Enhancing Veracity of IoT Generated Big Data in Decision Making

Xiaoli Liu, Satu Tamminen, Xiang Su,
Pekka Siirtola, Juha Röning, Jukka Riekki
Faculty of Information Technology and Electrical Engineering
University of Oulu
Oulu, Finland

Jussi Kiljander, Juha-Pekka Soininen
VTT Technical Research Centre of Finland
Oulu, Finland

***Abstract*—Data are crucial to support decision making. If data have low veracity, decisions are not likely to be sound. Internet of Things (IoT) generates big data with inaccuracy, inconsistency, incompleteness, deception, and model approximation. Enhancing data veracity is important to address these challenges. In this article, we summarize the key characteristics and challenges of IoT, which influence data processing and decision making. We review the landscape of measuring and enhancing data veracity and mining uncertain data streams. Moreover, we propose five recommendations for future development of veracious big IoT data analytics that are related to the heterogeneous and distributed nature of IoT data, autonomous decision-making, context-aware and domain-optimized methodologies, data cleaning and processing techniques for IoT edge devices, and privacy preserving, personalized, and secure data management.**

I. Introduction

As Internet of Things (IoT) expands into various industry areas and our everyday life, we are observing the generation of big IoT data. The big data have become the new oil of digital economy that need to be harnessed to reveal trends, unseen patterns, and hidden correlations, which will enable automated IoT systems to create new knowledge and to perform efficient and correct decision making.

One of the biggest problems in IoT data analytics and knowledge creation is that the data generated by IoT devices is often noisy, incomplete, imprecise, and even misleading. This leads to challenges in data cleaning, mining, contextualization and knowledge discovery. For example, sensor data are normally expected to have noise due to inaccuracy in data measurement, transmission, and possible power failures of sensor devices, which will jeopardize high quality decision making. Therefore, it is critical to both make the data as correct, accurate and truthful as possible, and to have a credible measure for the correctness that can be used in the decision making process.

Veracity, the fourth V of big data, is a term used in data analytics research to cover the topics including data quality, accuracy, correctness, and truthfulness [1]. Early research [2] [3] already states the importance of data veracity and the affect of low quality data on the validity of the results. Many proposals for addressing big data veracity have been introduced [4] [5]. However, several factors characterizing IoT systems, including short latencies, scalability, constrained resources, and heterogeneity of IoT data models, make the data veracity in IoT a unique research challenge when compared to the traditional big data systems such as governmental services, media systems, and healthcare. Moreover, as IoT systems directly interact with the physical world and the decision-making is performed mainly by machines, the data veracity plays even a bigger role, and is thus crucial for user engagement and acceptance of IoT services.

In this paper, we discuss the impact of data veracity in decision making processes, identify challenges of handling big data from the IoT viewpoint, and survey the state-of-the-art solutions and techniques for modelling and enhancing data veracity and mining data streams. Finally, we propose recommendations for future development for understanding the level of veracity and improving the veracity of data for big data analytics in IoT based systems.

The remainder of this article is organized as follows: Section II introduces definition and principles of IoT data veracity, discusses opportunities for business improvement based on veracious big data in decision making, and examines challenges for processing big data from IoT perspective. In Section III, we discuss, in more depth, techniques for measuring and enhancing veracious big data and for mining uncertain data streams. We enumerate recommendations for future development of veracious big data analytics in Section IV and conclude the paper in Section V.

II. Veracious big IoT data and Decision making

In this section, we introduce the definition of data veracity, discuss opportunities for business improvement based on big veracious data and IoT, and examine the challenges of typical IoT systems, such as smart city and smart health.

A. Data Veracity

Krotofil et al. define veracity as a property that an assertion truthfully reflects the aspect it makes a statement about [6]. In the IoT context, data veracity often refers to the problem associated with data usability and quality. Considering an IoT device that is producing and delivering data at some discrete intervals, it may malfunction during some time intervals. Moreover, the data could be corrupted or lost due to noise or as a defect in the communication mechanism. To enhance

the usability of IoT applications and services, it is important to know what information can be calculated instead of the missing data, and how does the missing information affect the actions produced by the IoT decision-making algorithms. In general, it is about how truly the measurement reflects the measurand. Being able to assess the level of veracity of data provides the foundation for trustworthy measurements and is crucial to make sound decisions. Vivek Kale introduces a 6-c characteristics for data and meta-data, including correct, clear, consistent, complete, certain, and confirm [7].

As the other most important aspect of data veracity, Data Quality (DQ) refers to how well data meets the requirements of data consumers. Based on this definition, data quality measurement can be subjective or objective; the first one is based on qualitative evaluations by data administrators and users and the second is based on quantitative metrics [8]. Regarding to subjective measurement, the level of quality of data represents the degree to which data meet the expectations of data consumers, based on their intended usage of the data. Thus, DQ is directly related to the perceived or established purposes of the data. Regarding to objective measurement, Klein et al. [9] defined five dimensions for assessing the quality of sensor data streams, including accuracy, confidence, completeness, data volume, and timeliness. Additional DQ dimensions for IoT cover ease of access, access security, representativeness, and interpretability of data.

B. Opportunities for business improvement based on veracious big data and IoT

Industrial decision making is based on the information gathered from the operational environment. For example in manufacturing industry, different sensors and measurement devices are widely deployed in industry processes. The advance of manufacturing technologies is based on information, but effective decisions do not depend only on reasoning techniques, but also on the quality and quantity of data [10]. To support numerous types of decision making of a manufacturing enterprise, complex systems require real-time data collected from machines, processes, and business environments. The veracity of the data plays an important role when ensuring the correctness of the supporting services. Furthermore, automated data processing and pre-processing become a necessity when the allowed time between data collection and decision making shortens.

Increasing agility is one alternative for manufacturers to address the challenges related to globalization and rapidly changing environments. In order to adapt and respond to changing environments, the industry needs a flexible network of independent units linked by information technology to share the knowledge. Data themselves do not have value, if they are not refined to information or knowledge. During this refinement, the importance of data veracity increases respectively (Fig. 1). Currently, the industry generally utilizes extensively only information for quality variability reduction and process optimization. To achieve higher levels (knowledge or even wisdom), new solutions for intelligent data pre-processing and

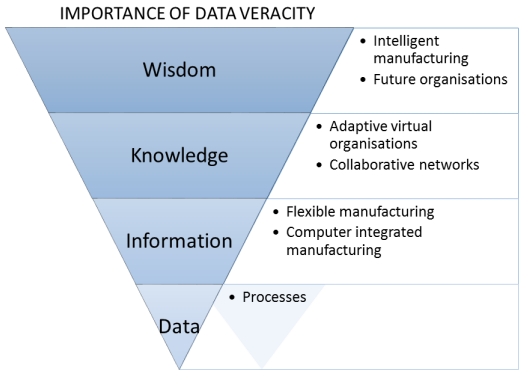

Fig. 1. Importance of the data veracity increases with refining data to knowledge..

analytics are required. Data pre-processing is an essential stage for data analytics and provides correct and useful data sets for applying data mining algorithms, which is an important step for enhancing data veracity. The increasing number of uncertain data sources needs to be taken into consideration in decision-making to ensure data veracity. Therefore, data veracity should be captured and presented to the user. Thus, there is a demand for new methods for veracity measurement and enhancement in data processing.

C. Challenges for processing veracious big IoT data

We summarize the key characteristics of IoT and the challenges for processing veracious IoT data in seven dimensions.

Big IoT Data stream processing: The challenge is to access data from a big amount of IoT devices, to prepare data and then to perform rich analytics. The methods and algorithms for processing big data need to have high performance for real-time analytics. The rate of generating data is high and often the data exhibit smooth variations, i.e. a small variation occurs between two consecutive time stamps. Data produced from monitoring various physical phenomena are continuous. Sampling is often used for achieving energy efficiency. Moreover, the sensor data related to various phenomena may present an inherent periodic pattern where the same values occur at specific intervals.

Data Processing Latency: Timely generation of information before it becomes outdated is critical for some IoT systems, for instance localization and navigation. IoT data are generated in real-time and allowed data processing time is short to produce useful information and to make the right decisions. Hence, the main challenge here is the development of efficient data pre-processing and analytics methods for different level of refinement, from simple to rich analytics based on requirements of use cases. Especially, lightweight data processing methods with low computation effort are required.

Scalability: Large IoT systems can be expected to have millions of devices producing data. Against this, most current data processing solutions are designed for centralized systems

and do not offer the required flexibility and scalability for large scale deployments. Hence, the relevant challenges are about distributed automatic data processing in IoT architectures. For example, when IoT devices are being automatically connected, the main challenges are to decide where to place the veracious big data processing components and which algorithms to use to process the data. For another, it is critical to avoid the cumulation of low accuracy (such as data error), which may leads large scale analytics unusable.

Completeness: Completeness refers to the exhaustiveness of the descriptions available for the IoT data, i.e. covering all the required aspects mentioned in the hyper-dimension sources, i.e. sources, data as well as the level of details of descriptions. The completeness and interpretability assessment of data should be covered for the evaluation at the input stage but also for the subsequent stages (throughput and output). Essentially, evaluation of this dimension will help to determine if veracity information is available about the data that may be critical at any subsequent stage.

Data Accuracy: The accuracy of information is the degree to which the information correctly describes the phenomena it is designed to measure [11]. It is usually characterized in terms of error in statistical estimates and is traditionally decomposed into bias (systematic error) and variance (random error) components. It may also be described in terms of the major sources of errors that potentially cause inaccuracy (e.g., coverage, sampling, non-response and response). A detailed survey on error sources is desirable when analyzing the accuracy of a potential dataset in regard to statistical analytics.

Complexity of IoT data models: Complexity of the data models and data formats means that IoT devices generate data with various structures, from simple and lightweight structures to complex and verbose structures. Data correlations reflect the extent of hierarchies, nested structures, and other possible correlations in the data. This makes non-uniformity and inconsistency big challenges when handling IoT data. This challenge covers the complexity of data structure, data format, data correlation, and data itself.

Privacy, consent, and security: Existing security and privacy solutions cannot provide complete security in big IoT data scenarios. When the number of sensors providing input increases, the possible attack surfaces for the system increase. Current solutions are mostly designed for static data sets, whereas IoT data streams are highly dynamic. Meanwhile, enhancing veracity often requires more information, which might pose a threat to privacy.

In a nutshell, big data have to be processed in real-time in order to obtain valid and useful information and to make the right decisions. These seven dimensions set big challenges to judge the data quality and handle veracity of data within reasonable amount of time as data volume is significant. The diversity of the data sources brings abundant data types and complex data structures, which increase the difficulty of data integration.

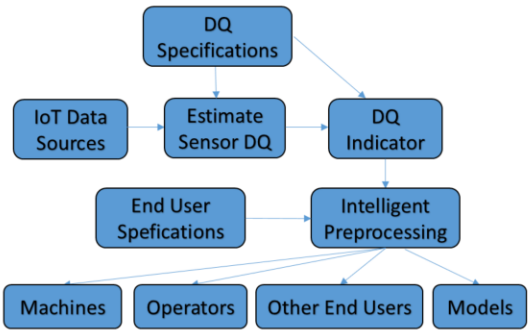

Fig. 2. A general process to assess the data veracity.

III. Assessment and Enhancement of Data Veracity

In this section, we survey the state-of-the-art techniques for assessing and enhancing big data veracity and algorithms for mining uncertain data streams. Figure 2 presents a general process for the assessment of data veracity. IoT devices generate a large volume of data with varying structures and high velocity. The sensor data quality needs to be guaranteed for the further data processing, and it requires to be measured according to the DQ specifications. Evaluation results can be described by DQ indicators, which are further utilized in data processing. Intelligent pre-processing enables the automated enhancement of the DQ, thus, the data can be distributed to different users according to their specifications for further data mining.

A. Methodologies for data assessment

Many factors need to be taken into consideration for building the DQ models, such as phases and steps, dimensions and metrics, data types, cost types, information systems, processes, and services [8]. Quality categories, criteria, indicators, and measures mainly characterize the DQ model. Each category can be associated with a particular property of data and each criterion can be associated to one or more indicators accordingly. A given indicator may correspond to a measure or a set of measures related to several quality criteria.

Assessment of data veracity denotes the quality of data collections on the relevant DQ dimensions, compared to the reference values and enables a diagnosis of quality. The assessment process commonly includes data analytics, DQ requirement analysis, identification of critical areas, and measurement of quality. DQ information can be stored in metadata, which provides complementary information on data.

Rodríguez et al. [12] propose an approach for monitoring applications by providing users with important DQ information. They focus on qualifying sensor data, instead of correcting or improving it. They present DQ assessment task in three steps, including specification of the information quality sources, estimation of the sensor DQ, and management of the DQ information. DQ information indicates the quality properties of data that are evaluated, such as criteria, measure,

and indicator. A set of criteria is selected to estimate the quality of raw sensor data at the acquisition layer, which allows users to estimate the quality of data sources, the context of acquisition and the transmission to the data management and processing center. The internal category is related to quality criteria such as consistency, currency, and volatility. The goal is to avoid inconsistent information and to maintain the temporality of sensor data at a processing level. The usage category is related to DQ criteria such as timeliness, availability and adequacy. The assessment of sensor DQ implies a strong correlation between sensor data and the information about dynamic changes of quality values. In this approach, quality information is considered as the complementary information describing the uncertainty of sensor data and helpful to understand sensor data. Metadata are data related to sensor behavior, the specificities of monitoring context and to DQ information. Batini et al. [8] summarize techniques to assess and improve the quality of data, compare thirteen DQ methodologies from several perspectives.

B. Data processing for enhancing data veracity

The sensor DQ can be enhanced by improving sensor infrastructures and behaviors, qualifying and maintaining sensor resources, using data pre-processing and uncertainty reasoning techniques. Commonly, data-driven and process-driven strategies can be used for improving DQ. Acquisition of new data, record linkage, error localization and correction, etc, are techniques used in data-driven strategy. Process control and process redesign are primary techniques used in process-driven strategy to improve DQ. Generally, process-driven techniques outperform data driven techniques from a long-term view [8].

Data preparation and data reduction are two main techniques used in data pre-processing. Data preparation includes data integration, cleaning, normalization and transformation. Data reduction is used to reduce the complexity of the data by feature selection, feature extraction, and instance selection [13]. There are several available techniques for data pre-processing. Pyle presents a proven approach for preparing the data [14]. García et al. [15] summarize the most influential data pre-processing algorithms covering missing values imputation, noise filtering, dimensionality reduction, instance reduction, discretization, and treatment of data for imbalanced pre-processing. They also discuss the characteristics and performance of the algorithms.

For missing value imputation, it is critical to differentiate between missing and empty values. In most cases, there is high information in noting the patterns of variables that are missing. The probability function of the data can be formulated by taking into account the mechanisms that induce missingness. Approximate probabilistic models can be sampled to fill the missing values by using maximum likelihood procedures [16]. García et al. [13] investigate the current development of big data pre-processing techniques, and they find that current development mainly focuses on feature selection and treatment of imbalanced data, while little work has been proposed for dealing the missing data in big data systems. A parallel data cleaning algorithm is designed by Chen et al. [17] for system data with missing information. Zhang et al. [18] use the rough set theory and introduce three different parallel matrix-based methods for processing large-scale incomplete data. A Streaming K-Nearest-Neighbors Imputation Framework (SKIF) is proposed to handle drifting large volume data streams [19]. It summarizes historical statistical information of complete records in some micro-resources and maintains these in a candidate pool as benchmark data. SKIF uses a novel hyprid-K-Nearest Neighbors imputation method to estimate the up-to-date incomplete records. In [20], a method is proposed for adaptive cleaning RFID data. It models the unreliability of RFID readings by viewing RFID streams as a statistical sample of tags in the physical world, and exploits techniques grounded in sampling theory to drive its cleaning processes. Through the use of tools such as binomial sampling and Π-estimators, the Statistical sMoothing for Unreliable RFid data (SMURF) filter continuously adapts the smoothing window size in a principled manner to provide accurate RFID data to applications.

Noise sometimes is present in the input attributes, which might affects the output attribute. We can leave the noise in, correct it or filter it out. Data polishing methods enable labeling of an instance and repair the values to appropriate ones. Noisy instances in the training data can be identified and removed the by noise filters without modifying the data mining techniques [21].

C. Mining uncertain data streams

Three main research areas concerning uncertain data are modeling, management, and mining of uncertain data. The major methods for handling the uncertainty include probability analysis, fuzzy analysis, bayesian analysis, soft computing technique (fuzzy logic, neural networks, and probabilistic reasoning) and rule based classification technique [22]. The general model for uncertain data is the possible world model. Graphical models are more sophisticated and can be used to model complex dependency. We need to choose the best model depending on the applications while considering the tradeoff between usability and expressiveness. Aggarwal et al. [23] [24] explore various uncertain data algorithms for data mining and management applications.

Research on mining data streams mainly focuses on classification, clustering and association rules extraction. It is a challenge for mining data streams as data streams suffer from many problems, such as bounded memory, single-pass, real-time response and so on. Single classifier based approach and ensemble based approach are two main approaches for data stream classification. Two well known algorithms for data stream classification are Very Fast Decision Tree learner (VFDT) and Concept-adapting Very Fast Decision Tree learner (CVFDT) [25]. CVFDT is an extension of VFDT for dealing with concept drift. Many methods are available for clustering data streams, such as STREAM, CluStream, E-stream, and ClusTree. Compared to traditional clustering techniques, clustering data stream methods are adapted using incremental learning or two-phase learning combined with different windows (Landmark, Sliding, Fading and Titled-time) [26].

In some use cases, sensor networks collect a large amount of uncertain data with high speed which requires to be processed in real-time. The methods for processing data streams need to be re-designed in order to take the uncertainty into account. Diao et al. [27] present a space and time efficient probabilistic modeling and inference based method for high-volume stream processing. First, they utilize graphical modeling to capture how a sensor produces data from the true phenomenon with various types of noise. Then they employ probabilistic inference to transform observed data into data of interest based on the data generation model. Advanced approximation techniques are explored in coping with high volume streams. They also present a Probabilistic Data Stream System (PODS) that supports relational query processing under uncertain data streams using continuous random variable [28]. PODS utilizes a unique data model based on Gaussian Mixture distribution, which is flexible and efficient. The widely used techniques for continuous random variables are multivariate integral and Monte Carlo simulation.

Aggarwal and Yu [29] propose a general method for clustering data streams where they assume only a few statistical measures of the uncertainty (such as the standard error) are available. They develop the UMicro algorithm with use of micro-clustering model, which was first proposed for large data sets, and subsequently adapted for deterministic data streams. Compared to a purely deterministic method, their approach is more effective and can greatly improve the quality of the underlying result even using modest information during the mining process.

Pan et al. [30] present two ensemble based algorithms, Static Classifier Ensemble and Dynamic Classifier Ensemble for data stream classification. Only class value of the sample is assumed to be uncertain, while attributes value is precise. Liang et al. [25] propose a Uncertainty-handling and Concept-adapting Very Fast Decision Tree algorithm (UCVFDT) for classifying high speed uncertain data streams. UCVFDT is based on Decision Tree Classification on Uncertain Data (DTU) and CVFDT. DTU is an extension of a well known classification algorithm C4.5 and is a well performance decision tree on static uncertain data sets. CVFDT is a data stream decision tree algorithm without considering uncertainty.

IV. Future research directions in veracious IoT data analytics

We propose following five important future research directions for veracious data analytics in IoT:

Direction 1. Data cleaning and veracity management technologies for heterogeneous and distributed IoT data: The data in IoT systems are often collected from distributed and heterogenous sources. There is a need for automated data cleaning and processing techniques for veracious IoT data that are able to take into account heterogeneity of data sources. Moreover, data cleaning and veracity management solutions are needed for managing veracity of data that are integrated from various heterogeneous sources. Finally, the proposed techniques should be able to handle different variables of interests to fulfill IoT applications' requirements which will likely provide complex services based on multiple parameters.

Direction 2. Approaches to support autonomous decision-making with veracity metadata: A key concept of IoT is that devices are able to execute complex tasks and make decisions without human intervention. In order to develop dependable and robust IoT systems, it is vital to design common approaches for integrating the veracity information into the decision-making process. There is a need for approaches to represent the veracity at different levels of abstraction (i.e. data, information, knowledge, and wisdom) and encode this metadata in a machine interpretable semantic format. Moreover, there is a need for common IoT agent design patterns to manage data veracity within the system control logic in a dependable and transparent way.

Direction 3. Context-aware and domain-optimized methodologies for enhancing data veracity: IoT systems rely on domain-knowledge provided by domain experts and often represented as facts and rules within a knowledge base. IoT systems need context-awareness to produce actions that match the situations. Because IoT systems already store a large amount of domain-specific knowledge and context data, it is natural to study how this knowledge could be used to enhance data veracity at different levels of abstraction. The context and domain knowledge can be also used to optimize data cleaning and pre-processing so that, for example, resources are not wasted in situations where high data veracity is not required.

Direction 4. Lightweight data cleaning and processing techniques for IoT edge devices: Implementing DQ control in "things" level would make the data processing and cleaning architecture capable of scaling and evolving with the same pace as IoT itself. IoT devices often have constrained resources. Deploying data cleaning and processing techniques on IoT devices require novel design of lightweight solutions that could be embedded in smart devices.

Direction 5. Privacy preserving, personalized, and secure veracious data management: As data veracity is often subjective, data consumers should be able to define their tradeoff of privacy and veracity based on their own requirements. DQ could be intentionally reduced to preserve privacy. Data consumers need to be able to manage their data veracity efficiently based on their own specifications and requirements. Moreover, there is a lack of solutions for security. We suggest to design and implement 1) an open ecosystem with standard APIs to avoid interoperability and reliability problems, and 2) the best security practices for IoT devices to protect against common security and privacy threats.

V. Conclusion

IoT presents challenges related to improving veracity in processing big data. Most current research of data veracity has been focused on and limited to DQ and data uncertainty so far, such as precision, plausibility, and timeliness. In this paper, we identify the challenges of handling veracity of big data from IoT viewpoint, perform a survey about

assessing and enhancing veracity by doing data pre-processing and mining uncertain data streams. Furthermore, we identify five research directions for future development of veracious big data analytics, including data cleaning and veracity management technologies for heterogeneous and distributed IoT data, approaches to support autonomous decision-making with veracity metadata, context-aware and domain-optimized methodologies for enhancing data veracity, lightweight data cleaning and processing techniques for IoT edge devices, and privacy preserving, personalized, and secure veracious data management.

Some standardization efforts have been done also in this area. For example, ISO 8000-8:2015 describes fundamental concepts of information and DQ, and how these concepts apply to quality management processes and quality management systems. Moreover, it specifies prerequisites for measuring information and DQ when executed within quality management processes and quality management systems.

In the future, our research will focus on 1) lightweight and real-time data stream pre-processing which can implement DQ control in "things" level; and 2) minimizing the latency in IoT stream pre-processing with developing automatic approaches for adding metadata as semantic annotations. With these approaches, rich information can be automatically added as metadata to veracious big data to enhance performance and scalability of IoT systems. Finally, we are interested in development of privacy-preserving big data processing frameworks based on General Data Protection Regulation [31].

ACKNOWLEDGEMENT

This research has been partly funded by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 646428). Dr. Xiang Su would like to thank Jorma Ollila Grant of Nokia foundation for funding his research.

REFERENCES

[1] B. Saha and D. Srivastava, "Data Quality: The other face of Big Data," in *Proc. of the 30th Int. Conf. of Data Engineering*. Chicago, IL, 2014, pp. 1294–1297.

[2] E. Rahm, and H. H. Do, "Data cleaning: Problems and current approaches," IEEE Data Eng. Bull., vol. 23, no. 4, pp. 3–13, 2000.

[3] D.M. Strong, Diane M., Yang W. Lee, and Richard Y. Wang, 'Data quality in context," Commun. ACM, vol.40, no.5, pp. 103–110, 1997.

[4] J. Kepner, V. Gadepally, P. Michaleas, N. Schear, M. Varia, A. Yerukhimovich, and R. Cunningham, "Computing on masked data: A high performance method for improving big data veracity", in *Proc. IEEE High Performance Extreme Computing Conference*. 2014, pp. 1–6.

[5] N. B. C. E. Jamil, I. B. Ishak, F. Sidi, L. S. Affendey, and A. Mamat, "A Systematic Review on the Profiling of Digital News Portal for Big Data Veracity," Procedia Computer Science, vol. 72, pp. 390–397, 2015.

[6] M. Krotofil, J. Larsen, and D. Gollmann, "The Process Matters: Ensuring Data Veracity in Cyber-Physical Systems," *in Proc. of the 10th ACM Symposium on Information Computer and Communications Security - ASIA CCS*. Singapore, 2015, pp. 133–144.

[7] V. Kale, Big Data Computing: *A Guide for Business and Technology Managers*, CRC Press, 2017.

[8] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," ACM Computing Surveys, vol. 41, no. 3, pp.1–52. 2009.

[9] A. Klein and W. Lehner, "Representing Data Quality in Sensor Data Streaming Environments," ACM Journal of Data and Information Quality, Vol. 1, No. 2, Article 10, 2009.

[10] I. Dumitrache and S.I. Caramihai, "The intelligent manufacturing paradigm in knowledge society," Knowledge Management, InTech, pp. 36–56, 2010.

[11] Statistics Canada, "Statistics Canada?s Quality Assurance Framework," Statistics Canada Catalogue no. 12-586-XIE, 2002.

[12] C.C.G. Rodríguez and S. Servigne, "Managing Sensor Data Uncertainty: a data quality approach," International Journal of Agricultural and Environmental Information Systems, vol.4, no.1, pp. 35–54, 2013.

[13] S. García, S. R. Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," Big Data Analytics, pp. 1–22, 2016. Big Data Computing: A Guide for Business and Technology Managers

[14] D. Pyle, *Data Preparation for Data Mining*, Morgan Kaufmann Publishers, Inc. 1999.

[15] S. García, J. Luengo and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," Knowledge-Based Systems 98, pp. 1–29, 2016.

[16] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 1st ed. New York: Wiley Series in Probability and Statistics-Wiley, 1987.

[17] F. Chen and L. Jiang, "A parallel algorithm for data cleansing in incomplete information systems using mapreduce," In *Proc. of 10th International Conference on Computational Intelligence and Security*. Kunmina, China, 2014, pp. 273–277.

[18] J. Zhang, j. S. Wong, Y. Pan, and T. Li , "A parallel matrix-based method for computing approximations in incomplete information systems," IEEE Trans Knowl. Data Eng., vol.27, no.2, pp. 326–339, 2015.

[19] P. Zhang, X. Zhu X, J. Tan, and L. Guo, "SKIF: a data imputation framework for concept drifting data streams," In *Proc. of 19th ACM international conference on Information and knowledge management*. Toronto, Canada, 2010, pp. 1869–1872.

[20] S. R. Jeffery, M. Garofalakis, and M. J. Franklin, "Adaptive Cleaning for RFID Data Streams," In *Proc. of of the 32nd international conference on Very large data bases* . Seoul, Korea, 2006, pp. 163–174.

[21] B. Kanagal and A. Deshpande, "Online filtering, smoothing and probabilistic modeling of streaming data," In *IEEE 24th International Conference on Data Engineering*. Mexico, 2008, pp. 1160–1169.

[22] L. S. Dutt and M. Kurian, "Handling of Uncertainty-A Survey," International Journal of Scientific and Research Publications, Vol. 3, no.1, ISSN 2250–3153, 2013.

[23] C. C. Aggarwal, and P. S. Yu, "A Survey of Uncertain Data Algorithms and Applications," IEEE Transactions on knowledge and data engineering, vol.21, No.5, May 2009.

[24] C. C. Aggarwal, *Managing and mining uncertain data*, Kluwer Academic Publishers, 2009.

[25] C. Liang, Y. Zhang, and Q. Song, "Decision Tree for Dynamic and Uncertain Data Streams," in *Proc. of 2nd Asian Conference on Machine Learning (ACML2010)*. Tokyo, Japan, 2010, pp. 209–224.

[26] H. L. Nguyen, Y. K. Woon, W. KeongNg, "A survey on data stream clustering and classification," Knowl. Inf. Syst. vol. 45, pp.535-569, 2015.

[27] Y. Diao, B. Li, A. Liu, L. Peng, C. Sutton, T. Tran, and M. Zink, "Capturing data uncertainty in high-volume stream processing," In *Proc. of fourth biennial Conference on Innovative Data Systems Research*. 2009.

[28] Y. Diao, B. Li, L. Peng, C. Sutton, T. Tran, and M. Zink, "PODS: a new model and processing algorithms for uncertain data streams," in *Proc. of ACM SIGMOD International Conference on Management of data*. Indianapolis, USA, 2010, pp. 159–170.

[29] C. C. Aggarwal and P. S. Yu, "A Framework for Clustering Uncertain Data Streams," in *Proc. of IEEE 24th International Conference on Data Engineering*. Cancun, Mexico, 2008, pp.150–159.

[30] S. Pan, K. Wu, y. Zhang, and X. Li, "Classifier Ensemble for Uncertain Data Stream Classification," In: Advances in Knowledge Discovery and Data Mining, 2010.

[31] X. Su, J. Hyysalo, M. Rautiainen, J. Riekki, J. Sauvola, A.I. Maarala, H. Hirvonsalo, P Li and H. Honko, "Privacy as a Service: Protecting the Individual in Healthcare Data Processing," Computer, vol. 49, no. 11, pp. 49–59, 2016.