Julkaistu teoksessa: Järvi, Ulla & Tammi, Tuukka (toim.) 2016: Maito tappaa ja muita outoja tiedeuutisia. Vastapaino: Tampere.

Algoritmin valta ja toimittajan vastahanka

Minna Ruckenstein

Googlen hakutuloksista voi piirtää reaaliaikaisen kartan, joka kertoo kuinka flunssa-aallot leviävät maailmanlaajuisesti. Näin uutisoitiin toukokuussa 2012 Helsingin Sanomissa. Big data ja ennakoivat algoritmit tulivat mediaan suurten lupausten saattelemana.

Googlen malli perustui siihen, että flunssa-aallon voimistuessa sairastuneet naputtelevat hakukoneeseen tautiin ja sen oireisiin liittyviä avainsanoja. Google sai nostetta ennakoivien algoritmien kehittäjänä, vaikka mallin todentamia arvioita taudin leviämisestä epäiltiin. Jo flunssakaudella 2012–2013 ennustukset flunssa-aaltojen liikkeistä olivat jonkin verran ylimitoitettuja. Sen jälkeen ne menivät vielä pahemmin pieleen. Sittemmin hakukoneen flunssa-aaltoja mallintavan algoritmin ennustavaa voimaa on pidetty vaatimattomana.

Algoritmeihin liittyvät onnistumistarinat ovat samantapaisia kuin muutkin teknologiatarinat. Teknologia esitellään suurelle yleisölle ratkaisuna. Algoritmit seulovat aineistosta hyödyllisen tiedon: algoritmi esitetään tiedon tuottajana pikemminkin kuin teknologian käyttäjät tai soveltajat.
Toimittajille ja muille tiedonvälittäjille ei ole vielä täysin hahmottunut, kuinka tarpeellisia he ovat datamaailman ja algoritmeille perustuvan vaikutusvallan avaamisessa. Tekniikan kehittyessä aineistojen tallentaminen ja työstäminen on aiempaa vaivattomampaa. Kukaan ei kuitenkaan vielä tiedä mitä kaikkea datan varastoimisesta ja uusista käytöistä seuraa. Toimittajia tarvitaan taustoittamaan, selvittämään asioiden kulkua ja osoittamaan niiden välisiä yhteyksiä. Siksi digitaalisiin aineistoihin liittyvistä asioista pitäisi kiinnostua, vaikka ne herättäisivät ristiriitaisia tunteita ja heittäisivät ihmisen henkilökohtaiselle epämukavuusalueelle.

Datamaailman ääripäät

Stanfordin yliopiston kampuksella Kaliforniassa järjestetyssä Quantified Self -konferenssissa meuhkattiin syyskuussa 2012 yksilöiden käyttäytymisestä kertovista tietomassoista, joita kertyy yhä kiihtyvään tahtiin. Käytännössä laajat aineistot voivat olla melkein mitä vain: potilastietoa, paikkatietoa, sähkön kulutukseen liittyvää tietoa, ostotapahtumia, nettisivujen lokitiedostoja tai sosiaalisen median sisältöä. Näitä aineistoja syntyy esimerkiksi television tai älypuhelimen käytön sivutuotteena, luottokortilla maksaessa tai kanta-asiakaskorttia höyläämällä.

Esitys toisensa jälkeen kuvaili tietomassojen mullistavan tutkimuskenttää, teollisuuden prosesseja, poliittista vaikuttamista ja kaupunkisuunnittelua. Suurten aineistojen visioitiin auttavan terveyden edistämisessä ja hoitojen vaikuttavuuden mittaamisessa. Lisäksi aineistoista luvattiin tukea liikennesuunnitteluun, kriminaalipolitiikkaan, ruoantuotantoon ja ilmastonmuutoksen hidastamiseen.

Parhaimmillaan suuret aineistot avaavat aiempaa laaja-alaisempia näkymiä tutkittavaan ilmiöön. Yksittäiset tapahtumat voi sijoittaa osaksi suurempaa kuvaa. Aineistojen avulla voi piirtää esiin ajallisia ja paikallisia yhteyksiä ja riippuvuuksia. Ne kuvaavat ihmisten parveilua, heidän välisiä sosiaalisia verkostojaan ja yhteisiä solmukohtia. Näin aineistoanalyysi pyrkii kertomaan mistä tulemme, mihin pyrimme ja miten tekemisiimme voisi vaikuttaa.

Kulttuuriantropologiksi koulutettuna tohtorina kuuntelin hämmentyneenä puhetta, joka todisteli, ettei ihmistieteilijöitä pian enää tarvita. Data kertoisi ihmisten käyttäytymisen ja aikomukset: teoretisointi tai merkitysten tulkinta olisi turhaa, kun aineistot paljastaisivat ihmisten toiminnan mielen ja suunnan. Aineistot auttaisivat ennaltaehkäisemään kroonisten sairauksien puhkeamista. Kenenkään ei enää tarvitsisi kärsiä kakkostyypin diabeteksesta, kun avun saisi jo ensimmäisten varoitusmerkkien ilmaantuessa. Aineistot erottelisivat onnettomuuksiin tai rikoksiin johtavat kehityskulut jo ennen kuin ne tapahtuvat. Poliisin voisi laittaa seisoskelemaan juuri siihen kadunkulmaan, jossa seuraava tappelu todennäköisesti tapahtuisi lauantai-iltana puolenyön aikaan.

Tiedeyhteisön vastareaktio syntyi saman tien: datauskovaisia pidettiin lähinnä harhaisina ja hurmoshenkisinä. Tieteen ja teknologian tutkimuksen konferenssissa San Diegossa vyörytettiin kritiikkiä massadatamaailmaa kohtaan. Data-analyysiin liitettyjä lupauksia pidettiin katteettomina. Suhteellisuudentajun puutetta pidettiin vakavana ongelmana: se veisi massadatailmiöltä laajempaa

uskottavuutta ja estäisi hedelmällisen yhteistyön syntymistä tieteenalojen välille. Puheenvuoroissa muistutettiin, että suuret aineistot ovat yhtä lailla alttiita inhimillisille vääristymille ja painotuksille kuin muukin tiedontuotanto. Numeroilla tehdään politiikkaa, vaikka ne esitetään epäpoliittisina. Ongelmallisena pidettiin sitäkin, että suuret aineistot latistavat käsitystä ihmisestä ja hänen sosiaalisista aikeistaan. Kiinnostava variaatio unohtuu, kun etsitään yhteisiä nimittäjiä. Sosiaalisia verkostokaavioita tarkastelemalla syntyy käsitys ihmisten liittymisestä toisiinsa, mutta kaaviot eivät paljasta vielä mitään siitä, mihin ihmiset sosiaalisuudellaan pyrkivät.

Tiedon manipulointia

Googlen flunssa-algoritmin väärät ennusteet huomioitiin mediassa epäilyillä, joiden mukaan sikainfluenssan kaltaiset poikkeukselliset epidemiat saivat ihmisjoukot naputtelemaan hakukoneita. Kun hakujen määrä kasvoi, algoritmi ei toiminutkaan entiseen tapaan. Epäiltiin, ettei Googlen flunssa-aaltomallissa huomioitu tarpeeksi vuodenaikojen vaihtelua. David Lazerin ja kumppaneiden Science-lehdessä vuonna 2014 julkaistussa artikkelissa käydään läpi perusteellisesti algoritmiin liittyviä ongelmia. He pitävät erityisen pulmallisena, ettei flunssa-algoritmia oltu päivitetty vuoden 2009 jälkeen. Verrattuna Googlen kaupalliseen algoritmiin ero oli huikea. Kaupallista algoritmia päivitetään ahkerasti. Yritys pyrkii vaikuttamaan algoritmin muutoksilla hakutulosten järjestykseen ja saatavuuteen. Googlea kiinnostaa, kuinka personoidut hakutulokset luovat mainonnalle tilaa, ohjailevat yritysten näkyvyyttä ja kuluttajien käyttäytymistä. Yrityksessä seurataan reaaliaikaisesti, kuinka tulosten esillepano vaikuttaa hakemisen tapoihin ja klikattaviin sivuihin.

Mitä paremmin ihmisten nettikäyttäytymistä ymmärretään, sitä houkuttelevampaa on hakutulosten ja ihmisten käyttäytymisen manipulointi. Facebook muunteli tutkimustarkoituksessa viikon ajan algoritmia, joka ohjaa millaisia päivityksiä käyttäjät näkevät uutisvirrassaan. Tavoitteena oli tutkia vaikuttaako positiivisten tai negatiivisten sanojen määrä käyttäjän lukemissa päivityksissä hänen itse kirjoittamiensa päivitysten tunnetilaan. Tunnesanoihin liittyvä manipulointi kohdistui satunnaisesti valittuihin englanninkielisiin käyttäjiin, joita oli kaiken kaikkiaan 689 003. Käytännössä tutkituilta käyttäjiltä piilotettiin heidän kavereidensa tekemiä päivityksiä, joissa oli käytetty joko negatiivisia tai positiivisia tunneilmaisuja. Adam Kramerin ja kumppaneiden tutkimus julkaistiin kesäkuussa 2014 arvostetussa *Proceedings of the National Academy of Sciences* –lehdessä. Se oli päätetty julkaista lehdessä, vaikka toimituskunta ilmaisee artikkelin yhteydessä huolensa siitä, ettei käyttäjiltä oltu pyydetty lupaa tutkimuksen tekemiseen. Tutkijat perustelivat toimimista ilman erillisiä tutkimuslupia

käyttäjiltä sillä, että käyttäjät olivat suostuneet tietojensa hyödyntämiseen hyväksymällä Facebookin käyttöehdot.

Tutkimuksen tuloksena oli, että tunnetilat tarttuivat. Käyttäjät todellakin alkoivat käyttää päivityksissään negatiivisia tai positiivisia sanoja sen mukaan, mille tunnetilalle heidät oli altistettu. Tutkimuksen mukaan tulokset kertoivat kuitenkin pienistä muutoksista, eivät mittavista käyttäytymisen muutoksista. Tutkimuksesta kerrottiin laajalti mediassa, myös suomalaisissa sanomalehdissä ja verkkomedioissa. Tunteiden tarttumista raportoitiin tavalla, josta ei käynyt ilmi muutosten luonnetta. Uutisissa selostettiin käyttäjien järkyttyneitä reaktioita ja toisteltiin Twitter-viestejä, joissa tutkimusta kuvailtiin pahaksi ja pelottavaksi. Sosiaalisen median reaktio viestivirran manipulointiin oli niin voimallinen, että sekä tutkijat että Facebookin edustajat joutuivat selittelemään.

Kramer ja muut kirjoittajat perustelivat tutkimustaan sillä, että sen avulla tehdään näkyväksi manipuloinnin todellisuutta, josta suurin osa ei koskaan tule julki. Epäilemättä tutkimukselle oli myös tiedeyhteisössä tilausta: se on kerännyt jo satoja siteerauksia. Vaikka tunteiden tarttuminen kertoi pienistä muutoksista käyttäjien tunneilmaisuissa, niilläkin voi saada aikaan merkittävää vaikutusta, kun sosiaalisen median käyttäjämäärät ovat suuria. Tutkija Ralph Schroeder arvioi *Big Data and Society* -lehdessä Facebook-tutkimusta ja pitää sitä jäävuoren huippuna. Hän kuvaa tutkimusta osana tieteellinen kokeilujen jatkumoa: vastaavanlaisia kokeiluja on tehty ja tehdään vastakin. Uutta on kokeilujen kytkeytyminen arkisiin median käyttöihin. Kansalaisten voi olla vaikea havaita digitaalisen tiedonkeruun laajuutta tai ymmärtää, kuinka poliittista ja kaupallista vaikuttamista tehdään sosiaalisen median areenoilla. Yrityksissä käyttäjistä koostetaan yksityiskohtaisia profiileja ja niiden avulla kohdennetaan mainontaa ja pyritään vaikuttamaan ostoaikeisiin. Schroederin mielestä olennaista on, etteivät käyttäjät saa tietää, kuinka paljon heistä tiedetään. Kun he eivät tiedä, he eivät voi reagoida tähän tietämiseen tai vastustaa sitä millään tavalla.

Tutkijoiden tulisi tuottaa tietoa, joka auttaisi näkemään datamaailman vääristymiä ja arvioimaan algoritmien osuvuutta. Ongelmana on kuitenkin se, ettei tutkijoilla ole pääsyä kaupallisten yritysten omistamiin ja hallinnoimiin aineistoihin ja algoritmeihin. Tiedon tuottamisesta tulee uudella tavalla harvainvaltaa, kun vain osalla tutkijoista on pääsy yritysten keräämään aineistoon ja aineistojen analyysiin.

Älykkäät kissaihmiset

Datan keruuseen ja käyttöön liittyvät jännitteet ja ristiriidat saivat minut kiinnostumaan datamaailmasta. Mitä enemmän olen oppinut, sitä tärkeämmältä kentän tuntemus tuntuu. Olen ollut linjaamassa Helsingin yliopiston Kuluttajatutkimuskeskukseen *Data, Self and Society* -tutkimusryhmää, johon liittyneet Krista Lagus ja Marjoriikka Ylisiurua avaavat ihmistieteilijöille data-analyysin periaatteita. Yhteinen tutkimushankkeemme käynnistyi, kun Kuluttajatutkimuskeskus avasi yhteistyössä Allerin, FIN-CLARINin ja CSC:n kanssa Suomi24-aineiston tutkimuskäyttöön. Kyseessä on suomalaisittain ainutlaatuinen avoimen datan hanke. Aineisto on kerännyt ympärilleen *Citizen Mindscapes* -tutkimuskollektiivin, johon on kytkeytynyt kymmeniä tutkijoita ja yhteistyötahoja. Aineistoanalyysi ja monialainen yhteistyö tuottaa tutkimusta ja toivon mukaan myös uudenlaisia sosiaalisen median käyttöjä.

Suuria aineistoja ei voi hallita, jos ei ymmärrä numeroita. Data-analyysi synnyttää uudenlaisia asiantuntija-asemia, kun osa tutkijoista hallitsee aineistoanalyysin uuden välineistön ja toiset eivät. Viimeisen kolmen vuoden aikana olen kuullut kymmeniä esityksiä datatalouden vallankumouksellisista mahdollisuuksista, uusista datainfrastruktuureista ja tarpeesta vapauttaa data. Power point -esitysten nuolikaaviot näyttävät datan liikkuvan vaivattomasti toimijalta toiselle. Seminaareissa puhutaan huomattavasti vähemmän työstä, joka vaaditaan datan liikuttamiseen tai kuinka mustasukkaisesti organisaatiot ja yritykset voivat aineistojaan varjella. Kaikkea aineistoa ei voi jakaa: niistä voisi paljastua asioita, jotka on parempi pitää salassa. Usein aineiston avaamisen esteenä on kuitenkin se, ettei jakamisen hyötyjä ymmärretä. Siksi pitäisi puhua enemmän siitä, miten avoin data voi edesauttaa maailman ymmärrystä ja uusia toimintatapoja, sekä mitä ovat uudet datakäytännöt, kuka niitä hallitsee ja mitä niistä seuraa.

Suomi24-keskusteluaineisto on ajallisesti pitkäkestoinen ja kansainvälisestikin poikkeuksellisen laaja sosiaalisen median aineisto. Viestejä on yli 70 miljoonaa ja niihin on tallentunut 15 vuoden ajalta suomalaista keskustelua. Parhaillaan käynnissä olevan työn tavoitteena on tunnistaa sosiaalisen median keskustelua ohjaavia aaltoja tai kaavamaisuuksia. Monialainen tutkimusryhmämme on tehnyt aineistokokeiluja ja yrittänyt löytää tapoja jäsentää valtavaa aineistomassaa. Samalla olemme pyrkineet hahmottamaan sosiaalisen median keskustelujen tuottamisen ehtoja. Suomi24:n kirjoittajia voi ajatella verkkoyhteisönä, jonka sisällä on lukuisia pienempiä yhteisöjä, joita muodostuu palveluun aihepiireittäin määriteltyjen keskustelupalstojen mukaan. Keskustelua vievät eteenpäin esimerkiksi harrastukset, sukupuolinen suuntautuneisuus, perheen perustaminen tai terveyteen liittyvät ongelmat. Palstoilla keskustellaan yhteisistä kiinnostuksenkohteista tai vakaumuksista,

saman elämäntilanteen tai automerkin jakavien tai samalla paikkakunnalla asuvien kanssa. Vilkkaita keskusteluja käydään esimerkiksi ikävästä ja lemmikkieläimistä: koiraihmiset juttelevat keskenään ja kissaihmiset keskenään. Osalle keskustelu on oman yrityksen tai poliittisen näkökulman edistämistä, toisille puhdasta ajanvietettä.

Yksi tapa hahmottaa aineistoanalyysin tavoitteita on pohtia, mihin sosiaalisen median tutkimuksella pyritään. Aineistoja syntyy reaaliaikaisesti, kun ihmiset osallistuvat keskusteluihin. Siinä mielessä ne ovat ainutlaatuinen tapa päästä käsiksi kansalaisten kiinnostuksenkohteisiin ja mielenliikkeisiin. Sosiaalista mediaa voi arvioida ja jäsentää osana tutkimusta, uuden oppimista, poliittista vaikuttamista ja osallistamista. Kun aineistosta löydetään esimerkiksi uudissanoja, voi nähdä keskustelun suunnan. Algoritmit voivat olla apuna osoittamassa keskustelun liikkeitä. Jännitteitä tunnistamalla niitä voidaan ehkä käsitellä ennen kuin niistä tulee avoimia konflikteja.

Oman työni näkökulmasta on selvää, että dataosaajat ja ihmistieteilijät tarvitsevat toisiaan. Välillä kiivaatkin keskustelut Kuluttajatutkimuskeskuksessa ovat vahvistaneet yhteistä tutkimusnäkemystä. Pyrimme edistämään data-analyysin keinoin eri tahojen välistä dialogia ja uuden oppimista, emme kansalaisten seurantaa, leimaamista ja ylhäältä ohjailua. Haluamme edistää datankäyttöä, joka ei pyri jäljittämään ja hallitsemaan ihmisiä vaan antaa heille välineitä hahmottaa omaa maailmaansa.
Dataosaajat ja ihmistieteilijät voivat yhteistyöllään osoittaa algoritmien määrittämän tiedon ja vallan rajat. Data-analyysin avulla voi tahattomasti johtaa harhaan tai tahallisesti harhauttaa. Matti Nelimarkka osoitti leikkimielisellä harjoituksellaan Suomi24-hackatonissa, että tarpeeksi kauan aineistoa pyörittämällä voi todistaa, että Suomi24:n palstoille kirjoittavat kissaihmiset ovat älykkäämpiä kuin koiraihmiset. Tämä tieto löytyi, kun saatiin todistettua sopivin laskutavoin, että kissaihmiset puhuvat keskustelupalstoilla enemmän matematiikasta ja koiraihmiset tisseistä.

Toimittajan vastahanka

Suurin hype on ohitse, eivätkä datauskovaiset enää julista suurten aineistojen kuljettavan kansalaisia maailmaan, jossa aineistot tuottaisivat täydellisen tiedollisen läpinäkyvyyden. Suuryritykset eivät ole muuttuneet data-analyysin myötä hyväntekijöiksi, eikä köyhyyttä saada maailmasta poistettua pelkällä datan pyörittelyllä. Älykellot ja sykemittarit eivät tehneet ihmisestä järkevää ja vastuullista terveystoimijaa. Datankeruu ja sen analyysi ovat kuitenkin osoittaneet voimansa. Näkökulmien kohdentaminen ja erilaisten kenttien avaaminen aineistotyön avulla on tullut yrityksiin ja yhteiskuntaan jäädäkseen.

Data-analyytikoiden tärkeänä tavoitteena on maksimoida laskennallinen voima ja algoritmitarkkuus, jotta yhä laajempia ja monimuotoisempia tietovarantoja saataisiin analysoitua ja vertailtua. Lentokentän logistiikka on saatu sujuvammaksi yhdistämällä tietoa sääolosuhteista ja lentokoneiden reaaliaikaisista liikkeistä. Keskosvauvoja voidaan hoitaa tehokkaammin yhdistämällä elintoiminnoista kerättyä aineistoa ja ennakoimalla sen avulla terveydentilan kehittymistä. Kuluttajaa puhutellaan aiempaa kohdistetummin jäljittämällä heidän käyttäytymistään verkon kauppapaikoilla ja sosiaalisen median palveluissa. Näin saadaan tietoa ihmisten tiedontarpeista tai esimerkiksi siitä, mihin aikaan vuorokaudesta he kirjautuvat verkkopalveluun, mistä he sinne tulevat, kauanko he sivuilla viihtyvät ja mitä he ostavat. Data-analyysi ja algoritmit tehostavat viestintää ja tiedonhakua. Suurin osa toimittajista ja tutkijoista ei taipuisi enää siihen, että työtä pitäisi tehdä ilman Googlea. Samalla yhden hakukoneen varassa toimimiseen liittyy tiedonvälityksen näkökulmasta ongelmia. Tavalliset käyttävät eivät pysty arvioimaan Google-hakujen luotettavuutta tai tapaa, jolla niitä personoidaan. Vain murto-osa verkon sisällöstä löytyy hakukoneilla.

Datamaailman mittasuhteita, ominaispiirteitä, hyviksiä ja pahiksia on vaikea hahmottaa. Mediassa on kerrottu Googlen ottaneen kritiikin vakavasti ja integroineen Yhdysvaltain tartuntatautiviraston (DCG) aineistot avuksi ennustamaan flunssakauden alkua, taudin leviämistä ja voimakkuutta. Google-kriittiset epäilevät edelleen flunssa-aaltomallia tai ainakin pitävät sen hehkuttelua turhana mainontana. Näkökulmista tulee helposti asetelmallisia, ollaan joko puolesta tai vastaan. Olennaista olisi löytää mitä tapahtuu harmaammalla alueella, joka harvoin nousee otsikoihin. Kohuihin keskittyvä journalismi kertoo yksittäisistä tapahtumasarjoista tai paljastuksista, mutta toimittajia tarvitaan avaamaan arkisempia datan keruuseen, analyysiin ja käsittelyyn liittyviä näkökulmia. Lisäksi on tärkeää kuvata datamaailmaa osana laajempaa poliittista, taloudellista ja ekologista kenttää. Esimerkiksi automaattisen datan keruun ympäristövaikutuksista puhutaan vähän. Kiinnittämällä huomio sähkönkulutukseen tai sensoreihin tarvittaviin raaka-aineisiin datamaailma avautuisi myös ympäristönäkökulmasta.

Digitaalisen maailman suurina uhkina puhutaan yksityisyyden ja tietoturvan menetyksistä tai identiteettivarkauksista. Yhtä lailla uhkaavaa on tietoverkkorakenteiden ja tiedonhaun monopolisoituminen sekä tiedontarpeiden yhdenmukaistuminen. Jos toimittajat googlaavat samoja sivuja, eikä kukaan enää poistu toimituksesta haastattelemaan kansalaisia ja asiantuntijoita, journalismi muuttuu merkityksettömäksi. Entinen lehdenlukija voi itse googlata samat sivut.

Toimittaja voi tehdä itsensä tarpeelliseksi antamalla askelmerkkejä digitaaliseen todellisuuteen. Hän voi asettua vastahankaan: välttää liiallista teknologiaintoilua ja toisaalta tarkastella kriittisesti massadataan ja algoritmeihin liitettyjä uhkia. Teknologiapiireissä puhutaan tahattomista seurauksista, joita uusien teknologioiden käyttöönotto aiheuttaa. Jos uusista teknologioista kertoo tavalla, jonka teknologiamaailmaan vihkiytymätönkin ymmärtää, voi tuoda esille miten vähän asiantuntijatkin saattavat ymmärtää siitä, minkälaista maailmaa he rakentavat.

Tarinoita datamaailmasta

Dataosaajien ja tilastollisten menetelmäosaajien rinnalla työskentely on opettanut, että tarvitsemme inhimillisiä ja realistisia tarinoita datamaailmasta. Kollegaltani Krista Lagukselta olen oppinut, kuinka teknologiaan ja suuriin aineistoihin liittyvää magiaa tuotetaan: tuloksia esitellään ikään kuin ne vain tupsahtaisivat maailmaan. Data-analyysistä ja sen tuloksista kerrotaan korostamalla analyysin helppoutta ja etuja. Helsingin Sanomat (13.3.2016) kuvailee data-analyytikosta kertovassa jutussa, kuinka Netflixin suosikkisarja *House of Cards* olisi voitu synnyttää pelkästään ihmisten katselutapoja analysoimalla. Mihin tarvitaan käsikirjoittajaa, kun on olemassa algoritmit! Data-analyysin tuloksista tuodaan esille vain kiinnostavimmat ja nekin voidaan kertoa liioitellen niiden osuvuutta ja esittävää voimaa.

Mediassa ja tutkimusjulkaisuissakin vähälle huomiolle jäävät menetelmien soveltamisen työläät vaiheet. Jo pelkästään aineistoon kiinnipääseminen voi olla kuukausien neuvottelujen tulos. Menetelmäosaaja pohtii teoreettisia viitekehyksiä, tekee lukuisia päätöksiä koskien aineiston esikäsittelyä. Hän valitsee työkaluja, joilla aineistoa lähestyy. Tästä näkökulmasta jopa algoritmia voi ajatella poliittisena tekona. Työkaluilla pyritään juuri tiettyihin päämääriin. Vähemmän hehkutusta sisältävien tarinoiden avulla voi avata työskentelyn vaiheita ja päätöksiä, jotka on tehtävä ennen kuin tilastollisten työkalujen soveltaminen esimerkiksi sosiaalisen median aineistoon on mahdollista.

Tutkimuskirjallisuudessa erotellaan pyrkimys tuntea yhä paremmin ihmiset ja ihmisjoukot: data-analyysi keskittyy profiloimiseen ja ihmisten tekemisten jäljittämiseen. Toinen pyrkimys edistää sitä, että kansalaiset voisivat käyttää heistä kerättyä tietoa siihen, että he tuntisivat itsensä ja pyrkimyksensä entistä paremmin. Tarinoita datamaailmasta voi järjestää tämän kahtiajaon avulla. Ihmiset toimivat usein varsin kaavamaisesti ja data-analyysi pystyy näyttämään tämän kaavamaisuuden: ihmisten sosiaalisten verkostojen, ostokäyttäytymisen tai päivittäisen liikkumisen pysyvyyden. Kun ihmisiä luokitellaan, heitä voidaan lähestyä oman luokkansa edustajana: yritysmaailmassa tämä luo edellytyksiä esimerkiksi dynaamiseksi tai aktiiviseksi kutsutulle

hinnoittelulle. Maksukykyiseksi tunnistetulle asiakkaalle myydään sama tuote kalliimmalla kuin pienituloisemmalle

Ihmisten profilointi, flunssa-aaltojen mallinnus tai tunteiden ohjailu sosiaalisessa mediassa ovat ilmiöitä, jotka kertovat mitä data-analyysiltä halutaan. Sen toivotaan auttavan ennakoimaan tulevaa, tekemään maailmasta ennustettavampi ja hallittavampi. Yhtä tärkeää olisi ymmärtää maailmaa niiden asioiden näkökulmasta, jotka eivät toteuta tiettyä kaavaa tai poikkeavat valtavirrasta. Data-analyysi voi olla myötäsukaista vallitsevalle poliittiselle tai taloudelliselle järjestykselle tai sen avulla voi pyrkiä löytämään uusia tapoja kertoa maailmasta. Uudenlaisten aineistojen yhdistely voi olla ravistelevaa tai uteliaisuusvetoista. Se voi tuottaa yllättäviä näkökulmia ja uudenlaista käsitystä tiedosta. Analyysin voi kohdentaa tavalla, joka pyrkii tuomaan esille omituisuuksia ja poikkeamia, joita ei muuten haluta nähdä. Avoimen datan hankkeilla voi tietoisesti pyrkiä kansalaisten etujen vaalimiseen tai demokraattisemman maailman edistämiseen. Tavoitteena voi olla esimerkiksi sen ymmärtäminen, miten ihmisten päivittäistä liikkumista voi kaupunkisuunnittelun avulla tukea tavalla, joka edistäisi kaikenlaisen liikkumisen sujuvuutta. Terveyteen liittyvien hankkeiden tavoitteena voisi olla yhdistää tietoa tavalla, joka tukee kansalaista terveydenhuollon asiakkaana. Terveyskeskuskäyntejä voi esimerkiksi tarkastella viikonpäivittäin. Hakeudutaanko tiettynä päivänä tai ajankohtana hoitoon tapaturmien tai epämääräisten kipujen takia? Säästöjä pyritään synnyttämään ennakoimalla hoidontarpeita ja tehostamalla viestintää ammattilaisten ja asiakkaiden välillä.

Toimittajien tehtävänä on tuottaa tiedonvälitykseen moninaisuutta. Tutkimusryhmällemme osoitetut haastattelupyynnöt odottavat meiltä vastauksia siihen, mitä vihapuheelle on tapahtunut viimeisen viidentoista vuoden aikana, millä palstoilla kirjoitetaan eniten sääntöjen vastaisia viestejä ja mitä sääntöjen vastaiset viestit käsittelevät. Pelkkään vihapuheeseen ja törkyviesteihin keskittyminen vääristää näkökulmaa Suomi24-keskusteluihin. Suomi24-aineistossa on myös paljon vastapuhetta: keskustelua, joka pyrkii rauhoittamaan kanssakirjoittajia tai ohjaamaan heitä ajattelemaan vierautta ja erilaisuutta myönteisemmin. Minkälaisia rooleja ja työnjakoa sosiaalisen median keskusteluissa ja tunteiden kuljetuksessa on tunnistettavissa? Voimmeko tunnistaa suurista keskustelumassoista työnjaollisia malleja, jotka järjestävät keskusteluavaruutta? Näiden kysymysten selvittäminen on työlästä ja aikaa vievää. Lopputuloksena on kuitenkin huomattavasti moninaisempi kuva siitä, mistä suomalaiset haluavat keskustella, miten he suhtautuvat muihin ja ilmaisevat tunteitaan.

Dataosaajan keskeneräinen maailma

Digitaaliset palvelut ja niiden avulla tuotetut aineistot vaikuttavat lukuisin tavoin kansalaisten ja kuluttajien elämään. Kun matemaatikot ja tietotekniikan osaajat alkavat mallintaa ihmisten käyttäytymistä ja toivovat ohjailevansa sitä, heille olisi tärkeää opettaa osana yliopisto-opintoja tiedontuotantoon liittyviä valta- ja vastuukysymyksiä. Tätä voi tehdä myös edistämällä uudenlaisia työnjaollisia malleja: tutkijat, toimittajat, päättäjät ja kansalaiset voivat tehdä yhteistyötä digitaalisen maailman avaajina ja selittäjinä.

Dataosaajat voivat auttaa toimittajia ja tutkijoita hahmottamaan, mitä data-analyysi vaatii aineistolta ja tekijältään. Samalla piirtyvät esiin data-analyysin rajat. Parhaimmillaan data-analyysi opettaa tiedon suhteellisuutta, aineistojen avulla voi peilata tutkittavan ilmiön eri puolia. Suomi24-aineistoa voi verrata tai rinnastaa esimerkiksi työttömyystilastoihin tai muihin sosiaalisen median aineistoihin. Kiinnostavan vertailukohdan aineistolle tuovat esimerkiksi terveystilastot. Suomi24 aineistolla voi näyttää miltä suomalaisten yleisimmät terveysongelmat, esimerkiksi masennus tai päihdeongelmat, tuntuvat kokemuksellisesti.

Datamaailman hehkuttelevien sankaritarinoiden rinnalle tarvitaan toisenlaisia tarinoita. Muualla maailmalla yliopistoissa on jo siirrytty pohtimaan miten data-analyysia voisi yhdistää laadullisemman otteen kanssa ja Suomessakin nuorempi tutkijapolvi on ottanut ensiaskeleita tähän suuntaan. Työtä voi tehdä pienimuotoisten pilottien tai kokeilujen avulla, jotka tutustuttavat erilaisiin aineistoihin, menetelmiin ja esittelevät tuloksia erilaisille yleisöille. Dataosaajat täytyy saada pohtimaan sitä, mitä he haluavat työllään saavuttaa ja mihin he tähtäävät osaamisellaan. Jos Googlen työntekijät eivät suostu kertomaan tekemisistään, tietoa voi kerätä tutkivan journalismin keinoin. Datamaailma on monin tavoin keskeneräinen ja se kehittyy suuntiin, joilla ei ole juuri mitään tekemistä toistensa kanssa.

Toimittajan on asetuttava vastahankaan ja kysyttävä tyhmiä kysymyksiä: asioiden ei pidä antaa normalisoitua, ilman että niitä yritetään taustoittaa ja ymmärtää. Digitaalisiin aineistoihin liittyvien kehityskulkujen ja aineistoista tehtyjen päätelmien dokumentoiminen on tärkeää, että opimme tunnistamaan millaisia kuvia ne luovat maailmasta, millaisena ne esittävät teknologian mahdin ja ihmisten toiminnan. On tärkeää, ettei valtaa ja vastuuta ulkoisteta algoritmeille. Teknologiat tulee määrätietoisesti ohjata omalle paikalleen, etteivät ne vie meitä mennessään.