Teemu Pinola

# DATATIETEILIJÄN KOMPETENSSIEN MÄÄRITTELEMINEN

JYVÄSKYLÄN YLIOPISTO
TIETOJENKÄSITTELYTIETEIDEN LAITOS
2015

# TIIVISTELMÄ

Pinola, Teemu
Datatieteilijän kompetenssien määritteleminen
Jyväskylä: Jyväskylän yliopisto, 2015, 53 s.
Tietojärjestelmätiede, pro gradu -tutkielma
Ohjaaja(t): Luoma, Eetu

Tässä pro gradu – tutkielmassa tarkasteltiin datatieteilijän kompetensseja. Kompetensseja lähestyttiin kahdesta eri näkökulmasta, tieteellisen kirjallisuuden pohjalta sekä empiirisesti työpaikkailmoitusten kautta. Tutkimusmenetelmänä käytettiin sisällönanalyysiä. Tieteellisen kirjallisuuden pohjalta määriteltiin keskeiset käsitteet massadata ja massadata-analytiikka, sekä valittiin tutkimuksessa käytetty viitekehys. Tutkimuksessa koottiin tieteellisessä kirjallisuudessa esiintyneet datatieteilijän kompetenssit. Empiirinen aineisto koostui 94 työpaikkailmoituksesta, joista eriteltiin datatieteilijältä vaaditut kompetenssit. Tieteellisestä kirjallisuudesta ja empiirisestä aineistosta kerättyjä kompetensseja vertailtiin yhtenäisen viitekehyksen avulla. Aiempaa tutkimustietoa datatieteilijän kompetensseista on hyvin vähän. Tulevaisuudessa datatieteilijöistä odotetaan olevan pulaa, joten on tärkeää tietää mitä datatieteilijän tulisi osata. Tämä on oleellista, jotta organisaatiot osaisivat rekrytoida oikein ja koulutusorganisaatiot opettaa oikeita asioita. Tutkimustulokset osoittavat, että tieteellisessä kirjallisuudessa ja työpaikkailmoituksissa datatieteilijältä vaadituissa kompetensseissa ei ollut kovin suuria eroja. Yllättävin ero oli datatieteilijältä odotetussa toimialaosaamisessa. Kirjallisuudessa datatieteilijältä odotettiin toimialaosaamista, mutta sitä ei vaadittu kuin yhdessä työpaikkailmoituksessa. Tutkimustulosten perusteella datatieteilijän tärkeimmät kompetenssit ovat tilastotieteellinen ja liiketoiminnallinen osaaminen, sekä analyyttiset taidot, ohjelmointi- ja kommunikointitaidot, koneoppiminen ja tiedonlouhinta. Lisäksi datatieteilijältä odotettiin intohimoa ja kykyä ratkaista liiketoiminnan ongelmia massadata-analytiikan avulla.

Asiasanat: datatieteilijä, kompetenssikehys, massadata, massadata-analytiikka

# ABSTRACT

Pinola, Teemu
Defining data scientist competencies
Jyväskylä: University of Jyväskylä, 2015, 53 p.
Information Systems, Master's Thesis
Supervisor(s): Luoma, Eetu

The purpose of this master's thesis is to study data scientist competencies, which were analyzed through scientific literature and job advertisements. The data was analyzed by using content analysis. Based on previous research the framework for this study was selected and key concepts big data and big data analytics were defined. The competencies of data scientist were collected from scientific literature. The empirical data consists of 94 job advertisements, from which the competencies required for data scientist were extracted. The competencies from these two sources where then compared by using the selected framework. Previous research considering data scientist competencies is lacking. In future there is expected to be a shortage of data scientists, so it is important to be aware of requirements for this job title. Furthermore this information is important for the recruitment as well as training of data scientists. The results of this study show that there were little differences between data scientist competencies from scientific literature and job advertisements. The most surprising difference was domain knowledge. In scientific literature domain knowledge was an important competency but it was mentioned only in one of the job advertisements. Based on the results of this study the most important competencies for data scientists are statistical and business knowledge, analytical, programming and communication skills, machine learning and data mining. In addition data scientist was required to be passionate and skilled in solving business problems through the use of big data analytics.

Keywords: big data, big data analytics, competency framework, data scientist

# KUVIOT



# TAULUKOT

# SISÄLLYS

# 1 JOHDANTO

Massadata (big data) on seuraavan sukupolven tietovarastointia (data warehousing) ja liiketoiminnan analytiikkaa (Minelli, 2013). Sille on ladattu todella suuret odotukset, sillä massadata-analytiikan avulla pyritään muun muassa vähentämään kustannuksia, tekemään nopeita ja parempia päätöksiä, sekä tuottamaan uusia tuotteita ja palveluita (Davenport, 2015). Käytännössä organisaatiot olettavat siis saavansa uusia liiketoimintamahdollisuuksia, sekä tuottaa lisäarvoa jo olemassa oleville liiketoiminnoille (Davenport, Bart & Bean, 2012). Organisaatiot tuottavat teratavuittain dataa lähes päivittäin. Tällaista dataa kertyy esimerkiksi erilaisista sensoreista ja lokeista. Data on yleensä sellaisessa muodossa, jota ei voi tallentaa perinteisiin relaatiotietokantoihin. Uusien sovellusten ja teknologioiden avulla myös relaatiotietokantoihin sopimattomasta datasta voidaan tuottaa liiketoimintahyötyä (Zikopoulos, 2012). Vuonna 2012 uutta dataa syntyi päivittäin noin 2.5 eksatavua. Päivittäin syntyvän datan määrän odotetaan kaksinkertaistuvan 40 kuukauden välein (McAfee & Brynjolfson, 2012).

Massadata-analytiikan oletetaan synnyttävän suuren tarpeen datatieteilijöille (data scientists). Yksistään Yhdysvalloissa odotetaan vuonna 2018 olevan pulaa 140000 – 190000 analyyttisiä taitoja omaavasta työntekijästä, sekä 1.5 miljoonasta data-tajun omaavasta johtajasta, jotka pystyvät käyttämään massadata-analytiikan tuotteena syntynyttä tietoa tehokkaasti hyödykseen päätöksen teossaan (Manyika, Chui, Brown, Bughin, Dobbs, Roxburgh, Byers & McKinsey Global Institute, 2011). Tämän takia on tärkeää tutkia, millaista osaamista organisaatioiden tulisi vaatia datatieteilijältä, jotta he osaisivat palkata oikean osaamisen omaavia henkilöitä tekemään massadata-analytiikkaa.

Massadata on suhteellinen käsite, eikä sillä pyritä pelkästään kuvaamaan datajoukkojen kokoa. Yleisimmän määrityksen mukaan sillä on kolme eri ulottuvuutta: määrä, monimuotoisuus ja nopeus. Määrällä kuvataan datajoukkojen suurta kokoa, monimuotoisuudella tuodaan esille että data on yleensä peräisin monesta eri lähteestä ja nopeudella tarkoitetaan datan vaihtuvuutta (Gantz & Reinsel, 2012). Yleisesti massadatalla kuitenkin tarkoitetaan datajoukkoja, jotka

ovat liian suuria perinteisille datan prosessointityökaluille ja näin ollen vaativat uusia prosessointiteknologioita.

Massadata-analytiikalla tarkoitetaan yleisemmin ennustavaa (predictive) ja ohjailevaa (prescriptive) analytiikkaa. Ennustavalla analytiikalla pyritään nimenmukaisesti ennustamaan todennäköisyyksiä tulevaisuudessa tapahtuville tapahtumille. Ohjailevalla analytiikalla pyritään valitsemaan tulevat toiminnot, joilla voidaan saavuttaa paras mahdollinen tulos (Minelli, 2013).

Datatieteilijä -tehtävänimeke kuvaa massadata-analyytikkoa. Massadata-analyytikon odotetaan tekevän löytöjä datasta, joiden avulla saavutetaan suuria liiketoiminnallisia hyötyjä (Davenport & Patil, 2012). Datatieteilijä -tehtävänimike on noussut otsikoihin massadatan suuren suosion avustuksella. Googlen pääekonomisti Hal Varian (2009) on sanonut: ”Seuraavat kymmenen vuotta tilastotieteilijä on seksikkäin työtehtävä. Ihmiset luulevat että vitsailen, mutta kuka olisi arvannut että tietokoneinsinöörit ovat 1990-luvun seksikkäin työtehtävä?”. Thomas Davenport ja D. Patil (2012) artikkelissaan ”Data Scientist: The Sexiest Job of the 21st Century” korostavat suurta kysyntää datatieteilijöille ja heidän harvinaisille taidoilleen.

Tässä tutkimuksessa pyritään selvittämään millaista osaamista ja millaisia taitoja datatieteilijällä tulisi olla. IDG Enterprise:n (2014) tekemässä tutkimuksessa havaittiin, että organisaatioilla on vaikeuksia löytää päteviä työntekijöitä massadatahankkeisiinsa. Tutkimuksen mukaan organisaatiot kuitenkin suunnittelivat työllistävänsä tai kouluttavansa uusissa massadatahankkeissa tarvittavat ammattilaiset. Tällä hetkellä myös Suomessa on pulaa datatieteilijöistä (Rastas & Esp, 2014). Davenportin ja Patilin (2012) mukaan vielä vuonna 2012 ei ollut olemassa yhtään koulutusohjelmaa datatieteilijälle. Nykyisin niitä löytyy jo jonkin verran. Suomessa ensimmäinen datatieteilijän koulutusohjelma käynnistyi viime syksynä (2014) Aalto-yliopistossa. Massadatan analysointi vaatii kuitenkin tekijältään uudenlaisia taitoja, joita ei vielä kunnolla tunneta. Rastaksen ja Espin (2014) ”Big datan hyödyntäminen” –raportissa korostetaan alan toimijoiden, yritysten ja koulutusorganisaatioiden välistä yhteistyötä, jotta datatieteilijän tarvitsemat kompetenssit saataisiin määriteltyä. On siis todella tärkeää tutkia millaista osaamista datatieteilijältä työelämässä odotetaan, jotta koulutus vastaisi mahdollisimman hyvin olemassa olevaan kysyntään.

## 1.1 Tutkimuskysymys

Massadata on käsitteenä suhteellisen tuore. Eri toimijoilla on erilaisia käsityksiä sen sisällöstä ja lopullinen määritelmä hakee muotoaan. Tästä syystä on oletettavaa, että eri toimijat viittaavat samalla epämääräisesti määritellyllä käsitteellä eri asioihin. Muun muassa Provost & Fawcett (2013) korostavat, että usein suurten datojen käsittely ja massadata menevät mediassa sekaisin. Myös Davenport (2014) pitää sekavuutta lisäävänä ongelmana sitä, että massadata -käsite on kuuma termi, jota erityisesti konsultit käyttävät palveluidensa myynninedistämisessä. Joissakin tapauksissa massadatalla saatetaan virheellisesti viitata jopa

perinteiseen analytiikkaan tai ääritapauksissa normaaliin raportointiin tai liiketoimintatiedon hallintaan (business intelligence) (Davenport, 2014).

Tässä tutkimuksessa keskitytään erityisesti siihen, täsmääkö kirjallisuudessa esiintyneet datatieteilijän kompetenssit työpaikkailmoituksissa vaadittuihin taitoihin. Aiemmassa kirjallisuudessa on epäjärjestelmällisesti lueteltu erinäisiä datatieteilijän kompetensseja ilman yhtenäistä viitekehystä. Tämän tutkielman tarkoitus on koota yhteen yhtenäiseen kehykseen nämä kompetenssit soveltaen Havelka & Merhoutin (2009) esittelemää kompetenssikehystä. Tutkimuksen tavoitteena on selvittää, datatieteilijältä vaadittuja kompetensseja.

1. Millaisella viitekehyksellä voidaan kuvata datatieteilijän kompetensseja?
2. Ovatko empiirisesti havaitut datatieteilijän kompetenssit linjassa kirjallisuudessa esiintyneiden kompetenssien kanssa?

Tämä tutkielma sisältää viisi sisältölukua. Seuraavassa luvussa käsitellään tarkemmin massadataa ja massadata-analytiikkaa, sekä esitellään niihin liittyviä teknologioita. Massadata–käsitettä tarkastellaan kolmesta eri näkökulmasta: määrä, monimuotoisuus ja nopeus. Teknologioista esitellään muutama tuote ja tekniikka yleisemmällä tasolla. Lopuksi annetaan yleiskuvas massadata-analytiikasta ja esitellään CRISP-DM (Cross Industry Standard Process for Data Mining) tiedonlouhinnan prosessimalli.

Kolmannessa luvussa määritellään kompetenssi, sekä valitaan kompetenssikehys kirjallisuudessa ja työpaikkailmoituksissa esiintyneiden kompetenssien esittämiseen yhteisessä viitekehyksessä. Kompetenssikehyksen valinta tehdään vertailemalla useita eri kehyksiä. Lisäksi luvussa käsitellään kirjallisuudessa esiintyneitä datatieteilijän kompetensseja.

Neljännessä luvussa kuvataan miten tutkimus on tehty ja esitetään tutkimuksen tulokset. Aluksi esitellään tutkimusmenetelmät ja tutkittava aineisto, sitten käydään läpi tutkittavasta aineistosta kerätyt kompetenssit ja lopuksi vertaillaan kirjallisuuden pohjalta toteutettua viitekehystä työpaikkailmoituksista saatuihin kompetensseihin.

Viidennessä luvussa pohditaan tieteellisessä kirjallisuudessa ja empiirisessä aineistossa havaittujen kompetenssien eroja. Kompetenssit listataan yhteiseen viitekehykseen vertailun helpottamiseksi. Eroavaisuudet nostetaan esiin ja pohditaan niiden merkitystä.

Viimeinen luku sisältää yhteenvedon tutkimuksesta. Se sisältää lyhyen tiivistelmän tutkimuksesta, vastaukset tutkimuskysymyksiin ja jatkotutkimusaiheet.

# 2 MASSADATA JA MASSADATA-ANALYTIIKKA

Tässä luvussa määritellään mitä massadata on ja millaisia sovelluksia ja teknologioita tarvitaan, jotta organisaatioiden on mahdollista harjoittaa massadata-analytiikkaa. Lisäksi luvussa käsitellään analytiikan ja massadata-analytiikan välistä suhdetta ja esitellään lyhyesti liiketoimintatiedon hallinta, joka liittyy oleellisesti massadata- analytiikkaan. Lopuksi määritellään mitä massadata-analytiikka oikeastaan on.

## 2.1 Määritelmä

Massadatatermille ei ole olemassa tarkkaa määritelmää. On todella vaikeaa tietää, mitä termillä eri lähteissä tarkoitetaan. John Gantz ja David Reinsel (2012) esittelevät massadatan koostuvan kolmesta eri ominaisuudesta: datasta, data-analytiikasta ja analytiikan tulosten kommunikoinnista sidosryhmille. Adam Jacobs (2009) kuvaa määritelmän ajasta riippuvaksi. 1980-luvulla massadatan määritelmä on ollut eri kuin mikä se nykyään on. Nykyään massadata voisi kuvailla dataksi, jota ei voi analysoida perinteisten tietokantatyökalujen avulla. Chenin ym. (2012) mukaan massadatalla tarkoitetaan datajoukkoja, jotka ovat kerätty useasta eri lähteestä, eikä niitä voi hallita perinteisin menetelmin. Massadata-käsitteeseen yhdistetään vahvasti myös data-analysoinnin tuotteena syntynyt liiketoimintahyöty (McAfee & Brynjolfson, 2012). Danah Boydin ja Kate Crawfordin (2012) mukaan massadatalla tarkoitetaan kykyä hakea, yhdistää ja viitata isojen datajoukkojen välillä.

Termiä massadata on kritisoitu paljon. Muun muassa Thomas Davenport (2014) on ennustanut massadatakäsitteelle lyhyttä elinkaarta sen vaikean määriteltävyyden ja sekavuuden vuoksi. Itse asiassa massadatateknologioita hyödyntävät yritykset eivät mielellään käytä massadatakäsitettä, vaan he pyrkivät kuvaamaan asian rakenteellisemmin esim. ”Analysoimme videodataa pankkiautomaateilta ja toimipisteiltämme, jotta saisimme paremman ymmärryksen asia-

kassuhteistamme.” on paljon kuvaavammin ilmaistu kuin ”Analysoimme massadataa.”. Kun datan yhteydestä jättää sana iso tai pieni pois, välttyy loputtomalta keskustelulta mikä on massadataa ja mikä ei. Alla oleva taulukko 1 havainnollistaa aineistojen analysoinnista käytettyjen käsitteiden evoluutiota:

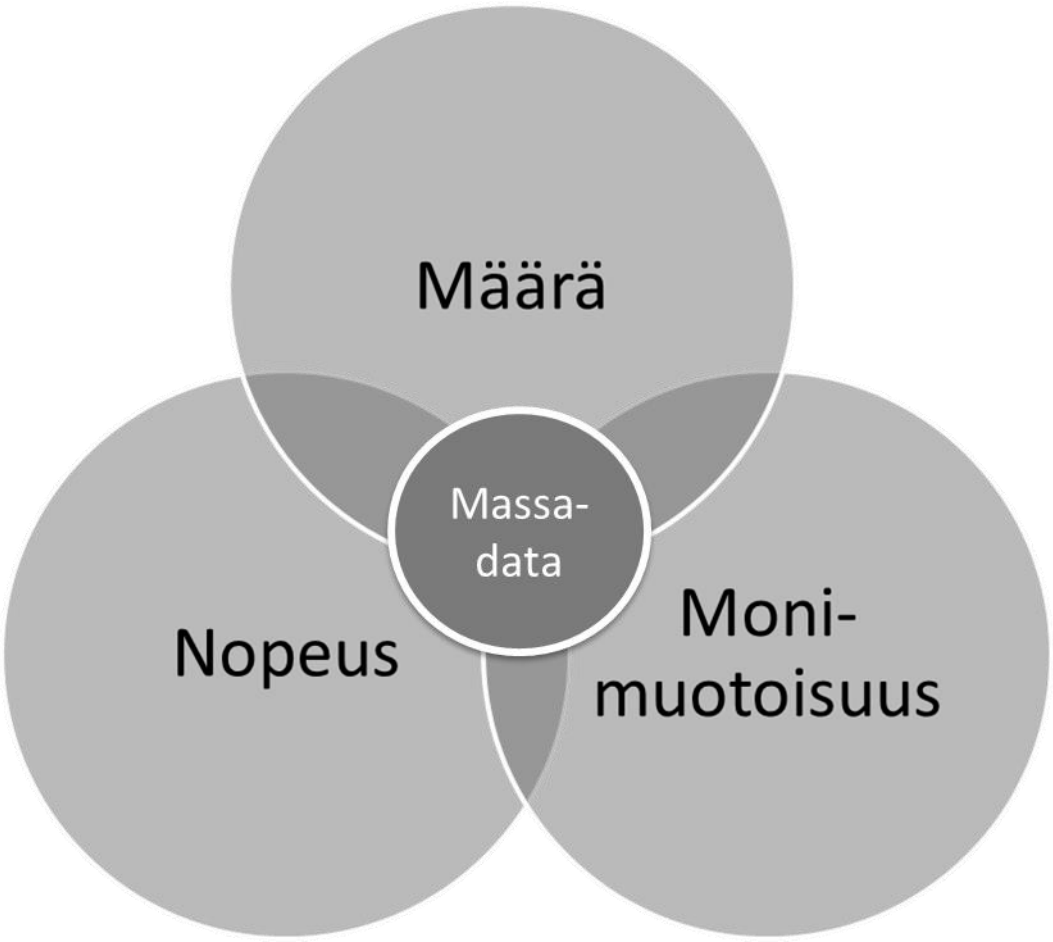
TAULUKKO 1 Datan analysoinnissa käytetyn käsitteistön kehitys 1970-luvulta tähän päivään (Davenport 2014, s.10)

| **Käsite** | **Aikaikkuna** | **Erityinen tarkoitus** |
|---|---|---|
| Päätöksenteon tuki (Decision support) | 1970-1985 | Datan analysointi päätöksenteon tukena |
| Johdon tuki (Executive support) | 1980-1990 | Keskittyminen data-analyysiin ylemmän johdon päätöksenteon tueksi |
| "Online-analyysi" (Online analytical processing, OLAP) | 1990-2000 | Ohjelmistot moniulotteisen datan analysointiin |
| Liiketoimintatiedon hallinta (Business intelligence) | 1989-2005 | Työkaluja tukemaan datalähtöisiä päätöksiä, painotus raportoinnissa |
| Analytiikka (Analytics) | 2005-2010 | Keskittyminen tilastolliseen ja matemaattiseen analyysiin päätöksenteossa |
| Massadata (Big data) | 2010- | Keskittyminen erityisen suuriin, strukturoimattomiin, "nopeasti liikkuviin” -aineistoihin |

NewVantagePartners teetti vuonna 2012 kyselyn massadatasta yli 50 suuren organisaation johtajalle. Kyselyn tulokset osoittavat, että aineiston rakenteettomuus on suurempi haaste kuin aineiston suuri määrä, ja että analysointitarpeet eroavat hyvinkin suuresti yritysten välillä. Esimerkiksi 30% vastaajista piti pääsyynä massadata-analytiikan käytölle useista eri lähteistä (monimuotoisuus) olevan tiedon analysointia, 22% keskittyi pääasiassa uusien aineistomuotojen analysointiin ja 12% reaaliaikaisen datavirran (streaming data) analysointiin (Davenport, 2014). Massadatakäsitteen vaikeaa määriteltävyyttä kuvastaa se, että kyseisessä NewVantagePartnersin tutkimuksessa vain 28% vastaajayrityksistä analysoi yli 1 TB:n aineistoa, mikä ei massadatastandardeilla ole kovin suuri aineisto (Davenport, 2014).

Yleisimmän käsityksen mukaan massadatalla on kolme eri ulottuvuutta: määrä (volume), nopeus (velocity) ja monimuotoisuus (variety) (Russom, 2011, Zikopoulos, 2012, McAfee & Brynjolfson 2012, Gantz & Reinsel, 2012, Gartner,

2013). Kuvio 1 kuvaa kuinka datan määrä, nopeus ja monimuotoisuus yhdistyvät massadatassa.




KUVIO 1 Massadatan kolme ulottuvuutta (Russom, 2011)

### 2.1.1 Määrä

Datan määrä on tuntomerkeistä ominaisin massadatalle. Yleensä massadatan tilavuuden määritellään olevan teratavusta ylöspäin. Määrä voidaan kuitenkin kuvata myös muina yksikköinä, kuten tiedostojen, asiakirjojen, taulukoiden tai transaktioiden lukumääränä (Russom, 2011). Eri organisaatioissa dataa kertyy vaihtelevia määriä. Esimerkiksi Facebookissa dataa kertyy 10 teratavua päivittäin, mutta joissakin organisaatioissa dataa saattaa kertyä jopa useita teratavuja tunnin välein ympäri vuoden (Zikopoulos, 2012).

Kertyvän datan määrä kasvaa jatkuvasti. Dataa tallentuu kaikkialta: ympäristöstä, taloudesta, lääketieteestä, valvonnasta ja kaikesta muusta mitä seurataan elektronisten apuvälineiden avulla. Lähes jokainen klikkaus tietokoneella generoi jonkinlaista dataa käytettävien järjestelmien lokeihin. Kuten edellä olevista esimerkeistä ja itse massadatatermistä käy ilmi, organisaatiolla on hallussaan valtavia määriä dataa. Mikäli dataa ei osata hallita, organisaatiot hukkuvat siihen. Suuressa datamäärässä on kuitenkin myös uusia mahdollisuuksia.

Oikeiden työkalujen ja teknologioiden avulla melkein kaikki organisaatioiden data pystytään analysoimaan. Analysoinnin tuotteena voi syntyä parempi ymmärrys omasta liiketoiminnasta, asiakkaista ja markkina-asemasta (Zikopoulos, 2012).

### 2.1.2 Monimuotoisuus

Monimuotoisuudella pyritään kuvaamaan datan olemusta. Massadatassa yhdistyy strukturoitu, osittain strukturoitu ja strukturoimaton data. (Russom, 2011). Strukturoidulla datalla tarkoitetaan perinteisiä datalähteitä, joilla on selkeä rakenne. Rakenne pysyy samana ajasta riippumatta ja se tekee strukturoidusta datasta helposti käsiteltävää. Strukturoimattomalla datalla ei ole rakennetta. Se voi olla esimerkiksi tekstiä, videoita tai ääntä. Osittain strukturoitu data pitää sisällään jonkin tasoista rakennetta, mutta sitä ei voi suoraan hyödyntää analytiikassa. Lokitiedostot ovat hyvä esimerkki osittain strukturoidusta datasta (Franks, 2012 s.14, Zikopoulos, 2012).

Data kertyy useista eri lähteistä, kuten sensoreista ja älylaitteista. Perinteisillä analytiikan sovelluksilla voi olla vaikeuksia käsitellä kaikkea generoituvaa dataa, myös niiden suorituskyky voi osoittautua liian heikoksi tietokantakyselyiden suorittamiseksi. Perinteiset relaatiotietokannat eivät myöskään pysty tallentamaan strukturoimatonta dataa (Zikopoulos, 2012).

Noin 80 prosenttia generoituvasta datasta on osittain strukturoitua tai strukturoimatonta dataa ja 20 prosenttia datasta on strukturoitua dataa, jota on mahdollista tallentaa relaatiotietokantoihin. Massadatan avulla organisaatioiden on mahdollista tuoda kaiken tyyppinen data analytiikan piirin. Organisaatiot pystyvät näin ollen hyödyntämään analytiikassaan niin strukturoitua kuin strukturoimatontakin dataa (Zikopoulos, 2012).

### 2.1.3 Nopeus

Yksinkertaistettuna nopeudella tarkoitetaan uuden datan muodostumisvauhtia. Perinteisesti nopeus on yhdistetty datan saapumis-, tallennus- ja hakuaikaan, mutta massadatan yhteydessä sillä on muutakin merkitystä. Nopeutta tarkastellessa onkin tärkeämpää kiinnittää huomiota datavirtojen vauhtiin kuin datavarastojen kasvuasteisiin. Nykypäivän organisaatiossa datavirrat ovat jatkuvia RFID-sensoreiden lisääntymisen ja muiden tietovirtojen johdosta (Zikopoulos, 2012).

Etu suhteessa muihin organisaatioihin syntyy tunnistamalla trendi eli kehityssuuntaus, ongelma tai mahdollisuus ennen muita kilpailijoita. Organisaatioiden tuleekin pystyä analysoimaan data lähes reaaliajassa, jotta kilpailuetu suhteessa muihin olisi mahdollista saavuttaa. Reaaliaikaista datan analysointia ei tule tehdä pelkästään kilpailuedun, vaan myös datan säilyvyyden vuoksi. Enenevissä määrin nykyisin tuotetun data säilyvyys/käytettävyysaika on hyvin lyhyt (Zikopoulos, 2012).

Monet eri organisaatiot ovat kehittäneen Russomin (2011) massadatan määritelmää eteenpäin lisäämällä siihen uusia ulottuvuuksia. Esimerkiksi IBM:n (2012) määritelmän mukaan massadatalla on myös todellisuus (veracity) - ulottuvuus. Sillä pyritään kuvaamaan liiketoimintajohdon luottamusta massadatasta jalostettuun informaatioon. Ongelma korostuu, kun monimuotoisuuden myötä datan lähteiden määrä kasvaa merkittävästi. Todellisuus – ulottuvuudella halutaan korostaa, että ilman luottamusta massadatasta jalostettuun tietoon, tietoa ei voida hyödyntää täysimääräisesti organisaatioiden päätöksenteossa. SAS (2013) on puolestaan lisännyt massadatamääritelmään kaksi eri tuntomerkkiä: vaihtelevuuden (variability) ja monimutkaisuuden (complexity). Vaihtelevuudella tarkoitetaan datavirtojen epäjohdonmukaisuutta. Datavirroissa esiintyy jaksoittaisia piikkejä. Monimutkaisuudella halutaan kuvata dataentiteettien yhteensovittamisen vaikeutta. Mikäli organisaation täytyy linkittää dataa eri järjestelmien ja liiketoiminta yksikköjen välillä, täytyy ymmärtää kuinka eri lähteistä tulevan datan voi liittää yhteen.

## 2.2 Tekniikat ja teknologiat

Perinteiset relaatiotietokannat eivät yksistään riitä massadatan hallintaan, koska massadata on yhdistelmä strukturoitua ja strukturoimatonta dataa. Strukturoimattomalla datalla tarkoitetaan dataa, jonka rakennetta ei ole ennalta määrätty, eikä sitä näin ollen voida tallentaa relaatiotietokantaan (Bakshi, 2012). Myös datajoukkojen koko on kasvanut niin suureksi, että uusia teknologioita ja tiedonhallintajärjestelmiä on jouduttu kehittämään massadatan hallintaan (Provost & Fawcett, 2013). Massadatan hallintajärjestelmät ja sovellukset tarvitsevat siis uuden ja kehittyneen tavan tallentaa, hallita, analysoida ja visualisoida dataa (Chen ym., 2012). Esimerkiksi tekniikat ja työkalut kuten MapReduce, Hadoop ja NoSQL ovat kehitetty vastaamaan massadatan vaatimuksiin (Bakshi, 2012).

### 2.2.1 MapReduce

MapReduce on Googlen kehittämä kehys (framework) suurten datajoukkojen prosessointiin. Sen avulla on helppo kehittää skaalauntuvia rinnakkaisajosovelluksia (parallel application), jotka prosessoivat isoja määriä dataa suurissa tietokoneklustereissa. MapReducen rajapinnassa on kaksi funktiota; kartoita (map) ja supista (reduce), jotka mahdollistavat mallin käyttämisen esimerkiksi datan prosessoinnissa, hakukoneissa ja koneoppimisessa (machine learning) (Yang, Dasdan, Hsiao & Parker, 2007).

Kehyksen tarkoituksena on piilottaa monimutkaiset asiat, kuten rinnakkaisajo (parallelization), vikasietoisuus (fault-tolerance), tiedonjakelu (data distribution) sekä kuormantasaus (load balancing) taustalle, ja mahdollistaa ohjelmoijan ilmaista pelkät suoritettavat laskutoimitukset. Laskenta suoritetaan

avain/arvo (key/value) parien avulla. Inspiraatio kehyksen luomiselle tuli useista funktionaalisista ohjelmointikielistä, jotka pitävät sisällään kartoita ja supista funktiot (Dean & Ghemawat, 2008).

Kartoita (map) -funktio ottaa syötteenä parin ja tuottaa joukon väliaikaisia (intermediate) avain/arvo pareja. MapReduce-kirjasto ryhmittää yhteen kaikki väliaikaiset arvot, jotka liittyvät samaan väliaikaiseen avaimeen. Tämän jälkeen avaimen arvo välitetään supista (reduce) -funktiolle. Supista -funktio hyväksyy väliaikaisen avaimen arvon ja joukon avaimeen liittyviä arvoja. Se yhdistää (merge) arvot mahdollisesti pienemmäksi arvojoukoksi. Supista kutsulla saadaan yleensä yksi tai ei yhtään tulosarvoa. Väliaikaiset arvot tuodaan supista funktioille iteratiivisesti. Näin ollen sillä voidaan käsitellä myös arvojoukkoja, jotka ovat liian suuria mahtuakseen tietokoneen muistiin (Dean & Ghemawat, 2008).

### 2.2.2 Hadoop

Hadoop on Javalla ohjelmoitu avoimen lähdekoodin toteutus Googlen kehittämästä MapReduce-kehyksestä. Siinä tiedostot jaetaan lohkoiksi, jotka replikoidaan ja levitetään kaikille verkon palvelimille, joissa Hadoop on käytössä. Hadoopissa myös jokainen pyyntö (job) jaetaan pienemmiksi kokonaisuuksiksi: tehtäviksi. Pyyntöä suoritettaessa tehtävät ajetaan yksittäisillä erillisillä palvelimilla ja tulos palautetaan vasta, kun kaikki yksittäiset tehtävät on suoritettu (Fischer, Su & Yin, 2011). Hadoopin yhteydessä yleensä esiintyvät myös termit Hbase, HDFS (Hadoop Distributed File System), Pig, Hive, ZooKeeper, Chukwa ja Avro.

Hbase on sarake-orientoitunut (column-oriented) jaettu (distributed) tietokanta, joka on toteutettu HDFS:N päälle. Se ei ole perinteinen relaatiotietokanta, eikä se siis näin ollen tue SQL-kieltä (Bashki, 2012). Sarake-oriontuneella tietokannalla HBase:n yhteydessä tarkoitetaan Googlen kehittämää Bigtable-mallia. Siinä jokainen taulu sisältää rivejä ja sarakkeita ja tallentuu moniulotteisena hajanaisena karttana (multidimensional sparse map). Lisäksi jokaisesta solusta löytyy aikaleima. HBase tuo siis HDFS:ään jaetun, vikasietoisen ja skaalautuvan tietokannan, joka sallii satunnaisen pääsyn (random access) tallennettuun dataan (Taylor 2010).

Hadoopin hajautettu tiedostojärjestelmä (HDFS) on suunniteltu tallentamaan erittäin suuria datajoukkoja ja välittämään ne suoratoistona (stream) käyttäjän sovelluksille. Isoihin palvelinklustereihin hajautettu datavarasto antaa mahdollisuuden suorittaa laskentaa yhtä aikaa monilla eri palvelimilla. HDFS:n kapasiteettiä pystytään kasvattamaan tarpeen vaatiessa. Tiedostojärjestelmässä metadata ja sovelluksen data tallennetaan eri paikkoihin. Metadata tallennetaan sille omistetulle palvelimella, jota kutsutaan NimiSolmuksi (NameNode). Sovelluksen data tallennetaan muille palvelimille, joita kutsutaan DataSolmuiksi (DataNode). Kaikki palvelimet keskustelevat toisensa kanssa TCP-pohjaisen protokollan avulla (Shvachko, Kuang, Radia & Chansler, 2010).

Hive on toteutettu Hadoopin päälle. Se on tietovarasto-ohjelmisto, joka mahdollistaa kyselyt isoihin datajoukkoihin, sekä niiden hallinnan. Hive sisältää SQL-kielen kaltaisen QL-kielen, joka helpottaa käyttäjiä, joille SQL-kieli on tuttu, tekemään erilaisia kyselyitä. Mikäli QL-kielen käyttäjät tuntevat myös MapReduce-kehyksen, he voivat tehdä omia kartoita- ja supista-skriptejä kyselyihinsä (Thusoo, Sarma, Jain, Sha, Chakka, Zhang Antony, Liu & Murthy, 2010).

## 2.3 Liiketoimintatiedon hallinta ja data-analytiikka

Yksinkertaistettuna liiketoimintatiedon hallinnalla (business intelligence) tarkoitetaan datan jalostamista tiedoksi. Organisaatiot haluavat analytiikan avulla monenlaista tietoa; asiakkaiden tarpeista ja päätöksentekoprosesseista, omista kilpailijoista, toimialan tilanteesta, taloudesta, sekä teknologista ja kulttuurillisista trendeistä (Golfarelli, Rizzi & Cella, 2004). Liiketoimintatiedon hallinnassa siis yhdistyy data ja analyyttiset työkalut ja sen tarkoituksena on jalostaa tietoja organisaation päätöksentekijöille (Negash, 2004). Yleisesti analytiikalla tarkoitetaan datan muuntamista merkitykselliseksi tiedoksi. Massadata tuo analytiikan piirin isomman määrän informaatiota, jolloin voidaan olettaa, että siitä syntyy myös määrällisesti enemmän ja laadullisesti parempia oivalluksia (Minelli ym., 2013 s.103).

Taulukossa 2 esitellään kolme analytiikan kategoriaa: kuvaileva analytiikka, ennustava analytiikka ja ohjaileva analytiikka. Minelli ym. (2013 s.99) mukaan liiketoimintatiedon hallinta pitää sisällään vain kuvailevan analytiikan. Kuvailevan analytiikan avulla saadaan informaatiota jo tapahtuneista tapahtumista. Ennustava ja ohjaileva analytiikka ovat puolestaan massadata-analytiikkaa. Ne esitellään tarkemmin seuraavassa kappaleessa.

TAULUKKO 2 Liiketoimintatiedon hallinta ja massadata-analytiikka (Minelli et al, 2013 s.100)

| **Kuvaileva analytiikka (Liiketoimintatiedon hallinta)** | **Ennustava analytiikka** | **Ohjaileva analytiikka** |
|---|---|---|
| Mitä tapahtui?<br>Milloin tapahtui?<br>Kuinka usein jotain tapahtui, mihin se vaikutti?<br>Mikä on ongelma? | Mitä todennäköisesti tapahtuu seuraavaksi?<br>Mitä jos nämä trendit jatkuvat?<br>Mitä jos? | Mikä on paras vastaus?<br>Mikä on paras tulos milläkin todennäköisyydellä?<br>Mitkä ovat mahdollisimman erilaiset ja parhaat vaihtoehdot? |
| Tilastotiede | Tiedonlouhinta<br>Ennustava mallintaminen<br>Koneoppiminen<br>Ennustaminen<br>Simulointi | Rajoituspohjainen optimointi<br>Monitavoiteoptimointi<br>Globaalioptimointi |

Dataan kohdistuvan analytiikan avulla voidaan luoda uusia sovelluksia, jotka mahdollistavat datan jalostamisen tiedoksi. Erityisesti seuraavilla osa-alueilla analytiikan tuomat mahdollisuudet ovat tehneet vaikutuksen niin teollisuuteen kuin akateemisiin tutkijoihinkin: sähköinen kaupankäynti ja markkinoiden tunteminen (e-commerce and market intelligence), sähköinen hallinto (e-goverment and politics), tiede ja teknologia (science and technology), terveys ja hyvinvointi (smart health and well-being), ja turvallisuus ja yleinen hyvinvointi (security and public safety) (Chen ym., 2012).

Kehitettyjen teknologioiden avulla toteutetulla analytiikalla on mahdollista saavuttaa liiketoimintahyötyä. Jatkuvasti kehittyvien analysointitekniikoiden tutkimus voidaan luokitella viiteen eri kategoriaan: (big) data-analytiikka, teksti-analytiikka (text analytics), web-analytiikka (web analytics), verkkojen analytiikka (network analytics) ja mobiilianalytiikka (mobile analytics) (Chen ym., 2012). Taulukossa 3 esitellään mitä eri analysointitekniikat pitävät sisällään.

TAULUKKO 3 Analytiikan teknologiat ja kehittyvät tutkimukset (Chen ym., 2012)

| | **Perustuu teknologioihin** | **Kehittyvät tutkimukset** |
|---|---|---|
| **(Big) data analytiikka** | RDMS, tiedonvarastointi, ETL, OLAP, BPM, tiedonlouhinta, klusterointi, regressio, luokittelu, ydistys analyysi (association analysis), poikkeusten tunnistus, hermoverkostot, geneettiset algoritmit, monimuut- | Tilastollinen koneoppiminen, peräkkäinen ja ajoittainen louhinta, paikkatietojen louhinta, datavirtojen ja sensoridatan louhinta, prosessien louhinta, yksityisyyden säilyttävä tiedonlouhinta, verkon louhinta, |

| | | |
|---|---|---|
| | tujamenetelmät, optimointi ja heurestinen haku | kolumni-orientoitunut DBMS, muistissa oleva DBMS (in-memory DBMS), rinnakkain DBMS, pilvilaskenta, Hadoop, MapReduce |
| **Tekstianalytiikka** | Tiedonhaku, asiakirjojen näyttö, hakuprosessointi, relevanssipalaute, käyttäjän mallit, hakukoneet, organisaatioiden hakujärjestelmät | Tilastollinen NLP, tiedon poiminta, aihe mallit, kysymys-vastaus järjestelmät, mielipiteen louhinta, näkemyksen analysointi, web stylometric analyysi, monikielinen analyysi, tekstin visualisointi, multimedia IR, mobiili IR, Hadoop, MapReduce |
| **Web-analytiikka** | Tiedonhaku, laskennallinen kielitiede, hakukoneet, webindeksointi, sivustojen ranking, hakulokin analysointi, suosittelujärjestelmät, verkkopalvelut, mashups | pilvipalvelut, pilvilaskenta, sosiaalinen haku ja louhinta, mainejärjestelmät, sosiaalisenmedian analysointi, webin visualisointi, internetpohjaiset huutokaupat, internetin kaupallistaminen, sosiaalinen markkinointi, tietoturva |
| **Verkostoanalytiikka** | Bibliometrinen analyysi, viittausverkosto, yhteisjulkaisuverkosto, lainaus, sosiaalisen verkon teoriat, verkon mittarit ja topologia, matemaattiset verkkomallit, verkon visualisointi | Link mining, yhteisöjen tunnistus, dynaamiset verkkomallit, agentti-pohjainen mallintaminen, sosiaalinen vaikuttaminen ja tiedon diffuusiomallit, ERGMs, virtuaaliyhteisöt, rikollisverkostot, sosiaalinen/poliittinen analyysi, luottamus ja maine |
| **Mobiilianalytiikka** | Verkkopalvelut, älypuhelin alustat | mobiilit verkkopalvelut, mobiilit sovellukset, mobiilit tunnistussovellukset, mobiili sosiaalinen innovaatio, mobiili sosiaalinen verkostoituminen, mobiili visualisointi, personalisointi ja käyttäytymisen mallintaminen, gamification, mobiili mainostaminen ja markkinointi |

## 2.4 Massadata-analytiikka

Massadata-analytiikka on paljon helpommin määriteltävissä kuin itse massadata – termi. Massadata-analytiikka on aiemmin määriteltyyn massadataan tehtyä analytiikkaa (Russom 2011). Termiä ei kuitenkaan käytetä vielä kovin yleisesti, toki se on yleistymään päin. TDWI:n (The Data Warehousing Institute) vuonna 2011 tekemässä tutkimuksessa, johon osallistui lähes 360 alan asiantuntijaa, kävi ilmi, että massadata-analytiikasta ei välttämättä käytetä samaa nimeä eri organisaatioissa. Massadata-analytiikkaa kuvattiin myös nimillä edistyksellinen analytiikka (advanced analytics), analytiikka, suurten datajoukkojen analytiikka (large-volume or large-data-set analytics), tietovarastointi (data warehousing), tiedonlouhinta (data mining) ja ennustava analytiikka (predictive analytics) (Russom, 2011). Frank J. Ohlhorstin (2013 s. 4) kirjassa ”Big Data Analytics” massadata kuvataan liiketoiminta-analytiikan kokoavana terminä, jonka alle kuuluvat seuraavat käsitteet; liiketoimintatiedon hallinta (business intelligence), tiedonlouhinta, tilastolliset sovellukset (statistical applications), ennustava analytiikka ja tietojen mallintaminen (data modeling). Chen ym. (2012) käsittelevät kuitenkin massadata-analytiikkaa liiketoimintatiedon hallinnan alle kuuluvana käsitteenä, toisin kuin Ohlhorst (2013) toteaa kirjassaan. Käsitteiden hierarkkinen järjestys näyttäisi riippuvan täysin lähteestä. Täytyy siis olla tarkkani, mitä termiä missäkin organisaatiossa käytetään kuvaamaan massadata-analytiikkaa tai mitä sillä milloinkin tarkoitetaan. Lähtökohtana massadata-analytiikkaan on kuitenkin se, että massadata tuo lisäarvoa organisaatioihin tuomalla analytiikan piiriin myös sen datan, jota aikaisemmin organisaatioissa ei ole pystytty hyödyntämään (Zikopoulos, 2012).

Massadata-analytiikassa hyödynnetään ennustavaa ja ohjailevaa analytiikkaa. Ennustavalla analytiikalla pyritään selvittämään todennäköisyyksiä jollekin tietylle tulevaisuuden tapahtumalle. Yhdistämällä ennustava analytiikka ja massadata saadaan aikaa hyviä tuloksia. Tällöin tutkitaan koko otosta ilman rajoittavia oletuksia, joita on perinteisesti pitänyt tehdä. Ennustavan analytiikan avulla saatujen tulosten laatu on huomattavasti parantunut ja tarkentunut massadatateknologioiden ansioista. Ohjailevan analytiikan avulla pyritään nimensä mukaisesti ohjamaan tulevia toimintoja. Olemassa olevasta datasta saatujen tulosten avulla pyritään ohjaamaan tulevat toiminnot siten, että ne suoritettaisiin mahdollisimman optimaalisesti ja tehokkaasti (Minelli ym., 2013 s.99).

Reaaliaikainen data-analytiikka antaa uudenlaista ymmärrystä organisaation liiketoimintaympäristöstä, sekä auttaa kehittämään uusia tuotteita ja palveluita sen mukaan, miten esimerkiksi erilaiset käyttäytymismallit heihin vaikuttavat. Massadata-analytiikka erottuu perinteisestä data-analytiikasta kolmella eritavalla:

1. Siinä kiinnitetään huomiota enemmän datavirtoihin kuin varastoihin.

2. Data-analyytikon sijasta dataa käsittelevät datatieteilijät ja tuote- ja prosessikehittäjät
3. Data-analytiikalla pyritään palvelemaan ydinliiketoiminta sekä tuotannon toimintoja.

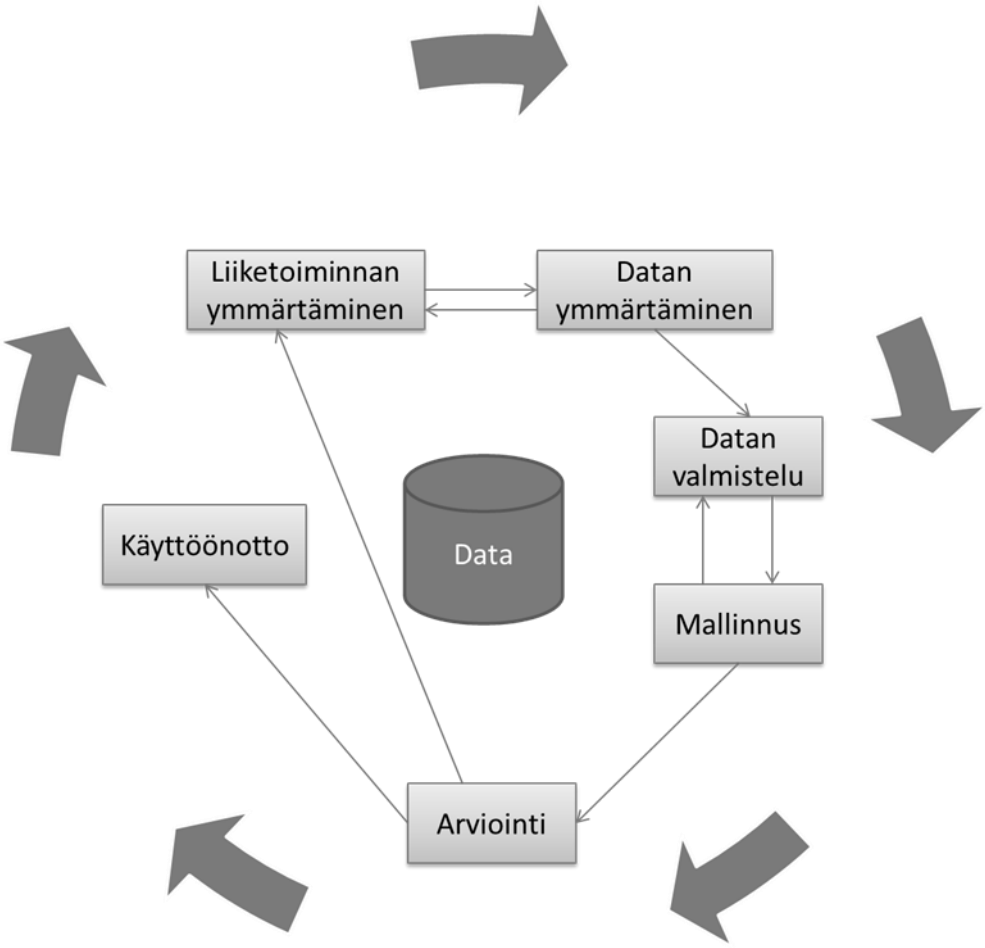
(Davenport, Barth & Bean, 2012).

Massadata-analytiikan tulokset esitetään usein visuaalisessa muodossa. Visuaalista analytiikkaa on suhteellisen helppoa tulkita ilman erityistä teknistä osaamista. Se ei kuitenkaan sovi kaikkeen. Tilastollisia malleja ja useiden muuttujien välisiä suhteita on todella vaikea kuvata visuaalisesti. Visuaalisesti esitetty data on enemmänkin kuvailevaa analytiikkaa kuin ennustavaa tai ohjailevaa analytiikkaa. Visualisoinnilla ei kuitenkaan aina saavuteta hyviä tuloksia, sillä sen avulla luodut esitykset voivat olla liian kompleksisia hyödynnettäviksi. On myös väitetty, että massadata on yhtä kuin yksinkertainen matematiikka. Oletamus perustuu siihen, että datan kerääminen ja sen muuntaminen strukturoituun muotoon on niin työlästä, että siihen ei enää jakseta tehdä monimutkaista tilastollista analytiikkaa. Tällöin tehdään vain yksinkertaisia esiintymistiheyslaskuja ja luodaan kuvat niiden pohjalta (Davenport, 2014).

### 2.4.1 CRISP –DM Tiedonlouhinnan prosessimalli

Massadata-analytiikassa ja yleisemmin data-analytiikassa tavoiteltava tieto löytyy käsiteltävästä datasta. Dataa pitää käsitellä ja jalostaa, jotta haluttu tieto saadaan siitä esille. CRISP-DM (Cross Industry Standard Process for Data Mining) on tiedonlouhinnan prosessimalli. Prosessi sisältää yhteensä kuusi eri vaihetta: liiketoiminnan ymmärtäminen, datan ymmärtäminen, datan valmistelun, mallintamisen, arvioinnin ja käyttöönoton.

Kuviosta 2 käy ilmi kuinka prosessin eri vaiheet linkittyvät toisiinsa. Tärkeää on myös huomata, että prosessin on iteratiivinen ja useimmissa tiedonlouhinta tapauksissa se suoritetaan useita kertoja ennen kuin haluttu lopputulos on saavutettu.




KUVIO 2 CRISP-DM tiedonlouhinnan prosessimalli (Provest & Fawcett, 2013)

CRISP-DM mallin ensimmäisessä vaiheessa tarkoituksena on selvittää mikä on ratkaistava ongelma. Ongelmaa määriteltäessä ensiksi keskitytään tarkastelemaan liiketoiminnan tavoitteita ja vaatimuksia. Kun liiketoiminta ongelma on saatu tunnistettua, siitä johdetaan varsinainen tiedonlouhintaongelma. Sen johtaminen liiketoimintaongelmasta on Provest ja Fawcettin (2013) mukaan haasteellista ja vaatii analyytikolta luovuutta. Tässä vaiheessa on erittäin tärkeää ymmärtää, mitä ongelman ratkaisulla pyritään saavuttamaan ja kuinka se tukee liiketoiminnan tarpeita (Provest & Fawcett, 2013).

Datan ymmärtämisvaiheessa pyritään selvittämään miten raakadatan avulla pystytään ratkaisemaan liiketoiminnan ymmärtämisvaiheessa yksilöity ongelma. Raakadata kerätään yleensä useista eri lähteistä, kuten asiakas- tai transaktiotietokannoista. Tällöin on mahdollista, että kerätty data sisältää päällekkäistä tietoa, joka saattaa heikentää datan luotettavuutta. On myös tärkeää ymmärtää raakadatan rajoitukset ja vahvuudet, sillä hyvin harvoin se sopii täsmällisesti ongelman ratkaisemiseen (Provest & Fawcett, 2013).

Jotta dataa on mahdollista käyttää analytiikassa, täytyy sen yleensä olla tietyssä muodossa. Data pitää valmistella hyvin ennen kuin sitä voidaan hyödyntää, sillä eri sovellukset odottavat saavansa datan tietyssä muodossa. Tyypillisesti dataa siistitään poistamalla tyhjät tiedot ja muuttamalla sitä eri muo-

toihin. Datan ymmärtämis- ja valmisteluvaihetta tehdään usein yhtä aikaa. Tällä pyritään mahdollistamaan datan muokkaus ja muuntaminen sellaiseen muotoon, että siitä saataisiin irti mahdollisimman hyvät tulokset (Provest & Fawcett, 2013).

Mallinnusvaiheessa ensiksi valitaan käytettävät mallinnustekniikat ja määritellään niiden käyttötavat. Valitulla mallinnustekniikalla luodaan malli tai kaava, joka kuvaa säännönmukaisuuksia datassa. Mallinnusvaiheessa tulee yleensä tarve palata takaisin datan valmisteluvaiheeseen, mikäli datan rakenteessa havaitaan ongelmia. Lisäksi mallinnusvaiheessa voidaan oivaltaa uusia tapoja muodostaa uuttaa tietoa (Wirth & Hipp, 2000, Provest & Fawcett, 2013).

Arviointivaiheessa tarkastellaan tuotettuja malleja, sekä niiden sopivuutta alkuperäisen liiketoimintaongelman ratkaisemiseksi. On tärkeää varmistua siitä, että raakadatasta tuotetut mallit ja kaavat kuvaavat todella todellisia säännönmukaisuuksia. Lisäksi arviointiin kuuluu mallien testaamista kontrolloidussa ympäristössä ennen niiden käyttöönottoa (Provest & Fawcett, 2013).

CRISP-DM -prosessin viimeinen vaihe on luodun mallin käyttöönotto. Mallin avulla saatu tieto ei itsessään ratkaise alkuperäistä liiketoimintaongelmaa, vaan tieto on pystyttävä esittämään käyttäjälle hyödynnettävässä muodossa. Riippuen alkuperäisen projektin vaatimuksista, hyödynnettävä tieto voi yksinkertaisimmillaan olla raportti tai jatkuva tiedonlouhintaprosessi. Tässä vaiheessa vastuu tiedon hyödyntämisestä jää loppukäyttäjälle (Provest & Fawcett, 2013).

# 3 KOMPETENSSI

Tässä luvussa määritellään, mitä kompetenssilla tarkoitetaan, valitaan tutkimuksessa käytettävä kompetenssikehys, sekä käsitellään datatieteilijän työssään tarvitsemia kompetensseja. Kompetenssikehys valitaan vertailemalla yhteensä viittä eri kompetenssikehystä. Lisäksi määritellään datatieteilijän työssään tarvitsemat kompetenssit, jotka on kerätty tieteellisestä kirjallisuudesta.

## 3.1 Kompetenssin määritelmä

Kompetenssilla tarkoitetaan henkilön kykyä suoriutua jostakin tietystä tehtävästä. CbBD (Competence Based Business Development) mallin mukaan kompetenssi koostuu henkilön tiedoista (eksplisiittinen/hiljainen), taidoista ja kyvyistä. Siihen vaikuttavat myös henkilön tarpeet, motiivit, tavoitteet, arvot ja asenteet. Kompetenssin omaava henkilö pystyy suoriutumaan määrätystä tehtävästä tehokkaasti ennalta määritellyssä ympäristössä (Schmiedinger, Valentin & Stephan, 2005). Yksinkertaistettuna työssä vaadittavalla kompetenssilla viitataan niihin taitoihin, jotka tekevät henkilöstä hyvän investoinnin työnantajalle (Bailey & Mitchell, 2006).

## 3.2 Kompetenssikehys

Kompetenssikehyksellä tarkoitetaan mallia, jota voidaan soveltaa yhteen tai useampaan työ- tai tehtävänimikkeeseen. Siinä määritellään millaisia kompetensseja (osaamista/taito) työntekijän tulee hallita selviytyäkseen työstään. Kompetenssikehyksen avulla voidaan helposti esittää mitä työntekijän työnkuvaan kuuluu. Näin ollen kaikilla on selvä käsitys siitä, mitä osaamista missäkin tehtävässä vaaditaan. Kehys helpottaa organisaation rekrytointia, työtehtävien sisäistä kiertoa sekä työntekijän ylenemistä ja urakehitystä organisaatiossa, koska kaikki tietävät mitä pitää osata missäkin työtehtävässä (Mansfield, 1996).

IT-ammattilaisten tarvitsemista kompetensseista on julkaistu useita tutkimuksia. Usein tutkimuksen yhteyteen on myös rakennettu kehys, jonka avulla voidaan paremmin havainnollistaa minkä tyyppisiä kompetensseja mikäkin tehtävä vaatii. Eri työtehtävät IT-alalla vaativat erilaisia kompetensseja. Näin ollen kompetenssikehyksen kategoriat täytyy olla melko abstraktit, jotta sillä pystyttäisi kuvaamaan jokaiseen työtehtävään vaadittavat tiedot, taidot ja kyvyt (Havelka & Merhout, 2009).

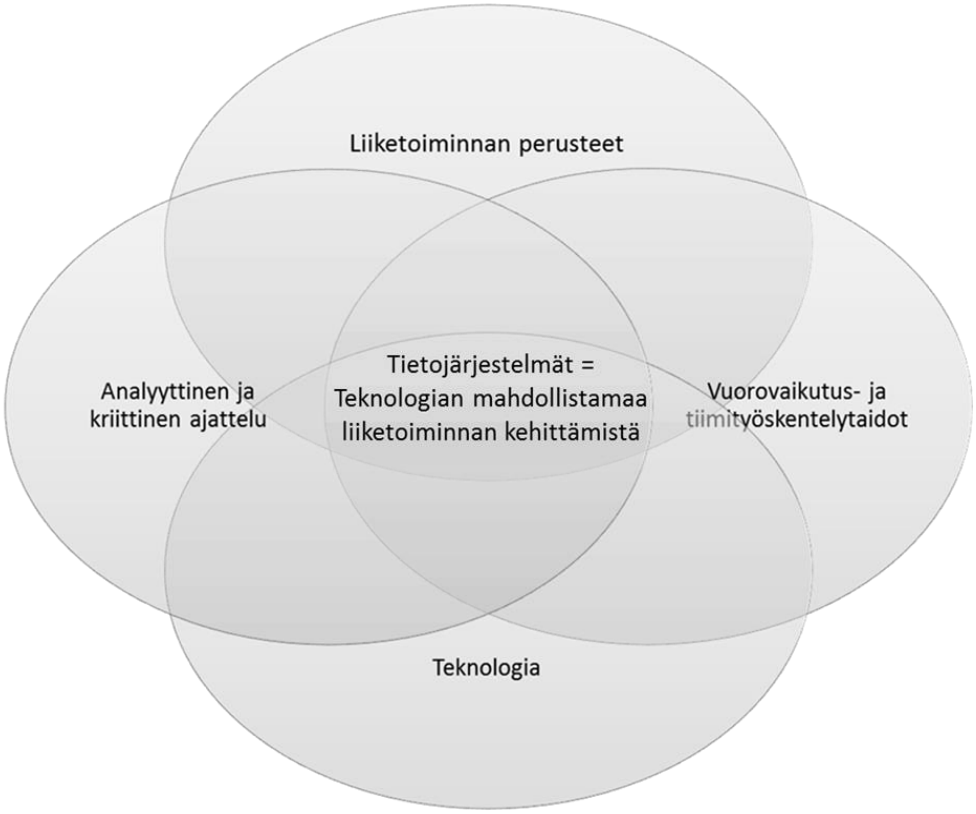
Todd ym. (1995) tutkivat IT-ammattilaisilta vaadittavaa osaamista analysoimalla IT-alan työpaikkailmoituksia vuosina 1970 - 1990. Tulosten esittämistä varten he kehittivät kompetenssikehyksen, johon listattiin kussakin ammatissa vaadittavia taitoja. Kehyksessä on kolme eri pääluokkaa: tekninen tietämys (technical knowledge), liiketoimintaosaaminen (business knowledge) ja järjestelmätietämys (systems knowledge). Luokat sisältävät kategorioita, joissa yksilöidään tarkemmalla tasolla millaisia kompetensseja kukin luokka pitää sisällään. ”Tekninen tietämys” -luokka on jaettu kahteen kategoriaan: laitteisto- ja ohjelmisto-osaamiseen. Liiketoimintaosaamisluokan alla on puolestaan kolme kategoriaa: vuorovaikutustaidot, liiketoiminta- ja johtamisosaaminen. Järjestelmätietämysluokka sisältää kaksi kategoriaa, jotka ovat ongelmien ratkaisukyky ja kehitysmenetelmät. Lee, Trauth ja Farwell (1995) suorittivat vastaavan kaltaisen tutkimuksen, jossa he selvittivät informaatiojärjestelmäammattilaisten muuttuvia osaamistarpeita. Tutkimuksessa vertailtiin, kohtaako IT-alalla annettu koulutus työtehtävissä vaadittavan osaamisen kanssa. Tutkimuksen kompetenssikehys sisälsi neljä eri kategoriaa: tekninen erityisosaaminen (technical specialities knowledge/skills), teknologioiden hallinta (technology management knowledge/skills), liiketoimintaosaaminen (business functional knowledge/skills) ja vuorovaikutustaidot (interpersonal and management knowledge/skills).

Todd ym. (1995) ja Lee ym. (1995) tutkimukset suoritettiin samana vuonna. Tutkimusten tulokset olivat hyvin samansuuntaisia, mutta pieniä eroja löytyi. Lee ym. (1995) mukaan IT-ammattilaisten työtehtävät keskittyvät aiempaa enemmän ohjelmistojen ympärille, joilla pyritään ratkaisemaan liiketoiminnan ongelmia. Näin ollen liiketoimintaosaaminen on välttämätöntä IT-alalla tehtävänkuvasta riippumatta. Todd ym. (1995) tutkimustulokset osoittavat, että työpaikkailmoituksissa esiintyneiden teknisten vaatimusten määrä suhteessa työntekijältä vaadittuun liiketoimintaosaamiseen oli kasvanut.

Lee ja Lee (2006) laajensivat Todd ym. (1995) kehittämää kompetenssikehystä. Tutkimuksessaan he tutkivat Fortune 500 -listalla olevien yritysten työpaikkailmoituksia. Todd ym. (1995) kompetenssikehykseen lisättiin yksi uusi kategoria: tietoliikenneosaaminen. Kategoria lisättiin, jotta pystyttäisiin tarkemmin erittelemään sähköisen liiketoiminnan mukana tuomat osaamisvaatimukset. Lee ja Lee (2006) tutkimuksen tulokset osoittavat, että IT-ammattilaisilla tulee olla myös muutakin kuin teknistä osaamista: hyviä vuorovaikutus- ja kommunikointitaitoja edellytettiin enemmän kuin kolmessa neljästä työpaikkailmoituksessa.

Havelka ja Merhout (2009) kehittivät oman ehdotuksensa IT-ammattilaisen kompetenssikehyksestä aiempien kehysten ja IT-alan johtajien haastattelujen pohjalta. Kehyksessä kompetenssit on määritelty neljään eri kategoriaan; henkilökohtaiset ominaisuudet, ammattitaito, liiketoimintaosaaminen ja tekninen osaaminen. Havelkan ja Merhoutin (2009) kompetenssikehys on yllä esitellyistä kompetenssikehyksistä kaikista laaja-alaisin. Se luokittelee IT-alan ammattilaisen vaadittavat taidot hyvin abstraktilla tasolla, minkä vuoksi se sopii hyvin monen eri tehtävänkuvan kompetenssikehykseksi.

ACM:n opetussuunnitelmassa vuonna 2002 todetaan IT-ammattilaisten tarvitsevan työssään liiketoimintaosaamista, analyyttista ja kriittistä ajattelukykyä, ihmissuhde- ja kommunikointi-taitoja, tiimityöskentelykykyä, sekä kykyä suunnitella ja toteuttaa tietojärjestelmiä, jotka parantavat organisaatioiden tehokkuutta (Gorgone ym., 2002). Kuviosta 2 käy hyvin ilmi, kuinka eri taidot ja osaamiset limittyvät keskenään, muodostaen kokonaisuuden IT-ammattilaisen tarvittavasta osaamisesta. Kokonaisuudesta muodostuu yksi kompetenssikehysmalli, johon voidaan listata ja kuvata IT-ammattilaiselta vaadittavia ominaisuuksia ja taitoja. ACM:n uusimmassa opetussuunnitelmassa 2010 käytetään samaa kehystä, mutta sen sisällä on painotuseroja verrattuna vuoden 2002 opetussuunnitelmaan (Topi ym., 2010). Topi ym. (2010) painottavat, että ”vaikka liiketoiminta tulee jatkossakin olemaan tietojärjestelmien pääsovellustoimiala, tarjoaa kyseinen oppihaara kriittisen tärkeää asiantuntemusta kasvavassa määrin myös muille aloille” ja esimerkkeinä he mainitsevat biologian, oikeustieteen ja terveydenhuollon.




KUVIO 3 ACM:n opetussuunnitelman esittämät IT-ammattilaisen osaamiskategoriat (Gorgone et al. 2002).

Taulukossa 4 vertaillaan jo esiteltyjä kompetenssikehyksiä suhteessa Havelka ja Merhout (2009) kehittämään kehykseen. Kuten taulukosta voi huomata, on Havelkan ja Merhoutin kehittämä kehys abstraktein esitellyistä kehyksistä. Tästä syystä se on valittu tämän tutkimuksen kompetenssikehyksesi, johon listataan datatieteilijän vaaditut taidot.

TAULUKKO 4 Kompetenssikehykset

| **Julkaisu/kehys** | **Havelka & Merhout (2009)** | **Lee, D.M.S, Trauth, E.M & Farwell, D. (1995)** | **Todd, McKeen & Gallupe (1995)** | **Lee S. M. & Lee, C. K. (2006)** | **Gorgone, Davis, Valacich, Topi, Feinstein & Longenecker (2002)** |
|---|---|---|---|---|---|
| **Kuvaus** | Aiemman kirjallisuuden ja haastattelujen pohjalta tehty synteesi | Tutkimus informaajärjestelmäammattilaisten muuttuvista osaamistarpeista (Society for Information Management) | Työpaikkailmoituksiin (1970-1990) pohjautuvat kategoriat | Laajennus Todd et al. (1995) kategorioihin | ACM:n IS2002 opetusohjelmasuositus. ACM:n päivitetty IS2010 (Topi, Valacich, Wright, |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Kaiser, Nunamaker, Sipior & Vreede (2010) perustuu samankaltaiseen kehykseen. |
| **Kompetenssikategoriat** | Henkilökohtaiset luonteenpiirteet (Personal traits) | | | | Luovuus (Analytical and critical thinking / Creativity) |
| | Ammattitaito (Professional skills) | Vuorovaikutus ja hallintaosaaminen (Interpersonal and management knowledge/skills) | Ongelmanratkaisutaito (Systems knowledge / Problem solving) | Ongelmanratkaisutaito (Systems knowledge / Problem solving) | Vuorovaikutus (Interpersonal, communication, and team skills / interpersonal) |
| | | | Kehittämismenetelmät (Systems knowledge / Development methodology) | Kehittämismenetelmät (Systems knowledge / Development methodology) | Ryhmätyö ja johtaminen (Interpersonal, communication, and team skills / Team work and leadership) |
| | | | | | Viestintä (Interpersonal, communication, and team skills / Communication) |
| | | | | | Ongelmanratkaisu (Analytical and critical thinking / Organizational problem solving) |
| | | | | | Etiikka ja ammattitaito (Analytical and critical thinking / Ethics and professionalism) |

| | | | | |
|---|---|---|---|---|
| Liiketoiminta-osaaminen (Business knowledge) | Liiketoimintaosaaminen (business functional knowledge/skills) | Liiketoiminta-osaaminen (Business knowledge / Business) | Liiketoiminta-osaaminen (Business knowledge / Business) | Liiketoiminta-mallit (Business fundamentals / Business models) |
| | | Johtamisosaaminen (Business knowledge / Management) | Johtamisosaaminen (Business knowledge / Management) | Liiketoiminnan funktionaaliset osa-alueet (Business fundamentals / Functional business areas) |
| | | Vuorovaikutustaidot (Business knowledge / Social) | Vuorovaikutustaidot (Business knowledge / Social) | Liiketoiminnan arviointi (Business fundamentals / Evaluation of business performance) |
| Tekninen osaaminen (Technical knowledge) | Tekninen erityisosaaminen (Technical specialties knowledge/skills) | Laitteisto-osaaminen (Technical knowledge / Hardware) | Laitteisto-osaaminen (Technical knowledge / Hardware) | Sovelluskehitys (Technology / Application development) |
| | Teknologian hallinta (Technology management knowledge/skills) | Ohjelmisto-osaaminen (Technical knowledge / Software) | Ohjelmisto-osaaminen (Technical knowledge / Software) | Verkkosovellusten arkkitehtuuri ja kehitys (Technology / Internet systems architecture and development) |
| | | | Tietoliikenneosaaminen (Architecture & Networks) | Tietokantojen suunnittelu ja ylläpito (Technology / Database design and administration) |
| | | | | Järjestelmäinfrastruktuuri ja integraatio (Technology / Systems infrastructure and integration) |

Bassellier ja Benbasat (2004) ovat kehittäneet kompetenssikehyksen IT-ammattilaisen liiketoimintaosaamisesta. Kehyksessä on kaksi kategoriaa organisaatiokohtainen ja ihmissuhteet ja johtaminen, joissa on yhteensä seitsemän eri osa-aluetta. Ensimmäiseen kategoriaan kuuluu organisaatiokohtaisia osa-alueita jotka ovat: organisaation yleiskatsaus, organisaation yksiköt, organisaation vastuut ja IT:n ja liiketoiminnan integrointi. Toisessa kategoriassa, ihmissuhteet ja johtaminen, on kolme eri osa-aluetta: kommunikointikyvyt, johtaminen ja tiedon verkostoituminen. Heidän kompetenssikehyksessä keskitytään pelkästään IT-ammattilaisen liiketoimintaosaamiseen, joten siitä puuttuu kokonaan IT-ammattilaisen tekninen osaaminen. Näin ollen Bassellierin ja Benbasatinn (2004) kehittämää kompetenssikehystä ei voida käyttää datatieteilijän kompetensseja kartoittaessa.

IT-ammattilaisen työssään tarvitsemat kompetenssit ovat hyvin laaja-alaisia. Liiketoimintaosaaminen on IT-ammattilaiselle yhtä tärkeää kuin tekniset taidot (Bassellier & Benbasat, 2004). Osaamista vaaditaan myös vuorovaikutustaitojen ja ongelmien ratkaisukyvyn osalta (Todd ym., 1995, Lee ym., 1995, Lee & Lee 2006). Nämä löydökset eivät sinänsä yllätä, sillä jo 1970-luvulla ACM :n opetussuunnitelma komitea tunnisti kuusi eri osa-aluetta/taitoa, jotka IT-ammattilaisen tulee hallita työskennelläkseen tehokkaasti: ihmiset, organisaatio, yhteiskunta, järjestelmät, tietokoneet ja mallit (Ashenhurst, 1972). On myös tärkeää huomata, että IT-ala kehittyy jatkuvasti, jonka seurauksena IT-ammattilaiselta vaaditaan myös hyvää oppimiskykyä. Uusien teknologioiden ja kehityssuuntien omaksumien on kriittisen tärkeä omaisuus IT- ammattilaiselle (Lu, Lo & Lin, 2011).

## 3.3 Datatieteilijän kompetenssit

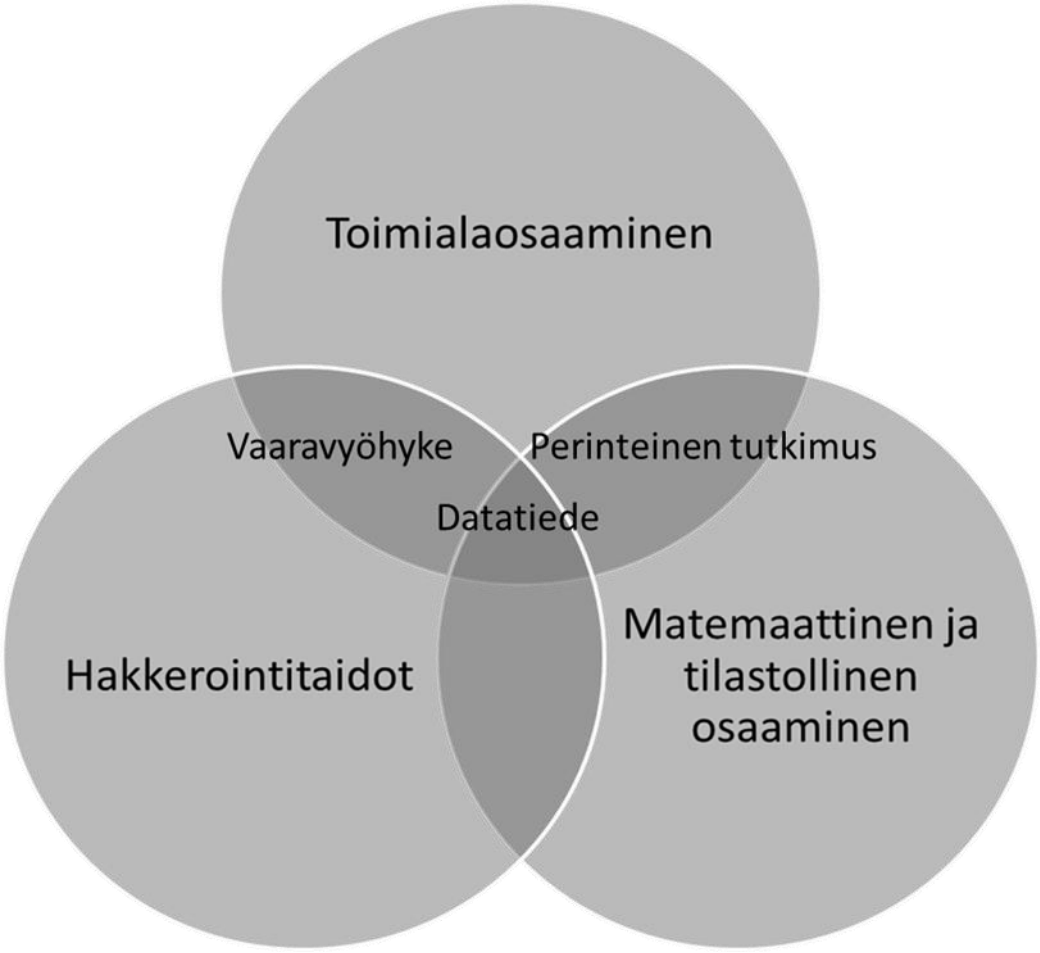
Datatieteilijän englanninkielinen tehtävänimike on "data scientist". Nimikkeen omaavia työntekijöitä kuvaillaan huippuammattilaisiksi, joilla on ammattitaito ja halu tehdä oivalluksia suurista datajoukoista. Datatieteilijän tulee osata yhdistellä eri datajoukkoja, sekä mahdollistaa strukturoimattoman datan analysointi. Analysoinnin tuotteena syntyneet tulokset tulee osata esittää ymmärrettävässä muodossa ja kertoa mitä tuloksista voidaan oppia, sekä miten tuloksia voidaan hyödyntää liiketoiminnassa (Davenport & Patil, 2012). Ohlhorstin (2013 s.30) mukaan datatieteilijän tulee osata eri analytiikan tekniikoita, kuten koneoppiminen ja tiedonlouhinta. Lisäksi hänen tulee hallita tilastotiedettä, ohjelmointia ja omata kokemusta algoritmeista. Tärkein datatieteilijän ominaisuus on kuitenkin kommunikointitaito. Datasta saatu tieto pitää pystyä kommunikoimaan selkeästi muille niin, että he ymmärtävät sen.

Datatieteilijä on joillain aloilla sijoitettu osaksi organisaation tuotekehitystä, kun taas perinteisesti analytiikan ammattilainen on sijoitettu organisaatioissa osaksi sisäistä konsultointia, avustamaan johtajia ja avainhenkilöitä sisäisessä

päätöksenteossa (Davenport, Barth & Bean, 2012). Datatieteilijä pyrkii vastamaan perinteisten data-ammattilaisten puutteisiin. Verrattaessa datatieteilijää kvantitatiiviseen analyytikkoon on huomioitava, että perinteinen kvantitatiivinen analyytikko voi olla hyvä analysoimaan dataa, mutta ei muuntamaan strukturoimatonta dataa analysoitavaan muotoon. Puolestaan tiedonhallinnan asiantuntija voi olla hyvä generoimaan ja organisoimaan strukturoitua dataa, mutta ei muuntamaan strukturoimatonta dataa strukturoiduksi tai suorittamaan itse analytiikkaa (Davenport & Patil, 2012). Toisin kuin Davenport ja Patil (2012), Salo (2013) toteaa, että perinteinen data-analyytikko soveltuu hyvin myös datatieteilijäksi. Yhtäläisyytenä perinteisen data-analyytikon osaamisalueisiin kuuluu liiketoimintaosaaminen, sekä tilastollisten ja muiden analyysitekniikoiden osaaminen. Ainoastaan teknologia on eri massadata-analytiikassa kuin data-analytiikassa. Salon (2013) mukaan data-analyytikko pystyy helposti omaksumaan uuden teknologian.

Conway (2011) mukaan datatieteilijän osaamisalueet jakautuu kolmeen pääryhmään: hakkerointitaidot, matemaattinen ja tilastollinen tietämys sekä toimialan tuntemus. Kuvio 3 kuvaa kuinka osaamisalueet limittäytyvät toistensa kanssa, muodostaen datatieteilijän tarvitsemat kompetenssit. Hakkerointitaidoilla ei viitata laittomaan toimintaan, vaan sillä pyritään yleisesti kuvaamaan tietokoneiden sujuvaa käyttötaitoa. Datatieteilijän täytyy pystyä keräämään data useista eri lähteistä ja järjestelmistä. Hänen tulee myös osata tekstitiedostojen manipulointi unix työkalujen avulla, kuten sed, awk ja grep. Edellä mainitut toimenpiteet mielettään jossain määrin epäjärjestelmällisiksi tavaksi toimia, jonka vuoksi osaamisalueen nimeksi valittiin hakkerointitaidot (Conway, 2011, Minelli, Chambers ja Dhiraj, 2013). Davenport (2014) nostaa hakkerointi taidoista esille ohjelmointitaidon, sekä kyvyn ymmärtää massadatateknologioiden arkkitehtuuria. Tärkeimpänä osana arkkitehtuuritietämystä on Hadoop/MapReduce - tuoteperheen tuntemus. Yleisesti hakkeri-termillä on negatiivinen sävy, joka yhdistetään laittomuuksiin, eikä aina suotta. On siis tärkeää, että datatieteilijän piirteistä hakkerin ominaisuudet eivät nouse hallitseviksi (Conway, 2011, Davenport, 2014).

Matemaattinen ja tilastollinen tietämys osaamisalueella viitataan datan analysointiin. Datatieteilijän tulee ymmärtää, kuinka dataa tulee lähestyä, ja mitä tekniikkaa sen analysointiin tarvitsee milloinkin käyttää, jotta datasta saataisiin esille oleellinen tieto. Viimeisenä osaamisalueena tulee toimialan tuntemus. On oleellisen tärkeää tietää, mitä ollaan tutkimassa ja mistä siinä on kyse. Ilman toimialan tuntemusta datatieteilijä ei voi tietää, mitä hän on tekemässä (Conway, 2011).




KUVIO 4 Big data analyytikon osaamisalueet (Conway 2011)

Davenport (2014) kuvailee datatieteilijän kompetenssien koostuvan viidestä eri piirteestä: hakkeri, tiedemies, luotettuneuvonantaja, kvantitatiivinen analyytikko ja liiketoiminnan asiantuntija. Hakkerin kompetenssit kuvattiin yllä Conwayn (2011) osaamiskehyksen yhteydessä. Tiedemiehen piirteistä halutaan korostaa kykyä ja asennetta suorittaa erinäisiä tehtäviä. Tarkemmin sillä tarkoitetaan taitoa muodostaa kokeita, suunnitella kokeellisia laitteita/sovelluksia, sekä kerätä, analysoida ja kuvailla datasta saatuja tuloksia. Datatieteilijän odotetaan myös osaavan tehdä päätöksiä saatujen tulosten pohjalta. Jotta analysoitavat tulokset saadaan muodostettua, tarvitaan päättäväisyyttä, improvisointitaitoja ja kykyä selvitä itse vastaan tulevista haasteista. Tämä ei kuitenkaan tarkoita, että datatieteilijältä vaaditaan jotain tiettyä tutkintoa. Monilla nykyisistä datatieteilijöistä ei ole ollenkaan akateemista tutkintoa (Davenport, 2014).

Luotettavana neuvonantajana datatieteilijällä tulee olla hyvät kommunikointi- ja ihmissuhdetaidot. Nämä taidot ovat kuitenkin hyvin harvinaisia datatieteilijöiden keskuudessa. Datatieteilijöitä käytetään yleisesti sisäisessä päätöksen teossa. Päätöksen tekijät eivät välttämättä ymmärrä kaikkia tärkeitä päätökseen liittyviä data ja analyyttisiä kysymyksiä. Tästä johtuen on tärkeää, että datatieteilijä kommunikoi itse päätöstentekijöiden kanssa (Davenport, 2014). Gartner:n tutkimuksessa havaittiin että 70 – 80 prosenttia liiketoimintatiedon

hallintaprojekteista epäonnistuu. Tutkimuksen mukaan se johtuu surkeasta kommunikoinnista IT- ja liiketoimintaosastojen välillä. Projekteissa ei ole osattu kysyä oikeita kysymyksiä tai ajatella liiketoiminnan oikeita tarpeita (Goodwin, 2011).

Kvantitatiivisen analyytikon taidoista on hyötyä, kun data on saatu siihen muotoon, että sitä voidaan perinteisin menetelmin analysoida. Matemaattiset ja tilastolliset tekniikat pitää siis olla hallussa ja ne pitää pystyä selittämään helposti myös ei-teknisille henkilöille. Massadata-analytiikan ja perinteisen data-analytiikan välillä on kuitenkin myös pieniä eroja. Esimerkiksi tilastollinen päättely (statistical inference) ei ole massadata-analytiikassa kovin oleellinen, sillä analytiikkaa tehdään koko otokseen, eikä vain pieneen osaan siitä, niin kuin perinteisessä data-analytiikassa. Massadata-analytiikassa ei tällöin myöskään tarvitse ottaa huomioon tilastollista merkittävyyttä (statistical significance) tai sitä, kuinka todennäköisesti saadut tulokset esittävät koko otosta. Toinen ero massadata-analytiikan ja perinteisen data-analytiikan välillä on, että massadata-analytiikassa tulokset pyritään usein esittämään visuaalisessa muodossa. Visuaalisesti esitetyt analytiikan tulokset ovat suhteellisen helposti kaikkien ymmärrettävissä. On kuitenkin tärkeää huomata, että visuaalinen analytiikka ei sovi kovin hyvin kuvaamaan useiden muuttujien välisiä suhteita tai tilastollisia malleja (Davenport, 2014).

Liiketoiminnan asiantuntijana datatieteilijän tule tietää, kuinka liiketoiminta toimii. Liiketoimintatietämyksen avulla massadata-analytiikalla voidaan mahdollisesti kehittää toimintatapoja ja ratkaista liiketoiminnan ongelmia. Datatieteilijän tulee olla kiinnostunut toimialastaan, sekä hänellä tulee olla kyky ratkaista liiketoiminnan ongelmia (Davenport, 2014).

Datatieteilijältä vaaditaan pitkä lista kompetensseja ja onkin todella harvinaista että ne löytyvät yhdeltä ja samalta henkilöltä. Todennäköisesti organisaatioissa massadata-analytiikka ei harjoita vain yksi henkilö vaan aiheen ympärille on kerätty tiimejä. Tiimi on rakennettu useista henkilöistä, jotka kukin vastaavat yhtä tai useampaa datatieteilijältä vaadittua kompetenssia. Tiimiä koottaessa on tärkeää, ettei ohjelmointitaitoa painoteta liikaa. Ohjelmointitaidon sijasta yksilön valinnassa tulisi painottaa enemmän analyyttisiä taitoja, kuten ongelmanratkaisukykyä (Provost ja Fawcett, 2013 s.35). Projektin luonne on myös hyvä ottaa huomioon tiimiä koottaessa. Jossain projekteissa tarvitaan ehkä enemmän datan käsittelyyn liittyvää osaamista, kuin analyyttistä osaamista ja taas jossain projekteissa tilanne on päinvastoin (Davenport, 2014). Provostin ja Fawcettin (2013) mukaan datatieteilijä tiimissä olevilta henkilöitä pitäisi joka tapauksessa vaatia enemmän analyyttisiä taitoja kuin perinteisiä ohjelmistokehitykseen liitettyä kompetensseja.

## 3.4 Datatieteilijän kompetenssien viitekehys

Datatieteilijän kompetenssien viitekehys muodostuu Halvekan ja Merhoutin (2009) kompetenssikehyksessä ja sen eri kategorioihin lisätyistä datatieteilijän

kompetensseista. Taulukko 5 kuvaa kirjallisuuskatsauksen pohjalta luotua viitekehystä. Viitekehyksen sisältämät kompetenssi kategoriat ovat henkilökohtaiset ominaisuudet, ammattitaito, liiketoiminta osaaminen ja tekninen osaaminen. Jokainen eri kategoria käydään seuraavaksi läpi omassa alaluvussaan.

TAULUKKO 5 Datatieteilijän kompetenssit Havelkan ja Merhoutin (2009) kompetenssikehyksessä.

| Kategoria | Kompetenssit |
|---|---|
| **Henkilökohtaiset ominaisuudet** | Intohimo ongelmien ratkaisemiseen (Davenport & Patil, 2012),<br>Halu oppia uutta<br>(Bhambhri, 2012),<br>Päättäväisyys (Daveport, 2014),<br>Kykyä työskennellä itsenäisesti (Davenport, 2014),<br>Improvisointitaito (Davenport, 2014) |
| **Ammattitaito** | Kommunikointitaito (Davenport & Patil, 2012, Chen ym., 2012, Ohlhorst, 2013, Davenport, 2014),<br>Ongelmanratkaisukyky (Davenport & Patil, 2012),<br>Analyyttiset taidot (Davenport, 2014),<br>Ihmissuhdetaidot (Davenport, 2014) |
| **Liiketoiminta-osaaminen** | Toimialaosaaminen (Davenport & Patil, 2012, Bhambhri, 2012, Chen ym., 2012, Conway, 2011, Davenport, 2014),<br>Liiketoimintaosaaminen (Bhambhri, 2012, Davenport, 2014) |
| **Tekninen osaaminen** | Ohjelmointitaito (Davenport & Patil, 2012, Bhambhri, 2012, Benda, 2012, Ohlhorst, 2013, Davenport, 2014),<br>Algoritmit (Ohlhorst, 2013),<br>Tilastotiede (analytiikka) (Benda, 2012, Bhambhri, 2012, Ohlhorst, 2013, Conway, 2011),<br>Tietokannat (Benda, 2012),<br>Visualisointityökalut (Benda, 2012),<br>Kyky toteuttaa tekninen järjestelmä (Chen ym., 2012),<br>Hakkerointitaidot (Conway, 2011, Davenport 2014),<br>Hadoop/MapReduce (Conway, 2011, Davenport, 2014),<br>Koneoppiminen (Ohlhorst, 2013),<br>Tiedonlouhinta (Ohlhorst, 2013) |

### 3.4.1 Henkilökohtaiset ominaisuudet

Tämä kategoria kuvaa yksilön ominaisuuksia ja henkilökohtaisia piirteitä, jotka tekevät hänestä menestyvän datatieteilijän. Tähän kategoriaan on listattu ominaisuuksia joita ei voi suoraan opiskella tai saavuttaa koulutuksen kautta. Kuitenkin tällaisten ominaisuuksien avulla yksilö voi olla kiinnostunut alasta (Havelka & Merhout 2009).

Kirjallisuudessa datatieteilijältä vaaditut henkilökohtaiset ominaisuudet eivät juurikaan IT-ammattilaisen kompetensseista, jotka käyvät ilmi kompe-

tenssien kehyksien vertailussa. Datatieteilijältä odotetaan intohimoa uuden oppimiseen ja ongelmien ratkaisemiseen (Bhambhri, 2012, Davenport ja Patil, 2012). Tätä voi pitää itsestään selvyytenä, sillä massadata teknologiat ovat vielä suhteellisen uusia ja ne uudistuvat ja kehittyvät jatkuvasti. Teknologioiden uutuudesta ja analyyttisestä tehtävänkuvasta johtuen datatieteilijältä odotetaan myös improvisointitaitoa. Hyvin harvaan ongelmaan on olemassa valmista ratkaisua, jolloin vaaditaan luovuutta ongelman selvittämiseksi. Davenportin (2014) mukaan myös kyky työskennellä itsenäisesti on tärkeää datatieteilijälle. Voidaan olettaa että tämän taidon pitää korostua silloin, kun datatieteilijä työskentelee yksin, eikä tiimissä.

### 3.4.2 Ammattitaito

Ammattitaito kategorialla pyritään kuvaamaan niitä taitoja ja kyvykkyyksiä, joita odotetaan kaikilta alan ammattilaisilta. Tämän kategorian sisältämiä ominaisuuksia ja taitoja voi oppia koulutuksen kautta (Havelka & Merhout, 2009).

Datatieteilijältä odotetaan hyvää kommunikointitaitoa. Se nousi useissa lähteissä yhdeksi tärkeimmistä datatieteilijän kompetensseista (Davenport & Patil, 2012, Chen ym., 2012, Ohlhorst, 2013, Davenport, 2014). Ei riitä, että asian osaa selittää selkeästi, vaan se pitää esittää niin, että jokainen eri sidosryhmän jäsen ymmärtää mistä kyseissä asiassa on kyse. Kommunikointitaitoihin liittyy myös ihmissuhde- ja vuorovaikutustaidot. Datatieteilijän odotetaan olevan pärjäävän kaikenlaisten ihmisten seurassa. Kommunikointitaitojen lisäksi datatieteilijän tulee osata ratkaista analyyttisiä ongelmia.

### 3.4.3 Liiketoimintaosaaminen

Tähän kategoriaan kuuluvat taidot liittyvät liiketoimintaan. Liiketoimintaosaamiseen sisältyy esimerkiksi ymmärrys siitä, kuinka liiketoimintaa harjoitetaan, mitä liiketoimintaprosessit ovat jne. Näitä taitoja omaksutaan koulutuksen ja kokemuksen kautta (Havelka & Merhout, 2009).

Massadataa valjastettaessa eri toimialoille täytyy datatieteilijältä löytyä toimialakohtaista osaamista, jotta datan analysoinnista saataisiin paras mahdollinen tulos. Myös liiketoimintakäytänteiden ymmärryksestä on hyötyä. Dataammattilainen osaa hyödyntää työssään organisaatioin tarjoamia erilaisia resursseja. Yhdistämällä esimerkiksi ihmiset, teknologia ja muut organisaation sisällä olevat resurssit, on mahdollista saada parempia tuloksia harjoitetusta analytiikasta (Bhambhri, 2012). Davenport (2014) painottaa liiketoimintaosaamisen tärkeyttä.

### 3.4.4 Tekninen osaaminen

Teknisen osaamisen kategoria sisältää yksilön alakohtaisen osaamisen, tässä tapauksessa informaatioteknologia-alalle kuuluvan erityisosaamisen. Nämä

tiedot ja taidot voidaan omaksua koulutuksen kautta, mutta ne vahvistuvat myös kokemuksen myötä (Havelka & Merhout, 2009).

Tilastotieteellinen tietämys ja analyyttinen osaaminen mahdollistavat massadata – analytiikan harjoittamisen. Datatieteilijän tulee osata eri analytiikan tekniikoita, kuten koneoppiminen ja tiedonlouhinta (Ohlhorst, 2013).

Ohjelmointitaito on yksi yleisimmistä ja perustavan laatuisimmista datatieteilijän ominaisuuksista (Davenport & Patil, 2012, Bhambhri, 2012, Benda, 2012, Ohlhorst, 2013, Davenport, 2014). Benda Radek (2012) suosittelee korkeamman tason ohjelmointikieliä, kuten Java, Pyhton, Perl, PHP, C, C++, C# ja Objective C, datatieteilijän tarpeisiin. Ohjelmointikielen valinnassa tulisi huomioida kielien eri ominaisuuksia. Esimerkiksi sellaiset ohjelmointikielet jotka käsittelevät dataa objekteina ja sallivat useita erilaisia toimintoja, vähentävät huomattavasti datan käsittelyyn käytettyä aikaa. Kuten jokaisen ohjelmoijan, myös datatieteilijän tulisi osata käyttää komentorivikäyttöliittymää (Command Line Interface). Myös erilaisten ohjelmointirajapintojen (Application Programming Interface) kanssa toimimisen tulisi olla tuttua (Benda, 2012). Analysoinnin tuotteena syntyneistä havainnoista datatieteilijän tulee osata esittää tieto visuaalisesti; kuvioiden tulee olla selkeitä ja vakuuttavia (Davenport & Patil, 2012). Kaiken tämän lisäksi datatieteilijän tulee osata käyttää ja hyödyntää massadata-analytiikkaan kehitettyjä sovelluksia, kuten Hadoopia (Conway, 2011, Davenport, 2014).

# 4 EMPIIRINEN OSIO

Tässä luvussa kuvataan miten tutkimus on tehty ja esitetään tutkimuksen tulokset. Aluksi kuvataan tutkimusmenetelmät ja tutkittava aineisto. Sitten käydään läpi tutkittavasta aineistosta kerätyt kompetenssit. Lopuksi kompetenssit listataan kirjallisuudenpohjalta valittuun viitekehykseen.

## 4.1 Tutkimusmenetelmät

Datatieteilijältä vaadittavia kompetensseja on tutkittu hyvin vähän (Benda, 2012). Rastaksen ja Espin (2014) mukaan eri tahojen, kuten yritysten ja koulutusorganisaatioiden, tulisi tehdä yhteistyötä, jotta datatieteilijän tarvitsemat kompetenssit saataisiin määriteltyä. Tämän tutkimuksen tavoitteena on selvittää millaisia kompetensseja datatieteilijältä odotetaan. Tutkimus toteutetaan vertailemalla kirjallisuudessa esiintyneitä kompetensseja työpaikkailmoituksista kerättyihin kompetensseihin.

Tämä tutkimus toteutettiin hyödyntäen sisällönanalyysimenetelmää. Sisällönanalyysimenetelmän avulla on tarkoituksena tuottaa tutkittavasta ilmiöstä kuvaus yleisessä ja tiivistetyssä muodossa. Menetelmän avulla lähdeaineistoa pyritään analysoimaan systemaattisesti ja objektiivisesti. Se sopii hyvin myös strukturoimattoman aineiston analysointiin. Analysoinnin tuloksena ei synny johtopäätöksiä, vaan se jäsentää aineiston muotoon josta johtopäätöksiä on mahdollista tehdä (Tuomi ja Sarajärvi, 2009 s. 103). Sisällönanalysointi voi olla kvalitatiivista tai kvantitatiivista (Seitamaa-Hakkarainen, 2000, Tuomi ja Sarajärvi, 2009 s.105). Kvantitatiivista sisällönanalysointia voidaan kutsua sisällön erittelyksi. Siinä keskitytään tiettyjen sanojen ja ilmaisujen esiintymistiheyteen käsiteltävässä aineistossa. Kvalitatiivisessa sisällönanalyysissä ei olla kiinnostuneita niinkään sanojen esiintymistiheydestä, vaan niiden merkityksestä. Aineisto luokitellaan kategorioihin, jotka helpottavat aineiston hahmottamista. Tavoitteena on saada systemaattinen ja kattava kuvaus aineistosta (Seitamaa-Hakkarainen, 2000). Yhdistämällä sisällönanalyysi ja sisällön erittely, voidaan

puhua kontekstianalyysistä. Tällöin aineistoa tarkastellaan jostain kontekstista, jossa tutkittavat asiat esiintyvät (Tuomi ja Sarajärvi, 2009).

Edellisessä luvussa esiteltiin kirjallisuuden pohjalta luotu viitekehys datatieteilijän kompetensseista. Se asettaa osittain raamit aineistoanalyysille. Tällaista sisällönanalyysiä kutsutaan teoriaohjaavaksi analyysiksi. Se tarkoittaa että luotu viitekehys ohjaa ja auttaa analyysin tekoa. Aluksia aineisto pelkistetään ja ryhmitellään. Tämän jälkeen aineisto yhdistetään olemassa olevaan viitekehykseen (Tuomi ja Sarajärvi, 2009 s.96-104).

## 4.2 Aineistoanalyysi

Tutkimuksen aineistoksi valittiin työpaikkailmoitukset. Niistä käy hyvin ilmi, mitä odotuksia organisaatioilla on hakemastaan henkilöstä. Työpaikkailmoituksista työntekijältä vaaditut kompetenssit ovat mainittu strukturoimattomassa tai strukturoidussa muodossa. Tässä tutkimuksessa tutkitaan datatieteilijän kompetensseja, joten työpaikkailmoitussivustolta ilmoituksia haettaessa käytettiin hakusanaa ”data scientist”. Aluksi työpaikkailmoituksia haettiin myös hakusanalla ”big data”, mutta hakusana esiintyi lähes jokaisessa IT-alan työpaikkailmoituksessa, joten piti valita kuvaavampi hakusana. Suomen työpaikkailmoitussivustoilta ei löytynyt kuin muutama ilmoitus, joten tutkittava aineisto kerättiin 21.7.2014 Monser.com – työpaikkailmoitussivustolta. Aineisto koostuu 94:sta työpaikkailmoituksesta. Haku toteutettiin hakemalla tehtävänimikkeen perusteella. Rajauksena käytettiin ”data scientist” -hakusanaa.

Sisältöanalyysin ensimmäinen vaihe on aineiston pelkistäminen. Aineistoa lähdettiin käsittelemään työpaikkailmoitus kerrallaan. Jokaisesta ilmoituksesta listattiin vaatimukset kyseistä työpaikkaa kohden. Vaatimukset sisälsivät työntekijältä odotettavia kompetensseja ja muita vaatimuksia, kuten vaadittu kokemus tai koulutus. Aineiston pelkistämisellä luodaan pohja analyysin seuraavalle ryhmittely vaiheelle (Tuomi ja Sarajärvi, 2009 s.101).

Vaatimukset työnhakijalle esiintyivät työpaikkailmoituksissa joko listana tai tekstin seassa. Esimerkiksi tapauksessa #1 ”Meillä on töitä intohimoiselle ja luovalle datatieteilijälle tai tutkimusinsinöörille.” Kyseisestä lauseesta on havaittavissa kaksi kompetenssia intohimo ja luovuus. Tapauksessa 44# ”Vahva puhuttu ja kirjoitettu kommunikointitaito” on pelkistettynä kommunikointitaito. Tapauksessa #5 ”Vähintään kolmen vuoden kokemus” tarkoittaa pelkistettynä samaa, työntekijältä odotetaan kolmen vuoden kokemusta. Lisäksi tapauksessa #5 on listattu useita teknologioita ja tekniikoita, kuten Python, Java, Hadoop ja MongoDB, nämä tiedon kerättiin myös ylös. Tapauksessa #46 kompetenssit on jo lähtökohtaisesti jaoteltu kahteen eri kategoriaan: vastuut ja vaatimukset. Vastuut kategoriassa on mainittu ”Kommunikoi tulokset tiimille ja yhtiön muille sidosryhmille.” Edellisestä lauseesta on yksi kompetenssi kommunikointitaidot. Vaatimukset kategoriasta on helposti tunnistettavissa eri ohjelmointikielet, kuten Java, C# ja C++. Vastaavasti ”Oma-aloittein ja saa asioita tehdyksi” tarkoittaa kompetensseina oma-aloitteinen ja aikaansaava. Osassa

ilmoituksista olikin jo alustavaa ryhmittelyä työntekijältä odotetuista kompetensseista ja vaatimuksista. Aineiston pelkistys vaiheessa työpaikkailmoituksista koottiin taulukko ilmoituksissa esiintyneistä vaatimuksista. Pelkistysvaihe sisälsi myös alustavaa ryhmittelyä, sillä vaatimuksia yhdenmukaistettiin kuvaaviksi avainsanoiksi. Esimerkiksi ”communication skills”- ja ”verbal communication skills”-kompetensseja vastaa avainsanan ”kommunikointitaidot”.

Aineistoanalyysin toinen vaihe on ryhmittely. Siinä pelkistämisvaiheessa kerätty tieto ryhmitellään loogisiksi kokonaisuuksiksi. Tämä vaihe on analyysin kannalta erittäin kriittinen, sillä kategorisointi on tulkinnan varaista (Tuomi ja Sarajärvi, 2009 s.101). Työpaikkailmoitusten vaatimuksista oli helposti erotettavassa kaksi kategoriaa: vaadittu kokemus ja koulutus. Kokemus-kategoriaan listattiin vaatimukset työkokemuksesta ja koulutuskategoriaan tutkintoaste. Työpaikkailmoituksissa olleet tutkintoasteet olivat kandidaatin, maisterin ja tohtorin tutkinnot. Koulutus ja kokemus vaatimusten kategorisoinnin jälkeen jäljelle jääneen tiedon ryhmittely ei ollut itsestään selvää. Loput vaatimukset ryhmiteltiin joko kompetensseihin tai teknisiin, teknologisiin ja tuote vaatimuksiin. Edellisessä luvussa määriteltiin kompetenssi tarkemmin, mutta yksinkertaistettuna se koostuu henkilön tiedoista (eksplisiittinen/hiljainen), taidoista ja kyvyistä (Schmiedinger, Valentin & Stephan, 2005). Kompetenssi kategoriaan listattiin abstraktit käsitteet, jotka eivät viittaa johonkin tiettyyn tekniikkaan, teknologiaan tai tuotteeseen. Jäljellä olevaan pelkistettyyn aineistoon jäi tekniikat, teknologiat ja tuotteet. Näin ollen siitä syntyi ryhmittely vaiheen viimeinen kategoria.

Aineistoanalysoinnin ryhmittelyvaiheessa tunnistettiin neljä eri luokituskategoriaa pelkistetylle aineistolle. Tunnistetut luokituskategoriat olivat, kompetenssit, tekniset taidot, vaadittu koulutus ja vaadittu kokemus. Kompetenssit kategoriaan listattiin abstrakteimmat käsitteet, kuten esimerkiksi kommunikointitaito, ohjelmointitaito, liiketoimintaosaaminen jne. Tekniset taidot kategoria sisältää työpaikkailmoituksissa esiintyneet vaatimukset tuotteiden, teknologioiden ja tekniikoiden osalta. Vaadittu koulutus ja kokemus kategoriat sisältävät nimensä mukaiset tiedot. Aineiston ryhmittelyn jälkeen kategoriaan kuuluneet tiedot yhdistettiin ja niiden esiintymiskerrat laskettiin yhteen. Näin saatiin tietää kuinka monessa eri työpaikkailmoituksessa mikäkin tieto esiintyi.

Analyysin viimeisessä vaiheessa kompetenssit kategorian tiedot yhdistetään kirjallisuuskatsauksessa esiteltyyn kompetenssikehykseen. Havelkan ja Merhourtin (2009) kompetenssikehys sisältää neljä eri kategoriaa. Kompetenssit oli helppoa jakaa näin ryhmiin, koska kompetenssikehyksessä on kuvailtu selvästi minkä tyyliset tiedot kuuluvat mihinkin kategoriaan. Myös tekniset taidot kategorian tiedot olisi ollut mahdollista yhdistää kompetenssikehykseen, mutta esitysteknisestä syystä tyydyttiin vain toteamaan, että kaikki tekniset taidot kuuluvat kompetenssikehyksen teknisen osaamisen kategoriaan.

Aineistoanalyysin avulla työpaikkailmoituksista saatiin oleellinen tieto esille, sekä ryhmiteltyä se loogisiin kokonaisuuksiin. Lisäksi aineistossa esiintyneet abstraktit kompetenssit liitettiin kirjallisuuskatsauksen pohjalta valittuun

kompetenssikehykseen. Seuraavaksi esitellään analysoinnin tuotteena syntyneet luokituskategoriat jokainen omassa alaluvussaan.

## 4.3 Tulokset

Taulukko 6 sisältää kymmenen yleisintä aineistossa esiintynyttä kompetenssia. Aineistosta löytyi yhteensä 95 erilaista kompetenssia tähän kategoriaan. Tilastotieteellinen ja analyyttinen osaaminen, koneoppiminen, sekä kommunikointitaidot mainittiin useammin kuin joka toisessa hakemuksessa. Liiketoimintaosaamista ja ohjelmointitaitoja odotettiin hakijalta noin 40 prosentissa ilmoituksista.

TAULUKKO 6 Aineiston top 10 kompetenssia

| | **Kompetenssi** | **Esiintymiskerrat** |
|---|---|---|
| 1. | Tilastotiede | 59 |
| 2. | Analyyttiset taidot | 53 |
| 3. | Koneoppiminen | 50 |
| 4. | Kommunikointitaidot | 49 |
| 5. | Liiketoimintaosaaminen | 39 |
| 6. | Ohjelmointitaidot | 38 |
| 7. | Tiedonlouhinta | 33 |
| 8. | Algoritmit | 28 |
| 9. | Ennustava analyysi | 27 |
| 10. | Tiimityöskentelytaidot | 25 |

Useimmissa työpaikkailmoituksissa listattiin vaatimuksina spesifisiä teknologioita, ohjelmistoja ja ohjelmointikieliä. Suhteessa muihin vaadittuihin taitoihin, teknisiä taitoja edellytettiin huomattavan paljon. Ohjelmointikielien osaamista vaadittiin useammin kuin tietyn tuotteen tai teknologian tuntemista. Massadata alusta Hadoop esiintyy huomattavasti useammassa ilmoituksessa kuin perinteiset relaatiotietokannat. Hadoopin ollessa noin joka kolmannessa ilmoituksessa, jäävät perinteiset relaatiotietokannat, kuten MySql ja SQLServer kumpikin noin seitsemään prosenttiin. Taulukko 7 sisältää kymmenen yleisintä työpaikkailmoituksissa esiintynyttä vaatimusta teknisten taidot kategoriassa. Aineistossa esiintyi yhteensä 89 erilaista vaatimusta tähän kategoriaan.

TAULUKKO 7 Aineiston top 10 teknistä taitoa

| | **Ohjelmointikieli / teknologia** | **Esiintymiskerrat** |
|---|---|---|
| 1. | R | 46 |
| 2. | Python | 46 |
| 3. | SQL | 37 |
| 4. | Java | 37 |
| 5. | Hadoop | 35 |
| 6. | SAS | 24 |

| | | |
|---|---|---|
| 7. | Hive | 22 |
| 8. | C++ | 21 |
| 9. | Pig | 19 |
| 10. | Matlab | 18 |

Suurimmassa osassa työpaikkailmoituksia on mainittu koulutusvaatimus. Vähintään kandidaatin tai maisterin tutkintoa edellytetään joka kolmannessa ilmoituksessa. Tohtorin tutkintoa vaaditaan huomattavasti vähemmän. Reilu neljäsosa ilmoituksista ei sisältänyt koulutusvaatimusta. Ilmoitusten perusteella voidaan päätellä, että mitä korkeampi koulutus, sen helpompi on työllistyä. Työkokemusta työnhakijalta odotettiin 57 prosentissa työpaikkailmoituksista. Tulosten perusteella neljän vuoden kokemus riittää 75 prosenttiin ilmoituksista.

Taulukossa 8 kompetenssit on listattu Havelkan ja Merhoutin (2009) kompetenssikehykseen. Kompetenssikehyksen kategorian yhteyteen on listattu kategorian kompetenssien yhteenlasketut esiintymiskerrat. Lisäksi jokaisen yksittäisen kompetenssin perässä on kyseisen kompetenssit esiintymiskerrat aineistossa. Kompetenssikehykseen listatuista työpaikkailmoituksissa kompetensseista teknisen osaamisen kategoriaan tuli selvästi eniten osumia, yhteensä 354. Toisiksi eniten osumia tuli ammattitaitokategoriaan. Yllättävintä tuloksessa on se, että liiketoimintaosaamisen kategoriaan tuli yhteensä vain 49 osumaa. Täytyy kuitenkin huomioida se, että liiketoimintaosaaminen mainittiin yhteensä 39 eri ilmoituksessa, joka noin 41 prosenttia koko aineistosta.

TAULUKKO 8 Aineistossa esiintyneet kompetenssit Halvekan ja Merhoutin kompetenssikehyksessä

| **Kategoria (esiintymiskerrat yht.)** | **Kompetenssit** | **Esiintymiskerrat** |
|---|---|---|
| **Henkilökohtaiset ominaisuudet (72)** | Luovuus | 16 |
| | Paineensietokyky | 10 |
| | Venymiskyky (joustavuus) | 10 |
| | Kyky toimia muuttuvassa ympäristössä | 5 |
| | Intohimoinen | 5 |
| | Uteliaisuus | 4 |
| | Oma-aloitteisuus | 4 |
| | Innovatiivisuus | 4 |
| | Aikaansaava | 4 |
| | Kyky oppia (oppimiskyky) | 4 |
| | Kyky työskennellä itsenäisesti | 4 |
| | Matkustusvalmius | 2 |
| **Ammattitaito (163)** | Analyyttiset taidot | 53 |
| | Kommunikointitaidot | 49 |
| | Tiimityöskentelytaidot | 25 |
| | Ongelmanratkaisukyky | 25 |
| | Johtamistaidot | 9 |
| | Raportointi | 2 |
| **Liiketoimintaosaaminen (49)** | Liiketoimintaosaaminen | 39 |
| | Yrittäjämäisyys | 4 |

| | | |
|---|---|---|
| | Prosessiosaaminen | 4 |
| | Strateginen suunnittelu | 2 |
| **Tekninen osaaminen (354)** | Tilastotiede | 59 |
| | Koneoppiminen | 50 |
| | Ohjelmointitaidot | 38 |
| | Tiedonlouhinta | 33 |
| | Algoritmit | 28 |
| | Ennustava analyysi | 27 |
| | Dokumentointitaidot | 22 |
| | Datan visualisointi | 13 |
| | ETL | 11 |
| | Projektinhallinta | 9 |
| | Optimointi | 9 |
| | NLP | 8 |
| | Matematiikka | 8 |
| | Ekonometria | 7 |
| | Verkostoanalyysi | 4 |
| | Open source | 4 |
| | Massadata – järjestelmät | 4 |
| | Suosittelujärjestelmät | 3 |
| | Ennustava mallintaminen | 3 |
| | Todennäköisyyksien analysointi | 2 |
| | Tietovarastot | 2 |
| | Suunnittelu | 2 |
| | Simulointi | 2 |
| | Pilvilaskenta | 2 |
| | Mallien tunnistaminen | 2 |
| | Ketterät kehitysmenetelmät | 2 |

# 5 POHDINTA

Tässä luvussa vertaillaan kirjallisuudessa esiintyneitä kompetensseja empiirisesti kerättyihin kompetensseihin. Kompetenssit listataan yhteiseen viitekehykseen vertailun helpottamiseksi. Lisäksi pohditaan vastaako tieteellinen käsitys datatieteilijän kompetensseista organisaatioiden vaatimiin taitoihin.

## 5.1 Datatieteilijän kompetenssit – ero kirjallisuuden ja empiirisen aineiston välillä

Tämän tutkimuksen tarkoituksena oli selvittää millaisia kompetensseja datatieteilijällä tulisi olla. Datatieteilijän kompetenssit kerättiin kirjallisuudesta ja työpaikkailmoituksista. Työpaikkailmoituksista kerätyt kompetenssit on listattu kirjallisuuden pohjalta luotuun viitekehykseen. Taulukko 9 sisältää kirjallisuudessa esiintyneet ja työpaikkailmoituksissa vaaditut kompetenssit. Selkeyden vuoksi taulukosta on jätetty pois tuotteet ja teknologiat. Kirjallisuudessa esiintyneet kompetenssit on huomattavasti abstraktimpia kuin työpaikkailmoituksista kerätyt. Lisäksi ne olivat vaikeammin tunnistettavissa kuin työpaikkailmoituksissa olleet kompetenssit. Työpaikkailmoituksissa kompetenssit esiteltiin vaatimuksina, kun taas kirjallisuudessa ne esiintyvät sekavasti tekstin seassa. Niiden tunnistaminen oli suhteellisen vaikeaa. Seuraavaksi jokainen kompetenssikehyksen kategoria käsitellään omassa alaluvussaan. Niissä käsitellään tarkemmin empiirisesti havaittujen ja kirjallisuudesta poimittujen kompetenssien yhtäläisyyksiä ja eroavaisuuksia.

TAULUKKO 9 Datatieteilijän kompetenssit Havelkan ja Merhoutin kompetenssikehyksessä.

| **Kategoria** | **Kirjallisuudessa esiintyneet kompetenssit** | **Empiirisestä aineistosta kerätyt kompetenssit** |
|---|---|---|
| **Henkilökohtaiset ominaisuudet** | Intohimo ongelmien ratkaisemiseen (Davenport & Patil, 2012),<br>Halu oppia uutta (Bhambhri, 2012),<br>Päättäväisyys (Daveport, 2014),<br>Kykyä työskennellä itsenäisesti (Davenport, 2014),<br>Improvisointitaito (Davenport, 2014) | Luovuus,<br>Paineensietokyky,<br>Venymiskyky (joustavuus),<br>Kyky toimia muuttuvassa ympäristössä,<br>Intohiominen,<br>Uteliaisuus,<br>Oma-aloitteisuus,<br>Innovatiivisuus,<br>Aikaansaava,<br>Matkustusvalmius,<br>Kyky työskennellä itsenäisesti,<br>Kyky oppia (oppimiskyky) |
| **Ammattitaito** | Kommunikointitaito (Davenport & Patil, 2012, Chen ym., 2012, Ohlhorst, 2013, Davenport, 2014),<br>Ongelmanratkaisukyky (Davenport & Patil, 2012),<br>Analyyttiset taidot (Davenport, 2014),<br>Ihmissuhdetaidot (Davenport, 2014) | Analyyttiset taidot,<br>Kommunikointitaidot,<br>Tiimityöskentelytaidot,<br>Ongelmanratkaisukyky,<br>Johtamistaidot,<br>Raportointi |
| **Liiketoiminta-osaaminen** | Toimialaosaaminen (Davenport & Patil, 2012, Bhambhri, 2012, Chen ym., 2012, Conway, 2011, Davenport, 2014),<br>Liiketoimintaosaaminen (Bhambhri, 2012, Davenport, 2014) | Liiketoimintaosaaminen,<br>Yrittäjämäisyys,<br>Prosessiosaaminen,<br>Strateginen suunnittelu |
| **Tekninen osaaminen** | Ohjelmointitaito (Davenport & Patil, 2012, Bhambhri, 2012, Benda, 2012, Ohlhorst, 2013, Davenport, 2014),<br>Algoritmit (Ohlhorst, 2013),<br>Tilastotiede (analytiikka) (Benda, 2012, Bhambhri, 2012, Ohlhorst, 2013, Conway, 2011),<br>Tietokannat (Benda, 2012),<br>Visualisointityökalut (Benda, 2012),<br>Kyky toteuttaa tekninen järjestelmä (Chen ym., 2012),<br>Hakkerointitaidot (Conway, 2011, Davenport 2014),<br>Hadoop/MapReduce (Conway, 2011, Davenport, 2014), | Tilastotiede,<br>Koneoppiminen,<br>Ohjelmointitaidot,<br>Tiedonlouhinta,<br>Algoritmit,<br>Ennustava analyysi,<br>Dokumentointitaidot,<br>Datan visualisointi,<br>ETL,<br>Projektinhallinta,<br>Optimointi,<br>NPL,<br>Matematiikka,<br>Ekonometria,<br>Verkostoanalyysi,<br>Open source,<br>Massadatajärjestelmät, |

| | | |
|---|---|---|
| | Koneoppiminen (Ohlhorst, 2013),<br>Tiedonlouhinta (Ohlhorst, 2013) | Suosittelujärjestelmät,<br>Ennustava mallintaminen,<br>Todennäköisyyksien analysointi,<br>Tietovarastot,<br>Suunnittelu,<br>Simulointi,<br>Pilvilaskenta,<br>Mallien tunnistaminen,<br>Ketterät kehitysmenetelmät |

### 5.1.1 Henkilökohtaiset ominaisuudet

Tämä kategoria sisältää datatieteilijän henkilökohtaisia ominaisuuksia ja luonteenpiirteitä. Tähän kategoriaan on listattu ne kompetenssit, joita ei voi suoraan opiskella tai saavuttaa koulutuksen kautta. Kategorian kompetensseista luovuus esiintyi useinten työpaikkailmoituksissa. Kirjallisuudessa luovuus - kompetenssia vastaa Daveportin (2014) mainitsema improvisointitaito. Yleisesti kirjallisuudessa painotettiin, että datatieteilijältä odotetaan omaa halua ja intohimoa toimia tehtävässään. Tällöin hän on myös päättäväinen ja haluaa saada ongelmat ratkaistua. Lisäksi kyky työskennellä itsenäisesti on tärkeää. Kaikki kirjallisuudessa mainitut kompetenssit löytyivät myös työpaikkailmoituksista. Työpaikkailmoituksissa datatieteilijältä odotettiin myös paineensietokykyä, venymiskykyä (joustavuutta) ja kykyä toimia muuttuvassa ympäristössä, joita ei mainittu ollenkaan kirjallisuudessa.

### 5.1.2 Ammattitaito

Ammattitaito kuvaa niitä taitoja ja kyvykkyyksiä, joita odotetaan kaikilta alan ammattilaisilta. Tämän kategorian sisältämät kompetenssit voi oppia koulutuksen kautta. Datatieteilijälle selvästi tärkeimmät kompetenssit tässä kategoriassa ovat kommunikointitaidot, analyyttiset taidot sekä ongelmanratkaisukyky. Näiden lisäksi kirjallisuudessa mainittiin myös ihmissuhdetaidot. Työpaikkailmoituksissa ihmissuhdetaidot esitettiin tarkemmalla tasolla. Niistä puhuttiin tiimityöskentely- ja johtamistaitoina. Lisäksi työpaikkailmoituksissa mainittiin raportointitaito, jota ei löytynyt kirjallisuudesta.

### 5.1.3 Liiketoimintaosaaminen

Liiketoimintaosaamiseen sisältyy esimerkiksi ymmärrys siitä, kuinka liiketoimintaa harjoitetaan, mitä liiketoimintaprosessit ovat jne. Näitä taitoja omaksutaan koulutuksen ja kokemuksen kautta. Yksiselitteisesti datatieteilijältä odotetaan liiketoimintaosaamista. Lisäksi kirjallisuudessa painotettiin toimialaosaamista, mitä ei mainittu työpaikkailmoituksissa. Työpaikkailmoituksissa liike-

toimintaosaamisen lisäksi datatieteilijältä odotettiin yrittäjämäisyyttä, prosessiosaamista ja strategista suunnittelukykyä, joita ei mainittu kirjallisuudessa.

### 5.1.4 Tekninen osaaminen

Teknisen osaamisen -kategoria sisältää datatieteilijän erityisosaamisen. Nämä tiedot ja taidot voidaan omaksua koulutuksen kautta, mutta ne vahvistuvat myös kokemuksen myötä. Tähän kategoriaan tuli ylivoimaisesti eniten kompetensseja niin kirjallisuudesta kuin työpaikkailmoituksistakin. Datatieteilijän tärkeimmiksi kompetensseiksi tässä kategoriassa nousivat ohjelmointitaito, tilastotieteellinen osaaminen, algoritmit, koneoppiminen ja tiedonlouhinta. Kirjallisuudessa puhuttiin myös paljon hakkerointitaidoista. Sillä tarkoitettiin tietokoneen käytön syvällistä osaamista. Lisäksi datatieteilijältä odotetaan osaamista tietokannoista ja datan visualisoinnista.

Työpaikkailmoituksissa esiintyi huomattavasti enemmän eri kompetensseja kuin kirjallisuudessa. Verrattaessa työpaikkailmoitusten ja kirjallisuuden kompetensseja voidaan todeta, että kirjallisuudessa mainitut kompetenssit ovat paljon abstraktimmalla tasolla. Tällöin voidaan olettaa että kirjallisuudessa mainittu kompetenssi kattaa useamman työpaikkailmoituksessa olleen kompetenssin. Esimerkiksi tilastotieteellinen osaaminen pitää sisällään eri analysointitekniikat. Työpaikkailmoituksissa jokainen analysointitekniikka on mainittu erikseen, mutta kirjallisuudessa on tyydytty toteamaan vain yläkäsite. Tämän oletuksen pohjalta voidaan todeta, että kirjallisuuden ja empiirisen aineiston kompetenssit eroavat ainoastaan dokumentointi- ja projektinhallintataitojen osalta. Ne esiintyivät pelkästään työpaikkailmoituksissa.

## 5.2 Johtopäätökset

Tutkimuksen tarkoituksena oli selvittää millaisia kompetensseja datatieteilijältä vaaditaan ja minkälaisella viitekehyksellä niitä voidaan kuvata. Tutkimuksessa tarkasteltiin datatieteilijän kompetensseja niin tieteellisen kirjallisuuden kuin empiirisen aineiston työpaikkailmoitusten näkökulmista. Datatieteilijän kompetensseja on tutkittu hyvin vähän, eikä vastaavanlaista tutkimusta datatieteilijän kompetensseista ole olemassa.

Tulokset osoittavat, että Havelkan ja Merhoutin (2009) kehittämä kompetenssikehys sopii parhaiten datatieteilijän kompetenssien viitekehykseksi. Tutkimuksessa vertailtiin useita eri kompetenssikehyksiä. Viitekehyksen valinta suoritettiin vertailemalla eri kompetenssikehysten sisältämiä kategorioita. Valintakriteeriksi muodostui kehyksen laaja-alaisuus ja sen sisäisten kategorioiden abstraktisuus.

Tutkimustulosten perusteella voidaan todeta, että datatieteilijän kompetenssit ovat melko yhtenäiset kirjallisuudessa ja empiirisessä aineistossa. Jos tietyn kompetenssin esiintymiskertoja ei oteta huomioon, olivat kirjallisuuden ja työpaikkailmoitusten kompetenssit hyvin pitkälti linjassa toistensa kanssa. Yllättävintä tuloksissa oli se, että toimialaosaamista ei vaadittu kuin yhdessä työpaikkailmoituksessa, vaikka kirjallisen aineiston mukaan se olisi datatieteilijälle tärkeä kompetenssi. Davenport (2014) ja Conway (2011) painottivat toimialaosaamisen tärkeyttä. Sen avulla datatieteilijä pystyy ymmärtämään kyseisen alan liiketoimintaa ja ratkaisemaan siihen liittyviä ongelmia massadata-analytiikan avulla. Tämä ei kuitenkaan saanut tukea empiirisessä aineistossa. Tämä saattaa johtua siitä, että yritykset näkevät datatieteilijän enemmän teknisenä kuin liiketoimintaa kehittävänä roolina. Provost & Fawcett (2013) ja Davenport (2014) painottivat, että yhden henkilön tarvitse osata kaikkea. Massadata-analytiikan ympärille onkin rakennettu tiimejä, jotka ratkovat yhdessä heille osoitettuja tehtäviä. Tämän perusteella työpaikkailmoituksissa onkin voitu hakea massadata-analytiikkatiimin jäsentä, jonka ei tarvitse tuntea toimialaa, mutta tuo tiimiin jotain muuta osaamista. Toimialaosaaminen on myös mahdollista omaksua työn kautta, joten sitä ei välttämättä tarvitse lähtökohtaisesti osata.

Analyyttinen osaaminen oli empiirisessä aineistossa yksi kysytyimmistä kompetensseista. Tämä saa tukea Manyika ym. (2011) tutkimuksesta, jonka mukaan analyyttisiä taitoja omaavista datatieteilijöistä on lähitulevaisuudessa pulaa. Myös Gorgone ym. (2002) ja Davenport (2014) painottivat analyyttisten taitojen tärkeyttä. Analyyttiset taidot voidaan nähdä myös ongelmanratkaisukykynä (Provost ja Fawcett 2013 s.35), joka nousi kompetenssina esille myös työpaikkailmoituksissa. Myös tilastotieteellisen osaamisen tärkeyttä painotettiin niin kirjallisuudessa, kuin työpaikkailmoituksissa. Verrattuna analyyttisiin taitoihin, tilastotieteellinen osaaminen on enemmän teknistä. Tilastotieteellistä mallien ja analysointitekniikoiden ymmärtäminen on välttämätöntä datatieteilijän tehtävissä. Tutkimustulokset osoittavat, että datatieteilijän tulee osata ajatella analyyttisesti, sekä soveltaa tilastotieteellistä osaamista massadata-analytiikassa. Tilastotieteellinen osaaminen liittyy useisiin muihin datatieteilijältä vaadittaviin kompetensseihin kuten esimerkiksi koneoppimiseen ja tiedonlouhintaan. Tämä perusteella tilastotieteellinen osaaminen on datatieteilijälle yksi tärkeimmistä kompetensseista.

Tulosten perusteella kommunikointitaito on erittäin tärkeää datatieteilijälle. Datatieteilijän tulee pystyä esittämään analysointinsa tulokset eri sidosryhmille siten, että myös kohde yleisö ymmärtää mistä on kyse. Davenportin ja Patilin (2012) mukaan datatieteilijä toimii tulkkina liiketoimintajohdon ja teknisten asiantuntijoiden välillä. Niin kirjallisuudessa kuin empiirisessä aineistossa kommunikointitaito nousi useasti esille. Kommunikoinnin tärkeydestä kertoo myös Gartner:n tutkimus, jossa havaittiin että 70 – 80 prosenttia liiketoimintatiedon hallintaprojekteista epäonnistuu. Tutkimuksen mukaan se johtuu surkeasta kommunikoinnista IT- ja liiketoimintaosastojen välillä. Projekteissa ei ole osattu kysyä oikeita kysymyksiä tai ajatella liiketoiminnan oikeita tarpeita. (Goodwin, 2011). Liiketoimintatiedon hallinta on esiaste massadata-

analytiikalle, joten tutkimustulosten voidaan olettaa pätevän myös tässä kontekstissa. Jotta kommunikointi onnistuu IT- ja liiketoimintaosastojen välillä datatieteilijän tulee osata niin tekniset kuin liiketoiminnalliset asiat. Liiketoimintaosaaminen on myös tärkeä osa datatieteilijän työkalupakkia.

Myös ohjelmointitaito näyttäytyi tutkimustulosten perusteella oleellisena kompetenssina datatieteilijälle. Ohjelmointitaito kuvaa yleisellä tasolla datatieteilijältä vaadittavaa teknistä osaamista. Kirjallisuudessa ohjelmointitaidosta puhuttiin osana hakkerointitaitoja. Conway (2011) ja Davenport (2014) kuvasivat hakkerointitaitojen sisältävän kyvyn toteuttaa teknisesti massadata-analytiikkaa. Yhdessäkään työpaikkailmoituksista ei kuitenkaan käytetty sanaa hakkerointitaidot. Todennäköisesti tämä johtuu siitä, että hakkerointitermillä on negatiivinen kaiku ja se usein yhdistetään laittomaan toimintaan. Yleisesti työpaikkailmoituksissa ohjelmointitaidon yhteydessä vaadittiin jonkun tietyn ohjelmointikielen, teknologian tai tuotteen tuntemista. Kirjallisuudessa ohjelmointitaitoa ei määritelty näin tarkalla tasolla. Tutkimistulosten perusteella datatieteilijällä tulee olla vahva tekninen osaaminen, joka painottuu kykyyn toteuttaa massadata-analytiikkaa. Provostin ja Fawcettin (2013) mukaan datatieteilijälle analyyttisten taitojen tulee olla hallitsevia suhteessa ohjelmointitaitoihin. Tämä ei saa kuitenkaan saa tukea empiirisestä aineistosta, sillä työpaikkailmoituksissa ohjelmointitaitoja painotettiin enemmän. Empiirisessä aineistossa teknisiin taitoihin listattiin ohjelmointikielet, tuotteet ja teknologiat, joita ei listattu kompetenssikehykseen. Yleisesti ottaen suurin osa näistä teknisistä taidosta on rinnastettavissa ohjelmointitaitoon, joten empiirisen aineiston osalta ohjelmointitaito mainittiin jollain tapaa lähes jokaisessa työpaikkailmoituksessa. Tämän perusteella vaikuttaisi siltä, että organisaatioissa datatieteilijän rooli nähdään enemmän tietoteknisenä.

Tutkimustulosten perusteella datatieteilijän tärkeimmät kompetenssit ovat tilastotieteellinen ja liiketoiminnallinen osaaminen, sekä analyyttiset taidot, ohjelmointi- ja kommunikointitaidot, koneoppiminen ja tiedonlouhinta. Ne antavat datatieteilijälle mahdollisuuden selvitä työtehtävistään. Datatieteilijän tärkeimmät kompetenssit esiintyivät toistuvasti niin kirjallisuudessa kuin empiirisessä aineistossa. Henkilökohtaisten ominaisuuksien näkökulmasta datatieteilijältä odotetaan paljoa ja intohimoa työtehtäviään kohtaan. Tämä tarkoittaa halua ratkaista liiketoiminnan ongelmia massadata-analytiikan avulla. Yhden henkilön on vaikea omaksua kaikkia datatieteilijältä vaadittavia kompetensseja. On varmasti mahdollista, että yksi henkilö omaisi kaikki datatieteilijältä vaaditut kompetenssit, mutta se on todennäköisesti todella harvinaista. Oletettavasti tämä johtaa siihen, että massadata-analytiikan ympärille kerätään tiimejä, jotka pitävät sisällään datatieteilijän tarvitsemat kompetenssit. Tällöin yhden henkilön ei tarvitse osata kaikkea.

Organisaatioiden näkökulmasta tutkimustulokset lisäävät ymmärrystä siitä, mitä datatieteilijän tulisi osata suoriutuakseen työtehtävistään. Kirjallisuuden ja empiirisen aineiston suurin ero oli toimialaosaaminen, jota ei vaadittu lainkaan työpaikkailmoituksissa. Tämä johtuu mahdollisesti siitä, että organisaatiot odottavat datatieteilijän omaksuvan toimialan työn kautta. Voi myös

olla, että organisaatiot eivät osaa vaati datatieteilijältä toimialaosaamista. Tutkimustulosten perusteella voidaan myös olettaa, että massadata-analytiikan ympärille on helpompi rakentaa tiimi, joka sisältää kaikki datatieteilijältä vaaditut kompetenssit, kun palkata yksi henkilö joka osaa kaiken. Pienemmillä organisaatioilla ei välttämättä ole mahdollista muodostaa tiimejä, joten heidän tulee olla erityisen huolellisia palkatessaan datatieteilijää. Tämän tutkimuksen tuloksia voidaan myös hyödyntää suunniteltaessa datatieteilijöiden koulutusta. Koulutuksessa on tärkeää keskittyä oikeiden ja tärkeiden asioiden opettamiseen.

## 5.3 Tutkimuksen yleistettävyys ja luotettavuus

Tämän tutkimuksen tarkoitus oli tuottaa uutta tietoa datatieteilijän kompetensseista. Tutkimuksen tulokset ovat yleistettävissä siihen kontekstiin, jossa tutkimus on toteutettu, eli yhden työpaikkailmoitussivuston tuottamiin työpaikkailmoituksiin. Kirjallisuuskatsauksen teossa hyödynnettiin alan kirjallisuutta ja useita eri hakukoneita.

Tutkimuksen luotettavuutta on pyritty parantamaan kuvaamalla tutkimuksen eri vaiheet mahdollisimman läpinäkyvästi. Tutkimuksen tuloksia on pyritty havainnollistamaan selkeiden ja kuvaavien taulukoiden avulla. Sisällönanalyysimenetelmän avulla tutkimuksessa on pyritty saamaan oleellinen tieto esille ilman, että merkityksellistä tietoa katoaa. Menetelmää on kuitenkin kritisoitu siitä, että järjestetty aineisto esitetään tuloksina ilman todellisia johtopäätöksiä (Tuomi ja Sarajärvi, 2009 s.103). Tutkimus on ollut pitkä prosessi ja tutkimuksen tulokset ovat muovautuneet uudelleen työstämisen kautta. Tutkittava aihealue on hyvin uusi tutkimuskentällä, joten luotettavien lähteiden löytäminen on ollut haastavaa. Kuitenkin tutkimusprosessin aikana on löytynyt useita uusia lähteitä, joita on voitu hyödyntää. Empiirinen aineisto sisälsi yhteensä lähes 100 eri työpaikkailmoitusta.

Tutkimuksen luotettavuutta olisi voitu parantaa etsimällä työpaikkailmoituksia useammasta lähteestä, sekä käyttämällä useampia hakusanoja. Esimerkiksi kaikki organisaatiot eivät välttämättä käytä massadata-analyytikosta termiä datatieteilijä, joten kerätty aineisto voi olla todellisuutta kapea-alaisempi. Useampien hakusanojen käyttö ei ollut kuitenkaan mahdollista tämän tutkimuksen puitteissa. Lisäksi tutkimuksen eri vaiheet olisi voitu kuvata tarkemmalla tasolla, erityisesti kirjallisuus katsauksen osalta. Kirjallisuuden pohjalta luotu viitekehys antoi raamit työpaikkailmoituksista kerätyille kompetensseille. Se osiltaan määritteli miten vaatimukset pelkistettiin ja ryhmiteltiin. Tämä saattoi johtaa siihen, että osa vaatimuksista ei sopinut viitekehykseen ja jäi näin ollen tutkimuksen ulkopuolelle. Massadata ja massadata-analytiikka ovat käsitteinä suhteellisen uusia ja niille ei ole muodostunut vakiintuneita määrityksiä. Tästä syystä eri lähteet kuvaavat käsitteet eri tavalla, mikä heikentää tutkimustulosten luotettavuutta.

# 6 YHTEENVETO

Tässä tutkimuksessa tutkittiin datatieteilijältä vaadittuja kompetensseja. Tutkimus tehtiin vertailemalla kirjallisuudessa ja työpaikkailmoituksissa esiintyneitä kompetensseja. Tieteellisen kirjallisuuden pohjalta määriteltiin yleisesti massadata ja massadata-analytiikka, sekä valittiin tutkimuksessa käytetty viitekehys. Viitekehyksen valinta tehtiin vertailemalla eri kompetenssikehyksiä. Tutkimuksessa koottiin tieteellisessä kirjallisuudessa esiintyneet datatieteilijän kompetenssit. Empiirinen aineisto koostui 94 työpaikkailmoituksesta, joista eriteltiin datatieteilijältä vaaditut kompetenssit. Tieteellisestä kirjallisuudesta ja empiirisestä aineistosta kerättyjä kompetensseja vertailtiin yhtenäisen viitekehyksen avulla.

Tutkimus suoritettiin, koska datatieteilijältä vaadittavia kompetensseja on tutkittu hyvin vähän (Benda, 2012). Lisäksi kysyntä datatieteilijöistä on kovaa, pelkästään Yhdysvalloissa odotetaan vuonna 2018 olevan pulaa 140000 – 190000 analyyttisiä taitoja omaavasta työntekijästä (Manyika ym., 2011). Myös Suomessa datatieteilijöiden kysynnän odotetaan kasvavan. Rastaksen ja Espin (2014) ”Big datan hyödyntäminen” –raportissa toivotaan koulutusorganisaatioiden ja yritysten yhteistyötä, jotta datatieteilijän kompetenssit saataisiin määriteltyä.

Massadata ja massadata-analytiikka ovat hyvin moniulotteiset termit. Massadatalle ei ole olemassa yksiselitteistä määritelmää ja onkin jo havaittavissa, että termi tullaan osittain korvaamaan kuvailevimmilla termeillä (Davenport 2014). Yleisimmän määritelmän mukaan massadatalla on kolme eri ulottuvuutta: määrä, monimuotoisuus ja nopeus. Massadata-analytiikalla tarkoitetaan massadataan tehtyä analytiikkaa. Analysointitekniikat ovat pitkälti samat kuin muussakin analytiikassa, mutta siinä ei analysoida jotain tiettyä otosta, vaan aina lähdedataa kokonaisuudessaan.

Havelkan ja Merhoutin (2009) kompetenssikehys soveltui parhaiten tutkimuksen viitekehykseksi. Viitekehyksen valinta tehtiin vertailemalla useita eri kompetenssikehyksiä. Havelkan ja Merhoutin kompetenssikehyksen sisältämät kategoriat olivat selvästi laaja-alaisimmat ja abstrakteimmat. Tämän perusteella se sopi parhaiten datatieteilijän kompetenssivertailun viitekehykseksi.

Datatieteilijältä vaaditut kompetenssit olivat pitkälti yhtenäisiä tieteellisessä kirjallisuudessa ja työpaikkailmoituksissa. Yllättävin ero koski toimialaosaamisesta. Kirjallisuudessa painotettiin toimialaosaamisen tärkeyttä, mutta sitä ei kuitenkaan vaadittu kuin yhdessä työpaikkailmoituksessa. Työpaikkailmoituksissa painotettiin enemmän teknistä osaamista kuin kirjallisuudessa. Tutkimustulosten perusteella datatieteilijän tärkeimmät kompetenssit olivat tilastotieteellinen ja liiketoiminnallinen osaaminen, sekä analyyttiset taidot, ohjelmointi- ja kommunikointitaidot, koneoppiminen ja tiedonlouhinta. Lisäksi datatieteilijältä odotettiin intohimoa ja kykyä ratkaista liiketoiminnan ongelmia massadata-analytiikan avulla.

## 6.1 Jatkotutkimusaiheet

Tämä tutkimus ei ota kantaa datatieteilijöiden osaamisprofiileihin, joka nousee selvästi esille yhtenä jatkotutkimusaiheena. Osaamisprofiili tarkoittaa yhdistelmää tietyistä kompetensseista. Organisaatiot käyttävät erilaisia tuotteita ja teknologioita massadata-analytiikkaan, joka määrittelee kussakin organisaatiossa datatieteilijältä vaaditun teknisen osaamisen. On myös mahdollista, että organisaatio on perustanut massadata-analytiikan ympärille tiimin, johon kuuluu yksi tai useampi datatieteilijä. Tällöin tiimi sisältää useita eri osaamisprofiileja ja tiimin jäseniltä vaaditaan erilaisia kompetensseja. Olisi tärkeää tutkia millaisia osaamisprofiileja tiimiin kuuluvilla datatieteilijöillä tulisi olla.

Tieteellisen kirjallisuuden perusteella toimialaosaaminen oli yksi tärkeimmistä kompetensseista datatieteilijälle, mutta sitä ei vaadittu kuin yhdessä työpaikkailmoituksessa. Ovatko organisaatiot havainneet, että toimiala opitaan tuntemaan työn kautta, eikä sitä näin ollen pidetä ennakkovaatimuksena tulevalle työntekijälle? Voi myös olla, että toimialaosaamista ei tarvita datatieteilijän tehtävissä tai organisaatiot eivät osaa vaatia sitä tulevilta työntekijöiltä. Olisi tärkeää tutkia mistä tämä ero johtuu.

Tämän tutkimuksen empiirinen aineisto on vain yhdestä lähteestä. Tätä tutkimusta voisi laajentaa niin, että empiirinen aineisto on kerätty useammasta lähteestä. Lisäksi voitaisiin tutkia miten datatieteilijöiden koulutusohjelmat vastaavat organisaatioiden vaatimiin kompetensseihin. Tämä tutkimus olisi tärkeää, jotta tuleville datatieteilijöille osattaisiin opettaa oikeita asioita.

# LÄHTEET


Ashenhurst, R. L. (1972). Curriculum recommendations for graduate professional programs in information systems. *Communications of the ACM, 15(5),* 363-398.

Bailey, J. & Mitchell, R. B. (2006). Industry perceptions of the competencies needed by computer programmers: Technical, business, and soft skills. *The Journal of Computer Information Systems, 47(2),* 28-34.

Bakshi, K. (2012). Considerations for Big Data: Architecture and Approach, 2012 *IEEE Aerospace Conference, March 3-10, 2012.*

Bassellier, G. & Benbasat, I. (2004). Business Competence of Information Technology Progessinals: Conceptual Development and Influence on IT-Business Partnerships. *MIS Quarterly, 28(4),* 673-694.

Benda, R. (2012). Science of Big Data: Background and Requirements, *Advances in Economics, Risk Management, Political and Law Science, November 2012, pp* 311-316.

Bhambhri, A. (2012). Six tips for students interested in big data analytics. *XRDS: Crossroads, The ACM Magazine for Students, 19(1),* 9-9.

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.

Chen, H., Chiang, R.H.L. & Storey V.C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly, 36(4),* 1165-1188.

Conway, D. (2011). Data Science in the U.S. Intelligence Community, *IQT Quarterly, 2(4),* 24-27.

Davenport, T.H, (2015) Three big benefits of big data analytics, http://www.sas.com/tr_tr/news/sascom/2014q3/Big-data-davenport.html (Noudettu 22.3.2015).

Davenport, T. H., Barth, P. & Bean, R. (2012). How 'Big Data'is Different. *MIT Sloan Management Review.*

Davenport, T. H. & Patil, D.J. (2012) Data Scientist: The Sexiest Job of the 21st Century, *Harvard business review, Octeber 2012*

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM, 51(1),* 107-113.

Fischer, M. J., Su, X., & Yin, Y. (2010) Assigning tasks for efficiency in Hadoop. *In Proceedings of the 22nd ACM symposium on Parallelism in algorithms and architectures, ACM,* 30-39.

Franks, B. (2012). *Taming the big data tidal wave: Finding opportunities in huge data streams with advanced analytics (Vol. 56).* John Wiley & Sons.

Gantz, J. & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the Future, 2007,* 1-16.

Gartner. (2013). Big Data. http://www.gartner.com/it-glossary/big-data/ (Noudettu 12.2.2013).

Golfarelli, M., Rizzi, S., & Cella, I. (2004) Beyond data warehousing: what's next in business intelligence?. *In Proceedings of the 7th ACM international workshop on Data warehousing and OLAP, ACM,* 1-6.

Goodwin, B. (2010). Poor Communication to blame for business intelligence failure, says Gartner, http://www.computerweekly.com/news/1280094776/Poor-communication-to-blame-for-business-intelligence-failure-says-Gartner (Noudettu 1.4.2015).

Gorgone, J. T., Davis, G. B., Valacich, J. S., Topi, H., Feinstein, D. L. & Longenecker Jr, H. E. (2002). Model curriculum and guidelines for undergraduate degree programs in information systems. *Association for Computing Machinery (ACM), Association for Information Systems (AIS), Association of Information Technology Professionals (AITP).*

Havelka, D. & Merhout, J.W. (2009). Toward a Theory of Iformation Technology Professional Competence. *The Journal of Computer Informaton Systems, 50(2),* 106-116.

IBM. (2013). What is big data?, http://www-01.ibm.com/software/data/bigdata (Noudettu 20.2.2013).

IDG Enterprice. (2014). Big Data Resource Report, http://www.idgenterprise.com/report/big-data-2 (Noudettu 15.2.2014).

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

Lee, D.M.S., Trauth E.M. & Farwell, D. (1995). Critical Skills and Knowledge Requirements of IS Professionals: A Joint Academic/Industry Investigation, *MIS Quarterly, 19(3),* 313 -340.

Lee, S.M. & Lee, C.K. (2006). IT Managers' Requisite Skills: Matching job seekers' qualifications with employers' skill requirements. *Communications of The ACM, 49(4),* 111-114.

Lu, H-K., Lo, C-H. & Lin P-C. (2011). Competence Analysis of IT Professionals Involved in Business Services – Using a Qualitative Method, Software Engineering Education and Training *(CSEE&T), 2011 24th IEEE-CS Conference on. IEEE 2011,* 61-70.

Mansfield, R. S. (1996). Building competency models: Approaches for HR professionals. *Human Resource Management, 35(1),* 7-18.

McAfee, A. & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harward Business Review, October 2012.*

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

Minelli, M., Chambers, M. & Dhiraj, A. (2013) *Big Data Big Analytics: emerging business intelligence and analytic trends for today's businesses.* New Jersey: John Wiley & Sons.

Negash, S. (2004) Business intelligence. *Communications of the Association for Information Systems, 13(1),* 177-195.

Ohlhorst, F.J. (2013). *Big Data Analytics: turning big data into big money*. New Jersey: John Wiley & Sons.

Provost, F. & Fawcett, T. (2013). Data Science and Its Relationship to Big Data and Data-Driven Decision Making, *Big Data, 1(1),* 51-58.

Rastas, T. & Asp, E. (2014) Big datan hyödyntäminen, *Liikenne- ja viestintäministeriön julkaisuja, 20/2014.*

Rowe, G. & Wright, G. (1999). The Delphi technique as a forecasting tool: issues and analysis, *International Journal of Forecasting 15(1999),* 353-375.

Russom, P. (2011). Big Data Analytics. *TDWI Best Practices Report, 4th quarter 2011.*

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

SAS. (2013). Big Data Meets Big Data Analytics, http://www.sas.com/resources/whitepaper/wp_46345.pdf (Noudettu 1.4.2013).

Schmiedinger, B. Valentin, K. & Stephan, E. (2005). Competence Based Business Development – Organizational Competencies as Basis for Successful Companies , *Proceeding of I-KNOW '05, Graz, Austria, June 29 – Juli 1, 2005.*

Seitamaa-Hakkarainen, P. (2000), Kvalitatiinen sisällön analyysi, https://www.academia.edu/589363/Kvalitatiivinen_sis%C3%A4ll%C3%B6n_analyysi (Noudettu 25.4.2015).

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

Taylor, R. C. (2010). An overview of the Hadoop/MapReduce/Hbase framework and its current applications in bioinformatics. *BMC bioinformatics, 11(12), S1.*

Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Zhang, N., Antony, S., Liu, H. & Murthy, R. (2010). Hive-a petabyte scale data warehouse using hadoop. *Data Engineering (ICDE), 2010 IEEE 26th International Conference,* 996-1005.

Todd, P.A., McKeen, J.D., and Gallupe, R.B. (1995) The evolution of IS job skills: A content analy-sis of IS job advertisements from 1970 to 1990. *MIS Quarterly, 19 (1),* 1–24.

Tuomi, J. & Sarajärvi, A (2009). *Laadullinen tutkimus ja sisällönanalyysi*, Helsinki: Tammi

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29-39.

Yang, H. C., Dasdan, A., Hsiao, R. L., & Parker, D. S. (2007). Map-Reduce-Merge: Simplified Relational Data Processing on Large Clusters, *Proceedings of the 2007 ACM SIGMOD international conference on Management of data. ACM, 2007,* 1029-1040.

Zikopoulos, P., Eaton, C., Deroos, D., Deutsch, T. & Lapis, G. (2012). *Unserstanding Big Data: Analytics for Enterprice Class Hadoop and Streaming Data.* USA: McGraw-Hill.