Neural Computing and Applications
https://doi.org/10.1007/s00521-021-06332-9

S.I. : DATA FUSION IN THE ERA OF DATA SCIENCE

Bio-inspired computation for big data fusion, storage, processing, learning and visualization: state of the art and future directions

Ana I. Torre-Bastida[1] · Josu Díaz-de-Arcaya[1] · Eneko Osaba[1] · Khan Muhammad[2] · David Camacho[3] · Javier Del Ser[4]

Received: 4 July 2020 / Accepted: 12 July 2021

Abstract

This overview gravitates on research achievements that have recently emerged from the confluence between Big Data technologies and bio-inspired computation. A manifold of reasons can be identified for the profitable synergy between these two paradigms, all rooted on the adaptability, intelligence and robustness that biologically inspired principles can provide to technologies aimed to manage, retrieve, fuse and process Big Data efficiently. We delve into this research field by first analyzing in depth the existing literature, with a focus on advances reported in the last few years. This prior literature analysis is complemented by an identification of the new trends and open challenges in Big Data that remain unsolved to date, and that can be effectively addressed by bio-inspired algorithms. As a second contribution, this work elaborates on how bio-inspired algorithms need to be adapted for their use in a Big Data context, in which data fusion becomes crucial as a previous step to allow processing and mining several and potentially heterogeneous data sources. This analysis allows exploring and comparing the scope and efficiency of existing approaches across different problems and domains, with the purpose of identifying new potential applications and research niches. Finally, this survey highlights open issues that remain unsolved to date in this research avenue, alongside a prescription of recommendations for future research.

Keywords Big data · Bio-inspired computation · Data fusion · Evolutionary computation · Swarm intelligence · Neural networks · Fuzzy logic

1 Introduction

Nowadays, the computational complexity of processes and decisions held on a daily basis depend on the availability of high-quality data, which often holds in practice thanks to the massive digitization of traditional activity sectors. Unfortunately, such information is often produced at rates never seen before and in a non-structured fashion, outstripping the scales at which it was collected and mined by traditional data management systems. This situation eventually originated the so-called *Big Data paradigm*, which refers to the collection, analysis and visualization of data at scales that surpass the capacities of traditional infrastructures for information storage and processing. The core concept of Big Data is the derivation of alternative and efficient computing means to ingest, retrieve, process and visualize large amounts of data [1, 2]. Actually, Internet of Things (IoT) and Cloud Computing are standard bearers of the current digitization process that is conducted in different sectors, as they support the connectivity and management of devices in charge of data gathering, delivery, processing, and computation under different architectural strategies. All in all, data play a paramount role in both

✉ Ana I. Torre-Bastida
isabel.torre@tecnalia.com

✉ Khan Muhammad
khan.muhammad@ieee.org

[1] TECNALIA, Basque Research and Technology Alliance (BRTA), 48160 Derio, Spain

[2] Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), Department of Software, Sejong University, Seoul 143-747, Republic of Korea

[3] Universidad Politécnica de Madrid, 28031 Madrid, Spain

[4] University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain

Published online: 03 August 2021

paradigms, the difference being the imposed requirements and specifications (e.g., processing latency or transmission bandwidth).

In this context, notable milestones in the past (e.g., Map-Reduce programming, complex event processing or NoSQL databases) have led to a relatively high degree of maturity of Big Data technologies. However, algorithms for information fusion, processing and data mining have not gone on a par with the aforementioned technologies. Indeed, only a fraction of classical approaches for drawing knowledge from data have been adapted to the new requirements and computing procedures brought by Big Data technologies. Although adaptations for these approaches keep growing at a continuous pace, many of them still remain unaddressed. The complexity, heterogeneity, dynamism and inherently distributed nature of Big Data technologies do not help either for this purpose. Even models enjoying a straightforward adaptability to Big Data computing environments (e.g., ensembles for predictive modeling) can be severely affected by the obsolescence of the information from where they are learned [3], or the failure of a node in a distributed Map-Reduce computing grid [4]. All in all, data fusion, processing, learning and visualization of Big Data require a major focus not only on tailoring the algorithmic steps underlying each model/technique to the computing technologies underneath, but also endowing them with higher levels of resilience against failures, adaptation to changes in data and the accommodation of unprecedented levels of data volume, heterogeneity and veracity. In short: coupling algorithmic adaptation with systems' adaptation.

In light of the above, Big Data environments call for computationally efficient techniques that meet such requirements by embracing self-learning and adaptation capabilities at the core of their design. This unchains a magnificent opportunity for bio-inspired computation, which has gained a remarkable momentum in the Big Data literature. Inspired by intelligent behavioral patterns observed in nature, many practitioners in the scientific community have emulated such bio-inspired processes in the form of computational algorithms, aiming at harnessing the adaptability and self-learning capabilities of such biological systems to face complex problems [5]. Consequently, an upsurge of inspirational sources has been historically considered for the design and development of bio-inspired methods for different computational problems. Some examples of this claim for optimization problems are the behavioral patterns of animals [6, 7], genetic inheritance mechanisms [8] or physical phenomena [9], among many others. In regards to modeling, connections among neurons in the brain have stimulated a flurry of neural network approaches, arriving at the current myriad of Deep Learning models, all sharing a similar bio-inspired rationale [10].

Bio-inspired computation can provide promising solutions for the acknowledged drawbacks of Big Data processing in IoT and Cloud Computing environments, such as poor scalability, security issues, task distribution, fault tolerance, or low performance in traditional information technology frameworks. New optimization, scaling and management approaches can largely be benefited from the adaptability of bio-inspired methods, even further when considering the different dimensions of Big Data (volume, variety, velocity, veracity and variability), which increase the complexity of the problems to be solved. Fortunately, the synergy among Big Data and bio-inspired computation is clear and meaningful. On the one hand, bio-inspired computation can act as a beacon for attaining near-optimal solutions for complex modeling and optimization problems that can be present in the Big Data paradigm. For instance, bio-inspired heuristic methods for optimization can efficiently accommodate the dynamic nature of objectives and constraints of an optimization problem characterizing the load balancing in a cloud computing grid [11]. Fuzzy logic can help accounting for the uncertainty of Big Data decision making, mostly when data sources are unreliable or the decision is held in a context subject to exogenous and non-considered factors [12]. The benefits resulting from this synergistic relationship are exposed by new Big Data infrastructures, tools and technologies that have adopted bio-inspired algorithms to reach a higher level of efficiency in their tasks. Some few examples of technologies that take advantage of the capabilities of bio-inspired algorithms are, among many others, NoSQL databases [13–15], load planners/schedulers [16], or tools assisting analytical tasks such as feature selection [17], dimensionality reduction [18] or data fusion [19]. On the other hand, through bio-inspired computation perspective, Big Data provides the possibility of great volumes and varieties of data and the efficient implementation of solvers through new technologies, which offer parallel, distributable and scalable workloads. In this context, there are numerous studies and surveys focused on Big Data analytics [20]. All evidences confirm that efforts conducted in this topic are growing lately, which calls for a reference material to organize achievements so far, and connect them with a prospect of valuable research directions.

The goal of this survey is to answer this call by enumerating and thoroughly examining the principal points of connection between Big Data technologies and bio-inspired computation. To this end, we undertake several interconnected tasks, all departing from a critical assessment of the recent literature:

- First, we review the main concepts related to Big Data and bio-inspired computation, settling common grounds for an adequate understanding of our study.
- We examine contributed works where Big Data infrastructure, tools and technologies have been improved through bio-inspired computation approaches.
- We exhaustively review how bio-inspired algorithms have enhanced the Big Data domain, classifying them into different steps of the Big Data life cycle (i.e., data fusion, processing, learning and visualization).
- We explore and compare to each other the specific scope of problems tackled so far by the community, identifying further applications that can be addressed in the future.
- Finally, we provide our envisioned future for this research in the form of a prospect of challenges, trends and research directions that can be pursued for stepping further in this research topic.

This work is structured in the following way: In Sect. 2 we present in detail both Big data and bio-inspired computing concepts. Section 3 delves into the synergies between these two paradigms, providing a taxonomy to classify advances reported so far and a critical review of the existing literature. Next, we introduce current challenges and open opportunities in 4. Section 5 ends the survey by summarizing the main conclusions and by providing an outlook towards the future of this exciting field.

2 Big data and bio-inspired computation: first concepts

As has been anticipated in the introduction, this section first defines concepts underneath Big Data (Sect. 2.1) and bio-inspired computation (Sect. 2.2). On the one hand, we focus on the Big Data life cycle phases, along with their associated technologies. On the other hand, we classify bio-inspired algorithms as per the kind of problems they can solve, as well as their biological source of inspiration. This allows detecting which bio-inspired algorithms have demonstrated a better off-the-shelf applicability to large data volumes, or have been specifically designed for such a purpose.

2.1 Big data paradigm

Briefly explained, Big Data is a concept that encloses large volumes of high-speed, complex, variable and heterogeneous data, along with advanced technologies and techniques that enable their collection, storage, processing/analysis and visualization. This specific definition expands the one provided by Gartner in [1]. In this subsection, we first discuss the relationships among Big Data and bio-inspired computation, which have stimulated the research that has hitherto been made in this field. We next describe in detail the Big Data life cycle, which is of capital importance for properly understanding the investigation carried out in this area and the subsequent analysis of the literature.

2.1.1 Big data dimensions and bio-inspired computation

There is a clear consensus within the community that Big Data relies on five different main features of data: volume, velocity, variety, variability and veracity [21]. All these characteristics are critical and define the way data is managed across the environment, which can be defined as follows [22]: (1) *volume* represents the magnitude of the data in terms of size; (2) *velocity* refers to the speed at which the information is produced, received and processed; (3) *variety* is related to the heterogeneity of data produced by different domains; (4) *variability* refers to changes in non-stationarity events that affect data, which require accommodating their effects on the system and/or models over time; and (5) *veracity* deals with the provenance and reliability of the collected information. These five dimensions are cross-domain, and unless properly resolved, can hinder the adoption of data-based operational workflows in a diversity of applications.

Fortunately, bio-inspired computation can effectively help legacy technologies to cope with challenges stemming from the above features. In terms of *volume*, for example, bio-inspired optimization metaheuristics can contribute to the feasibility of traditional data mining models for large datasets under assorted strategies, including instance reduction, feature selection, or model simplification [23]. Indeed the compliance of the optimization problems formulated in these strategies with the typical volumes of Big Data is among the motivations for the upsurge of large-scale global optimization, a subarea within bio-inspired optimization that deals with problems of very high dimensionality (thousands to millions of decision variables [24]). Bio-inspired solvers have also been proven to excel at data integration, aggregation and fusion [25, 26], outstanding as essential drivers to deal with the *variety* and *variability* dimensions of Big data. Lastly, *velocity* and *veracity* dimensions affect data and service quality, as well as monitoring and security problems. Examples of bio-inspired optimization algorithms dealing with these issues can be found in [27, 28], whereas elements from fuzzy logic have also been utilized in Big Data environments subject to data uncertainty (see [29, 30] and references in the comprehensive overview in [31]).

Figure 1 summarizes graphically each of the five dimensions of Big Data described above, as well as

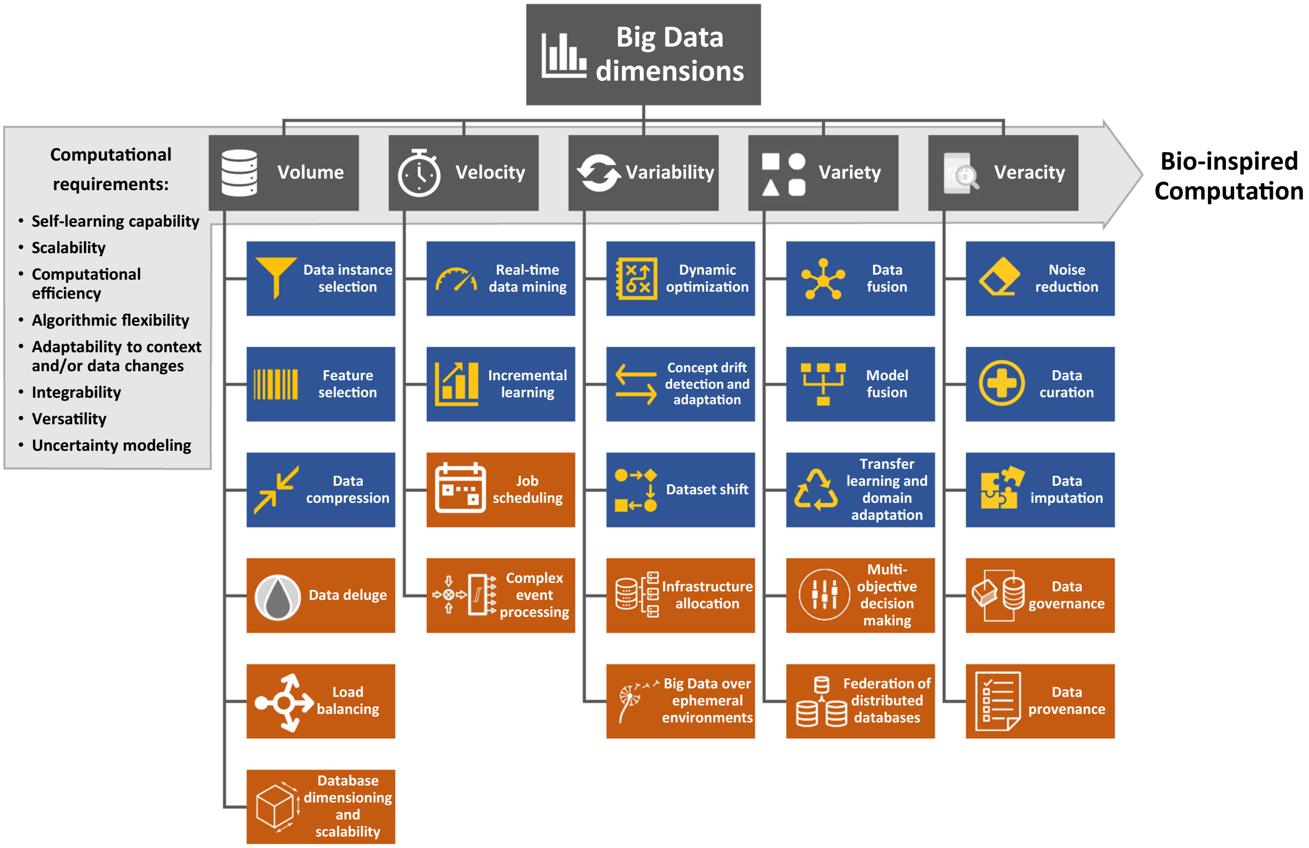


Fig. 1 Big Data dimensions associated with typical problems liable to be solved by bio-inspired computation. Nodes colored in blue correspond to computational tasks, whereas those colored in light brown indicate specific applications where the Big Data dimension indicated in their parent nodes are particularly relevant. Computational requirements enabled by bio-inspired computation are indicated in the gray box set on the background

problems typically arising from each of them. Along with this information, we include citations to several landmark reviews gravitating on how bio-inspired computation has managed to overcome the barriers imposed by the Big Data paradigm.

2.1.2 Big data life cycle

A logical line of thinking springing from the aforementioned dimensions is that Big Data requires highly adaptive techniques to efficiently process large quantities of data within tolerable computational times. Following [32], three are the questions that must be formulated in regard to the management and treatment of data: (1) Is it technologically affordable to capture and store all data? (2) is it possible to clean, enrich, and analyze the data? and (3) is it possible to retrieve, search, integrate, and visualize the data?. Answering these three questions (which can be summarized as the *store-process-manage* triplet) is essential for extracting valuable insights from data in practical use cases.

Considering these three technological concerns, a common way to orchestrate the heterogeneity of technologies under the Big Data paradigm is around the Big Data life cycle, which comprises data storage, data fusion, data learning, searching, sharing, transferring, visualization, querying, updating and information privacy. Among these new areas, the ones that best fit with the main philosophy of bio-inspired computation, and those in which solutions of greater value can be provided, are the following:

– *Data Fusion*: This phase represents the process of merging multiple data sources, towards producing consistent, accurate, and useful information. Data fusion is clearly related to the *variety* feature of Big Data, and its complexity stems from the large volumes of data that must be fused. In this sense, bio-inspired algorithms inherently provide great benefits for this purpose, with an increasing prevalence of model-based data fusion based on Deep Learning neural network models. In fact, the main concept of data fusion originates from the human and animal ability to incorporate information from multiple senses to improve their monitoring capabilities. This being said, the design flexibility and unified learning framework that current Deep Learning models provide is currently one of the enablers of the so-called model-based data fusion. Indeed, the fact that hierarchical features can be nowadays learned from image, video, text, and other forms of data in the space and/or sequential (time) domains permit to learn them together by assembling neural parts devoted to each domain. In these areas, information sharing is realized through the exchange and sharing of parts of the neural networks, which are trained together for the task at hand. Therefore, Deep Learning methods can effectively implement Data Fusion by implementing multi-modal feature extraction over a mixture of neural units specialized for sequential (e.g., LSTM or GRU cells) and space domains (convolutional filters). Once assembled, the training process of the overall neural network (gradient backprop) tunes the parameters of these units for them to learn what features to extract and fuse for solving the task at hand. Emerging learning paradigms such as transfer learning, domain adaptation and multitask learning are also largely harnessing the possibilities brought by neural computation [33].

- *Data Storage*: This stage refers to the need for effective repositories capable of storing and efficiently managing huge volumes of data. This process poses a remarkable challenge in terms of distribution, scalability and performance. Some additional problems to face in this regard are the concurrency and consensus derived from writing and accessing data in the repositories. In this context, bio-inspired algorithms are appropriate for this purpose, since most of them consider distribution and parallelism intrinsically in their design. Furthermore, data reduction has also leveraged bio-inspired computation in a number of representative works [34, 35].
- *Data Processing*: This phase regards the proper processing of all the merged and stored data. In this sense, any technique developed for this purpose must accommodate the great amount of information available in Big Data context and the rate at which it is produced. Once again, the inherent parallelism of bio-inspired methods makes them promising alternatives for managing the distribution of large volumes of the data, particularly in what refers to feature selection, instance filtering and data imputation, as well as in streaming environments [36]. Likewise, a large fraction of data in the context of Big Data is composed of images/videos. Consequently, image prioritization/video summarization technologies are key stakeholders to contribute to data reduction.
- *Data Learning*: This step regards all processes aimed at retrieving relevant knowledge from the available Big Data. At this point we stress on the paramount relevance that bio-inspired computation has held in data mining, with a plethora of studies exhaustively reviewing the activity in this confluence of technologies over the years. However, the interest in extrapolating these prior achievements around bio-inspired data mining to the scales, speeds and variety of Big Data has not fully exploded to reach its potential. Obviously, neural computation relies extensively on the biological mechanisms inside the human brain. Modern variants such as convolutional neural networks hinge on how the visual cortex operates when fed with an image. Modern neural computation, collectively referred to as Deep Learning, can be conceived as a family of bio-inspired computation techniques by themselves that require heavy loads of data for learning their constituent parameters. However, as we will show later, the possibilities of bio-inspired computation span far beyond the biological principle of models and algorithms currently utilized.
- *Data Visualization*: Once the learning model has produced an insight from Big Data, this phase undertakes the visualization of large volumes of data and information, coupled with the added knowledge extracted by the learning models in use. Visualization is actually a challenge that has not yet been as addressed as other phases of the life cycle, possibly due to the strong link between Artificial Intelligence, computer graphics and cognitive sciences [37, 38].

With all this, Fig. 2 showcases the described five phases of the data life cycle, which are used to convert simple and raw data into valuable knowledge. Through the conduction of these steps, the sixth and last dimension of the Big Data is attained: *value*.

From the technological point of view, these phases need to be efficiently implemented using suitable tools and mechanisms. Techniques and technologies involved in this process are jointly integrated into a single system, forging

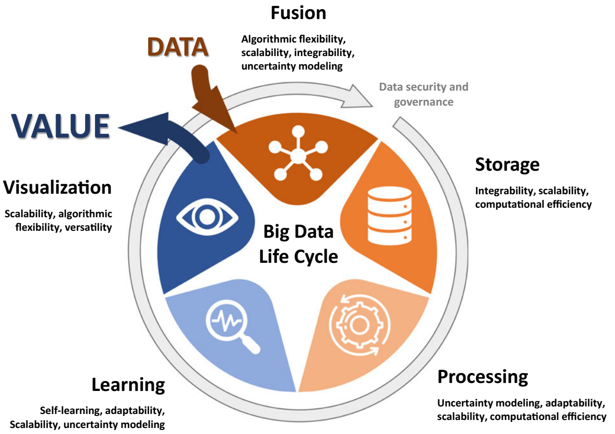

Fusion
Algorithmic flexibility, scalability, integrability, uncertainty modeling
DATA
Data security and governance
VALUE
Storage
Integrability, scalability, computational efficiency
Visualization
Scalability, algorithmic flexibility, versatility
Big Data Life Cycle
Learning
Self-learning, adaptability, Scalability, uncertainty modeling
Processing
Uncertainty modeling, adaptability, scalability, computational efficiency

Fig. 2 Phases of the big data life cycle

what is called *Big Data platform*, which resides in complex server infrastructures. Additional technologies being applied to Big Data include massively parallel-processing systems (MPP), search-based applications, data mining grids, distributed file systems, distributed databases or NoSQL databases and cloud-based infrastructure (applications, storage and computing resources).

All the components that comprise a Big Data architecture have different technological requirements and characteristics, which depend on the purpose they should cover in the ecosystem. In accordance with the increase in these requirements, adopted solutions usually tend to be a set of integrated and suitable tools for data analytics and Big Data. These combined systems are called *Big Data suites*. In the specific context of security [39], several technologies can be found in the Big Data technology stack. In this paper, we analyze the initiatives proposed to improve any of the above technologies (from cloud technologies to analysis assistance tools) by means of bio-inspired computation.

In what refers to infrastructure, Big Data technologies [40] support three options: on-premise, cloud and hybrid. Thus, depending on the approach, the infrastructure management complexity and the needed tools vary significantly. In this case, bio-inspired metaheuristics have demonstrated a remarkable performance when solving complex problems associated with infrastructure and technologies, such as resource allocation and management [41, 42], job scheduling [43], log synchronization and information security [44], or anomaly detection in the management and health of the IT infrastructure [45]. We will later examine them thoroughly.

2.2 Fundamentals of bio-inspired computation

In a nutshell, bio-inspired computation [46] can be defined as the combination of computational intelligence [47] and collective intelligence behaviors [48]. Usually, computations methods classified in this category are conceived for efficiently solving highly complex problems. These solvers are designed using as source of inspiration a wide variety of principles and phenomena encountered in nature and biological systems. The main reason for mimicking such observed behaviors for solving complex computational tasks is to harness the adaptive, reactive and distributed features of these natural systems. In this way, every aspect that defines the solving method is modeled mirroring the living phenomena and biological systems, such as the evolution of species [8], immune systems [49], the human brain [50], or the collective behavior of animals [6, 51, 52], among others.

In this survey, we focus our attention on four specific areas that can be placed within the wider field of bio-inspired computation: neural networks [53, 54], Evolutionary Computation [55, 56], Swarm Intelligence [57] and Fuzzy Systems [58, 59]. All these four concepts are fully related to the Big Data paradigm and the main problems arisen in this field, due to the suitability of their application to this area [60]. Our decision to undertake this study departs from our findings recently drawn in [60]. In this paper we present a taxonomy of bio-inspired computational intelligence, highlighting four major families: Natural Computing, Artificial Immune Systems, Fuzzy Systems and Neural Networks. In our case, we do not consider Artificial Immune Systems given the lack of works reporting advances in the application of this family of algorithms to Big Data systems. This scarcity, however, unveils an interesting research direction in security that we will later discuss in detail.

A convenient criterion to organize all techniques under the bio-inspired computation umbrella is the kind of computational problems that can be solved. As such, computational intelligence techniques and methods can undertake three generic problems (Fig. 3), which differ from each other depending on the unknown information to be solved by the technique at hand [56]:

1. Modeling or system identification, in which given a prior set of inputs and their corresponding outputs, the goal is to determine the model that best relates both, so that a new output can be produced for any given input. All predictive modeling techniques belong to this first category.
2. Simulation, in which given an input data and an assumed expression for the system, the goal is to observe the properties of its produced output. A clear example of *simulation* in the wide sense is clustering: Given an input data, a clustering algorithm is applied towards observing whether the output shows up a certain group structure.
3. Optimization, in which given a system and a measure of quality of its output, the goal is to find the input that maximizes the quality of its output. This is actually what is done by bio-inspired meta-heuristic algorithms.

We define now the four aforementioned large families of bio-inspired computation methods and their connection to the above generic problems. Table 1 complements these explanations with an excerpt of the particular problems in

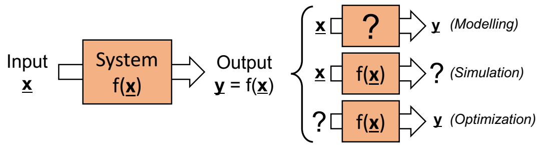


Fig. 3 Conceptual diagram showing the three tasks that can be tackled with Computational Intelligence

Table 1 Relationship between the four families of bio-inspired computation approaches, typical problems and dimensions in the Big Data context

Family	Task	Big data problems	Dimensions
Neural network	Modeling	Server load forecasting	Volume
	Simulation (clustering)	Intrusion detection	Variety
		Data compression	
		Predictive maintenance	
		Manifold learning	
EC	Optimization	Task scheduling	Variability
SI		Design of data filters	Velocity
		Resource allocation	
		Server placement	
		Database sizing	
Fuzzy systems	Modeling	Predictive control	Volume
	Simulation (clustering)	Multi-criteria decision making	Veracity

the context of Big Data that such families can address, as well as the affected Big Data dimensions:

2.2.1 Neural networks

Neural networks are computational models inspired by brain modeling studies. It consists of a set of units, called artificial neurons, connected together to transmit signals. The smallest unit of analysis of neural networks in the computational domain is what is called neuron or perceptron. An important feature of neural networks is their ability to learn from their environment. Neural networks have been widely applied on supervised, unsupervised, hybrid and reinforcement learning [61]. For this reason they have been extensively applied to modeling problems such as classification, regression or matching, as well as to simulation problems via unsupervised neural approaches such as Kohonen maps, auto-encoders, Hebbian learning and the like.

2.2.2 Evolutionary computation

Evolutionary Computation (EC) comprises a family of algorithms for global optimization inspired by biological evolution. Some recurrent ideas that have been used as inspiration up to now are, among others, the survival of the fittest, natural selection, reproduction, mutation, competition or symbiosis. For properly emulating the processes involved in nature and the natural selection mechanism, candidate solutions are organized in a population, and the fitness function determines how good they are adapted to the environment in which solutions *live*. This fitness should be strictly related with the problem at hand, being proportional to the quality of the solution solving that problem. Most representative EC techniques, which differ in the way in which they represent and evolve individuals, are as follows: (1) *genetic programming*, in which individuals are represented as executable programs [62]; (2) *evolutionary programming*, phenotype-oriented [63]; (3) *evolutionary strategies*, which can be deemed as the *evolution of evolution* [64]; (4) *differential evolution*, population-based search strategy in which the modification of individuals is based on the difference between them [65]; (5) *genetic algorithms*, population-based techniques based on the Darwinian evolution of species theory [8]; (6) *cultural evolution*, adaptation to the environment at faster rates than biological evolution [66], (7) *co-evolution*, distribute solvers in which multiple subpopulations evolve in a joint way [67].

Up to now, EC has been applied in a wide spectrum of knowledge fields. For interested readers, we suggest the findings reported in works such as [68–70] for the analysis of recent research trend in some specific applications.

2.2.3 Swarm intelligence

Swarm Intelligence (SI) is a specific branch of Computational Intelligence also dedicated to the optimization of complex problems through the study and adaptation of the collective behavior of decentralized, self-organized agents. This way, SI methods usually consist of a population (*swarm*) of simple agents, which evolve jointly along time through local interactions with one another, and with their environment. Furthermore, despite the interactions among individuals are determined beforehand, social interaction plays a key role in the resulting behavior of the swarm towards achieving a global objective. In other words, although every agent relies on local interactions impacting on the resulting behavior of the swarm, the global performance of the group simultaneously determines the conditions under which individual agents perform. As previously mentioned, a wide spectrum of inspirational sources has been embraced over the last couple of decades for producing SI methods. We can highlight among such sources the behavioral patterns of animals such as bees [7], cuckoos [51], fireflies [71], or cats [72]. Other inspiring motifs for SI methods are physical processes, such as the electromagnetic theory [73], optic systems [74], or general relativity [75]. Social human behaviors have also served as inspiration for modeling novel metaheuristics, with renowned examples such as anarchic societies [76].

One of the main features that make SI methods specially efficient for solving optimization problems is their ability for distributing the optimization tasks, decentralizing in this way the evolution of solutions. This feature makes them particularly appealing for their implementation in Big Data ephemeral environments, in which computation resources are intermittently available. Other acknowledged differences of this optimization paradigm with respect to EC are the behavioral mechanisms by which the swarm evolves towards the best solution of the problem at hand, which are driven by one-to-one simple interaction rules rather than by population-based selection and crossover operators (see Fig. 4 for a diagram illustrating such differences).

2.2.4 Fuzzy systems

Fuzzy systems are specific mechanisms within Computational Intelligence which faithfully adapts to the human reasoning model and to the real-world. This logic introduces a better understanding of clauses of the type `it is hot`, `it is high` or `it is fast`. In this context, the term *fuzzy* refers to the fact that the logic involved can deal with concepts that cannot be expressed as `true` or `false`, but rather as `partially true`. For reaching this goal, the core concept of fuzzy systems is to understand the quality quantifiers for inferences and human reasoning. In this way, fuzzy systems are usually used as mechanisms inside other methods, but also as monolithic methods. Up to now, many real-world applications have been benefited from these paradigms, mainly control (optimization), prediction (modeling) and decision support [77–79].

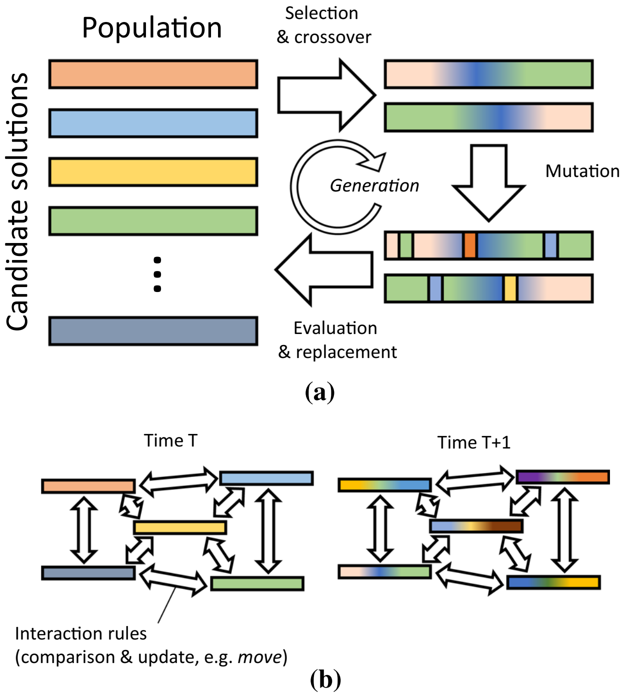


Fig. 4 Diagram depicting the differences between **a** EC and **b** SI

3 A joint perspective on bio-inspired computation and big data

This section is devoted to presenting and describing the main synergies between both paradigms studied in this paper: Big Data and bio-inspired computation. Several reviews and surveys have so far addressed this intersection from different perspectives, domains or applications. Table 2 summarizes the essential information of such works carried out during the last two years, including the period of time covered by the articles analyzed in it, the number of reviewed works, the proposal of a taxonomy to organize them, the phases of the Big Data life cycle covered, families of bio-inspired algorithms under scope and, finally, whether a critical analysis, challenges and research directions are given. The comparison made in these terms with the present work reveals several aspects of improvement:

- A self-contained introduction to the concepts underneath Big Data and bio-inspired computation (Sect. 2), helping the reader understand their synergies and complementarities, as reflected in Table 1 and Fig. 1.
- A significantly higher number of reviewed works (324), which nearly triples the amount of references considered in other surveys alike.
- A wider domain coverage than other similar studies that focus only on a reduced subset of the phases of the Big Data life cycle, collection and processing/analysis. In our case, we use a three-fold criterion when designing our taxonomy: Big Data infrastructure, Big Data technologies and Big Data life cycle phases.
- A more extensive taxonomy to classify the works under analysis in terms of the families of bio-inspired algorithms used in every reviewed work.
- A critical analysis dissecting what has been done so far in the field, along with a set of future challenges that are tightly connected to bio-inspired computation and Big Data, avoiding to fall into generalistic formulations.

For this purpose, two separated biases have been used: (i) the adoption of bio-inspired computation for modifying

Table 2 Recent overviews on Big Data connected to bio-inspired computation, and their comparison to this work

Survey	Period	# Reviewed works	Literature taxonomy (classification criterion)	Big data dimension and lifecycle	Families of bio-inspired computing approaches	Critical analysis, trends and challenges
[12]	2002–2018	83	Yes (uncertainty modeling/ mitigation approach)	Big data collection, fusion and processing	Evolutionary algorithms, artificial neural networks and Fuzzy Logic	Real-time machine learning and meta-heuristic algorithms for mitigating uncertainty
[20]	2001–2019	96	No	Big data processing and learning	Fuzzy logic, Evolutionary algorithms and artificial neural networks	Development of modern smart cities
[80]	2014–2019	35	Yes (family of bio-inspired algorithms)	Big data Processing and Learning	Evolutionary, swarm-based and ecological algorithms	Containers, serverless computing, blockchain, software-defined clouds and quantum computing
[81]	2014–2019	116	Yes (Big data analytics for Industrial IoT)	Big data collection, fusion and processing	No specific focus on algorithmic solutions	Security, privacy and concentric computing
This work	2010–2020	231	Yes (Infrastructure management, Big data technologies, Big data phases)	Big data fusion, storage, processing, learning and visualization	Evolutionary computation, swarm intelligence, neural networks and fuzzy systems	Lack of reference problems, unrealistic use cases, algorithmic novelty, operationalization, real-time Big data, XAI and security (Sect. 4)

different technologies of the Big Data stack, in terms of infrastructure and life cycle technologies; and (ii) the evolution of bio-inspired algorithms adapted to the Big Data life cycle and its features, such as programming models and Big Data volumes. To this end, we divide the analysis into three subsections. The first two are associated with Big Data infrastructure (Sect. 3.1) and Big Data technologies (Sect. 3.2), elaborating on how they can leverage the adoption of bio-inspired computation approaches. Section 3.3 rounds up this joint perspective by outlining bio-inspired computing algorithms that have been adapted to the Big Data domain, emphasizing on the life cycle phases involved in each bibliographic item. Figure 5 summarizes the recent literature noted in the field, in which the combination of these technologies has reported remarkable performance and efficiency gains so far.

3.1 Bio-inspired computation for big data infrastructures

Generally, Big Data platforms can be deployed into two different kinds of infrastructures: *on-premise* or *in the cloud* [82]. Furthermore, a third approach hybridizing these two concepts is also possible. The existence of these types makes necessary the existence of tools for the systematization of the deployment, used as a guide for the system administrator. In this specific point is where the optimization capabilities of bio-inspired computation solvers acquire relevance, allowing for the automatization of these tasks in an efficient fashion. The main goal of the system administrator is to achieve a smart system management, which can lead to significant improvements in resource usage, such as provisioning; virtualization and allocation [41]; scheduling and optimization; balancing and reservation; and anomaly detection, among many others. In this regard, it should be clarified that resources are conceived as the elements that make up the infrastructure, such as virtual machines, containers, network elements, physical servers or computer nodes.

In addition to the above heterogeneity of resources and tasks, the inherent characteristics of new approaches to Big Data Analytics (speed, non-stationarities, and resilience to failure/ephemeral computing resources) have opened up new challenges in terms of adaptability, learning and self-organization. Analytical models are nowadays deployed on hybrid, volatile, highly scalable and rapidly reconfigurable resources. It is within this complex ecosystem of computation technologies where it becomes essential to ensure that systems and processes meet the aforementioned capabilities, paving the way for bio-inspired computation to become an enabler for this purpose.

To properly categorize the analysis of the study, we follow the previously mentioned classification, which is the most commonly used within the Big Data context: on-

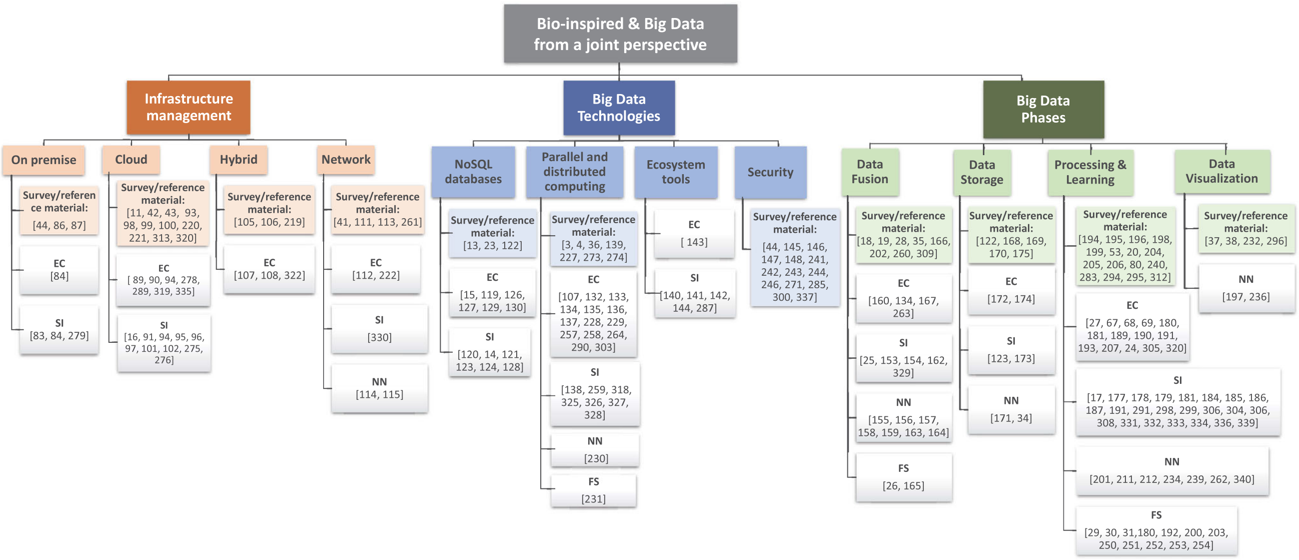

Bio-inspired & Big Data from a joint perspective
Infrastructure management
On premise
Survey/reference material: [44, 86, 87]
EC [84]
SI [83, 84, 279]
Cloud
Survey/reference material: [11, 42, 43, 93, 98, 99, 100, 220, 221, 313, 320]
EC [89, 90, 94, 278, 289, 319, 335]
SI [16, 91, 94, 95, 96, 97, 101, 102, 275, 276]
Hybrid
Survey/reference material: [105, 106, 219]
EC [107, 108, 322]
Network
Survey/reference material: [41, 111, 113, 261]
EC [112, 222]
SI [330]
NN [114, 115]
Big Data Technologies
NoSQL databases
Survey/reference material: [13, 23, 122]
EC [15, 119, 126, 127, 129, 130]
SI [120, 14, 121, 123, 124, 128]
Parallel and distributed computing
Survey/reference material: [3, 4, 36, 139, 227, 273, 274]
EC [107, 132, 133, 134, 135, 136, 137, 228, 229, 257, 258, 264, 290, 303]
SI [138, 259, 318, 325, 326, 327, 328]
NN [230]
FS [231]
Ecosystem tools
EC [143]
SI [140, 141, 142, 144, 287]
Security
Survey/reference material: [44, 145, 146, 147, 148, 241, 242, 243, 244, 246, 271, 285, 300, 337]
Big Data Phases
Data Fusion
Survey/reference material: [18, 19, 28, 35, 166, 202, 260, 309]
EC [160, 134, 167, 263]
SI [25, 153, 154, 162, 329]
NN [155, 156, 157, 158, 159, 163, 164]
FS [26, 165]
Data Storage
Survey/reference material: [122, 168, 169, 170, 175]
EC [172, 174]
SI [123, 173]
NN [171, 34]
Processing & Learning
Survey/reference material: [194, 195, 196, 198, 199, 53, 20, 204, 205, 206, 80, 240, 283, 294, 295, 312]
EC [27, 67, 68, 69, 180, 181, 189, 190, 191, 193, 207, 24, 305, 320]
SI [17, 177, 178, 179, 181, 184, 185, 186, 187, 191, 291, 298, 299, 306, 304, 306, 308, 331, 332, 333, 334, 336, 339]
NN [201, 211, 212, 234, 239, 262, 340]
FS [29, 30, 31,180, 192, 200, 203, 250, 251, 252, 253, 254]
Data Visualization
Survey/reference material: [37, 38, 232, 296]
NN [197, 236]

Fig. 5 Taxonomy of works related to the application of bio-inspired computation to the big data domain, classified as per the different application areas under consideration

premises infrastructures (Sect. 3.1.1), cloud infrastructures (Sect. 3.1.2) and hybrid approaches (Sect. 3.1.3).

3.1.1 Bio-inspired computation applied to on-premise infrastructures

Briefly explained, *on-premise* regards to the software and technology located within the physical confines of an organization. This concept opposes running the system remotely on hosted servers or in the cloud. Thus, by installing and running software on hardware located within the premises of the company, full physical access to the data is available. Furthermore, the configuration, management and security of the computing infrastructure can be carried out directly in the system.

Regarding the configuration and management, bio-inspired computation can resolve problems related to task allocation and resource scheduling. In [83], for example, the authors present an approach based on distributed SI mechanisms that mimic the behavior of social insects to solve problems such as overlay management, routing, task allocation, and resource discovery. Through this approach, the authors of [83] construct an adaptive and robust management system for peer-to-peer networks.

The use of Graphics Processing Units (GPUs) and cluster-based parallel computing techniques is also a research trend, aiming at accelerating the process of extracting the correlations between items in sizeable data instances. In [84], for instance, authors propose four different population-based metaheuristics for efficiently mining association rules, which benefit from the cluster intensive computing and massive GPU threading.

On another vein, a special case of Big Data *on-premise* infrastructure is the so-called High Performance Computing (HPC, [85]), which refers to hardware and programming models specialized in solving highly complex problems mainly via parallelization. In this sense, using HPC solutions requires new techniques for memory management. An interesting recent survey published by Pupykina et al. [86] discusses the challenges of memory management in HPC and Cloud Computing, including a review of bio-inspired optimization methods to increase memory utilization.

In the security context, referring to the application level security as well as advanced protection against malware, the paper presented by Mthunzi et al. [44] proposes a comprehensive review of the benefits that the application of bio-inspired algorithms brings to the specific field of cybersecurity. It is also interesting the work of Rauf et al. [87], which highlights and discusses challenges and open opportunities in the intersection of cybersecurity and bio-inspired computation. Lastly, another totally different approach can be found in [88], where several management problems related to the increase in complexity and the need for energy are addressed in detail. For achieving the planned objectives, a bio-inspired self-organized technique is proposed for the redistribution of load among servers in data centers.

Reflecting on the activity noted so far on bio-inspired computation applied to the design, management and operation of on-premise Big Data infrastructures, we stress on the lack of informed evidences whether bio-inspired algorithms can meet realistic complexity scales of large computing farms. Furthermore, even if resource utilization does not vary as dynamically as in other alternative shared computing environments, most works reviewed in this strand of literature do not inform about the latencies induced by the usage of bio-inspired methods for, e.g., resource balancing or fast evolving computing tasks, which could hinder their practical adoption in Big Data environments subject to timing constraints. This criticism mostly refers to optimization methods: Biologically inspired modeling solutions suited for their deployment over Big Data infrastructure are far more mature than their optimization counterparts.

3.1.2 Bio-inspired computation applied to cloud computing infrastructures

In few words, Cloud Computing infrastructure can be defined as the collection of hardware and software elements needed to enable the remote management of the whole Big Data system. These concepts include computing power, networking and storage. It also contemplates an interface for users to access their virtualized resources, like cloud management software, deployment software and platform virtualization. In the Big Data context, the ability of Cloud Computing to offer fully scalable technical resources adapted to the needs of each project is crucial. Thanks to that, limitations of traditional physical servers are avoided. However, appropriate management tools are needed in order to efficiently take care of tasks such as resource virtualization or services deployment optimization.

In the current literature, works in this line of research can be classified into two main strands: (i) approaches related to the resource provisioning and allocation in Cloud Computing environments, and (ii) tasks related to the deployment, planning and optimization of services and applications:

- On the one hand, the allocation and scheduling of multiple virtual resources, such as virtual machines (VMs), is a well-known research field in Cloud Computing. In [89], for example, a Genetic Algorithm is proposed for the optimization of VM distribution across a federated cloud. Similar is the approach followed by Rocha et al. in [90], which presents a hybrid optimization model that allows a cloud service provider to establish VM placement strategies. This way, the energetic efficiency and network quality of service are jointly optimized. More recent is the work presented in [91], which solves the same problem by means of an ant colony system. In addition, the research introduced in [92] hybridizes a Firefly Algorithm with fuzzy logic for server consolidation and VM placement in cloud data centers. Also interesting is the study presented in [93], which focuses on Hadoop Big Data technology. In that work, authors implement a bio-inspired solver for optimizing the placement of VMs in OpenStack. In [94], Pires et al. propose a novel multi-objective formulation of the VM placement problem, which is addressed by means of a novel multi-objective memetic algorithm. Additionally, in [95] an Ant Colony Optimization and dynamic forecast scheduling is combined for solving the VM placement problem, showing a remarkable efficiency in terms of less wasted resources and better load balancing. Finally, an interesting approach based on Cuckoo Search is proposed in [96] for data center resource provisioning in the cloud.
- On the other hand, task scheduling over distributed and virtual resources is a main concern which can affect the performance of Big Data system. In [97], a meta-heuristic algorithm called Chaotic Social Spider Algorithm is developed for solving the task scheduling problems in virtual machines. The authors of this work focused on minimizing the overall makespan, while leveraging load balancing. Additionally, in the survey presented in [98], different bio-inspired approaches are analyzed for tackling the aforementioned problem. A work closer to Big Data technologies is conducted in [99], in which authors theorize on how the Map Reduce programming model performs the assignment of tasks in Cloud Computing environments. This analysis is carried out by resorting to assorted algorithms, including bio-inspired techniques.

It is also worth mentioning that one of the key goals in cloud environments is the optimal use of resources, for which load balancing techniques are often applied. This has been a particularly profitable playground for bio-inspired optimization techniques, yielding extensive surveys such as the one in [100], which provides a wide coverage of nature-inspired meta-heuristic techniques applied in the area of cloud load balancing. In this line [101] addresses the problem of load balancing in cloud environments by proposing a hybrid Cuckoo Search and Firefly Algorithm, showing a promising performance. An additional approach for load balancing is described in [102], focused on both Fog and Cloud Computing environments. The authors compare the performance of several bio-inspired computation methods, including Cuckoo Search, Flower Pollination and Bat Algorithm.

Our review of the literature related to Cloud Computing infrastructure has revealed that in most cases, the conditions under which algorithmic proposals are validated are largely uncoupled from the constraints and computation budgets that such algorithms would encounter in practical settings. This criticism refers not only to the scales by which, e.g., load balancing methods are validated (regime of tasks/users being concurrently handled), but also when it comes to the variability in time of the tasks under computation. Furthermore, very scarce to null attention is paid to the *efficiency* of the bio-inspired algorithm itself, mainly due to the simplicity of the simulation settings under which algorithms are validated. We advocate for a closer look taken at the implications of using bio-inspired algorithms, taking a step aside common practice, and informing the community of bio-inspired methods that can *truly* be adopted under computation-intensive regimes.

3.1.3 Bio-inspired computation applied to hybrid big data infrastructures

As mentioned, hybrid infrastructures comprise a blend of private clouds, public clouds and on-premise data centers. Thus, Big Data systems and applications can be deployed on any of these environments, depending on several business strategies, such as the main objective of the system, its tactical requirements and the required outcome. This is the case for heterogeneous distributed systems, in which environments and resources such as cluster computing, grid computing, peer-to-peer computing, cloud computing and ubiquitous computing are mixed [103, 104]. This particular scenario brings the necessity of efficiently managing a large variety of tools and software. This need motivates the development of new algorithms schemes for events and tasks scheduling. Thus, new methods for resource management should also be designed for increasing the performance of such systems. In [105], for example, a valuable survey is presented revolving around the advances on scheduling algorithms, energy-aware models, self-organizing resource management, dataware service allocation, Big Data management and performance analysis. All this analysis is conducted from the perspective of bio-inspired computation. In [106], a review of biological concepts and principles to solve service provisioning problems is presented, along with the proposal of a bio-inspired cost minimization mechanism for data-intensive scenarios where such problem emerges. The proposed method utilizes bio-inspired mechanisms to search and find the optimal data service solution in Big Data environments, considering data management and service maintenance costs. Finally, in [107], a preliminary work is presented on the deployment of evolutionary algorithms on Hybrid Big Data infrastructures. To do that, authors widen the functionality of the well-known ECJ tool [108] for fulfilling their purpose.

On a short reflexive note, here we foresee an increasing prevalence of bio-inspired algorithms capable of bringing together multiple conflicting objectives. Such objectives emerge as a result of the hybridization of different infrastructures, both private and public, which may have some goals in common (e.g., energy efficiency), but others that delineate an interesting Pareto trade-off to be balanced (correspondingly, cost of service versus fairness in the distribution of shared public computing resources). This paves the way towards a magnificent opportunity for multi-criteria decision making algorithms suited to deal with multiple confronted objectives, such as multi-objective meta-heuristics. Our examination of the literature uncovers that this is a niche of opportunity that should attract more efforts in the near future.

3.1.4 Bio-inspired computation applied to big data networks

We finish this subsection turning our attention towards a particularly significant element within the infrastructure: the network. In fact, different computing models can configure their operation based on the network topology and the associated communication latency. Examples of these models are Fog [109] and Edge Computing [110]. In this area, there are multiple open opportunities and a wide room for improvement, by means of optimization techniques used for orchestrating the deployment of elements depending on the features and distribution of the network. It is in this specific stream in which bio-inspired algorithms can emerge as an efficient approach for the aforementioned orchestration. For instance, in [111] a scheduling method for application modules in a fog computing environment is proposed using bio-inspired solving schemes such as Genetic Algorithm, Particle Swarm Optimization and Ant Colony Optimization for the reduction in the energy consumption and execution time. A similar approach is proposed in [112], in which a framework for the optimal deployment in Fog/Edge Computing environments via bio-inspired algorithms is described.

Another cornerstone task related to the infrastructure network is the security in communications. For this problem, bio-inspired algorithms can also be very useful, as shown in [113]. In that paper, authors propose a semi-class intrusion detection method which combines multiple classifiers to arrange exceptions and typical exercises in a computer system. Another axis of interest is the scalability of the network, which is also an aspect of utmost relevance in Big Data scenarios. In [114, 115], for example, authors propose and utilize a framework that supports simulation and testbed experiments to investigate the scalability and adaptability of ant routing algorithms in networking.

In this application area, there is a notable inertia towards the use of bio-inspired techniques for network security purposes. However, Big Data networks, *stricto sensu*, has so far not been risen much interest in the use of bio-inspired computation to address inherent problems such as latency minimization, routing or network dimensioning.

We nevertheless envision that the extrapolation of the Big Data paradigm towards ephemeral computing will span further opportunities due to the intermittency of the network, the variability of task completion schedules and the uncontrolled availability of computation nodes. It is only under these circumstances when the complexity of governing ephemeral computing resources will require the flexibility and adaptability granted by bio-inspired computation.

3.2 Bio-inspired computation applied to big data technologies

The fast evolution and the emergence of new technologies in the Big Data stack, along with the adhesion of a growing number of organizations to this paradigm, causes the appearance of new challenges and opportunities in this field. Usually, these challenges are associated with the development, management and operation of new functionalities. In this regard, one of the essential aspects related to the Big Data technology stack is the non-functional requirements that the solution and tools need to consider. Singh et al. explain in [116] some of the most representative ones: (i) scalability; (ii) data I/O performance; (iii) fault tolerance; (iv) real-time processing; (v) supported data size; and (vi) iterative task support. Based on these six criteria, we can classify Big Data tools into three large groups [117]: *NoSQL databases*, *parallel and distributed programming models* and *ecosystems of tools*. We now analyze them in detail:

3.2.1 Bio-inspired computation and NoSQL databases

In a nutshell, a NoSQL [118] database provides a mechanism for the storage and retrieval of data, which is modeled in means other than the traditional tabular relations used in relational databases. This kind of database presents different points of improvements which can be addressed through the application of bio-inspired algorithms. Some of these applications are related to the horizontal scalability (choice of cluster topology), availability and replication of the data (assignment of the replicas to the nodes), or the consistency level of the information (ensuring the writing optimization), among many others.

In [119], for example, authors present a framework that allows Hadoop to manage the distribution of the data and its placement based on cluster analysis of the data itself. This work is not directly related to NoSQL databases, but it arguably represents an interesting approach for optimal data distribution in physical storage using evolutionary clustering techniques. The paper presented by Nowosielski et al. [120] is a good example of how bio-inspired solvers can aid in the achievement of horizontal scalability, specifically the Flower Pollination and the Krill Herd metaheuristic algorithms. In the specific context of data availability and replication, the work published in [14] presented an adaptive distributed database replication technique based on the application of an algorithm based on colonies of *pogo antsis*. An additional valuable research can be found in [121], in which the Firefly Algorithm is applied for the positioning and optimization of traffic in NoSQL database system, modeled with exponentially distributed service and vacation. Bio-inspired computation can also contribute to the design of the logical data schema. The research presented in [122] is an example of this trend, proposing a design repository for storing and retrieving biological (and engineering) design strategies.

Another interesting investigation is also shown in [123], in which a data warehouse schema design optimization is optimized by means of a Particle Swarm Optimization approach. In [124] a mathematical model of a column-oriented database performance was presented. Authors propose the use of Flower Pollination Algorithm for regression equation coefficients optimization. Furthermore, they highlight its accuracy and sophistication, which makes it appropriate for the foundation of database performance optimization.

Another highly relevant field of study combining NoSQL databases and bio-inspired computing is the so-called query optimization [125]. The work presented by Rani et al. in [126], for example, proposes the use of a bio-inspired algorithm based on the antibody-antigen clonal selection scientific theory for the efficient modeling of distributed query plans. The same author presents in [127] a study revolving around the distributed query processing optimization based on artificial immune systems, which is among the few references identified so far where immune systems have been utilized in Big Data scenarios.

Furthermore, there are situations in which bio-inspired techniques assist in the extraction of association rules over databases, as can be seen in [128]. In that study, authors showcase an approach for extracting association rules by applying a Bee Swarm Optimization meta-heuristic algorithm to a large database using the massively parallel threads of a GPU processor. An additional valuable approach is proposed in [129] for association rule mining, in which the JAYA algorithm is applied to big database instances.

Finally, an additional possible viewpoint can also be highlighted in this section, which evinces even further how bio-inspired optimization methods can take advantage of NoSQL technologies. This is the concrete proposal of Jordan et al. in [130]. In this paper, authors showcase how a system benefits from optimization knowledge persisted on a NoSQL database, serving as associative memory to better guide the optimizer through dynamic environments. This supports our claim that bio-inspired computation can not only benefit non-conventional databases, but can also leverage conversely the storage capabilities of such databases to store history information that can be retrieved and exploited by the bio-inspired algorithm upon requiring it, as in, e.g., recurrently changing concepts modeled by neural networks (*continual learning*) or dynamic optimization with bio-inspired meta-heuristics. This synergy is worth to be explored further by prospective studies around recurrent evolving learning environments.

3.2.2 Bio-inspired computation for parallel and distributed computing models

The significant rise of distributed and parallel processing techniques has dramatically transformed the use case landscape, improving existing levels of processing performance. In this context, two clear approaches can be spotted: batch programming models and those adapted to real-time or streaming environments. As in other situations discussed before, problems arising in these two scenarios can be tackled through the perspective of bio-inspired computation.

On the one hand, regarding batch parallel programming models, two main challenges can be found: (i) improvements over existing programming models (such as MapReduce [131]), or (ii) the development of new improved computing approaches under bio-inspired computation techniques. In the first case, we find interesting works such as [132] and [133]. In those studies, the former introduces improvements into the programming model regarding the efficient distribution of tasks, whereas the latter showcases more precise locations of the distributed data. Another remarkable research work can be found in [134], which provides a Big Data scheme based on Spark to handle highly imbalanced datasets. They successfully validated their approach over several datasets composed of up to 17 million instances. In [135], Hans et al. present details about reshaping the DEAP library for Evolutionary Computation by parallelizing the costly evaluation of encoded programs (individuals) on a Spark cluster. It is interesting to highlight also the work presented in [136], where authors focus on the Cloud Computing paradigm with emerging programming models, such as Spark, to prove how several parallel differential evolutionary algorithms can perform well in this situation. Obtained outcomes demonstrate the existence of a competitive speedup against serial implementations, along with a remarkable horizontal scalability. Finally, we can find new programming models such as the one proposed in [107], in which a new approach to deploy computing intensive runs of enterprise applications on Big Data infrastructures is presented.

On the other hand, a streaming system can be referred to as *real-time* if it guarantees a response within tight deadlines. Furthermore, depending on the specific context of the application, *tight times* can be a matter of minutes, seconds, or even milliseconds. Nowadays, due to the velocity dimension of Big Data, these systems are cornerstones of the technology stack in the treatment of large volumes of data, and they can take advantage from the characteristics of bio-inspired computation, such as its speed and efficiency when solving complex problems. A proof validating this claim is the existence of the so-called Software Model for Distributed Incremental Closeness Factor-Based Algorithms (SMDICFBA), in which incremental clustering models are proposed to learn dynamically about embedded patterns from raw data [137]. An additional example for supporting this statement can be found in [138], in which a new approach to stream computing is introduced. For achieving online optimization and scheduling, a particle swarm optimization algorithm hybridized with back-propagation and an immune clonal algorithm are used in that work. Lastly, we pause at the term *Organic Computing* [139], which behaves and interacts with humans in a bio-inspired manner. All in all, it is important to ensure that the efficiency of bio-inspired algorithms do not clash with the stringent computational requirements imposed by avant-garde parallel computing setups. Connecting back with our reflections offered previously, there is little evidence of implementations of bio-inspired optimization algorithms that can perform within realistic computational boundaries.

3.2.3 Bio-inspired computation for big data ecosystem tools

Big Data Ecosystem can be defined as a framework for solving Big Data problems, comprised by a suite of cluster management and task/jobs scheduling/assignment tools, which encompasses a number of valuable services (ingesting, storing, analyzing and maintaining). An example of this kind of ecosystems optimized by bio-inspired computation can be found in [140], which presents a hybrid Particle Swarm Optimization–Genetic Algorithm for solving the task assignment problem. Another case is presented in [141], in which a bio-inspired method based on ant systems is developed for optimizing the distribution of service deployment. Regarding scheduling, we can find works such as [142], in which multi-stage multi-machine multi-product scheduling problem is resolved using the Bat Algorithm. In [143] energy-aware cloud task scheduling is studied by resorting to the same method. Finally, a task scheduler on diverse computing systems is described in [144]. In that case, the system is developed as a hybridization of the bat algorithm and the artificial bee colony. Apart from these reviewed works, we have not found any further contributions showcasing tools for Big Data ecosystems empowered by bio-inspired algorithms.

3.2.4 Bio-inspired computation for security

We finish this section by devoting a few lines to works related to security technologies. Interesting investigations on this context can be found in [145–147]. Being strict, these works are not directly associated with Big Data environments, but they are used for paradigms such as Cloud Computing or Internet of Things. All these papers adopt the use of bio-inspired algorithms for solving different problems such as access control or intrusion detection, which are common to any complex networked system. Big Data is by no means an exception, and should embrace advances in bio-inspired computation for security purposes in future evolutions of its technology stack, including all applications for which this area of Artificial Intelligence has a long history of successes in network security.

3.3 Big data life cycle bolstered by bio-inspired computation

In Sect. 2.1.2 we introduced the Big Data life cycle, which is made up of different phases. Bio-inspired computation can improve each of such phases in terms of efficiency and fulfillment of non-functional requirements. In this section, we outline a significant group of valuable works for each phase, which arguably help understand the importance of the consideration of bio-inspired algorithms over each of these life cycle phases.

The relevance of bio-inspired methods applied to the Big Data paradigm has been previously studied, but always associated with specific algorithm categories or under the prism of specific problems. For example, a survey on data science with population-based algorithms is presented in [149]. Authors of this work focus on EC and SI, and they acknowledge the need for new techniques in the field to appropriately deal with the problems, scales and requirements arising from Big Data. Likewise, the work in [150] paves the way towards using genetic programming in Big Data problems. This work shows and discusses different ways of configuring Big Data training evaluations and parallelization, and demonstrates their impact on efficient problem solving.

For the sake of comprehensiveness, we show in Fig. 6 the different life cycle phases and solutions that bio-inspired computation provides for each of them. We proceed now to overview the research conducted up to now on each of these life cycle steps: data fusion (Sect. 3.3.1), data storage (Sect. 3.3.2), data processing and learning (Sect. 3.3.3) and data visualization (Sect. 3.3.4).

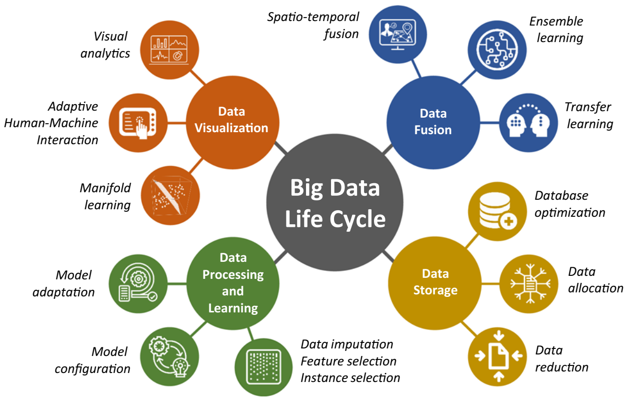


Fig. 6 Application areas of bio-inspired computation for each Big Data life cycle phase

3.3.1 Data fusion and bio-inspired computation

Data fusion is the process of integrating multiple data sources to produce more consistent and useful information. In the Big Data paradigm, this is a crucial procedure due to the large amount and heterogeneity of the data sources that currently can be found in a given use case. From the perspective of bio-inspired computation, this is a problem that has been tackled in the literature before, as can be seen in valuable reviews such as [151, 152] or [153]. Furthermore, there is a clear consensus that the relevance of this topic increases along with the volume of information becoming larger.

The heterogeneity of the data and the diversity of their sources cause difficulties when accessing and understanding their underlying structure. Users identify a problem for properly representing and interpreting the same real-world objects recovered from different data sources. In this context, [154] presents an approach to solve the dynamic feature selection based on Big Data fusion with multi-objective particle swarm optimization. Another example is proposed by Dong et al. in [155], in which authors determine security threats in power grid by making full use of heterogeneous data sources in power big data. In that paper, researchers map heterogeneous data in different formats to a unified embedded vector space with deep restricted Boltzmann machine, achieving the efficient fusion of heterogeneous data sources. Furthermore, Zhang et al. have published several recent works related to Big Data Fusion techniques using ensemble learning and Neural Networks as their core of research [156, 157]. As a matter of fact, ensemble learning can also be conceived as a fusion of decisions made by the constituent models in the ensemble. Bearing this in mind, the automatic construction of ensembles has also largely leveraged the use of bio-inspired optimization algorithms [158, 159], with recent examples of their application to Big Data scenarios [134, 160].

Data fusion techniques can be applied to multiple domains such as culture, health, language analysis, and transportation and mobility in Smart Cities. In the cultural heritage domain, Piccialli et al. [161] present and discuss the application of a clustering approach for behavioral classification of IoT cultural data collected in the National Archaeological Museum of Naples (Italy). In the Health domain, for example, we find studies like [162], in which e-health data is collected from patients suffering from different diseases, and the optimal attributes are chosen by using an improved Dragonfly Algorithm for an enhanced classification. In the text analysis domain, the research introduced in [163] proposes and compares effective fusion matching methods using neural networks for automatic removing semantic collision of files. In Smart Cities, Wang et al. [164] present an interesting approach about urban Big Data fusion based on Deep Learning. The investigation detailed in [165] is also centered in Smart Cities, focusing on the management of natural disasters using fuzzy models. In transportation domain, the work [166] presents a study related to train transport, revolving around delay prediction by means of Big Data fusion techniques based on bio-inspired techniques. Finally, we note the profitable strand of literature revolving on rule mining with bio-inspired methods, which has also permeated to the Big Data field. An example is [167], which proposes an efficient associative classifier for large imbalanced datasets based on an evolutionary algorithm that efficiently discovers rare yet reliable association rules.

Without a doubt, the main algorithmic player in bio-inspired computation when it comes to data fusion is Deep Learning. The flexibility of neural architectures to blend together features extracted from different information domains has stepped further over the state of the art as a form of model-based information fusion. Other subfamilies of bio-inspired computation have also been used for this purpose, but rather for auxiliary tasks that help—yet not realize on their own—the fusion of different information flows (e.g., meta-heuristics for neural architecture search).

3.3.2 Data storage and bio-inspired computation

The case of Big Data storage is closely linked to the correct selection and optimization of persistence tools and technologies, which have been already seen in Sect. 3.2.1. Indeed, there are specific tasks associated with this phase of the life cycle which are also likely to be improved by virtue of bio-inspired algorithms. Additionally, these tasks do not only relate with the storage technology itself. An example is the conceptual design of the database schema, with multiple related works such as [122, 123] or [168]. The management and maintenance of large volumes of data is also subject to improvement. This research trend is exemplified by [169], where a biologically inspired algorithm is proposed to identify and mitigate the impact of misbehavior on the performance of data management in social networks. Finally, it is also interesting to highlight [170], which introduces a bio-inspired approach combining Big Data with data intensive computing issues in the future vision of a smart healthcare data management.

A further interesting work related with data persistence is [171], in which authors propose a new algorithm inspired from the working principle of human memory for storing Hierarchical Temporal Memory features detected from an image. A few explorations of data allocation and reduction using bio-inspired methods have been reported in [172, 173] and [34, 174], respectively. Finally, it is interesting to point that there are studies also dedicated to secure sharing of large volumes of data using bio-inspired computing approaches, such as the one presented by Ogiela et al. in [175]. Unfortunately, our bibliography analysis has not yielded any further evidences of biologically inspired mechanisms used for improving the data management efficiency of modern data storage technologies. The plethora of works dealing with relational databases enhanced by bio-inspired mechanisms seem not to have been extrapolated to the Big Data realm, even if the diversity of data and the confluence of spatial and temporal information flows open up large possibilities for the research domain targeted in this survey.

3.3.3 Bio-inspired computation for data processing and learning

These are arguably the most important phases within the Big Data life cycle, since they are the ones in charge of converting data into knowledge. There are many works to consider in this specific area [46]. For this reason, we split these works into two groups: (i) techniques based on bio-inspired concepts for the pre- or post-processing of data, and (ii) adaptation of bio-inspired algorithms to be capable of responding and solving the requirements and dimensions of the Big Data paradigm:

– *Bio-inspired pre- and post-processing techniques* have been widely utilized in the literature for an assorted of possibilities, from data imputation to instance selection, noise filtering, dimensionality reduction or model output simplification [176]. A growing corpus of works can be found in the literature with new algorithmic proposals that undertake the aforementioned tasks in scenarios and setups that could be considered close to the computational requirements imposed by the Big Data paradigm [177, 178]. However, a closer inspection to the literature reveals that an open challenge emerges from the extrapolation of such bio-inspired approaches to the scales of Big Data, which we later discuss in depth in Sect. 4.

- *Bio-Inspired algorithms adapted to Big Data*: in this case, two computational problems have been actively investigated in Big Data environments: clustering (simulation) and prediction (modeling). For clustering purposes, a manifold of research studies have been conducted using different bio-inspired methods, such as [27, 179–181] or [182]. In [183], a technique based on the Whale Optimization solver is presented as a clustering technique to be used in the Big Data domain. Authors evaluate their research against four alternative clustering techniques, obtaining promising results. In prediction, many interesting works can be found in the current literature. In [184], for example, an ant colony-based algorithm is used, in which prediction over data streams is performed. In [185], an Ant Colony Optimization method is also employed for Big Data distribution considerations. The same method is used in [186], where decision analysis is studied over mobile Big Data.
- Another notable group of works to mention are those in which *distributed and parallelizable programming models* are used for the implementation of the bio-inspired algorithms. An example of this trend can be found in [187], using MapReduce for developing a particle swarm optimization-back-propagation neural network algorithm. In [188], Spark is used for developing a Particle Swarm Optimization and a Differential Evolution algorithm. Finally, authors of [189] introduce a parallel population-based optimization algorithm with Spark. Another interesting work along this line is [190], in which a scalable Genetic Algorithm is developed using Apache Spark. To do that, authors maintain the population diversity and minimize the materialization and shuffles in resilient distributed datasets.

Finally, it is interesting to highlight that bio-inspired computation can also be used in conjunction with other techniques, such as time series analysis [191], for the calculation of similarity functions [192]. Furthermore, novel bio-inspired approaches can be created specifically focused on this field of application, such as the Danger Theory presented in [193].

We end up this glimpse at the literature with a notable mention to the prominence of bio-inspired methods used for automating the hyper-parametric tuning process, which have lately grown towards covering the design of the entire data mining pipeline [194, 195]. As we will later expose, the popularity and track of recent success cases of the so-called AutoML research area [196] unleashes a vast research niche for the extension of the functionalities of existing tools and frameworks to Big Data scenarios. The possibility of federating models without compromising the privacy and confidentiality of Big Data from where they learned (also referred to as *Federated Learning*) is another research line with a narrow connection to bio-inspired learning models. However, the practical totality of federated learning scenarios reported to date has gravitated on neural network models, as they easily allow for privacy-aware knowledge sharing, aggregation and redistribution among peers. Furthermore, even though many of these studies resort to the Big Data term in their introduction and claims, they lag notably behind the scales expected for *realistic* Big Data use cases, nor do they generalize to other models for which the federation of knowledge is not that clear to perform. We will later revolve on these issues and their implications towards effective Big Data governance.

3.3.4 Data visualization and bio-inspired computation

On a concluding point for this section, we underscore that techniques for the efficient visualization of large volumes of data are in a relatively less mature point of development. The same happens about their synergy with bio-inspired computation, since works related to both areas of research are scarce. The closest work that falls in this intersection is the one presented by Gritsenko et al. [197]. In that work, a visualization method itself is not presented, but a neural network approach coined as Extreme Learning Machines for visualization is proposed for improving the output of results so that it can be visualized more easily. The difficulty to measure the level of visual perception by the user, his/her cognitive assimilation of the visualization, and the strong case-specific nature of the visualization has hitherto yielded largely ad-hoc tools and techniques. However, we foresee that the current momentum of eXplainable Artificial Intelligence (XAI) tools spawn a new visualization era in which insights about the data are produced by explaining and understanding the knowledge captured by models constructed during the learning phase. The need for coupling the explanatory information embedded in the generated explanations with the cognitive capabilities of the audience becomes very relevant in Big Data contexts. In our targeted application domain, spatial and temporal data often collide together (especially in applications related to Smart Cities, Earth observation or digital twins of large industrial assets), requiring explanations that require a higher degree of sophistication when presenting them to non-specialized users. We will elaborate on this claim in Sect. 4.4.

4 Critical analysis, open challenges and research directions

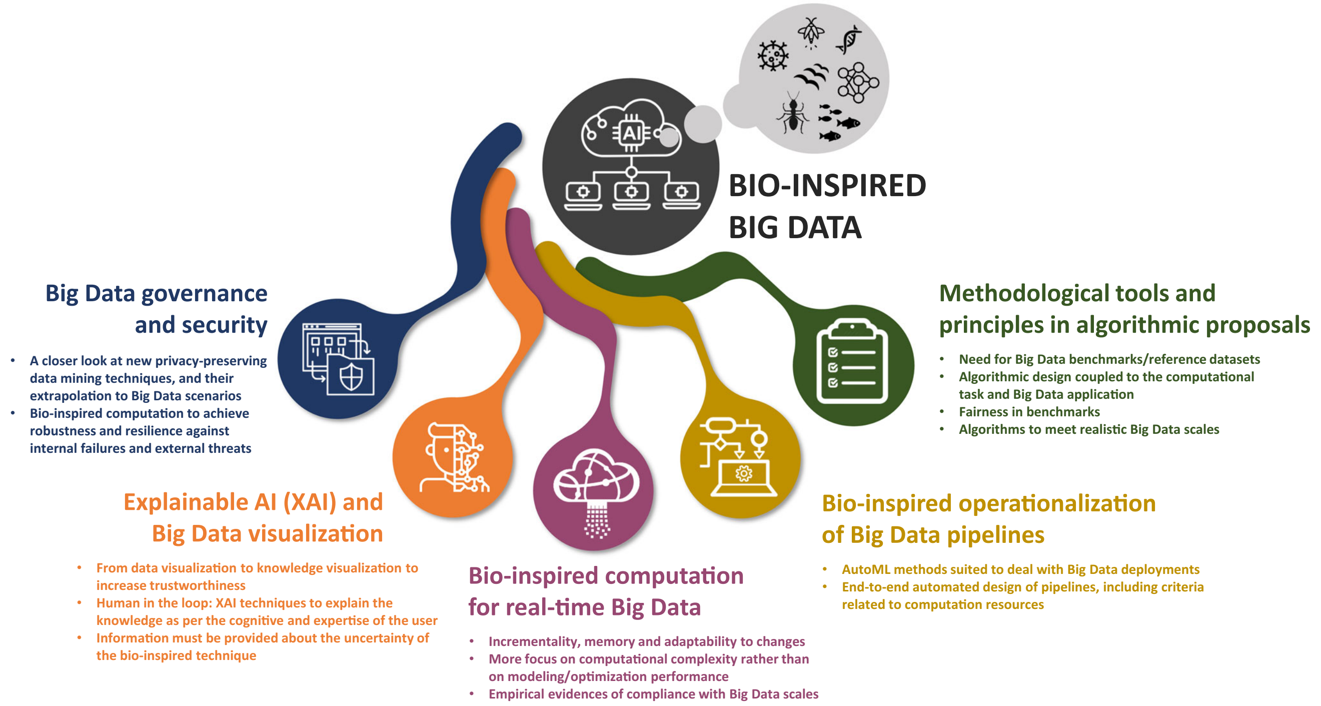
The vast activity noted in the literature is a clear representation of the technical advances attained lately with bio-inspired computation applied to Big Data. Indeed, manifold domains have capitalized bio-inspired computation in data-based applications, including energy [198, 199], transport and mobility [60], health [200], industry [201, 202], agriculture [203], cyber-physical systems [20], social networks [204–206] or sensor networks [207], among many others. Recent worldwide developments around the COVID-19 pandemic have also ignited research activity on Big Data and Artificial Intelligence (in many cases, using deep neural networks for CT scan-based diagnosis), yet without much evidence that the scales of studies claiming to be Big Data so far can be considered as such [208, 209].

In this section we summarize several weak and promising aspects detected at the merger between Big Data and bio-inspired computation. As a result of our literature assessment, we have observed that there are still many questions to investigate when hybridizing both paradigms. In what follows several research niches are enumerated and discussed with respect to the previously analyzed literature. Figure 7 summarizes graphically our prospects of the future of the field.

4.1 Is the community really focusing on big data?

To begin with, a pause of reflection must be first made at the short albeit rich history of bio-inspired computation and Big Data. To quantitatively buttress this statement, Fig. 8 depicts the number of yearly publications retrieved from the Scopus database when being queried with the term `Big Data` and different concepts related to bio-inspired computation. The corpus of literature is impressive, and keeps growing steadily over the years. However, this seemingly vigorous momentum of the field must be assessed with caution: A large proportion of the works encountered during our examination of the literature revealed insufficiently justified usages of the term *Big Data*, reporting algorithmic advances and designing experimental setups far from achieving the scales assumed for Big Data scenarios. No evidences were given on the implementation of the algorithm in question in Big Data frameworks, nor were the datasets in use large and/or fast produced enough to justify the *Big Data* label.

Among the reasons for the fact identified above, we underscore the lack of real public datasets and problems that match the scales assumed for Big Data scenarios, either in terms of volume, variety or velocity. For example, Mann et al.[210] have already detected this problem in the health domain, identifying that there is a wide mismatch between the optimism surrounding the solutions implemented by Big Data technologies and the real existence of



Big Data governance and security

- A closer look at new privacy-preserving data mining techniques, and their extrapolation to Big Data scenarios
- Bio-inspired computation to achieve robustness and resilience against internal failures and external threats

Explainable AI (XAI) and Big Data visualization

- From data visualization to knowledge visualization to increase trustworthiness
- Human in the loop: XAI techniques to explain the knowledge as per the cognitive and expertise of the user
- Information must be provided about the uncertainty of the bio-inspired technique

Bio-inspired computation for real-time Big Data

- Incrementality, memory and adaptability to changes
- More focus on computational complexity rather than on modeling/optimization performance
- Empirical evidences of compliance with Big Data scales

Bio-inspired operationalization of Big Data pipelines

- AutoML methods suited to deal with Big Data deployments
- End-to-end automated design of pipelines, including criteria related to computation resources

Methodological tools and principles in algorithmic proposals

- Need for Big Data benchmarks/reference datasets
- Algorithmic design coupled to the computational task and Big Data application
- Fairness in benchmarks
- Algorithms to meet realistic Big Data scales

Fig. 7 Challenges envisioned in the crossroads between bio-inspired computation and Big Data

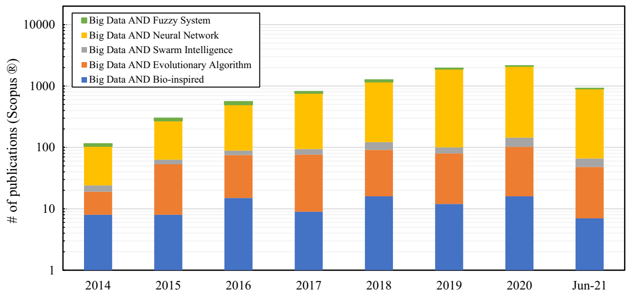


Fig. 8 Yearly publications retrieved from Scopus by submitting the queries indicated in the legend (as per June 1, 2021). The vertical axis is in logarithmic scale

Big Data problems to test them. This remarkable absence of reference problems puts to question the veracity of bio-inspired solutions, and reduces the impact and soundness of contributed tools and software frameworks reported to date. A benchmark comprising several Big Data problems/tasks of realistic scales could be very helpful to set a reference for the assessment of the relative gains claimed by new bio-inspired Big Data solutions. Such a benchmark should bring technical challenges, along at least one of the Big Data dimensions formulated in Sect. 2.1.1, that cannot be tackled by conventional computers and programming tools. Some recent compendiums [211–213] can be of help to discern whether new studies on bio-inspired are *indeed Big Data* or, instead, embrace the term in a less demanding setup. In these works several metrics are also defined, which could also be used in prospective studies (particularly those related to efficiency all along the Big Data life cycle),

Furthermore, most works have been focused on a very narrow portfolio of application scenarios, with the optimization of cloud environments at the forefront of the application of bio-inspired methods. Other Big Data areas such as security and governance, undoubtedly unleash new opportunities that at present, remain largely uncharted by the community.

Another aspect that buries the field in shadows of doubt is the justification of the novelty of the bio-inspired algorithm just by the metaphor that inspires its design. This is a widely acknowledged concern in bio-inspired computation [5], igniting controversial debates around the convenience of these practices for the knowledge advance in the field. As in other application domains, we have identified evidences that such poor practices also prevail in bio-inspired computation for Big Data: many contributions in this line design biased experimental benchmarks favoring their proposed algorithm and penalizing others, by, e.g., tuning the parameters only for selected counterparts in the benchmark, or by varying the conditions under which each algorithm is evaluated (different machines, datasets and/or software implementations). Disregarding the true intentions underneath these poor practices, it should be enforced that prospective studies provide the means to validate the results by third parties, embracing recommendations elicited by recent works on this topic [214].

On a constructive note, it is our firm belief that the community should welcome new biological metaphors for improving the efficiency of Big Data systems along their different dimensions. Nevertheless, it is necessary that new works conform firmly to methodological principles: fairness in the comparisons, experimental replicability and a solid justification why the design of the algorithm is driven by the requirements of the Big Data task to be solved [215]. Implementations of bio-inspired computation approaches in high-performance languages and platforms are largely available nowadays (including GPU versions of optimization algorithms [216, 217]). Furthermore, large-scale global optimization solvers are also a subject of intense investigation [24]. This settles a solid stepping stone and an unprecedented opportunity for bio-inspired computation to meet the scales of Big Data, leaving behind studies of loose connection to Big Data requirements and questionable scientific impact.

4.2 Towards a bio-inspired operationalization of big data pipelines

Traditionally, the scientific community in the field of Artificial Intelligence has focused on the development of new algorithms and techniques over the years. These activities are often conducted under laboratory or experimental settings, overlooking real world potentials and risks. An important challenge can be found in this regard, focused on the life cycle management of Artificial Intelligence approaches and their implementation and maintenance in production environments.

Related to this, Big Data technologies are complex and numerous, and the lack of adequate tools to automatize and operationalize their use and management is a clear problem. In this context, the AIOps concept [218] becomes relevant. AIOps aims to improve and automate all tasks of the software operation phase by employing Artificial Intelligence techniques. As we have analyzed throughout this study, it is clear that the self-learning capabilities of bio-inspired computation techniques have a lot to say in this research direction, given that they are widely used in the development of key tasks of the operationalization process, such as optimization tasks [219] and resource planning [220]. Furthermore, the versatility of bio-inspired algorithms can solve complex problems for highly configurable systems [221], as is the case of the Big Data technology stack specialized in analysis and deployment in Cloud Computing infrastructure.

At this point a relevant point of distinction must be made between (i) the automatic configuration of data-based pipelines (which are collectively referred to as AutoML methods), and (ii) the automated deployment of such pipelines over the resources available in Big Data infrastructures. Both tasks have been recently tackled in isolation, e.g., AutoML has no regards to the available computing resources underneath, nor do deployment tools consider the chance to redesign the data-based pipeline as per the needs and the restrictions of the deployment itself. We definitely advocate for more research efforts invested in blending together requirements imposed at the software (data mining, visualization) and hardware (latency, memory, time) levels. Some recent advances have been done in this direction with the proposal of new specification languages that incorporate elements and requirements from both realms for the distribution of analytical pipelines [222]. Nevertheless, there is still a long road ahead to reach enough maturity for the adoption of these advances in real-world production environments.

4.3 Feasibility of bio-inspired computation for real-time big data

A widely acknowledged problem of bio-inspired algorithms is that in their seminal form, they do not accommodate stringent time constraints as those emerging in streaming contexts. By contrast, the original form of optimization and modeling approaches are better suited to deal with stationary data contexts, in which all the information from where knowledge is extracted is made available before the data processing and learning phases (*batch setting*). However, when information flows continuously, in large volumes and at a fast pace, bio-inspired techniques must be endowed with the features (incrementality, resiliency to data changes, efficiency in the consumption of resources, model memory) required to sustain their analysis and produce outcomes in a similar fashion to the batch setting. Renowned benchmarks for Big Data streaming such as Yahoo! Streaming Benchmark [223] and other recent proposals [224–226] are designed to pose complex challenges for Big Data processing systems in terms of throughput and latency that permeate to the upper layers, e.g., efficient implementations of algorithms that learn incrementally from data that is available for very short periods of time.

In this regard, it would be interesting to investigate new developments or reimplementations of existing algorithms to adapt them to real-time Big Data contexts, even if it is necessary to consider new strategies and methodologies for the deployment of analytical models in streaming systems [227]. For this to occur in the future, a closer look should be paid to emerging paradigms in bio-inspired computation that are specifically suited to real-time scenarios, such as extremely optimized versions of EC and SI solvers, new forms of neural computation for non-stationary streams, or studies in which the operation of the bio-inspired technique is driven not only by the quality of its output, but also by the complexity of its implementation.

Interestingly, the community has already dedicated notable efforts towards anticipating the above needs in the design of algorithms, yielding research areas of utmost relevance such as dynamic optimization [228, 229], learning models over non-stationary data streams [230] or evolving fuzzy systems [231]. Unfortunately, we note very few evidences that such algorithmic developments can be deployed effectively in Big Data contexts, either for the processing, learning and visualization phases of the Big Data cycle, or for supporting the underlying processes of data fusion, storage and governance (in particular, load balancing, dynamic resource allocation or task scheduling, which are often performed in real-time). This is a research niche that should be addressed in the future to shed light on the potentiality of bio-inspired computation for real-time Big Data platforms.

4.4 Explainable AI (XAI) and big data visualization

Big Data often lies in the core of critical decisions, which in some domains of application may entail severe consequences. Health diagnosis is arguably the most enlightening example supporting this statement. A wrong diagnosis of the patient can lead to a wrongly prescribed therapy. Conversely, if Big Data models fail to detect an illness, the patient at hand might undergo fatal consequences. A similar observation can be made in other domains (e.g., defense, law, state administration), mostly in those where decisions affect directly human life anyhow. When this is the case, veracity rises as the Big Data dimension on which a primary focus must be placed, allowing for the quantification of the uncertainty, accountability and the delivery of explanations of the insights drawn from data. In other words: for decisions to be fully informed, opaque models should be avoided or, at least, complemented with techniques that allow understanding the reasons why they were made.

Traditionally, visualization tools have been at the forefront of inspecting large volumes of Big Data, seeking new forms of data representation that allow understanding relationships between heterogeneous data and their evolution over space and time. The term visual analytics was actually forged to highlight the potential that a good visualization has to explore and analyze data without resorting to additional models [232]. However, the scales, variety and veracity of current Big Data scenarios make visualization not enough any longer. Powerful bio-inspired modeling approaches such as Deep Learning networks are in many cases the only viable option to analyze Big Data, surpassing in some cases over-human performance. However, the superior modeling capability of such models clashes with their black-box nature, hindering any chance to explain what they observe in their input data to produce their outputs.

Based on the above rationale, bio-inspired computation for Big Data should massively embrace explainability as one of their main design drivers, either by developing new approaches from scratch that are more algorithmically transparent than their predecessors, or by incorporating tools that provide such explanations. The design of these explainability tools is the motivation of the upsurge of XAI [233, 234] witnessed in the last couple of years. Specifically, XAI refers to methods and techniques developed to ease the interpretation and understanding of decisions made by Artificial Intelligence models by humans, disregarding their expertise or background in this discipline. Other akin research areas that contribute to the trustworthiness of Big Data decisions is confidence estimation, namely, the quantitative evaluation of the epistemic uncertainty of Artificial Intelligence models. Since most bio-inspired learning algorithms are controlled by stochastic processes (for instance, stochastic gradient descent in neural networks, or the search operators in EC- and SI-based search meta-heuristics), a very relevant side information is to compute the variability of the output with respect to the input data and the distribution of the stochastic components of the model.

When endowing Big Data applications with functionalities to explain decisions and estimate the confidence of the deployed algorithms, the entire Big Data life cycle could be trustworthy, ensuring that the *veracity* dimension is appropriately considered. Nonetheless, most existing work published nowadays focuses on new algorithms and applications, stressing on performance rather than on usability and interpretability of real users. We envision that it is now the time to go beyond performance and focus on practical value, bridging the gap between achievements reported by the academia and the real-world problems faced by practitioners in their respective sectors [235]. For this purpose, and in accordance with recent studies [236, 237], Big Data visualization must enter the XAI arena, and help depicting highly dimensional explanations of outputs produced by bio-inspired models in an understandable manner. For this to occur, we foresee that XAI functionalities currently underway in the XAI research field should grow in mature and adapted suitably to deal with models distributed over computing nodes, each learning from different data silos. Specifically, the multimodality of data present in a significant segment of Big Data applications (those capturing data over both space and time, e.g., Smart Cities, transport, Earth observation) requires a new generation of explainability tools that allow human reasoning of patterns and explanations held over such domains simultaneously [238].

4.5 Big data governance and security

Another challenge emerges from all those activities necessary for the data to be correctly and fairly managed, secured and traced, which is called data governance. The characteristics of modern bio-inspired Deep Learning models—in particular, their capability to ingest and fuse different information flows along the learning process—usually pose a severe threat to data governance approaches, specially in what refers to privacy regulation and informed consent. Enhanced governance techniques and tools are required to help preserve the autonomy and rights of individuals to control their personal information, and to guarantee that protected data remains as such over the entire Big Data cycle. There are already works focused on studying the maintenance of privacy in the analysis of personal data [239], and the achievement of traceability of the data flow during the analysis process [80, 240]. It is undeniable that techniques such as differential privacy, federated learning and homomorphic encryption are expected to play a major role in Big Data governance for years to come. However, a question remains whether current bio-inspired computation techniques will smoothly accommodate the assumptions and restrictions imposed by these upsurging privacy-preserving methods.

A related research direction is that of security. In recent years, a vibrant activity has been noted around the development of algorithms for ensuring confidentiality, integrity, and availability in complex data-based systems. It is a consolidated fact that the existing cyber-infrastructure has numerous inherent limitations that make the maintenance of the current network security devices not scale well, and provide the adversary with asymmetric advantages. For example cybersecurity, with problems such as spam filtering [241] or intrusion detection in real time [242–244], is a research area in which numerous studies are undertaken trying to adapt the advantages of bio-inspired computation to this kind of systems. The reality is that security is an indispensable and complex requirement in any system, for which bio-inspired approaches can yield a competitive advantage. This claim can be easily confirmed by reviewing the current literature, where bio-inspired algorithms are a promising approach currently yielding great results in Cloud Computing environments [245]. The huge amount of logging information generated by complex Big Data infrastructures is, without a doubt, a rich substrate for detecting, identifying and counteracting security threats. The self-organizing nature of bio-inspired computation can provide the required level of robustness and resilience against such threats, specially those inspired by artificial immune systems for authentication and access control systems [246], evolutionary algorithms as constituent parts of intrusion detection systems relying on predictive modeling [247], or swarm intelligence methods for forensic analysis [248]. The record of successes around the application of bio-inspired methods to the security of complex networked systems is a motivational evidence towards embracing them massively in the Big Data realm.

5 Conclusions and outlook

We live in the era of digitization, which has caused an explosion of data in sectors that had traditionally lagged behind in the adoption of information and communication technologies. Consequently, multiple opportunities to generate value from data have spawned in almost all sectors. In this context, Big Data encompasses all tools and technologies that support the efficient materialization of data analysis when produced at volumes, rates and heterogeneity levels that cannot be managed by traditional means. Big Data systems are being increasingly adopted by the enterprises exploiting applications to manage data-driven processes, practices and systems in a business wide context. Specifically, Big Data systems and their underlying applications empower enterprises with analytical decision making for optimizing organizational productivity and competitiveness.

Despite the above benefits, the stringent operational conditions under which Big Data platforms operate demand several capabilities to their underlying processes, technologies and algorithms. Among them, in this survey we have focused on adaptability to data changes, scalability, computational efficiency, flexibility, integrability and uncertainty modeling. All these requirements address renowned issues arising from different phases of the Big Data life cycle. In this regard, we have stressed on the capital role that bio-inspired computation can play for Big Data technologies to acquire and effectively provide such functionalities. Indeed, modeling, simulation and optimization tasks can be formulated at different phases of the life cycle wherein biologically inspired methods have been applied. To properly inform the audience about the history of bio-inspired Big Data, we have performed a critical literature analysis along different axis: i) the Big Data technology that benefits from the application of bio-inspired methods (infrastructure, NoSQL database technology, network and parallel/distributed computing model); and ii) the Big Data life cycle phase in question (data fusion, storage, processing, learning and visualization). Relevant references have been thoroughly discussed, unveiling research trends and niches that remain open in the field.

As a result of our critical examination of the literature, we have outlined several research directions that may effectively deal with the main challenges in bio-inspired Big Data. Three of them stand out as those that deserve more research efforts in years to come:

- Common methodological grounds in the proposal of new bio-inspired algorithms for Big Data, including the adoption of good practices and recommendations to ensure their scientific and practical value.
- An explicit consideration of complexity in the design of new algorithms, specially those for real-time environments, avoiding at all means the use of the term Big Data to refer to problems and scenarios that do not correspond to the expected scales of this paradigm.
- A close look at the possibilities brought by avant-garde research areas in bio-inspired computation, such as XAI as a core element adding value to the data visualization phase of the Big Data life cycle.

This survey intends to serve as a smooth entry point for practitioners and newcomers interested in performing research around bio-inspired Big Data technologies. Inspirational behaviors behind bio-inspired computation techniques accumulate thousands of years of accumulated experience in addressing complex modeling, simulation and optimization tasks. It is straightforward to think that the scales, variability and uncertainty of problems tackled nowadays by Big Data technologies should leverage the capabilities offered by bio-inspired methods. Nature knows how to best adapt to changes, scale up nicely under environmental pressure and resiliently react against threats. Bio-inspired Big Data is, on balance, a natural choice.

Acknowledgements This work has received funding support from the Basque Government (*Eusko Jaurlaritza*) through the Consolidated Research Group MATHMODE (IT1294-19), EMAITEK and ELKARTEK programs. D. Camacho also acknowledges support from the Spanish Ministry of Science and Education under PID2020-117263GB-100 grant (FightDIS), the Comunidad Autónoma de Madrid under S2018/TCS-4566 grant (CYNAMON), and the CHIST-ERA 2017 BDSI PACMEL Project (PCI2019-103623, Spain).

Author contributions Conceptualization was performed by A. I. Torre-Bastida and J. Del Ser. Methodology was performed by A. I. Torre-Bastida, J. Díaz-de-Arcaya and J. Del Ser. Formal analysis, investigation and writing–original draft preparation were performed by A. I. Torre, J. Díaz-de-Arcaya and E. Osaba. Writing—review and editing was performed by K. Muhammad, D. Camacho and J. Del Ser. Funding acquisition was performed by D. Camacho and J. Del Ser. Supervision was performed by J. Del Ser.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Beyer MA, Laney D (2012) The importance of 'big data': a definition. Stamford, CT: Gartner, pp 2014–2018
2. Ward JS, Barker A (2013) Undefined by data: a survey of big data definitions. arXiv preprint arXiv:1309.5821
3. Tidke B, Mehta R (2018) A comprehensive review and open challenges of stream big data. Soft computing: theories and applications. Springer, Berlin, pp 89–99
4. Memishi B, Ibrahim S, Pérez MS, Antoniu G (2016) Fault tolerance in Map-Reduce: A survey. Resource Management for Big Data Platforms. Springer, Berlin, pp 205–240
5. Del Ser J, Osaba E, Molina D, Yang X-S, Salcedo-Sanz S, Camacho D, Das S, Suganthan PN, Coello CAC, Herrera F (2019) Bio-inspired computation: Where we stand and what′s next. Swarm Evolut Comput 48:220–250
6. Dorigo M, Birattari M, Stutzle T (2006) Ant colony optimization. IEEE Comput Intell Mag 1(4):28–39
7. Karaboga D, Basturk B (2007) A powerful and efficient algorithm for numerical function optimization: artificial bee colony (abc) algorithm. J Global Optim 39(3):459–471
8. Holland JH (1992) Genetic algorithms. Sci Am 267(1):66–73
9. Salcedo-Sanz S (2016) Modern meta-heuristics based on nonlinear physics processes: A review of models and design procedures. Phys Rep 655:1–70
10. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
11. Minxian X, Tian W, Buyya R (2017) A survey on load balancing algorithms for virtual machines placement in cloud computing. Concurr Comput: Practice Experience 29(12):e4123
12. Hariri RH, Fredericks EM, Bowers KM (2019) Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data 6(1):44
13. Melita L, Ganapathy G, Prakash P (2010) Bio-inspired algorithms for query optimization in biological databases. In 2010 2nd International Conference on Computer Technology and Development, pp 370–373. IEEE
14. Abdul-Wahid S, Andonie R, Lemley J, Schwing J, Widger J (2007) Adaptive distributed database replication through colonies of pogo ants. In IEEE International Parallel and Distributed Processing Symposium, pp 1–8. IEEE
15. Zulqar ALI, Kiran HM, Shahzad W (2018) Evolutionary algorithms for query optimization in distributed database sys-tems: A review. ADCAIJ Adv Distrib Comput Artif Intell J 7(3):115–128
16. Abualigah L, Diabat A (2020) A novel hybrid antlion optimization algorithm for multi-objective task scheduling problems in cloud computing environments. Cluster Computing, pp 1–19
17. Manikandan RPS, Kalpana AM (2019) Feature selection using fish swarm optimization in big data. Cluster Comput 22(5):10825–10837
18. Pop F, Negru C, Ciolofan SN, Mocanu M, Cristea V (2016) Optimizing intelligent reduction techniques for big data. Big Data Optimization: Recent Developments and Challenges. Springer, Berlin, pp 49–70
19. Hausler S, Chen Z, Hasselmo ME, Milford M (2020) Bio-inspired multi-scale fusion. Biological Cybernetics, 1–21
20. Iqbal R, Doctor F, More B, Mahmud S, Yousuf U (2020) Big data analytics: Computational intelligence techniques and application areas. Technol Forecast Soc Chang 153:119253
21. Laney D (2012) Deja vvvu: others claiming gartners construct for big data. Gartner Blog 14:1
22. Faraway JJ, Augustin NH (2018) When small data beats big data. Stat Probab Lett 136:142–145
23. Wang L, Shen J (2014) Bio-inspired cost-effective access to big data. ISNGI sponsors, pp 243
24. Mahdavi S, Shiri ME, Rahnamayan S (2015) Metaheuristics in large-scale global continues optimization: A survey. Inf Sci 295:407–428
25. Sung W-T, Tsai M-H (2012) Data fusion of multi-sensor for iot precise measurement based on improved pso algorithms. Comput Math Appl 64(5):1450–1461
26. Wang J, Tawose O, Jiang L, Zhao D (2019) A new data fusion algorithm for wireless sensor networks inspired by hesitant fuzzy entropy. Sensors 19(4):784
27. Forestiero A, Pizzuti C, Spezzano G (2009) Flockstream: a bio-inspired algorithm for clustering evolving data streams. In Tools with Artificial Intelligence, 2009. ICTAI'09. 21st International Conference on, pp 1–8. IEEE
28. Yang Q, Yoo S-J (2018) Optimal uav path planning: Sensing data acquisition over iot sensor networks using multi-objective bio-inspired algorithms. IEEE Access 6:13671–13684
29. Bing L, Chan KCC (2014) A fuzzy logic approach for opinion mining on large scale twitter data. In 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, pp 652–657. IEEE,
30. Ghosh G, Banerjee S, Yen NY (2016) State transition in communication under social network: An analysis using fuzzy logic and density based clustering towards big data paradigm. Future Gener Comput Syst 65:207–220
31. Wang H, Zeshui X, 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
32. Mills S, Lucas S, Irakliotis L, Rappa M, Carlson T, Perlowitz B (2012) Demystifying big data: a practical guide to transforming the business of government. TechAmerica Foundation, Washington
33. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q (2019) A comprehensive survey on transfer learning. arXiv preprint arXiv:1911.02685,
34. Jiang J (1999) Image compression with neural networks-a survey. Sig Process Image Commun 14(9):737–760
35. Rehman MH, Liew CS, Abbas A, Jayaraman PP, Wah TY, Khan SU (2016) Big Data reduction methods: a survey. Data Sci Eng 1(4):265–284
36. Ramrez-Gallego S, Krawczyk B, Garca S, Woniak M, Herrera F (2017) A survey on data preprocessing for data stream mining. Neurocomputing 239(1):39–57
37. Agrawal R, Kadadi A, Dai X, Andres F (2015) Challenges and opportunities with big data visualization. In Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems, pp 169–173,
38. Keim D, Huamin Q, Ma K-L (2013) Big-data visualization. IEEE Comput Graphics Appl 33(4):20–21
39. Sharma PP, Navdeti CP (2014) Securing big data hadoop: a review of security issues, threats and solution. Int J Comput Sci Inf Technol 5(2):2126–2131
40. Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU (2015) The rise of 'big data' on cloud computing: Review and open research issues. Inf Syst 47:98–115
41. Yadav V, Natesha BV, Guddeti RMR (2019) Ga-pso: Service allocation in fog computing environment using hybrid bio-inspired algorithm. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON), pp 1280–1285. IEEE
42. Masdari M, Gharehpasha S, Ghobaei-Arani M, Ghasemi V (2019) Bio-inspired virtual machine placement schemes in cloud

computing environment: taxonomy, review, and future research directions. Cluster Computing, 1–31

43. Gamal M, Rizk R, Mahdi H, Elhady B (2019) Bio-inspired based task scheduling in cloud computing. Machine Learning Paradigms: Theory and Application. Springer, Berlin, pp 289–308
44. Mthunzi SN, Benkhelifa E, Bosakowski T, Hariri S (2019) A bio-inspired approach to cyber security. Machine Learning for Computer and Cyber Security: Principle, Algorithms, and Practices, p 75
45. Lorbeer B, Deutsch T, Ruppel P, Küpper A (2019) Anomaly detection with hmm gauge likelihood analysis. In 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), pp 1–8. IEEE
46. Kar AK (2016) Bio inspired computing-a review of algorithms and scope of applications. Expert Syst Appl 59:20–32
47. Fulcher J (2008) Computational intelligence: an introduction. Computational intelligence: a compendium. Springer, Berlin, pp 3–78
48. Malone TW, Bernstein MS (2015) Handbook of collective intelligence. MIT Press, Cambridge
49. Dasgupta D, Attoh-Okine N (1997) Immunity-based systems: A survey. In 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation, vol 1, pp 369–374. IEEE
50. Hecht-Nielsen R (1988) Neurocomputing: picking the human brain. IEEE Spectr 25(3):36–41
51. Yang XS, Deb S (2009) Cuckoo search via lévy flights. In 2009 World congress on nature & biologically inspired computing (NaBIC), pp 210–214. IEEE
52. Kennedy J, Eberhart R (1995) Particle swarm optimization. In Proceedings of ICNN'95-International Conference on Neural Networks, vol 4, pp 1942–1948. IEEE
53. Beale HD, Demuth HB, Hagan MT (1996) Neural network design. Pws, Boston
54. Hassoun MH et al (1995) Fundamentals of artificial neural networks. MIT press, Cambridge
55. Bäck T, Fogel DB, Michalewicz Z (1997) Handbook of evolutionary computation. CRC Press, Florida
56. Eiben AE, Smith JE et al (2003) Introduction to evolutionary computing. Springer, Berlin
57. James Kennedy (2006) Swarm intelligence. Handbook of nature-inspired and innovative computing. Springer, Berlin, pp 187–219
58. Sugeno M, Asai K, Terano T (1992) Fuzzy systems theory and its applications. Tokyo Institute of Technology, Meguro
59. Kruse R, Gebhardt JE, Klowon F (1994) Foundations of fuzzy systems. Wiley, New Jersy
60. Del Ser J, Osaba E, Sanchez-Medina JJ, Fister I (2019) A long road ahead. IEEE Transactions on Intelligent Transportation Systems, Bioinspired computational intelligence and transportation systems
61. Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H (2018) State-of-the-art in arti-cial neural network applications: A survey. Heliyon 4(11):e00938
62. Banzhaf W, Nordin P, Keller RE, Francone FD (1998) Genetic programming. Springer, Berlin
63. Yao X, Liu Y, Lin G (1999) Evolutionary programming made faster. IEEE Trans Evol Comput 3(2):82–102
64. Beyer H-G, Schwefel H-P (2002) Evolution strategies-a comprehensive introduction. Nat Comput 1(1):3–52
65. Price K, Storn RM, Lampinen JA (2006) Differential evolution: a practical approach to global optimization. Springer Science & Business Media, NY
66. Reynolds RG (1994) An introduction to cultural algorithms. In Proceedings of the third annual conference on evolutionary programming. World Scientific pp 131–139
67. Paredis J (1995) Coevolutionary computation. Artif life 2(4):355–375
68. An J, She J, Chen H, Wu M (2018) Applications of evolutionary computation and artificial intelligence in metallurgical industry. General Conference on Emerging Arts of Research on Management and Administration. Springer, Berlin, pp 77–87
69. Smith SL (2018) Medical applications of evolutionary computation. In Proceedings of the Genetic and Evolutionary Computation Conference Companion, pp 1141–1169
70. Cuevas E, Zaldívar D, Perez-Cisneros M (2016) Applications of evolutionary computation in image processing and pattern recognition. Springer, Berlin
71. Yang X-S et al (2008) Firefly algorithm. Nature-inspired Metaheuristic Alg 20:79–90
72. Chu SC, Tsai PW, Pan JS (2006) Cat swarm optimization. Pacific Rim international conference on artificial intelligence. Springer, Berlin, pp 854–858
73. Birbil SI, Fang SC (2003) An electromagnetism-like mechanism for global optimization. J Global Opt 25(3):263–282
74. Kaveh A, Khayatazad M (2012) A new meta-heuristic method: ray optimization. Comput Struct 112:283–294
75. Beiranvand H, Rokrok E (2015) General relativity search algorithm: a global optimization approach. Int J Comput Intell Appl 14(03):1550017
76. Ahmadi-Javid A (2011) Anarchic society optimization: A human-inspired method. In 2011 IEEE Congress of Evolutionary Computation (CEC), pp 2586–2592. IEEE
77. Ghosh S, Razouqi Q, Schumacher HJ, Celmins A (1998) A survey of recent advances in fuzzy logic in telecommunications networks and new challenges. IEEE Trans Fuzzy Syst 6(3):443–447
78. Wang XZD, Keane JA (2006) A survey of hierarchical fuzzy systems. Int J Comput Cogn 4(1):18–29
79. Kar S, Das S, Ghosh PK (2014) Applications of neuro fuzzy systems: A brief review and future outline. Appl Soft Comput 15:243–259
80. Gill SS, Buyya R (2019) Bio-inspired algorithms for big data analytics: a survey, taxonomy, and open challenges. Big Data Analytics for Intelligent Healthcare Management. Elsevier, Amsterdam, pp 1–17
81. Rehman MH, Yaqoob I, Salah K, Imran M, Jayaraman PP, Perera C (2019) The role of big data analytics in industrial internet of things. Future Gener Comput Syst 99:247–259
82. Villars RL, Olofson CW, Eastwood M (2011) Big data: What it is and why you should care. White Paper, IDC 14:1–14
83. Brocco A, Baumgart I (2012) A framework for a comprehensive evaluation of ant-inspired peer-to-peer protocols. In Parallel, Distributed and Network-Based Processing (PDP), 2012 20th Euromicro International Conference on, pp 303–310. IEEE
84. Djenouri Y, Djenouri D, Habbas Z, Belhadi A (2018) How to exploit high performance computing in population-based metaheuristics for solving association rule mining problem. Distributed and Parallel Databases, pp 1–29
85. Hager G, Wellein G (2010) Introduction to high performance computing for scientists and engineers. CRC Press, Florida
86. Pupykina A, Agosta G (2019) Survey of memory management techniques for hpc and cloud computing. IEEE Access 7:167351–167373
87. Rauf U (2018) A taxonomy of bio-inspired cyber security approaches: existing techniques and future directions. Arab J Sci Eng 43(12):6693–6708
88. Barbagallo D, Di Nitto E, Dubois DJ, Mirandola R (2009) A bio-inspired algorithm for energy optimization in a self-organizing data center. International Workshop on Self-Organizing Architectures. Springer, Berlin, pp 127–151

89. Agostinho L, Feliciano G, Olivi L, Cardozo E, Guimaraes E (2011) A bio-inspired approach to provisioning of virtual resources in federated clouds. In Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on, pp 598–604. IEEE
90. Rocha LA, Cardozo E (2014) A hybrid optimization model for green cloud computing. In Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, (pp 11–20). IEEE Computer Society
91. Pacini E, Mateos C, Garino CG (2015) Balancing throughput and response time in online scientific clouds via ant colony optimization (sp2013/2013/00006). Adv Eng Softw 84:31–47
92. Perumal B, Murugaiyan A (2016) A firefly colony and its fuzzy approach for server consolidation and virtual machine placement in cloud datacenters. Adv Fuzzy Syst 2016:5
93. Thaha AF, Singh M, Amin AHM, Ahmad NM, Kannan S (2014) Hadoop in openstack: Data-location-aware cluster provisioning. In Information and Communication Technologies (WICT), 2014 Fourth World Congress on, pp 296–301. IEEE
94. Pires FL, Barán B (2013) Multi-objective virtual machine placement with service level agreement: A memetic algorithm approach. In 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing, pp 203–210. IEEE
95. Seddigh M, Taheri H, Sharifian S (2015) Dynamic prediction scheduling for virtual machine placement via ant colony optimization. In 2015 Signal Processing and Intelligent Systems Conference (SPIS), pp 104–108. IEEE
96. Kumar KSS, Jaisankar N (2017) Towards data centre resource scheduling via hybrid cuckoo search algorithm in multi-cloud environment. Int J Intell Enterprise 4(1–2):21–35
97. Xavier VMA, Annadurai S (2019) Chaotic social spider algorithm for load balance aware task scheduling in cloud computing. Cluster Comput 22(1):287–297
98. Singh P, Dutta M, Aggarwal N (2017) A review of task scheduling based on meta-heuristics approach in cloud computing. Knowl Inf Syst 52(1):1–51
99. Althebyan Q, Jararweh Y, Yaseen Q, AlQudah O, Al-Ayyoub M (2015) Evaluating map reduce tasks scheduling algorithms over cloud computing infrastructure. Concurr Comput: Practice Experience 27(18):5686–5699
100. Milan ST, Rajabion L, Ranjbar H, Navimipoir NJ (2019) Nature inspired meta-heuristic algorithms for solving the load-balancing problem in cloud environments. Computers & Operations Research
101. Kumar KP, Ragunathan T, Vasumathi D, Prasad PK (2020) An efficient load balancing technique based on cuckoo search and firefly algorithm in cloud. Algorithms, 423
102. Javaid N, Butt AA, Latif K, Rehman A (2019) Cloud and fog based integrated environment for load balancing using cuckoo levy distribution and flower pollination for smart homes. In 2019 International Conference on Computer and Information Sciences (ICCIS), pp 1–6. IEEE
103. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M et al (2016) Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467,
104. Chen T, Li M, Li Y, Lin M, Wang N, Wang M, Xiao T, Xu B, Zhang C, Zhang Z (2015) Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274
105. Pop F, Iosup A, Prodan R (2018) High performance scheduling for heterogeneous distributed systems, Hps-hds
106. Wang L, Shen J (2012) Towards bio-inspired cost minimisation for data-intensive service provision. In Services Economics (SE), 2012 IEEE First International Conference on, pp 16–23. IEEE,
107. Chávez F, Fernández F, Benavides C, Lanza D, Villegas J, Trujillo L, Olague G, Román G (2016) Ecj+ hadoop: An easy way to deploy massive runs of evolutionary algorithms. European Conference on the Applications of Evolutionary Computation. Springer, Berlin, pp 91–106
108. Luke S, Panait L, Balan G, Paus S, Skolicki Z, Bassett J, Hubley R, Chircop A (2006) Ecj: A java-based evolutionary computation research system. Downloadable versions and documentation can be found at the following https://cs.gmu.edu/eclab/projects/ecj
109. Stojmenovic I, Wen S (2014) The fog computing paradigm: Scenarios and security issues. In 2014 federated conference on computer science and information systems, pp 1–8. IEEE
110. Shi W, Cao J, Zhang Q, Li Y, Lanyu X (2016) Edge computing: Vision and challenges. IEEE Internet Things J 3(5):637–646
111. Kabirzadeh S, Rahbari D, Nickray M (2017) A hyper heuristic algorithm for scheduling of fog networks. 2017 21st Conference of Open Innovations Association (FRUCT), pp 148–155
112. Diaz-de AJ, Minon R, Torre-Bastida AI (2019) Towards an architecture for big data analytics leveraging edge/fog paradigms. In Proceedings of the 13th European Conference on Software Architecture-Vol 2, pp 173–176
113. Zou X, Cao J, Guo Q, Wen T (2017) A novel network security algorithm based on improved support vector machine from smart city perspective. Computers & Electrical Engineering
114. Frey M, Große F, Günes M (2013) libara: a framework for simulation and testbed based studies on ant routing algorithms in wireless multi-hop networks. In Proceedings of the 7th International Conference on Performance Evaluation Methodologies and Tools, pp 304–309
115. Frey M, Günes M (2015) Follow the pheromone trail: on studying ant routing algorithms in simulation and wireless testbeds. In Proceedings of the 18th Symposium on Communications & Networking, pp 68–74
116. Singh K, Kaur R (2014) Hadoop: addressing challenges of big data. In 2014 IEEE International Advance Computing Conference (IACC), pp 686–689. IEEE
117. Prasad YL (2016) Big data analytics made easy. Notion Press, Chennai
118. Cattell R (2011) Scalable sql and nosql data stores. Acm Sigmod Record 39(4):12–27
119. Hajeer MH, Dasgupta D (2017) Handling big data using a data-aware hdfs and evolutionary clustering technique. IEEE Transactions on Big Data
120. Nowosielski A, Kowalski PA, Kulczycki P (2015) The column-oriented database partitioning optimization based on the natural computing algorithms. In Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on, pp 1035–1041. IEEE
121. Woźniak M, Gabryel M, Nowicki RK, Nowak BA (2016) An application of firefly algorithm to position traffic in nosql database systems. Knowledge, Information and Creativity Support Systems. Springer, Berlin, pp 259–272
122. Wilson J, Chang P, Yim S, Rosen DW (2009) Developing a bio-inspired design repository using ontologies. In ASME 2009 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers, pp 799–808
123. Derrar H, Ahmed-Nacer M, Boussaid O (2012) Particle swarm optimisation for data warehouse logical design. Int J Bio-Inspired Comput 4(4):249–257
124. Nowosielski A, Kowalski PA, Kulczycki P (2017) A database performance polynomial multiple regression model. In 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), pp 743–474. IEEE

125. Ioannidis YE (1996) Query optimization. ACM Comput Surv (CSUR) 28(1):121–123
126. Rani R (2016) An efficient bio-inspired approach to generate distributed query plans. In 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), pp 1–5. IEEE
127. Rani R (2019) Distributed query processing optimization in wireless sensor network using artificial immune system. Computational Intelligence in Sensor Networks. Springer, Berlin, pp 1–23
128. Djenouri Y, Fournier-Viger P, Lin JC-W, Djenouri D, Belhadi A (2019) Gpu-based swarm intelligence for association rule mining in big databases. Intell Data Anal 23(1):57–76
129. Amraoui H, Mhamdi F, Elloumi M (2019) Association rule mining using discrete jaya algorithm
130. Jordan J, Cheng W, Scheuermann B (2017) Advancing dynamic evolutionary optimization using in-memory database technology. In European Conference on the Applications of Evolutionary Computation, pp 156–172. Springer
131. Dean J, Ghemawat S (2010) Mapreduce: a flexible data processing tool. Commun ACM 53(1):72–77
132. Rhine R, Bhuvan NT (2016) Locality aware mapreduce. Innovations in Bio-Inspired Computing and Applications. Springer, Berlin, pp 221–228
133. Pradeep D, Sundar C (2020) Qaoc: Novel query analysis and ontology-based clustering for data management in hadoop. Future Gener Comput Syst 108:849–860
134. Triguero I, Galar M, Merino D, Maillo J, Bustince H, Herrera F (2016) Evolutionary undersampling for extremely imbalanced big data classification under apache spark. In 2016 IEEE Congress on Evolutionary Computation (CEC), pp 640–647. IEEE
135. Hans N, Mahajan S, Omkar S (2015) Big data clustering using genetic algorithm on hadoop mapreduce. Int J Sci Technol Res 4:58–62
136. Teijeiro D, Pardo XC, González P, Banga JR, Doallo R (2016) Implementing parallel differential evolution on spark. In European Conference on the Applications of Evolutionary Computation, pp 75–90. Springer
137. Joshi RR, Mulay P, Chaudhari A (2020) Smdicfba: Software model for distributed incremental closeness factor based algorithms. In A Journey Towards Bio-inspired Techniques in Software Engineering, pp 1–28. Springer
138. Sun D, Tang H (2017) Fast-ffa: a fast online scheduling approach for big data stream computing with future features-aware. Int J Bio-Inspired Comput 10(3):205–217
139. Müller-Schloer C, Schmeck H, Ungerer T (2011) Organic computing-A paradigm shift for complex systems. Springer Science & Business Media, Berlin
140. Sadasivam GS, Selvaraj D (2010) A novel parallel hybrid pso-ga using mapreduce to schedule jobs in hadoop data grids. In Nature and Biologically Inspired Computing (NaBIC), 2010 Second World Congress on, pp 377–382. IEEE
141. Csorba MJ, Meling H, Heegaard PE (2011) A bio-inspired method for distributed deployment of services. New Gener Comput 29(2):185
142. Musikapun P, Pongcharoen P (2012) Solving multi-stage multi-machine multi-product scheduling problem using bat algorithm. In 2nd international conference on management and artificial intelligence, vol 35, pp 98–102. IACSIT Press Singapore
143. Kaur T, Chana I (2018) Greensched: An intelligent energy aware scheduling for deadline-and-budget constrained cloud tasks. Simul Model Pract Theory 82:55–83
144. Arunarani AR, Manjula D (2016) Babc task scheduler: hybridisation of bat and artificial bee colony for deadline constrained task scheduling. Inte J Business Intell Data Mining 11(4):379–399
145. Namasudra S, Roy P, Vijayakumar P, Audithan S, Balusamy B (2017) Time efficient secure dna based access control model for cloud computing environment. Future Gener Comput Syst 73:90–105
146. Cui Z, Cao Y, Cai X, Cai J, Chen J (2018) Optimal leach protocol with modified bat algorithm for big data sensing systems in internet of things. Journal of Parallel and Distributed Computing
147. Alshehri MD, Hussain FK, Hussain OK (2018) Clustering-driven intelligent trust management methodology for the internet of things (citm-iot). Mobile Networks and Applications, pp 1–13
148. Ogiela MR, Ko H (2018) Bio-inspired and cognitive approaches in cryptography and security applications. Concurrency and Computation: Practice and Experience, 30(2)
149. Cheng S, Liu B, Ting TO, Qin Q, Shi Y, Huang K (2016) Survey on data science with population-based algorithms. Big Data Anal 1(1):3
150. Brandejsky T (2020) Preconditions of gpa-es algorithm application to big data. Artificial Intelligence and Evolutionary Computations in Engineering Systems. Springer, Berlin, pp 485–492
151. Nordmann B (2012) Bio-inspired computing, information swarms, and the problem of data fusion. In Technological Innovations in Sensing and Detection of Chemical, Biological, Radiological, Nuclear Threats and Ecological Terrorism, (pp 35–44). Springer
152. Hall DL, Llinas J (2017) Multisensor data fusion. Handbook of multisensor data fusion. CRC Press, Florida, pp 21–34
153. Kim H, Suh D (2018) Hybrid particle swarm optimization for multi-sensor data fusion. Sensors 18(9):2792
154. Aboud A, Fdhila R, Alimi AM (2016) Mopso for dynamic feature selection problem based big data fusion. In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp 003918–003923. IEEE
155. Dong-Lan L, Xin L, Hao Y, Wen-Ting W, Xiao-Hong Z, Jian-Fei C (2017) A multilevel deep learning method for data fusion and anomaly detection of power big data. In 3rd Annual International Conference on Electronics, Electrical Engineering and Information Science (EEEIS 2017). Atlantis Press,
156. Zhang Y, Zhao P, Zhao Y, Yan Z, Li B (2018) The model research on location fusion algorithm with big data selection and accuracy correction. DEStech Transactions on Computer Science and Engineering, (wicom),
157. Yini Z Ping Z, Yan Z, Yang M, Li B (2019) The research for a kind of information fusion model based on bp neural network with multi position sources and big data selection. In 2019 IEEE 2nd International Conference on Electronics Technology (ICET), pp 619–623. IEEE
158. Chandra A, Yao X (2006) Ensemble learning using multi-objective evolutionary algorithms. J Math Model Alg 5(4):417–445
159. Sagi O, Rokach L (2018) Ensemble learning: A survey. Wiley Interdiscip Rev: Data Mining Knowl Discovery 8(4):e1249
160. Triguero I, Galar M, Vluymans S, Cornelis C, Bustince H, Herrera F, Saeys Y (2015) Evolutionary undersampling for imbalanced big data classification. In 2015 IEEE Congress on Evolutionary Computation (CEC), pp 715–722. IEEE
161. Piccialli F, Cuomo S, di Cola VS, Casolla G (2019) A machine learning approach for iot cultural data. Journal of Ambient Intelligence and Humanized Computing, pp 1–12
162. Lakshmanaprabu SK, Shankar K, Ilayaraja M, Nasir AW, Vijayakumar V, Chilamkurti N (2019) Random forest for big data classication in the internet of things using optimal features. Int J Machine Learn Cybern 10(10):2609–2618
163. Ruo H, Zhao H, Yantai W (2019) The methods of big data fusion and semantic collision detection in internet of thing. Cluster Comput 22(4):8007–8015

164. Liu J, Li T, Xie P, Shengdong D, Teng F, Yang X (2020) Urban big data fusion based on deep learning: An overview. Inf Fusion 53:123–133
165. Banisakher M, Omar M, Hong S, Adams J (2020) A human centric approach to data fusion in post-disaster management. J Business Manage Sci 8(1):12–20
166. Wang P, Zhang Q (2019) Train delay analysis and prediction based on big data fusion. Transp Saf Environ 1(1):79–88
167. Almasi M, Abadeh MS (2018) A new mapreduce associative classifier based on a new storage format for large-scale imbalanced data. Cluster Comput 21(4):1821–1847
168. Wei H (2016) A bio-inspired integration method for object semantic representation. J Artif Intell Soft Comput Res 6(3):137–154
169. Ahmed AM, Kong X, Liu L, Xia F, Abolfazli S, Sanaei Z, Tolba A (2017) Bodmas: bio-inspired selfishness detection and mitigation in data management for ad-hoc social networks. Ad Hoc Netw 55:119–131
170. Di Stefano A, La Corte A, Lió P, Scatá M (2016) Bio-inspired ict for big data management in healthcare. Intelligent Agents in Data-intensive Computing. Springer, Berlin, pp 1–26
171. Krestinskaya O, James AP (2016) Bioinspired memory model for htm face recognition. In 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp 1528–1532. IEEE
172. Gai K, Qiu M, Zhao H (2016) Cost-aware multimedia data allocation for heterogeneous memory using genetic algorithm in cloud computing. IEEE Transactions on Cloud Computing
173. Shabeera TP, Kumar SDM, Salam SM, Krishnan KM (2017) Optimizing vm allocation and data placement for data-intensive applications in cloud using aco metaheuristic algorithm. Eng Sci Technol, Int J 20(2):616–628
174. Ullah F, Yahya KM (2012) A new data compression technique using an evolutionary programming approach. International multi topic conference. Springer, Berlin, pp 524–531
175. Ogiela MR, Ogiela L (2015) Bio-inspired approaches for secret data sharing techniques. In 2015 International conference on intelligent informatics and biomedical sciences (ICIIBMS), pp 75–78. IEEE
176. García S, Luengo J, Herrera F (2015) Data preprocessing in data mining. Springer, Berlin
177. Manikandan RPS, Kalpana AM (2017) Feature selection using fish swarm optimization in big data. Cluster Computing, 1–13
178. Fong S, Wong R, Vasilakos AV (2016) Accelerated pso swarm search feature selection for data stream mining big data. IEEE Trans Serv Comput 9(1):33–45
179. Villar-Rodriguez E, Gonzalez-Pardo A, Del Ser J, Bilbao MN, Salcedo-Sanz S (2016) A novel adaptive density-based aco algorithm with minimal encoding redundancy for clustering problems. In Evolutionary Computation (CEC), 2016 IEEE Congress on, pp 3139–3145. IEEE
180. Reddy KSS and Bindu CS (2017) A review on density-based clustering algorithms for big data analysis. In 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pp 123–130. IEEE
181. Eerlapati A (2017) Electromagnetsim Based K-Means Clustering for Big Data. Ph.D thesis, Texas A&M University-Corpus Christi
182. Kim S-S, McLoone S, Byeon J-H, Lee S, Liu H (2017) Cognitively inspired artificial bee colony clustering for cognitive wireless sensor networks. Cogn Comput 9(2):207–224
183. Kulkarni O, Jena S, Sanjay CH (2019) Fractional fuzzy clustering and particle whale optimization-based mapreduce framework for big data clustering. Journal of Intelligent Systems, 1(ahead-of-print),
184. Pan QK, Tasgetiren MF, Suganthan PN, Chua TJ (2011) A discrete artificial bee colony algorithm for the lot-streaming flow shop scheduling problem. Inf Sci 181(12):2455–2468
185. Saritha K, Abraham S (2017) Prediction with partitioning: Big data analytics using regression techniques. In Networks & Advances in Computational Technologies (NetACT), 2017 International Conference on, pp 208–214. IEEE
186. Banerjee S, Badr Y (2018) Evaluating decision analytics from mobile big data using rough set based ant colony. Mobile Big Data. Springer, Berlin, pp 217–231
187. Cao J, Cui H, Shi H, Jiao L (2016) Big data: A parallel particle swarm optimization-back-propagation neural network algorithm based on mapreduce. PLoS ONE 11(6):e0157551
188. Fan D, Lee J (2019) A hybrid mechanism of particle swarm optimization and differential evolution algorithms based on spark. KSII Trans Internet Inf Syst (TIIS) 13(12):5972–5989
189. Jedrzejowicz P, Wierzbowska I (2020) Apache spark as a tool for parallel population-based optimization. In Intelligent Decision Technologies 2019, pp 181–190. Springer: Berlin
190. Maqbool F, Razzaq S, Lehmann J, Jabeen H (2019) Scalable distributed genetic algorithm using apache spark (s-ga). In International Conference on Intelligent Computing, (pp 424–435). Springer
191. Guigou F, Collet P, Parrend P (2017) The artificial immune ecosystem: a bio-inspired meta-algorithm for boosting time series anomaly detection with expert input. In European Conference on the Applications of Evolutionary Computation. Springer, pp 573–588
192. Radhakrishna V, Aljawarneh SA, Kumar PV, Janaki V (2017) A novel fuzzy similarity measure and prevalence estimation approach for similarity profiled temporal association pattern mining. Future Generation Computer Systems,
193. Lu L, Liang Y, He Y, Yang C (2012) Danger theory: a new approach in big data analysis. In IET Conference Proceedings. The Institution of Engineering & Technology
194. Zöller MA, Huber MF (2019) Survey on automated machine learning. arXiv preprint arXiv:1904.12054, 9
195. He X, Zhao K, Chu X (2019) Automl: A survey of the state-of-the-art. arXiv preprint arXiv:1908.00709
196. Zöller MA, Huber MF (2019) Benchmark and survey of automated machine learning frameworks. arXiv preprint arXiv:1904.12054,
197. Gritsenko A, Akusok A, Baek S, Miche Y, Lendasse A (2017) Extreme learning machines for visualization+ r: Mastering visualization with target variables. Cognitive Computation, pp 1–14
198. Andreou AG, Figliolia T, Sanni K, Murray TS, Tognetti G, Mendat DR, Molin JL, Villemur M, Pouliquen PO, Julian P et al. (2016) Bio-inspired system architecture for energy efficient, bigdata computing with application to wide area motion imagery. In 2016 IEEE 7th Latin American Symposium on Circuits & Systems (LASCAS), pp 1–6. IEEE,
199. Wang T, Liu W, Zhao J, Guo X, Terzija V (2020) A rough set-based bio-inspired fault diagnosis method for electrical substations. Int J Electr Power Energy Syst 119:105961
200. Munir K, de Ramón-Fernández A, Iqbal S, Javaid N (2019) Neuroscience patient identification using big data and fuzzy logic-an alzheimer's disease case study. Expert Syst Appl 136:410–425
201. Yan H, Wan J, Zhang C, Tang S, Hua Q, Wang Z (2018) Industrial big data analytics for prediction of remaining useful life based on deep learning. IEEE Access 6:17190–17197
202. Diez-Olivan A, Del Ser J, Galar D, Sierra B (2019) Data fusion and machine learning for industrial prognosis: Trends and perspectives towards industry 4.0. Information Fusion 50:92–111

203. Aghelpour P, Bahrami-Pichaghchi H, Kisi O (2020) Comparison of three different bio-inspired algorithms to improve ability of neuro fuzzy approach in prediction of agricultural drought, based on three different indexes. Comput Electron Agric 170:105279
204. Yadav A, Vishwakarma DK (2020) A comparative study on bio-inspired algorithms for sentiment analysis. Cluster Computing, pp 1–21
205. Bello-Orgaz G, Jung JJ, Camacho D (2016) Social big data: Recent achievements and new challenges. Information Fusion 28:45–59
206. Camacho D, Panizo-LLedot A, Bello-Orgaz G, Gonzalez-Pardo A, Cambria E (2020) The four dimensions of social network analysis: An overview of research methods, applications, and software tools. Information Fusion 63:1–33
207. Raychaudhuri A, De D (2020) Bio-inspired algorithm for multi-objective optimization in wireless sensor network. In Nature Inspired Computing for Wireless Sensor Networks. Springer, pp 279–301
208. Zhou C, Fenzhen S, Pei T, Zhang A, Yunyan D, Luo B, Cao Z, Wang J, Yuan W, Zhu Y et al (2020) Covid-19: challenges to gis with big data. Geography Sustain 1(1):77–87
209. Wang CJ, Ng CY, Brook RH (2020) Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing. JAMA 323(14):1341–1342
210. Mann RP, Mushtaq F, White AD, Mata-Cervantes G, Pike T, Coker D, Murdoch S, Hiles T, Smith C, Berridge D et al (2016) The problem with big data: operating on smaller datasets to bridge the implementation gap. Front Publ Health 4:248
211. Bajaber F, Sakr S, Batarfi O, Altalhi A, Barnawi A (2020) Benchmarking big data systems: A survey. Comput Commun 149:241–251
212. Han R, John LK, Zhan J (2017) Benchmarking big data systems: A review. IEEE Trans Serv Comput 11(3):580–597
213. Han R, Lu X, Xu J (2014) On big data benchmarking. In Workshop on Big Data Benchmarks, Performance Optimization, and Emerging Hardware. Springer, pp 3–18
214. LaTorre A, Molina D, Osaba E, Del Ser J, Herrera F (2020) Fairness in bio-inspired optimization research: A prescription of methodological guidelines for comparing meta-heuristics. arXiv preprint arXiv:2004.09969
215. Molina D, Poyatos J, Del Ser J, García S, Hussain A, Herrera F (2020) Comprehensive taxonomies of nature-and bio-inspired optimization: Inspiration versus algorithmic behavior, critical analysis and recommendations. Cogn Comput 12:897–939
216. Talbi E-G, Hasle G (2013) Metaheuristics on gpus. J Parallel Distrib Comput 73(1):1–3
217. Essaid M, Idoumghar L, Lepagnot J, Brévilliers M (2019) Gpu parallelization strategies for metaheuristics: a survey. Int J Parallel Emergent Distrib Syst 34(5):497–522
218. Dang Y, Lin Q, Huang P (2019) Aiops: real-world challenges and research innovations. In 2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), pp 4–5. IEEE
219. Barba-González C, Nebro AJ, Benítez-Hidalgo A, García-Nieto J, Aldana-Montes JF (2020) On the design of a framework integrating an optimization engine with streaming technologies. Future Gener Comput Syst 107:538–550
220. Ahmed NO, Bhargava B (2020) Bio-inspired formal model for space/time virtual machine randomization and diversification. IEEE Transactions on Cloud Computing,
221. Ali A, Hafeez Y, Hussainn SM, Nazir MU (2020) Bio-inspired communication: A review on solution of complex problems for highly configurable systems. In 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp 1–6
222. Diaz-de AJ, Miñon R, Torre-Bastida AI, Del Ser J, Almeida A (2020) PADL: a language for the operationalization of distributed analytical pipelines over edge/fog computing environments. In accepted for its presentation in the 5th International Conference on Smart and Sustainable Technologies
223. Chintapalli S, Dagit D, Evans B, Farivar R, Graves T, Holderbaugh M, Liu Z, Nusbaum K, Patil K, Peng BJ et al (2016) Benchmarking streaming computation engines: Storm, flink and spark streaming. In 2016 IEEE international parallel and distributed processing symposium workshops (IPDPSW), pp 1789–1792. IEEE
224. Karimov J, Rabl T, Katsifodimos A, Samarev R, Heiskanen H, Markl V (2018) Benchmarking distributed stream data processing systems. In 2018 IEEE 34th International Conference on Data Engineering (ICDE), pp 1507–1518. IEEE
225. Shukla A, Chaturvedi S, Simmhan Y (2017) Riotbench: An iot benchmark for distributed stream processing systems. Concurr Comput: Practice Experience 29(21):e4257
226. Wang L, Zhan J, Luo C, Zhu Y, Yang Q, He Y, Gao W, Jia Z, Shi Y, Zhang S et al (2014) Bigdatabench: A big data benchmark suite from internet services. In 2014 IEEE 20th international symposium on high performance computer architecture (HPCA), pp 488–499. IEEE,
227. Sun D, Gao S, Liu X, Li F, Buyya R (2020) Performance-aware deployment of streaming applications in distributed stream computing systems. Int J Bio-Inspired Comput 15(1):52–62
228. Mavrovouniotis M, Li C, Yang S (2017) A survey of swarm intelligence for dynamic optimization: Algorithms and applications. Swarm Evolut Comput 33:1–17
229. Nguyen TT, Yang S, Branke J (2012) Evolutionary dynamic optimization: A survey of the state of the art. Swarm Evolut Comput 6:1–24
230. Ditzler G, Roveri M, Alippi C, Polikar R (2015) Learning in nonstationary environments: A survey. IEEE Comput Intell Mag 10(4):12–25
231. Baruah RD, Angelov P (2011) Evolving fuzzy systems for data streams: a survey. Wiley Interdiscip Rev: Data Mining Knowl Discover 1(6):461–476
232. Keim DA, Mansmann F, Schneidewind J, Thomas J, Ziegler H (2008) Visual analytics: Scope and challenges. In Visual data mining. Springer, pp 76–90
233. Arrieta AB, Diaz-Rodriguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Herrera F (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58:82–115
234. Samek W (2019) Explainable AI: interpreting, explaining and visualizing deep learning, vol 11700. Springer Nature, Berlin
235. Zhu J, Liapis A, Risi S, Bidarra R, Youngblood GM (2018) Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation. In 2018 IEEE Conference on Computational Intelligence and Games (CIG), pp 1–8. IEEE
236. Choo J, Liu S (2018) Visual analytics for explainable deep learning. IEEE Comput Graphics Appl 38(4):84–92
237. Reis T, Bruchhaus S, Vu B, Bornschlegl MX, Hemmje ML (2021) Towards modeling ai-based user empowerment for visual big data analysis. In BIRDS+ WEPIR@ CHIIR, pp 67–75
238. Rojat T, Puget R, Filliat D, Del Ser J, Gelin R, Díaz-Rodríguez N (2021) Explainable artificial intelligence (xai) on time series data: A survey. arXiv preprint arXiv:2104.00950
239. Alguliyev RM, Aliguliyev RM, Abdullayeva FJ (2019) Privacy-preserving deep learning algorithm for big personal data analysis. J Ind Inf Integr 15:1–14
240. Xiao Y, Lu H, Chen G, Mao W (2019) A bio-inspired method to realize fault-tolerance online. In 2019 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS), pp 169–175. IEEE

241. Bouarara HA, Hamou RM, Amine A (2020) New bio inspired techniques in the filtering of spam: Synthesis and comparative study. In Robotic Systems: Concepts, Methodologies, Tools, and Applications, pp 693–726. IGI Global
242. Husain MS (2020) Nature inspired approach for intrusion detection systems. Design and Analysis of Security Protocol for Communication, pp 171–182
243. Elhadj HB, Jmal R, Chelligue H, Fourati LC (2020) A2isdiot: Artificial intelligent intrusion detection system for software defined iot networks. In Workshops of the International Conference on Advanced Information Networking and Applications. Springer, pp 798–809
244. Miloslavskaya N (2020) Stream data analytics for network attack's prediction. Procedia Comput Sci 169:57–62
245. Ahsan MM, Gupta KD, Nag AK, Pouydal S, Kouzani AZ, Mahmud MAP (2020) Applications and evaluations of bio-inspired approaches in cloud security: A review. IEEE Access
246. Fernandes DAB, Freire MM, Fazendeiro PAP, Inacio PRM (2017) Applications of artificial immune systems to computer security: A survey. J Inf Secur Appl 35:138–159
247. Thakkar A, Lohiya R (2020) Role of swarm and evolutionary algorithms for intrusion detection system: A survey. Swarm Evolut Comput 53:100631
248. Khan S, Shiraz M, AW Abdul Wahab, Gani A, Han Q, Bin Abdul Rahman Z (2014) A comprehensive review on adaptability of network forensics frameworks for mobile cloud computing. The Scientific World Journal
249. Bragazzi NL, Dai H, Damiani G, Behzadifar M, Martini M, Jianhong W (2020) How big data and artificial intelligence can help better manage the covid-19 pandemic. Int J Environ Res Publ Health 17(9):3176
250. Castro JL, Delgado M, Medina J, Ruiz-Lozano MD (2011) An expert fuzzy system for predicting object collisions. its application for avoiding pedestrian accidents. Expert Syst Appl 38(1):486–494
251. Adhikary P, Roy PK, Mazumdar A (2012) Safe and efficient control of hydro power plant by fuzzy logic. IJESAT 2(5):1270–1277
252. Muller R, Nocker G (1992) Intelligent cruise control with fuzzy logic. In Proceedings of the Intelligent Vehicles, pp 173–178. IEEE
253. Kaur A, Kaur A (2012) Comparison of fuzzy logic and neuro-fuzzy algorithms for air conditioning system. Int J Soft Comput Eng 2(1):417–20
254. Zhang X, Onieva E, Perallos A, Osaba E, Lee VCS (2014) Hierarchical fuzzy rule-based system optimized with genetic algorithms for short term traffic congestion prediction. Transp Res Part C: Emerg Technol 43:127–142
255. Hassanien AE, Emary E (2018) Swarm Intell: Principles, Adv, Appl. CRC Press, Florida
256. Shi Y (2011) Brain storm optimization algorithm. In International conference in swarm intelligence. Springer pp 303–309
257. Ramírez-Gallego S, García S, Benítez JM, Herrera F (2018) A distributed evolutionary multivariate discretizer for big data processing on apache spark. Swarm Evolut Comput 38:240–250
258. Canaval SG, Rubio BO, Mozo A (2015) Npepe: massive natural computing engine for optimally solving np-complete problems in big data scenarios. In East European Conference on Advances in Databases and Information Systems. Springer, pp 207–217
259. Abdualrhman MAA, Padma MC (2019) Cs-ibc: Cuckoo search based incremental binary classifier for data streams. J King Saud Univ-Comput Inf Sci 31(3):367–377
260. Casey MC, Damper RI (2010) Special issue on biologically-inspired information fusion. Inform Fusion 11(1):2–3
261. Alam S, De D (2019) Bio-inspired smog sensing model for wireless sensor networks based on intracellular signalling. Information Fusion 49:100–119
262. Barredo-Arrieta A, Del Ser J (2020) Plausible counterfactuals: Auditing deep learning classifiers with realistic adversarial examples. arXiv preprint arXiv:2003.11323,
263. Banharnsakun A (2019) Multi-focus image fusion using best-so-far abc strategies. Neural Comput Appl 31(7):2025–2040
264. Ghomeshi H, Gaber MM, Kovalchuk Y (2020) A non-canonical hybrid metaheuristic approach to adaptive data stream classification. Future Gener Comput Syst 102:127–139
265. Sekhar SRM, Siddesh GM, Anand S, Laxmi D (2020) Nature-inspired techniques for data security in big data. In Security, Privacy, and Forensics Issues in Big Data, pp 189–216. IGI Global,
266. White T (2012) Hadoop: The definitive guide. "O'Reilly Media, Inc.",
267. Monteith JY, McGregor JD, Ingram JE (2013) Hadoop and its evolving ecosystem. In 5th International Workshop on Software Ecosystems (IWSECO 2013), p 50
268. Erraissi A, Belangour A, Tragha A (2017) A big data hadoop building blocks comparative study. International Journal of Computer Trends and Technology
269. Zikopoulos P, Deroos D, Parasuraman K, Deutsch T, Corrigan D, Giles J (2013) Harness the power of big data: The IBM big data platform. McGraw-Hill New York, NY
270. Shanmuganathan Subana (2016) Artificial neural network modelling: An introduction. In Artificial neural network modelling. Springer, pp 1–14
271. McKinnon AD, Thompson SR, Doroshchuk RA, Fink GA, Fulp EW (2013) Bio-inspired cyber security for smart grid deployments. In 2013 IEEE PES Innovative Smart Grid Technologies Conference (ISGT), pp 1–6. IEEE
272. Zhan Z-H, Liu X-F, Gong Y-J, Zhang J, Chung HS-H, Li Y (2015) Cloud computing resource scheduling and a survey of its evolutionary approaches. ACM Comput Surv (CSUR) 47(4):63
273. Forestiero A, Leonardi E, Mastroianni C, Meo M (2010) Self-chord: a bio-inspired p2p framework for self-organizing distributed systems. IEEE/ACM Trans Netw (TON) 18(5):1651–1664
274. Mirchandaney R, Towsley D, Stankovic JA (1990) Adaptive load sharing in heterogeneous distributed systems. J Parallel Distrib Comput 9(4):331–346
275. Feller E, Rilling L, Morin C (2011) Energy-aware ant colony based workload placement in clouds. In Proceedings of the 2011 IEEE/ACM 12th International Conference on Grid Computing, pp 26–33. IEEE Computer Society
276. Chiang F, Braun R (2006) A nature inspired multi-agent framework for autonomic service management in pervasive computing environments. In Network Operations and Management Symposium, 2006. NOMS 2006. 10th IEEE/IFIP, pp 1–4. IEEE
277. Yeom K (2010) Bio-inspired self-organization for supporting dynamic reconfiguration of modular agents. Math Comput Modell 52(11–12):2097–2117
278. Joseph CT, Chandrasekaran K, Cyriac R (2015) A novel family genetic approach for virtual machine allocation. Procedia Comput Sci 46:558–565
279. Pop CB, Anghel I, Cioara T, Salomie I, Vartic I (2012) A swarm-inspired data center consolidation methodology. In Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, p 41. ACM
280. Barbagallo D, Di Nitto E, Dubois DJ, Mirandola R (2010) A bio-inspired algorithm for energy optimization in a self-organizing data center. In Self-Organizing Architectures. Springer, pp 127–151

281. Xu Z (2017) Algorithm and Hardware Co-design for Learning On-a-chip. Ph.D thesis, Arizona State University
282. Wang L, Ma Y, Yan J, Chang V, Zomaya AY (2018) pipscloud: High performance cloud computing for remote sensing big data management and processing. Future Gener Comput Syst 78:353–368
283. Singh AK, Dziurzanski P, Mendis HR, Indrusiak LS (2017) A survey and comparative study of hard and soft real-time dynamic resource allocation strategies for multi-/many-core systems. ACM Comput Surv (CSUR) 50(2):24
284. Mishra AK Cogpro: Cognitive processor for astronomical big data analysis
285. Mahmud M, Kaiser MS, Rahman MM, Rahman MA, Shabut A, Al-Mamun S, Hussain A (2018) A brain-inspired trust management model to assure security in a cloud based iot framework for neuroscience applications. arXiv preprint arXiv:1801.03984,
286. Shyamala CK, Chandran Ashwathi (2018) An autonomous trust model for cloud integrated framework. In Computational Vision and Bio Inspired Computing. Springer, pp 31–44
287. Acharjya DP, Ahmed K (2016) A survey on big data analytics: challenges, open research issues and tools. Int J Adv Comput Sci Appl 7(2):511–518
288. Pacini E, Mateos C, Garino CG (2014) Distributed job scheduling based on swarm intelligence: A survey. Comput Electr Eng 40(1):252–269
289. Teijeiro D, Pardo XC, González P, Banga JR, Doallo R (2016) Towards cloud-based parallel metaheuristics: a case study in computational biology with differential evolution and spark. The International Journal of High Performance Computing Applications, p 1094342016679011,
290. Bernábe-Loranca MB, Velazquez RG, Analco ME, Ruíz-Vanoye J, Penna AF, Sánchez A (2016) Bioinspired tabu search for geographic partitioning. In Advances in Nature and Biologically Inspired Computing. Springer, pp 189–199
291. García-Magariño I, Lacuesta R, Lloret J (2017) Agent-based simulation of smart beds with internet-of-things for exploring big data analytics. IEEE Access
292. Aceto G, Persico V, Pescapé A (2020) Industry 4.0 and health: Internet of things, big data, and cloud computing for healthcare 4.0. Journal of Industrial Information Integration, p 100129
293. Adadi A, Berrada M (2018) Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access 6:52138–52160
294. Taheri J, Zomaya AY (2009) Bio-inspired algorithms for mobility management. J Interconnect Netw 10(04):497–516
295. Osaba E, Yang XS, Del Ser J (2020) Is the vehicle routing problem dead? an overview through bioinspired perspective and a prospect of opportunities. In Nature-Inspired Computation in Navigation and Routing Problems. Springer, pp 57–84
296. Tuhtan JA, Nag S, Kruusmaa M (2020) Underwater bioinspired sensing: New opportunities to improve environmental monitoring. IEEE Instrum Measurement Mag 23(2):30–36
297. Feng R, Chen X, Song F, Wang F, Wang X-L, Wang Y-Z (2020) A bioinspired slippery surface with stable lubricant impregnation for efficient water harvesting. ACS Appl Mater Interfaces 12(10):12373–12381
298. Hussain I, Ullah M, Ullah I, Bibi A, Naeem M, Singh M et al (2020) Optimizing energy consumption in the home energy management system via a bio-inspired dragonfly algorithm and the genetic algorithm. Electronics 9(3):406
299. Ullah I, Hussain I, Singh M (2020) Exploiting grasshopper and cuckoo search bio-inspired optimization algorithms for industrial energy management system: Smart industries. Electronics 9(1):105
300. Johnson AP, Al-Aqrabi H, Hill R (2020) Bio-inspired approaches to safety and security in iot-enabled cyber-physical systems. Sensors 20(3):844
301. Tassone J, Choudhury S (2020) A comprehensive survey on the ambulance routing and location problems. arXiv preprint arXiv:2001.05288
302. Wenbin G, Li Y, Zheng K, Yuan M (2020) A bio-inspired scheduling approach for machines and automated guided vehicles in flexible manufacturing system using hormone secretion principle. Adv Mech Eng 12(2):1687814020907787
303. Hamdad L, Ournani Z, Benatchba K, Bendjoudi A (2020) Two-level parallel cpu/gpu-based genetic algorithm for association rule mining. Int J Comput Sci Eng 22(2–3):335–345
304. Alfarraj O, Alzubi A, Tolba A (2019) Optimized feature selection algorithm based on fireflies with gravitational ant colony algorithm for big data predictive analytics. Neural Comput Appl 31(5):1391–1403
305. Hajeer MH, Dasgupta D (2016) Distributed genetic algorithm to big data clustering. In 2016 IEEE Symposium Series on Computational Intelligence (SSCI), pp 1–9. IEEE
306. Senthilkumar M (2018) Energy-aware task scheduling using hybrid firefly-bat (ffabat) in big data. Cybern Inf Technol 18(2):98–111
307. Ganesan R, Raajini XM, Nayyar A, Sanjeevikumar P, Hossain E, Ertas AH (2020) Bold: Bio-inspired optimized leader election for multiple drones. Sensors 20(11):3134
308. Slowik A, Kwasnicka H (2017) Nature inspired methods and their industry applications-swarm intelligence algorithms. IEEE Trans Industr Inf 14(3):1004–1015
309. Jayaratne M, Alahakoon D, De Silva D, Yu X (2018) Bio-inspired multisensory fusion for autonomous robots. In IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society, pp 3090–3095. IEEE
310. Martínez-Álvarez F, Asencio-Cortés G, Torres JF, Gutiérrez-Avilés D, Melgar-García L, Pérez-Chacón R, Rubio-Escudero C, Riquelme JC, Troncoso A (2020) Coronavirus optimization algorithm: A bioinspired metaheuristic based on the covid-19 propagation model. arXiv preprint arXiv:2003.13633
311. Zhang Q, Yang J, Liu X, Guo L (2020) A bio-inspired navigation strategy fused polarized skylight and starlight for unmanned aerial vehicles. IEEE Access 8:83177–83188
312. Silva J, Herazo-Beltrán Y, Marín-González F, Varela N, Lezama Omar BP, Palencia P, Mercado CV (2020) Comparison of bio-inspired algorithms applied to the hospital mortality risk stratification. In International Conference of Research Applied to Defense and Security. Springer, pp 177–185
313. Singh S, Chana I (2016) A survey on resource scheduling in cloud computing: Issues and challenges. J Grid Comput 14(2):217–264
314. Chen CLP, Zhang C-Y (2014) Data-intensive applications, challenges, techniques and technologies: A survey on big data. Inf Sci 275:314–347
315. Wang J, Yilang W, Yen N, Guo S, Cheng Z (2016) Big data analytics for emergency communication networks: A survey. IEEE Commun Surv Tutorials 18(3):1758–1778
316. Gill SS, Chana I, Singh M, Buyya R (2019) Radar: Self-configuring and self-healing in resource management for enhancing quality of cloud services. Concurr Comput: Practice Experience 31(1):e4834
317. Singh I, Singh KV, Singh S (2017) Big data analytics based recommender system for value added services (vas). In Proceedings of Sixth International Conference on Soft Computing for Problem Solving. Springer, pp 142–150
318. Ilango SS, Vimal S, Kaliappan M, Subbulakshmi P (2019) Optimization using artificial bee colony based clustering approach for big data. Cluster Comput 22(5):12169–12177

319. Kune R, Konugurthi PK, Agarwal A, Chillarige RR, Buyya R (2014) Genetic algorithm based data-aware group scheduling for big data clouds. In 2014 IEEE/ACM International Symposium on Big Data Computing, pp 96–104. IEEE
320. Gandomi AH, Sajedi S, Kiani B, Huang Q (2016) Genetic programming for experimental big data mining: A case study on concrete creep formulation. Automation in Construction 70:89–97
321. Elsayed S, Sarker R (2016) Differential evolution framework for big data optimization. Memetic Comput 8(1):17–33
322. Kashan AH, Keshmiry M, Dahooie JH, Abbasi-Pooya A (2018) A simple yet effective grouping evolutionary strategy (GES) algorithm for scheduling parallel machines. Neural Comput Appl 30(6):1925–1938
323. Mafarja MM, Mirjalili S (2017) Hybrid whale optimization algorithm with simulated annealing for feature selection. Neurocomputing 260:302–312
324. Barbu A, She Y, Ding L, Gramajo G (2016) Feature selection with annealing for computer vision and big data learning. IEEE Trans Pattern Anal Mach Intell 39(2):272–286
325. Tayal A, Singh SP (2018) Integrating big data analytic and hybrid firefly-chaotic simulated annealing approach for facility layout problem. Ann Oper Res 270(1–2):489–514
326. Saida IB, Nadjet K, Omar B (2014) A new algorithm for data clustering based on cuckoo search optimization. In Genetic and Evolutionary Computing. Springer, pp 55–64
327. Raj ED, Babu LDD (2015) A firefly swarm approach for establishing new connections in social networks based on big data analytics. Int J Commun Netw Distrib Syst 15(2–3):130–148
328. Wang H, Wang W, Cui L, Sun H, Zhao J, Wang Y, Xue Yu (2018) A hybrid multi-objective firefly algorithm for big data optimization. Appl Soft Comput 69:806–815
329. Wang L, Geng H, Liu P, Lu K, Kolodziej J, Ranjan R, Zomaya AY (2015) Particle swarm optimization based dictionary learning for remote sensing big data. Knowl-Based Syst 79:43–50
330. Hossain MS, Moniruzzaman M, Muhammad G, Ghoneim A, Alamri A (2016) Big data-driven service composition using parallel clustered particle swarm optimization in mobile environment. IEEE Trans Serv Comput 9(5):806–817
331. Lin K-C, Zhang K-Y, Huang Y-H, Hung JC, Yen N (2016) Feature selection based on an improved cat swarm optimization algorithm for big data classification. J Supercomput 72(8):3210–3221
332. Cheng S, Zhang Q, Qin Q (2016) Big data analytics with swarm intelligence. Industrial Management & Data Systems
333. Pan XM (2016) Application of improved ant colony algorithm in intelligent medical system: from the perspective of big data. Chem Eng Trans 51:523–528
334. Bin H, Dai Y, Yun S, Moore P, Zhang X, Mao C, Chen J, Lixin X (2016) Feature selection for optimized high-dimensional biomedical data using an improved shuffled frog leaping algorithm. IEEE/ACM Trans Comput Biol Bioinf 15(6):1765–1773
335. Elsherbiny S, Eldaydamony E, Alrahmawy M, Reyad AE (2018) An extended intelligent water drops algorithm for workflow scheduling in cloud computing environment. Egypt Informatics J 19(1):33–55
336. Neeba EA, Koteeswaran S (2019) Bacterial foraging information swarm optimizer for detecting affective and informative content in medical blogs. Cluster Comput 22(5):10743–10756
337. Ahmad K, Kumar G, Wahid A, Kirmani MM (2015) Intrusion detection and prevention on flow of big data using bacterial foraging. In Handbook of Research on Securing Cloud-Based Databases with Biometric Applications, pp 386–411. IGI Global
338. Schmidt B, Al-Fuqaha A, Gupta A, Kountanis D (2017) Optimizing an artificial immune system algorithm in support of flow-based internet traffic classification. Appl Soft Comput 54:1–22
339. George G, Parthiban L (2015) Multi objective hybridized firefly algorithm with group search optimization for data clustering. In 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), pp 125–130. IEEE
340. Xun P, Chen SX, XianPing Yu, Zhang L (2018) Developing a novel hybrid biogeography-based optimization algorithm for multilayer perceptron training under big data challenge. Sci Program 1–7(03):2018
341. Rezaei P, Solimanpur M, Rezaee MJ (2016) Solving multi-objective portfolio optimization problem using invasive weed optimization. Swarm Evolut Comput 28:01

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.