Data-analyysin monet mahdollisuudet

Ari Turunen

Mittalaitteet tuottavat nykyään huikeita määriä dataa. Onneksi tietokoneiden tehon kasvu mahdollistaa entistä monipuolisempien datankäsittelymenetelmien käytön. Algoritmisen data-analyysin huippuyksikössä luodaan tehokkaita analyysimenetelmiä, joilla datasta seulotaan olennaista informaatiota eri tieteenalojen käyttöön. Informatiikka on saanut monia etuliitteitä, kuten bio, geo tai neuro. Suomen Akatemian rahoittamassa huippuyksikössä on tutkijoita sekä Helsingin yliopistosta että Aalto-yliopistosta.

Helsingin yliopiston tietojenkäsittelytieteen professori Esko Ukkonen on lähes 200 tieteellisessä artikkelissaan käsitellyt algoritmeja, datastruktuureja, koneoppimista ja bioinformatiikkaa. Vuosina 2004–08 hän oli Helsingin yliopiston ja Aalto-yliopiston yhteisen tietotekniikan tutkimuslaitoksen (HIIT) tutkimusjohtajana. Ukkonen on johtavia niin sanotun kombinatorisen hahmontunnistuksen kehittäjiä. Tutkimuksen lähtökohtana on luoda algoritmeja, jotka ratkaisevat avainsanahakujen ja muiden merkkijonojen käsittelyyn liittyviä perustehtäviä.

Ukkosen työ on hyvä esimerkki siitä, miten matemaattisten ja tilastollisten menetelmien käyttö on edistänyt tieteenalat ylittävää tutkimusta. Sovellusaloja ovat olleet esimerkiksi tautigenetiikka, uutisvirtojen seuranta, kielen kehitys, paleontologia, tähtitiede, ja ympäristötutkimus.

Ilman datan hallintaa ei pärjätä

Informaation määrä kasvaa dramaattisesti. NASAn EOSDIS -satelliittitietojärjestelmä (*Earth Observing System Data and Information System*) vastaanottaa päivässä kolme teratavua uutta dataa. Eri puolilla maailmaa oleville tutkijoille se jakelee sitä kaksi teratavua. Pelkästään tämä yksittäinen tietojärjestelmä tuottaa vuodessa 1,1 petatavua dataa. Luku on huikea, sillä maailman kirjallisen tuotannon arvioidaan olevan noin kaksi petatavua vuodessa. Vuosina 1999–2003 datan kasvu oli 30 % vuodessa. Vuonna 2007 ihmiskunnan kaikki tallentama data saavutti lähes 300 eksatavun määrän, arvioi Etelä-Kalifornian yliopisto tutkija Martin Hilbert. Uusien mittalaitteiden käyttöönoton myötä kasvu on yhä huimempaa. Maailman informaation tuotanto on nyt arviolta 5–10 eksatavua vuodessa.

Moderni tutkimus ei pärjää ilman datan hallintaa ja sen analyysia. Oman tutkimusaiheensa tuntemuksen lisäksi tutkijan on osattava käyttää tilastotieteen ja tietojenkäsittelytieteiden menetelmiä. Suurista datamääristä on löydettävä tutkimukselle olennainen tieto, jolloin puhutaan datan- tai tiedonlouhinnasta. Ukkosen mielestä laskennallinen lähestymistapa on jo itsestäänselvyys kokeellisen ja teoreettisen rinnalla.

– Monimutkaisista ilmiöistä yhä helpommin kerätty runsas data on johtanut datavetoiseen tutkimusasetelmaan, jossa ilmiön monimutkainen malli pyritään oppimaan automaattisilla menetelmillä suoraan datasta. Riittävän tarkoista malleista tulee niin mutkikkaita, että niiden käsittely on täysin mahdotonta ilman tehokkaita algoritmeja ja tietokoneita.

Informatiikat – tutkimusalojen menetelmäkokonaisuudet

Algoritmisen data-analyysin yksikkö kehittää algoritmeja, joiden avulla tietokoneohjelmat pystyvät seulomaan tehokkaasti tietoa hyvin suurista datamääristä.

– Laskennallisen tieteen perinteinen, lähinnä numeeriseen matematiikkaan perustuva työkalupakki on saanut rinnalleen laajenevan joukon tietojenkäsittelytieteen piirissä kehitettyjä tiedonhallinnan, erityyppisten tietojen yhdistämisen ja visualisoinnin sekä tiedonlouhinnan, koneoppimisen ja data-analyysin algoritmiikan menetelmiä. Näin eri tutkimusaloille on viime vuosina kehittynyt ”informatiikoiksi” kutsuttuja menetelmäkokonaisuuksia.

Hyvän esimerkin tiedonlouhinnan mahdollisuuksista tarjoaa tähtitiede, jossa tiedonlouhinta on tuonut tutkimukseen tehokkuutta.

– Tähtitieteen tutkimuksessa yksikkömme kehitti tiedonlouhintaa käyttävän data-analyysilinjan, jonka avulla etsitään tähtisikermiä uusimmista tähtiluetteloista. Sikermät ovat potentiaalisia tähtien syntyalueita. Tämän voi nähdä tutkimustyön tuottavuuden parannuksena: automaattinen perusmenetelmä löysi noin 15 000 sikermäehdokasta, josta käsityövaltainen jälkiprosessointi karsi lopulta esille hieman toistasataa kiinnostavaa uutta löydöstä.

Yksikköön kuuluva professori Heikki Mannilan ryhmä on tarkastellut myös paleontologista dataa Helsingin yliopiston paleontologien kanssa. Fossiililöydöksiä on ajoitettu algoritmien avulla.

– Tarkastelimme Euroopan ja Euraasian nisäkäsfossiileja viimeisten 20 miljoonan vuoden ajalta. Niiden automaattinen ajoittaminen onnistui niin, että löydetty järjestys vastasi suurilta osin asiantuntijoiden entiseen näkemykseen perustuvaa ryhmittelyä. Joukossa oli joitakin kiinnostavia poikkeavuuksia.

Ukkosen mukaan merkittäviä aloja ovat nykyään molekyylibiologiaan ja genetiikkaan liittyvä bioinformatiikka sekä geo- ja neuroinformatiikka.

Neuroinformatiikasta on saatu uusia, mielenkiintoisia tutkimustuloksia.

– Yksikössämme professori Aapo Hyvärinen analysoi tilastollisia koneoppimisalgoritmeja käyttäen, kuinka kahden henkilön aivot synkronoituvat toistensa kanssa, kun nämä henkilöt keskustelevat keskenään. Molempien koehenkilöiden aivotoimintaa mitataan samaan aikaan magnetoenkefalografialla (MEG). Synkronointi on hyvin heikkoa, eikä sitä voi nähdä mittausdatasta ilman juuri tätä tarkoitusta varten kehitettyjä koneoppimismenetelmiä. Koneoppimismenetelmillä voidaan nähdä tarkkaan rajatuilla aivoalueilla tapahtuvaa aivoaaltojen vahvuuksien synkroniaa.

Bioinformatiikan mahdollisuudet

Esko Ukkosen ominta aluetta on monitieteinen bioinformatiikka.

– Molekyylibiologinen sekvenssidata, ennen kaikkea DNA-jonot, on äärimmäisen kiehtova ja hedelmällinen algoritmitutkimuksen sovellusalue. Aloitimme DNA-jonojen käsittelyyn tarkoitettujen kombinatoristen merkkijonomenetelmien kehitystyön ryhmässäni jo 1980-luvulla. 1990-luvulla alkaneen DNA-datan voimakkaan kasvun seurauksena bioinformatiikan algoritmien ja mallinnusmenetelmien kansainvälinen kehitystyö on jatkunut voimakkaana.

Eliöiden DNA-merkkijonojen selvittäminen on tyypillinen esimerkki analyysista, joka hyödyntää isoa määrää dataa ja edellyttää hyviä algoritmeja.

– Yksikkömme osallistuu akatemiaprofessori Ilkka Hanskin ryhmän keskeisen malliorganismin, täpläverkkoperhosen, genomin *de novo* -sekvenssointiin kehittämällä suurtehosekvenssoinnin vaatimia algoritmeja. Täpläverkkoperhosen genomi on kooltaan noin 300 miljoonaa emästä. Haasteellisena laskentatehtävänä on rekonstruoida tällainen DNA-jono mahdollisimman hyvin, kun siitä on käytettävissä vain lyhyitä, korkeintaan muutaman sadan emäksen pituisia näytteitä. Täpläverkkoperhosesta tulee ensimmäinen korkeampi organismi, jonka genomi on selvitetty Suomessa – tässä on hyvä syy julistaa täpläverkkoperhonen Suomen kansallisorganismiksi!

Ukkosen mielestä bioinformatiikka tarjoaa uudentyyppisiä ja mielenkiintoisia haasteita suomalaisille algoritmitutkimukselle ja teollisuudelle.

– Jatkuvasti tehostuva DNA-sekvenssointi on tärkein bioinformatiikan kysyntään vaikuttava tekijä. Olemme pian tilanteessa, jossa kunkin

yksilön genomi voidaan selvittää edullisesti ja nopeasti. Tämä tuottaa valtavasti sekvenssidataa, jonka hallinnassa ja analysoinnissa riittää haastetta lähivuosikymmeniksi. Miten DNA:n koodaama ohjelma on rakennettu ja miten se toimii – mikä on sen syntaksi ja semantiikka?

Kun ihmisen genomi selvitettiin, tutkijat ovat keskittyneet selvittämään, kuinka geenit säätelevät toisiaan ja miten geenivirheet syntyvät. Voidaan selvittää myös, mitkä muut tekijät, tunnettujen altistavien geenien lisäksi, ovat vaikuttaneet sairauden puhkeamiseen. Tilastollisten menetelmien avulla voidaan kartoittaa eri vaihtelulähteet ja löytää niiden ja sairauksien välisiä riippuvuussuhteita.

Akatemiaprofessoreiden Lauri Aaltosen ja Jussi Taipaleen ryhmien kanssa tehty tutkimus osoitti, että ihmisillä, joiden DNA-jonossa on tietynlainen mutaatio, on kohonnut riski sairastua paksusuolisyöpään. 75 prosenttia eurooppalaisista kantaa perimässään riskiä lisäävää muotoa.

Suomen Akatemian ja Tekesin *Finnsight 2015* -raportin mukaan Suomen menestymisen mahdollisuuksia tulevaisuudessa ovat muun muassa bio-osaaminen ja tätä tukevat tietotekniset palvelut. Suomella olisi jo nyt mahdollisuuksia tarjota osaamistaan vientituotteeksi asti.

– Lähtökohdat ovat hyvät, koska korkeatasoista tutkimusta on. Tarvittaisiin kuitenkin lisää ”Tieteestä tuotteeksi”-mentaliteettia ja hyviä esikuvia. Valitettavasti biotekniikkalähtöisen teollisuuden kehittäminen on osoittautunut haasteelliseksi eikä Euroopan poliittinen ilmasto suosi sitä. Silti voi kysyä, miksi kehitystä eteenpäin vievät molekyylibiologian mittaustekniikat ja -laitteet tulevat jokseenkin poikkeuksetta Euroopan ulkopuolelta. Näiden kylkiäisenä voisi myydä myös algoritmeja ja tietokoneohjelmia.

Ihmistieteiden informatiikka

Tietokonelingvistiikassa Suomi on ollut perinteisesti hyvä. EU:n valtavat dokumenttikokoelmat tarjoavat mielenkiintoisia sovelluskohteita. Automaattinen kielenkääntäminen paranee, kun eri käännöspareista alkaa olla valtavia tietokantoja. Tietokoneelle pystytään opettamaan, millaiseksi eri hahmot kussakin kielessä kääntyvät. Algoritmisen data-analyysin yksikössä kehitetty käännösohjelma on nopea.

– Automaattinen kielenkääntäminen on erittäin kilpailtu ja varsin kypsä ala siinä mielessä, että huomattavia käytännön edistysaskelia on enää vaikea tehdä. Modernit konekääntämisjärjestelmät (esim. *Google Translate*) perustuvat tilastollisiin malleihin, jotka opetetaan käyttäen massiivisia monikielisiä aineistoja ja suurta laskentakapasiteettia. Yksikössämme kehitetty menetelmä on käytännön käännöslaadultaan samaa tasoa kuin kilpailijat, ja sen etu on suuri käännösnopeus. Järjestelmä ei juurikaan sisällä lingvististä esitietoa, vaan käännöstulos syntyy kahden tilastollisen mallin yhdistelmänä: käännösmallin, joka liittää kaksikielisiin n-grammeihin todennnäköisyyden sekä kielimallin, joka liittää kohdekielen lauseisiin todennäköisyyden.

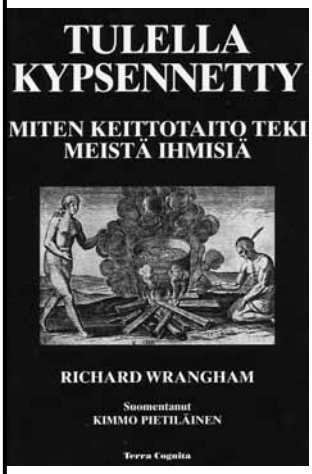
Geologian tutkimuskeskuksen ansiosta Suomi on yksi maailman parhaiten kartoitettu maa kallio- ja maaperältään. Valtakunnan metsien inventoinnissa yhdistetään satelliittikuvat numeeriseen dataan, jolloin saadaan täsmälliset tiedot metsien vuotuisesta kasvusta. Metsätraktoreihin saadaan paikkatietojärjestelmien avulla heti tieto siitä, mitä saadaan hakata ja mitä ei.

Dataa syntyy paljon, mutta sitä ei päästä hyödyntämään niin hyvin kuin pitäisi. Eri hallinnonalojen reviirit estävät tämän toistaiseksi. Kun Maanmittauslaitos halusi antaa kaikille suomalaisille mahdollisuuden käyttää laitoksen kokoamaa paikkatietoaineistoa ilmaiseksi, päätös meinasi tyssätä valtiovarainministeriöön.

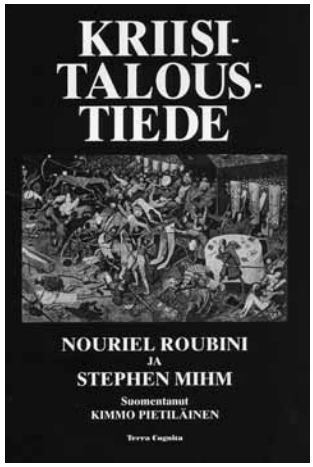
– Julkisin varoin tuotetun datan julkisuus ja vapaa saatavuus on yleensä ihmisten edun mukaista ja voi innoittaa yrityksiä kehittämään hyviä tuotteita. Mutta kaikkea yritysten kannalta kiinnostavaa dataa ei voi mennä julkistamaan. Esimerkiksi yksityisyyden suoja tulee varsin pian vastaan, Ukkonen huomauttaa.

Kirjoittaja on tietokirjailija.

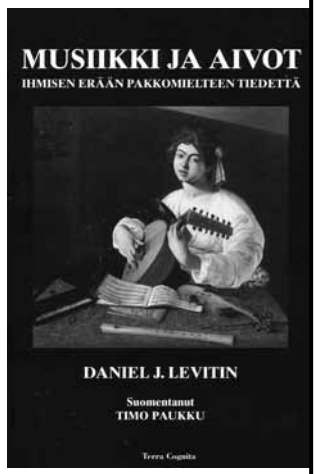
Parasta suomalaista tietokirjallisuutta



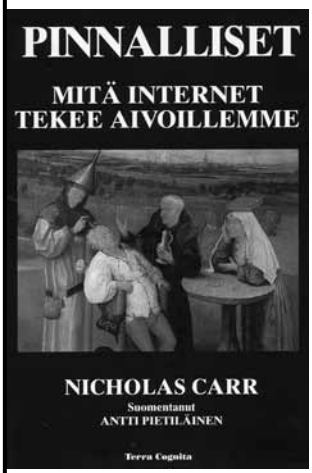
Richard Wrangham
Tulella kypsennetty. Miten keittotaito teki meistä ihmisiä.
Ovh. 40,–



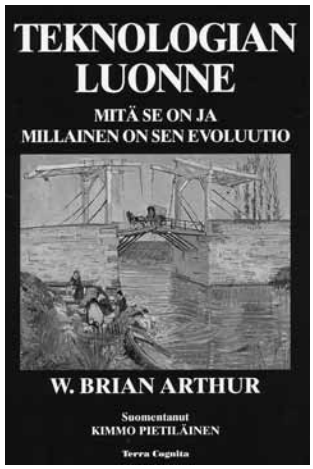
Nouriel Roubini ja Stephen Mihn
Kriisitaloustiede
Ovh. 40,–



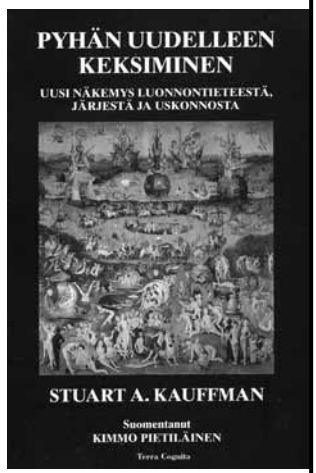
Daniel J. Levitin
Musiikki ja aivot. Ihmisen erään pakkomielteen tiedettä
Ovh. 40,–



Nicholas Carr
Pinnalliset. Mitä internet tekee aivoillemme
Ovh. 40,–



W. Brian Arthur
Teknologian luonne. Mitä se on ja millainen on sen evoluutio
Ovh. 40,–



Stuart A. Kauffman
Pyhän uudelleen keksiminen. Uusi näkemys luonnontieteestä, järjestä ja uskonnosta
Ovh. 40,–

Kirjakaupasta tai suoraan kustantajalta
TERRA COGNITA OY
www.terracognita.fi