

TILASTOTIEDE

TILASTO NÄKEE TULEVAN

Tilastot ovat parhaimmillaan taidetta, joiden avulla luonnon ja yhteiskunnan mutkikkaat ilmiöt aukeavat kuin itsestään. Pahimmillaan ne vääristävät tietoa ja johtavat harhaan – tahattomasti tai tahallaan.

Teksti Lauri Nummenmaa Kuvitus Riku Koskelo

Mikä yhtye on myynyt eniten levyjä maailmassa? (The Beatles, 183 miljoonaa levyä). Entäpä kuka pelaaja on tehnyt urallaan eniten maaleja NHL:ssä? (**Wayne Gretzky**, 894 kappaletta). Tai koska Suomen valtion velka on ollut alimmillaan? (vuonna 1974, kolme prosenttia bruttokansantuotteesta).

Ihmiset rakastavat tietoa ja sen hallitsemista. Jo tuhansia vuosia sitten tallensimme savesta tehtyihin tauluihin lukuja kaupankäynnistä, sodasta ja verotuksesta. Nykypäivänä käytämme samaan tehtävään tietokoneen laskenta- tai tietokantataulukoita. Numerotiedon järjestelmällinen tallentaminen on tärkeää, koska se ker-

too meille, kuinka asiat ovat aikaisemmin olleet. Lisäksi numeroiden avulla voi tehdä laskutoimituksia ja esittää tietoa tarkasti.

Jos sanomme yhden lapsen olevan ”lyhyt” ja toisen ”pitkä”, ymmärtävät kaikki ihmiset lasten pituuden hieman eri tavalla. Emme myöskään voi sanoa kuinka paljon toinen lapsi on toista pidempi. Jos sen sijaan kerromme toisen lapsen olevan 50 senttimetriä pitkä ja toisen 60, hahmottavat kaikki lasten pituudet samalla tavalla. Ymmärrämme, että kyse on vastasyntyneistä, eikä esimerkiksi koululaisista. Lisäksi voimme laskea, että toinen lapsi on 10 senttiä toista pidempi.

Mittaamalla lasten pituudet neuvolakäyntien yhteydessä ja vertaamalla niitä lasten tyypillisistä pituuksista muodostettuihin historiallisiin kasvukäyriin lääkäri voi arvioida, onko lap-

sen pituuskasvu normaalilla uralla vai pitäisikö mahdollisia kasvua hidastavia tekijöitä alkaa selvittää. [Toisin sanoen](#) tietojen järjestelmällinen kerääminen, siis tilastointi, voi auttaa meitä näkemään tulevaan.

Näkömätön muuttuu näkyväksi

Ihmisen aistit hahmottavat vain pienen osan luonnossa olevasta tiedosta ja pystyvät tallentamaan vain murto-osan siitä muistiin. Aivot ovat erikoistuneet keräämään ympäristöstä käden ulottuvilla olevaa tietoa, koska se on auttanut lajiamme selviytymään eloonjäämiskamppailussa.

Arkihavainnointi on kuitenkin niin tilannesidonnaista ja pienimuotoista, ettei se riitä ~~millään~~ suurten tietomasojen hahmottamiseen ja ymmärtämiseen. Niinpä suuri osa meille tärkeistä luonnonilmiöistä jää aivoiltamme



Numeroiden avulla
tietoa on mahdollista
esittää tarkasti.

havaitsematta – emme esimerkiksi huomaa aistinvaraisesti raskasmetallien tuhoavan keskushermostoa, mutta tilastojen perusteella ilmiö on päivänselvä. Tilastojen avulla luonnon ja yhteiskunnan monimutkaisuus voidaan ottaa haltuun.

Yksi esimerkki varhaisten tilastojen voimasta löytyy 1800-luvulta. Vuosisadan puolivälissä kolera oli tappava tauti, jonka tartuntamekanismia ei tunnettu. Jotkut tutkijat ajattelivat sen leviävän vaarallisten höyryjen, ”miasmojen” välityksellä. Osa tutkijoista taas ajatteli, että tauti leviää kosketuksen välityksellä.

Lääkäri **John Snow** alkoi merkitä Lontoon Sohossa koleratartunnan saaneiden asuinpaikkoja karttaan. Hän huomasi, että tartunnat eivät jakautuneet satunnaisesti, vaan ne keskittyivät erään Broad Streetillä sijainneen kaivon ympäristöön. Kaivon vesi tuli Thamesin saastuneelta alajuoksulta. Snow huomasi, että puhtaammalta yläjuoksulta vetensä saavilla kaupungin alueilla koleratartunnat olivat yli kymmenen kertaa harvinaisempia. Snow sai taivuteltua kaupungin virkamiehet sulkemaan

Satunnaisuudesta huolimatta ilmiöt eivät ole täysin kaoottisia, vaan niitä voidaan ennustaa tietyissä rajoissa.

saastuneen kaivon, ja tartunnat vähenivät alueella välittömästi.

Snowta pidetään nykyään modernin epidemiologian isänä – hänen karttansa paljasti selvän säännönmukaisuuden likaisen veden ja koleratartuntojen välillä. Kartan avulla Snow teki silmälle näkymättömät bakteerit näkyviksi.

Nykypäivänä aistiemme apuna on monenlaisia mittalaitteita, joilla voimme havaita ihmisaistien tavoittamattomissa olevia ilmiöitä. Ja mikä tärkeämpää, meillä on käytössä tehokkaita laskennallisia menetelmiä, joiden avulla laitteistojen tuottaman signaalivirran joukosta voidaan poimia merkityksellisiä muutoksia.

Auttaako psykoterapia masennuksen hoidossa? Onko sähköallergia pelkkää mielikuvituksen tuotetta?

Periytyykö älykkyys? Vastaus kaikkiin kysymyksiin on ”kyllä”, ja tällainen tieteellinen tieto perustuu tilastolliseen tutkimukseen.

Mistä sattuma syntyy?

Fysiikan ja matematiikan lait ovat universaaleja. Pythagoraan lause pätee niin Maan pinnalla kuin Marsissa, ja aineen häviämättömyyden laki on voimassa niin Kouvolassa kuin Kuala Lumpurissa.

Kaikkia luonnon tapahtumia on ~~silti~~ mahdotonta kuvata tarkkojen matemaattisten ja fysikaalisten lakien avulla, koska sattumalla on suuri vaikutus siihen, miten luonto toimii. Emme esimerkiksi tiedä varmasti, kuinka lämmintä ensi vuoden juhannusaattona on – säähän voi kesäkuussa olla joko helteinen tai hyisen kylmä.

Satunnaisuudesta huolimatta ilmiöt eivät ole täysin kaoottisia, vaan niitä voidaan ennustaa tietyissä rajoissa. Vaikkei juhannussää ole tarkkaan tiedossa etukäteen, voimme olla melko varmoja siitä, että kesällä lämpötila pysyttelee kymmenen asteen paremmalla puolella.

Samanlainen satunnaisuus liittyi myös Broad Streetin koleratapauksiin. Kaikki alueen kaivoista juoneet eivät sairastuneet, eivätkä kaikki tapaukset keskittyneet samalle alueelle. Tilastotieteen avulla satunnainen ja systemaattinen vaihtelu oli kuitenkin mahdollista erottaa toisistaan: John Snown kartta osoitti, että tapaukset olivat todennäköisempiä Broad Streetillä kuin muualla Sohossa. Tarkan lopputuloksen sijaan kyse oli todennäköisimmän lopputuloksen arvioimisesta.

Tilastotieteen avulla samanlaisia laskelmia voidaan tehdä lääkkeiden tehosta, eläinten käyttäytymisestä tai vaikkapa ihmisten ruokavalion muutosten vaikutuksesta lihomiseen. Tilastotiede onkin ennen kaikkea riskinhallintatiede.

Todennäköisesti mahdollista

Todennäköisyyksiin liittyvä matemaattinen teoria alkoi muotoutua 1700-luvun aikana. Sysäyksen sille antoivat uhkapeleihin liittyvien todennäköisyyksien hallinta ja vakuutus-

Kartta paljasti piilottelevat bakteerit

John Snow merkitsi koleratartunnan saaneiden asuinpaikkoja karttaan. Hän huomasi, että tartunnat keskittyivät saastuneen vesikaivon ympäristöön.



● Koleratapaus ● Vesikaivo



Lauri Markkasen vapaaheittojen onnistumisen todennäköisyydeksi on laskettu 85 prosenttia.

yhtiöiden tarve arvioida vakuutettujen riskien toteutumista ja vakuutus-ten hinnoittelua. Varsinainen tilastollinen laskenta pysyi kuitenkin pitkään sivuroolissa luonnontieteissä.

Vasta 1900-luvun alussa biologian ja psykologian tieteidenalojen kehittyminen alkoi edellyttää menetelmiä, joilla mittaustulosten näennäisestä vaihtelusta ihmisten ja organismien välillä pystyttiin paikallistamaan säännönmukaisuuksia. Tilastollisen analyysin jättiläiset **Sir Francis Galton**, **Karl Pearson** ja **Sir Ronald Fisher** loivat perustan matemaattiselle tilastotieteelle, jossa tilastotietojen kerääminen ja todennäköisyyslaskenta valjastettiin luonnon, talouden ja yhteiskunnan ilmiöiden mallintamiseksi ja ennustamiseksi. Lopputuloksena oli

Jos kolikkoa heittäisi äärettömästi, tuloksista puolet olisi kruunია, puolet klaavoja. Sattuma tasoittuu pitkässä juoksussa.

supertieteenala, joka tunnetaan nyky-päivänä tilastotieteenä.

Mistä lopulta puhumme, kun puhumme todennäköisyyksistä? Todennäköisyyden tarkka määrittäminen on yllättävän monimutkaista, vaikka ihmisaivot ovat lopulta eräänlainen intuitiivinen todennäköisyyslaskin. Niiden tehtävänä on jatkuvasti ennustaa geeneihin ja muistiin tallennetun tiedon perusteella tulevien tapahtu-

mien todennäköisyyttä – onko pomo tänään niin hyvällä tuulella, että voisi pyytää palkankorotusta, nouseeko tuuli vielä niin paljon, että veneilemään ei kannata lähteä?

Todennäköisyyksiä voidaan hahmottaa myös muodollisesti pitkän aikavälin esiintymisen avulla. Yksinkertaisimmillaan voidaan tutkia jonkin satunnaisia lopputuloksia tuottavan mekanismin, kuten nopan tai kolikon heiton tuloksia. Kaikkitietävälle olen- nolle vastaus olisi aina etukäteen selvä, mutta ihmisille lopputulos selviää vasta, kun prosessi on tapahtunut.

Jos esimerkiksi kolikkoa heitettäisiin äärettömän monta kertaa, puolet lopputuloksista on kruunია ja puolet klaavoja. Vaikka kolme peräkkäistä heittoa voikin olla sattumalta olla klaavoja, pitkässä juoksussa sattuman vaikutus tasoittuu.

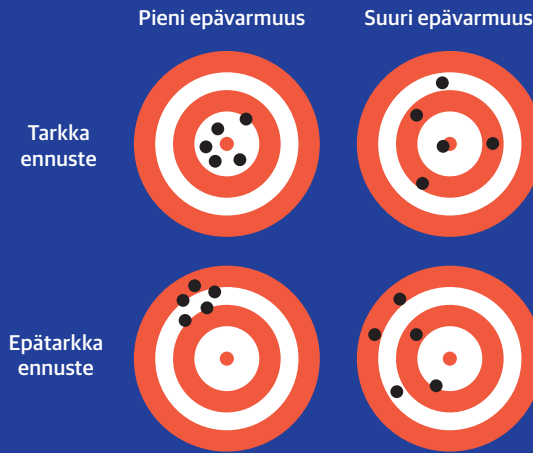
Oikeansuuntainen riittää

Kolikon heittoon liittyvä todennäköisyys on kolikon geometrian ominaisuus, koska muotonsa vuoksi kolikko tippuu puolet kerroista kruunan ja puolet kerroista klaavan puolelle. Mutta monien muidenkin tapahtumien todennäköisyys voidaan selvittää keräämällä ja tarkastelemalla pitkän aikavälin tilastoja.

Esimerkiksi **Lauri Markkasen** vapaaheittojen onnistumistodennäköisyys on 85 prosenttia, eli sadasta heitosta Markkanen upottaa 85 ja 15 menee korin ohi. Todennäköisyys on saatu yksinkertaisesti laskemalla kaikki Markkasen uransa aikana heittämät koriin ja sen ohi menneet vapaahetot.

Samaan tapaan lääketutkimuksessa ”toistoja” voidaan tehdä antamalla testattavaa lääkettä tuhannelle ihmiselle, jolloin jokainen ihminen on uusi toisto lääkkeen tehosta. Lopuksi voidaan laskea, kuinka monella ihmisellä lääke helpotti oireita verrattuna lumelääkkeeseen. Tulosten perusteella voidaan tämän jälkeen vaikkapa todeta, että lääke vaikutti 700 potilaalla. Tällöin voidaan päätellä, että keskimäärin seitsemän kymmenestä potilaasta saa apua lääkkeestä, eli todennäköisyys lääkkeen toivotulle vaikutukselle on 7/10.

Epätarkkuus ei ole aina ongelma



Kuinka monta potilasta lääketutkimukseen pitää ottaa, että lääkkeen teho voidaan arvioida luotettavasti? Intuitiomme sanoo, että suurempi määrä potilaita antaa tarkemman arvion. Käytännössä tulosten tarkkuuteen vaikuttaa kuitenkin tutkittujen potilaiden lisäksi monta tekijää, kuten vaikkapa se, kuinka järjestelmällisesti lääke eri potilaisiin vaikuttaa.

Tilastotieteen teorian avulla tällaiset tekijät voidaan huomioida ja laskea, kuinka suuri epävarmuus tutkimusten tuloksiin liittyy.

Intuition vastaisesti tutkimuksissa saatavista ennusteista on hyötyä myös silloin, kun ne eivät ole täysin tarkkoja – riittää, että ne ovat oikeansuuntaisia. Esimerkiksi statiineista on keskimäärin hyötyä kolesterolin hoidossa vaikka ne eivät auttaisikaan joka ikistä potilasta. Pörssimeklarille on hyödyllistä pystyä arvioimaan edes suunnitteen tulevien viikkojen kurssikehitys, vaikkei hän pystyisikään arvioimaan täsmällisesti tulevia pörssikursseja.

Pienetkin oikeansuuntaiset tiedonmuruset ovat aina arvokkaampia kuin epätietoisuudessa eläminen.

Maalaisjärkeä numeroilla

Entä jos haluamme arvioida todennäköisyyden sellaiselle tapahtumalle,

Hyvä ennuste on tarkka ja siihen liittyy vähän epävarmuutta. Myös epätarkka ennuste on hyödyllinen, jos se on oikeansuuntainen.

jota emme voi toistaa? Tällöin voimme perustaa todennäköisyysarvion ihmisten arvioihin. Esimerkiksi uutta Marsiin lähetettävää luotainta ei laukaista kuin yhden ainoan kerran, joten arviota Mars-lennon onnistumisesta ei voida määrittää toistojen avulla.

Tällöin voimme saada arvion todennäköisyydestä vaikkapa siten, että pyydämme kaikkia luotaimen suunnitteluun osallistuneita insinöörejä lyömään vetoa lennon onnistumisen puolesta tai sitä vastaan. Jos arvaus on oikein, vedonlyöjä saa 100 euroa ja jos ehdokas häviää, hän menettää rahansa. Lopuksi lasketaan, kuinka paljon puolesta ja vastaan -ääniä saatiin.

Tällä tavoin todennäköisyys ei ole kolikon geometrian tapaan puhtaasti objektiivinen asia, vaan subjektiivinen ominaisuus, joka kuvaa tapahtuman lopputulokseen liittyvää epävarmuutta yksittäisillä ihmisillä. Mutta lopulta sama subjektiivisuus koskee myös

rahan heittoa: Jos heitän rahaa ja peitän sen saman tien kädelläni, tiedän rahan heiton lopputuloksen varmasti. Mutta jos kysyn sinulta, mikä lopputulos on, joudut arvaamaan tai sanomaan että 50 prosentin todennäköisyydellä kruuna. Tällainenkin todennäköisyys on lopulta subjektiivinen, koska se riippuu siitä, mitä tietoa kulloinkin on käytettävissä.

Ranskalainen matemaatikko **Pierre-Simon de Laplace** kutsuikin todennäköisyyksiä ”numeroihin perustuvaksi maalaisjärjeksi”. Tämä kuvaa hyvin sitä, miten todennäköisyyksiä käytetään päätöksenteossa: lopulta kaikki tilastollinen tieto kulkee ihmisaivojen päätöksentekomekanismien läpi ja auttaa meitä epävarmuuden hallitsemisessa.

Esimerkiksi [Nasa](#) arvioi laskelmisaan Apollo 11 -ohjelman onnistumistodennäköisyydeksi 73 prosenttia ja miehistön selviämisen todennäköisyydeksi 95 prosenttia. Astronautti **Neil Armstrong** puolestaan arvioi vastavien todennäköisyyksien olevan 50 prosenttia ja 90 prosenttia, minkä perusteella hän uskalsi nousta Saturn V -raketin matkassa Kuun kiertoradalle ja laskeutua Kotka-moduulin ovesta ensimmäisenä ihmisenä vieraan taivaankappaleen pinnalle.

Tieto on valtaa

Teollinen vallankumous vapautti ihmisen fyysisestä työstä, mutta henkinen ponnistelu kulutti vielä pitkään mielen lihaksia. Esimerkiksi numerotiedon järjestely vaatii ~~valtavien tietomäärien käsittelyä~~. Nyt myös henkinen työtaakka on keventynyt tietokoneiden ja tekoälyn avulla.

Tekoäly on ajatuksena ikivanha: jo 1900-luvun alkupuolella ensimmäisten tietokoneiden kehittäminen sai tutkijat pohtimaan, voisiko inhimillisen ajattelun ”mekanisoida”. Tavoitteena on, että koneet pystyvät tekemään tylsiä rutiinitöitä ilman ihmisten työpanosta ja tunnistamaan tietomassasta säännönmukaisuuksia, joita ihminen ei pysty havaitsemaan.

Vielä 1950-luvulla tutkijat työskentelivät mekaanisten laskinten ja laskutikkujen kanssa. Taskulaskimet yleis-

ALGORITMI OPPII ILMAN IHMISTÄ

Tekoäly tunnistaa aineistosta säännönmukaisuuksia.

Tietokoneohjelmat perustuvat algoritmeihin eli automatisoituihin ohjeisiin. Tavallisesti ihminen antaa algoritmilte etukäteen

säännöt siitä, kuinka kussakin tilanteessa pitää toimia. Koneoppiminen on tekoälyn uudempi alalaji. Siinä algoritmi parantaa itse omaa suorituskyykyään. Se oppii aineistosta omatoimisesti ilman ihmisen antamia uusia toimintaohjeita.

Nykypäivänä koneoppimista on kaikkialla, puheentunnistuksesta konenäköön, osakemarkkinoiden analysoimiseen ja itsestään ajaviin autoihin. Sen avulla voidaan ennustaa, missä päin kaupunkia tapahtuu todennäköisimmin rikoksia tai arvioida, kuinka voimakasta kipua ihminen kokee.

Kone on sokea aineistolle

Koneoppimisalgoritmi pitää optimoida siten, että se toimii mahdollisimman tehokkaasti ja tarkasti. Tämä voidaan tehdä kahdella tavalla. Ohjatussa oppimisessa algoritmilte opetetaan, millaisia yksiköitä kuhunkin luokkaan, esimerkiksi omenoihin, kuuluu. Algoritmilte annetaan omenan kuvia esittävä aineisto, josta se opettelee kyseisiin kuviin liittyviä säännönmukaisuuksia. Tämän perusteella toimiva algoritmi pystyy tunnistamaan myös uudet omenan kuvat, joita ei ole ollut opetusaineistossa.

Lähestymistapa toimii silloin, kun tiedetään etukäteen, millaista tehtävää algoritmin halutaan suorittavan. Usein algoritmia käyttävä ihminen ei kuitenkaan tarkasti tiedä, mitä hän haluaa algoritmin oppivan, tai mikä olisi paras tapa järjestää aineisto. Ohjaamattomassa oppimisessä algoritmi etsii tietovirrasta säännönmukaisuuksia ilman, että sille kerrotaan, mitä sen tulisi etsiä – esimerkiksi kuvia erilaisista hedelmistä, joita se yrittää itse jaotella luokkiin

niiden samankaltaisuuden perusteella. Tilannetta voi verrata vaikkapa uuden videopelin opettelemiseen. Harvassa pelissä tulee mukana yksityiskohtaisia ohjeita, koska niiden kirjoittaminen olisi todella työlästä. Niinkin yksinkertaisen pelin kuin Pac-Manin ohjeet olisivat tekstiksi kirjoitettuna monta sivua, uudempien ja monimutkaisempien pelien ohjeista puhumattakaan. Siksi ohjelmoijat luottavat siihen, että ihmiset oppivat pelimekaniikan havainnollisella oman pelaamisensa ja pelitapah- tumien välisiä lainalaisuuksia.

Koneoppiminen on tieteenkin lähtökohtaisesti täysin sokeaa sille, mitä aineisto on. Jos vaikkapa lääkärit koettavat tunnistaa aivojen magneettikuvien perusteella mahdolliset aivoverenvuotopotilaat terveistä verrokeista, voi algoritmi näyttää aluksi toimivan täydellisesti. Mutta koska aivoverenvuoto-

potilaat ovat tyypillisesti terveitä potilaita vanhempia, voi algoritmi perustaa "arvauksensa" pelkästään aivojen yleiseen iän mukana rapistuvaan kuntoon eikä niinkään siihen, onko kuvassa todellisuudessa havaittavaa aivoverenvuotoa. Tämän vuoksi algoritmien toimintaa on jatkuvasti seurattava ja niiden toiminta on varmistettava monenlaisissa ympäristöissä.

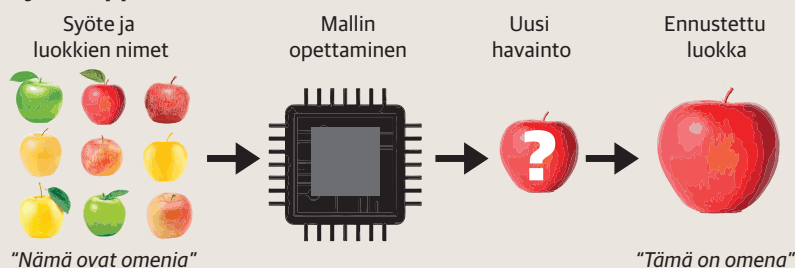
Uusi kielimalli mullistaa

Viime vuosien suurin läpimurto tekoälytutkimuksessa on Chat GPT:n kaltaisten suurten kielimallien kehittyminen. Syväoppimiseen perustuvat mallit on opetettu valtavalla määrällä luonnollista kieltä, josta ne oppivat sanojen, käsitteiden ja lauseiden välisiä suhteita.

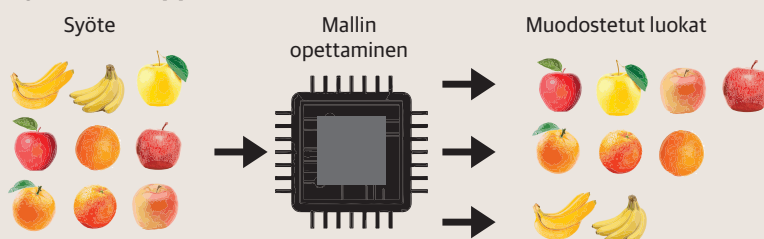
Malleja voi käyttää esimerkiksi kielten kääntämiseen englannista ohjelmointikieleksi, tai luonnollisen kielen, kuten lehtiartikkelien tai proosatekstien tuottamiseen ja tiedon hakemiseen. Tällainen tekoäly mullistanees asiantuntijatyön tulevaisuudessa, koska mallit pystyvät teoriassa hallitsemaan kaiken digitaalisessa muodossa olevan tekstin ja luomaan sen perusteella uutta tietoa.

Kaksi tapaa kouluttaa

Ohjattu oppiminen



Ohjaamaton oppiminen



tyivät toimistoissa 1970-luvulla, ja mikrotietokoneet valtasivat yliopistot ja tutkimuslaitokset 1980-luvun alussa. Henkilökohtaiset tietokoneet olivat kuitenkin hitaita vielä pitkälle 1990-luvulle asti, ja yksinkertaistenkin tilastollisten laskutoimitusten tekeminen saattoi kestää vielä tunteja tai päiviä. Siksi tutkijoiden piti valikoida tarkkaan, millaisia laskelmia he halusivat aineistoistaan tehdä.

Tällä vuosituhannella mikroprosessorien nopeus on viimeinkin poistanut suuren osan laskentatehon rajoituksista. Rajat tilastotieteen, matematiikan ja tietojenkäsittelytieteiden välillä ovatkin pitkälti hävinneet, ja usein puhutaan yleisesti joko laskennallisista tieteistä tai datatieteestä.

Supernopeat tietokoneet ja oppivat algoritmit auttavat louhimaan informaatiota ison tietomassan seasta. Vaikka yksittäiset tiedonmuruset vaikuttaisivat aluksi mitättömän heikoilta signaaleilta, voi niistä yhdistellä tilastotieteen keinoin ällistytävään tarkkoja ennusteita. Esimerkiksi Ukrainan sodan aikana venäläisten matkapuhelinliittymien sijaintitiedon avulla pystyttiin tietoa Venäjän armeijan liikkeistä, koska sotilaat

Ei ole sattumaa, että verkkokaupat tarjoavat sinulle tiettyjä tuotteita: huomaamaton tiedon louhinta on tehokasta.

eivät ole malttaneet pitää puhelimiään pois päältä edes taistelukentällä.

Sosiaalisen median yhtiöt puolestaan keräävät suuret määrät tietoa ihmisten mielenkiinnon kohteista heidän surffailunsa ja klikkailunsa avulla, ja samalla tavalla kaupan ala kerää tietoa ostokäyttäytymisestä bonuskorttien ostotapahtumia seuraamalla. Ei ole sattumaa, että verkkokaupat tarjoavat sinulle tiettyjä tuotteita: asiakkaiden valtava tietovirta sisältää heikkoja mutta riittävän voimakkaita signaaleita siitä, millaisia ostoksia eri ihmiset verkkokaupassa tekevät.

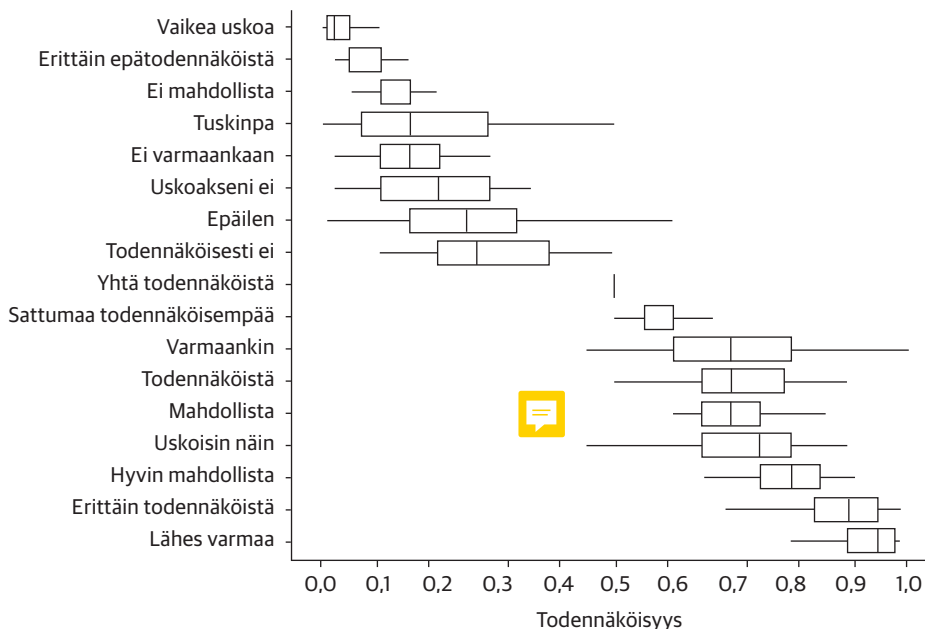
Huomaamaton tiedon louhinta voi olla ällistytävään tehokasta. Esimerkiksi ihmisten persoonallisuuden arvioimiseen käytetään perinteisesti pitkiä kyselylomakkeita, jotka vastajien pitää täyttää joko netissä tai paperilla. Tutkimuksessa on kuitenkin osoitettu, että koneoppimisalgorit-

mien avulla ihmisten persoonallisuus voidaan ennustaa tarkasti pelkästään heidän somekäyttäytymisensä perusteella. Ei siis ole yhdentekevää, mitä tietoa itsestään sosiaaliseen mediaan laittaa – algoritmi voi paljastaa ihmisestä paljon henkilökohtaisempaa tietoa kuin kuvittelemmekaan.

Toisaalta samanlaisia algoritmeja voi käyttää myös lääketieteessä ennustamaan sairastumista erilaisista heikoista signaaleista. Belgialaiset tutkijat ovat esimerkiksi huomanneet, että masennuksen puhkeaminen voidaan ennustaa seuraamalla ihmisen mielialaa mobiililaitteilla jo paljon ennen hoitoon hakeutumista. Tällaisten kevyiden biomarkkereiden paikallistaminen ihmisten itse tuottamasta datasta on arvokasta erityisesti vaativaa ja monimutkaista diagnostiikkaa vaativien sairauksien, kuten Alzheimerin taudin kohdalla. Mitä aikaisemmin potilas ohjataan tutkimuksiin ja hoitoon, sitä parempi hoitovaste tutkimuksissa voidaan saada.

Liian tarkka mikroskooppi

Kun mittalaitteet tarkentuvat, saatamme nähdä myös tarpeetonta tai haitallista tietoa. Uusien kuvantamislaitteiden avulla radiologi voi nähdä esimerkiksi ihmisen päässä monia poik-



Kuinka todennäköinen on tapahtuma, jonka toteutuminen on "mahdollista"?

Ihmiset ovat hyvin yksimielisiä siitä, mitä erilaiset todennäköisyyttä kuvaavat sanonnat tarkoittavat. Tutkimuksessa koehenkilöitä pyydettiin liittämään sanontoihin numeerisia todennäköisyyksiä (asteikolla 0–1). **Puolet vastauksista osui laatikoiden alueelle, 90 prosenttia viivojen väliin.**



keavia, joskin harmittomia muutoksia. Samoin tilastollisessa tutkimuksessa voidaan suurissa aineistoissa löytää sellaisiakin yhteyksiä, joilla ei ole käytännön kannalta merkitystä.

Eräässä tutkimuksessa selvitettiin älylaitteiden ja sosiaalisen median käytön vaikutusta nuorten hyvinvointiin. Satojentuhansien vastaajien aineistossa havaittiin negatiivinen yhteys: paljon sosiaalista mediaa käyttävät nuoret voivat keskimäärin huonommin kuin somea vähän käyttävät.

Efektin suuruus oli kuitenkin häviävän pieni. Digitaalitekniikan käyttö selitti vain 0,4 prosenttia hyvinvoinnin vaihtelusta. Tutkijat havainnollistivat vaikutuksen suuruutta vertailemalla digitaalitekniikan käytön vaikutuksia hyvinvointiin täysin mielettömien

Ihmisten toiminnan ennustaminen on monimutkaista, koska erilaisia tuntemattomia tekijöitä on niin valtavasti.

selittäjien vaikutuksiin. He havaitsivat, että esimerkiksi perunoiden syöminen tai silmälasien käyttäminen ennusti hyvinvointia aivan yhtä hyvin kuin digitaalitekniikan käyttäminen. Vertailun vuoksi säännöllinen liikunta saattaa vähentää masennukseen sairastumisen riskiä jopa 30 prosenttia!

Tilastollisia tutkimuksia lukiessa pitääkin olla aina tarkkana sen suhteen, kuinka suuria vaikutuksia esi-

merkiksi sosiaalipoliittisilla päätöksillä on lasten terveyteen tai kuinka paljon uusi lääke vähentää potilaiden kipua.

Laskentatehon ja tekoälyn kehittyminen ei ole vielä syrjäyttänyt ihmistä päätöksentekijänä – tutkijan, insinöörin tai poliitikon täytyy edelleen tulkitella tilastolliset tulokset sekä arvioida, millaisia päätöksiä niiden perusteella kannattaa tehdä.

Epävarmuutta ei voi poistaa

Ihmismieli yrittää jatkuvasti nähdä tulevaan, oli sitten kyse meille heitetyn pallon lentoradan arvioimisesta tai taloudellisesta varautumisesta eläkepäivien viettoon.

Omaan kokemukseen perustuvat ennusteet ovat lähes aina epätarkkoja ja rajallisia, koska emme osaa yhdistellä suuria tietomääriä tai painottaa tärkeää ja vähemmän tärkeää tietoa päätöksenteossa. Tilastollisten menetelmien avulla tällaisten ennusteiden tekeminen onnistuu, koska tietokoneet seulovat ennusteiden laatimisen kannalta oleellisen tiedon epäoleellaisuuksien joukosta.

On hämmästyttävää ajatella, että kuuraketin lentoradan laskeminen onnistui jo 1960-luvulla tietokoneella, jonka tehot jäivät merkittävästi jälkeen nykyaikaisista taskulaskimista. Taivaankappaleiden liikkeet sekä raketin lento avaruudessa ovat kuitenkin säännönmukaisia.

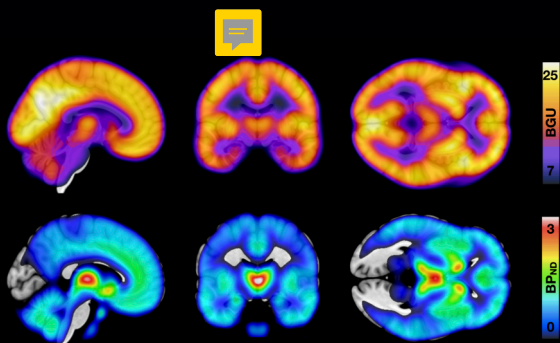
Ihmisten toiminnan ennustaminen on paljon monimutkaisempaa, koska erilaisia tuntemattomia tekijöitä on paljon enemmän. Moni on esimerkiksi yrittänyt ennustaa pörssimarkkinoiden kehittymistä epäonnistuen surkeasti, ja vain ani harva suuria palkkioita nostavista salkunhoitajista onnistuu saamaan markkinoilta niiden keskimääräistä tuottoa enempää.

Vaikka ihmiset ovat vuosi vuodelta pystyneet ottamaan paremman ja paremman niskalenkin luonnon satunnaisuudesta, ei kaikkea epävarmuutta kenties voida koskaan poistaa. Ehkä tämä lopulta on myös hyvä asia – elämähän olisi tuskin elämisen arvoista, jos tietäisimme tarkalleen mitä jokainen hetki tuo meille tullessaan.

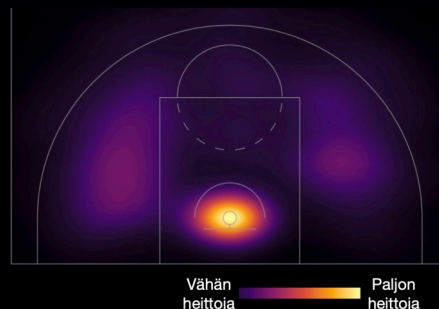
Aukeaa kuin itsestään

Tilastollinen kartta on parhaimmillaan taidetta, joka paljastaa kuvatun ilmiön ensisilmäyksellä.

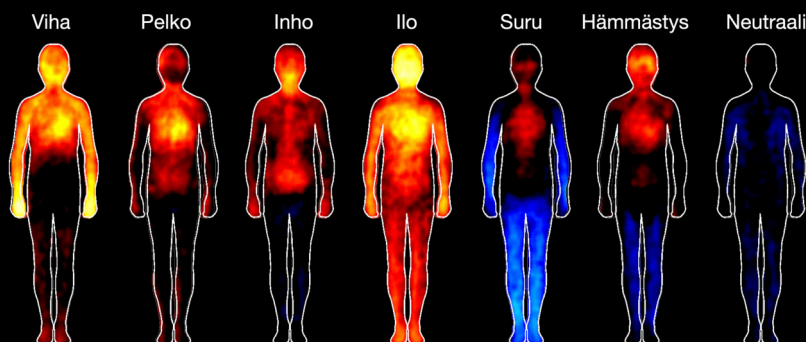
Välittäjäaineiden sijoittuminen aivoihin



Michael Jordanin heittopaikat 1997-1998



Tunteiden vaikutus kehossa



KUVAAJA HELPOTTAA JA HARHAUTTAA

Pylväät ja piirakat voivat vääristää tietoa, myös tarkoituksella.

Monesti kuulee sanottavan, että mikään ei valehtele niin paljon kuin tilasto. "Vale, emävale, tilasto", kuuluu myös pitkään toisteltu sananlasku. Todellisuudessa huolellisesti laaditut tilastot eivät valehtele lainkaan, päinvastoin. Ne ovat luotettavan päätöksenteon kulmakivi niin tekniikassa, lääketieteessä kuin politiikassa.

Tavalliset kansalaiset näkevät tilastoista kuitenkin usein vain tiivistelmän, tyypillisesti kuvaajan.

Ihminen on visuaalinen eläin: havain-

noimme ympäristöä pitkälti näköaistin varassa. Kuva kertoo enemmän kuin tuhat sanaa, ja tämän tietävät myös tilastotieteilijät. Suurten numerotaulukoiden tai abstraktien kaavojen hahmottaminen on jopa asiantuntijoille vaikeaa, joten usein tutkimustulokset pyritään esittämään visuaalisessa muodossa.

Parhaimmillaan kuvaajat ovat kuin taidetta, josta monimutkaiset luonnonilmiöt ja suuret tietomäärät hahmottuvat kuin itsestään. Pahimmillaan ne vääristävät tietoa ja antavat ilmiöstä harhaanjohtavan kuvan – tahattomasti tai tahallaan.

Muutamalla perusohjeella epäselvä tai harhauttava kuvaaja on kuitenkin helppo tunnistaa.

1. Mitä kuvaajassa on esitetty?

Hyvästä tilastollisesta kuvaajasta pitää aina löytyä tieto siitä, mitä kuvassa on esitetty. Tarkista kuvaajan pysty- ja vaakakselien selitteet: kuvataanko niissä esimerkiksi kokonaismäärää vai prosentuaalista osuutta? Mieti, mitä kuvaajassa on tärkeintä: halutaanko sillä havainnollistaa mittaustulosten keskinäisiä suuruuksia vai kenties tulosten muuttumista ajan myötä?

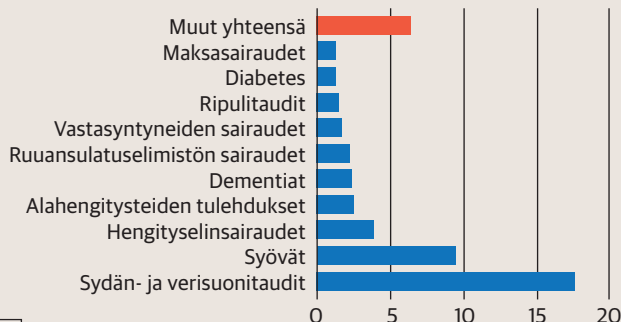
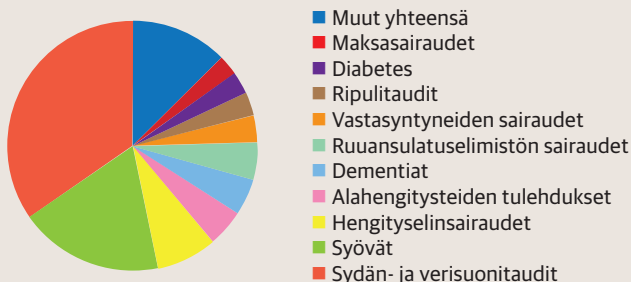
2. Miten tieto on esitetty?

Sama tieto voidaan esittää lukemattomilla eri tavoilla, joista osa on parempia ja osa huonompia. Esimerkiksi seuraavassa kuvassa on esitelty yleisimpien

Monta tapaa tulkita

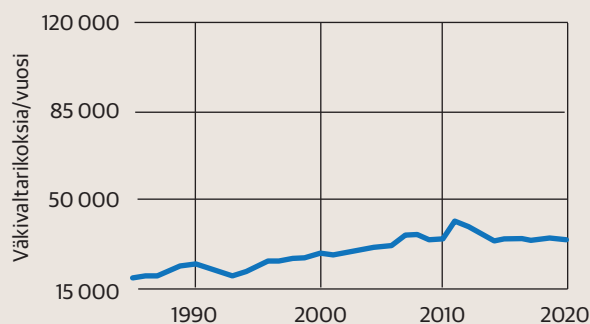
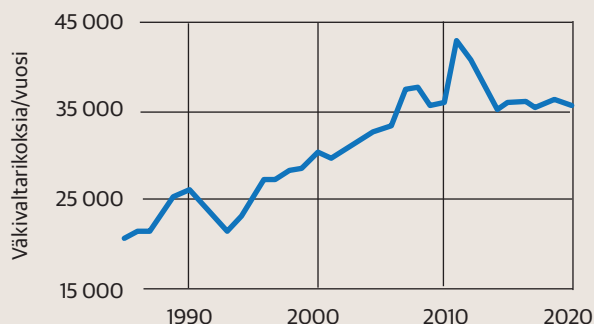
Kuvaajan tyyppi vaikuttaa siihen, millaisia päätelmiä sen perusteella voi tehdä.

Yleisimmät kuolinsyyt (Suomessa?)



A	B
C	D

Väkivaltarikosten määrä vuodessa



kuolinsyiden jakaumaa sektori- eli pii-rakkakuvaajana (A) ja pylväskuvaajana (B). Sektorikuvaajassa kuolinsyiden yleisyys selviää sektorien kokoja vertaamalla. Ihmissilmälle erojen hahmottaminen on kuitenkin vaikeaa: kuvaajasta erottuu lähinnä kaksi yleisintä kuolinsyytä (sydän- ja verisuonitaudit sekä syövät), mutta muiden kuolinsyiden suhteellisia osuuksia on liki mahdotonta vertailla. Pylväskuvaajassa sama tieto on esitetty helpommin hahmotettavassa muodossa.

3. Onko tietoa yritetty vääristää tai sumentaa?

Varo vaaraa. Sektorikuvaajaa käytetään usein tarkoituksella todellisten tietojen sumentamiseen, joten hälytyskellojen pitää aina soida, kun sellaisen näkee.

Jos et hahmota, minkä kokoisia sekto-

rit ovat, tarkista, löytyykö tieto numeroina.

Myös pylväskuvaaja voi olla harhaanjohtava. Pylväskuvaajassa numeerisen akselin pitää aina alkaa nolasta, jolloin pylväiden pinta-alat ovat suoraan verrannollisia esitettävien arvojen suuruuteen. Katkaistu akseli vääristää kuvaajaa. Viivakuvaajassa akselin katkaisu ei haittaa, koska kuvaajan tulkinta perustuu viivan korkeuteen. Viivakuvaajassa harhauttaminen perustuu sen sijaan pystyakselin laajuuden säätämiseen.

Verrataan esimerkiksi kahta viivakuvaajaa (C ja D), joissa on esitetty väkivaltarikosten muuttuminen aikasarjana. Kuvasta C on helppo nähdä, että vuosittainen rikollisuuden määrä on kasvanut seurantajaksolla huomasti. Samalla tavalla "oikein" piirretty kuvaaja D puolestaan antaa virheellisen vaikutelman

siitä, että rikosten määrä olisi pysynyt suunnilleen vakiona. Kun pystyakselin asteikko muutetaan riittävän laajaksi, aikasarja näyttää käytännössä vaakasuoralta, mikä hämärtää muutoksen suuntaa.

4. Onko kuvaaja selkeä?

Yksinkertainen on kaunista, myös tilastollisessa grafiikassa. Mitä enemmän koristeellisia elementtejä tai monimutkaisia kolmiulotteisia muotoja ja palkkeja kuvaajassa on, sitä hankalampi siitä on saada selvää. Tällaisten kuvaajien käyttö viittaa usein tahalliseen lukijan harhauttamiseen tai vähintäänkin graafikon ammattitaidon puutteeseen. ●

Lauri Nummenmaa

on professori valtakunnallisessa PET-keskuksessa.