Journal of Computer Science

Original Research Paper

Data Streams Curation for Better Machine Learning Functionality and Result to Serve IoT and other Applications: A Survey

Haya Salah, Islam Al-Omari, Jaber Alwidian, Rashed Al-Hamadin and Tariq Tawalbeh

King Abdullah I School of Graduate Studies and Scientific Research, Princess Sumaya University for Technology (PSUT), Amman, Jordan

Article history:
Received: 08-05-2019
Revised: 30-08-2019
Accepted: 28-10-2019

Corresponding Authors:
Rashed Al-Hamadin
King Abdullah I School of Graduate Studies and Scientific Research, Princess Sumaya University for Technology (PSUT), Amman, Jordan
Email: rashedhamadin@gmail.com

Abstract: Data Curation on data streams is effective in operating and reducing costs of BIG DATA analytic. Basically, analytic preparation requires data curation of available heterogeneous data sets available in big data clusters and such analytic process becomes harder when it comes to the concept of conducting the curation process on Data-on-Motion, in order to come at actionable insights and valuable analytic on a real-time basis including the Machine Learning further analytic and processing. In our paper, we identified and surveyed the different issues and challenges among different areas that are related to the big data. In addition to investigate, the most common techniques and methods followed through the implementations including Streams Curation, the Machine Learning Different Algorithms used in such implementations and the Feature Engineering different techniques that can be considered as curation pre-processing paradigm for data streams analytic. Furthermore, our paper shows the different application areas were data curation concept plays a critical role. Finally, we draw the map between the techniques and methods that are related to the data curation field to emphasize on its main critical role among Business, Retails, Culture, Arts, Health, Medicine, Social Media, Wireless Sensor Networks, Natural Language Processing (NLP) and Automated Feature Engineering (FE). On other hand, we identified the different issues and challenges among different areas including the IoT and Media Streams Curation to help the scholars in this region accordingly.

Keywords: Data Curation, Data Streaming, Data Ingestion, Big Data, Machine Learning

Introduction

One of the main requirements related to the data analytic in big data platforms is the quality of the data ingested in the big data storage pool, as this quality in a moment has the tangible impact on the business operations, retails, health, medicine and many other areas that all are looking for having the best analytic quality for the data located at the Big Data platforms. With Data Curation, the addressing for the data quality issues through the different techniques and methods increase the data management process and enhance the way to make the data usable regardless the number of the data sources, types or behaviors when looking for the data quality (Cavanillas *et al*., 2016). From another point of view, data streams ingestion techniques and requirements are becoming every day more dynamic and easy to implement, but drawing any actions at the moment at which data is ingested is becoming a more challenging part of any big data platforms, as the data-driven decision-making systems are required to build the actionable intelligence for the data being on motion rather than processing it after landing on the big data storage (Yang *et al*., 2017; Padhy *et al*., 2015). Accordingly, the key points to consider when analyzing such sort of data is the quality, types and different sources of the data which will lead to distinguished benefits on the business operations and machine learning processes to be taken on a real-time basis. Also, the increase in the number of the data sources defines" a long tail of data variety", which is a moment provides the adaptive models to be presented to the data consumers that considers covering the data management environments in order to be less frequently used and less structured and more decentralized, thus, improve the data curation on big data platforms on the basis of the cost and number of the curators for the best minimal time (Fig. 1).

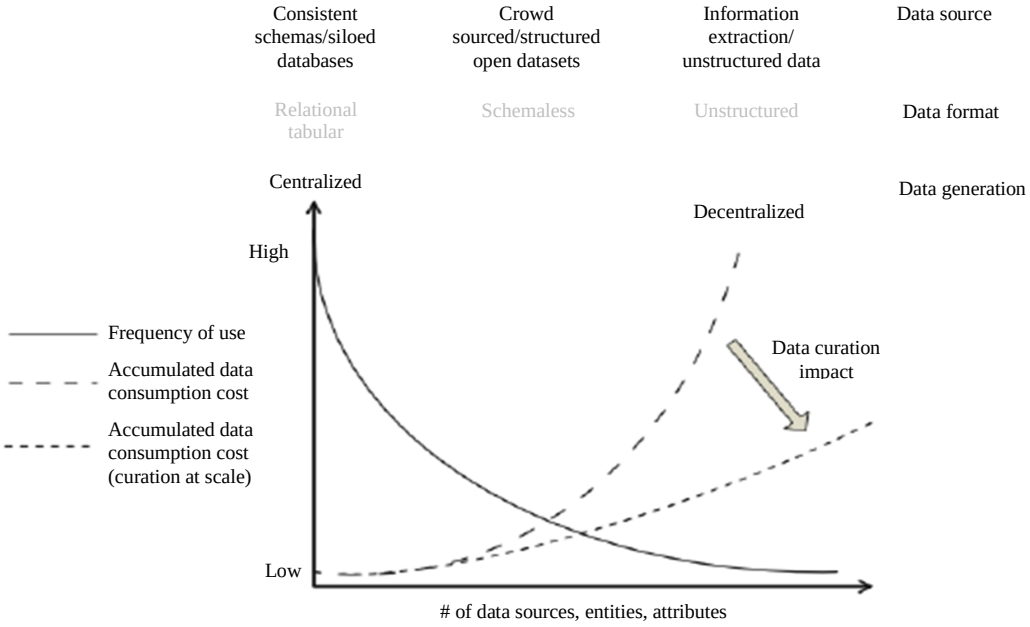

Fig. 1: The long tail of data curation and the scalability of data curation activities. Graph from (Cavanillas *et al*., 2016)

The current study introduced in this paper, will focus on the data curation different aspects, including the main issues, the different techniques followed through the literature and the applications related to data streams and data-on-motion curation concepts and it aims to analyze and show the relationships between such aspects with the Machine Learning best functionality and results, in order to summarize the best practices that can be followed to develop automated Machine Learning pipelines with respect to the data curation different processes that affect the data journey in the big data platforms. Accordingly, this study identifies the data curation issues that prone the machine learning different processes through the IoT generated streams, the automated decision-making analytic, disaster automated responses, anomaly detection, media streaming and feature engineering different implementations. The rest of this study is organized as the following: Detailed data curation issues are summarized in section II, section III shows the different data curation techniques that can be followed, the applications and real-world related examples are summarized in section IV, Section V list future directions and our own discussions and Section VI concludes this study.

Issues

IoT Generated Streams Curation

Issues that can affect the processes automation to curate the data for the sake of implementing high-quality data-driven learning models on the data being ingested in big data platforms are various, for instance structured and unstructured physical IoT devices with the characteristics of high volume, fast data rate and unreliable value would require minimizing the data volume, enhancing the quality of the data and providing fast retrieval process for the data and the curation results from the meta-data and other related caching storage, are all data curating issues that needs consideration when it comes to build such a mechanism. Yang *et al*. (2017), states that enhancing the error detection on the IoT generated streams for the purpose of data cleaning is required and is a critical issue, while enhancing the error recovery, minimizing the data redundancy and preserving the constancy of the data and its quality for the IoT generated streams are critical issues to be addressed when considering data curation at the scale of big data in the cloud data centers. In addition to that, the authors also propose an on-cloud IoT curation framework for IoT generated data streams to solve the issues related alongside with their thoughts for the IoT generated data streams curation issues based on Map Reduce logic. Authors in (Yasumoto *et al*., 2016), provide a very rich survey for the key challenges of the data curation facing the existing technologies related to IoT generated data based on selective use case scenarios and they arrive at identifying six challenges related to the IoT generated data streams curation from different points of view and they also propose a curation framework named Information Flow of Things (IFoT-Curator) that considers the processing, analyzing and curating the IoT streams generated data on real-time bases using distributed processing among IoT devices for the identified issues Fig. 2.

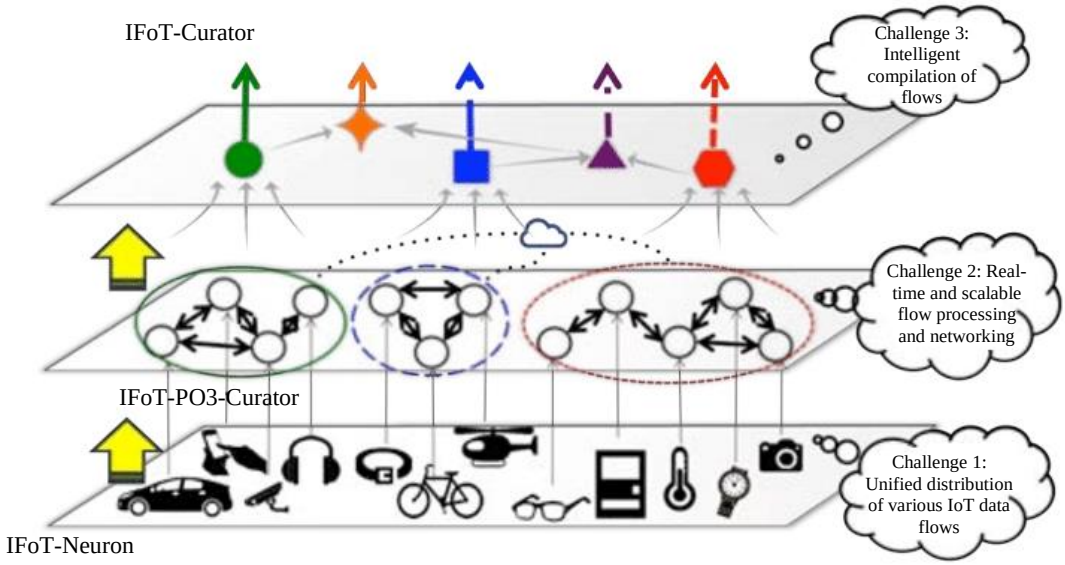

Fig. 2: IFoT Challenges and approaches. Graph taken from reference (Yasumoto *et al*., 2016)

Automating Decision Making Analytics

Challenges that can be found in combining data, its meta-data and a description of the goal needed from the data can lead to the automation of data analytic and this would require using analytic ontology for assisting in data cleaning, suggestions, variable selection and the goal interpretation for the models to use and not only the meta-data learning, which will lead to draw precious explanation for the decision making process based on logical inferences and different machine learning techniques such as Random Forest Algorithm for fast learning on real-time data being ingested (Miller *et al*., 2017; Fujisawa *et al*., 2016). Utilizing the predictive data analytic (Functional Regression, Second-Order Autoregressive time-series model, Functional Data Analysis and Principle Data Analysis) can assess the process of drawing theories from the data on motion or the data that rely on the time (e.g., predicting stock prices from income statement data) and this will enhance the process of analytic automation, especially in the case meta-data are including references to other pre-calculated knowledge based on the successful learning examples (Yasumoto *et al*., 2016), or using wiki technology to make the formation for the knowledge based on the historical data listed in the documents (Tebbakh, 2014).

As a consequence to the precious benefits that we can get from the pre-processed data with the aid of Machine Learning, further automation for the sake of gaining insights from the data based on the pre-processed data in a very valuable area are required. Reference (Choi *et al*., 2017), authors propose a chronological order approach to curate the relations between the topics of the data to enhance the information retrieval from the big data systems and this approach would ensure that data preserves and maintains its value over time. Also Reference (Caeiro-Rodrıguez *et al*., 2013), proposed AREA platform, which is a curation platform that can be used in the educational environments in order to encourage using technologies in the learning environments, where teachers can use it to develop their guidelines and lessons plans in order to share ideas and resources with their students. Such curation processes in the aforementioned areas would be further enhanced with implementing machine learning on real-time basis to enhance the whole approach and raise the reliability of the data with the further inquiries to be made in later stages by the big data end users in such domains, especially in the domains that include the old data Underwood *et al*. (2017).

Disaster Automated Responses

From another point of view, (Zhang *et al*., 2016; Scaramozzino *et al*., 2014) that the high values, high outcome and the diversity of the data types and distribution for the scientific data push toward data curation and because the scientific information agencies require a lot of maintenance operations for the scientific data it includes, a new trend of development and transformation appeared to meet the increasing demand for high computational and data-intensive environments. Developing enhanced and well-prepared management planning for the national scientific data is a basic requirement for revolutionizing the national technologies developed by the government, which requires a vast and high available infrastructure for the big data-intensive environments. And for the actions that should be decided

once the data ingestion starts (Sowe and Zettsu, 2013), then motivating the disaster response scientists taking part in the data curation process is another issue related to this field, as the disasters information would require actionable intelligence based on curated and consisted data on real-time bases.

Streams Anomaly Detection

Poonsirivong and Jittawiriyanukoon, (2017), state that Anomaly Detection in data streams is another data curation issue that would require advanced machine learning techniques in order to identify the data that do not conform to normal data behavior and identify it as an outlier, by the addition to recover data that has errors. This approach push data curation challenges to care about collecting the knowledge from the data and understanding the data attributes changing dynamically with the time, such attributes can include outliers that will prone the statistics datasets, on which, insights to be drawn in actionable intelligent way by machine learning different processes, in some moment, this would be handled by enhancing the error detection and recovery for the IoT generated streams, for instance, (Yang *et al*., 2017).

Media Streams Curation

Manual curation for the images being uploaded on social media is another challenge that requires data curation processing and in order to avoid failing in unprocessed/unnoticed images content, further machine learning approaches using Convolutional Neural Network (CNNs) and transfer learning is proposed by the authors (Tous *et al*., 2016), were they state that content curation in other terms with the aid of machine learning on real-time bases can be implemented by facilitating the mechanism of the generated content searching and discovering functionality for enhancing images quality being uploaded to the social media websites by using more than one CNN in consequence on the images streaming data.

Feature Engineering

As the most tackling part in the machine learning and deep learning projects is the feature listing and extraction that would require advanced techniques to be implemented in order go over such issue (Najafabadi *et al*., 2015; Bode *et al*., 2019), many approaches where followed through the literature in order to minimize the time required to handle such type of data curation on IoT streams and data in motion. Banerjee *et al*. (2018) conduct a very rich survey on people involved in IoT analytic application development in order to identify the analysis pain areas related to IoT analytic tasks, the survey analysis shows that, domain knowledge and technical knowledge are required at the maximum limit when it comes to feature listing, selection and reduction, authors continue in identifying the steps involved in IoT applications processing and analytic, finally they propose and test a feature recommendation architecture to tackle the IoT analytic related to feature engineering.

Techniques

Streams Curation Techniques

Different techniques and methods were followed through the literature in order to prepare the data for learning and analytic further processing phases during the data creating and streaming stages and those methods vary depending on the phase and the purpose of the data ingestion stage, i.e., for example, some methods care about the data quality while others annotate the data with the learning goals to be achieved (Cavanillas *et al*., 2016). In mining social media images and movies, available features extracted from the social curation lists using k-mean and bag-of-visual-words (representing BoW of text), can lead to valuable media content understanding and learning processes, such curation lists can be maintained on a weekly basis due to the heavy processing of the data at time required to handle such type of data curation on once and to enhance this processing and minimize the time needed to perform it, this would require data to be curated at the ingestion stage (i.e., creation time). Ishiguro *et al*. (2012), states three types of users in the social data curation process, the first type is Content Creators, who generate social media content, including text messages, images and videos, the second is Content Curators, who evaluate and collect posted social media content according to their, or the domain's perspective and finally, Content Consumers who consume the creators social media content. Having the functional regressions for the variables that are created at some moment during the ingestion process (Miller *et al*., 2017), would save the manual efforts done by the Content Curators and save time in this regards.

When it comes to heavy streaming, more advanced curation techniques to be considered, (Fujisawa *et al*., 2016), for example, do a video content curation for the cameras on real-time basis in order to create a video consisting of scenes from the different cameras located at different corners at each point of time for sport games is another challenging application for such data curation process, which can be conducted by training a machine learning model on the different images features, the game data and the camera different positions at each point of time, this would require dividing the videos into small segments that facilitate

the process of applying machine learning technique. Such learning process implemented by collecting the multiple video streams, which are in turn divided into small segments that are assigned metadata describing them as a curation process, followed by building a time-based estimation model to generate the automatic cameras switching hence curating the games' video content by using the estimated right corner to capture the scenes at the right time based on machine learning and the curated streams. Implementing such advanced mechanism for streaming and preparing the data to take prompt decisions in switching the cameras, need more steps to taken into account and in this context Salarian *et al*. (2015), proposed mechanism that can handle the massive data amount processing at the fly including the data streaming, the data cleaning and the needed data operations to prepare it for the next processing stages and grips mobility context processing based (Apache Hadoop) cloud environment and MapReduce (Dean and Ghemawat, 2004).

Machine Learning Techniques

Introducing the Massive Online Analysis (MOA) to many of the nowadays streaming applications raises the need for different methodologies and techniques to be followed in order to handle the Machine Learning on real-time basis (Bifet *et al*., 2018). Miller *et al*. (2017), lists different studies and opinions about using advanced machine learning techniques in data curation, starting with the Second-Order Autoregressive time-series modeling (Adhikari and Agrawal, 2013), for Predictive Analytic combined with the Functional Data Analysis that can capture the phenomena that evolves over the time, e.g. using two input and output functions to describe the rate of schange in the output function using Ordinary Differential Equations (ODE) (Kolokolnikov, 2003). Functional regressions (Greven and Scheipl, 2017), can be used when measuring the covariates variables influences in responses to other variables in the case of the variables are functional variables. Finally, Principle Differential Analysis (Ramsay, 2006) to detect the cases when describing the change in the output functions using ODE. Random Forest Learning Algorithm, which can be considered the fastest and the most suitable algorithm for learning on a real-time basis is considered by the authors in (Fujisawa *et al*., 2016).

Feature Engineering

Supervised Machine algorithms require labeled data set, but when it comes for data on streams curation in terms of features extraction and selection as a pre-processing step for the automated machine learning, unsupervised machine learning techniques show better results in improving the classification process for the time series data. Bode *et al*. (2019) shows a detailed comparison for the accuracies gained for some selective unsupervised machine learning techniques in regards to unsupervised feature extraction and clustering against the normal supervised approaches for the statistical feature selection Fig. 3.

Li *et al*. (2017; Alnuaimi *et al*., 2019) State a valuable surveys studies on the feature selection different algorithms that can be used with different types of data Fig. 4.

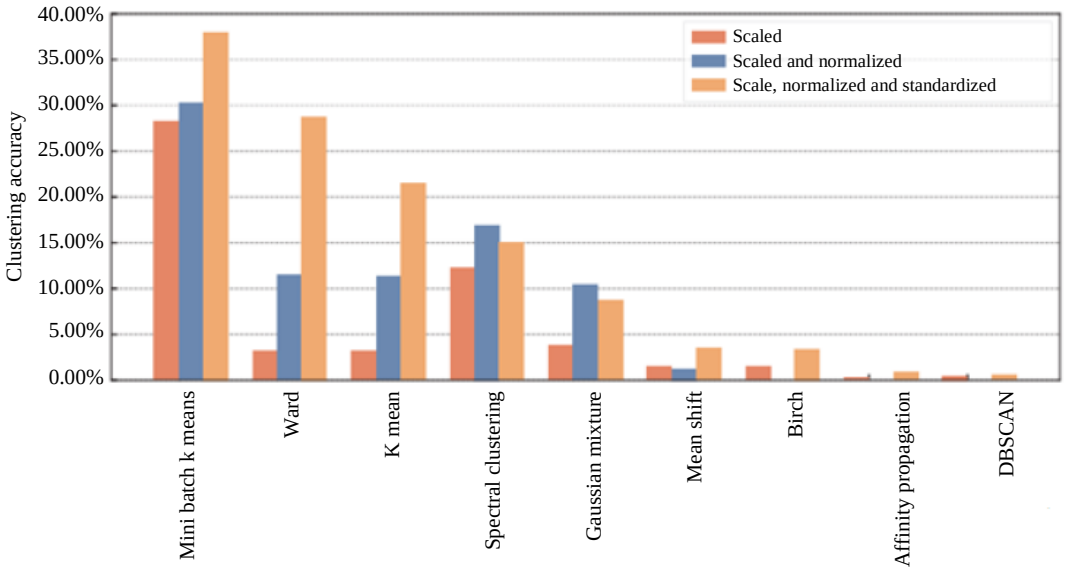

Fig. 3: Accuracy of clustering algorithms using unsupervised extracted features. Graph taken from (Bode *et al*., 2019)

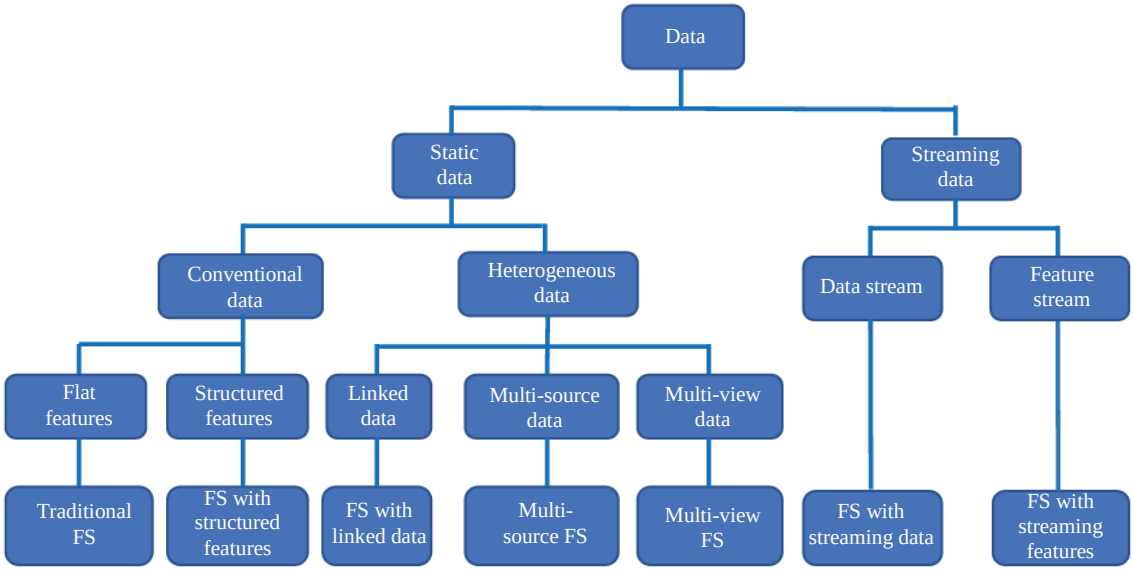

Fig. 4: Feature selection algorithms from the data perspective. Graph taken from (Li *et al*., 2017)

Applications

Business and Retail

The tangible impact of data curation on the business and retail world is noticeable through the direct positive effect it exhibits on an enterprises' return on investment and, ultimately, its generated revenues. The introduction of data curation into the business domain largely contributes to the delivery of higher quality information concerning a customers' interaction with any of the offered goods and services and is hence an asset to businesses as it helps ensure the success of its' objectives (Abe, 2014). Abe (2012) highlights the impact of the implementation of a proper data curation strategy on the success of marketing promotion. Multiple marketing approaches, such as digital marketing, viral marketing and salespeople, are deployed to guarantee the success of a promotion. Another hidden and often unstudied factor is the influential customers who are able to affect other customers decisions and hence present a marketing opportunity to further increase uptake, as more information concerning customers adoption of offered goods and services is obtained. Data regarding such influences is often difficult to obtain, hence creating a gap in any performed analytic, therefore, an abduction strategy in conjunction with data curation is proposed by the author to determine any gaps in the available data and further improve the accuracy of marketing analytic by improving data quality.

An important big data application in the business world is the analysis of event logs, which helps optimize business strategies as more accurate insights into the business processes are obtained. Often, event logs are analyzed by third-party business analysts, hence the obstacles of data privacy and the existence of differing event log formats at different enterprises are encountered. Furthermore, business analysts rarely have previous programming experience and hence present yet another obstacle. Kudo *et al*. (2016) as a result introduces a privacy-preserving curation scheme, achieved through the use of machine learning, that allows third-party business analysts with little programming experience to perform analytic on event logs from different enterprises while ensuring that confidential data is not disclosed.

Culture and Arts Preservation

The presearvation of local cultural heritage is beneficial for upcoming generations and can be achieved by the exploitation of artifacts found at local museums as well as any local cultural activities. Access to big data nowadays presents numerous opportunities for the protection and preservation of local culture, as huge volumes of information become more readily available for use. As a result, sources of knowledge regarding culture are no longer restricted to historical objects found at local museums or local events, but also includes social media user-generated content. Yoshioka and Iida (2016) demonstrates the role of data curation in the preservation of local culture, by developing a platform for the curation of local music. Cui *et al*. (2013) proposes the use of data curation tools to combine tourism and culture data from various sources, including local museums and web sources such as blogs and local tourism portals for the purpose of cultural preservation, by making the curated data available on the web through an "online museum". Thus, the value of data curation in the process of local

culture and art preservation is found in its ability to enhance the results of machine learning technologies implemented for the purpose of promoting tourism experiences online. Hence, enhancing the quality of data helps ensure the enrichment and preservation of knowledge of local cultural heritage.

Health and Medicine

Quality of data is a crucial factor to the accuracy of analytic; Any depreciation in data quality renders it useless as it negatively impacts any conducted analysis. This issue is particularly important in the healthcare sector, as important decisions are based on data analytic. Pezoulas *et al*. (2019), which proposes a data curation framework consisting of three layers; data evaluation module, data quality control module and the data standardization module, ensures a curated dataset of high quality as an output. The introduced framework is particularly useful in the medical industry and can be easily integrated with big data platforms deployed by health care institutions, to ensure the continual assessment of the quality of electronic medical and health records. Hence, it's important to ensure the availability of curated datasets in the healthcare industry, which is currently driving research regarding the creation of platforms specifically designed for the curation of healthcare data (Landis *et al*., 2015; Wollatz *et al*., 2018; Sultanum *et al*., 2018).

Progress in machine learning yields advances in big data analytic; furthermore, higher levels of prediction and classification accuracy are obtained as the volume of available data increases. The successful deployment of machine learning systems, therefore, requires large data volumes and computational resources, which are often not readily available for individual use. Dao *et al*. (2018) hence, introduce the DATABRIGHT data curation platform for machine learning. The platform transforms traditional crowd-sourcing, which lacks proper control of data quality and diversity, into data contribution by offering "data shares" to individuals who contribute data to the platform, hence creating a repository of curated data, exhibiting quality and diversity that is suited for machine learning activities. Successful applications of the platform include the use of curated microscopy cellular tissue images for the building of a drug discovery prediction model. This is due to the peer-review introduced by the DATABRIGHT platform which supports the peer-review of domain experts, ensuring the quality of any curated data and improving the accuracy of predictive models, than when used on already existing data sets. Another example on the effect of data curation on machine learning processes in the healthcare industry is presented by (Kotsampasakou *et al*., 2017), which demonstrate how the use of a curated dataset improves the accuracy of a drug-induced liver injury prediction model, due to the enhanced quality of curated datasets compared to other un-curated sets, which in turn reflects itself on the accuracy of the employed machine learning model.

Social Media

The vast amounts of information present on social networking sites present various opportunities in the area of big data analytic as valuable knowledge can be extracted from this data. Data curation in the area of social media revolves around the compilation of social media content, including text, images and videos. The adoption of social media as a regular part of the lives of millions of people around the world has led to the empowerment of consumers in the market place as their communication and decision-making processes continue to be influenced by social media daily. Consumers can continually assess what products suit their needs by leveraging their access to massive amounts of information available through social networks, as well as which advertisements to adopt or avoid. The aforementioned consequences along with consumers ability to generate user-content empower them further as they are able to influence the opinions of their peers (Rafailidis *et al*., 2014). Tous *et al*. (2016) demonstrate the use of data curation in the social networking world for business applications, by proposing the use of deep convolutional neural networks for the curation and filtering of user-generated content on social networking sites, for the aid of digital marketing, by introducing the identification of a brand's visual identity in SNS user-generated content. The introduction of machine learning into the curation process increases the size of the resulting dataset, enhances the quality of the data and enables the discovery of UGC in near real-time.

Ishiguro *et al*. (2012) employ data curation in the area of social media as a source of information for the automatic data mining of image content found on social media. The paper studies the prediction of image and video content view count and demonstrates the improvement introduced in the model's achieved accuracy through the use of a curated dataset. Ishiguro *et al*. (2012) introduce the CrisisTracker system which utilizes Twitter user's activity for awareness reports creation during disasters, with the use of data curation. Access to areas affected by disasters is often difficult hence it's necessary to employ alternate communication mechanisms (Rogstadius *et al*., 2013). It is common that individuals residing in areas afflicted by disaster often use Social Networking Sites to share information about their own or their surrounding conditions, in various forms, including text, images and videos. Hence, the exploitation of such information presents great value and offers support to the afflicted communities when used for the purpose of

awareness reports creation. CrisisTracker, therefore, is designed to combine data curated from the Twitter platform with automated data collection and real-time text clustering to provide a real-time overview of the content generated on Twitter by individuals during disasters. Hence, the use of curated social media data with a machine learning technique, namely clustering in this case, guarantees the improvement of achieved scalability and accuracy and most importantly, offering support to people afflicted by disasters.

Wireless Sensor Networks

Wireless Sensor Networks (WSN) generate vast amounts of sensor data, which present a challenge to the traditional analytical capabilities, techniques and tools. IoT sensor datasets are complex and are often characterized by their huge volume, availability under different structured and unstructured formats and high rate of change, as well as their unreliable quality, which is highly affected by the used wireless communication technology, signal processing and sensors. As a result, the introduction of data curation into the deployment of WSN in IoT is essential to any industrial and scientific applications benefiting from the technologies (Yasumoto *et al*., 2016). As demonstrated by (Yang *et al*., 2017) the use of curated WSN streamed datasets in prediction models improves the obtained model accuracy as it positively impacts the quality of the dataset. The efficiency of the curation process is enhanced through the introduction of data analytic for feedback generation, to be implemented in combination with any data filtering processes. Furthermore, the proposed curation scheme is not only restricted to the improvement of the collected dataset, but is optimized to generate and send feedback to the wireless sensors used in the network to update their locations ensuring that the strength of the network is maintained, hence improving the quality of data prior to its' generation and collection.

Natural Language Processing

Natural Language Processing Algorithms nowadays are becoming more popular and valuable in textual data processing and data curation combined with this sort of algorithms may yield valuable insights for the data being on the motion. Miyamoto *et al*. (2017), introduce a system for enhancing Question-Answering automation based on NLP and data curation that is intended for detecting the language, validating the desired pairs of the questions and answers based on predefined patterned-based topic segmentation for enhanced next levels of data processing.

And when it comes to the journalists reviewing the history work, Fulda *et al*. (2016) introduce a data curation techniques that can help in accelerating the authoring for the time-based events from unstructured data in minutes and rule out unsuitable data parts in seconds by the authors, in addition, it visualizes those events form the unstructured data to help to compare and browsing this type of data simultaneously. This would require different techniques that consider the tense of the governing verb to determine whether the temporal expressions refer to the future or the past and it determines if a four-digit number refers to a year depending on the context using different tools and techniques.

Automated Feature Engineering

Applications proposed to automate the process of feature extraction and selection are vary depending on the purposes and its role in ML pipelines, for instance, (Kanter and Veeramachaneni, 2015) propose an algorithm for deep features construction named FeatureTools for relational data. TPOT is another open source platform for automated features engineering including construction and selection for the features (Olson *et al*., 2016a; 2016b), (Gijsbers *et al*., 2019). Balaji and Allen, (2019) states a bench-marking methodology for different automated machine learning open source platforms including TPOT.

Comparative Study

In this work, the identified and surveyed different issues and challenges among different areas including the IoT and Media Streams Curation and Anomaly Detection, Automating Decision Making and Disaster Actionable Responses Analytic and Automated Feature Engineering can be found in Table 1. In Table 2, a comparative study for the most common techniques and methods followed through the different applied implementations through the literature including Streams Curation, the Machine Learning Different Algorithms used in such implementations and the Feature Engineering different techniques that can be considered as curation pre-processing paradigm for data streams analytic. The number of challenges distribution among the suitable applications as a comprehensive content for the comparative study can be found in Fig. 5. Table 3, highlights the different application areas were data curation concept plays a critical role among Business, Retails, Culture, Arts, Health, Medicine, Social Media, Wireless Sensor Networks, Natural Language Processing and Automated Feature Engineering (FE), also it provide more details about the application and what are the main concepts and features for each one that leads to facilitate the selection of the suitable application based on the recommended features that are required.

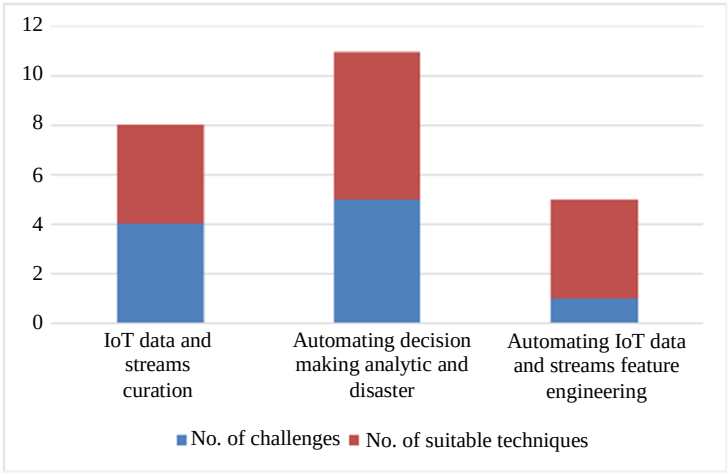

Fig. 5: The number of challenges distribution among the suitable applications.

Table 1: Most highlighted challenges (Issues) to data streams curation and pre-processing

Field	Challenges	References
IoT Data and StreamsCuration	1. Structured and unstructured physical IoT devices (Minimizing Volume, Enhancing Quality and Fast Retrievals) (Cavanillas *et al*., 2016) 2. Distributed IoT Flows (Unification, Real-Time Scalable Flows Processing and Networking and Intelligent Flows Compilation) (Yasumoto *et al*., 2016) 3. Data streams anomaly detection and error recovery for data on streams (Cavanillas *et al*., 2016; Poonsirivong and Jittawiriyanukoon, 2017). 4. Manual curation for the images being uploaded on social media, automated images content curation and generated content discovery on a real-time basis using CNN and transfer learning (Tous *et al*., 2016)	Cavanillas *et al*. (2016; Yasumoto *et al*., 2016; Poonsirivong and Jittawiriyanukoon, 2017; Tous *et al*., 2016).
Automating Decision Making and Disaster Automated Responses	1. Enhancing data reliability, cleaning, suggestion, selection and goal interpretation using Machine Learning (Miller *et al*., 2017; Fujisawa *et al*., 2016; Underwood *et al*., 2017). 2. Drawing theories and knowledge from data-on-motion using predictive analytics (Yasumoto *et al*., 2016) or wiki technology (Tebbakh, 2014) 3. Enhancing information retrieval and data reliability from historical data using Machine Learning (Choi *et al*., 2017; Caeiro-Rodrıguez *et al*., 2013; Underwood *et al*., 2017). 4. Maintain the scientific data to guarantee high values, outcome and diversity (Zhang *et al*., 2016; Scaramozzino *et al*., 2014). 5. Actionable intelligence based on curated and consisted of data on real-time bases for disaster automated responses. (Sowe and Zettsu, 2013)	Yasumoto *et al*. (2016; Miller *et al*., 2017; Fujisawa *et al*., 2016; Tebbakh, 2014; Choi *et al*., 2017; Caeiro-Rodrıguez *et al*., 2013; Underwood *et al*., 2017; Zhang *et al*., 2016; Scaramozzino *et al*., 2014; Sowe and Zettsu, 2013)
IoT Data and Streams Engineering	1. Minimizing the time of features listing, extraction, selection and reduction using machine learning and deep learning (Najafabadi *et al*., 2015; Bode *et al*., 2019; Banerjee *et al*., 2018)	Najafabadi *et al*. (2015; Bode *et al*., 2019; Banerjee *et al*., 2018).

Table 2: Top applied areas and the most suitable techniques (methods) for each one.

Field	Techniques	References
IoT Data and Streams Curation	1. Social Media and Images: K-Mean and BOW (Ishiguro *et al*., 2012). 2.Content Manual Curation: Variables Functional Regressions (Miller *et al*., 2017). 3.Video Streaming: Dividing video content and assign Meta-data to feed Time-based estimation model in real time (Fujisawa *et al*., 2016). 4. Massive Mobility Context Processing at run time (Salarian *et al*., 2015)	Ishiguro *et al*. (2012; Miller *et al*., 2017; Fujisawa *et al*., 2016; Salarian *et al*., 2015)
IoT Data and Streams Machine Learning	1. Massive Online Analysis (MOA) (Kolokolnikov, 2003). 2. Phenomena that evolve over time: Second-Order Autoregressive Time-Series Modeling, Functional Data analysis and Ordinary Differential Equation (ODE) (Miller *et al*., 2017; Theodore, 2003). 3. Covariates Variables Influences to others in case functional variables are represented (Miller *et al*., 2017; Ramsay, 2006). 4. Detect cases when describing change: Principle Differential Analysis using ODE (Miller *et al*., 2017; Li *et al*., 2017). 5. Real-Time Learning: Random Forest Learning Algorithm (Fujisawa *et al*., 2016). 6. Classification process for time series data: Unsupervised learning (Bode *et al*., 2019).	Miller *et al*. (2017; Fujisawa *et al*., 2016; Bode *et al*., 2019; Adhikari and Agrawal, 2013; Kolokolnikov, 2003; Ramsay, 2006; Li *et al*., 2017)
IoT Data and Streams Feature Engineering	1. Feature extraction using unsupervised learning (Bode *et al*., 2019). 2. Feature streams and data streams FE (Alnuaimi *et al*., 2019; Abe, 2014). 3. Heterogeneous data, Multi-source data and linked data (Alnuaimi *et al*., 2019; Abe, 2014). 4. Conventional Data: Flat features and structured features (Alnuaimi *et al*., 2019; Abe, 2014).	Bode *et al*. (2019; Alnuaimi *et al*., 2019; Abe, 2014).

Table 3: Most highlighted applications on data curation

Application type	Brief Desc.	References
Business and Retail	1.Delivery of higher quality information concerning a customers' interaction with any of the offered goods and services (Abe, 2014). 2.The success of a marketing promotion (Abe, 2012). 3.Privacy-Preserving confidential scheme for event logs analysis (Kudo *et al*., 2016).	Abe (2014; 2012; Kudo *et al*., 2016) Yoshioka and Iida (2016;
Culture and Arts Preservation	1. Local music curation (Yoshioka and Iida, 2016). 2. Combine tourism and culture from different sources (Cui *et al*., 2013)	Cui *et al*., 2013)
Health and Medicine	1. Ensure the continual assessment of the quality of electronic medical and health records (Pezoulas *et al*., 2019). 2. Healthcare data curation (Landis *et al*., 2015; Wollatz *et al*., 2018; sultanum *et al*., 2018). 3. Machine Learning for the use of microscopy cellular tissue images for building of a drug discovery prediction model – DATABRIGHT (Dao *et al*., 2018). 4. The effect of data curation on machine learning processes in the healthcare industry, and improving accuracy (Kotsampasakou *et al*., 2017).	Pezoulas *et al*. (2019; Landis *et al*., 2015; Wollatz *et al*., 2018; Sultanum *et al*., 2018; Dao *et al*., 2018; Kotsampasakou *et al*., 2017).
Social Media	1. The empowerment of consumers in the market place as their communication and decision-making processes continue to be influenced by social media daily (Rafailidis *et al*., 2014). 2. The use of data curation in the social networking world for business applications, by proposing the use of deep CNN for the curation and filtering of user-generated content on social networking sites (Kolokolnikov, 2003). 3. Data curation in the area of social media as a source of information for the automatic data mining of image content found on social media (Ishiguro *et al*., 2012). 4. The CrisisTracker system which utilizes Twitter user's activity for awareness reports creation during disasters, with the use of data curation (Ishiguro *et al*., 2012).	Rafailidis *et al*. (2014; Kolokolnikov, 2003; Ishiguro *et al*., 2012; Rogstadius *et al*., 2013; Reymondet *et al*., 2016)
Wireless Sensor Networks	1. The introduction of data curation into the deployment of WSN in IoT is essential to any industrial and scientific applications benefiting from the technologies (Yasumoto *et al*., 2016). 2. The use of curated streamed datasets in prediction models improve the obtained model accuracy as it positively impacts the quality of the dataset (Yang *et al*., 2017).	Yasumoto *et al*. (2016; Yang *et al*., 2017).
Natural Language Processing	1.Question Answering System based on NLP (Reymonde *et al*., 2016). 2. Journalist reviewing the history textual content and accelerating the authoring process for time-based events (Fulda *et al*., 2016).	Reymonde *et al*. (2016; Fulda *et al*., 2016)
Automated Feature Engineering	1. Algorithm for deep features construction for relational datasets (FeatureTools) (Kanter and Veeramachaneni, 2015), 2. TPOT open source platform for automated FE including construction and selection, (Olson *et al*., 2016a; 2016b; Gijsbers *et al*., 2019). 3. Bench-Marking for different automated machine learning tools including TPOT (Balaji and Allen, 2019).	Kante and Veeramachaneni. (2015; Olson *et al*., 2016a; 2016b; Gijsbers *et al*., 2019; Balaji and Allen, 2019)

Conclusion

Our paper shows that the data curation concept plays a critical role in enhancing the today's working data systems and the key points to consider upon employing such role are the valuable benefits that fit the Data-on-Motion and Machine Learning characteristics on Real-Time basis. However, existing challenges that need collaboration and further proposals to be solved in order to enhance the future processes related to automating the analytic and machine learning different schemes. Our comparative study investigates the existing aspects related to data curation processes in term of the different issues and challenges, the used techniques and methods and finally the applied areas it employed in. This paper aims to aid the researchers in related fields to advance their research toward having curated data-driven advancements, as it highlights the most important and common recent literature references related to such research fields.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abe, A., 2012. Data mining in the age of curation. Proceedings of the IEEE 12th International Conference on Data Mining Workshops, Dec. 10-10, IEEE Xplore Press, Brussels, Belgium, pp: 273-279. DOI: 10.1109/ICDMW.2012.114

Abe, A., 2014. Data mining considering curation. Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, Aug. 11-14, IEEE Xplore Press, Warsaw, Poland, pp: 272-283. DOI: 10.1109/WI-IAT.2014.123

Adhikari, R. and R. Agrawal, 2013. An introductory study on time series modeling and forecasting. arXiv preprint arXiv: 1302.6613, 2013.

Alnuaimi, N., M.M. Masud, M.A. Serhani and N. Zaki, 2019. Streaming feature selection algorithms for big data: A survey. Applied Comput. Inform. DOI: 10.1016j.aci.2019.01.001.

Apache Hadoop. http://hadoop.apache.org/

Balaji, A. and A. Allen, 2019. Benchmarking automatic machine learning frameworks. arXiv.org.

Banerjee, S., T. Chattopadhyay, A. Pal and U. Garain, 2018. Automation of feature engineering for IoT analytics. ACM SIGBED Rev., 15: 24-30. DOI: 10.1145/3231535.3231538

Bifet, A., R. Gavalda, G. Holmes and B. Pfahringer, 2018. Machine Learning for Data Streams: with Practical Examples in MOA. 1st Edn., MIT Press, ISBN-10: 0262037793, pp: 288.

Bode, G., T. Schreiber, M. Baranski and D. Muller, 2019. A time series clustering approach for building automation and control systems. Applied Energy. 238: 1337-1345. DOI: 10.1016/j.apenergy.2019.01.196

Caeiro-Rodrıguez, M., R. Perez-Rodrıguez, J. Garcıa-Alonso, M. Manso-Vazquez and M. Llamas-Nistal, 2013. AREA: A social curation platform for open educational resources and lesson plans. Proceedings of the IEEE Frontiers in Education Conference, Oct. 23-26, IEEE Xplore Press, Oklahoma City, OK, USA, pp: 795-801. DOI: 10.1109/FIE.2013.6684935

Cavanillas, J.M., E. Curry and W. Wahlster, 2016. New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe. 1st Edn., Springer International Publishing AG, ISBN-10: 3319215698, pp: 303.

Choi, S., J. Seo, M. Kim, S. Kang and S. Han, 2017. Chrological big data curation: A study on the enhanced information retrieval system. IEEE Access, 5: 11269-11277. DOI: 10.1109/ACCESS.2016.2642979

Cui, B., W. Wang, W. Zhou and S. Yokoi, 2013. An exploration of protecting local culture via content curation in local online museum. Proceedings of the International Conference on Signal-Image Technology and Internet-Based Systems, Dec. 2-5, IEEE Xplore Press, Kyoto, Japan, pp: 391-395. DOI: 10.1109/SITIS.2013.70

Dao, D., D. Alistarh, C. Musat and C. Zhang, 2018. Data bright: Towards a global exchange for decentralized data ownership and trusted computation. arXiv:1802.04780

Dean, J. and S. Ghemawat, 2004. MapReduce: Simplified data processing on large clusters. Proceedings of the 6th conference on Symposium on Opearting Systems Design and Implementation, (SDI' 04), San Francisco CA.

Fujisawa, K., Y. Hirabe, H. Suwa, Y. Arakawa and K. Yasumoto, 2016. Automatic content curation system for multiple live sport video streams. Proceedings of the IEEE International Symposium on Multimedia, Dec. 14-16, IEEE Xplore Press, Miami, FL, USA, pp: 541-546. DOI: 10.1109/ISM.2015.17

Fulda, J., M. Brehmel and T. Munzner, 2016. TimeLineCurator: Interactive authoring of visual timelines from unstructured text. IEEE Trans. Vis. Comput. Graph., 22: 300-309. DOI: 10.1109/TVCG.2015.2467531

Gijsbers, P., J. Vanschoren and R. Olson, 2019. Layered TPOT: Speeding up tree-based pipeline optimization. arXiv.org.

Greven, S. and F. Scheipl, 2017. A general framework for functional regression modelling. Stat. Modell., 17: 1-35. DOI: 10.1177 1471082x16681317.

Ishiguro, K., A. Kimura and K. Takeuchi, 2012. Towards automatic image understanding and mining via social curation. Proceedings of the IEEE 12th International Conference on Data Mining, Dec. 10-13, IEEE Xplore Press, Brussels, Belgium, pp: 906-911. DOI: 10.1109/ICDM.2012.37

Ishiguro, K., A. Kimura and K. Takeuchi, 2012. Towards automatic image understanding and mining via social curation. Proceedings of the IEEE 12th International Conference on Data Mining, Dec. 10-13, IEEE Xplore Press, Brussels, Belgium, DOI: 10.1109/icdm.2012.37.

Kanter, J.M. and K. Veeramachaneni, 2015. Deep feature synthesis: Towards automating data science endeavors. Proceedings of the IEEE International Conference on Data Science and Advanced Analytics, Oct. 19-21, IEEE Xplore Press, Paris, France. DOI:10.1109/dsaa.2015.7344858.

Kolokolnikov, T., 2003. First-order ordinary differential equations, symmetries and linear transformations. Eur. J. Applied Math., 14: 231-246. DOI: 10.1017/S0956792503005126

Kotsampasakou, E., F. Montanari and G.F. Ecker, 2017. Predicting drug-induced liver injury: The importance of data curation. Toxicology, 389: 139-145. DOI: 10.1016/j.tox.2017.06.003.

Kudo, M., K. Maeda and F. Satoh, 2016. Adaptable privacy-preserving data curation for business process analysis services. Proceedings of the IEEE International Conference on Services Computin, Jun. 27-27, IEEE Xplore Press, San Francisco, CA, USA, pp: 411-418. DOI: 10.1109/SCC.2016.60.

Landis, D., W. Courtney, C. Dieringer, R. Kelly and M. King *et al*., 2015. COINS data exchange: An open platform for compiling, curating and disseminating neuroimaging data. NeuroImage, 124: 1084-1088. DOI: 10.1016/j.neuroimage.2015.05.049.

Li, J., K. Cheng, S. Wang, F. Morstatter and R.P. Trevino *et al*., 2017. Feature selection. ACM Comput. Surveys, 50: 1-45. DOI: 10.1145 3136625

Miller, J.A., H. Peng and M.E. Cotterell, 2017. Adding support for theory in open science big data. Proceedings of the IEEE World Congress on Services, Jun. 25-30, IEEE Xplore Press, Honolulu, HI, USA, pp: 251-255.
DOI: 10.1109/SERVICES.2017.20

Miyamoto, K., A. Koseki and M. Ohno, 2017. Effective data curation for frequently asked questions. Proceedings of the IEEE International Conference on Service Operations and Logistics and Informatics, Sep. 18-20, IEEE Xplore Press, Bari, Italy, pp: 7-12. DOI: 10.1109/SOLI.2017.8120960

Najafabadi, M.M. T.M. Khoshgoftaar, N. Seliya and R. Wald *et al*., 2015. Deep learning applications and challenges in big data analytics. J. Big Data.
DOI: 10.1186/s40537-014-0007-7

Olson, R.S., N. Bartley, R.J. Urbanowicz and J.H. Moore, 2016b. Evaluation of a tree-based pipeline optimization tool for automating data science. Proceedings of the Genetic and Evolutionary Computation Conference, Jul. 20-24, Denver, Colorado, USA, ACM, pp: 485-492.
DOI: 10.1145/2908812.2908918

Olson, R.S., R.J. Urbanowicz, P.C. Andrews, N.A. Lavender and L.C. Kidd *et al*., 2016a. Automating biomedical data science through tree-based pipeline optimization. Proceedings of the European Conference on the Applications of Evolutionary Computation, (AEC' 16), Springer, Cham, pp: 123-137.

Padhy, S., S. Padhy, G. Jansen, J. Alameda and E. Black *et al.*, 2015. Brown dog: Leveraging everything towards autocuration. Proceedings of the IEEE International Conference on Big Data, Oct. 29-Nov. 1, IEEE Xplore Press, Santa Clara, CA, USA, pp: 493-500. DOI: 10.1109/BigData.2015.7363791

Pezoulas, V.C., K. Kourou, K. Fanis and T.P. Exarchos, 2019. Medical data quality assessment: On the development of an automated framework for medical data curation. Comput. Biol. Med.
DOI: 10.1016/j.compbiomed.2019.03.001.

Poonsirivong, K. and C. Jittawiriyanukoon, 2017. A rapid anomaly detection technique for big data curation. Proceedings of the 14th International Joint Conference on Computer Science and Software Engineering, Jul. 12-14, IEEE Xplore Press, Nakhon Si Thammarat. DOI: 10.1109/JCSSE.2017.8025900

Rafailidis, D., A. Nanopoulos and E. Constantinou, 2014. With a little help from new friends: Boosting information cascades in social networks based on link injection. J. Syst. Software, 98: 1-8.
DOI: 10.1016/j.jss.2014.08.023

Ramsay, J.O., 2006. Principal Differential Analysis. In: Encyclopedia of Statistical Sciences, pp: 165- 28.

Reymondet, L., A.M. Ross and D.H. Rhodes, 2016. Considerations for model curation in model-centric systems engineering. Proceedings of the IEEE Systems Conference, Apr. 18-21, IEEE Xplore Press, Orlando, FL, USA.
DOI: 10.1109/syscon.2016.7490560.

Rogstadius, J., M. Vukovic, C.A. Teixeira V. Kostakos and E. Karapanos *et al*., 2013. Crisis tracker: Crowdsourced social media curation for disaster awareness. IBM J. Dev., 57: 41-413.
DOI: 10.1147/jrd.2013.2260692.

Salarian, M., A. Manavella and R. Ansari, 2015. Accurate localization in dense urban area using google street view images. Proceedings of the SAI Intelligent Systems Conference, Nov. 10-11, IEEE Xplore Press, London, UK.
DOI: 10.1109/intellisys.2015.7361184

Scaramozzino, J.M., M.L. Ramırez and K.J. McGaughey, 2014. A study of faculty data curation behaviors and attitudes at a teaching-centered university. Coll. Res. Libr., 73: 349-365.
DOI: 10.5860/crl-255

Sowe, S.K. and K. Zettsu, 2013. The architecture and design of a community-based cloud platform for curating big data. Proceedings of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Dec. 19-19, IEEE Xplore Press, Beijing, China, pp: 171-178. DOI: 10.1109/CyberC.2013.35

Sultanum, N., D. Singh, M. Brudno and F. Chevalier, 2018. Doccurate: A curation-based approach for clinical text visualization. IEEE Trans. Visualizat. Comput. Graph., 25: 1-1.
DOI: 10.1109/TVCG.2018.2864905

Tebbakh, A., 2014. Network of semantic wikis (Wicri) and data curation. Proceedings of the 4th International Symposium ISKO-Maghreb: Concepts and Tools for Knowledge Management, Nov. 9-10, IEEE Xplore Press, Algiers, Algeria, pp:
DOI: 10.1109/ISKO-Maghreb.2014.7033482

Tous, R., O. Wust, M. Gomez, J. Poveda and M. Elena *et al*., 2016. User-generated content curation with deep convolutional neural networks, Proceedings of the IEEE International Conference on Big Data, Dec. 5-8, IEEE Xplore Press, Washington, DC, USA, pp: 2535-2540. DOI: 10.1109/BigData.2016.7840893

Underwood, W., R. Marciano, S. Laib, C. Apgar and L. Beteta *et al*., 2017. Computational curation of a digitized record series of WWII Japanese-American internment. Proceeding of the International Conference on Big Data, Dec. 11-14, IEEE Xplore Press, Boston, MA, USA.
DOI: 10.1109bigdata.2017.8258184

Wollatz, L., M. Scott, S.J. Johnston, P.M. Lackie and S.J. Cox, 2018. Curation of image data for medical research. Proceedings of the 14th International IEEE eScience Conference, Oct. 29-Nov. 1, IEEE Xplore Press, Amsterdam, Netherlands. DOI: 10.1109/eScience.2018.00026

Yang, B.C., D. Puthal, S.P. Mohanty and E. Kougianos, 2017. Big-sensing-data curation for the cloud is coming: A promise of scalable cloud-data-center mitigation for next-generation IoT and wireless sensor networks. IEEE Consumer Electr. Magazine, 6: 48-56. DOI: 10.1109/MCE.2017.2714695

Yasumoto, K., H. Yamaguchi and H. Shigeno, 2016. Survey of real-time processing technologies of IoT data streams. J. Inf. Process., 24: 195-202. DOI: 10.2197/ipsjjip.24.195

Yoshioka, H. and K. Iida, 2016. Design and experimental verification of platform for the local music curation using IBeacon and apps. Proceedings of the International Seminar on Application for Technology of Information and Communication, Aug. 5-6, IEEE Xplore Press, Semarang, Indonesia. DOI: 10.1109 isemantic.2016.7873814.

Zhang, J., Y. Liu, H. Li and J. Zhao, 2016. Role definition of STI agencies in data curation. Proceedings of the International Conference on Progress in Informatics and Computing, Dec. 23-25, IEEE Xplore Press, Shanghai, China, pp: 518-523. DOI: 10.1109/PIC.2016.7949555