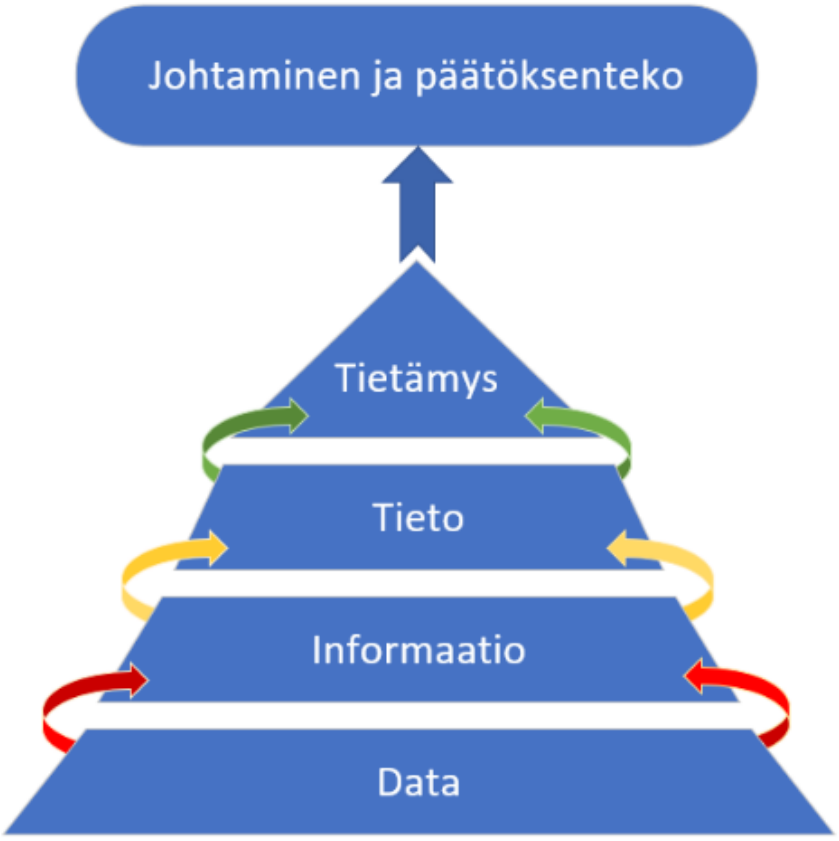
Datan käytettävyys päätöksissä ja tekoälymalleissa

7.12.2021

Nyt kasvussa olevaan, mutta jo pitkään esillä olleeseen tekoälytrendin äärelle on jälleen palattu. Viime kerran trendi nosti päätään vuosikymmen sitten. Tekoälykehittämisen rintamalla on viimeisten vuosien aikana otettu suuria teknologisia harppauksia. Tekoälykehityksestä voidaan hyötyä yhtälailla robotisaation ja automatisaation saralla. Valmistavalla teollisuudenaloilla on VTT-tutkimuksen mukaan haasteita osaavan työvoiman löytämiseksi. Robotit ja automatisointi ovat tutkimuksen mukaan PK-yrityksille menetelmä säilyttää ja vahvistaa kustannustehokkuutta ja toimitusvarmuutta. Yleisesti ottaen automaattisten toimintojen käyttöönoton vaikutukset ovat kilpailukykyä parantavia kehittyvillä markkinoilla.

Roboteista ja erilaisista prosessien ohjausjärjestelmistä on mahdollista kerätä suuri määrä hyödyllistä dataa. *Dataa* analysoimalla ja prosessoimalla sitä tulkittavaan muotoon informaatioksi, mahdollistetaan tiedolla johtaminen.

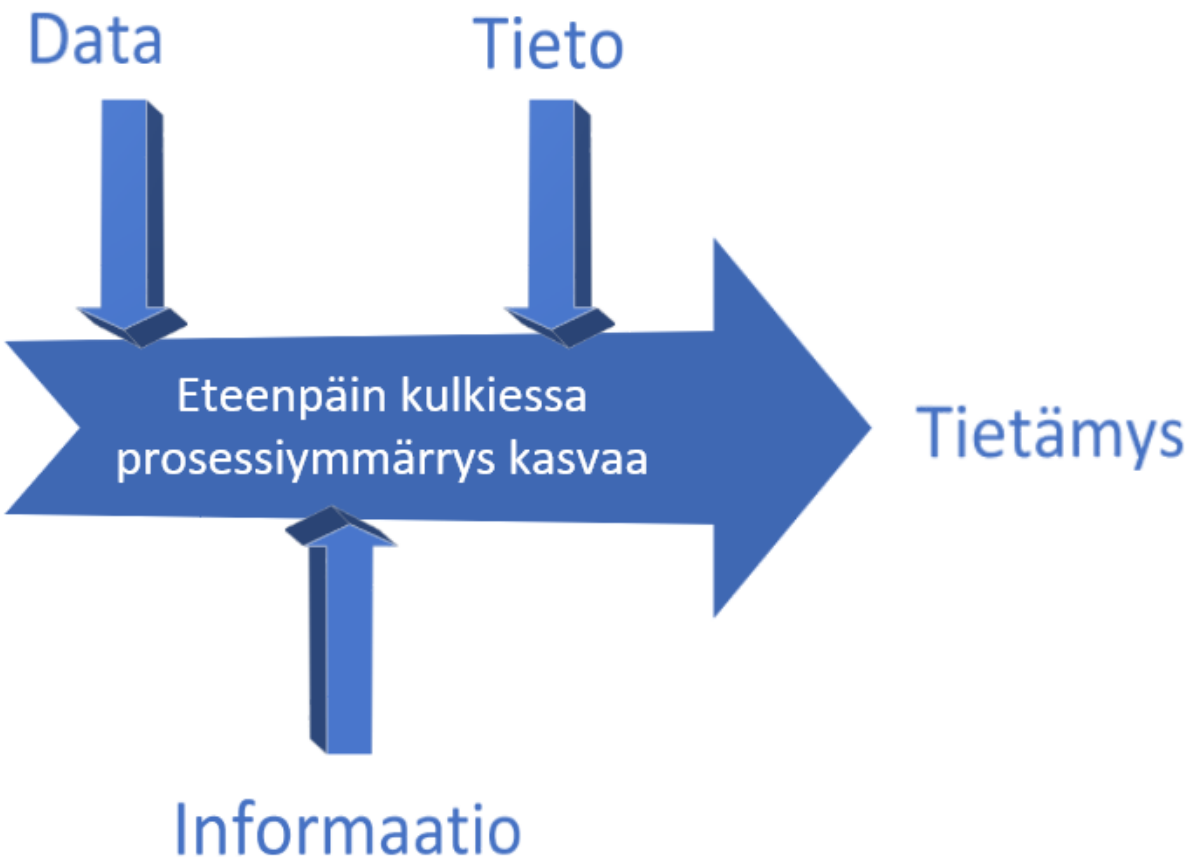


Kuva 1. DIKW-malli (Data, Information, Knowledge, Wisdom).

Päätöksien laatuun vaikuttaa keskeisesti datan laatu ja kuinka siitä johdettu informaatio esitetään. Näin ollen raporttien, eli datan analyysien laadulla on merkittävä vaikutus tiedolla johtavan päättäjän päätöksentekoon. Päätöksentekijän tietämys ja kausaliteettien ymmärrys myös kasvaa ja kehittyy päätösten seuraamuksista. Ihmisen kontekstuaalista älykkyyttä ja prosessointikykyä tarvitaan edelleen tulkitsemaan piilevää

informaatiota, joka ei välttämättä aina välity kerätystä datasta.*Informaatio* puolestaan tuo mahdollisuuden tehdä *tietoon* perustuvia strategisia päätöksiä lyhyellä ja pitkällä aikavälillä. Tiedon onnistunut käyttöönotto ja siitä muotoutuvista ohjeistuksista, uudesta toiminnasta ja uusista toimintaprosesseista voidaan kutsua DIKW-mallin mukaisesti *tietämyksen* saavuttamiseksi.

Data on näin ollen merkittävin tiedolla johtajan päätösten raaka-aine olemassa olevaa tietämyksen soveltamista unohtamatta.



Kuva 2. Prosessiymmärryksen roadmap.

Edellä mainitusta on johdettu hyvin data-analysointiin perehtyneiden keskuudessa tuttuun *”garbage in, garbage out”* -lausahdus. Suomeksi tämä tarkoittaa, että huonosta datasta tietosaanto on parhaimmillaankin heikkoa ja epäedustavaa. Tämä pätee myös siitä johdettuihin analyyseihin, kuin myös tekoälymallinnuksiin.Myös tekoälyn kohdalla data on merkittävin raaka-aine yksittäisten tapahtumien ja monikärkisten kausaliteettien mallinnuksessa. Tekoälymallin tuottamaan analyysiin ja erityisesti mallin tuottamiin ennusteisiin, sekä niiden tarkkuuteen voidaan vaikuttaa merkittävästi datan laatua parantamalla. Saatavilla olevan datan laadun merkitys kaikkien kehitettävien tekoälymallien –teknologiasta riippumatta, on keskeisin asia koko tekoälyä käsittävällä tieteenalalla. Kun datan laatu on varmistettu, voidaan keskittyä erikoistaviin tekoälytekniikoihin, ts. käytettäviin tekoäly-algoritmeihin.

Datan laadun ja käytettävyyden varmistamiseksi, keräyksessä on hyvä noudattaa suunnitelmallisuutta. Datan keräämisessä on tärkeää olla ennakkoon määriteltynä seurattavaa tapahtumaa tai tiedonkeruukohdetta tukeva ETL-prosessi. ETL-prosessi kuvaa menettelyä, jossa tunnistetaan mitä raakaa prosessoimatonta

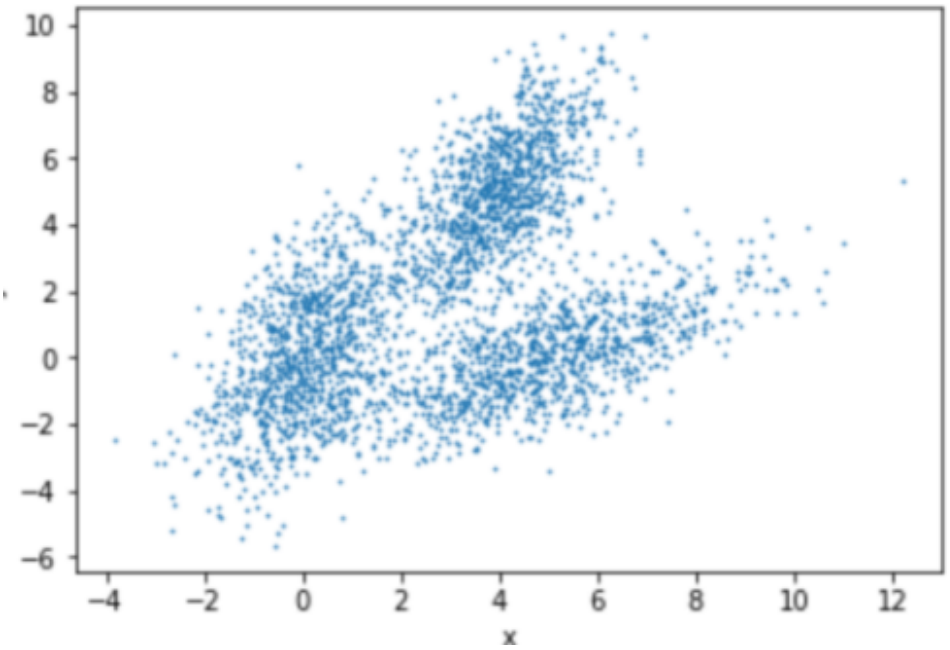
dataa, esimerkiksi sensoreista kerättävää dataa, yhdistetään muuhun tietystä tapahtumaketjusta kerättävään dataan ja missä kyseinen data tai informaatio on saatavilla.



Kuva 3. ETL-prosessin rakenne termin osalta.

”Extract, transform, load – Käännetty englannista-Laskennassa, poimiminen, muuntaminen, lataus on yleinen menetelmä tietojen kopioimiseksi yhdestä tai useammasta lähteestä kohdejärjestelmään, joka edustaa tietoja eri tavalla kuin lähde tai eri kontekstissa kuin lähde.” – Wikipedia, ETL

On olemassa myös tilanteita, joissa raasta datasta voidaan yrittää koostaa informaatiota hajontakuvio-tyyppisien analyysien avulla. Hajonnasta pyritään etsimään datasta tai sen koosteista yhdenmukaisuuksia. Asiaan tarkemmin perehtyneille usein puhutaan, että etsitään datassa esiintyviä luokkia.



Kuva 4. Eräs datapisteistä koostettu scatter-plot eli hajontakuvio.

Datan keräämisessä, teknisestä näkökulmasta katsoen datan varastointiin on useita vaihtoehtoja.

Tietovarastot voivat tyypillisesti rakentua joko toisistaan erillistä atomista dataa sisältävistä, tapahtumakokonaisuuksia sisältävistä tai asiayhteyksiä sisältävistä data ja tyyppikohtaisesti informaatio- ja tietoalkioista. Yleisemmin kaikkien tietovarastotyyppien yksittäisestä datapisteestä käytetään vain termiä tietoalkio. Erilaisten tietovarastojen sisältämää dataa voidaan myös koostaa ETL-prosessissa ja tässä prosessissa voidaan prosessoida myös useiden tietovarastojen dataa ja koostaa tietoa kunkin analysointikohteen tarpeen mukaan.

Datan keräämisessä keskeistä on selvittää ja ymmärtää kuinka yksittäinen datapiste ja siihen mahdollisesti liittyvät muut datapisteet muodostavat kausaliteetteja. Kausaliteetit ovat datassa esiintyviä syy-seuraus tyyppisesti toisiinsa sidonnaisia osakokonaisuuksia kerättyyn dataan nähden.

Datan keräyksessä ja ETL-prosessin toteutuksessa suurena apuna on, mikäli yrityksessä on tietokanta-asiantuntemusta ja sopivan lämpimät suhteet järjestelmätoimittajiin. Erilaiset tietokantatyypit mahdollistavat monimutkaistenkin asiayhteyksien datan tallentamisen prosessista ja saadun datan hyödyntämisen jatkojalostusvaiheeseen. Usein järjestelmätoimittajat ovat halukkaita avaamaan yritykselle heidän omistaman dataa tai sen turvallisen kopion.

ETL-prosessilla varmistetaan laadukkaan informaation saanto ja siten mahdollistetaan tulevaisuudessa tarkempien, tehokkaampien ja laadukkaampien raporttien, kuin myös mahdollisesti harkinnassa olevien tekoälymallien rakentaminen.

Muistisääntönä voi todeta, että data ei itsessään ole kovin tärkeää, jos sille ei ole määritetty prosesseissa merkitystä. Kerätään siis dataa siten, että sen käyttötarkoitus on tunnistettu ja sovelletaan datan tallennukseen ETL-prosessissa määritettyjä ja kehitettyjä menetelmiä. Käytetään myös tarpeen tullen asiantuntija-apua tietovarastojen rakentamisessa ja tarvittavien järjestelmien käyttöönotoissa.

Kun dataa on ja se on laadukasta, datasta saatavasta informaatiosta voidaan tehdä laadukkaita johtopäätöksiä ja mallinnuksia tekoälyn avulla.

Toni Takala
TKI-asiantuntijaSeAMK Tekniikka

Lähteet:

Ohjelmisto ja tekoälyinvestoinnit PK-yrityksissä

https://www.vttresearch.com/en/news-and-ideas/survey-smes-reluctant-invest-new-technologies-help-availablehttps://pages.vistage.co.uk/hubfs/Vistage%20Research%20-%20Artificial%20Intelligence%20for%20SMEs-%20UK-IRL%20Version.pdf

DIKW

https://towardsdatascience.com/rootstrap-dikw-model-32cef9ae6dfbhttps://otec.uoregon.edu/data-wisdom.htm

ETL

https://en.wikipedia.org/wiki/Extract,_transform,_loadhttps://www.geeksforgeeks.org/etl-process-in-data-warehouse/

Datan laatu

https://towardsdatascience.com/data-quality-garbage-in-garbage-out-df727030c5ebhttps://www.researchgate.net/profile/Marco-Valtorta-2/publication/220918649_The_Effects_of_Data_Quality_on_Machine_Learning_Algorithms/links/09e4150a506b15472b000000/The-Effects-of-Data-Quality-on-Machine-Learning-Algorithms.pdf