

Tiedonlouhinta ja sen mahdollisuudet

Henry Joutsijoki

Sisältö

Johdanto

Tiedonlouhinta

Koneoppiminen ja tiedonlouhinta

Tiedonlouhinnan tulevaisuus

Alustusta

- Nyky-yhteiskunnassamme käsitteet tehokkuus, nopeus, tarkkuus, maksimointi ja minimointi korostuvat yhä useammin monissa asiayhteyksissä.
- Taloudellinen tilanteemme johtaa siihen, että olemassaolevia resursseja pitää käyttää mahdollisimman monipuolisesti.
- Tästä seuraa edelleen, että uusia aluevaltauksia ja innovaatioita tulee kehittää perinteisten alojen rinnalle.
- Kilpailu kovenee kuitenkin jatkuvasti joka sektorilla, sillä samankaltaisia ideoita syntyy ympäri maailmaa.

Alustusta

- Viimeisen vuosikymmenen aikana ympäri maailmaa on kasvanut suurta kiinnostusta erilaisten datojen analysointiin.
- Digitaalisessa muodossa olevan tiedon määrä on valtava ja kasvaa erittäin nopeasti vuosi vuodelta.
- Teknologian kehittymisen myötä dataa kertyy lukuisista eri lähteistä ja kerätyt aineistot muodostavat muun muassa hyvin suuria tietokantoja.
- Viime vuosina erityisesti big data on noussut pinnalle ja sen ympärillä tehdään aktiivisesti tutkimusta.

Mistä saamme dataa?

- Esimerkkejä datalähteistä:
 1. Pankkien maksutapahtumat
 2. Kännykkäkamera ym. muut kamerat
 3. Supermarkettien ostotapahtumatiedot
 4. Terveystieteiden tutkimusten potilaiden mittaukset
 5. Liikenne
 6. Ilmasto
 7. Sensorit yleisesti ottaen
 8. Teollisuusprosessit
 9. Lokitiedot
 10. Sosiaalinen media ja muut web-aineistot
 11. Hakukoneiden keräämä data

Datan luonne

- Data voidaan jakaa karkeasti kahteen eri tyyppiin riippuen sen lähteestä:
 1. Järjestetty data (esimerkiksi kyselylomake) rajoittuu tietylle toimialalle ja data on luonteeltaan diskreettiä. Järjestettyä dataa säilytetään monesti Excel-tiedostossa tai relaatiotietokannassa.
 2. Järjestämätöntä dataa ei voida suoraviivaisesti lokeroida johonkin luokkaan. Järjestämätöntä dataa ovat esimerkiksi kuvat, videot, pdf-tiedostot, sähköpostit ja sosiaalisen median tuottamat aineistot.

Miksi keräämme dataa?

- Datan sisältämässä informaatiossa nähdään potentiaalia ja mahdollista taloudellista hyötyä muun muassa uusien palveluiden ja liiketoiminnan muodossa.
- Datan käsittelymenetelmien kehittäminen luo uusia työmahdollisuuksia. Suurten tietomassojen käsittely vaatii uudenlaisia teknisiä toteutustapoja.
- Datasta voidaan löytää uusia yhteyksiä eri asioiden välille.
- Datasta saatavan tietämyksen avulla pystytään vahvistamaan olemassaolevia ennakkokäsityksiä.

Sisältö

Johdanto

Tiedonlouhinta

Koneoppiminen ja tiedonlouhinta

Tiedonlouhinnan tulevaisuus

Lähtökohta

- Peruslähtökohtana tiedonlouhinnalle on, että kerätty data muokataan havaintomatriisiksi (taulukoksi).
- Havaintomatriisissa rivit edustavat havaintoja (esim. asiakas) ja sarakkeet mitattavaa ilmiötä kuvaavia muuttujia (esim. ikä, sukupuoli tai tieto siitä, kuinka montaa e-aineistoa asiakas on käyttänyt).
- Havaintomatriisin yleinen muoto:

$$\begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{pmatrix}$$

Datasta tiedonlouhintaan

- Kerätyn “raakatiedon” suuren määrän vuoksi on pitänyt kehittää tietojenkäsittelymenetelmiä (erityisesti *tiedonlouhinnan menetelmiä* [data mining]), jotta oleellinen informaatio saadaan esille datasta [4].
- Tiedonlouhinnalle on olemassa lukuisia määritelmiä.
- Yksi mahdollinen määritelmä tiedonlouhinnalle [4]:

Tiedonlouhinta on (usein laajojen) havaintotietojen, datajoukkojen tietojen välisten suhteiden etsimistä ja datan ominaisuuksien yhteenvedon tuottamista uusin tavoin, jotka ovat sekä ymmärrettäviä että hyödyllisiä käyttäjän näkökulmasta.

Tiedonlouhinta

- Tiedonlouhinnalla saatuja ominaisuuksia ja yhteenvetoja kutsutaan yleensä *malleiksi* tai *hahmoiksi* [4]. Esimerkkejä malleista ovat assosiaatiosäännöt, klusterit (rypäs, ryväs), graafit ja puurakenteet [4].
- Tiedonlouhinta käsittelee monesti dataa, joka on kerätty muuhun tarkoitukseen kuin itse tiedonlouhintaan (esim. lokitiedot) [4].
- Tiedonlouhinta eroaa tilastotieteestä, jossa aineistoa kerätään erityisesti vastaamaan tiettyihin kysymyksiin [4].

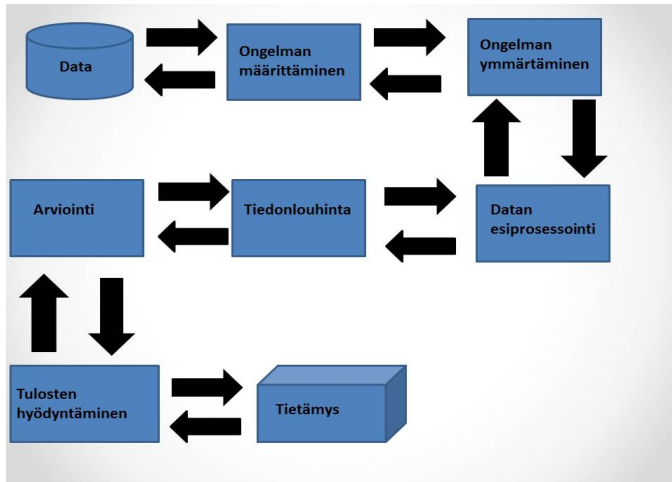
Tiedonlouhinta ja data-analyysi

- Tiedonlouhinnan määritelmän perusteella datajoukot voivat olla usein hyvinkin laajoja (vrt. big data).
- Tiedonlouhinta on data-analyysin yksi osa-alue, joka on keskittynyt ilmiön mallintamiseen ja tietämyksen muodostamiseen erityisesti ennustamisen näkökulmasta [4].
- Tiedonlouhinta ei suinkaan ole ainoa data-analyysin osa-alue, vaan data-analyysi on monipuolinen kokonaisuus lähtien datan keräämisestä ja datan siivoamisesta ulottuen datan visualisointiin ja tulosten analysointiin.

Datasta tietämykseen

- Tiedonlouhinta on yksi vaihe tietämyksen muodostamisessa.
- Lopullinen tavoite käytännössä on, että ymmärryksemme datasta lisääntyy ja pystymme hyödyntämään dataa uusilla tavoilla.
- Tietämyksen muodostaminen on monivaiheinen prosessi, joka alkaa datasta ja ongelman määrittämisestä päättyen tulosten hyödyntämiseen ja tietämyksen muodostamiseen.
- Käytännössä suurten tietomassojen analysointi ja niistä saatavan uuden tietämyksen muodostaminen ei ole helppo tehtävä ja se voi viedä huomattavan paljon aikaa.

Datasta tietämykseen



Kuva : Tietämyksen muodostamisen prosessi. Kuva mukailee lähdettä [1].

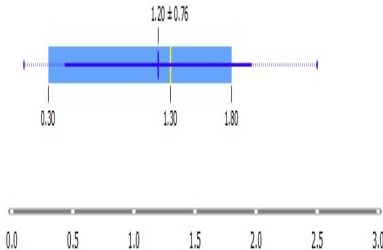
Tiedonlouhinnan tehtäviä

- Tiedonlouhinnan tehtävät voidaan jakaa viiteen ryhmään [4]:
 1. Tutkiva data-analyysi
 2. Deskriptiivinen eli kuvaava mallintaminen
 3. Ennustava mallintaminen
 4. Hahmojen ja sääntöjen etsiminen
 5. Haku sisällön perusteella

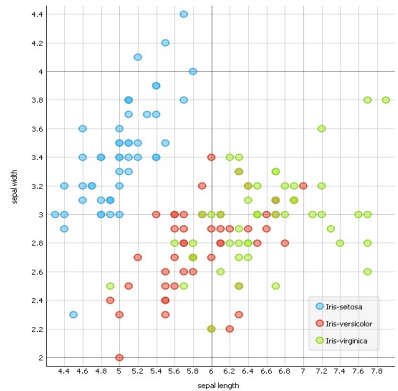
Tutkiva data-analyysi

- Perusideana on tarkastella dataa ilman tarkkaa suunnitelmaa etsittävästä [4].
- Tutkiva data-analyysi on luonteeltaan vuorovaikutteista ja visuaalista [4].
- Pieniä aineistoja varten on olemassa monipuolisia graafisia esitystapoja (histogrammi, sirontakuviot, laatikkokaavio).
- Kun muuttujien lukumäärä kasvaa, datan visualisointi tulee hankalammaksi. Tähän tarkoitukseen on olemassa menetelmiä, jotka tiivistävät alkuperäisen aineiston sisältävän informaation pienempään määrään muuttujia.

Esimerkki



Kuva : Esimerkki muuttujasta määritetystä laatikkokaavio.

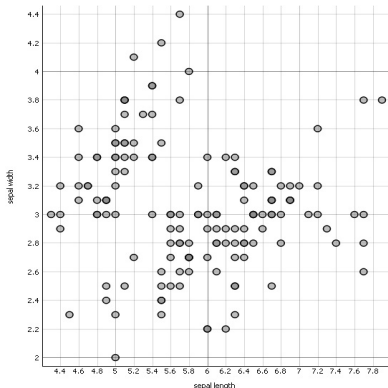


Kuva : Esimerkki kahdesta muuttujasta määritetystä sirontakuviosta.

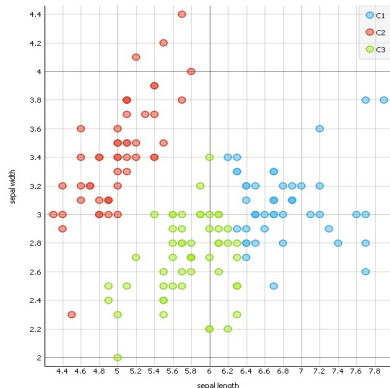
Kuvaava mallintaminen

- Perusideana kuvaavassa mallintamisessa on kuvata koko data (tai sen generoiva prosessi) [4].
- Esimerkkejä kuvaavasta mallintamisesta ovat klusterit/segmentointi ja muuttujien välisten suhteiden kuvausmallit (riippuvuuden mallintaminen) [4].
- Esimerkiksi klusteroinnissa tavoitteena on löytää “luonnolliset ryhmät datasta”.
- Klusterointia voidaan hyödyntää esimerkiksi asiakasprofiilien löytämisessä ja lehtien tai kirjojen ryhmittelyssä, kun niistä on erotettu niitä kuvaavia numeerisia muuttujia.

Esimerkki



Kuva : Alkuperäinen data.



Kuva : Data on klusteroitu kolmeen klusteriin. Klusterit on erotettu toisistaan värikoodauksella.

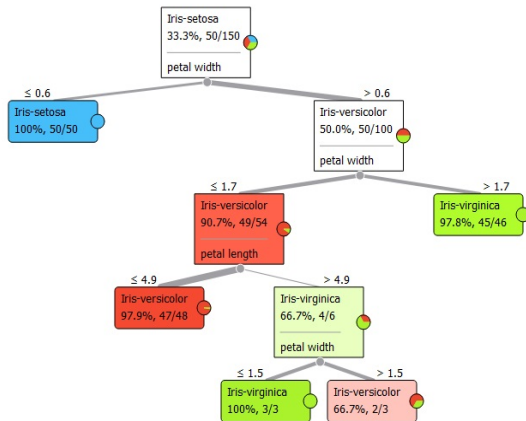
Ennustava mallintaminen

- Tavoitteena muodostaa malli, jonka avulla pystytään ennustamaan yhden muuttujan luokka tai kvantitatiivinen arvo datassa olevien muiden muuttujien arvojen perusteella [4].
- Ennustavassa mallintamisessa erotetaan kaksi tapausta:
 1. luokittelu
 2. regressio.
- Luokittelussa ennustettava muuttuja on luokkatyyppinen ja regressiossa kvantitatiivinen.
- Ennustavaa mallintamista varten on olemassa laaja kokoelma tilastollisia menetelmiä sekä koneoppimismenetelmiä.

Esimerkkejä ennustavasta mallintamisesta

- Oletetaan, että palvelua käyttävät asiakkaat on jaettu kolmeen ryhmään profiilien perusteella.
- Luokittelutehtävässä palvelua käyttävän uuden asiakkaan ryhmä voidaan ennustaa muodostamalla ennustava malli olemassaolevan datan perusteella.
- Potilaan diagnoosin ennustaminen on myös luokittelutehtävä.
- Yksinkertainen esimerkki regressio-ongelmasta on tutkia pituuden ja painon välistä suhdetta ja osoittaa, kuinka muutos pituudessa vaikuttaa painoon.

Esimerkki ennustavasta mallista



Kuva : Esimerkki päätöspuualgoritmin antamasta mallista.

Hahmojen ja sääntöjen etsiminen

- Tavoitteena on muodostaa hahmoja ja sääntöjä laajoista datajoukoista.
- Hahmolla tarkoitetaan lokaalia käsitettä, joka kuvaa datan jostakin erityisestä näkökulmasta, kun taas malli antaa globaalin kokonaiskuvauksen datasta [4].
- Esimerkki: Jos palvelun käyttäjät on jaettu erillisiin ryhmiin heidän profiiliensa perusteella, voidaan tutkia, millaisia e-aineistoja kunkin ryhmän sisällä luetaan. Löytyykö kullekin ryhmälle oma hahmo/sääntö?
- Ostoskoridata on hyvä esimerkki sääntöjen etsimisestä, jota käytetään käytännössäkin.

Haku sisällön perusteella

- Tavoitteena on löytää käyttäjän antaman hahmon (esim. joukko avainsanoja) perusteella samanlaisia hahmoja tietojoukosta [4].
- Hakua sisällön perusteella esiintyy erityisesti teksti- ja kuva-aineistojen tapauksessa.
- Esimerkiksi tekstiaineistojen tapauksessa käyttäjä voi etsiä annettujen avainsanojen (hahmon) avulla keskeisiä dokumentteja verkosta/palvelusta.

Sisältö

Johdanto

Tiedonlouhinta

Koneoppiminen ja tiedonlouhinta

Tiedonlouhinnan tulevaisuus

Koneoppimisen suhde tiedonlouhintaan

- Koneoppimisen suosio on noussut valtavasti viime vuosina ja se on keskeisessä roolissa tiedonlouhintaprosessin onnistumisen kannalta.
- Suurista tietomassoista saatava tietämys perustuukin usemmiten koneoppimisalgoritmien “älykkyyteen”.
- Koneoppiminen ja tiedonlouhinta ovat hyvin läheisiä käsitteitä keskenään. Tiedonlouhinta voidaan käsittää koneoppimisen sovellusalaksi. [4]
- Koneoppimisen piirissä kehiteltyjä menetelmiä (algoritmeja) voidaan hyödyntää tiedonlouhinnassa [4].
- Keskeisintä koneoppimismenetelmien kohdalla on niiden yleispätevyys. Koneoppimisalgoritmien käyttäminen ei ole kiinnitetty johonkin tiettyyn alaan, vaan niitä voidaan hyödyntää kaikkialla, missä dataa on saatavilla.

Koneoppiminen

- Koneoppimiselle on olemassa lukuisia määritelmiä riippuen tarkasteltavasta lähteestä.
- Koneoppimisen lähtökohta on, että kone/tietokone suorittaa oppimisprosessin ihmisten asemesta.
- Mitä on oppiminen?
- Oppimiselle voidaan antaa seuraava määrittely [4]:
Oppiminen on algoritmin kyky parantaa sen suorituskyykyä aikaisemman kokemuksen perusteella.
- Toisin sanoen oppiminen on opetetun algoritmin kyky tehdä ennustuksia.

Koneoppiminen

- Koneoppiminen puolestaan voidaan määritellä seuraavasti [5]:
Koneoppiminen on sellaisten algoritmien ja menetelmien systemaattista tutkimista, jotka parantavat niiden tietämystä tai suoritusta kokemuksen perusteella.
- Mitä tarkoitetaan kokemuksella?
- Aikaisempi kokemus koneoppimisalgoritmien tapauksessa on olemassaolevaa (“historiallista”) dataa, joka on talletettuna esimerkiksi tietokantaan [4].

Koneoppimisalgoritmien pääjaottelu

1. Ohjattu oppiminen
2. Ohjaamaton oppiminen
3. Vahvistusoppiminen

Ohjattu oppiminen

- Ohjatussa oppimisessä datan havaintojen luokat ovat etukäteen tiedossa.
- Jokainen aineiston havainto on siis asetettu kuuluvaksi ennaltamäärättyyn luokkaan/kategoriaan.
- Jos luokkia ei ole etukäteen tiedossa, voidaan ohjatun oppimisen algoritmeja käyttää, jos datassa on klusteroituva (ryhmittyvä) rakenne [4]. Tällöin datalle ensin suoritetaan klusterointi, jonka jälkeen sovellusalan asiantuntija määrittää kullekin klusterille ja sen sisältämille havainnoille luokkaleiman. Tämän jälkeen ohjatun oppimisen algoritmeja voidaan hyödyntää.

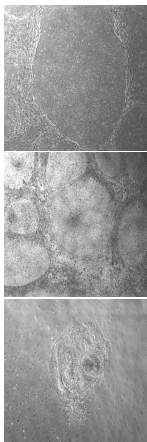
Ohjattu oppiminen

- Jokaisen ohjatun oppimisen piiriin kuuluvalla koneoppimisalgoritmille täytyy antaa opetusjoukko, jota käyttäen algoritmi muodostaa mallin ennustamista varten [2].
- Opetusjoukko koostetaan olemassaolevasta datasta.
- Opetusvaiheen jälkeen algoritmille syötetään testijoukko, joka on riippumaton opetusjoukosta.
- Testijoukon perusteella määritämme, kuinka hyvin algoritmi on onnistunut uusien havaintojen ennustamisessa eli mikä on algoritmin yleistämiskyky.
- Koneoppimismenetelmien kohdalla yleistämiskyky on keskeistä, jotta menetelmää voidaan hyödyntää sovelluksessa nyt ja tulevaisuudessa luotettavasti.

Ohjatun oppimisen menetelmiä

- K :n lähimmän naapurin menetelmä
- Naiivi Bayes -luokittelija
- Lineaarinen diskriminanttianalyysi
- Tukivektorikone
- Neuroverkot (mm. syväoppiminen)
- Päättöpuut ja satunnaismetsät

Esimerkki ohjatun oppimisen ongelmasta



- Luokkien määrä: 3
- Piirteet: Kuvista irrotettuja intensiteettihistogrammeja
- Luokkatieto: hyvä, puolihyvä, huono
- Tavoite: Ennusta oikea luokkatieto uusille kantasolukoloniakuville.
- H. Joutsijoki et al., "Histogram-based classification of iPSC colony images using machine learning methods," *Proceedings of the 2014 IEEE Conference on Systems, Man, and Cybernetics*, pp. 2611-2617, 2014.

Kuva : Kantasolukoloniakuvia kolmesta luokasta: Hyvä (ylin), puolihyvä (keskimmäinen), huono (alin).

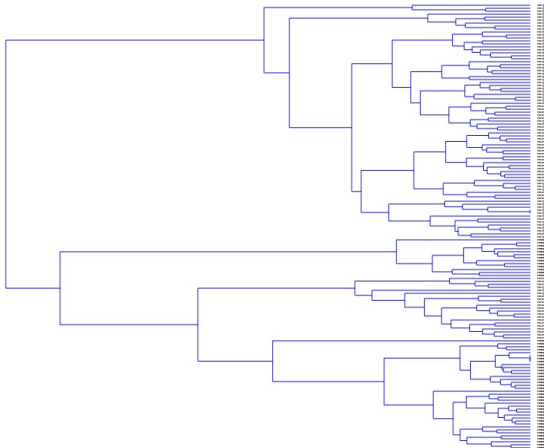
Ohjaamaton oppiminen

- Ohjaamatonta oppimista usein sanotaan klusteroinniksi, mutta myös assosiaatiosääntöjen etsiminen kuuluu ohjaamattoman oppimisen piiriin [3].
- Klusteroinnissa luokkatietoja havainnoille ei tarvita, vaan dataa voidaan analysoida ilman niitä.
- Klusteroinnin tavoitteena on löytää datasta ryhmät, joiden sisäinen samankaltaisuus on suuri ja eri ryhmien välinen samankaltaisuus on mahdollisimman pientä.

Ohjaamattoman oppimisen menetelmiä

- Itseorganisoituvat kartat
- K-Means-algoritmi
- Hierarkkinen klusterointi
- EM-algoritmi

Esimerkki

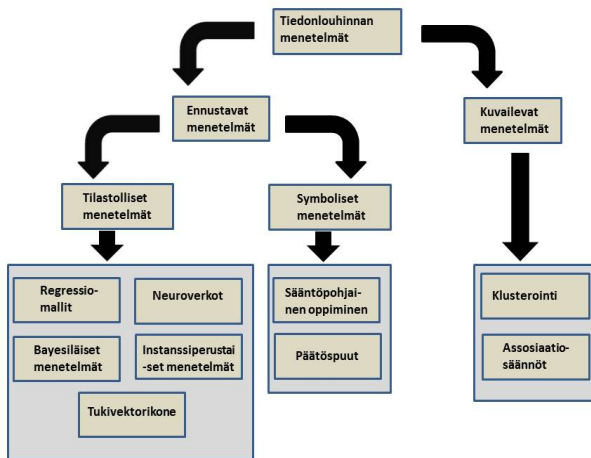


Kuva : Esimerkki hierarkkisen klusteroinnin lopputuloksena syntyvästä dendrogrammista.

Vahvistusoppiminen

- Vahvistusoppiminen sijoittuu ohjatun ja ohjaamattoman oppimisen välimaastoon.
- Vahvistusoppimisessa käyttäjä antaa palautetta (positiivista/negatiivista) algoritmille ja saadun palautteen perusteella algoritmi muokkaa nykyistä ratkaisua ja pyrkii löytämään paremman ratkaisun.
- Q-oppimisalgoritmi on kuuluisin vahvistusoppimiseen kuuluva menetelmä.
- Vahvistusoppimista voidaan hyödyntää esimerkiksi peliohjelmoinnissa.

Tiedonlouhintamenetelmien jaottelu



Kuva : Hierarkkinen esitys tiedonlouhintamenetelmistä. Kuva mukailee lähdettä [1].

Sisältö

Johdanto

Tiedonlouhinta

Koneoppiminen ja tiedonlouhinta

Tiedonlouhinnan tulevaisuus

Tulevaisuus

- Tiedonlouhinnan, koneoppimisen, data-analyysin tai yleisesti ottaen tekoälyn tulevaisuus näyttää varsin valoisalta, sillä digitaalisessa muodossa olevan tiedon määrä kasvaa erittäin nopeasti.
- Yhä useampi sovellusala tulee hyödyntämään oppivia algoritmeja, joiden avulla mekaaniset työtehtävät pystytään automatisoimaan ja monissa paikoin myös turvallisuutta pystytään lisäämään.
- Koneoppimisalgoritmeja tullaan hyödyntämään jatkossa enenevässä määrin eri laitteissa ja kulkuneuvoissa (autot, puhelimet, televisiot,...).

Tulevaisuus

- Suurten tietomassojen tiedonlouhinnan myötä pystytään asiakasryhmiä kohdentamaan paremmin sekä tarjoamaan kuluttajille tarkemmin kohdennettuja palveluja, informaatiota ja mainontaa.
- Tiedonlouhintaan, koneoppimiseen ja tekoälyyn liittyy paljon myös uhkakuvia, joista osa juontaa juurensa elokuvista ja tieteiskirjallisuudesta. Tulevatko menetelmät “liian älykkäiksi”? Rikkovatko älykkäät algoritmit yksityisyyden rajat? Voimmeko luottaa älykkäisiin menetelmiin tarpeeksi? Kenellä on vastuu, jos esimerkiksi koneoppimisalgoritmeja hyödyntävä itseohjautuva auto ajaa kolarin?

Kiitos!

Kysymyksiä?



S. Garcia et al., *Data Preprocessing in Data Mining*, Springer-Verlag, 2015.



J. Han et al., *Data Mining: Concepts and Techniques*, 3rd edition, Morgan Kaufmann, (Elsevier), 2012.



K.J. Cios et al., *Data Mining: A Knowledge Discovery Approach*, Springer, 2007.



M. Juhola, *Tiedonlouhinta kurssin luentomateriaali*, Tampereen yliopisto, 2014.



T.M. Mitchell. *Machine Learning*. Mcgraw-Hill, New York, USA, 1997.