SYNTEETTISEN DATAN HYÖDYNTÄMINEN TEOLLISUUSYRITYKSISSÄ

Synthetic data utilization in manufacturing companies

Lappeenrannan–Lahden teknillinen yliopisto LUT

Tuotantotalouden kandidaatintyö

2024

Valtteri Anttonen

Tarkastaja: Dosentti Jouni Koivuniemi

TIIVISTELMÄ

Lappeenrannan–Lahden teknillinen yliopisto LUT

LUT Teknis-luonnontieteellinen

Tuotantotalous

Valtteri Anttonen

Synteettisen datan hyödyntäminen teollisuusyrityksissä

Tuotantotalouden kandidaatintyö

2024

42 sivua, 12 kuvaa ja 1 taulukko

Tarkastaja: Dosentti Jouni Koivuniemi

Avainsanat: Synteettinen data, datapohjainen arvontuotanto, CRISP-DM, teollinen massadata, koneoppiminen

Key words: Synthetic data, data-based value creation, CRISP-DM, industrial big data, machine learning

Tämän kandidaatintyön tavoitteena on selvittää kuinka teollisuusyritykset pystyvät hyödyntämään synteettistä dataa liiketoiminnassaan. Työssä tarkastellaan datapohjaiseen liiketoimintaan liittyviä haasteita datan keräämisen ja analysoinnin kautta, sekä tutkitaan mitä arvoa teollisuusyritys voi saada datan hyödyntämisestä ja mitä ongelmia tähän liittyy. Työ on toteutettu integroivana kirjallisuuskatsauksena.

Teollisuusyritysten toiminnassa voidaan havaita monia dataan liittyviä haasteita, joihin synteettinen data voi tarjota ratkaisuja. Synteettisellä datalla tarkoitetaan keinotekoisesti, esimerkiksi tekoälymallilla luotua dataa ja se on tällä hetkellä murrosvaiheessa diffuusionsa suhteen. Voidaan olettaa, että monet teollisuusyritykset tulevat hyödyntämään synteettisen datan tarjoamia mahdollisuuksia lähivuosien aikana. Synteettiseen dataan liittyy paljon mahdollisuuksia, mutta myös riskejä ja rajoitteita liittyen sen aitouteen ja puolueellisuuteen.

Työn johtopäätöksenä voidaan sanoa, että teollisuusyrityksillä on monia mahdollisuuksia hyödyntää synteettistä dataa ja tällä hetkellä kirjallisuudessa suosituimmat tapaukset liittyvät koneoppimismallien kouluttamiseen. Synteettinen data ei kuitenkaan tarjoa ratkaisua suoriltaan kaikkiin ongelmiin. Yritysten tulee tunnistaa omien datastrategioidensa heikkoudet ja integroida synteettisen datan hyödyntäminen osaksi yrityksen toimintoja, esimerkiksi osaksi datalouhinnan prosessia.

Sisällysluettelo

Tiivistelmä

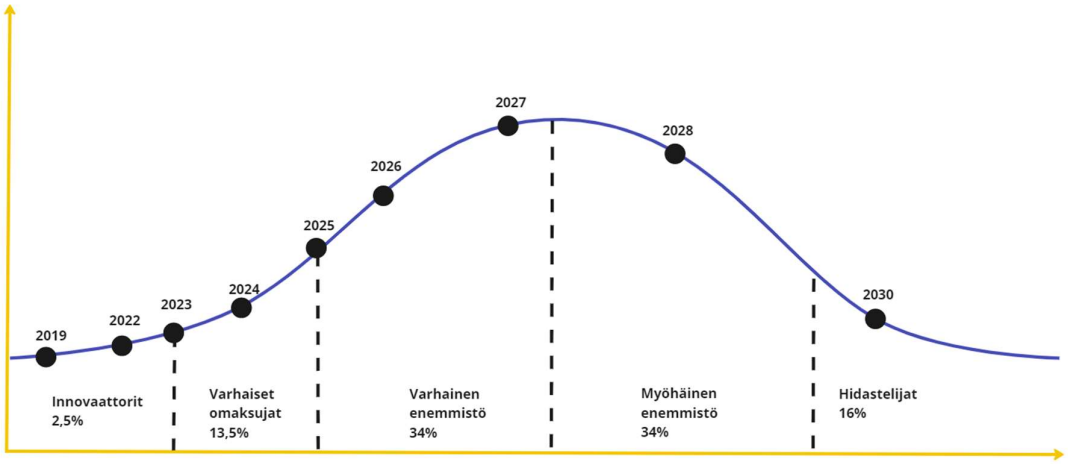
1 Johdanto ... 3
1.1 Työn tavoite ja tutkimuskysymykset ... 4
1.2 Tutkimusmenetelmät ja rajaukset ... 5
1.3 Tutkimuksen rakenne ... 6
2 Datapohjainen liiketoiminta ... 7
2.1 Teollinen massadata ... 7
2.2 Datapohjainen arvontuotanto ... 11
2.3 Teolliset älykkäät tuotepalvelujärjestelmät ... 15
3 Synteettinen data ... 19
3.1 Synteettisen datan määritelmä ... 20
3.2 Synteettisen datan riskit ja rajoitteet ... 23
3.3 Synteettisen datan tuottaminen ... 24
4 Synteettisen datan sovellutukset teollisuusyrityksissä ... 29
4.1 Synteettinen data käytännössä ... 29
4.2 Synteettinen data mahdollistajana ... 31
5 Johtopäätökset ... 33
5.1 Vastaukset tutkimuskysymyksiin ... 33
5.2 Tulevaisuudennäkymät ja jatkotutkimusaiheet ... 37
Lähteet ... 38

1 Johdanto

Yritykset pyrkivät jatkuvasti lisäämään datan hyödyntämistä jokapäiväisessä liiketoiminnassaan. Neljännen teollisen vallankumouksen myötä teollisuusyritykset keräävät ja hyödyntävät dataa yhä enemmän ja suuria kysymyksiä ovat kuinka datan keräämisestä saadaan tehokkaampaa, sekä kuinka tätä dataa pystytään hyödyntämään liiketoiminta-arvon tuotannossa enemmän. Hercik & Svoboba (2023) mukaan yksi neljännen teollisen vallankumouksen kulmakivistä on datan kerääminen koneista ja tuotantoprosesseista. Datan merkityksellisyyttä korostaa myös nykyajan vaatimukset monilla aloilla siirtyä kohti palvelupohjaista liiketoimintaa. Carillo (2017) mukaan data ohjaa liiketoimintaa ja on muuttunut uudeksi digitaaliseksi öljyksi, joka virtaa organisaatioiden suonissa. Perinteiset teollisuusyritykset eivät tuota pelkästään koneita, vaan pyrkivät tarjoamaan erinäköisiä palveluita pelkkien tuotteiden rinnalla, kuten huolto- ja kunnossapidon palvelut.

Datapohjaisessa liiketoiminnassa on kuitenkin omat haasteensa. Datan kerääminen nykyisistä teollisista prosesseista tuottaa haasteita liittyen esimerkiksi datan oikeellisuuteen ja laatuun (Hercik & Svoboba 2023), ja yrityksen toiminnasta kerätyn datan valmistelu analysointia varten voi viedä jopa 80 % data-analyytikon ajasta (Fleckenstein & Fellows 2018, 15). Myös datan omistajuuteen ja yksityisyydensuojaan liittyy paljon haasteita, sekä datan riittävyyteen esimerkiksi tekoälyn kouluttamiseen. Näihin ongelmiin voisi synteettinen data tarjota ratkaisuja. CloudTDMS:s (2024) tekemän ennusteen mukaan synteettinen data tulee olemaan yleisesti hyväksytty ja käytetty teknologia yrityksissä vuoteen 2030 mennessä siten, että ”varhainen enemmistö” tulee integroimaan synteettisen datan käytön omiin prosesseihin vuosien 2025–2026 aikana (kuva 1). Grand View Research:n (2024) tekemän markkinaennusteen mukaan synteettisen datan tuottamisen maailmanlaajuiset markkinat kasvavat vuoden 2022 164 miljoonasta Yhdysvaltain dollarista 1,8 miljardiin Yhdysvaltain dollariin vuoteen 2030 mennessä, vuotuisella 35 % kasvuvauhdilla. Puolestaan Allied Market Research (2024) ennustaa synteettisen datan tuottamisen maailmanlaajuisten markkinoiden kasvua 3,5 miljardiin Yhdysvaltain dollariin 2031 mennessä. Synteettinen data on parhaillaan

murrosvaiheessaan sen omaksumisen suhteen, mikä tekee siitä hyvin mielenkiintoisen aiheen tutkia.



Kuva 1: Synteettisen datan diffuusio (mukaillen CloudTDMS 2024)

1.1 Työn tavoite ja tutkimuskysymykset

Tämän kandidaatintyön tavoitteena on selvittää miten teollisuusyritykset pystyvät hyödyntämään synteettistä dataa liiketoiminnassaan ja sen kehittämisessä. Työssä tarkastellaan myös datapohjaisen liiketoiminnan mahdollisuuksia ja ongelmia, joihin synteettinen data voi tarjota ratkaisuja. Työssä vastataan yhteen päätutkimuskysymykseen kolmen osakysymyksen avulla. Päätutkimuskysymys työssä on:

- *Kuinka teollisuusyritys voi hyödyntää synteettistä dataa liiketoiminnassaan?*

Päätutkimuskysymyksen lisäksi työssä vastataan seuraaviin osatutkimuskysymyksiin:

- *Kuinka data muuttaa teollisuusyritysten liiketoimintaa ja mitä haasteita tähän liittyy?*
- *Mitä on synteettinen data sekä sen hyödyt ja rajoitteet?*
- *Kuinka teollisuusyritys pystyy luomaan arvoa synteettisen datan avulla?*

Tämän työn pohjalta lukija saa kattavan yleiskuvan mitä tarkoitetaan synteettisellä datalla. Lukija saa yleiskuvan synteettisen datan mahdollisuuksista teollisuusyrityksessä, sekä mitä riskejä ja rajoitteita synteettiseen dataan liittyy. Ensimmäisellä osatutkimuskysymyksellä haetaan vastausta siihen, kuinka teollisuusyritykset pystyvät hyödyntämään suuria määriä teollista massadataa ja mitä uusia liiketoimintamahdollisuuksia suurten datamäärien hyödyntäminen tuottaa. Toisen osatutkimuskysymyksen avulla selitetään mitä synteettisellä datalla tarkoitetaan, mitä mahdollisia hyötyjä sekä rajoitteita siihen liittyy. Kolmas osatutkimuskysymys vastaa siihen, mihin teollisuusyritys pystyy käyttämään synteettistä dataa liiketoiminnassaan ja mitä mahdollisia hyötyjä se pystyy saavuttamaan. Vastaamalla näihin osatutkimuskysymyksiin saadaan kattava kuva kokonaisuudesta siitä, miten suuret teolliset datamäärät muuttavat teollisuusyritysten liiketoimintaa, mitä ongelmia tähän murrokseen liittyy, sekä kuinka synteettinen data pystyisi ratkaisemaan näitä ongelmia.

1.2 Tutkimusmenetelmät ja rajaukset

Tämä kandidaatin työ toteutetaan kuvailevana integroivana kirjallisuuskatsauksena. Kuvaileva kirjallisuuskatsaus on yleiskatsaus aiheesta eikä aineistojen valintaa rajaa metodiset säännöt (Salminen 2011, 6). Työ rakentuu aiheeseen liittyvän teoriakirjallisuuden ympärille, työssä käytetyt lähteet ovat suurimmaksi osaksi tiedelehtiartikkeleita sekä kirjoja. Taulukkoon 1 on koottu työn tärkeimmät hakutyökalut sekä hakusanat. Salmisen (2011, 8) mukaan integroiva kirjallisuus on hyvä tapa tuottaa uutta tietoa jo tutkitusta aiheesta sekä se sallii erilaisin metodisin lähtökohdin tehdyt tutkimukset analyysin pohjaksi. Tiedonhakua on toteutettu pääosin englanniksi, sillä suomenkielisillä hakusanoilla hakutulosten määrä on hyvin pieni tai osassa tapauksia olematon. Tämän vuoksi työtä varten on käännetty monia keskeisiä avainsanoja englannista suomeksi. Joillekin avainsanoille löytyy vakiintunut suomennos, mutta tapauksissa, joissa vakiintunutta suomennosta ei ole ollut saatavilla, on kirjoittaja suomentanut termin parhaaksi näkemällään tavalla. Työssä on hyödynnetty myös Chat GPT:tä työn alkuvaiheessa auttamaan aiheen ideoinnissa sekä sen rajaamisessa. Chat GPT:tä ei ole käytetty tiedon etsimiseen, vaan kirjoittajan omien ajatusten ja ideoiden jalostamisen tukena.

Taulukko 1: Työn tärkeimmät hakutyökalut ja hakusanat

Tutkimuksen tärkeimmät hakutyökalut	Tärkeimmät hakusanat
• Lut Primo • Google Scholar • Scopus	• ”Synthetic data” AND (”Industry” OR ”manufacturing) • ”Industrial big data” • ”Synthetic data utility” • ”Synthetic data fidelity” • ”Industrial smart PSS” AND ”Synthetic data” • CRISP-DM

Työ on rajattu tutkimaan synteettisen datan hyödyntämismahdollisuuksia tarkemmin teollisuusyrityksen näkökulmasta, vaikkakin työssä mainitaan muitakin mahdollisia yhteiskunnallisia aloja, joilla synteettistä dataa voidaan hyödyntää. Työssä ei paneuduta syvällisesti synteettisen datan luomisprosesseihin ja niiden takana oleviin algoritmeihin, vaan keskitytään enemmän synteettisen datan sovelluksiin. Työn alussa tarkastellaan myös datan käsitettä, mutta tämä tarkastelu on rajattu koskemaan suuria teollisia datamääriä.

1.3 Tutkimuksen rakenne

Tämä kandidaatintyö koostuu viidestä pääluvusta, joista ensimmäinen on johdanto, kolme seuraavaa ovat aiheen käsittelykappaleita ja viides luku on työn johtopäätökset. Johdannossa avataan työn taustoitusta perustellen aihe synteettisen datan ajankohtaisuudella. Johdantoluku käsittelee myös työn tutkimuskysymyksiä, tutkimusmenetelmiä sekä aiheen rajausta. Toinen pääluku käsittelee datan hyödyntämistä teollisuusyrityksissä. Luvussa kerrotaan mitä haasteita ja mahdollisuuksia suuriin datamääriin liittyy, sekä mitä ominaispiirteitä suurilla teollisilla datamäärillä on. Kolmannessa pääluvussa käsitellään synteettistä dataa, mitä se on, sekä mitä riskejä ja mahdollisuuksia siihen liittyy. Luvussa käydään myös läpi synteettisen datan tuottamisprosessia. Neljäs pääluku käsittelee synteettisen datan soveltamismahdollisuuksia teollisuusyrityksissä ja esimerkkejä, kuinka yritys tästä hyötyy. Työn viimeisessä, eli viidennessä pääluvussa käydään läpi työn johtopäätökset.

2 Datapohjainen liiketoiminta

Nykypäivänä yritykset pyrkivät hyödyntämään dataa ja digitaalisia teknologioita yhä enemmän. Ollakseen mahdollisimman tehokkaita, yritykset haluavat mahdollisimman paljon dataa, mahdollisimman tarkasti ja oikeaan aikaan (Fleckenstein & Fellows 2018, 15). Sah (2022, 1) mukaan oikean datan saaminen oikeassa muodossa ja oikeaan aikaan on ensimmäinen kriittinen askel kohti parannettua asiakaskokemusta, uusien tuotteiden julkaisua, tuottavuuden lisäämistä ja parempaa riskienhallintaa.

2.1 Teollinen massadata

Kirjallisuus antaa datalle moninaisia määritelmiä, mutta Keith (2007, 3) tiivistää, että suurimmassa osassa määritelmissä toistuvat samat piirteet, data on tosiasioita, tapahtumia ja vastaavia, jotka on tallennettu. Kansainvälinen standardisoimisjärjestö (ISO/IEC 2382:2015) määrittelee datan niin, että data on uudelleentulkittava esitys informaatiosta, jota voidaan soveltaa viestintään, tulkintaan sekä käsittelyyn. Dataa voidaan käsitellä joko ihmisten tai koneiden toimesta. Datan määrittelemisessä on huomioitavaa, että määritelmä voi hieman vaihdella tieteenaloittain sekä kontekstin mukaan. Useita määritelmiä kuitenkin yhdistää yhteinen näkemys siitä, että data on merkkijonoja, numeroita tai muita symboleja, jotka edustavat faktoja.

Maailma upottaa itseänsä tietoon ja dataan. Tätä uutta massiiviseen datamäärään liittyvää aikakautta voidaan kutsua massadatan aikakaudeksi (eng. Big data) (Fink, Guzzo, King, Landis & Tonidandel 2015). Suuria määriä dataa kerätään monilta aloilta, kuten mediasta, terveydenhuollosta, tieteestä, liikkumisesta ja ihmisten kuljettamisesta sekä teollisuudesta. Yleisesti ottaen massadatan ominaisuuksia voidaan tarkastella viiden ominaisuuden avulla, joista englanninkielisessä kirjallisuudessa käytetään nimitystä 5V. Nämä massadatan ominaisuudet ovat määrä (volume), monimuotoisuus (variety), nopeus (velocity), totuudenmukaisuus (veracity) ja arvo (value). (Babic, Gace, Pevec, Podobnik, Sabolic & Vdovic 2019)

Teollisuuden kontekstissa suurista datamääristä puhutaan nimellä teollinen massadata (eng. industrial big data) ja sillä viitataan teollisuuden prosesseista saataviin suuriin datamääriin.

Teollisuudesta tuotettua massadataa voidaan käsitellä ja ymmärtää samoilla viiden V:n malleilla kuin perinteisempääkin massadataa lisäten siihen vielä kaksi ulottuvuutta, vaihtelevuus (variability) ja näkyvyys (visibility) (Basanta-Val 2018). Li, Lu, Meng & Yan (2017) ehdottavat samaa viiden V:n mallia teollisen massadatan tarkasteluun kuin Babic et al. (2019). Tässä työssä teollista massadataa tarkastellaan viiden V:n mallin avulla, sillä se näyttää olevan suosituin malli kirjallisuudessa. Teollisen massadatan piirteisiin kuuluu, että se on usein hyvin rakenteellista ja korreloitunutta, sillä teollisuuden massadata syntyy automatisoiduista laitteista hallituissa ympäristöissä. (Basanta-Val 2018) Teollisuusyrityksestä voidaan dataa kerätä lukuisista eri toiminnoista, kuten koneiden toimintatietoja, henkilöstön käyttäytymistietoja, tuotteiden kustannustietoja, logistiikan tietoja, tuotelaadun tietoja, asiakastietoja sekä koneiden ja tuotteiden käyttötietoja sisältäen vikaprosentit ja korjausasteet (Li et al. 2017).

Li et al. (2017) mukaan 5V ominaisuudet teollisesta massadatasta mahdollistavat kattavan ja systemaattisen tiedon tarjoamisen datan louhinnan ja analytiikan kautta. Datalouhintaa voidaan käyttää esimerkiksi laite- ja tuotesuunnitteluvirheiden löytämiseksi, tuotantoprosessien vikojen löytämiseksi sekä henkilöstön ja asiakkaiden käyttäytymistapojen havainnollistamiseksi. Ihmisistä, laitteista, tuotantoprosessista ja ympäristöstä kerätty data voi mahdollistaa tuotantokustannusten alentamista ja parantaa tuotelaatua. (Li et al. 2017)

Määrä

Tätä ensimmäistä massadataan liittyvää käsitettä pidetään jopa tärkeimpänä ja merkittävimpänä massadatan ominaisuutena. Tällä ominaisuudella viitataan massiivisiin datamääriin, joita tuotetaan nykyään. Massiiviset datamäärät asettavat lisävaatimuksia kaikille työkaluille ja tekniikoille, joita käytetään. (Demchenko, Grosso & Laat 2013) Li et al. (2017) mukaan myös teollisuudesta kerätyn datan määrä on kasvanut todella paljon lisääntyvien esineiden internet (eng. internet of things) ratkaisuiden myötä. Myös Kamradt, Nassif & Tekli (2024, 13) yhtyvät väitteisiin siitä, että teollisesta ympäristöstä kerätyn datan määrä on jo nyt valtava ja tulee kasvamaan edelleen tulevaisuudessa.

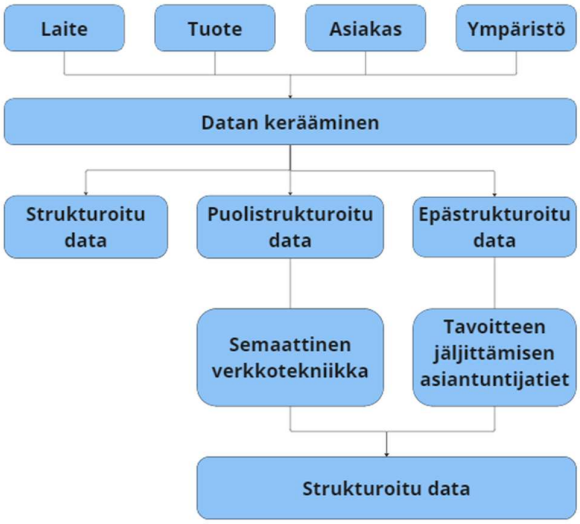
Nopeus

Toinen myös teolliseen massadataan liittyvä piirre on datan luomisen ja sen analysoinnin nopeus (Babic et al. 2019). Teollisesta ympäristöstä dataa voidaan kerätä monella eri nopeudella eli datan keräämisen frekvenssi ei ole vakio. Käytännössä tämä tarkoittaa, että datahavaintoja voidaan saada laitteista satoja tai jopa tuhansia kertoja sekunnissa. (Li et al. 2017)

Calic & Ghasemaghaei (2019) mukaan datan nopeus voidaan määritellä sen mukaan, kuinka nopeasti dataa on prosessoitu ja analysoitu. Heidän mukaansa edistyneet älykkäät metodit ja ratkaisut mahdollistavat nopealla tahdilla luodun datan analysoimisen ja hyödyntämisen reaaliajassa.

Monimuotoisuus

Yksi massadatan merkittävimmistä haasteista on käsitellä suuria määriä dataa eri lähteistä. Datan eri lähteet tuottavat tietoa eri muodoissa, kuten strukturoitua-, puolistrukturoitua- sekä epästrukturoitua dataa. (Babic et al. 2019) Strukturoidulla datalla tarkoitetaan tuotantolaitteista ja koneista saatua dataa, joka on tietyssä formaatissa, esimerkiksi Excel muotoisena tai SQL tietokantana. Tällaisessa muodossa olevaa dataa on ihmisten ja tietokoneiden helppo käsitellä ja jalostaa. Puolistrukturoitu data on esimerkiksi internetistä ja asiakaspalautteista saatua dataa XML muodossa ja epästrukturoitu data pitää esimerkiksi sisällään ääniä, videoita ja kuvia. Puoli- ja epästrukturoitu data tulisi muuntaa strukturoituun muotoon ennen niiden hyödyntämistä (kuva 2). (Li et al.2017) Li et al. (2017) mukaan data täytyy olla strukturoidussa muodossa, jotta sitä voidaan hyödyntää analyyttisesti.



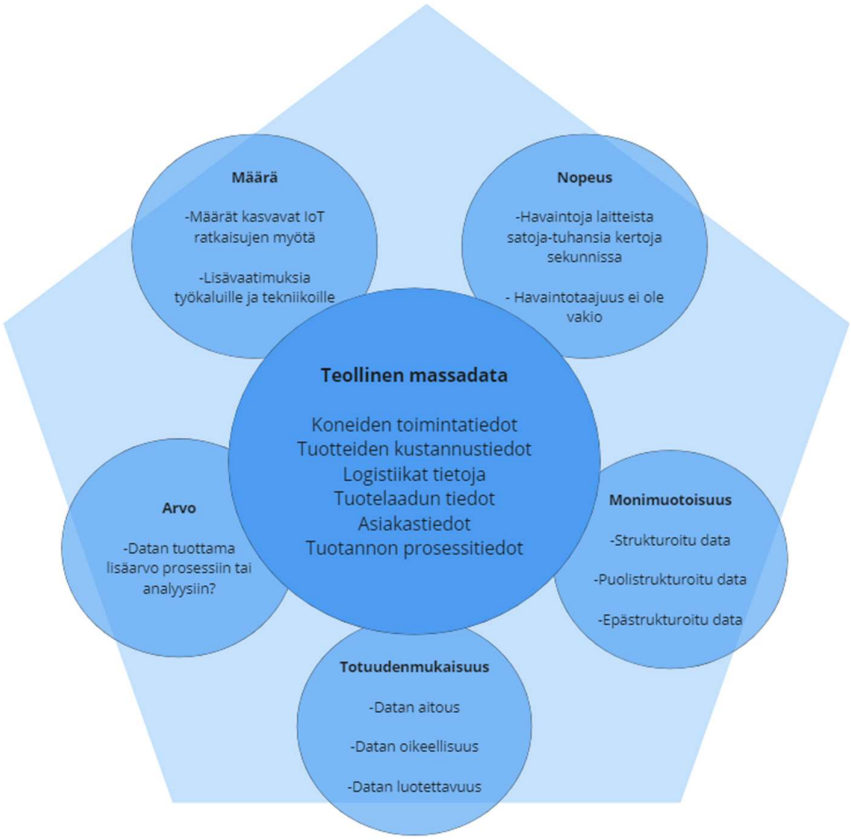
Kuva 2: Teollisen datan strukturointi (mukaillen Li et al. 2017)

Totuudenmukaisuus

Tällä massadatan ominaisuudella viitataan datan luotettavuuteen, oikeellisuuteen ja aitouteen. Datan luotettavuuden varmistaminen ja datan suojaaminen luvattomalta käytöltä ja muokkaukselta muodostavat suuria haasteita, kun dataa on kerätty useista lähteistä. (Babic et al. 2019)

Arvo

Massadatan arvo-ominaisuus on hyvin tärkeä osa dataa ja se määritellään kerätyn datan tuottamasta lisäarvosta prosessiin, toimintaan tai analyysiin. Big Datan yhtenä ensisijaisena tehtävänä onkin jalostaa valtavista tietomääristä arvokkaita päätelmiä ja havaintoja. Tämä arvo voi ilmetä monella eri tapaa, esimerkiksi parempana päätöksentekona ja prosessien optimointina. (Demchenko et al. 2013) Nämä teollisen massadatan viisi ominaisuutta on koottu yhteenvetona alla olevaan kuvaan 3.



Kuva 3: Teollinen massadata

Paljon puhutaan datasta ja sen tärkeydestä, mutta yleensäkään datalla itsessään ei ole mitään itseisarvoa, vaan siitä tulee hyödyllistä, kun se muutetaan informaatioksi ja yritys käyttää tätä informaatiota päätöksentekonsa tukena. Jotta yritys voi saavuttaa datan avulla liiketoiminnallisia tavoitteita, niiden täytyy ymmärtää, minkälaista dataa heillä on, minkälaista dataa he haluavat tulevaisuudessa ja kuinka organisoida tehokkaasti tätä dataa (Fleckenstein & Fellows 2018, 15). Sah (2022, 1) mukaan data ja data-analytiikka muodostavat yrityksen digitaalisen strategian ytimen ja näin ollen epäonnistuminen data- ja analytiikka-aloitteessa johtaa myös muihin epäonnistumisiin yrityksen digitalisaatiossa. Tämä epäonnistumisien määrä data ja data-analytiikka aloitteissa on yllättävän korkea. Monet analyytikot ovat tulleet hieman eri johtopäätöksiin näistä onnistumisprosenteista, mutta yleinen konsensus on, että jopa puolet näistä aloitteista epäonnistuu saavuttamaan liiketoiminnalliset tavoitteensa (Sah 2022, 1). Fleckenstein & Fellows (2018, 18) yhtyvät myös lausuntoihin data-strategioiden tärkeydestä ja heidän mukaan vain noin kymmenesosalla organisaatioista on virallinen data-strategia, joka kattaa datan hallinnan. COVID-19 pandemia paljasti merkittäviä puutteita ja eroja erityisesti eri teollisuudenalojen digitaalisen kypsyyden tasoissa, koskien esimerkiksi reaaliaikaisen dataprosessoinnin kykyä, digitaalisten prosessien kypsyyttä, sekä sopeutumisen nopeutta ja halukkuutta sopeutumiseen (Luttgens, Mertens, Nitsch, Piller, Putz & Van Dyck 2022, 2). Ilman oikeanlaista viitekehystä dataprojektin onnistuminen riippuu todella paljon vain ja ainoastaan sitä suorittavasta työryhmästä, eikä onnistuneita projekteja saada toistettua organisaation sisällä (Hipp & Wirth 2000).

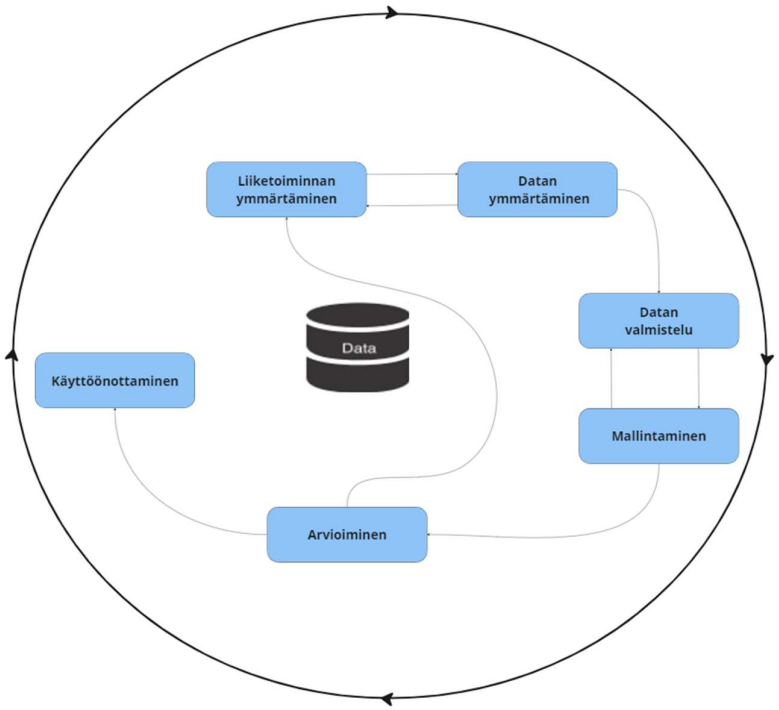
Dataa täytyy siis käsitellä ja hyödyntää, jotta sillä pystytään tuottamaan arvoa yritykselle. Mikään yritys ei hyödy datan keräämisestä ja sen varastoinnista, mikäli sitä ei käytetä minkään liiketoimintaprosessin tukena. Seuraavien kappaleiden tarkoituksena on käsitellä perinteisiä data-analytiikan prosesseja, joilla voidaan konkretisoida arvontuotantoa datan avulla.

2.2 Datapohjainen arvontuotanto

Yksi tällainen prosessikehys, jota yritykset voivat hyödyntää datan louhinnassa ja data-analytiikan projektien toteuttamisessa on nimeltään CRISP-DM (Cross-industry standard process for data mining). Tämä tiedonlouhintamenetelmä on alun perin kehitetty vuonna 1999 ja on tänä päivänäkin yksi suosituimmista ja käytetyimmistä analytiikkametodeista, sekä

kirjallisuudessa eniten viitattu metodi. CRISP-DM prosessia pidetään myös täydellisimpänä tiedonlouhintaprosessina etenkin teollisuuden tarpeisiin. Muita vastaavia metodeja on esimerkiksi IBM:n kehittämä ASUM-DM sekä SAS:n kehittämä SEMMA, sekä alkuperäisestä CRISP-DM prosessista laajennetut CRISP-DM 2.0 sekä CASP-DM, joka on erityisesti tarkoitettu vastaamaan koneoppimisen tuomiin haasteisiin ja mahdollisuksiin. (Contreras-Ochando, Ferri, Flach, Hernandez-Orallo, Kull, Lachiche, Martinez-Plumed & Ramırez-Quintana, 2021). Tässä työssä käytetään kuitenkin CRISP-DM prosessikehystä sen yleisyyden ja suosion takia.

CRISP-DM prosessimalli koostuu yhteensä kuudesta eri vaiheesta: liiketoiminnan ymmärtäminen, datan ymmärtäminen, datan valmistelu, mallintaminen, arvioiminen sekä toteuttaminen (Kamal & Pahuja, 2023). Kuva 4 kuvaa CRISP-DM prosessimallin eri vaiheita ja sitä, kuinka ne liittyvät toisiinsa erilaisten takaisinkytkentöjen seurauksina. Näiden kuuden vaiheen järjestys ei ole tiukasti määritelty, vaan oikeassa projektissa edellisen vaiheen tulos määrittää mikä vaihe täytyy suorittaa seuraavaksi (Hipp & Wirth 2000). Hipp & Wirth (2000) mukaan datalouhinta ei myöskään ole suoraviivainen, yhdeltä istumalta suoritettava prosessi, vaan se on syklistä sillä prosessin aikana opitut asiat herättävät usein uusia kysymyksiä liiketoimintaprosessiin liittyen.



Kuva 4: CRISP-DM prosessimalli (mukaillen Contreras-Ochando et al. 2021)

Liiketoiminnan ymmärtäminen

Tässä ensimmäisessä vaiheessa pyritään ymmärtämään projektin tavoitteet ja vaatimukset liiketoiminnan näkökulmasta. Nämä tiedot siten muutetaan datalouhinnan ongelman määritelmäksi ja alustavaksi projektisuunnitelmaksi. (Hipp & Wirth 2000) Keith (2007, 2) mukaan datasta tulee informaatiota ja näin ollen tietoa, kun sitä käsitellään niin että se on vastaanottajansa käytettävissä, data asetetaan kontekstiin ja sille annetaan merkitys. Tätä datan, informaation ja tiedon välistä yhteyttä on kuvattu kuvassa 5. Tietämystä voidaan pitää organisaation suurimpana strategisena vahvuutena. Tiedon avulla yritykset pystyvät tekemään parempia päätöksiä, tehostamaan toimintojaan ja prosessejaan, edistämään innovaatioita, tunnistamaan uusia liiketoimintamahdollisuuksia sekä pystyvät parempaan riskienhallintaan. (Feng, Guo & Tu 2024)

Kuva 5: Datan, informaation ja tiedon välinen yhteys

Liiketoiminnan ymmärtäminen voidaan jakaa vielä neljään tarkempaan vaiheeseen, joista kustakin saadaan jokin tuotos ulos. Ensimmäisessä vaiheessa määritetään liiketoiminnan tavoitteet keräämällä tietoa liiketoimintaympäristöstä ja yrityksen tai prosessin nykytilasta. Tavoitteena on tunnistaa projektin tavoitteet ja ymmärtää, mitä projektilla halutaan saavuttaa. Toisessa vaiheessa arvioidaan tilannetta, tunnistetaan pääresursseja, kuten projektiin kuuluvaa henkilöstöä, sekä laitteita ja ohjelmistoja, joita on mahdollista käyttää projektissa. Olennaisena osana on myös projektin riskienhallinta. Kolmannessa vaiheessa määritetään tietojen louhinnan tavoitteet teknisestä näkökulmasta, tuotetaan suunnitelma teknisin termein, eli mistä dataa saadaan ja miten. Neljännessä vaiheessa tuotetaan yksityiskohtainen projektisuunnitelma, jossa kuvataan yksityiskohtaisesti datan louhinnan vaiheet ja tekniikat. (Blanco, Illarramendi & Nino 2015)

Datan ymmärtäminen

Datan ymmärtämisen vaihe alkaa datan keräämisellä ja jatkuu toimilla, joiden tarkoituksena on tulla tutuksi datan kanssa. Näiden toimien tarkoituksena on tunnistaa datan laatuun liittyviä ongelmia, saada ensimmäisiä käsityksiä datasta, sekä tunnistaa mahdollisia osajoukkoja hypoteesien muodostamiseksi. Liiketoiminnan ymmärtäminen ja datan ymmärtäminen ovat tiiviisti yhteydessä toisiinsa; datalouhinnan ongelman laatiminen ja projektisuunnitelman laatiminen vaativat ymmärrystä käytettävästä olevasta datasta. (Hipp & Wirth 2000)

Teollisessa kontekstissa datan ymmärtäminen pitää sisällään ymmärrystä datan lähteistä. Dataa voidaan saada esimerkiksi eri tuotantokoneista sensoreiden avulla sekä tuotannonohjausjärjestelmistä. (Bekar, Nyqvist, Skoogh 2020) Bekar et al. (2020) mukaan haasteita aiheuttaa se, ettei eri datalähteiden välillä välttämättä ole yhteyttä ja näin ollen niiden aikasynkronointi voi aiheuttaa suuria haasteita datan valmistelussa.

Datan valmistelu

Datan valmistelu sisältää ne toimenpiteet, joilla raakadatasta rakennetaan lopullinen datasetti, joka syötetään mallinnustyökaluihin. Datan valmistelu sisältää useita vaiheita, joita ovat esimerkiksi taulukoiden ja attribuuttien valinta, datan puhdistus, uusien attribuuttien luominen sekä datan muuntaminen mallinnustyökalua varten. (Hipp & Wirth 2000)

Bekar et al. (2020) sanovat valmisteluvaiheen sisältävän esimerkiksi datan puhdistamista, missä epäkelpoja ja vääristyneitä datapisteitä siivotaan pois. Datan valmisteluun voi sisältyä myös rivien poistamista, joista puuttuu jokin datapiste, kaksinkertaisten datapisteiden siivoamista, sekä poikkeavien datapisteiden poistamista (Cortez, Iliadis, Macintyre & Maglogiannis 2022). Bekar et al. (2020) sekä Cortez et al. (2022) käyttivät myös datan standardointitekniikoita valmisteluvaiheessa. Molemmissa standardoinneissa datan arvot muutettiin uudelle asteikoille, jossa keskiarvo on 0 ja keskihajonta on 1.

Mallintaminen

Mallintamisvaiheessa valitaan eri mallinnustekniikoita sekä niiden parametreja säädetään optimaalisimpiin arvoihin. Jotkin tekniikat edellyttävät tiettyjä datamuotoja ja yleensä samalle datan louhintaongelman tyypille on useita eri mallinnustekniikoita. Myös datan valmistelun ja mallintamisen välillä on vahva yhteys, sillä yleensä datan ongelmia huomataan

myös mallintamisen aikana, tai vastaavasti uusia ideoita datan rakentamiseksi. (Hipp & Wirth 2000)

Arvioiminen

Tässä vaiheessa projektia on luotu yksi tai useampi data-analytiikan malli, jotka ovat toimivia ja laadukkaita datan analyysin näkökulmasta. Ennen siirtymistä mallin lopulliseen käyttöönottoon, täytyy mallia arvioida vielä perusteellisemmin ja käydä läpi mallin rakentamisen vaiheet, jotta varmistutaan siitä, että malli saavuttaa halutut liiketoiminnan tavoitteet. Tärkeää on arvioida, onko jotain liiketoiminnan ongelmia, joita ei huomioitu tarpeeksi. Arviointivaiheessa tulisi myös tehdä päätös datan louhinnan tulosten käytöstä. (Hipp & Wirth 2000)

Käyttöönottaminen

Mallin luominen ei yleensä ole projektin loppu, vaan saatua tietoa on yleensä tarve järjestellä ja esittää muodossa, jossa asiakas voi sitä hyödyntää. Käyttöönoton vaihe voi olla niinkin yksinkertainen kuin raportin luominen tai yhtä monimutkainen kuin louhintaprosessin toteuttaminen. Monissa tapauksissa käyttäjä, ei data-analyytikko, suorittaa käyttöönottoon liittyvät vaiheet. Joka tapauksessa on kuitenkin tärkeä ymmärtää jo etukäteen, mitä toimenpiteitä tarvitaan luodun mallin käyttöönottamiseksi käytännössä. Tässä vaiheessa on myös tärkeä arvioida, kuinka koko projekti saavutti sille asetetut liiketoiminnan tavoitteet. (Hipp & Wirth 2000)

2.3 Teolliset älykkäät tuotepalvelujärjestelmät

COVID-19 pandemia haastoi yhteiskuntaa, politiikkaa ja taloutta odottamattomasti ja toimi katalyyttina vahvistaen olemassa olevia trendejä yritysten digitalisaatioon liittyen (Luttgens et al. 2022, 2). Yleisestikin digitalisaatioon ja edistyneeseen massadata-analytiikkaan liittyvät teknologiat ovat käynnistäneet laajamittaisen yhteiskunnallisen muutoksen, synnyttäen uuden datavetoisen liiketoiminta-aikakauden, mikä vaikuttaa yritysten strategioihin, liiketoimintamalleihin ja prosesseihin, sekä luo uusia työskentelytapoja (Carillo 2017). Luttgens et al. (2022, 2) mukaan pandemiasta seuranneiden toimitusketjujen häiriöiden ja työvoimapulan seurauksena yritykset kaikilla aloilla pyrkivät tulevaisuudessakin automatisoimaan tuotantoaan enemmän ja muuttamaan palveluitaan digitaalisiksi, ne liiketoimintamallit

menestyvät parhaiten, jotka pystyvät vastaamaan muuttuviin taloudellisiin, sosiaalisiin ja yhteiskunnallisiin vaatimuksiin.

Myös teollisuusyritykset ovat omaksuneet muutoksen kohti palvelukeskeisempää liiketoimintaa ja muuttavat toimintaansa kohti tuotepalvelujärjestelmä-painotteista liiketoimintamallia (Chowdhury, Haftor & Pashkevich 2018). Tuotepalvelujärjestelmä on liiketoimintamalli, joka menee pidemmälle kuin pelkän tuotteen myynti, yhdistämällä ratkaisuun myös palvelun tarjoamisen (Kundu & McKay 2014). Chowdhury et al. (2018) mukaan perinteisistä tuotepalvelu-järjestelmistä on teknologien kehityksen myötä kehittynyt älykkäitä tuotepalvelujärjestelmiä. Teollisuuden kontekstissa älykkäät tuote-palvelu-järjestelmät tunnetaan nimellä industrial smart product service systems. Älykkäiden teollisten tuotepalvelujärjestelmien tarkoituksena on integroida teolliset laitteet ja palvelut vastaamaan asiakkaan elinkaarivaatimuksia business-to-business-liiketoimintamalleissa. (Niu, Song, Tang, Wang, Zheng & Zheng 2022)

Arvonluonti älykkäillä tuotepalvelujärjestelmillä

Teollisuusyrityksissä etädiagnostiikaan perustuvat älykkäät tuotepalvelujärjestelmät (Smart PSS) mahdollistavat kaksi erilaista arvoketjua. Ensimmäisessä arvoketjussa valmistaja pystyy keräämään etänä koneeseen liittyviä tietoja ja näin ollen tarjoamaan asiakkaalle palveluita liittyen koneen korjaukseen ja ylläpitoon. Toisessa arvoketjussa valmistaja voi asentaa ohjausjärjestelmän asiakkaan tiloihin, jolloin asiakas pystyy tarkastelemaan laitteen tuottamia tietoja ja käyttää niitä esimerkiksi oman tuotantonsa optimointiin. (Chowdhury et al. 2018) Asiakkaan ja laitevalmistajan välisessä arvonluonnissa esiintyy myös ongelmia liittyen tiedon saamiseen ja sen jakamiseen. Grubic & Peppard (2016) sanovat, että käytännössä tämä tarkoittaa esimerkiksi sitä, että laitedataa tutkiva analyytikko on tuhansien kilometrien päässä ja työskentelee datalla, joka antaa merkittävästi suppeamman kuvan todellisuudesta. Tämä voi johtua siitä, että laitedataa on yksinkertaisesti vaikea kerätä, tai asiakas ei ole yhteistyökykyinen datan jakamisessa (Grubic & Peppard 2016).

Nämä arvonluontitavat edellyttävät siis yhteistyötä ja toistensa ymmärtämistä asiakkaan ja laitevalmistajan välillä. Grubic & Peppard (2016) mukaan on tärkeää, että myös asiakas ymmärtää tämän tiedonjakamisen hyödyt laitevalmistajan kanssa ja että ilman asiakkaan tukea ei laitevalmistaja pysty välttämättä yksin tarjoamaan näitä arvolupauksia. Heidän mukaansa

näitä arvonluontiprosesseja tulisi lähestyä arvon yhteiskehittämisprojekteina asiakkaan ja laitevalmistajan tai palveluntarjoajan välillä.

Case Valmet Oyj

Valmet on Suomen johtavia teknologiayrityksiä, joka tuottaa teknologiaa, palveluita ja automaatiota paperi-, sellu-, ja energiateollisuudelle. Yli 40 % maailmassa valmistetusta paperista ja kartongista valmistetaan Valmetin koneilla. Yritys on lanseerannut datapohjaisista palveluista palveluportfolion, mutta nämä palvelut tuottavat vielä prosentuaalisesti pienen osan sen liikevaihdosta. Näiden palveluiden tarkoituksena yhdessä datan kanssa on auttaa esimerkiksi tunnistamaan mahdolliset vikaantumiset ennalta ja estämään laitteiden rikkoutumista, parantamaan laitteiden käyttöastetta ja optimoimaan laitteiden huoltoa. Dataa saadaan paperikoneissa olevista mittalaitteista ja yhdessä paperikoneessa voi olla 15 000 erilaista mittapistettä. (Eloranta, Paavola & Seppänen 2021, 80–83)

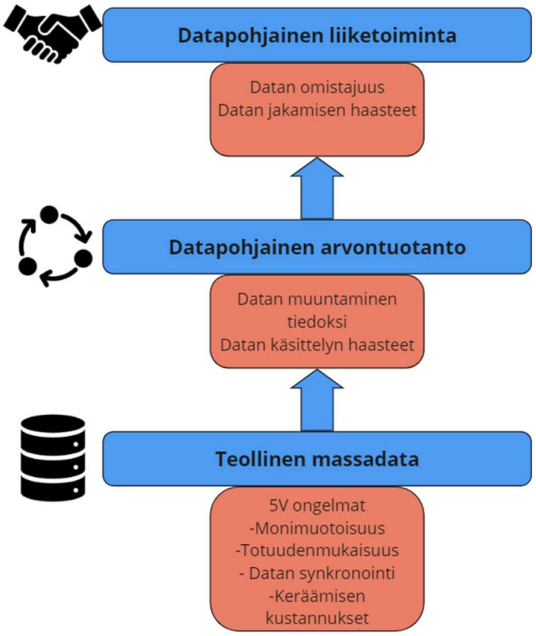
Valmetilla datapohjaiset palvelut ovat integroituneet osaksi muuta liiketoimintaa ja suurin osa datapohjaisesta liiketoiminnasta tulee automaatioliiketoiminnasta, joka oli vuonna 2020 10 % yrityksen liikevaihdosta. Tyypillisiä datapohjaisia palveluita ovat esimerkiksi asiantuntijapalvelut, joissa ratkotaan prosessiin tai laitteeseen tullutta ongelmaa etäyhteyden kautta. Datan analysoinnin, yhdistämisen ja koneen toimintaa koskevan ymmärryksen kautta tehdään ratkaisuehdotus, kuinka pitäisi toimia. Korona-aika vaikutti teollisen internetin palveluihin myönteisesti muuttamalla asiakkaiden asennetta etätekemisestä ja datan hyödyntämisestä myönteisempään suuntaa. Myös Valmetin näkökulmasta näyttää siltä, että korona-aika vauhditti digitalisaatiota, kun kaikki joutuivat totuttelemaan etätyön tekemiseen. Yrityksessä uskotaan datapohjaisen liiketoiminnan kasvavan tulevaisuudessa erityisesti tekoälyn hyödyntämisen kautta kaikilla liiketoiminta-alueilla. (Eloranta et al. 2021, 80–83)

Valmetilla datapohjaisen liiketoiminnan kehittymisen haasteet liittyvät suurilta osin dataan ja kysymys datan omistajuudesta on keskeinen hidaste liiketoiminnan kehittämisessä. Laitedataan on helpompi saada käyttöoikeus kuin prosessidataan ja sitä käytetäänkin vapaammin. Laitedata ei ole asiakkaalle kilpailutekijä niin kuin prosessidata on. Valmetin asiakkaat eivät halua jakaa prosessidataa, joka sisältää kilpailuedulle oleellisia tietoja, sillä se paljastaa missä asiakas toimii paremmin kuin kilpailijansa. Prosessidatan saaminen tarjoaisi kuitenkin Valmetille monia mahdollisuuksia palveluiden ja tuotteiden kehittämiseen, mutta toistaiseksi sitä ei ole voitu käyttää vapaasti. (Eloranta et al. 2021, 80–83) Eloranta et al. (2021,

84) pohtivat kuinka datan saatavuutta voidaan edistää esimerkiksi yhteisten alustojen kautta tarjoamalla prosessidataa anonyymin tietopankin kautta. He kuitenkin toteavat, että tästä saatava hyöty olisi vain rajallista, sillä datan arvo syntyy usein vasta asiakaskontekstin ymmärtämisen kautta.

Yhteenveto

Edellisten kappaleiden perusteella voidaan siis todeta, että data on yhä tärkeämmässä roolissa teollisuusyrityksissä. Datan avulla voidaan parantaa yrityksen sisäistä tehokkuutta ja löytää uusia datavetoisia liiketoimintamahdollisuuksia. Mahdollisuuksia on loputtomiin, mutta näitä kaikkia haastaa yhteiset ongelmat. Näitä datan aiheuttamia haasteita yrityksen eri tasoilla on kuvattu alla olevassa kuvassa 6.



Kuva 6: Datan haasteita teollisuusyrityksessä

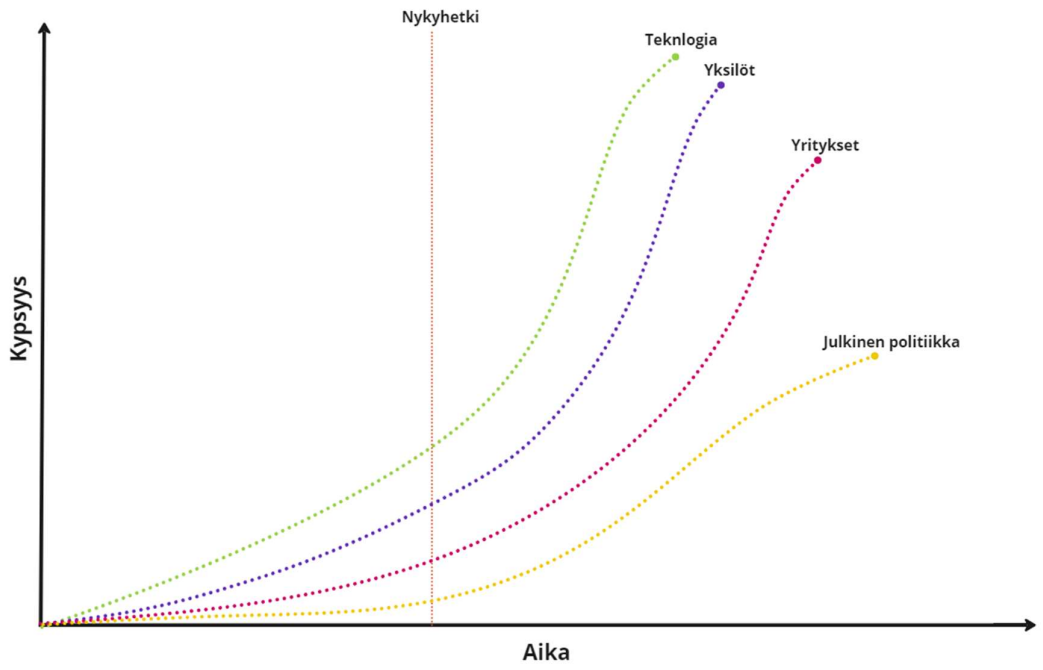
Kuvan alimmalla tasolla on kuvattu yhteenvedon omaisesti punaisessa laatikossa teollisen massadatan ongelmia. Kuvan keskimmäisellä tasolla kuvataan datan prosessointiin liittyviä haasteita ja ylimmällä tasolla, mitä ongelmia liiketoimintamahdollisuuksien näkökulmasta data voi aiheuttaa.

3 Synteettinen data

Nopeiden, monimuotoisten ja korkean volyymin datasettien saatavuus yhdistettynä edistyneisiin tilastollisiin työkaluihin tarjoaa mahdollisuuden kiihdyttää tutkimustyötä ja innovaatiota, sekä parantaa päätöksentekoa. Myös monet suuret datasetit voivat olla arkaluontoisia ja niiden käyttäminen ja jakaminen saattaa rikkoa perusoikeuksia, jotka on suojattu yksityisyyttä koskevilla säädöksillä. (Bottarelli, Cohen, Cherubin, Houssiau, Jordon, Maple, Szpruch & Weller 2022) Toisaalta, oikean datan määrä maailmassa kasvaa eksponentiaalisesti ja koko maailman datamäärästä 90 % on luotu viimeisen kahden vuoden aikana, pääasiallisesti IoT -ratkaisuiden ja niissä olevien sensoreiden lisäännyttyä (Patel 2024). Näin massiivisen datatulvan keskellä herää kysymys, onko synteettiselle datalle tarvetta maailmassa? Tämän luvun tarkoituksena on antaa lukijalle teoriakatsaus siitä, mitä tarkoitetaan synteettisellä datalla, mitä mahdollisuuksia ja rajoitteita siihen liittyy, sekä kuinka se pystyisi ratkaisemaan dataan liittyviä nykyisiä ongelmia.

Birisci, Celik & Gürsakal (2023, 2–9) kertovat esimerkkejä, miksi synteettinen data on tärkeää ja kuinka se on joissain tapauksissa parempaa kuin perinteinen data. Heidän mukaansa synteettistä dataa voidaan käyttää kouluttamaan ja testaamaan koneoppimismalleja, kun oikeaa dataa ei ole saatavilla, johtuen esimerkiksi sen vaikeudesta ja kalleudesta kerätä. Synteettistä dataa voidaan käyttää myös korvaamaan sellaisia datajoukkoja, jotka pitävät sisällään paljon arkaluontoista informaatiota, esimerkiksi henkilötietoja. He mainitsevat, että toinen tapa hallita näitä arkaluontoisia datajoukkoja on anonymisoida niissä olevaa dataa. Kuitenkaan tämä ei aina takaa täysin anonyymia datajoukkoa, joten yksittäisten datapisteiden tunnistaminen on periaatteessa mahdollista. Anonymisoinnissa on myös riskinä, että tärkeää tietoa menetetään tai sitä ei pystytä hyödyntämään. (Birisci et al. 2023, 2–9) Synteettisellä datalla on monia hyötyjä verrattuna oikeaan dataan. Birisci et al. (2023, 3) mukaan synteettinen data on kontrolloidumpaa ja paremman laatuista kuin oikea data, sekä sitä voidaan tuottaa juuri sen verran kuin nähdään tarpeelliseksi. Kuva 7 kuvaa, kuinka synteettisen datan luomiseen tarkoitettu teknologia on kypsynyt nopeammin kuin sen hyödyntäminen organisaatioissa. Käytännössä tämä tarkoittaa sitä, että synteettisen datan luomiseen on olemassa teknologiaa, mutta ihmiset ja organisaatiot eivät ole vielä oppineet käyttämään synteettistä

dataa. Synteettisen datan luominen ei ole siis esteenä synteettisen datan käytön lisääntymiselle.



Kuva 7: Synteettisen datan täysimittainen omaksuminen eri sektoreittain (mukaillen Branson, Harbron, James & Sundler 2021)

3.1 Synteettisen datan määritelmä

Synteettinen data on keinotekoisesti tuotettua dataa, jonka tarkoituksena on heijastella oikean datan tilastollisia ominaisuuksia (Branson et al. 2021). Birisci et al. (2023, 1) mukaan yksinkertaisimmin sanottuna, synteettinen data on luotu tietokoneohjelmalla, sen sijaan että se kerättäisiin todellisista tapahtumista maailmassa. Niin kuin oikeankin datan kanssa, ei synteettiselle datalle ole yhtä yksiselitteistä, laajasti hyväksyttyä määritelmää. Bottarelli et al. (2022) määrittelevät synteettisen datan siten, että synteettinen data on dataa, joka on luotu tarkoitukseen suunnitellun matemaattisen mallin tai algoritmin avulla, tavoitteenaan ratkaista yksi tai useampi datatieteen tehtävä. Synteettisen datan määritelmissä toistuvat kuitenkin seuraavat keskeiset piirteet: synteettinen data luodaan algoritmin avulla eikä sitä

hankita suoraan mittauksella, sen luominen perustuu matemaattiseen tai tilastolliseen malliin, sekä se jäljittelee todellista dataa (Birisci et al. 2023, 2).

Bottarelli et al. (2022) mukaan synteettisen datan yleisin käyttötarkoitus on toimia oikean datan korvikkeena ja vastuullisesti käytettynä se voi mahdollistaa oppimisen eri datasettien välillä, kun alkuperäinen data on puutteellista, niukkaa tai vinoutunutta. Synteettisen datan tarkoitus on luoda mahdollisimman realistinen datajoukko säilyttäen alkuperäisen datan vivahteet niin, ettei se paljasta arkaluontoista informaatiota (Boedihardjo, Strohmer & Vershynin 2023). Jotta synteettinen data olisi merkityksellistä, sen täytyy olla samanlaista kuin oikea data, mutta kuitenkin erilaista. Oikean datan korvaamiselle täytyy olla peruste, jotakin mistä synteettinen data pystyy suoriutumaan paremmin kuin alkuperäinen data (Bottarelli et al. 2022). Bottarelli et al. (2022) jaottelevat synteettisen datan ominaisuudet kolmeen peruspilariin: hyödyllisyys (eng. utility), tarkkuus (eng. fidelity) ja yksityisyys (eng. privacy)

Hyödyllisyys

Synteettisen datan hyödyllisyydellä tarkoitetaan sen käyttökelpoisuutta hoitamaan sille tarkoitettu tehtävä (Bottarelli et al. 2022). Päättelyt synteettisestä datasta ovat päteviä vain, jos synteettisen datan luomiseen käytetyt mallit vastaavat malleja, joilla voitaisiin luoda alkuperäinen datajoukko (Dibben, Nowok, Raab, Slavkovic & Snoke 2018). Dibben et al. (2018) mukaan synteettisen datan hyödyllisyyttä arvioidaan usein analyyseilla, jotka vertailevat synteettisen datan tuloksia alkuperäisen datan tuloksiin. Myös Bottarelli et al. (2022) ehdottavat, että mallien suorituskykyjä tulisi vertailla konkreettisilla mittareilla, kuten tarkkuuden ja neliöjuurivirheen tarkastelu. Myös Dibben et al. (2018) mukaan neliöjuurivirheen tarkastelu on hyvä keino hyvä mitata synteettisen datajoukon hyödyllisyyttä. He myös toteavat, että mikäli alkuperäisestä datasta ja synteettisestä datasta tehdyt päätelmät ovat yhteneväisiä, voidaan sanoa synteettisellä datalla olevan korkea hyödyllisyys.

Tarkkuus

Tarkkuus ja hyödyllisyys usein yhdistetään yhdeksi ja samaksi, mutta niiden välillä on kuitenkin ero, joka on hyvä tunnistaa. Tarkkuus mittaa sitä, kuinka lähellä synteettiset datapisteet ovat todellisia datapisteitä, eli kuinka laadukkaita synteettiset datapisteet ovat (Marathe, Mortveit, Swarup, Thorve & Vullikanti 2022). Synteettisen datan tarkkuuden määrittelemiseksi käytetään mittareita, jotka suoraan vertailevat synteettistä datajoukkoa oikeaan datajoukkoon. Korkeammalla tasolla tämä tarkoittaa sitä, kuinka hyvin synteettinen

datajoukko vastaa tilastollisesti alkuperäistä datajoukkoa. (Bottarelli et al. 2022) Marathe et al. (2022) mukaan synteettisen datan tarkkuuteen liittyy tiukasti myös datan moninaisuuden käsite. Synteettisen datan moninaisuus mittaa, kuinka hyvin synteettinen data kattaa oikean datan vaihtelevuuden, eli kuinka hyvin oikean datan luonne ja rakenteet toistuvat synteettisessä datassa (Marathe et al. 2022; Choi, Naeem, Oh, Uh & Yoo 2020). Bottarelli et al. (2022) mukaan korkea tilastollinen yhteneväisyys, eli korkea tarkkuus oikean datasetin ja synteettisen datasetin välillä tarkoittaa, että monet tehtävät mitkä suoritettaisiin oikealla datalla, pystytään suorittamaan synteettisellä datalla.

Yksityisyys

Synteettisen datan yksityisyydellä tarkoitetaan sitä, kuinka paljon se paljastaa tietoa alkuperäisestä datasta, jota käytettiin sen luomiseen (Bottarelli et al. 2022). Synteettisen datan luomiseen käytetyt mallit täytyy kouluttaa oikealla datalla, joten suuri kysymys on, kuinka varmistutaan siitä, ettei synteettinen data paljasta arkaluontoista tietoa oikeasta datasta (Nikolenko 2021, 269). Bottarelli et al. (2022) mukaan perussääntönä on, kun synteettisen datan tarkkuus kasvaa, niin sen yksityisyys laskee. Heidän mukaansa on mahdotonta luoda täysin yksityistä synteettistä dataa, mikä olisi hyödyllistä kaikissa käyttötarkoituksissa. Ratkaisuna voidaan joutua luomaan useita eri synteettisiä datajoukkoja, joista osassa yksityisyys on korkeammalla tasolla ja osassa synteettisen datan tarkkuus on korkeammalla tasolla (Bottarelli et al. 2022).

Synteettisen datan kolme luokkaa

Synteettinen data voidaan kategorisoida kolmeen eri luokkaan, jotka ovat täysin synteettinen data, osittain synteettinen data sekä hybridimuotoinen synteettinen data (Sano 2022). Sano (2022) mukaan täysin synteettisessä datassa kaikki tiedot on luotu keinotekoisesti, eikä luotu datasarja sisällä alkuperäistä dataa yhtään. Puhtaasti synteettinen data tarjoaa erinomaisen yksityisyyden suojan (Mathur, Reiter & Si 2024). Täysin synteettistä dataa tuottava malli tunnistaa todellisen datajoukon tilastollisia ominaisuuksia ja luo näiden perusteella datasarjoja vastaamaan näitä tunnistettuja ominaisuuksia. Täysin synteettistä dataa on myös mahdollista luoda ilman oikeaa dataa. Tällöin synteettisen datan tarjoama yksityisyydensuoja on todella korkea, mutta luotu datajoukko ei välttämättä vastaa todellisen maailman tapahtumia. (Bottarelli et al. 2022)

Osittain synteettisessä datassa vain herkimpien muuttujien tilalle tuodaan synteettistä dataa ja loput alkuperäisestä datasta julkaistaan sellaisenaan (Sano 2022). Taulukkomuotoisessa osittain synteettisessä datassa attribuutit voidaan jakaa kahteen kategoriaan, herkkiin ja ei-herkkiin. Ei-herkkiin arvoihin sovitetaan tilastollinen mallin ja tämän mallin avulla tuotetaan uudet arvot herkille muuttujille. Tämän jälkeen herkät attribuutit korvataan alkuperäisestä datasta näillä uusilla arvoilla ja tuloksena on osittain synteettinen datajoukko. Datan hyödyllisyyden kannalta osittain synteettisellä datalla on etu verrattuna täysin synteettiseen dataan, sillä vain osa alkuperäisen datan arvoista on keinotekoisesti tuotettuja. (Bottarelli et al. 2022) Hybridimuotoista synteettistä dataa saadaan, kun yhdistetään täysin synteettistä ja oikeaa dataa. Datajoukon luomiseen käytetystä mallista riippuen hybrididatajoukko voi edustaa enemmän oikeaa datajoukkoa tai synteettistä datajoukkoa. (Sano 2022) Tähän mennessä onnistuneimmat tapaukset synteettisen datan käyttöön ovat ne, joissa olemassa olevaa datasettiä laajennetaan synteettisen datan avulla (Bottarelli et al. 2022).

3.2 Synteettisen datan riskit ja rajoitteet

Tässä luvussa käsitellään synteettiseen dataan liittyviä riskejä ja rajoitteita, joita kirjallisuudessa esiintyy. Useimmat riskit liittyvät synteettisen datan laatuun, sen luomiseen käytetyn datan laatuun sekä synteettisen datan paljastamiin tietoihin oikeasta datasta. Nämä ovat ongelmia, jotka nousevat esiin lähes kaikessa synteettistä dataa koskevassa kirjallisuudessa. Patel (2024) mukaan yritysten on huolehdittava vastuullisuudesta jo synteettisten datasettien suunnitteluvaiheessa, varmistamalla ihmisen valvonnan synteettisten datajoukkojen suunnittelussa ja toteutuksessa. Hän korostaa vielä, että säännölliset tarkastukset, selkeä dokumentointi ja jatkuva riskienarviointi ovat välttämättömiä yrityksille, jotka käyttävät joko sisäisesti tuotettua, tai ulkopuoliselta palveluntarjoajalta hankittua synteettistä dataa.

Synteettisen datan laatu

Synteettisen datan hyödyntämismahdollisuuksista monet liittyvät koneoppimismallien kouluttamiseen. Koneoppimismallien kouluttamisessa koulutusdatan laatu on tärkeässä roolissa ja ensimmäisissä synteettisen datan sovellutuksissa olikin ongelmana datan heikko laatu. Nykyään synteettisellä datalla haetaan ratkaisuja myös näihin koneoppimismallien datan laatuongelmiin ja synteettinen data voikin olla lupaava teknologia tämän ongelman ratkaisussa. (Birisci et al. 2023, 3–9) Haasteena kuitenkin on varmistua synteettisen datan

laadusta, sillä synteettiset datapisteet eivät ole suoraan oikeasta maailmasta. Myöhemmin tässä luvussa esitellään synteettisen datan luomisprosessia ja tässä prosessissa viimeisenä vaiheena onkin luodun datajoukon laadun varmistaminen.

Mahdollinen puolueellisuus

Yksi yleinen synteettiseen dataan liittyvä rajoite on synteettisen datan mahdollinen puolueellisuus (eng. bias). Birisci et al. (2023, 15–17) sanovat, että suurin osa oikeasta maailmasta kerätystä datasta on jollain tapaa puolueellista. Puolueellisuudella he tarkoittavat, että data voi olla epätarkkaa johtuen virheistä sen keräämisessä tai se voi olla epätarkkaa kattaen vain pienen osan siitä kokonaisuudesta, jota se edustaa, eli jotkin osajoukot näyttäytyvät suurempina tai pienempinä datasetissä kuin ne oikeassa maailmassa ovat. Koneoppimismallit ovat tunnettuja siitä, että ne perivät puolueellisuuksia niiden koulutusdatasta ja koska synteettisen datan luomiseen käytetyt mallit koulutetaan oikean maailman datalla, saattavat nämä oikeassa datassa olevat vinoutumat periytyä synteettiseen datajoukkoon (Bottarelli et al. 2022).

Yksityisyydensuojaongelmat

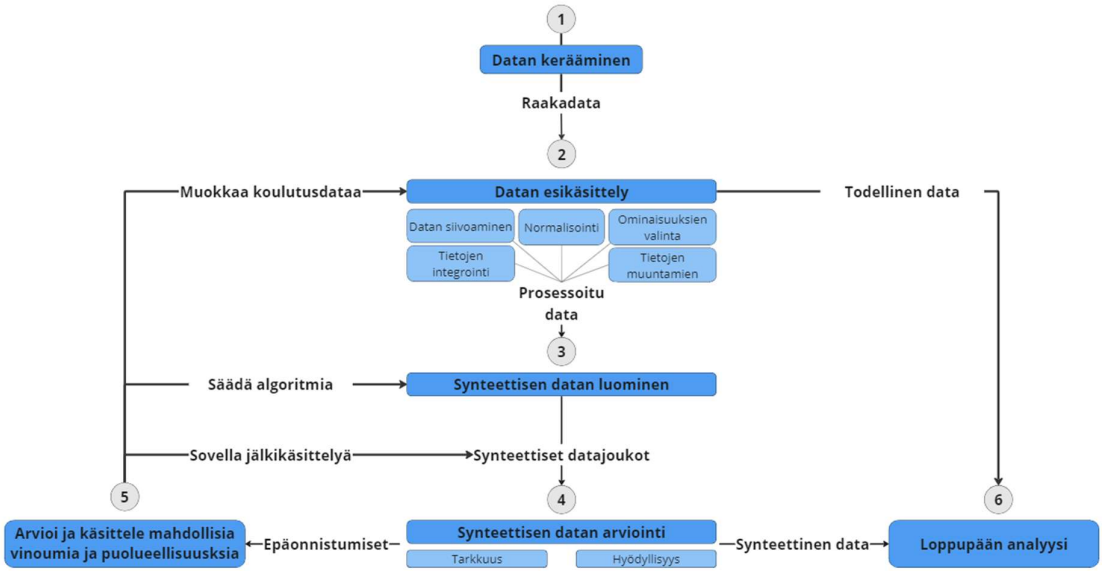
Synteettisen datan yksityisyyteen liittyy myös mahdollisia ongelmia. Kuten kappaleessa 3.1 todettiin, täytyy synteettisen datan kohdalla tehdä valintoja datan yksityisyyden ja tarkkuuden välillä. Kuten Bottarelli et al. (2022) sanovatkin, niin synteettisen datan tarkkuuden kasvaessa sen yksityisyys yleensä heikkenee. Yksityisyyden heikkeneminen tarkoittaa, että synteettisestä datasta pystytään tunnistamaan alkuperäisen datajoukon pisteitä, esimerkiksi tunnistamaan ja yhdistämään henkilöitä alkuperäisestä datajoukosta (Birisci at al. 2023, 7–9). Patel (2024) mukaan suuri osa yrityksistä käyttääkin matalan tarkkuuden dataa välttääkseen mahdolliset riskit liittyen datan yksityisyysongelmiin. Hänen mukaansa yleisimmät kaupalliset synteettisen datan tarjoajat tarjoavatkin strukturoitua matalan tarkkuuden synteettistä data, jonka he mainostavat olevan ”riskivapaata”.

3.3 Synteettisen datan tuottaminen

Synteettisen datan luominen on vielä kehittyvä tutkimusaihe ja vielä puuttuu systemaattisia viitekehyksiä, jotka mahdollistavat tämän teknologian turvallisen ja vastuullisen käytön (Bottarelli et al. 2022) Kuten perinteistäkin dataa, niin myös synteettistä dataa on olemassa erilaista. Birisci et al. (2023, 203–205) mukaan yleiset muodot mitä synteettistä dataa

tuotetaan ovat tekstimuotoinen data, mediamuotoinen data sisältäen kuvat ja videot, sekä taulukkomuotoinen data.

Birisci et al. (2023, 23) jakavat synteettisen datan luomisprosessin kolmeen eri vaiheeseen, joista ensimmäisessä vaiheessa todellisesta datasta luodaan malli, jota käytetään toisessa vaiheessa synteettisen datan luomiseen. Kolmannessa vaiheessa synteettisesti luotua datajoukkoa verrataan alkuperäiseen oikeaan datajoukkoon, jotta varmistutaan niiden tilastollisesta samankaltaisuudesta. Buggineni, Camelio & Chen (2024) ovat kehittäneet kuusivaiheisen prosessimallin synteettisen datan luomiseen, mikä karkealla tasolla mukailee tätä Birisci et al. (2023, 23) kehittämää prosessikehystä. Buggini et al (2024) luomaa prosessimallia havainnollistetaan kuvassa 8. Ensimmäisessä vaiheessa kerätään dataa halutusta prosessista tai toimintaympäristöstä. Teollisessa ympäristössä datan keräämistä on käsitelty tarkemmin kappaleessa 2.1 . Prosessin toisessa vaiheessa kerätty data esikäsitellään, jotta varmistutaan sen sopivuudesta synteettisen datan luomiseen. Bottarelli et al. (2022) mukaan tuotetun synteettisen datajoukon laatu riippuu pitkälti koulutusdatan laadusta, mitä käytettiin kouluttamaan tekoälymalli, joka luo synteettisen datajoukon. Mikäli koulutusdatassa on paljon vinoumia, voivat nämä vinoumat periytyä helposti myös tuotettuun synteettiseen datajoukkoon (Bottarelli et al. 2022).



Kuva 8: Synteettisen datan luomisprosessi (mukaillen Buggineni et al. 2024)

Prosessin kolmannessa vaiheessa kehitetään tekoälymalli luomaan synteettistä dataa kerätyn oikean datan pohjalta. Seuraavissa kahdessa kappaleessa käydään hieman tarkemmin läpi näitä tekoälymalleja.

Tutkimus generatiivisista malleista, joita käytetään synteettisen datan luomiseen, on räjähtänyt ja suosituin lähestymistapa tällä hetkellä ovat niin sanotut GAN- mallit, suomeksi generatiiviset kilpailevat neuroverkot (Bottarelli et al. 2022). Nämä eri GAN-mallit sopivat erityyppisten synteettisten datajoukkojen luomiseen, kuten kuvien, videoiden, tekstin ja taulukkomuotoisen datan (Birisci et al. 2023, 61-71). Tämän tyyppisiä tekoälymalleja on lukuisia erilaisia, esimerkiksi WGAN, MMD-GAN, CGAN ja DCGAN (Bottarelli et al. 2022; Birisci et al. 2023, 61–72), mutta tämän työn laajuuden puitteissa on hyvä tutkia näitä malleja vain hyvin yleisellä tasolla. Bottarelli et al. (2022) mukaan ihmiset eivät täysin itsekään ymmärrä näiden neuroverkkojen tarkkaa toimintaa, joten aiheeseen on hyvä suhtautua varovaisuudella. Synteettisen datan luominen ei rajoitu pelkästään generatiivisiin kilpaileviin neuroverkkoihin, vaan vastaavia malleja löytyy lukuisia muitakin. Tässä työssä keskitytään kuitenkin GAN-malleihin niiden suosion takia.

Generatiiviset kilpailevat verkostot (GANs) koostuvat kahdesta neuroverkosta, jotka ovat generaattori sekä diskriminaattori. Nämä verkot kilpailevat keskenään ja yrittävät siten huijata toisiaan. Ensimmäinen neuroverkon eli generaattorin tehtävä on tuottaa dataa, joka on jakautunut tilastollisesti tietyllä tapaa. Toinen neuroverkko eli diskriminaattori yrittää tunnistaa näistä generaattorin luomista datapisteistä ja alkuperäisestä datasta, mitkä ovat oikeita ja mitkä ovat generaattorin luomia. (Birisci et al. 2023, 61) Tätä prosessia toistettaessa molemmat verkot kehittyvät, generaattori oppii luomaan yhtä aidompia datapisteitä ja diskriminaattori oppii tunnistamaan yhä paremmin mitkä datapisteistä ovat generaattorin luomia ja mitkä ovat oikeita datapisteitä. Tässä prosessissa generatiivinen kilpaileva neuroverkko oppii koko ajan paremmaksi, minkä tuloksena malli voi luoda hyvin realistista dataa. (Bottarelli et al. 2023) Birisci et al. (2023, 203–204) mukaan generatiivisia kilpailevia neuroverkkoja voidaan käyttää luomaan dataa, jota ei ole saatavilla, sekä niitä voidaan käyttää suojaamaan todellisen datan yksityisyyttä. Malleja voidaan käyttää luomaan synteettinen datakaksonen, joka jäljittelee todellista dataa, mutta ei sisällä arkaluontoista tietoa. Näiden mallien käyttöön liittyy myös haasteita. Generatiiviset kilpailevat neuroverkot voivat olla vaikeita kouluttaa, ne voivat olla hitaita tuottamaan synteettistä dataa, sekä ne voivat luoda virheellisesti dataa, joka ei vastaa alkuperäistä datajoukkoa. (Birisci et al. 2023, 203–204) Patel (2022)

muistuttaakin, että synteettisen datan luojien tulisi varmistua synteettisen datan laadusta, sekä heidän tulisi pystyä kuvailemaan kuinka se on luotu ja miten datajoukko voitaisiin luoda uudestaan, sekä miten se eroaa alkuperäisestä datajoukosta.

Buggineni et al. (2024) luoman prosessimallin neljännessä vaiheessa arvioidaan synteettistä dataa sen hyödyllisyyden ja tarkkuuden perusteella, joita käsiteltiin jo aikaisemmin kappaleessa 3.1 . Buggineni et al. (2024) prosessimallin viidennessä vaiheessa arvioidaan synteettisen datajoukon mahdollisia vinoutumia sekä puolueellisuuksia. Heidän mukaansa näitä vinoumia ja puolueellisuuksia voidaan korjata kolmella eri tavalla. Ensimmäinen vaihtoehto on muokata koulutusdataa ja luoda synteettinen datajoukko uudelleen. Esimerkiksi jos synteettinen data on vääristynyt tiettyyn suuntaan, voi koulutusdatan laajentaminen auttaa vähentämään tätä vinoumaa. Toinen vaihtoehto on synteettisen datan luomiseen käytetyn mallin muokkaaminen. Jos esimerkiksi synteettisen datajoukon uskollisuus on riittämätön, voidaan mallia säätää sisällyttämään alkuperäisen datan piirteitä enemmän. (Buggineni et al. 2024) Säätämällä mallia esimerkiksi tähän suuntaan, on hyvä palauttaa mieleen aikaisemmin käsitellyt yhteydet synteettisen datan hyödyllisyyden, tarkkuuden ja yksityisyyden välillä. Buggineni et al. (2024) ehdottavat kolmanneksi ratkaisuksi vinoumien ja puolueellisuuksien korjaamiseksi synteettisen datan jälkikäsittelyä. Tässä vaihtoehdossa synteettistä datajoukkoa muokataan tunnistettujen ongelmien pohjalta, mutta datajoukkoa ei luoda uudestaan mallin avulla.

Prosessimallin kuudennessa vaiheessa luotua korkealaatuista synteettistä datajoukkoa hyödynnetään organisaation suunnittelemalla tavalla. Buggineni et al. (2024) mukaan synteettisen datan luominen on iteratiivinen prosessi, jossa saatua datajoukkoa testataan ja arvioidaan jatkuvasti ja näiden testausten perusteella prosessia kehitetään ja säädetään.

Buggineni et al. (2024) luomassa prosessimallissa voidaan huomata paljon samankaltaisuuksia luvussa 2.2 esitettyyn CRISP-DM prosessimalliin. CRISP-DM prosessimallin kolme ensimmäistä vaihetta ovat liiketoiminnan ymmärtäminen, datan ymmärtäminen sekä datan valmistelu. Näissä kolmessa vaiheessa on monia yhtymäkohtia Buggineni et al. (2024) luomaan synteettisen datan prosessimalliin. Kun synteettistä dataa luodaan hyödynnettäväksi liiketoiminnan tueksi, on tärkeä pitää mielessä datan lopullinen käyttötarkoitus ja miettiä kuinka organisaatio kokonaisuudessaan hyötyy tästä prosessista. Tämän lisäksi molempien viitekehyksien yhtenä kohtana on datan käsittely ennen sen hyödyntämistä. Buggineni et al. (2024) mukaan synteettisen datan yhtenä vahvuutena voidaankin pitää sen tuomaa mahdollisuutta

vähentää datan esikäsittelyyn kulutettua aikaa. Niin kuin synteettisen datan luomisprosessi, on myös CRISP-DM iteratiivinen prosessi, jossa tarkoituksena on palata aiempaan vaiheeseen aina kun huomataan mahdollisuus kehittää prosessia.

Alla olevaan kuvaan 9 on koottu yhteenvedonomaisesti tässä luvussa käsiteltyjä synteettiseen dataan liittyviä ominaisuuksia. Kuvan ylimmässä osassa kerrataan synteettisen datan vahvuuksia verrattuna oikeaan dataan, keskellä on mainittu synteettisen datan kolme muotoa ja tämän alla on mainittu synteettisen datan ominaisuuksien kolme peruspilaria. Alimpana kuvassa on mainittu vielä synteettiseen dataan liittyviä epävarmuustekijöitä, joita on korostettu punaisella värillä.

Kuva 9: Synteettisen datan ominaisuuksien yhteenveto

4 Synteettisen datan sovellutukset teollisuusyrityksissä

Tässä luvussa käsitellään synteettisen datan hyödyntämistä teollisuusyrityksissä. Luvussa 2 keskityttiin hahmottamaan kuinka teollinen massadata muuttaa teollisuusyritysten liiketoimintaa ja mitä mahdollisuuksia ja haasteita datapohjaiseen liiketoimintaan voi liittyä. Luvun perusteella voidaan sanoa, että teollisuusyrityksessä dataan liittyviä ongelmakohtia ovat datan kerääminen ja sen käsittelyyn ja valmisteluun kulunut aika, datan huonoon laatuun liittyvät ongelmat sekä datan omistajuuteen liittyvät haasteet. Teollisuusyritykset voivat omistaa tuotantolaitteisiin liittyvää laitedataa, mutta heillä ei ole pääsyä asiakasyrityksen prosessidataan, mikä olisi erittäin hyödyllistä omien tuotteiden ja palveluiden kehittämisessä. Kamradt et al. (2024, 15) sanovat, että teollisen massadatan aiheuttamiin haasteisiin ei ole olemassa yksittäisiä yksiselitteisiä ratkaisuja, vaan teknologioita ja tekniikoita täytyy soveltaa tapauskokohtaisesti. Patel (2024) mukaan synteettisen datan tuomat hyödyt organisaatioille ovat käytännössä rajattomat.

4.1 Synteettinen data käytännössä

Nikolenko (2021, 2) kuvaa koneoppimisprojektin rakenteen yksinkertaisimmillaan kuvassa 10. Hänen mukaansa koneoppisprojektit jonkin tietyn ongelman ratkaisemiseksi koostuvat neljästä vaiheesta, joista ensimmäinen on ongelmaan liittyvän datan kerääminen, toinen on datan merkitseminen, eli datapisteiden tunnistaminen ja niille merkityksen antaminen. Kolmantena vaiheena on koneoppismallien koulutus ja niiden testaus ja neljäntenä vaiheena on mallin käyttöönotto. Koneoppimiseen liittyvästä kirjallisuudesta suurin osa käsittelee koneoppimismallien kouluttamista ja käyttöönottoa, mutta Nikolenko (2021, 2) mukaan datan kerääminen ja sen merkitseminen ovat vaiheita, jotka voivat viedä jopa 80 % projektin ajasta. Tämän työn puitteissa ei ole tarpeellista tietää koneoppimisprojekteista tämän tarkemmalla tasolla, sillä ne yleensäkin ovat hyvin yksilöllisiä riippuen ongelmasta, jota projektilla pyritään ratkaisemaan.

Kuva 10: Koneoppimisprojektin rakenne (mukaillen Nikolenko 2021, 2)

Synteettinen datan koneoppisprojektissa

Buggineni et al. (2024) esittelevät tutkimuksessaan case-tapauksen, jossa synteettistä dataa hyödynnettiin kouluttamaan satunnaisen metsän (eng. a random forest model) koneoppimismalli tunnistamaan oikean kokoisia ja väärän kokoisia valmistettuja koneen osia. Alkuperäinen datajoukko sisälsi 500 osan tiedot, joista 480 osaa oli oikean kokoisia ja 20 osaa oli väärän kokoisia, eli viallisia. Osia arvioitiin niiden korkeuden, leveyden ja syvyyden perusteella. Ongelma alkuperäisen datajoukon kanssa oli, että datajoukko oli vahvasti puolueellista, sillä suurin osa datapisteistä oli näitä ”oikean” kokoisia osia. Bugginieni et al. (2024) käyttivät kappaleessa 3.3 esiteltyä synteettisen datan tuottamisen viitekehystä laajentamaan tätä oikeaa datajoukkoa niin, että puolet datapisteistä edusti oikean kokoisia osia ja puolet väärän kokoisia. Heidän tutkimuksessaan tätä satunnaisen metsän koneoppismallia koulutettiin ja testattiin kolmella eri tavalla. Ensimmäisessä testissä koulutukseen ja testaukseen käytettiin pelkästään oikeaa dataa, toisessa testissä malli koulutettiin hybridillä datalla, eli oikean datan ja synteettisen datan yhdistelmällä ja testattiin oikealla datalla. Viimeisessä vaihtoehdossa malli koulutettiin ja testattiin hybrideillä datajoukoilla. Case tutkimuksen tuloksena Buggineni et al. (2024) toteavat niiden mallien, jotka koulutettiin hybridillä datalla ja testattiin synteettisellä datalla suoriutuvan paremmin kuin mallien, joissa hyödynnettiin pelkästään oikeaa dataa. Heidän mukaansa aiheesta tarvitaan kuitenkin vielä lisää tutkimusta monimutkaisemmista skenaarioista valmistavat teollisuuden yhteydestä.

Birisci et al. (2023, 13–14) toteavat, että teollisuusyrityksessä koneoppimismalleja voidaan kehittää parantamaan laadunhallintaa, ennustamaan valmistusprosessien tehokkuutta, uusien tuotteiden suorituskykyä sekä ennustamaan tuotteiden kysyntää. Yritys voi näiden kautta saavuttaa kustannussäästöjä ja tehostaa prosessejaan, mikäli se onnistuu tekemään näistä koneoppimismalleista riittävän tarkkoja.

Jia, Lee, Pahren & Thomas (2022) tutkivat synteettisen datan mahdollisuutta parantaa yrityksen tuotteiden laadunhallintaa kouluttamalla ja käyttämällä koneoppimismallia tunnistamaan viallisia tuotteita kuvista. Tämä tutkimus tehtiin koskemaan kuluttajatuotteita valmistavan Procter & Gamblen (P&G) tuotteita. Jia et al. (2022) kokivat haasteeksi sen, että viallisista tuotteista oli saatavilla paljon vähemmän kuvia kuin ei-viallisista tuotteista, sekä näiden kuvien tuottaminen viallisista tuotteista olisi kallista. He loivat synteettisesti kuvia mahdollisista viallisista tuotteista yhdistelemällä ei-viallisten tuotteiden kuviin osia viallisten tuotteiden kuvista. Näin he saivat vähemmän puolueellisen koulutusdatajoukon, joka

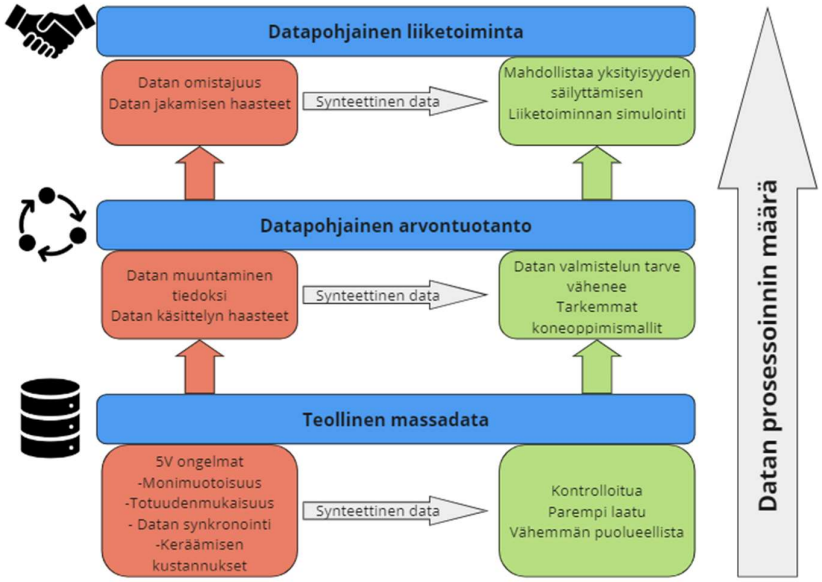
jakaumaltaan edusti paremmin viallisia ja ei-viallisia tuotteita. Tällä hybridillä datajoukolla koulutettu koneoppimismalli suoriutui viallisten tuotteiden tunnistamisesta parhaiten saavuttaen 100 % tarkkuuden. Muut mallit saavuttivat 96 % ja 99 % tarkkuudet, eli mallien välillä suurta eroa ei ollut.

Synteettisen datan potentiaali digitaalisissa kaksosissa

Digitaalinen kaksonen on virtuaalinen ja reaaliaikainen vastine todellisessa maailmassa olevalle fyysiselle prosessille, tuotteelle tai palvelulle. Digitaalinen kaksoisolento mahdollistaa reaaliaikaisen vuorovaikutuksen reaalimaailman ja tämän virtuaalisen maailman välillä (Kamradt et al. 2024, 64). Kamradt et al (2024, 64–67) vievät ajatuksen digitaalisista kaksosista hyvin laajalle tasolle, siten että tulevaisuudessa kokonaisista tehtaista pystyttäisiin tekemään virtuaalisia ja näin ollen testaamaan yrityksen eri prosesseja koko tehtaan tasolla. Aquino, Lopes, Ribeiro, Silveira, Skoogh & Verri (2024) tutkivat digitaalisten kaksosten mahdollisuuksia tuotantoprosesseista ja korostavat kuinka puutteellinen data tuotantoympäristöstä aiheuttaa haasteita näiden virtuaalisten ratkaisuiden kehittämisessä. Vaihtoehdoksi he pohtivat synteettisen datan mahdollisuutta ja tulivat lopputulokseen, että synteettisen datan avulla pystyttiin luomaan ja testaamaan tuotantoprosessien digitaalisia kaksosia.

4.2 Synteettinen data mahdollistajana

Oikein käytettynä synteettinen data tarjoaa siis paljon mahdollisuuksia sen hyödyntämiseen liiketoiminnassa. Soveltaen synteettisen datan ominaisuuksia, on kuvaan 11 täydennetty mahdollisuuksia, joita se tarjoaa teollisuusyritysten kohtaamiin haasteisiin. Näitä ratkaisuja käsiteltiin myös aikaisemmin konkreettisella tasolla kirjallisuudesta löytyvien esimerkkien avulla.



Kuva 11: Synteettisen datan potentiaali teollisuusyrityksessä

Kuvan vasemmassa reunassa on esitetty punaisilla laatikoilla aikaisempien lukujen havaintoja liittyen datan aiheuttamiin haasteisiin teollisuusyrityksessä. Vihreät laatikot kuvan oikeassa reunassa korostavat niitä mahdollisia hyötyjä, joita synteettinen datan voisi tarjota näihin jo aikaisemmin esitettyihin haasteisiin. Kuvan oikeassa reunassa oleva nuoli kuvaa mitä korkeammalle tasolle kuvaajassa siirrytään, niin sitä enemmän dataa täytyy prosessoida hyödyn saavuttamiseksi. Teollisuusyritykset pystyvät siis hyödyntämään synteettistä dataa hyvin monella eri tasolla aina datan keräämisestä, sen hyödyntämiseen uusiin liiketoimintamahdollisuuksiin.

5 Johtopäätökset

Tämän kandidaatintyön tavoitteena oli tutkia kuinka teollisuusyritykset pystyvät hyödyntämään synteettistä dataa liiketoiminnassaan. Työssä tutkittiin, mitä ongelmia datapohjaiseen liiketoimintaan liittyy teollisuusyrityksessä ja kuinka synteettinen data pystyisi tarjoamaan ratkaisuita näihin ongelmiin. Tässä luvussa vastataan työn alussa esitettyihin tutkimuskysymyksiin, sekä esitetään aiheeseen liittyviä jatkotutkimusaiheita. Työn tutkimustuloksien rajoitteena voidaan pitää sitä, että ne pohjautuvat pelkästään teoriakirjallisuudesta löytyvään tietoon, eikä oikeiden yritysten suorittamiin tekoihin.

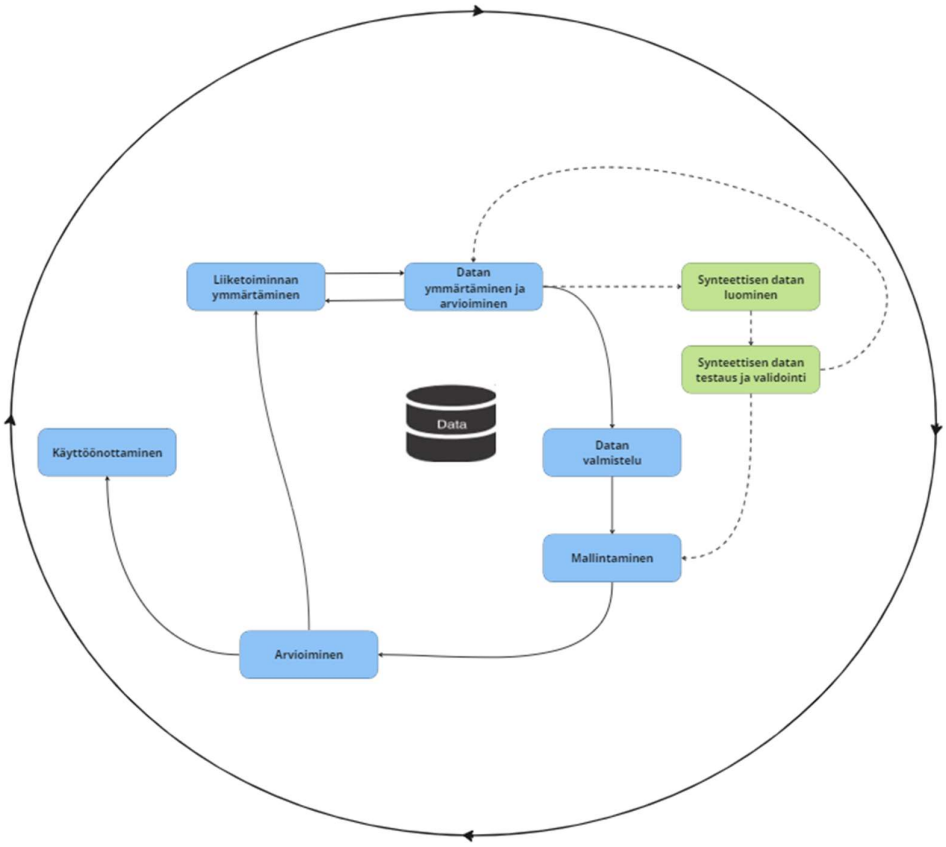
5.1 Vastaukset tutkimuskysymyksiin

Työn pohjalta pystytään vastaamaan jokaiseen työn alussa esitettyyn tutkimuskysymykseen. Työssä esitetty päätutkimuskysymys oli:

- *Kuinka teollisuusyritys voi hyödyntää synteettistä dataa liiketoiminnassaan?*

Kuten ei oikean datan kanssa, ei myöskään synteettiselle datalle ole olemassa vain oikeita viitekehyksiä, miten sitä voidaan hyödyntää. Teollisuusyritysten toiminnasta voidaan tunnistaa monia dataan liittyviä ongelmia, joista Babic et al. (2019) mainitsevat datan monimuotoisuuteen totuudenmukaisuuteen liittyvät haasteet, ja joihin ainakin teoriatasolla synteettinen data voi tarjota ratkaisuja. Työn perusteella voidaan sanoa, että teollisuusyritykselle on monia mahdollisuuksia hyödyntää synteettistä dataa ja tällä hetkellä suosituimmat kohteet liittyvät koneoppimismallien kouluttamiseen. Yrityksen kannalta on kuitenkin tärkeintä tunnistaa oman datastrategiansa heikkoudet ja puutteet ja näiden perusteella arvioida synteettisen datan tarvetta ja käyttötarkoitusta. Synteettistä dataa ei voida pitää hopealuotina yrityksen ongelmien ratkaisussa, vaan on tärkeä integroida sen hyödyntäminen olemassa olevien toimintojen tueksi. Yksi tällainen vaihtoehto on integroida synteettisen datan tuottaminen ja hyödyntäminen osaksi yrityksen datalouhinnan prosessia, esimerkiksi osaksi Hipp & Wirth (2000) kuvailemaa CRISP-DM-prosessia. Kuva 12 on päivitetty versio CRISP-DM-prosessissa, jossa on otettu mukaan synteettinen data täydentämään liiketoiminnasta kerättyä dataa. Prosessin toiseen, datan ymmärtämisen vaiheeseen on lisätty arviointivaihe,

jossa arvioidaan käsillä olevaa datajoukkoa ja sen mahdollisuuksia yksinään täyttää sille asetettu tehtävä. Mikäli arvioinnin lopputulos on, että datajoukko pystyy täyttämään sille asetetun tehtävän, voidaan siirtyä suoraan seuraavaan vaiheeseen datan valmisteluun. Jos arvioinnin lopputulos on, ettei kyseinen datajoukko yksinään suoriudu tehtävästä, voidaan datajoukkoa käyttää luomaan synteettistä dataa. Synteettisen datan luomiseen voidaan hyödyntää kappaleessa 3.3 esiteltyä prosessia. CRISP-DM -mallin toiseksi viimeisessä vaiheessa, jossa arvioidaan luodun mallin soveltuvuutta, tulisi arvioida myös kuinka synteettinen data on mahdollisesti vaikuttanut mallin rakentamiseen ja mallin antamiin tuloksiin.



Kuva 12: Synteettisen datan integrointi CRISP-DM prosessikehykseen

Päätutkimuskysymyksen lisäksi työssä esitettiin kolme osatutkimuskysymystä, joiden pohjalta pystytään vastaamaan työn päätutkimuskysymykseen. Työn ensimmäinen osatutkimuskysymys oli:

- *Kuinka data muuttaa teollisuusyritysten liiketoimintaa ja mitä haasteita tähän liittyy?*

Fleckenstein & Fellows (2018, 15) mukaan teollisuusyritykset pyrkivät keräämään oikeanlaista dataa oikeassa muodossa, mitä ne pyrkivät hyödyntämään omien prosessien kehittämisessä sekä uudenlaisten arvonluontitapojen keksimisessä. COVID-19 pandemia vahvisti olemassa olevia trendejä liittyen yritysten digitalisaatioon, ja massadata analytiikkaan liittyvät teknologiat ovat synnyttäneet laajamittaisen yhteiskunnallisen muutoksen kohti palvelukeskeisempää liiketoimintaa. Teollisuusyritykset ovat vastanneet tähän muutokseen muuttamalla omia liiketoimintamalleja kohti älykkäitä tuotepalvelujärjestelmiä. (Chowdhury, Haftor & Pashkevich 2018; Carillo 2017) Tämä muutos vaatii suuria määriä laadukasta dataa, mihin liittyy myös ongelmia. Teollisuusyritykset keräävät dataa organisaatiossaan monelta eri tasolta, kuten tuotannon prosesseista, logistiikasta, tuotteista sekä asiakkaista. Nämä datamäärät voivat olla valtavia, niiden kerääminen saattaa olla kallista, mutta silti datajoukot voivat olla puutteellisia niille tarkoitettuun tehtävään. Dataprojektiin käytetystä ajasta suurin osa saattaa kulua tämän puutteellisen datajoukon siivoamiseen ja muokkaukseen. Teollisuuden kontekstissa on esitelty monia datalouhinnan prosesseja, joista tässä työssä käytiin tarkemmin läpi CRISP-DM prosessikehystä. Prosessin eri vaiheisiin liittyy haasteita liittyen datan keräämiseen, sen määrään, oikeellisuuteen ja laatuun.

Työn toisen osatutkimuskysymyksen tarkoituksena on avata lukijalle mitä tarkoitetaan synteettisellä datalla, sekä mitä haasteita ja mahdollisuuksia sen käyttöön liittyy. Työn toinen osatutkimuskysymys oli:

- *Mitä on synteettinen data sekä sen hyödyt ja rajoitteet?*

Synteettinen data on keinotekoisesti, esimerkiksi tekoälymallilla tuotettua dataa, jonka tarkoituksena on heijastella oikean datan tilastollisia ominaisuuksia. Synteettiselle datalle ei ole yhtä yksiselitteisesti hyväksyttyä määritelmää, mutta sen eri määritelmissä toistuvat seuraavat kolme piirrettä: synteettistä dataa ei hankita suoraan mittausten avulla, sen luominen perustuu matemaattiseen tai tilastolliseen malliin ja se jäljittelee todellista dataa. Bottarelli et al. (2022) ehdottavat, että synteettistä dataa voidaan yleisesti tarkastella kolmen

ominaisuuden avulla; hyödyllisyys, tarkkuus ja yksityisyys. Hyödyllisyydellä tarkoitetaan sen käyttökelpoisuutta hoitaa sille tarkoitettu tehtävä, tarkkuudella tarkoitetaan synteettisen datan laatua, eli kuinka lähellä synteettiset datapisteet ovat oikean joukon datapisteitä. Yksityisyydellä tarkoitetaan puolestaan sitä, kuinka hyvin synteettisestä datasta voidaan tunnistaa alkuperäisen datajoukon pisteitä.

Synteettisellä datalla on monia hyötyjä verrattuna oikeaan dataan. Sitä voidaan käyttää laajentamaan alkuperäistä datajoukkoa, mikäli oikeaa dataa on saatavilla niukasti. Synteettistä dataa voidaan myös käyttää korjaamaan oikean datajoukon puolueellisuuksia niissä tilanteissa, joissa oikeassa datajoukossa tietyt ryhmät ovat vahvemmin tai heikommin edustettuna kuin toiset. Birisci et al. (2023, 3) myös korostavat, että synteettinen data voi olla paremman laatuista kuin oikea data, sitä voidaan tuottaa juuri sen verran kuin katsotaan tarpeelliseksi, sekä sitä voidaan käyttää parantamaan oikeiden datajoukkojen yksityisyydensuojaa. Synteettisen datan rajoitteet voidaan jakaa kolmeen luokkaan, joista ensimmäinen koskee datan laatua. Synteettiset datapisteet eivät ole oikeasta maailmasta, joten niihin liittyy aina riski, etteivät ne edusta riittävän tarkasti oikean maailman tapahtumia. Toisena ongelmana on synteettisen datan mahdollinen puolueellisuus, mikä tarkoittaa, että oikeassa datassa olevat vinoumat ja puolueellisuudet saattavat periytyä myös synteettiseen datajoukkoon. Kolmantena rajoitteena voidaan nostaa esille synteettisen datan yksityisyyden ongelmat, eli pystytäänkö synteettistä datajoukosta tunnistamaan alkuperäisen datajoukon pisteitä, sekä tasapainoilu yksityisyyden ja tarkkuuden välillä. Työn kolmas osatutkimuskysymys liittyi synteettisen datan hyödyntämismahdollisuuksiin teollisuusyrityksissä:

- *Kuinka teollisuusyritys pystyy luomaan arvoa synteettisen datan avulla?*

Kirjallisuuden perusteella teollisuusyrityksen kontekstissa synteettisen datan hyödyntämismahdollisuudet liittyvät pitkälti koneoppimismallien kouluttamiseen. Jotta tekoälymallit pystyvät tekemään hyviä päätöksiä, täytyy niitä kouluttaa laadukkaalla, monipuolisella ja laajoilla datajoukoilla. Vaikkakin työn yhtenä tuloksena voidaan sanoa, että teollisuusyrityksissä datan määrä on viime vuosina lisääntynyt, niin ei tämä data ole välttämättä laatunsa tai monipuolisuutensa puolesta riittävää tekoälymallien kouluttamiseen. Synteettinen data tarjoaa kustannustehokkaan ja mahdollisesti myös nopeamman vaihtoehdon koulutusdatan luomiseen (Birisci et al. 2023, 13–14). Näiden tekoälymallien avulla teollisuusyritykset pystyvät tehostamaan omia prosessejaan, parantamaan laadunhallintaa, sekä testaamaan uusien tuotteiden suorituskykyä ja ominaisuuksia. Synteettisellä datalla on myös potentiaalia

yrityksen prosessien ja liiketoimintamallien kehittämisessä. Synteettistä dataa voidaan hyödyntää prosessien digitaalisten kaksosten luomisessa ja niiden simuloinnissa, sekä myös esimerkiksi koko tehtaan tasolla tapahtuvassa tuotantoprosessien simuloinnissa.

5.2 Tulevaisuudennäkymät ja jatkotutkimusaiheet

Synteettinen data on tällä hetkellä murrosvaiheessa omaksumisensa suhteen. Sen luomiseen tarkoitettu teknologia on kypsynyt nopeasti, mutta yritykset eivät ole omaksuneet synteettisen datan hyödyntämistä yhtä nopealla tavalla. Tulevaisuudessa on kuitenkin oletettavissa, että organisaatiot alkavat hyödyntämään synteettistä dataa enemmän, siinä tahdissa kuin muutkin tekoälyratkaisut yrityksissä yleistyvät. Teollisuusyritysten näkökulmasta voi tulevaisuudessa olettaa, että synteettistä dataa tullaa yhä laajemmin käyttämään tekoälymallien kouluttamiseen tukemaan ja tehostamaan yrityksen eri prosesseja.

Kirjallisuuskatsauksen pohjalta voidaan sanoa, että synteettisen datan tuottamisesta, sen hyödyistä, riskeistä ja haitoista löytyvä kirjallisuusmateriaali on paljon laajempaa kuin sen soveltamiseen liittyvä materiaali. Synteettisen datan hyödyntämiseen liittyvä kirjallisuus teollisuusyritysten tapauksessa painottuu hyvin vahvasti tekoälymallien kouluttamiseen. Teollisuusyritysten perinteisimpiin data-analytiikan prosesseihin liittyy myös paljon haasteita koskien datan määrää ja laatua. Yksi jatkotutkimusaihe voisi täten koskea kuinka synteettisellä datalla pystytään parantamaan niitä data-analytiikan prosesseja, joihin ei liity tekoälyä. Tässä painopisteenä voisi erityisesti olla data-analytiikkaprosessin tehostaminen ajallisesti sekä prosessin tarkkuuden parantaminen. Toinen suositeltu jatkotutkimusaihe liittyy koneoppimismallien implementoimiseen osaksi yritysten prosesseja sekä niiden saavuttamien hyötyjen arvioimista. Monessa kirjallisuuslähteessä mainittiin, kuinka koulutettua tekoälymallia täytyy testata ja arvioida vielä monimutkaisemmissa ympäristöissä.

Lähteet

Allied Market Research (2022) Synthetic Data Generation Market Size, Share, Competitive Landscape and Trend Analysis Report by Component, by Deployment Mode, by Data Type, by Application, by Industry Vertical : Global Opportunity Analysis and Industry Forecast, 2021-2031. [verkkosivu]. [viitattu 24.3.2024]. Saatavilla: https://www.alliedmarketresearch.com/synthetic-data-generation-market-A31749

Babic, J., Gace, I., Pevec, D., Podobnik, V., Sabolic, M. &Vdovic, H. (2019) Distributed Data Platform for Automotive Industry: A Robust Solution for Tackling Big Challenges of Big Data in Transportation Science. IEEE International Conference on Telecommunications. nro. 15, s. 1–8.

Basanta-Val, P. (2018) An Efficient Industrial Big-Data Engine. IEEE Transactions on Industrial Informatics. Vol. 14, nro. 4, s. 1361–1369.

Bekar, E. T., Nyqvist, P. & Skoogh, A. (2020) An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study. Advances in Mechanical Engineering. Vol. 12, nro. 5

Birisci, E., Celik, S. & Gürsakal, N. (2023) Synthetic Data for Deep Learning. Yhdysvallat. Apress L. P. 220 s.

Blanco, J. M., Illarramendi, A. & Nino, M. (2015) Business understanding, challenges and issues of Big Data Analytics for the servitization of a capital equipment manufacturer. Teoksessa: 2015 IEEE International Conference on Big Data (Big Data). 2015 IEEE. s. 1368–1377.

Boedihardjo, M., Strohmer, T. & Vershynin, R. (2023) Privacy of Synthetic Data: A Statistical Framework. IEEE Transactions on Information Theory. Vol. 69, nro. 1, s. 520–527.

Bottarelli, M., Cohen, S., Cherubin, G., Houssiau, F., Jordon, J., Maple, C., Szpruch, L. & Weller, A. (2022) Synthetic Data -- what, why and how?. Ithaca.

Branson, J., Harbron, C., James, S. & Sundler, M. (2021) Synthetic data use: exploring use cases to optimise data utility. Discover Artificial Intelligence. Vol. 1, nro. 1, s. 1–13.

Buggineni, V., Camelio, J. & Chen, C. (2024) Enhancing manufacturing operations with synthetic data: a systematic framework for data generation, accuracy and utility. Frontiers in Manufacturing Technology. Vol. 4, s. 1-12.

Calic, G. & Ghasemaghaei, M. (2019) Does big data enhance firm innovation competency? The mediating role of data-driven insights. Journal of Business Research. Vol. 104, s. 69–84.

Carillo, K. D. A. (2017) Let's stop trying to be "sexy" – preparing managers for the (big) data-driven business era. Business Process Management Journal. Vol. 23, nro. 3, s. 598–622.

Choi, Y., Naeem, M. F., Oh, S. J., Uh, Y. & Yoo, J. (2020) Reliable Fidelity and Diversity Metrics for Generative Models. Ithaca

Chowdhury, S., Haftor, D. & Pashkevich, N. (2018) Smart product-service systems (Smart PSS) in industrial firms: a literature review. Procedia CIRP. Vol. 73, s. 26–31.

CloudTDMS. (2024) The Rise of Synthetic Data: A look into the Future. [verkkoaineisto]. [viitattu 15.2.2024]. Saatavilla: https://www.cloudtdms.com/explore/synthetic-data-future/

Contreras-Ochando, L., Ferri, C., Flack, P., Hernandez-Orallo, J., Kull, J. Lachiche, M., Martinez-Plumed, F. & Ramırez-Quintana, M. (2021) CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering. Vol. 33, nro. 8, s. 3048–3061.

Cortez, P., Iliadis, L., Macintyre, J. & Maglogiannis, I. (2022) Production Time Prediction for Contract Manufacturing Industries Using Automated Machine Learning. Teoksessa: Artificial Intelligence Applications and Innovations. Switzerland: Springer International Publishing AG. s. 262–273.

Demchenko, Y., Grosso, P. & Laat, C. (2013) Addressing big data issues in Scientific Data Infrastructure. Teoksessa: 2013 International Conference on Collaboration Technologies and Systems (CTS). 2013 IEEE. s. 48–55.

Dibben, C., Nowok, B., Raab, G., Slavkovic, A. & Snoke, J. (2018) General and specific utility measures for synthetic data. Journal of the Royal Statistical Society. Series A, Statistics in Society. Vol. 181, nro. 3, s. 663–688.

Eloranta, V., Paavola, M. & Seppänen, M. (2021) Datapohjaisen arvonluonnin strategiset vaihtoehdot. Helsinki. Työ- ja elinkeinoministeriö. nro. 3.

Feng, N., Guo, F. & Tu, S. (2024) Big-data analytics capability, value creation process, and collaboration innovation quality in manufacturing enterprises: A knowledge-based view. Computers & Industrial Engineering. Vol. 187

Fink, A. A., Guzzo, R., King, E., Landis, R. S. & Tonidandel, S. (2015) Big Data Recommendations for Industrial–Organizational Psychology. Industrial and Organizational Psychology. Vol. 8, nro. 4, s. 491–508.

Fleckenstein, M. & Fellows, L. (2018) Modern Data Strategy. Sveitsi. Cham: Springer International Publishing. 263 s.

Grand View Research (2024) Synthetic Data Generation Market Size, Share & Trends Analysis Report By Data Type, By Modeling Type, By Offering, By Application, By End-use, By Region, And Segment Forecasts, 2023 – 2030. [verkkosivu]. [viitattu 24.3.2024]. Saatavilla: https://www.grandviewresearch.com/industry-analysis/synthetic-data-generation-market-report

Grubic, T. & Peppard, J. (2016) Servitized manufacturing firms competing through remote monitoring technology. Journal of Manufacturing Technology Management. Vol. 27, nro. 2, s. 154–184.

Hercik, R. & Svoboda, R. (2023) Collecting and Pre-Processing Data for Industry 4.0 Implementation Using Hydraulic Press. Data (Basel). Vol. 8, nro. 4, s. 1-14.

Hipp, J. & Wirth, R. (2000) CRISP-DM: Towards a Standard Process Model for Data Mining. Teoksessa: Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining. Vol. 1, s. 29-39.

ISO/IEC 2382. (2015). Information technology Vocabulary. International Organization for Standardization. [verkkodokumentti]. [viitattu 21.2. 2024]. Saatavilla: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:2382:ed-1:v2:en

Jia, X., Lee, J., Pahren, L. & Thomas, P. (2022) A Novel Method in Intelligent Synthetic Data Creation for Machine Learning-based Manufacturing Quality Control. IFAC-Papers Online. Vol. 55, nro. 19, s. 73-78.

Kamal, A. & Pahuja, L. (2023) EnLEFD-DM : Ensemble Learning based Ethereum Fraud Detection using CRISP-DM framework. Expert Systems. Vol. 40, nro. 9.

Kamradt, M., Nassif, J. & Tekli, J. (2024) Synthetic Data Revolutionizing the Industrial Metaverse. 1. p. Cham. Springer Nature Switzerland. 186 s.

Keith, G. (2007) Principles of data management facilitating information sharing. 1. p. Swindon. BCS Learning & Development Limited. 223 s.

Kundu, S. & McKay, A. (2014) A representation scheme for digital product service system definitions. Advanced Engineering Informatics. Vol. 28, nro. 4, s. 479–498.

Li, L., Lu, L., Meng, Y. & Yan, J. (2017) Industrial Big Data in an Industry 4.0 Environment: Challenges, Schemes, and Applications for Predictive Maintenance. IEEE. Vol. 5, s. 23484–23491.

Luttgens, D., Mertens, A., Nitsch, V., Piller, F. T., Putz, S. & Van Dyck, M. (2022) Forecasting next generation manufacturing : digital shadows, human-machine collaboration, and data-driven business models. 1. p. Cham. Springer Nature Switzerland. 158 s.

Mathur, S., Reiter, J. & Si, Y. (2024) Fully Synthetic Data for Complex Surveys. Ithaca.

Marathe, M. V., Mortveit, H. Swarup, S., Thorve, S. & Vullikanti, A. (2022) Fidelity and diversity metrics for validating hierarchical synthetic data: Application to residential energy demand. Teoksessa: 2022 IEEE International Conference on Big Data (Big Data). 2022 IEEE. s. 1377–1382.

Nikolenko, S. I. (2021) Synthetic data for deep learning. Cham. Springer Nature Switzerland. 354 s.

Niu, Z., Song, W., Tang, Y., Wang, Q., Zheng, j. & Zheng, P. (2022) Risk evaluation for industrial smart product-service systems: An integrated method considering failure mode correlations. Advanced Engineering Informatics. Vol. 54.

Patel, P. (2024) Synthetic data. Business Information Review.

Sah, P. (2022) Defining Enterprise Data and Analytics Strategy. 1. p. Singapore. Springer Nature Singapore. 174 s.

Salminen, A. (2011) Mikä kirjallisuuskatsaus? Johdatus kirjallisuuskatsauksen tyyppeihin ja hallintotieteellisiin sovelluksiin. Vaasa. Vaasan yliopisto. 44 s.

Sano, N. (2022) Utility and Risk Evaluation of Synthetic Data by Orthogonal Transformation. The Review of Socionetwork Strategies. Vol. 16, nro 1, s. 71–79.