Generating Synthetic Healthcare Records Using Generative Adversarial Networks

Amirsina Torfi and Mohammadreza Beyki

Virginia Tech, Department of Computer Science
Blacksburg, VA, 24061

Final Project Presentation (CS 6604 - Digital Libraries)

Course Instructor:
Dr. Edward A. Fox

5 December 2019

Motivation

- Electronic Health Records (EHRs) & Big Data in healthcare → Calls for employing data-driven methods with *Artificial Intelligence* (AI)
- De-identification of EHR data employed for mitigation of privacy risks → NOT SECURE! [1, 2, 3]
- Need for synthetic healthcare records for Machine Learning

- Introduction
- **Specific Aims**
- Background
- Accomplishments
- References

Aim 1: Develop the Generative Model

- Capturing spatial-temporal information
- Handling discrete data
- Evaluation of synthetic data quality using **statistical** analysis

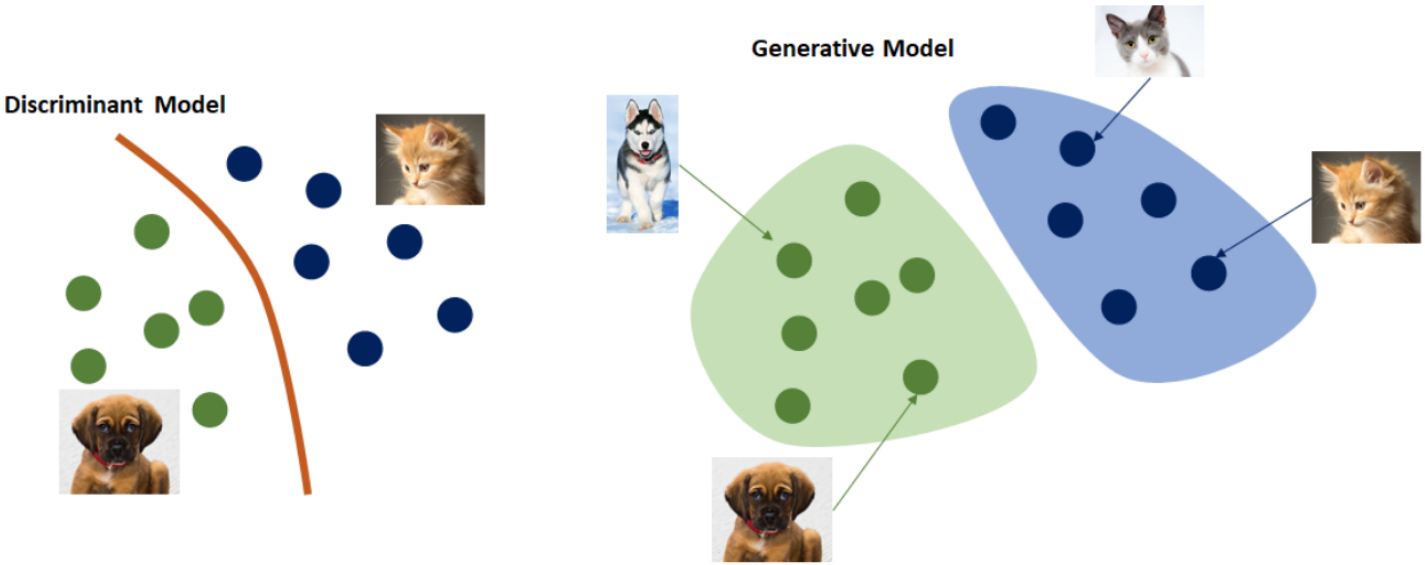


Figure 1: Generative/Discriminative Models [Link]

Aim 1: Develop the Generative Model

- **Hypothesis 1:** Generative Adversarial Networks (GANs) perform better than other generative models.
- **Hypothesis 2:** Convolutional Neural Networks (CNNs) outperform Multilayer Perceptrons $\rightarrow$ *Capturing and integrating more temporal and spatial information from healthcare records*

Aim 2: Measuring Realistic Characteristics

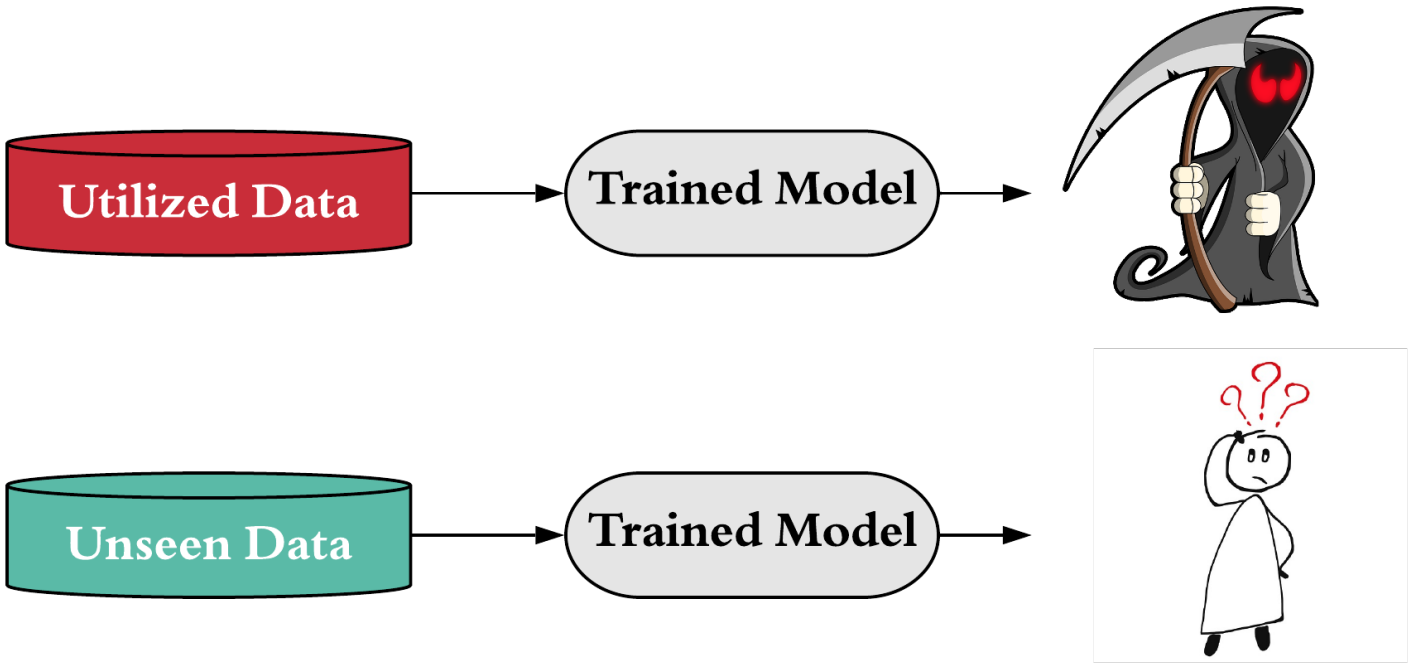
- Propose a discriminative model to measure the realistic characteristics of the data (**unique contribution**)
- Use machine learning instead of statistics
- Can we replace real data with synthetic data?

Aim 3: Privacy

- Assess privacy by **Membership Inference Attack**

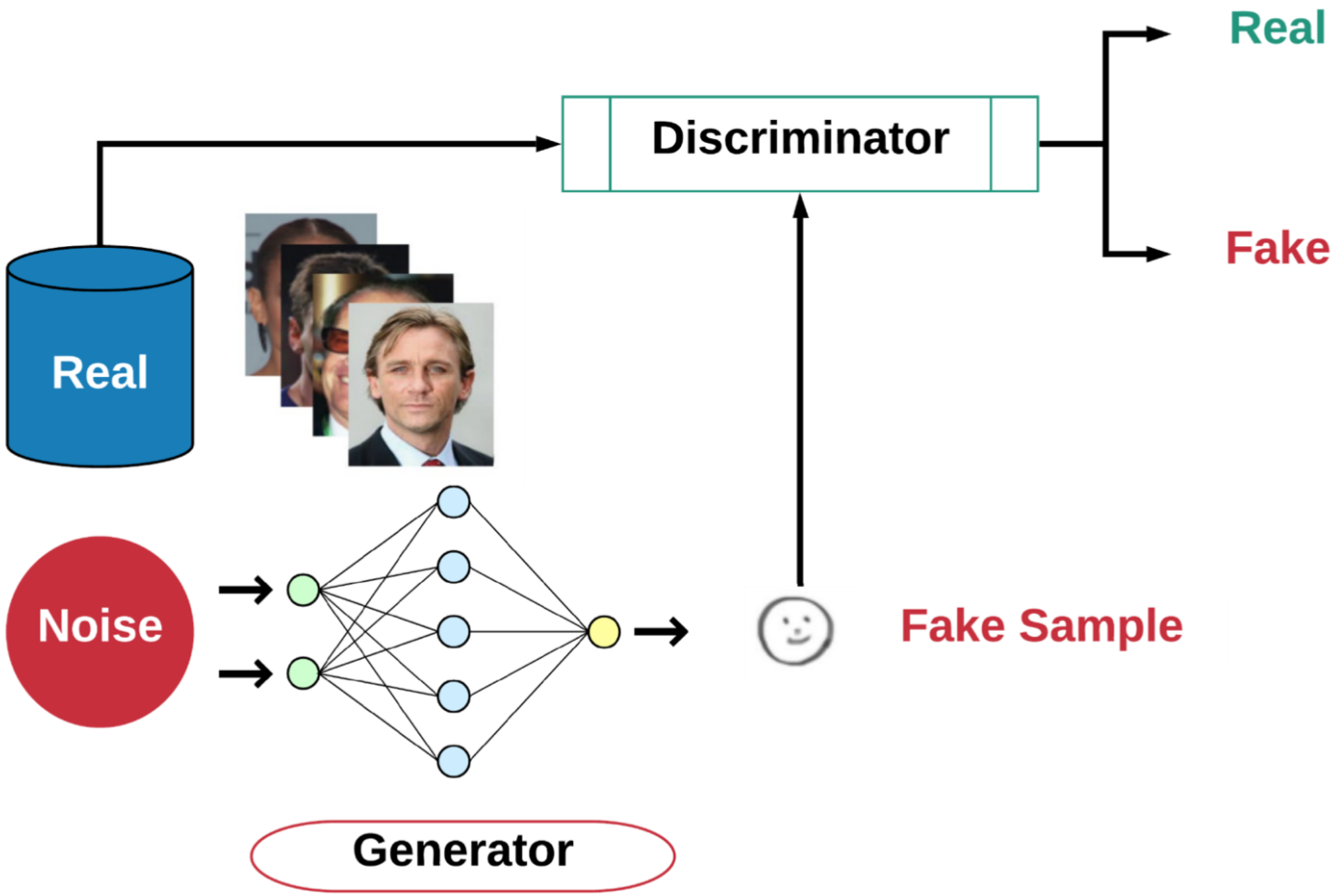
Aim 3: Privacy

- **Hypothesis:** Machine Learning models responding differently to data they saw or never saw in training



- Introduction
- Specific Aims
- **Background**
- Accomplishments
- References

Generative Adversarial Networks



Power of GANs

Figure 2: Example of Fake images [4]

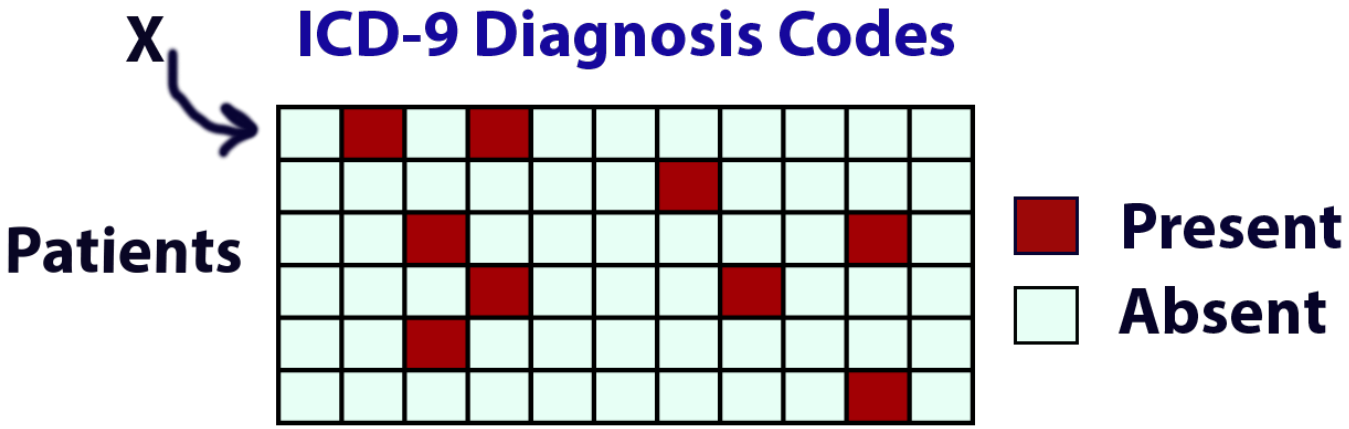
- Introduction
- Specific Aims
- Background
- **Accomplishments**
- References

Accomplished Goals

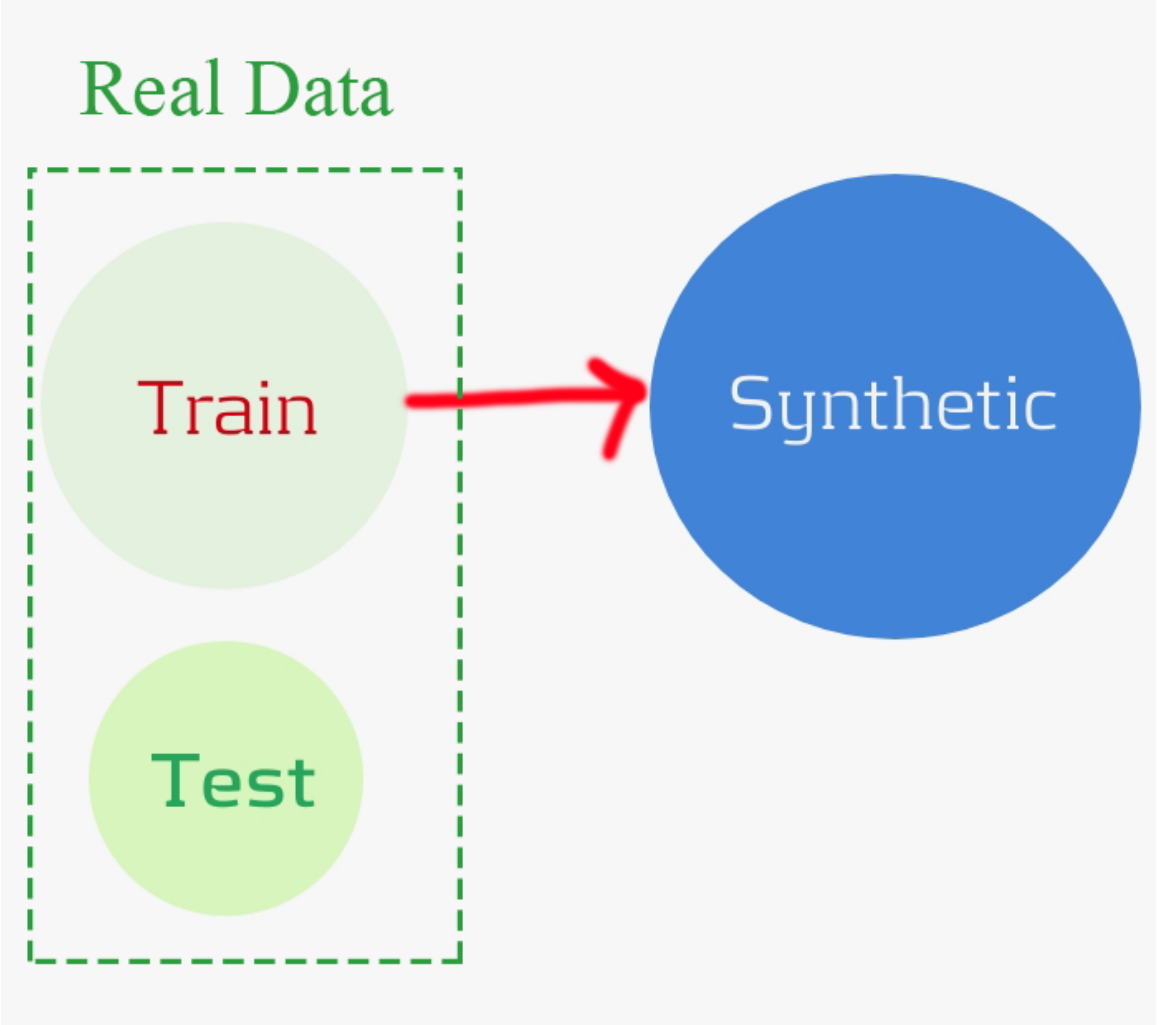
- Proposed an efficient architecture to generate synthetic healthcare records using Convolutional GANs and *Convolutional Autoencoders* $\rightarrow$ *"COR-GAN"*
- The effectiveness of utilizing Convolutional Neural Networks (CNNs) is proved empirically $\rightarrow$ Capturing inter-correlation between features
- Privacy is assessed $\rightarrow$ Membership Inference Attack

EHR data

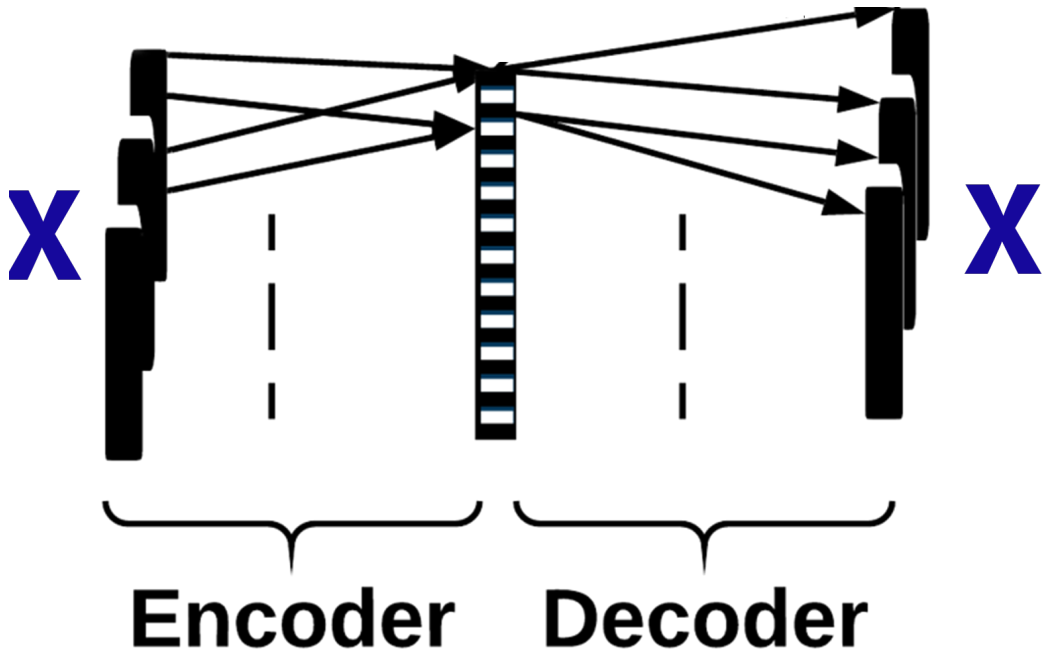
- There are $|\mathcal{M}|$ discrete variables (e.g., diagnosis, medication, or procedure codes)
- **EHR data of a particular patient:** A fixed-size vector $\mathbf{X} \in \mathbb{Z}_+^{|\mathcal{M}|}$, $\mathbb{Z}_+ = 0, 1, 2, ...$
- The i^{th} dimension $\rightarrow$ Number of occurrences (*i.e.*, counts) of i-th variable in patient record
- **Binary** representation $\mathbf{X} \in \{0, 1\}^{|\mathcal{M}|} \rightarrow i^{th}$ dimension indicates absence or occurrence of i^{th} variable



Train/Test Data



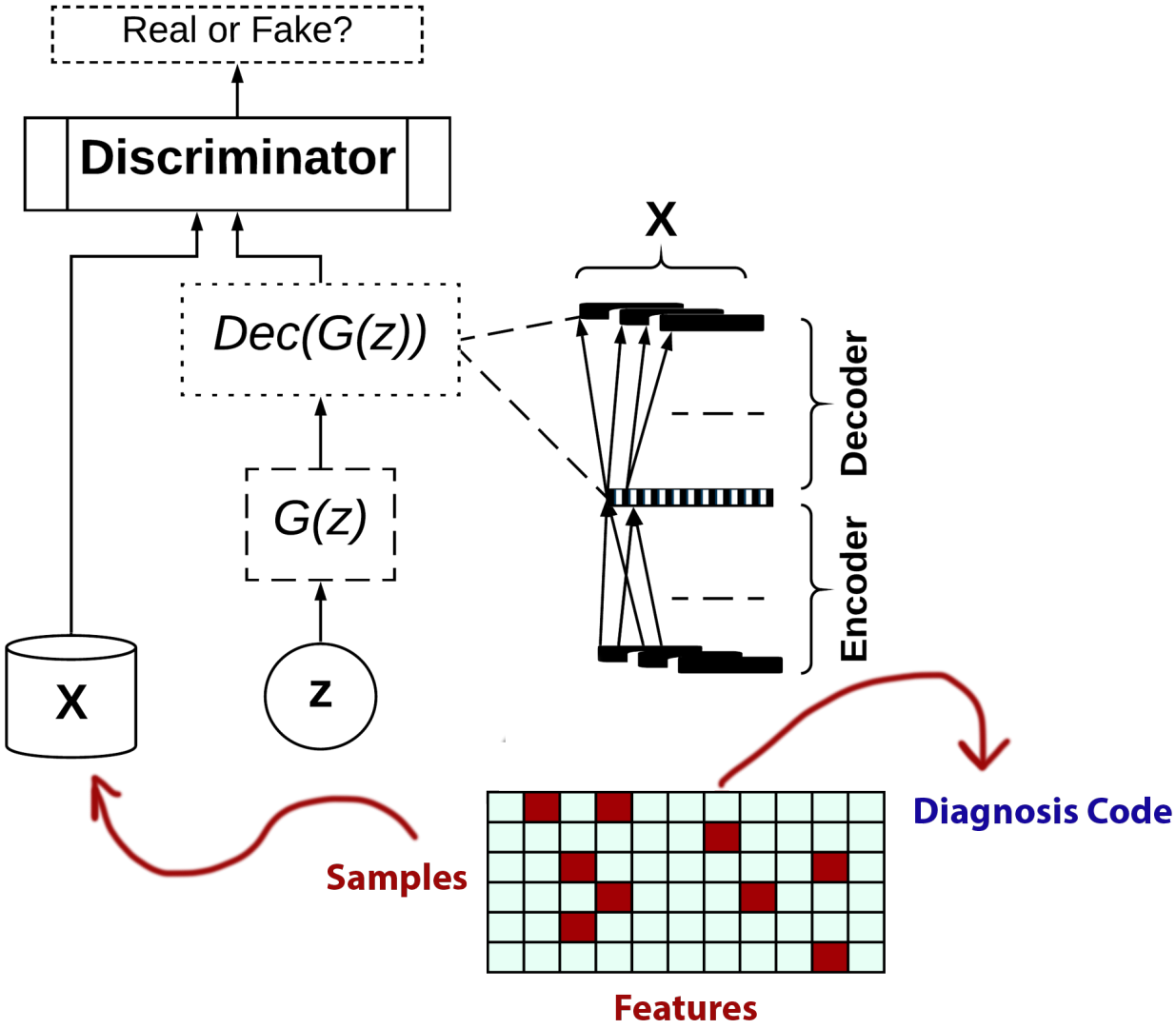
Autoencoder Training



$$\textit{Autoencoder} : BCE_{loss} = -\frac{1}{N}\sum_{i=1}^{N} x_i log(y_i) + (1 - x_i) log(1 - y_i)$$

$$y_i = \boldsymbol{Dec}(\boldsymbol{Enc}(x_i))$$

Proposed Architecture



Dataset

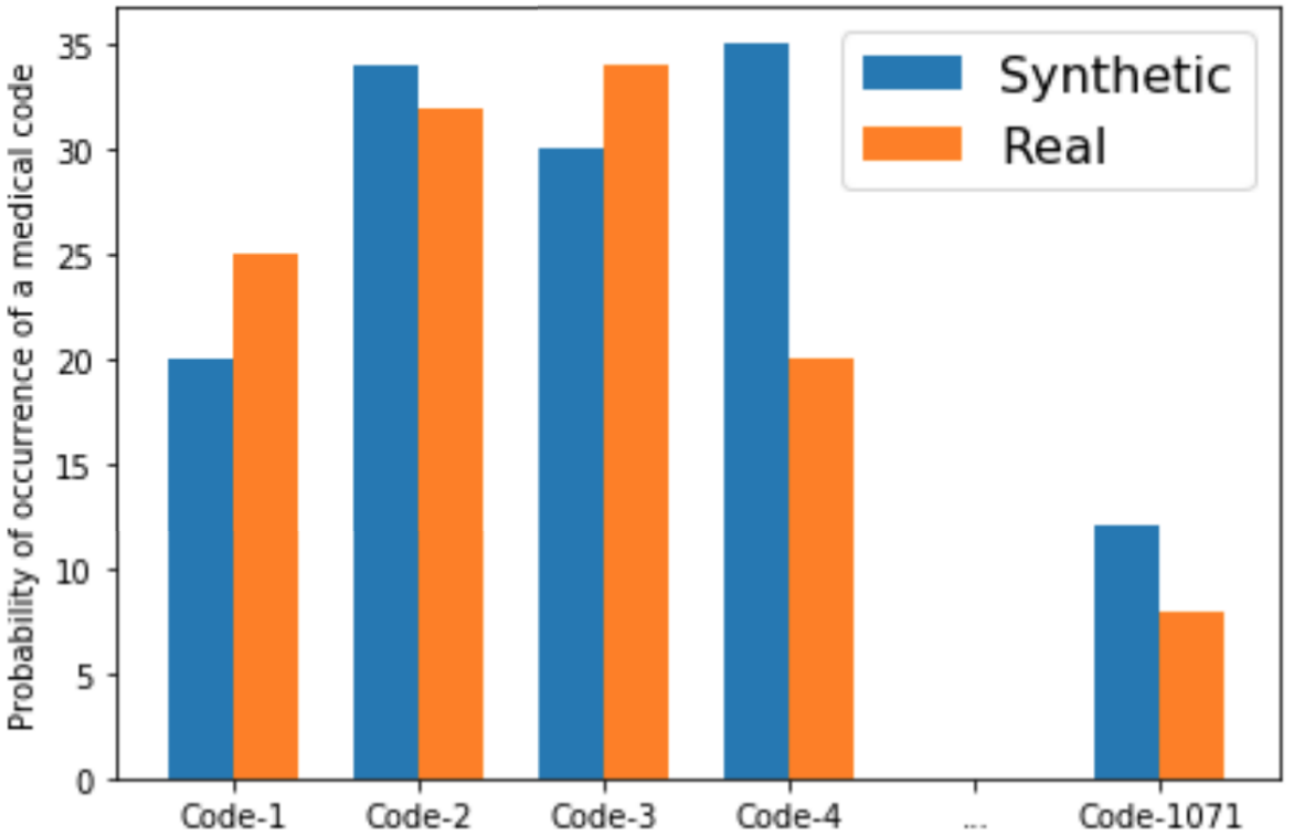
- The *MIMIC-III* dataset [5]
- Medical records of almost 46K patients
- Extracted ICD-9 codes only
- Represent patient records as a fixed-size vector
- 1071 entries for each patient record
- Dataset is used for experiments associated with binary discrete variables

Baseline Models

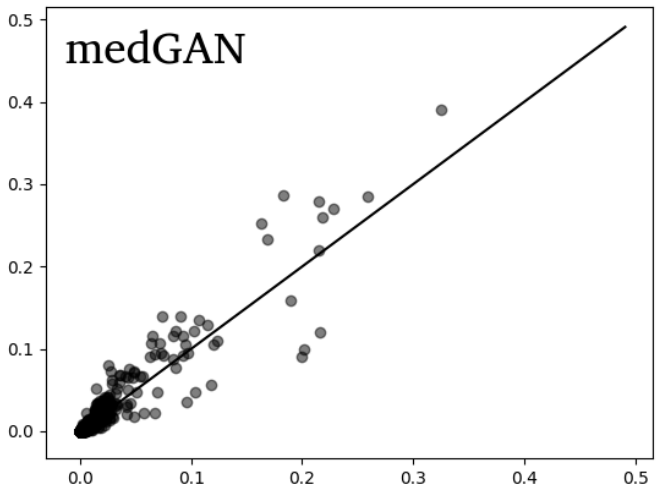
Table 1: Comparison of different baseline architectures.

Name	Decoder (Pretrained)	Generator	Technique
GAN	Autoencoder (NO)	MLP	Regular Training
GAN_{pre}	Autoencoder (YES)	MLP	Regular Training
GAN_{pre}	Autoencoder (YES)	MLP	MD
medGAN [6]	Autoencoder (YES)	MLP	MA + BN
corGan[Ours]	Autoencoder (YES)	1-D CNN	MD + BN

Dimension-Wise Probability



Dimension-Wise Probability



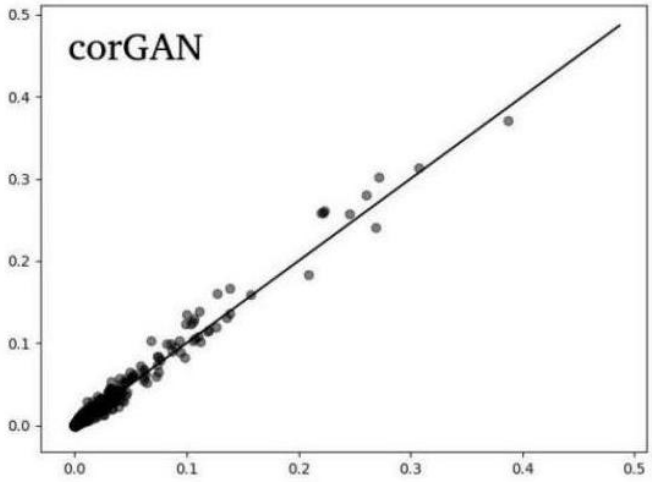


Figure 3: x- and y-axes represent Bernoulli success probability for real and synthetic datasets. Diagonal line shows ideal case.

Discrete Synthetic Data Quality Evaluation

- Maximum Mean Discrepancy
  - Represents similarity between two distributions → Distance between mean feature embeddings
  - Distributions P_R and P_G are defined over set $\mathbb{X}$
  - Used **Kernel MMD**, with isotropic Gaussian
  - For 100 runs

Table 2: Distinguishing between real and synthesized samples by employing Maximum Mean Discrepancy metric.

Name	Score
GAN	0.0064 ± 0.00035
GAN_{pre}	0.0048 ± 0.00022
GAN_{pre+mb}	0.0043 ± 0.00018
medGAN [6]	0.0032 ± 0.00021
corGan [Ours]	$\mathbf{0.0008 \pm 0.00015}$

- Introduction
- Specific Aims
- Background
- Accomplishments
- **References**

[1] V. Janmey and P. L. Elkin, "Re-identification risk in HIPAA de-identified datasets: The MVA attack," in *AMIA Annual Symposium Proceedings*, vol. 2018, p. 1329, American Medical Informatics Association, 2018.

[2] M. Scaiano, S. Korte, A. Baker, G. Green, K. El Emam, and L. Arbuckle, "Re-identification risk measurement estimation of a dataset," Apr. 26 2018.
US Patent App. 15/320,240.

[3] A. Baker, L. Arbuckle, K. El Emam, B. Eze, S. Korte, S. Rose, and C. Ilie, "Method of re-identification risk measurement and suppression on a longitudinal dataset," June 5 2018.
US Patent 9,990,515.

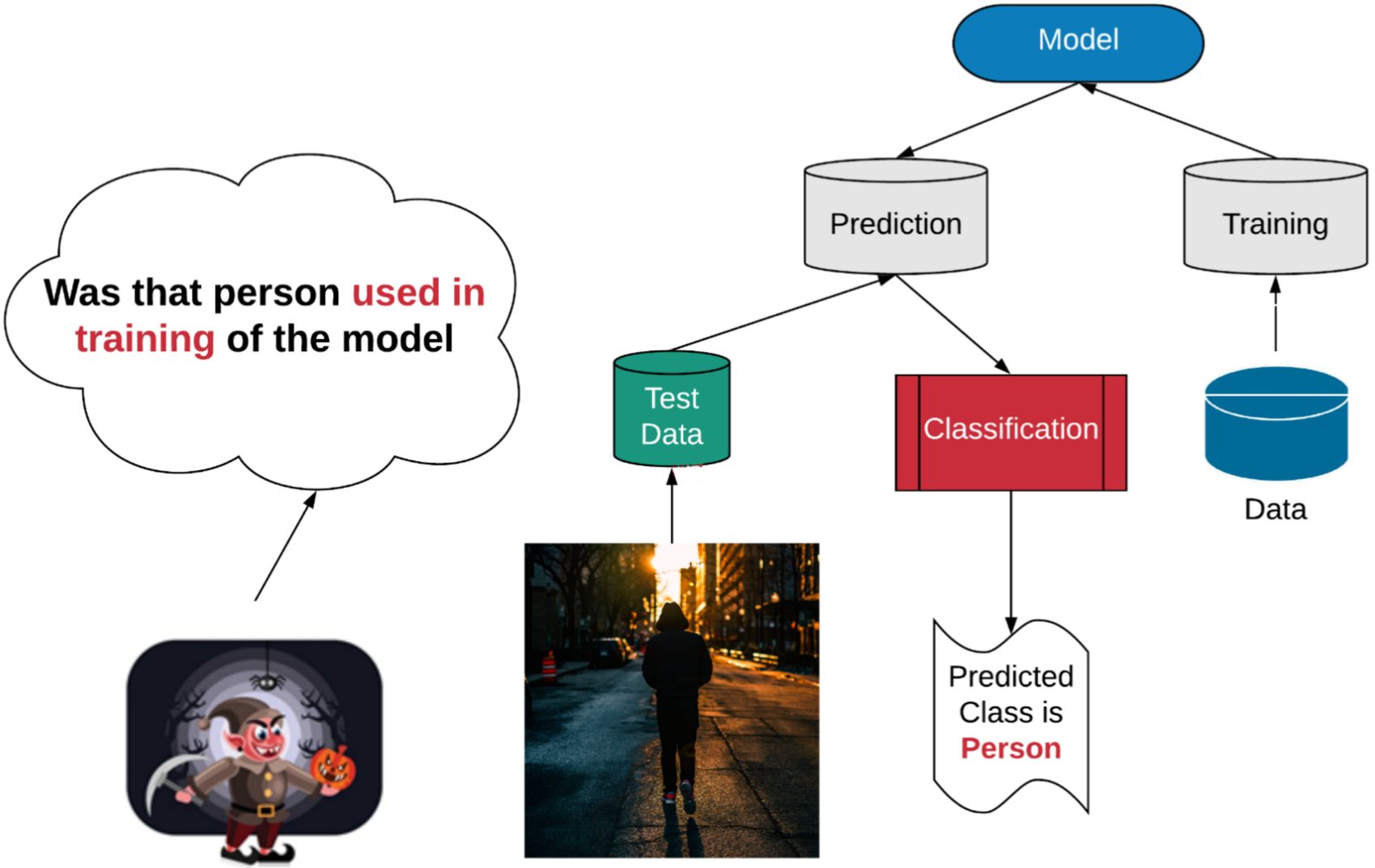
[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

[5] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[6] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," *arXiv preprint arXiv:1703.06490*, 2017.

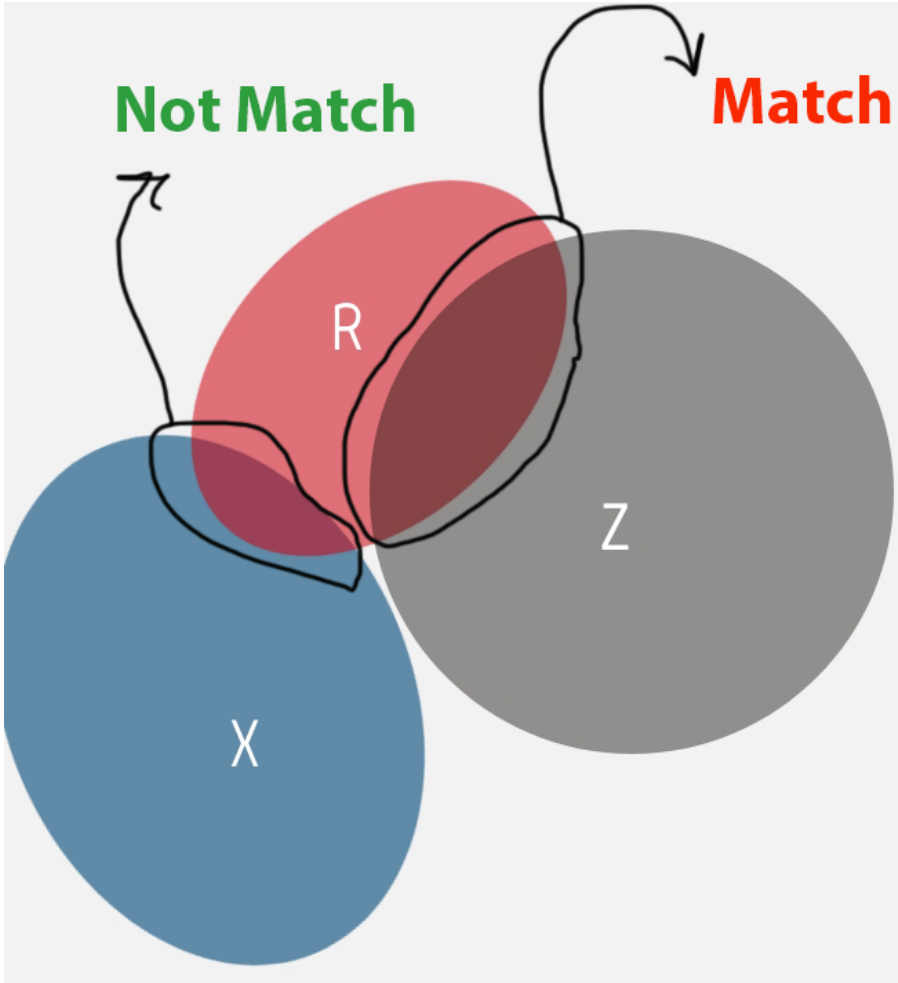

Thank You

Privacy


Was that person used in training of the model
Model
Prediction
Training
Test Data
Classification
Data
Predicted Class is Person

Privacy Assessment

- Pick ($\mathbb{Z}$) and ($\mathbb{X}$) from real training and random data source. Pick ($\mathbb{R}$) from synthetic set.
- Compared each sample in set of $\mathbb{X} + \mathbb{Z}$ with each sample in set of $\mathbb{R}$
- Calculate **Cosine Similarity**
- If similarity is higher than threshold: **Match**



Measure Privacy

- Assessing the **effect of number of records known by attacker** $\rightarrow$ **Assumption:** $|\mathbb{R}| = |\mathbb{X}| = |\mathbb{Z}|$
- **Precision:** For matches identified by adversary, **only a portion of them actually used**
- **Recall:** Adversary has **successfully determined a portion of known records being used in training**

Table 3: $\mathcal{U}$: # of records known to attacker.

$\mathcal{U}$	100	1k	2k	3k	4k	5k
Precision	0.60	0.51	0.41	0.40	0.40	0.39
Recall	0.05	0.10	0.19	0.28	0.27	0.28