Big Data and Information Analytics
©American Institute of Mathematical Sciences
Volume **2**, Number **2**, April **2017**
doi:10.3934/bdia.2017002
pp. **127–139**

BIG DATA COLLECTION AND ANALYSIS FOR MANUFACTURING ORGANISATIONS

Pankaj Sharma

Department of Mechanical Engineering
Hauz Khas, Indian Institute of Technology Delhi
New Delhi, 110016, India

David Baglee

Faculty of Applied Science
Department of Computing, Engineering and Technology
Industry Centre, Hylton Riverside, Sunderland, UK

Jaime Campos

Department of Informatics
Linnaeus University
SE-351 95 Växjö, Sweden

Erkki Jantunen

VTT Technical Research Centre of Finland
P.O.Box 1000, FI-02044 VTT, Finland

(Communicated by Jianhong Wu)

Abstract. Data mining applications are becoming increasingly important for the wide range of manufacturing and maintenance processes. During daily operations, large amounts of data are generated. This large volume and variety of data, arriving at a greater velocity has its own advantages and disadvantages. On the negative side, the abundance of data often impedes the ability to extract useful knowledge. In addition, the large amounts of data stored in often unconnected databases make it impractical to manually analyse for valuable decision-making information. However, an advent of new generation big data analytical tools has started to provide large scale benefits for the organizations. The paper examines the possible data inputs from machines, people and organizations that can be analysed for maintenance. Further, the role of big data within maintenance is explained and how, if not managed correctly, big data can create problems rather than provide solutions. The paper highlights the need to have advanced mining techniques to enable conversion of data into information in an acceptable time frame and to have modern analytical tools to extract value from the big datasets.

1. **Introduction.** Data is the new oil for the future. Organizations that are able to exploit data analytics will have a clear edge over the competition. The process of data acquisition has become simpler and economically affordable with the technological advances in hardware systems and software. A large amount of data is generated and collected during asset maintenance. Data is an important tool in

2010 *Mathematics Subject Classification.* 00B10.
Key words and phrases. Big data, CBM, manufacturing.

the hands of the asset maintenance personnel. However, the utilization of this data is often not up to the desired levels. There is evidence that most organisations have far more data than they possibly use; yet, at the same time, they do not have the data they really need [1]. There are associated problems of bad data quality which leads to undesired data discarded by the organizations. An increasing number of embedded systems within machines and assets have led to an explosion in the generation of data. In spite of this, it appears that, at the management level, executives are not confident that they have enough correct, reliable, consistent and timely data upon which to make decisions [2]. Companies in every industry have struggled-and sometimes failed outright- as a result of poor data access and management, an inability to translate data into valuable information, and poor data quality [3].

There are several serious difficulties often encountered in completing the full circle of condition-based maintenance (CBM). As there is generally a huge amount of data generated, there might be a need to gather the data from the assets dispersed over a large geographical area, *i.e.* data acquisition problems. There may be a need to integrate this data to provide any useful information, with time the need may be felt for data acquisition from additional sources, its integration with the rest for more meaningful interpretation and finally; the availability of an expert for converting data into useful information for maintenance. In addition, good experts are rare, and therefore, even if a condition-monitoring programme is in operation, failures still occur frequently, defeating the very purpose for which investment in CBM is made [4, 5].

The decision making in asset maintenance is based on the analysis of parameters monitored by the sensors placed on the machines. The decision is essentially a predetermined action taken by the maintenance personnel when the monitored parameter breaches a threshold. However, in certain cases, the decision can also be taken dynamically with the analysis of the streaming data.

To solve these problems/issues in maintenance, different kinds of Information Communication technologies (ICTs) have grown in popularity. Artificial intelligence (AI) for decision making for maintenance started to appear in the 1980's in form of expert systems. In the 1990's, complex techniques including artificial neural networks and fuzzy logic were used to support maintenance decisions. Reviews on artificial intelligence techniques [6, 7] and their application to condition monitoring [8, 9] are available. Distributed artificial intelligence (DAI) has also been used in condition monitoring with the advent of Internet during the late 1990s [10, 11, 12, 13]. AI led to DAI and subsequently to agent technology. Recently, researchers have begun to apply web and agent technologies in maintenance and condition monitoring. A review on the subject was published in 2006 [14] and an extended and updated version was published in 2008 [15]. These technologies gained wider acceptance because of the agents' capability to operate on distributed open environment using the Internet or corporate Intranet and access heterogeneous and geographically distributed databases and information sources [16, 17]. The Internet, which became popular in civilian use in the beginning of the 1990's, utilises standard communication protocols that provide the transfer of data throughout the world [18]. The Intranet utilises web technology for creating and sharing knowledge in an enterprise [19].

Currently with increased capacity of the databases the companies are able to store large amounts of data. To be able to gain insight into the produced data is

a challenge as well as a competitive advantage for companies which are able to use it optimally for their benefits [20]. In maintenance, the use of big data is in the introductory phases where, for instance, frameworks are suggested and tested in different sectors of the domain of interest [21, 22, 23, 24]. However, there are huge potentials and several challenges with the implementation of big data systems because of data complexity, which are highlighted in this work.

The demand for big data from a number of sources is creating a range of problems for the organizations. The problems of data acquisition have multiplied exponentially in this era of big data. The deluge of data has increased rapidly that in a recent announcement from Google, MapReduce is abandoned as it is unable to handle the amounts of data Google wants to analyze these days [27]. The only thing that we are sure of today is that this trend is irreversible. A relevant question here is to ask; what are the volumes, velocity or variety of data beyond which it becomes big? The answer to it is that there are no benchmarks that define big data in terms of a quantified V. These limits vary depending on the size, sector and location of the organization. Furthermore, these limits are growing with time. These Vs are not independent of each other. As one of the dimension changes, the likelihood increases that another dimension will also change as a result. However, a 'three-V tipping point' exists for every firm beyond which traditional data management and analysis technologies become inadequate for deriving timely intelligence. The Three-V tipping point is the threshold beyond which firms start dealing with big data [40]. At this tipping point, there arises a need to conduct a cost-benefit trade-off by the organizations to decide about the implementation of a big data analytics programme.

However, this does not mean that Big Data is always better. It depends if the data is noisy i.e. corrupt or meaningless data, or not, and if it is representative of what we are looking for [28]. Claims to accuracy about Big Data are misleading. When the numbers of variables grow, the numbers of fake correlations also grow [29]. Maintenance analysts must guard against all chances of incorrect deductions through wrong analysis of poor quality data. A detailed discussion on the challenges of using big data in asset management will be discussed later in this paper.

2. **Big data in a CBM centric maintenance startegy.** Condition Based Maintenance (CBM) can be considered the stronghold of modern maintenance i.e. it is a strategy that can lead to economically optimal results. As the name suggests CBM is based on the idea of carrying out maintenance when a condition outside of the norm, is detected. The main drawback in this is that the maintenance actions that are not needed would lower the availability of the machinery i.e. the machinery cannot be used for production when maintenance is taking place. Unnecessary maintenance actions would create the need for additional maintenance since a reason for failure is previous maintenance actions that have not been carried out appropriately. It is interesting that the maintenance related standards typically present corrective maintenance as the opposite to condition based maintenance although logically thinking corrective maintenance is condition based maintenance i.e., if something is broken it has to be fixed. In other words, the famous principle "don't fix it unless it is broken" fits here to describe the ideology behind corrective maintenance. Naturally the main problem with corrective maintenance is that when following that strategy, all failures come as surprises and even very small and easily predictable faults can stop the whole production process. Consequently, the aim of CBM is to be proactive i.e. to be able to predict when the components of

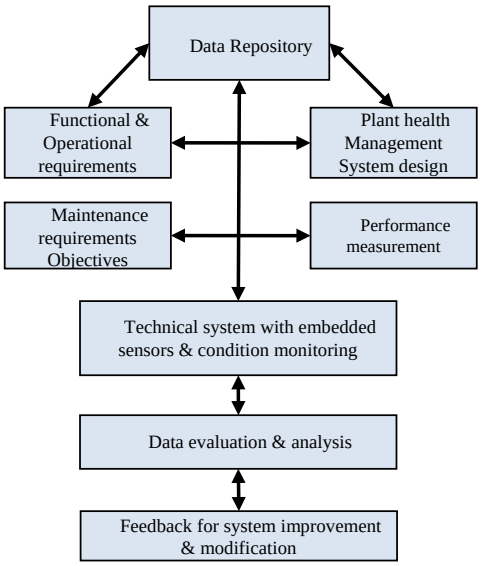

FIGURE 1. A Conceptual architecture of complex data analysis for maintenance decisions (Adapted from Health management of Complex technical systems, [26])

production machinery would fail and carry out maintenance actions beforehand in order to avoid unplanned stoppages of production.

A stakeholder's requirements based health management system (HMS) framework is given at Figure 1. With increasing use of condition monitoring, data collection, and internet in management of maintenance process, the information logistic is required to be streamlined. Condition monitoring uses various intelligent health monitoring techniques to monitor and control the health status of plant and machineries by analysing the data after it has been collected. The identification of effective and efficient strategies for the maintenance of a plant and machineries is of a major importance from global competition, safety and financial point of view. Today, most of the organizations are trying to follow the condition based preventive maintenance, shown in Figure 1 based on the state of the component degradation. However, in reality, the relevant parameters behind the degradation process are very complex, and needs to be undertaken analytically. In addition, it is important to link and integrate maintenance management data systems and to quantify the importance of each data field [25].

Big Data is an assimilation of data from three sources; Machines/sensors, People and Organizations. In the domain of maintenance management too, the data gets generated at either or all of these three sources. Traditional asset management plans use limited data collected from the sensors to carry out analysis and decision making for maintenance interventions. However, in a big data approach, people and organization data will also contribute towards garnering intelligent information.

The **machine or sensor data** will be generated when measurements are carried out while following a CBM strategy. With these measurements the idea is to follow the development of wear of the components of the machinery. The three techniques to follow the development of wear are watching time, monitoring load and measuring

FIGURE 2. Old traditional gearbox opened for checking the wear of gear teeth

wear. Of these three techniques the time based method is not a reliable method although it is very widely used. If the wear of the components of machinery would only depend on time it would be possible to follow the wear by having similar components that are inside the machine on a table beside the machine and when they become worn one could know that now it is the time to take action. These kinds of wear monitoring tables are not commonly used in the industry although time based maintenance is commonly used.

Monitoring load is another way to define how worn the components of the machinery are since in general load is physical reason for wear. In case the load is stable following the running hours i.e. following time can be used to define the need for maintenance. In case the load is unstable i.e. varying, the issue becomes much more complicated. The main reason for this is that wear is not typically a linear function of load i.e. doubling the load would not double the wear instead it could make it ten times higher. Consequently, there is need to record how much time the machine has been used at each load level and from there with a model to calculate the sum of wear.

In the real industrial world the most reliable way of monitoring wear is to measure wear as such (Figure 2). Unfortunately, this approach cannot often be followed since the components of interest are hidden and thus cannot be monitored with low cost solutions. In fact the wear monitoring solutions could become extremely expensive and also be prone to fail easily. Due to the fact that it is not easy to monitor wear as such indirect methods to define the amount of wear are often used. The most common method is the measurement of vibration acceleration. A good CBM application on the assets will result in a large amount of sensor data. This data can take various forms including velocity, acceleration, acoustic emission, etc. In addition, utilising the Internet of Things (IOT) techniques will allow for a greater variety and volume of machine data.

Although, most of the CBM decision making will be based on the analysis of the sensor data, people and organization data too becomes significant when a Big Data model is used. People data is a direct result of increased number of smart devices in circulation today. Smartphones, Tablets and Computers are able to transmit large amounts of people data. These are in the form of tweets, Facebook posts, complaints, reviews etc. by both the customers and the employees. To complicate this approach this data is unstructured i.e. it is not in traditional relational tables.

These smart devices are allowing people to generate data in large volumes, of diverse variety and at great velocity. Organizations too are adding to the data. Close-Circuit cameras and other security means of the organizations generate data that can be used for intelligent decision making. Analysis of the video files can reveal information about the machines that require maximum attention of the operator and maintenance crew, even in the absence of any documentation. Employee information held with the organizations with their geographical locations can be of use while planning maintenance activities. The skill sets of the employees, stored in the organizations can help in deputing right man for the maintenance job. All this data comes in different forms. Text, videos, Java Script Object Notation (JSON) data, relational tables, Comma separated Values (CSV), images, graphs etc. are some of the types of data that will be collected. Suitable data analysis packages from the Big Data Ecosystem should be used to analyse these different types of data.

3. **Problems of data in asset management (Pre big data era).** The complexity of data in asset management has been a problem area for persons managing the assets that may be located together or distributed geographically. The data issues start often with an inappropriate or unnecessary selection of assets and measurement parameters. There is a tendency that industrial companies measure what is easy to measure, not what is required [30]. This results in the collection of useless data that has no analytical value. This bad data needs to be understood in the business concept and how it leads to wasted expense, unproductive labour, service interruptions and impact on customers [3]. The major problem is to convert this data into information that can be used for meaningful management decisions.

There is a compatibility problem that organizations face because of separate computer software packages for company administration and asset management systems. These packages are not able to talk to each other i.e. easily share and analyse data. This has led to the notion of islands of information. Such disconnects make it extremely difficult to bring real-time information from the plant into business systems [2].

Human influence in data acquisition is a double edged sword that cuts both ways. Humans have larger tolerance to ambiguity and can assimilate heterogeneity. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and are poor at understanding nuances [31]. On one hand, it can lead to problems where the operators that lack training analyse irrelevant and incorrect data that leads to incorrect conclusions. On the other hand, most of literature finds human intervention better than simply depending on the sensors [41]. Human factors involved in the collection of data are more reliable, as these data are more closely related to indicators of ownership and responsibility. Technicians and operators will collect data only if they believe it is worthwhile, and the results are made available for consultation and use [32]. Recently, there has been a surge in the different types of sensors that are able to read and feed data from multiple systems; however, they can entail additional expense for the organization. Therefore, in many practices, data collection as well as data recording and updating in information systems is still largely manual or only semi-automated. A technological solution in itself may not improve data integrity as systematic institutional changes are often required [33].

The reasons of data complexity arising due to bad quality of data are summarized by Koronios et al. [2] as: inadequate management structures for ensuring complete,

timely and accurate reporting of data; inadequate rules, training, and procedural guidelines for those involved in data collection; fragmentation and inconsistencies among the services associated with data collection; and the requirement for new management methods which utilize accurate and relevant data to support the dynamic management environment [2].

4. **Bigger problems due to big data?** The magnitude of the size and complexity of Big Data can be understood through a few examples. For every second that the Large Hadron Collider at the Conseil Européen pour la Recherche Nucléaire, or European Council for Nuclear Research, commonly known as CERN run an experiment, the instrument can generate 40 terabytes of data. For every 30 minutes that a Boeing jet engine runs, the system creates 10 terabytes of operations information. For a single journey across the Atlantic Ocean, a four-engine jumbo jet can create 640 terabytes of data. Furthermore, there are more than 25,000 flights that fly each day [34]. Data overload is now becoming a problem. At the organizational level, there are incredibly large amounts of data, including structured and unstructured, enduring and temporal, content data and an increasing amount of structural and discovery metadata [2]. The design of big data systems will continue to evolve when we need to handle larger scale of data and more challenging user demands [27]. Data can be filtered and compressed by orders of magnitude without compromising our ability to reason about the underlying activity of interest. One challenge is to define these "on-line filters in such a way they do not discard useful information since the raw data is often too voluminous to even allow the option of storing it all [31]. Big Data comprises of 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 [35]. High Dimensionality and a large sample size of Big Data raise the following three problems [35];

1. Noise accumulation and spurious correlations due to high dimensionality;
2. Heavy computational cost and algorithmic instability due to high dimensionality in large sample size;
3. Multiple sources of data at different times and collected using different technologies creating issues of heterogeneity, experimental variations and statistical biases, and requires us to develop more adaptive and robust procedures.

Data complexity has increased as a result of emergence of Big Data. As the sources of data are autonomous (No centralized control on the geographically dispersed sensors and agents), it is unwise to transfer this data to centralized location for analysis. On the other hand, analysis at these autonomous independent sites will lead to local biases of each site. Under such a circumstance, a Big Data mining system has to enable an information exchange and fusion mechanism to ensure that all distributed sites (or information sources) can work together to achieve a global optimization goal [36].

Big Data Analysis for decision making often includes querying over massive datasets. This aggregation of large datasets consumes time and is often very expensive because the query processing by default consumes the entire dataset, which is often hundreds or thousands of terabytes [43]. This blocks the computational resources from being used for other more important jobs. This problem can be

obviated through sampling where queries are evaluated against a small randomly-sampled data, returning an approximated result with an error bound. This sampling leads to the problem of *Sparseness of data* [43]. A medium or large population may still be deemed sparse by the theory of statistical sampling, if the value is distributed over a very wide range. Sparse data results in uneven frequency distributions [42]. Such high-dimensional spare data significantly deteriorate the reliability of the models derived from the data [44].

Machine Maintenance data that is collected is often with skewed distribution. This happens when the data in one class is much higher in number than in other classes [46, 47]. This is termed as *class imbalance problem* or *rare event detection*. The minority class usually represents the most important concept to be learned, and it is difficult to identify it since it might be associated with exceptional and significant cases [48]. In addition to other situations like rare disease detection and bank fraud identification, this condition is common in managing risk and predicting failures of technical equipment [45]. Another notable problem that needs to be addressed when using big data for equipment maintenance is that the patterns and relations in such data often evolve over time, thus, models built for analyzing such data quickly become obsolete over time [52]. This is termed as *Concept Drift*. Widmer and Kubat [49] introduced concept drift and defined it as a change in the concept with time. The concept could be the classification boundary or clustering centres [51]. Concept Drift can be further divided into loose concept drift and rigorous concept drift. In the former, concepts in adjacent data chunks are sufficiently close to each other; in the latter, genuine concepts in adjacent data chunks may randomly and rapidly changed [50].

Human perspective to Big Data has also changed when the data suddenly became Big. User-friendly visualization techniques are necessary to narrow the gap between big data system and its users. The visualization techniques should display the analytic results in an intuitive way, so that users can identify the interesting results effectively. To enable a fast response, the back-end system is expected to provide a real-time performance and the visualization algorithms need to transform the users event into a proper and optimized query [27, 31].

Big Data has also impacted the conventional statistical methods and techniques. There is an emerging need to modify these methods to suit the Big Data requirements. There are many reasons that indicate towards the need to develop new statistical techniques. First, the concept of statistical significance that is not completely relevant to a big data scenario. Traditionally, small samples are chosen from the population and a model is formulated. The results of the model are compared with actual happenings and the significance of the relation is established. The conclusion is then generalized to the entire population. In contrast, big data samples are massive and represent the majority of, if not the entire, population. As a result, the notion of statistical significance is not that relevant to big data [40]. Moreover, these traditional techniques that use small sample data often do not scale up to big data. These factors lead to a situation where there has to be a complete rethink on the statistical techniques that are to be used.

5. **Value of big data and importance of advanced data mining.** Big data is revolutionizing a number of fields. It has found large acceptance in medical applications like health parameters monitoring and transmission of data [37, 38], Economics and Finance [35], predicting stock market [39], etc. Big data and its

analysis have great potential in the field of asset maintenance as well. In the previous section, the challenges that arise in big data analysis due to high dimensionality and large sample size were discussed. But high dimensionality of the data also means that it can be used to develop effective methods that can accurately predict the future observations and at the same time to gain insight into the relationship between the features and response for scientific purposes. In addition, large sample size helps in ; firstly, exploring the hidden structures of each subpopulation of the data, which is traditionally not feasible and might even be treated as 'outliers' when the sample size is small; and secondly, extracting important common features across many subpopulations even when there are large individual variations [35]. When data are already present, the limiting factors to gaining these advantages are accessibility of those data for analysis and the availability of analytics techniques themselves [38]. The main focus of any analytics must be to arrive at conclusions about the data in a timeframe in which it is still relevant. In an asset management scenario, if the analysis of machine health parameters to arrive at prognosis of the impending failure is not completed before the machine eventually fails, it is of no use. These limits on the "elapsed time" between data collection and deductions have increased the focus on data mining techniques.

In order to be able to do so, the computational efficiency has to take a giant leap to keep up with the expectations from big data analytics. Big data, because of the shear dimensionality may also lead to statistical inaccuracies. This may result in spurious correlations and "incidental endogeneity". Spurious correlations refer to the fact that many uncorrelated random variables may have high sample correlations. Incidental endogeneity refers to the genuine existence of correlations between variables unintentionally. Both spurious correlations and incidental endogeneity happen because of high dimensionality. Such a paradigm change has led to significant progresses on developments of fast algorithms that are scalable to massive data with high dimensionality. This forges cross-fertilizations among different fields including statistics, optimization and applied mathematics [35]. It can be understood that advances in data mining techniques will assist in covering the gap between data and information. Figure 3 summarizes the problems which are incidental due to the peculiar characteristics of the Big data. The figure also emphasizes the need to have security measures and data mining techniques in order to exploit the full potential of Big Data.

6. **Conclusion.** The current work has shown that different kind of ICTs has been developed which provide organisations with new possibilities to organise different processes. When it comes to ICTs such as the Data Mining and Big Data systems, it is important to have a deep understanding of the domain of interest and its data lifecycle as well as objectives and data requirements to be able to use and perform the proper analytics method. Consequently, experts in the area of big data analytics are becoming crucial to be able to understand various aspects of the data during its life cycle as well as requirements of the data for decision making purposes when big data analytics are performed. In order to adapt an approach to using big data tools and techniques to support an organisation, it is important to take full advantage of recent advances in information technologies related to CBM, software and semantic information to develop an effective information and communication infrastructure.

Acknowledgments. The authors would like to extend their gratitude to FIMECC Ltd (Finnish Metals and Engineering Competence Cluster) for project promotion

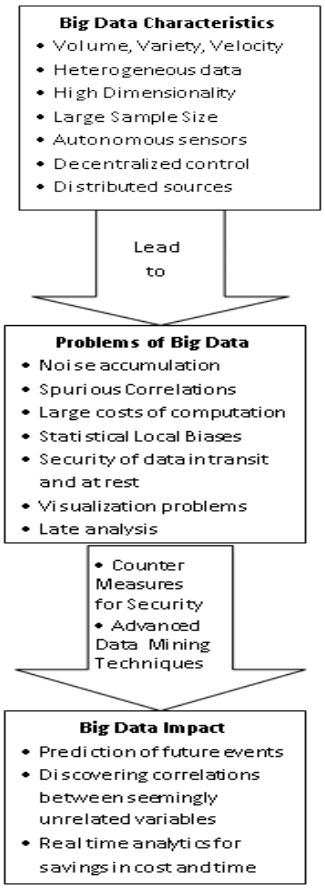

FIGURE 3. Big Data in Asset Management

and Tekes (the Finnish Funding Agency for Technology and Innovation) for its financial support to the research project S4Fleet – Service Solutions for Fleet Management.

REFERENCES

[1] A. V. Levitan and T. C. Redman, Data as a resource: Properties, implications and prescriptions, *Sloan Management Review*, **40** (1998), 89–101.

[2] A. Koronios, S. Lin and J. Gao, A data quality model for asset management in engineering organisations, *Proceedings of the 10th International Conference on Information Quality*, Massachusetts Institute of Technology, Cambridge, USA, 2005.

[3] G. Gilliland, S. K. Barger, V. Bhatia and R. Nicol, Creating value through data integrity: A pragmatic approach, *BCG Perspectives*, (2011), Available at http://www.bcgindia.com/documents/file83320.pdf.

[4] J. S. Rao, M. Zubair and C. Rao, Condition monitoring of power plants through the Internet, *Integrated Manufacturing Systems*, **14** (2003), 508–517.

[5] O. Prakash, Asset management through condition monitoring - How it may go wrong: A case study, *Proceedings of the 1st World Congress on Engineering Asset Management*, (WCEAM) 2006, Gold Coast, Queensland, Australia, July 11–14, 2006.

[6] S. A. Kalogirou, Artificial intelligence for the modeling and control of combustion processes: A review, *Progress in Energy and Combustion Science*, **29** (2003), 515–566.

[7] S. H. Liao, Expert system methodologies and applications - A decade review from 1995 to 2004, *Expert Systems with Applications*, **28** (2005), 93–103.

[8] K. Warwick, A. O. Ekwue and R. Aggarwal, *Artificial Intelligence Techniques in Power Systems*, Institution of Electrical Engineers, Stevenage, UK, 1997.

[9] K. Wang, *Intelligent Condition Monitoring and Diagnosis System a Computational Intelligent Approach*, Frontiers in Artificial Intelligence and Applications, **93**, 2003.

[10] M. Rao, H. Yang and H. Yang, Integrated distributed intelligent system for incident reporting in DMI pulp mill, success and failures of knowledge-based systems in real-world applications, *Proceedings of the First International Conference*, (1996), 169–178.

[11] M. Rao, J. Zhou and H. Yang, Architecture of integrated distributed intelligent multimedia system for on-line real-time process monitoring, *SMC'98 Conference Proceedings, 1998 IEEE International Conference on Systems, Man, and Cybernetics*, **2** (1998), 1411–16.

[12] M. Rao, J. Zhou and H. Yang, Integrated distributed intelligent system architecture for incidents monitoring and diagnosis, *Computers in Industry*, **37** (1998), 143–151.

[13] K. M. Reichard, M. Van Dyke and K. Maynard, Application of sensor fusion and signal classification techniques in a distributed machinery condition monitoring system, *Proceedings of SPIE - The International Society for Optical Engineering*, **4051** (2000), 329–336.

[14] J. Campos and O. Prakash, Information and communication technologies in condition monitoring and maintenance, in Dolgui, A., Morel, G and Pereira, C.E. (Eds.) *Information Control Problems in Manufacturing*, Elsevier, **39** (2006), 3–8.

[15] J. Campos, Survey paper: Development in the application of ICT in condition monitoring and maintenance, *Computers in Industry*, **60** (2009), 1–20.

[16] K. P. Sycara, MultiAgent Systems, *AI Magazine*, **19** (1998).

[17] J. Q. Feng, D. P. Buse, Q. H. Wu and J. Fitch, A multi-agent based intelligent monitoring system for power transformers in distributed substations, *International Conference on Power System Technology Proceedings*, **3** (2002), 1962–1965.

[18] A. C. Weaver, The internet and the world wide web, *23 rd International Conference on Industrial Electronics, Control and Instrumentation*, **4** (1997), 1529–1540.

[19] D. Stenmark, Designing the new intranet, Gothenburg Studies in Informatics, Report 21, March 2002.

[20] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto and R. Buyya, Big data computing and clouds: Trends and future directions, *Journal of Parallel and Distributed Computing*, Special Issue on Scalable Systems for Big Data Management and Analytics, (2015), 79–80, 3–15.

[21] B. Xu and S. Kumar, *Big Data Analytics Framework For System Health Monitoring*, Presented at the 2015 IEEE International Congress on Big Data (BigData Congress), IEEE Computer Society, 2015.

[22] E. Fumeo, L. Oneto and D. Anguita, Condition based maintenance in railway transportation systems based on big data streaming analysis, *Procedia Computer Science*, **53** (2015), 437–446.

[23] A. Mohamed, M. S. Hamdi and S. Tahar, A machine learning approach for big data in oil and gas pipelines, Presented at the *2015 International Conference on Future Internet of Things and Cloud (FiCloud)*, 2015 International Conference on Open and Big Data (OBD), IEEE Computer Society, 2015.

[24] A. Nunez, J. Hendriks, L. Zili, B. De Schutterand and R. Dollevoet, Facilitating maintenance decisions on the dutch railways using big data: The ABA case study, Presented at the *2014 IEEE International Conference on Big Data (Big Data)*, IEEE. 2014.

[25] A. Parida and U. Kumar, Managing information is key to maintenance effectiveness, in *Proceedings of Intelligent Maintenance System*, Arles, France, 15-17 July, 2004.

[26] P. Soderholm, *Continuous Improvement of Complex Technical System: Aspects of Stakeholder Requirements and System Functions*, Licentiate Thesis, Division of Quality and Environmental Management, Lulea University of Technology, Lulea, 2003.

[27] G. Chen, S. Wua and Y. Wang, The evolvement of big data systems: From the perspective of an information security application, *Big Data Research*, **2** (2015), 65–73.

[28] W. Fan and A. Bifet, Mining big data: Current status, and forecast to the future, *SIGKDD Explorations*, **14** (2012), 1–5.

[29] N. Taleb, *Antifragile: How to Live in a World We Don't Understand*, Penguin Books Limited, 2012.

[30] A. Parida, Role of condition monitoring and performance measurements in asset productivity enhancement, 20th International Conference on Condition Monitoring and Diagnostic Engineering Management, Faro, Portugal, 2007.

[31] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan and C. Shahabi, Big data and its technical challenges, *Communications Of The ACM*, **57** (2014), 86–94.

[32] U. Kumar, D. Galar, A. Parida, C. Stenström and L. Berges, Maintenance performance metrics: A state-of-the-art review, *Journal of Quality in Maintenance Engineering*, **19** (2013), 233–277.

[33] W. J. Orlikowski and S. R. Barley, Technology and institutions: What can research on information technology and research on organizations learn from each other?, *MIS Quarterly*, **25** (2001), 145–165.

[34] S. Rogers, Big data is scaling bi and analytics, Available at `http://www.information-management.com/issues/21-5/big-data-is-scaling-bi-and-analytics-10021093-1.html`, 2011.

[35] J. Fan, F. Han and H. Liu, Challenges of big data analysis, *National Science Review*, **1** (2014), 293–314.

[36] X. Wu, X. Zhu, G.-Q. Wu and W. Ding, Data mining with big data, *IEEE Transactions on Knowledge and Data Engineering*, **26** (2014), 97–107.

[37] J. C. Hsieh, A. H. Li and C. C. Yang, Mobile, cloud, and big data computing: Contributions, challenges, and new directions in telecardiology, *International Journal of Environmental Research and Public Health*, **10** (2013), 6131–6153.

[38] ISACA, *Generating Value From Big Data Analytics*, White Paper, Retrieved from (`http://www.isaca.org`), 2014.

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

[40] A. Gundami and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, *International Journal of Information Management*, **35** (2015), 137–144.

[41] P. Oborski, Man-machine interactions in advanced manufacturing systems, *The International Journal of Advanced Manufacturing Technology*, **23** (2004), 227–232.

[42] J. Horák, I. Ivan, T. Inspektor and J. Tesla, Sparse big data problem: A case study of czech graffiti crimes, In: *Ivan I., Singleton A., Horák J., Inspektor T. (eds) The Rise of Big Spatial Data*, Lecture Notes in Geoinformation and Cartography, Springer, 2017.

[43] Y. Yan, L. J. Chen and Z. Zhang, Error bounded sampling for analytics on big sparse data, *Proceedings of the VLDB Endowment*, **7** (2014), 1508–1519.

[44] P. K. Kumar, P. C. Rao, R. Changala, T. J. Rao and P. H. Shankar, Data mining challenges with big data, *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, **3** (2015), 148–150.

[45] R. Longadge, S. S. Dongre and L. Malik, Class imbalance problem in data mining: Review, *International Journal of Computer Science and Network (IJCSN)*, **2** (2013).

[46] H. He and E. A. Garcia, Learning from imbalanced data, *IEEE Transactions on Knowledge and Data Engineering*, **21** (2009), 1263–1284.

[47] Y. Sun, A. K. C. Wong and M. S. Kamel, Classification of imbalanced data: A review, *International Journal of Pattern Recognition and Artificial Intelligence*, **23** (2009), 687–719.

[48] G. M. Weiss, Mining with rarity: A unifying framework, *SIGKDD Explorations*, **6** (2004), 7–19.

[49] G. Widmer and M. Kubat, Learning in the presence of concept drift and hidden contexts, *Machine Learning*, **23** (1996), 69–101.

[50] P. Zhang, X. Zhu and Y. Shi, Categorizing and mining concept drifting data streams, In Proceedings of *the 14th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*, (2008), 812–820.

[51] M. B. Chandak, Role of big data in classification and novel class detection in data streams, *Journal of Big Data*, **3** (2016), 1–9.

[52] I. Zliobaite, M. Pechenizkiy and J. Gama, An overview of concept drift applications, *Big Data Analysis: New Algorithms for a New Society*, **16** (2015), 91–114.

E-mail address: pankajtq@gmail.com
E-mail address: david.baglee@sunderland.ac.uk
E-mail address: jaime.campos@lnu.se
E-mail address: erkki.jantunen@vtt.fi