributed stream mining that will combine S4 and Storm with MOA [38].
- **R [3, 29]:** open source programming language and software environment designed for statistical computing and visualization. R was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand beginning in 1993 and is used for statistical analysis of very large data sets [38].
- **Vowpal Wabbit [2]:** it is an open source project started at Yahoo! Research and continuing at Microsoft Research to design a fast, scalable, useful learning algorithm. VW is able to learn from terafeature datasets. It can exceed the throughput of any single machine network interface [38] when doing linear learning, via parallel learning.

More specific to Big data/ Graph mining, found the following open source tools:

- **GraphLab [2, 3, and 4]:** it is a high-level graph-parallel system that built without using MapReduce. GraphLab computes over dependent records which are stored as vertices in a large distributed data-graph. Algorithms in GraphLab are expressed as vertex-programs which are executed in parallel on each vertex and can interact with neighboring vertices.
- **Pegasus:** it is a big graph mining system that built on top of MapReduce. It allows to find patterns and anomalies in massive real-world graphs.

Hence Section 6, contain information about important tools to mining big data to discover the future events. Now next section dealt with issues related to forecast the future.

## VII. Forecast to the Future

There are many future important challenges in Big Data management and analytics that arise from the nature of big data: large, diverse, and evolving. These are some of the other challenges

that researchers and practitioners will have to deal in next years:

- **Analytics Architecture**: It is not clear yet how an optimal architecture of an analytics systems should be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz [15, 39, and 40]. The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in real-time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer [40]. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general, and extensible, allows ad-hoc queries, minimal maintenance, and debuggable [41].
- **Time evolving data:** Data may be evolving over time, so it is important that the Big Data mining techniques should be able to adapt and in some cases to detect change first for e.g. the data stream mining field has very powerful techniques for this task.
- **Distributed mining:** Many data mining techniques are not trivial to paralyze [3, 40]. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods to predict future.
- **Visualization:** A main task of Big Data analysis is "how to visualize the results". As the data is so big, it is very difficult to find user-friendly visualizations. New techniques, and frameworks to tell and show stories will be needed for e.g. the photographs, infographics and essays in the beautiful book "The Human Face of Big Data" [41].
- **Compression:** Dealing with Big Data, the quantity of space needed to store it is very relevant. There are two main approaches: [41] compression where we don't loose anything, or sampling where we choose what is the data, which is more representative. Using compression, we may take more time and less space, so we can consider it as a transformation from time to space [41]. Using sampling, we are loosing information, but the gains in space may be in orders of magnitude for e.g. Feldman et al. use coresets to reduce the complexity of Big Data problems. Coresets are small sets that provably approximate the original data for a given problem. Using merge-reduce the small sets can then be used for solving hard machine learning problems in parallel [41].
- **Statistical significance:** It is important to achieve significant statistical results, and not be fooled by randomness. As Efron explains in his book about Large Scale Inference, it is easy to go wrong with huge data sets and thousands of questions to answer at once [3].
- **Hidden Big Data:** Large quantities of useful data are getting lost since new data is largely untagged file based and unstructured data [7, 40, 41, and 43]. The 2012 IDC study on Big Data explains that in 2012, 23% (643 exabytes) of the digital universe would be useful for Big Data if tagged and analyzed [7, 11, 40, 41, and 42]. However, currently only 3% of the potentially useful data is tagged, and even less is analyzed [34, 36, 43, 44, 45 and 46].

Hence this section dealt with challenges issues arises related to forecast the future. Finally section 8, conclude this paper in brief.

## VIII. Conclusion

Due to increasing the size of data day by day, Big Data is going to continue growing during the next years and becoming one of the exciting opportunities in future. This paper insights about the topic, and controversy, and the main challenges etc. for the future. Hence Big Data is becoming the new Final Frontier for scientific data research and for business applications. And on challenging side, Securely Management of Big Data with today's threat spectrum is a big issue. Because today's we have an overwhelming growth of data in terms of volume, velocity and variety.

So from a security and privacy standpoint, the threat landscape and security and privacy risks have also seen an unprecedented growth. So as for future research is needed to build a generic architectural framework towards addressing these security and privacy challenges in a holistic manner. Now we are in a new era where Big Data mining will help us to discover knowledge that no one has discovered before. So everybody is warmly invited to participate in this intrepid journey to discover the future views.

## References

[1.] http://big-data-mining.org/
[2.] Ms. Neha A. Kandalkar , Prof. Avinash Wadhe," Extracting Large Data using Big Data Mining", International Journal of Engineering Trends and Technology (IJETT) – Volume 9 Number 11 - Mar 2014.
[3.] Wei Fan, Albert Bifet," Mining Big Data: Current Status, and Forecast to the Future"
[4.] James Joshi, Balaji Palanisamy," Towards Risk-aware Policy based Framework for Big Data Security and Privacy", 2014.
[5.] http://www.phibetaiota.net/2012/06/patrick-meier-un-report-on-big-data-for-development-highlights/
[6.] http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopmentGlobalPulseMay2012.pdf
[7.] Bharti Thakur, Manish Mann," Data Mining for Big Data: A Review", International Journal of Advanced Research in Computer Science and

Software Engineering, Volume 4, Issue 5, May 2014, ISSN: 2277 128X.
[8.] http://www.ijser.org/paper/Data-Mining-and-Data-Pre-processing-for-Big-Data.html
[9.] Aarshi Jain, Ankit Chadha," International Journal of Engineering & Science Research", IJESR/November 2013/ Vol-3/Issue-11/5015-5019 ISSN 2277-2685.
[10.] http://albertbifet.com/big-data-mining/
[11.] https://medium.com/big-data science/big-data-science-ba016521d3a6
[12.] Dattatray Raghunath et al. "A Survey on Big Data mining Applications and different Challenges", IJARCET, Volume 3 Issue 11, November 2014.
[13.] http://www.bls.gov/careeroutlook/2013/fall/art01.pdf
[14.] D. Pratiba, G. Shobha, Ph. D," Educational BigData Mining Approach in Cloud: Reviewing the Trend", International Journal of Computer Applications (0975 – 8887) Volume 92 – No.13, April 2014.
[15.] Mrs. Deepali Kishor Jadhav," Big Data: The New Challenges in Data Mining", IJIRCST, ISSN: 2347- 5552, Volume-1, Issue-2, September, 2013.
[16.] Harshawardhan S. Bhosale , Prof. Devendra P. Gadekar," A Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014 1 ISSN 2250-3153
[17.] Miss. Neha V. Deshmukh Miss. Surbhi G. Atal Prof. Pushpanjali Chauragade," Big Data Mining: An Overview", International Journal for Engineering Applications and Technology, January 2015 ISSN: 2321-8134.
[18.] Melwin Devassy , Gera.Praveen Kumar," Enhancing Data Privacy In Data Extraction With Big Data", International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 122– 127 (2014)
[19.] Miss. Punde Archana, et al." A Review:Data Mining for Big Data" IJARCET, Volume 3 Issue 10, October 2014.
[20.] A. V. N. S. Jyothirmayee , Dr. G. Sreenivasula Reddy , K. Akbar "Understanding Big Data & DV2", International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 7, July 2014).
[21.] http://albertbifet.com/global-pulse-big-data-for-development/
[22.] T.Rathika, J.Senthil Murugan,"FP Tree Algorithm and Approaches in Big Data", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 9, September 2014.
[23.] http://www.ciard.net/sites/default/files/bb40_backgroundnote.pdf
[24.] http://operationalrisk.blogspot.in/2014/07/global-pulse-resilience-in-development.html
[25.] http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment GlobalPulseMay2012.pdf
[26.] http://wpdi.org/sites/default/files/ipi_epub_new_technology_final.pdf
[27.] http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopmentUNGlobalPulseJune2012.pdf
[28.] https://brusselsbriefings.files.wordpress.com/2015/01/note.pdf
[29.] David J. Hand," Mining the past to determine the future: Problems and possibilities", International Journal of Forecasting 25 (2009) 441–451.
[30.] Kaushika Pal, Dr. Jatinderkumar R. Saini," A Study of Current State of Work and Challenges in Mining Big Data", International Journal of Advanced Networking Applications (IJANA) ISSN No. : 0975-0290
[31.] http://www.nscb.gov.ph/statfocus/2013/SF_102013_OSG_bigData.asp
[32.] http://paris21.org/newsletter/fall2013/big-data-dr-jose-ramon-albert
[33.] Xindong Wu, Xingquan Zhu , Gong-Qing Wu, Wei Ding," Data Mining with Big Data"
[34.] http://www.rcrwireless.com/20121212/big-data-analytics/huge-big-data-gap-only-0-5-data-analyzed
[35.] http://albertbifet.com/author/abifet/
[36.] http://spotfire.tibco.com/blog/?p=16824
[37.] http://albertbifet.com/big-data-mining-tools/
[38.] Bharti Thakur Manish Mann," Data Mining With Big Data Using C4.5 and Bayesian Classifier", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 8, August 2014.
[39.] Jainendra Singh," Big Data Analytic and Mining with Machine Learning Algorithm", International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 4, Number 1 (2014).
[40.] http://albertbifet.com/big-data-mining-future-challenges/
[41.] http://www.emc.com/about/news/press/2012/20121211-01.htm
[42.] http://india.emc.com/about/news/press/2012/20121211-01.htm
[43.] http://www.forbes.com/sites/marketshare/2012/12/21/behold-the-untapped-big-data-gap/
[44.] http://www.informationweek.in/informationweek/news-analysis/177856/percent-world analyzed-study
[45.] http://datascienceseries.com/blog/digital-universe-will-grow-to-40zb-in-2020-with-a-62-share-for-emerging-markets
[46.] http://precisionmatch.net/blog/big-data-in-2020/