
Big data ja tiedonlouhinta

Veikko Eranti & Tuukka Ylä-Anttila, 4.12.2017



UNIVERSITY
OF TAMPERE

Luennoitsijat

Tuukka Ylä-Anttila (VTT): Väittelin lokakuussa 2017 HY:llä populismista politiikanteon tapana, nyt Tampereen yliopiston tutkijatohtori ja tutkin populismia, politiikanvastaisuutta ja osallistumattomuutta; käytän ja kehitän menetelmällistä suuntausta jossa yhdistyy laskennallinen tekstianalyysi ja sosiologinen tulkinta.

Veikko Eranti (VTT, FM): Väittelin Helsingin yliopistosta 2016, väitöskirjan otsikkona *Individuals Doing Politics*. Tutkijatohtori, Tampereen yliopisto. Tutkimusaiheina sosiaalinen media, teoria, demokratia, osallistuminen...

Sisältö

'Big Data' ja digitalisaatio

Laskennallinen yhteiskuntatiede (*Computational Social Science*)

'Kvalitatiiviset' ja 'kvantitatiiviset' asenteet

Verkostoanalyysi

Koneoppiminen

Tiedonlouhinta

Empiirisiä esimerkkejä meidän ja muiden tutkimuksesta

Kriittisiä huomioita

Mitä on "Big Data"?

- Sekä yhteiskunnallinen että yhteiskuntatieteellinen ilmiö
 - Entistä suurempi osa sosiaalisesta elämästä tapahtuu digitaalisilla alustoilla, mikä itsessään vaikuttaa yhteiskuntaan
 - Samalla tietoa siitä kertyy enemmän kuin ennen
 - Samaan aikaan menetelmät sen analysoimiseksi ovat kehittyneet
 - Data ei ole tietoa: kun (big) dataa käytetään tutkimuksessa, tavoitteena on jalostaa siitä tietoa
 - "Big data" -käsite sisältää siis useita ajatuksia:
 - Datan lisääntyminen
 - Datan kerääminen (ja menetelmät: scraping, API, Open Data)
 - Datan analysoiminen (menetelmät: tiedonlouhinta, mallintaminen)
 - 'Datafikaation' uhat ja mahdollisuudet
 - 1) yhteiskunnalle
 - 2) yhteiskuntatieteille
-

Kolme V:tä

*“Big data is **high-volume, high-velocity and high-variety** information assets that **demand cost-effective, innovative forms of information processing** for enhanced insight and decision making.” (Gartner 2014)*

= paljon erityyppistä dataa, jonka määrä lisääntyy koko ajan.
Olennaista on, että jollakulla on pääsy dataan ja että sitä voidaan analysoida

Datan määrä ja saatavuus

“Teknologian yhä kasvava integroituminen jokapäiväiseen elämäämme tuottaa jatkuvasti lisääntyvää määrää dataa yhteiskunnan kaikkien osa-alueiden arkipäiväisestä toiminnasta. Tämä datan määrä avaa uusia mahdollisuuksia ymmärtää monimutkaisia ihmisjärjestelmiä”.

Conte, R. 2012. Manifesto of Computational Social Science. The European Physical Journal Special Topics. November 2012: Vol. 214, Issue 1, pp. 325-346.

Suurin osa datasta täysin merkityksetöntä hölinää. Ehkä.

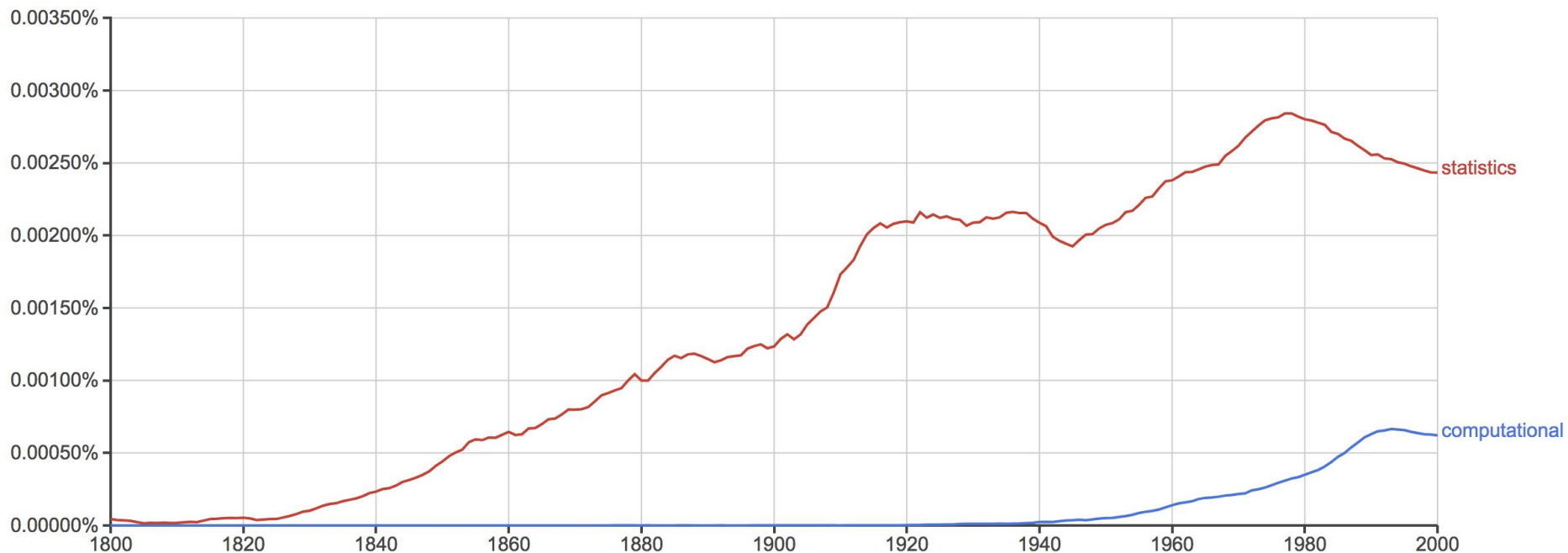
Data ja miten sitä käytetään

- Pelkkä jossain olemassa oleva ”iso datamäärä” ei lohduta ketään ollenkaan
- Koko big datan käsite edellyttää, että tätä kaikkea tietomäärää myös hyödynnetään jotenkin
- Erilaisilla kaupallisilla toimijoilla käytössään uskomattoman paljon dataa ja mahdollisuuksia

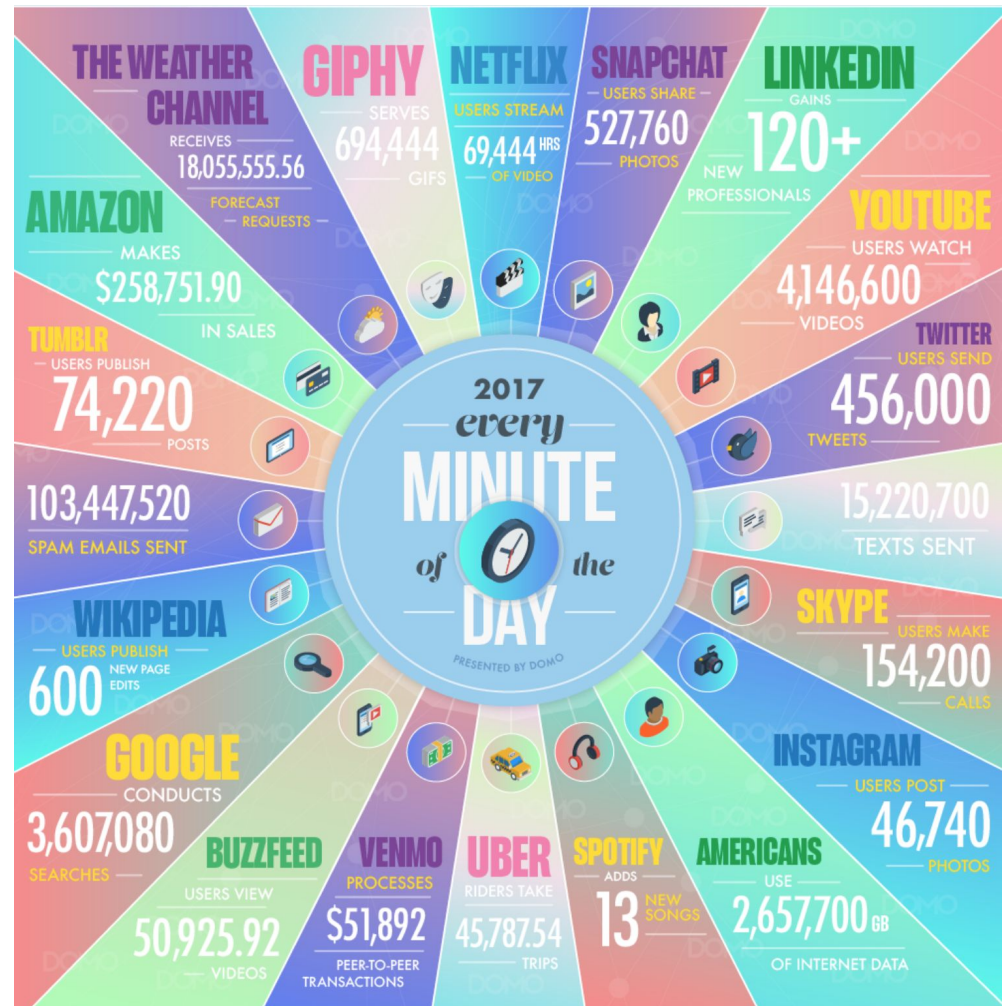


Digitalisaatio ja datan määrä

- Yhä enemmän sensoreita joka puolella
- Yhä enemmän vanhempaakin kirjallista materiaalia digitoituna

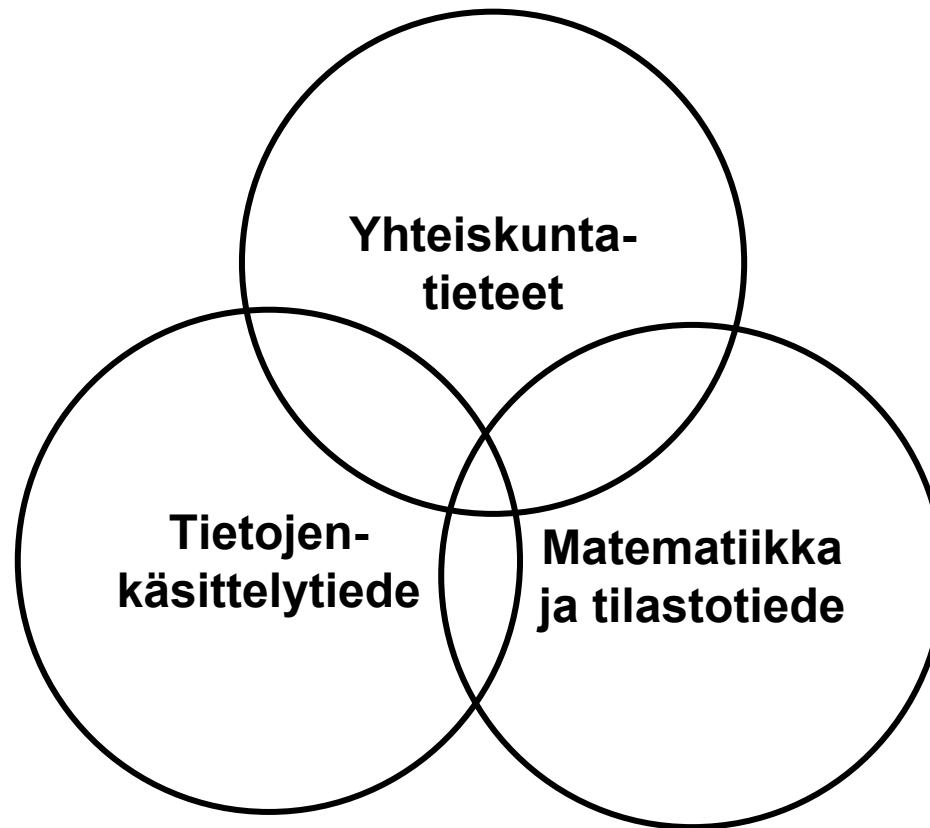


Paljonko "dataa" syntyy joka hetki?

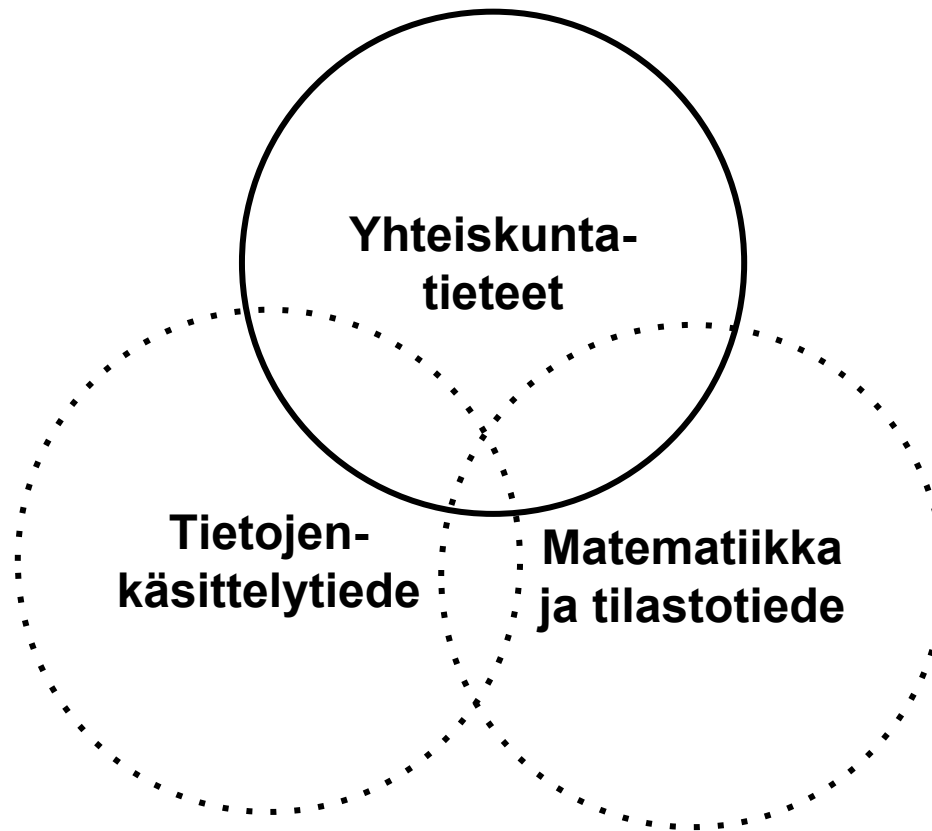


<https://www.domo.com/learn/data-never-sleeps-5>

Monitieteinen laskennallinen yhteiskuntatiede



Monitieteinen laskennallinen yhteiskuntatiede



Eri lähestymistapoja CSS:ään

- Vanhoja asioita uudella tavalla
 - Esim. tekstin automaattinen luokittelu algoritmeilla, mutta samoihin luokkiin kuin ennen tehtiin käsin; tai kvantitatiiviset analyysit uusista aineistoista
 - Tyypillisesti yhteiskuntatieteilijät, opettelevat uusia menetelmiä
 - Uusien menetelmien mahdollistamia uusia asioita
 - Esim. simulaatiomallit, verkostanalyysi
 - Yhteiskuntatieteilijöitä, mutta joskus myös tietojenkäsittelytieteilijöitä, jotka alkavat tehdä omilla menetelmillään ”yhteiskuntatiedettä”, kritiikkinä ks. esim. Grimmer 2015: ”We Are All Social Scientists Now”
 - Koska lähestymistavat eri tieteistä tulevilla erilaiset, voi olla hyödyllisempää että samat ihmiset osaavat eri asioita kuin vain kohtauttaa eri alojen ihmisiä
-

Mitä halutaan tehdä?

- Selittää?
 - Kausaaliset väitteet, *miksi* jokin tapahtuu, A aiheuttaa B:n – tyypillisempi luonnontieteille, myös “kvantitatiiviselle” sosiaalitieteelle
 - Tulkita?
 - Ymmärtää, *miten* jokin tapahtuu: esim. ihmisten kokemukset, merkitykset asioista – tyypillisempi (kvalitatiivisesta) sosiaalitieteistä ponnistaville
 - Raja hämärtymässä, koska uudet menetelmät lupaavat esim. *tulkintojen mittaamista* tai jopa sen selittämistä, *miksi* jokin tulkitaan jollain tavalla (esim. Bail: Terrified)
-

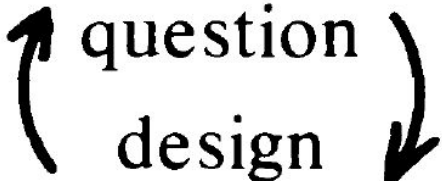
Exploratory vs. confirmatory analysis

- Tilastotieteessä käytetty erottelu
- *Confirmatory*: klassinen tiede, hypoteesien testaus

(*) question → design → collection →

analysis → answer

- *Exploratory*: hypoteesit eivät synny tyhjästä, emme voi tietää mikä on kiinnostavaa ennen tutkimusta

(*) idea →  question
design → collection →

analysis → answer

Koneoppiminen (*machine learning*)

- Tietokoneohjelmia, jotka lukevat jotain dataa ja muuttavat käyttäytymistään sen perusteella
 - Esim. kännykän näppis, joka oppii ehdottamaan sanoja
 - Suuri osa tiedonlouhinnasta (*data mining*) perustuu koneoppimiseen: ohjelma oppii aineiston rakenteen
 - Ohjaamaton (*unsupervised*) koneoppiminen
 - Ohjelma oppii täysin ilman ihmisen inputtia
 - Esim. luokittelee läjän kuvia eri kategorioihin
 - Ohjattu (*supervised*) koneoppiminen
 - Ihminen kertoo koneelle, miten pitää luokitella
 - Esim. ihminen antaa koneelle läjän kuvia koirista ja toisen läjän kissoista, kone oppii “miltä koira näyttää” ja “miltä kissa näyttää”, tämän jälkeen osaa luokitella “kissoja” ja “koiria”
-

Tiedonlouhinta (*data mining*), 1/2

- Tiedonlouhinta mahdollistaa ison datan *käyttämisen*
 - Menetelmät, joilla voidaan automatisoidusti löytää valtavista data-aineistoista jotain mielenkiintoista
 - Etsii datan kohinasta *rakennetta*, jotain joka toistuu, yhteyksiä asioiden välillä
 - Tavoitteena siis louhia *datasta tietoa* (jotain joka merkitsee jotain)
 - Suomennos “tiedonlouhinta” kuvaa ajatusta paremmin kuin “data mining” (‘datanlouhinta’), koska kyseessä ei ole *datan* kerääminen vaan *tiedon* jalostaminen *datasta*
-

Tiedonlouhinta (*data mining*), 2/2

Induktiiviset vs. deduktiiviset tutkimusasetelmat:

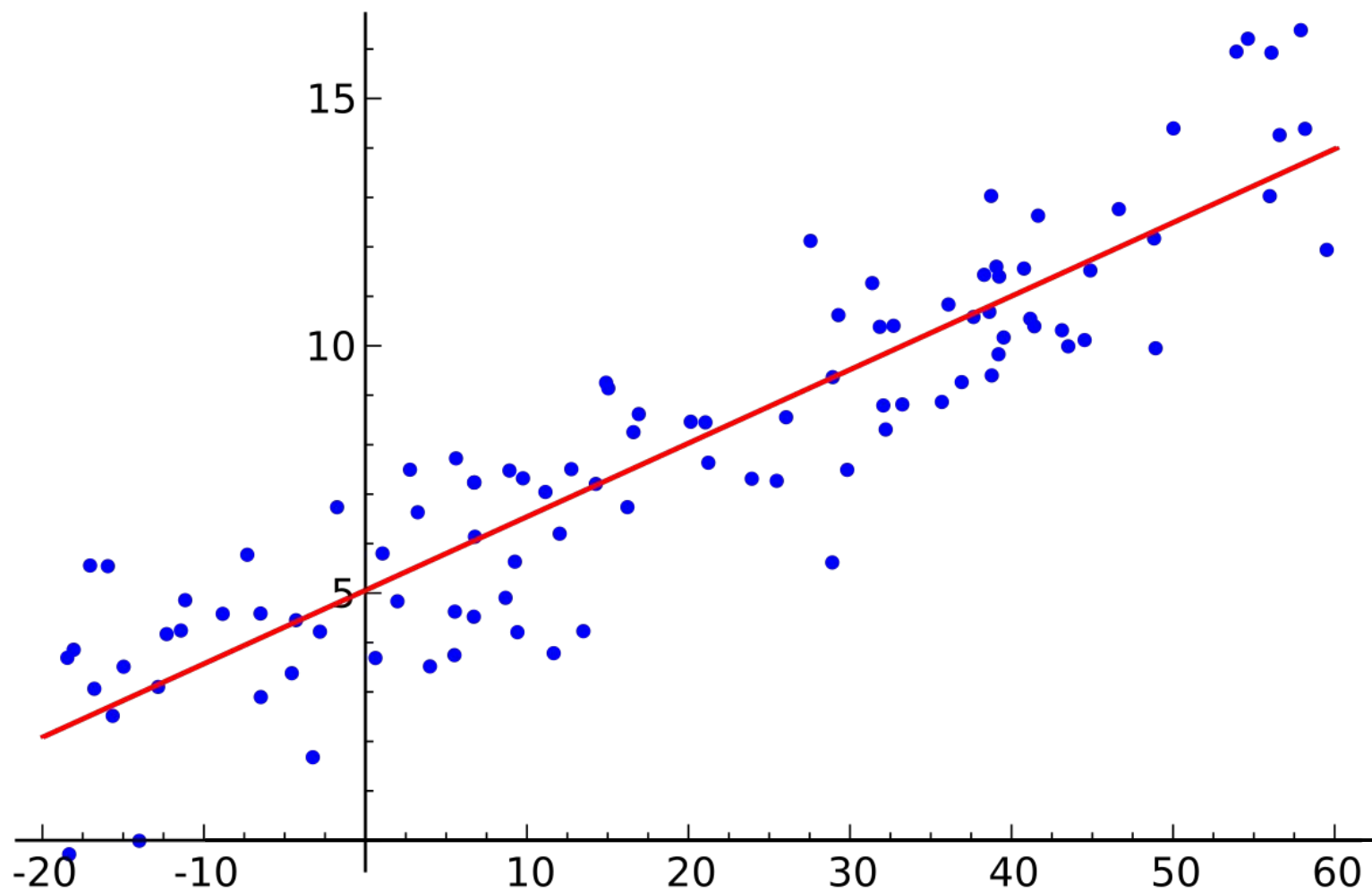
- Induktiivinen: data on ensin, siitä muodostetaan teoria
- Deduktiivinen: teoria on ensin, sitä testataan datalla
- Tiedonlouhinta yleensä **induktiivista**

Pelkkä “tiedonlouhinta” (algoritmit) harvoin riittää tiedon muodostamiseksi, niiden outputteja pitää myös osata *tulkita* ja asettaa kontekstiin: mikä on tärkeää?

Teorian rooli: ohjaa kysymään oikeita kysymyksiä, ohjaa tulkitsemaan saatuja tuloksia ja niiden merkitystä

”Täysi induktiivisuus” on siis täyttä utopiaa

Tiedonlouhinta käyttää koneoppimista



Etusivu

Ilmoitukset

Vii



Big Data Borat

@BigDataBorat

Learnings of Big Data for Make Nation of
Kazakhstan #1 Leading Data Scientist Nation



Big Data Borat

@BigDataBorat

Seuraa



Data Science is statistics on a Mac.

Käännä kielestä englanti

16.32 - 27. elokuuta 2013

626 uudelleentwiittausta 322 tykkäystä



20

626

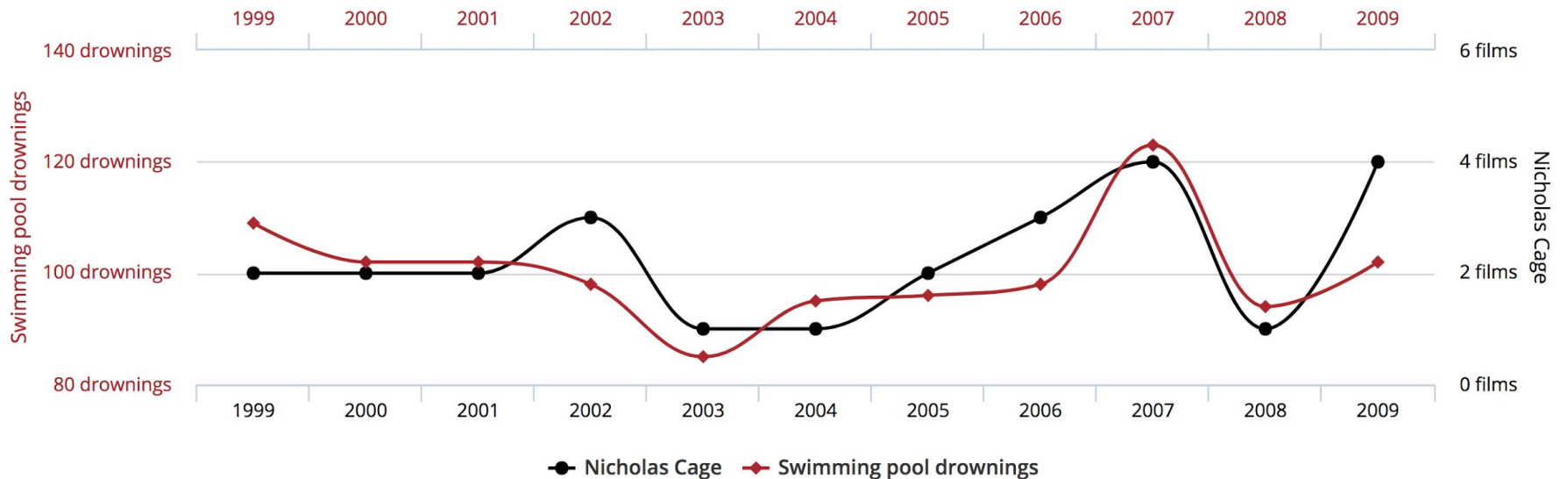
322



Twiiittaa vastauksesi

Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in

Correlation: 66.6% ($r=0.666004$)



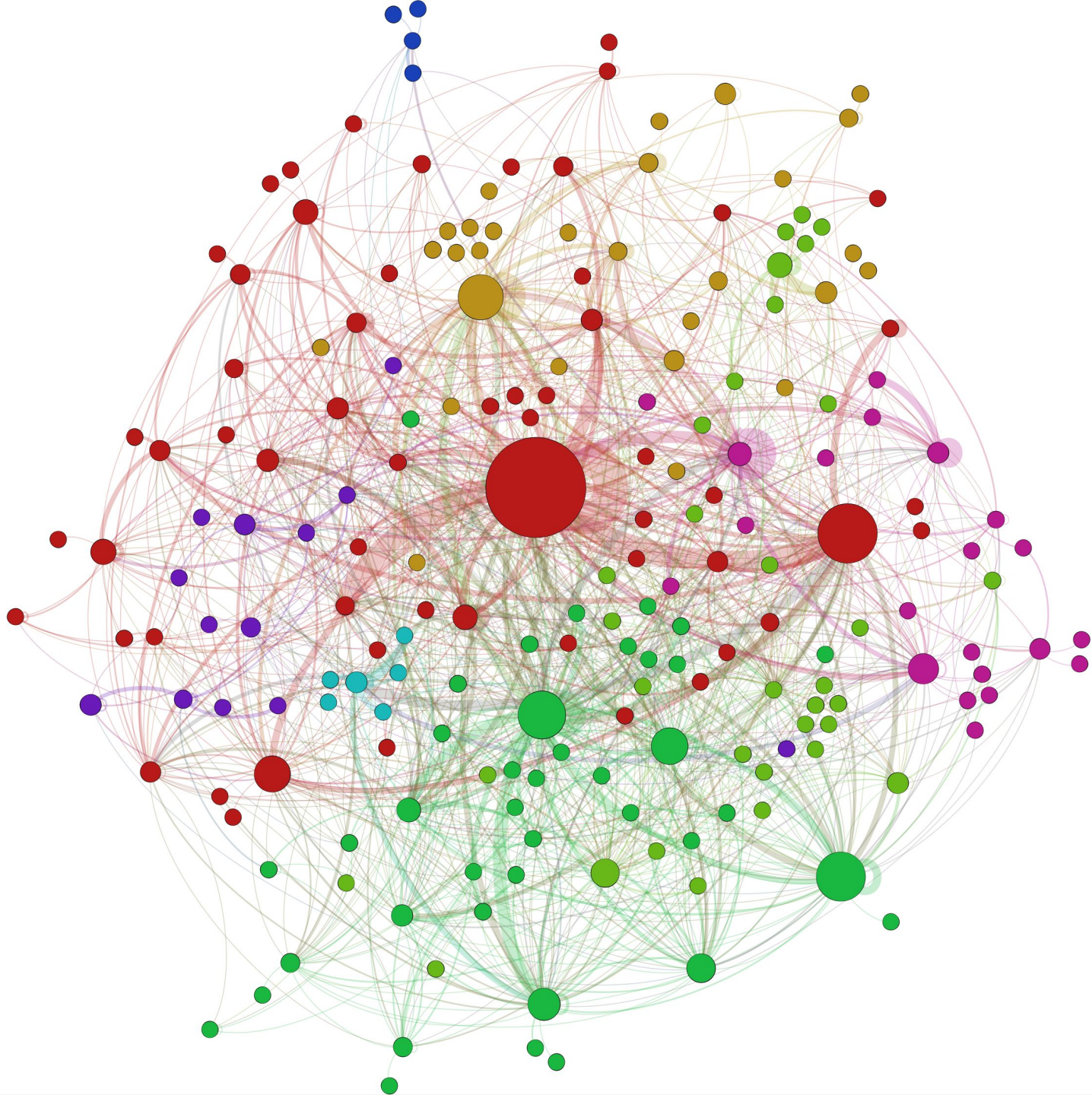
tylervigen.com

Data sources: Centers for Disease Control & Prevention and Internet Movie Database

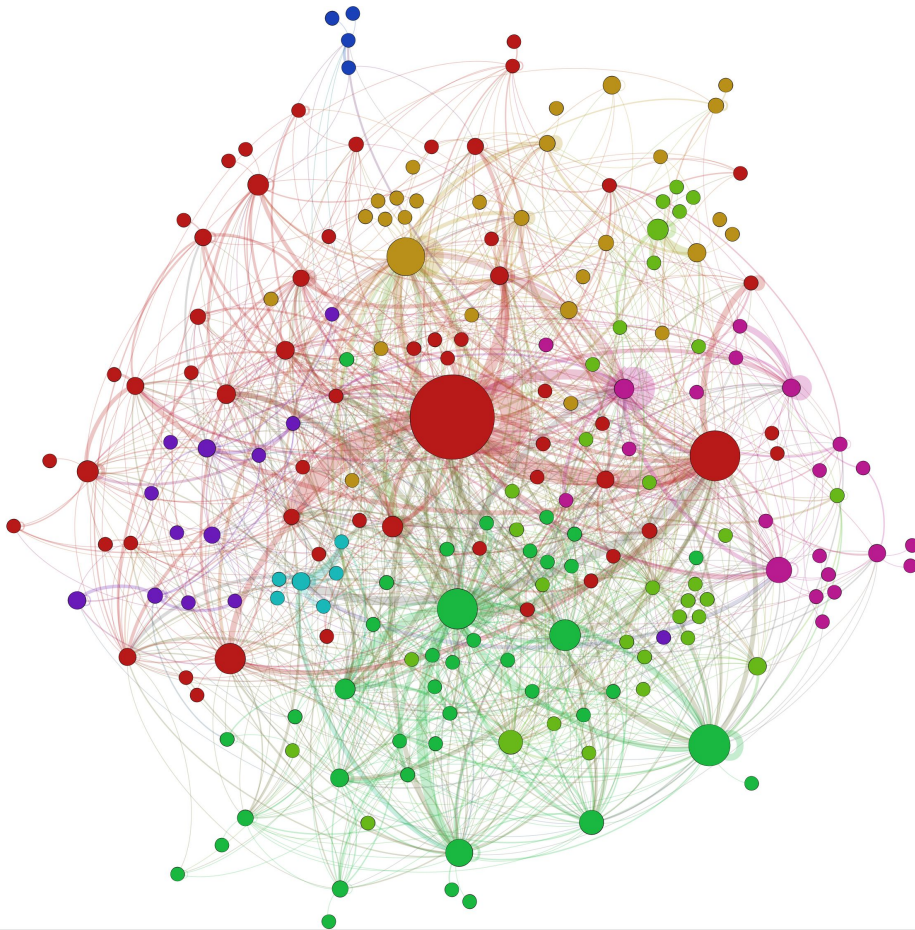
<http://www.tylervigen.com/spurious-correlations>

Verkostoanalyysi

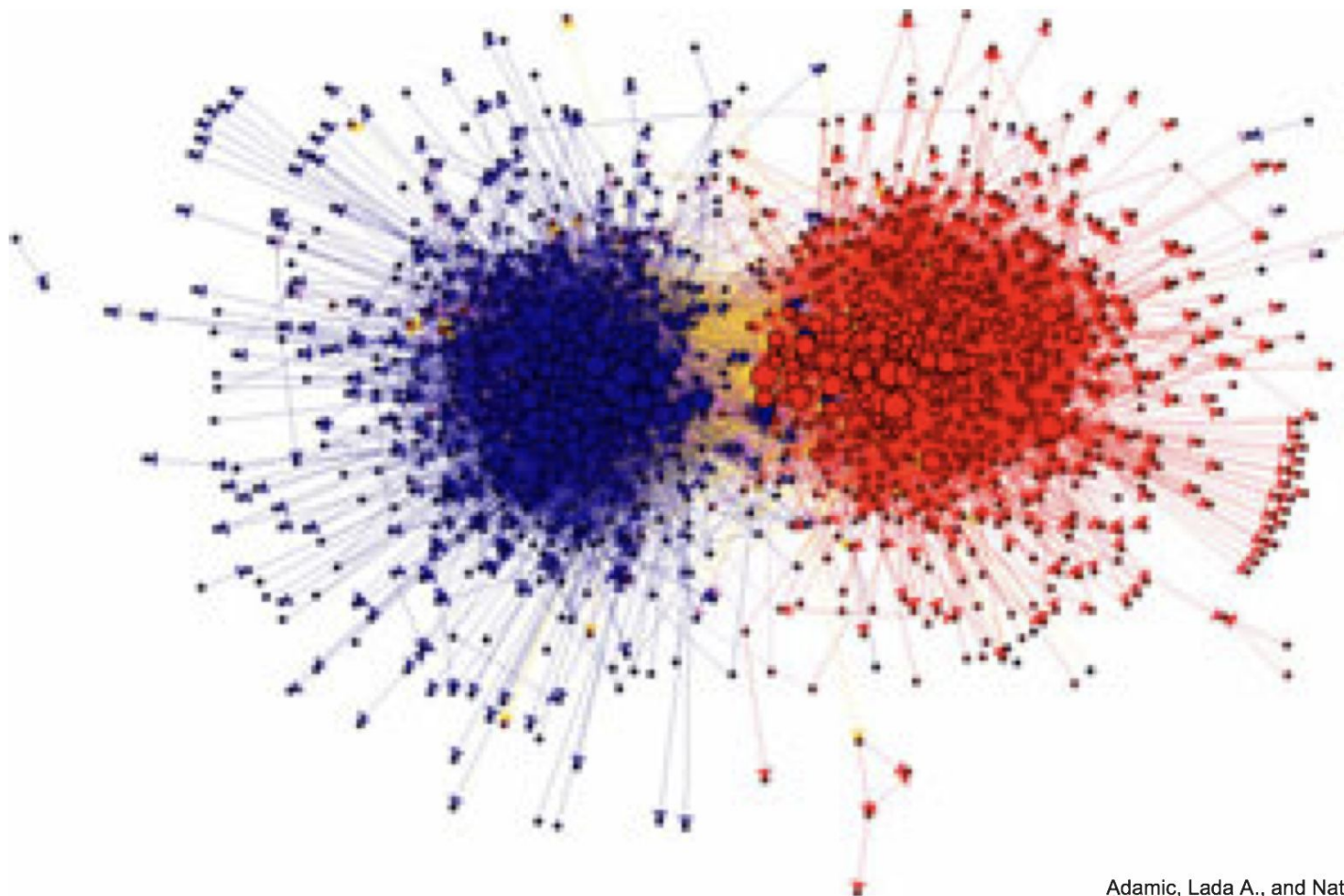
- Tapa tutkia ihmisjoukon sisäistä rakennetta ja ihmisten välisiä yhteyksiä
 - Hyödynnetään matemaattista verkostoanalyysiä ja graafiteoriaa
 - Ihmisryhmät hahmotetaan kokoelmana noodeja (ihmiset) ja linkkejä (tai *edgejä*) jotka kuvaavat heidän välisiään suhteita
 - 2000-luvun isoja menetelmiä, kiitos 9/11 ja Facebook
 - Enemmän perspektiivi ja joukko matematiikan menetelmiä kuin sinänsä yhtenäinen teoriapohja
-



Verkostoanalyysi



- Erään kaupunginosan Facebook-ryhmä
 - Nodet ovat postiaajia, viivat (eli edget) palluroiden välillä ovat interaktiota postiaajien välillä (tykkäys, kommentti jne.)
 - Värit osoittavat ihmisryhmiä jotka ovat eniten keskenään tekemisissä
 - Käytettiin ohjaamaan haastateltavien valintaa tutkimuksessa
-



Adamic, Lada A., and Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog." *Proceedings of the 3rd international workshop on Link discovery*. ACM, 2005.

Mitä luennolla ei käsitellä: kompleksiset järjestelmät & simulaatiot

Laskennallisen yhteiskuntatieteen sisällä on myös monia suuntauksia, jotka eivät ole käsittelyn keskiössä tällä kertaa.

Kompleksiset järjestelmät

Näitä tutkivat insinöörit – systeemidynamiikka jne.

Simulaatiot ja mallinnus

Rakennetaan toimijoille käyttäytymissääntöjä ja mallinnetaan niiden pohjalta järjestelmiä.

Yhteiskunnan mittaamista tekstiaineistoista

- Viime vuosien suuria edistysaskeleita on, että laskennallisia menetelmiä voidaan käyttää myös tekstimassojen tutkimukseen
 - Monet kysymyksenasettelut, jotka aikaisemmin nojasivat siihen että tutkijat lukevat nipun tekstiä, voidaan nykyään toteuttaa laskennallisesti isoilla tekstimassoilla
 - Tähän on päästy etenkin kielitieteen ja tietojenkäsittelytieteen yhteistyöllä
 - Hyödyttää sellaista yhteiskuntatiedettä, joka usein tutkii *kielivälitteisiä* asioita (kuten kulttuuria, politiikkaa)
 - Tällöin voidaan mitata sellaisia asioita, joita aiemmin on vain tulkittu (esim. yhteiskunnallista keskustelua)
-

Miten käytännössä louhitaan?

- Toistaiseksi melkein kaikki tiedonlouhintamenetelmät vaativat ohjelmointikokemusta (Python tai R)
 - Valmiita ohjelmistojakin on ja tulee koko ajan lisää (MALLET, SPSS Modeler, verkosto-/GIS-ohjelmat)
 - Data ja menetelmä kannattaa valita tutkimuskysymyksen mukaan, joka taas valitaan teorian ja aiemman tutkimuskirjallisuuden perusteella
 - Käytännössä ”datahype” johtaa siihen että usein saadaan aineisto tai keksitään menetelmä, sitten mietitään mitä niillä voisi tutkia
 - Tai jopa lähdetään vaan eksploraatiivisesti tutkimaan aineistoja – eikä sekään väärin ole!
-

"Data carpentry" (datakäsityötaito)

- Vaikka "algoritmit automatisoivat kaiken", oikeasti tiedonlouhinnassa paljon tietojen käsittelemistä käsin, pikkujuttujen ohjelmointia
 - Esim. "näissä 100 000 tiedostossa on datapisteiden välillä pilkku eikä rivinvaihto, korjaa"
 - Tai "ääkköset meni rikki"
 - Tai "99% tästä aineistosta on random-huutelua"
 - *Data carpentry*, arkisen datanpyörittelyn taito
 - Datan "prosessointi" ennen sen "analyysiä"
 - Filtteröidään pois ylimääräinen
 - Kun analysoidaan sanojen esiintymistä tekstiaineistoissa, kieliaineistojen käsittelyn erityisvaatimukset
 - Lemmatisointi tai stemmaaminen (sanat perusmuotoon)
-

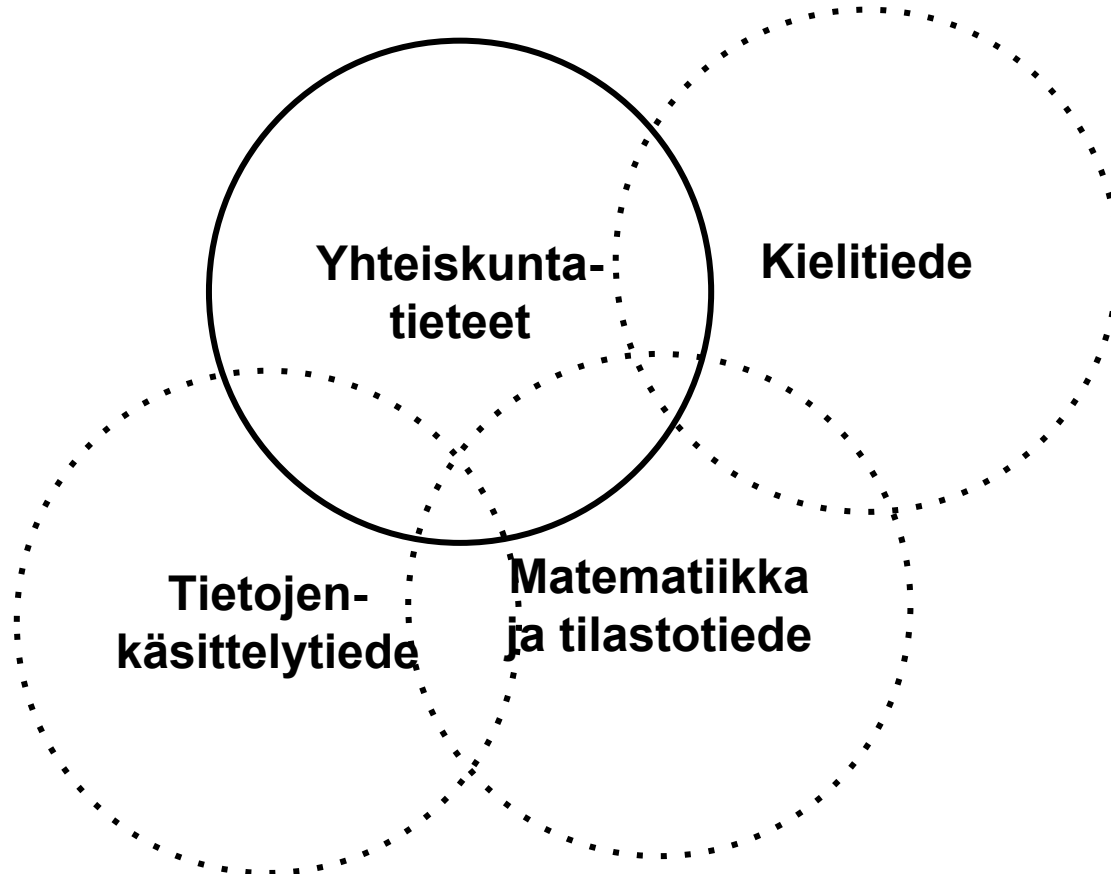
Louhintä vs. käsityö



Kielitieteen rooli

- Kun halutaan käsitellä kielellisiä ilmiöitä, tarvitaan usein taustatukea kielitieteilijöiltä
 - Valmiita korpuksia (esimerkiksi corp.csc.fi)
 - Stemmaus / lemmatisointi
 - Win / Winning / winner -> win
 - Voita? Palaa? Kuusi?
 - Sentiment analysis – positiivisuuden / negatiivisuuden tunnistaminen sanalistojen avulla
 - Sanaluokat - tekeminen / oleminen
 - Stop-words
-

Monitieteinen laskennallinen yhteiskuntatiede



Väliharjoitus!

- Tiedonlouhinnan ja laskennallisen yhteiskuntateiteen perusteita on helppo kokeilla myös itse
- Yksi helppo työkalu on Google Booksin Ngrams-toiminto, jolla voi etsiä termien mainintoja kaikista Googlen digitoimista kirjoista (yli 25 miljoonaa)
- Kokeillaan! Tutkikaa kolmen toisiinsa liittyvän termien suosiota julkaistuissa kirjoissa. Miettikää mahdollisia selityksiä suosion vaihtelulle.

books.google.com/ngrams

Empiirisiä esimerkkejä

- Seuraavaksi muutama maistiainen meidän omista tutkimuksistamme
 - Tuukka esittelee analyysia siitä, miten Hommafoorumilla ja MV-lehdessä keskustellaan *tiedosta*
 - Veikko ilmastokeskustelua New York Timesissä ja The Hindussa
 - Molemmat perustuvat LDA-menetelmään ja aihemallinnukseen
-

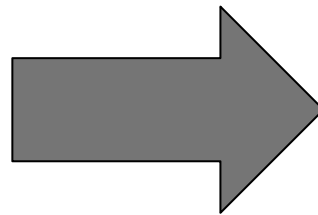
Aihemallinnus

- *Unsupervised* machine learning
- Algoritmi, joka etsii tekstiaineistoista *aiheita* mallintamalla sanojen esiintymistä yhdessä

Dokumentit:

Koirille voi opettaa temppuja.
Koirat haukkuvat ja heiluttavat häntää.

Kissoille ei voi opettaa temppuja.
Kissat kehräävät ja nukkuvat.
Kissoillakin on häntä.



Aiheet:

Koira (2), haukkua (1), heiluttaa (1)

Kissa (3), kehrätä (1), nukkua (1)

Opettaa (2), temppuja (2), häntä (2)

Topic modeling in action

Rubiinkuutio
Jäsen^^^
0000

Vs: Etniset geneettiset intressit
• Vastaus #244 : 09.06.2016, 21:15:02 •

Lainaus käyttäjältä: Brandis - 09.06.2016, 21:11:56

Et

Eikö kaikenlainen valinta tässä suhteessa ole mainitsemaasi rodunjalostusta? Olettaen, että pyrkii (omasta näkövinkkelistään katsottuna) parhaaseen mahdolliseen lopputulokseen kumppanin valinnassa ja sitä kautta jälkeläisten maailmaan tuomisessa.

On. Kaikki tiedotiset pohdinnat oman jälkikasvunsa geneeistä ovat jo tietynlaista rodunjalostusta, jos siinä on pyrkimys johonkin parempaan. Jos hankit lapsen älykkään/vahvan miehen kanssa saataksesi lapselle älykkäitä/vahvoja geneejiä, niin se on jo ensimmäinen rodunjalostusmekaniikka. Tietysti biologisesti rodunjalostusta tekevät tässä mielessä epätietoisesti ihan eläimetkin ja toki ihmisetkin vaihtomaisesti. Mutta varsinaista rodunjalostusta se on minusta vasta, kun tietoisesti pyrkii tavalla tai toisella parempiin geneihin. Eli jos alitajuisesti ihastuu vahvaan/älykkääseen mieheen omien geneiensä takia, niin minusta se ei ole varsinaisesti rodunjalostusta, vaan lähinnä viettien biologiaa.

Lainaus

Et

Kuten sanoit; se on mitä suurimmassa määrin moraalinen mielipide. On hyvä toimia itse kuten tahtois maailman ympärillään toimivan.

Yksi perinteinen ja hyvin laajasti hyväksytty moraalinen mielipide tuokin.

Kirjattu

"Sananvapaus ei ole lupa loukata tai olla eri mieltä."

Brandis
Jäsen^^^
0000

Vs: Etniset geneettiset intressit
• Vastaus #243 : 09.06.2016, 21:11:56 •

Lainaus käyttäjältä: Rubiinkuutio

Et

Mutta siis noin pohjustukseksi esittämälläni väitteille, niin tuohan on moraalinen mielipide ja makuaasia ja jos joku tahtoo jalostaa rotua, niin siitä vaan. Jos joku pitää sitä vääränä, niin siinähan pitää. Evoluutio ei välitä, vaan jatkaa toimintaansa molempien mielipiteiden kanssa.

Eikö kaikenlainen valinta tässä suhteessa ole mainitsemaasi rodunjalostusta? Olettaen, että pyrkii (omasta näkövinkkelistään katsottuna) parhaaseen mahdolliseen lopputulokseen kumppanin valinnassa ja sitä kautta jälkeläisten maailmaan tuomisessa.

Kuten sanoit; se on mitä suurimmassa määrin moraalinen mielipide. On hyvä toimia itse kuten tahtois maailman ympärillään toimivan.

• Viimeksi muokattu: 09.06.2016, 21:13:29 kirjoittanut Brandis •

Kirjattu

Meidän sirkuksemme on määrätty maan ja taivaan välillä. Meaperä antaa sille turvan ja taivas unelmat. Jos toinen niistä katsoo, toistakaan ei voisi olla. "Jalot maassa ja sydän taivaassa" - niinhän sitä ennen sanottiin. ~ Martti Linqvist, *Linnean rokoasta elämästä 1992*, s. 137 kappale 7 ~

törö
Jäsen^^^
0000

Vs: Etniset geneettiset intressit
• Vastaus #193 : 26.02.2013, 18:34:49 •

Lainaus käyttäjältä: matkamiehii - 26.02.2013, 17:30:09

Et

Lainaus

Et

Itselläni on geneejiä sekä idästä että lännestä, joten tulen hyvin juttuun sekä ruotsalaisten että venäläisten kanssa. En minä ole ruotsalaisille mikään "snälla du" niin kuin jonkun varsinaissuomalaisen postinkantajan jälkeläinen suoraan alenevassa povessa, vaikkain paikkoruotsia hyväksyäkään.

Oltiin mielenkiintoista tehdä joskus DNA-testi, isän puolen suvussa vaikuttaisi olevan lieviä mongoli-rotumaisia piirteitä ja äidinpuolen suvussa on jonkin verran ruotsalaista verta. Itse näytän jokoikin stereotyyppiseltä ruotsalaiselta enemmän kuin suomalaiselta.

Suomalaiset on varsin mielenkiintoinen etninen ryhmä noin geneettisessä mielessä.

Suomi oli pitkään periferia, joten tänne valui sopeutumattomia aineksia vähän joka suunnalta. Laikoja ja rikollisia se vapaus ei niinkään viehättänyt, koska täällä oli myös aika köyhää ja suhtautuminen loiseläjiin siitä syystä melko nuivaa, mutta omaperäistä porukkaa tuli Ruotsista, Venäjältä, Preussista jne. ja vähän kauempaakin, koska rajojen ylittäminen oli melkein yhtä helppoa kuin turhapaikanhakijoille nykyisin eikä "asylumiakaan" tarvinnut käydä sanomassa poliisiesdille. Melkein pä riitti kun ei vaikuttanut irtolaiselta.

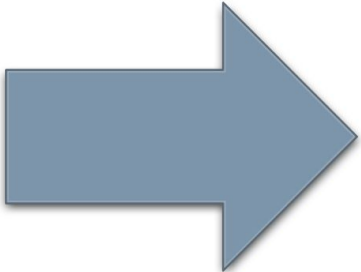
Tasapaksuun elämään näistä pakolaisista ja heidän jälkeläisistään ei oikein ole ollut, joten konsensusta vierastava yleissopeutumaton hompansi voi olettaa olevansa geeniperimältään hyvinkin monikansallinen. Jos pottunokan alkuperää aletaan joskus tutkia kunnolla, en laikaan ylitä jos se paljastuu tuontitavaraksi.

Monikulttuurista kohtaavat konsensushyypiöt ja muut omissa ympyröissään pyörivät elitistit sen sijaan voivat ihan oikeasti olla vieraiden geenien tarpeessa.

Mustalaiset taas ovat tutkitut niin sisäsiiloitettuja, että heidän kannattaisi harkita fuusioitumista vaikka taikaseinänsomalien kanssa.

Kirjattu

Topic modeling in action



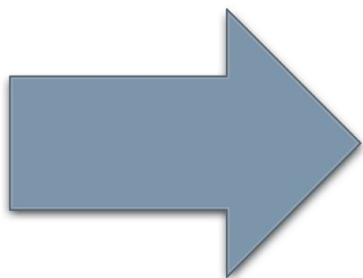
```
1 En
2 lähetä
3 mihinkään
4 ennakkotarkastukseen
5 tms
6 vaan
7 laitan
8 ne
9 tänne
10 Hommafoorumille
11 teidän
12 kommentoitaviksi
13 Sen
14 jälkeen
15 perustava
16 kokous
17 hyväksyy
18 ne
19 Ei
20 ole
21 vielä
22 ollut
23 perustettavaa
24 kokousta
25 Jep
26 sitten
27 vasta
28 PRH
29 een
```

```
1 Kyllä
2 Persut
3 näytti
4 selkeästi
5 persettä
6 maahanmuutokysymyksille
7 eikä
8 ole
9 eduskunnassakaan
10 ollut
11 asiassa
12 aktiivinen
13 paitsi
14 äänestyksessä
15 viime tipassa
16 Millään
17 puolueella
18 ei
19 ole
20 selkeää
21 visiota
22 asiaan
23 ja
24 kaikki
25 hyväksyvät
26 sen
27 että
28 suomalaiset
29 on
30 omassa
31 maassaan
32 vähemmistö
33 n
34 v
35 kuluttua
36 tai
37 nopeamminkin
38 nykyinenolla
39 Puolueen
40 ideologia
41 voisi
42 olla
43 realismi
44 Maahanmuuttokysymyksissään
45 kukaan
46 nykypuolueista
47 ei
48 ole
49 uskaltanut
50 esittää
51 seuraavia
52 loogisia
53 kysymyksiä
```

```
1 Mitenkäs
2 puoluetuki
3 toimii
4 Valitsijajyhdistys
5 tuskin
6 olisi
7 siihen
8 oikeutettu
9 vai
```



Topic modeling in action



$$p(D|\alpha, \beta) = \prod_{d=1}^M \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d.$$



Topic modeling in action

KNOWLEDGE

truth (#73)	understanding (#110)	beliefs (#151)	research (#163)	academia (#186)
truth (2028)	understand (1278)	believe (1385)	research (1748)	university (1786)
conversation (1909)	complete (710)	actually (905)	relate (752)	professor (1167)
media (1796)	reason (681)	realise (878)	example (742)	researcher (1029)
opinion (1743)	thought (542)	understand (866)	result (726)	research (noun) (783)
present (1546)	problem (531)	feel (759)	effect (634)	expert (447)
claim (noun) (1527)	reality (491)	live (739)	reason (595)	helsinki (432)
fact (1450)	habit (467)	reason (737)	found (582)	science (307)
claim (verb) (1291)	about (426)	know (731)	difference (554)	political (282)
try (1108)	think (388)	reality (709)	often (538)	research (verb) (261)
speak (1076)	belief (377)	business (692)	phenomenon (508)	discipline (234)

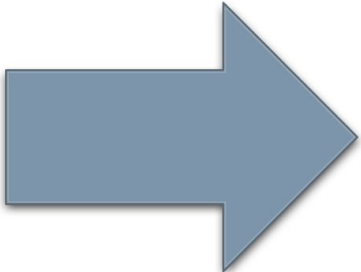
COUNTERKNOWLEDGE

climate change (#7)	medicine (#108)	quantification (#133)	intelligence (#162)	wtf (#166)
temperature (522)	doctor (1241)	amount (3171)	intelligence (795)	wtf (6506)
degree (305)	treatment (781)	percentage (2668)	race (508)	newspaper (6394)
rise (341)	patient (695)	number (2402)	gene (457)	mainstream media (1790)
climate (322)	disease (586)	great (1566)	evolution (445)	story (1611)
cause (303)	hospital (577)	last (1545)	intelligent (441)	media (1602)
warm (288)	illness (472)	finland (1400)	african (421)	reader (1270)
warming (260)	medicine (438)	statistics (1199)	species (410)	publish (1080)
atmosphere (257)	treat (421)	calculate (1039)	iq (363)	ilja ⁵ (1056)
water (257)	cause (397)	part (908)	test (348)	write (1029)
climate change (254)	hiv (304)	research (822)	africa (345)	finland (943)

CONSPIRACY

power (#11)	jews (#37)	mainstream media (#128)	eurabia (#183)
political (4203)	jew (2656)	media (5251)	europa (3175)
power (1970)	israel (2072)	journalist (4273)	people (1030)
politician (1690)	nazi (419)	story (3601)	world (845)
act (1361)	palestinian (370)	yle ⁶ (2782)	great (816)
politics (1287)	arab (273)	article (2662)	european (803)
great (1200)	palestine (262)	news (2157)	war (793)
people (1043)	holocaust (256)	hesari ⁷ (2066)	new (731)
benefit (962)	israeli (254)	newspaper (1501)	power (691)
leader (914)	germany (249)	read (1362)	destroy (600)
action (893)	area (244)	write (1227)	country (556)

Topic modeling in action

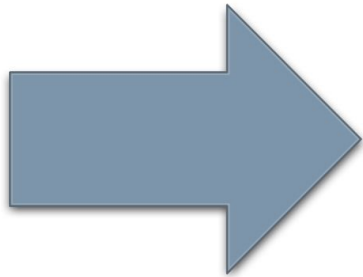


alanjl-mannym746.txt	0.087257
lanjl-tinnitus125.txt	0.086269
alanjl-mannym891.txt	0.086214
alanjl-mannym234.txt	0.085998
alanjl-mannym513.txt	0.084075
alanjl-mannym321.txt	0.077267
alanjl-mannym112.txt	0.074024
alanjl-mannym100.txt	0.072939
alanjl-mannym012.txt	0.072862
alanjl-mannym079.txt	0.072032
alanjl-mannym089.txt	0.069585
alanjl-mannym123.txt	0.069362

alanjl-mannym746.txt	0.087257
lanjl-tinnitus125.txt	0.086269
alanjl-mannym891.txt	0.086214
alanjl-mannym234.txt	0.085998
alanjl-mannym513.txt	0.084075
alanjl-mannym321.txt	0.077267
alanjl-mannym112.txt	0.074024
alanjl-mannym100.txt	0.072939
alanjl-mannym012.txt	0.072862
alanjl-mannym079.txt	0.072032
alanjl-mannym089.txt	0.069585
alanjl-mannym123.txt	0.069362



Topic modeling in action



Rubiikinkuutio
Jäsen^^^

Vs: Etniset geneettiset intressit
« Vastaus #244 : 09.06.2016, 21:15:02 »

Lainaus käyttäjältä: Brandis - 09.06.2016, 21:11:56

Viestejä: 7 247
Liked: 2434

A fonte puro pura defluit aqua.

On. Kaikki tietoiset pohdinnat oman jälkikasvunsa geneeistä ovat jo tietynlaisia rodunjalostusta, jos siinä on pyrkimys johonkin parempaan. Jos hankit lapsen älykkään/vahvan miehen kanssa saadaksesi lapselle älykkäitä/vahvoja geneejiä, niin se on jo ensimmäinen rodunjalostusmekaniikka. Tietysti biologisesti rodunjalostusta tekevät tässä mielessä epätietoisesti ihan eläimetkin ja toki ihmisetkin vaihtomaisesti. Mutta varsinaista rodunjalostusta se on minusta vasta, kun tietoisesti pyrkii tavalla tai toisella parempiin geneeihin. Eli jos alitajuisesti ihastuu vahvaan/älykkääseen mieheen omien geneiensä takia, niin minusta se ei ole varsinaisesti rodunjalostusta, vaan lähinnä viettien biologiala.

Lainaus

Kuten sanoit; se on mitä suurimmassa määrin moraalinen mielipide. On hyvä toimia itse kuten tahtois maailman ympärillään toimivan.

Yksi perinteinen ja hyvin laajasti hyväksytty moraalinen mielipide tuokin.

"Sananvapaus ei ole lupa loukata tai olla eri mieltä."

Kirjattu

Brandis
Jäsen^^^

Vs: Etniset geneettiset intressit
« Vastaus #243 : 09.06.2016, 21:11:56 »

Lainaus käyttäjältä: Rubiikinkuutio

Viestejä: 1 765
Liked: 1729

Villasukkainen,
Siikihansikkainen

Mutta siis noin pohjustukseksi esittämilleni väitteille, niin tuohan on moraalinen mielipide ja makuasia ja jos joku tahtoo jalostaa rotua, niin siitä vaan. Jos joku pitää sitä vääränä, niin siinähan pitää. Evoluutio ei välitä, vaan jatkaa toimintaansa molempien mielipiteiden kanssa.

Eikö kaikenlainen valinta tässä suhteessa ole mainitsemaasi rodunjalostusta? Olettaen, että pyrkii (omasta näkövinkkelistään katsottuna) parhaaseen mahdolliseen lopputulokseen kumppanin valinnassa ja sitä kautta jälkeläisten maailmaan tuomisessa.

Kuten sanoit; se on mitä suurimmassa määrin moraalinen mielipide. On hyvä toimia itse kuten tahtois maailman ympärillään toimivan.

« Viimeksi muokattu: 09.06.2016, 21:13:29 kirjoittanut Brandis »

Kirjattu

Meidän sirkuksemme on määrätty maan ja taivaan väliin. Maaperä antaa sille turvan ja taivas unelmat. Jos toinen niistä katoaisi, toistakaan ei voisi olla. "Jalat maassa ja sydän taivaassa" - niinhän sitä ennen sanottiin. ~ Martti Lindqvist, Uhelma rohkeasta elämästä 1992, s. 137 kappale 7 ~

törö
Jäsen^^^

Vs: Etniset geneettiset intressit
« Vastaus #193 : 26.02.2013, 18:34:49 »

Lainaus käyttäjältä: matkamiehii - 26.02.2013, 17:30:09

Viestejä: 18 610
Liked: 6607

Lainaus

Itselläni on geneejiä sekä idästä että lännestä, joten tulen juttuun sekä ruotsalaisien että venäläisten kanssa. En minä ole ruotsalaisille mikään "snalla du" niin kuin jonkun varsinaissuomalaisen postintekijän jälkeläinen suoraan alenevassa polvessa, vaikkain pakkoruotsia hyväksyään.

Olisi mielenkiintoista tehdä joskus DNA-testi, isän puolen suvussa vaikuttaisi olevan lieviä mongolirotumaisia piirteitä ja äidinpuolen suvussa on jonkin verran ruotsalaista verta. Itse näytän jonnekin stereotyyppiseltä ruotsalaiselta enemmän kuin suomalaiselta.

Suomalaiset on varsin mielenkiintoinen etninen ryhmä noin geneettisessä mielessä.

Suomi oli pitkään periferia, joten tänne valui sopeutumattomia aineksia vähän joka suunnalta. Laikoja ja rikollisia se vapaus ei niinkään viehättänyt, koska täällä oli myös aika köyhää ja suhtautuminen toisellaan siitä syystä melko nuivaa, mutta omaperäistä porukkaa tuli Ruotsista, Venäjältä, Preussista jne. ja vähän kauempaakin, koska rajojen ylittämisen oli melkein yhtä helppoa kuin turhapaikanhakijoille nykyisin eikä "asylumiakaan" tarvinnut käydä sanomassa poliisiedille. Melkein pä riitti kun ei vaikuttanut itolaiselta.

Tasapaksuun elämään näistä pakolaisista ja heidän jälkeläisistään ei oikein ole ollut, joten konsensusta vierastava yleissopeutumaton hompanssi voi olettaa olevansa geeniperimältään hyvinkin monikansallinen. Jos pottunokan alkuperää aletaan joskus tutkia kunnolla, en lainkaan ylläty jos se paljastuu tuontitavaraksi.

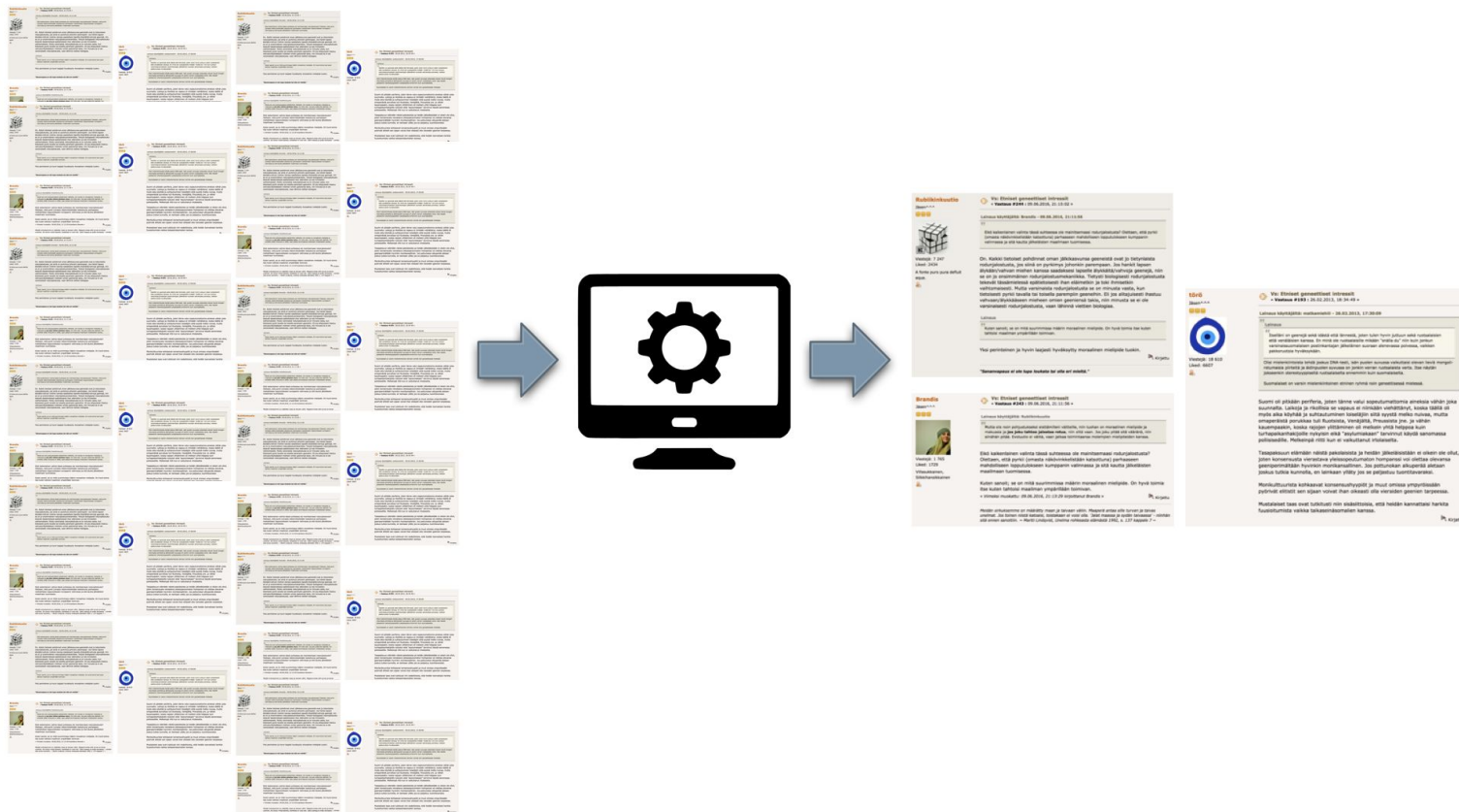
Monikulttuurista kohkaavat konsensusshyppiöt ja muut omissa ympyröissään pyörivät elitistit sen sijaan voivat ihan oikeasti olla vieraiden geenien tarpeessa.

Mustalaiset taas ovat tutkitusti niin sisäistioitsoita, että heidän kannattaisi harkita fuusiotumista vaikka taikaseinäsomalien kanssa.

Kirjattu

‘Distilling data’

- The role of computational analysis is to *distil the data* in terms of substantive interest for qualitative analysis



Tuloksia

- Kun MV-lehdessä puhutaan totuudesta, siellä toistellaan että MV kertoo totuuden, valtamedia valehtelee; voimakas erottelu *faktojen* ja *tunteiden* välillä
 - Kun Hommafoorumilla puhutaan tiedosta, puhutaan yleensä myös sukupuolesta
 - Naiset tunteellisia, miehet rationaalisia
 - Siinä missä MV-lehti ohittaa ilmastonmuutoksen huijauksena, Homman keskustelu on moniäänistä
 - Hommalaiset usein tuomitsevat MV:n ja perustelevat sitä sillä, että MV kirjoittaa uskomushoidoista ja huuhaasta
 - Salaliittoteorioita löytyy molemmista
 - Huom. nämä tulokset pääosin ”kvalitatiivisia” (tulkitsevia), voisi myös mitata näitä keskusteluja
-

Ilmastopaperi: taustaa

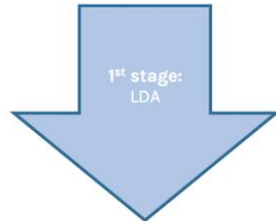
- YK:n ilmastokokousten mediahuomio tutkimuskohteena
 - Perustuu aikaisempaan tutkimusprojektiin joka tutki globaalia kansalaisyhteiskuntaa ja ilmastomuutoskeskustelun mediarepresentaatioita
 - Tutkimuskysymyksemme oli, missä keskustelunaiheissa on potentiaalia kompromisseille – siis mistä aiheista puhuvat useamman eri taustan puhujat
-

Ilmastopaperi: Aineisto

- New York Timesin ja The Hindun (Intia) YK:n ilmastokokouksien aikana julkaistut ilmastonmuutosta tai itse kokouksia käsitelleet artikkelit
 - Kioto (1997), Kööpenhamina (2009), Durban (2011)
 - Yhteensä 677 artikkelia (yli 400 000 sanaa)
 - Näistä poimittiin pelkästään suorat sitaatit tai sellaisten parafraseeraukset
 - Perustuu aikaisemman projektin aineistoon, jonka takia käsinkoodattu aineisto oli olemassa.
-



736 claims



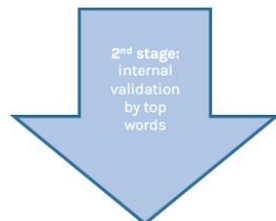
1st stage:
LDA

warm
global
scientist

energi
fund
billion

china
target
reduct

30 topics

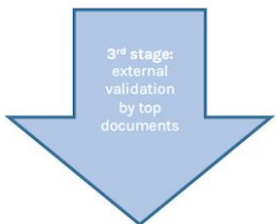


2nd stage:
internal
validation
by top
words

warm
global
scientist

energi
fund
billion

17 topics



3rd stage:
external
validation
by top
documents

Green Growth

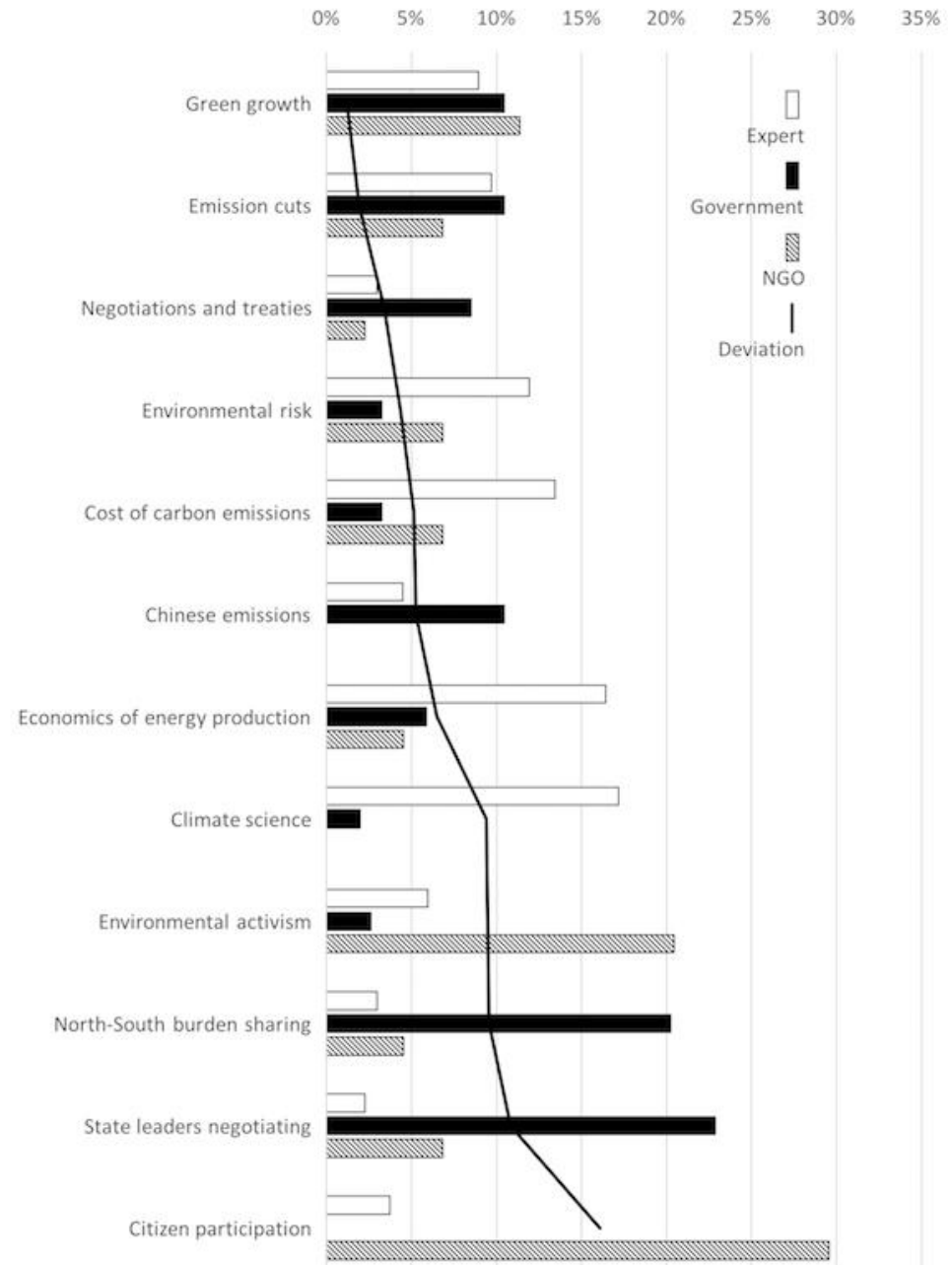
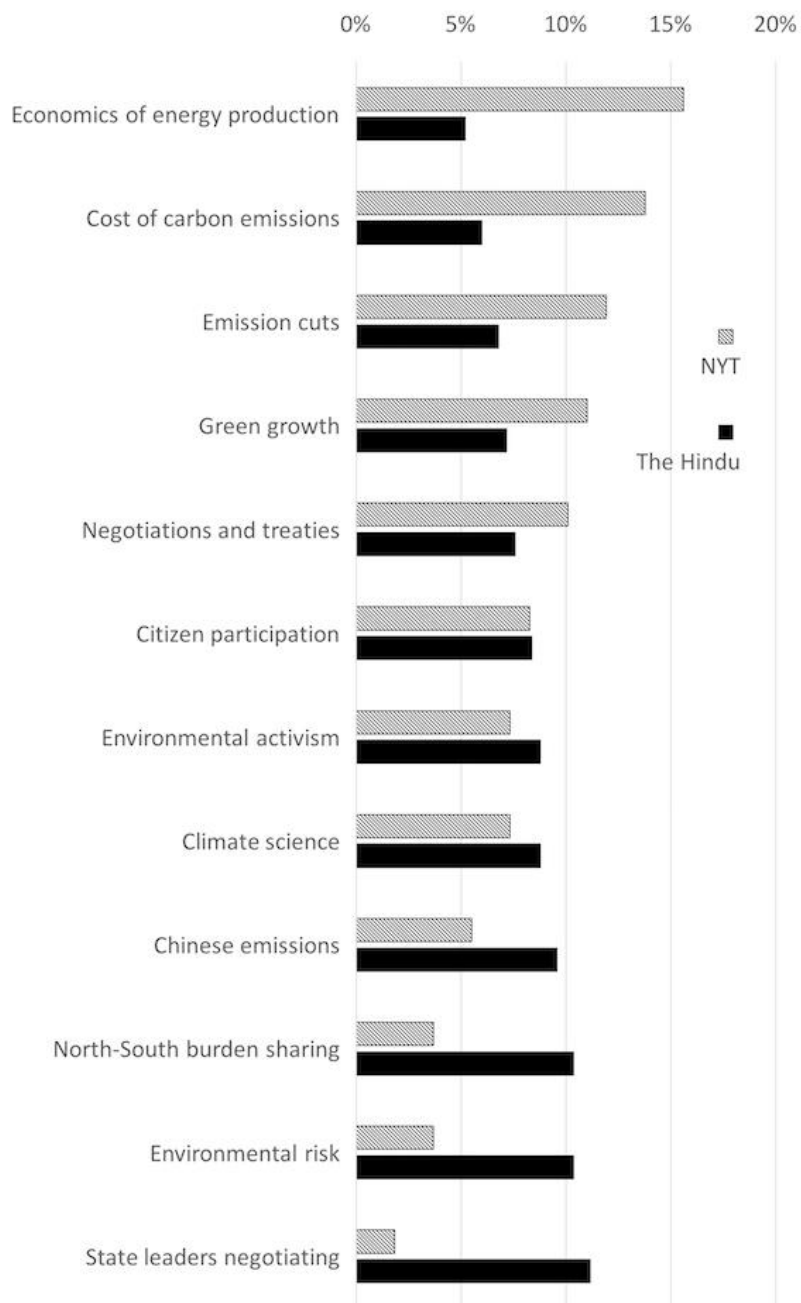
Emission Cuts

Climate Science

12 frames

Validoiminen ja tulkinta

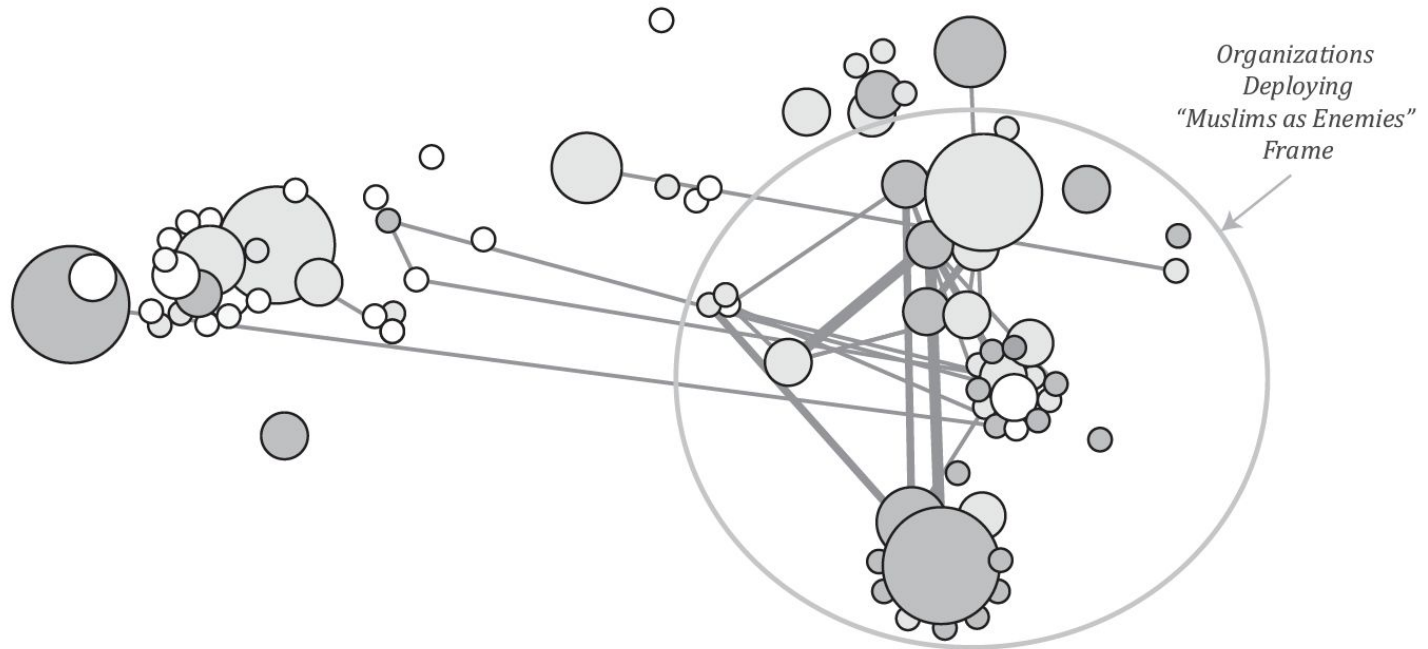
Green Growth	Emission Cuts	Negotiations and Treaties	Environmental Risk
<u>energi</u>	<u>emiss</u>	nation	<u>indian</u>
fund	cut	unit	sea
billion	<u>greenhous</u>	state	water
state	<u>industri</u>	<u>treati</u>	<u>increas</u>
public	gas	commit	forest
clean	<u>gase</u>	<u>economi</u>	today
invest	call	major	ocean
renew	system	<u>respons</u>	region
<u>creat</u>	adopt	<u>american</u>	risk
govern	<u>declar</u>	recent	rate
Cost of Carbon Emissions	Chinese Emissions	Economics of Energy Production	Climate Science
carbon	china	<u>econom</u>	warm
<u>emiss</u>	target	<u>technolog</u>	global
<u>reduc</u>	<u>reduct</u>	cost	scientist
<u>dioxid</u>	chines	<u>compani</u>	research
<u>pollut</u>	growth	<u>energi</u>	<u>univers</u>
trade	<u>intens</u>	fuel	<u>atmospher</u>
cap	current	power	<u>caus</u>
coal	<u>reduc</u>	money	<u>studi</u>
power	plan	price	effect
product	<u>oblig</u>	<u>mani</u>	<u>releas</u>



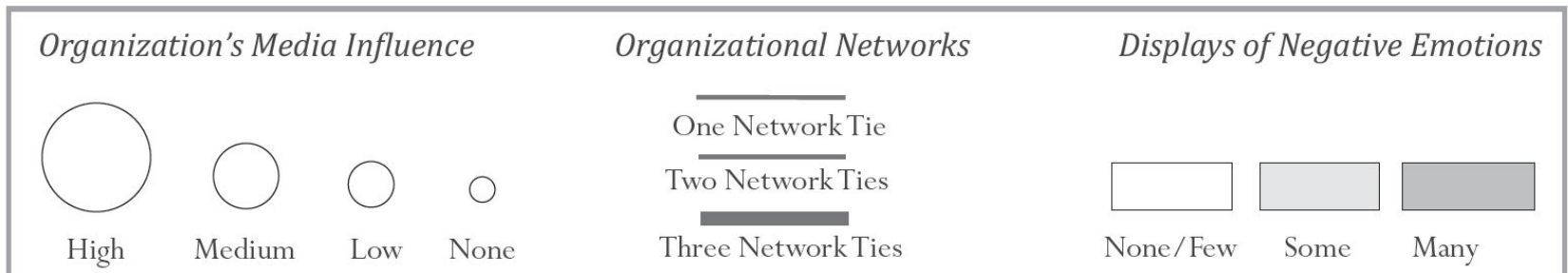
Ilmastopaperi: tuloksia

- Vihreän kasvun ja kestävä teknologian aiheet, päästöleikkaukset sekä itse kokousprosessi yhdistivät eri puhujia
 - Kansalaisten osallistuminen kiinnosti ainoastaan kansalaisia itseään
 - The Hindussa puhuttiin paljon enemmän ympäristöriskeistä sekä ilmastonmuutokseen vastaamisen moraalisesta taakanjaosta
 - New York Timesissa enemmän energian tuotannosta, sen hinnasta ja taloudellisista kytkennöistä sekä hiilipäästöjen hinnasta
-

Menetelmien yhdisteleminen



KEY



Christopher Bail 2012: "The Fringe Effect", American Sociological Review

Tulkinta ja ymmärtäminen

- Monet uusista luennolla esitetyistä malleista tarvitsevat käytännön tutkimustyössä tuekseen kvalitatiivista ja ymmärtävää tulkintaa
 - Aihemallinnus ja muut klusterointiin perustuvat menetelmät tarvitsevat taakseen ymmärrystä
 - Monet aineiston analyysitavat nojaavat vuorotellen laskennalliseen ja tulkinnalliseen analyysiin.
-

Etiikka

- Kaikkea saatavilla olevaa dataa ei voi noin vain käyttää
- Vaikka yksittäiset muruset olisivat julkisia, kun ne kerätään yhteen kasaan, datasta voi saada sellaista tietoa yksilöistä, jota tämä ei tarkoittanut asettaa julkiseksi
- Kaikki mikä on sallittua ei välttämättä ole eettistä, mutta kaikki mikä on eettistä ei myöskään välttämättä ole sallittua

ON MAY 8, a group of Danish researchers publicly released a dataset of nearly 70,000 users of the online dating site OkCupid, including usernames, age, gender, location, what kind of relationship (or sex) they're interested in, personality traits, and answers to thousands of profiling questions used by the site.

When asked whether the researchers attempted to anonymize the dataset, Aarhus University graduate student Emil O. W. Kirkegaard, who was lead on the work, replied bluntly: "No. Data is already public."

Etiikka 2: datasettien yhdistely

- Anonymisointi kuulostaa hyvältä idealta, muttei käytännössä aina auta
- Yhdistämällä anonymisoitua dataa muuhun salattuun tai julkiseen dataan, nimet ovat usein löydettävissä

Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

February 5, 2008

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We

Mistä dataa?

- Open Data (esim. HS Vaalikone)
 - Sosiaalinen media: API
 - “Scraping”
 - Yrityksiltä
 - Ks. Suomi24 ja Citizen Mindscapes -projekti, Elisan paikkatietodata
 - Huom. mahdolliset eettiset ja juridiset kysymykset
 - Yhteiskunnallinen tietoaarkisto (www.fsd.uta.fi), CSC.fi, FIN-CLARIN kielipankki, Tilastokeskus, Helsinki Region Infoshare
-

Don't believe the hype

- Hyphen mukana tulee joskus katteettomia lupauksia
 - Mitä big data -aineistot edustavat?
 - Voiko Twitterillä tutkia yhteiskunnallista keskustelua Suomessa?
 - Onko henkilön FB-verkosto hänen "sosiaalinen verkostonsa"?
 - Big data -aineistoja ei ole koottu vastaamaan tutkimuskysymyksiin
 - Mitä kulttuuristen asioiden määrät tarkoittavat?
 - Paljonko on 0,7 populismia?
 - Onko datatiede tarpeeksi läpinäkyvää?
 - Algoritmit "mustina laatikkoina"
 - Lopulta kun "big dataa" opitaan käyttämään tarpeeksi, siitä tulee yksi datan tyyppi muiden joukossa
 - "Big datan" sijaan aletaan puhua some-datasta, keskusteluaineistoista, verkostoaineistoista...
 - "Computational social science" -> social science
 - Sama pätee myös "digitaalisiin ihmistieteisiin"
-

Kiitos!
