

Data-analytiikan hyödyntäminen valmistavan teollisuuden yrityksissä

Mika Rantonen, Janne Alatalo,
Tarja Ajo & Eppu Heilimo (toim.)



Euroopan unionin
osarahoittama

jamk

Jyväskylän
ammattikorkeakoulu

**Data-analytiikan
hyödyntäminen valmistavan
teollisuuden yrityksissä**

JYVÄSKYLÄN AMMATTIKORKEAKOULU

JULKAISUJA 346

Kirjoittajat

Mika Rantonen

Yliopettaja ja projektipäällikkö
Jyväskylän ammattikorkeakoulu

Janne Alatalo

Asiantuntija
Jyväskylän ammattikorkeakoulu

Tarja Ajo

Asiantuntija
Jyväskylän ammattikorkeakoulu

Eppu Heilimo

Asiantuntija
Jyväskylän ammattikorkeakoulu

Vesa Vertainen

Asiantuntija
Jyväskylän ammattikorkeakoulu

Mika Valkama

Asiantuntija
Seinäjoen ammattikorkeakoulu

Saku Kaarlejärvi

Asiantuntija
Seinäjoen ammattikorkeakoulu

Juha-Matti Arola

Senior Project Manager
Seinäjoen ammattikorkeakoulu

Mikko Pakkasela

Tutkimuspäällikkö
Satakunnan ammattikorkeakoulu

Jari Kyngäs

Erikoistutkija
Satakunnan ammattikorkeakoulu

Juha Stenfors

Erityisasiantuntija
Satakunnan ammattikorkeakoulu

Mika Rantonen
Janne Alatalo
Tarja Ajo
Eppu Heilimo
(toim.)

Data-analytiikan hyödyntäminen valmistavan teollisuuden yrityksissä



ENSTO

LS CARGO
LOGISTICS



VALTRA

SEAMK
Seinäjoen ammattikorkeakoulu
Seinäjoki University of Applied Sciences



Euroopan unionin
osarahoittama



Elinkeino-, liikenne- ja
ympäristökeskus

jamk | Jyväskylän
ammattikorkeakoulu

JYVÄSKYLÄN AMMATTIKORKEAKOULU
JULKAISUJA 346

Sarjan vastaava toimittaja • Jari Parkkinen

Data-analytiikan hyödyntäminen valmistavan teollisuuden yrityksissä

Mika Rantonen, Janne Alatalo, Tarja Ajo & Eppu Heilimo (toim.)

Kannen kuva, taitto ja kuvitus • Heli Sutinen, Jamk
Paino • Punamusta Oy • 2025

ISBN 978-951-830-791-7 (Painettu)

ISBN 978-951-830-792-4 (PDF)

ISSN-L 1456-2332

Kustantaja

Jyväskylän ammattikorkeakoulu
www.jamk.fi/julkaisut
[sarjajulkaisut\(a\)jamk.fi](mailto:sarjajulkaisut(a)jamk.fi)

©2025

Tekijät ja Jyväskylän ammattikorkeakoulu



Julkaisu noudattaa
Creative Commons Nimeä (CC BY 4.0) -lisenssiä.

SISÄLTÖ

KIRJOITTAJAT	8
TIIVISTELMÄ.....	11
ABSTRACT	12
LUKIJALLE	13

Mika Rantonen, Janne Alatalo, Tarja Ajo & Eppu Heilimo,
Jyväskylän ammattikorkeakoulu

POHJATIEDOT KUNTOON - DATATALOUDEN PERUSTEITA

DATATALOUS JA PK-YRITYKSEN DATASTRATEGIA.....	16
--	-----------

Mikko Pakkasela, Satakunnan ammattikorkeakoulu

DATAN HYÖDYNTÄMISEN HAASTEET	20
---	-----------

Mika Rantonen, Jyväskylän ammattikorkeakoulu

DATA-AVARUUDET – KOHTI DATAN HYÖDYNTÄMISEN SEURAAVAA TASOA	24
---	-----------

Tarja Ajo, Jyväskylän ammattikorkeakoulu

DATA-ANALYTIIKAN JAOTTELUA.....	32
--	-----------

Tarja Ajo & Mika Rantonen,
Jyväskylän ammattikorkeakoulu

KÄDET SAVEEN - DATAN HYÖDYNTÄMINEN KÄYTÄNNÖSSÄ

TIEDON JALOSTAMINEN RAPORTOITAVAAAN MUOTOON	39
--	-----------

Juha Stenfors, Satakunnan ammattikorkeakoulu

DATAN VARASTOINNIN TERMIT HALTUUN	44
--	-----------

Mika Rantonen, Jyväskylän ammattikorkeakoulu

DATAN VISUALISOINNIN SELKEYTTÄMINEN SADEPILVIKUVAAJILLA...50	
---	--

Tarja Ajo & Janne Alatalo, Jyväskylän ammattikorkeakoulu

**TEKOÄLY KÄYTTÖÖN PK-YRITYKSISSÄ – KYBERTURVALLISUUDESTA
TINKIMÄTTÄ.....55**

Vesa Vertainen, Jyväskylän ammattikorkeakoulu

KOHTI ÄLYKKÄÄMPÄÄ DATANKERUUTA TEKOÄLYN TUELLA60

Saku Kaarlejärvi & Juha-Matti Arola, Seinäjoen ammattikorkeakoulu

**GENERATIIVISEN TEKOÄLYN MAHDOLLISUUDET DATAN
HYÖDYNTÄMISESSÄ66**

Janne Alatalo, Jyväskylän ammattikorkeakoulu

OPITAAN ESIMERKISTÄ - KÄYTTÖTAPAUKSIA TEOLLISUUDESTA

RFID-TEKNOLOGIA METALLIALAN YRITYKSISSÄ 72

Mika Valkama, Seinäjoen ammattikorkeakoulu

**MATEMAATTINEN MALLINTAMINEN: JUNAN OVIEN AVAUTUMINEN JA
SULKEUTUMINEN79**

Jari Kyngäs, Satakunnan ammattikorkeakoulu

**KULJETUSPÄÄSTÖJEN MITTAAMINEN KILPAILUETUNA:
DATALÄHTÖINEN LÄHESTYMISTAPA.....82**

Eppu Heilimo, Jyväskylän ammattikorkeakoulu

KIRJOITTAJAT



Mika Rantonen

Yliopettaja ja projektipäällikkö, Jyväskylän ammattikorkeakoulu
| mika.rantonen@jamk.fi

Mika Rantonen toimii Jamkin IT-instituutissa data-analytiikan ja tekoälyyn liittyvissä projekteissa projektipäällikkönä. Koulutukseltaan Rantonen on tekniikan lisensiaatti sekä filosofian maisteri. Lisäksi Rantonen toimii data-analytiikan ja tekoälyn yliopettajana ja kyseisen YAMK-koulutusohjelman tukivastaavana.



Janne Alatalo

Asiantuntija, Jyväskylän ammattikorkeakoulu
| janne.alatalo@jamk.fi

Janne Alatalo toimii asiantuntijana Jamkin IT-instituutissa. Hän on valmistunut tekniikan alalta insinööriksi (AMK) vuonna 2014 ja insinööriksi (YAMK) vuonna 2023. Hän on työskennellyt Jamkin hanketoiminnan parissa projekti-insinöörinä ja asiantuntijana jo 10 vuotta. Hänen asiantuntijuusalueensa on ohjelmistotekniikka, sovellettu tekoäly ja data-analytiikka sekä sovellettu kyberturvallisuus.



Tarja Ajo

Asiantuntija, Jyväskylän ammattikorkeakoulu
| tarja.ajo@jamk.fi

Tarja Ajo toimii Jamkin IT-instituutissa data-analytiikan ja tekoälyn asiantuntijana, sekä opinnäytetöiden ohjaajana. Koulutukseltaan Ajo on filosofian maisteri (matematiikan aineenopettajakoulutus) ja tieto- ja viestintätekniikan insinööri (AMK).



Eppu Heilimo

Asiantuntija, Jyväskylän ammattikorkeakoulu
| eppu.heilimo@jamk.fi

Eppu Heilimo toimii Jamkin IT-instituutissa VauhtiData-, KE-MAR-, ja Resilmesh-hankkeissa data-analytiikan ja tekoälyn asiantuntijana. Koulutukseltaan hän on ohjelmistotekniikan insinööri (AMK).



Vesa Vertainen

Asiantuntija, Jyväskylän ammattikorkeakoulu
| vesa.vertainen@jamk.fi

Vesa Vertainen työskentelee Jamkin IT-instituutissa kyberturvallisuuden TKI-projekteissa asiantuntijana, sekä Kyberosaake-hankeessa projektipäällikkönä. Hän on koulutukseltaan tieto- ja viestintäteknikan insinööri (AMK) sekä automaatio-tekniikko. Parhaillaan hän työstää myös kyberturvallisuuden YAMK-tutkintoa.



Mika Valkama

Asiantuntija, Seinäjoen ammattikorkeakoulu
| mika.valkama@seamk.fi

Mika Valkama on ohjelmistokehitykseen, automaatiojärjestelmiin ja teollisuuden digitalisaatioon erikoistunut asiantuntija. Hänellä on pitkä kokemus palveluiden kehittamisestä, esimerkiksi suunnittelusta, käyttöönotosta ja ongelmanratkaisusta. Lisäksi Valkama on osaava mobiili- ja websovellusten kehittäjä.



Saku Kaarlejärvi

Asiantuntija, Seinäjoen ammattikorkeakoulu
| saku.kaarlejarvi@seamk.fi

Saku Kaarlejärvi on tietotekniikan insinööri, joka on erikoistunut ohjelmointiin, tekoälyyn ja sulautettuihin järjestelmiin. Hänellä on kokemusta tuotannonohjauksen mittauksessa käytetyistä ratkaisuista, Power BI -raporttien rakentamisesta sekä älyteknologioiden mahdollisuuksien tutkimisesta hyvinvointi- ja terveysaloilla.



Juha-Matti Arola

Senior Project Manager, Seinäjoen ammattikorkeakoulu
| juha-matti.arola@seamk.fi

Juha-Matti Arola on kokenut projektipäällikkö, joka on erikoistunut EU-rahoitteisiin ESR-, JTF- ja EAKR-hankkeisiin, virtuaalisten työ- ja oppimisympäristöjen kehittämiseen sekä tiedolla johtamisen hyödyntämiseen. Hän yhdistää vahvan teknologisen ja pedagogisen osaamisen tukemaan pk-yritysten ja aluekehityksen innovaatioita.



Mikko Pakkasela

Tutkimuspäällikkö, Satakunnan ammattikorkeakoulu
| mikko.pakkasela@samk.fi

Mikko Pakkasela toimii Satakunnan ammattikorkeakoulun Tiedolla johtamisen keskuksen tutkimuspäällikkönä ja on koulutukseltaan tuotantotalouden diplomi-insinööri. Hänellä on taustalla laaja osaaminen teollisuuden eri aloilta, kuten prosessi-, ohjelmisto- ja konepajateollisuudesta sekä viestinnästä. Hän on kiinnostunut yritysten liiketoiminnan ja tiedolla johtamisen kehittämisestä erityisesti pk-yrityksissä.



Jari Kyngäs

Erikoistutkija, Satakunnan ammattikorkeakoulu
| jari.kyngas@samk.fi

Jari Kyngäs toimii erikoistutkijana SAMKin Tiedolla johtamisen keskuksessa. Kyngäs on filosofian tohtori, pääaineena tietojenkäsittelytiede ja sivuaineena tilastotiede. Hän on ollut SAM-Kissa 26 vuotta, joista 24 opettajana ja viimeiset 2 erikoistutkijana.



Juha Stenfors

Erityisasiantuntija, Satakunnan ammattikorkeakoulu
| juha.stenfors@samk.fi

Juha Stenfors toimii asiantuntijana Samkin Tiedolla johtamisen keskuksessa ja kehityspalvelussa. Stenfors on valmistunut tietojenkäsittelyn tradenomiksi (AMK) vuonna 1995 ja tradenomiksi (YAMK) vuonna 2013. Hän on työskennellyt Samkin hanketoiminnan parissa asiantuntijana kolme vuotta, hallinnollisissa tehtävissä noin yhdeksän vuotta ja opettajana noin kymmenen vuotta. Stenforsin asiantuntijuusalue hankkeissa on tiedon kerääminen, jalostaminen ja raportointi.

TIIVISTELMÄ

Mika Rantonen, Janne Alatalo, Tarja Ajo & Eppu Heilimo (toim.)

Data-analytiikan hyödyntäminen valmistavan teollisuuden yrityksissä.

Jyväskylän ammattikorkeakoulun julkaisuja, 346

VauhtiData-hankkeen teos kokoaa yhteen kolmetoista data-analytiikkaa ja tekoälyä käsittelevää artikkelia. Kokoelmateos on suunnattu erityisesti valmistavan teollisuuden yrityksille, jotka haluavat kehittää toimintaansa dataperusteiseksi. Vaikka teollisuusyritykset jo tyypillisesti keräävät runsaasti tietoa prosesseistaan, jää suuri osa datan tarjoamasta kehitys- ja kasvupotentiaalista usein hyödyntämättä. Teoksessa tarkastellaan, miten data-analytiikkaa ja tekoälyä voidaan tehokkaasti hyödyntää kilpailukyvyn vahvistamiseen ympäristössä, jossa tiedolla johtaminen on noussut keskeiseksi kilpailutekijäksi. Esille nostetaan kilpailukyvyn vahvistaminen piilossa olevia mahdollisuuksia tunnistamalla, tuotantoprosesseja tehostamalla ja nopeammalla päätöksenteolla.

Artikkelit valottavat monipuolisesti data-analytiikan käsitteitä, terminologiaa, konkreettisia käyttötapauksia ja teknologisia ratkaisuja. Sisältö kattaa teemoja yleisestä datastrategiasta ja tiedolla johtamisen perusteista käytännönläheisiin menetelmiin kuten visualisointiin, tekoälyn sovelluksiin ja tietovarastoihin. Teos esittelee todellisia teollisuusympäristöissä toteutettuja ratkaisuja, jotka osoittavat datapohjaisten menetelmien hyödyt käytännössä. Nämä lähestymistavat tarjoavat yrityksille konkreettisia työkaluja kannattavuuden parantamiseen ja asiakastyytyväisyyden nostamiseen sekä tuotannon, laadun ja uusien liiketoimintamallien kehittämiseen. Jakamalla tietoa ja VauhtiData-hankkeessa syntynyttä käytännön kokemusta teos tukee teollisuusyrityksiä niiden matkalla kohti aidosti dataohjattua toimintaa.

Avainsanat: tekoäly, data, teollisuus, analyysi, digitalisaatio

ABSTRACT

Mika Rantonen, Janne Alatalo, Tarja Ajo & Eppu Heilimo (Eds.)

Utilisation of data analytics in manufacturing industry companies.

Publication of Jamk University of Applied Sciences, 346

The edited volume from the VauhtiData project brings together thirteen articles addressing data analytics and artificial intelligence. This collection is aimed at manufacturing industry companies seeking to develop data-driven operations. Although industrial companies typically already collect abundant information about their processes, much of the development and growth potential this data offers often remains untapped. The volume examines how data analytics and artificial intelligence can effectively strengthen competitiveness in an environment where data-driven management has become a key competitive factor. The focus is on strengthening competitiveness by identifying hidden opportunities, improving production processes, and enabling faster decision-making.

The articles comprehensively illuminate data analytics concepts, terminology, concrete use cases, and technological solutions. The content covers themes ranging from general data strategy and fundamentals of data-driven management to practical methods such as visualisation, artificial intelligence applications, and data warehousing. The edited volume presents real solutions implemented in industrial environments that demonstrate the benefits of data-based methods in practice. These approaches offer companies concrete tools for improving profitability and customer satisfaction, as well as developing production, quality, and new business models. By sharing knowledge and practical experience gained in the VauhtiData project, the work supports industrial companies on their journey toward truly data-driven operations.

Keywords: artificial intelligence, data, industry, analysis, digitalisation

LUKIJALLE

MIKA RANTONEN, JANNE ALATALO, TARJA AJO & EPPU HEILIMO,
JYVÄSKYLÄN AMMATTIKORKEAKOULU

Data-analytiikka ja tekoäly ovat nousemassa yhdeksi keskeisimmistä tekijöistä valmistavan teollisuuden kilpailukyvyn vahvistamisessa.

Tehokas datan hyödyntäminen voi paljastaa piilossa olevia mahdollisuuksia, tehostaa tuotantoprosesseja, parantaa laatua ja nopeuttaa päätöksentekoa. Käytännössä tämä tarkoittaa parempaa kannattavuutta, ketterämpää toimintaa ja tyytyväisempiä asiakkaita.

Valmistavan teollisuuden yrityksissä on usein hyvä tilanne datan keräämisen suhteen. Yritykset keräävät dataa muun muassa valmistusprosessistaan prosessinohjauksen ja valvonnan vuoksi. Siten valmistusprosessit ovat yksi sovelluskohde, johon valmistavan teollisuuden yrityksissä voidaan käyttää data-analytiikan menetelmiä ja tekoälyä yrityksen tuottavuuden lisäämiseksi.

Monissa yrityksissä ei kuitenkaan vielä hyödynnetä datan koko potentiaalia. Yrityksissä on jo hyvä käsitys siitä, että data on arvokasta ja sitä tulisi kerätä. Käsitys datan konkreettisista hyödyntämismahdollisuuksista voi kuitenkin olla hämärä. Tämän vuoksi on tärkeää jakaa oppeja, kokemuksia ja osaamista oikean elämän tapaustutkimuksista, joissa dataa on hyödynnetty yritysten ongelmien ratkaisemiseen. Vaikka jokainen valmistavan teollisuuden yritys ei kohtaa täsmälleen samoja ongelmia, tapaustutkimuksissa opituista asioista on usein hyötyä myös muille, sillä oppeja voi tavallisesti soveltaa muihinkin ongelmiin.

Jyväskylän ammattikorkeakoulu, Seinäjoen ammattikorkeakoulu ja Satakunnan ammattikorkeakoulu muodostivat yhdessä korkeakoulukonsortion, joka toteutti VauhtiData-nimisen hankkeen vuosien 2023–2025 aikana. Hanke toteutettiin läheisessä yhteistyössä paikallisten yritysten kanssa. VauhtiData-hankkeen kumppaniyrityksiä olivat Polarteknik Oy, BMH Technology Oy, Valtra Oy Ab, LS Cargo Finland, Ensto Finland Oy, Mäkelä Alu Oy ja Alumech Oy. Hanke oli Euroopan unionin osarahoittama.

Hankkeessa tehtiin tapaustutkimuksia ongelmiin, jotka oli tunnistettu hankkeen kumppaniyritysten kanssa. Yrityslähtöinen soveltava tutkimustyö takaa sen, että tutkimuksella ratkaistaan yritysten oikeita eikä oletettuja ongelmia. Yritysten kanssa yhteistyössä tehtävä soveltava tutkimus myös mahdollistaa sen, että tutkimusta voidaan tehdä yritysten oikealla datalla.

Tämä teos on koottu tukemaan valmistavan teollisuuden yrityksiä matkalta kohti datan ohjaamaa tai dataan pohjautuvaa toimintaa, josta voidaan puhua tiedolla johtamisena. Teoksessa on kolmetoista artikkelia data-analytiikkaan liittyvistä käsitteistä, terminologiasta, konkreettisista käyttötapauksista ja teknologisista ratkaisuista. Teoksessa jaetaan VauhtiData-hankkeessa syntynyttä käytännön kokemusta ja oppeja. Artikkeleissa yhdistyy käytännön kokemus, ajankohtainen tieto sekä näkemykset siitä, miten data-analytiikka ja tekoäly voivat auttaa ja palvella liiketoimintaa parhaalla mahdollisella tavalla.

Teoksen alussa on neljä artikkelia, jotka pohtivat aihetta yleisellä tasolla. Teos alkaa Mikko Pakkaselan artikkelilla, jossa johdatetaan lukija pk-yritysten datastrategiaan. Toisessa artikkelissa Mika Rantonen avaa datan hyödyntämisen haasteita, minkä jälkeen Tarja Ajon artikkelissa kerrotaan data-avaruuksista – datan hyödyntämisen mahdollisesta tulevaisuudesta. Yleisen osion loppuksi Tarja Ajon & Mika Rantosen artikkelissa käsitellään vielä data-analytiikan jaottelua eri näkökulmista.

Yleistä tilannekuvaa valottavien tekstien jälkeen teos jatkuu kuudella artikkelilla, jotka jakavat käytännön osaamista aiheen ympäriltä. Juha Stenfors kirjoittaa tiedon jalostamisesta, Mika Rantonen datan varastoinnin termeistä ja Tarja Ajo & Janne Alatalo sadepilvikuvaajamenetelmästä datan visualisoinnissa. Käytännön osaamista käsittelevistä artikkeleista kolme viimeistä keskittyvät tekoälyn hyödyntämiseen eri näkökulmista: Vesa Vertainen kirjoittaa tekoälyn käytöstä pk-yrityksissä kyberturvallisuuden näkökulmasta, Saku Kaarlejärvi & Juha-Matti Arola tarkastelevat tekoälyn hyödyntämistä datankeruussa ja Janne Alatalo kirjoittaa generatiivisen tekoälyn mahdollisuuksista datan hyödyntämisessä.

Teoksen viimeiset kolme artikkelia käsittelevät tapaustutkimuksia, jotka on toteutettu osana VauhtiData-hanketta. Mika Valkama kirjoittaa RFID-tekniologian hyödyntämisestä teollisuusympäristössä, Jari Kyngäs matemaattisesta mallintamisesta junien ovien ohjausyksikön hajoamisen ennustamisessa ja Eppu Heilimo datalähtöisestä lähestymistavasta kuljetuslogistiikan alalla kuljetuspäästöjen mittaamiseen.

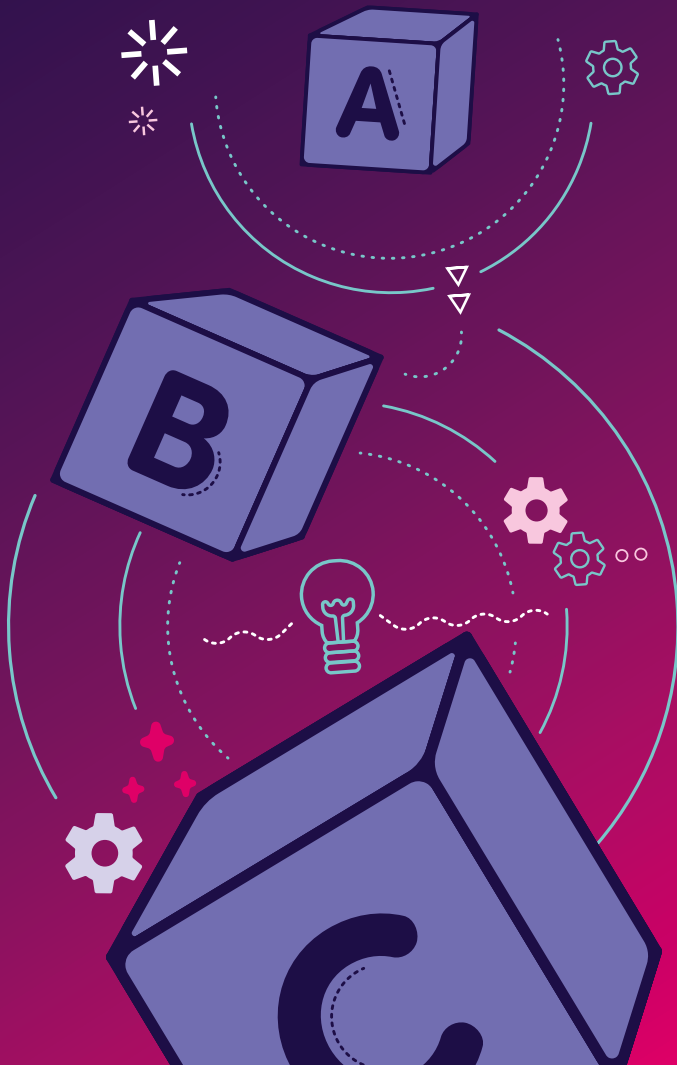
Olitpa vasta aloittelemassa data-analytiikan hyödyntämistä tai etsimässä keinoja nykyisten ratkaisujen kehittämiseen, toivomme, että löydät oppaasta ajatuksia herättäviä ideoita ja konkreettisia työkaluja oman organisaatiosi tueksi. Tervetuloa mukaan matkalle kohti datalla johdettua tulevaisuutta!

Jyväskylässä 30.4.2025, teoksen toimituskunta





POHJATIEDOT KUNTOON - DATATALOUDEN PERUSTEITA



DATATALOUS JA PK-YRITYKSEN DATASTRATEGIA

MIKKO PAKKASELA, SATAKUNNAN AMMATTIKORKEAKOULU

Suomi on maailman digitalisaation kärkimaa. Tästä huolimatta datan hyödyntäminen edistyy suomalaisyrityksissä hitaasti ja kilpailukyky sekä tuottavuus ovat jäljessä digivertailujen huipputuloksista. Tässä artikkelissa tarkastellaan datatalouden nykytilannetta sekä datastrategian merkitystä innovaatiotoiminnan vauhdittajana.

Suomi on ollut jo useiden vuosien ajan digitaalisten ja innovatiivisten ratkaisujen kärkimaa. Euroopan komission vuoden 2024 innovaatiovertailussa Suomi sijoittui kolmanneksi EU-maiden joukossa, heti Tanskan ja Ruotsin jälkeen. Erityisenä vahvuutena nähdään yhteistyö julkisen ja yksityisen sektorin sekä innovatiivisten pk-yritysten välillä. Toisaalta kehitystä jarruttavat muun muassa työn tuottavuuden heikko kasvu, vähäinen korkean teknologian vienti ja valtion niukka panostus tutkimukseen ja kehitykseen (European Commission, 2024, s. 18, 81).

Digibarometri 2023 -julkaisun mukaan Suomi nousi maailman digitalisaation kärkeen 78,9 pisteellään. Se ohitti niukasti Tanskan (78,6 pistettä), ja jätti Alankomaat (71,4) ja Ruotsin (70,5) selvästi taakseen (Ali-Yrkkö ym., 2023). Tästä huolimatta kilpailukyky ja tuottavuus eivät vastaa innovaatio- ja digivertailujen huipputuloksia. Big datan hyödyntäminen suomalaisyrityksissä etenee hitaasti, ja yrityskokoluokkien välillä on merkittäviä eroja. ICT-pääomien vaikutus talouskasvuun on ollut Suomessa viimeisen vuosikymmenen aikana, mikä on huolestuttavaa, koska datan hyödyntämisellä on selkeä yhteys tuottavuuden parantamiseen (Ali-Yrkkö ym., 2023, s. 13, 37).

Elinkeinoelämän tutkimuslaitoksen mukaan työn tuottavuuden yleisindeksi on pysynyt paikallaan vuodesta 2017 ja laskenut vuodesta 2021. Yksityisten yritysten kokonaistuottavuus on heikentynyt 10 % vuoden 2007 jälkeen (Koski & Maliranta, 2023, s. 3; Pohjola, 2025, s. 12). Samalla kilpailijamaat kuten Ruotsi, Tanska ja Saksa ovat parantaneet omaa kilpailukykyään. Erityisen heikko tuottavuus koskee alle 50 työntekijän yrityksiä, jotka jäävät jopa 20 % jälkeen suuremmista yrityksistä (Koski ym., 2023, s. 23).

Alle 10 työntekijän mikroyritykset muodostavat 95,5 % suomalaisista yrityksistä ja alle 50 työntekijän yritykset 3,6 %. Näin ollen yli 99 % yrityksistä kuuluu näihin kahteen ryhmään ja ne työllistävät 45 % työvoimasta (Suomen Yrittäjät, ei pvm.). Koska juuri näissä yrityksissä on myös merkittävä osa uusista työpaikoista, tuottavuuden kehittäminen niissä tarjoaa valtavan potentiaalin (Elinkeinoelämän keskusliitto, ei pvm.).

TIEDOLLA JOHTAMINEN JA INNOVAATIOTOIMINTA

Yritysten innovaatio voi tarkoittaa uuden tai parannetun tuotteen, palvelun, prosessin tai markkinoinnin kehittämistä. Tärkeintä on, että innovaatiot vastaavat asiakkaiden tarpeisiin ja tuovat lisäarvoa, joista ollaan valmiita maksamaan.

Tilastokeskuksen mukaan vuosina 2020–2022 innovaatiotoimintaa harjoitti 62 % suomalaisyrityksistä. Vuonna 2022 innovaatiomenot olivat 7,1 miljardia euroa ja niiden osuus kokonaisliikevaihdosta oli 16 % (Tilastokeskus, 2024). Innovaatiotoiminta painottuu Tilastokeskuksen (2024) mukaan suurempiin yrityksiin: 10–49 työntekijän yrityksissä 59 % osallistui innovaatiotoimintaan, kun suurissa yrityksissä osuus oli 82 %. Toimialoista teollisuus erottuu edukseen: pienissä teollisuusyrityksissä 63,1 % ja suurissa jopa 93,3 % harjoittaa innovaatiotoimintaa. Tämä on erityisen merkityksellistä palvelualalla, joka muodostaa noin 40 % BKT:sta ja lähes kolmanneksen viennistä (Saarinen, 2024).

Tilastokeskuksen (2024) tutkimuksessa korostuu kolme innovaatiotoiminnan keskeistä tekijää: asiakas- ja markkinatarpeiden tunnistaminen, digitalisaation ja datan hyödyntäminen sekä T&K-toiminnan kehittäminen. Pienissä yrityksissä painottuvat erityisesti kaksi ensimmäistä, joissa data-analytiikka näyttelee keskeistä roolia.

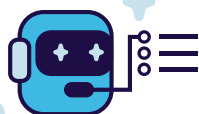
Tiedolla johtaminen puolestaan mahdollistaa perustelun päätöksenteon ja investointien vaikutusten seuraamisen, mikä auttaa hallitsemaan riskejä ja parantamaan tuottavuutta.



TIEDOLLA JOHTAMINEN, DATASTRATEGIA JA INNOVAATIOITOIMINTA

Datan tehokas hyödyntäminen edellyttää tiedossa olevia datalähteitä, järjestelmällistä keruuta, laadunvarmistusta ja osaamisen ylläpitämistä. Datastrategia auttaa jäsentämään nämä osa-alueet. Pk-yrityksissä on kuitenkin yleistä, että datasta vastaa yksi henkilö muun työn ohella ilman dokumentointia. Satakunnan ammattikorkeakoulun hallinnoimassa ja EU:n osarahoittamassa Tiedolla johtamisen edelläkävijä -hankkeessa toteutetun kyselyn mukaan vain 6 % vastanneista yrityksistä oli laatinut datastrategian (n=64), mikä kertoo suu-rista hyödyntämättömistä mahdollisuuksista (Pakkasela, 2024).

Pk-yritysten datastrategiat voivat olla yksinkertaisia ja kehittyä ajan myötä. Keskeistä on, että strategia tukee yrityksen liiketoimintatavoitteita. Pienissä yrityksissä hyvä tapa on keskustella datastrategian luomisesta ChatGPT:n kanssa. Yleisen sparrauksen lisäksi se osaa avustaa datastrategian laatijaa sekä prioriteettijärjestyksessä että laajuudessa. Kun ChatGPT tekoälymallia syötteistää viestillä ”Mitä pk-yrityksen datastrategia tulisi pitää sisällään?”, niin tekoälymalli ehdottaa datastrategian keskeisiksi alueiksi (OpenAI, 2025):



- Datan keruu ja hallinta
- Tavoitteet ja liiketoimintahyödyt
- Datan hyödyntäminen ja analyytiikka
- Tietoturva ja tietosuojat
- Teknologiat ja infrastruktuuri
- Datan hyödyntämisen kulttuuri ja osaaminen
- Datan jakaminen ja yhteistyö

Datastrategia tukee pitkäjänteistä innovaatioitoimintaa, jonka vaikuttavuuden seuranta edellyttää useiden toimenpiteiden kokonaisarviointia.

Tehokas innovaatioitoiminta vaatii myös toimivia tietojärjestelmiä. Tärkeää on, miten nämä järjestelmät tukevat innovaatioitoiminnan keskeisiä osa-alueita: asiakastarpeiden tunnistamista, digitalisaatiota ja T&K-kehittämistä. Pk-yrityksille tämä on erityisen haasteellista, koska niillä on usein rajallisesti sekä resursseja että osaamista. Pk-yritykset hyötyvät erityisesti pienistä ohjelmistotaloista, jotka tarjoavat skaalautuvia ja kustannustehokkaita ratkaisuja. Nämä toimijat voivat auttaa yrityksiä myös rajapintojen ja datan omistajuuden kysymyksissä.

Kaiken kaikkiaan tiedolla johtaminen ja innovaatiokyky kulkevat käsi kädessä. Pk-yrityksille tämä voi olla suuri mahdollisuus – mutta myös riski, jos tiedolla johtamista ei oteta tosissaan. Laiminlyönnit näkyvät ennen pitkää liiketoiminnassa ja markkina-asemassa.

Yritysten tulee olla tietoisia toimintaympäristön muutoksista ja mahdollisuuksista. Korkeakoulujen hankkeet, kuten VauhtiData - Datasta vauhtia valmistavan teollisuuden liiketoimintaan ja Tiedolla johtamisen edelläkävijä -hankkeet, tarjoavat näissä teemoissa tukea erityisesti pk-yrityksille. Hankkeissa jaetaan ajankohtaista, puolueetonta tietoa, toteutetaan pilotointeja sekä rakennetaan hyödyllisiä verkostoja. Molempia hankkeita hallinnoi Satakunnan ammattikorkeakoulu ja niitä osarahoittaa Euroopan aluekehitysrahasto (EAKR).

LÄHTEET

Ali-Yrkkö, J., Kässi, O., Pajarinen, M. & Rouvinen, P. (2023). *Digibarometri 2023: Data, tekoäly ja talouskasvu*. Taloustieto.

European Commission. (2024). *European Innovation Scoreboard 2024*. Publications Office of the European Union. <https://doi.org/10.2777/779689>

Elinkeinoelämän keskusliitto. (Ei pvm.). *Yritykset*. Haettu 4.3.2025 osoitteesta <https://ek.fi/tutkittua-tietoa/tietoa-suomen-taloudesta/yritykset/>

Koski, H. & Maliranta, M. (2023). *Työkaluja tuottavuuskasvun vahvistamiseen Suomessa*. ETLA Muistio No 118. <https://pub.etla.fi/ETLA-Muistio-Brief-118.pdf>

Koski, H., Maliranta, M., Fornaro, P., Juuti, T., Kiema, I. & Pajarinen, M. (2023). *Yritysten tuottavuuserot ja tuottavuuden eturintama*. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja, 42. <http://urn.fi/URN:ISBN:978-952-383-270-1>

OpenAI. (2025). *ChatGPT 4o* (Helmikuun versio) [Generatiivinen tekoälymalli]. <https://chatgpt.com>

Pakkasela, M. (2024). *Tiedolla johtaminen pk-yritysten innovaatiotoiminnassa*. Turun yliopisto, kauppa- ja talousakademi, Porin yksikkö.

Saarinen, M. (2024). Suomen tuottavuus sukeltaa, ja siihen löytyy yllättävä selitys. *Helsingin Sanomat*. <https://www.hs.fi/talous/art-2000010775819.html>

Suomen Yrittäjät. (Ei pvm.). *Yrittäjyys Suomessa*. Haettu 4.3.2025 osoitteesta <https://www.yrittajat.fi/yrityksien-yrittajien-tietoa-yrityksistä/yrityksien-yrittajien-suomessa/>

Tilastokeskus (2024). *Innovaatiotoimintaa harjoitti 62 % yrityksistä vuosina 2020–2022*. Haettu 22.3.2025 osoitteesta <https://stat.fi/julkaisu/cIn1gzxf03sv20bw1rc1nnwua>

DATAN HYÖDYNTÄMISEN HAASTEET

MIKA RANTONEN, JYVÄSKYLÄN AMMATTIKORKEAKOULU

Nykyisin yritykset keräävät hurjan määrän dataa, mutta datan hyödyntämisessä voi olla haasteita. Tässä artikkelissa kuvataan dataan liittyviä haasteita, joihin olemme törmänneet aikaisemmissa dataan pohjautuvissa Jyväskylän ammattikorkeakoulun hankkeissa.



Dataa on hyödynnetty yrityksien liiketoiminnassa niin kauan kuin sitä on ollut mahdollista kerätä, mutta ennen tietokoneiden ja sähköisen tallentamisen tuloa se on ollut työlästä ja haastavaa. Digitalisaation myötä datan keräys ja tallentaminen digitaalisessa muodossa helpottui ja datan jatkojalostaminen mahdollistui tuoden mukanaan lähes rajattomat mahdollisuudet datan hyödyntämiseen liiketoiminnassa. Nykyisin lähes kaikki yritykset ja organisaatiot tuottavat ja käyttävät dataa erilaisten tietojärjestelmien kautta. Vaikka datan tuottaminen ja käyttäminen on yrityksille jo arkipäivää, datan hyödyntäminen liiketoiminnan tukemisessa tai uuden liiketoiminnan kehittämisessä dataperusteisesti on haasteellista. Yksittäisiä käyttötapauksia tai Proof-of-Concept (PoC)- tyyppisiä ratkaisuja on joissakin yrityksissä tehty. Datapohjaisten projektien suunnittelussa ja kehittämisessä kohdataan haasteita, jotka toistuvat samankaltaisina eri organisaatioissa.

Haasteiksi on tunnistettu (Sivula ym., 2023, s. 43):

- Datan saatavuus ja siiloutuminen
- Datan laatu
- Datan omistajuus ja hallinta puuttuu

DATAN SAATAVUUS JA SILOUTUMINEN

Tallennetun datan määrä on kasvanut räjähdysmäisesti digitalisaation seurauksena (Gupta & Giri, 2018). Dataa tallennetaan suuria määriä tietojärjestelmiin, joka voi luoda väärän mielikuvan datan käyttömahdollisuuksista. Haasteeksi usein kuitenkin voi muodostua datan saaminen tietojärjestelmistä muuhun käyttöön, kuten vaikka keskitettyyn tietovarastoon. Perinteiset tietojärjestelmät silouttavat datan ja vaikeuttavat datan saatavuutta yli organisaattiorajojen. Datat saatavuuteen hallintaan liittyy **osaaminen** ja **tekniikka**.

Dataan pohjautuvissa projekteissa tyypillisesti data haetaan suoraan tietojärjestelmien tietokannoista. Tämä vaatii tietokantaosaamista ja tietokannan sisällön tuntemista, esimerkiksi tietokannan taulujen välisten suhteiden (relaatioiden) tunteminen. Tällainen osaaminen on yleensä keskittynyt yhdelle tai muutamalle tietokanta-asiantuntijalle yrityksessä, jolloin datan saatavuuteen voi tulla viiveitä. Osaamisen haasteet voivat liittyä myös käytettäviin teknologioihin, koska useiden isojen kaupallisten palveluntarjoajien palvelut voivat erota toisistaan. Data voi olla saatavilla hyvin vaihtelevassa formaatissa, jolloin datan käsittely vaatii tiettyjä ohjelmointitaitoja, jotta datan saa esikäsiteltyä analysoitavaan muotoon. Tekniikan osalta haasteena voi olla datan määrään liittyviä asioita, kuten tiedonsiirtoväylän asettamat rajoitteet viiveelle, nopeudelle tai kapasiteetille. Teollisuuden alalla on omat haasteensa. Datat saanti teollisuuden koneista voi vaatia erikoisosaamista, tai tiedonsiirtoväylä puuttuu kokonaan, jolloin ainoa tapa siirtää dataa on esimerkiksi muistitikku. (Sivula ym., 2023, s. 46)

DATAN LAATU

Dataan pohjautuvissa projekteissa usein haasteena on datan laatu eli heikko- ja huonolaatuinen data. Huonolaatuisen datan käyttö johtaa vääristyneeseen lopputulokseen ja edelleen väärin johtopäätöksiin liiketoiminnassa. Tästä voi seurata katastrofaalisia taloudellisia vaikutuksia. Tästä syystä on tärkeää, että dataan ja siitä tehtyihin analyyseihin sekä johtopäätöksiin voidaan luottaa (Moses ym., 2022). Datat laatuun vaikuttaa suuri joukko eri tekijöitä, joihin tulisi kiinnittää huomiota dataan pohjautuvan projektin alkuvaiheessa ja suunnittelussa.



Datan laatuun vaikuttavat sen täydellisyys, tarkkuus, analysoitavuus, ajallisuus, asiaankuuluvuus, saatavuus sekä paikkansapitävyys (Hawker, 2023, s. 5). Kuitenkaan kaikki yrityksen data ei ole hyödyllistä, ja hyödytöntä dataa ei kannata kerätä ja varastoida, koska datan varastointi maksaa varsinkin datan määrän kasvaessa. Yrityksellä täytyy olla prosessien standardit määritettynä, jotta datan laatu pysyy hyvänä. Datan käytettävyys on tärkeää, koska dataa voivat hyödyntää yrityksen eri organisaatiot.

Datan tulee olla paikkansa pitävää ja virheetöntä, jotta liiketoimintaa kuvaavien mittareihin, tiedolla johtamiseen tai tekoälyratkaisuihin voidaan luottaa. Jos opetusdata koneoppimis- tai syväoppimisalgoritmin kehityksessä on huonolaatuista tai vinoutunutta, kehitetty tekoälyratkaisu tekee systemaattisesti vääriä tulkintoja (Sivula ym., 2023, s. 59). Tällä saattaa olla vaikutusta yrityksen toimintaan varsinkin silloin, jos tekoälyratkaisua käytetään päätöksenteon tukena (Hawker, 2023, s. 5).

Datan esikäsittelyvaiheessa datan korjaaminen käyttökelpoiseksi on aikaa vievää ja voi olla mahdotonta.

Dataan liittyvät laatuongelmat periytyvät eli juurisyöt ovat olemassa jo edeltävissä prosesseissa ja niihin liittyvissä tietojärjestelmissä. Datan esikäsittelyvaiheessa datan korjaaminen käyttökelpoiseksi on aikaa vievää ja voi olla mahdotonta. Datan laatuongelman juurisyöt pitää selvittää ja pyrkiä korjaamaan. Juurisyiden korjaaminen olisi kaikkein kustannustehokkainta, sillä niiden korjaamisen ansiosta välttyttäisiin tulevaisuudessa vastaavilta tilanteilta. (Sivula ym., 2023, s.49)



Dataan pohjautuvissa projekteissa dataa tarvitaan yleensä useasta eri lähteestä ja tällöin puhutaan datan rikastamisesta (Moses ym., 2022). Tämä tarkoittaa datan yhdistelemistä ja yhteismitallistamista, mikä on pääsääntöisesti dataan pohjautuvan projektin suurin ja kallein työvaihe. Ilman tämän työvaiheen toteutusta dataa ei pystytä hyödyntämään suunnitellulla tavalla.

DATAN OMISTAJUUS JA HALLINTA

Datan omistajuuksien määrittäminen voi olla haastavaa ja sen suhteen voidaan olla epätietoisia. Yleisesti tietojärjestelmille asetetaan omistajat liiketoiminnan tai tietohallinnon puolelta tai molemmista. Mutta vastaavanlaista pääomien omistajuutta on mietitty harvoin tai omistajaksi on voitu asettaa esimerkiksi johtotason liiketoimintajohtaja. On erittäin tärkeää, että datan omistajalla on käytännön näkemystä datan laadusta ja saatavuudesta (Hawker, 2023, s. 17).

Tämän takia tietopääomille pitäisi määrittää omistajat, jotka omistavat kaiken prosesseissa liikkuvan datan ja vastaavat datan/tietopääomien laadusta. Näillä henkilöillä on tärkeä operatiivinen rooli osallistua dataan pohjautuviin projekteihin, joissa tietopääomia käsitellään. Haasteeksi saattaa muodostua myös data, joka ylittää yrityksen eri liiketoiminta-alueet ja tietojärjestelmät. Esimerkiksi toiminnanohjausjärjestelmässä voidaan käsitellä dataa tuotannosta, jakelusta, varastonhallinnasta, laskutuksesta ja kirjanpidosta (Sivula ym., 2023, s. 50). Sekaannusta datan omistajuudesta voi syntyä, jos ei ole sovittu omistaako datan yrityksen tuotteen ostaja/käyttäjä vai tuotteen valmistaja.

LÄHTEET

Gupta, S. & Giri, V. (2018). Introduction to Enterprise Data Lakes. Teoksessa *Practical Enterprise Data Lake Insights: Handle Data-Driven Challenges in an Enterprise Big Data Lake* (s. 1–31). https://doi.org/10.1007/978-1-4842-3522-5_1

Hawker, R. (2023). *Practical Data Quality: Learn Practical, Real-world Strategies to Transform the Quality of Data in Your Organization*. Packt Publishing.

Moses, B., Gavish, L. & Vorwerck, M. (2022). *Data Quality Fundamentals: A Practitioner's Guide to Building Trustworthy Data Pipelines* (s. 1–10). Yhdysvallat: O'Reilly Media.

Sivula, A., Aho, M. & Laukkanen, M. (2023). *Datasta liiketoimintaan – 10 tehokasta työkalua*. Alma Talent.

DATA-AVARUUDET – KOHTI DATAN HYÖDYNTÄMISEN SEURAAVAA TASOA

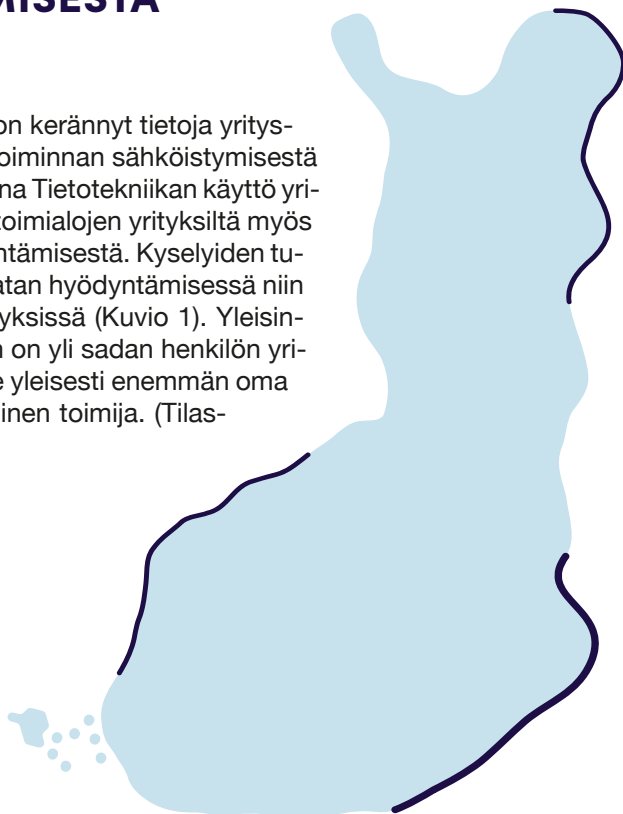
TARJA AJO, JYVÄSKYLÄN AMMATTIKORKEAKOULU

Data-avaruuDET on kuvattu datan hyödyntämisen kehitysaskelena, joka mahdollistaa yrityksille uusia innovaatioita ja liiketoiminnan kasvua. Tässä artikkelissa käsitellään datan hyödyntämisen nykytilaa Suomessa, yleisesti data-avaruuksia ja näkökulmia data-avaruuksiin tai -ekosysteemeihin liittymiselle.

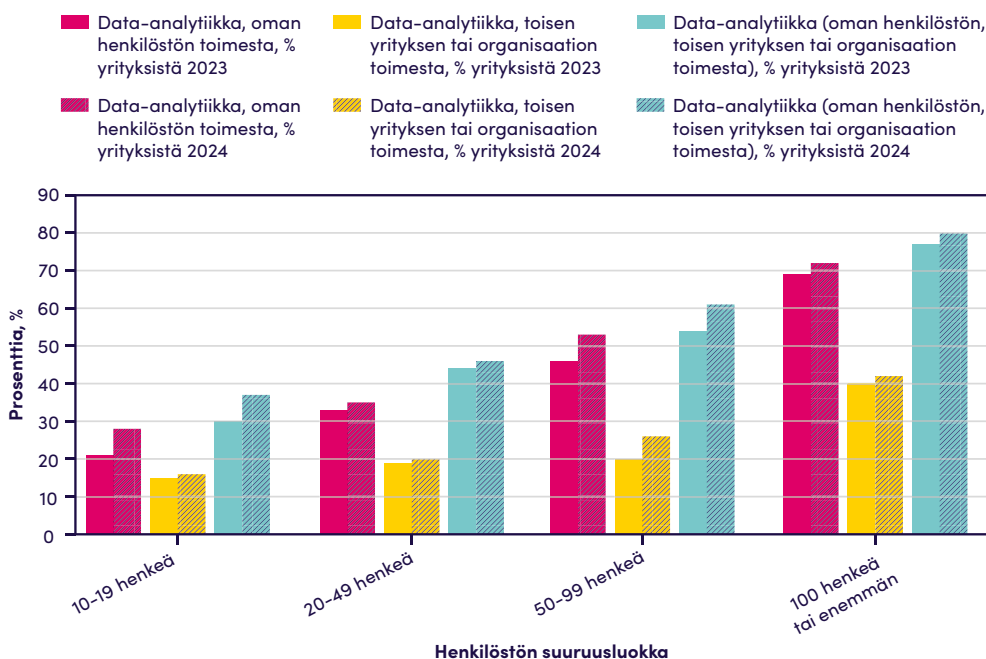


DATAN HYÖDYNTÄMISESTÄ SUOMESSA

Suomessa Tilastokeskus (2024b) on kerännyt tietoja yritysten tietotekniikan käytöstä ja liiketoiminnan sähköistymisestä vuodesta 2000 lähtien. Viime vuosina Tietotekniikan käyttö yrityksissä -kyselyssä on kysytty eri toimialojen yrityksiltä myös data-analytiikan ja tekoälyn hyödyntämisestä. Kyselyiden tulokset näyttävät selkeää kasvua datan hyödyntämisessä niin pienissä kuin suuremmisakin yrityksissä (Kuvio 1). Yleisintä data-analytiikan hyödyntäminen on yli sadan henkilön yrityksissä, ja data-analytiikkaa tekee yleisesti enemmän oma henkilöstö kuin yrityksen ulkopuolinen toimija. (Tilastokeskus, 2024a.)



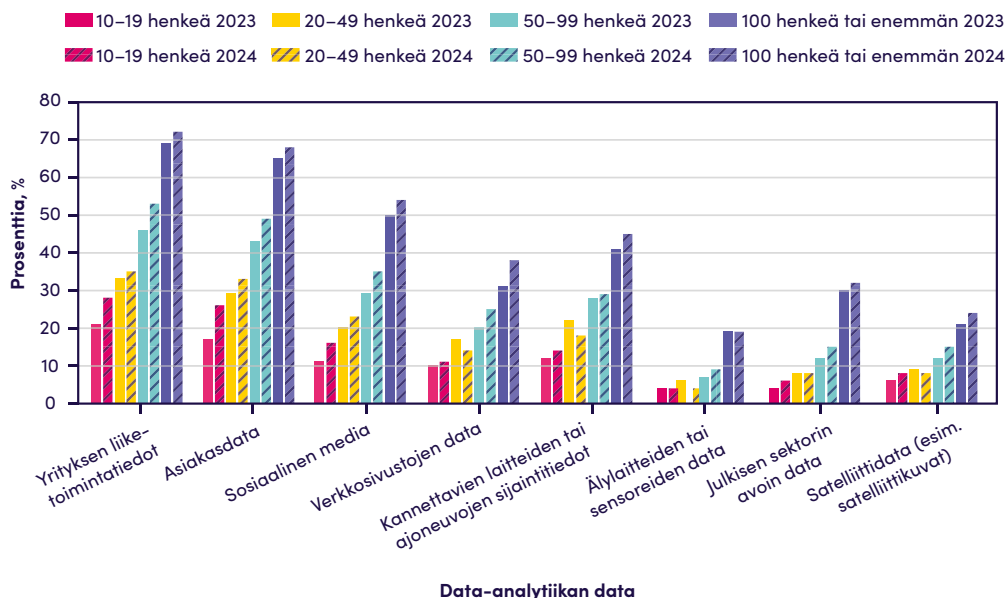
Data-analytiikan hyödyntäminen yrityksissä, %



Kuvio 1. Data-analytiikan hyödyntäminen yrityksissä vuosina 2023 ja 2024 henkilös-
töluokittain. Koottu Tilastokeskuksen (2024a) tietojen perusteella.

Kyselyn tuloksista saadaan viitettä myös analysoitavan datan luonteesta (Kuvio 2). Suurimmaksi osaksi yritykset näyttävät analysoivan omia liiketoimintatietojaan, minkä jälkeen yleisintä on asiakasdatan analysointi. Yrityksen itse tuottaman datan analysointi voidaan katsoa olevan matalalla roikkuva hedelmä, josta on tavallisesti helpointa ja hyödyllisintä lähteä liikkeelle, ellei analytiikkaa ole aikaisemmin tehty. Siten ei ole yllättävää, että kyselyn kategorioista julkisen sektorin avoin data ja muut luokat, jotka todennäköisesti eivät ole yritysten itse keräämää dataa, ovat erityisesti pienempien henkilöstöluokkien yrityksissä vähemmän edustettuja. Kun data-analytiikan hyötyihin ja menetelmiin on saatu ensikosketusta helposti saatavilla olevasta datasta, on tavalista, että data-analytiikan mahdollisuuksia tutkitaan lisää ja analysointia laajennetaan muun datan hyödyntämiseen.

Data-analytiikan data (oman henkilöstön toimesta), %



Kuvio 2. Yritysten data-analytiikkaan hyödyntämä data vuosina 2023 ja 2024 henkilöstöluokittain. Koottu Tilastokeskuksen (2024a) tietojen perusteella.

Tulevaisuudessa mielenkiinnon kohteeksi muodostunee yritysten hyödyntämä data ja erityisesti muu kuin yritysten oma, itse tuotettu data. Dataekosysteemit ja -avaruudet, joissa jaetaan dataa yrityksiltä ja organisaatioilta toisille hyödynnettäväksi, ovat viime vuosina olleet Euroopassa kehityksen kohteena ja strategisena tavoitteena eurooppalaisten yritysten kilpailukyvyyn parantamiseksi. Myös Suomessa on enenevässä määrin tehty data-avaruuksiin liittyvää kehitys-, tiedotus- ja tutkimustyötä.

Sitran ja Innolinkin Dataa hyödyntävien ekosysteemien ja verkostojen kehitystaste 2023 -kyselytutkimuksessa (Sievers ym., 2023) kartoitettiin datataloutta harjoittavien yritysten kokemuksia ja näkemyksiä datan jakamisesta. Kyselyssä 67 % vastaajista (n=109) kertoi saaneensa hyötyä datan jakamisesta. Konkreettisia hyötyjä olivat mm. uusien palveluiden kehittäminen, liiketoimintamahdollisuudet ja -kumppanit. Suurin osa tutkimukseen osallistuneista yrityksistä ja organisaatioista tähtäsi myös edistyneempään datan jakamiseen, kuin tutkimushetkellä ekosysteemin tilanne oli.

Ongelmatonta datan jakaminen ei kuitenkaan ole. Hyötyjen ohella haastattelussa nousi esille useita haasteita, kuten rahoitus, datan jakaminen ja laatu, lainsäädäntöön liittyvät haasteet, intressiristiriidat sekä ymmärryksen ja resurssien

puute (Sievers ym., 2023). Yleisesti edellä mainitut ovat asioita, jotka on myös laajemmin tunnistettu ongelmiksi datan hyödyntämisessä yritys ja organisaatio-
rajojen yli (Euroopan komissio, 2024c).

DATAVERKOSTOT, -EKOSYSTEEMIT JA -AVARUUDET

Monelle data-avaruuksia tutumpia termejä lienevät dataverkosto ja -ekosysteemi, joilla viitataan useamman tahon muodostamaan verkostoon, jossa jollakin tavalla jaetaan dataa. Lähteestä riippuen dataekosysteemillä voidaan tarkoittaa yksittäisten tahojen verkostoa tai verkostojen verkostoa. Käytännössä näissä datan jakaminen voi olla mitä vain rajapintojen kautta tapahtuvasta järjestelmällisestä datan jakamisesta sähköpostilla satunnaisesti välitettäviin tiedostoihin. Myös säännöt datan jakamiselle voivat olla hyvin vaihtelevia tai puuttua kokonaan. Dataa hyödyntävien ekosysteemien ja verkostojen kehitysaste 2023 -kyselytutkimuksessa 19 % dataa verkostoissa jakavista ei ollut sopimuksia datan jakamiselle ja 7 % ilmoitti sopimusten olevan epätasa-arvoisia (kysymykseen vastanneita yrityksiä ja organisaatioita 94) (Sievers ym., 2023).

Data-avaruus (eng. *data space*) termillä tarkoitetaan dataverkostoja ja -ekosysteemejä organisoidumpaa ja säännellympää datan jakamista. Euroopan unionin rahoittaman Data Spaces Support Centren tuottamassa sanastossa termi *data space* määritellään infrastruktuurina, joka mahdollistaa datan liikumisen osapuolten välillä kyseisen data-avaruuden hallintokehyksen mukaisesti (Poikola ym., 2023). Käytännössä siis järjestelmässä on oltava yhteisesti sovitut säännöt datan jakamiselle, jotta kyseessä olisi data-avaruus.

EUROOPAN DATASTRATEGIA – VISIO EUROOPAN YHTEISISTÄ DATAN SISÄMARKKINOISTA

Suurin syy data-avaruuksien kehittämiseksi on datan hyödyntämisessä piilevä potentiaali ja tarve saada data innovaatioita tekeville. Euroopan datastrategiassa (Euroopan komissio, 2020) keskeisenä tavoitteena pidetään yhtä suurta data-avaruutta, jonka kautta eri toimialojen data olisi vaivattomasti saatavilla. Suurten datamassojen tunnistetaan toistaiseksi olevan muutamien suuryritysten hallussa, mikä ei kannusta teknologiakehitykseen yrityksissä. Jos dataa ei saada yritysten ja muiden kehitystyötä tekevien käyttöön, uhkana nähdään kehityksestä jälkeen jääminen ja kilpailukyvyyn heikentyminen Yhdysvaltojen ja Kiinan rinnalla.

Monien mahdollisuuksien ohella datan jakamiseen tunnistetaan kuitenkin liittyvän useita haasteita. Yksi merkittävä ongelma sekä yritysten että yksilöiden näkökulmasta on luottamus sopimusten noudattamiseen ja arkaluontoisten tietojen käsittelemiseen tietoturvallisesti ja eettisesti. Lähtökohtaisesti yksilöiden tulisi voida päättää, mitä dataa heistä jaetaan, eikä yksilöä tulisi voida tunnistaa datasta. Yritysten puolestaan voi olla vaikea päästä yhteisymmärrykseen datan jakamiseen liittyvistä ehdoista, ja pienemmät yritykset ovat vaarassa joutua yliajetuiksi näistä neuvoteltaessa. (Euroopan komissio, 2020.)

Data-avaruuksien tavoitteena on teknisen ratkaisun lisäksi luoda sopimus-pohjaa ja normistoa, johon voidaan tukeutua, jotta datan jakaminen tapahtuisi turvallisesti ja EU:n periaatteiden mukaisesti (Euroopan komissio, 2020). Data-avaruuksien hallinnon rungoiksi on kirjoitettu datahallintosäädös (eng. Data Governance Act) ja datasäädös (eng. *Data Act*), joista ensimmäisen on tarkoitettu tarjoavan puitteet julkisen sektorin datan hyödyntämiseen (Euroopan komissio, 2024c). Myöhemmin julkaistu datasäädös täydentää datahallintosäädöstä yritysten näkökulmasta määrittelemällä muun muassa aina kohtuuttomina pidettäviä sopimusehtoja yritysten välillä (Euroopan komissio, 2024b).

DATA-AVARUUKSIEN TILANNEKUVAA MEILTÄ JA MUUALTA

Matkalla kohti yhtä yhteistä data-avaruutta välietappeina pidetään toimialakohtaisia data-avaruuksia. Eri toimialoilla on omat ominaispiirteensä, ja näillä digitaalinen kehitys kulkee omaa vauhtiaan, minkä takia toimialakohtaiset data-avaruudet ovat luonnollinen kehitysaskel. EU-tasolla (Euroopan komissio, 2024a) yhteisiä data-avaruuksia on listattu neljältätoista eri sektorilta. Kaiken kaikkiaan eri sektoreiden data-avaruuksia EU:n Common European Data Spaces -verkkosivulla on listattu useita kymmeniä, minkä lisäksi muutamia on kehitteillä. Valmistavassa teollisuudessa data-avaruuksia on toistaiseksi kolme: Data Space 4.0, SM4RTENANCE ja UNDERPIN.

Suomessakin data-avaruushankkeita on käynnissä useita. Sitran Suomalaisen data-avaruuksien tilannekuva -selvityksessä (Poikola ym., 2024) tunnistettiin yli 30 suomalaista tai eurooppalaista data-avaruushanketta, joissa suomalaisyrityksiä oli mukana. Tunnistetuista hankkeista kuitenkin todettiin, ettei kaikista näistä ole selvää, ovatko ne varsinaisesti data-avaruushankkeita. Eniten meneillään olevia projekteja selvitystä tehtäessä oli liikenteen ja logistiikan toimialalla sekä osaamisen toimialoilla, joissa molemmissa oli käynnissä neljä suomalaista ja kaksi eurooppalaista hanketta. Toistaiseksi valmistavan teollisuuden alalla on yksi verkosto – FAMN Finnish Advanced Manufacturing Network, jonka omilla verkkosivuilla Make in Finland -kokonaisuudesta käytetään nimitystä ekosysteemi (Poikola ym., 2024; FAMN Finnish Advanced Manufacturing Network, ei pvm.).

MUKAAN DATA-AVARUUKSIIN?

Yritysten ja organisaatioiden kyvykkyyttä datan hyödyntämiseen on ryhdytty kuvaamaan pyramidina. Pyramidin perustukset muodostuvat yrityksen tai organisaation oman datan ja sen mahdollisuuksien tuntemisesta. Huipulla ovat data-avaruuDET – viimeisin kehitysmuoto datan jakamiselle. Ennen pyramidin huipulle tavoittelemista on tärkeää käydä läpi edeltävät kehitysvaiheet ja kartuttaa ymmärrystä datasta ja oman yrityksen kehitysasteesta. (Poi-kola ym., 2024.)



Oman datan ja kyvykkyyksien ymmärrys auttavat selvittämään edellytykset, tarkoituksenmukaisuuden ja riskit dataekosysteemeihin ja -avaruuksiin liittymiselle. Toisinaan tämä voi tarkoittaa kantapään kautta saatua oppia. Kokeilemalla esimerkiksi tekoälymalleja saatetaan havaita, että jokin tärkeä ennustava piirre ei ole yrityksen oman toiminnan kautta saatavilla vaan on toisen yrityksen hallussa. Vaihtoehtoisesti voidaan huomata, että tuotetulla datalla voi olla arvoa monen muun toimijan näkökulmasta.

Vaikka data-avaruuksien on kuvattu datan hyödyntämisen huipulle, aina ei ole tarkoituksenmukaista saavuttaa pyramidin huippua. Yrityksen toiminnalle voi riittää oman datan hyödyntäminen tai pienempi dataekosysteemi: verkosto, jossa data liikkuu osapuolia hyödyttävällä tavalla. Data-avaruuksissa ei myöskään tarvitse olla kiinteästi mukana, jotta siinä liikkuvasta datasta voisi hyötyä. Yritys voi olla loppukäyttäjä, joka hakee data-avaruudesta dataa mahdollista korvausta vastaan. Onpa lopputulos millainen tahansa, tilannekuva oman yrityksen tai organisaation tarpeisiin datan näkökulmasta tulisi olla selvä, jotta voitaisiin edes vastata kysymykseen, saadaanko data-avaruuksiin liittymisestä hyötyä.

DATA-AVARUUKSIEN TULEVAISUUS JA RISKIT

Euroopan yhteisen data-avaruuden edistämiseksi tehdään paljon töitä ja tämän tavoitteen saavuttamiseksi on asetettu selkeitä askelmerkkejä. Näistä ei kuitenkaan ole hyötyä, jos yksityiset henkilöt ja yritykset eivät halua jakaa dataansa. Epäluottamus ja epävarmuus tietojen päättymisestä väriin käsiin voivat nousta kynnyksysymykseksi.

Maailman turvallisuustilanne sekä aikaisemmat väärinkäytökset ja epäselvyydet horjuttavat datastrategialle tärkeää luottamusta. Jos mekanismeihin ei luoteta, saatetaan päätyä tilanteeseen, jossa dataa ei haluta jakaa tai jaetaan vain sellaista, millä ei saada merkittäviä innovaatioita aikaan. Vähän data-analytiikkaa ja tekoälyä soveltaneiden yritysten voi olla vaikea tunnistaa datan jakamiseen liittyviä liiketoimintariskejä. On myös vaikea hahmottaa, voiko data päätyä tahojen käyttöön, joiden ei sitä kuuluisi käyttää.

Tuntemattoman pelkoon ja väriin mielikuviin lääkkeinä toimivat tieto ja kokemus. Pienimuotoisenkin testaus ja datapohjaiset projektit saattavat avata silmiä uusille vaihtoehdoille ja riskeille, mitä datalla voidaan tehdä. Tämä voi osaltaan hälventää pelkoja, selvittää keinoja, miten riskejä vältetään tai vahvistaa näkemystä siitä, että data pitäisi ainakin toistaiseksi pitää itsellä. EU:n datastrategian sanoin tarvitaan tasapaino hyötyjen saavuttamiseen ja riskeihin vastaamiseen välille.

LÄHTEET

- Euroopan komissio. (2020). *Euroopan datastrategia*. <https://eur-lex.europa.eu/legal-content/FI/TXT/HTML/?uri=CELEX:52020DC0066>
- Euroopan komissio. (2024a). *Common European Data Spaces*. Haettu 25.11.2024 osoitteesta <https://digital-strategy.ec.europa.eu/en/policies/data-spaces>
- Euroopan komissio. (2024b). *Data Act explained*. Haettu 25.11.2024 osoitteesta <https://digital-strategy.ec.europa.eu/en/factpages/data-act-explained>
- Euroopan komissio. (2024c). *Data Governance Act explained*. Haettu 25.11.2024 osoitteesta <https://digital-strategy.ec.europa.eu/en/policies/data-governance-act-explained>
- FAMN Finnish Advanced Manufacturing Network. (Ei pvm.) *Make in Finland - The Manufacturing Industry Ecosystem*. Haettu 25.11.2024 osoitteesta <https://www.famn.fi/>
- Poikola, A., Verdonck, B. & Joosten, R. (2023). *DSSC Glossary* (Versio 1.0). Data Spaces Support Centre. Viitattu 25.11.2024. <https://dataspacesupportcentre.refined.site/space/Glossary/55443460/DS-SC+Glossary+%7C+Version+1.0+%7C+March+2023?attachment=%2Fdownload%2Fattachments%2F55443460%2FDSSC-Data-Spaces-Glossary-v1.0.pdf&type=application%2Fpdf&filename=DSSC-Data-Spaces-Glossary-v1.0.pdf>
- Poikola, A., Wong, D., Lähteenoja, V. & Turpeinen, M. (2024). *Suomalaisten data-avaruuksien tilannekuva*. Sitra. <https://www.sitra.fi/julkaisut/suomalaisten-data-avaruuksien-tilannekuva/>
- Sievers, H., Ulander, M., Helkala, I. & Saari, S. (2023). *Dataa hyödyntävien ekosysteemien ja verkostojen kehitysaste 2023*. Sitra. <https://www.sitra.fi/wp/wp-content/uploads/2023/09/sitra-dataa-hyodyntavien-ekosysteemien-ja-verkostojen-kehitysaste-kyselytutkimus-2023.pdf>
- Tilastokeskus. (2024a). *Tietotekniikan käyttö yrityksissä henkilöstön suuruusluokittain, 2022–2024*. Haettu 7.1.2025 osoitteesta https://pxdata.stat.fi/PxWeb/pxweb/fi/StatFin/StatFin__icte/statfin_icte_pxt_13vg.px/
- Tilastokeskus. (2024b). *Tietotekniikan käyttö yrityksissä: Tilaston dokumentaatio*. Haettu 5.3.2024 osoitteesta <https://stat.fi/tilasto/dokumentaatio/icte>

DATA-ANALYTIIKAN JAOTTELUA

TARJA AJO & MIKA RANTONEN,
JYVÄSKYLÄN AMMATTIKORKEAKOULU

Data-analytiikka on käsitteenä laaja ja siksi usein epämääräinen. Jotta voitaisiin paremmin ymmärtää, mitä data-analytiikalla missäkin yhteydessä tarkoitetaan, on tarpeen avata termistöä ja selvittää aiheen luokitteluja sekä ajatusmaailmaa näiden taustalla. Tässä artikkelissa keskitytään kuvaamaan liiketoimintalähtöistä jaottelua ja tämän jaottelun osa-alueita.

DATA-ANALYTIikka SATEENVARJOTERMINÄ

Data-analytiikka (eng. *data analytics*) -termiä käytetään kirjallisessa ja puhutussa kielessä usein suvereenisti sekaisin käsitteen data-analyysi (eng. *data analysis*) kanssa. Tämä ei ole ihme, sillä data-analytiikalle ei löydy yhtä vakiintunutta määritelmää, ja termipankeissa sille voidaan ehdottaa synonyymiä data-analyysi. Jos kuitenkin halutaan olla tarkkoja, data-analytiikkaa käytetään useimmiten ja enenevässä määrin sateenvarjotermiä, jolla viitataan laajempaan prosessiin datan keräämisestä ja esikäsittelystä datan analysointiin ja mahdollisiin jatkotoimenpiteisiin asti. Tavoitteista riippuen data-analytiikassa voidaan myös hyödyntää lukemattomia eri menetelmiä tutuista tilastotieteen tunnusluvuista tekoälyyn ja simulointiin asti. Datan analysointi voidaan siten nähdä olevan osa data-analytiikkaa.

Aiheen laajuuden takia toisinaan tarkennetaan, minkä aihealueen analytiikasta on kyse. Yrity maailmassa puhutaan liiketoiminta-analytiikasta (eng. *business analytics*) ja suurten datamäärien tapauksessa big data -analytiikasta. Näiden lisäksi data-analytiikkaa on tapana jaotella osa-alueisiin analytiikan luonteen mukaan. Vaikka luokittelut ovat osittain karkeita ja epämääräisiä, auttavat ne jäsentämään ja hakemaan tietoa sekä etsimään ratkaisuja ongelmiin valtavan tietomäärän seasta.

LUOKITTELUA TUTKIMUKSEN LUONTEEN JA RATKAISTAVAN ONGELMAN MUKAAN

Data-analytiikkaa ja data-analyysijä näkee jaoteltavan tutkimuksen luonteen ja tutkimusta ohjaavien kysymysten mukaan. Bolton (2022) esittää yleisesti data-analytiikalle kolme karkeata pääryhmää analyysien tavoitteiden ja analysoitavan datan mukaan.

Data-analytiikan kolme pääryhmää ovat:

- eksploratiivinen,
- konfirmatorinen ja
- kvalitatiivinen data-analyysi.

Luokista ensimmäistä, eksploratiivista data-analytiikkaa kuvataan lähestymistavaksi, jossa pyritään löytämään datassa piileviä rakenteita, yhteyksiä muuttujien välillä ja mahdollisia uusia näkökulmia tutkittavaan aiheeseen (Bolton, 2022). Haettaessa kokonais kuvaa ilmiöistä eksploratiivinen data-analytiikka on tärkeä, perustavanlaatuinen vaihe, jolla luodaan ymmärrystä ja pohjaa myöhemmälle tekemiselle. Tässä vaiheessa on tavallista, että dataa voidaan joutua keräämään, yhdistämään ja esikäsittämään useampaan kertaan tehtyjen havaintojen pohjalta, minkä vuoksi ei puhuta ainoastaan eksploratiivisesta datan analysoinnista vaan analytiikasta.



Konfirmatorisessa data-analytiikassa haetaan varmistusta havainnoille ja testataan perusjoukolle asetettuja hypoteeseja. Tällainen tutkimus lienee tutuinta biologiasta ja lääketieteestä, missä tehdyille oletuksille ja havainnoille on tarvetta hakea tukea tilastollisin menetelmin. Tilastollinen testaus voi esimerkiksi koskea jakaumien sijaintia tai muuttujien riippuvuutta. Kvalitatiivisella analyysillä puolestaan tarkoitetaan ei-numeerisen datan analysointia, jonka pohjalta voidaan edelleen tehdä päätöksiä (Bolton, 2022). Välttämättä laadullinen analyysi ei tarkoita ihmisen tekemää tekstien tulkintaa, vaan esimerkiksi asiakaspalautteita voidaan yrittää luokitella ohjelmallisesti positiivisiksi ja negatiivisiksi sisällön mukaan.

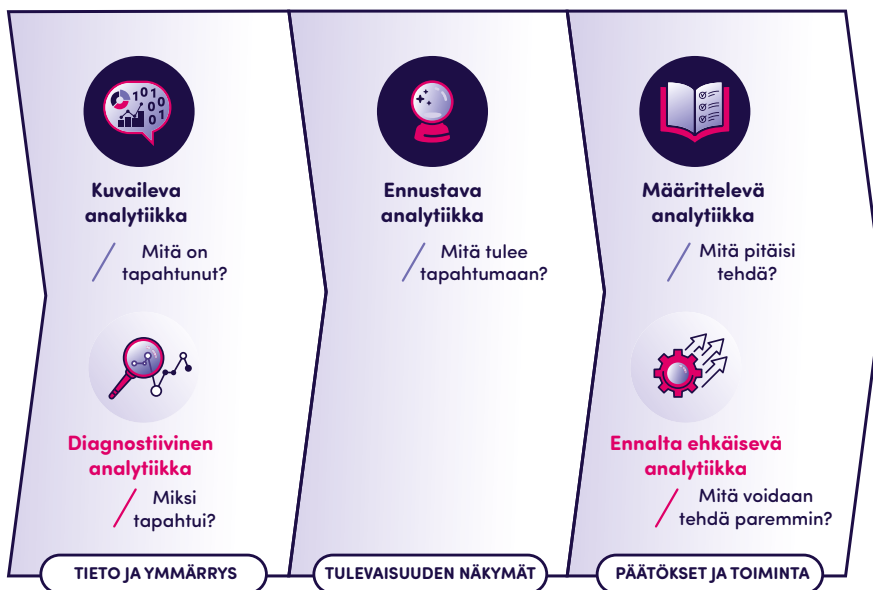
Boltonin (2022) esittämän luokittelun ohella data-analytiikkaa jaotellaan myös ratkaistavan ongelman mukaan. Tämä luokittelutapa kumpuaa yrityskentältä ja lienee helpommin lähestyttävä, sillä tässä data-analytiikan tyypeille voidaan tunnistaa selkeitä kysymyksiä, joihin ryhmän menetelmillä pyritään vastaamaan. Sivarajah, Kamal, Irani ja Weerakoddy (2017) laajassa kirjallisuuskatsauksessa big datan analysointiin tarkoitetuille menetelmille tunnistettiin kolme selkeää pääryhmää, joihin menetelmät voitiin luokitella:

Dig datan analysointiin tarkoitettujen menetelmien pääryhmät ovat:

- deskriptiivinen (eng. *descriptive*), kuvaileva analytiikka – mitä on tapahtunut?
- prediktiiivinen (eng. *predictive*), ennustava analytiikka – mitä tulee tapahtumaan?
- preskriptiivinen (eng. *prescriptive*), määrittelevä analytiikka – mitä pitäisi tehdä?

Näiden lisäksi kirjallisuudesta nousi esille muita yritystoiminnasta lähtöisin olevia ongelmamuotoja, joihin ei kuitenkaan ollut tunnistettu omia pääryhmiä erillisiä menetelmiä ja jotka johdonmukaisuuden vuoksi saatetaan katsoa kuuluvan osaksi pääryhmää. Nämä luokat ovat diagnostiivinen (eng. *diagnostic*) ja ennalta ehkäisevä (eng. *pre-emptive*) analytiikka, joista diagnostiivisen analytiikka usein luetaan osaksi kuvailevaa analytiikkaa ja ennalta ehkäisevä osaksi määrittelevää analytiikkaa. (Savarajah ym., 2017; Lepeniöti ym., 2020.)

Edellä luetellut pää- ja alaryhmät voidaan järjestää Kuvion 1 mukaisesti vaiheittaiseksi prosessiksi, missä edetään tiedon ja ymmärryksen keräämisen pohjalta ennusteisiin ja lopulta päätösten tukemiseen ja toimintavaihtoehtojen selvittämiseen.



Kuvio 1. Data-analytiikan pääryhmät ongelmakuvauksen mukaan jaoteltuina.

DESKRIPTIIVINEN JA DIAGNOSTIIVINEN ANALYTIikka

Deskriptiivisessä eli kuvailevassa analytiikassa mielenkiinnon kohteena ovat menneet tapahtumat sekä prosessien ja asioiden nykytila: Mitä on tapahtunut? Miten asiat ovat olleet ja miten ne ovat nyt? Mikä on ollut muutoksen suunta? Näihin kysymyksiin pyritään vastaamaan historia- ja nykydatan pohjalta koostamalla tietoa tunnusluvuiksi ja kuvaajiksi. Dataa analysoimalla voidaan myös pyrkiä tunnistamaan trendejä, poikkeavuuksia, kausivaihtelua ja syklistä vaihtelua. Kuvailevaa analytiikkaa kutsutaan joskus yksinkertaisimmaksi data-analytiikan muodoksi, sillä ilmiöiden kuvailussa päästään ainakin alkuun yksinkertaisilla tilastollisilla tunnusluvuilla sekä tietoa koostavilla kuvaajilla (Sivarajah ym., 2017). Myös datassa piilevät trendit on usein helppoja määrittää joko kuvaajista tai erilaisten apuohjelmien avulla. Kuitenkin piilossa olevien rakenteiden tai ryhmien tunnistaminen voi vaatia edistyneempiä menetelmiä, ja tässä voidaan hyödyntää tekoälyä.

Diagnostiivisessa analytiikassa puolestaan pyritään selvittämään ilmiöiden syitä ja muuttujien välisiä yhteyksiä. Tätä analytiikan tyyppiä voidaan pitää seuraavana vaiheena kuvailevan analytiikan jälkeen, ja useimmiten seuraava looginen kysymys asioiden nykytilan kuvaamisen jälkeen onkin kysyä, miksi näin on tapahtunut tai mitkä tekijät ilmiöön vaikuttavat. Koska deskriptiivinen ja diagnostiivinen analytiikka liittyvät vahvasti toisiinsa ja harvemmin tyydyttään vain asioiden menneen tai nykytilan kuvailuun, näitä kahta ei välttämättä eroteta toisistaan. Yhdessä näiden voi katsoa pitkälti vastaavan aikaisemmin kuvattua eksploratiivista data-analytiikkaa.

Konkreettisesti diagnostiivinen analytiikka voi olla ihmisen tekemää tunnuslукujen ja kuvaajien tulkintaa ja päätelmien tekemistä. Se voi olla myös korrelaatioiden laskemista tai pääkomponenttianalyysin tekemistä, yleisesti ottaen yhteyksien tutkimista. Vaikka diagnostisessa analytiikassa ei varsinaisesti tavoiteltaisi ennusteiden tekemistä, voidaan aineistoon pyrkiä sovittamaan malli ja katsoa, miten hyvin malli tai sen sisältämät tekijät selittävät ilmiötä. Päinvastoin mallista harvemmin saadaan hyvää, mikäli aiheen ja selittävien tekijöiden ymmärtäminen on heikkoa. Tämän vuoksi deskriptiivistä ja diagnostiivista analytiikkaa on hyvä tehdä, jotta pohja ennusteille ja toimenpidesuosituksille on vahva.

PREDIKTIIVINEN ANALYTIikka

Prediktiivinen, toisin sanoen ennustava analytiikka, tähtää tulevien tapahtumien ennustamiseen. Käytännössä tämä voi tarkoittaa esimerkiksi tapahtumien suunnan ennustamista (kasvu/lasku), asiakkaan päätöksen ennustamista pysyä asiakkaana tai siirtyä toisen palvelun pariin, tai numeerisia arvojen, kuten sähkönkulutuksen arviointia. Ennusteiden pohjalle tarvitaan ilmiöön liittyvää historiadataa, josta tunnistetaan ennusteeseen vaikuttavia tekijöitä: trendejä, kausivaihtelua ja muita muuttujia.

Yksi tunnetuimmista tavoista tehdä ennusteita on regressioanalyysi. Regressioanalyysi on yksi tilastollisista analyyseistä, jossa ennustettava muuttuja kirjoitetaan selittävien muuttujien lineaarisena lausekkeena. Muotoiltua lauseketta voidaan käyttää ennustamiseen, mikäli selittävien muuttujien arvot ovat tiedossa. Tilastollisten analyyseiden lisäksi keinoja ennusteiden tekemiseen löytyy koneoppimisen ja todennäköisyysmallien saralta. Erityisesti koneoppiminen tarjoaa monia mahdollisuuksia ennusteiden tekemiseen, joista esimerkkejä ovat päätöspuut, klusterointialgoritmit ja neuroverkot. Viime vuosina prediktiivisen analytiikan tutkimus onkin pitkälti keskittynyt koneoppimismenetelmien soveltamiseen, tilastollisten analyyseiden ja todennäköisyysmallien jäädessä taka-alalle (Lepeniöti ym., 2020).

PRESKRIPTIIVINEN ANALYTIikka

Preskriptiivinen analytiikka on analytiikan osa-alue, jossa pyritään tukemaan päätöksentekoa etsimällä parasta toimintavaihtoehtoa vallitsevassa tilanteessa. Lepeniotin ja muiden (2020) mukaan suurin hyöty preskriptiivisestä analytiikasta saadaan yhdessä prediktiivisen analytiikan kanssa, kun tehtyjen ennusteiden pohjalta selvitetään toimintavaihtoehtoja. Mikäli optimoituja päätöksiä on mahdollista tehdä etukäteen, voidaan pyrkiä vaikuttamaan itse tapahtumaan, valmistautua siihen ja oikea-aikaisesti kohdistaa tarvittavia toimenpiteitä. Verrattuna kuvailevaan ja ennustavaan data-analytiikkaan preskriptiivinen analytiikka on uudempi data-analytiikan osa-alue ja sitä on tutkittu vähemmän. Preskriptiivisen analytiikan tutkimuksesta Lepenioti ja muut (2020) toteavat, että toistaiseksi valtaosa tutkimuksista on keskittynyt matemaattiseen ohjelmointiin ja erityisesti optimointimenetelmiin ja -algoritmeihin.

Selittävä tekijä preskriptiivisen analytiikan tutkimuksen vähyyteen suhteessa kuvailevaan ja ennustavaan analytiikkaan löytyy siinä, että edeltävissä askeleissa – datan keräyksessä ja laadussa sekä kuvailevassa ja ennustavassa analytiikassa – on yrityskentällä vielä harppauksia otettavana. Niin Suomessa kuin monessa muussakin maassa pienissä ja keskisuurissa yrityksissä ollaan toistaiseksi opettelemassa datan hyödyntämistä ja selvittämässä sen mahdollisuuksia. Myös suurissa yrityksissä ongelmia voi olla tehtyjen ennusteiden tarkkuudessa, ja näiden parantamiseksi keskitytään tarvittavan datan hankkimiseen tai datan laadun parantamiseen. Erityisesti jos tavoitellaan preskriptiivisen analytiikan tekemistä ennusteiden pohjalta, tehtyjen ennusteiden tulisi olla riittävän tarkkoja, jotta niiden pohjalta olisi mielekästä tehdä preskriptiivistä analytiikkaa.

LÄHTEET

Bolton, J. (2022). *Data-analytics: Techniques and Applications*. States Academic Press.

Lepenioti, K., Bousdekis, A., Apostolou, D. & Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, 50, 57–70. <https://doi.org/10.1016/j.ijinfomgt.2019.04.003>

Sivarajah, U., Kamal, M. M., Irani, Z. & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70, 263–286. <https://doi.org/10.1016/j.jbusres.2016.08.001>

KÄDET SAVEEN - DATAN HYÖDYNTÄMINEN KÄYTÄNNÖSSÄ



TIEDON JALOSTAMINEN RAPORTOITAVAAN MUOTOON

JUHA STENFORS, SATAKUNNAN AMMATTIKORKEAKOULU

Tietovarastot ovat keskeinen osa nykyaikaista tiedolla johtamista. Ne mahdollistavat tiedon keräämisen, yhdistämisen ja analysoinnin eri lähteistä, jotta voidaan tuottaa arvokasta tietoa päätöksenteon tueksi. Tiedon jalostaminen alkaa raakadatan keräämisestä eri lähteistä ja päättyy valmiiden, raportoitavien tietojen esittämiseen. Seuraavassa käydään läpi, miten tietoa jalostetaan eri lähteistä raportoitavaan muotoon yritysten näkökulmasta.



TIEDON LÄHTEET JA HAASTEET

Data, joka on käyttökelpoista yritysten kannalta, on aika usein monessa eri lähteessä. Pitää ratkaista, miten tietoa saadaan siirrettyä tietovarastoon ja miten dataa pitää muokata tai rikastaa ennen kuin se on raportointiin vaadittavassa muodossa.

Tietoa voidaan hakea esimerkiksi seuraavanlaisista lähteistä:

- Sisäiset järjestelmät: ERP, CRM, taloushallinnon järjestelmät, HR-järjestelmät
- Ulkoiset tietolähteet: avoin data, markkinadata, sidosryhmien järjestelmät
- Manuaaliset lähteet: kyselyt, manuaalisesti ylläpidetyt Excel-tiedostot
- IoT- ja sensoridata: fyysisten laitteiden keräämä data

(Tietokaira, ei pvm.)

Tiedon kerääminen ja sen käyttö voi olla joissakin tapauksissa haasteellista. Yritysten omissa ympäristöissä olevat tietolähteet saattavat olla helpommin käyttöönotettavissa, mutta joissain tilanteissa näissäkin kohdataan haasteita.

Voi esimerkiksi olla niin, ettei järjestelmä mahdollista tiedon saantia rajapintojen avulla. Rajapintojen olemassaolosta huolimatta voi tulla tilanteita, että ne eivät tarjoa kaikkea tarvittavaa tietoa. Tietokanta voi olla niin vaikeaselkoinen ja dokumentoimaton, ettei tiedon saanti ole mahdollista. Nämä asiat voivat korostua varsinkin tilanteissa, joissa järjestelmä on ostettu palveluna eikä tätä tiedon saantia olla huomattu varmistaa sopimuksin. Lisäksi vastaan voi tulla tilanteita, ettei tietoa ole mahdollista saada muutoin kuin raporttien avulla.

ETL-PROSESSI

ETL-prosessi (*Extract, Transform, Load*) on keskeinen menetelmä tiedonhallinnassa, erityisesti tietovarastojen rakentamisessa ja ylläpidossa.

ETL-prosessi koostuu kolmesta päävaiheesta:

- Tiedon kerääminen (Extraction)
- Tiedon muuntaminen (Transformation)
- Tiedon lataaminen (Loading)

Tähän prosessiin on käytettävissä erilaisia välineitä. ETL-työkaluja ovat esimerkiksi Azure Data Factory, Talenda ja Informatica. Pilvipalveluita Google BigQuery, AWS ja Microsoft Synapse. Tietokantaratkaisuja on esimerkiksi SQL Server, PostgreSQL ja Snowflake. Lisäksi tätä vaihetta voidaan tehdä olemassa olevilla skriptikielillä tehdyillä kokonaisuuksilla kuten Python ja käyttämällä tietokantoja tiedon varastointiin. Valittu ratkaisumalli riippuu usein siitä, kuinka isosta tietovarastosta on kysymys. (IBM, ei pvm.)

Tiedon keräämisen vaiheessa, dataa kerätään eri lähdejärjestelmistä, kuten tietokannoista, sovelluksista tai tiedostoista. Tämän poiminnan tavoitteena on hakea tarvittava data haluttuja käyttötarkoituksia varten.

Muuntamisessa poimitulle datalle suoritetaan tarvittavat muunnokset, kuten puhdistus, yhdistäminen ja uudelleenjärjestely, jotta se saadaan yhtenäiseen ja analysoitavaan muotoon. Tämä vaihe varmistaa datan laadun ja yhteensopivuuden kohdejärjestelmän kanssa. Vaihe on myös todella tärkeä, koska tällöin on hyvä varmistaa, että dataa on ja se on muutettu mahdollisimman lähellä sitä muotoa, että se on raportoinnissa hyvin käytettävissä.

DATAN LAADUN TARKASTUS

Kun data tulee monista lähteistä, datan laatu ja yhdenmukaisuus on varmistettava ennen raportointia. Datan laadun varmistus on tärkeää aloittaa datan puhdistamisesta, joka on virheellisten, puuttuvien tai ristiriitaisten tietojen korjaamista.

Tarkastusta jatketaan standardoimalla data. Tiedoissa voi esiintyä erilaisia muotoiluja esimerkiksi valuutoille, päivämäärille, tunnuksille ja koodeille, jotka on hyvä saattaa yhtenäiseen muotoiluun. Datassa esiintyviä muotoiluja voidaan muokata myöhemminkin raportointityökaluissa, kuten PowerBI:ssä, jossa on datalle hyvät muokkausmahdollisuudet.

Keskeisten tietojen yhdenmukaisuus pitää varmistaa ja tätä kutsutaan termillä *Master Data Management* (MDM). Keskeisillä tiedoilla tarkoitetaan muun muassa asiakkaita, tuotteita, organisaatioita ja koodeja. Tätä työtä voidaan myös tehdä raportointivälineessä korvaustyökalulla, jos koetaan, että se on riittävää eikä tietovarastoon ole tarvetta tehdä tällaisia muutoksia. Jos kuitenkin dataa käytetään muuallakin, joissa tuota mahdollisuutta ei ole, niin silloin tieto parasta muuttaa jo tässä vaiheessa. (lab, 2024.)

Yleisimpiä haasteita data laadussa ovat:

- Puuttuva data voi johtaa virheellisiin johtopäätöksiin
- Datan päällekkäisyys voi aiheuttaa raportointiongelmia
- Tiedon päivityssyklit voivat vaihdella eri järjestelmissä

(lab, 2024.)

Nämä kaikki yhteensä tai erikseen voivat tuottaa suuriakin ongelmia raportoinnin luotettavuudessa. Puuttuvaa dataa voi olla haastavaa tarkistaa, jos datamäärät ovat suuria. Tällöin pitää vain tehdä mahdollisimman paljon ja laajasti datan tarkastuksia. Päällekkäisyydet voivat olla myös samankaltaisia ongelmia, ja tämän todentaminen vaatii samaa tarkistelua kuin puuttuvan datan osalta.

Tiedon päivityssyklit pitää miettiä tilanteen mukaan. Aika useasti massa-ajoyöllä voi olla riittävä taso. Kuitenkin jos raportoinnin tarve vaatii tiheämpää päivityssykliä, täytyy toimia tämän tarpeen mukaan. Joskus myös tietolähteet voivat asettaa rajoja tiedonsiirrolle. Esimerkiksi järjestelmien rakenne tai resurssit saattavat olla sellaisia, etteivät ne mahdollista raportointitietovaraston tiedon jatkuvaa päivittämistä.

TIEDON MUOKKAAMINEN JA JALOSTUSPROSESSI

Periaatteessa datan laaduntarkastuksessa tehdään jo eräänlaista muokkaamista ja jalostamista. Kuitenkin edellisten lisäksi voi olla tarvetta jalostaa dataa muullakin tavoin. Yksi osa jalostamista on datan normalisointi, jossa muutujan arvot skaalataan välille nolla (0) ja yksi (1).

Edellisen tyyppisten skaalausten lisäksi voi olla tarvetta jalostaa tietoa. Tiedon jalostamista voidaan tehdä tietovarastossa eri taulujen välillä tai sitten vaikka täysin eri tietolähteiden välillä esimerkiksi Excel-tilit tai vastaavat. Tiedon jalostamiseksi voidaan lukea myös tiedon aggregointi. Tällöin luodaan erilaisia raportointeja helpottavia laskennallisia kenttiä. Tällaisia ovat muun muassa summa-sarakkeet, myynnit per kuukausi, asiakkaan keskimääräinen ostosten summa ja ostettujen tuotteiden lukumäärä. Vastaavien tietojen lisäksi raportointia varten voi olla tarve tallentaa mediaani-, minimi- ja maksimiarvoja. Monissa raportointityökaluissa on kuitenkin toimintoja laskennallisten tietojen muodostamiseksi, minkä vuoksi näiden etukäteen laskemisen tarpeellisuutta kannattaa harkita käytön mukaan. (Niemi, 2024, s. 11.)

ERILAISET TIETOVARASTORATKAISUT

Keskeisessä roolissa tiedon tallennuksessa raportoinnin tarpeisiin ovat tietovarastot. Ne mahdollistavat suuren datamäärän tallentamisen, käsittelyn ja analysoinnin tehokkaasti. Tarpeiden ja datan määrän mukaan pitää valita sovelia tietovarastoratkaisu. Seuraavassa lyhyesti muutamia vaihtoehtoja.

Perinteiset relaatiotietokantapohjaiset tietovarastot (Data Warehouses, DWH)

- rakennettu relaatiotietokantojen (SQL) päälle
- data tallennetaan rakenteellisessa muodossa (taulukot, indeksit, avaimet)
- hyvä historiallisen datan säilyttämiseen ja tehokkaaseen raportointiin.

Pilvipohjaiset tietovarastot (Cloud Data Warehouses)

- pilvipohjaisia tietovarastoja, jotka skaalautuvat dynaamisesti
- voi sisältää sekä relaatiotietoja että jäsentämätöntä dataa
- hyvä vaihtoehto yrityksille, jotka haluavat välttää kalliit paikalliset laitteistot.

Datalake-ratkaisut (Data Lakes)

- tallentaa jäsentämätöntä, puolijäsentynyttä ja jäsenneltyä dataa (esim. sensoridata, lokit, kuvat)
- voidaan yhdistää tietovarastoon, jolloin raportointia ja Big Data -analytiikkaa voidaan yhdistää.

Hybridiratkaisut (Data Lakehouse, HTAP)

- yhdistää datalaken ja perinteisen tietovaraston ominaisuudet
- tuki sekä Big Data -analytiikalle että perinteisille raportointiratkaisuille
- voi mahdollistaa reaaliaikaisen analytiikan ja nopean raportoinnin samassa järjestelmässä. (Unite.AI, 2023)

YHTEENVETO

Tiedon kerääminen ja varastointi ovat moninainen kokonaisuus, jonka suunnittelu ja toteutus vaativat paljon resursseja. Jos yrityksessä on epävarmuutta siitä, mitä ja miten kokonaisuus pitäisi toteuttaa, on suositeltavaa ostaa konsultointi- ja tarvittaessa myös toteutusapua. Tällä tavoin päästään varmemmin hyvään lopputulokseen.

LÄHTEET

Iab. (2024). *Datan laadun arviointi: Tekniikoita ja käytäntöjä*. Haettu 25.3.2025 osoitteesta <https://www.iab.fi/uutiset-blogi/markkinaluvut/datan-laadun-arviointi-tekniikoita-ja-kaytantoja.html>

IBM. (Ei pvm.). *What is ETL (extract, transform, load)?* Haettu 24.3.2025 osoitteesta <https://www.ibm.com/think/topics/etl>

Niemi, A. (2024). *Datan validointi- ja normalisointimenetelmät modernissa tietovarastoinnissa* [opinnäytetyö, Hämeenlinnan ammattikorkeakoulu]. Theseus-julkaisuarkisto. <https://urn.fi/URN:NBN:fi:amk-2023082825020>

Tietokaira. (Ei pvm.). *Tietovarastointi ja tietokannat*. Haettu 24.3.2025 osoitteesta <https://tietokaira.fi/palvelut/tiedonhallinta/>

Unite.AI. (2023). *Aloittelijan opas tietovarastointiin*. Haettu 27.3.2025 osoitteesta <https://www.unite.ai/fi/Tietovarastoinnin-aloittelijaopas/>

DATAN VARASTOINNIN TERMIT HALTUUN

MIKA RANTONEN, JYVÄSKYLÄN AMMATTIKORKEAKOULU

Datan määrän räjähdysmäinen kasvu tuo haasteita datan tallentamiselle perinteisin menetelmin ja datan jatkojalostamiselle. Markkinoille on tullut paljon erilaisia ja eri teknologioilla olevia ratkaisuita, joiden käyttötarkoituksiin ja ominaisuuksiin perehdytään tässä artikkelissa.

DATAMÄÄRÄN KASVU

Digitalisaation myötä digitaalisesti tallennetun datan eli tiedon määrä on kasvanut räjähdysmäisesti viimeisen vuosikymmenen aikana ja se edelleen jatkuu. Viime vuosikymmenenä on noussut esiin termi Big Data (massadata), jolla viitataan erittäin suuriin, monimutkaisiin ja nopeasti kasvaviin tietomääriin (Gupta & Giri, 2018). Datan määrän kasvu luo haasteita perinteisille tietojenkäsittelymenetelmille, joten uudentyypisille tiedon varastointijärjestelmille on tarvetta.

Tiedon varastointi on prosessi, jossa tieto eli data tallennetaan järjestelmällisesti ja helposti käsiteltävässä muodossa erilaisiin järjestelmiin. Tässä vaiheessa törmätään melkoiseen sanaviidakkoon, jossa nimet ja termit vaihtelevat suuresti. Tämän artikkelin tarkoitus on avata lukijalle terminologiaa ja päästä selvyteen eri termien välillä.

DATA

Data voi olla monenlaisessa muodossa, ja se voi tulla monenlaisesta eri lähteestä. Data voidaan kategorisoida sen rakenteen perusteella seuraavasti (Lehmäheiu ym., 2018, s. 13):

1. Strukturoitu

- Strukturoitu data on hyvin järjestettyä ja tallennettu ennalta määriteltymiin muotoihin, kuten taulukoihin
- Helposti käsiteltävissä ja analysoitavissa koneellisesti

Esimerkkejä:

- Excel-tilit, joissa on asiakkaiden nimiä, osoitteita ja tilausnumeroita
- Tietokantataulukset, joissa on esimerkiksi tuotteiden varastotiedot

2. Puolistrukturoitu (semistructured)

- Puolistrukturoitu data sisältää joitakin rakenteellisia elementtejä, mutta ei ole täysin järjestettyä. Se voi sisältää tunnisteita tai merkintöjä, jotka helpottavat tietojen järjestämistä ja analysointia.

Esimerkkejä:

- PDF-, logi-, XML- ja JSON-tiedostot, joissa on tietoja hierarkisessa muodossa.
- Sähköpostit, joissa on sekä strukturoituja osia (esim. lähettäjä, vastaanottaja, päivämäärä) että strukturoimattomia osia (viestin sisältö).

3. Täysin rakenteetonta eli strukturoimatonta

- Strukturoimatonta dataa ei noudata mitään ennalta määriteltyä rakennetta, mikä tekee sen käsittelystä ja analysoinnista haastavaa.

Esimerkkeinä binääriset tiedostot kuten kuvat, ääni tai videot.

TIETOKANNAT

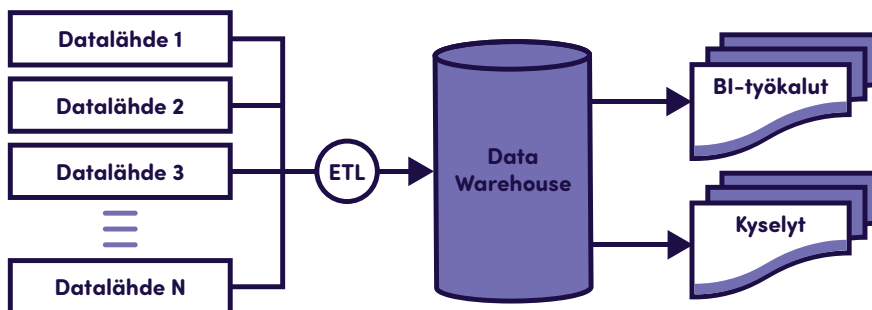
Edellä puhuttiin tiedosta eli datasta ja sen tallentamisesta. Data tallennetaan digitaalisessa muodossa yleisimmin tietokantaan, jota voidaan pitää järjestettynä kokoelmana dataa. Tietokantoja on erilaisia riippuen niiden käyttötarkoituksesta ja millaista dataa niihin tallennetaan. Tietokantaratkaisuilla on erilaisia ominaisuuksia: toisissa korostuu nopeat haut ja päivitykset, toisissa tietyntyyppisen datan käsittely. Tämän takia tietokannat voidaan jakaa relaatiotietokantoihin ja NoSQL- tietokantoihin (Lemahieu ym., 2018, s. 30). Relaatiotietokanta koostuu tauluista. Taulukossa on rivejä ja sarakkeita, jotka vastaavat tietynlaista tietoa. Esimerkiksi asiakastietokannassa voi olla taulu, jossa on sarakkeet nimelle, osoitteelle ja puhelinnumerolle. Jokainen asiakas on yksi rivi tässä taulussa. Esimerkkejä relaatiotietokannoista on MySQL, PostgreSQL ja Oracle Database. NoSQL- tietokannoissa käytetään erilaisia tietomalleja, ja tietokannan rakenne on joustavampi erilaisien datojen tallentamiseen. Esimerkkejä näistä on MongoDB, Cassandra ja Redis. Tietokannat käyttävät **tietokantajärjestelmiä**, jotka ovat ohjelmistoja, joiden avulla tietokantaa hallitaan.

DATA WAREHOUSE

Data warehouse (datavarasto/tietovarasto) syntyi aivan 1980-luvun lopulla ja 1990-luvun alussa perustuen useisiin erilaisiin tiedonhallintateorioihin, jotka yhdistyivät. IBM:n tutkijoiden Barry Devlinin ja Paul Murphyn uskotaan olleen ensimmäiset, jotka loivat ”liiketoimintatietovaraston” 1980-luvun lopulla (Devlin ym., 1988). Kun ensimmäiset yritykset hajautettujen tietokantojen hallintajärjestelmiin kariutuivat, Bill Inmon määritteli tietovarastoinnin säännöt, jotka saivat jalansijaa koko 1990- ja 2000-luvun ajan. Samaan aikaan myös Ralph Kimball antoi merkittävän panoksen tieteenalalle, erityisesti tietovarastojen suunnitteluun ulottuvuusmallinnuksen (*dimension modelling*) avulla. Käsite tietovarasto on sittemmin kehittynyt sellaiseksi kuin se on nykyään, monimutkaiseksi järjestelmäksi, joka on kasvanut tietotekniikan kehittyessä.

Tietovarasto ei ole sama asia kuin tietokanta, koska siinä datat yhdistetään useista eri lähteistä yhdeksi tietovarastoksi. Tietovarasto hakee tietoja eri datalähteistä, ja tieto kulkee erilaisten käsittelyjen jälkeen hyvin mallinnettuna, käsiteltynä ja jäsennettynä tallennettavaksi (Gupta & Giri, 2018). Tallennusmuoto eli skeema pitää määritellä etukäteen, mikä tekee tästä huonosti muokautuvan. Yksinkertaisesti sanottuna tietokanta tallentaa tietoja, kun taas tietovarasto voidaan ajatella ympäristöksi tai järjestelmäksi, joka yhdistää useita fyysisiä tietokantoja yhdeksi loogiseksi kokonaisuudeksi. Mitä enemmän tietolähteitä on, sitä monimutkaisemmaksi tietovarasto tulee. Tietovaraston tiedonsiirrossa käytetään yleensä ETL-menetelmää, joka tarkoittaa datan siirtämistä,

muokkaamista ja lataamista. Menetelmässä tiedot haetaan (*Extract*) lähdejärjestelmästä, niitä muokataan (*Transform*) ja ladataan (*Load*) lopulta data tietovarastoon (ks. Kuvio 1). Latausprosessissa tiedot muunnetaan tietovaraston rakenteen muotoon integroiden samalla eri lähtöjärjestelmien tietoja. Nykyään suosituimpia tietovarastotoimittajia ovat muun muassa Amazon Redshift, Google BigQuery, Snowflake, Oracle ja IBM Db2 Warehouse.



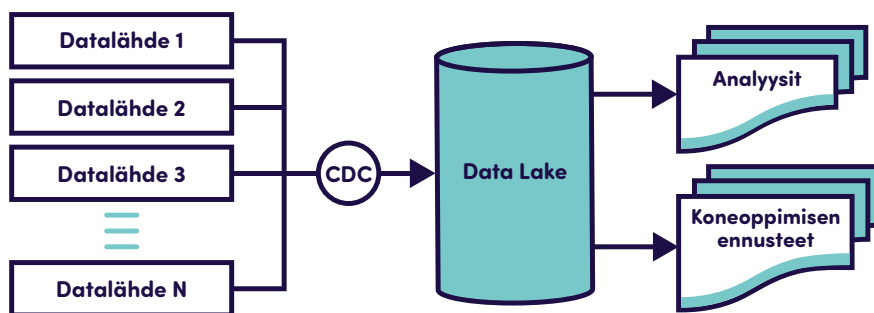
Kuvio 1. Tietovaraston tiedonsiirto prosessissa käytetään ETL-menetelmää.

Tietovarasto tarjoaa tietoa siitä, mitä yrityksen tiedot voivat tarkoittaa ja mahdollistaa tietoon perustuvan päätöksenteon. Esimerkiksi yrityksen johto tarkastelee tietovarastoon kerätyistä tiedoista liiketoiminnan tilannekuvaa ja mitä trendejä on mahdollisesti havaittavissa. Tilannekuvan pohjalta voidaan tehdä liiketoimintaan strategisia muutoksia, jotka voivat tarjota kilpailuetua ja lisätä yrityksen pitkäikäisyyttä. Tilannekuvan luomiseen käytetään BI (Business Intelligence) -työkaluja datan visualisointiin ja ymmärtämiseen.

DATA LAKE

Data lake (tietoallas/data-allas) termin keksi James Dixon vuonna 2010 (Foot, 2020), kun hän halusi vähentää datan siiloutumista. Tietoallas on keskitetty tietovarasto, jonne kerätään ja tallennetaan suuria määriä raakadataa, eli tieto on alkuperäisessä muodossa (Gupta & Giri, 2018). Tallentamisen jälkeen tietoa voidaan käsitellä ja tallentaa esikäsittelyyn uudestaan erilaisiin käyttötarkoituksiin kuten data-analytiikan, koneoppimisen tai syväoppimisen kehittämiseen. Tietoallas on skaalautuva, ja sinne tallennettu data voi olla strukturoitua, puolistrukturoitua tai täysin rakenteetonta (Gupta & Giri, 2018). Tietoallas on myös ketterä alusta, jonka tarkoituksena on säilyttää kaikki yrityksen tiedot keskitetysti yhdellä alustalla. Tämän avulla voidaan tehdä kattavia raportteja, visualisointeja ja analytiikkaa liiketoiminnan osalta.

Tietoallasta voidaan pitää tietyllä tavalla tietovaraston seuraajana. Useimmissa organisaatioissa tiedon jalostamisen perustana on edelleen tietovarastot, mutta ne käyttävät sitä yhdessä tekoälyn, koneoppimisen ja muiden kehittyneiden analyysien kanssa, joita tietoallas ohjaa. Tietoaltaassa tietojen haku on muutostietojana (*Change Data Capture, CDC*), jossa tunnistetaan ja tallennetaan tietokannassa oleviin tietoihin tehdyt muutokset ja toimitetaan nämä muutokset reaaliaikaisesti jatkojalostusprosessiin tai -järjestelmään (ks. Kuvio 2).



Kuvio 2. Tietoaltaan tiedonsiirto prosessissa käytetään CDC-menetelmää.

lhannetapauksessa organisaatioilla olisi oltava hyvin arkkitehtuuriltaan toimiva integraatio datavaraston ja tietoaltaan välillä, jotta vältetään tarpeeton pirstaloituminen näiden kahden välillä. Tietoaltaan etuna tietovarastoon on, että tallennusmuotoa eli skeema voidaan muuttaa, ja tämä tekee tietoaltaasta hyvin joustavan.

DATA LAKEHOUSE

Data LakeHouse (DLH) on yksinkertainen yhdistelmä tai data-arkkitehtuuri, jossa data-allas ja tietovarasto yhdistyvät. Data LakeHouse on avoimiin standardeihin perustuva tallennusratkaisu, joka on luonteeltaan monipuolinen. Data lakehouset mahdollistavat koneoppimisen, business intelligenen ja ennakoivan analytiikan, ja ne antavat organisaatioille mahdollisuuden hyödyntää edullista ja joustavaa tallennusta kaikentyyppisille tiedoille (strukturoidulle, strukturoimattomalle ja puolistrukturoidulle) tarjoten samalla tietorakenteita ja tiedonhallintaominaisuuksia.

Esimerkkeinä Data Lakehouseista ovat Databricks Lakehouse Platform ja Amazon Redshift Spectrum. Teknologioiden kehittyessä ja data lakehouse -ratkaisujen yleistyttyä on kuitenkin siirrytty siihen, että lakehouse-komponentteja ei enää kytketä tiettyyn tietoaltaaseen. Esimerkiksi Google Cloudin lähestymistapa on ollut yhdistää yritysten datatoimintojen, tietoaltaan ja tietovaraston keskeiset ominaisuudet. Tässä toteutuksessa BigQueryn tallennus- ja laskentatehot ovat data lakehouse -arkkitehtuurin ytimessä. Sen jälkeen voidaan soveltaa yhtenäistä hallintatapaa ja muita tietovaraston kaltaisia ominaisuuksia Dataplexin ja Analytics Hubin avulla. BigQuery ei ole ainoastaan integroitu Google Cloud -ekosysteemiin, vaan sen avulla voi myös käyttää kumppaneiden ja avoimen lähdekoodin teknologioita. Näin voidaan yhdistää allas- ja varasto-ominaisuuksien parhaat puolet yhteen järjestelmään.

LÄHTEET

Devlin, B. A. & Murphy, P. T. (1988). *An architecture for a business and information system*. IBM Systems Journal, 27(1), 60–80. <https://doi.org/10.1147/sj.271.0060>

Foote, K. D. (2020). *A Brief History of Data Lakes*. Haettu 9.4.2025 osoitteesta <https://www.dataversity.net/brief-history-data-lakes/>

Gupta, S. & Giri, V. (2018). Introduction to Enterprise Data Lakes. Teoksessa *Practical Enterprise Data Lake Insights: Handle Data-Driven Challenges in an Enterprise Big Data Lake* (s. 1–31). https://doi.org/10.1007/978-1-4842-3522-5_1

Lemahieu, W., vanden Broucke, S. & Baesens, B. (2018). *Principles of Database Management: The Practical Guide to Storing, Managing and Analyzing Big and Small Data*. Cambridge University Press.

DATAN VISUALISOINNIN SELKEYTTÄMINEN SADEPILVIKUVAAJILLA

TARJA AJO & JANNE ALATALO, JYVÄSKYLÄN AMMATTIKORKEAKOULU

Datan visualisointi on olennainen osa sekä datan analysointia että havaintojen viestintää. Selkeä ja objektiivinen viestintä voi kuitenkin olla haasteellista aineistoa kuvaavien tunnuslukujen tai yksittäisten kuvaajien avulla. Tässä artikkelissa kerrotaan sadepilvikuvaajista, joilla saadaan sisällytettyä tietoa eri näkökulmista samaan kuvaajaan.

Tilastolliset tunnusluvut ja yleisesti lukuarvot ovat tiivis tapa esittää informaatiota. Toisinaan nämä voivat kuitenkin olla harhaanjohtavia ja liian yleistäviä – numeroita, jotka peittävät taakseen tarpeellista tietoa. Mikäli aineistossa on paljon ääripäiden lukuarvoja, ei pelkkien tunnuslukujen esittäminen ole edes perusteltua, sillä näillä ei välitettäisi lukijalle oikeaa kuvaa todellisesta tilanteesta. Tällaisissa tilanteissa voidaan kuitenkin käyttää kuvaajia, joilla tunnuslukujen rinnalle tuodaan tarvittavaa lisäinformaatiota tilanteen tulkitsemiseen. Sanotaan, että kuva kertoo enemmän kuin tuhat sanaa, ja hyvien kuvaajien tapauksessa tämä pitää paikkansa. Visualisointi on merkittävä osa sekä datan analysointia että tulosten esittämistä. Valitulla visualisointitavalla voidaan vaikuttaa siihen, mitä kuvaajista voidaan lukea. Sadepilvikuvaajien (eng. *raincloud plot*) avulla tietoa voidaan esittää useasta eri näkökulmasta yhdellä kuvaajalla (Allen ym., 2021).

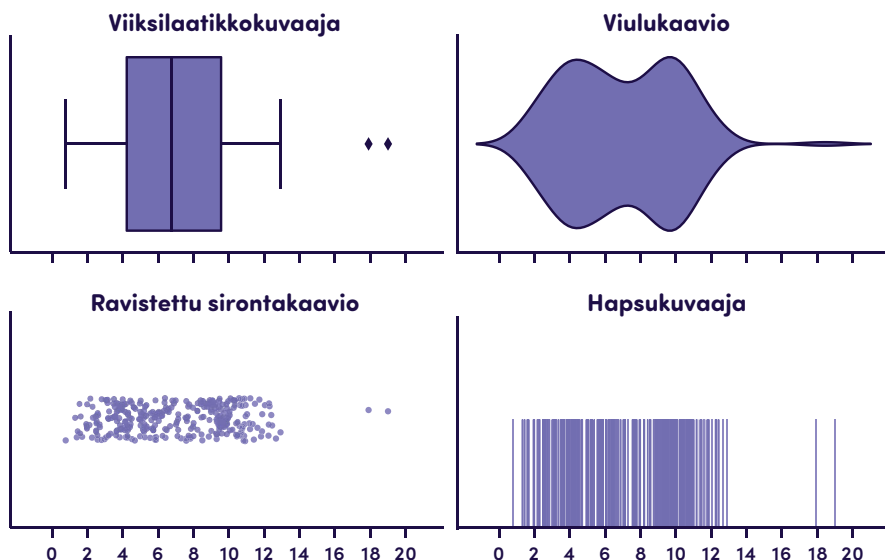
KUVAAJIEN KÄYTTÖ JA TAVOITTEET

Visualisointi on monella tapaa tehokas ja mielekäs tapa esittää tietoa: kokonaisuus, joka olisi tekstiksi kirjoitettuna tai sanallisesti selitettynä pitkä, epäselvä ja raskas, voidaan esittää kuvaajan avulla helposti ja nopeasti ymmärrettävässä muodossa. Kuvaajat ovat myös hyvin pitkälti käytetystä kielestä riippumaton tapa esittää havaintoja, mikä helpottaa kommunikaatiota monikulttuurisessa työyhteisössä. Siten kuvaajat toimivat hyvin sekä artikkeleiden että esitysten tukena. Vaikka kuvaajat olisi tarkoitettu tekstin tai esityksen tueksi, tulisi niiden olla ymmärrettäviä sellaisenaan. Sen lisäksi että nämä saattavat olla ainoita mitä teksteistä ja esityksistä luetaan, kuvaajia saatetaan käyttää uudelleen muualla tai ne voidaan tarkoituksella irrottaa asiayhteydestä, minkä vuoksi on tärkeää, että ne edelleen välittävät saman informaatio kuin alkuperäisessä asiayhteydessä.

Hyvän kuvaajan luominen ei kuitenkaan aina ole helppoa. Ensimmäinen ongelma on tavallisesti kuvaajan tyypin valinta: eri tyypeillä on omat hyvät ja huonot puolensa, ja tiedon esittäminen yhdellä tavalla voi antaa puutteellisen kuvan datasta. Seuraavaksi voidaankin päätyä miettimään, miten yhteen kuvaajaan saadaan kaikki tarpeellinen informaatio, jotta esityksestä ei tehtäisi vääriä päätelmiä. Yksi vaihtoehto on muodostaa useita kuvaajia, mutta tässä ongelmaksi muodostuu sekä kuvaajien keskinäinen vertailu ja kokonaisuuden hahmottaminen että edellä mainittu asiayhteydestä irrottaminen. Usein kuvaajien piirtäminen onkin tasapainoilua yksinkertaisuuden sekä informaation välittämisen ja tulkittavuuden välillä.

VIIKSILAATIKKOKUVAAJA, VIULUKAAVIO JA RAAKADATA: SADEPILVIKUVAAJA

Kuvio 1 esittelee erilaisia tapoja visualisoida sama data. Eri kuvaajatyypit viestivät saman asian eri tavoin korostamalla eri asioita. Viiksilaatikkokuvaaja (eng. *box plot*) ja viulukaavio (eng. *violin plot*) ovat erityisesti tilastotieteestä tuttuja yhden muuttujan tunnuslukuja ja sen jakaumaa kuvaavia kaavioita. Ravistettu sirontakaavio (eng. *jitter plot*) ja hapsukuvaaja (eng. *rug plot*) puolestaan visualisoivat raakadataa esittämällä jokaisen näytteen arvon kuvaajassa.

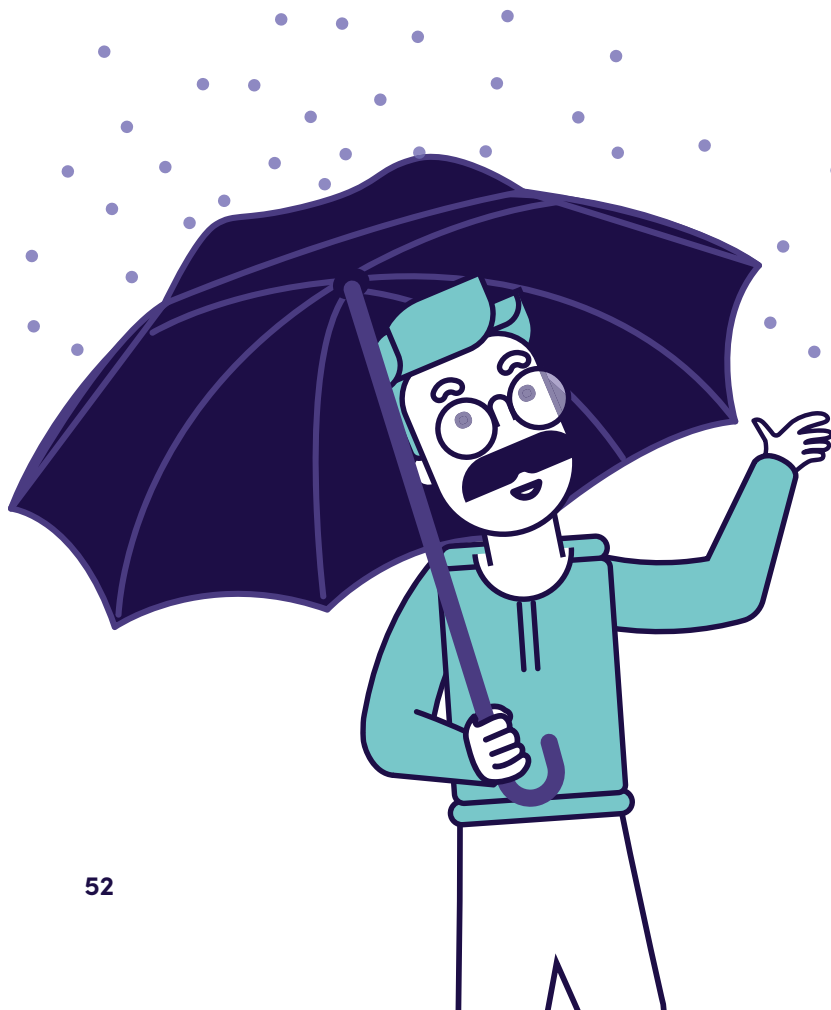


Kuvio 1. Neljä eri tapaa visualisoida sama data: viiksilaatikkokuvaaja, viulukaavio, ravistettu sirontakaavio ja hapsukuvaaja.

Viiksilaatikkokuvaajassa esitetään tilastolliset tunnusluvut: minimi ja maksimi, ylä- ja alakvartiilit, mediaani sekä usein myös laskennalliset oudokit (eng. *outlier*) visuaalisessa muodossa (Nuzzo, 2016). Viiksilaatikkokuvaajan etuja ovat sen selkeys ja luettavuus riippumatta havaintojen lukumäärästä – kuvaaja säilyy luettavana, oli havaintojen lukumäärä (n) sitten 50 tai 500. Viiksilaatikkokuvaajan huonoja puolet ovat kuitenkin samat kuin numeeristen tunnuslukujen kohdalla: samoihin laskennallisiin tunnuslukuihin voidaan päätyä eri havainnoilla.

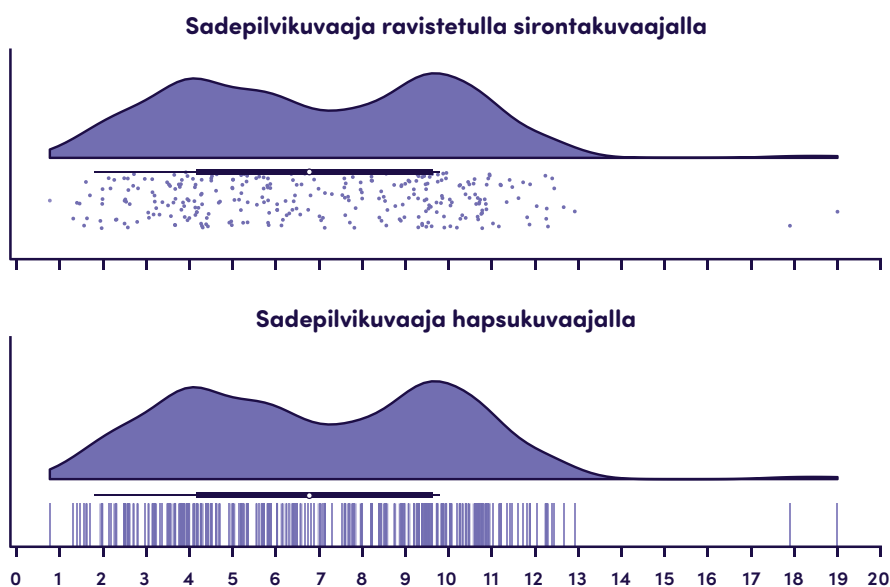
Viulukaavio puolestaan pyrkii havainnollistamaan muuttujan tiheysfunktioita, joka on mallinnettu otoksesta KDE-menetelmällä (eng. *kernel density estimation*) (Hintze & Nelson, 1998). Perinteisesti horisontaalisesti ja y-akselin positiiviselle puolelle piirretty tiheysfunktion kuvaaja on viulukaavioissa usein käännetty y-akselin suuntaiseksi ja peilattu x-akselin arvon molemmin puolin. Usein viulukaavioita esittävän osan sisällä on myös mukana pieni versio viiksilaatikkokuvaajasta, joka tuo lisätietoa datan ymmärtämiseen.

Tällaisten kuvaajien lisäksi on usein hyödyllistä tarkastella myös raakadataa, sillä se voi kielä sisäisistä rakenteista datassa. Yhden muuttujan raakadatan visualisointiin soveltuvia visualisointimenetelmiä ovat muun muassa edellä



olevassa kuvassa esitetyt ravistettu sirontakaavio- ja hapsukuvaaja. Ravistettu sirontakaavio perustuu normaaliin sirontakaaviomenetelmään, jossa toisen akselin arvo on satunnaistettu, jotta samassa kohtaa olevat näytteet eivät mene päällekkäin. Hapsukuviomenetelmässä näytteiden arvot on visualisoitu ohuella viivalla.

Kuvio 2 esittää kaksi esimerkkiä sadepilvikuvaajista. Sadepilvikuvaajassa yhdistetään viulu- ja viiksikaavioista sekä raakadatan saatava informaatio samaan kuvaajaan. Raakadatan visualisointiin voi käyttää edellä mainittuja ravistettua sirontakaaviota tai hapsukuvaajaa.



Kuvio 2. Kaksi eri versiota sadepilvikuvaajasta. Ensimmäisessä kuvaajassa on käytetty ravistettua sirontakuvaajaa raakadatan visualisoimiseen ja toisessa siihen on käytetty hapsukuvaajaa.

Sadepilvikuvaajan nimi tulee monen muun kuvaajatyypin tapaan sen ulkoasusta. Kuvaajassa viulukaaavion puolikas on asemoitu vaakasuoraan, jolloin tämä näyttää kumpupilveltä. Kun viulukaaavion, eli pilven alle vielä lisätään yksittäiset havainnot ravistetulla sirontakuvaajalla tai hapsukuvaajalla, näyttää kaavio sadepilveltä. Koska silmämääräisesti eroja vastaavien kuvaajien välillä voi olla vaikea erottaa, käy järkeen lisätä kuvaajaan vielä jakauman sijaintia kuvaavia tunnuslukuja viiksilaatikkokuvaajan muodossa.

YKSITYISKOHDILLA SELKEYTTÄ JA LUETTAVUUTTA KUVAAJIIN

Nyrkkisääntönä kuvaajien piirtämiselle annetaan usein yksinkertaisuus ja tehokkuus. Joskus puhutaan myös data-muste-suhteesta. Tällä tarkoitetaan sitä, että kuvaajassa mustetta tulisi käyttää vain datan kuvaamiseen ja välttää sen käyttämistä tarpeettomiin elementteihin. Tämä ei kuitenkaan tarkoita sitä, etteikö kuvaajaan voisi tai pitäisi lisätä pieniä yksityiskohtia, joilla helpotetaan tiedon lukemista ja tulkintaa.

Sadepilvikuvaajien tapauksessa lisättäviä yksityiskohtia voivat olla muun muassa ravistuksen määrä sekä läpinäkyvyys ja reunat yksittäisille pisteille. Erityisesti tilanteissa, joissa havaintojen määrä on suuri, nämä voivat parantaa raakadatan arvojen erottuvuutta. Jos taas saman muuttujan jakaumia halutaan vertailla esimerkiksi eri ryhmien kesken, ja asettaa näiden sadepilvikuvaajat samalle akselille päällekkäin, tulee tällöin kiinnittää huomiota myös kuvaajien erottuvuuteen toisistaan. Tätä voidaan edesauttaa selitteillä, väreillä, läpinäkyvyydellä ja käyttämällä raakadatan havaintojen esittämiseen erilaisia merkkejä. Houkutusta sisällyttää kuvaajaan liikaa tietoa ja “koristeita” tulee kuitenkin välttää.

LÄHTEET

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., van Langen, J. & Kievit, R. A. (2021). Raincloud plots: a multi-platform tool for robust data visualization [version 2; peer review: 2 approved]. *Wellcome Open Research*, 4(63). <https://doi.org/10.12688/wellcomeopenres.15191.2>

Hintze, J. L. & Nelson, R. D. (1998). Violin Plots: A Box Plot–Density Trace Synergism. *The American Statistician*, 52(2), 181–184. <https://doi.org/10.1080/00031305.1998.10480559>

Nuzzo, R. L. (2016). The Box Plots Alternative for Visualizing Quantitative Data. *PM&R*, 8(3), 268–272. <https://doi.org/10.1016/j.pmrj.2016.02.001>

TEKOÄLY KÄYTTÖÖN PK-YRITYKSISSÄ – KYBERTURVALLISUUDESTA TINKIMÄTTÄ

VESA VERTAINEN, JYVÄSKYLÄN AMMATTIKORKEAKOULU

Tekoälymallit ovat tuoneet kyberturvallisuuteen erityispiirteitä, jotka voivat osaltaan lisätä organisaation hyökkäyspinta-alaa. Mitä pk-yrityksissä tai valmistavan teollisuuden yrityksissä olisi kyberturvallisuuden kannalta otettava huomioon tekoälyn käyttöä suunniteltaessa?

Kun organisaatiot suunnittelevat tekoälyn käyttöönottoa, vähintä mitä ne voivat tehdä, on laatia selvät linjaukset liittyen tietosuojaan ja salassapitoon. Arkaluontoista dataa käsiteltäessä on salassapito ja tietosuoja aina huomioitava. Esimerkiksi julkisissa chattiboteissa on syytä varmistaa, että kyselyissä syötetty tieto on sellaista, jota voidaan jakaa. Vaarana on, että tieto ajautuu kielimallin koulutusmateriaaliin ja voi tulla sitä kautta jaetuksi muille mallin käyttäjille. Myös organisaation itsensä kouluttaman kielimallin tapauksessa datan laadusta on huolehdittava; erityisesti jos data sisältää arkaluontoisia yrityksen tietoja, esimerkiksi liikesalaisuuksia, tai asiakkaiden tietoja (Department for Science, Innovation & Technology, 2024).

Myös EU:n tekoälyasetus, jota aletaan soveltaa vaiheittain 2025 alkaen, asettaa omat vaatimuksensa tekoälyn käytölle yrityksissä. Asetus luokittelee tekoälyjärjestelmät ja -mallit niiden riskien perusteella ja asettaa eri luokille erilaisia vaatimuksia. Tämän vuoksi yritysten on tärkeää tiedostaa kehitettyjen tai omassa käytössä olevien järjestelmiensä riskiluokka. Esimerkiksi valmistavassa teollisuudessa tekoälyjärjestelmät, joita käytetään tuotantoprosesseissa, ovat useimmiten vähäriskisiä, mutta henkilöstöhallintoon liittyvät tekoälyjärjestelmät voivat sen sijaan kuulua suuririskisten luokkaan. (Teknologiateollisuus, 2025.)

TEKOÄLYMALLEIHIN KOHDISTUVIA UHKIA

Kyberturvallisuuden kannalta erityisen herkkiä tekoälymalleja ovat kielimallit. Omia tekoälyjärjestelmiä rakennettaessa onkin kiinnitettävä huomiota kyberturvallisuuteen ja varmistettava riittävän vahva tietoturva-arkkitehtuuri, johon kuuluvat muun muassa vahva pääsynhallinta, turvalliset suunnittelukäytännöt ja verkkoympäristön asianmukainen konfigurointi. Riittämättömät suojaustoimenpiteet voivat johtaa muun muassa järjestelmien tai käyttäjätiedon luottamuksellisuuden vaarantumiseen tai mallin käyttäytymisen peukalointiin. Seurauksena voi yritykselle tulla taloudellisia tappioita, mainehaittoja tai tietosuojarikkomuksia. (Department for Science, Innovation & Technology, 2024.)

Valmiit mallit saattavat sisältää haavoittuvuuksia, joiden kautta pahantahtoinen toimija voi tehdä hyökkäyksiä vihamielisten syötteiden avulla.

Kun suunnitellaan kielimallin kouluttamista omiin käyttötarkoituksiin, käytävä malli on valittava huolellisesti. Valmiit mallit saattavat sisältää haavoittuvuuksia, joiden kautta pahantahtoinen toimija voi tehdä hyökkäyksiä vihamielisten syötteiden avulla (ns. adversiaaliset hyökkäykset, eng. *adversarial attacks*). Samoin mallin koulutuksessa käytettävät kirjastot on valikoitava ja tarkastettava huolellisesti. Kirjastot, jotka usein ovat avointa lähdekoodia, voivat periaatteessa mahdollistaa toimitusketjuhyökkäyksen. (Department for Science, Innovation & Technology, 2024; Vähä-Sipilä ym., 2021, s. 18.)

Vihamielisissä syötteissä on usein kyse ”myrkyttämisestä” (eng. *prompt injection*). Jos pahantahtoisella toimijalla on esimerkiksi pääsy yrityksen chattibottiin, voidaan siltä yrittää urkkia yrityksen tai käyttäjien dataa ja saada se suoritamaan toimintoja, jotka ovat alkuperäisen tarkoituksen ulkopuolella (Ama, 2025). Pohjimmiltaan *prompt injection* muistuttaa SQL-injektiota, jossa käyttäjän syöttämä tieto yhdistyy luotettuihin SQL-komentoihin. Samoin kehotteen myrkyttämisessä käyttäjän syöte ja kielimalliin rakennettu järjestelmäkehote eli mallin kehittäjän määrittelemät ohjeet ja rajoitteet sekoittuvat toisiinsa. (Kosinski & Forrest, 2024.)

Lähes samanlainen menetelmä on myös *jailbreaking*, mutta siinä tällaista sekoittumista ei tapahdu, vaan sopivilla kehoitteilla pyritään vain ohittamaan kielimallin turvatoimeksi rakennetut suodattimet. Suojatoimena voi olla esimerkiksi järjestelmäkehote, joka kieltää loukkaavan puhetavan. Tunnettu *jailbreaking*-tekniikka on ns. DAN-kehote (*Do Anything Now*), jossa kielimallia pyydetään ottamaan roolihahmo tai olemaan mukana pelissä, jossa kaikki on

sallittua. (Kosinski & Forrest, 2024.) *Jailbreaking*-strategia voi olla myös jonkun vähemmän tunnetun kielen (johon kielimalli ei ole saanut paljoa koulutusmateriaalia) tai esimerkiksi morse-koodin käyttäminen kehotteessa (Bengio ym., 2025, s. 198).

Hyökkääjät saattavat pyrkiä myös koulutusdatan myrkyttämiseen (eng. *data poisoning*), joka voi johtaa vääristyneisiin tuloksiin tai päätöksenteon vaarantumiseen. Hyökkääjä voi yrittää esimerkiksi korruptoida koulutusdatan luokittelun, joka on uhka erityisesti ohjatun oppimisen malleille tai pyrkiä lisäämään malliin tai sen koulutukseen käytettävään datasettiin tarkoituksellisen vääristymän. Tästä käytetään termiä *bias injection*. Lopputulemana voisivat olla vääristyneet tai ”epäreilut” tulokset päätöksentekoprosessissa. (Department for Science, Innovation & Technology, 2024.)

KEINOJA TURVALLISUUDEN KEHITTÄMISEEN

Ennen tekoälyn käyttöönottoa on sille syytä tehdä perusteellinen tietoturvatästä sekä uhkamallinnus, jonka avulla tiedetään järjestelmiin ja datan säilytykseen liittyvät potentiaaliset uhkat, haavoittuvuudet ja hyökkäysvektorit. Lisäksi on olennaista luoda selkeät käyttöohjeet siitä, miten käyttäjät saavat mallia käyttää, jotta minimoidaan tietoturvariskit ja vikatilanteet. Syötteiden puhdistus ja validointi ovat myös tärkeitä käytäntöjä, joilla voidaan vaikeuttaa sellaisten kommentojen perillemeno, jotka saisivat mallin suorittamaan epätarjoituksenmukaisia tehtäviä ja vaarantaisivat järjestelmän eheyden. (Department for Science, Innovation & Technology, 2024.)

Syötetyt kehotteet on suodatettava niin, etteivät esimerkiksi koodinpätkät tai haitalliset komennot mene läpi. Lisäksi salauksen käyttö sekä dataa lähetettäessä että säilöittäessä on olennainen tietoturvatoinenpide. Esimerkkejä tällaisesta salauksesta ovat homomorfinen salaus, jossa dataa voidaan käsitellä salattuna, purkamatta sitä missään vaiheessa, tai SMPC (*Secure Multi-Party Computation*), jonka avulla eri osapuolet voivat yhdessä analysoida dataa paljastamatta yksityisiä tietoja toisilleen tai chattibotille. (Ama, 2025.)



Bengio ym. (2025, s.199) esittävät vastakeinona *adversarial attack*-tyyppisille hyökkäyksille ns. *adversarial training*-koulutusta. Tällöin kehitetään ensin hyökkäyksiä, joilla malli saadaan toimimaan ei-toivotusti, ja sen jälkeen opetetaan järjestelmä tunnistamaan ja käsittelemään hyökkäykset niin, että niiden vaikutus saadaan kumottua. Iso-Britannian tieteen, innovaation ja teknologian ministeriön tuottaman selvityksen mukaan tämän tyyppistä koulutusta käytetään nykyään yleisesti suurimpien tekoälyjärjestelmien kehityksessä, mutta se ei kuitenkaan yksinään ole riittävä keino hyökkäysten estämiseksi. Tekoälyjärjestelmien mahdollisten syötteiden eksponentiaalisen suuri määrä tekee mahdottomaksi löytää kaikkia mahdollisia hyökkäyksiä, joita vastaan malli olisi koulutettava.

Mallien poisoppimista ei-toivotuista käyttäytymistavoista (eng. *machine unlearning*) on myös esitetty. Se on kuitenkin osoittautunut haastavaksi ja aiheutta tutkitaan edelleen. Käytännössä vaikuttaa siltä, että mallit ennemmin tukahduttavat haitallista tietoa kuin poistavat sitä lopullisesti, jolloin tieto saattaa olla edelleen esiin kaivettavissa. Tärkeää onkin jatkuvasti valvoa mallien toimintaa ja reagoida tarpeellisin toimenpitein. (Bengio ym., 2025, s. 199–200.)

Organisaatioiden on luotava selkeät säännöt siitä, mitä tekoälyjärjestelmiä saa käyttää ja millä tavoin.

Tekoälyn käyttöönottoa ei ole syytä pelätä, mutta varautuminen ja riskien huomioon ottaminen on ehdottoman tärkeää. Organisaatioiden on luotava selkeät säännöt siitä, mitä tekoälyjärjestelmiä saa käyttää ja millä tavoin. Omien tekoälymallien riskienhallinnan tukena voi käyttää esimerkiksi Liikenne- ja viestintävirasto Traficomin Kyberturvallisuuskeskuksen julkaisemaa opasta Tekoälyn soveltamisen kyberturvallisuus ja riskienhallinta, johon sisältyy myös riskien itsearviointityökalu (Vähä-Sipilä ym., 2021). Lisäksi EU:n tekoälyasetuksessa on vahvistettu riskiperusteiset säännöt tekoälyjärjestelmien kehittäjille ja käyttöönottajille (Euroopan komissio, 2025). Asetuksen velvoitteet kannattaakin selvittää omassa käytössä olevien tekoälymallien osalta.

LÄHTEET

Ama, E. B. (2025). *Chatbot Security Essentials: Safeguarding LLM-Powered Conversations*. Haettu 28.3.2025 osoitteesta <https://www.lakera.ai/blog/chatbot-security>

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., ... Zeng, Y. (2025). *International AI Safety Report*. <https://www.gov.uk/government/publications/international-ai-safety-report-2025>

Department for Science, Innovation & Technology. (2024). *Cyber security risks to artificial intelligence*. GOV.UK. Haettu 4.3.2025 osoitteesta <https://www.gov.uk/government/publications/research-on-the-cyber-security-of-ai/cyber-security-risks-to-artificial-intelligence>

Euroopan komissio. (2025). *EU:n tekoälysäädös*. Haettu 15.5.2025 osoitteesta <https://digital-strategy.ec.europa.eu/fi/policies/regulatory-framework-ai>

Kosinski, M., & Forrest, A. (26.3.2024). *What is a prompt injection attack?* lbm.com. Haettu 7.3.2025 osoitteesta <https://www.ibm.com/think/topics/prompt-injection>

Teknologiateollisuus. (24.2.2025). *Tekoälyasetus - tekoälyjärjestelmien ja yleiskäyttöisten tekoälymallien turvallisuussäädös*. Haettu 28.3.2025 osoitteesta <https://teknologiateollisuus.fi/jasenille/tietopankki/digitalisaatio-ja-tekoaly/tekoalyasetus/>

Vähä-Sipilä, A., Marchal, S. & Aksela, M. (2.11.2021). *Tekoälyn soveltamisen kyberturvallisuus ja riskienhallinta*. Liikenne- ja viestintävirasto Traficom, Kyberturvallisuuskeskus. <https://www.kyberturvallisuuskeskus.fi/sites/default/files/media/publication/Teko%C3%A4lyn%20soveltamisen%20kyberturvallisuus%20ja%20riskienhallinta.pdf>

KOHTI ÄLYKKÄÄMPÄÄ DATANKERUUTA TEKÖÄLYN TUELLA

SAKU KAARLEJÄRVI & JUHA-MATTI AROLA, SEINÄJOEN AMMATTIKORKEAKOULU

Digitalisaatio on kasvattanut organisaatioiden datan määrää. Datan tehokas hyödyntäminen organisaatioissa on keskeinen tapa kasvattaa kilpailukykyä. Tässä artikkelissa pohditaan, miten tekoälyä voi soveltaa datan hyödyntämisen tehostamiseen.

Digitalisaation myötä organisaatioiden keräämän datan määrä on kasvanut nopeasti. Tiedon tehokas kerääminen, käsittely ja analysointi ovat muodostuneet keskeisiksi kilpailutekijöiksi niin julkisella kuin yksityisellä sektorilla. Tekoäly tarjoaa uusia mahdollisuuksia näiden prosessien tehostamiseen mahdollistaen automaattisen datankeruun, virheiden tunnistamisen ja datan laadun parantamisen (Suomi.fi, 2024). Seinäjoen ammattikorkeakoulun julkaisemassa artikkelissa Tekoälyn hyödyntäminen datankeruussa kuvataan käytännön esimerkkejä siitä, miten tekoälyä voidaan hyödyntää erityisesti rakennetun ympäristön tietojen keräämisessä. Etenkin ChatGPT:n kaltaiset kielimallit ovat osoittautuneet käyttökelpoisiksi datan käsittelyssä ja suodattamisessa (Kaarlejärvi & Arola, 2024).

TEKOÄLY DATANKERUUN APUNA

Digitalisaatio ja teknologinen murros ovat nostaneet datan roolin keskeiseksi osaksi organisaatioiden toimintaa. Samaan aikaan datan määrä on kasvanut räjähdysmäisesti – sitä syntyy sensoreista, verkkopalveluista, asiakasrajapinnoista ja koneista. Tämän valtavan tietovirran hallinta ja hyödyntäminen on osoittautunut haasteelliseksi. Avuksi on tullut tekoäly, joka tuo uusia keinoja tiedon keräämiseen, järjestämiseen ja hyödyntämiseen (Suomi.fi, 2024).

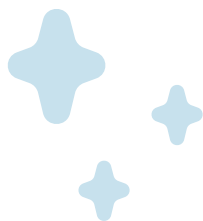
Datankeruun merkitys ei rajoitu vain teknisiin sovelluksiin, vaan se on keskeinen osa strategista johtamista ja innovaatiotoimintaa. Esimerkiksi reaaliaikainen asiakasdata mahdollistaa palveluiden jatkuvan optimoinnin, ja tuotantolaitoksista kerätty data auttaa ennaltaehkäisemään vikatilanteita.

Tekoälyn avulla tämä keruu voidaan toteuttaa skaalautuvasti ja tehokkaasti ilman että inhimilliset resurssit muodostuvat pullonkaulaksi. Samalla tekoäly muuttaa käsitystämme tiedosta: sen sijaan että ihmiset keräävät ja käsittelevät tietoa, koneet alkavat osallistua itse tiedon tuottamiseen ja arviointiin. Lisäksi tekoälyn kehittyessä se kykenee tunnistamaan datasta yhteyksiä ja trendejä, joita ihmisanalyttikot eivät huomaisi, mikä tarjoaa kilpailuetua tiedolla johtamiseen perustuvassa toimintaympäristössä (Chatterjee ym., 2021).

Seinäjoen ammattikorkeakoulun artikkelissa Tekoälyn hyödyntäminen datankeruussa (2024) kuvataan konkreettisesti, kuinka ChatGPT:tä hyödynnettiin datan käsittelyyn rakennustyömaan tietoaineistoissa. Kokemus osoitti, että kielimallien käyttö voi tehostaa työtä merkittävästi, mutta samalla se nosti esiin tietoturvan, rajallisen kontekstittajun ja teknisen osaamisen merkityksen. Tämä havainnollistaa, kuinka tekoäly voi olla hyödyllinen työkalu myös sellaisten tietomuotojen käsittelyssä, joita perinteiset järjestelmät eivät pysty tehokkaasti prosessoimaan. On kuitenkin tärkeää, että käyttäjillä on valmiudet arvioida tekoälyn tuottaman datan laatua ja luotettavuutta – pelkkä teknologia ei takaa tiedon oikeellisuutta (Kaarlejärvi & Arola, 2024; Honkanen, 2024).

DATAN ESIKÄSITTELY JA LAADUNVARMISTUS

Kerätty data ei yleensä ole automaattisesti käyttökelpoista. Tekoäly voi suodattaa virheellistä tai epäolennaista tietoa, tunnistaa puuttuvia arvoja ja ehdottaa datan täydentämistä (Halenius & Larsio, 2019). Esikäsittelyvaiheen merkitys korostuu erityisesti suurten tietomassojen yhteydessä, joissa manuaalinen laadunvarmistus ei ole realistista. Kaarlejärven & Arolan (2024) artikkelissa kuvattiin, kuinka tekoälyllä tuotettiin koodia, joka käsitteli ja muotoili PDF-datan Excel-muotoon. Tämä mahdollisti datan hyödyntämisen jatkokehittyneessä analytiikassa, kuten visualisoinnissa ja raportoinnissa. Tiedon laadun parantaminen tekoälyn avulla ei ole vain tekninen kysymys, vaan sillä on myös liiketoiminnallista merkitystä. Laadukas data luo perustan tehokkaalle päätöksenteolle, kehittämiselle ja automaatiolle (Suomi.fi, 2024).



Elämme maailmassa, jossa dataa on enemmän kuin koskaan.

Me elämme maailmassa, jossa dataa on enemmän kuin koskaan. Usein kuitenkin unohdetaan, että suuri datamäärä ei ole itseisarvo – laadukas ja relevantti data on. Ilman huolellista esikäsittelyä tekoäly voi tehdä virheitä, aiheuttaa väärinymmärryksiä tai jopa vahvistaa olemassa olevia vinoumia (Chatterjee ym., 2021). Esimerkiksi sosiaali- ja terveyssektorilla tekoälyä käytetään yhä enemmän potilaskertomusten ja anturitiedon yhdistämiseen. Näissä järjestelmissä datan eheys ja aikaleimojen oikeellisuus ovat kriittisiä, sillä pienikin virhe voi vaikuttaa hoitopäätöksiin. AIDRIN-kehys (Hiniduma ym., 2024) tarjoaa hyvän mallin arvioida datan valmiutta tekoälykäyttöön: – se tarkastelee datan saatavuutta, yhtenäisyyttä, keräystapaa ja eettisiä reunaehjoja.

Uudet sensoriteknologiat ja älylaitteet avaavat täysin uudenlaisia mahdollisuuksia datan keruuseen. Rakennusosalalla käytetään jo droneteknologiaa ja konenäköä työmaiden valvonnassa, logistiikassa tekoäly valvoo kylmäketjuja lämpötila-antureiden avulla, ja vähittäiskaupassa asiakkaiden liikkumista seurataan videoanalytiikalla (Mikalef ym., 2021). Näissä sovelluksissa tekoäly ei ainoastaan kerää dataa – se ohjaa datankeruuta. Se päättää, milloin poikkeama on merkityksellinen ja milloin kannattaa tallentaa lisätietoa. Tämä itsenäinen havaintokyky tekee tekoälystä erityisen kiinnostavan myös tutkimus- ja kehitysympäristöissä, joissa datan ajantasaisuus on ratkaisevaa.

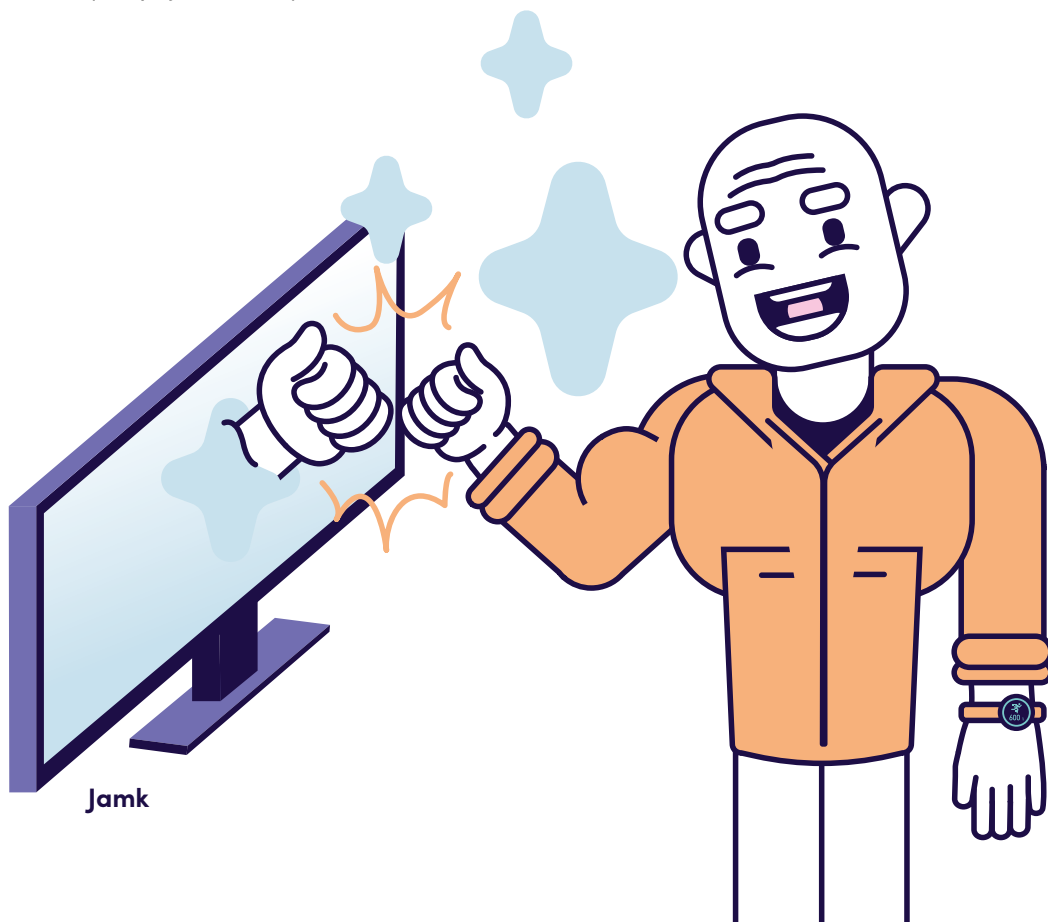
TIETOTURVA, YKSITYISYYS JA LÄPINÄKYVYYS

Vaikka tekoäly tarjoaa merkittäviä hyötyjä, sen käyttöön liittyy myös eettisiä ja lainsäädännöllisiä haasteita. Kaarlejärven & Arolan (2024) artikkelissa korostettiin tietoturvan merkitystä: tekoälyä hyödyntäessä ei tule syöttää järjestelmään henkilötietoja tai luottamuksellisia tietoja. Vastaavasti Haleniuksen & Larsion (2019) mukaan eettisesti kestävä tekoäly vaatii eettistä dataa: dataa, joka on kerätty läpinäkyvästi, suostumukseen perustuen ja vastuullisesti hallinnoituna. Myös yksilön näkökulmasta on tärkeää, että datankeruun tavoitteet, rajat ja käyttötavat ovat ymmärrettävissä. EU:n datastrategia (Euroopan komissio, ei pvm.) korostaa yksilön oikeutta hallita omaa dataansa ja vaatii läpinäkyviä sopimuksia osapuolten välillä. Lisäksi on syytä todeta, että jos tekoäly kouluttaa itseään tai ei ole paikallinen, on hyvä käyttää vain yleistä tai julkista dataa, sillä tietojen leviämässä on riskinsä. Paikallisissa tekoälymallissa, joissa tietoja käsitellään paikallisesti omalla palvelimella tai omalla tietokoneella, voidaan myös käyttää yksityisempääkin dataa, sillä se ei leviä paikallisesta koneesta muualle.

Tiedon omistajuus, käyttöoikeudet ja yksityisyys ovat nousseet tekoälyn myötä aivan uuteen valoon. Erityisesti tekoälypalvelut, kuten ChatGPT, tallentavat ainakin osan käyttäjän syöttämästä datasta kehitystarkoituksiin – ellei käytössä ole ns. yksityinen instanssi. Kaarlejärven & Arolan (2024) artikkelissa nostettiin esiin huoli datan hallinnan siirtymisestä käyttäjältä pilvipalveluntarjoajalle. UNESCO:n (2021) tekoälyn eettisiä periaatteita käsittelevä suositus painottaa ihmiskeskeisyyttä, oikeudenmukaisuutta ja läpinäkyvyyttä. On tärkeää, että organisaatiot ymmärtävät teknologian eettisen vaikutuspiirin eivätkä keskity pelkästään tehokkuuden maksimointiin.

IHMINEN JA TEKÖÄLY – YHTEISTYÖ, EI KILPAILU

Tekoäly ei korvaa ihmistä datankeruussa, vaan toimii parhaimmillaan asiantuntijan työparina. Ihminen ohjaa tekoälyn käyttöä, tulkitsee tulokset ja tekee päätökset. Esimerkiksi oppilaitoksissa tekoälyä on hyödynnetty automatisoimaan tehtävien tarkastamista, jolloin opettajat voivat keskittyä vuorovaikutukseen ja yksilölliseen ohjaukseen (Honkanen, 2024). Samalla tavalla yrityksissä tekoäly voi auttaa nostamaan esiin asiakaspalautteesta trendejä ja ilmiöitä, joita manuaalinen tarkastelu ei tavoita. Lopullinen vastuu pysyy kuitenkin ihmisellä – tekoäly toimii välineenä, ei itsenäisenä päätöksentekijänä (Luopajarvi, 2023).



Datalukutaito nousee nyt ydintaidoksi: ymmärrys siitä, mitä data on, mistä se tulee, mitä se kertoo – ja mitä ei.

Tekoälyn käyttöönotto vaatii osaamista – ei vain teknistä, vaan myös kriittistä ja eettistä. Datalukutaito nousee nyt ydintaidoksi: ymmärrys siitä, mitä data on, mistä se tulee, mitä se kertoo – ja mitä ei. Vasta kun tämä on hallussa, tekoälyä voidaan käyttää vastuullisesti ja tehokkaasti. Organisaatioilla on vastuu kouluttaa henkilöstöään ymmärtämään datan keruun perusteet ja sen eettiset reunaehdot. Tämä tarkoittaa muun muassa sitä, että työntekijät osaa- vat kysyä oikeita kysymyksiä tekoälyn ehdotuksille: ”Mistä tämä tieto on peräisin?”, ”Mihin se perustuu?”, ”Miten se voi olla väärässä?”

KOHTI JATKUVASTI OPPIVAA JÄRJESTELMÄÄ

Tulevaisuudessa tekoäly ei ainoastaan kerää ja suodata dataa, vaan voi myös ehdottaa, mitä dataa kannattaa kerätä. Tällaiset dynaamiset järjestelmät oppivat jatkuvasti kontekstista ja voivat auttaa organisaatioita kohdentamaan resurssinsa tehokkaammin (Suomi.fi, 2024). Kehityksen rinnalla on tärkeää säilyttää kriittinen tarkastelu: keitä tekoälyratkaisut lopulta hyödyttävät ja miten ne palvelevat yhteistä hyvää? Tarvitaan selkeitä pelisääntöjä, teknologista ymmärrystä ja ennen kaikkea rohkeutta kokeilla ja oppia.

Tulevaisuuden tekoäly ei pelkästään tee sitä, mitä käskee, vaan se tekee aloitteita. Jo nyt kehittyneet järjestelmät voivat ehdottaa uusia datan keruun lähteitä, tuottaa kyselylomakkeita tai automaattisesti yhdistää ja harmonisoida eri tietolähteistä saatua dataa.

Beam.ai:n (2024) kaltainen työkalu edustaa tätä uutta suuntaa. Se kokoaa, siivoaa ja järjestää verkkolähteistä, tiedostoista ja tietokannoista poimittua dataa – ilman manuaalista kosketusta. Kun järjestelmät oppivat, ne optimoivat myös sen, mitä dataa ylipäättään kannattaa kerätä. Tällainen jatkuva parantaminen tuo mukanaan tehokkuutta, mutta myös eettisiä kysymyksiä: onko käyttäjällä enää täyttää näkyvyyttä siihen, mitä kerätään ja miksi?

LÄHTEET

- Beam.ai. (2024). *Data Collection AI Agent*. <https://beam.ai/agents/data-collection-agent>
- Chatterjee, S., Rana, N. P., Tamilmani, K. & Sharma, A. (2021). The next generation of data-driven decision-making: A review of literature and future research agenda. *International Journal of Information Management*, 58, 102331. <https://doi.org/10.1016/j.ijinfomgt.2020.102331>
- Euroopan komissio. (Ei pvm.). *Euroopan datastrategia*. Haettu 21.5.2025 osoitteesta https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_fi
- Halenius, L. & Larsio, A. (14.1.2019). *Eettinen tekoäly tarvitsee eettistä dataa*. Haettu 21.5.2025 osoitteesta <https://www.sitra.fi/blogit/eettinen-tekoaly-tarvitsee-eettista-dataa/>
- Hiniduma, K., Byna, S., Bez, J. L. & Madduri, R. (2024). AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI. Teoksessa *Proceedings of the 36th International Conference on Scientific and Statistical Database Management* (s. 1-12). <https://doi.org/10.1145/3676288.3676296>
- Honkanen, J. (2024). *Tekoälyn käyttö tehtävien tarkastamiseen* [Opinnäytetyö, Hämeen ammattikorkeakoulu]. Theseus. <https://urn.fi/URN:NBN:fi:amk-2024112931283>
- Kaarlejärvi, S., Arola, J. (2024). *Tekoälyn hyödyntäminen datankerauksessa*. @SEAMK-verkkolehti. <https://urn.fi/URN:NBN:fi-fe20241205100009>
- Luopajarvi, M. (2023). *Tekoälyn hyödyntäminen tietoverkoissa* [Pro gradu -tutkielma, Vaasan yliopisto].
- Mikalef, P., Krogstie, J., Pappas, I. O. & Giannakos, M. (2021). Investigating the effects of big data analytics capabilities on firm performance: The mediating role of dynamic capabilities. *Information & Management*, 58(2), 103434. <https://doi.org/10.1016/j.im.2020.103434>
- Suomi.fi. (2024). *Toimiva tekoäly vaatii laadukasta dataa*. Haettu 21.5.2025 osoitteesta <https://kehittajille.suomi.fi/oppaat/vastuullinen-tekoaly/maaritle-datapolitiikka/toimiva-tekoaly-vaatii-laadukasta-dataa>
- UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>

GENERATIIVISEN TEKOÄLYN MAHDOLLISUUDET DATAN HYÖDYNTÄMISESSÄ

JANNE ALATALO, JYVÄSKYLÄN AMMATTIKORKEAKOULU

Generatiiviset tekoälyteknikat ovat viimevuosina mahdollistaneet monia asioita, jotka ovat olleet aiemmin vaikeita toteuttaa. Tämä artikkeli ennakoii muutosta, jonka generatiiviset tekoälyteknikat aiheuttavat eri aloille, ja kertoo mahdollisuuksista, joita nämä teknikat tuovat datan hyödyntämisessä.

Tekoäly otti suuren teknologisen harppauksen vuoden 2022 lopulla, kun yhdysvaltalainen OpenAI yritys julkaisi ChatGPT-nimisen tekoälypalvelun (OpenAI, 2022). Palvelu oli ensimmäinen suurelle yleisölle suunnattu tuote, jossa käyttäjä pääsi viestittelemaan älykkään tekoälyavustajan kanssa luonnollisella kielellä ja jossa tekoäly selvästi ymmärsi käyttäjää sekä tarjosi hyödyllisiä vastauksia takaisin käyttäjälle.

ChatGPT-palvelu oli ensimmäinen tekoälyavustaja markkinoilla, ja sen julkaisu aiheutti paljon kohinaa muissa teknologiajäteissä. Muut teknologiajätit seurasivat kuitenkin nopeasti perässä julkaisemalla omat palvelunsa. Suurimmat ja tunnetuimmat ainakin länsimaissa ovat tällä hetkellä Microsoftin Copilot, Googlen Gemini ja Anthropic Claude. Kaikki nämä palvelut perustuvat samaan teknologiaan, josta yleisesti käytettyjä termejä on generatiivinen tekoäly (Generative AI) tai suuri kielimalli (Large Language Model).

ChatGPT-palvelun julkaisu oli ennennäkemätön harppaus tekoälyn sovellusmahdollisuuksissa, eikä mitään vastaavaa ole tekoälyn saralla aikaisemmin nähty. Vaikka tänä päivänä monelle tekoäly voi tarkoittaa älykästä tekoälyavustajaa, niin näin ei aina kuitenkaan ole ollut. Ennen ChatGPT-aikaa tekoäly tarkoitti älykästä ratkaisua, joka oli luotu ratkaisemaan jokin todella tarkasti rajattu ongelma. Tietty tekoälyratkaisu toimi vain siihen tiettyyn ongelmaan mihin se oli toteutettu. Niissä ei ollut käyttöliittymää, jonka kautta ihminen olisi päässyt ohjeistamaan niitä. Esimerkiksi kuvia luokitteleva tekoäly otti sisään pelkkiä kuvasyötteitä, eikä niissä ollut mitään mekanismia, millä niitä pystyisi hyödyntää mihinkään muuhun kuin kuvien luokitteluun. Tämän vuoksi voidaan sanoa, että niissä ei ollut sisällä yleisesti sovellettavaa älyä.

GENERATIIVISEN TEKOÄLYN MAHDOLLISTAJAT

OpenAI oli ensimmäinen yritys, joka sai viitteitä siitä, että suuri datamäärä ja so-pivat syväoppimisteknologiat pystyvät kouluttamaan tekoälymallin, joka sisältää yleisesti sovellettavaa älyä (Radford ym., 2018). Lopulliseen generatiivisen tekoälyn läpimurtoon johti pitkä kehityskaari, jossa monet teknologiajätit ja yli-opistot ympäri maailman edistivät teknologiaa luonnollisen kielen käsittelyyn. Lopullisena mahdollistavana teknologiana voidaan pitää laskentatehon kasvua sekä syväoppimisteknologioiden kehitystä, mitkä mahdollistivat erittäin suurien kielimallien kouluttamisen massiivisella datamäärällä (Chowdhery ym., 2023).

Massiivisen datamäärän käyttö tekoälyn opetusvaiheessa johtaa siihen, että malli oppii opetusdatasta paljon yleisesti sovellettavaa tietoa, joka mahdollistaa mallin käytön monenlaisiin käyttökohteisiin. Ennen generatiivisen tekoälyn aikaa perinteisten tekoälymallien kouluttaminen alusta lähtien vaati paljon opetusdataa. Opetusdatan luominen oli usein kallista ja aikaa vievää, koska se valmisteltiin ihmisen avulla. Ihminen prosessoi ja merkitsi tekoälylle menevän koulutusdatan käsin. Nykyään kuitenkin generatiiviset tekoälymallit oppivat paljon yleisesti sovellettavaa älyä jo alkuperäisestä opetusdatasetistä. Sen avulla nämä mallit pystyvät suoriutumaan uusista tehtävistä suoraan tai vain muutama ihmisen antaman esimerkin perusteella (Brown ym., 2020; Wei ym., 2022).



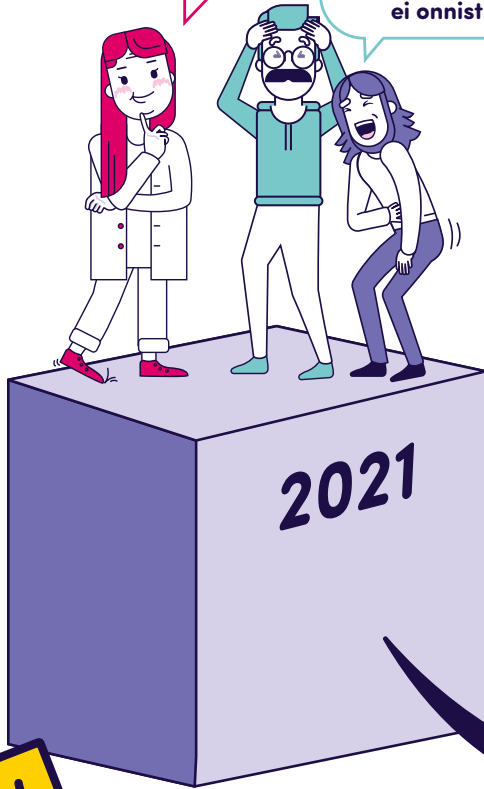
Kehitys on johtanut siihen, että joitain aiemmin mahdollisena pidettyjä asioita onkin nykyään täysin mahdollista toteuttaa.

Tämän vuoksi generatiivinen tekoäly voi mahdollistaa täysin uusia datan hyödyntämisen käyttökohteita, jotka eivät olleet vielä mahdollisia muutama vuosi sitten. Esimerkkinä tästä on Jamkin ja Valtran yhteistyöprojekti, jossa Valtran laadunvalvontadatan vapaatekstikuvauksia luokiteltiin paikallisen kielimallien avulla pelkästään kuvailemalla kielimallille halutut luokat (Alatalo, 2024). Luokittelija, joka pystyy laadukkaaseen tekstin luokitteluun pelkällä luokkien kuvauksella ja tehtävänkuvauksella, olisi ollut täysin science fictionia vielä vuonna 2020, mutta nykyään tämä on jo täysin arkipäivää.

Tämä kuvaa hyvin generatiivisen tekoälyn äärimmäisen nopeaa kehitystah-tia. Kehitys on johtanut siihen, että joitain aiemmin mahdollisena pidettyjä asioita onkin nykyään täysin mahdollista toteuttaa.

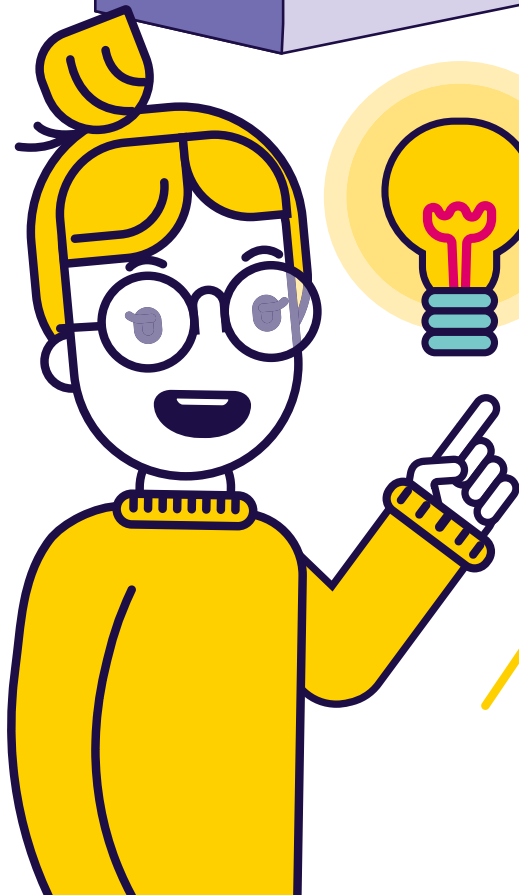
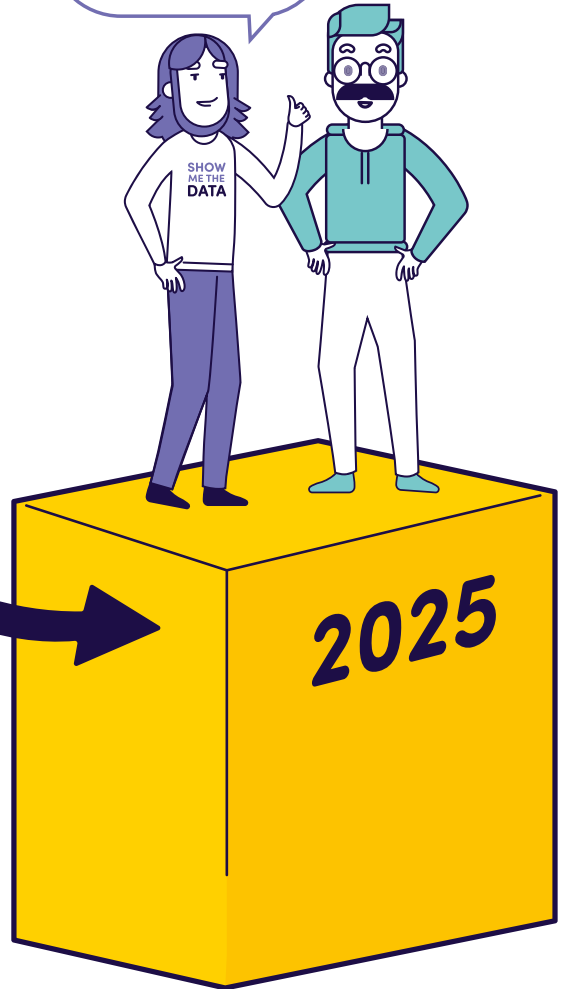
Suunnitellaan sovellus, joka muuttaa käsinkirjoitetut reseptit digitaalisiksi ohjeiksi!

Melkoinen työmaa, ei onnistu!



Helposti! Anna puoli tuntia, niin on eka versio.

Kielimalleja käyttämällä onnistuu!

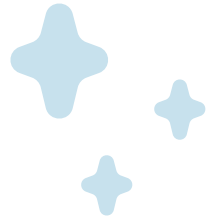


Kielimallit ovat tehneet lähes mahdottomasta helppoa vain muutamassa vuodessa.

TULEVAISUUDEN NÄKYMÄT

Alun perin suuret kielimallit ottivat syötteenä pelkästään tekstiä ja loivat ulostulona pelkkää tekstiä, mutta kehitys on myös tällä saralla nopeaa. Niin kutsutut multimodaaliset (*multimodal*) kielimallit ymmärtävät syötteenä kuvia, videota ja ääntä pelkän tekstin lisäksi, ja jotkut voivat jopa luoda kuvia itse, vaikkakin usein kuvien luomiseen käytetään vielä erillistä siihen erikoistunutta mallia. Tämä tarkoittaa, että generatiivista tekoälyä voi soveltaa enenevässä määrin erilaisiin datatyyppeihin. Tämä kasvattaa entisestään niiden mahdollisuuksia monenlaisten ongelmien ratkaisemiseen.

Multimodaaliset kielimallit ovat tekemässä läpimurtoa ja lisäävät näiden teknologioiden käyttömahdollisuuksia entisestään erilaisten ongelmien ratkaisuun.



ChatGPT palvelu julkaistiin vuoden 2022 lopussa, ja OpenAI on julkaissut uusia älykkäämpiä malleja siitä lähtien vuosittain. Muut teknologiajätit ovat samaan aikaan kehittäneet omia palvelujaan, ja kilpailu käyttäjistä on erittäin kovaa. Multimodaaliset kielimallit ovat tekemässä läpimurtoa ja lisäävät näiden teknologioiden käyttömahdollisuuksia entisestään erilaisten ongelmien ratkaisuun. On vaikea uskoa, että tämä kehityskulku pysähtyy lähiaikoina.

Datan hyödyntämiseen liittyen generatiivisen tekoälyn nopea kehitys aiheuttaa sen, että datan käyttömahdollisuuksia tulisi arvioida uudelleen nykyisten tekoälymallien mahdollisuuksiin peilaten. Ennen generatiivisen tekoäly aikaa tekoälyä oli hankala soveltaa moneen käyttökohteeseen laadukkaan datan vähyden vuoksi. Generatiiviset tekoälymallit sisältävät kuitenkin yleisesti sovellettavaa älyä, joka voi joissain tilanteessa johtaa siihen, että vaaditun datan määrä ja laatuvaatimukset tippuvat. Tämä voi tarkoittaa sitä, että aikaisemmin arvoton data on nykyään täysin käyttökelpoista monenlaisten ongelmien ratkaisemiseen ja voi tuottaa yrityksille uutta arvoa.

LÄHTEET

Alatalo, J. (15.11.2024). Avoimet kielimallit suomenkielisen tekstin luokitteluun. *Jamk Arena Pro*. <https://urn.fi/urn:nbn:fi:jamk-issn-2984-0783-137>

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language Models are Few-Shot Learners. Teoksessa H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (toim.). *Advances in Neural Information Processing Systems*, 33, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bf8ac142f64a-Paper.pdf

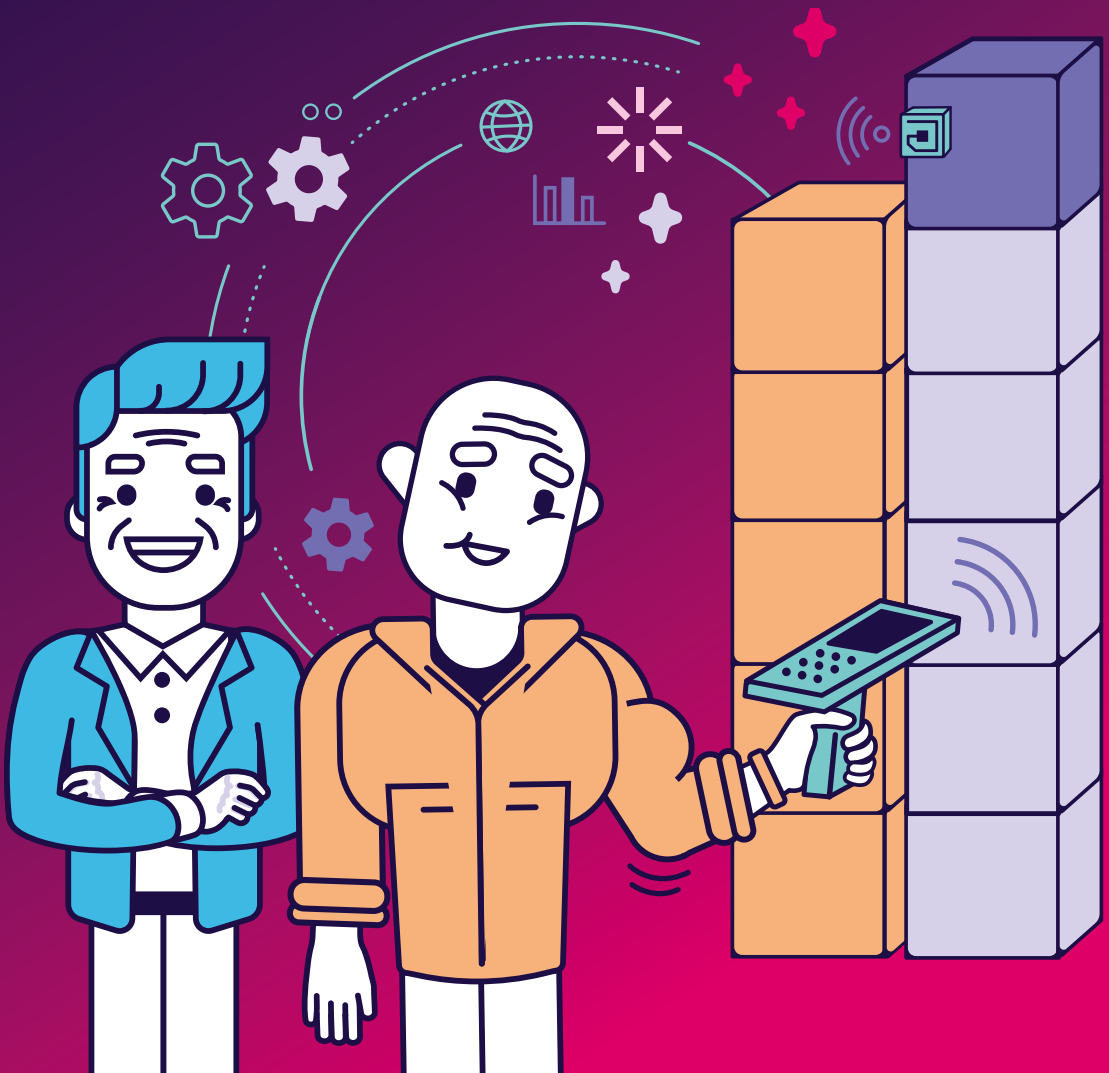
Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... Fiedel, N. (2023). PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240), 1–113.

OpenAI. (30.11.2022). *Introducing ChatGPT*. Haettu 30.4.2025 osoitteesta <https://openai.com/index/chatgpt/>

Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M. & Le, Q. V. (2022). *Finetuned Language Models are Zero-Shot Learners*. International Conference on Learning Representations. <https://doi.org/10.48550/arXiv.2109.01652>

OPITAAN ESIMERKISTÄ - KÄYTTÖTAPAUKSIA TEOLLISUUDESTA



RFID-TEKNOLOGIA METALLIALAN YRITYKSISSÄ

MIKA VALKAMA, SEINÄJOEN AMMATTIKORKEAKOULU

RFID-teknologia lupaa mullistaa kohteiden seurannan, mutta miten se toimii haastavissa olosuhteissa? Tämä artikkeli esittelee käytännön kokeiluja, joissa selvitettiin mahdollisia ongelmia teollisuusympäristössä ennen suoraa käyttöönottoa.

Metallialan yrityksissä on monia prosesseja ja materiaalivirtoja, joiden tarkempi seuranta voisi parantaa toiminnan tehokkuutta ja nopeutta. Digitalisaation myötä datan määrä kasvaa jatkuvasti, ja erityisesti laadukas, reaaliaikainen data mahdollistaa parempia päätöksiä, prosessien optimointia ja resurssien säästöä. Esimerkiksi datavetoiset yritykset voivat tehostaa varastonhallintaa ja vähentää hukkaa, mikä parantaa sekä kilpailukykyä että kestävyttä (European Commission, 2020; McKinsey & Company, 2022).

Tämä artikkeli perustuu VauhtiData-hankkeessa toteutettuihin kokeiluihin, joissa selvitettiin RFID-teknologian soveltuvuutta varastoitavien kohteiden – kuten kuljetushäkkien, pulverimaalilaatikoiden ja leikattujen metallikappaleiden – seurantaan metallialan yrityksissä. Tavoitteena oli tunnistaa teknologian vahvuudet ja haasteet sekä löytää parhaat käytännöt sen hyödyntämiseksi.

Kokeilut toteutettiin yhteistyössä kahden paikallisen metallialan yrityksen kanssa. Testattavat kohteet valikoituivat yritysten tarpeiden perusteella: haluttiin seurata kuljetushäkkien lähtöä maailmalle ja paluuta yritykseen; pulverimaalilaatikot olivat jo varastojärjestelmässä, mutta niistä kaivattiin tarkempaa tietoa sijainnista varaston ulkopuolella ja mahdollista liitääntä maalausprosessiin; viimeisenä kohteena keskityttiin leikattujen metallikappaleiden kulkuun leikkauksesta lavojen ja varaston kautta kuljetukseen. Testeihin saimme lainaksi Siemensiltä antennia, lukijoita ja tunnisteita.

JOHDANTO RFID-TEKNOLOGIAAN

Radio Frequency Identification (RFID) on teknologia, joka hyödyntää radiotaajuuksia esineiden tunnistamiseen ja seurantaan ilman suoraa näköyhteyttä, toisin kuin perinteiset viivakoodit. RFID-järjestelmä koostuu tunnisteista (*tags*), lukijoista (*readers*) ja antenneista. Tunnisteet voivat olla passiivisia, jolloin ne saavat energiansa lukijan radioaalloista, tai aktiivisia, joissa on oma virtalähde pidempää lukuetaisyyttä ja suurempaa tiedon tallennusta varten (Finkenzeller, 2010).

Teknologia toimii kolmella päätaajuusalueella:

- Matalataajuus (LF, 125–134 kHz) tarjoaa tyypillisesti 10–50 cm:n lukuetaisyyden ja sopii lyhyen kantaman sovelluksiin.
- Korkea taajuus (HF, 13.56 MHz) yltää noin 1 metriin.
- Ultrakorkea taajuus (UHF, 860–960 MHz) voi saavuttaa jopa 10–15 metrin etäisyyden, mikä tekee siitä suosittuun teollisuudessa ja logistiikassa sen nopean tiedonsiirron ja laajan kantaman vuoksi.

Kuluttajille tuttu *Near Field Communication* (NFC) on HF-taajuuteen perustuva RFID:n alamuoto, jota käytetään esimerkiksi lähimaksamisessa ja tiedonsiirrossa älypuhelimilla. NFC:n lukuetaisyys on rajoitettu muutamaan senttimetriin, mikä parantaa turvallisuutta, mutta rajoittaa sen käyttöä laajamittaisessa seurannassa.

Tunnisteisiin voidaan tallentaa monenlaista tietoa, kuten yksilöllisiä tunnistenumeroita, tuotekoodeja tai lisätietoja kohteesta. Passiivisten tunnisteen muisti vaihtelee tyypillisesti 64 bitistä muutamaan kilotavuun (esim. UHF-tunnisteissa jopa 8 kB), kun taas aktiiviset tunnisteet voivat sisältää kymmeniä kilotavuja dataa riippuen mallista (Weinstein, 2005). Jokaisella tunnisteella on aina yksilöllinen tunniste-ID (esim. EPC-koodi UHF-tunnisteissa), joka säilyy kiinteänä riippumatta muusta tallennetusta datasta mahdollistaen kohteen erottamisen muista.

METALLIOSIEN KANSSA ILMENEVÄT ONGELMAT

Teollisuusympäristöissä RFID-tekniikan käyttöön liittyy haasteita, sillä erilaiset materiaalit ja olosuhteet voivat häiritä radiotaajuussignaaleja (Zhang ym., 2017). Keskeiset ongelmat ovat:

1. Signaalin heijastuminen: Metallipinnat ja muut heijastavat materiaalit, kuten kiiltävät pinnoitteet, heijastavat RFID-signaaleja, mikä voi vääristää signaalia ja estää tunnisteen luotettavan lukemisen (Finkenzeller, 2010). Tämä on erityisen ongelmallista UHF-taajuuksilla, joita suositaan niiden pitkän luoketäisyyden vuoksi.

2. Signaalin vaimennus: Tiheät materiaalit, kuten metalli, vaimentavat radiotaajuuksia, mikä heikentää lukematarkkuutta ja lyhentää luoketäisyyttä (Weinstein, 2005). Esimerkiksi tiiviisti pakatut osat voivat estää signaalin pääsyn tunnisteseen, mutta myös muut tiiviit materiaalit, kuten nesteet tai paksumat komposiitit, voivat aiheuttaa samankaltaisia ongelmia.

3. Interferenssi: Useiden esineiden läheisyys, erityisesti metallisten, aiheuttaa signaalien sekoittumista, mikä vaikeuttaa yksittäisten tunnisteen erottamista (Zhang ym., 2017). Metallialalla tämä korostuu, mutta myös muut teolliset ympäristöt, joissa on paljon koneita tai elektronisia laitteita, voivat kärsiä vastaavista häiriöistä.

Näiden haasteiden vuoksi RFID-tekniikan soveltaminen teollisuusympäristöissä edellyttää huolellista suunnittelua ja optimointia. Ratkaisuja voivat olla esimerkiksi häiriöitä kestävien tunnisteen käyttö, antennien strateginen sijoittelu ja lukijoiden tehon säätö (Finkenzeller, 2010). Haasteet voivat nostaa implementointikustannuksia ja hidastaa tekniikan käyttöönottoa, mikä korostaa tarvetta ympäristökohtaisille ratkaisuille ja testaukselle.

KOKEILUT METALLIALAN YRITYKSISSÄ

Hankkeemme puitteissa toteutimme kokeiluja RFID-tekniikan soveltuvuudesta metallialan yrityksissä kolmessa eri käyttökohteessa: leikkattujen metalliosien, kuljetushäkkien ja pulverimaalilaatikoiden seurannassa. Ennen yrityksissä tehtyjä kokeiluja tutustuimme Siemensiltä lainattuihin antenneihin, lukijoihin ja tunnisteesiin SEAMKin laboratoriossa, jossa suunnitelimme testit yritysten tarpeiden pohjalta. Kokeilut toteutettiin lopulta kahden paikallisen metallialan yrityksen kanssa hyödyntäen saamiamme laitteita.

KULJETUSHÄKKIEN SEURANTA

Kuljetushäkeistä yritys halusi tietää, milloin häkit lähtevät laitoksesta asiakkaille ja milloin ne palaavat takaisin. Täydet häkit kuormataan rekoille, jotka vievät ne asiakkaille, ja ne palautuvat useamman häkin nipuissa. Niput siirretään trukilla varastoalueelle, missä niistä voidaan lukea viivakoodit. Tämä käsin tehtävä luenta on hidasta ja vaivalloista.

Kokeilussa selvitettiin, voisiko trukkiin asennettava lukija tai lukija-antenni lukea RFID-tunnisteet siirron aikana. Testit osoittivat, että tunnisteiden lukeminen onnistui kohtalaisesti, mutta se vaati tarkkaa antennin asennon ja etäisyyden optimointia. Metallirakenteiset häkit aiheuttivat häiriöitä, ja luotettavuuden parantamiseksi tunnisteita tulisi todennäköisesti asentaa häkkiin useampia. Lisäksi tunnisteiden tulee kestää ulko-olosuhteita, kuten säätä ja mekaanista rasitusta. Tulokset viittaavat siihen, että RFID voisi korvata viivakoodit, mutta järjestelmän käyttöönotto edellyttää lisätestausta ja investointeja laitteistoon sekä sopivan tunnisteiden etsimisen.

PULVERIMAALILAATIKOIDEN SIJAINNIN SELVITTÄMINEN VARASTON ULKOPUOLELLA

Pulverimaalilaatikoiden seurannassa tavoitteena oli ratkaista ongelma, jossa laatikot poistetaan varastosta, mutta niiden tarkka sijainti sen jälkeen jää epäselväksi. Laatikoita saatetaan säilyttää maalaus koneen lähellä pitkiäkin aikoja, vaikka ne on merkitty varastosta poistetuiksi, mikä hankaloittaa varastohallintaa.

Kokeilussa testattiin RFID-tunnisteiden lukemista käsiskannerilla. Tulokset olivat lupaavia: tunnisteet toimivat luotettavasti ja käsiskanneri mahdollisti laatikoiden sijainnin nopean toteamisen. Lisäksi tunnisteita voitaisiin hyödyntää maalausprosessissa varmistamaan, että koneeseen ladataan oikea maali oikeaan työhön, mikä vähentäisi virheitä. Käsiskanneri tarvitsee rinnalleen kehitystyötä, jotta käyttökokemus saadaan sujuvaksi, integraatiota taustajärjestelmiin sekä käyttöliittymän joko lukijaan itseensä tai lähellä oleviin näyttöihin. Tämä osoitti, että RFID-teknologia sopii hyvin ympäristöihin, joissa metallin häiriövaikutus on vähäisempi, ja se voisi parantaa sekä varastohallintaa että prosessien tehokkuutta.

LEIKATTUJEN METALLIOSIEN SEURANTA

Leikattujen metalliosien seurannassa tavoitteena oli parantaa osien löydettävyyttä eri työvaiheiden jälkeen. Osia kuljetetaan lavoilla ilman järjestelmällistä seurantaa, ja niitä etsitään käsin varastohyllyiltä. Osien tai osanippujen tunnistetarrat ovat jo käytössä, mutta manuaalinen haku on hidasta ja tehotonta.

Kokeilussa testattiin, voisiko RFID-tunniste tarrassa mahdollistaa osien reaaliaikaisen sijainnin selvittämisen. Tulokset osoittivat, että tarratyypiset RFID-tunnisteet eivät toimineet luotettavasti suoraan metallikappaleen pinnalla. Kun tunnisteiden ja metallipinnan väliin lisättiin etäisyyttä, lukeminen onnistui, mutta käytännössä tämä ei ollut toimiva ratkaisu: osat pinotaan lavoille päällekkäin, jolloin alemmat tunnisteet jäävät peittoon, mikä estää niiden lukemisen. Tämä korostaa metalliympäristöjen haasteita ja tarvetta kehittää erikoistunnisteita tai vaihtoehtoisia menetelmiä pinottujen osien seurantaan.

VAIHTOEHTOISET TEKNOLOGIAT

RFID-teknologia tarjoaa merkittäviä etuja reaaliaikaisessa seurannassa ja automatisoinnissa, mutta sen haasteet tekevät vaihtoehtoisten teknologioiden tarkastelusta tarpeellista.

ULTRA-WIDEBAND (UWB)

UWB-teknologia hyödyntää laajaa taajuuskaistaa ja tarjoaa erittäin tarkan paikannuksen, usein senttimetrin tarkkuudella. Toisin kuin RFID, se on vähemmän altis metallipintojen aiheuttamille häiriöille, mikä tekee siitä lupaavan vaihtoehdon teollisuusympäristöihin (Sahinoglu ym., 2008). Teknologian häittäpuolena on kuitenkin sen korkeampi kustannus ja tarve erilliseen infrastruktuuriin, kuten ankkuripisteisiin, mikä voi rajoittaa sen käyttöönottoa pienemmissä yrityksissä.

1D/2D-VIIVAKOODIT JA QR-KOODIT

Viivakoodit ovat yksinkertainen ja kustannustehokas ratkaisu esineiden tunnistamiseen. 1D-viivakoodit tallentavat rajallisesti tietoa (esim. tunnistenumeron), kun taas 2D-viivakoodit, kuten QR-koodit, voivat sisältää laajempia tietoja (Gao ym., 2007).

Toisin kuin RFID, viivakoodit vaativat näköyhteyden lukijaan, mikä hidastaa toimintaa verrattuna automaattiseen lukemiseen, mutta ne tarjoavat edullisen ja helposti integroitavan vaihtoehdon.

IOT-TEKNOLOGIAT

IoT-pohjaiset teknologiat, kuten mobiiliverkkoja (esim. 4G/5G) ja satelliittipainnusta hyödyntävät laitteet, mahdollistavat esineiden seurannan laajalla alueella (Mounika & Chepuru, 2019). Esimerkiksi kappaleiden liikkeistä asiakkaiden ja laitoksen välillä ulkotiloissa, korvaten tai täydentäen RFID:tä pidemmillä matkoilla. Sisätiloissa niiden tarkkuus on kuitenkin rajallinen, ja signaali voi heikentyä tai katketa rakennusten sisällä. Lisäksi IoT-laitteet vaativat usein virtalähteen ja datayhteyden, mikä nostaa kustannuksia verrattuna passiivisiin RFID-tunnisteisiin. Tämä teknologia sopii parhaiten ulkoseurantaan, mutta ei vastaa kokeilujemme sisätilojen tarpeisiin yhtä hyvin kuin RFID tai viivakoodit.

YHTEENVETO

Tarkastelimme RFID-teknologian soveltuvuutta metallialan yrityksissä osana VauhtiData-hanketta keskittyen materiaalivirtojen seurantaan teollisuusympäristöissä. Teoria kertoo, että RFID-teknologia voi merkittävästi tehostaa reaaliaikaista seurantaa ja varastonhallintaa. Käytännön kokeilut kuitenkin osoittivat, että käyttöympäristö voi asettaa haasteita, jotka vaativat huolellista suunnittelua ja teknisten ratkaisujen kokeilua. Tulokset korostavat, että teknologian käyttöönotto on ympäristökohtaista: joissakin tapauksissa se toimii erinomaisesti, kun taas toisissa tarvitaan lisäoptimointia tai vaihtoehtoisia menetelmiä.

Metallialan yritysten digitalisaation edistämiseksi on tärkeää jatkaa erilaisten teknologioiden testaamista ja kehittää kustannustehokkaita ratkaisuja, jotka vastaavat teollisuuden tarpeisiin. RFID:n ja muiden seurantateknologioiden yhdistelmä voi tarjota joustavan tavan parantaa materiaalivirtojen hallintaa, vähentää hukkaa ja tukea kestävästä kehityksestä. Tämä tukee yritysten kilpailukykyä ja avaa uusia mahdollisuuksia datavetoiseen päätöksentekoon.

LÄHTEET

European Commission. (2020). *A European strategy for data (COM(2020) 66 final)*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0066>

Finkenzeller, K. (2010). *RFID handbook: Fundamentals and applications in contactless smart cards, radio frequency identification and near-field communication* (3. p.). Wiley. <https://doi.org/10.1002/9780470665121>

Gao, J. Z., Prakash, L. & Jagatheesan, R. (2007). Understanding 2D-barcode technology and applications in m-commerce: Design and implementation of a 2D barcode processing solution. *31st Annual International Computer Software and Applications Conference (COMPSAC 2007)*, 2, 49–56. <https://doi.org/10.1109/COMPSAC.2007.229>

McKinsey & Company. (28.1.2022). *The data-driven enterprise of 2025*. Haettu 31.3.2025 osoitteesta <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-data-driven-enterprise-of-2025>

Mounika A. & Chepuru A. (2019). IOT Based Vehicle Tracking and Monitoring System Using GPS and GSM. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(2S11), 2399–2403. <https://doi.org/10.35940/ijrte.b1275.0982s1119>

Sahinoglu, Z., Gezici, S. & Guvenc, I. (2008). *Ultra-wideband positioning systems: Theoretical limits, ranging algorithms, and protocols*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511541056>

Weinstein, R. (2005). RFID: A technical overview and its application to the enterprise. *IT Professional*, 7(3), 27–33. <https://doi.org/10.1109/MITP.2005.69>

Zhang, J., Tian, G. Y., Marindra, A. M. J., Sunny, A. I. & Zhao, A. B. (2017). A review of passive RFID tag antenna-based sensors and systems for structural health monitoring applications. *Sensors*, 17(2), 265. <https://doi.org/10.3390/s17020265>

MATEMAATTINEN MALLINTAMINEN: JUNAN OVIE AVAUTUMINEN JA SULKEUTUMINEN

JARI KYNGÄS, SATAKUNNAN AMMATTIKORKEAKOULU

Matemaattinen mallinnus on yksi tapa etsiä datassa esiintyviä yhteyksiä ja riippuvuuksia. Matemaattisia malleja voidaan käyttää poikkeuksellisten tilanteiden havaitsemiseen tarkastelemalla, miten hyvin todellisen elämän mitatut arvot sopivat yhteen matemaattisen mallin kanssa. Tässä artikkelissa tarkastellaan tapaustutkimusta, jossa matemaattista mallinnusta käytettiin junan ovien ohjausyksikön hajoamisen ennustamiseen.

Datasta voidaan etsiä yhteyksiä ja riippuvuuksia useilla erilaisilla tavoilla. Yksinkertaisimmillaan data käydään lävitse etsien muuttujien kaikki arvot ja mahdollisesti luokittelemalla ne. Erilaisten löydösten perusteella voidaan tehdä yhteenvetoja ja mahdollisesti erilaisia päätelmiä. Monimutkaisimmillaan proseduuri voi sisältää edellisen lisäksi sopivien tilastollisten menetelmien tai tekoälyn käyttämistä, simulointia ja mallintamista. Tässä paperissa keskitytään matemaattiseen mallintamiseen ilman valmiiden (tilastollisten) menetelmien käyttämistä.

Matemaattisella mallintamisella voidaan tarkoittaa useita asioita (Heiliö & Pohjolainen, 2010), mutta yleisesti ottaen se vaatii jonkinlaisen laskennan mukana olemista.

Matemaattisen mallintamisen prosessi etenee suurin piirtein näin:

1. Meillä on ongelma, joka meidän tulee ratkaista ja johon ei ole olemassa olevaa teoreettista ratkaisua saatavilla.
2. Rakennamme matemaattisen mallin kuvaamaan ongelmaa. Malli voi olla aluksi hyvinkin yksinkertainen, koska sitä voidaan parannella, kun tieto lisääntyy.

3. Mallia testataan erilaisilla syötteillä – vaihetta kutsutaan simuloinniksi.

4. Simuloinnin tuloksia verrataan todellisen ilmiön tuottamiin tuloksiin eli validoidaan malli. Mikäli todellisia tuloksia ei ole, mallin hyvyuden arviointi jää tekijöiden ja mahdollisten asiantuntijoiden harteille.

5. Toistetaan vaiheita 2–4 kunnes tulos on tyydyttävä.

Tämän artikkelin esimerkissä käytetään hyvin yksinkertaista mallintamista. Matemaattisella mallintamisella voidaan pyrkiä selvittämään todellisen ilmiön taustalla oleva matemaattinen yhteys mahdollisimman tarkasti, mutta esimerkin tapauksessa on riittänyt päästä hyvään lopputulokseen.

CASE: POLARTEKNIK

Polarteknik valmistaa junan ovia, ja näissä ovissa on erilaisia teknisiä ratkaisuja, jotka kuluessaan ja/tai ikääntyessään menevät epäkuntoon. Ovissa on hihna, jota moottori liikuttaa ja minkä seurauksena ovet avautuvat ja sulkeutuvat. Hihna voi käytössä löystyä, tai hihnarakenteen sekaan voi joutua likaa ja roskaa. Polarteknik halusi selvittää, voiko oven epäkuntoon menemistä ennakoida.

Ongelmaa lähdettiin ratkaisemaan niin, että aluksi ”nauhoitettiin” täysin kunnossa olevan junan oven avautumista ja sulkeutumista. Nauhoitukset toteutettiin kahdella eri menetelmällä, joista ensimmäisessä hyödynnettiin D56-junan ovien ohjausyksikköä (Polarteknik, 2025) ja toisessa oskilloskooppia (Pico Technologies, 2025). Oskilloskooppi nauhoitti oven avauskäskyn, kotirajalla olamisen (ovi kiinni), avausmoottorin voltit, ampeerit ja avausmoottorin pulssit.

Oskilloskoopilla tehdyt moottorin jännite- ja virtamittaukset epäonnistuivat laitteessa jälkikäteen havaitun vian vuoksi. Niinpä näistä arvoista ainoastaan avausmoottorin pulsseista oli hyötyä. Pulssien avulla saatiin luotua yksinkertainen ennustusfunktio oven rikkoutumiselle, mutta arvojen epäluotettavuuden takia mallia ei otettu käyttöön. Lopulta kaikki arvot ja mittaukset jätettiin huomiotta niiden epäluotettavuuden vuoksi. Syytä epäluotettavuuteen emme onnistuneet selvittämään.

D56-junan ovien ohjausyksikön mittauservot sen sijaan vaikuttivat olevan hyvinkin luotettavia. Laitte mittasi 28 eri asiaa, joista tärkeimmät olivat laitteen moottorin voltit ja ampeerit sekä kymmenkunta oven paikkaan liittyvää muutujaa. Ohjausyksikön antamien arvojen perusteella pystyttiin tekemään yksinkertainen matemaattinen malli, jonka avulla oven rikkoutumisen saattoi ennakoida 100 %:n varmuudella.

OVIENTÄRKKÖTUMISEN MALLINTAMINEN

Mallinnuksessa pyrittiin selvittämään mitattujen muuttujien jakauma – mikäli eri mittaustilanteet johtavat muuttujien erilaisiin jakaumiin se tarkoittaisi sitä, että laitteistossa on tapahtunut jotakin (epäedullista). Muuttujien jakauma selvitetiin keskiarvon, keskihajonnan, huipukkuuden (*kurtosis*) ja vinouman (*skewness*) avulla.

Hihnan kireyttä mitataan hertseillä. Arvot voidaan oviin laittaa välille 5–15 hertsiä. Mitä isompi arvo on, sitä kireämpi on hihna ja sitä paremmin ovi toimii. Alle viiden hertsin arvoilla junan ovi ei enää avaudu ollenkaan. Hihnan löysyys havaittiin pulsseista hertsiarvojen 5–8 välillä suhteellisen helposti. Tästä oli helppo luoda matemaattinen malli, joka osaa varoittaa hihnan liiallisesta löysyydestä. Kahdeksan hertsin kohdalla tarvittiin 20–30 oven avausta ja sulkeamista ja viiden hertsin kohdalla vain 2–5.

Oven tiukkuutta kontrolloitiin esteillä, jotka kiinnitettiin oven ja kiskon väliin. Tiukkuus voi saada arvoja väliltä 25–70 N, ja mitä isompi arvo on, sitä tiukempi ovi on sulkea. Arvoina käytettiin lukemia 25, 30, 42, 52 ja 70. Tässä kohden pulssit alkoivat ilmoittaa liian tiukasta kiskosta 52 N kohdalla. Tähän oli myös helppo tehdä matemaattinen kaava, joka ilmoitti oven liiallisesta tiukkuudesta.

LOPPUSANAT

Tilastotieteilijä (tai data-analyytikko) on yleensä tilastomenetelmien ja mallintamisen ekspertti, mutta hän ei useimmiten tiedä tutkimuksen kohteesta tai luonteesta mitään. Niinpä jokainen projekti tarvitsee myös asiantuntijan, joka osaa kertoa mittaustulosten tarkoituksesta ja/tai erilaisten laitteiden toiminnasta. Asiantuntijan ja tilastotieteilijän yhteistyön seurauksena syntyy hyvää ja merkityksellistä tutkimusta, joka voi auttaa yritystä eteenpäin.

LÄHTEET

Heiliö, M. & Pohjolainen, S. (2010). *Matemaattinen mallinnus*. WSOYpro.

Polarteknik. (2025). *NEW Control Unit D56*. <https://www.polarteknik.fi/wp-content/uploads/D56-control-unit-d56-borchure.pdf>

Pico Technologies. (2025). *PicoScope 4000A Series*. <https://www.picotech.com/oscilloscope/4000/picoscope-4000-series>

KULJETUSPÄÄSTÖJEN MITTAAMINEN KILPAILUETUNA: DATALÄHTÖINEN LÄHESTYMISTAPA

EPPU HEILIMO, JYVÄSKYLÄN AMMATTIKORKEAKOULU

Kuljetuspäästöjen mittaaminen on siirtynyt vapaavalintaisesta lisäarvosta keskeiseksi kilpailutekijäksi valmistavan teollisuuden alalla. Tässä artikkelissa käsitellään kuljetuslogistiikan hiilijalanjäljen laskentaa standardien mukaisesti sekä käytännön ratkaisuja, jotka tarjoavat yrityksille kilpailuetua ympäristötietoisilla markkinoilla.

KULJETUSPÄÄSTÖJEN SEURANTA KILPAILUETUNA

Ympäristövastuu on muodostunut välttämättömäksi osaksi nykyaikaista liiketoimintaa, erityisesti valmistavan teollisuuden alalla. Kuljetuslogistiikan hiilijalanjäljen mittaaminen ja vähentäminen on siirtynyt vapaavalintaisesta lisäarvosta keskeiseksi kilpailutekijäksi (European Environment Agency, 2022). Tämä muutos heijastaa sekä kiristyviä ympäristösäädöksiä että asiakkaiden kasvavia odotuksia.

Valmistavan teollisuuden yrityksille CO₂-päästöjen seuranta on kehittynyt merkittäväksi strategiseksi vahvuudeksi. Kehittyneitä dataratkaisuja hyödyntävät yritykset saavuttavat kyvyn mitata toimitusketjunsä ympäristövaikutuksia tarkasti ja reaaliaikaisesti. Ympäristötietoisuuden kasvaessa kuluttajien keskuudessa myös yritysten välinen kilpailu kiristyy vastuullisuuden saralla.

Valmistusyritykset, joiden toimitusketjut ulottuvat usein globaalisti, kohtaavat erityisiä haasteita päästöjen seurannassa. Kuljetuspäästöjen mittaaminen ei kuitenkaan vaadi monimutkaisia järjestelmiä. Modernit datapohjaiset ratkaisut tarjoavat kustannustehokkaan tavan päästölaskentaan kansainvälisten standardien mukaisesti (ISO, 2023).

Automatisoitu päästölaskenta mahdollistaa standardoidut mittarit jokaiselle lähetykselle. Tämä tuo läpinäkyvyyttä, joka on yhä merkittävämpi kilpailuetu tarjouspyynnöissä ja hankintaprosesseissa. Monet yritykset käyttävät vertailuvia päästöanalyyskejä myyntivalttina ja erottautuvat kilpailijoista, joilta puuttuu tämä ympäristödatan hyödyntämisen kyky.

CO₂-PÄÄSTÖLASKENNAN TEKNISET PERUSTEET

Kuljetuspäästöjen laskenta perustuu kansainväliseen ISO 14083 -standardiin ”Kasvihuonekaasut. Kuljetusketjujen kasvihuonekaasupäästöjen määrittäminen ja raportointi” (ISO, 2023). Tämä standardi tarjoaa yhtenäisen viitekehyksen päästölaskennalle ja varmistaa tulosten vertailukelpoisuuden.

Kyseisen standardin mukaisesti kuljetuspäästöjen peruskaava on
 $CO_2\text{-päästöt} = \text{Etäisyys} \times \text{Päästökerroin} \times \text{Rahdin paino}.$

Standardin mukaiset päästökertoimet huomioivat kuljetusmuodon (maantie, meri, rautatie, lento), kaluston tekniset ominaisuudet sekä polttoainetyypit. Kuljetusmuotojen välillä on merkittäviä eroja: lentorahdin päästöt ovat tyypillisesti moninkertaiset merikuljetuksiin vertaillen samantapista matkaa.

Ajantasaisen reittidatan ja liikennetietojen yhdistäminen laskentamalliin voi huomattavasti parantaa päästölaskennan tarkkuutta verrattuna yksinkertaisiin etäisyyspohjaisiin malleihin. Tällöin huomioidaan esimerkiksi todelliset kuljetusreitit, liikennetiheys ja maasto-olosuhteet.



HAASTEET

Datapohjaisen päästölaskennan toteuttamisessa kohdataan tyypillisiä haasteita. Yleisimpiä ongelmia ovat datan laadun vaihtelut, puutteelliset lähtötiedot sekä standardien ja menetelmien jatkuva muuttuminen. Tulevaisuuden kehityskulkuja määrittää vahvasti tutkimus- ja yritys yhteistyö. Satelliittipaikkannuksen, reittinavigoinnin ja tekoälypohjaisten ennustemallien integrointi päästölaskentaan tulee todennäköisesti parantamaan laskennan tarkkuutta merkittävästi tulevina vuosina. Tämä jatkuva kehitys tekee päästölaskennasta hyvinkin monimutkaista.

”CO₂-päästölaskennan tekniset perusteet”-kappaleessa mainittu päästöjen laskukaava vaikuttaa yksinkertaiselta, mutta käytännön sovelluksissa haasteita aiheuttavat erityisesti tarkat etäisyyslaskelmat eri kuljetusmuodoilla. Esimerkiksi merikuljetusten todellisten reittien mallintaminen, rautatiekuljetusten energiankulutuksen vaihtelu eri maastonmuodoissa tai maantiekuljetusten optimaalisten reittien laskeminen ruuhka-aikoja huomioiden on niin monimutkaista, että oman laskentajärjestelmän kehittäminen ei ole useimmille yrityksille kannattavaa.

Tämän vuoksi käytännön päästölaskennassa hyödynnetään tyypillisesti valmiita karttapalveluita, kuten Google Maps, HERE Location Services tai Map-Box, jotka tarjoavat rajapintoja tarkkojen kuljetusreittien ja -etäisyyksien laskentaan. Vaihtoehtoisesti yritykset voivat käyttää kokonaisvaltaisia päästölaskentapalveluita kuten EcoTransIT World, NTM Calc tai BigMile, jotka yhdistävät reittitiedot suoraan päästökertoimiin eri kuljetusmuodoille.

KÄYTÄNNÖN IMPLEMENTOINTI: CASE-ESIMERKKI PÄÄSTÖLASKENNASTA

Kuljetuspäästöjen laskentaan on olemassa erilaisia työkaluja, joiden käyttöönotto voidaan toteuttaa yrityksen tarpeiden mukaan joko kevyesti tai integroidummin. Tässä esimerkissä käsittelemme EcoTransIT World -palvelun hyödyntämistä, joka on yksi markkinoilla olevista vaihtoehdoista.

Päästölaskennan voi aloittaa yksinkertaisimmillaan käyttämällä verkkoselaimessa toimivia laskureita. Näissä käyttäjä syöttää kuljetuksen perustiedot, lähtö- ja määräpaikat, kuljetusmuodon ja rahdin painon, minkä jälkeen laskuri tuottaa arvion kuljetuksen hiilijalanjäljestä. Tällainen lähestymistapa soveltuu yksittäisten kuljetusten tai kuljetusvaihtoehtojen vertailuun, mutta muuttuu nopeasti työlääksi suuremmilla kuljetusmäärillä.

Systemaattisempaan päästölaskentaan tarvitaan automatisoitu ratkaisu. Koska Excel on yleisesti käytössä monissa logistiikkayrityksissä, pilotoimme tutkimuksessamme EcoTransIT-palvelun käyttöä Excel-tiedostoista käsin. Pilotissa mallinsimme logistiikkaketjun kolmessa vaiheessa: alkukuljetus, pääkuljetus ja loppukuljetus. Käyttäjä syöttää kaikkien vaiheiden tiedot (lähtö- ja määräpaikat sekä kuljetusmuodon) yhdelle riville, minkä jälkeen ohjelma laskee automaattisesti kunkin vaiheen päästöt, yhteispäästöt ja kokonaismatkan pituuden.



Satoja kuljetuksia voidaan analysoida kerralla, mikä mahdollistaa erilaisten kausittaisten vertailujen ja trendien tunnistamisen.

Pilottitoteutuksessa Excel VBA -koodia käytettiin käsittelemään monimutkaisen kuljetusketjun rakennetta. Testasimme, miten standardoituja XML-kyseilyjä voidaan muodostaa rajapintaan ja jäsentää vastauksista oleelliset hiilidioksidipäästötiedot. Erityistä huomiota kiinnitettiin paikkatietojen käsittelyyn, sillä järjestelmä vaatii maakoodien ja kaupunkien nimet standardoidussa muodossa.

Käyttäjystävällisyyttä voitaisiin lisätä integroinnilla Google Maps -palveluun, jossa kuljetuspisteet näkyisivät hyperlinkkeinä suoraan Excel-taulukossa. Tämä auttaisi nopeasti visualisoimaan kuljetusreitit ja tunnistamaan mahdollisia optimointikohtia.

Tällaisen lähestymistavan vahvuus on sen kyky käsitellä suuria datamääriä tehokkaasti. Satoja kuljetuksia voidaan analysoida kerralla, mikä mahdollistaa erilaisten kausittaisten vertailujen ja trendien tunnistamisen. Huomionarvoista on kuitenkin, että automatisoitujen rajapintaratkaisujen käyttö on tyypillisesti maksullista, kun taas verkkopohjaiset laskurit ovat usein ilmaisia. Vaikka Excel VBA -työkalut eivät välttämättä edusta nykyaikaisia ohjelmistoratkaisuja, niiden etuna on suora integroitavuus yritysten jo olemassa oleviin Excel-taulukoihin, mikä madaltaa käyttöönoton kynnystä huomattavasti.

YHTEENVETO

Datapohjainen lähestymistapa kuljetuspäästöjen laskentaan tarjoaa valmistavan teollisuuden yrityksille konkreettisen työkalun ympäristövastuun ja liiketoiminnan kehittämiseen. Standardeihin pohjautuva päästölaskenta mahdollistaa kuljetusketjun hiilijalanjäljen luotettavan mittaamisen ja vertailun.

Artikkelissa kuvatut käytännön ratkaisut, kuten verkkopohjaiset laskurit tai rajapintaintegraatiot, osoittavat, että päästölaskentaa voidaan toteuttaa eri tasoilla yrityksen tarpeiden mukaan. Huolimatta teknisistä haasteista, kuten etäisyyslaskelmien monimutkaisuudesta tai datan laadun vaihteluista, datapohjaiset ratkaisut tekevät päästöjen seurannasta entistä saavutettavampaa.

Kuljetuspäästöjen analytiikka ei ole vain vastuullisuuskysymys, vaan kasvavassa määrin strateginen kilpailuetu markkinoilla. Läpinäkyvä päästölaskenta vahvistaa yrityksen asemaa tarjousprosesseissa ja heijastaa toiminnan tehokkuutta. Tulevaisuudessa päästölaskennan integrointi päätöksentekoon on keskeinen osa valmistusyritysten strategiaa yhä ympäristötietoisemmässä liiketoimintaympäristössä.

LÄHTEET

European Environment Agency. (2022). *Transport and environment report 2022*. <https://doi.org/10.2800/47438>

International Organization for Standardization (ISO). (2023). *ISO 14083:2023 Greenhouse gases – Quantification and reporting of greenhouse gas emissions arising from transport chain operations*. <https://www.iso.org/standard/78864.html>

Jyväskylän ammattikorkeakoulun julkaisut.

Avoimet julkaisumme tekevät tunnetuksi
Jamkin laajaa ja monialaista
opetus-, tutkimus- ja kehittämistyötä.

▶ jamk.fi/julkaisut



Jyväskylän ammattikorkeakoulu

PL 207, 40101 Jyväskylä
Rajakatu 35,
40200 Jyväskylä
Puh. +358 20 743 8100
Fax. +358 14 449 9694

jamk.fi

ISBN 978-951-830-791-7 (Painettu)
ISBN 978-951-830-792-4 (PDF)
ISSN-L 1456-2332



Euroopan unionin
osarahoittama

jamk | Jyväskylän
ammattikorkeakoulu

