Profilointi ja OmaData
– miten kirjoja suositellaan digitaalisessa ympäristössä

Laajan tarjolla olevan aineiston huonona puolena on, että käyttäjälle saattaa iskeä runsaudenpula ja kiinnostavia sisältöjä voi olla vaikeaa löytää. Käyttömukavuutta lisäisi, jos hakutoiminnallisuuden lisäksi tarjottaisiin käyttäjälle aineistoja myös muilla tavoin. Tätä sisältöjen tarjoamista voidaan tehdä esimerkiksi erilaisilla suosittelujärjestelmillä.

Mikäli käyttäjä on täysin anonyymi, eli hänestä ei tiedetä juuri mitään, voidaan hänelle suositella sisältöjä yleisellä tasolla. Tällaisia aineistopohjaisia suosituksia voisivat olla esimerkiksi 10 lainatuinta kirjaa tai tietokirjallisuuden uusimmat julkaisut. Myös kirjastoammattilaisten tekemät kuratoidut suosituslistaukset toimisivat ilman pohjatietoja käyttäjistä.

Jos käyttäjälle halutaan antaa tätä henkilökohtaisempia suosituksia, vaatii se ainakin jossain määrin käyttäjän profilointia. Käyttäjälle räätälöidyt palvelut ja käyttäjän yksityisyyden suoja liittyvät suoraan toisiinsa ja niiden suhdetta täytyy punnita kun suosittelujärjestelmiä rakennetaan. Voidaan miettiä esimerkiksi voiko käyttäjä kieltäytyä suosittelusta ja tietojensa keräämisestä? Mikä määrä kerättyä tietoa on vähintään tarpeen suosittelun tarjoamiseksi? Kuinka paljon käyttäjän henkilötietojen kerääminen parantaa hänelle kohdistettavaa suosittelua? Onko käyttäjällä itsellään roolia suosittelun parantamisessa ja jos, niin millä tavalla?

Vielä keskeneräisen diplomityötutkimukseni haastatteluaineistossa on noussut esiin, että haastatellut arvostavat kirjastoammattilaisten asiantuntemusta aineistojen suosittelijoina ja että lähipiirin, esimerkiksi ystävien tai sosiaalisen median ryhmien kautta tapahtuva vertaissuosittelu on heille tärkeää. E-kirjastopalveluiden suosittelua ei kannatakaan ajatella pelkästään teknisenä ja automaattisena toimenpiteenä, vaan kirjastoammatillisella kuratoinnilla voisi olla merkittäväkin rooli palvelun osana.

Automaattiseen suositteluun liittyy helposti myös eettisiä ongelmia, mikäli suosittelujärjestelmä ei ole läpinäkyvä ja käyttäjälle ei ole selvää, mihin suosittelu perustuu ja mihin tietoja käytetään. Onko suosittelulla kaupallisia intressejä, myykö joku suosittelujärjestelmän toimittaja esimerkiksi sisällön nostoa suosituksiin rahalla, samaan tapaan kuin Google nostaa sponsoroituja mainoksia esiin hakutuloksissa? Suositellaanko juuri julkaistuja sisältöjä enemmän kuin arkiston helmiä? Käykö niin, että lukumahdollisuutemme kaventuvat kaventumistaan, kun innokkaalle fantasiakirjallisuuden lukijalle suositellaan vain lisää fantasiaa? Lähettääkö rautakauppa henkilölle kohdennetun mainoksen hänen lainattuaan remonttikirjallisuutta?

Yksi keskeinen käyttäjäongelma liittyen suositteluihin on, että jokaiseen uuteen palveluun kirjautuessaan käyttäjä joutuu aloittamaan puhtaalta pöydältä. Mitä enemmän palvelussa lukee kirjoja, sitä paremmin palvelu osaa suositella lisää luettavaa. Kuitenkin palvelusta toiseen siirryttäessä tiedot lukemisista eivät nykyisellään siirry mukana, joten tietojen kerryttäminen pitää aloittaa alusta. Samoin jos käytössä on samaan aikaan useampi palvelu, eivät nämä tiedä käyttäjän toiminnasta muualla ja saattavat helposti suositella sellaisia sisältöjä, jotka käyttäjä on juuri toisessa palvelussa lukenut.

Suomesta lähtöisin oleva, nykyään jo kansainvälisesti tunnettu OmaData (MyData) -periaate asettaa ihmisen henkilötiedon hallinnan keskiöön. Se tarkoittaa, että jokaisella henkilöllä, jonka tietoja käsitellään, pitäisi olla oikeuksia omaan tietoonsa ja sen käyttöön. Nykyään jokainen organisaatio ajattelee käyttäjää omana asiakkaanaan, mutta OmaData -periaatteen mukaan organisaatioiden pitäisi ajatella pikemminkin olevansa yksi käyttäjän monista palveluntarjoajista. Tällainen ajattelumalli sopii luontevasti juuri julkisille toimijoille. Esimerkiksi kirjastot ja Yleisradio eivät kilpaile keskenään asiakkaista, vaan voisivat täydentää toistensa palveluita käyttäjän eduksi. OmaData -periaatteita ovat ihmiskeskeinen henkilötiedon hallinta, ihmisten voimaantuminen, datan siirrettävyys ja uudelleenkäyttö, läpinäkyvyys, luotettavuus, yhdistettävyys ja yhteentoimivuus.

Yleistyessään OmaData -malli henkilötietojen hallinnassa voisi ratkaista tarkasti tietosuojaan suhtautuvien kirjastojen ja vastaavien tahojen suosittelu- ja profilointitarpeet. Esimerkiksi käyttäjä voisi kerryttää lukutietojaan itse hallitsemaansa (tai OmaData -operaattorin ylläpitämään) mediaprofiiliin, jolloin suosittelut pysyisivät yhtenäisempinä palveluiden välillä ja käyttäjä voisi itse päättää mitä tietoja jakaa palveluiden kanssa. Tarvittaessa onnistuisi myös profiilin muokkaus. Esimerkiksi käyttäjä, joka pari vuotta sitten luki innokkaasti rikosromaaneja ja haluaisi nyt saada muunlaisia suosituksia, voisi tehdä profiilistaan version ilman dekkaritietoja, jolloin ne eivät vaikuttaisi uusiin suosituksiin.

OmaData -ajattelun yleistyessä voisimme tulevaisuudessa käyttäjinä voimaistua omasta tiedostamme ja käyttää sitä omien suositustemme parantamiseen. Tällä hetkellä Helsingin kaupunki kehittää omia OmaData -kyvykkyyksiään ja vuonna 2020 valtiovarainministeriö on myöntänyt Helsingin, Espoon, Turun ja Oulun kaupungeille yhteensä reilut 2 miljoonaa euroa OmaData-hankkeisiin.

Profilointi lukusuosituksia varten

- miten kirjoja suositellaan digiaikana

Emilia Hjelm

28.4.2021

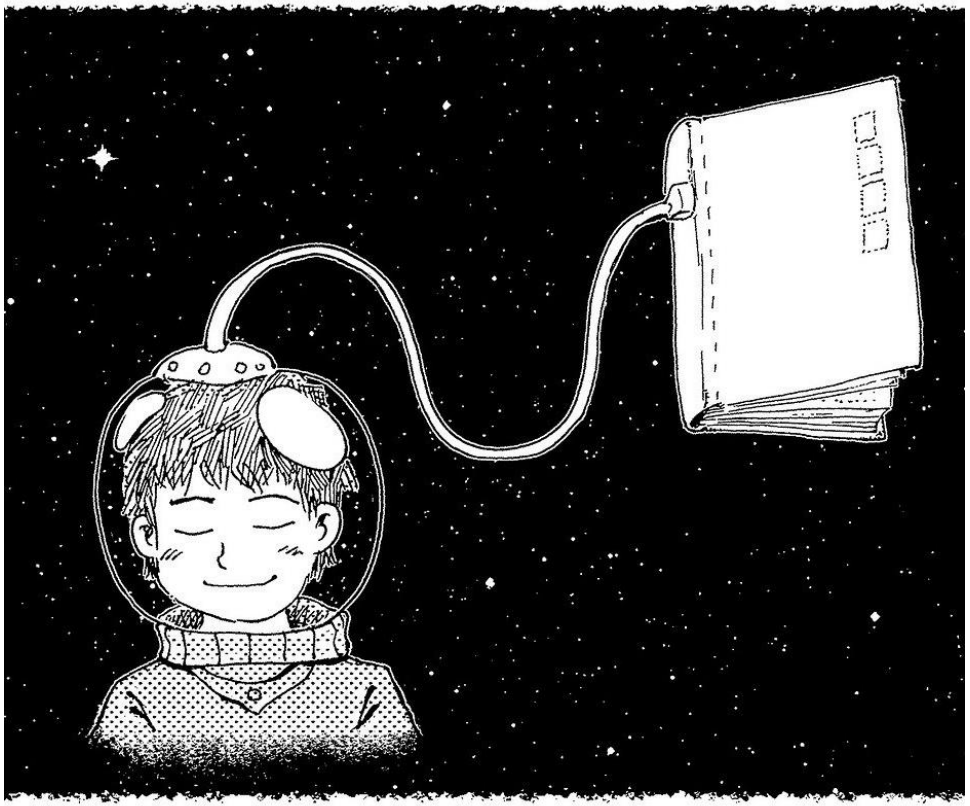
Digimediahanke loppuseminaari

Emilia Hjelm

- **FM (tietojenkäsittelytiede) Helsingin yliopisto**
  - Tiedekasvatus, miten lapsille opetetaan ohjelmointia
- **Aalto -yliopisto**
  - Diplomityötutkimus meneillään profiloinnista ja automaattisista suosituksista lukemisen palveluissa
  - Kirjallisuuskatsaus + haastattelututkimus (N=20)
  - Mahdollinen julkaisu
- **Open Knowledge Finland ry (2016)**
  - MyData 2016 -konferenssi
  - Helsingin kaupungin innovaatiorahaston projekti “Sähköinen asiointi ja henkilötieto)
- **Elisa Oyj 2017 ->**
  - Tekninen tuoteomistaja, erityisfokuksessa tietoturva ja tietosuoja

Esityksen aiheet

1. Suosittelun lyhyt historia
2. Ennustavat ja pyydystävät metriikat
3. MyData



Suosittelun lyhyt historia

- Suosittelujärjestelmät kehitettiin alun perin 1990-luvulla internet-sisältöjen navigointia varten
- Nopeasti kaupalliset yritykset ottivat suosittelut käyttöön lisätäkseen myyntiään
- Aluksi yritykset käyttivät ennustavia metriikoita (predictive metrics) yrittääkseen ymmärtää käyttäjien toiveita
- Muutos ennustavista metriikoista arvioiviin metriikoihin (approximation metrics) tapahtui 2009-2010 kun yritysten liiketoimintamalli muuttui fyysisten tuotteiden toimittamisesta internetin kautta tapahtuvaan jakeluun.

Ennustava malli (Predictive Paradigm)

- Pyrkii arvioiden osuvaan ennustamiseen
- Perustuu käyttäjien antamille arvioille, esim. 0-5 tähteä
  - Explicit ratings
- Sisältöön perustuva suodattaminen
(engl. Content-based filtering (CBF))
  - Mitkä sisällöt muistuttavat toisiaan?
- Yhteistyösuodatus
(Collaborative filtering (CLF))
  - Käyttäjien väliset samanlaisuudet
- Laatumetriikoiden käyttö
  - Neliöllinen keskiarvo (RMSE root mean square error).
Käytetään suosittelun osuvuuden arviointiin. Ensin lasketaan ennustus siitä millaisen arvion käyttäjä antaa ja jälkikäteen verrataan ennustetta lopulliseen arvioon. Lasketaan populaatiosta, ei yksittäiselle käyttäjälle.

Arvioiden läpi katsottuna käyttäjä on mieltymystensä muovaama epätarkka muotokuva

Pyydystävät metriikat

Pyydystävät metriikat (engl. Captivation metrics) eivät pyri ennustamaan arvioita vaan mittaamaan miten hyvin järjestelmä saa ylläpidettyä käyttäjänsä mielenkiinnon.

- Implisiittinen tekniikka
- Kaikki käyttäjän vuorovaikutukset järjestelmän kanssa tallennetaan
- Teot nähdään totuudenmukaisempina kuin sanat
  - Toistuva käyttö == tykkääminen
  - Palvelun käytön jatkaminen == tyytyväisyys palveluun
- Metriikat ovat arvioita

Tallennetun käytöksensä läpi katsottuna käyttäjä on haamu, joka jättää jälkiä ajan myötä

Muutama haaste

suositteluissa ja profiloinnissa

Haasteita

Puolueettomuuden vale
- algoritmit ja suosittelujärjestelmät usein esitetään koneellisina ja puolueettomina

Neuroverkkojen käyttö
- edes suunnittelijat itse eivät välttämättä osaa sanoa mihin suosittelut / luokittelut perustuvat

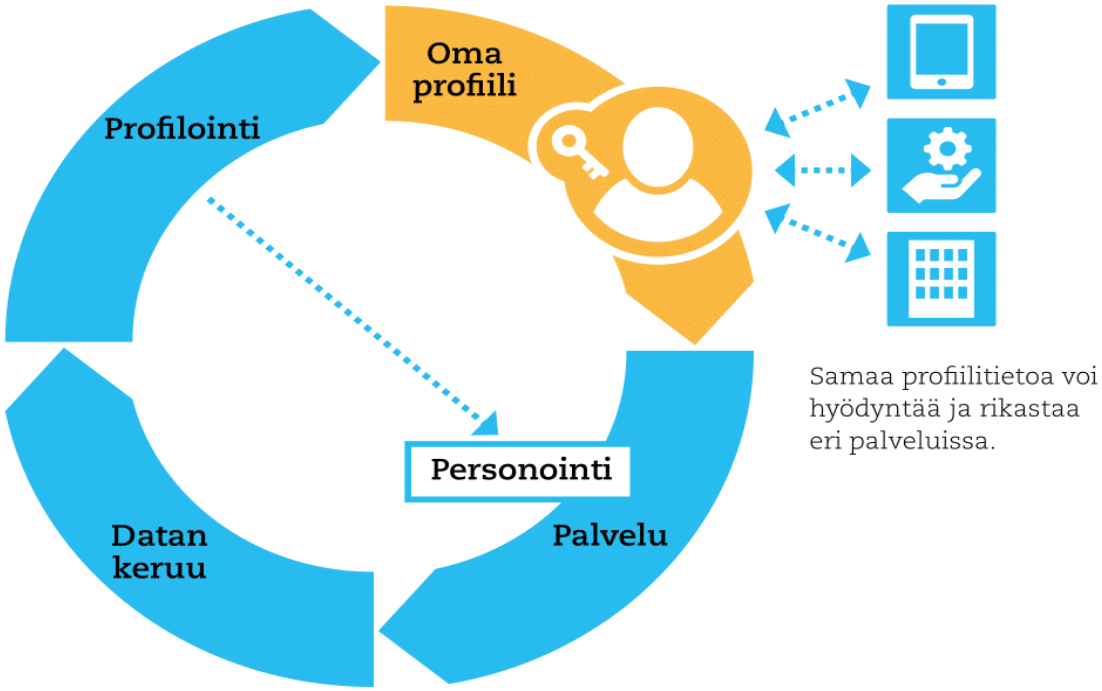
Läpinäkyvyyden puute
- mahdotonta erottaa “luonnollista” suositusta sponsoroidusta sisällöstä
- mahdotonta korjata suosittelua tai antaa palautetta

Käytöksen arviointiin perustuvat metriikat yksinkertaistavat liikaa
- ihminen on liian monimutkainen ymmärrettäväksi pelkästään tykkäysten perusteella

Aiempiin lukemisiin perusteleva **suosittelu voi käydä koko ajan kapeammaksi ja kapeammaksi** (jolloin “lukumahdollisuudet” kaventuvat)

Suosittelu tapahtuu aina yhden palvelun sisällä
- pitkäaikaiset käyttäjät nauttivat osuvammista suosituksista, uudet aloittavat tyhjästä
- tekee palvelun vaihtamisesta vaikeaa

MyData mediaprofiili



Lähde: **MyData – johdatus ihmiskeskeiseen henkilötiedon hyödyntämiseen**

MyData tavoitteet: Minkä tulee muuttua?

Muodollisista käytännöllisiin oikeuksiin

Tavoitteena on, että pääsy omiin tietoihin, tietojen oikaiseminen ja siirrettävyys, sekä oikeus tulla unohdetuksi kehittyvät “yhden klikkauksen oikeuksiksi”, jotka ovat yhtä yksinkertaisia ja tehokkaita käyttää kuin tämän päivän ja huomisen parhaat verkkopalvelut.

Tietosuojasta tiedolla voimaantumiseen

Tietosuoja-sääntely ja yritysten yksityisyyskäytännöt on suunniteltu suojaamaan ihmisiä, etteivät organisaatiot väärinkäyttäisi heidän henkilötietojaan. Tämä on tärkeää tulevaisuudessakin ja lisäksi yleisiä toimintatapoja tulee muuttaa suuntaan, jossa yksilöitä sekä suojellaan, mutta myös voimaannutetaan käyttämään dataa, jota organisaatioilla on heistä.

Suljetuista avoimiin ekosysteemeihin

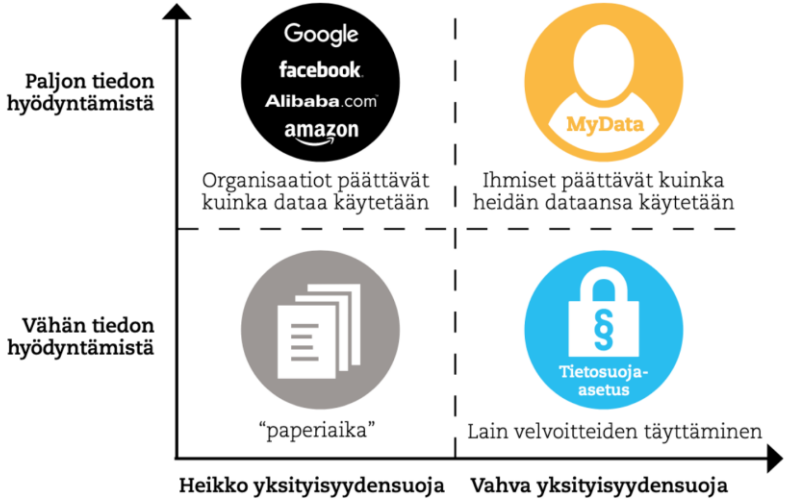
Tämän päivän datatalous tuottaa verkostoefektejä, jotka hyödyntävät alustatoimijoita, joilla on mahdollisuus kerätä ja käsitellä suuria määriä henkilötietoja. Antamalla yksilöiden määrätä mitä heidän datalleen tapahtuu, pyrimme luomaan todellista datan vapaata liikkuvuutta, tasapainoa, oikeudenmukaisuutta, monipuolisuutta ja kilpailua digitaaliseen talouteen.

MyData-periaatteet

Ihmiskeskeinen henkilötiedon hallinta	**Datan siirrettävyys ja uudelleenkäyttö**
Ihminen oman datansa yhdistäjänä	**Läpinäkyvyys ja luotettavuus**
Ihmisten voimaantuminen	**Yhteentoimivuus**

https://mydata.org/declaration

MyData - ajattelun perusteita



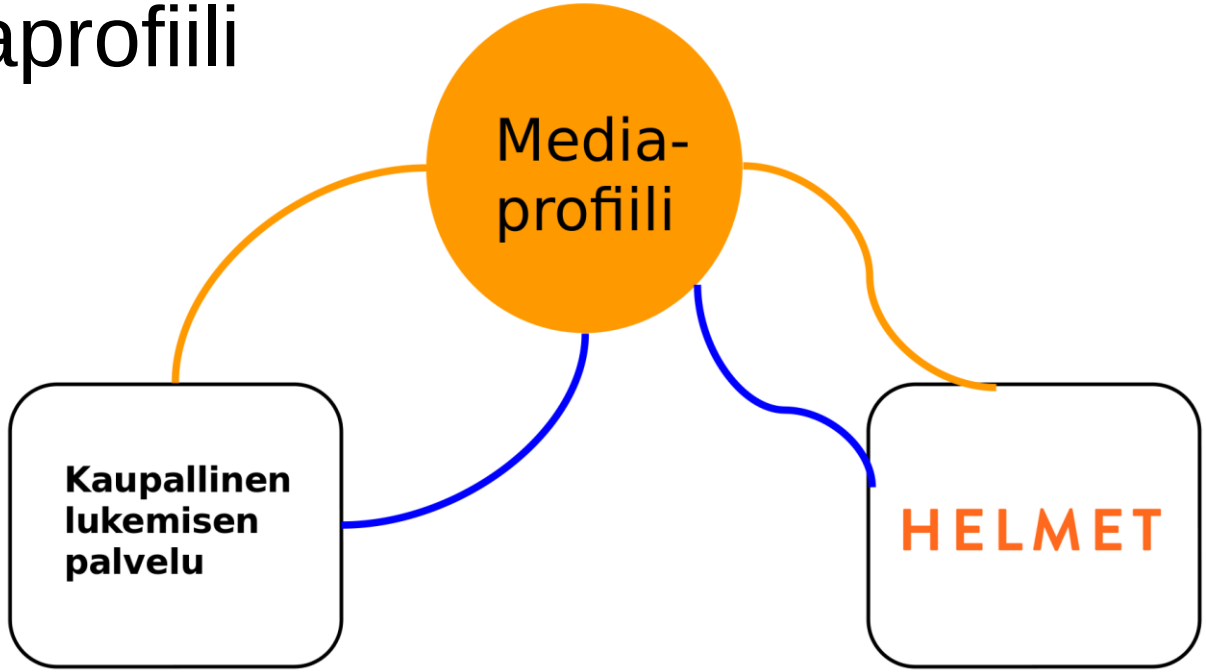
MyData -ajattelun perusteita

Ihmisellä on oikeus:

- **tietää**, mitä henkilötietoa hänestä on
- **nähdä** itseään koskeva henkilötietosisältö
- **oikaista** väärät henkilötiedot
- **valvoa** kuka hänen henkilötietoaan käsittelee
- **saada omat tietonsa**
- **siirtää omat tietonsa** eri toimijoille
- **poistaa** omat tiedot

MyData-malli vertailussa

Mediaprofiili

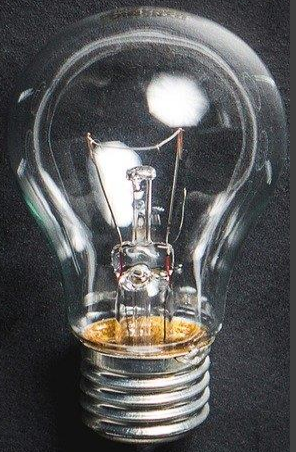


Oranssi luku
Sininen kirjoitus

MyData -tyyppinen mediaprofiili

TARVITAAN:

- Suostumusarkkitehtuuri (MyData -operaattori)
- Jaetut semantiikat ja käännökset (translations) yhdistettävien palveluiden välillä



RATKAISEE:

- Läpinäkyvyys siihen mihin suositukset perustuvat
- Käyttäjä voi muokata profiiliaan tai käyttää eri profiileja eri palveluissa
- Toimittajaloukku: käyttäjän on helpompaa vaihtaa palvelua tai käyttää montaa palvelua samanaikaisesti
- Tasa-arvoisempi asema käyttäjälle suhteessa palveluntarjoajaan
- Vähemmän fokusta arvailussa

Kiitokset!

emilia.hjelm@iki.fi

References

Boyd, Danah, and Alice E. Marwick. "Social privacy in networked publics: Teens' attitudes, practices, and strategies." (2011).

Brunton, Finn, and Helen Nissenbaum. "Political and ethical perspectives on data obfuscation." Privacy, due process and the computational turn: The philosophy of law meets the philosophy of technology (2013): 164-188.

Gillespie, Tarleton. "The relevance of algorithms." Media technologies: Essays on communication, materiality, and society 167 (2014): 167.

Angwin, Julia. Dragnet nation: A quest for privacy, security, and freedom in a world of relentless surveillance. Macmillan, 2014.

Matt, Christian, et al. "Escaping from the filter bubble? The effects of novelty and serendipity on users' evaluations of online recommendations." (2014).

Ekstrand, Michael D., and Martijn C. Willemsen. "Behaviorism is not enough: better recommendations through listening to users." Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 2016.

Poikola, Antti, Hjelm, Emilia, Schildt, Daniel. "Sähköinen asiointi ja henkilötieto" https://www.okf.fi/projects/mydata-helsinki/ 2017.

Seaver, Nick. "Captivating algorithms: Recommender systems as traps." Journal of Material Culture (2018): 1359183518820366.

Poikola, Antti, et al. "MyData-johdatus ihmiskeskeiseen henkilötiedon hyödyntämiseen." (2018).

West, Sarah Myers. "Data capitalism: Redefining the logics of surveillance and privacy." Business & society 58.1 (2019): 20-41.