# Proposal ‘Contributed Paper Session for 63rd ISI World Statistics Congress 2021

## On

## Title- Big Data Mining from Big Data Sources

**Author- Md. Shariful Islam**

**Member-International Statistical Institute (Membership No-19241)**

**E-mail-Orphi456@gmail.com**

**Phone-8801960567408**

**B.Sc.(Honors), M. Sc. (Statistics)**

**Jahangirnagar University, Bangladesh**

**Cypher Officer**

**Department of Cypher,**
**Ministry of Defense People’s Republic of Bangladesh**

**And Statistician**

**<u>Affiliation</u>:**

**1. Socio Economic Condition of Fisherman and at their Different Aspect in Bangladesh.**

**Research Paper, Department of Statistics Jahangirnagr University, Dhaka, Bangladesh**

2. ‘Aims to Expand Statistical Capacity in Bangladesh’
Contributed Paper Presentation, WSC (ISI, 2017), Marrakech, Morocco

**3. ‘Big Data’ Contributed Paper Presentation,**
WSC (ISI, 2019), Kualalumpur, Malaysia

# Abstract

Big Data creates significant new opportunities for organizations to derive new value and create competitive advantages from their most valuable asset information. Enormous amounts of data are generated every minute. Some sources of data, such as those found on the Internet are obvious. Social networking sites, search and retrieval engines, media sharing sites, stock trading sites, and news sources continually store enormous amounts of new data throughout the day. Big Data helps drive efficiency, quality, and personalized products and services producing levels of customer satisfaction and profit. Consider the vast quantity data collected from sensors in methodological and climate systems, or patient monitoring systems in hospitals. Data acquisition and control systems, such as those found in cars, airplanes, cell towers, and power plants, all collect unending streams of data. The healthcare industry is inundated with data from patient records alone. Regardless of the source of the data contained within them are nuggets of knowledge that can potentially improve our understanding of the world around us.

The field of data mining is interdisciplinary. The core of the field relies data structures and algorithm from computer science, and probability and statistics from mathematics. The field focuses on the application numerous methods that came out of research in machine learning-a field whose is to develop new algorithms that is improve some performance metric from data. While machine learning tends to focus on theory and algorithm development, data mining focuses on the application of these methods toward large-scale. Data mining would not have made the great strides it has today without accomplishments of machine learning.

Data Mining focuses on the process of discovering new patterns of Data sets while Data science focuses on the data that is data mining, machine learning, deep learning statistics and much more. So Big Data Mining that is new pattern Data Sets keeps great significance in Data Science. So Big Data mining is data sets of Big Data that contribute great significance in storing, collecting and distributing information to clients by modern technology.

Big Data Analytics, Big data, Data Mining Techniques, Data Sets, New Patterns of Data

## Introduction

Data Mining focuses on the process of discovering new patterns of Data sets while Data science focuses on the data that is data mining, machine learning, deep learning statistics and much more. So Big Data Mining that is new pattern Data Sets keeps great significance in Data Science.

Big Data defines enormous amounts of data that are generated every minute. Some sources of data, such as those found on the Internet are obvious. Social networking sites, search and retrieval engines, media sharing sites, stock trading sites, and news sources continually store enormous amounts of new data throughout the day. For example, study in Google Data warehouse, every minute, google receives over 2 million queries, e-mail users send over 200 million messages, YouTube users upload 48 hours of video, Facebook users share over 680000 pieces of content and Twitter users generate 100000 tweets. Some sources of Data are not obvious. Consider the vast quantity data collected from sensors in methodological and climate systems, or patient monitoring systems in hospitals. Data acquisition and control systems, such as those found in cars, airplanes, cell towers, and power plants, all collect unending streams of data. The healthcare industry is inundated with data from patient records alone. Regardless of the source of the data contained within them are nuggets of knowledge that can potentially improve our understanding of the world around us. All the system of data generation are part of Big Data and data mining is new pattern of Data Sets.

## Description

Big Data is a catch phrase that describes aspects of the data itself IBM, a major player in this field, has four descriptions that are used to determine if data are classified as Big Data's:

Volume : The sheer size of the data is enormous, making traditional data processing methods impossible.

Variety: the data can be represented in a wide range of types , structures or unstructured including text, sensor, streaming audio or video or user click streams, to name a few.

Veracity : The resulting information for the analyses need to be accurate and trustworthy

Velocity : Many Data need to be processed and analyzed in near real time. This is a huge problem considering the wide range of sources that data comes from.

Besides on above there are many other related fields that have provided an important, long standing foundation in this area. The broad topic Database Management system(DBMS) focuses on collection, storage, management and retrieval of data. It has become increasingly common for businesses to have multiple data bases from multiple sources often with their own formats. Such as a data warehouse that is a type of database that focuses on the aggregation and integration of data from sources, usually for analysis and reporting purposes. Many fields in Big Data focus on the extraction of information. Often Business Intelligence (BI) systems and related systems manage their data in the form of Data Cubes, which are multi-dimensional of data managed and modeled in a way for rapid query and analysis. Online analytical processing, or OLAP, is an important part of BI systems that focuses creating views and queries from data cubes for the purposes of analyzing business data. Data Visualization is an important field in Big Data that focuses on the development of methods that can help analysis visualize interesting patterns or relationship in data

## Data Mining

The field of data mining is interdisciplinary. The core of the field relies data structures and algorithm from computer science, and probability and statistics from mathematics. The field focuses on the application numerous methods that came out of research in machine learning-a field whose is to develop new algorithms that is improve some performance metric from data. While machine learning tends to focus on theory and algorithm development, data mining focuses on the application of these methods toward large-scale. Data mining would not have made the great strides it has today without accomplishments of machine learning.

Data mining is the computational process of finding patterns in given data and making predictions for the new data. The main techniques for finding patterns are association rule/frequent-pattern mining, clustering, classification and regression analysis. The data mining process consists of 3 stages: getting and cleaning data (or ETL - Extract, Transform, and Load), Model building and Deployment

The data mining models can be divided into 2 categories: *unsupervised and supervised* models. Unsupervised models are concerned with finding patterns/clusters in the given data, whereas supervised models deal with training the system with historical data and making predictions or classifying new data. Various statistical models like regression analysis, Bayes model, maximum-likelihood estimation, etc. and machine learning models like Neural Network, Decision tree, etc. are used for data mining.

Classification is a data mining function that assigns items in a collection to target categories or classes. Falling under the field of supervised learning, of learning from labeled data, the goal of classification is to develop a computational model from existing data that can accurately predict the target class for each new datum that is yet to be observed. A classification task begins with a data set in which the class assignments are known, called labeled data.

Cluster and Outlier Analysis unsupervised mining function for discovering grouping in data. It falls under unsupervised learning because the data is unlabeled. Clustering analysis finds clusters of data objects that are similar in some sense to one another. The members of a cluster are more like each other than they are like members of other clusters. The goal of clustering analysis is to find high quality clusters such that the inter-cluster similarly is low and the intra cluster similarly in high. Clustering can also be used for anomaly detection. Once the data has been segmented, it might find that some cases do not fit well any clusters. These cases are anomalies or outliers.

Big Data mining is the new pattern of Data sets which the discovery of interesting, unexpected or valuable structures in large Datasets. Data mining is also knowledge application of data science fundamental principles. That means analyzing big data sets that is, Big Data Mining is involved the application of data science fundamental principles.

In experiments to be seen, cluster and classify a large dataset on a private cloud, which can be scaled up to handle the growing dataset. The time taken to cluster the dataset and to generate the models is also quite satisfactory. It is also observed the turn-around –time of the results improved by increasing the size of the cloud.

The data mining algorithm studied so far are not able to handle data streams. So, in order to obtain useful information from data streams.

## Conclusion

Much of the work is based on collective research and industrial experience with data mining, which culminated in the strong interest and desire to teach data mining to a new generation of in stem fields. Learning in the era of Big Data are covering nearly every aspect of world around has data being collected. Computers, cars, appliances, mobile devices, gaming consoles, and shops visit all are collecting data with every activity for perform. The vast majority of industries are seeking careers in have a plethora of opportunities to use data mining to further potential employers. Traditionally, data mining has been widely associated with programs in computer science. Now is the time in all stems fields to work hard to provide knowledge required to competently work in the emerging field of Big Data.

## References

[1] A. Bifet, “Mining Big Data in Real Time,” Informatica, Vol.37, pp. 15–20, 2013.

[2] G. Krempl, I. Zliobaite, D. B. Nski, E. H. Ullermeier, et. al., “Open Challenges for Data Stream Mining Research,” ACM SIGKDD Explorations, Vol. 16, No. 1, pp. 1-10, 2013.

[3] D.-H. Tran, M. M. Gaber, K.-U. Sattler, “Change detection in streaming data in the era of big data: models and issues,” ACM SIGKDD Explorations, Vol. 16, No. 1, pp. 30-38, 2014.

[4] W. Fan, A. Bifet, “Mining Big Data: Current Status, and Forecast to the Future,” ACM SIGKDD Explorations, Vol. 14, No. 2, pp. 1-5, December 2012.

[5] Y. Demchenko, P. Grosso, C. D. Laat, P. Membrey, “Addressing Big Data Issues in Scientific Data Infrastructure,” 2013 International Conference on Collaboration Technologies and Systems (CTS), 20-24 May 2013, San Diego, CA, USA, pp. 48-55, 2013