Tuomas Niemelä

STRATEGISEN TIEDUSTELUN TUKEMINEN TIEDONLOUHINNALLA – UUTISDATASTA TIEDUSTELUTIEDOKSI

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2021

TIIVISTELMÄ

Niemelä, Tuomas Ville Hermanni
Strategisen tiedustelun tukeminen tiedonlouhinnalla – uutisdatasta tiedustelutiedoksi
Jyväskylä: Jyväskylän yliopisto, 2021, 82 s.
Kyberturvallisuus, pro gradu -tutkielma
Ohjaaja(t): Kari, Martti J

Strateginen tiedustelu tuottaa tietoa maailmasta ympärillämme kansallisen ja kansainvälisen tason suunnittelun ja päätöksenteon tueksi. Strategisella tasolla tiedustelun kohteita ovat kansainväliset suhteet, konfliktit, terrorismi, järjestäytynyt rikollisuus ja muut merkittävät ilmiöt. Tietoa kerätään esimerkiksi henkilölähteistä, viestiliikenteestä, satelliittikuvista ja avoimista lähteistä, kuten mediasta. Merkittävä osa strategisen tiedustelun tiedoista voidaan hankkia avoimista lähteistä. Teknologian kehityksen myötä tiedustelujärjestelmät kykenevät keräämään dataa niin suuria määriä, että ihmiset eivät ehdi käsittelemään niitä. Informaatioteknologian hyödyntämistä ihmisten tekemän tiedusteluanalyysin tukena on kuitenkin tutkittu vain vähän. Tieto on tiedustelun keskeisiä resursseja ja sitä käsitellään pääsääntöisesti tietojärjestelmien avulla, joten on luontevaa tutkia ja kehittää tiedustelun toimintaa hyödyntämällä tietojärjestelmätieteen teorioita ja tutkimusmenetelmiä. Tiedonlouhinta on tietojärjestelmätieteen tutkimusalueella kehitetty prosessi, jonka avulla voidaan tuottaa tietoa suuresta määrästä dataa. Tässä tutkielmassa selvitetään, miten tiedonlouhinnan avulla voidaan tukea strategista tiedustelua tuottamalla tietoa uutisdatasta. Vastaus kysymykseen selvitetään suunnittelututkimuksen metodologiaa hyödyntäen. Kirjallisuuskatsauksen avulla muodostetaan tietopohja, johon perustuen suunnitellaan ja kehitetään prosessimalli, jossa yhdistetään tiedonlouhinnan ja tiedustelun toimintoja. Prosessimallin toimivuus todennetaan prototyyppisovelluksen avulla. Tutkimusaineistona käytetään Global Database of Events and Tone -tietokantaa, josta prototyyppisovelluksen avulla tuotetaan tiedustelutietoa viiden strategisen tiedustelun toimintaa kuvailevan skenaarion ohjaamana. Skenaarioissa hyödynnetään erilaisia laskennallisia ja koneoppimiseen perustuvia menetelmiä. Tulokset osoittavat, että tiedonlouhinnan avulla voidaan tuottaa informaatiota strategisen tiedusteluanalyysin tueksi sekä automatisoida tiedustelutietojen keräystä ja prosessointia. Tutkielma tarjoaa käytännöllisen esimerkin, miten tiedusteluorganisaatiot voivat hyödyntää tiedonlouhintaa toimintansa tukena. Lisäksi tutkielma osoittaa, että tietojärjestelmätiede soveltuu hyvin tiedustelun tutkimukseen. Tiedonlouhinnan soveltaminen tiedustelun kontekstiin voidaan nähdä uutena tutkimusalueena, jossa kehitetään menetelmiä eri tiedustelun keräysmenetelmillä hankitun datan tutkimiseen.

Asiasanat: tiedonlouhinta, tiedustelu, suunnittelututkimus

ABSTRACT

Niemelä, Tuomas Ville Hermanni
Supporting strategic intelligence with knowledge discovery and data mining – from news data to intelligence
Jyväskylä: University of Jyväskylä, 2021, 82 pp.
Cyber Security, Master's Thesis
Supervisor(s): Kari, Martti J

Strategic intelligence produces information and knowledge about the world around us to support planning and decision-making at the national and international levels. At the strategic level, intelligence interests include international relations, crises, conflicts, terrorism, organized crime, and other significant phenomena. Intelligence is collected, for example, from personal sources, communications, satellite images, and open sources such as the media. A significant part of strategic intelligence can be obtained from open sources. With the development of technology, intelligence systems can collect such large amounts of data that people do not have time to process them. The use of technology to support people in intelligence has been identified, but little research has been done on its use. Information is a key resource for intelligence, so it is natural to study and develop the operation of intelligence by utilizing theories and research methods in information systems science. Knowledge discovery in databases (KDD) is a process developed in the research area of information systems. It can be used to produce information and knowledge from a large amount of data. This thesis explores how KDD can be applied to support strategic intelligence by generating information from news data. The answer to the question is formed through design science research methodology. A process model was developed to present the KDD application to strategic intelligence. The knowledge base for design and development was formed with the help of a literature review. The functionality of the developed process is demonstrated and evaluated with a prototype application. The Global Database of Events and Tone is used as data, from where the prototype application is used to produce intelligence guided by five scenarios describing the operation of strategic intelligence. Scenarios were developed around key strategic intelligence targets, and each scenario utilizes a variety of computational and machine learning-based methods for data mining. The results show that KDD can be used to produce information to support strategic intelligence analysis and to automate the collection and processing of intelligence. This thesis provides a practical example of how intelligence organizations can utilize information technology to support their operations. In addition, this thesis shows that information systems science is well suited for intelligence studies. The application of KDD to intelligence can be seen as a new area of research, where methods are being developed to study data obtained by various intelligence collection methods.

Keywords: knowledge discovery, data mining, intelligence, design science research

KUVIOT

TAULUKOT

SISÄLLYS

1 JOHDANTO

Strateginen tiedustelu tuottaa tietoa maailmasta ympärillämme kansallisen ja kansainvälisen tason suunnittelun ja päätöksenteon tueksi. Strategisen tason tiedustelun kohteita ovat kansainväliset suhteet, konfliktit, terrorismi sekä merkittävät ja laajat ilmiöt. Tiedustelun tärkein tehtävä on vähentää epävarmuutta tuottamalla tietoa. Tiedustelutiedon tuotantoa kutsutaan tiedusteluprosessiksi, joka koostuu viidestä toiminnosta: suunnittelusta ja ohjauksesta, keräyksestä, prosessoinnista, analyysista ja tuotannosta sekä jakamisesta. Tietoa kerätään esimerkiksi henkilölähteistä, viestiliikenteestä, satelliittikuvista ja avoimista lähteistä, kuten mediasta.

Tarkan, hyödyllisen ja ajankohtaisen tiedustelutiedon tuottaminen ei kuitenkaan ole kovin helppoa. Ensimmäisen haasteen tiedustelulle asettaa teknologian kehitys. Dupontin (2003) mukaan teknologian kehityksen myötä tiedustelun keräysmenetelmät ja -järjestelmät ovat kehittyneet niin, että dataa saadaan kerättyä ja tallennettua yhä kasvavia määriä. Kerättyä dataa tai informaatiota on niin paljon, että ihminen ei ehdi käsittelemään ja tuottamaan siitä merkityksellistä tietoa. Uutta teknologiaa ja älykkäitä järjestelmiä on hyödynnetty tiedustelutietojen keräyksessä ja prosessoinnissa, mutta niiden hyödyntäminen muissa prosessin vaiheissa on jätetty vähemmälle huomiolle (Dupont, 2003). Poikkeavien ja kiinnostavat ilmiöiden hahmottaminen suuresta datamassasta on erittäin haastavaa (Quiggin, 2007). Toisen haasteen tiedustelulle asettaa strategisen ympäristön muutos, epävarmuus ja aikapaine. Strateginen ympäristö on muuttunut niin, että käsiteltävien asioiden kohteiden määrä on aikaisempaa suurempi (Treverton, 2003; Clark, 2019). Lisäksi päätöksentekijät odottavat vastauksia nopeasti (George, 2014). Ihmisellä on luonnostaan kognitiivisia rajoitteita (ks. Tversky & Kahneman, 1974), jotka vaikuttavat myös tiedusteluanalyysiin (Heuer, 2009). Kognitiivisia rajoitteita voidaan purkaa hyödyntämällä rakenteellisia analyysimenetelmiä (Heuer & Pherson, 2020), mutta aikapaineen vuoksi ne jäävät usein hyödyntämättä (Chin, Kuchar & Wolf, 2009).

Tiedustelun haasteita voidaan kuitenkin ratkaista tutkimuksen kautta. Tiedustelun tutkimusperinne on alkanut kehittymään toisen maailman sodan jälkeen. Vaikka Kentin (1993) mukaan tiedustelu on oma tieteenalansa, sitä ei toistaiseksi ole sellaiseksi yleisesti tunnustettu. Marrin (2014) esittää, että tie-

dustelun tutkimus (intelligence studies) on ennemmin akateeminen suuntaus, joka tutkii kansallisen turvallisuuden ja tiedustelun käytäntöjä. Tiedustelun tutkimukseen ei ole kehitetty omia tutkimusmenetelmiä. Gill ja Phythian (2016) ovat tunnistaneet tiedustelun tutkimuksesta neljä pääsuuntausta, joissa sovelletaan tutkimusmenetelmiä eri tieteenaloilta yhteiskuntatieteistä oikeustieteisiin. Nämä suuntaukset ovat: tiedustelun historian tutkimus, tiedustelua määrittelevä tutkimus, tiedusteluorganisaatioiden ja -toiminnan tutkimus sekä tiedusteluhallinnon ja -politiikan tutkimus.

Jos ajatellaan, että tieto on tiedustelun keskeisiä resursseja ja teknologian kehityksen myötä tietoa käsitellään yhä enemmän tietojärjestelmien avulla, on luontevaa tutkia ja kehittää tiedustelua hyödyntämällä tietojärjestelmätieteen teorioita ja tutkimusmenetelmiä. Tässä tutkielmassa tiedustelun tutkimusta lähestytään tietojärjestelmätieteen kautta.

Tiedonlouhinta (Knowledge Discovery in Databases, KDD) on tietojärjestelmätieteen tutkimusalueella kehitetty prosessi, jonka avulla voidaan tuottaa tietoa suuresta datamassasta (Fayyad, Piatetsky-Shapiro & Smyth, 1996). Tietojärjestelmätieteessä tapa tehdä ja esittää tutkimuksia on periytynyt käyttäytymis- ja yhteiskuntatieteistä. Suunnittelututkimus (Design Science Research, DSR) on kehitetty tietojärjestelmätieteiden tutkimusmenetelmäksi, joka palvelee paremmin tieteenalan erityispiirteitä (March & Smith, 1995; Hevner, March, & Park, 2014; Peffers, Tuunanen, & Niehaves, 2018). Tässä tutkielmassa hyödynnetään suunnittelututkimuksen metodologiaa ja selvitetään, miten tiedustelun haasteita voidaan ratkaista tiedonlouhinnan avulla.

1.1 Aikaisempi tutkimus

Kansainvälisen politiikan tutkimuskentässä on tehty jo 60-luvulta lähtien kansainvälisten suhteiden kvantifiointia sekä geopoliittisten tapahtumien analyysiä (Sillanpää, 2010). Yhdysvaltojen puolustusministeriön tutkimusorganisaatio DARPA on yhteistyössä Lockheed Martin -yrityksen kanssa kehittänyt järjestelmää, jonka tavoitteena on tukea kansallisen tason päätöksentekijöitä tuottamalla ennakkovaroituksia kriiseistä ja konflikteista. Järjestelmä kerää uutistekstejä, joista se tuottaa arvioita ja ennusteita maailman tapahtumien seurauksista. (O'Brien, 2010.) Järjestelmään liittyen on julkaistu kattavasti tutkimusta, joissa käsitellään laskennallisten menetelmien soveltamista konfliktien ja kriisien tulkinnassa (Lockheed Martin, 2020). Vastaavaa tutkimusta on tehty myös Global Database of Events, Language and Tone (GDELT) -uutisdataa hyödyntäen. GDELT:n datan avulla on ennustettu väkivaltaisuustasoja Afganistanissa (Yonamine, 2013), valtioiden epävakautta (Qiao, Zhang, Ding, Cheng & Wang, 2017), sosiaalista levottomuutta (Galla & Burke, 2018) ja valtioiden välisten suhteiden kehittymistä (Chen, Jatowt & Yoshikawa, 2020), sekä tulkittu alueellisten konfliktien intensiteettiä (Levinn, Ali & Crandall, 2018). Tiedonlouhinnan avulla on selvitetty kansainvälisen politiikan normeja mallintamalla valtioiden välisiä tapahtumia (Murali, Patnaik & Cranefield, 2020). Tiedonlouhintaa tai vastaavien menetelmien soveltamista on tutkittu rikostiedustelun (McCue, 2014) ja

liiketoimintatiedustelun näkökulmasta (Zanasi, 1998; Dey, Haque, Khurdiya & Shroff, 2011; Khan, 2012; Ziegler 2012). Laajemmin data-analyysin soveltamista tiedustelun kontekstiin on tutkittu big data -käsitteen kautta useissa julkaisuissa (Lim, 2016; Jani & Soni, 2018; Van Puyvelde, Coulthart & Hossain, 2017). Big datan, tekoälyn ja koneoppimisen hyödyntäminen tiedustelussa on laajempi tutkimusalue, jota tutkitaan mm. Yhdysvaltojen kansallista tiedustelua kehittävän The Intelligence Advanced Research Projects Activityn eri projekteissa (IARPA, 2020). Konseptitasolla informaatioteknologian tiedustelulle tarjoamat mahdollisuudet on tunnistettu, mutta teknologian hyödyntämistä tiedusteluanalyysissä on tutkittu varsin vähän (Eldridge, Hobbs & Moran, 2017). Akateemisen tutkimuksen lisäksi, kansainvälisen politiikan tutkimusta tekevät organisaatiot ovat kehittäneet erilaisia tekoälyä hyödyntäviä menetelmiä oman työnsä tueksi. Esimerkiksi, Center for Strategic and International Studiesin (CSIS) alainen Beyond Parellel -organisaatio kerää tietoa avoimista lähteistä ja tuottaa data-analytiikkaa hyödyntäen tietoa Korean suhteista (Center for Strategic and International Studies, 2020).

Aikaisemmassa tutkimuksessa on tunnistettu, että uutisteksteistä voidaan tuottaa dataa, jota käsittelemällä erilaisten laskennallisten menetelmien avulla voidaan tuottaa tietoa kriiseistä, konflikteista ja muista kansainvälisesti merkittävistä tapahtumista. Tiedonlouhinnan hyödyt on tunnistettu rikos- ja liiketoimintatiedustelussa, mutta strategisen tason tiedustelun kontekstissa tiedonlouhintaa ei ole suoraan hyödynnetty. Tiedustelun näkökulmasta on tunnistettu konseptina, että data-analyysin avulla voidaan tukea tiedustelua, mutta käytännön ratkaisuja on esitelty hyvin vähän. Tämän tutkielman tarkoitus on täyttää aikaisemmasta tutkimuksesta tunnistettu puute yhdistämällä tiedonlouhinta strategisen tiedustelun kontekstiin ja esittelemällä käytännön ratkaisu sen hyödyntämisestä.

1.2 Tavoitteet, tutkimusongelma ja -kysymykset

Tutkielmalla on useita eri tasoisia tavoitteita. Ensimmäisenä tavoitteena on esitellä konkreettinen ratkaisu, miten tiedonlouhinnan avulla voidaan tukea strategista tiedustelua. Toiseksi, tutkielman avulla halutaan osoittaa, että tiedustelun tutkimusta voidaan tehdä tietojärjestelmätieteen avulla. Kolmanneksi, tavoitteena on muodostaa ymmärrystä informaatioteknologian tarjoamista mahdollisuuksista suhteessa tiedustelun asettamiin tarpeisiin. Tutkimuskysymys on:

- Miten tiedonlouhinnan avulla voidaan tuottaa uutisdatasta tietoa strategisen tiedustelun tarpeisiin?

Tutkimuskysymystä on jäsennetty seuraavilla alakysymyksillä:

- Millaista tiedustelutietoa uutisdatasta voidaan tuottaa?
- Mitä resursseja tiedonlouhinnan avulla voidaan säästää?
- Mitä tiedustelun toimintoja tiedonlouhinnan avulla voidaan tukea?

Tutkimuskysymyksiin selvitetään vastaus Peffersin, Tuunasen, Rothenbergerin ja Chatterjeen (2007) kehittämää suunnittelututkimuksen metodologiaa (Design Science Research Methodology, DSRM) noudattaen. DSRM:n keskiössä on artefaktin suunnittelu, kehittäminen ja evaluointi (Peffers ym., 2007). Tässä tutkielmassa suunnitellaan ja kehitetään prosessimalli tiedonlouhinnan soveltamisesta strategisen tiedustelun tueksi. Prosessin suunnittelua ja kehittämistä ohjaava tietopohja muodostetaan kirjallisuuskatsauksen avulla. Tiedustelun kirjallisuudesta selvitetään, miten tiedustelu toimii, mistä asioista strateginen tiedustelu tuottaa tietoa ja mistä tiedot kerätään. Tiedustelun tutkimukselle on ominaista sen anglo-amerikkalainen perinne (Gill & Phythian, 2016), jonka vuoksi tässäkin tutkielmassa käsitellään hyvin paljon tiedustelua Yhdysvaltojen näkökulmasta. Tiedonlouhinnan kirjallisuudesta selvitetään, miten tiedonlouhinta toimii ja millaista tietoa eri louhintamenetelmien avulla datasta voidaan tuottaa. Kirjallisuuskatsauksen lähdeaineisto hankittiin pääsääntöisesti Google Scholar -hakukonetta hyödyntäen. Täydentäviä hakuja tehtiin lisäksi Jyväskylän yliopiston kirjaston JYKDOK-palvelun avulla ja joitakin lähdeaineistoja poimittiin löydettyjen aineistojen lähdeviitteistä. Tiedonhankinta aloitettiin hakusanoilla ”strategic intelligence”, ”national intelligence”, ”knowledge discovery in databases” ja ”data mining”, mutta useita tarkentavia hakukierroksia tehtiin eri hakusanoilla kun tietopohja ja siihen liittyvät käsitteet alkoivat muodostumaan.

Prosessin toimivuutta demonstroidaan ja sen käyttökelpoisuutta evaluoidaan prototyyppisovelluksen avulla, jota käytetään tiedustelun toimintaa kuvailevien skenaarioiden kehystämänä. Skenaarioita ohjaavat tiedustelun kohteet ja tarpeet ovat johdettu tiedustelun kirjallisuudesta. Prototyyppisovelluksen avulla tuotetaan uutisdatasta tietoa strategisen tiedustelun kannalta merkittävistä kohteista. Tutkimusaineistona käytetään GDELT-tietokantaa, joka on muodostettu keräämällä eri uutislähteistä tekstejä ja tunnistamalla niistä tapahtumia, toimijoita, paikkoja sekä sävyjä. Aineisto sisältää yli 500 miljoonaa riviä rakenteellista dataa. Prototyypin avulla todennetaan, että tiedonlouhinnan soveltaminen strategisen tiedustelun tarpeisiin on mahdollista. Lopuksi prosessin hyödyllisyyttä ja tehokkuutta arvioidaan laadullisesti tarkastelemalla prototyypin toimintaa ja sen avulla tuotettua tietoa.

Tutkimuksen tulokset osoittavat, että tiedonlouhinnan avulla voidaan tuottaa tietoa strategisen tiedustelu tarpeisiin ja tukea tiedusteluanalyysia. Uutisdatasta tuotettu tieto on pääsääntöisesti tiedusteluanalyysia tukevaa informaatiota, jota voidaan hyödyntää ajattelun tukena tai liittää osaksi tiedustelutuotetta visualisoimaan monimutkaisia ilmiöitä. Prosessin avulla voidaan automatisoida esimerkiksi valtioiden välisten suhteiden seurantaa tai löytää suuresta datajoukosta nopeasti poikkeamia, klustereita, verkostoja tai toistuvia ilmiöitä. Kehitetty prosessi on hyvin joustava ja sitä voidaan hyödyntää erilaiseen dataan ja sen avulla voidaan vastata erilaisiin tiedustelukysymyksiin. Tutkielma osoittaa, että tietojärjestelmätiede soveltuu hyvin tiedustelun tutkimukseen. Tulokset täydentävät sekä tiedustelun että tietojärjestelmätieteen tutkimusta. Tulokset tarjoavat mahdollisuuksia jatkaa tiedustelun tutkimista tietojärjestelmätieteen avulla ja suunnitella ja kehittää uusia artefakteja, joiden avulla voidaan tehostaa ja kehittää tiedusteluorganisaatioiden toimintaa.

Tutkimuksen rakenne etenee seuraavasti. Toisessa luvussa esitellään strategisen tiedustelun kirjallisuuskatsaus, jonka tavoitteena on selvittää miten strateginen tiedustelu toimii ja millaista tietoa strateginen tiedustelu käsittelee. Kolmannessa luvussa esitellään tiedonlouhinnan kirjallisuuskatsaus, jonka tavoitteena on selvittää miten tiedonlouhinta toimii ja millaista tietoa datasta voidaan tuottaa eri menetelmien avulla. Neljännessä luvussa esitellään tutkimusmenetelmä ja -aineisto. Viidennessä luvussa esitellään tutkimuksen tuloksena muodostettu prosessimalli tiedonlouhinnan soveltamisesta strategisen tiedustelun tarpeisiin sekä demonstroidaan ja evaluoidaan prosessin toimivuus ja hyödyllisyys viiden tiedustelun toimintaa kuvaavan skenaarion kehystämänä. Kuudennessa luvussa vastataan tutkimuskysymyksiin sekä tulkitaan ja pohditaan tulosten käytännöllistä ja tieteellistä merkitystä. Lisäksi luvussa esitellään mahdollisia jatkotutkimuksen aiheita. Lopuksi, tutkielman yhteenveto esitellään luvussa seitsemän.

2 STRATEGINEN TIEDUSTELU

> Tunne itsesi ja tunne vihollinen, sadassakaan taistelussa et ole vaarassa. Kun et tunne vihollista, mutta tunnet itsesi, ovat mahdollisuutesi voittaa tai hävitä yhtäläiset. Jos et tunne sen paremmin vihollista kuin itseäsi, häviät varmasti jokaisen taistelun. – Sun Tzu

Tässä luvussa esitellään strategista tiedustelua sekä tiedusteluprosessia. Tiedusteluprosessin eri toiminnot esitellään omissa alaluvuissaan. Avointen lähteiden tiedustelua käsitellään muita keräysmenetelmiä tarkemmin omassa alaluvussa. Tiedusteluanalyysiin on keskitytty muita toimintoja laajemmin, koska sen voidaan katsoa olevan tiedon tuotannon kannalta keskeisimmässä roolissa.

2.1 Strategisen tiedustelun käsite

Tiedusteluanalyysin isäksi kutsuttu Sherman Kent (1949) kuvailee kirjassaan ”Strategic intelligence for American world policy” strategista tiedustelua kolmesta näkökulmasta: tiedon, organisaation ja prosessin. Strateginen tiedustelutieto on hänen mukaansa kuvailevaa, raportoivaa ja spekulatiivista tietoa valtion ulkoisista asioista. Strategista tiedustelutietoa tuottavat tiedusteluorganisaatiot. Strategisen tiedustelutiedon tuotanto on järjestelmällistä toimintaa, jota kutsutaan tiedusteluprosessiksi. (Kent, 1949.)

Jensen, McElreath ja Graves (2017) esittävät, että tiedustelulle ei ole olemassa yhtä yhteisesti tunnistettua määritelmää, koska toimijoita ja toiminnan tasoja on useita. Toimijoita ovat esimerkiksi valtionhallinto, asevoimat tai yksityisen sektorin organisaatiot. Tasoja ovat strateginen, operatiivinen ja taktinen. (Jensen ym., 2017.) Dupont (2003) kuitenkin esittää, että näiden tasojen rajat ovat tiedustelussa häilyviä, johtuen teknologian kehityksen tarjoamista mahdollisuuksista ja päätöksentekijöiden tarkemmista tietotarpeista. Strategisella tasolla viitataan usein kansallisen tason toimintaan. Johnson (2010) käyttääkin teoksessaan strategisesta tiedustelusta käsitettä kansallisen turvallisuuden tiedustelu (national security intelligence). Hän määrittelee sen olevan laajaa tietoutta ja

ennakkotietoutta maailmasta ympärillämme, joka on presidentillisen päätöksenteon johdanto. Goldmanin (2011, s. 241) määritelmän mukaan strateginen tiedustelu tuottaa tietoa kansallisen ja kansainvälisen tason suunnittelun ja päätöksenteon tueksi. Jensen ym. (2017) täsmentävät, että tiedustelussa strategisella tasolla tarkoitetaan laajoja ja pitkäaikaisia asioita, joilla on suurta merkittävyyttä ja seurauksellisuutta. Tämänkaltaisia asioita ovat esimerkiksi maailmantalous tai ydinaseohjelman kehitys (Jensen ym., 2017). Kansallisen tason tiedustelun keskeisiä kiinnostuksen kohteita ovat toisten valtioiden tai tärkeiden eivaltiollisten ryhmien, kuten kansainvälisten organisaatioiden tai terroristijärjestöjen, toiminta, politiikka ja suorituskyvyt (Lowenthal, 2019). Strateginen tiedustelutieto käsittelee henkilöitä, taloutta, sosiologiaa, logistiikkaa, tietoliikenneyhteyksiä, geografiaa, politiikkaa ja tiedettä (Goldman, 2011, s. 241).

Clarkin (2020, s. 13) mukaan tiedustelussa on yleisesti kyse epävarmuuden vähentämisestä konfliktissa. Konfliktilla hän ei tarkoita ainoastaan aseellista yhteenottoa, vaan konfliktitilanteella voidaan ymmärtää mikä tahansa kilpailuasetelma, johon osallistuu kaksi tai useampia osapuolia. Tämä laajentaa käsitettä valtioiden ulkopuolelle, kuten yrityksiin tai muihin organisaatioihin. Tällöin puhutaan tyypillisesti liiketoimintatiedustelusta (Competitive Intelligence, CI tai Business Intelligence, BI) (kts. Liebowitz, 2006).

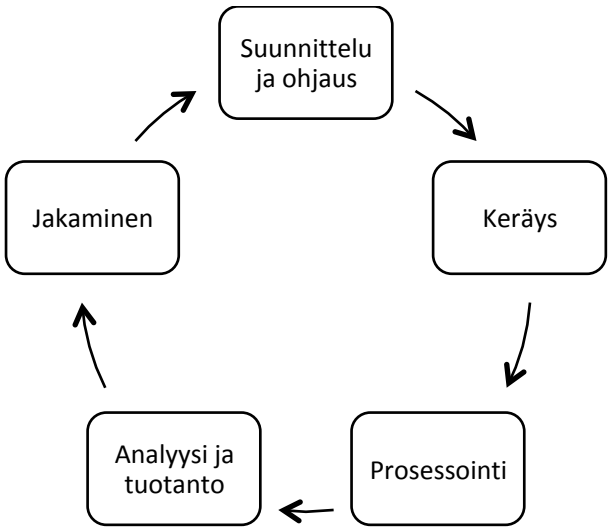
Strategista tiedustelutietoa tuottavat pääsääntöisesti kansalliset tiedusteluorganisaatiot, joista esimerkkejä ovat Yhdysvaltojen keskustiedustelupalvelu CIA (ks. Johnson, 2010), Britannian ulkomaantiedustelupalvelu MI6 (ks. Davies, 2004) ja Venäjän federaation turvallisuuspalvelu FSB (ks. Pringle, 2010). Suomen tiedusteluorganisaatioita ovat Suojelupoliisi (ks. Suojelupoliisi, 2021) ja Puolustusvoimien tiedustelulaitos (ks. Puolustusvoimat, 2021). Kansallisten tiedustelupalveluiden lisäksi strategisen tason tiedustelutietoa tuottavat kaupalliset toimijat, itsenäisten tutkijoiden muodostamat kollektiivit ja kansainvälisen politiikan ja turvallisuuden tutkimusta tekevät organisaatiot. Esimerkiksi Stratfor Enterprises on kaupallinen toimija, joka tarjoaa yrityksille ja globaalin tason päätöksentekijöille Strafor Worldview -tuotetta, jonka tavoitteena on tuottaa geopoliittista tiedustelutietoa ja analyysejä maailman tapahtumien taustamerkityksistä ja tulevaisuuden näkymistä (ks. Stratfor, 2021). Itsenäisten tutkijoiden kollektiivista esimerkkinä voidaan käyttää Bellingcatia, joka on selvittänyt mm. vuonna 2014 Ukrainassa alas ammutun malesialaisen matkustajakoneen ampujien attribuution (ks. Sienkiewicz, 2015). The Center for Strategic and International Studies (CSIS) on esimerkki tutkimusorganisaatiosta, joka tutkii ja julkaisee tietoa Yhdysvaltojen kansalliseen turvallisuuteen liittyvistä asioista (ks. CSIS, 2021). Strategisen tason tiedustelutietoa tuotetaan myös ennusteturvauksissa. Ennusteturnauksien tarkoitus on selvittää mitkä yksilöt, ryhmät tai algoritmit tuottavat tarkimpia ja osuvampia ennusteita erilaisista geopoliittisista aiheista, kuten Kiinan ja Japanin väliset konfliktit tai Venäjän johdon vaihtuminen (Tetlock, Mellers, Rohrbaugh & Chen, 2014).

Tässä tutkielmassa strategisella tiedustelulla tarkoitetaan prosessia, jossa tuotetaan tietoa maailmasta kansallisen ja kansainvälisen tason asioista, ilmiöistä ja tapahtumista suunnittelun ja päätöksenteon tueksi. Strategisella tasolla kuvataan tiedustelun kontekstia. Strategisen tason tiedustelun kohteita ovat

toiset valtiot, laajat ilmiöt ja kansainvälisesti merkittävät asiat ja tietoa tuotetaan kansallisen tason päätöksentekijöille, kuten valtionhallinnolle.

2.2 Tiedusteluprosessi

Eri tiedusteluorganisaatiot mallintavat tiedusteluprosessin eri tavalla (Tropotei, 2018). Usein kuitenkin tiedustelun toimintaa kuvataan syklisen prosessin (kuvio 1) mukaan, johon kuuluu viisi vaihetta: suunnittelu- ja ohjaus, keräys, prosessointi, tuotanto ja analyysi sekä jakaminen (Johnson, 1986; Phythian, 2013). Syklisen tiedusteluprosessin historia ulottuu aina 1700-luvulle saakka Ranskan vallankumouksen aikaan, josta se on jatkanut eri sotilasorganisaatioiden doktriineissa aina tähän päivään saakka (Warner, 2013).



KUVIO 1 Tiedustelusykli (Johnson, 1986). Suomennettu alkuperäisestä kuvasta.

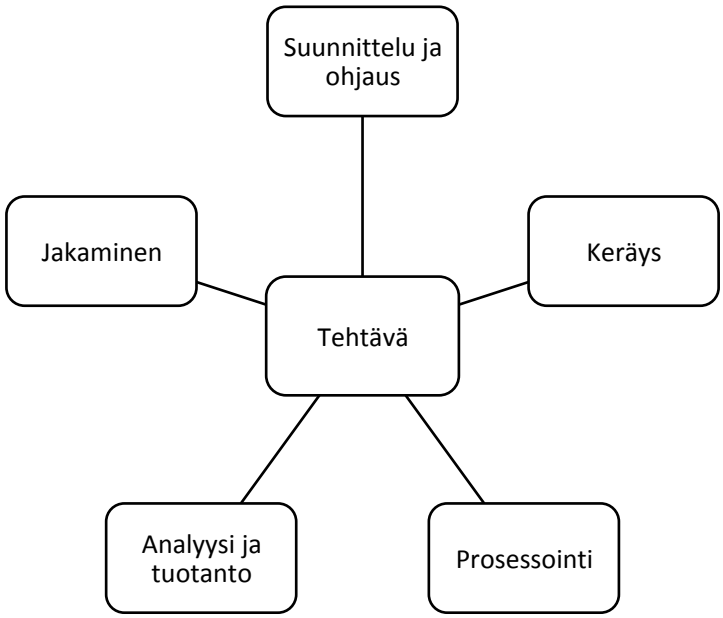
Clark (2020) kuvailee kuvitteellisen skenaarion avulla, miten täydellisessä maailmassa tämä perinteinen tiedusteluprosessi voisi toimia. Prosessi käynnistyy kun asiakas esittää kysymyksen: Miten vakaa on Etiopian valtionhallinto? Seuraavaksi eri keräysmenetelmillä hankitaan tietoa Etiopian valtionhallinnosta: Avointen lähteiden tietoa kerätään mediasta, signaalitiedustelutietoa Etiopian valtionhallinnon viestiliikenteestä ja henkilötiedustelijat kysyvät lähteiltänsä Etiopian sisäpoliittisista suhteista. Kerätyt tiedot prosessoidaan ja yhdistetään aiemmin kerättyyn tietoon, jonka jälkeen aloitetaan analyysin tuottaminen. Analyysi sisältää esimerkiksi henkilöprofiilien muodostamista Etiopian valtionhallinnossa työskentelevistä ihmisistä ja arvioita heidän käyttäytymisestä mahdollisissa tulevissa skenaarioissa. Analyysin jälkeen tiedoista kirjoitetaan raportti, joka toimitetaan tai esitellään asiakkaalle. Lopulta Clark kuitenkin toteaa, että todellisuudessa tiedustelu ei ole aivan näin yksinkertaista. (Clark, 2020, s. 34-39.)

Hulnickin (2006) mukaan syklinen malli ei vastaa todellisuutta, koska tiedustelun asiakkaat harvoin ohjaavat tiedonhankintaa, keräys ja analyysi toimivat omana syklinään ja asiakkaat kaipaavat tiedustelutuotteita jatkuvasti päätöksenteon tueksi ilman monivaiheista ja hidasta tuotantoprosessia, joka käynnistyy heidän ohjauksestaan. Tropotein (2018) mukaan syklisessä prosessissa on perustavanlaatuisia ongelmia: Asiakkaat eivät osaa kysyä oikeita kysymyksiä; Tiedusteluorganisaatio ei voi määritellä mitä tietoa asiakas tarvitsee; Tarkoilla kysymyksillä ja ohjauksella saadaan vain yksipuolisia vastauksia; Asiakkaiden tietotarpeet voivat olla suppeita ja laajoja. Hänen mukaansa yksi prosessi ei sovellu näiden kaikkien eri tarpeiden täyttämiseen. (Tropotei, 2018.)

Phythian (2013) on toimittanut kirjan, jossa tarkastellaan perinteistä tiedustelusykliä kriittisesti strategisen-, sotilas-, rikos-, liiketoiminta- ja kybertiedustelun näkökulmista. Hänen mukaansa perinteinen malli tulee nähdä ennemmin ohjaavana konseptina, kuin tarkkana kuvauksena tiedustelun toiminnasta eri toimintaympäristöissä. Teoksessa on esitetty tiedustelusyklin korvaajaksi mm. verkkoa (Gill & Phythian, 2013), Venn-diagrammia (Richards, 2013), matriisia (Hulnick, 2013) ja monisuuntaista silmukkaa (Davies, Gustafson & Rigden, 2013). Clark (2020) mallintaa tiedusteluprosessin kohdekeskeisesti. Osa tutkijoista ajattelee, että perinteinen sykli on virheellinen, kun taas toiset tulkitsevat sitä väljänä, kuvailevana ja suuntaa antavana viitekehyksenä (Marrin, 2018). Kritiikki kohdistuu suurimmaksi osaksi prosessimallin syklisyyttä kohtaan. Tiedustelun perinteinen prosessimalli kuvaa hyvin tiedustelun organisoitumista ja funktioita, mutta ei todellista vaiheittain etenevää toimintaa, jossa edellinen vaihe antaa syötteen seuraavalle (Johnson, 2010; Clark, 2020).

Yhdysvaltojen asevoimien yleisesikunnan tiedustelua ohjaavassa doktriinissa (US Joint Chiefs of Staff, 2013) tiedusteluprosessi kuvataan kuutena rinnakkaisena toimintona: suunnittelu- ja ohjaus, keräys, käsittely ja hyödyntäminen, analyysi ja tuotanto, jakaminen ja integraatio sekä arviointi ja palaute. Perinteiseen malliin verrattuna, tässä prosessimallissa hyödyntäminen on lisätty käsittelyn rinnalle ja jakamista on täydennetty integraatiolla. Mallissa keskeisenä yhdistävänä tekijänä on tehtävä, eli toiminnan päämäärä. Toiminnot ovat rinnakkaisia ja toisiaan täydentäviä. Esimerkiksi ohjaus ja suunnittelu voi antaa syötteen suoraan analyysille ja tuotannolle, jos ei ole tarpeen kerätä uutta dataa ja prosessoida siitä informaatiota.

Tässä työssä tiedusteluprosessilla tarkoitetaan viittä rinnakkaista toimintoa, joita yhdistää yhteinen tiedustelutehtävä. Prosessi on esitetty kuviossa 2. Prosessi käynnistyy tyypillisesti suunnittelulla ja ohjauksella, mutta se ei välttämättä etene vaiheesta toiseen, vaan joissakin tapauksissa vaiheita (toimintoja) voidaan ohittaa.



KUVIO 2 Tiedusteluprosessi (US Joint Chiefs of Staff, 2013). Mukailtu ja suomennettu alkuperäisestä.

2.3 Tiedustelun suunnittelu ja ohjaus

> We don't know what we don't know – Donald Rumsfeld, former US Secretary of defense

Kentin (1949, s. 151–152) mukaan tiedusteluprosessi voi käynnistyä kahdesta eri lähtökohdasta: Päätöksentekijä esittää tietopyynnön tiedusteluorganisaatiolle tai tiedusteluorganisaatio seuraa systemaattisesti mitä maailmassa tapahtuu ja tämä käynnistää tarpeen raportoida asiakkaille. Johnsonin (2008) mukaan tyypillisesti tiedustelutuotteen tuotantoprosessin käynnistää tiedusteluyhteisö asiakkaan sijasta. Odom (2008) kuvailee artikkelissaan tiedustelun ohjauksen monitahoisuutta. Päätöksentekijät harvoin tietävät mitä tiedustelu voi tuottaa, joten he eivät kykene määrittelemään tietotarpeita tehokkaasti. Tämä tarkoittaa sitä, että tiedusteluanalyytikon pitää ohjata päätöksentekijöitä: mitä tietoa on mahdollista tuottaa ja keneltä voi kysyä ja mitä kysymyksiä. Tiedusteluanalyytikot ovat tiedustelutiedon keskiössä, joten heidän pitää ohjata myös tiedonhankintaa, koska tiedonhankkijat eivät tiedä mitä heidän pitäisi kerätä, jotta tiedusteluanalyytikko voi laatia päätöksentekijälle hänen tarpeitaan palvelevan tuotteen. (Odom, 2008.)

Tiedustelua ohjataan tietotarpeilla, jotka tyypillisesti muotoillaan kysymyksen muotoon. Kysymysten avulla määritellään tiedustelun kohteita. Strategisen tiedustelun kontekstissa, kysymykset koskevat tyypillisesti kansainvälistä turvallisuutta ja uhkia. Tiedustelukysymyksiä voivat olla esimerkiksi: Missä

laajuudessa Al Qaedan solut toimivat Pakistanin valtion alueella ja sieltä käsin (Johnson, 2010)? Kuinka monta strategista ohjusta Neuvostoliitolla on, ja kuinka tarkkoja ne ovat (Treverton & Gabbard, 2008, s. 3-12.)?

Treverton ja Gabbard (2008) kuvailevat, että tiedustelukysymykset voidaan jakaa palapeleihin ja mysteereihin. Palapelit voidaan ratkaista, kun riittävästi kysymykseen liittyvää informaatiota saadaan hankittua ja yhdistettyä. Mysteerit ovat haastavampia ratkaista. Niiden ratkaisu perustuu ajatteluun, ei tiedonhankintaan. Tyypillisesti palapelit liittyvät asioihin ja mysteerit ihmisiin (Treverton & Gabbard, 2008, s. 3-12.) Tiedustelukysymykset määrittelevät mistä kohteista ja lähteistä tietoa kerätään, ja miten analyysi toteutetaan.

Tietotarpeita ja tiedustelun kohteita voidaan hahmottaa myös tarkastelemalla kansainvälisten uhkien määritelmiä. Yhdistyneiden kansakuntien korkean tason paneelin (2004) mukaan mikä tahansa tapahtuma, joka johtaa suuriin määriin kuolemia tai heikentää elämän mahdollisuutta on uhka kansainväliselle turvallisuudelle. He jakavat tapahtumat kuuteen kategoriaan (United Nations, 2004):

- Taloudelliset ja sosiaaliset uhat, kuten köyhyys, tartuntataudit ja ympäristön pilaantuminen
- Valtioiden väliset konfliktit
- Valtion sisäiset konfliktit, kuten sisällissodat, kansanmurhat ja muut mittavat levottomuudet
- Ydin-, kemialliset- ja biologiset aseet
- Terrorismi
- Kansainvälinen järjestäytynyt rikollisuus

Valtiollisessa kontekstissa voidaan hahmottaa kansallisen tiedustelun kohteita tarkastelemalla valtionhallinnon asiakirjoja. Yhdysvalloissa strategisen tiedustelun kohteet listataan ja priorisoidaan uhkaavuutensa mukaan salaiseksi luokiteltuun dokumenttiin (Johnson, 2010). Yhdysvaltojen kansallisen tiedustelun strategiassa (Office of the Director of National Intelligence, 2019) mainitaan perinteiseksi vastustajiksi Venäjä, Kiina, Pohjois-Korea ja Iran. Kehittyviksi uhkiksi kuvataan kyberuhat ja väkivaltaiset ääriryhmät. Suomen ulko- ja turvallisuuspoliittisessa selonteossa (Valtioneuvosto, 2020) kuvaillaan Suomen toimintaympäristöä määrittäviksi ilmiöiksi suurvaltojen keskinäisiä suhteita sekä ilmastonmuutoksen ja pandemioiden kaltaiset globaalit haasteet. Edellä mainitut ilmiöt voidaan ymmärtää strategisen tiedustelun kiinnostuksen kohteiksi.

Suomen tiedustelulaeissa (Laki sotilastiedustelusta 1:4 §; Poliisilaki 5a:3 §) tiedustelun kohteiksi määritellään:

- Terrorismi
- Ulkomainen tiedustelutoiminta
- Suomen maanpuolustukseen kohdistuva tiedustelutoiminta
- Joukkotuhoaseiden suunnittelu, valmistaminen, levittäminen ja käyttö
- Kaksikäyttötuotteet
- Kansanvaltaista yhteiskuntajärjestystä uhkaava toiminta

- Suuren ihmismäärän henkeä tai terveyttä taikka yhteiskunnan elintärkeitä toimintoja uhkaavasta toiminta
- Vieraan valtion toiminta, joka voi aiheuttaa vahinkoa Suomen kansainvälisille suhteille, taloudellisille tai muille elintärkeille eduille
- Vieraan valtion asevoimien ja niihin rinnastuvien järjestäytyneiden joukkojen toiminta ja toiminnan valmistelu
- Kansainvälistä rauhaa ja turvallisuutta uhkaavasta toiminta
- Kriisinhallintaoperaatiota uhkaava toiminta
- Suomen kansainvälisen avun antamisen ja muun kansainvälisen toiminnan turvallisuutta uhkaava toiminta
- Yhteiskuntajärjestystä uhkaava kansainvälinen järjestäytynyt rikollisuus
- Vieraan valtion sotatarvikkeiden kehittäminen ja levittäminen
- Kansainvälistä rauhaa ja turvallisuutta vakavasti uhkaava kriisi
- Kansainvälisten kriisinhallintaoperaatioiden turvallisuutta vakavasti uhkaava toiminta
- Suomen kansainvälisen avun antamisen ja kansainvälisen muun toiminnan turvallisuutta vakavasti uhkaava toiminta.

Tämän tutkielman näkökulmasta on tärkeä ymmärtää, miten tiedustelua ohjataan ja mistä asioista strateginen tiedustelu tuottaa tietoa. Tiedustelua ohjataan kysymyksillä tai määrittelemällä kohteita, joita tiedustelu seuraa. Kysymykset koskevat usein ulkomaita tai merkittäviä organisaatioita, kuten terroristijärjestöjä. Tiedustelun kohde voi olla myös jokin ilmiö, kuten joukkotuhoaseiden valmistaminen. Kansainväliset kriisit, suurvaltojen toiminta sekä niiden väliset keskinäiset suhteet voidaan tunnistaa strategisen tiedustelun keskeisimmiksi kiinnostuksen kohteiksi.

2.4 Tiedustelutiedon kerääminen

Tiedustelutietoa kerätään eri menetelmien avulla, jotka tyypillisesti jaetaan viiteen kategoriaan: geo-, signaali-, henkilö-, mittaus- ja tunnusmerkkitiedusteluun sekä avointen lähteiden tiedusteluun (Johnson, 2010; Lowenthal & Clark, 2015).

Geotiedustelussa (Geospatial Intelligence, GEOINT) tiedustelutietoa tuotetaan geospatiaalisen datan ja kuvien avulla. Tiedon lähde on tyypillisesti kuva, joka otetaan satelliittiin tai lentokoneeseen asennetun sensorin avulla. Sensorit voidaan jakaa sen mukaan, millä sähkömagneettisen spektrin alueella ne toimivat. Sensorit voidaan luokitella passiivisiksi tai aktiivisiksi. Passiivinen sensori vastaanottaa säteilyä, kun taas aktiivinen sensori sekä vastaanottaa että lähettää säteilyä. Elektro-optinen kamera on esimerkki passiivisesta sensorista. Aktiivisia sensoreita ovat esimerkiksi synthetic-aperture radar (SAR) ja light detection and ranging (LIDAR). Aktiivisten sensorien avulla voidaan tuottaa valokuvan kaltaisia representaatioita. (Murdock & Clark, 2015.)

Signaalitiedustelussa (Signals Intelligence, SIGINT) tietoa kerätään sähköisistä lähetteistä. Signaalitiedustelu voidaan jakaa kahteen alalajiin: kommunikaatiotiedusteluun ja elektroniseen mittaustiedusteluun. Kommunikaatiotiedustelussa (Communications Intelligence, COMINT) kerätään sähköisiä lähetteitä, jotka sisältävät ihmisten kommunikaatiota. Näitä lähetteitä ovat esimerkiksi sähkeet, radioliikenne, puhelut ja sähköpostit. Elektronisessa mittaustiedustelussa (Electronic Intelligence, ELINT) kerätään sähköisten laitteiden, kuten tutkien, lähettämiä signaaleja. (Nolte, 2015.)

Henkilötiedustelussa (Human Intelligence, HUMINT) ihmiset keräävät tietoa itse tai toisten ihmisten kautta. Tyypillisesti henkilötiedustelussa erikseen siihen koulutettu henkilö toimii peitetysti ja rekrytoi itselleen tietolähteen, joka hankkii salassa pidettäviä tietoja kohteesta ja toimittaa tiedustelijalle. (Althoff, 2015.)

Mittaus- ja tunnusmerkkitiedustelu (Measurement and Signature Intelligence, MASINT) on teknisesti toteutettua tiedustelua, joka käsittää muuttuvien ja pysyvien kohteiden havaitsemisen, paikantamisen, seuraamisen, identifioimisen ja yksilöllisten ominaisuuksien määrittelemisen. Menetelmä on integroitu muihin keräysmenetelmiin. (Morris & Clark, 2015.)

Nykyaikana tiedustelutietoja kerätään jatkuvasti kehittyvän teknologian avulla. Keräysjärjestelmät ovat automatisoituja ja niitä asennetaan erilaisiin alustoihin maalle, merelle, ilmaan ja avaruuteen. Avaruudessa toimivat satelliitit ovat alustoista yhä kasvavassa roolissa. Huolimatta teknologian kehittymisestä, ihmislähteisiin perustuva henkilötiedustelu on säilyttänyt edelleen tärkeän asemansa. (Dupont, 2003.) Eri keräysmenetelmillä kerätään hyvin monimuotoista dataa: kuvia, tekstiä, ääntä, signaaleja. Tämän tutkielman näkökulmasta on hyvä ymmärtää, että eri keräysmenetelmillä tuotettava data on monimuotoista: kuvia, tekstiä, ääntä, signaaleja, yms. Tiedonlouhinnan soveltaminen eri keräysmenetelmiin vaatii tapauskohtaisen lähestymistavan. Avointen lähteiden tiedustelu on tämän tutkielman keskiössä, joten se kuvataan seuraavassa luvussa muita keräysmenetelmiä tarkemmin.

2.5 Avointen lähteiden tiedustelu

Tämän tutkielman näkökulmasta keskeisin keräysmenetelmä on avointen lähteiden tiedustelu, koska tutkimusaineistona käytetään avoimista lähteistä kerättyä dataa. Lisäksi avointen lähteiden tiedustelu on strategisen tiedustelun kontekstissa yksi merkittävimmistä keräyslajeista. Suurin osa strategisen tason tiedustelutiedosta muodostetaan seuraamalla ja tutkimalla avoimia lähteitä (Kent, 1949; Dupont, 2003; Odom, 2008). Avointen lähteiden tiedustelulla voidaan hankkia tietoa kansainvälisistä turvallisuusuhkista, kuten valtioiden välisistä tai sisäisistä konflikteista, terrorismista, joukkotuhoaseista, rikollisuudesta sekä taloudellisista ja sosiaalisista uhkista (Steele, 2007).

Avointen lähteiden tiedustelun kehitys on alkanut toisen maailmansodan aikaan, jolloin tiedonhankintaa tehtiin ulkomaisista sanomalehdistä ja radiolähetyksistä. Internetin kehityksen myötä avointen lähteiden tiedustelu kehittyi

kuitenkin aivan uudelle tasolle. (Glassman & Kang, 2012.) Avointen lähteiden tiedustelussa tietoa hankitaan lähteistä, jotka ovat kaikille vapaasti ja laillisesti saatavilla, kuten internetistä, kirjoista, dokumenteista tai havainnoimalla. Avointen lähteiden tiedustelu on kuitenkin passiivista, eli aktiivinen toiminta, kuten kommunikointi sosiaalisen median palveluissa luokitellaan henkilötiedusteluksi. (Jardines, 2015.) Muihin keräysmenetelmiin verrattuna avointen lähteiden tiedustelu tarjoaa edullisen, käytännöllisen ja yhteisöllisesti jaettavan vaihtoehdon tiedustelutiedon tuotantoon. Suurin etu on sen avoin luonne, joka mahdollistaa yhteistyön eri organisaatioiden välillä. (Steele, 2007.) Avointen lähteiden tiedustelu on prosessina joustava, läpinäkyvä, käyttäjäystävällinen, avoin ja yhteinen. Keräysmenetelmän kanssa työskentelevien ihmisten tärkeimpiä taitoja ovat tiedon etsintä, lajittelu, luokittelu ja yhdistäminen. (Glassman & Kang, 2012.)

Avointen lähteiden tiedustelua kohtaan on osoitettu paljon kritiikkiä. Odomin (2008) mukaan tiedusteluanalyytikot kutsuvat usein avointen lähteiden kautta hankittuja tietoja rinnakkaisiksi tiedustelutiedoiksi. He eivät usein pidä avointen lähteiden tiedustelua todellisena tiedusteluna, eivätkä osaa hyödyntää sitä analyysi- ja tuotantotyössään. (Odom, 2008.) Hulnick (2010) pohtii artikkelissaan, että voidaanko avointen lähteiden tiedustelua pitää edes tiedusteluna ja mikä on sen merkitys Yhdysvaltojen tiedusteluyhteisölle. Hänen mukaansa avointen lähteiden tiedustelu on kärsinyt arvonlaskua vuoden 2001 syyskuun 11. päivän iskujen jälkeen. Sitä on pidetty tarpeettomana kansallisen tason tiedustelulle, koska yleisesti siinä hyödynnettävien lähteiden luotettavuus on heikompi kuin muissa keräysmenetelmissä. Hän päätyy kuitenkin johtopäätökseen, että avointen lähteiden kautta hankittu tieto oikein tulkittuna voi olla yhtä hyödyllistä, kuin muidenkin menetelmien kautta hankittu tieto. (Hulnick, 2010.) Mercado (2009) pohtii artikkelissaan, miksi avointen lähteiden tiedustelua ei hyväksytä muiden tiedonhankintamenetelmien rinnalle. Hänen sen potentiaalia ei osata hyödyntää ja tämä johtaa hyväksynnän puutteeseen. Tehokas avointen lähteiden tiedustelu vaatii vieraiden kielten osaamista ja ymmärrystä vieraista valtioista ja niiden mediasta. Potentiaalin hyödyntämiseksi avointen lähteiden tiedusteluun tulee kohdentaa henkilöstöresursseja vastaavasti kuin muihinkin keräysmenetelmiin. Hajaantuneet avointen lähteiden tiedustelua tekevät organisaatiot tulee keskittää yhtenäiseen ohjaukseen. Avointen lähteiden tiedustelu vaatii omat järjestelmät tiedon etsimiseen, kääntämiseen ja siirtämiseen. (Mercado, 2009.)

Dupont (2003) esittää, että tiedusteluyhteisöjen tulee lopettaa avointen lähteiden tiedustelun merkityksen kyseenalaistaminen ja keskittyä pohtimaan, miten avointen lähteiden tiedustelu saadaan integroitua osaksi muuta tiedustelujärjestelmää. Odomin (2008) mukaan tiedusteluyhteisö, etenkin strategisella tasolla voi oppia paljon uutistoimitusten raportoinnista ja työprosesseista. Hänen mukaansa uutistoimitukset toimivat hyvin samankaltaisesti kuin tiedusteluorganisaatiot, mutta ne osaavat hyödyntää teknologiaa paremmin kerätessään, analysoidessaan ja jakaessaan avointen lähteiden kautta kerättyä tietoa (Odom, 2008). Teknologian hyödyntäminen on keskeistä avointen lähteiden tiedustelun tehokkaassa hyödyntämisessä. Layton ja Watters (2015) tarjoavat

kirjassaan käytännöllisiä ratkaisuja avointen lähteiden tiedustelun automatisointiin.

Yhteenvetona, avointen lähteiden tiedustelussa kerätään dataa julkisesti saatavilla olevista lähteistä, kuten perinteisestä tai sosiaalisesta mediasta. Avointen lähteiden tiedustelun kautta kerätty data on hyvin monipuolista: uutistekstejä, twiittejä, raportteja, videota, kuvia tai äänitteitä. Data on tyypillisesti ihmisen ymmärrettävässä muodossa, jolloin se ei vaadi erityistä prosessointia. Avointen lähteiden tiedustelun luotettavuutta ja hyödyllisyyttä vähätellään, mutta panostamalla keräysmenetelmään vastaavalla tavalla kuin muihinkin menetelmiin voidaan sen hyödyllisyyttä ja arvostusta nostaa. Merkittävä osa strategisen tason tiedustelutiedosta voidaan kerätä avoimista lähteistä.

2.6 Prosessointi

> I have three major problems: processing, processing, processing. – Admiral Mike McConnell, former director of the National Security Agency

Tiedustelujärjestelmän kyky kerätä dataa on kehittynyt kylmän sodan jälkeen merkittävästi. Kerättyä dataa tai informaatiota on niin paljon, että ihminen ei ehdi käsittelemään ja tuottamaan siitä merkityksellistä tietoa. Tätä ongelmaa kutsutaan informaatiotulvaksi. (Dupont, 2003; Mandel, 2019.) Informaatiotulvan hallitsemiseksi data tulee tulkita, lajitella, luokitella ja tallentaa yhteisesti tunnistettuun formaattiin, jotta sen hyödyntäminen on tehokkaampaa. Tätä kokonaisuutta kutsutaan prosessoinniksi.

Prosessointi tarkoittaa mahdollisesti tiedustelutiedoksi soveltuvan datan kehittämistä käytettävään muotoon analyysiin ja tiedustelutuotteisiin (Goldman, 2011). Kerätty data, kuten puhelu, kuva, uutinen tai dokumentti käännetään, tulkitaan tai puretaan luettavaan muotoon (Johnson, 2010).

Tehokas useita lähteitä yhdistävä analyysi edellyttää, että informaatio on standardoitu, integroitu sekä sille on määritelty paikkatieto (Steele, 2007). Yhdysvaltojen kansallisen tiedustelun johtajan James Clapperin (2016) mukaan tiedustelutiedot voidaan organisoida sen mukaan mitä tutkitaan: henkilöä, paikkaa tai asiaa. Johnston, Wright, Bice, Almendarez ja Creekmore (2015) esittävät, että tiedustelutietojen hallintaa voidaan kehittää objektipohjaisen tuotannon avulla, jossa muodostetaan kiinnostuksen kohteista objekteja, joihin sitoen tuotetaan tiedustelutietoa. Heidän mukaansa objekti voi olla henkilö, paikka tai asia, ja siihen voidaan sitoa aiheita, tapahtumia tai toimintaa. Objektipohjaisen tuotannon tarkoitus on yhtenäistää eri tiedustelutoimijoiden käsitteistö samasta aiheesta, organisoida ja tallentaa hankitut tiedot objekteihin sitoen, jotta analyysiä tehtäessä tieto on paremmin löydettävissä (Johnston ym., 2015). Osana objektipohjaista tuotantoa on kehitetty myös aktiivisuuskeskeisen tiedustelun konseptia. Sen tarkoituksena on ymmärtää toiminnan kautta tunnettujen ja tuntemattomien objektien toimintaa. Konsepti edellyttää, että jokaisella tiedustelutiedolla on georeferenssi, eli ne sidotaan paikkaan ja aikaan, sekä tiedot integroidaan heti keräyksen jälkeen ennen analyysiä. Laadukkaasti prosessoidusta

informaatiosta voidaan löytää merkityksellistä tiedustelutietoa ilman erillistä ohjausta tiedustelukysymyksen muodossa. (Atwood, 2015; Biltgen & Ryan, 2016.)

Eräänlaisena ideaalina, Dupont (2003) esittää ajatuksen holistisesta järjestelmästä, johon kaikki tiedustelutieto, julkinen ja salaiseksi luokiteltu, signaalitiedustelulla siepatut kommunikaatiot, satelliitilla otetut kuvat sekä valmiit raportit ovat tallennettu ja josta ne ovat sekä tiedusteluyhteisön että päätöksentekijöiden käytössä. Teknologia mahdollistaa kaiken tämän, ainakin tulevaisuudessa, mutta siihen liittyy paljon institutionaalisia ongelmia. (Dupont, 2003.)

Yksi keskeinen tiedusteluprosessin toiminnoista on tiedon luotettavuuden arviointi. Informaation luotettavuuden arviointi voidaan jakaa lähteen luotettavuuden arviointiin ja informaation sisällön uskottavuuden arviointiin. Heidän mukaansa yksinkertaisempi ja parempi tapa on arvioida informaation luotettavuutta yhdellä todennäköisyyttä kuvaavalla arvolla eri tasoilla. Esimerkiksi tiedon kerääjä arvioi, että informaatio on 60 % todennäköisyydellä tarkkaa ja analyytikon arvion mukaan tarkkuus on 80 %. Arvioiden seurauksena informaation luotettavuus on 70 %. (Irwin & Mandel, 2019.)

Yhteenvetona, prosessointi tiedustelun kontekstissa tarkoittaa seuraavia toimintoja:

- Datan tulkinta tai muuttaminen tekstiksi yhteisesti ymmärrettävälle kielelle – informaatioksi
- Lajitellaan ja luokitellaan informaatio yhteisen standardin mukaan: aika, paikka, henkilö, asia, aihe, tapahtuma, toiminta
- Arvioidaan luotettavuus
- Tallennetaan ja yhdistetään aiemmin prosessoituun informaatioon

2.7 Tiedusteluanalyysi

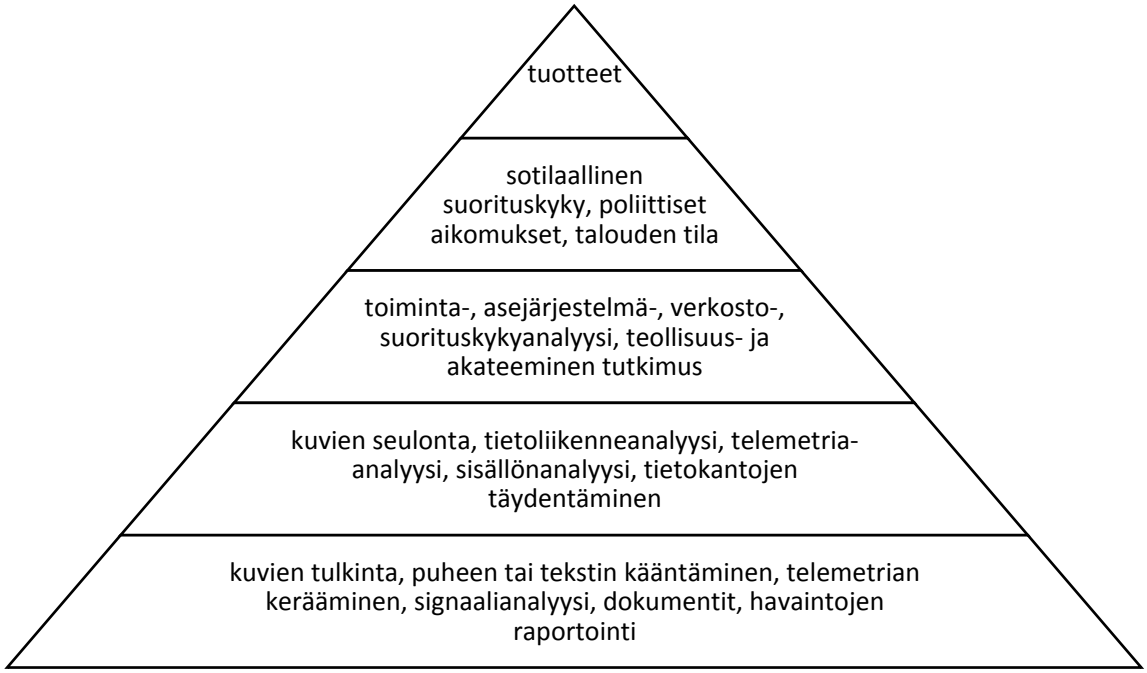
> Do you like doing jigsaw puzzles if you don't know the picture, you only have a quarter of the pieces and the president wants to know what the picture is in five minutes because he needs to make a consequential decision? – Sue Gordon, former principal executive of National Intelligence

Tiedusteluanalyysissä pyritään löytämään vastaus ongelmaan vaiheittaisen prosessin kautta (Kang & Stasko, 2011). Brucen ja Georgen (2014) mukaan tiedusteluanalyysi on tiedusteluprosessin osuus, jossa ajatellaan. Siinä seurataan tärkeitä valtioita, trendejä, ihmisiä, tapahtumia, ilmiöitä, ja tunnistetaan kaavoja tai poikkeamia sekä tulkitaan mennyttä ja tehdään arvioita tulevasta. Analyysin tuotoksena syntyy arvioita, ennusteita ja katsauksia: Arvio on epätäydellisestä tai epävarmasta informaatiosta analysoitu johtopäätös, ennuste on tulevaisuudesta tehty arvio ja katsaus on uudenlainen näkökulma johonkin asiaan tai ongelmaan (Bruce & George, 2014). Tiedusteluanalyysi ei pyri tuottamaan tieteellistä totuutta, joten sen logiikkaa on erilainen kuin tieteellisessä tutkimuksessa.

Kuten tiedustelulle, ei tiedusteluanalyysillekään ole muodostettu yhtä yhtenäistä käsitettä. Tiedusteluanalyysi on erittäin monimuotoista (Treverton & Gabbard, 2008; Marrin, 2011; Bruce & George, 2014). Monimuotoisuutta voidaan havainnollistaa lähteiden, analysoitavan kohteen tai ilmiön sekä toiminnan tason mukaan. Marrinin (2011) mukaan tiedusteluanalyysin eri muotoja ovat:

- Yksilähdeanalyysi, jossa analysoidaan yhden keräysmenetelmän tuottamaa dataa tai informaatiota. Esimerkkinä signaalianalyysi, joka on hyvin teknistä.
- Monilähdeanalyysi, jossa yhdistetään ja analysoidaan eri tiedonhankintamenetelmillä kerättyä tietoa.
- Kohdeanalyysi, jossa analysoidaan esimerkiksi yhden maantieteellisen kohteen tapahtumia. Tämänkaltaisen analyysin keskiössä on analyytikko, jolla on paljon substanssitietoa analysoitavasta kohteesta.
- Toimintoanalyysi, jossa analysoidaan jotakin tiettyä toimintaa, kuten esimerkiksi taloutta, johtamista tai asevoimia. Tämänkaltainen analyysi on hyvin vastaavaa kuin kohdeanalyysi, mutta analyytikolla on kohteen sijaan tietopohja analysoitavasta toiminnasta.
- Taktinen analyysi, jossa pyritään vastaamaan faktoilla kysymyksiin: kuka, mitä, missä ja milloin.
- Strateginen analyysi, jossa pyritään hahmottamaan laajoja kysymyksiä, monimutkaisia konsepteja sekä abstraktioita.

Tiedusteluanalyysin monitasoisuutta voidaan havainnollistaa pyramidihierarkian avulla (kuvio 3). Alimmalla tasolla eri tiedonhankintamenetelmillä kerätty data tulkitaan sellaiseen muotoon, että sitä voidaan hyödyntää analyysissä. Seuraavaksi, esianalyysissä tallennetusta informaatiosta seulotaan sellaiset tiedot, jotka tunnistetaan oleelliseksi jatkoanalyysin kannalta. Seulonnan ja informaation etsimisen jälkeen yhdistetään eri keräysmenetelmillä tuotettua informaatio. Tällä tasolla muodostuu tietoa toiminnasta, järjestelmistä ja organisaatioista. Seuraavalla tasolla tuotettu tieto yhdistetään laajempaan tietopohjaan ja käsitykseen maailmasta. Lopulta tieto muotoillaan tuotteeksi, joka palvelee päätöksentekijän tarpeita. (Treverton & Gabbard, 2008.)



KUVIO 3 Tiedusteluanalyysin tasot (Treverton & Gabbard, 2008). Mukailtu ja suomennettu alkuperäisestä.

Tiedusteluanalyysin suorituskyky muodostuu ihmisistä ja työvälineistä, joita he hyödyntävät (Treverton & Gabbard, 2008). Vaikka tiedusteluanalyysistä voidaan tunnistaa erilaisia funktioita, sisältyy kaikkien analyytikkojen työhön tutkimista, lukemista, ajattelua, kirjoittamista ja esittelyä (Marrin, 2011). Erilaiset analyysityövälineet ja teknologia tarjoavat hyötyjä alemmilla tasoilla, mutta mitä ylemmäksi hierarkiassa noustaan, korostuu ihmisen ajattelu ja analyyttiset kyvyt (Treverton & Gabbard, 2008).

Tiedusteluanalyytikko

Brucen ja Georgen (2014) mukaan täydellisessä tiedusteluanalyytikossa yhdistyy historioitsijan, journalistin, tutkimusmenetelmien asiantuntijan, tiedonhankinnan johtajan ja skeptikon tiedot ja taidot. Tiedusteluanalyytikon tulee hallita jokin tietty substanssi ja ymmärtää sen merkityksen osana kansainvälistä politiikkaa, osata hankkia ja analysoida laadullista tietoa avoimista lähteistä, ymmärtää salaisten tiedonhankintamenetelmien suorituskyvyt, ymmärtää miten kognitiivinen psykologia vaikuttaa omaan toimintaan, tuntee salaamisen ja harhauttamisen periaatteet, oppia käytännöstä sekä osata tehdä yhteistyötä. (Bruce & George, 2014.) Tiedusteluanalyytikot ja heidän tehtävänsä ovat hyvin erilaisia. Osa heistä on erikoistunut syvällisesti vain yhteen kapeaan aiheeseen ja osa kapeasti moniin eri aiheisiin, vallitsevan tilanteen ja priorisoinnin mukaisesti. Osa analyytikoista on asiantuntijoita, eli he tuntevat jonkin tai jotkin asiat syvällisesti, ja ovat tutkineet kyseistä substanssia pitkään. Osa analyytikoista omaa tietoa ja taitoa asioiden yhdistelyssä ja he kykenevät tuottamaan informaatiota yhdistelemällä vastauksen monimutkaisiinkin ongelmiin. Joihinkin tehtäviin voidaan hankkia osaamista tiedusteluyhteisön ulkopuolelta esimer-

kiksi opiskelemalla substanssia, kuten kansainvälistä politiikkaa tai analyyttistä taitoa, kuten tilastotiedettä, mutta joissakin tehtävissä oppiminen tapahtuu kokemuksen ja työn kautta tiedusteluyhteisön sisällä. (Treverton & Gabbard, 2008.)

Työvälineet

Tiedusteluanalyysissä hyödynnettävillä työvälineillä tarkoitetaan teknologiaa, tuotteita ja prosesseja, jotka tukevat analyytikon työtä kolmella tapaa: Ne helpottavat informaation ja tiedon löytämistä datasta, mahdollistavat hypoteesien kehittelyn ja testaamisen sekä helpottavat kommunikaatiota keräyksen ja asiakkaiden kanssa (Treverton & Gabbard, 2008). Phersonin ja Heuerin (2014) mukaan tiedusteluanalyysiä voidaan tehdä neljällä tavalla: asiantuntija-arviona, rakenteellisena analyysinä, lähes kvantitatiivisesti tai empiirisenä analyysinä. Asiantuntija-arviota kutsutaan myös perinteiseksi tavaksi tehdä analyysiä. Se perustuu tiedusteluanalyytikon itsenäiseen ajatteluun ja intuitioon. Dhami, Mandel, Mellers ja Tetlock (2015) esittävät, että tiedusteluanalyysi nojaa voimakkaasti asiantuntijuuteen, jonka ongelmana on liiallinen itsevarmuus, joka johtaa epätarkkoihin arvioihin. Heidän mukaansa tiedusteluanalyysin tarkkuutta ja laatua voidaan parantaa, jos ennusteiden toteutumista mitataan ja pelkän asiantuntijuuden lisäksi tiedusteluyhteisö tukeutuisi analyysityössään tieteellisiin menetelmiin, kuten tilastotieteeseen datan käsittelyssä ja käyttäytymistieteeseen muodostaessaan tiimejä.

Ihmisen kognitiivista rajoittuneisuutta tehdä arvioita epävarmoissa tilanteissa ovat tutkineet mm. Tversky ja Kahneman (1974). Yhdysvaltain keskustiedustelupalvelun analyytikko Richards J. Heuer, Jr. (2009) kiinnostui Tverskyn ja Kahnemanin innoittamana kognitiivisesta psykologiasta ja havaitsi, että tiedusteluanalyysissä ei huomioitu riittävästi ihmisen psykologiaa ja kognitiivisia vinoumia. Ratkaisuksi hän kehitti rakenteellisia menetelmiä tiedusteluanalyysin tarpeisiin (Heuer, 2009). Suurin osa ihmisen kognitiivista rajoitteisuutta helpottavista työvälineistä perustuu kahteen periaatteeseen: asian purkamiseen pienempiin osiin ja visualisointiin, eli asian ulkoistamiseen ihmisen mielestä paperille, seinälle tai tietokoneen ruudulle (Pherson & Heuer, 2014.)

Pherson ja Heuer (2020) esittelevät kirjassaan ”Structured analytic techniques for intelligence analysis” viisikymmentä tiedusteluanalyysiin soveltuvaa rakenteellista menetelmää. Ne voidaan organisoida kahdeksaan kategoriaan (Pherson & Heuer, 2014):

- purkaminen ja visualisointi
- Ideointitekniikat
- Skenaariot ja indikaattorit
- Hypoteesien kehittäminen ja testaaminen
- Syyn ja seurauksen arviointi
- Analyysin haastaminen
- Konfliktien hallinta
- Päätöksentekoa tukeva analyysi

Useimmin hyödynnettyjä tekniikoita ovat rakenteellinen aivoriihi, avainoletusten tarkastaminen, kilpailevien hypoteesien menetelmä, indikaattorit ja entä-jos analyysi (Pherson & Heuer 2014). Myös Chin, Kuchar ja Wolf (2009) ovat havainneet, että tiedusteluanalyytikot hyödyntävät usein kilpailevien hypoteesien menetelmää. Tiedusteluanalyytikot työskentelevät usein aikapaineessa epäselvän, epätäydellisen ja harhauttavan informaation kanssa, jonka vuoksi tiedusteluanalyysi on erittäin altis virheille (Pherson & Heuer 2014). Tiedusteluanalyytikot kuitenkin hylkäävät rakenteellisten menetelmien hyödyntämisen usein aikapaineen vuoksi (Chin ym., 2009). Osa rakenteellisista menetelmistä vaativat vain vähän aikaa, ja siitä huolimatta ne parantavat analyysin laatua (Pherson & Heuer 2014).

Pirolli ja Card (2005) ovat tutkineet tiedusteluanalyytikkojen ajatteluprosessia. Heidän mukaansa prosessi voidaan jakaa karkeasti kahteen rinnakkaiseen osaprosessiin: etsintään ja ymmärtämiseen. Etsintäprosessissa haetaan, suodatetaan, luetaan ja poimitaan informaatiota. Ymmärtämisen prosessissa muodostetaan konsepti, joka tukee todistusaineistoa. Prosessi voi käynnistyä konseptista, joka kertoo tarinan, ja johon etsitään tukevaa informaatiota. Prosessi voi myös käynnistyä informaatiosta, joka käynnistää tarpeen muodostaa tarina ja tiedustelutuote. (Pirolli & Card, 2005.) Kang ja Stasko (2011) ovat selvittäneet tiedusteluanalyytikoiden työprosessia. Heidän mukaansa ensimmäinen vaihe on konseptin muodostaminen, jota seuraa tiedonhankinta, analyysi ja tuotanto. He tunnistivat myös kaksi erilaista tapaa tehdä analyysiä: intuitiivisesti ja rakenteellisesti. (Kang & Stasko, 2011.) Chin ym. (2009) ovat myös selvittäneet tiedusteluanalyysin työprosessia, jotta voidaan kehittää parempia työvälineitä ja teknologiaa tukemaan analyytikoiden työtä. Heidän mukaansa, ensimmäinen vaihe analyysiprosessissa on tiedon kerääminen tai kokoaminen, jonka jälkeen tietoihin tutustutaan, ne lajitellaan ja luokitellaan sekä niitä suodatetaan. Tämän jälkeen analyytikot pyrkivät muodostamaan aineistosta kaavoja ja trendejä. Analyytikot hyödynsivät työssään erilaisia työvälineitä. Jotkut piirsivät paperille tai seinälle, kun taas toiset pitivät taulukkolaskentaohjelmaa perustyövälineenä. Analyytikot tuottivat muun muassa graafeja, aikajanoja tai huomautuskynällä merkittyjä kalentereita tai karttoja kuvaamaan löytämiänsä kaavoja (Chin, Kuchar & Wolf, 2009). Esimerkkejä työvälineistä, joita analyytikot käyttävät työnsä tukena ovat: Wikispaces, Google Sites, Mindmeister, Zotero, Dax Norman Trust Scale ja Analyst Notebook (Kang & Stasko, 2011).

Dupontin (2003) mukaan nykyaikana tiedusteluanalyytikko työskentelee päätelaitteen avulla, josta hänellä on pääsy tietojärjestelmään, joka sisältää integroitua eri tiedonhankintamenetelmin hankittua tietoa. Teknologian kehitys ja älykkäät järjestelmät ovat vaikuttaneet merkittävämmin tiedustelutietojen hankintaan ja prosessointiin. Vaikka tiedusteluanalyytikot hyödyntävätkin tietokoneita työssään, tiedusteluanalyysi tehdään edelleen ihmismielen avulla (Dupont, 2003). Eldridge ym. (2017) ovat tunnistaneet, että teknologian hyödyntäminen tiedusteluanalyysissä on varsin vähän tutkittu alue. He esittelevät yhteisen kognitiivisen järjestelmän konseptin, jonka mukaan tietojärjestelmä tulee nähdä ihmisten ja tietokoneiden kokonaisuutena. Tehokkain tapa kehittää järjestelmiä tiedusteluanalyysin tueksi on eliminoida ihmisten heikkouksia koneiden vahvuuksilla ja koneiden heikkouksia ihmisten vahvuuksilla. Ohjelmistot

eivät ainakaan toistaiseksi kykene korvaavaan ihmisen ajattelua, joten tietokoneet eivät tule ainakaan lähitulevaisuudessa korvaamaan tiedusteluanalyytikon työtä. Tietokoneita pitää kuitenkin osata hyödyntää ihmisten tekemän tiedustelun tukena. (Eldridge ym., 2017.)

Tulevaisuuden ennustaminen

Päätöksenteon kannalta on hyödyllistä tietää päätöksentekoon vaikuttavat asiat mahdollisimman nopeasti ja ennemmin etukäteen, joten tulevien tapahtumien ennustamisen voidaan tulkita olevan tiedustelun keskeisiä toimintoja.

Tetlock ja Gardner (2016) kuvailevat kirjassaan ”Superforecasting: The Art and science of prediction” tulevaisuuden ennustamisen mahdollisuuksia ja haasteita. Heidän mukaansa tulevaisuus on äärettömän epävarma, koska tapahtumat eivät muodostu normaalijakaumaa noudattaen. Tulevaisuutta voidaan kuitenkin ennustaa ja arvioida, mutta se vaatii oikeanlaisia ihmisiä ja oikeanlaisia työtapoja. (Tetlock & Gardner, 2016.) Clark (2020) esittelee kirjassaan ennustamisen kolme muotoa: ekstrapoloinnin, projektoinnin ja ennalta arvioinnin. Kaksi ensimmäistä on tilastotieteeseen perustuvia menetelmiä ja kolmas intuitiivinen, ihmisen ajatteluun perustuva menetelmä.

Jotkin tapahtumat ovat kuitenkin mahdottomia ennustaa. Taleb (2007) esittää kirjassaan teorian mustasta joutsenesta. Hänen mukaansa ”musta joutsen” on metafora yllätykselliselle tapahtumalle, jolla on merkittäviä vaikutuksia ja sitä yritetään jälkikäteen virheellisesti järkeistää. Miller (2014) esittää, että tapahtumien ennustaminen vaati toistuvia ilmentymiä ja selkeitä kaavoja. Ongelmallista on, että kaikki tapahtumat eivät noudata historiallista jatkumoa, joka mahdollistaisi ennustamisen. Koska tulevaisuus on epälineaarinen ja kaoottinen, yllätyksellisten tapahtumien tunnistamisen suhteen on merkityksellisempää etsiä poikkeamia jatkumoissa, trendien ja kaavojen mallintamisen sijaan (Miller, 2014). Leskovec, Rajaman ja Ullman (2020 s. 5–6) esittävät, että tilastotieteellä on kuitenkin rajansa poikkeavien tapahtumien tunnistamisen suhteen. Tätä rajallisuutta kuvaa heidän esittelemä Bonferronin periaate, jonka mukaan sattumanvaraisesta datasta voidaan löytää osumia erilaisiin asioihin, ja jos datan määrä kasvaa, kasvaa myös virheellisten löydösten määrä. He käyttävät esimerkkinä vuoden 2001 syyskuun 11. päivän terrori-iskuja, joiden ennakkomerkeistä oli olemassa dataa, mutta tiedon tunnistamista ei kyetty tekemään.

Tulevaisuutta voidaan ennustaa, mutta usein merkittävät tapahtumat ovat mahdottomia ennustaa. Ennustamisen sijaan, tiedustelun näkökulmasta merkityksellisempää on tunnistaa poikkeamia historiallisista jatkumoista mahdollisimman nopeasti.

Väärinkäsityksiä

Odom (2008) selvittää artikkelissaan tiedusteluanalyysiä ja siihen liittyviä väärinkäsityksiä: Tiedusteluanalyytikot työskentelevät kansallisella tasolla eri organisaatiossa kuin keräys tai päätöksentekijät. Tiedusteluanalyytikot eivät näin ollen ole mukana päätöksentekoprosessissa, jonka vuoksi heille ei ole välttämättä käsitystä, mistä päätöksentekijät ovat kiinnostuneita tai mitä väärinkäsi-

tyksiä heillä on; Tiedusteluanalyytikon tärkein rooli on tehdä asiakkaat tietoiseksi tiedustelutiedosta, joita he eivät tiedä tai eivät halua tietää, mutta josta heidän pitäisi tietää; Tiedusteluanalyysiä ei voida automatisoida ja toteuttaa ainoastaan tietokoneiden avulla, etenkin tuotteiden laatiminen vaatii ihmisen työtä; Tiedusteluanalyysissä voidaan arvioida vastustajan toimintamahdollisuuksia arvioimalla sen tahtotilaa tai suorituskykyä. Suorituskyvyn arviointi on helpompaa, koska se perustuu faktoihin, kuten asevoimien organisaatioon ja vahvuuteen. Tahtotilan arviointi on haastavampaa ja siihen liittyy paljon epävarmuuksia, koska se perustuu ihmisen vapaaseen tahtoon. Tiedusteluanalyysin tavoite on yllätysten välttäminen tätä epävarmuutta vähentämällä. Vastustajan tahtotilaa voidaan selvittää tutkimalla sen johtajien ja päätöstentekijöiden rakenteellisia rajoitteita, jotka ovat usein sidottuja instituutioihin ja organisaatioihin, joiden toimintaa voidaan selvittää historian kautta. Miten instituutiot ja organisaatiot ovat toimineet aiemmin; Ei ole olemassa yhtä tapaa tehdä tiedusteluanalyysiä, vaan tiedusteluanalyysin tekniikat ja menetelmät riippuvat tiedonhankintamenetelmästä, kenelle tietoa tuotetaan ja missä (organisaatiossa) tietoa tuotetaan; Tiedusteluanalyysin tarkoitus ei ole totuuden selvittäminen, vaan epävarmuuden vähentäminen. Tiedusteluanalyysi ei kansallisella tasolla ole kiinteä osa päätöksenteon organisaatiota, joten kommunikaatio tiedusteluorganisaation ja päätöksentekijöiden välillä on tärkeää. (Odom, 2008.)

Tässä tutkielmassa tiedusteluanalyysilla tarkoitetaan tiedusteluprosessin vaihetta, jossa informaatiosta tuotetaan tietoa. Tiedusteluanalyysia tehdään monella eri tavalla ja tasoilla. Lopullisena tavoitteena on tuottaa raportti tai esitys päätöksen tueksi. Tiedusteluanalyysin erityispiirteitä ovat aikapaine ja hajanainen informaatio. Vaikka analyysi perustuukin vahvasti ihmisen ajatteluun, voidaan sitä tukea informaatioteknologian avulla purkamalla, kokoamalla ja visualisoimalla informaatiota.

2.8 Tiedustelutiedon jakaminen

> Well, you warned me, but you didn't convince me. – Henry Kissinger, National security Advicer

Tiedustelutuote on valmis tiedustelutieto, joka jaetaan niille, jotka tarvitsevat sitä päätöksentekoon (Goldman 2011, s. 215). Esimerkiksi talouden päätöksentekijöille jaetaan ennusteita öljyn ja energian trendeistä, ulkopolitiikan edustajille henkilötietoja ulkomaiden vertaisista ja sotilaille potentiaalisen vastustajan asejärjestelmistä tai organisaatiosta (Dupont, 2003). Tiedustelutuotteita voidaan jakaa "työntämällä" tai "vetämällä". Perinteisessä, työntävässä jakelussa tiedusteluanalyytikko tekee päätöksen, mitä tietoa hän olettaa päätöksentekijän tarvitsevan, ja jakaa tiedon päätöksentekijälle. Vetävässä jakelussa päätöksentekijällä on pääsy tiedustelutietoja sisältävään tietojärjestelmään, josta hän itse käy lukemassa sinne tuotettuja tietoja. (Sharfman, 1995; Dupont, 2003.)

Tiedustelutuotteet eroavat rakenteeltaan ja sisällöltään akateemisista teksteistä seuraavin tavoin (Major, 2014, s. 21–23):

- Ne keskittyvät tulevaisuuteen
- Raporttien lukijat eivät ole asiantuntijoita
- Teksti on yleistävää ja kuvailevaa
- Raportti alkaa johtopäätöksillä ja päättyy siihen mitä niistä seuraa.

Tiedustelutuotteet jaetaan usein kolmeen kategoriaan (Kent, 1949; Major, 2014; Clark, 2020):

- Perustieto (basic intelligence)
- Tilannetieto (current intelligence)
- Arvioiva tieto (estimative and predictive intelligence)

Hulnick (2006) kuvailee artikkelissaan tiedustelutuotteita vastaavalla tavalla, mutta hän lisää ennakkovaroituksen neljänneksi tuotteeksi täydentämään perinteistä kolmijakoa. Ennakkovaroituksella tarkoitetaan tuotetta, joka nimensä mukaisesti varoittaa mahdollisesta vihollisen hyökkäyksestä tai väkivaltaisten toimien mahdollisuudesta (Goldman 2011, s. 266). Ennakkovaroitustiedustelu on oma lähestymistapa, jolla koko tiedustelun kokonaisuutta voidaan jäsentää (ks. Grabo, 2010; Gentryn & Gordonin, 2019).

Perustietoa voidaan kuvailla myös tausta- tai kirjastotiedoksi. Se sisältää tietoa esimerkiksi tietyn hävittäjätyypin teknisistä ominaisuuksista (Major, 2014) tai merkittävien organisaatioiden, kuten valtioiden, historiasta, hallinnosta, taloudesta, asevoimista, energiasta, maantieteestä (CIA, 2021). Perustietoa sisältävästä tuotteesta hyvä esimerkki on CIA:n tuottama ”The World Factbook” (ks. CIA, 2021).

Yhdysvaltain tiedusteluyhteisö tuottaa kansallisen tason päätöksentekijöille kahta tunnettua tuotetta, jotka kuvaavat hyvin tilanne- ja arvioivaa tietoa. Nämä ovat: päivittäistä, tilannetietoa sisältävää Presidential daily brief (PDB) ja pidemmän aikavälin, arvioivaa ja laajaa tietoa sisältävä National intelligence estimate (NIE). PDB tuotetaan kuutena päivänä viikossa ja NIE keskimäärin kaksikymmentä kertaa vuodessa. (Johnson, 2008.)

PDB koostuu tärkeimmistä ja arkaluonteisista asioista, jotka esitellään ja lähetetään Yhdysvaltain presidentille, valikoiduille johtajille sekä Valkoisen talon avustajille (Goldman 2011, s. 213). Sen tarkoitus on kertoa päätöksentekijöille, mitä maailmalla tapahtui edellisen 24 tunnin aikana ja mitä odotetaan tapahtuvan seuraavan 24 tunnin aikana. Tuote on sisältänyt tietoa mm. terroriiskusta Tokion metrossa, Kiinan ja Pakistanin ohjuskaupoista, Burundin levottomuuksista ja Neuvostoliiton ydinohjusten siirrosta Kuubaan. (Johnson, 2008.)

NIE sisältää joko ennusteen jostakin talous- tai turvallisuuspolitiikkaa koskettavasta tilanteesta tai arvion ulkomaiden suorituskyvyistä, haavoittuvuuksista ja toimintavaihtoehdoista (Goldman 2011, s. 187). Tarve aloittaa tiedustelutuotteen tuotanto voi käynnistyä asiakkaan kysymyksestä, mutta useimmin tiedusteluyhteisö esittää aiheita (Johnson, 2008). Yhdysvaltojen kansallinen tiedustelu julkaisi vuonna 2007 julkisen version Iranin ydinaseohjelmaa käsittelevästä National intelligence estimatesta (Office of the Director of National Intelligence, 2007). Se on hyvä esimerkki kyseisen tuotteen sisällöstä ja analyysin laadusta. Raportissa arvioidaan, että Iranin ydinaseohjelma on lop-

punut vuonna 2003 ja kohtalaisella varmuudella sitä ei ole käynnistetty uudestaan vuoden 2007 puoliväliin mennessä. Raportissa arvioidaan myös, että Iran ei erittäin todennäköisesti kykene tuottamaan ydinaseiden valmistamiseen soveltuvaa uraania ennen vuotta 2009. Muita National intelligence estimaten aiheita ovat olleet mm. Yhdysvaltojen ja Neuvostoliiton strategisten ydinasejoukkojen tasapaino, konventionaalisen asevoiman tasapaino Euroopassa, näkymiä Neuvostoliiton ja Kiinan suhteiden kehitykseen, näkymä Atlantin liittoutuman yhtenäisyydestä, kehitysmaiden kansainvälisten velkaongelmien merkittävyys, Neuvostoliiton aikeet ja kyvyt, Neuvostoliiton hyökkäys Afganistaniin, Jugoslavian hajoaminen, Kiinan atomipommitesti ja OPEC-maiden sijoitusstrategiat (Johnson, 2008).

Muita esimerkkejä päivittäisistä tiedustelutuotteista ovat: Daily digest, Daily intelligence summary ja Early report. Daily digest on noin 10–15 sivun tiedusteluraportti, jossa esitellään yhtä asiaa maailmanlaajuisesta näkökulmasta. Se jaetaan sadoille valtionhallinnon viranomaisille, jotka työskentelevät ulkopolitiikan parissa eri ministeriöissä. Daily intelligence summary on raportti, joka sisältää päivittäisen analyysin mahdollisesta kriisitilanteesta ja yhteenvedon relevantista tiedustelutiedoista viimeisen 24 tunnin ajalta. Early report käsittelee päivän ajankohtaisia asioita. Se pohjautuu suurten uutistoimitusten pääkirjoituksiin ja jaetaan maanantaista perjantaihin joka aamu kello 8 korkean tason viranomaisille Yhdysvaltain valtionhallinnossa. (Goldman, 2011, s. 87 ja 107.)

Tuotteiden jakaminen luo yhteyden tiedustelun ja päätöksentekijöiden välille. PDB jaetaan kirjallisena tuotteena, mutta sen lisäksi tiedustelupalvelun edustajat esittelevät sen suullisesti presidentille. Wolfberg (2017) esittelee artikkelissaan tuotteen esittelijöiden ja päätöksentekijöiden suhdetta. Hän kuvaa esittelyn tarkoitusta käsitteellä “sensegiving”, joka voidaan suomentaa ”järkeistämiseksi”. Käsite kuvaa prosessia, jossa tiedustelupalvelun asiantuntija selittää ja asettaa kontekstiin monimutkaisia ja teknisiä asioita, jotta päätöksentekijä saa tiedustelutuotteesta parhaan mahdollisen hyödyn. Esittelijän tehtävänä on toimia välittäjänä ja tulkkina tiedusteluyhteisön ja poliitikkojen välillä, mutta hän ei itse kirjoita tiedustelutuotetta tai anna toimenpidesuosituksia päätöksentekijöille. Esittelijät ohjaavat myös tiedustelua kommunikoimalla päätöksentekijöiden kiinnostuksen kohteita tai tietotarpeita tuotetta kirjoittaville analyytikoille. (Wolfberg, 2017.)

Tässä tutkielmassa tiedustelutiedon jakamisella tarkoitetaan valmiin tiedustelutuotteen toimittamista tai esittelyä päätöksentekijälle. Jakaminen on vuorovaikutuksellinen toiminto, joka luo yhteyden tiedusteluorganisaation ja päätöksentekijän välille. Jakamiseen liittyy tiedustelutiedon kommunikointi ja selittäminen päätöksentekijälle sekä tarkentavat kysymykset ja palaute tiedusteluorganisaatiolle.

3 TIEDONLOUHINTA

> Information is the oil of the 21st century, and analytics is the combustion engine. – Peter Sondergaard, former executive vice president, research and advisory at Gartner

Tässä luvussa esitellään tutkielman keskeisintä teoriaa – tiedonlouhintaa. Tiedonlouhinta on ratkaisu, jonka avulla strategisen tiedustelun informaatiotulvaa pyritään ratkaisemaan. Luvussa esitellään aluksi tiedonlouhinnan käsitettä sekä taustaa. Tämän jälkeen kuvaillaan datan, informaation ja tiedon hierarkiaa sekä luodaan lyhyt katsaus big dataan ja tekoälyyn. Lopuksi esitellään tarkemmin tiedonlouhintaprosessin vaiheita erilaisia louhintamenetelmiä.

3.1 Tiedonlouhinnan käsite

Fayyad, Piatetsky-Shapiro ja Smyth (1996) ovat kehittäneet tiedonlouhintaprosessin (Knowledge Discovery in Databases, KDD), jonka tavoitteena on tuottaa datasta tietoa. Tiedonlouhinta on kehitetty hyödyntämällä useita eri tutkimusalueita: tietokantoja, tilastotiedettä, tekoälyä, koneoppimista ja datan visualisointia. Vastaavaa prosessia on kuvataan myös käsitteillä: data mining (DM), knowledge extraction, information discovery, information harvesting, data archeology ja data pattern processing. Verrattuna datan louhintaan, tiedonlouhinnan tarkoitus on korostaa prosessin lopputuotteena syntyvän tiedon merkitystä, syötteenä toimivan datan sijasta. Fayyad ym. (1996) korostavat, että tiedonlouhinta on laajempi prosessi, jonka yksi vaiheista on datan louhinta. Vaikka he ovat pyrkineet erottamaan datan- ja tiedonlouhinnan käsitteet toisistaan, useat tutkijat käyttävät kuitenkin usein datan louhinnan -käsitettä kuvaamaan laajaa, tiedon tuottamiseen keskittyvää prosessia (Mariscal, Marban & Fernandez, 2010). Tiedonlouhintaprosessin rinnalle on kehitetty muitakin vaihtoehtoisia prosessimalleja, kuten CRoss-Industry Standard Process (CRISP-DM) ja Sample, Explore, Modify, Model, Access (SEMMA), mutta Fayyadin, ym. (1996) kehittämä tiedonlouhintaprosessi on yksityiskohtaisempi, jonka vuoksi useimmat tutkijat ja asiantuntijat suosivat sitä (Shafique & Qaiser, 2014).

Tiedonlouhintaprosessin keskeisin toiminto on datan louhinta. Datan louhinnassa algoritmien avulla tuotetaan datasta malleja (Fayyad ym., 1996). Leskovecin, Rajaramanin ja Ullmanin (2020, s. 1–4) mukaan datasta voidaan muodostaa tilastollisia, koneoppimisen, laskennallisia, referoivia tai ominaisuuksia tunnistavia malleja. Zakin ja Meiran (2014, s. 1) mukaan datan louhinnan tavoitteena on löytää oivaltavia, mielenkiintoisia, kuvaavia, ymmärrettäviä, uusia tai ennustavia malleja suuresta datamassasta. Tiedustelun käsitteistössä datan louhinnalla tarkoitetaan hyödyllisen tiedon erottamista suurista datajoukoista tai tietokannoista (Goldman, 2011, s. 88).

Tiedonlouhintaa on alun perin sovellettu mm. astronomian tutkimukseen ja liiketoiminnan tarpeisiin (Fayyad ym., 1996). Tänä päivänä suosittuja tiedonlouhinnan soveltamisalueita ovat terveydenhuolto (ks. Jothi & Husain, 2015) sekä opetus ja koulutus (ks. Romero & Ventura, 2013) sekä liiketoimintatiedustelu. Zanasin (1998) mukaan liiketoimintatiedustelussa voidaan hyödyntää tiedonlouhintaa, kun tavoitteena on ymmärtää markkinoita: kilpailijoiden toimintaan, teknologian kehittymistä tai trendejä. Sen avulla voidaan automaattisesti tuottaa analyysi ja synteesi internetistä saatavilla olevasta datasta (Zanasi, 1998). Dey, Haque, Khurdiya ja Shroff (2011) esittelevät erilaisia tekniikoita, joiden avulla voidaan tuottaa erilaisista verkkolähteistä kerätystä tekstistä tietoa liiketoimintatiedustelun tarpeisiin. Sosiaalisesta mediasta ja uutisartikkeleista voidaan kerätä tekstiä, jota prosessoimalla voidaan esimerkiksi indikoida tuotemerkin suosiota ja arvioida myyntilukuja. Tekstidatan muuttaminen rakenteelliseen muotoon ja yhdistäminen muuhun informaatioon on keskeistä. Loppukäyttäjä hyödyntää tätä lajiteltua ja luokiteltua informaatiota. (Dey ym., 2011.) Khan (2012) on selvittänyt tiedonlouhinnan hyödyntämistä liiketoimintatiedustelussa konseptitasolla. Hänen mukaansa tiedonlouhinta on yksi pitkälle kehitetty prosessi, jota voidaan hyödyntää erittäin hyvin liiketoimintatiedustelun osana. Suurin hyöty tiedonlouhinnasta liiketoimintatiedustelussa on kyky tuottaa tarkkaa informaatiota, kuten tilannetietoa yrityksen suorituskyvystä. Liiketoimintatiedustelun konsepti ja tiedonlouhinnan prosessi luovat pohjan, jonka päälle voidaan kehittää menetelmiä ja tekniikoita, joilla informaatiotulvaa saadaan hallittua ja tuotettua siitä arvokasta tietoa. (Khan, 2012.) Ziegler (2012) esittelee kirjassaan liiketoimintatiedustelun tarpeisiin kehitettyjä automatisoituja tiedonkeräysmenetelmiä. Hän on koonnut kirjaansa julkaisuja, jossa menetelmiä sovelletaan eri toiminta-alueille, kuten trendi- ja aiheanalyysiin, brändin ja maineen automaattiseen seurantaa internetissä, asiakaspalautteen automatisoituun lajitteluun ja luokitteluun, automatisoituun uutistekstien tulkintaan, sisällön poimintaan uutisartikkeleista, informatiivisen sisällön suodattamiseen uutissivustoilta, teknologian synergiaetujen tunnistamiseen ja tekstin semanttisten suhteiden laskentaan.

Tässä tutkielmassa tiedonlouhinnalla tarkoitetaan laajaa prosessia, joka jäsentää datalähtöisen tiedontuotantoprosessin eri toimintoja. Tiedonlouhinnan ja tiedustelun tavoitteet ovat samankaltaiset – datasta tuotetaan tietoa. Tiedonlouhinta tarjoaa teknisen ratkaisun, miten tiedustelun informaatiotulva saadaan hallittua.

3.2 Data, informaatio ja tieto

> An ounce of information is worth a pound of data. An ounce of knowledge is worth a pound of information. An ounce of understanding is worth a pound of knowledge. – Russell Ackoff

Data on symboleita, jotka kuvaavat objekteja ja tapahtumia (Ackoff, 1989). Data on osittaista ja ehdollista, se muodostetaan tilannesidonnaisten konseptien, tallennusten ja käytäntöjen kautta (Jones, 2019).

Data esitetään usein matriisina (kuvio 4), jossa on rivejä ja sarakkeita. Riippuen kontekstista, rivien tiedoista voidaan käyttää termejä entiteetti, instanssi, esimerkki, tallenne, objekti, piste tai ominaisvektori. Sarakkeista voidaan käyttää termejä attribuutti, ominaisuus, ulottuvuus, muuttuja tai kenttä. Rivit määrittävät datan koon ja sarakkeet ulottuvuuksien määrän. Kaikki data, kuten kuvat, tekstit, ääni, ym. eivät ole kuitenkaan alun perin matriisin muodossa, mutta ne voidaan muuttaa sellaiseksi. (Zaki & Meira, 2014, s. 1–3.)

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

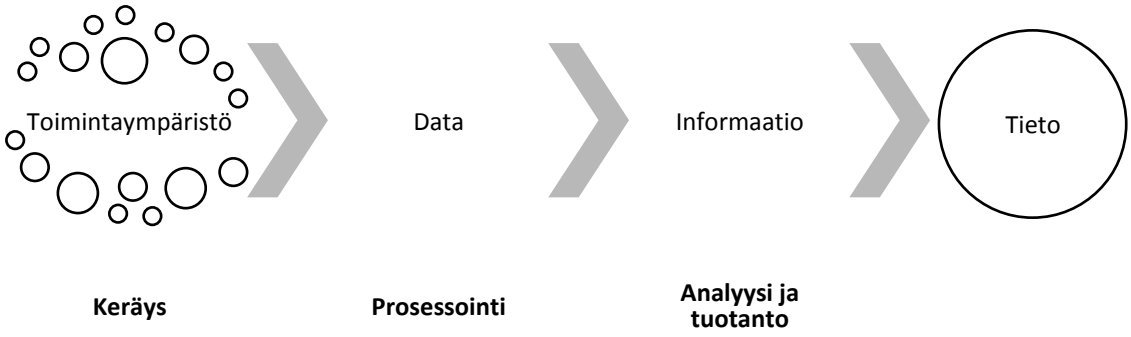
KUVIO 4 Datamatriisi (Zaki & Meira, 2014, s. 1).

Data voidaan jakaa kolmeen luokkaan: diskreetti, jatkuva ja kategorinen. Diskreetti data koostuu yksittäisistä luvuista. Jatkuva data sisältää mitattua dataa, joka voi muodostaa äärettömän määrän arvoja. Kategorinen (tai symbolinen) data on tekstiä, eli merkkijonoja. Se on rikas datatyyppi, jolla voi ilmaista sellaisia arvoja, johon numeraalinen data ei pysty. Tekstidataa ei kuitenkaan voi käsitellä matemaattisesti ilman, että se ensin muutetaan numeraaliseksi dataksi. (Kantardzic, 2011, s. 26–28.)

Mitä enemmän datassa on ulottuvuuksia, sitä harvempaa se tulee olemaan. Harvaa ja moniulotteista dataa on vaikea tulkita. Tämä muodostaa myös haasteen tietokoneelle, jos datasta on tarkoitus etsiä rakenteita. Ongelmaa kutsutaan ulottuvuuksien kiroukseksi (eng. curse of dimensionality) (Verleysen & Francois, 2005).

Data muuttuu hyödylliseksi, kun sitä prosessoidaan. Prosessoitua dataa voidaan kutsua informaatioksi. Informaatiolla voidaan vastata kysymyksiin kuka, mitä, milloin, missä ja kuinka monta. Yhdistelemällä informaatiota muodostetaan tietoa, jonka kautta voidaan muodostaa vastaus kysymykseen: miten.

Selittämällä tietoa muodostetaan ymmärrys, jonka avulla vastataan kysymykseen: miksi. (Ackoff, 1989.) Tieto on informaatiota, joka on asetettu kontekstiin, se on merkityksellistä ja sitä voidaan hyödyntää toiminnassa, kuten ongelmanratkaisussa (Khan, 2012) tai suunnittelussa ja päätöksenteossa. Yhdysvaltojen asevoimien yleisesikunta (US Joint Chiefs of Staff, 2013) kuvaa tiedustelua ohjaavassa doktriinissa datan, informaation ja tiedon suhdetta tiedusteluprosessiin seuraavalla tavalla: toimintaympäristöstä hankitaan dataa, joka prosessoidaan informaatioksi, josta tuotetaan analyysin kautta tietoa. Tämä mallinnus on esitetty kuviossa 5.



KUVIO 5 Datan, informaation ja tiedon suhde (US Joint Chiefs of Staff. 2013). Mukailtu ja suomennettu alkuperäisestä.

Tämän tutkielman näkökulmasta on hyvä ymmärtää, että dataa voidaan kerätä lähes mistä tahansa todellisen maailman ilmiöstä. Tietokoneet pystyvät käsittelemään vain numeraalista dataa, koska niiden toiminta perustuu binääriseen laskentaan. Tietokoneiden avulla voidaan muuttaa suuri määrä dataa ihmisen hyödynnettäväksi. Tietokoneet ovat hyviä prosessoimaan numeroita ja dataa, mutta ihminen on hyvä ymmärtämään monimutkaisia representaatioita ja sitomaan informaatiota kontekstiin.

3.3 Suuri määrä dataa vai Big Data?

Big data on trendikäs käsite, kun puhutaan suurien tietomassojen käsittelystä ja data-analyysistä. Koska tämäkin tutkielma käsittelee vastaavia aiheita, on hyvä luoda katsaus, mitä Big datalla todellisuudessa tarkoitetaan. Miten voidaan erottaa suuri määrä dataa ja Big data? Mauro, Greco ja Grimaldi (2016) esittävät, että Big data on tietovaranto, jonka volyymi, nopeus ja moninaisuus on niin suuri, että sen hyödyntäminen edellyttää erityistä teknologiaa ja analyyttisiä menetelmiä. Gartnerin (2020) määritelmän mukaan se on tietovaranto, joka ominaisuuksiensa vuoksi asettaa uudenlaisia vaatimuksia datan prosessoinnille. Madden (2012) yksinkertaistaa määritelmän kuvaamalla, että Big data on liian isoa, nopeaa ja vaikeaa, jotta sen käsittely tavanomaisilla menetelmillä olisi mahdollista. Tämä tarkoittaa esimerkiksi petatavuittain, eri lähteistä ja jatku-

vasti kerättävää dataa, josta pitää kyetä tuottamaan nopeasti tietoa moniulotteisen analyysin avulla (Madden, 2012).

Big data ei ole todellisuudessa niin vallankumouksellisen suurta, universaalia ja kattavaa kuin mitä usein esitetään (Jones, 2019). Tämä tarkoitta, että Big datan käsitettä käytetään virheellisesti, vaikka tarkoitetaan vain suurta datajoukkoa. Leskovec ym. (2020) kuvailevat, että heidän kirjassaan keskitytään massiivisten datajoukkojen käsittelyyn. Heidän määritelmänsä mukaan massiivinen datajoukko on niin iso, että se ei mahdu keskusmuistiin. Käytännössä tämä tarkoittaa yli kymmenen gigatavun suuruista datajoukkoa. Vaikka big data kuulostaa massiivista datajoukkoa pienemmältä, tarkoitetaan sillä käsitteenä kuitenkin tätä suurempaa ja moniulotteisempaa datajoukkoa.

Yhteenvetona voidaan todeta, että alle kymmenen gigatavun suuruinen datajoukko on niin pieni, että sen hallinta ja prosessointi ei vaadi kovin monimutkaisia menetelmiä. Massiiviset datajoukot ovat yli kymmenen gigatavun suuruisia, mutta niiden hallinta ja prosessointi voidaan toteuttaa tavanomaisilla menetelmillä. Big data on jatkuvasti kasvava useiden petatavujen kokoinen tietovaranto, jonka käsittely vaatii erityisiä menetelmiä. Tässä tutkielmassa käsitellään massiivisia datajoukkoja.

3.4 Tekoäly

> Googlen tutkimusryhmä kertoo luoneensa tekoälyn, joka oppi itsekseen go-pelin mestariksi – "Se pystyy luomaan tietoa itse" – Yle uutiset, 19.10.2017

Tässä luvussa luon lyhyen katsauksen tekoälyyn, koneoppimiseen ja syväoppimiseen, koska niillä on oleellinen liityntäpinta tiedonlouhintaan.

Tekoälyn tutkimus voidaan katsoa alkaneeksi jo 50-luvulla. McCarthy, Minsky, Rochester ja Shannon (2006) arvioivat vuonna 1955, että oppiminen ja älykkyys voidaan kuvata niin tarkasti, että niitä voidaan simuloida koneen avulla. Tätä arviota voidaan pitää tekoälytutkimuksen käynnistäjänä. Tutkimus ei ole edennyt vielä niin pitkälle, että olisi kehitetty yleinen tai laaja – ihmisen kognitiiviset kyvyt omaava tekoäly.

Usein, kun puhutaan tekoälystä, puhutaan tarkemmin kapeasta tekoälystä, koneoppimisesta tai sen kehittyneemmästä muodosta syväoppimisesta. AlphaGo Zeroksi nimetyn ohjelman on kehittänyt Googlen tutkimusryhmä (Silver ym, 2017). He esittelevät artikkelissaan ohjelman arkkitehtuuria ja algoritmeja: Ohjelma perustuu syvä- ja vahvistusoppimiseen, jossa se pelaa Go-peliä itseään vastaan lukuisia kertoja sekä havainnoi mustien ja valkoisten pelimerkkien sijaintia pelilaudalla. Omat ja vastustajan siirrot, sekä niistä seuraavat voitot ja häviöt muokkaavat ohjelman algoritmia ja käytöstä. Ohjelma oppii toistojen kautta, kokeilemalla ja erehtymällä kuten ihminen.

AlphaGo ja valtaosa nykypäivän tekoälysovelluksista luokitellaan kapeaksi tekoälyksi, joka on hyvin kaukana yleisestä tekoälystä. Kapea tekoäly kykenee ratkaisemaan rajatun ja ennalta määritellyn ongelman, kun taas yleinen tekoäly kykenee ihmisen kognition kaltaiseen ajatteluun. Yleisen tekoälyn kehi-

tys on vasta lähtökuopissaan, kun kapean tekoälyn tutkimus ja reaalimaailman sovellukset ovat yleisiä. (ks. Adams ym., 2012)

Koneoppimisella tarkoitetaan menetelmiä, jotka tunnistavat datasta automaattisesti rakenteita, ja hyödyntävät näitä rakenteita ennustaessaan tulevaa dataa (Murphy, 2012). Oppiva algoritmi suorittaa tehtävän, kuten luokittelun ja parantaa suorituskykyään oppimansa kokemuksen, eli sille syötetyn opetusdatan avulla (Jordan & Mitchell, 2005).

LeCun, Bengio ja Hinton (2015) esittelevät artikkelissaan syväoppimisen ja koneoppimisen eroja. Syväoppiminen jatkaa siitä, mihin konventionaalinen koneoppiminen ei kykene. Koneoppimisessa datan esiprosessointi vaatii enemmän työtä, koska malli vaatii syötteeksi erikseen valitun ominaisuusvektorin. Syväoppimisen malliin voidaan syöttää suoraan dataa, josta se automaattisesti tunnistaa edustavat ominaispiirteet. Syväoppimisen algoritmit ovat useita prosessointikerroksista koostuvia laskennallisia malleja, jotka voivat oppia datasta useita representaatioita monilla abstraktiotasoilla (LeCun, ym., 2015.)

Oppivat algoritmit, sekä kone- että syväoppimisen, opetetaan yleensä ohjattuna, eli algoritmille syötetään suuri määrä luokiteltua dataa ja se tuottaa mallin, joka osaa luokitella uutta dataa muodostamiensa vektoreiden avulla. Ohjatun oppimisen vastakohta on ohjaamaton oppiminen, joka mukailee ihmisten ja eläinten tapaa oppia: mallille syötetään dataa ja se muodostaa niistä rakenteita, kuten kategorioita tai klustereita itse. (LeCun, ym., 2015.)

Kone- ja syväoppimista hyödynnetään mm. kuvantunnistuksessa, puheen muuttamisessa tekstiksi, käyttäjää profiloivissa suosittelijoissa ja hakukoneissa. Syväoppimisen kehittyviä tutkimusalueita ovat konenäkö ja luonnollisen kielen ymmärtäminen. (LeCun, ym., 2015.)

Tässä tutkielmassa hyödynnetään sekä tekoälyksi luokiteltavia koneoppimiseen perustuvia algoritmeja.

3.5 Tiedonlouhinta prosessina

Fayyad ym. (1996) jakavat tiedonlouhintaprosessin yhdeksään vaiheeseen:

1. Toimintaympäristön analyysi
2. Datajoukon muodostaminen
3. Datan esiprosessointi
4. Datan ulottuvuuksien vähentäminen
5. Datan louhintamenetelmän valinta
6. Algoritmin valinta ja mallin muodostaminen
7. Datan louhinta
8. Tulosten tulkinta
9. Tiedon hyödyntäminen

Ensimmäisessä vaiheessa muodostetaan ymmärrys kohdeympäristöstä, johon tiedonlouhintaprosessia sovelletaan ja määritellään tavoite asiakkaan näkökulmasta. Toisessa vaiheessa valitaan datajoukko, tai sen osajoukko tai datanäyte,

johon tiedonlouhinta kohdistetaan. Kolmannessa vaiheessa valittu data esikäsitellään tarvittavin osin. Tämä tarkoittaa esimerkiksi kohinan poistamista tai puuttuvien arvojen täydentämistä. Neljännessä vaiheessa tunnistetaan tehtävän kannalta oleelliset ominaispiirteet vähentämällä datan ulottuvuuksia. Viidennessä vaiheessa valitaan tehtävän kannalta sopiva datanlouhintamenetelmä, joita ovat esimerkiksi luokittelu, klusterointi, yhteenveto, regressio. Kuudennessa vaiheessa valitaan hyödynnettävät algoritmit ja menetelmät rakenteiden tunnistamiseen. Tässä vaiheessa päätetään mitkä mallit ja parametrit soveltuvat lopullisen tavoitteen saavuttamiseen. Seitsemännessä vaiheessa suoritetaan datan louhinta, jossa datasta etsitään kiinnostuksen kohteita. Kahdeksannessa vaiheessa tulkitaan löydettyjä kaavoja. Tässä vaiheessa voidaan visualisoida löydettyjä rakenteita tai niiden tuottamaa informaatiota. Vaiheen tulkintojen perusteella voidaan iteroida takaisin aikaisempiin vaiheisiin. Yhdeksännessä vaiheessa tuotettu tieto hyödynnetään suoraan, yhdistetään toiseen järjestelmään myöhempää hyödyntämistä varten tai dokumentoidaan ja raportoidaan tiedosta kiinnostuneille osapuolille. Tässä vaiheessa tarkastetaan myös mahdolliset ristiriidat suhteessa aiemmin tuotettuun tietoon. (Fayyad ym., 1996.)

Seuraavissa alaluvuissa esitellään tarkemmin prosessin vaiheita. Luvussa 3.4.1 esitellään tarkemmin tekniikoita, joita hyödynnetään datajoukon muodostamisessa, esiprosessoinnissa ja ulottuvuuksien vähentämisessä. Luvussa 3.4.2 esitellään tarkemmin datanlouhintamenetelmiä ja algoritmeja. Luvussa 3.4.3 esitellään tarkemmin tulosten tulkinnan ja hyödyntämisen mahdollistavia visualisointitekniikoita.

3.5.1 Datan esiprosessointi eli valmistelu ja vähentäminen

Kun datajoukko on valittu ja muodostettu pitää se valmistella sellaiseen muotoon, että datan louhinta on mahdollista. García, Luengo ja Herrera (2015) esittelevät kirjassaan tiedonlouhinnassa tarvittavia datan esiprosessointitekniikoita. Datajoukoissa on tyypillisesti tiedonlouhintatehtävän kannalta haitallisia ominaisuuksia, kuten epäoleellista tietoa, kohinaa ja puuttuvia arvoja. Datan laatua voidaan parantaa ja haitallisia ominaisuuksia poistaa erilaisten esiprosessointitekniikoiden avulla. Tekniikat voidaan jakaa valmistelu- ja vähentämistekniikoihin. (García ym., 2015.)

Datan valmistelutekniikoita ovat (García ym., 2015):

- datan puhdistaminen (cleaning)
- datan imputaatio (imputation), eli puuttuvan tiedon paikkaaminen
- kohinan tunnistaminen (noise identification)
- datan muodon muuttaminen (transformation)
- datan integrointi (integration)
- datan normalisointi (normalization)

Datan puhdistamisella tarkoitetaan huonon datan korjaamista, väärän datan suodattamista ja liian yksityiskohtaisen datan poistamista. Datan puhdistamisella on paljon yhteistä imputaation ja kohinan tunnistamisen kanssa. Datan imputaatiossa täydennetään datan puuttuvia arvoja. Muuttujille, jotka ovat tyh-

jiä annetaan keinotekoinen arvo esimerkiksi arvion, tilastollisen päättelyn tai laskutoimituksen perusteella. Kohinan tunnistaminen tarkoittaa sattumanvaraisten virheiden tunnistamista muuttujista. Kohinan tunnistamisen jälkeen voidaan suorittaa erillisiä korjaavia toimenpiteitä kohinan poistamiseksi. Kohinan poistaminen voidaan tulkita myös tasoittamiseksi. Datan muodon muuttamisella tarkoitetaan tasoittamista, ominaispiirteiden muodostamista, yhdistämistä, normalisointia, diskretisointia ja yleistämistä. Data muutetaan sellaiseen muotoon, jotta datanlouhinta on mahdollista tai tehokkaampaa. Datan integroinnilla tarkoitetaan useampien datajoukkojen yhdistämistä eri tietovarannoista. Integroinnissa tunnistetaan ja yhdenmukaistetaan muuttujat ja ulottuvuudet, analysoidaan attribuuttien korrelaatio, tunnistetaan ja poistetaan päällekkäisyydet sekä hallitaan ristiriidat eri tietolähteiden arvoissa. Datan normalisoinnissa attribuuttien mittayksiköt yhtenäistetään, jolloin niillä on yhteinen painoarvo. Tämä on tarpeellista, kun hyödynnetään tilastollisia menetelmiä. (García ym., 2015.)

Datan vähentämistekniikoita ovat (García ym., 2015):

- ominaispiirteiden valinta (feature selection)
- instanssin valinta (instance selection)
- diskretisointi (discretisation)
- ominaispiirteiden muodostaminen (feature extraction)

Ominaispiirteiden valinnalla tarkoitetaan merkityksettömien tai tarpeettomien ominaispiirteiden tai ulottuvuuksien poistamista datasta. Tavoitteena on löytää mahdollisimman pieni, mutta tiedonlouhintatehtävän kannalta riittävän kattava otos datasta. Instanssin valinnalla tarkoitetaan datajoukon alijoukon valintaa, joka tehdään käyttämällä jotakin sääntöä tai sattumanvaraisesti. Tavoitteena on ottaa sellainen näyte, joka edustaa riittävästi koko datajoukkoa. Diskretisointi muuttaa datajoukon numeeriset attribuutit nominaalisiksi. Diskretisointi voidaan ymmärtää kuuluvan joko datan vähentämis- tai valmistelutekniikaksi. Ominaispiirteiden muodostaminen on jatkoa ominaispiirteiden ja instanssien valinnalle. Siinä poistamisen sijaan voidaan yhdistää ja luoda keinotekoisia attribuutteja. (García ym., 2015.)

3.5.2 Datan louhinta eli laskennallinen mallintaminen

Datan louhinta on tiedonlouhintaprosessin ydin, joten kuvaan sen tässä luvussa prosessin muita vaiheita tarkemmin.

Datan louhinnassa ei ole ainoastaan kyse tekoälystä tai koneoppimisesta. Datan louhinnassa voidaan hyödyntää koneoppimisen algoritmeja, mutta se ei ole välttämätöntä, ja joissakin tapauksissa perinteiset tilastolliset mallit ovat jopa tehokkaampia, kuin älykkäät koneoppimiseen perustuvat mallit (Leskovec, ym., 2020). Datan louhinta perustuu logiikkaan, matematiikkaan ja algoritmeihin. Algoritmien avulla datajoukosta muodostetaan malli, eli tunnistetaan datasta rakenteita (Fayyad ym., 1996). Rakenne voi olla esimerkiksi rypäs, eli klusteri datapisteitä, joilla on samankaltaisia ominaisuuksia tai arvoja. Datasta voidaan tunnistaa erilaisia rakenteita riippuen tehtävästä ja datasta. Tutkimalla

dataa ja analysoimalla tehtävää valitaan tehtävään soveltuva datan louhintamenetelmä.

Datan louhinnalla on tyypillisesti kaksi tavoitetta: ennustaminen ja kuvaileminen. Ennustamisella tarkoitetaan, että datan perusteella ennustetaan uutta dataa. Kuvailemisella tarkoitetaan, että datasta poimitaan ja esitetään kuvailevaa informaatiota. (Fayyad ym., 1996.)

Zaki ja Meira (2014, s. 25–30) jakavat datan louhintamenetelmät neljään kategoriaan:

- Tutkivassa data-analyysissa (exploratory data-analysis) tunnistetaan datan sisältämien attribuuttien ilmentymiä itsenäisenä tai toisistaan riippuvaisina. Usein käytetty tutkivan data-analyysin menetelmä on pääkomponenttianalyysi (principle component analysis, PCA), jonka avulla voidaan vähentää datan ulottuvuuksien määrää ja tunnistaa datajoukosta merkittävää informaatiota.
- Toistuvien kaavojen tunnistamisessa (frequent pattern mining) poimitaan hyödyllisiä ja informatiivisia toistuvia rakenteita suurista ja moniulotteisista datajoukoista. Usein käytetty menetelmä on Apriori-algoritmi. Sitä voidaan soveltaa esimerkiksi ruokakorianalyysissa (market basket analysis), jossa pyritään tunnistamaan ostettujen tuotteiden perusteella ostokäyttäytymistä. Menetelmä tuottaa yksinkertaistettuna seuraavaa tietoa: Jos asiakas ostaa tuotteet A ja B, ostaa se myös todennäköisesti tuotteen C.
- Klusteroinnissa (clustering) muodostetaan datajoukosta luonnollisia ryhmiä, eli klustereita. Usein käytettyjä menetelmiä ovat klusterin keskiarvon menetelmä (K-means), tiheyspohjainen skannaus (DBSCAN) ja lähimmän naapurin menetelmä (nearest neighbour). Tyypillisesti klusteroinnissa malli opetetaan ohjaamattomana, eli klusterit muodostetaan ilman etukäteen määriteltyjä parametreja.
- Luokittelussa (classification) ennustetaan uuden datapisteen kuuluminen johonkin ennalta määriteltyyn kategoriaan tai luokkaan. Usein käytettyjä menetelmiä ovat todennäköisyyslaskentaan ja Bayesin teoreemaan perustuva bayeslainen luokittelu (Bayes classifier), päätöspuu (decision tree) ja tukivektorikone (SVM, support vector machine). Tyypillisesti luokittelu tehdään ohjattuna, eli luokittelun suorittava malli opetetaan etukäteen luokitellulla datalla.

Fayyad ym. (1996) lisäävät edellä esiteltyjen neljän kategorian jatkeeksi regression ja poikkeaman tunnistamisen. Zaki ja Meira (2014) kuitenkin esittävät, että regressio sisältyy luokitteluun ja poikkeaman tunnistaminen on tehtävä, joka voidaan saavuttaa esimerkiksi luokittelun, klusteroinnin tai toistuvien kaavojen tunnistamisen kautta.

Luokittelu on selkeästi yksi käytetyimmistä menetelmistä. Wu, ym. (2008) ovat selvittäneet kymmenen suosituinta datan louhintaan käytettyä algoritmia. Olen lajitellut algoritmit taulukkoon 1. Suosituimmat menetelmät ja algoritmit ovat hyvin yksinkertaisia, joten voidaankin arvioida, että soveltamalla monimutkaiseen dataan yksikertaisia menetelmiä saadaan aikaiseksi hyviä tuloksia.

TAULUKKO 1 Kymmenen suosituinta datan louhinnassa hyödynnettyä algoritmia (Wu, ym., 2008).

	Algoritmi	Kategoria	Esimerkki
1.	Decision tree	luokittelu	Mainosten personointi
2.	K-means	klusterointi	Opiskelijoiden akateemisen menestyksen ennustaminen
3.	Support Vector Machine	luokittelu	Mekaanisten virheiden diagnostiikka
4.	Apriori	toistuvien kaavojen tunnistaminen	Tulva-alueiden tunnistaminen
5.	Expectation-Maximation	klusterointi	Kuvien segmentointi
6.	PageRank	luokittelu ja klusterointi	Googlen hakukoneen sivujärjestyksen priorisointi
7.	AdaBoost	luokittelu	Rintasyövän tunnistaminen kuvista
8.	K-Neareast Neighbor	luokittelu	Taloudellisten tapahtumien ennustaminen
9.	Naive Bayes	luokittelu	Petoksen tunnistaminen
10.	Classification And Regression Trees	luokittelu	Hepatiitin tunnistaminen

Tiedonlouhinnassa datan louhintaprosessia ajetaan ja iteroidaan tyypillisesti useita kertoja, jotta datasta saadaan tunnistettua tehtävän kannalta hyödylliset rakenteet ja malli saadaan toimimaan mahdollisimman optimaalisesti (Fayyad ym., 1996).

Kuten luvussa 2.6 esiteltiin, tiedusteluanalyysissä pyritään hahmottamaan monimutkaisista ilmiöistä kaavoja, tunnistamaan poikkeamia sekä arvioimaan tulevaa. Tiedonlouhinnassa hyödynnettävät menetelmät vastaavat hyvin tähän tarpeeseen. Esimerkiksi tutkivan data-analyysin ja klusteroinnin avulla voidaan jäsentää moniulotteista informaatiota, toistuvien kaavojen tunnistamisen avulla voidaan selvittää ilmiöiden rakenteita ja luokittelun avulla voidaan tunnistaa poikkeamia tai ennustaa tulevaa.

3.5.3 Visualisointi

Mikäli datajoukolla ei ole luontaista kaksi- tai kolmiulotteista semantiikka, sitä ei voida esittää fyysisellä näyttöruudulla. Visualisointi mahdollistaa tämänkaltaisen datajoukon esittämisen. Datan visuaalista tulkintaa hyödynnetään prosessin tiedonlouhintaprosessin eri vaiheissa. Datan visualisoinnin tarkoitus on esittää data jossakin visuaalisessa muodossa, jotta ihminen kykenee ymmärtämään ja tekemään johtopäätöksiä datasta. Visualisointi tarjoaa ihmiselle katsauksen dataan ja mahdollistaa uusien hypoteesien muodostamisen. Visualisoimalla voidaan käsitellä helposti dataa, joka on moniulotteista ja sisältää kohinaa. Visualisoinnin tulkinta ei vaadi matemaattista tai tilastotieteellistä osaamista, tai ymmärrystä algoritmeista ja parametreista. (Keim, 2002.)

Visualisointitekniikoita on erittäin paljon. Tyypillisemmät ja yksinkertaisempia menetelmiä ovat kaksi- ja kolmiulotteiset pylväs-, piste- ja viivakaaviot. Moniulotteista dataa voidaan visualisoida kehittyneimmillä tekniikoilla, kuten rinnakkaiskoordinaattitekniikalla. Kaikkea dataa ei kuitenkaan voida määritellä

ulottuvuuksien kautta. Tekstidataa voidaan visualisoida esimerkiksi sanalaskurin tai sanapilven avulla. Datan hierarkiaa ja yhteyksiä voidaan visualisoida esimerkiksi verkostograafien avulla. (Keim, 2002.)

4 TUTKIMUSMENETELMÄ JA AINEISTO

Tässä luvussa esitellään suunnittelututkimuksen tutkimusmenetelmä ja sen taustoja lyhyesti. Tämän jälkeen esitellään DSRM-prosessi ja sen soveltaminen tässä tutkielmassa. Lopuksi esitellään tutkimusaineistona käytetty GDELT-tietokanta.

4.1 Suunnittelututkimus

Suunnittelututkimus (Design science research, DSR) on konstruktiivinen tutkimusmenetelmä, jossa tietoa ja ymmärrystä muodostetaan rakentamalla artefakti ja soveltamalla sitä ongelmaan (Hevner ym., 2014). Artefakti tarkoittaa ihmisen keinotekoisesti (artificially) muodostamaa asiaa, joka täydentää luonnollisesti muodostuneita asioita (Simon, 1996).

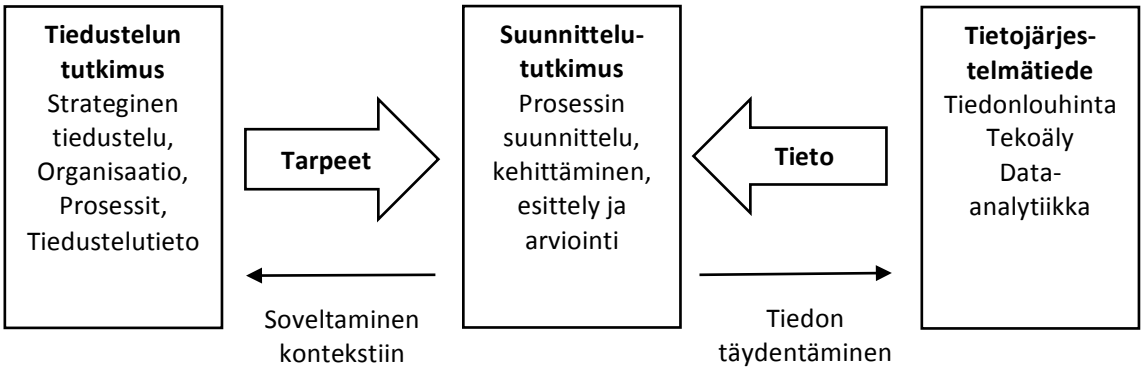
Tässä tutkielmassa tutkimusongelmaan selvitetään ratkaisu hyödyntäen Peffersin ym. (2007) kehittämää suunnittelututkimuksen metodologiaa (Design science research methodology, DSRM). Heidän metodologiassa kehitetään artefakti, joka tuottaa konkreettisen ratkaisun määriteltyyn ongelmaan. Artefaktin suunnittelu, kehittäminen ja evaluointi tuottaa ymmärrystä tutkimusongelmasta ja sitä koskettavasta aihepiiristä. Menetelmän lopullisena tavoitteena on tuottaa tieteellistä tietoa (Peffers ym., 2007).

DSRM:n lisäksi on useita erilaisia tapoja tehdä suunnittelututkimusta (Peffers ym., 2018):

- Tietojärjestelmien suunnittelun teoriassa keskitytään kehittämään, kokoamaan ja esittelemään tietojärjestelmien suunnittelun teoriaa.
- Suunnittelukeskeisessä tietojärjestelmätutkimuksessa keskitytään tietojärjestelmien suunnittelun toimintojen kuvaamiseen ja innovatiivisten toimintatapojen esittelyyn.
- Selittävässä suunnitteluteoriassa keskitytään suunniteltujen ominaisuuksien vaikutuksiin ympäristössä tai käyttäjissä.

- Toiminnallisessa suunnittelututkimuksessa keskitytään käytäntöön ja ongelman ratkaisevan artefaktin suunnitteluun.

DSRM on edellä esiteltyihin menetelmiin verrattuna prosessiltaan joustava ja se keskittyy soveltavaan artefaktin kehittämiseen (Peffers ym., 2018). DSRM on valittu tämän tutkielman menetelmäksi, koska joustavuutensa vuoksi se soveltuu muita menetelmiä paremmin tieteidenväliseen tutkimukseen. Tässä tutkielmassa yhdistetään tietojärjestelmätiede tiedustelun tutkimukseen. Tutkielma tarjoaa esimerkin, miten tiedustelun tutkimusta voidaan lähestyä tietojärjestelmätieteen kautta. Soveltamisalueen tutkimus määrittää tarpeet, tietojärjestelmätiede muodostaa tietopohjan ja suunnittelututkimus yhdistää teorian todelliseen maailmaan (Hevner ym, 2014). Tämän tutkielman viitekehys on esitelty kuviossa 6.



KUVIO 6 Tutkimuksen viitekehys (Hevner, March, Park & Ram 2004). Sovellettu ja suomennettu alkuperäisestä.

4.2 Suunnittelututkimuksen metodologian soveltaminen tässä tutkielmassa

DSRM-prosessi jaetaan seuraaviin vaiheisiin (Peffers ym., 2007):

1. Ongelman määrittely ja motivaatio ratkaisulle
2. Ratkaisun määrittely ja tavoitteiden asettaminen
3. Suunnittelu ja kehittäminen
4. Demonstraatio
5. Evaluointi
6. Kommunikointi

Tämän tutkielman tutkimusprosessi käynnistyi tavoitekeskeisesti. Taustalla oli ajatus, että uuden teknologian kuten tekoälyn ja koneoppimisen hyödyntäminen tiedustelussa on mielenkiintoinen ja ajankohtainen aihe. Oletuksena oli, että internetistä voidaan kerätä uutistekstejä, muuttaa ne rakenteelliseksi dataksi ja tuottaa datasta informaatiota ja tietoa, jota voidaan hyödyntää tiedustelun tarpeisiin. Tiedustelu- ja tiedonlouhintaprosessien yhtäläisyydet oli alustavasti

tunnistettu. Alkuperäinen tavoite oli selvittää, miten tiedonlouhintaa voidaan soveltaa tiedustelussa. Tavoitetta tarkennettiin prosessin edetessä.

Ongelman määrittely ja motivaatio ratkaisulle

Ensimmäisessä vaiheessa määritellään tutkimusongelma ja oikeutetaan ratkaisun tärkeys. Tutkimusongelman määrittelyssä laajempi ongelma puretaan pienempiin osiin. Tämä purkaminen tai erittely ohjaa prosessin myöhempiä vaiheita. Ratkaisun kehittäminen tulee oikeuttaa, koska se motivoi tutkijaa ja yleisöä kiinnostumaan ratkaisusta, ja se auttaa ymmärtämään tutkijan ajatusta ongelman olemuksesta. (Peffers ym., 2007.) Gregor ja Hevner (2013) esittävät, että suunnittelututkimuksen avulla voidaan tuottaa erilaisia kontribuutioita riippuen tutkittavan ongelman ja ratkaisujen maturiteetista. Matalan maturiteetin tutkimusalueelle voidaan soveltaa korkean maturiteetin tutkimusalueella kehitettyjä ratkaisuja. Koska tietojärjestelmätieteen maturiteetti on korkea ja tiedustelun tutkimuksen maturiteetti on matala, voidaan suunnittelututkimuksen avulla muodostaa ehdotus tiedonlouhinnan (tutkittu ja sovellettu paljon tietojärjestelmätieteessä) soveltamisesta tiedustelun kontekstiin.

Perehtymällä kirjallisuuteen hahmotettiin tarkemmin strategisen tiedustelun toimintaympäristöä ja siihen liittyviä ongelmia. Kirjallisuudesta tunnistettiin, että teknologian kehitys on johtanut siihen, että käsiteltävän datan ja tiedon määrä tiedustelussa on niin suuri, että sen käsittely on haastavaa. Tämä muodostaa kaksi merkittävää ongelmaa: Osa kerätystä tiedosta jää hyödyntämättä ja ihmisten resursseja käytetään sellaiseen työhön, joka voidaan tehdä koneen avulla. Aikaisemmassa tutkimuksessa oli tunnistettu, että tietojärjestelmien avulla voidaan kehittää tiedusteluorganisaatioiden toimintaa, mutta konkreettisia ratkaisuja löydettiin vain vähän.

Ratkaisun määrittely ja tavoitteiden asettaminen

Toisessa vaiheessa määritellään tavoitteet ratkaisulle. Tutkimusongelmasta johdetaan rajatut tavoitteet, jotka ovat mahdollisia ja saavutettavissa. Tavoitteet voivat olla määrällisiä tai laadullisia. Tavoitteiden asettaminen vaatii ymmärrystä tutkimusongelmasta ja olemassa olevista ratkaisuista eli aikaisemmasta tutkimuksesta. (Peffers ym., 2007.)

Perehtymällä aikaisempaan tutkimukseen selvitettiin, millaisia ratkaisuja ongelmaan on kehitetty. Erilaisia ratkaisuja uutisdatan keräykseen ja hyödyntämiseen erilaisissa konteksteissa on kehitetty varsin paljon. Tiedonlouhintaa ja strategista tiedustelua ei kuitenkaan ole aiemmassa tutkimuksessa konkreettisesti yhdistetty. Tässä tutkielmassa tavoitteena on yhdistää tiedonlouhintaprosessi suoraan tiedusteluprosessin eri vaiheisiin ja selvittää miten tiedusteluprosessin eri vaiheita voidaan tukea ja millaista tietoa tiedonlouhinnan avulla voidaan tuottaa. Konkreettisten ja avointen ratkaisujen avulla pienetkin organisaatiot voivat helposti kehittää omaa toimintaansa ilman suuria ja kalliita hankkeita. Tutkielman ensimmäiseksi tavoitteeksi määriteltiin konkreettisen ratkaisun esittely tiedonlouhinnan soveltamisesta strategisen tiedustelun tueksi. Lisäksi tutkielman avulla halutaan osoittaa, että tiedustelun tutkimusta voidaan tehdä

tietojärjestelmätieteen avulla, ja muodostaa ymmärrystä informaatioteknologian tarjoamista mahdollisuuksista kehittää tiedustelun toimintaa.

Suunnittelu ja kehittäminen

Kolmannessa vaiheessa suunnitellaan ja kehitetään artefakti, joka ratkaisee ongelman tai osan siitä. Artefakti voi olla algoritmi, sovellus, malli, metodi, konsepti tai järjestelmä, jonka suunnitteluun on yhdistetty tutkimuksellista työtä. Suunnittelu ja kehittäminen sisältää artefaktin toiminnallisuuksien ja arkkitehtuurin määrittelyn sekä itse artefaktin luomisen. Tämä vaihe vaatii perehtymistä teorioihin, joita voidaan soveltaa ratkaisuksi. (Peffers ym., 2007.)

Hevnerin ym. (2004) mukaan informaatioteknologian artefaktit voidaan jakaa neljään kategoriaan: konstruktiot, mallit, menetelmät ja järjestelmät. Konstruktiot ovat sanastoa ja symboleja. Niiden avulla kuvaillaan ongelmat ja ratkaisut. Konstruktiot luovat pohjan mallien sanoittamiselle. Mallit ovat abstraktioita ja representaatioita, kuten käyttäjien vaatimukset tai evaluointikriteerit. Mallien avulla konstruktiot yhdistetään todellisen maailman ilmiöihin. Menetelmien avulla ratkaistaan ongelmia. Menetelmä voi olla esimerkiksi algoritmi, käytäntö tai prosessi. Järjestelmät voivat olla joko valmiita tietojärjestelmiä, kuten sovelluksia tai niiden prototyyppejä ja esimerkkiratkaisuja.

Kirjallisuuskatsauksen avulla selvitettiin, miten strateginen tiedustelu ja tiedonlouhinta toimivat, sekä miten tiedonlouhinta on sovellettavissa osaksi tiedusteluprosessia. Tämä muodostaa tietopohjan, jonka avulla suunnitellaan ja kehitetään artefakti, Tässä tutkielmassa kehitettävä artefakti on prosessimalli, joka yhdistää tiedonlouhinnan tiedustelun kontekstiin. Prosessimalli on konkreettinen ehdotus tiedonlouhinnan soveltamisesta tiedustelun tueksi.

Demonstraatio

Neljännessä vaiheessa demonstroidaan eli esitellään artefaktia käytännössä, kun sillä ratkaistaan yksi tai useampi tutkimusongelman osa-alueista. Artefaktia voidaan käyttää kokeellisissa olosuhteissa, simulaationa, tapaustutkimuksena, todisteena tai muussa soveltavassa toiminnassa. Tämä vaihe vaatii ymmärrystä, kuinka artefaktilla ratkaistaan todellinen ongelma. (Peffers ym., 2007.)

Järjestelmien ja toimivien sovellusten avulla voidaan osoittaa, että konstruktiot, mallit tai menetelmät voidaan toimeenpanna toimivaksi järjestelmäksi. Niiden avulla voidaan osoittaa sekä suunnitteluprosessin että itse suunniteltujen artefaktien konkreettista soveltuvuutta. Rakentamalla järjestelmä, joka automatisoi prosessia voidaan osoittaa, että prosessin automatisointi on mahdollista. (Hevner ym., 2004.)

Tässä tutkielmassa kehitetyn prosessin toimivuutta demonstroidaan erikseen suunniteltavan ja kehitettävän prototyyppisovelluksen avulla. Sovellus rakennetaan kehitetyn tiedonlouhintaprosessin ohjaamana ja sen tarkoituksena on todentaa prosessin toimivuutta. Sovelluksen avulla tuotetaan uutisdatasta tiedustelutietoa kuvitteellisten, tiedustelun toimintaa kuvaavien skenaarioiden kehystämänä. Skenaariot johdetaan strategisen tiedustelun kohteista, jotka selvitetään kirjallisuuskatsauksen avulla. Tässä vaiheessa kehitetty artefakti sido-

taan todelliseen maailmaan ja ongelma-alueeseen tiedustelun asettamien vaatimusten ja empiirisen datan avulla. Demonstraatio kytkeytyy kiinteästi tutkimusprosessin seuraavaan vaiheeseen – evaluointiin.

Evaluointi

Viidennessä vaiheessa evaluoidaan, eli tarkkaillaan ja mitataan, kuinka hyvin artefakti soveltuu ongelman ratkaisuun. Tässä vaiheessa ratkaisun tavoitteita verrataan demonstraatiossa saavutettuihin tuloksiin. Evaluointi vaatii soveltuvien mittareiden ja analyysimenetelmien valinnan, jotka riippuvat tutkimusongelman luonteesta ja kehitetyn artefaktin tyypistä. Konseptina evaluointi voi sisältää mitä tahansa empiiristä näyttöä tai loogista todistusaineistoa. (Peffers ym., 2007.)

Evaluointi on suunnittelututkimuksen keskeisimpiä toimintoja (March & Smith, 1996; Hevner ym., 2004). Marchin ja Smithin (1995) mukaan artefaktin evaluoinnin tarkoitus on vastata kysymykseen, miten hyvin se toimii? Venablen ym. (2016) mukaan artefaktia voidaan evaluoida joko luonnollisessa tai keinotekoisessa ympäristössä. Luonnollisella ympäristöllä tarkoitetaan esimerkiksi todellista organisaatiota tai tilannetta. Keinotekoisella ympäristöllä tarkoitetaan esimerkiksi testiympäristöä, jossa artefaktin toimintaa simuloidaan ilman todellista käyttökohdetta ja todellisia käyttäjiä.

Venable ym. (2016) esittävät neljä evaluointistrategiaa:

- Nopea ja yksinkertainen, joka soveltuu yksinkertaisten artefaktien arviointiin, joilla alhainen sosiaalinen ja tekninen vaikuttavuus.
- Ihmis- ja vaikutuskeskeinen, joka soveltuu artefaktien arviointiin, jotka ovat riippuvaisia käyttäjien toiminnasta.
- Tekniikka- ja tehokkuuskeskeinen, joka soveltuu artefaktien arviointiin joiden suurin hyöty muodostuu teknisen ulottuvuuden kautta.
- Täysin tekninen, joka soveltuu teknisten tai tulevaisuudessa hyödynnettävien artefaktien arviointiin.

Strategian jälkeen tulee valita tarkempi menetelmä. Menetelmät mukailevat muiden tieteenalojen tutkimusasetelmia. Hevner ym. (2004) jakavat artefaktien evaluointimenetelmät tarkkaileviksi, analyyttisiksi, kokeellisiksi, testaaviksi ja kuvaileviksi. Menetelmät on esitelty taulukossa 2.

TAULUKKO 2 Suunnittelututkimuksen evaluointimenetelmät (Hevner ym., 2004).

Menetelmä	**Tarkenne**
Tapaustutkimus	Artefaktia arvioidaan todellisessa ympäristössä
Kenttätutkimus	Artefaktia monitoroidaan useissa projekteissa
Staattinen analyysi	Artefaktin rakennetta tutkitaan ja selvitetään sen staattisia ominaisuuksia, kuten monimutkaisuutta
Arkkitehtuurianalyysi	Arvioidaan artefaktin sopivuutta tekniseen tietojärjestelmään
Optimointi	Osoitetaan artefaktille ominaiset optimaaliset ominaisuudet tai rajoitteet

(jatkuu)

Taulukko 2 (jatkuu)

Dynaaminen analyysi	Artefaktia tutkitaan käytössä ja selvitetään sen dynaamisia ominaisuuksia, kuten suorituskykyä
Kontrolloitu koe	Artefaktia arvioidaan kontrolloidussa ympäristössä
Simulaatio	Artefaktia käytetään keinotekoisella datalla
Funktionaalinen testaus	Artefaktia käytetään virheiden ja vikojen etsintään
Rakenteellinen testaus	Artefaktia testataan jollakin teknisellä mittarilla
Looginen argumentti	Teoreettisen taustatiedon avulla muodostetaan vakuuttava argumentti, jolla perustellaan artefaktin käytettävyyttä
Skenaario	Artefaktin ympärille muodostetaan yksityiskohtainen skenaario, jolla perustellaan sen käytettävyyttä

Peffers, Rothenberger, Tuunanen ja Vaezi (2012) ovat selvittäneet, miten eri artefaktityyppejä on evaluoitu. Heidän mukaansa tyypillisesti suunnittelututkimuksen avulla kehitetty artefakti on algoritmi, jota on evaluoitu teknisen kokeen avulla, joka mittaa algoritmin suorituskykyä kokeellisissa olosuhteissa, ei niinkään todellisessa maailmassa. Kuvailevat skenaariot ovat teknisten kokeiden jälkeen seuraavaksi käytetyin evaluointimenetelmä ja sitä on hyödynnetty kaikkien eri artefaktityyppien arvioinnissa. Kuvailevassa skenaarioissa artefaktin käyttöä sovelletaan keinotekoisissa tai todellisissa tilanteissa osoittaakseen sen käyttökelpoisuutta (Peffers ym., 2012).

Marchin ja Smithin (1995) mukaan menetelmän valinnan lisäksi artefaktin evaluointiin tulee kehittää metriikka, jolla artefaktia mitataan. Mitattavat kriteerit ovat jokaiselle artefaktille omanlaiset, mutta kriteerien määrittelyssä voidaan hyödyntää valmiiksi kehitettyjä mittaristoja (Venable ym., 2016). March ja Smith (1995) esittävät mitattaviksi kriteereiksi: kokonaisvaltaisuuden, helppokäyttöisyyden, vaikuttavuuden, tehokkuuden, eleganssin, uskollisuuden todellisen maailman ilmiöihin, yleisyyden, vaikutus ympäristöön ja käyttäjiin, sisäisen johdonmukaisuuden, yksityiskohtaisuuden, operatiivisuuden, kestävyyden, yksinkertaisuuden ja ymmärrettävyyden. Smithson ja Hirscheim (1998) esittävät kriteereiksi: laatu-tehokkuuden, resurssitehokkuuden ja ymmärrettävyyden. Rosemann ja Vessey (2008) esittävät artefaktin arviointikriteereiksi: tärkeyden, soveltuvuuden ja saavutettavuuden. Aierin ja Fisherin (2010) esittämiä kriteerejä ovat: käyttökelpoisuus, sisäinen johdonmukaisuus, ulkoinen johdonmukaisuus, laaja soveltuvuus, yksinkertaisuus ja hedelmällisyys uusille löydöksille. Wixom ja Watson (2010) esittävät liiketoimintatiedustelun hyötyjä tuottaviksi kriteereiksi: rahan ja ajan säästämisen, paremman informaation laadun ja määrän, paremmat päätökset, liiketoimintaprosessien kehityksen ja tuen liiketoiminnan strategisten tavoitteiden saavuttamiseksi. Venable ym. (2016) mukaan evaluoinnin tarkoitus on selvittää, miten hyödyllinen kehitetty artefakti on.

Tässä tutkielmassa evaluointistrategiaksi on valittu nopea ja yksinkertainen, koska artefaktia ei tämän tutkielman rajoissa pääse käyttämään tai esittelemään todellisessa ympäristössä. Artefaktin sosiaaliset näkökulmat ovat merkittävät, joten puhtaasti tekninen evaluointi ei ole vaihtoehto. Artefaktin toimintaa evaluoidaan muodostamalla prototyyppijärjestelmän ympärille tiedus-

telun toimintaa kuvailevia skenaarioita (Hevner ym., 2004; Peffers ym., 2012). Evaluoinnin tarkoitus on todistaa artefaktin hyödyllisyys ja tehokkuus (Venable ym., 2012). Tässä tutkielmassa prosessin hyödyllisyyttä ja tehokkuutta arvioidaan laadullisesti seuraavien kriteerien avulla:

- Prosessin avulla voidaan säästää aikaa tai resursseja
- Prosessin avulla voidaan tuottaa tietoa tiedustelun tarpeisiin
- Prosessia voidaan soveltaa erilaisiin lähteisiin tai kohteisiin

Kommunikointi

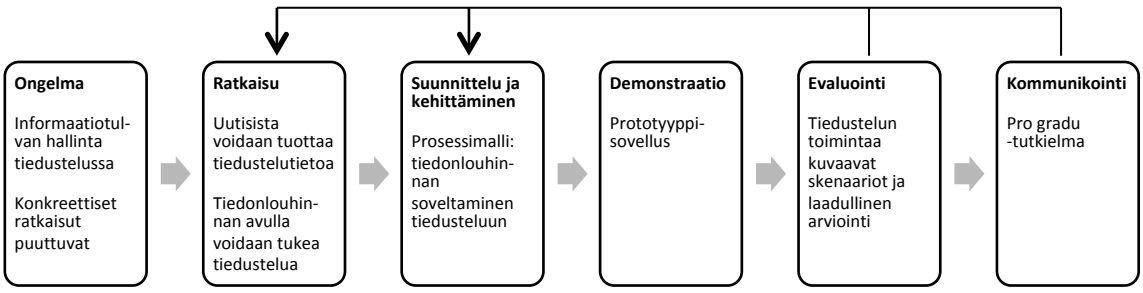
Kuudennessa vaiheessa kommunikoidaan tutkimusongelma, artefakti, sen käyttö ratkaisuna, tehokkuus ja uutuusarvo tiedeyhteisölle ja muille soveltuville yleisöille. Kommunikointi vaatii ymmärrystä tieteenalan kulttuurista. Julkaisuformaatti voi noudattaa tieteenalan tavanomaisia käytäntöjä tai julkaisu voidaan jäsentää tämän kuusivaiheisen prosessin mukaisesti. (Peffers ym., 2007.)

Prosessin vaiheet on muotoiltu etenevän peräkkäisesti, mutta todellisuudessa ei ole tarpeen aloittaa ensimmäisestä vaiheesta, ja edetä vaihe vaiheelta päättäen kuudenteen vaiheeseen. Prosessi voi käynnistyä eri lähtökohdista: ongelma-, tavoite-, suunnittelu ja kehitys- tai tarvelähtöisesti. Prosessi voi myös edetä iteratiivisesti, eli esimerkiksi arvioinnin jälkeen palataan takaisin määrittelemään ratkaisun tavoitteita tai muokkaamaan artefaktin suunnittelua ja kehittämistä. (Peffers ym., 2007.)

Tutkimusprosessi raportoidaan Jyväskylän yliopiston informaatioteknologian tiedekunnan raportointiohjeen mukaisesti ja julkaistaan pro gradu -tutkielmana. Ongelman määrittely, motivaatio, tavoitteet ja mahdolliset ratkaisut esitellään johdannossa. Suunnittelun ja kehittämisen tietopohja, eli kirjallisuuskatsaus esitellään luvuissa kaksi ja kolme. Artefaktin suunnittelu ja kehittäminen sekä sen toiminnan demonstraatio ja evaluointi esitellään luvussa viisi.

Yhteenveto

Tieteellisen suunnittelututkimuksen soveltaminen tässä tutkielmassa on esitetty yhteenvetona seuraavassa kuviossa (kuvio 7).



KUVIO 7 Tieteellisen suunnittelututkimuksen soveltaminen

4.3 Aineisto

Tutkimusaineistoksi on valittu Global Database of Events, Language and Tone (GDELT) uutisdata-aineisto. Aineisto sisältää yli 200 miljoonaa tapahtumaa paikkatietoineen vuodesta 1979 alkaen. Se on muodostettu keräämällä uutistekstejä kansainvälisistä uutislähteistä, kuten AfricaNews, Agence France Presse, BBC Monitoring, Facts on File, Foreign Broadcast Information Service, United Press International, the Washington Post, New York Times, Associated Press ja Google News. (Leetaru & Schrodt, 2013.) Tietokannan toisen version (GDELT 2.0) muodostaminen on aloitettu helmikuusta 2015 alkaen. Toisessa versiossa uutisdataa kerätään yli sadasta lähteestä ja 65 eri kieleltä käännettynä. Tietokannassa on yli 500 miljoona riviä lajiteltua ja luokiteltua uutisdataa, ja päivittäin tietokantaan kerätään noin 250 000 uutta riviä. (GDELT Project, 2021.)

Tietokannan muodostaminen toimii seuraavasti. Uutisteksteistä tunnistetaan tapahtuma, paikka ja sävy algoritmien avulla. Uutistekstistä poimitut tapahtuman lajitellaan ja luokitellaan Conflict and Mediation Event Observations (CAMEO) koodiston avulla (ks. Schrodt, 2012). Koodisto muodostuu monitasoisesta numeraalisesta koodauksesta, jota vastaa sanallinen kuvaus toiminnan luonteesta. Ensimmäisellä tasolla (QuadClass) koodisto jakautuu neljään luokkaan: sanallinen yhteistoiminta, materiaalinen yhteistoiminta, sanallinen konflikti ja materiaalinen konflikti. Toisella tasolla (EventRootCode) on kaksikymmentä luokkaa (taulukko 3), jotka ovat johdettu ensimmäisen tason luokista. Esimerkiksi sanallinen yhteistoiminta sisältää luokat 01–05. Toisen tason luokitusta on vielä täsmennetty kahdella alemmalla tasolla, jotka sisältävät vaihtelevan määrän luokkia. Esimerkiksi kolmannen tason koodi 134 tarkoittaa uhkaamista neuvottelujen lopettamisella ja koodi 138 uhkaamista sotilaallisella voimalla. Neljännen tason koodeja on muodostettu vain osasta kolmannen tason koodeista.

TAULUKKO 3 CAMEO tapahtumakoodit: EventRootCode-taso (Schrodt, 2012).

Tiedustelu	Tiedonlouhinta
01	Antaa julkilausuma
02	Vedota
03	Ilmaista aikomus tehdä yhteistyötä
04	Kysyä neuvoa, konsultoida
05	toteuttaa diplomaattista yhteistyötä
06	toteuttaa materiaalista yhteistyötä
07	tarjota apua
08	antaa apua
09	tutkia
10	vaatia
11	olla erimieltä

(jatkuu)

Taulukko 3 (jatkuu)

12	hylätä
13	uhata
14	protestoida
15	osoittaa sotilaallista voimaa
16	heikentää, loitontaa suhteita
17	pakottaa
18	hyökätä
19	taistella, kamppailla
20	harjoittaa epätavanomaista massamaista väkivaltaa, joukkotuhoa

Muita samankaltaisia uutisdata-aineistoja ovat The KEDS Levant ja ICEWS. Näistä KEDS Levantin ajallinen kattavuus on vastaava kuin GDELT:n, mutta sen tiedot on hankittu vain kahdesta uutislähteestä: Reutersista ja AFP:stä. (Leetaru & Schrodt, 2013.) ICEWS on Yhdysvaltain puolustusministeriön kehittämä tietokanta, joka on tarkoitettu vain viralliseen käyttöön (For official use only), jonka vuoksi sen saatavuus on rajoitettu. GDELT on kaikille avoimesti saatavilla. ICEWS suodattaa, lajittelee ja luokittelee datan tarkemmin kuin GDELT, joka sisältää enemmän dataa, mutta samalla enemmän kohinaa. GDELT on kuitenkin tarkempi tapahtuman paikannuksen suhteen, joka mahdollistaa esimerkiksi alueellisten konfliktien tarkemman seurannan. (Ward ym., 2013.)

Tutkielmassa kehitetty prototyyppisovellus hyödyntää GDELT 2.0 tietokantaa, joka on vapaasti saatavilla Google BigQuery -palvelun kautta. GDELT valittiin aineistoksi, koska se on avoimesti ja ilmaiseksi saatavilla. Lisäksi GDELT on samankaltaisiin aineistoihin verrattuna monipuolisempi ja laajempi. GDELT on tutkielman empiirinen aineisto, jonka avulla muodostetaan yhteys kehitetyn artefaktin ja todellisen maailman välille. Tiedustelun näkökulmasta aineiston valinta on perusteltua, koska strategisen tason tiedustelutietoa voidaan tuottaa avoimista lähteistä. Uutiset ja niistä kerättävä data edustaa hyvin avointen lähteiden tiedustelua ja niiden avulla voidaan tuottaa tietoa strategisen tiedustelun kohteista.

5 TULOKSET

Tutkielman tavoitteena on suunnitella ja kehittää artefakti: prosessimalli tiedonlouhinnan soveltamisesta tiedusteluun. Tässä luvussa esitellään aluksi lyhyesti artefaktin suunnittelu ja kehittäminen, jonka jälkeen demonstroidaan artefaktin toimintaa tuottamalla uutisdatasta tiedustelutietoa prototyyppisovelluksen avulla. Lopuksi evaluoidaan artefaktin käyttökelpoisuutta.

5.1 Tiedonlouhintaprosessin suunnittelu ja kehittäminen

Tiedonlouhintasovelluksen suunnittelun lähtökohtana oli tiedonlouhinta- ja tiedusteluprosessien yhteensovittaminen. Prosessien yhteensovittaminen on luontevaa, koska niissä on hyvin paljon samankaltaisuuksia ja vastaavuuksia. Prosessimallit ja niiden toimintojen vastaavuudet on esitetty taulukossa 4.

TAULUKKO 4 Tiedustelu- ja tiedonlouhintaprosessien vastaavuudet

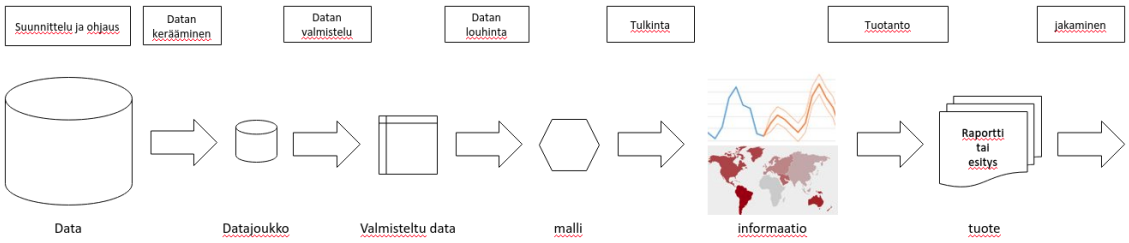
Tiedustelu	**Tiedonlouhinta**
Suunnittelu ja ohjaus	Toimintaympäristön analyysi
Keräys	Datajoukon muodostaminen
Prosessointi	Datan esiprosessointi
Prosessointi	Datan muuttaminen ja ulottuvuuksien vähentäminen
Suunnittelu ja ohjaus	Datan louhintamenetelmän valinta
Suunnittelu ja ohjaus	Algoritmin valinta
Prosessointi	Datan louhinta
Analyysi ja tuotanto	Tulosten tulkinta
Jakaminen	Tulosten hyödyntäminen

Yhteensovittamisen lopputuloksena muodostettiin prosessimallien synteesi:

1. Suunnittelu ja ohjaus: Perehdytään tiedustelutehtävän kannalta oleelliseen toimintaympäristöön ja saatavilla olevaan dataan. Määritellään tiedustelukysymykset sekä valitaan hyödynnettävä datan louhintamenetelmä alustavasti. Määrittelyssä keskeistä on tunnistaa ja artikuloida tie-

dustelun kohde selkeästi: Mistä tietoa halutaan ja miltä aikaväliltä. Suunnittelussa ja ohjauksessa määritellään lisäksi aikamääre tuotannolle, milloin tuote pitää jakaa päätöksentekijälle.
2. Datan kerääminen: Kerätään ja tallennetaan data sen hyödyntämistä varten. Datasta valitaan datajoukko, joka on tehtävän toteuttamisen kannalta relevantti. Mikäli data on ulkoisessa tietovarannossa, tulee kerätty datajoukko tallentaa järjestelmään, jolla datan valmistelu ja louhinta suoritetaan.
3. Datan valmistelu: Valmistellaan data sellaiseen muotoon, jotta datan louhinta on mahdollista.
4. Datan louhinta: Tuotetaan datasta malli algoritmien avulla. Mallinnettu data tai mallin avulla käsitelty data visualisoidaan informaatioksi, joka on ihmisen tulkittavissa.
5. Tulkinta ja tuotanto: Tiedusteluanalyytikko tulkitsee tuotettua informaatiota ja tuottaa sen avulla tiedustelutuotteen jaettavaksi tiedustelun asiakkaalle. Analyytikko voi hyödyntää informaatiota ajattelunsa tukena sekä liittää visualisoituja tuloksia suoraan tiedusteluraportin osaksi.
6. Tiedon jakaminen: Valmis tiedustelutuote jaetaan asiakkaalle, joko dokumenttina tai suullisena ja visuaalisena esityksenä.

Tiedustelun kontekstiin sovellettu tiedonlouhinnan prosessimalli on esitetty kuviossa 8.



KUVIO 8 Prosessimalli tiedonlouhinnan soveltamisesta tiedusteluun

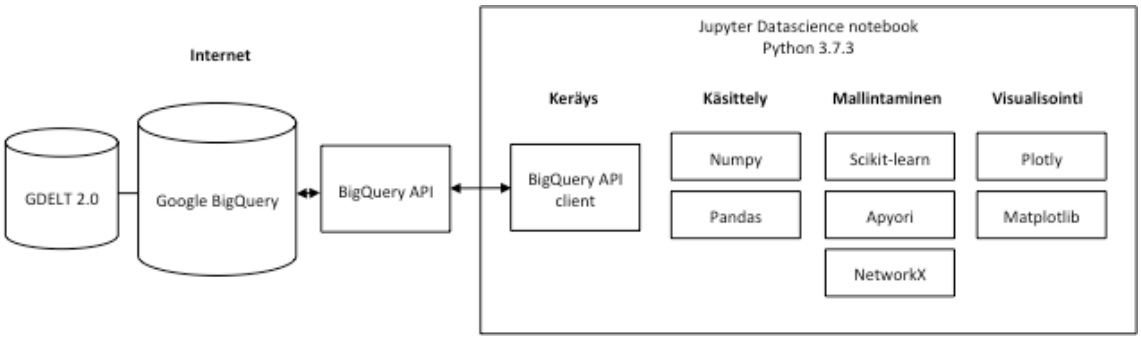
Kirjallisuuskatsauksen perusteella suunniteltua ja kehitettyä kuusivaiheista prosessimallia voidaan hyödyntää tiedonlouhintajärjestelmien ja -sovellusten suunnittelun ja kehittämisen perustana. Seuraavaksi prosessin toimivuutta esitellään erikseen suunniteltavan ja kehitettävän prototyyppisovelluksen avulla. Prototyyppisovelluksen rakentaminen käynnistyi järjestelmän arkkitehtuurin suunnittelulla, joka jäsentää sovelluksen rakennetta ja osien suhdetta toisiinsa. Prosessimallin perusteella voidaan esittää, että tiedonlouhintasovelluksen tulee sisältää seuraavat toiminnot:

- Datan keräys ja tallentaminen
- Datan käsittely ja muokkaaminen
- Datan mallintaminen algoritmien avulla
- Datan visualisointi

Sovelluksen kehitysympäristöksi valittiin Jupyter Notebook[1] ja ohjelmointikieleksi Python. Kehitysympäristöä täydennettiin seuraavilla kirjastoilla:

- BigQuery API client tarjoaa helpon tavan datajoukon keräämiseen ja muodostamiseen varsinaisesta GDELT 2.0 tietokannasta, joka on internetissä Googlen BiqQuery palvelussa. Kirjasto mahdollistaa datajoukon lataamisen ja tallentamisen internetistä kehitysympäristön välimuistiin tai kovalevylle.
- Numpy tarjoaa tuen suurien ja moniulotteisten datamatriisien käsittelylle ja laskennalle.
- Pandas tarjoaa tuen datamatriisien nopealle käsittelylle. Kerätty datajoukko tallennetaan kirjaston dataframe-tietorakenteeseen.
- Scikit-learn tarjoaa valikoiman datan louhinnassa hyödynnettäviä koneoppimisen algoritmeja, joiden avulla voidaan luokitella ja klusteroida dataa.
- Apyori tarjoaa helppokäyttöisen ja yksinkertaisen Apriori algoritmin, jota hyödynnetään toistuvien rakenteiden tunnistamisessa.
- NetworkX:n avulla voidaan mallintaa data graafiksi ja tehdä erilaisia verkostoanalyyseja.
- Plotly ja Matplotlib ovat visualisointiin tarkoitettuja kirjastoja. Niiden avulla voidaan tuottaa erilaisia kuvaajia datasta.

Tiedonlouhintasovelluksen järjestelmäarkkitehtuuri on esitelty kuviossa 9.



KUVIO 9 Tiedonlouhintasovelluksen arkkitehtuuri

5.2 Tiedonlouhintaprosessin käyttökelpoisuuden demonstrointi

Tässä luvussa demonstroidaan kehitetyn prosessin toimivuutta käyttämällä sen perusteella kehitettyä prototyyppisovellusta strategisen tiedustelun kontekstis-

[1] Jupyter Notebook on avoimeen lähdekoodiin perustuva selainpohjainen sovellus, jonka avulla voidaan kirjoittaa, muokata, ajaa ja esittää eri ohjelmointikielillä tuotettua koodia (ks. https://jupyter.org/).

sa. Demonstraatiossa lähestytään kontekstia tiedustelun kohteiden kautta - mistä asioista tai ilmiöistä strateginen tiedustelu tuottaa tietoa? Keskeisimmät strategisen tiedustelun kohteet ovat muodostettu luvun 2.2 yhteenvetona ja ne ovat seuraavat:

- Terrorismi
- Valtioiden sisäiset konfliktit
- Valtioiden väliset suhteet ja konfliktit
- Taloudelliset ja sosiaaliset uhkat
- Joukkotuhoaseet, asekaupat ja sotatarvikkeiden levittäminen
- Kansainvälinen rikollisuus

Keskeisimmistä strategisen tiedustelun kohteista johdettiin viisi skenaariota, joiden ympärille demonstraatiot muodostettiin. Skenaarioissa pyrittiin kuvaamaan mahdollisimman kattavasti erilaisia tiedustelun kohteita ja hyödyntämään erilaisia datan louhintamenetelmiä. Skenaarioiden muodostamista edelsi alustava data-analyysi, jossa selvitettiin löytyykö GDELT-aineistosta tietoa mahdollisista kohteista. Alustavassa data-analyysissä selvisi, että joukkotuhoaseista ja kansainvälisestä järjestäytyneestä rikollisuudesta ei ollut saatavilla soveltuvaa dataa, jotta valinta tiedustelun kohteeksi olisi järkevää. Lopulliset viisi skenaariota on esitelty taulukossa 5.

TAULUKKO 5 Skenaariot, tiedustelutehtävät ja niissä hyödynnetyt menetelmät

Nro.	Tehtävä	Menetelmä
1.	Terrorismin alueellinen tarkastelu	Klusterointi
2.	Valtioiden sisäiset konfliktit	Tutkiva data-analyysi
3.	Valtioiden väliset konfliktit	Toistuvien kaavojen tunnistaminen
4.	Valtioiden sotilaallisen tuen verkostot	Tutkiva data-analyysi
5.	Poikkeman tunnistaminen valtioiden välisistä suhteista	Luokittelu

Seuraavaksi esitellään tarkemmin jokainen skenaario. Niiden esittely on jäsennetty kehitetyn prosessimallin vaiheiden 1–5 mukaisesti. Tiedon jakamista ei käsitellä, koska prosessin avulla ei muodostettu lopullisia tiedustelutuotteita. Tämä rajaus on perusteltua, koska lopulliset tiedustelutuotteet laaditaan ja jaetaan ihmisen toimesta ja tässä työssä keskitytään teknisen tiedonlouhintaprosessin hyödyntämiseen osana tiedusteluprosessia.

5.2.1 Skenaario 1: Terrorismin alueellinen tarkastelu

Ohjaus ja suunnittelu

Tiedustelutehtävänä on hankkia tietoa terrorismin alueellisesta jakautumisesta ja keskittymisestä. Tehtävän tarkastelujakso rajattiin vuoteen 2020. Tehtävän perustella muodostettiin tiedustelukysymys: Mitkä olivat terroristisen toiminnan keskeisimmät tapahtuma-alueet vuonna 2020? Aluksi selvitettiin CAMEO-

koodistosta, mitä terrorismiin liittyviä toimijoita tai tapahtumatyyppejä GDELT-aineistosta on mahdollista kerätä. CAMEO-koodistossa ei ole eriteltynä uusimpia terroristiorganisaatioita, kuten ISIS tai Daesh, mutta vanhemmat organisaatiot, kuten Al Qaeda ja Taleban löytyvät koodistosta. Koodistosta löytyy myös yleisempi luokitus terroristiselle toimijalle: "TERRORIST". Alustavan data-analyysin kautta selvisi, että yleisen luokituksen avulla oli löydettävissä huomattavasti enemmän tapahtumia, kuin organisaation mukaisella luokituksella. Jokaisesta tapahtumasta löytyy sijaintitiedot latitudina ja longitudina, sekä paikan nimenä. Tehtävän toiminta-ajatuksena on, että terrorismin keskittymiä voidaan tunnistaa yhdistämällä tapahtumia, joiden sijainti on lähellä toisiaan. Mitä enemmän tapahtumia on samalla alueella, sitä tiheämmän klusterin ne muodostavat. Datan louhintamenetelmäksi valittiin tällä perusteella tiheyspohjainen klusterointi.

Datan Keräys

Tietokannasta ladattiin seuraavat muuttujat:

- aika (SQLDATE)
- toimija (Actor1Name)
- toiminnan kohde (Actor2Name)
- tapahtumapaikan nimi (ActionGeo_FullName)
- tapahtumapaikan latitudi (ActionGeo_Lat)
- tapahtumapaikan longitudi (ActionGeo_Long)

Rivit rajattiin aikarajauksella: 1.1.2020-31.12.2020 ja toimijan nimellä: "TERRORIST". Kerätty datajoukko sisälsi 74329 riviä.

Datan valmistelu

Datajoukko indeksoitiin päivämäärän mukaan ja datajoukosta poistettiin rivit, joissa muuttujien latitudi tai longitudi arvo oli tyhjä. Tyhjien arvojen poistamisen jälkeen datajoukkoon jäi 70444 riviä.

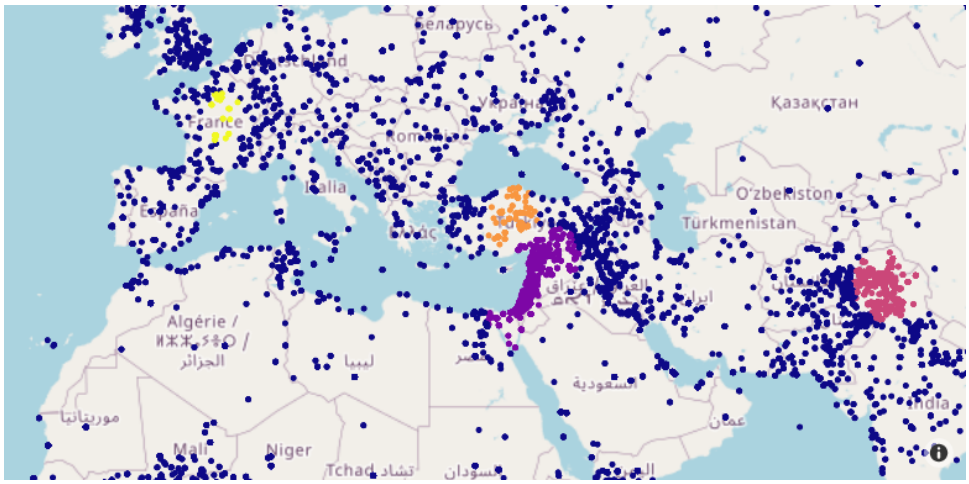
Datan louhinta

Datan louhinnassa käytettiin klusterin tiheyteen perustuvaa Density-based spatial clustering of applications with noise (DBSCAN) -algoritmia (Ester, Kriegel, Sander & Xu, 1996). Algoritmia ohjataan kahdella parametrilla: tarkasteluetäisyydellä (epsilon) ja naapuruston määrällä (minpts). Tarkasteluetäisyys määrittelee jokaisen datapisteen ympärille laskettavan ympyrän säteen pituuden. Jos pisteet ovat tarkasteluetäisyyden päässä toisistaan ovat ne klusterin ydinpisteitä. Naapuruston määrällä säädetään ydinpisteiden reunoilla olevien datapisteiden sisältyvyyttä klusteriin. Klusteriin hyväksytään pisteet, jotka eivät suoraan kuulu ydinjoukkoon, mutta ovat jonkin ydinjoukon pisteen lähellä. (Zaki &

Meira, 2014 s. 375–378.) Algoritmin parametreiksi valittiin testiajojen jälkeen: tarkasteluetäisyys (Epsilon) = 2 ja naapuruston määrä (minpts) = 2000.

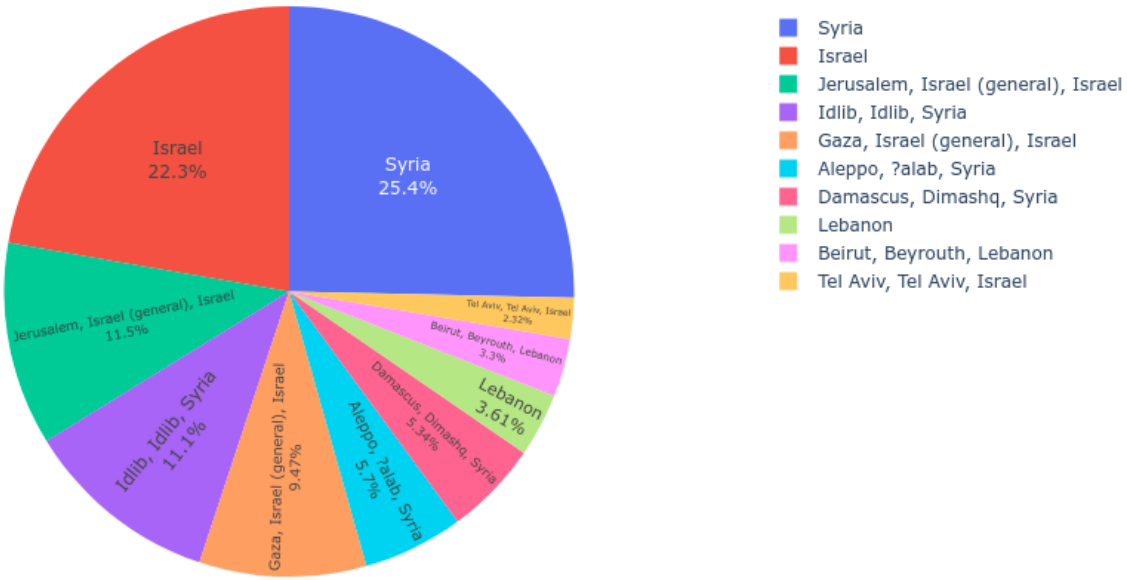
Tulkinta ja tuotanto

Datan louhinta tuotti tulokseksi neljä klusteria, jotka on esitelty seuraavalla karttakuviolla (kuvio 10). Tihein klusteri muodostui Välimeren itärannikolle ja kolme seuraavaa klusteria Pakistanin, Turkin ja Ranskan alueille. Siniset pisteet ovat tapahtumia, jotka algoritmi tulkitsi kohinaksi, koska ne eivät ympäröivien pisteiden kanssa muodostaneet riittävän tiheää klusteria.



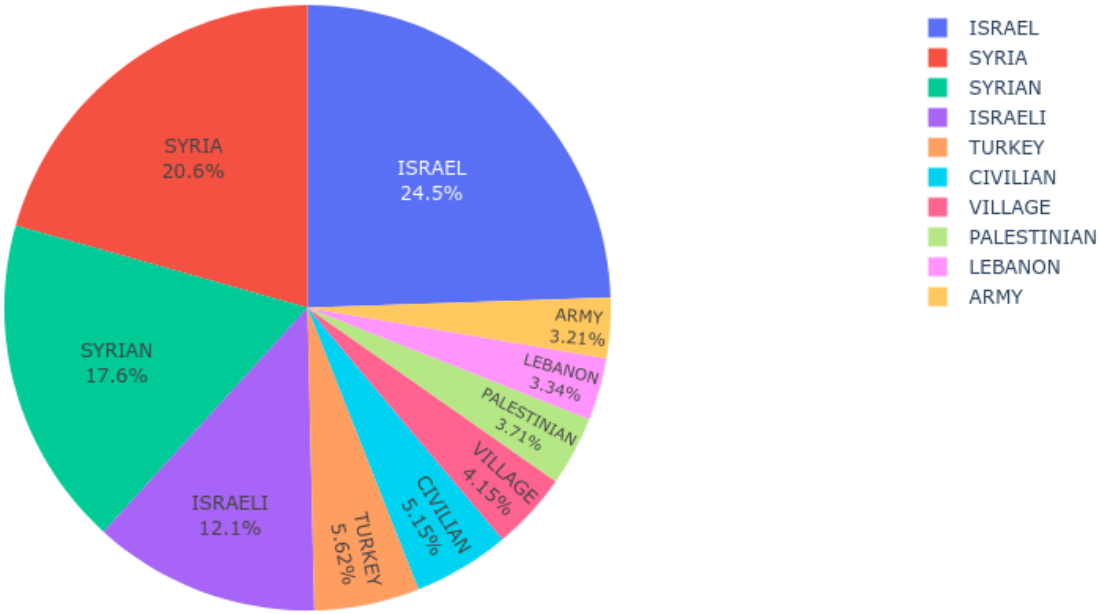
KUVIO 10 Terrorismin alueelliset klusterit

Jokaista muodostettua klusteria voidaan tarkastella irrallaan toisistaan. Erillistarkasteluun valittiin klusteri 1. Karttakuviosta voidaan tulkita yleisesti tapahtumien alueellista levinneisyyttä, mutta tarkempaa tulkintaa varten voidaan tuottaa kuvaaja, jonka avulla hahmotetaan paremmin keskeisiä tapahtuma-alueita. Klusterin 1 keskeisimmät tapahtumapaikat on esitetty kuviossa 11. Kuviosta voidaan tulkita, että keskeisimpiä tapahtuma-alueita ovat Syyrian ja Israelin valtioiden alueet, tarkemmin Jerusalemin, Idlibin ja Gazan alueet.



KUVIO 11 Terrorismin tapahtumapaikat

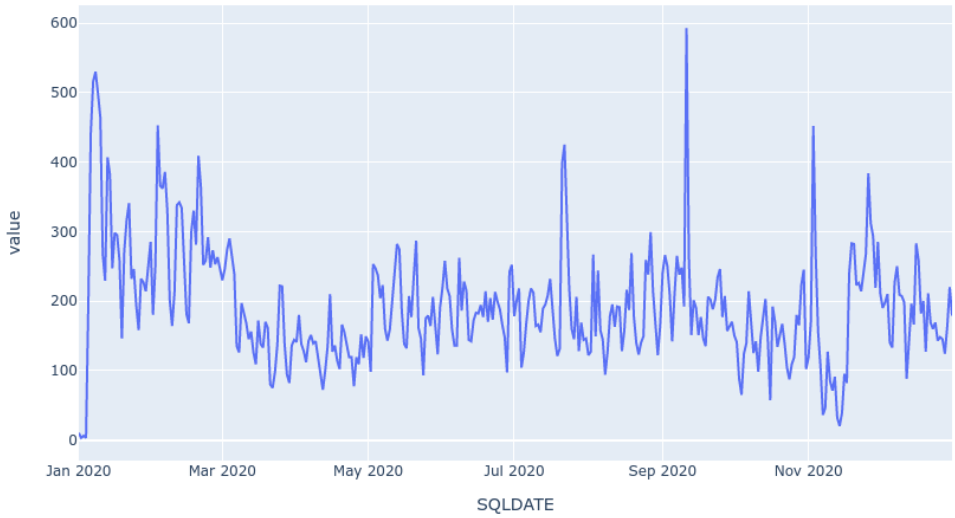
Alueellisen informaation lisäksi klusterista voidaan tarkastella toiminnan kohteita. Klusterin 1 keskeisimmät toiminnan kohteet on esitetty kuviossa 12. Kuviosta voidaan tulkita, että terroristinen toiminta on kohdistunut melko tasaisesti sekä israelilaisiin että syyrialaisiin.



KUVIO 12 Terrorismin kohteet

Lisäksi terroristisen toiminnan intensiteettiä voidaan tarkastella ajan ulottuvuudessa. Klusterin 1 tapahtumien määrällinen intensiteetti on esitetty kuviossa 13. Kuviosta voidaan tulkita, että marras–huhtikuussa ja lokakuun alussa tapahtumien määrä oli keskimääräistä vähäisempää. Aikasarjalta on tunnistet-

tavissa selkeitä lyhyitä intensiivisiä jaksoja, jotka todennäköisesti selittyvät terrori-iskuilla, jotka ovat saaneet suurta mediahuomiota.



KUVIO 13 Terroristisen toiminnan intensiteetti aikasarjalla

5.2.2 Skenaario 2: Valtioiden sisäiset konfliktit

Ohjaus ja suunnittelu

Tiedustelutehtävänä on tuottaa päivittäistä tietoa valtioiden sisäisistä konflikteista. Tarkastelujaksoksi valittiin sattumanvaraisesti päivämäärä 31.5.2020. Tehtävän perustella muodostettiin tiedustelukysymys: Missä valtioissa oli käynnissä sisäinen konflikti 31.5.2020? Tehtävän selvittämiseksi päätettiin hyödyntää tutkivaa data-analyysiä, koska datajoukosta löytyi soveltuvat muuttujat sisäisten konfliktien tunnistamiseksi. QuadClass-muuttujan avulla voidaan poimia konfliktia kuvaavat tapahtumat ja yksinkertaisella laskutoimituksella voidaan muodostaa indeksi, joka kuvaa valtion sisäisen konfliktin. Monimutkaisia dataa mallintavia algoritmeja ei tässä tapauksessa ole tarvetta hyödyntää.

Datan kerääminen

Tietokannasta ladattiin muuttujat:

- aika (SQLDATE)
- toimija (Actor1CountryCode)
- toiminnan kohde (Actor2CountryCode)
- tapahtuman luokka (QuadClass)
- tapahtumapaikan nimi (ActionGeo_FullName)

Rivit rajattiin aikarajauksella: 31.5.2020 ja ehdolla Actor1CountryCode = Actor2CountryCode.

Datan valmistelu

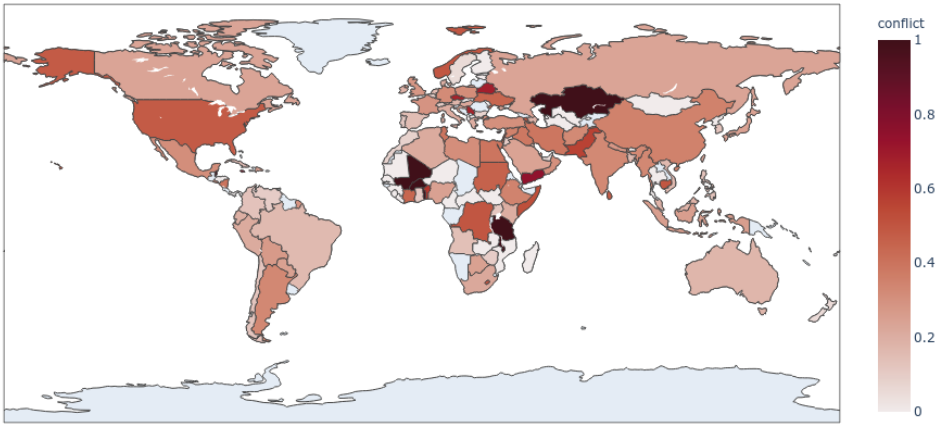
Datajoukosta poistettiin rivit, joissa Actor1CountryCode tai Actor2CountryCode olivat tyhjiä. Datajoukko indeksoitiin päivämäärän mukaan. QuadClass purettiin arvon mukaan kahdeksi binääriseksi muuttujaksi: QuadClass = 1 tai 2 muodostivat muuttujan ”yhteistyö” ja QuadClass = 3 tai 4 muodostivat muuttujan ”konflikti”.

Datan louhinta

Jokaiselle valtiolle laskettiin konflikti-indeksi, joka muodostettiin jakamalla valtion päivittäiset konfliktitapahtumat valtion päivittäisten tapahtumien kokonaismäärällä.

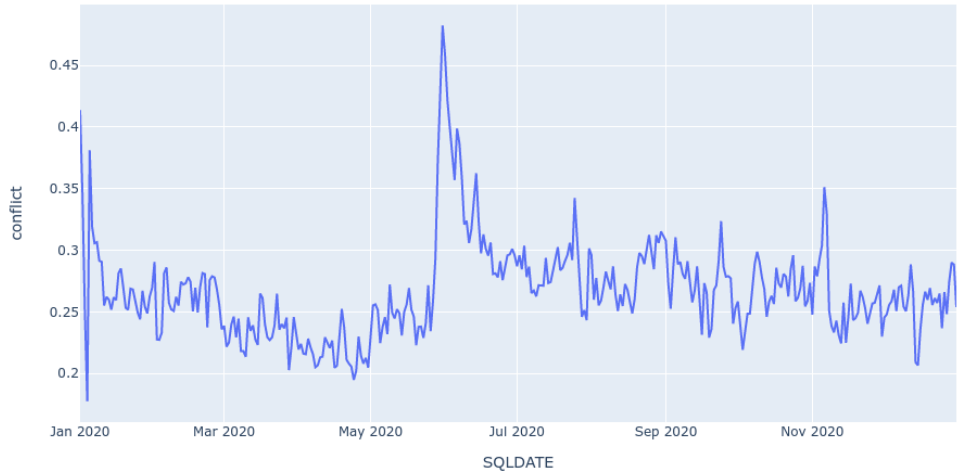
Tulkinta ja tuotanto

Valtioiden sisäisiä konflikteja voidaan tulkita visualisoimalla konflikti-indeksit karttakuvaajana (kuvio 14).



KUVIO 14 Valtioiden sisäiset konfliktit 31.5.2020

Konfliktin kontekstin hahmottamista voidaan parantaa tarkastelemalla tietyn valtion sisäisen konfliktin tasoa pidemmällä aikavälillä. Erityistarkasteluun valittiin Yhdysvallat, jonka sisäisen konfliktin astetta tarkasteltiin ajamalla tiedonlouhintaprosessi uudelleen aikarajauksella 1.1.2020-31.12.2020. Yhdysvaltojen sisäiset konfliktit vuonna 2020 visualisoitiin aikasarjaksi, joka on esitetty kuviossa 15.



KUVIO 15 Yhdysvaltojen sisäiset konfliktit vuonna 2020

Toukokuun lopulla on tunnistettavissa selkeä piikki, joka selittyy George Floydin kuolemaa seuranneista levottomuuksista ja mellakoista ympäri Yhdysvaltoja (ks. Heiskanen, 2021).

5.2.3 Skenaario 3: valtioiden väliset konfliktit

Ohjaus ja Suunnittelu

Tiedustelutehtävänä on hankkia päivittäin tietoa valtioiden välisistä konflikteista ja tuottaa raportti päivittäin. Tarkastelujaksoksi valittiin sattumanvaraisesti päivämäärä 31.5.2020. Tehtävän perustella muodostettiin tiedustelukysymys: Mitkä valtiot olivat konfliktissa keskenään 13.7.2020 ja millainen konflikti on kyseessä? Tehtävän selvittämiseksi päätettiin hyödyntää toistuvien kaavojen tunnistamista. Päivittäin ladattavasta datasta valitaan konfliktiksi luokiteltavat tapahtumat ja tunnistetaan mitkä kaavat toistuvat muuttujissa: toimija, tapahtuman luokitus ja toiminnan kohde.

Datan kerääminen

Tietokannasta ladattiin muuttujat:

- toimija (Actor1CountryCode)
- toiminnan kohde (Actor2CountryCode)
- tapahtuman luokitus (EventRootCode)

Rivit rajattiin aikarajauksella: 13.7.2020 sekä tapahtuman luokituksen perusteella sisältämään vain konflikteja käsittelevät tapahtumat (QuadClass = 3 tai 4).

Datan valmistelu

Datajoukosta poistettiin rivit, joissa toimija (Actor1CountryCode) on sama kuin toiminnan kohde (Actor2CountryCode). Lisäksi poistettiin rivit, joissa toimija

tai toiminnan kohde olivat tyhjiä. Datan valmistelun jälkeen datajoukkoon jäi 6852 riviä.

Datan louhinta

Datan louhinnassa käytettiin Agrawalin ja Srikantin (1994) kehittämää assosiaatiosääntöjen ja toistuvien kaavojen tunnistamiseen tarkoitettua Apriori algoritmia. Algoritmin avulla datajoukosta lasketaan kolme arvoa: tuki (support), luottamus (confidence) ja noste (lift). Tuki selittää kohteiden esiintyvyyttä datajoukossa. Jos tuen arvo on esimerkiksi 0.2, esiintyy kyseinen kohde 20 % kaikista datajoukon riveistä. Luottamus selittää kohteiden esiintyvyyttä yhdessä. Jos luottamuksen arvo on esimerkiksi 0.2, esiintyvät kohteet A ja B 20 % riveissä, joissa kohde A esiintyy. Noste selittää toistuvan säännön voimakkuutta. Algoritmin parametreja ovat miniarvot tuelle, luottamukselle ja nosteelle. Parametreiksi valittiin testiajojen jälkeen:

- Tuki: 0.01
- Luottamus: 0.1
- Noste: 1

Tulkinta ja tuotanto

Louhinta tuotti tulokseksi viisi riviä, joista on tunnistettavissa kaksi valtioiden välistä konfliktia: Armenian ja Azerbaidzhanin välillä sekä Kiinan ja Yhdysvaltojen välillä. Armenian ja Azerbaidzhanin välillä on tunnistettavissa taistelua (EventRootCode = 19) ja erimielisyyttä (EventRootCode = 11). Kiinan ja Yhdysvaltojen välillä on tunnistettavissa lievempi konflikti, joka sisältää erimielisyyttä (EventRootCode = 11), suhteiden heikentämistä (EventRootCode = 16) ja pakottamista (EventRootCode = 17). Algoritmin tuottamat tulokset on lajiteltu laskevasti tuen mukaan ja esitetty taulukossa 6.

TAULUKKO 6 Valtioiden väliset konfliktit 13.7.2020

Toistuva kaava	Tuki	Luottamus	Noste
19, ARM, AZE	0.052	0.210	2.438
11, ARM, AZE	0.027	0.103	1.199
CHN, 16, USA	0.016	0.242	3.949
CHN, 11, USA	0.013	0.316	1.199
CHN, 17, USA	0.012	0.418	1.583

5.2.4 Skenaario 4: Valtioiden sotilaallisen tuen verkostot

Ohjaus ja suunnittelu

Tiedustelutehtävänä on hankkia tietoa valtioiden antamasta ja vastaanottamasta sotilaallisesta tuesta, kuten asekaupoista ja sotatarvikkeiden levittämisestä. Tehtävän tarkastelujakso rajattiin vuoteen 2020. Tehtävän perustella muodos-

tettiin tiedustelukysymys: Mitkä valtiot tukivat toisia valtioita sotilaallisesti vuonna 2020, ja mitkä olivat tuetut valtiot? Aluksi selvitettiin soveltuva tapahtumaluokka, jonka perusteella rajattiin käsiteltävä datajoukko. CAMEO-koodisto sisältää koodin 072, joka käsittää sotilaallisen- tai poliisivoiman antamisen, sisältäen aseet ja henkilöstön. Tehtävän selvittämiseksi päätettiin hyödyntää verkosto-analyysiä, joka voidaan luokitella tutkivaksi data-analyysiksi.

Datan kerääminen

Ladataan tietokannasta muuttujat:

- Toimija (Actor1CountryCode)
- Toiminnan kohde (Actor2CountryCode)

Rivit rajataan aikarajauksella 1.1.2020-31.12.2020 ja tapahtumakoodilla 072.

Datan valmistelu

Poistetaan rivit, joissa toimija ja toiminnan kohde ovat samoja. Esimerkiksi rivit, joissa Yhdysvallat antaa sotilaallista apua itselleen. Lasketaan yhteen tapahtumien määrä ja muodostetaan uusi ulottuvuus: painoarvo. Tuotoksena saadaan datajoukko, jossa on kuvattu yhteistoiminnan suunta (toimija–toiminnan kohde) sekä painoarvo (weight). Lopuksi painoarvo normalisointiin pakottamalla arvot suurimman ja pienimmän mukaan välille 0–1. Valmisteltu datajoukko sisälsi 2608 riviä. Datajoukon ensimmäiset ja viimeiset viisi riviä on esitelty kuviossa 16.

	Actor1CountryCode	Actor2CountryCode	weight
0	TUR	LBY	1.000000
1	TUR	SYR	0.743075
2	RUS	ITA	0.349723
3	CHN	IRQ	0.323407
4	USA	AFG	0.268698
...	...	...	...
2603	SGP	CHE	0.000000
2604	SGP	PHL	0.000000
2605	HTI	MDG	0.000000
2606	SGP	POL	0.000000
2607	SYR	MMR	0.000000

2608 rows × 3 columns

KUVIO 16 Valtioiden antama sotilaallinen tuki vuonna 2020.

Datan louhinta

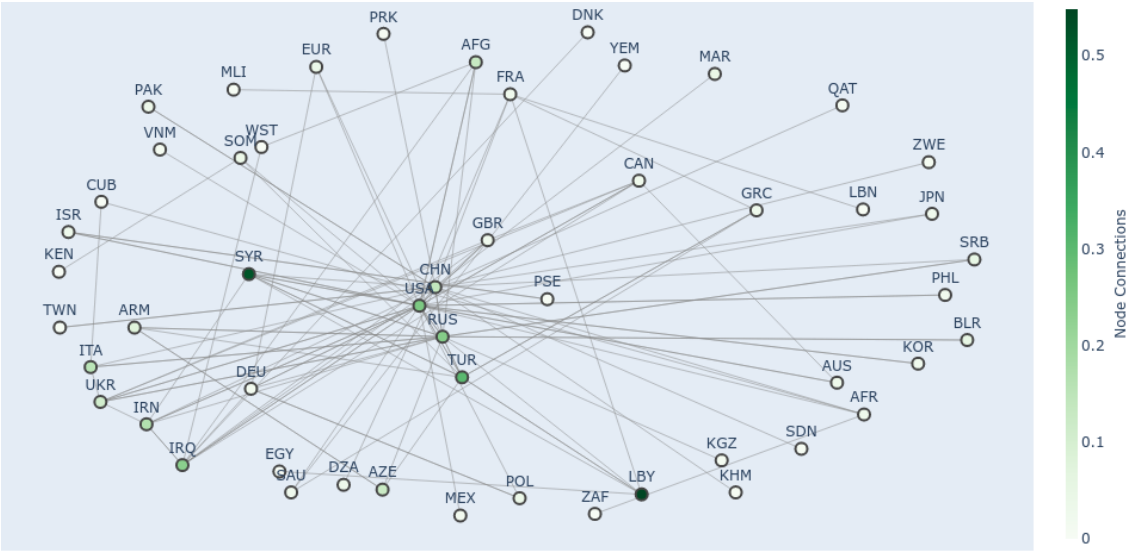
Datan louhinnassa hyödynnettiin verkostoanalyysiä, joka toteutettiin networkx-kirjaston avulla. Datajoukosta muodostettiin aluksia graafi, joka muodostuu solmuista (node) ja niiden välisistä yhteyksistä (edge). Graafin yhteyksillä on painoarvo (weight). Graafiin valittiin rivit, joiden painoarvo on yli 0,03.

Graafin muodostamisen jälkeen laskettiin verkoston keskeisimmät toimijat Bonacichin (1987) esittelemän menetelmän mukaisesti. Menetelmässä solmun keskeisyys määritetään laskemalla yhteen solmun yhteyksien määrä painottaen yhdistettyjen solmujen keskeisyyttä. Keskeisimmät solmut ovat yhdistetty muihin keskeisiin solmuihin.

Lisäksi verkostosta laskettiin Kleinbergin (1999) esittelemän Hyperlink-induced topic Search (HITS) -algoritmin avulla verkoston auktoriteetit ja keskukset. Auktoriteetti on solmu, johon yhdistyy paljon yhteyksiä muista solmuista. Keskus on solmu, josta lähtee ulos useita yhteyksiä. Menetelmän avulla saadaan selville, minkä valtioiden kautta sotilaallista tukea välitetään ja minne sotilaallista tukea toimitetaan.

Tulkinta ja tuotanto

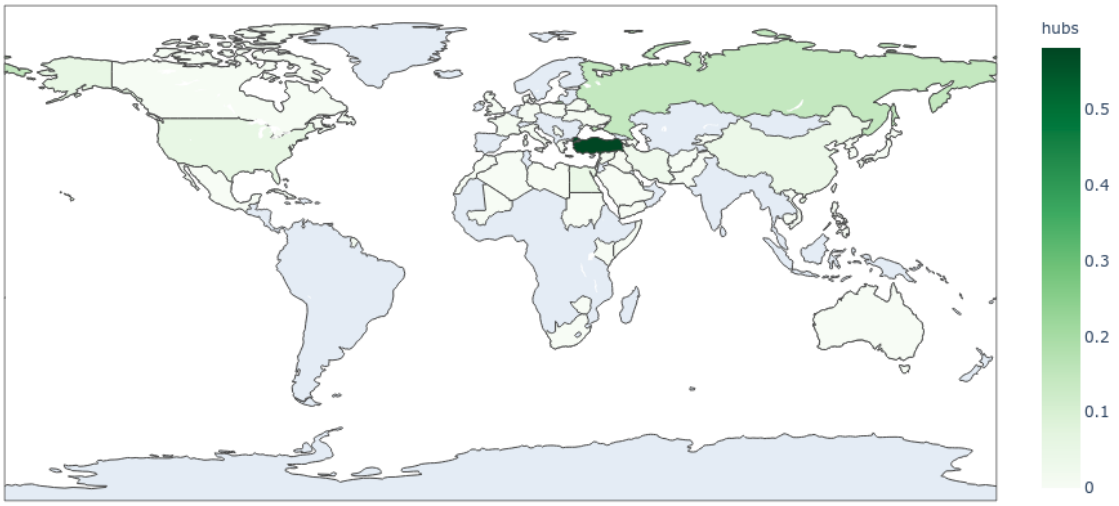
Sotilaallisen yhteistyön verkosto on visualisoitu verkostokuvaajana kuviossa 17. Yhteyksien määrä on visualisoitu vihreän värin tummuudella ja keskeisyys verkoston rakenteelle. Keskeisimmät toimijat on sijoitettu kuvaajan keskelle. Kuvion perusteella voidaan tulkita, että Kiina, Yhdysvallat, Venäjä, Turkki, Iso-Britannia, Palestiina ja Syyria ovat keskeisiä toimijoita, joilla on suhteita useisiin eri toimijoihin. Syyriaan ja Libyaan muodostuu suurin määrä yhteyksiä.



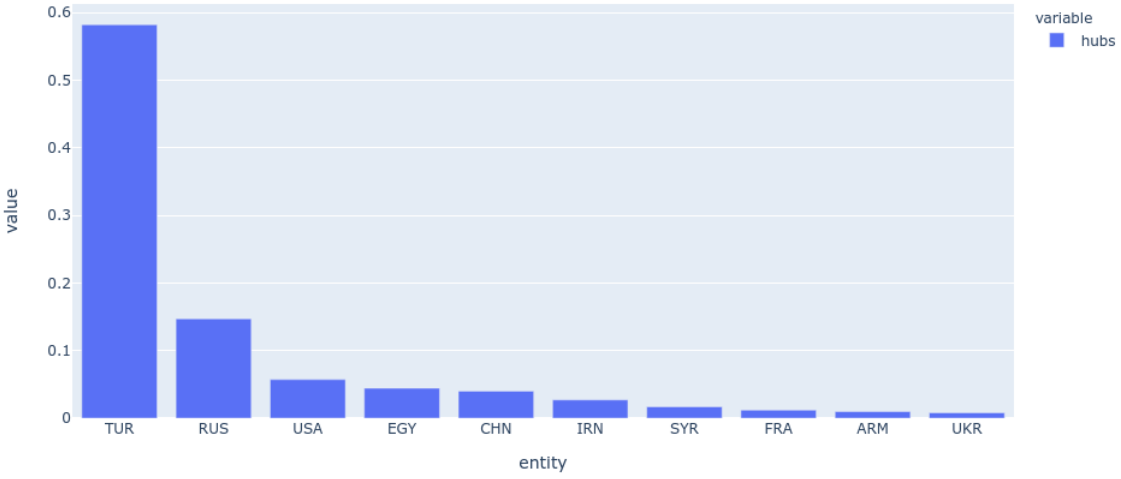
KUVIO 17 Sotilaallisen tuen verkosto vuonna 2020

Verkostokuvaajan lisäksi tuotettiin kartta- ja pylväskuvaajat (kuvio 18, kuvio 19, kuvio 20 ja kuvio 21) visualisoimaan sotilaallisen tuen antajia ja vastaanottajia.

Kuvioista 18 ja 19 voidaan tulkita, että Turkki on ollut keskeisin sotilaallisen tuen antamisen keskus vuonna 2020.

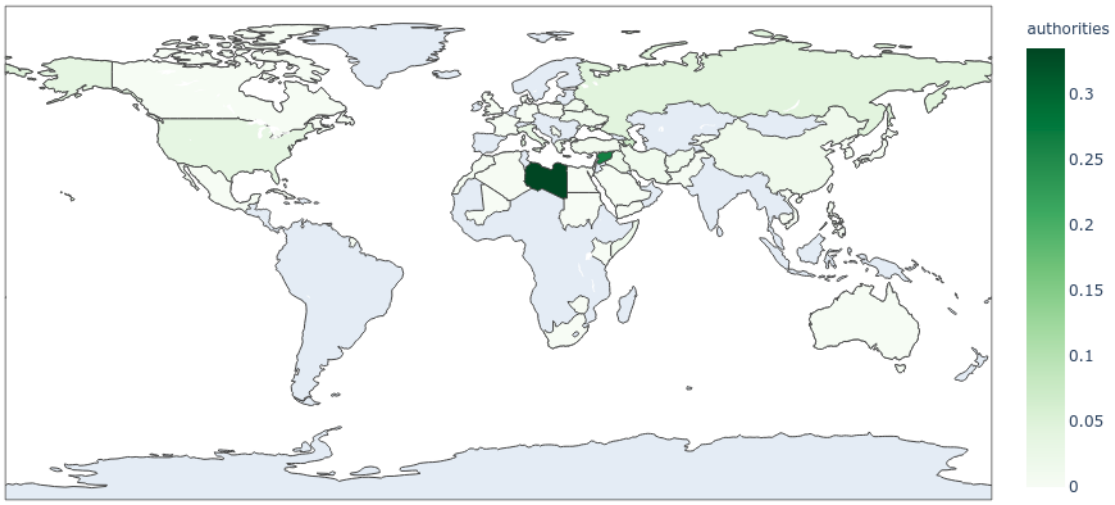


KUVIO 18 Kartta: Sotilaallista tukea antavat valtiot vuonna 2020

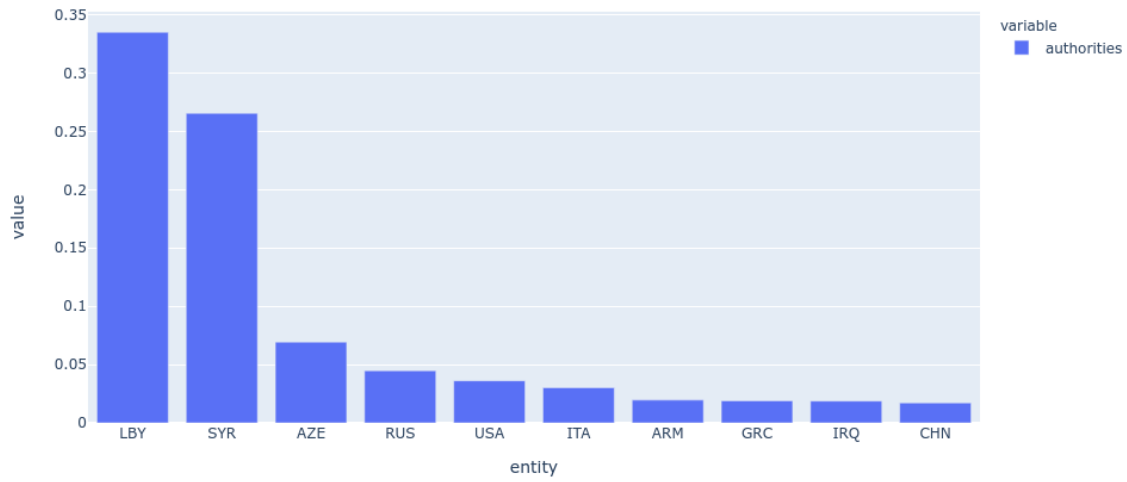


KUVIO 19 Pylväskuvaaja: Sotilaallista tukea antavat valtiot vuonna 2020.

Kuvioista 20 ja 21 voidaan tulkita, että Libya ja Syyria ovat keskeisiä auktoriteetteja, eli valtioita joita on tuettu sotilaallisesti muiden valtioiden toimesta.

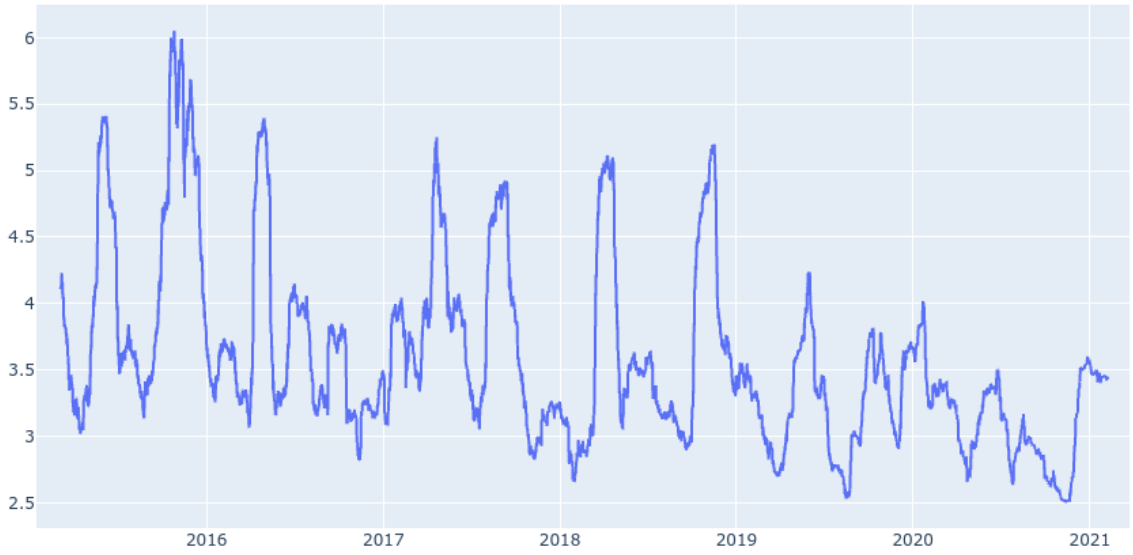


KUVIO 20 Kartta: sotilaallisen tuen vastaanotto vuonna 2020.



KUVIO 21 Pylväskuvaaja: sotilaallisen tuen vastaanotto vuonna 2020.

Turkilla oli ennakko-odotuksia merkittävämpi rooli verrattuna Yhdysvaltoihin ja Venäjään, joten jotta kontekstia voidaan hahmottaa paremmin tuotettiin kaksi aikasarjakuvaajaa, joissa visualisoidaan Turkin (kuvio 22) ja Yhdysvaltojen (kuvio 23) sotilaallisen tuen antamisen määrällistä intensiteettiä vuodesta 2015 alkaen. Kuvioista voidaan tulkita, että Yhdysvaltojen sotilaallisen tuen antaminen on vähentynyt etenkin vuodesta 2019 alkaen. Turkin sotilaallisen tuen antamisessa on taas selkeä piikki vuosien 2019–2020 vaihteessa, joka selittää sen keskeisyyttä vuoden 2020 datassa.



KUVIO 22 Yhdysvaltojen sotilaallisen avun antaminen vuosina 2015-2021

KUVIO 23 Turkin sotilaallisen avun antaminen vuosina 2015-2021

5.2.5 Skenaario 5: Poikkeaman tunnistaminen valtioiden välisistä suhteista

Ohjaus ja Suunnittelu

Tiedustelutehtävänä on hankkia tietoa valtioiden välisistä suhteista ja tunnistaa niistä poikkeamia. Tehtävän tarkastelujakso rajattiin alkamaan vuodesta 2015 ja tarkastelun kohteeksi valittiin Yhdysvaltojen ja Venäjän suhde. Tehtävän perusteella muodostettiin tiedustelukysymys: Miten Venäjän ja Yhdysvaltojen suhde on muuttunut vuodesta 2015 alkaen ja mitä konflikteja valtioiden välillä on tunnistettavissa? Tehtävän selvittämiseksi päätettiin hyödyntää tutkivaa data-analyysiä valtioiden välisten suhteiden kuvailuun ja luokittelua poikkeamien tunnistamiseksi.

Datan kerääminen

Tietokannasta ladattiin muuttujat:

- aika (SQLDATE)
- toimija (Actor1CountryCode)
- toiminnan kohde (Actor2CountryCode)
- tapahtuman luokka (EventRootCode)

Rivit rajattiin aikarajauksella: 19.2.2015, josta alkaen GDELT 2.0 -tietokantaan on alettu keräämään dataa. Lisäksi rivit rajattiin tapahtuman luokituksen perusteella sisältämään vain konflikteja käsittelevät tapahtumat (QuadClass = 3 tai 4).

Datan valmistelu

Datajoukko indeksoitiin päivämäärän mukaan ja poistettiin virheelliset rivit, joiden SQLDATE oli aikaisempi kuin 19. Helmikuuta 2015, ja rivit, joissa Actor1CountryCode oli yhtä kuin Actor2CountryCode. Valmistelun jälkeen datajoukkoon jäi 530257 riviä. Datajoukko ryhmiteltiin päivämäärän mukaan laskemalla jokaiselle päivämäärälle rivien lukumäärä. Näin muodostettiin "konflikti" -muuttuja, joka kuvaa valtioiden välisen konfliktin intensiteettiä. Muuttuja normalisoitiin vähentämällä keskiarvo ja jakamalla keskihajonnalla.

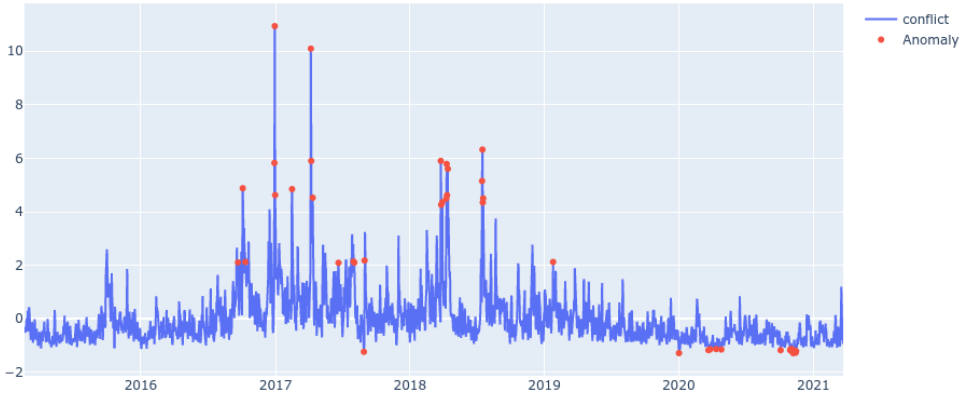
Datan louhinta

Datan louhinnassa käytettiin Schölkopfin, Plattin, Shawe-Taylorin, Smolan ja Williamsonin (2001) kehittämää One-Class SVM -algoritmia. Braein ja Wagnerin (2020) mukaan One-Class SVM soveltuu hyvin poikkeaman tunnistamiseen aikasarjadatasta, joka sisältää vain yhden muuttuja. Heidän kokeidensa perusteella se suoriutui tehtävästä paremmin, kuin monimutkaisemmat syväoppimiseen perustuvat algoritmit, jotka soveltuvat paremmin moniulotteisemman datan käsittelyyn (Braei & Wagner, 2020). One-Class SVM on koneoppimiseen perustuva algoritmi, joka laskee sille syötetystä datajoukosta tukivektorin, johon se vertaa uutta dataa määritellyn raja-arvon mukaan. Poikkeamiksi luokitellaan arvot, jotka ovat laskennallisen ytimen ulkopuolella. Algoritmin parametreja ovat: nu, ydin ja gamma. Nu-arvo määrittää oletettujen poikkeamien määrän. Mitä suurempi arvo, sitä enemmän poikkeamia tunnistetaan. Ydinparametrin avulla määritellään käytettävän laskennallisen ytimen tyyppi. Gamma on ytimen kerroin, jonka avulla säädetään ytimen kokoa. (Schölkopf ym., 2001.) Parametreiksi valittiin testiajojen jälkeen:

- nu: 0.02
- ydin: rbf
- gamma: 0.1

Tulkinta ja tuotanto

Tunnistetut poikkeamat ja valtioiden välisen konfliktin intensiteetti on esitetty kuviossa 24. Algoritmin malli tunnisti 43 poikkeamaa, jotka ovat joko poikkeavan korkeita tai matalia yksittäisiä arvoja.

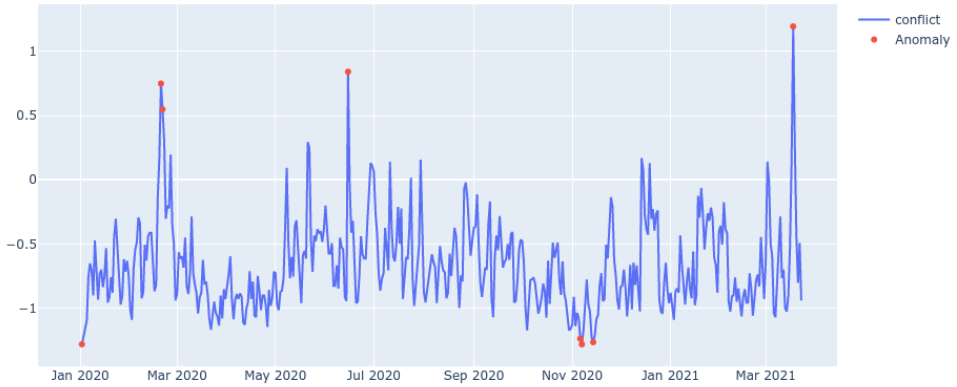


KUVIO 24 Poikkeamat Venäjän ja Yhdysvaltojen suhteissa vuosina 2015–2021.

Korkeimmat poikkeamat ovat 29.–31. joulukuuta vuonna 2016 ja 7.–8. huhtikuuta vuonna 2017. Vuoden 2019 jälkeen trendi on laskeva ja vuonna 2020 tunnistetaan useita poikkeavan alhaisia arvoja.

Ylen aamussa (YLE, 2021) kerrottiin 22.3.2021, että "Venäjän ja Yhdysvaltojen välit kiristyvät." Vaikka YLE uutisoikin välien kiristymisestä, ei algoritmi tunnista nousua vielä poikkeamaksi, koska suhteellisesti maiden välinen konflikti on aikaisempia vuosia alhaisemmalla tasolla.

Datan louhinta suoritettiin uudelleen, opettamalla malli matalamman intensiteetin ajanjaksolla, vuodesta 2020 alkaen. Tulokset ovat esitetty kuviossa 25. Uudelleen opetettu malli tunnisti poikkeaman päivämäärältä 18.3.2021, joka kuvaa valtioiden välisen konfliktin intensiteetin nopeaa nousua.



KUVIO 25 Poikkeamat Venäjän ja Yhdysvaltojen suhteissa vuodesta 2020–kevät 2021.

18.3.2021 tunnistettu poikkeama selittyy Yhdysvaltojen ja Venäjän presidenttien välisestä sanallisesta konfliktista, joka alkoi 17.3.2021, kun Yhdysvaltain presidentti Joe Biden kritisoi Venäjää ja presidentti Putinia, joka vastasi kritiikkiin seuraavana päivänä (ks. Uosukainen, 2021).

5.3 Evaluointi: prosessin hyödyllisyys ja tehokkuus

Tässä luvussa evaluoidaan kehitetyn prosessin hyödyllisyyttä ja tehokkuutta laadullisena arviona kolmen kriteerin näkökulmasta:

- Resurssit: Prosessin avulla voidaan säästää resursseja, kuten aikaa.
- Tiedon laatu: Prosessin avulla voidaan tuottaa tietoa, joka on hyödynnettävissä strategisessa tiedustelussa.
- Yleistettävyys: Prosessia voidaan soveltaa erilaisiin lähteisiin tai kohteisiin.

5.3.1 Resurssit

Tiedonlouhintaprosessin ajaminen vaatii aineellisena resurssina tietokoneen, jossa on riittävästi laskentatehoa ja muistia. Datan koko määrittää resurssien tarpeen. Tässä työssä hyödynnetty data ja menetelmät eivät vaatineet tietokoneelta merkittäviä resursseja. Uutisdata vaatii levyresursseja satoja gigatavuja. Käytetty data oli tallennettu Googlen BigQuery -palveluun, joten erillisiä levyresursseja ei datan tallentamiseen vaadittu, mutta datan kerääminen vaati internet yhteyden. Datan sujuva käsittely vaatii keskusmuistia 8–16 gigatavua. Työssä esiteltyjen algoritmien ajaminen kesti muutamia sekunteja 3,6 GHz:n neliydinprosessorilla. Aineettomana resurssina tiedonlouhinta vaatii keskitason osaamista ohjelmoinnista, datan käsittelystä sekä ymmärrystä tilastotieteestä ja algoritmeista. Käytetyt ohjelmistot olivat kaikki avoimesti ja ilmaiseksi saatavilla, joten rahallisia resursseja ei vaadita. Tiedonlouhintaprosessin ajaminen on nopeaa, mutta sen suunnittelu ja kehittäminen vaatii aikaa. Jokaista skenaariota varten tiedonlouhintaprosessin ajaminen vaati eri toimintojen yhdistämistä ja ohjelmointia. Toimivan ajon suunnitteluun ja kehittämiseen käytettiin aikaa useita tunteja. Monivaiheisissa skenaarioissa (skenaariot 5 ja 6) toimivan ajon suunnittelu ja kehittäminen kesti useita päiviä. Kun järjestelmä on ohjelmoitu, sen ajaminen datan keräyksestä louhintaan saakka kesti kaikissa skenaarioissa vain muutamia sekunteja.

Skenaarioissa käsiteltiin tietokoneen avulla suuri määrä dataa. Vastaavan määrän käsittely olisi ihmiseltä vienyt huomattavasti enemmän aikaa. Esimerkiksi Skenaariossa viisi valtioiden välisen konfliktin intensiteettiä kuvaavan aikasarjakuvaajan luominen olisi vaatinut ihmiseltä yhden vuoden uutistekstien lukemisen, tulkinnan, lajittelun ja luokittelun. Kun tiedustelutehtävän mukainen tiedonlouhinta on suunniteltu ja kehitetty, voidaan se ajastaa esimerkiksi tuottamaan automaattisesti päivittäinen karttakuva tai taulukko edellisen vuorokauden aikana tapahtuneista kiinnostavista ilmiöistä, kuten skenaarioissa kaksi ja kolme. Poikkeaman tunnistamisen kuvaajasta voi ihminen tehdä yhtä nopeasti kuin tietokonekin. Tiedonlouhinnan avulla voidaan kuitenkin säästää resursseja, jos tarkasteltavia kohdevaltioita on useita.

5.3.2 Tiedon laatu

Ensimmäisessä skenaariossa tunnistettiin klustereita toimijan ja sijainnin perusteella koko maailman alueelta. Alueet, joissa havaintoja on enemmän muodostavat oman klusterin. Asiat, jotka tapahtuvat lähellä toisiaan, liittyvät tyypillisesti toisiinsa. Klusteri tarjoaa mahdollisuuden tarkastella toisiinsa liittyviä asioita ilman keinotekoisia rajauksia. Klusterin avulla voidaan paljastaa myös ilmiön todellinen laajuus, kun tarkastelu ei rajoitu esimerkiksi valtioiden rajoihin. Muodostamalla klustereita voidaan rajata datajoukkoja, joita tutkimalla voidaan tarkastella ilmiötä eri näkökulmista, kuten intensiteetin tai toiminnan kohteiden. Tiedonlouhinnan avulla tuotettu kartta ei sellaisenaan ole hyödynnettävissä tiedustelutuotteiden osana, mutta sen avulla voidaan hahmottaa kokonaisuutta ja ohjata tiedusteluanalyysiä tai tiedonhankintaa. Klustereiden dataa eri tavalla visualisoituna voidaan hyödyntää mahdollisesti osana tiedustelutuotteita, mutta paremmin niiden tarkastelu soveltuu tiedusteluanalyysin taustamateriaaliksi.

Skenaariossa kaksi tarkasteltiin valtioiden sisäisiä konflikteja hyvin yksinkertaisen laskennan kautta: Jos valtion sisäisestä toiminnasta kertovat uutiset kertovat konflikteista enemmän kuin yhteistyöstä, voidaan todeta, että valtio on sisäisessä konfliktissa. Päivittäinen kartta konflikteista ei ilman taustakontekstia kerro vielä todellista tarinaa. Kartan rinnalla esitettävä aikasarjakuvaaja, jossa tarkastellaan yksittäisen valtion konflikti-indeksiä pidemmällä aikavälille kuvaa kontekstia paremmin. Tiedonlouhinnan tuottamia visualisointeja voidaan hyödyntää osana tiedustelutuotteita, mutta toimintaa selittävän tarinan kirjoittaminen vaatii perehtymistä konfliktin syihin lukemalla esimerkiksi uutisia.

Skenaariossa kolme selvitettiin toistuvia kaavoja yhden vuorokauden aikana tapahtuneista valtioiden välisistä konflikteista. Tuotetusta informaatiosta voidaan tulkita minkä maiden välillä konflikti on ja mikä on konfliktin luonne: ovatko valtiot eri mieltä, uhkaako toinen valtio toista vai onko valtioiden välillä aseellinen taistelu. Tieto sopii hyödynnettäväksi tiedusteluprosessin vaiheessa ohjaus ja suunnittelu. Sen perusteella voidaan selvittää tarkemmin mitä on tapahtunut. Tietoa ei voida sellaisenaan hyödyntää tiedusteluraportin osana.

Skenaariossa neljä selvitettiin valtioiden välistä sotilaallisen avun antamista. Tiedonlouhinnan avulla tuotettiin verkostokuvaaja, kartta, pylväsdiagrammeja sekä aikasarjakuvaajia. Informaatio on monipuolisesti käyttökelpoista. Visualisointeja voidaan hyödyntää hyvin tarkemmassa analyysissä tai osana tiedustelutuotteita.

Skenaariossa viisi tuotettiin kvantitatiivista tietoa valtioiden välisistä suhteista. Tunnistetun poikkeaman avulla voidaan käynnistää raportointi kohteesta, josta poikkeama on tunnistettu. Poikkeamien visualisointi soveltuukin tiedustelun suunnittelun ja ohjauksen tueksi, kun valitaan kiinnostavia aiheita tilannetietoraporttien, kuten päiväraporttien sisällöksi. Valtioiden välisten suhteiden aikasarjakuvaajaa, jossa poikkeama on esitetty suhteessa aikaisempaan dataan, voidaan hyödyntää hyvin osana tiedustelutuotetta.

5.3.3 Yleistettävyys

Data määrittää sen, millaista tietoa tiedonlouhinnan avulla voidaan tuottaa. Esimerkiksi terroristijärjestö Daeshin toiminnasta ei voitu tuottaa tietoa, koska uutisteksteistä ei ole tunnistettu ja poimittu kyseistä järjestöä datajoukkoon oman entiteettinään.

Skenaariossa yksi tuotettiin klusteroimalla tietoa terroristisen toiminnan alueellisesta jakaantumasta. Vastaavalla tavalla tiedonlouhintaprosessia voidaan soveltaa esimerkiksi signaalitiedustelulla kerättävään paikannusdataan. Jos tietyltä alueelta paikannetaan useita signaaleja, voidaan olettaa, että alueella tapahtuu jotakin merkittävää. Tämänkaltaisten alueellisten keskittymien tunnistaminen voidaan toteuttaa klusteroinnin avulla.

Skenaariossa kaksi tuotettiin tutkivan data-analyysin ja hyvin yksinkertaisten laskennallisten menetelmien avulla konfliktin tasoa kuvaava indeksi. Joissakin tapauksissa ainoastaan yhteen laskemalla ja visualisoimalla kerättyä dataa, voidaan saavuttaa haluttu lopputulos. Jos data sisältää paikkatiedon ja jonkin intensiteettiä kuvaavan muuttujan, voidaan samaa menetelmää hyödyntää erilaisten ilmiöiden kuvaamiseen.

Skenaariossa kolme hyödynnettiin Apriori-algoritmia toistuvien kaavojen tunnistamiseen. Algoritmin avulla voidaan helposti tunnistaa trendejä sekä tunnistaa syy-seuraussuhteita.

Skenaariossa neljä tiedonlouhinnan avulla tehtiin verkostoanalyysiä. Vastaavaa menetelmää voidaan soveltaa tietoliikennetiedustelun avulla kerättyyn tietoon. Verkostoanalyysin avulla voidaan tunnistaa tietoliikenteestä keskeisimpiä toimijoita ja analysoida, kuka viestii ja kenen kanssa.

Skenaariossa viisi hyödynnettiin luokittelua poikkeaman tunnistamisessa. Luokittelun avulla voidaan helposti automatisoida tiedustelutietojen esianalyysiä. Luokittelun avulla voidaan käsitellä uusi kerätty data, tunnistaa siitä mielenkiintoiset asiat ja tarjota niitä tiedusteluanalyytikolle. Luokittelualgoritmi voidaan opettaa erilaisella datalla tunnistamaan poikkeamia eri konteksteista.

Kokonaisuutena voidaan tunnistaa, että tiedonlouhinta soveltuu prosessina erittäin hyvin tiedustelun työvälineeksi. Se on joustava, ja sen avulla voidaan tuottaa hyvin monipuolisesti tietoa erilaisiin tarpeisiin. Tiedonlouhinnan avulla voidaan myös automatisoida tiedusteluprosessin eri toimintoja.

6 TULKINTA JA POHDINTA

Tässä luvussa esitetään tutkimuksen tulokset vastaamalla tutkimuskysymyksiin sekä arvioidaan tulosten käytännöllistä ja tieteellistä merkitystä. Lisäksi arvioidaan menetelmän ja tulosten luotettavuutta sekä niiden käytettävyyttä jatkotutkimuksen näkökulmasta.

Tutkimuskysymys ja kolme alakysymystä ovat: Miten tiedonlouhinnan avulla voidaan tuottaa uutisdatasta tietoa strategisen tiedustelun tarpeisiin?

- Millaista tiedustelutietoa uutisdatasta voidaan tuottaa?
- Mitä resursseja tiedonlouhinnan avulla voidaan säästää?
- Mitä tiedustelun toimintoja tiedonlouhinnan avulla voidaan tukea?

Miten tiedonlouhinnan avulla voidaan tuottaa uutisdatasta tietoa strategisen tiedustelun tarpeisiin?

Uutisdatasta voidaan tuottaa tietoa strategisen tiedustelun tarpeisiin kuusivaiheisen prosessin avulla, joka on kuvattu luvussa 5.1. Prosessin toimivuutta on esitelty viiden skenaarion avulla luvussa 5.2.

Kirjallisuuskatsauksen kautta kävi selväksi, että tiedonlouhinnalla ja tiedustelulla on paljon yhteistä. Molempien tarkoitus on tuottaa tietoa. Tiedustelu tarjoaa kontekstin: miten dataa kerätään ja mitä ovat tiedon tarvitsijoiden kiinnostuksen kohteet. Tiedustelun kirjallisuus esittelee rakenteellisia menetelmiä, joiden avulla datasta tuotetaan tietoa, mutta tietojärjestelmien ja teknisten prosessien hyödyntäminen tiedustelun kontekstissa on varsin vähän tutkittu alue. Tiedonlouhinta tarjoaa mallin, joka toimii tietojärjestelmien ja tiedontuotantoprosessin kehittämisen pohjana. Tiedonlouhinta- ja tiedusteluprosessien yhdistäminen sekä soveltaminen strategisen tiedustelun kontekstiin uutisdata-analyysin kautta esiteltiin viidennessä luvussa. Tutkimuksen tuloksena suunniteltu ja kehitetty prosessimalli on hyvä keskustelunavaus, kun pohditaan miten uutta teknologiaa – tekoälyä, koneoppimista ja syväoppimista – voidaan hyödyntää tiedustelussa.

Tutkimuksen tuloksena suunniteltiin ja kehitettiin prosessimalli, joka vastaa kysymykseen, miten tiedonlouhinnan avulla voidaan tuottaa uutisdatasta

tietoa strategisen tiedustelun tarpeisiin. Mallin toimintaa esiteltiin ja sen käyttökelpoisuutta arvioitiin viiden skenaarion avulla, joissa kuvattiin erilaisia tiedustelutehtäviä ja erilaisten datan louhintamenetelmien hyödyntämistä.

Millaista tiedustelutietoa uutisdatasta voidaan tuottaa?

Soveltamalla prosessia uutisdataan selvisi, että tiedonlouhinnan avulla voidaan tuottaa tietoa strategisen tiedustelun kiinnostuksen kohteista rajoitetusti. Data määrittää sen, millaista tietoa voidaan tuottaa. Huonosta datasta ei saada aikaiseksi laadukasta tietoa. Jos datasta ei löydy tietoa joukkotuhoaseista, ei tietoa voida luonnollisesti tuottaa. Kerätyn datan lajittelulla ja luokittelulla on suuri merkitys. GDELT-aineisto on lajiteltu CAMEO-koodiston avulla, jossa on puutteita ajankohtaisuutensa vuoksi. Esimerkiksi uusimpia terroristiorganisaatioita ei koodistoon ole luokiteltu.

Uutisdatasta voidaan tuottaa helposti tietoa valtioiden välisistä suhteista ja konflikteista pitkällä aikavälillä. Data-analyysin vahvuus on sen kvantitatiivinen lähestymistapa, joka täydentää tiedusteluanalyysia, joka on usein laadullista, kun arvioidaan laajoja asioita, kuten geopolitiikkaa. Uutisdatasta voidaan selvittää valtioiden välisten suhteiden kehittymistä ja visualisoida trendejä aikasarjakuvaajana. Uutisdatasta tiedonlouhinnan avulla tuotettava tieto on melko pintapuolista informaatiota. Tiedustelutuotteiden näkökulmasta tuotettavan tiedon avulla voidaan tukea kaikkien tuotetyyppien – tilanne, perus ja arvioiva – tuotantoa. Esimerkiksi tiedonlouhinnan avulla tuotettavia visualisointeja voidaan liittää osaksi tiedustelutuotteita. Uutisdata muodostuu uutisista, jotka toimittajat ovat kirjoittaneet, joten uuden ja mullistavan tiedon löytäminen uutisdatasta on lähtökohtaisesti paradoksaalista.

Chin ym. (2009) esittelevät, kuinka tiedusteluanalyytikot keräsivät, lajittelivat, luokittelivat ja suodattivat informaatiota, josta he tuottivat graafeja, aikajanoja ja karttoja kuvaamaan kaavoja ja trendejä. Analyytikot työskentelivät paperilla, piirtämällä seinälle tai taulukkolaskentaohjelman avulla. Tässä työssä esitelty prosessi tarjoaa tiedusteluanalyysiin modernimman ja tehokkaamman tavan tuottaa vastaavia tuloksia, sillä tiedonlouhinnan avulla voidaan käsitellä nopeasti suuri määrä dataa.

Rakenteellisten menetelmien avulla voidaan parantaa tiedusteluanalyysin laatua (Pherson & Heuer 2014). Rakenteellisten menetelmien tarkoitus on purkaa asiat pienemmäksi, jonka jälkeen osat kokoamalla voidaan tunnistaa oleelliset tiedot. Purkaminen voidaan ymmärtää tiedon kvantifiointina. Uutisdataan kohdistetun tiedonlouhinnan avulla voidaan kvantifioida geopoliittisia ilmiöitä. Heuristisella arvioinnilla voidaan tulkita esimerkiksi, että Venäjän ja Yhdysvaltojen välit ovat heikentyneet, mutta määrällisen datan avulla voidaan tuottaa aikasarjakuvaaja, kuten skenaariossa: suurvaltojen väliset suhteet, on esitelty. Kuvaajan avulla voidaan vastata kysymykseen, kuinka paljon välit ovat heikentyneet suhteessa historiaan.

Tiedustelun pitäisi kyetä tuottamaan ennakoivaa tai ennustavaa tietoa. Tutkijat ovat pyrkineet kehittämään menetelmiä, jotka ekstrapoloivat uutisdatasta trendejä ja ennustavat tulevaisuutta (Yonamine, 2013; Qiao ym., 2017; Galla & Burke, 2018; Chen ym., 2020). Geopolitiikan ja merkittävien ilmiöiden en-

nustaminen tilastotieteen avulla on kuitenkin ongelmallista, koska tulevaisuus ei noudata täysin historiallista jatkumoa (Taleb, 2007; Tetlock & Gardner, 2010). Tiedustelun näkökulmasta merkityksellisempää on etsiä poikkeamia jatkumoissa (Miller, 2014). Tiedonlouhinnan avulla voidaan tunnistaa poikkeamia esimerkiksi valtioiden välisistä suhteista, kuten skenaariossa viisi on esitelty. Poikkeamien tunnistamista voidaan hyödyntää, kun halutaan löytää kiinnostavat asiat nopeasti suuresta määrästä informaatiota. Sen sijaan, että ihminen lukisi uutismediaa, voidaan toiminto automatisoida tiedonlouhinnan avulla.

Mitä resursseja tiedonlouhinnan avulla voidaan säästää?

Aika voidaan tunnistaa tiedustelun keskeisimmäksi resurssiksi. Kun maailmassa tapahtuu jotakin merkittävää, päätöksentekijät odottavat tiedustelutuotteita nopeasti (George, 2014). Kerättyä dataa ja informaatiota on kuitenkin niin paljon, että analyytikot eivät ehdi sitä käsitellä (Clark, 2019). Aikapaineen vuoksi analyytikot hylkäävät rakenteellisten menetelmien hyödyntämisen (Chin ym., 2009). Tiedonlouhintaprosessin avulla voidaan saavuttaa nopeutta, kun ihmiselle työläät ja hitaat vaiheet automatisoidaan tietokoneen tekemän laskennan avulla. Tiedonlouhinnan avulla voidaan automatisoida tiedusteluprosessin vaiheita ja saavuttaa sekä ajallista että henkilöstöresurssien säästämistä. Sen sijaan, että ihminen lukee läpi kaikki päivän uutiset ja tunnistaa niistä kiinnostavat ilmiöt, voidaan tiedonlouhinnan avulla tunnistaa kiinnostavia kohteita, kuten skenaarioissa kaksi ja kolme on esitelty.

Tiedonlouhinnan avulla voidaan säästää aikaa ja ihmisresursseja, mutta samalla prosessin toimeenpano asettaa omat vaatimuksensa. Tiedonlouhintaprosessin toimeenpano vaatii ohjelmoinnin ja data-analytiikan osaamista sekä tilastotieteen ymmärtämistä. Tiedusteluorganisaatioon tarvitaan ihmisiä, jotka toimeenpanevat tiedonlouhintaprosessia osana tiedusteluprosessia yhteistyössä tiedon kerääjien ja analyytikoiden kanssa. Brucen ja Georgen (2014) mukaan täydellisessä tiedusteluanalyytikossa yhdistyy historioitsija, journalisti, tutkija, keräysmenetelmien asiantuntija ja skeptikko. Tämän tutkielman perusteella voidaan esittää, että listaa jatketaan data-analyytikolla tai tiedonlouhijalla. Tiedusteluorganisaatiot voivat tehostaa toimintaansa palkkaamalla historioitsijoiden, yhteiskuntatieteilijöiden ja lingvistien lisäksi data-analyytikoita. Data-analyytikko ei kuitenkaan korvaa perinteisempien asiantuntijoiden työpanosta, vaan täydentää tiimiä omalla osaamisellaan.

Prototyyppijärjestelmä, joka suunniteltiin ja kehitettiin prosessin esittelyn ja arvioinnin mahdollistamiseksi, tarjoaa esimerkin tiedonlouhinnan vaatimista teknisistä resursseista. Datajoukon koko ja algoritmien monimutkaisuus määrittelevät tietokoneelta vaadittavan levytilan, keskusmuistin ja laskentatehon määrän. Tässä tutkielmassa toteutettu tiedonlouhinta GDELT-uutisdatasta yksinkertaisia koneoppimisen algoritmeja hyödyntäen onnistui helposti ja nopeasti keskitehoisella pöytätietokoneella. Järjestelmän kehittäminen ja ajaminen vaatii perusosaamista ohjelmoinnista, datan käsittelystä sekä ymmärrystä algoritmien toimintaperiaatteista. Tiedonlouhinta ei vaadi organisaatiolta merkittäviä teknisiä resursseja, mutta osaamisvaatimuksia henkilöstölle se asettaa. Toisaalta tiedonlouhinnan avulla voidaan vapauttaa henkilöstöresursseja yksinkertaisista

tehtävistä, jotka liittyvät datan tulkintaa, informaation lajitteluun ja luokitteluun tai tietovirtojen seurantaan.

Mitä tiedustelun toimintoja tiedonlouhinnan avulla voidaan tukea?

Presidential daily brief on nopeasti laadittava tiedustelutuote edellisen 24 tunnin aikana tapahtuneista asioista ja sen aiheet valitaan päivittäin tiedusteluyhteisön sisällä (Johnson, 2008). Tiedonlouhintaprosessin avulla voidaan tukea tämänkaltaisten tuotteiden suunnittelua ja ohjausta. Prosessin avulla voidaan tunnistaa automaattisesti poikkeamia edellisen vuorokauden aikana kerätystä datasta. Poikkeamat ovat tyypillisesti sellaisia ilmiöitä, jotka herättävät kiinnostusta. Poikkeamat voidaan valita raportin aiheiksi ja ohjata niiden avulla sekä keräystä että analyysia hankkimaan ja tuottamaan aiheesta lisää tietoa.

Tiedonlouhintaa voidaan hyödyntää myös datan keräyksessä ja prosessoinnissa. Tässä työssä hyödynnettiin valmiiksi lajiteltua ja luokiteltua uutisdataa. Tiedonlouhintaprosessin avulla voidaan myös kerätä tekstiä ja prosessoida siitä luonnollisen kielen käsittelymenetelmien avulla prosessoitua informaatiota, jota voidaan hyödyntää tiedustelussa. Tiedonlouhinnan avulla voidaan automatisoida sekä keräystä että prosessointia.

Tiedusteluanalyysissä hyödynnettävillä työvälineillä tarkoitetaan teknologiaa, tuotteita ja prosesseja, jotka tukevat analyytikon työtä kolmella tapaa: Ne helpottavat informaation ja tiedon löytämistä datasta, mahdollistavat hypoteesien kehittelyn ja testaamisen sekä helpottavat kommunikaatiota keräyksen ja asiakkaiden kanssa (Treverton & Gabbard, 2008). Tiedonlouhintaprosessin avulla voidaan tukea analyytikon työtä jokaisella kolmella tavalla. Ensiksi, tiedon löytämiseksi tiedusteluanalyytikon tulee käydä läpi suuria aineistoja etsiessään mielenkiintoisia ilmiöitä. Sen sijaan, että tiedusteluanalyytikko lukisi uutisia asekaupoista ja pyrkisi niiden avulla selvittämään sotilaallisen yhteistyön suhdeverkostoja, voidaan etsintä automatisoida tiedonlouhinnan avulla, kuten esimerkiksi skenaariossa kaksi on esitelty. Vastaavasti, myös terroristisen toiminnan keskeisimpiä alueita voidaan tunnistaa tiedonlouhinnan avulla, kuten skenaariossa yksi on esitelty. Toiseksi, tiedonlouhinnan avulla voidaan testata hypoteeseja. Esimerkiksi, jos hypoteesina on: että Venäjän ja Yhdysvaltojen suhde on heikentynyt, voidaan datan avulla laskea hypoteesia tukevaksi informaatioksi suhdekuvaaja, joka kertoo määrällisesti tarkasteltuna valtioiden välisten suhteiden kehittymisen, kuten skenaariossa viisi on esitelty. Kolmanneksi, tiedonlouhinnan avulla voidaan helpottaa kommunikaatiota, koska prosessin tuotteena syntyy visualisoitua aineistoa, joka helpottaa monimutkaisten ilmiöiden kuvailua.

6.1 Tulosten käytännöllinen ja tieteellinen merkitys

Tutkijat ovat tunnistaneet, että tekoälyn, data-analyysin ja kehittyvän informaatioteknologian avulla voidaan tukea tiedustelua (Lim, 2016; Jani & Soni, 2018; Van Puyvelde ym., 2017; Eldridge ym., 2017), mutta julkisesti esiteltyjä käytän-

nön ratkaisuja on tunnistettavissa hyvin vähän. Tässä tutkielmassa esitellään käytännöllinen ratkaisu, miten tiedusteluorganisaatiot voivat hyödyntää uutta teknologiaa ja älykkäitä menetelmiä toimintojensa kehittämisessä. Tutkielman tuloksena esiteltiin prosessimalli, jonka pohjalta voidaan suunnitella ja kehittää menetelmiä ja järjestelmiä tiedusteluorganisaatioiden työn tueksi. Tulokset osoittavat, miten tiedonlouhinnan avulla voidaan automatisoida tiedusteluprosessin toimintoja ja saavuttaa ajallisia säästöjä.

Tämä tutkielma täydentää tietojärjestelmätieteen tutkimuskenttää esittelemällä uuden kontekstin, johon tiedonlouhintaa voidaan soveltaa. Tiedonlouhintaa on sovellettu useilla eri tutkimusalueilla, kuten terveydenhuollossa (Jothi & Husain, 2015) ja koulutuksessa (Romero & Ventura, 2013). Opetus- ja koulutusdatan louhinta (educational data mining, EDM) on oma tutkimusalueensa, jossa kehitetään menetelmiä koulutukseen ja opetukseen liittyvän datan tutkimiseen. Tiedusteluun liittyvällä datalla on omat erityispiirteensä, jotka määrittyvät keräysmenetelmien (Luvut 2.3 & 2.4) mukaisesti. Tämän tutkielman perusteella voidaan esittää, että tiedustelutiedonlouhinta on oma tutkimusalueensa, jossa kehitetään menetelmiä eri tiedustelun keräysmenetelmillä hankitun datan tutkimiseen. Tutkielma luo perustaa tiedustelun ja tietojärjestelmätieteen tutkimuksen yhdistämiselle.

Tiedustelun tutkimuksen näkökulmasta tutkielma osoittaa, että tiedustelussa ja tietojärjestelmätieteissä on vastaavuuksia ja samankaltaisuuksia. Tieto on keskeinen yhdistävä käsite, joka yhdistää tutkimusalueiden konstruktioita ja konsepteja. Tiedustelu on varsin nuori tutkimusalue, jolla ei ole omia tutkimusmenetelmiä tai tieteellistä perinnettä. Tutkielma osoittaa, että tietojärjestelmätiede soveltuu hyvin tiedustelun tutkimukseen. Tutkielma avaa mahdollisuuksia jatkaa tiedustelun tutkimista tietojärjestelmätieteen avulla ja suunnitella ja kehittää uusia artefakteja, joiden avulla voidaan tehostaa ja kehittää tiedusteluorganisaatioiden toimintaa.

6.2 Luotettavuuden arviointi

Larsen ym. (2020) ovat kehittäneet viitekehyksen suunnittelututkimuksen luotettavuuden arviointiin. Se koostuu kolmesta osiosta:

- Suunnittelua ohjaavan teorian ja vaatimusten luotettavuus
- Kehittämisen ja kontekstiin soveltamisen luotettavuus
- Suunnittelun lopputuloksen luotettavuus

Suunnittelua ohjaavan teorian ja vaatimusten luotettavuus

Warner (2007) esittää, että tiedustelun tutkimus voidaan jakaa kahteen osaan tutkijan position mukaan: Tutkija on joko tiedusteluorganisaation sisällä tai ulkopuolella. Mikäli tutkija on tiedusteluorganisaation sisällä ja hänellä on pääsy salaiseen tietoon, saa hän kerättyä paremmin totuutta kuvaavan tutkimusaineiston, mutta tutkitun tiedon julkaiseminen on haastavaa. Mikäli tutkija on

organisaation ulkopuolella, tarkastelee hän vain tiedustelun julkista osaa, joka ei kuvaa koko todellisuutta. Tässä tutkielmassa strategisen tiedustelun toimintaa ja kohteita selvitettiin kirjallisuuskatsauksen avulla, joka ohjasi artefaktin suunnittelua ja kehittämistä. Vaatimusten luotettavuutta olisi voitu parantaa, jos tiedustelutehtävät olisi asetettu tiedusteluorganisaation toimesta, jolloin vaatimukset olisivat vastanneet paremmin todellista kontekstia.

Kehittämisen ja kontekstiin soveltamisen luotettavuus

Tutkielmassa kehitetyn tiedonlouhintaprosessin toimivuutta esiteltiin prototyyppijärjestelmän avulla, joka hyödynsi GDELT-aineistoa. Wang, Kennedy, Lazer ja Ramakrishan (2016) ovat esittäneet kritiikkiä GDELT-uutisdatan validiteetista ja reliabiliteetista. Heidän mukaansa aineistossa on useita duplikaatteja samoista tapahtumista ja uutisteksteistä on virheellisesti koodattuja, jonka seurauksena noin 20 % aineiston tapahtumista vastaa todellisuutta 80 % ovat virheellisiä. He vertailivat myös eri aineistojen tarkkuuksia keskenään. Aineistojen data korreloi heikosti keskenään, etenkin päivätasolla ja sellaisten tapahtumien suhteen, joita oli määrällisesti vähän. Reliabiliteetti kuitenkin hyvälle tasolle tapauksessa, jolloin tapahtumasta oli kerätty aineistoon suuri määrä osumia. Tällaisia tapahtumia ovat merkittävät ja suurta huomiota mediassa saavat tapahtumat. (Wang, Kennedy, Lazer & Ramakrishan, 2016.) Luotettavuutta voidaan parantaa hyödyntämällä useita aineistoja. Tässä tapauksessa luotattavuutta olisi voitu parantaa, jos GDELT-uutisdatan rinnalla oli hyödynnetty esimerkiksi Twitteristä kerättyä dataa. Luotettavuuden kyseenalaistaminen kohdistuu tässä suhteessa tutkielman alakysymykseen: Millaista tiedustelutietoa uutisdatasta voidaan tuottaa? Mikäli data on virheellistä, on siitä tuotettu tietokin virheellistä. Tiedonlouhinnan avulla tuotetun tiedon oikeellisuus ei ole tämän tutkielman tavoitteiden kannalta merkityksellistä, koska keskiössä oli prosessin toimivuuden ja käyttökelpoisuuden osoittaminen, varsinaisen tiedustelutiedon tuottamisen sijaan.

Suunnittelun lopputuloksen luotettavuus

Venablen ym. (2016) mukaan artefaktia voidaan arvioida joko luonnollisessa tai keinotekoisessa ympäristössä. Tässä tutkielmassa artefaktin arviointi toteutettiin täysin keinotekoisessa ympäristössä. Skenaarioissa hyödynnettiin todellista uutisaineistoa tiedustelutiedon tuotannossa, mutta tiedustelukysymykset olivat johdettu kirjallisuuskatsauksen perusteella. Luotettavuutta olisi voitu parantaa, mikäli artefaktia olisi käytetty todellisessa ympäristössä, kuten tiedusteluorganisaatiossa tai tiedustelutehtävät olisivat tulleet todelliselta tiedusteluorganisaatiolta. Evaluointi luonnollisessa ympäristössä olisi vaatinut resursseja tiedusteluorganisaatiolta. Tutkielman tavoitteiden kannalta riittävä luotettavuus saavutettiin evaluoimalla artefaktia keinotekoisessa ympäristössä.

Cleven, Gubler ja Hüner (2009) esittävät, että suunnitteletutkimuksen lopputuloksena tuotettua artefaktia voidaan arvioida joko sisäisesti tai ulkoisesti. Sisäisessä arvioinnissa tutkija tai tutkimusryhmä itse evaluoi artefaktin toiminnallisuuksia, käyttökelpoisuutta tai suorituskykyä. Ulkoisessa arvioinnissa artefakti evaluoidaan hyödyntäen todellisia käyttäjiä esimerkiksi kyselytutkimuksen avulla. Tämän tutkielman lopputuloksen arvioinnin luotettavuutta olisi voitu parantaa ulkoisen arvioinnin avulla. Nyt arviointi toteutettiin ajan säästämiseksi tutkijan toimesta. Käytännössä tutkija arvioi itse kehittämäänsä artefaktia, joten arviointi on altis subjektiivisille vinoumille. Arviointiprosessi on kuitenkin dokumentoitu avoimesti ja se perustuu loogisiin argumentteihin, joten sen uskottavuus on lukijan arvioitavissa.

6.3 Jatkotutkimus

Tutkielman kautta muodostui useita ajatuksia jatkotutkimukselle. Tiedonlouhintaa on sovellettu eri tutkimusalueilla varsin kattavasti, mutta hyödyntämistä tiedustelussa on tutkittu varsin vähän. Tämän tutkielman tulokset osoittavat, että soveltaminen tarjoaa mahdollisuuksia tiedustelun kehittämiselle ja tutkimiselle.

Tässä tutkielmassa tiedustelun tarpeet ja konteksti muodostettiin kirjallisuuskatsauksen avulla. Kiinnostavaa olisi tutkia tiedonlouhinnan hyödyntämistä tapaustutkimuksena tiedusteluorganisaatiossa todellisiin tarpeisiin vastaten. Tämä vaatisi neuvottelua tutkimuksen julkisuuden suhteen. Akateeminen tutkimus kun on lähtökohtaisesti julkista ja tiedusteluorganisaatioiden keräämä data ja tarkat tiedustelun kohteet ovat pääsääntöisesti turvaluokiteltua tietoa.

Tämän tutkielman kehyksessä käsiteltiin ainoastaan avoimista lähteistä kerättyä uutisdataa. Mielenkiintoista olisi laajentaa tutkimusta tarkastelemalla tarkemmin tiedonlouhinnan soveltamista muihin datalähteisiin. Kehitetty prosessimalli on joustava ja sen avulla voidaan tuottaa tietoa erilaisesta datasta. Avointen lähteiden osalta seuraava laajennus voisi olla tiedustelutiedon tuottaminen sosiaalisen median teksteistä. Tämä vaatisi erilaisten luonnollisen kielen käsittelymenetelmien hyödyntämistä. Muiden lähteiden osalta kiinnostavaa olisi tutkia tiedonlouhinnan hyödyntämistä esimerkiksi geo- ja signaalitiedustelussa. Geotiedustelun suhteen hedelmällistä olisi tutkia, miten tiedonlouhinnan avulla voidaan automatisoida satelliittitiedustelun kuvantulkintaa ja kohteiden seurantaa. Signaalitiedustelun osalta mielenkiintoista olisi tutkia esimerkiksi, miten tiedonlouhinnan avulla voidaan tunnistaa verkostoja tai tunnistettuja entiteettejä viestiliikenteestä.

Tiedonlouhintaa sovellettiin tässä tutkielmassa vain yhteen datalähteeseen. Tiedustelussa kerätään tietoa useista lähteistä ja informaation yhdistäminen on tärkeää. Informaatio voidaan yhdistää standardoitujen objektien avulla. Jatkotutkimuksena voidaan tutkia, miten tiedonlouhinnan avulla voidaan tunnistaa edellä mainittuja objekteja erilaisesta datasta, ja yhdistää eri keräysmenetelmillä hankittu data yhteiseen tietomalliin.

Tässä tutkielmassa eri menetelmiä hyödynnettiin vain pintapuolisesti, koska tarkoitus oli todistaa prosessin toimivuutta esimerkkien avulla. Poikkeaman tunnistaminen todettiin keskeiseksi tehtäväksi tiedustelun näkökulmasta. Jatkotutkimuksessa olisi hyödyllistä keskittyä poikkeaman tunnistamiseen ja selvittää, miten erilaisesta datasta voidaan tunnistaa poikkeamia, ja mitkä ovat tehokkaimmat menetelmät poikkeaman tunnistamiseen eri tapauksissa.

Lopuksi voidaan todeta, että tiedustelun tutkimuksen ja tietojärjestelmätieteiden tutkimuksen yhdistäminen on loogista, koska tieto on molemmilla tutkimusalueilla keskeisessä roolissa. Tietojärjestelmätiede on laaja ja kehittyvä tutkimusalue, joka tarjoaa soveltuvia konstruktioita ja menetelmiä tiedustelun tutkimiseen ja kehittämiseen.

7 YHTEENVETO

Tämän tutkielman tavoitteena oli tutkia tiedustelua tietojärjestelmätieteen avulla sekä esitellä, miten tiedonlouhinnan avulla voidaan tukea strategista tiedustelua. Tiedonlouhinnan soveltamista uutisdata-analyysiin on tutkittu aiemmin ja informaatioteknologian tarjoamat hyödyt tiedustelun tarpeisiin on tunnistettu, mutta konkreettista ja käytäntöön sitovaa tutkimusta tiedonlouhinnan soveltamisesta tiedustelun kontekstiin ei ole julkisesti esitelty. Tutkimuskysymyksenä oli: Miten tiedonlouhinnan avulla voidaan tuottaa uutisdatasta tietoa strategisen tiedustelun tarpeisiin?

Tutkimusmenetelmänä hyödynnettiin Peffersin ym. (2007) kehittämää suunnittelututkimuksen metodologiaa (DSRM). Tutkimusongelmaan selvitettiin vastaus suunnittelemalla ja kehittämällä erityisesti strategisen tiedustelun kontekstiin sovellettu tiedonlouhinnan prosessimalli. Suunnittelu ja kehittäminen toteutettiin strategisen tiedustelun ja tiedonlouhinnan kirjallisuuteen pohjautuen. Prosessimallin toimivuutta demonstroitiin ja evaluoitiin prototyyppijärjestelmän avulla, jota käytettiin viidessä strategista tiedustelua kuvailevassa skenaariossa. Tutkimusaineistona käytettiin GDELT-tietokantaa, joka sisältää avoimista lähteistä kerättyä uutisdataa. Skenaariot johdettiin kirjallisuuskatsauksen avulla tunnistetuista keskeisimmistä strategisen tiedustelun kohteista. Prototyyppisovelluksen avulla tuotettiin eri datanlouhintamenetelmiä hyödyntäen tiedustelutietoa terrorismista, valtioiden välisistä konflikteista, kansainvälisestä sotilaallisesta yhteistyöstä sekä valtioiden sisäisistä konflikteista. Prosessimallin hyödyllisyyttä evaluoitiin sen avulla tuotetun tiedon laadun, vaatimien resurssien ja yleistettävyyden näkökulmasta. Evaluoinnin tarkoitus oli selvittää prosessin hyödyllisyyttä.

Tutkimuksen tuloksena muodostettiin kuusivaiheinen prosessimalli, jonka avulla voidaan tuottaa uutisdatasta tietoa strategisen tiedustelun tarpeisiin. Skenaarioiden avulla osoitettiin, että prosessin avulla voidaan tuottaa informaatiota erilaisista strategisen tiedustelun kohteista. Informaatio on laadultaan visuaalista, joten sitä on ihmisen helppo tulkita. Tiedonlouhinnan avulla voidaan säästää tiedusteluorganisaation resursseja automatisoimalla toimintoja, jotka vievät ihmisiltä paljon aikaa, mutta tietokoneilta vain vähän. Esimerkiksi tiedonlouhinnan avulla voidaan automatisoida kansainvälisten suhteiden seuran-

taa tunnistamalla poikkeamia niistä. Tiedonlouhinnan avulla voidaan automatisoida keräystä ja prosessointia, mutta sen avulla ei kuitenkaan voida korvata tiedusteluanalyytikoiden tekemää tiedusteluraporttien tai esitysten tuotantoa. Tiedusteluanalyytikot voivat hyödyntää tiedonlouhintaa analyysinsä tukena ja liittää prosessin kautta tuotettuja visualisointeja osaksi tuotteita. Suurin hyöty syntyy ajan säästämisessä. Tiedonlouhinnan avulla on tehokasta käsitellä suuria aineistoja, joiden läpikäymisessä ihmisellä kestäisi useita tunteja tai päiviä.

Tiedusteluanalyysin suorituskyky muodostuu ihmisen ajattelusta ja työvälineistä, joita ovat erilaiset rakenteelliset menetelmät ja tietojärjestelmät. Tämän tutkielman tulokset osoittavat, että informaatioteknologian avulla voidaan tukea tiedusteluanalyysia, mutta kehittyneimmätkään algoritmit eivät korvaa ihmisen ajattelua tiedustelutiedon tuotannossa. Informaatioteknologian avulla voidaan kuitenkin automatisoida sellaisia prosessin osia, jotka ovat ihmiselle työläitä, mutta tietokoneelle helppoja.

Aikaisemmassa tutkimuksessa on tunnistettu, että informaatioteknologian avulla voidaan tukea tiedustelua ja tietojärjestelmät muodostavat merkittävän osan tiedustelun suorituskyvystä. Tiedonlouhinnan sekä vastaavien menetelmien avulla on tuotettu uutisteksteistä tietoa erilaisiin tarpeisiin. Tämän tutkielman tulokset tukevat aikaisemmassa tutkimuksessa esiteltyjä ratkaisuja. Aikaisempaan tutkimukseen verrattuna, tämä tutkielma esittelee konkreettisen ratkaisun tiedonlouhinnan soveltamisesta strategisen tiedustelun tarpeisiin.

Useissa tutkimuksissa on pyritty kehittämään menetelmiä, joiden tarkoitus on ennustaa uutisdatasta tulevaisuuden kehitystä. Tämän tutkielman tulosten perusteella voidaan kuitenkin esittää, että strategisen tiedustelun näkökulmasta tutkittava ilmiöt, kuten kansainväliset suhteet ovat lähtökohtaisesti kaoottisia, ja niiden ennustaminen uutisten perusteella on hyvin epäluotettavaa. Uutisteksteistä tai -datasta voidaan kuitenkin tunnistaa tiedonlouhinnan avulla kansainvälisissä suhteissa ilmeneviä poikkeamia, kuten kriisejä ja konflikteja, ja kohdentaa niihin tarkempaa tiedonhankintaa ja analyysiä.

Tässä tutkielmassa tutkitaan strategista tiedustelua ainoastaan avoimista lähteistä kerätyn uutisdatan avulla. Vaikka uutisten avulla voidaankin tuottaa merkittävä osa strategisen tason tiedustelutiedosta, on ainoastaan yhden keräysmenetelmän ja lähteen kautta toteutettu tutkimus varsin pintapuolinen. Tämä asettaa rajoituksia tulosten yleistettävyydelle. Lisäksi tutkielman asetelmassa prosessin suunnittelu, kehittäminen ja arviointi toteutettiin keinotekoisessa ympäristössä. Tämä heikentää tutkimuksen luotettavuutta. Luotettavuutta olisi voitu parantaa, jos suunnittelua ohjaavat tarpeet olisi hankittu todelliselta tiedusteluorganisaatiolta oikeaan tiedustelutehtävään liittyen, tai prototyyppisovellusta olisi käytetty todellisessa tiedusteluorganisaatiossa oikealla datalla tiedustelutyötä tekevien analyytikkojen arvioimana.

LÄHTEET

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis, 16(1),* 3–9.

Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., . . . Schlesinger, M. (2012). Mapping the landscape of human-level artificial general intelligence. *AI Magazine, 33(1)*, 25–42. https://doi.org/10.1609/aimag.v33i1.2322

Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. Teoksessa *Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215*, 487–499).

Althoff, M. (2015). Human intelligence. Teoksessa M. Lowenthal Mark, & R. M. Clark (toim.), *The five disciplines of intelligence collection (45–80),* Thousand oaks, CA: Sage

Atwood, C. P. (2015). Activity-based intelligence: Revolutionizing military intelligence analysis. *Joint Force Quarterly, 77 (2nd Quarter).* Haettu osoitteesta https://ndupress.ndu.edu/Portals/68/Documents/jfq/jfq-77/jfq-77_24-33_Atwood.pdf

Biltgen, P., & Ryan, S. (2016). *Activity-based intelligence: Principles and applications*. Norwood: Artech House.

Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology, 92(5)*, 1170–1182. https://doi.org/10.1086/228631

CIA. (2021). The world factbook. Haettu osoitteesta https://www.cia.gov/the-world-factbook/

Chin Jr, G., Kuchar, O. A., & Wolf, K. E. (2009). Exploring the analytical processes of intelligence analysts. Teoksessa *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 11–20. https://doi.org/10.1145/1518701.1518704

Clapper, J. (2016). DNI Clapper's as delivered remarks at the 2016 geoint symposium. Haettu osoitteesta https://www.dni.gov/index.php/newsroom/speeches-interviews/speeches-interviews-2016/item/1594-dni-clapper-s-as-delivered-remarks-at-the-2016-geoint-symposium

Clark, R. M. (2020). *Intelligence Analysis: A Target-Centric Approach.* (Sixth edition). Thousand Oaks: CQ Press.

Cleven, A., Gubler, P., & Hüner, K. M. (2009, Toukokuu). Design alternatives for the evaluation of design science research artifacts. Teoksessa *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology,* (1–8). https://doi.org/10.1145/1555619.1555645

CSIS. (2021). Center for strategic and international studies. Haettu osoitteesta https://www.csis.org/

Davies, P. (2004). *MI6 and the machinery of spying: Structure and process in britain's secret intelligence.* New York: Routledge.

Davies, P., Gustafson, K., & Rigden, I. (2013). The intelligence cycle is dead, long live the intelligence cycle: Rethinking intelligence fundamentals for a new intelligence doctrine. Teoksessa M. Phythian (toim.), *Understanding the intelligence cycle* (56–76). New York: Routledge.

De Mauro, A., Greco, M. & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features, *Library Review, Vol. 65 No. 3,* 122–135. https://doi.org/10.1108/LR-06-2015-0061

Dey, L., Haque, S. M., Khurdiya, A., & Shroff, G. (2011, September). Acquiring competitive intelligence from social media. Teoksessa *Proceedings of the 2011 joint workshop on multilingual OCR and analytics for noisy unstructured text data* (1–9). https://doi.org/10.1145/2034617.2034621

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science, 10*(6), 753–757. https://doi.org/10.1177/1745691615598511

Dupont, A. (2003). Intelligence for the twenty-first century. *Intelligence and National Security, 18*(4), 15–39. https://doi.org/10.1080/02684520310001688862

Eldridge, C., Hobbs, C., & Moran, M. (2018). Fusing algorithms and analysts: Open-source intelligence in the age of 'Big data'. *Intelligence and National Security, 33*(3), 391–406. https://doi.org/10.1080/02684527.2017.1406677

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. Teoksessa *Kdd, Vol. 96, No. 34,* (226–231).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17(3),* 37. https://doi.org/10.1609/aimag.v17i3.1230

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* Springer. https://doi.org/10.1007/978-3-319-10247-4

Gartner. (2021). Definition of big data - gartner information technology glossary. Haettu osoitteesta https://www.gartner.com/en/information-technology/glossary/big-data

GDELT Project. (2021). Haettu osoitteesta https://www.gdeltproject.org/

Gentry, J. A., & Gordon, J. S. (2019). *Strategic warning intelligence: History, challenges, and prospects.* Washington: Georgetown University Press.

George, R. Z., & Bruce, J. B. (2014). *Analyzing intelligence: National security practitioners' perspectives.* Washington: Georgetown University Press.

Gill, P., & Phythian, M. (2013). From intelligence cycle to web of intelligence: Complexity and the conceptualisation of intelligence. Teoksessa M. Phythian (toim.), *Understanding the intelligence cycle* (21–42). New York: Routledge.

Gill, P., & Phythian, M. (2016). What is intelligence studies? *The International Journal of Intelligence, Security, and Public Affairs, 18(1),* 5–19. https://doi.org/10.1080/23800992.2016.1150679

Glassman, M., & Kang, M. J. (2012). Intelligence in the internet age: The emergence and evolution of open source intelligence (OSINT). *Computers in Human Behavior, 28*(2), 673–682. https://doi.org/10.1016/j.chb.2011.11.014

Goldman, J. (2011). *Words of intelligence: An intelligence professional's lexicon for domestic and foreign threats.* Lanham: Scarecrow Press.

Grabo, C. (2010). *Handbook of warning intelligence: Assessing the threat to national security* Lanham: Scarecrow Press.

Heiskanen, H. (2021, Toukokuun 28). Mielenosoitukset mustan miehen kuolemasta poliisin käsissä jatkuvat Minneapolisissa – Pormestari vaatii syytteitä poliisia vastaan. *Yle uutiset.* https://yle.fi/uutiset/3-11372204

Heuer Jr, R. J. (2009). The evolution of structured analytic techniques. *Presentation to the National Academy of Science, National Research Council Committee on Behavioral and Social Science Research to Improve Intelligence Analysis for National Security,* , 529–545.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly,* 75–105. https://doi.org/10.2307/25148625

Hulnick, A. S. (2006). What's wrong with the intelligence cycle. *Intelligence and National Security, 21*(6), 959–979. https://doi.org/10.1080/02684520601046291

Hulnick, A. S. (2010). The dilemma of open sources intelligence: Is OSINT really intelligence? Teoksessa Johnson, L. K. (toim.), *The oxford handbook of national security intelligence* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195375886.003.0014

Hulnick A. S. (2013). Intelligence theory: Seeking better models. Teoksessa M. Phythian (toim.), *Understanding the intelligence cycle* (163–174). New York: Routledge.

Irwin, D., & Mandel, D. R. (2019). Improving information evaluation for intelligence production. *Intelligence and National Security, 34(4),* 503–525. https://doi.org/10.1080/02684527.2019.1569343

Jardines, E. A. (2015). Open source intelligence. Teoksessa M. Lowenthal Mark, & R. M. Clark (toim.), *The five disciplines of intelligence collection (5–43),* Thousand oaks, CA: Sage

Jensen III, C. J., McElreath, D. H., & Graves, M. (2017). *Introduction to intelligence studies.* Boca Raton: Routledge.

Johnson, L. K. (1986). Making the intelligence "Cycle" work. *International Journal of Intelligence and Counter Intelligence, 1*(4), 1–23. https://doi.org/10.1080/08850608608435033

Johnson, L. K. (2008). Glimpses into the gems of american intelligence: The president's daily brief and the national intelligence estimate. *Intelligence and National Security, 23*(3), 333–370. https://doi.org/10.1080/02684520802121257

Johnson, L. K. (Ed.). (2010). *The oxford handbook of national security intelligence*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195375886.003.0001

Johnston, C., Wright Jr, E. C., Bice, J., Almendarez, J., & Creekmore, L. (2015). Transforming defense analysis. *Joint Forces Quarterly, 79*(4), 12–18. Haettu osoitteesta https://ndupress.ndu.edu/Portals/68/Documents/jfq/jfq-79/jfq-79_12-18_Johnston-et-al.pdf

Jones, M. (2019). What we talk about when we talk about (big) data. *The Journal of Strategic Information Systems*, *28*(1), 3–16. https://doi.org/10.1016/j.jsis.2018.10.005

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Jothi, N., & Husain, W. (2015). Data mining in healthcare–a review. *Procedia computer science, 72,* 306–313. https://doi.org/10.1016/j.procs.2015.12.145

Kang, Y., & Stasko, J. (2011). Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. Teoksessa *2011 IEEE Conference on Visual Analytics Science and Technology (VAST),* 21–30. https://doi.org/10.1109/VAST.2011.6102438

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms.* New Jersey: John Wiley & Sons.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics, 8*(1), 1–8. https://doi.org/10.1109/2945.981847

Kent, S. (1949). *Strategic intelligence for american world policy* Princeton University Press. https://doi.org/10.1515/9781400879151

Kent, S. (1993). The need for an intelligence literature. *Studies in Intelligence*, *1*(1), 1–11. Haettu osoitteesta https://www.cia.gov/static/539a695e2365ed5a17422139fa14e3cf/Need-for-Intelligence-Literature.pdf

Khan, R. A. (2012). KDD for business intelligence. *Journal of Knowledge Management Practice, 13(2),* 134. Haettu osoitteesta http://www.tlainc.com/articl304.htm

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM), 46(5),* 604–632. https://doi.org/10.1145/324133.324140

Laki sotilastiedustelusta 26.4.2019/590.

Larsen, K. R., Lukyanenko, R., Mueller, R. M., Storey, V. C., VanderMeer, D., Parsons, J., & Hovorka, D. S. (2020, December). Validity in Design Science Research. Teoksessa *International Conference on Design Science Research in Information Systems and Technology,* (272–282). Springer, Cham. https://doi.org/10.1007/978-3-030-64823-7_25

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Leetaru, K., & Schrodt, P. A. (2013). Gdelt: Global data on events, location, and tone, 1979–2012. Teoksessa *ISA Annual Convention, , 2(4)* 1–49.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). *Mining of massive data sets* Cambridge: Cambridge university press.

Liebowitz, J. (2006). *Strategic intelligence: Business intelligence, competitive intelligence, and knowledge management* Boca Raton: Auerbach publications.

Lowenthal, M. M. (2019). *Intelligence: From secrets to policy*. Thousand Oaks: CQ press.

Lowenthal, M. M., & Clark, R. M. (2015). *The five disciplines of intelligence collection.* Thousand oaks, CA: Sage.

Mainwaring, S., & Aldrich, R. J. (2021). The secret empire of signals intelligence: GCHQ and the persistence of the colonial presence. *The International History Review, 43*(1), 54-71. https://doi.org/10.1080/07075332.2019.1675082

Major, J. S. (2014). *Communicating with intelligence: Writing and briefing for national security* Lanham: Rowman & Littlefield.

Mandel, R. (2019). Global data shock: strategic ambiguity, deception, and surprise in an age of information overload. CA: Stanford University Press.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision support systems, 15(4),* 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review, 25*(2), 137. https://doi.org/10.1017/S0269888910000032

Marrin, S. (2011). *Improving intelligence analysis*. London: Routledge. https://doi.org/10.4324/9780203810200

Marrin, S. (2016). Improving intelligence studies as an academic discipline. *Intelligence and National Security, 31(2),* 266-279. https://doi.org/10.1080/02684527.2014.952932

Marrin, S. (2018). Evaluating intelligence theories: Current state of play. *Intelligence and National Security, 33*(4), 479-490. https://doi.org/10.1080/02684527.2018.1452567

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine, 27*(4), 12. https://doi.org/10.1609/aimag.v27i4.1904

Mercado, S. C. (2009). Sailing the sea of OSINT in the information age. Teoksessa Andrew, C., Aldrich, R. J., & Wark, W. K. (toim.), *Secret intelligence: A reader (78–90),* New York: Routledge.

Miller, B. H. (2014). US strategic intelligence forecasting and the perils of prediction. *International Journal of Intelligence and CounterIntelligence, 27*(4), 687–701. https://doi.org/10.1080/08850607.2014.924810

Morris, J. L. & Clark, R. M. (2015). Measurement and Signature intelligence. Teoksessa M. Lowenthal Mark, & R. M. Clark (toim.), *The five disciplines of intelligence collection (159–208),* Thousand oaks, CA: Sage

Murdock, D & Clark, R. M. (2015). Geospatial intelligence. Teoksessa M. Lowenthal Mark, & R. M. Clark (toim.), *The five disciplines of intelligence collection (111–158),* Thousand oaks, CA: Sage

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective* Cambridge: MIT press.

Nolte, W. M. (2015). Signals intelligence. Teoksessa M. Lowenthal Mark, & R. M. Clark (toim.), *The five disciplines of intelligence collection (81–110),* Thousand oaks, CA: Sage

Office of the Director of National Intelligence. (2007). National intelligence estimate, Iran: Nuclear intentions and capabilities. Haettu osoitteesta https://www.dni.gov/files/documents/Newsroom/Reports%20and%20Pubs/20071203_release.pdf

Office of the Director of National Intelligence. (2019). The National Intelligence Strategy of the United States of America. Haettu osoitteesta https://www.dni.gov/files/ODNI/documents/National_Intelligence_Strategy_2019.pdf

Odom, W. E. (2008). Intelligence analysis. *Intelligence and National Security, 23*(3), 316-332. https://doi.org/10.1080/02684520802121216

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435(7043),* 814–818. https://doi.org/10.1038/nature03607

Peffers, K., Rothenberger, M., Tuunanen, T., & Vaezi, R. (2012, May). Design science research evaluation. Teoksessa *International Conference on Design Science Research in Information Systems (398–410).* Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-29863-9_29

Peffers, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research, *European Journal of Information Systems, 27:2,* 129–139. https://doi.org/10.1080/0960085X.2018.1458066

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24(3),* 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pherson, R. H., & Heuer Jr, R. P. (2014). Structured analysis techniques: A new approach to analysis. Teoksessa George, R. Z., & Bruce, J. B. (2014). *Analyzing intelligence: National security practitioners' perspectives. (444–477).* Washington: Georgetown University Press.

Pherson, R. H., & Heuer Jr, R. J. (2020). *Structured analytic techniques for intelligence analysis* Thousand Oaks: CQ Press.

Phythian, M (2013). *Understanding the intelligence cycle*. New York: Routledge.

Pirolli, P., & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. Teoksessa *Proceedings of International Conference on Intelligence Analysis, 5* 2–4.

Poliisilaki 26.4.2019/581

Puolustusvoimat. (2021). Puolustusvoimien tiedustelulaitos. Haettu osoitteesta https://puolustusvoimat.fi/tietoa-meista/tiedustelulaitos

Richards Julian. (2013). Pedalling hard: Further questions about the intelligence cycle in the contemporary era. Teoksessa M. Phythian (toim.), *Understanding the intelligence cycle* (57–69). New York: Routledge.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.* https://doi.org/10.1002/widm.1075

Madden, S. (2012). From databases to big data. *IEEE Internet Computing, 16(3),* 4–6. https://doi.org/10.1109/MIC.2012.50

Sienkiewicz, M. (2015). Open BUK: Digital labor, media investigation and the downing of MH17. *Critical Studies in Media Communication, 32(3),* 208–223. https://doi.org/10.1080/15295036.2015.1050427

Schrodt, P. A. (2012). *Cameo: Conflict and mediation event observations event and actor codebook.* Pennsylvania State University. Haettu osoitteesta http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf

Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, *12*(1), 217–222.

Sharfman, P. (1995). Intelligence analysis in an age of electronic dissemination. *Intelligence and National Security, 10*(4), 201–211. https://doi.org/10.1080/02684529508432333

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., . . . Bolton, A. (2017). Mastering the game of go without human knowledge. *Nature, 550*(7676), 354-359. https://doi.org/10.1038/nature24270

Steele, R. D. (2007). Open source intelligence. Teoksessa M. Phythian (toim.), *Handbook of Intelligence Studies (129-147)*. New York: Routledge.

Stratfor. (2021). Stratfor. Haettu osoitteesta https://www.stratfor.com/

Suojelupoliisi. (2021). Suojelupoliisi. Haettu osoitteesta https://supo.fi/

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural computation, 13(7), 1443-1471. https://doi.org/10.1162/089976601750264965

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable.* New York: Random house.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction.* New York: Random House.

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science, 23*(4), 290-295. https://doi.org/10.1177/0963721414534257

Treverton, G. F., & Gabbard, C. B. (2008). *Assessing the tradecraft of intelligence analysis.* Santa Monica: Rand Corporation.

Tropotei, T. O. (2018). Criticism against the intelligence cycle. *Scientific Research & Education in the Air Force-AFASES,* 77–88. https://doi.org/10.19062/2247-3173.2018.20.9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124

United Nations. (2004). *A more secure world: Our shared responsibility - report of the secretary-general's high-level panel on threats, challenges and change* https://www.un.org/peacebuilding/sites/www.un.org.peacebuilding/files/documents/hlp_more_secure_world.pdf

Uosukainen, S. (2021, Maaliskuun 18). Putin vastasi Bidenille sananlaskulla: "Joka toista haukkuu, itse on" – myöhemmin Putin ehdotti Bidenille verkkokeskustelua mahdollisimman pian. *Yle uutiset.* https://yle.fi/uutiset/3-11843939

US Joint Chiefs of Staff. (2013). *Joint publication 2-0 joint intelligence*. Haettu osoitteesta https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp2_0.pdf

Valtioneuvosto. (2020). *Valtioneuvoston ulko- ja turvallisuuspoliittinen selonteko.* Haettu osoitteesta http://urn.fi/URN:ISBN:978-952-287-876-2

Verleysen, M., & François, D. (2005). The curse of dimensionality in data mining and time series prediction. Teoksessa *International Work-Conference on Artificial Neural Networks,* 758–770. https://doi.org/10.1007/11494669_93

Wang, W., Kennedy, R., Lazer, D., & Ramakrishnan, N. (2016). Growing pains for global monitoring of societal events. *Science, 353(6307),* 1502–1503. https://doi.org/10.1126/science.aaf6758

Ward, M. D., Beger, A., Cutler, J., Dickenson, M., Dorff, C., & Radford, B. (2013). Comparing GDELT and ICEWS event data. *Analysis, 21(1),* 267–297. Haettu osoitteesta https://www.researchgate.net/profile/Andreas-Beger/publication/303211430_Comparing_GDELT_and_ICEWS_event_data/links/57f7d9bb08ae886b89836115/Comparing-GDELT-and-ICEWS-event-data.pdf

Warner Michael. (2013). The past and future of the intelligence cycle. Teoksessa M. Phythian (toim.), *Handbook of Intelligence Studies (23-34)*. New York: Routledge.

Wixom, B., & Watson, H. (2010). The BI-based organization. *International Journal of Business Intelligence Research (IJBIR)*, *1*(1), 13–28. https://doi.org/10.4018/jbir.2010071702

Wolfberg, A. (2017). The president's daily brief: Managing the relationship between intelligence and the policymaker. *Political Science Quarterly, 132*(2), 225–258. https://doi.org/10.1002/polq.12616

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., . . . Philip, S. Y. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14*(1), 1–37. https://doi.org/10.1007/s10115-007-0114-2

YLE. (22.3.2021). Ylen aamu tänään: Venäjän ja Yhdysvaltojen välit kiristyvät – mikä maiden välejä hiertää? Haettu osoitteesta https://yle.fi/uutiset/3-11847006

Zaki, M. J., & Meira, W. (2014). *Data mining and analysis: Fundamental concepts and algorithms.* Cambridge: Cambridge University Press.

Zanasi, A. (1998). Competitive intelligence through data mining public sources. *Competitive Intelligence Review: Published in Cooperation with the Society of Competitive Intelligence Professionals*, *9*(1), 44-54. https://doi.org/10.1002/(SICI)1520-6386(199801/03)9:1<44::AID-CIR8>3.0.CO;2-A

Ziegler, C. (2012). *Mining for strategic competitive intelligence : Foundations and applications*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-27714-6