Big Data Analytics: Machine Learning and Bayesian Learning Perspectives -- What is done? What is not?

By: Shan Suthaharan

This is the peer reviewed version of the following article:

Suthaharan S. Big data analytics: Machine learning and Bayesian learning perspectives—What is done? What is not? *WIREs: Data Mining and Knowledge Discovery.* 2019;9:e1283. https://doi.org/10.1002/widm.1283

which has been published in final form at https://doi.org/10.1002/widm.1283. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Abstract:

Big data analytics provides an interdisciplinary framework that is essential to support the current trend for solving real-world problems collaboratively. The progression of big data analytics framework must be clearly understood so that novel approaches can be developed to advance this state-of-the-art discipline. An ignorance of observing the progression of this fast-growing discipline may lead to duplications in research and waste of efforts. Its main companion field, machine learning, helps solve many big data analytics problems; therefore, it is also important to understand the progression of machine learning in the big data analytics framework. One of the current research efforts in big data analytics is the integration of deep learning and Bayesian optimization, which can help the automatic initialization and optimization of hyperparameters of deep learning and enhance the implementation of iterative algorithms in software. The hyperparameters include the weights used in deep learning, and the number of clusters in Bayesian mixture models that characterize data heterogeneity. The big data analytics research also requires computer systems and software that are capable of storing, retrieving, processing, and analyzing big data that are generally large, complex, heterogeneous, unstructured, unpredictable, and exposed to scalability problems. Therefore, it is appropriate to introduce a new research topic—transformative knowledge discovery—that provides a research ground to study and develop smart machine learning models and algorithms that are automatic, adaptive, and cognitive to address big data analytics problems and challenges. The new research domain will also create research opportunities to work on this interdisciplinary research space and develop solutions to support research in other disciplines that may not have expertise in the research area of big data analytics. For example, the research, such as detection and characterization of retinal diseases in medical sciences and the classification of highly interacting species in environmental sciences can benefit from the knowledge and expertise in big data analytics.

Keywords: big data analytics | machine learning | Bayesian optimization | knowledge discovery | deep learning

Article:

1 INTRODUCTION

The scientific term “big data analytics” may describe an analytical framework that provides approaches to extract knowledge from a big data environment and characterize the data source that produced the big data environment which is large, complex, heterogeneous, unstructured, unpredictable, and exposed to scalability problems (Suthaharan, 2015). This analytical framework not only provides theory and methods, but also facilitates the selection of big data systems and software. In general, the big data analytics adopt machine learning as one of the supporting tools to formulate this analytical framework; hence, it is dependent upon the successful advancements of its companion field of machine learning and other alternatives, such as data mining. When necessary, the machine learning models are parametrized with two types of parameters: hyperparameters and learned-parameters (Thornton, Hutter, Hoos, & Leyton-Brown, 2013). The suitable values for the hyperparameters are determined before training; hence, it requires proficiency in machine learning, whereas the learned-parameters are derived by the machine learning algorithms from the training data sets. When interdisciplinary applications are merged in big data analytics framework, a big question comes to everyone's mind is that, with the huge spectrum of users of big data, how the big data analytics framework will serve them better by grouping them into individuals with different levels of proficiency, ranging from nonexpert users to data analysts to data scientists, for both big data analytics research and applications. The aforementioned parametrization approaches in machine learning models, adopted by big data analytics, create problems and challenges to several disciplines when, in general, the expertise in machine learning is lacking. Similarly, the adaptive selection of machine learning algorithms for different applications is also a challenge to nonexpert users of big data analytics and machine learning. Therefore, the expectations of the individuals from different disciplines are the availability of a flexible big data analytics framework that is more intelligent and can minimize the need for user expertise for tuning model parameters and selecting suitable machine learning models and algorithms for their applications.

One of the recent advances in machine learning is the automated machine learning (AutoML) technique which adopts Bayesian optimization (Shahriari, Swersky, Wang, Adams, & De Freitas, 2016) and enables simultaneous selection and optimization of hyperparameters of machine learning models, which include logistic regression, support vector machine, decision tree, and random forest (Thornton et al., 2013). It is a very useful technique that enables the use of machine learning models in an interdisciplinary setting; however, it still suffers from performance degradation issues due to large, complex, unstructured, unpredictable, scalable, and heterogeneous data characteristics in big data analytics. The Bayesian optimization is not fully understood in a big data analytics settings; hence, it is still questionable whether the AutoML approach will provide desired results with big data analytics framework under interdisciplinary settings. What is required now is a newly defined research domain that allows exploration of big data analytics with the advancement of AutoML toward developing smart machine learning that is fully automated, adaptive, and cognitive in nature. This new research domain can make big data analytics more interdisciplinary by changing big data analytics framework much smarter through the adoption of FullAutoML techniques.

The purpose of this review article is to report the recent progress in big data analytics, machine learning, and Bayesian learning, and the way these three areas work together as companions

toward meeting the expectation of interdisciplinary research and applications. In particular, this article reviews and reports the research progress in big data analytics and in the efforts made to make machine learning much smarter so that the interdisciplinary research can benefit from big data analytics techniques and technologies.

2 BIG DATA ANALYTICS

The current progress in big data analytics can be appreciated by observing its widespread applications in many scientific and nonscientific disciplines. The recent major conference in the discipline—the 2017 I.E. International Conference on Big Data, held in Boston, MA, USA, on December 11–14, 2017—incorporated several topics that highlight the trend of big data analytics and the interests of research community in the field. Using the compiled conference program and the proceedings published in this research forum, we can divide the current progress in big data analytics into the following focused-driven research domains:

- *Descriptive big data analytics*: It addresses the theoretical and design aspects of modeling and algorithms of big data analytics and associated big data characteristics;
- *Predictive big data analytics*: It focuses on the topics of machine learning that helps the study of predictive and classification models and algorithms for big data analytics;
- *Visual big data analytics*: It defines the preprocessing and visualization techniques that help us understand the big data characteristics through exploration analysis;
- *Streaming big data analytics*: It describes the theory and methods required to study spatiotemporal characteristics of big data and machine learning models and algorithm;
- *Graphical big data analytics*: It helps to study big data environment using graphical models and network analysis under machine learning and big data analytics paradigm;
- *Big data systems and software*: It addresses the designing and building of big data systems and software that focus on efficient resource utilization and big data processing.

Although the progress of big data analytics can be grouped into above categories, a closer and careful understanding of the current research activities in each of these categories suggest that the companion field of machine learning contributes significantly to the successful progress of big data analytics which include brain network analysis (Khazaee, Ebrahimzadeh, & Babajani-Feremi, 2016), social network analysis (Cybenko, 2017), and ecological network analysis (Stephens, Sánchez-Cordero, & González Salazar, 2017). Hence, it cannot be separated from big data analytics. In addition, Bayesian inference is also emerging with big data analytics, because of the parametrization and optimization requirements of the machine learning models and algorithms (Klein, Bartels, Falkner, Hennig, & Hutter, 2015; Shahriari et al., 2016). The research in systems and software also focuses on the optimization of the performance of a computing environment; hence, the parametrization and parallelization of the computing processes and resources are explored (Dean & Ghemawat, 2010; White, 2012). The systems and software are essentials to house the big data analytics framework; hence, their progress is first discussed, and then the research progress in machine learning and Bayesian optimization, focusing on big data analytics, is discussed.

3 SYSTEMS AND SOFTWARE

The current big data systems and software have been developed as a parametric platform using a divide-and-conquer algorithm so that the distributed resources (e.g., cores, memory, executors, and software modules) can be parallelized to achieve optimal performance for big data analytics. It also focuses on providing both local and global (i.e., cloud) computing resources as an affordable mechanism for different levels of users and applications—an essential feature that is required for an interdisciplinary setting. Figure 1 illustrates these two types of big data analytics framework (local systems and software, and cloud computing resources) that are observable clearly in the current progress. It mainly shows the currently available cloud computing resources such as the Microsoft Azure (Azure: https://azure.microsoft.com/en-us/), Amazon Web Services (AWS: https://aws.amazon.com/), and Google Cloud Platform (GCP: https://cloud.google.com/), and the way they are utilized with Apache Hadoop (http://hadoop.apache.org/) and Spark (http://spark.apache.org/) systems, and the machine learning framework TensorFlow (https://www.tensorflow.org/) along with the programming languages such as R (https://www.r-project.org/), and Python (https://www.python.org/).

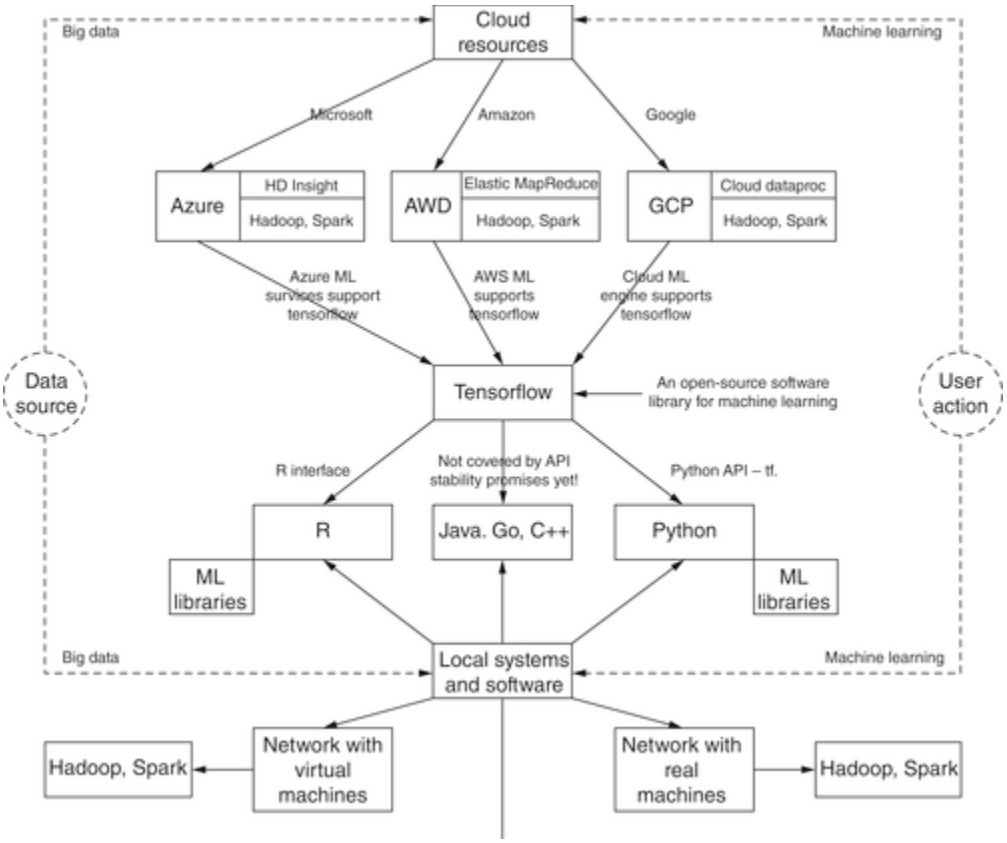

Figure 1. A possible set of systems and software for big data analytics framework

3.1 Parametrization

Presently, the big data systems and software adopt the concept called (*key*, *value*)-pair to parametrize big data computing environments. The use of this concept can be seen in the currently available computing platforms, such as the Hadoop distributed file systems, MapReduce framework, and Spark in-memory computing system that adopt the programming languages, including R, Python, Java, and Scala with machine learning software libraries

(Pääkkönen & Pakkala, 2015). These big data processing computing platforms were initially developed to perform big data analytics in local computing environment; however, as the research in big data analytics progresses, a lack of resources has been realized and the ideas were extended to cloud architecture for addressing scalability issues with other big data characteristics. Figure 1 also illustrates some of the latest tools developed for big data systems and software (Bhatt, Patwa, & Sandhu, 2017; Komarek, Pavlik, & Sobeslav, 2017).

3.2 Parallelization

The aforementioned parametrization techniques enable the parallelization of computing resources with automated resource allocation and utilization. A simple example of systems and software parametrization and parallelization can be found in chapter 5 of book by the author (Suthaharan, 2015). In the current big data system, this parallelization is not transparent and it performs optimal resource allocation and utilization automatically. However, in some situations, it is important for the users to set the system parameters such as the number of cores, executors and memory size. Therefore, the progress in this domain has also shifted to allow user intervention and suggest some mechanism to select a correct combination of system parameters (Sundaravarathan, Martin, Rope, McRoberts, & Statchuk, 2016).

4 MACHINE LEARNING

The initial progress in machine learning has defined the standard methodological processes that include the development of parametric models and development of algorithms that help optimize the model parameters using training, validation, and testing processes, and given (labeled) data. These methodological processes may be grouped into two categories: interpretable knowledge discovery (Suthaharan, 2015) and actionable knowledge discovery (Cao, 2015). Additionally, the machine learning models are defined using two types of parameters, namely, hyperparameters and learned-parameters. The hyperparameter values must be selected at the beginning of the training process—requires some expertise in the research domain—and the learned-parameters are optimized by machine learning algorithms.

4.1 Parametrization

To date, the machine learning models are parametrized in three groups: mathematical, hierarchical, and layered models. The mathematical models include the models developed using the concept of support vectors and statistical regression—examples are support vector machine, logistic regression, and lasso regression. The hierarchical models include the models developed based on the concept of decision tree—examples are random forest models. The layered models include the models proposed based on the concept of neural networks—examples include deep learning models. The first two types of approaches have been proposed to address batch learning. Later, a need for on-line learning (i.e., learning from a single point) was realized for big data analytics; hence, the layered models have been proposed.

4.2 Optimization

The optimization has been performed in a conventional manner, targeting the extraction of exact and complete knowledge from data; hence, strong mathematical techniques have been deployed to optimize model parameters. These approaches are based on the concepts of gradient descent or stochastic gradient descent (Luketina, Berglund, Greff, & Raiko, 2016). However, the latest trend is in the use of Bayesian optimization—a probabilistic approach—to address the problems and challenges evolve from the big data characteristics, such as data heterogeneity, unpredictability, and scalability (Shahriari et al., 2016; Wainer & Cawley, 2017), when the optimization of model parameters is the main objective.

5 SMART MACHINE LEARNING

As deep learning technique progresses—of course! as a promising solution to solve big data analytics problems—researchers have been trying to make machine learning techniques more intelligent as possible to meet the requirements of interdisciplinary research and applications. Hence, the expected features of machine learning techniques are currently divided into three groups: automatic, adaptive, and cognitive features. As a result, the current trend in big data analytics focuses on three types of machine learning—AutoML, adaptive machine learning, and cognitive machine learning. In essence, Bayesian learning, Bayesian optimization, and related approaches can help us develop smart machine learning.

5.1 Bayesian learning

The Bayesian learning, which includes Bayesian optimization and Bayesian mixture models, is especially adopted to optimize hyperparameters because of its ability to defend the difficulties that come from the big data characteristics using probabilistic approaches. Some of the latest progress in the integration of Bayesian optimization in big data analytics and machine learning are the techniques proposed in (Polson & Sokolov, 2017), (Shahriari et al., 2016), and (Snoek, Larochelle, & Adams, 2012). These approaches replace mathematical optimization with probabilistic optimization—a preferable approach for big data settings.

5.2 Bayesian mixture model and machine learning

The Bayesian mixture models assume data heterogeneity and represent a data domain in action as a composition of multiple sub domains (or clusters) with finite or infinite number of mixture distributions (Tafaj, Kasneci, Rosenstiel, & Bogdan, 2012). An example of its implementation in software systems for a big data environment is that the master node in a data sharing network can distribute these sub domains to worker nodes by associating them with general prior distributions and then the worker nodes process them and generate corresponding posterior distributions to specify hyperparameters for machine learning algorithms, such as the kernel and regularization parameters in Support Vector Machine (SVM) (Klein et al., 2015), and weights and learning-rate parameters in deep learning (Suthaharan, 2015).

5.3 Automated machine learning

A machine learning technique can be claimed to be truly automatic, only if every step of the machine learning technique is automated such that it can minimize the problems of

interdisciplinary applications. The research in AutoML for big data analytics is still in progress and it includes the research work reported in (Feurer et al., 2015) and (Luo, 2016). The current approaches mainly focus on the optimization of hyperparameters and the selection of models and algorithms to create suitable AutoML approaches that can be adopted by any types of users of interdisciplinary applications.

One aspect of an AutoML is the optimization of hyperparameters, such as the regularization parameter in lasso regression (Suthaharan, 2015), kernel and regularization parameters in SVM (Nguyen, Gupta, Rana, & Venkatesh, 2017; Wainer & Cawley, 2017), and learning rate in deep learning. This objective requires efficient learning models and algorithms. The current research progress in this area shows a significant interest and an appropriate use of computational techniques based on Bayesian (Klein et al., 2015) and radial basis function (Ilievski, Akhtar, Feng, & Shoemaker, 2017).

Another aspect of an AutoML is the automatic selection of models and algorithms for optimization and learning. As we know, the hyperparameter selection requires the automatic selection of suitable models and algorithms from a large pool of models and algorithms for optimization with respect to a given data set. However, in big data analytics environment, this selection process is problematic and very challenging because of the data characteristics that include complexity, scalability, and heterogeneity. A significant research has been performed—including (Thornton et al., 2013) and (Sparks et al., 2015)—in this problem space; however, the research is still progressing because of the emerging problems and challenges that evolve from the interdisciplinary nature of big data analytics.

5.4 Adaptive machine learning

The adaptive machine learning is not new and it also includes AutoML concept. The concept of adaptive machine learning can be dated back to 1990s (Blum, 1998; Littlestone & Warmuth, 1994), as stated in the paper (Torabi, Sayad, & Balke, 2005). Today, it can be observed that the big data characteristics and the current widespread interdisciplinary applications have enforced new constraints and requirements that triggered the exploration of novel approaches for adaptive machine learning.

One aspect of adaptive machine learning is the selection and optimization of the learned model parameters with respect to changing data characteristics between interdisciplinary domains. The examples include the techniques published in (Anagnostopoulos, Anagnostopoulos, & Hadjiefthymiades, 2011) and (Azodi, Gawron, Sapegin, Cheng, & Meinel, 2015). In other words, we can say that the progress in adaptive machine learning focuses on learning techniques that are adaptive to unseen data. Another aspect of adaptive machine learning is the revision of machine learning models and algorithms, with a minimum effort, to work with changing data characteristics between interdisciplinary domains. For example, in a recent paper, the authors have proposed a learning-based approach that generates an adaptive best-fit algorithm from a set of supervised learning algorithms to detect silent errors that occur in a high-performance computing environment (Subasi et al., 2017). They described the silent errors as silent data corruption in high performance computing systems, and these errors corrupt the execution results with no warning and undetectable by hardware or software.

5.5 Cognitive machine learning

Cognitive machine learning is an emerging field and it includes both the adaptive and AutoML concepts. In big data analytics, the use of exact or complete knowledge for making decisions is impractical because of the big data characteristics that include data heterogeneity, unpredictability, and scalability. Hence, the machine learning research community has realized the need for developing techniques and technologies that mimic the cognitive processes that the humans use to solve complex environmental problems and make decisions through approximation, hypothesis, and reasoning. Hence, computational intelligence and cognitive computing have been studied in recent years by focusing on machine learning and big data analytics (Suthaharan, 2016; Wang et al., 2018).

5.6 Intelligent computing (Computational Intelligence)

The intelligent computing (or computational intelligence) is another concept that has been integrated in big data analytics to help computers learn from data in the similar way that the human brain learns from data (Modha et al., 2011). This concept includes the techniques such as the fuzzy sets, genetic algorithms, and neural networks–The latest book chapter by Samanpour, Ruegenberg, and Ahlers (2018) discusses clearly about the integration of these evolutionary algorithms with machine learning, especially in the interdisciplinary domain.

5.7 Cognitive computing

This research focuses on resource-efficient, cost-effective, and cognition-enabled computing platforms for big data analytics (Suthaharan, 2016). In this computing platform, cognition-enable means the availability of methodologies and techniques that mimic human cognitive process that utilize object-recognition, speech-recognition, and functional brain networks (machine learning, reasoning, and analysis) to make cognitive decisions. Similarly, the cost-effective feature includes computational cost as well as affordable systems and software that support interdisciplinary settings. Finally, resource-efficient feature includes the automated resource allocation and resource utilization with no transparency to users—It is especially useful to nonexpert users of the learning models (Hurwitz, Kaufman, & Bowles, 2015).

5.8 Smart deep learning

A literature review suggests that machine learning for big data analytics is converging to deep learning techniques. For example, based on the proceedings of the 2017 International Conference on Machine Learning held in Sydney, Australia, on August 6–11, 2017, we can clearly see that the research interest of the machine learning community has shifted toward deep learning significantly. The deep learning, compared to other machine learning techniques, focuses on single-data-point analysis (i.e., it enables both on-line learning and batch learning), which is a highly preferred option for a big data analytics framework. Hence, it is expected in the current research of big data analytics that the deep learning will be studied, focusing on automated, adaptive, and cognitive approaches to create smart deep learning.

6 CONCLUSIONS

The progress of research in big data analytics and machine learning was reviewed and understood in three phases: past, present, and future. In Phase 1, a significant amount of research has been done on data analytics and machine learning by defining data-driven approaches. In other words, the focus of research in Phase 1 was mainly on *interpretable knowledge discovery* theme (i.e., the development of models and algorithms that enhance the discovery of interpretable knowledge from a given set of data). In essence, the data-driven approaches focused on the discovery of patterns that help understand the source that produced the data.

In Phase 2, the focus of data analytics research shifted to big data analytics based on the redefinition of data that describe big data characteristics. Hence, the research has started to shift toward domain-driven approaches (which of course included data-driven approaches). In other words, the focus of the research in Phase 2 was on *actionable knowledge discovery* theme. The actionable knowledge discovery defines the practical significance of the knowledge discovered from data through domain-driven data mining. Hence, it defines the flow of knowledge from data-driven approaches to domain-driven interdisciplinary applications.

It is now well understood that big data analytics research spans across multiple disciplines where AutoML models and algorithms are required to solve problems without the help of data science experts. Hence, as we have seen, the selection and optimization of hyperparameters became a very strong research component among machine learning (or artificial intelligence) and data science research community. A significant research still has to be done on hyperparameter selection and optimization using Bayesian optimization to develop AutoML approaches that are useful for interdisciplinary big data analytics.

In Phase 3, over the next two decades, a significant research is expected on FullAutoML approaches, toward achieving Smart Machine Learning techniques that are fully automated, adaptive, and cognitive under big data characteristics. However, to advance this research, it is important to study approaches that help discover patterns that are sensitive to changes between interdisciplinary domains. Hence, the progress reported in this paper suggests an establishment of a new research theme—*transformative knowledge discovery*. It means that, when a learning model is developed for a big data environment (big data domain or a discipline), it is important to discover the knowledge that can cause a noticeable change to the model, in addition to discovering the knowledge that is interpretable and actionable. This knowledge will help the model to become adaptive to a new environment. Hence, when a learning model is built for a data domain, it is important to study all the potential alternatives, which include both the real and hypothetical alternatives.

The discovery of transformative knowledge from big data can significantly benefit from Bayesian learning, including Bayesian optimization and Bayesian mixture models. In other words, we need to define a new research discipline called *Transformative Data Science and Analytics*, or *Transformative Data Science and Big data*, or *Transformative Big Data Analytics* that study Bayesian optimization approaches and Bayesian mixture models extensively.

References

Anagnostopoulos, T., Anagnostopoulos, C., & Hadjiefthymiades, S. (2011). An adaptive machine learning algorithm for location prediction. International Journal of Wireless Information Networks, 18(2), 88–99.

Azodi, A., Gawron, M., Sapegin, A., Cheng, F., & Meinel, C. (2015). Leveraging event structure for adaptive machine learning on big data landscapes. In S. Boumerdassi, S. Bouzefrane, & É. Renault (Eds.), Proceedings of the international conference on mobile, secure and programmable networking (pp. 28–40). Switzerland: Springer.

Bhatt, S., Patwa, F., & Sandhu, R. (2017). Access control model for AWS internet of things. In Z. Yan, R. Molva, W. Mazurczyk, & R. Kantola (Eds.), Proceedings of the international conference on network and system security (pp. 721–736). Cham: Springer.

Blum, A. (1998). On-line algorithms in machine learning. In A. Fiat & G. J. Woeginger (Eds.), Online algorithms. Lecture notes in computer science (Vol. 1442, pp. 306–325). Berlin, Heidelberg: Springer-Verlag.

Cao, L. (2015). Actionable knowledge discovery and delivery. In L. Cao (Ed.), Metasynthetic computing and engineering of complex systems (pp. 287–312). London: Springer.

Cybenko, G. (2017). Parallel computing for machine learning in social network analysis. In Proceedings of the international parallel and distributed processing symposium work-shops (pp. 1464–1471). Lake Buena Vista, FL: IEEE.

Dean, J., & Ghemawat, S. (2010). MapReduce: A flexible data processing tool. Communications of the ACM, 53(1), 72–77.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), Proceedings of the advances in neural information processing systems (pp. 2962–2970). Curran Associates, Inc.

Hurwitz, J., Kaufman, M., & Bowles, A. (2015). Cognitive computing and big data analytics. Indianapolis, USA: John Wiley & Sons, Inc.

Ilievski, I., Akhtar, T., Feng, J., & Shoemaker, C. A. (2017). Efficient hyperparameter optimization for deep learning algorithms using deterministic RBF surrogates. In Proceedings of the AAAI (pp. 822–829). Retrieved from https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14312

Khazaee, A., Ebrahimzadeh, A., & Babajani-Feremi, A. (2016). Application of advanced machine learning methods on resting-state fMRI network for identification of mild cognitive impairment and Alzheimer's disease. Brain Imaging and Behavior, 10(3), 799–817.

Klein, A., Bartels, S., Falkner, S., Hennig, P., & Hutter, F. (2015). Towards efficient Bayesian optimization for big data. In Proceedings of the NIPS 2015 workshop on Bayesian optimization.

Komarek, A., Pavlik, J., & Sobeslav, V. (2017). Performance analysis of cloud computing infrastructure. In M. Younas, I. Awan, & I. Holubova (Eds.), Proceedings of the international conference on mobile web and information systems (pp. 303–313). Cham: Springer.

Littlestone, N., & Warmuth, M. K. (1994). The weighted majority algorithm. Information and Computation, 108(2), 212–261.

Luketina, J., Berglund, M., Greff, K., & Raiko, T. (2016). Scalable gradient-based tuning of continuous regularization hyperparameters. In Proceedings of the international conference on machine learning (pp. 2952–2960). Retrieved from JMLR.org

Luo, G. (2016). A review of automatic selection methods for machine learning algorithms and hyper-parameter values. Network Modeling Analysis in Health Informatics and Bioinformatics, 5(1), 18.

Modha, D. S., Ananthanarayanan, R., Esser, S. K., Ndirango, A., Sherbondy, A. J., & Singh, R. (2011). Cognitive computing. Communications of the ACM, 54(8), 62–71.

Nguyen, D. T., Gupta, S., Rana, S., & Venkatesh, S. (2017). Stable Bayesian optimization. In J. Kim, K. Shim, L. Cao, J. G. Lee, X. Lin, & Y. S. Moon (Eds.), Proceedings of the Pacific-Asia conference on knowledge discovery and data mining (pp. 578–591). Cham: Springer.

Pääkkönen, P., & Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems. Big Data Research, 2(4), 166–186.

Polson, N. G., & Sokolov, V. (2017). Deep learning: A Bayesian perspective. Bayesian Analysis, 12(4), 1275–1304.

Samanpour, A. R., Ruegenberg, A., & Ahlers, R. (2018). The future of machine learning and predictive analytics. In C. Linnhoff-Popien, R. Schneider, & M. Zaddach (Eds.), Digital marketplaces unleashed (pp. 297–309). Berlin, Heidelberg: Springer.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & De Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE, 104(1), 148–175.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), Proceedings of the 25th international conference on neural information processing systems (pp. 2951–2959). USA: Curran Associates Inc.

Sparks, E. R., Talwalkar, A., Haas, D., Franklin, M. J., Jordan, M. I., & Kraska, T. (2015). Automating model search for large scale machine learning. In Proceedings of the 6th ACM symposium on cloud computing (pp. 368–380). New York, NY, USA: ACM.

Stephens, C. R., Sánchez-Cordero, V., & González Salazar, C. (2017). Bayesian inference of ecological interactions from spatial data. Entropy, 19(12), 547.

Subasi, O., Di, S., Balaprakash, P., Unsal, O., Labarta, J., Cristal, A., … Cappello, F. (2017). MACORD: Online adaptive machine learning framework for silent error detection. In Proceedings of the international conference on cluster computing (pp. 717–724). Honolulu, HI: IEEE.

Sundaravarathan, K., Martin, P., Rope, D., McRoberts, M., & Statchuk, C. (2016). MEWSE: Multi-engine workflow submission and execution on apache yarn. In B. Jones (Ed.), Proceedings of the 26th annual international conference on computer science and software engineering (pp. 194–200). Riverton, NJ, USA: IBM Corp.

Suthaharan, S. (2015). Machine learning models and algorithms for big data classification: Thinking with examples for effective learning. New York, USA: Springer.

Suthaharan, S. (2016). A cognitive random forest: An intra-and intercognitive computing for big data classification under cune condition. In V. N. Gudivada, V. V. Raghavan, V. Govindaraju, & C. R. Rao (Eds.), Handbook of statistics (Vol. 35, pp. 207–227). Elsevier.

Tafaj, E., Kasneci, G., Rosenstiel, W., & Bogdan, M. (2012). Bayesian online clustering of eye movement data. In S. N. Spencer (Ed.), Proceedings of the symposium on eye tracking research and applications (pp. 285–288). New York, NY, USA: ACM.

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In R. Ghani, T. E. Senator, P. Bradley, R. Parekh, & J. He (Eds.), Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 847–855). New York, NY, USA: ACM.

Torabi, K., Sayad, S., & Balke, S. T. (2005). On-line adaptive Bayesian classification for inline particle image monitoring in polymer film manufacturing. Computers & Chemical Engineering, 30(1), 18–27.

Wainer, J., & Cawley, G. (2017). Empirical evaluation of resampling procedures for optimising SVM hyperparameters. Journal of Machine Learning Research, 18(15), 1–35.

Wang, Y., Howard, N., Kacprzyk, J., Frieder, O., Sheu, P., Fiorini, R. A., … Widrow, B. (2018). Cognitive informatics: Towards cognitive machine learning and autonomous knowledge manipulation. International Journal of Cognitive Informatics and Natural Intelligence, 12(1), 1–13.

White, T. (2012). Hadoop: The definitive guide. California, USA: O'Reilly Media, Inc.