Privacy for Free: How does Dataset Condensation Help Privacy?

Tian Dong[1*], Bo Zhao[2], Lingjuan Lyu[3]

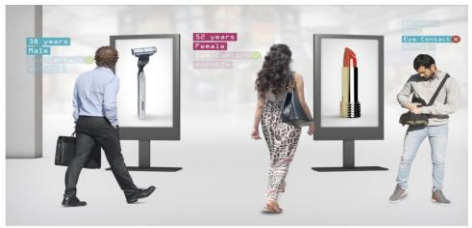
[1] Shanghai Jiao Tong University
[2] The University of Edinburgh
[3] Sony AI
*Work done during internship at Sony AI

Sony AI

SONY

Privacy is important

Personal data is generated everywhere



Smart Retail

Smart City

Social Network

Smart Home

Regulations

GDPR

CALIFORNIA CONSUMER PRIVACY ACT

中华人民共和国
个人信息保护法

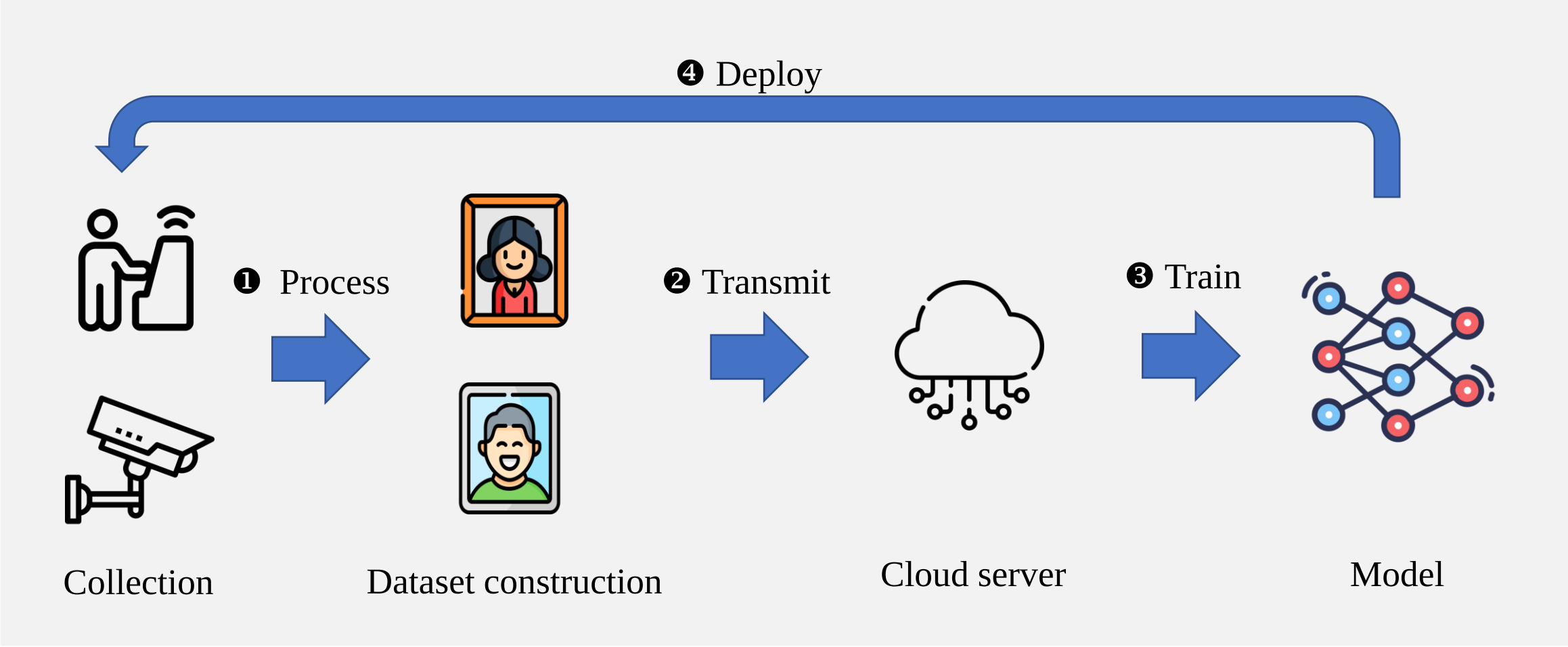
2021年8月20日
第十三届全国人民代表大会常务委员会
第三十次会议通过

COOKIEFIRST

APPI

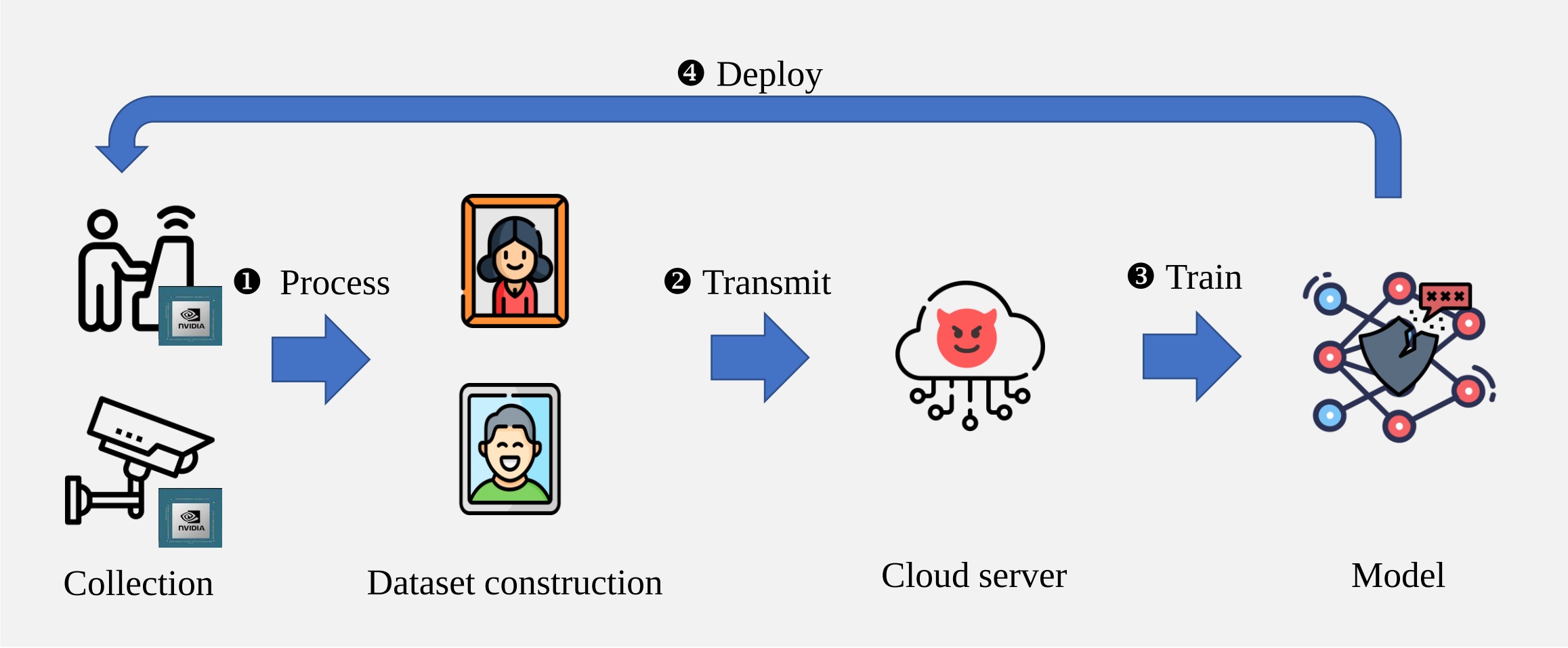
Act on Protection of Personal Information
Japanese Data Protection Law updated 2022

Typical ML pipeline



Machine learning (ML) pipeline

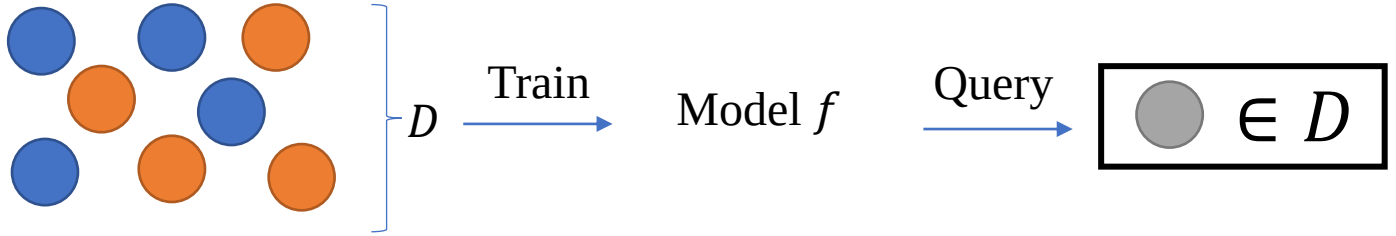
Potential privacy issues



Machine learning pipeline

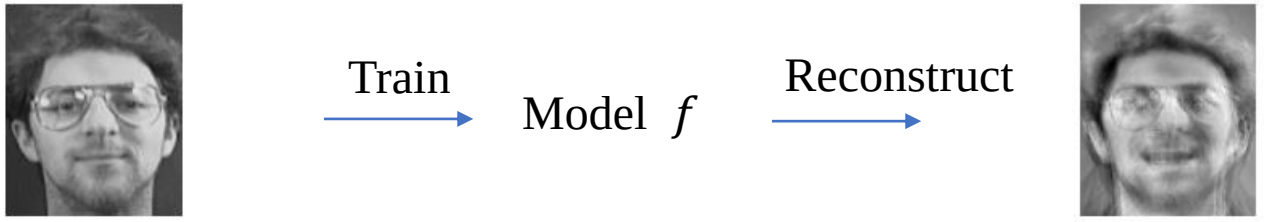
Privacy attacks

Membership Inference Attack (MIA) (Shokri et al., S&P'17, etc.)



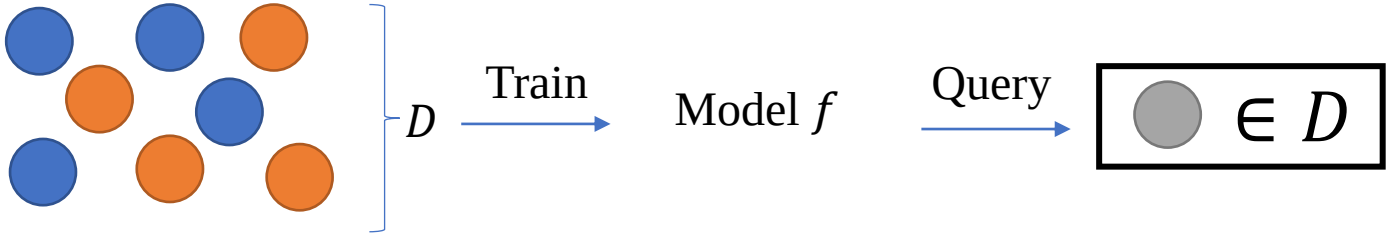
Attribute Inference Attack (Melis et al., S&P'17, etc.)

Model Inversion Attack (Fredrikson et al., CCS'15)



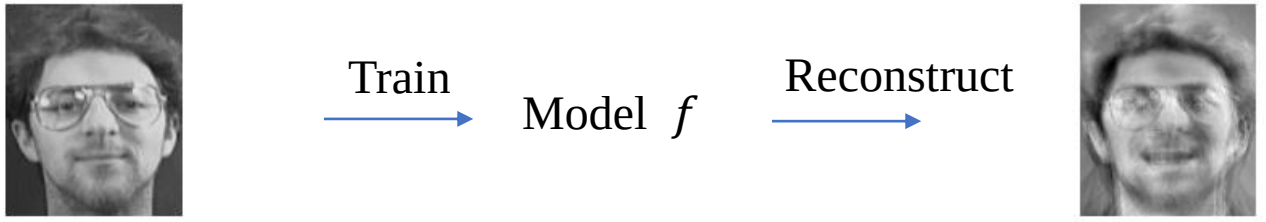
Privacy attacks

Membership Inference Attack (MIA) (Shokri et al., S&P'17, etc.)



Attribute Inference Attack (Melis et al., S&P'17, etc.)

Model Inversion Attack (Fredrikson et al., CCS'15)



Main idea and threat model

Idea: Generate surrogate dataset $\mathcal{S}$ to protect privacy of raw dataset $\mathcal{T}$

Adversary's Goal:

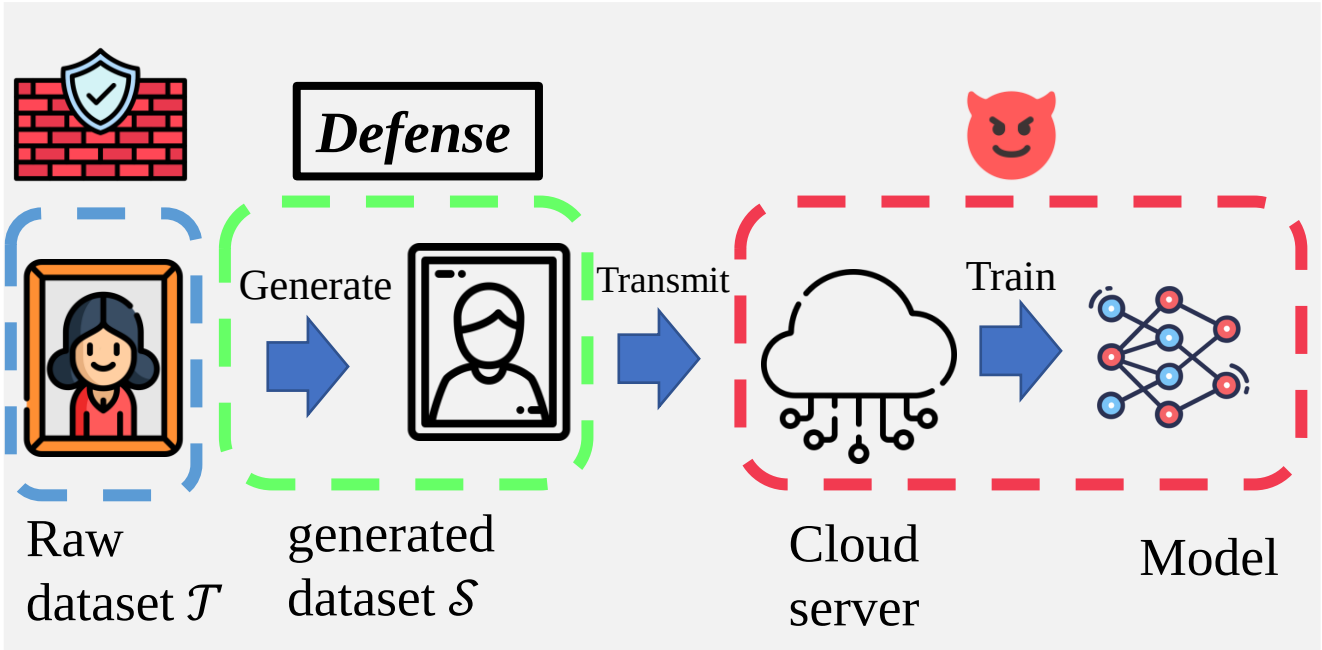
- Infer membership of raw data, i.e., whether $x \in \mathcal{T}$.

Adversary's knowledge:

- Data distribution of $\mathcal{T}$
- No access to raw dataset
- White-box access to
  - generated (privacy-preserving) dataset $\mathcal{S}$
  - Models trained on generated data

Adversary's Capacity:

- Produce shadow generated data with same distribution as $\mathcal{T}$
- Train shadow models on shadow generated data



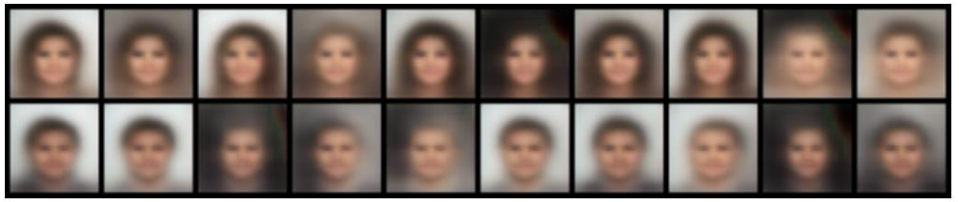
Existing solutions and limitations

Differential privacy (DP)-based generator

- (ϵ, δ)-DP: for randomized algorithm $\mathcal{M}$, (D, D') neighbor datasets

$$\mathbb{P}(\mathcal{M}(D) \in S) \leq e^{\epsilon}\mathbb{P}(\mathcal{M}(D') \in S) + \delta$$

- DP-GAN (Xie et al., 2018)
- GS-WGAN (Chen et al., 2020)
- DP-MERF (Harder et al., 2021)
- DP-Sinkhorn (Cao et al., 2021)



Images generated via DP-Sinkhorn for gender classificaiton (Cao et al., 2021)

Existing solutions and limitations

Limitations:

- Requiring sufficient computing power → generators are hard to train on edge devices (e.g., smart camera).
- Introduced noises (e.g., by Gaussian mechanism) lower the utility of generated data → more generated data are needed for training → lower model training efficiency (i.e., sample-efficiency).

Existing solutions and limitations

Limitations:

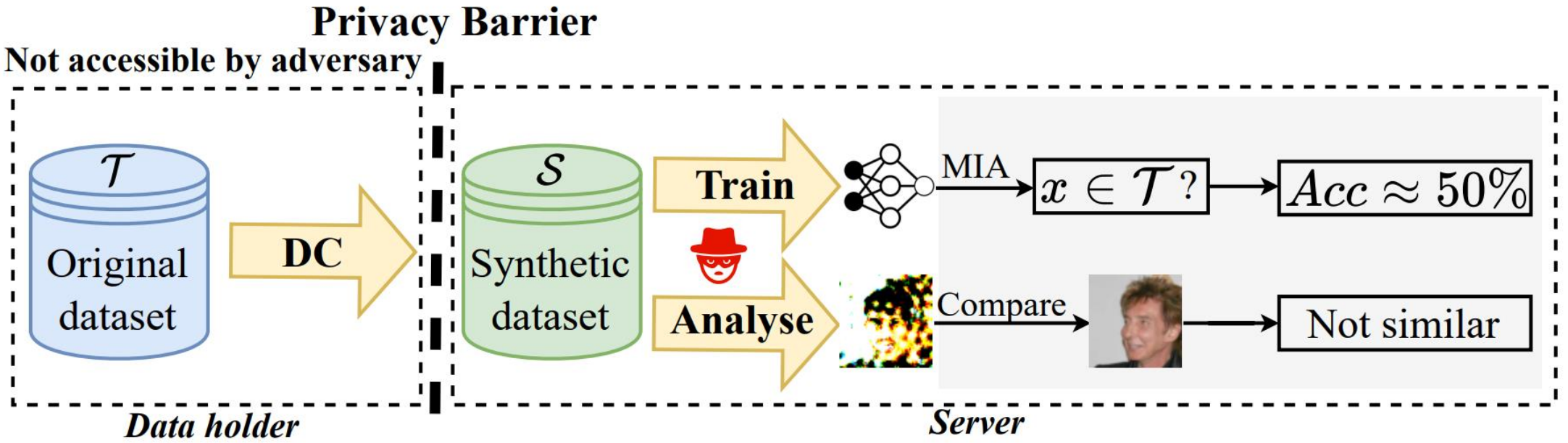
- Requiring sufficient computing power → generators are hard to on
- of generated data → more generated data are needed for training

  → lower model training efficiency (i.e., data-efficiency).

How to generate privacy-preserving data for data-efficient model training?

Apply DC for privacy

Our solution: Apply dataset condensation (DC) to synthesize surrogate data for privacy-preserving model training.



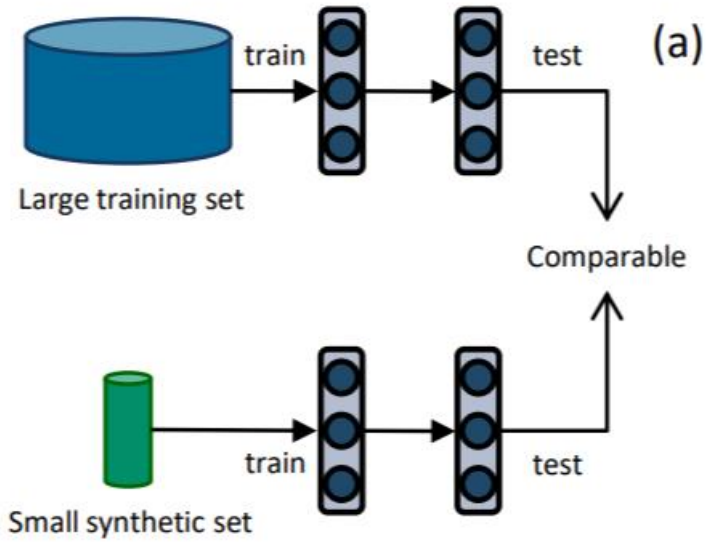
The adversary on cloud can neither recover the raw data through visual comparison analysis nor infer raw data membership from DC-synthesized data.

What is Dataset Condensation?

Objective: distill knowledge from a large training set into a small (high-quality) synthetic set.

Main approaches:

- Dataset Distillation (Wang et al., 2018)
- Gradient Matching (Zhao et al., 2021)
- Differentiable Siamese Augmentation (**DSA**) (Zhao & Bilen, 2021b)
- Kernel Inducing Point (**KIP**) (Nguyen et al., 2021a;b).
- Distribution Matching (**DM**) (Zhao & Bilen, 2021a)
- Matching Training Trajectories (Cazenavette et al., 2022)
- Contrastive Signals (Lee et al., 2022)



Dataset Condensation aims to generate a small set of synthetic images that can match the performance of a network trained on a large image dataset (Zhao et al., 2021) .

Motivation

Observation: Data synthesized by Dataset Condensation (DC) are visually different from original data and enable models to achieve high accuracy.

Original Synthesized

Female

Original Synthesized

Male

Main Results

Overview

- **Theoretical findings**
  - Connection between raw and synthetic data
  - Analysis on visual privacy and membership privacy

- **Empirical validations**
  - Visual privacy analysis by image similarity comparison.

  - Membership privacy analysis against loss-based and likelihood-based attack LiRA (Carlini et al., 2022).

  - Utility and sample-efficiency comparison between DC and data generators (cGAN and differentially private generators)

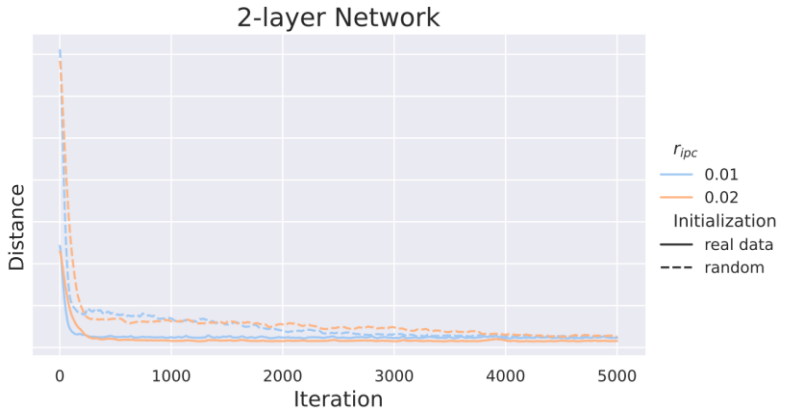
Main Results

Theoretical findings:

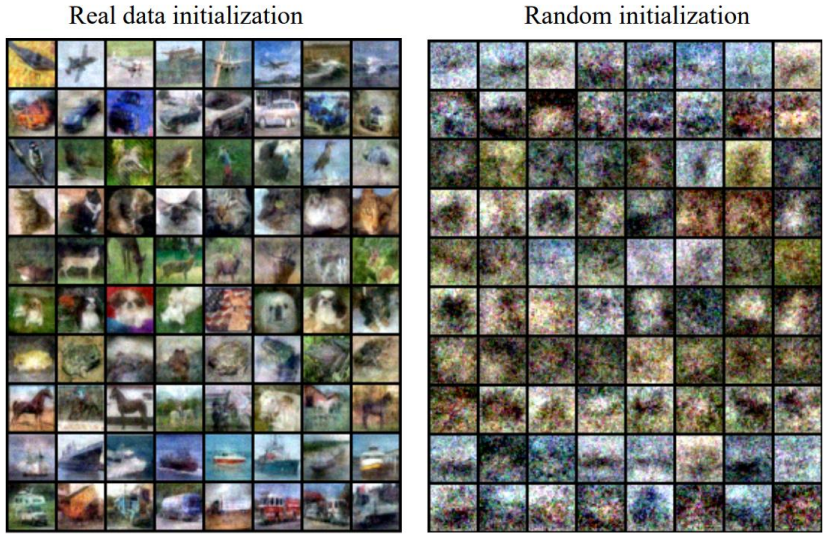
- Connect raw data with synthetic data (Proposition 4.3):
  - Barycenters of synthetic dataset $\mathcal{S}$ and raw dataset $\mathcal{T}$ coincide after condensation.

- **Visual privacy of synthetic data** for different initializations (Proposition 4.4):
  - Real data: initialization data can be leaked
  - Random: No membership information can be leaked

- **Membership privacy for models trained on synthetic data** (Proposition 4.10):
  The existence of one sample in raw dataset has limited impact ($O(\frac{|\mathcal{S}|}{|\mathcal{T}|})$) on models trained on synthetic data (idea of DP).



Empirical verification of Proposition 4.3.



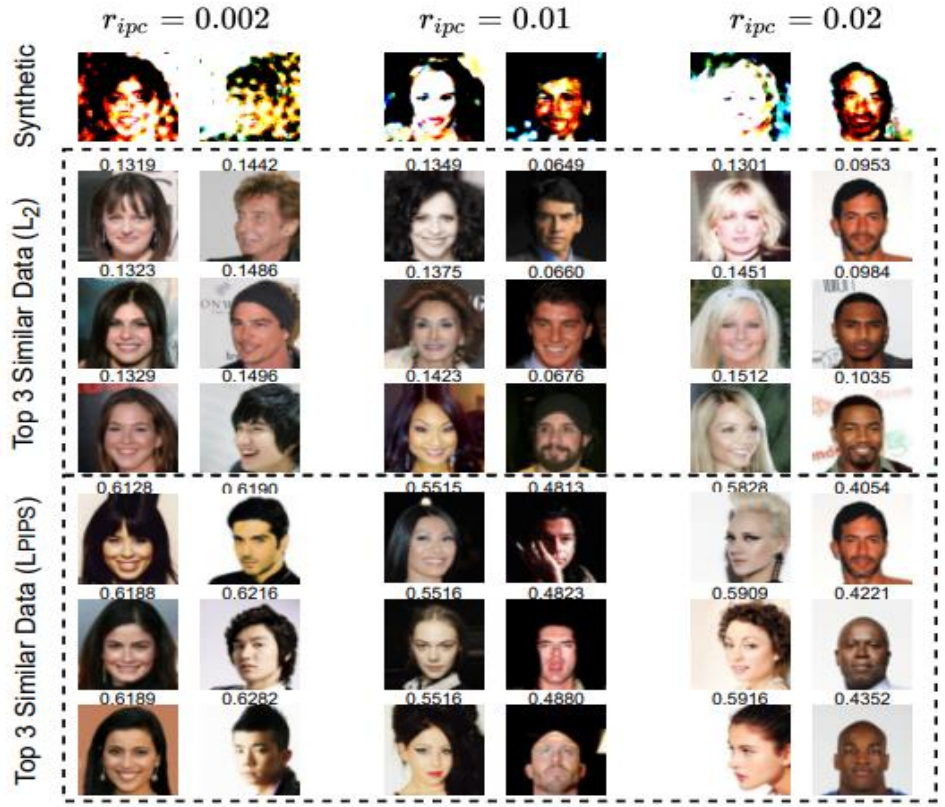
Empirical verification of Proposition 4.4.

Main Results

Visual privacy:
Find the top 3 most similar images via comparison (L_2 norm & LPIPS).

Compression ratio
(ratio of images per class):

$$r_{ipc} = \frac{|\mathcal{S}|}{|\mathcal{T}|}$$



Takeaway: The adversary cannot recover raw data from synthetic data by visual comparison.

Main Results

Membership privacy (Loss-based MIA, DC with **real data initialization**)

Advantage (%) = $2 \times \left(\frac{|\cup_x\{x|x\in\mathcal{T}_{init}, l(x)<\tau\}|+|\cup_x\{x|x\in\mathcal{T}_{init}^C, l(x)\geq\tau\}|}{|\mathcal{T}_{init}|+|\mathcal{T}_{init}^C|} - 50\%\right)$, where $\mathcal{T}_{init} \subset \mathcal{T}$ is used for initialization $\mathcal{T}_{init}^C \subset \mathcal{T} \setminus \mathcal{T}_{init}$ and $|\mathcal{T}_{init}| = |\mathcal{T}_{init}^C|$.

Takeaway: Data used for DC images can still be leaked because of high similarity between synthetic data and raw data.

Method	r_{ipc}	FashionMNST	CIFAR-10	CelebA
Real (baseline, non-private)	0.002	46.67 ± 16.33	72.00 ± 24.00	100.00 ± 0.00
	0.01	21.00 ± 3.67	92.80 ± 5.31	84.00 ± 5.06
	0.02	17.33 ± 2.91	82.60 ± 5.59	77.00 ± 6.71
DM	0.002	78.17 ± 3.20	49.80 ± 5.83	37.00 ± 12.69
	0.01	83.67 ± 2.77	64.20 ± 4.77	47.00 ± 19.52
	0.02	83.00 ± 2.56	68.20 ± 7.35	53.00 ± 14.18
DSA	0.002	74.40 ± 2.65	55.40 ± 8.20	30.50 ± 8.16
	0.01	81.60 ± 2.27	56.60 ± 2.95	28.00 ± 3.74
KIP (w/o ZCA)	0.002	67.83 ± 4.54	42.40 ± 4.80	23.00 ± 11.87
	0.01	70.00 ± 2.47	51.40 ± 5.73	25.00 ± 15.65
KIP (w/ ZCA)	0.002	67.67 ± 4.42	50.40 ± 5.35	23.00 ± 15.52
	0.01	64.00 ± 4.23	48.40 ± 6.62	17.00 ± 18.47

Main Results

Membership privacy (Loss-based MIA, DC with **random initialization**)

Advantage (%) = $2 \times \left(\frac{|\cup_x\{x|x\in\mathcal{T}_{mem}, l(x)<\tau\}|+|\cup_x\{x|x\in\mathcal{T}_{mem}^C, l(x)\geq\tau\}|}{|\mathcal{T}_{mem}|+|\mathcal{T}_{mem}^C|} - 50\%\right)$, where $|\mathcal{T}_{mem}| = |\mathcal{T}_{mem}^C|$, $\mathcal{T}_{mem} \cap \mathcal{T}_{mem}^C = \emptyset$ and $\mathcal{T}_{mem} \cup \mathcal{T}_{mem}^C = \mathcal{T}$

cGAN model can still leak privacy (Chen et al., 2020). Not private!

Takeaway: The advantage of loss-based MIA is close to 0, indicating the attack cannot effectively infer data membership privacy.

Methods	r_{ipc}	FashionMNST	CIFAR-10	CelebA
cGAN (baseline, non-private)	0.002	0.29 ± 0.89	-0.44 ± 1.88	-0.57 ± 0.97
	0.01	0.18 ± 1.21	-0.58 ± 2.09	-0.81 ± 0.95
	0.02	0.04 ± 0.70	-0.77 ± 1.59	-0.47 ± 1.22
DM	0.002	-0.34 ± 0.42	0.31 ± 1.93	-0.66 ± 1.44
	0.01	-0.29 ± 0.48	1.06 ± 1.20	-0.56 ± 1.52
	0.02	0.18 ± 0.53	0.72 ± 0.70	-0.67 ± 1.18
DSA	0.002	0.09 ± 0.51	0.39 ± 1.04	-0.39 ± 1.90
	0.01	0.52 ± 0.55	1.27 ± 1.71	-1.16 ± 0.90
KIP (w/o zca)	0.002	-1.13 ± 1.84	0.25 ± 1.20	-0.56 ± 1.07
	0.01	-0.95 ± 0.96	0.25 ± 1.80	-1.51 ± 0.69
KIP (w/ zca)	0.002	-0.56 ± 2.02	-0.64 ± 1.86	-1.06 ± 1.10
	0.01	-1.69 ± 1.96	-0.22 ± 1.27	-1.80 ± 1.91

Main Results

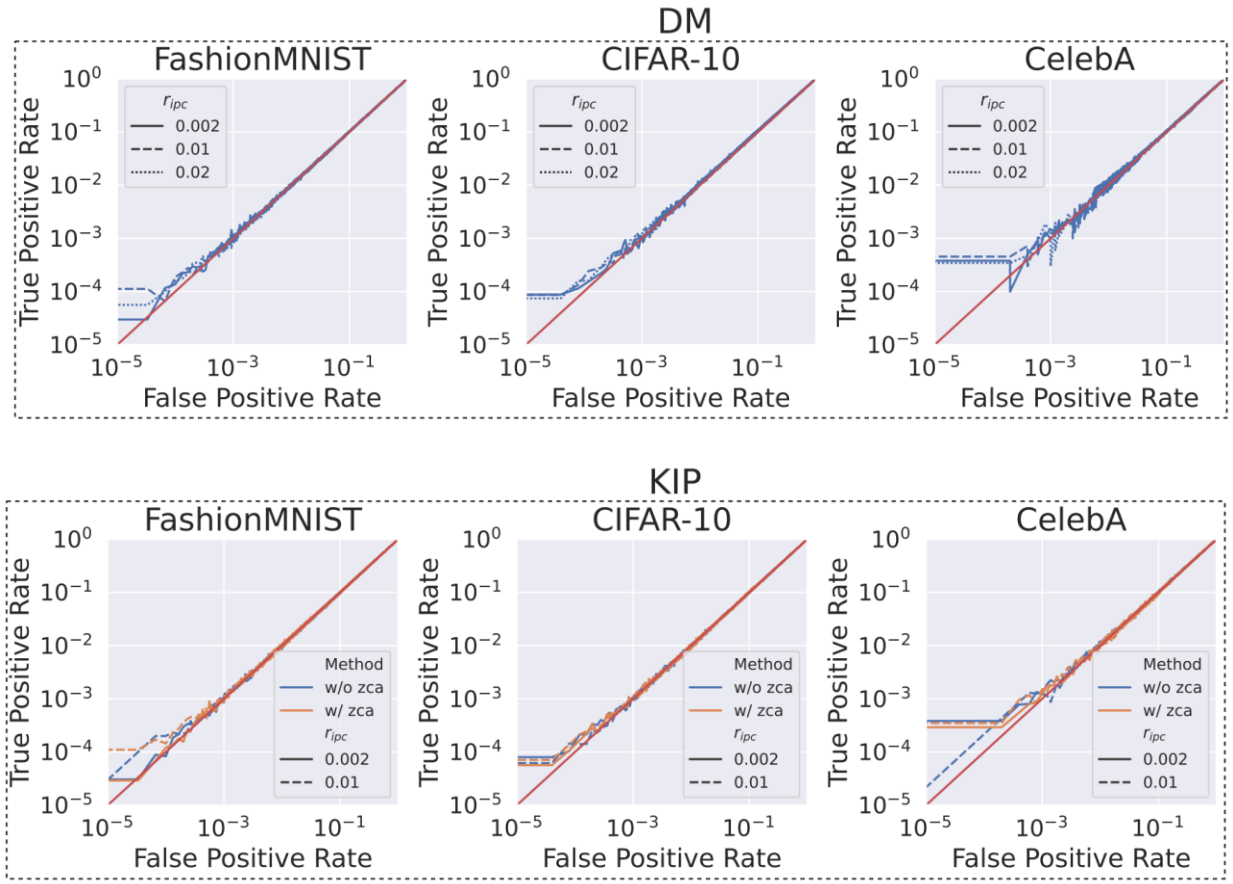
Membership privacy (Likelihood-based MIA, DC with **random initialization**)

Likelihood-based MIA (LiRA) (Carlini et al., 2022):

Thresholding the likelihood Λ:

$$\Lambda = \frac{p(\text{conf}_{obs}|\mathcal{N}(\boldsymbol{\mu}_{in}, \boldsymbol{\sigma}^2_{in}))}{p(\text{conf}_{obs}|\mathcal{N}(\boldsymbol{\mu}_{out}, \boldsymbol{\sigma}^2_{out}))}$$

Takeaway: LiRA cannot effectively infer membership privacy for models trained on synthetic data.



Main Results

Utility comparison between DM and differentially private data generator on FashionMNIST

Takeaway: **DM-synthesized data enable models to achieve higher accuracy than private & non-private generators.**

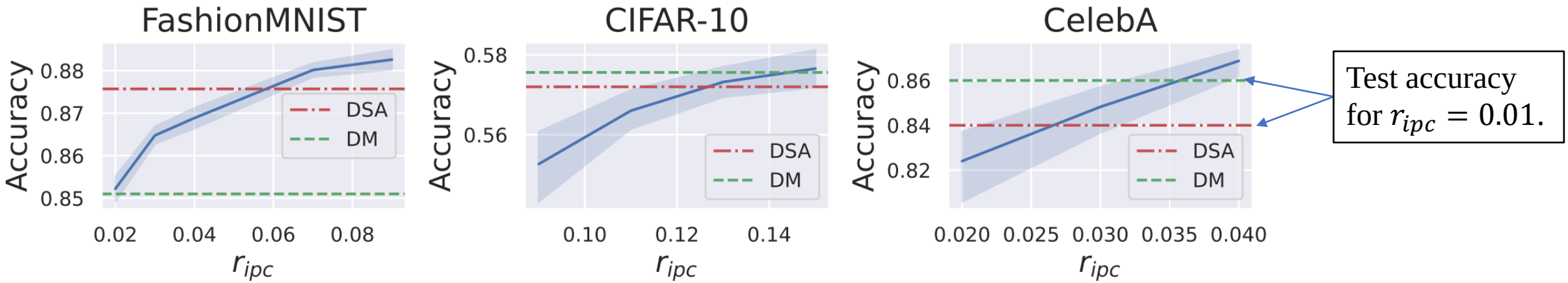
Method	DP Budget	r_{ipc}		
		0.002	0.01	0.02
GS-WGAN	$\epsilon = 10$	53.53 ± 0.42	51.85 ± 0.54	50.10 ± 0.32
DP-MERF	$\epsilon = 10$	52.18 ± 0.37	52.88 ± 0.75	50.73 ± 0.66
	$\epsilon = 2$	60.41 ± 0.78	55.14 ± 0.61	56.39 ± 0.45
DP-Sinkhorn	$\epsilon = 10$	-	-	70.9^*
KIP (w/o zca)	$\hat{\epsilon} = 1.25$	73.70 ± 1.13	68.11 ± 1.33	-
KIP (w/ zca)	$\hat{\epsilon} = 2.07$	74.37 ± 0.96	70.03 ± 0.84	-
DM	$\hat{\epsilon} = 2.30$	80.59 ± 0.62	85.10 ± 0.51	86.13 ± 0.34

* Results reported in the paper (Cao et al., 2021) ($r_{ipc} = 1$).

Note: The empirical budget $\hat{\epsilon}$ and ϵ are provided not for comparison but only to show robustness against MIA.

Main Results

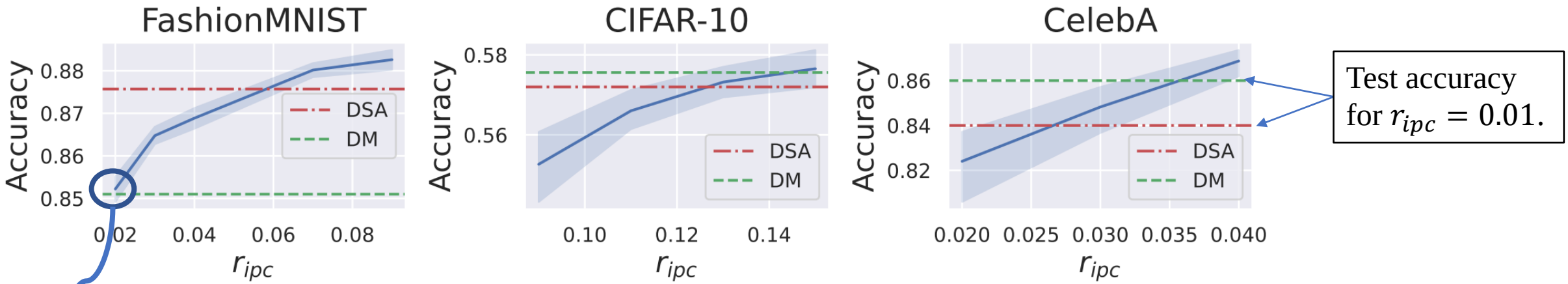
Sample-efficiency comparison between DM, DSA and cGAN (Baseline):
The amount of generated data (measured by r_{ipc}) needed to achieve certain accuracy.



Test accuracy of models trained on cGAN-generated data for different r_{ipc}.

Main Results

Sample-efficiency comparison between DM, DSA and cGAN (Baseline):
The amount of generated data (measured by r_{ipc}) needed to achieve certain accuracy.



Test accuracy of models trained on cGAN-generated data for different r_{ipc}.

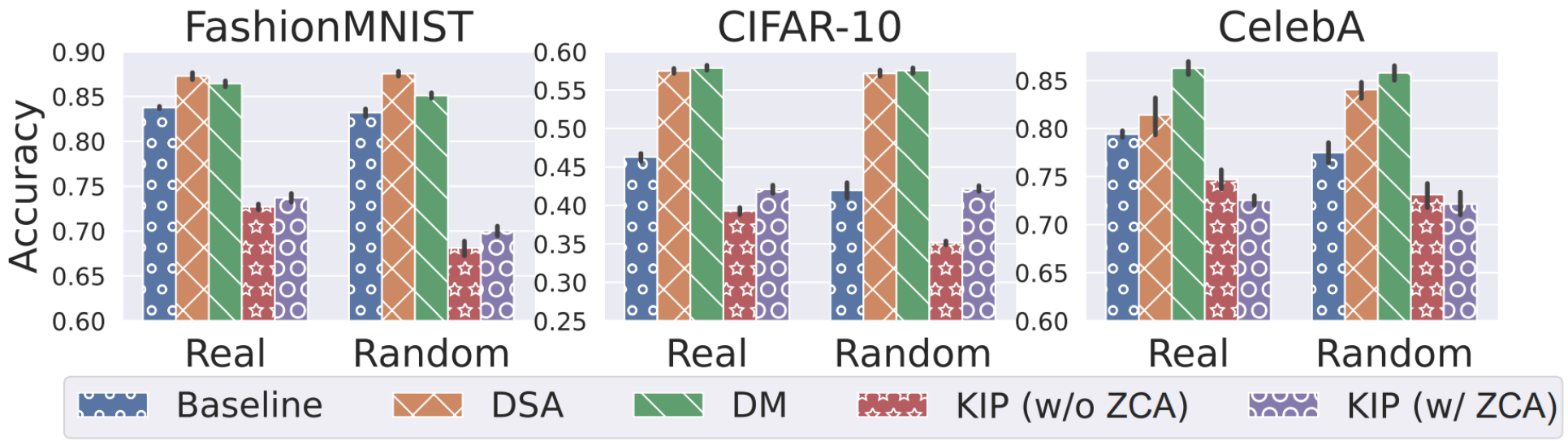
Example: On FashionMNIST, cGAN needs to generate data with $r_{ipc} = 0.02$ to achieve the same test accuracy (0.85) as DM method with $r_{ipc} = 0.01$

=> 2 times efficiency improvement

Takeaway: To achieve the same accuracy, DC needs (at least 2 times) fewer samples, thus is more sample-efficient.

Main Results

Utility comparison between DM, DSA, KIP and cGAN (Baseline) under $r_{ipc} = 0.01$.



Takeaway: DM and DSA outperform the other methods in generating high-quality data.

Summary

- We identify the privacy benefit of DC and propose to use DC for efficient and privacy-preserving data generation in machine learning pipeline.
- We theoretically analyze why DC can help protect visual and membership privacy.
- We empirically validate the privacy benefit brought by DC with two MIAs (loss-based and likelihood-based) on three image datasets.
- We envision this work as a milestone for data-efficient and privacy-preserving machine learning.