Arttu Mäkelä

MASSADATA JA INFORMAATION YKSITYISYYDEN SÄÄNTELY VAKUUTUSALALLA - HAASTATTELUTUTKIMUS

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2022

TIIVISTELMÄ

Mäkelä, Arttu
Massadata ja informaation yksityisyyden sääntely vakuutusalalla - haastattelututkimus
Jyväskylä: Jyväskylän yliopisto, 2022, 64 s.
Tietojärjestelmätiede, pro gradu -tutkielma
Ohjaaja: Koskelainen, Tiina

Digitaalisessa muodossa olevan datan määrä on kasvanut räjähdysmäisesti viime vuosikymmenten aikana, ja organisaatiot toimialasta riippumatta pyrkivät löytämään suurista datajoukoista päätöksentekoa tukevaa merkityksellistä informaatiota. Vakuutusalalla on erityisen suuri potentiaali massadatan hyödyntämisessä toimialan erityispiirteiden, kuten datapainotteisuuden vuoksi. Suuren potentiaalin rinnalla myös hyödyntämisessä esiintyy myös merkittäviä haasteita. Massadatalle ominaisten piirteiden lisäksi haasteita aiheuttaa informaation yksityisyyden sääntely, kuten Euroopan Unionin yleinen tietosuoja-asetus. Tämän tutkielman tarkoitus on kartoittaa massadatan käyttöä, hyödyntämisen potentiaalia sekä käytössä esiintyviä haasteita vakuutusalalla. Lisäksi tutkielmassa selvitetään informaation yksityisyyttä koskevan sääntelyn vaikutuksia datan käyttöön vakuutusalalla. Kirjallisuuskatsauksen lisäksi kysymyksiin pyritään vastaamaan teemahaastatteluilla, joissa haastateltiin vakuutusalalla toimivia data-analytiikan asiantuntijoita. Haastattelujen pohjalta tunnistettiin massadatan hyödyntämisen potentiaali ja nykytila vakuutusalalla, sekä hyödyntämisen keskeisimmät haasteet. Lisäksi tunnistettiin informaation yksityisyyteen liittyvän sääntelyn keskeisimmät vaikutukset. Massadatan potentiaali vakuutusalalla on suuri, mutta siihen kohdistetut odotukset eivät ole toistaiseksi realisoituneet. Keskeisimmät hyödyntämisen haasteet liittyvät datan ominaisuuksiin, datan saatavuuteen sekä resursseihin. Informaation yksityisyyttä koskeva sääntely nähdään tarpeellisena ja ajantasaisena nykypäivän dataympäristössä. Aiemmasta kirjallisuudesta poiketen tutkielmassa tunnistettiin vakuutusalan erityispiirteeksi korostunut tarve datan käsittelyn läpinäkyvyydelle, sekä havaittiin että kolmansien osapuolten rajoitukset vaikeuttavat datan saantia enemmän kuin informaation yksityisyyden sääntely.

Asiasanat: massadata, vakuutusala, informaatio, data-analytiikka, informaation yksityisyys.

ABSTRACT

Mäkelä, Arttu
Big data and the regulation of the privacy of information in the insurance industry – Thematic interview
Jyväskylä: University of Jyväskylä, 2022, 64 pp.
Information Systems, Master's thesis
Supervisor: Koskelainen, Tiina

The quantity of digital data has been growing exponentially during the last decades, and organizations from various line of businesses are seeking to find meaningful insights from the large sets of data to help their decision making. Insurance industry has exceptionally high potential in the usage of Big Data based on the special characteristics of the industry. Along with the potential, the challenges are also significant. In addition to the challenges caused by the basic nature of Big Data, some of the challenges are caused by the regulation regarding on the privacy of information, such as General Data Protection Regulation (GDPR). The goal of this thesis is to map out how the big data is used in the insurance industry, what is the potential of usage, and what are the main challenges of the usage. In addition, thesis strives to examine how the regulation regarding on the privacy of information influences the business of insurance companies. In addition to literature review, theme interview was conducted to interview experts of the topic. The current state of the Big Data usage on the insurance industry were clarified alongside with the fundamental challenges occurring in the utilization of Big Data. Also, the effects of the regulations regarding on the privacy of information were identified. The potential of the usage of Big Data is remarkable, but most of the expectations has not been met yet. The key challenges that were identified were related to properties and availability of Big Data, and resources needed in the utilization. The regulation of the privacy of information was perceived to be up-to-date and necessary in the data environment of today's insurance business. The interviews revealed that insurance companies have a high need for the transparency of their data utilization, caused by the bad reputation of insurance business.

Keywords: big data, insurance business, information, data analytics, information privacy

KUVIOT

TAULUKOT

SISÄLLYS

1 JOHDANTO

Digitaalisessa muodossa olevan datan määrä ja datajoukkojen koko on kasvanut eksponentiaalisesti viimeisten vuosikymmenten aikana. Eaton ym. (2012) arvioivat vuonna 2012, että vuoteen 2020 mennessä koko maailman datamäärä tulee ylittämään 35 zettatavua. Datan määrän kasvua kuvastaa toteamus, jonka mukaan heidän oma arvionsa on vanhentunut jo ennen kuin lukija on saanut teoksen käsiinsä. Arvio osui oikeaan, sillä vuonna 2022 datan määrän kasvu on nopeampaa kuin koskaan. Organisaatiot pyrkivät löytämään suurista datajoukoista merkityksellistä informaatiota, jota hyödynnetään päätöksenteon tukena (Gandomi & Haider, 2015). Suurista datajoukoista puhuttaessa käytetään yleisesti termiä massadata (*eng. Big Data),* jolle ei aiheen merkittävyydestä ja ajankohtaisuudesta huolimatta ole olemassa vakiintunutta määritelmää.

Massadataan yhdistetään usein kolme ominaisuutta, jotka erottavat sen muusta datasta. Nämä ominaisuudet ovat volyymi (*eng. volume),* nopeus (*eng. velocity)*, sekä moninaisuus (*eng. variety)* (Kitchin & McArdle; Eaton ym. 2012). Gandomi & Haider (2015) nostavat lisäksi esiin kolme olennaista massadataa kuvaavaa ominaisuutta, jotka ovat totuudenmukaisuus (*eng. veracity),* vaihtelevuus (*eng. variability),* sekä arvo (*eng. value)*. Kitchinin & McArdlen (2012), sekä Eatonin ym. (2012) mainitsemat ominaisuudet kuvastavat sitä, että massadataa syntyy jatkuvasti suuria määriä monista eri lähteistä ja tämä tapahtuu suurella nopeudella. Gandomin & Haiderin (2015) nostamat kolme ominaisuutta taas kuvastavat sitä, että jotkut massadatan lähteet tuottavat epäluotettavaa dataa, massadataa syntyy vaihtelevalla nopeudella, sekä sitä että massadatan arvotiheys on matala. Matalalla arvotiheys viitataan siihen, että dataa on analysoitava suuria määriä arvokkaan tiedon löytämiseksi.

Massadatan mahdollisuudet ovat erityisen lupaavia juuri vakuutusalalle sen erityispiirteiden, kuten liiketoiminnan datapainotteisuuden ja vakuutustuotteiden abstraktiuden vuoksi (Corbett, Schroek & Shockley, 2013; Hussain & Prieto, 2016). Vakuutusalan liiketoiminta perustuu datan analysoinnin kautta saavutettuun riskin ymmärrykseen ja arviointiin. Corbettin, Schroekin & Shockleyn (2013) mukaan 74 % vakuutusalan yhtiöistä pystyvät luomaan eroa kilpailijoihinsa nähden informaation ja analytiikan keinoin, kun vastaava luku muiden

alojen osalta on 61 %. Odotukset massadatan vaikutuksista vakuutusalalle ovat siis suuret, mutta odotusten täyttyminen on osoittautunut haastavaksi muun muassa massadatan perusominaisuuksien, kuten volyymin ja matalan arvotiheyden vuoksi. Lisäksi datan hyödyntämistä vaikeuttaa informaation yksityisyyden ja sitä koskevan sääntelyn huomioonottaminen (Hussain & Prieto, 2016). Verkkoliikenteen analysointityökalujen nopea kehitys on johtanut asiakkaista kerätyn datan volyymin, laadun ja nopeuden kasvuun. Kerätyn asiakasdatan kasvusta huolimatta yksilöillä on heikko käsitys mitä dataa heistä kerätään, ja miten sitä hyödynnetään (Richards & King, 2014). Informaation yksityisyyteen liittyvän tietoisuuden ja sääntelyn määrän lisääntyminen on vaikuttanut myös vakuutusyhtiöiden toimintaan. Vakuutusyhtiöt ovat tottuneet hallinnoimaan ja analysoimaan suuria määriä dataa, joka sisältää yksityisyyden loukkauksille altista sensitiivistä informaatiota. Lisääntynyt keskustelu ja sen aiheuttama sääntelyn tiukentuminen on pakottanut sensitiivistä tietoa käsittelevät organisaatiot tarkastelemaan datanhallinnan prosessejaan. Tunnetuin esimerkki valtioiden rajoja ylittävästä informaation yksityisyyttä koskevasta sääntelystä on Euroopan Unionin Tietosuoja-asetus eli GDPR.

Tässä tutkielmassa selvitetään massadatan mahdollisuuksia ja hyödyntämisen haasteita vakuutusyhtiöiden käytössä. Lisäksi pyritään selventämään informaation yksityisyyteen liittyvän sääntelyn vaikutuksia datan hyödyntämiseen vakuutusalalla. Vastauksen selvittämiseksi tutkielmassa vastataan seuraaviin tutkimuskysymyksiin:

- Miten vakuutusyhtiöt voivat hyödyntää massadataa?
- Mitä ovat vakuutusyhtiöiden olennaisimmat haasteet massadatan hyödyntämisessä?
- Miten informaation yksityisyyteen liittyvä sääntely vaikuttaa datan hyödyntämiseen vakuutusyhtiöissä?

Tutkielmassa lähdetään liikkeelle aiempaa teoriaa kartoittavasta kirjallisuuskatsauksesta, jonka avulla pyrittiin taustoittamaan tutkimusta perehtymällä aihepiirin käsitteisiin ja aiempaan tutkimukseen. Tutkielman empiirinen osuus toteutettiin laadullisena tutkimuksena. Aineiston keräämiseksi toteutettiin viisi teemahaastattelua.

Tutkielma koostuu johdannon lisäksi kuudesta luvusta, jotka ovat massadata, informaation yksityisyyden sääntely ja dataetiikka, tutkimusmenetelmät, tulosten pohdinta ja johtopäätökset, sekä yhteenveto. Ensimmäinen sisältöluku käsittelee massadatan käsitettä ja aihepiirin aiempaa tutkimusta. Toisessa sisältöluvussa keskitytään informaation yksityisyyden ja sen sääntelyn tutkimuksen ympärille, sekä luodaan katsaus dataetiikkaan. Kolmannessa luvussa esitellään tutkielmassa käytetyt tutkimusmenetelmät. Neljännessä luvussa esitellään empiirisen osion keskeiset tulokset. Viidennessä luvussa tarkastellaan tutkimuksen tuloksia, tehdään niistä tulkintoja ja johtopäätöksiä, sekä mainitaan tutkielman rajoitteet ja relevantit jatkotutkimusaiheet. Viimeinen luku käsittää tutkielman yhteenvedon.

2 MASSADATA

Datan analysoinnista on muodostunut keskeinen tekijä, joka siivittää innovaatiota ja kilpailua. Suurten datajoukkojen analysoinnin hyödyntäminen päätöksenteon, prosessien ja tuotteiden edistämisessä on muodostunut selviytymisen elinehdoksi. (Kitchens, Dobolyi, Li & Abbasi, 2018) Tässä luvussa pureudutaan massadatan käsitteeseen, sekä aihepiirin tutkimuksen avainteemoihin. Lisäksi luvussa käsitellään massadatan eri muotoja. Lopuksi luodaan katsaus massadatan hyödyntämisen haasteisiin ja mahdollisuuksiin vakuutusyhtiöiden liiketoiminnassa.

2.1 Massadata ja sen ominaisuudet

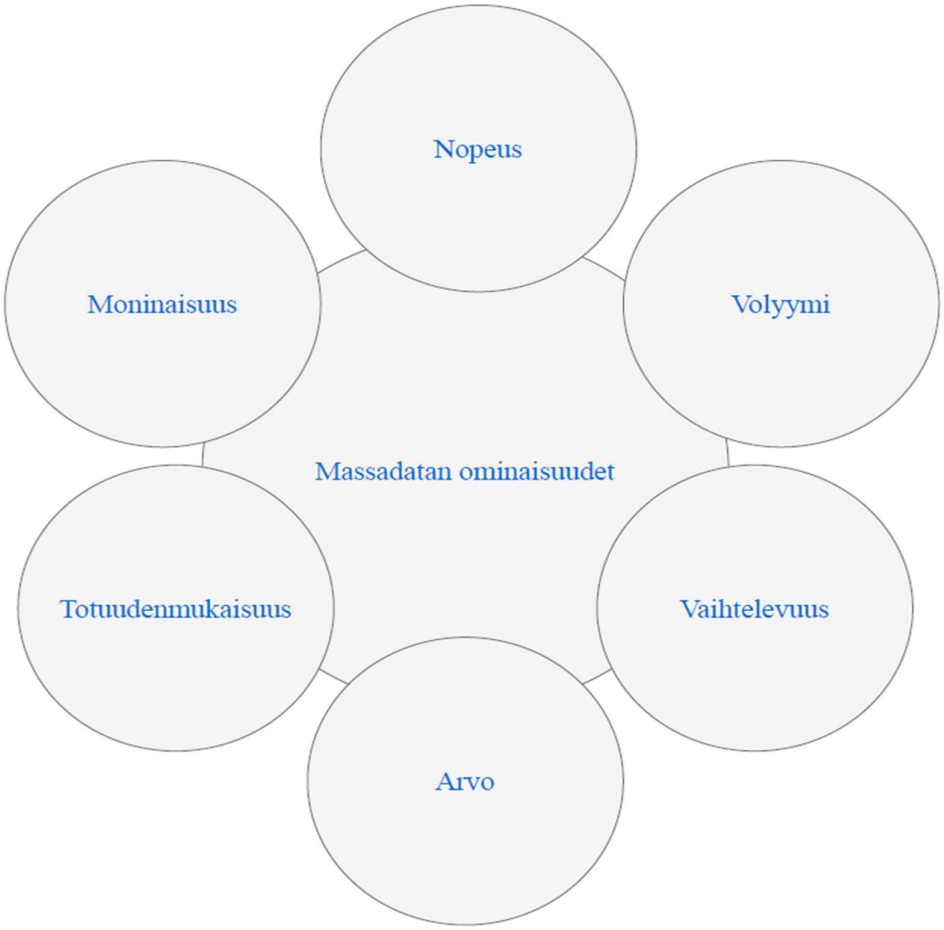
Datan, informaation ja tiedon käsitteet ovat toisilleen läheisiä ja keskinäisriippuvaisia, mutta käsitteiden välillä on kuitenkin selkeitä ja perustavanlaatuisiakin eroavaisuuksia. Kyseiset käsitteet menevät usein sekaisin etenkin arkikielessä, ja on tavallista, että datana puhutaan asiasta, joka on todellisuudessa informaatiota, ja toisinpäin. Tietojärjestelmätieteen tutkimuksessa data, informaatio ja tieto ovat koko tieteenalan ydinkäsitteitä, ja kyky niiden erojen tunnistamiseen ja ymmärtämiseen on tärkeää. Zins (2007) mukaan tietojärjestelmätieteen tieteenalan (*eng. Information Systems Science)* akateeminen kirjallisuus on määritellyt näitä käsitteitä moninaisin vivahtein, ja tiedeyhteisö ei ole saavuttanut yhteisymmärrystä käsitteiden riippuvuuksien luonteesta, tai käsitteiden merkityksistä ylipäätään. Data, informaatio ja tieto nähdään tutkijoiden toimesta usein toinen toisensa mahdollistavina käsitteinä, jossa data on informaation raaka-aine, ja informaatio vuorostaan tiedon raaka-aine. (Zinz, 2007) Zins (2007) huomauttaa, että mikäli näin on, tietojärjestelmätieteen) tutkimukseen ei pitäisi sisällyttää lainkaan tiedon tutkimista, koska tieto on tässä hierarkiassa informaation yläpuolella, ja siten englanninkielisen nimen mukaisesti ei sisälly tietojärjestelmätieteen (*eng Information Systems Science)* tutkimuksen piiriin. Loseen (1998) esittämän tieteenalasta

riippumattoman määritelmän mukaan informaatio on syvimmältä olemukseltaan prosessin tuotoksen tunnusmerkkejä, jotka kertovat prosessista sekä prosessin syötteistä.

Dataa on olemassa enemmän kuin koskaan aiemmin sen historian aikana (Zwitter, 2014). Jatkuvasti kasvavan datamäärän hallinnoiminen ja analysoiminen on muodostunut mahdottomaksi perinteisten datahallinnan menetelmien avulla sen suuren volyymin ja hajanaisuuden vuoksi (Davenport, Barth & Bean, 2012) Datamäärän suuren kasvun eräs keskeisistä syistä on se, että nykypäivänä informaation tuottaminen ja levittäminen on varsin vaivatonta (De Mauro, Greco & Grimaldi, 2015). Dataa, jota ei voida käsitellä perinteisin tiedonkäsittelyn menetelmin kutsutaan yleisesti massadataksi (*eng. Big Data).* Massadataan tutustuminen on käytännöllistä aloittaa sen käsitettä tarkastelemalla. Vaikka massadata on ilmiönä hyvin pinnalla ja ajankohtainen, sille ei kuitenkaan ole olemassa yleisesti käytössä olevaa määritelmää. De Mauron ym. (2015) mukaan vakiintuneen määritelmän puute johtuu massadatan nopeasta ja epäjohdonmukaisesta kehityksestä. Kitchinin & McArdlen (2016) mukaan käsitteen etymologia juontaa juurensa 1990-luvun puoliväliin, jolloin John Mashey käytti käsitettä ensimmäisen kerran. Mashey viittasi massadatan käsitteellä suurten datajoukkojen käsittelyyn ja analysointiin (Kitchin & McArdle, 2016). Eaton ym. (2012) määrittelevät massadatan olevan yksinkertaisesti joukko dataa, jota ei pystytä analysoimaan perinteisin menetelmin ja työkaluin. Sekä Kitchinin ja McArdlen (2016) että Eatonin ym. (2012) mukaan massadataa koskevat määritelmät sisältävät usein kolme olennaista ominaisuutta, jotka erottavat sen muusta datasta. Tätä kutsutaan englanniksi kolmen V:n malliksi, joka sisältää volyymin (volume), nopeuden (velocity) ja moninaisuuden (variety).

Lycettin (2017) mukaan volyymilla kuvastetaan sitä, että kyky käsitellä suuria datamääriä tuottaa organisaatiolle merkittävää hyötyä, ja että suuret datamäärät ovat toimivia datamalleja tärkeämpiä. Nopeudella taas painotetaan sitä, että datavirran tahti ja reaaliaikaisuus on olennainen tekijä hyötyjen talteen saamiseksi. Reaaliaikaisesta datavirrasta saatua informaatiota voidaan hyödyntää nopeasti, parhaimmillaan reaaliaikaisesti organisaatioiden päätöksenteossa. (Lycett, 2017) Nykypäivän yritysten välisessä kilpailussa kilpailuedun saavuttaminen voi olla kiinni sekunneista, ja siksi kyky reaaliaikaiselle analysoinnille on tärkeää (Eaton, ym. 2012). Moninaisuudella taas kuvastetaan massadatan varsin sekavaa luonnetta, sillä data on usein epäjohdonmukaista, virheellistä, sekä se tulee lukuisista lähteistä monessa eri muodossa (Lycett, 2017). Datan vaihtelevuuden kasvun myötä organisaatioita pakotetaan implementoimaan myös strukturoimatonta tai puolistrutkturoitua dataa päätöksenteossaan (Eaton ym. 2012). Strukturoitu data on jäsenneltyä ja helposti ennakoitavaa, kuten esimerkiksi asiakkaiden osoitetietoja ylläpitävän tietokannan tiedot. Strukturoimaton data on taas luonteeltaan hyvin vaihtelevaa ja epäjärjestelmällistä. Strukturoimaton data on vaihtelevaa niin sisällöltään, kooltaan kuin formaatiltaan, mikä tekee sen analysoinnista huomattavasti haastavampaa strukturoituun, tai puolistrukturoituun dataan nähden.) Suurin osa massadatasta on strukturoimatonta dataa (Gandomi & Haider, 2015; Das, T & Kumar, P, 2013).

Gandomi & Haider (2015) nostavat näiden ominaisuuksien lisäksi esille kolme olennaista massadataa kuvaavaa ominaisuutta, jotka ovat totuudenmukaisuus (veracity), vaihtelevuus (variability) sekä arvo (value). Todenmukaisuudella viitataan joillekin massadatan lähteille ominaiseen ominaisuuteen, eli epäluotettavuuteen (Gandomi & Haider, 2015). Totuudenmukaisuus vaikuttaa suoraan datan hyödynnettävyyteen, ja on siksi olennainen datalta vaadittava ominaisuus (Rubin & Lukoianova, 2013). Esimerkkinä epäluotettavasta, mutta silti arvoa sisältävästä datasta on ihmisten jakamat ajatukset sosiaalisessa mediassa. Vaihtelevuudella viitataan datavirran nopeuden muutoksiin. Datavirran nopeus ei tyypillisesti ole tasainen, vaan sillä on huippunsa ja suvantovaiheensa. Oraclen mukaan massadatan ”arvotiheys” on matala, mikä tarkoittaa, että datasta saatavilla oleva arvo on melko vähäistä datan määrään nähden. Siksi arvon seulominen massadatasta edellyttää kykyä käsitellä suuria määriä dataa. (Gandomi & Haider, 2015) Massadataan olennaiset ominaisuudet Gandomin & Haiderin (2015) mukaan esitelty kuviossa 1 (KUVIO 1).



KUVIO 1 Massadatan ominaisuudet (Gandomi & Haider, 2015)

2.2 Massadatan tutkimuksen avainteemat

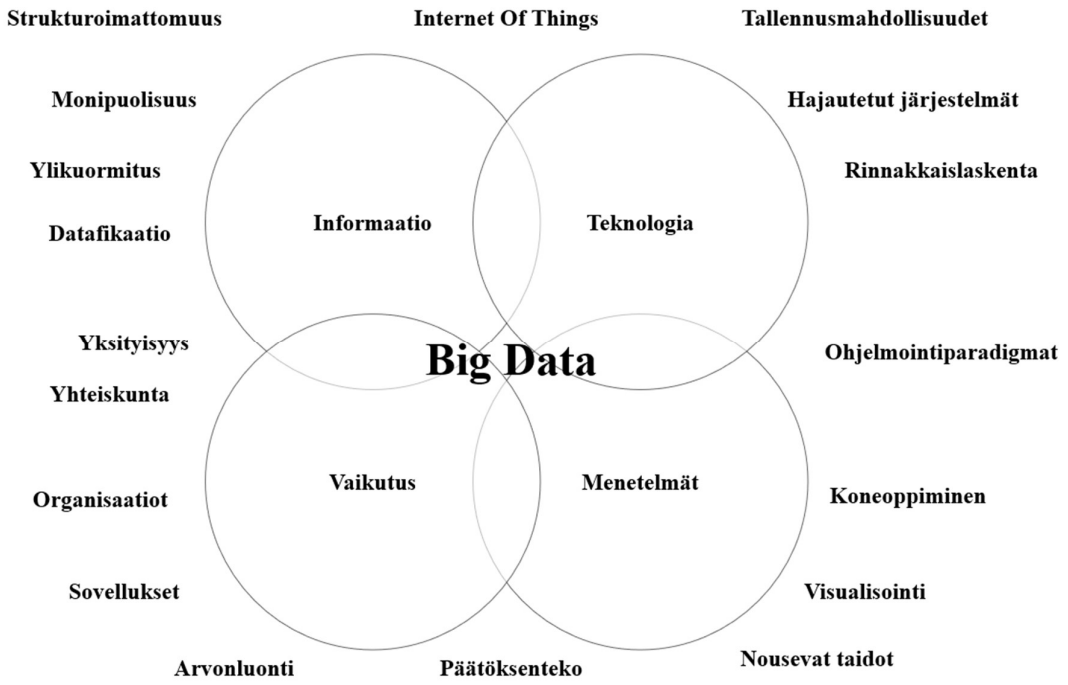
De Mauro ym. (2015) tarkastelivat massadatan käsitettä tutkimalla aihetta koskevaa tutkimusmateriaalia. He esittelivät tutkimuksensa perusteella neljä eri avainteemaa, joita käsiteltiin toistuvasti massadataa koskevissa tutkimuksissa. Nämä neljä teemaa olivat informaatio, teknologia, menetelmät ja vaikutus. Informaation helppo saatavuus on yksi merkittävimmistä syistä massadata-ilmiön olemassaololle, ja myös yksi tutkimuksissa toistuvista avainteemoista. Analogisen informaation konvertoimista digitaaliseen muotoon kutsutaan digitoinniksi. Digitointi mahdollistaa analogisen informaation siirtämisen ja varastoinnin vaivattomasti digitaalisessa muodossa. Informaation digitointi yleistyi ensimmäisten massadigitointi-projektien myötä, joissa kokonaisten kirjastojen kirjojen sisältö muutettiin kirjaimia tunnistavia optisia tunnistimia hyödyntäen digitaaliseen muotoon. Toinen olennainen massadata-ilmiötä ruokkinut tekijä on datafikaatio. (De Mauro ym. 2015) Datafikaatio ja digitointi ovat käsitteinä varsin lähellä toisiaan ja ovat täten helppo sekoittaa, mutta niillä on kuitenkin perustavanlaatuisia eroja. De Mauron ym. (2015) mukaan datafikaatio voidaan nähdä askeleena, joka seuraa informaation digitointia. Datafikaatio pyrkii järjestelemään ja suodattamaan digitaaliseen muotoon muutetun analogisen informaatioon siten, että siitä on mahdollista poimia hyödyllistä tietoa ja tehdä johtopäätöksiä (Demauro, ym. 2015). Lycettin, (2017) mukaan datafikaation käsitettä voidaan havainnollistaa kolmen eri konseptin avulla, jotka ovat dematerialisaatio, likvidikaatio, ja tiheys. Dematerialisaatiolla tarkoitetaan kykyä erottaa resurssin informaatioaspekti sen fyysisestä olomuodosta. Datan likvidaatiolla tarkoitetaan taas nykypäivän työkaluja, joilla informaatiota voidaan vaivattomasti manipuloida, liikuttaa ja tulkita tavoilla jotka olivat aikaisemmin kalliita, aikaa vieviä ja hankalia. Datan tiheydellä taas tarkoitetaan kykyä erottaa datan sekamelskasta parhaat datalähteet, joiden dataa voidaan hyödyntää sillä hetkellä tarvittavassa asiayhteydessä. (Lycett, 2017) Digitointi ja datafikaatio ovat kasvaneet merkittäviksi ilmiöiksi niitä edesauttavien laitteiden helpon saatavuuden vuoksi (Demauro, ym. 2015). Digitaalisia sensoreita sisältävät laitteet jotka ovat yhdistetty Internetiin ovat yleistyneet räjähdysmäisesti 2000-luvun aikana. Laitteiden digitaaliset sensorit mahdollistavat datan automatisoidun digitoinnin, ja digitoitu data voidaan kerätä ja järjestellä helposti hyödynnettävään muotoon. (De Mauro ym. 2015)

Massadata-termiä voidaan tarkastella valitsemalla lähtökohdaksi hallintaan vaadittavat työkalut, eli teknologiat. De Mauron ym. (2015) mukaan massadataa tarkastellaan usein sen hyödyntämisen mahdollistavien teknologioiden kautta, ja tämä onkin toinen heidän tutkimuksissa esiintyneistä avainteemoista. Massadatan hallintaa varten on täytynyt löytää uusia tapoja käsitellä dataa. Datajoukkojen massiivinen koko ja sen prosessointiin tarvittavien toimien monimutkaisuus asettavat suuret vaatimukset tallennustilalle ja laskentateholle, joihin perinteiset teknologiat eivät pysty vastaamaan. (De Mauro ym. 2015). Lisäksi suurten datamäärien hallinta perinteisillä teknologioilla on kallista ja aikaa vievää. Jossa massadatassa piilevä arvo saataisiin valjastettua hyötykäyttöön, sen

hallinnointiin vaaditaan tavallista suorituskykyisempiä, skaalautuvia ja joustavia teknologioita. (Oussos, Benjelloun, Lahcen & Belfkih, 2018) Massadatan käsittelyyn tarkoitetut teknologiat mahdollistavat datan analysoinnin ja johtopäätösten tekemisen parhaimmillaan reaaliajassa (Gandomi & Haider, 2015). De Mauron ym. (2015) mukaan tallennustila on eräs olennainen teknologian osaalue, joka asettaa rajoja datan määrälle. Datamäärän valtavan kasvun myötä yhä suurempia datamääriä tulisi saada pakattua fyysiseltä kooltaan pieniin laitteisiin. Tallennustilan on ennustettu kasvavan eksponentiaalisesti, mutta kasvun pitämien datamäärän kasvun vauhdissa vaatii kallista jatkuvaa tutkimusta sekä tuotekehitystä. (De Mauro ym. 2015)

Kolmas De Mauron ym. (2015) tutkimuksissa esiintyvä avainteema on massadatan käsittelyn menetelmät. Datamäärän kasvu on pakottanut kehittämään perinteisten tilastollisten menetelmien tilalle uudenlaisia keinoja ja tapoja, joilla massiivisesta ja hajanaisesta datajoukosta saadaan ulosmitattua arvoa. (De Mauro ym. 2015) Datajoukkojen suuren koon aiheuttamien haasteiden lisäksi, hyötyäkseen datasta organisaatioiden tulee pystyä analysoimaan dataa reaaliaikaisesti (Gandomi & Haider, 2015). De Mauron ym. (2015) mukaan analysointiin sopivien menetelmien hyödyntäminen edellyttää spesifiä osaamista ja tuntemusta erityisesti menetelmien potentiaalista ja rajoitteista. Tämänkaltaista osaamista on kuitenkin saatavilla työmarkkinoilla rajallisesti. (De Mauro ym. 2015)

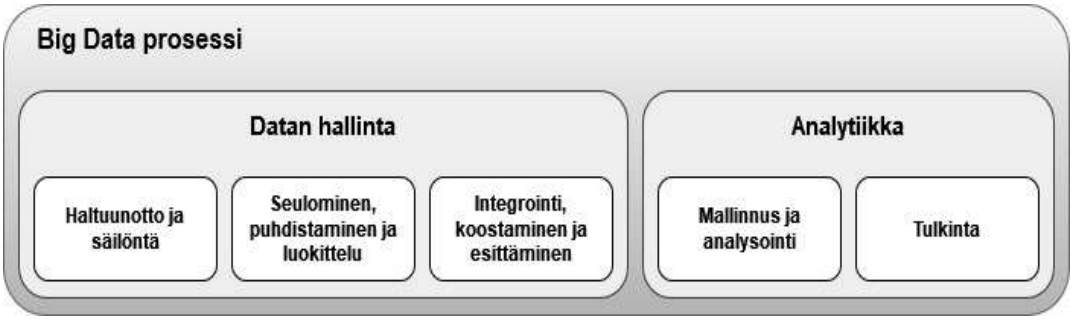
Neljäntenä tutkimuksissa esiintyvänä avainteemana De Mauron ym. (2015) tutkimuksessa nousee esiin vaikutus, eli millainen vaikutus massadatalla on yritysten ja yhteiskunnan toimintaan. Informaatiokeskeisyys korostuu nykypäivän yrityksissä alasta riippumatta, ja siksi usein samat informaatiota koskevat ratkaisut voivat olla toimivia myös konteksteissa, jotka ovat näennäisesti hyvin kaukana toisistaan. Esimerkiksi Googlen hakukoneessa tehtyjen hakujen korrelaatioanalysoinnin avulla on tehty päätelmiä liittyen influenssaepidemioihin, työttömyyteen sekä inflaatioon. Tyypillisesti massadatan vaikutuksia yrityksiin ja yhteiskuntaan kuvataan kertomalla tarinoita menetelmien ja teknologioiden implementoinnin onnistumisista. Kuitenkaan kaikki massadatan vaikutukset eivät ole myönteisiä. Eräs keskeisistä huolta aiheuttavista aiheista on sen vaikutus ihmisten yksityisyyteen, sillä datajoukon moninaisista lähteistä huolimatta tieto voi olla joissain tapauksissa pääteltävissä yksittäiseen ihmiseen. Lisäksi huolta aiheuttaa käyttäytymismallien analysoinnin keinoin saatavat ennusteet ihmisten käyttäytymisestä tulevaisuudesta, joka voi vaarantaa ihmisten vapaata tahtoa ja vapautta. Datan lähteiden hallinnan keskittyminen muutamien organisaatioiden varaan vaarantaa taas vapaata kilpailua ja nostaa saattaa lähteitä hallitsevat organisaatiot epäreiluun asemaan kilpailijoihinsa nähden. Esimerkkinä sosiaalisen median yritykset, joilla on ainoina yrityksinä pääsy luonteeltaan laajaan ja määrältään suureen sosiaaliseen dataan. Heillä on myös itseoikeus päättää ketkä kyseistä informaatiota voivat hyödyntää. (De Mauro ym. 2015) Demauron ym. (2015) tutkimuksessa löydetyt avainteemat ja niiden sisällöt esitellään kuviossa 2 (KUVIO 2)



KUVIO 2. Massadatan tutkimusta koskevat avainteemat ja niiden sisällöt (mukaillen De Mauro ym. 2015, s. 5)

2.3 Jatkuvan ja monimuotoisen datavirran hallintamenetelmät – Massadatan analytiikka

Suuret datajoukot itsessään ovat arvottomia, jos dataa ei pystytä muuttamaan päätöksenteon apuna käytettäväksi informaatioksi (Gandomi & Haider, 2015). Massadata on noussut tärkeäksi päätöksenteon välineeksi 2010- luvulla, ja organisaatiot ovat alkaneen hyödyntämään massadataa ennakoidakseen tulevaisuuden tapahtumia ja tapahtumien todennäköisyyksiä (George, Haas & Petland, 2014). Massadatan myötä organisaatiot ovat joutuneet vastaamaan uudenlaisiin suuren ja jatkuvasti kasvavan datamäärän aiheuttamiin haasteisiin. (Oussous ym. 2017). Gandomin ja Haiderin (2015) mukaan toimivat ja tehokkaat prosessit ovat edellytyksiä nopeasti liikkuvan ja moninaisen datavirran muuntamiseksi liiketoiminnan kannalta hyödyllisiksi johtopäätöksiksi. Johtopäätösten haravoimisen prosessi voidaan jakaa viiteen eri vaiheeseen, ja nämä vaiheet voidaan jakaa kahteen eri alaprosessiin: datan hallintaan ja analytiikkaan. Datan hallinta käsittää prosessit ja teknologiat joiden avulla data otetaan haltuun, säilötään sekä valmistellaan analytiikkaa varten. Toinen alaprosessi eli analytiikka taas käsittää tekniikat, joiden avulla datajoukosta saadaan tehtyä liiketoiminnan kannalta arvokkaita johtopäätöksiä. (Gandomi & Haider, 2015)



KUVIO 3 Massadatan analysoinnin prosessi. (mukaillen, Gandomi & Haider, 2015)

Davenportin, Barthin ja Beanin (2012) mukaan massadataa liiketoiminnassaan hyödyntävien organisaatioiden datan hallinnan menetelmät eroavat perinteisistä menetelmistä kolmella tavalla:

- He tarkkailevat datavirtaa samalla tavalla kuin pörssikursseja
- He luottavat data-analysoinnissa tietojenkäsitteytieteilijöihin, sekä tuote- ja prosessikehittäjiin data-analytiikkojen sijaan
- Analyytikot eivät toimi pelkästään IT:hen liittyvissä toiminnoissa, vaan heitä hyödynnetään myös ydinliiketoiminnassa, sekä operatiivisissa ja tuotekeskeisissä toiminnoissa.

Perinteisesti dataa on analysoitu ottamalla käsiteltäväksi tietty datajoukko, jolle suoritetaan useita analyyseja. Samoja menetelmiä ei voida soveltaa massadataan, sillä se virtaa jatkuvasti, ja useissa sen sovelluskohteissa datavirtaa täytyy pystyä analysoimaan reaaliaikaisesti haluttujen hyötyjen saavuttamiseksi. Reaaliaikaisen analysoinnin mahdollisuudet ovat moniulotteisia, sillä se mahdollistaa jo käynnissä olevien tapahtumien lopputuloksen ennustamisen (Davenport ym. 2012). Reaaliaikaisen analysoinnin mahdollisuuksia hyödynnetään yhä enenevissä määrin esimerkiksi terveydenhuolto-, logistiikka-, ja finanssialoilla (Oussous ym. 2017).

2.3.1 Tekstianalytiikka

Tekstianalytiikka on yhdistelmä tilastollisen analyysin, koneoppimisen ja tietokonelingvistiikan menetelmiä. Tekstimuotoisen datan lähteitä voivat olla esimerkiksi sähköpostit, keskustelufoorumeille lähetetyt viestit, tai verkossa julkaistut uutiset. Tekstin analysoinnilla tarkoitetaan tekniikoita, jotka pyrkivät louhimaan informaatiota tekstimuotoisesta datasta. (Gandomi & Haider, 2015) Morenon ja Redondon (2016) mukaan tekstin analysointi voidaan nähdä eräänlaisena tiedon louhinnan muotona, jonka pyrkimyksenä on löytää toistuvia kuvioita luonteeltaan strukturoimattomasta tekstistä. Lähteeltään ja luonteeltaan monipuoliseen tekstiin sisältyy paljon informaatiota maailmasta, sekä maailman entiteeteistä ja niiden välisestä vuorovaikutuksesta. Tekstin analysointi voi hyödyttää organisaatioita toimialasta riippumatta. Saavutettavia hyötyjä voi olla esimerkiksi

markkinoiden parempi ymmärrys, trendien ennakointi sekä parantunut riskien hallinta. (Moreno & Redondo, 2016)

Morenon ja Redondon (2016) mukaan suurten tekstimäärien analytiikan suurimpia ongelmia on tekstien suuren koon lisäksi niiden kyseenalainen laatu ja epätodenmukaisuus. Toisin kuin perinteisessä tekstianalytiikassa, massadatan tapauksessa analysoitava teksti ei ole esimerkiksi tietokoneohjelman tuottamaa tekstiä, joka on luotu analysoitavaksi, vaan teksti on ihmisten tuottamaa, ja siten hyvin vapaata niin muodoltaan kuin sisällöltään. Toisena keskeisenä ongelmana he mainitsevat datan saatavuuden. Esimerkkinä he mainitsevat Twitterin, joka omistaa hyvin suuren määrän ihmisten tuottamaa dataa twiittien muodossa, mutta mahdollistaa pääsyn vain murto-osaan käyttöliittymänsä kautta. Lisäksi palvelun esittämät twiitit ovat algoritmin valitsemia, mikä tekee johtopäätösten tekemisestä hankalaa. (Morenon & Redondon, 2016)

2.3.2 Äänianalytiikka

Äänimuotoisen datan analytiikka, eli äänianalytiikka, pyrkii poimimaan informaatiota äänen muodossa olevasta datasta. Äänianalytiikassa käytettävät teknologiat tallentavat, kategorisoivat, varastoivat sekä analysoivat ääntä (Shim ym. 2016). Puheluiden analysointi voi tapahtua myös reaaliaikaisesti (Gandomi & Haider, 2015). Kuten massadata yleensäkin, myös äänidata on luonteeltaan strukturoimatonta. Äänianalytiikka pyrkii useimmissa tapauksissa poimimaan informaatiota analysoimalla tallenteita ihmisen tuottamasta puheesta, ja siksi termejä äänianalytiikka ja puheanalytiikka käytetään usein ristiin. Shimin ym, (2016) äänianalytiikka on vähemmän tutkittu aihe verrattuna muihin liiketoiminta-analytiikan osa-alueisiin.

Äänianalytiikan tyypillisimmät soveltamisalat ovat puhelinasiakaspalvelukeskukset, sekä terveydenhoitoala (Gandomi & Haider, 2015). Shimin ym. (2016) mukaan myös sosiaalisessa mediassa, kuten Youtubessa oleva videomuotoinen data on yksi äänianalytiikan mahdollinen soveltamisala tukevaisuudessa. He listaavat äänianalytiikan hyödyiksi muun muassa asiakaspalvelupuheluiden laadun parantumisen, operationaalisen puolen ongelmien havaitsemisen sekä jopa organisaation tuotteisiin liittyvien puutteiden tai ongelmien automaattisen tunnistamisen (Shim ym. 2016). Gandomin & Haiderin (2015) mukaan saavutettavia hyötyjä ovat lisäksi asiakaspalvelijan työsuoritteen arviointi, myynnin kasvu, sekä organisaation käytäntöjen noudattamisen valvominen. Hirschberg, Hjalmarssom & Elhadad (2010) kertovat äänianalytiikan mahdollisuuksiksi terveydenhuollossa sekä potilaiden tilan monitoroinnin, että myös meneillään olevan hoidon tehokkuuden arvioinnin sellaisten sairauksien osalta jotka vaikuttavat potilaan tuottamaan puheeseen.

2.3.3 Videoanalytiikka

Videomuotoisen datan analysointi, eli videoanalytiikka, monitoroi, analysoi sekä poimii informaatiota videomuotoisesta datasta. Videoanalytiikka on vielä kehit-

tymätöntä muiden datamuotojen analysointiin nähden, mutta esimerkiksi verkkoon yhdistettyjen valvontakameroiden sekä videoiden jakamiseen perustuvien palvelujen nopean kasvun myötä myös videoanalytiikalla on nopea kasvutahti. Massadataa pidetään eräänä videoanalytiikkaa eteenpäin ajavana tekijänä (Gandomi & Haider, 2016) Ananthanarayanan ym. (2017) mukaan vuonna 2015 kehittyvillä markkinoilla asennettuja ja toiminnassa olevia valvontakameroita oli yksi kappale jokaista 29 ihmistä kohden, kypsillä markkinoilla taas yksi kamera kahdeksaa ihmistä kohden. Arvioiden mukaan valvontakameroiden määrä tulee kasvamaan 20 prosentin vuositahdilla. Videoanalytiikkaa hyödynnetään lukuisilla soveltamisalueilla, kuten liikenteen valvonnassa, sekä valvonnassa niin yksityisissä kuin julkisissa tiloissa. (Anthanarayanan, 2017) Videoanalytiikan hyödyntäminen valvontatehtävissä on tehokkaampaa niin kustannuksiltaan kuin vaikutuksiltaan. Gandomin & Haiderin (2016) mukaan turvallisuustehtävissä työskentelevän ihmisen keskittyminen valvontatehtäviin lakkaa noin 20 minuutin jälkeen. Automatisoitu valvonta taas kykenee tehokkaasti suorittamaan valvontatehtäviä ympäri vuorokauden.

Valvontakameroiden suuresta määrästä voidaan päätellä, niiden tuottaman datan määrä on hyvin suuri. Kuten tiedetään, videomuodossa olevan datan yksi suurimmista ongelmista on sen suuri tilan tarve. Yksi sekunti hd-tason videokuvaa vaatii saman verran tallennustilaa kuin 2000 sivua tekstiä (Gandomi & Haider, 2016). Videoanalytiikan haasteena on sen käyttökohteiden vaatimukset videon viiveettömyydelle, sekä korkeat suoritustehovaatimukset videota käsittelevältä laitteistolta. Datan volyymin, vaadittavan suoritustehon sekä viiveettömyyden vuoksi kamerat ovat haastavimpia massadataa tuottavia laitteita esineiden internetissä (Anthanarayanan ym, 2017).

2.3.4 Sosiaalinen media ja massadata

Sosiaalisen median massadatasta on muodostunut tärkeä tiedon lähde niin julkishallinnollisille kuin yksityisille, voittoa tavoitteleville organisaatioille. Informaation diffuusiolla tarkoitetaan prosessia, jossa informaatio leviää paikasta toiseen ihmisten välisen kanssakäymisen välityksellä. Sosiaalisen median massadatan analytiikka pyrkii merkityksellisiin oivalluksiin informaation diffuusiosta, ihmisten mielipiteistä sekä ajatuksista analysoimalla sosiaalisen median käyttäjien tuottamaa dataa, ja havaitsemalla yhteyksiä käyttäjien välillä (Stieglitz & Dang-Xuan, 2013). Tufecki (2014) vertaa sosiaalisen median tuottaman massadatan vaikutusta ihmisen käyttäytymisen tutkimukseen yhtä merkittävänä asiana, kuin mitä mikroskoopin keksiminen merkitsi astrologialle ja biologialle. Sosiaalinen media erottautuu muista massadatan lähteistä datan tuottavan ihmisjoukon valtavalla määrällä ja kirjavuudella. Eräs sosiaalisen median suosion syistä on sen tarjoama mahdollisuus julkiseen viestintään matalilla kustannuksilla (Stieglitz ym. 2018). Siten sosiaalisesta mediasta on muodostunut käyttäjilleen olennainen informaatiokanava. Facebook ilmoitti vuonna 2021, että vuoden ensimmäisellä neljänneksellä konsernin ydinpalveluita, Whatsappia, Facebook Messengeriä ja Instagramia käytti yli 3.5 miljardia ihmistä. (Statista, 2021). Sosi-

aalisen median laaja käyttäjämäärä tuottaa suuria määrä dataa suurella volyymilla, ja on siten merkittävä massadatan lähde. Suurten käyttäjämäärien tuottaman datan volyymi houkuttelee organisaatioita panostamaan sosiaalisen median massadatan analysointiin.

Sosiaalisen median dataa kerätään verkkosivuilta ja blogeista eri liiketoiminnan alojen liiketoiminnan ja päätöksenteon tueksi. Sosiaaliseksi mediaksi voidaan lukea suosituimpien sosiaalisen median alustojen, kuten Twitterin lisäksi myös keskustelufoorumit, wikikirjastot, podcastit, striimit, sekä verkkoyhteisöt (Zeng, Chen, Lusch, Li, 2010). Boyd ja Ellison (2007) määrittelee sosiaalisen median kolmen ominaisuuden kautta. Ensimmäisenä määrittävänä ominaisuutena on käyttäjien mahdollisuus luoda julkisia tai puolijulkisia profiileja. Toiseksi, käyttäjät pystyvät muodostamaan yhteyksiä toisiin käyttäjiin, ja siten luomaan verkostoja. Kolmantena ominaisuutena on käyttäjien mahdollisuus samaistua muihin käyttäjiin tarkastelemalla heidän profiilejaan ja tekemisiään. (Boyd & Ellison, 2007) Sosiaalinen media sisältää kaikkia edellisissä kappaleissa esiteltyjä datamuotoja, sillä käyttäjät tuottavat dataa niin tekstin, videon, äänen kuin kuvienkin muodossa (Ghani ym. 2019). Käyttäjien tuottaman datan lisäksi sosiaalinen media tuottaa transaktiodataa, joka syntyy käyttäjien vuorovaikutuksen seurauksan. Transaktiodata syntyy palvelussa olevien entiteettien, kuten ihmisten ja organisaatioiden vuorovaikutuksesta, ja on siten toinen informaation lähde sosiaalisessa mediassa (Gandomi & Haider, 2016; Felt, 2016). Lisäksi dataan voi sisältyä erilaista palveluiden keräämää tietoa, kuten dataa käyttäjien sijaintitiedoista. Esimerkkinä vuorovaikutuksen tuottamasta datasta käy käyttäjän tykkäys jostakin sivustosta, tai data joka kertoo profiileista, joilla tietty käyttäjä on vieraillut.

Sosiaalisessa mediasta saatava data voi olla luonteeltaan strukturoitua tai strukturoimatonta (Stieglitz ym. 2018). Esimerkkinä strukturoimattomasta datasta käy esimerkiksi twiitin tekstisisältö, kun taas strukturoitu data on vaikkapa kahden käyttäjän välinen parisuhdetieto. Hargittai (2018) tutki artikkelissaan sosiaalisen median massadatan mahdollisia vääristymiä, ja teki löydöksen jonka mukaan korkeassa sosiaaliekonomisessa asemassa olevat käyttävät todennäköisemmin useaa eri alustaa. Siten sosiaalisesta mediasta kerätty data korostaa hyvin toimeentulevien näkemyksiä ja mielipiteitä. (Hargittai, 2018)

Sosiaalinen media mahdollistaa datan suoran keräämisen sen käyttäjistä ilman perinteisten kyselytutkimuksien käyttöä, joka on ollut perinteisesti käyttäjien mieltymyksiin ja käyttäytymiseen liittyvän datan hankinnan edellytys (Abkenar ym. 2021). Sosiaalisen median tuottamaa massadataa pidetään porttina olennaiseen dataan ja johtopäätöksiin ihmisen käyttäytymisestä, jonka vuoksi sitä hyödynnetään niin tutkijoiden, yritysten, poliitikkojen toimittajien kuin myös hallitusten toimesta. (Tufekci, 2014). Kuten data yleensäkin, myös sosiaalisesta mediasta kerätty data on arvotonta, kunnes se saadaan muutettua tiedoksi, joka ohjaa organisaatioita parempaan päätöksentekoon (Ghani ym. 2019). Sosiaalista median datan analysoinnista saatavaa tietoa pyritään siis hyödyntämään niin voittoa tavoittelevien kuin sitä tavoittelemattomienkin organisaatioiden toi-

mesta. Sosiaalisen mediasta saatavan massadatan tyypillisimmät sovellutuskohteet ovat trendien havaitseminen, ja käyttäjien tunnetilojen sekä mielipiteiden merkitseminen ja analysointi (Ghani ym. 2019; Stieglitz ym. 2018). Käyttäjien tuottama data voi sisältää tietoa heidän ajatuksistaan, käyttäytymisestään tai mielipiteistään, ja tätä tietoa organisaatiot pyrkivät louhimaan sosiaalisen median massadatasta (Ghani ym. 2019). Tämänkaltaista tietoa voidaan käyttää hyödyksi päätöksenteossa, kuten esimerkiksi tuotekehityksen vieminen eteenpäin sosiaalisesta mediasta saadun asiakaspalautteen perusteella. Organisaatioille on myös tärkeää saada tietoa käyttäjästä josta data on peräisin, ja siksi datan suuren määrän lisäksi organisaatiot hyötyvät myös siitä, että sosiaalinen media mahdollistaa datan yhdistämisen sen tuottaneeseen käyttäjään (Stieglitz ym. 2018) Houkuttelevan datan lisäksi sosiaalinen media voi olla myös alustana tärkeä osa yritysten liiketoimintaa, ja edesauttaa monia liiketoiminnan osa-alueita, kuten markkinointia, tuotesuunnittelua, innovaatioita, sekä asiakassuhteiden ylläpitoa.

Ghanin ym. (2019) sosiaalisen mediasta saatavan massadatan analytiikan tekniikat ovat sidoksissa luonnollisen kielen prosessointiin, sentimenttianalyysiin, sosiaalisten verkostojen analyysiin, sekä uutisanalytiikkaan. Sentimenttianalyysi on eräänlainen yhdistelmä luonnollisen kielen prosessointia, tekstianalytiikkaa sekä statistiikkaa. Sosiaalisen median kontekstissa luonnollisen kielen prosessoinnilla tarkoitetaan pyrkimystä erotella relevanttia informaatiota suuresta määrästä ihmisten tuottamaa tekstiä, joka on kirjoitettu eri kielillä ja on muutenkin muodoltaan vapaata (Farzindar & Inkpen, 2015). Sentimenttianalyysin perimmäinen tarkoitus on löytää tekstistä, kuten esimerkiksi twiitistä, sen sisältämä tunnelataus, ja luokitella teksti sen perusteella sävyltään positiiviseksi, negatiiviseksi tai neutraaliksi. Sentimenttianalyysi luokittelee siten sosiaalisen median käyttäjien asenteita, tunteita ja mielipiteitä liittyen esimerkiksi tietyn yrityksen tuotteisiin, brändiin tai palveluihin. (Ghani ym. 2019) Sosiaalisten verkostojen analyysi (*eng. Social Network Analysis)* on joukko tekniikoita, jotka analysoivat sosiaalisten verkostojen rakennetta sekä kokonaisuutena, että myös yksittäisen nystyrän (*eng. Node*) osalta (Cheong & Cheong, 2011). Sosiaalisen verkoston analyysi on muodostunut tärkeäksi työkaluksi sosiaalisen median analysoinnissa esimerkiksi sosiologian ja huijauksenestonsovelluksissa (Ghani ym. 2019). Uutisanalytiikka kokoaa yhteen reaaliaikaisesti uutisia kaikista saatavilla olevista uutislähteistä, ja analysoi ne saadakseen tietoa esimerkiksi tiettyyn yritykseen liittyvästä uutisoinnista ja uutisoinnin sävystä (Chowdhury, Routh & Chakrabarti, 2014). Uutisanalytiikka on siten samankaltaista sentimenttianalyysin kanssa, mutta analysoitavan datan lähteenä on sosiaalisen median käyttäjien sijaan uutistoimistot.

2.4 Massadata vakuutusalalla

Vakuutusala on luonnostaan hyvin datakeskeinen, sillä sen liiketoiminta perustuu datan analysoinnin kautta saavutettuun riskin ymmärrykseen ja arviointiin. Muiden alojen joukossa myös vakuutusala on suuressa muutoksessa,

ja on suuntautumassa entistäkin vahvemmin dataperustaiseksi. Vakuutusalan organisaatiot eivät tuota ja myy lainkaan fyysisiä tuotteita ja myös siksi data on niiden toiminnassa keskeisessä roolissa. Vakuutusyhtiöt ovat jo pitkään keränneet dataa useasta eri lähteestä, kuten asiakkailtaan ja muilta yrityksiltä. Kerättyä dataa on hyödynnetty usealla liiketoiminnan osa-alueella, kuten tuotehinnoittelussa ja riskien määrittelyssä, ja käytännössä jokainen vakuutusalan toimijoiden tekemä tärkeä päätös on datan ohjaamaa (Corbett, Schroek & Shockley, 2013). Vakuutusalan toimijat toimivat nykyään markkinoilla, jossa asiakkailla on reaaliaikainen pääsy yhä moninaisimpiin vakuutustuotteisiin ja palveluihin. Lisäksi kentälle on tullut internet-perustaisia toimijoita, jotka haastavat perinteisiä yhtiöitä hyödyntämällä informaatiota ja analytiikkaa tarjoten uusia hinnoittelumalleja. (Corbett ym. 2013)

Vakuutusyhtiöiden perinteisesti hyödyntämä data on luonteeltaan strukturoitua dataa, josta ulosmitataan arvoa perinteisin data-analytiikan menetelmin. Suurissa, puolistrukturoiduissa ja strukturoimattomissa datajoukoissa piilee kuitenkin merkittävä määrä arvoa, jota ei kyetä vielä täysimääräisesti ulosmittaamaan. Datajoukkoja, jotka sisältävät runsaasti vielä hyödyntämätöntä kaupallista arvoa ovat esimerkiksi korvaustoimintaan liittyvä tekstimuotoinen dokumentointi (Hussain & Prieto, 2016). Jiang & Song (2016) uskovat että massadatan analysoinnin kyky tulee olemaan yhä keskeisempi kilpailua määrittävä tekijä vakuutusyhtiöiden välillä. Massadatan analytiikka onkin eräs merkittävä vakuutusalan organisaatioita ja niiden prosessien muutosta eteenpäin vievä ajuri, joka lupaa ratkaisuja useisiin alan pitkäkestoisiin haasteisiin. (Corbett ym. 2013). Datan ja analytiikan hyödyntämisen tehostumisen uskotaan lopulta muuttavan koko alan kehityksen suunnan (Boobier, 2016). Korkeista odotuksista ja aiheen merkittävyydestä alalle huolimatta aihetta on tutkittu tiedeyhteisöissä kovin vähän.

2.4.1 Mahdollisuudet

Massadatan mahdollisuudet ovat erityisen lupaavia juuri vakuutusalalle sen erityispiirteiden, kuten datapainotteisuuden vuoksi (Corbett ym. 2013; Hussain & Prieto, 2016). Corbettin ym. (2013) mukaan 74 % prosenttia vakuutusalan yhtiöistä pystyvät luomaan eroa kilpailijoihinsa nähden informaation ja analytiikan keinoin, kun vastaava luku muiden alojen osalta on 61 %. Finanssialan organisaatiot erottautuvat muista aloista heidän hallussa olevan datan perusteella. Vakuutusyhtiöillä on käytössään suuret varastot strukturoitua dataa, joka on pääosin tuotettu organisaation sisällä. (Hussain & Prieto, 2016). Massadatan sovelluskohteita vakuutusalalla on lukuisia, ja se tuo uusia mahdollisuuksia lähes jokaiselle sen liiketoiminnan osa-alueelle. Hussainin & Prieton (2016) mukaan massadatan avulla alan toimijat voivat saavuttaa paremman ymmärryksen markkinoista, asiakkaistaan, kanavistaan, tuotteistaan, sääntelystä, kilpailijoista, toimittajista sekä työntekijöistään. Cavanillas, Curry ja Wahlsterin (2016) mukaan yksi tärkeimmistä massadatan sovelluskohteista on asiakaskokemuksen parantaminen ja asiakkaan sitouttaminen hyödyntämällä uusia strukturoimattomia data-

lähteitä, kuten sosiaalista mediaa. Massadataa voidaan hyödyntää myös vakuutusten väärinkäytön havaitsemisessa, riskinvalinnassa, sekä tuotehinnoittelussa (Fang, Jiang & Song, 2016).

Massadata mahdollistaa suurille asiakasjoukoille luotujen vakuutustuotteiden ja sopimusten yksilöinnin, jolloin vakuuttamisessa voidaan ottaa huomioon asiakkaiden yksilölliset tarpeet, ja toisaalta saadaan mahdollisimman tarkkaa tietoa vakuutettavista riskeistä. Käyttöön perustuva hinnoittelu ei ole uusi ilmiö vakuutusalalla, mutta massadata mahdollistaa niiden laajamittaisen hyödyntämisen ja parantaa tuottavuutta (Baesens ym. 2016). Tämä mahdollistaa muun muassa tarkan asiakaskohtaisen hinnoittelun. Vakuutusyhtiöt ovat alkaneet hyödyntää esimerkiksi asiakkaiden käyttämien puettavien älylaitteiden keräämää dataa terveysvakuutustensa yksilöimiseksi. Ajoneuvovakuutuksissa taas on hyödynnetty ajoneuvoihin asennettuja sensoreita, jotka keräävät dataa kuljettajan ajotyylistä, jonka avulla vakuutusta voidaan hinnoitella ja räätälöidä asiakkaan tarpeiden ja vakuutettavan riskin mukaisesti.

Suuret yhtiöt ovat tavallisesti halukkaampia ottamaan käyttöön uusia teknologioita nopeasti, mutta data-analytiikkaan panostaminen voi mahdollistaa nopeaa kasvua myös pienemmille alan toimijoille. Toimivat data-analytiikan prosessit voivat tehdä pienemmistä yrityksistä kannattavampia, sekä vähemmän alttiita markkinoiden heilahtelujen aiheuttamille vahingoille. Lisäksi data voi tarjota heille syvemmän ymmärryksen omasta liiketoiminnastaan. (Boobier, 2016)

2.4.2 Haasteet

Massadatan haasteista vakuutusalalla on löydettävissä yhteneväisyyksiä edellisessä luvuissa mainittuihin massadatan yleisiin haasteisiin, kuten informaation yksityisyyden aiheuttamat pulmat. Vakuutusyhtiöt hallitsevat ja analysoivat suuria määriä dataa, joka pitää sisällään runsaasti yksityisyyden loukkauksille altista informaatiota, kuten tietoja asiakkaidensa terveydentilasta tai taloudellisesta tilanteesta. Täyttääkseen informaation yksityisyyttä koskevat vaatimukset, alan toimijoiden tulee noudattaa lakeja, asetuksia ja käytänteitä, kuten esimerkiksi Euroopan Unionin yleistä tietosuoja-asetusta eli GDPR:ää. Asetusten vaikutukset finanssialan yrityksille ovat merkittävät, sillä asiakkaat voivat vaatia heitä koskevien henkilötietojen poistamista tai hyödyntämättä jättämistä (Hussain & Prieto, 2016). Asetusten noudattaminen ja massadatan täysimittainen hyödyntäminen osoittautuvat usein haastavaksi tai jopa mahdottomaksi. Hussainin & Prieton (2016) mukaan sääntelyn epäselvyys siitä, miten asiakkaista kerättyä dataa saadaan hyödyntää omistajien toimesta hidastaa massadatan hyödyntämisen yleistymistä.

Boobierin (2016) mukaan pienten ja suurten vakuutusyhtiöiden välillä on eroja uusien teknologioiden käyttöönoton halukkuudessa. Pienet vakuutusyhtiöt kohtaavat kokonsa vuoksi suurempia riskejä investoidessaan uusiin teknologioihin. Siksi pienemmät yhtiöt saattavat olla haluttomampia olemaan teknologisesti aallonharjalla, ja sen sijaan odottavat, että teknologiat on testattu toimiviksi suu-

rempien yhtiöiden toimesta. Toisaalta suuremmat yhtiöt ovat rakenteeltaan pienempiä monimutkaisempia, joka vaikeuttaa nopeiden muutosten toteuttamista, vaikka siihen olisi halukkuutta. (Boobier, 2016)

Massadatalle asetetaan suuria odotuksia vakuutusalan organisaatioiden ja yritysten toimesta, mutta odotusten täyttäminen osoittautuu usein haastavaksi massadatalle ominaisten ominaisuuksien, kuten volyymin vuoksi. Vakuutusyhtiöt kykenevät yhä paremmin tunnistamaan tuottamansa datan arvon, mutta kompastuvat usein haasteisiin yrittäessään selvittää miten löytää suurista datamääristä juuri heitä hyödyttävää dataa, ja miten muuntaa tämä data hyödylliseksi informaatioksi (Boobier, 2016; Hussain & Prieto, 2016). Toinen massadatan keskeiseen ominaisuuteen liittyvä haaste on epävarmuus siitä, että tietty datajoukko edustaa koko populaatiota (Hussein & Prieto, 2016). Suuresta koosta huolimatta dataa analysoivat vakuutusyhtiöt eivät voi olla varmoja, että datajoukkoon sisältyy tarkasteltava kohderyhmä kokonaisuudessaan. Lisäksi, hyötyjen ja odotusten täyttymiseen vaadittavia ominaisuuksia kyetään tunnistamaan yrityksissä heikosti ja siksi odotusten ja todellisten saavutettujen hyötyjen välistä eroa on vaikea kaventaa. (Tiefenbacher & Olbrich, 2015). Hussain & Prieto (2016) kertovat eräänä haasteena, että vakuutusalan yhtiöt eivät ole kyenneet tehokkaasti hyödyntämään massadatan eri tyyppejä muihin aloihin verrattuna. Äänimuotoinen ja sosiaalisen median tuottama massadata ovat datatyyppejä, jotka sisältävät vielä paljon hyödyntämätöntä arvoa. Tämän he katsovat ainakin osittain johtuvan siitä, että organisaatiot ovat vielä keskittyneet omien suurikokoisten strukturoidusta datasta koostuvien varastojen hyödyntämiseen. (Hussain & Prieto, 2016)

Boobierin (2016) mukaan eräs massadatan hyödyntämistä hidastava tai estävä tekijä on analytiikkaa varten tarvittavan organisaation perustaminen ja hallinnointi. Massadatan hyödyntäminen vaatii nykyaikaisen, toimivan organisaation, järjestelmät ja ohjelmistot. Hussainin & Prieton (2016) mukaan finanssi- ja pankkialan yritysten data-analytiikan kehittämisen hidasteena on heidän käyttämät luotettavat, mutta vanhanaikaiset järjestelmät. Massadatan hyödyntämiseen tarkoitettujen sovellusten integrointi näihin järjestelmiin on usein tehotonta verrattuna siihen, että massadataa varten rakennetaan oma itsenäinen järjestelmä. Lisäksi hyödyntäminen edellyttää erityisosaamista ja tietotaitoa, jonka avulla data ja siitä tehdyt havainnot voidaan muuntaa käyttökelpoiseksi ratkaisuksi (Boobier, 2016; Hussain & Prieto, 2016). Tietotaidon hankkimisen ja kehittämisen esteenä on usein organisaatiokulttuuri, joka ei tunnista massadatan tuomia mahdollisuuksia yrityksen ydinliiketoiminnalle (Hussain & Prieto, 2016).

3 INFORMAATION YKSITYISYYDEN SÄÄNTELY JA DATAETIIKKA

Verkkoliikenteen analysointityökalujen nopea kehitys on johtanut asiakkaista kerätyn datan volyymin, laadun ja nopeuden kasvuun ja datan analysoinnista on muodostunut keskeinen tekijä joka siivittää innovaatiota ja kilpailua.) Kerätyn ja analysoidun datan määrän kasvusta huolimatta yksilöillä on edelleen heikko käsitys siitä, mitä dataa heistä kerätään ja mitä tietoja jaetaan kolmansien osapuolten kanssa (Richards & King, 2014). Digitalisaation ja datafikaation väitetään arkikeskustelussa usein lopulta hävittävän yksityisyyden. Richardsin ja Kingin (2014) mukaan tämä ei pidä paikkaansa, vaan häviämisen sijaan yksityisyyden luonne kehittyy ja muuttuu yhteiskunnan mukana. Kappaleessa 2.1 käsiteltiin datan, informaation ja tiedon käsitteitä. Kirjallisuuden perusteella data on informaation raaka-aine ja alakäsite. Tässä tutkielmassa käytetään termiä informaation yksityisyys puhuttaessa datan tai informaation yksityisyydestä, sekä niihin liittyvästä sääntelystä. Keskustelun lisääntyessä ja tietoisuuden kasvaessa informaation yksityisyyden sääntelyyn on alettu kohdistamaan enemmän huomiota. Tässä luvussa luodaan katsaus datan ja informaation yksityisyyden käsitteisiin, sekä informaation yksityisyyden sääntelyyn, kuten Euroopan Unionin tietosuoja-asetukseen GDPR:ään. Lisäksi luvussa käsitellään dataetiikan käsitettä, ja tarkastellaan dataetiikkaa massadatan kontekstissa.

3.1 Datan ja informaation yksityisyys

Belangiern ja Crossierin (2011) mukaan Mason (1986) ennusti 1980-luvun puolivälissä informaatioteknologian käytön lisääntymisen ja digitalisaation aiheuttavan ongelmia neljällä osa-alueella, jotka olivat yksityisyys, virheettömyys, omistajuus sekä saavutettavuus. Belangier ja Crossier (2011) toteavat näiden ennustuksien osuneen oikeaan jokaisen osa-alueen osalta, mutta erityisen hyvin koskien yksityisyyttä. Yksityisyys on ollut tutkimuksen ja keskustelun aiheena jo

vuosisatojen ajan. Digitalisaation myötä yksityisyydelle on syntynyt paljon tutkittu alakäsite informaation yksityisyys, joka käsittää yksilöiden kommunikaation yksityisyyden sekä datan yksityisyyden. Kuten tietojärjestelmätieteen tutkimus yleensä, informaation yksityisyys on aiheena poikkitieteellinen, ja sitä on tutkittu tietojärjestelmätieteen lisäksi myös esimerkiksi markkinoinnin, oikeustieteiden, johtamisen ja psykologian tieteenalojen toimesta. Westin (1970) määrittelee yksityisyyden olevan oikeus, joka antaa yksilölle vallan määrittää milloin, mitä, ja minkä verran itseään koskevaa informaatiota saa levittää. Richards & King (2014) puolestaan määrittelevät yksityisyyden olevan informaatiota koskevia laaja-alaisia sääntöjä jotka eettisin menetelmin ohjaavat informaation virtaa soveliaaseen suuntaan. Belangier ja Crossier (2011) toteavat informaation yksityisyydelle olevan olemassa useita määritelmiä, mutta määritelmien avainsisällössä ei esiinny suurta vaihtelua. Tyypillisesti määritelmä viittaa yksilöiden haluun hallita heitä koskevan informaation sekundääristä käyttöä. Informaation sekundäärisellä käytöllä tarkoitetaan tapausta, jossa kerättyä informaatiota käytetään johonkin muuhun kuin sen alkuperäiseen tarkoitukseen. (Belangier & Crossier, 2011)

3.2 Massadata ja yksityisyys

Datan määrän kasvu on kasvattaa myös riskiä datan yksityisyyden loukkauksille, ja sitä pidetäänkin eräänä massadatan merkittävimmistä haittapuolista (Soria-Comas & Domingo-Ferrer, 2015). Määrän nopean kasvun lisäksi myös ihmisistä kerätyn datan luonteen muutokset lisäävät riskiä arkaluonteisen datan yksityisyyden loukkauksille. Massadata muodostaa kenties digitaalisen ajan suurimman haasteen dataa ja yksityisyyttä koskevan lainsäädännön laatimiselle (Zarsky, 2017). Soria-Comasin & Domingo-Ferrerin (2015) mukaan massadata ja se hyödyntäminen on monilta osin ristiriitaista olemassa olevien henkilökohtaista dataa koskevien periaatteiden, kuten suostumuksellisuuden, käytön rajaamisen oikeuden sekä läpinäkyvyyden ja avoimuuden kanssa. Sosiaalisen median suosion kasvu on eräs merkittävä yksittäinen tekijä, joka on lisännyt yksityisyyden loukkausten riskejä, sillä sosiaalinen media kerää suuren määrän henkilökohtaista tietoa käyttäjiensä elämästä ja sosiaalisista suhteista. (Mehmood, Natgunanathan, Ciang, Hua, Guo., 2016) Perera ym. (2015) nostavat myös internetiin liitettyjen esineiden arkipäiväistymisen, eli esineiden internetin, kasvun yhdeksi seikaksi joka mahdollisesti johtaa käyttäjistä kerätyn datan yksityisyyden vaarantumiseen.

Mehmoodin ym, (2016) mukaan käyttäjien yksityisyyttä voidaan loukata kolmenlaisissa tilanteissa. Ensiksi yksityisyyttä voidaan loukata tilanteissa, joissa käyttäjän henkilötiedot yhdistetään sensitiivistä tietoa sisältävään dataan. Tämänkaltainen tilanne voi syntyä esimerkiksi silloin kun tietyn käyttäjän henkilötiedot yhdistetään hänen ostomieltymyksiin verkkokaupoissa. Myös Nunan & Di Domenico (2013) sekä Herschel & Miori (2017) mainitsevat datan uudelleenkäytön, eli sekundäärisen käytön eräänä merkittävänä käyttäjien yksityisyyttä

vaarantavana tekijänä. Tietty datajoukko itsessään ei välttämättä sisällä käyttäjiä loukkaavaa dataa, mutta asia voi muuttua, kun datajoukko yhdistetään toiseen datajoukkoon. Esimerkkinä tämänkaltaisesta tapauksesta Nunan & Di Domenico (2013) mainitsevat tapauksen, jossa pystyttiin tunnistamaan seuranhakupalveluiden käyttäjien henkilöllisyys hyödyntämällä kasvojentunnistusteknologiaa ja Facebookista julkisesti saatavilla olevia profiilikuvia. Toinen käyttäjien yksityisyyden loukkaamisen riskiä lisäävä tekijä on henkilökohtaisen informaation kerääminen ja valjastaminen tuottamaan arvoa jollekin liiketoiminnalle. Kolmantena käyttäjien yksityisyyden loukkaamisen riskiä lisäävänä tekijänä Mehmood ym. (2016) mainitsevat tietoturvaltaan puutteelliset käytännöt liittyen datan varastointiin ja prosessointiin. Nunan & Di Domenico (2014) mainitsevat niin ikään tietoturvallisuuden yhdeksi massadatan yksityisyyden haasteeksi. Massadatajoukot eivät ole usein kadotessaan korvattavissa. (Nunan & Di Domenico, 2014). Soria-Comas & Domingo-Ferrer (2015) mainitsevat lisäksi yksityisyyttä uhkaaviksi tekijöiksi dataa hallinnoivien organisaatioiden työntekijöiden sisäiset väärinkäytökset, luvattoman datan sekundäärisen käytön, yritysten muuttuvat käytännöt sekä hallitusten pääsyn dataan ilman laillista perustetta. Datan yksityisyyden ja massadatan väliset ristiriidat ovat herättäneet keskustelua siitä, että millä kustannuksilla yksityisyyttä tulisi vaalia, sillä liian tarkka yksityisyyden suojaaminen mahdollisesti hidastaa teknologian kehittymistä (Soria-Comas & Domingo-Ferrer, 2015). Kaikki käyttäjistä kerätty data ei kuitenkaan ole kovinkaan arkaluonteista. Soria-Comas & Domingo-Ferrer (2015) kertovatkin että erään näkemyksen mukaan yksityisyyttä voidaan vaalia ilman kehityksen hidastamista keskittymällä vain sellaiseen datan yksityisyyden ongelmiin, jonka väärinkäyttö voi aiheuttaa harmia niille, joista data on kerätty.

Nunan & Di Domenico (2014) mainitsevat jatkuvasti yleistyvän automaattisen tiedonkeruun olennaisena yksityisyyttä vaarantavana tekijänä. Datan automaattinen kerääminen on osittain ihmisestä riippumaton prosessi. Heidän mukaan laitteiden automaattisesti keräämän datan suuri volyymi ja sen analysoinnin nopeuden vaatimukset kertovat yksiselitteisesti siitä, että käyttäjien suostumusta ei oteta prosessissa huomioon. Myöskään Herschelin ja Miorin (2017) mukaan automaattisen tiedonkeruun menetelmät eivät ota huomioon yksityisyyden säilyttämiseen ja luottamukselliseen liittyviä seikkoja lainkaan. Ihmisen osallistuminen prosessiin ei muuta tilannetta, sillä yksityisyyden huomioiminen tekee datan suuren volyymin vuoksi koko tiedonkeruuprosessista epäkäytännöllisen.

3.3 Euroopan Unionin yleinen tietosuoja-asetus (GDPR)

Dataan ja sen hyödyntämiseen liittyvä eettinen valvonta on pitkälti datan valmistajien, toimittajien käyttäjien käsissä, sillä sitä koskevien linjausten laatimisesta vastaavaa absoluuttista auktoriteettia ei ole olemassa (Hand, 2018). Eräs

tunnetuimpia esimerkkejä tietojenkäsittelyn laajamittaisesta, maiden rajoja ylittävästä valvonnasta on Euroopan Unionin laatima yleinen tietosuoja-asetus (*Eng. GDPR, General Data Protection Regulation)*, jota on sovellettu kaikissa unionin jäsenmaissa keväästä 2018 lähtien (Europa.eu, 2021). GDPR on laadittu siitä lähtökohdasta, että yksityisyys lukeutuu perustavanlaatuisiin ihmisoikeuksiin, ja sen piiriin lukeutuu kaikki EU:n alueella asuvat ihmiset riippumatta siitä missä heitä koskevan datan prosessointi itsessään tapahtuu (Goddard, 2017).

3.3.1 GDPR:n periaatteet

GDPR asettaa yrityksille ja organisaatioille luonnollisia henkilöitä koskevan henkilökohtaisen datan eli henkilötietojen prosessoinnin vaatimukset (EU) 2016/679, 5. artikla). GDPR määrittelee henkilötiedon olevan mitä tahansa informaatiota, joka liittyy tunnistettuun tai tunnistettavissa olevaan luonnolliseen henkilöön. Tarkemmin, henkilötiedoilla tarkoitetaan dataa, jonka perusteella voidaan suorasti tai epäsuorasti tunnistaa yksittäinen ihminen, sekä erityisesti dataa joka sisältää suoranaisia tunnistetietoja, kuten IP-osoitteita tai sijaintitietoja (Goddard, 2017). GDPR edellyttää suurimmassa osassa tapauksia, että henkilö josta dataa kerätään, on antanut suostumuksensa. GDPR sisältää kuusi datan yksityisyydensuojaamisen periaatetta:

- Oikeudenmukaisuus
- Käyttötarkoituksen rajaaminen
- Datan minimointi
- Eheys ja luottamuksellisuus
- Säilytyksen rajaaminen
- Datan virheettömyys

Malgieri (2020) erittelee oikeudenmukaisuuden periaatteen kahteen nyanssiin, jotka ovat läpinäkyvyys ja lainvoimaisuus. Myös Goddardin (2017) mukaan datan käsittely voi olla oikeudenmukaista ainoastaan silloin kun dataa hallinnoiva taho kommunikoi selkeästi ja yleiskielisesti datan käsittelyyn liittyvistä käytännöistä, sekä datan kohteen oikeuksista. Malgierin (2020) johtopäätöksen mukaan oikeudenmukaisuuden periaatteella tarkoitetaan GDPR:n kontekstissa tasapainottelua datan kohteen ja datan hallinnoijien intressien huomioimisessa datan käsittelyä koskevassa sääntelyssä. Tämän näkemyksen mukaan GDPR keskittyy sovittujen käytäntöjen noudattamisen korostuksen sijaan ennaltaehkäisemään tilannetta, jossa datan kohteen ja hallinnoijien intressit on huomioitu epäsuhtaisesti ja joka aiheuttaa näin ollen haavoittuvuuksia jommallekummalle osapuolelle. (Malgieri, 2020)

Kuten aiemmassa luvussa todettiin, datan käyttö muuhun kuin alkuperäiseen tarkoitukseen eli sekundäärinen käyttö on uhka käyttäjien yksityisyydelle (Nunan & Di Domenico, 2013; Herschel & Miori, 2017). GDPR:ssä tätä pyritään ehkäisemään käyttötarkoituksen rajaamisella, jolla varmistetaan että dataa kerätään vain tiettyä laillista tarkoitusta varten, eikä dataa käsitellä myöhemmin tästä tarkoituksesta poikkeavasti (Euroopan parlamentin ja neuvoston asetus (EU)

2016/679, 5. artikla).Basin & Debois (2018) kertovat käyttötarkoituksen rajaamisella tarkoitettavan sitä, että dataa keräävien ja käsittelevien organisaatioiden tulee pyrkiä läpinäkyvyyteen datan käyttötarkoituksen suhteen, ja huolehtia että sitä käytetään nimenomaisesti vain siihen tarkoitukseen, kun käyttäjille on kerrottu.

Tietojen minimoinnilla tarkoitetaan, että kerättävän tiedon tulee olla olennaista ja rajattua. Tietoa tulee kerätä vain sen käyttötarkoituksen palvelemiseen, joka käyttäjille on ilmoitettu (EU) 2016/679, 5. artikla). Pfitzmann & Hansen (2010) katsovat datan minimoinnilla tarkoitettavan kerättävän datan rajaamisen lisäksi myös sitä, että mahdollisuus muita koskevan henkilökohtaisen datan keräämiseen ylipäätään tulisi minimoida. Lisäksi kerätyn henkilökohtaisen datan säilyttämisaika tulee olla pienin mahdollinen (Pfitzmann & Hansen, 2010). Tämä on GDPR:ssä mainittu erillisenä periaatteena, eli säilytyksen rajaamisena. Goldsteen ym. (2021) huomauttavat että kerättävän tiedon rajaaminen ja tarvittavan tiedon määrittäminen etukäteen voi olla haasteellista etenkin edistyneemmissä sovelluskohteissa, kuten vaikkapa koneoppimisen malleissa.

GDPR:n mukaan kerätyn henkilötietoja sisältävän datan tulee olla luottamuksellista ja eheää. Luottamuksellisuudella viitataan rekisterinpitäjän velvollisuuteen huolehtia tietoturvasta, jolla ehkäistään luvaton pääsy henkilötietoihin tai niiden käsittelyyn käytettyyn laitteistoon. Riittävällä tietoturvalla data suojataan luvattomalta tai lainvastaiselta käytöltä, sekä katoamiselta tai vahingoittumiselta (EU) 2016/679, 5. artikla).

Säilytyksen rajaamisella pyritään minimoimaan aika, jolloin henkilötietoja säilötään muodossa, joka mahdollistaa käyttäjän yksilöimisen. GDPR ei seikkaperäisesti määrittele vaadittavaa tiedon anonymisoinnin tasoa, vaan katsoo datajoukon olevan anonyymia, kun käyttäjän yksilöiminen mahdollista vain suurella vaivannäöllä tai epätodennäköisin keinoin (Gruschka ym. 2018). Henkilötietoja voidaan säilöä tätä pidempään vain, jos tietoja kerätään tätä vaativiin käyttötarkoituksiin, kuten tieteelliseen tutkimukseen. (EU) 2016/679, 5. artikla). Henkilötietoja sisältävä data tulee siten poistaa välittömästi, kun sille ei ole enää laillista tarvetta. Yksilöitävissä olevan datan säilyttäminen ilman hyväksyttyä käyttötarkoitusta on kielletty edellä mainittuja poikkeuksia lukuun ottamatta. (Arfelt, Basin & Debois, 2019).

Virheettömyyden periaatteen mukaan henkilötietoja sisältävä data tulee tarpeen mukaan pitää ajantasaisensa ja virheettömänä. Mikäli datan havaitaan olevan virheellistä, virheellinen data tulee poistaa ilman viivettä, mikäli virheellä on vaikutus siinä käytössä johon data on alun perin kerätty (EU) 2016/679, 5. artikla).

3.3.2 GDPR ja massadata

GDPR on saanut osakseen kritiikkiä jo ennen asetuksen valmistumista ja käyttöönottoa, sekä sen jälkeen. GDPR:n katsotaan olevan joiltain osin huonosti yhteensopiva nykypäivän dataympäristön ja erityisesti massadatan kanssa. Zarskyn (2018) mukaan GDPR on epäsopiva nykypäivän dataympäristöön ja sen sisältö on jo valmistuessaan vanhentunutta monilta osin. Tätä perustellaan sillä,

että massadata on muuttanut dataympäristöä merkittävästi, ja GDPR epäonnistuu ottamaan huomioon massadatan ja siihen liittyvät käytännöt. Sisällön vanhentumisen vaihtoehtona on, että massadatan analysoinnin menetelmiä muutetaan GDPR:n mukaisiksi, jolloin on mahdollista, että menetelmistä tulee tehottomampia. (Zarsky, 2018) Pormester (2017) käsittelee artikkelissaan säilytyksen rajaamisen periaatetta, ja huomauttaa että asetus ei määrittele kuinka kauan hyväksyttyä käyttötarkoitusta, kuten tutkimusta varten kerättyä dataa voi säilyttää. Tutkimuksen lavean määrittelyn vuoksi tutkimusta tekevät organisaatiot, kuten lääkeyhtiöt voivat käytännössä säilyttää henkilötietoja niin kauan kuin haluavat tekemäänsä tutkimukseen vedoten (Pormeister, 2017). Myös Gruschka ym. (2018) huomauttavat että GDPR ja massadata on joissain tapauksissa huonosti yhteensopivia. Massadatan analysointi pohjautuu suuriin määrin dataa, ja tämä on ristiriidassa datan minimoinnin periaatteen kanssa. Lisäksi datan analysoinnille on ominaista, että tutkittavat hypoteesit vaihtuvat datan keräämisen jälkeen tutkimuksen edetessä, jolloin datan käyttötarkoitus muuttuu (Gruschka ym. 2018). GDPR:n mukaan dataa saa käyttää yksinomaan siihen käyttöön johon keräyksen kohteena olevat henkilöt ovat antaneet suostumuksensa. Tämän vuoksi datan analysointi tulisi tehdä anonyymilla datalla aina kun mahdollista (Gruschka ym. 2018).

3.4 Dataetiikka

Valmiin datan vaivaton saatavuus, ja varastointiin sekä analysointiin tarkoitettujen työkalujen nopean kehityksen aiheuttama uhka käyttäjien yksityisyydelle vaatii datan käytön tarkastelua myös eettisestä näkökulmasta (Hand, 2018). Datan aiheuttamat yksityisyyden loukkaukset voivat tapahtua tahallisesti tai tahattomasti (Herschel & Miori, 2017). Dataetiikka tietojenkäsittelyn ja informaation etiikkaan pohjautuva etiikan osa-alue, joka tutkii ja arvioi dataan liittyviä moraalisia pulmia datan käytön eri osa-alueilla, kuten esimerkiksi valmistuksessa, tallentamisessa, jakamisessa ja käytössä. Dataetiikkaan sisältyy myös esimerkiksi tekoälyn koneoppimisen kehityksen sekä vallitsevien IT-alan käytäntöjen tutkimista ja arviointia moraalisesta näkökulmasta (Floridi & Taddeo, 2016.) Dataetiikka keskittyy odotetusti ihmisistä kerättyyn dataan (Hand, 2018). Dataetiikka on tieteenalana melko tuore, ja siten itse dataetiikkaan keskittyvä tutkimus on vielä suhteellisen vähäistä. Massadatan etiikka on tutkimusaiheena kattavammin tutkittu, ja tämä näkyy selvästi saatavilla olevan akateemisen kirjallisuuden määrässä. Floridin & Taddeon (2016) mukaan dataetiikan tavoitteena on luoda moraalista tarkastelua kestäviä ratkaisuja perusteltujen käytäntöjen ja arvojen pohjalta. Kuten todettua, dataetiikka pohjautuu informaation ja tietojenkäsittelytieteiden etiikkaan, jotka ovat tutkineet digitaalisiin teknologioihin liittyviä ongelmia 30 vuoden ajan. Kuitenkin dataetiikka tarjoaa niihin verrattuna pidemmälle jalostuneen lähestymistavan, sillä se keskittyy vain informaation sijaan da-

taan. Tämä lähestymistapa tuo eettisen tarkastelun alle myös datan joka ei suoranaisesti jalostu informaatioksi, mutta jota kuitenkin käytetään jonkin toiminnan tukena. (Floridi & Taddeo, 2016)

Datan analysoinnin käytänteiden, tutkimuksen ja etiikan väliset pulmat kuvastaa aiheena hyvin myös laajempia haasteita, joita liittyy dataan, valtaan ja oikeudenmukaisuuteen tai sen puutteeseen (Richterich, 2018). Datan eettisyyteen liittyvien pulmien ratkaisu on muiden teknologien aiheuttamien haasteisiin verrattuna ongelmallisempaa, koska data ja datatiede ovat luonteeltaan ubiikkeja, ja siten nämä ongelmat koskettavat kaikkia datan ja sen käytön parissa toimivia ihmisiä (Floridi & Taddeo, 2016). Handin (2018) mukaan eräs dataetiikan keskeisistä ongelmista liittyy datan monikäyttöisyyteen ja uudelleenkäytettävyyteen. Yksittäiselle datajoukolle on olemassa käytännössä loputtomasti eri sovelluskohteita, ja tulevaisuudessa siitä voidaan saada irti sellaista tietoa mitä nykypäivän menetelmillä ei vielä pystytä saamaan. Tämän vuoksi käyttäjien on mahdotonta tietää miten heitä koskevaa dataa tullaan käyttämään tulevaisuudessa. (Hand. 2018) Myös Herschel & Miori (2017) nostavat käyttäjistä kerätyn datan ja siitä jalostetun informaation uudelleenkäytön eräänä yksityisyydenloukkausten riskiä lisäävänä tekijänä. Uudelleenkäyttö hankaloittaa käyttäjien mahdollisuutta rajoittaa heistä kerätyn datan käyttöä, sekä altistaa käyttäjät ennakoimattomille yksityisyyden loukkauksille. Lisäksi uudelleenkäyttö voi uhata dataa käsittelevän organisaation legimiteettiä (Herschel & Miori, 2017).

3.5 Massadatan etiikka

Dataetiikkaa koskeva tutkimus käsittelee paljolti nimenomaan massadatan etiikkaa. Massadatan tuomat tieto ja valta luovat tarpeen massadatan hyödyntämistä ohjaavien eettisten sääntöjen laatimiselle (Richards & King, 2014). Massadatan etiikka on poikkitieteellinen tutkimusala, joka tutkii massadatan etiikkaa sekä politiikka, sekä sen vaikutusta yksityisyyteen käyttäjien ja organisaatioiden näkökulmasta (Chen & Quan-Haas, 2018). Massadatan etiikan tutkijat pyrkivät löytämään vastauksia ja keinoja siihen, miten ihmisten tulisi elää datan kyllästämässä maailmassa, sekä miten massadatan haitallista hyväksikäyttöä voidaan ehkäistä (Zwitter, 2014). Massadatan etiikan perimmäisin tarkoitus on suojella yksilöiden oikeuksia (Richard & King, 2014). Chen & Quan-Haase (2018) kertovat että massadatan etiikan tutkimuksen on olennaista selvittää ja valistaa ihmisiä massadatan vääristymistä ja rajoituksista, sillä julkinen keskustelu ei nosta näitä asioita pinnalle riittävän kattavasti.

Massadatan mahdollisuuksia saatetaan liioitella erityisesti puhuttaessa sen kyvystä ennustaa tulevaa, ja sen kielteiset aspektit jäävät helposti ennustavuuden varjoon julkisessa päätöksenteossa sekä keskustelussa (Quan-Haase & Chen, 2018). Yksilöillä onkin usein varsin heikko käsitys siitä mitä tietoja heistä kerätään, kuka tietoa käyttää ja miten, sekä mitkä ovat heidän oikeutensa näihin tietoihin liittyen. Richardin ja Kingin (2014) mukaan massadataan sisältyy kaikista arkaluonteisimpia ja paljastavimpia datajoukkoja, kuten käyttäjien sijaintitietoja,

puheluhistorioita, sekä sosiaalisen median verkostoista kertovaa data. Tämä arkaluonteinen data on sekä julkishallinnon että liikemaailman toimijoiden saatavilla ja käytettävissä, ja tämänkaltaisen datan kerääminen yleistyy vuosi vuodelta (Richard & King, 2014). Arkaluonteisuutensa vuoksi datajoukot ovat siten alttiita yksityisyyden loukkauksille. Massadatan käytön, keräämisen ja jakamisen yksityisyyden loukkaukset ovatkin yleisiä, ja niiden kohteeksi ovat joutuneet myös valtaosa suurista teknologiayrityksistä, kuten Facebook, Google ja Apple (Chen & Quan-Haase, 2018). Richardsin ja Kingin (2014) näkemyksen mukaan yksityisyys ja massadata eivät kuitenkaan ole toisensa poissulkevia tekijöitä, vaikka digitalisaation ja datafikaation usein väitetään tappavan yksityisyyden. Massadata kyllä vaarantaa yksityisyyttä, mutta jotta massadataa voidaan hyödyntää optimaalisesti, on olennaista ymmärtää, että informaatiota on mahdollista jakaa luottamuksellisesti ilman yksityisyyden vaarantamista. (Richards & King, 2014). Massadatan eettisten linjausten kehittämisen lisäksi on siis myös tärkeää valistaa ja kehittää yksilöiden näkemystä yksityisyydestä.

Herschelin ja Mioirin (2017) mukaan datan eettinen käyttö edellyttää tietämystä siitä, miten dataa käytetään ja miten datan yksityisyys ja luottamuksellisuus säilytetään. Yksityisyyden ja luottamuksellisuuden säilyttämisen edellytyksenä on tietämys miten yksilöiden tunnistamisen mahdollistavat tiedot poistetaan, sekä kellä on pääsy dataan ja milloin he pääsevät siihen käsiksi (Herschel & Moiri, 2017) Zwitter (2014) puolestaan mainitsee yksityisyyden ohella tärkeänä tutkimuksen aiheena ei-henkilökohtaisen dataan ja siihen liittyvien toimien eettisen tarkastelun. Tätä massadatan eettistä aspektia kutsutaan ryhmäyksityisyydeksi (*eng. Group Privacy)*. Ryhmäyksityisyyden eettiset kysymykset liittyvät tietoihin, jotka määrittävät kuulumisemme johonkin ryhmään, kuten ikä ja sukupuoli. Esimerkkinä tämänkaltaisesta datasta ja sen käyttötapauksesta käy Twitter-feedin sentimenttianalyysi poliittisten tavoitteiden edistämisen apuvälineenä. Ryhmittelevää dataa voidaan käyttää ihmisten ohjaamiseen haluttuun suuntaan. Massadata voi paljastaa käyttäytymisen korrelaatioita, jotka ovat olleet aiemmin tunnistamattomia, ja siten mahdollistaa käyttäytymisen ohjaamisen uusilla tavoilla. (Zwitter, 2014) Zwitter (2020) mainitsee ihmisryhmän tai yksilön alttiuden ja tähän liittyvän ennakoivan politiikan erääksi olennaiseksi massadatan eettiseksi kysymykseksi. Massadataa voidaan valjastaa ennustamaan mitä ihmiset todennäköisesti tekevät, ja tämä voi johtaa yksilön tai ihmisryhmän oikeuksien loukkauksiin, kuten ennakoituun vangitsemiseen ilman että he ovat tehnyt rangaistavaa rikosta. (Zwitter, 2020) Richards & King (2014) mainitsevat hyödyntämisen läpinäkyvyyden olevan avainasemassa. Datan hyödyntämisen tulee olla avointa ja läpinäkyvää niin julkishallinnollisten kuin yksityisten organisaatioiden tapauksessa. Läpinäkyvyys ehkäisee massadatan väärinkäyttöä sekä edesauttaa organisaatioiden ja yksilöiden välisen luottamuksen syntymistä. Luottamus taas vaikuttaa yksilöiden halukkuuteen jakaa itsestään olennaista dataa, joka hyödyttää päätöksentekoa. (Richards & King, 2014)

4 TUTKIMUSMENETELMÄT

Tämä luku keskittyy tutkimuksen taustoitukseen ja toteutuksen kuvailemiseen. Luvun alussa esitellään valittu tutkimusmenetelmä ja perustellaan tutkimusmenetelmän valintaa. Tutkimusmenetelmän esittelyn jälkeen kuvataan tutkimuksen taustaa, sekä kerrotaan sen järjestelyistä sekä toteutuksesta. Tämä jälkeen kuvataan, kuinka aineisto on analysoitu ja mitä menetelmiä analysoinnissa on käytetty. Tutkimuksen luotettavuutta ja laatua arvioidaan luvussa 7.

Aihepiirin aikaisempaan kirjallisuuteen tutustuttiin kirjallisuuskatsauksen avulla. Kirjallisuuskatsauksen keskeiset käsitteet olivat massadata sekä informaation yksityisyys. Katsauksessa käsitelty kirjallisuus on pääosin akateemista kirjallisuutta, jota haettiin pääosin Google Scholarista, sekä Jyväskylän yliopiston Jykdok-tietokannasta. Käytettyjä hakusanoja oli muun muassa ”big data” ”information privacy” ”insurance industry”, sekä näiden yhdistelmät. Valittavien lähdemateriaalien valinnassa kiinnitettiin huomiota julkaisuajankohtaan, sitaattien määrään sekä julkaisijaan.

Tutkielman tutkimusmenetelmäksi valikoitui kvalitatiivinen, eli laadullinen menetelmä, ja aineistonkeruumenetelmäksi teemahaastattelu. Laadullisella tutkimusmenetelmällä tarkoitetaan aineiston ja analyysin ei-numeraalista kuvausta. (Eskola & Suoranta, 1998). Tutkimusmenetelmäksi valikoitui kvalitatiivinen menetelmä, koska tarkoituksena oli kartoittaa aihepiirin kanssa läheisesti tekemisissä olevien henkilöiden mielipiteitä ja näkemyksiä, jotka ovat mittaamattomia muuttujia. Tutkimuksessa käytettäviä haastattelumenetelmiä jaotellaan usein sen mukaan, miten ennalta määrätty niiden rakenne on. Teemahaastattelu sijoittuu tässä suhteessa tarkasti ennalta määrätyn ja avoimen haastattelun välimaastoon. Aineistonkeruumenetelmänä teemahaastattelu on joustava, sillä se mahdollistaa kysymysten esittämisen laajemmalta alueelta kuin tiukasti kysymyspohjaan sidotut strukturoidut haastattelut (Hirsjärvi & Hurme, 2000). Kvalitatiivinen menetelmä sopii parhaiten mittaamattomista muuttujista koskevan tiedon keräämiseen, ja valittiin siksi tutkielman tutkimusmenetelmäksi. Teemahaastattelun tavoitteena

oli kerätä vakuutusalalla toimivien data-analytiikan ja massadatan asiantuntijoiden näkemyksiä massadatan hyödyntämisestä, mahdollisuuksista sekä haasteista vakuutusalalla. Lisäksi haastattelulla pyrittiin selvittämään informaation yksityisyyden sääntelyn vaikutuksia data-analytiikkaan vakuutusalalla. Siten haastattelujen tavoitteena oli kartoittaa haastateltavien ammattilaisten näkemyksiä tutkielman alussa määriteltyihin tutkimuskysymyksiin. Teemahaastattelussa haastateltiin vaihtelevan taustan omaavia aihepiirin asiantuntijoita eri organisaatioista. Tutkimuskysymykset olivat:

- Miten vakuutusyhtiöt voivat hyödyntää massadataa?
- Mitä ovat vakuutusyhtiöiden olennaisimmat haasteet massadatan hyödyntämisessä?
- Miten informaation yksityisyyteen liittyvä sääntely vaikuttaa datan hyödyntämiseen vakuutusyhtiöissä?

4.1 Aineistonkeruu

Tutkielman aihepiiri liittyy data-analytiikkaan, massadataan sekä informaation yksityisyyden sääntelyn vaikutuksiin datan hyödyntämisessä vakuutusalalla. Tutkimuksessa haastatellut henkilöt ovat data-analytiikkaa hyvin tuntevia vakuutusalan asiantuntijoita. Kerättävän aineiston laadun varmistamiseksi tutkielmaan pyrittiin löytämään haastateltavia, jotka täyttävät seuraavat kriteerit.:

1. Haastateltavat edustavat useaa organisaatiohierarkian tasoa
2. Haastateltavat edustavat useaa eri organisaatiota
3. Haastateltavilla tulee olla vankkaa kokemusta ja tietämystä vakuutusalasta
4. Haastateltavien tulee omata kokemusta ja tietämystä data-analytiikan ja massadatan hyödyntämisestä vakuutusalalla

Ensimmäinen ja toinen kriteeri varmistettiin haastattelijan toimesta jo haastateltavia valikoitaessa. Kolmas ja neljäs kriteeri varmistettiin kertomalla haastattelukutsussa lyhyesti tehtävästä tutkimuksesta, ja etsittäviltä haastateltavilta edellytettävän asiantuntemuksen aihepiiristä.

Teemahaastatteluun valittavien haastateltavien määrän arviointi voi olla haastavaa, sillä suositukset vaihtelevat eri lähteiden välillä. Masonin (2010) mukaan laadullisessa tutkimuksessa datan määrä ei välttämättä korreloi informaation määrän kanssa. Laadulliset tutkimukset keskittyvät merkitysten löytämiseen, ja siksi esimerkiksi esiintymistiheydet ovat usein toisarvoisia laadullisessa tutkimuksessa. Lisäksi laadulliset menetelmät ovat tutkimuksen teon kannalta työläitä, ja siksi suuren otannan käyttö on usein epäkäytännöllistä.

(Mason, 2010) Guest, Bruce & Johnson (2006) totesivat, että perustavanlaatuisimmat teemat voidaan löytää jo kuuden haastattelun tuloksena. Myös Francis ym. (2010) toteavat, että merkittävimmät teemat voidaan löytää jo 5-6 haastattelun tuloksena. Tutkimukseen valikoidut haastateltavat olivat pitkän linjan ammattilaisia, joilla oli kokemus lukuisista eri vakuutusalan tehtävistä data-analytiikan parissa. Haastattelukutsun saaneet henkilöt valikoitiin heidän ammatillisen taustan, työnimikkeen sekä saatujen suosittelujen perusteella. Haastattelukutsussa esiteltiin tutkimuksen aihepiiri ja tutkimuskysymykset. Henkilöitä, jotka kokivat olevansa epäpäteviä vastaamaan aihepiirin kysymyksiin ei valittu haastateltaviksi. Lopulta haastateltavaksi valikoitui viisi eri organisaatioissa työskentelevää vaihtelevan taustan omaavaa aihepiirin asiantuntijaa, joiden tutkimuksen kannalta olennaiset tiedot on esitelty taulukossa 1.

Taulukko 1 Laadulliseen tutkimukseen osallistuneiden haastateltavien yhteenveto

Organisaatiohierarkian taso	**Työkokemus vakuutusalalta (Vuosia)**	**Työkokemus data-analytiikan parista (Vuosia)**
Johtaja	18	18
Johtaja	>30	>30
Keskijohto	25	36
Asiantuntija	10	13
Asiantuntija	6	13

Haastattelut tallennettiin tietokoneelle myöhempää tekstiksi muuttamista, eli litterointia varten. Haastattelut toteutettiin yksilöhaastatteluina, eli tilasuuteen osallistuivat vain haastattelija ja haastateltava. Haastattelussa käytettävä haastattelurungon teemat laadittiin tutkimuskysymysten ja teoriaosuudessa tehtyjen löydösten perusteella (Liite 1). Haastattelurunkoa laatiessa haluttiin varmistua, että kysymällä nämä kysymykset saadaan vastaukset tutkielman tutkimuskysymyksiin. Teemahaastattelulle ominaiseen tyyliin kysymysrungosta oli mahdollista poiketa niin kysyttävien kysymysten, kuin myös kysymysten järjestyksen osalta. Siten kaikilta haastateltavilta ei kysytty samoja kysymyksiä. Haastattelurunko on liitetty tutkielman loppuun. Haastattelut kestivät keskimäärin 45 minuuttia.

4.2 Aineiston analysointi

Kerätty aineisto analysoitiin temaattisen analyysin keinoin. Temaattinen analyysi on systemaattinen metodi, joka mahdollistaa jaettujen kokemusten ja tarkoitusten havaitsemisen datajoukosta (Braun & Clarke, 2012). Braun & Clarke (2012) jakavat temaattisen analyysin prosessin kuuteen vaiheeseen, joita

käytetään aineiston analysoimiseen myös tässä tutkielmassa. Analyysin vaiheet ovat:

1. Dataan tutustuminen
2. Alustavien luokittelujen generointi
3. Teemojen etsintä
4. Potentiaalisten teemojen arviointi
5. Teemojen määrittely ja nimeäminen
6. Raportin tuottaminen (Braun & Clarke, 2012)

Haastatteluista kerätyn aineiston analysointi aloitettiin muuttamalla nauhoitetut haastattelut tekstimuotoon, eli litteroimalla nauhoitteet. Jokainen haastattelunauhoite muutettiin yksittäiseksi tekstitiedostoksi Google Docs-palveluun, johon käyty keskustelu kysymyksineen ja vastauksineen litteroitiin. Hirsjärven ja Hurmeen (2000) mukaan ei ole olemassa yleistä ohjeistusta litteroinnin tarkkuuden määrittämiseen, vaan tämä riippuu tutkimuksesta. Hirsjärven ja Hurmeen (2000) mukaan huolellinen aineistoon tutustuminen on edellytys onnistuneelle analyysille. Kerättyyn aineistoon tutustuminen aloitettiin lukemalla litteroitu teksti useaan otteeseen. Aineistoa lukiessa haastatteluista pyrittiin tunnistamaan jo alustavasti tärkeimpiä lauseita erityisesti tutkimuskysymyksiin vastaamisen kannalta. Havaittuja huomioita, kuten toistuvuuksia, yhtäläisyyksiä tai ristiriitoja aiempaan tutkimukseen luokiteltiin alustavasti eli koodattiin, ja merkittiin kommentein Google Docs -dokumenttiin. Alustavat luokittelut tehtiin tarkasti havainnollistettavaa asiaa tai ilmiötä kuvaaviksi, esimerkiksi haastatteluissa esiin nousseita massadatan käyttökohteita vakuutusalalla luokiteltiin yksittäisiin käyttökohteisiin, kuten ”Aktuaaritoiminta” tai ”Vakuutustuotteiden personointi”. Tämän jälkeen luokitteluista etsittiin ja luokittelut jaettiin laajempiin teemoihin. Teemojen luonnissa hyödynnettiin tutkimuskysymyksiä, sekä haastattelurungon aihepiirejä. Esimerkiksi edellä mainitut luokittelut asetettiin teeman ”Massadatan vaikutukset ja potentiaali” alle. Luokittelut ”Kolmansien osapuolten rajoitukset” ja ”Sääntely” taas yhdistettiin teeman ”Datan saatavuus ja saavutettavuus” alle. Erittelyn jälkeen teemat jaoteltiin kolmeen eri tekstitiedostoon tutkimuskysymyksittäin. Lopuksi, kaikki haastattelumateriaaleista poimitut otteet jaoteltiin teemoittain. Haastatteluista nousi esiin myös teemoja, jotka eivät nousseet esille aiempaa tutkimusta tarkastellessa, kuten vakuutusyhtiöiden korostunut tarve toiminnan läpinäkyvyydelle alaa vaivaavan mainehaitan vuoksi. Seuraava luku on omistettu tutkimustulosten kuvaamiselle.

5 TUTKIMUKSEN TULOKSET

Tässä luvussa kuvataan haastattelututkimuksesta saatuja tuloksia. Tulosten esittely on jaoteltu tutkimuskysymyksittäin teemoihin, jotka ovat massadata vakuutusalalla, massadatan hyödyntämisen haasteet vakuutusalalla, sekä informaation yksityisyyden sääntelyn vaikutukset vakuutusliiketoimintaan.

5.1 Massadata vakuutusalalla

Jo ennen digitalisaatiota vakuutusalan erityispiirteenä on ollut datan merkittävä rooli liiketoiminnan keskiössä, sillä dataa on tuotettu ja hyödynnetty yhtiöiden toimesta kauan. Siten vakuutusyhtiöillä on paljon kokemusta datan käsittelystä ja käsittelyn haasteista, sekä huomioonotettavista seikoista kuten sensitiivisen datan käsittelyn tietosuojasta huolehtimisesta. Haastateltavat olivat yhtä mieltä siitä, että data on kasvamassa yhä merkittävämmäksi tekijäksi vakuutusalalla, ja tunnistettuja käyttötapauksia massadatan hyödyntämiselle on runsaasti. Nykyisistä käyttötapauksista kysyttäessä aktuaaritoiminta, eli vakuutusmatematiikka nousi esiin ensimmäisenä lähes jokaisen haastateltavan kohdalla. Muita toistuvasti esiin tulleita nykypäivän käyttötapauksia oli asiakashallinta, asiakaspalvelu, asiakasanalytiikka, vahingontorjunta myynti ja markkinointi, riskinvalinta, sekä petosanalytiikka. Vakuutusyhtiöt ovat keränneet asiakkaistaan dataa jo kauan, mutta kerätty data on ollut luonteeltaan karkeaa ja niukkaa. Niukkuuden lisäksi datan hyödyntämistä on aikaisemmin estänyt tehokkaiden työkalujen puute. Yksi haastateltavista kuvasi datan volyymin ja monikanavaisuuden kasvun lisäämää ennusteen tarkkuutta näin:

"Se viittas siihen aikaan jossa data oli kallista, sitä ei ollut ja meillä ei ollut tietojenkäsittelykapasiteettia, eli vaikka dataa ois ollukin, niin ei ollu kapasiteettia käsitellä sitä. Periaatteessa näiden yksinkertaisten kun iän ja sukupuolen ohella jos meillä on suunnilleen henkilötason tietona, että mitä henkilö on viimesen kymmenen vuoden aikana ostanu kau-

pasta, supermarketista, kuinka usein hän käy kuntosalilla, mitä hän on ehkä ostanu apteekista viime vuosina, niin mä melkein ajattelisin että tota on tuotavissa paljon paremmin ennuste henkilön riskillisyydestä kuin että jos on ikä ja sukupuoli."

Vakuutusala on kokenut muiden alojen ohella suuria muutoksia teknologian kehityksen seurauksena. Datan, järjestelmien sekä palvelukanavien määrän kasvu on tehnyt vakuutusyhtiöiden toiminnasta aiempaa kompleksisempaa, ja prosessien automatisoiminen on vaikeampaa. Vakuutusala on myös hyvin säänneltyä, ja vakiintuneet toimijat ovat suuria yhtiöitä mikä aiheuttaa jäykkyyttä ja hidasta reagointikykyä ympäristön muutoksiin. Voidaan todeta, että riskit ja haasteet ovat kasvaneet mahdollisuuksien kasvun rinnalla. Kysyttäessä massadatan tähän mennessä aiheuttamista vaikutuksista eräs haastateltava vastasi:

"No ei nyt oo mitään hirveetä muutosta tapahtunu, että no varmaan tuolla verkkopuolella pystytään nyt enemmän tekemään kun ennen, mut ei mitää suurta mullistusta, se on lopulta kuitenki sitten se niin että tää toimintaympäristö on muuttunu aika paljon monimutkaisemmaks, et ollaan selkeesti monikanavaisempia, paljon enemmän dataa, paljon enemmän järjestelmiä, tavallaan se et me pystytään automatisoimaan niitä, ni se on muuttunu vaikeemmaks."

Vakuutusyhtiöt omaavat kokemusta analytiikan ja ennusteiden laatimisen haasteellisuudesta jo pitkältä ajalta. Eräs haastateltava luonnehti alan kokeneisuutta datan hyödyntämisessä seuraavasti:

"[...] joissakin puheenvuoroissa olen sanonut, että aktuaarit tai vakuutusmatemaatikot kannattaa ottaa mukaan tähän, uuteen datan maailmaan ja mallintamiseen sen takia että me ollaan tehty kaikki virheet jo aikaisemmin.

5.1.1 Datan hyödyntäminen ja vaikutukset vakuutusalalla

Kuten aikaisemmin on todettu, vakuutusyhtiöillä on ollut jo kauan hallussaan suuria määriä dataa. Analogista dataa on kerätty koko vakuuttamisen käsitteen olemassaolon ajan, mutta datan muuttaminen digitaaliseen muotoon eli digitointi, sekä digitoidun datan jalostaminen informatiiviseen muotoon, eli datafikaatio on mahdollistanut datan tehokkaan hyödyntämisen. Yksi haastateltavista totesi:

"Mut periaattesssahan sen voi sanoa että se data joka meistä syntyy, tai dataahan on ollu ennen, mutta digitaalisessa muodossa oleva data on se, et se data on sellasessa muodossa että sitä voi käsitellä. Niin niin sehän nyt se varsinainen juttu on."

Toinen haastateltava kuvasi vakuutusyhtiöiden datan käsittelyn historiaa ja hyödyntämisen tärkeyttä yhtiöiden päätöksenteossa seuraavasti:

"Mehän ollaan 150 vuotta hilluttu tuolla sinikantisista vihoista alkaen, ja erinäkösiin muihin tallennusmedioihin viety sitä tietoo, ja sitä pitäisi vaan vielä enemmän hyödyntää päätöksenteossa"

Vakuutusyhtiöiden hyödyntämä data on ollut perinteisesti pääasiallisesti sisäsyntyistä, eli heidän omista asiakkaistaan kerättyä dataa. Ulkoisia datanlähteitä on alettu kuluneiden vuosien aikana hyödyntämään yhä enemmän, kun datan tuomat mahdollisuudet on ymmärretty ja data on otettu olennaiseksi osaksi yhtiöiden strategiaa. Ulkoiset toimijat joilta dataa ostetaan ovat pääasiassa julkishallinnollisia toimijoita, kuten Tilastokeskus. Vakuutusyhtiöiden omia datan lähteitä on pääasiassa heidän omat järjestelmät, kuten asiakashallintajärjestelmät (*eng. CRM, Customer Relationship Management),* sekä verkkoon sulautetut evästeiden kautta toimivat asiakasanalytiikkadataa keräävät järjestelmät joita käytetään esimerkiksi verkon personointiin. Nämä järjestelmät mahdollistavat datan keräämisen asiakkaan verkkokäyttäytymisestä ja myös heidän vieraillessaan muilla kuin kyseisen yhtiön sivuilla. Kerättyä tietoa voidaan hyödyntää myynnissä ja markkinoinnissa niin sisäisissä kuin ulkoisissa kanavissa, sekä raportoinnissa. Eräs useasti esiin tullut käyttötapaus on asiakashallintajärjestelmien tekemät toimintasuositukset, joka mainittiin usean haastateltavan toimesta. Yksi haastateltavista kertoi seuraavasti:

"[…] siinä on sitte monikanavamarkkinoinnissahan se on valtavassa roolissa että voidaan asiakkaan tietoja yhdistämällä nähdä mitä se on tehny, ja sen perusteella laukasta toimenpiteitä vaikka kampanjahallinnassa tai saadaan CRM:ään toimenpidesuosituksia. Että sä oot käyny tällä sivulla ja sulla ei oo kotivakuutusta, niin sä oot ehkä potentiaalinen kotivakuutusasiakas. Voidaan tehdä sellasia nostoja sinne […]"

Kerättyä asiakasanalytiikkadataa voidaan käyttää myös asiakkaalle näkyvien verkkosivujen personointiin:

"[…] ja sitte verkkosivuja voidaan myös sillä personoida tietyllä analytiikan segmentillä, että jos sä putoot tiettyyn segmenttiin niin sille segmentille voidaan näyttää erityyppistä sisältöö. Muumuassa etusivulla meillä on siellä suosittelualgoritmi, joka on tehty sillain, että kun sä oot käyny tuotesivulla, niin se viimesimmän tuotesivun mukanen nosto tehään siihen etusivulle. Se personoituu sun käyntihistorian perusteella [..]"

Massadatan hankkiminen ulkoisilta toimijoilta on vähäistä, sillä aiemmin datan hyötyjä ei ole nähty niin merkittäviksi, että dataan olisi haluttu investoida. Datan korkean hinnan aiheuttamat resurssiongelmat mainittiin usean haastateltavan toimesta. Kysyttäessä ulkoisten palveluiden ostamisesta, yksi haastateltavista kertoi:

"Siis, osa massasta ostetaan ja ehkä enenevissä määrin. Me ei olla vielä viety kaikkee tuotantoon mitä nyt ollaan tässä rakentamassa, et ehkä lähdetään enemmän sieltä että asiakasymmärryksen kautta tullaan ulos ja lähdetään ihan sieltä massaa pyörittää läpi."

Massadatan ominaisuuksiin lukeutuu sen suuri volyymi, joka tarkoittaa, että dataa syntyy suurella nopeudella, ja usein arvon ulosmittaaminen edellyttää reaaliaikaista analysointia. Haastateltavilta kysyttiin tiedossa olevia käyttötapauksia massadatan reaaliaikaiselle analysoinnille, ja kaikki haastateltavat mainitsivat petosanalytiikan yhtenä käyttötapauksena reaaliaikaiselle data-analytiikalle. Lisäksi yhtiöillä on ollut kokeiluja tai suunnitelmia mm. putkistojen vuotoja havaitsevien sensorien hyödyntämisestä vahingontorjunnassa, sekä säädataan ja asiakkaiden tiedottamiseen perustuvasta vahinkojen ennaltaehkäisemisessä. Muutama haastateltava mainitsi myös puheluiden reitittämisen asiakkaiden ominaisuuksien perusteella yhdeksi reaaliaikaisen analysoinnin käyttökohteeksi. Haastattelujen perusteella vakuutusyhtiöt eivät siten juurikaan hyödynnä reaaliaikaista datan analysointia tällä hetkellä. Teknologian, kuten sensorien ja tekoälyn kehittyessä tulevaisuuden käyttötapauksia kuitenkin nähtiin runsaasti.

Teoriaosiossa tutkitun perusteella sosiaalisen median massadataa pidetään yleisesti merkittävänä massadatan lähteenä, mutta haastattelujen perusteella vakuutusyhtiöt hyödyntävät sosiaalisen median massadataa niukasti. Sosiaalinen media nähdään ennemminkin markkinointikanavana kuin hyödyllistä tietoa tarjoavana datan lähteenä. Haasteltavilta kysyttiin miten heidän organisaatiota hyödyntävät sosiaalisen median massadata, ja eräs vastasi:

Joo, tällä hetkellä hyvin vähän. […] Kun meillä on tässä rakenteilla sellanen uus maailma missä yhdistetään useita sosiaalisen median massadatapalveluita ja voidaan yhdistää sitä asiakkaaseen.[.]Ja meille tulee sinne oiskohan 7 tai 8 eri lähdettä mistä me pyöritetään sitä. ”

[.]Meilläki se on nyt mitä (sosiaalista median datan hyödyntämistä) tässä ollaan aikasemmin kokeiltu, niin ne työkalut on ollu hyvin perustyökaluja miten sitä ollaan hyödynnetty, tai yritetty höydyntää, sanotaanko näin päin.

Haastateltavia pyydettiin arvioimaan sosiaalisen median massadatan merkittävyyttä. Sosiaalinen media nähdään näkyvyyttä lisäävänä markkinointialustana, mutta sosiaalisen median yhtiöiden omistamaa dataa ei pidetty merkittävänä vakuutusyhtiöiden liiketoiminnan kannalta, vaan sosiaalinen media nähtiin lähes yksinomaan näkyvyyttä lisäävänä markkinointialustana, eikä merkittävän datan lähteenä. Sosiaalinen media nähtiin ongelmallisena ja epäsopivana vakuutusyhtiöiden käyttöön. Eräältä haastattelevalta kysyttiin, että näkeekö hän sosiaalisen median datassa piilevän sellaista arvoa, jota vakuutusyhtiöt voisivat hyödyntää jos pääsisivät siihen käsiksi. Haastateltava vastasi:

”Yleensä kun mainostetaan sosiaalisessa mediassa, niin on helpompi mainostaa jotain tuotetta mikä on käsinkosketeltavaa [.], mutta vakuutus on semmonen pitkän harkinnan tuote, että mä en tiä saatasko me somesta semmosta dataa mikä välttämättä meitä niinkään hyödyntäis just se meidän tuotteen ominaisuuden takia, et oisko siinä sit semmosta hyötyä.”

Kuten todettua, vakuutusyhtiöt ovat enimmiltä osin kiinni vielä perinteisemmissä datan käsittelyn menetelmissä, ja pääosa käsiteltävästä datasta on tekstimuotoista. Haastatteluista ei käynyt ilmi, että yhtiöt hyödyntäisivät esimerkiksi ääni- tai videomuotoista dataa. Yhtiöt hyödyntävät myös transaktiodataa, jota esimerkiksi asiakashallintajärjestelmät tuottavat aina kun siellä suoritetaan jokin toiminto ihmisen toimesta tai automatisoidusti. Transaktiodata mainittiin esimerkiksi tässä vastauksessa:

"[...] niin itseasiassa siitä vakuutusyhtiön prosesseistakin syntyy sitä digitaalista dataa mitä voidaan analysoida. Et jos meillä aikasemmin jossain vakuutuksen myöntämisessä, korvauskäsittelyssä ihminen teki juttuja, niin siinä ei kaikki aspektit suinkaan tullut dokumentoiduks. Sit jos meillä on tekoälysovellus niin siitähän jää digijälki, mitä se on päättänyt."

5.1.2 Massadatan potentiaali

Dataan investoimiseen ei ole menneisyydessä suhtauduttu vakuutusyhtiöissä myönteisesti, mutta halut investointeihin ovat kasvaneet datan potentiaalin ymmärtämisen myötä. Data onkin otettu keskeiseksi osaksi yhtiöiden strategiaa. Tiedolla johtamisen tärkeyttä ja potentiaalia kuvattiin erään haastateltavan toimesta seuraavasti:

[…] siellä on vielä paljon tehtävää, mutta ollaan mun mielestä oikeella polulla siinä. Sit ollaan myös ymmärretty se tiedon merkitys et se tiedolla johtamisen, tiedon arvo on ymmärretty ja mitä sillä voidaan saada aikaseks on ymmärretty, ja asioita tehdään, asiat perustuu tietosiin päätöksiin. Näen että polku on varmasti oikea ja, ollaan menossa oikeeseen suuntaan siinä suhteessa. "

Haastateltavien näkemykset massadatan hyödyntämisen nykytilasta ja tulevaisuuden potentiaalista olivat yhteneväisiä. Kaikilta haastateltavilta kysyttiin, onko massadataan kohdistuvat suuret odotukset vielä realisoituneet vakuutusalalla. Kaikkien haastateltujen mielestä vakuutusala ei ole vielä kokenut suuria teknologian aiheuttamia mullistuksia, mutta uskoivat että mullistuksia tulee tapahtumaan tulevaisuudessa. Muutoksiin johtaa sekä teknologian että ympäristön muutokset. Eräs haastateltava kuvasi asiaa näin:

"Ei vielä. Ei missään nimessä, et kyllä me perinteisessä datamaailmassa ehkä suurimmilta osin välineistä pyöritään. Et ehkä yks juttu ollu että niiden tehokkaiden työkalujen käyttö on ollu aika vaatimatonta, et mitä sieltä koittaa löytää tuolta massasta ulos. Ja ehkä enemmän ollaan tukeuduttu tohon omissa tietokannoissa oleviin datoihin ja luotu sieltä niitä ennusteita."

Puhuttaessa massadatan tulevaisuudennäkymistä ja sen potentiaalista vakuutusalaa muuttavana tekijänä, kaikki haastateltavat olivat siinä uskossa, että data todennäköisesti aiheuttaa suuria muutoksia alalle. Massadatan tuomia liiketoiminnallisia hyötyjä pidettiin myös keskeisenä kilpailutekijänä. Haastattelussa kysyttiin miten tärkeänä kilpailutekijänä haastateltavat kokevat datan. Eräs haastateltava vastasi:

"No mun mielestä se on ihan elintärkeetä että me tunnetaan meidän asiakkaat, meidän pitäis tuntee meidän asiakkaat ja meidän pitäs tuntee se data, sen pitäs olla ehdottomasti oikeellista ja siihen pitäis kaikkien päätösten ja tekojen tavallaan perustua, siihen tiedolla johtamiseen. Et mun mielestä se on ihan vitaali tieto se data, et ilman sitä on hankala ehkä pysyä kilpailussa mukana. Et mun mielestä se on ihan elintärkee."

Massadata mahdollistaa paremman ymmärryksen asiakkaistaan ja vakuutettavista riskeistä, joka lisää liiketoiminnan kannattavuutta. Massadatan tulevaisuuden potentiaali pohjautuu paljolti teknologian kehitykseen, erityisesti koneoppimisen ja tekoälyn hyödyntämiseen. Näiden teknologioiden hyödyntäminen yhdessä suurten datamäärien kanssa nähtiin merkittäväksi alan suunnanohjaajaksi tulevaisuudessa. Joidenkin vastausten mukaan on esitetty, että vakuutusten tarve voi jopa kokonaan hävitä datan määrän lisääntyessä. Yksi haastateltavista vastasi seuraavasti kysyttäessä massadatan tulevaisuudennäkymistä vakuutusalalla:

"Siis mä olen ja elän siinä uskomuksessa, että mitä enemmän vakuutusyhtiöllä on dataa käytettävissä, niin sen paremmin ja tehokkaammin vakuutus toimii. Että voisin ajatella, että massadatan käyttö auttaa profiloimaan ihmisiä ja riskityyppejä, ja sitten se voi olla lähinnä tekoälysovellus, joka ikäänkuin tekee matchin yksittäisen vakuutushakemuksen tai vakuutuksen hakijan ja sen massadatan tyypittelyn välillä, ja pystyy siitä sitten tekemään hyvän ennusteen, että mikä on tämänkin henkilön riskillisyys."

Kattava yksilöllisen datan keruu mahdollistaa vakuutustuotteiden individualisoinnin niin sisällöltään kuin hinnaltaan. Tämänkaltaisten ratkaisujen etuihin lukeutuu hinnan ja riskin kohtaaminen, jolloin asiakkaat maksavat oikeansuuruista vakuutusmaksua suhteessa toisiinsa. Vakuutusyhtiöt pystyvät datan avulla toteuttamaan tehokasta ja tarkempaa riskienhallintaa, ja voivat halutessaan valita asiakkaikseen vain matalan riskin asiakkaita. Tuotteiden dataan pohjautuvaa yksilöintiä ei ole tehty vielä suomalaisissa yhtiöissä, sillä haastateltavat eivät tienneet, että yhtäkään tämänkaltaista tuotetta olisi tuotannossa. Suomalaiset vakuutusyhtiöt ovat kuitenkin pilotoineet hankkeita, joissa käytetään autoissa olevaa telematiikkaa tai esimerkiksi terveysteknologian keräämää dataa. Ulkomailla on kuitenkin jo olemassa niin sanottuja pistevakuutusyhtiöitä, jotka ovat dataan pohjautuvilla innovatiivisilla tuote- ja palveluratkaisuilla kyenneet haastamaan myös vakiintuneet toimijat. Pistevakuutusyhtiöille on tyypillistä tavallista suppeampi tuotevalikoima, ja tarkat asiakasvalinnan kriteerit. Haastatteluissa asia ilmeni seuraavasti:

"Mut sit tuolla jenkeissä ja EU-alueella on semmosia pieniä ns. pistevakuutusyhtiöitä, jotka tekevät jonkun asian, analysoivat pirun tarkkaan ketä haluavat asiakkaiksi ja tekee älyttömän kannattavaa liiketoimintaa."

"Joo, erittäin suppee tuotevalikoima ja ei kaikkea tuotteita missään nimessä, vain jotain tiettyä riskiä vastaan, ja ne tekee sen pirun hyvin et valikoi tosi tarkkaan riskin et ei ees halua tarjota kaikille, ja sit kaikki mitä ne tekee markkinointponnisteluita ja muita niin se kohdejoukko rajataan niin et siellä ei tuu niitä ei-toivottuja riskejä. "

Prosesseja automatisoimalla vakuutusyhtiöt voivat saavuttaa korkeamman tehokkuuden ja paremman kulusuhteen. Reaaliaikaisen analysoinnin avulla vakuutusyhtiöt kykenevät ennakoimaan vahinkoja, ja ennaltaehkäisemään niitä asiakkaita tiedottamalla. Analysoitava data voi olla peräisin esimerkiksi sensoreista, kuten vuotovahdeista, tai ulkoisilta toimittajilta ostettua säädataa. Yhtä, vielä suunnittelun tasolla olevaa käyttötapausta kommentoitiin haastatteluissa seuraavasti:

"Sensoreista on ollu myös vihje siihen, että esimerkiks tilanteessa josssa autonvalmistajat laittas sensorit niihin autoihin, ja me saatas vaikka kiveniskemistä heti signaali ja se autojärjestelmä lähettäs sen signaalin ihan keskitettyyn tietovarastoon. mitä vois sitten vakuutusyhtiöt ostaa ja hyödyntää. Niin me nähtäs että keväällä maaliskuun 7. päivä lähti nää signaalit lisääntymään, ja me voitas sit asiakasta tiedottaa että muistakaa se varoväli kun ajatte motarilla, tai jos on nyt liian myöhästä niin meidän kumppanikorjaamot auttaa teitä et siellä ilman ajanvarausta että ei tarvii kun mennä sisään. Et tän tyylistä ollaan väläytelty et vois olla hienoja keissejä mitä voitas dataa hyödyntää jatkossa tulevaisuudessa ehkä paremminkin."

5.2 Massadatan hyödyntämisen haasteet

Massadatan potentiaali vakuutusalalla on hyvin suuri, mutta mahdollisuuksien rinnalla myös hyödyntämisen haasteet ovat moninaisia ja merkittäviä. Hyödyntäminen on kokonaisuutena monimutkainen, ja sen onnistumiseen tai epäonnistumiseen vaikuttavien tekijöiden joukko on kirjava. Keskeisiä haastatteluissa esiin nousseita hyödyntämisen haasteita aiheuttavia teemoja ovat resurssit, datan ominaisuudet, sekä datan saatavuus. Datan saatavuuteen vaikuttaa olennaisesti myös eri tasoinen informaation yksityisyyden sääntely, mutta siihen keskitytään syvemmin luvussa 7.3.

5.2.1 Resurssit

Vajavaiset resurssit mainittiin massadatan hyödyntämisen haasteina useaan otteeseen haastateltavien toimesta. Panostukset dataan ovat usein kalliita, ja siksi dataan ja tiedolla johtamiseen liittyviin investointeihin on perinteisesti suhtau-

duttu epäsuotuisasti. Dataan käytettävät resurssit on nähty ennemminkin pakollisena kulueränä kuin hyödyllisenä investointina. Dataan kohdistettujen resurssien puute on siten aiheutunut organisaatioiden päätöksentekijöiden skeptisistä asenteista datan mahdollisuuksiin. Käyttöön saatavien resurssien puutteeseen voi myös johtaa datasta saatavien hyötyjen viestimisen vaikeus datan parissa toimivien yksiköiden ja johtoportaan välillä. Kehittyneet ja monimutkaiset dataratkaisut voivat olla välillä vaikeaa ymmärtää niitä käyttävienkin ihmisten toimesta, jolloin hyötyjen esittäminen on haastavaa aihetta heikommin tunteville. Haasteltavien mukaan asenteissa on tapahtunut muutosta, ja nykyään data on keskeisemmässä roolissa yhtiöiden strategiaa.

"Se mikä on varmaan semmonen vakuutusyhtiöiden synti, tai tämmönen, niin sitä dataahan on kerätty ihan käsittämättömän pitkän aikaa, ja sit kun se (data) vaan pitäs laittaa töihin entistä enemmän. Meillä hyödyntämispotentiaalia on todella todella paljon, et mitä me voidaan sillä datalla tehdä, et pitää ymmärtää se investointina koska se ei oo aina halpaa toi datahomma, tai mitä tehdään datan ympärillä niin se ei oo aina halpaa."

Haastateltavien mukaan teknologian saatavuus on nykypäivänä hyvää, mutta kehittyneiden työkalujen ja datamaailman monimutkaisuus ovat johtaneet pulaan työkalujen hyödyntämisen taitavista osaajista. Joissakin tapauksissa työtaakka kasautuu, kun yksittäiset henkilöt joutuvat hoitamaan lukuisia eri rooleja. Osaamisen puute tunnistettiin haastatteluissa keskeiseksi resursseihin liittyväksi haasteiksi. Vaikka investointihaluja dataan löytyy, ei investoinnit välttämättä tuota haluttua lopputulosta osaamisen puutteen vuoksi. Eräs haastateltava kertoi osaavien tekijöiden puutteesta seuraavasti:

"No joo, varmaan resurssit on sitten varmaan (haaste), et resurssit on tietyn kokoset ja niillä ei voida tehdä kuin tietyn verran asioita. Sitten ehkä pitäs valjastaa enemmän tekijöitä että saatais tän tyyppisiä ratkaisuja, että voitas vaikka yhdistää tekoälyllä sitä dataa […] et se vaatii tekijöitä ja resursseja ja rahaa, eli näkisin et siitä se on kiinni. Aika paljon taistellaan, samat henkilöt tekee useita eri asioita niin se on hankalaa se resurssointi massadatan hyödyntämisessä ja monessa muussakin asiassa että tekemistä on pirusti ja tekijöitä on aina liian vähän. "

Edistyneiden työkalujen olemassaolo ja saatavuus ei takaa niiden tehokasta käyttöä. Osaamisen puute työkaluja kohtaan johtaa niiden alihyödyntämiseen, tai työkalujen avulla luodut ennusteet ovat virheellisiä. Datan valtavaan määrään ja monimutkaisuuteen liittyvää osaamisen ja tietotaidon puutetta kuvattiin erään haastateltavan vastauksessa seuraavasti:

"Ja ihan niinku keskeisiä ja isoja (haasteita) on kun tota dataa katselee, niin pysyykö homma hanskassa siinä, ettei mee kausaliteetti ja korrelaatio sekaisin. Ja tulee siis se että datan valtava määrä ja mallien monimutkaisuus, niin se tekee kyllä helposti sokeeks sille, että siinä pystyy tekemään isoja virheitä. Ja tekijä rupee liikaa luottaa siihen malliin minkä hän on tehnyt. Ja käyttää ehkä malliansa semmosella alueella mihin sitä ei oo tehty

ja kalibroitu. Että tulee aikalailla mahdollisuuksia, mutta aikalailla myös uusia riskejä siitä että homma menee pieleen."

Datan suuri määrä ja monimutkaisuus nostaa alttiutta virheiden tekemiselle, joilla voi olla mittavat seuraukset, jos mallinnuksen tekijä ei osaa suhtautua skeptisesti omaan tuotokseensa. Haastatteluissa ilmeni myös, että joissain tapauksissa aihepiirin ymmärryksen vajavaisuudesta johtuva epätietoisuus siitä, mihin data-analytiikan osa-alueeseen tulisi keskittää resursseja voi aiheuttaa panostuksia, jotka eivät tuota halutun kaltaista lopputulosta.

"Kyl se enemmän on se ihmisen tyhmyys, eli se semmonenm, että kun on hienostuneet ja monimutkaiset työkalut ja sitten sieltä löytää, saa mallinnettua jotakin niin on syytä olla hyvin kriittinen siihen tekeleeseensä ja hyvin analysoida se, että että missä on rajat, mihin asti siihen voi luottaa ja millon sitten pitää hälytyskellojen soida."

"Et nythän meillä jos mennään data-analytiikan puolella, niin meillähän on erilaisia rooleja. Et kun tässä ollaan kouhkattu hirveesti AI:sta, data sciencesta ja koneoppimisesta, niin kaikki on nyt sitten palkanneet data scienteistieja. Ja he ovat nyt sitten tuhranneet vuosikausia kaikennäköstä, ja varmaan aika monessa firmassa on todetettu, et kyl ne jotain malleja saa, mutta tuotantoon ne on aika hankala viedä. [.] nyt sit ollaan huomattu et tarvitaanki datainsinöörejä, jotka huolehtii siitä et ne ennustemallit saa sitä dataa millä niitä koulutetaan, mitä niillä ennustetaan, ja sit kun ollaan siirrytty pilveen niin sitä pilviosaamista et miten siellä pilvessä ne koneoppimisympäristöt et tarvitaan niitä koneoppimisinsinöörejä, ja noistahan on nyt huutava pula. Nythän noi yliopistot kouluttaa noita data scientisteja paljon, mut tota alaosaa ei oikein osaa kukaan, et ei siellä niinku ton takia olla sitten hirveesti saatu mitään mullistusta aikaan että, tavallaan teknologia mahdollistaa, mutta ei nää oo niin hirveen helppoja asioita."

5.2.2 Datan ominaisuudet

Massadata on perusluonteeltaan vaikeasti hyödynnettävää sen ominaisuuksien, kuten volyymin, monikanavaisuuden ja matalan arvotiheyden vuoksi. Nämä ominaisuudet tekevät johtopäätösten tekemisen ihmisaivoin mahdottomaksi, ja sulkevat pois myös perinteiset datan käsittelyn menetelmät. Massadatassa piilevä arvo ymmärretään ja tiedostetaan hyvin, mutta arvon ulosmittaaminen on osoittautunut haastavaksi massadatan ominaisuuksien vuoksi. Massadatassa piilevän hyödyn ulosmittaaminen vaatii usein uusien teknologioiden, kuten tekoälyn valjastamista. Tämä taas johtaa edellisessä kappaleessa mainittuihin osaamisen ja tietotaidon puutteisiin, sillä osaaminen on vielä verrattain vähäistä. Kysyttäessä massadataan liitettyjen ominaisuuksien aiheuttamista haasteista, saatiin eräältä haastateltavalta vastaukseksi:

" Joo, väittäisin että se kohina ja siitä kohinan keskeltä niiden merkityksellisten signaalien löytyminen, niin ehkä se on massadatasta selkein ongelma et ihmissilmä tai ihmissilmä ei pysty usein tulkitsemaan sitä dataa ja löytämään sieltä semmosta merkityksellistä, et se vaatii usein sitä koneistoa [.] Et sen mä näkisin et se on isoin, et yksinkertaisesta datasta

on helppo nähdä signaaleja, mutta isosta datasta mikä tulee monesta eri lähteestä, yhestä asiakkaasta niin siinä menee niinku ihmisen ymmärrys ei riitä siihen."

Jotkut vakuutusyhtiöt ovat alkaneet käyttämään dataa hinnoittelun perusteena hyvin tarkalla tasolla. Datan oikeellisuuteen tulee kiinnittää huomiota heti datan luontihetkellä, sillä virheellinen data voi johtaa suuriin virheisiin koko hyödyntämisen prosessissa. Perustavanlaatuisella tavalla virheellinen data johtaa virheelliseen riskinmäärittelyyn ja lopulta hinnoitteluun. Yksi haastateltavista kuvasi ongelmaa seuraavasti:

"Eli siihenhän (liikennevakuutukseen) on tullu näitä hinnoittelukomponentteja vaikka kuinka ja paljon. Et siellä aikasemmin monella yhtiöllä oltiin jossain kuntatasolla, nyt oltiin postinumerotasolla. Joissain ollaan viety vieläkin tarkemmalle, et kummalla puolella katua se asunto sijaitsee. Ja sitten on noi, että vakuutuksenottajasta halutaan tietää kaikki mahdollinen veriryhmästä alkaen, ja sit sitä autoo kuitenkin saattaa käyttää useampi kuljettaja. Et me viedään se riskinvalinta ja hinnoittelu äärimmäisen tarkalle tasolla, ja sit me ei oikeesti välttämättä hinnotella sitä riskiä mikä me ostetaan. Et joku vanhempi hankkii auton, ja sit sillä ajaa perheen lapset."

5.2.3 Datan saatavuus ja saavutettavuus

Datan saatavuuden puute ja saavutettavuuden esteet tunnistettiin keskeisiksi massadatan hyödyntämistä estäviksi haasteiksi. Haastattelujen perusteella ymmärryksen puute datan hyödyntämisen ja hyödyntämisen mahdollisuuksista aiheuttaa haasteita datan saatavuuteen liittyen. Datan saatavuuden varmistaminen edellyttää ymmärrystä datan hyödyntämisestä läpi organisaation, aina toimihenkilötasolta johtoon asti. Muutama haastateltavista totesi toimihenkilöiden tietämyksen puutteen aiheuttavan esteitä asiakasdatan saavutettavuudelle. Toimihenkilöt vastaavat asiakasdatan keräämisen suostumuksen keräämisestä asiakkailta, ja puutteellinen ymmärrys voi johtaa välinpitämättömyyteen suostumusten keräämisessä. Suostumuksen puuttuminen evää yhtiöiden asiakasdatan keräämisen, jolloin heidän hallussa olevan datan edustavuus on vajavainen.

"Et jos miettii jotain CRM:ää, niin siellä pääsääntöisesti oo käyttäjälle yhtään turhaa tietolaatikkoo. Niis mitä paremmin niitä on täytetty, niin sitä paremman laatusta dataa meillä on. Voi tuntua tyhmältä käyttäjästä ja asiakaspalvelusta, et halutaan kerätä jotain tiettyä raksia johonkin CRM-järjestelmään, ja jonka ne yleensä ohittaa koska ei näe sitä oleellisena. [.] käyttäjälle saattaa tuntua tyhmälle jonkun muun tiedon kerääminen tietyssä muodossa, mutta se lyhentää sitä datan käsittelyn tarvetta sit melkoisesti."

Datan saavutettavuuteen ja saatavuuteen vaikuttaa voimakkaasti kolmansien osapuolten, kuten sovellustoimittajien tai verkkoselaimista vastaavien yhtiöiden vaatimukset ja rajoitukset datan hyödyntämiselle. Verkkoselainten evästeet ovat keskeisessä roolissa asiakasdatan keräämisessä, sillä ne sallivat mm. asiakkaan verkkokäyttäytymisen seurannan, ja saattaa tarjota vakuutusyhtiöillä tärkeää tie-

toa heidän asiakkaistaan. Kolmansien osapuolten evästedatan kerääminen loppuu kokonaan, kun lähitulevaisuudessa evästetuki häviää kaikista selaimista. Evästedatan saavutettavuuden esteet voivat hankaloittaa kokonaiskuvan muodostamista yhtiön asiakaskunnasta, ja tämä voi edelleen johtaa virheellisiin päätöksiin. Eräs haastateltava kuvasi ongelmaa näin:

"Jos 20-30 prosenttia asiakkaista kieltää evästeet ja analytiikan seurannan, niin sehän tarkoittaa sitä, että sitten meiltä lähtee se kokonaiskuva siitä pois ja se tarkkojen trendien havaitseminen muuttuu aika vaikeeks."

Haastatteluissa kerrottiin, että myös sovellustoimittajat voivat esittää ehtoja heidän sovellusten käytölle, jotka estävät vakuutusyhtiöitä kiinnostavan datan hyödyntämisen. Tämä kävi ilmi muun muassa seuraavassa vastauksessa, jossa haastateltava kertoi organisaatiossaan meneillään olevasta projektista:

"Kun meillä on tässä rakenteilla sellanen uus maailma missä yhdistetään useita sosiaalisen median massadatapalveluita ja voidaan yhdistää sitä (dataa) asiakkaaseen. Ja tai sitten käytännössä ei pystytä yhdistämään, koska kolmannen osapuolet estää sen, se on vähän niinkun moi moi. Sit meidän täytyy jotain muita avaimia löytää sinne, et pystyttäis ennustamaan sen massadatan perusteella."

Haastateltavien mukaan datan saavutettavuuden esteet aiheuttavat ongelmia suunnitelluille tulevaisuuden käyttötapauksille, kuten autovakuutuksien yksilölliselle, sensorien keräämään dataan perustuvalle hinnoittelulle. Eräs toimiala, joka omistaa suuria määriä vakuutusyhtiöitä kiinnostavaa dataa, mutta eivät luovuta sitä ulkopuolelle ovat autonvalmistajat. Nykyautojen tietojärjestelmät sensoreineen keräävät dataa joka hetki, mutta datan saatavuus on heikkoa. Yksi haastateltavista kuvasi asiaa näin:

"Se ois mielenkiintoista saada ajoneuvovalmistajien data, koska niillähän on kaikki uudet autot varsinki onlinessa kokoajan ja ne kerää sitä dataa kokoajan. Se ois mielenkiintonen saada aito käyttötieto, mutta eihän ne tietenkään anna hyödyntää sitä, koska ne haluaa jossain vaiheessa myydä sen."

Datan saatavuuteen vaikuttaa voimakkaasti myös informaation yksityisyyden sääntely, ja sen vaikutuksia käsitellään seuraavassa alaluvussa.

5.3 Informaation yksityisyyden sääntelyn vaikutukset vakuutusliiketoimintaan

Haastattelujen perusteella vakuutusyhtiöt noudattavat erityistä varovaisuutta käsitellessään sensitiivistä tietoa, ja huolehtivat datan anonymisoinnista jo datan luontihetkellä. Kaikkien haastateltavien mukaan informaation yksityisyyden

sääntely on ajantasaista nykypäivän vakuutusalan dataympäristöön nähden. Vakuutusyhtiöillä ei ole halua varastoida tarpeettomasti arkaluonteista dataa, sillä se altistaa yksityisyyden loukkauksen riskeille, mikä puolestaan voi aiheuttaa muun muassa mainehaittoja. Vakuutusyhtiöt ovat keränneet, säilyttäneet ja hyödyntäneet jo pitkään arkaluonteista tietoa asiakkaistaan. Informaation yksityisyyttä koskevan sääntelyn lisääminen ei ole siten vaikuttanut merkittävästi haittaavasti vakuutusyhtiöiden toimintaan. Vakuutusyhtiöillä on vakuutustoimintaan liittyviä lain suomia erityisoikeuksia- ja toisaalta velvollisuuksia säilyttämäänsä tietoon liittyen.

5.3.1 Vakuutusalan erityispiirteet ja alakohtainen sääntely

Haastattelujen perusteella vakuutusyhtiöillä on alan maineen vuoksi erityinen tarve toimintansa, kuten myös asiakkaitaan koskevan datan hyödyntämisen läpinäkyvyyteen. Vakuutusalan toimijoilla on haastattelujen perusteella epäluotettava maine, ja siksi datan hyödyntämisen läpinäkyvyys on vakuutusalalla erityisen tärkeää mainehaittojen minimoimiseksi. Yksi haastateltavista kuvasi tätä erityispiirrettä seuraavasti:

"Vakuutusyhtiöillä on se perinteinenkin ongelma, että näkyvyyttä ei oo tarpeeksi ja se aiheuttaa sitä, että vakuutusyhtiöihin kohdistuu epäluuloja. Eli kyllä siis voin sanoa, että tämmösestä historiallisesta näkökulmasta vakuutusyhtiöillä on ihan hyvä, että tulee vastuuta."

Vakuutusyhtiöillä on erivapauksia säilyttää sensitiivistä tietoa pidempään, kuin muilla aloilla. Eri datatyyppejä on sallittua säilyttää eripituisia aikoja, ja säilytysajan umpeuduttua datan hävittäminen tapahtuu automatisoidusti. Sääntely nähtiin myös osittain epäselvänä, ja tämä on lisännyt eri yksiköiden viestintää sääntelyn noudattamisesta huolehtivien compliance-yksiköiden välillä. Tietosuojaa koskevan sääntelyn noudattamisen huolehtiminen nähtiin tulevaisuudessa keskeiseksi tehtäväksi vakuutusyhtiöiden compliance-yksiköissä. Erityyppisen datan säilytysajat nähtiin myös monimutkaisena yhden haastateltavan näkökulmasta. Hän kuvasti datan säilytystä organisaatiossaan seuraavasti:

"Se on valtavan kompleksista, että koska datan saa hävittää asiakkaasta, no se riippuu tosi paljon asiakassuhteesta, et jos on vaikka vahinkoja ollu niin laki velvottaa meitä säilyttämään esimerkiksi henkilövahinkojenkin tapauksessa 50 vuotta sitä dataa. Et se ei oo ihan niin, et kun asiakas poistuu, ni me voidaan tuhota sitä dataa, et sitä voidaan tietyin periaattein säilyttää. Mut sitä varten on omat siivousajonsa tietokantoihin […]"

Informaation yksityisyyden sääntely vaikuttaa vakuutusalalla datan keräämiseen, hyödyntämiseen ja varastoimiseen. Sääntelyn nähtiin aiheuttavan myös haasteita datan hyödyntämisessä, mutta sääntelyn ei koettu varsinaisesti estävän datan hyödyntämistä tai data-analytiikan kehitystä. Sääntely nähtiin myös tasapuolisena tekijänä alan eri toimijoita kohtaan. Haastateltavat esittivät myös nä-

kemyksiä, joiden mukaan pakottava sääntely on mahdollisesti myös innovaatiota lisäävä tekijä, kun vakuutusyhtiöt joutuvat pohtimaan, miten kerätä ja hyödyntää dataa siten, että sääntelyä noudatetaan.

Yleisesti ottaen, informaation yksityisyyden sääntely nähtiin haastateltavien näkökulmasta tärkeänä ja aiheellisena. Toisaalta, erään haastateltavan näkemyksen mukaan sääntelyn edellyttämän tiedonantovelvollisuuden täyttäminen ei välttämättä tarkoita, että asiakas ymmärtää mitä tietoa hänestä kerätään, ja ei siten täytä tarkoitustaan:

"Periaatteessa kai voisi jopa sanoa, että onhan meillä aika hyväkin sääntely [.] mutta voihan itsekin jo sanoa, että on jossakin sovelluksessa ja se sitten sanoo, että oletko lukenut nää käytön ehdot ja hyväksytkö että sovellus selvittää sinusta sitä sun tätä, kylhän siin aika helposti lukematta ja enempää miettimättä laittaa rastin ruutuun. Enkä mä tiedä onko se, että miten siinä sääntelykään auttaa, sääntelyhän nyt on ajanut tähän, että tulee että pitää laittaa rasti ruutuun ja luvata että olen ymmärtänyt mukamas mitä tapahtuu."

5.3.2 GDPR

Vakuutusyhtiöt noudattavat tietoa käsitelleessään yleistä- ja alakohtaista sääntelyä, kuten Suomen vakuutussopimuslakia. Lainsäädännön lisäksi yhtiöt noudattavat myös ylemmän tason sääntelyä, kuten Euroopan Unionin yleistä tietosuojaasetusta GDPR:ää. GDPR:n vaikutuksia vakuutusalaan hillitsi se, että vakuutusyhtiöt ovat tottuneita sensitiivisen tiedon käsittelijöitä, ja noudattavat toiminnassaan GDPR:ää tiukempaa sääntelyä. GDPR nähtiin siten ylemmän, laveamman tason sääntelynä. GDPR:n koettiin tuoneen sensitiivisen tiedon käsittelyä koskevat ohjeistukset yleiseen tietoon. Yksi haastateltavista avasi GDPR:n vaikutuksia seuraavasti:

"No, tota noin yks kalleimpia direktiivejä mitä on viety tuotantoon, ja todella tarpeellinen, se on nyt ihan selkee juttu. Se on aiheuttanu aika paljon tekemistä tässä 2018 jälkeen, et mitä pitää tehdä erilailla, ja miten pitää tehdä tietoja ja miten niitä suojataan ja mitä kerätään ja mitä säilytetään. [.]Ja sit tietenki meidän toimiala on semmonen, että meillä on paljon sellasta tietoo asiakkaista, niin nykyisistä kuin entisistä, mitkä ovat taas sitten paljon paljon tiukemman sääntelyn piirissä kuin mitä GDPR on ja mitä se tarjoilee siihen sääntelyn ylimpään kerrokseen. Et siellä on luonnollisesti sit joku vakuutussalaisuus tulee tietoon, ja sit meillä on paljon sensitiivistä tietoo asiakkaista et on terveydentilatietoo, talouden tietoja, tämmösiä mitkä on ollu jo ennen GDPR:ää vahvasti piilossa pidettäviä tietoja ja anonymisoitua tietoa."

Eräs haastateltava vastasi GDPR:n aiheuttaman datan käsittelyn hankaloittamista koskevaan kysymykseen seuraavasti:

"No, ehkä sillon aluksi tuntu, että se hankaloittaa, mut nyt tavallaan ehkä tottunut sen vaatimuksiin, et ei, emmä nyt nää et se mitenkään hankaloittais erityisesti."

Yksi konkreettinen ilmi tullut GDPR:n aiheuttama muutos oli tarkempi käyttöoikeuksien rajaaminen, jolloin arkaluonteiseen dataan pääsevät käsiksi vain henkilöt, joille käyttöoikeus on olennainen heidän työtehtävän hoitamisessa. Yksi haastateltava kuvasi vaikutusta näin:

[.] sit siitä pääsee lähestulkoon joka päivä keskustelemaan meidän loppukäyttäjien kanssa, et miks he ei saa sitä jotain tietoo nähdä, ja sit otetaan aina toi GDPR-kortti mukaan et sinulla ei oo oikeutta tähän tietoon [.]

5.3.3 Dataetiikka

Sääntelyn noudattamisen lisäksi vakuutusyhtiöt miettivät myös datan hyödyntämiseen liittyviä eettisiä pulmia. Sääntelyn noudattamisen ohella dataetiikan huomioonottaminen nähtiin tärkeänä uusia data-analytiikan kehityskohteita suunnitellessa. Yksi haastateltavista kuvasi eettistä problematiikkaa seuraavasti, kun häneltä kysyttiin miten vakuutusyhtiöt ottavat dataetiikan huomioon:

”Kyllä, se on tietysti kans tärkee, että lainausmerkeissä se mikä on sallittua ei välttämättä ole oikein, eli tän tyyppinen ajattelu on kyllä mukana [.]Et se on aina hyvin selkee että me tehdään hyvinkin varovasti, ehkä jopa liiankin varovasti asioita. Et toi on niissä kyllä mukana toi etiikkapuoli, mulla on tossa etiikkakin kirjotettu tohon viereen, ihan sellaset eettiset ohjeet tossa vieressä että kyllä se on tärkee osa. ”

Vakuutusyhtiöiden omistaman sensitiivisen tiedon, kuten terveydentilatietojen hyödyntäminen nähtiin eettisesti oikeana, kunhan dataa käsitellään massana ja sitä ei voida jäljittää yksilöön.

”Siitä on nyt paljon keskusteltu eettisestä näkökulmasta, että mitä tietoa saamme hyödyntää esimerkiks hinnoittelussa, et saadaanko hyödyntää esimerkiks terveydentilatietoja, no ok, tietyllä rajoituksilla saadaan kyllä hyödyntää. Mut sit kun lähdetään rakentaa jotain uutta tuotetta, niin kyllähän meillä saa olla käytössä ne kaikki terveydentilatiedot, kunhan me katotaan niitä möykkynä [.] niin sitä saadaa käyttää, ja se on myös mun mielestä eettisesti oikein käyttää sitä tietoo. ”

6 TULOSTEN TULKINTA JA POHDINTA

Tässä luvussa tarkastellaan empiirisen osion tuloksia, sekä verrataan niitä tutkielman teoriaosuudessa esiteltyyn aiempaan kirjallisuuteen ja pyritään tunnistamaan näiden välisiä eroja ja yhtäläisyyksiä. Luvussa esitetään empiirisen tutkimuksen keskeisimmät tulokset, sekä vastataan tutkielman tekoa ohjanneisiin tutkimuskysymyksiin. Tulosten esittelyn lomassa esitetään relevantteja jatkotutkimusaiheita. Lopuksi pohditaan tutkimuksen merkittävyyttä vakuutusalalle ja aiheen piirissä työskentelevälle tiedeyhteisölle, sekä osoitetaan tutkimuksen rajoitteita. Tutkimuskysymykset olivat:

- Miten vakuutusyhtiöt voivat hyödyntää massadataa?
- Mitä ovat vakuutusyhtiöiden olennaisimmat haasteet massadatan hyödyntämisessä?
- Miten informaation yksityisyyteen liittyvä sääntely vaikuttaa datan hyödyntämiseen vakuutusyhtiöissä?

Tutkielman teoriaosiossa pyrittiin tutustumaan tutkimuksen aihepiirin tärkeimpiin käsitteisiin ja teorioihin, kuten massadataan, sekä informaation yksityisyyteen. Empiirisessä osiossa kerättiin aineistoa teemahaastattelujen avulla. Haastattelun teemat olivat massadatan hyödyntäminen vakuutusalalla, massadatan hyödyntämisen haasteet, sekä informaation yksityisyyteen liittyvän sääntelyn vaikutukset datan hyödyntämiseen.

6.1 Tulkinta

6.1.1 Massadatan hyödyntäminen ja potentiaali vakuutusalalla

Massadatan mahdollisuudet vakuutusalalla ovat merkittäviä ja moninaisia, mutta toistaiseksi se ei ole saanut aikaan suuria mullistuksia. Massadataan on

liitetty vakuutusalalla suuria odotuksia, mutta ne eivät ole vielä realisoituneet. Tämä on linjassa aiemmasta kirjallisuudesta tehtyihin havaintoihin, kuten Hussain & Prieto (2016), joiden mukaan vakuutusala on kyennyt hyödyntämään massadatan koko potentiaalia heikosti muihin aloihin verrattuna. Dataan pohjautuvan päätöksenteon odotetaan kuitenkin vielä aiheuttavan mullistuksia alalle. Dataan panostamiseen ja tiedolla johtamiseen kohdistuvat asenteet ovat muuttuneet lähivuosien aikana, ja data on otettu keskeiseksi osaksi yhtiöiden strategiaa, ja siten myös dataan kohdistuvat investoinnit ovat yleistyneet. Massadatan hyödyntämisen mahdollistavat työkalut ovat verrattain uusia, ja niiden täyttä potentiaalia ei ole pystytty vielä hyödyntämään vakuutusyhtiöiden toimesta muun muassa osaamisen puutteen vuoksi.

Vakuutusyhtiöt voivat hyödyntää massadataa aktuaaritoiminnassa, asiakkuudenhallinnassa, asiakaspalvelussa, asiakasanalytiikassa, vahingontorjunnassa, riskinvalinnassa, petosanalytiikassa sekä myynnissä ja markkinoinnissa. Haastattelut tarjosivat tukea Hussainin & Prieton (2016) väitteelle, jonka mukaan vakuutusyhtiöiden hyödyntämä data on pääasiallisesti heidän itse tuottamaansa. Kuitenkin, datan merkityksen ja mahdollisuuksien ymmärryksen myötä ulkoista dataa on alettu hyödyntämään entistä enemmän. Ulkoisia datan lähteitä ovat pääasiallisesti julkishallinnolliset organisaatiot, kuten Tilastokeskus. Sisäisiä datan lähteitä ovat vakuutusyhtiöiden käyttämät tietojärjestelmät, kuten asiakkuudenhallintajärjestelmät.

Teoriaosion löydöksiä vahvistaen, sosiaalisen median massadataa ei juurikaan hyödynnetä vakuutusyhtiöiden toimesta, vaan sosiaalinen media nähdään ennemminkin markkinointikanavana kuin merkittävänä datan lähteenä. Cavanillasin, Curryn ja Wahlsterin (2016) mukaan sosiaalisen media tarjoaa mahdollisuuksia asiakaskokemuksen parantamiseen, mutta haastateltavat eivät tunnistaneet tätä potentiaalia. Mainittuja syitä tälle olivat vakuutusalan erityispiirteet kuten tuotteiden abstraktius, sekä se että vakuutusyhtiöillä ei ole ollut halua investoida ulkoiseen dataan. Haastatteluista ei käynyt myöskään ilmi, että vakuutusyhtiöt hyödyntäisivät muuta kuin tekstin muodossa olevaa dataa. Tämä tarjoaa vahvistusta Hussainin ja Prieton (2016) väitteelle, jonka mukaan vakuutusala kykenee hyödyntämään massadatan eri muotoja heikosti muihin aloihin verrattuna. Gandomi & Haider kertovat puhelinasiakaspalvelua harjoittavien yritysten hyötyvän erityisen paljon äänianalytiikan hyödyntämisestä. Suomalaiset vakuutusyhtiöt käyttävät puhelinta keskeisenä palvelukanavanaan, ja velvollisuus puheluiden nauhoittamiselle tuottaa suuret määrät äänimuotoista dataa. Vakuutusalalla hyödynnettävä massadata on siten haastatteluiden perusteella monikanavaista, mutta ei niinkään monimuotoista. Massadatalla tyypillinen ominaisuus on sen volyymi, joka edellyttää usein reaaliaikaista analysointia hyötyjen saavuttamiseksi (Davenport ym. 2007). Haastattelujen perusteella käyttötapaukset reaaliaikaisesti tapahtuvalle massadatan analyysille ovat vähäisiä, ja ainoana tänä päivänä käytössä olevana käyttötapauksena nousi esiin vakuutuspetosten ehkäisyyn ja havaitsemiseen tähtäävä petosanalytiikka.

Aiemmassa kirjallisuudessa esitettyjen näkemysten, kuten Corbettin ym. (2013) mukaan massadatan hyödyntämiseen vakuutusalalla liittyy suurta potentiaalia, ja haastateltavien näkemykset tarjosivat tälle vahvistusta. Vakuutusyhtiöillä on runsaasti kokemusta datan analysoinnista ja siihen pohjautuvista päätöksistä, mutta nykyajan dataympäristö tarjoaa yhä paremmat puitteet tietoon perustuviin päätöksiin. Suuren volyymin omaava, useista eri kanavista tuleva data tarjoaa vakuutusyhtiöille kyvyn tehdä datasta tarkempia ennusteita. Haastateltujen näkemyksen tukivat Jiang & Songin (2016) näkemystä siitä, että data nousee yhä keskeisemmäksi kilpailutekijäksi vakuutusalalla. Vakuutusyhtiöiden kyky hyödyntää tehokkaasti dataa nähdään nousevan tärkeäksi kilpailutekijäksi alalla. Datan tehokas hyödyntäminen mahdollistaa paremman ymmärryksen asiakkaista sekä vakuutettavista riskeistä. Tulevaisuuden potentiaali pohjautuu pitkälti teknologioihin, joita tarvitaan hyödyllisten johtopäätösten tekemiseksi suuresta datamassasta, kuten tekoälyyn ja koneoppimiseen. Dataan pohjautuva vakuutusten individualisointi parantaa kannattavuutta ja tehostaa hinnoittelua. Corbettin ym. (2013) mukaan pienet internet-perustaiset toimijat kykenevät haastamaan myös suurempia vakuutusalan organisaatioita. Boobierin (2016) mukaan data-analytiikkaan tehtävät investoinnit voivat lisätä pienempien yhtiöiden kannattavuutta. Haastateltavat tunnistivat, että innovatiivisia dataan pohjautuvia tuote- ja hinnoitteluratkaisuja käyttävistä uusista vakuutusyhtiöistä on maailmalla jo esimerkkejä, ja ne pystyvät haastamaan myös vakiintuneemmat alan toimijat ja tekemään kannattavaa liiketoimintaa. Aiemmassa kirjallisuudessa Baesens ym. (2016) kertovat massadatan mahdollisuuksista parantaa tuottavuutta käyttöön perustuvan hinnoittelun avulla.

6.1.2 Massadatan hyödyntämisen haasteet vakuutusalalla

Suuren potentiaalin vastapainona massadatan hyödyntämisen haasteet vakuutusalalla ovat moninaisia ja moniulotteisia, ja esiin nousseet haasteet ovat suurimmilta osin tunnistettu myös aiemman massadataa koskevan kirjallisuuden toimesta. Keskeisimmät esiin nousseet haasteet liittyvät resursseihin, datan ominaisuuksiin sekä datan saatavuuteen. Haastattelujen perusteella vakuutusyhtiöillä on käytössään työkaluja jotka mahdollistavat massadatan hyödyntämisen, mutta kärsivät osaamisen puutteesta. Monimutkainen nykypäivän dataympäristö ja kehittyneet työkalut vaativat erityisosaamista, jota on tänä päivänä harvassa. Hussainin ja Prieton (2016) mukaan tietotaidon ja kehittämisen esteenä on organisaatiokulttuuri, joka ei tunnista massadatan mahdollisuuksia liiketoiminnalle. Tiefenbacherin & Olrbrichin (2015) mukaan taas hyötyjen saavuttamiseksi vaadittavia ominaisuuksia ei kyetä tunnistamaan riittävän hyvin, ja siksi massadatan potentiaalia on vaikea saavuttaa. Boobier (2016) taas mainitsi massadatan ympärille perustetun nykyaikaisen, toimivan organisaation ja työkalujen puutteen keskeiseksi vakuutusalan haasteeksi. Haastattelujen perusteella vakuutusyhtiöiden päätöksentekijöiden asenteissa dataa kohtaan on tapahtunut muutoksia, ja data otetaan usein keskeiseksi osaksi yhtiön strategiaa. Asenteiden muutosten myötä myös halukkuus dataa koskeviin investointeihin on lisääntynyt,

vaikka taloudelliset resurssit ovat edelleenkin yksi hyödyntämisen estävistä haasteista.

Haastattelujen perusteella vakuutusyhtiöissä tunnistetaan massadatassa piilevä arvo, mutta arvon ulosmittaaminen on hankalaa massadatan ominaisuuksien vuoksi. Nämä ominaisuudet tekevät johtopäätösten tekemisen ihmisaivoin mahdottomaksi, ja sulkevat pois myös perinteiset datan käsittelyn menetelmät. Datan volyymi, moninaisuus ja todenmukaisuus nousivat esiin haasteita aiheuttavina massadatan ominaisuuksina. Volyymi ja monikanavaisuus poissulkevat perinteiset datan käsittelyn menetelmät, ja edellyttävät kehittyneiden teknologioiden, kuten tekoälyn valjastamista. Datan suuri määrä ei kuitenkaan takaa ennusteiden toteutumista, mikäli data on perustavanlaatuisesti virheellisestä, eli epätotuudenmukaista. Kyseiset ominaisuudet ja niiden aiheuttamat haasteet ovat laajasti tunnistettuja aiemmassa kirjallisuudessa esimerkiksi Gandomin & Haiderin (2015), sekä Eatonin ym. (2012) mukaan, ja kuvastavat hyvin massadataan usein liitettyjä ominaispiirteitä.

Dataa on yleisesti ottaen hyvin saatavilla, mutta silti saatavuuden tai saavutettavuuden ongelmat aiheuttavat haasteita vakuutusyhtiöissä. Saatavuuteen vaikuttaa kolmansien osapuolten rajoitteet tai haluttomuus jakaa dataa, organisaatioissa piilevä heikko ymmärrys datan käyttötarkoituksista, sekä informaation yksityisyyden sääntely. Informaation yksityisyyden sääntelyn on todettu vaikeuttavan massadatan hyödyntämistä vakuutusalalla mm. Hussainin & Prieton (2016) toimesta. Kerätyn aineiston perusteella sääntelyn lisäksi kolmannet osapuolet, kuten sovellustoimittajat tai selainyhtiöt rajoittavat massadatan käyttöä esimerkiksi evästeasetusten välityksellä. Haastateltavien näkemysten mukaan kaupallisten, kolmansien osapuolten rajoitukset koettiin jopa voimakkaammin datan käyttöä määrittäviksi kuin informaation yksityisyyden sääntely. Kolmansien osapuolten rajoituksia ei tunnistettu massadatan hyödyntämisen haasteeksi aiemman kirjallisuuden perusteella, joten tulos on yllättävä ja vaatii syvempää paneutumista. Datan saatavuuteen voi vaikuttaa se, että datan hyödyistä ja käyttötarkoituksista ei ole onnistuttu viestimään koko organisaatioille. Tämä voi aiheuttaa työntekijöiden keskuudessa välinpitämättömyyttä asiakkaiden tiedonkeruulupien keräämistä kohtaan, mikä johtaa edempänä ongelmiin datan oikeellisuuden suhteen.

6.1.3 Informaation yksityisyyden sääntelyn vaikutukset massadatan hyödyntämiseen vakuutusalalla

Herschelin & Miorin (2017) mukaan datan eettinen käyttö ei ole mahdollista ilman tietotaitoa datan luottamuksellisesta ja yksityisyyttä varjelevasta käsittelystä. Informaation yksityisyyttä koskevaa sääntelyä, kuten GDPR:ää, pidettiin yleisesti tarpeellisena, sekä ajantasaisena nykypäivän dataympäristöön nähden, eikä sen koettu estävän massadatan hyödyntämistä tai analytiikan menetelmien kehitystä. Soria-Comas & Domingo-Ferrerin mukaan (2015) liika informaation yksityisyyden sääntely voi hidastaa teknologian kehitystä, joten tämä ei ole käynyt haastattelujen perusteella toteen vakuutusalalla. Lisäksi GDPR:n on katsottu olevan joiltain osin vanhentunut heti valmistuessaan, ja siten epäyhteensopiva

nykyisen dataympäristön kanssa, erityisesti massadatan suhteen (Zarsky, 2018). Siten tämä tulos poikkeaa aiemmassa kirjallisuudessa esitetyistä näkemyksistä. Vakuutusalan yritykset ovat tottuneet käsittelemään sensitiivistä tietoa ja noudattamaan sitä koskevaa lainsäädäntöä kuten vakuutussopimuslakia, jonka nähtiin olevan tiukempaa kuin esimerkiksi GDPR. Lisäksi vakuutusalan yrityksille on toimintansa mahdollistamiseksi ja myös asiakkaiden turvaamiseksi asetettu yleisestä sääntelystä poikkeavia erityismyönnytyksiä ja velvollisuuksia tiedon säilyttämisen suhteen. Tämä on eräs mahdollinen syy siihen, että lisääntynyt informaation yksityisyyden sääntely ei vaikuta voimakkaasti alaan datan käsittelyyn.

Sääntelyn noudattamisen lisäksi vakuutusyhtiöt huolehtivat tiedonkäsittelyn eettisyydestä, sillä vaikka jokin olisi sallittua, se ei tarkoita, että se olisi eettisen tarkastelun valossa oikein. Tiedonantovelvollisuuden täyttäminen ei välttämättä tarkoita, että asiakkaalla on tarvittava tieto myytävästä tuotteesta. Kyseinen ongelma liittyy sekä vakuutusalan, että datamaailman monimutkaisuuden aiheuttamiin viestimisen vaikeuksiin. Kuten aiemmin todettiin, datan käsittelyn hyödyistä on vaikeaa viestiä jo organisaation eri yksiköiden välillä ratkaisujen monimutkaisuudesta ja vaikeasti selitettävistä arvonluonnin mekanismeista johtuen. Siten ei ole yllättävää, että selkeä viestintä myös organisaation ulkopuolisille ryhmille on haasteellista. Richardsin & Kingin (2014) mukaan hyödyntämisen läpinäkyvyys eettisen datan käsittelyn kulmakiviä ja edesauttaa asiakkaiden ja yhtiöiden välistä luottamusta. Luottamus taas vaikuttaa yksilöiden halukkuuteen jakaa itseään koskevaa dataa. (Richards & King, 2014) Haastatteluiden perusteella eettisyydestä huolehtiminen liittyy vakuutusalan maineen aiheuttamaan korostuneeseen tarpeeseen toiminnan läpinäkyvyydestä. Haastateltavat kokivat, että vakuutusyhtiöillä on korostunut tarve datan käsittelyn läpinäkyvyydelle alan kärsimien mainehaittojen vuoksi. Tätä piirrettä ei tunnistettu teoriaosiossa aiemman kirjallisuuden perusteella vakuutusalan erityispiirteeksi, ja on siten mielenkiintoinen jatkotutkimusaihe.

6.2 Pohdinta

Tehty tutkimus ja sen tulokset valottavat suomalaisten vakuutusyhtiöiden dataanalytiikan ja massadatan hyödyntämisen nykytilaa sekä tulevaisuudennäkymiä. Aiempaan teoriaan keskittyvässä osioissa pureudutaan aiheeseen liittyvään aiempaan kirjallisuuteen. Empiirisessä osiossa kartoitetaan aihepiirin asiantuntijoiden kokemuksia massadatasta, hyödyntämisen haasteista sekä informaation yksityisyyden sääntelyn vaikutuksista datan hyödyntämiseen vakuutusyhtiöissä. Massadatan ennustavan aiheuttavan koko alaa muovaavia muutoksia, mutta akateeminen yhteisö on tutkinut aihetta toistaiseksi niukasti. Siten tehty tutkimus hyödyttää sekä akateemista tiedeyhteisöä, että vakuutusalalla toimivia yrityksiä tarjoten tietoa nykytilasta, tulevaisuudennäkymistä sekä ehdottaen relevantteja jatkotutkimusaiheita. Aiempaan kirjallisuuteen nähden mielenkiintoisia

tutkimuksessa esiin tulleita seikkoja oli suomalaisten vakuutusyhtiöiden maineen aiheuttama korkea tarve datan käsittelyn läpinäkyvyydelle, sekä kolmansien osapuolten rajoitukset datan käsittelyn suhteen, joiden todettiin hankaloittavan käsittelyä jopa enemmän kuin julkishallinnollinen sääntely. Haastatteluiden perusteella vakuutusyhtiöt eivät hyödynnä heillä hallussa olevaa äänimuotoista dataa, kuten puhelutallenteita. Kirjallisuuskatsauksen mukaan äänianalytiikassa piilee hyötyä erityisesti puhelinasiakaspalvelua harjoittaville yrityksille, kuten vakuutusyhtiöille. Siten äänianalytiikan mahdollisuuksien tunnistaminen ja hyödyntäminen suomalaisissa vakuutusyhtiöissä on varteenotettava jatkotutkimusaihe.

Kuten yleensäkin, tutkimuksella on myös rajoitteita. Tutkimusta varten oli tarkoitus suorittaa useampia haastatteluita, mutta aihepiirin parissa työskenteleviä ihmisiä oli lopulta vaikea löytää ja tavoittaa. Otantaa kasvattamalla voitaisiin nostaa tutkimuksen luotettavuutta, sekä löytää näkemyksiä joita näissä haastatteluissa ei käynyt ilmi. Tutkimuksessa ei huomioida sitä, minkälaisissa organisaatioissa haastateltavat työskentelevät. Massadatan hyödyntämiseen ja sitä koskeviin asenteisiin voi vaikuttaa mm. organisaation koko (Boobier, 2016). Siten haastateltavien edustamien organisaatioiden tyypittely voisi tarjota hyödyllistä tietoa massadatan käytöstä erityyppisissä vakuutusalan organisaatioissa. Toisaalta, organisaatioiden tyypittely voi aiheuttaa ongelmia aineiston anonymisoinnin suhteen. Haastateltavat toimivat suomalaisissa vakuutusalan organisaatioissa, ja otannan levittäminen myös ulkomaisiin toimijoihin parantaisi luotettavuutta. Eräänä rajoitteena nähdään puuttumattomuus haastatteluissa käytettäviin käsitteisiin. Kuten kirjallisuusosiossa todettiin, massadatan käsitteelle ei ole olemassa vakiintunutta määritelmää. Haastattelussa ei keskitytty lainkaan massadatan määritelmään, jolloin massadatan käsite jää haastateltavien oman arvioinnin varaan. Haastateltavilta ei myöskään suoraan kysytty suoraan minkä muotoista dataa he hyödyntävät, jolloin vastauksen saaminen edellytti vastausten tulkitsemista epäsuorasti. Haastattelurungon päivittäminen voisi tuoda siten kattavampia tuloksia. Eräs keskeinen informaation yksityisyyden riski on datan sekundäärinen käyttö. Haastatteluista saatiin viitteitä lainsäädännön ja muun sääntelyn rajoissa tapahtuvasta datan sekundäärisestä käytöstä, mutta tähän ei haastattelussa pureuduttu enempää jatkokysymysten avulla.

7 YHTEENVETO

Tässä tutkielmassa haluttiin kartoittaa massadatan hyödyntämisen nykytilaa ja potentiaalia vakuutusalalla. Lisäksi haluttiin tietää, miten informaation yksityisyyden sääntely vaikuttaa datan hyödyntämiseen vakuutusyhtiöiden toiminnassa. Tutkimuksessa lähdettiin liikkeelle kolmesta tutkimuskysymyksestä, jotka olivat *"Miten vakuutusyhtiöt voivat hyödyntää massadataa?"*, *"Mitä ovat vakuutusyhtiöiden olennaisimmat haasteet massadatan hyödyntämisessä?"*, sekä *Miten informaation yksityisyyden sääntely vaikuttaa datan hyödyntämiseen vakuutusalalla?"*.

Vastausta tutkimuskysymyksiin ryhdyttiin selvittämään luomalla katsaus massadataa ja informaation yksityisyyttä koskevaan aiempaan kirjallisuuteen. Tämän jälkeen aihetta kartoitettiin empiirisen tutkimuksen avulla. Empiirisen tutkimusmenetelmänä käytettiin laadullista tutkimusta, ja aineistonkeruumenetelmäksi valikoitui teemahaastattelu. Haastattelukysymysten laatimisessa otettiin huomioon tutkielman tutkimuskysymykset, sekä teoriaosiossa tehdyt löydökset.

Kirjallisuuden perusteella todettiin, että termille massadata ei ole olemassa ajankohtaisuudestaan huolimatta yleisesti hyväksyttyä määritelmää. Massadatailmiön eräs olennainen syy on informaation tuottamisen ja levittämisen helpottuminen, joka on johtanut datamäärien räjähdysmäiseen kasvuun. Datanhallinta perinteisin menetelmin on mahdotonta datan alati kasvavan määrän, volyymin sekä hajanaisuuden vuoksi. Massadatan ominaisuuksista puhuttaessa käytetään usein kolmen V:n mallia, johon sisältyy volyymi (volume), nopeus (velocity) sekä moninaisuus (variety). Kolmen V:n mallia on jatkokehitetty kuuden V:n malliksi ja siihen on lisätty edellä mainittujen ominaisuuksien lisäksi totuudenmukaisuus (veracity), vaihtelevuus (variability) sekä arvo (value). Massadata on vaikuttavuutensa vuoksi tärkeä tutkimusaihe. Massadatan tutkimuksen avainteemat ovat informaatio, teknologia, metodit ja vaikutus. Massadatan tutkimus käsittelee aihetta usein datan analysointiin vaadittavien teknologioiden näkökulmasta. Perinteiset menetelmät eivät toimi massadatan hallinnoimisen ja analysointiin, vaan arvon ulosmittaamiseksi tarvitaan suorituskykyisiä, skaalautuvia ja joustavia teknologioita. Datamäärän suuruuden lisäksi haasteita aiheuttaa tarve datan

reaaliaikaiselle analysoinnille. Sosiaalisen median tuottamasta massadatasta on muodostunut tärkeä tiedonlähde niin yksityisten kuin julkihallinnollistenkin organisaatioiden päätöksenteossa.

Vakuutusala on ollut koko olemassaolonsa ajan hyvin datapainotteista, ja datan merkitys on kasvamassa yhä suuremmaksi. Datan digitointi ja datafikaatio mahdollistavat datan tehokkaamman hyödyntämisen ja massadatan odotetaan muuttavan alaa perinpohjaisesti. Toimintaympäristön muutoksien seurauksena järjestelmien, palvelukanavien ja datan määrä on kasvanut, mikä tekee vakuutusyhtiöiden liiketoiminnasta aiempaa kompleksisempaa. Massadatan mahdollisuuksien ohella hyödyntämiseen liittyy myös merkittäviä haasteita. Lisäksi toimialan toimintakenttää ovat muuttaneet pienemmät toimijat, jotka pystyvät horjuttamaan vakiintuneiden toimijoiden jalansijaa innovatiivisilla ratkaisuilla ja teknologiaa hyödyntämällä.

Kerätyn datan määrän nopeasta kasvusta huolimatta yksilöillä on usein heikko käsitys siitä, mitä dataa heistä kerätään. Datan määrän nopean kasvun ja datajoukkoihin sisältyvän datan arkaluonteisuuden todettiin yksityisyyden loukkauksien riskiä, ja massadata muodostaakin kenties digitaalisen ajan suurimman haasteen yksityisyyttä koskevan lainsäädännön laatimiselle. Monilähteisyydestä, volyymista ja yleisestä sekavuudesta huolimatta massadatan lähde voi olla pääteltävissä yksittäiseen ihmiseen. Massadatan sekundäärinen käyttö, henkilökohtaisen datan valjastaminen liiketoiminnallisiin tarkoituksiin sekä puutteellinen tietoturva lisäävät yksityisyyden loukkausten riskejä. Sosiaalinen media on osaltaan kasvattanut yksityisyyden loukkausten alttiutta. Informaation yksityisyydestä käyty keskustelu on herättänyt huolia myös siitä, että liiallinen yksityisyyden varjeleminen hidastaa teknologian kehitystä.

Datan ja sen hyödyntämisen eettinen valvonta on pitkälti datan valmistajien, käyttäjien ja toimittajien käsissä, sillä tästä vastaavaa auktoriteettia ei ole olemassa. Euroopan Unionin yleinen tietosuoja-asetus eli GDPR on tunnetuimpia esimerkkejä maiden rajoja ylittävästä tietojenkäsittelyn valonnasta. GDPR asettaa yrityksille ja organisaatioille luonnollisia henkilöitä koskevan henkilötiedon prosessoinnin säädökset. Asetus on herättänyt kritiikkiä, ja sen katsotaan sopivan huonosti nykypäivän dataympäristöön erityisesti massadatan osalta. Massadata on muuttanut dataympäristöä nopeasti, ja siksi osan GDPR:n sisällöstä on esitetty olleen vanhentunutta jo valmistuessaan.

Empiirisessä osuudessa toteutetussa teemahaastattelussa haastateltiin vakuutusalalla toimivia data-analytiikan asiantuntijoita. Haastattelun tulokset mukailevat pääosin aiemmasta kirjallisuudesta saatuja havaintoja. Vakuutusyhtiöt ovat analysoineet hallitsemaansa dataa koko alan olemassaolon ajan. Massadatan potentiaali vakuutusalalla nähdään hyvin suurena, mutta siihen liitetyt odotukset eivät ole toistaiseksi realisoituneet. Tämänhetkinen massadatan käyttö on suhteellisen rajoittunutta etenkin hyödynnettävän datan muodon osalta. Aiemman kirjallisuuden perusteella tehdyistä havainnoista poiketen, sosiaalisen median dataa ei nähdä alalla kovinkaan potentiaalisena tiedon raaka-aineena. Massadatalle tyypillistä reaaliaikaista analyysia ei juurikaan hyödynnetä petosanalytiikan lisäksi. Data nähdään alalla keskeisenä kilpailutekijänä, ja se on osattu

osaksi yhtiöiden strategiaa. Tulevaisuuden potentiaali nojaa pitkälti kehittyneiden teknologioiden, kuten tekoälyn ja koneoppimisen valjastamiseen. Uusien teknologioiden ja innovatiivisten ratkaisujen ympärillä toimivat pienemmät, uudet toimijat ovat jo nyt muovanneet alaa Suomen ulkopuolella.

Massadatan suuren potentiaalin rinnalla todettiin massadatan hyödyntämiseen liittyvän suuria haasteita, joiden vuoksi massadataan kohdistuvat odotukset eivät ole realisoituneet vakuutusalalla. Massadatan hyödyntäminen vakuutusalalla nähtiin jokseenkin rajoittuneena etenkin datan eri muotojen osalta. Keskeisimpien haasteiden todettiin liittyvän myös aiemmassa kirjallisuudessa tunnistettuihin teemoihin, eli resursseihin, datan ominaisuuksiin sekä datan saatavuuteen. Dataa kohtaan tehtävien investointien puute, sekä osaamisen ja tietämyksen puute nykyaikaisesta dataympäristöstä nähtiin hyödyntämistä rajoittavina tekijöinä. Aiemmasta kirjallisuudesta tehdyistä löydöksistä poiketen datan saatavuuteen vaikutti voimakkaasti informaation yksityisyyden sääntelyn lisäksi kolmansien osapuolten rajoitteet datan käytölle.

Informaation yksityisyyden sääntelyn todettiin olevan tarpeellista ja ajantasaista nykyiseen dataympäristöön nähden, eikä sen nähty vaikeuttavan merkittävästi datan hyödyntämistä tai menetelmien kehittämistä. Mahdolliseksi vakuutusalaa suojaavaksi tekijäksi todettiin alan kokeneisuus arkaluonteisen tiedon käsittelyssä, sekä alalle asetetut erityisoikeudet ja velvollisuudet informaation yksityisyyttä kohtaan. Sääntelyn noudattamisen lisäksi datan käsittelyä ohjaavan eettisen tarkastelun todettiin olevan keskeisessä roolissa vakuutusyhtiöiden toiminnassa. Vakuutusalan maineen todettiin olevan datan käsittelyn läpinäkyvyyden tarvetta kasvattava tekijä. Tätä seikkaa ei tunnistettu aiemmasta kirjallisuudesta.

Tutkimuksessa esiintyneiksi relevanteiksi jatkotutkimusaiheiksi todettiin sosiaalisen median massadatan hyödyntäminen vakuutusalalla, vakuutusalan maineen aiheuttama korostunut tarve datan käsittelyn läpinäkyvyydelle, sekä kolmansien osapuolten rajoitukset datan hyödyntämistä estävinä tekijöinä. Rajoitteiksi todettiin haasteltujen asiantuntijoiden suhteellisen pieni määrä, puutteellinen käsitemäärittely haastattelutilanteessa, sekä haastateltavien edustamien organisaatioiden tyypittelyn puute.

LÄHTEET

Abkenar, S.B., Kashani, M.H., Mahdipour, E., Jameii, S.M. (2021) Big data analytics meets social media: A systematic review of techniques, open issues and future directions. *Telematics and Informatics, Volume 57.*

Ananthanarayanan, G., Bahl, P., Bodik, P., Chintalapudi, K., Philipose, M.,Ravindranath, L., Sinha, S. (2017) Real-Time Video Analytics: The Killer App for Edge Computing. *Computer, Volume 50.* 58-67.

Arfeit, E., Basin, D., Debois, S. (2019)Monitoring the GDPR. European Symposium on Research in Computer Security. *ESCORICS 2019, European Symposium on Research in Computer Security;* (681 - 699) Luxembourg, September 23-27, 2019.

Baesens, B., Bapna, R., Marsden, J., Vanthienen, J., Zhao, J. (2016) Transformational Issues of Big Data and Analytics in Networked business. *MIS Quarterly, 38,* 629-631.

Bélanger, F., & Crossler, R. E. (2011). Privacy in the Digital Age: A Review of Inf ormation Privacy Research in Information Systems. *MIS Quarterly, 35 (4),* 1017-1041.

Boobier, T. (2016) *Analytics for Insurance. The Real Business of Big Data*. Wiley. Haettu osoitteesta 10.1002/9781119316244.fmatter

Boyd, D., Ellison, N. (2007) Social network sites: Definition, history, and schola rship. *Journal of Computer-Mediated Communication, Volume 13, Issue 1.*

Braun, V., Clarke, V. (2012) Thematic Analysis. *Encyclopedia of Critical Psychology,* 1947-1952.

Cavanillas, J.M., Curry, E., Wahlster, W. (2016) *New Horizons for a Data-Driven Economy.* Springer Open, Haettu osoitteesta DOI:10.1007/978-3-319-21569-3

Chen, W., Quan-Haase, A. (2018) Big Data Ethics and Politics: Toward New Understandings. *Social Science Computer Review 2020, Vol 2020, Vol. 38 (1) 3-9.*

Cheong, F., Cheong, C. (2011) Social Media Data Mining: A Social Network Analysis Of Tweets During The 2010-2011 Australian Floods. *Pac ific Asia Conference on Information Systems (PACIS) 9.7.2011*

Corbett, P., Schroek, M., Shockley, R. (2013)
Analytics: The real-world use of big data in insurance. *IBM Global Business Services. October 2012.*

Das, T.K, Kumar, P.M. (2013) Big Data Analytics: A Framework for Unstruct ured Data Analysis. *School of Information Technology and Engineeri ng.* 153-156

Davenport, T. H., & Harris, J. G. (2007). The dark side of customer analytics. *Harvard Business Review*, *85*(5), 37–48.

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. *AIP Conference Proc eedings*, *1644*, 97–104.

Eaton, C., Deutch, T., Deroos, D, Lapis, G., Zikopoulos,P.(2012) *Understanding Big Data; Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw Hill

Eskola, J.& Suoranta, J. (1998). *Johdatus laadulliseen tutkimukseen.* Vastapaino

Euroopan parlamentin ja neuvoston asetus (EU) 2016/679, annettu 27 päivänä huhtikuuta 2016, luonnollisten henkilöiden suojelusta henkilötietojen käsittelyssä sekä näiden tietojen vapaasta liikkuvuudesta ja direktiivin 95/46/EY ku-moamisesta (yleinen tietosuoja-asetus). Euroopan unionin virallinen lehti 4.5.2016. https://eur-lex.europa.eu/legal-content/FI/TXT/HTML/?uri=CE-LEX:32016R0679&from=FI

Fang, K., Jiang, Y., Song, M. (2016) Customer profitability forecasting using Big Data analytics: A case study of the insurance industry. *Computers and industrial Engineering, Vol. 101.* 554-564.

Farzindar, A., Inkpen, D. (2015) Natural Language Processing for Social Media. *Synthesis Lectures on Human Language Technologies.*

Floridi, L, Taddeo, M. (2018) What is data ethics? *Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences, December 2016.*

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P., & Grimshaw, J. M. (2010). What is an adequate sample size?

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, metho ds, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

George, G., Haas, M. R., & Pentland, A. (2014). Big Data and Management. *Https://Doi.Org/10.5465/Amj.2014.4002*, *57*(2), 321–326. https://doi.org/10.5465/AMJ.2014.4002

Ghani, N., Hamid, S., Hashem, I., Ahmed, E. (2019) Social media big data anal ytics: A survey. *Computers in Human Behavior. Volume 101, Dec ember 2019,* 417-428

Goddard, M. (2017) The EU General Data Protection Regulation (GDPR): European regulation that has a globat impact. *International Journal of Market Research, Vol 59 (6).*

Goldsteen, A., Ezov, G., Shmelkin, R., Moffle, M., Farkas, A. (2021) Data Minimization for GDPR compliance in Machine Learning Models.

Gruschka, N., Mavroeidis, V., Vishi, K., Jensen, M. (2018) Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR. *2018 IEEE International Conference on Big Data*

Guest, G., Brunce, A., Johnson, L. (2006) How Many Interviews are Enough?: An Experiment with Data Saturation and Variability

Hand, D.J. (2018) Aspects of Data Ethics in a Changing World: Where Are We Now? *Big Data, Vol. 6, No. 3*

Hargittai, E. (2018) Potential Biases in Big Data: Omitted Voices on Social Media. *Vol 38, Issue 1, 2020.*

Herschel, R,. Miori, V.M. (2017) Ethics & Big Data. *Technology in Society. Volume 49, May 2017, Pages 31-36.*

Hirschberg, J., Hjalmarsson, A., Elhadad, N. (2010) "You're Sick as You Sound": Using Computational Approaches for Modeling Speakers State to Cauge Illness and Recovery,

Hirsjärvi, S., Hurme, H. (2000) Tutkimushaastattelu: Teemahaastattelun teoria ja käytäntö.

Hussain, K., Prieto, E. (2016) Big data in the finance and insurance sectors

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets: *Http://Dx.Doi.Org/10.1177/2053951716631130*, *3*(1). https://doi.org/10.1177/2053951716631130

Lycett, M. (2013). "Datafication": Making sense of (big) data in a complex world. *Eur opean Journal of Information Systems*, *22*(4), 381–386. https://doi.org/10.1057/ejis.2013.10

Malgieri, C. (2020) The concept of Fairness in the GDPR. A linguistic and contextual interpretation

Mason, M. (2010). Sample size and saturation in PhD studies using qualitative interviews.

Mehmood, A., Natgunanathan, I., Xiang, Y., Member, S., Hua, G., & Guo, S. (2016). *Protection of Big Data Privacy*. https://doi.org/10.1109/ACCESS.2016.2558446

Oussous, A., Benjelloun, F.Z., Lahcen, A., Belfkih, S. (2018) Big Data technologies: A survey

Perera, C., Ranjan, R., Wang, L., Khan, S., Zomaya, A. (2015) Big Data Privacy in the Internet of Things Era. IT Professional, Vol. 16, Issue 3.

Pfitzmann, A., Hansen, M. (2010) A terminology for talking about privacy by data minimization: Anonymity, Unlikability, Undetectability, Unobservability. Pseudonymity and Identity Management.

Pormeister, K. (2017) Genetic data and the research exemption: is the GDPR going too far? *International Data Privacy Law, 2017, Vol. 7, No. 2*

Richards, N.M. ., & King, J. H. (2014). Big Data Ethics. *Wake Forest Law Review*, *49*.https://heinonline.org/HOL/Page?handle=hein.journals/wflr49&id=405&div=&collection=

Richards, N.M., King, J.H. (2014) Big Data Ethics

Richterich, A. (2018) The Big Data Agenda: Data Ethics and Critical Data Studies

Rubin, V., Lukoianova, T. (2013). Veracity Roadmap: Is Big Data Objective, Truthful and Credible? Advances in Classification Research Online

Shim. J.P., Koh, J., Fister, S., Seo, H.Y. (2016) Phonetic analytics technology and big data: real world cases

Soria-Comas. J., Doming-Ferrer, J.(2015) Big Data Privacy: Challenges to Privacy Principles and Models.

Statista (2021) Global social networks ranked by number of users 2021.

Stieglitz, S., Dan-Xuan. L. (2013) Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining, 3, pp. 1277-1291*

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data

preparation. *International Journal of Information Management*, *39*, 156–168. https://doi.org/10.1016/J.IJINFOMGT.2017.12.002

Tiefenbacher, K., Olbrich, S. (2015) Increasing the level of customer orientation – A big data case study from insurane industry. ECIS 2015

Tufekci, Z. (2014) Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls.

Wattal, S., Schuff, S., Mandvivwalla, M., Williams, C.B. (2010) Web 2.0 and Polit ics: The 2008 U.S Presidential Election and E-politics Research Agenda. *MIS Quarterly Vol 34. No. 4 pp. 669-688*

Westin, A. (1970) Privacy and Freedom

Zarsky, T.Z. (2017) Incompatible: The GDPR in the Age of Big Data

Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and int elligence. *IEEE Intelligent Systems*, *25*(6), 13–16. https://doi.org/10.1109/MIS.2010.151

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, *58*(4), 479–493. https://doi.org/10.1002/ASI.20508

Zwitter, A. (2014) Big Data Ethics. *Big & Society July -December 2014:* 1-6

LIITE 1 ENSIMMÄINEN LIITE

Haastattelurunko

1. Esittely
    - Esittele itsesi ja kiitä haastateltavaa
    - Esittele tutkielma
    - Informoi haastateltavaa haastattelun luottamuksellisuudesta
    - Informoi haastateltavaa aineiston anonymisoinnista
    - Informoi haastateltavaa hänen oikeudesta olla vastaamatta kysymykseen
    - Informoi haastateltavaa hänen oikeudesta keskeyttää haastattelu
    - Pyydä suostumus haastattelun nauhoittamiseen

2. Kysymykset haastateltavan ammatillisesta taustasta ja nykytilanteesta
    - Minkälainen on ammatillinen historianne?
    - Mikä on nykyinen työnimikkeenne?
    - Kuinka pitkä ammatillinen taustanne on?
    - Kuvailisitteko, miten nykyinen toimenkuvanne liittyy data-analytiikkaan?

3. Kysymykset massadatasta vakuutusalalla
    - Yleisesti ottaen, miten massadataa hyödynnetään vakuutusalalla tänä päivänä?
    - Missä eri liiketoiminnan osa-alueissa massadataa voidaan hyödyntää?
    - Mitkä ovat massadatan tyypillisimpiä sovelluskohteita?
    - Mihin tarkoitukseen kerättyä ja analysoitua massadataa käytetään?
    - Mistä käytettävä massadata hankitaan?
    - Miten sosiaalisen median massadataa hyödynnetään?
    - Miten dataa säilytetään ja miten se hävitetään?
    - Mitkä ovat massadatan hyödyntämisen tulevaisuudennäkymät?

4. Kysymykset massadataan liittyvistä haasteista vakuutusalalla

- Millaisia haasteita massadatan hyödyntämiseen vakuutusalalla liittyy?
- Mitä ominaisuuksia organisaatiolta vaaditaan massadatan hyödyntämisen onnistumiseksi?
- Miten näet massadataan liitettyjen odotusten realisoitumisen vakuutusalalla?

5. Kysymykset liittyen informaation yksityisyyttä koskevan sääntelyn vaikutuksiin massadatan hyödyntämisessä vakuutusalalla
    - Miten datan ja informaation sekä niiden analysoinnin yksityisyyttä säädellään?
    - Yleisellä tasolla, miten informaation yksityisyyteen liittyvä sääntely vaikuttaa massadatan hyödyntämiseen vakuutusalalla?
    - Miten sääntely on vaikuttanut massadatan hyödyntämiseen organisaatiossanne?
    - Onko datan ja informaation yksityisyyteen liittyvä sääntely ajantasaista vakuutusalan nykytilaan nähden?
    - Miten GDPR otetaan huomioon organisaatiossanne?
    - Miten vakuutusyhtiöt ottavat huomioon dataetiikan?

6. Yhteenveto ja haastattelun päättäminen
    - Kiitä haastateltavaa hänen ajastansa
    - Informoi miten tutkimus etenee tästä
    - Kehoita olemaan matalalla kynnyksellä yhteydessä, jos tulee mitään kysyttävää