CORE Metadata, citation and similar papers at core.ac.uk
Provided by Jurnal Online Universitas Pertahanan (Indonesian Defense University)

DATA SCIENCE, BIG DATA, AND PREDICTIVE ANALYTICS: A PLATFORM FOR CYBERSPACE SECURITY INTELLIGENCE

SAINS DATA, BIG DATA, *DAN ANALISIS PREDIKTIF: SEBUAH LANDASAN UNTUK KECERDASAN KEAMANAN SIBER*

Dicky R. M. Nainggolan[1]
Defense Management Study Center, Indonesia Defense University
(dickronez@gmail.com)

Abstract– *Data are the prominent elements in scientific researches and approaches. Data Science methodology is used to select and to prepare enormous numbers of data for further processing and analysing. Big Data technology collects vast amount of data from many sources in order to exploit the information and to visualise trend or to discover a certain phenomenon in the past, present, or in the future at high speed processing capability. Predictive analytics provides in-depth analytical insights and the emerging of machine learning brings the data analytics to a higher level by processing raw data with artificial intelligence technology. Predictive analytics and machine learning produce visual reports for decision makers and stake-holders. Regarding cyberspace security, big data promises the opportunities in order to prevent and to detect any advanced cyber-attacks by using internal and external security data.*

Keywords: Big Data, Cyber Security, Data Science, Intelligence, Predictive Analytics

Abstrak – *Data merupakan unsur terpenting dalam setiap penelitian dan pendekatan ilmiah. Metodologi sains data digunakan untuk memilah, memilih dan mempersiapkan sejumlah data untuk diproses dan dianalisis. Teknologi big data mampu mengumpulkan data dengan sangat banyak dari berbagai sumber dengan tujuan untuk mendapatkan informasi dengan visualisasi tren atau menyingkapkan pengetahuan dari suatu peristiwa yang terjadi baik dimasa lalu, sekarang, maupun akan datang dengan kecepatan pemrosesan data sangat tinggi. Analisis prediktif memberikan wawasan analisis lebih dalam dan kemunculan machine learning membawa analisis data ke tingkat yang lebih tinggi dengan bantuan teknologi kecerdasan buatan dalam tahap pemrosesan data mentah. Analisis prediktif dan machine learning menghasilkan laporan berbentuk visual untuk pengambil keputusan dan pemangku kepentingan. Berkenaan dengan keamanan siber, big data menjanjikan kesempatan dalam rangka untuk mencegah dan mendeteksi setiap serangan canggih siber dengan memanfaatkan data keamanan internal dan eksternal.*

Kata Kunci: Analisis Prediktif, Big Data, Intelijen, Keamanan Siber, Sains Data

[1] Author holds a Bachelor in Informatics Engineering degree from UNLA, received his Master of Science degree in Defense and Security Management from Indonesia Defense University in joint curriculum with Cranfield Defence and Security, Cranfield University, England and earned the Defense Management Course certificate from Naval Postgraduate School, Monterey, California, United States of America.

Introduction

Big Data utilisation is becoming more real to help in decision-making process and analysing future trends in this day and age. The term of Big Data it first appeared in 1997 in the journal by Michael Cox and David Ellsworth[2], but people have begun to pay attention since the publication of the McKinsey Global Institute.[3] Since then, many scientists are developing Big Data opportunities for government policies, health care, law enforcement, cyberspace security, research and development, estimating economics productivity, energy management, natural disaster analysis and many more.

Promoting Big Data technology is proven by successfully helping many world-class companies like Google, Facebook, Twitter, Yahoo, and other data researcher companies. This fact implies not only about Big Data's potential but also significant challenges in each of its analytical steps.[4] Interestingly, Big Data is not always used by big companies, but with the convergence between Big Data and Cloud Computing technology, this can be useful for small and medium enterprises in innovation, competition and productivity either.

R. Agrawal, et al. (2014) explained that Big Data requires a critical analysis process in detecting and revealing the meaningful motives contained in the data.[5] Thus, the Big Data potential will be explored and can be accounted in almost all human activities. On the one hand, Big Data makes it possible to uncover the information contained in the data set. On the other hand, the policies and demands for Information Security and Data Privacy are increasingly being promoted. Information Security Issues and Big Data look like they were contradictions. In fact, these subjects are two sides of the same coin. Big Data can be exercised to formulate the right policies and strategies in order to support and to improve data security and data privacy. There were just

[2] G. Press, "A very short history of big data", *Forbes*, 2013, https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/, retrieved 20 May 2017

[3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, & A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity", McKinsey Publication, 2011.

[4] C. P. Chen & C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", *Information Sciences* Vol. 275 2014, pp. 314-347.

[5] R. Agrawal, A. Imran, C. Seay, & J. Walker, "A layer based architecture for provenance in big data", In *Big Data (Big Data), 2014 IEEE International Conference on* Big data, 2014, pp. 1-7.

a small number of publications and research regarding this issue [6].

The national development programme cannot be separated from the development of vital infrastructures that they are now increasingly dependent on the use of information and communication technology (ICT). The benefits of ICT were proven to improve performance in management process becoming effective and efficient. Therefore, it can alleviate financial burden or can increase profits. ICT promises an unlimited connection but also unleashes unlimited potential attacks as well. It raises the vulnerability systems up to various direction of cyber attack. Thus, the connection between infrastructure development and semantic technology mechanisms, the form of a number of different technologies collaboration, are highly needed.[7]

This paper explores a means of improving cybersecurity using Big Data and determining the direction of relevant predictive analysis research.

Data Science

Data-Based Science or Data Science is an interdisciplinary study that explores a scientific method and how to extract knowledge or insight from data clans in various forms, not only structured but also unstructured.[8] Stages in the data science are similar to those in 3 stages of Knowledge Discovery in Database (KDD):

(1) Pre-processing Data to analyse multivariate data sets prior to data mining. Pre-processing also aims to clean the data by eliminating data from observations that noise or data lost.

(2) Data Mining generally involves six task groups[9]:

- Anomaly detection (deviation detection)
-Identify uncommon data records, interesting data or data errors that require further investigation.
- Association rule learning (dependency modelling)
- Finding relationships between variables For example, supermarkets can collect data about consumer buying habits By using association rules of study,

[6] L Xu, C. Jiang, J. Wang, J. Yuan, & Y. Ren, "Information security in big data: privacy and data mining", *IEEE Access*, 2, 2014, pp. 1149-1176.

[7] Wu, X., X. Zhu, X., G. Q. Wu, and W. Ding, "Data Mining With Big Data." *IEEE Transactions On Knowledge And Data Engineering 26 (1)*, 2014, pp. 97-107.

[8] Vashant Dhar, "Data science and prediction", *Communications of the ACM*, Vol. 56, No.12, 2013, pp. 64-73.

[9] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases" *AI magazine*, 1996, Vol. 17, No.3, pp. 37.

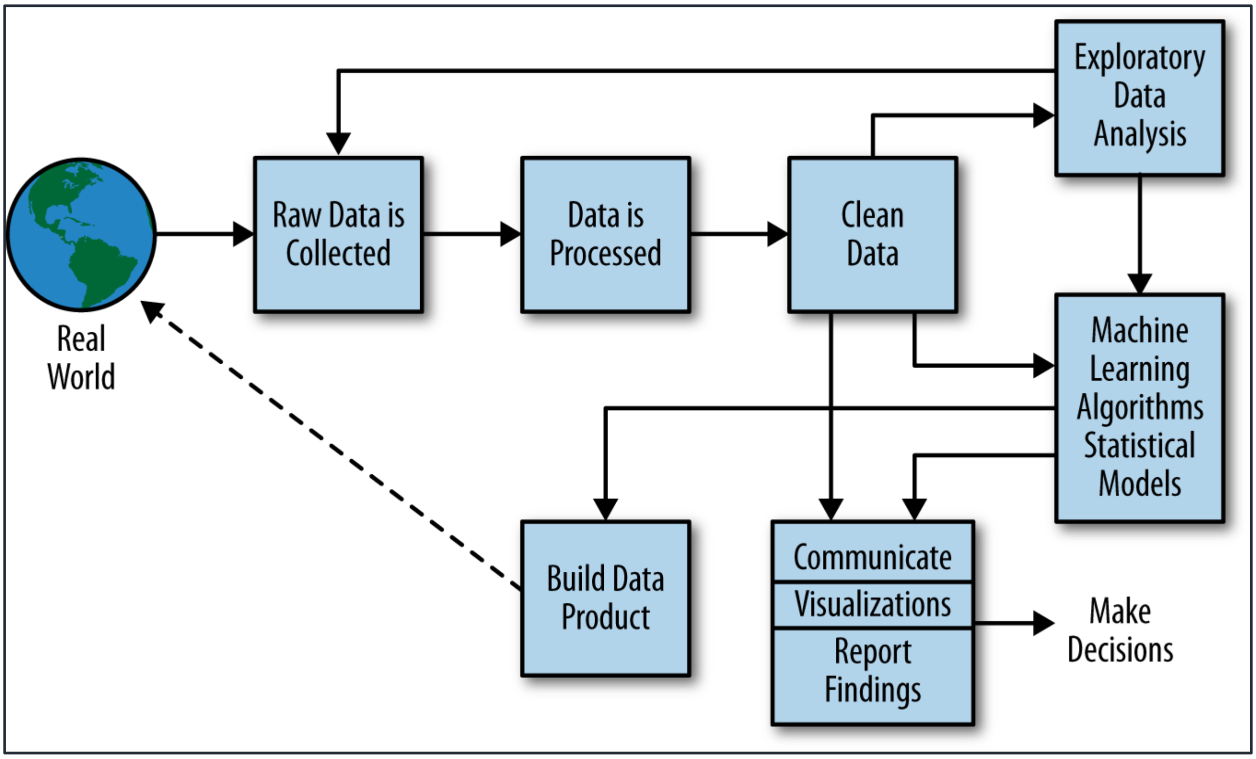
supermarkets can determine which products are often purchased together and use this information for marketing purposes. Referred to as a market basket analysis.

- Clustering - is the task of finding groups and structures in data that are in various ways.
- Classification - is the generalisation task of known structures to be applied to new data.
- Regression - trying to find functions that model data with little error.
- Summarisation - provides a solid data representation, including visualisations and reports.

(3) Result Validation is a structured understanding and giving meaning to the results of the same type of prediction or find the same general pattern type regardless of the test data provided.

The figure 1 below displays visualisation of data science flowchart.[10] Data collected from environment, represented by the world. Data is "purged" or processed to generate a data

Figure 1. Data science process flowchart.



Source: R Schutt & C. O'Neil, "Doing Data Science: Straight Talk from the Frontline", O'Reilly Media, Inc, 2013.

[10] R. Schutt, and C. O'Neil, "*Doing data science: Straight talk from the frontline*", O'Reilly Media, Inc.", 2013, pp. 41.

set (usually a data table) that can be used for processing. Then, explorative data analysis and statistical modelling are being performed. "Data products" are used to suggest decisions based on activity histories. Lastly, it creates data and delivers feedback into the environment.

C. Hayashi, et al. (2013) explained a concept to unify statistics, data analysis and related methods are used in data science in order to understand and to analyse actual phenomena.[11] It uses techniques and theories drawn from various fields in the field of study covering mathematics, statistics, information science, and computer science.

Big Data

In May 2011, McKinsey Global Institute describes the big data main elements and ecosystem. Such as:

1. Techniques for analysing data, such as A/B testing, machine learning and natural language processing.
2. Big data technologies, like business intelligence, cloud computing and databases.
3. Visualisation, such as charts, graphs and other displays of the data.

With regard to these components, big data often fetches three stages of data processing. Firstly, there are several techniques to conduct data mining. Secondly, special technologies are applied to collect, to process and to analyse vast amount of data. Lastly, predictive analysis result are often displayed by visual to illustrate the trend.

There is no single absolute definition of big data. People identify the same major characteristics of the big data as “3V”. Such as, Volume, Velocity, and Variety (structured, semi-structured, and unstructured data).[12] In 2013 IBM added the fourth V, called Veracity (quality and data integrity)[13], whilst Value (highlighting the added value) was proposed by Oracle into big data characteristics in the same year.[14]

[11] C. Hayashi, K. Yajima, H. Bock, N, Ohsumi, Y, Tanaka, and Y, Baba, (Eds.), “Data Science, Classification, and Related Methods”, Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27–30, 1996, Springer Science & Business Media, 2013.

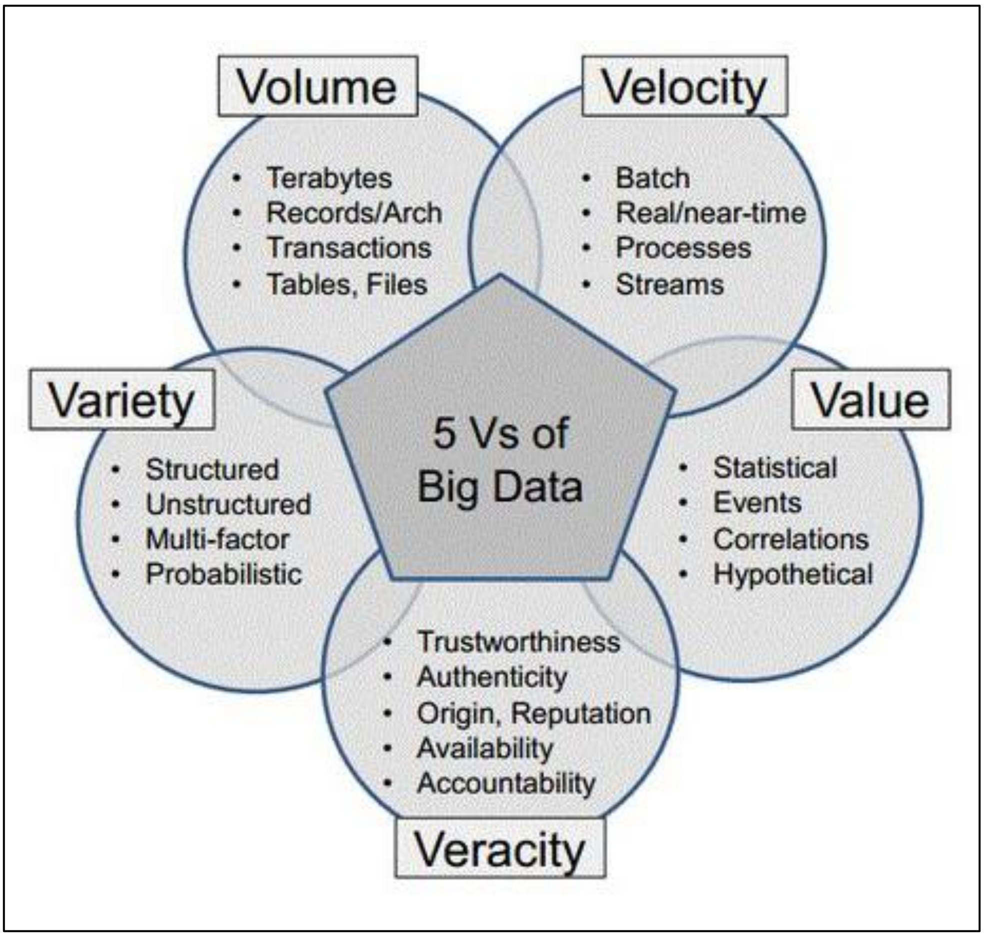
[12] P. Zikopoulos, & C. Eaton, “Understanding big data: Analytics for enterprise class hadoop and streaming data”, McGraw-Hill Osborne Media, 2011.

[13] S. B Siewert, “Big data in the cloud: Data velocity, volume, variety, veracity”, IBM Developerworks, 2013.

[14] Doug Cackett,"Information Management and Big data: A Reference Architecture", *Oracle: Redwood City*, CA, USA, 2013.

Nowadays, many researchers and scholars describe big data characteristics with "5V". Figure 2 depicts 5Vs of Big Data characteristics below.

Figure 2. What is Big data (Courtesy of Data Technocrats)



Source: D. Kim, 제3회사내기술세미나-hadoop (배포용)-dh kim-2014-10-1, 2014.

Some of the advantages that drive big data usage:

1. Data storage costs have decreased over the years. Traditional data storage is only able to store data in a certain period of time. Big data can save indefinitely so it is useful for long term trends.
2. Some big data Tools like Hadoop currently provide query processing technology and analysis with fast and complex processing.
3. The data warehouse scheme is flexible that the Extract, Transform, and Load (ETL) process can vary for structured and unstructured data. This is hard to do in traditional data warehouses.

[15] D. Kim, 제3회사내기술세미나-hadoop (배포용)-dh kim-2014-10-1, https://www.slideshare.net/jihanie/3-hadoopdh-kim2014101, retrieved May 28, 2017.

Groups of big data technology divided by two major types of data processing:

Table 1. Comparison sheet of Big Data Processing

	Batch Processing	Interactive Analysis	Stream Processing
Query Runtime	Minutes to Hours	Milliseconds to Minutes	Never-ending
Data Volume	TBs to PBs	Gbs to PBs	Continous Stream
Programming Model	MapReduce	Queries	DAG
User	Developers	Analyst and Developers	Developers
Originating Project	Google MapReduce	Google Dremel	Tweeter Storm
Open Source Project	Hadoop / Spark	Drill / Shark / Impala / HBase	Storm / Apache S4 / Kafka

Source: A. Solanki, "Big data–Back to the Basics", 2013.

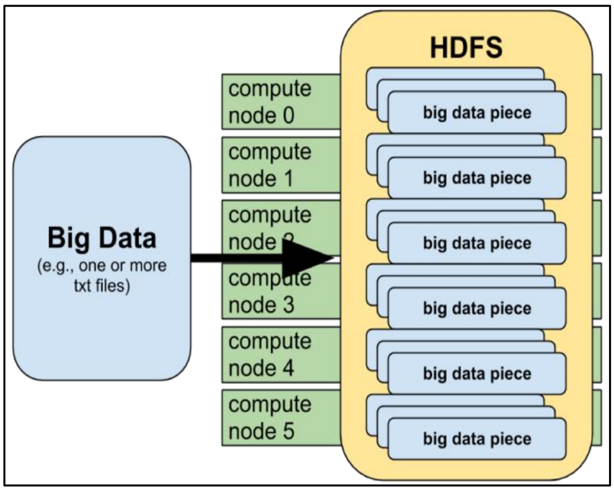
1. Batch processing, data is analysed when all data has been collected and processed all at once. The most popular batch processing technology is Hadoop.
2. Stream processing, data is analysed as stream data entry.

There is one minor type of data processing called Interactive Analysis that provides an alternative choice. These processing comparison is displayed concisely in table 1 below.

Hadoop Distributed File System (HDFS) is the Hadoop framework to distribute devided data file system in data warehouse into data blocks and allocates them to specified server locations via TCP/IP. Data servers have a damaged data tolerance system. It means if there is data failure, it does not affect the general performance of the system. This illustrates in figure 3.[17]

In addition to HDFS, Hadoop has a vast software ecosystem. Many of Hadoop's components are Apache project development that produces open source software so it will be frequently updated and serves the needs of big data

Figure 3. HDFS distributed files across multiple physical compute nodes



Source: G. Lockwood, "Conceptual Overview of Map-Reduce and Hadoop", 2015.

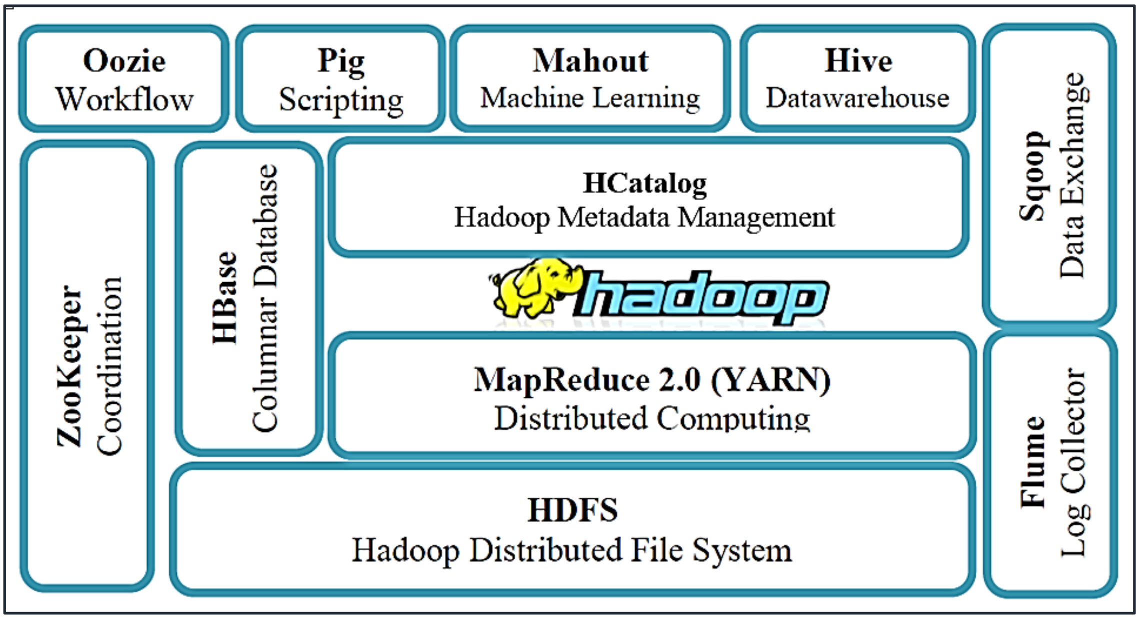
[16] A. Solanki, "Big data–Back to the Basics", 2013, http://trinityordestiny.blogspot.co.id/2013/07/big-databack-to-basics.html, retrieved June 8, 2017.

[17] G. Lockwood, "Conceptual Overview of Map-Reduce and Hadoop", 2015, http://www.glennklockwood.com/data-intensive/hadoop/overview.html, retrieved June 5, 2017.

analysis. Some of them, which is shown in figure 4 below[18], are:

- MapReduce programming technology (in Java) of Google in customising some large-scale data processing problems by means of parallel and distribution processing. Two steps processing in MapReduce are Map and Reduce.

Figure 4. Hadoop Ecosystem



Source: R. Alguliyev & Y. Imamverdiyev, "Big data: big promises for information security, In Application of Information and Communication Technologies (AICT), 2014

- Apache Pig component consists of a compiler that generates a MapReduce programme sequence with a "Pig Latin"
- HCatalog functions for management and data tables created in Hadoop and supports the ongoing functionality of Pig, MapReduce, Streaming and Hive etc.

- HBase (Hadoop Database) is a Distributed Column Database (Google's BigTable).

- Zookeeper stores coordination information, distributed synchronisation, naming, and group services. This is important in distributed systems method.

- Mahout is machine learning software, including key algorithms, such as classification, grouping, and recommendations and collaborative services. This is important in distributed systems method.

- Mahout is machine learning software, including key algorithms, such as classification, grouping, and

[18] R. Alguliyev & Y. Imamverdiyev, "Big data: big promises for information security, In *Application of Information and Communication Technologies (AICT), IEEE 8th International Conference on Oct 15, 2014, pp. 1-4.*

recommendations and collaborative filtering implemented with the Map/Reduce paradigm at the top-level of Hadoop.

- Components like Sqoop and Flume, included in the ecosystem used to transmit data to the Hadoop cluster and vice versa.

In 2015, big data has become a tool in organisations to operate efficient by analysing internal and external data collections. By implementing the big data framework, machine learning concepts and deep computing, organisations can predict potential problems in the future and it capable to take immediate respond to avoid or to minimise the negative impacts that will arise.[19] The integration of management system with big data provide a platform for data mining from the collection of individual data and maps the insights of the entire system.

Predictive Analytics

Various statistical techniques such as data mining, predictive modelling and machine learning are used to analyse the latest phenomena and historical track record to predict future events or to study a mystery event. Predictive analytics provides a level of probability based on information retrieval, the meaning of the relationship between multiple data variables, and information enrichment, thus it predicts patterns of behaviour in a certain period of time (past, present, or future) that accurately depends on the quality of the analysis and assumptions applied.

Predictive analysis differs from forecasting because it provides a predictive score for each element of its analysis. The probability value of this analysis can be integrated into prescriptive analytics in the decision-making process and preventing potential problems to attain near-zero failure.

In carrying out predictive analytics procedures, rigorous data analysis is a necessity that it results the scientific recommendations in order to make make decisions or long-term policies. There are three popular models, namely:

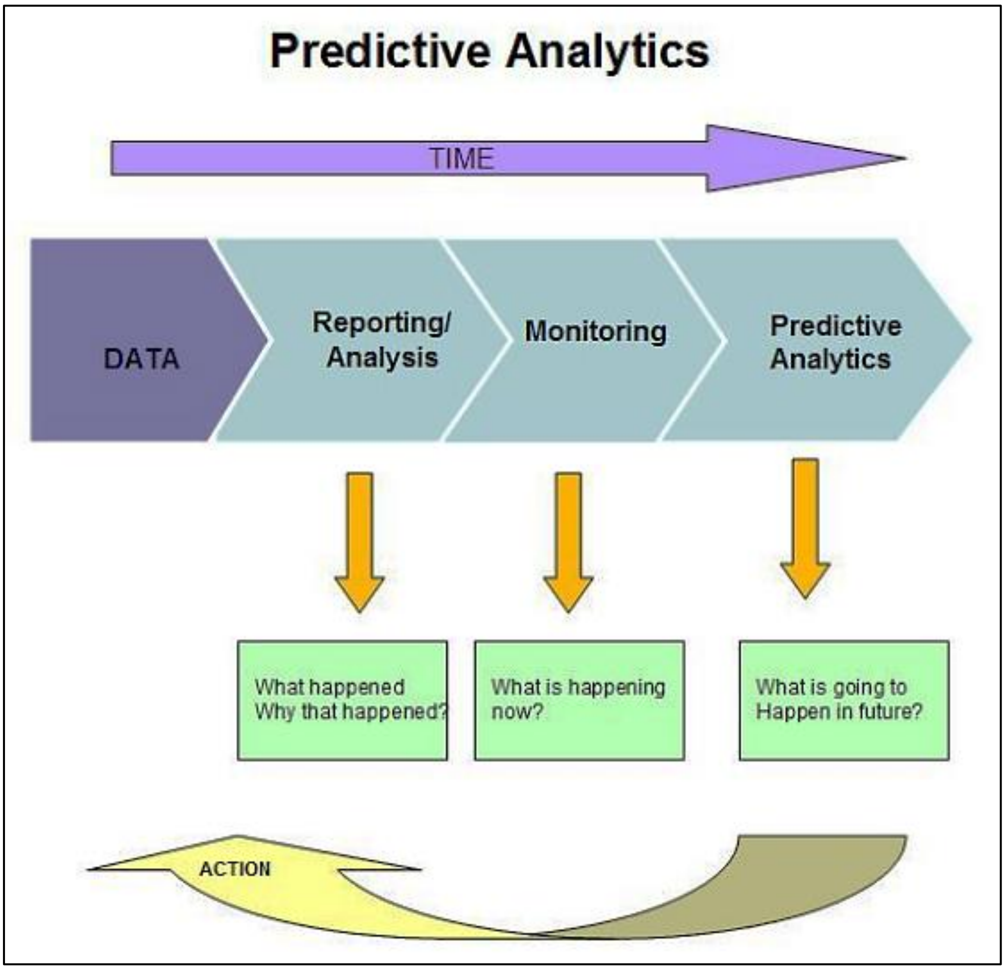
1. Predictive Model that builds relationship model to analyse a specific unit performance with several different samples.
2. Descriptive Model that is modelling group behavioural patterns based on decision preferences.

[19] R. Solnik,"The Time Has Come: Analytics Delivers for IT Operations", 2013, http://www.datacenterjournal.com/time-analytics-delivers-operations/, retrieved June 3, 2017.

3. Decision Model which describes the relationship of all analytic elements with many variables in predicting and developing the logic final decision.

The predictive analytics cycle can be illustrated in the figure 5 below.[20]

Figure 5. Predictive Analysis Cycle



Source: Imanuel,"What Is Predictive Analytics ?" Predictive Analytics Today, 2017.

Analytical Techniques

Analytical for predictive analytics techniques are extensively grouped into regression and machine learning techniques.

Regression Techniques

The regression model technique is done by forming mathematical equations as representative of the interaction model amongst the various variables applied in certain situations. Therefore, they produce qualified predictive analysis. There are ten main models of these regression techniques, namely Linear Regression Model, Discrete Choice Models, Logistic Regression, Multinomial Logistic Regression, Probit Regression, Logit Versus Probit, Time Series Models, Survival or Duration Analysis,

[20] Imanuel,"What Is Predictive Analytics ?" *Predictive Analytics Today* 2017, http://www.predictiveanalyticstoday.com/what-is-predictive-analytics, retrieved June 6, 2017.

Classification and Regression Trees (CART), and Multivariate Adaptive Regression Splines (MARS).

Machine Learning Techniques

Machine learning is one branch of artificial intelligence technology enabling computers to learn with the ability to process numbers of advanced statistical methods for regression and classification. In certain cases, this technique can directly predict the dependent variable without the need to define the base relationship between variables that can formed in complex states and the mathematical calculation of the dependencies is unknown. This technology synthesises human cognitive algorithms and able to learn from several examples of cases in its programming syntaxes. Seven machine learning common methods are Neural Networks (NN), Multilayer Perceptron (MLP), Radial Basis Functions, Support Vector Machines, Naïve Bayes, *k*-Nearest Neighbours (KNN), and Geospatial Predictive Modeling.

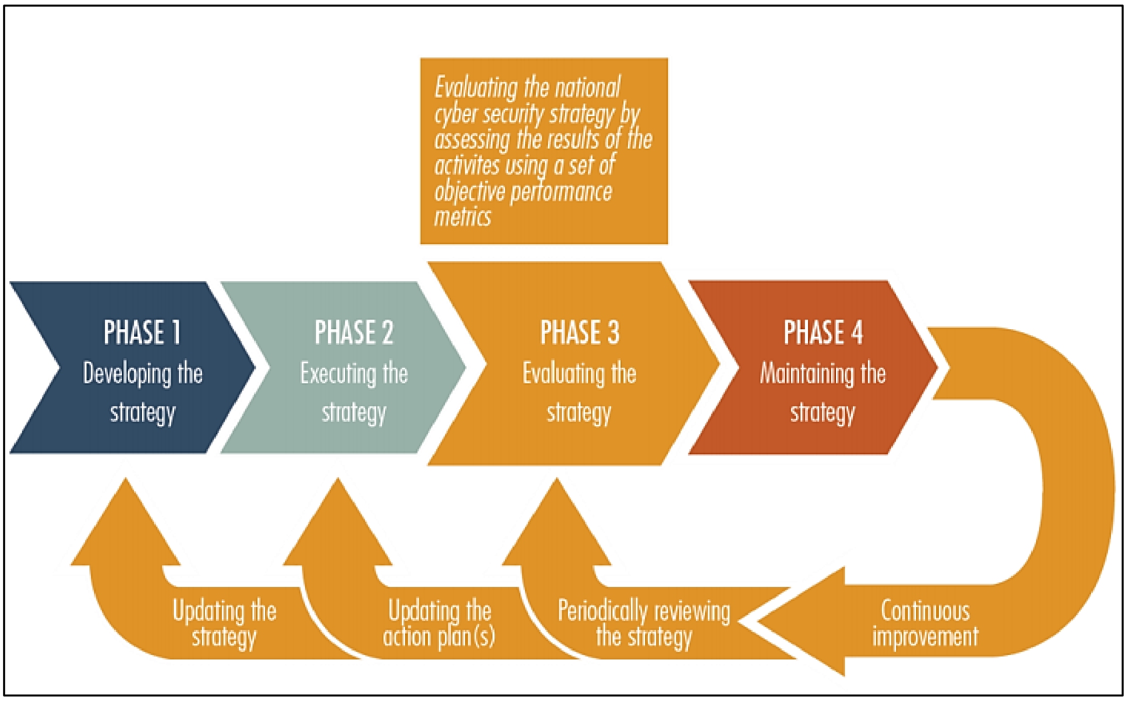
Cyberspace Security

Cyberspace security or ICT security protects the data and its network system from unauthorised access. This security prevents the entire system from whether being compromised from any illegal mining and modifications. The Big Data technology offers high-speed, large-scale heterogeneous dataset analysis capabilities that can be used to expand conventional information security systems with storage, data maintenance facilities. The technology also enables complex data analysis, data comparison, anomaly detection in the system, and the prevention of cyber threats in real-time.

Cyber threats and cyber attacks might come from any directions or sources such as, domestic or foreign countries, individual or groups, and terrorists or state/non-state actors. Cyber intrusions are become more often, more advanced, and more fatal by pointing their targets to companies or organisations, financial institutions, government agencies, schools, and citizens. European Union Agency for Network and Information Security (ENISA) published a national cybersecurity strategies life cycle that consist of 4 phases from developing, executing, evaluating, to maintaining the strategy.[21]

[21] N Falessi, R Gavrila, MR Klejnstrup, K Moulinos. National cyber security strategies: practical guide on development and execution. *European Network and Information Security Agency (ENISA) Publication*, 2012.

Figure 6. Life cycle of a national cybersecurity strategy


Evaluating the national cyber security strategy by assessing the results of the activites using a set of objective performance metrics
PHASE 1
Developing the strategy
PHASE 2
Executing the strategy
PHASE 3
Evaluating the strategy
PHASE 4
Maintaining the strategy
Updating the strategy
Updating the action plan(s)
Periodically reviewing the strategy
Continuous improvement

Source: N Falessi, et al. National cyber security strategies: practical guide on development and execution. European Network and Information Security Agency (ENISA), 2012.

This life cyle model that shown in figure 6 above, describes not only one-off evaluations, but also ongoing.

S. Curry, et al. (2013) from RSA, the cyber security company, recommend the Intelligence-Driven Technology model that has a much larger data analysis capability and is more diverse than the conventional SIEM (Security Information and Event Management) model.[22] Broadly speaking, the principles of Intelligence-Driven Security parameters for large scale systems are:

- using advanced monitoring subsystems in order to monitor multiple layer of sources. They are also capable in combining informations from different sources;
- having an automated tool that accumulate and process Big Data then it can produce standardise format reports. These outputs are accessible to subsystems as well;
- integrating with central repository, great analytical and visualisation tools to exploit information and knowledge from raw data.

Despite the advantages of using Big Data to address cyber security, there are some challenges that worth to consider, which are:

1) **Privacy.** The privacy preserving application in big data is not yet fully

[22] S Curry, E Kirda, E Schwartz, WH Stewart, & A Yoran. Big data fuels intelligence-driven security. *RSA Security Brief Report*, 2013.

developed, therefore the exposure of the privacy data to end-user makes privacy violation easier.

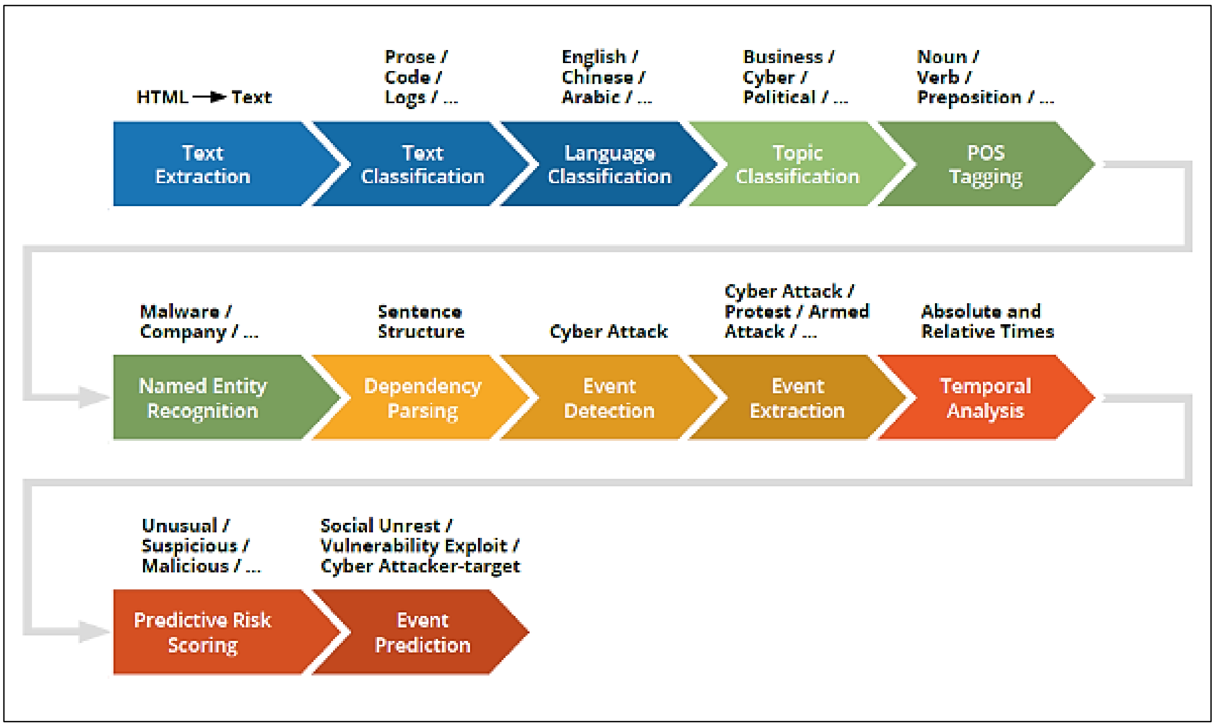
2) **Big Data analytics for APT detection.** Detecting Advanced Persistent Threat (APT) requires a more sophisticated real-time sensors including new detection algorithms and reliable multi-source data processing.

3) **Supercharged cryptography.** Algorithms of encryption and decryption; attribute-based encryption, encrypted data search; availability, reliability, and integrity attacks. [23]

4) **Security research on Big Data datasets.** It is extremely difficult to understand the number of significant cybersecurity data, and it is almost impossible to research from organically collected datasets.

5) **Data provenance problem.** It is difficult to ensure that the quality of each data source from the aspect of authenticity and integrity is in order to meet the need for an analytical algorithm. In order to produce accurate results, machine learning and robust statistics contributions are vital in this role.

Figure 7. A combination of rule-based, statistical, and machine-learning techniques


HTML → Text
Prose / Code / Logs / ...
English / Chinese / Arabic / ...
Business / Cyber / Political / ...
Noun / Verb / Preposition / ...
Text Extraction
Text Classification
Language Classification
Topic Classification
POS Tagging
Malware / Company / ...
Sentence Structure
Cyber Attack
Cyber Attack / Protest / Armed Attack / ...
Absolute and Relative Times
Named Entity Recognition
Dependency Parsing
Event Detection
Event Extraction
Temporal Analysis
Unusual / Suspicious / Malicious / ...
Social Unrest / Vulnerability Exploit / Cyber Attacker-target
Predictive Risk Scoring
Event Prediction

Source: Staffan Truvé, "Machine Learning in Cyber Security: Age of the Centaurs", Threat Intelligence Whitepaper, Recorded Future, 2017

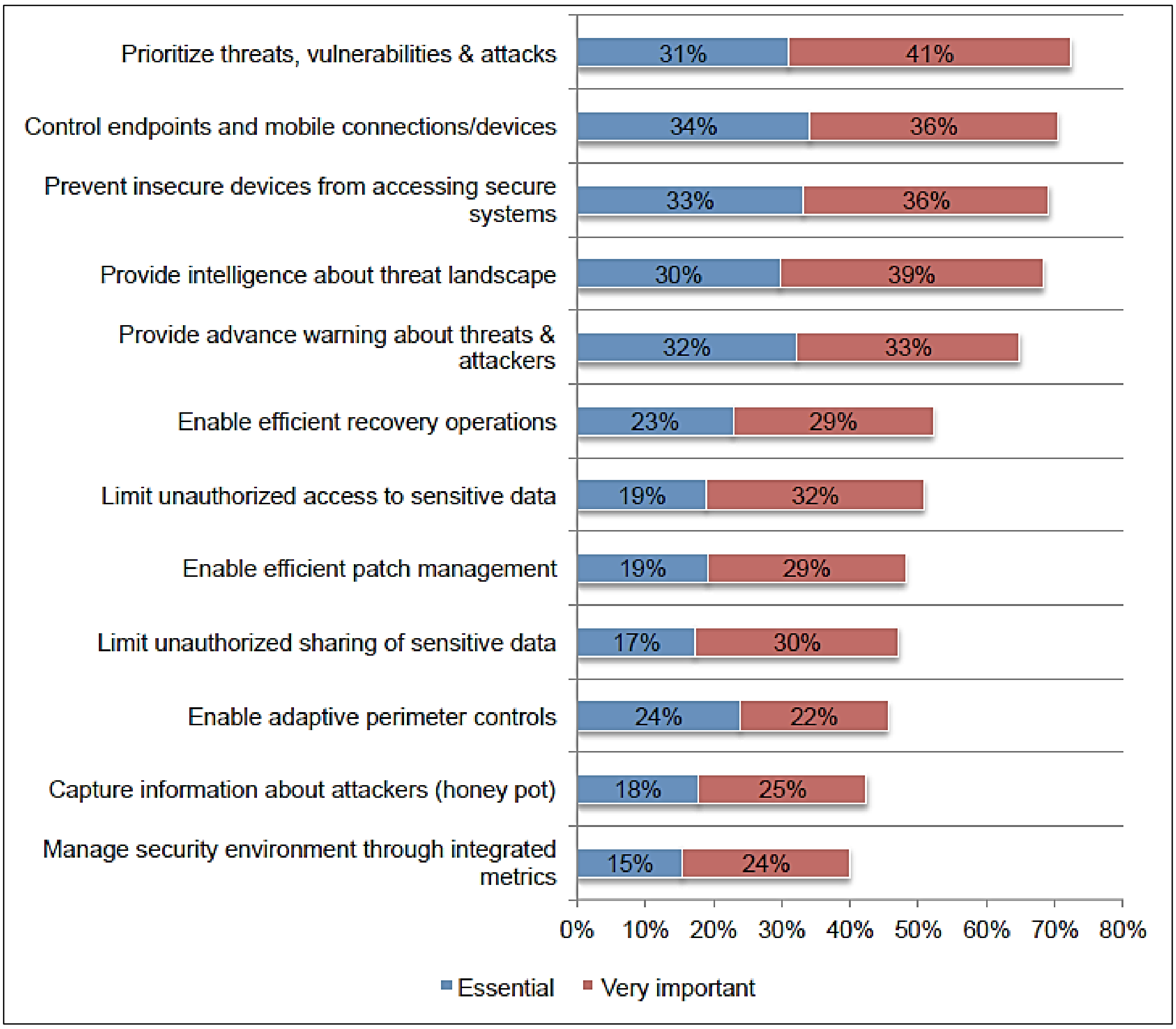
[23] U. Ganugula, & A. Saxena, "High Performance Cryptography: Need of the Hour" *CSI Communications Magazine*, 2012, pp.15.

6) **Security visualisation.** Big data visualisation remains at elementary level, dominated by graphs, pie charts, and spreadsheet pivot tables. Soon there will be an improvement to visualise big data analytics reports.

7) **Skilled personnel.** A collaborative team is needed as an important element of successful implementation of big data technology for information security. The team contains staff with specialised expertise such as data processing expertise, datasets analysis, cyber attack analysis, and decision scientist. This collaborative team and big data technology provide optimal results to the organisation.

The Recorded Future processing pipeline uses machine learning and rule-based algorithms to transform

Figure 8. The most important features used today for security technologies



Source: Anonymous, "Big Data Analytics in Cyber Defense", Ponemon Institute Research Report, 2013.

unstructured information from the web into actionable threat intelligence,[24] as shown in figure 7 below.

In 2013, Ponemon Institute LLC published a research report called Big Data Analytics in Cyber Defense[25]. They surveyed 706 IT security practitioners in numbers of fields including financial services, manufacturing, and government sectors with an average of 10 years experience. All respondents were familiar with their organisation's defence against cyber security attacks and had some level of responsibility for managing the cyber security activities within their organisation. The Institute organised research regarding to the four main topics:

- Perceptions about cyber readiness;
- Cyber security risks, vulnerabilities and consequences;
- Big data analytics & cyber security solutions;
- Industry differences.

Relating to big data analytics and cyber security topic, the survey discovered the most significant features for security technologies, according to respondents are: ability to prioritise threats, vulnerabilities and attacks; control of endpoints and mobile connections and devices; prevent insecure devices from accessing secure systems, provide intelligence about threat landscape and an advance warning about attackers as shown in figure 8[26].

It was also noticeable that in term of strengthening securities, respondents would most like big data analytics to be combined with anti-virus/anti-malware, anti-DoS/DDoS, security intelligence systems (SIEM) and content aware firewalls as shown in Figure 9[27]. Respondents want to know potential future threats and future data exfiltration as well.

People are now in the planning stages for advanced research and development of Big Data analysis to better recognise, to understand, and to respond to cyber-attacks completely. The key to success lies in focusing on one narrow aspect of cyber defence. If this is successfully implemented and shows an accurate quality of analysis, the scale can be expanded and be used to estimate the resources needed for development and implementation in many larger

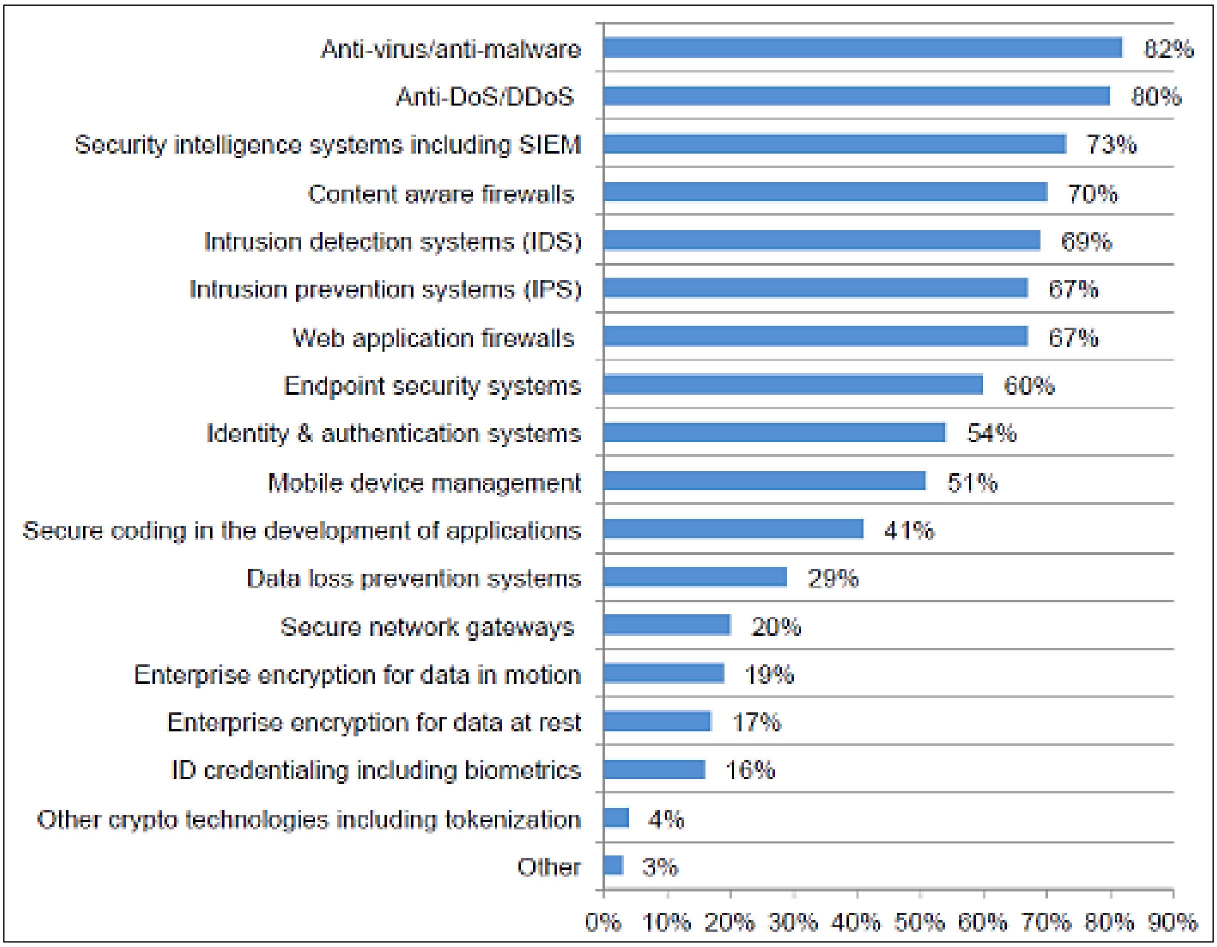
[24] Staffan Truvé, "Machine Learning in Cyber Security: Age of the Centaurs", *Threat Intelligence Whitepaper, Recorded Future*, 2017, pp.6.

[25] Anonymous, "Big Data Analytics in Cyber Defense", *Ponemon Institute Research Report*, 2013.

[26] Ibid, p.11

[27] Ibid, p.13

Figure 9. Enabling technologies combined with big data analytics for better



Source: Anonymous, "Big Data Analytics in Cyber Defense", Ponemon Institute Research Report, 2013.

organisational environments.

Conclusion

The data-based science technology of big data and predictive analytics tools for cyberspace security intelligence has a sophisticated capability to mine information from the structured and unstructured data at a low cost or even free. Cyber risk would continuelly increase following of a new technology invention. Big Data for security allows researchers detect and respond to any threat in real-time. In spite of all the advantages, big data has also some issue that need to cope with. For example, Data Privacy and Provenance, Datasets Management, and Human-computer interaction.

This highly promising analytics paradigm of large heterogeneous data volumes changes the information security and cyber world game plans. Advancing the big data features for advanced persistent threat detection, data leakage detection, computer forensic, cyber crime intelligence, and behaviour or trend visualisation receive a lot of researchers attentions. This appreciation would bring a significant impact to industry development, including cyber security policy and security company. Governments, armed forces and other organisations that have responsibility for cyberspace security should figure to harvest the vast amounts of multi-layer data for intelligent decision-making.

References

Books

P. Zikopoulos, & C. Eaton, "Understanding big data: Analytics for enterprise class hadoop and streaming data", McGraw-Hill Osborne Media, 2011.

R. Schutt, and C. O'Neil, "*Doing data science: Straight talk from the frontline*", O'Reilly Media, Inc.", 2013.

Journals

C. P. Chen & C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", *Information Sciences* Vol. 275, 2014.

L Xu, C. Jiang, J. Wang, J. Yuan, & Y. Ren, "Information security in big data: privacy and data mining", *IEEE Access*, 2, 2014.

Vashant Dhar, "Data science and prediction", *Communications of the ACM*, Vol. 56, No.12, 2013.

Wu, X., X. Zhu, X., G. Q. Wu, and W. Ding, "Data Mining With Big Data." *IEEE Transactions On Knowledge And Data Engineering* 26 (1), 2014.

Publications

Anonymous, "Big Data Analytics in Cyber Defense", *Ponemon Institute Research Report*, 2013.

Doug Cackett,"Information Management and Big data: A Reference Architecture", *Oracle: Redwood City*, CA, USA, 2013.

J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, & A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity", McKinsey Publication, 2011.

N Falessi, R Gavrila, MR Klejnstrup, K Moulinos.National cyber security strategies: practical guide on development and execution. *European Network and Information Security Agency (ENISA) Publication*, 2012.

S. B Siewert, "Big data in the cloud: Data velocity, volume, variety, veracity", IBM Developerworks, 2013.

S Curry, E Kirda, E Schwartz, WH Stewart, and A Yoran, A. Big data fuels intelligence-driven security. *RSA Security Brief Report*, 2013.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases" *AI magazine*, 1996.

U. Ganugula, & A. Saxena, "High Performance Cryptography: Need of the Hour" *CSI Communications Magazine*, 2012.

Staffan Truvé, "Machine Learning in Cyber Security: Age of the Centaurs", *Threat Intelligence Whitepaper, Recorded Future*, 2017.

Conferences

C. Hayashi, K. Yajima, H. Bock, N, Ohsumi, Y, Tanaka, and Y, Baba, (Eds.), "Data Science, Classification, and Related Methods", Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27–30, 1996, Springer Science & Business Media, 2013.

R. Agrawal, A. Imran, C. Seay, & J. Walker, "A layer based architecture for provenance in big data", In *Big Data (Big Data), 2014 IEEE International Conference on* Big Data, 2014.

R. Alguliyev & Y. Imamverdiyev, "Big data: big promises for information security, In *Application of Information and Communication Technologies (AICT), IEEE 8th International Conference on* *Oct 15, 2014.*

Websites

A. Solanki, "Big data–Back to the Basics", 2013, http://trinityordestiny.blogspot.co.id/2013/07/big-databack-to-basics.html, retrieved June 8, 2017.

D. Kim, 제3회사내기술세미나-hadoop (배포용)-dh kim-2014-10-1, https://www.slideshare.net/jihanie/3-hadoopdh-kim2014101, retrieved May 28, 2017.

G. Lockwood, "Conceptual Overview of Map-Reduce and Hadoop", 2015, http://www.glennklockwood.com/data-intensive/hadoop/overview.html, retrieved June 5, 2017.

G. Press, "A very short history of big data", *Forbes*, 2013, https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/, retrieved 20 May 2017.

Imanuel,"What Is Predictive Analytics ?" *Predictive Analytics Today* 2017, http://www.predictiveanalyticstoday.com/what-is-predictive-analytics, retrieved June 6, 2017.

R. Solnik,"The Time Has Come: Analytics Delivers for IT Operations", 2013, http://www.datacenterjournal.com/time-analytics-delivers-operations/, retrieved June 3, 2017.