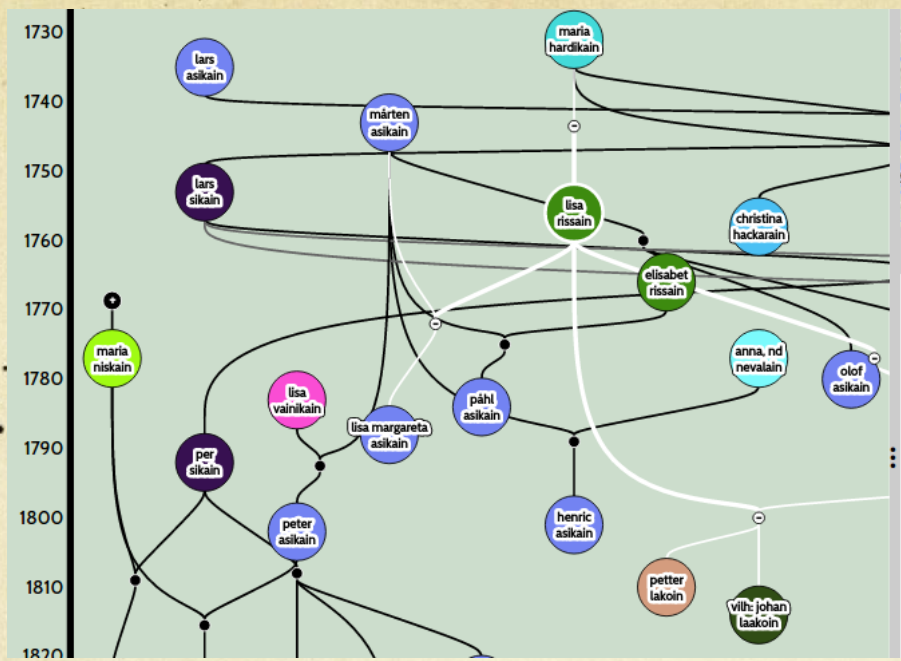
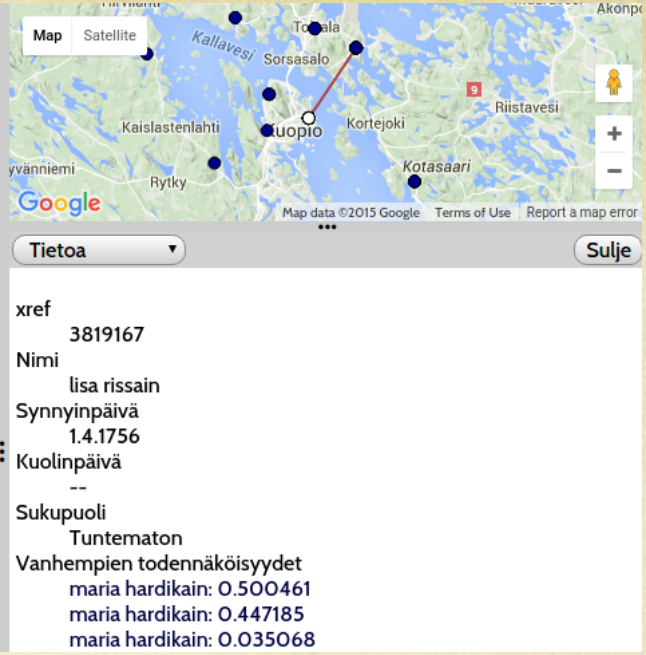




Nimien linkitys historiallisissa kirkonkirja-aineistoissa

Eric Malmi
Tietotekniikan laitos, Aalto-yliopisto

Aiheet

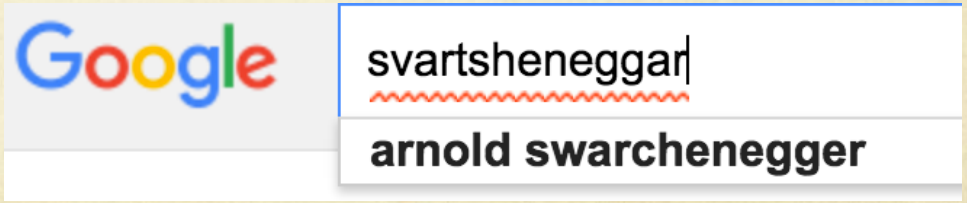
1. Ongelman kuvaus
2. Nimien normalisointi ja samankaltaisuus
3. Sukupuiden automaattinen päättely
4. AncestryAI – tekoälysovellus sukututkijan avuksi

Ongelman kuvaus

- Viittaavatko annetut nimet samaan henkilöön?

  Juho vs. Johannes Maija vs. Raija

- Engl. *entity resolution*, *record linkage*, *deduplication*
- Sovellukset
  - Sukututkimus
  - Tietokantojen yhdistäminen
  - Tiedonhaku
- Kollektiivinen linkitys



Nimien samankaltaisuus

- Lukuarvo kahden nimen samankaltaisuudelle
- Levenshtein
  - “Pienin määrä operaatioita, joiden avulla toinen merkkijono voidaan muuttaa toiseksi.”
  - LevenshteinDist(“eerik”, “eric”) = 2
- Jaro-Winkler
  - Kehitetty nimien vertailuun
  - Laskee yhtenevät kirjaimet ja transpositiot
  - Painottaa nimien alkua
  - JaroWinklerSim(“maija”, “raija”) = 0.867
  - JaroWinklerSim(“maija”, “maria”) = 0.893
- Neuroverkkomenetelmät

Nimien normalisointi

Eerikki -> ERIK
Eric -> ERIK
eerik -> ERIK

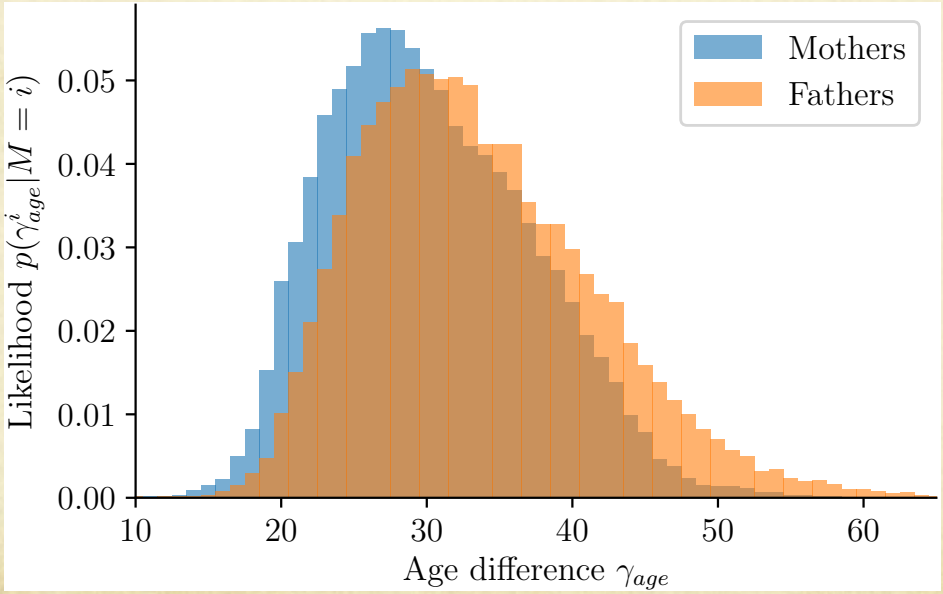
- Mahdollistaa suunnilleen samannimisten henkilöiden hakemisen tehokkaasti
- Nimiryhmien automaattinen päättely
  - Pohjana HisKin nimiryhmät
  - Uusi nimi lisätään lähimpään näistä ryhmistä tai luodaan uusi ryhmä

Sukupuiden päättely

- Aineistona HisKin 5 milj. kastetapahtumaa
- Tehtävä: Linkit kastetapahtumasta vanhempien kastetapahtumiin
- Kastetapahtuma
  - Lapsen nimi + vanhempien nimet
  - Syntymäpaikka
  - Päivämäärä
- Haasteet
  - Duplikaattinimet
  - Puuttuvat tiedot
  - Opetusaineisto / validointi

Sukupuiden päättely

- Todennäköisyys perustuen nimiin, syntymäaikoihin ja paikkoihin
- Todennäköisyysjakaumat opitaan datasta



AncestryAI

- Web-sovellus, jolla voi visualisoida sekä hakea automaattisesti pääteltyjä sukupuita
- Tarkoitettu tukemaan sukututkijan työtä
- Saattaa auttaa erityisesti muualta muuttaneiden henkilöiden löytämisessä
- Laskee todennäköisimmät vanhemmat koko HisKi-aineistolle noin 30 minuutissa

Demo

ancestryai.cs.hut.fi

Kiitos!

- ancestryai.cs.hut.fi
- eric.malmi@aalto.fi
- Auttakaa algoritmin kehityksessä lisäämällä kommentteja
- Kiitokset: Marko Rasa, Pekka Valta, Juha Mäkeläinen, Matti Juhala, ym.