Critical Analysis On Data Science And Big Data Avenues

Mirza Ghazanfar Baig, Sandeep Kumar Nayak

Abstract: In the current scenario of digitalization of volumes data used to process and store in a bulky variety. These kind of volumes data can be the operational and non-operational whereas regular transition for multiple operation use to perform which to analyze the structuring of data in huge amount. Regular updation also hazardous activity for operational data and regular deposition of past data are helpful for future prospective. Data specification may be extracted with the labeled data called supervised. An unsupervised learning is helpful for identifying equally likely entity for the future perspective. In this research paper, it is being tried to identify that the data is a fuel for Data Science, Artificial Intelligence, Machine Learning and Deep Learning with various equally likely item with the help of suggestive measures of supervised and unsupervised learning of data. The key attribute for handling errors such as environment and performance of error identification are also presented in the Machine Learning security, which shows the importance of environment for better performance of error management.

Keywords : Data Science, Big Data, Artificial Intelligence, Machine Learning, Deep Learning, Supervised Learning, Unsupervised Learning and Reinforcement Learning.

1. INTRODUCTION

THE enormous amount of data is generating in every minute, there is a need to extract useful data from a repository inside for business in today's word. Now this is where Data Science and Big Data comes in to picture. It is observed that, every day the world produces around 2.5 quintillion bytes of data, with 90% of these data generated in the world being unstructured. It proclaim that by 2020, over 40 Zettabytes of data will have been generate, imitated and consumed. Data cannot be discard only it stored, 70-80% of whole data is being generate in last 2-3 years [1]. Data generate in these days are not only structured, but also semi structure or unstructured, imagine how any one going to process that much amount and verity of data using simple business tool [2]. So simple business intelligence tool cannot do work anymore we need complex and effective Architecture and Algorithms to analysis and it is hard to extract useful inside from the Big Data using Unsupervised and Supervised Learning Data Techniques.

Data science can be define as "the set of fundamental principles that support and guide the principle extraction of information and knowledge from data" [3]. A closely related term is data mining, which is the actual extraction of knowledge from data via machine learning algorithms that incorporate data science principles [4]. Machine learning is the field of study that focuses on how computers learn from data and the development of algorithms that make this learning possible [5]. Finally, another important concept in data science is domain expertise, which in healthcare can be define as the understanding of real-world clinical problems and the realities of patient care that help frame and contextualize the application of data science to healthcare problems [6].

The term Data Science and Big Data may scare you, but understanding it is very easy. Let us break it down in to Data and Science. Data can be anything (Fact or Figure), which can store or process and the structure or row information collected from number of data generating sources. Science is the process of exploring, observing and making sense out of something in systematic manner to enhance the human knowledge.

- *Mirza Ghazanfar Baig, Department of Computer Application, Integral University, Lucknow, India, PH-+91, 7376742180. E-mail: mirza.gb@gmail.com*
- *Dr. Sandeep Kumar Nayak, Associate Professor, Department of Computer Application, Integral University, Lucknow, India, PH-+91, 9935513331, E-mail: nayak.kr.sandeep@gmail.com*

2 DATA SCIENCE

2.1 Definition

It is the process of extracting useful knowledge and insight form data we collected by using scientific methods dealing with unstructured and structured data, Data Science is a field that comprises of everything that related to data cleansing, preparation, and analysis [7]. Data Science is the combination of mathematics, programming, statistics, analytic, data visualization & communication, capturing data in ingenious ways, the ability to look at things differently, and the activity of cleansing, preparing and aligning the data.

2.2 Cyclic Process of Data Science

In the data science life cycle, there are six step in the whole process, its starts with Defining-

- Business Understanding & Requirements
  It is the process of discovering, analyzing and defining the requirements that are relate to a specific business objective [8].
- Data Understanding & Acquisition
  It is the process that has been understood as the process of gathering, filtering, and cleaning data before the data is put in a data warehouse or any other storage solution.
- Data Exploration & Processing
  Data exploration is the process of comprises missing value imputation, outliers, feature engineering, variable creation in data.
- Modeling
  Modeling is the process of producing a descriptive diagram of relationships between various types of information that are to be stored in a database.
- Evaluation:
  To evaluating data using logical reasoning or analytical to examine each component of the data provided.
- Deployment:
  Concept of deployment in to data science refers as

the application of a model for prediction using a new dataset.

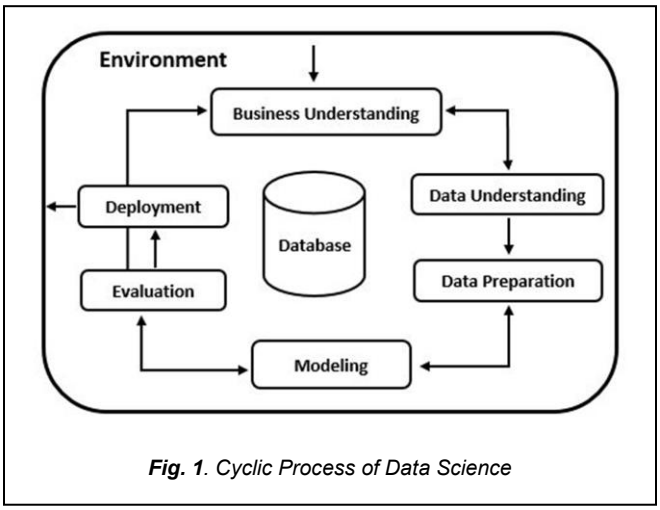

***Fig. 1**. Cyclic Process of Data Science*

3 BIG DATA NOTIONS

Big Data is a collection of data set, which is large and complex for that system it referees, and difficult to access or process using traditional database application with in a given time-frame. The data, which is bound the storage capacity and processing power of that machine it refers to, called the Big Data. It is a huge volume of data, which cannot simply store or process [9]. Big data is just a term that refers to deadly combination of data sets whose Volume (size), Variability (complexity), and Velocity (rate of growth) make them difficult to be store, access, process or analyze through conventional technologies and simple business intelligence tools [10]. The database mechanism, statistics algorithm and visualization packages, within the time necessary to make them valuable. In simple terms, it is the umbrella of techniques used when trying to extract new insights and useful information from data [11].

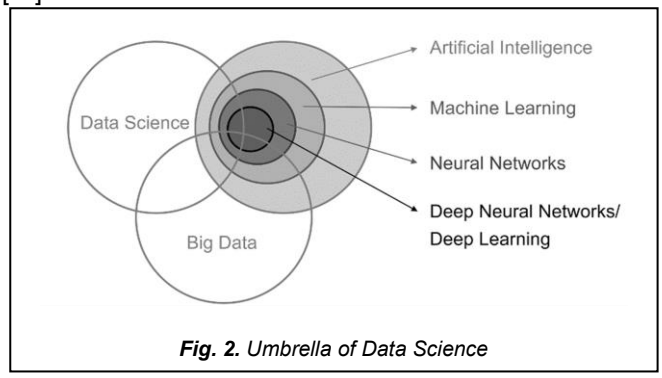

***Fig. 2.** Umbrella of Data Science*

4 ARTIFICIAL INTELLIGENCE

It is one of the main stream of Data Science. American computer scientist John McCarthy first coins the term Artificial Intelligence in 1956 at Dartmouth Conference. He defined the AI as the science and engineering of making intelligent machine. When the computer able to perform action usually requiring human intelligence, such as speech recognition, decision-making, languages translation and visual perception [12]. Machine performed like human generating activities and behavior through Digital Intelligence. Types of Human Behavior are Thinking Rationally and Acting Rationally. AI systems are autonomous, can operate without human intervention, and can learn and identify patterns to make decisions and to reach different conclusions based on the analysis of different situations [13]. When the machine is able to produce the acting behavior as a human is call artificial intelligence means that intelligence made by humans not by God. As the satellites on the earth are AI made by humans and moon is natural, made by god. Machine Learning and Deep Learning aids AI by providing set of algorithms and neural network networks to solve data driven problem. Investment in new AI-based technologies has been one of the critical strategies of the public sector at various levels of government in several countries around the world [14]. The main AI Approaches are- Reasoning and Deduction, Knowledge Representation Technique, Planning and Learning.

5 MACHINE LEARNING

In Machines Learning, models are need to trained behave like humans enabling them to mimic advanced psychological features or functions like decision-making, reasoning, and inferences [15]. It is a subset of AI, which use statically methods to enable machine to improve with experiences. It can learn from statically methods, which enable it to model the scenario in any mathematical form. ML is required for Navigation, Recognition, Prediction or Description. A good use case for machine learning that has to be perform in our day-to-day activity is spam filters, which characteristically determine whether a text or message is junk based on however closely it matches emails with the same tag [16]. Suppose the agent has a specific task to perform in the environment on the bases of the perception of the environment observed through preceptors and the agent performs the action of the task through actuator.

Agent	=	A
Environment	=	E
Task	=	T
Output	=	O
Performance	=	P
Error	=	e
Max Error	=	e+
Min Error	=	e-
High Performance Output	=	Po+
Low Performance Output	=	Po-

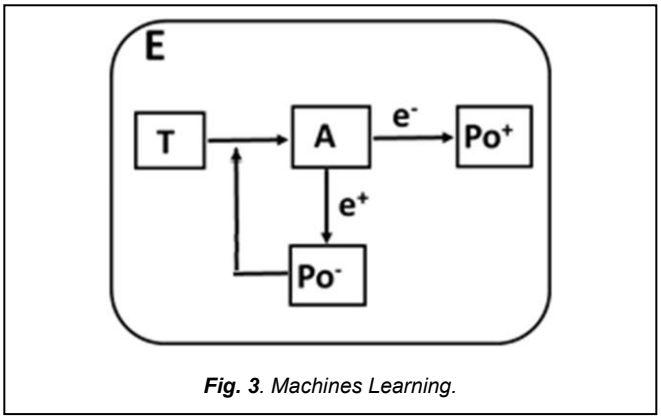

***Fig. 3**. Machines Learning.*

Hence, the Performance of the task is directly proportional to the improvement in the specific environment will either increasing or decreasing.

$$P \propto E$$

As prefigure-3 when the tusk enter into the agent then the performance of error handling will increase or decrease based on the specific environment.

5.1 Supervise Learning

Supervise Learning techniques rely on the set of past transactions for which the label (also referred to as outcome or class) of the transaction is known [17]. It is technique in which teacher or trainer, teach or train the machine using data, which is well label.

- Learning agent and trainer both takes input from surrounding environment but the trainer has the desired output or result.
- There is a learning agent, which process the given input and gets the output.
- The actual output subtracted from the desired output by the adder, which shows the error in actual output.
- The error injected as input to learning agent so that it can learn from the error while generating output from the next time and can remove the error.
- Data are label and we may already know the output with the help of trainer.
- Techniques of Supervised learning are- Regression, Classification and Deviation Detection.
- Algorithms of Supervised Learning are-
- Regression, Decision tree, Random forest, SVM and Naïve-Bayes.

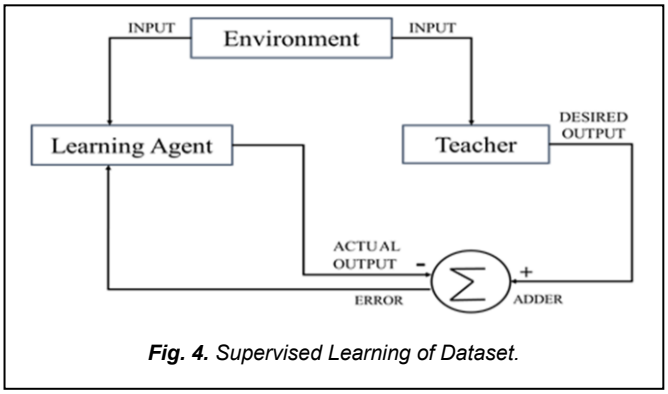

Fig. 4. *Supervised Learning of Dataset.*

5.2 Unsupervised Learning

It is a process of training machine using information unlabeled and allowing the algorithm to act on that information without guidance. It does not have any trainer, which means that correct answer is unknown to the learning agent. UL is becoming more and more popular because it does not need manually annotated data, especially in the current rapid growth of data [18].

- Unsupervised Learning does not have any teacher.
- The learning agent does not know actual output.
- In Unsupervised Learning agent have to learn itself from the patterns and meaningful correlations without the corresponding output.
- Information clusters based on feature similarities.
- Techniques used in unsupervised learning are- Clustering, Association Rule Mining and Sequential-Pattern.
- Algorithms of Unsupervised Learning are-
- K-Means, Market Basket Analysis, Apriori Algorithm

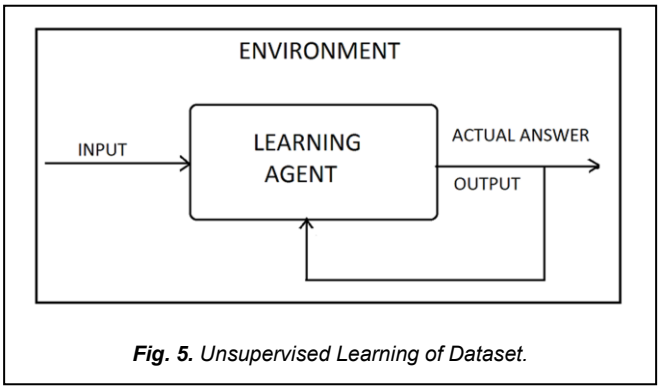

Fig. 5. *Unsupervised Learning of Dataset.*

5.3 Reinforcement Learning

The entity that performs the configuration can be either the driver itself (agent) or a track engineer (supervisor). With the phrase environment configuration, it is being refer to the activity of altering some environmental parameters to improve the performance of the agent's policy [19]. Reinforcement learning is a part from advances Machine Learning where an agent placed in an environment and it learns to behave in this environment by performing positive actions through observation and get the rewards, which it gets from those actions.

- It is the part of advances Machine Learning.
- The technique we used in Reinforcement Learning is totally reward based.
- It learns from the environment.
- It takes the data from the environment and analyses it.
- Algorithms of Reinforcement Learning are-
  Queue Learning, Reward Matrix, Maximization, Macron Decision Process.

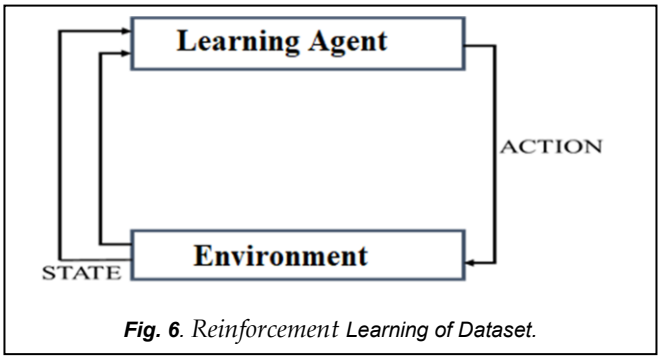

Fig. 6*. Reinforcement Learning of Dataset.*

Q- Learning

It is a model-free reinforcement-learning algorithm. Main aim of Q-learning is to understand the policy, enabled by an agent, which tells that what action need to take under certain circumstances. The connotation is "model-free" because it does not requires model from the environment, and without the requirement of adaptations, model can handle the problems through rewards and positive transitions.

Reward Matrix

In reward-based matrix, the reward function is initially unknown to the agent. Agent have to learn the reward and

transition function or can combine the both by learning instead Q-Values.

Expectation-Maximization Algorithm
Expectation-Maximization Algorithm is use for the latent variables in order to predict their values with the condition when the probability distribution general form of governing to the latent variables. This algorithm is actually at the base of many unsupervised clustering algorithms in the field of ML.

Markov Decision Process
A specific type of problem defines Reinforcement Learning, and all its solution known as Reinforcement Learning algorithms. According the problem, Agents are supposed to take the best action on bases of its current state, and when the action get repeated, then the problem is called as Markov Decision Process.

6 DEEP LEARNING

Deep Learning is a subset of Machine Learning, which make the computation of multi-layer neural networks feasible. It uses the concept of neural network to solve the complex problem. If there is any scope to say the agent do not touch a hot vessel in ML it will be do not touch any vessel which is hot, but in Deep Learning it will be do not touch anything which is hot. The most uncomplicated model in deep learning corresponds to a fully connected layer. This layer takes a vector as input x, performs transformation of the input wx+b, and then applies a non-linear activation task, e.g., a sigmoid function (σ) to produce the final output σ (wx + b). Multi-Layer Neural Networks are simply a generalization of this basic idea where Neural Network Layers are stacked in sequence [20, 21].

7 LIMITATION

As with many aspects of Data Science, data is a human creation, so it has some limits on its usability when you first obtain it. Here are some limitations that are likely to encounter-

7.1 Algorithms Require Massive Stores of Training Data

Models are 'trained', not programmed. This means that the algorithms requires an enormous amount of data to perform multiple tasks at the level of humans. Despite of the fact that data set is being created at an enhanced pace and the robust computing power is needed to efficiently process it is available, massive data sets are not easy to create or obtain for business houses. Moreover, every slight variation in an assigned task calls for another large data set to conduct additional training.

7.2 Labeling to the Training Data is a Tedious Process

To know that, what is in the data set, a time-consuming process of manually spotting and labeling items is required. However, promising new techniques are coming up, like in-stream supervision, where data is label during natural usage [17]. Application designers can accomplish this by 'sneaking in' features in the design that inherently grow training data. High-quality data collection from users used to enhance machine learning over time.

7.3 Agent and Models Cannot Explain Themselves

Whether the decision is right or wrong, knowing the visibility into how and why it made is vital, so that the human expectation are brought up in line with how the algorithm actually executed. There are techniques that can used to interpret complicated machine learning but Model and Agent are the entity that needed to define them itself.

7.4 There is Massive Predisposition in the Data

In some of the instances, models that are apparently performing well maybe actually picking up noise in to data set. As much as accuracy is important, unbiased decision-making builds trust. The infallibility of an AI solution based on the quality of its inputs. If the training data is not neutral, the outcomes will inherently amplify the discrimination and predisposition that lies in the data set.

7.5 Learning Algorithms do not collaborate to each other

In the face of the multiple innovations in learning algorithms, Models still lack the ability to generalize conditions that vary from the ones they encountered in training. Models have difficulty conveying their understandings from one set of environments to the other.

7.6 The Decision Boundary is Over Trained

The decision boundaries over trained. Which means that, data set for training is not including some examples that you want to have in a class, when you use those examples after training, you might not get the correct class label.

7.7 Data inputting after classification gets wrong class label

The input, which is not from any of the classes in reality, then it, might get a wrong class label after classification.

7.8 Classification of big data is a big challenge

We have to select many good examples from each class while you are training the classifier. If consider classification of big data that can be a big challenge.

8 SUGGESTIVE MEASURES

8.1. Distributed high-end data repository is needed. Due to the high-end computing capabilities required for performing such a large amount of analytics, the Data Science analytics capabilities are leveraged in this architecture.

8.2. Strong and Secure Labeling algorithm is required. Because the labels are intended for use when objects are stored, transmitted between systems, and when they are being handled by applications that act on labels.

8.3. Self-Explanatory Module and Agents are deadly required. An Agent must be situated in some environment and that is capable of autonomous action in this environment in order to meet its design objectives [22].

8.4. Training dataset must be natural unbiased. The Training dataset provides an unbiased evaluation of a model to fit while tuning the model's hyper parameters that is number of hidden units in a neural network.

8.5. Collaborative learning algorithm is required. It refers to an instruction method in which learners at various performance levels work together in small groups toward a common goal.

8.6. Concrete and Decision Boundary conditions are required. To find the decision boundary that maximizes the margin or the width separating the positive from the negative training data points [23].

9 DISCUSSION

In the face of the multiple innovations in learning algorithms, Models still lack the ability to generalize conditions that vary from

the ones they encountered in training. Models have difficulty conveying their understandings from one set of environments to the other. This means that anything models get rewarded for a specific use case will only be applicable to the particular use case. The business houses forced incessantly require resources to train other models, even though the use cases are directly similar. The solution of this scenario comes in the form of learning transmission. Knowledge is obtained from the task can be used in situations where the labeled data set is obtainable. The other general approaches need to established that will have the ability to build new models more rapidly.

9.1. We can get very specific about the definition of the models, which means that you can train through classifier in the model building, which has a perfect decision boundary to distinguish different model accurately.

9.2. After training, it is not necessarily store the training examples in a memory. We can just store the decision boundary in the form of mathematical equations and that would be sufficient for classifying future inputs.

Amazing technological innovations in the field of Big Data, Data Science and its sub-field Artificial Intelligence and Machine Learning has made in the last couple of years. Agents are behaving just like humans is no longer scientific fiction, but in the reality multiple industries are practicing today. Now the today's fact is that, the human society is gradually shifting or to more trusting on smart machines to solve day-to-day challenges and making their decisions. However, these evolutions heavily effects the study of the Data Science and Big Data Analytic, which might be enabling to more techniques to complete complex tasks or to overcome form the current limitations with significant implications for the way business is to conducted. In all the hype surrounding these game-changing technologies, the reality that often times gets lost amidst both the fears and the headline victories like Cortana, Alexa, Google Duplex, Siri and Sophia, is that AI technologies have several limitations that will still need a substantial amount of effort to overcome.

10 CONCLUSION

We are in an era of Data Science and Big Data. The paper describes the concept of Data Science and Big Data along with the concept of Artificial Intelligence, Machine Learning, Deep Learning, Supervised Learning, Unsupervised Learning and Reinforcement Learning with the help of cases and figures. The paper also focuses on Big Data processing problems and the 3Vs Concept (Volume, Velocity and variety) of Big Data. These technical challenges must needed efficient and fast processing for Big Data Analytic. The limitation include not just the obvious issues of scale, but also multidimensionality, lack of structuration, labialization, error-handling, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical limitations are common across a huge variety of presentation domains of data set, and therefore not cost-effective to address in the context of one domain alone.

REFERENCES

[1] Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70, 263-286.

[2] Anuragi, M. (2013). Information Cooperation and User Service among Technical Institute Libraries, Reference to Ghaziabad Region. International Research: Journal of Library and Information Science, 3(4).

[3] Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

[4] Elgendy, N., & Elragal, A. (2014, July). Big data analytics: a literature review paper. In Industrial Conference on Data Mining (pp. 214-227). Springer, Cham.

[5] Xindong Wu,Fellow IEEE, Xingquan Zhu, Senior Member IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE, "Data Mining with Big Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1, January 2014.

[6] Xu, Y., Zhang, D., Song, F., Yang, J. Y., Jing, Z., & Li, M. (2007). A method for speeding up feature extraction based on KPCA. Neurocomputing, 70(4-6), 1056-1061.

[7] Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51-59.

[8] Harbers, M. (2011). Explaining agent behavior in virtual training. Utrecht University.

[9] Bhosale, H. S., & Gadekar, D. P. (2014). A review paper on big data and hadoop. International Journal of Scientific and Research Publications, 4(10), 1-7.

[10] Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. Technological Forecasting and Social Change, 130, 99-113.

[11] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., & Kavukcuoglu, K. (2016, June). Asynchronous methods for deep reinforcement learning. In International conference on machine learning (pp. 1928-1937).

[12] Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited.

[13] Čerka, P., Grigienė, J., & Sirbikytė, G. (2017). Is it possible to grant legal personality to artificial intelligence software systems?. Computer law & security review, 33(5), 685-699.

[14] de Sousa, W. G., de Melo, E. R. P., Bermejo, P. H. D. S., Farias, R. A. S., & Gomes, A. O. (2019). How and where is artificial intelligence in the public sector going? A literature review and research agenda. Government Information Quarterly, 101392.

[15] Gu, S., Holly, E., Lillicrap, T., & Levine, S. (2017, May). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In 2017 IEEE international conference on robotics and automation (ICRA) (pp. 3389-3396). IEEE.

[16] Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., & Raghavendra, V. (2018, May). Deep learning for entity matching: A design space exploration. In Proceedings of the 2018 International Conference on Management of Data (pp. 19-34). ACM.

[17] Carcillo, F., Le Borgne, Y. A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2019). Combining unsupervised and supervised learning in credit card

fraud detection. Information Sciences.
[18] Mao, X., Yang, H., Huang, S., Liu, Y., & Li, R. (2019). Extractive summarization using supervised and unsupervised learning. Expert Systems with Applications, 133, 173-181.
[19] Big Data Working Group, “Big Data Taxonomy”, Cloud Security Alliance, 2014, https://cloudsecurityalliance.org/research/big-data/
[20] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
[21] Metelli, A. M., Ghelfi, E., & Restelli, M. (2019, May). Reinforcement Learning in Configurable Continuous Environments. In International Conference on Machine Learning (pp. 4546-4555).
[22] Fukuchi, Y., Osawa, M., Yamakawa, H., & Imai, M. (2017, October). Autonomous self-explanation of behavior for interactive reinforcement learning agents. In Proceedings of the 5th International Conference on Human Agent Interaction(pp. 97-101). ACM.
[23] Myers, A. C., & Myers, A. C. (2009, January). JFlow: Practical mostly-static information flow control. In Proceedings of the 26th ACM SIGPLAN-SIGACT symposium on Principles of programming languages (pp. 228-241). ACM.