Privacy Preserving Data Analysis

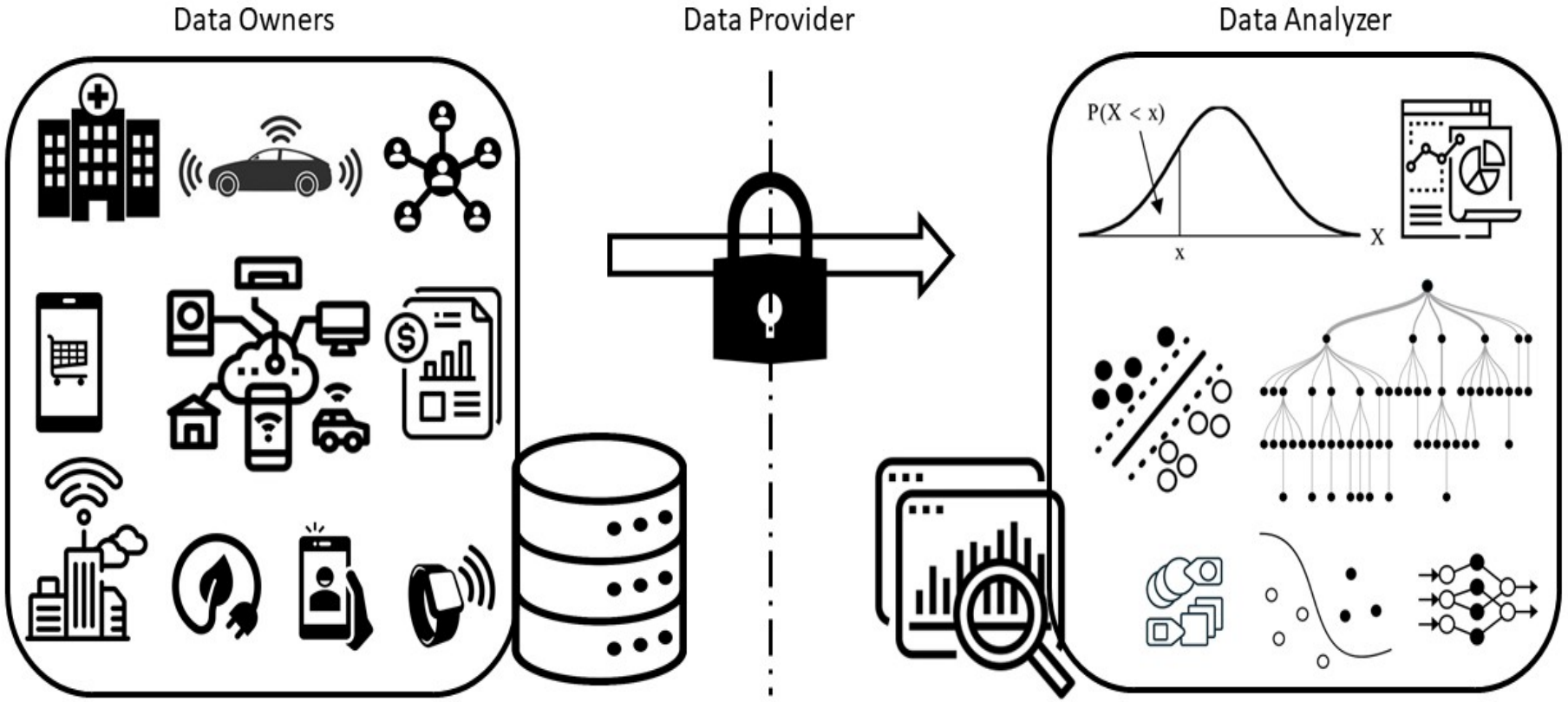
Mina Alishahi
October 10, 2023

Open Universiteit
www.ou.nl

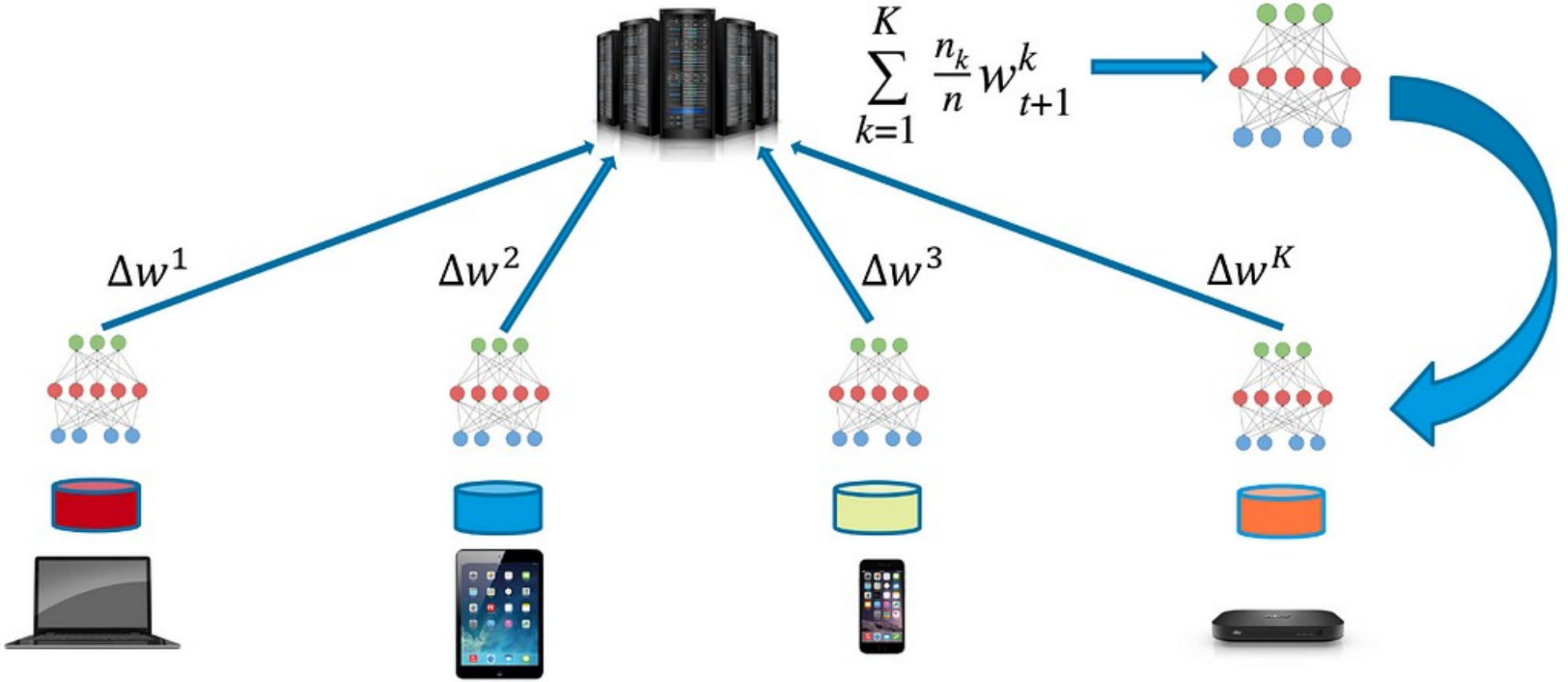
Outline

- What in general I am (was) doing!
- Privacy & Feature selection
- Privacy & Clustering
- Ongoing research
  - Privacy & indoor location data
  - Privacy & ML fairness
  - Game theory in federated learning

General overview



Mainly in federated learning setting

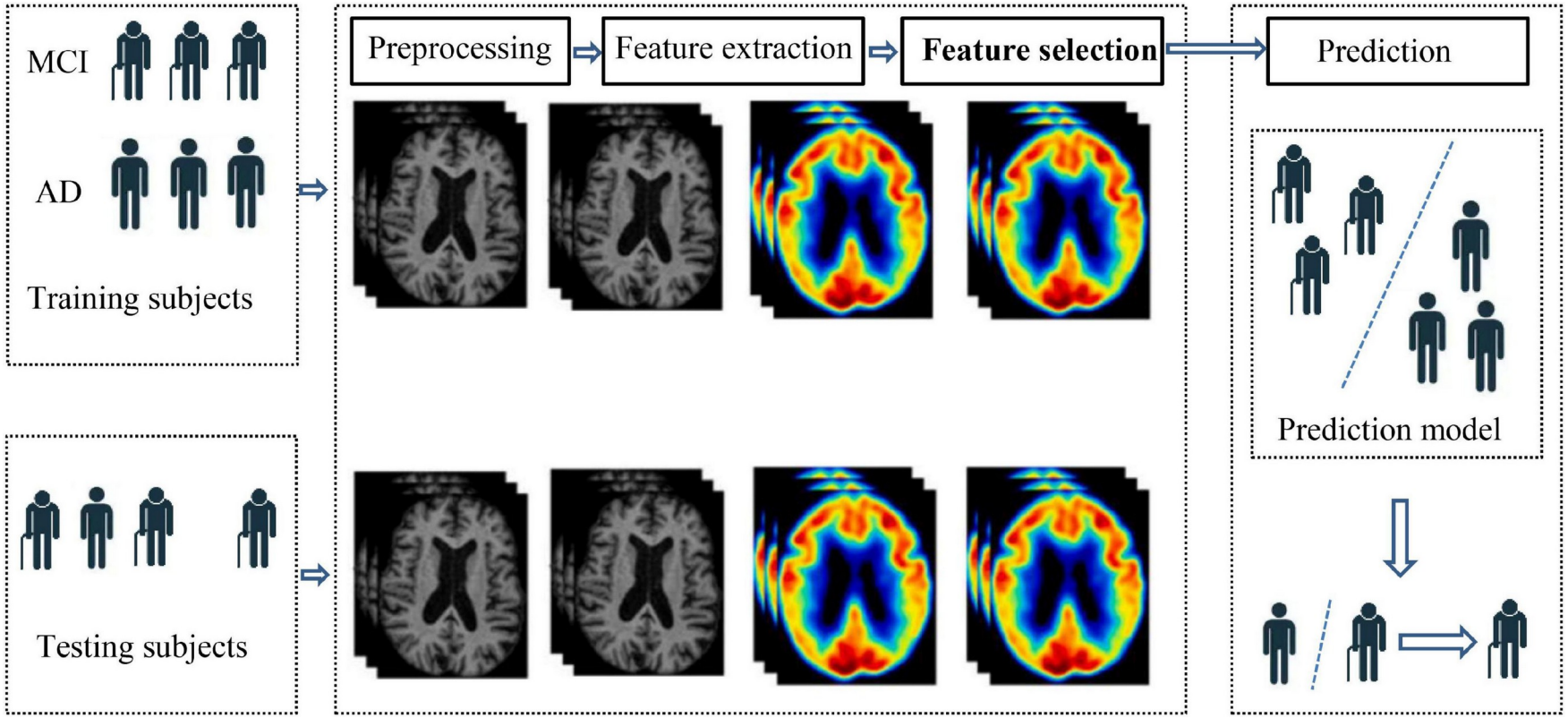


https://proandroiddev.com/federated-learning-e79e054c33ef

Privacy & Feature selection

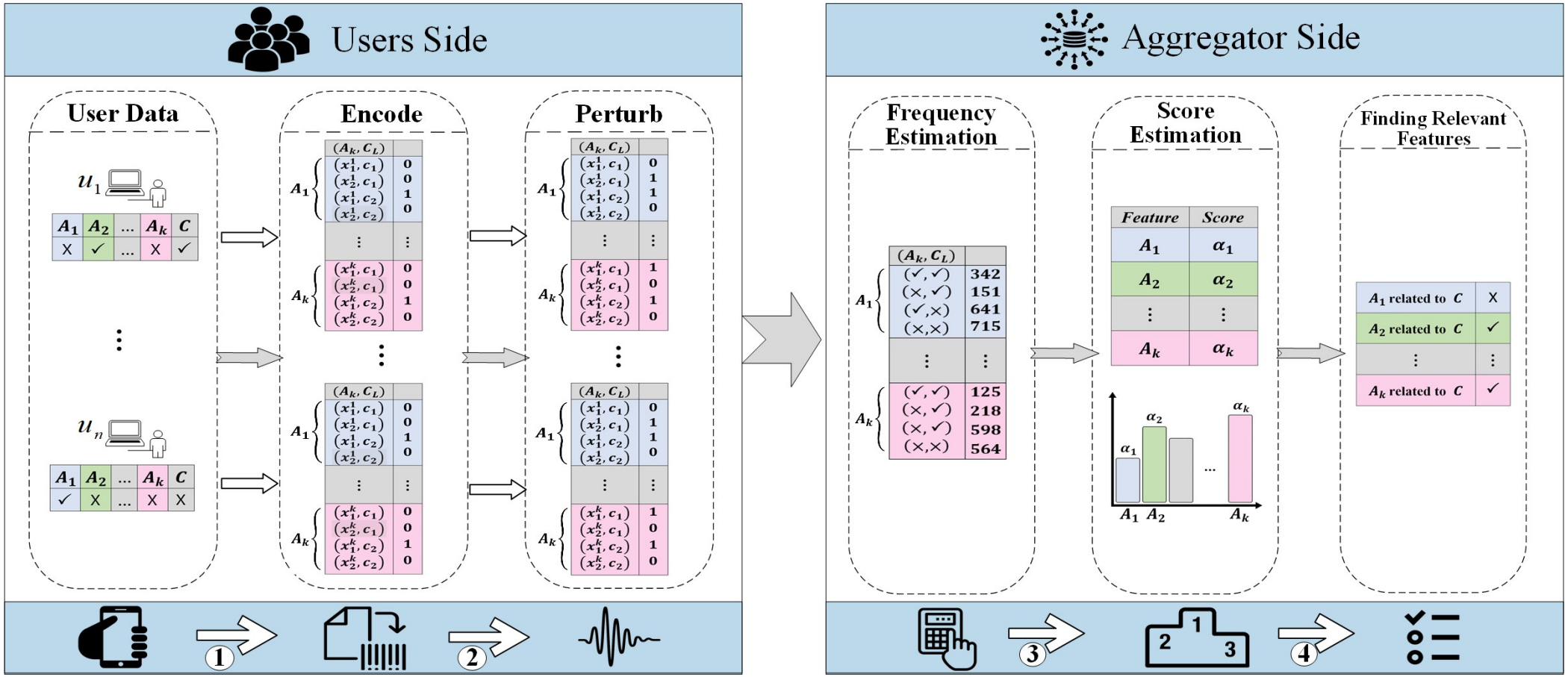
Motivation

What is feature selection?



https://www.frontiersin.org/articles/10.3389/fnagi.2022.924113/full

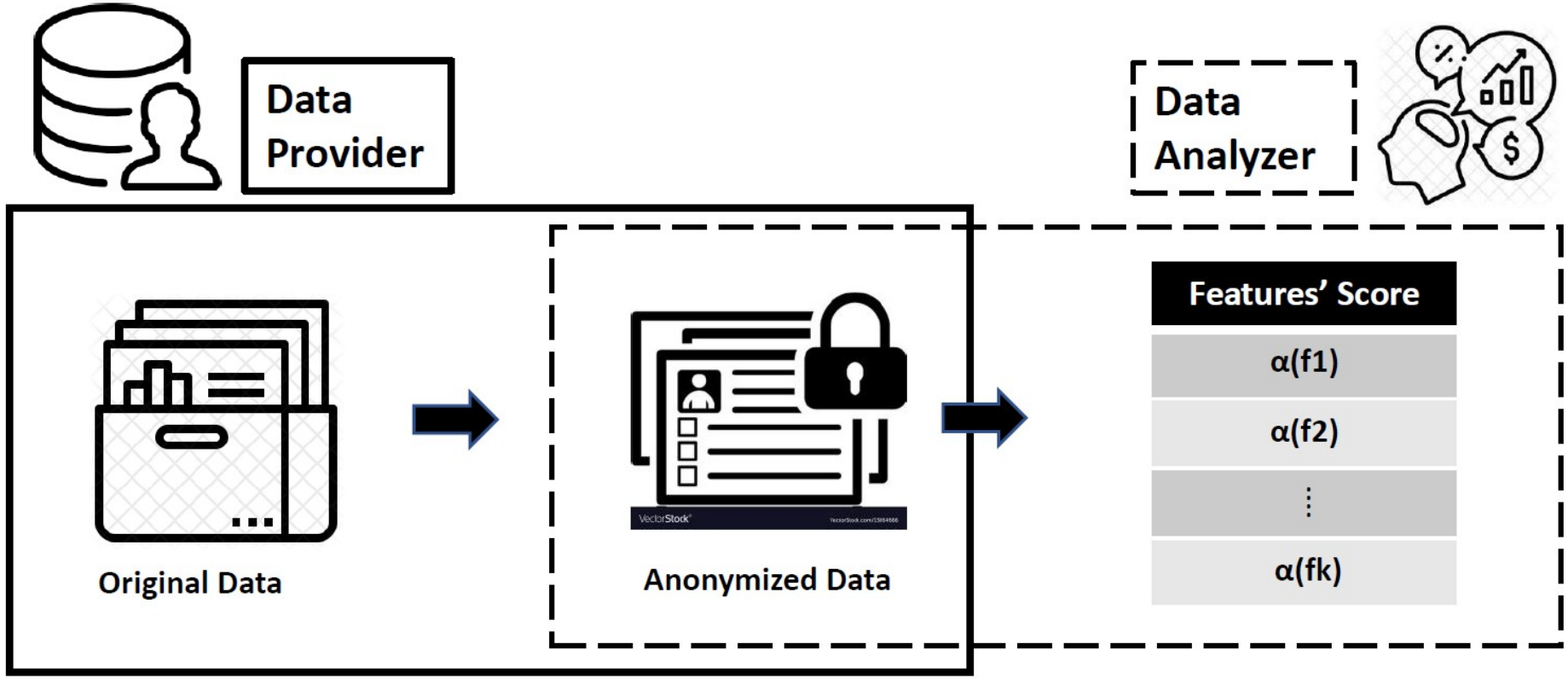
1) Private feature selection using LDP



Add Noise to Remove Noise: Local Differential Privacy for Feature Selection, Mina Alishahi, Vahideh Moghtadaiee, Hojjat Navidan, Computers&security, 2022

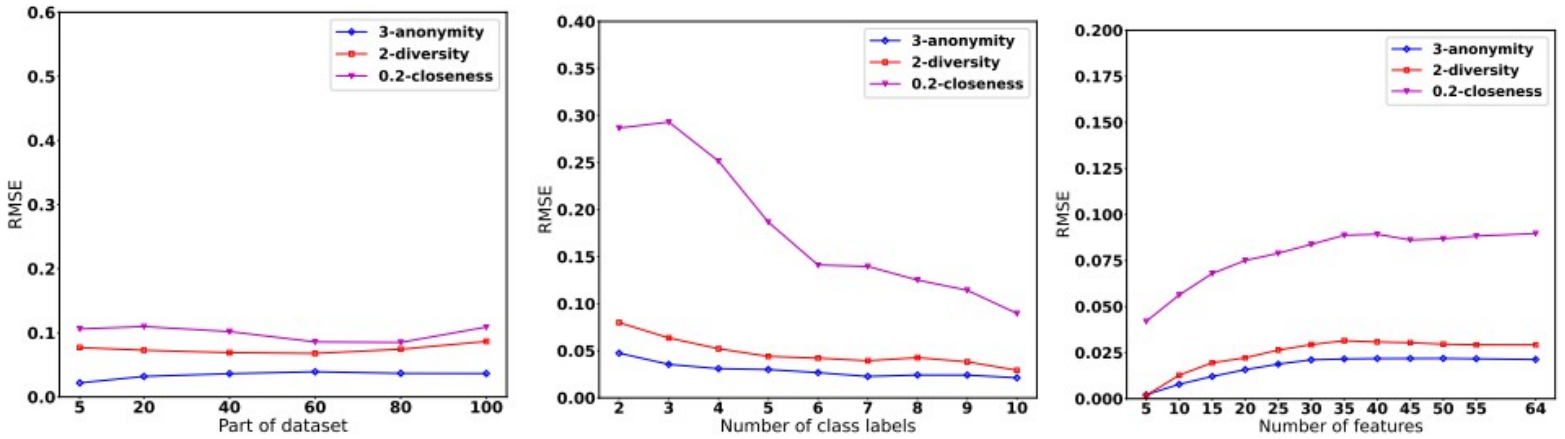
2) Feature selection on anonymized dataset

Person	First name	Account type	Subscription date	Tickets submitted
1	Luke	Pro	13 May 2017	2
2	John	Enterprise	25 Feb 2016	3
3	Nathan	Free	17 Sep 2014	5
4	Aaron	Free	2 May 2018	2
5	Daniel	Pro	13 Aug 2018	0
6	Michael	Pro	13 Dec 2018	1



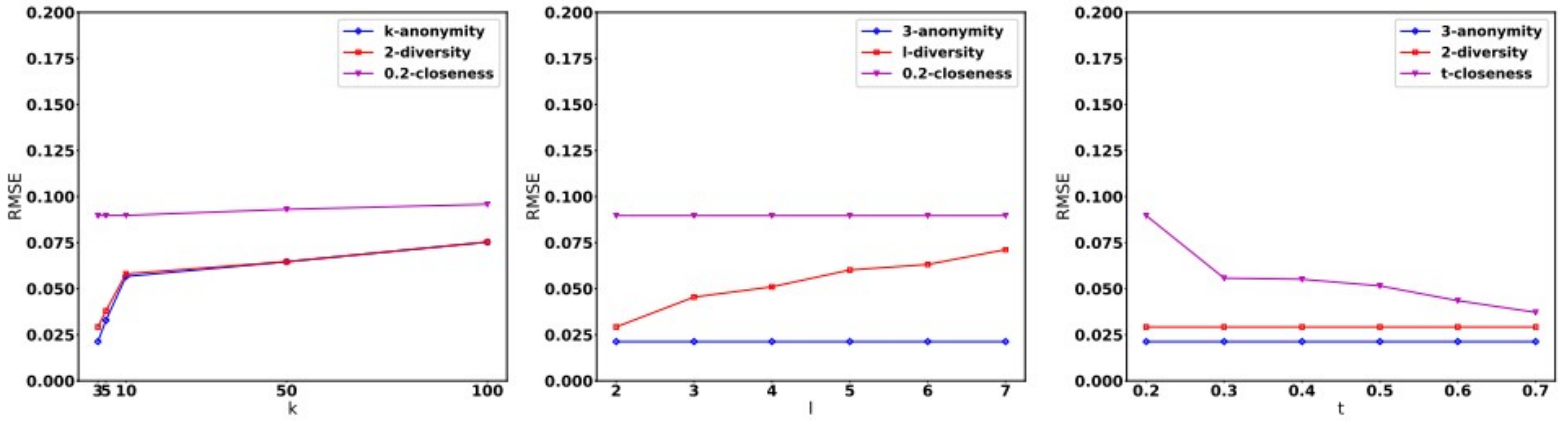
Feature Selection on Anonymized Datasets, Mina Alishahi, Vahideh Moghtadaiee, The 21st IEEE International Conference on Dependable, Autonomic & Secure Computing, (DASC 2023)

2) Feature selection on anonymized dataset



(a) Dataset size (b) Number of class labels (c) Number of features

Fig. 7: The RMSE of Pearson correlation on the Optical Digits dataset when dataset properties are varied



(a) (b) (c)

Fig. 8: The RMSE of Pearson correlation for various values of a) k, b) ℓ, and c) t on the Optical Digits dataset

2) Feature selection on anonymized dataset

Findings:

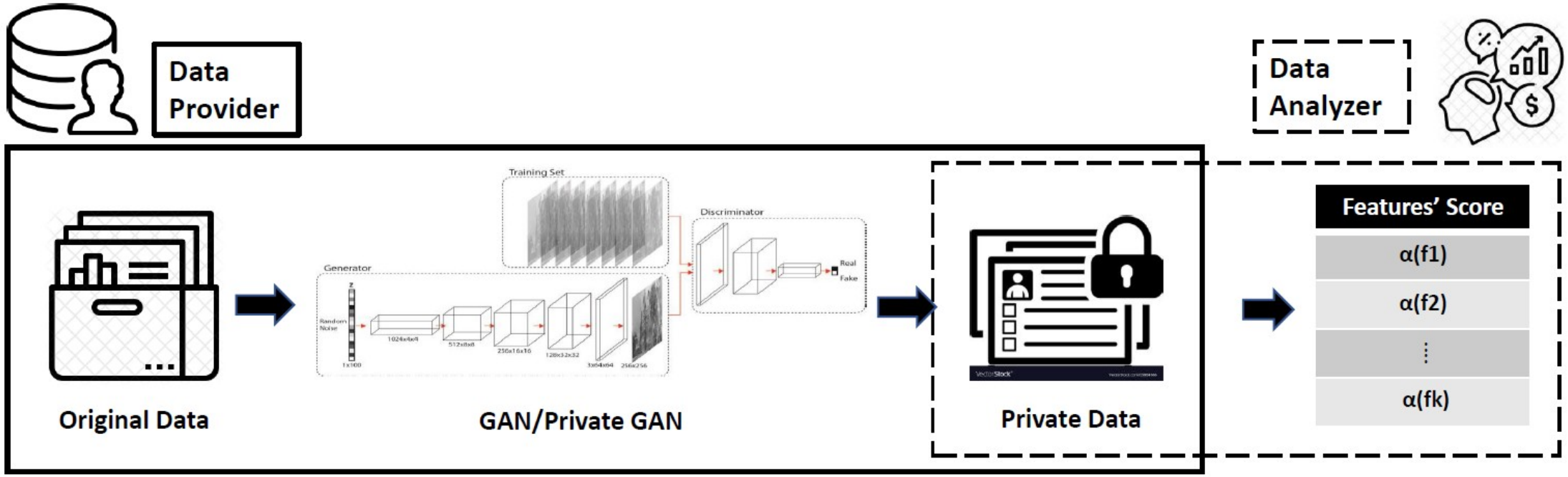
- In general, all feature selection techniques perform worse on anonymized versions of categorical datasets rather than numerical and mixed ones.

- The size of the dataset shows no impact on preserving the features' significance in anonymized datasets. The increment in the number of class labels improves the accuracy of t-close datasets, while it does not show considerable effect on k-anonymous and ℓ-diverse datasets. Increasing the number of features has a negative impact on preserving the features' significance in anonymized datasets.

- While increasing k after a threshold shows negligible influence on preserving the features' importance in k anonymous dataset, the increment of ℓ has a direct impact in reducing the accuracy. On the other hand, by increasing t, the error rate is reduced.

- When it comes to training the accurate classifiers on a subset of features selected in anonymized datasets, the multi-labeled datasets are not a desirable source of input.

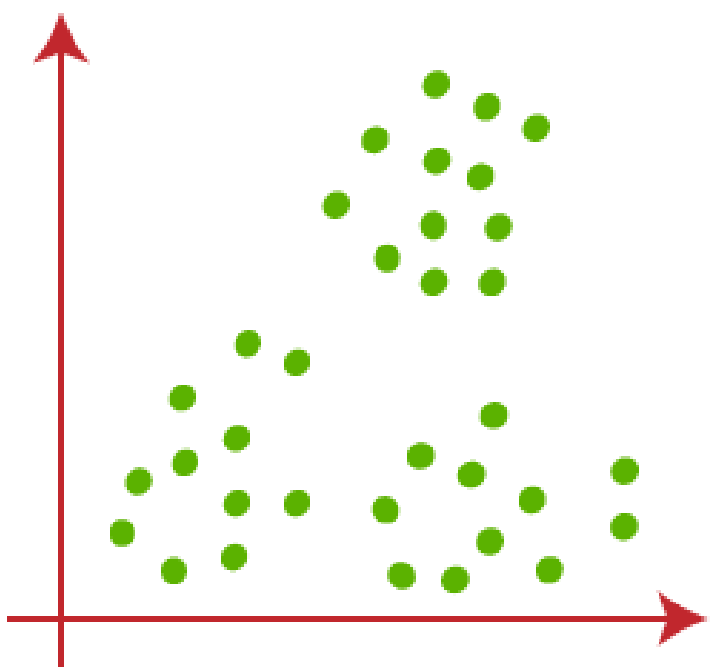
3) Feature selection & GAN

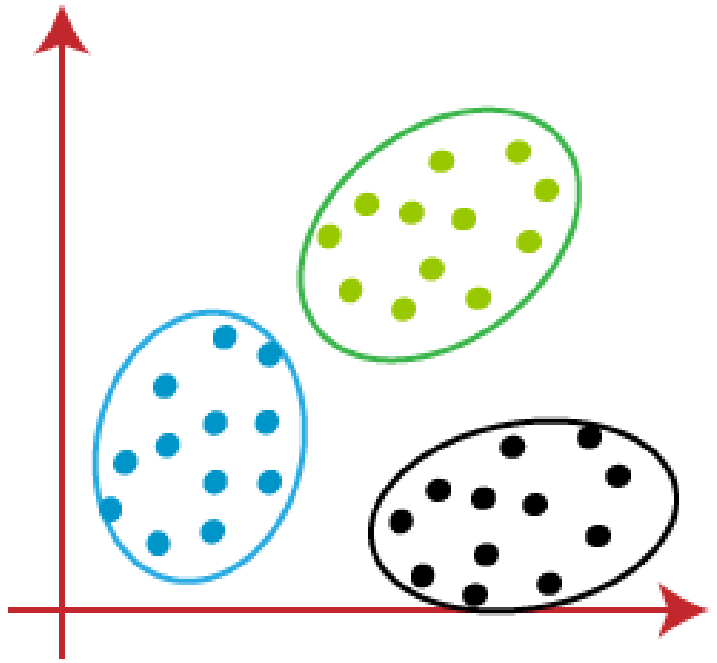


4) A survey on privacy-preserving feature selection

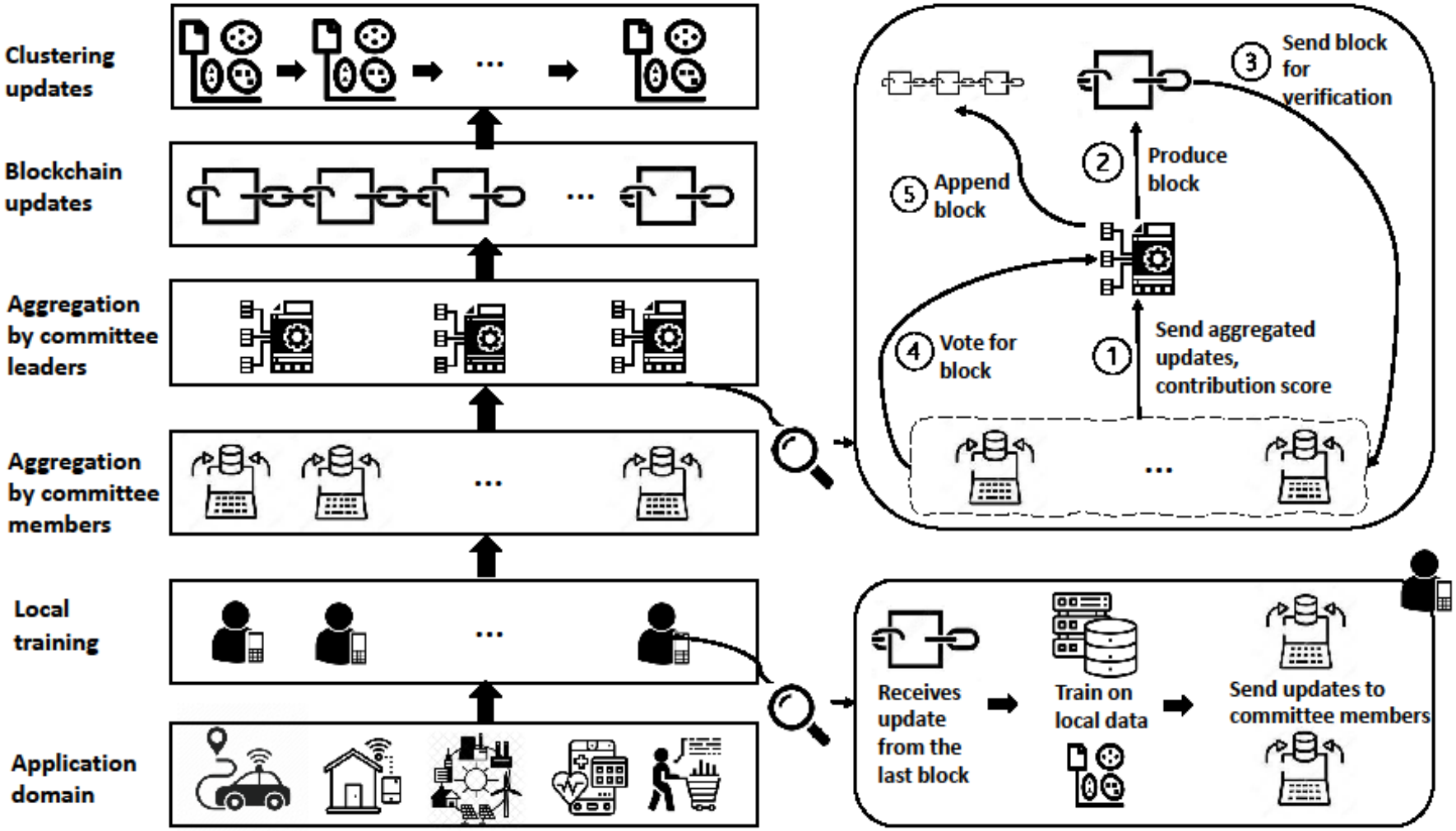
Privacy & Clustering

What is clustering?



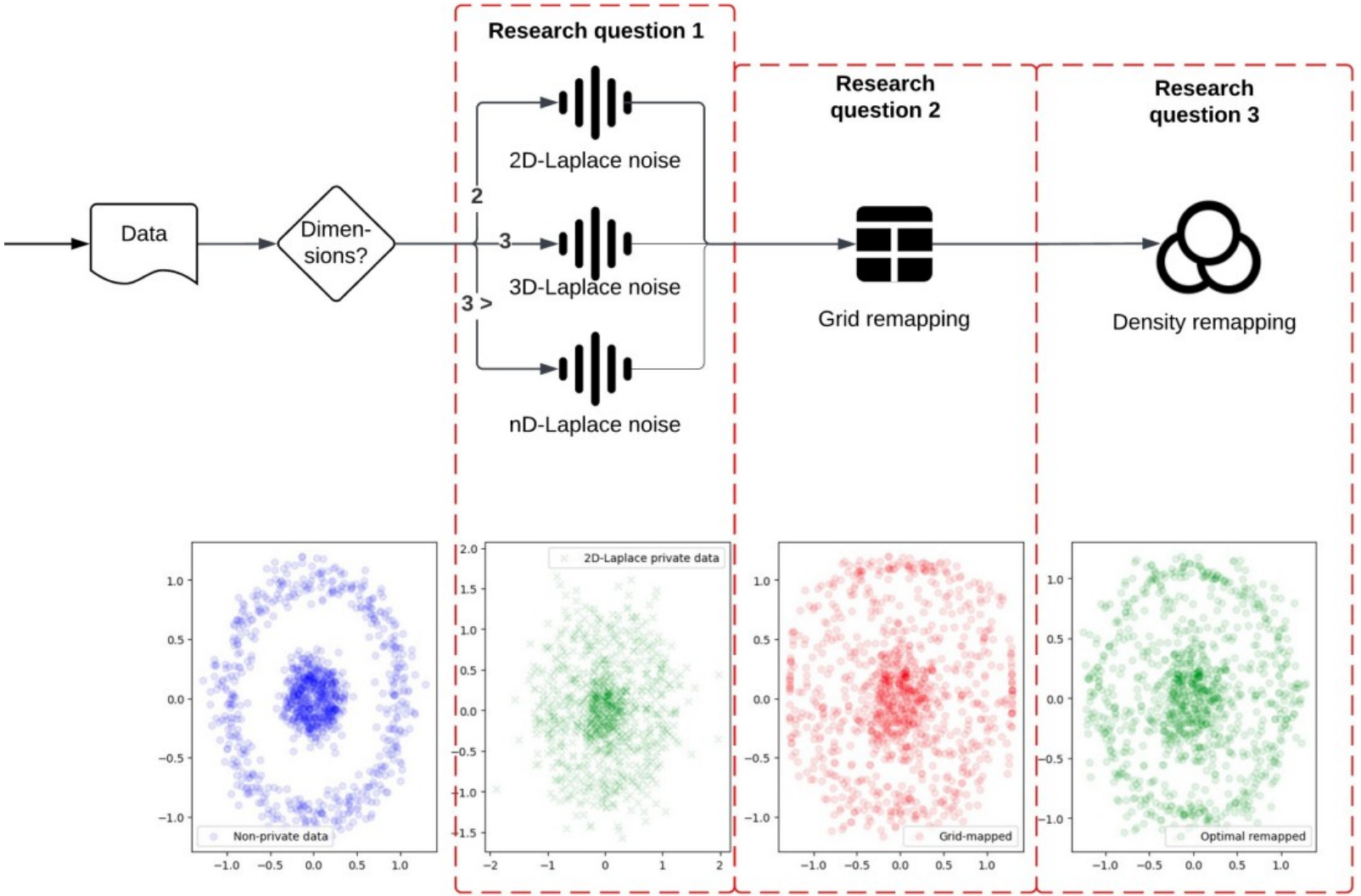


1) Blockchain & Federated clustering

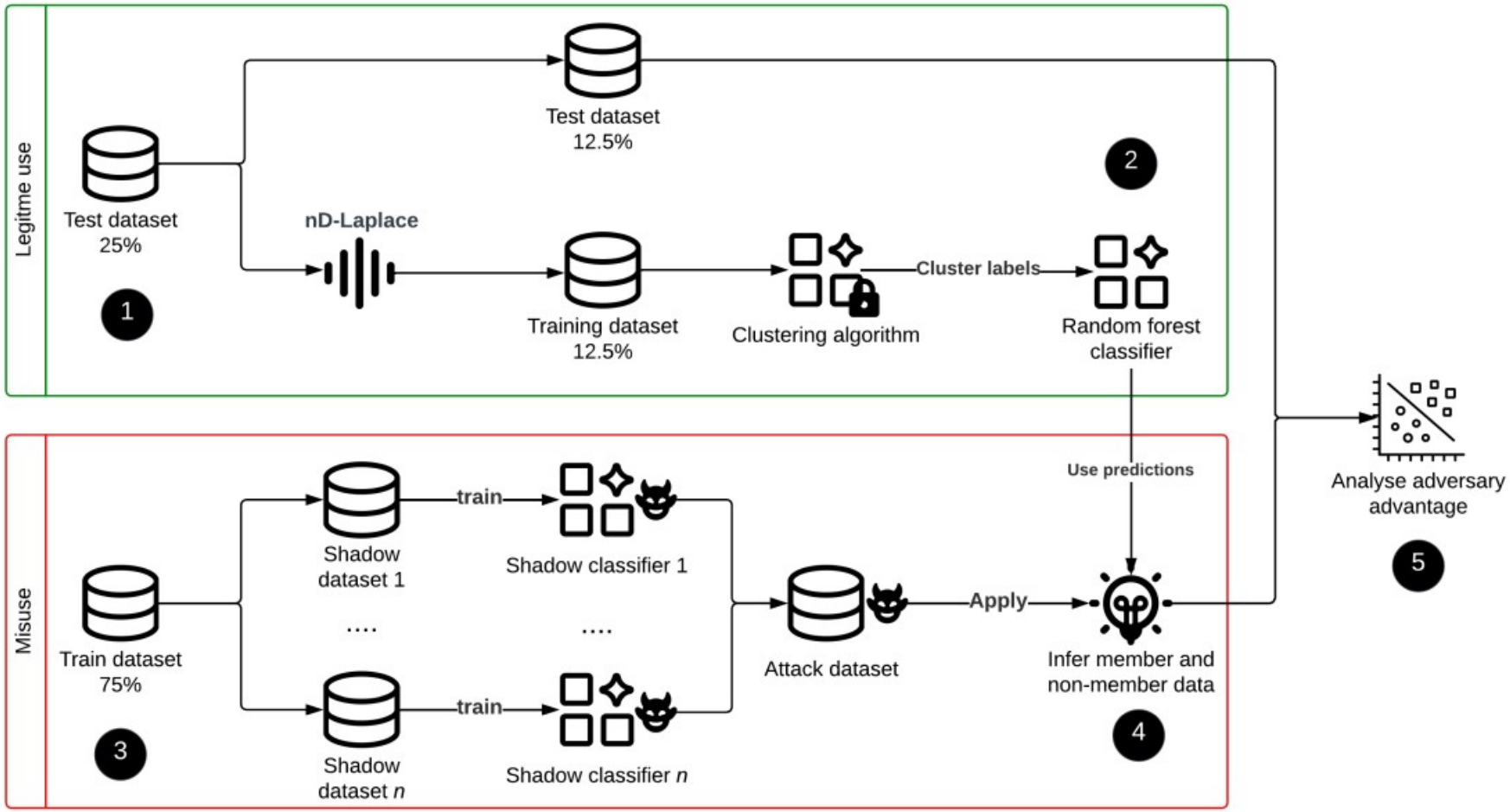


BC-FL k-means: A Blockchain-based Framework for Federated Clustering, Mina Alishahi, Wouter Leeuw, Nicola Zannone, TrustCom 2023

2) Geo-indistinguishability for clustering



2) Geo-indistinguishability for clustering



Ongoing research ....

1) Privacy & indoor location positioning

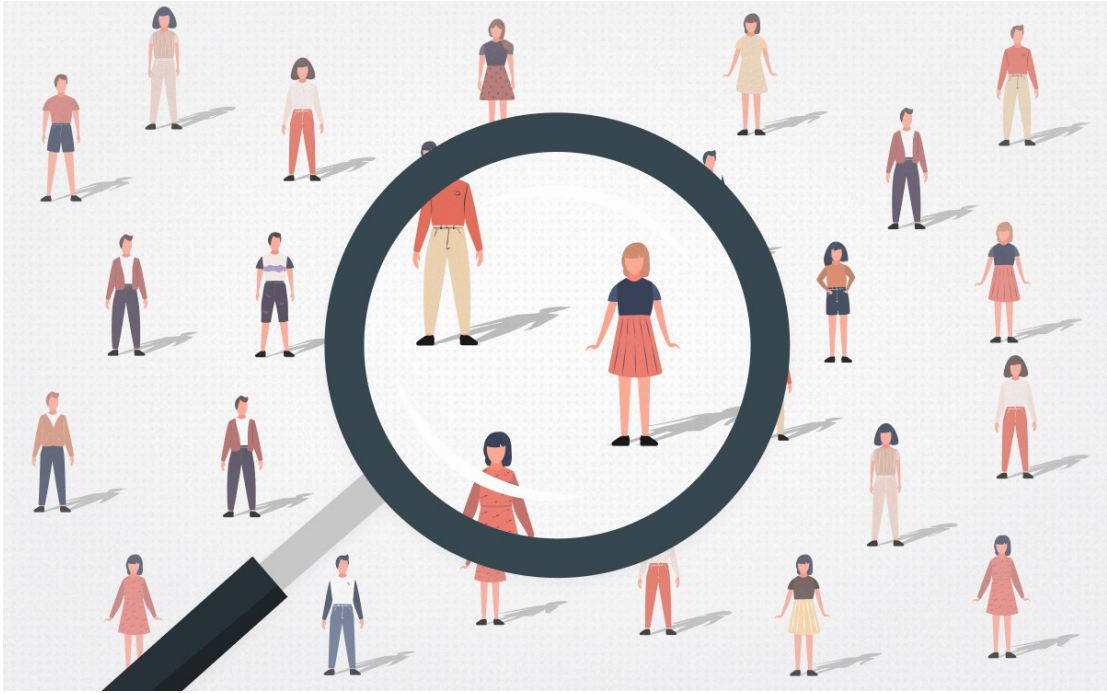
- On the privacy protection of **indoor location** dataset using **anonymization**, Computers&security, 2022

- Indoor Geo-Indistinguishability: Adopting **Differential Privacy** for **Indoor Location** Data Protection, IEEE Transactions on Emerging Topics in Computing, 2022

- Hide me Behind the Noise: Local Differential Privacy for Indoor Location Privacy, IEEE S&P workshop

2) Privacy & ML Fairness

rivate evaluation of fairness
enerate datasets that respect privacy and fairness requirements
earn unbiased ML models over distributed private data (fair federated learning)



3) Game theory in federated learning

- Federated learning that balances utility, privacy, trust, and fairness

Question?

mina.sheikhalishahi@ou.nl

www.minaalishahi.com