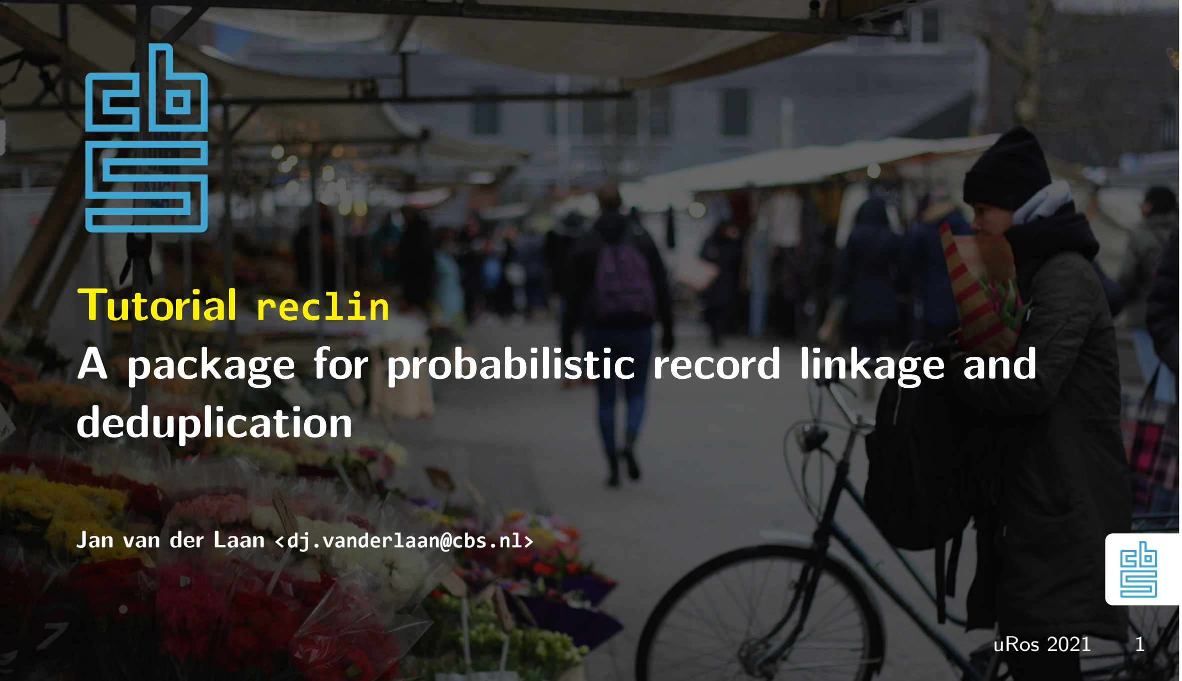


Tutorial reclin

A package for probabilistic record linkage and deduplication

Jan van der Laan <dj.vanderlaan@cbs.nl>

About me

Methodologist/data scientist at Statistics Netherlands (CBS)

Author of a couple of R-packages: `reclin`, `simplermarkdown`, `LaF`, `ldat`, `lvec`

Give multiple courses at CBS: multivariate analysis methods in R, using R in statistical processes.

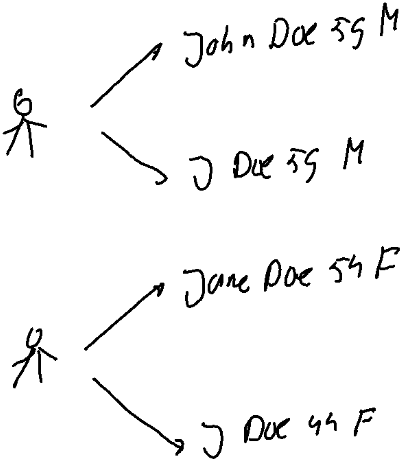
Current projects: social network analysis, measuring social economic status, measuring segregation.

Outline

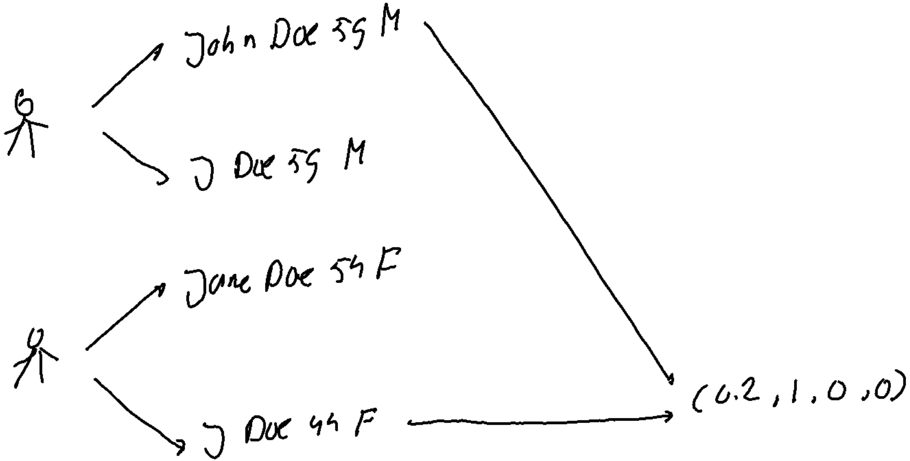
- Overview record linkage
- Traditional record linkage / EM
- String similarity / custom similarity functions
- Machine learning

Each item 10-15 min explanation + 10-15 min working on excercises

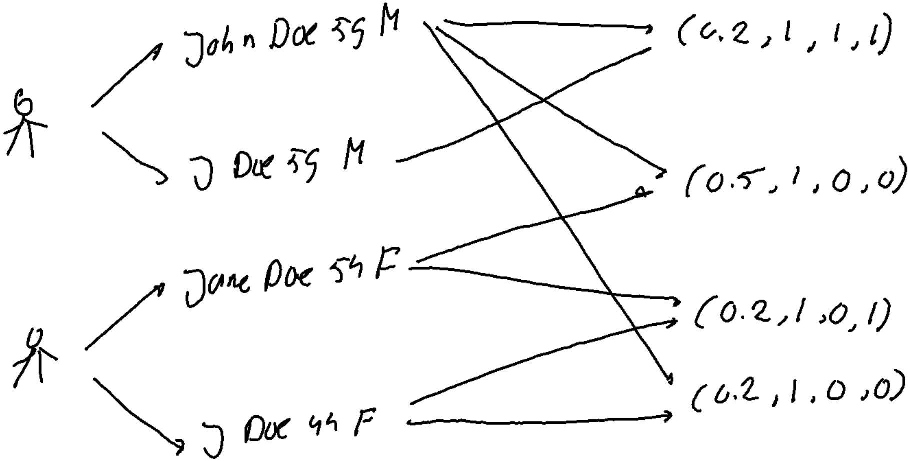
Overview record linkage



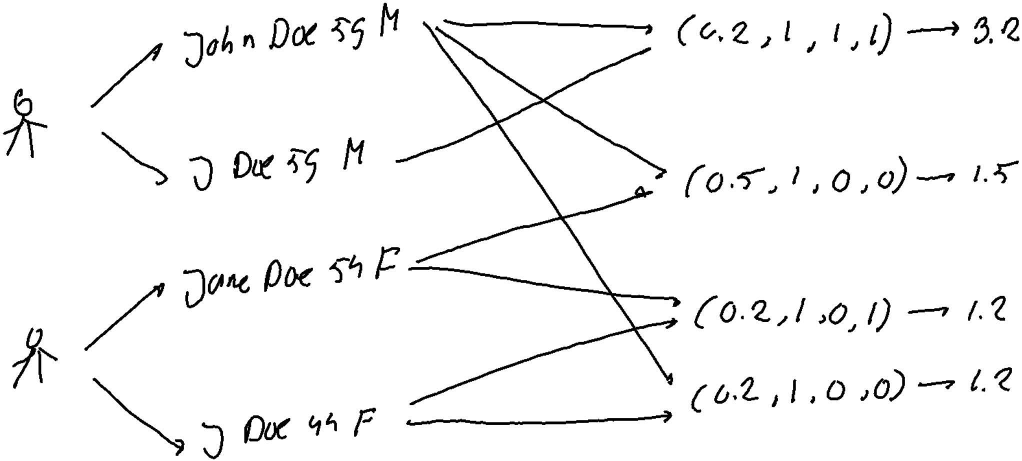
Overview record linkage



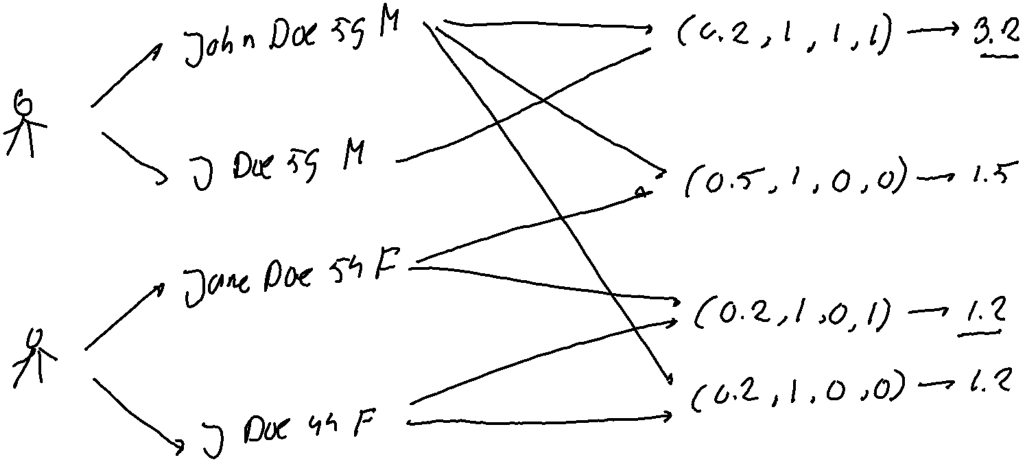
Overview record linkage



Overview record linkage



Overview record linkage



The record linkage process

1. Generate record pairs
   - Blocking
2. Generate comparison vectors
3. Translate comparison vector in a score measuring likelihood of both records in a pairs belonging to the same object.
   - *Classical* probabilistic linkage using EM
   - Machine learning
   - Simple scoring functions
4. Select pairs with a high enough score
5. Generate linked dataset
   - Force one-to-one linkage

Enough with the sheets, let’s go to R ...

Contact and more information

Jan van der Laan <dj.vanderlaan@cbs.nl>

https://cran.r-project.org/package=reclin

https://github.com/djvanderlaan/reclin