Antti Ruippo

EETTISET ONGELMAT MASSADATAN JA TIEDON-LOUHINNAN HYÖDYNTÄMISESSÄ

JYVÄSKYLÄN YLIOPISTO
TIETOJENKÄSITTELYTIETEIDEN LAITOS
2016

TIIVISTELMÄ

Ruippo, Antti
Eettiset ongelmat massadatan ja tiedonlouhinnan hyödyntämisessä
Jyväskylä: Jyväskylän yliopisto, 2016, 39 s.
Tietojärjestelmätiede, kandidaatintutkielma
Ohjaaja: Seppänen, Ville

Niin julkiset kuin yksityisetkin organisaatiot keräävät ja hyödyntävät yhä enemmän dataa osana toimintaansa. Big data eli massadata on keskeinen teknologia suurten datamäärien hallitsemiseksi. Massadataa analysoidaan puolestaan usein tiedonlouhinnan menetelmin. Molempien innovaatioiden suosio on kasvanut trendinomaisesti ja niiden käytön ennustetaan yleistyvän jatkossakin. Massadataa ja tiedonlouhintaa on hyödynnetty myös arveluttaviin tarkoituksiin. Julkisuudessa on ollut esillä tapauksia, joissa yritykset ovat kohdentaneet markkinointiaan selvittämällä arkaluonteista tietoa asiakkaistaan. Tämä yhdistettynä ajatukseen uusiin teknologioihin liittyvistä käytäntötyhjiöistä toimii motiivina aihevalinnalle. Tämän tutkielman tarkoituksena on selvittää, mitä eettisiä haasteita liittyy massadatan ja tiedonlouhinnan hyödyntämiseen. Lisäksi tutkielmassa taustoitetaan molempia teknologioita sekä niiden käsittelyssä sovellettavia etiikan teorioita. Tutkielman tutkimusmetodi on kirjallisuuskatsaus. Tuloksina havaittiin, että moniin massadatan ja tiedonlouhinnan hyödyntämisen vaiheisiin liittyy erinäisiä eettisiä ongelmia. Sekä yksilöä että yhteiskuntaa koskevia haasteita löydettiin. Tulokset koostettiin taulukkomuotoiseen esitykseen, joka toimii luonnoksena kyseisten teknologioiden hyödyntämisen eettiseksi viitekehykseksi.

Asiasanat: big data, tiedonlouhinta, etiikka

ABSTRACT

Ruippo, Antti
Ethical issues in the utilization of Big Data and Data Mining
Jyväskylä: University of Jyväskylä, 2016, 39 p.
Information Systems, Bachelor's Thesis
Supervisor: Seppänen, Ville

Both public and private organizations collect and utilize more and more data as part of their activities. Big data is a key technology for managing large amounts of data. Big data is often analyzed through data mining methods. Both innovations have become increasingly popular, and their use is predicted to become even more common in the future. Big data and data mining have also been used in dubious ways. Cases have been reported where companies have targeted their marketing efforts by identifying sensitive information of their customers. This, combined with the thought of policy vacuums related to new technologies, serves as motif for the choice of topic. The purpose of this thesis is to find out what ethical challenges are related to the utilization of big data and data mining. In addition, the background of these technologies and ethical theories applied in reviewing them are summarized. The study is performed through a literature review. The results showed that there are a number of ethical problems related to different stages in the use of big data and data mining. Both problems dealing with individuals and society were found. The results were compiled in a table, which acts as an ethical framework for the use of these technologies.

Keywords: big data, data mining, ethics

TAULUKOT

SISÄLLYS

1 JOHDANTO

Big datan eli massadatan hyödyntäminen tavalla tai toisella organisaatioiden toiminnassa yleistyy kiihtyvällä vauhdilla. Suurten datamassojen analysoimiseen käytetty menetelmä, tiedonlouhinta, kasvattaa myös suosiotaan. (Hashem ym., 2015; Liikenne- ja viestintäministeriö, 2014; Tene & Polonetsky, 2012.) Samanaikaisesti huoli yksityisyydensuojasta kasvaa. Esimerkki arveluttavasta liiketoiminnasta massadataan liittyen on yhdysvaltalaisen Target-tavaratalon halu ja kyky saada selville tietoja naisasiakkaidensa raskauksista analysoimalla heidän ostokäyttäytymistään (Duhigg, 2012).

Moor (1985) argumentoi, että informaatioteknologian kehitys tuo mukanaan uusien toimintamahdollisuuksien vuoksi ”käytäntötyhjiöitä”. Näitä puutteita täydennetään siis usein vasta teknologioiden käyttöönoton jo alettua. Ajatus siitä, että uudet teknologiat vaativat eettistä pohdintaa jonkinlaisten sääntöjen laatimiseksi, on yksi perusteluistani tämän tutkielman tarpeelle. Moor (1985) väittää kuitenkin, että eettisten ohjeiden kehittäminen edellyttäisi jonkinlaista viitekehystä käsillä olevan teknologian eettisen tarkastelun pohjaksi. Tästä syystä tavoitteenani on hahmotella tällainen eettisen ulottuvuuden sisältävä viitekehys massadatan ja tiedonlouhinnan ongelmien tarkastelua varten.

Massadataan ja tiedonlouhintaan liittyvien eettisten ongelmien kartoittaminen on mielestäni erityisen mielekästä siksi, että uusien teknologioiden hyödyntämiseen liittyvä lainsäädäntö laahaa edellä kuvailtujen käytäntötyhjiöiden vuoksi teknisen kehityksen perässä. Tämän vuoksi pelkkä lakiin nojaaminen hyvien käytänteiden erottamiseksi huonoista ei ole tarkoituksenmukaista. Mason, Mason ja Culnan (1995, 161) korostavatkin, että tietojenkäsittelyn parissa työskentelevien vastuulla on pohtia toimiensa eettisyyttä suhteuttamalla työnantajiensa, asiakkaidensa ja muiden ihmisten oikeudet ja tarpeet toisiinsa. Tämän tutkielman tuloksena syntyvää viitekehystä on siis tarkoitus pystyä käyttämään pohjana ammattieettisten ongelmien kartoittamiseen massadataan ja tiedonlouhintaan liittyen.

Han, Kamber ja Pei (2011, 32–33) pitävät tärkeänä, että tiedonlouhinnan yhteiskunnallisia vaikutuksia tutkitaan, sillä kyseessä on kaikkia ihmisiä arjessa koskettava teknologia. Davis (2012, 62) puolestaan väittää massadatan olevan

yksi aikamme mullistavimmista teknologioista, jolla on vaikutusta muiden muassa liike-elämään, yhteiskuntaan ja terveydenhuoltoon. Nämä näkemykset tukevat tämän tutkielman tarpeellisuutta.

Culnan ja Armstrong (1999) esittävät empiiristen tutkimustulosten pohjalta, että organisaatiot voivat saada liiketoiminnallista hyötyä tiedon keräämisestä eettisesti ja läpinäkyvästi. Kyseisessä tutkimuksessa havaittiin, että henkilökohtaisen informaation keräämiseen suhtauduttiin suopeimmin silloin, kun koehenkilöille kerrottiin tavoitteesta tutkia reiluja tiedonkeruun käytäntöjä (Culnan & Armstrong, 1999). Juuri tietojen kerääminen ihmisistä on olennainen osa massadatan tarkoitusta, joten eettinen toiminta voi hyvinkin olla organisaation etujen mukaista massadataa kerättäessä. Täten etu liiketoiminnalle on yksi painava peruste eettisesti kestävän massadatan ja tiedonlouhinnan käytön tutkimiselle.

Tämän kandidaatintutkielman tarkoitus on siis kartoittaa massadataan ja tiedonlouhintaan liittyviä eettisiä ongelmia. Tutkimus suoritetaan kirjallisuuskatsauksen keinoin. Tutkielman lähdemateriaalina on pyritty hyödyntämään monipuolisesti massadataa ja tiedonlouhintaa käsittelevää tutkimusta sekä tietojenkäsittelyn etiikan kirjallisuutta. Lisäksi eettistä pohdintaa on täydennetty etiikan klassikoiden erilaisilla näkökulmilla moraaliseen toimintaan.

Tutkielmani ensimmäinen tutkimuskysymys on: ”Mitä eettisiä ongelmia liittyy massadatan ja tiedonlouhinnan hyödyntämiseen organisaation toiminnassa?” Tähän kysymykseen vastaamalla pyrin tuomaan esille kattavasti huomionarvoisia aiheeseen liittyviä eettisiä ongelmia. Vastatakseni ensimmäiseen tutkimuskysymykseen minun täytyy myös taustoittaa kyseisiä ilmiöitä asianmukaisella tarkkuudella.

Toinen tutkimuskysymykseni on: ”Mitä yksilöitä ja yhteiskuntaa koskevia teemoja tulisi huomioida massadataa ja tiedonlouhintaa koskevassa eettisessä pohdinnassa?” Tähän kysymykseen vastaaminen vaatii tutkimuskirjallisuudessa esitettyjen huolenaiheiden koostamista ja jäsentämistä sovellettavissa olevaan muotoon. Käytännössä tämä tarkoittaa siis jo mainitun eettisen viitekehyksen laatimista massadataa ja tiedonlouhintaa varten. Tarkoitukseni ei ole suinkaan pyrkiä kertomaan, mikä kaikki toiminta on moraalisesti hyväksyttävää ja mikä ei liittyen näiden teknologioiden hyödyntämiseen. Viitekehys tarjoaa ennen kaikkea työkalun eettisen toiminnan pohtimiseksi. Onkin hyvä muistaa, että eettiseen pohdintaan sisältyy vahvasti ajatus siitä, että erilaisilla ajattelutavoilla voi perustellusti päätyä eri valintoihin (Mingers & Walsham, 2010).

Tutkimuskysymysten luonteen vuoksi tutkielmassa ei ole tarkoitus mennä teknisessä mielessä yksityiskohtaiselle tasolle tutkittavien teknologioiden osalta. Koska tarkoituksena on kehittää massadatan ja tiedonlouhinnan eettisiä ongelmia kartoittava viitekehys, jää teknisiä ratkaisuja koskevien johtopäätösten tekeminen viitekehystä mahdollisesti hyödyntävälle ammattilaiselle. Pyrkimykseni on siis osoittaa ja koostaa joitakin tutkimuskirjallisuuden tuntemia aiheeseen liittyviä eettisiä kysymyksiä, jotka voisivat toimia pohjana ammattieettiselle pohdinnalle ja lisätutkimukselle.

Seuraavassa eli toisessa luvussa tarkastellaan kirjallisuudessa esitettyjä erilaisia määritelmiä massadatalle ja tiedonlouhinnalle. Lisäksi kerrotaan näiden teknologioiden toiminnasta ja käyttötarkoituksista sekä taustoitetaan niiden historiaa. Kolmannessa luvussa käydään läpi etiikan keskeisiä teorioita ja niiden kysymyksenasetteluja sekä perehdytään tarkemmin tietojenkäsittelyn etiikkaan. Neljännessä luvussa koostetaan ja pohditaan massadataan ja tiedonlouhintaan liittyviä eettisiä ongelmia. Tässä yhteydessä kerrotaan myös aiheeseen liittyvistä tosielämän esimerkkitapauksista sekä esitellään viitekehys massadataa ja tiedonlouhintaa koskevien eettisten ongelmien tarkastelemiseksi. Viidennessä luvussa tehdään yhteenveto tutkielman tärkeimmästä sisällöstä ja havainnoista. Lisäksi arvioidaan, miten tutkimuskysymyksiin vastattiin ja mitkä ovat tutkielman vahvuuksia ja heikkouksia.

2 MASSADATA JA TIEDONLOUHINTA

Tässä luvussa tarkastelen, mitä massadata ja tiedonlouhinta tarkemmin ottaen ovat ja mitä sukulaiskäsitteitä näillä on. Tarkoituksena on antaa kattava joskin tiivis selvitys siitä, mihin näitä teknologioita käytetään ja mihin periaatteisiin näiden toiminta perustuu. Käsitemäärittelyn pohjana hyödynnän aiheesta tehtyä tutkimustietoa.

2.1 Massadatan eli big datan määrittelystä

Big data on käsite, jolla viitataan niin suuriin tietomassoihin, ettei niiden analysointiin ja hallintaan ole mielekästä käyttää perinteisiä tietokantojen hallintamenetelmiä (Jacobs, 2009; Ward & Barker, 2013). *Massadata* on suositeltu ja tässäkin tutkielmassa käytettävä suomennos termille, joka voidaan kääntää myös muotoon *iso data* (Sanastokeskus TSK, 2013a). Crawfordin ja Schultzin (2014) mukaan big data on epätarkka ja yleisluontoinen termi. Vaihtelevien käsitysten vuoksi on tarpeen selvittää, millaisia tuntomerkkejä massadataan on liitetty. Seuraavaksi tarkastellaan joitakin ehdotuksia massadatan tarkemmaksi määritelmäksi.

Tutkimuslaitos Gartnerin suositussa määritelmässä massadataan liitetään neljä keskeistä tunnusmerkkiä. Tähän ”neljän V:n” määritelmään kuuluvat sanat volume (määrä), velocity (nopeus), variety (monipuolisuus) ja veracity (totuudenmukaisuus). Viimeisenä mainitulla totuudenmukaisuudella tarkoitetaan itse asiassa massadatalle tyypillistä luotettavuuden kyseenalaisuutta. (Gartner 2012 Wardin & Barkerin 2013, 1, mukaan.) Ward ja Barker (2013) huomauttavat, että Gartnerin kehittämä määritelmä massadatalle on luonteeltaan varsin anekdoottinen eikä tarjoa kvantitatiivista mittaristoa väitteiden tueksi. Näin ollen kyseistä määritelmää ei ole tarkoituksenmukaista käyttää akateemisen tarkastelun pohjana.

Kitchin (2013) erottaa tutkimuskirjallisuudesta seitsemän tuntomerkkiä, jotka liitetään yleisimmin massadataan, ja ne ovat: suuri koko, nopeus, moni-

puolisuus, perusteellisuus, yksityiskohtaisuus, suhteellisuus ja joustavuus. Big dataan liitetään siis tutkimuksessa hyvinkin monia erilaisia ominaisuuksia. Floridi (2012) huomauttaa, että massadatan määrittelyyn perinteisesti liitetty ajatus hyödyllisen datan liian suuresta määrästä suhteessa käytettävissä oleviin analysointiresursseihin on epälooginen. Analysointiongelmat kumpuavat pikemminkin vaikeudesta määritellä, mikä osa suuresta datamassasta on käyttökelpoista ja mikä ei (Floridi, 2012).

Toisenlaisen näkökulman massadatan määrittelyyn tuovat boyd ja Crawford (2012), joiden mukaan kyseessä on teknologian, analysointimenetelmien ja mytologiaa muistuttavien uskomusten vuorovaikutuksesta koostuva ilmiö. Merkillepantavaa tässä määritelmässä on datan analysoinnin sisällyttäminen massadatan käsitteeseen, mikä vaikuttaa olevan erottava tekijä monen määritelmän välillä. Tässä näkökulmassa kyseenalaistetaan myös laajalle levinnyt usko massadatan analysoinnin tarjoamien näkemyksien ylivertaisuuteen aiempiin analysointimenetelmiin nähden (boyd & Crawford, 2012).

Massadatan määrittelyn vaikeudesta kertoo omaa kieltään sekin, etteivät monet organisaatiot, jotka kertovat hyödyntävänsä kyseistä teknologiaa, tiedä mitä massadata oikeastaan on. Kitchin ja McArdle (2016) argumentoivat empiirisen tutkimuksen pohjalta, että vain harvat massadataksi kutsutut tietovarannot itse asiassa täyttävät tärkeimpiä kriteerejä massadatalle ja termiä käytetään kevyin perustein.

Vaikka kookkaita datamassoja on ollut olemassa jo ennen massadatan aikakautta, ei historian tuntemia valtavia tietovarantoja voida useinkaan luonnehtia termillä massadata. Havainnollistavana esimerkkinä tästä voidaan käyttää väestönlaskentaprosessin tuottamia valtavia datamassoja, jotka eivät kuitenkaan täytä koon lisäksi muita massadatan kriteerejä, kuten joustavuutta ja nopeutta (Kitchin & McArdle, 2016). Massadatalla viitataan siis ennen kaikkea sellaisiin datamassoihin, joiden keräämisessä, tallentamisessa ja organisoinnissa hyödynnetään hyvin modernia teknologiaa.

Davenport, Barth ja Bean (2012) esittävät, että organisaation tavoitteiden kannalta massadata ja sen analysointimenetelmät eroavat perinteisemmästä data-analytiikasta kolmesta syystä. Ensinnäkin massadatassa huomio on yhä enemmän datavirroissa eikä -varastoissa. Toisekseen massadataa analysoivat datatieteilijät data-analyytikoiden sijaan. Kolmanneksi massadatan analysointi on tiiviimmin kytköksissä organisaation ydinliiketoimintaan kuin perinteinen data-analytiikka. (Davenport, Barth & Bean, 2012.)

Edellä olevan taustoituksen pohjalta tässä tutkielmassa massadatalla tarkoitetaan teknologista kokonaisuutta, jossa kerätään, kootaan ja talletetaan niin suuria datamassoja niin nopealla tahdilla, etteivät perinteiset tietokantaproseduurit riitä niitä hallitsemaan. Seuraavassa alaluvussa tarkastelen lähemmin, miten ja mihin tarkoituksiin massadataa hyödynnetään organisaatioissa.

2.2 Massadatan käytännön hyödyntäminen

Massadataa hyödynnetään käytännössä hyvin monilla aloilla. Liikenne- ja viestintäministeriö (2014) kuvailee big data -strategiassaan merkittäviä sovellusalueita olevan mm. terveysala, markkinointi, tutkimus ja julkishallinnon tehostaminen. Massadataa onkin kuvailtu teknologiana ainutlaatuisen monimuotoiseksi monien käyttötarkoitustensa ja datamassojen mittasuhteiden vuoksi (Davis, 2012).

Pilvipalveluita tarjoavien yritysten suosima liiketoimintamalli on tarjota yritysasiakkailleen palvelintilaa ja analysointipalveluita massadataa varten (Kshetri, 2014). Tällaisessa mallissa asiakasyrityksen ei tarvitse rakentaa itse tarvittavaa infrastruktuuria vaan se voi ulkoistaa massadatan keräämisen, säilyttämisen ja analysoinnin ulkoiselle palveluntarjoajalle.

Massadataa voidaan hyödyntää lääketieteessä muun muassa eri lääkkeiden yhteisvaikutusten tutkimiseksi aiempaa tehokkaammin. Vertailemalla lääkkeiden käyttäjistä kerättyjä suuria datamassoja ristiin on onnistuttu esimerkiksi huomaamaan jopa miljoonan yhdysvaltalaisen potilaan olevan vaarassa sairastua diabetekseen lääkkeiden sivuvaikutusten vuoksi (Tene & Polonetsky, 2012). Massadataa analysoimalla voidaan siis paitsi tehdä liikevoittoa myös edistää kansanterveyttä.

Einav ja Levin (2014) esittävät massadatan tarjoavan ennennäkemättömän mahdollisuuden ekonomisteille päästä käsiksi suuriin ja yksityiskohtaisiin kansantaloutta koskeviin datamassoihin muun muassa aiempaa parempien ennusteiden laatimiseksi. Suuria tutkimusaineistoja hyödyntävänä tieteenalana taloustiede on luonteva sovellusalue massadatalle sekä sen analysointimenetelmille.

Taloustieteen lisäksi myös monilla muilla tieteenaloilla voidaan tunnistaa hyödyllisiä sovellutustapoja suurten datamassojen analysoimiselle. Kitchin (2014) kuvailee massadataa disruptiiviseksi innovaatioksi, jonka merkittävät vaikutukset tutkimuksen tekemiseen vaativat lisätutkimusta. Tällä sovellusalueella massadatan hyödyntämisen problematiikkaa on tarpeen tarkastella erityisesti tutkimusetiikan näkökulmasta.

On arvioitu, että massadataa hyödyntämällä voidaan tehostaa julkishallinnon toimintaa Suomessa. Massadata-analytiikan uskotaan tarjoavan mahdollisuuden mitata julkisten toimijoiden tehokkuutta entistä tarkemmin. Lisäksi yksilöllisempien palveluiden tarjoaminen ja kansalaisten osallistaminen näiden kehitykseen ovat massadatan ennakoituja hyötyjä. (Liikenne- ja viestintäministeriö, 2014.)

2.3 Tiedonlouhinta eli data mining käsitteenä

Data mining tarkoittaa menetelmiä, joilla pyritään löytämään mielenkiintoisia säännönmukaisuuksia ja muuta uutta tietoa erityisesti suurista datamassoista

(Han ym., 2011; Hand, Mannila & Smyth, 2001). Tässä tutkielmassa termistä käytetään Sanastokeskus TSK:n (2013b) suosittelemaa suomennosta *tiedonlouhinta*, jonka lisäksi synonyymina voidaan käyttää myös muotoa *tiedonrikastus*.

Hanin ym. (2011) mukaan nimitys ”data mining” on harhaanjohtava, sillä tavoitteena ei ole datan louhinta sinänsä vaan tietämyksen louhinta datasta. Tässä mielessä suomennos ei ole suora käännös englanninkielisestä termistä. Sanan ”data” korvaaminen ”tiedolla” suomenkielisessä termissä kuvaakin totuudenmukaisemmin, millaista toimintaa termillä tarkoitetaan.

Tiedonlouhinnan avulla pyritään löytämään datamassoista ennestään tuntemattomia säännönmukaisuuksia, kuten toistuvia kuvioita aikasarjoista, klustereita ja puurakenteita. Tiedonlouhinnan erityispiirre on lisäksi se, että louhittavaa dataa ei useinkaan ole kerätty varta vasten tiedonlouhintaa varten vaan johonkin muuhun tarkoitukseen. (Hand ym., 2001.) Quinnin (2014) mukaan tiedonlouhinta on yhden tai useamman tietokannan läpikäymiseen käytetty menetelmä, jonka tarkoituksena on tuottaa uutta tietoa hyvin suuresta määrästä erilaisia transaktioita. Usein toistuvia piirteitä tiedonlouhinnan määrittelyissä ovat siis käytetyn aineiston suuruus ja louhimalla saadun tiedon hyödyllisyys.

Erään näkemyksen mukaan tiedonlouhinta on vain yksi monista askelista prosessissa, jonka tavoitteena on uuden tiedon löytäminen datamassoista. Tätä yläprosessia kutsutaan nimellä *knowledge discovery in databases* (KDD), mikä tarkoittaa suomeksi tietämyksen löytämistä tietokannoista. (Fayyad, Piatetsky-Shapiro & Smyth, 1996.) Sittemmin termi tiedonlouhinta on vakiintunut laajempaan käyttöön, ja Han ym. (2011) perustelevatkin selvyydellä käsitteen käyttämistä viittaamaan koko prosessiin.

Edelleen hämmennystä voi aiheuttaa muun muassa Mitchellin (1999) esittämä näkemys siitä, että tiedonlouhinnan prosessi käsittää myös datan kokoamisen yhteen tietokantaan ennen varsinaista analysointia. Näin tiedonlouhinnan voidaan nähdä limittyvän massadatan käsitteen kanssa. Selvyyden vuoksi tässä tutkielmassa tiedonlouhinnan menetelmän ei katsota sisältävän datan keräämiseen liittyviä toimenpiteitä.

Machine learning eli koneoppiminen on toinen läheisesti tiedonlouhintaan liittyvä käsite. Koneoppiminen tarkoittaa tietokonejärjestelmän ohjeistamista siten, että järjestelmä voi muuttaa toimintaansa jonkin tehtävän ratkaisun edetessä ilman erillisiä ihmisen antamia lisäkäskyjä. Termillä viitataan sekä itse ilmiöön että sitä tutkivaan tietotekniikan tutkimushaaraan. (Carbonell, Michalski & Mitchell, 1983.) Koneoppimisen ja tiedonlouhinnan suhde toisiinsa on varsin tiivis, sillä toisinaan näitä termejä kerrotaan käytettävän jopa toistensa synonyymeina (Mitchell, 1999). Nämä ovat monista yhtäläisyyksistään huolimatta kuitenkin toisistaan erotettavissa olevia ilmiöitä. Eroavaisuuksina mainitaan muun muassa tiedonlouhinnan parempi skaalautuvuus ja keskittyminen suurempiin datamassoihin kuin koneoppimisen tapauksessa. (Han ym., 2011.)

Lisäarvon tuottamiseen liiketoiminnalle käytettävistä data-analytiikan työkaluista käytetään usein nimitystä *business intelligence* (BI). Tämä käsite sisältää monien analyysimenetelmien joukossa myös tiedonlouhinnan. (Negash, 2004.) Tässä tutkielmassa tarkasteltavat teknologiat rajautuvat kuitenkin mas-

sadataan ja tiedonlouhintaan. Näin ollen business intelligencen kokonaisuus ei kuulu tutkielman aihepiiriin, vaikka massadata ja sen analysoiminen ovatkin tärkeitä BI:n osatekijöitä.

Tiedonlouhinnalla tarkoitetaan tässä tutkielmassa siis teknologiaa ja analysointimenetelmiä, joilla voidaan löytää tilastollisia säännönmukaisuuksia erityisesti massadatasta. Tiedonlouhinnassa voidaan hyödyntää tämän määrittelyn mukaan esimerkiksi koneoppimisen metodeja, joskin nämä ovat kaksi omaa käsitettään. Seuraavassa alaluvussa käydään lyhyesti läpi teknologiaa ja menetelmiä, joita tiedonlouhinta hyödyntää massadatan analysoimisessa.

2.4 Tiedonlouhinnan hyödyntäminen massadatan analysoimiseksi

Tiedonlouhinta on monesta vaiheesta koostuva ja monimutkainen prosessi. Alla kuvaillaan lyhyesti seuraavat siihen liittyvät käsitteet: esiprosessointi, tietovarastointi, OLAP, datakuutio ja klusterointi. Tämä ei ole missään nimessä tyhjentävä luettelo tiedonlouhintaan liittyvistä tekniikoista, vaan ennemmin lyhyt katsaus tiedonlouhinnan luonteeseen teknologiana. Lisäksi selitetään lyhyesti, millaisia löydöksiä tiedonlouhinta tuottaa.

Koska massadata on usein hankalasti hyödynnettävissä sellaisenaan, on dataa tavallista esiprosessoida ennen varsinaista tiedonlouhintaa (Han ym., 2011). Tässä vaiheessa datasta poistetaan kohinaa ja yhdenmukaistetaan eri lähteistä kerättyä dataa. Esiprosessointi koostuu tyypillisesti datan puhdistamisesta, integroimisesta ja muuntamisesta sekä datasetin pienentämisestä. (Han ym., 2011.)

Data warehouse eli tietovarasto on organisaation toiminnalle tärkeää informaatiota kokoava tietovaranto, joka on yleensä erillään organisaation operatiivisista tietokannoista. Tietovarasto on koottu erityisesti analysointitarkoituksiin, ja tätä koontia kutsutaan tietovarastoinniksi. (Han ym., 2011.)

OLAP (lyh. sanoista *Online Analytical Processing*) on erityisesti tietovarastojen analysoimiseen tarkoitettu prosessointimenetelmä, joka kykenee käsittelemään erittäin suuria määriä dataa. Analyysitulokset ovat erityisesti organisaation päätöksenteon tueksi tarkoitettuja, ja ne ovat jokseenkin kaukana asiakasrajapinnasta. (Han ym., 2011.)

Data cube eli datakuutio tarkoittaa tapaa organisoida suuri datamassa siten, että datan eri ominaisuudet muodostavat omia ulottuvuuksiaan. Tämä mahdollistaa kyselyiden muodostamisen ja datan analysoimisen sen ominaisuuksien mukaan. Datakuutio on yksi tapa järjestää tietovarasto. (Han ym., 2011.)

Klusteroinnilla tarkoitetaan tiedonlouhinnassa menetelmää, jossa keskenään samankaltaisia piirteitä omaavat datan osaset järjestetään omiksi osajoukoikseen, klustereiksi. Saman datamassan eri klusterit ovat siis toisistaan poikkeavia ominaisuuksiltaan. (Han ym., 2011.) Tätä menetelmää hyödyntämällä

voidaan siis esimerkiksi tutkia, mitä muita ominaisuuksia yhden ominaisuuden yhdistämillä osajoukoilla on.

Muita tärkeitä löydöstyyppejä tiedonlouhinnan tavoitteiden kannalta ovat usein toistuvat säännöllisyydet (engl. *frequent patterns*), assosiaatiot ja korrelaatiot. Ensimmäisenä mainitut ovat datasetissä useasti esiintyviä säännönmukaisuuksia. Assosiaatiot ovat todennäköisyyksiin perustuvia sääntöjä, jotka liittyvät kahden tai useamman eri datakomponentin keskinäiseen suhteeseen. Korrelaatiot puolestaan ovat tilastollisesti merkittäviä riippuvuussuhteita, joita datasetistä saattaa löytyä eri tekijöiden väliltä. (Han ym., 2011.)

2.5 Muita tutkielman kannalta tärkeitä käsitteitä

Yksi keskeinen käsite tälle tutkielmalle on yksityisyys, joka on yksilön näkökulmasta ehkä suurin huolenaihe massadataan liittyen. Yksityisyyden käsite on varsin monitulkintainen ja laaja, eikä tämän tutkielman rajoissa ole mielekästä tarkastella eettisiä ongelmia monen eri yksityisyyden määritelmän näkökulmasta. Tämän johdosta määritelmä perustuu tässä yhteydessä Quinnin (2014) esittämään määrittelyyn, joka on pohjana hänen omassa informaatioteknologian ja yksityisyyden välisessä pohdinnassa. Hänen mukaan yksityisyys on sosiaalinen järjestely, joka mahdollistaa yksilöiden vaikuttamisen siihen, ketkä pääsevät käsiksi heidän yksityistietoihin ja fyysiseen minään (Quinn 2014, 176). Puhuttaessa yksityisyyden vaarantavista toimista tarkoitetaan keinoja, jotka heikentävät yksilön vaikutusmahdollisuuksia näihin asioihin.

Anonymisointi on sekä perinteisissä tietokannoissa että massadatassa käytetty yksityisyyden suojelun periaate, johon voidaan pyrkiä eri menetelmin (Sedayao, Bhardwaj & Gorade, 2014). Tietosuojaryhmän (2014) mukaan näiden menetelmien tärkeimpiä luokkia ovat satunnaistaminen ja luokituksen karkeistaminen. Satunnaistaminen koostuu joukosta menetelmiä, joilla lisätään epävarmuutta dataan yksilöitä koskevan päättelyn vaikeuttamiseksi. Luokituksen karkeistaminen puolestaan perustuu datan erilaisten luokittelujen muuttamiseen epätarkemmiksi. (Tietosuojaryhmä, 2014.) Näiden yläkäsitteiden piiriin kuuluu siis lukuisia eri menetelmiä, joilla on omat vahvuutensa ja heikkoutensa. Tämän tarkempaan tekniseen selvitykseen ei ole kuitenkaan tämän tutkielman rajoissa mielekästä mennä.

Differentiaalinen yksityisyys (engl. *differential privacy*) on satunnaistamismenetelmä, jolla pyritään takaamaan henkilön anonymiteetti tietokannassa (Dwork, 2006). Toisin sanoen tietokannasta ei pitäisi pystyä yksilöimään kyselyillä tai analyyseilla tiettyyn henkilöön liittyviä datajoukkoja. Esimerkkinä menetelmän tavoitteesta mainitaan, että henkilöltä ei voida evätä oikeutta saada vaikkapa vakuutusta vain tietokannassa esiintymisen vuoksi (Dwork, 2006).

Yksityisyyden säilyttävä tiedonlouhinta (engl. *privacy-preserving data mining, PPDM)* on nimensä mukaisesti nimitys, jota käytetään erityisesti yksityisyyden varjeluun tähtäävistä tiedonlouhintamenetelmistä (Aggarwal & Yu, 2008). Mutasen (2007) mukaan yksityisyyden säilyttävän tiedonlouhinnan käy-

tännön menetelmät eivät sinänsä eroa tavallisista, vaan erona on yksityisyyden varjelun painotus työkalujen valinnassa. Motivaationa niiden kehittämiselle on toiminut huoli siitä, etteivät tiedonlouhinnasta kiinnostuneet organisaatiot ole olleet valmiita vaarantamaan tietovarastoissaan olevien käyttäjätietojen yksityisyyttä (Huang, Du & Chen, 2005). Tällaisen huolenaiheen olemassaolo kertoo tiedonlouhinnan eettisten ongelmien pohdinnan olevan tärkeää.

Varjopuolena yksityisyyden säilyttävän tiedonlouhinnan menetelmät tyypillisesti muuntelevat tietoa anonymisointimenetelmin yksityisyyden säilyttämiseksi, jolloin tuloksena saadun informaation tarkkuus voi kärsiä (Aggarwal & Yu, 2008). Tämä seikka voi vähentää yksityisyyden säilyttävän tiedonlouhinnan houkuttelevuutta organisaation näkökulmasta. Huang ym. (2005) väittävät sen olevan kuitenkin tärkeä tutkimuskohde yhteiskunnan, tieteen ja jopa kansallisen turvallisuuden näkökulmista.

3 ETIIKKA JA TIETOJENKÄSITTELY

Tässä luvussa kuvaillaan tietojenkäsittelyn etiikkaa tutkimusalana ja selvitetään, miten se sijoittuu laajemmassa perspektiivissä etiikan kentälle. Lisäksi tarkastellaan, miten tietojenkäsittelyn etiikka soveltuu tutkimusongelman ratkaisemiseen. Viimeisessä alaluvussa käsitellään IT-alan ammattietiikkaa ja miten kirjallisuutta aiheesta sovelletaan myöhemmin esiteltävään eettiseen viitekehykseen. Ensin on kuitenkin syytä taustoittaa etiikan eri haaroja ja teorioita.

3.1 Etiikasta yleisesti

Etiikka ymmärretään yleisesti moraalin filosofiseksi tutkimukseksi, joka on kiinnostunut ajatusrakenteista ja perusteluista ihmisten moraalisen käyttäytymisen taustalla (Quinn, 2014; Sterba, 1998). Etiikan kenttä jakautuu kolmeen eri haaraan, jotka ovat metaetiikka, soveltava etiikka ja normatiivinen etiikka. Lisäksi moraalifilosofian alaan voidaan lukea kuuluvan myös deskriptiivinen etiikka, joka on kiinnostunut ihmisten eettisen toiminnan ja arvojen kartoittamisesta. (Stahl, Timmermans & Mittelstadt, 2016.) Seuraavissa alaluvuissa tarkastellaan lähemmin, mitä nämä etiikan eri osa-alueet pitävät sisällään.

3.1.1 Metaetiikka

Metaetiikka tutkii ennen kaikkea etiikan teorioita ja sen pyrkimyksenä on arvioida niiden keskinäisiä suhteita (Fieser, 2000). Tämän tutkielman kannalta metaetiikan olennaisinta antia on juurikin eri etiikan teorioiden vertailu tutkimusongelmien ratkaisun näkökulmasta. Pääpainotus neljännen luvun eettisessä pohdinnassa on kuitenkin soveltavalla ja normatiivisella etiikalla, joita taustoitetaan seuraavaksi.

3.1.2 Soveltava etiikka

Soveltava etiikka tarkastelee, miten joissakin tietyissä käytännön eettisissä ongelmatilanteissa tulisi toimia. Tarkoituksena on soveltaa etiikan eri teorioita monimutkaisten ja ratkaisemattomissakin olevien ongelmien pohdintaan ja tuottaa joitakin ohjenuoria näissä tilanteissa toimimista varten. (Moor, 1985; Marturano, 2002.)

Tietojenkäsittelyn etiikka ja monet ammattietiikan osa-alueet lukeutuvat soveltavan etiikan piiriin (Marturano, 2002). Tämän tutkielman tuloksena syntyvän eettisen viitekehyksen on tarkoitus toimia pohjana juuri IT-ammattilaisten pohdinnalle massadataan ja tiedonlouhintaan liittyen. Luvussa 3.2 taustoitetaan tarkemmin tietojenkäsittelyn etiikan teorioita ja historiaa.

3.1.3 Normatiivinen etiikka

Normatiivisen etiikan tavoitteena on tuottaa moraalisen toiminnan pohjaksi soveltuvia arvoja ja käytäntöjä (Fieser, 2000). Quinnin (2014) mukaan normatiivisen etiikan teorioista informaatioteknologian ongelmien pohtimiseen soveltuvat parhaiten velvollisuusetiikka, seurausetiikan kaksi eri lajia sekä yhteiskuntasopimusteoria. Näitä ennen esitellään kuitenkin lyhyesti hyve-etiikan periaatteita.

Hyve-etiikka on Aristoteleen oppeihin perustuva teoria, jonka mukaan kaiken moraalisen toiminnan täytyy pohjautua joihinkin perustavanlaatuisiin hyveisiin. Näiden hyveiden tulisi olla mahdollisimman ”puhtaita” ja yleisesti hyväksyttäviä. (Quinn, 2014.) Aristoteleen (1989) näkemyksen mukaan hyveisiin pohjautuvan toiminnan motivaationa on pyrkimys hyvään elämään ja onnellisuuteen. Quinnin (2014) mukaan hyve-etiikka voi toimia erityisen hyvin muiden etiikan teorioiden täydentäjänä.

Velvollisuusetiikka (kutsutaan myös nimityksellä deontologinen etiikka) taas korostaa eettisen toiminnan perustelemista tärkeiksi katsottuihin velvollisuuksiin vedoten (Quinn, 2014). Kant (1998) katsoo velvollisuuden kumpuavan halusta toimia lakia kunnioittaen ja että tällaisessa toiminnassa on tärkeintä teon perusteltavuus eikä niinkään päämäärät. Velvollisuusetiikan periaatteet kiteytyvät Kantin kehittämään usein lainattuun kategoriseen imperatiiviin. Tämä lauseke on muotoiltu englanniksi seuraavasti: ”Act only from moral rules that you can at the same time will to be universal moral laws.” (Quinn 2014, 21.) Ihmisten tulisi siis tämän säännön perusteella toimia siten, kuin he voisivat olettaa muidenkin moraalisesti oikein toimivien ihmisten toimivan.

Seurausetiikan teorioissa painotetaan tekojen hyvyyden arviointia niiden seurauksien perusteella. Esimerkki seurausetiikan opista on utilitarismi, jota on kahta eri lajia. (Quinn, 2014.) Kantin deontologiselle etiikalle vastakkaista näkemystä edustava tekoutilitarismi lähtee ajatuksesta, että hyvä teko maksimoi yhteistä hyvää teon itsensä laadusta huolimatta (Mill, 1998). Sääntöutilitarismi on tämän pohjalta kehitetty erillinen teoriansa. Sen ytimessä on moraalisääntöjen valitseminen niiden yleisen käyttöönoton tuottaman maksimaalisen hyödyn

perusteella. (Quinn, 2014.) Molemmille ajattelutavoille on siis oleellista hyvien seurauksien mahdollisimman suuri osuus huonoihin seurauksiin nähden.

Yhteiskuntasopimusteorian ydinajatus on, että moraalisäännöt hyväksyttävälle ja ei-hyväksyttävälle toiminnalle perustuvat yhteisen hyvän edistämisen periaatteeseen (Quinn, 2014). Merkittävä tämän etiikan haaran teoreetikko Rawls (1998) on kehittänyt kaksi oikeusperiaatetta. Ensinnäkin ihmisillä tulisi olla mahdollisimman paljon vapauksia ja oikeuksia, jotka ovat kaikille samoja. Toisekseen yhteiskunnan eriarvoisuuden tulisi hyödyttää eniten heikoimmassa asemassa olevia. (Rawls, 1998.) Yhteiskuntasopimusteorian mukaan moraalisesti hyväksyttävä toiminta on siis linjassa yhteisön arvojen kanssa. Teoria ottaa myös kantaa siihen, mihin näiden arvojen tulisi perustua.

3.1.4 Deskriptiivinen etiikka

Deskriptiivinen etiikka tutkii ihmisten moraalikäsityksiä ja on täten kiinnostunut kuvailemaan, millaisia käsityksiä oikeasta ja väärästä esimerkiksi jollakin tutkittavalla joukolla yksilöitä on. (Stahl ym., 2016.) Esimerkiksi informaatioteknologia-alan ammattilaisten arvoja ja moraalikäsityksiä kartoittava tutkimusartikkeli olisi luonteeltaan deskriptiivinen. Tällainen lähestymistapa voisi hyvinkin olla hedelmällinen myös massadatan ja tiedonlouhinnan eettisten ongelmien tutkimiseksi. Tähän kandidaatintutkielmaan tämä ei kuitenkaan olisi soveltunut tutkimusmetodiksi, sillä tutkimus suoritetaan kirjallisuuskatsauksena ilman empiiristä osuutta.

3.2 Tietojenkäsittelyn etiikasta

Tietojenkäsittelyn etiikalla tarkoitetaan etiikan haaraa, joka tutkii erityisesti informaatioteknologiaan liittyvää moraalista problematiikkaa (Floridi, 1999; Moor, 1985; Stahl ym., 2016). Floridi (1999) erottaa kaksi kokonaisuutta tästä etiikan alasta: käytännönläheisemmän *tietokone-etiikan* sekä tämän filosofisena pohjana toimivan *informaatioetiikan*. Tässä tutkielmassa termillä tietojenkäsittelyn etiikka viitataan näiden kahden muodostamaan kokonaisuuteen.

Masonin (1986) mukaan tietojenkäsittelyyn liittyviä suuria eettisiä ongelmia ovat yksityisyys sekä tiedon tarkkuus, omistajuus ja saatavuus. Hieman samaan tapaan Davis (2012) tunnistaa erityisesti massadatan eettisten ongelmien olevan luonteeltaan identiteettiin, yksityisyyteen, omistajuuteen ja maineeseen liittyviä. Tietojenkäsittelyä koskevat eettiset kysymyksenasettelut voivat siis pysyä hyvinkin samankaltaisina vuosikymmenestä toiseen.

Tietojenkäsittelyn etiikan kentän on kuvailtu kärsivän monista ongelmista, joita ovat ainakin liiallinen yksilökeskeisyys, normatiivisen näkökulman puute, heikko yhteys etiikan klassikoihin ja muuhun etiikan tutkimukseen sekä ajankohtaisten ongelmien liiallinen painotus (Laudon, 1995). Ongelmallisena on pidetty myös vankan filosofisen perustan puutetta (Floridi, 1999). Lisäksi yhtenä huolenaiheena on ollut IT-etiikan sisäinen hajanaisuus ja suoranainen epä-

määräisyys tutkimuskenttänä (Smith & Hasnas, 1999). Stahl ym. (2016) peräänkuuluttavat samaan tapaan johdonmukaisuutta tietojenkäsittelyn etiikkaan, jonka tulisi heidän mukaan kehittyä tieteenalana huomattavasti nykyisestä tilastaan. Näiden näkemysten valossa tutkimusongelmien ratkaisemiseen lienee suotavaa hyödyntää myös näkökulmia muualta etiikan kentältä.

Smith ja Hasnas (1999) esittävät liiketoimintaetiikan näkökulman tuomista tietojärjestelmiä koskevaan eettiseen ajatteluun organisaatioissa. Tällaisessa ajattelussa on piirteitä liiketoimintaetiikan kolmesta kilpailevasta haarasta, jotka ovat osakkeenomistajien, sidosryhmien ja yhteiskuntasopimuksen näkökulmat. Jälkimmäinen haara painottaa liiketoiminnan yhteiskuntavastuuta ja yleisen hyvinvoinnin edistämistä. Kaksi edellistä näkökulmaa puolestaan korostavat joidenkin rajattujen ihmisryhmien etujen ajamista. (Smith & Hasnas, 1999.) Tämän tutkielman näkökulmasta on hyvä huomata, että massadataa ja tiedonlouhintaa hyödynnetään liiketoiminnassa taloudellisen edun saavuttamiseksi. Näin ollen yhteiskuntavastuun ja osakkeenomistajien näkökulman välille voi tulla eettinen ristiriita. Taloudellinen kannustin eettisesti arveluttavan toiminnan taustalla on tässä yhteydessä erittäin olennainen ongelma, jota on syytä pohtia.

3.3 Tietojenkäsittelyn etiikan asemoituminen etiikan kentälle

Kuten edellä todettiin, tietojenkäsittelyn etiikan ymmärretään yleisesti kuuluvan soveltavan etiikan piiriin (Marturano, 2002). Kuitenkin muun muassa Floridi (1999) on puhunut tietojenkäsittelyn etiikan oman filosofisen pohjan puolesta. Seuraavaksi tarkastellaan, millaisia käsityksiä kirjallisuudessa on tietojenkäsittelyn etiikan asemasta etiikan kentällä.

Moorin (1985) mukaan tietojenkäsittelyn etiikka on täysivaltainen oma etiikanalansa, jossa yhdistyy informaatioteknologian ilmiöiden konseptointi ja monipuolinen eettinen pohdinta. Tämän näkökulman mukaan IT-innovaatioiden eettisiä ongelmia ei siis voida lähestyä pelkästään soveltamalla niihin suoraan etiikan teorioita, vaan tarkastelu täytyy kytkeä tieteellisiin konsepteihin innovaatioista. Täten on loogista ajatella, että tällaisten kysymysten pohdintaan keskittynyt suuntaus mielletään omaksi etiikan haarakseen.

Johnson (2008) väittää kuitenkin, että koko tietojenkäsittelyn etiikan alan voidaan nähdä tarkastelevan pohjimmiltaan ongelmatilanteita, joiden eettiseen luonteeseen informaatioteknologialla ei ole sinänsä vaikutusta. Tässä ajattelutavassa lähdetään oletuksesta, että minkä vain eettistä pohdintaa vaativan ongelmatilanteen voi purkaa etiikan klassikoiden esittämiin peruskysymyksiin. Itse ongelman kontekstilla ei siis ole kovin suurta painoarvoa tällaisessa ajatusmallissa. Tämä näkemys kyseenalaistaa toisin sanoen tietojenkäsittelyn etiikan tarpeen omana tieteenalanaan. Vaihtoehtoisesti informaatioteknologiaan liittyvät eettiset ongelmat voitaisiin siis nähdä jonkinlaisina ”yleiseettisinä” ongelmina.

Floridi (1999) argumentoi varsin kattavasti, miksi etiikan klassikot eivät sovellu riittävän hyvin tietojenkäsittelyyn liittyvien eettisten ongelmien pohdintaan. Hänen mukaan muiden muassa velvollisuus-, seuraus- ja hyve-etiikka ovat painotuksiltaan liian yksioikoisia monimuotoisten informaatioteknologian käyttöä koskevien moraalisten ongelmien ratkaisemiseksi. Perinteisten etiikan teorioiden rajoitteiksi Floridi mainitsee esimerkiksi ihmiskeskeisyyden sekä kyvyttömyyden ymmärtää virtuaalisessa ympäristössä tapahtuvien tekojen luonnetta. (Floridi, 1999.) Tämä suuntaus siis painottaa juuri kontekstin merkitystä eettisessä pohdinnassa. Smith (2002) kuitenkin väittää, että informaatioteknologiaan liittyvään eettisen pohdintaan soveltuu käytännössä kaikki perinteiset etiikan teoriat kunhan niitä käytetään perustellusti.

3.4 Tutkimusongelmien ratkaisusta etiikan avulla

Johnson (2008) tunnistaa kolme eri tapaa organisoida informaatioteknologiaan liittyviä eettisiä kysymyksiä: teknologian tyypin, teknologiaa hyödyntävän alan ja kyseessä olevaan ongelmaan sovellettavien etiikan konseptien mukaan. Koska olen rajannut tutkielmani koskemaan massadataa ja tiedonlouhintaa, on ensimmäisenä mainittua organisointitapaa jo sovellettu tämän tutkimuksen puitteissa. Lisäksi lähtökohtanani on tutkia eettisiä ongelmia nimenomaan tietojenkäsittelyn etiikan pohjalta.

Johnson (2008) esittää myös, että tietojenkäsittelyyn liittyviä eettisiä ongelmia pohdittaessa on tehtävä synteesiä, josta käy ilmi sekä kyseessä olevan teknologian että eettisen ongelmatilanteen ymmärrys. Nämä kriteerit pyrin täyttämään taustoittamalla massadataa ja tiedonlouhintaa teknologioina sekä kytkemällä eettisen tarkastelun näiden ilmiöiden olennaisimpiin toiminnallisuuksiin.

Eettiselle pohdinnalle on leimallista lopullisten vastausten puuttuminen. Quinn (2014) toteaa olevan täysin mahdollista, että esimerkiksi kaksi utilitaristia voi olla yhtä mieltä eettisen ongelman lähtökohdista mutta päätyä silti täysin erilaisiin johtopäätöksiin. Tämän etiikan ulottuvuuden vuoksi tutkielmassa esitellään erilaisia näkökulmia massadatan ja tiedonlouhinnan ongelmiin. Kuten totesin aiemmin kohdassa 3.1.2, tutkielmassa esiteltävän viitekehyksen käyttäjällä on velvollisuus vetää itse johtopäätökset ongelmatilanteissa. Valmiiden vastausten puuttuessa soveltajalla on myös vapaus hyödyntää valitsemaansa näkökulmaa pohtiessaan viitekehyksen esittelemiä eettisiä teemoja.

Alaluvussa 3.3 esiteltiin eriäviä näkemyksiä informaatioteknologiaan liittyvien eettisten kysymysten luonteesta. Tässä tutkielmassa eettinen pohdinta pohjautuu niin tietojenkäsittelyn etiikan esittämiin tarkempiin kysymyksiin kuin kulloinkin aiheellisiin etiikan klassikoiden perustavanlaatuisiin kysymyksiin. Koska etiikan teorioiden kirjo on valtava, ei massadataa ja tiedonlouhintaa tarkasteltaessa ole järkevää soveltaa kovin montaa teoriaa yhtä eettistä ongelmatilannetta kohden. Quinn (2014) toteaa pitkällisen etiikan teorioiden taustoituksen pohjalta, että informaatioteknologian eettisiin ongelmiin soveltuvat par-

haiten seuraavat neljä teoriaa: deontologia eli velvollisuusetiikka, teko- ja sääntöutilitarismi sekä yhteiskuntasopimusteoria. Nämä neljä teoriaa ovat täten tämän tutkielman oleellisimmat eettiset näkökulmat, joilla täydennetään tietojenkäsittelyn etiikan kirjallisuudessa esitettyjä huomioita.

On huomionarvoista, että edellä luetelluista teorioista yhteiskuntasopimusteoria soveltuu ennen kaikkea yhteiskunnallisten asioiden eettiseen pohdintaan. Tämän tutkielman tavoitteena on juurikin tällaisten ongelmien löytäminen massadatan ja tiedonlouhinnan hyödyntämisestä yksilönäkökulman lisäksi. Tutkielman yhden näkökulman lepääminen vain yhden teorian varassa on kuitenkin jokseenkin ongelmallista itsessään. Ylipäätään teorioiden soveltaminen käyttötapauksen mukaan ei ole aivan yksioikoisesti tarkoituksenmukaista. Koen joka tapauksessa, että velvollisuusetiikka ja teko- sekä sääntöutilitarismi eivät anna riittävästi eväitä yhteiskunnallisten ongelmien tarkasteluun eettisessä mielessä. Tämän vuoksi yhteiskuntasopimusteorian soveltaminen pohdinnassa on mielestäni oikeutettua. Kuitenkin tämän teorian hyödyntäminen yksilöön liittyvien ongelmien pohdinnassa olisi hankalaa, joten tässä tutkielmassa tyydytään ratkaisuun soveltaa yksilön ja yhteiskunnan näkökulmiin eri etiikan teorioita.

3.5 IT-ammattilaisen eettisestä toiminnasta

Informaatioteknologia-alan ammattilaisille on kehitetty monia eettisiä ohjeistoja työhön liittyviä ongelmatilanteita varten. Alan vaikutusvaltaisen järjestön Association for Computing Machineryn (ACM) ammattietiikan koodi tunnistaa näitä tilanteita ja antaa yleisohjeita niissä toimimiseen (Anderson, 1992). Suomessa vastaavanlaisen koodin on kehittänyt Tieto- ja viestintätekniikan ammattilaiset TIVIA ry (TIVIA, 2002). Kummassakaan ohjeistossa ei kuitenkaan kytketä ohjeita eettisiin periaatteisiin. Esimerkiksi Quinn (2014) huomauttaa analyysissaan ACM:n koodin pohjautuvan peräti viiteen täysin eri etiikan teoriaan. Tässä tutkielmassa on syytä hyödyntää yleisluontoisempaa ja tiiviimpää esitystä IT-ammattilaisen eettisiksi periaatteiksi.

Quinn (2014, 361–362) on muotoillut ACM:n monitahoisen eettisen koodin pohjalta yleisluontoiset eettiset periaatteet ohjelmistokehitysalan ammattilaisille. Allaolevassa taulukossa (taulukko 1) on esitettynä nämä periaatteet lyhyiden selitysten saattamina suomeksi käännettynä.

TAULUKKO 1. Ohjelmistokehitysammattilaisen eettiset periaatteet Quinnia (2014, 361–362) mukaillen. Suomennettu englannista.

Eettinen periaate	Tiivistetty selitys
1. Ole puolueeton.	Henkilökohtainen etusi on yhtä tärkeä kuin yhteiskunnan ja organisaatiosi etu.
2. Jaa sellainen tieto, joka muiden kuuluu tietää.	Kerro eturistiriidoista. Älä anna vääriä lausuntoja. Älä salaa tärkeää tietoa muilta.
3. Kunnioita muiden oikeuksia.	Älä loukkaa muiden oikeuksia yksityisyyteen tai henkiseen ja fyysiseen omaisuuteen.
4. Kohtele muita oikeudenmukaisesti.	Kohtele työtovereitasi tasa-arvoisesti. Älä rankaise muita eettisten periaatteiden noudattamisesta.
5. Ota vastuu teoistasi ja tekemättä jättämistäsi.	Olet vastuussa sekä hyvistä että huonoista asioista, jotka ovat seurausta toimistasi ja pidättäytymisestäsi toimia.
6. Ota vastuu alaistesi teoista.	Esimiehet ovat vastuussa alaistensa työtehtäviin liittyvien riskien hallinnasta.
7. Säilytä integriteettisi.	Ole lojaali työnantajallesi ja suoriudu sitoumuksistasi lakien puitteissa.
8. Kehitä kykyjäsi jatkuvasti.	Kehitä ohjelmistokehitysosaamistasi ja kykyjäsi panna näitä eettisiä periaatteita toimeen.
9. Jaa tietämyksesi, asiantuntemuksesi ja arvosi.	Jaa osaamistasi ohjelmistokehityksestä ja tietouttasi ammattietiikasta muiden kanssa.

Ylläolevia periaatteita voidaan soveltaa jäljempänä esiteltävän massadatan ja tiedonlouhinnan hyödyntämistä koskevan eettisen viitekehyksen käytön tukena. Quinn (2014) toteaa, ettei kaikkia yhdeksää periaatetta ole tarkoituksenmukaista soveltaa jokaisen ongelmatilanteen ratkaisun pohtimisessa. Näin ollen näitä periaatteita soveltavan pohdittavaksi jää, mitkä niistä kannattaa kulloinkin huomioida.

Leonard ja Cronan (2001) esittävät, että IT-alan ammattilaisen eettiseen käyttäytymiseen vaikuttavat henkilökohtaiset uskomukset, eettiset asenteet, päättelytavat, sukupuoli ja egon vahvuus. Päätöksentekoprosessi toimintatavan valitsemiseksi IT-kontekstissa voi siis olla hyvin monimutkainen, ja eettinen

ulottuvuus on vain yksi osa tätä prosessia. Samoja eettisiä koodeja soveltavat ihmiset voivat siis päätyä erilaisiin valintoihin myös henkilökohtaisten ominaisuuksien vuoksi. Lisäksi, kuten johdannossa todettiin, Mingers ja Walsham (2010) huomauttavat eettisen pohdinnan mahdollistavan luonteenomaisesti monet erilaiset lopputulemat.

4 MASSADATAN JA TIEDONLOUHINNAN EETTISET ONGELMAT

Tässä luvussa käydään läpi massadataa ja tiedonlouhintaa koskevaa tutkimuskirjallisuutta ja tutkitaan tietojenkäsittelyn etiikan kirjallisuuden valossa, mitä eettisiä ongelmia näiden menetelmien hyödyntämiseen organisaation toiminnassa liittyy. Eettinen pohdinta perustuu Quinnin (2014) näkemykseen IT-kontekstiin sopivista etiikan teorioista. Viimeisessä alaluvussa koostetaan tulokset ja luonnostellaan niiden pohjalta eettinen viitekehys. Lisäksi pohditaan tuloksia sekä esitetään johtopäätöksiä.

4.1 Massadataan liittyvät eettiset ongelmat

Tässä alaluvussa tarkastellaan erityisesti massadatan keräämiseen, säilyttämiseen, julkistamiseen ja myymiseen liittyviä eettisiä kysymyksiä. Sekä yksilön että yhteiskunnan näkökulmia aiheeseen tuodaan esille.

4.1.1 Datan kerääminen

Ensimmäinen askel massadatan hyödyntämiseen tähtäävässä toiminnassa on datan kerääminen. Suuret datamassat kootaan tyypillisesti tietokantoihin, jotka eroavat ominaisuuksiltaan perinteisistä relaatiotietokannoista. Dataa voidaan kerätä useasta eri lähteestä useilla eri menetelmillä, ja uutta dataa voi virrata tietokantaan erittäin suurella nopeudella. (Hashem ym., 2015.)

Valtavien datamassojen kerääminen ihmisistä ja heidän suorittamistaan erilaisista transaktioista voi tapahtua joko kyseessä olevien ihmisten tietoisesti antamalla suostumuksella tai ilman sitä. Luvatta tapahtuvan datan keräämisen voidaan ajatella olevan jo määritelmällisesti epäeettistä. Luvattoman tietojen keräämisen eettisten ongelmien pohtiminen ei siis liene relevanttia. Tarkkailtavan suostumuksella tapahtuvaan datankeruuseen kuuluu sen sijaan jonkin verran moraalisessa mielessä harmaata aluetta.

Edes massadatan kerääminen luvalla ei nimittäin ole eettisessä mielessä aivan mutkatonta. Custers (2016) esittääkin aiheellisen kysymyksen: milloin yrityksen oikeus kerätä dataa käyttäjästä umpeutuu tämän hyväksyttyä tietosuojakäytänteen? Lienee kohtuutonta olettaa, että käyttäjä ymmärtää luettuansakaan suostuvansa mahdollisesti vuosikausia kestävään tiedonkeruuseen. Joidenkin ohjelmistojen tai palveluiden käytön suhteen käyttäjällä ei välttämättä edes ole muuta realistista vaihtoehtoa kuin hyväksyä käytänne saadakseen välttämättömän palvelun käyttöönsä (Custers, 2016). Tilanteessa, jossa valinnanvaraa ei käytännössä anneta, Kantin kategorisen imperatiivin ehdot eivät täyty. Käyttäjän kannalta huonon mutta pakon edessä solmitun sopimuksen ei voida ajatella perustuvan yleistettävissä olevaan moraalisesti hyvään käytäntöön.

Quinn (2014) huomauttaa, että luvan saaminen datan keräämiseen jonkin palvelun käyttäjistä voi tapahtua kahdella eri tavalla. Ensinnäkin käyttäjältä voidaan kysyä erikseen, haluaako hän antaa luvan kerätä itsestään tietoja. Toinen keino on datan kerääminen, jollei käyttäjä sitä erikseen kiellä. Jälkimmäinen menettely on edellistä yleisemmin käytössä. (Quinn, 2014.) Eettisessä mielessä näillä kahdella käytännöllä on merkittävä ero, jonka havainnollistamiseen sopii sääntöutilitarismi. Kuvitellaan, että moraalisesti hyväksyttävä sääntö olisi kerätä aina tietoja kieltäytymättä jättämisen perusteella. Sääntö on heikko, koska näin ajateltuna datan kerääjien etu olisi hankaloittaa kieltäytymistä esimerkiksi tekemällä prosessista mahdollisimman pitkä ja vaikeasti suoritettava. Loogisesti paremmalla pohjalla olisikin moraalisääntö, jonka mukaan käyttäjiltä tulisi kysyä erikseen lupa datan keräämistä varten.

Massadatalle tyypillisen automatisoidun datankeruun tapauksessa on ongelmallista määritellä, kuka kantaa moraalisen vastuun toiminnasta. Floridi (2013) puhuukin jaetun moraalin (engl. *distributed morality*) käsitteestä: informaatioteknologian hyödyntämisessä moraalisuus voi jakautua useille mukana oleville neutraaleille toimijoille. Tällä tarkoitetaan toisin sanoen sitä, että erittäin suuren moraalisen latauksen omaava toiminta voi koostua joukosta pienempiä, moraalisesti neutraaleja tekoja (Floridi, 2013). Viime kädessä automatisoidustakin datankeruusta on kuitenkin vastuussa joku tai jotkut organisaation ihmisedustajat. Näiden toimijoiden harkittavaksi jää kokonaisuuden eettisyyden hahmottaminen.

Johnson (2008) nostaa esiin datan keräämisestä mahdollisesti aiheutuvia kauaskantoisia vaikutuksia. Hänen mukaan datan kerääminen ihmisten kaikesta toiminnasta voi johtaa kollektiiviseen tarkkailtavana olemisen tunteeseen. Tästä johtuen ihmiset saattavat rajoittaa – mahdollisesti syystäkin – tekemisiään vain tarkkailun mahdollisuuden vuoksi. Tämä voisi johtaa yksilönvapauksien toteutumisen heikkenemiseen ja täten vaikuttaa suuresti yhteiskuntaan. (Johnson, 2008.) Yhteiskuntasopimusteorian mukaan yksilöillä tulisi olla mahdollisimman suuret ja yhtäläiset vapaudet: tämä ihanne ei toteutuisi kuvaillun kaltaisessa yhteiskunnassa, jossa ihmiset ajautuisivat osin omaehtoisesti rajoittamaan vapauksiaan.

Lerman (2013) esittää huolensa massadatan katveessa olevista ihmisistä. Tällä tarkoitetaan usein huono-osaisia ihmisiä, joista ei kerätä dataa parempiosaisista eriävien elämäntapojen vuoksi. Massadatan analysointiin perustuvat yhteiskunnalliset päätökset saattavatkin suosia hyväosaisia ja heikentää datankeruun ulkopuolelle jäävien asemaa entisestään. (Lerman, 2013.) Massadatan keräämisen yleistyessä on siis huomioitava myös eettiset ongelmat liittyen datankeruun otantaan. Tämä koskee erityisesti tilanteita, joissa massadataa analysoimalla tehdään suuria ihmisjoukkoja koskevia päätöksiä.

4.1.2 Datan säilyttäminen ja julkistaminen

Tenen ja Polonetskyn (2012, 251) mukaan suurista datamassoista voi löytää laillisinkin keinoin vertailemalla yksityisiä tietoja. Tämän huomion vuoksi on syytä tarkastella, millaisia esimerkkejä tällaisesta toiminnasta on kerrottu kirjallisuudessa. Lisäksi pohditaan, mitä ongelmia liittyy ihmisistä kerätyn datan pitkäaikaiseen säilyttämiseen.

Edellä kuvailtiin eettisiä ongelmia liittyen tietosuojakäytänteiden hyväksymisestä seuraavaan datan keräämiseen. Tähän asiaan liittyvät ongelmat eivät rajaudu pelkästään keruuhetkeen vaan ne ulottuvat aina massadatan säilyttämiseen saakka. Custers (2016) huomauttaa, että monet dataa keräävät organisaatiot eivät poista käyttäjistä kerättyä dataa käytännössä koskaan. Tällöin käyttäjistä kerätyt tiedot voivat jäädä yrityksen tietokantoihin ja altistua esimerkiksi tietomurroille vaikka käyttäjä olisi lopettanut palvelun käyttämisen kauan sitten (Custers, 2016). Tämä herättää kysymyksiä datan soveliaasta säilytysajasta.

Yksi keino lisätä käyttäjistä kerätyn datan yksityisyydensuojaa massadatassa on anonymisointi, jonka periaatteita käytiin läpi alaluvussa 2.5. Anonymisoiduistakin datamassoista voidaan kuitenkin löytää erilaisin menetelmin yksittäisiä henkilöitä koskeva dataa. Tästä ilmiöstä käytetään nimitystä deanonymisointi (engl. *de-anonymization*) (Narayanan & Shmatikov, 2008). Koska ihmisiä voidaan tunnistaa dataseteistä, ei anonymisointi ole millään muotoa varma keino taata yksityisyys. Silti tunnetaan tapauksia, joissa jokin dataa kerännyt taho on julkistanut suuria datamassoja luullen niiden olevan mahdottomia deanonymisoida. Zimmer (2010) mainitsee esimerkkinä Harvardin ja UCLA:n tutkijoiden julkistaman datasetin, joka koski anonyymin yliopiston anonyymien opiskelijoiden Facebookin käyttöä neljän vuoden ajalta. Yliopisto paljastui kuitenkin aineiston perusteella hyvin pian Harvardiksi, ja tämän jälkeen datasetistä pystyi identifioimaan yksittäisiä opiskelijoita. (Zimmer, 2010.)

Analysoimalla ristiin erilaisia julkisesti saatavilla olevia datamassoja on mahdollista saada selville hyvinkin arkaluonteista tietoa. Narayanan ja Shmatikov (2008) ovat onnistuneet selvittämään Netflix-käyttäjien palvelun käytön perusteella selville hyvin henkilökohtaisia tietoja käyttäjistä, kuten poliittisia näkemyksiä. Tutkittu data oli anonymisoitua, ja taustatietojen selvittämiseen tutkijoiden tarvitsi käyttää vain elokuvasivusto IMDB:tä (Narayanan & Shmatikov, 2008). Hieman samaan tapaan Acquisti ja

Gross (2009) havaitsivat Yhdysvaltain kansalaisten sosiaaliturvatunnusten ennustamisen olevan mahdollista soveltamalla tilastotieteellisiä menetelmiä täysin julkisesti saatavilla olevaan dataan. Tämä herättää kysymyksiä siitä, mitä kaikkia tietomassoja on syytä avata kaikkien käytettäviksi. Jälkimmäisessä tapauksessa sosiaaliturvatunnukset pystyttiin ennustamaan henkilöiden syntymäaikojen ja -paikkojen perusteella (Acquisti & Gross, 2009). Tutkittavan datan ei siis tarvitse olla luonteeltaan lainkaan arkaluonteista yksityisten asioiden selvittämiseksi.

On ongelmallista, että esiteltyjen tapauksien tapaan harmittomien tietojen varassa voidaan tehdä niinkin pitkälle meneviä analyyseja, että voidaan saada selville arkaluonteista tietoa. Eettisestä näkökulmasta on luontevaa todeta, että moraalinen vastuu tällaisen ennustamisen mahdollisuudesta kuuluu ennemmin analysoijille kuin harmittoman tiedon julkistajille. Zimmerin (2010) esittelemä esimerkki arkaluonteista tietoa sisältävän datasetin julkistamisesta on kuitenkin asia erikseen. Sokea luottaminen jonkin aineiston anonymisoinnin pitävyyteen on heikko peruste yksityistä tietoa sisältävän materiaalin julkistamiselle. Vaikka jostakin julkiseksi tehdystä datasetistä ei voisikaan nykyisillä menetelmillä tunnistaa ihmisiä, on analysointimenetelmien kehityksen vuoksi syytä epäillä tämän olevan mahdollista tulevaisuudessa.

Massadatan säilyttämiseen liittyen on herännyt huolta kerätyn datan siirtämisestä erilaisten lainsäädäntöjen piiriin kuuluvien alueiden välillä. Kansainvälisten sopimusten puuttuessa organisaatiota voi houkutella mahdollisuus hyödyntää dataa muualla päin maailmaa tavalla, joka olisi keräyspaikassa laitonta. (Johnson, 2008.) Tässä tapauksessa velvollisuusetiikan mukaan samansisältöinen teko olisi aivan yhtä tuomittava huolimatta sen tekopaikasta. Tilanne on myös malliesimerkki siitä, että informaatioteknologian kehitys luo käytäntötyhjiöitä, joita arveluttava toiminta voi hyödyntää.

Yksi massadataan liittyvä eettinen kysymys liittyy yksilön oikeuteen saada nähtäviinsä itsestään jonkin tahon toimesta kerätyt tiedot (Tene & Polonetsky, 2012). Paljon mediahuomiota saanut itävaltalaisopiskelija Max Schremsin toteutunut pyyntö saada Facebookilta asiakirja, josta selviää kaikki yrityksen hänestä keräämänsä tiedot, liittyy tähän ongelmaan (Hill, 2012). Velvollisuusetiikan näkökulmasta voidaan ajatella, että tietoja keräävällä organisaatiolla on moraalinen velvollisuus kertoa käyttäjälleen, mitä tietoja tästä on kerätty. Seurausetiikalla perustellen organisaatio voisi kuitenkin jättää tiedot antamatta, jos niiden salassapito johtaisi suurempaan yhteiseen hyvään. Esimerkiksi kansallinen turvallisuus voisi olla tällainen peruste.

4.1.3 Datan myyminen

Davisin (2012) tekemän selvityksen mukaan 40 Forbesin listaamista 50:stä Yhdysvaltain suurimmasta yrityksestä ilmoittaa jakavansa käyttäjistä keräämäänsä dataa kolmansien osapuolien kanssa. Lisäksi 34 mainitsi pidättyvänsä myymästä dataa ilman suostumusta, kun taas loput 16 eivät ilmoittaneet käytäntöään tässä suhteessa (Davis, 2012). On siis hyvin paljon

mahdollista, että suuryritykset käyvät luvallisesti käyttäjistään keräämällä datalla kauppaa ilman näiden lupaa.

Quinn (2014) huomauttaa, että asiakkaista kerätystä datasta on itsestään tullut tuote. Täten dataa keräävän organisaation liiketoiminnan kannalta on järkevintä vaatia asiakkaalta erillistä kieltoa tätä koskevan datan myymiselle. Näin organisaatiolla on enemmän dataa myytävänään kuin tilanteessa, jossa asiakkaan tulisi hyväksyä datan myyminen erikseen. (Quinn, 2014.) Davisin (2012) mukaan tässä on kyse eettisestä arvovalinnasta ihmisten tuntemattoman pelon kunnioittamisen ja liiketoimintamallin tuettavuuden välillä. Tässä eettiseen pohdintaan soveltuu liiketoiminnan etiikan kolme näkökulmaa, jotka ovat osakkeenomistajien, sidosryhmien ja yhteiskuntasopimuksen näkökulmat. Osakkeenomistajien ja sidosryhmien etujen palvelemiseksi organisaation tulisi toimia ennen kaikkea voiton maksimoimiseksi. Yhteiskuntasopimuksen lähestymistavassa tärkeää on kuitenkin myös yhteiskuntavastuu. (Smith & Hasnas, 1999.) Viimeiseksi mainittua ajattelua soveltaen ihmisten datan käsittelyyn liittyviä pelkoja tulisi kunnioittaa. Kahden edellisen näkökulman perusteella hyväksyttävän toimintatavan valinta on mutkikkaampaa. Epäselvien datakäytäntöjen johdosta voi nimittäin aiheutua haittaa organisaation imagolle, jolloin pitkän tähtäimen taloudellinen hyöty voi vaarantua.

Lisäksi Davisin (2012) mukaan datan siirtyminen organisaatiosta toiseen voi tarkoittaa erilaisten tietoturvakäytänteiden vuoksi datan yksityisyyden vaarantumista. Vaikka organisaatio väittäisi kaiken kerätyn ja myytävän datan olevan täysin anonymisoitua, on tämän varotoimen aukottomuus deanonymisointitekniikoiden tehokkuuden valossa kyseenalainen (Davis, 2012).

4.2 Tiedonlouhintaan liittyvät eettiset ongelmat

Tässä alaluvussa keskitytään massadatan analysoimiseksi käytettävään tiedonlouhintaan liittyviin eettisiin ongelmiin. Näitä ovat itse analyysia ja sen tulosten hyödyntämistä koskevat ongelmat. Lisäksi kiinnitetään huomiota yhteiskunnallisiin huoliin, joita on esitetty kirjallisuudessa.

4.2.1 Datan analysoiminen

Tiedonlouhintaan liittyen on herännyt kysymyksiä massadata-analyysien tarkkuudesta. Lazer, Kennedy, King ja Vespignani (2014) käsittelevät esimerkkinä Google Flun tapausta, jossa Googlen hakudataan perustuvat ennusteet influenssatapauksista ovat olleet erittäin epätarkkoja. Tämän kaltaisissa massadatan analysoimiseen perustuvissa menetelmissä tutkimusten läpinäkyvyys ja toistettavuus ovat myös kyseenalaisia (Lazer ym., 2014).

Eräs eettinen ongelma tiedonlouhinnassa on moraalisen vastuun määrittelyn vaikeus useiden neutraalien toimijoiden, kuten robottien, suorittaessa massadatan analysoimisen. Kuten edellä jo kuvailtiin, tämä ongelma liittyy

myös datan automatisoituun keräämiseen. (Floridi, 2013.) On myös esitetty huolia siitä, että automaattinen päätöksenteko ihmisistä kerätyn datan pohjalta voi johtaa syrjintään ja valinnanvapauden kaventumiseen. Esimerkiksi ihmisten luottopäätösten ja vakuutusmaksujen määräytyminen data-analyysin kautta voi edistää tätä kehitystä. (Tene & Polonetsky, 2012, 252.) Tällaisen taloudellisen eriarvoisuuden voidaan nähdä vaikuttavan eniten pienituloisiin eli heikossa asemassa oleviin. Yhteiskuntasopimuksen näkökulmasta eriarvoisuuden tulisi hyödyttää eniten huono-osaisia, mikä ei tässä skenaariossa toteudu. Toisaalta tiedonlouhinnan avulla saadut tarkemmat tiedot esimerkiksi luottokelpoisuudesta voivat tasa-arvoistaa luottomenettelyjä.

Crawfordin ja Schultzin (2014) mukaan on esitetty huolia siitä, että massadatan analysoimiseksi kehitettyjen algoritmien vaikutuksista yksityisyyteen eivät voi olla varmoja edes niiden kehittäjät. Tämä on erityisen suuri huolenaihe oppivien algoritmien osalta, sillä ne kehittyvät analysoimisessa itsestään (Crawford & Schultz, 2014). Näin ollen massadatan analysointiin käytetyillä tehokkailla menetelmillä voi olla arvaamattomia vaikutuksia.

Eettisestä näkökulmasta tiedonlouhintaa hyödynnettäessä olisi syytä pohtia mahdollisuuksia käyttää yksityisyyden säilyttävän tiedonlouhinnan menetelmiä. Hanin ym. (2011) mukaan alan tutkimus on johtanut tiedonlouhinnan yhteyteen integroitujen anonymisointimenetelmien tehokkuuden kehitykseen. Samaan tarkoitukseen tähtää myös yksityisyyden varjelemiseksi kehitetty data-analyysialusta PINQ (lyh. sanoista *privacy integrated queries*), joka hyödyntää anonymisointitekniikoita (McSherry, 2009). Tietosuojaryhmä (2014, 24) kuitenkin väittää, että mikään käytössä olevista anonymisointitekniikoista ei ole aukoton. Joka tapauksessa organisaatioiden on syytä miettiä, voisivatko ne hyödyntää yksityisyyden säilyttävän tiedonlouhinnan menetelmiä vaarantamatta data-analyysin tarkkuutta.

On hyvä huomata, että on olemassa monia sovellusalueita, joilla tiedonlouhintaa ei käytetä lainkaan henkilöistä kerättyyn dataan. Tällaisia alueita ovat muun muassa luonnonmullistusten ennustaminen, geologia ja astronomia. (Han ym., 2011, 620.) Tiedonlouhinta näillä saroilla ei nähdäkseni kaipaa eettistä pohdintaa yhteiskunnan tai yksilön näkökulmasta. Mahdollisesti kyseeseen tulevat tieteeneettiset ongelmat eivät kuulu tämän tutkielman sisältöön.

4.2.2 Analyysitulosten hyödyntäminen

Tiedonlouhinnan avulla saatuja organisaation kannalta mielenkiintoisia tietoja voidaan hyödyntää monilla eri tavoilla. Yksi suosittu sovellusalue on markkinointi: massadataa analysoimalla voidaan kohdentaa tuotteiden markkinointia ja ottaa selvää eri tuotteiden ostamisen välisistä riippuvuussuhteista asiakaskohtaisesti (Han ym., 2011; Quinn, 2014). Tämän lisäksi käsitellään tiedonlouhinnan hyödyntämistä muun muassa valtionhallinnon ja lainvalvonnan toimesta.

Duhiggin (2012) mukaan yhdysvaltalainen kauppaketju Target on kohdentanut vauvatuotteiden markkinointiaan naisasiakkaille, joiden yritys on ennustanut tiedonlouhinnan avulla olevan raskaana. Crawford ja Schultz (2014)

korostavat, että Target ei missään vaiheessa ollut kerännyt dataa, joka olisi osoittanut jonkin asiakkaan olevan raskaana: ennustemalli vain kertoi, milloin tiettyjä ostoksia tekevä asiakas oli tätä suurella todennäköisyydellä. Ostoskombinaatioita tutkivassa menetelmässä on jälleen kyse moraalisesti latautuneesta toiminnasta, joka koostuu sinänsä moraalisessa mielessä neutraaleista toimista. Menettelystä vastuussa oleva taho on kuitenkin tietoisesti halunnut selvittää asiakkaan raskauden, jolloin ostosten tutkimisen näennäiseen eettiseen neutraaliuteen ei ole mielestäni uskottavaa vedota.

Crawford ja Schultz (2014) nostavat esiin yritysten tavan kiertää markkinointilakeja massadatan analysoimisen avulla. Internetissä tapahtuvaa lainojen ja asuntojen mainontaa on nimittäin kohdennettu yritysten kannalta suotuisille asiakasehdokkaille, jolloin epäsuotuisat ehdokkaat eivät näe vastaavia mainoksia. Näin organisaatiot voivat siis syrjiä ihmisiä salaa vaikkapa rodun tai sukupuolen perusteella. (Crawford & Schultz, 2014.) Tässä on siis pohjimmiltaan kyse tietoisesta lain rikkomisesta, jonka paljastuminen on tehty teknologian keinoin epätodennäköiseksi. Sääntöutilitarismin ja velvollisuusetiikan näkökulmasta tämä ei ole eettisesti perusteltavissa, sillä tällöin toimijat hyväksyisivät yleiseksi säännöksi sääntöjen rikkomisen, kunhan siitä ei jää kiinni.

Quinnin (2014) mukaan Yhdysvaltain hallinnon alaiset toimijat käyttävät tiedonlouhintaa muun muassa veronkierron, tautiepidemioiden ja terrorismin torjuntaan. Näihin tarkoitetut prosessit louhivat tietoa monista eri lähteistä, kuten rahoituslaitosten dokumenteista, internet-hauista, hätäpuheluista sekä suurista puhelutietorekistereistä (Quinn, 2014). Tällainen yksityishenkilöiden tarkkailu valtionhallinnon toimesta herättää monia eettisiä kysymyksiä. Utilitaristisesta näkökulmasta hyödyt veropetosten, epidemioiden ja terroriaikeiden huomaamisesta voivat hyvinkin olla suuremmat kuin tarkkailun haitat kansalaisten yksityisyydelle. Lyon (2014) huomauttaa kuitenkin, että esimerkiksi terroriaikeiden valvonta massadataa analysoimalla voi johtaa menetelmän arvaamattomuudesta johtuen vääriin hälytyksiin, jotka voivat olla vahingollisia syyttömille epäillyille.

Crawfordin ja Schultzin (2014) mukaan joissakin Yhdysvaltain kaupungeissa poliisi allokoi partiointia louhimalla dataa rikosten tekopaikoista. Näin partiot tekevät tehostetun partioinnin alueilla todennäköisesti enemmän pidätyksiä suhteessa muihin alueisiin. Rikostilastoihin voi siis tulla tätä kautta vinoumia joidenkin alueiden osalta. (Crawford & Schultz, 2014.)

4.3 Viitekehysluonnos massadatan ja tiedonlouhinnan eettisten ongelmien hahmottamiseksi

Edellä on kuvailtu tutkimuskirjallisuuden esittämiä eettisiä kysymyksiä massadatan ja tiedonlouhinnan hyödyntämiseen liittyen. Sekä yksilöön että yhteiskuntaan liittyviä ongelmakohtia on havaittu. Seuraavaksi esitellään

koostava taulukko löydöksistä (taulukko 2). Ongelmat on jaoteltu sen mukaan, mihin massadatan tai tiedonlouhinnan hyödyntämisen osa-alueeseen ne erityisesti liittyvät. Lisäksi löydökset on luokiteltu edelleen sen mukaan, koskettavatko ne yksilöä vai yhteiskuntaa tai molempia näistä. Taulukko toimii luonnoksena kyseisten teknologioiden eettisten ongelmien viitekehykseksi.

TAULUKKO 2. Koostava esitys massadatan ja tiedonlouhinnan eettisistä ongelmista tutkimuskirjallisuuteen perustuen.

<table>
<tr><td></td><td colspan="2">Eettisiä ongelmakohtia</td></tr>
<tr><td>Massadatan/tiedonlouhinnan hyödyntämisen osa-alue</td><td>Yksilön näkökulma</td><td>Yhteiskunnan näkökulma</td></tr>
<tr><td rowspan="2">Massadatan kerääminen</td><td>Datankeruuluvan voimassaolo (Custers, 2016)</td><td>Datankeruun ulkopuolelle jääminen (Lerman, 2013)</td></tr>
<tr><td>Datankeruu erillisellä luvalla/kieltäytymättä jättämisen perusteella (Quinn, 2014)</td><td>Yksilönvapauden rajoittuminen tarkkailun tunteen vuoksi (Johnson, 2008)</td></tr>
<tr><td rowspan="2">Massadatan säilyttäminen</td><td>Säilytyksen sovelias kesto (Custers, 2016)</td><td>Siirtäminen erilaisen lainsäädännön piiriin (Johnson, 2008)</td></tr>
<tr><td>Oikeus nähdä itsestä kerätty data (Tene & Polonetsky, 2012)</td><td rowspan="5"></td></tr>
<tr><td rowspan="2">Massadatan julkistaminen</td><td>Anonymisoinnin teho (Zimmer, 2010)</td></tr>
<tr><td>Arkaluonteisen tiedon johtaminen ei-arkaluonteisesta datasta (Acquisti & Gross, 2009)</td></tr>
<tr><td rowspan="2">Massadatan myyminen</td><td>Organisaatioiden erilaiset datakäytännöt (Davis, 2012)</td></tr>
<tr><td>Myynti erillisellä luvalla/kieltäytymättä jättämisen perusteella (Quinn, 2014)</td></tr>
<tr><td rowspan="2">Tiedonlouhinnan analyysimenetelmät</td><td>Tiedonlouhinnan tarkkuus (Lazer ym., 2014)</td><td rowspan="2"></td></tr>
<tr><td colspan="2">Oppivien algoritmien arvaamaton toiminta (Crawford & Schultz, 2014)</td></tr>
</table>

Tiedonlouhinnan tulosten hyödyntäminen	Ihmisiä koskevien tärkeiden päätösten automatisointi (Tene & Polonetsky, 2012)	
	Hyödyntäminen laittoman liiketoiminnan peittämiseksi (Crawford & Schultz, 2014)	
	Tulosten käyttö markkinoinnissa (Crawford & Schultz, 2014)	
	Kohdennettu lainvalvonta (Crawford & Schultz, 2014); terroriaikeiden ennustaminen (Lyon, 2014)	

Edellä oleva taulukko on siis tiivistetty kooste tämän tutkielman tärkeimmistä tuloksista. Esityksestä saa karkean yleiskuvan massadataan ja tiedonlouhintaan liittyvistä eettisistä ulottuvuuksista. Näin se toimii eräänlaisena luonnoksena aiheen hahmottamisen viitekehykseksi. Karkeuden vuoksi yksittäisten esimerkkien informatiivisuus on matala. Tästä syystä jokaista eettistä ongelmaa kohden on merkitty kirjallisuusviitteet, jotta lukija voi halutessaan etsiä käsiinsä alkuperäistekstit. Tämän viitekehyksen hyödyntämisen apuna voi käyttää sivulla 22 esiteltyjä ohjelmistokehitysammattilaisen eettisiä periaatteita.

4.4 Tulosten pohdintaa

Massadataan ja tiedonlouhintaan liittyy tulosten perusteella monimuotoisia eettisiä haasteita. Esitellyt tulokset eivät ole missään nimessä tyhjentävä luettelo kaikesta aiheeseen liittyvästä tutkimuksesta. Tämän kandidaatintutkielman puitteissa ei olisi ollut mielestäni järkevää käydä seikkaperäisesti läpi jokaista eettiseksi ongelmaksi tulkittavissa olevaa yksityiskohtaa. Käsiteltyjen teemojen kirjo on kuitenkin mielestäni riittävän kattava yleiskäsityksen saamiseksi aiheesta.

Pyrkimykseni oli kartoittaa sekä yksilöön että yhteiskuntaan liittyviä ongelmatilanteita. Kuten taulukosta näkyy, kirjallisuudesta löytyi enemmän yksilön näkökulmasta esitettyjä huolia. Toisaalta rajanveto yksilöä ja yhteiskuntaa koskettavien teemojen välillä on itse tekemäni eikä se ole ehdottoman tarkka. Periaatteessa on mahdollista ajatella kaikkien yksilöä koskevien huolien olevan skaalautuvia siinä määrin, että ne voitaisiin luokitella myös yhteiskunnallisiksi haasteiksi.

On huomioitava, että tekemäni jako massadatan ja tiedonlouhinnan hyödyntämisen eri osa-alueiden välille ei ole poissulkeva. Monet esitellyistä eettisistä ongelmista voivat siis käytännössä kuulua useampaan eri luokkaan.

Johtopäätöksenä voidaan todeta, että esitettyjen eettisten huolenaiheiden määrä ja vakavuus kertovat massadatan ja tiedonlouhinnan vastuullisen hyödyntämisen olevan tärkeää. Puutteellisen lainsäädännön vuoksi organisaatioiden oman arvostelukyvyn varaan jää paljon tässä suhteessa. Erilaisten yksityisyyden säilyttävän tiedonlouhinnan menetelmien ja anonymisointitekniikoiden

käyttöä tulisi suosia silloin kun se on mahdollista. Laajamittaisen datankeruun synnyttämien huolien vuoksi organisaatioissa on myös syytä pohtia, milloin massadataa tarvitsee ylipäätään kerätä.

5 YHTEENVETO

Tämän tutkielman tavoitteena oli selvittää, mitä eettisiä ongelmia liittyy massadatan ja tiedonlouhinnan hyödyntämiseen. Yhtenä tärkeänä motiivina oli Moorin (1985) esittämä ajatus käytäntötyhjiöistä, joita seuraa informaatioteknologian kehityksestä. Kitchin (2014) kuvailee massadataa disruptiiviseksi innovaatioksi, joten sen tutkiminen tästä näkökulmasta nähtiin aiheelliseksi. Lisäksi Hanin ym. (2011) mukaan tiedonlouhinnan yhteiskunnalliset vaikutukset vaativat selvittämistä, sillä kyseessä on arjen mullistava teknologia. Näin ollen katsoin tutkielman aiheen olevan mielekäs toteutettavaksi.

Ensimmäinen tutkimuskysymykseni oli: ”Mitä eettisiä ongelmia liittyy massadatan ja tiedonlouhinnan hyödyntämiseen organisaation toiminnassa?” Tähän kysymykseen vastatakseni kävin ensin läpi kyseisten teknologioiden erityispiirteitä ja taustaa. Tämän jälkeen käsittelin etiikan keskeisimpiä teorioita ja niiden merkitystä tutkielmani kannalta. Viimeisessä sisältöluvussa kävin läpi aineistosta löytyneitä mainintoja massadatan ja tiedonlouhinnan eettisistä ongelmista. Lopuksi koostin löydökset yhden taulukon alle viitekehysluonnokseksi.

Tutkielman toinen tutkimuskysymys oli: ”Mitä yksilöitä ja yhteiskuntaa koskevia teemoja tulisi huomioida massadataa ja tiedonlouhintaa koskevassa eettisessä pohdinnassa?” Tämän kysymyksen muotoilu on huomioitu tärkeimmät löydökset koostavassa taulukossa. Pyrin luvussa 4 jäsentämään kaikki kirjallisuudessa esitetyt huomiot teemoittain sekä sen mukaan, koskevatko ne ensisijaisesti yksilöä vai yhteiskuntaa. Luonnostelemassani viitekehyksessä nämä jaottelut ilmentyvät mielestäni selkeästi. Itse taulukon suppean esitysmuodon vuoksi yksittäiset ongelmat ovat vain karkeasti muotoiltuja, minkä takia jokaista ongelmaa kohden on ilmoitettu lähde.

Tutkielman vahvuutena onnistuin mielestäni vastaamaan tutkimuskysymyksiin hyvin. Niihin vastaavia löydöksiä esiteltiin melko kattavasti. Koen tutkielman aiheen taustoituksen olevan myös riittävän laaja aiheen rajausta silmälläpitäen.

Tutkielman heikkoutena voidaan pitää sitä, että tietojenkäsittelyn etiikan kirjallisuudessa oli varsin vähän syvällistä pohdintaa aiheeseen liittyvästä

eettisestä problematiikasta. Sovelsinkin löytyneisiin ongelmakuvauksiin jonkin verran etiikan klassisia teorioita. Käytettyjen teorioiden valinta perustui Quinnin (2014) näkemykseen parhaiten soveltuvista teorioista informaatioteknologian eettisten ongelmien pohdintaan. Kaiken kaikkiaan tieteellistä aineistoa aiheesta löytyi mielestäni riittävästi kirjallisuuskatsauksen tekemistä varten.

Tämän tutkielman kontribuutiona toimii luonnos viitekehyksestä, joka soveltuu massadataa ja tiedonlouhintaa koskevien eettisten ongelmien tarkasteluun. Tätä viitekehystä voi käyttää yleiskuvan saamiseksi aiheesta. Lisäksi näiden teknologioiden parissa työskentelevä voi hyödyntää viitekehystä ja esittelemiäni Quinnin (2014) eettisiä periaatteita pohdinnan tukena. Kumpikaan ei kuitenkaan anna valmiita vastauksia käytännön ongelmatilanteita varten. Perusteena tälle käyttötarkoitukselle toimii Masonin ym. (1995, 161) ajatus siitä, että IT-alan ammattilaisten tulee itse pohtia eettisiä ratkaisuja epäselvissä ongelmatilanteissa.

Yksi mielenkiintoinen jatkotutkimusaihe tutkielman tiimoilta voisi olla massadataa ja tiedonlouhintaa hyödyntävien organisaatioiden toimintatapojen kartoitus näiden teknologioiden käyttöön liittyen. Myös IT-alan ammattilaisten eettiset asenteet koskien massadataa ja tiedonlouhintaa voisivat olla hyvä aihe empiiriselle tutkimukselle.

LÄHTEET

Acquisti, A., & Gross, R. (2009). Predicting Social Security numbers from public data. *Proceedings of the National academy of sciences*, *106*(27), 10975-10980.

Aggarwal, C. C., & Yu, P. (2008). *A general survey of privacy-preserving data mining models and algorithms* (s. 11-52). Springer US.

Anderson, R. E. (1992). ACM code of ethics and professional conduct. *Communications of the ACM*, *35*(5), 94-99.

Aristoteles (1989). *Nikomakhoksen etiikka*. (S. Knuuttila, toim.) Helsinki: Gaudeamus.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662-679.

Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). An overview of machine learning. In *Machine learning* (s. 3-23). Springer Berlin Heidelberg.

Crawford, K., & Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.*, *55*, 93.

Culnan, M. J. (1993). "How Did They Get My Name?": An Exploratory Investigation of Consumer Attitudes toward Secondary Information Use. *MIS quarterly*, 341-363.

Culnan, M. J., & Armstrong, P. K. (1999). Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization science*, *10*(1), 104-115.

Custers, B. (2016). Click here to consent forever: Expiry dates for informed consent. *Big Data & Society*, *3*(1), 2053951715624935.

Davenport, T. H., Barth, P., & Bean, R. (2012). How big data is different. *MIT Sloan Management Review*, *54*(1), 43.

Davis, K. (2012). *Ethics of Big Data: Balancing risk and innovation*. Sebastopol: O'Reilly Media, Inc..

Duhigg, C. (2012, 16. helmikuuta). How companies learn your secrets. The New York Times. Haettu 3.3.2016 osoitteesta: http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?_r=0

Dwork, C. (2006). Differential privacy. In *Automata, languages and programming* (s. 1-12). Springer Berlin Heidelberg.

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*, *346*(6210), 1243089.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Fieser, J. (2000). *Moral philosophy through the ages*. Kalifornia: Mayfield Publishing Company.

Floridi, L. (1999). Information ethics: On the philosophical foundation of computer ethics. *Ethics and information technology*, *1*(1), 33-52.

Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy & Technology*, 1-3.

Floridi, L. (2013). Distributed morality in an information society. *Science and engineering ethics*, *19*(3), 727-743.

Floridi, L. (toim.). (2008). *The Blackwell guide to the philosophy of computing and information*. Malden: John Wiley & Sons.

Gartner (2012). The importance of 'big data': a definition. *Stamford, CT: Gartner*, 2014-2018.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques*. 3. painos. Waltham: Elsevier.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. MIT press.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, *47*, 98-115.

Hill, K. (2012, 7. helmikuuta). Max Schrems: The thorn in Facebook's side. Forbes. Haettu 19.4.2016 osoitteesta: http://www.forbes.com/sites/kashmirhill/2012/02/07/the-austrian-thorn-in-facebooks-side/#4954e856b309

Huang, Z., Du, W., & Chen, B. (2005). Deriving private information from randomized data. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data* (s. 37-48). ACM.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, *52*(8), 36-44.

Johnson, D. (2008). Computer Ethics. Teoksessa L. Floridi (toim.), *The Blackwell guide to the philosophy of computing and information*. Malden: John Wiley & Sons.

Kant, I. (1998). Duty and Categorical Rules. Teoksessa J.P. Sterba (toim.), *Ethics: The Big Questions* (s. 171–185). Oxford: Blackwell Publishers.

Kitchin, R. (2013). Big data and human geography Opportunities, challenges and risks. *Dialogues in human geography*, *3*(3), 262-267.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 2053951714528481.

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 2053951716631130.

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

Laudon, K. C. (1995). Ethical concepts and information technology. *Communications of the ACM*, *38*(12), 33-39.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science*, *343*, 1203–1205.

Leonard, L. N., & Cronan, T. P. (2001). Illegal, inappropriate, and unethical behavior in an information technology context: A study to explain influences. *Journal of the Association for Information Systems*, *1*(1), 12.

Lerman, J. (2013). Big data and its exclusions. *Stanford Law Review Online*, *66*.

Liikenne- ja viestintäministeriö (2014). Big data –strategia. Haettu 3.3.2016 osoitteesta: http://www.lvm.fi/lvm-mahti-portlet/download?did=139030

Lyon, D. (2014). Surveillance, snowden, and big data: capacities, consequences, critique. *Big Data & Society*, *1*(2), 2053951714541861.

Marturano, A. (2002). The role of metaethics and the future of computer ethics. *Ethics and Information Technology*, *4*(1), 71-78.

Mason, R. O. (1986). Four ethical issues of the information age. *MIS Quarterly*, *10*(1), 5-12.

Mason, R. O., Mason, F. M., & Culnan, M. J. (1995). Ethics of Information Management. Thousand Oaks: SAGE Publications, Inc.

McSherry, F. D. (2009). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data* (s. 19-30). ACM.

Mingers, J., & Walsham, G. (2010). Toward ethical information systems: the contribution of discourse ethics. *MIS Quarterly*, *34*(4), 833-854.

Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*, *42*(11), 30-36.

Moor, J. H. (1985). What is computer ethics?. *Metaphilosophy*, *16*(4), 266-275.

Mutanen, T. (2007). Consumer Data and Privacy in Ubiquitous Computing. *VTT Publications 647*. Helsinki: VTT.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on* (s. 111-125). IEEE.

Negash, S. (2004). Business intelligence. *The communications of the Association for Information Systems, 13*(1), 54.

Quinn, M. (2014). *Ethics for the information age*. Essex: Pearson.

Rawls, J. (1998). Welfare Liberalism. Teoksessa J.P. Sterba (toim.), *Ethics: The Big Questions* (s. 222–237). Oxford: Blackwell Publishers.

Sanastokeskus TSK (2013a). Big data. Haettu 13.4.2016 osoitteesta: http://www.tsk.fi/cgi-bin/netmot.exe?UI=figr&qfind=big+data

Sanastokeskus TSK (2013b). Tiedonlouhinta. Haettu 13.4.2016 osoitteesta: http://www.tsk.fi/cgi-bin/netmot.exe?UI=figr&qfind=tiedonlouhinta

Sedayao, J., Bhardwaj, R., & Gorade, N. (2014). Making Big Data, Privacy, and Anonymization Work Together in the Enterprise: Experiences and Issues. In *Big Data (BigData Congress), 2014 IEEE International Congress on* (s. 601-607). IEEE.

Smith, H. J. (2002). Ethics and information systems: Resolving the quandaries. *ACM SIGMIS Database*, *33*(3), 8-22.

Smith, H. J., & Hasnas, J. (1999). Ethics and information systems: the corporate domain. *MIS Quarterly*, 109-127.

Stahl, B. C., Timmermans, J., & Mittelstadt, B. D. (2016). The Ethics of Computing: A Survey of the Computing-Oriented Literature. *ACM Computing Surveys (CSUR)*, *48*(4), 55.

Sterba, J.P. (1998). *Ethics: The Big Questions*. Oxford: Blackwell Publishers.

Tene, O., & Polonetsky, J. (2012). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology & Intellectual Property*, *11*, xxvii.

Tietosuojaryhmä (2014). Lausunto 5/2014 anonymisointitekniikoista. 0829/14/FI WP216.

TIVIA (2002). Tietotekniikan ammattilaisen etiikan ohjeisto. Haettu 26.4.2016 osoitteesta: http://www.tivia.fi/julkaisut/etiikan-ohjeet

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

Zimmer, M. (2010). "But the data is already public": on the ethics of research in Facebook. *Ethics and information technology*, *12*(4), 313-325.