Utilising Differential Privacy and Synthetic Data to Defend Against Adversarial Attacks

Sebastian Kobler

Supervisor: Dr Clément Canonne

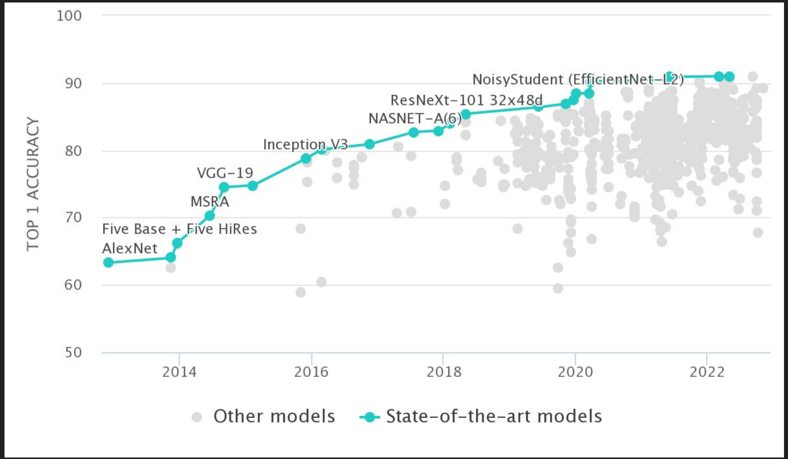
November 16, 2022

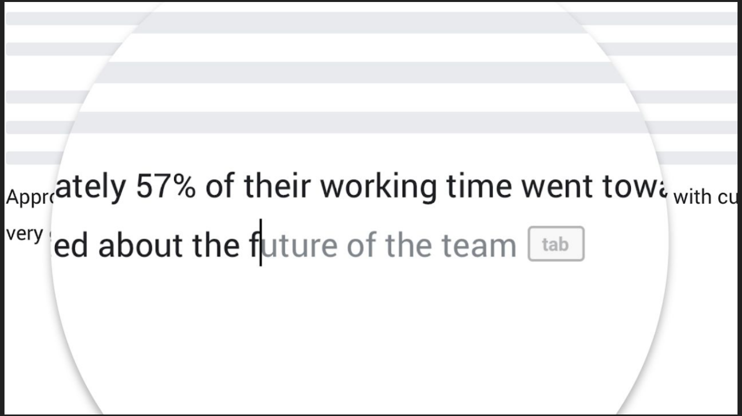
THE UNIVERSITY OF SYDNEY

Outline

1. Motivation
2. Background
3. Experiments and Results
4. Discussion
5. Conclusion

Machine Learning in 2022…





Deep-Learning Algorithm Can Detect Cancer Spread Following Surgery

New research shows that a deep-learning algorithm that uses computed tomography scan images can help predict the spread of head and neck cancer outside of the neck lymph nodes.

Adversarial Machine Learning in 2022…

This Real-Life "Invisibility Cloak" Hides You From Person-Detecting Machine Learning Models

Designed to attack detectors rather than classifiers, these wearable "universal patches" delete you right out of a machine's vision.

Adversarial AI and the dystopian future of tech

New Go-playing trick defeats world-class Go AI—but loses to human amateurs

Adversarial policy attacks blind spots in the AI—with broader implications than games.

BENJ EDWARDS - 11/8/2022, 6:43 AM

Adversarial attacks can cause DNS amplification, fool network defense systems, machine learning study finds

Ben Dickson 25 July 2022 at 11:33 UTC

What can we do?

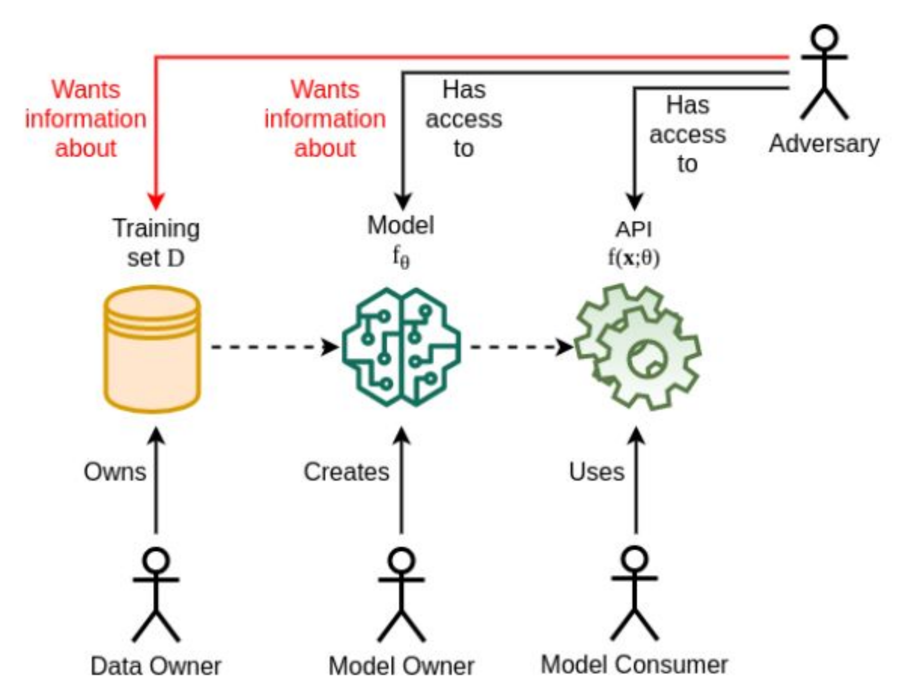
Outline

1. Motivation
2. Background
3. Experiments and Results
4. Discussion
5. Conclusion

Machine learning is not inherently private…

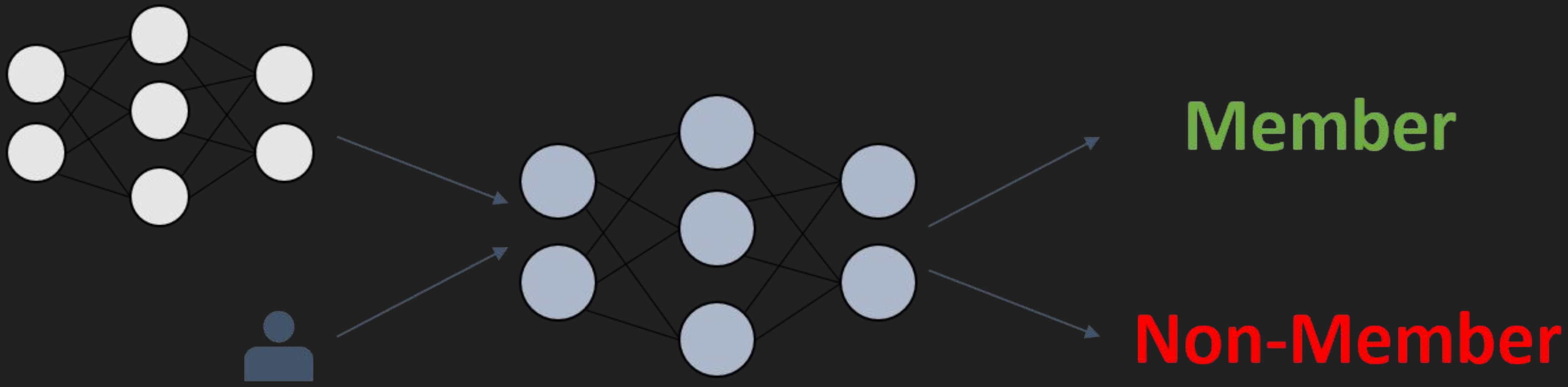
What happens if we introduce the presence of **adversaries**?

Adversarial Attacks



M . Rigaki et al, “A Survey of Privacy Attacks in Machine Learning”

Membership Inference Attacks (MIA)



Given a model and some data record, was this record used in the training process?

Other Attacks

Attacks can target model:

Privacy:

- Membership Inference Attack (MIA)
- Attribute Inference Attack (AIA)
- Reconstruction Attack

Integrity:

- Poisoning Attack

Adversarial Attacks

ties of these algorithms. If we seriously consider taking the human doctor completely 'out of the loop' (which now has legal sanction in at least one setting via the FDA, with many more to likely follow), we are forced to also consider how adversarial attacks may present new opportunities for fraud and harm. In fact, even with a human in the loop, any clinical system that leverages a machine learning algorithm for diagnosis, decision-making, or reimbursement could be manipulated with adversarial examples.

These attacks are both possible and practical (not just a theoretical idea)

S. Finlayson et al. "Adversarial Attacks Against Medical Deep Learning Systems"

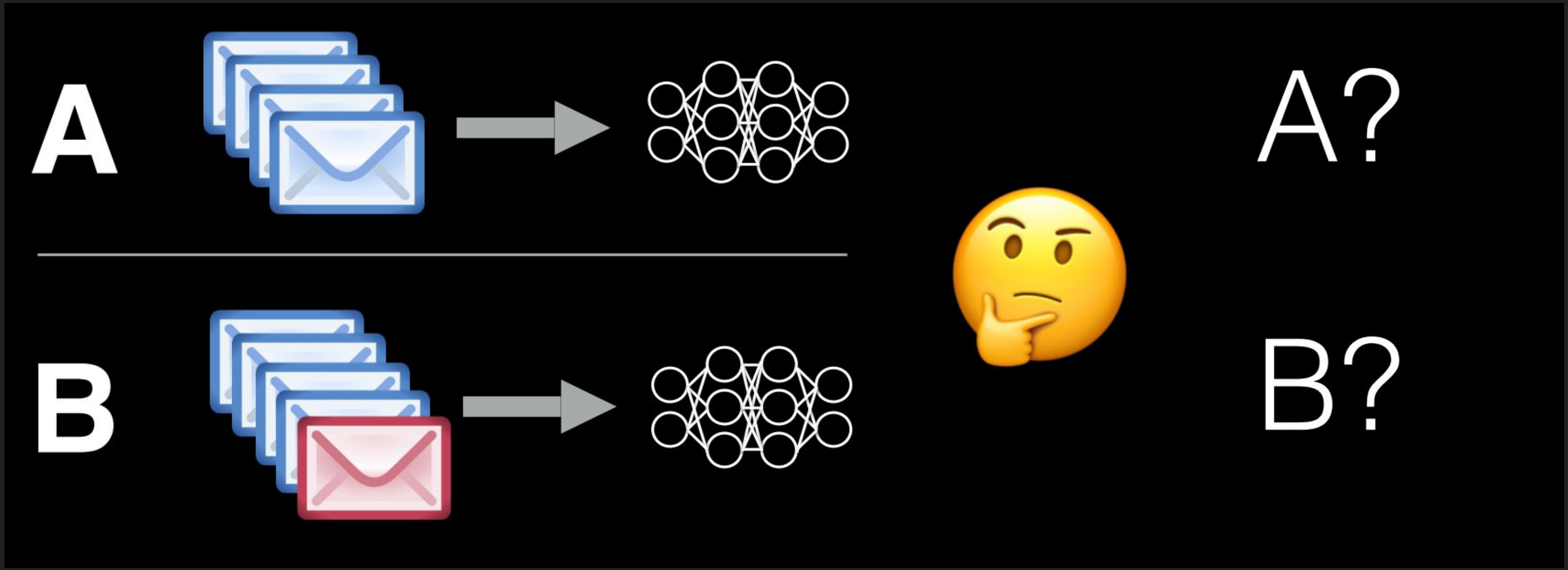
Now that we know attacks against machine learning are possible, can we implement some defences?

Yes!

But it comes at a cost …

Differential Privacy (DP)

"The output of a differentially private analysis will be roughly the same, whether or not you contribute your data"



Nicholas Carlini

Differential Privacy : Mathematical Definition

"The output of a differentially private analysis will be roughly the same, whether or not you contribute your data"

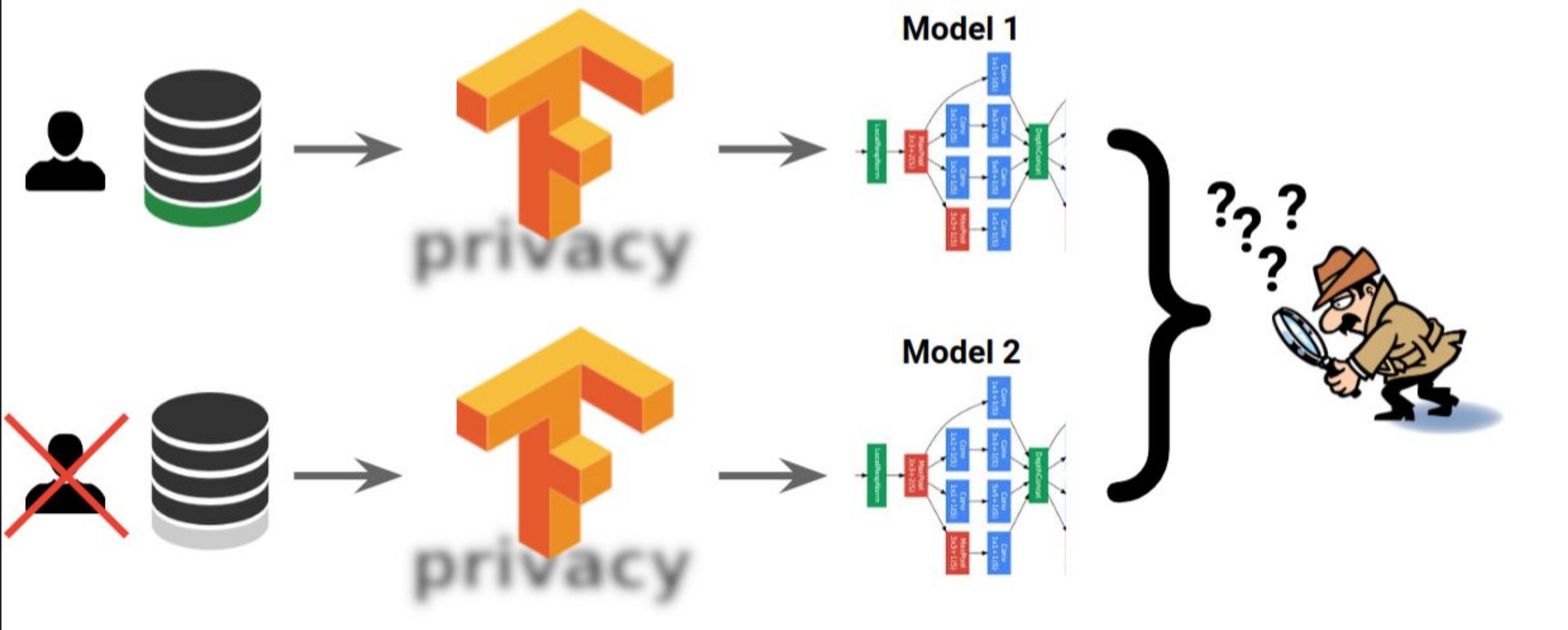
Definition (ϵ - Differential Privacy) Let ϵ be a real number and $\mathcal{M} : \mathcal{X}^n \to \mathcal{T}$ be some randomised algorithm that takes a dataset $\mathcal{D}$ as input. The algorithm $\mathcal{M}$ is (ϵ, δ) - *differentially private* if for any pair of neighbouring datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ and all subsets $S \subset \mathcal{T}$,

$$\mathbb{P}[\mathcal{M}(\mathcal{D}_1) \in S] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{M}(\mathcal{D}_2) \in S] + \delta$$

C. Dwork et al. "Calibrating noise to sensitivity in private data analysis.

ε > 0 quantifies the privacy loss between neighbouring datasets. Therefore, **the smaller the better**

Deep Learning with Differential Privacy



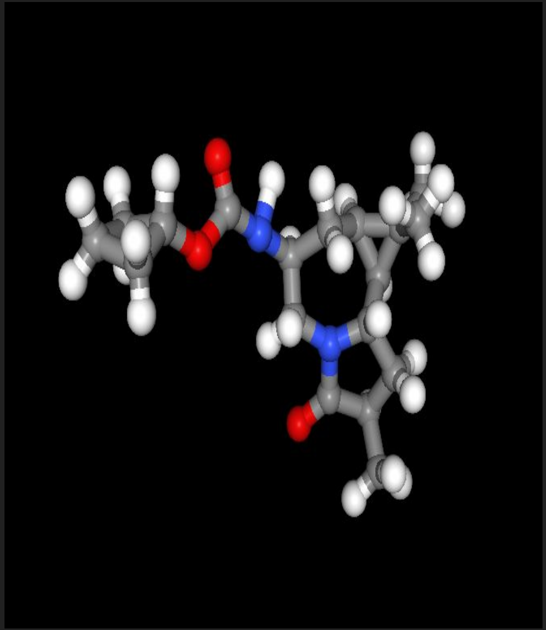
Abadi et al. (2016)

Image Credit*: Tensorflow blog*

But can we also attempt to protect privacy without altering the training process?

Welcome to the world of **Synthetic Data Generation**!

Synthetic Data



This “X” does not exist

How is it done?

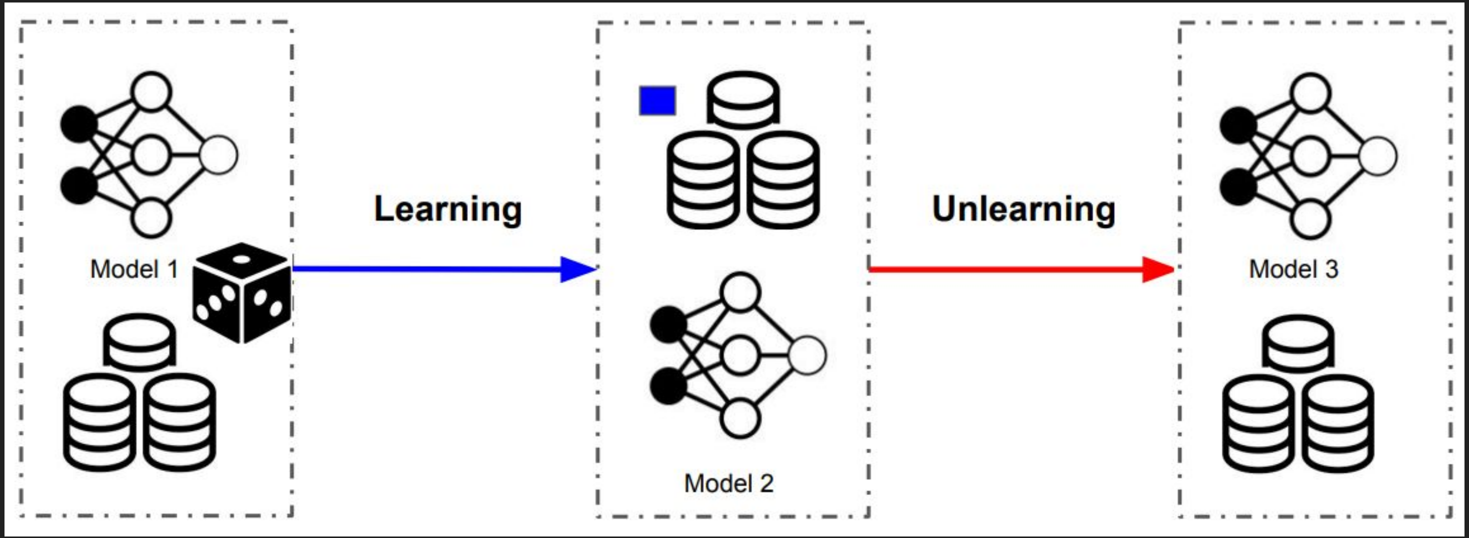
- **Fit** a probabilistic model to the underlying distribution of data
- Sample from it
- Enjoy your new ~~private~~, synthetic data.

Synthetic Data **does not** imply Private Data

Differentially Private Synthetic Data!

Machine Unlearning

Machine Unlearning

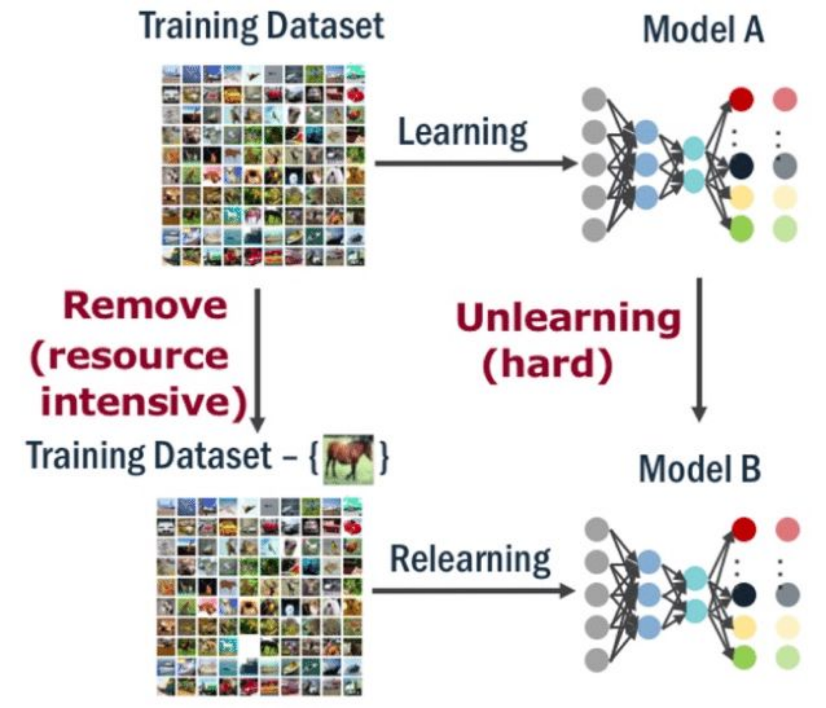


Emmy Fang, Nick Jia.
University of Toronto

But why do we need to unlearn?

“The right to be forgotten”

How do we Unlearn?



Credit: *Class Clown: Data Redaction in Machine Unlearning at Enterprise Scale*

Outline

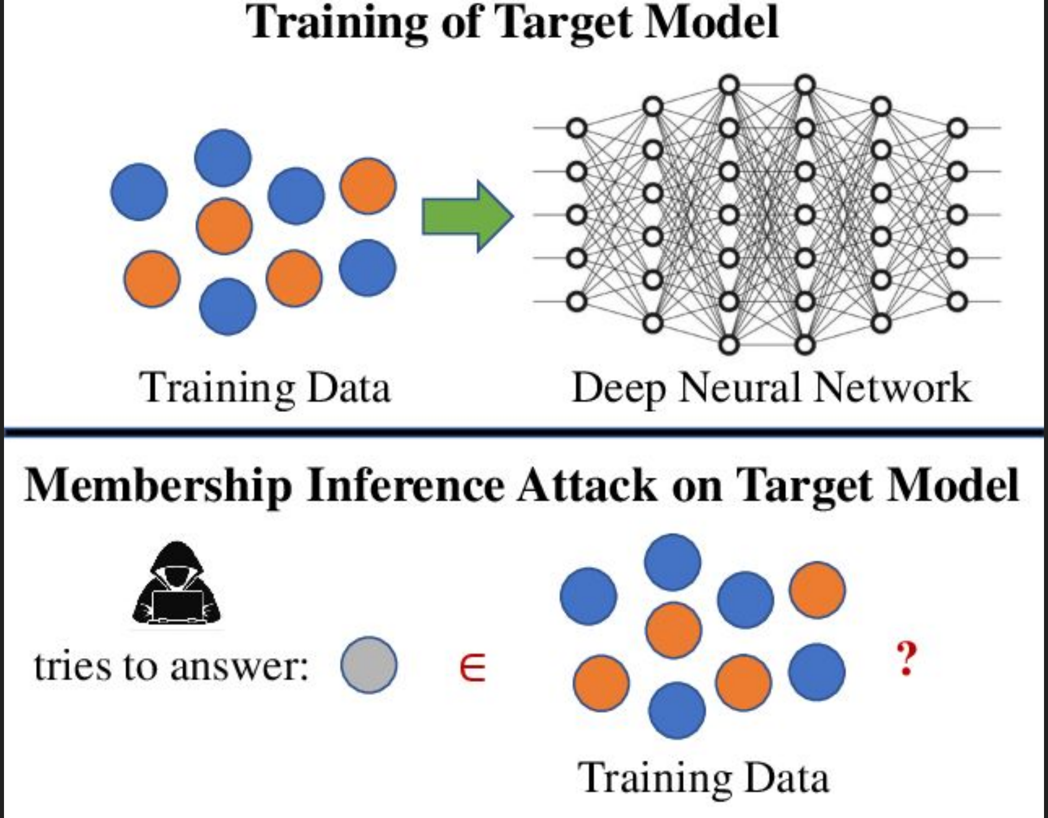
1. Motivation
2. Background
3. Experiments and Results
4. Discussion
5. Conclusion

So why do we need Differentially Private Synthetic Data?

Experiments

1. Attacking Machine Learning Models
2. Generating Synthetic Data
3. Generating Differentially Private Synthetic Data

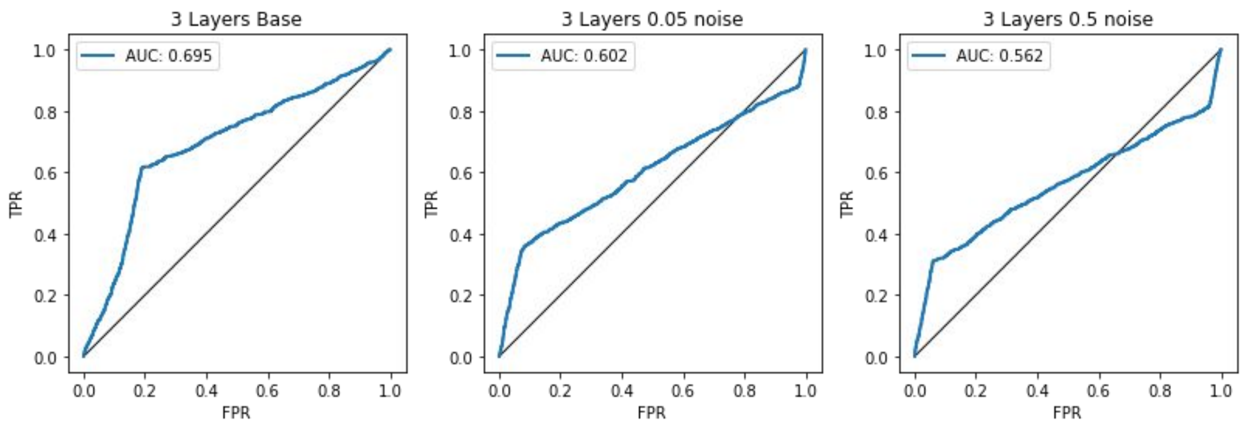
Experiment 1: Attacking Machine Learning Models



- Run MIA on Neural networks trained with varying levels of ε.
- What's the tradeoff between ε and model accuracy?

Experiment 1: Results

Model Name	Training Time (s)	ϵ	Test Accuracy (%)	Max AUC	Max Attacker Adv
3 Layers 0.5 noise	1003	10.7	0.33	0.56	0.26
3 Layers 0.1 noise	991	1398288	0.45	0.57	0.26
3 Layers 0.05 noise	986	10773104	0.53	0.60	0.29
3 Layers 0.01 noise	971	310773104	0.67	0.66	0.37
3 Layers 0.001 noise	968	31248273104	0.70	0.69	0.42
3 Layers Base	832	∞	0.70	0.69	0.42



Can we generate synthetic data that maintains the properties of the original dataset?

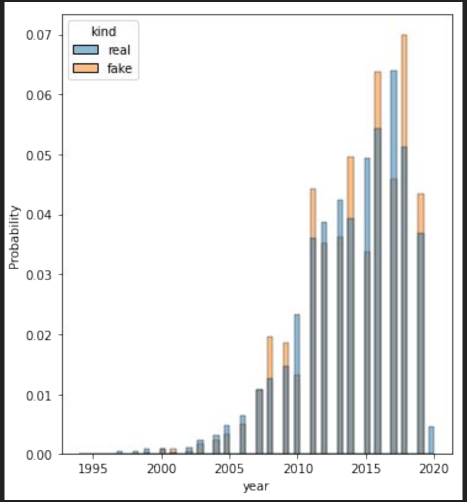
Experiment 2: Generating Synthetic Data

- How can we determine its success at replacing real data?

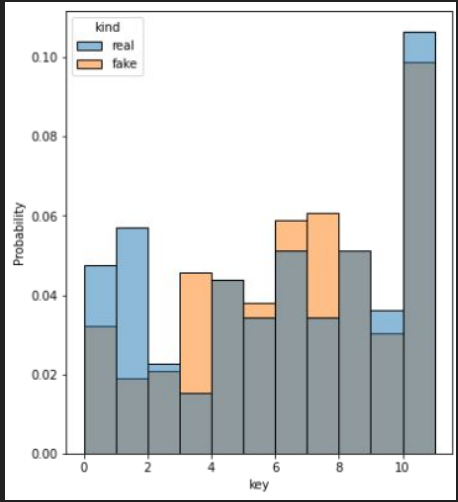
- **Metrics:**
  - Statistical
  - Likelihood
  - Detection
  - Efficacy
  - Privacy

- **Datasets:**
  - Vehicle
  - TikTok
  - Twitter

Experiment 2: Results



Vehicle dataset “year” feature



TikTok dataset “key” feature

Experiment 2: Results

Dataset	Statistical	Likelihood	Detection	Efficacy	Privacy	Training Time (min)
Vehicle	0.926 (KS)	0.99	0.92	0.49 / 0.64	0.07	42
TikTok	0.72 (KS)	0.99	0.71	0 / 0.08	0.24	11
Twitter	0.99 (CS) 0.94 (KS)	0.99	0.70	0.29 / 0.54	0.20	243

	ID	Game	Sentiment	Tweet
0	0	Fortnite	Irrelevant	great stream, today. thank you all for coming ...
1	1	TomClancysGhostRecon	Negative	CS:GO: Adds new bench on mirage. . Smileybs (o...
2	2	Overwatch	Positive	This is really interesting for indie RPGs with...
3	3	PlayerUnknownsBattlegrounds(PUBG)	Negative	RT @richardturrin: Amazon and Goldman partner....
4	4	Hearthstone	Positive	Miss U Pubg . . . .

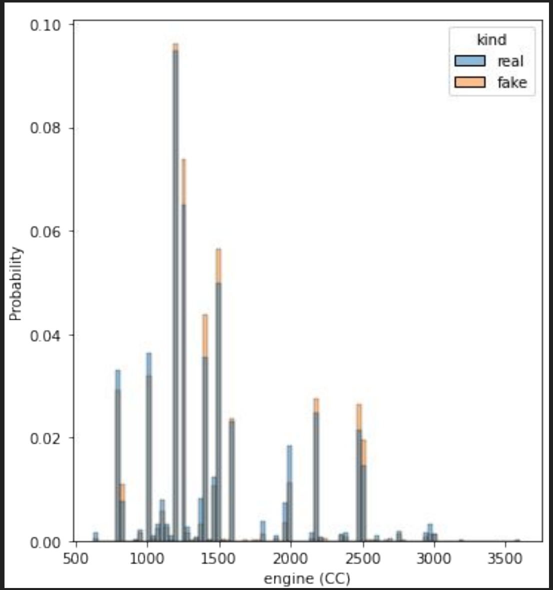
Some issues here…

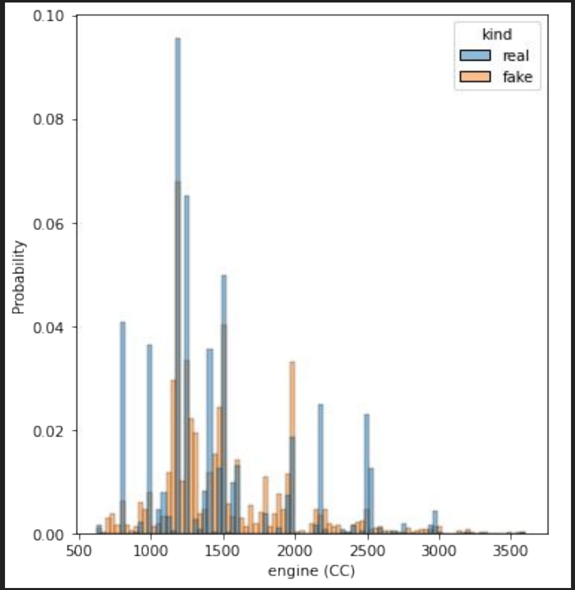
So we can make **accurate** synthetic data, but how about making it **private**?

Experiment 3: Generating Differentially Private Synthetic Data

- DPSGD in the training of the Generative Model
- Vary the level of noise in training.
- What's the tradeoff between ε and data accuracy?

Experiment 3: Results





0.001 vs 0.1

Experiment 3: Results

Dataset	Statistical	Likelihood	Detection	Efficacy	Privacy	Training Time (min)
0.001 DP	0.95 (KS)	0.99	0.89	0.61 / 0.64	0.09	15
0.01 DP	0.91 (KS)	0.99	0.78	0.34 / 0.64	0.12	16
0.1 DP	0.86 (KS)	0.99	0.72	0.1 / 0.64	0.14	18

	name	year	selling_price	km_driven	mileage (kmpl)	engine (CC)	max_power (bhp)	seats
0	K.t izleR XCpr	2017.0	1000000	50000.0	12.1	2179.0	79.0	5.0
1	VfgengNole Sag VX	2012.0	450000	10000.0	11.2	1428.0	72.4	5.0
2	CritXrga 1jGpro VSeS	2015.0	155000	15000.0	23.0	1449.0	81.6	5.0
3	Adslirziamrlrc Bl Vp	2011.0	370000	100000.0	23.0	1198.0	63.2	5.0
4	ycisf)dorNel Elvutle 1	2012.0	620000	15000.0	2.0	799.0	120.0	7.0

Outline

1. Motivation
2. Background
3. Experiments and Results
4. Discussion
5. Conclusion

So what’s the story here?

We showed Machine Learning models are easily attacked without protection…

But this protection does not come cheaply.

We showed how fake data can effectively replace real data…

But only from an accuracy standpoint

Can Differentially Private Synthetic Data fix these issues?

Discussion

Advantages:

- Maintains accuracy at low noise
- Privacy guarantees
- Infinite generation
- Privacy-preserving property of DP
- Negates the need for Machine Unlearning
- Low noise acts as regulariser

Disadvantages:

- Efficacy decays at high noise
- Training of the generative model

Outline

1. Motivation
2. Background
3. Experiments and Results
4. Discussion
5. Conclusion

Conclusion

Contributions

- Explored the advantages and disadvantages of current defence mechanisms against adversarial attacks.
- Analysed the tradeoff between ε (privacy budget) and accuracy.
- Assessed the effectiveness of Differentially Private Synthetic Data at :
  - Defending against adversarial attacks
  - Solving issues involved in “The right to be forgotten”

Future Work

Privacy

- Train models on DP Synthetic Data
- Run attacks on these new models to see if its possible to extract information from the original dataset
- Test the effectiveness of privacy mechanisms in the generation process (similarity and outlier filters)

Future Work

Data Types

- Construct DP Synthetic data on other data types such as images etc

Models

- Test DPSGD in other generative model types (e.g GANs for images)

Q&A