PAROMA-MED

Privacy Aware and Privacy Preserving Distributed and Robust Machine Learning

Blockchain-Enhanced Reversible Watermarking Framework for End-to-End Data Traceability in Federated Learning Systems

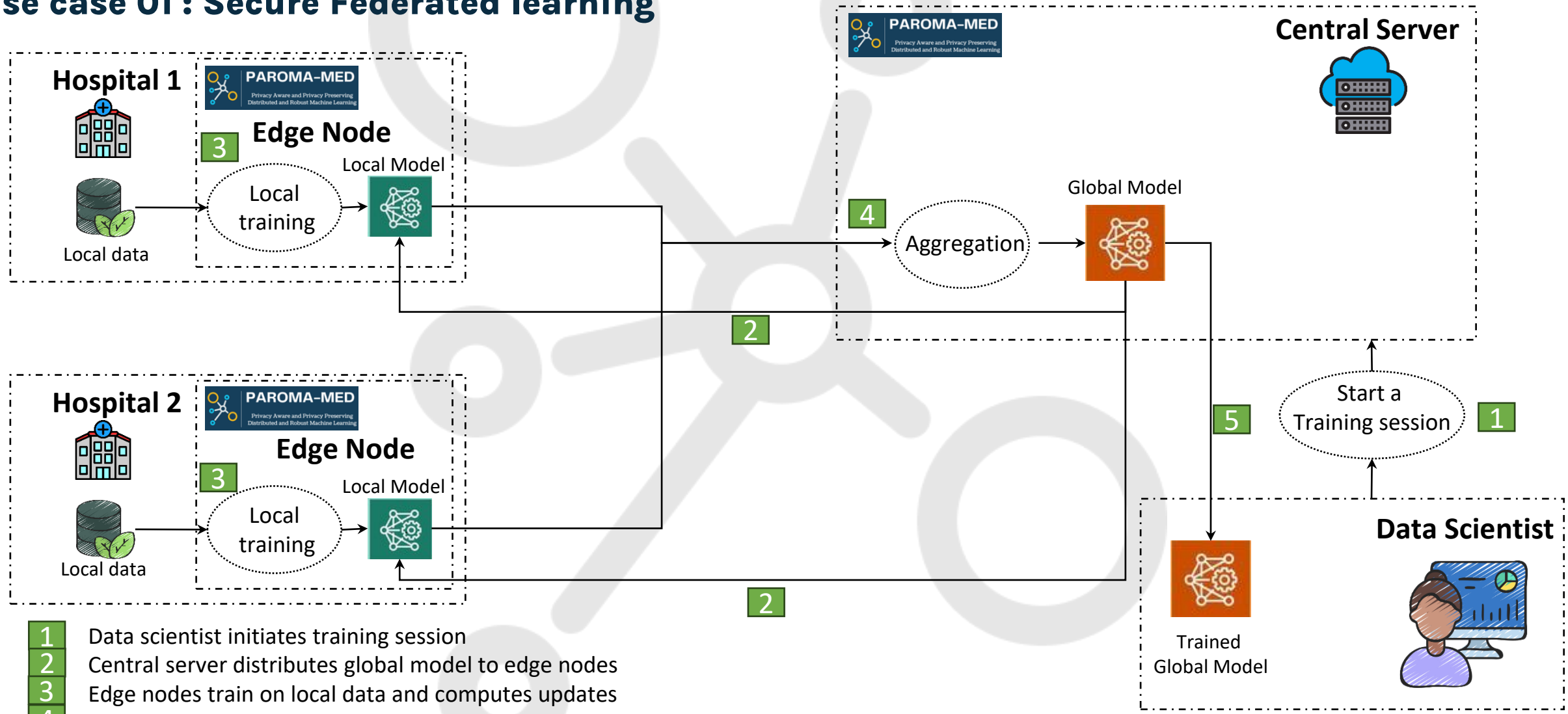
Reda Bellafqira, Gouenou Coatrieux, Chloé Berton.

IMT Atlantique, Inserm, UMR 1101 LaTIM, 29238 Brest, France.

CSP2025, 27/04/2025, Okinawa, Japan.

Paroma-Med Framework

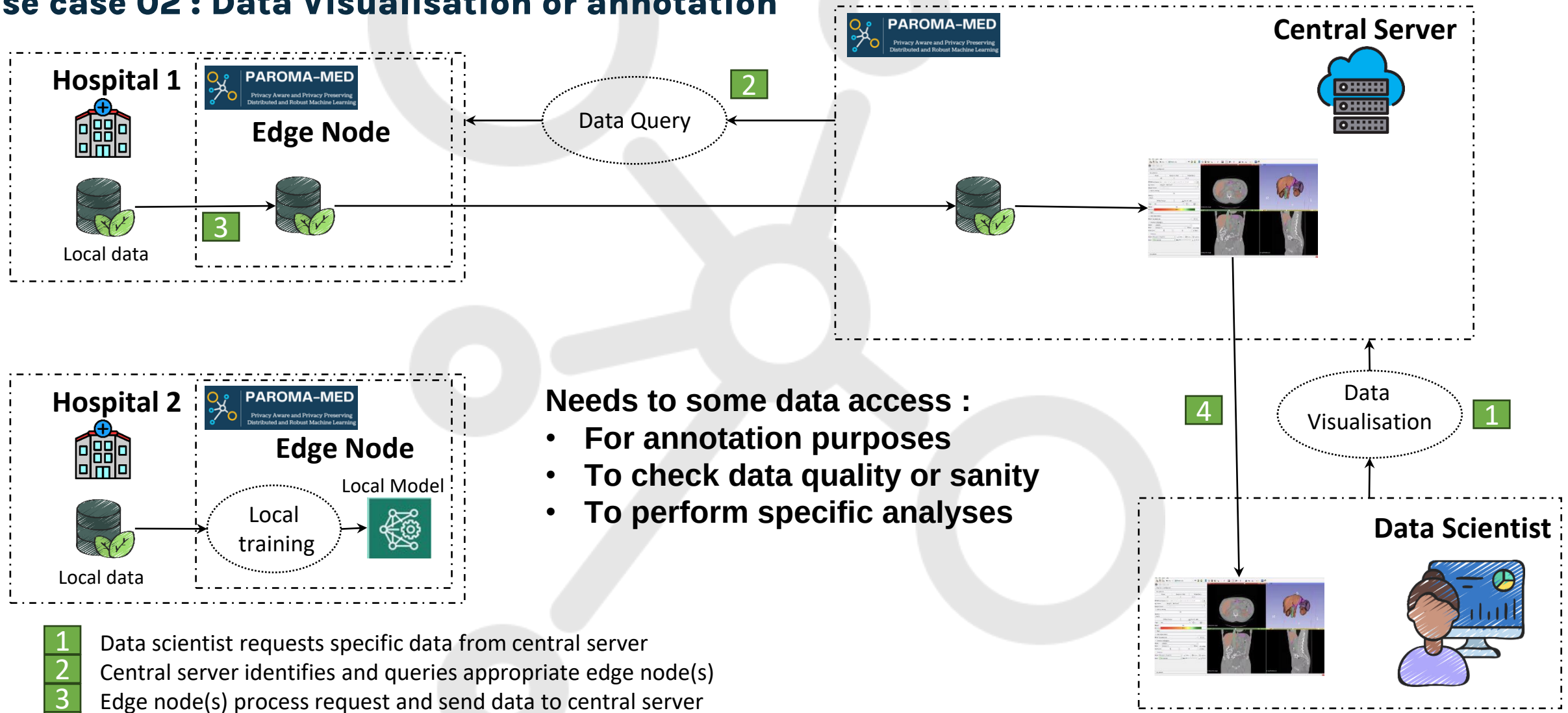
Use case 01 : Secure Federated learning



1. Data scientist initiates training session
2. Central server distributes global model to edge nodes
3. Edge nodes train on local data and computes updates
4. Nodes send updates to central server to aggregate them (Repeat steps 2-5 for multiple rounds)
5. Final global model sent to the data scientist

Paroma-Med Framework

Use case 02 : Data Visualisation or annotation



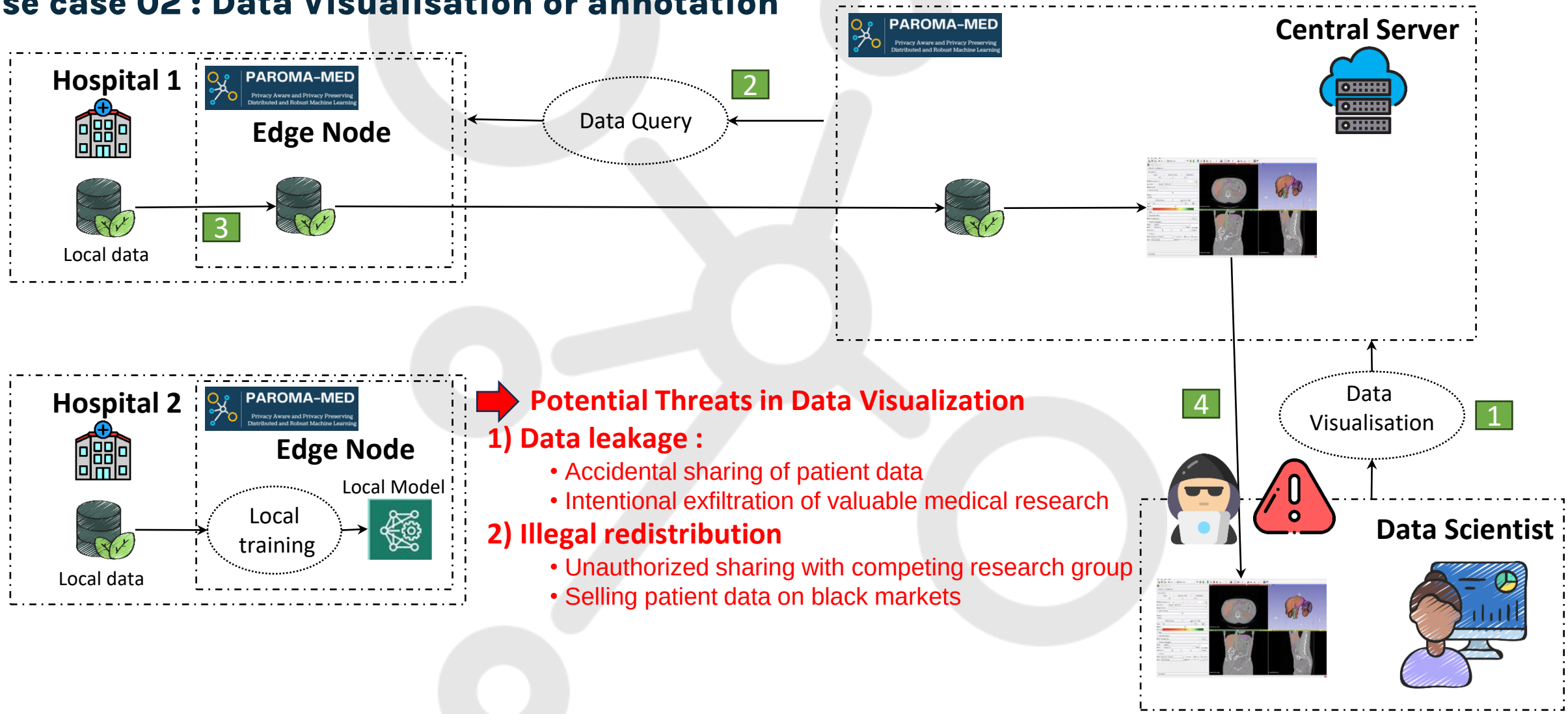
Needs to some data access :

- **For annotation purposes**
- **To check data quality or sanity**
- **To perform specific analyses**

1. Data scientist requests specific data from central server
2. Central server identifies and queries appropriate edge node(s)
3. Edge node(s) process request and send data to central server
4. Central server grants data scientist to visualize data using tools (e.g., Slicer 3D)

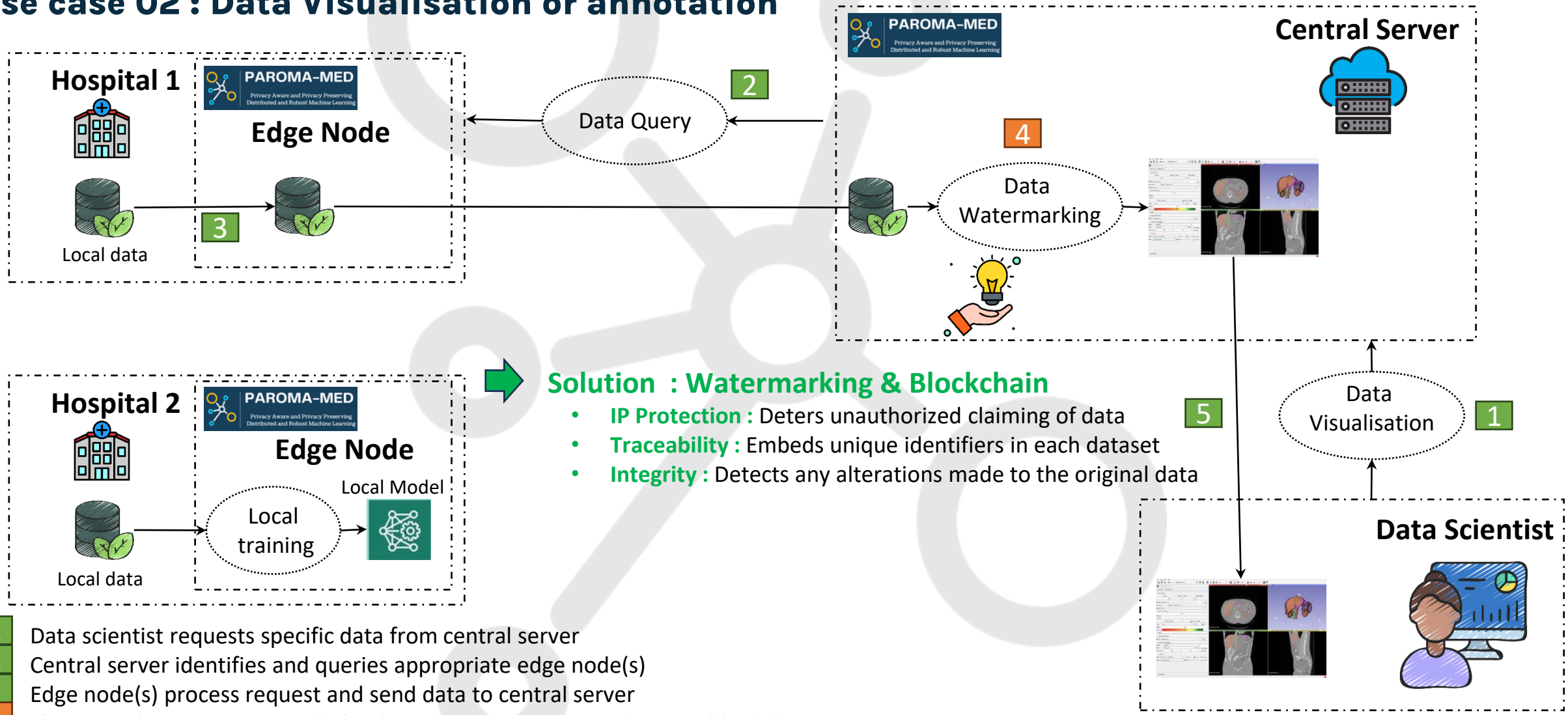
Paroma-Med Framework

Use case 02 : Data Visualisation or annotation



Paroma-Med Framework

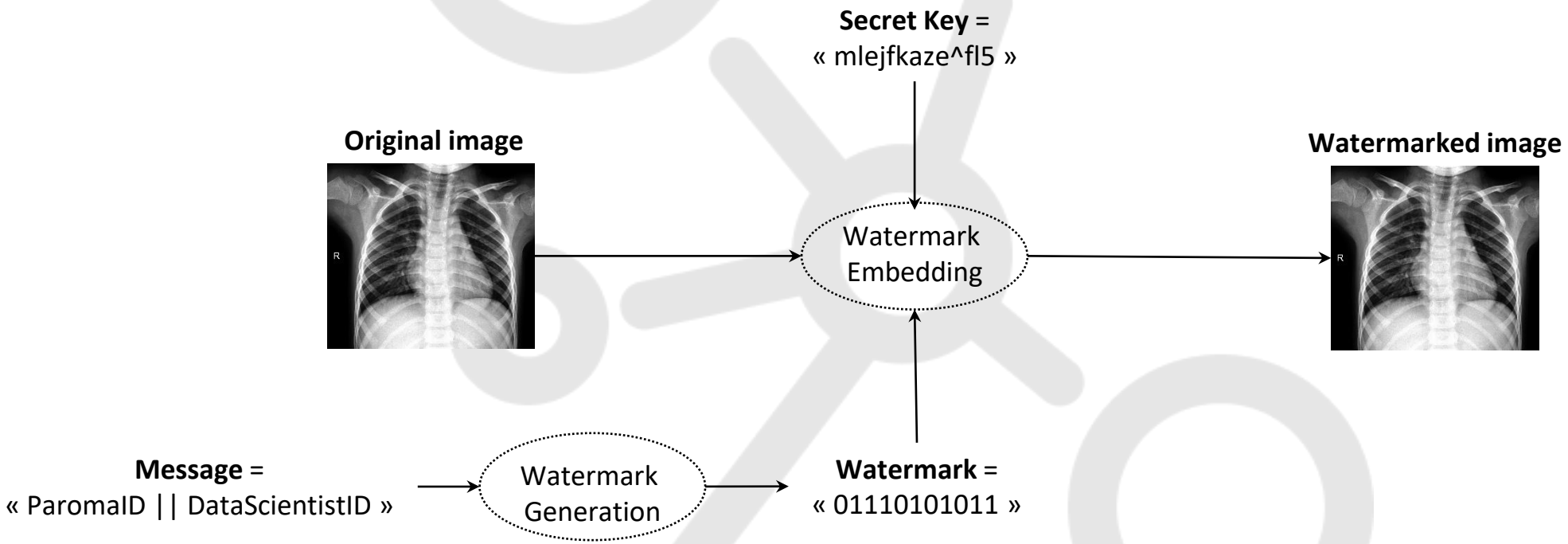
Use case 02 : Data Visualisation or annotation



1. Data scientist requests specific data from central server
2. Central server identifies and queries appropriate edge node(s)
3. Edge node(s) process request and send data to central server
4. **The central server watermark the data and store the metadata in a blockchain**
5. Central server grants data scientist to visualize data using tools (e.g., Slicer 3D)

Principle of image watermarking

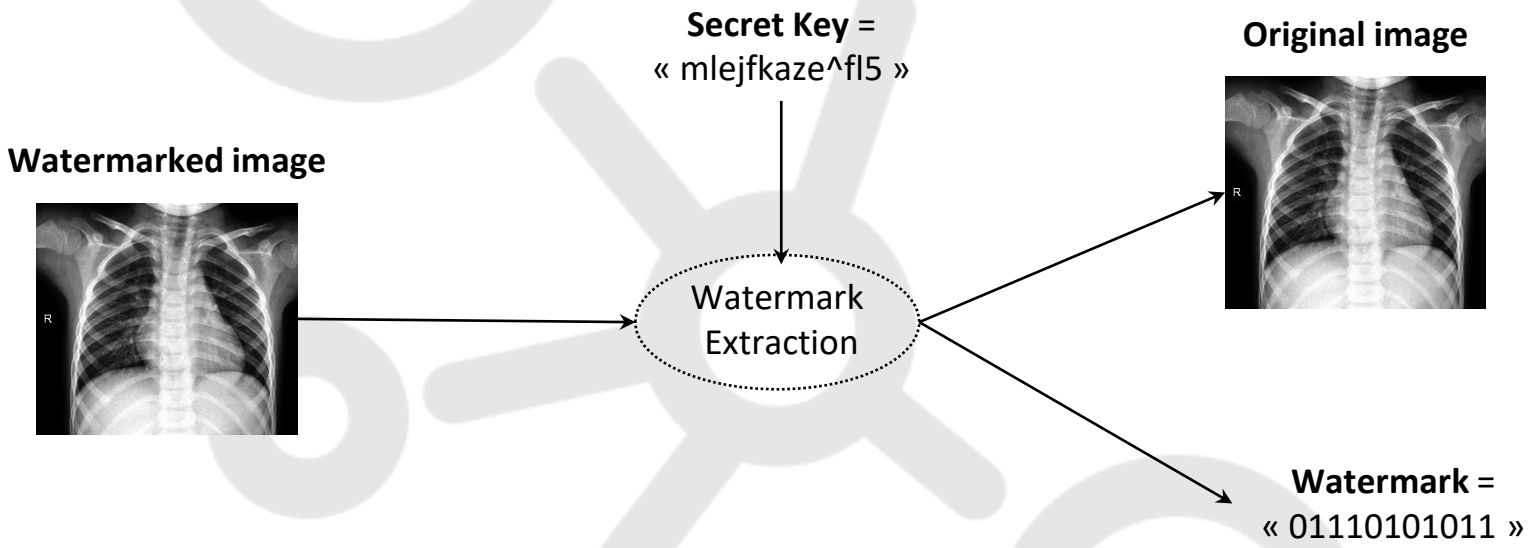
1. Embedding Stage



https://github.com/Bellafqira/histogram_shiffting_predictions

Principle of image watermarking

2. Detection/Extraction Stage



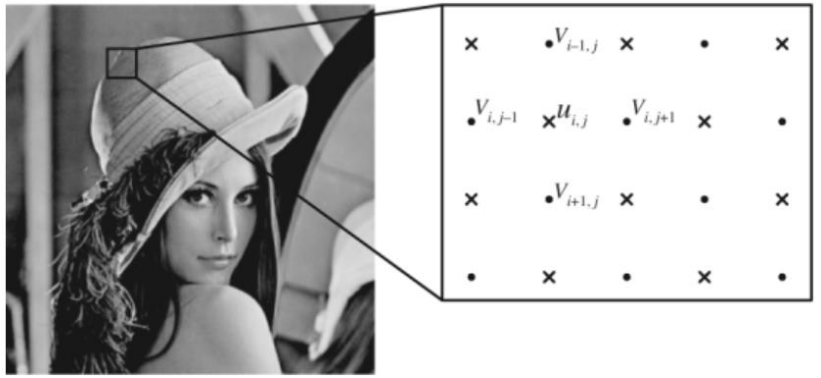
Why Reversible Watermarking?

- Preserves data integrity for medical diagnosis and research
- Allows complete removal of the watermark when needed
- Ensures no loss of original data quality
- Critical for maintaining the accuracy of AI model training
- Complies with medical data regulations requiring data preservation

Histogram shifting over prediction errors

Center pixel $u_{i,j}$ of the cell can be predicted from the four neighboring pixels $v_{i,j-1}$, $v_{i+1,j}$, $v_{i,j+1}$, and $v_{i-1,j}$. The predicted value $u'_{i,j}$ is computed as follows:

$$u'_{i,j} = \left\lfloor \frac{v_{i,j-1} + v_{i+1,j} + v_{i,j+1} + v_{i-1,j}}{4} \right\rfloor$$



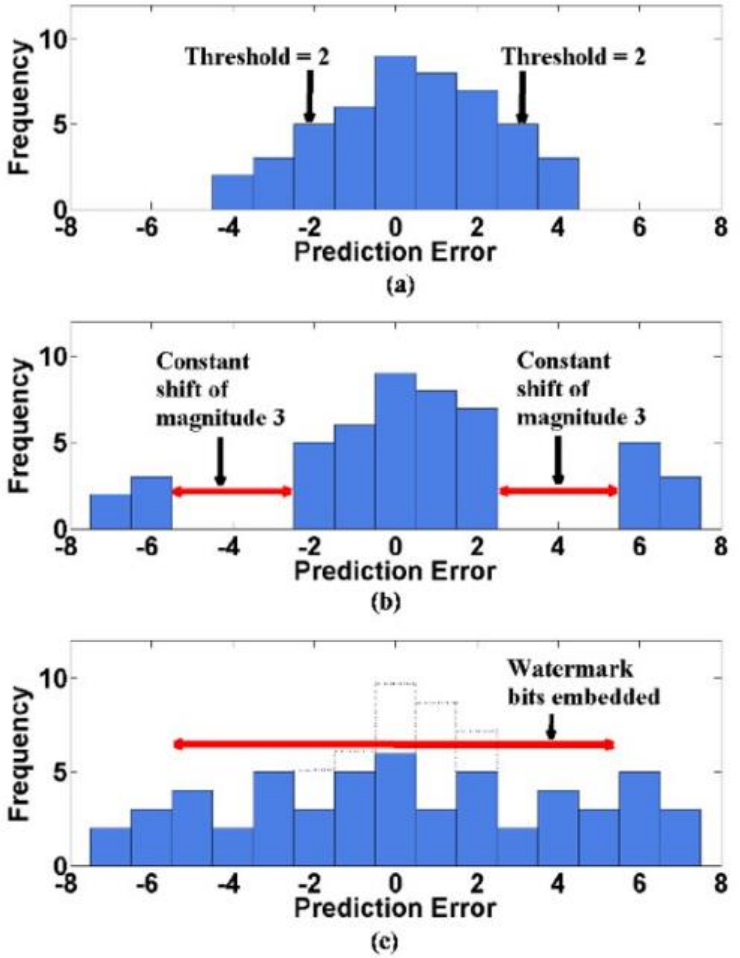
Based on the predicted value $u'_{i,j}$ and original value $u_{i,j}$, the prediction error $d_{i,j}$ is computed as

$$d_{i,j} = u_{i,j} - u'_{i,j}$$

Histogram shifting over prediction errors

The histogram shift encoding algorithm modifies prediction errors d as follows:

$$D_{i,j} = \begin{cases} 2d_{i,j} + b, & \text{if } d_{i,j} \in [T_n; T_p] \\ d_{i,j} + T_p + 1, & \text{if } d_{i,j} > T_p \text{ and } T_p \geq 0 \\ d_{i,j} + T_n, & \text{if } d_{i,j} < T_n \text{ and } T_n < 0 \end{cases}$$



T_n and T_p are used, where T_n is the negative threshold value, and T_p is the positive threshold value.

Histogram shifting over prediction errors

The decoder recovers original prediction errors $d_{i,j}$ and bits of the embedded data b from $D_{i,j}$ according to the following:

$$d_{i,j} = \begin{cases} \lfloor D_{i,j}/2 \rfloor, & \text{if } D_{i,j} \in [2T_n; 2T_p + 1] \\ D_{i,j} - T_p - 1, & \text{if } D_{i,j} > 2T_p + 1 \text{ and } T_p \geq 0 \\ D_{i,j} - T_n, & \text{if } D_{i,j} < 2T_n \text{ and } T_n < 0 \end{cases}$$

$$b = D_{i,j} \bmod 2, \quad D_{i,j} \in [2T_n; 2T_p + 1]$$

The original pixel's value is computed as

$$u_{i,j} = u'_{i,j} + d_{i,j}$$

Histogram shifting over prediction errors

- **Overflow Problem:**
  - When adding 1 to edge pixels with value 255, an overflow occurs, causing data loss.
- **Solution:** We generate a binary overflow vector that contains:
  - 0 when center of block contains 254
  - 1 when center of block contains 255
- This vector is embedded starting from the last block of the image and is crucial for reversibility of the scheme.

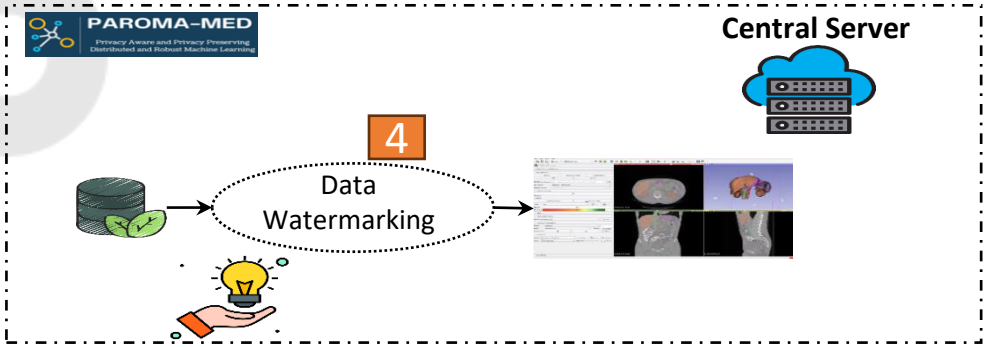
Histogram shifting over prediction errors

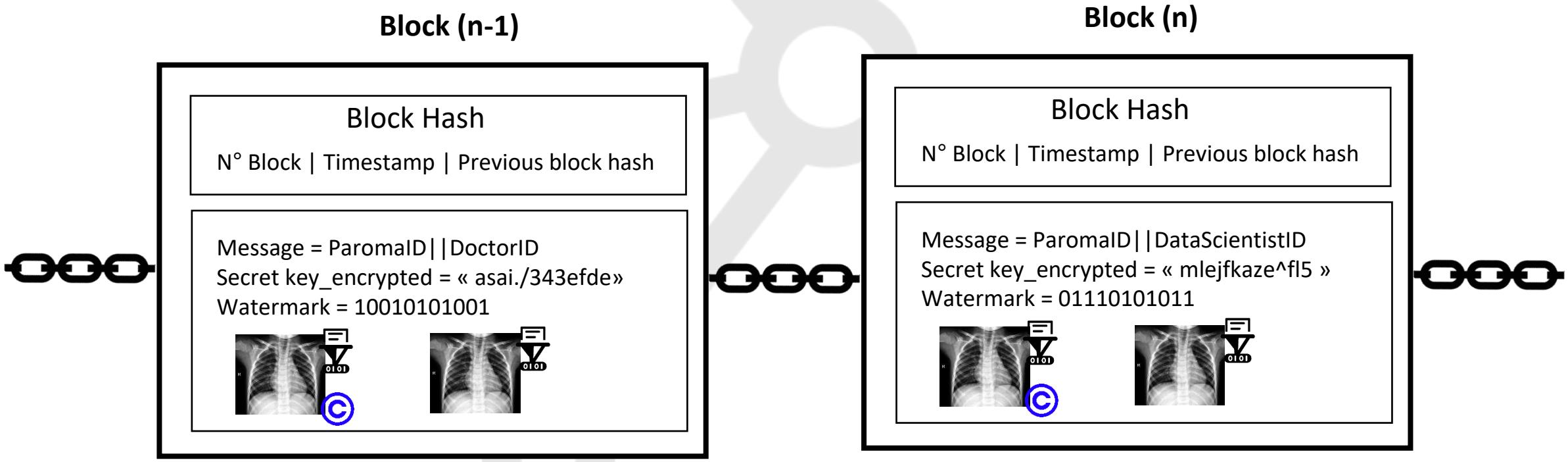
- **Extraction Process:**
  - Extract the watermark from blocks
  - Count the number of 255s to determine overflow vector size
  - Use binary values to reverse modifications in blocks where centers equal 25

Watermarking & Blockchain

Blockchain Integration Benefits

- Immutable record of all data access and watermarking operations
- Enables quick verification of data ownership and integrity
- Decentralized storage enhances security and reliability
- Supports transparent data sharing in research collaborations

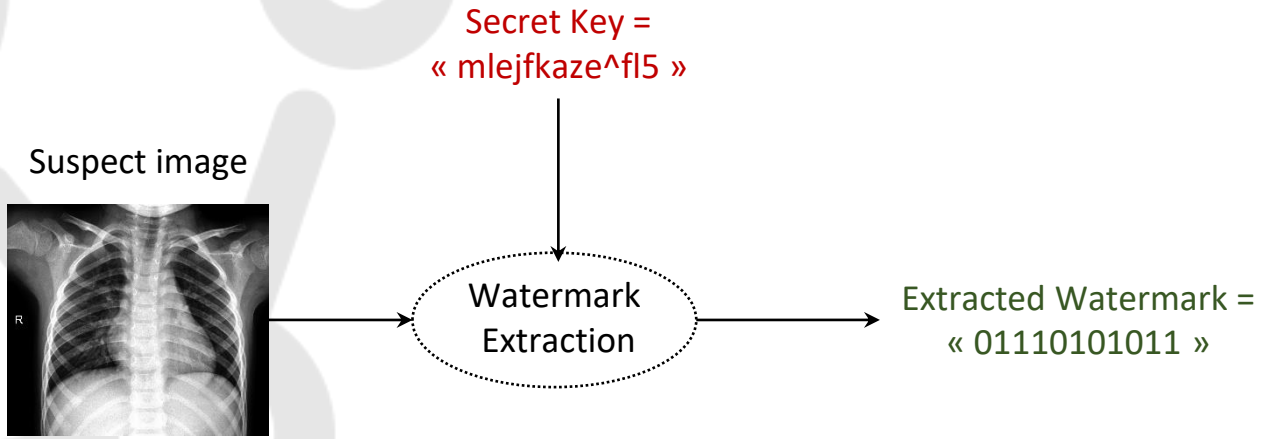




: Hash Function : Watermark

Principle of image watermarking

2. Detection/Extraction Stage



Block (n)

Block Hash

N° Block | Timestamp | Previous block hash

Message = ParomaID||DataScientistID
Secret key_encypted = « mlejfkaze^fl5 »
Watermark = 01110101011

1. **Compute the hash of the suspect image**
2. **For each** block in the Blockchain:
    1. If the **computed hash** exists in the current block:
        1. Return the **message** stored in the block
    2. **Extract** the **watermark** from the image using the block's **secret key**
        1. If the **extracted watermark** equals the **watermark in the block**:
            1. Return the **message** in the block
    3. If **no match**, continue to the **next block**
3. If **no matching** block is **found**:
    1. Return "**No watermark detected**"

Experimental results

- Dataset and Experimental Setup:
  - Chest X-ray images for COVID-19 positive cases along with Normal and Viral Pneumonia comprising 544_893 images of size (299,299) encoded into 8 bits.
  - VGG16 model for classification and performed comparative analysis between watermarked and non-watermarked datasets.

Experimental results: Utility

Fig. 1. Training and Validation Loss vs. Epochs

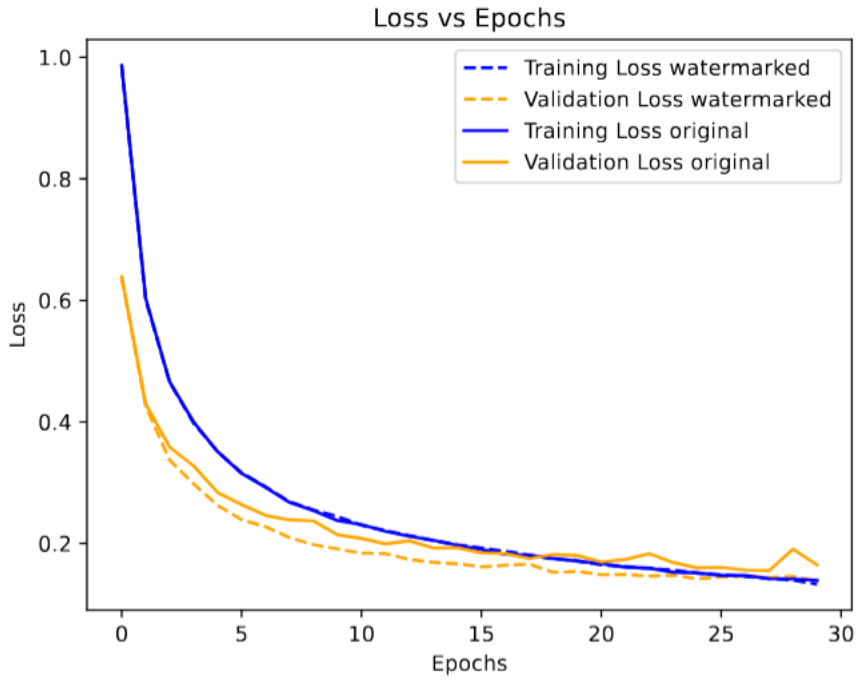
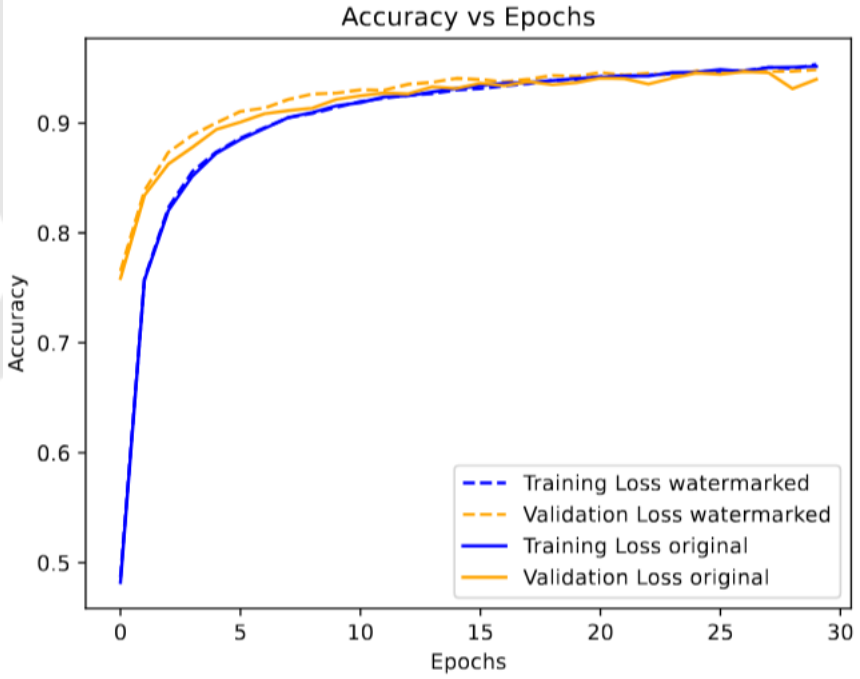


Fig. 2. Training and Validation Accuracy vs. Epochs



- A slight improvement in the training performance on the watermarked dataset.
- Watermarking adds noise to the data, potentially helping the model generalize better.

Experimental results : Robusteness

- We used the bit error rate (BER), which represents the proportion of bits that are incorrectly decoded during the watermark extraction process.
- A BER equal to zero -> a perfect match between the embedded and extracted watermarks
- A BER of 0.5 -> there is no correlation between the watermarks.
- the watermark is the HMAC of the message. Due to the confusion and diffusion properties of the HMAC, it is very difficult to find two different messages whose HMAC values have a BER smaller than 0.3.

BIT ERROR RATE (BER) UNDER DIFFERENT ATTACKS

Attack Type	BER
Histogram Shift	0
Contrast Increase	0
Contrast Decrease	0
Gamma Correction	0
Histogram Equalization	0
Adaptive Histogram Equalization	0
Gaussian Noise	0.20703125
No Attack (Original)	0

Conclusion & Futur works

- Combined reversible watermarking with blockchain for medical image traceability in federated learningTechnical
- Histogram shifting on prediction errors with overflow management for guaranteed reversibility
- Blockchain validation, digital signatures, and cryptography ensure data integrity and authenticity
- Preserved VGG16 model performance with slight accuracy improvements and robust attack protection
- Limitations -> High computational cost of blockchain operations; vulnerability to geometric transformations
- Future Work -> Develop geometric-invariant watermarking; implement on Ethereum with smart contracts; enhance scalability for healthcare applications

PAROMA-MED partners

EURESCOM

agentscape

ERICSSON

Institut Mines-Télécom

Qualtek
Thinking Forward

THALES
Building a future we can all trust

UBITECH
ubiquitous solutions

UNIVERSITÉ DE RENNES 1

6G Health Institute

PAROMA-MED

Privacy Aware and Privacy Preserving Distributed and Robust Machine Learning

Thank you!

Reda Bellafqira – reda.bellafqira@imt-atlantique.fr