Jan Hokkanen

YKSITYISYYS JA YKSITYISYYDEN SUOJAAMINEN BIG DATAN AIKAKAUDELLA: TEKNOLOGINEN NÄKÖKULMA

JYVÄSKYLÄN YLIOPISTO
TIETOJENKÄSITTELYTIETEIDEN LAITOS
2015

TIIVISTELMÄ

Hokkanen, Jan
Yksityisyys ja yksityisyyden suojaaminen big datan aikakaudella: teknologinen näkökulma
Jyväskylä: Jyväskylän yliopisto, 2015, 54 s.
Tietojärjestelmätiede, kandidaatintutkielma
Ohjaaja(t): Halttunen Veikko

Elämme tiedon aikakautta – teknologian kehityksen myötä sekä digitaalisesta että analogisesta maailmasta kertyvien tietojen määrä on kasvanut eksponentiaalisesti viime vuosien aikana. Big data mahdollistaa monia koko yhteiskuntaa hyödyttäviä etuja aina kilpailukyvyn parantamisesta kehittyneempään terveydenhuoltoon. Massiivisen tietojen keräämisen ja analysoinnin on kuitenkin pelätty kaventavan yksityisyyden suojaa. Tässä kirjallisuuskatsaukseen pohjautuvassa tutkielmassa perehdyttiin yksityisyyttä, big dataa, massadata-analytiikkaa sekä yksityisyyden suojaamiseen käytettäviä teknisiä menetelmiä käsittelevään kirjallisuuteen. Tutkielman tarkoituksena oli selvittää millaisia yksityisyyteen kohdistuvia riskejä big datan hyödyntämiseen voi liittyä sekä arvioida teknisten menetelmien kykyä suojata yksityisyyttä. Tutkielmassa selvisi, että big datan hyödyntämiseen voi liittyä yksityisyyteen kohdistuvia riskejä. Riskit vaikuttaisit olevan seurausta hyvin laajamittaisesta tietojen keräämisestä sekä tehokkaasta prosessoinnista, minkä seurauksena datasta kyetään löytämään enemmän yksityisyyttä uhkaavaa informaatiota kuin useat ihmisistä olettavat tietoja luovuttaessaan. Läpinäkyvyyden puutteen havaittiin entisestään lisäävän tilanteeseen liittyvää ongelmallisuutta. Havaintojen pohjalta tehtiin johtopäätös siitä, että teknisten menetelmien ohella tarvitaan entistä enemmän julkisen vallan sääntelyä sekä organisatorisia toimenpiteitä yksityisyyden suojaamiseksi.

Asiasanat: anonymiteetti, big data, identifiointi, kryptografia, yksityisyys

ABSTRACT

Hokkanen, Jan
Privacy and privacy protection in the era of big data: technological perspective
Jyväskylä: University of Jyväskylä, 2015, 54 p.
Information systems science, bachelor's thesis
Supervisor(s): Halttunen Veikko

We live in an information age - the ubiquity of computing and electronic communication technologies has led to the exponential growth of data from both digital and analog sources. Big data drives also big benefits, from innovative products and services to new ways to treat diseases. The problem is that new capabilities to gather, analyze, process, and preserve vast quantities of data raise new concerns about the nature of privacy and the means by which individual privacy might be compromised. Purpose of this bachelor's thesis was to study big data, big data analytics, privacy and technologies used to protect privacy. Objective was to find out privacy issues of big data and analyze technologies capability to protect privacy. Study was carried out as review of literature. It seems that the challenges to privacy arise because technologies collect so much data (e.g., from sensors in everything from phones to thermostats) and analyze them so efficiently (e.g., through data mining and other kinds of analytics) that it is possible to learn far more than most people had anticipated or can anticipate given continuing progress, and the lack of transparency has made this problem even bigger. Challenges are also compounded by limitations on technologies used to protect privacy (especially traditional technologies such as de-identification). This is why thesis concludes that technology alone cannot protect privacy so we need organizational policies and legislation to protect privacy.

Keywords: anonymity, big data, cryptography, identification, privacy

KUVIOT

Kuvio 1: Systemaattisen kirjallisuuskatsauksen eri vaiheet (Brereton, Kitchenham, Budgen, Turner & Khalil, 2007, 2, suomennettu). 7
Kuvio 2: Altmanin yksityisyyden säätelymalli (Laukka, 2000, 6, suomennettu) 15
Kuvio 3: MapReduce-mallin mukaisen prosessin kulku (Salmi, 2015, 16). 21

TAULUKOT

Taulukko 1: Anonymisointitekniikoiden vahvuudet ja heikkoudet (Tietosuojaryhmä, 2014, 24)..... 40

SISÄLLYS

TIIVISTELMÄ
ABSTRACT
KUVIOT
TAULUKOT

1 JOHDANTO ... 6

2 YKSITYISYYS ... 9
2.1 Yksityisyyden alue ... 9
2.2 Yksityisyyden funktiot ... 12
2.3 Yksityisyyteen liittyvä problematiikka ... 13
2.4 Yksityisyys tämän tutkielman kannalta ... 14

3 BIG DATA JA YKSITYISYYS ... 17
3.1 Mitä on big data? ... 17
3.1.1 Big datan 3V-käsitemalli ... 18
3.1.2 Teknologiat ja tekniikat big datan takana ... 20
3.1.3 Massadata-analytiikka ... 22
3.2 Big datan hyödyntäminen ja yksityisyyteen kohdistuvat riskit ... 25
3.2.1 Esimerkkejä ja skenaarioita ... 25
3.2.2 Yksityisyyteen kohdistuvat riskitekijät ja niiden merkitys ... 29
3.3 Yhteenveto ... 32

4 YKSITYISYYDEN SUOJAAMINEN ... 34
4.1 Tietosuoja, tietoturva ja yksityisyys ... 34
4.2 Erilaiset suojausmenetelmät ... 35
4.2.1 Kryptografia ... 36
4.2.2 Anonymisointi ... 38
4.2.3 Ilmoitus ja hyväksyntä ... 41
4.3 Johtopäätökset ja pohdinta ... 42

5 YHTEENVETO ... 45

LÄHTEET ... 49

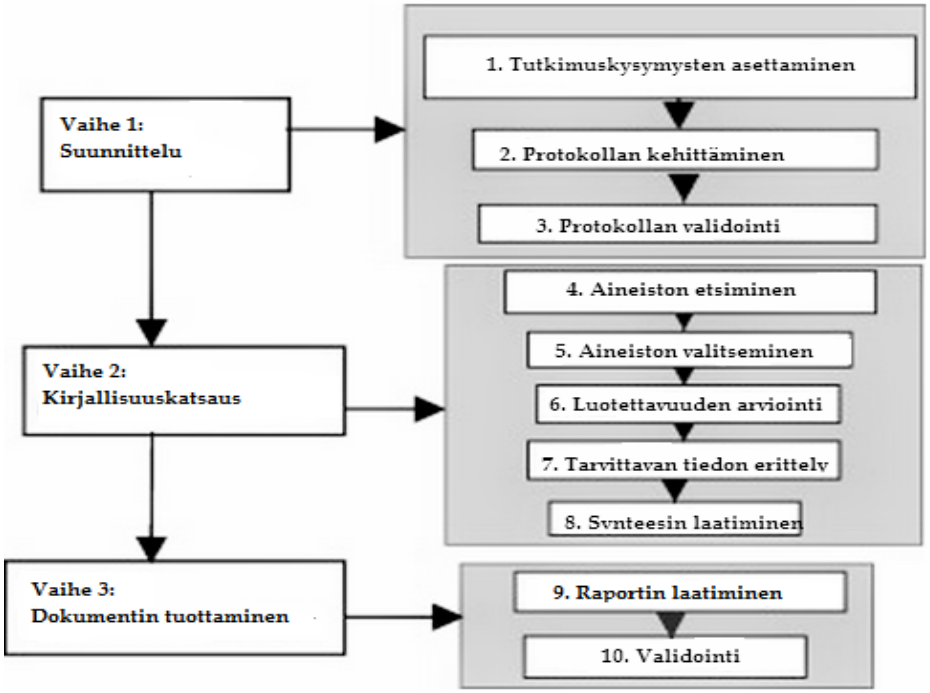
1 JOHDANTO

Elämme tiedon aikakautta – yhä useammat ja useammat laitteet sekä erilaiset sensorit keräävät mitä monipuolisempia tietoja käyttäjistään ja käyttöympäristöstään. Tuotamme itsestämme koko ajan tietoa useisiin järjestelmiin erilaisina klikkaus- ja lokitietoina. Yritykset keräävät kuluttajista valtavat määrät tietoa omiin rekistereihinsä ja tietoa kerääntyy myös useista yritysten sisäisistä järjestelmistä. Tietoja voidaan kerätä lisäksi sosiaalisen median kautta sekä monista muista eri lähteistä, mitkä eivät välttämättä ole edes suunniteltuja tiedon keräämistä varten (Franks, 2012).

Tietojen tehokas kerääminen ja hyödyntäminen ovat yhä tärkeämmässä roolissa myös koko yhteiskunnan hyvinvoinnin kannalta. Viime vuosien aikana onkin syntynyt tarve kehittää uusia menetelmiä ja teknologioita, mitkä kykenevät tehokkaammin prosessoimaan yhä suurempia, hajautuneempia sekä nopeaa tahtia kasvavia tietomassoja. Tällaisten datamassojen sekä niiden keräykseen, prosessointiin, analysointiin ja esittämiseen tarkoitettujen tekniikoiden yhteisiksi nimittäjiksi ovat hiljalleen muodostuneet big data ja massadata-analytiikka (engl. big data analytics) (Sagiroglu & Sinanc, 2013.). Tässä tutkielmassa käytetään juuri big datan ja massadata-analytiikan käsitteitä, mutta myös esimerkiksi iso datan ja data-analytiikan käsitteitä on käytetty samassa merkityksessä. Datalla taas viitataan tietoon, jolla ei itsessään välttämättä ole semanttista merkitystä tai informatiivista järjestystä. Datasta voidaan kuitenkin löytää informaatiota tai se voi myös olla järjestymätöntä kohinaa. Edelleen jos informaatiolle voidaan antaa semanttinen merkitys, niin on se myös tietoa.

Big data mahdollistaa monia koko yhteiskuntaa hyödyttäviä etuja aina kilpailukyvyn parantamisesta kehittyneempään terveydenhuoltoon, mutta massiivinen tietojen kerääminen ja analysointi ovat kuitenkin herättäneet huolestuneisuutta yksityisyyden suojan säilymisestä. Huolestuneisuutta ovat herättäneet erityisesti yksilöiden toimintaan liittyvien tietojen kerääminen, kuten aikeet kerätä paikkainformaatiota tai analysoida internetin selauskäyttäytymistä. Tällaisia hankkeita on usein myös vastustettu voimakkaasti (Kshetri, 2014.). Yksityisyyteen liittyvät huolet vaikeuttavat näin ollen big datan hyödyntämistä ja luovat epävarmuutta, mikä saattaa esimerkiksi näkyä kuluttajien halutto-

muutena siirtyä uusien teknologioiden tai palveluiden käyttäjiksi ja osaltaan hankaloittaa myös datan hyödyntämistä. Vuonna 2012 hyödynnettiinkin vain noin 3 % kaikesta kertyneestä datasta, vaikka huomattavasti suurempi osa (23 %) olisi ollut hyödynnettävissä (Gantz & Reinsel, 2012). Uhkana onkin se, että big datan mahdollistamat hyödyt jäävät ainakin osittain saavuttamatta jos yksityisyyteen liittyviä huolia ja niistä kumpuavia epävarmuustekijöitä ei selvitetä (Franks, 2012). Tästä huolimatta big datan vaikutusta yksityisyyteen on tutkittu verrattain vähän, joskin viimevuosien aikana aihepiiri on ollut tiedepiireissä kasvavan mielenkiinnon kohteena (Cuzzocrea, 2014). Valtaosassa aiemmista tutkimuksista ei myöskään Kshetrin (2014) mukaan ole juuri eritelty miten big data oikeastaan vaikuttaa yksityisyyteen, vaan on tyydytty vain toteamaan, että big datan ja massadata-analytiikan hyödyntäminen nostavat esille lukuisia eettisiä ongelmia ja kysymyksiä. Aihepiirin tarkemmalle tutkimiselle on näin ollen sekä eettisiä että taloudellisia syitä. Tässä kirjallisuuskatsaukseen pohjautuvassa tutkielmassa perehdytään yksityisyyteen, big dataan ja tähän läheisesti liittyvään massadata-analytiikkaan sekä yksityisyyden suojaamisessa hyödynnettäviin teknisiin menetelmiin. Kuviossa 1 havainnollistetaan systemaattiseen kirjallisuuskatsaukseen liittyviä vaiheita sekä osa-alueita, minkä mukainen lähestymistapa toimii myös tämän tutkielman perustana.



Kuvio 1: Systemaattisen kirjallisuuskatsauksen eri vaiheet (Brereton, Kitchenham, Budgen, Turner & Khalil, 2007, 2, suomennettu).

Valtaosa kuviossa 1. esitetyistä vaiheista on itsensä selittäviä, mutta suunnitteluun liittyviä kohtia 2. sekä 3. lienee tarpeellista avata hieman enemmän. Nämä vaiheet liittyvät Brereton ym. (2007) mukaan kirjallisuuskatsauksen toteuttamiseen sekä aineiston keräämisessä noudatettaviin menettelyihin ja ovat näin ollen keskeisessä asemassa myös harhaisuuden välttämisen kannalta. Täs-

sä kirjallisuuskatsaukseen pohjautuvassa tutkielmassa lähdemateriaalin hakeminen toteutetaan hyödyntämällä erilaisia hakukoneita sekä tietokantoja. Hakuprosessi pohjautuu tutkielman aihepiiriin liittyviin asiasanoihin, minkä jälkeen julkaisujen hyödyntämiskelpoisuus pyritään selvittämään arvioimalla julkaisun oleellisuutta tutkielman tavoitteiden kannalta. Lähdemateriaalin luotettavuus pyritään varmistamaan vertailemalla julkaisuissa esitettyjä tuloksia vastaavanlaisiin tutkimuksiin sekä muihin aihepiiriä koskeviin julkaisuihin. Lähdemateriaalien luotettavuuden arviointiin käytetään lisäksi seuraavia mittareita: 1) julkaisussa käytetyt viittaukset (lukumäärä ja tunnettavuus), 2) tekijöiden ja/tai julkaisijan tunnettavuus ja arvostettavuus, 3) julkaisuihin tehtyjen viittausten lukumäärä sekä 4) yleinen vaikutelma luotettavuudesta.

Tutkielman tavoitteena on arvioida millaisia yksityisyyteen kohdistuvia riskejä big datan ja tähän läheisesti liittyvän massadata-analytiikan hyödyntämiseen voi liittyä sekä pohtia näiden riskien merkitystä yksityisyyden kokemisen kannalta. Tutkielman tavoitteena on myös pohtia teknisten menetelmien mahdollisuuksia suojata yksityisyyttä sekä analysoida menetelmien käyttöön liittyviä haasteita big datan aikaudella. Tutkimusongelma voidaan edelleen tiivistää seuraaviin tutkimuskysymyksiin:

- Millaisia yksityisyyteen kohdistuvia riskejä big datan hyödyntämiseen voi liittyä ja kuinka nämä riskit voivat vaikuttaa yksityisyyden kokemiseen?
- Kykenevätkö tekniset menetelmät suojaamaan yksityisyyden big datan aikakaudella ja millaisia haasteita teknisten menetelmien käyttöön voi liittyä?

Tutkielman luvussa kaksi tarkastellaan yksityisyyteen liittyvää monitulkintaista semantiikkaa, yksityisyyden merkitystä, ulottuvuuksia sekä yksityisyyteen liittyvää problematiikkaa. Luvussa kolme siirrytään käsittelemään big dataa ja tähän läheisesti liittyvää massadata-analytiikkaa sekä näihin liittyviä teknologioita ja tekniikoita. Kolmannen luvun tarkoituksena on muodostaa havainnollistava käsitys tämän tutkielman kannalta keskeisistä big dataan ja massadata-analytiikkaan liittyvistä teknologioista ja tekniikoista sekä etenkin arvioida näiden hyödyntämiseen liittyviä yksityisyyteen kohdistuvia riskejä teknologisen näkökulman pohjalta. Tutkielman neljännessä luvussa siirrytään tarkastelemaan yksityisyyden suojaamisessa hyödynnettäviä teknisiä menetelmiä sekä pohtimaan menetelmien kykyä suojata yksityisyyttä ja arvioimaan näiden käyttöön liittyviä haasteita big datan aikaudella. Tarkastelun kohteeksi otettavien menetelmien valinta perustuu menetelmän tunnettavuuteen (kuinka laajalti menetelmää käytetään), tutkielmassa käytettyyn aineistoon (kuinka tärkeäksi menetelmä on kirjallisuudessa koettu) sekä lopulta tutkielmalle asetettuihin tavoitteisiin ja tutkielman rajaukseen. Tutkielman päättää yhteenveto, missä tarkastellaan tutkimusongelmaan saatuja vastauksia sekä muita tutkielman aikana esiin tulleita havaintoja. Yhteenvedossa arvioidaan lisäksi tutkielman merkitystä ja saatujen tulosten yleistettävyyttä sekä esitetään mahdollisia jatkotutkimisaiheita.

2 YKSITYISYYS

Termi ”yksityisyys” on usein esillä niin puhekielessä kuin erilaisissa julkaisuissa sekä poliittisissa ja lainsäädännöllisissä keskusteluissa. Yksityisyydelle ei kuitenkaan ole olemassa yhtä eksaktia määritelmää ja sellaista on myös jossain määrin mahdotonta antaa (Tavani, 1999). Yksityisyyttä ja yksityisyyden käsitteen alla kulkevia ilmiöitä on lisäksi ajan mittaan tutkittu hyvinkin erilaisista näkökulmista mikä on saanut aikaan sen, että tutkimuksia on leimannut erilaisten määritelmien käyttö. Sekaannusta on omalta osaltaan lisännyt myös se, että eri kulttuureissa yksityisyydellä voi olla erilaisia merkityksiä ja joistakin kielistä jopa puuttuu yksityisyyttä tarkoittava sana (Newell, 1998; Traver, 1984). Tässä luvussa tarkastellaan lähinnä filosofisen ja antropologisen kirjallisuuden pohjalta yksityisyyden aluetta, yksityisyyden merkitystä sekä yksityisyyteen liittyvää problematiikkaa. Aihepiirin käsittelyä painotetaan kuitenkin mahdollisimman paljon tutkielman teknologisen näkökulman mukaisesti ja luvussa sivutaan myös hieman lainsäädännöllistä näkökulmaa yksityisyyteen. Luvun päätteeksi määritellään tarkemmin kuinka aihepiiriä aiotaan lähestyä tässä tutkielmassa ja esitellään Altmanin (1975) yksityisyyden säätelyn prosessimalli sekä kuvataan mallin merkitystä tämän tutkielman tavoitteiden kannalta.

2.1 Yksityisyyden alue

Periaate julkisen toiminnan piiristä ja yksityiselämästä sekä niiden eriävyydestä on ikivanha, mikä käy ilmi myös Aristoteleen kirjoituksista. Aristoteles nimittäin erotti toisistaan julkisen poliittisen toiminnan (polis) sekä perheen sisäisen elämän (oikos). Käsite yksityisyydestä on myös maailmanlaajuisesti ymmärretty (Harris, Werner, Brown & Ingebritsen, 1995). Ajan mittaan yksityisyyden piiriin on koettu kuuluvan erilaisia asioita ja käsite siitä mitä yksityisyys on, muuttuu yhä yhteiskunnan sekä teknologian kehityksen mukana. Näin ollen yksityisyyden eksakti määritteleminen ei ole edes tarkoituksenmukaista tämän tutkielman kannalta. Edelleen voidaan kuitenkin havaita, että yksityisyyden käsite on hyvin moniulotteinen. Saarenpään (2009) mukaan voidaan erottaa

ainakin seuraavat kymmenen välittömästi itsemääräämisoikeuteemme kuuluvaa osa-aluetta: 1. *fyysinen yksityisyys,* 2. *alueellinen yksityisyys*, 3. *sosiaalinen yksityisyys*, 4. *mediayksityisyys*, 5. *anonymiteetti*, 6. *yksityisyys henkilötietojen käsittelyssä*, 7. *tiedollinen omistusoikeus*, 8. *oikeus tulla arvioiduksi oikeassa valossa,* 9. *potilasyksityisyys* sekä 10. *viestintäyksityisyys*.

Kirjallisuutta tarkastelemalla voidaan havaita monia muitakin ja osin edellä esitetystä eroavia yksityisyyden alueen jaotteluja. Burgoon ym. (1989) esimerkiksi jaottelevat yksityisyyden neljään eri alueeseen, joita he nimittävät fyysiseksi, interaktionaaliseksi, psyykkiseksi sekä informaation yksityisyydeksi. Hieman toisenlaisen yksityisyyden alueen jaottelun ovat esittäneet Westin (1967) ja Pedersen (1979), jotka erottelevat seuraavat neljä aluetta: yksinolo, intimiteetti, anonymiteetti ja pidättyväisyys (reserve). Pedersen (1979) tosin lisäsi vielä faktorianalyysien perusteella viidennen alueen ”eristyneisyys” (isolation) sekä jakoi intimiteetin intimiteettiin ystävien ja intimiteettiin perheen kanssa. Huomionarvoista edellä esitetyissä yksityisyyden alueen jaotteluissa on se, että erilaisia jaotteluja on useita ja voidaan päätyä toisistaan eroaviin jaotteluihin riippuen valitusta lähestymistavasta.

Tämän tutkielman kannalta olennaisimmat ulottuvuudet ovat *alueellinen yksityisyys, sosiaalinen yksityisyys, mediayksityisyys, anonymiteetti, yksityisyys henkilötietojen käsittelyssä, tiedollinen omistusoikeus* sekä *oikeus tulla arvioiduksi oikeassa valossa*, joita tarkastellaan seuraavaksi hieman tarkemmin. Tarkastelu pohjautuu pitkälti Saarenpään (2009) tulkintaan kustakin yksityisyyden alueesta, joten käsittelyn painotus on näin ollen hieman enemmän Suomen lainsäädännössä. Vastaavanlaisia yksityisyyden alueiden käsityksiä on kuitenkin havaittavissa myös muuta kirjallisuutta tarkastelemalla, joskin erilaisten lähestymistapojen vuoksi alueiden jaottelu sekä niille annettava merkitys saattavat poiketa seuraavaksi esitettävästä. Tavoitteena on havainnollistaa yksityisyyden käsitteen monipuolisuutta sekä avata lyhyesti kuhunkin yksityisyyden alueeseen liittyvää semantiikkaa. Näin ollen esitettävä jaottelu ei ole tarkoitettu tyhjentäväksi yksityisyyden osa-alueiden kuvaukseksi ja tulee muistaa, että käsitteenä yksityisyys elää jatkuvasti teknologian kehityksen sekä yhteiskunnan mukana.

Alueellinen yksityisyys

Alueellinen yksityisyys tarkoittaa oikeutta olla fyysisesti yksin kotirauhan piirissä ja erilaisen valvonnan sekä häirinnän ulottumattomissa (Saarenpää, 2009). Saarenpään (2009) mukaan Suomessa ei ole erityislakia videovalvonnasta, mutta salakatselua ja salakuuntelua koskevien säännösten lisäksi sitä koskee kuitenkin myös henkilötietolaki, sillä sekä kuva että ääni lukeutuvat tuossa laissa tarkoitettuihin henkilötietoihin. Kolmannessa luvussa osoitetaan, että big datan hyödyntämiseen voi liittyä myös alueellista yksityisyyttä uhkaavia tekijöitä, mitkä ovat pitkälti seurausta analogisesta maailmasta tallennettavasta datasta ja sen hyödyntämisestä.

Sosiaalinen yksityisyys

Sosiaalinen yksityisyys koskee oikeuttamme pitää ihmissuhteemme, elämäntapamme, harrastuksemme ja vakaumuksemme yksin tiedossamme valvonnan ja

tiedonvälityksen ulottumattomissa. Näin ollen kysymys on niistä rajapinnoista, joita meillä on muihin yksilöihin tai yhteisöihin sekä yhteiskuntaan (Saarenpää, 2009). Kolmannessa luvussa tullaan osoittamaan, että big datan hyödyntämiseen voi liittyä sosiaalista yksityisyyttä uhkaavia tekijöitä, sillä yksi merkittävimmistä big datan tiedonlähteistä ovat sosiaalisen median palveluista kertyvät tiedot, joiden perusteella voidaan löytää sosiaalista yksityisyyttä loukkaavaa informaatiota.

Mediayksityisyys

Jokaisella on lähtökohtaisesti oikeus pysyä tuntemattomana yhteiskunnassa suhteessa joukkoviestintään eli oikeus pidättäytyä julkisuudesta. Yksityiselämäämme koskevien tietojen julkistaminen, joukkoviestinnässä ja myös muutoin, on pitkälti kiellettyä (Saarenpää, 2009.). Mediayksityisyys onkin ollut erityisesti yhdysvaltalaisen yksityisyyskeskustelun varhaisimpia kysymyksiä ja yksittäisenä käsitteenä ″privacy″ juontaa osin juurensa Warrenin ja Brandeisin (1890) ″The Right to Privacy″ artikkeliin, mikä käsitteli sananvapauden ja yksityiselämän välisiä jännitteitä. Mediayksityisyys on yhä tärkeämpää, sillä big data mahdollistaa entistä mittavamman tietojen keräämisen ja yhdistelyn, jonka myötä esille saattaa tulla entistä enemmän yksityisyyden suojaa uhkaavia tietoja. Näin ollen big data menetelmiä hyödyntämällä saatavien tietojen julkinen käsittely saattaisi olla aiempaa haitallisempaa yksilöille.

Anonymiteetti

Anonymiteetti voidaan nähdä oikeutenamme toimia ja vaikuttaa tuntemattomana yhteiskunnassa suhteessa julkiseen valtaan. Näin ollen sen tulisi siis ensi sijassa olla yksilön oikeus yhteiskuntaa eikä toista yksilöä kohtaan (Saarenpää, 2009.). Anonymiteetti on mielenkiintoisessa asemassa tämän tutkielman kannalta, sillä myöhemmin luvuissa 3. ja 4. tullaan osoittamaan big data menetelmien mahdollistavan joissain tapauksissa tietojen kohdentamisen niin, että tiedot pystytään jälleen yhdistämään tiettyyn yksilöön tai ryhmään. Tällainen uudelleentunnistaminen vaarantaa ilmiselvästi yksilöiden anonymiteetin ja asettaa uudenlaisia haasteita yksityisyyden suojaamiselle. Kolmannessa luvussa tullaan lisäksi osoittamaan, että anonymiteetti voi vaarantua myös analogisesta maailmasta kerättävien tietojen vuoksi.

Yksityisyys henkilötietojen käsittelyssä

Tietosuojalainsäädännön avulla toteutettava henkilötietojen suoja on jo vuosia ollut informaatioyhteiskunnalle tyypillinen, pitkälti teknologian hyödyntämismahdollisuuksiin liittyvä, yksityisyyden osa-alue. Suomen tavoin monissa muissa maissa tietosuojalainsäädäntö koskee muutakin kuin automaattisen tietojenkäsittelyn avulla tapahtuvaa henkilötietojen käsittelyä, mutta juuri tietotekniikan kehityksen mahdollistamat henkilötietojen käsittelyyn liittyvät menetelmät vauhdittavat tietosuojalainsäädännön kansainvälistä kehitystä. Viime kädessä Euroopan vuoden 1995 henkilötietodirektiivi on merkittävällä tavalla yhdistänyt eri maiden tietosuojalainsäädäntöjä. Direktiivin pyrkimyksenä on ollut rajoittaa sekä ohjata henkilötietojen käsittelyä (Saarenpää, 2009.). Myö-

hemmin luvuissa 3. ja 4. nähdään, että big datan hyödyntäminen saattaa asettaa uudenlaisia haasteita myös henkilötietojen käsittelylle.

Tiedollinen omistusoikeus
Tiedollinen omistusoikeus on jokaista koskeva oikeus ja takaa oikeuden omaan nimeen, kuvaan ja hahmoon sekä erilaiseen kaupalliseen hyödyntämiseen (Saarenpää, 2009). Tiedollinen omistusoikeus onkin Saarenpään (2009) mukaan informaatioyhteiskunnassa merkittävä yksityisyyteen ja siitä luopumiseen liittyvä taloudellinen oikeus. Tutkielman kannalta tämä on merkittävää siinä mielessä, että luovuttaessamme erinäisistä syistä tietojamme muiden käyttöön eivät ne siirry luovutuksensaajan rajoittamattomaan käyttöön. Näin ollen luovuttaessamme tietoa esimerkiksi palveluiden käytön yhteydessä ei luovutettujen tietojen omistusoikeus siirry palveluntarjoajan rajoittamattomaan käyttöön. Kolmannessa luvussa osoitetaan, että tilanteeseen saattaa liittyä tietyissä tilanteissa ongelmia, mistä kenties merkittävimpänä esimerkkinä ovat tietojen poistamisen vaikeutuminen sekä läpinäkyvyyden puute.

Oikeus tulla arvioiduksi oikeassa valossa
Tämä Suomessa hieman tuntemattomampi yksityisyyden osa-alue on Yhdysvalloissa saanut suuremman huomion ja siinä on Saarenpään (2009) mukaan kyse niin kunnian kuin henkilötietojen suojaamiseen sekä medioiden toimintaan liittyvistä kysymyksistä. Kyse on siis viimekädessä yksilön kunnioittamisesta häneen liittyvää informaatiota eri tavoin käsiteltäessä sekä hänestä viestittäessä (Saarenpää, 2009). Tämä yksityisyyden osa-alue on erityisen mielenkiintoinen tutkielman kannalta, sillä big datalle on tyypillistä tietojen automatisoitu kerääminen erilaisista lähteistä, mistä saattaa seurata suurta vaihtelevuutta myös tietojen laadun suhteen. Näin ollen tämän yksityisyyden alueen suojaaminen saattaa olla erityisen haastavaa, mikä tulee esille myös kappaleessa 3.2 tarkastellessa lähemmin yksityisyyteen kohdistuvia riskitekijöitä.

2.2 Yksityisyyden funktiot

Tähän mennessä on käynyt ilmi, että yksityisyys on maailmanlaajuisesti ymmärretty ja hyvin moniulotteinen käsite, jonka merkitys vaihtelee eri kulttuureissa sekä muuttuu alati yhteiskunnan ja teknologian kehityksen mukana. Tässä kappaleessa tarkastellaan lähemmin yksityisyyden merkitystä yksilöiden kannalta.

Kenties tärkein yksityisyyden funktio on suojan tarjoaminen, mikä käy ilmi myös aihepiiriä käsittelevässä kirjallisuudessa. Jourard (1966) esimerkiksi kokee yksityisyyden keskeiseksi ominaisuudeksi suojan tarjoamisen erilaisia ulkoisia uhkia vastaan. Samoilla linjoilla on myös Berardo (1975), joka pitää yksityisyyden säätelyn tavoitteena Newellin (1995) mukaan reagointiajan ja eräänlaisen puskurin tarjoamista sosiaalisten normien hahmottamiseksi. Yksityisyys voidaan nähdä myös eräänlaisena haavoittuvuuden säätelynä, sillä ulkopuoliset tahot saavat luovuttamamme informaation mukana valtaa, mikä tekee meidät

haavoittuviksi näiden tahojen edessä eli yksityisyyden ylläpitäminen toimii ikään kuin suojamuurina eri tahoja vastaan (Newell, 1995). Tämä on huomionarvoinen seikka myös tämän tutkielman kannalta, sillä big datan ja massadataanalytiikan menetelmiä hyödyntämällä on mahdollista yhdistää valtavat määrät erilaisia tietoja. Tietoja yhdistelemällä taas saatetaan löytää uudenlaista informaatiota, mikä saattaa kaventaa yksityisyyden mahdollistamaa suojamuuria. Kolmannessa luvussa tullaan lisäksi osoittamaan, että datan keräämiseen ja käyttöön liittyvä läpinäkyvyyden puute hankaloittaa tilannetta entisestään.

Toinen tärkeä yksityisyyden funktio on toimia vapaiden valintojen mahdollistajana. Tämän funktion merkitystä korostaa esimerkiksi Negley (1966), joka toteaa yksityisyyden mahdollistavan vapaat valinnat sekä moraalin. Myös Saarenpään (2009) mukaan yksityisyys on tärkeä tekijä vapaan valinnan mahdollistajana. Yksityisyyden merkitystä vapaan valinnan kannalta korostaa esimerkiksi anonymiteettiin perustuva äänestäminen – ilman anonymiteettia erilaiset tekijät, kuten ryhmän paine saattaisivat vaikuttaa ehdokkaan valintaan. Tällöin valinta ei välttämättä olisi täysin autonomisesti tehty. Yksityisyyteen liittyy läheisesti myös itsearvioinnin mahdollisuus sekä rajoitettu ja suojattu kommunikaatio (Westin, 1967).

Ilman yksityisyyttä meillä ei siis esimerkiksi olisi mahdollisuutta suojautua ulkopuolisten tahojen vaikutukselta emmekä pystyisi välttämättä toimimaan autonomisesti emmekä rajoittamaan tai suojaamaan kommunikaatiotamme. Kuitenkin myös monia muita funktioita; merkityksiä on havaittavissa kirjallisuutta tarkastelemalla. Pedersen (1997) on esimerkiksi havainnut kirjallisuudesta jopa 20 eri ehdotusta yksityisyyden funktioiksi. Newell (1995) toteaakin, että yksityisyyden merkityksestä ei ole päästy täyteen yhteisymmärrykseen ja, että yksityisyydelle voidaan löytää useita eri funktioita. Tässä tutkielmassa ei kuitenkaan rajauksen vuoksi voida tarkemmin pohtia yksityisyyden merkitystä, vaan aihepiirin käsittelyssä painotetaan enemmän suojan tarjoamisen sekä autonomian funktioita, mikä tukee myös tutkielman teknologista näkökulmaa aihepiirin tarkasteluun.

2.3 Yksityisyyteen liittyvä problematiikka

Yksityisyyden monitulkintaisen semantiikka on jo yksinään merkittävä ongelmakohta mikä yksityisyyteen liittyy, sillä se tekee hankalaksi yksityisyyden arvottamisen. Aiemmissa tutkimuksissa on myös osoitettu, että jokainen kokee yksityisyyden suojan eri tavalla (Brown & Muchira, 2004). Tämä taas asettaa omalta taholtaan haasteita yksityisyyden suojaamiselle, sillä ei tiedetä tarkkaan mitä oikeastaan tulisi suojella. Monitulkintaisen semantiikan vuoksi ei myöskään ole helppoa arvioida yksityisyyden todellisia vaikutuksia yksilöön, muihin yksilöihin tai yhteiskuntaan ennen kuin yksityisyyttä kavennetaan tai se kaventuu riittävästi. Tästä seuraa se, että yksityisyyteen liittyvien positiivisten ja/tai negatiivisten seurauksien selvittäminen on usein haastavaa. Thomson (1975) toteaakin, että yksityisyys on korvattavissa muilla oikeuksilla, kuten omistusoikeudella ja oikeudella fyysiseen turvallisuuteen. Näin ollen yksityi-

syyden suojan käsitteestä voitaisiin laillisesta näkökulmasta katsoen jopa luopua korvaamalla se muilla oikeuksilla ja/tai säännöksillä. Tämä saattaisi myös yksinkertaistaa lainsäädäntötyötä sekä selkeyttää yksityisyyden suojaan liittyvää epämääräisyyttä ja näin ollen edesauttaa myös yksityisyyden suojaamista.

Toinen merkittävä yksityisyyteen liittyvä ongelmakohta on se, että yksityisyyden varjoon voidaan pyrkiä kätkemään myös vahingollista toimintaa. Tällaisesta vahingollisen toiminnan kätkemisestä esimerkkinä ovat kotona tapahtuvat pahoinpitelyt (Allen, 1988). Tämä on hyvin pulmallinen seikka yksityisyyden arvottamisen kannalta, sillä yksilön edut eivät saisi mahdollistaa toisten yksilöiden tai yhteiskunnan vahingoittamista. Yksilöiden oikeudet tulisikin tasapainottaa muiden yksilöiden sekä yleisen hyvän kanssa (Kelly & Rowland, 2000). Kuitenkin toisaalta yksityisyys, kuten aiemmin kävi ilmi, tarjoaa monenlaista suojaa yksilöille sekä mahdollistaa myös autonomisen toiminnan. Näin ollen yksityisyyden tapauksessa on hyvin haastavaa löytää oikea tasapaino yksilön oikeuksien ja yleisen hyvän kanssa.

Vahingollisen toiminnan kätkemisen mahdollisuus ja pyrkimys tilanteen ratkaisemiseksi ovat mielenkiintoisia aihepiirejä myös tämän tutkielman kannalta, sillä nykyaikaiset datan keräämiseen ja analysointiin liittyvät menetelmät voisivat mahdollistaa sen, että tapahtuvasta vahingollisesta toiminnasta saataisiin jonkinlainen ennakkovaroitus. Näin ollen voitaisiin parhaimmillaan jopa estää pahantekijän aikeet tai lieventää tapahtuvaa vahinkoa. Samalla kuitenkin voidaan päätyä rankaisemaan yksilöä, vaikka tämä ei olisi vielä tehnyt mitään rikollista. Tällainen menettely edellyttää lisäksi yksilöiden laajamittaista tarkkailua ja analyysien perusteella saatavien tietojen luotettavuus voi, kuten myöhemmin osoitetaan, vaihdella erilaisista syistä. Tällainen menettely herättääkin useita eettisiä kysymyksiä sekä asettaa mahdollisesti yksityisyyteen kohdistuvia riskejä, joita pohditaan lähemmin tutkielman kolmannessa luvussa.

2.4 Yksityisyys tämän tutkielman kannalta

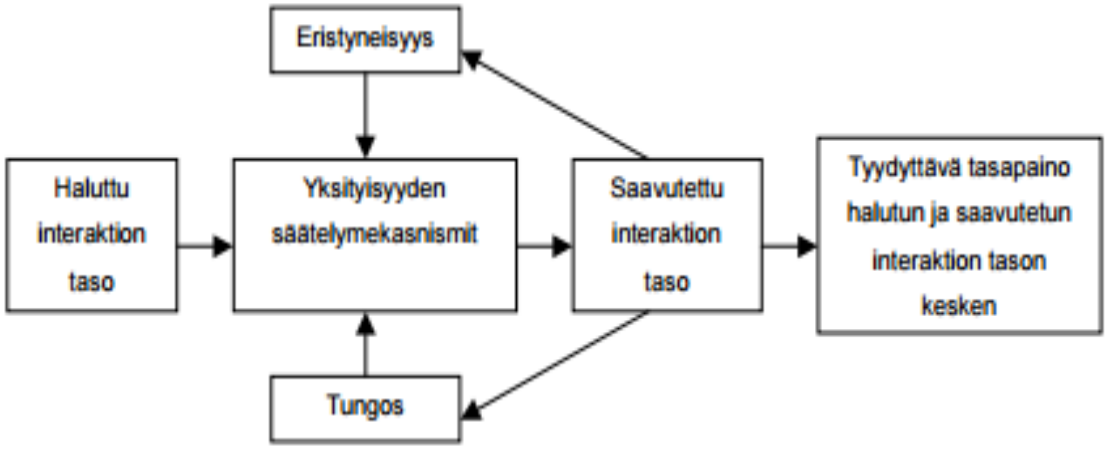
Kappaleessa 2.1 kävi ilmi, että yksityisyys on maailmanlaajuisesti ymmärretty sekä hyvin moniulotteinen käsite, minkä merkitys vaihtelee eri kulttuureissa sekä muuttuu alati yhteiskunnan ja teknologian kehityksen mukana. Kappaleessa 2.2 tarkasteltiin yksityisyyden merkitystä ja tällöin ilmeni, että yksityisyys kytkeytyy monin tavoin yksilöiden elämään sekä yksilöiden ja muiden tahojen väliseen kommunikointiin. Yksityisyyteen havaittiin myöhemmin liittyvän myös pulmallisia seikkoja, joita käsiteltiin tarkemmin kappaleessa 2.3. Tässä vaiheessa lieneekin siis ilmennyt, että yksityisyys on erittäin laajalle ulottuva käsite, jonka täsmällinen määrittely on hyvin haastavaa. Myös yksityisyyden todellisen merkityksen sekä vaikutusten selvittäminen voi tehtyjen havaintojen perusteella osoittautua haastavaksi tehtäväksi.

Tässä tutkielmassa keskitytään käsittelemään enemmän tietosuojan sekä tietoturvan näkökulmaa yksityisyyteen. Näin ollen informaation yksityisyys korostuu muita yksityisyyden alueita enemmän. Tutkielmassa ei tästä johtuen tarkastella lähemmin esimerkiksi kulttuurisia tai yksilöllisiä eroja yksityisyyden

kokemisen kannalta. Toisena lähtökohtana yksityisyyden tarkasteluun pidetään yksityisyyden suojaa, minkä Berman ja Mulligan (1998) näkevät odotuksena anonymiteetista, oikeudenmukaisuudesta, omien tietojen kontrolloinnista sekä luottamuksellisuudesta. Tutkielman yhtenä tavoitteena on myös, kuten johdannossa mainittiin, havainnollistaa big datan hyödyntämiseen liittyvien yksityisyyteen kohdistuvien riskien vaikutusta yksityisyyden kokemiseen. Tämän tavoitteen kannalta on tärkeää ymmärtää yksityisyyden säätelyyn liittyvää mekaniikkaa sekä yksityisyyden tason säätelyn merkitys. Tässä tutkielmassa yksityisyyden tason säätelyä lähestytään kontrolliteorian tai tarkemmin ottaen Altmanin (1975) yksityisyyden säätelyn prosessimallin pohjalta, mitä käsitellään seuraavaksi hieman lähemmin.

Altmanin yksityisyyden säätelyn prosessimalli

Yksityisyystutkimuksen varsin pirstoutuneella kentällä yksi vaikutusvaltaisimmista yksityisyyden prosessikuvauksista on jo pitkään ollut Altmanin (1975) yleinen viitekehys yksityisyyden säätelylle, joka syntyi Laukan (2000,6) mukaan yrityksenä yhdistää henkilökohtaisen alueen, tungoskäyttäytymisen sekä yksityisyyden tutkimuksen erillään olleet alueet. Altmanin (1975) teoria on lähemmin katsoen sovellus yleisestä kontrolliteoriasta, jonka keskeiset piirteet perustuvat toiminnan ohjaamiseen eräänlaisen palautekehän avulla. Ohjaaminen tapahtuu tarkemmin ottaen vertaamalla tavoitetilaa saavutettuun tilaan ja pyrkimyksenä on näiden kahden tilan välisen eron pienentäminen (Carver & Scheier, 1982). Kun tätä teoriaa sovelletaan ihmisten toimintaan, voidaan kontrolliprosessissa ajatella olevan kognitiivinen ja emotionaalinen puoli (Laukka, 2000). Kognitiivinen osa koostuu erilaisista sisäisistä tavoitteista, nykytilan analyysista sekä näiden tilojen vertailusta. Emotionaalinen komponentti taas puolestaan koostuu tarkemmin ottaen eroista tavoitetilan sekä nykytilanteen välillä. Emotionaalinen komponentti toimii myös eräänlaisena säätelytoiminnan motivaattorina (Carver & Scheier, 1982.). Kuvio 2 havainnollistaa Altmanin (1975) yksityisyyden säätelyn prosessimallia.



Kuvio 2: Altmanin yksityisyyden säätelymalli (Laukka, 2000, 6, suomennettu)

Yksilöllä tai ryhmällä on tietyllä hetkellä tietty halutun interaktion taso, joka muuttuu tilannetekijöiden ja henkilöiden välisten suhteiden muuttuessa (Altman, 1975). Interaktiolla tarkoitetaan tässä Laukan (2000) mukaan tarkemmin ottaen toiselle henkilölle tai ryhmälle annetun syötteen ja palautteen määrää. Laukan (2000, 7) mukaan Altman (1975) esittää edelleen, että haluttu interaktion taso pyritään saavuttamaan erilaisten yksityisyydensäätelymekanismien, kuten kognitiivisten mekanismien, behaviorististen mekanismien, verbaalisten mekanismien tai non-verbaalisten mekanismien, avulla. Tässä tutkielmassa ei kuitenkaan rajauksen vuoksi pystytä käsittelemään mekanismeja tarkemmin, vaan olennaista on se, että säätelymekanismeja käytetään kunnes haluttu interaktion taso vastaa haluttua tai kunnes yksilö on halukas muuttamaan tavoiteltua interaktion tasoa. Epätyydyttävä interaktion taso taas koetaan tungoksena tai eristyneisyytenä (ks. kuvio 2).

Kontrollintunteen on myös osoitettu olevan keskeinen tekijä yksityisyyden kokemisen kannalta, mikä puoltaa kontrollimallin käyttöä yksityisyyden säätelyn kuvaamiseksi (Laukka, 2000). Yleisesti ottaen yksityisyyden tason säätelyn onnistumiseksi on tärkeää, että yksilöillä on mahdollisuus valita haluamansa yksityisyyden (interaktion) taso ja, että heillä on käytettävissä erilaisia mekanismeja (kuten edellä mainitut mekanismit), joiden avulla yksityisyyden tasoa voidaan säädellä. Lisäksi yksilöiden tulee olla tietoisia halutun yksityisyyden tason sekä nykyisen yksityisyyden tason välisistä eroista eli heillä tulee olla mahdollisuus seurata vallitsevaa yksityisyyden tasoa sekä verrata tätä haluamaansa yksityisyyden tasoon (Laukka, 2000).

Altmanin (1975) teoria on kohdannut myös kritiikkiä käsitteiden huonon määrittelyn sekä liian yleisen käsittelytasonsa vuoksi, mutta se tarjoaa mahdollisuuden sitoa yhteen erilaisia tutkimustuloksia sekä tarkastella yksityisyyden säätelyn kannalta keskeistä kontrollin käsitettä. Näin ollen sitä pystytään käyttämään hypoteettisena mallina myös informaation yksityisyyden säätelyn kuvaamisessa (Laukka, 2000.). Tähän juuri sovelletaan Altmanin (1975) teoriaa myös tässä tutkielmassa, sillä tarkoituksena on käsittelyn edetessä havainnollistaa kuinka yksilöt pyrkivät säätelemään yksityisyyden tasoaan sekä erityisesti pohtia miten big datan hyödyntäminen voi vaikuttaa yksityisyyden säätelyyn. Tarkoituksena on myös arvioida kuinka tyydyttävästä yksityisyyden tasosta poikkeaminen vaikuttaa koettuun yksityisyyteen. Altmanin (1975) yksityisyyden säätelyn prosessimalliin palataan myöhemmin kappaleissa 3.2 ja 3.3.

3 BIG DATA JA YKSITYISYYS

Viime vuosien aikana on syntynyt tarve kehittää uusia innovatiivisia tekniikoita, jotka kykenisivät prosessoimaan tehokkaammin valtavia, hajautuneita ja nopeasti kasvavia datamassoja (Lublinsky, Smith & Yakubovich, 2013). Tällaisten valtavien, hajautuneiden ja nopeaa tahtia kasvavien datamassojen keräykseen, prosessointiin, analysointiin sekä esittämiseen tarkoitettujen tekniikoiden yhteisiksi nimittäjiksi ovat hiljalleen muodostuneet big data ja massadata-analytiikka (engl. big data analytics) (Sagiroglu & Sinanc, 2013). Massiivisten tietomassojen keräämisen ja analysoinnin on kuitenkin pelätty vaarantavan yksityisyyden suojan. Yksityisyyteen liittyvät ongelmat koetaan usein myös merkittäväksi haasteeksi big datan hyödyntämisen kannalta (Franks, 2012). Lisäksi big datan ja massadata-analytiikan käsitteiden merkityksestä ei ole toistaiseksi päästy täyteen yhteisymmärrykseen (Ward & Barker, 2013). Näin ollen lienee tarpeellista tutkia aihepiiriä hieman lähemmin. Tässä luvussa avataan ensin big datan ja massadata-analytiikan käsitteitä sekä tarkastellaan tämän tutkielman tavoitteiden kannalta keskeisiä big dataan ja massadata-analytiikkaan liittyviä tekniikoita ja teknologioita. Luvun loppupuolella siirrytään analysoimaan tarkemmin millaisia yksityisyyteen kohdistuvia riskejä big datan hyödyntämiseen voi liittyä sekä arvioidaan havaittujen riskien merkitystä yksilöiden ja yksityisyyden kokemisen kannalta.

3.1 Mitä on big data?

Kansainvälisen ICT-tutkimus- ja konsultointiyritys Gartnerin (2015) mukaan big data viittaa valtaviin, järjestelemättömiin sekä nopeasti kasvaviin datamassoihin, joiden tehokas prosessointi edellyttää uusia innovatiivisia tekniikoita. Gartnerin määritelmä luonnehtii hyvin big dataa yleisellä tasolla, sillä big datalle ominaisia piirteitä ovat juuri valtavat ja hajautuneet tietomassat, mitkä kasvavat nopealla tahdilla. Big datan tiedonlähteet ovat lisäksi mitä erilaisimpia, sillä tietoa kertyy muun muassa RFID-tunnisteista eli radiotaajuisista etätunnisteista, erilaisista antureista ja sensoreista, genomi-informaatiosta, sosiaalisen median syötteistä, hakuohjelmiin syötetyistä hakusanoista, internet-

klikkauksista, kannettavien laitteiden paikkatiedoista sekä kanta-asiakaskortille kirjatuista ostoksista ja myös erilaisten ohjelmien lokitiedoista. Franks (2012) on edelleen havainnut neljä big datalle tyypillistä seikkaa, mitkä yleensä ilmenevät big datan yhteydessä:

- Data on jonkin laitteen automaattisesti tuottamaa.
- Tietoa kerätään uudenlaisista lähteistä.
- Useat tiedonlähteet eivät ole suunniteltuja tiedonkeräämistä varten.
- Suuri osa kertyvästä datasta on hyödytöntä ja turhaa.

Tietojen hyödyntämisen suurimmat haasteet liittyvät usein kerättävän tiedon hallintaan sekä siihen missä käyttötarkoituksissa kerättyä tietoa voidaan hyödyntää, sillä osa kertyvästä tiedosta voi olla luonteeltaan arkaluonteista ja näin ollen mahdollisesti yksityisyyden suojaa loukkaavaa. Useista erilaisista lähteistä kerätyt tiedot ovat myös varsin kompleksisia ja vaativat huomattavan suuren tallennustilan. Näin ollen kertyvän tiedon siistimistä, käsittelyä tai jakamista ei voida enää suorittaa tehokkaasti pelkästään perinteisesti käytettävien teknologioiden avulla (Franks, 2012). Perinteiset relaatiotietokantoihin ja tietovarastoihin pohjautuvat tallennus-, haku- ja analysointimenetelmät eivät siis enää palvele big data – konseptia. Näin ollen tiedonhallintaa varten täytyy löytää uusia innovatiivisia ratkaisuja, joiden avulla dataa voidaan hallita tehokkaammin sekä edullisemmin. Seuraavissa kappaleissa tarkastellaan lähemmin big dataan ja massadata-analytiikkaan liittyviä teknologioita ja tekniikoita sekä havainnollistetaan big datalle sekä big datan tiedonlähteille ominaisia piirteitä 3V-käsitemallin pohjalta. Tavoitteena on luoda havainnollistava käsitys big datasta ja massadata-analytiikasta ja näin ollen luoda edellytykset peilata yksityisyyteen kohdistuvia riskejä teknologista taustaa vasten.

3.1.1 Big datan 3V-käsitemalli

Big dataa ja big datan tiedonlähteitä luonnehditaan usein 3V-käsitemallin avulla, joka koostuu kolmesta eri komponentista. Nämä komponentit kuvaavat big datalle sekä big datan tiedonlähteille tyypillisiä piirteitä. Näin ollen voidaankin ajatella, että jos tietolähde täyttää 3V-käsitemallissa esitetyt tunnusomaiset piirteet voidaan se tulkita big datan tiedonlähteeksi (Patil & Thia, 2013). Big datan ja big data tiedonlähteiden kolme tunnusomaista piirrettä; komponenttia ovat: *volyymi* (engl. volume), *vauhti* (engl. velocity) sekä *vaihtelevuus* (engl. variety), jotka esitti alun perin Laney (2001) viitatessaan näillä datamäärän lisääntymiseen sähköisessä kaupankäynnissä. Seuraavaksi tarkastellaan lähemmin kutakin 3V-käsitemallin komponenttia.

Volyymi

Volyymilla eli ”määrällä” viitataan big datalle ominaiseen datan suuruuteen ja sen kasvamiseen (Salo, 2013). Tyypillistä on, että dataa syntyy jatkuvasti ja, että erilaiset datatyypit tulevat tallennettavaksi erikokoisina. Valtavista datamas-

soista on etua analyysejä ja ennustuksia tehdessä, mutta haittapuolena on datan säilömiseen edellytetty suuri tallennustila, tehoton prosessointi sekä suuri ”turhan” datan määrä. Esimerkkinä volyymista ovat muutamien kilotavujen kokoiset blogitekstit sekä useiden gigatavujen kokoiset loki- ja klikkaustiedot (Krishnan, 2013). Data, mikä aiemmin volyyminsa puolesta luokiteltiin big dataksi, ei välttämättä ole sitä enää muutaman vuoden kuluttua, sillä teknologian kehityksen myötä yhä suurempia datamääriä kyetään tallentamaan ongelmitta (Franks, 2012). Näin ollen volyymin voitaisiin ajatella olevan enemmänkin suhteellinen käsite, joka on sidottu datan tallennuskykyyn.

Vauhti

Vauhti viittaa kiihtyvään nopeuteen, jolla dataa syötetään sekä otetaan käyttöön tietojärjestelmissä (Salo, 2013). Vauhdin avulla voidaan määritellä kohteen nopeus sekä liikkeen suunta (Krishnan, 2013). Tietoa kerätään myös usein reaaliajassa (esimerkiksi sulautetut järjestelmät), mikä edellyttää hyvin tehokasta ja nopeaa tiedonkäsittelyä. Vauhdista havainnollinen esimerkki on analyysin tekeminen sosiaalisen median synnyttämästä datavirrasta (Prajapati, 2013). Esineiden internetin ja internetverkkojen nopeutumisen myötä kertyvän datan vauhti luultavasti kiihtyy entisestään, sillä yhä useammat laitteet syöttävät dataa järjestelmiin entistä nopeampien verkkojen välityksellä.

Vaihtelevuus

Vaihtelevuus tai ”monimuotoisuus” kuvastaa datan muuttumista yhä epäyhtenäisemmäksi (Salo, 2013). Vaihtelevuudella viitataan monimuotoisiin datatyyppeihin, joissa voi esiintyä niin kuvia ja tekstiä kuin ääntä ja videoita (Prajapati, 2013). Dataa syntyy myös erilaisissa rakenteisissa muodoissa aina sähköposteista sosiaalisen median tilapäivityksiin (Krishnan, 2013). Tietoja kerätään näin ollen lukuisista ja mitä erilaisimmista lähteistä sekä formaateista, mistä seuraa edelleen datan suuri vaihtelevuus. Valtaosa tiedonlähteistä on myös tyypiltään sellaisia, että dataa ei kyetä keräämään strukturoidussa (järjestellyssä) muodossa (Franks, 2012). Massadata-analytiikan keskeisimpiä tavoitteita onkin käsitellä juuri tällaista strukturoimatonta (järjestelemätöntä) dataa ja löytää siitä hyödyllistä informaatiota. Massadata-analytiikkaa käsitellään kuitenkin itsenäisenä osa-alueena lähemmin kappaleessa 3.1.3.

Edellä esitettyjen kolmen V:n rinnalle on esitetty myös neljäs V-kirjain – ”Value” (suom. arvo), jonka on ajateltu kuvaavan datan tuomaa hyötyä liiketoiminnalle. Neljäs V ei kuitenkaan ole vakiinnuttanut asemaansa, sillä esimerkiksi datan luotettavuuteen liittyvää termiä ”Veracity” sekä datan olennaisuuteen liittyvää termiä ”Volatility” on ehdotettu neljänneksi V:ksi ”Valuen” ohella. Näin ollen 3V-käsitemalli on edelleen toimiva ja myös varsin laajalti käytetty tapa hahmottaa big datan luonnetta (Chen, Chiang & Storey, 2012).

3.1.2 Teknologiat ja tekniikat big datan takana

Valtavien, järjestelemättömien ja nopeaa tahtia kasvavien datamassojen tallentaminen ja käsittely sekä analysointi edellyttävät uudenlaisia tehokkaampia työkaluja (Manyika ym., 2011). Seuraavaksi käsitellään tämän tutkielman tavoitteiden kannalta keskeisiä teknologioita ja tekniikoita big datan takana, sillä kaikkien teknologioiden ja tekniikoiden käsittely ei ole aihepiirin laajuuden vuoksi mahdollista tämän tutkielman rajauksen sisällä.

Hadoop

Yksi keskeisimmistä teknologioista big datan takana on Java -pohjainen (Apache) Hadoop ohjelmistokehys, jonka päälle voidaan rakentaa erilaisia toimintoja. Hadoop soveltuu skaalautuvuutensa ansiosta hyvin suurteen datamassojen käsittelyyn ja se käsittää suuren määrän erilaisia keskenään yhteensopivia työkaluja, joten sitä voidaan käyttää moniin eri toimenpiteisiin ja käyttäjät voivat myös itse valita haluamansa sovelluskohteen (Lublinsky ym., 2013). Hadoopin laskennallinen tehokkuus perustuu datan käsittelyn hajauttamiseen useista tavallisista palvelimista muodostuviin klustereihin, joihin data toisinnetaan (kopioidaan) vikasietoisuuden lisäämiseksi (Salo, 2013). Hadoop koostuu käytännössä neljästä ydinkomponentista, jotka ovat: Hadoop Distributed File System (HDFS), Hadoop MapReduce, Hadoop Common sekä Hadoop YARN. Edellä mainittujen ydinkomponenttien lisäksi on olemassa myös muita oheisprojekteja, kuten Ambari, Avro, Cassandra, Chukwa, Hbase, Hive, Mahout, Pig, Spark, Tez ja Zookeeper (Apache, 2015.). Tässä tutkielmassa tarkastellaan kuitenkin vain HDFS:ää sekä MapReducea, sillä ne ovat tutkielman tavoitteiden kannalta oleellisimmassa asemassa.

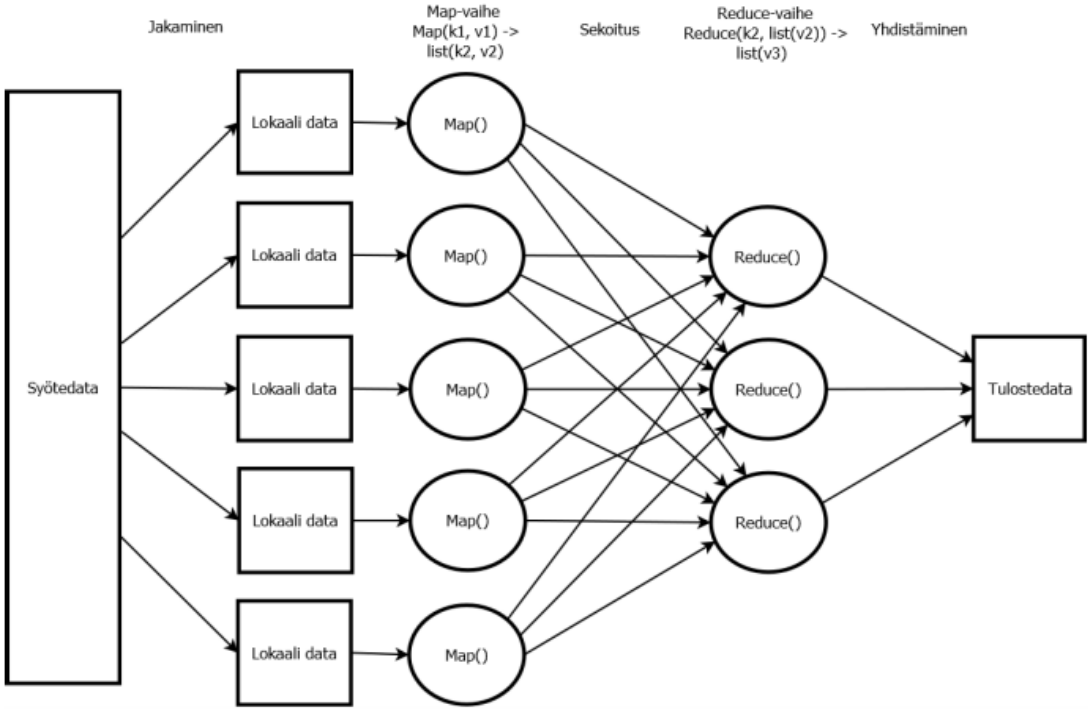
HDFS

HDFS eli Hadoop Distribudet File System on yksi Hadoopin ydinkomponenteista (Salo, 2013). HDFS on hajautettu tiedostojärjestelmä, minkä tehtävänä on hallinnoida palvelinklustereihin tallennettua dataa (Shvachko, Kuang, Radia & Chansler, 2010). HDFS on suunniteltu mahdollistamaan todella suurien datamassojen hajauttamisen klusterin palvelimille. Tyypillistä on myös, että data kirjoitetaan vain kerran, mutta luetaan useita kertoja esimerkiksi analysointia varten. Tallennettava data jaetaan edelleen lohkoihin, jotka hajautetaan palvelin klusteriin. Hajauttaminen mahdollistaa tehokkaamman käsittelyn sekä myös datan rinnakkaisprosessoinnin, mikä on hyödyllistä esimerkiksi analyyseja tehdessä (Lublinsky ym., 2013.). Hajauttamisella saavutetaan lisäksi korkeampi vikasietoisuus ja näin ollen myös parempi luotettavuus, sillä tiedostot paloitellaan pienempiin osiin, jotka vielä toisinnetaan usealle eri palvelimelle. Tällaisiin hajautettuihin tallennusrakenteisiin liittyy kuitenkin yksityisyyden kannalta ongelmallisia seikkoja. Hajauttaminen saattaa nimittäin johtaa yksityisyyden tason säätelyn vaikeutumiseen, sillä tällainen menettely voi vaikeuttaa tietojen poistamista sekä monimutkaistaa tietojen käsittelyyn liittyviä prosesseja ja näin

ollen hankaloittaa informaation yksityisyyden kontrollointia. Tilanteeseen liittyvä ongelmallisuutta käsitellään lähemmin kappaleissa 3.2 ja 3.3.

MapReduce

Apache Hadoop MapReduce on YARN-pohjainen järjestelmä massiivisten tietomäärien rinnakkaiseen käsittelyyn (Apache, 2015). MapReducen sanotaan usein muodostavan Hadoopin ytimen, sillä sen avulla voidaan käsitellä ja analysoida hyvin suuria klustereihin tallennettuja datamassoja sekä pilkkoa klustereihin kertyneet datamassat pieniksi palasiksi, jotka jaetaan edelleen klusterin muodostavaksi solmuiksi (Dittrich & Quiané-Ruiz, 2012; Schmidt & Phillips, 2013). Tarkemmin ottaen MapReducen voidaan ajatella koostuvan kahdesta erillisestä toiminnosta ”Map” ja ”Reduce”, joista ensimmäinen tapahtuu aina ennen jälkimmäistä (IBM, 2015). Näistä kahdesta funktiosta ”Map” ottaa parametrinaan syötedatan avain-arvoparina, jonka pohjalta se järjestelee ja suodattaa joukon uusia avain-arvo-pareja. ”Reduce” funktio puolestaan saa parametrikseen yhden Map-funktion tuottaman avaimen sekä kaikki kyseistä avainta vastaavat arvot, joille se suorittaa käyttäjän valitseman yhteenveto operaation. Tämän jälkeen Reduce-vaiheen tuottamat paluuarvot yhdistetään lopulliseksi ulostuloksi, mikä vastaa alun perin asetettuun ongelmaan (Salmi, 2015). Kuviossa 3 havainnollistetaan MapReduce-mallin mukaisen prosessin kulkua.



Kuvio 3: MapReduce-mallin mukaisen prosessin kulku (Salmi, 2015, 16).

MapReduce-mallin mukaiset hakuprosessit saattavat kuitenkin joissain tilanteissa vaarantaa yksityisyyden, sillä tällainen menettely mahdollistaa tietojen haun samanaikaisesti useilta eri palvelimilta ja ulostuloon saattaa yhdiste-

lyn vuoksi muodostua myös yksityisyyden suojan kannalta ongelmallisia tietoja. Kappaleissa 3.2 sekä 3.3 käsitellään tarkemmin big datan hyödyntämiseen liittyviä yksityisyyteen kohdistuvia riskejä.

Pilviteknologia
Pilviteknologiaan pohjautuvat järjestelmät toimivat perustana useille big data –palveluille sekä ympäristöille, sillä harvalla yrityksellä on olemassa olevaa infrastruktuuria tai valmiuksia investoida kertaheitolla edellytettyyn tallennuskapasiteettiin ja laitteistoon. Pilviteknologia tarjoaa lisäksi usein paremman skaalautuvuuden, jonka myötä tallennustilaa pystytään melko vaivattomasti kasvattamaan tarpeen mukaisesti. Erilaiset pilvipalvelut, kuten Microsoftin Azure HDInsight ja Googlen BigQuery, tarjoavatkin näin ollen keinot myös tera-, peta- sekä eksabittien hallintaan sekä prosessointiin (Microsoft, 2015; Google, 2015). Salon (2013) mukaan palveluntarjoajien aiemmin myymät pilvipalveluratkaisut ovatkin muuntautuneet hiljalleen eräänlaisiksi big data pilvipalveluiksi. Tämän tutkielman kannalta pilviteknologiaan pohjautuvat tallennusratkaisut ovat varsin mielenkiintoisessa asemassa, sillä pilvipalvelut hajauttavat yleensä tallennustilan varsin tehokkaasti. Näin ollen tietoja saatetaan tallentaa myös EU:n ulkopuolelle, mikä saattaa rikkoa EU:n henkilödirektiiviä (Andreasson, Koivisto, & Ylipartanen, 2014). Myös Wangin, Zhaon, Jiangin ja Len (2009) mukaan hajautettujen tietojärjestelmien täysimääräinen hyödyntäminen edellyttää yksityisyyteen ja tietoturvaan liittyvien asioiden tarkkaa huomioimista, sillä arkaluonteiset tiedot ovat hajautuneet eri puolille maailmaa. Pilviteknologiaan perustuviin tallennusratkaisuihin saattaa siis liittyä tietosuojan ja yksityisyyden kannalta pulmallisia seikkoja, joita käsitellään lähemmin kappaleissa 3.2 ja 3.3.

3.1.3 Massadata-analytiikka

Dataa ei kannata tallentaa suuria määriä jos sitä ei kyetä hyödyntämään tehokkaasti. Dataa voitaisiinkin näin ollen verrata öljyyn, millä ei jalostamattomana ole juuri käyttöä, mutta jalostettuna siitä tulee hyvin arvokasta kauppatavaraa. Tämän vuoksi tarvitaan menetelmiä, jotka mahdollistavat tehokkaan datan hyödyntämisen. Yhtenä ratkaisuna tähän ongelmaan on massadata-analytiikka, jonka tavoitteena on löytää hyödyllistä tietoa hyvin suurista, järjestelemättömistä ja nopeaa tahtia kasvavista datamassoista. Hyödyllisen tiedon löytäminen tällaisista datamassoista edellyttää uusia innovatiivisia teknologioita ja menetelmiä, sillä perinteiset analyysi menetelmät eivät yksinään kykene käsittelemään dataa riittävän nopeasti (Franks, 2012). Massadata-analytiikassa käytetäänkin useita eri menetelmiä ja teknologioita, kuten kehittynyttä tiedonlouhintaa, datan fuusiointia ja informaation integrointia, koneoppimista, hajautettua laskentaa sekä NoSQL-tietokantoja. Seuraavaksi käsitellään lähemmin tutkielman tavoitteiden kannalta keskeisiä massadata-analytiikassa hyödynnettäviä menetelmiä ja teknologioita, sillä kaikkien menetelmien ja tekniikoiden käsittely ei ole aihepiirin laajuuden vuoksi mahdollista tämän tutkielman rajauksen sisällä.

Tiedonlouhinta

Tiedonlouhinta saatetaan joskus sekoittaa analyysien tekemiseen, mutta on vain käytännössä osa sitä (Graham ym., 2014). Tiedonlouhintamenetelmien avulla pyritään löytämään laajoista tietoaineistoista piilevää informaatiota, mitä voitaisiin edelleen hyödyntää vaikkapa liiketoiminnassa. Louhintaan käytettävät menetelmät vaihtelevat aina matemaattisista menetelmistä puoliautomaattisiin tietojenkäsittelymenetelmiin (Hovi, Hervonen & Koistinen, 2009). Big datan tapauksessa datan hyödyntämiskohteet, määrä ja luonne sekä formaatti kuitenkin vaihtelevat jatkuvasti, mistä seuraa alituisen sovittamisen sekä hienosäädön tarve. Näin ollen, vaikka myös big datan tiedonlouhinnassa pätevät pitkälti samat periaatteet, prosessit sekä käytännöt, voivat louhittavan datan luonne ja louhintaan sovellettavat menetelmät sekä teknologia poiketa huomattavasti organisaation aiemmista tiedonlouhintaan liittyvistä käytännöistä. Big datan tiedonlouhinnassa yhdistyykin niin koneoppiminen ja tekoälyn käyttö kuin erilaiset tilastotieteelliset ja laskennalliset menetelmät. Lisäksi tiedonlouhintaan käytettävät menetelmät ja teknologiat riippuvat myös tiedonlouhinnalle asetetuista tavoitteista, sillä asetetut tavoitteet vaikuttavat esimerkiksi käytettäviin algoritmeihin (Graham ym., 2014.). Tiedonlouhinnan menetelmät ja teknologiat ovat myös jatkuvan tutkimuksen kohteena, joten ne kehittyvät koko ajan. Viimeaikaisimpia kehitystyön ja tutkimuksen saavutuksia ovat esimerkiksi yhä automaattisemmat ja itsestään oppivat tiedonlouhintamenetelmät (Le ym., 2012). Tiedonlouhinnan ja datan keräämisen automatisoinnilla voidaan entisestään tehostaa analyysien tekemistä. Näin ollen kehitystyö automatisoinnin parissa jatkuukin luultavasti myös tulevien vuosien aikana. Yksityisyyden kannalta suurimmat riskit liittyvät louhittavien aineistojen mittavuuteen ja monipuolisuuteen (Graham ym., 2014; Jensen, 2013).

Datan fuusiointi ja informaation integrointi

Datan fuusioinnille on olemassa useita toisistaan eroavia määritelmiä, mutta käytännössä datan fuusioinnin voidaan ajatella viittaavan erillisten tietoaineistojen (engl. data sets) yhdistämiseen yhtenäisemmäksi aineistoksi (Khaleghi, Khamis, Karray & Razavi, 2013). Informaation integroinnin voidaan myös ajatella tarkoittavan erillisten aineistojen yhdistämistä, mutta usein datan fuusiointi ja informaation integrointi toteutetaan hieman eri tavoin. Datan fuusiointi nimittäin tapahtuu yleensä järjestelmien välisellä kommunikaatiolla ja pelkistämällä sekä vaihtelemalla aineistojen dataa, kun taas informaation integroinnissa pikemminkin yhdistellään laajemmin aineistoissa olevaa dataa ja säilytetään kattavampi aineisto (Graham ym., 2014.). Tulevaisuudessa dataa kertyy yhä useimmista erilaisista sensoreista, jonka myötä datan fuusioinnin merkitys kasvaa entisestään (Khaleghi ym., 2014). Tällaiset yhdistelemällä luodut aineistot ovat kiinnostavassa asemassa myös tämän tutkielman kannalta, sillä niihin liittyy useita yksityisyyden kannalta pulmallisia seikkoja, joita käsitellään lähemmin tutkielman kappaleissa 3.2, 3.3 sekä 4.2. Toisaalta aineistoja yhdistelemällä saavutetaan huomattavia etuja, kuten tiedonlouhinnan tehostuminen ja

aineistojen helpompi hallittavuus. Kehitteillä on myös menetelmiä, joilla voidaan paremmin säilyttää yksityisyyden suoja myös tällaisissa yhdistetyissä aineistoissa (Navarro-Arribas & Torra, 2012). Tutkielman neljännessä luvussa tarkastellaan lähemmin yksityisyyden suojaamisessa hyödynnettäviä teknisiä menetelmiä, mutta jo tässä vaiheessa voidaan todeta anonymisoinnin sekä salauksen olevan merkittävässä asemassa yksityisyyden suojaamisen kannalta.

Kuvien ja äänten tunnistaminen

Erilaiset kuvien ja äänten tunnistamiseen käytettävät teknologiat voivat aistia monenlaista informaatiota, niin kuvista, videoista kuin puheestakin, joissain tilanteissa jopa yhtä hyvin kuin ihmiset (Graham ym., 2014). Näin ollen kehittyneiden teknologioiden ja tekniikoiden avulla on mahdollista saada monenlaista informaatiota myös reaalimaailman ilmiöistä, kuten ihmisten liikkeistä tai liikenteen sujuvuudesta. Kuvien pohjalta voidaan edelleen koostaa esimerkiksi erilaisia visualisointeja helpottamaan kaupunkisuunnittelua. Myös useat kuluttajille suunnatut kartta- sekä visualisointiapplikaatiot, kuten Google Earth ja GPS-navigointijärjestelmät, hyödyntävät digitaalisesti koostettuja kaupunkimalleja (Agarwal ym., 2011). Reaalimaailmasta kerättäviin tietoihin ja etenkin niiden yhdistämiseen muiden tietoaineistojen kanssa voi liittyä useita yksityisyyteen kohdistuvia riskejä, mutta toisaalta palveluiden ja tuotteiden (esimerkiksi karttapalvelujen tai suositusjärjestelmien) tarjoaminen saattaa edellyttää henkilökohtaisten tietojen luovuttamista. Näin ollen ongelman korjaaminen ei ole aivan yksiselitteinen prosessi, mikä käy ilmi myös kappaleissa 3.2 ja 3.3.

Sosiaalisten verkkojen analysointi

Sosiaalisten verkkojen analysoinnilla tarkoitetaan informaation louhimista toisiinsa kytkeytyneistä yksiköistä olettaen, että yksiköiden väliset suhteet ovat merkityksellisiä ja, että nämä yksiköt eivät toimi autonomisesti (Wasserman, 1994). Nykyään sosiaaliset verkot ilmenevät usein yhä selvemmin internetin välityksellä. Digitaalisen maailman sosiaalisista verkoista esimerkkeinä ovat sosiaalisen median palveluissa, kuten Facebookissa, LinkedInissä ja Twitterissä, ilmenevät sosiaaliset verkostot. Kuitenkin myös analogisen maailman sosiaaliset verkot saattavat jättää digitaalisia jälkiä, kuten soitettuja ja vastaanotettuja puheluita koskevia tietoja. Sosiaalisen verkkojen analysointia varten on olemassa myös useita erilaisia työkaluja, kuten Allegrograph, GraphVis, R, ja Wolfram Alpha. Lisäksi digitaalisen datan kasvun vuoksi voidaan metadataa hyödyntämällä tehdä entistä tehokkaammin analyyseja, mitkä saattavat uhata yksilöiden sosiaalista yksityisyyttä. (Graham ym. 2014.). Sosiaalisia verkkoja analysoimalla voidaan esimerkiksi päätellä seikkaperäisiä asioita myös yksilöiden mieltymyksistä, sillä ihmisillä on tapana assosioida niiden ihmisten kanssa, joilla heillä on jotain yhteistä (Zheleva & Getoor, 2008). Näin ollen sosiaalisten verkkojen analysointiin voi liittyä merkittäviä riskejä yksityisyyden kannalta (Smith, Szongott, Henne & von Voigt, 2012). Seuraavassa kappaleessa siirrytään tarkistelemaan lähemmin millaisia yksityisyyteen kohdistuvia riskejä big datan ja massadata-analytiikan hyödyntämiseen voi liittyä.

3.2 Big datan hyödyntäminen ja yksityisyyteen kohdistuvat riskit

Massiivisten tietomassojen keräämisen ja analysoinnin on pelätty vaarantavan yksityisyyden suojan. Huolestuneisuutta ovat herättäneet etenkin yksilöiden toimintaan liittyvien tietojen kerääminen eli esimerkiksi paikkainformaation tallentaminen tai internetin selauskäyttäytymisen seuraaminen. Tällaisia hankkeita on myös vastustettu voimakkaasti (Kshetri, 2014.). Yksityisyyteen liittyvät ongelmat koetaan usein myös merkittäväksi haasteeksi big datan hyödyntämisen kannalta (Franks, 2012). Tässä kappaleessa pyritään lähdemateriaalin sekä erilaisten skenaarioiden ja esimerkkien perustalta arvioimaan millaisia yksityisyyteen kohdistuvia riskejä big datan ja massadata-analytiikan hyödyntämiseen voi liittyä. Samalla pohditaan havaittujen riskien taustaa sekä riskien merkitystä yksilöiden ja yksityisyyden kokemisen näkökulmasta. Riskien ohella halutaan havainnollistaa myös datan hyödyntämiseen liittyviä mahdollisuuksia yksilöiden ja yhteiskunnan näkökulmasta, sillä ne kytkeytyvät yksityisyyden tason säätelyyn sekä yksityisyyden suojaamiseen liittyvään haasteellisuuteen ja ovat näin ollen tärkeässä asemassa myös tämän tutkielman kannalta.

3.2.1 Esimerkkejä ja skenaarioita

Tämän kappaleen tavoitteena on täsmällisen tulevaisuudenkuvan esittämisen sijaan havainnollistaa esimerkkien ja skenaarioiden avulla niin mahdollisuuksia kuin yksityisyyteen kohdistuvia riskejä, joita datan tehokkaaseen hyödyntämiseen voi liittyä. Kappaleessa pohditaan lisäksi niin riskien kuin myös mahdollisuuksien merkitystä yksilöiden, yksityisyyden kokemisen sekä yhteiskunnan kannalta.

Terveydenhuollon mullistuminen

Big datan on arvioitu mullistavan terveydenhuoltoa monella tapaa. Terveydenhuollossa voidaan ensinnäkin saavuttaa merkittäviä kustannussäästöjä big datan ja massadata-analytiikan menetelmiä hyödyntämällä, (Groves, Kayyali, Knott & Van Kuiken, 2013; Liikenne- ja viestintäministeriö, 2014). Big data mahdollistaa myös tehokkaampien hoitokeinojen kehittämisen sekä entistä tarkemman potilaiden seurannan (Graham ym., 2014; Groves ym., 2013). Datan laajamittainen kerääminen ja analysointi mahdollistavat lisäksi perinnöllisiin sairauksiin johtavien tekijöiden selvittämisen sekä näin ollen tarkempien ennusteiden laatimisen, jolloin hoitotoimenpiteet voidaan siis aloittaa jopa ennen varsinaisten oireiden alkamista. Myös epidemioiden puhkeamista ja kehittymistä voidaan seurata big dataa ja massadata-analytiikkaa hyödyntämällä, mikä taas on hyödyllistä epidemioiden torjunnan sekä pysäyttämisen kannalta (Mayer-Schönberger & Cukier, 2013).

Kuitenkin tällainen menettely edellyttää useimmissa tapauksissa hyvin laajamittaista tietojen keräämistä sekä analysointia ja mahdollisesti myös eri aineistojen yhdistelyä mittavampien analyysien aikaansaamiseksi. Tulee myös muistaa, että tässä tapauksessa merkittävä määrä kerättävästä ja/tai käsiteltä-

västä datasta on luonteeltaan arkaluonteista, sillä se sisältää tietoja yksilön terveydentilasta ja elintavoista. Edelleen voidaan havaita, että terveyttä koskevien tietojen hyödyntämiseen liittyy useita yksityisyyteen kohdistuvia riskejä, jotka kohdistuvat erityisesti yksityisyyteen henkilötietojen käsittelyssä, potilasyksityisyyteen sekä yksilöiden oikeuteen tulla arvioiduksi oikeassa valossa. Riskit ovat seurausta tietojen arkaluoneisuuden ohella myös uudelleentunnistamiseen liittyvistä riskeistä, sillä tiedot saatetaan pystyä kohdentamaan tiettyyn yksilöön tai ryhmään (Ohm, 2010). Uudelleentunnistaminen on näin ollen merkittävä riski yksityisyydelle ja se vaikeuttaa myös anonyymien aineistojen luomista. Uudelleentunnistamiseen liittyvää problematiikkaa tarkastellaan lähemmin kappaleessa 4.2.2. Tulee myös muistaa, että terveyttä koskevien analyysien ja ennusteiden laatimisessa pitää noudattaa erityistä huolellisuutta, sillä tietojen siistiminen tai käsittely saattaa johtaa harhaisuuteen, vaikka käytettävät tiedot olisivat paikkansapitäviä (Benjamini, 2010; Graham ym., 2014). Nykyään valtaosa terveydenhuollon tiedoista onkin lähtökohtaisesti salassa pidettäviä ja niiden käsittelyä sitoo yleensä vaitiolovelvollisuus (Andreasson ym., 2013). Näin ollen dataa ei voida tällä hetkellä hyödyntää täysin avoimesti ja terveydenhuolto on muutoinkin erityisasemassa tiedon käsittelyn kannalta, sillä luottamus tietojen oikeudenmukaisesta käytöstä on palvelu- ja potilassuhteen toiminnan edellytys. Potilastietojen sähköistymisen ja keskittämisen seurauksena tietosuojariskit tuleekin ottaa entistä paremmin huomioon, mutta toisaalta on tärkeää huomioida myös tietojen sähköistymisen ja keskittämisen mahdollisuudet (Andreasson, Koivisto & Ylipartanen, 2014.).

Sensorimaailma

Jo nykyään ihmisten liikkeistä voidaan kerätä monenlaista tietoa esimerkiksi matkapuhelimien tuottamista lokitiedoista (Becker ym., 2013). Lisäksi yhä useammat laitteet varustetaan erilaisilla sensoreilla, jotka tallentavat huomattavat määrät tietoa reaalimaailman ilmiöistä. Tulevaisuudessa tällaisten sensoreilla varustettujen laitteiden määrän on ennustettu kasvavan entisestään. Lisäksi yhä useammat näistä laitteista ovat yhteydessä verkkoon. Tietoliikenne- ja elektroniikkasuuryritys Ciscon (2011) esittämän arvion mukaan vuoteen 2020 mennessä maailmassa olisi jopa 50 miljardia internetiin kytkettyä laitetta. Samalla laitteiden sensorit luultavasti kehittyvät yhä tarkemmiksi ja kykenevät näin ollen tallentamaan yhä tarkempaa sekä monipuolisempaa dataa ympäristöstään (Graham ym., 2014). Tulevaisuudessa erilaiset laitteet ja sensorit saattavatkin tallentaa liikkeiden ohella monenlaista tietoa myös ihmisten toiminnasta sekä käytöksestä (Chen ym., 2012; Graham ym., 2014). Tällaisen analogisesta maailmasta kerättävän datan avulla voidaan saavuttaa koko yhteiskuntaa hyödyttäviä etuja, sillä saatavat tiedot voivat mahdollistaa esimerkiksi älykkäämmän infrastruktuurin, kuten kehittyneen katuvalaistuksen sekä automatisoidumman liikenteen ohjauksen (Liikenne- ja viestintäministeriö, 2014). Tulevaisuudessa ihmisten käytöksestä sekä toiminnasta saatavien tietojen perusteella voidaan luultavasti laatia ennusteita myös yksilöiden odotettavissa olevasta käyttäyty-

misestä, mikä voi auttaa esimerkiksi rikollisen toiminnan ennaltaehkäisemisessä (Graham ym., 2014).

Edellä kuvatun kaltaisiin menettelyihin voi kuitenkin liittyä ongelmallisia seikkoja yksityisyyden suojan kannalta, sillä sensoreiden tallentama ja erityisesti video- ja pysäytyskuvien tuottama data voi sisältää myös arkaluonteisia tietoja yksilöiden toiminnasta (Matusek, 2014). Pysäytys- ja videokuva rinnastetaankin tästä johtuen henkilötietolaissa varovaisuusperiaatteen noudattamisen edellyttämiin henkilötietoihin (Saarenpää, 2009). Näin ollen tällaisen datan hyödyntämisessä tulisi noudattaa samoja varovaisuuden periaatteita kuin muita henkilötietoja käsitellessä, joka saattaa osaltaan hankaloittaa tietojen hyödyntämistä. Tulee myös muistaa, että tehtävien ennusteiden paikkansapitävyyden varmistaminen saattaa osoittautua haasteelliseksi (Boyd & Crawford, 2011). Näin ollen ennusteet saattavat olla harhaisia, mikä taas rikkoo yksilöiden oikeutta tulla arvioiduksi oikeassa valossa. Tietojen harhaisuus on näin ollen syytä huomioida myös dataa hyödynnettäessä.

Kodin asema

Erilaisia sensoreita asennetaan koko ajan enemmän myös kodinkoneisiin ja muuhun kotona käytettävään elektroniikkaan. Lisäksi nämä laitteet kytkeytyvät yhä useammin internetiin (Atzori, Lera & Morabito, 2010; Mainetti, Patrono & Vilei, 2011.). Tällaisesta esineiden internetin kiihdyttämästä kehityksestä seuraa se, että myös kodin tapahtumista on mahdollista tallentaa paljon erilaisia tietoja. Kertyviä tietoja voidaan käyttää monenlaisiin käyttötarkoituksiin, joista kuluttajien kannalta yksi oleellisemmista käyttökohteista ovat älykkäämmät kodinkoneet ja laitteet. Kodista on näin ollen tulossa älykkäämpi - tulevaisuudessa jääkaappi voisi esimerkiksi analysoida asukkaiden ruokavalion tasapainoisuutta tarkastelemalla jääkaapissa olevia tuotteita ja puhelin saattaisi tarkkailla käyttäjänsä elintapojen aktiivisuutta. Lisäksi kosteus- ja lämpötilamittarien tuottamasta datasta voidaan jo nykyään päätellä seikkaperäisiä tietoja niin sisäilman laadukkuudesta kuin myös esimerkiksi asunnossa tapahtuvasta tupakoinnista (Graham ym., 2014).

Tällaiseen kehitykseen liittyy kuitenkin yksityisyyden kannalta ongelmallisia puolia (Atzori ym., 2010). Yksityisyyden suojan kannalta ongelmallista on ensinnäkin se, että ainakin osa laitteiden keräämistä tiedoista saatetaan siirtää käsiteltäväksi palveluntarjoajien palvelimille tehostetumman prosessoinnin sekä laajemman sisällöntuotannon vuoksi (Graham ym., 2014.). Tiedonsiirto kodin laitteiden ja palveluntarjoajan välillä taas altistaa kerättävän datan niin mahdollisille hyökkäyksille kuin myös kolmannen tahon toimesta tapahtuvalle datan väärinkäytölle. Yhtenä ratkaisuna tiedonsiirron turvaamiseen on salaus, jonka roolia yksityisyyden suojaamisessa käsitellään tarkemmin tutkielman neljännessä luvussa. Kuitenkin salausta hyödynnettäessäkin saattaa tällaisten verkon aktiivilaitteiden tietoturvasta löytyä aukkoja, joita hyödyntämällä hyökkääjät voivat päästä tarkkailemaan esimerkiksi lähiverkon liikennettä tai lukemaan laitteiden tallentamaa dataa. Pahimmillaan hyökkääjä voisi päästä käsiksi kodin turvajärjestelmään, jolloin hän kykenisi tarkkailemaan kameroiden tallentamaa

video – tai pysäytyskuvaa (Andreasson ym., 2014.). Huomionarvoista on myös, että Gartnerin (2012) arvion mukaan jopa reilu kolmannes datasta saatetaan säilyttää vuoteen 2016 erilaisissa pilvitallennusjärjestelmissä, mikä asettaa omalta osaltaan uudenlaisia haasteita tietoturvan suhteen. Näin ollen tällainen kodin virtualisoituminen saattaa asettaa uudenlaisia yksityisyyteen kohdistuvia riskejä sekä vaarantaa myös alueellisen yksityisyyden koskemattomuuden.

Koulutuksen uudistuminen

Laajamittainen datan kerääminen ja analysointi voivat mahdollistaa koulutuksen uudistamisen tulevaisuudessa niin, että oppilailla on mahdollisuus saada entistä yksilöllisempää opetusta. Tämä on mahdollista, sillä oppilaiden yksilöllisestä oppimisesta voidaan saada aiempaa tarkempaa tietoa ja tätä tietoa voidaan käyttää hyväksi opetuksen kehittämisessä (Daries ym., 2014; Graham ym., 2014). Tulevaisuudessa voitaneen analysoida yksityiskohtaisemmin myös opetettujen asioiden yhteyttä työelämässä menestymiseen (Graham ym., 2014). Näin ollen koulutusta voitaisiin parhaimmillaan muuttaa niin, että yhä useammat yksilöt kykenevät oppimaan työelämässä menestymiseen vaaditut taidot.

Tällainen menettely edellyttää kuitenkin suurten datamassojen keräämistä ja analysointia. Data saattaa tällaisten tietojen tapauksessa olla lisäksi luonteeltaan arkaluonteista, sillä se sisältää yksilöiden älykkyyttä koskevaa tietoa. Näin ollen datan käsittelyssä tulee noudattaa erityistä varovaisuutta, mikä saattaa osaltaan hankaloittaa tietojen hyödyntämistä. Yksityisyyttä voitaisiin mahdollisesti suojata muuttamalla aineistot anonyymiksi, jolloin tietoja voitaisiin myös hyödyntää vapaammin. Kuitenkin anonyymeihin aineistoihin saattaa liittyä uudelleentunnistamisen riski, kuten kappaleessa 4.2.2 tullaan osoittamaan, mikä vaarantaa tietoja koskevien henkilöiden yksityisyyden. Lisäksi myös tällaisen menettelyn ongelmana on analyysien luotettavuuden varmistaminen, joka on ensiarvoisen tärkeää, sillä harhaiset analyysit uhkaavat yksilöiden oikeutta tulla arvioiduksi oikeassa valossa.

Palvelut ja tuotteet 2.0

Kuluttajia analysoimalla voidaan tarjota parempia tuotteita ja palveluita, jotka vastaavat entistä tarkemmin kuluttajien mieltymyksiä (Boyd & Crawford, 2011; Franks, 2012). Useissa verkkokaupoissa hyödynnetäänkin jo suositusjärjestelmiä, jotka perustuvat esimerkiksi kuluttajan ostohistoriaan tai jonkinasteiseen kategorisointiin vaikkapa iän, sukupuolen tai kiinnostusten kohteiden mukaisesti. Tällaisella menettelyllä voidaan siis saavuttaa useita niin yrityksiä kuin kuluttajia hyödyttäviä etuja. Suomessa S-ryhmä onkin jo vuoden 2014 alusta siirtynyt keräämään entistä tarkempia tietoja S-etukorteilla tehdyistä ostoksista, joita hyödynnetään muun muassa kohdennetussa markkinoinnissa (Andreasson ym., 2014).

Kuluttajien analysointiin ja profilointiin voi liittyä kuitenkin myös yksityisyyden kannalta ongelmallisia puolia, sillä kulutuskäyttäytymisen ja sosiaalisten suhteiden pohjalta voidaan tehdä johtopäätöksiä esimerkiksi yksilön sukupuolisesta suuntautumista. Lisäksi tehtävien johtopäätöksien tai profiloinnin

luotettavuudesta ei voida olla täysin varmoja (Graham ym., 2014; Park, Huh, Oh & Han, 2012.). Tällainen menettely voi siis vaarantaa etenkin kuluttajien sosiaalisen yksityisyyden sekä yksilöiden oikeuden tulla arvioiduksi oikeassa valossa. Erityisen ongelmallisia seurantaan ja profilointiin liittyvät menettelyt ovat tilanteissa, joissa kuluttajat eivät ole näistä tietoisia. Ttällöin ongelmaksi voi muodostua se, että kuluttajat luovuttavat ymmärtämättömyyttään tai tietämättään heidän yksityisyyttään uhkaavia tietoja. Tämä on esimerkki datan keräämiseen ja käsittelyyn liittyvästä läpinäkyvyyden puutteesta. Myöhemmin osoitetaan, että läpinäkyvyyden puute saattaa hankaloittaa myös yksityisyyden tason säätelyä ja vaikuttaa näin ollen negatiivisesti yksityisyyden kokemiseen. Seuraavassa kappaleessa tarkastellaan yksityisyyteen kohdistuvien riskitekijöiden ohella lähemmin myös läpinäkyvyyden puutteen taustaa sekä tämän vaikutusta yksityisyyden säätelyyn ja yksityisyyden kokemiseen.

3.2.2 Yksityisyyteen kohdistuvat riskitekijät ja niiden merkitys

Tähän mennessä tehtyjen havaintojen myötä lienee selvää, että mahdollisimman laaja mittainen datan kerääminen ja hyödyntäminen ovat oleellisia tekijöitä koko yhteiskuntaa hyödyttäviä etujen saavuttamiseksi, mutta tällaiseen menettelyyn voi liittyä yksityisyyden kannalta ongelmallisia puolia. Lisäksi, kuten jo aiemmin osoitettiin, kokee jokainen yksityisyyden suojan eri tavoin (Brown & Muchira, 2004). Näin ollen ei ole itsestään selvää mitä tulisi suojata ja kuinka tarkkaan, mikä kävi ilmi myös edellisessä kappaleessa esitetyistä skenaarioista ja esimerkeistä. Tässä kappaleessa pyritään lähemmin analysoimaan big datan hyödyntämiseen liittyviä yksityisyyteen kohdistuvia riskitekijöitä sekä näiden taustaa. Kappaleessa pohditaan myös havaittujen riskitekijöiden merkitystä yksilöiden ja erityisesti yksityisyyden kokemisen näkökulmasta. Käsittelyn selkeyttämiseksi aihepiirin käsittely on jaettu riskitekijöiden piirteiden mukaisesti kahteen eri kategoriaan, jotka ovat: *kontrollin menetys* sekä *datan väärinkäyttö.*

Kontrollin menetys

Big datan hyödyntäminen saattaa hankaloittaa yksityisyyden tason säätelyä ja johtaa yksilöiden näkökulmasta jonkinasteiseen kontrollin menetykseen, sillä massiivisen datan keräämisen ja hyödyntämisen yhteydessä dataa saatetaan ensinnäkin käyttää eri tarkoitukseen kuin se on alun perin kerätty. Tällainen menettely on jopa melko ominaista big data -konseptille, sillä dataa syntyy lukuisista eri lähteistä ja tallennuksen yhteydessä tiedot yleensä toisinnetaan useille eri palvelimille (ks. kappale 3.1). Eri aineistoihin tallennettua dataa voidaan myös yhdistellä kattavampien analyysien aikaansaamiseksi (Christen, 2013). Kappaleessa 3.1 osoitettiin myös, että aineistojen yhdistäminen yksinkertaistaa lisäksi kerätyn datan hallintaa. Viime aikoina onkin entistä kattavammin pyritty yhdistämään erilaisia tiedonlähteitä (Doan, Halevy & Ives, 2012; Dong & Srivastava, 2013). Lisäksi, kuten jo aiemmin todettiin, tallennusratkaisut pohjautuvat yhä useammin kolmannen osapuolen tarjoamaan pilvipalveluratkaisuun (Salo, 2013). Pilviteknologian mahdollistamien suurempien ja usein myös edullisempien tallennustilojen ansiosta kertyvä data on myös mahdollista säilyttää

periaatteessa ikuisesti (Jensen, 2013). Datan pitkäaikainen säilytys onkin melko ominaista big data – konseptille. Tämä poikkeaa osaltaan aiemmista käytännöistä, jotka keskittyivät usein tarpeellisen datan säilyttämiseen ja hyödyntämiseen. Näin ollen datan hävittäminen omalta osaltaan saattaa jopa heikentää mahdollisuuksia löytää hyödyllistä informaatiota, sillä big data -konseptin yhtenä keskeisenä periaatteena on löytää uudenlaista informaatiota valtavista ja monipuolisista datamassoista. Tällainen menettely on kuitenkin ristiriidassa yksityisyyden tason säätelyn kannalta, sillä tietyissä tilanteissa yksilöillä tulisi olla oikeus heitä koskevien tietojen poistamiseen (Jensen, 2013).

Kontrollin menetystä puoltaa yksilöiden näkökulmasta myös se, että datan keräämiseen ja käsittelyyn liittyvät toimenpiteet sekä käyttötarkoitukset saattavat olla varsin epäselviä. Tämä johtuu osittain siitä, että näihin liittyvistä menettelyistä ja toimenpiteistä saatetaan tiedottaa vain vaikeaselkoisten ja epäselvien tietosuojalausekkeiden avulla (Pollach, 2007). Hajautuneet tallennusratkaisut sekä datan toissijainen käyttö saattavat myös vaikeuttaa todellisen tallennuspaikan sekä käyttötarkoituksen selvittämistä. Tämä saattaa johtaa läpinäkyvyyden puutteeseen, jonka myötä datan todellinen tallennuspaikka ja käyttötarkoitus hämärtyvät.

Läpinäkyvyyden puute sekä datan keräämiseen ja käsittelyyn liittyvät kompleksiset prosessit saattavat edelleen merkittävästi heikentää yksilöiden mahdollisuuksia kontrolloida omaa yksityisyyttään, sillä ne hankaloittavat halutun yksityisyyden tason sekä vallitsevan yksityisyyden tason välisten erojen hahmottamista. Tämä taas rikkoo yksityisyyden suojaa, sillä omien tietojen kontrollointi on yksi keskeisimmistä yksityisyyden suojan komponenteista (Berman & Mulligan, 1998). Tietojen kontrolloinnin vaikeutuminen saattaa edelleen vaikuttaa negatiivisesti myös yksityisyyden tason säätelyyn. Yksityisyyden tason säätelyn vaikeutuminen taas voi Altmanin (1975) yksityisyyden säätelyn prosessimallin mukaan johtaa eristyneisyyteen tai tungokseen. Tällöin yksilö ei kykene saavuttamaan haluamaansa yksityisyyden tasoa, minkä on aiemmissa tutkimuksissa osoitettu johtavan subjektiiviseen stressitason nousuun sekä samalla mielialan laskuun (Aiello & Kob, 1995; Davidson & Henderson, 2000). Näin ollen läpinäkyvyyden puute sekä datan keräämiseen ja käyttöön liittyvät kompleksiset prosessit saattavat vaikuttaa negatiivisesti myös koettuun yksityisyyteen sekä mielialaan. Yksityisyyden kontrolloinnin hankaloituminen saattaa johtaa lisäksi yksityisyyden suojan kaventumiseen jos yksilö luovuttaa tietämättään ja/tai ymmärtämättömyyttään yksityisyyttään vaarantavia tietoja, kuten paikkatietoja tai kulutuskäyttäytymiseen liittyviä tietoja. Sosiaalisen median palvelut ovat Smithin ym. (2012) mukaan omalta osaltaan vaikeuttaneet tilannetta entisestään.

Läpinäkyvyyden mahdollistaminen ei kuitenkaan ole Jensenin (2013) mukaan täysin ongelmatonta, sillä datan keräykseen ja käsittelyyn liittyy usein kilpailukyvyn kannalta oleellisia yrityssalaisuuksia, joiden paljastaminen olisi vastoin yritysten etuja. Lisäksi yksilöiden informoiminen datan keräämiseen ja käsittelyyn liittyvistä menettelyistä saattaisi edellyttää kohtuuttoman suurta panostusta yrityksiltä. Näin ollen läpinäkyvyyden mahdollistamiseksi sekä yksi-

tyisyyden ja yksityisyyden tason säätelyn mahdollistamiseksi tarvitaan myös muita toimenpiteitä, joita käsitellään tutkielman luvussa 4.

Datan väärinkäyttö

Datan väärinkäyttö on merkittävä riski yksityisyyden suojalle, sillä datan väärinkäytöllä rikotaan odotus oikeudenmukaisuudesta sekä luottamuksesta, jotka Berman ja Mulligan (1998) näkevät merkittävinä yksityisyyden suojan komponentteina anonyymiuden ja tietojen kontrolloinnin ohella. Datan väärinkäytökset ovat erityisen ongelmallisia silloin kun data sisältää arkaluonteisia, kuten terveyttä, älykkyyttä tai varallisuutta, koskevia tietoja (Graham ym., 2014). Tulee kuitenkin muistaa, että datan väärinkäyttö rikkoo lähes aina ilmiselvästi luvattoman tietojen tarkastelun tai käsittelyn kohteeksi joutuvan yksityisyyttä (Andreasson ym., 2014). Datan väärinkäytöksiin voivat johtaa useat erilaiset tilanteet, kuten tietomurrot tai muussa tilanteessa tapahtuva luvaton datan tarkastelu tai käsittely. Tietojen luvattomaan tarkasteluun voi syyllistyä kolmannen osapuolen ohella myös aineiston käyttöön näennäisesti valtuutettu henkilö, joka tahallaan tai ymmärtämättömyyttään tarkastelee työtehtävänsä kannalta tarpeettomia tietoja. Jälkimmäisenä mainittu esimerkki lienee yleisin datan väärinkäyttöön johtava tekijä terveydenhuollon puolella, jossa työntekijät voivat esimerkiksi uteliaisuuttaan vilkuilla työtovereidensa potilastietoja (Andreasson ym., 2014).

Pilviteknologiaan pohjautuvat tallennusratkaisut voivat myös omalta osaltaan lisätä datan väärinkäytösten riskiä, sillä tallennustilan tarjoamiseen osallistuu tällöin tavallisesti jokin kolmas osapuoli pilvipalveluratkaisun tarjoajan roolissa. Näin ollen on mahdollista, että pilvipalveluratkaisun tarjoaja tarkastelee tai käyttää tietoja luvattomasti (Cachin, Keidar & Shraer, 2009). Kolmannen osapuolen tai pahantahtoisten hyökkääjien toimesta tapahtuva luvaton datan tarkastelu tai käyttö eivät suinkaan ole ainoita syitä datan väärinkäytökselle, vaan väärinkäytös voi johtua myös ohjelmistovirheestä tallennustilan tarjoavan osapuolen palvelussa (Cachin ym., 2009). Lisäksi entisen CIA:n ja NSA:n työntekijän Edward Snowdenin paljastusten sekä julkisuudessa myöhemmin esitettyjen tietojen perusteella yhdysvaltalaiset pilvipalvelujen tarjoajat ovat velvoitettuja luovuttamaan tietoja tai pääsyn palveluihinsa NSA:lle (National Security Agency) (Andreasson ym., 2014). Tämä taas asettaa useiden pilvipalveluiden tietoturvan kyseenalaiseksi, sillä useat palveluntarjoajista, kuten Amazon ja Google, ovat yhdysvaltaisia suuryrityksiä. Tietoturvan kannalta ongelmallista on myös se, että pilviteknologian myötä tallennusratkaisun tarjoaminen saattaa keskittyä yhä harvempien tahojen käsiin, mikä taas voi kasvattaa merkittävien tietomurtojen riskiä. Näin ollen pilviteknologiaan pohjautuvien tallennusratkaisujen tapauksessa on perinteisten tietoturvaongelmien ohella mukana merkittävä joukko uusia ongelmia, mitkä voivat lisätä riskiä datan väärinkäytöksiin.

Datan väärinkäytösten riskiä lisäävät myös melko vaivattomasti saatavilla olevat tiedot, kuten erilaisten sosiaalisen median palveluiden tuottama data (Smith ym., 2012). Sosiaalisen median palveluiden tuottaman datan ja muun helposti saatavilla olevan datan tapauksessa voi myöskin olla entistä vaikeam-

paa määritellä, missä menee raja datan oikeudenmukaisen käytön ja väärinkäytön välillä, sillä datan käyttö voi joissain tilanteissa olla laillista, mutta samalla eettisesti arveluttavaa (Boyd & Crawford, 2011). Datan väärinkäytöksiä voidaan kuitenkin ennaltaehkäistä esimerkiksi salauksen tai anonymisoinnin avulla, joita käsitellään lähemmin tutkielman luvussa 4.

3.3 Yhteenveto

Kappaleessa 3.1 käsiteltiin big dataa ja tähän läheisesti liittyvää massadata-analytiikkaa. Aihepiiriä lähestyttiin big datan 3V-käsitemallin sekä big dataan ja massadata-analytiikkaan liittyvien teknologioiden ja tekniikoiden pohjalta. Tavoitteena oli muodostaa havainnollistava käsitys big datasta ja massadata-analytiikasta sekä näin ollen alustavasti avata teknologisen taustan valossa big datan hyödyntämiseen liittyviä yksityisyyteen kohdistuvia riskejä. Kappaleessa 3.2 siirryttiin analysoimaan tarkemmin millaisia yksityisyyteen kohdistuvia riskejä big datan hyödyntämiseen voi liittyä sekä pohtimaan havaittujen riskien merkitystä yksilöiden ja erityisesti yksityisyyden kokemisen näkökulmasta.

Kappaleissa 3.1 ja 3.2 esitettyjen havaintojen sekä lähdemateriaalin pohjalta voidaan tehdä johtopäätös siitä, että yksityisyyteen sekä turvallisuuteen liittyvät huolet vaikuttaisivat pitkälti olevan kytköksissä sosiaalisten verkkojen tiedonlouhintaan, ulkoistettuihin tietokantoihin (pilviteknologiaan pohjautuviin tallennusjärjestelmiin ja niihin liittyviin tietoturvariskeihin), valtavien data aineistojen hyödyntämiseen, datan välitykseen (esimerkiksi järjestelmien välillä), graafien tiedon analysointiin ja louhintaan sekä tiedonhakuoperaatioihin pilviteknologiaan pohjautuvista järjestelmistä (Cuzzocrea, 2014). Kappaleessa 3.2 havaittiin myös, että tallennustilan keskittymiseen liittyy uudenlaisia tietoturvariskejä, mitkä voivat edellyttää erityistä huomiota tietosuojan suhteen.

Kappaleessa 3.2 esitettyjen seikkojen perusteella voidaan myös todeta, että analyysien, ennusteiden tai profilointien tekeminen vahvistamattomien data aineistojen pohjalta vaikuttaisi lisäävän yksityisyyteen kohdistuvia riskejä, sillä tällöin tietojen luotettavuuden arviointi on hankalaa. (Boyd & Crawford, 2011; Graham ym., 2014; Rubinstein, 2012). Näin ollen voidaan päätyä tekemään johtopäätöksiä väärien tietojen pohjalta, mikä taas vaarantaa ilmiselvästi yksilöiden oikeuden tulla arvioiduksi oikeassa valossa. Kappaleessa 3.2.1 havaittiin myös, että jos kaikkia aineiston tietoja ei huomioida voivat tehtävät valinnat johtaa harhaisiin tuloksiin etenkin terveyttä koskevissa analyyseissa, vaikka itse tiedot olisivat paikkansapitäviä (Benjamini, 2010). Kappaleessa 3.2 esitettyjen esimerkkien ja skenaarioiden perusteella voidaan lisäksi todeta, että palveluiden ja tuotteiden käyttöön saattaa liittyä erilaisten tietojen keräämistä, joita saatetaan hyödyntää niin markkinoinnissa kuin tehostetumassa sisällöntarjonnassa. Näin ollen palveluiden ja tuotteiden käytön todellisena hintana saattaa olla jonkinasteinen yksityisyyden menetys. Erityisen ongelmallista tämä vaikuttaisi olevan sellaisissa tilanteissa, joissa kuluttajat ja/tai käyttäjät eivät ole tietoisia datan keräämiseen ja käsittelyyn liittyvistä toimenpiteistä.

Big datan hyödyntämiseen liittyvää ongelmallisuutta vaikuttaisi esille tulleiden havaintojen perusteella lisäävän omalta osaltaan myös läpinäkyvyyden puute. Läpinäkyvyyden puute vaikuttaisi olevan pitkälti seurausta kompleksisista datan keräämiseen ja käyttöön liittyvistä prosesseista sekä siitä, että näihin liittyviä menettelyjä saatetaan valottaa vain tietosuojalausekkeiden avulla. Aiemmissa tutkimuksissa on edelleen todettu, että tietosuojalausekkeet ovat usein pitkiä ja vaikeaselkoisia (Pollach, 2007). Smithin ym. (2012) mukaan tämä on entistä ongelmallisempaa sosiaalisen median palveluiden myötä. Läpinäkyvyyden puutteen sekä datan keräämiseen ja käyttöön liittyvien kompleksisten prosessien todettiin kappaleessa 3.2.2 vaikeuttavan tietojen kontrollointia sekä heikentävän näin ollen yksilöiden mahdollisuuksia säädellä omaa yksityisyyden tasoaan. Yksityisyyden tason säätelyn vaikeutuminen taas voi Altmanin (1975) yksityisyyden säätelyn prosessimallin mukaisesti johtaa eristyneisyyteen tai tungokseen. Näin ollen yksilöt eivät välttämättä kykene saavuttamaan haluamaansa yksityisyyden tasoa, minkä on aiemmissa tutkimuksissa osoitettu johtavan subjektiiviseen stressitason nousuun sekä samalla mielialan laskuun (Aiello & Kob, 1995; Davidson & Henderson, 2000).

Big datan ja tähän läheisesti liittyvän massadata-analytiikan hyödyntäminen voi näin ollen vaarantaa yksityisyyden sekä vaikuttaa negatiivisesti yksityisyyden kokemiseen. Suurimmat riskit vaikuttaisivat edellä esitettyjen havaintojen sekä tehtyjen johtopäätösten perusteella kohdistuvan alueelliseen yksityisyyteen, sosiaaliseen yksityisyyteen, anonymiteettiin, yksityisyyteen henkilötietojen käsittelyssä, yksilöiden oikeuteen tulla arvioiduksi oikeassa valossa sekä tiedolliseen omistusoikeuteen. Havaintojen perusteella myös yksityisyyden suoja saattaa joissain tilanteessa kaventua ja yksilöt voivat menettää lisäksi yksityisyyden mahdollistaman etulyöntiaseman, sillä erilaisia tietoaineistoja yhdistelemällä ja analogisesta maailmasta tallennettavan datan myötä esille saattaa tulla yksilöiden salassa pitämiä tietoja. Tällaisia tietoja saattavat olla esimerkiksi tupakointia tai muutoin epäterveellisiä elämäntapoja koskevat tiedot. Voidaan havaita, että tilanteen oikeudenmukaisuus riippuu osittain siitä mistä näkökulmasta aihepiiriä lähestytään, sillä osa on varmasti valmis hyväksymään sen, että epäterveellisiä elämäntapoja noudattavat maksavat enemmän vakuutuksistaan kuin terveellisesti elävät yksilöt. Tilanteesta haasteellisemman tekee kuitenkin se, että yksityisyyden suojan kaventuminen saattaa vaikuttaa negatiivisemmin jo valmiiksi heikossa asemassa oleviin yksilöihin (Kshetri, 2014).

Edellä esitetyt yksityisyyteen kohdistuvat riskit eivät suinkaan muodosta täydellistä listausta kaikista yksityisyyteen kohdistuvista riskeistä, joita big datan ja massadata-analytiikan hyödyntämiseen voi liittyä. Täydellisen listauksen laatiminen on hyvin haastavaa niin yksityisyyden ja yksityisyyden suojan määrittelyyn liittyvän haasteellisuuden kuin myös käytettävien teknologioiden ja tekniikoiden jatkuvan kehittymisen vuoksi. Näin ollen tarkan ja kaiken kattavan listauksen laatiminen jokaisesta mahdollisesta riskitekijästä ei olisi edes tarkoituksenmukaista saati mahdollista näin lyhyen tutkielman sisällä. Tästä johtuen tarkoituksena oli pikemminkin arvioida aihepiiriin liittyvää ongelmallisuutta yksilöiden ja yksityisyyden kokemisen näkökulmasta.

4 YKSITYISYYDEN SUOJAAMINEN

Tässä luvussa tarkastellaan yksityisyyden suojaamista teknologisesta näkökulmasta, joten aihepiirin käsittely painottuu tietoturvaan sekä erityisesti teknisiin toimenpiteisiin. Luvussa tutustutaan näin ollen lähemmin erilaisiin teknisiin menetelmiin, kuten anonymisointiin ja salaukseen, joilla yksityisyyttä voidaan suojata. Tarkoituksena ei kuitenkaan ole muodostaa täydellistä listausta kaikista olemassa olevista suojausmenetelmistä, vaan pikemminkin havainnollistaa erityyppisiä menetelmiä ja arvioida näiden käyttöön liittyviä haasteita sekä analysoida teknisten menetelmien kykyä suojella yksityisyyttä big datan aikakaudella.

4.1 Tietosuoja, tietoturva ja yksityisyys

Tietosuojan tarkoituksena on Andreassonin, Koiviston ja Ylipartasen (2013, 14) mukaan ohjata rekisterinpitäjiä hyviin henkilötietojen käsittelykäytäntöihin sekä turvata tiedon kohteen yksityiselämää, etuja ja oikeuksia. Tietosuoja on myös perustuslaillinen oikeus ja se takaa ihmisen oikeuden elää elämäänsä niin kuin tahtoo ilman kenenkään perusteetonta puuttumista siihen (Andreasson ym., 2013). Tietosuoja voitaisiinkin hieman filosofisemmin ajatella olevan eräänlaista tiedollisen kotirauhan kunnioittamista. Järvinen (2010) on pitkälti samoilla linjoilla, sillä hänen mukaan tietosuojan tarkoituksena on varmistaa henkilöihin tai heihin liittyvien tietojen oikeudenmukainen kerääminen sekä käyttö. Näin ollen tietosuoja on hyvin merkittävä tekijä yksityisyyden suojaamisen kannalta.

Tietoturva liittyy läheisesti tietosuojaan, mutta käsitteitä ei kuitenkaan tule pitää toistensa synonyymeina. Tietoturvan voidaan tarkemmin ottaen ajatella tarkoittavan tietosuojan näkökulmasta niitä toimenpiteitä, joilla rekisteröidyn yksityisyyden, etujen ja oikeuksien turvaamiseen sekä suojaamiseen pyritään. Tietoturvaan liittyy niin organisatorisia käytäntöjä kuin erilaisia teknisiä toimintoja, joiden tarkoituksena on varmistaa tiedon saatavuus, eheys sekä luot-

tamuksellisuuden säilyminen (Andreasson ym., 2013). Hakala, Vainio ja Vuorinen (2006) taas korostavat luotettavuutta, käytettävyyttä ja eheyttä tietoturvan peruspilareina. Tarkasteltaessa tietoturvaa tämän tutkielman kannalta voidaan havaita, että etenkin luotettavuus ja käytettävyys liittyvät keskeisesti tutkielman aihepiiriin, sillä ne ovat olennaisia tekijöitä pilviteknologian ja pilvipalveluiden tietoturvan kannalta, mitkä liittyvät läheisesti big datan takana oleviin teknologioihin ja tekniikkoihin (ks. kappale 3.1). Erityisesti luotettavuuden merkitystä voidaan perustella pilviteknologialle luonteenomaisella palvelurakenteella ja tähän liittyvällä tietojenkäsittelyn ulkoistamisella (Lehtinen, 2010.).

Myös kyberturvallisuutta voidaan pitää yhtenä tietoturvallisuuden alalajina, mutta tietoturvallisuudessa ja kyberturvallisuudessa painotetaan kuitenkin eri alueita, sillä kyberturvallisuus keskittyy yhteiskunnan toimintojen sekä väestön hyvinvoinnin suojaamiseen kyberavaruudessa (digitaalisessa maailmassa) tapahtuvia hyökkäyksiä vastaan, kun taas tietoturvallisuudessa pyritään varmistamaan toiminnan jatkuvuus suojaamalla tietojen luottamuksellisuutta, eheyttä sekä saatavuutta. Tietoturvan tapauksessa tiedot voivat myös olla niin paperilla, tietojärjestelmissä kuin henkilön muistissakin, kun taas kyberturvallisuudessa ajatellaan tietojen olevan sähköisessä muodossa (Andreasson ym., 2013.). Tulevien vuosien aikana kyberturvallisuuden merkitys tietoturvan ja näin ollen myös yksityisyyden suojaamisen kannalta korostunee entisestään, sillä tietojärjestelmiä kohtaan tehtyjen hyökkäysten määrä on ollut kasvussa 2000-luvulla (Graham ym., 2014). Hyökkäykset taas lisäävät datan väärinkäytösten riskiä, minkä todettiin olevan merkittävä uhka yksityisyydelle (ks. kappale 3.2.2).

Tietosuoja sekä tietoturvaan liittyvät toimenpiteet ovat siis hyvin merkittävässä asemassa niin toiminnan jatkuvuuden ja luottamuksen säilymisen kuin myös yksityisyyden suojaamisen kannalta. Tässä tutkielmassa keskitytään tarkastelemaan teknologisen näkökulman vuoksi juuri teknisiä menetelmiä, joiden avulla yksityisyyttä voidaan suojata. Näin ollen organisatoriset käytännöt sekä toimenpiteet, kuten tietoturvapolitiikka, tietoturvallisuuden organisoiminen, pääsyoikeuksien valvonta sekä henkilöstöturvallisuus, jäävät tutkielman rajauksen ulkopuolelle. Kuitenkin on hyvä muistaa, että organisatoriset toimenpiteet liittyvät oleellisesti tietoturvaan ja näin ollen myös yksityisyyden suojaamiseen.

4.2 Erilaiset suojausmenetelmät

Tässä kappaleessa tarkastellaan erilaisia teknisiä suojausmenetelmiä sekä arvioidaan näiden soveltuvuutta yksityisyyden suojaamiseen big dataa ja massadata-analytiikkaa hyödynnettäessä. Soveltuvuuden arvioinnissa painotetaan erityisesti menetelmän tarjoamaa yksityisyyden suojaa, mutta toisaalta myös aineistossa olevan datan hyödyntämismahdollisuuksia. Kappaleessa havainnollistetaan myös big datan ja massadata-analytiikan hyödyntämisestä seuraavia haasteita, joita teknisten menetelmien käyttöön voi liittyä. Kappaleen ensisijaisena tavoitteena on kuitenkin arvioida teknisten menetelmien kykyä ratkaista

yksityisyyteen liittyvät huolet, joita big datan hyödyntämiseen todettiin liittyvän tutkielman kolmannessa luvussa.

4.2.1 Kryptografia

Kryptografiaa on jo pitkään käytetty kommunikoinnin salaamiseen ulkopuolisten tarkastelulta (Rivest, 1990). Kommunikoinnin salaaminen on ollut ensisijaisen tärkeää etenkin sotatilanteiden aikana, jolloin käskyt tuli tehokkaasti välittää omille joukoille niin, että niiden tulkitseminen olisi mahdollisimman hankalaa vastapuolelle. Käytettävät salausmenetelmät ovat vaihdelleet varsin huomattavasti ja ajan mittaan on kehitetty useita erilaisia koodeihin perustuvia menetelmiä sekä järjestelmiä, kuten toisen maailmansodan aikainen Enigma, salaamaan viestit vastapuolen urkinnalta. Vastaavasti myös koodien purkamiseen on kehitetty useita erilaisia menetelmiä.

Toisen maailmansodan myötä salaukseen käytettävien menetelmien kehittely ja automatisointi nopeutuivat entisestään. Toista maailmansotaa voidaankin Järvisen (2003) mukaan pitää hyvin merkittävänä tekijänä kryptografian kehityksen kannalta. Samoihin aikoihin käyttöön tulivat myös ensimmäiset laskukoneet, mitkä kiihdyttivät automatisointia entisestään. Kryptografia alkoi näin ollen liittyä entistä läheisemmin tietojenkäsittelytieteisiin sekä matemaattiseen laskentaan (Rivest, 1990.). Kansainvälisen rahaliikenteen sähköistyminen 1970-luvulla lisäsi salauksen merkitystä entisestään ja lopulta 1990-luvulla matkaviestintä sekä internet liikenne mullistivat salaukseen kohdistuneen kaupallisen kysynnän (Järvinen, 2003). Nykyään kryptografia on Katzin ja Lindelin (2015) mukaan läsnä jokapäiväisessä elämässä.

Modernissa kryptografiasta käytetään laajalti algoritmisia sekä matemaattisia menetelmiä turvaamaan digitaalista informaatiota, järjestelmiä sekä erinäisiä operaatioita, uhkaavien tahojen hyökkäyksiltä ja se on oikein toteutettuna yksi vahvimmista keinoista turvata digitaalinen informaatio. (Graham ym., 2014; Katz & Lindell, 2015). Kryptografiaa hyödyntämällä voidaan mikä tahansa selkokielinen data (engl. plaintext) muuntaa salausalgoritmilla sellaiseen muotoon, mikä on käyttökelvotonta eli niin kutsuttua kryptotekstiä (engl. cryptotext; ciphertext). Tällainen kryptoteksti koostuu tarkemmin ottaen käytännössä satunnaisista bittijoukoista (Katz & Lindell, 2015). Salattua dataa ei voida hyödyntää mihinkään tarkoitukseen ennen kuin se puretaan, joten sitä ei näin ollen ainakaan periaatteessa tarvitse suojata millään muulla menetelmällä tai tekniikalla.

Kuitenkin salausmenetelmien vahvuuksissa on havaittavissa selviä eroja ja etenkin lyhempiin salausavaimiin (käytännössä siis vähemmän bittejä salaukseen käytettävässä avaimessa) perustuvat menetelmät ovat vaarassa joutua murretuiksi laskentakyvyn kasvaessa, sillä tällöin salaus voidaan pystyä murtamaan esimerkiksi kokeilemalla eri avaimia kunnes salattu data saadaan palautettua selkokieliseksi (Järvinen, 2003). Salaus pystytään murtamaan myös hyökkäämällä avaimen haltijaa vastaan ja kaappaamalla salaukseen käytettävä avain. Jälkimmäisenä mainittu hyökkäysmenetelmä on ongelmallinen etenkin salattuun avaimeen perustuvien eli niin kutsuttujen symmetristen salausmene-

telmien tapauksessa, sillä salaukseen käytettävä avain tulee toimittaa jokaiselle salausta tarvitsevalle taholle. Näin ollen salaisen avaimen paljastuminen tekee salauksen käyttökelvottomaksi. Symmetrisiä salausmenetelmiä ovat esimerkiksi DES (Data Encryption Standard), AES (Advanced Encryption Standard) sekä IDEA (International Data Encryption Algorithm).

Toinen salauksen toteuttamiseen käytetty tapa perustuu sekä julkiseen avaimeen että salaiseen avaimeen. Tällaisissa menetelmissä julkista avainta käytetään datan salaamiseen, mutta purkamiseen tarvitaan toista salassa pidettävää avainta. Näin ollen käytännössä kuka tahansa voi julkisen avaimen avulla salata datan, mutta sen purkamisen voivat tehdä vain ne tahot, jotka tietävät salaisen avaimen (Katz & Lindell, 2015; Järvinen, 2003). Tällaisissa menetelmissä salaukseen käytettävästä avaimesta sopiminen on siis vain ilmoitusasia eikä sitä tarvitse turvallisesti toimittaa lähettäjälle, mikä taas omalta osaltaan parantaa turvallisuutta (Järvinen, 2003). Tällaisista sekä julkista että salaista avainta käyttävistä salausmenetelmistä käytetään myös nimitystä epäsymmetrinen salaus. Epäsymmetriseen salaukseen perustuvia salausmenetelmiä ovat esimerkiksi RSA (Rivest, Shamir, Adleman) ja ECC (Elliptic Curve Cryptography). Epäsymmetrisen salauksen menetelmät saattavat kuitenkin olla hitaampia käyttää ja vaativat usein myös enemmän tallennustilaa kuin symmetriseen salaukseen perustuvat menetelmät, joten ne eivät välttämättä sovellu käytettäväksi kaikkiin tarkoituksiin. Lisäksi epäsymmetriseen salaukseen pohjautuvat menetelmät saattavat, etenkin lyhyiden viestien tapauksessa, olla alttiita Forward Search-hyökkäyksille. Forward-Search hyökkäyksessä kaapataan jokin salattu viesti ja pyritään kokeilemalla tuottamaan samanlainen viesti - jos kokeillusta tekstistä saadaan tuotettua julkista avainta käyttämällä samanlainen salattu teksti kuin kaapattu, niin saadaan tällöin selville myös alun perin lähetetty selkokielinen teksti. Epäsymmetrisen salauksen tapauksessa viestin lähettäjä voi lisäksi olla kuka tahansa, jolloin syntyy uudenlainen luottamusongelma lähettäjän luotettavuudesta (Järvinen, 2003.).

Salausteknologioita voidaan käyttää myös kerroksittain, jolloin muodostuu eräänlainen sipulinkuorta muistuttava suojamuuri datan suojaksi. Tällaisessa rakenteessa selkokielinen data on keskellä ja tätä saattaa ympäröidä useat erilaisia salausmenetelmiä hyödyntävät kerrokset. Kerroksittaisen rakenteen etuna on usein parempi turvallisuus sekä joustavampi käyttäjien valtuuttaminen, sillä käyttäjä voidaan valtuuttaa käyttämään vain osaan kerroksista ja näin ollen hän ei välttämättä koskaan pääse käsiksi selkokieliseen dataan (Popa, Redfield, Zeldovich & Balakrishnan, 2011). Kerroksittaista suojausta voidaan tietokantojen datan suojaamisen ohella soveltaa tietoliikenteen salaamiseen, mistä tunnettu esimerkki on anonyymin internetin käytön mahdollistava ohjelmisto Tor.

Viimeisimpiä kryptografian saavutuksia ovat salausmenetelmät, jotka mahdollistavat myös salatulla datalla operoinnin (Katz & Lindell, 2015). Näin ollen tietoturvariskit esimerkiksi hajautettua laskentaa tai analyyseja tehdessä pienenevät huomattavasti, sillä dataa ei tarvitse salata ja/tai purkaa yhä uudelleen. Tällaiset salausmenetelmät voivatkin parhaimmillaan vaikuttaa huomat-

tavan positiivisesti myös yksityisyyden suojan säilymiseen (Attrapadung, Libert & Peters, 2012). Kuitenkin tällaiset menetelmät saattavat olla hitaampia ja toiminnallisuudeltaan puutteellisia, joten ne eivät vielä sovellu kaupalliseen käyttöön (van Dijk, Gentry, Halevi, Vaikuntanathan, 2010). Kehitystä on tapahtunut myös paikkatietojen salaamisen puolella, mutta tällä hetkellä paikkatietojen käsittely ja suojaaminen tapahtuu vielä useimmiten luotetun kolmannen tahon avulla (Graham ym., 2014.). Tulevaisuudessa paikkatietojen salaaminen ilman kolmannen tahon osallistumista voi muodostua yhä tärkeämmäksi tavoitteeksi, sillä kuten aiemmin todettiin tallentavat yhä useammat laitteet tietoa yksilöiden sijainnista ja kolmannen osapuolen osallistuminen saattaa lisätä väärinkäytön mahdollisuuksia (ks. kappale 3.2). Lisäksi lähitulevaisuudessa erilaisten sensoreiden ja laitteiden tuottaman datan määrä tulee lisääntymään huomattavasti esineiden internetin myötä (Cisco, 2011). Tämä taas saattaa asettaa uudenlaisia yksityisyyteen sekä tietoturvaan kohdistuvia riskejä, mitkä ovat seurasta analogisen maailman ilmiöiden tallentamisesta sekä lisääntyvästä tiedonsiirrosta (ks. kappaleet 3.2 ja 3.3). Salaus on entistä tärkeämpää myös pilviteknologiaan perustuvien tallennusratkaisujen vuoksi, sillä tällaiset ratkaisut edellyttävät tiedonsiirtoa tallennusratkaisun tarjoajan ja palvelun käyttäjän välillä. Wangin, Lin, Owensin & Bhargavan (2009) mukaan turvalliseen pilviteknologiaan liittyykin olennaisesti vahva salaus. Myös Spiekermann ja Cranor (2009) korostavat tiedonsiirron turvaamista asianmukaisin keinoin. Näin ollen on entistä tärkeämpää huomioida salauksen kasvava merkitys niin datan siirrossa kuin myös tallennuksen ja käsittelyn yhteydessä.

Salauksen merkittävänä etuna on se, että data säilyy käyttökelpoisena. Näin ollen datan käyttöön valtuutetut tahot voivat hyödyntää sitä lähes yhtä hyvin kuin silloin jos se olisi tallennettu selkokielisinä. Salausta hyödynnettäessä tulee kuitenkin huomioida kolmansien tahojen mahdollisuudet päästä käsiksi dataan ennen kuin tämä ehditään muuntaa salattuun muotoon, sillä tällöin tiedot voidaan lukea selkokielisinä. Salauksen käytön unohtaminen saattaakin näin ollen muodostaa merkittävän uhan yksityisyydelle. Viimeaikaiset tietomurrot ovat myös osoittaneet, että salaukseen ei tule luottaa täysin sokeasti, sillä aina voi olla olemassa tietoturva-aukko, mitä hyödyntämällä hyökkääjä pääsee käsiksi tallennettuun dataan. Salausmenetelmät eivät myöskään täysin poissulje valtuutetun tahon toimesta tapahtuvaa datan väärinkäytön mahdollisuutta eivätkä ne itsessään ota huomioon muita datan käyttöön mahdollisesti liittyviä eettisiä ongelmia. Näin ollen salaus on merkittävän tärkeässä asemassa tietoturvan ja yksityisyyden suojaamisen kannalta, mutta se ei kuitenkaan tarjoa täydellistä ratkaisua ongelmaan.

4.2.2 Anonymisointi

Anonymisointi on jo pitkään ollut laajalti käytetty keino yksityisyyden suojaamiseksi erityisesti terveydenhuollossa sekä tutkimuskäyttöön suunnatuissa aineistossa. Anonymisointi toteutetaan käytännössä käsittelemällä rekisteröityjen henkilötietoja erilaisin menetelmin (esimerkiksi poistamalla tai salaamalla elementtejä) niin, että tiedot eivät ole enää yksilöitävissä (Graham ym., 2014). Ta-

voitteena on näin ollen saattaa tiedot sellaiseen muotoon, jossa niitä ei kyetä enää kohdentamaan tiettyyn yksilöön tai ryhmään. Hyvänä lähtökohtana anonyymille aineistolle voitaneen pitää sitä, että aineistosta ei tulisi pystyä eristämään tietueita niin, että yksilö tai ryhmä voidaan uudelleentunnistaa. Tietueita (samasta tietokannasta tai eri tietokannoista) ei myöskään tulisi pystyä yhdistämään koskemaan samaa rekisteröityä tai rekisteröityjen ryhmää. Lisäksi attribuuttien arvoa ei tulisi pystyä helposti päättelemään muiden attribuuttien sarjan arvoista. Jos anonymisointiin käytettävä menetelmä tarjoaa suojan kaikkia edellä mainittuja uudelleentunnistamiseen liittyviä riskejä vastaan, muodostaa se vahvan suojan uudelleentunnistamiselta, jonka rekisterinpitäjä tai kolmas osapuoli pystyisivät tekemään käyttämällä kohtuullisesti toteutettavissa olevia keinoja (Tietosuojaryhmä, 2014.). Yleisesti ottaen anonymisointi perustuu kahteen tekniikkaryhmään, jotka ovat: *satunnaistaminen* ja *luokituksen karkeistaminen.*

Satunnaistaminen perustuu tietojen totuudenmukaisuuden muuttamiseen, jolloin tietojen ja yksilön vahva yhteys voidaan parhaimmillaan poistaa. Tämä edellyttää kuitenkin, että tiedot ovat riittävän epävarmoja. Satunnaistaminen ei sinällään vähennä tietueiden ainutlaatuisuutta mikä tarkoittaa sitä, että jokainen tietue voidaan edelleen kohdentaa yhteen rekisteröityyn. Satunnaistaminen suojaa kuitenkin päättelyhyökkäyksiltä/-riskeiltä. Satunnaistamista voidaan käyttää myös yhdessä luokituksen karkeistamisen kanssa, jolloin tietosuojatakeet vahvistuvat. Kuitenkin myös tällöin lisätekniikat saattavat olla tarpeen, jotta tietueesta ei voida tunnistaa yksittäistä henkilöä. Satunnaistamiseen perustuvia menetelmiä ovat esimerkiksi kohinan lisääminen, permutaatio ja differentiaalinen yksityisyys (Tietosuojaryhmä, 2014.).

Luokituksen karkeistamisen tavoitteena on karkeistaa rekisteröityjen attribuuttien luokitusta niin, että joukosta erottaminen estyy. Tämä tapahtuu käytännössä yleistämällä attribuuttien arvoja mittakaavaa tai suuruusluokkaa muuttamalla. Luokituksen karkeistaminen ei kuitenkaan kaikissa tapauksissa takaa tehokasta anonymiteettia ja varsinkin yhdistettävyyden sekä päättelyn estäminen edellyttää monimutkaista kvantitatiivista lähestymistapaa. Luokituksen karkeistamiseen pohjautuvia menetelmiä ovat esimerkiksi aggregointi ja k-anonymiteetti sekä l-diversiteetti/t-läheisyys (Tietosuojaryhmä, 2014.).

Anonymisointia tehdessä kenties suurimpia virheitä on pitää peitenimillä suojattuja tietoja anonyymeina tietoina, sillä peitenimillä suojaaminen mahdollistaa melko todennäköisesti uudelleentunnistettavuuden. Peitenimellä suojaaminen tarkoittaa käytännössä tietueen attribuutin korvaamista toisella (esimerkiksi muutetaan henkilötunnus joksikin toiseksi yksilöiväksi tunnisteeksi x). Peitenimellä suojaamisen tavoitteena on vähentää tietoaineiston yhdistettävyyttä rekisteröidyn alkuperäiseen identiteettiin, mutta on osoitettu, että luonnollinen henkilö voidaan edelleen todennäköisesti yhdistää tietueeseen (Graham ym., 2014; Ohm, 2010). Tämän vuoksi peitenimellä suojaaminen ei yksinään johda anonyymien tietoaineiston luomiseen, mutta se voi vähentää tietoaineiston yhdistettävyyttä rekisteröidyn alkuperäiseen identiteettiin ja on näin ollen sellaisenaan hyödyllinen turvatoimi. Peitenimeen perustuvia menetelmiä ovat esimerkiksi salaus salaisella avaimella, tiivistysfunktio, avaimeen perustuva

tiivistysfunktio (avain tallennetaan), deterministinen salaus eli avaimeen perustuva tiivistysfunktio (avain poistetaan) ja alkioiden korvaus eli tokenisaatio (Tietosuojaryhmä, 2014.). Kuitenkin, kuten jo aiemmin todettiin, peitenimellä suojattu aineistoja ei tulisi pitää anonyymeina, vaan niiden käsittelyssä tulisi noudattaa pitkälti samoja varovaisuuden periaatetta kuin henkilötietoja käsiteltäessä. Edellä esiteltyjen anonymisointitekniikoiden vahvuuksia ja heikkouksia havainnollistetaan taulukossa 1.

	Onko joukosta erottaminen edelleen riski?	**Onko yhdistettävyys edelleen riski?**	**Onko päättely edelleen riski?**
Peitenimillä suojaaminen	Kyllä	Kyllä	Kyllä
Kohinan lisääminen	Kyllä	Ehkä ei	Ehkä ei
Korvaaminen	Kyllä	Kyllä	Ehkä ei
Aggregointi tai k-anonymiteetti	Ei	Kyllä	Kyllä
L-diversiteetti	Ei	Kyllä	Ehkä ei
Differentiaalinen yksityisyys	Ehkä ei	Ehkä ei	Ehkä ei
Tiivistysfunktio/tokenisaatio	Kyllä	Kyllä	Ehkä ei

Taulukko 1: Anonymisointitekniikoiden vahvuudet ja heikkoudet (Tietosuojaryhmä, 2014, 24)

Euroopan unionin tietosuoja direktiivin 95/46/EY ja muiden aihepiiriin kuuluvien EU:n säädösten mukaan anonymisointi tulisi tapahtua niin, että henkilön tunnistaminen estyy peruuttamattomasti. Näin ollen rekisterinpitäjän olisi siis otettava anonymisointia tehdessään huomioon kaikki ”kohtuudella toteutettavissa” olevat keinot, joita rekisterinpitäjä tai jokin kolmas osapuoli voivat tunnistamiseen käyttää. Anonymisoinnissa tulee myös noudattaa oikeudellisia rajoituksia. Oikeudellisten rajoitusten noudattaminen on tarpeen jo siksi, että tiedot säilyisivät tunnistettavassa muodossa, mikä on edellytys rekisteröityjen oikeudelle käyttää tiedonsaantioikeuttaan (Tietosuojaryhmä, 2014.).

Anonyymien aineistojen käsittelyssä ei tarvitse noudattaa samanlaista varovaisuutta tai säännöksiä kuin henkilötietojen käsittelyssä, sillä ne eivät lukeudu henkilötietojen piiriin. Näin ollen anonymisointi mahdollistaa tältä osin datan tehokkaan hyödyntämisen. Ongelmaksi ovat kuitenkin viime aikoina muodostuneet uudelleentunnistamiset, jonka myötä aineistossa olevat tiedot pystytään uudelleen yksilöimään tiettyyn rekisteröityyn tai rekisteröityjen ryhmään (Graham ym., 2014; Jensen, 2013; Tietosuojaryhmä, 2014). Uudelleentunnistamisen riski on erityisen suuri peitenimillä suojatuissa aineistossa, sillä uudelleentunnistaminen voidaan tehdä esimerkiksi yhdistämällä aineistoja keskenään, jolloin yksi tai useampi henkilö voi olla tunnistettavissa (Tietosuojaryhmä, 2014). On myös viitteitä siitä, että big data ja massadata-analytiikka menetelmiä hyödyntämällä uudelleentunnistaminen on entistä tehokkaampaa (Graham ym., 2014; International Working Group on Data Protection in Telecommunications, 2014). Lisäksi anonyymeja aineistoja käsiteltäessä tulee huomioida se, että vaikka anonyymit tiedot eivät kuulu tietosuojalainsäädännön piiriin voi rekisteröidyillä olla oikeus suojaan muiden säännösten, kuten vies-

tinnän luottamuksellisuutta koskevien säännösten, perustalta (Tietosuojaryhmä, 2014.).

Anonyymien aineistojen luominen onkin uudelleentunnistamiseen liittyvien riskien lisääntymisen sekä datan hyödyntämisen merkityksen korostumisen myötä entistä haastavampi tehtävä, sillä elementtejä riittävästi poistamalla, salaamalla tai totuusarvoa muuttamalla voidaan saavuttaa tehokkaampi suoja, mutta tällöin tietojen hyödyntämiskelpoisuus voi kärsiä tai tiedoista voi tulla hyödyntämiskelvottomia. Toisaalta jos elementtejä ei poisteta, salata tai muuteta riittävästi, on mahdollista, että yksilöt tunnistetaan uudelleen (Graham ym., 2014; Ohm, 2010; Tietosuojaryhmä, 2014). Näin ollen voidaan joutua tekemään jonkinlainen kompromissi yksityisyyden suojan ja datan hyödyntämiskelpoisuuden välillä. Yleisesti ottaen voidaan sanoa, että mitä enemmän dataa on saatavilla ja mitä monipuolisempaa data on, niin sitä todennäköisemmin yksilöt voidaan uudelleen tunnistaa anonyymeista aineistoista. Anonymisointia toteuttaessa tulee myös muistaa, että tunnistamisella ei tarkoiteta pelkästään mahdollisuutta selvittää henkilön nimi ja/tai osoite, vaan tämä kattaa myös tunnistettavuuden joukosta erottamalla, yhdistelemällä tai päättelemällä (Tietosuojaryhmä, 2014).

Anonymisointi voi näin ollen mahdollistaa tehokkaan datan hyödyntämisen sekä myös yksityisyyden suojan säilyttämisen, mutta anonyymien aineistojen luomisessa tulee noudattaa yhä enemmän erityistä tarkkaavaisuutta. Uudelleentunnistamisen riskiä tulisi myös jatkuvasti arvioida sekä tarvittaessa tehdä aineistoon muutoksia anonyymiuden säilyttämiseksi. Tietosuojaryhmän (2014) antaman lausunnon mukaan anonymisointitekniikoilla voidaankin taata tietosuoja vain jos niitä sovelletaan suunnitellusti. Näin ollen anonymisointia tehdessä tulisi siis arvioida käyttämänsä menetelmän mahdollistamaa suojaa vähintäänkin edellä mainittujen kolmen (erotettavuus, yhdistettävyys ja päättely) kriteerin mukaisesti, mikä taas edellyttää, että anonynimisointiprosessin edellytykset ja tavoitteet on määritelty selkeästi. Tämä on kuitenkin entistä hankalampaa big datalle ominaisen datan toissijaisen käytön (ks. luku 3) sekä uudelleentunnistamiseen liittyvien riskien vuoksi.

4.2.3 Ilmoitus ja hyväksyntä

Ilmoitus ja hyväksyntä (engl. notice and consent) on laajalti käytetty menetelmä yksityisyyden suojaamiseksi (Graham ym., 2014). Se perustuu käytännössä yksilön valintaan siitä hyväksyykö hän palvelun tai tuotteen, kuten ladattavien applikaatioiden, käyttämiseksi edellytetyt ehdot. Periaatteessa ilmoitukseen ja hyväksyntään pohjautuvat menetelmät mahdollistavat siis yksilöille kontrollin omasta yksityisyyden tasostaan, sillä he voivat hyväksymällä tai hylkäämällä valita ovatko valmiita luovuttamaan käyttöön edellytetyt tiedot. Ideaali tapauksessa yksilö lukeekin tuotteen ja/tai palvelun käyttöön liittyvät ehdot huolellisesti ja punnitsee ehtojen hyväksymiseen liittyviä seurauksia tarkkaan. Lienee kuitenkin selvää, että näin ei useimmissa tapauksissa ole, vaan yksilö saattaa tehdä päätöksensä ehtojen hyväksymisestä hetkessä sen kummemmin ajattele-

matta mahdollisia seurauksia (Barocas & Nissenbaum, 2009; Graham ym., 2014; Jensen, 2013).

Ilmoitukseen ja hyväksyntään perustuvien menetelmien ongelmallisuutta lisää entisestään se, että palvelun tai tuotteen käyttämiseen liittyvistä ehdoista eli esimerkiksi kerättävistä tiedoista sekä tarvittavista resursseista saatetaan ilmoittaa vain tietosuojalausekkeiden avulla, joiden on aiemmissa tutkimuksissa todettu olevan usein pitkiä ja vaikeaselkoisia (Pollach, 2007). Lisäksi big datan ja massadata-analytiikan hyödyntämisen myötä yksilöillä on todellisuudessa hyvin pienet mahdollisuudet arvioida ehtojen hyväksymisestä aiheutuvia seuraamuksia (ks. kappaleet 3.2.2 ja 3.3). Näin ollen ilmoitukseen ja hyväksyntään perustuvat menetelmät tarjoavat usein todellisuudessa vain laillisesti korrektin tavan ilmoittaa käyttöön liittyvistä ehdoista. Tästä johtuen yksilöiden ymmärrykseen perustuva hyväksyntä jää usein puuttumaan, mikä käytännössä tekee tällaiset menetelmät käyttökelvottomiksi yksityisyyden suojaamisen näkökulmasta.

Ilmoitukseen ja hyväksyntään perustuvia menetelmiä tulisikin muuttaa niin, että ne mahdollistavat yksilöiden ymmärrykseen perustuvan hyväksynnän saavuttamisen. Ymmärrykseen perustuvan hyväksynnän saavuttamisen kannalta on tärkeää lisätä datan keräämiseen ja käyttöön liittyvää läpinäkyvyyttä. Tämä voitaisiin toteuttaa esimerkiksi selkeyttämällä ehtoihin liittyviä tietosuojalausekkeita sekä tarjoamalla keskitetymmät ja automatisoidut työkalut yksityisyydentason mukauttamiseen. Graham ym. (2014) ehdottavat ratkaisuksi kolmannen osapuolen tarjoamaa alustaa yksityisyyden säätelyyn, mikä käytännössä perustuisi yksityisyys preferenssien tallentamiseen, jonka jälkeen tallennettuja preferenssejä voitaisiin käyttää automatisoidussa päätöksenteossa ehtojen hyväksymiseksi tai hylkäämiseksi. Näin ollen yksilöitä kuormittava päätöksenteko vähentyisi, mikä omalta osaltaan edesauttaisi todelliseen ymmärrykseen perustuvan sopimuksen aikaansaamisessa ja näin ollen myös yksityisyyden tason säätelyssä. Tällaisia keskitetyn ja automatisoidun yksityisyydentasonsäätelyn mahdollistavia alustoja ei kuitenkaan ainakaan tällä hetkellä ole kaupallisessa käytössä. Lisäksi myös tällaisiin keskitettyihin ratkaisuihin voi liittyä yksityisyyden kannalta ongelmallisia seikkoja, kuten kolmannen osapuolen toimesta tapahtuva datan luvaton tarkastelu tai käsittely. Myös päätöksenteon automatisoinnin luotettavuuden varmistaminen saattaa olla haastava tehtävä.

4.3 Johtopäätökset ja pohdinta

Kappaleessa 4.1 todettiin tietosuojan olevan perustuslaillinen oikeus, joka kytkeytyy tiiviisti tietoturvaan ja näin ollen edelleen yksityisyyden suojaamiseen. Samalla havaittiin myös, että tietoturva koostuu sekä organisatorisista että teknisistä toimenpiteistä, joilla tietoja pyritään suojaamaan. Tällöin havaittiin myös, että molemmat ovat tärkeitä tekijöitä tietosuojan kannalta, mutta tutkielman teknologisen näkökulman vuoksi organisatoristen toimenpiteiden käsittely ei ollut mahdollista tässä tutkielmassa. Kappaleessa 4.2 siirryttiin käsittelemään

erilaisia teknisiä menetelmiä, joiden avulla yksityisyyttä voidaan suojata. Tavoitteena oli arvioida teknisten menetelmien kykyä ratkaista big datan hyödyntämiseen liittyvät yksityisyyden suojaamiseen kohdistuvat haasteet. Tällöin havaittiin, että kryptografia sekä anonymisointi ovat oikein sovellettuna vahvoja menetelmiä yksityisyyden suojaamiseksi, mutta niiden käyttöön liittyy big datan ja massadata-analytiikan hyödyntämisen myötä yhä useampia haasteita.

Merkittävimmät haasteet vaikuttaisivat olevan seurausta valtavista ja monipuolisista tietoaineistoista sekä näiden tehokkaasta prosessoinnista ja yhdistämisestä sekä lisääntyvästä tiedonsiirrosta. Lisäksi big datan ja massadata-analytiikan menetelmät saattavat havaintojen perusteella lisätä uudelleentunnistamisen riskiä, mikä tekee anonyymien aineistojen luonnista entistä hankalampaa. Samalla havaittiin, että laskentakyvyn lisääntyessä yhä vahvempia salausmenetelmiä saatetaan pystyä murtamaan. Huomionarvoista on myös se, että big datan hyödyntämisestä seuraavien haasteiden ohella merkittävä riskitekijä vaikuttaisi tämän tutkielman perusteella olevan anonymisoinnin ja salauksen toteuttamiseen liittyvät inhimilliset virheet. Tällaisia virheitä ovat esimerkiksi salauksen merkityksen unohtaminen tai se, että peitenimellä suojattuja aineistoja pidetään aidosti anonyymeinä aineistoina (ks. kappaleet 4.2.1 ja 4.2.2). Näin ollen yksityisyyden suojaamisen epäonnistuminen ei aina välttämättä ole seurasta suojaamiseen käytetystä teknisistä menetelmistä, vaan inhimillisistä virheistä, joita esimerkiksi rekisterinpitäjä tai jokin muu taho tekee tietoja tarkastellessaan tai käsitellessään. Inhimillisten virheiden yksityiskohtainen tutkiminen ei tosin ollut mahdollista tämän tutkielman rajauksen sisällä. Uskoisin kuitenkin, että inhimillisten virheiden ennaltaehkäisyn kannalta on ensiarvoisen tärkeää laatia mahdollisimman kattavat ja yhdenmukaiset käytännöt, joilla datan oikeudenmukaiseen ja turvalliseen käyttöön pyritään. Edelleen voidaan havaita, että big datan ja massadata-analytiikan hyödyntämisen myötä on entistä tärkeämpää tietää millaisia yksityisyyteen kohdistuvia riskejä tietojen keräämiseen ja käsittelyyn voi liittyä, sillä riskien huomioimatta jättämisellä voi olla vakavia seuraamuksia rekisteröityjen yksityisyyden sekä luottamuksellisuuden säilymisen kannalta.

Tutkielmassa havaittiin lisäksi, että ilmoitukseen ja hyväksyntään perustuvien menetelmien suurin ongelma vaikuttaisi olevan seurausta ymmärrykseen perustuvan hyväksynnän uupumisesta. Tämän havaittiin johtuvan pitkistä ja vaikeaselkoisista tietosuojalausekkeista sekä datan keräämiseen ja käyttöön liittyvistä kompleksisista prosesseista, josta seuraa läpinäkyvyyden puute (ks. kappaleet 3.2.2, 3.3 ja 4.2.3). Tilanteen korjaamiseksi tietosuojalausekkeita tulisi selkeyttää ja yksilöille tulisi tarjota päätöksentekoa helpottavia palveluita, joita he voisivat hyödyntää yksityisyys preferenssiensä tallentamiseen sekä edelleen yksityisyyden tason säätelyyn liittyvissä asioissa. Tarkemmin ottaen tämä voisi tapahtua esimerkiksi niin, että yksityisyys preferenssit tallennettaisiin, minkä jälkeen niitä voitaisiin käyttää tuotteen ja/tai palvelun käyttöön liittyvien ehtojen hylkäämiseen tai hyväksyntään (ks. kappale 4.2.3). Näin ollen päätöksenteko tapahtuisi preferenssien tallentamisen jälkeen automatisoidusti, joka voisi vähentää ehtojen hyväksyntään liittyvää päätöksenteon kuormittavuutta ja

omalta osaltaan edesauttaa yksityisyyden säätelyyn liittyvää haasteellisuutta. Luottamuksellisuuden kannalta voisi olla hyödyllistä jos julkinen sektori osallistuisi tällaisen palvelun tarjoamiseen. Julkisen sektorin tarjoama identifiointi- ja yksityisyysalusta voisi parhaimmillaan tarjota luotettavan palvelun sekä yksityisyyden tason säätelyyn että tunnistautumiseen. Tällainen identifiointi- ja yksityisyysalusta saattaisi myös ainakin osittain ratkaista läpinäkyvyyden mahdollistamiseen liittyvän ongelman, sillä yritysten ei tarvitsi paljastaa yksityiskohtaisia spesifikaatioita datan keräämiseen ja käsittelyyn liittyvistä algoritmeista tai muista prosesseista kaikille osapuolille, vaan ainoastaan ilmoittaa ne alustan tarjoajalle eli jollekin luotetulle taholle – edellä esitetyn esimerkin tapauksessa siis julkiselle sektorille. Tällaisen alustan kehittämiseen liittyy luultavasti useita haasteita, joiden ratkaiseminen vaatii jatkoselvityksiä. Kuitenkin vaihtoehtoisille yksityisyyden tason säätelyä helpottaville teknologioille saattaa tulevaisuudessa olla entistä enemmän kysyntää, sillä yksityisyyden tason säätely vaikuttaisi havaintojen valossa olevan entistä haastavampi tehtävä. Käytettävät teknologiat ja menetelmät sekä yksityisyyden merkitys muuttuvat kuitenkin jatkuvasti, joten tarkkaa kehityssuuntaa on hankala ennustaa.

Tutkielman havaintojen perusteella tarvitaan vaihtoehtoisten menetelmien ohella entistä enemmän myös organisaatioiden itsesäätelyä sekä lainsäädäntöä ja valvontaa, jotta datan oikeudenmukainen käyttö saadaan varmistettua. EU:n tietosuojadirektiivin uudistus tähtää omalla panoksellaan vastaamaan kasvaviin haasteisiin joita datan hyödyntämiseen voi liittyä, sillä tietosuojadirektiivin uudistuksen myötä organisaatioilta tullaan edellyttämään muun muassa tarkkaavaisempaa tietojen käsittelyä sekä vaatimaan tietosuojavastaavien nimittämistä organisaatioihin, joissa käsitellään hiemankin suurempia henkilörekisterejä (Andreasson ym., 2014). Tämän tutkielman puitteissa ei tosin voida tarkemmin tutkia organisatoristen käytäntöjen tai lainsäädännön mahdollisuuksia yksityisyyden suojaamiseksi. Voidaan kuitenkin havaita, että datan tehokkaan hyödyntämisen mahdollistava ja samalla yksityisyyden suojan säilyttävän viitekehyksen rakentaminen edellyttää jatkoselvityksiä sekä tiivistä yhteistyötä tiedeyhteisön, lainsäätäjien sekä muiden tahojen kesken. Kehitystyön tukena voisivat toimia jo olemassa olevat standardit, joissa määritellään tietoturvan ja turvamekanismien vaatimuksia. Tällaisia standardeja ovat esimerkiksi ISO27001 sekä ISO27002, mitkä kuuluvat ISO27000 standardisarjaan. Kyseinen standardisarja on kehitetty parantamaan yritysten kykyä hallita tietoturvallisuutta, riskejä sekä valvontaa (Evans, Tsohou, Tryfonas & Morgan, 2010.). Näin ollen ne voisivat tarjota hyvän lähtökohdan kehitystyölle. Kansainvälisten standardien kehittäminen ja yhteensovittaminen lainsäädännön kanssa tukee myös yhtenäisempiä käytäntöjä, joista sopimisen havaittiin olevan tärkeää niin datan oikeudenmukaisen käytön kuin myös luottamuksen kannalta.

5 YHTEENVETO

Tässä kirjallisuuskatsaukseen pohjautuvassa tutkielmassa tarkasteltiin yksityisyyttä, big dataa ja massadata-analytiikkaa sekä yksityisyyden suojaamiseen käytettäviä teknisiä menetelmiä. Tutkielmassa pyrittiin selvittämään millaisia yksityisyyteen kohdistuvia riskejä big datan ja tähän läheisesti liittyvän massadata-analytiikan hyödyntämisestä voi aiheutua sekä pohdittiin havaittujen riskien merkitystä yksilöiden ja erityisesti yksityisyyden kokemisen näkökulmasta. Toisena painopisteenä olivat yksityisyyden suojaamiseen käytetyt tekniset menetelmät sekä näiden käyttöön liittyvät haasteet big datan aikakaudella. Tutkimusongelma tiivistettiin edelleen seuraaviin tutkimuskysymyksiin:

- Millaisia yksityisyyteen kohdistuvia riskejä big datan hyödyntämiseen voi liittyä ja kuinka nämä riskit voivat vaikuttaa yksityisyyden kokemiseen?
- Kykenevätkö tekniset menetelmät suojaamaan yksityisyyden big datan aikakaudella ja millaisia haasteita teknisten menetelmien käyttöön voi liittyä?

Tutkielman toteutus perustui johdannossa esiteltyyn systemaattisen kirjallisuus katsauksen malliin. Näin ollen tutkimusongelmaa lähestyttiin suunnittelun ja protokollan validoinnin jälkeen lähdemateriaalin koostamisella, mikä toteutettiin sovitun protokollan mukaisesti eli avainsanojen sekä erilaisten hakukoneiden ja tietokantojen avulla. Suoraan käyttökelpoisen lähdemateriaalin löytäminen osoittautui ajoittain haastavaksi tehtäväksi, minkä uskon johtuvan aihepiiriä koskevan tutkimuksen verrattain vähäisestä määrästä sekä lähdemateriaalille asetuista laatuvaatimuksista. Lähdemateriaalia jouduttiin tämän vuoksi keräämään myös aihepiiriin ulkopuolisesta kirjallisuudesta, joka kuitenkin liittyi läheisesti tutkielman aihepiiriin. Tutkielman aineisto koostuu näin ollen niin yksityisyyttä, big dataa, yksityisyyden suojaamiseen käytettäviä menetelmiä kuin sähköisiä palveluita käsittelevistä julkaisuista. Tutkielmassa esitetyt tulokset perustuvat niin tutkielmassa käytettyyn lähdemateriaaliin kuin myös havaintojen pohjalta tehtyihin johtopäätöksiin.

Lähdemateriaalin koostamisen ja validoinnin jälkeen tutkimusongelmia lähestyttiin big datan, yksityisyyden sekä yksityisyyden suojaamiseen käytettävien teknisten menetelmien näkökulmasta. Toisessa luvussa käsiteltiin yksityisyyttä sekä tähän liittyvää monitulkintaista semantiikkaa ja esiteltiin myös tarkemmin tutkielmassa käytetty lähestymistapa aihepiiriin. Kolmannessa luvussa muodostettiin asetettujen painopisteiden perustalta havainnollinen käsitys big datasta ja tähän läheisesti liittyvästä massadata-analytiikasta sekä näihin liittyvistä teknologioista ja tekniikoista. Tämän jälkeen siirryttiin arvioimaan lähemmin big datan hyödyntämiseen liittyviä yksityisyyteen kohdistuvia riskitekijöitä ja näiden merkitystä yksilöiden sekä erityisesti yksityisyyden kokemisen näkökulmasta.

Kappaleessa 3.2.1 ja 3.2.2 esitettyjen skenaarioiden sekä näiden pohjalta havaittujen riskitekijöiden perusteella tehtiin johtopäätös siitä, että yksityisyyteen sekä turvallisuuteen liittyvät huolet vaikuttaisivat olevan pitkälti kytköksissä sosiaalisten verkkojen tiedonlouhintaan, ulkoistettuihin tietokantoihin (pilviteknologiaan pohjautuviin tallennusjärjestelmiin ja niihin liittyviin tietoturvariskeihin), valtavien tietoaineistojen hyödyntämiseen, datan välitykseen (tiedonsiirtoon), graafien tiedon analysointiin ja louhintaan sekä tiedonhakuoperaatioihin pilviteknologiaan pohjautuvista järjestelmistä (Cuzzocrea, 2014). Edelleen havaittiin myös, että tallennustilan keskittymiseen liittyy uudenlaisia tietoturvariskejä, mitkä saattavat edellyttää erityistä huomiota tietosuojan suhteen.

Tutkielmassa todettiin myös, että analyysien, ennusteiden tai profilointien tekeminen vahvistamattomien aineistojen pohjalta vaikuttaisi lisäävän yksityisyyteen kohdistuvia riskejä, sillä näiden luotettavuuden arviointi on hankalaa. (Boyd & Crawford, 2011; Graham ym., 2014; Rubinstein, 2012). Tällöin näiden pohjalta tehtävät johtopäätökset saattavat perustua vääriin tietoihin, mikä taas vaarantaa ilmiselvästi yksilöiden oikeuden tulla arvioiduksi oikeassa valossa. Edelleen havaittiin myös, että jos kaikkia aineiston tietoja ei huomioida voivat tehtävät valinnat johtaa harhaisiin tuloksiin etenkin terveyttä koskevissa analyyseissa, vaikka itse data olisi paikkansapitävää (Benjamini, 2010). Dataa saatetaan havaintojen perusteella käyttää lisäksi muihin tarkoituksiin kuin se on alun perin kerätty. Näin ollen yksilöt saattavat ymmärtämättömyyttään luovuttaa heidän yksityisyyttään uhkaavia tietoja, mikä on Smithin ym. (2012) mukaan entistä ongelmallisempaa sosiaalisen median myötä.

Tilanteen ongelmallisuutta vaikuttaisi lisäävän omalta osaltaan myös läpinäkyvyyden puute, mikä hankaloittaa erityisesti yksityisyyden tason säätelyä. Yksityisyyden tason säätelyn vaikeutumisen todettiin Altmanin (1975) yksityisyyden säätelyn prosessimallin mukaisesti johtavan eristyneisyyteen tai tungokseen. Aiemmissa tutkimuksissa on todettu, että epätyydyttävä yksityisyyden taso johtaa subjektiiviseen stressitason nousuun sekä samalla mielialan laskuun (Aiello & Kob, 1995; Davidson & Henderson, 2000). Näin ollen big datan ja massadata-analytiikan hyödyntäminen voivat havaintojen perusteella vaikuttaa negatiivisesti myös koettuun yksityisyyteen sekä mielialaan. Tutkielmassa havaittiin lisäksi, että big data ja massadata-analytiikka menetelmät voivat

mahdollistaa yksilöiden tai ryhmän uudelleentunnistamisen anonyymeista aineistoista, mikä vaarantaa ilmiselvästi tunnistettujen yksityisyyden. Samanlaisia tuloksia on saatu myös muissa julkaisuissa (Graham ym., 2014; Ohm, 2010; Tietosuojaryhmä, 2014). Tutkielmassa tehtyjen havaintojen valossa big datan hyödyntämiseen näyttäisi näin ollen liittyvän yksityisyyteen kohdistuvia riskejä, mitkä kohdistuvat erityisesti alueelliseen yksityisyyteen, sosiaaliseen yksityisyyteen, anonymiteettiin, yksityisyyteen henkilötietojen käsittelyssä, oikeuteen tulla arvioiduksi oikeassa valossa sekä tiedolliseen omistusoikeuteen.

Tutkielman toinen painopiste oli yksityisyyden suojaamiseen käytettyjen teknisten menetelmien tarkastelussa. Tavoitteena oli arvioida teknisten menetelmien kykyä ratkaista yksityisyyden suojaamiseen liittyvät haasteet. Menetelmien valinta perustui johdannossa esiteltyihin kriteereihin, joiden perusteella tarkastelun kohteeksi valikoitui kolme yksityisyyden suojaamiseen käytettyä teknistä menetelmää. Valittujen menetelmien todettiin tukevan osaltaan myös tutkielmalle asetettuja tavoitteita. Valitut menetelmät olivat 1) kryptografia, 2) anonymisointi sekä 3) ilmoitus ja hyväksyntä.

Kryptografian merkittävimpänä etuna havaittiin olevan datan säilyminen käyttökelpoisena. Samalla todettiin kuitenkin, että salausta hyödynnettäessä tulee huomioida datan väärinkäytökset sekä erityisesti mahdollisuudet päästä käsiksi salaamattomiin aineistoihin. Havaintojen pohjalta tehtiin edelleen johtopäätös siitä, että salaukseen ei myöskään tulisi luottaa liikaa, sillä viimeaikaiset tietomurrot ja laskentakyvyn lisääntyminen ovat osoittaneet erityisesti heikompien salausmenetelmien haavoittuvuuden. Salausmenetelmät eivät tutkielman aikana tehtyjen havaintojen tai käytetyn lähdemateriaalin perusteella myöskään täydellisesti poissulje datan käyttöön valtuutetun tahon toimesta tapahtuvaa datan väärinkäytön mahdollisuutta eivätkä ne itsessään vaikuttaisi ottavan huomioon muita datan käyttöön mahdollisesti liittyviä eettisiä ongelmia.

Toinen tarkastelun kohteeksi valittu menetelmä oli anonymisointi, minkä todettiin kryptografian ohella mahdollistavan yksityisyyden suojaamisen sekä datan hyödyntämisen, mutta samalla sen käyttöön havaittiin liittyvän uudelleentunnistamisen riski, joka hankaloittaa anonyymien aineistojen luomista (Graham ym., 2014; Ohm, 2010; Tietosuojaryhmä, 2014; International Working Group on Data Protection in Telecommunications; 2014). Edelleen tehtiin havaintojen valossa johtopäätös siitä, että big datan ja massadata-analytiikan hyödyntämisen myötä uudelleentunnistamisen riski usein kasvaa.

Kolmantena tarkasteltiin ilmoitukseen ja hyväksyntään perustuvia menetelmiä, joiden havaittiin usein epäonnistuvan yksityisyyden suojaamisessa. Tällaisten menetelmien ongelmana todettiin olevan erityisesti ymmärrykseen perustuvan hyväksynnän puuttuminen (Barocas & Nissenbaum, 2009; Graham ym., 2014). Havaintojen perusteella tehtiin johtopäätös siitä, että ilmoitukseen ja hyväksyntään perustuvien menetelmien ongelmallisuus saattaa lisääntyä big datan ja massadata-analytiikan myötä. Tällöin ehdotettiin myös, että menetelmiä voitaisiin kehittää selkeyttämällä tietosuojalausekkeita sekä automatisoimalla käyttöehtojen hyväksyntään liittyvä päätöksenteko.

Tutkielman tarkoituksena ei ollut tarjota valmista viitekehystä ja esille tulleiden havaintojen valossa sellaisen rakentaminen on hyvin työläs prosessi. Havaintojen perusteella vaikuttaisi kuitenkin siltä, että esitettyjen teknisten menetelmien ohella tarvitaan entistä enemmän vaihtoehtoisia teknisiä menetelmiä sekä organisatorisia toimenpiteitä ja lainsäädäntöä yksityisyyden suojaamiseksi. Erityistä tarkkaavaisuutta tulee tässä tutkielmassa tehtyjen havaintojen perusteella kiinnittää liiallisen säätelyn välttämiseen, mutta samalla tulee löytää keinot yksityisyyden suojan säilyttämiseksi. Hyvänä käytäntönä voitaneen tässä tutkielmassa esitettyjen havaintojen perusteella pitää myös varovaisuuden noudattamista – oli sitten kyse henkilötietojen tai anonyymien aineistojen käsittelystä. Lisäksi läpinäkyvyyden mahdollistaminen sekä käyttöehtojen hyväksyntään liittyvän päätöksenteon kuormittavuuden lieventäminen vaikuttaisivat olevan merkittäviä tekijöitä niin luottamuksen kuin yksityisyyden tason säätelyn kannalta.

Saadut tulokset olivat odotettuja ja tutkimusprosessi täytti sille asetetut tavoitteet. Tutkielman suurimpana kontribuutiona voitaneen sen kokoavan ja pohtivan luonteen vuoksi pitää keskustelun nostattamista sekä big datan hyödyntämiseen liittyvien ongelmien ja toisaalta myös yksityisyyden suojaamiseen liittyvien haasteiden havainnollistamista. Empiirisen aineiston puuttumisen sekä tutkijan valintaan perustuvan lähdemateriaalin vuoksi esitettyjen tulosten yleistettävyyteen tulee kuitenkin suhtautua tietyllä varauksella, vaikka harhaisuutta pyrittiin välttämään systemaattisella lähestymistavalla.

Jatkotutkimukset ovat tarpeen jo aihepiirin laajuuden vuoksi, sillä tässä tutkielmassa pystyttiin käsittelemään hyvin rajautuneesti vain teknologista näkökulmaa aiheeseen. Tulee myös muistaa, että käytettävät teknologiat ja menetelmät kehittyvät jatkuvasti, mikä lisää omalta osaltaan jatkotutkimusten tarpeellisuutta. Jatkotutkimuksissa voitaisiin selvittää lainsäädännön ja organisatoristen toimenpiteiden mahdollisuuksia yksityisyyden suojaamisen sekä datan käytön mahdollistamisen kannalta. Lainsäädännön ja organisatoristen toimenpiteiden tarkastelussa voitaisiin painottaa riskienhallinnan näkökulmaa, minkä uskon olevan yhä olennaisemmassa roolissa datan käyttöön liittyvien ongelmien ratkaisemiseksi. Myös yksityisyyden määrittely jäi tässä tutkielmassa teoreettiselle tasolle. Lisäksi, kuten tutkielmassa havaittiin, yksityisyyden käsite muuttuu alati teknologien kehityksen ja yhteiskunnan mukana. Näin ollen jatkotutkimukset voisivat suuntautua myös yksityisyyden tutkimiseen sekä erityisesti kulttuuristen ja yksilökohtaisten eroavaisuuksien selvittämiseen. Tutkielmassa ei myöskään pystytty käymään lävitse kaikkia olemassa olevia suojausmenetelmiä tai kuvaamaan tarkasteltuja menetelmiä yksityiskohtaisesti. Näin ollen kiinnostavana jatkotutkimuksen kohteena voisivat olla myös tekniset menetelmät, joita voidaan hyödyntää yksityisyyden suojaamisessa.

LÄHTEET

Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S. M. & Szeliski, R. (2011). Building rome in a day. *Communications of ACM, 54*(10), 105-112.

Aiello, J. & Kolb, K. (1995). Electronic performance monitoring and social context: Impact on productivity and stress. *80*(3), 339-353.

Allen, A. (1988). *Uneasy access: Privacy for women in a free society*. Totowa: Rowman & Littlefield.

Altman, I. (1975). *The environment and social behavior: Privacy, personal space, territory, and crowding*. Monterey: Brooks/Cole Publishing Company.

Andreasson, A., Koivisto, J. & Ylipartanen, A. (2013). *Tietosuojavastaavan käsikirja*. Helsinki: Tietosanoma Oy.

Andreasson, A., Koivisto, J. & Ylipartanen, A. (2014). *Tietosuojavastaavan käsikirja 2*. Helsinki: Tietosanoma Oy.

Apache. (2015, 7. elokuuta). What is apache hadoop? Haettu 14.8.2015 osoitteesta https://hadoop.apache.org/

Attrapadung, N., Libert, B. & Peters, T. (2012). Computing on authenticated data: New privacy definitions and constructions. Teoksessa X. Wang, & K. Sako (toim.), (s. 367-385) Springer Berlin Heidelberg.

Atzori, L., Iera, A. & Morabito, G. (2010). The internet of things: A survey. *Computer Networks, 54*(15), 2787-2805.

Barocas, S. & Nissenbaum, H. (2009). On notice: The trouble with notice and consent. *Proceedings of the Engaging Data Forum: The First International Forum on the Application and Management of Personal Electronic Information, October 2009.*

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., . . . Volinsky, C. (2013). Human mobility characterization from cellular network data. *Communications of the ACM, 56*(1), 74-82.

Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal, 52*(6), 708-721.

Berardo, F. M. (1975). Marital invisibility and family privacy. Teoksessa D. H. Carson (toim.), *Man environment interactions: Evaluations and applications* (). Stroudsburg: Dowden, Hutchinson & Ross.

Berman, J. & Mulligan, D. (1998). Privacy in the digital age: Work in progress. *Nova Law Review, 23*(2), 551-582.

Boyd, D. & Crawford, K. (2011). Six provocations for big data. *Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011.* Oxford: Oxford Internet Institute.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M. & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software, 80*(4), 571-583.

Brown, M. & Muchira, R. (2004). Investigating the relationship between internet privacy concerns and online purchase behavior. *Journal of Electronic Commerce Research, 5*(1), 62-70.

Burgoon, J. K., Parrott, R., Le Poire, B. A., Kelley, D. L., Walther, J. B. & Perry, D. (1989). Maintaining and restoring privacy through communication in different types of relationships. *Journal of Social and Personal Relationships, 6*(2), 131-158.

Cachin, C., Keidar, I. & Shraer, A. (2009). Trusting the cloud. *ACM SIGACT News, 40*(2), 81-86.

Carver, C. S. & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality–social, clinical, and health psychology. *Psychological Bulletin, 92*(1), 111-135.

Chen, H., Chiang, R. & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165-1188.

Christen, P. (2014). Privacy aspects in big data integration: Challenges and opportunities. *Proceedings of the First International Workshop on Privacy and Security of Big Data,* Shanghai, China (1-1). New York, NY, USA: ACM.

Cisco. (2011). *The internet of things how the next evolution of the internet is changing everything.* (White Paper nro 04/11). Cisco Internet Business Solutions Group.

Cuzzocrea, A. (2014). Privacy and security of big data: Current challenges and future research perspectives. *Proceedings of the First International Workshop on Privacy and Security of Big Data,* Shanghai, China (45-47). New York, NY, USA: ACM.

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., . . . Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. *Communications of the ACM, 57*(9), 56-63.

Davidson, R. & Henderson, R. (2000). Electronic performance monitoring: A laboratory investigation of the influence of monitoring and difficulty on task performance, mood state, and self-reported stress levels. *Journal of Applied Social Psychology, 30*(5), 906-920.

Dittrich, J. & Quiane-Ruiz, J. A. (2012). Efficient big data processing in hadoop MapReduce. *Proceedings of the VLDB Endowment, 5*(12), 2014-2015.

Doan, A., Halevy, A. & Ives, Z. (2012). *Principles of data integration*. Waltham: Morgan Kaufmann.

Dong, X. L. & Srivastava, D. (2013). Big data integration. *Data Engineering (ICDE), 2013 IEEE 29th International Conference on,* (1245-1248).

Evans, R., Tsohou, A., Tryfonas, T. & Morgan, T. (2010). Engineering secure systems with ISO 26702 and 27001. *System of Systems Engineering (SoSE), 2010 5th International Conference on,* (1-6).

Franks, B. (2012). *Taming the big data tidal wave : Finding opportunities in huge data streams with advanced analytics*. Hoboken New Jersey: John Wiley & Sons, Inc.

Gantz, J. & Reinsel, D. (2012). *THE DIGITAL UNIVERSE IN 2020: Big data, bigger digital shadows, and biggest growth in the far east.* (nro 12/12). IDC IVIEW.

Gartner. (2012, 25. kesäkuuta). Gartner says that consumers will store more than a third of their digital content in the cloud by 2016. Haettu 14.8.2015 osoitteesta http://www.gartner.com/newsroom/id/2060215

Gartner. (2015, 14. kesäkuuta). What is big data? Haettu 14.6.2015 osoitteesta http://www.gartner.com/it-glossary/big-data

Google. (2015, 14. kesäkuuta). BigQuery - google cloud platform. Haettu 14.6.2015 osoitteesta https://cloud.google.com/bigquery/

Graham, S. L., Press, W., Gates, S. J., Gorenberg, M., Holdren, J., Lander, E. S., . . . Blumenthal, M. S. (2014). *Big data: A technological perspective.* (nro 01/05).President's Council of Advisors on Science and Technology.

Groves, P., Kayyali, B., Knott, D. & Van Kuiken, S. (2013). *The 'big data' revolution in healthcare: Accelerating value and innovation.* (nro 01/13). McKinsey & Company.

Hakala, M., Vainio, M. & Vuorinen, O. (2006). *Tietoturvallisuuden käsikirja*. Jyväskylä: Docendo.

Harris, P. B., Werner, C. M., Brown, B. B. & Ingebritsen, D. (1995). Relocation and privacy regulation: A cross-cultural analysis. *Journal of Environmental Psychology, 15*(4), 311-320.

Hovi, A., Hervonen, H. & Koistinen, H. (2009). *Tietovarastot ja business intelligence*. Jyväskylä: WSOYpro.

IBM. (2015, 14. kesäkuuta). What is MapReduce? Haettu 14.6.2015 osoitteesta http://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/

International Working Group on Data Protection in Telecommunications. (2014). *Working paper on big data and privacy privacy principles under pressure in the age of big data analytics. 55th meeting, 5 – 6 may 2014, skopje.* (nro 675.48.12).

Järvinen, P. (2003). *Salausmenetelmät*. Jyväskylä: Docendo.

Järvinen, P. (2010). *Yksityisyys : Turvaa digitaalinen kotirauhasi*. Jyväskylä: Docendo.

Jensen, M. (2013). Challenges of privacy protection in big data analytics. *Big Data (BigData Congress), 2013 IEEE International Congress on,* (235-238).

Jourard, S. M. (1966). Some psychological aspects of privacy. *Law and Contemporary Problems, 31*(2), 307-318.

Katz, J. & Lindell, Y. (2015). *Introduction to modern cryptography, second edition.* Boca Raton: CRC Press Taylor & Francis Group.

Kelly, E. P. & Rowland, H. C. (2000). Ethical and online privacy issues in electronic commerce. *Business Horizons, 43*(3), 3-12.

Khaleghi, B., Khamis, A., Karray, F. O. & Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion, 14*(1), 28-44.

Krishnan, K. (2013). *Data warehousing in the age of big data*. Amsterdam: Morgan Kaufmann.

Kshetri, N. (2014). Big data's impact on privacy, security and consumer welfare. *Telecommunications Policy, 38*(11), 1134-1145.

Laney, D. (2001, 6. helmikuuta). 3D data management: Controlling data volume, velocity and variety. Haettu 14.6.2015 osoitteesta http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Laukka, M. (2000). *Proceedings of the Fifth Nordic Workshop on Secure IT Systems (Nordsec 2000), Reykjavik, October 12-13, 2000.*

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on,* (8595-8598).

Lehtinen, V. (2010). *Tietoturvan ja tietosuojan kehittäminen pilviteknologiassa : Standardit ja kehysmallit sekä riskienhallinnan näkökulma.* Tietojärjestelmätieteen Pro gradu -tutkielma. Jyväskylän yliopisto.

Liikenne- ja viestintäministeriö. (2014). *Big datan hyödyntäminen*. Liikenne- ja viestintäministeriön julkaisuja 20/2014.

Lublinsky, B., Smith, K. T. & Yakubovich, A. (2013). *Professional hadoop solutions*. Indianapolis: John Wiley & Sons.

Mainetti, L., Patrono, L. & Vilei, A. (2011). Evolution of wireless sensor networks towards the internet of things: A survey. *Software, Telecommunications and Computer Networks (SoftCOM), 2011 19th International Conference on,* (1-6).

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburg, C. & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Matusek, F. (2014). *Selective privacy protection for video surveillance.* Tieto- ja sähkötekniikan tiedekunnan väitöskirja. Oulun yliopisto.

Mayer-Schönberger, M. & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work and think* . New York: Houghton Mifflin Harcourt Publishing Company.

Microsoft. (2015, 14. kesäkuuta). HDInsight – hadoop big data solution |Microsoft azure. Haettu 14.6.2015 osoitteesta http://azure.microsoft.com/en-us/services/hdinsight/

Murthy, A. (2012, 15. elokuuta). Apache hadoop YARN – concepts and applications . Haettu 14.6.2015 osoitteesta http://hortonworks.com/blog/apache-hadoop-yarn-concepts-and-applications/

Navarro-Arribas, G. & Torra, V. (2012). Information fusion in data privacy: A survey. *Information Fusion, 13*(4), 235-244.

Negley, G. (1966). Philosophical views on the value of privacy. *Law and Contemporary Problems, 31*(2), 319-325.

Newell, P. B. (1995). Perspectives on privacy. *Journal of Environmental Psychology, 15*(2), 87-104.

Newell, P. B. (1998). A cross-cultural comparison of privacy definitions and functions: A system approach. *Journal of Environmental Psychology, 18*(4), 357-371.

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review, 57*, 1701-1777.

Park, S. H., Huh, S. Y., Oh, W. & Han, S. P. (2012). A social network-based inference model for validating customer profile data. *MIS Quarterly, 36*(4), 1217-1237.

Patil, M. R. & Thia, F. (2013). *Pentaho for big data analytics*. Birmingham: Packt Publishing.

Pedersen, D. M. (1979). Dimensions of privacy. *Perceptual and Motor Skills, 48*(3), 1291-1297.

Pedersen, D. M. (1997). Psychological functions of privacy. *Journal of Environmental Psychology, 17*(2), 147-156.

Pollach, I. (2007). What's wrong with online privacy policies? *Communications of the ACM, 50*(9), 103-108.

Popa, R. A., Redfield, C. M. S., Zeldovich, N. & Balakrishnan, H. (2011). CryptDB: Protecting confidentiality with encrypted query processing. *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles,* Cascais, Portugal (85-100). New York, NY, USA: ACM.

Prajapati, V. (2013). *Big data analytics with R and hadoop*. Birmingham: Packt Publishing.

Rivest, R. L. (1990). Cryptography. *Handbook of theoretical computer science: Algorithms and complexity* (1. painos, s. 719-756). Amsterdam: Elsevier Science Publishers B.V.

Rubinstein, I. (2012). *Big data: The end of privacy or a new beginning?* (Public Law Research Paper nro 12-56). NYU School of Law.

Saarenpää, A. (2009). *Oikeusjärjestys 2000* (6 painos, s. 299-451). Rovaniemi: Lapin yliopisto.

Sagiroglu, S. & Sinanc, D. (2013). Big data: A review. *Collaboration Technologies and Systems (CTS),* San Diego, CA (42-47).

Salmi, M. (2015). *Massadatan matalavasteaikainen analyysijärjestelmä.* Tietotekniikan koulutusohjelma Diplomityö. Oulun yliopisto.

Salo, I. (2013). *Big data - tiedon vallankumous*. Jyväskylä: Docendo.

Schmidt, K. & Phillips, C. (2013). *Programming elastic MapReduce*. Sebastopol: O'reilly Media.

Shvachko, K., Kuang, H., Radia, S. & Chansler, R. (2010). The hadoop distributed file system. *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on,* (1-10).

Smith, M., Szongott, C., Henne, B. & von Voigt, G. (2012). Big data privacy issues in public social media. *Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on,* Campione d'Italia (1-6).

Spiekermann, S. & Cranor, L. F. (2009). Engineering privacy. *IEEE Transactions on Software Engineering, 35*(1), 67-82.

Tavani, H. T. (1999). Privacy online. *ACM SIGCAS Computers and Society, 29*(4), 11-19.

Thomson, J. (1975). The right to privacy. *Philosophy and Public Affairs, 4*(4), 295-314.

Tietosuojaryhmä. (2014). *Lausunto 5/2014 anonymisointitekniikoista.* (nro 0829/14/FI WP 216).

Traver, H. (1984). Orientations toward privacy in hong kong. . *Perceptual and Motor Skills, 59*, 635-644.

van Dijk, M., Gentry, C., Halevi, S. & Vaikuntanathan, V. (2010). Fully homomorphic encryption over the integers. Teoksessa H. Gilbert (toim.), (s. 24-43) Springer Berlin Heidelberg.

Wang, J., Zhao, Y., Jiang, S. & Le, J. (2009). Providing privacy preserving in cloud computing. *Test and Measurement, 2009. ICTM '09. International Conference on,* (213-216).

Wang, W., Li, Z., Owens, R. & Bhargava, B. (2009). Secure and efficient access to outsourced data. *Proceedings of the 2009 ACM Workshop on Cloud Computing Security,* Chicago, Illinois, USA (55-66). New York, NY, USA: ACM.

Ward, J. S. & Barker, A. (2013). *Undefined by data: A survey of big data definitions.* arXiv preprint arXiv:1309.5821.

Warren, S. & Brandeis, L. (1890). The right to privacy. *Harvard Law Review, 4*(5), 193.

Wasserman, S. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Westin, A. (1967). *Privacy and freedom*. New York: Atheneum.

Zheleva, E. & Getoor, L. (2008). Preserving the privacy of sensitive relationships in graph data. Teoksessa F. Bonchi, E. Ferrari, B. Malin & Y. Saygin (toim.), (s. 153-171) Springer Berlin Heidelberg.