REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
05/20/2020	Technical Report	

4. TITLE AND SUBTITLE
Basic Understanding of Artificial Intelligence, Machine Learning, Big Data, and Cyberoperations: Challenges and Opportunities

5a. CONTRACT NUMBER
FA8702-20-C-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Montanez Rodriguez, Rosana

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

8. PERFORMING ORGANIZATION REPORT NUMBER
PRS-20-1262

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Special Operations Command

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
UTSA-MITRE Collaboration

14. ABSTRACT
The focus of this paper is to provide a common understanding of applications of artificial intelligence and its related fields (DL and ML) into cyberoperations, their differences and their relation to data analytics, and the role of Big Data. This understanding is fundamental to the design and architecture decision in support of Unified Platform.

15. SUBJECT TERMS
Artificial Intelligence; Informatics; Computer Security; deep learning; cyber security; artificial intelligence; cyberoperations; bid data; data analytics; machine learning;

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**		10	Susan Carpenito
					19b. TELEPHONE NUMBER *(Include area code)* 781-271-7646

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD™

Basic Understanding of Artificial Intelligence, Machine Learning, Big Data, and Cyberoperations

Challenges and Opportunities

May, 2020

Basic Understanding of Artificial Intelligence, Machine Learning, Big Data, and Cyberoperations

Challenges and Opportunities

Introduction

Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), and Big Data are the popular buzzwords in today's technology world. Their applications in Data Analytics is driven by the need for insight from Big Data. The advent of Big Data made large scaled computations feasible and created an opportunity to study complex systems. It also led to a "data-hugging" frenzy in which organizations stored as much data as they produced without discrimination. For many, what started as data lakes quickly transformed into data swaps and many found themselves with the cost of storing data that is unusable because it cannot be interpreted. As the realities of Big Data starts to sink in, we are at a good point to reevaluate our data strategies in a way that allows us to leverage complex analytics that are required to understand hard problems in cyber operations without the pitfall of hoarding data [1].

There are several differences between the application of AI, ML, and DL in cyberoperations and business domains. First, most cyberoperation data is unlabeled and datasets are rich in features that requires expert advice for context, selection, and engineering of features. Additionally, most of the commercial application of AI, ML, and DL focuses on the accuracy of the answer. In cyberoperations, the accuracy of the answer and features selected by the algorithm are equally important for proactive defense mechanisms. For example, most neural network applications focus on the correctness of the answer. In cyber security, understanding which features were selected, and the weights assigned to each feature can provide insight into the problem beyond the accuracy of the result.

The focus of this paper is to provide a common understanding of applications of artificial intelligence and its related fields (DL and ML) into cyberoperations, their differences and their relation to data analytics, and the role of Big Data. This understanding is fundamental to the design and architecture decision in support of Unified Platform

Fundamentals

This section provides a definition of the difference of each data analysis technique. Although these techniques are built in the same mathematical principles, the terms are not interchangeable. Understanding their differences, limitations, and tradeoffs can assist with identifying the best solution for a problem. In addition to domain knowledge, different techniques require different levels of skills with some of them requiring programming knowledge.

Data Analytics

Data Analytics is the process of extracting knowledge from large datasets to support decision. Most of the techniques used in data analytics use basic statistical analysis – forecasting and prediction is the only exception.

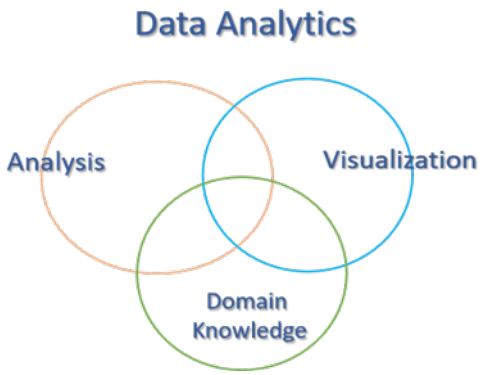

Figure 1. Areas That Comprise Data Analytics

Three activities are central to data analytics:

a) Data visualization. The graphical representation of data to facilitate understanding to support a decision.

b) Data analysis. The process of identifying data and relationships that support decision.

c) Data mining. The extraction of knowledge, insight, and patterns in data.

Problems that are a good fit for data analytics are:

1. Trend over time problems
2. Scales and performance problems
3. Comparison problems
4. Forecast and prediction of simple problems

Artificial Intelligence, Machine Learning, and Deep Learning

AI, ML, and DL use the same mathematical principles: logistic regression, neural networks, Naïve Bayes, among other advanced statistics techniques. They defer on the focus of each discipline.

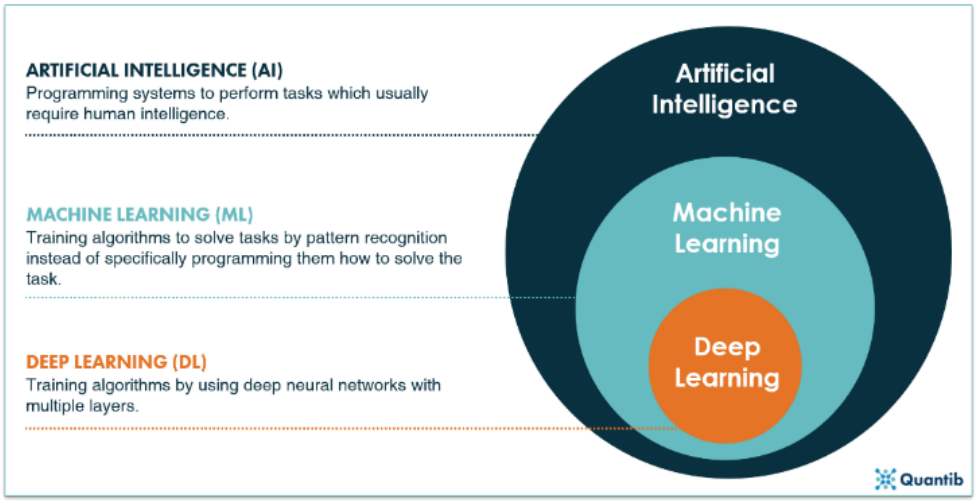

Figure 2. Relationship Between AI, ML, and DL[1]

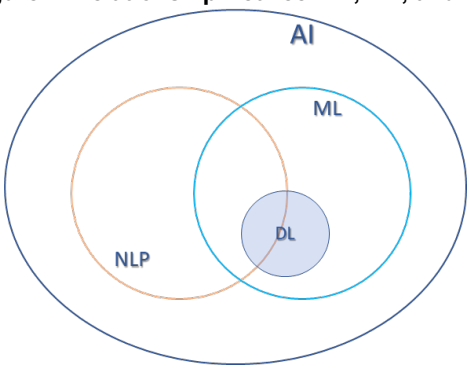

Figure 3. Relationship Among Analytics Disciplines

a) Machine Learning focuses on learning patterns to generate predictions; predictions are based on past events.
b) Natural Language Processing focuses on extracting knowledge from human speech and language elements.
c) Deep learning is a combination of machine learning techniques and neural network that results in multiple hidden layers.
d) Artificial Intelligence is a combination of techniques that allows a system to extract knowledge and learn from new experiences. The focus of AI is to develop autonomous, human-like decisions.

[1] Q. B.V and O. Six, "The ultimate guide to AI in radiology." https://www.quantib.com/the-ultimate-guide-to-ai-in-radiology

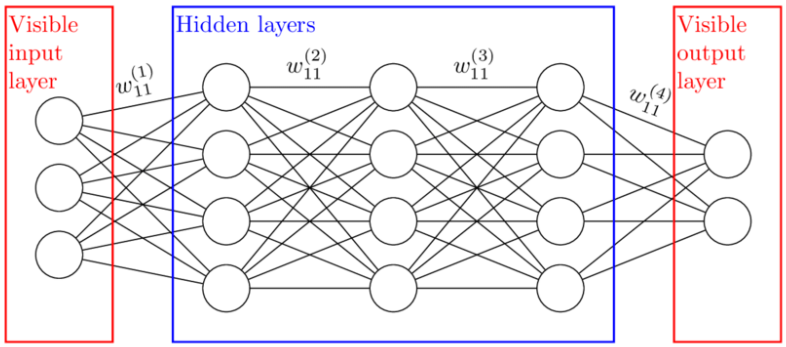

Figure 4. Representation of a Deep Neural Network

Similarities and Differences

Table 1 depicts the areas of similarities and difference between Data Analytics and Artificial Intelligence.

Table 1. Comparison of Data Analytics and Artificial Intelligence

Differences	
Data Analytics	Artificial Intelligence
• Focuses on extracting knowledge from large datasets to support decision	• Extracting knowledge to enable human-like decision making
• Central activities: data visualization, data analysis, and data mining	• Can make assumptions, test, and learn autonomously
• Other activities: prediction and forecasting	• Capable of passing the Turing test
Similarities	
Overlap of techniques between the two areas	

Limitations

The weakness of data analytics and AI applications is that they are developed around a specific problem. This specificity limits their applications to other problems. Other limitations are,

- Data analytics are dependent on the past to predict the future. In a simple system, this assertion is usually true, but complex or multidimensional problems usually do not abide by this rule. One of the reasons is because different

conditions might return the same outcome. Also, most multidimensional problems are poorly understood by nature, therefore there is a never-ending discovery cycle. Some early work in ML suggest that cyber events are a Time Series Function [2], which means that future events are dependent on recent past events.

- AI and Data Analytics focus on augment human knowledge or assist with a task. They still require humans with domain knowledge to make decisions when conflicts arise.
- Complex problems require a data scientist to assist with application development.

Big Data

Big Data analytics can help reveal relationships that are difficult to uncover by an analyst. Big data is data that is characterized by four properties:

1. Velocity describes the speed at which data is generated.
2. Volume describes the size of the data, usually terabytes, petabytes, or exabyte sizes.
3. Variety describes diversity of data formats (structured), no format (unstructured), or a combination of both (semi-structured).
4. Veracity describes the origins, quality, and accuracy of the data.

All data goes through a pre-processing step of refinement called Extraction, Transform, Load (ETL). Each stage in this process performs a different function:

1. Extract. Collecting the data from multiple sources, or streams.
2. Transform. This stage includes several processes, among them
    a. Selecting the data
    b. Cleaning the data to handle inconsistence, incompleteness or missing data'
    c. Normalizing data,
    d. Discretizing and reducing data,
    e. Ensuring statistical quality of data, and
    f. Understanding the data through descriptive statistics.
3. Load. Writing the data to storage.

In cyberoperations, most of the data is a byproduct of normal operations of the devices. Data diversity is big as a single network device might generate multiple data formats, for example logs, audits, and communication between devices in the network. Because the

number of devices in a network is large, and every device generates data, cyber data is a big data.

Limitations

Lack of planning, strategy, and lack of optimization can doom big data systems. Some of the common issues are [3], [4]

1. Big data systems are complex systems; therefore, they require a simple approach to engineering. For the system to be optimal, it must be decoupled, preferably following a microservice approach.
2. Some data transformation is required before storage. Unless the data has a common schema, it will be difficult to use for exploration.
3. Expect failures. Failures are common in a complex transaction. It is important that the system provides a mechanism to observe the status of any transaction.

Volume can be a challenge. Data storage can become expensive. The system needs a data retention policy based on consumer's needs, which specifies the lifecycle of the data.

Bibliography

[1] M. Zweben, "What Happened to Hadoop? What Should You Do Now?," 09 Aug 2019. [Online]. Available: https://dzone.com/articles/what-happened-to-hadoop-what-should-you-do-now. [Accessed 25 Oct 2019].

[2] Z. Zhan, A statistical framework for analyzing cyber attacks, The University of Texas at San Antoion, 2014.

[3] J. Klein, R. Buglak, D. Blockow, T. Wuttke and B. Cooper, "A reference architecture for big data systems in the national security domain," in *2016 IEEE/ACM 2nd International Workshop on Big Data Software Engineering (BIGDSE)*, 2016.

[4] Y. Zhao, A. Polk, S. Kallis, L. Jones, R. Schwamm and T. Kendall, "Big Data and Deep Models Applied to Cyber Security Data Analysis.," in *" workshop Adversary-aware Learning Techniques and Trends in Cybersecurity (ALEC) of the AAAI Fall Symposium*, 2018.

[5] R. E. Samuel, G. Cormier, S. Fascendini, C. M. Stubanas and K. A. Yacko, "Four It/Is Pillars for Artificial Intelligence Machine Learning/Deep Learning Applications," *Issues in Information Systems,* p. 149, 2018.

[6] P. Guo, "Data science workflow: Overview and challenges," *Communications of the ACM,* 2013.

[7] A. Maheshwari, Data analytics made accessible, Seattle: Amazon Digital Services, 2014.

[8] P. Ongsulee, "Artificial intelligence, machine learning and deep learning," in *15th International Conference on ICT and Knowledge Engineering (ICT&KE)*, 2017.

[9] K. Grace, C. Nunnery, J. Kraunelis, M. Colosimo, K. Nolan, C. Miniter and R. Charland, "Cyber Defense Analytics for the USAF Data Analytics Pathfinder," MITRE, Bedford, 2018.

Appendix A: Data Roles in Support of AI, ML, AND DL

This is a list of common roles that support data analytics and AI efforts. For an expanded list on technology skills and competency skills refer to [5], [6].

1. Data Architect. Assist with the holistic view of the problem domain. Defines the data that will be stored, consumed, integrated, and managed by different entities and Information Technology (IT) systems. Gather high-level business needs and requirements and design a solution. Required broad understanding on AI, ML, and DL technologies.
2. Data Engineer. Identifies and ingest data, build reliable data pipelines, prepare features for a predictive model, and transforms the data into formats that data scientist can consume. Focus on data preparation and manipulation. From ingestion to formatting/transformation to storage for consumption
3. Data Scientist. Combines data inferences and algorithm development to discover insight from structure and unstructured data.
4. Data Analyst. Retrieves, organizes, and performs quantitative analysis on data to generate reports that will identify trends to support decision making.

Appendix B – Glossary

Terms defined from a data science perspective.

Term	Definition
Data Lake	Single store of all enterprise data, typically unstructured or semi-structured data.
Attributes	Qualities or properties in the data that can be used for analysis.
Features	In ML/DL/AI, are variables selected from a dataset that have the most impact on the outcome. Features can be generated by combining several features. For example, a model used to predict the likelihood of an account defaulting might include a feature that is the difference between the savings balance and owed credit.
Training Model	In ML/DL/AI, a training model is a collection of coefficients that optimize predictions for a specific mathematical transformation. Examples of mathematical transformations are Bayesian method, logistic regression, neural networks, and perceptron.
Data Lake	Brochure level abstraction for data storage
Data Warehouse	Functional level abstraction for data storage