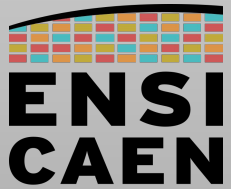
Generating Private Data Surrogates for Vision Tasks

Ryan Webster
Julien Rabin
Loic Simon
Frederic Jurie

rwebstr@gmail.com



Privacy in Generative Models

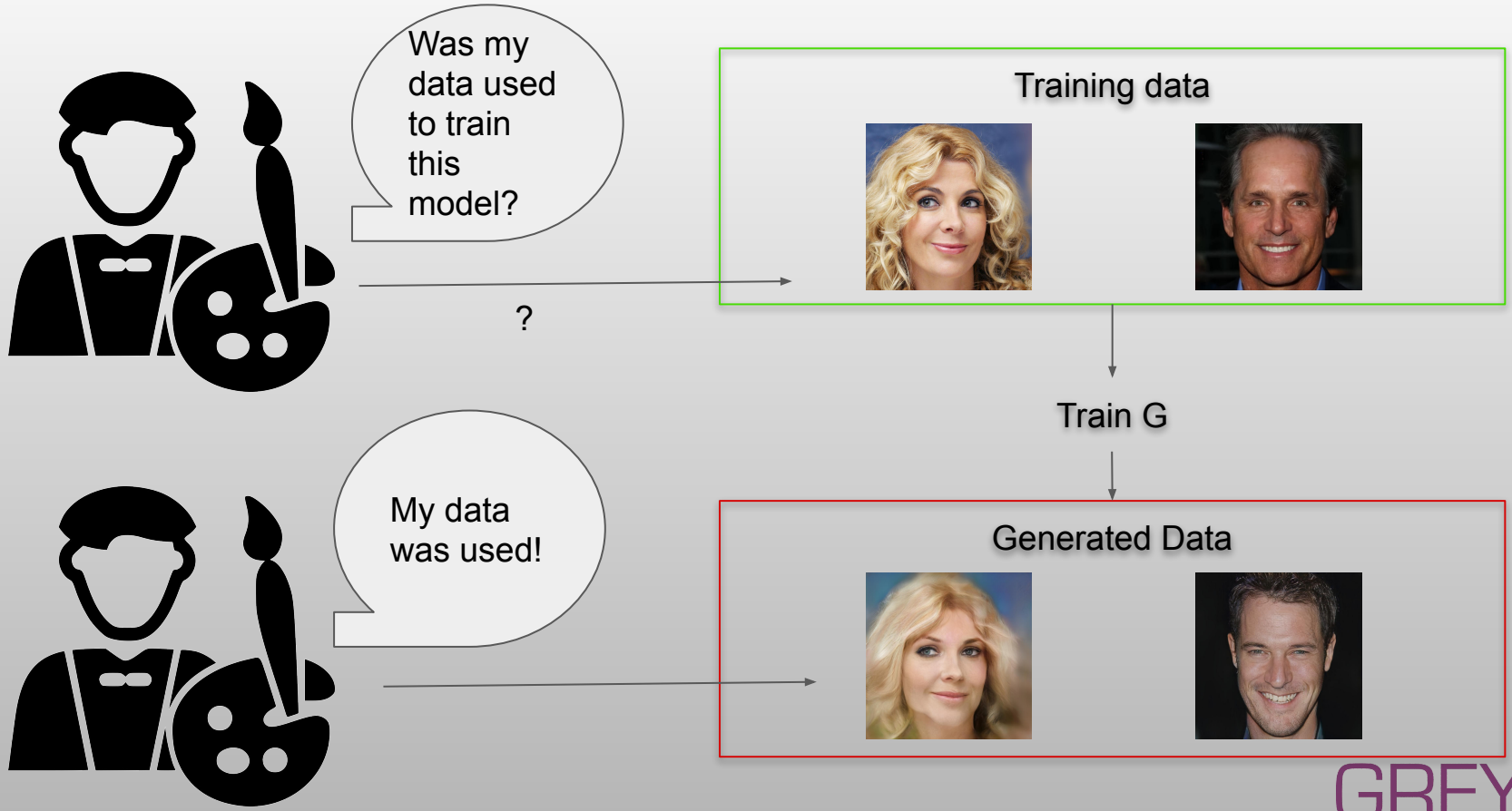
- Deep networks can contain information about their training data.
- Membership attacks involve discerning which samples were used given a model.
- Recently, several successful attacks have been demonstrated against generative models.
- GANs appear to be resistant to such attacks. In this work, we utilize this observation to create surrogate data, giving privacy for other tasks (such as classification).

GREYC

Example of Privacy in Image Generation

- In the medical domain, a publicly released model could be used to determine which patients records were used during training.
- Generative models have particular interest in the artistic domain. An artist may be able to discern if his work was used to train a model. This may be a problem if the model was trained without copyright permission.

Example of Generative Privacy: Copyright



Outline

1. Membership attacks
    a. General description
    b. Attacks against generative models
        i. Discriminative Attacks
        ii. Recovery Attacks
2. Creating Data Surrogates with GANs
3. Results

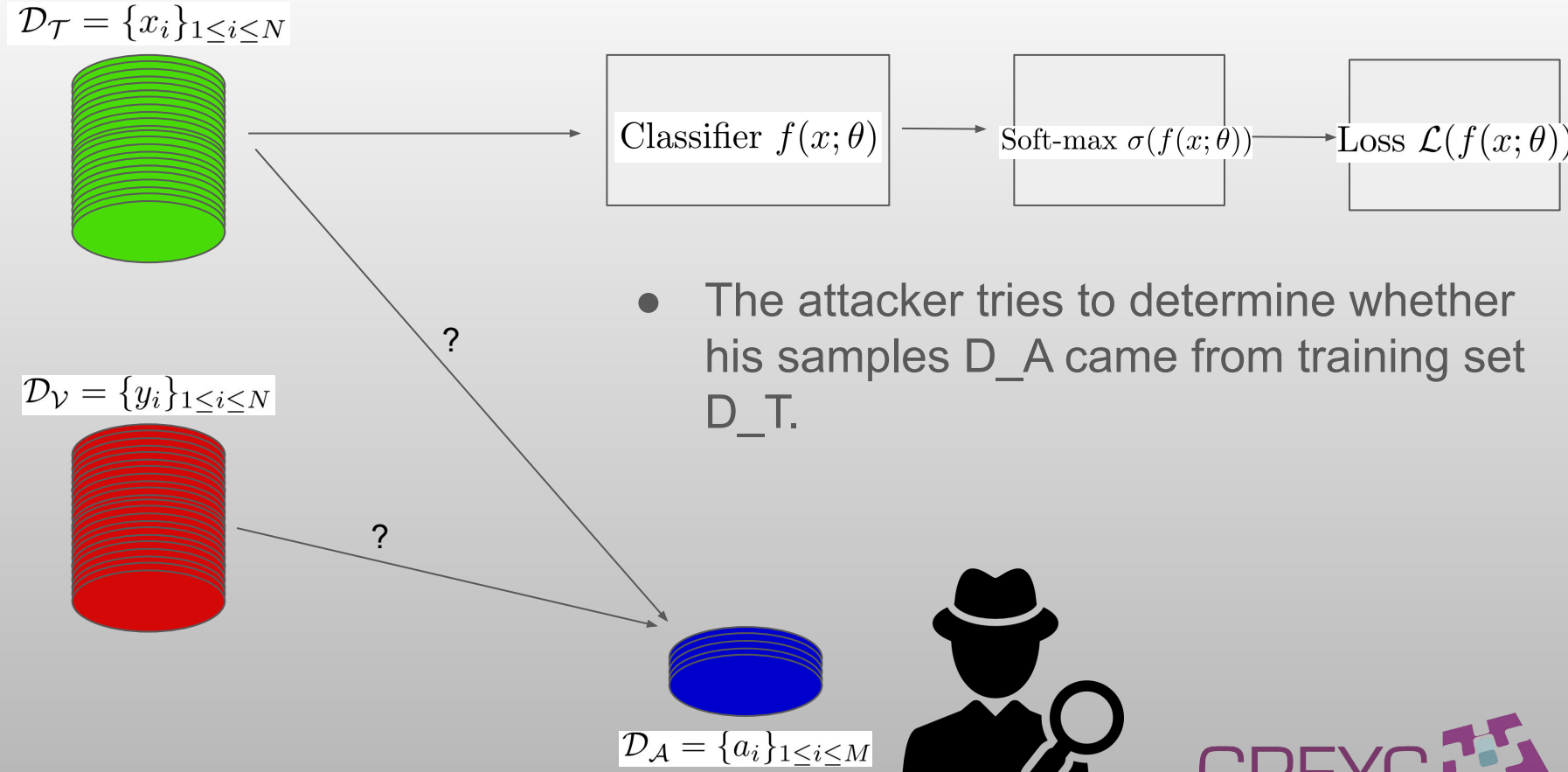
GREYC

Membership Attacks

- An attacker holds some data suspected to have been used to train a model. They posses samples D_A, which is comprised of samples from both the training set D_T, and validation set D_V. He may have some knowledge about the model.
- In a whitebox attack, the attacker has access the model parameters θ, in a blackbox attack, they only have access to the model outputs f(x; θ).

GREYC

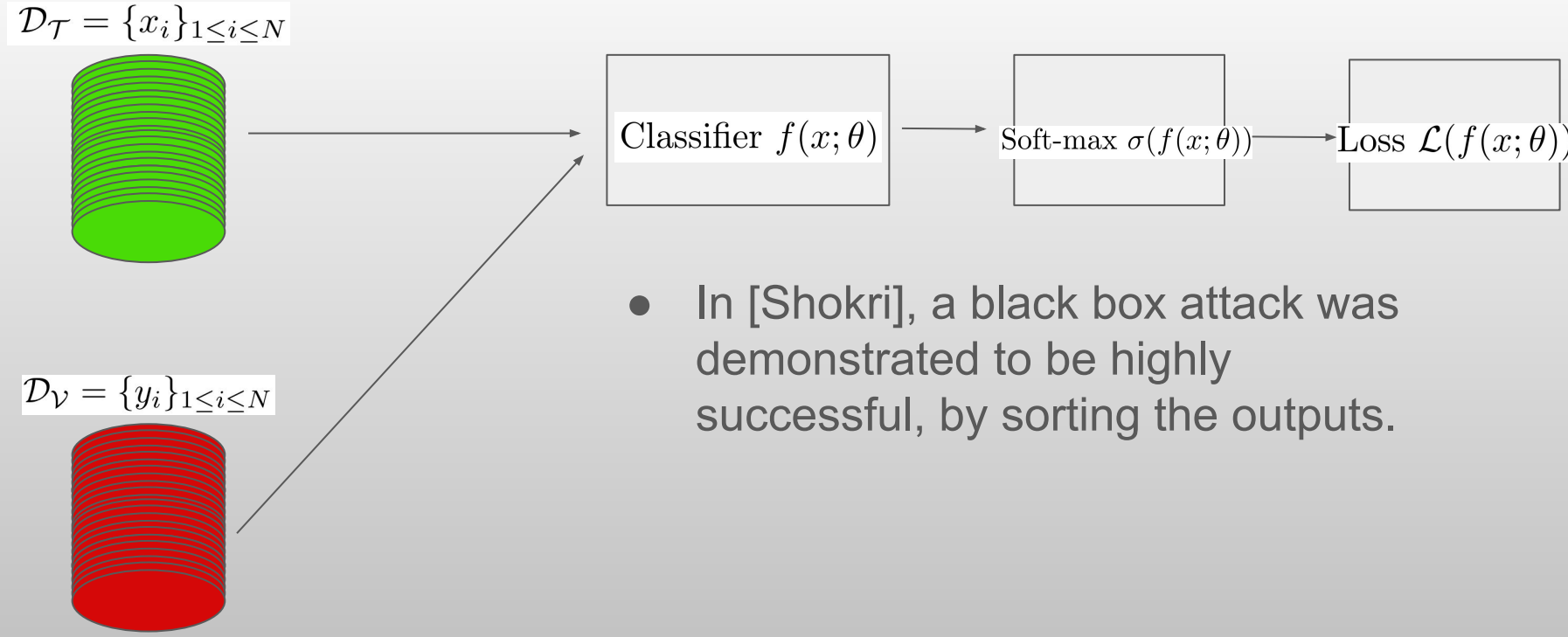
Membership Attacks



- The attacker tries to determine whether his samples D_A came from training set D_T.

GREYC

Membership Attacks



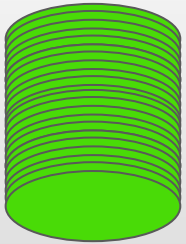
- In [Shokri], a black box attack was demonstrated to be highly successful, by sorting the outputs.

"Membership Inference Attacks Against Machine Learning Models" Shokri et al, 2017

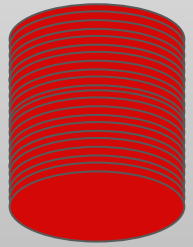
GREYC

Membership Attacks

$\mathcal{D}_{\mathcal{T}} = \{x_i\}_{1 \leq i \leq N}$



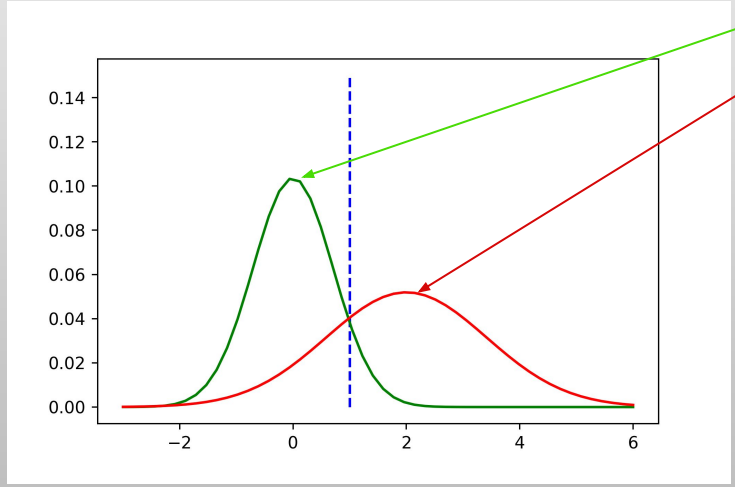
$\mathcal{D}_{\mathcal{V}} = \{y_i\}_{1 \leq i \leq N}$



Classifier $f(x;\theta)$

Soft-max $\sigma(f(x;\theta))$

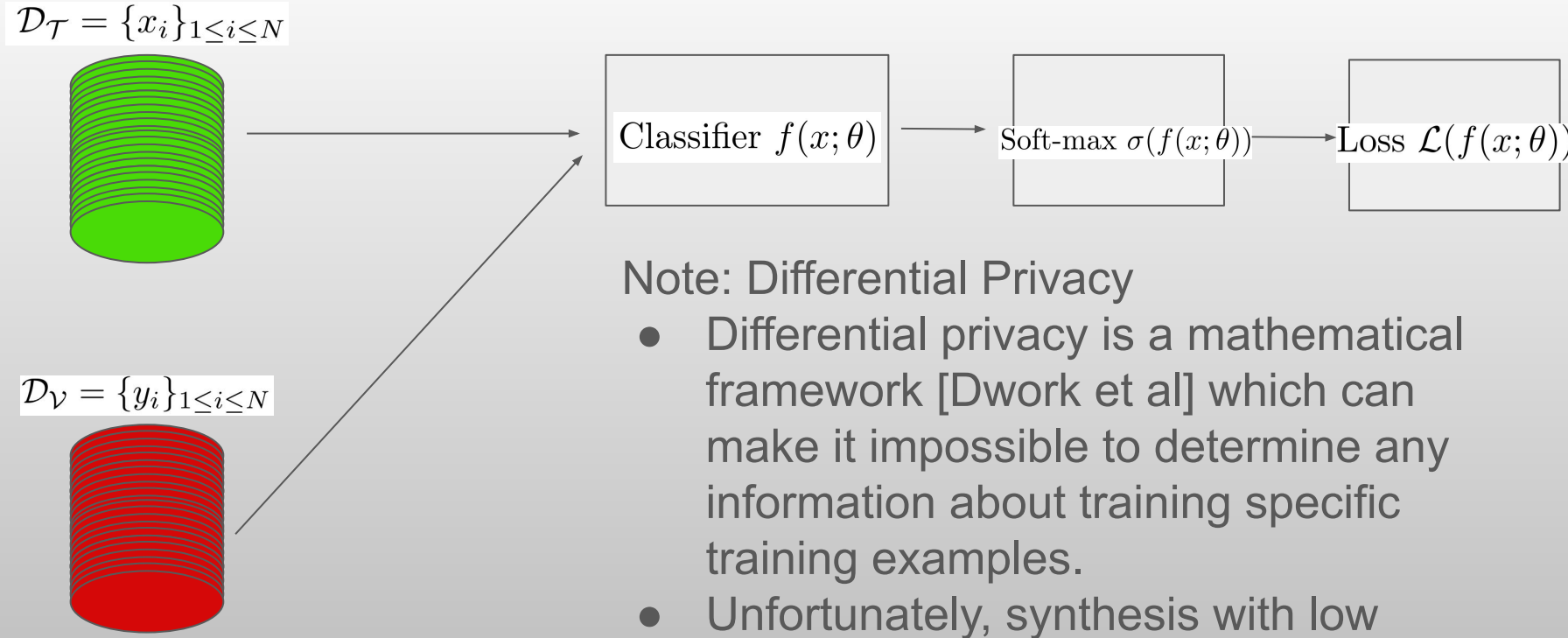
Loss $\mathcal{L}(f(x;\theta))$



"Membership Inference Attacks Against Machine Learning Models" Shokri et al, 2017

GREYC

Membership Attacks



Note: Differential Privacy

- Differential privacy is a mathematical framework [Dwork et al] which can make it impossible to determine any information about training specific training examples.
- Unfortunately, synthesis with low epsilon has not been demonstrated.

"The Algorithmic Foundations of Differential Privacy" Cynthia Dwork, (2006)

GREYC

Membership attacks: Generic Algorithm

Algorithm 1 Membership attack

Input: Training set $\mathcal{D}_{\mathcal{T}}$, validation set $\mathcal{D}_{\mathcal{V}}$

1: Set the attack score function A, either from the recovery loss function in Eq. (1) or the discriminator D.

2: Let $x_i \in \mathcal{D}_{\mathcal{T}} \cup \mathcal{D}_{\mathcal{V}}$, such that

$$\begin{cases} x_i \in \mathcal{D}_{\mathcal{T}} & \text{if } i \leq N \\ x_i \in \mathcal{D}_{\mathcal{V}} & \text{if } N+1 \leq i \leq 2N \end{cases}$$

3: Sorted indices: $I \leftarrow \arg\operatorname{sort}\{A(x_i)\}_{1 \leq i \leq 2N}$

Output:

4: Estimated set of training images: $\mathcal{T} \leftarrow \{x_{I(i)}\}_{1 \leq i \leq N}$

5: Membership attack accuracy: $Acc \leftarrow |I \cap \{i : 1 \leq i \leq N\}|/N$

GREYC

Membership Attacks Against Generative Models

Generative Adversarial Networks (GAN)

$$\max_D \min_G \mathbb{E}_{z\sim\mathcal{N}(0,1),x\sim p_{data}}[log(D(x)) + log(1 - D(G(z)))]$$

Generative Latent Optimization (GLO)

$$\min_G \sum_{(z_i,x_i)} \mathcal{L}_{rec}(G(z_i), x_i) = \|G(z_i) - x_i\|_2^2$$

- How do we attack a generative model, when there is no obvious attack function, only a generator?
- LOGAN (Hayes et al, 2019), shows moderate success just using the discriminator, but this is unrealistic, as the discriminator is usually discarded after training.
- What about in a more realistic scenario, when we only have the generator?

GREYC

Discriminator Attacks

- In LOGAN [Hayes et al], the discriminator is used as the attack function.
- This assumes access to discriminator parameters, which is unlikely as the discriminator is not used in most application settings.

“LOGAN: Membership Inference Attacks Against Generative Models,” Hayes et al.

GREYC

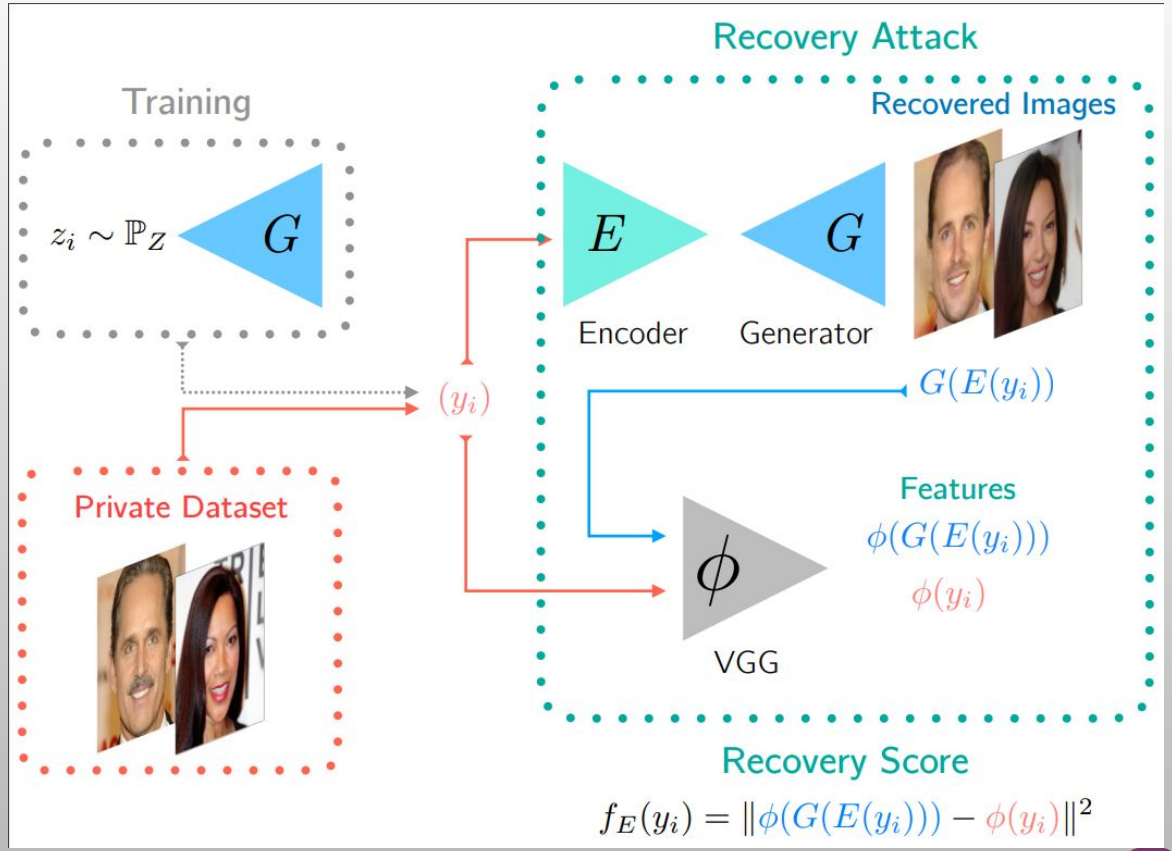
Recovery Attacks

- In [Webster et al.], a recovery attack was highly effective against some generative models.
- Recovery attacks work by solving an inverse problem for the generator and directly generate certain samples.
- In this work, we introduced *encoder* based recovery attacks.
- Encoder parameters are used to invert images into latent codes. The resulting recovery error is used as the attack function, like in [Webster et al].

“Detecting overfitting in Deep Generative Networks via Latent Recovery,” Webster et al, CVPR 2019.

GREYC

Recovery Attacks



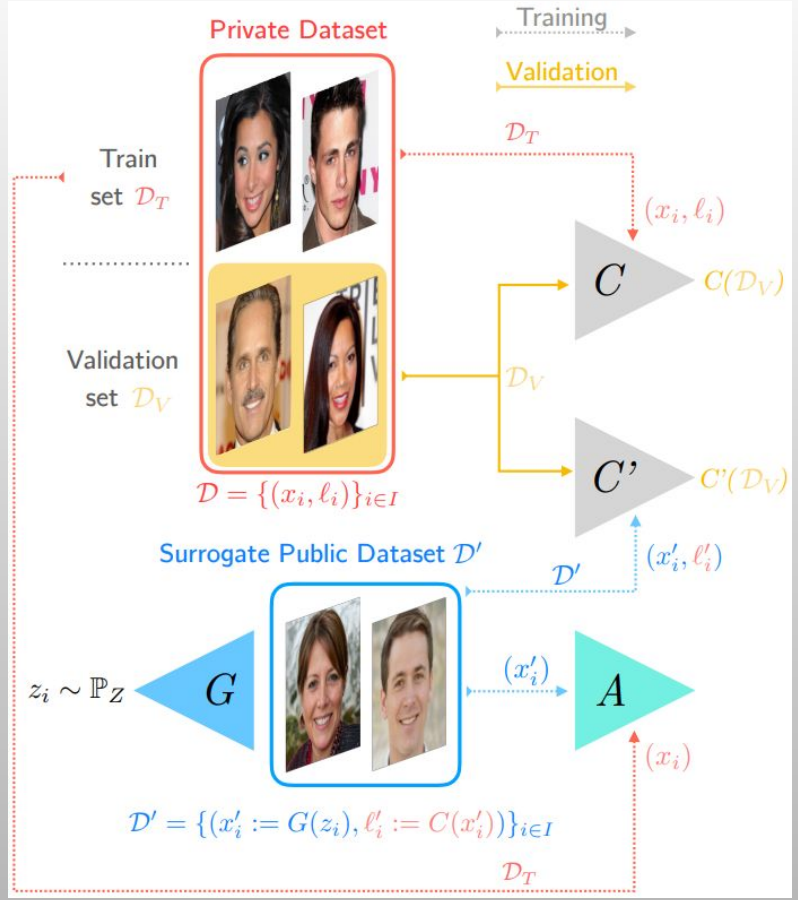
Generated Private Data Surrogates

- In [Webster et al.], recent GANs (such as Progressive GAN), was shown to be resistant to membership attacks.
- We build on this observation and explore the use of generated data in lieu of real data.
- Generated surrogates are empirically resistant to membership attacks, while having similar performance.

“Detecting overfitting in Deep Generative Networks via Latent Recovery,” Webster et al, CVPR 2019.

GREYC

Generated Private Data Surrogates



Results

- We measure the drop in performance when using surrogate data, alongside FID.
- More realistic generated data results in better classification.

			Gender	Smiling	Average (5 attributes)	Drop in Performance	FID
VGG-Face Features	C	Real Data	94.50	85.20	90.64	-	-
	C'	DCGAN	91.90	82.10	86.50	4.14	67.07
		MESCH	92.60	81.45	88.90	1.74	26.31
		LSGAN	92.10	80.80	88.35	2.29	42.01
		PGGAN	93.10	83.05	89.35	**1.29**	**19.17**

Results

- Below are the accuracies of various membership attacks. Here, training data comprises 50% of the attackers data, so that 50% represents random guessing.
- Note that membership attacks are all barely above guessing, with PGGAN being the most resistant to attacks.

	L_2 Recovery	VGG-Face Recovery	VGG-19 Recovery	Discriminator D
DCGAN	54.1	54.5	51.6	57.1
MESCH	53.9	50.8	52.5	50.1
LSGAN ($\|\mathcal{D}_T\| = 26k$)	54.8	54.1	54.0	62.9
LSGAN ($\|\mathcal{D}_T\| = 5k$)	58.1	56.2	57.8	99.4
PGGAN	52.0	50.3	52.1	N/A

GREYC

Conclusion

- Deep learning models are susceptible to membership attacks.
- Recent GAN generated images seem to be resistant against two types of membership attacks.
- Generated data can be used in lieu of real data, with similar classification performance on several tasks.

GREYC