ISBN: 978-1-948012-16-4

Contemporary Issues in Computing (CIC)

DOI: <http://doi.org/10.26480/cic.01.2020.54.58>

CrossMark

IMPROVING DATA VALIDATION USING MACHINE LEARNING, A NEW WAYS OF SEEING BIG DATA

Tahamina Yesmin^a and Kosalaya Chakrabarti^b^aHaldia Institute of Management, Haldia, 721657, India^bHaldia Institute of Technology, Haldia, 721657, India

This is an open access article distributed under the Creative Commons Attribution License CC BY 4.0, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ARTICLE DETAILS

Article History:

Received 26 October 2020

Accepted 27 November 2020

Available online 03 December 2020

ABSTRACT

The new Data Innovation Strategy gives a chance to us to bring the thought into a progressively solid acknowledgment and to consider how it could turn out to be a piece of the generation procedure. Big data is an awesome stockpile of data and information from the frameworks to opposite end-clients. Big data's are growing in all science and graphically areas very rapidly. While the capability of these gigantic data is without a doubt critical, completely comprehending them requires new ways of thinking and novel learning strategies to address the different difficulties. Although for operating such quality of knowledge requires equipment and it motivate a simulation of data to made and machine learning technique. In this paper, we portray the idea of the data factualness contents with respect to big data and machine learning pertinent works and inspect machine learning strategies, tools, and statistical quality models.

KEYWORDS

Data Quality Management, Measurement Errors, Data Gaps, Machine Learning, Supervised Learning, Data quality, Data Clean, Quality Control.

1. INTRODUCTION

Late advancements in data and correspondence innovation (ICT) have prompted large scale manufacturing of data by informal organizations, sensor systems, and other Internet-based uses of different spaces like human services. The huge measure of data that is created with rapid from different web sources is called Big Data (Mehmood et al., 2016). Huge Data is all around depicted by volume, grouping, and speed. In any case, veracity is another property of Big Data which is creating in notoriety and concerns the rising issue of conviction or quality related with the usage of information. Information quality (DQ) is explained as information 'Comfort to apply' (Panahy et al., 2013), this wide definition passes on the idea that data is utilized for specific destinations, and data of high caliber would be data which is sufficiently satisfactory to permit the clients of data to meet their targets. In the domain of Big Data, investigate on quality is still at an early stages organize, setting off the requirement for more research in this area. Data quality appraisal is a fundamental essential for data improvement, and the motivation behind data quality assessment is to decide the quality degree of data (Philip et al., 2010).

Absence of data quality showed through missing data, copy data, profoundly connected factors, enormous number of factors, and anomalies. Low quality data can present critical issues for building ML models just as big data applications. Statistical strategies, for example, missing data ascription, anomaly recognition, data changes, dimensionality decrease, powerful measurements, cross-approval and bootstrapping assume a basic job in data quality administration.

Over the previous decade, machine learning procedures have been generally received in various gigantic and complex data-escalated fields, for example, drug, stargazing, science, etc, for these strategies give potential answers for mine the data covered up in the data. In any case, as the ideal opportunity for big data is coming, the assortment of data sets is so huge and complex that it is hard to manage utilizing customary learning strategies since the set up procedure of learning from traditional datasets was not intended to and won't function admirably with high volumes of data. For example, most conventional machine learning calculations are intended for data that would be totally stacked into memory (Chen et al., 2014), which doesn't hold anything else with regards to big data. Thus, regardless of the way that gaining from this different information is required to bring basic science and planning progresses close by improvements in nature of our life it brings colossal challenges all the while (Slavakis et al., 2014).

Affiliations as often as possible overestimate information quality and underplay the consequences of low quality information. The results of horrible information may run from significant to sad. Information quality issues can make adventures miss the mark, realize lost wages and diminished customer associations, and customer turnover. Affiliations are routinely fined for not having a reasonable regulatory consistence process. Top notch data is at the core of administrative consistence. The Data Warehousing Institute (TDWI) gauges that poor data quality costs organizations in the United States over \$700 billion every year (TDWI, 2016).

All things considered, the field of AI is isolated into three sub regions:

Quick Response Code	This paper was presented at	Access this article online
International Conference on Contemporary Issues in Computing (ICIC-2020) - Virtual IETE Sector V, Salt Lake, Kolkata From 25th-26th July 2020	Website: www.intelcomp-design.com	DOI: 10.26480/cic.01.2020.54.58

controlled learning, independent learning, and bolster learning (Adam et al., 2008). Immediately, oversaw learning requires getting ready with named information which has inputs and needed yields. On the other hand with the controlled learning, independent learning doesn't require named planning information and the earth just gives commitments without needed targets. Stronghold taking in enables gaining from analysis traversed correspondences with an external circumstance.

Machine Learning has unfathomable potential results for and is a vital piece of large information assessment (Sun et al., 2014). Figure 1. shows the method old enough of Machine learning. AI is a mix of Computer Science, Engineering and Statistics. Any field that necessities to decode and catch up on information can benefit by AI.

Figure 1: Formation of Machine Learning

2. DEFINITION AND CLASSIFICATION OF MACHINE LEARNING

Machine inclining is a field of research that officially centers on the hypothesis, execution, and properties of learning frameworks and calculations. It is a profoundly interdisciplinary field expanding upon thoughts from a wide range of sorts of fields, for example, man-made brainpower, improvement hypothesis, data hypothesis, measurements, intellectual science, ideal control, and numerous different orders of science, designing, and arithmetic. In light of its execution in a wide scope of uses, machine learning has secured pretty much every logical space, which has expedited incredible effect the science and society. It has been utilized on an assortment of issues, including proposal motors, acknowledgment frameworks, informatics and data mining, and self-ruling control frameworks.

By and large, the field of machine learning is isolated into three sub spaces: directed learning, unaided learning, and support learning. Quickly, directed learning requires preparing with named data which has inputs and wanted yields. Interestingly with the managed learning, solo learning doesn't require named preparing data and the earth just gives contributions without wanted targets. Fortification learning empowers learning from criticism got through connections with an outside domain. In view of these three basic learning ideal models, great deals of hypothesis components and application administrations have been proposed for managing data errands.

From data preparing point of view, administered learning and unaided learning mostly center on data examination while fortification learning is favored for basic leadership issues. Another point is that most customary machine-learning based frameworks are structured with the suspicion that all the gathered data would be totally stacked into memory for concentrated handling. Nonetheless, as the data continues getting bigger and bigger, the current machine learning procedures experience extraordinary troubles when they are required to deal with the exceptional volume of data. These days, there is an extraordinary need to create productive and astute learning techniques to adapt to future data preparing requests (Mitchell, 1977; Russell et al., 1995; Cherkassky et al., 2007; Mitchell, 2006; Rudin et al., 2014; Bishop, 2006; Adam et al., 2008; Jones, 2014; Langford, 2005; Bekkerman et al., 2003; Weiss et al., 2016).

3. HOW ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING ALGORITHMS USE DATA

Artificial Intelligence is one sweeping field of AI, where most talks revolve around learning figuring that use information for working up plans that can be applied to new, unnoticeable information. A count learns its benchmarks subject to the models associated with the readiness information. Most discourse, as the one right now, around supposed coordinated AI which uses named information.

The discourse would likewise apply to independent realizing, where no stamped information is open and groupings of information are being made. The habits by which estimations are readied, through splitting getting ready information into planning, test and endorsement sets is respectably perplexed and not part of the talk of the paper. Moreover, the usage of fortress realizing, where a section of experimentation is associated with the arrangement organize, isn't considered right now.

Huge advances have been made for picture affirmation, where a count is worked by figuring out how to organize pictures. For instance, various photographs including houses are dismembered to make estimation for recognizing houses. The standards reliant on the examination are then attempted against another course of action of pictures, which moreover join houses, for precision. Various models include: information from past bad behavior events that help to get where and when in all likelihood, bad behavior occurs, or information from online life posts (for example the use of unequivocal words or mixes of words) are used to anticipate in case someone is most likely going to tap on an advancement. Information from past prosperity appraisals and their outcomes can help with appreciation and anticipate which patients have a particular disorder or are all things considered raised peril to encounter the evil impacts of savage contamination. Information on scrutinizing history could be used to foresee the pay of people.

It is imperative to take note of that most learning calculations (can only with significant effort) go past the data they use to take in designs from the purported preparing data. While PC researchers put forth all potential attempts to gain from the current data, when a machine learning model is sent and utilized, in actuality, it isn't generally (effectively) conceivable to confirm its prosperity or potential harm.

When utilizing machine learning calculations, there are at any rate three distinct data sets included. This is a rearranged depiction of one sort of machine learning calculations, which should assist with showing signs of improvement comprehension of the kind of data engaged with regularly utilized instances of machine learning calculations.

It is essential to see that most learning estimations (can just with vital effort) go past the data they use to take in plans from the showed arranging data. While PC investigators put forward every single potential undertaking to get from the present data, when an AI model is sent and utilized, actually, it isn't all around (effectively) conceivable to attest its flourishing or potential harm.

When utilizing AI tallies, there are at any rate three undeniable enlightening lists included. This is an adjusted outline of one sort of AI figuring, which should help with giving signs of progress impression of the kind of data attracted with ordinarily utilized events of AI tallies.

Exactly when a count is sent; it is energized with new, unnoticeable features (input information). These are surveyed against the model parameters for taking exercises or choosing. For example, the examining history of people without yet knowing whether they would tap on a particular advancement.

The figuring by then delivers understood checks (or gauges, deducing, determined exercises or out-put information) that are made when the disguised information are empowered into the AI computation.

4. DATA QUALITY FUNDAMENTALS

Because of an expanded dependence upon data to help business choices, inquire about in the territory of data quality has made incredible walks as of late. Scientists have been looking to characterize what is implied by such terms as exactness, fulfillment, and conceivably, and to figure out what quality elements are most significant when evaluating the quality of our data. [9] Shows the different contrasts in both expert and the scholarly perspective on the parts of data quality. While a few scientists have more than 20 measurements for deciding data quality, we will concentrate exclusively on data exactness for this examination. Most specialists remember precision for their rundown of data quality segments, and we will follow the definition and thoughts of (Wand et al., 1996).

Precision is viewed as how close an estimation, or data record, is to this present reality circumstance it speaks to. It is regularly viewed as a natural data quality segment, free of its setting inside the framework. While

substantially more mind boggling and incredible measurements exist to detail our data quality, we decided to concentrate on precision since it is the most effortless to control in our test data sets and it is of all inclusive worry to analysts and chiefs.

Data quality research is principally best in class by software engineering and data frameworks specialists. Software engineering analysts address data quality issues identified with the recognizable proof of copy data, settling irregularities in data, ascription for missing data, connecting and incorporating related data got from different sources (Ganti et al., 2013). PC researchers utilize algorithmic methodologies dependent on statistical strategies to address the above issues (Osborne, 2013). Data frameworks scientists, then again, study data quality issues from a frameworks point of view (McGilvray, 2008). For instance, they research the commitment of user interfaces towards data quality issues. Despite the fact that analysts additionally stand up to data quality issues, the size of their data sets fail to measure up to big data and machine learning situations.

5. DATA VALIDATION

National statistical organizations (NSI) perform Data Validation to test the dependability of conveyed data. Data that appear to be either clearly off-base or conceivably wrong is sent back to the data providers for remedy or remark. This is especially significant in the space of regulatory data that is sent by data providers, (for example, collectives, cantons/states, government units, colleges or schools), and where a two-route correspondence with these is conceivable and fundamental as these data are not gathered for statistical purposes in the primary moment. Up to this point, such Data Validation was acted in for the most part two different ways: either physically by eye-balling through the got data, or consequently through edges and intelligent tests. In specific cases the automation (not performed by machine learning) is adequately exceptional with the goal that the data providers need to address all the apparent blunders and need to give their OK for questionable cases. Therewith, the data that lands at the National statistical organizations is of a more excellent, all the more immediately prepared, and the procedure is more assets productive. We anticipate comparable outcomes with the use of machine learning in this space in a way that probably won't supplant however supplement the current programmed instruments as shown in Figure 2.

The plan to perform machine learning on Data Validation emerged rather by some coincidence. Associates from a similar unit were talking about a specific astonishing example in their data in a get-together with data providers. The astonishing example ended up being a minor wonder that had valid justifications to exist. This prompted some reflex particles about examples in the data. Of the few ten thousand data focuses every year (and in other regulatory data we are talking even of 6 or 7 digits for each year) there may be 'correct' designs, there may be orderly errors, and there may be botches that are recognizable by people to not be right. The inquiry in this manner developed, regardless of whether such examples and 'exceptions' can be recognized by a machine learning calculation – or at the end of the day by a programmed instrument that goes past ex-bet characterized edges or sensible tests (Ruiz et al., 2018).

Figure 2: Big Data Validation Process

6. BIG DATA

Conventional data utilize brought together database engineering in which enormous and complex issues are unraveled by a solitary PC framework. Brought together engineering is exorbitant and inadequate to process huge measure of data. Big data depends on the conveyed database design where a huge square of data is tackled by separating it into a few littler sizes. At that point the answer for an issue is registered by a few distinct PCs present in a given PC arranges. The PCs convey to one another so as to discover the answer for an issue (Weiss et al., 2016). The conveyed

database gives better registering, lower cost and furthermore improve the presentation when contrasted with the brought together database framework. This is on the grounds that incorporated design depends on the centralized servers which are not as financial as microchips in circulated database framework. Additionally, the dispersed database has progressively computational force when contrasted with the unified database framework which is utilized to oversee conventional data.

6.1 Assessing data quality for improvement of data

Many criteria can be looked into when assessing the quality of data for AI applications. In general, data quality includes many different issues; for example questions of completeness, accuracy, consistency, timeliness, duplication, validity, availability and provenance (Burt et al., 2018).

One of the problems of big data is that the sheer size of the data has the tendency to convince us that the findings based on such large-scale data must be accurate. However, if the data quality is not taken into account, this assumption might not hold. Data quantity is only one criterion for the accuracy of measuring or predicting something. To tackle statistical accuracy based on data and determine how well we can represent the real world, it needs to be assessed alongside data quality (Caliskan et al., 2017).

6.2 Data Quality Challenges in Big Data

Data quality is an issue that has been read for quite a few years now. Be that as it may, fundamentally the attention has been on the data in operational databases and data distribution centers. Recently, analysts have started researching data quality issues past the operational and warehousing data. In contrast to the instance of social databases, No SQL (Structured Query Language) frameworks for big data utilize a wide arrangement of data models. The specialist question is: Do we need a different data quality administration approach for each No SQL (Structured Query Language) framework? In big data and machine learning spaces, data is procured from different sellers. Data is likewise produced by publicly supporting, which is supplemented by client contributed data through portable and web applications. How would we survey the veracity and exactness of publicly supported and client contributed data? The expansion of advanced channels and versatile processing is creating more data than any time in recent memory.

What is the effect of cloud organizations on data quality? Should data quality examinations move past segment investigation in social databases and address issues identified with complex data changes, joining of data from differing data sources, and collections that give bits of knowledge into data? (Venkat et al., 2017).

6.2.1 Dealing with Missing Data

Missing data is a significant worry in big data. From a statistical point of view, missing data is ordered into one of the three classifications: Missing Completely At Random (MCAR), missing At Random (MAR), Missing Not At Random (MNAR) (Allison, 2001). In Missing Completely At Random, as the name suggests, there is no example to the missing data. Data is missing freely of both watched and in secret data. The absent and watched values have comparable circulations. As such, Missing Completely At Random is only a subset of the data. Blemish is a misnomer and another name like Missing Conditionally at Random better catches its significance. Given the watched data, data is missing autonomously of imperceptibly data. It is conceivable that there are orderly contrasts between the absent and watched values, and these distinctions can totally be clarified by other watched factors. Missing Completely At Random suggests MAR, yet the other way around. In Missing Not At Random, missing perceptions are identified with the in secret data. Thusly, watched data is a one-sided test. High paces of missing data require cautious consideration, paying little mind to the examination strategy utilized. On the off chance that Missing Completely At Random and MAR cases win for a variable, such factors can be dropped over all perceptions in the data set. Be that as it may, dropping Missing Not At Random factors may prompt outcomes that are emphatically one-sided.

6.2.2 Data Quality Monitoring

This unit expands on the Data Quality Assessment segment's usefulness. It's essential capacity is to constantly screen data quality and report results through dashboards and alarms. Data quality is affected by forms that procure data from the outside world through clump and ongoing

feeds. For instance, an association may obtain data sustains from numerous merchants. Issues emerge when various qualities are accounted for similar data thing. For example, budgetary administrations organizations get continuous value data from various merchants. On the off chance that the occasions packed cost for a money related instrument got from merchants contrast, how does the framework select one of the qualities as the right cost? Seller source chain of command is a way to deal with taking care of this issue. It requires forcing a requesting on the sellers dependent on their notoriety. On the off chance that various costs are accessible for an instrument, select the cost from that seller who has the most elevated notoriety. A way to deal with improving data quality utilizing the clashing data esteems about a similar item starting from different sources is exhibited in (Muller et al., 2012).

Another wellspring of data quality debasement is when frameworks are blended or redesigned. The frameworks being consolidated might be working under various arrangements of business rules and User Interfaces. The data may likewise cover between the frameworks. Surprisingly more terrible, the covering data might be opposing.

Frameworks consolidating commonly require considerable manual exertion and choices made in settling data issues must be recorded as Meta data. Ultimately, changes, reconciliation, and collection likewise add to data quality rot. These procedures may involve loss of data accuracy and scale and supplanting of old identifiers with new ones.

6.2.3 Data Storage, Retrieval & Purging

This part gives tenacious stockpiling, question systems, and the board usefulness to verify and recover data. Both social and No SQL (Structured Query Language) database frameworks are utilized to understand this usefulness (Gudivada et al., 2016). A few data models and question dialects are given to productively store and inquiry organized, semi-organized, and unstructured data (Gudivada et al., 2016). Rules-driven preparing is utilized to cleanse lapsed and stale data.

Here in Table 1, shown below we present some big data tools' performances that will efficiently compare data validation processes in a convenient way.

Table 1: Big Data Tools' Performance Chart

Tools' Name Features	Apache Hadoop	MongoDB	Cassandra	Storm	SAS
Description On Data Validation	The center quality of Hadoop is its HDFS (Hadoop Distributed File System) which can hold all sort of information – video, pictures, JSON, XML, and plain content over a similar record framework.	Provides support for various advances and stages. No hiccups in establishment and upkeep. Reliable and ease.	No single purpose of disappointment. Handles huge information rapidly. Log-organized capacity, Automated replication, Linear versatility and Simple Ring engineering.	Apache Storm is a cross-platform, distributed stream processing and fault-tolerant real-time computational framework. It is free and open-source. The developers of the storm include Backtype and Twitter. It is written in Clojure and Java.	SAS can peruse information documents made by other factual bundles. SAS permits information documents made by SPSS, Excel, Minitab, Stata, Systat, and others to be consolidated into a SAS program legitimately or through record change programming.
Developer	Apache Software Foundation	MongoDB, Inc	Apache Software Foundation	Canonical Ltd.	SAS Institute
Operating System Server	Cross Platform	Linux, Windows, OSX, Solaris	BSD	Cross Platform	Windows, IBM mainframe, Unix/Linux, OpenVMS Alpha
Programming Language	Java	C++	Java	Clojure and Java	C Programming
Usage	It is a software structure utilized for grouped record framework and treatment of huge information. It forms informational indexes of large information by methods for the Map Reduce programming model.	MongoDB is a NoSQL, archive arranged database written in C, C++, and JavaScript. It is allowed to utilize and is an open source instrument that underpins different working frameworks including Windows Vista (and later forms), OS X (10.7 and later forms), Linux, Solaris, and FreeBSD.	Apache Cassandra is liberated from cost and open-source conveyed No SQL DBMS built to oversee immense volumes of information spread over various item servers, conveying high accessibility.	Reliable at scale, Very fast and fault-tolerant, Guarantees the processing of data. It has multiple use cases – real-time analytics, log processing, ETL (Extract-Transform-Load), continuous computation, distributed RPC, machine learning.	SAS has a huge part to play later on for examination and enormous information. Today everybody in the business world must know about the advantages of having SAS abilities and comprehend why these aptitudes are popular in present and future markets. As somebody with SAS skill, here are my top reasons why you ought to learn SAS.

7. CONCLUSION

Machine learning is basic to address the troubles posed by big data and uncover covered patterns, data, and bits of information from huge data remembering the ultimate objective to change the ability of the last into veritable motivation for business essential administration and consistent examination. The quality of data can give the ascent of unfair or in any case incorrect machine learning frameworks. Subsequently, a comprehension of the data quality can assist with comprehension and moderate potential issues of such frameworks. Machine learning frameworks and calculations that utilize data require more extensive and progressively adaptable ways for surveying and tending to data quality corresponding to a few, and interconnected, essential rights.

REFERENCES

Adam, B., I.F.C. Smith, F. Asce. 2008. Reinforcement learning for structural control, J Comput Civil Eng, 22(2), 133–139.

Adam, B., I.F.C. Smith., F. Asce. 2008. Reinforcement learning for structural control, J Comput Civil Eng, 22(2), 133–139.

Allison, P.D. 2001. *Missing Data*, SAGE Publications.

Bekkerman, R., E.Y. Ran, N. Tishby, Y. Winter. 2003. Distributional word clusters vs. words for text categorization, J Mach Learn Res, 3, 1183–1208.

Bishop, C.M. 2006. Pattern recognition and machine learning, (Springer, New York).

Burt, A., et al. 2018. Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models. In, The Future of Privacy Forum White Paper.

Caliskan, A., J.J. Bryson, A. Narayanan, 2017. Semantics derived automatically from language corpora contain human-like biases, Science, 356(6334), 183-186.

- Chen, X.W., X. Lin. 2014. Big data deep learning: challenges and perspectives, *IEEE Access*, 2, 514–525.
- Cherkassky, V., F.M. Mulier. 2007. Learning from data: concepts, theory, and methods, (John Wiley & Sons, New Jersey).
- Ganti, V., A.D. Sarma. 2013. Data Cleaning: A Practical Perspective, ser. Synthesis Lectures on Data Management, Morgan & Claypool Publishers.
- Gudivada, V., D. Rao, V. Raghavan. 2016. Renaissance in database management: Navigating the landscape of candidate systems, *IEEE Computer*, 49(4), 31 – 42.
- Gudivada, V., N.D. Rao, V.V. Raghavan. 2014. NoSQL systems for big data management, in 2014 IEEE World Congress on Services. Los Alamitos, CA, USA: IEEE Computer Society, 190–197.
- Jones, N. 2014. Computer science: the learning machines, *Nature*, 505(7482), 146–148.
- Langford, J. 2005. Tutorial on practical prediction theory for classification, *J Mach Learn Res*, 6(3), 273–306.
- Lee, Y., D. Strong, B. Kahn, R. Wang. 2002. AIMQ: A Methodology for Information Quality Assessment, *Information and Management*. 133-146.
- McGilvray, D. 2008. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information, San Francisco, CA: Morgan Kaufmann Publishers Inc.
- Mehmood, A., I. Natgunanathan, Y. Xiang, G. Hua, S. Guo. 2016. Protection of Big Data Privacy, *IEEE Access*, 4, 1821–1834.
- Mitchell, T.M. 1997. Machine learning (McGraw-Hill, New York).
- Mitchell, T.M. 2006. The discipline of machine learning, (Carnegie Mellon University, School of Computer Science, Machine Learning Department).
- Müller, H.J., C. Freytag, U. Leser. 2012. Improving data quality by source analysis, *J. Data and Information Quality*, 2(4), 1–15, 38.
- Osborne, J.W. 2013. Best practices in data cleaning: a complete guide to everything you need to do before and after collecting your data. Thousand Oaks, CA: SAGE.
- Owen, S., R. Anil, T. Dunning, E. Friedman. 2011. Mahout in Action: Manning Publications Co.
- Panahy, P., Sidi, F., Affendey, L., Jabar, M., Ibrahim, H., Mustapha, A. 2013. Discovering dependencies among data quality dimensions: A validation of instrument. *J. Appl. Sci.*, 13, 95–105.
- Philip, A.K., P. Woodall. 2010. A Hybrid Approach to Assessing Data Quality.
- Rudin, C., K.L. Wagstaff. 2014. Machine learning for science and society, *Mach Learn*, 95(1), 1–9.
- Ruiz, C., S. Federal. 2018. Improving Data Validation using Machine Learning, Neuchâtel, Switzerland, 18-20 September.
- Russell, S., P. Norvig. 1995. Artificial intelligence: a modern approach, (Prentice-Hall, Englewood Cliffs).
- Slavakis, K., G.B. Giannakis, G. Mateos. 2014. Modeling and optimization for big data analytics: (statistical) learning tools for our era of data deluge, *IEEE Signal Proc Mag*, 31(5), 18–31.
- Sun, Y. et al. 2014. Organizing and Querying the Big Sensing Data with Event-Linked Network in the Internet of Things, *International Journal of Distributed Sensor Networks*, 14, 11.
- TDWI. 2016. The data warehousing institute. Available: <https://tdwi.org/Home.aspx>.
- Venkat, N., Gudivada, A., Apon, J. Ding. 2017. Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations, Department of Computer Science, East Carolina University, USA, *International Journal on Advances in Software*, 10, 1-2.
- Wand, Y., R. Wang. 1996. Anchoring Data Quality Dimensions in Ontological Foundations, *Communications of the ACM*, 39, 86-95.
- Weiss, K., T.M. Khoshgoftaar, D.D. Wang, Weiss, et al. 2016. A survey of transfer learning, *J Big Data*, 3, 9. DOI: 10.1186/s40537-016-0043-6