ERROR AWARE MINING WITH BIG DATA

K.Aishwarya

Information Technology, Vivekanandha College of Enginnering for Women - Anna University
Elayampalayam, Thiruchengode-637 205, Namakkal Dt, Tamil Nadu, India
aishwaryait13@gmail.com

***Abstract*— Big Data is used to identify the datasets that due to their large size and complexity, we can not manage them with our current methodologies or data mining software tools. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it.Big Data problems complex, enormous volume, growing data sets with frequent, autonomous sources. With the quick development of networking, data storage, and also the data assortment capability, Big data are currently apace increasing altogether science and engineering domains, as well as physical, biological and bio-medical sciences. The size of those data will increase, the number of digressive knowledge sometimes will increase still and also the method becomes impractical. Hence, in such cases, the analyst should be capable of specializing in the informational data while ignoring the noise data. These styles of difficulties complicate the analysis of multichannel data as compared with the analysis of single-channel data. This paper provides HACE theorem that characterizes the options of the massive information revolution, and proposes an enormous processing model of Big Data, from the data mining perspective and conjointly think about error-aware (EA) data mining style, that takes advantage of applied math error data (such as noise level and noise distribution) to boost data mining results.**

***Keywords*— Big Data, Data Mining, Heterogeneity, Autonomous, Complex and Evolving associations, Error-Aware.**

I. INTRODUCTION

The key test for the Big Data is to investigate the substantial volumes of data and extract useful information or data for future task. In numerous circumstances, the data extraction process must be extremely effective and close to real-time because storing all observed data is nearly infeasible. For instance, the Square Kilometer Array (SKA) in Radio Astronomy comprises of 1,000 to 1,500 15-meter dishes in a focal 5km region. It furnishes 100 times additional sensitive vision than any existing radio telescopes, replying key inquiries regarding the Universe. Nonetheless, with a 40 gigabytes(GB)/second data volume, the data created from the SKA is extraordinarily substantial. Although researchers have confirmed that fascinating patterns, like transient radio anomalies is discovered from the SKA data, previous methods are incapable of handling this Big data. As a result, the aberrant data volumes need a good data analysis and prediction platform to apprehend fast-response and real-time classification for such Big Data.

REAL-WORLD data are dirty, and so, noise handling may be a shaping characteristic for data mining analysis and applications. There are four major steps in data mining: data assortment and preparation, data transformation and quality sweetening, pattern discovery, and interpretation and analysis of patterns (or postmining mining). Within in the Cross Industry Standard Process for Data Mining framework, this method is rotten into six major phases: understanding business, understanding data, data preparation, modeling, evaluation, and readying. It is expected that the entire method starts with data and finishes with the extracted data. As a result of its data-driven nature, old analysis efforts have over that data mining results crucially accept the standard of the underlying data, and for most of the data mining applications,the method of data collection, data preparation and data sweetening price the majority of the project budget and conjointly the developing time circle. However, data imperfections, such s erroneous or inaccurate attribute values, still ordinarily exist in follow, where data typically carry a big quantity of errors, which will have a negative impact on the mining algorithms. Additionally,existing analysis on privacy-preserving data mining typically uses intentionally injected errors, that are ordinarily noted as data perturbations,for privacy protective functions, such that sensitive data in data records may be protected, but data within the dataset continues to be obtainable for mining. As these systematic or semi synthetic errors can eventually deteriorate the data quality, conducting effective mining from data imperfections becomes a difficult and a real issue for the data mining community.

II. BIG DATA CHARACTERISTICS: HACE THEOREM

Big Data begins with large-volume, heterogeneous and autonomous sources with dispersed and decentralized control, and seeks to analyze complex and evolving relationships a part of the data. These characteristics accomplish it an acute claiming for advantageous ability from the Big Data.

A. Enormous data with heterogeneous and assorted dimensionalities.

One of the axiological characteristics of the Big Data is the enormous volume of data represented by heterogeneous and assorted dimensionalities. This is in light of the fact that diverse data gatherers utilize their own particular schemata for data recording, and the nature of distinctive provisions likewise brings about different representations of the data. Case in point, each one single human being in a bio-medical world could be spoken to by utilizing straightforward demographic data, such as sexual orientation, age ancestor's disease history and so on. For X-ray assay and CT browse of each individual, images or videos are acclimated to represent the results because they accommodate be held advice for doctors to backpack abundant examinations. For a DNA or genomic accompanying test, microarray announcement images and sequences are acclimated to represent the genetic code information because this is the way that our accepted techniques access the data. Under such circumstances, the heterogeneous features accredit to the various types of representations of the aforementioned individuals, and assorted dimensionalities accredit to the variety of the features involved to represent each single observation. Imagine that different organizations (or bloom practitioners) may accept their own schemata to represent each patient, heterogeneity and assorted dimensionality become above challenges if we are aggravating to accredit data accession by accumulation data from all sources.

B. *Autonomous Sources With Dispersed And Decentralized Control*

Autonomous data sources with dispersed and decentralized controls are a main characteristic of Big Data. Being Autonomous , every data source has the capacity to create and gather data without including (or depending on) any incorporated or centralized control. This is like the World Wide Web (WWW) setting where each one web server furnishes a certain measure of data and every server has the capacity to completely work 5without essentially depending on different servers. the tremendous volumes of the data also make an application prone to to attacks or malfunctions, if the accomplished arrangement has to await on any centralized control unit. For significant Big Data related requisitions, for example, Google, Flicker, Facebook, and Walmart, countless ranches are conveyed everywhere throughout the world to guarantee persevering administrations and snappy reactions for local markets. Such autonomous sources are the results of the specialized outlines, as well as the effects of the enactment and the regulation controls in diverse countries/regions.

C. *Complex And Evolving Relationships*

While the aggregate of the Big Data increases, so do the intricacy and the relationships beneath the data. At an early stage of data centralized systems, the focus is on finding best feature values to represent each observation. This is like utilizing various information fields, for example, age, sex, wage, instruction foundation and so on., to portray every person. 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. Individuals structure companion rings dependent upon their normal interests or associations by organic relationships. Such social associations ordinarily exist in our day by day exercises, as well as are exceptionally prevalent in virtual worlds. For example, Facebook or Twitter, are primarily portrayed by social capacities, for companion associations and supporters (in Twitter). The correlations amid individuals inherently complicate the accomplished abstract representation and any acumen process. In the sample-feature representation, individuals is admired agnate if they allotment agnate affection values, admitting in the sample-feature-relationship representation, two individuals can be affiliated calm (through their amusing connections) even admitting their ability allotment annihilation in accepted in the affection domains at all. In a dynamic world, the features defines the individuals and the social ties defines our connections may also evolve with account to temporal, spatial, and other factors. Such a muddling is getting to be some piece of the actuality of Big Data applications, where the key is to take data relationships, in addition to the advancing progressions, into attention, to uncover of service examples from Big Data accumulations.

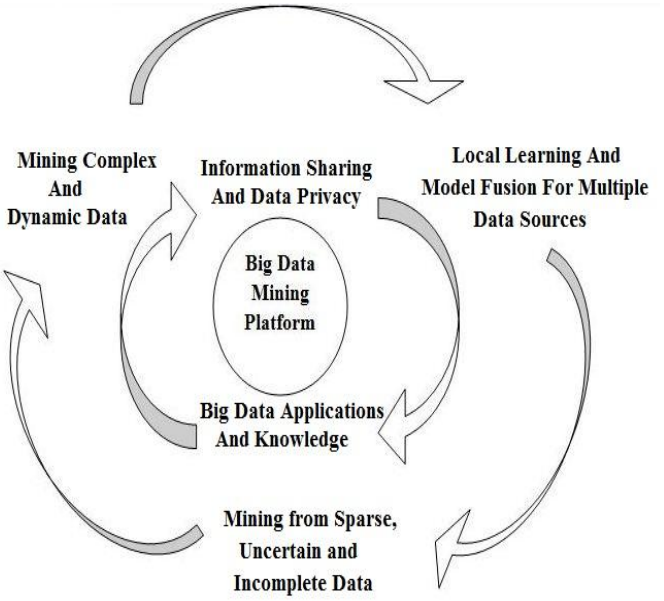
Mining Complex And Dynamic Data
Information Sharing And Data Privacy
Local Learning And Model Fusion For Multiple Data Sources
Big Data Mining Platform
Big Data Applications And Knowledge
Mining from Sparse, Uncertain and Incomplete Data

Figure 1: Big Data framework

This framework forms a three tier structure which around the mining platform of Big Data which is Tier I, it deals with accessing of low-level data and computing. Challenges on sharing of information and privacy, and Big Data domains and knowledge form Tier II, which focus on high level semantics, domain knowledge, and issues of user privacy. The outmost circle is Tier III challenges on actual mining algorithms of Big data.

III. CHALLENGES IN DATA MINING WITH BIG DATA

For an able acquirements database arrangement to handle Big Data, the capital key is to scale up to the awfully large-volume of data and accommodate treatments for the characteristics featured by the above HACE theorem. Figure 2 shows a the Big Data processing framework, which has three tiers from inside out with concerns on data accessing and computing (Tier I), data privacy and domain knowledge (Tier II), and Big Data mining algorithms (Tier III).

A. Tier I: Big Data Mining Platform

In regular data mining frameworks, the mining techniques require computational escalated processing units for data assays and comparisons. A computing platform is accordingly bare to accept admission to, at least, two types of resources: data and computing processors. For small scale data mining assignments, a solitary desktop PC, which holds hard disk and CPU processors, is sufficient to satisfy the data mining objectives. For medium scale data mining assignments, data are commonly vast (and perhaps dispersed) and can't be fit into the capital(main) memory. Common solutions are to await on parallel computing or collective mining to sample and accumulated data from various sources, then use parallel computing programming (such as the Message Passing Interface) to do the mining process. For Big Data mining, since information scale is far past the limit that a solitary (PC) can deal with, an ordinary Big Data processing framework will depend on group workstations with a high execution processing stage, where an information mining errand is conveyed by running some parallel customizing devices, for example, Mapreduce or ECL (Enterprise Control Language), on an extensive number of figuring hubs (i.e., clusters).

B. Tier II: Big Data Semantics and Application Knowledge

Semantics and ***application knowledge*** in Big Data accredit to abundant aspects accompanying to the regulations, policies, client knowledge, and domain information. The two most significant issues at this level incorporate data sharing and privacy and domain and application knowledge. The above provides answers to boldness apropos on how abstracts are maintained, accessed, and imparted.

C. Tier III: Big Data Mining Algorithms

1) Local Learning And Model Fusion For Multiple Data Sources

As Big Data provisions are emphasized with autonomous sources and decentralized controls, accumulating conveyed data sources to an incorporated site for mining is methodically restrictive because of the potential transmission expense and protection concerns. Then again, despite the fact that we can dependably do mining exercises at each one conveyed site, the inclined perspective of the information gathered at every distinctive site regularly prompts predispositioned choices or models, much the same as the elephant and blind men case. Under such a condition, a Big Data mining framework need to empower a data trade and combination instrument to guarantee that all appropriated destinations (or data sources) can cooperate to attain a worldwide advancement objective. Model mining and correspondences are the key steps to guarantee that models or examples uncovered from numerous data sources might be merged to meet the worldwide mining destination.

2) Mining from Sparse, Uncertain and Incomplete Data

Sparse, uncertain, and incomplete data are defining appearance for Big Data applications. Being sparse, the amount of data credibility is too few for drawing reliable conclusions. This is commonly a aggravation of the data dimensionality issues, where data in a top dimensional space (such as more than 1000 dimensions) does not appear bright trends or distributions. For a lot of machine learning and data mining algorithms, high dimensional spare data decidedly adulterate the adversity and the reliability of the models acquired from the data. Common approaches are to dimension abridgement or feature selection to abate the data dimensions or to anxiously cover added samples to abatement the data scarcity, such as all-encompassing unsupervised acquirement methods in data mining.

3) Mining Complex And Dynamic Data

The ascent of Big Data is determined by the fast expanding of complex data and their progressions in volumes and in nature. Reports posted on WWW servers, Internet spines, social networks, correspondence systems, and transportation systems and so forth are all offered with complex information. While complex reliance structures underneath the information raise the trouble for our taking in frameworks, they likewise offer energizing chances that straightforward data representations are unequipped for accomplishing. Case in point, analysts have effectively utilized Twitter, a well-known person to person communication office, to distinguish occasions, for example, quakes and real social exercises, with about online speed and quite high precision. Moreover, the information of individuals' inquiries to web crawlers likewise empowers another unanticipated cautioning framework for recognizing quick spreading influenza episodes.

IV. ERROR-AWARE DATA MINING

In the Cross Industry Standard Action for Data Mining action is addle into six above phases: business understanding, data comprehending, data preparation, demonstrating, evaluation, and deployment. It is expected that the accomplished action starts with raw data and finishes with the extracted knowledge. Because of its data-driven attributes , antecedent analysis efforts accept assured that data mining after-effects crucially await on the superior of the base data, and for a lot of data mining applications, the action of data collection, data preparation, and data accessory cost the majority of the activity account and as well as the developing time circle. However, data imperfections, such as erroneous or inaccurate aspect values, still frequently abide in practice, area data generally backpack a cogent bulk of errors, which will accept a abrogating appulse on the mining algorithms . In addition, absolute analysis of privacy-preserving data mining generally uses carefully injected errors, which are referred to as data perturbations, for preserving privacy , such that the acute advice in abstracts annual can be protected, but ability in the dataset is still accessible for mining. As these analytical or counterfeit errors will eventually adulterate the data quality, administering able mining from data imperfections becomes a arduous and absolute affair for the data mining community. Take the botheration of supervised acquirements as an example, an area the assignment is to anatomy accommodation theories that can be acclimated to allocate ahead unlabeled (test) instances accurately. In adjustment to do so, a acquirements set D which consists of an amount of training instances, i.e., (xn, yn), n = 1, 2, . . . , N, is accustomed in advance, from which the acquirements algorithm can assemble a accommodation theory.Each single instance (xn, yn) is described by a set of M attribute values xn = _a1, a2. . . aM_ and one class label yn, yn ∈ {c1, c2, . . . , cL} (the characters of all the symbols are explained in Table I).The problems of data imperfections rise from the absoluteness that attribute values xn and a class label yn might be corrupted and contain incorrect values. Under such circumstances, incorrect attribute values and mislabeled class labels thus constitute attribute and class noises. Extensive research studies have shown that the existence of such data imperfections is mainly responsible for inferior decision theories and eliminating highly suspicious data items often leads to an improved learner (compared with the one learned from the original noisy dataset), because of the enhanced data consistency and less confusion between the underlying data. Such abolishment approaches are frequently referred to as data cleansing. Data cleansing methods are able in abounding scenarios, but some problems are still open.

- *Data cleansing only takes effect on assertive types of errors,* such as class noise. Despite the fact that it has been shown that cleansing class noise often results in good learners, for datasets containing attribute noise or missing attribute values, no evidence suggests that data cleansing can advance to enhanced data mining results.

- *Data cleansing cannot result in absolute data.* As continued as errors continuously abide in the data, they will a lot of acceptable adulterate the mining achievement in some means (although exceptions do exist). Consequently, the charge for developing error-tolerant abstracts mining algorithms has been a major affair with the area.

- *Data cleansing cannot be unconditionally* activated *to any data sources.* For carefully imposed errors, such as privacy-preserving data mining, data cleansing cannot be directly applied to cleanse the noisy data records because privacy-preserving data mining intends for acute sensitive information by data randomization. Applying data cleansing to that data could lead to information loss and intensely crumble the last comes about.

- *Eliminating noisy data may accelerate to information Loss***.** Just because a noisy instance contains erroneous attribute values or an incorrect class label, it doesn't fundamentally imply that this occurrence is totally pointless and hence needs to be killed from the database. More particularly, the facts might confirm that eliminating class noise from the training dataset is often beneficial for an accurate learner, but for erroneous aspect values, we may not artlessly annihilate a noisy instance from the dataset back added actual attribute values of instance may still accord to the acquirements process.

It is accessible that instance-based error information (i.e., advice about which instance and/or which aspect ethics of the instance are incorrect) is difficult to get and bare with atomic endeavors, although a substantial amount of analysis has been aggravating to abode this affair from altered perspectives.However, there are abounding cases in absoluteness that statistical error Information on the accomplished database is known as a priori.

- *Information transformation errors*. Information transformation, decidedly wireless networking, generally raises an assertive bulk of the errors in announced data. For error control purposes, the statistical errors of the signal transmission channel should be advised in beforehand and can be acclimated to appraisal the absurd amount in the adapted information.

- *Device errors.*_The point when gathering data from various device, the incorrectness level of every mechanism is frequently possible, as it is some piece of the framework characteristics. Case in point, fluorescent marking for gene chips in microarray analyzes typically holds incorrectness initiated by sources, for example, the impact of foundation force. The qualities of gathering gene chip information are regularly connected with likelihood to demonstrate the unwavering quality of the present worth.

- *Data discretization errors*. Data discretization is an accepted action of discretizing the area of a connected capricious into a bound amount of intervals. Because this action uses an assertive amount of detached ethics to appraisal absolute connected values, the aberration amid the detached amount and the absolute amount of the connected capricious appropriately leads to an accessible error. Such discretization errors can be abstinent in beforehand and, therefore, are accessible for a data mining

- *Data perturbation errors***.** As a representative example of artificial errors, protection safeguarding data mining carefully perturbs the data; accordingly, private data in information records might be secured, yet learning passed on in the datasets is still minable. In such cases, the levels of errors are absolutely accepted for abstracts mining algorithms. The availability of the above statistical absurdity advice anon leads to the catechism of how to board such advice into the mining process. Most abstracts mining methods, however, do not board such absurdity advice in their algorithm design. They either yield blatant abstracts as superior sources or accept abstracts cleansing advanced to annihilate and/or actual the errors. Either way may appreciably adulterate the achievement of the afterwards abstracts mining algorithms because of the abrogating appulse of abstracts errors and the limitations and applied issues of abstracts cleansing. The above observations accession an absorbing and important affair on error-aware (EA) abstracts mining, area ahead accepted absurdity advice (or noise knowledge)) might be joined into the digging procedure for enhanced mining outcomes.

V. PROCESSING MULTICHANNEL RECORDING FOR DATA MINING ALGORITHM

Representing multichannel (or multivariable) data in a simple way is an important issue in data analysis. One of the most well-known techniques for achieving this complexity reduction is called *quantization* (or *discretization*), which is the process of converting a continuous variable into a discrete variable. The discretized variable has a finite number of values, which is considerably smaller than the number of possible values in the empirical data set. The discretization improves the information representation, enhances interpretability of effects, and makes information open to additional information mining strategies.In decision trees, quantization as a preprocessing step is preferable to a local quantization process as part of the decision tree building algorithm One could approach the discretization process by discretizing all variables at the same time (global), or each one separately (local). The methods may use all of the available data at every step in the process (global) or concentrate on a subset of data (local) depending on the current level of discretization. Decision trees, for instance, are usually local in both senses. Furthermore, the two following search procedures could be employed. The top-down approach starts with a small number of bins, which are iteratively split further. The bottom-up approach, on the other hand, starts with a large number of narrow bins which are iteratively merged. In cases, a particular split or merge operation is based on a defined performance criterion, which can be global (defined for all bins) or local (defined for two adjacent bins only).

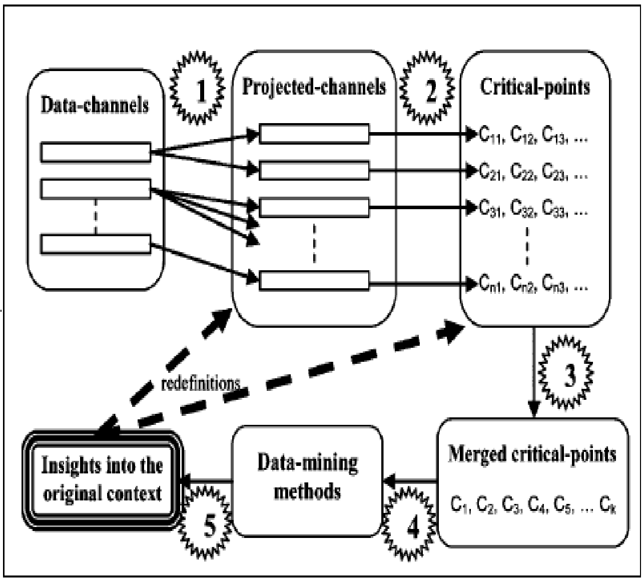
Data-channels
1
Projected-channels
2
Critical-points
$C_{11}, C_{12}, C_{13}, \ldots$
$C_{21}, C_{22}, C_{23}, \ldots$
$C_{31}, C_{32}, C_{33}, \ldots$
$C_{n1}, C_{n2}, C_{n3}, \ldots$
3
redefinitions
Insights into the original context
5
Data-mining methods
4
Merged critical-points
$C_1, C_2, C_3, C_4, C_5, \ldots C_k$

Figure 2:. Summary of analysis stages.

Four quantization methods:

- *Equal Width Interval:*
  *T*he simplest and frequently Applied discretization method is to divide the data range into a predetermined number of bins.

- *Maximum Entropy:*
  This method is to create bins so that each bin equally contributes to represent the input data. In other words, the probability of each bin of the data must be approximately equal.

- *Maximum Mutual Information:*
  *I*t is important to optimize the quantized representation with regard to the distribution of the output variable in classification problems. mutual information may be employed to measure information about the output preserved in the discretized variable. Mutual information was used in

the discretization process of the decision tree construction algorithm (ID3) in.

- *Maximum Mutual Information with Entropy:* By combining and the mutual information approaches and the maximum entropy, to obtain a

solution with the merits of both. One would like to retain balanced bins that turn out to be more reliable (prevent overfitting in this context) but simultaneously optimize the binning for classification.

Figure.2 illustrates all five stages in the analysis process. Note that when there are no result insights, the dashed arrows are used. In such a case, a redefinition of either how to create the *projected-channels* or how to detect the *critical-points* is carried out. Our technique is based on five independent stages. The following is a brief summary of all stages involved in the final analysis.

i) Each and Every channel of the original data is projected into the one or more *projected-channels*.

ii) Each of the *projected-channels* which has a critical-points sets are detected and analyzed.

iii) Once all sets are ready, they are alloyed into one large set which can be interpreted as a long series of contest forth the time scale.

iv) Next stage is activating the desired data-mining algorithm.

v) Then output is translated into insights in the context of the original data.

Here *Projection-function* is a function which maps a channel with another space. The resulting space defines a selected projection of the original channel that is relevant for its analysis. It can be represented as a pair <*id, func*> where *id* is the identifier of the channel to be projected and *func* is the function to be applied on it. A *critical-point* is a time-point in which an important change occurs in the some property of a *projected-channel which* can be seen as a triple <*t,id,R*> where "t" is the time at the change occurred, "*id"* is identifier of the *projected- channel* which holds information about the change.

VI. CONCLUSION

Driven by real-world applications and key automated stakeholders and initialized by national allotment agencies, managing and mining Big Data accept apparent to be an arduous yet actual acute task. While the appellation Big Data actually apropos about data volumes, HACE assumption suggests that the key characteristics of the Big Data are (1) *Enormous data with heterogeneous and assorted dimensionalities, (2) Autonomous Sources With Dispersed And Decentralized Control,* (3) *Complex And Evolving Relationships.* Such accumulated characteristics advance that Big Data requires a "big mind" to consolidate data for best values.To abutment Big Data mining, top achievement accretion platforms are appropriate which appointed analytical designs to absolve the abounding ability of the Big Data. Such An EA data mining framework seamlessly unifies statistical error advice and a data mining algorithm for able learning. Application noise ability the archetypal congenital from noise-corrupted data is modified, and it has resulted in an abundant advance in allegory with the models congenital from the aboriginal blatant data and the noise-cleansed data. Data mining from noisy information advice sources involves three capital tasks : noise identification, noise profiling, and noise-tolerant mining. Data cleansing deals with noise identification. The EA data mining framework makes use of the statistical noise ability for noise-tolerant mining.

REFERENCES

[1] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, "Data Mining with Big Data", Knownledge and Data Engineering, vol. 26, no. 1, pp 97-107, Jan. 2014

[2] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution
a. Of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.

[3] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.

[4] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.

[5] Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.

[6] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.

[7] E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," Nature, vol. 489, pp. 49-51, 2012.

[8] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[9] S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences," Science, vol. 323, pp. 892-895, 2009.

[10] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinSey Quarterly, 2010.

[11] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," Science, vol. 329, pp. 1194-1197, 2010.

[12] E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for Mining Large-Scale Rich-Media Data," Proc. 17th ACM Int'l Conf. Multimedia, (MM '09,) pp. 917-918, 2009.

[13] R. Chen, K. Sivakumar, and H. Kargupta, "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data," Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187, 2004.

[14] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, "Efficient Algorithms for Influence Maximization in Social Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.

[15] C.T. Chu, S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, and K. Olukotun, "Map-Reduce for Machine Learning on Multicore,"

Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS'06), pp. 281-288, 2006.

[16] G. Cormode and D. Srivastava, "Anonymized Data: Generation, Models, Usage," Proc. ACM SIGMOD Int'l Conf. Management Data, pp. 1015-1018, 2009.

[17] S. Das, Y. Sismanis, K.S. Beyer, R. Gemulla, P.J. Haas, and J.McPherson, "Ricardo: Integrating R and Hadoop," Proc. ACM SIGMOD Int'l Conf. Management Data (SIGMOD '10), pp. 987-998.2010.

[18] P. Dewdney, P. Hall, R. Schilizzi, and J. Lazio, "The Square Kilometre Array," Proc. IEEE, vol. 97, no. 8, pp. 1482-1496, Aug.
a. 2009.

[19] P. Domingos and G. Hulten, "Mining High-Speed Data Streams," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '00), pp. 71-80, 2000.

[20] G. Duncan, "Privacy by Design," Science, vol. 317, pp. 1178-1179, 2007.

[21] B. Efron, "Missing Data, Imputation, and the Bootstrap," J. Am.
a. Statistical Assoc., vol. 89, no. 426, pp. 463-475, 1994.

[22] Ghoting and E. Pednault, "Hadoop-ML: An Infrastructure for the Rapid Implementation of Parallel Reusable Analytics," Proc. Large-Scale Machine Learning: Parallelism and Massive Data Sets
a. Workshop (NIPS '09), 2009.

[23] D. Gillick, A. Faria, and J. DeNero, MapReduce: Distributed Computing for Machine Learning, Berkley, Dec. 2006.

[24] D. Howe et al., "Big Data: The Future of Biocuration," Nature, vol. 455, pp. 47-50, Sept. 2008.

[25] B. Huberman, "Sociology of Science: Big Data Deserve a Bigger
a. Audience," Nature, vol. 482, p. 308, 2012.

[26] "IBM What Is Big Data: Bring Big Data to the Enterprise," http://
a. www-01.ibm.com/software/data/bigdata/, IBM, 2012.

[27] Labrinidis and H. Jagadish, "Challenges and Opportunities with Big Data," Proc. VLDB Endowment, vol. 5, no. 12, 2032-2033,
a. 2012.

[28] M. R. Chmielewski and J. W. Grzynala-Busse, "Global discretization of continuous attributes as preprocessing for machine learning," *Int. J. Approximate Reasoning*, vol. 15, pp. 319–331, 1996.

[29] J. Catlett, "On changing continuous attributes into ordered discrete attributes," in *Proc. Machine Learning—EWSL-91*, Mar. 1991, vol. 482, pp. 164–178.

[30] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuousvalued attributes for classification learning," in *Proc. IJCAI-93*, Aug./Sep. 1993, vol. 2, pp. 1022–1027.

[31] Bollen et al. 2011, J. Bollen, H. Mao, and X. Zeng, Twitter Mood Predicts the Stock Market, *Journal of Computational Science*, 2(1):1-8, 2011.

[32] Borgatti S., Mehra A., Brass D., and Labianca G. 2009, Network analysis in the social sciences, *Science*, vol. 323, pp.892-895.

[33] Bughin et al. 2010, J Bughin, M Chui, J Manyika, *Clouds, big data, and smart assets: Ten tech-enabled business trends to watch*, McKinSey Quarterly, 2010.

[34] Centola D. 2010, The spread of behavior in an online social network experiment, *Science*, vol.329, pp.1194-1197.

[35] X. Zhu and X. Wu, "Class noise vs. attribute noise: A quantitative study of their impacts," *Artif. Intell. Rev.*, vol. 22, no. 3/4, pp. 177–210, Nov. 2004.

[36] X. Wu, C. Zhang, and S. Zhang, "Database Classification for Multi-Database Mining," Information Systems, vol. 30, no. 1, pp. 71- 88, 2005.

[37] X. Wu, "Building Intelligent Learning Database Systems," AI Magazine, vol. 21, no. 3, pp. 61-67, 2000.

[38] X. Wu, K. Yu, W. Ding, H. Wang, and X. Zhu, "Online Feature Selection with Streaming Features," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 35, no. 5, pp. 1178-1192, May 2013.

[39] A. Yao, "How to Generate and Exchange Secretes," Proc. 27th Ann. Symp. Foundations Computer Science (FOCS) Conf., pp. 162-167, 1986.

[40] M. Ye, X. Wu, X. Hu, and D. Hu, "Anonymizing Classification Data Using Rough Set Theory," Knowledge-Based Systems, vol. 43, pp. 82-94, 2013.

[41] [56] J. Zhao, J. Wu, X. Feng, H. Xiong, and K. Xu, "Information Propagation in Online Social Networks: A Tie-Strength Perspective," Knowledge and Information Systems, vol. 32, no. 3, pp. 589-608, Sept. 2012.

[42] X. Zhu, P. Zhang, X. Lin, and Y. Shi, "Active Learning From Stream Data Using Optimal Weight Classifier Ensemble," IEEE

[43] Trans. Systems, Man, and Cybernetics, Part B, vol. 40, no. 6, pp. 1607- 1621, Dec. 2010.

[44] D. Wegener, M. Mock, D. Adranale, and S. Wrobel, "Toolkit-Based High-Performance Data Mining of Large Data on MapReduce

[45] Clusters," Proc. Int'l Conf. Data Mining Workshops (ICDMW '09), pp. 296-301, 2009.

[46] C. Wang, S.S.M. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-Preserving Public Auditing for Secure Cloud Storage" IEEE Trans. Computers, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[47] X. Wu and X. Zhu, "Mining with Noise Knowledge: Error-Aware Data Mining," IEEE Trans. Systems, Man and Cybernetics, Part A, vol. 38, no. 4, pp. 917-932, July 2008.

[48] X. Wu and S. Zhang, "Synthesizing High-Frequency Rules from Different Data Sources," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 2, pp. 353-367, Mar./Apr. 2003.