Towards an Adaptive Multi-Agent System for Dynamic Big Data Analytics

Elhadi Belghache, Jean-Pierre Georgé, Marie-Pierre Gleizes

Towards an Adaptive Multi-Agent System for Dynamic Big Data Analytics

Elhadi Belghache, Jean-Pierre Georgé, Marie-Pierre Gleizes
IRIT, University of Toulouse, France
Email: {elhadi.belghache, jean-pierre.george, marie-pierre.gleizes}@irit.fr

***Abstract*—The big data era brought us new data processing and data management challenges to face. Existing state-of-the-art analytics tools come now close to handle ongoing challenges and provide satisfactory results with reasonable cost. But the speed at which new data is generated and the need to manage changes in data both for content and structure lead to new rising challenges. This is especially true in the context of complex systems with strong dynamics, as in for instance large scale ambient systems. One existing technology that has been shown as particularly relevant for modeling, simulating and solving problems in complex systems are Multi-Agent Systems. This article aims at exploring and describing how such a technology can be applied to big data in the form of an *Adaptive Multi-Agent System* providing dynamic analytics capabilities. This ongoing research has promising outcomes but will need to be discussed and validated. It is currently being applied in the *neOCampus* project, the ambient campus of the University of Toulouse III.**

***Keywords*—Big Data; Adaptive Multi-Agent Systems; Dynamic Analytics; Big Data Challenges**

1. Rising Challenges in the Big Data World

[1], [2] and others provide recent surveys of the main concepts of big data analytics with a slight highlight on big data analysis (mining). In fact, most of the big data publications focus more on big data analysis and its applications than on other important concepts. In this scope of big data literature analysis, [3] extracted and validated from the literature the main dimensions or concepts that characterize the big data topic, which are: data dimension, IT-infrastructure dimension, method dimension, application dimension. This analysis showed that very few publications amongst the literature tackle the data selection, the results visualization and interpretation.

1.1. Big Data Analytics

One critical part of our endeavor is to go to the root of the Big Data Analytics topic, its definitions and concepts, its challenges and its goals, so as to approach it with a new eye. From the literature we managed to construct a global picture of the Big Data Analytics with enough details in order to have a fine overview of it (see figure 1).

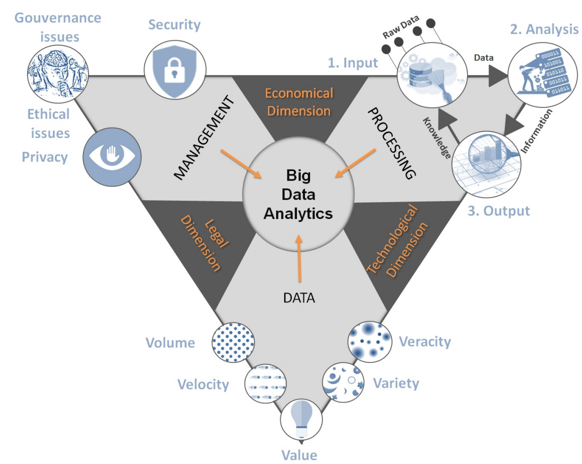

Figure 1. An overview of Big Data Analytics.

We define big data analytics by means of **three concepts** and **three dimensions**, which hold the main keys that one should keep in mind when it comes to big data analytics.

Data. Obviously the core of big data analytics, data consists here in a huge amount of potentially private and critical data generated rapidly from multiple sources in various shapes; data may be volatile, non persistent and untrustworthy due to the possible uncertainties within them. These features of big data are the basis of the main challenges of big data analytics that prompted the scientific to renew the data processing concept.

Processing. The most known processing model is *Knowledge Discovery from Databases (KDD)*. Nevertheless each community (Databases, Machine Learning, Business Intelligence, etc.) has its own way to model the data processing pipeline. They are all similar with small variations to highlight the steps related to the community strength.

The main model is the simplest one. **Input** or *pre-processing* is the transformation step of the incoming *raw data* into *data* which means putting them in a ready-to-process form by gathering, integrating, cleaning, reducing the raw data. **Analysis**, also known as *data mining*, is the center of the process, where a plethora of mining algorithms extract relevant *information*. **Output**, called *post-processing*, is the last step of the processing pipeline, in which the user produces his own *knowledge* about the incoming data by interpreting or annotating the extracted *information* by means of an intuitive visualization after that *information* has been evaluated and selected.

Since these three steps are sequential, each one influences the next for the best or the worst. Indeed, the more clean are the *data* the more relevant the extracted *information* will be, the more relevant and well presented are the *information* the best should the produced *knowledge* be, and this *knowledge* will help the pre-processing in the next processing cycle. This is why the *pre-processing*, the *mining* and the *post-processing* have to be considered with an equivalent concern.

Management. People's concern about malicious or unethical use of their personal, private or critical data led big data analytics designers to set rules for safe data management that ensures *privacy*, *security* and *ethical governance* of such personal and sensitive data.

A widespread technique to achieve this is to encrypt the *Personal Identifiers (PID)*, like the name, the birth date and the zip code, during the integration task in the pre-processing step and remove them for the analysis step.

Technological dimension. Or IT-infrastructure dimension; relies on the *data* concept and the *processing* concept. Thus, this dimension represents the set of tools, software and hardware architectures used for data storage and data processing, like for example: *DBMS, MapReduce, Grid Computing, Cloud Computing* and so on.

Economical dimension. Joins the *management* concept and the *processing* concept, and embodies the commitment of the data analysts to discover new processing methods that get more relevant and less costly information in order to earn as much benefits as possible in a safe management scope.

Legal dimension. Depends on *management* and *data*. It expresses the obligation of ethical and safe use of the data by for example giving to the owner the full rights on his data and also giving him a clear description of how his data are used and for what purpose. Many organizations and councils were founded to establish and enforce such laws.

1.2. Ongoing Challenges

At the beginning of the big data era three main challenges inherent to the characteristics of big data appeared (the initial '3Vs' of big data): **Volume:** data sets with tremendous size and high complexity (a lot of features); **Velocity:** rapid generation of data that arrives in continuous streams; **Variety:** different types of data come in various forms. These challenges, also known as the 'data flood', drove the storage systems and processing techniques at that time to their limits.

After becoming familiarized with the first challenges, new techniques started to get good results, but soon the data flood overwhelmed these techniques. Indeed, as the volume of data grew and the sources multiplied the raw data were becoming poorer and useful information becoming rarer ("thirst despite the flood"). Increasingly usefulness and reliability of the data and their sources were questioned. Hence the apparition of two new challenges bringing the challenges of big data to '5Vs'. [4] define the new 'Vs' as follows: **Value:** the usefulness of the data or more precisely the amount of useful information among the data flood; **Veracity**: reliability and confidence attributed to the data and their sources.

1.3. Rising challenges

Many of the current analytics tools [5] [6] can handle ongoing challenges and provide satisfactory results with reasonable cost. For example the recent *Data Science Machine* [7] is based on a *Deep Feature Synthesis* algorithm and an automatically optimized ensemble learning pipeline in the interest of providing a fully automated analytics tool.

However, with the recent increase of the number of smart and wearable devices and other measuring instruments in ambient applications, we barely begin to deal with *all* the aspects of those new big data. As a result, the importance of the ongoing challenges is renewed, and additional complementary data processing and data management challenges appear. As [2] expected we move forward to the next stage of big data analytics.

- **Genericity:** as shown in [5] and [6], most of the analytics tools are domain-dependent and require domain-specific expertise to build these tools. Hence, in order to adapt these tools to other application domains, the designer has to reconstruct his processing technique (data mining algorithm). Thus, designing a generic big data analytics tool should be a new challenge.
- **Super Velocity:** since the computing power is growing, the chips are shrinking and networks capabilities are increasing; the generation and the acquisition rates of data is way beyond the storing rate and the processing rate of the current systems. This is especially true when unpredictable burst-type data generation occurs.
- **Variability:** in addition to the changes (growth and evolution) to data content, the data structure (features) may change as well over time. Thence, we consider that managing the dynamicity of the data (content and **structure**) must be acknowledged as a new crucial challenge.

1.4. The need of new dynamic approaches

As the big data community shows a real interest in handling dynamic data (evolving content), new data mining techniques known as *stream mining* were developed [8] [9]. These stream mining algorithms usually sample the data stream (pick up some points) in a certain manner and process them in a incremental or on-line way. Despite the outcomes delivered by these new techniques, the data stream is under-exploited which potentially leads to a leak of useful

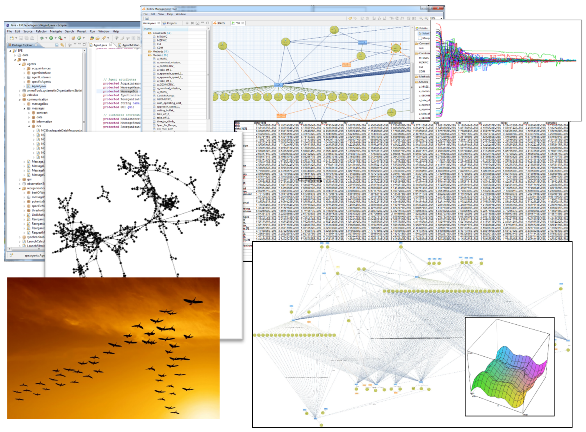

Figure 2. Multi-Agent Systems : from the inspiration from natural systems to artificial systems tackling complex and decentralized tasks.

information on one hand and forget what was previously discovered on the other hand.

Thence, our aim is to design new big data analytics techniques that can manage truly dynamic big data (content and structure), in a domain-agnostic way, and adapt itself to the changes that occur over time without having to shut down the data processing to take into account these changes by updating the process and restarting all over again.

2. A New technology for Big Data : Adaptive Multi-Agent Systems (AMAS)

2.1. Multi-Agent Systems

Multi-Agent System (MAS) [10] is defined as a macro-system composed of autonomous agents which pursue individual objectives and which interact in a common environment to solve a common task. It is often viewed as a paradigm to design complex applications (see figure 2). The *autonomy* of an agent is a fundamental characteristic : an agent is capable of reacting to its environment and displaying pro-activity (activity originating from its own decision). As such, it is the building brick of a paradigm which can be used to model a complex reality in a bottom-up way, relying only on a limited and localized knowledge of the environment for each agent. And indeed, agents have been used in a great variety of fields, a fact which can contribute to explain the difficulty to produce a unified definition of the concept.

While it is not true for all MAS, some interesting properties can be achieved when taking advantage of the autonomy of the agents. This autonomy, coupled with an adequate behavior of the agents, can lead to systems able to adjust, organize, react to changes, *etc.* without the need for an external authority to guide them. These properties are gathered under the term self-* capabilities [11] (self-*tuning*, self-*organizing*, self-*healing*, self-*evolving*...). Not all MAS necessarily present all of these self-* capabilities but, as a result of building a system from autonomous and locally situated agents, many MAS will exhibit them to some degree. Consequently, MAS are often relevant for dynamically taking into account changes in their environment. For example, a MAS in charge of regulating the traffic of packets in a computer network could be able to react efficiently to the disappearance of some of the relay nodes.

MAS have been applied to a great variety of fields: social simulation, biological modelling, systems control, robotics, *etc.* and agent-oriented modelling can be seen as a programming paradigm in general, facilitating the representation of a problem.

2.2. Adaptive Multi-Agent Systems

A particular approach to MAS relying strongly on self-* properties is the *AMAS* technology and underlying theory [12]. A designer following this approach focuses on giving the agent a local view of its environment, means to detect problematic situations and guidelines to act in a *cooperative* way, meaning that the agents will try to achieve their goals while respecting and helping the other agents around them as best as they can. The fact that the agents do not follow a global directive towards the solving of the problem but collectively build this solving, produces an *emergent problem solving process* that explores the search space of the problem in original ways.

Cooperation is the engine of the self-organisation processes taking place in the system and the heart of our bottom-up method. Cooperation is classically defined by the fact that two agents work together if they need to share resources or competences. We add to this definition, the fact that an agent locally tries on one hand, to anticipate problems and on the other hand to detect cooperation failures called Non Cooperative Situations (*NCS*, see definition 1) and try to repair these NCS [13]. To anticipate NCSs, the agent always chooses the actions which disturb other agents it knows the less. In others words, the agents, by trying to always have a cooperative attitude, act by reorganising their acquaintances and interactions with the others agents.

Definition 1. An agent is in a *Non Cooperative Situation (NCS)* when: $(\neg c_{per})$ a perceived signal is not understood or is ambiguous; $(\neg c_{dec})$ perceived information does not produce any new decision; $(\neg c_{act})$ the consequences of its actions are not useful to others.

The objective is to design systems that do the best they can when they encounter difficulties. The designer has to describe not only what an agent has to do in order to achieve its goal but also which locally detected situations must be avoided and when they are detected how to suppress them.

A cooperative agent in the *AMAS* theory has the four following characteristics. First, an agent is autonomous. Secondly, an agent is unaware of the global function of the system; this global function emerges (from the agent level towards the multi-agent level). Thirdly, an agent can detect NCSs and acts to return in a cooperative state. And finally, a cooperative agent is not altruistic (it does not always seeks to help the other agents), but benevolent (it seeks to achieve its goal while being cooperative).

Agents have to be able to detect when they are in an NCS and how they can act to come back in a cooperative situation. Agents also always try to stay in a cooperative situation and so the whole system converges to a cooperative state within and with its environment.

The main information an AMAS agent uses for its decision process is a specific measure called *criticality*. This measure represents the state of dissatisfaction or urgency of the agent regarding its local goal. Each agent is in charge of estimating its own criticality and providing it to the other agents[1]. The role of this measure is to aggregate into a single comparable value all the relevant indicators regarding the state of the agent. Having a single indicator of the state of the agent simplifies the reasoning of the agents. In addition, this mechanism has the interesting property of limiting the information transmitted to the others agents, which can be of interest in case of a large distributed systems where data privacy, data volume and computational complexity are issues.

With this additional information, each agent can and has to choose to cooperate with the most critical agent he is aware of. This leads to a very powerful heuristic to cut through a search space so as to drive the system to the expected state, effectively achieving a decentralized process that can be qualified as *emergent collective problem solving*.

This describes the typical decision process of a generic *AMAS* agent. But the NCS and the actions which could be applied to solve them are not generic: designers have to write their own specific NCS set and related actions for each kind of agent they wish the system to contain. Moreover, designers have the task to provide the agents which adequate means to calculate their criticality. But the main idea here is that this is far more manageable and realistic at the local level of each agent than at the global level of the whole complex system.

3. Managing Data in Dynamic Complex Systems with an AMAS

There is currently an increasing interest in MAS technologies and their applications on Big Data analytics [14]. Several try to use concepts like swarm intelligence, self-organising maps, etc. However, all these new techniques are still domain-dependent and do not handle changes in the data. We aim to tackle this by applying the AMAS technology and its mechanism of adaptation through cooperation.

3.1. Applicability on the Big Data analytics pipeline

The conventional Big Data analytics process (figure 3) is a rigid straightforward pipeline which doesn't allow the modification of already loaded data (content and most importantly **structure**) on the fly.

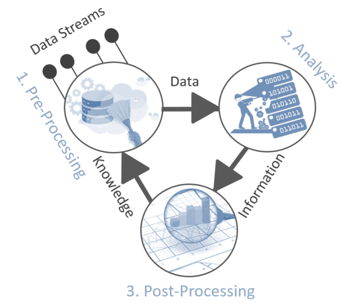

Figure 3. Conventional Big Data Analytics pipeline.

Another process model, the *distributed* pipeline (figure 4), was proposed to bypass this rigidity through processing time reduction by means of parallelism; for example with the help of the *MapReduce* pattern and its famous *Hadoop* framework. However, the main issue about dynamics still remains.

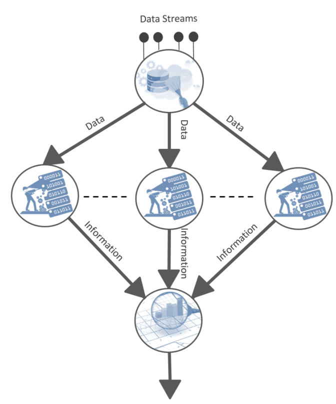

Figure 4. Distributed Big Data Analytics pipeline.

The AMAS technology, with the cooperative interaction process of its autonomous agents, gives us the means to break down this rigidity and *decentralize* the Big Data analytics process (see figure 5). This results in Big Data analytics tasks interaction, mainly through communication, and then each task can help and work together with other tasks for the sake of the continuous real-time adaptation of the analytic process to data changes.

Another way to achieve this goal, is to use the retro-action property of System-of-Systems (SoS) by designing one or several AMASs for each step of Big Data analytics and agentify them (represent them with an agent) in one super AMAS (see figure 6).

3.2. A first tool : continuous real-time detection of correlations, dependencies, relationships...

Our first aim in this scope is to build up a tool for continuous real-time detection of data correlations, dependencies

1. In open and untrusted environments, there exists several mechanisms to tackle uncertainty on exchanged information. This is often the case in System of Systems approaches. Inside a given system where each agent has been designed for the same stakeholder, each agent is assumed to provide the most trustful and accurate information

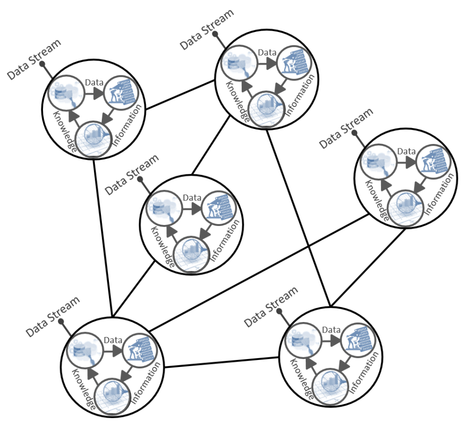

Figure 5. AMAS based Big Data Analytics 'pipeline'.

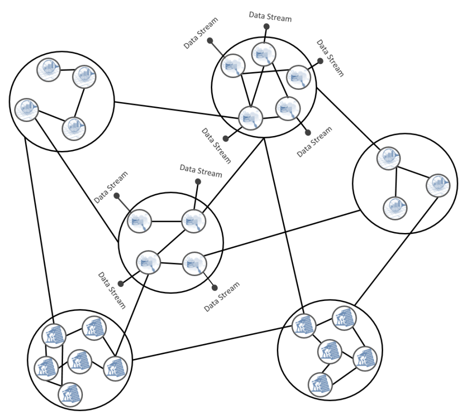

Figure 6. AMAS of AMAS based Big Data Analytics 'pipeline'.

and relationships in broad sense, in any dynamic complex system (i.e. without requiring any specific knowledge of the application domain).

When one finds strong correlation (positive or negative) he may think of a *causal* relation between the features, but often it is a weaker relation (like *influence* or *association*) or a more complex relation that may involve deep dependencies through the data like a *'Genetic Regulation Network'* where proteins inhibit and activate the generation of other ones. In addition, those relations might exist for a short time and for a particular subset or cluster of the data.

To find as much subtle relationships as possible from the data flood in real-time, we design our detection tool on the AMAS technology and we describe briefly its architecture and the agents behavior that achieve this goal.

3.2.1. The architecture. Our AMAS is composed of agents that represent the data features or attributes. Each agent receive its feature stream (the part of the data stream that conveys the values of the feature) and is, initially, randomly related to some of the other agents which compose its 'neighborhood'. For example if we have to manage a smart city, we can plug an agent on each of its sensors without gathering the sensors data in a database.

So, if there are changes in the structure of the data, it will be automatically reflected onto the AMAS mainly by the creating or removal of the agents related to the changes. As AMAS are open systems, these new agent can 'enter' and immediately take part in the activity.

3.2.2. The agents behavior. Two neighbor agents can interact in order to detect potential correlations between the features they represent by sharing the values collected from the data stream of these features, or more synthetic markers when appropriate.

If correlations are found, the agents decide which type of relation could produce these correlations and they put a confidence on this relation between them. Then they will do same with other neighbors.

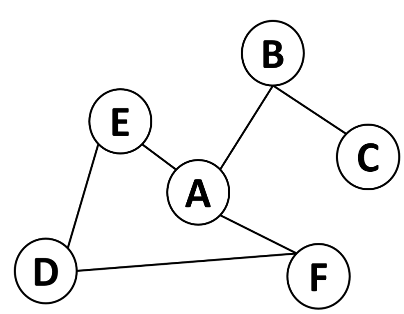

Figure 7. An example of a possible AMAS for detection of relationships.

For instance, when the agent B (see figure 7), has interacted with both his neighbors A and C, he will search for new agents to put them in his neighborhood.

However, when a new neighbor agent as E is already in the neighborhood of one of its own neighbors (A), then B will increase the confidence of his relation with A, if E and B are related, otherwise B will decrease this confidence. Thus, if the confidence of B about his relation with A reaches a certain (higher or lower) bounds, B will update (strengthening or weakening) the type of his relation with A. Also, the relation between B and E may evolve over time due to data changes and then these changes will be propagated to the A-B relation and the A-E relation. This may lead some of the relations to oscillate between strengthening and weakening which can be detected as emanating from a complex multi-feature relation.

As a result of this confidence propagation and relation updates, the whole system will adapt itself to dynamic data in an progressive and organic way.

3.3. Expected outcome

The characteristics of AMAS systems let us expect several promising outcomes, such as fast real-time correlation detection in dynamic environments and non exponential 'complex-system-like' scaling up; subtle dependencies detection that would be hidden in statistic-based analytics and fine context-learning capabilities; limited memory and storage requirement, etc. This approach would not replace current techniques, but would be applicable in specific situation where these techniques would give unsatisfactory result (for instance hidden or surprising correlations), fail or even break down (for instance, no storage or long term memory possibilities due to critical or private data, need for the removal of a data source during run-time or introduction of new and dynamic data, etc.). Moreover, current techniques and several AMAS tools could be composed in a System-of-Systems architecture to constitute a complete decentralised data 'pipeline', their individual strength being combined.

Two main projects in which our laboratory is involved will directly benefit from the bigdata analytics presented in this paper.

neOCampus. This research project is supported by the University of Toulouse III. It's aim is to demonstrate the skills of researchers of different domains of the University towards the design of the campus of the future. Three majors goals are identified : ease the life of campus user, reduce the ecological print, control the energy consumption. The campus is seen as a smart city where several thousands of data streams come from heterogeneous indoor and outdoor sensors (CO2, wind, humidity, luminosity, human presence, energy and fluids consumption...). Artificial intelligence techniques are used to understand the aims and behaviour of citizens from manually selected subsets of data. For scaling up, we need to automatically create these subsets.

3PEgase. The increasing ageing of the population is a major problem, and they have to be assisted in their daily life. Technological advances allow efficient and relevant home monitoring (such as fall detection, under nutrition, geolocation). More complicated situations and alerts require big data analytics. Our artificial intelligence approach enables real-time learning from indoors sensors data in order to detect abnormal situations for each specific housing. But we need a new tool for correlating similar situations from hundreds of different users to improve the learning process.

4. Conclusion and future work

The speed at which new data is generated, and the need to manage changes in data both for content and structure lead to new rising challenges in what can be called Dynamic Big Data Analytics. This is especially true in the context of complex systems with strong dynamics, as in for instance large scale ambient systems. One existing technology that has been shown as particularly relevant for modeling, simulating and solving problems in complex systems are Multi-Agent Systems. We described and discussed in this article how such a technology can be applied to big data in the form of an *Adaptive Multi-Agent System* where local analytics agents interact in a self-organised way.

This is ongoing research that we think has really promising outcomes, such as fast real-time correlation detection in dynamic environments, fine context-learning capabilities, non exponential 'complex-system-like' scaling up, limited memory requirement, etc. As the architecture of the systems and the behaviors of the agents (the algorithms) have been designed, we move into the implementation and validation phases.

This technology is currently being applied to several problems that will show its genericity (i.e. it does not require domain-specific expertise from the engineer that applies it) and validate its interests. The first is the *neOCampus* project, the ambient campus of the University of Toulouse III that is being iteratively equipped with pervasive ambient sensors and effectors. Student activity will be one of the main generator of data. The second is the *3Pegase* project in which we work with Orange and hospitals in Toulouse among others. The aim is an end-to-end predictive platform for elderly people staying at their own pervasively equipped homes. The third will be the performance and quality validation in well known big data on-line competitions.

References

[1] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," *Journal of Big Data*, vol. 2, no. 1, 2015.

[2] W. Fan and A. Bifet, "Mining big data: current status, and forecast to the future," *ACM sIGKDD Explorations Newsletter*, vol. 14, no. 2, pp. 1–5, 2013.

[3] T. Hansmann and P. Niemeyer, "Big data-characterizing an emerging research field using topic models," in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, vol. 1. IEEE, 2014.

[4] H. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big data and its technical challenges," *Communications of the ACM*, vol. 57, no. 7, 2014.

[5] V. Bhatnagar and S. Srinivasa, *Big Data Analytics: Second International Conference, BDA 2013, Mysore, India, December 16-18, 2013, Proceedings*. Springer, 2013, vol. 8302.

[6] S. Srinivasa and S. Mehta, *Big Data Analytics: Third International Conference, BDA 2014, New Delhi, India, December 20-23, 2014. Proceedings*. Springer, 2015, vol. 8883.

[7] J. M. Kanter and K. Veeramachaneni, "Deep feature synthesis: Towards automating data science endeavors," in *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*. IEEE, 2015, pp. 1–10.

[8] A. Paprotny and M. Thess, *Realtime Data Mining: Self-learning Techniques for Recommendation Engines*. Springer Science & Business Media, 2013.

[9] T. Palpanas, "Data series management: the road to big sequence analytics," *ACM SIGMOD Record*, vol. 44, no. 2, pp. 47–52, 2015.

[10] G. Weiß, *Multiagent Systems, A modern Approach to Distributed Artificial Systems*. MIT Press, 1999.

[11] G. Di Marzo Serugendo, M.-P. Gleizes, and A. Karageorgos, "Self-organising software: From natural to artificial adaptation," 2011.

[12] J. P. Georgé, B. Edmonds, and P. Glize, *Making Self-organizing adaptive multi-agent systems work*. Kluwer Publishing, 2004, ch. 16, pp. 321–340.

[13] G. Picard and M.-P. Gleizes, "Cooperative Self-Organization to Design Robust and Adaptive Collectives," in 2^{nd} *International Conference on Informatics in Control, Automation and Robotics (ICINCO'05), Barcelona, Spain, Volume I*. INSTICC Press, 2005.

[14] L. Cao, Y. Zeng, A. L. Symeonidis, V. Gorodetsky, S. Y. Philip, and M. P. Singh, *Agents and Data Mining Interaction: 8th International Workshop, ADMI 2012, Valencia, Spain, June 4-5, 2012, Revised Selected Papers*. Springer, 2013, vol. 7607.