Janne Voutilainen

MACHINE LEARNING AND INTELLIGENCE CYCLE: ENHANCING THE CYBER INTELLIGENCE PROCESS

UNIVERSITY OF JYVASKYLA
FACULTY OF INFORMATION TECHNOLOGY
2019

ABSTRACT

Voutilainen, Janne
Machine Learning And Intelligence Cycle: Enhancing The Cyber Intelligence Process
Jyvaskyla: University of Jyvaskyla, 2019, 63 pp.
Computer Science (Cyber Security), Master's Thesis
Supervisor: Lehto, Martti

Finding an indication from open sources to reveal a malicious cyber phenomenon is a demanding task. The information that is produced from the strategic cyber intelligence processes with, large-scale organizations can better prepare for cyber-attacks. The study aims to answer the question: Can Machine Learning (ML) be utilized for strategic open source cyber intelligence.

In 2019, e-criminals have adopted new tactics to demand enormous ransoms in bitcoins from large-scale organizations by using malicious ransomware software. The phenomenon is called Big Game Hunting. In the study, Big Game Hunting was used as an example for a target that was investigated with strategic cyber intelligence.

The answers to the research questions were achieved with The Design Science Research Process. The Design Science Cycle was conducted two times. In the first solution, a custom ML model was created precisely for the intelligence direction. The queried data was a limited dataset that was provided by the National Cyber Security Centre of Finland. The model returned correct data, but in the perspective of intelligence direction, the information was insufficient. In the second solution, the queries were made from the IBM Watson Discovery News data-set. The results offered enough valuable intelligence information about Big Game Hunting.

When the intelligence cycle and ML were combined, the main findings were that in information collection, the correct queries offered the best information. Furthermore, the short sentences, passages created by the Watson algorithm in the first solution proved to be useful. In information procession with unsupervised learning, the Watson algorithm was able to label the data in entities. The entities enabled the ability to analyse the data and find new, hidden information. The conclusion from the research was that ML could be utilised in strategic cyber intelligence.

Keywords: Cyber Security, Intelligence, Machine Learning

TIIVISTELMÄ

Voutilainen, Janne
Koneoppiminen ja tiedusteluympyrä: kybertiedustelun parantaminen
Jyväskylä: Jyväskylän yliopisto, 2019, 63s.
Tietojenkäsittelytiede (Kyberturvallisuus), pro gradu-tutkielma
Ohjaaja: Lehto, Martti

Vihamieliseen kyberilmiöön viittavan indikaation löytäminen avoimista lähteistä on vaativa tehtävä. Tieto, jota strateginen kybertiedustelu tuottaa, mahdollistaa suurten yritysten varautumisen kyberhyökkayksiin. Tutkimuksessa vastataan kysymykseen: Voidaanko koneoppimista hyödyntää strategisessa avoimen lähteiden kybertiedustelussa?

Vuonna 2019 kyberrikolliset alkoivat käyttää uutta taktiikkaa, jossa he vaativat suuria rahasummia yrityksiltä käyttämällä kiristyshaittaohjelmia. Ilmiön nimi on Big Game Hunting. Tutkimuksessa ilmiötä käytettiin strategisen kybertiedustelun esimerkkikohteena.

Tutkimustulokset saavutettiin suunnittelututkimuksella. Tutkimuksessa tehtiin kaksi suunnittelututkimuksen kierrosta. Ensimmäisen kierroksen tuloksena syntyi koneoppimismalli, joka suunniteltiin tiedusteluohjauksen mukaisesti. Kyberturvallisuuskeskus antoi rajoitetun datan, josta mallilla etsittiin tietoa Big Game Hunting ilmiöstä. Malli kykeni löytämään tietoa, mutta tiedusteluohjauksen kannalta tieto oli riittämätöntä. Toisen kierroksen tuloksena syntyneessä ratkaisussa tietoa haettiin IBM Watson Discovery News tietokannasta. Haut tuottivat riittävästi tiedustelutietoa ilmiöstä.

Kun koneoppimen ja tiedusteluprosessi yhdistettiin, tärkeimmät havainnot olivat, että oikeanlaiset kyselyt tuottavat parhaan tiedon tiedonkeräykseen. Lisaksi lyhyet Watson-algoritmin tuottamat virkkeet osoittautuivat hyödyllisiksi. Koneoppiminen helpotti tiedon prosessointia luomalla ohjaamattomalla oppimisella dokumentteihin metatietoa, jonka perusteella tieto jaettiin sopiviin kokonaisuuksiin. Kokonaisuudet mahdollistavat tiedon analysoinnin ja uuden tiedon löytämisen. Tutkimuksen johtopäätöksenä voidaan todeta, että koneoppimista voidaan hyödyntää strategisessa avointen lähteiden kybertiedustelussa.

Avainsanat: kyberturvallisuus, tiedustelu, koneoppiminen

FIGURES

TABLES

TABLE OF CONTENTS

1 STRATEGIC OPEN SOURCE CYBER INTELLIGENCE, MACHINE LEARNING SOLUTION

The idea of the research was provided by the University of Jyvaskyla, Faculty Of Information Technology. Even though there are applications where Artificial Intelligence (AI) and its subtype Machine Learning (ML) is used to monitor cyberspace dataflows, what would be the new possibilities of AI and ML in the domain of cybersecurity?

In recent decades, cyber threat agents such as nation-state actors, cyber-terrorists, cybercriminals, and malicious individuals have created a significant threat to the economy and safety. From the end of 2018 to the first half of 2019, a new ransomware campaign appeared in a cyber domain. The cybercriminals demanded an enormous amount of money in bitcoins for large-scale organisations essential data that was encrypted by the attacker. The name of the phenomenon is Big Game Hunting.

Would it be possible to find signals from open-source data with ML to improve predictions and knowledge of such malicious cyber events? If the threats can be predicted, the interception of the danger becomes more efficient.

The ultimate goal of the research is to improve cybersecurity. In the study, by combining ML to intelligence processes information collection, procession, and analysing phases, we explore the suitability of ML to the Strategic Open Source Cyber Intelligence (SOSCI).

Three key findings were achieved in the study: First, the short sentences called passages that include valuable information produced by ML. Second, the ability to label the documents with metadata and third, the hidden knowledge that was revealed through the analysis of numerous reports.

In the next chapters, the research method, Design Science Research Process, is demonstrated. Then, the following essential concepts are presented: the basics of machine learning, SOSCI, cyberspace, the IBM Cloud platform, and the National Cyber Security Centre – Finland (NCSC-FI). Finally, the evaluation of potential use cases leads to the intelligence direction that enables the design science research cycle.

Best acknowledgments to The Support Foundation on the Finnish Air Force[1] for sponsoring the research.

1.1 Research questions

The goal of the study is to find out the answer to the main research question:

Can machine learning be utilised for SOSCI?

The three sub-questions support finding the answer to the main question:

1. Can machine learning be utilised in information collection?
2. Can machine learning be utilised in information procession?
3. Can machine learning be utilised in information analysis?

The research is restricted to information collection, procession, and the analysis phases of the SOSCI process. The intelligence direction that is presented in chapter 4.1 was created during the research to enable the intelligence cycle. The dissemination of intelligence products is not included in the study.

Cyberspace is presented only to classify the collected information about Big Game Hunting in the layers of cyberspace. The data, for example, the malicious program code in the logical network–layer, is not included in the research. The devices in the physical network layer or user accounts in the Cyber Persona Layer are not included in the study.

Two separate datasets are used in the research. NCSC-FI provided a dataset that was used in the first solution, and in the second solution, the information was collected from IBM Watson Discovery News. Both datasets are understood as open-source data.

The validity and the reliability of the answers to the intelligence questions are not evaluated in the research.

[1] Ilmavoimien tukisäätiö

2 RESEARCH METHOD

The Design Science Research Process (DSRP) is a set of analytical techniques and perspectives for Information Systems (IS) research. There are two activities in DSRP for improving knowledge in the IS domain: The generation of knowledge through the design of new and innovative things or processes and the analysis of things and processes via reflection and abduction. In the scope of DSRP, an example of things and processes means algorithms, human/computer interfaces, and system design methods or languages. A common term for things and processes in DSRP is an artifact (Vaishnavi, Kuechler, & Petter, 2004).

The key element for design science research process is the contribution of new knowledge. A five-step process forms new knowledge. The process is called the Design Science Research (DSR) cycle. The DSR cycle is repeated continuously until the results are satisfactory (Vaishnavi et al., 2004).. The DSR cycle presented is in figure 1.

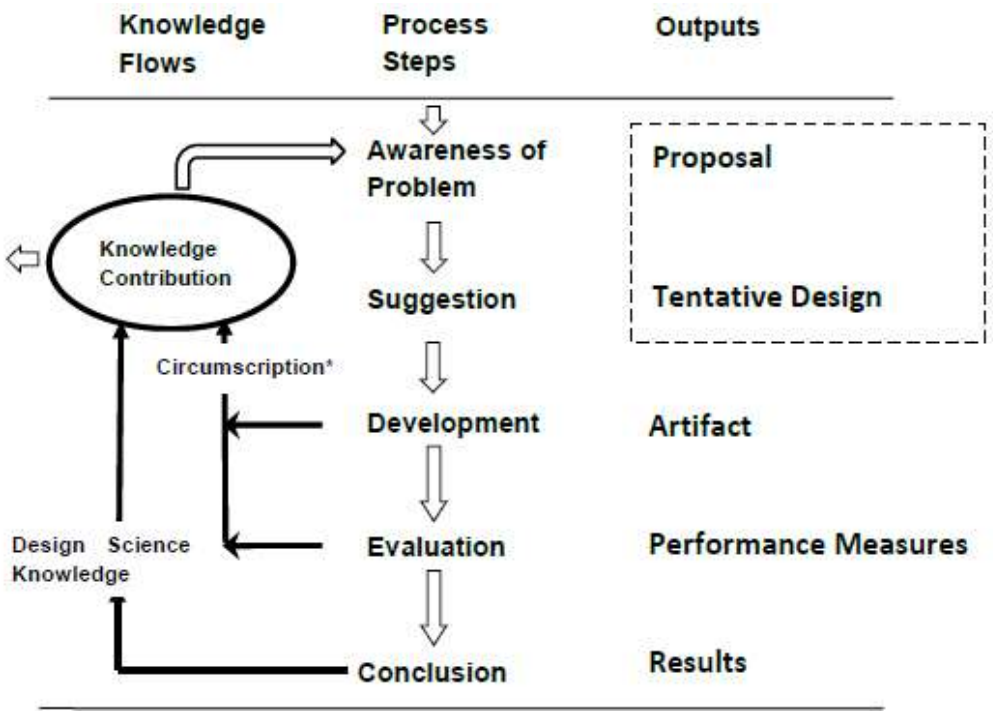

FIGURE 1 The DSR cycle (Vaishnavi et al., 2004)

The first phase is the Awareness of the Problem. Typically, in DSRP, the awareness of the problem comes from different sources, for example, from industry or information technology (Vaishnavi et al., 2004). The defined problem will be used to develop an artifact. It might be reasonable to split the problem into subparts so that the solution can answer the complexity of the problem (Peffers et al., 2006).

The second phase, Suggestion, is closely connected to the proposal and Awareness of problem. The output for the suggestion is the Tentative Design. Suggestion is a step where new and tentative solutions for the problem are innovated. According to Vaishnavi, the first and second output phases of the Design Science Cycle are closely connected, and for that reason, in figure 1, the outputs are surrounded in dotted line. The innovations might be new functionalities like in this case, the combination of the intelligence cycle and ML (Vaishnavi et al., 2004).

In the third phase, Development, the ideas, and innovations from Tentative Design will be further developed and planned in detail. The development depends on the artifact to be created. The important thing in the development is that the invention or novelty is the design of the artifact, not necessarily the construction of the artifact. (Vaishnavi et al., 2004).

The fourth phase is Evaluation. In the Evaluation phase, the function of the artifact should be measured and observed by how well the artifact supports the solution to the problem. The Evaluation depends on the artifact and the nature of the problem. One way to evaluate the artifact is the comparison of functionality with the solution objectives (Peffers et al., 2006). It is essential for gaining new information about the design and construction of the artifact. The obtained knowledge may lead to new suggestion and eventually, a new Design Science Research Cycle (Vaishnavi et al., 2004).

Depending on the literature source, there might be a phase between development and evaluation. In the article: The design science research process: A model for producing and presenting information systems research, the phase is called demonstration. In demonstration, the artifacts efficiency to solve the problem is proved(Peffers et al., 2006). Demonstration is not included in the study. Instead, demonstration is included in the Development phase.

The final and fifth phase of the DSR cycle is the Conclusion and the Results. If the results compared for the previous phase's criteria are fulfilled, the DSR cycle ends. The conclusions should be reported continuously. The findings might be facts that repeat continuously, or in some cases, it may not be possible to find such results. If there are no proper facts, it might be a subject for new research. At the latest in the final phase, the decision of a new DSR cycle is made. If the results are satisfactory, the DSR cycle ends for writing the results (Vaishnavi et al., 2004).

Detailed description of the phases of the research is presented in table 1.

TABLE 1 The phases of the research

Process step	Output
Awareness of the problem: A requirement of a comprehensive analysis of cyber-space phenomenon called Big Game Hunting	**Proposal:** Research: Can ML support intelligence process?
Suggestion: Use the Watson algorithm in IBM Cloud	**Tentative Design:** Combine Intelligence cycle and Watson. Could ML be utilised in information collection, processing and analysing?
Development: How to take advantages of Watson's cognitive capabilities? What are the requirements for AI training data? Which are the correct entities for the model and how to find them? Which Machine Learning type is used? Define data for queries. What are the relations between entities? How to find trends? What kind of visualization of the results would be the most useful? Possibilities to create a distinct database from the collected data? How to use Application Programming Interface (API) in the input and output of the data?	**Artifact:** A combination of ML and SOSCI.
Evaluation: Gaining new information and consideration of a new suggestion and DSR cycle.	**Performance Measures:** Comparison to the pre-created specific criterion of success: • Can the artifact find related information from the data? • Can the artifact analyse and process the collected information? • Can the artifact provide enough information to the intelligence direction and its sub-questions?
Conclusion: State the suitability of Watson for creating intelligence reports.	**Results:** Success: Implementation of IBM Watson for creating a strategic level of cyber reports. The estimate for further research. If the results are not sufficient, start a new DSR cycle.

3 LITERATURE REVIEW AND ESSENTIAL DEFINITIONS

3.1 Machine Learning

Kulkarni compares ML in his book: Reinforcement and systemic machine learning for decision making, for the human learning process. Learning is a holistic process, and in almost every case, it is somehow related to decision making. The results of learning come from processing data by sorting, storing, classifying and mapping (Kulkarni, 2012).

There are three ways of learning: First, learning happens inputs from more experienced persons such as professor at the University. The first way of learning can be understood as supervised learning. In the second case, learning forms from personal experience. Third, peoples learn from disruption based on experiences. The same principles apply in ML.

Kulkarni introduces three ML subtypes. Supervised, unsupervised, and semi-supervised. Depending on the literature sources, the names and classification for subtypes vary (Kulkarni, 2012). For example, in Jyvaskyla university's material: Basics and Applications of Artificial Intelligence[2], a sub-type called reinforcement learning is introduced (Lehto et al., 2019). Reinforcement learning is not included in the study.

3.1.1 Supervised learning

The objective for supervised learning is that the algorithm can divide data, for example, documents, into the correct subsets. In supervised learning, learning takes place by classifying data. The learner learns based on the available documents and the labels. The label is referred to as a class (Kulkarni, 2012). As Alpaydin mentions in the book: Machine Learning: The new AI, the supervisor decides the correct output for a given input (Alpaydin, 2016).

IBM Watson Discovery uses supervised learning for training the Watson algorithm. In supervised learning, the training data that is fed into the algorithm usually includes the correct results, labels. The labels are decided and provided by human or another algorithm(IBM, 2018a).

In figure 2, there is an example of data classification. The different materials are in class A or class B. The classifier is a program that tries to assign the content in the correct group. During supervised learning, the line between the classes is calculated. If there is an unknown document, the classification depends on the

[2] Tekoalyn perusteita ja sovelluksia

distance to the separator line (Kulkarni, 2012). It is worth to note that supervised learning is a process for finding an appropriate result, instead of only data classification.

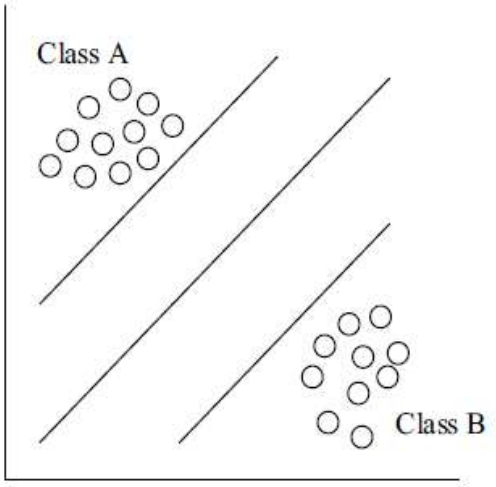

FIGURE 2 The data classification in supervised learning (Kulkarni, 2012)

3.1.2 Unsupervised learning

In unsupervised learning, the system tries to find and recognise similarities and data patterns without any external teaching. (IBM, 2018a). Unsupervised learning is based on mathematically calculated similarities and differences. As a result, in unsupervised learning, the data is clustered in particular classes as in figure 3. Usually, in unsupervised learning, the algorithms create hierarchical structures to arrange the objects (Kulkarni, 2012).

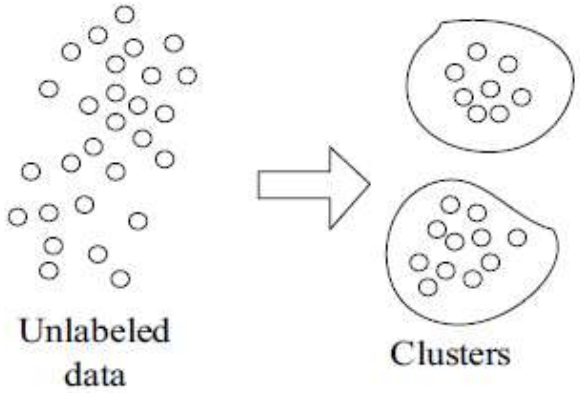

FIGURE 3 The data classification in unsupervised learning (Kulkarni, 2012)

3.1.3 Semi-supervised learning

There are characteristics of supervised and unsupervised learning in semi-supervised learning. Part of the teaching-data is labeled but not all(IBM, 2018a). The goal for semi-supervised learning is to find the best of both characteristics of the before mentioned paradigms (Kulkarni, 2012).

3.2 Intelligence

According to Watson, in military science, intelligence means information about an enemy or an enemy area. In the modern world, nations intelligence agencies have their collection and processing systems that provide rapid and accurate raw information refinement to knowledge (Watson, 1998).

The intelligence disciplines that are recognised within the United States Intelligence community are Open Source Intelligence (OSINT), Human Intelligence (HUMINT), Signals Intelligence (SIGINT), Geospatial Intelligence (GEOINT) and Measurement and Signature Intelligence (MASINT) (Clark & Lowenthal, 2016).

Each source produces an enormous amount of data and would presumably be suitable for processing with ML. HUMINT, SIGINT, GEOINT, and MASINT use such techniques that are unavailable for the needs of the university research.

The data for this research comes from open sources. The primary source for the data being processed with ML in the study is the Internet. In other cases, sources might be any available information for the general public such as television, newspapers, journals, and radio (Goldman, 2011). For that reason, the selected intelligence discipline for the research is OSINT.

3.2.1 Strategic intelligence

Strategic intelligence produces knowledge and estimates from the future (Joint Chiefs Of Staff, 2013). In the scope of the research, the intention is to combine IBM Watson's capabilities to the strategic intelligence cycle.

The idea of combining AI to Strategic Intelligence is not new. It is described in Liebowitz's book: Strategic Intelligence: Business Intelligence, Competitive Intelligence, and Knowledge Management. According to Liebowitz, AI techniques could be used in Strategic Intelligence, and it could enhance knowledge management. Strategic intelligence provides valuable information towards making strategic decisions in the organizations (Liebowitz, 2006).

In Don McDowell's book Strategic Intelligence, the strategic intelligence process is similar to the basic intelligence cycle and concept. Strategic intelligence provides information for executive-level clients. The information that is refined during strategic intelligence provides information concerning the purpose, construction and nature of the investigated phenomenon so that the client's

organisation can develop strategies on how to deal with it in the long term (McDowell, 2009).

3.2.2 Open Source Intelligence

OSINT is intelligence produced from publicly available information that is processed promptly to answer to a specific intelligence requirement (Bazzell, 2018). In Clark and Lowenthal's book (2016), the definition of OSINT is almost similar, but they bind legal issues to the description. OSINT should be done by lawful means. Any activity that requires theft, hacking, or overriding individual rights does not belong in the framework of OSINT (Clark & Lowenthal, 2016).

An essential requirement for OSINT is a proper source of criticism of the investigated data. The information collected from open sources should be carefully vetted and evaluated. The individuals and organisations who are targets for OSINT might provide disinformation and misinformation through their information-sharing channels (Clark & Lowenthal, 2016).

Finally, OSINT should satisfy the information required from the direction of the customer. The target of the OSINT depends on the case, but generally, OSINT works against individuals, organisations, technologies, locations, or governments. OSINT is an excellent tool for providing background information about the investigated target, and it might reveal the existing atmosphere. OSINT is a suitable intelligence discipline for providing early warning signals on incoming events (Clark & Lowenthal, 2016).

3.2.3 Intelligence cycle

The intelligence cycle is a five-step or six-step process, where during the process, the raw data changes to complete intelligence information (George & Bruce, 2008). Once data has been collected, processed, analysed, and assessed in the final phase, it is disseminated to the client. In the feedback phase, the client returns the feedback to the intelligence organization. The feedback might include new intelligence direction (Goldman, 2011). The intelligence cycle is presented in figure 4.

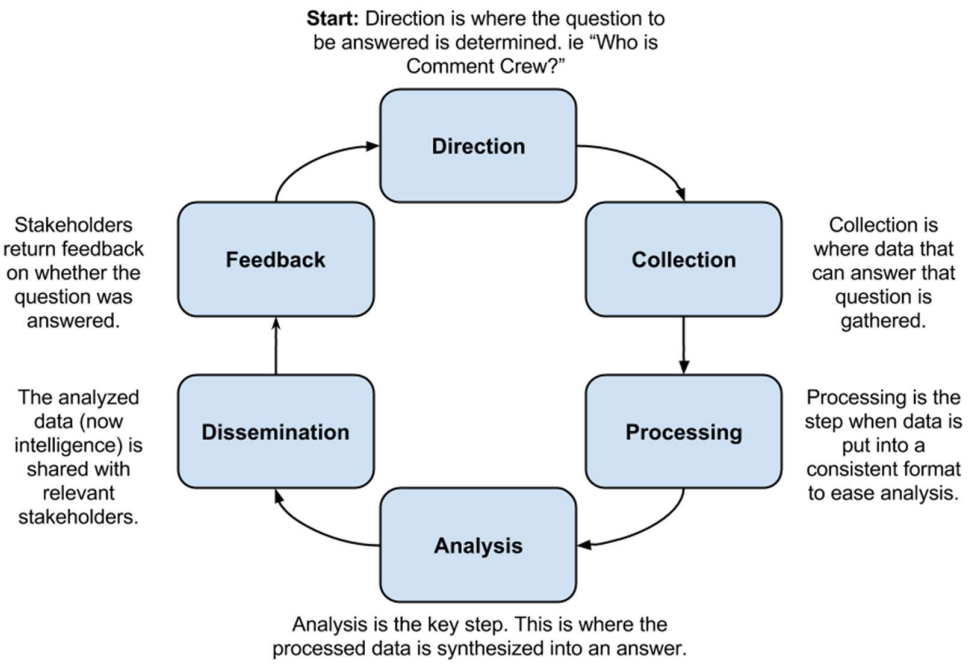

FIGURE 4 The Intelligence cycle (Roberts, 2015)

The Intelligence cycle begins from the direction (figure 4), the needs of intelligence consumers, or the intelligence client. The client might be a policymaker, military official, or another decision-maker who needs intelligence information for conducting their tasks or responsibilities (Goldman, 2011).

The definition of the collection in Goldman's book Words of Intelligence: An Intelligence Professional's Lexicon for Domestic and Foreign Threats, is:

> The obtaining of information or intelligence information in any manner, including direct observations, liaison with official agencies, or solicitation from official, unofficial, or public sources, or quantitative data from the test or operation of foreign systems (Goldman, 2011, p.60).

For strategic intelligence, the information collection aims to a deepen the understanding of the phenomenon and its large-scale impacts in the near and far future. The data collection should be comprehensive from all possible sources, because in strategic intelligence the goal is to build a deep understanding of the phenomenon, make forecasts of effects in future and give options for stakeholders and executives. The nature collected and analyzed information is qualitative, anecdotal and even impressionistic (McDowell, 2009).

The OSINT collection should begin with the goal of the intelligence task in mind. Also, attention should be targeted to the following questions: What exactly is the question of that is trying to be answered? What are the critical elements of the question? What are the best sources? How is the information collected and what is the required time for the whole information cycle? (Clark & Lowenthal, 2016).

Due to the nature of the collected data, the volume might be significant, and the measurement is difficult in traditional ways (McDowell, 2009). As McDowell notices, for strategic intelligence, the requirements for the data are complicated.

The analysis in the intelligence area means a systematic approach to problem-solving. First, the data is dived into distinct elements and examined to find essential parameters (Goldman, 2011).

Due to the complexity and the structure of the data, strategic intelligence analysis planning should be planned carefully. The understanding of the meaning of the investigated phenomenon from qualitative data there is no statistical reliability or reliance might require innovative procedures (McDowell, 2009).

Dissemination is the release of the information in the defined protocol; the distribution of intelligence products might be oral, written, or graphics in a suitable format (Goldman, 2011).

3.2.4 Strategic Open Source Cyber Intelligence

In this research, the definition of Strategic Open Source Cyber Intelligence (SOSCI) is derived from strategic intelligence, open-source intelligence and cyberspace.

SOSCI provides analysed information from open sources to organisations' executive-level decision-makers and stakeholders about the threats in cyberspace in the near-far future and long term. The information concerns threat-actors, their capabilities and motivations.

3.3 Cyberspace

The definition of cyberspace is not simple, and many definitions depend on the point of view of what cyberspace is. Generally, cyberspace can be understood as a collection of devices that are connected via a network. The information is stored, collected, and utilised with computational power. The purpose of cyberspace is to process, manipulate and exploit information. People interact with information. It is essential to note that both people and information are in a vital role of cyberspace (Rantapelkonen & Salminen, 2013). According to JP 3-12, Cyberspace Operations, the definition of cyberspace is:

> A global domain within the information environment consisting of the interdependent networks of information technology infrastructures and resident data, including the Internet, telecommunications networks, computer systems, and embedded processors and controllers (Joint Chief Of Staff, 2018).

In JP 3-12 cyberspace is described in three interrelated layers. The purpose of the model is to assist in planning and operations in cyberspace (Joint Chief Of Staff,

2018). The three-level model is suitable for defining targets for intelligence in the scope of this research.

3.3.1 The Physical Network Layer

The Physical Network Layer consists of the Information Technology (IT) devices, such as computers, network routers, and data servers. In the Physical Network Layer, the data is stored, transported, and processed. The layer includes hardware and infrastructure. An entity, public or private own every physical component of cyberspace (Joint Chief Of Staff, 2018).

3.3.2 The Logical Network Layer

The elements of the Logical Network Layer consist of network that is related to the physical layer and based on the code that drives and is used by the physical components. The individual links and nodes are represented in the Logical Network Layers as well as data, applications, and network processes. The elements of the Logical Network Layer can be targeted only by cyberspace capabilities (Joint Chief Of Staff, 2018).

3.3.3 The Cyber-Persona Layer

The Cyber-Persona Layer consists of network user accounts. The accounts might be related to an actual person, company, entity, or they can be automated. The accounts are associated with each other, and they have relationships. The accounts include data related to the connected owner, and they have connected personal or organisational data such as e-mails, IP- addresses, web-pages, phone-numbers, Web forum logins, or passwords to different accounts. The unique cyber persona might have several users; for example, one malicious hacker group might use the same malware command alias. Vice versa, one individual or entity might have multiple cyber personas connected to many accounts around cyberspace. Because the cyber personas and their relationships can be complicated, it makes the intelligence collection and analysis in the Cyber-Persona Layer a challenging mission. Another issue that makes the understanding of the cyber persona layer challenging is the fact that the Cyber – Persona's virtual location is not necessarily connected to a geographical location (Joint Chief Of Staff, 2018).

3.4 IBM Cloud platform

IBM Cloud is a cloud computing service that offers Infrastructure as a service (IaaS) and Platform as a service (PaaS). In the Cloud, customers can access multiple services, including the Watson AI algorithm in various ways (Rouse, 2017).

In this chapter, the IBM Cloud properties that are used in the research are described.

3.4.1 IBM Watson Discovery

The IBM Watson Discovery service is a cognitive analytics engine that can search and find data patterns. It is a part of the IBM Cloud platform. With Watson Discovery, it is possible to train AI to understand documents of the specific domain and find the most relevant answers from the data (IBM, 2019a).

When data is uploaded to Discovery, the service adds cognitive metadata to the documents. There is a total of nine enrichment available (IBM, 2019b). In the research, the following were selected for further use:

- Entity extraction; returns items that are present in the data. Discovery can automatically recognise entities from data (IBM, 2019c). Another option is to use a custom model. The custom model is used in the first solution of DSR cycle, and it is explained in detail in chapter 5.1.
- Relation extraction; recognises when two entities are related and identifyies the relation type (IBM, 2019). Also, in relation extraction, there is an option to use a custom model. The custom model is used in the research, and it is explained in chapter 5.1.
- Keyword extraction; important topics that exist in the data. Discovery automatically identifies keywords (IBM, 2019c).
- Category classification; categorises input data into a hierarchical taxonomy to five levels. The property allows more accurate classification of the data (IBM, 2019c).
- Concept tagging; Identifies concepts how the input text is associated based on other entities and relations that are present in the text. Property enables a better level of analysis than basic keyword identification (IBM, 2019c).

3.4.2 IBM Watson Knowledge Studio

The IBM Watson Knowledge Studio is an application in the IBM Cloud, where the custom ML model is created. The benefit of a custom ML model is that it is specially designed for the required purpose. After the ML model is ready, it is moved to Watson Discovery, where the model searches the data for the defined task (IBM, 2016).

3.5 The National Cyber Security Centre

NCSC – FI is part of the Finnish Communications Regulatory Authority (FICORA). The primary task of NCSC – FI is the creation, maintenance and dissemination of the cybersecurity situation picture. Other duties include maintaining

the cyber risk threat assessment with the co-operation with different administrative instances and actors.

Furthermore, NCSC – FI supports other authorities and private sector actors in the management of widespread cyber incidents. NCSC – FI collects and analyses relevant information to fulfill the information requirements of different actors. The analysis of risk assessment is created with international partners, and it produces forecasts of the consequences of the cyber threats to Finland (Secretariat of the Security Committee, 2013).

4 INITIATION OF DSR CYCLE, DEFINING THE PROBLEM

Four options for the use case of the ML raised from the discussion with NCSC – FI. The memo from the meeting is in appendix 1. Initially, the alternatives were:

- Cyber Weather report or part of its subchapters.
- In-depth analysis of the Big Game hunting phenomenon. Analysis based on an exact intelligence question with "5WH[3]." Trying to obtain the trends of the Big Game Hunting, the development of the phenomenon and the correlation with the event with time.
- Try to find trends from the source data. When signals from the event exist more often, it might predict the incoming cyber campaign. Trends might reinforce analysts' observations of the rising cyber event.
- Keyword extraction from a specific article group. Creating a database for keywords considering the investigated cyber issue

The first alternative, Cyber Weather report or part of its sub-chapters would have been suitable, but when compared with the available resources and the scope of the research, the full report would have been too large of an entirety. In turn, a part of the Cyber Report would have been a suitable use case.

The third option, finding trends, is an exciting and useful use case. The product from this option is: warning from the incoming cyber event. The challenge is that finding a trend would have required more time than available during the research process. Another reason why assumption this option was not chosen was the lack of ability to follow source the data in real-time. The available resources provided for university student does not include the visualization of data. It would have been an essential feature in the finding trend use case.

The keyword extraction exists as a property in the Watson Discovery service. The service can recognise keywords from the user's data. This option would have been too shallow for the research and considering the research questions; it would have been challenging to demonstrate the benefits of ML.

The selected option for the research is the analysis in depth of the Big Game hunting phenomenon. Big Game hunting is a relatively new phenomenon in cyberspace (Infradata, 2019). Also, this option enables the possibility to create a machine learning model that is designed precisely for the investigated issue.

[3] Who, What, Where, When, Why and How

4.1 The intelligence direction

The strategic level intelligence direction that is used as an example in the research is:

Provide information about Big Game Hunting:

- Who are the adversaries in Big Game Hunting?
- What are the target organisations?
- Where does Big Game Hunting occur in cyberspace and geographically?
- When does the attack occur?
- Why do adversaries select a particular organisation?
- How has Big Game Hunting changed during the first half of 2019?

It should be noted that the intelligence cycle direction for the research is imaginary; its purpose is to enable the intelligence process and intelligence cycles different phases. Also, the validity and the reliability of the answer to the intelligence question is not essential in the framework of the research and not evaluated.

Nevertheless, the intelligence direction is realistic; stakeholders or executives in the cyber domain might require information since, according to Infradata, the phenomenon has continued to rise in cyberspace (Infradata, 2019). Furthermore, in the first meeting with NCSC -FI, they mention the possibility to investigate the trend of Big Game Hunting. Also, the discussions raised the issue that during similar new large scales cyber events such as NotPetya and WannaCry, the information about the phenomenon at the beginning of the campaign is confusing and it is difficult to obtain correct information about the event.

5 THE FIRST SOLUTION

5.1 Building the ML model

The ML model building process follows the instructions provided by IBM. The model is built in the Watson Knowledge Studio. IBM recommended that there should be more than one person to participate in the development of the machine learning model (IBM, 2018b). The IBM cloud resources that are provided to university students, an academic license, limit the number of persons to one, so the whole process was implemented by the researcher.

There are two types of models available in the IBM Knowledge Studio. The ML model uses a statistical approach to find entities and relationships from the data, and it can learn and adapt when the amount of the data grows. Another option is a rule-based model that is more predictable and easier to maintain, but it cannot learn from the new data, and it only finds patterns that it has been taught to find (IBM, 2018c).

The selected alternative for the research is the ML model because during the time the investigated phenomenon Big Game – Hunting might change, and there will be a possibility to add new data about the phenomenon when the collected information enables the insertion of new documents to the model.

5.1.1 Type system

The type system requires a collection of entities. In the type system, entities describe how things are categorised in the real world. The *role*s define the context where the mention occurs, and as in the real world, entities *relate* to each other (IBM, 2017). For example, in the framework of the research, entity *CYBER_THREAT* arises from *CYBERSPACE,* and the relation between *RANSOMWARE* and *CYBER_THREAT* is *instrumentOf.*

The goal of the ML model is to provide information for a strategic intelligence direction that is related to cyberspace and Big Game Hunting. The selected approach in the study concentrates on the intelligence direction since the goal is to find essential information about the investigated phenomenon.

The type-system building began with finding suitable entities, and at the same time, the proper documents for training were initially observed. First, cybersecurity-related entities were selected from the Vocabulary of Cyber Security (Sanastokeskus TSK ry, 2018). The approach was that all entities that might have something familiar with intelligence direction were selected.

When each entity from the Vocabulary of Cyber Security was selected, the observation was that there were initially too many and too specific objects, and

the type system could be simplified. The type system was modified repeatedly until it was reasonable in the framework of the research and the intelligence direction.

Even though the intelligence direction is strategic level; it was observed that part of the entities were tactical level, because during the document selection, that is described in detail in the chapter 5.1.2 it came clear that there are not enough strategic level documents from Big Game Hunting for annotation tasks. The reason for that might be the fact that Big Game Hunting is a relatively new phenomenon in cyberspace. Also, only the strategic level entities are not enough to describe the Big Game Hunting. On the other hand, the assumption was that extending entities and documents to the operational level might provide more correct documents from the data. In the final version, there were ten entity types; the detailed descriptions of the entities are in table 2.

TABLE 2 Description of entities

Entity name	Description
TARGET_ORGANIZATION	The target for Big Game hunting, no names included, organisations that have been a target for ransomware attacks: police, government agencies, and similar organisations. Words that relate to targets, such as "victim" and affected users.
RANSOM	Ransom types that Target organization pays: cryptocurrency, bitcoins, money
THERAT_AGENTS	Nation-state, criminal, attacker, Names on malicious groups included, e.g., Fancy Bear
RANSOMWARE	Ransomware programs: NotPetya, GrandCrab, WannaCry.
CYBER_THREAT	Threat-related words: ransomware, attack, Big Game Hunting
ATTACK_VECTOR	Known attack vectors for ransomware: phishing, RDP, spam, EDP
VULNERABILITY	SBM, Eternal Blue
LOGICAL_NETWORK_LAYER	According to Cyberspace operations: files, network, data
PHYSICAL_NETWORK_LAYER	According to Cyberspace operations: physical devices.
CYBER_PERSONA_LAYER	According to Cyberspace operations: user accounts.

The final task in the type system creation is adding relations between entities. According to IBM instructions, relation type defines the binary and ordered relationship between two entities (IBM, 2017). Initially, the relations were created as entities relating to each other. For example, the relation between RANSOMWARE and ATTACK_VECTOR was *usesForAttack*. When the model

was evaluated, it did not reach any relation score. For that reason, the relations were changed similar to how Watson recognises by using relation extraction property described in chapter 3.4.1. In figure 5 is the used type system with entities, relationships, and roles. The roles are marked with dissimilar boxes: cyberspace with dotted box, a threat with a white box and target with a gray box.

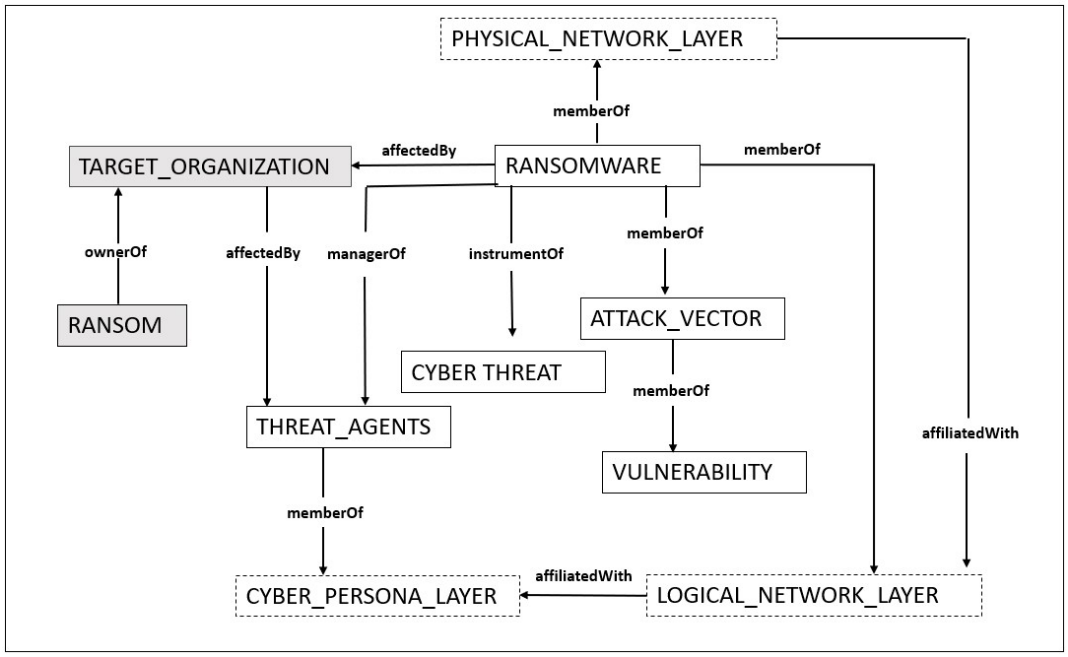

FIGURE 5 Type-system

5.1.2 Initial documents for model

The data for the creation of the ML model was selected from documents concerning cyberspace threats, ransomware, and Big Game Hunting. The primary document is a European Network and Information Security Agency (ENISA) Threat Landscape Report that offers a comprehensive perspective of the area of interest. According to IBM recommendations, the length of the document should be between 1.000 and 2.000 words. The maximum length for the documents is 40.000 words. The source documents were more extensive than the recommended length. For that reason, the original documents were separated in suitable length chapters that fit in the type system. The following documents were used to build the ML model:

- Threat Landscape Report 2018 (ENISA, 2018)
- WannaCry Ransomware Outburst (ENISA, 2017)
- Enterprise is the target of 'big game hunting' (Loeb, 2019)
- PINCHY SPIDER adopts "Big game hunting" to distribute GandCrab (Feeley, Hartley, & Frankoff, 2019)
- Global Cyber Threat Report (Infradata, 2019)

5.1.3 Annotation

Annotation is a task where the type system is connected to the documents. It is a task related to supervised learning that is described in chapter 3.1.1. A human annotator prepares the data for classification. During the annotation, the labels are created. With the created labels, the words are classified to the correct entity; in other words; the entity is a label.

Correct annotation requires a good understanding of the documents and the type system. There is a possibility to pre-annotate documents with Watson Natural Language Understanding (NLU) property. The service annotates the documents with a predefined set of entity types. The automatic annotation was used, but it provided on incorrect annotations since the annotated documents were specifically cybersecurity-related. The automatic annotation can be used in standard documents.

In manual annotation, words are attached to a correct entity. For Example, in figure 6, the word **ransomware** is attached to entity THREAT and word **Ceber** to RANSOMWARE.

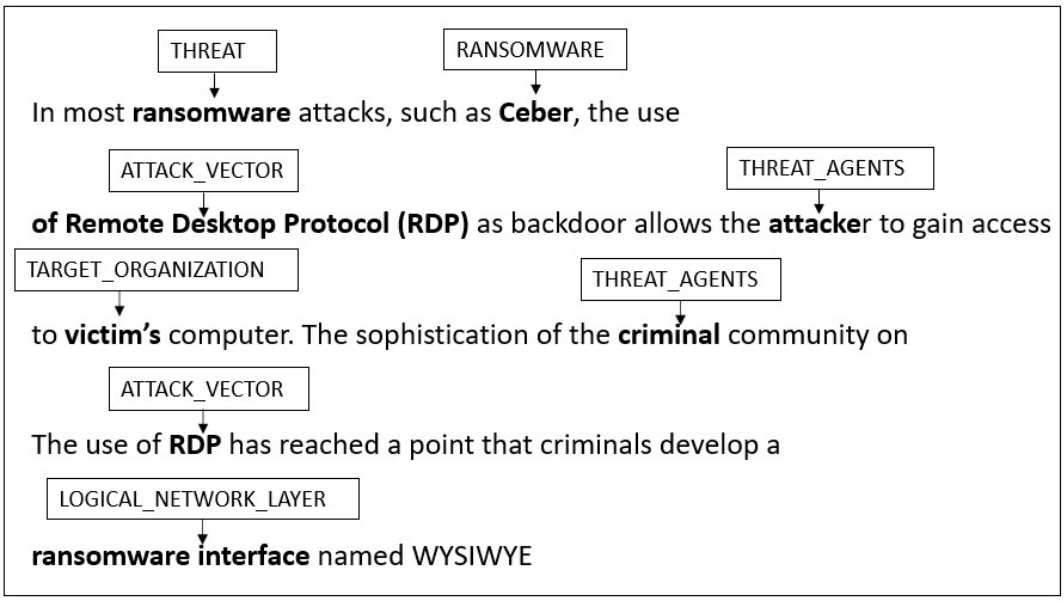

FIGURE 6 Annotation for words

When words were connected to the correct entity, the annotation continued by selecting the correct mention. The alternatives for mentions were: name, noun, pronoun, or none. Also, the class of the word was selected: specific, negative or general. The final task for the annotation process was applying the relations that were created during type system building. The annotation of relations is in figure 7.

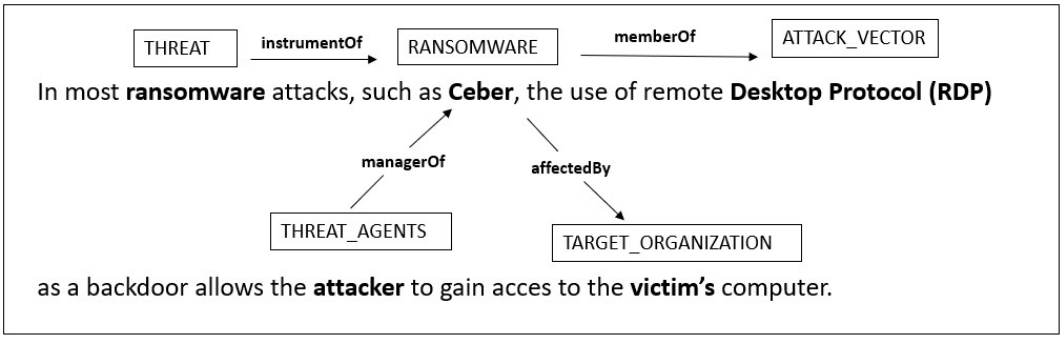

FIGURE 7 Annotation for relations

The observation during annotation was that it should be carefully considered which words and sentences belong to the correct entity. In the academic license version on IBM cloud, one person annotates each document. For paid versions, it is possible to share annotation tasks with several annotators. If many peoples annotate the documents, it improves the validity of the investigated results since it secures multiple times that the documents and entities correlate.

When the first documents were manually annotated, the model itself was used for the annotating task. The model uses both unsupervised and semi-supervised learning (chapter 3.1.2 and 3.1.3). It means, that the model recognises entities and words automatically from the new documents and creates the connections. After each automatic annotation, the entities and relations were manually checked by a human and corrected if needed.

In the beginning, the results were inaccurate. The model connected some words to an incorrect entity. In those cases, the annotations were adjusted manually. When the model was used again for annotation in version 1.1 and 1.2, the automatic annotation was accurate, and no corrections were needed. It means that the model was able to learn correct entities and related words.

5.1.4 Training and analyzing the model

When the manual annotation was completed, the ML model was trained until the results were satisfactory. Training means adding new data after the model automatically finds correlations by unsupervised learning from the documents. For the training, a part of the documents was separated as comparison dataset, called ground truth.

The evaluation is based on the model's ability to find entities and relations from the new data (IBM, 2018d). Each time a new document was added, the model was trained, and a new version of the model was created and evaluated. The first version 1.0 was based on documents mentioned in chapter 5.1.2. The versions of the model added materials and references are in table 3.

TABLE 3 Versions and documents

Version	Document name	Reference
1.1	Ransomware	(Microsoft, 2019)
1.2	Ransomware	(ENISA, 2019)
1.3	Threat Landscape Report 2018, chapter 4, Threat agents.	(ENISA, 2018)

The IBM Knowledge Studio provides an evaluation score for each version. Because the academic license limited the number of training, the model was trained to a sufficient performance range that is defined by IBM. Since the research questions do not deal with the model accuracy; the evaluation and training were conducted until the sufficient model accuracy was reached. In figure 8 are the versions and the development of the model. In version 1.3, the accuracy lowered. One reason for that might be that the added documents did not include enough proper words for entities and relations on the automatic annotation. The trained model did not reach any relation or conference scores due to research restrictions. The academic license restricts the number of training to 30 / month, and the available time for the research was limited. For those reasons, version 1.2 was selected.

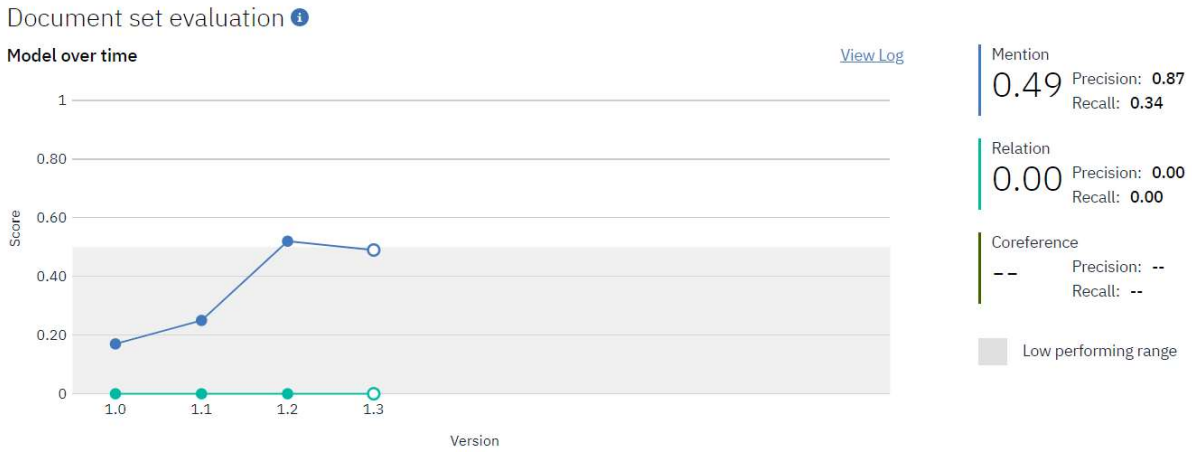

FIGURE 8 The evaluation scores

5.1.5 Using the machine learning model in Watson Discovery

The ML model that was created and evaluated in the Watson Knowledge Studio was deployed to Watson Discovery before the data for queries was uploaded. Unsupervised learning was used when the data was moved to the Watson Discovery. The algorithm recognized the similarities in the documents and created a variety of clusters.

The enrichments for the data that is described in chapter 3.4.1 was selected to the model before data was uploaded to Discovery service.

NCSC-FI provided 1.310 cybersecurity-related news documents from 1.1.2019 to 31.5.2019. During the upload, it was noticed that Discovery accepted only 840 documents due to academic license restrictions.

As mentioned in chapter 5.1.4 the model did not reach any relation score due to IBM cloud academic version limitations, but It was observed that during the data upload that Discovery automatically created relations between entities using relation enrichment property. When the relations were inspected in detail, it was observed that they were dissimilar compared to intended ones.

5.1.6 Training Watson Discovery for intelligence direction

Before the actual queries for intelligence direction were done, the Watson Discovery was trained once more to find information about Big Game Hunting according to the custom model. The training was completed with the NLU query. According to IBM guidelines, the query should be written in a way that the user would ask the question, and some term in the query should overlap between the query and desired answer (IBM, 2019d).

The intelligence direction of the study concerns finding information of Big Game Hunting campaign. The assumption was that the data did not include the straight correct answer for the questions, and essential information might be found in previous ransomware attacks. For that reason, the phrase for natural language query that was used for training Watson was: "Big Game Hunting and ransomware." The training itself took place by rating the correct documents from the data that NCSC-FI provided. The model was trained with 100 documents from the data. Fourteen documents were relevant, and 86 documents not relevant. Part of the documents did not include words at all about Big Game Hunting, but they included the context of malware and big ransoms. It requires the human ability to find the correct context from the documents that did not include related words or phrases.

In the figure 9 is the flowchart of supervised learning in the solution. It should be noted that even the chart is about unsupervised learning, in processing-phase, the Watson Discovery uses unsupervised learning when the new data is ingested to the system, Watson discovery recognises automatically predefined enrichments.

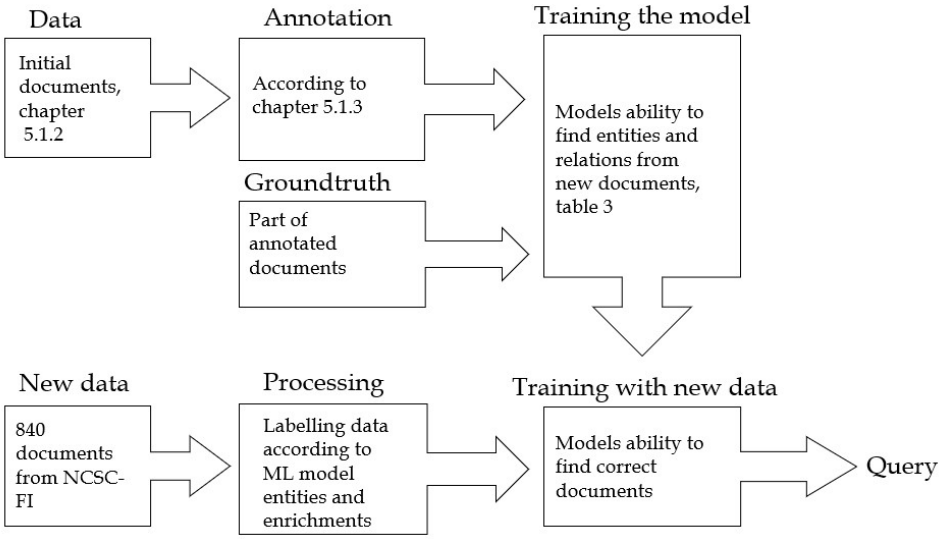

FIGURE 9 Flowchart of supervised learning in the solution (redrawn from Lehto et al., 2019)

5.2 Querying Data

When a query is created in Discovery, the engine observes each result and tries to match them with predefined paths. Results will be added to the result set. A query can be detailed or comprehensive, depending on the investigated issue. The more specific the query is, the more accurate the results are (IBM, 2019c).

There are two different query concepts in the Watson Discovery. A natural language query is an option, where the question is asked in a plain language such as "What is Big Game Hunting?" Another option is Discovery Query Language, where the query is written in the Discovery Query Language. It enables the ability to build more targeted queries. Also, it is possible to aggregate and filter the results and write nested queries (IBM, 2019c).

Query search parameters enable searching the data, identifying the correct results, and performing analysis on the result set (IBM, 2019e). There are multiple parameters for queries.

The first query was made with natural language with the sentence: "Big Game Hunting." The analysis was not included. Watson returned a total of 398 documents and five passages. According to IBM, passages are generated with sophisticated algorithms to determine the best paragraphs from all of the documents returned by the query. The passages are in figure 10:

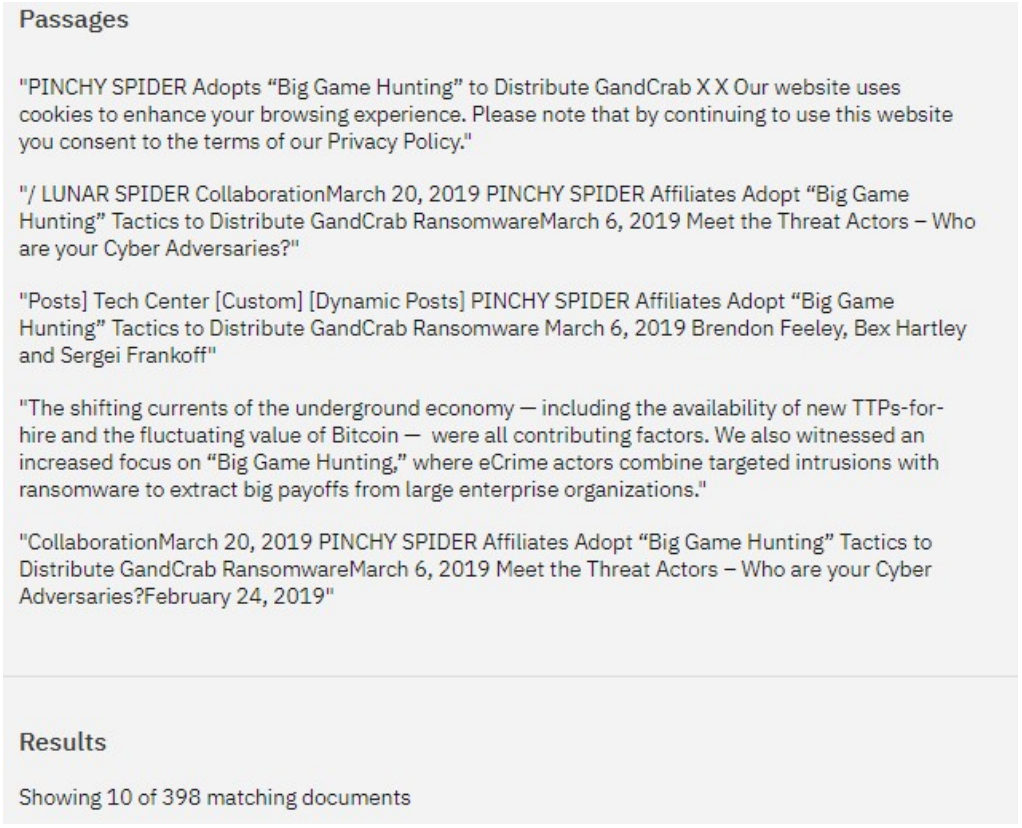

FIGURE 10 Passages for the first query

The returned passages include usable information about Big Game Hunting in the perspective of the intelligence direction about the adversary group called PINCHY SPIDER. Furthermore, the name of the ransomware: GandGrab was obtained, and the passages included some dates. The results provide information to the intelligence directions first and second sub-questions:

- Who are the adversaries: ECrime actor, Pinchy Spider
- What are the target organisations: Large enterprise organisations

It was noticed that the number of returned documents 398, was significant compared to the total amount of 840. The assumption was that there might be two probable reasons: either the queried data includes 398 documents of Big Game Hunting or the Watson includes the words “Big” “Game” and “Hunting” separately for the answers.

The second query was made with Natural Language Query with the words “Ransomware” and “Big Game Hunting.”. The passages that Discovery returned are in figure 11.

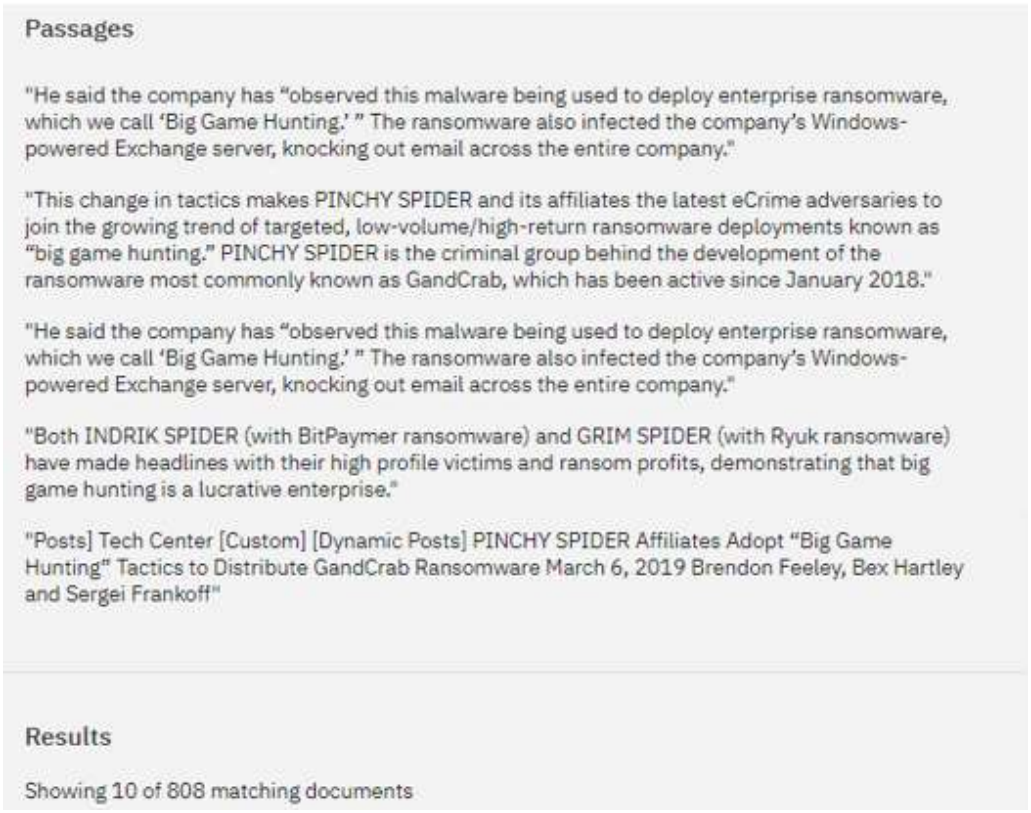

FIGURE 11 Passages for the second query

The new information was obtained from the second query concerned about malicious group INDRIK SPIDER and PINCHY SPIDER. Also, two new ransomware names were found: BitPaymer and Ruyk. As well, information about the Big Game Hunting operation was found: the Windows-powered Exchange server is a vulnerable operating system, and emails might be an attack vector. Again, the number of matching documents was substantial; at this point, a total of 808 from 840 documents was returned. The second query offered usable information about Big Game Hunting.

The third query was conducted with natural language query with the word "Big Game Hunting" and "Ransomware." The text was analyzed with the top values of enriched keywords and filtered by the entity type THREAT_AGENTS. The used Discovery Query Language code was: *term(enriched_text.entities.type,count:10).*
The exact definition for aggregation clause *term* according to IBM is:

> Returns the top values (by score and by frequency) for the selected enrichments. All enrichments are valid values. You can optionally use count to specify the number of terms to return. The count parameter has a default value of 10. This example returns the full text and enrichment of the top values with the concept enrichment and specifies to return ten terms (IBM, 2018e).

The query means in plain language: "*Find documents that include the words Ransomware and Big Game Hunting, arrange the words with the ML model entities, show top 10 matching entities.*"
The result for the query is in figure 12 below:

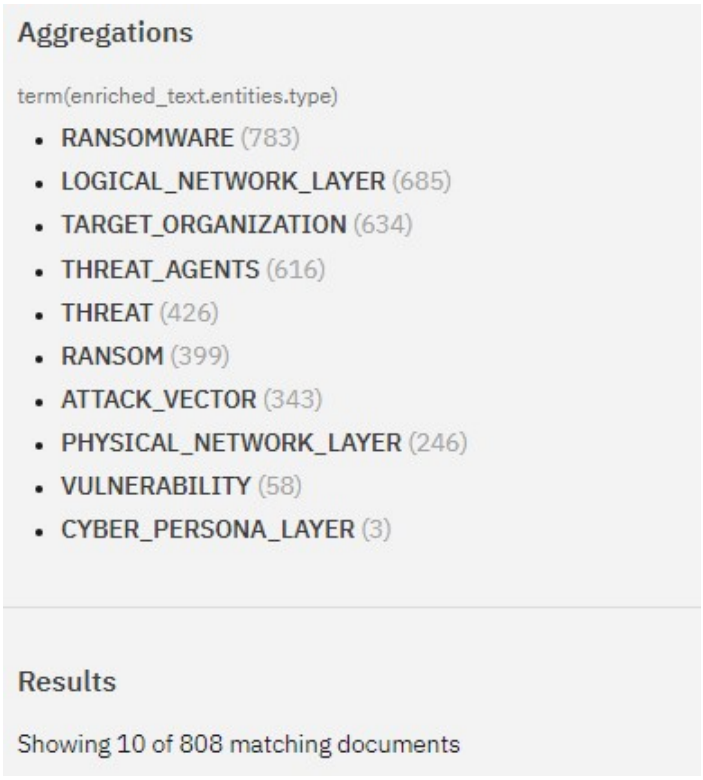

FIGURE 12 Aggregations classified according to type-system

The knowledge obtained from the query concerns about layers of cyberspace. Since the top values are in the logical network layer, it might help to understand the Ransomware phenomenon better. The query returned a total of 808 documents that match the defined query rules.
Considering cyberspace, the results provide information for the intelligence direction, the thirds question:

- Where does the Big Game Hunting occurs in cyberspace and geographically: Logical Network Layer

The fourth query was similar, but t Big Game Hunting and Ransomware were compared against the keywords (chapter 3.4.1) that Watson Discovery generated during data ingestion. The used Discovery Query Language code was: *term(enriched_text.keywords.text,count:10)*. The results are in figure 13.

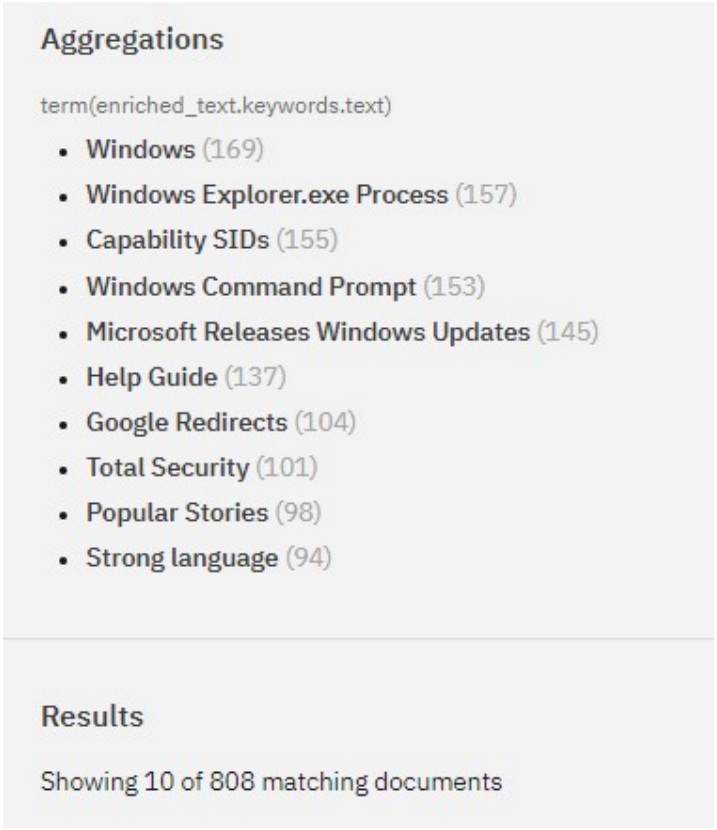

FIGURE 13 Aggregations classified with keywords

The new knowledge gained from the fourth query concerns Windows Explorer.exe and Capability SID. The query returned a total of 808 documents that match the defined query rules.
The same words were used in the fourth query, but the words "Ransomware" and "Big Game Hunting" were compared against enriched HTML categories' labels. The results are in figure 14.

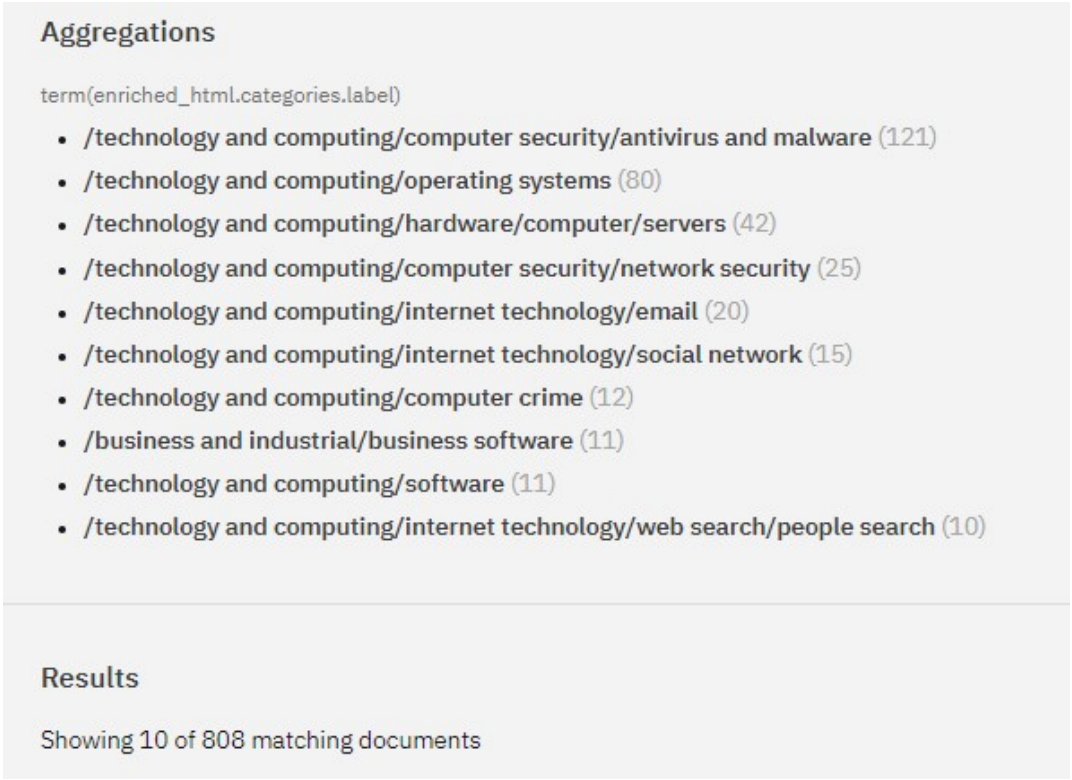

FIGURE 14 Aggregations classified by HTML categories

At this point, it was noticed that the number of matching documents is still relatively large even when the data was queried in various ways. The reason was found when observing the returned documents in detail from each previous queries. Even the model was trained to find Big Game Hunting words in context on the phenomenon it still obtained distinct words (figure 15).

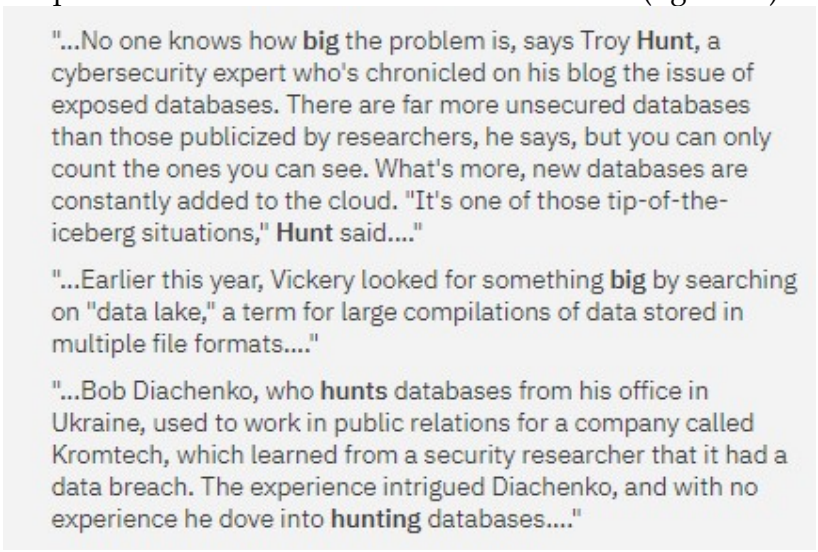

FIGURE 15 An irrelevant document

The final query from the data was made with Discovery Query Language with following query code: *enriched_text.entities.text::"Big Game Hunting"* In plain language The query means: *"Find from given text documents that include exactly the words "Big Game Hunting." arrange the results with custom model entities"* The *exactly* comes from the operator ::. Watson returned the following results (figure 16).

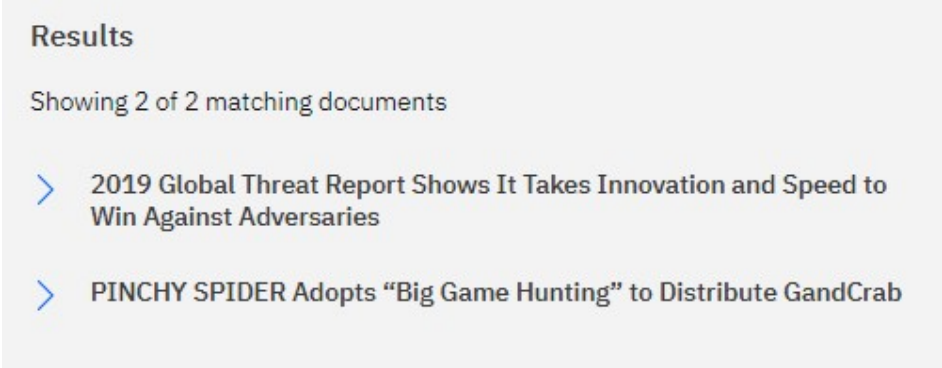

FIGURE 16 Two matching documents

The results confirmed the assumption about the ML model. If the data is queried without the exact operator::, the queries return all documents that include the words "Big" "Game" and "Hunting." Also, the data included only two documents that concern Big Game Hunting as cyberspace phenomenon.

5.3 Conclusion

The artifact that was created during the Design Science Cycle is: *A ML model that is planned according to intelligence direction to provide information about Big Game Hunting.*

5.3.1 Documents for ML training

The documents for cybersecurity and threat roles were easy to obtain because of the ENISA's Threat Landscape Report, which is a reliable and well-known document in the cybersecurity domain. Since Big Game Hunting is a relatively new phenomenon in cyberspace, the source documents were challenging to find. First, by using Internet search engines, there were some hits concerning the phenomenon. Eventually, the documents that were used for ML training concerning Big Game Hunting were created by combining contents form three cybersecurity-related web-pages. The reliability of the authors was challenging to evaluate, but the information on the web–pages was congruent, and it was cross-checked against each other. For those reasons, the documents were accepted as training material for the ML model.

During the document selection, the observation was that it is vital that a person who selects the documents for ML training needs to understand the domain deeply. In the research case, the document selection was conducted by the researcher. A better alternative would probably have been to use an expert from the NCSC – FI.

When the model was developed, new documents were added. The material was from ENISAs and Microsoft's web pages. Since Microsoft is a known actor in information technology, and the information was congruent with ENISA, the Microsoft document about ransomware was accepted for training. The concern about the reliability of Microsofts document was the commercial point of view. The document was evaluated, and there was no bias in that perspective.

Since the size of the documents for the creation of the ML model is limited 40.000 words, and the recommendation is 1.000-2.000, the original documents were separated into suitable lengths. Before the documents were ingested to Watson Knowledge studio, they were converted to the correct format that is UTF-8. The editing of the documents was conducted by copying pdf to Microsoft Word, where the number of words was counted, and then the suitable length chapter was moved to Notepad ++ after which the document was saved in UTF-8 format.

There is a possibility during the editing and conversion that words might disappear, and for that reason, the context of the document might alter. It was observed that in some cases, single words were missing. After the observation, each document was carefully checked for mistakes. Still, it cannot be guaranteed that all documents were undamaged. In future research, the possibility of automatic conversion should be considered.

The critical finding in the perspective of the research was that the selection of the documents requires careful information analysis. The connection to the ML model is visible. The more accurate the data for ML training is, the more precise the results the ML provides are.

5.3.2 Data for queries

NCSC-FI provided 1.301 documents that were a total of 50 megabytes of news data from various cybersecurity-related Internet sources. For that reason, the data is biased in the perspective of cybersecurity. The purpose of the model was to find indications and information about a phenomenon that concerned ransomware. When the ML model was used for the query of the data, then some was acquired. During the queries, it was not confirmed how the model would function with the data that includes other than cybersecurity-related information.

The academic license limited the amount of data to 1.000 documents and 200 megabytes of data. During the data ingestion, Discovery accepted only 840 documents. The reason was the size of the data. When the data is fed in, Discovery creates metadata that consumes the available disk space in the cloud service.

According to recommendations, the resources that are included in the academic license should only be used for testing the model. If the created model is

used further, there should be more available disk space for the data. Even the amount of the data was limited by the number of the documents and the size of the files; 840 documents was enough to test the model's ability to find data about the investigated phenomenon.

5.3.3 ML model

The approach to how the type-system was created proved to be complicated. The type-system that was created is a researcher's perspective of cyberspace and Big Game Hunting. Initially, there were multiple entities, and during the building, it was soon noticed that the amount of the entities needed to be decreased. One reason for the reduction was the difficulty to find enough useful documents for each entity. The reduction of entities was repeated until there were ten entities left. The reason for the approach where multiple entities were selected from the Vocabulary of Cybersecurity depended on the research knowledge about cyberspace. It was easier to add more entities and then take the useless ones away.

The intelligence direction guided the development of the type system during the process but afterward, the solution would have been easier to achieve if the intelligence direction and questions would have been more of a guideline to the type system.

An important observation during the development was the level of the entities. As the intelligence direction was strategic, should all entities be strategic as well? The reason why there are entities that are not strategic is the fact that it was difficult to find documents that concern the investigated phenomenon only on the strategic level. On the other hand, if the required information in the intelligence direction is strategic, the selection can be made after the model provides the answers.

The relations that are used in the type-system did not initially function at all. The IBM instructions were partly unclear how the relations function in the ML model. In the first versions of the type-systems, the relations were written as the entities relate in the real world. For example, in the early versions, the relation between CYBER_ATTACK and VULNERABILITY was *isDirectedTo*. When the model was initially tested it did not reach any relation score in the Watson Knowledge Studio. During the development, the relations were changed according to WATSON NLU relations. Even the relations were changed, the model did not reach any relation score. Finally, the academic license limited the number of tests so that it was accepted that there would not be a score for relations. The problem with relations was solved when the model was deployed to Watson Discovery. The Discovery was able to find the relations during the deployment automatically.

IBM recommends that the people who create the system need to be experts of the domain, and there should be multiple developers for the system. That is because then there would be a comprehensive and multi-sided view of the area of interest. The recommendation for annotation where words are connected to the entities is that multiple persons annotate the same documents because then

the correct connection with the words and documents is secured numerous times. Due to the limitations of an academic license, there was only one account available for the annotation. The researcher annotated all documents. Because of the restriction, there was not any cross-checking of the annotation. It might affect the model's ability to provide correct answers.

The layers of cyberspace were included because it would be possible to observe where in cyberspace the phenomenon occurs.

5.3.4 The queries

The queries provided results about Big Game Hunting. The passages that were introduced in the first and the second query were valuable in the perspective of the intelligence direction and the research; the Watsons cognitive capabilities appeared during the first and the second query. The algorithm was able to find the best paragraphs from the data.

In the third query, the words were tested with the entities that exist in the ML model. The model was able to classify the data with the entities.

In the fourth and the fifth queries, the words were queried against the keywords and the HTML labels that the algorithm created during the data ingestion.

Because the number of documents and the returned hits in the aggregations in the first five queries were plentiful, there was a suspicion that the model cannot return reasonable answers or the queries need to be modified.

The final two queries confirmed the assumption. The queries included separate words Big, Game, and Hunting. Furthermore, the queried data was insufficient to answer the intelligence direction because only two documents dealt with Big Game Hunting.

5.3.5 Evaluation of the artifact

1. Can the artifact find related information from the data?

Referred to the first and the second query, the artifact can find related information.

2. Can the artifact analyse the collected information?

Referred to the third, fourth, and fifth query, the collected information can be analysed with the model.

3. Can the artifact provide enough reasonable information to the intelligence direction and its sub-questions?

The number of returned documents in the fourth and fifth queries revealed that the artifact could not provide enough reasonable information to answer the intelligence direction's questions.

Due to the evaluation results, the new DSR cycle was initiated. According to Design science methodology, the gained knowledge from the solution is used in the next DSR cycle.

6 THE SECOND SOLUTION

The first and second process steps in the Design Science Research Cycle, Awareness of the problem, and Suggestion remain similar than in the first solution. The development of the artifact is modified with the knowledge gained in the first solution. The reason why performance measures failed in the first solution was too limited data for the intelligence direction. Furthermore, the ML model could not recognise Big Game Hunting as a cyberspace phenomenon. Instead, the queries returned documents that included separate words and no intended context. The use case in the second solution is identical compared to the first solution. Also, the intelligence direction stays unchangeable. The development of the artifact in the second solution is described in table 4.

TABLE 4 The development of the second solution

Development	Artifact
How to take advantages of the knowledge gained in the first solution? Can the ML model in the first solution be used as is? How to improve the data for queries? How to take advantages of Watson's cognitive capabilities? How to develop the queries? How to pay more attention to intelligence direction? Are there possibilities to expand the data used in the first solution?	An improved solution that passes the performance measures of the artifact.

6.1 Watson Discovery News

In the second solution, the data for queries is the Watson Discovery News, open pre–enriched and labeled dataset. The data is public, and the sources for the data set are in five languages. The major part of the documents is in English news sites that are updated continuously. The oldest information in Watson Discovery Data

is 60 days old. The data changes every day; IBM adds approximately 300.000 new articles from news and blogs daily (IBM, 2018a).

It was found out from the IBM documentation that the custom model created in the Watson Knowledge Studio cannot be used in the Watson Discovery News. Besides, Watson Discovery news cannot be trained, there is no option to add data, or the dataset cannot be configured for the specific use case. According to IBM, the proposed use cases for Watson Discovery News are news alerting, event detection, and trending topics in the news (IBM, 2019f).

Even though there is no possibility to use a custom model it seems that the proposed use cases might support the requirements of the research. The properties of data for the queries at 1.7.2019 are in figure 17.

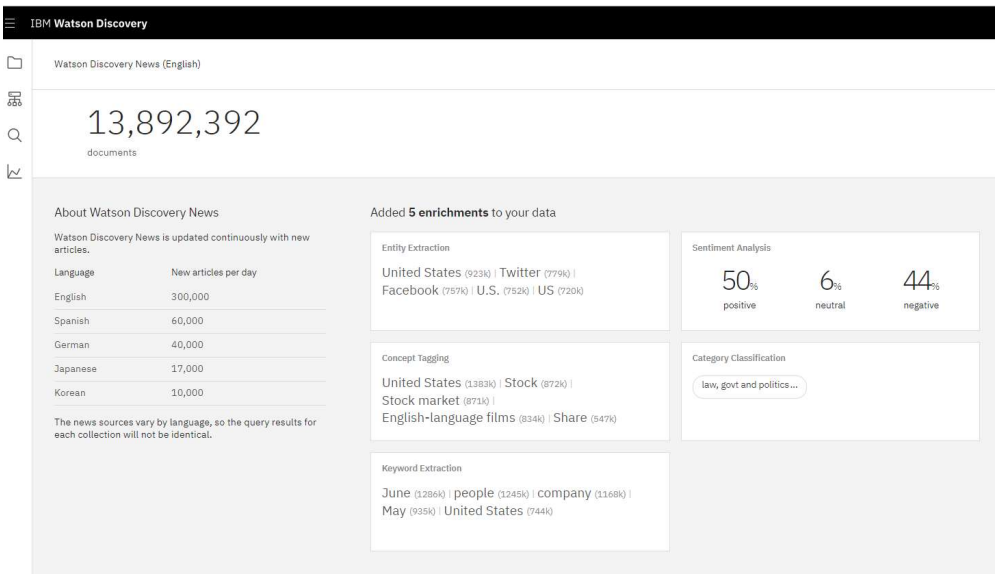

FIGURE 17 The features of Watson Discovery News

There are over 13.000.000 documents available for the queries. The total amount of English documents cannot be confirmed. Compared to the 840 documents in the first solution, the amount of the data is multiple. From the perspective of the intelligence direction, it should be remembered that the data in Watson Discovery News is not only cybersecurity-related. Instead, it includes various documents from various domains. Referred to figure 17, when the suitability of the data was initially investigated, it was noticed that the amount of the documents changed several times during an hour. It means the data is updated continuously. In figure 18 is the flowchart of unsupervised learning. The ML takes place in the processing phase, where the data is labelled with selected enrichments.

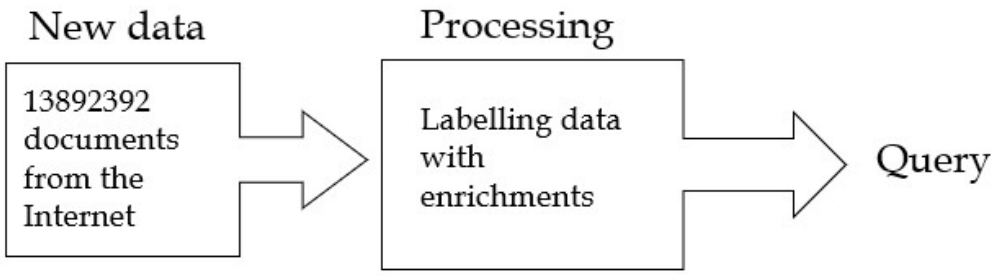

FIGURE 18 Flowchart of unsupervised learning in the solution (redrawn from Lehto et al., 2019)

6.2 Querying data

The first query was similar than in the first solution. The used query type was the natural language query with the words "Big Game Hunting." Watson returned 3.047.080 documents. The same issue exists than in the first solution. In figure 19, the results included documents that are not suitable in the context of the intelligence question.

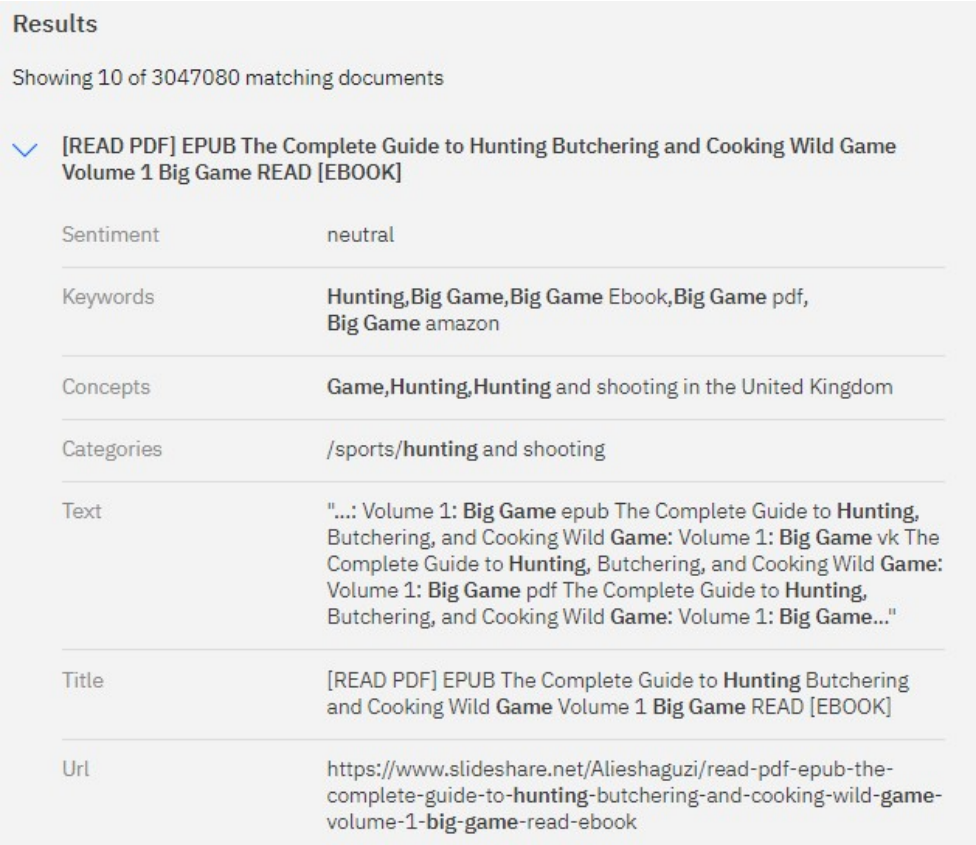

FIGURE 19 A document that includes irrelevant information

The second query was similar than in the first solution. The used words for natural language query were "Big Game Hunting" and "Ransomware." Watson returned 3.055.450 documents. It can be noticed from the figure 20 below that each top ten documents included valuable information about the Big Game Hunting Ransomware campaign.

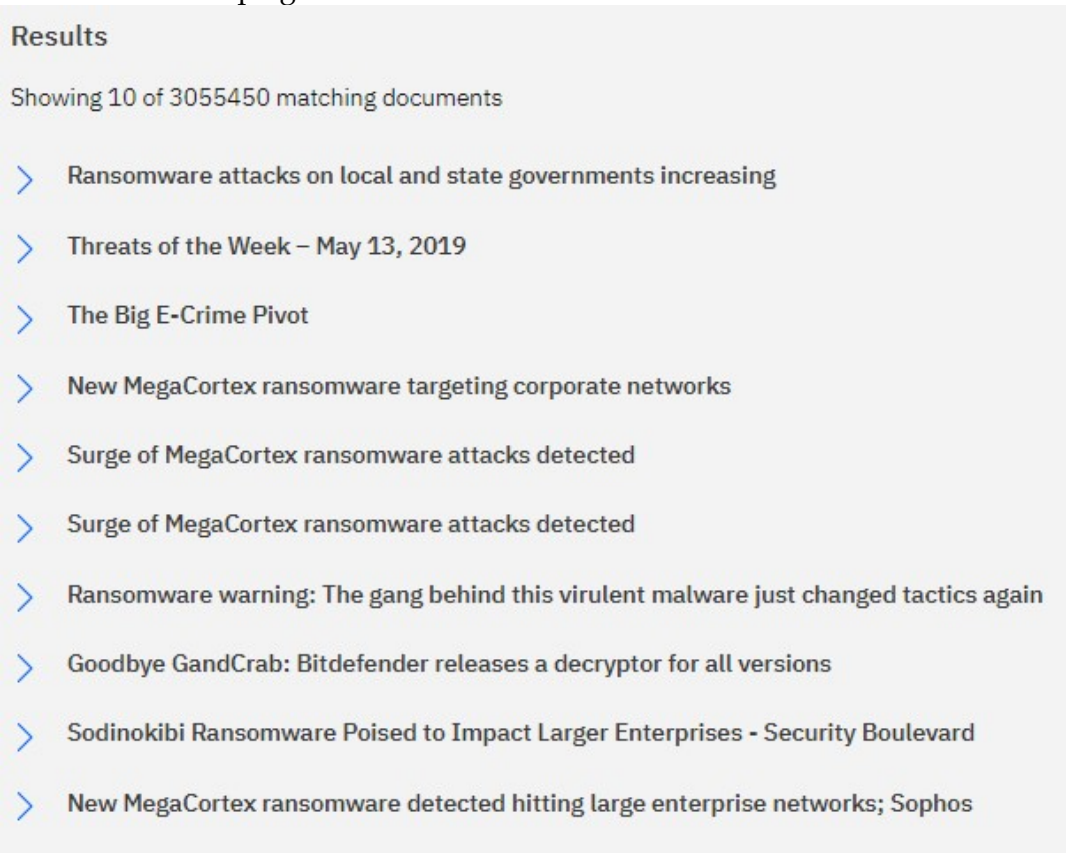

FIGURE 20 Relevant results

The second query concludes that the correct keywords for querying the phenomenon have found. When the first obtained document was investigated in detail, it was found that the result also includes the Uniform Resource Locator (URL) of the news (figure 21).

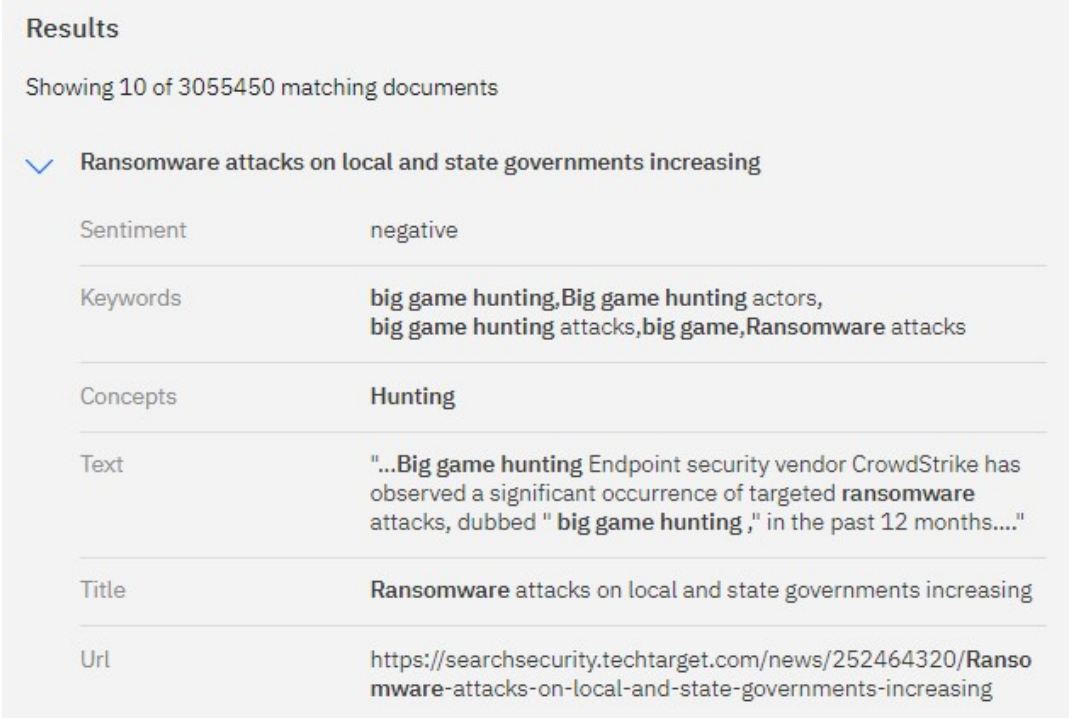

FIGURE 21 Document with URL

The third document (figure 22) from the second query, provided information that relates to the intelligence direction's fifth question:

- Why adversaries select a particular organization: For having high dollar ransoms from manufacturing and industrial companies.

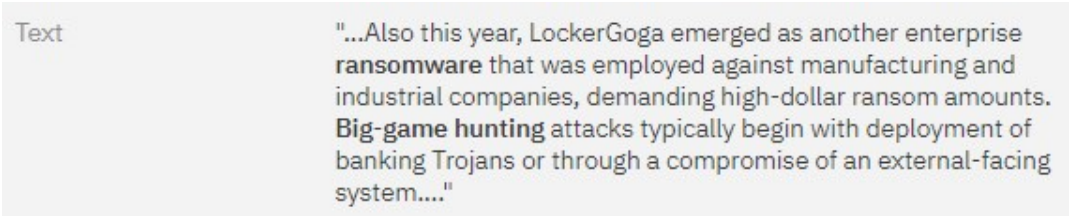

FIGURE 22 Matching document

In the third query, the results were analysed with time using the Discovery Query Language. The code for the query was: *timeslice(publication_date,1month)*. The used words for the natural language query were "Big Game Hunting" and "Ransomware." In plain language query means: "*Find the number of documents that include words Big Game Hunting and Ransomware. Aggregate the results monthly by the publish date.*" The first date that the answer included was 1.12.2001 and the last date is the same as when the query was created 1.7.2019. It is possible to filter the

queries by date. In figure, 23 is part of the results from the beginning of 2018 to the query date. The information from the query that can be obtained from the results is the trend considering the Big Game Hunting Phenomenon, if the assumption is that the number of documents correlates to the relevance of the phenomenon (figure 23).

FIGURE 23 Data classified with time and the number of documents

In the fourth query, the keywords were similar than in the previous queries. The results were aggregated with a geographical location using the following query language code*: term(country,count:10)*. In the plain language: *Find the number of documents that include words Big Game Hunting and Ransomware. Aggregate the results by the country where the document has been published. Give the top ten results.* The results of the query are in figure 24.

FIGURE 24 Data classified by geographical location

The final query for the second solution was a combination of the third and fourth queries. The used Discovery Query Language code was: *timeslice(publication_date,1month).term(country,count:5).* Again, the words for the query were "Big Game Hunting" and "Ransomware." In plain language, the meaning of the query is: *Find the number of documents that include the words Big Game Hunting and Ransomware. Aggregate the results by the publishing date and the country where the document has been published and show the top five results* (figure 25).

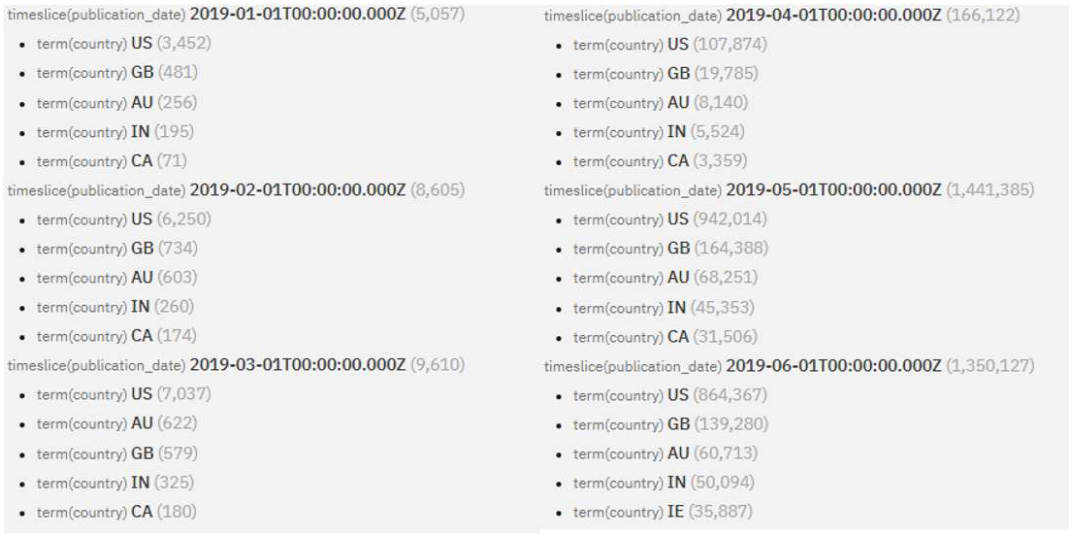

FIGURE 25 Data classified by geographical location, time and number of documents

6.3 Conclusion

The created artifact from the second solution is: *Targeted queries from Watson Discovery News to answer the intelligence questions concerning Big Game Hunting.*

Even though the Watson Discovery News is a dataset that includes news from the various domains, it proved to be suitable data for investigating the Big Game Hunting phenomenon. Compared to the first solution, there was no possibility to use a custom model for the queries, but in perspective of the acquired results, the relevance of the custom model is minor. Even the lack of ability to train the Watson or the lack of possibility to add one's own data did not prevent from finding consistent data for the intelligence direction. An important observation of the data was the fact that the amount of documents changes almost continuously. It means that the latest news is available for the queries.

6.3.1 The queries

In the first query, Discovery returned documents that were not suitable for the intelligence direction. Discovery did not create similar passages than in the first solution. When the IBM instructions were searched, it was confirmed that there is no such property available (IBM, 2019).

The second query with natural language using the words "Big Game Hunting" and "Ransomware" returned over 3.000.000 documents that included the queried words. Due to the amount of the documents, it was not confirmed if the results included documents that are not related to Big Game Hunting as a cyberspace phenomenon, but the first ten most relevant documents were all suitable from the perspective of the intelligence direction. The essential information for the intelligence direction was found in the third document. A critical issue that was found out in the second query was the availability of the document URL. When the URL is available, the document can easily be downloaded and added to an intelligence organization's separate database for further use.

The final three queries were not similar than the first solution. That is because the custom model is not available in the Watson Discovery news, and the intention was to find queries that serve the intelligence direction best. Furthermore, the data in Discovery News is labeled differently and for that reason, it is not possible to create similar queries than in the first solution.

The results from the third query were essential. The query was created with a time perspective that is one key element in the intelligence direction. The results of the query were created based on finding answers to the "When" and "How" intelligence questions. Even though the general properties of the Watson Discovery News states that the data is 60 days old, the first documents were from the year 2001. The data from 2018 to 2019 was used to create a diagram (figure 26). It is easy to obtain a trend of Big Game Hunting and Ransomware related documents in the dataset.

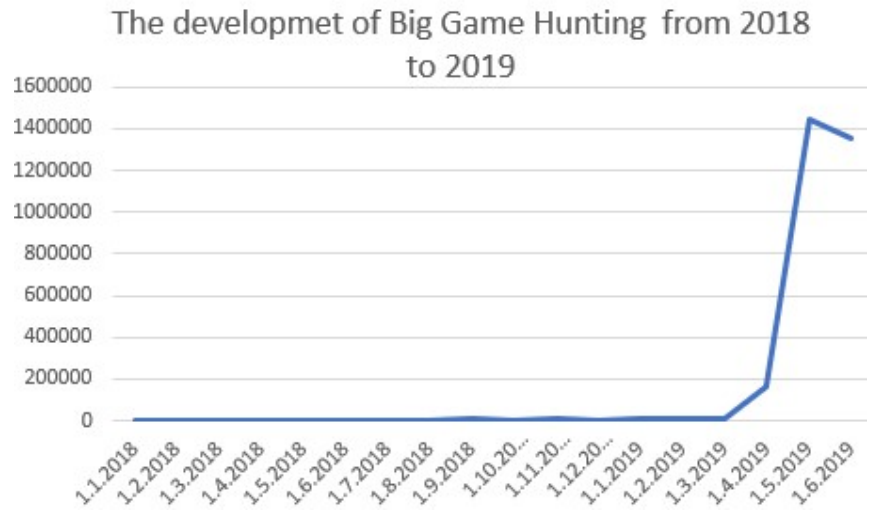

FIGURE 26 The trend of Big Game Hunting

In the third query, the interval of time was one month. When the query was created, it was found out that the shortest interval is one minute. As mentioned, the data on Discovery changes continuously. If the query would have been written in the following code: *timeslice(publication_date,1minute)*, the result would have provided almost real-time tracking of Big Game Hunting Ransomware campaign.

In the fourth and fifth query, the amount of the published documents was attached to time and geographical location. Afterward, the data was moved to excel. The results are in figure 27.

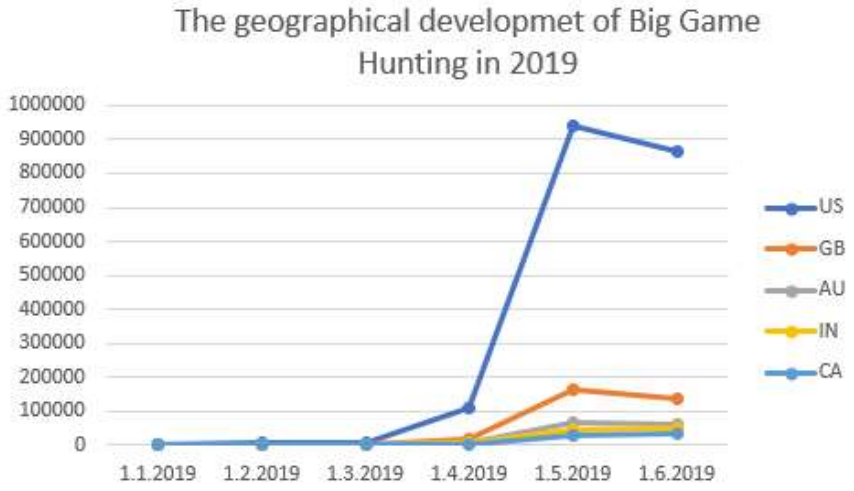

FIGURE 27 The geographical development of Big Game Hunting in 2019

When the data was visualised, it is easy to obtain information in order to answer the intelligence direction sub-questions “Where” and “When” and “How.”

6.3.2 Evaluation of the artifact

1. Can the artifact find related information from the big data?

Referred to the second query the artifact can find related information

2. Can the artifact analyze the collected information?

Referred to the third and fourth query, the artifact can be used to analyse the collected information.

3. Can the artifact provide enough reasonable information to the intelligence direction and its sub-questions?

Referred to the fourth and fifth query, the artifact can provide reasonable information to the intelligence direction and its sub-questions. Watson Discovery News is suitable for collecting, processing and analysing information in strategic cyber intelligence.

7 DISCUSSION AND CONCLUSIONS

7.1 The suitability of the research method

Initially, there were two options for research methods of the study: qualitative case study or DSRP. The Qualitative case study method gives tools for researching complex phenomena in their framework. It might be a suitable method when answering the questions of how and why (Baxter & Jack, 2008). Observing the context of the research, research questions it is easy to conclude that qualitative case study method would be suitable for the examination. The goal of the study would have been similar; to find new points of views for the phenomenon of how AI would ease collecting, analyzing and reporting the information about the cyberspace strategic situation and how AI should facilitate phases of the intelligence cycle.

The qualitative case study was an option for the study, but in the framework of the research, there would have been a risk that the research questions would have been too diffuse. In the case study, it should be remembered that there should not be too many objectives for the research (Baxter & Jack, 2008). ML, Intelligence cycles' phases, and their combination include too much research objectives for a master's level thesis. For those reasons, a qualitative case study was not selected as a research method.

The DSRP was a suitable method for the study. The first and second phases were planned in co-operation with the University of Jyvaskyla and IBM. The actual research began in the third phase, where questions were created to support the development of the artifact. Proper development requires good knowledge about the properties and restrictions that are present. Before the actual building of the artifacts began, the IBM documentation was studied precisely.

The evaluation was based on the artifacts' ability to obtain information from the data. The selected approach for evaluation was loose and depended on the extent of the research. A detailed and rigor evaluation would have required more time and resources than were available. The essential element of Design Science was fulfilled during the study. The new knowledge was obtained through DSRP.

7.2 The reliability and validity of the literature and interviews

The written materials were used for describing the elements of ML, Intelligence process, and IBM Cloud service characteristics and properties.
Since intelligence is closely related to the military domain, part of the used documents were public military field manuals. The reliability of military sources is relatively decent in the scope of the research. It should be noted that a part of

military literature and manuals are classified, but issues that concern the study are commonly known principles, and their reliability is sufficient. The authors of the books: Open-source intelligence techniques: Resources for searching and analyzing online information, The five disciplines of intelligence collection and Strategic intelligence: A handbook for practitioners, managers, and users have all rigid background from governmental intelligence in the U.S. The authors have worked in universities as professors. The data collected from intelligence are general principles, and there is only minor variation in the information concerning the intelligence process. The information about intelligence is verified against multiple sources. The objectivity of intelligence sources seems to be useful since the data is based on facts instead of opinions. The only issue that influences the objectivity of intelligence literature is the military or governmental background of the authors. The experiences might cause some bias; on the other hand, the availability of information about intelligence from different sources than the military is minor. The used sources for intelligence are relatively new, and the information is current since the principles of intelligence have remained similar during the last decades. The information concerning the requirements of this research is deep enough, and even though it provides the basics of intelligence, the coverage is in-depth.

The collected information about ML is accurate. The principles of ML is reliable since it is based on proven facts. There is no variation concerning ML compared to IBM sources and Kulkarni or Murphys's publications. Also, Kulkarni's book is a source from Jyvaskyla University's study material: Basics and Applications of Artificial Intelligence.[4]

Because the IBM Cloud is in a vital role in the research, there is material from manuals of the service. Since IBM is a commercial company, the information might be biased and should be carefully evaluated. Other sources about ML are reliable because they concern commonly known facts, and there is no variation. The information about ML is from the current decade; it is on-date. The information provides in-depth coverage of ML and meets the requirements of the study.

The primary sources for information about cyberspace are the US. Field manual: Cyberspace Operations and book: The fog of cyber defense. The definition of cyberspace is complicated, so it is difficult to prove the reliability of the sources. The information is verified against multiple sources, and variation exists on the appearance of cyberspace. There are some common characteristics in each definition of cyberspace, but as mentioned, cyberspace is a complex phenomenon, and there is not only one correct definition. The authors of the sources used in the chapter concerning cyberspace have an excellent reputation in the subject field. The Cyberspace operations field manual is one point of view of cyberspace, and it is the way that the US military understands the cyber environment. There is some bias in the field manual since it is written information from a military perspective. However, the information about cyberspace is current and fulfills the requirements of the research.

[4] Tekoalyn perusteita ja sovelluksia

In the research, one unstructured interview was used with NSCS-FI to map the use case of ML and to support the DSR cycle to create a background for the researched artifact.
The procedure for the interview began with a briefing where the issue was presented. During the presentation, the requirements for the meeting were expressed in a slideshow. After the slideshow, there was a discussion about the topic. The results of the review were recorded as a memo. The memo is attached as appendix 1 of the research. The participants of the meeting were specialists of cybersecurity from NCSC – FI. The reliability of the interviews is good. The reliability might vary because the used method is the unstructured interview. The collected information met the requirements of the research.

7.3 Combining ML and strategic open-source intelligence

The intelligence direction for the study was imaginary, but organisations such as NCSC-FI might require the same kind of information. The direction was built based on the meeting. One option was to create an in-depth analysis of cyberspace phenomenon. Another alternative would have been to ask NCSC- FI to provide an intelligence direction, but in that case, the nature of the study would have changed to an intelligence task. Another reason for this kind of approach was the possibility to use the direction in the evaluation of the artifacts and as a guideline to build a suitable ML model.

Even the first solution did not pass the evaluation, it can be used for data collection in restricted data that is not available for everyone. The data in the first solution included only 840 documents, and in that perspective, it is not possible to evaluate the suitability of the solution to strategic intelligence collection since the correct estimate of the investigated phenomenon requires comprehensive data collection from all potential sources. Also, the type-system did not support strategic information in the best possible way, as stated in chapter 5.3.3. When the private data is ingested to Watson Discovery, the service labeled the data with unsupervised ML. The labeling of the data with properties mentioned in the chapter 3.4.1 Watson Discovery, and with custom ML model entities fulfill the requirement of processing the data into a consistent format.

Instead, the second solution is suitable for strategic cyber intelligence with similar arguments. The second solution fulfilled the requirements of the information collection. The volume of the data was enormous, almost 14.000.000 documents. The automatic data labeling property is a crucial observation in the research. Large amounts of the data are easily dived in the elements that can further be analysed with queries. Watson Discovery creates labels for the sources of the Discovery News dataset similarly.

As mentioned in chapter 3.2.3 intelligence cycle, the analysis of the collected data requires a systematic approach to the problem-solving — the information-analysis of the collected information, based on various queries. The queries were written in a combination of natural language and Watson Discovery Query

Language. In both solutions, only a minor part of the available query properties was tested. Nevertheless, the used queries provided sufficient information to answer the research questions of the study. In both solutions, the data was qualitative, and the queries revealed hidden knowledge. An excellent example of the ML capabilities in information analysis is the second solutions final query where time, location, and the intensity of the news about Big Game hunting were combined. Without ML, it would have been complicated to obtain the same knowledge.

Since the reliability and validity of the answers to the intelligence questions are not included in the research, there is no source criticism on the documents that the solutions found. In the first solution, the data was initially vetted by NSCS – FI, but in the second solution, it was only checked that the artifact could find suitable documents from the Watson Discovery News dataset.

Both solutions are suitable for OSINT. The first solution is more targeted than the second and as noticed from the results that the first solution provided more accurate and detailed results.

During the research, it was found out that that ML would have been suitable for other three use cases in chapter 4. If the type-system used in the first solution would have been modified according to cyber-weather reports overview chapter so that entities matches to network functionality, cyber espionage, vulnerabilities, and other subchapters (National Cyber Security Centre, 2019). It would be possible to label the data for further analysis.

It was found out in the second solution (figure 19) that the URL of the matching document is available in the results. The documents can be saved and added for further for own datasets, such as the documents in chapter 5.1.2.

In the second solution, the possibility to obtain trends was found. Even though in the definition of The Watson Discovery News Database, the age of documents is 60 days, some documents were published in 2001. The query properties that was used in the second solution fifth query were basic. There is a possibility to use multi-leveled aggregations and filters for the queries. If all properties are used, the queries are targeted, and more detailed information about the trends can be obtained.

Keyword extraction was tested in the first solutions fourth query. When the private data was ingested to Watson Discovery, according to chapter 3.4.1, the system automatically creates the keywords from the data. In figure 13, keywords that Watson Discovery created are present. The keywords can be used further for targeted queries.

If the custom model entities are compared to the HTML labels (figure 14), it can be observed that basically, the labels are almost similar to custom model entities. For example, the antivirus and malware matches to entities RANSOMWARE and operating system are a subtype of LOGICAL_NETWORK_LAYER. The building of the custom model was an enormous task. Watson Discovery was able to create suitable labels for queries with ML automatically. The property can help vastly in the case of private data.

7.4 Limitations

The academic license restricted the study in three ways. First, in the IBM Knowledge Studio, the number of available ML model training sessions was limited so, that the ML model just reached the minimum acceptable score for deployment to Watson Discovery as in figure 8.

Second, in the Watson Discovery, the available disk space was only 200 megabytes, and the amount of the documents was limited to 1.000. Due to the limitation of disk space, approximately 300 documents that NCSC-FI provided were left out of the data. The size of the data was approximately 50 MB, but during the upload, the Discovery creates metadata, that uses the available space. For that reason, the Discovery accepted only 840 documents.

The third restriction was the number of queries in the academic license; the amount is limited to 200 queries monthly. Due to the limitation, only essential query properties in the perspective of the use cases were discovered.

7.5 Answers to the research questions

Answers to the sub-questions:

1. Can machine learning be utilised in information collection?

Machine Learning can be utilised in information collection for strategic cyber open-source intelligence. If the queried data is private data or Watson Discovery News, by using natural language queries in Watson Discovery service, information for intelligence direction can be collected.

2. Can machine learning be utilised in information procession?

Machine Learning can be utilized in information procession. When data is fed into the Watson Discovery service by using unsupervised learning, the service creates metadata that is used for information classification. If the custom ML model is used, the service also creates the classification by custom model entities.

3. Can machine learning be utilised in information analysis?

Machine Learning can be utilised in information analysis. By using diverse queries in the Watson Discovery service, the collected and processed information can be analysed according to intelligence direction.

Answer to the main research question:

Can machine learning be utilised for strategic cyber intelligence?

Machine learning can be utilised for strategic cyber intelligence.

7.6 Further research

The key finding from the study is that ML can be used in intelligence cycles collection, processing, and analysis. The queries in the Watson Discovery were made in two ways: With the natural language and with Discovery Query Language. IBM offers Application Programming Interface in two directions in the Watson Discover: to ingest data into a system and to output the result to the application. In future research, the requirements of an application for Open Source Strategic Intelligence should be studied. The dissemination phase of the intelligence should be included for the application and plan the output of intelligence information compatible with international threat repositories.

The reliability and validity of the intelligence questions answers were not evaluated. There are multiple ways of taking advantage of artificial intelligence in the IBM Cloud. In future research, the validity and reliability should also be included. Furthermore, using the AI capabilities to evaluate the answers might be one part of future research.

In cyber intelligence, the potential of IBM Cloud should be studied in cyberspaces logical network layer where the ML capabilities would be used to explore the code of the malicious programs.

REFERENCES

Alpaydin, E. (2016). *Machine learning : The new AI*. Cambridge, Massachusetts ; London, England: The MIT Press. Retrieved 15.5.2019 from https://ebookcentral.proquest.com/lib/jyvaskyla-ebooks/detail.action?docID=4714219

Clark, R. M., & Lowenthal, M. M. (2016). *The five disciplines of intelligence collection*. Thousand Oaks, California: CQ Press.

ENISA. (2017). WannaCry ransomware outburst Accessed. June 16, 2019 https://www.enisa.europa.eu/publications/info-notes/wannacry-ransomware-outburst

ENISA. (2018). *ENISA threat landscape report 2018*. Accessed June 4, 2019 https://www.enisa.europa.eu/publications/enisa-threat-landscape-report-2018

ENISA. (2019). Ransomware. Accessed Jun 16, 2019 https://www.enisa.europa.eu/topics/csirts-in-europe/glossary/ransomware

Feeley, B., Hartley, B. & Frankoff, S. (2019). PINCHY SPIDER adopts "Big game hunting" to distribute GandCrab. Accessed June 5, 2019 https://www.crowdstrike.com/blog/pinchy-spider-adopts-big-game-hunting/

George, R. Z., & Bruce, J. B. (Eds.). (2008). *Analyzing intelligence : Origins, obstacles, and innovations*. Washington D.C.: Georgetown University. Retrieved 13.5.2019 https://ebookcentral.proquest.com/lib/jyvaskyla-ebooks/detail.action?docID=547760

Goldman, J. (2011). *Words of intelligence : An intelligence professional's lexicon for domestic and foreign threats* (2nd ed ed.). Lanham, Md.: Scarecrow Press. Retrieved 26.5.2019 https://ebookcentral.proquest.com/lib/jyvaskyla-ebooks/detail.action?docID=686277

IBM. (2016). Knowledge studio - IBM cloud. Accessed June 2, 2019 https://cloud.ibm.com/catalog/services/knowledge-studio

IBM. (2017). IBM cloud docs. Accessed June 6, 2019 https://cloud.ibm.com/docs/services/knowledge-studio?topic=knowledge-studio-typesystem

IBM. (2018a). The IBM advantage for cognitive discovery cloud architecture - IBM-advantage-paper-for-cognitive-discovery.pdf. Accessed June 1, 2019

https://www.ibm.com/cloud/garage/files/IBM-Advantage-Paper-for-Cognitive-Discovery.pdf

IBM. (2018b). IBM cloud docs. Accessed June 17, 2019 https://cloud.ibm.com/docs/services/watson-knowledge-studio?topic=watson-knowledge-studio-team

IBM. (2018c). IBM cloud docs. Accessed June 2, 2019 https://cloud.ibm.com/docs/services/watson-knowledge-studio?topic=watson-knowledge-studio-create-project

IBM. (2018d). IBM cloud docs. Accessed June 17, 2019 https://cloud.ibm.com/docs/services/watson-knowledge-studio?topic=watson-knowledge-studio-train-ml

IBM. (2018e). IBM cloud docs. Accessed June 26, 2019 https://cloud.ibm.com/docs/services/discovery?topic=discovery-query-aggregations

IBM. (2019a). IBM cloud docs. Accessed June 2, 2019 https://cloud.ibm.com/docs/services/discovery?topic=discovery-about

IBM. (2019b). IBM cloud docs. Accessed June 23, 2019 https://cloud.ibm.com/docs/services/discovery?topic=discovery-configservice#adding-enrichments

IBM. (2019c). IBM cloud docs. Accessed June 24, 2019 https://cloud.ibm.com/docs/services/discovery?topic=discovery-query-concepts

IBM. (2019d). IBM cloud docs. Accessed June 24,2019 https://cloud.ibm.com/docs/services/discovery?topic=discovery-improving-result-relevance-with-the-tooling

IBM. (2019e). IBM cloud docs. Accessed June 25, 2019 https://cloud.ibm.com/docs/services/discovery?topic=discovery-query-parameters

IBM. (2019f). IBM cloud docs. Accessed July 1, 2019 https://cloud.ibm.com/docs/services/discovery?topic=discovery-watson-discovery-news

Infradata. (2019). Global cyber threat report 2019 | infradata. Accessed June 1, 2019 https://www.infradata.com/news-blog/global-cyber-threat-report-2019/

Joint Chiefs Of Staff. (2013). Joint publication 2-0, joint intelligence. Accessed May 12 2019 https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp2_0.pdf

Joint Chief Of Staff. (2018). JP 3-12, cyberspace operations, 8 june 2018 - jp3_12.pdf. Accessed May 3, 2019 https://fas.org/irp/doddir/dod/jp3_12.pdf

Kulkarni, P. (2012). *Reinforcement and systemic machine learning for decision making*. Hoboken New Jersey: John Wiley & Sons. Retrieved 18.4.2019 http://ieeexplore.ieee.org/xpl/bkabstractplus.jsp?bkn=6266787

Lehto, M., Neittanmäki, P., Niinimäki, E., Nyrhinen, R., Ojalainen, A., Pölonen, I., . . . Äyrämö, S. (2019). *Tekoälyn perusteita ja sovelluksia.* Retrieved 15.5, 2019. https://tim.jyu.fi/view/kurssit/tie/tiep1000/tekoalyn-sovellukset/kirja#zC9FEBTcviMC

Liebowitz, J. (2006). *Strategic intelligence : Business intelligence, competitive intelligence, and knowledge management*. London: Auerbach Publications. Retrieved 27.4.2019 http://ebookcentral.proquest.com/lib/jyvaskyla-ebooks/detail.action?docID=290094

Loeb, L. (2019). Enterprise is the target of 'big game hunting'. Accessed June 5, 2019 https://www.securitynow.com/author.asp?section_id=649&doc_id=750049

McDowell, D. (2009). *Strategic intelligence : A handbook for practitioners, managers, and users* (Rev. ed ed.). Lanham (Md.): Scarecrow Press. Retrieved 16.4.2019 https://www.dawsonera.com/guard/protected/dawson.jsp?name=https://login.jyu.fi/idp/shibboleth&dest=http://www.dawsonera.com/depp/reader/protected/external/AbstractView/9780810862852

Microsoft. (2019). Ransomware. Accessed June 16, 2019 https://docs.microsoft.com/en-us/windows/security/threat-protection/intelligence/ransomware-malware

National Cyber Security Centre. (2019). Cyber weather. Accessed June 3, 2019 https://www.kyberturvallisuuskeskus.fi/en/ncsc-news/cyber-weather

Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., & Bragge, J. (2006). The design science research process: A model for producing and presenting information systems research. Paper presented at the *Desrist 2006,* 83-106. Retrieved 27.5.2019 https://jyx.jyu.fi/handle/123456789/63435 http://www.urn.fi/URN:NBN:fi:jyu-201904092111

Rantapelkonen, J., & Salminen, M. (Eds.). (2013). *The fog of cyber defence*. Helsinki: National Defence University. Retrieved 10.6.2019 http://urn.fi/URN:ISBN:978-951-25-2431-0

Roberts, S. J. (2015, -02-16T00:00:00.000Z). Intelligence concepts — the intelligence cycle. Accessed May 14, 2019 https://medium.com/@sroberts/intelligence-concepts-the-intelligence-cycle-f25ec067f1d6

Rouse, M. (2017). What is IBM cloud (formerly IBM bluemix and IBM SoftLayer)? - definition from WhatIs.com. Accessed July 13, 2019 https://searchcloudcomputing.techtarget.com/definition/IBM-Bluemix

Sanastokeskus TSK ry. (2018). *Vocabulary of cyber security* Helsinki: Accessed June 6, 2019 https://turvallisuuskomitea.fi/wp-content/uploads/2018/06/Kyberturvallisuuden-sanasto.pdf

Secretariat of the Security Committee. (2013). *Finland´s cyber security strategy.* Helsinki: Accessed May 25, 2019 https://turvallisuuskomitea.fi/wp-content/uploads/2018/09/Cyber-Strategy-for-Finland.pdf

Vaishnavi, V., Kuechler, B., & Petter, S. (2004). Design science research in information systems.(2004/17), 1-62. Accessed May 20, 2019 from http://desrist.org/desrist/content/design-science-research-in-information-systems.pdf

Watson, B. W. (1998). Intelligence | military science. Accessed May 26, 2019f https://www.britannica.com/topic/intelligence-military

APPENDIX 1

Meeting with NCSC-FI 21.5.2019
MEMORANDUM
The meeting started at NCSC-FI's premises in Helsinki at 10:05

Participants:
Janne Voutilainen
Riikka Valtonen
Juhani Eronen
Tomi Kinnari
Ilkka Sovanto

Initially, Janne Voutilainen introduced the current situation of research. The presentation looked at the features, capabilities, and requirements of the IBM Cloud service. After the presentation, it was found that the features of the service are suitable for the intended use.

After the presentation, Ilkka Sovanto and Tomi Kinnari introduced the April 2019 cyberspace report and its various sub-sections. It was found that old cyberspace reports and their sources can be used as training data for the AI algorithm. The Cyber Weather report found that machine learning could be suitable for tracking the trends of the cyber weather subsets listed below.

- Network functionality
- Information breaks & leaks
- Malware & vulnerabilities
- Espionage
- Fraud and fishing
- IoT and automation

It would be good to follow the top 5 threats in Cyber Weather, but the problem is the lack of training data because the TOP 5 list has only been released since the beginning of 2019.It was found that it would be possible to link the comparative data to the system from the RSS feed of the NCSC – FI. One way to use Discovery is to explore a specific topic that aims to deepen the understanding of the subject.

NCSC-FI uses many news sources as sources, and other sources include real-time data from the HAVAR system. HAVAR observations are not included in the study due to the confidentiality of the information.
Tomi Kinnari introduced the definition of strategic cyberspace situation awareness. According to Kinnari, there is no one universally applicable definition, and the cyber situational awareness is very different from the traditional snapshot tied to the "god-eye" chart and is not comparable.
Twitter is a good source for cyber-tracking, improving the validity of

knowledge if you can follow the right person at the right time. The challenge is to find a person when tweets are important for cybersecurity.
Riikka Valtonen suggested that during a campaign such as NotPetya it would be good if artificial intelligence could be used to detect campaign activation.

Until now, the event has been "closed" only when the intensity of the observations is decreasing. The news feed could have an indicator that predicts the activation of the event, it was found that Discovery's properties could be well suited for this purpose. In this case, use historical information to find indicators. At the time of the NotPetya activation, finding the right information was tricky due to a large flood of information. False findings and news weaken the s snapshot. A pre-prepared artificial intelligence application could help a person to detect true and relevant information.

The object under investigation could be selected by industry where cyber threats to energy supply, for example, are started to be monitored. The Big Game Hunting phenomenon, which is well suited for analysis, emerged from the Cyber Weather report. The service could be used so that news related to the phenomenon is used as training data for artificial intelligence. Links to this phenomenon are available from the Cyber Security Center.
It was noted that the University's IBM Discovery - LITE resources may be inadequate. It was agreed that Janne Voutilainen would find out whether it is possible to get additional resources for free. The Cybersecurity Center will find out whether it would be possible to obtain funding for additional resources. It was noted that the conditions are good because the research topic is new and relevant. The idea was that it would be a good idea to get geographic locations for events.

It was found that the use of artificial intelligence could improve the snapshot by investigating the victims of cyber-attacks and the choice of means used by the attacker. By discovering the principle or knowledge that an attacker chooses a victim or a victim organization, it could be possible to reveal and anticipate cyber-attacks.

It was found that Discovery's properties are suitable for finding event trends, but the period of the research is so short that trends may not be possible. One product is a remarkable product that Discovery could produce is to include cyber-attacks in response to shortening the response time, i.e. the temporal dimension.

Also, an application was found to revise and strengthen their own hypotheses.
Another case of use was a situation where a good source article is found on the phenomenon under investigation, keywords are extracted from the article, and new articles are sought based on keywords.
At the end of the meeting, the table below shows the main uses, information, and product. Janne Voutilainen will discuss in more detail the best-suited application with IBM and supervisor of the research

Possible Usage	Training / Base / New Data Source	Product
Cyberweather report or part of it: Phenomenon: e.g. vulnerabilities Industries: e.g. energy supply	Old cyberspace portals, news sources, energy reports	Analysis of the phenomenon
BIG GAME HUNTING-type phenomenon. Showing the exact question "5WH." Explore trends, changes, and subtleties of the phenomenon. Correlation between time and subject	TTM - Publication Articles related to the phenomenon or topic	Answer to the question possible "network." visualization
Finding a trend Observing the upward trend, for example, in the case of BIG GAME HUNTING. Confirming Your Own Observation	Public News, blogs, cyber security newsletter Watson Discovery News Database Virus total service	Early warning, early detection, broadening understanding, understanding the importance of meaning Finding Keywords
A set of articles from keywords, keyword	extension News Twitter hashtag # blogs	If the phenomenon being investigated is an entity→ new keywords “Snapshot” about the situation

Also, it was agreed that the Cyber Security Center would provide Janne Voutainen with news data for the current year, adapted to the format to continue the investigation.

The meeting ended at 13:00