ENCODING OF NUMERICAL DATA FOR PRIVACY-PRESERVING RECORD LINKAGE

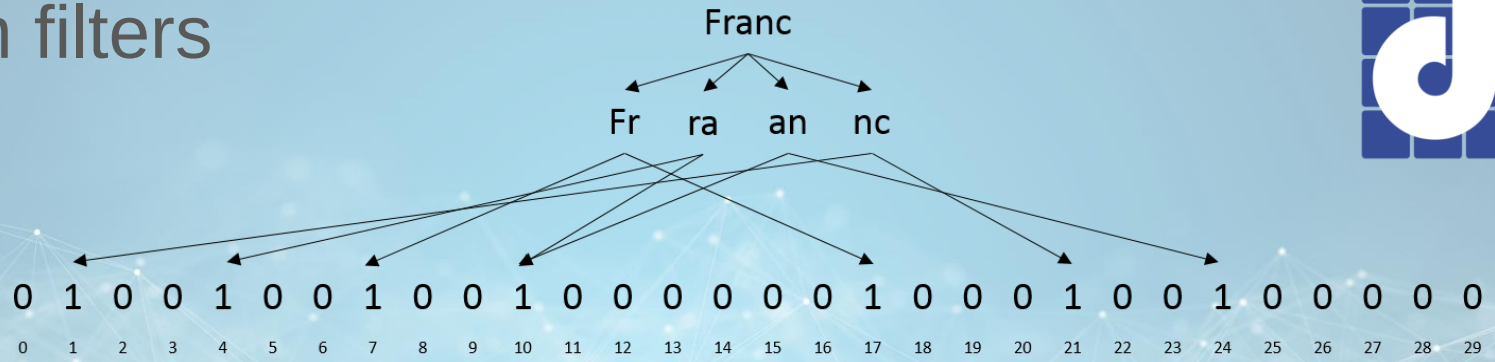
Lea Demelius

Overview

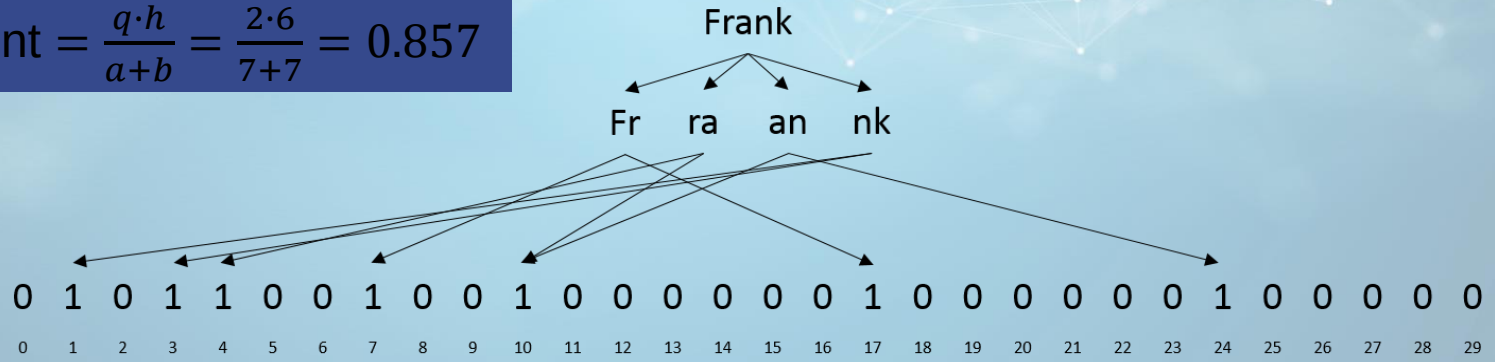
- Introduction
  - Privacy-Preserving Record Linkage
  - Bloom filters and Cryptographic Longterm Keys
  - Why numerical encoding?
- Methods
  - Experimental setup
  - Numerical encoding method
- Results

Privacy-preserving Record Linkage

Bloom filters

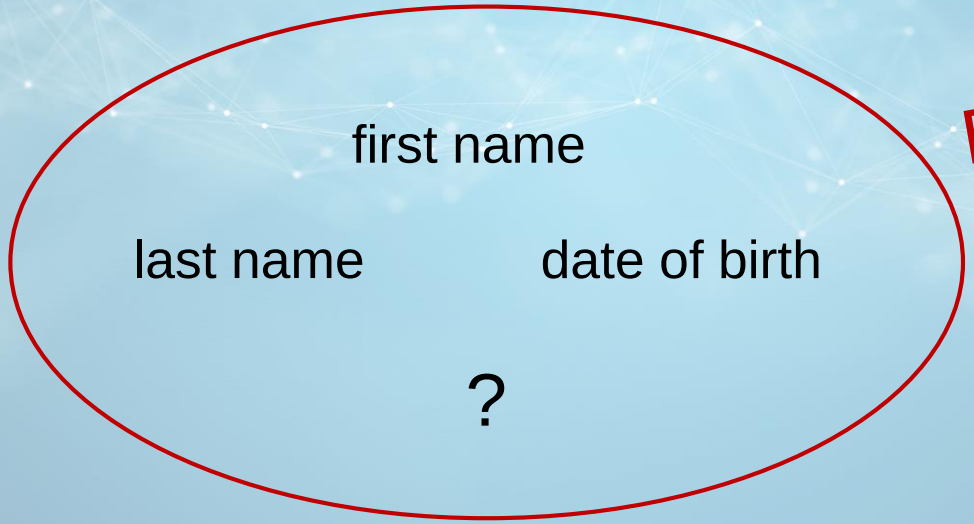


Dice coefficient $= \frac{q \cdot h}{a+b} = \frac{2 \cdot 6}{7+7} = 0.857$



Cryptographic Longterm Keys (CLKs)

= record-level Bloom filters

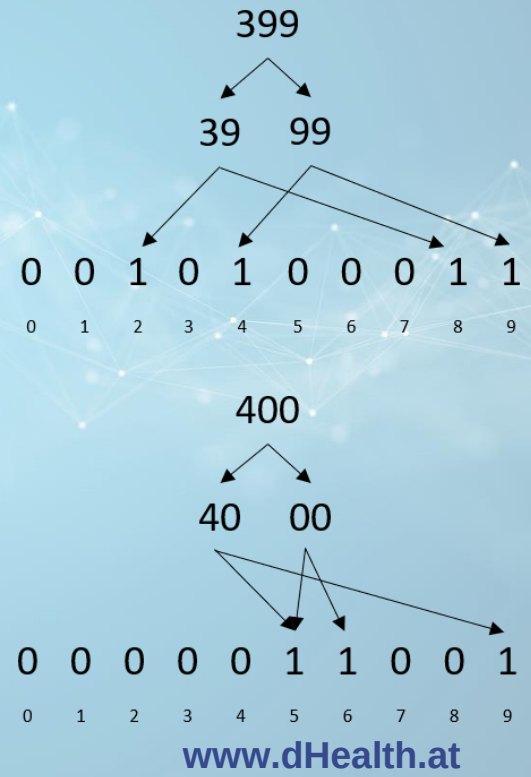


EUPID
(European Patient Identity Service)

An additional identifier: Place of birth

Geocoordinates:

- independent of language
- stable
- allow distance measurements

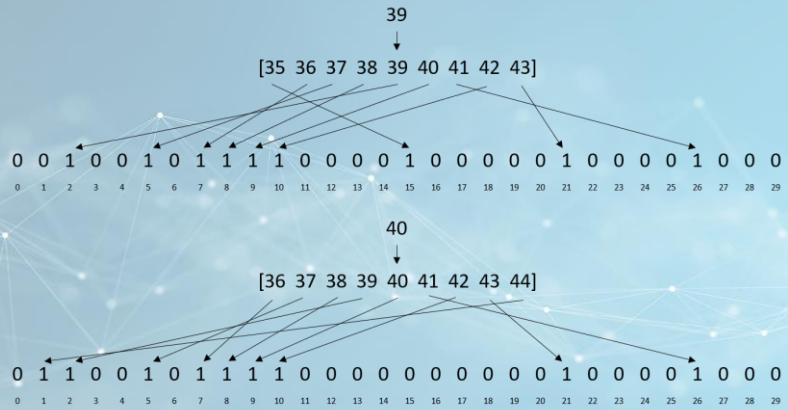


Methods

- Python
- synthetic data (3000 records)
- comparison of 3 methods:
  - string encoding
  - numerical encoding
  - string encoding with shortened geocoordinates

Numerical encoding

Numerical encoding



$\forall i \in [0,2b]$:

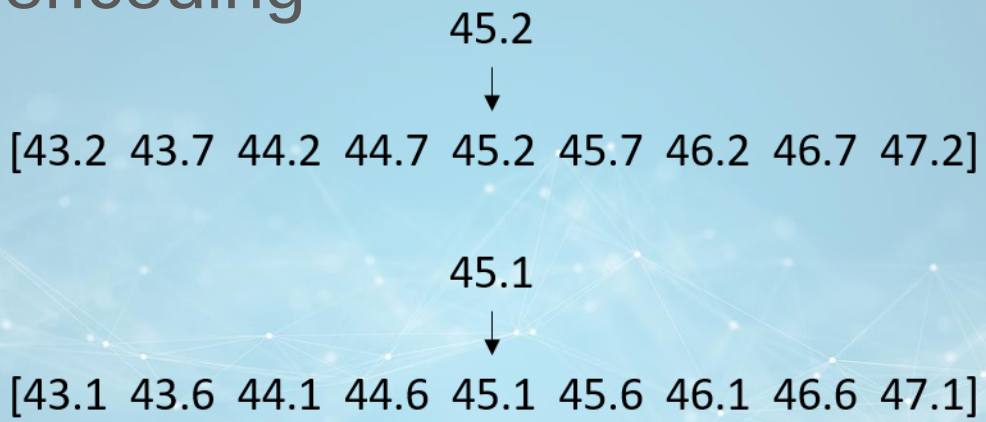
$$x_i = g + (i - b) \cdot d_{intv}$$

g … *geocoordinate*

d_{intv} … *interval width*

$2b + 1$ … *length of list*

Numerical encoding



Numerical encoding

45.2

↓

[43.0 43.5 44.0 44.5 45.2 45.5 46.0 46.5 47.0]

45.1

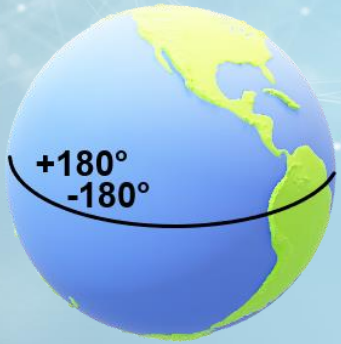
↓

[43.0 43.5 44.0 44.5 45.1 45.5 46.0 46.5 47.0]

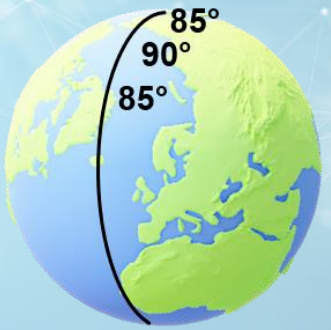
$$new\_x_i = \begin{cases} x_i, & x_i \bmod d_{intv} = 0 \\ x_i - (x_i \bmod d_{intv}), & x_i \bmod d_{intv} < \frac{d_{intv}}{2} \\ x_i + (d_{intv} - (x_i \bmod d_{intv}), & x_i \bmod d_{intv} \geq \frac{d_{intv}}{2} \end{cases}$$

Numerical encoding

longitude

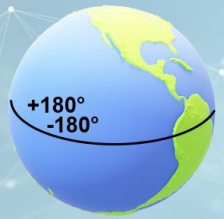


latitude



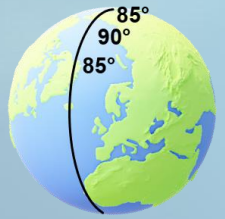
Numerical encoding

longitude



$$final\_x_i = \begin{cases} min + new\_x_i \bmod max, & new\_x_i > max \\ new\_x_i \bmod max, & new\_x_i < min \\ new\_x_i, & else \end{cases}$$

latitude



$$final\_x_i = \begin{cases} 2 \cdot max - new\_x_i, & new\_x_i > max \\ 2 \cdot min - new\_x_i, & new\_x_i < min \\ new\_x_i, & else \end{cases}$$

Results

Test run No.	Test run description	Recall	Precision	F1 score
1	string encoding	0.863	0.977	0.917
2	numerical encoding	0.997	0.997	0.997
3	string encoding with shortened geocoordinates	0.960	0.990	0.975

Results

Numerical Encoding using CLKs leads to:

- better quality
- more privacy

ENCODING OF NUMERICAL DATA FOR PRIVACY-PRESERVING RECORD LINKAGE

Lea Demelius