ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

HARNESSING BIG DATA: ADVANCES AND CHALLENGES IN DATA MINING

*Mungara S V V Satya Naryana[1],Mahammad Pasha[2] ,B. Jayapal[3],M. Ravinder[4]

[1]*UG Scholar, St. Martin's Engineering College, Secunderabad, Telangana – 500100*

[2]Assistant Professor, KPRIT, Hyderabad, Telangana – 500088

[3,4] UG Scholar , Siddhartha Engineering College, Hyderabad, Telangana – 500088

*Corresponding Author

Email: Mungarasurya01@gmail.com

Abstract-Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This article presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

Keywords: Big Data, Data Mining, Source Of Information, Data Storage, User Privacy, Technological Domains, Big Data Processing, Learning Algorithms, Complex Network, Cognitive Domains, Complex Data, Kinetic Data, Massive Data, Knowledge Discovery, Data Mining Methods, Decentralized Control, Map Reduce, Data Mining Algorithms**.**

I. INTRODUCTION

HACE Theorem: Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. In a naïve sense, we can imagine that a number of blind men are trying to size up a giant elephant (see Figure 1), which will be the Big Data in this context. The goal of each blind man is to draw a picture (or conclusion) of the elephant according to the part of information he collected during the process. Because each person's view is limited to his local region, it is not surprising that the blind men will each conclude independently that the elephant "feels" like a rope, a hose, or a wall, depending on the region each of them is limited to. To make the problem even more complicated, let's assume that (a) the elephant is growing rapidly and its pose also changes constantly, and (b) the blind men also learn from each other while exchanging information on their respective feelings on the elephant. Exploring the Big Data in this 4 scenario is equivalent to aggregating heterogeneous information from different sources (blind men) to help draw a best possible picture to reveal the genuine gesture of the elephant in a real-time fashion. Indeed, this task is not as simple as asking each blind man to describe his feelings about the elephant and then getting an expert to draw one single picture with a combined view, concerning that each individual may speak a different language (heterogeneous and diverse information sources) and they may even have privacy concerns about the messages they deliberate in the information exchange process.

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This is because different information collectors use their own schemata for data recording, and the nature of different applications also results in diverse representations of the data. For example, each single human being in a bio-medical world can be represented by using simple demographic information such as gender, age, family disease history etc. For X-ray examination and CT scan of each individual, images or videos are used to represent the results because they provide visual information for doctors to carry detailed examinations. For a DNA or genomic related test, microarray expression images and sequences are used to represent the genetic code information because this is the way that our current techniques acquire the data. Under such circumstances, the heterogeneous features refer to the different types of representations for the same individuals, and the diverse features refer to the variety of the features involved to represent each single observation. Imagine that different organizations (or health practitioners) may have their own schemata to represent each patient, the data heterogeneity and diverse dimensionality issues become major challenges if we are trying to enable data aggregation by combining data from all sources.

ISSN 2454-9940
INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
www.ijasem.org
Vol 19, Issue 1, 2025

Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data sources is able to generate and collect information without involving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) setting where each web server provides a certain amount of information and each server is able to fully function 5 without necessarily relying on other servers. On the other hand, the enormous volumes of the data also make an application vulnerable to attacks or malfunctions, if the whole system has to rely on any centralized control unit. For major Big Data related applications, such as Google, Flicker, Face book, and Wal-Mart, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets. Such autonomous sources are not only the solutions of the technical designs, but also the results of the legislation and the regulation rules in different countries/regions. For example, Asian markets of Wal-Mart are inherently different from its North American markets in terms of seasonal promotions, top sell items, and customer behaviours. More specifically, the local government regulations also impact on the wholesale management process and eventually result in data representations and data warehouses for local markets.

While the volume of the Big Data increases, so do the complexity and the relationships underneath the data. In an early stage of data centralized information systems, the focus is on finding best feature values to represent each observation. This is similar to using a number of data fields, such as age, gender, income, education background etc., to characterize each individual. This type of sample-feature representation inherently treats each individual as an independent entity without considering their social connections which is one of the most important factors of the human society. People form friend circles based on their common hobbies or connections by biological relationships. Such social connections commonly exist in not only our daily activities, but also are very popular in virtual worlds. For example, major social network sites, such as Face book or Twitter, are mainly characterized by social functions such as friend connections and followers (in Twitter). The correlations between individuals inherently complicate the whole data representation and any reasoning process. In the sample-feature representation, individuals are regarded similar if they share similar feature values, whereas in the sample-feature-relationship representation, two individuals can be linked together (through their social connections) even though they might share nothing in common in the feature domains at all. In a dynamic world, the features used to represent the individuals and the social ties used to represent our connections may also evolve with respect to temporal, spatial, and other factors. Such a complication is becoming part of the reality for Big 6 Data applications, where the key is to take the complex (non-linear, many-to-many) data relationships, along with the evolving changes, into consideration, to discover useful patterns from Big Data collections.

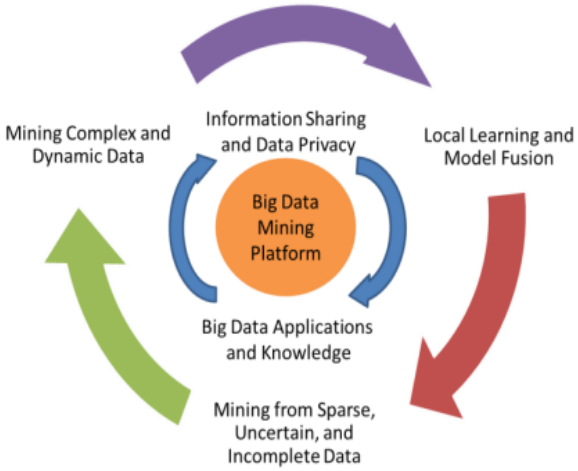

Figure 1: A Big Data processing framework: The research challenges form a three tier structure and center around the "Big Data mining platform" (Tier I), which focuses on low-level data accessing and computing. Challenges on information sharing and privacy, and Big Data application domains and knowledge form Tier II, which concentrates on high level semantics, application domain knowledge, and user privacy issues. The outmost circle shows Tier III challenges on actual mining algorithms.

For an intelligent learning database system (Wu 2000) to handle Big Data, the essential key is to scale up to the exceptionally large volume of data and provide treatments for the characteristics featured by the aforementioned HACE theorem. Figure 1 shows a conceptual view of the Big Data processing framework, which includes three tiers from inside out with considerations on data accessing and computing (Tier I), data privacy and domain knowledge (Tier II), and Big Data mining algorithms (Tier III). The challenges at Tier I focus on data accessing and actual computing procedures. Because Big Data are often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing. For example, while typical data mining algorithms require all data to be loaded into the main memory, this is becoming a clear technical barrier for Big Data because moving data across different locations is expensive (e.g., subject to intensive network communication and other IO costs), even if we do have a super large main memory to hold all data for computing.

The challenges at Tier II centre around semantics and domain knowledge for different Big Data applications. Such information can provide additional benefits to the mining process, as well as add technical barriers to the Big Data access (Tier I) and mining algorithms (Tier III). For example, depending on different domain applications, the data privacy and information sharing mechanisms between data producers and data consumers can be significantly different. Sharing sensor network data for applications like water quality monitoring may not be discouraged, whereas releasing and sharing mobile users' location information is clearly not acceptable for majority, if not all, applications. In addition to the above privacy issues, the application domains can also provide additional information to benefit or guide Big Data mining algorithm designs. For example, in market basket transactions data, each transaction is considered independent and the discovered knowledge is typically represented by finding highly correlated items, possibly with respect to different temporal and/or spatial restrictions. In a social network, on the other hand, users are linked and share dependency structures. The knowledge is then represented by user communities, leaders in each group, and social influence modelling etc. Therefore, understanding semantics and application knowledge is important for both low-level data access and for high level mining algorithm designs.

II. LITERATURE SURVEY

Data mining with big data has gained extensive attention due to the increasing complexity and volume of data generated from various sources, including social media, healthcare, finance, and sensor networks. This literature survey provides an overview of the key challenges, techniques, and future directions in data mining with big data. Big data is characterized by the "5 V's": Volume, Variety, Velocity, Veracity, and Value, each posing unique challenges in data mining. Volume refers to the massive amounts of data generated; Variety indicates diverse data types (text, video, images); Velocity refers to the high rate of data inflow; Veracity emphasizes data accuracy and consistency; Value underscores the need to extract meaningful information. Data mining in big data involves extracting useful information from extensive, often unstructured datasets. It uses statistical, machine learning, and pattern recognition methods to reveal hidden patterns.

Storing vast amounts of data, especially in real-time, is challenging. Traditional databases are insufficient, leading to the adoption of distributed systems like Hadoop and Apache Spark, which can handle large datasets across multiple nodes. Computational complexity of processing big data, especially with high dimensionality, requires advanced techniques for scalable and efficient data mining. Big data contains sensitive information that can lead to privacy concerns if not managed properly. Techniques like anonymization, encryption, and differential privacy are often used to protect user data. Integrating data from diverse sources can be complex due to differences in formats, structures, and quality. Ensuring data accuracy and completeness is challenging due to the potential for errors, noise, and missing values in massive datasets.

Key techniques in data mining for big data include pre-processing, which cleans and prepares data for analysis. Pre-processing steps like data cleaning, normalization, transformation, and reduction enhance data quality and algorithm efficiency. Techniques such as k-means clustering, decision trees, and support vector machines are widely employed to classify and group data in applications like customer segmentation, anomaly detection, and recommendation systems. Algorithms like Apriori and FP-Growth are modified for big data to identify frequently occurring patterns, useful in market basket analysis and fraud detection. Additionally, frameworks like Apache Hadoop, Apache Spark, and NoSQL databases provide distributed environments for efficient, large-scale data mining.

Recent advances include deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which are applied in big data analytics for image processing, natural language processing, and predictive analytics. Graph mining has also become essential, especially for analyzing relationships and interactions within social networks. Incremental and stream mining techniques address real-time data needs, updating models dynamically, which is

useful in fields like stock trading and fraud detection. Applications of data mining in big data are widespread. In healthcare, it enables personalized treatment recommendations, early disease detection, and predictive diagnostics, using data from electronic health records and genomics. In finance, it aids in fraud detection, credit scoring, and risk assessment through classification, clustering, and anomaly detection techniques. Social media analysis relies on big data mining for sentiment analysis, trend detection, and recommendation systems by analyzing user behaviour and preferences. IOT and smart city applications, such as traffic management, energy optimization, and environmental monitoring, also leverage data mining to handle data generated by IOT devices.Data mining with big data provides numerous opportunities across fields but presents significant technical and ethical challenges.

III. PREVIOUS RELATED WORK DONE

1. Big Data Mining Platforms (Tier I):

Due to the multi-source, massive, heterogeneous and dynamic characteristics of application data involved in a distributed environment, one of the important characteristics of Big Data is computing tasks on the petabytes (PB), even the Exabyte (EB)-level data with a complex computing process. Therefore, utilizing a parallel computer infrastructure, its corresponding programming language support, and software models to efficiently analyze and mine the distributed PB, even EB-level data are the critical goal for Big Data processing to change from “quantity” to “quality”. Currently, Big Data processing mainly depends on parallel programming models like Map Reduce, as well as providing a cloud computing platform of Big Data services for the public. Map Reduce is a batch oriented parallel computing model. There is still a certain gap in performance with relational databases. How to improve the performance of Map Reduce and enhance the real-time nature of large-scale data processing is a hot topic in research. The Map Reduce parallel programming model has been applied in many machine learning and data mining algorithms.

2. Big Data Semantics and Application Knowledge(Tier II):

In privacy protection of massive data, Ye, et al, (2013) proposed a multi-layer rough set model, which can accurately describe the granularity change produced by different levels of generalization and provide a theoretical foundation for measuring the data effectiveness criteria in the anonymization process, and designed a dynamic mechanism for balancing privacy and data utility, to solve the optimal generalization / refinement order for classification. A recent paper on confidentiality protection in Big Data (Machana vajjhala and Reiter 2012) summarizes a number of methods for protecting public release data, including aggregation (such as k-anonymity, I-diversity etc.), suppression (i.e., deleting sensitive values), data swapping (i.e., switching values of sensitive data records to prevent users from matching), adding random noise, or simply replacing the whole original data values at a high risk of disclosure with values synthetically generated from simulated distributions. For applications involving Big Data and tremendous data volumes, it is often the case that data are physically distributed at different locations, which means that users no longer physically possess the storage of their data.

3. Big Data Mining Algorithms (Tier III):

In order to adapt to the multi-source, massive, dynamic Big Data, researchers have expanded existing data mining methods in many ways, including the efficiency improvement of single-source knowledge discovery methods [Chang et al., 2009], designing a data mining mechanism from a multi-source perspective [Wu and Zhang, 2003; Wu et al., 2005; Zhang et al., 2005], as well as the study of dynamic data mining methods and the analysis of convection data [Domingos and Hulten, 2000; Chen et al, 2005]. The main motivation for discovering knowledge from massive data is improving the efficiency of single-source mining methods. On the basis of gradual improvement of computer hardware functions, researchers continue to explore ways to improve the efficiency of knowledge discovery algorithms to make them better for massive data.

Knowledge evolution is a common phenomenon in real-world systems. For example, the clinician's treatment programs will constantly adjust with the conditions of the patient, such as family economic status, health insurance, the course of treatment, treatment effects, and distribution of cardiovascular and other chronic epidemiological changes with the passage of time. In the knowledge discovery process, concept drifting aims to analyze the phenomenon of implicit target concept changes or even fundamental changes triggered by context changes in data streams. According to different types of concept drifts, knowledge evolution can take forms of mutation drift, progressive drift, and data distribution drift, based on single features, multiple features, and streaming features [Wu et al., 2013].

IV. THE PROPOSED WORK

The proposed work focuses on addressing core challenges in data mining with big data, aiming to enhance data processing efficiency, accuracy, and scalability. Key areas of focus include developing optimized algorithms, implementing advanced frameworks, and integrating secure, privacy-preserving methods. The work will target the following objectives:

1. Efficient Data Pre-processing and Integration
to handle the variety and volume of big data, the proposed work will develop robust pre-processing techniques for cleaning, transforming, and normalizing data from diverse sources. A hybrid data integration approach will be designed to unify different data formats (structured, semi-structured, unstructured) while maintaining data quality. This integration will facilitate seamless data preparation for mining processes and help ensure consistency and accuracy.

2. Scalable Classification and Clustering Algorithms
existing data mining algorithms often struggle with the scale of big data. The work will adapt and enhance traditional classification (e.g., decision trees, support vector machines) and clustering algorithms (e.g., k-means, hierarchical clustering) for scalability using distributed computing. Utilizing frameworks like Apache Spark, these optimized algorithms will support real-time data analysis across various fields such as finance, healthcare, and social media.

3. Advanced Deep Learning Techniques for Pattern Recognition
To improve pattern recognition and predictive capabilities, the work will implement deep learning models such as convolutional neural networks (CNNs) for image and video analysis and recurrent neural networks (RNNs) for time-series data. These models will leverage big data frameworks to handle high-dimensional data, supporting applications in real-time surveillance, anomaly detection, and trend analysis.

4. Privacy Preserving Data Mining Methods:
With growing privacy concerns, the proposed work will incorporate privacy-preserving techniques such as differential privacy and federated learning. Differential privacy will ensure user anonymity during data mining processes, while federated learning will enable decentralized training of models on local devices. This dual approach will support secure data mining without compromising user privacy.

5. Real-Time Data Stream processing
As big data often includes high-velocity data streams (e.g., IoT devices, social media feeds), the work will develop a real-time stream processing module. This module will use incremental learning techniques to update models dynamically, enabling timely insights and predictions in applications such as traffic management, financial trading, and health monitoring.

6. Evaluation and Performance Metrics
to assess the effectiveness of the proposed methods, evaluation metrics will include accuracy, precision, recall, F1 score, processing time, and scalability. The performance of these models will be benchmarked against standard big data datasets, with improvements analyzed across different domains.

7. Implementation on Big Data Frameworks
The entire proposed system will be implemented on distributed frameworks like Apache Hadoop and Spark, leveraging their parallel processing and storage capabilities to achieve high performance. This implementation will help evaluate how effectively these platforms manage and process extensive datasets while maintaining accuracy and response times.

V. IMPLEMENTATION

The implementation of data mining in big data environments requires a robust, multi-layered approach to handle the scale, variety, and velocity of the data involved. Big data projects typically start with a structured approach to data collection, pre-processing, and integration. Given the often diverse sources of big data—ranging from transactional systems and IoT devices to social media platforms and cloud repositories—data collection is a complex process that demands careful planning. This stage involves setting up pipelines for automated data extraction from these sources, leveraging APIs or connectors to ingest data efficiently and without manual intervention. Following collection, the data must go through rigorous cleaning to address

common quality issues, including missing values, duplicates, and outliers, as well as noise reduction, which is critical for ensuring reliable analysis. Standard data cleaning techniques, like imputation for missing values and normalization for scaling, play an important role here. Data integration is then achieved through Extract, Transform, Load (ETL) tools, which not only streamline the merging of data from various sources but also perform essential transformations to ensure compatibility across formats, thus preparing the data for efficient storage and processing.

Once the data is prepared, the next stage involves selecting the appropriate storage and processing infrastructure to handle the large-scale data efficiently. Distributed file systems, particularly the Hadoop Distributed File System (HDFS), are often foundational in this regard, as they provide a fault-tolerant means of storing large data volumes across multiple nodes in a cluster. The distributed nature of HDFS enables faster access to data and ensures data redundancy, which is essential for resilience in big data projects. Alongside storage solutions, parallel processing frameworks like Apache Hadoop and Apache Spark play a crucial role in executing data mining algorithms at scale. Hadoop's Map Reduce paradigm enables the division of tasks across different nodes, optimizing computation times for batch processing workloads. Apache Spark, known for its in-memory computing capabilities, provides an edge when handling iterative algorithms, making it a popular choice for big data projects that demand quick, real-time analysis. Additionally, NoSQL databases like MongoDB, Cassandra, and HBase allow for flexible data storage, which is highly suited to unstructured data, supporting high-throughput reads and writes essential in real-time applications.

At the core of the data mining process are algorithms tailored to specific analytical goals, such as clustering, classification, pattern mining, and anomaly detection. Clustering and classification algorithms are widely used in big data environments to segment data into meaningful groups and to predict categories based on features, respectively. Clustering methods, like k-means and hierarchical clustering, are frequently implemented in Spark and Hadoop environments, capitalizing on their parallel processing capabilities to handle large datasets efficiently. Classification algorithms, such as decision trees, random forests, and support vector machines, can be run on distributed frameworks to identify classes or labels for data points. In cases where discovering relationships between data points is necessary, frequent pattern mining algorithms like Apriori and FP-growth are deployed to identify associations and correlations within large datasets, which can reveal insights like market basket patterns in retail. For anomaly detection, algorithms like DBSCAN, Isolation Forest, and one-class SVM are employed to identify outliers, which is essential in applications like fraud detection and network security. The use of these data mining algorithms, especially on distributed platforms, allows the system to scale with the data, ensuring that insights remain actionable even as data volumes grow.

Finally, the results generated by data mining must be interpreted, visualized, and shared with relevant stakeholders. Visualization frameworks such as Tableau, Power BI, and D3.js are often employed to present findings in an accessible manner, making it easier for decision-makers to leverage data insights for strategic purposes. Reporting is often automated, with dashboards updated in real-time to reflect the latest trends and anomalies, ensuring that users have current insights at their fingertips. Feedback loops are often integrated into the system, allowing continuous refinement of algorithms and models based on changing data trends. This end-to-end process enables the transformation of raw data into valuable insights, empowering organizations to make informed decisions based on the patterns and trends uncovered through data mining.

In sum, implementing data mining with big data requires careful orchestration of multiple technologies and techniques across the data lifecycle, from ingestion and processing to analysis and visualization. Distributed computing frameworks, storage systems, and data mining algorithms work together to extract valuable insights from vast datasets, enabling real-time analytics and predictive capabilities essential in today's data-driven environment.

VI. CONCLUSION

Driven by real-world applications and key industrial stakeholders and initialized by national funding agencies, managing and mining Big Data have shown to be a challenging yet very compelling task. While the term Big Data literally concerns about data volumes, our HACE theorem suggests that the key characteristics of the Big Data are (1) huge with heterogeneous and diverse data sources, (2) autonomous with distributed and decentralized control, and (3) complex and evolving in data and knowledge associations. Such combined characteristics suggest that Big Data requires a "big mind" to consolidate data for maximum values (Jacobs 2009). In order to explore Big Data, we have analyzed several challenges at the data, model, and system levels. To support Big Data mining, high performance computing platforms are required which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection

environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise and errors can be introduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is a major challenge. At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. This requires carefully designed algorithms to analyze model correlations between distributed sites, and fuse decisions from multiple sources to gain a best model out of the Big Data. At the system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and data sources, along with their evolving changes with time and other possible factors. A system needs to be carefully designed so that unstructured data can be linked through their complex relationships to form useful patterns, and the growth of data volumes and item relationships should help form legitimate patterns to predict the trend and future. We regard Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineering domains. With Big Data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real-time. We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. The era of Big Data has arrived.

VII. REFERENCES

[1] Ahmed, R., & Karypis, G. (2012). Algorithms for mining the evolution of conserved relational states in dynamic networks. *Knowledge and Information Systems, 33*(3), 603-630.

[2] Alam, M. H., Ha, J., & Lee, S. (2012). Novel approaches to crawling important pages early. *Knowledge and Information Systems, 33*(3), 707-734.

[3] Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science, 337*, 337-341.

[4] Machanavajjhala, A., & Reiter, J. P. (2012). Big privacy: protecting confidentiality in big data. *ACM Crossroads, 19*(1), 20-23.

[5] Banerjee, S., & Agarwal, N. (2012). Analyzing collective behavior from blogs using swarm intelligence. *Knowledge and Information Systems, 33*(3), 523-547.

[6] Birney, E. (2012). The making of ENCODE: Lessons for big-data projects. *Nature, 489*, 49-51.

[7] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1-8.

[8] Borgatti, S., Mehra, A., Brass, D., & Labianca, G. (2009). Network analysis in the social sciences. *Science, 323*, 892-895.

[9] Bughin, J., Chui, M., & Manyika, J. (2010). Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. *McKinsey Quarterly*.

[10] Centola, D. (2010). The spread of behavior in an online social network experiment. *Science, 329*, 1194-1197.

[11] Chang, E. Y., Bai, H., & Zhu, K. (2009). Parallel algorithms for mining large-scale rich-media data. In *Proceedings of the 17th ACM International Conference on Multimedia (MM '09)* (pp. 917-918). New York, NY, USA.

[12] Chen, R., Sivakumar, K., & Kargupta, H. (2004). Collective mining of Bayesian networks from distributed heterogeneous data. *Knowledge and Information Systems, 6*(2), 164-187.

[13] Chen, Y.-C., Peng, W.-C., & Lee, S.-Y. (2012). Efficient algorithms for influence maximization in social networks. *Knowledge and Information Systems, 33*(3), 577-601.

[14] Chu, C. T., Kim, S. K., Lin, Y. A., Yu, Y., Bradski, G. R., Ng, A. Y., & Olukotun, K. (2006). MapReduce for machine learning on multicore. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS '06)* (pp. 281-288). MIT Press.

[15] Cormode, G., & Srivastava, D. (2009). Anonymized data: Generation, models, usage. In *Proceedings of SIGMOD* (pp. 1015-1018).

[16] Das, S., Sismanis, Y., Beyer, K. S., Gemulla, R., Haas, P. J., & McPherson, J. (2010). Ricardo: Integrating R and Hadoop. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data (SIGMOD '10)* (pp. 987-998).

[17] Dewdney, P., Hall, P., Schilizzi, R., & Lazio, J. (2009). The square kilometre array. *Proceedings of the IEEE, 97*(8).

[18] Domingos, P., & Hulten, G. (2000). Mining high-speed data streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '00)* (pp. 71-80).

[19] Duncan, G. (2007). Privacy by design. *Science, 317*, 1178-1179.

[20] Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association, 89*(426), 463-475.

[21] Ghoting, A., & Pednault, E. (2009). Hadoop-ML: An infrastructure for the rapid implementation of parallel reusable analytics. In *Proceedings of the Large-Scale Machine Learning: Parallelism and Massive Datasets Workshop (NIPS 2009)*.

[22] Gillick, D., Faria, A., & DeNero, J. (2006). MapReduce: Distributed computing for machine learning. Berkeley, December 18, 2006.

[23] Helft, M. (2008). Google uses searches to track flu's spread. *The New York Times*. Retrieved from http://www.nytimes.com/2008/11/12/technology/internet/12flu.html

[24] Howe, D., et al. (2008). Big data: the future of biocuration. *Nature, 455*, 47-50, September 2008.

[25] Huberman, B. (2012). Sociology of science: Big data deserve a bigger audience. *Nature, 482*, 308.

[26] IBM. (2012). What is big data: Bring big data to the enterprise. Retrieved from http://www01.ibm.com/software/data/bigdata/

[27] Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM, 52*(8), 36-44.

[28] Kopanas, I., Avouris, N., & Daskalaki, S. (2002). The role of domain knowledge in a large-scale data mining project. In Vlahavas, I. P., & Spyropoulos, C. D. (Eds.), *Methods and Applications of Artificial Intelligence*, Lecture Notes in AI, LNAI 2308 (pp. 288-299). Springer-Verlag, Berlin.

[29] Labrinidis, A., & Jagadish, H. (2012). Challenges and opportunities with big data. In *Proceedings of the VLDB Endowment, 5*(12), 2032-2033.

[30] Lindell, Y., & Pinkas, B. (2000). Privacy preserving data mining. *Journal of Cryptology*, 36-54.

[31] Liu, W., & Wang, T. (2012). Online active multi-field learning for efficient email spam filtering. *Knowledge and Information Systems, 33*(1), 117-136.

[32] Lorch, J., Parno, B., Mickens, J., Raykova, M., & Schiffman, J. (2013). Shoroud: Ensuring private access to large-scale data in the data center. In *Proceedings of the 11th USENIX Conference on File and Storage Technologies (FAST'13)*, San Jose, CA.

[33] Luo, D., Ding, C., & Huang, H. (2012). Parallelization with multiplicative algorithms for big data mining. In *Proceedings of the IEEE 12th International Conference on Data Mining* (pp. 489-498).

[34] Mervis, J. (2012). U.S. science policy: Agencies rally to tackle big data. *Science, 336*(6077), 22.

[35] Michel, F. (2012). How many photos are uploaded to Flickr every day and month? Retrieved from http://www.flickr.com/photos/franckmichel/6855169886/

[36] Mitchell, T. (2009). Mining our reality. *Science, 326*, 1644-1645.

[37] Nature Editorial. (2008). Community cleverness required. *Nature, 455*(7209), September 4, 2008.

[38] Papadimitriou, S., & Sun, J. (2008). Disco: Distributed co-clustering with MapReduce: A case study towards petabyte-scale end-to-end mining. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM '08)* (pp. 512-521).

[39] Ranger, C., Raghuraman, R., Penmetsa, A., Bradski, G., & Kozyrakis, C. (2007). Evaluating MapReduce for multi-core and multiprocessor systems. In *Proceedings of the 13th IEEE International Symposium on High Performance Computer Architecture (HPCA '07)* (pp. 13-24).

[40] Rajaraman, A., & Ullman, J. (2011). Mining of massive datasets. *Cambridge University Press*.

[41] Reed, C., Thompson, D., Majid, W., & Wagstaff, K. (2011). Real-time machine learning to find fast transient radio anomalies: A semi-supervised approach combining detection and RFI excision. *International Astronomical