Kynä ja kone keskustakampuksella

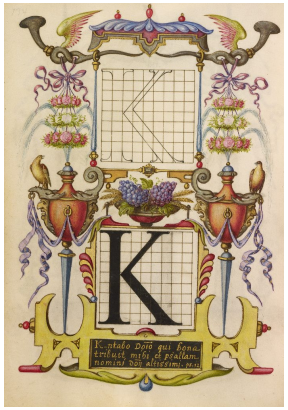
“Kaikki siitä puhuvat, mutta mitä se on: digitaalisuus?”

Kynä ja kone:
Menetelmät ja analyysit

UNIVERSITY OF HELSINKI

1640 THE NATIONAL LIBRARY OF FINLAND

Timo Honkela
timo.honkela@helsinki.fi



15.9.2016

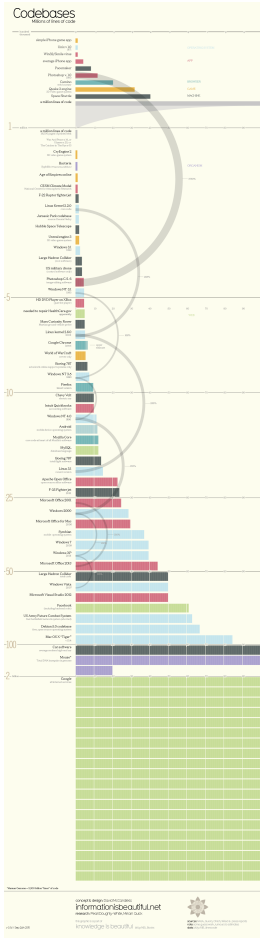
Järjestäjät

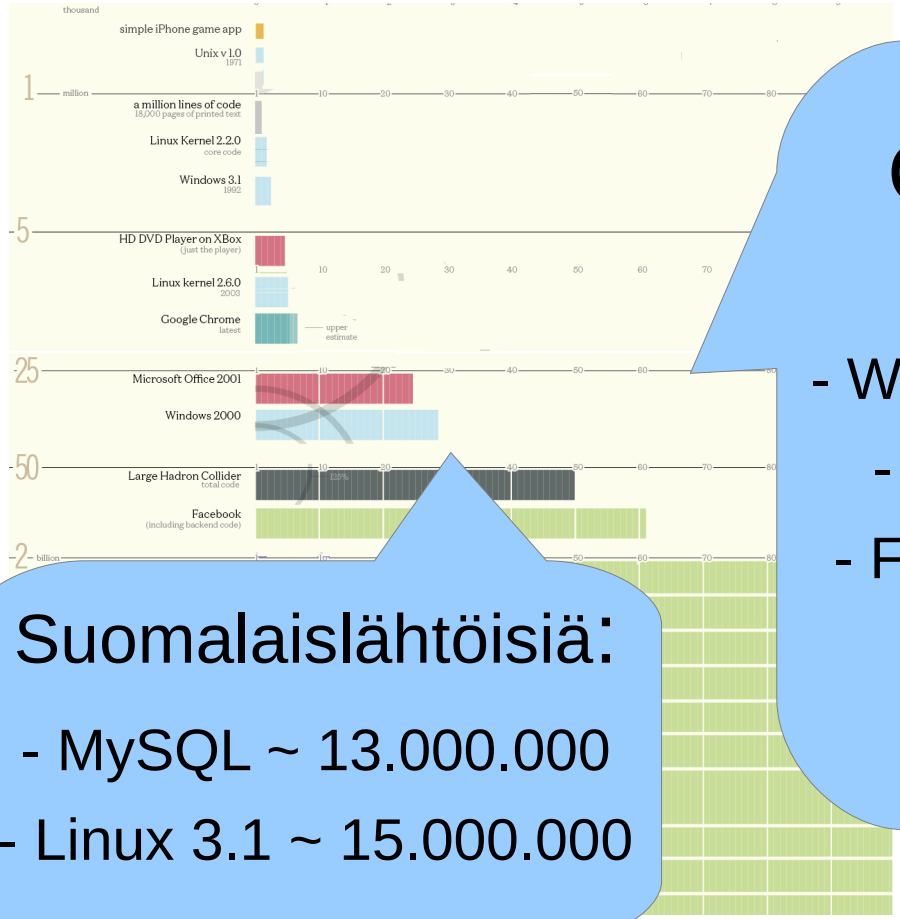
Pirjo Hiidenmaa

Timo Honkela

Maija Paavolainen

Valtavat ohjelmistomassat nyky-yhteiskunnan peruspilareina





Ohjelmarivejä:

- Unix 1.0 ~ 10.000
- Windows 3.1 ~ 2.000.000
- Firefox ~ 10.000.000
- Facebook ~ 60.000.000
- Googlen palvelut ~ 2.000.000.000

Suomalaislähtöisiä:

- MySQL ~ 13.000.000
- Linux 3.1 ~ 15.000.000

http://www.informationisbeautiful.net/visualizations/million-lines-of-code/

http://375humanistia.helsinki.fi/humanistit

Humanististen tieteenalojen ja tutkimusaiheiden moninaisuus ja tärkeys

Andrew Chesterman

Käännöstiede

Pirjo Hiidenmaa

Tiedeviestintä

Ilkka Niiniluoto

Filosofia

Terttu Nevalainen

Historiallinen sosiolingvistiikka

René Gothoni

Uskontotiede

Pirjo Kolbe

Euroopan historia

Jaakko Hintikka

Tietoteoria

Anna Mauranen

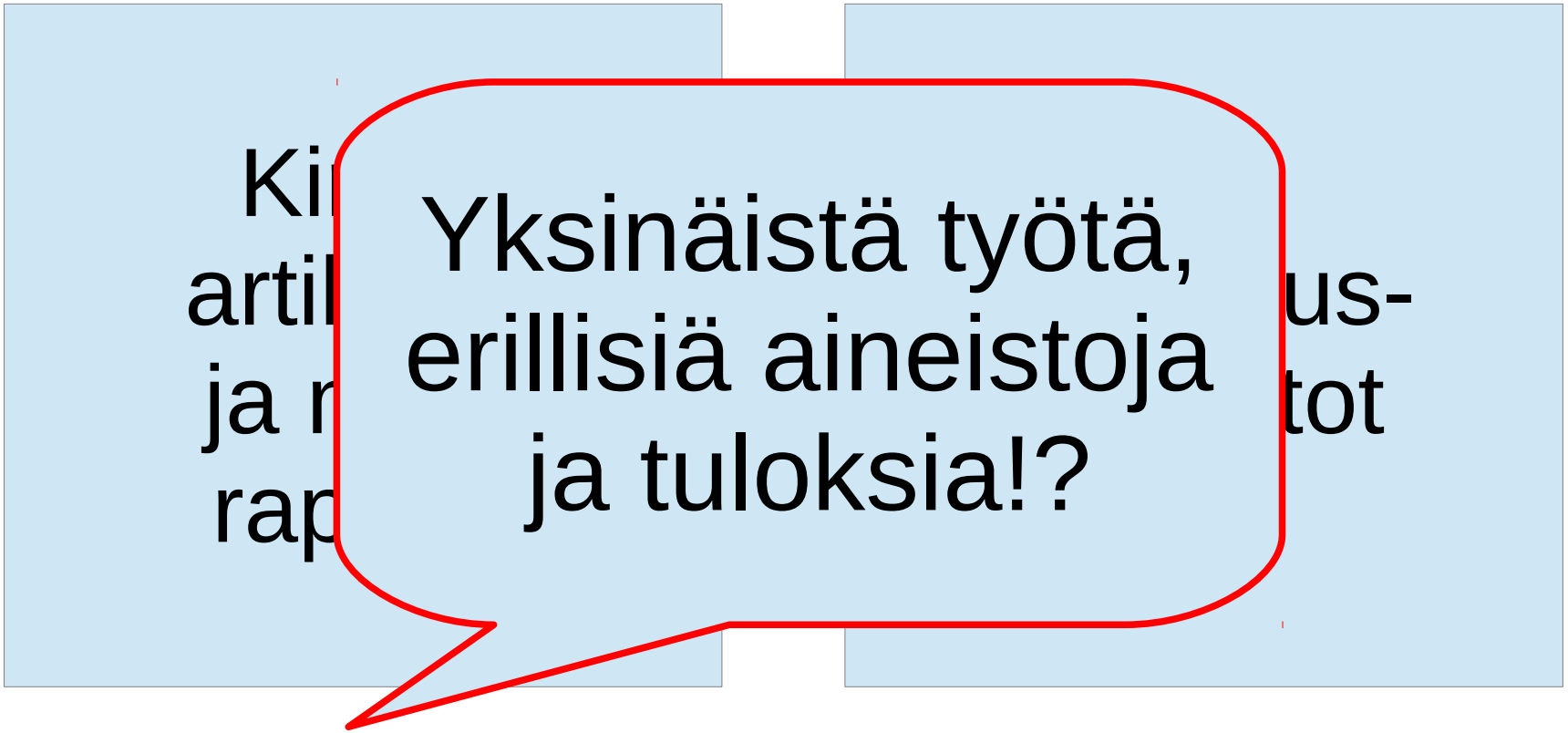
Englanti lingua francana

Humanistiset tieteet: valtavat aineistot

Kirjat, artikkelit ja muut raportit

Tutkimus-aineistot

Humanistiset tieteet: valtavat aineistot



Tietoteknisiä muutostekijöitä

- Laskentakapasiteetti kasvaa
- Muistitila kasvaa
- Koneoppimisen ja hahmontunnistuksen lisääntyvä käyttö
  - Uusia menetelmiä kehitetään, vanhoja löydetään uudelleen
- Saatavilla olevien aineistojen määrä ja koko kasvavat

  (“Big Data”, “Open Data”, “Open Linked Data”)

Miksi laskenta- ja muistikapasiteetilla on merkitystä?

- Voidaan tarkastella digitaalisessa muodossa olevan puheen, musiikin tai kuva- ja videomateriaalinen laatua, jos käytettävissä on niukasti tai runsaasti tietokoneen muistia

Edes tekstiaineistojen analyysi
ei ollut aikoinaan kunnolla mahdollista
kapasiteettirajoitusten takia.

Niinpä monia hyviä analyysimenetelmiä
on voitu keksiä jo vuosikymmeniäkin
sitten ja nykyään niitä “keksitään uudelleen”

Kapasiteetin merkitystä

- Esimerkiksi digitaalinen kuva tai musiikkikappale voidaan esittää mielekkäästi vasta, kun kapasiteettia on riittävästi
- Kognitiivisten tulkintaprosessien simulointi vaatii vielä paljon enemmän resursseja

Tieteenfilosofisia näkökulmia

- Deduktiivinen, abduktiivinen ja induktiivinen päättely
- Teoria- versus aiheistolähtöisyys

Rakennetaan teoria tai esitysmuoto sisäisen näkemyksen varassa

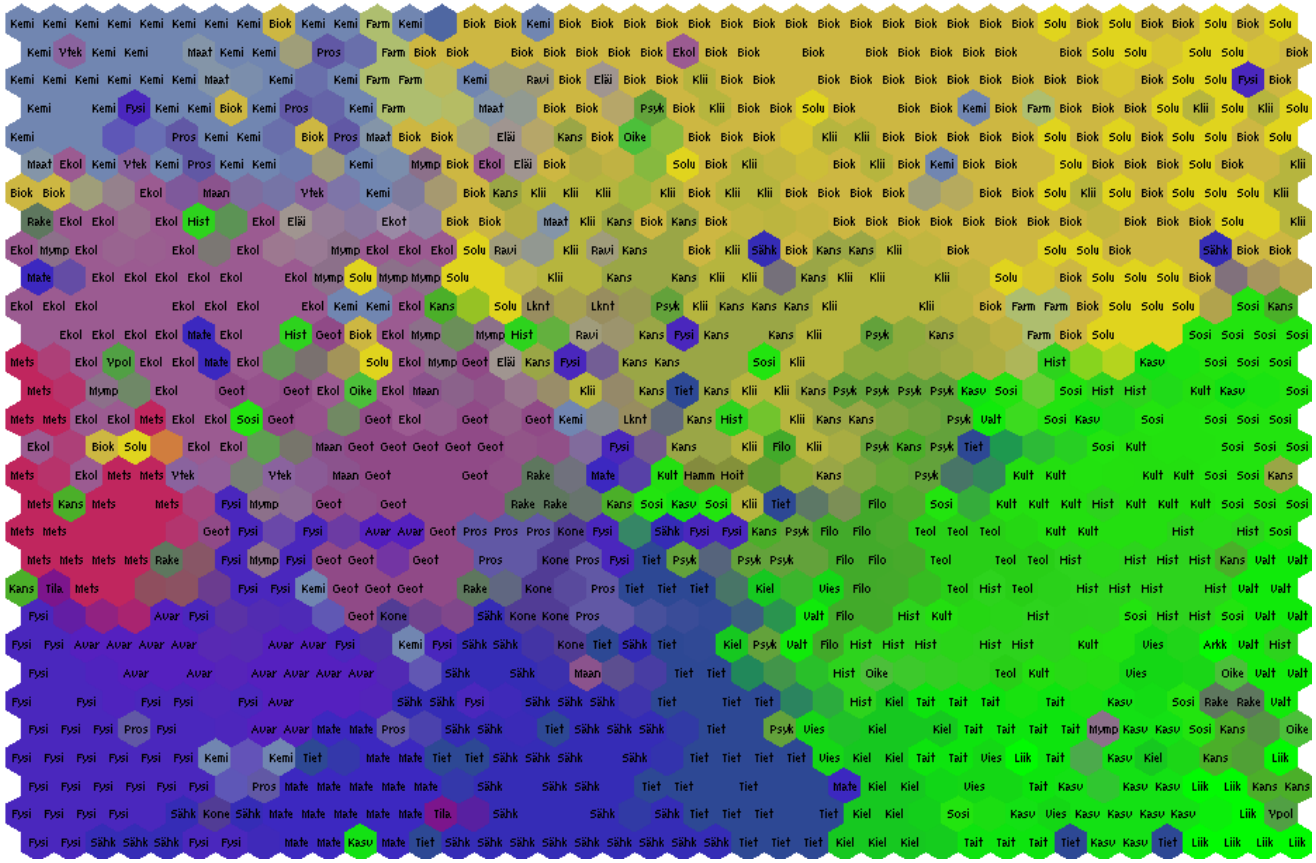
Rakennetaan teoria tai esitysmuoto teorialähtöisesti mutta varmistaen sen toimivuus aineistoilla

Rakennetaan teoria tai esitysmuoto aineistolähtöisesti esimerkiksi koneoppimista hyödyntäen

Tilastotiede “von oben” ja “von unten”

- Tilastotiedettä voidaan käyttää menetelmänä hypoteesien statuksen selvittämiseksi
- Tilastotiedettä ja todennäköisyyslaskentaa voidaan hyödyntää myös “automaattiseen teorianmuodostukseen” eli siihen, että tilastollisen koneoppimisen avulla muodostetaan tieteellisiä malleja

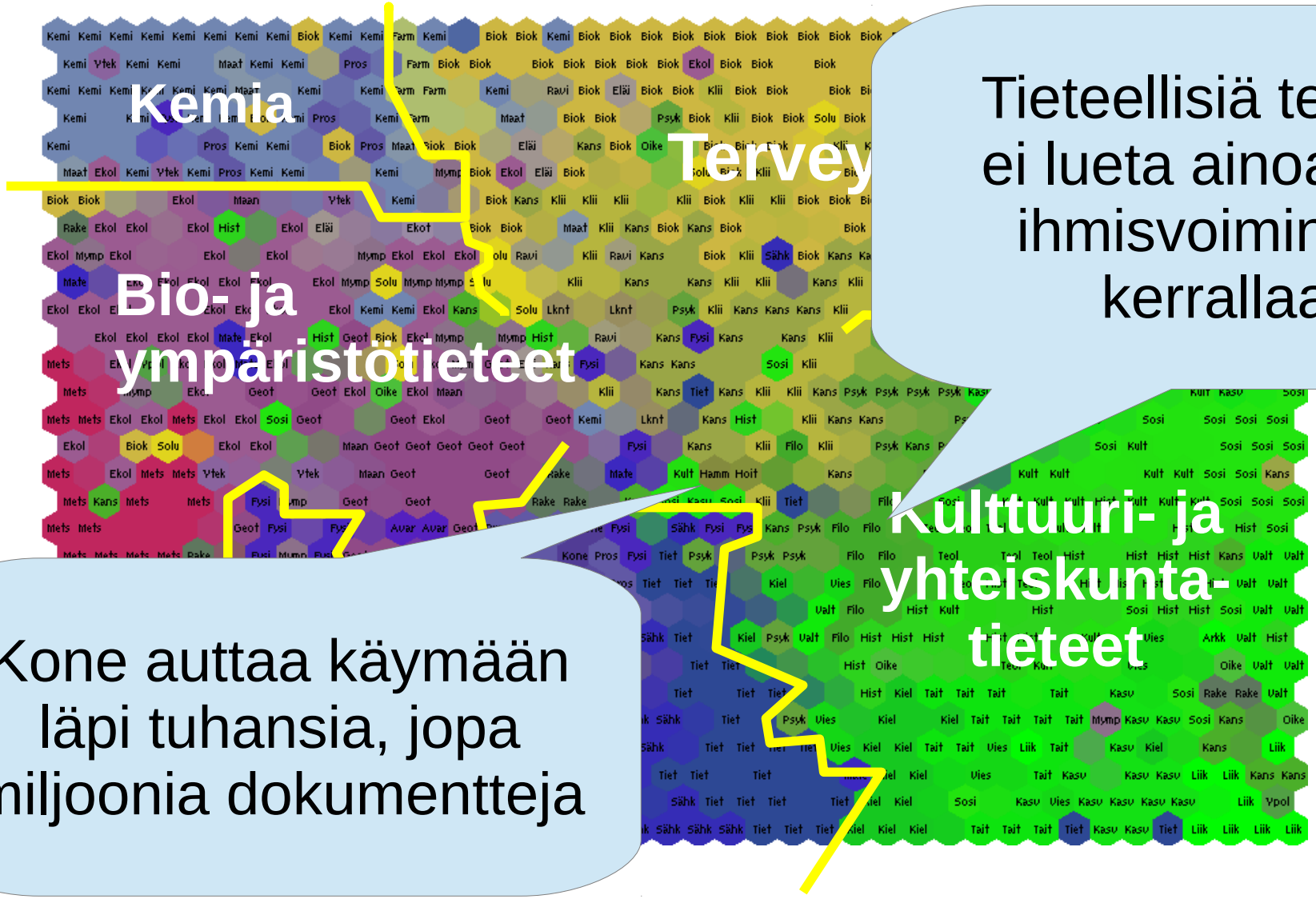
Suomen tieteen analyysi “von unten”: Suomen Akatemian aineiston louhinta



(Honkela & Klami, 2008)

Kohonen: Self-Organizing Map -neuroverkkoalgoritmi & Paukkeri et al.: Likey / datapohjainen termien valinta aineistosta

Suomen tieteen analyysi “von unten”: Suomen Akatemian aineiston louhinta



Tieteellisiä tekstejä ei lueta ainoastaan ihmisvoimin yksi kerrallaan

Kone auttaa käymään läpi tuhansia, jopa miljoonia dokumentteja

(Honkela & Klami, 2008)

Sanojen suhteet paljastuvat niiden käytöstä

- Kun käytettävissä on suuria tekstiaineistoja, mielivaltaisen kielen sanojen välisiä suhteita voidaan selvittää tilastollisesti
- Perusidea on se, että kahta sanaa käytetään tyypillisesti samaan tapaan (samanlaisessa lauseyhteydessä), jos niiden merkitykset ja/tai kieliopillinen rooli on samankaltainen

Tekoäly ja koneoppiminen kirjastossa

- Automaattinen asiasanoitus
- Dokumenttien automaattinen luokittelu
- Kunkin dokumentin sijoittaminen yhteen tai usempaan luokkaan; ehkä erilaisilla jäsenyysasteilla
- Dokumenttien ryhmittely luokittelun sijaan tai lisäksi
- Virtuaalinen kirjasto

WEBSOM: Honkela, Kaski, Kohonen, Lagus (1996...)

AINEISTOT

- Numeerinen data
- Tekstikokoelma
- Vahvasti hahmoluonteinen (ääni, puhe, kuva, jne.)
- Sekamuodot

ANALYYSITAPOJA

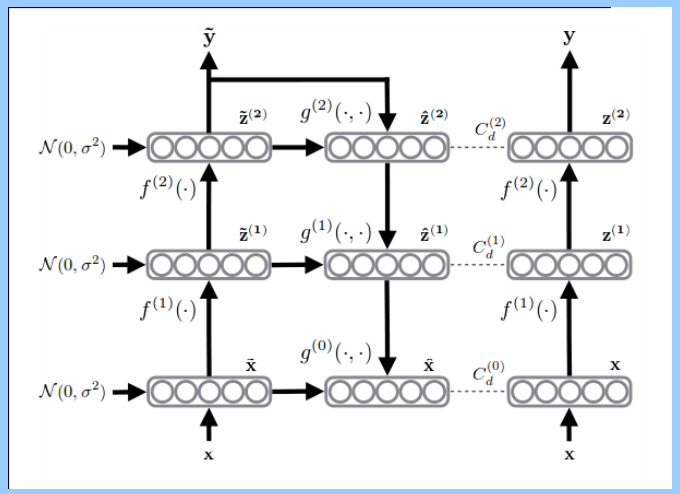
- Teorian testaaminen
- Teorian tai representaation kehkeyttäminen datan pohjalta
- Sekamuodot

Digitaaliset ihmistieteet tieteen kartalla

- Ihmistieteiteiden ja yhteiskuntatieteiden ydinkysymyksiä ei voi lähestyä yksioikoisesti luonnontieteiden ihanteiden varassa ja niiden menetelmillä; siihen ne ovat liian yksinkertaisia ja yksinkertaistavia
- Tietokoneavusteisuus antaa kuitenkin mahdollisuuden rakentaa uusia siltoja ja uudenlaista meta-analyysia
- Suuria humanistisia aineistoja voidaan analysoida olettamatta, että niiden pohjalla olevat tulkinnat ovat luonnontieteellisen yksinkertaisia vaan pohjautuvat rikkaaseen ja monimutkaiseen sosiokulttuuriseen ja yksilöpsykologiseen puitteeseen

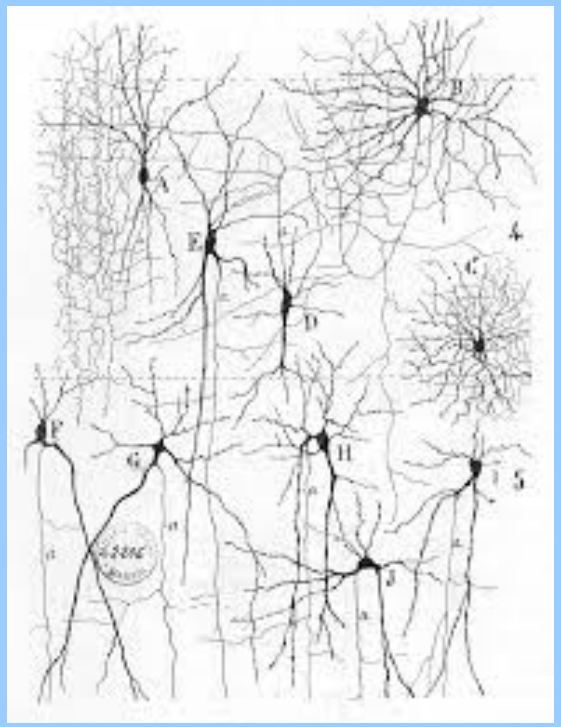
Koneen lisääntyvä intuitio?!

Honkela (2000)



Esim. Rasmus, Valpola, Honkala. Berglund, Raiko

http://arxiv.org/pdf/1507.02672v1.pdf



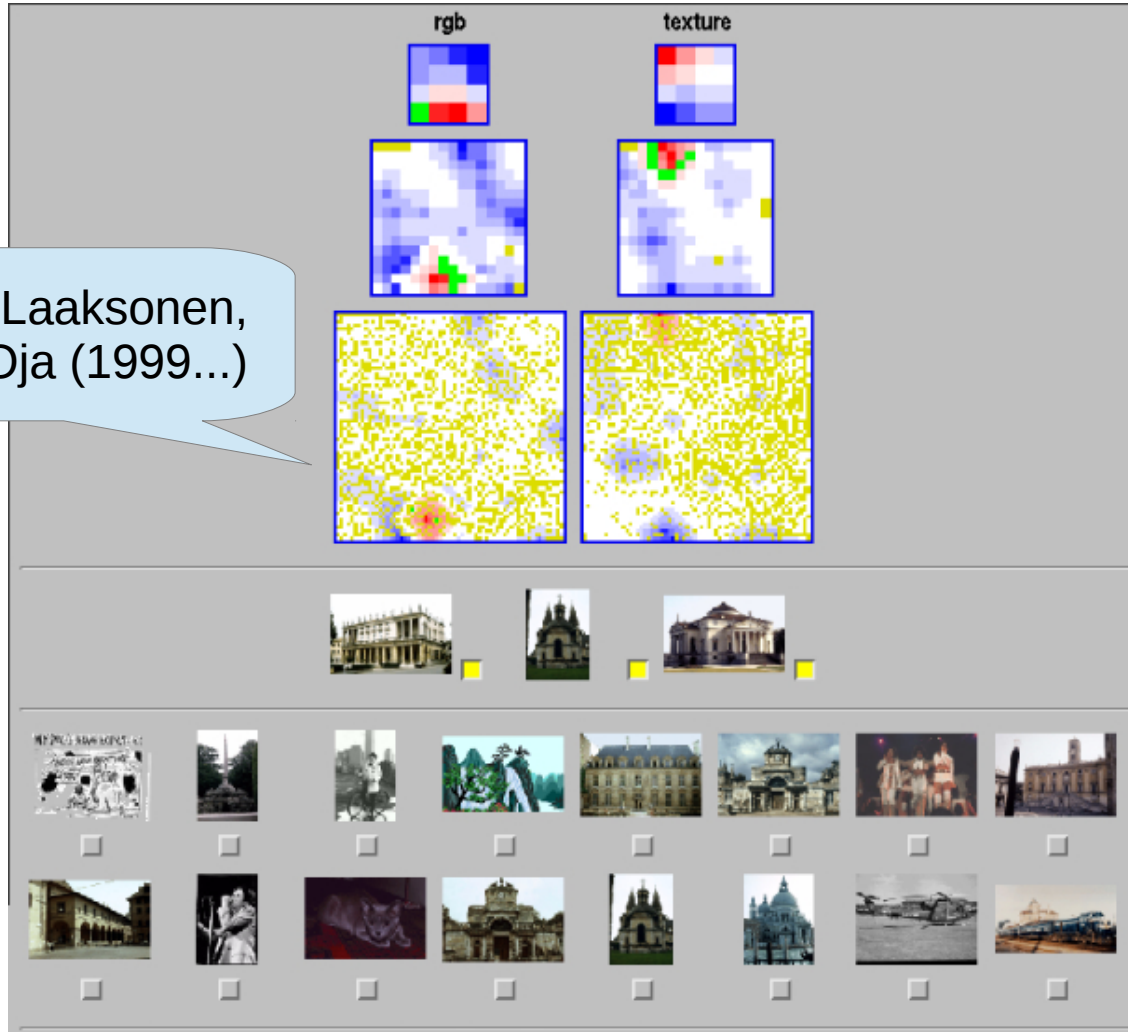
https://en.wikipedia.org/wiki/Biological_neural_network

(Kieli)aineistojen analyysin erityiskysymyksiä

- Tekstiaineistojen teorialähtöinen koodaus
- Metadatan käyttö ilmiöiden tutkimuksessa
- Ko(n)tekstidatan hyödyntäminen analyysissa (esim. topiikkimallit)
- Yhteiskunnallinen analyysi
- Tunteet ja kieli
- Kieli ja kuvallinen informaatio
- Kulttuurikonteksti, tulkinnan yksilöllisyyys

Kuvia katselevat ja tuottavat koneet

- Myös kuvallinen data voi olla koneoppimis-menetelmien kohteena
- Kaupalliset sovellukset tunnistavat esimerkiksi ihmisiä kuvista

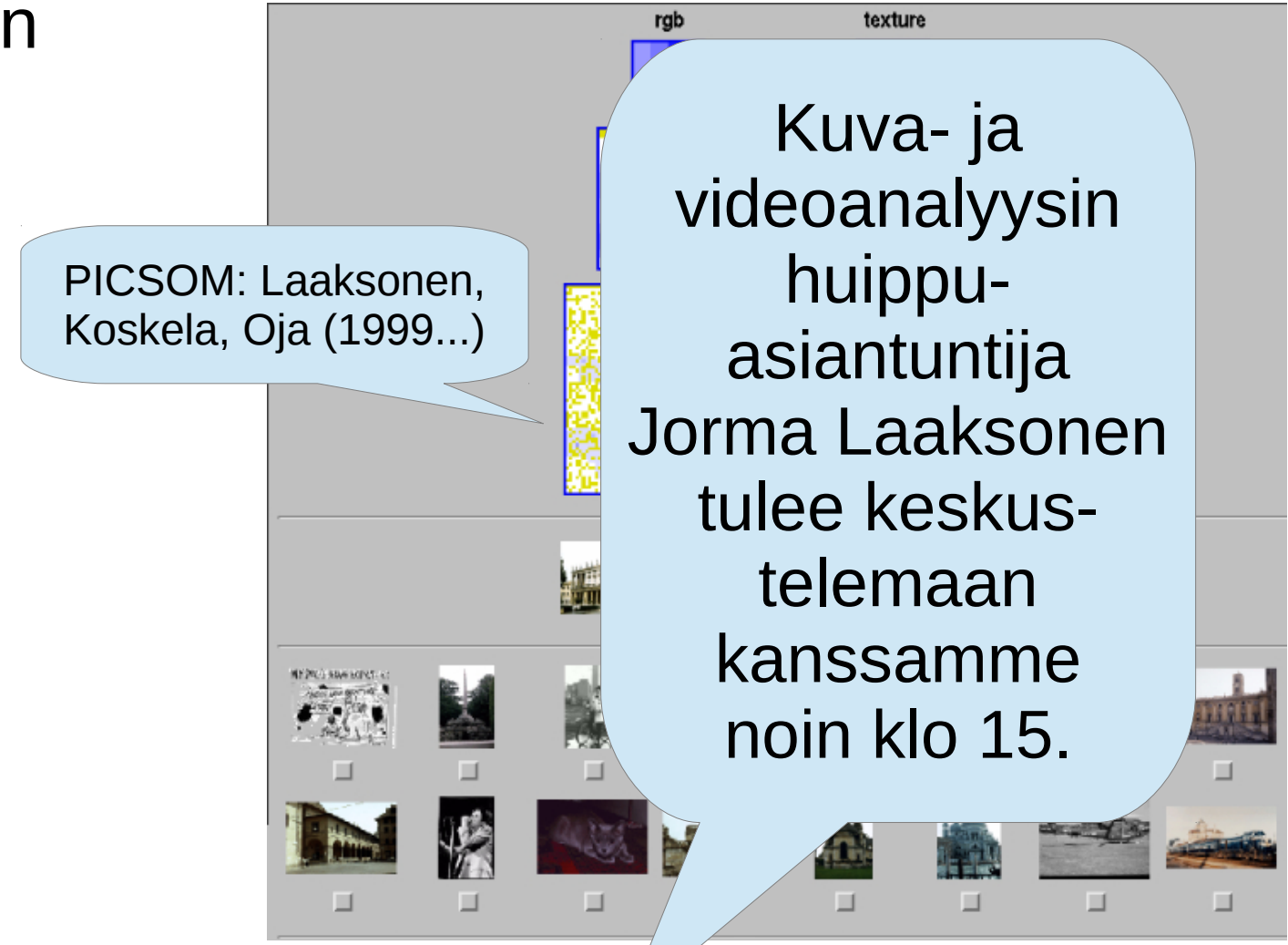


PICSOM: Laaksonen, Koskela, Oja (1999...)

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.6021&rep=rep1&type=pdf

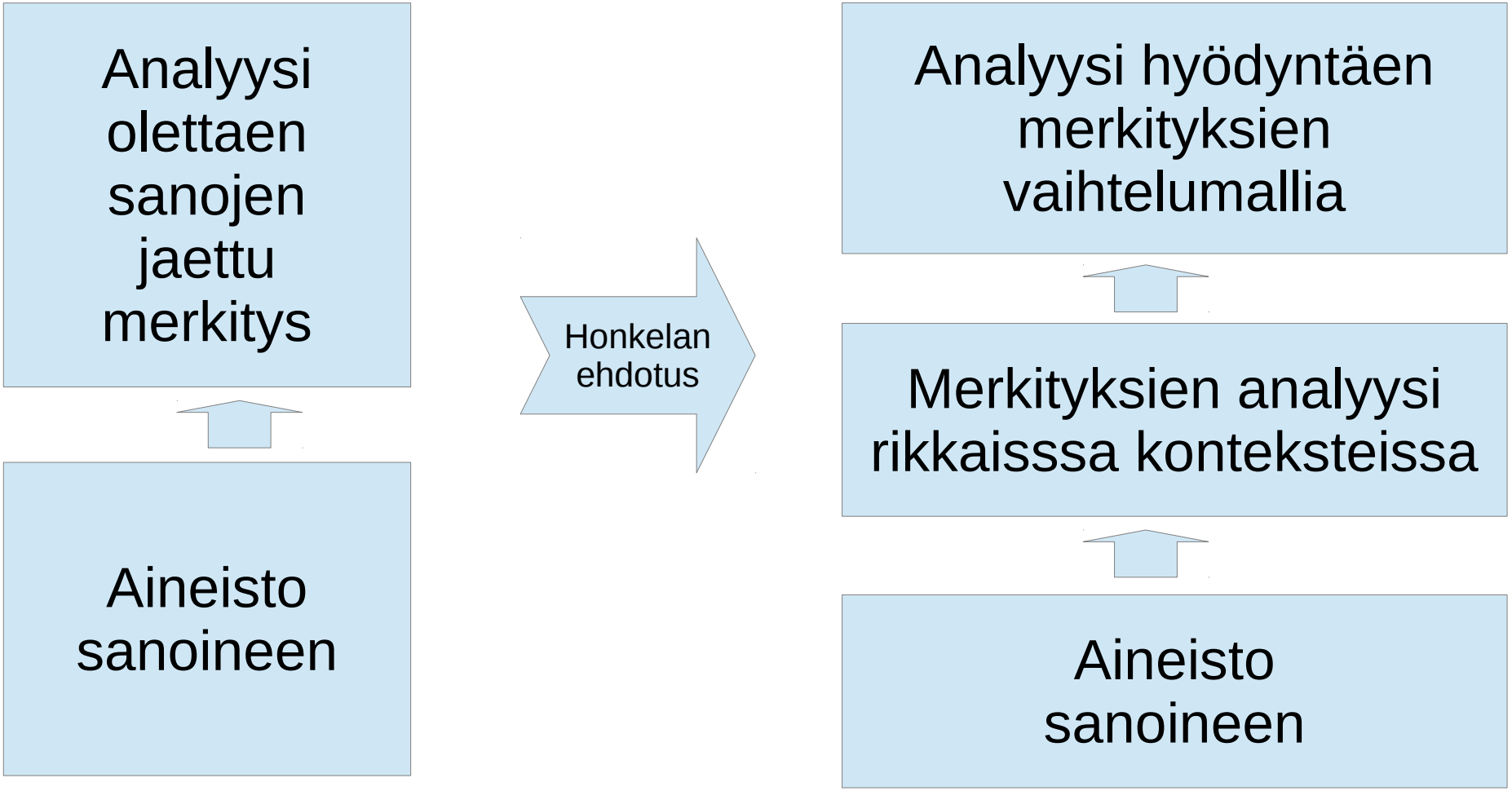
Kuvia katselevat ja tuottavat koneet

- Myös kuvallinen data voi olla koneoppimis-menetelmien kohteena
- Kaupalliset sovellukset tunnistavat esimerkiksi ihmisiä kuvista

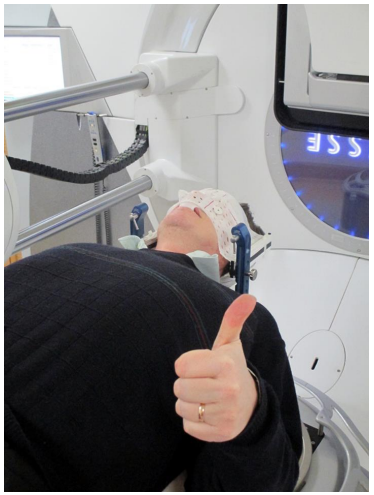


http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.6021&rep=rep1&type=pdf

Humanististen tieteiden meta-analyysi



Kiitos!



http://375humanistia.helsinki.fi/humanistit/timo-honkela

http://www.slideshare.net/timohonkela

https://www.youtube.com/watch?v=UXwkGPMMZdk