Datatieteen data

Pauli Miettinen, datatieteen professori, Itä-Suomen yliopisto, 11.5.2023

Datatieteen monta merkitystä

- Datatieteellä tarkoitetaan montaa eri asiaa
- Tämä esitys: datatiede on tietojenkäsittelytieteen ja tilastotieteen leikkauksessa oleva metodologinen tiede
  - Oma taustani:
    - Tietojenkäsittelytieteilijä, tarkemmin tiedonlouhija
    - Tutkimus lähinnä menetelmänkehitystä, viime aikoina enemmän myös soveltavaa

Datatiede – data = tiede?

- Datatiede on menetelmätiede = kehittää menetelmiä
- Data ei ole ole tutkimuksen tulos
  - Menetelmät ovat
- Käytetty data on usein ”löydettyä” (”found data”)
  - Kerätty muuhun tarkoitukseen muiden toimesta
  - Hyvä ja kattava metadata helpottaa meitä

”Perinteinen” datatieteessä luotava data

- Joitain datasettejä kerätään & julkaistaan
  - Usein koneellista louhintaa esim. internetistä
    - Käyttöoikeuksia ei välttämättä ole selvitetty
    - ”se oli internetissä…”
- Metadata hyvin vaihtelevaa
  - Alalla ei ole vakiintuneita käytänteitä
- Myös synteettisiä datoja

Datan tyypillisiä ongelmia

- Keräysprosessia ei dokumentoitu
- Dataformaattia ei dokumentoitu
- Lukuisia eri dataformaatteja
  - csv, arff, mmx, txt, tab, tsv, dat, fimi, …
  - Yleensä tekstimuotoista dataa suositaan
  - Oletus: kaikki osaavat muuttaa datan tarvittavaan muotoon

Koodi = datatieteen metadata

- Tyypillisin metadatan muoto on lyhyt kuvaus siitä, mitä data on ja jonkinlainen ohjelmalistaus
  - Joko esimerkki datan käytöstä tai koodi jolla testit voi toistaa
  - Tai koodi joka prosessoi muualta saatavan datan käytettyyn muotoon
- Yhtäältä täydellinen dokumentaatio siitä, mitä on tehty
- Toisaalta erittäin vaikeasti käsiteltävää metadataa
  - Koodi ei sinällään sisällä juuri mitään rakenteisia ominaisuuksia

Koodi = datatieteen data

- Tyypillinen tutkimuksen tulos on menetelmä
  - menetelmä = idea, miten ongelma voidaan ratkaista
  - Koodi = menetelmän konkreettinen instanssi
- Tänä päivänä koodi usein julkaistaan
  - data + koodi = toistettava tutkimus (?)
- Koodin edellytyksiä ei välttämättä julkaista
  - kirjastojen & kääntäjän versiot, ympäristömuuttujat, yms.

Suutarin lapsilla ei ole kenkiä

- Koodin metadatan ongelmat:
  - Oletuksia ei ole dokumentoitu
  - Toimintaa ei ole dokumentoitu
- Koodin metadataan paljon erilaisia ajatuksia
  - JavaDoc, PyDoc, …
  - CMake, autoconf/automake, Pipfile/requirements.txt, deb, rpm, …
- Yleensä jossain on README, jossa kerrotaan jotain…
- *Read the source, Luke*

Summa summarum

- Testien toistettavuutta ei valvota
  - Yrityksiä on ollut, liian rajoitetut ympäristöt rajoittavat mahdollisuuksia
- Koodin julkaisu on yhtäältä paljon parempi ja toisaalta paljon huonompi tilanne kuin monella muulla alalla
- Varsinaista dataa julkaistaan vähemmän, mutta koodi on monesti datan asemassa
  - Tätä ei aina ymmärretä