Datan louhinta ei korvaa tieteellistä tutkimusta

Kuva: Kelan kuva-arkisto

JENNI BLOMGREN
Kirjoittaja on Yhteiskuntapolitiikka-lehden toimittaja ja Kelan tutkimuspäällikkö

Tutkimusaineistojen maailmassa ollaan jo jonkin aikaa oltu suuren murroksen äärellä. Todella suuret aineistot (iso data, *big data*) ovat tätä päivää, ja datan määrä kasvaa edelleen kiihtyvällä vauhdilla: se mikä oli eilen isoa, on tänään enää keskikokoista. Jokaisesta suomalaisesta kertyy valtavasti henkilötietoa erilaisiin viranomaisten rekistereihin. Näitä on totuttu yhdistelemään keskenään tutkimuskäyttöä varten, mutta kasvavaa kiinnostusta on perinteisen rekisteritiedon yhdistämiselle myös uudentyyppisiin tietokokonaisuuksiin. Tällaisia ovat esimerkiksi kauppaketjujen ostorekisterit ja älypuhelinten tallentamat paikkatiedot.

Tehokkailla datan louhintamenetelmillä rikkaiden aineistojen analysointi on nopeaa. Tekoäly, koneoppiminen ja neuroverkot alkavat tulla osaksi yhteiskuntatieteilijöidenkin työkalupakkia. Viime vuosien aikana datatiede (*data science*) on noussut muodikkaaksi tieteenalaksi. Datatieteen yhtenä keskeisenä ideana on etsiä tietomassoista riippuvuuksia tilastotieteen ja tietojenkäsittelytieteen menetelmin. Datatieteilijää on *Harvard Business Review*'ssa jopa kutsuttu 2000-luvun seksikkäimmäksi ammatiksi!

Isojen aineistojen ja datan louhinnan huumassa on muistettava, että tieteellinen tieto kertyy kuitenkin pienin askelin. Suurten aineistojen louhinnalla voi saada nopeasti vastauksia mitä-kysymyksiin, mutta miksi-kysymyksiin löytyy vastaus vain tarkempien, hyvin rajattujen tutkimuskysymysten ja huolella mietittyjen asetelmien kautta. Syvimmälle menevät tulkinnat löytyvät laadullisen, ymmärtämään pyrkivän tutkimuksen kautta. Tutkimusalueeseensa ja sen teoriaan paneutunut tutkija tietää, mitä aineistolta kannattaa kysyä – tai toisin päin: millaista aineistoa pitää kerätä saadakseen kysymykseensä vastauksen. Myös yhteiskunta- ja terveystieteilijöillä on paineita muodostaa yhä laajempia tutkimusaineistoja osin sillä ajatuksella, että kunhan dataa on paljon, kiinnostavat havainnot nousevat ikään kuin itsestään suoraan aineistosta. Tietosuojalainsäädännön kannalta tällainen lähestymistapa henkilötietojen keräämiseen ei kuitenkaan ole aivan asianmukainen: henkilötietoja saa kerätä vain siinä määrin kuin on tarpeen tutkimuskysymyksiin vastaamiseksi.

Tietojen ennakkoluulottomalla yhdistelyllä ja isoja datamassoja tehokkaasti hyödyntävillä menetelmillä – tietosuojasääntely huomioiden – voi löytyä uusia yllättäviäkin asioiden välisiä yhteyksiä. Niiden perusteella voi puolestaan syntyä oivalluksia tarkemmiksi tutkimuskysymyksiksi syvällisempään tieteelliseen tarkasteluun. Molempia lähestymistapoja tarvitaan.

*

Aloitin viime vuonna toimittajana *Yhteiskuntapolitiikka*-lehden laajennetussa toimituskunnassa. Tuntuu hienolta päästä mukaan tekemään yhtä Suomen johtavista yhteiskuntatieteellistä tutkimusta julkaisevista lehdistä!