Improved Fully Homomorphic Encryption and Its Application to Private AI

Joon-Woo Lee

School of Computer Science and Engineering

Chung-Ang University

August 7, 2023

Contents

✓ Introduction

✓ Recent Research Results

✓ Future Research Plans

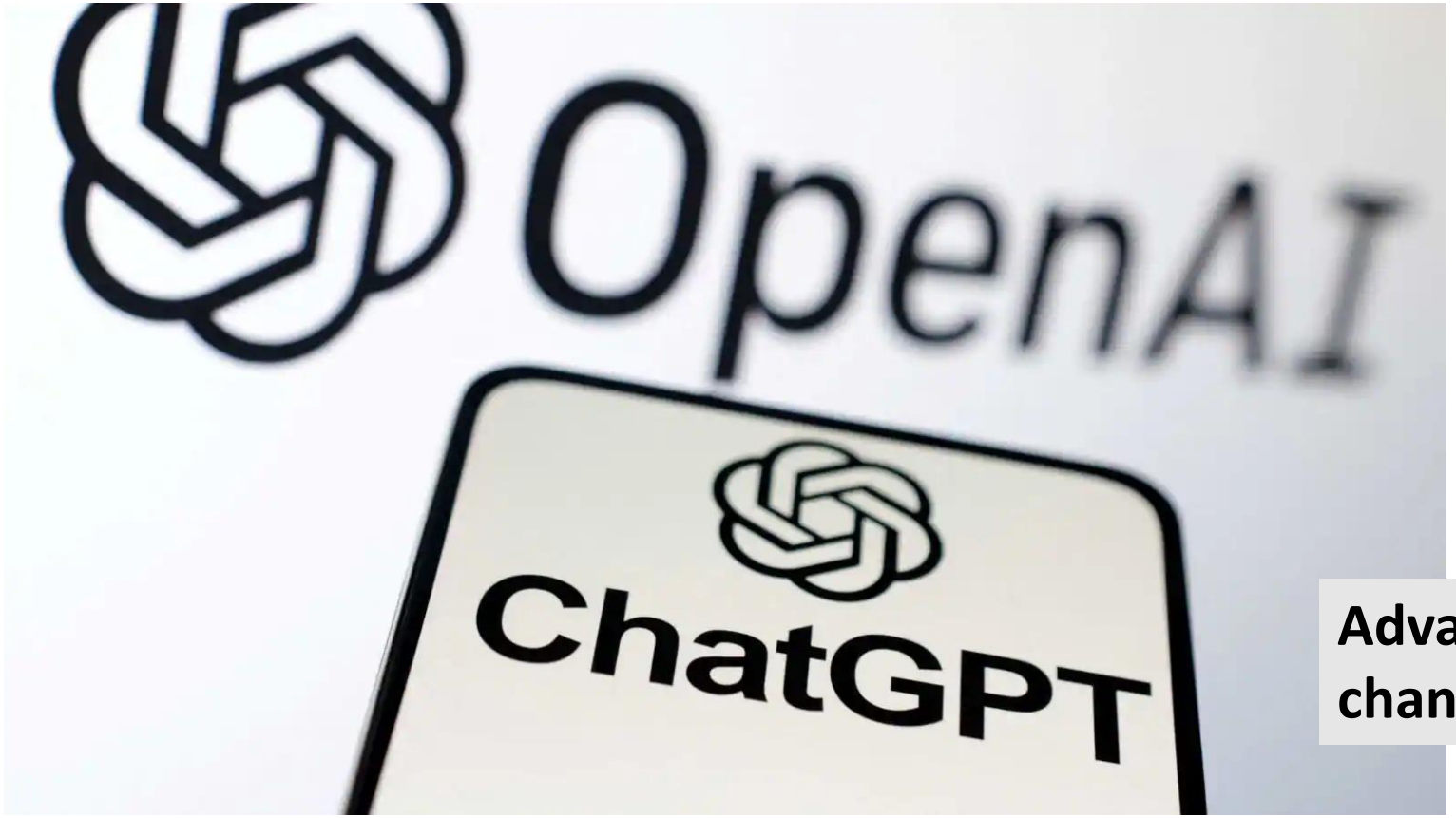
Contents

✓ Introduction

✓ Recent Research Results

✓ Future Research Plans

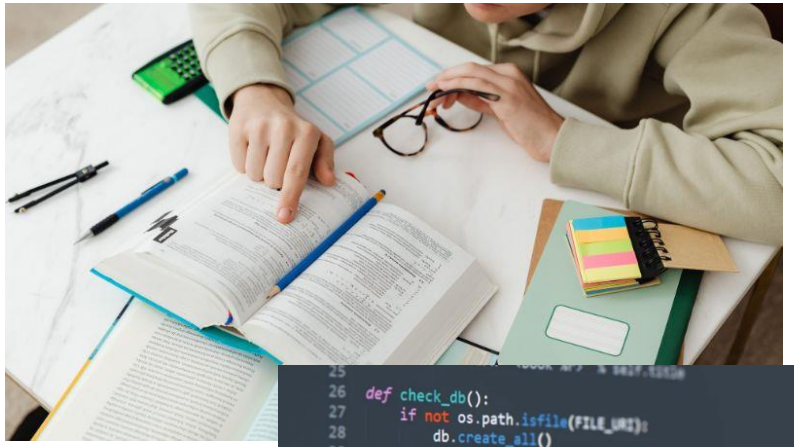
Necessity of Private AI



Advanced AI models drastically changes our lifes and industry!

Necessity of Private AI

When we study…



When we develop…

```
def check_db():
    if not os.path.isfile(FILE_URI):
        db.create_all()

@app.route("/")
def home():
    check_db()
    all_books = db.session.query(Book).all()
    return render_template("index.html", books=all_books)

@app.route("/edit", methods=["GET","POST"])
def edit():

    if request.method == "POST":
        book_id = request.form["id"]
        book_to_update = Book.query.get(book_id)
        book_to_update.rating = request.form["rating"]
        db.session.commit()
```

When we make ppt…

AI service can help our lifes with various tasks…

ChatGPT

Examples

"Explain quantum computing in simple terms" →

"Got any creative idea… year old's birthday…

"How do I make an HTT… in Javascript?" →

Capabilities

Remembers what user said earlier in the conversation

…requests

Limitations

May occasionally generate incorrect information

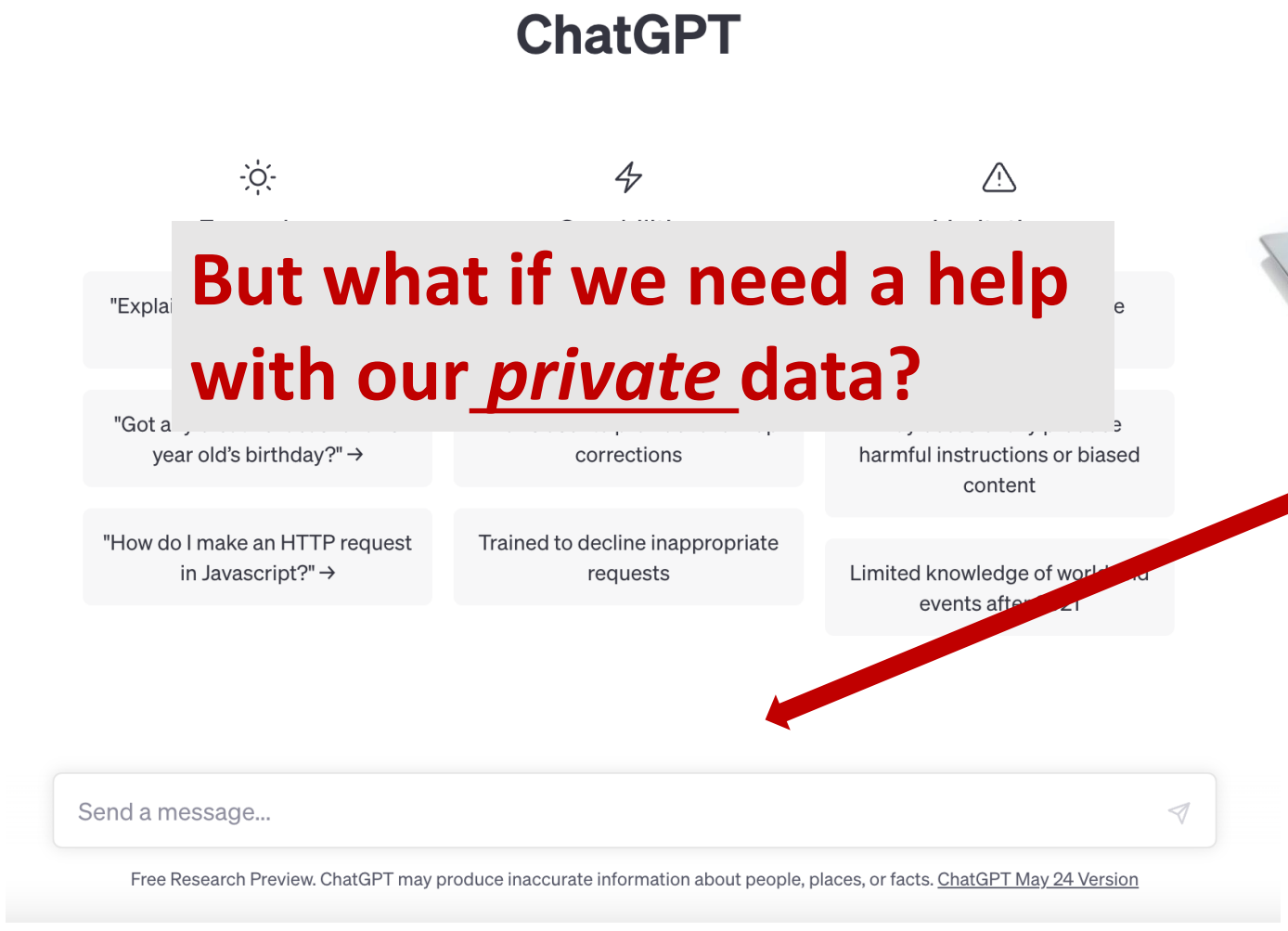
…produce …s or biased

Limited knowledge of world… events after…21

Send a message...

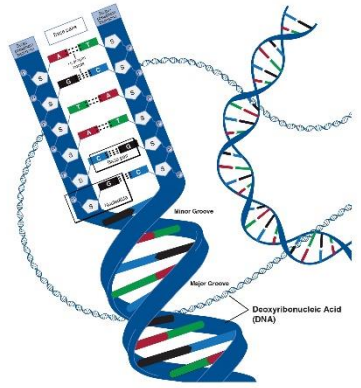
Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 24 Version

Necessity of Private AI



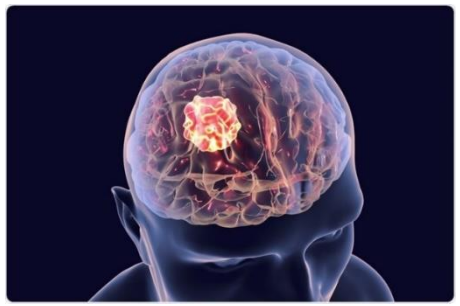
But what if we need a help with our *private* data?

Financial Data

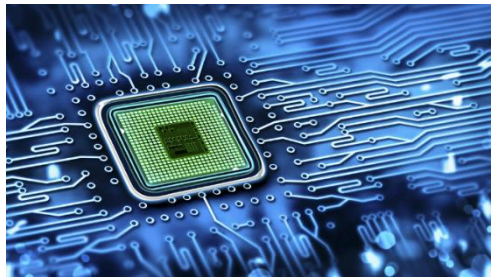


DNA sequence

Health condition



Semiconductor yield



Necessity of Private AI

Privacy ⟷ **Productivity**

We should not allow giving our private data to service provider!

But it sounds great if we receive AI service with the private data…

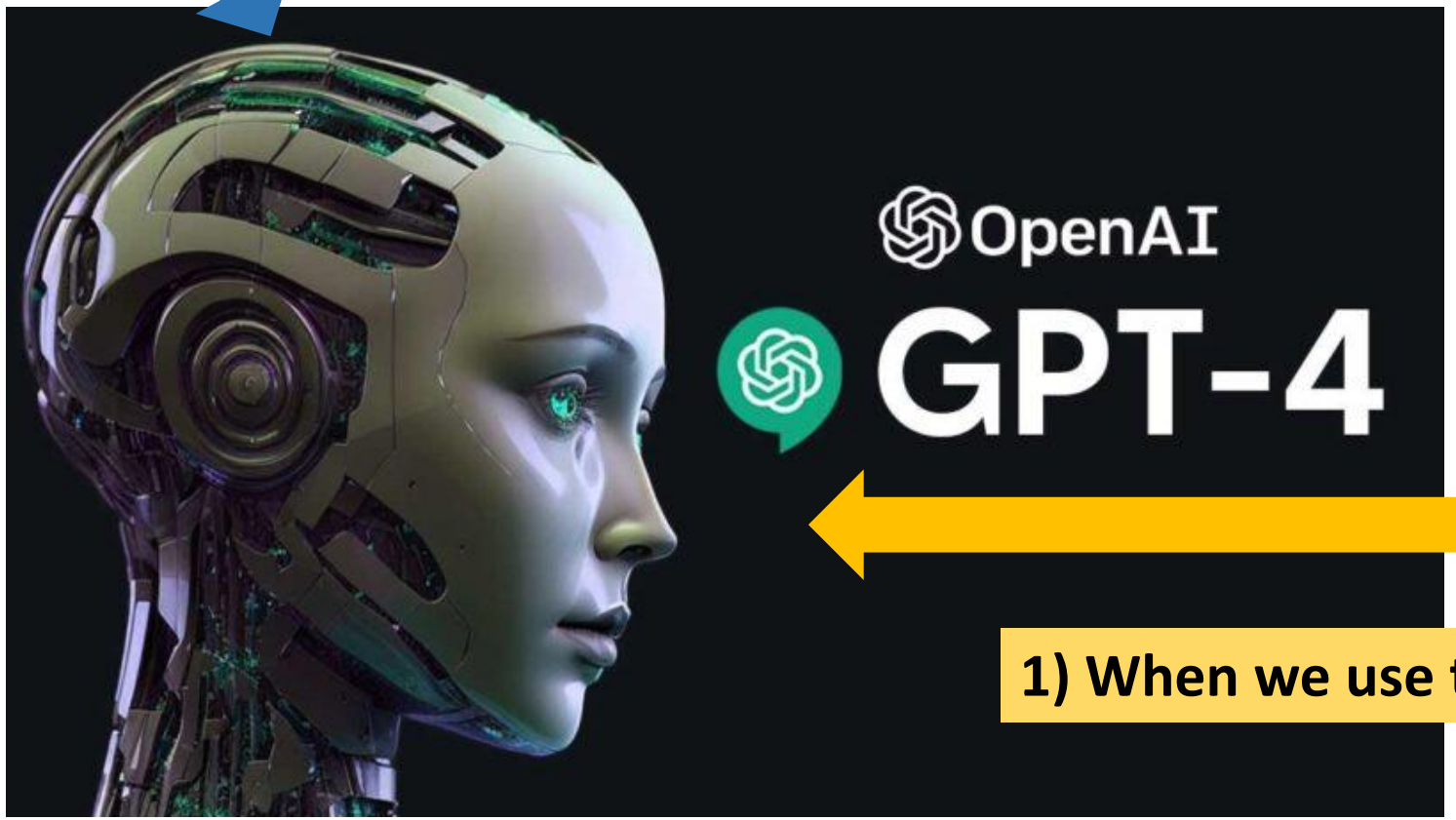
We face the dilemma!

Can we get the AI service without privacy infringement? **Private AI**

Two Types of Privacy Issue

Private dataset

2) When the company develops an improved AI model

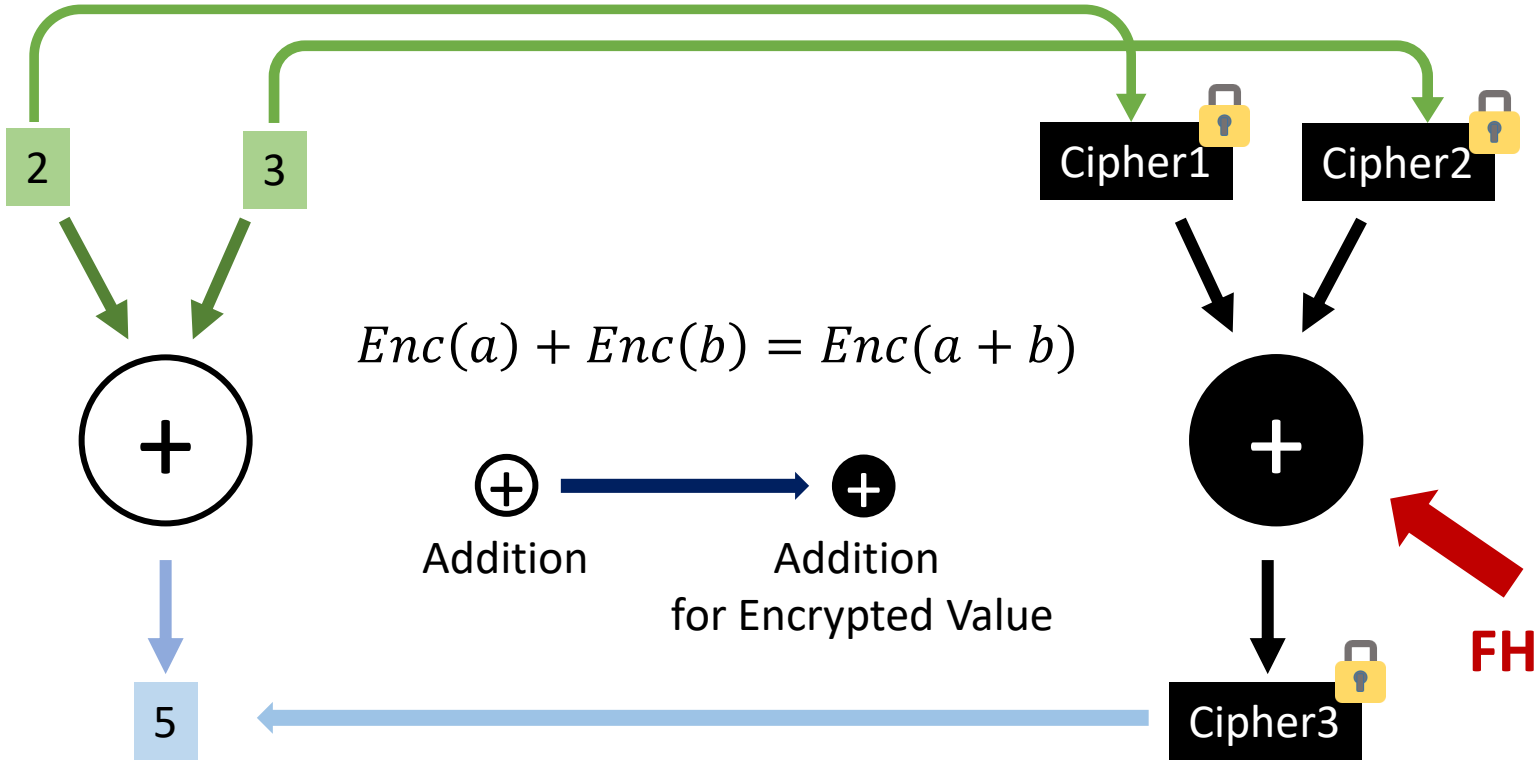


There are two types of privacy in private AI.

Private data

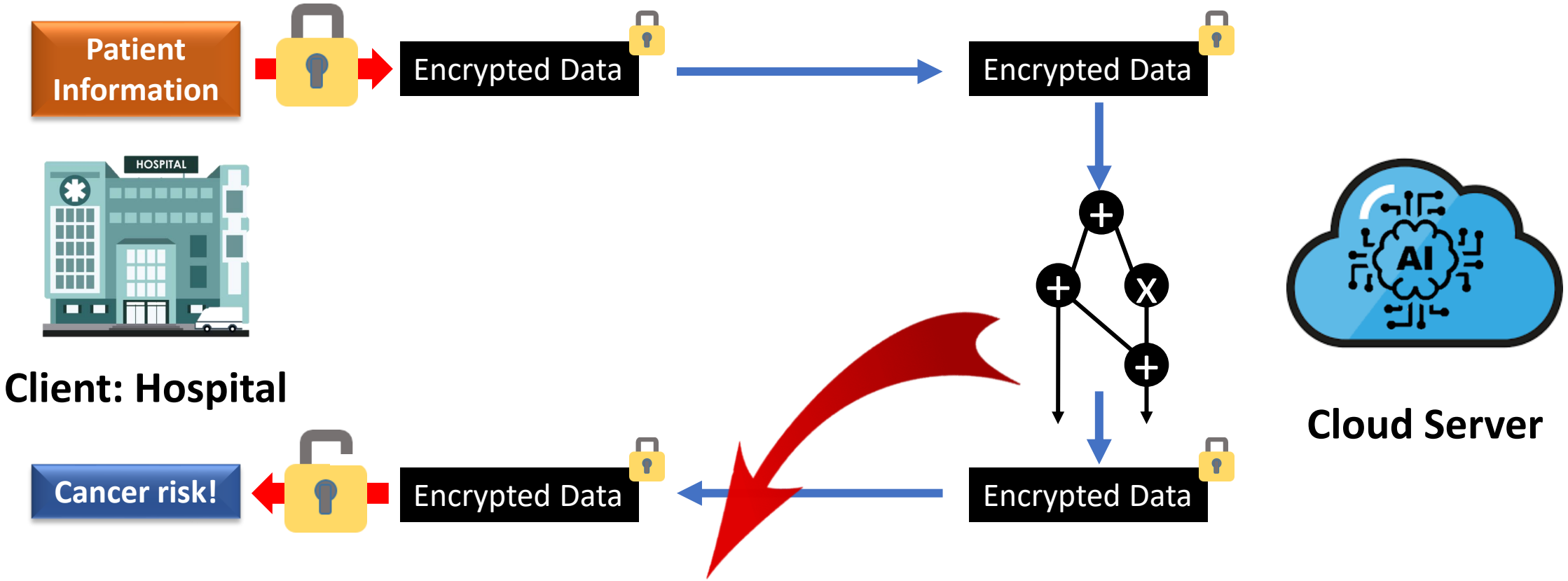
1) When we use the service with our private data

Fully Homomorphic Encryption (FHE)



Encryption scheme supporting **unlimited** **addition and multiplication on encrypted data**

Private AI System with Encrypted Data:
Fully Homomorphic Encryption



Computation with Encrypted Data

Fully Homomorphic Encryption!

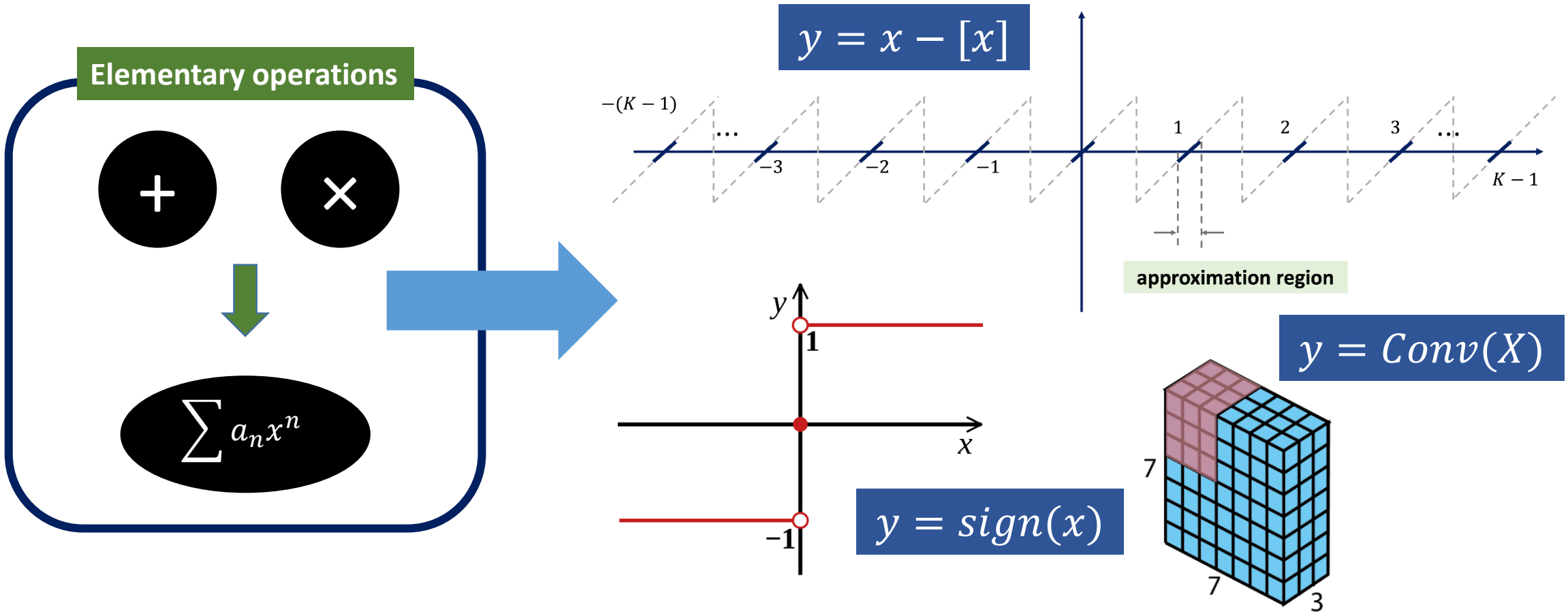
Main Problem in FHE

- However, the industry currently does not offer AI services with FHE scheme for the following three reason:

  1. Too long computation time
  2. Insufficient precision in computation
  3. Too large required memory in computation

- **Hence, the improvement and optimization of FHE is required for private AI services with FHE.**

Implementation of Advanced Operations



For enhanced crypto-based services, it is crucial to implement advanced operations on encrypted data.

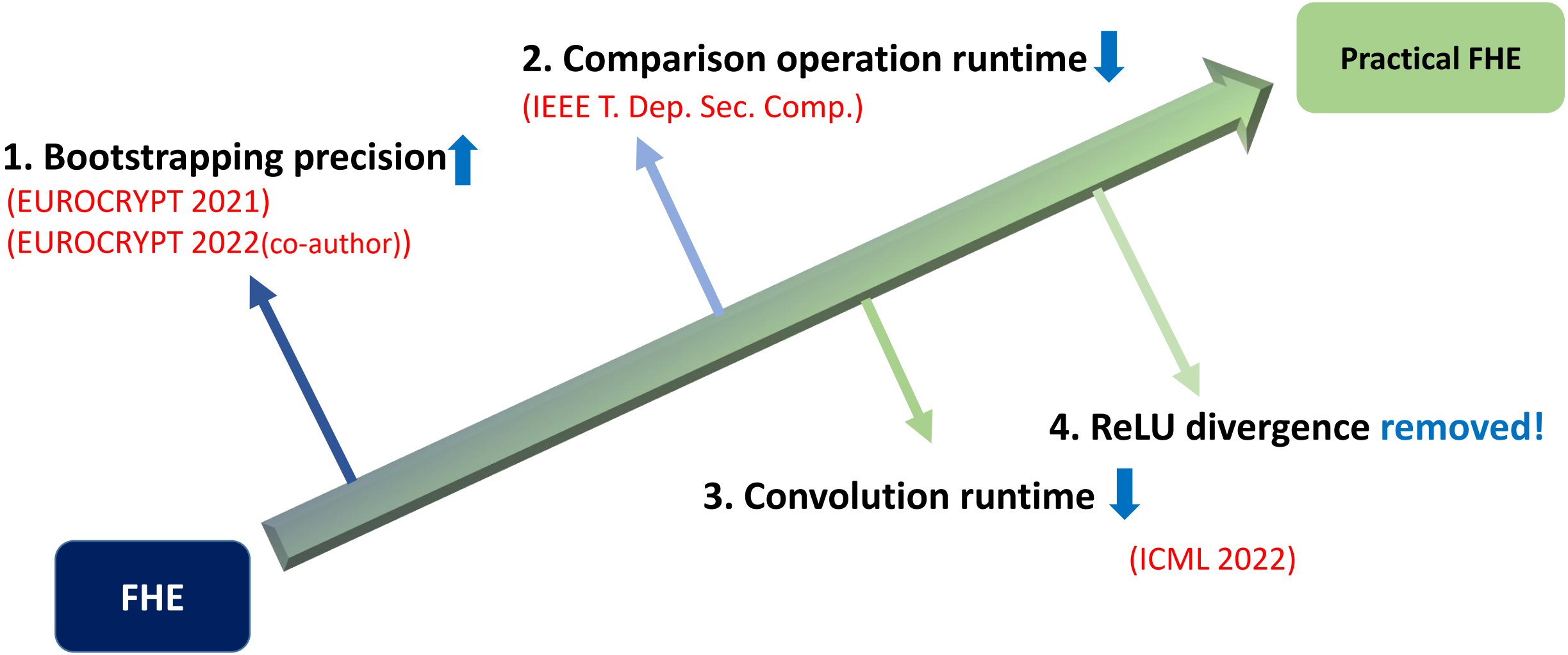
Contents

✓ Introduction

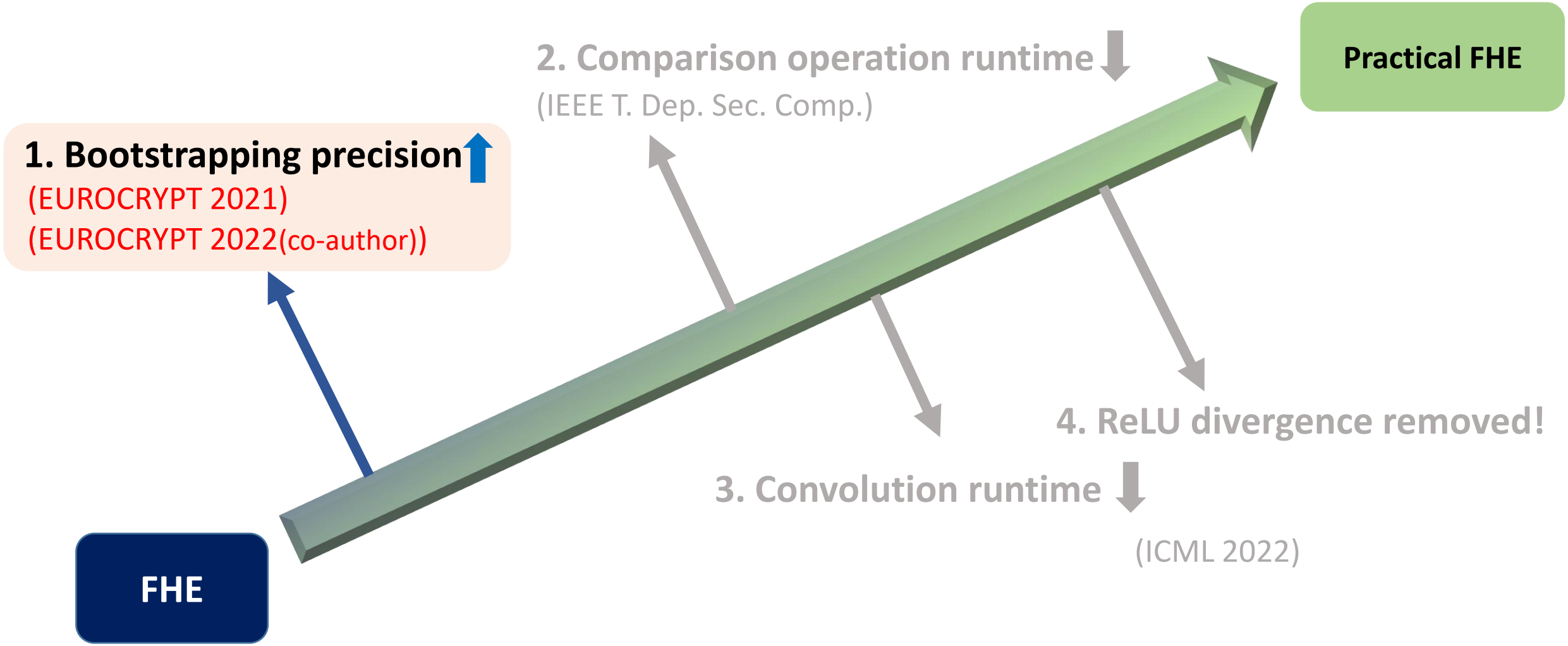
✓ Recent Research Results

✓ Future Research Plans

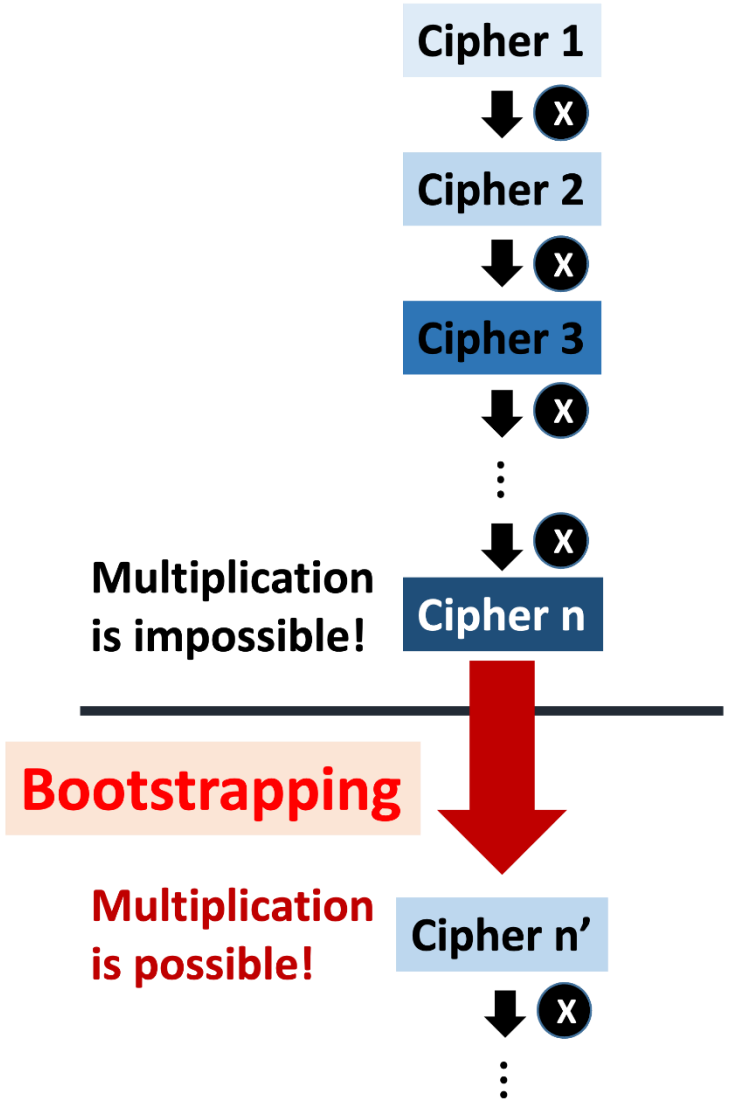
Research Goal: Improvement of FHE



Research Goal: Improvement of FHE

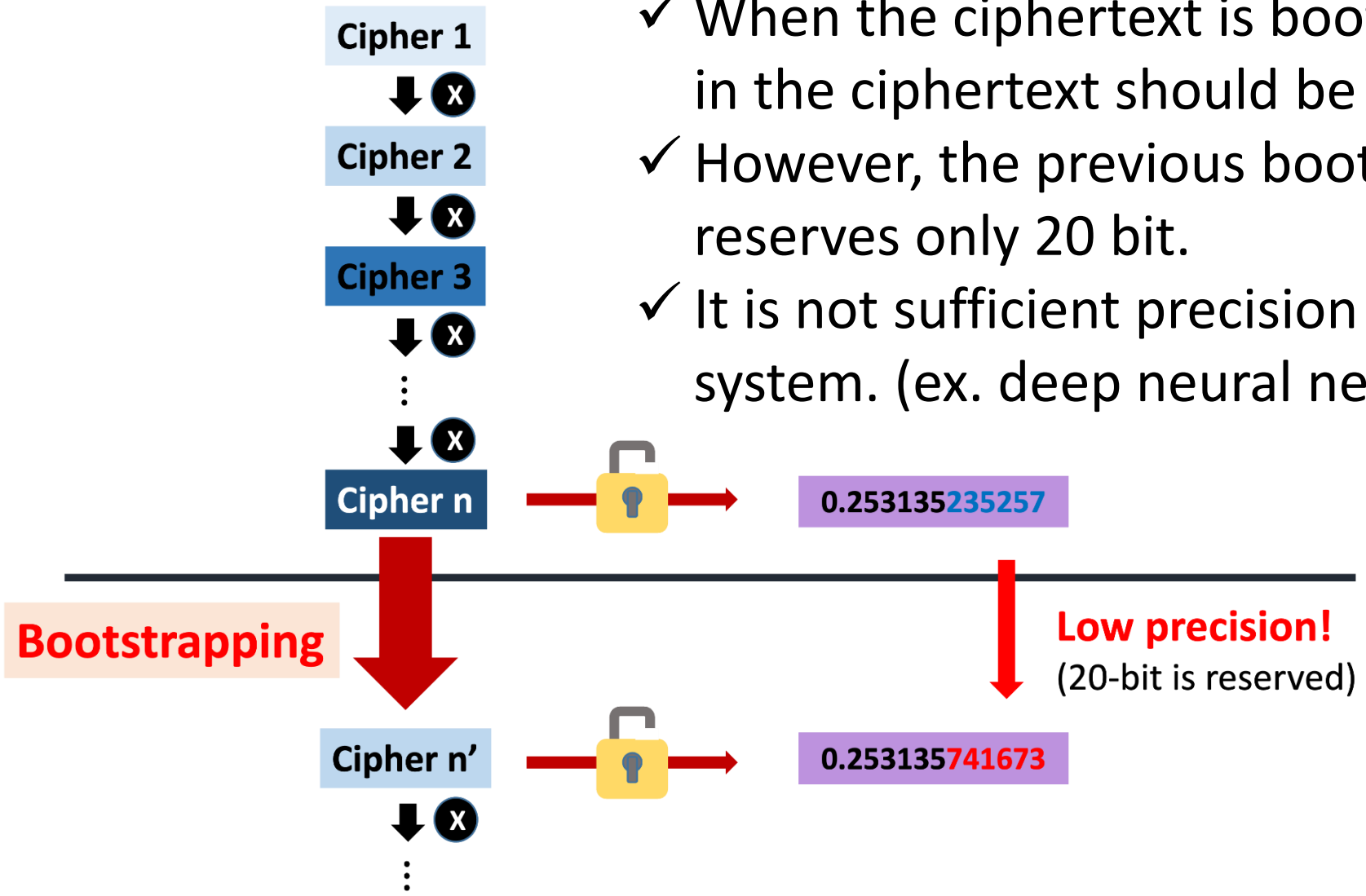


Background: Bootstrapping operation in FHE



- ✓ **Bootstrapping** is an important operation in FHE that allows operations on encrypted data to continue.
- ✓ A harmed ciphertext is refreshed with the bootstrapping so that the further multiplication is possible.

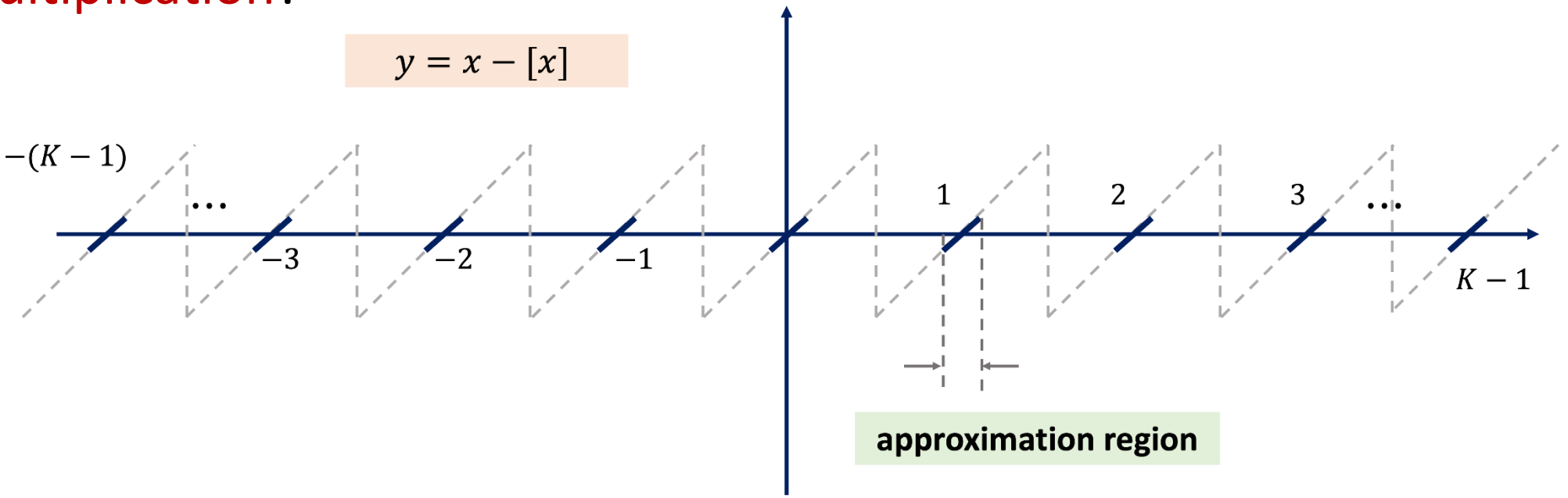
Background: Low Precision in Bootstrapping



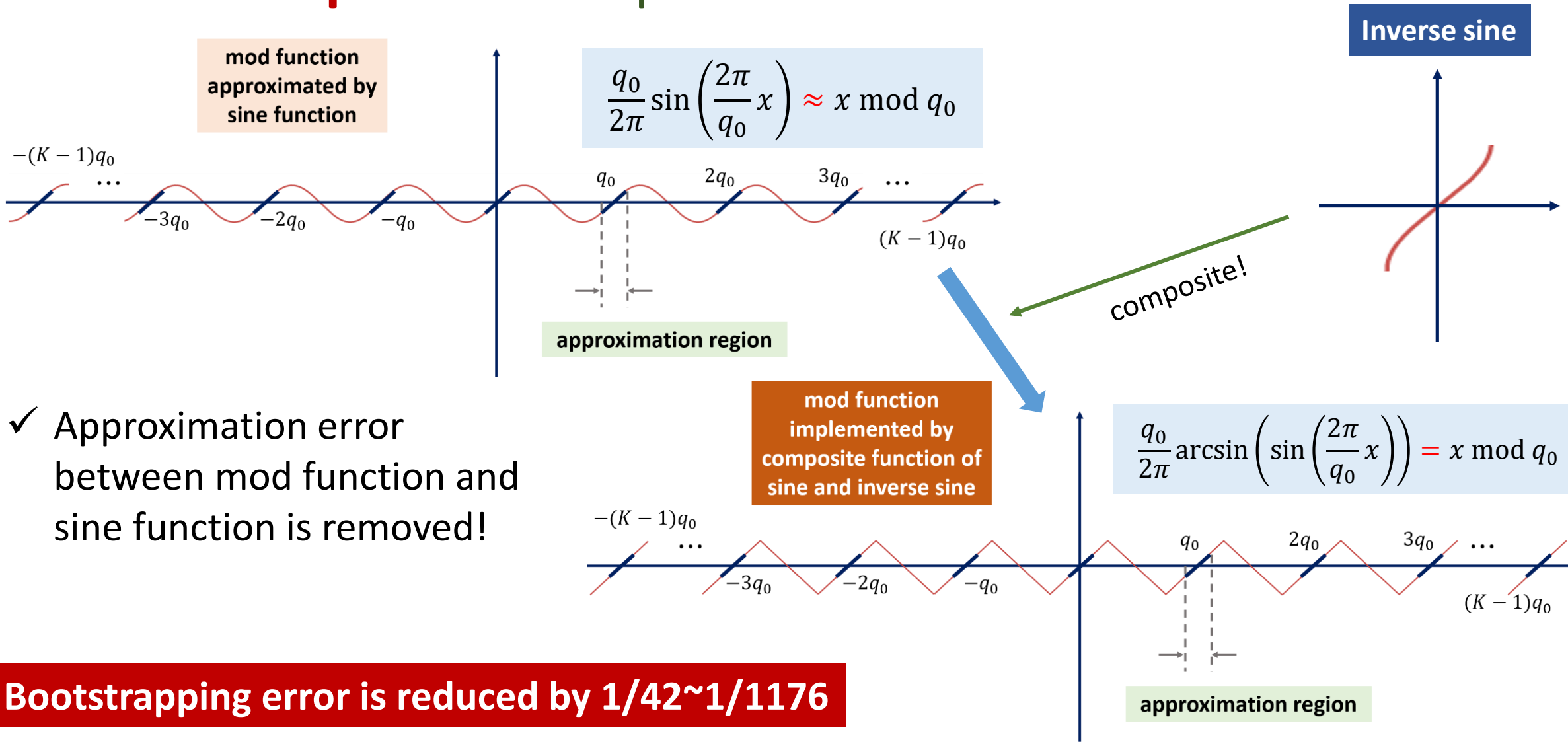
- ✓ When the ciphertext is bootstrapped, the values in the ciphertext should be reserved.
- ✓ However, the previous bootstrapping only reserves only 20 bit.
- ✓ It is not sufficient precision for many complex system. (ex. deep neural network)

Issue: Main Problem in Low-Precision Bootstrapping

- ✓ Precise evaluation of **modular reduction function** on FHE determines the precision of the bootstrapping.
- ✓ How to evaluate the modular reduction function with only addition and multiplication?



Proposed: Composition with Inverse Sine[8]



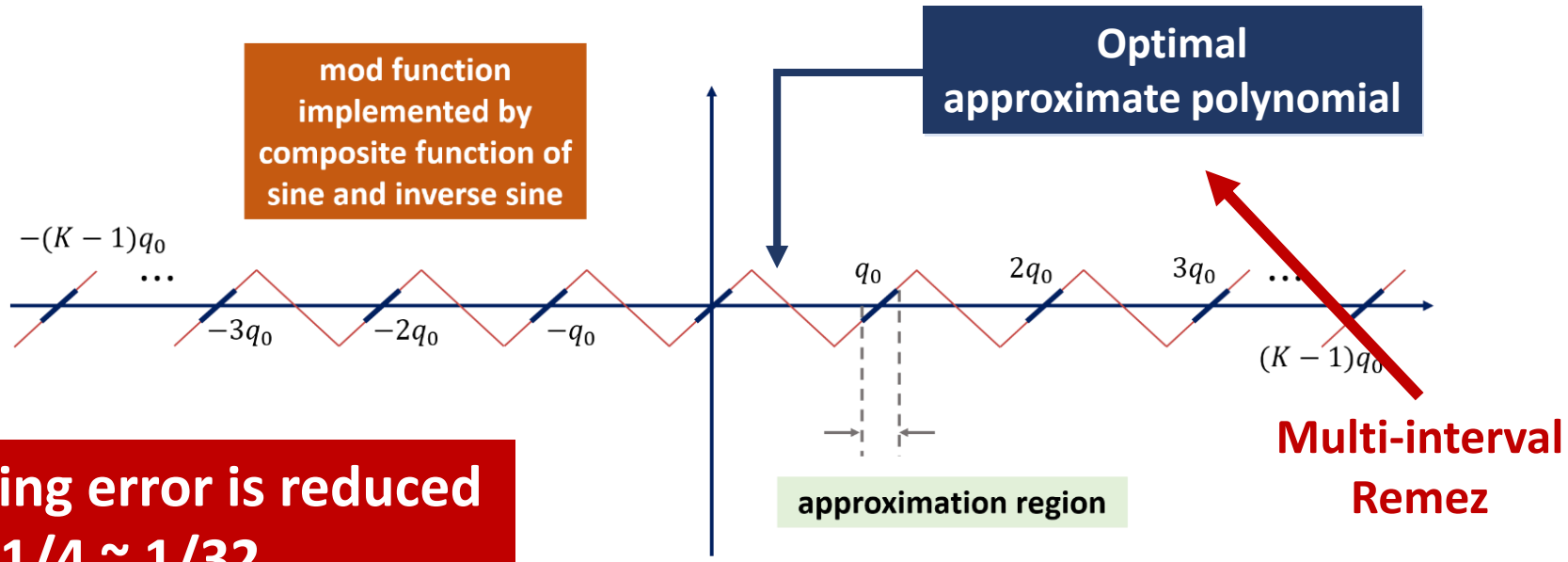
✓ Approximation error between mod function and sine function is removed!

Bootstrapping error is reduced by 1/42~1/1176

[8] Joon-Woo Lee*, Eunsang Lee, Yongwoo Lee, Young-Sik Kim, and Jong-Seon No, "High-precision bootstrapping of RNS-CKKS homomorphic encryption using optimal minimax polynomial approximation and inverse sine function," ***EUROCRYPT 2021***, pp. 618-647, Springer, Cham, 2021 (**Top-tier conference**, acceptance ratio : 19.5%).

Proposed: Optimal Minimax Polynomial Approximation[8]

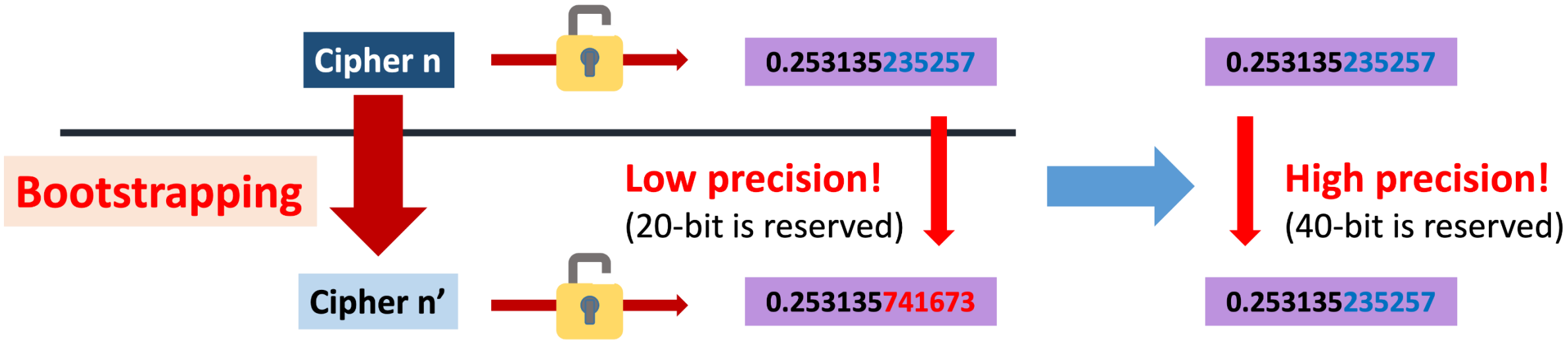
- ✓ **Remez algorithm** : finding the optimal minimax approximate polynomial for one interval.
- ✓ Practical **multi-interval Remez algorithm** is proposed to obtain the optimal polynomial.



Bootstrapping error is reduced by 1/4 ~ 1/32.

[8] Joon-Woo Lee*, Eunsang Lee, Yongwoo Lee, Young-Sik Kim, and Jong-Seon No, "High-precision bootstrapping of RNS-CKKS homomorphic encryption using optimal minimax polynomial approximation and inverse sine function," ***EUROCRYPT 2021***, pp. 618-647, Springer, Cham, 2021 (**Top-tier conference**, acceptance ratio : 19.5%).

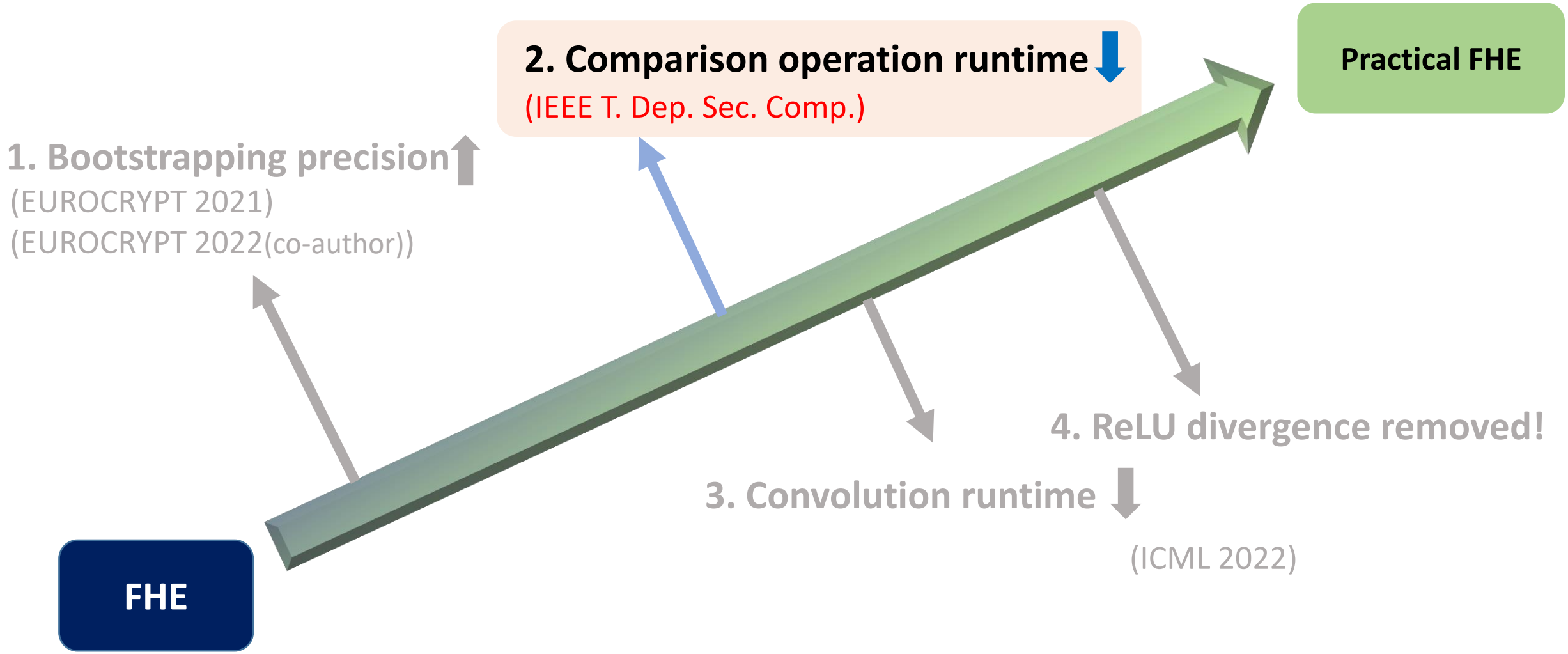
Proposed: High-precision Bootstrapping[8]



✓ **Bootstrapping precision is sufficiently improved!**

[8] Joon-Woo Lee*, Eunsang Lee, Yongwoo Lee, Young-Sik Kim, and Jong-Seon No, "High-precision bootstrapping of RNS-CKKS homomorphic encryption using optimal minimax polynomial approximation and inverse sine function," ***EUROCRYPT 2021***, pp. 618-647, Springer, Cham, 2021 (**Top-tier conference**, acceptance ratio : 19.5%).

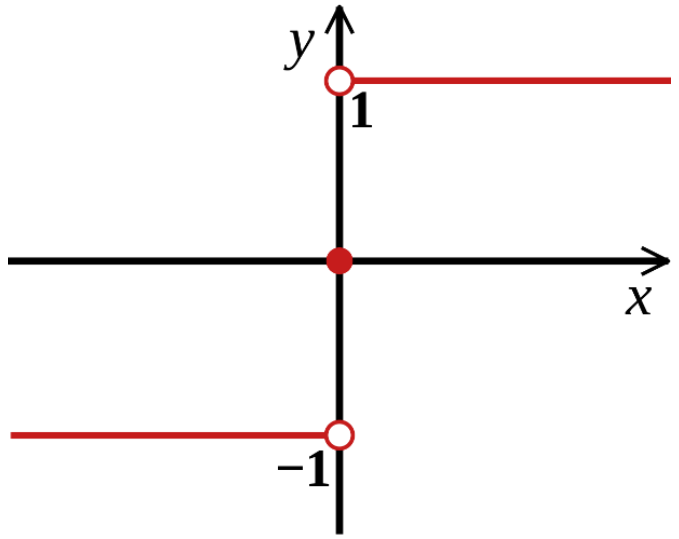
Research Goal: Improvement of FHE



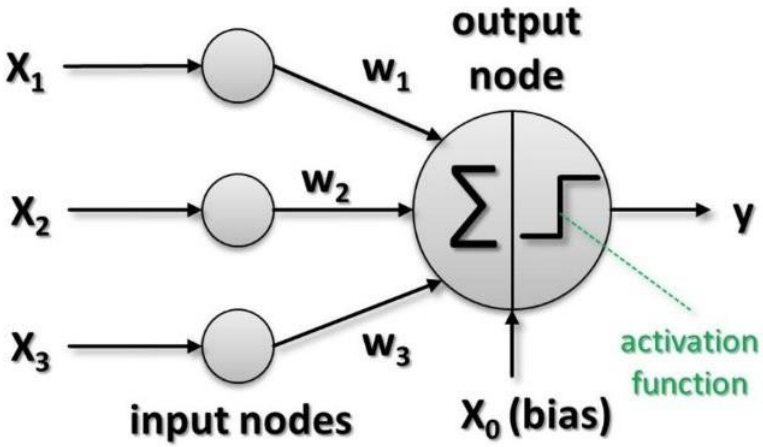
Comparison Operation on FHE

- ✓One of the useful elementary operation is the comparison operation.
  - ✓max function, min function, sign function…
- ✓These operations are based on the sign function.
- ✓How to evaluate the sign function on FHE only with addition and multiplication?

$$\mathrm{sign}(x) = \begin{cases} -1, & x < 0 \\ 1, & x > 0 \end{cases}$$



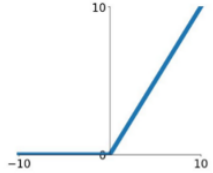
Application: Non-Linear Activation Function on FHE



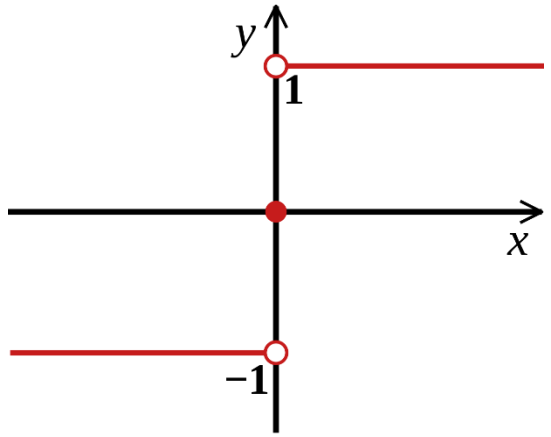
- ✓Non-linear activation function allows the deep learning model to learn **non-linear boundaries.**
- ✓It is required to evaluate **ReLU function** on FHE.
- ✓The elementary function in ReLU function is **the sign function**.

ReLU

$\max(0, x)$

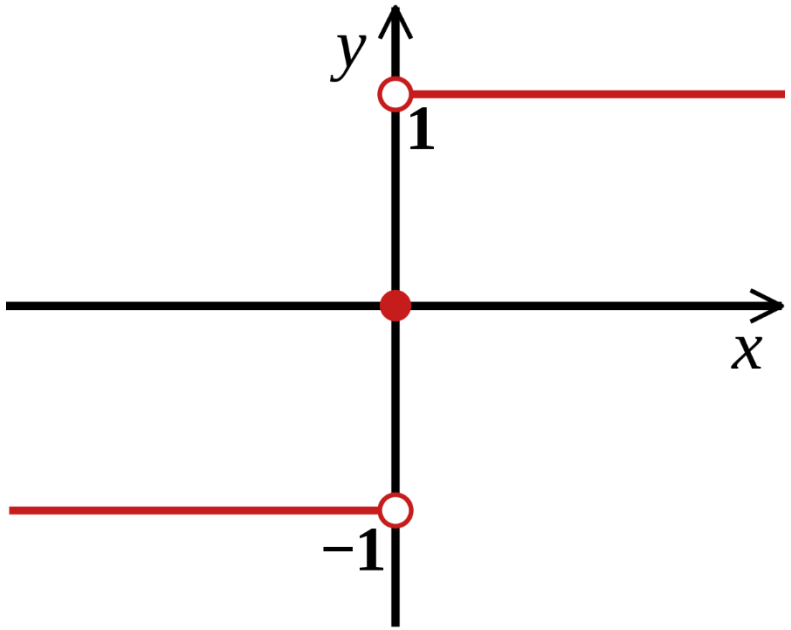


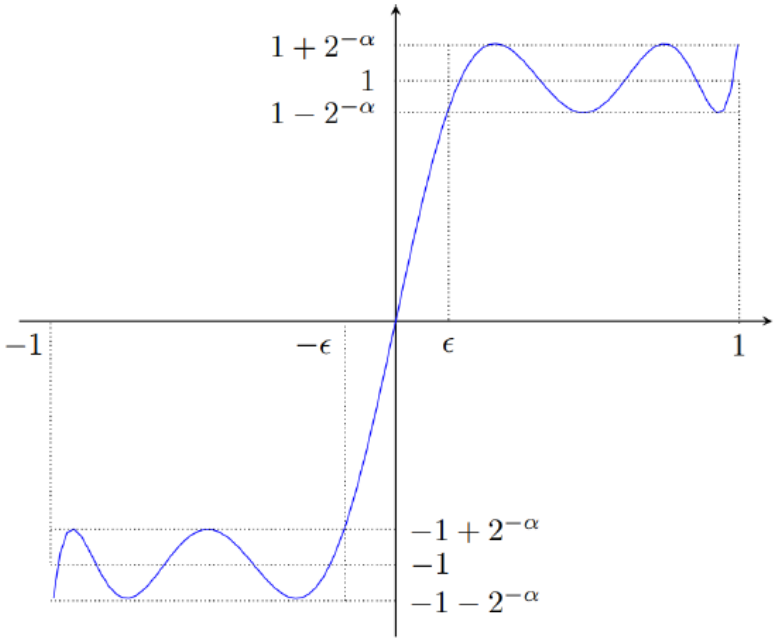
$$\text{sign}(x) = \begin{cases} -1, & x < 0 \\ 1, & x > 0 \end{cases}$$



Issue: Approximation of Sign Function

✓ Trivial solution: minimax approximation

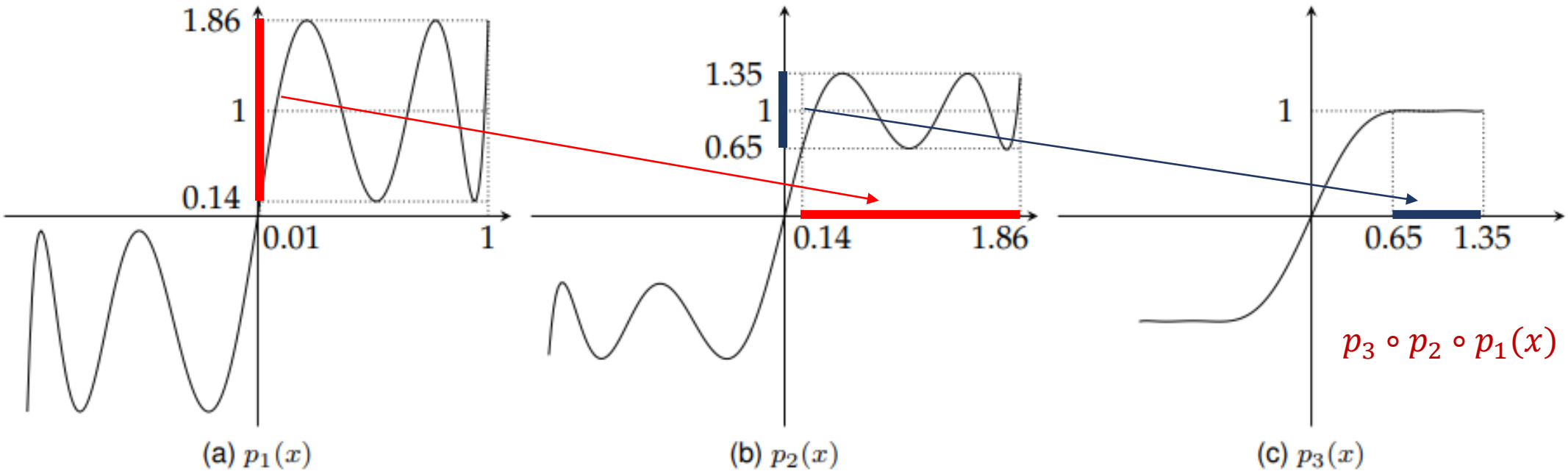




Polynomial with degree **300**

Too large degree of minimax polynomial for large precision!

Proposed: Minimax Composition of Sign Function[2]



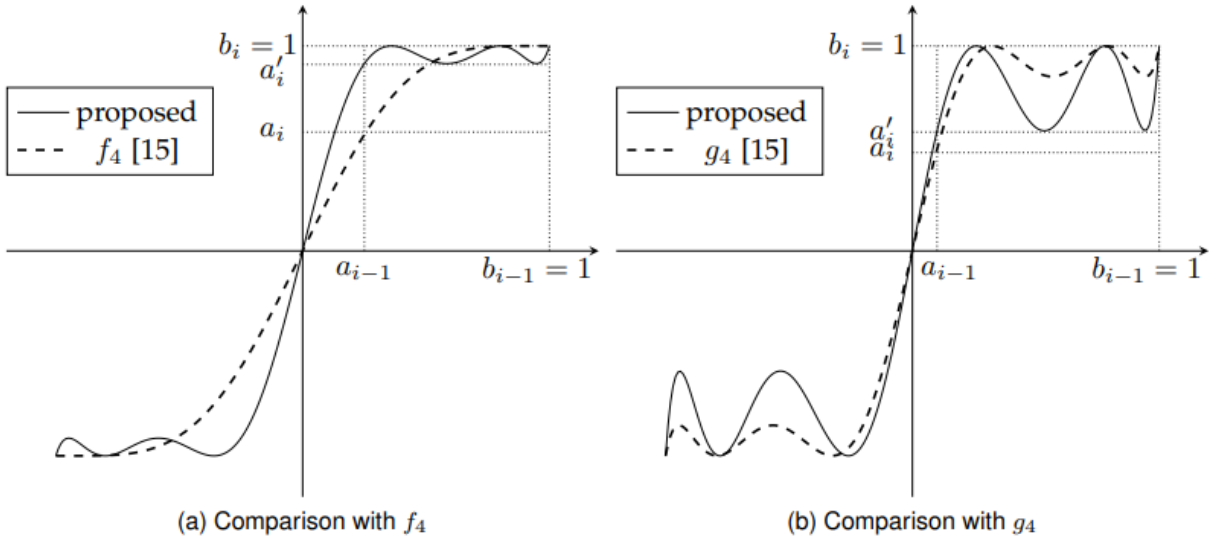
$$p_3 \circ p_2 \circ p_1(x) \approx \text{sign}(x)$$

Three polynomials with degree **7**

Approximation by the composition of several small-degree minimax polynomials

[2] Eunsang Lee, **Joon-Woo Lee***, Young-Sik Kim, and Jong-Seon No, "Minimax approximation of sign function by composite polynomial for homomorphic comparison," ***IEEE Trans actions on Dependable and Secure Computing,*** 19(6), pp. 3711-3727, 2022 (**IF : 6.791,** JCR 2021 **Top 8.6%**).

Proposed: Optimality of Minimax Composition[2]

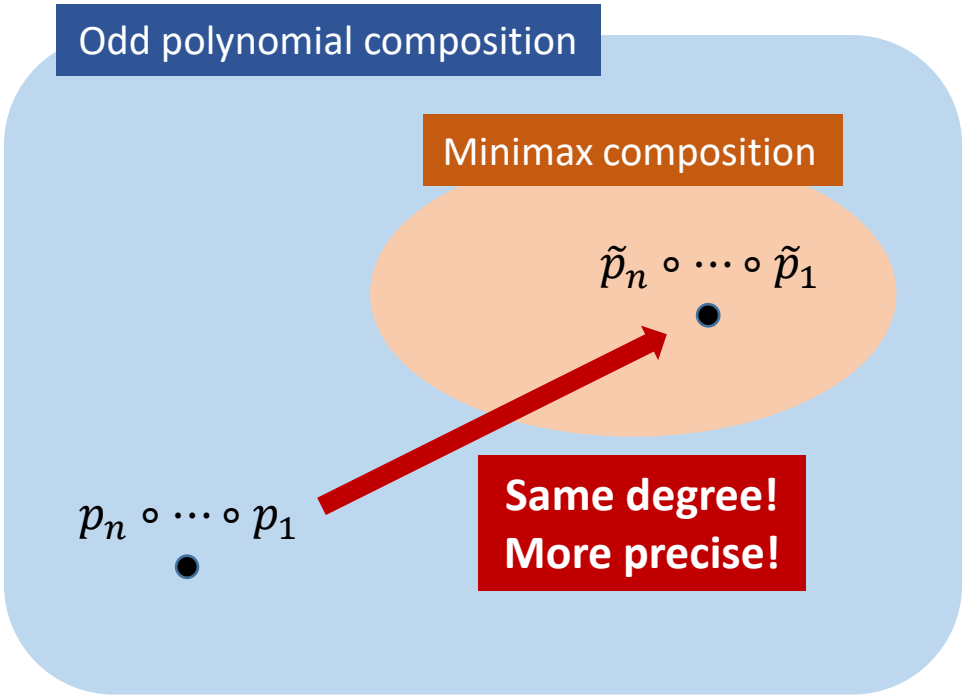


(a) Comparison with f_4 (b) Comparison with g_4

Theorem: Optimality of minimax composition

Given any composition of polynomials approximating the sign function, we can always find a minimax composition of polynomials with the same degrees approximating the sign function better.

Is there any better method using composition of some odd functions?
→ No!



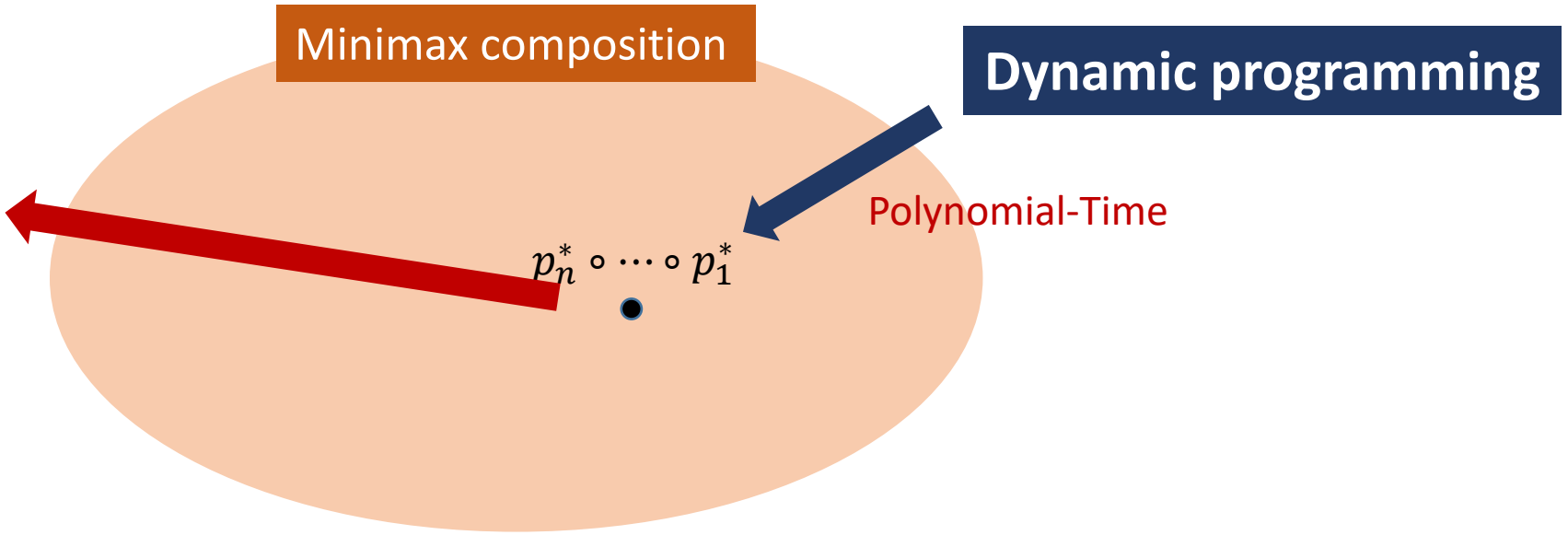
[2] Eunsang Lee, **Joon-Woo Lee***, Young-Sik Kim, and Jong-Seon No, "Minimax approximation of sign function by composite polynomial for homomorphic comparison," ***IEEE Transactions on Dependable and Secure Computing,*** 19(6), pp. 3711-3727, 2022 (**IF : 6.791,** JCR 2021 **Top 8.6%**).

Proposed: Dynamic Programming for Minimal Operations[2]

How to obtain the minimax composition with minimal operation?

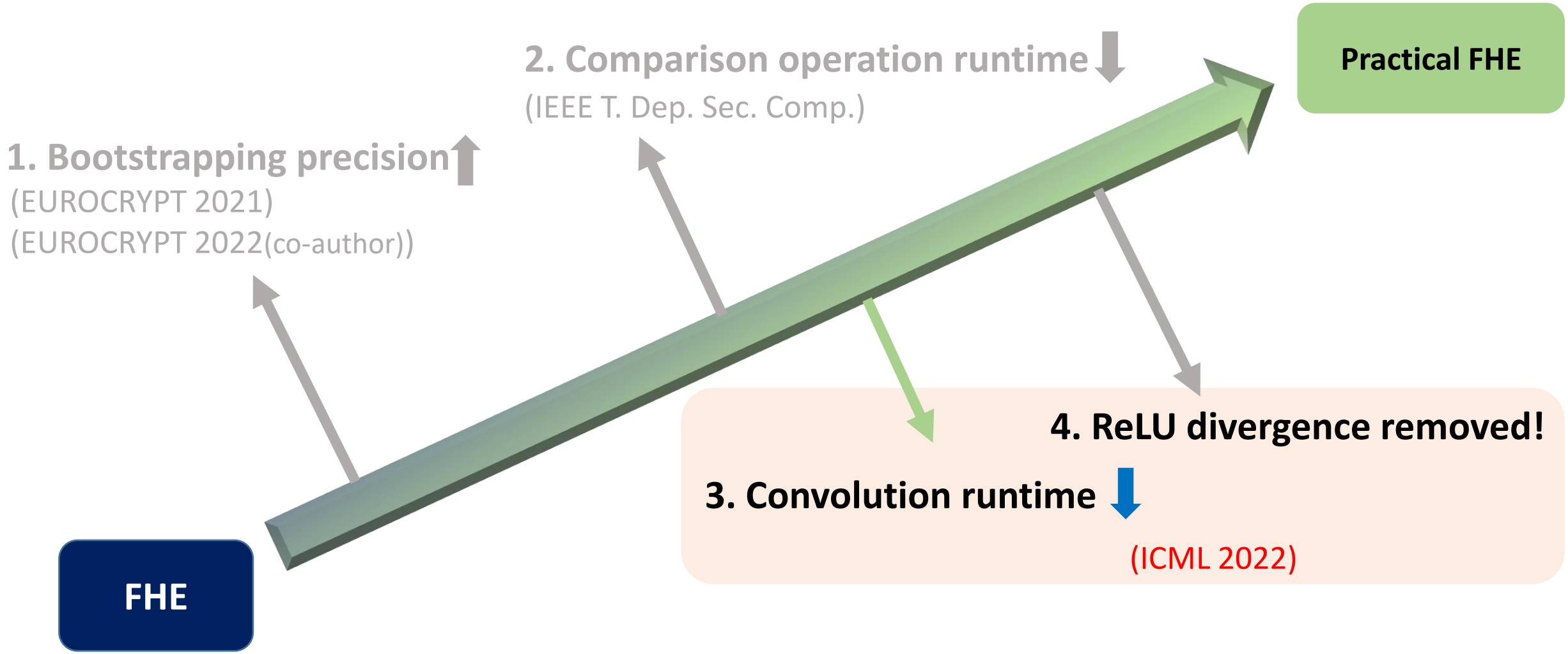
Optimal composition!

45% runtime reduced!

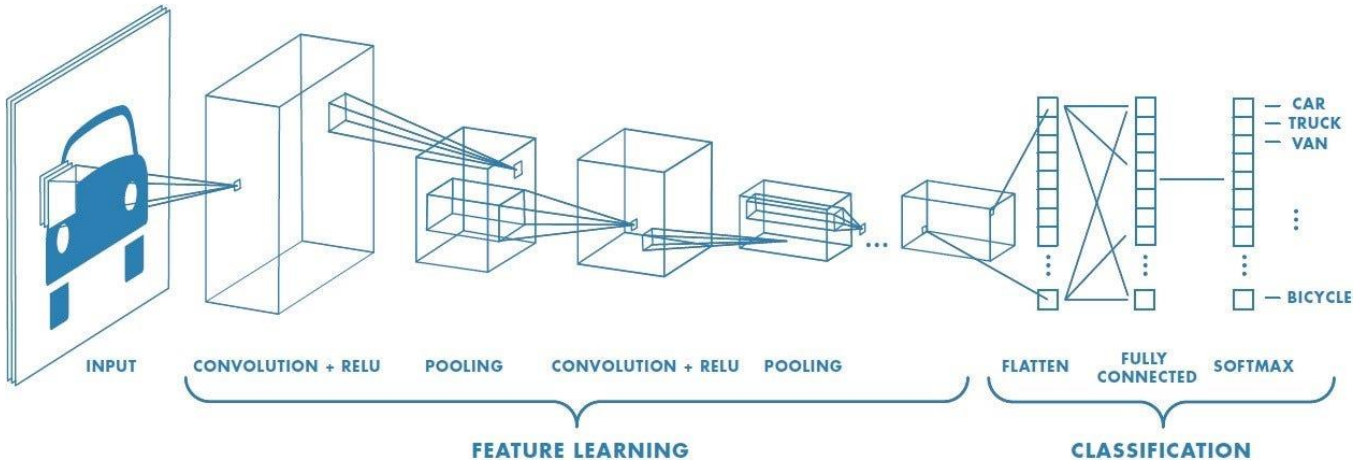


[2] Eunsang Lee, Joon-Woo Lee*, Young-Sik Kim, and Jong-Seon No, "Minimax approximation of sign function by composite polynomial for homomorphic comparison," ***IEEE Transactions on Dependable and Secure Computing,*** 19(6), pp. 3711-3727, 2022 (**IF : 6.791,** JCR 2021 **Top 8.6%**).

Research Goal: Improvement of FHE



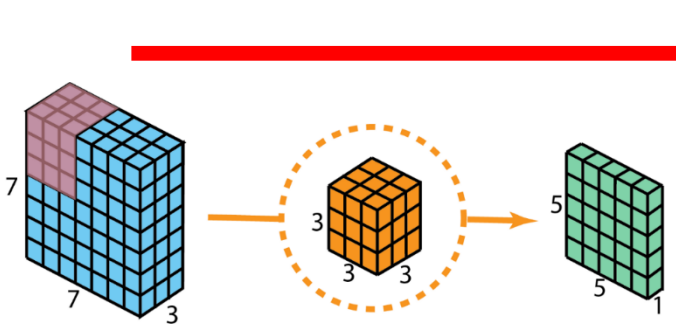
Background: Implementation of CNN Model on FHE



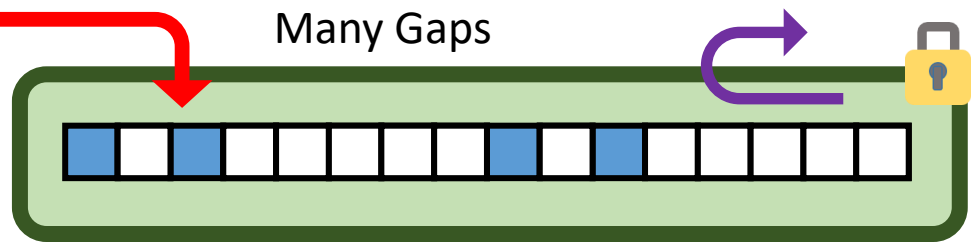
- ✓We implemented the ResNet CNN model with 20 layers on FHE **for the first time**![4]
- ✓However, there are two problems in CNN on FHE.
  1. Too long computation time with many computational resources.
  2. The deeper CNN model on FHE has low classification accuracy.

[4] **Joon-Woo Lee**(co-first with Hyungchul Kang), Hyungchul Kang, Yongwoo Lee, Wooseok Choi, Jieun Eom, Maxim Deryabin, Eunsang Lee, Junghyun Lee, Donghoon Yoo, Young-Sik Kim*, and Jong-Seon No, "Privacy-preserving machine learning with fully homomorphic encryption for deep neural network," ***IEEE Access***, 10, pp. 30039-30054, 2022. **(Google Scholar Citation: 85)**

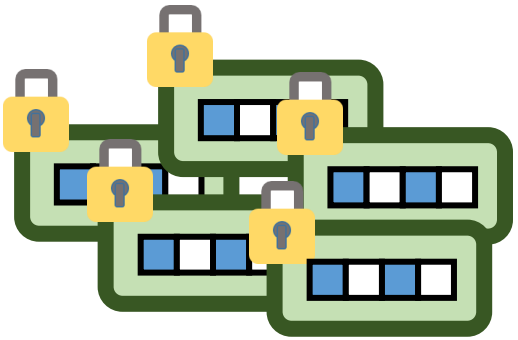
Issue: Convolution on 1D FHE



Convolution with 3D Tensor



1D vector in FHE scheme
and only cyclic shift

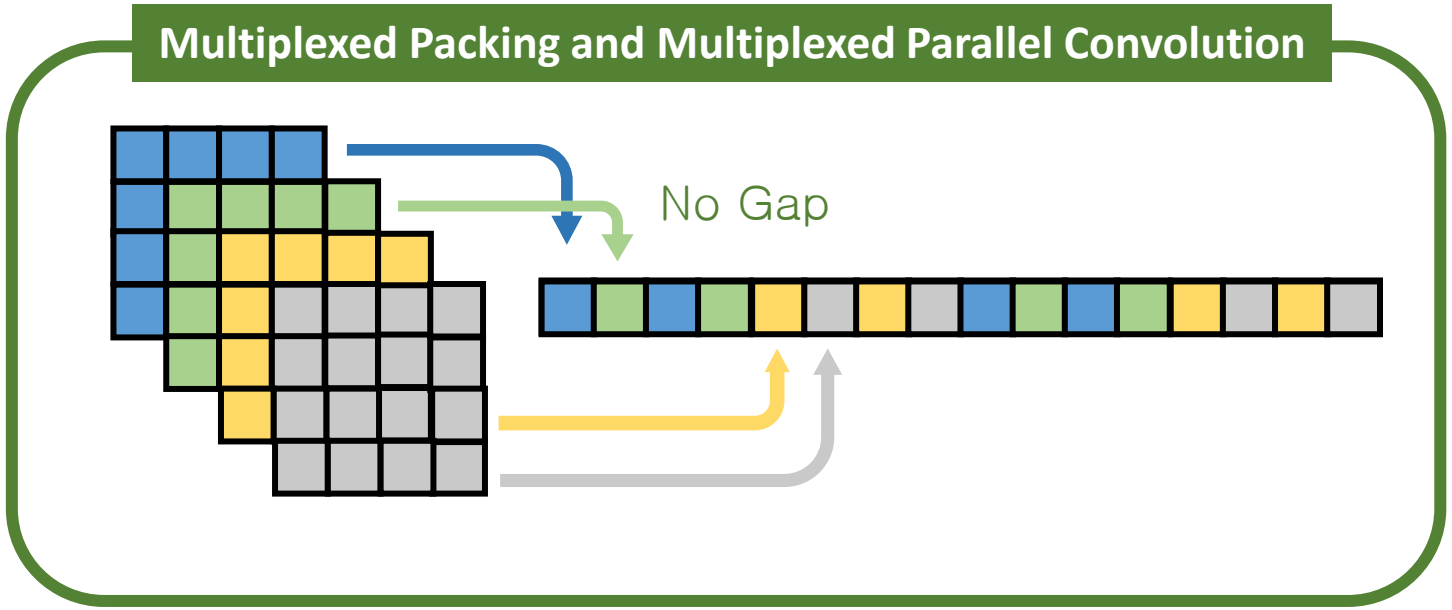


Many intermediate ciphertexts

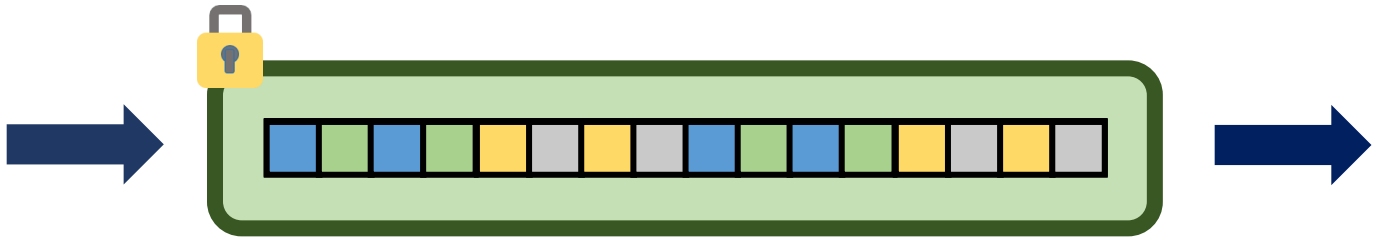
- ✓CKKS scheme only supports **1-dim data structure** and **cyclic shift** data movement.
- ✓Trivial technique should make **gaps** in the ciphertext.
- ✓It causes many intermediate ciphertext to be bootstrapped.

→ **3 hours per image with 64 threads**
Too much runtime and resources!

Proposed: Multiplexed Parallel Convolution[3]



- ✓**Systematic packing method** of 3D tensors in 1D vector in FHE.
- ✓Compatible with **convolution using any strides**.



Only one intermediate ciphertext

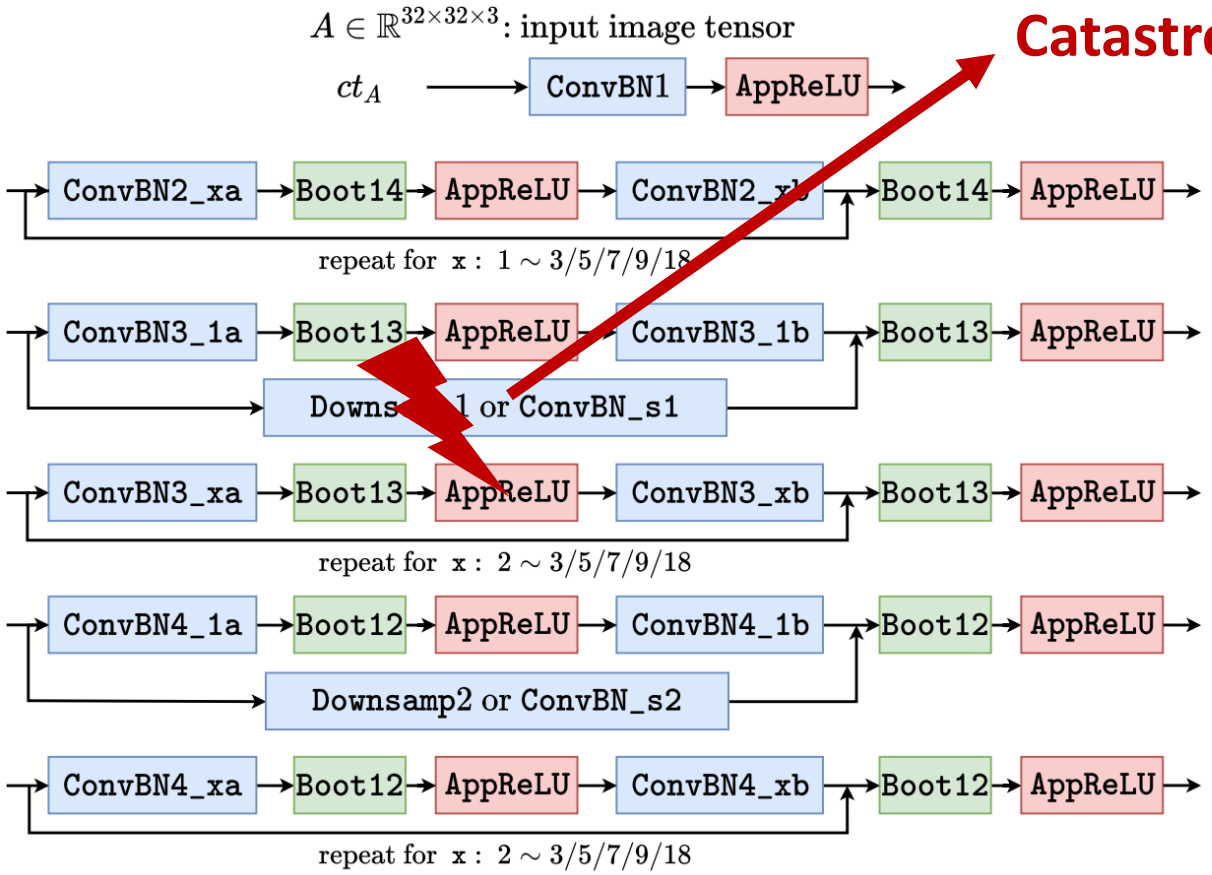
3 hours per image with 64 threads

# of operations is reduced by 1/134

40 minutes per image with single thread

[3] Eunsang Lee, **Joon-Woo Lee***, Junghyun Lee, Yongjune Kim, Young-Sik Kim, Jong-Seon No, and Wooseok Choi, "Low-complexity deep convolutional neural networks on fully ho momorphic encryption using multiplexed parallel convolutions," ***ICML 2022***, pp. 12403-12422, 2022 (**Top-tier conference**, acceptance ratio : 21.9%).

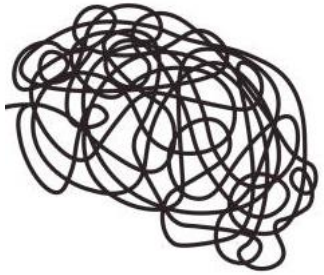
Observed: Divergence Problem in Deep CNN on FHE[3]



Structure of ResNet CNN model on FHE

Catastrophic divergence in ReLU

✓When performing deep neural network on FHE, catastrophic divergence phenomenon occurs at least one layer with probability of 25%.

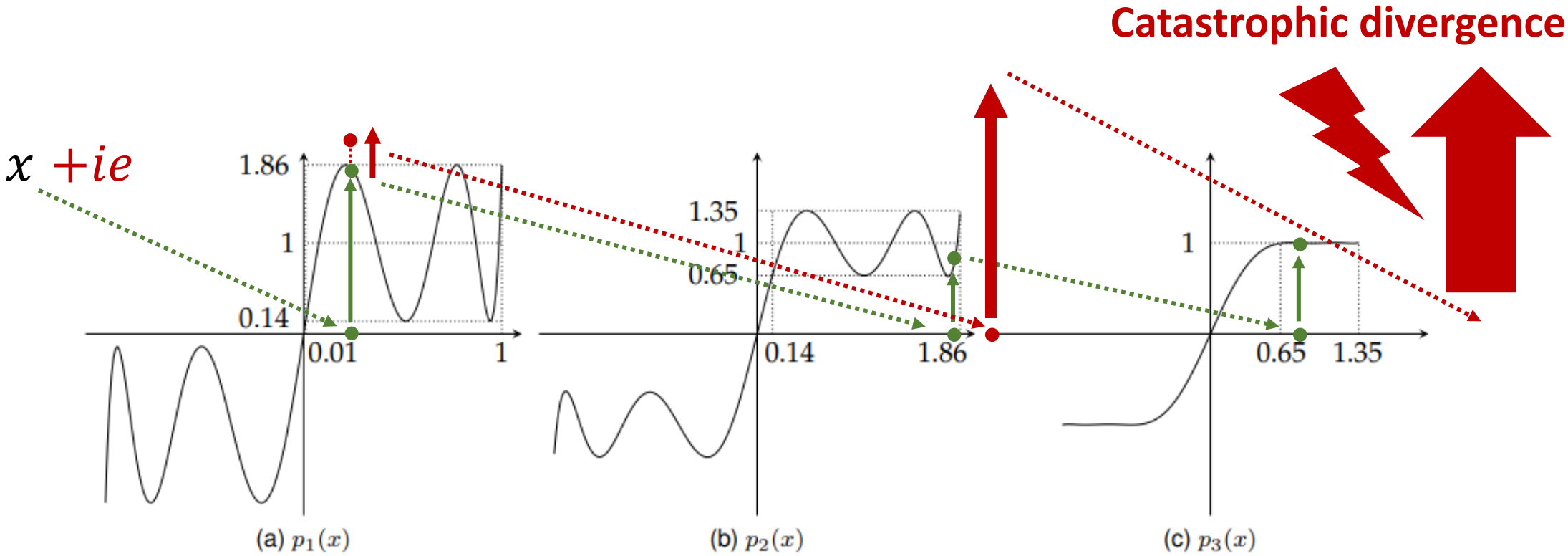


Fatal error!

[3] Eunsang Lee, **Joon-Woo Lee***, Junghyun Lee, Yongjune Kim, Young-Sik Kim, Jong-Seon No, and Wooseok Choi, "Low-complexity deep convolutional neural networks on fully ho momorphic encryption using multiplexed parallel convolutions," ***ICML 2022***, pp. 12403-12422, 2022 (**Top-tier conference**, acceptance ratio : 21.9%).

Analyzed: Effect of Imaginary Error on Divergence Problem[3]

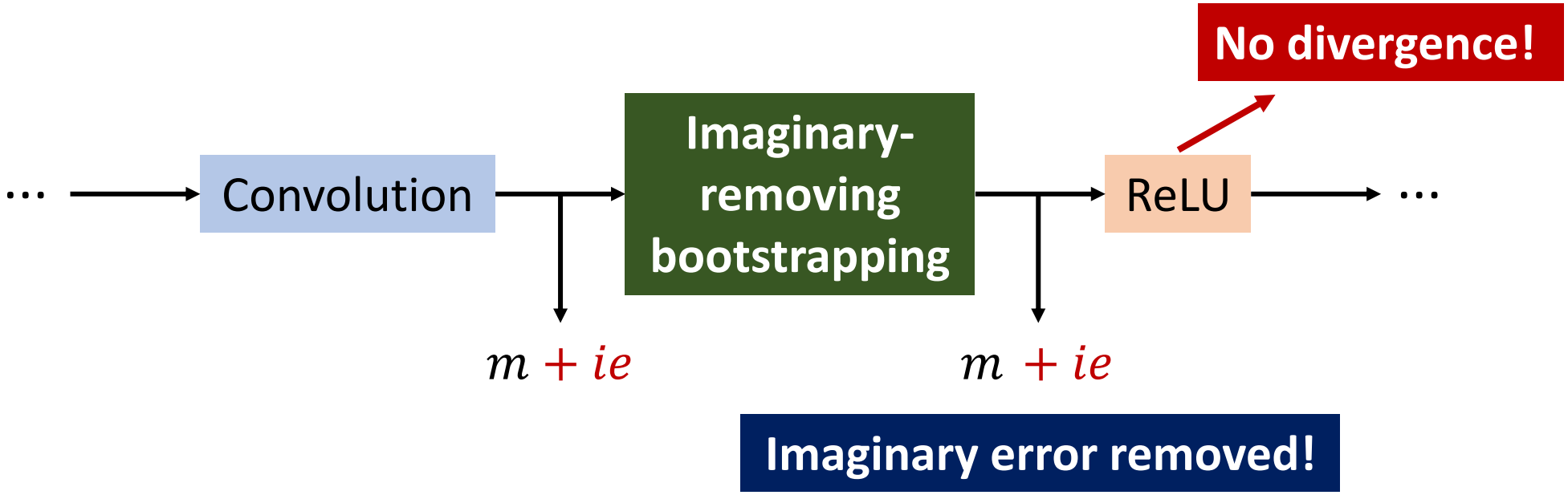
✓Small imaginary error can occur divergence problem in ReLU.



[3] Eunsang Lee, **Joon-Woo Lee***, Junghyun Lee, Yongjune Kim, Young-Sik Kim, Jong-Seon No, and Wooseok Choi, "Low-complexity deep convolutional neural networks on fully ho momorphic encryption using multiplexed parallel convolutions," ***ICML 2022***, pp. 12403-12422, 2022 (**Top-tier conference**, acceptance ratio : 21.9%).

Proposed: Imaginary-Removing Bootstrapping[3]

✓We propose imaginary-removing bootstrapping to remove imaginary error before ReLU.



[3] Eunsang Lee, **Joon-Woo Lee***, Junghyun Lee, Yongjune Kim, Young-Sik Kim, Jong-Seon No, and Wooseok Choi, "Low-complexity deep convolutional neural networks on fully ho momorphic encryption using multiplexed parallel convolutions," ***ICML 2022***, pp. 12403-12422, 2022 (**Top-tier conference**, acceptance ratio : 21.9%).

Implemented: Privacy-Preserving Deep ResNet on FHE[3]

Depth of ResNet on FHE : **20 layers**
3 hours with **64 threads** for 20 layers

1. **Multiplexed Parallel Convolution**
2. **Imaginary-Removing Bootstrapping**

Depth of ResNet on FHE : **20~110 layers**
40 minutes with **1 thread** for 20 layers

We implemented deep CNN model on FHE *for the first time*!

[3] Eunsang Lee, **Joon-Woo Lee***, Junghyun Lee, Yongjune Kim, Young-Sik Kim, Jong-Seon No, and Wooseok Choi, "Low-complexity deep convolutional neural networks on fully ho momorphic encryption using multiplexed parallel convolutions," ***ICML 2022***, pp. 12403-12422, 2022 (**Top-tier conference**, acceptance ratio : 21.9%).

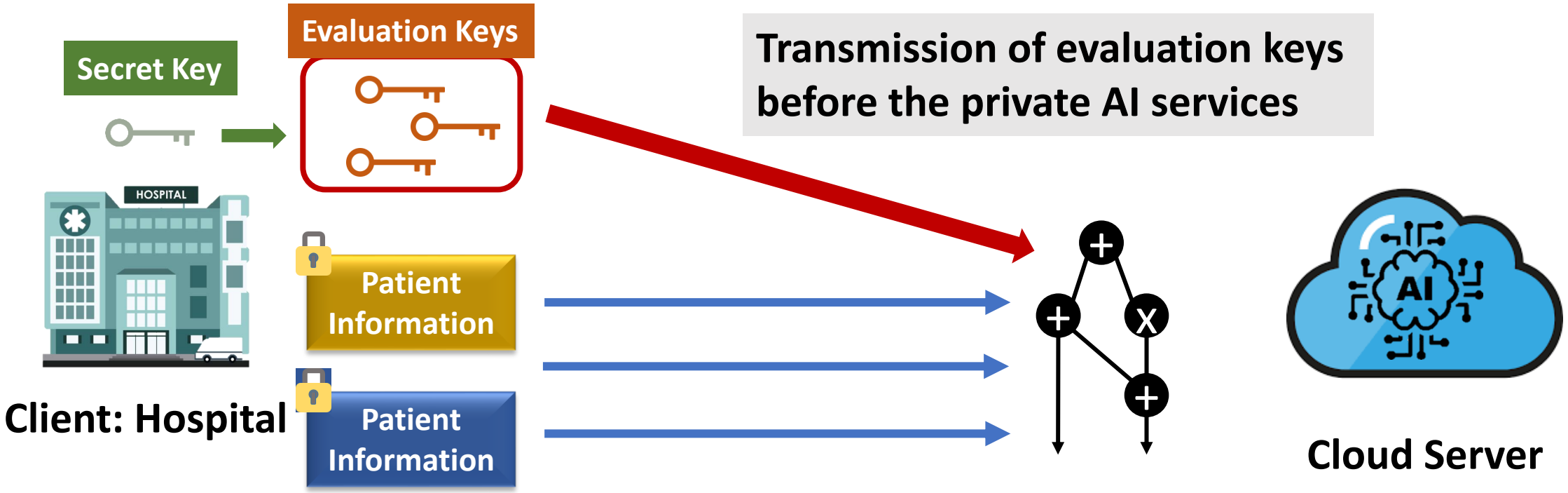
Contents

✓ Introduction

✓ Recent Research Results

✓ Future Research Plans

Problem: Evaluation Keys in FHE

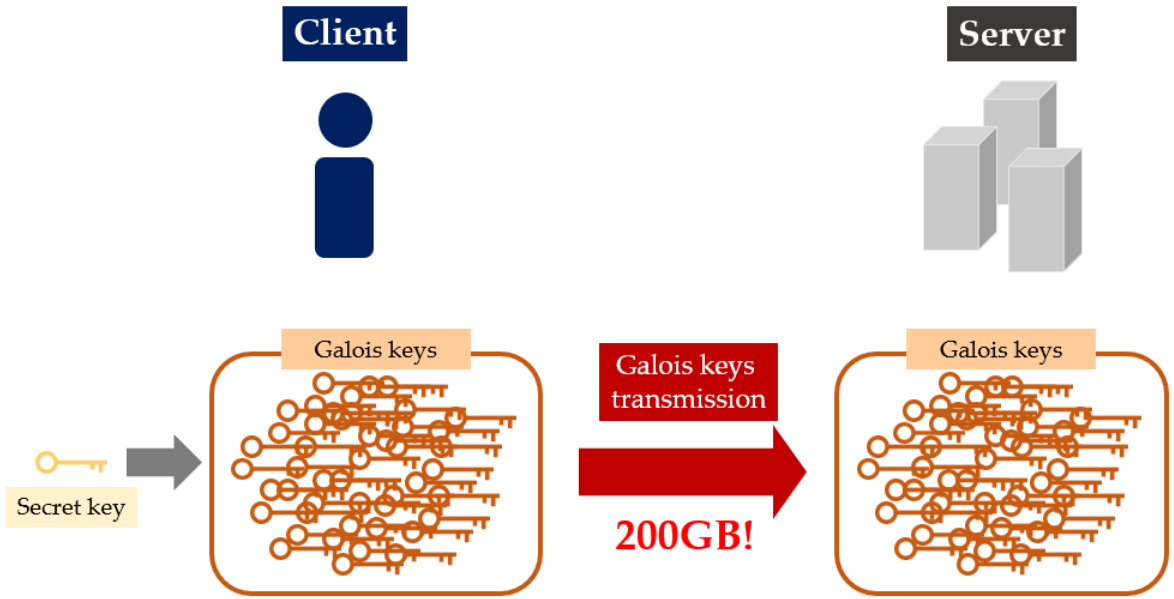


✓Evaluation keys are required for the server to perform operations on encrypted data.

✓In the previous scheme, these keys are generated from the secret key in the client side.

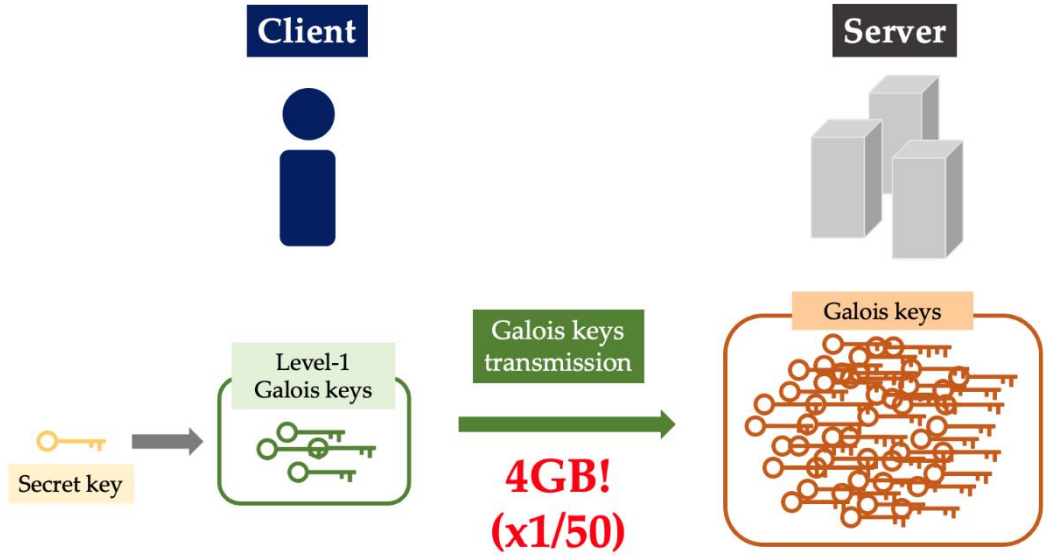
✓These should be transmitted from the client to the server before the services.

Problem: Communication Costs for Evaluation Keys



- ✓Many types of operations is required to perform the PPML model.
  - ✓Rotation operation with different cyclic shift.
- ✓Very large number of the evaluation keys (617 rotation keys)
- ✓The amount of communication dramatically increases!
  - ✓ResNet-20 for CIFAR-10: 265 rotation keys and 105.6GB transmission
  - ✓ResNet-18 for ImageNet: 617 rotation keys and 197.6GB transmission

Solution: Key Idea for Solution of Evaluation Key Issue[12]

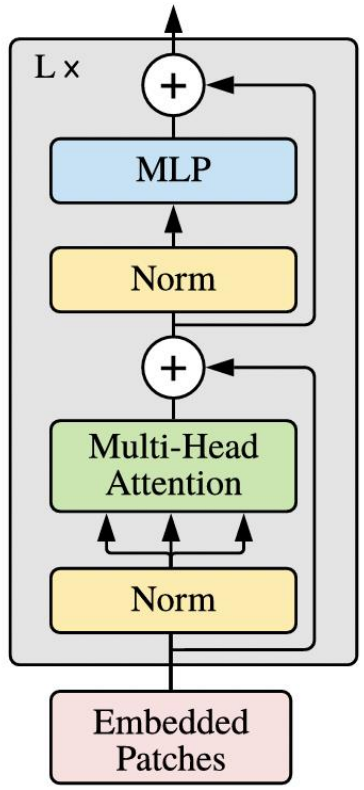


- What if all rotation keys can be generated by the server itself?
  - The client creates a small number of so-called “master rotation keys”.
  - The server can then generate all required rotation keys for each service by using the “master rotation keys.”
- 200GB key transmission is reduced to **3.9GB or lower.**

[12] **Joon-Woo Lee**, Eunsang Lee*, Young-Sik Kim*, Jong-Seon No, “Rotation Key Reduction for Client-Server Systems of Deep Neural Network on Fully Homomorphic Encryption,” submitted to ***ASIACRYPT 2023*** (**Top-tier conference**).

Future work: Transformer on FHE

- ✓Transformer is replacing the CNN model in various AI area.
- ✓Since the transformer network uses different operation block from CNN model, the novel technique on FHE for this new operation blocks is required.
  - ✓Transposition and matrix multiplication
  - ✓High-precision softmax function
  - ✓Minimax-composition GeLU function
  - ✓Layer normalization



$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

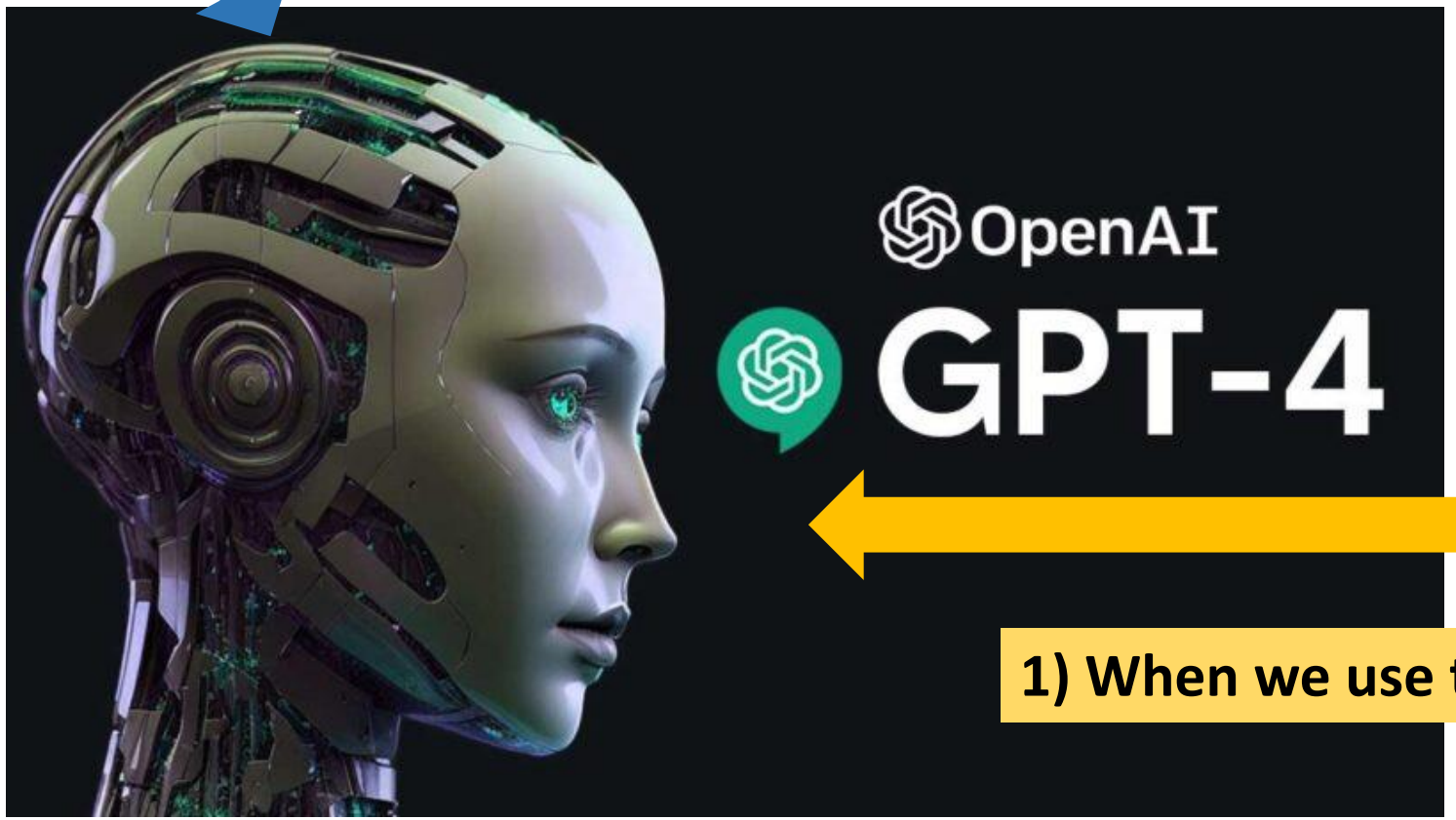
$$GeLU(x) = x\Phi(x) \approx 0.5x(1+\tanh(0.798x + 0.0357x^3))$$

$$LayerNorm(\{x_i\}_i) = \left\{\gamma \cdot \frac{x_i - \bar{x}}{\sqrt{Var(\{x_i\})}} + \beta\right\}_i$$

Two Types of Privacy Issue

Private dataset

2) When the company makes the better AI model

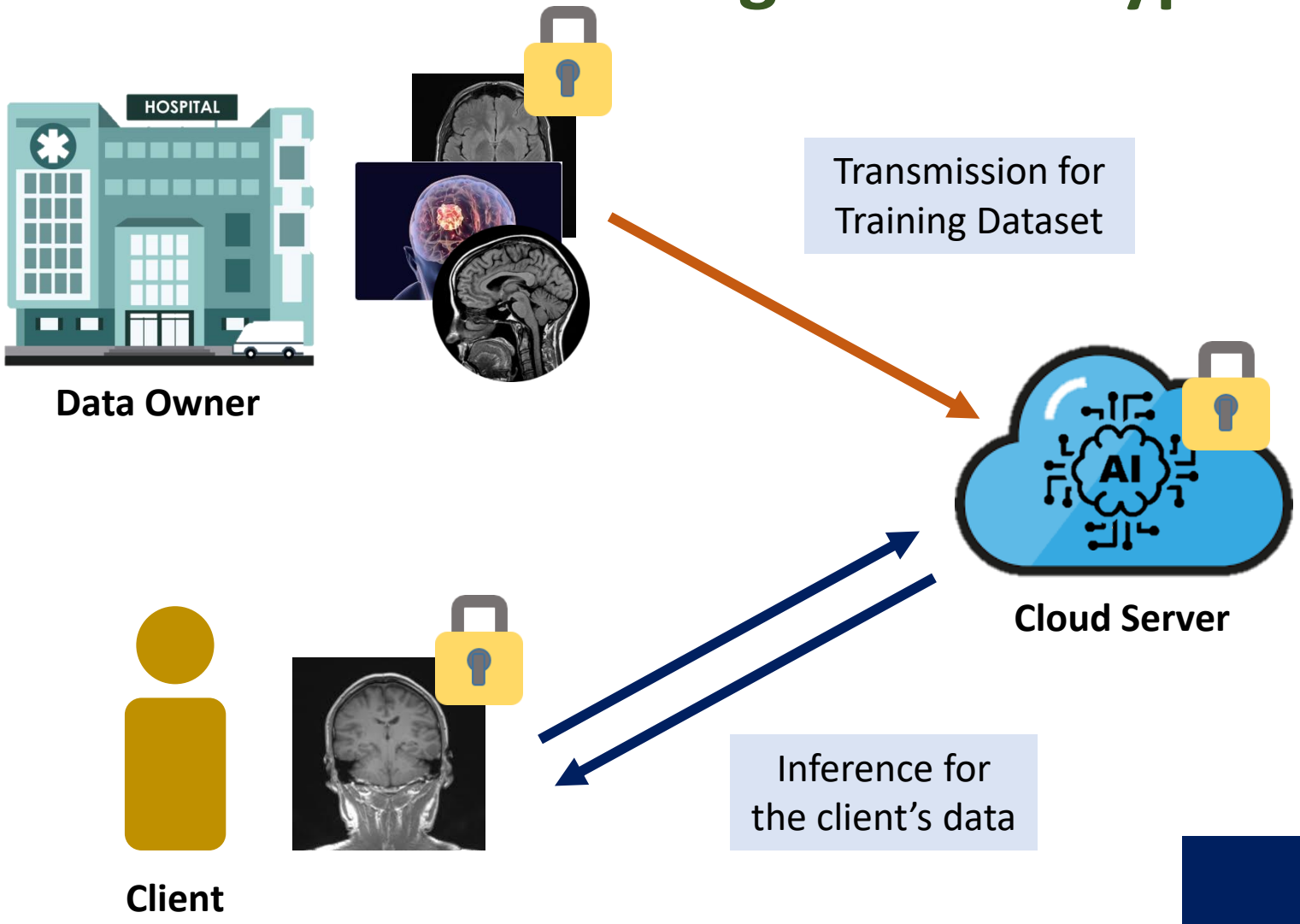


There are two types of privacy in private AI.

Private data

1) When we use the service with our private data

Training with Encrypted Data



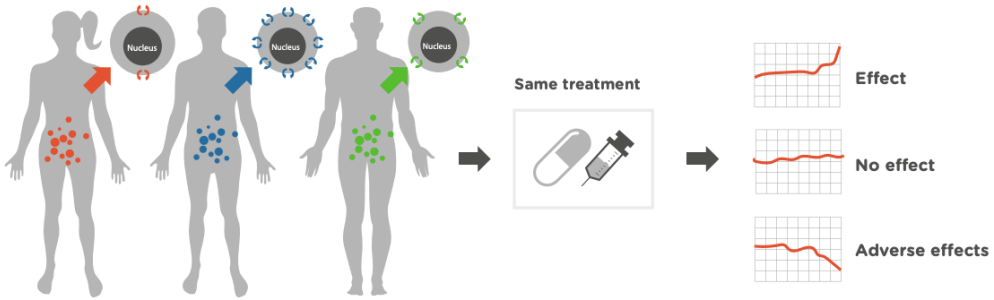
- Data owner can offer massive training dataset to cloud server to train the ML model.
- However, the data for training can be private and sensitive.
- If the privacy issue is not solved, advanced DNN cannot be trained.

Training of PPML with encrypted data is needed!

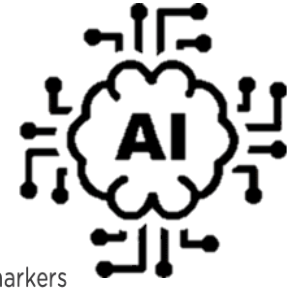
Application: Personalized Medicine

TRADITIONAL MEDICINE: SAME TREATMENT FOR ALL

Cancer patients with e.g. colon cancer receive the same therapy even though they have different biomarkers

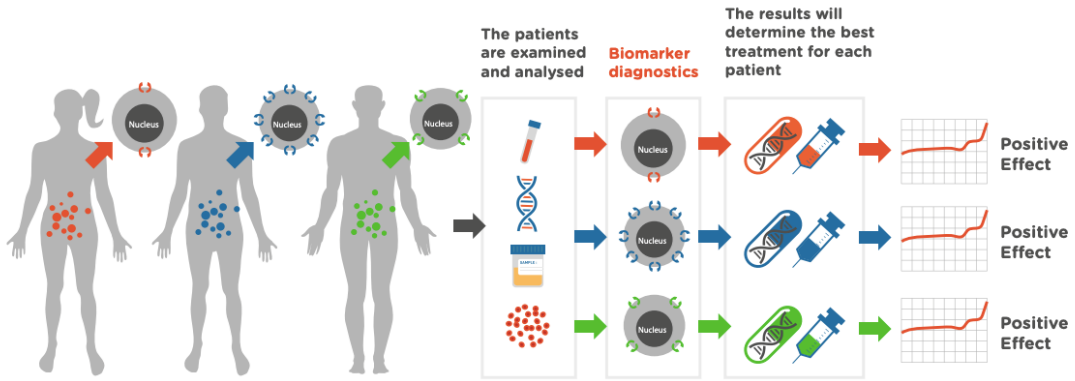


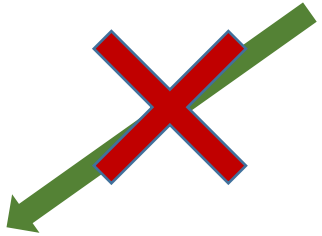
Personalized medicine with personal health information



INNOVATIVE MEDICINE: PERSONALISED MEDICINE

Cancer patients with e.g. colon cancer receive a personalised therapy based on their biomarkers





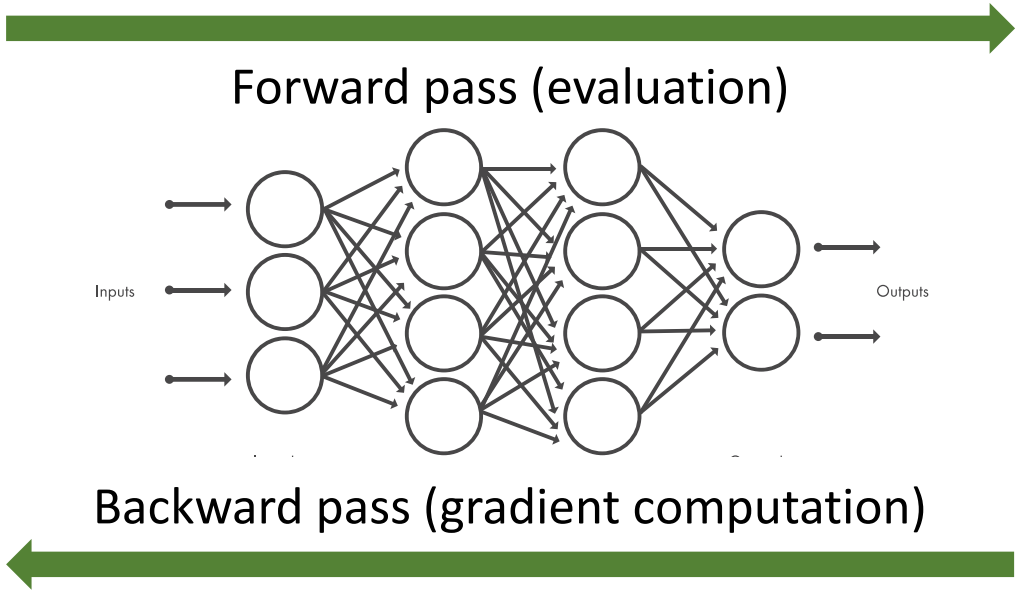
HIRA
HEALTH INSURANCE REVIEW & ASSESSMENT SERVICE

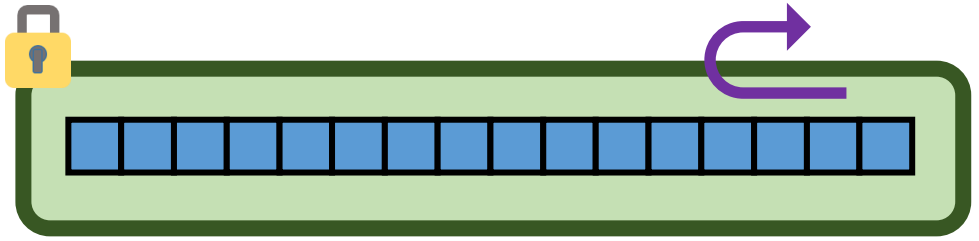
Patient Information Dataset

But we cannot use the dataset for training with privacy issue

- With training of the dataset on FHE, the **personalized medicine** can be possible!

Problem: Packing and Forward/Backward Pass





1D vector in FHE scheme
and only cyclic shift

- In training process, there are both forward pass and backward pass.
- When using FHE, the one-dim vector structure and cyclic shift data movement is quite a huge limitation.
- Fully streamlining this procedure with FHE will be a breakthrough for training of DNN with FHE.

Thank you!