August 23, 2024

# PRIVACY ENHANCING TECHNOLOGIES

# PETs
## Definition

Technologies that allow to utilise personal data while minimising privacy risks

Office, I. C. 2023. Privacy-enhancing technologies (PETs). en. 2023 https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf

# PETs
## Definition

Technologies that allow to utilise personal data while minimising privacy risks

# PETs

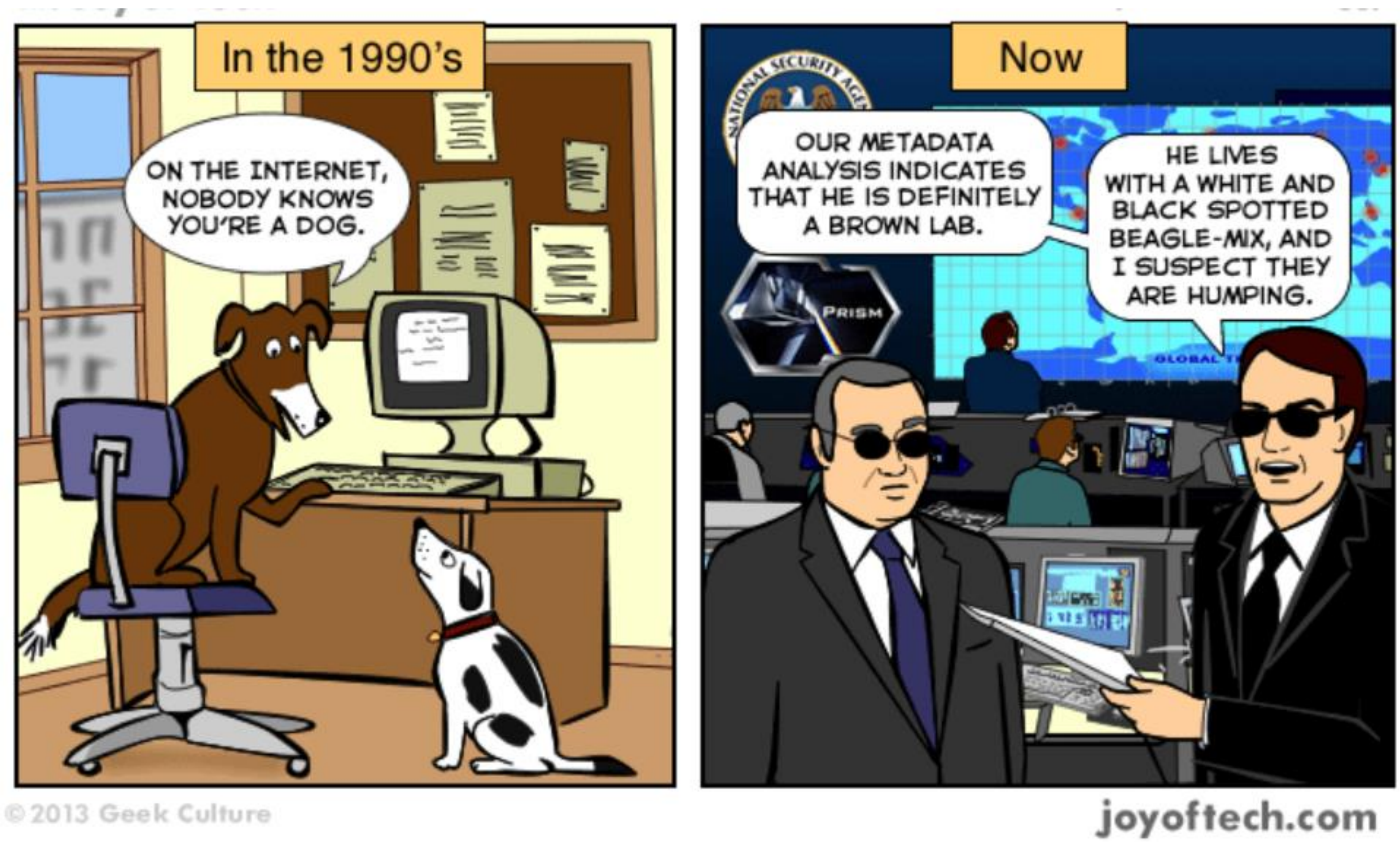
- The concept of PETs was first discussed in a report published by the Information and Privacy Commissioner of Ontario, Canada (1995) "Privacy-Enhancing Technologies: The Path to Anonymity"

- From the end-user perspective : any technology that allow them to interact with digital systems in a privacy-preserving manner

- From the organisational perspective : any technology that allow them to develop a privacy-enhancing product or a service

- Every PET **does not** guarantee anonymisation, but they enhance data privacy by minimizing the personal data usage

# Anonymity


In the 1990's
ON THE INTERNET, NOBODY KNOWS YOU'RE A DOG.
Now
OUR METADATA ANALYSIS INDICATES THAT HE IS DEFINITELY A BROWN LAB.
HE LIVES WITH A WHITE AND BLACK SPOTTED BEAGLE-MIX, AND I SUSPECT THEY ARE HUMPING.
PRISM
© 2013 Geek Culture
joyoftech.com

# Anonymity

- The state of being unknown / undetectable

- Anonymisation – the technique of removing the link between data and the data subject

Example

| Name | Age | Disability |
|---|---|---|
| Kate | 21 | Vision impairment |

→

| Name | Age | Disability |
|---|---|---|
| ILBer# | 20-30 | Vision impairment |

- Is anonymous data governed by GDPR?

- Can you think of negative implications of anonymity in the digital world?

# PETs

## How do PETs achieve data protection

- Transform data or derive new data
- Hide or shield data
- Split datasets

Office, I. C. 2023. Privacy-enhancing technologies (PETs). en. 2023 https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf

# PETs

## Some Popular PETs

- Pseudonymisation
- K-Anonymity
- Differential Privacy
- Federated Learning
- Homomorphic Encryption
- Zero-Knowledge Proof
- Synthetic Data

There are more…….any technique that fits the definition

# PETs

## Some Popular PETs

- Pseudonymisation
- K-Anonymity
- Differential Privacy
- Federated Learning
- Homomorphic Encryption
- Zero-Knowledge Proof
- Synthetic Data

# PETs

## Pseudonymisation

- Converting personal data in a way that data can no longer be linked to a person, without using additional information

- Why it is different from anonymisation?

- 2 Steps (Most of the time)

  1. Replace direct identifier(s) with an artificial identifier -> Always
  2. Save actual identifiers in a separate secure table (mapping table) -> Depends on technique used in 1

- Mapping table and pseudonimised tables are linked through the pseudonym column

# PETs
## Pseudonymisation

| Name | Age | ZIP | Gender | Disease |
|---|---|---|---|---|
| Jhon | 29 | 1981 | M | Cancer |
| Kate | 22 | 1980 | F | Infection |
| Alice | 35 | 1975 | F | AIDS |
| Ellen | 38 | 1978 | F | AIDS |
| Mary | 41 | 1974 | F | Cancer |

Raw Data

| ID | Age | ZIP | Gender | Disease |
|---|---|---|---|---|
| P001 | 29 | 1981 | M | Cancer |
| P002 | 22 | 1980 | F | Infection |
| P003 | 35 | 1975 | F | AIDS |
| P004 | 38 | 1978 | F | AIDS |
| P005 | 41 | 1974 | F | Cancer |

Pseudonymised Data

| ID | Name |
|---|---|
| P001 | Jhon |
| P002 | Kate |
| P003 | Alice |
| P004 | Ellen |
| P005 | Mary |

Needs Security

Mapping Table

# PETs

## Generate the pseudonyms

- Counter

  Replace identifiers with a number or a string

  Increment by 1 for each identifier

  Each identifier has a unique pseudonym : No collisions

| ID | Name |
|---|---|
| P001 | Jhon |
| P002 | Kate |
| P003 | Alice |

| ID | Name |
|---|---|
| 0 | Jhon |
| 1 | Kate |
| 2 | Alice |

| ID | Name |
|---|---|
| 10@gmail.com | jhon@gmail.com |
| 11@abc.ac.nz | bob@abc.ac.nz |
| 12@de.nz | mary@de.nz |

# PETs

## Generate the pseudonyms

- Random Number Generator (RNG)
  - Produce values within a set
  - Collision issue

```
import random

for i in range(10):
    print(random.randint(0,9))
```

| ID | Name |
|---|---|
| 9 | Jhon |
| 6 | Kate |
| 3 | Alice |
| 9 | Mary |
| 4 | Dirk |

# PETs

## Generate the pseudonyms

- Cryptographic Hashing
  - Create a fixed-length output from an arbitrary-length input
  - Cannot be reversed to the original value
  - No collisions for different inputs

```python
import hashlib

# identifiers
names = ["Jhon", "Kate", "Alice", "Mary", "Dirk"]

# hashing using SHA256
for name in names:
    result = hashlib.sha256(str.encode(name))
    print(result.hexdigest())
```

| ID | Name |
|---|---|
| ee0ace6e8f5dc17dc271cb6e7c0cdc2de39f84c84541461b95ed7c59414becf0 | Jhon |
| 1a5d06a170dde413475957ca2b63396aa5e8fbecb7d379fcb668da3753fca5b6 | Kate |
| 3bc51062973c458d5a6f2d8d64a023246354ad7e064b1e4e009ec8a0699a3043 | Alice |
| aebac53c46bbeff10fdd26ca0e2196a9bfc1d19bf88eb1efd65a36151c581051 | Mary |
| 764e8ab23aba697ce8365352a91b8e3c57b8f6672c6873f7d9de17f254a31cce | Dirk |

# PETs

## Generate the pseudonyms

- Message Authentication Code (MAC)

  Create a fixed-length output from an arbitrary-length input

  A secret key is needed

  No collisions for different inputs

```python
import hmac
import hashlib

# identifiers
names = ["Jhon", "Kate", "Alice", "Mary", "Dirk"]

# secret key
key = "c189026b-62b0-43946-8a38*e911"

# using HMAC (Keyed Hash MAC)
for name in names:
    result = hmac.new(str.encode(key), str.encode(name), hashlib.sha256)
    print(result.hexdigest())
```

| ID | Name |
|---|---|
| 9c2975017f9a1c469c35e44346b1b728b5a260c86bbfb5e7cac1c941f63af3d4 | Jhon |
| 5d5993a32fb5b290791bcdc87fdda58ef0c71b60ea104a43eb7c38a8260b0fa2 | Kate |
| b09a9911748cfe38306191122d8f3c2b363cac875b8c1920e7406b33628880f7 | Alice |
| 99cd4f78b2fdbdd01bd5d0c44651ab4201a64cf97ae576ac4b813c3bb18cb2d2 | Mary |
| f7ed755df953d154a9953d521adfcd6956f7bd2ef31be6e4374158a7f058f513 | Dirk |

# PETs

## Generate the pseudonyms

- Encryption

  Two way

  A secret key is needed

  No collisions

    Deterministic encryption on different values
    Non-deterministic encryption

```
from Crypto.Cipher import AES
from Crypto.Random import get_random_bytes

# identifiers
names = ["Jhon", "Kate", "Alice", "Mary", "Dirk"]

#secret key
key = get_random_bytes(16)

cipher = AES.new(key, AES.MODE_CTR)

#using AES128 encryption
for name in names:
    ciphertext = cipher.encrypt(str.encode(name))
    print(ciphertext.hex())
```

| ID | Name |
|---|---|
| 6ae9b676 | Jhon |
| 4672d45f | Kate |
| 87967253b7 | Alice |
| dbaa6bf9 | Mary |
| 018f1600 | Dirk |

# PETs

## Think……..

- Does all Pseudonymisation techniques require a mapping table?

- Scalability

- What if the same identifier is repeating in multiple tables?

    Ex: A university student can also be in the university employee table

    1. Deterministic Pseudonymisation
    2. Document Randomised Pseudonymisation
    3. Fully Randomised Pseudonymisation

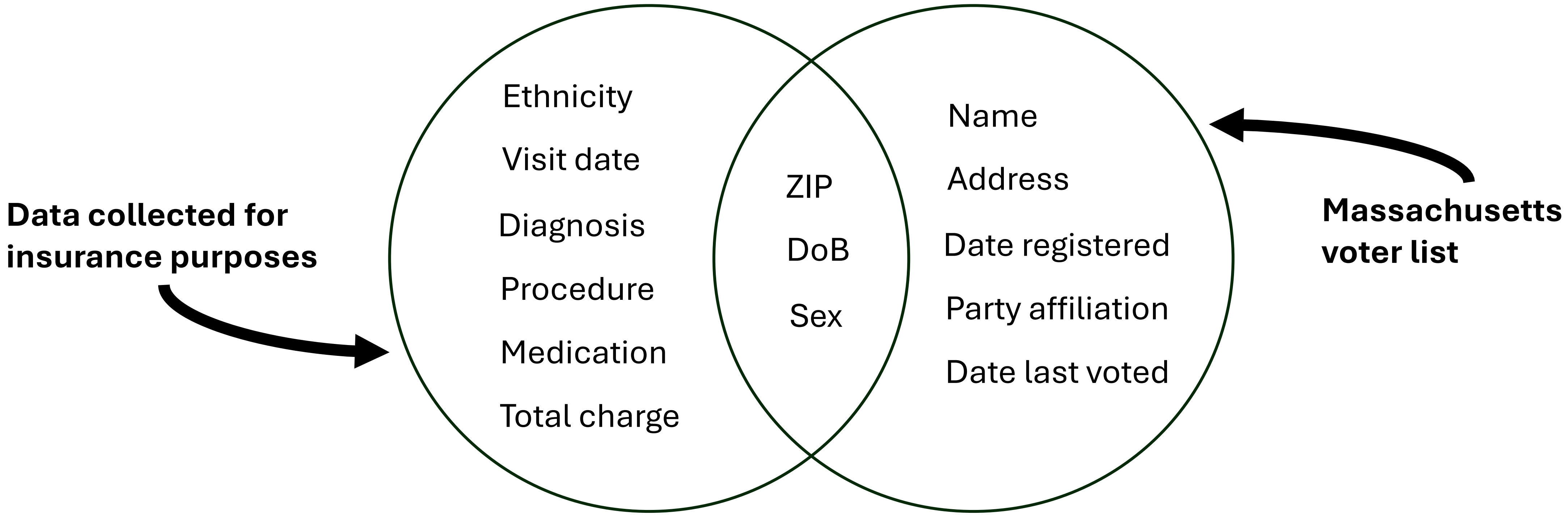
- Can we represent multiple identifiers from one pseudonym

    {name, driver_licence , email}

# ____________ can escape GDPR. Which is/are correct?

| 0% | 0% |
|---|---|
| ✅ Anonymisation | ❌ Pseudonymisation |

# Linkage Attacks




*"87% of the US population can be uniquely identified by only {ZIP, DoB, Sex} data fields"*

- Latanya Sweeney

Latanya Sweeney. 2002. K-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10, 5 (October 2002), 557–570. https://doi-org.ezproxy.auckland.ac.nz/10.1142/S0218488502001648

# PETs

## K-Anonymity

- Making an individual indistinguishable from k-1 individuals in the same dataset : “hiding in a crowd”

- It is achieved through generalisation or suppression of quasi-identifiers

- Terminology

  Quasi-identifiers - data attributes, excluding PII, that in combination can uniquely identify a person
  (ex: age, zip code, race, and gender of a patient)

  Generalisation - replacing an individual data value with a more general value
  (ex: age 20 replaced by age range 20-30)

  Suppression - removing data values in a dataset to achieve anonymity
  (ex: Zip 1234 replaced by 123*, removing an entire column)

- K-anonymity provided syntactic anonymisation

# PETs

## K-Anonymity

| ID | Age | Zip | Gender | Disease |
|---|---|---|---|---|
| P001 | 22 | 2141 | M | Cancer |
| P002 | 24 | 2141 | F | Infection |
| P003 | 31 | 2138 | F | AIDS |
| P004 | 32 | 2139 | F | AIDS |
| P005 | 41 | 2243 | M | Cancer |
| P006 | 41 | 2245 | M | Infection |
| P007 | 48 | 6534 | M | Infection |

Quasi Identifiers

Sensitive Data

2-Anonymous Data (k=2)

| ID | Age | Zip | Gender | Disease |
|---|---|---|---|---|
| P001 | 21 – 30 | 2141 | Human | Cancer |
| P002 | 21 – 30 | 2141 | Human | Infection |
| P003 | 31 – 35 | 213* | F | AIDS |
| P004 | 31 – 35 | 213* | F | AIDS |
| P005 | 41 - 50 | * | M | Cancer |
| P006 | 41 - 50 | * | M | Infection |
| P007 | 41 - 50 | * | M | Infection |

Generalisation

Suppression

k = 2 => **at least** 2 rows in each group

# PETs

## K-Anonymity

Is it still possible to know what disease a patient had if you know background details of that patient?

Alice is a 32-year-old female who lives in area 2138......

Kate lives in area of 2141, and she purchased antibiotics on her way to home after the hospital visit...

2-Anonymous Data (k=2)

| ID | Age | Zip | Gender | Disease |
| --- | --- | --- | --- | --- |
| P001 | 21 – 30 | 2141 | Human | Cancer |
| P002 | 21 – 30 | 2141 | Human | Infection |
| P003 | 31 – 35 | 213* | F | AIDS |
| P004 | 31 – 35 | 213* | F | AIDS |
| P005 | 41 - 50 | * | M | Cancer |
| P006 | 41 - 50 | * | M | Infection |
| P007 | 41 - 50 | * | M | Infection |

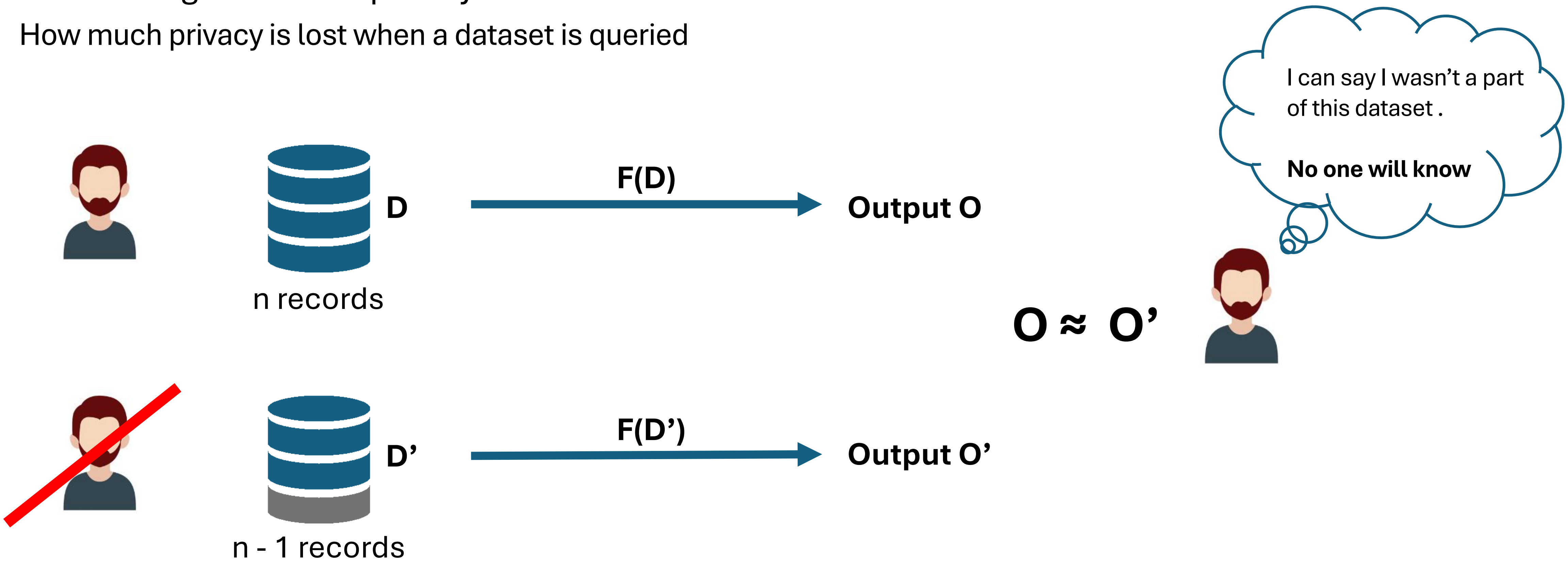
# PETs

## Differential Privacy (DP)

Adding random noise to the data so that it is hard to tell whether an individual is participating in a dataset or not

A mathematical guarantee of privacy

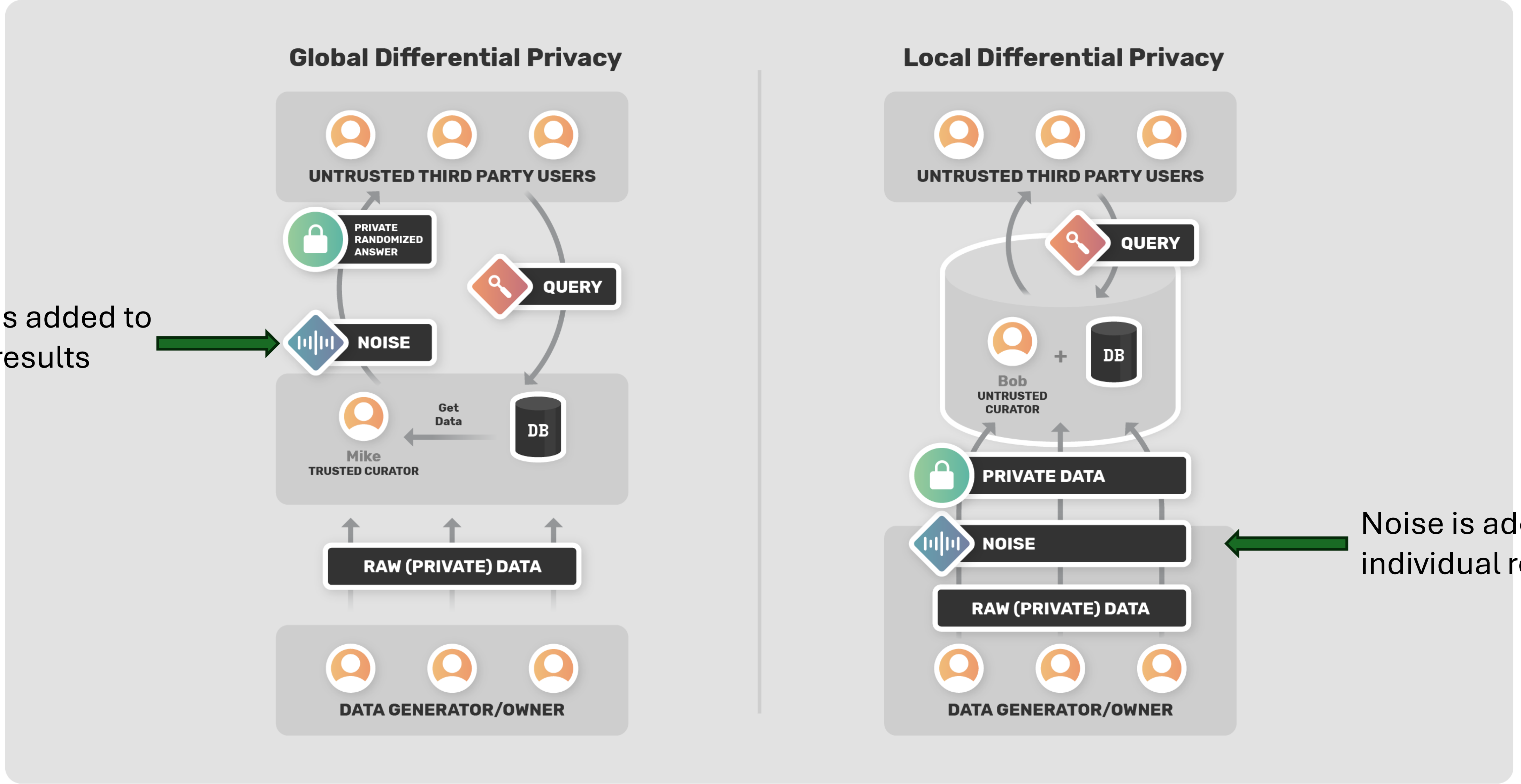
How much privacy is lost when a dataset is queried



# PETs

## Differential Privacy (DP)

Noise is added to query results



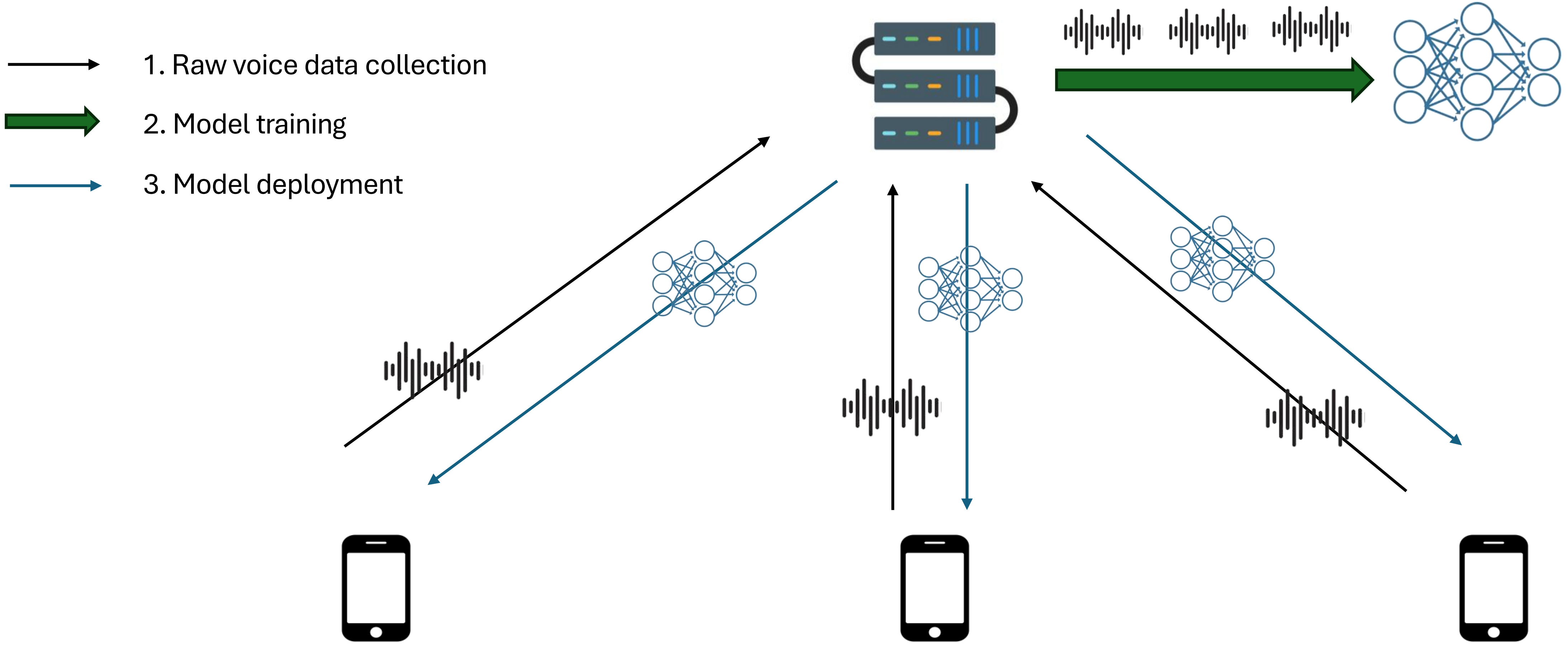

Noise is added to individual records

LOCAL VS GLOBAL DIFFERENTIAL PRIVACY (OpenMined 2021)

# PETs

## Federated Learning (FL)

Scenario : Apple using machine learning with **less privacy** to train Siri in identifying only the iPhone owner's voice



# PETs

## Federated Learning (FL)

- Privacy preserving machine learning
- Data is not sent to a centralised location to train a machine learning model
- Setup
  - There is a central server
  - Nodes are connected to the central server (e.g., mobile phones)
  - Data is collected at the local nodes
  - Nodes can connect and disconnect from the server at any time

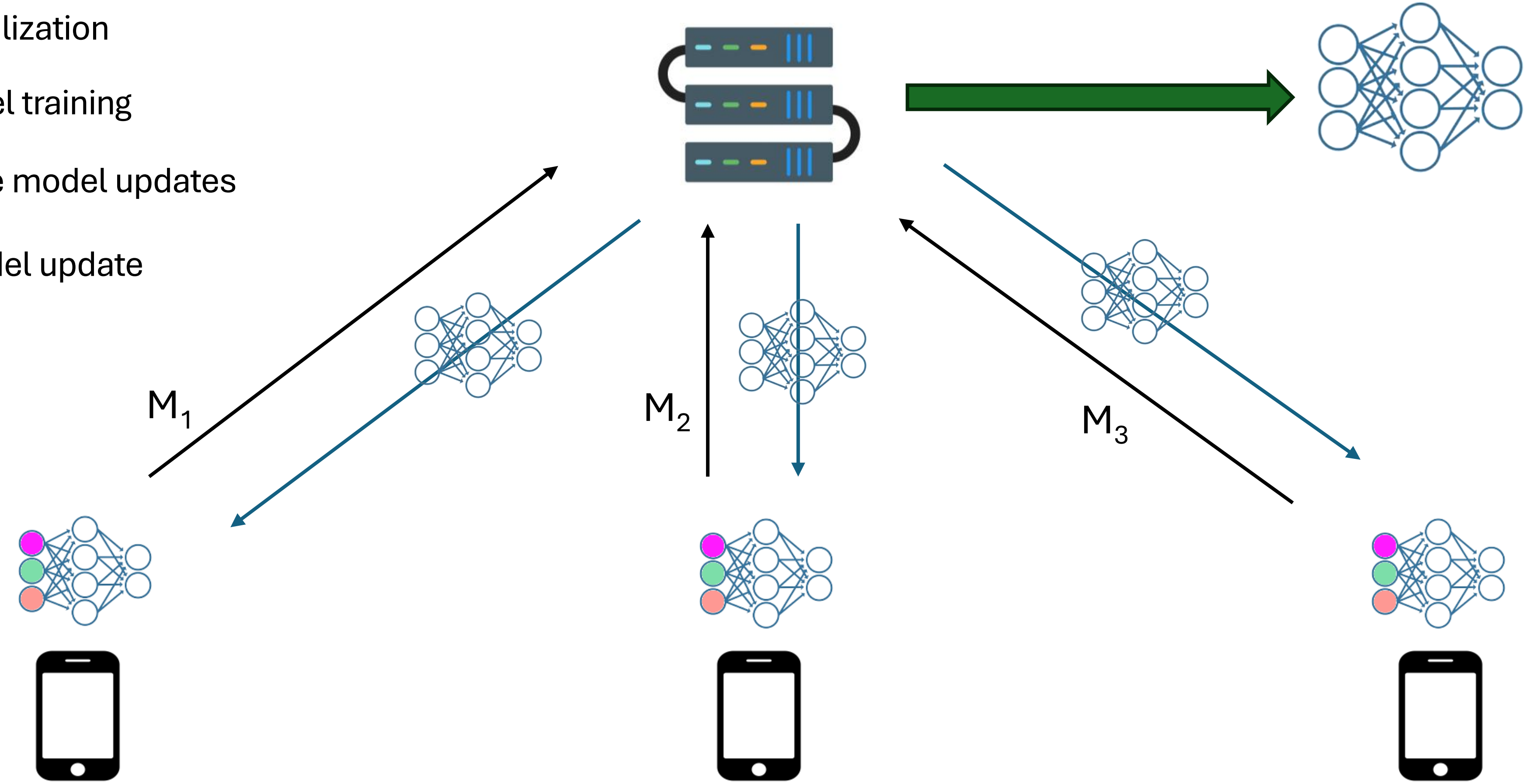
Data never leave the source !!

# PETs

## Federated Learning (FL)

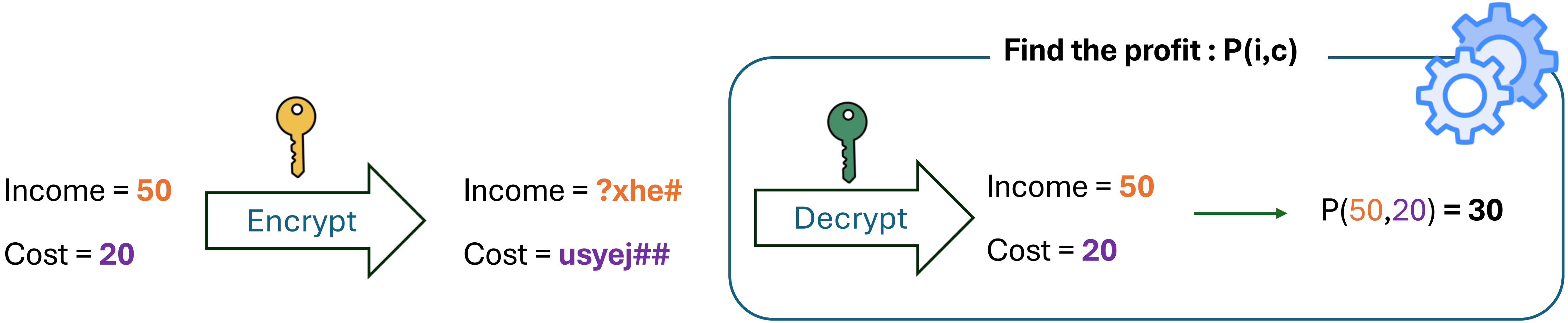
Scenario : Apple using **privacy preserving** machine learning to train Siri in identifying only the iPhone owner's voice

1. Model initialization
2. Local model training
3. Sending the model updates
4. Global model update



# PETs

## Homomorphic Encryption (HE)

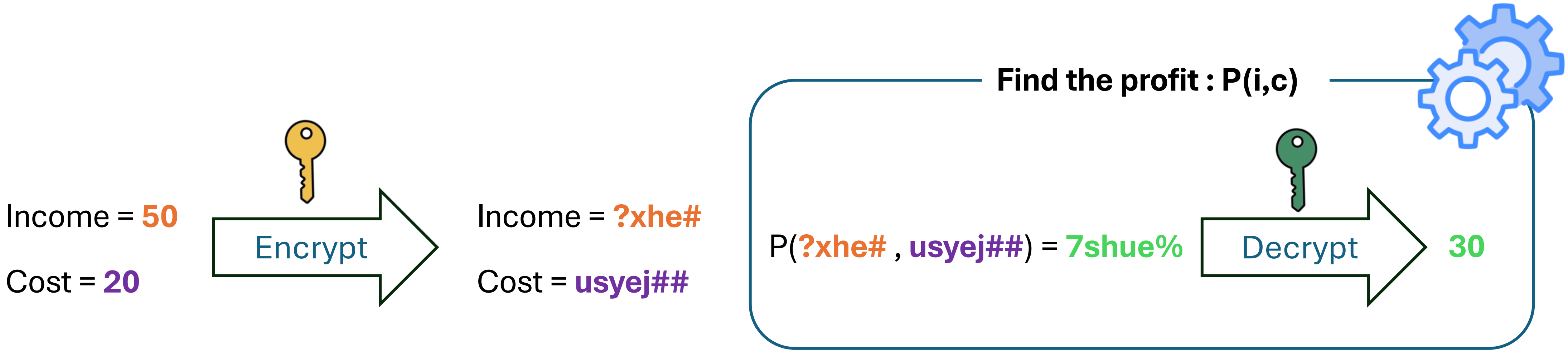



- If we want to process the encrypted data
  - The data must be reversed to the original format
  - Privacy violations are possible

# PETs
## Homomorphic Encryption (HE)

- What if we can perform operations on encrypted data without decrypting them

  Homomorphic Encryption




- For arithmetic operations : addition and/or multiplication

# PETs
## Homomorphic Encryption (HE)

- Types of Homomorphic Encryption

| Types | Operation (+, *) | Number of times |
|---|---|---|
| Partially | One | Unlimited |
| Somewhat | Both | Limited |
| Fully | Both | Unlimited |

- Can we perform other arithmetic operations using + and *, if we have unlimited chances to use them?

Almost

4/2
= 4 + (-1 * 2) = 2
= 2 + (-1 * 2) = 0

How many times => 2 (answer)

Alexander Wood, Kayvan Najarian, and Delaram Kahrobaei. 2020. Homomorphic Encryption for Machine Learning in Medicine and Bioinformatics. ACM Comput. Surv. 53, 4, Article 70 (July 2021), 35 pages. https://doi-org.ezproxy.auckland.ac.nz/10.1145/3394658

# PETs
## Zero Knowledge Proof (ZKP)

- Allow one party to prove the validity of a claim to another party without revealing the data bound to the claim
- A cryptographic technique

# PETs

## Zero Knowledge Proof (ZKP)

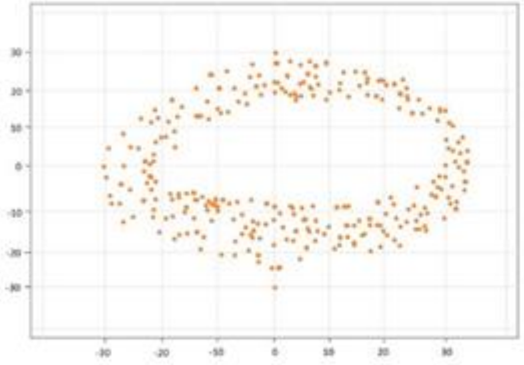
- There are 2 roles in ZKP
  1. **Prover** : wants to show they have knowledge of something
  2. **Verifier** : check the prover's claim

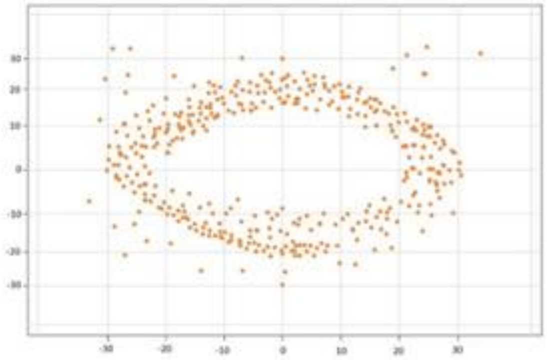
- **Proof size** : the amount of information passed between prover and verifier during the process

- There are 3 criteria to be met in ZKP
  1. **Completeness** : An honest prover always convince the verifier
  2. **Soundness** : Verifier can reject the false proofs shared by the dishonest prover
  3. **Zero-knowledge** : Prover knows nothing beyond the claim made by the prover

# PETs
## Synthetic Data

- Artificially generated data that mimics the statistical properties of real data


Original Data

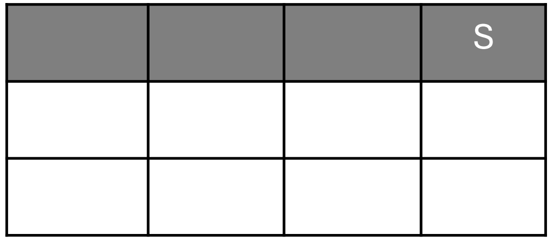

Synthetic Data

Navigating The Potential And Perils Of Synthetic Data In Healthcare (Shashank Agarwal 2024)

**When do we need it?**

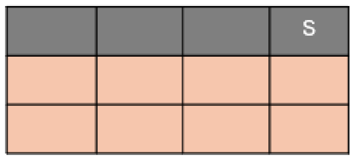
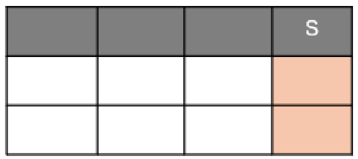
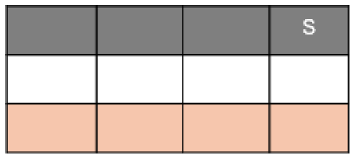
# PETs
## Synthetic Data

- Types of synthetic data




original dataset

S = sensitive data

| Types | Synthetic Data Usage |
|---|---|
| Fully |   |
| Partially |   |
| Hybrid |   |

# PETs
## Synthetic Data

- How to generate Synthetic Data?

  - Statistical Distribution

    Using algorithms that can generate synthetic data using the statistical properties (ex: mean, variance etc) of the original data

  - Generative Adversarial Networks

    A deep learning technique that use 2 neural networks to generate new data

    Generator Network – generating the new data

    Discriminator Network – Deciding how close the data to the real data

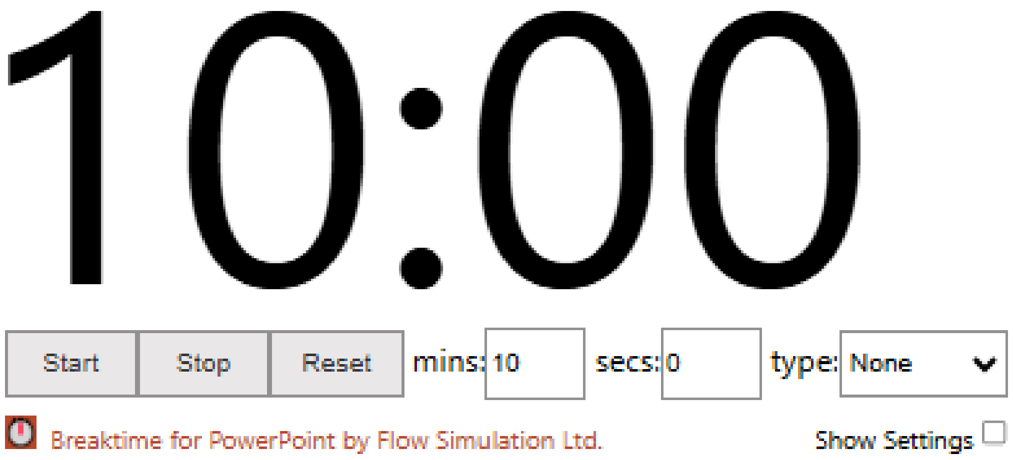
  - Variational Autoencoders (VAEs)

    Encoded-decoded architecture

    VAE transforms data into a lower-level representation (encode) then try to regenerated it back (decode)

# PETs

## Canvas – More Than Privacy Protection



# PETs
## Secondary Benefits

- Safe collaboration with untrusted parties

  pseudonymisation

- Data can be used for secondary purposes without explicit consent of the data subjects

- Reduced regulatory pressure

- Competitive advantage “Our product offers robust privacy measures than others”

- Mitigate the difficulties in obtaining datasets for product testing or model training purposes

# PETs

## How do PETs achieve data protection

- Transform data or derived data
    - Pseudonymisation
    - K-Anonymity
    - Differential Privacy
    - Synthetic Data
- Hide or shield data
    - Pseudonymisation
    - Homomorphic Encryption
    - Zero-Knowledge Proof
- Split datasets or control access to datasets
    - Pseudonymisation
    - Federated Learning

What about the utility?

# PETs
## Maturity

- Cutting edge does not mean the technology is applicable

- Standards in application might not have been developed

  Ex: [https://homomorphicencryption.org/standard/](https://homomorphicencryption.org/standard/)

- Some PETs can work better in theoretical settings

- How can we decide the maturity
  - ✓ Standard measurement systems

    Ex: Technology Readiness Levels (TRL)

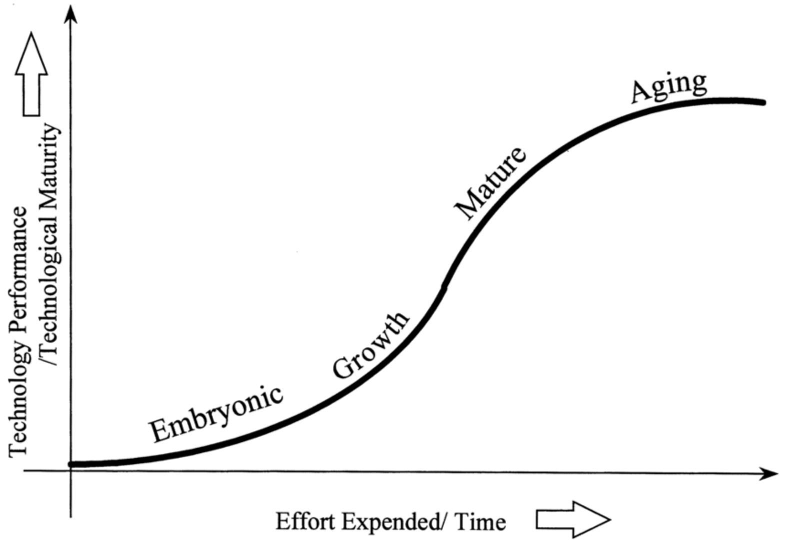
  - ✓ Quality measures

    Scalability

    Versatility

    Level of protection

    Robustness




Performance analysis of technology using the S curve model: The case of digital signal processing (DSP) technologies (Nieto, Mariano & Lopéz, Francisco & Cruz-Roldan, Fernando. (1998). )

# PETs

## Known Weaknesses

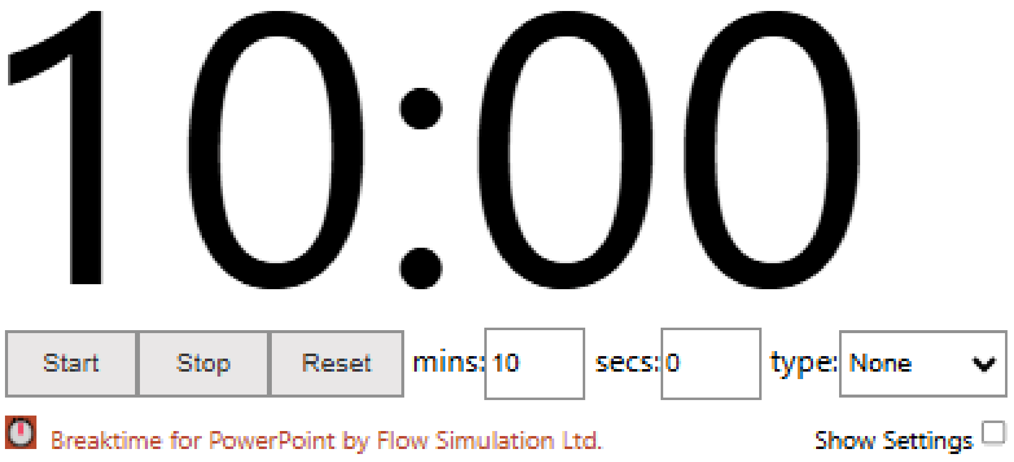
- Pseudonymisation
  - Risk of re-identification
- K-Anonymity
  - Risk of re-identification
- Differential Privacy
  - No consensus over the optimal privacy vs utility tradeoff
  - Performance overhead
  - Require higher computational power
  - Limited chances of gaining insights
- Federated Learning
  - Depends on architectural specifications
  - Data quality and format at the local nodes matter
  - Performance overhead

# PETs
## Known Weaknesses

- Homomorphic Encryption
  - Performance overhead
  - Limited functionality
- Zero-Knowledge Proof
  - Interactive protocols are vulnerable to side channel attacks
  - Performance overhead
- Synthetic Data
  - May not present outliers in the original dataset
  - Risk of biased data
  - Difficulty to mimic complex data

# PETs

## Canvas – Which PET(s) Do You Need?



# PETs

## How to choose?

- Involve personal data handling
- Nature of the data processing : *individual level or aggregate analysis?*
- Scope of the data processing : *internal, collaboration, sharing?*
- Size of the data : *"is it worth to apply a selected PET on 5 records?..."*
- Possible threats to data : *"Is the curator trusted?.."*
- Sufficient resources : *time, manpower, computational power, money,……*
- Maturity

# Summary

- Anonymity
- Privacy Enhancing Technologies (PETs)
    - Several PETs
    - Maturity
    - Weaknesses
    - How to choose a PET