Syfer: Neural Obfuscation for Private Data Release

Central challenges for Clinical AI

- **Data sharing** is a major obstacle to Clinical AI

reproducibility | rare diseases | diversity

- Key tension protecting patient privacy v.s. advancing care
- Need tools to enable **secure and privacy preserving ML**

HIPAA's standard of de-identification

- HIPAA establishes the standard to protect individuals' medical records (PHI)

- HIPAA defines two methods for *de-identification* of PHI:

1. Removing specific identifiers

or

2. Using statistical tools to render information not individually identifiable

- Names
- Geographic subdivisions smaller than a state
- All elements of dates (except year) related to an individual (including admission and discharge dates, birthdate, date of death, all ages over 89 years old, and elements of dates (including year) that are indicative of age)
- Telephone, cellphone, and fax numbers
- Email addresses
- IP addresses
- Social Security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Device identifiers and serial numbers
- Certificate/license numbers
- Account numbers

Existing approaches are not enough

- **Homomorphic encryption**
    - Requires building with crypto primitives. 100-1000x overhead
    - Too cumbersome for training modern DL models

SecureML: A System for Scalable Privacy-Preserving Machine Learning

Payman Mohassel* and Yupeng Zhang†
*Visa Research, †University of Maryland

Oblivious Neural Network Predictions via MiniONN transformations

Jian Liu
Aalto University
jian.liu@aalto.fi

Mika Juuti
Aalto University
mika.juuti@aalto.fi

Yao Lu
Aalto University
yao.lu@aalto.fi

N. Asokan
Aalto University
asokan@acm.org

Existing approaches are not enough

- **Homomorphic encryption**
    - Too cumbersome for training modern DL models
- **Differential Privacy**
    - Private at the cost of a large utility loss, especially for healthcare applications

Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings

Vinith M. Suriyakumar, Nicolas Papernot, Anna Goldenberg, Marzyeh Ghassemi
vinith@cs.toronto.edu
University of Toronto, Vector Institute

Differential Privacy Has Disparate Impact on Model Accuracy

Eugene Bagdasaryan Cornell Tech eugene@cs.cornell.edu

Omid Poursaeed* Cornell Tech op63@cornell.edu

Vitaly Shmatikov Cornell Tech shmat@cs.cornell.edu

Existing approaches are not enough

- **Homomorphic encryption**
    - Too cumbersome for training modern DL models
- **Differential Privacy**
    - Private at the cost of a large utility loss, especially in healthcare
- **Lightweight encoding schemes**
    - Allow downstream training of DL models but are not private

Dauntless: Data Augmentation and Uniform Transformation for Learning with Scalability and Security

Hanshen Xiao and Srinivas Devadas

MIT, {hsxiao,devadas}@mit.edu

InstaHide: Instance-hiding Schemes for Private Distributed Learning*

Yangsibo Huang† Zhao Song‡ Kai Li§ Sanjeev Arora¶

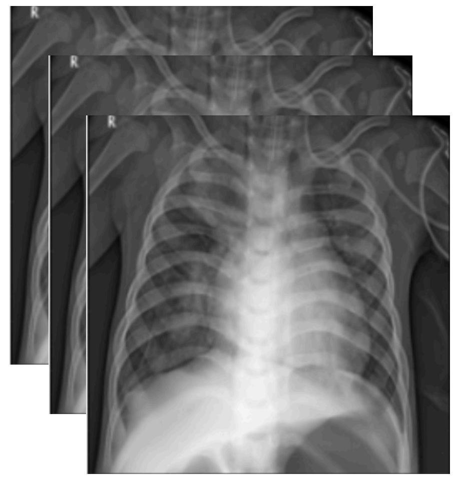
Existing approaches are not enough

- **Homomorphic encryption**
  - Too cumbersome for training modern DL models
- **Differential Privacy**
  - Private at the cost of a large utility loss, especially in healthcare
- **Lightweight encoding schemes**
  - Allow downstream training of DL models but are not private

⟶ **Need a method to evaluate the privacy of encoding schemes**

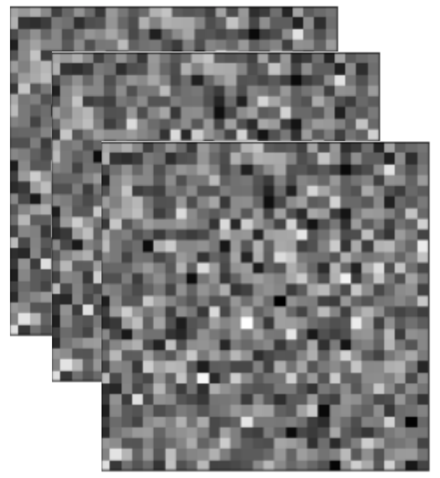
Ideal use case

Private (PHI)



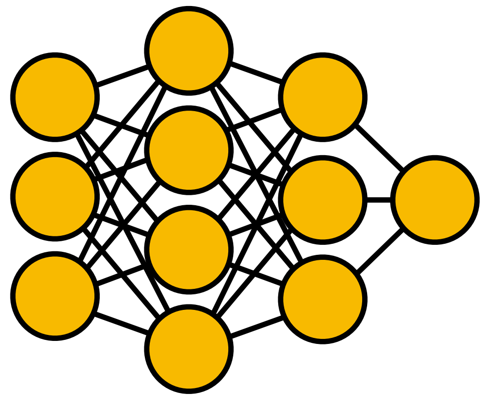
encode

Secure encodings

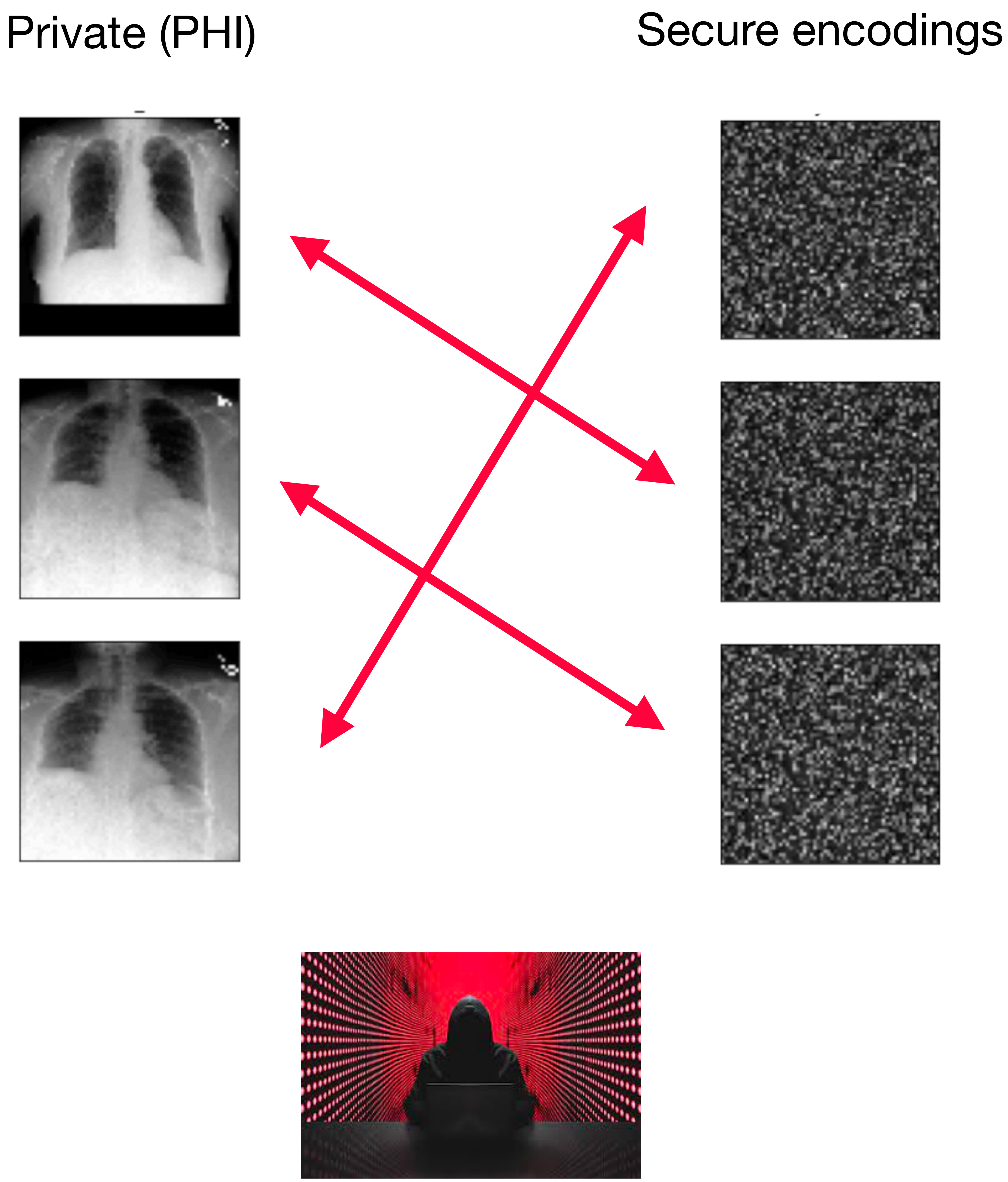


train

classifier



Threat model

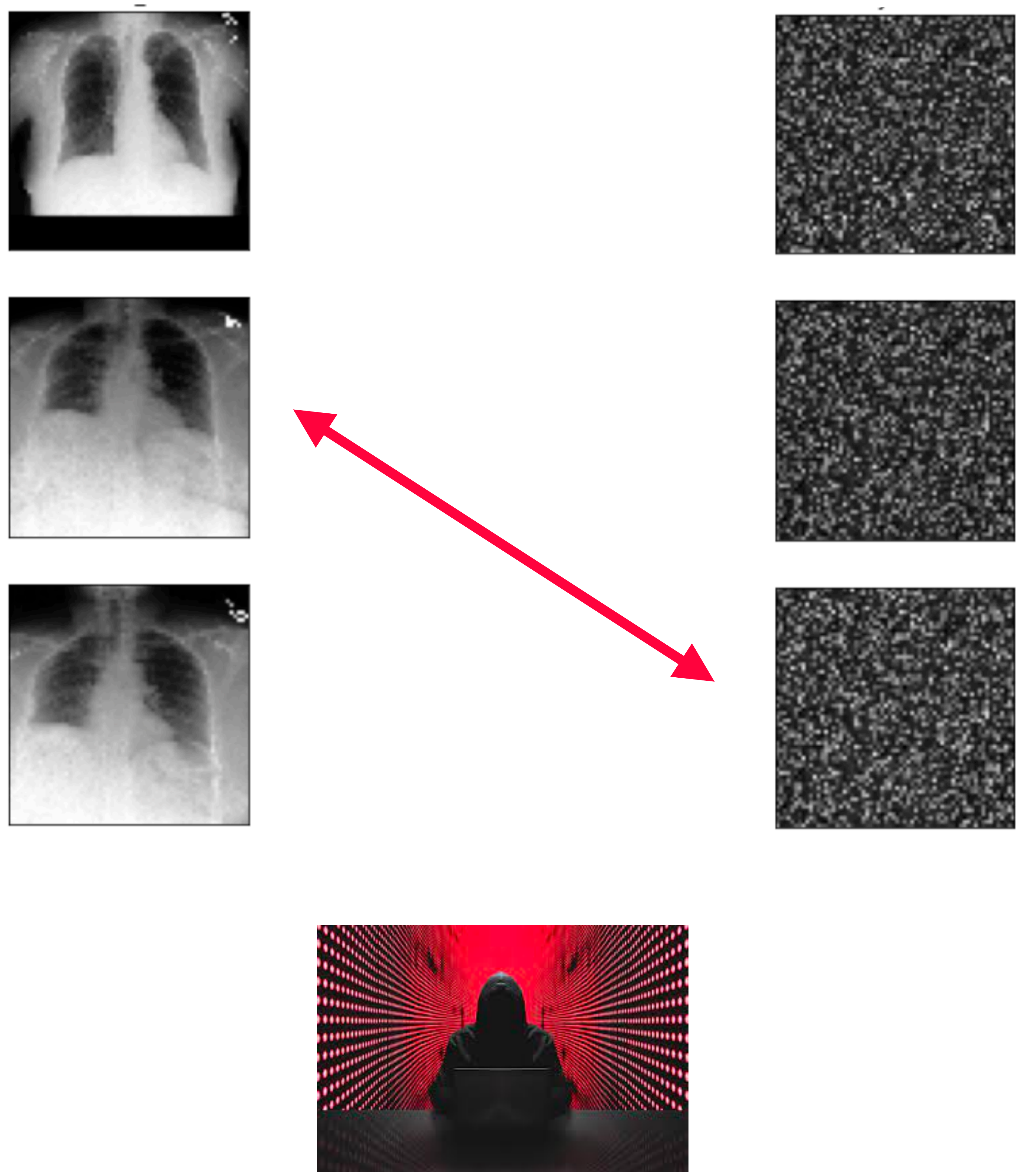


An attacker who observes the plaintext data and the encoded data should not be able to reconstruct the matching.

Threat model

Private (PHI)

Secure encodings

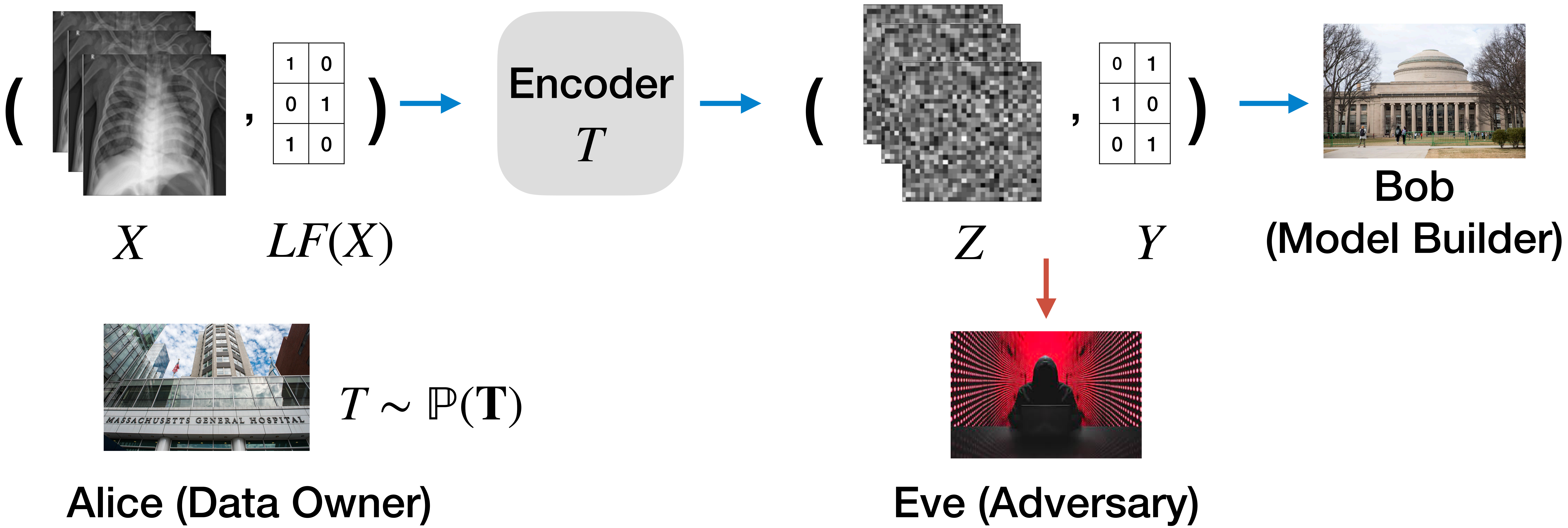


An attacker who observes the plaintext data and the encoded data should not be able to retrieve a ***single matching pair***.

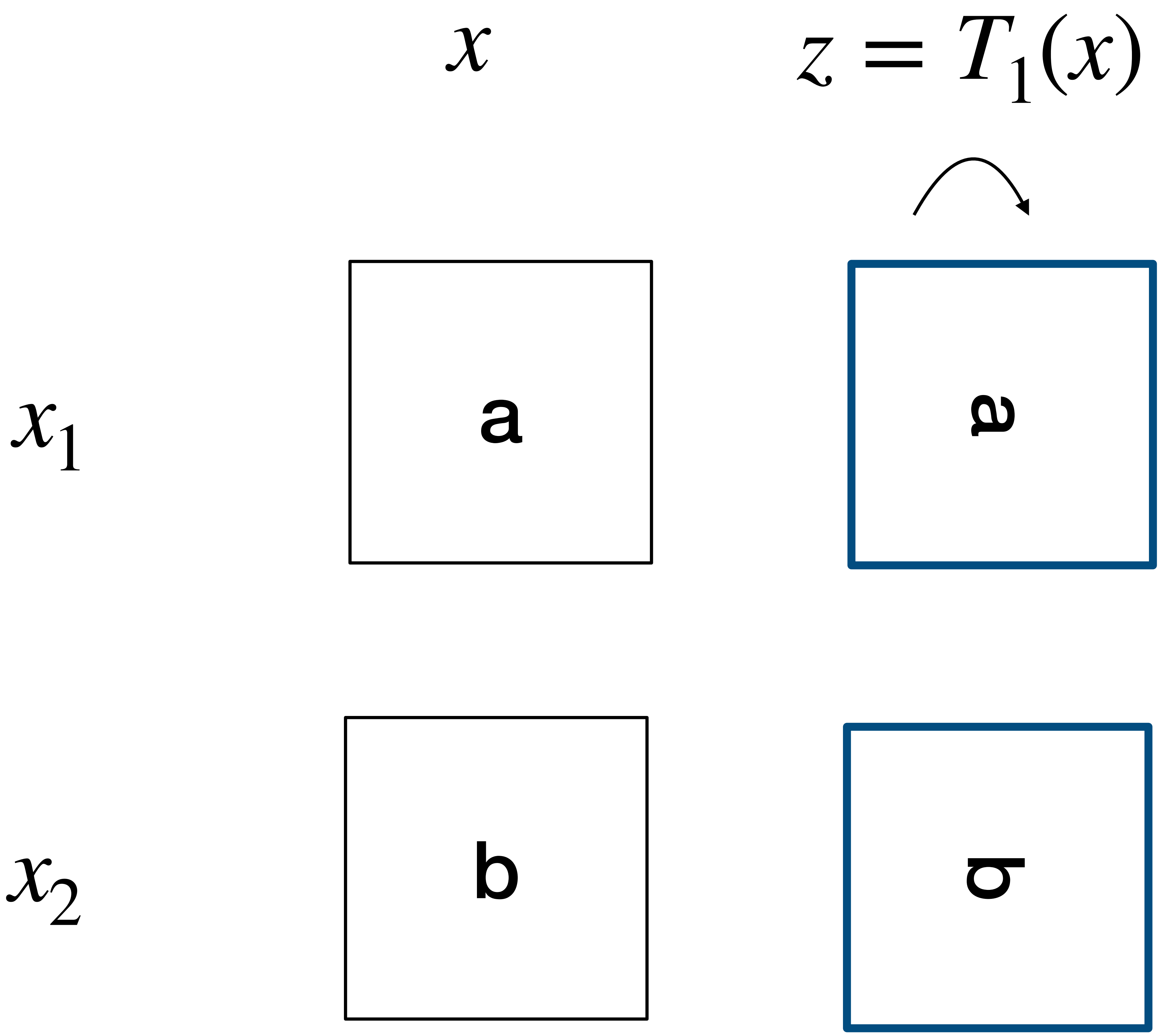
Attacker task = police line-up

What is the plaintext image corresponding to ()

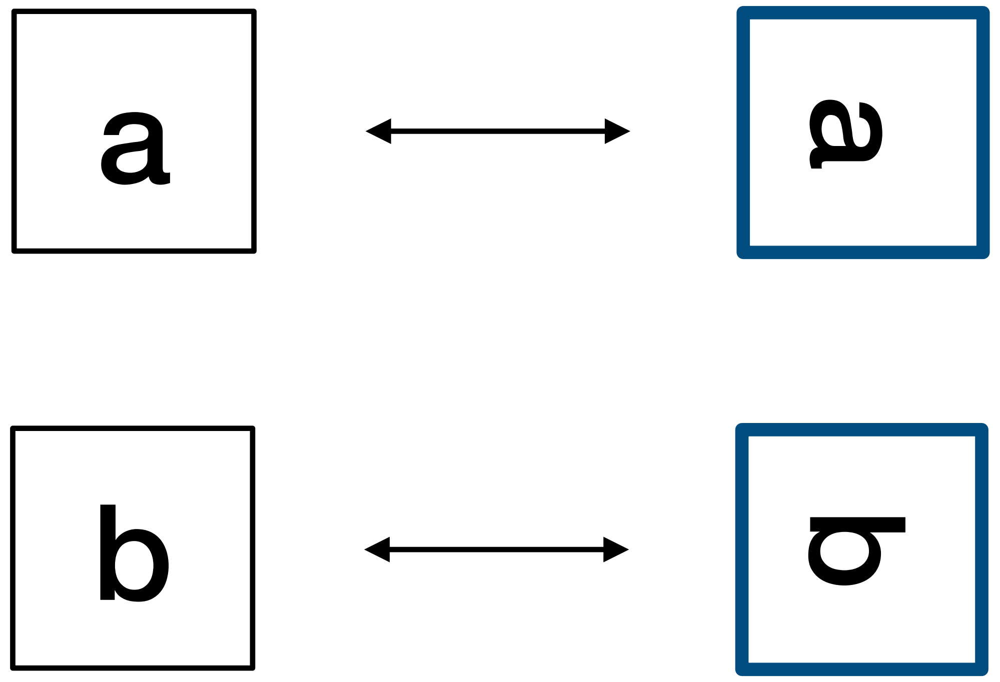
Formal setting



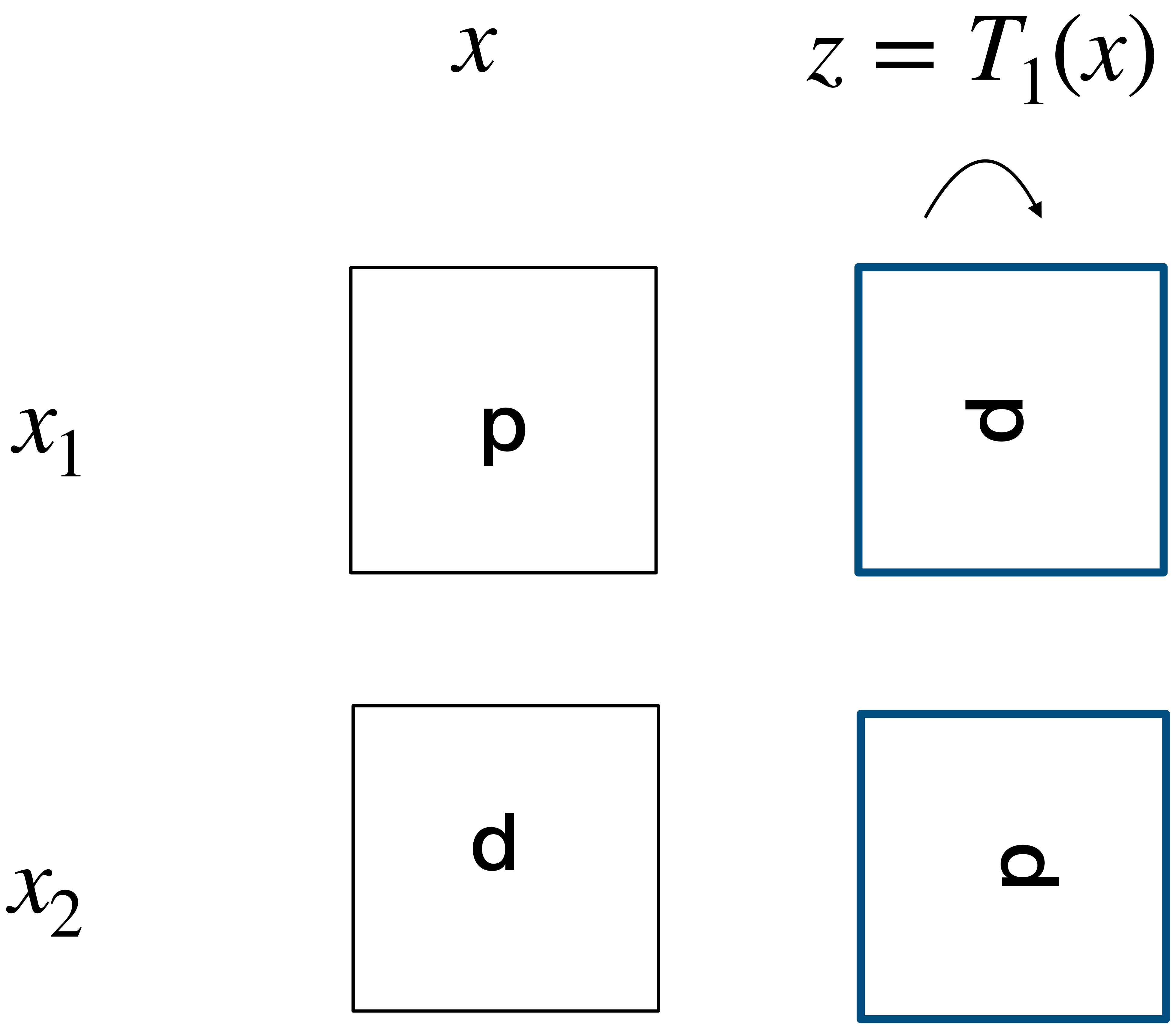
Toy example and intuition



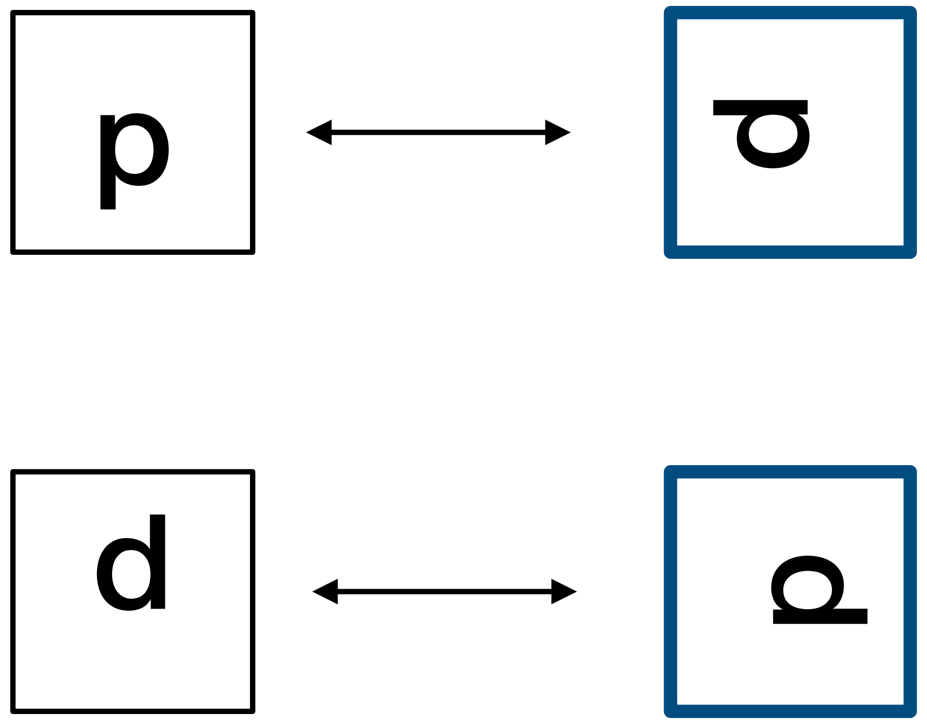
If Alice only ever uses T_1 the $-90°$ rotation, Eve can deduce the matching:



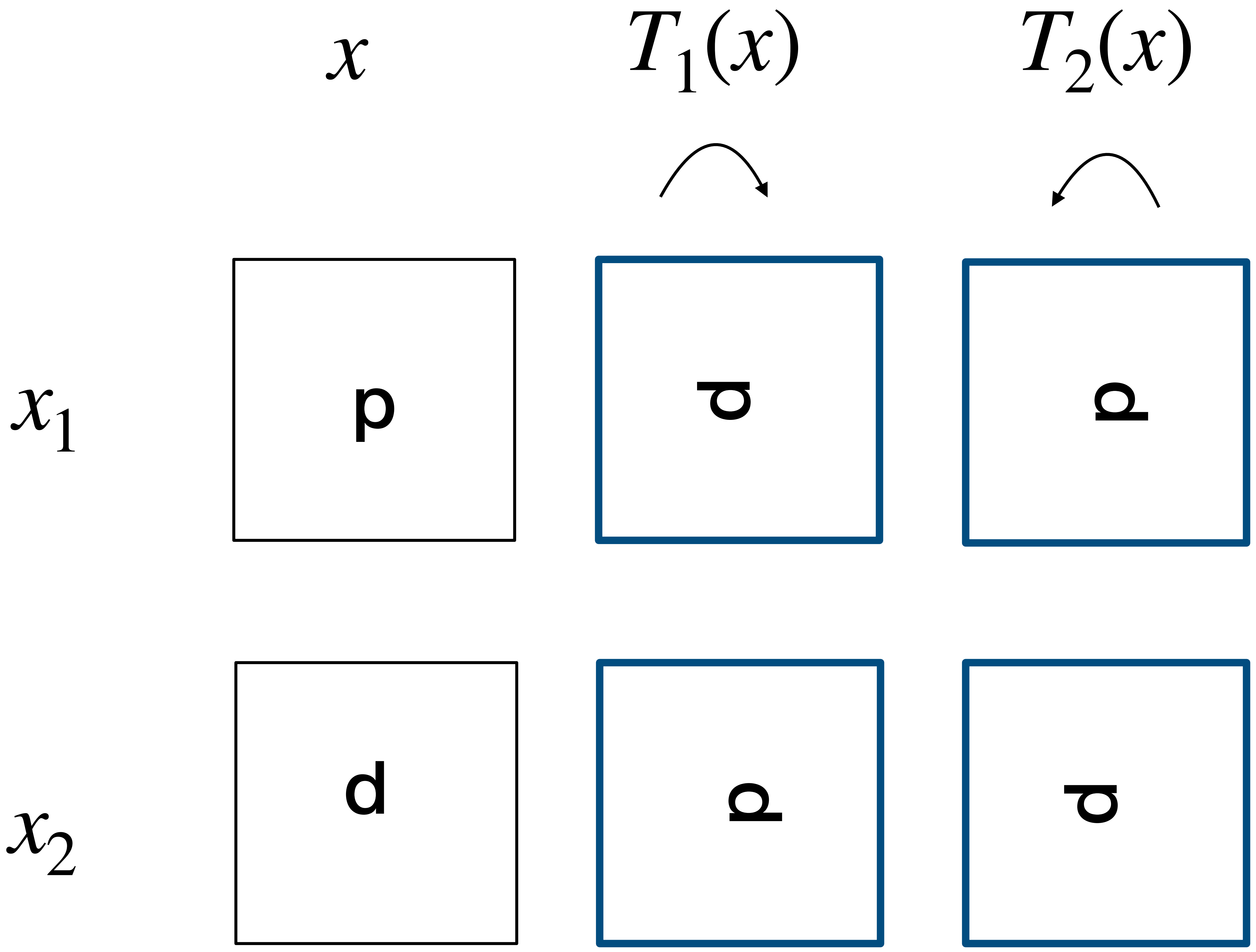
Toy example and intuition



If Alice only ever uses T_1 the $-90°$ rotation, Eve can deduce the matching:

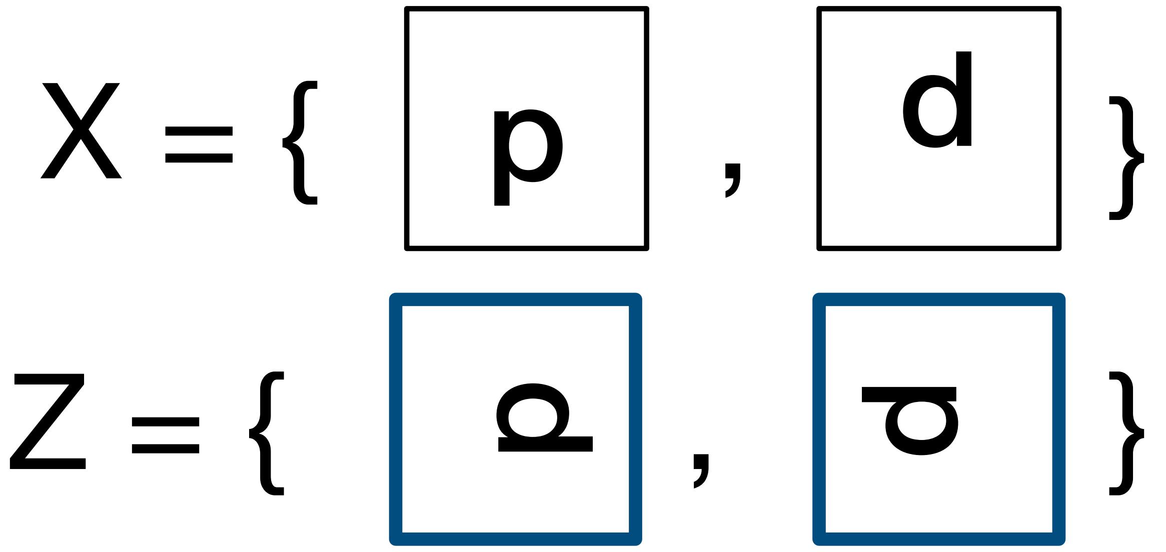


Toy example and intuition



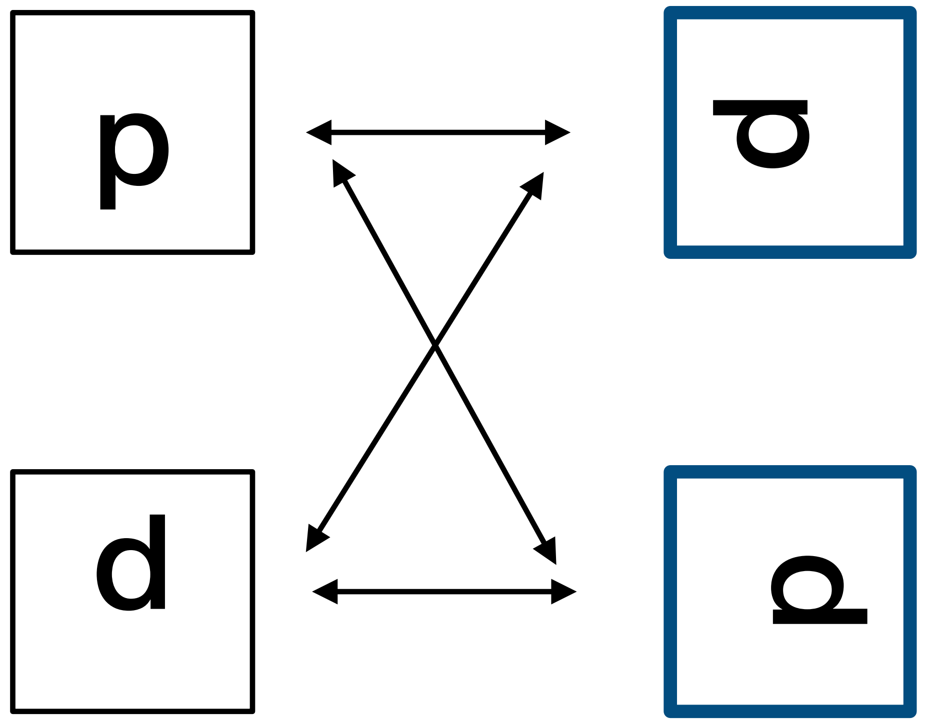
Now Alice uses T_1 and T_2 with equal probability.

Eve observes:

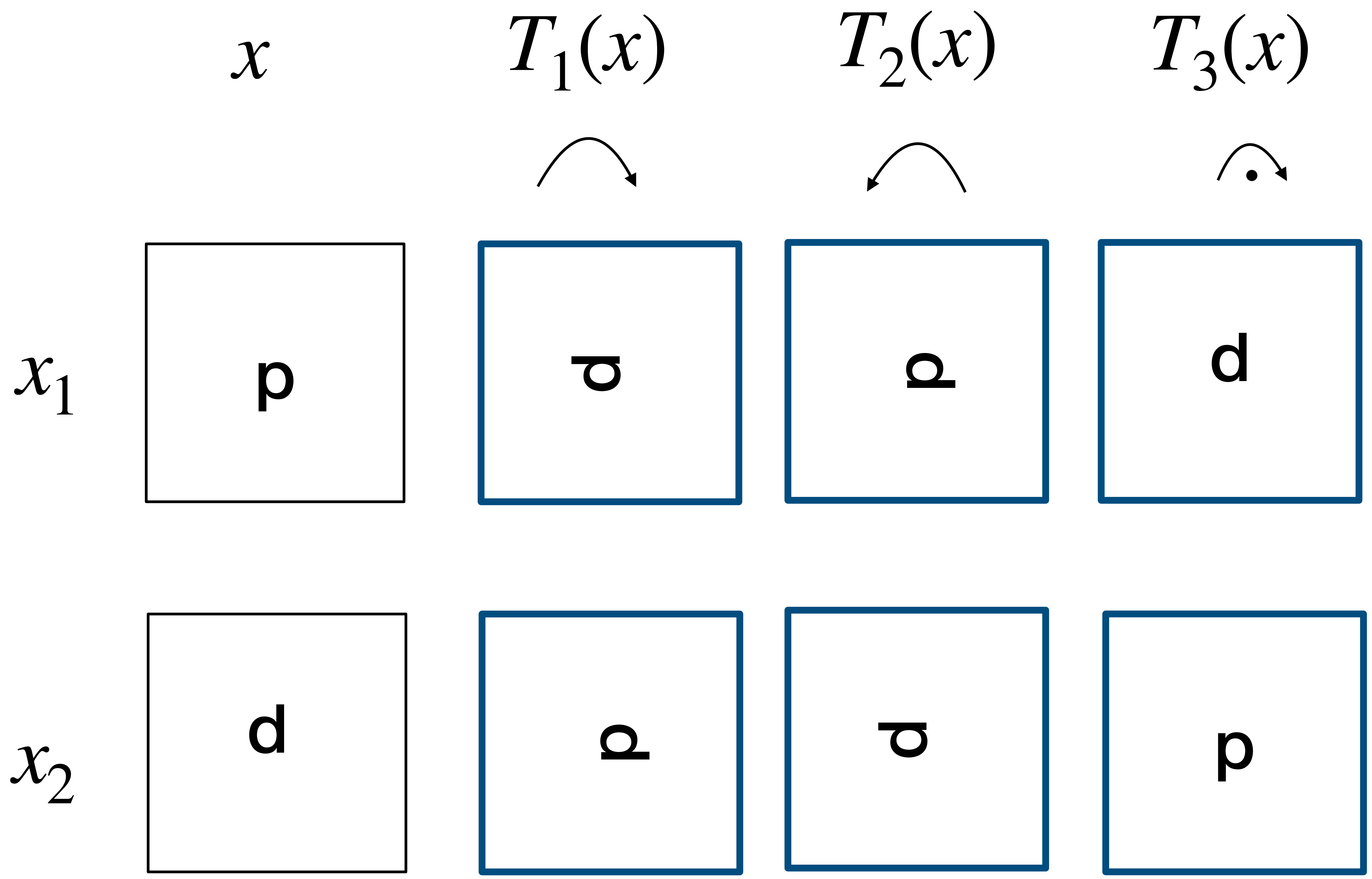


and $\mathscr{T} = \{T_1, T_2\}$

There are two possible matchings



Toy example and intuition



Alice uses $\mathscr{T} = \{T_1, T_2, T_3\}$

With probability 1/3, Eve observes:

$X = \{ \boxed{p}, \boxed{d} \}$

$Z = \{ \boxed{d}, \boxed{p} \}$

She would then deduce

$\boxed{p} \longleftrightarrow \boxed{d}$

$\boxed{d} \longleftrightarrow \boxed{p}$

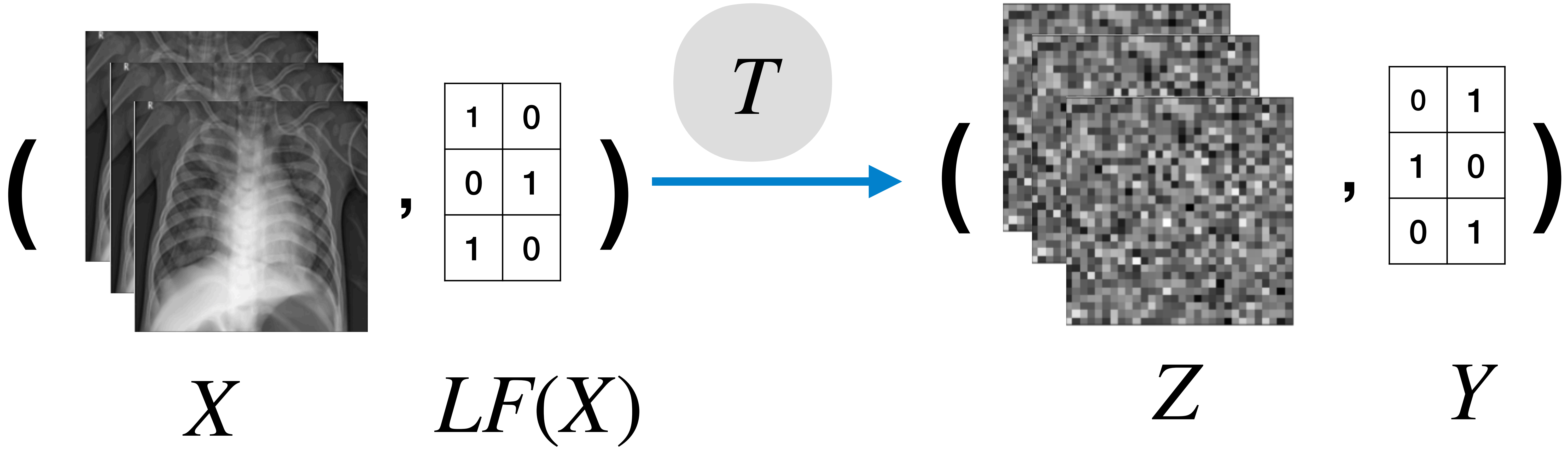
Toy example and intuition

- **Takeaways**
    - Whether T is private or not depends on $\mathscr{T}$ (set of transformations used by Alice) and more generally on the distribution $\mathbb{P}(\mathbf{T})$ that Alice uses to sample T
    - Adding more Ts *does not* make an encoding scheme more private
    - Designing an encoding scheme = finding a *good* distribution $\mathbb{P}(\mathbf{T})$

Formal setting - Alice Data Owner

$T \sim \mathbb{P}(\mathbf{T})$

Alice (Data Owner)

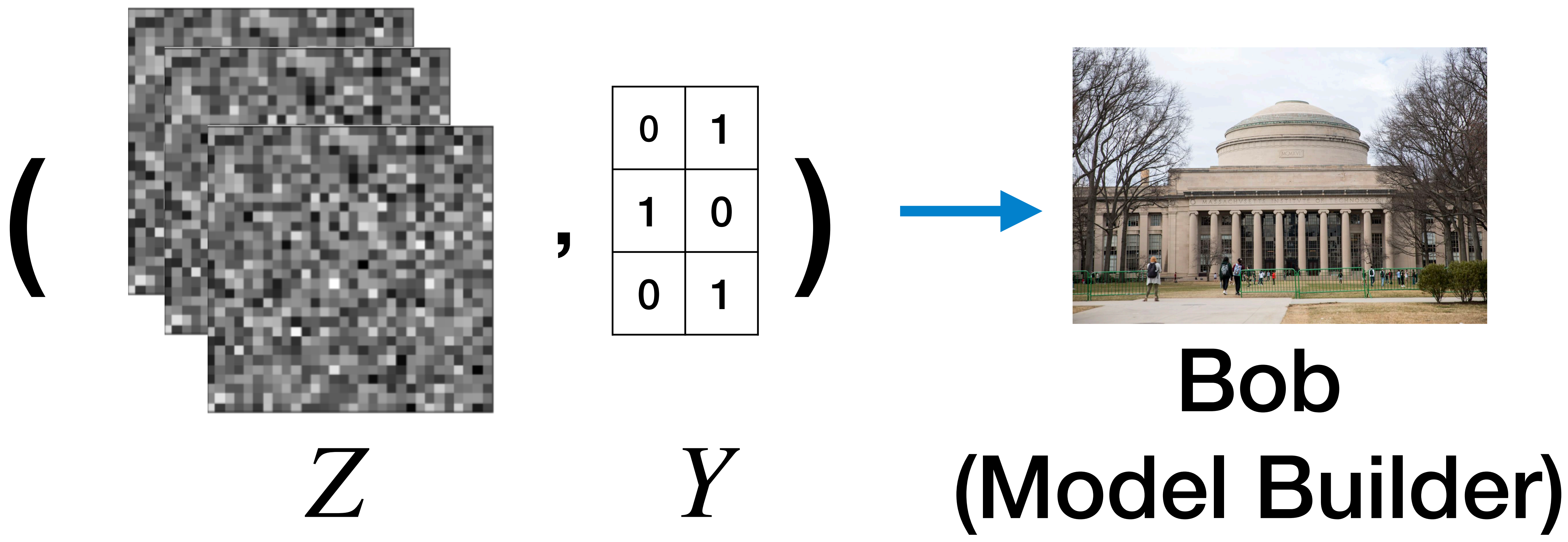


Alice owns a dataset X with labels $LF(X)$

She samples a transformation $T \sim P(\mathbf{T})$ and

releases the encoded data $(Z, Y) = T(X, LF(X))$

Formal setting - Bob Model Builder



Bob receives the encoded data $(Z, Y) = T(X, LF(X))$

Bob trains a classifier C_T to minimize generalization error on the test set $(Z^{test}, Y^{test}) = T(X^{test}, LF(X^{test}))$

Bob sends C_T to Alice for usage on new data

Formal setting - Eve Adversary

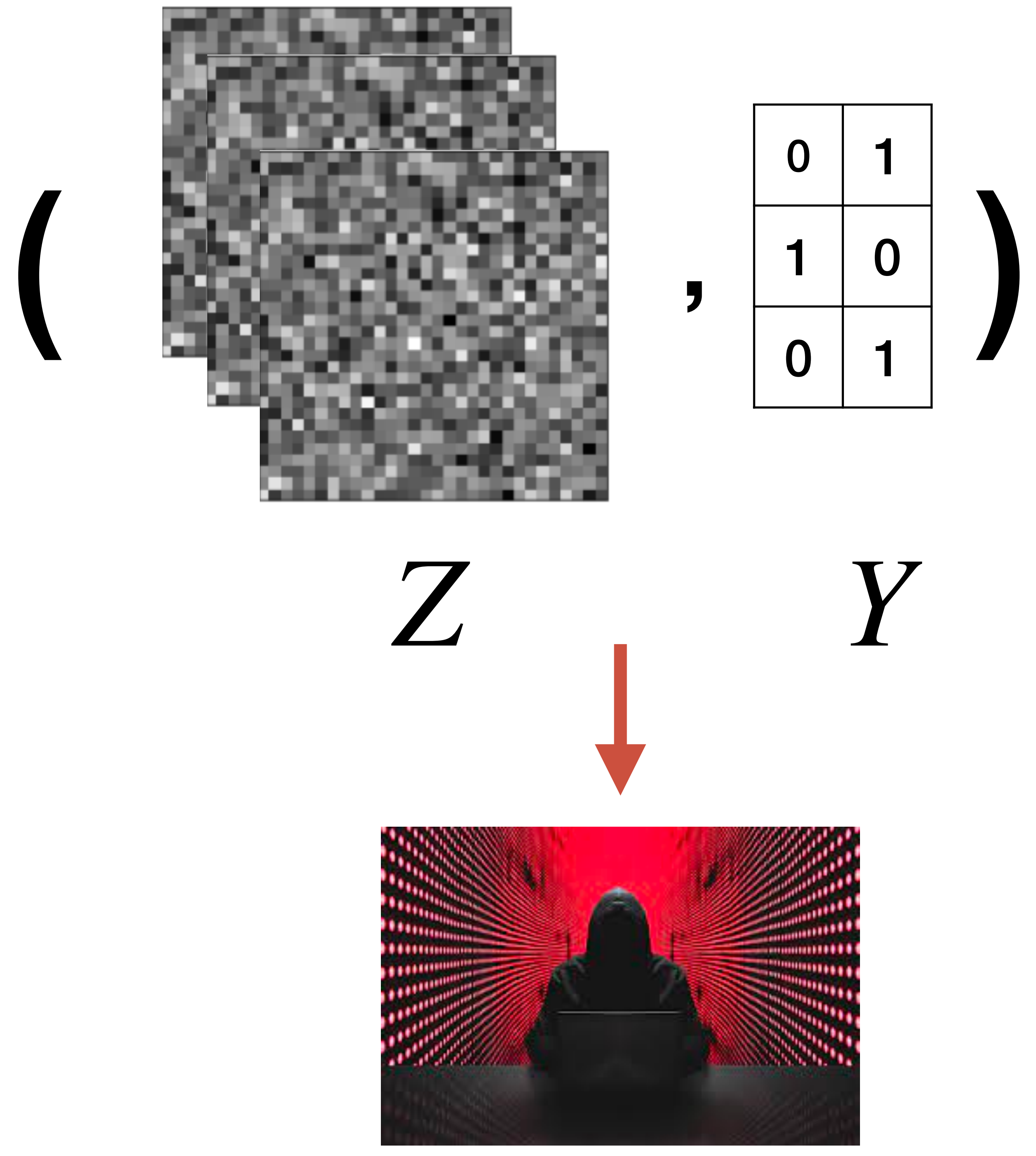
Eve observes the encoded data
$(Z, Y) = T(X, LF(X))$

Eve knows the encoding scheme used by Alice, i.e $\mathbb{P}(\mathbf{T})$

Eve possesses a $X_E \supseteq X$, and more generally a prior $\mathbb{P}(\mathbf{X_A} = X)$

Eve ***does not know*** T, which acts as Alice's private key.

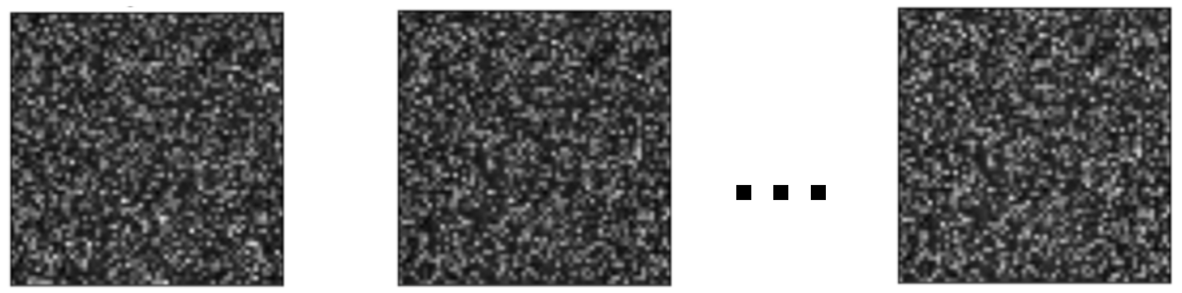
Goal: re-identify any one private image



Eve (Adversary)

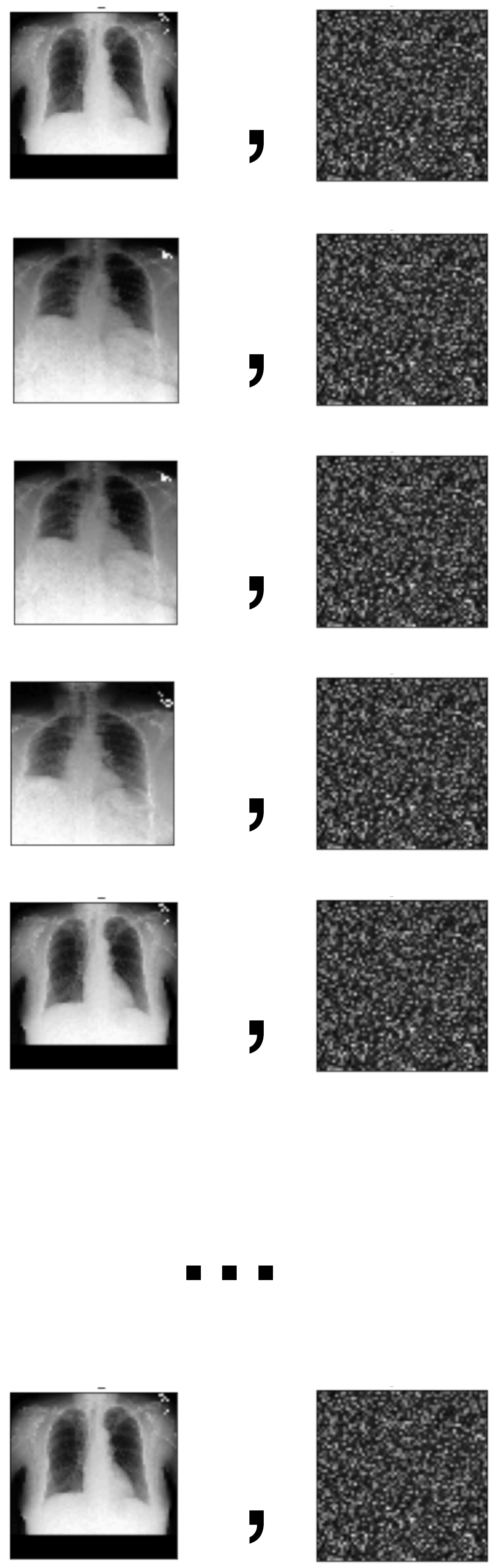
Privacy definition - Guesswork

Given $X_E \supseteq X_A = \{$  $\}$

and $Z = \{$  $\}$

A computationally unbounded Eve uses her knowledge of $\mathbb{P}(T)$ to compute for each pair (x, z) the probability that they match and ranks the *all possible pairs* from most likely to least likely.

Eve's guesses:



1. (,)
2. (,)
3. (,)
4. (,)
5. (,)

...

$|X_E| \times |Z|$ (,)

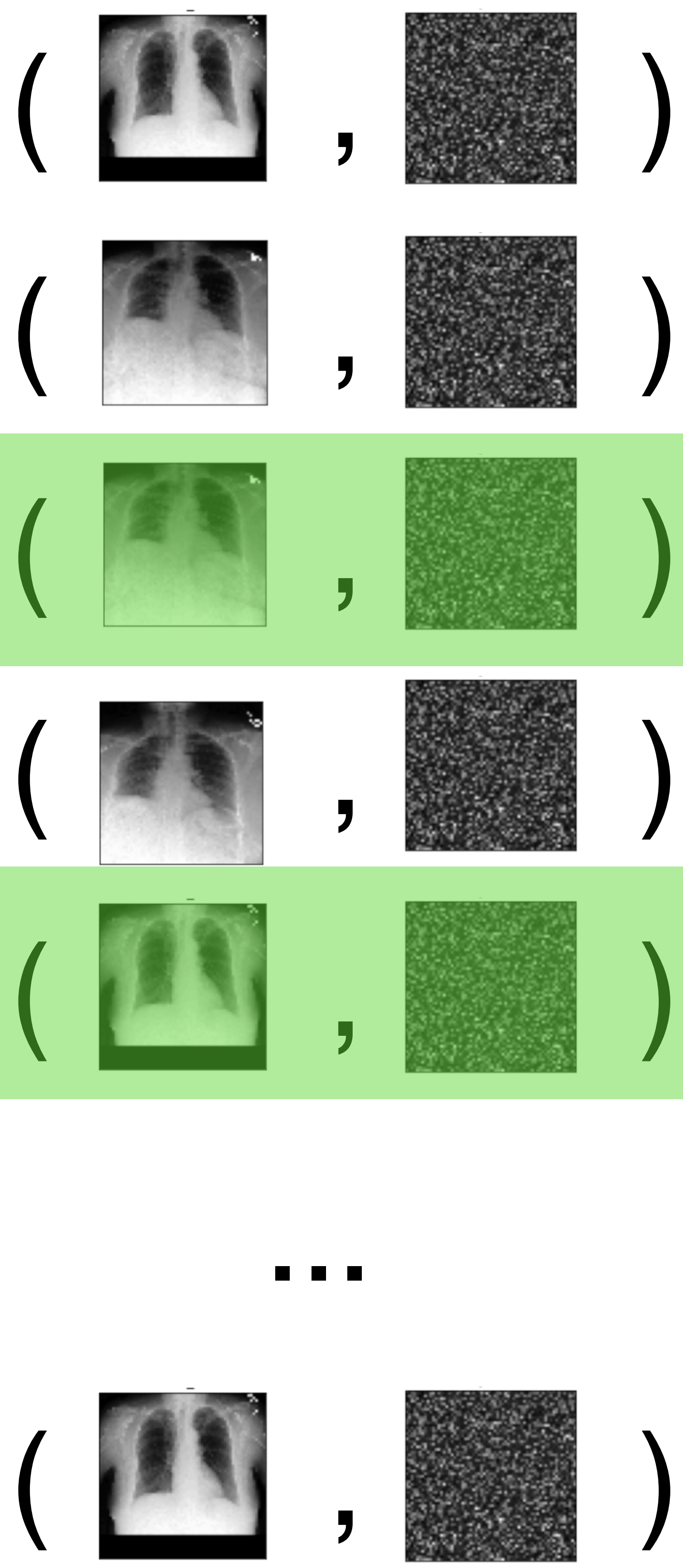
Privacy definition - Guesswork

Eve's guesses:

1.
2.
3.
4.
5.

...

$|X_E| \times |Z|$



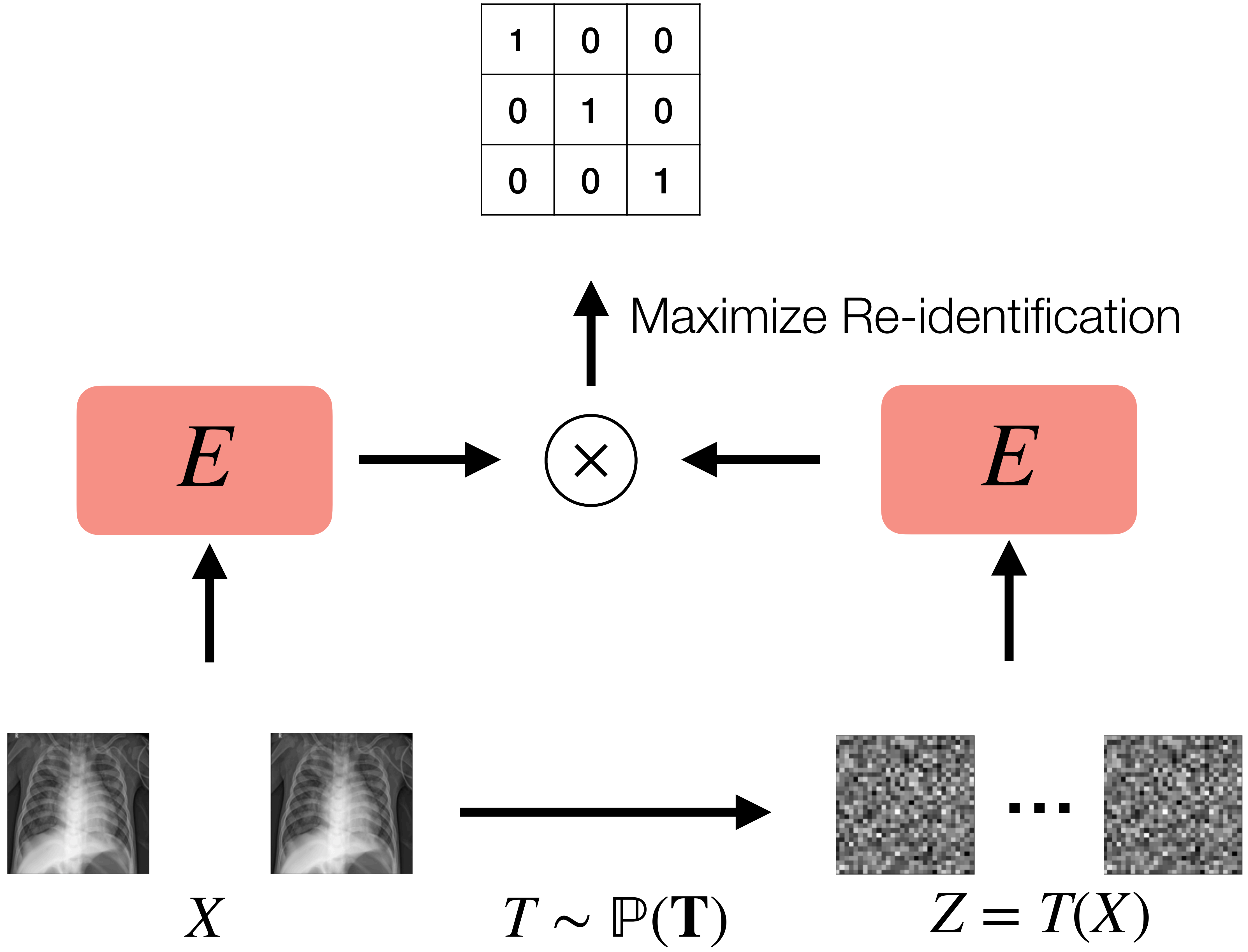
In Eve's list, exactly $|Z|$ pairs are correct.

We define ***guesswork*** as the index of the ***first correct guess.***

$\longleftarrow \mathscr{G} = 3$

Privacy estimation

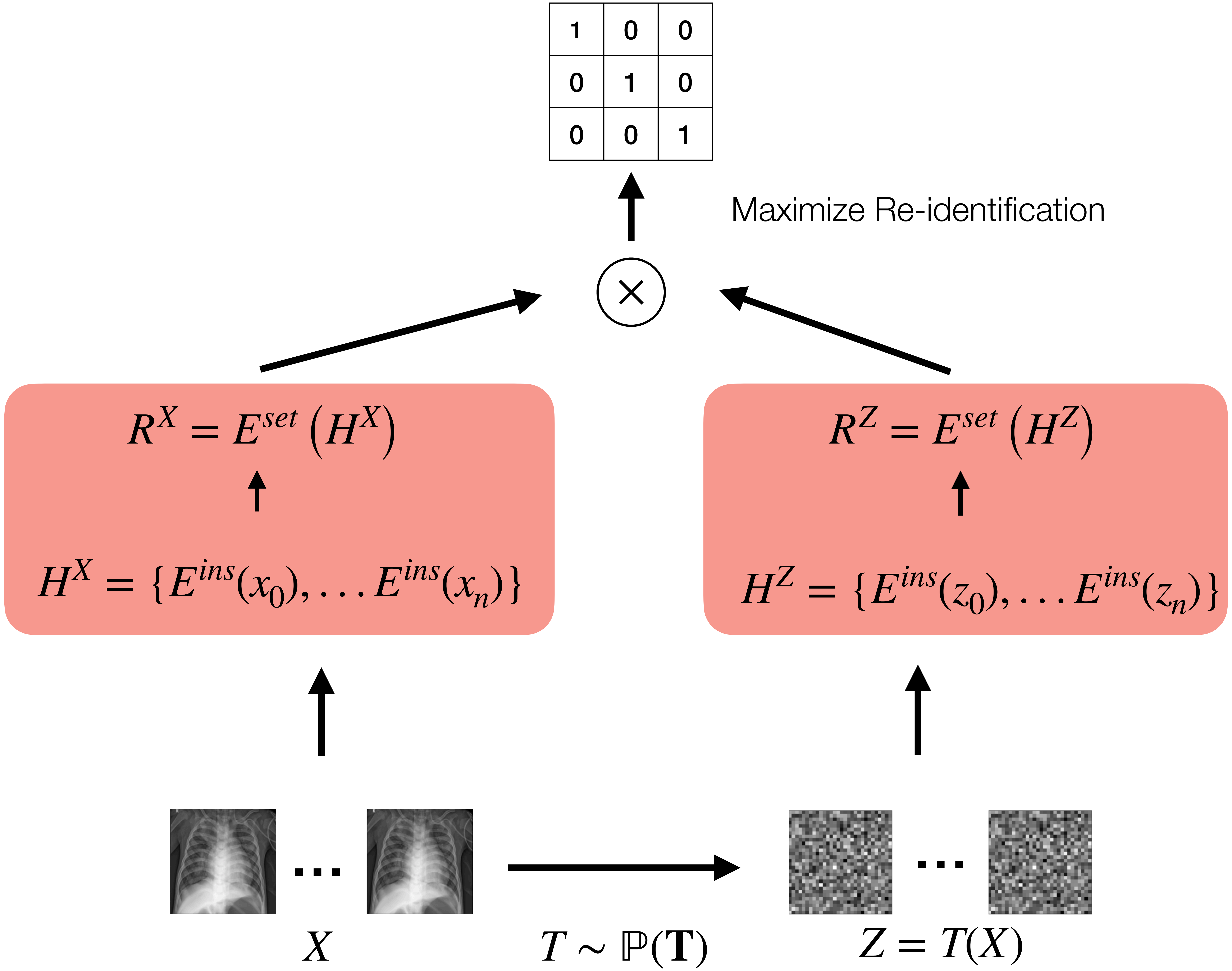
- Can’t simulate true computationally unbounded Eve
- We estimate privacy with contrastive learning
- Our model-based attacker learns to estimate the probability that a pair (x, z) is a correct match $P((x, z) \in M_T)$



Privacy estimation - details

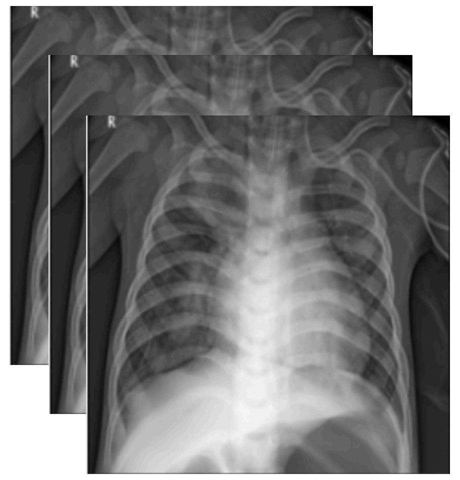
$$\hat{p}(x_i, z_j) = \frac{\exp(\mathrm{sim}(r_i^X, r_j^Z))}{\sum_{k,l} \exp(\mathrm{sim}(r_k^X, r_l^Z))}$$

$$\mathscr{L}_{reidentification} = -\sum_{(x,z)\in M_T} \log(\hat{p}(x, z))$$



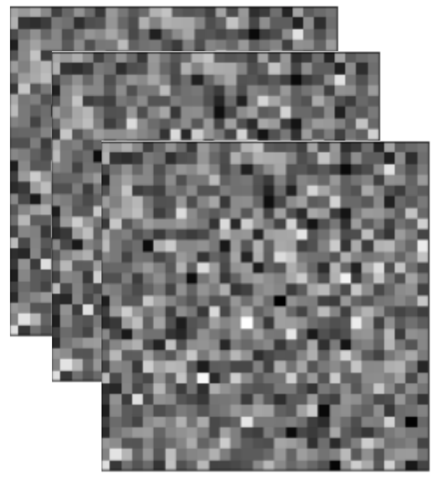
Ideal use case

Private (PHI)



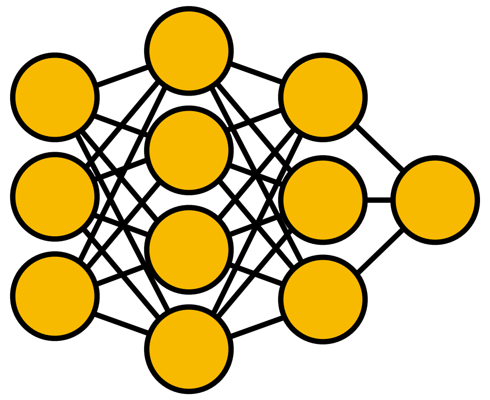
encode

Secure encodings



train

classifier



- **Desiderata:**
  - Protect raw data identity (HIPAA), i.e. **achieve high guesswork**
  - Support **any downstream task** with standard ML tools
  - Data owner **does not train** any models
  - No centralized coordination, publish encoded dataset

Main challenge

How to build a distribution $\mathbb{P}(\mathbf{T})$

... that achieves privacy

... while maintaining
downstream utility on
tasks of interest

... without knowing the
tasks *a priori*

... nor having access to the
private data

Main challenge

How to build a distribution $\mathbb{P}(\mathbf{T})$

... that achieves privacy → Always output 0 as the "encoded data", i.e. $\mathscr{T} = \{T : x \mapsto 0\}$

... while maintaining downstream utility on tasks of interest → Train a classifier and output predicted labels as the "encoded data"

... without knowing the tasks *a priori* → ***Syfer***: we model T as a neural network and learn a "good" distribution $\mathbb{P}(\mathbf{T})$ using public data

... nor having access to the private data

Proposed Encoding scheme: Syfer

$T = (T^X, T^Y)$

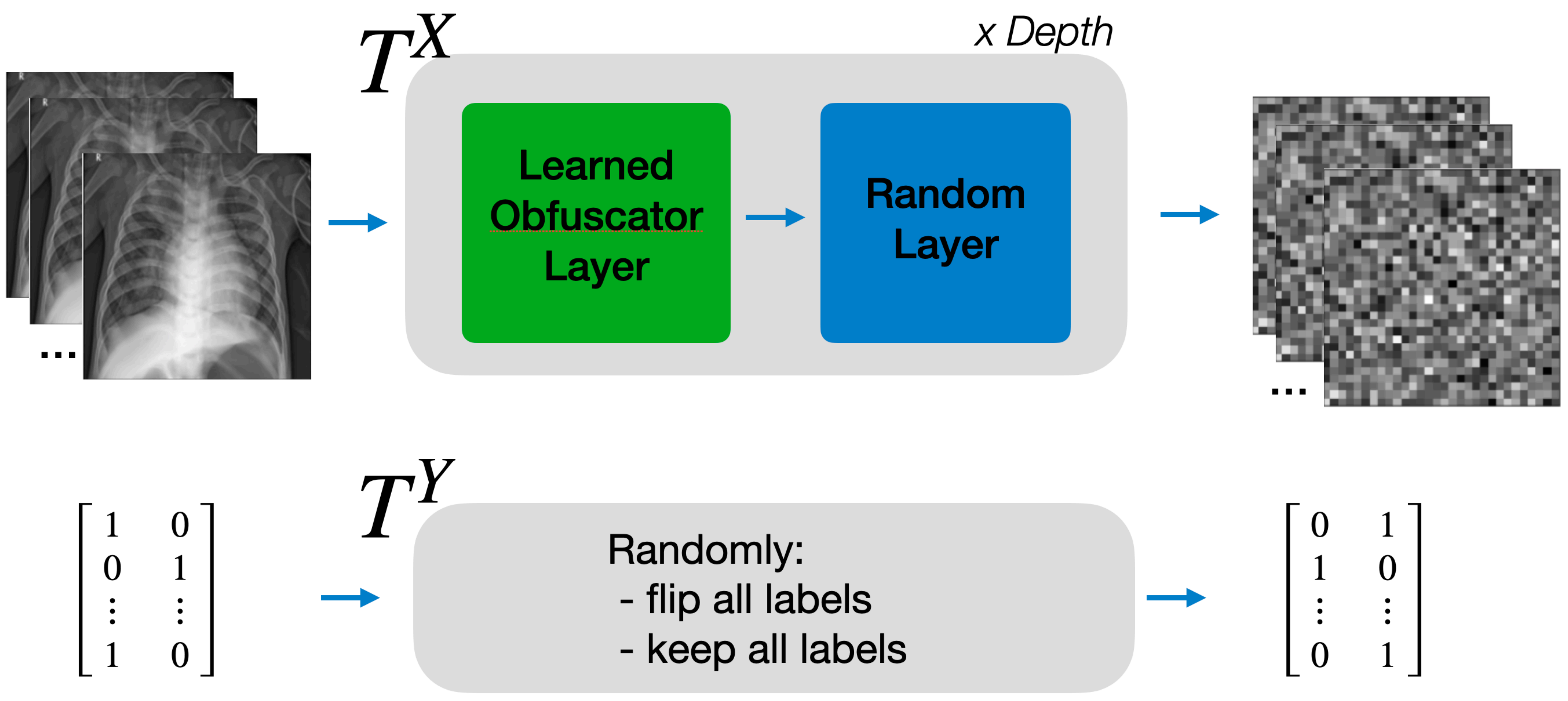
Neural encoder T^X

We decompose T^X in blocks of **obfuscator layers** and **random layers**.

In practice:

- The **learned obfuscator weights** are known to all actors (Alice and Eve)

- To construct a T^X, Alice samples **random layer weights**

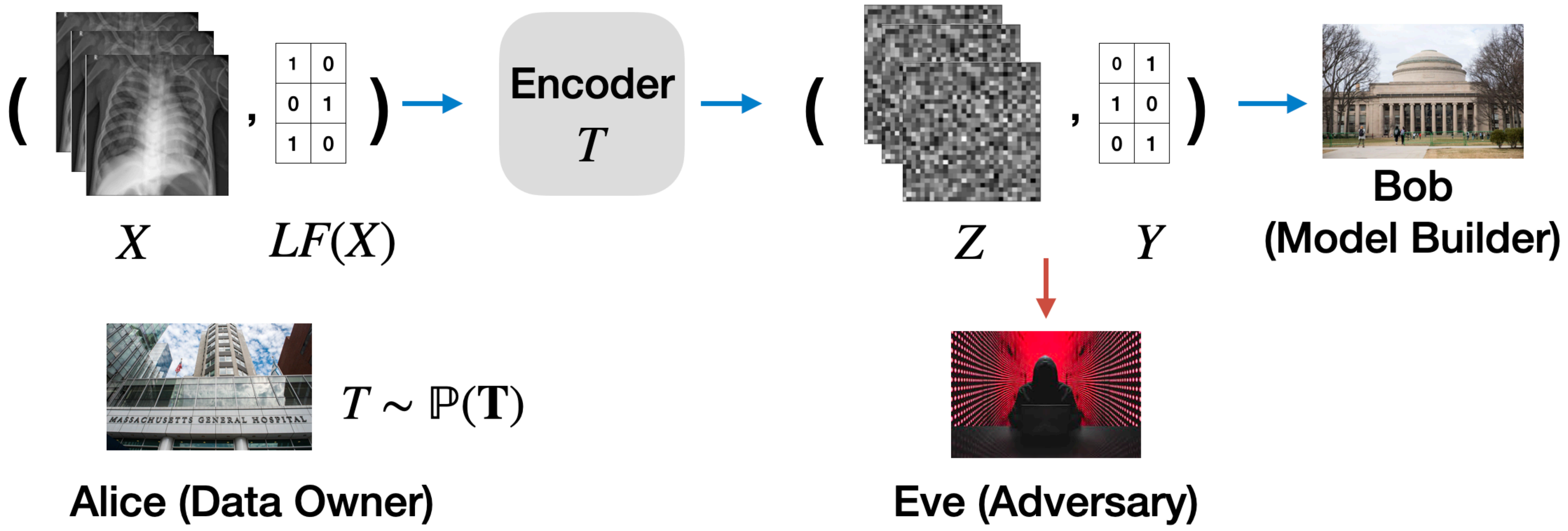


Label encoder T^Y

In practice: Alice **randomly** decides to flip the labels or not.

Motivation for training

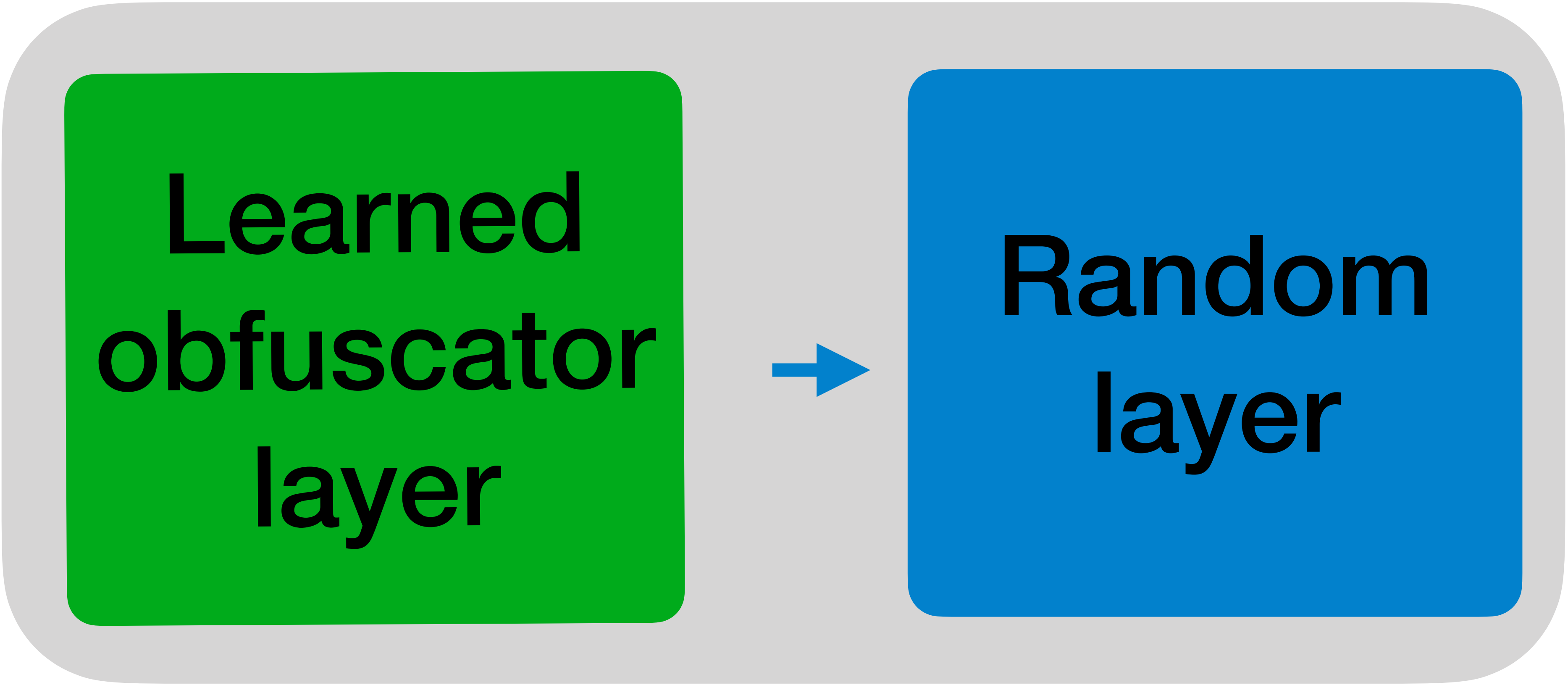
Reminder:How do we evaluate Syfer?



Eve knows X and needs to generalize to **unknown** T

Bob sees $Z^{train} = T(X^{train})$ and needs to generalize to **unknown** X^{test}

Syfer Training Algorithm



The **obfuscater layers** are trained on a public unlabeled dataset X_{public} to optimize

$$\mathscr{L}_{Syfer} = \mathscr{L}_{reconstruction} - \mathscr{L}_{reidentification}$$

where

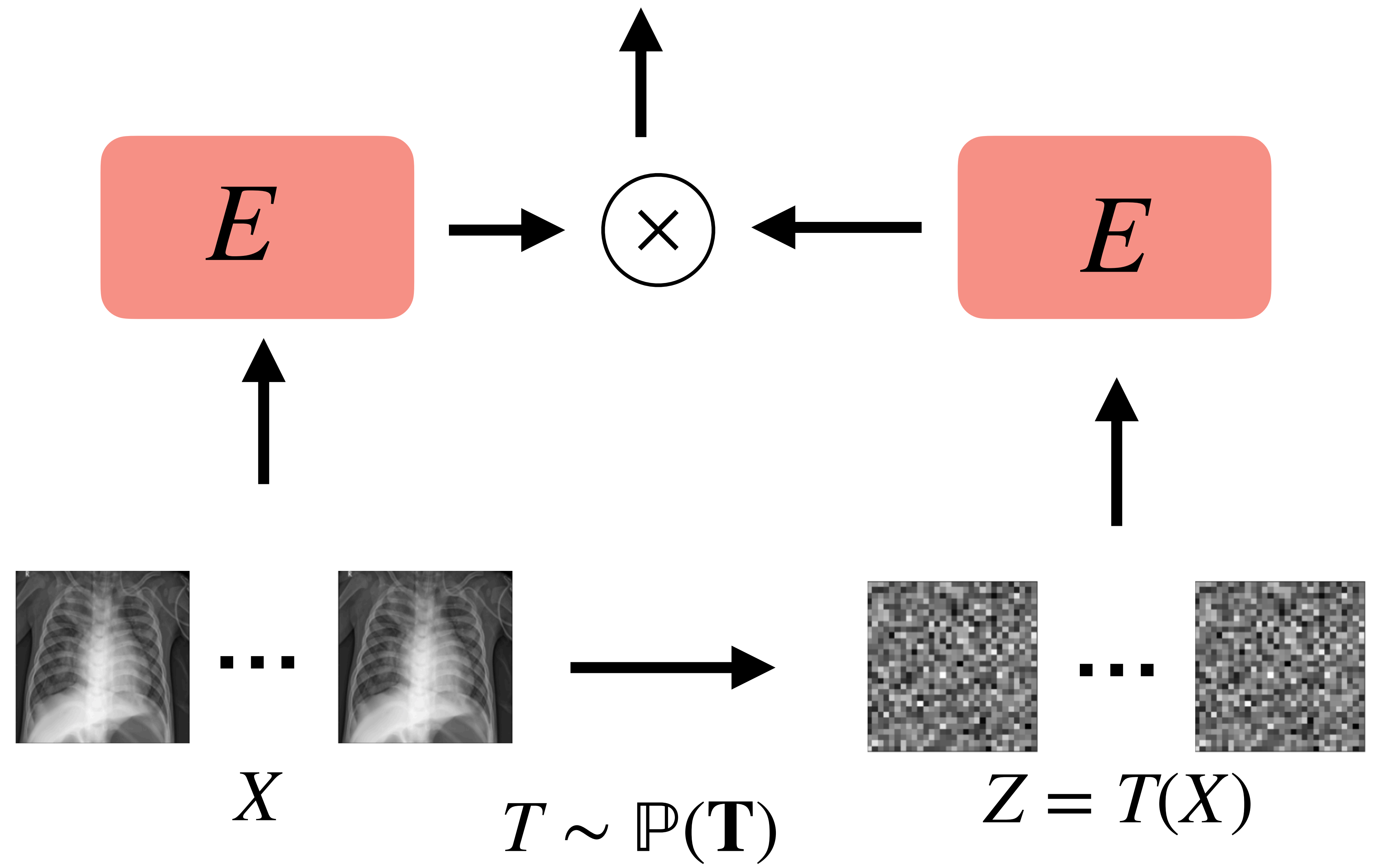
$\mathscr{L}_{reidentification}$ = Re-identification loss of **an adversary**

$\mathscr{L}_{reconstruction}$ = Reconstruction loss of a decoder D_T for a **fixed choice of random layers**

Syfer Training Algorithm

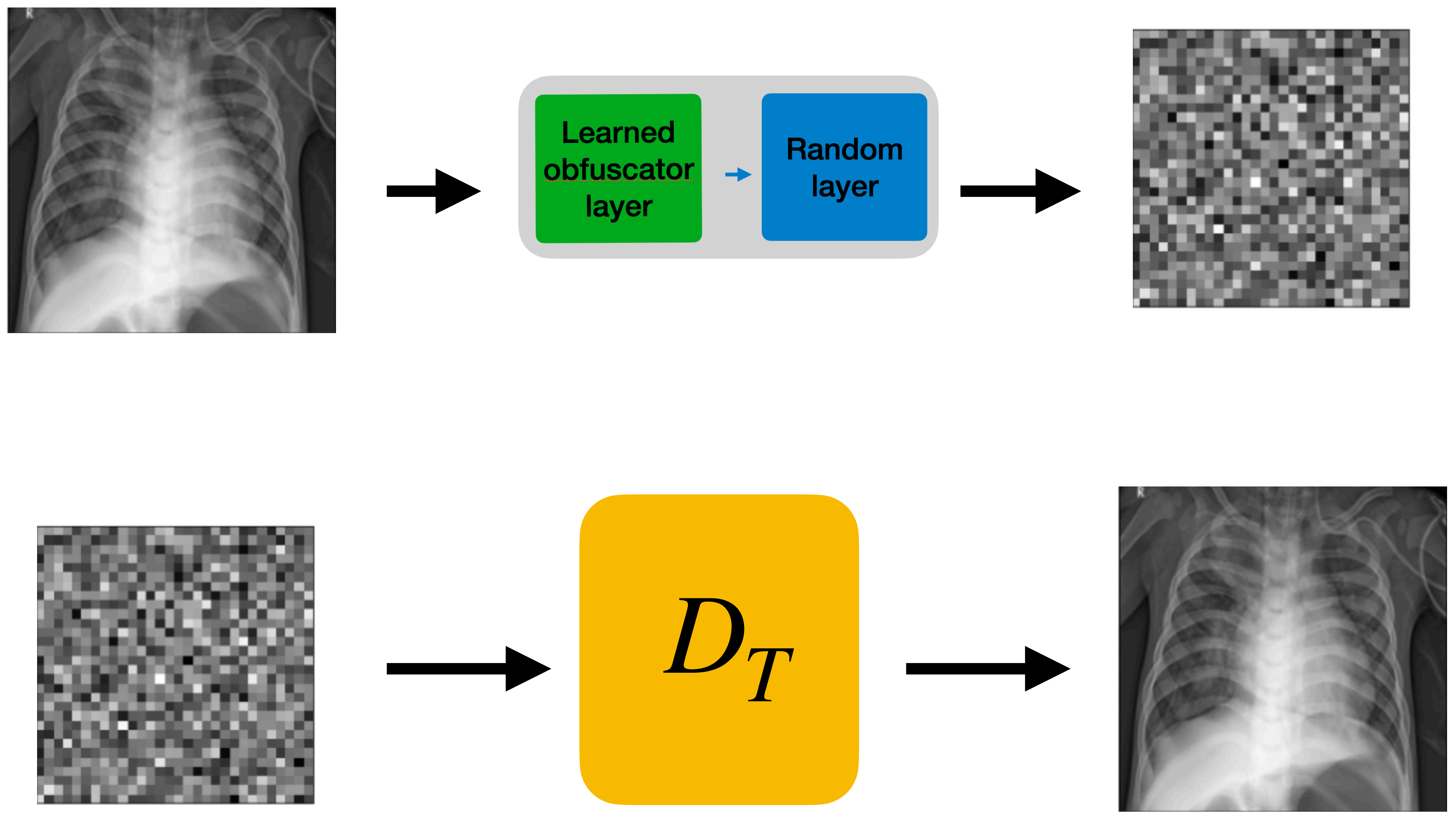
Re-identification loss of **an adversary**

$$\mathscr{L}_{reidentification} = -\sum_{(x,z)\in M_T} \log(\hat{p}(x,z))$$



For a **fixed choice of random weights**, train a decoder D_T to minimize a reconstruction loss

$$\mathscr{L}_{reconstruction} = \mathbb{E}_T[\mathbb{E}_X[\,||x - D_T(T(x))||^2]]$$



Syfer Training Algorithm

The **obfuscater layers** are trained on a public unlabeled dataset X_{public} to optimize

$$\mathscr{L}_{Syfer} = \mathscr{L}_{reconstruction} - \mathscr{L}_{reidentification}$$

The **adversary model** E is alternatively updated to minimize $\mathscr{L}_{reidentification}$

The **decoder model** D_T is alternatively updated to minimize $\mathscr{L}_{reconstruction}$

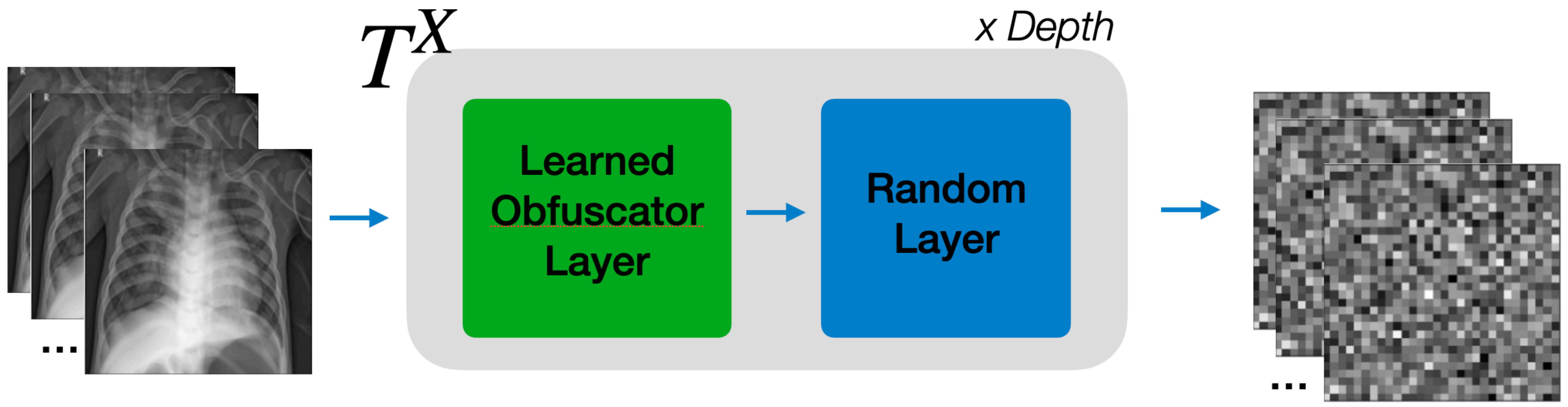
Algorithm 1 *Syfer* training

Initialize obfuscator parameters θ_{Syfer}
Initialize attacker E with parameters $\varphi = (\varphi^{\text{ins}}, \varphi^{\text{set}})$
Initialize decoders $D_1, \dots D_s$ with parameters $\beta_1, \dots, \beta_s$
For each decoder, sample random layer weights $\theta^1_{key}, \dots \theta^s_{key}$ (fixed throughout training)
Set flag *optimize_estimators* $\leftarrow$ true
repeat
    Sample a batch of datapoints X from X^{public}
    ▷ Step 1: Compute re-identification loss
    Sample a set of random layer weights $\theta^{\text{batch}}_{key}$
    Using obfuscator parameters θ_{Syfer} and key $\theta^{\text{batch}}_{key}$:
    $T^{\text{batch}} \leftarrow f(\theta_{Syfer}, \theta^{\text{batch}}_{key})$
    $\left(Z^{\text{batch}}, Y^{\text{batch}}\right) \leftarrow T^{\text{batch}}(X, LF(X))$
    $R^Z \leftarrow E_\varphi\left(Z^{\text{batch}}, Y^{\text{batch}}\right)$
    $R^X \leftarrow E_\varphi(X, LF(X))$
    $\mathcal{L}_{\text{reid}} \leftarrow \text{contrastive\_loss}\left(R^X, R^Z\right)$
    ▷ Step 2: Compute reconstruction loss
    $\mathcal{L}_{\text{rec}} \leftarrow 0$
    for $i \in \{1, \dots s\}$ **do**
        Using obfuscator parameters θ_{Syfer} and fixed key θ^i_{key}:
        $T^i \leftarrow f(\theta_{Syfer}, \theta^i_{key})$
        $(Z^i, Y^i) \leftarrow T^i(X, LF(X))$
        $\mathcal{L}_{\text{rec}} \leftarrow \mathcal{L}_{\text{rec}} + \text{MSE}\left(D_i\left(Z^i\right), X\right)$
    end for
    ▷ Step 3: Alternatively update parameters
    if *optimize_estimators* **then**
        $\varphi \leftarrow \varphi - \nabla_\varphi \mathcal{L}_{\text{reid}}$
        $\beta_i \leftarrow \beta_i - \nabla_{\beta_i} \mathcal{L}_{\text{rec}}$ {for $i \in \{1, \dots s\}$}
        optimize_estimators $\leftarrow$ false
    else
        $\theta_{Syfer} \leftarrow \theta_{Syfer} - \nabla_{\theta_{Syfer}}(\lambda_{\text{rec}} \cdot \mathcal{L}_{\text{rec}} - \lambda_{\text{reid}} \cdot \mathcal{L}_{\text{reid}})$
        optimize_estimators $\leftarrow$ true
    end if
until convergence

Experimental Setup

- Train Syfer and baselines on NIH Chest X-Ray dataset.
  - X_{public} = NIH
  - **Obfuscator** implemented as Simple Attention Unit (SAU)
  - **Rand Layer** implemented as Linear layer + SELU activation + LayerNorm
  - **Attacker** E, and **Decoder** D_T, implemented with SAUs.
- Test for Privacy and Utility on MIMIC Chest X-Ray dataset
  - Evaluate Syfer on **held out datasets** $(X, LF(X))$ and **held out attacker** architectures.

Experiments - Privacy evaluation

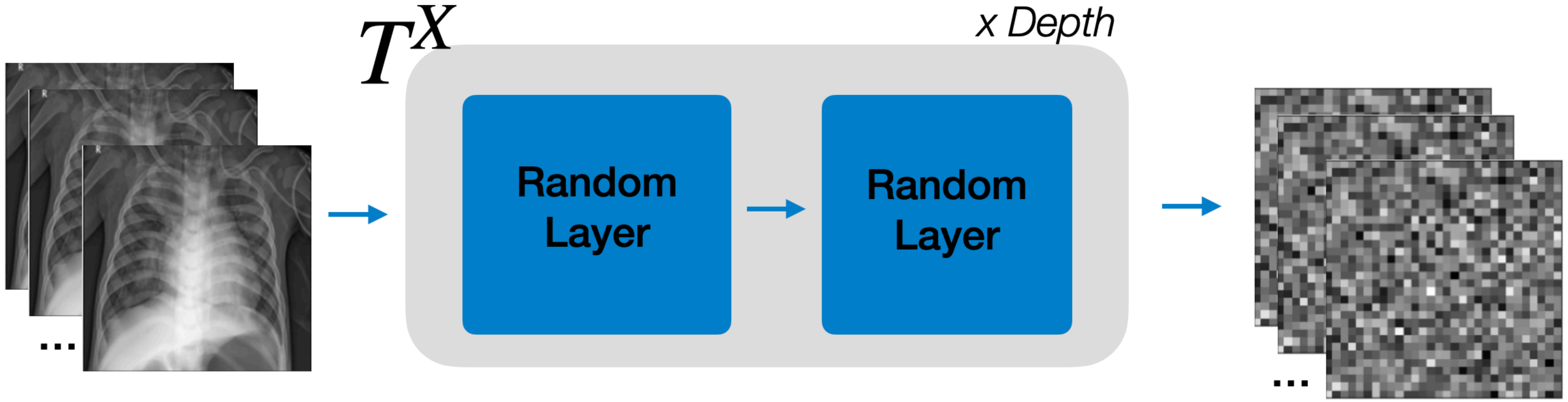


- **Generalized Privacy**: How secure are encodings Z when released alone (without labels)?
    - Guesswork $\mathscr{G}$
    - ReID AUC
        - ROC AUC of the attacker E, when viewed as binary classification
- We sample (10k examples, a T), evaluate $\mathscr{G}$ and ReID AUC repeat 100 times

i.e. can we securely release unlabeled data?

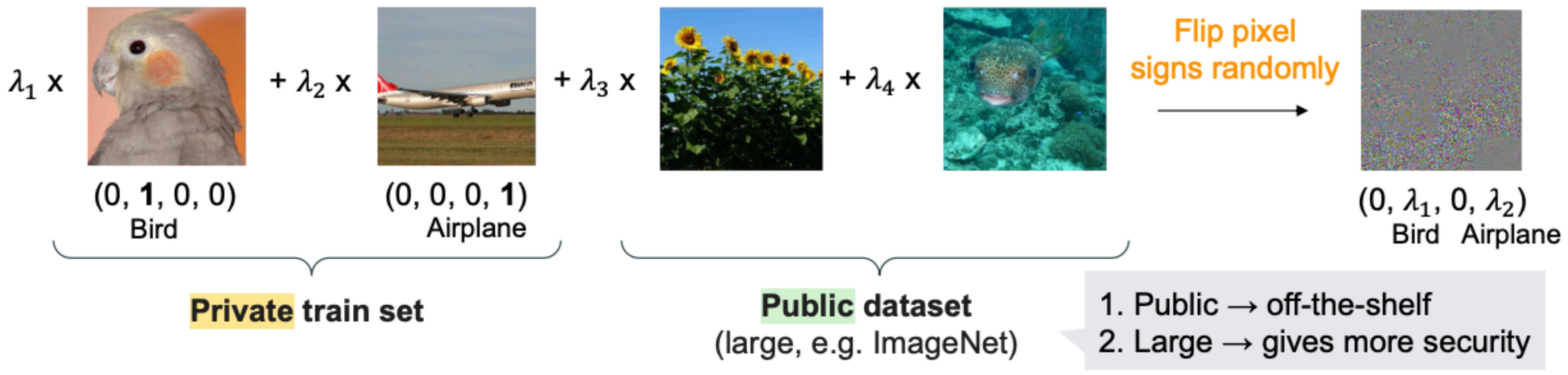
Experiments - Baselines

Syfer-random ablation where the obfuscator layers are not trained

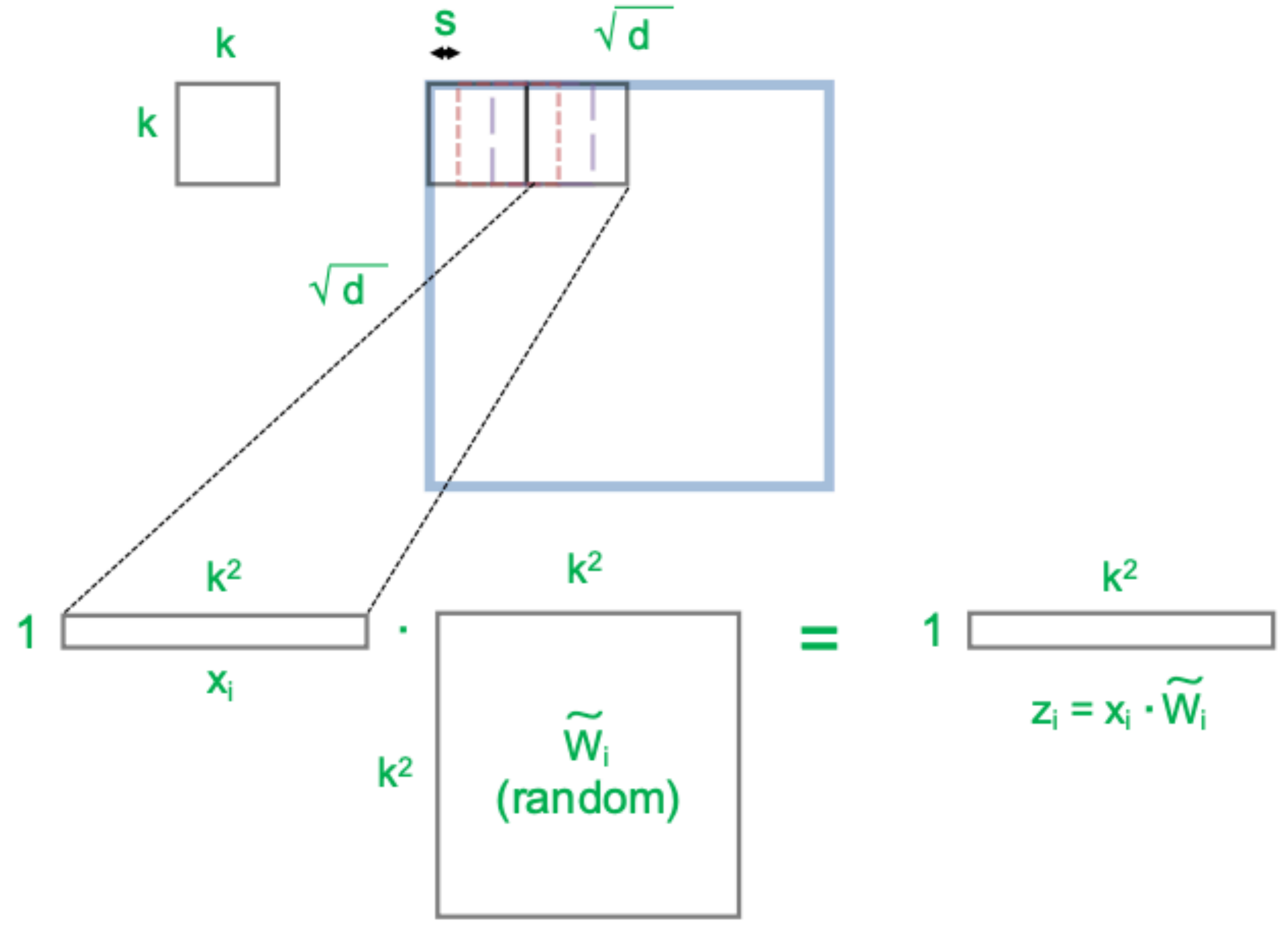


Experiments - Lightweight encoding baselines

InstaHide [Huang et al, 2020], linear image mixing with bit flip

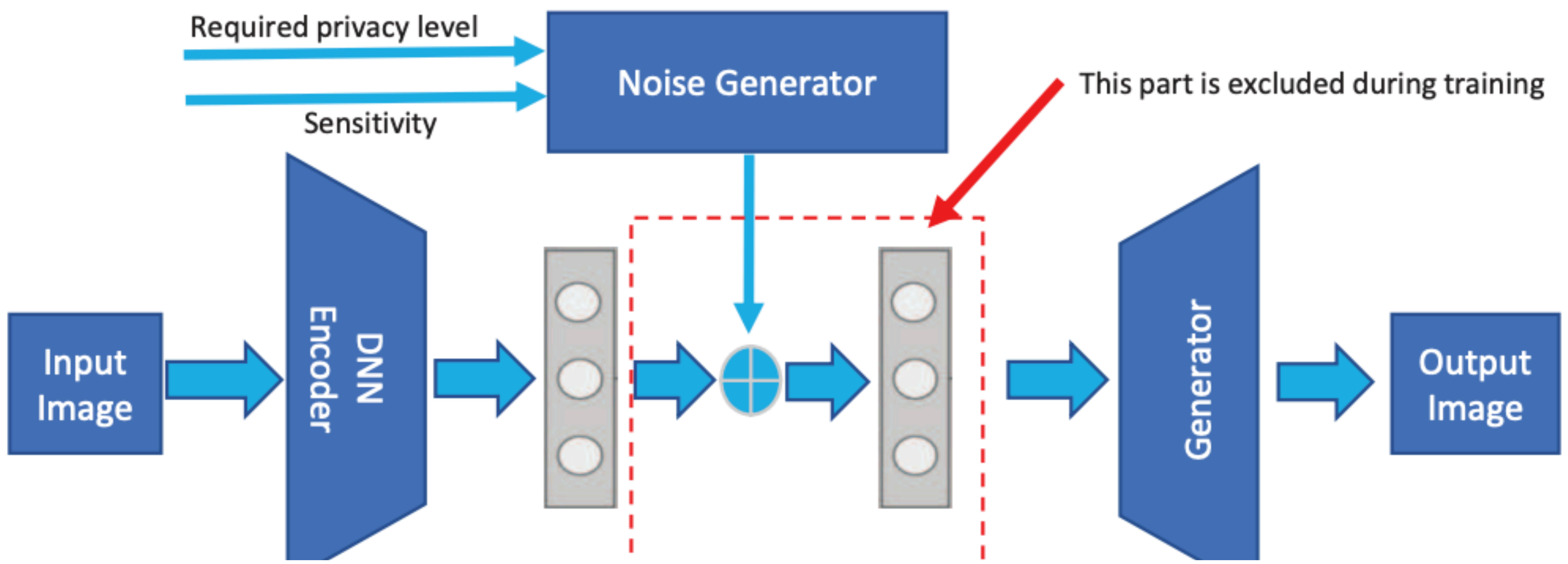


Dauntless [Xiao et al, 2021]**,** separate linear layer applied to each patch. Provably secure if assume X is Gaussian



Experiments - Diff privacy baselines

DP-Image [Lui et al 2021], Differential Privacy Methods on auto encoder.
Add laplacian noise to latent space



Experiments - Privacy Evaluation

Generalized Privacy (no label released)

	Guesswork ↑	ReID AUC ↓
Dauntless	1	100
InstaHide	1	100
DP-Image b =1	3	89
DP-Image b = 5	1379	73
Syfer-Rand	2	99
Syfer (w/o label encoding)	**8476**	**50**

Experiments - Privacy Evaluation

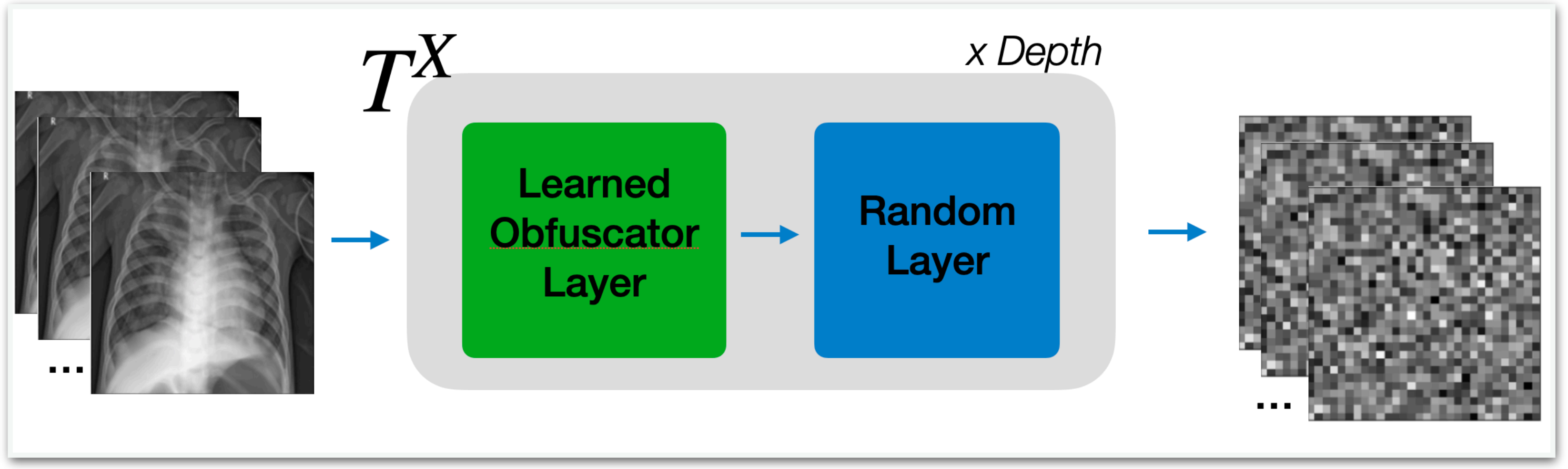
Syfer Privacy across attacker architectures

	Guesswork ↑	ReID AUC ↓
SAU	8477	50
ViT	8411	50
ResNet-18	10070	89

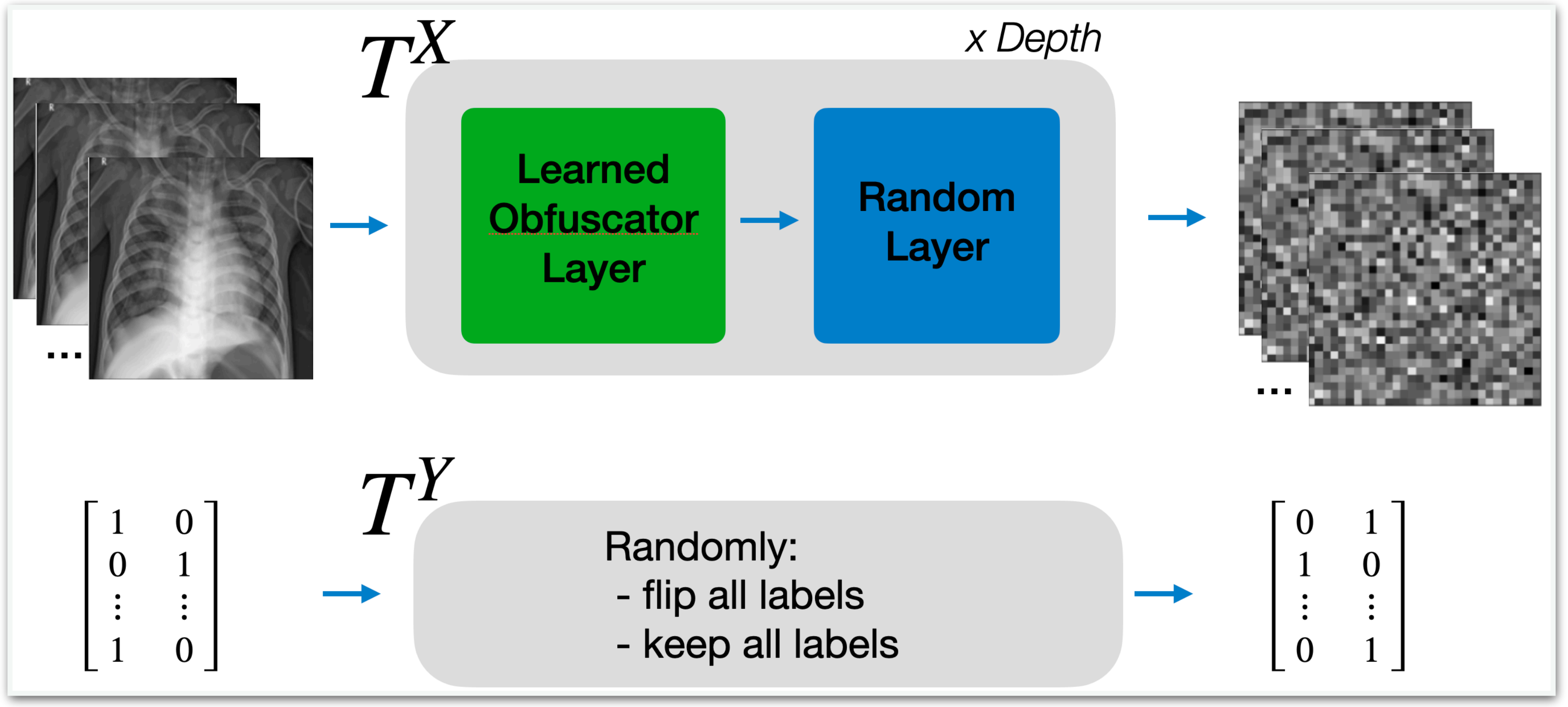
Syfer maintains privacy across **heldout datasets**, **heldout attackers.**

Experiments - Privacy evaluation

- Now, we release the data with labels
- **Privacy with labels:** How secure is $(Z, Y) = (T^X(X), T^Y(LF(X))$?
- Privacy can only get worse (non-private schemes remain non-private)



Syfer w/o label encoding



Syfer

Experiments - Privacy evaluation

Syfer Privacy when released with labels Edema, Atelectasis, Cardiomegaly, Consilidation

	Guesswork ↑	ReID AUC ↓
Edema	3617	50
Actel	1697	55
Cons	9834	51
Cardio	13189	50

Ablation:

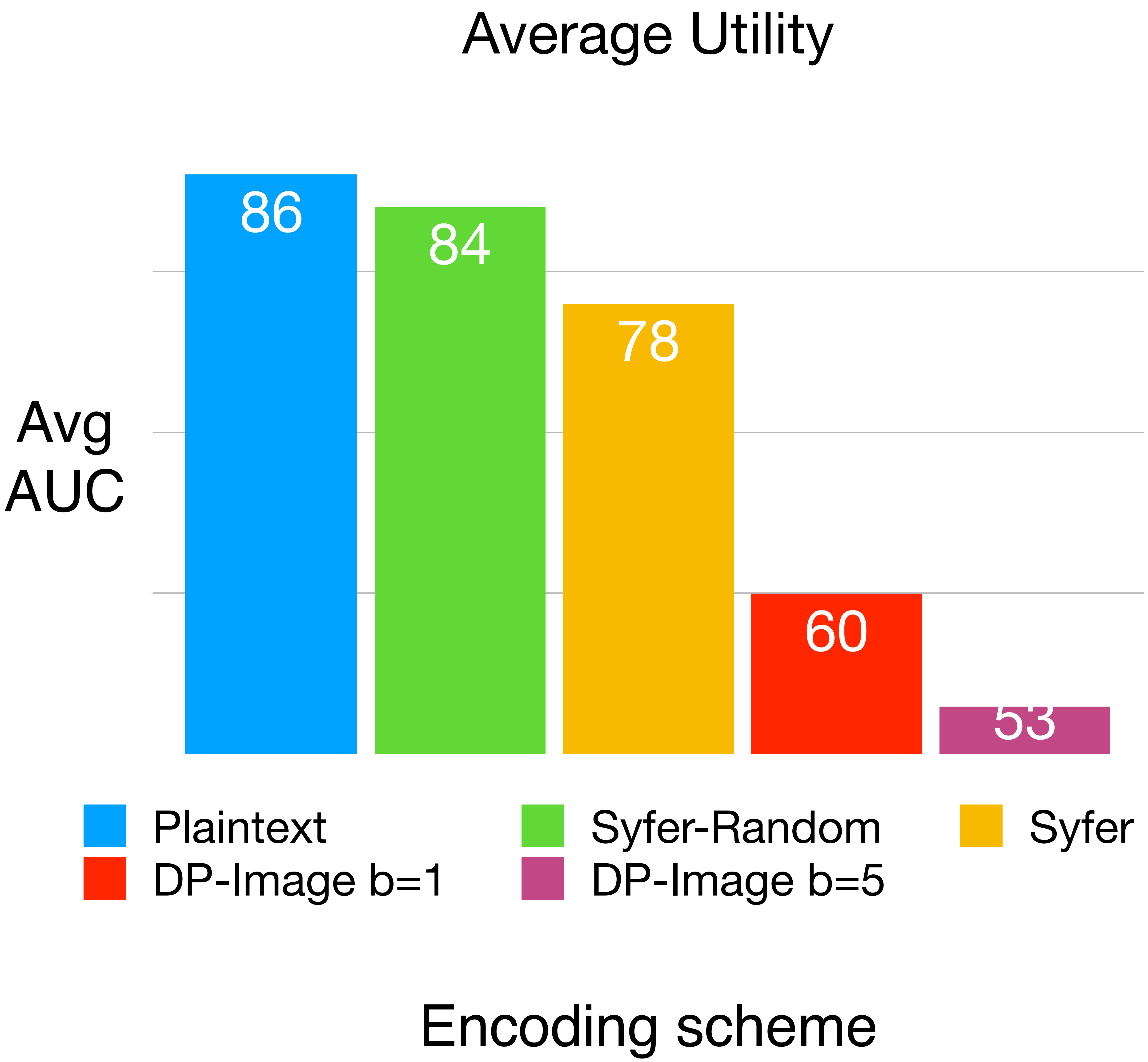
Syfer with no label encoding

$T^{Y}(l) = l$

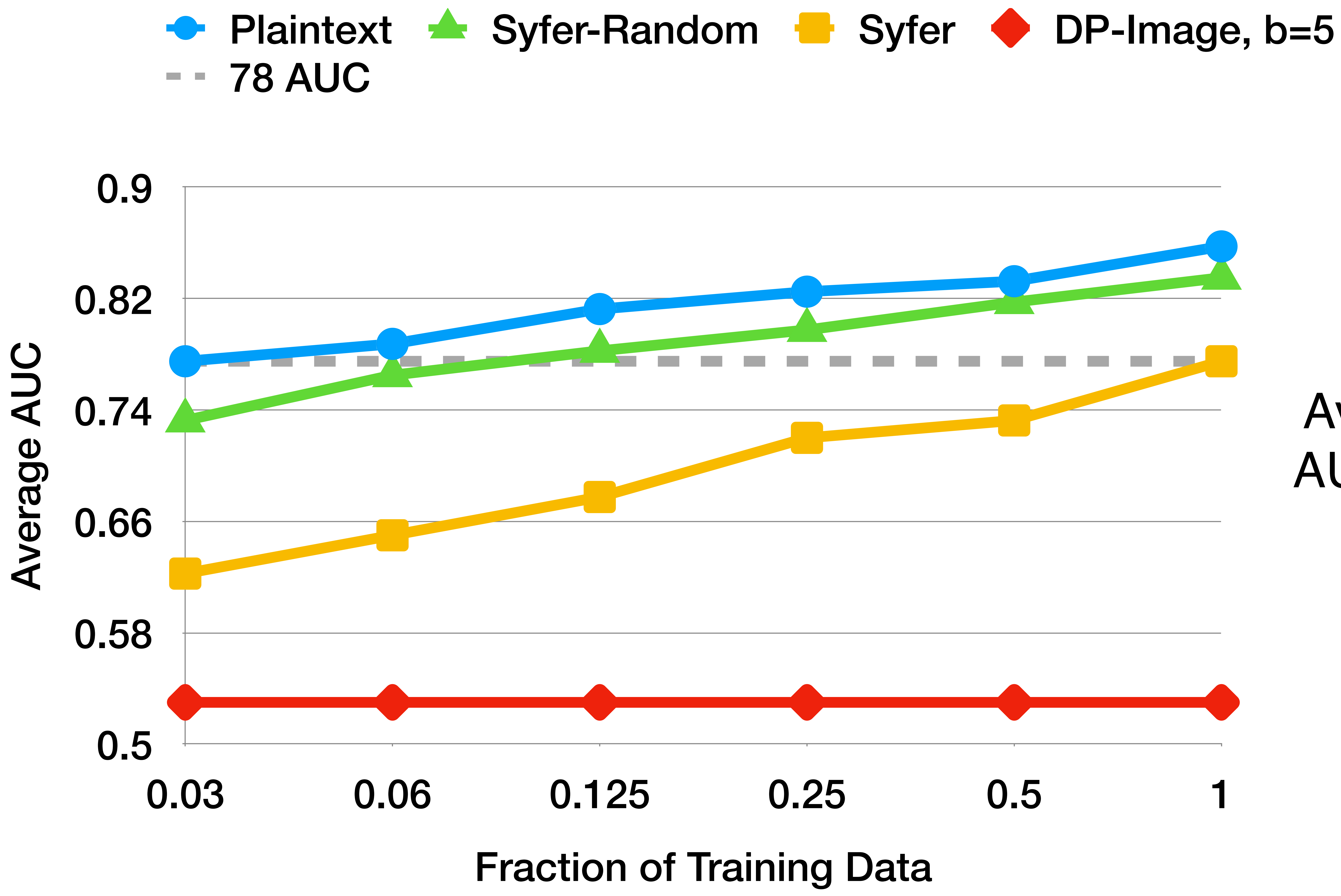
	Guesswork ↑	ReID AUC ↓
Edema	47	76
Actel	36	76
Cons	42	75
Cardio	80	75

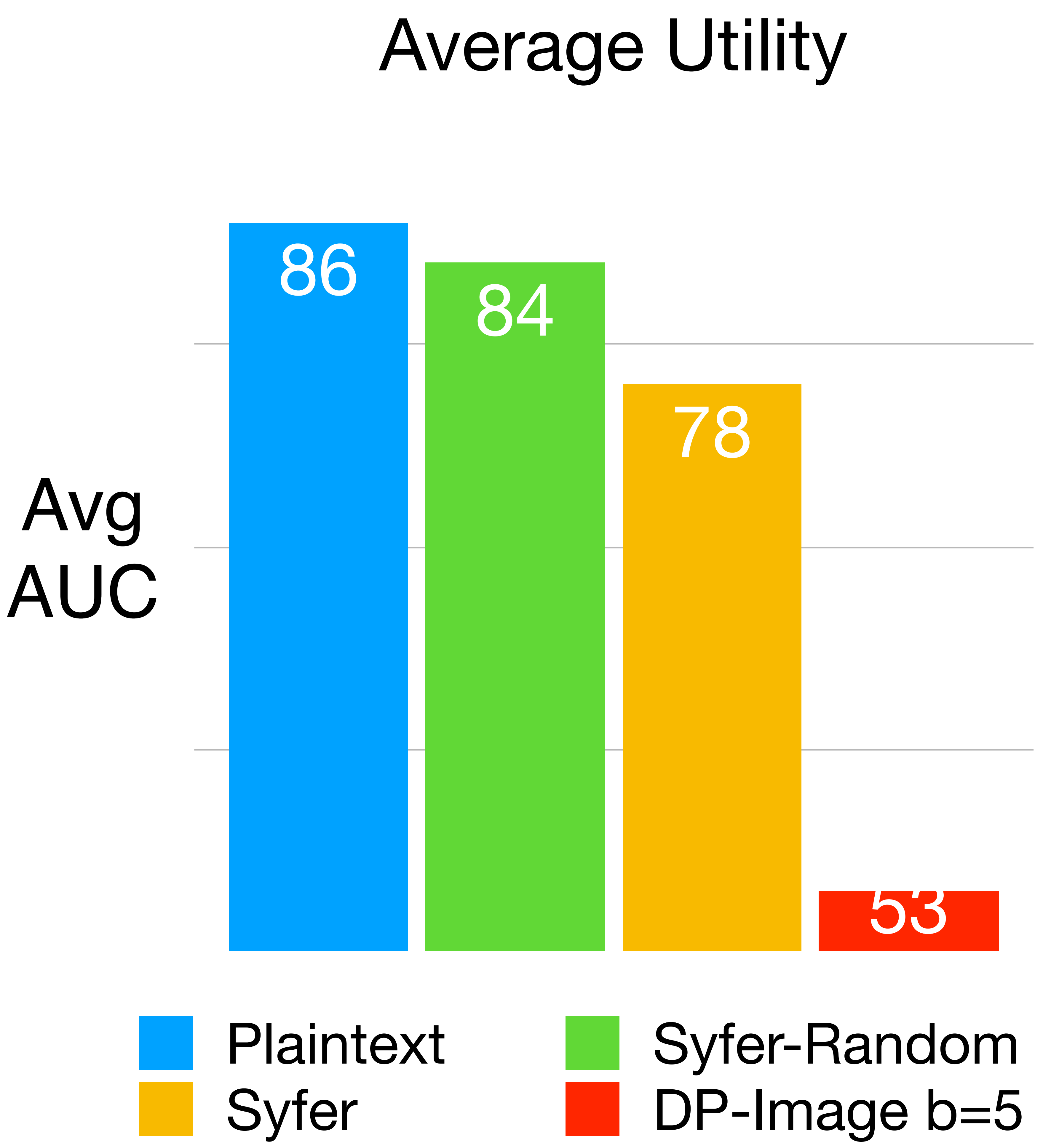
Experiments - Utility evaluation

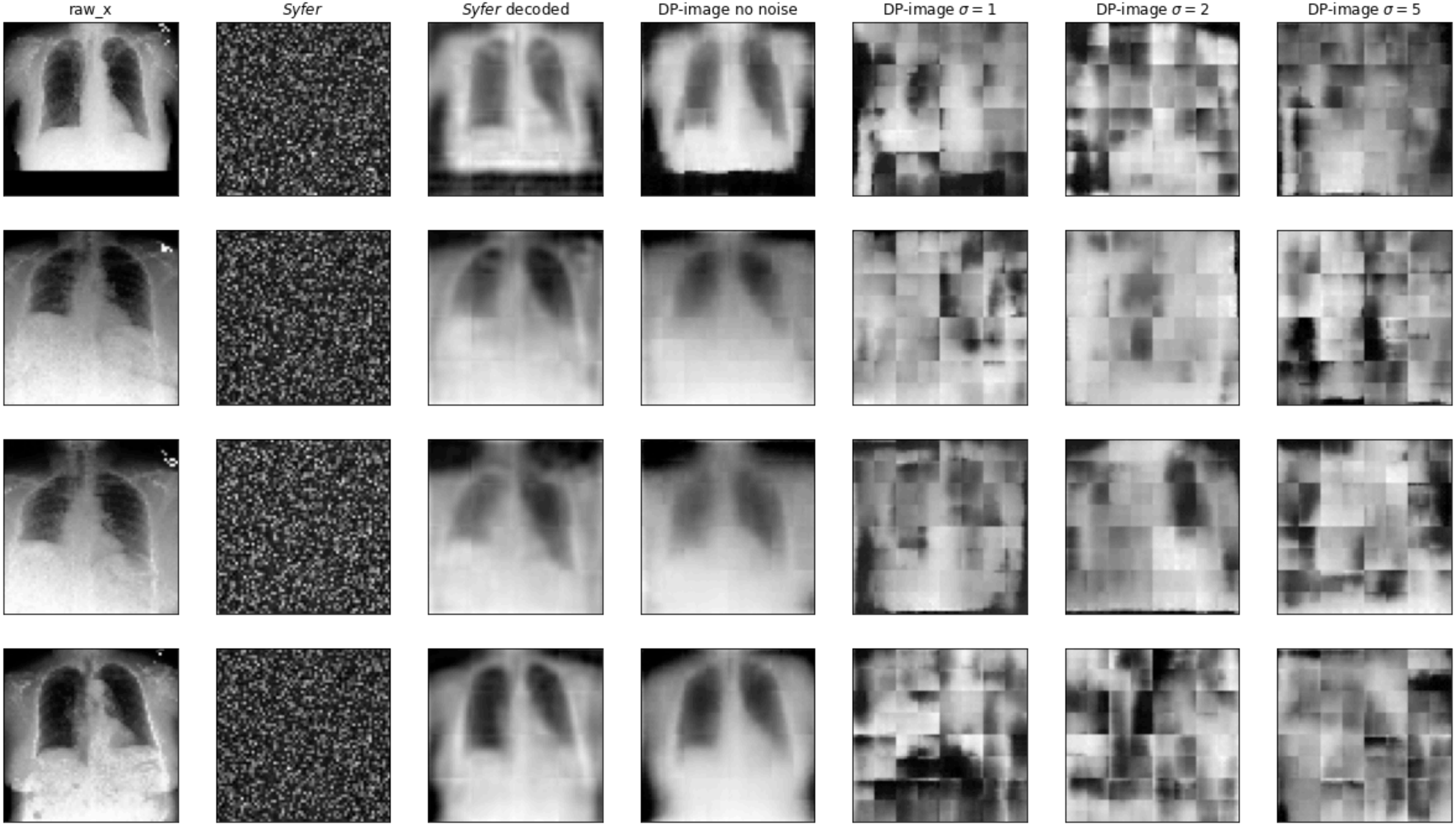
- **Utility Evaluation**:
    - ROC AUC of classifiers trained on encoded MIMC data
- Achieves much better utility than DP-Image
    - **+25** points AUC relative to DP
    - **- 8** points relative to plaintext baseline
    - **- 6** points relative to random Syfer baseline
- How does it impact sample complexity?



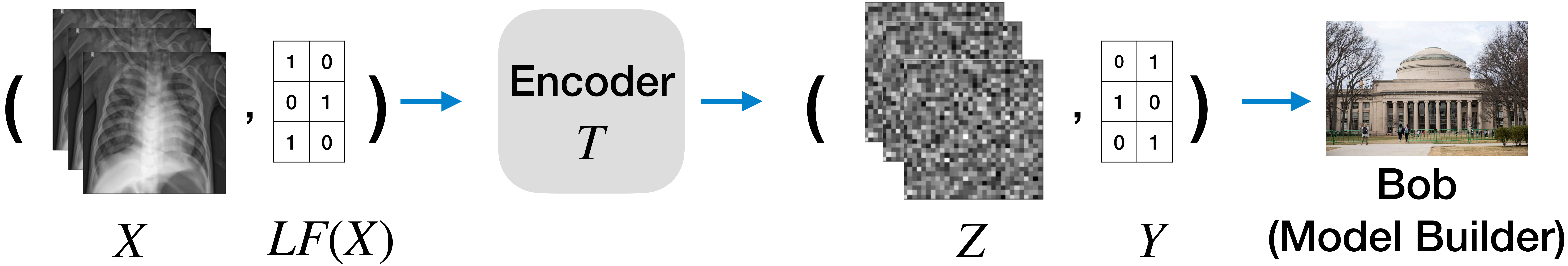
Experiments - Utility evaluation





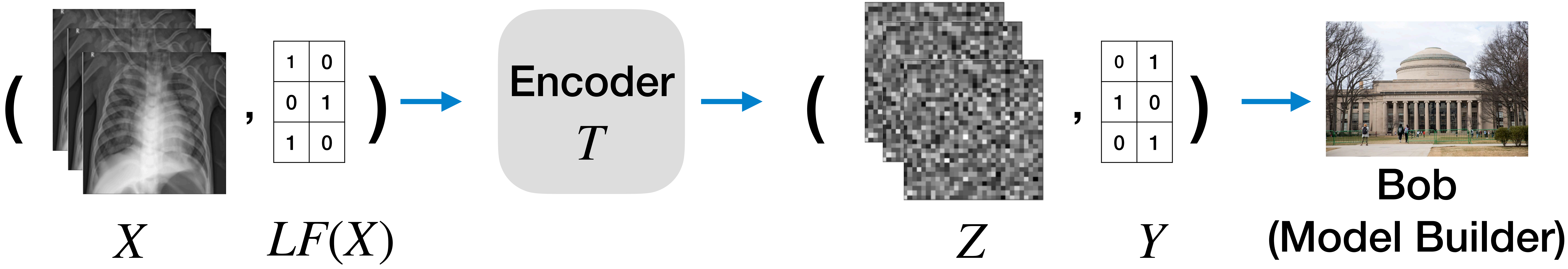

raw_x
Syfer
Syfer decoded
DP-image no noise
DP-image $\sigma = 1$
DP-image $\sigma = 2$
DP-image $\sigma = 5$

Takeaways



- New direction of private ML based on preconditioning random networks
- **Properties:**
  - Protect raw data identity (HIPAA), i.e. **achieve high guesswork**
  - Support **any downstream classification task** with standard ML tools
  - Data owner **does not train** Syfer. Syfer trained on X_{public}

Takeaways

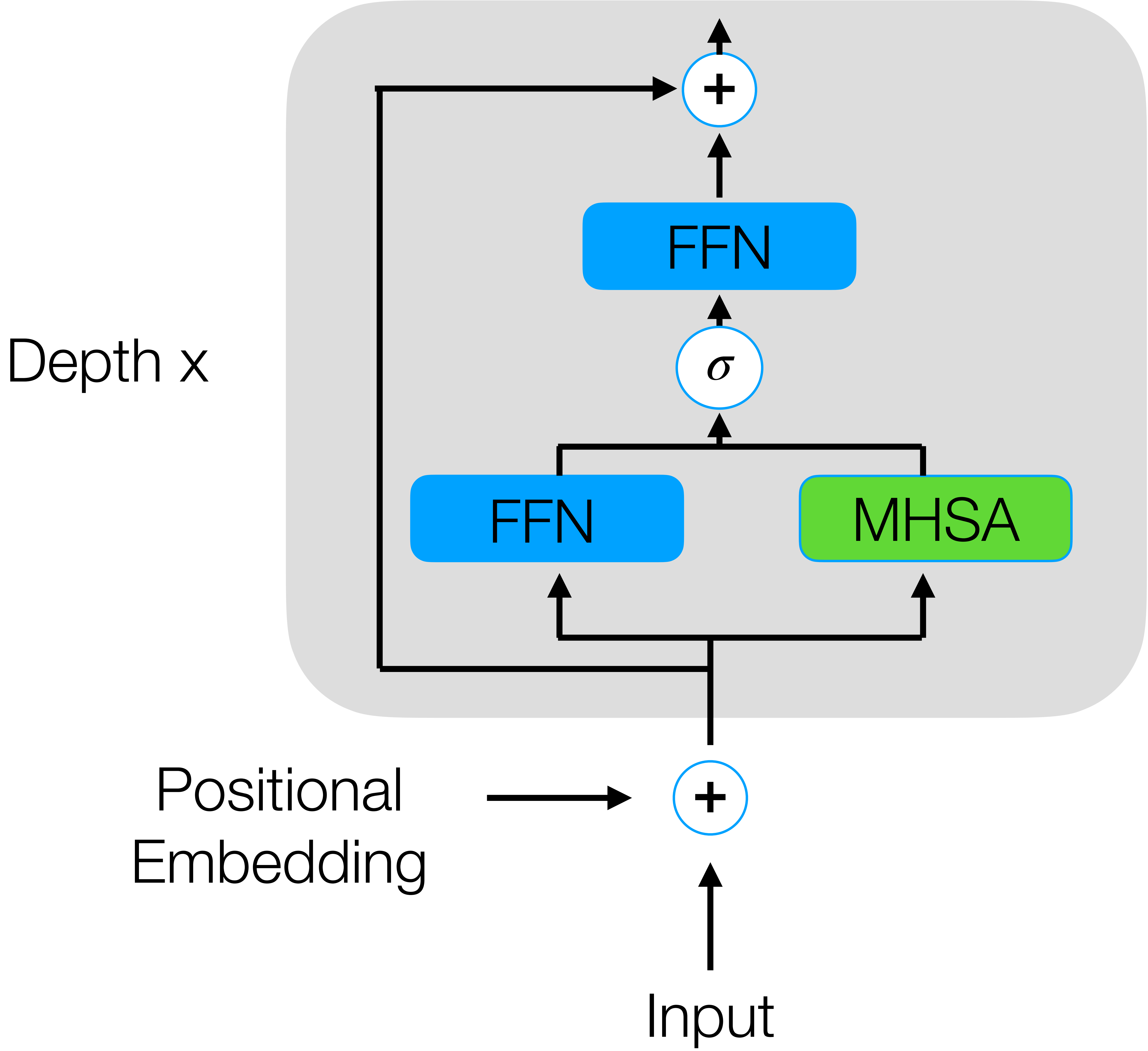


- New direction of private ML based on preconditioning random networks
- **Future work:**
  - Improved architectures + training can further improve utility
  - Support multi-hospital training
  - Applications to other modalities

Appendix Slides

SAU: Simple Attention Unit

- Attention based layer
- Interpolate with learnable gate between:
  - FFN
  - Multi-head self attention (MHSA)
- Empirically more stable than transformers



Encoding	Guesswork	ReId AUC
Dauntless	1 (1,1)	100 (100, 100)
InstaHide	1 (1,1)	100 (100, 100)
DP-S, $b = 10$	1 (1, 2)	98 (98, 98)
DP-S, $b = 20$	4 (1, 14)	86 (85, 86)
DP-S, $b = 30$	68 (2, 189)	70 (70, 70)
DP-I, $b = 1$	3 (1, 8)	89 (88, 89)
DP-I, $b = 3$	97 (7, 296)	73 (73, 73)
DP-I, $b = 5$	1379 (49, 4135)	59 (59, 60)
Syfer-Random	2 (1, 4)	99 (99, 99)
Syfer	8476 (1971, 20225)	50 (49, 52)

Table 1. Privacy evaluation of different encoding schemes against an SAU based attacker on the unlabeled MIMIC-CXR training set. DP-S and DP-I stand for DP-Simple and DP-Image respectively. All metrics are followed by 95% confidence intervals.

Attacker	Guesswork	ReId AUC
SAU	8476 (1971, 20225)	50 (49, 52)
ViT	8411 (5219, 12033)	50 (49, 51)
Resnet-18	10070 (9871, 10300)	50 (47, 53)

Table 2. Privacy evaluation of for *Syfer* across different attacker architectures on the unlabeled MIMIC-CXR training set. All metrics are followed by 95% confidence intervals.

Diagnosis	Guesswork	ReId AUC
Syfer		
Edema	3617 (94, 11544)	50 (49, 51)
Consolidation	1697 (83, 5297)	55 (53, 57)
Cardiomegaly	9834 (2072, 15766)	51 (49, 53)
Atelectasis	13189 (2511, 28171)	50 (48, 52)
Ablation: *Syfer* with no label encoding ($T^Y(l) = l$)		
Edema	47 (12, 83)	76 (76, 76)
Consolidation	36 (2, 104)	76 (76, 76)
Cardiomegaly	42 (17, 57)	75 (75, 75)
Atelectasis	80 (65, 98)	75 (75, 75)

Table 3. Privacy evaluation of *Syfer* when released with different diagnoses in MIMIC-CXR training set. All metrics are followed by 95% confidence intervals.

Dataset	Train	Dev	Test
Unlabeled			
NIH	40365	NA	NA
MIMIC-CXR	57696	NA	NA
Labeled			
MIMIC-CXR E	3660	1182	12125
MIMIC-CXR Co	1120	375	11031
MIMIC-CXR Ca	11724	3876	12791
MIMIC-CXR A	2164	3992	12129

Table 5. Dataset statistics for all datasets. The training and development sets of MIMIC CXR Edema, Consolidation, Cardiomegaly

Encoding	E	Co	Ca	A	*Avg*
Plaintext	91	78	89	85	86
DP-Simple, $b = 10$	51	51	52	52	52
DP-Simple, $b = 20$	50	50	50	50	50
DP-Simple, $b = 30$	49	49	50	51	50
DP-Image, $b = 1$	60	59	60	59	60
DP-Image, $b = 2$	54	50	55	55	54
DP-Image, $b = 5$	53	55	51	52	53
Syfer-Random	89	75	86	84	84
Syfer	82	69	81	78	78

Table 4. Impact of *Syfer* on chest X-ray prediction tasks across different encoding schemes. All metrics are ROC AUCs across the MIMIC-CXR test set. Guides of abbreviations for medical diagnosis: (E)dema, (Co)nsolidation, (Ca)rdiomegaly and (A)telectasis.