# Try before You Buy: Privacy-preserving Evaluation on Cloud-based Machine Learning Data Marketplace

**Qiyang Song**, Jiahao Cao, Kun Sun, Qi Li, and Xu Ke

# What is Data Marketplace ?

GEORGE MASON UNIVERSITY

- A good deep learning model relies on huge good-quality data.
  - ➢ Trainers want to enrich their internal data sets with external data.
- As a result, data marketplaces emerge,
  - ➢ providing data exchanging platforms for both enterprises and individuals.

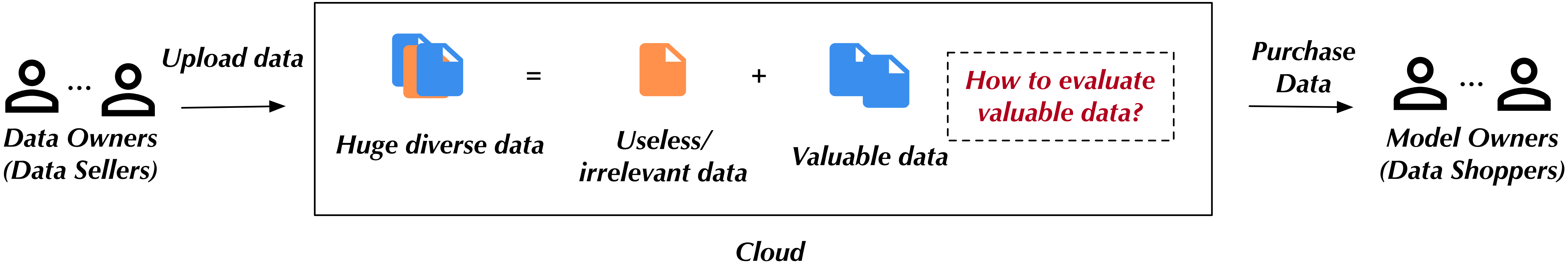
BDEX

Windows Azure Marketplace

DAWEX SELL BUY SHARE DATA

# Cloud-based Data Marketplace

- Traditional Cloud-based Data Marketplace

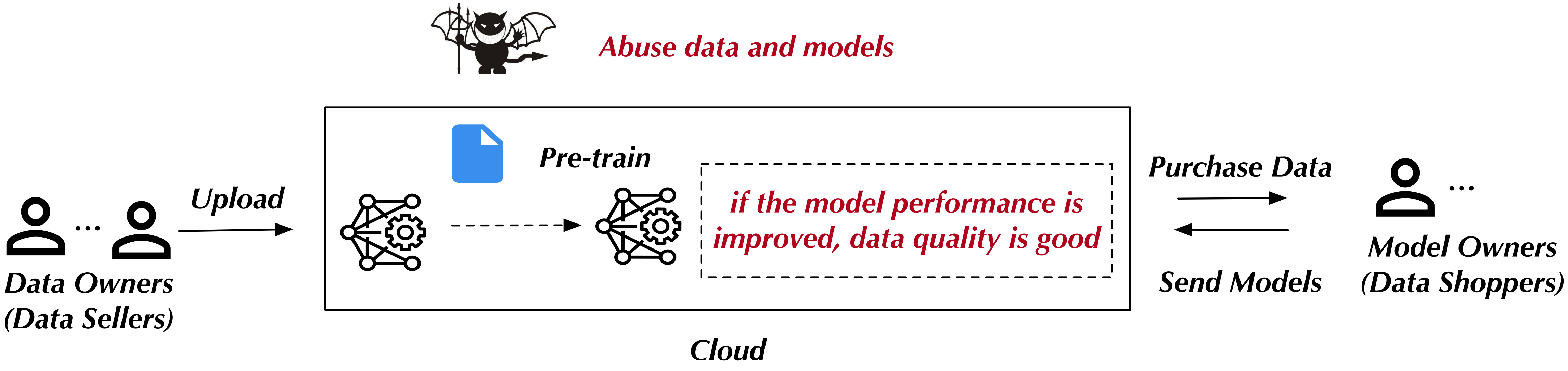



- ➢ Model owners want to **purchase the most valuable data** to improve their models,
- ➢ but data owners may provide **useless/irrelevant data** that do not improve model performance.

How to evaluate the most valuable data for data shoppers' models?

# Intuitive ML Data Evaluation

- Cloud needs to access both sellers' data and shoppers' models, but it is untrusted.
- Data and models may be sensitive for both sellers and shoppers!



# Our Goal:

Provide privacy-preserving ML data evaluation on data marketplaces

# Existing Privacy Protection Solutions

- Existing privacy protection solutions
  - Homomorphic Encryption (HE), Secure Multi-party Computation (MPC)
  - Can preserve both the privacy and functionality of data/models on the cloud
  - **Limitations**
    - high computational and communication overhead
    - not specially designed for ML data evaluation

We need a lightweight encryption approach that is specially designed for ML data evaluation.

# Our Solution

- We design a lightweight encryption approach to protect the privacy of data/models.
  - So, the encrypted data/models cannot be directly evaluated by the cloud.

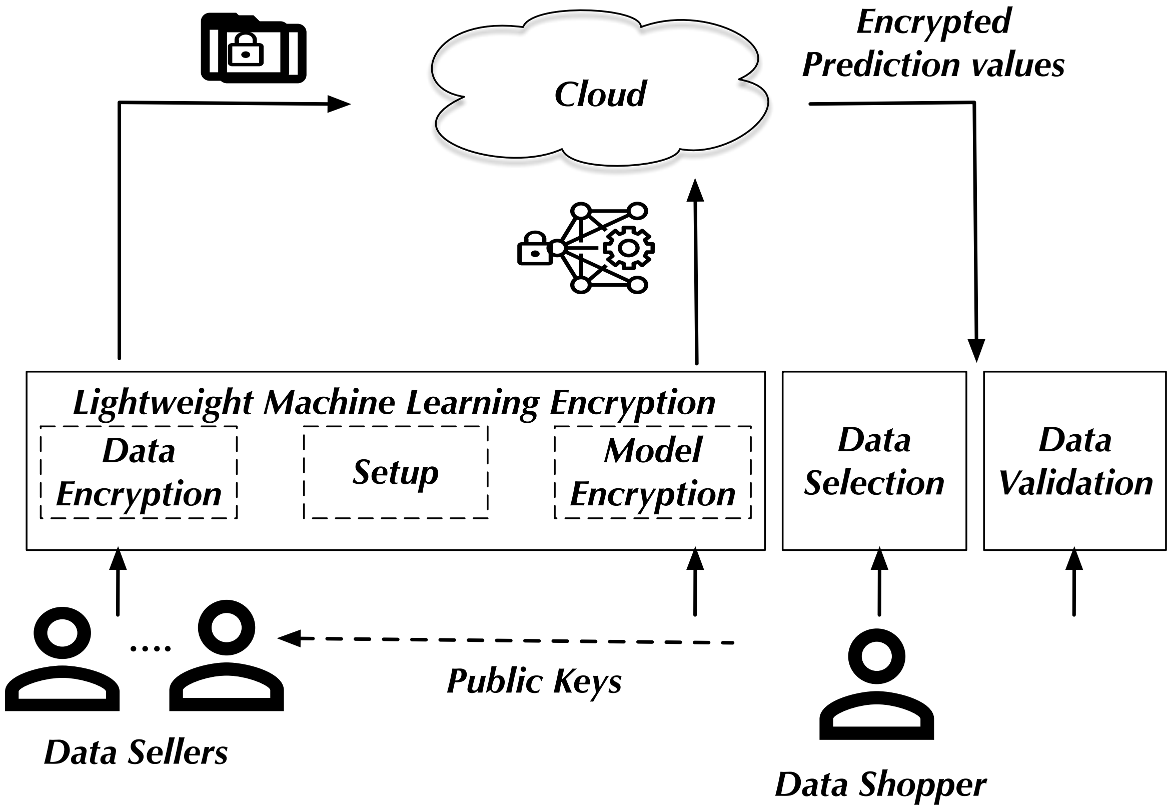
- We provide a ML evaluation approach that is compatible with our lightweight encryption approach
  - Instead of accessing the original data, we need extra information and mechanisms to evaluate valuable data.

# Our System

- Data sellers upload encrypted data to the cloud for sale.

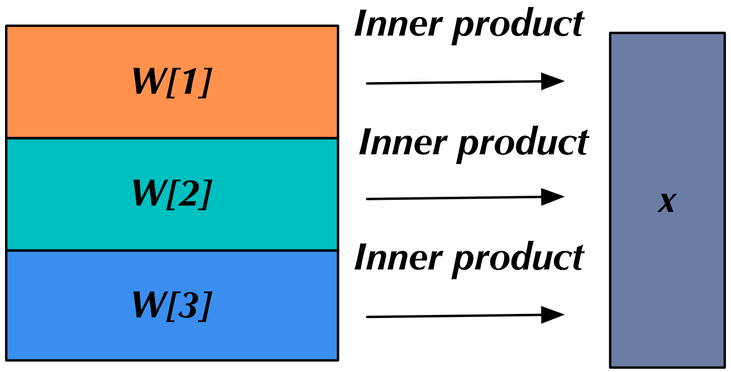
- Data shopper uploads encrypted model and retrieves prediction values to select/validate data.

- The cloud helps the shopper to evaluate sellers' encrypted data.

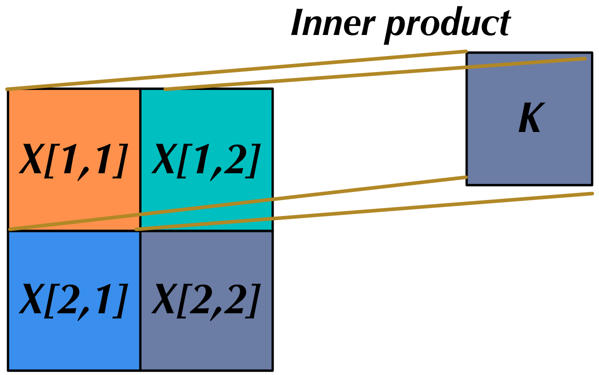


# Rationale behind Our ML Encryption

- We need **inner product computation over ciphertexts**.
  - For most neural networks, both common matrix and convolution computation can be decomposed to **inner product computation.**




*Decomposing matrix computation Wx*




*Decomposing convolution computation between K and X*

# Lightweight ML Encryption

- We use lightweight **inner-product functional encryption (IFE) and matrix transformation** to encrypt data/models.
  - ➢ Still can use encrypted model to predict/train encrypted data

# Lightweight ML Encryption

- We use lightweight **inner-product functional encryption (IFE) and matrix transformation** to encrypt data/models.
  - Still can use encrypted model to predict/train encrypted data

*The first layer*

*The second layer*

# Lightweight ML Encryption

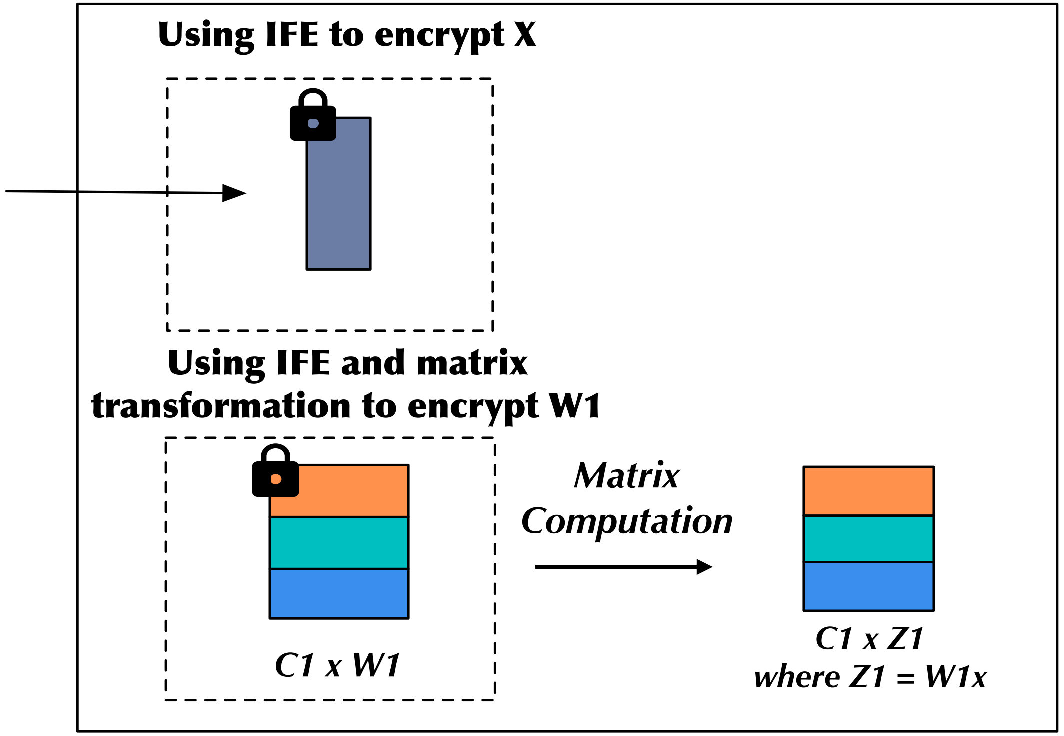
- We use lightweight **inner-product functional encryption (IFE) and matrix transformation** to encrypt data/models.
    - ➢ Still can use encrypted model to predict/train encrypted data


*The first layer*

*The second layer*

# Lightweight ML Encryption

- We use lightweight **inner-product functional encryption (IFE) and matrix transformation** to encrypt data/models.
  - ➢ Still can use encrypted model to predict/train encrypted data

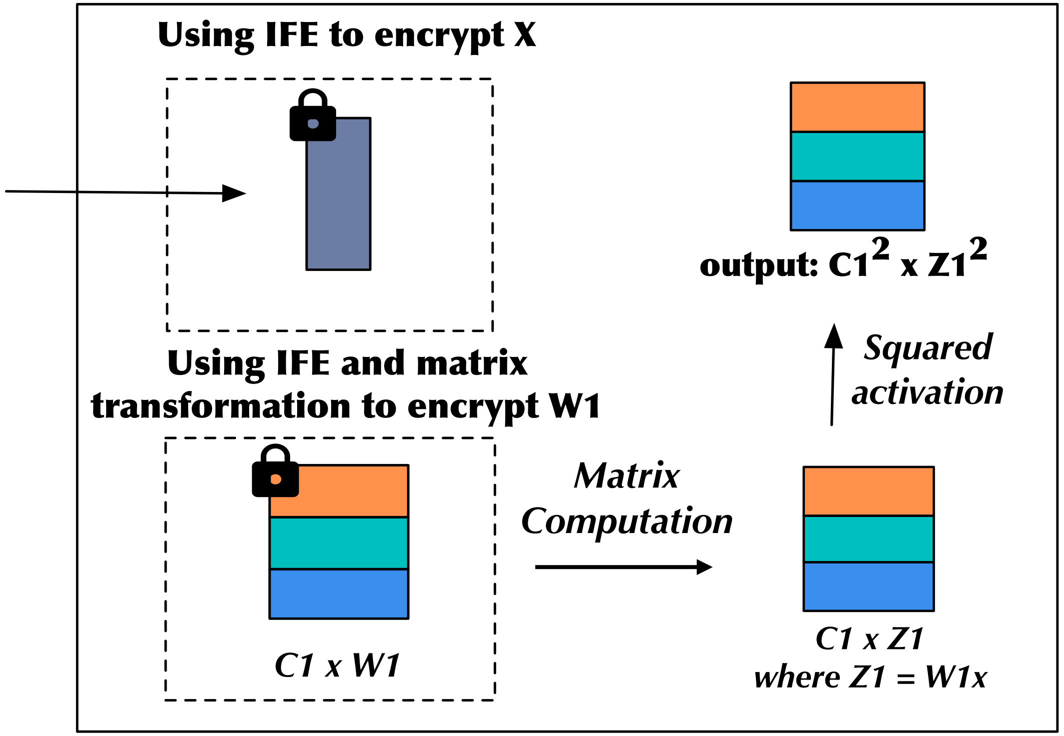



*The first layer*

*The second layer*

# Lightweight ML Encryption

- We use lightweight **inner-product functional encryption (IFE) and matrix transformation** to encrypt data/models.
  - ➢ Still can use encrypted model to predict/train encrypted data

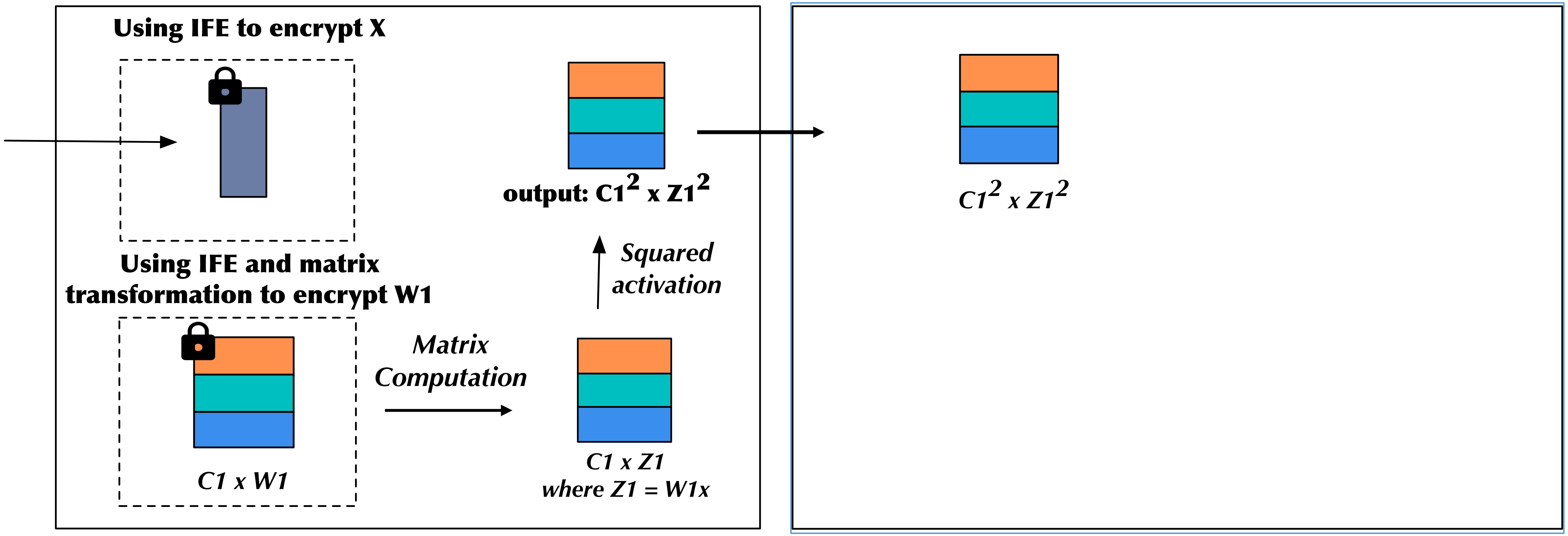



*The first layer*

*The second layer*

# Lightweight ML Encryption

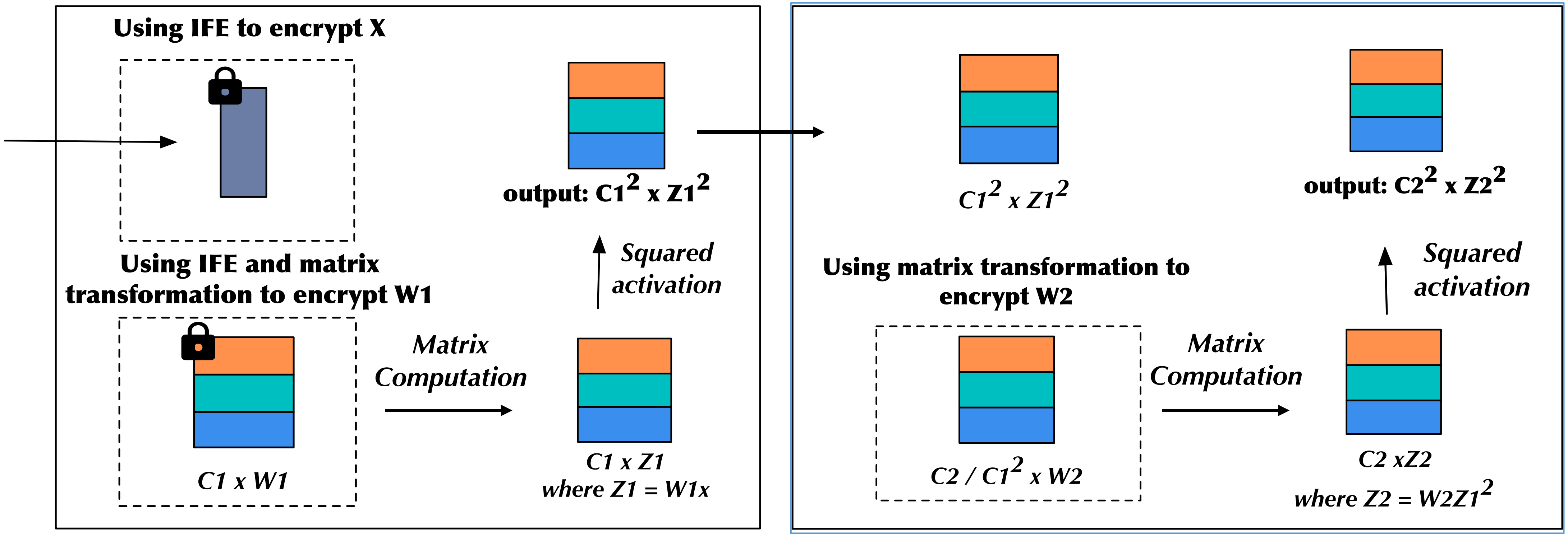
- We use lightweight **inner-product functional encryption (IFE) and matrix transformation** to encrypt data/models.
  - ➢ Still can use encrypted model to predict/train encrypted data




*The first layer* *The second layer*

# Lightweight ML Encryption

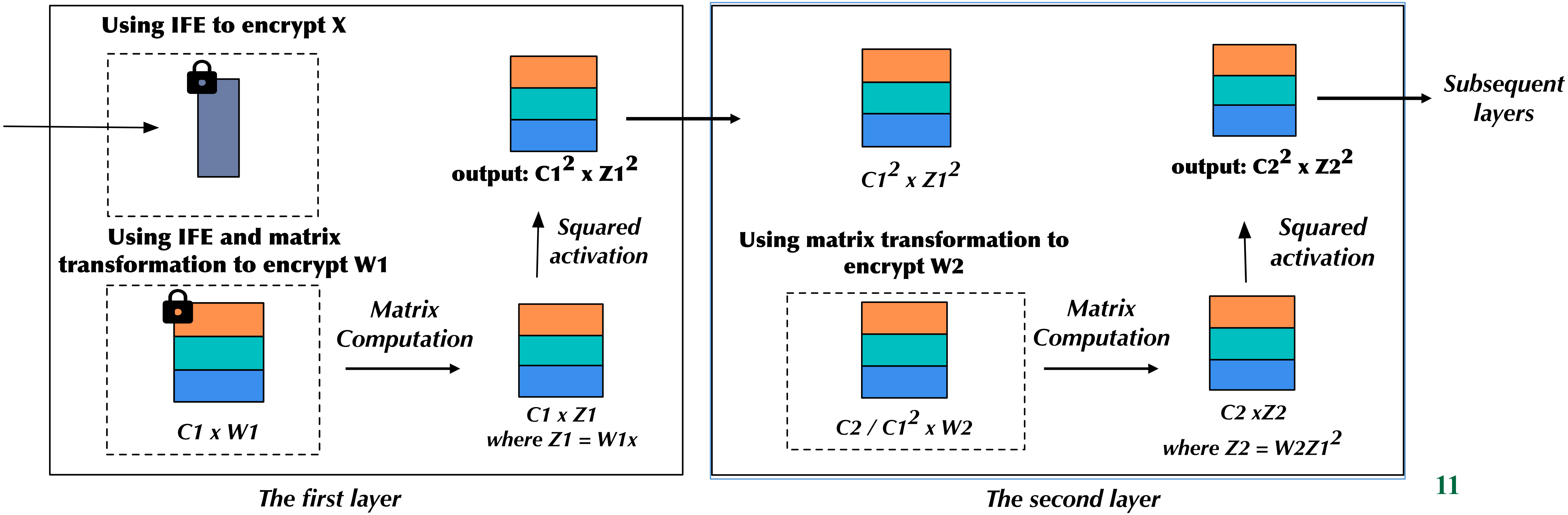
- We use lightweight **inner-product functional encryption (IFE) and matrix transformation** to encrypt data/models.
  - ➢ Still can use encrypted model to predict/train encrypted data




*The first layer* *The second layer*

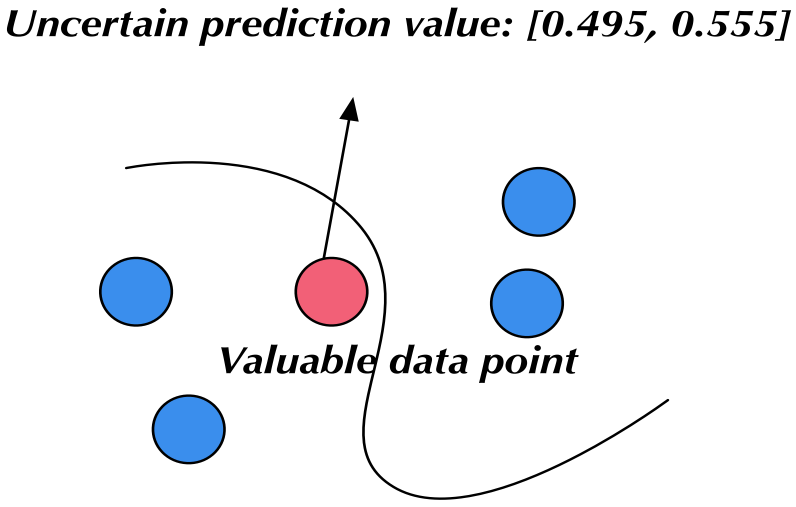
# Lightweight ML Encryption

- We use lightweight **inner-product functional encryption (IFE) and matrix transformation** to encrypt data/models.
  - ➢ Still can use encrypted model to predict/train encrypted data



# Rationale behind Data Selection

- Data selection is based on active learning.
- Active learning uses prediction values **(not original data)** to evaluate data.
- Valuable data have **uncertain prediction values.**
  - ➢ located near the decision boundary, i.e., provide more information



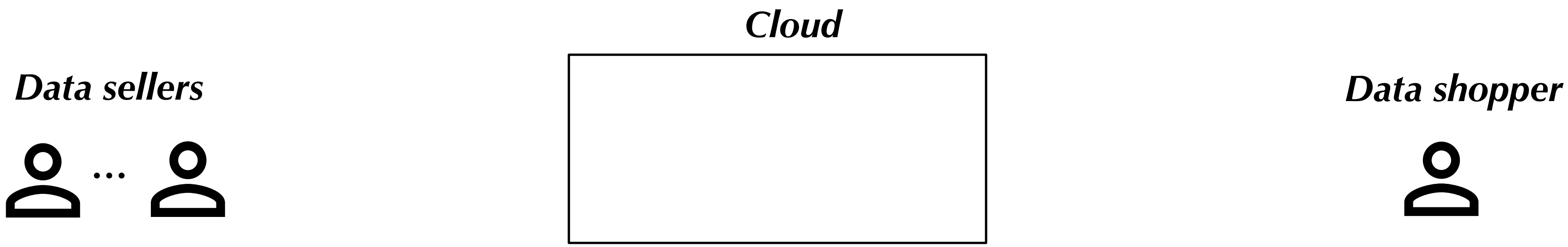

*Active learning*

# Data Selection

1. Data sellers and shopper upload their encrypted data and model.
2. The cloud performs prediction operations.
3. Data shopper collects encrypted prediction values to select valuable data.

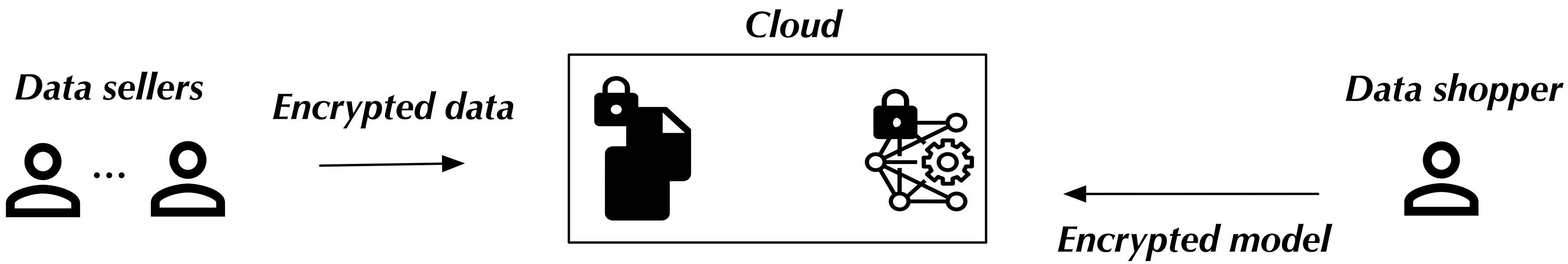
# Data Selection

1. Data sellers and shopper upload their encrypted data and model.
2. The cloud performs prediction operations.
3. Data shopper collects encrypted prediction values to select valuable data.



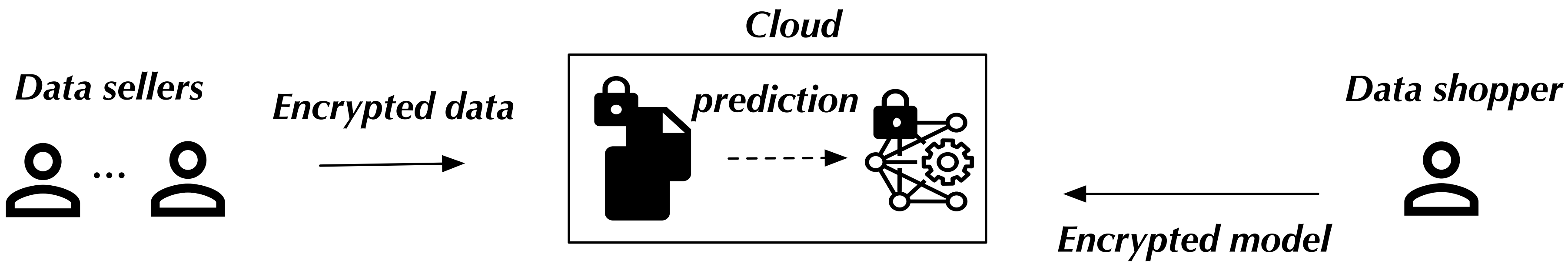
# Data Selection

1. Data sellers and shopper upload their encrypted data and model.
2. The cloud performs prediction operations.
3. Data shopper collects encrypted prediction values to select valuable data.



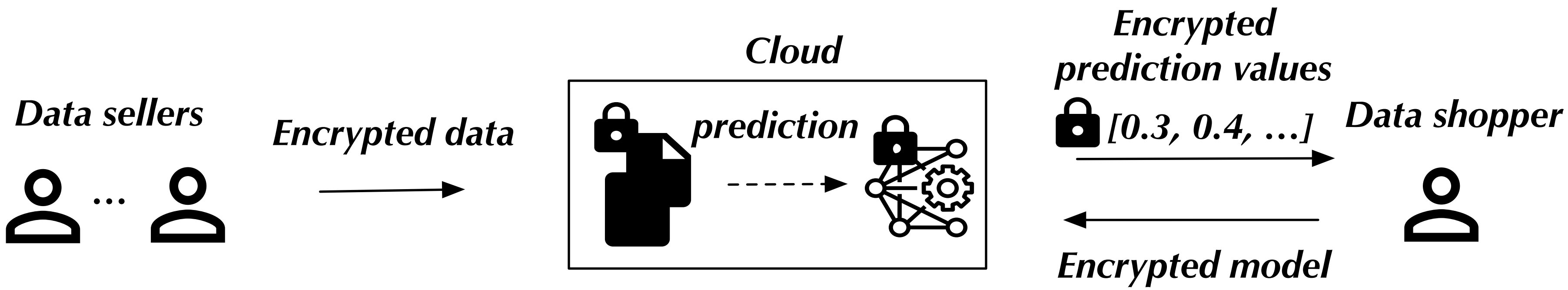
# Data Selection

1. Data sellers and shopper upload their encrypted data and model.
2. The cloud performs prediction operations.
3. Data shopper collects encrypted prediction values to select valuable data.



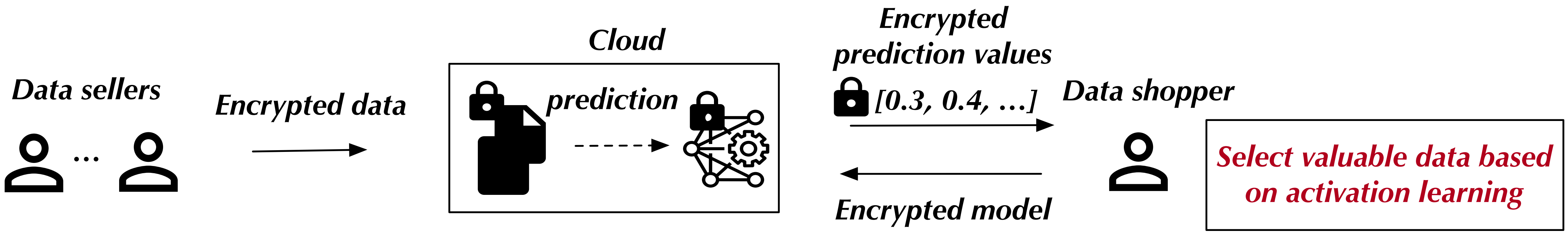
# Data Selection

1. Data sellers and shopper upload their encrypted data and model.
2. The cloud performs prediction operations.
3. Data shopper collects encrypted prediction values to select valuable data.



# Data Selection

1. Data sellers and shopper upload their encrypted data and model.
2. The cloud performs prediction operations.
3. Data shopper collects encrypted prediction values to select valuable data.



# Another Problem:

Data selection only considers the informativeness of data, but not labels, not relevance.

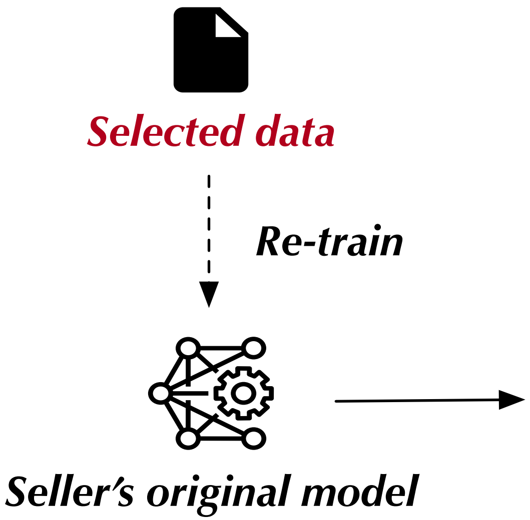
What if the selected data contain **unintentionally mislabeled data or irrelevant data?**

# Rationale behind Data Validation

- The shopper and cloud **cannot directly see selected data** to estimate quality.
- Indirect approach:  let the model "try" data and check model performance.
  - "try" : use the selected data to retrain the shopper' s model.

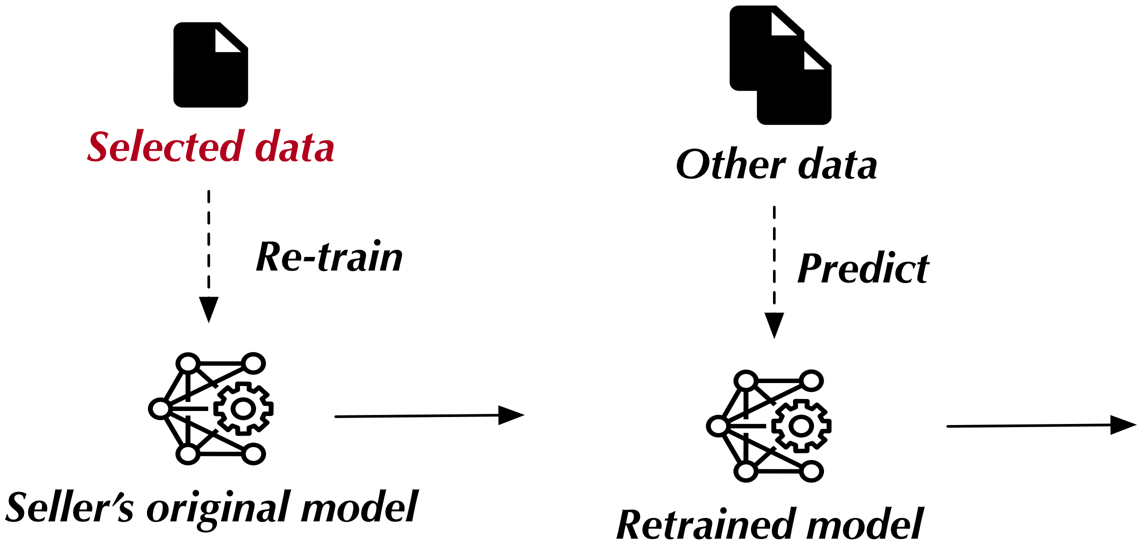
# Rationale behind Data Validation

- The shopper and cloud **cannot directly see selected data** to estimate quality.
- Indirect approach:  let the model "try" data and check model performance.
  - "try" : use the selected data to retrain the shopper' s model.



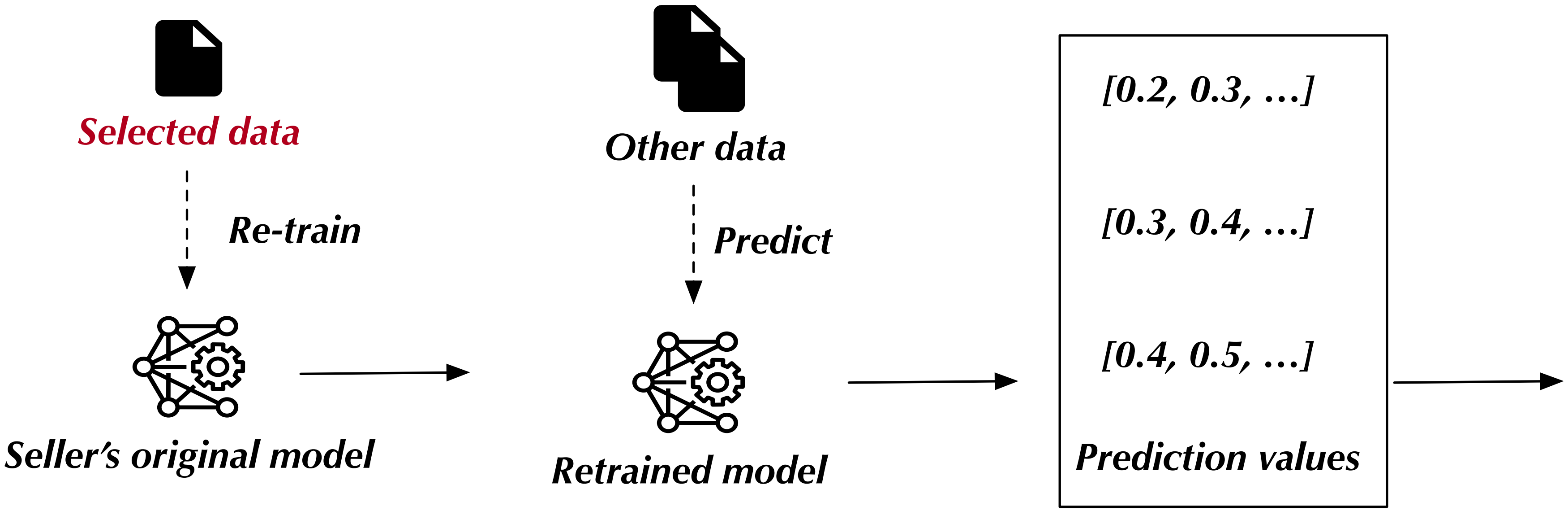
# Rationale behind Data Validation

- The shopper and cloud **cannot directly see selected data** to estimate quality.
- Indirect approach:  let the model "try" data and check model performance.
  - ➢ "try" : use the selected data to retrain the shopper' s model.



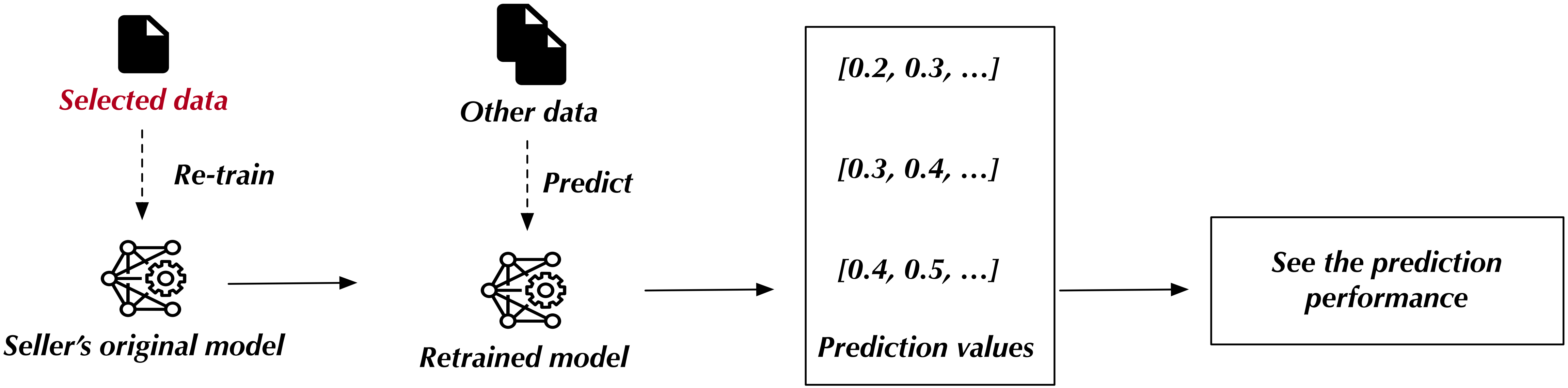
# Rationale behind Data Validation

- The shopper and cloud **cannot directly see selected data** to estimate quality.
- Indirect approach: let the model "try" data and check model performance.
  - "try" : use the selected data to retrain the shopper' s model.



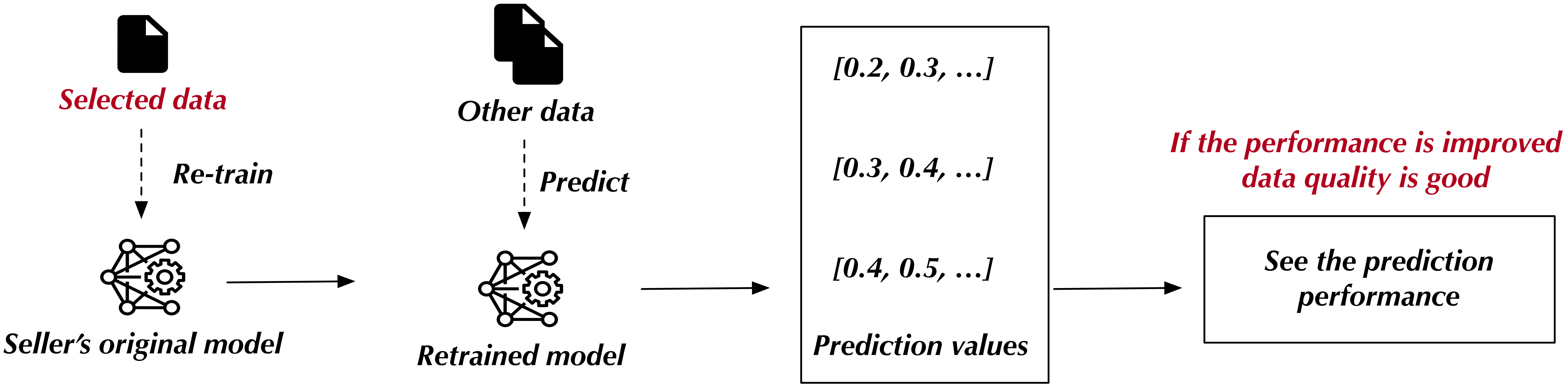
# Rationale behind Data Validation

- The shopper and cloud **cannot directly see selected data** to estimate quality.
- Indirect approach: let the model "try" data and check model performance.
  - ➢ "try" : use the selected data to retrain the shopper' s model.



# Rationale behind Data Validation

- The shopper and cloud **cannot directly see selected data** to estimate quality.
- Indirect approach:  let the model "try" data and check model performance.
  - ➢ "try" : use the selected data to retrain the shopper' s model.



# Data Validation

1. Cloud uses the selected data to retrain the shopper' s encrypted model.
2. Cloud uses the retrained model to predict uniformly distributed data.
3. The shopper **collects encrypted prediction values** to estimate data quality.

# Data Validation

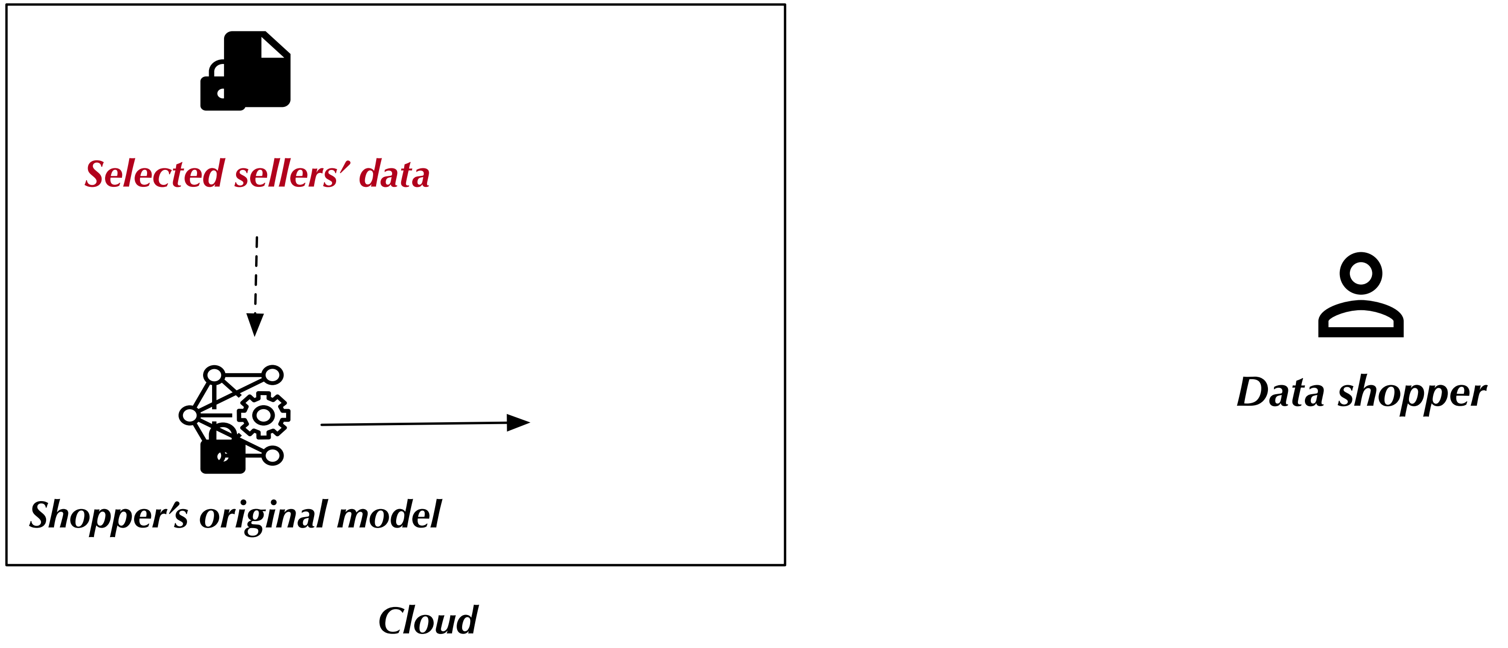
1. Cloud uses the selected data to retrain the shopper' s encrypted model.
2. Cloud uses the retrained model to predict uniformly distributed data.
3. The shopper **collects encrypted prediction values** to estimate data quality.

*Data shopper*

*Cloud*

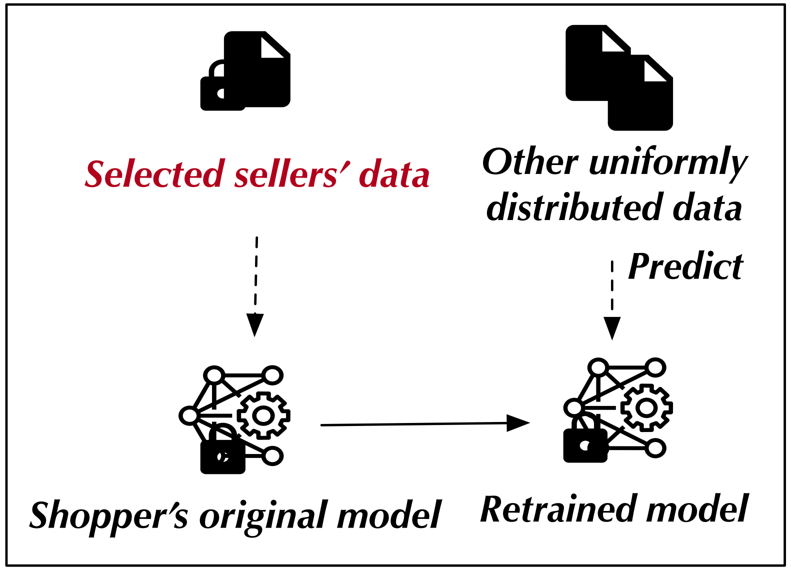
# Data Validation

1. Cloud uses the selected data to retrain the shopper' s encrypted model.
2. Cloud uses the retrained model to predict uniformly distributed data.
3. The shopper **collects encrypted prediction values** to estimate data quality.

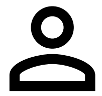


# Data Validation

1. Cloud uses the selected data to retrain the shopper' s encrypted model.
2. Cloud uses the retrained model to predict uniformly distributed data.
3. The shopper **collects encrypted prediction values** to estimate data quality.



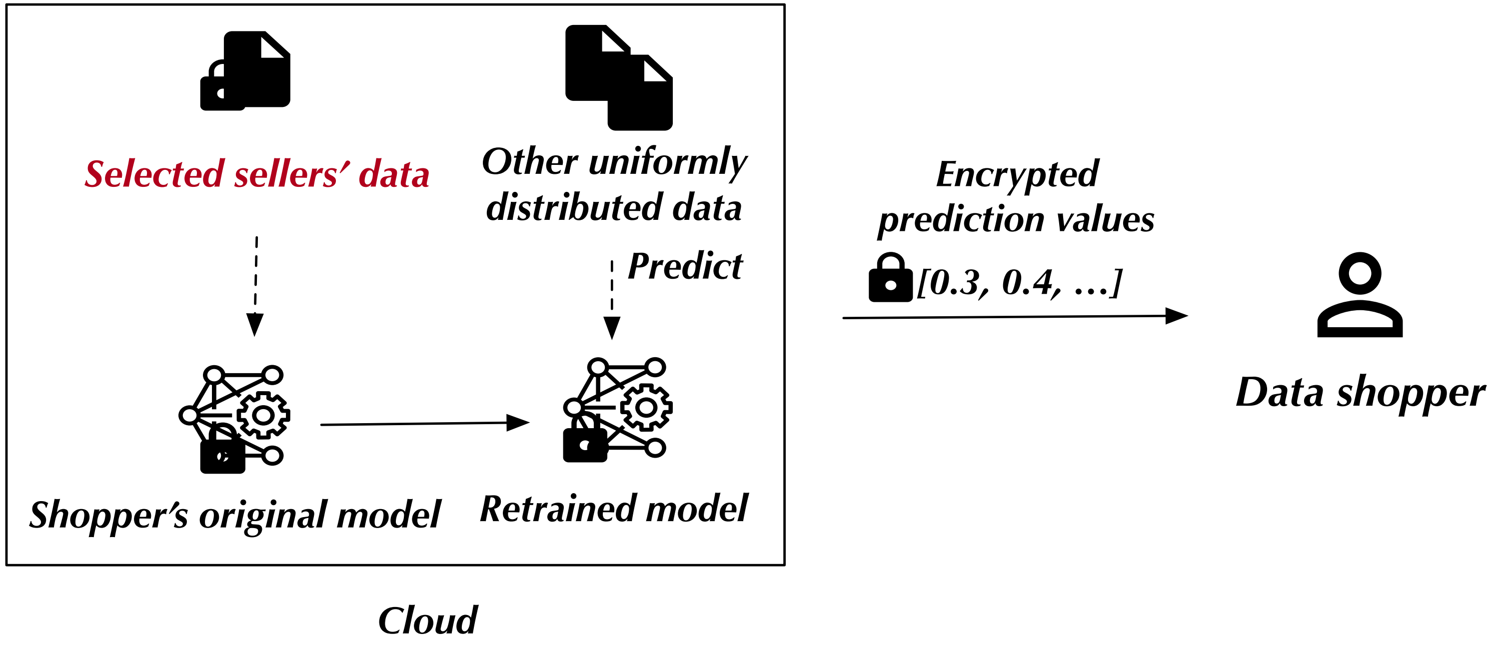

Cloud



Data shopper

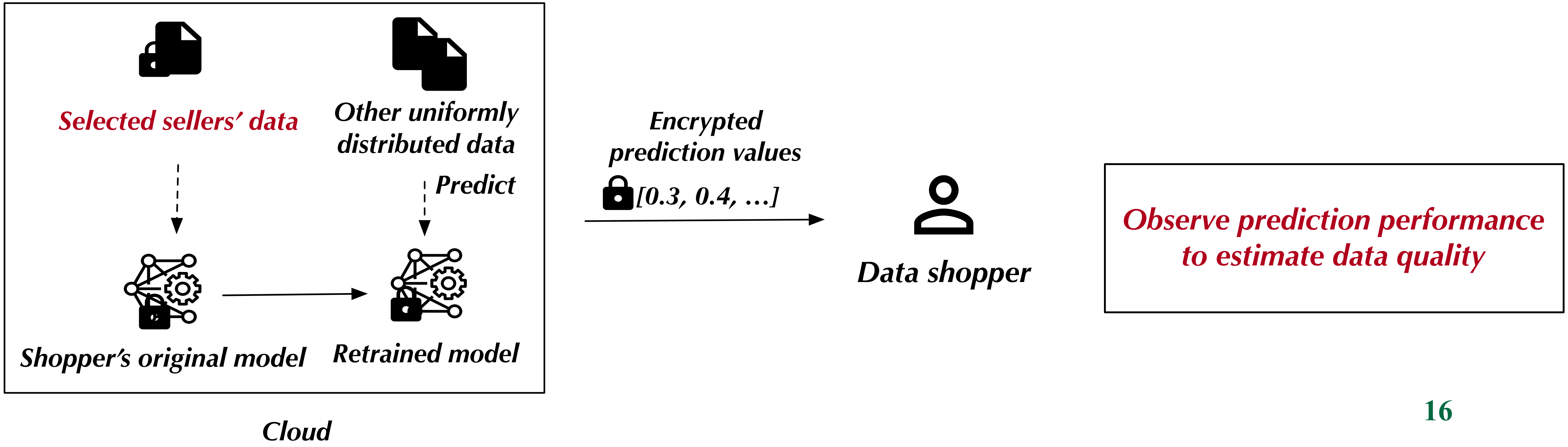
# Data Validation

1. Cloud uses the selected data to retrain the shopper' s encrypted model.
2. Cloud uses the retrained model to predict uniformly distributed data.
3. The shopper **collects encrypted prediction values** to estimate data quality.



# Data Validation

1. Cloud uses the selected data to retrain the shopper' s encrypted model.
2. Cloud uses the retrained model to predict uniformly distributed data.
3. The shopper **collects encrypted prediction values** to estimate data quality.

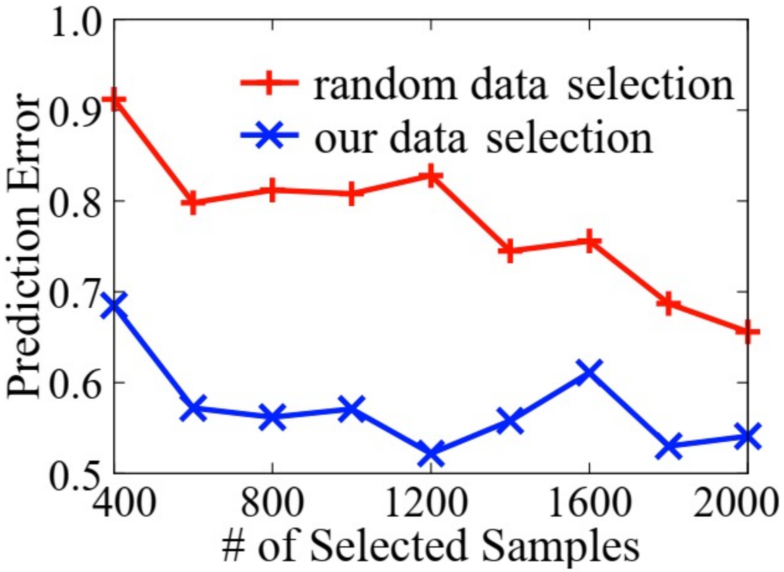


# Experiment Setup

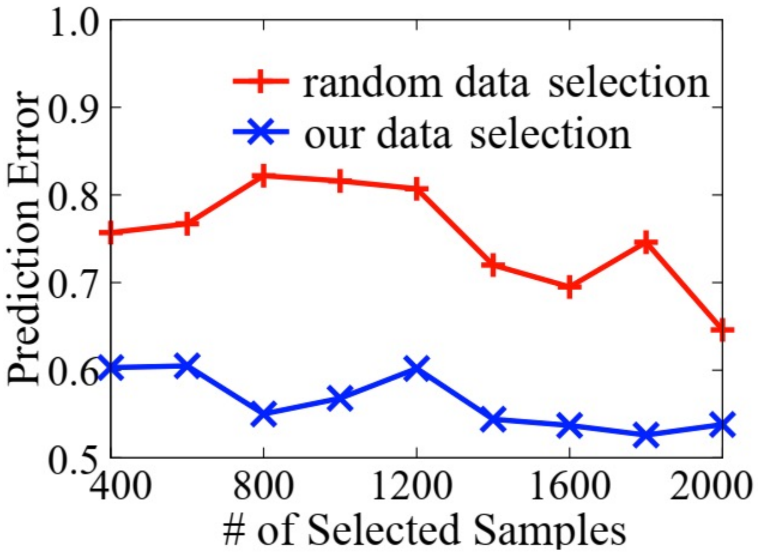
- We simulate
  - 100 sellers and 1 shopper
  - divide MINIST to 101 subsets, assigned to sellers and shopper
- We evaluate
  - benefits of our data selection
  - the accuracy of our data validation
  - computational overhead

# Benefits of Our Data Selection

- Compared with random selection, our data selection can reduce about 60% prediction errors.




(a) Model 1

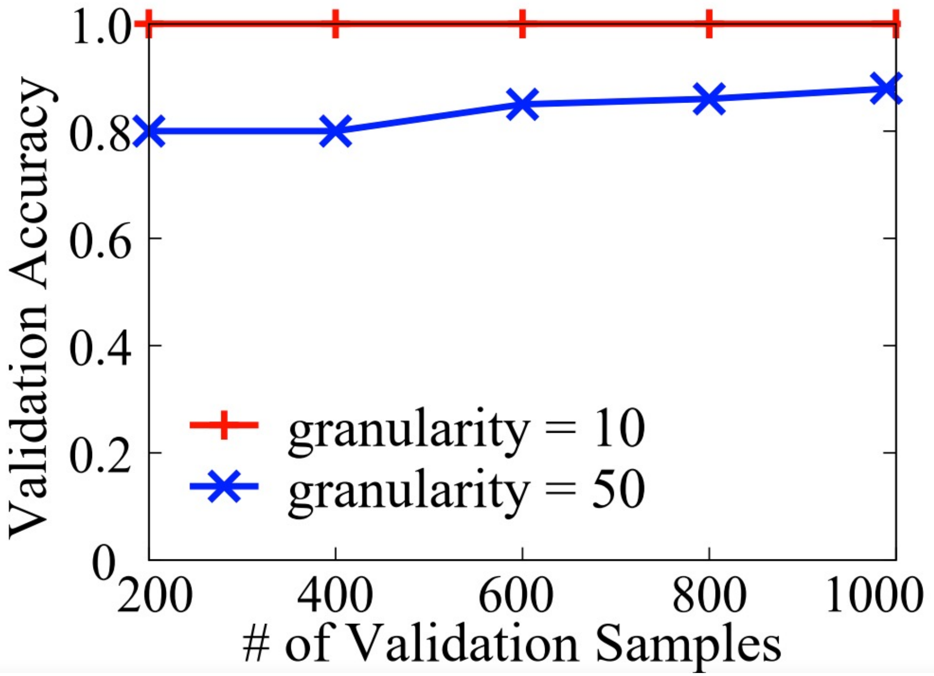



(b) Model 2

* Model 1 and 2 are trained with 5500 and 55000 samples, respectively.

# Accuracy of Our Data Validation

- Simulate low-quality samples that are most likely to evade data validation




* We split samples into multiple subsets and validate them one by one.
* Validation granularity means that the size of validation subsets.

# Computational Overhead

- Compared with homomorphic encryption based approach (E2DM)

**Table 1: Execution Time of CNN models**

| Operations | Execution Time (second) | |
|---|---|---|
| | E2DM | Ours |
| Data Encryption | 0.40 | 0.48 |
| Model Encryption | 0.14 | 0.20 |
| Feed Forward | 35.88 | 2.59 |
| Back Propagation | N/A | 0.05 |

* We encrypt a six-layer CNN model and measure relevant operations

# Conclusion

- A privacy-preserving and efficient ML data evaluation framework on data marketplaces

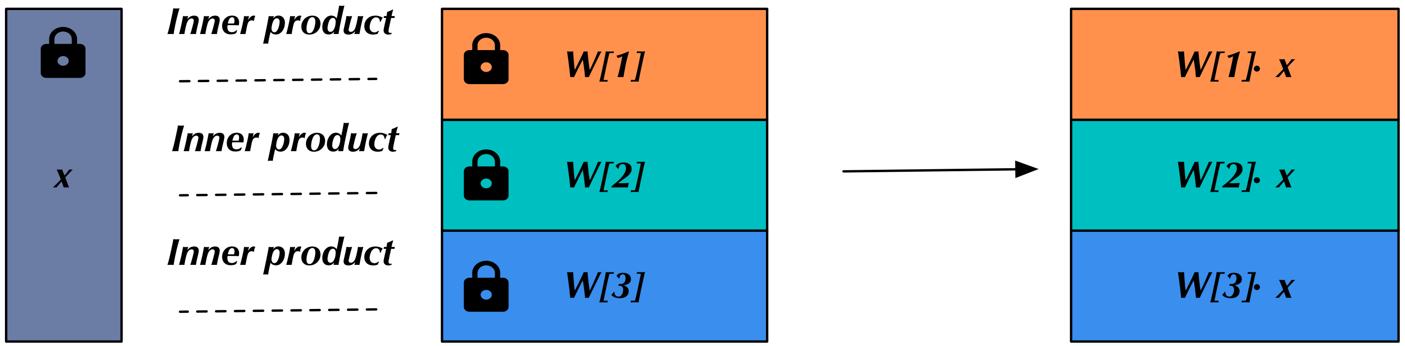
- A new lightweight ML encryption protocol that can preserve both privacy and functionality of data/models on the cloud
  - Based on IFE and matrix transformation

- Privacy-preserving Data Selection and Validation
  - Can select valuable data and validate the data quality
  - Do not disclose the original data and models

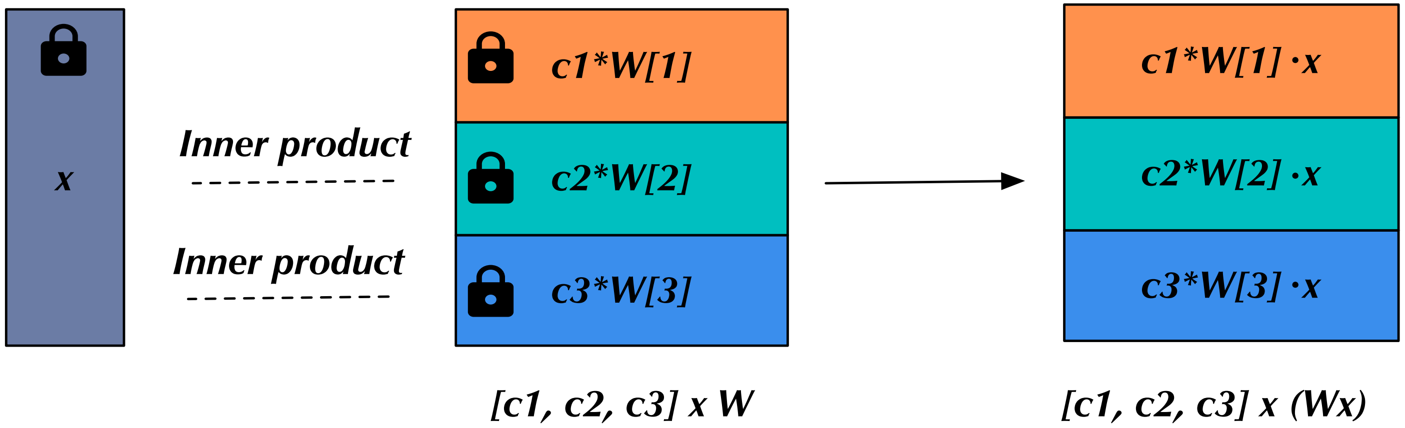
# Thank you!

# Backup: IFE-based Matrix Encryption

- We can use **inner product functional encryption** to enable matrix or convolution computation over ciphertexts.

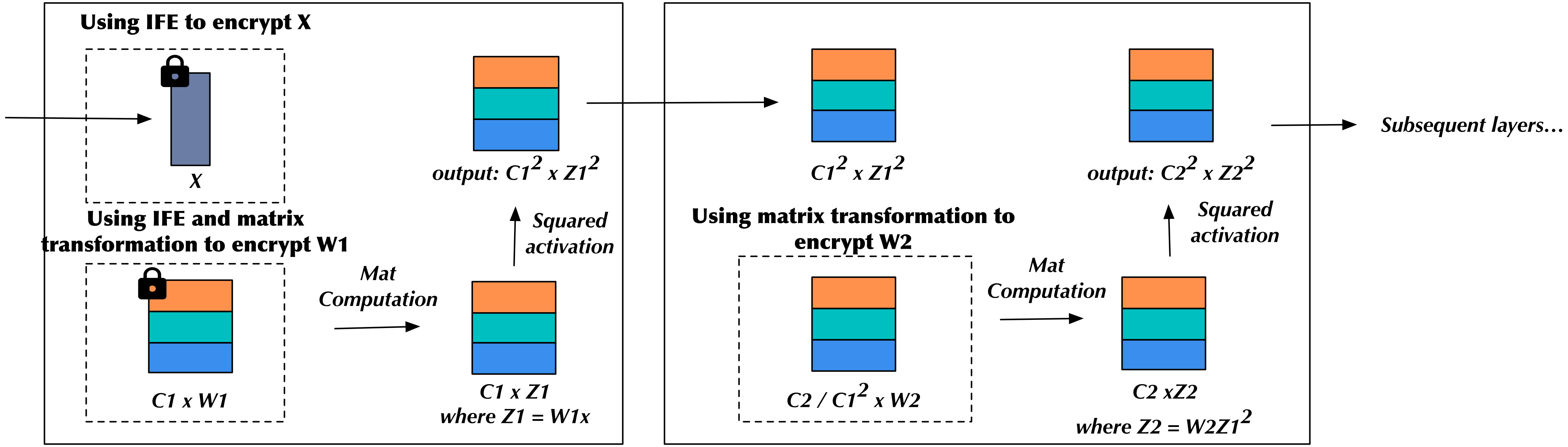



- The result is plaintext, we apply **matrix transformation** to hide the result.



# Backup: Our ML Encryption

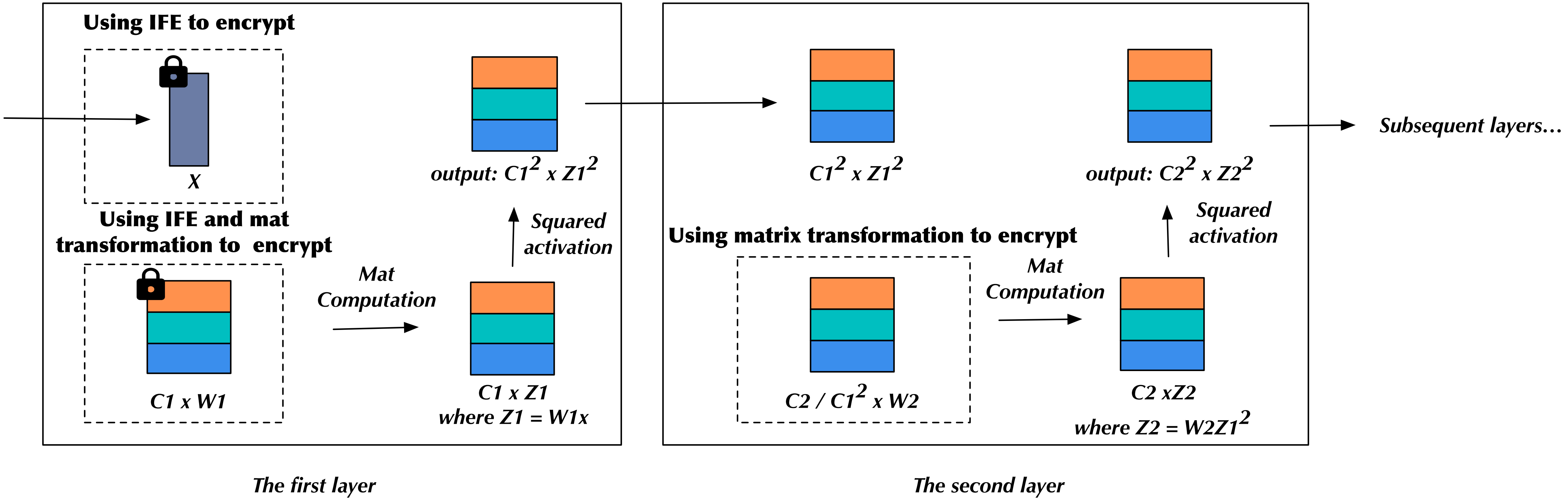
- IFE is only used to encrypt the first layer since it only support simple inner product computation.
- Remaining layers are encrypted by matrix transformation (see our paper).




*The first layer* *The second layer*

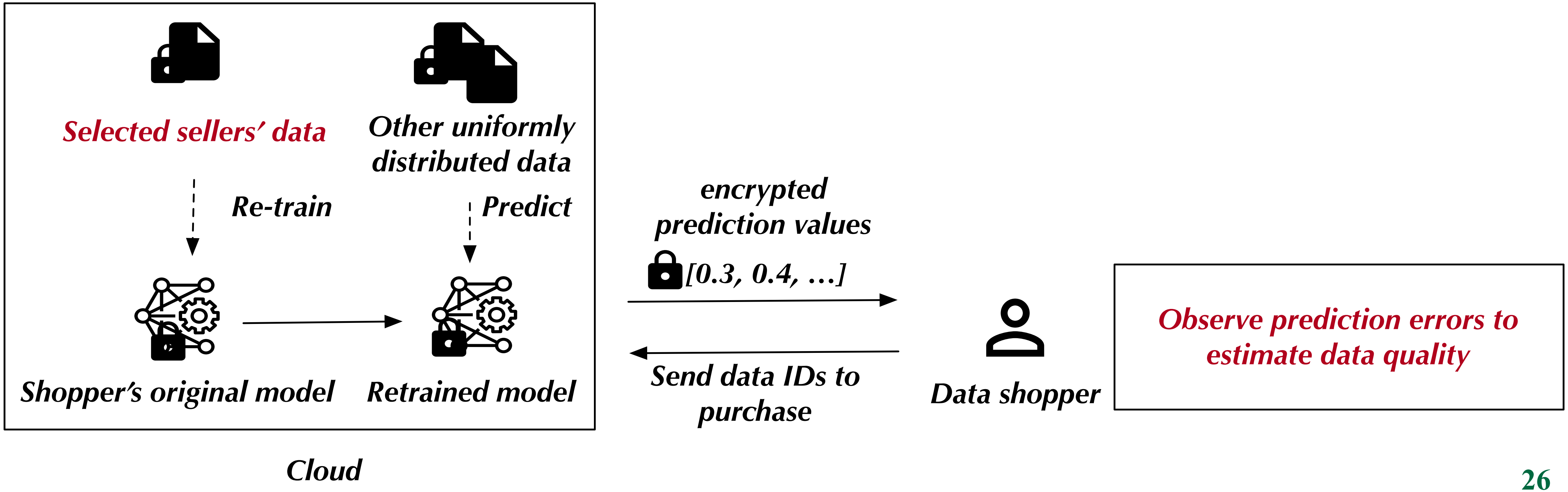
# Backup: Our ML Encryption

- During prediction, the output of each layer is $C_i^2 \times Z_i^2$ ( $Z_i^2$ is original output).
- We can decrypt the output by multiply $C_i^{-2}$.



# Backup: Data Validation

- To evaluate data of different values, we set a variable threshold T.
- T is often the previous prediction errors. If the current prediction errors < T, we can say the performance is improved, and the data quality is good.



# Backup: TEE

- TEE may leak some sensitive information.
  - Cache Attacks
  - Fault injection attacks
- TEE has some memory limits.
  - For SGX, 128 MB