Big Data and Computational Intelligence: Background, Trends, Challenges, and Opportunities

Sukey Nakasima-López, Mauricio A. Sanchez and Juan R. Castro

Abstract The boom of technologies such as social media, mobile devices, internet of things, and so on, has generated enormous amounts of data that represent a tremendous challenge, since they come from different sources, different formats and are being generated in real time at an exponential speed which brings with it new necessities, opportunities, and many challenges both in the technical and analytical area. Some of the prevailing necessities lie on the development of computationally efficient algorithms that can extract value and knowledge from data and can manage the noise within in it. Computational intelligence can be seen as a key alternative to manage inaccuracies and extract value from Big Data, using fuzzy logic techniques for a better representation of the problem. And, if the concept of granular computing is also added, we will have new opportunities to decomposition of a complex data model into smaller, more defined, and meaningful granularity levels, therefore different perspectives could yield more manageable models. In this paper, two related subjects are covered, (1) the fundamentals and concepts of Big Data are described, and (2) an analysis of how computational intelligence techniques could bring benefits to this area is discussed.

1 Introduction

The emergence of the third industrial revolution was in mid-1990s, through internet technology, it brought with it a new and powerful structure that would change the world in communication and knowledge generation, with the arrival of new technologies and the transition to a new economy based on data, where it now represents a great importance, since it is considered an economic and social engine [1, 2].

Excessive data has been generated from technologies such as internet of things, social networks, mobile devices, among others. As Helbing [3] indicates, there will

S. Nakasima-López (✉) · M. A. Sanchez · J. R. Castro
Facultad de Ciencias Químicas e Ingeniería (FCQI), Universidad Autónoma de Baja California, 22390 Tijuana, Baja California, Mexico
e-mail: sukey.nakasima@uabc.edu.mx

M. A. Sanchez et al. (eds.), *Computer Science and Engineering—Theory and Applications*, Studies in Systems, Decision and Control 143,
https://doi.org/10.1007/978-3-319-74060-7_10

be more machines connected to the internet than human users. In accordance with International Data Corporation (IDC), in 2011 were created and copied 1.8 zettabytes (ZB) in the world, up to 2003, 5 exabytes (EB) data were created for humanity, today that amount of information is created in two days [4].

All these technologies forming a vital part in each of the activities we carry out and it assume more important roles in our lives [5]. As is the case that Rodríguez-Mazahua et al. [6] highlights about of how Google predicted in the health public sector the propagation of flue, based searches that users did, in terms of "flu symptoms" and "treatments of flu" in a couple of weeks before there would an increase in patients arriving in a certain region with flu. This search reveals a lot about the searchers: their wants, their needs, their concerns, extraordinarily valuable information.

This exchange of information that is being generated and stored at great velocity and in exponential quantities has never before been seen throughout history, as argued by Helbing [3], the necessity emerges to mine and refine data in order to extract useful information and knowledge and to be able to make more precise forecasts, where standard computational methods can not cope with.

At present, companies and industries are aware that in order to be competitive, data analysis must become a vital factor in discovering new ideas and delivering personalized services [7]. There are a lot of potentials and high value hidden in a huge data volume, that are demanding computing innovation technologies, techniques, and methodologies to model different phenomena, with extreme precision [8].

2 Evolution of Data Analysis

With the progressive evolution of informatization, we have gone from the necessity of only storage and data management to the possibility of value extraction from data. And it is on this way where data analysis and analytics have emerged and evolved.

In 1958, Hans Peter Luhn made the first reference of business intelligence at the field of business data analysis. But it was in 1980 when Howard Dresner consolidated the term, making references to a set of software to support the business decision making based on the collection of data and descriptive analysis, showing events that have already occurred and based on insight into the past and what is happening in the present. It is estimated that 80% of generated analytics results are descriptive, and are considered of low complexity and hindsight value [5, 9].

Subsequently the necessity for a predictive analysis was required to extract knowledge from data in the form of patterns, trends and models, and it was at the late of 1980s, when the expression of data mining emerged, whose origin is artificial intelligence, that is defined as the process of discovering patterns (that must be meaningful) in data and its focus is on predictive analysis, at the same time, the expression knowledge discovery in database (KDD) also begins to be used. Soon

these technique with machine learning would allow us build predictive models to determinate the outcome of an event that might occur in the future, this can lead to the identification of both risk and opportunities [5, 9, 10].

3 Emergence of Big Data

Given the arrival of ubiquitous technologies, the popularization of the world wide web and affordable personal computers, as well as other devices, we are facing a new phenomenon where the challenge is not only the storage and processing but also the analytics techniques and methodologies, that could cope to variables such as volume, velocity, and variety, the term that define these characteristics is known as Big Data.

The term was popularized in 2011 by IBM initiatives that invested in the analytics market. One of the first companies to face the problem of Big Data was Google when it had the necessary to classify its pages by quality, importance, and authority, analyzing that direct clicks and that coming from other intermediate links. For this reason, Google created PageRank algorithm, but this algorithm required running in a parallel environment to cope with the challenges of volume, velocity, and variety at the same time, for that it utilized MapReduce algorithm altogether [11, 12].

On the other hand, Gandomi and Haider [11] emphasize that information that has the basic dimensions of big data are an asset for the firms that would like to create real-time intelligence from data that are changing exponentially in time. All these necessities demand cost-effective, innovative forms of information processing for enhanced insight and decision-making, that traditional data management system is not capable of handling.

There is valuable information hidden in that sea of data that could be leveraged in making-decisions, originated for clickstream from users that reveal its behavior and browsing patterns, their needs, wishes, and worries. For all these reasons, we can define that Big Data as immense data size that include heterogeneous formats (structured, unstructured and semi-structured data), that cannot be handled and analyzed by traditional databases, because the generation speed and representation limits their capacity and require advanced techniques and powerful technologies to enable the capture, storage, distribution, management, and advanced algorithms to analysis information [7, 11, 13, 14].

Another definition that was made by Hashem et al. [15] say that Big Data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.

Big Data has been described for many dimensions, where each dimension explains its complexity.

Volume: it is a major feature but not the only one, the minimum data size is considered in terabytes (equivalent to 16 millions of pictures stored in Facebook servers) and petabytes (equivalent to 1024 terabytes), the future trend is that the volume is going to increase. Transnational corporations as walmart generated batches more than one million transactions per hour, the e-commerce, sensors and social media has contributed to the generation of this huge volume of data, is estimated that Facebook has stored 260 billions of photographs, using more than 20 petabytes of storage, also Google processed a hundred of petabytes in data, as well as the electronic store Alibaba that generates dozens of transactions in terabytes per day [7, 11, 13, 16].

The existence of more data allow us to create better models since we can discover many critical variables that can help us to describe in a better way the phenomena [14, 15].

Velocity: the speed in that data is generated and processed, such that it represents a huge challenge when analyzing data in real time and having the ability to get indicators at this rate that can be useful to the decision-makers. For example, customizing the daily offer according to profile and behavior of customers, this is possible thanks to the popularity and affordability of different devices which are present in all places, like as smartphone and other mobile devices that are constantly connected to the network and provide relevant and updated information that could be leveraged to create new business strategies, such as, geospatial location, demographic information and purchasing patterns [14, 15, 17].

The tendency of improving the ability of data transmissions will continue to accelerate the velocity. In 2016, it was estimated that 6.4 billion devices were connected to the internet, using and sharing information from many sectors. Also, 5.5 billion of new devices were added in this same year, and all of them were creating and exchanging data. It is expected that for 2020 the number of devices in use and connected will be about 20.8 billion [16].

Variety: it referrers to the heterogeneous dataset that is composed of structured data (e.g. relational database and spreadsheets, among others) and these represent about 5% of all existing data, unstructured (corresponds to images, audio, and video, among others), and the semi-structured (which do not conform to a strict standard such as XML, emails, social media, and the information recovery from different devices, among others). Requiring more computing power for its efficient handling [11, 14, 16].

Garner Inc, introduced these three dimensions that are the major characterization of Big Data and are known as 3V's, however, other V's have been added that complement the complex description of this phenomena, according to Gandomi and Haider [11] and Lee [16], they identified and described the follow V's:

Value: a feature introduced by Oracle, raw data has low-value density with respect to the volume, however, we could get high value from analyzing enormous amounts of data and transforming it into strategies for the organization. The most representative benefits that obtain value are increments in revenues, reduction of operational cost, improvement to customer services, among others.

Veracity: IBM coined this characteristic that represents the distrust and latent uncertainty on the source of data generated by incompleteness, vagueness, inconsistencies, subjectivities, and latency present in data. For instance, analyzing the sentiments of people with voting age about their perception, judgment, from all kinds of comments from candidates done through social networks.

Variability: a term introduced by SAS, it refers to the variability in the rate of data streaming because of its inconsistency by intermittency, or peaks of traffic, for periods of times. Also, it denotes the complexity by connecting, combining, cleaning, and converting data collected from many sources.

4 Big Data Value Chain

All activities that taking place in an organization are known as the value chain and they have the aim to deliver a product or service to the market. The categorization of generic activities that add value to the chain allows for the organization have a better and optimized understanding about what happened in each area. All this conceptualization can be applied in virtual value chain environments such as an analytics tools that will allow us understanding of the creation of value from data technology. Doing a mapping of each phase present in the flow of information, in big data its value chain can be seen as a model of high level activities that comprise an information system [18]. The activities identified as part of value chain of Big Data are [17, 18]:

Data acquisition: It is the first activity in the chain of value, and it is referring to the collection process, filtration and cleanness of data before to placing it in whatever storage solution exists and from there do different kinds of analyses. According to Lyko et al. [19] data acquisition in the context of Big Data is governed by 4V's, where it is assumed that: a high volume, variety, velocity and a low value from the initial stage, with the aims to increase the value from collected data.

Data curation: a continuous activity of data management, because operating in each phase from process lifecycle of data, to ensure the quality and maximizing its effective use, these features are relevant and represent a strong impact on the business operations, since they influence in the process of the making-decision of an organization [20]. In this process, we seek that the data be reliable, accessible, reusable and adequate to the purpose for they were created.

Data storage: according to Chen et al. [13] in the environment of Big Data is refers to the storage and management of a huge volume of data in scalable platforms, that in turn provide reliability, fault tolerance, and availability of the data for its subsequent access. As Strohbach et al. [21] points out, the ideal characteristics of a storage system in big data will be: virtual capacity of unlimited data, high rate of random access to writing and reading, flexible and efficient, manage different models of data, support to structured and not structured data, and that can work with encrypted data for major security.

Data analysis: collected data can be interpreted in a clear form to the decision-makers, according to Yaqoob et al. [22] techniques of big data are required to make for efficient the analytics of enormous data volumes in a limited time period.

Data use: All of those tools that allow us the integration of data analysis with the business activities that are required for decision-making, to provide the organizations the opportunities to be competitive through the reduction of costs, increasing value, monitoring for many parameters that are relevant to the good functioning of the organization, generation of predictions, simulations, visualizations, explorations, and data modeling in specific domains. The sectors where data analytics have been implemented and successful are manufacturing, energy, transport, health and so on, all of them are known as industry 4.0.

5 Challenge of Big Data

Due to the complex nature of the environment of big data, many challenges are present in each stage of the data lifecycle. Also, the development of new skills, updating or the replacement of more powerful IT technologies to obtain greater performance in the process that is required. Below are listed some of them [13, 23]:

Data complexity: it is related to the characteristics that describe big data, the diversification of types, structures, sources, semantics, organization, granularity, accessibility and complicated interrelations, make it difficult to represent, understand and process the data in this context. A good representation of data allows us to obtain greater meaning, and a bad representation reduces the value of data, impeding its effective analysis. Due to the constant generation of data from different source, collection, integration and data integrity, with optimized resources both hardware and software it has become in one of the major challenges [6, 7]. To ensures quality in data, it is necessary to establish control processes, such as metrics, data evaluation, erroneous data repair strategies and so on [16].

Process complexity: related to isolating noise contained in data from errors, faults or incomplete data, to guarantee its value and utility. Reducing redundancy and compressing data to enhance its value and make them easily manageable. As Sivarajah et al. [24] points out, some of the areas that present challenges in the process are in data aggregation and integration, modeling, analysis, and interpretation.

System complexity: when it is desired to design system architectures, computing frameworks, and processing systems, it is necessary to take into account the challenge of high complexity of big data, therefore increasing requirements of processing, due to its volume, structure, and dispersed value. And that it must support large work cycles and that their analysis and delivery of results must be in real time, having as main objective the operational efficiency and the energy consumption.

6 Areas of Application of Big Data

The sectors where big data has had application and a strong positive impact, and where has overcome the storage and analysis challenges, has been:

Internet of Things: it currently represents one of the major markets. Its devices and sensors produce large amounts of data and have the potential to generate trends and investigate the impact of certain events or decisions. The development and application has been given in intelligent buildings, cyber-physical systems, as well as traffic management systems [25].

Smart grid: big data analytics allows the identification of electrical grid transformers at risk, and the detection of abnormal behavior of connected devices. It allows establishing preventive strategies with the purpose of reducing costs by correction, as well as more approximate forecasts of the demand that allow to make a better balance of the energy loads [7].

Airlines: employ hundreds of sensors on each aircraft to generate data about their entire aircraft fleet, its objective is to monitor their performance and apply preventive maintenance resulting in significant savings for the company [26].

E-health: used to customize health services, doctors can monitor the symptoms of patients in order to adjust their prescriptions. Useful for optimizing hospital administrative operations and reducing costs. One of these solutions is offered by CISCO [7].

Services: the tools of big data allow to analyze the current behavior of clients, to cross the information with historical data and their preferences, in order to offer a more effective service and to improve their marketing strategies, such as Disneyland park, who have introduced a bracelet equipped with radio frequency, which allows visitors to avoid waiting in lines and book rides, this allows to create a better experience for the visitor, attracts many more customers and increases their income [26].

Public uses: used in complex water supply systems, to monitor its flow and detect leaks in real time, illegal connections and control of valves for a more equitable supply in different parts of the city. As is the case of Dublin city council where one of the most important services is the transport and for that purpose it has equipped its buses with GPS sensors to collect geospatial data in real time and with this, through its analysis, it can optimize their routes and use of their transport, allowing fuel savings and decrease the level of pollution in the air emitted by the transport system [7, 26].

7 Computational Intelligence

Due to the increase in the complexity surrounding the data, since they are being generated in an excessive way and in very short time periods, it requires both powerful computing technology, as well as robust algorithms from which we can

extract knowledge and value. In this context, a solution that can encompass the representative characteristics of the big data phenomenon is computational intelligence (CI).

According to Hill [27], CI focuses on replicating human behavior more than on the mechanisms that generate such behavior, through computation. The algorithms based on CI, allows modeling of human experience in specific domains in order to provide them with the capacity to learn and then adapt to new situations or changing environments [28–30]. Some of the behaviors it includes are; abstraction, hierarchies, aggregation of data for the construction of new conclusions, conceptual information, representation and learning from symbols. The use of CI requires a focus on problems rather than on technological development [27].

Bio-inspired algorithms are increasingly used to work and give solutions to problems with a high level of complexity, since they are intelligent algorithms that can learn and adapt as would biological organisms, these algorithms have the characteristic that they can be tolerant to incomplete, imprecise, and implicit uncertain data. They can also increase the range of potential solutions with better approximation, manageability and robustness at the same time [29–31]. Some of the technologies often associated with CI are described below:

Artificial Neural Networks (ANN): a discipline that tries to imitate the processes of learning of the brain, replicating the inaccurate interpretation of information obtained from the senses taking advantage of the benefits of the fast processing offered by computer technology. The first step toward artificial intelligence came from neurophysiologist Warren McCulloch and the mathematician Walter Pitts that in 1943 wrote a paper about how the neurons work and they established the precedent of creation of a computational model to neural network [32]. ANN are also defined as adaptive algorithms of non-linear and self-organized processing, with multiple processing units connected known as neurons, in a network with different layers. They have the ability to learn based on their inputs and adapted according to the feedback obtained from their environment [31].

Different neural network architectures have been developed, some of them are [29]:

- Hopfield network, it is the simplest of all, because it is a neuron network with a single layer. This model was proposed in 1982 by John Hopfield [33].
- Feedforward multilayer network, executes its passage forward, has an input layer, another layer for output, and in an intermediate way has the known hidden layers in which can be defined N number of them. This design was made by Broomhead and Lowe in 1988 [34].
- Self-organized networks, such as Kohonen's self-organizing feature maps and the learning vector quantizer. A paper by Kohonen was published in 1982 [34].
- Supervised and unsupervised, some networks with radial basis functions.

Its main advantages are in the handling of noise in data and good control, achieving low error rates. Neural network methods are used for classification, clustering, mining, prediction, and pattern recognition. They are broadly divided into three types, of which they are recognized; feedforward network, feedback network and

self-organized networks. The characteristics of artificial neural networks are; distributed information storage, parallel processing, reasoning and self-organized learning, and have the ability to adjust non-linear data quickly. Neural networks are trained and not programmed, are easy to adapt to new problems and can infer relationships not recognized by programmers [35].

Genetic Algorithms (GA): inspired by the principles of genetics and natural selection, in order to optimize a problem, were first proposed by John Holland in 1960 [36]. Through a cost function, it tries to find the optimal value either in maximum or minimum of a given set of parameters [31]. These types of algorithms have succeeded in optimizing search systems that are difficult to quantify, such as in financial applications, industrial sector, climatology, biomedical engineering, control, game theory, electronic design, automated manufacturing, data mining, combinatorial optimization, fault diagnosis, classification, scheduling, and time series approximation [29].

In big data, GA have been applied to generate clustering, in order to have a better management of data volume, dividing data into small groups that are considered its population. Also, one of its great benefits is that they are highly parallelizable. It can be combined with K-means algorithms (created in 1957 by Stuart Lloyd [37]), the combination of GK-means will take less memory and process large volumes of data in less time and achieve very good results [38].

Fuzzy logic: all human activities have implicit uncertainty, our understanding is largely based on imprecise human reasoning, and this imprecision could be useful to humans when they must make decisions, but in turn are complex processing for computers. Fuzzy logic is a method to formalize the human capacity of imprecise reasoning. It is partial or approximate, assigning to a fuzzy set degree of truth in a set between 0 and 1 [39].

In the context of big data have the ability to handle various type of uncertainties that are present at each phase of big data processing. Also, fuzzy logic techniques with other Granular Computer techniques can be employed to the problem which can be reconstructed to a certain granular level. It would be more efficient if they are associated with other decision-making techniques, such as probability, rough sets, neural networks, among others [17]. Some of the applications of fuzzy systems have been given in; control systems, vehicle braking system, elevator control, household appliances, traffic signal control, so on [29].

Its relevance in a big data environment lies in its ability to provide a better representation of the problem through the use of linguistic variables, which facilitates the handling of volume and variety when datasets are growing exponentially and dynamically. In addition, experts can benefit from the ease of interpreting results associated with these linguistic variables [40].

Other applications are in intelligent hybrid systems, where the benefits of both fuzzy systems and neural networks are combined, enhancing the ability of the latter to discover through learning the parameters needed to process data. Being these; fuzzy sets, membership functions, and fuzzy rules. It has also been proposed to integrate to these hybrid intelligent systems, genetic algorithms to optimize

parameters, adjust control points of membership functions and fine tune their fuzzy weights [41].

Granular Computing: computing paradigm for information processing, tries to imitate the way in which humans process information obtained from their environment in order to understand a problem. Can be modeled with principles of fuzzy sets, rough sets, computation with words, neural networks, interval analysis, among others. In 1979, Zadeh introduced the notion of information granulation and suggested that fuzzy set theory might find possible applications with respect to this. Its powerful tools are vital for managing and understanding the complexity of big data, allowing multiple views for data analysis, from multiple levels of granularity [42, 43].

Granules may be represented by subsets, classes, objects, clusters, and elements of a universe. These sets are constructed from their distinctions, similarities, or functionalities. It has become one of the fastest growing information paradigms in the fields of computational intelligence and human-centered systems [17].

Fuzzy logic techniques together with granular computing concepts are considered one of the best options for the process of decision-making. A fuzzy granule is defined by generalized constraints, said granules can be represented by natural language words. The fuzzification of its granules together with its values that characterize it, is the way in which human constructs its concepts, organizes, and manipulates them. They can be used to reconstruct problems with a certain level of granularity (from the finest, which could be at the level of an individual, to the coarse granules that could be at community level), the objective would be to focus on the volume, feature of big data, reducing its size and creating different perspectives that can later be analyzed and become indicators of relevance for decision-making [17, 42, 43].

Granular computing has represented an alternative solution to obtaining utility and value of big data in spite of its complexity. Because of their integration with computational intelligence theories, it can help effectively support all the operational levels that include; acquisition, extraction, cleaning, integration, modeling, analysis, interpretation, and development [42].

Machine Learning: branch of artificial intelligence, which focuses on the theory, performance, and properties of algorithms and learning systems. It is an interdisciplinary field which is related to artificial intelligence, optimization theory, information theory, statistics, cognitive science, optimal control and other disciplines of science, engineering and mathematics. Its field is divided into three subdomains, which are [44]:

- **Supervised learning**: requires training from input data and the desired output. Some of the tasks performed in data processing are classification, regression, and estimation. Of the representative algorithms in this area are: support vector machine that was proposed by Vladimir Vapnik in 1982 [45], hidden Markov model was proposed in 1966 by Baum and Petrie [46], Naives Bayes [47], bayesian network [48], among others.

- **Unsupervised learning**: only requires input data, without indicating the desired objective. The tasks of data processing that it performs are clustering and prediction, of which some of existing algorithms are: Gaussian mixture model that was created by Karl Pearson's in 1984 [49], X-means [50], among others.
- **Reinforced learning**: allows learning from the feedback received through the interaction obtained from an external environment. They are oriented to decision making and some algorithms are: Q-learning that was introduced by Zdzislaw Pawlak in 1981 [51], TD learning proposed by R. D. Sutton in 1988 [52], and Sarsa learning that was proposed by Sutton and Barton in 1998 [53].

The application of machine learning can be carried out through three phases [54]:

- **Preprocessing**: helps to prepare raw data which by its nature consists of the unstructured, incomplete, inconsistent and noisy data, through cleaning of data, extraction, transformation, and fusion of data which can be used in the learning stage as input data.
- **Learning**: uses learning algorithms to fine-tune the model parameters and generate desired data outputs from pre-processed input data.
- **Evaluation**: the performance of learning models is determined here, characteristics to which special attention is given are: performance measurement, dataset selection, error estimation, and statistical testing. This will allow you to adjust the model parameters.

The objective of machine learning is to be able to discover knowledge and to serve the decision makers to be able to generate intelligent decisions. In real life, they have applications in search engines for recommendation, recognition systems, data mining for discovery of patterns and extraction of value, autonomous control systems, among others [7]. Google uses machine learning algorithms for large volumes of disordered data.

8 Conclusions

The Big Data phenomenon confronts us with great opportunities, but also great challenges in order to achieve its benefits. Its complex characteristics invite us to rethink the ways data is managed, processed and analyzed for value, quality and relevance. It is necessary to overcome the superficial analysis that only describes a historical event and evolving to deeply analyses that allow us creating predictions and prescription that support actions and strategies to improve the decision-making in the organizations.

Traditional models and technologies cannot cope with this effectively, so it is necessary to design and develop new technologies, with greater computing power, advanced algorithms, new techniques and methodologies that serve as support to be able to have greater control of the volume, variety, and speed that are part of the complex nature of big data.

For this reason, we believe that the design and development of algorithms based on computational intelligence can be a very good way to face these challenges that characterizes Big Data by its complex nature, for it will be necessary to adapt them to the existing platforms and make many experiments, to demonstrate that they are really efficient, effective, and help us to discover new patterns, ideas and knowledge in this context as well.

References

1. Brynjolfsson E, Kahin B (2000) Understanding the digital economy: data, tools and research. Massachusetts Institute of Technology
2. Rifkin J (2011) The third industrial revolution: how lateral power is transforming energy, the economy, and the world
3. Helbing D (2015) Thinking ahead—essays on big data, digital revolution, and participatory market society
4. Akoka J, Comyn-Wattaiau I, Laoufi N (2017) Research on big data—a systematic mapping study. Comput Stand Interfaces 54:105–115
5. Thomson JR (2015) High integrity systems and safety management in hazardous industries
6. Rodríguez-Mazahua L, Rodríguez-Enríquez CA, Sánchez-Cervantes JL, Cervantes J, García-Alcaraz JL, Alor-Hernández G (2016) A general perspective of big data: applications, tools, challenges and trends. J Supercomput 72(8):3073–3113
7. Oussous A, Benjelloun FZ, Ait Lahcen A, Belfkih S (2017) Big data technologies: a survey. J King Saud Univ Comput Inf Sci
8. McKinsey & Company (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute, p 156
9. Niño M, Illarramendi A (2015) Entendiendo el Big Data: antecedentes, origen y desarrollo posterior. DYNA NEW Technol 2(3), p [8 p]–[8]
10. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, vol 2
11. Gandomi A, Haider M (2015) Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manage 35(2):137–144
12. Srilekha M (2015) Page rank algorithm in map reducing for big data. Int J Conceptions Comput Inf Technol 3(1):3–5
13. Chen M, Mao S, Liu Y (2014) Big data: a survey. Mobile Netw Appl 19(2):171–209
14. Kacfah Emani C, Cullot N, Nicolle C (2015) Understandable big data: a survey. Comput Sci Rev 17:70–81
15. Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Ullah Khan S (2015) The rise of 'big data' on cloud computing: review and open research issues. Inf Syst 47:98–115
16. Lee I (2017) Big data: dimensions, evolution, impacts, and challenges. Bus Horiz 60(3):293–303
17. Wang H, Xu Z, Pedrycz W (2017) An overview on the roles of fuzzy set techniques in big data processing: trends, challenges and opportunities. Knowl Based Syst 118:15–30
18. Curry E (2016) The big data value chain: definitions, concepts, and theoretical approaches. In: New horizons for a data-driven economy: a roadmap for usage and exploitation of big data in Europe, pp 29–37
19. Lyko K, Nitzschke M, Ngomo A-CN (2016) Big data acquisition
20. Freitas A, Curry E (2016) Big data curation
21. Strohbach M, Daubert J, Ravkin H, Lischka M (2016) Big data storage. In: New horizons for a data-driven economy, pp 119–141

22. Yaqoob I et al (2016) Big data: from beginning to future. Int J Inf Manage 36(6):1231–1247 Pergamon
23. Jin X, Wah BW, Cheng X, Wang Y (2015) Significance and challenges of big data research. Big Data Res 2(2):59–64
24. Sivarajah U, Kamal MM, Irani Z, Weerakkody V (2017) Critical analysis of big data challenges and analytical methods. J Bus Res 70:263–286
25. Ahmed E et al (2017) The role of big data analytics in internet of things. Comput Netw
26. Alharthi A, Krotov V, Bowman M (2017) Addressing barriers to big data. Bus Horiz 60 (3):285–292
27. Hill R (2010) Computational intelligence and emerging data technologies. In: Proceedings—2nd international conference on intelligent networking and collaborative systems, INCOS 2010, pp 449–454
28. Jang J, E M, Sun CT (1997) Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence. Autom Control IEEE 42(10):1482–1484
29. Engelbrecht AP (2007) Computational intelligence: an introduction, 2nd edn
30. Kruse R, Borgelt C, Klawonn F, Moewes C, Steinbrecher M, Held P (2013) Computational intelligence. Springer, Berlin
31. Kar AK (2016) Bio inspired computing—a review of algorithms and scope of applications. Expert Syst Appl 59:20–32
32. Kumar EP, Sharma EP (2014) Artificial neural networks—a study. Int J Emerg Eng Res Technol 2(2):143–148
33. Elmetwally MM, Aal FA, Awad ML, Omran S (2008) A hopfield neural network approach for integrated transmission network expansion planning. J Appl Sci Res 4(11):1387–1394
34. Negnevitsky M (2005) Artificial intelligence: a guide to intelligent systems. In: Artificial intelligence: a guide to intelligent systems. Pearson Education, pp 87–113
35. Biryulev C, Yakymiv Y, Selemonavichus A (2010) Research of ANN usage in data mining and semantic integration. In: MEMSTECH'2010
36. Mitchell M (1995) Genetic algorithms: an overview. Complexity 1(1):31–39
37. Govind Maheswaran JJ, Jayarajan P, Johnes J (2013) K-means clustering algorithms: a comparative study
38. Jain S (2017) Mining big data using genetic algorithm. Int Res J Eng Technol 4(7):743–747
39. Ross TJ et al (2004) Fuzzy logic with engineering applications. IEEE Trans Inf Theory 58 (3):1–19
40. Fernández A, Carmona CJ, del Jesus MJ, Herrera F (2016) A view on fuzzy systems for big data: progress and opportunities. Int J Comput Intell Syst 9:69–80
41. Almejalli K, Dahal K, Hossain A (2007) GA-based learning algorithms to identify fuzzy rules for fuzzy neural networks. In: Proceedings of the 7th international conference on intelligent systems design and applications, ISDA 2007, pp 289–294
42. Pal SK, Meher SK, Skowron A (2015) Data science, big data and granular mining. Pattern Recogn Lett 67:109–112
43. Yao Y (2008) Human-inspired granular computing 2. Granular computing as human-inspired problem solving, No. 1972, pp 401–410
44. Qiu J, Wu Q, Ding G, Xu Y, Feng S (2016) A survey of machine learning for big data processing. EURASIP J Adv Sign Process 2016(1):67
45. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
46. Baum LE, Petrie T (1966) Statistical inference for probabilistic functions of finite state Markov chains. Ann Math Stat 37(6):1554–1563
47. Rish I (2001) An empirical study of the Naïve Bayes classifier. IJCAI 2001 Work Empir Meth Artif Intell 3
48. Zarikas V, Papageorgiou E, Regner P (2015) Bayesian network construction using a fuzzy rule based approach for medical decision support. Expert Syst 32:344–369
49. Erar B (2011) Mixture model cluster analysis under different covariance structures using information complexity

50. Pelleg D, Pelleg D, Moore AW, Moore AW (2000) X-means: extending K-means with efficient estimation of the number of clusters. In: Proceedings of the seventeenth international conference on machine learning, pp 727–734
51. Pandey D, Pandey P (2010) Approximate Q-learning: an introduction. In: 2010 second international conference on machine learning and computing, pp 317–320
52. Desai S, Joshi K, Desai B (2016) Survey on reinforcement learning techniques. Int J Sci Res Publ 6(2):179–2250
53. Abramson M, Wechsler H (2001) Competitive reinforcement learning for combinatorial problems. In: Proceedings of the international joint conference on neural networks IJCNN'01, vol 4, pp 2333–2338
54. Zhou L, Pan S, Wang J, Vasilakos AV (2017) Machine learning on big data: opportunities and challenges. Neurocomputing 237:350–361