Data and Network Security

(Master Degree in Computer Science and Cybersecurity)

Lecture 6

Outline for today

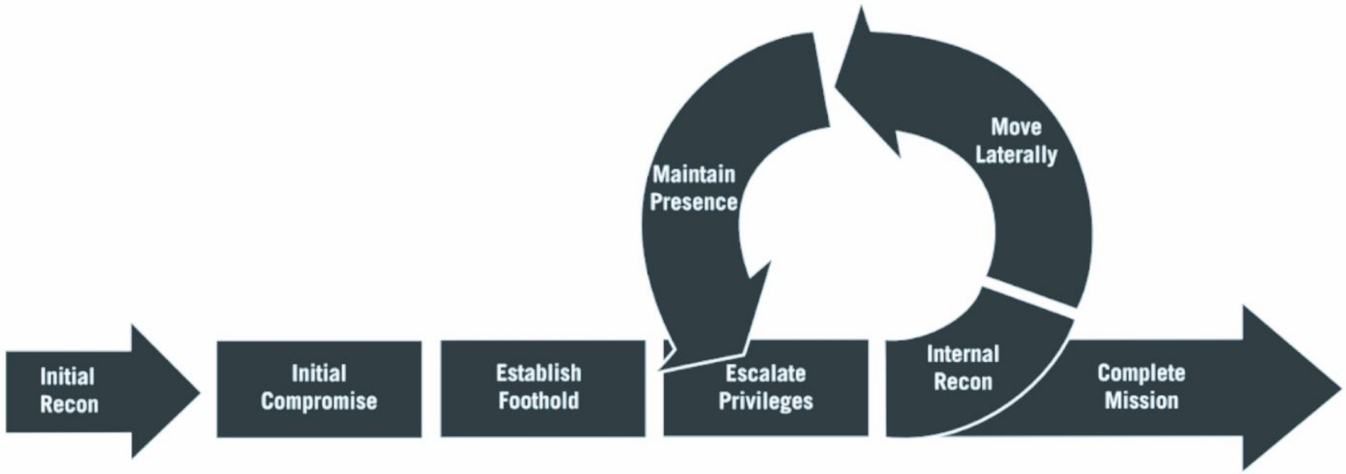
- Recap last lecture
- Property leakage from ML models
- Privacy preserving ML

Advanced Persistent Threats

Sophisticated, targeted cyberattack in which an unauthorized entity gains access to a network and remains undetected for an extended period of time.

- **APT attacks are characterized by:**
  - **advanced tactics,**
  - **stealthy infiltration methods,**
  - **persistent presence within the targeted network.**

APT - Life Cycle



Detecting APTs

To counteract this threat, an entity/organization needs to put some active defense mechanisms in place.

- **Cyber Threat Hunting**
  - Process that is put in place in order to tackle (hunt) this kind of sophisticated threat.

MITRE ATT&CK

Curated knowledge base and framework that categorizes the tactics, techniques, and procedures used by adversaries during cyber attacks.

Developed by MITRE Corporation, a nonprofit organization that operates federally funded research and development centers, ATT&CK provides a comprehensive taxonomy of cyber threats based on real-world observations and expert analysis.

Started in 2013 with the purpose of documenting common tactics, techniques and procedures against Windows enterprise networks and nowadays it spans almost all main enterprise solutions and also provides mitigations strategies.

MITRE ATT&CK - The TTP trio

Tactics, Techniques and Procedures

MITRE ATT&CK - The TTP trio

Tactics

Tactics represent the high-level objectives or goals that adversaries aim to achieve during a cyber attack. They describe the strategies employed by attackers to accomplish their mission.

Example:

- **gaining initial access to a target network,**
- **establishing persistence,**
- **escalating privileges,**
- **exfiltrating data,**
- **disrupting operations.**

Tactics serve as the primary categories for organizing and classifying adversary behavior.

MITRE ATT&CK - The TTP trio

Techniques

Techniques are the specific methods or procedures used by adversaries to achieve each tactic. They describe the step-by-step actions taken by attackers to accomplish their objectives.

Example:

Techniques under the "initial access" tactic may include:

- **phishing emails,**
- **exploiting software vulnerabilities,**
- **leveraging stolen credentials to gain entry into a target network**

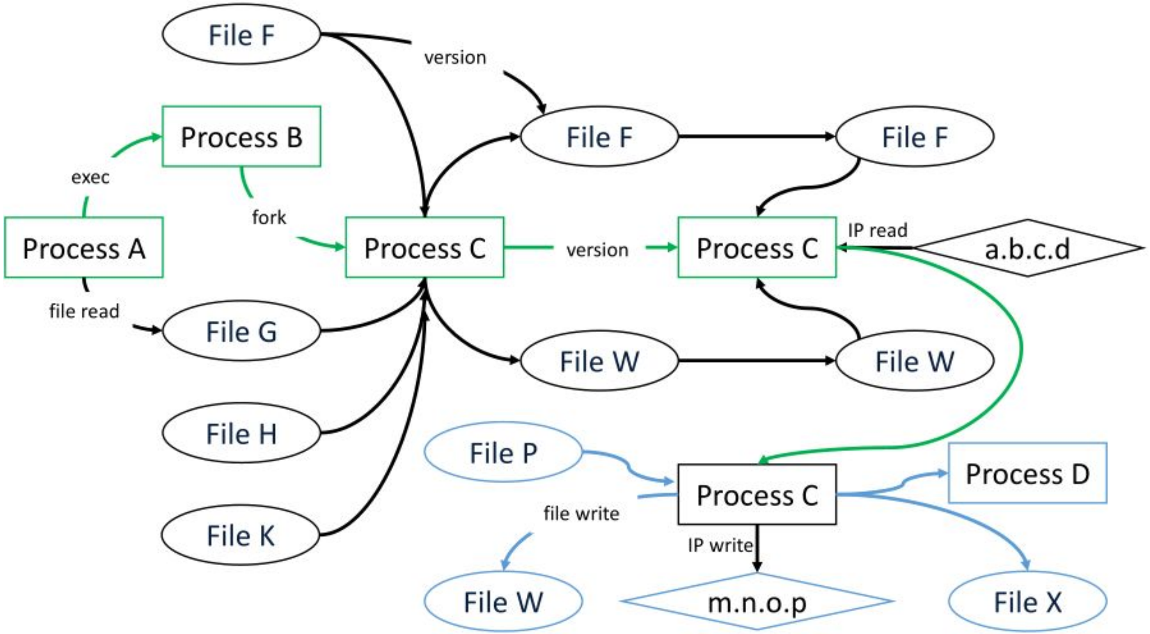
MITRE ATT&CK - The TTP trio

Procedures (sub-techniques)
Variations or specific implementations of techniques that further refine the behaviors observed in cyber attacks.
They provide additional granularity and detail to techniques, allowing for more precise analysis and detection of adversary activity. Procedures describe specific ways in which techniques are executed or customized by attackers to suit their objectives or adapt to the target environment.

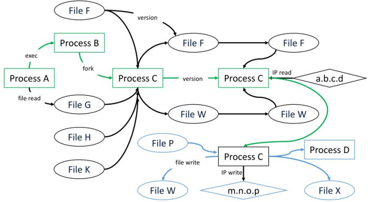
Example:
A procedure under the "exploitation of remote services" technique may involve exploiting a specific vulnerability in a web server software to gain unauthorized access.

What is needed? Provenance

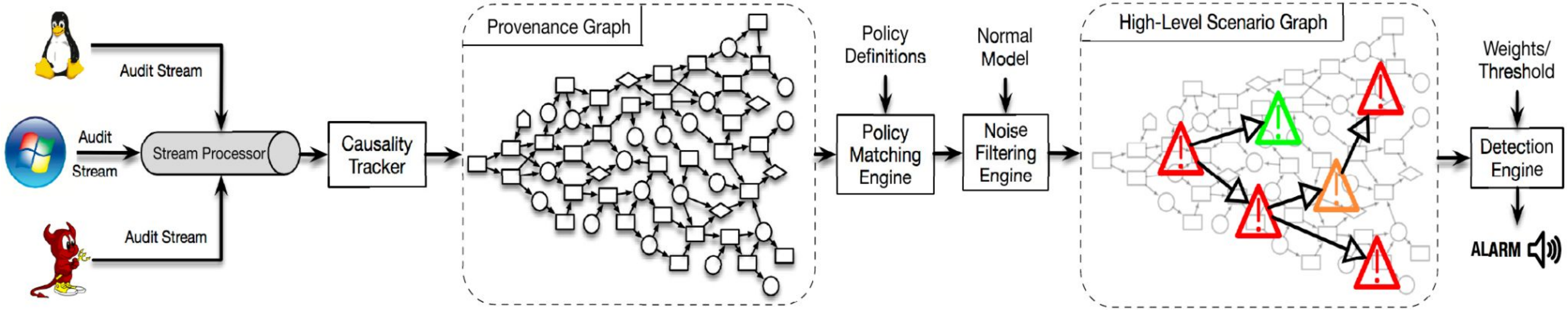
Provenance Graphs: Represent system execution as a Directed Acyclic Graph that describes information flow and causality (edges) between kernel objects (vertices, e.g., processes, files, sockets).



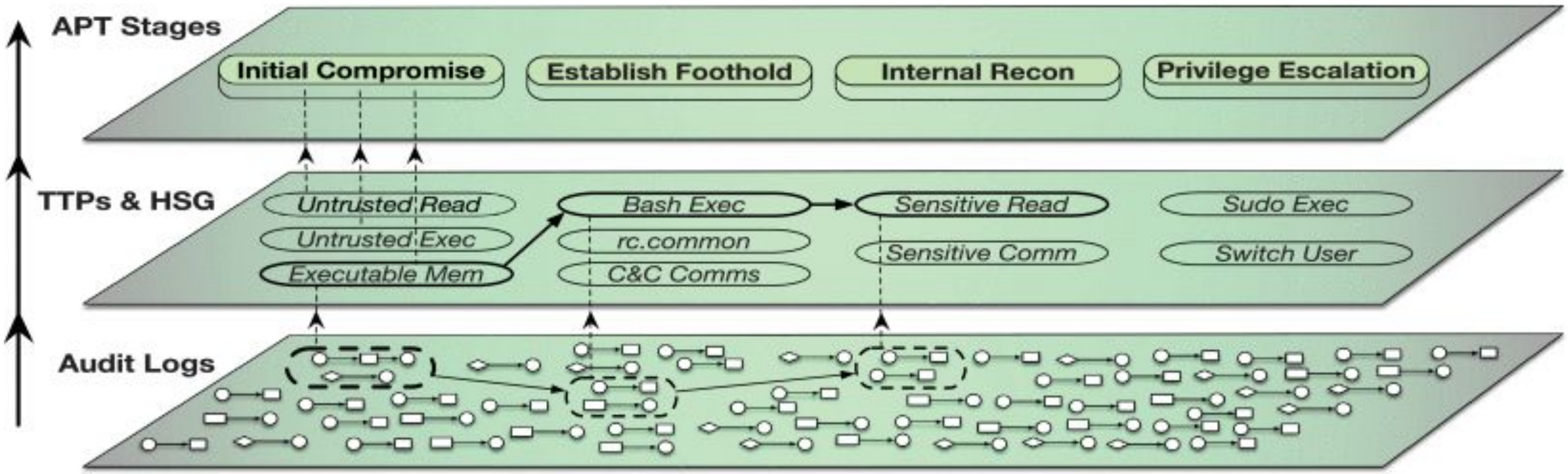
Approaches relying on provenance graphs



Aim: Generation or a high-level graph that represents the attacker actions and thus makes it easier to spot and mitigate (possibly in real-time).



Mapping the information



Utilize Mitre ATT&CK framework to map low-level system events to an intermediate high-level representation that can be then easily mapped to an APT campaigns' phases.

Outline for today

- Recap last lecture
- Information leakage from ML models
- Privacy preserving learning issues

Sensitive property

DATASET

- **Demographic Information:**
  - Age, gender, ethnicity, income level.
- **Behavioral Patterns:**
  - Shopping habits, browsing history, social interactions.
- **Personal Preferences:**
  - Political affiliations, health conditions, lifestyle choices.

Sensitive property

DATASET

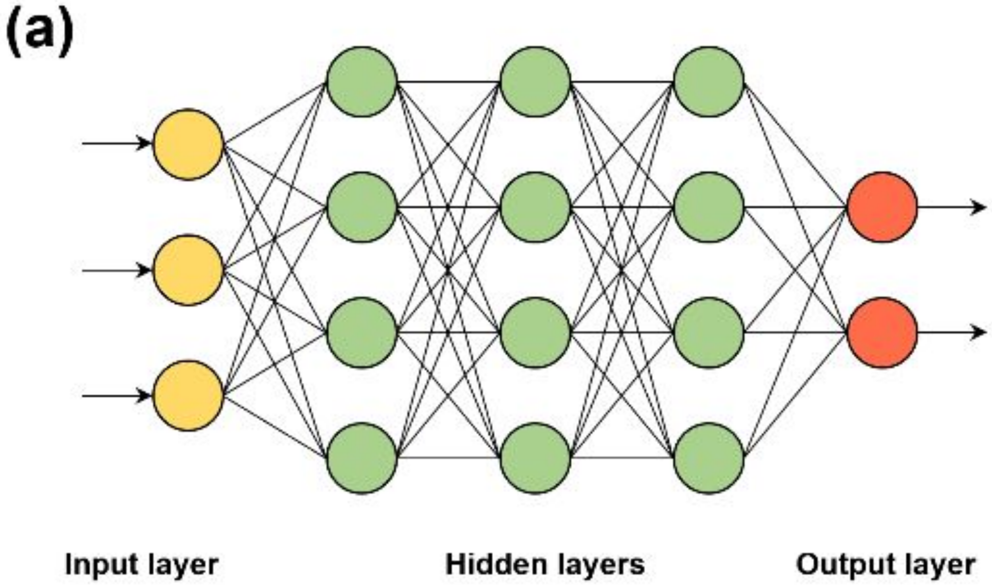
- Demographic Information:
  - Age, gender, ethnicity, income level.
- Behavioral Patterns:
  - Shopping habits, browsing history, social interactions.
- Personal Preferences:
  - Political affiliations, health conditions, lifestyle choices.

Disclosure of such properties can lead to privacy breaches, discrimination, or manipulation of individuals.

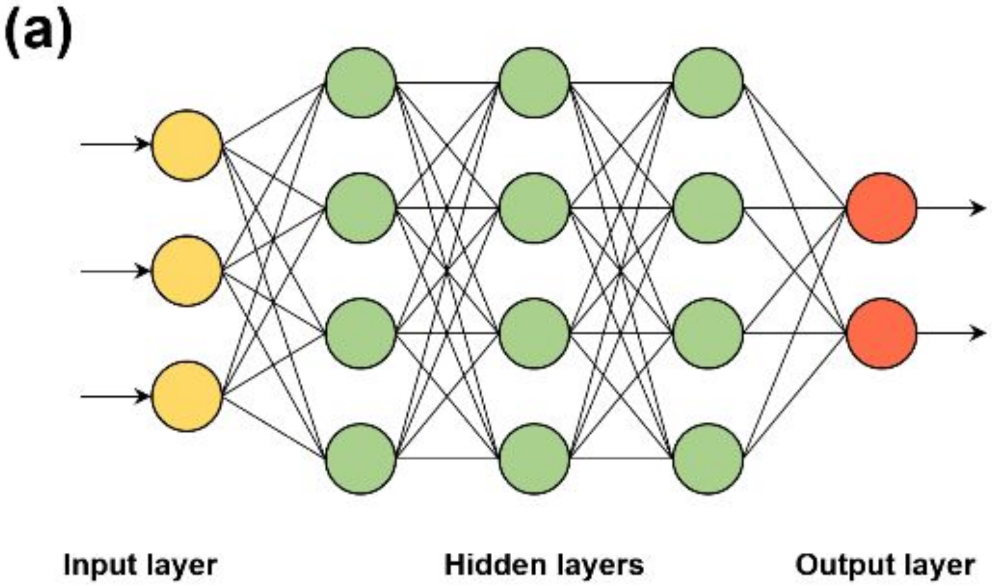
Information leakage from ML models

Can I infer some (sensitive) property of the dataset used to train an ML model?

What can ML models tell?

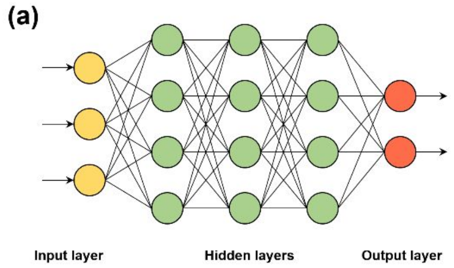


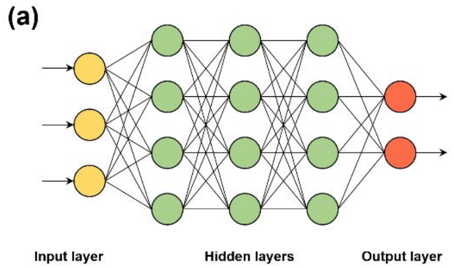
Does this dataset have this property?



But you have only the model…

Does this dataset have this property P?

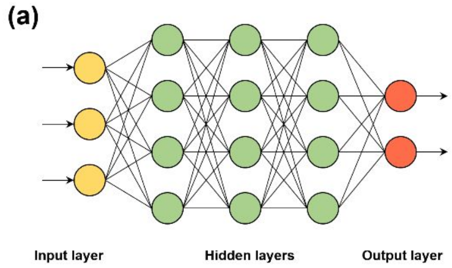




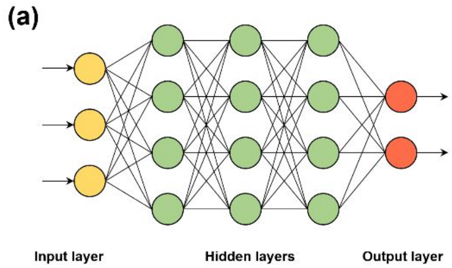
Construct N different ML models, similar to the target model.

- The dataset of some has property P, and the others dont.

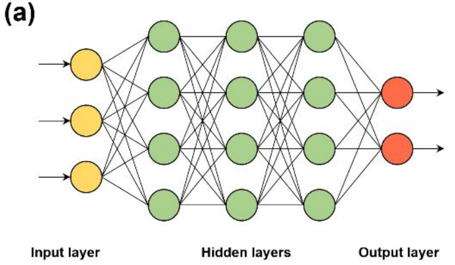
Does this dataset have this property?



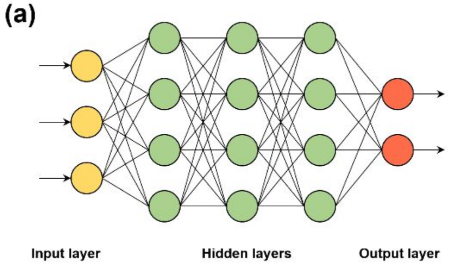
Represent all these ML models as a “feature vector”



Does this dataset have this property?

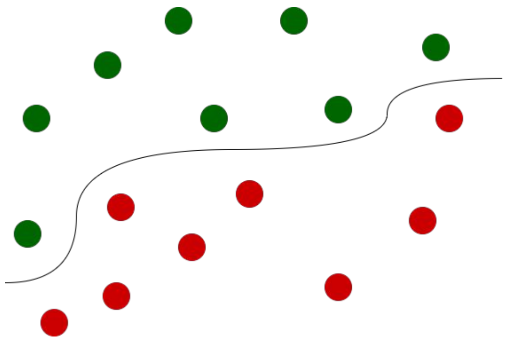


[f1,f2,f3,...fx]

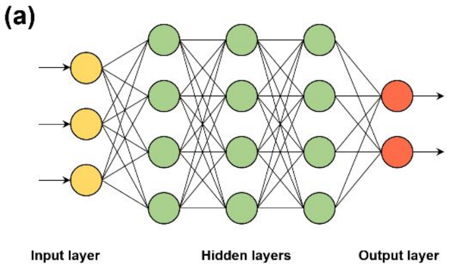


[f1,f2,f3,...fx]

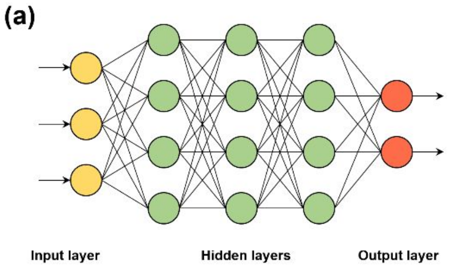
Train a binary classifier on these features



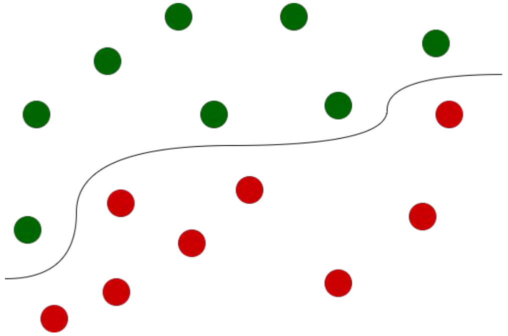
Does this dataset have this property?



[f1,f2,f3,...fx]



[f1,f2,f3,...fx]



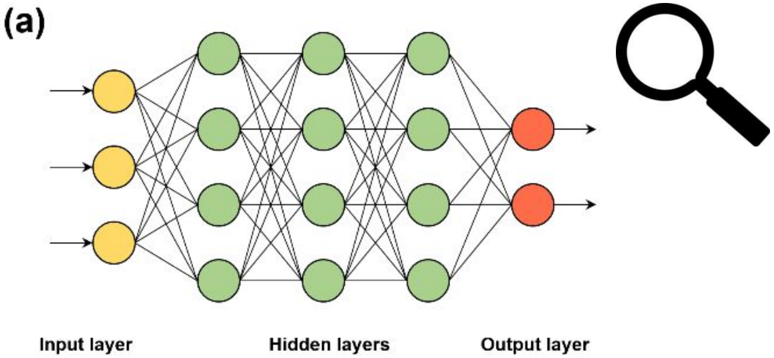
Represent and test the target classifier.

Information leakage from ML models

Was this datapoint part of this dataset?

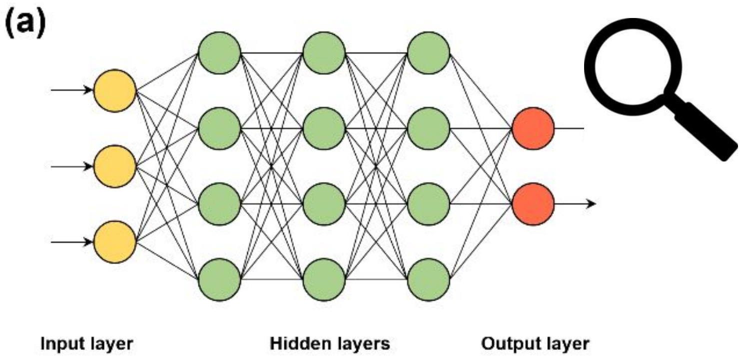
Information leakage from ML models

The model's responses provide valuable information that adversaries can leverage to infer whether a particular data point was part of the training dataset.



Information leakage from ML models

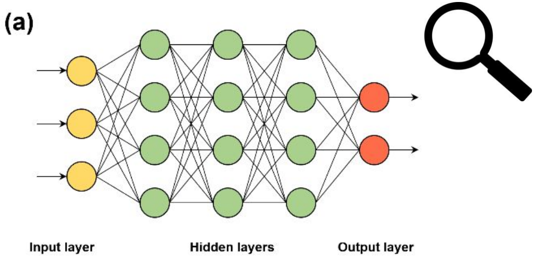
During the training phase, a machine learning model learns to generalize patterns and relationships from the training dataset to make predictions on unseen data. As a result, the model's behavior may vary depending on whether it has seen a particular data point during training.



Information leakage from ML models

When we query the model with input data, we observe the model's responses:

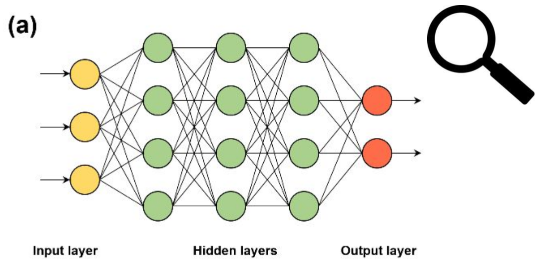
- predicted labels,
- probabilities, or scores assigned to different classes.



Information leakage from ML models

When we query the model with input data, we observe the model's responses:

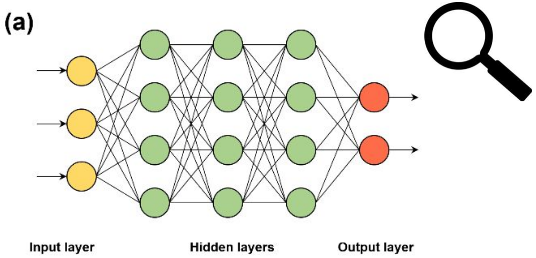
- **predicted labels,**
- **probabilities, or scores assigned to different classes.**



*distinguishing features in these responses that can indicate whether the input data was likely part of the training dataset

Distinguishing features?

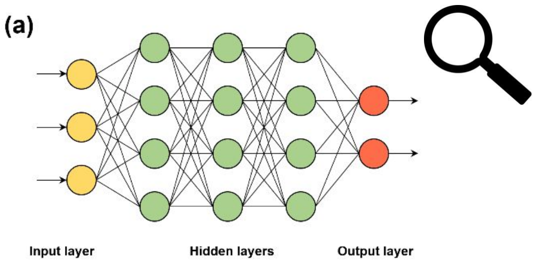
One key indicator that a data point was part of the training dataset is overfitting.



Distinguishing features?

One key indicator that a data point was part of the training dataset is overfitting.

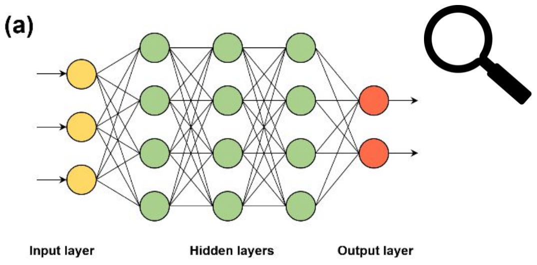
Overfitting occurs when a model learns to memorize specific examples from the training data rather than capturing general patterns.



Distinguishing features?

One key indicator that a data point was part of the training dataset is overfitting.

Overfitting occurs when a model learns to memorize specific examples from the training data rather than capturing general patterns.

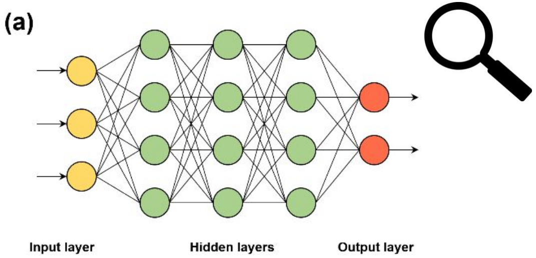


If a model exhibits overfitting, it may produce responses that are overly confident or precise for data points seen during training but less accurate for unseen data.

Distinguishing features?

Another indicator - confidence discrepancy.

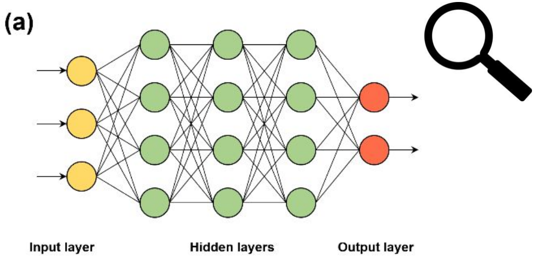
If the model's confidence is significantly higher for certain inputs compared to others, it may suggest that those inputs were present in the training dataset.



Distinguishing features?

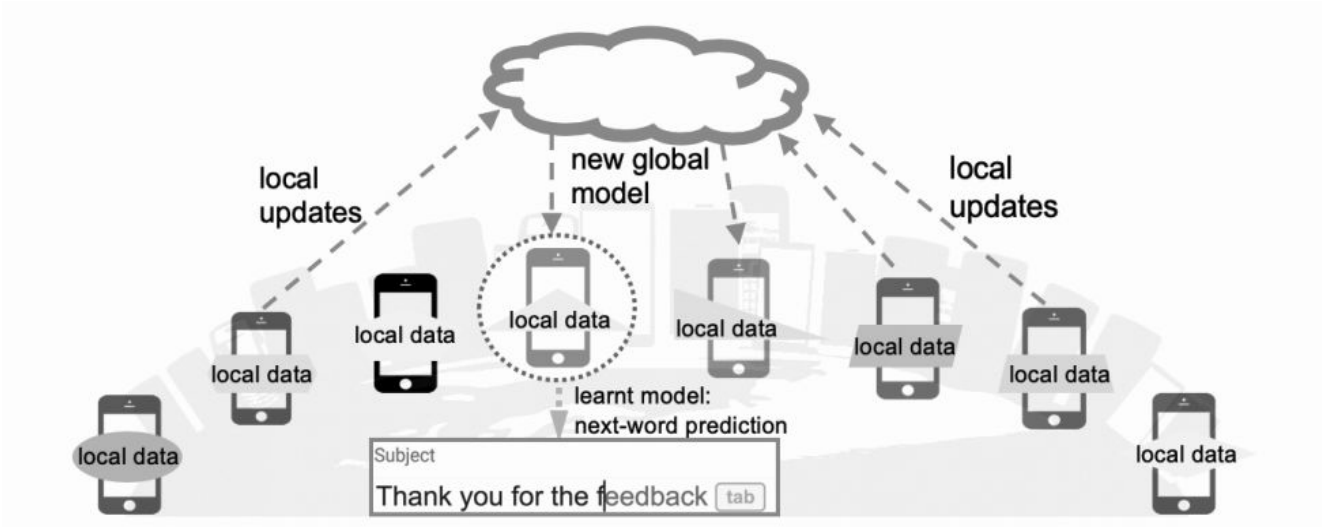
Another indicator - bias in model responses.

If the model consistently predicts certain classes or outputs for specific inputs, it may indicate that those inputs were overrepresented in the training dataset.



Is there any leakage in privacy-preserving learning?

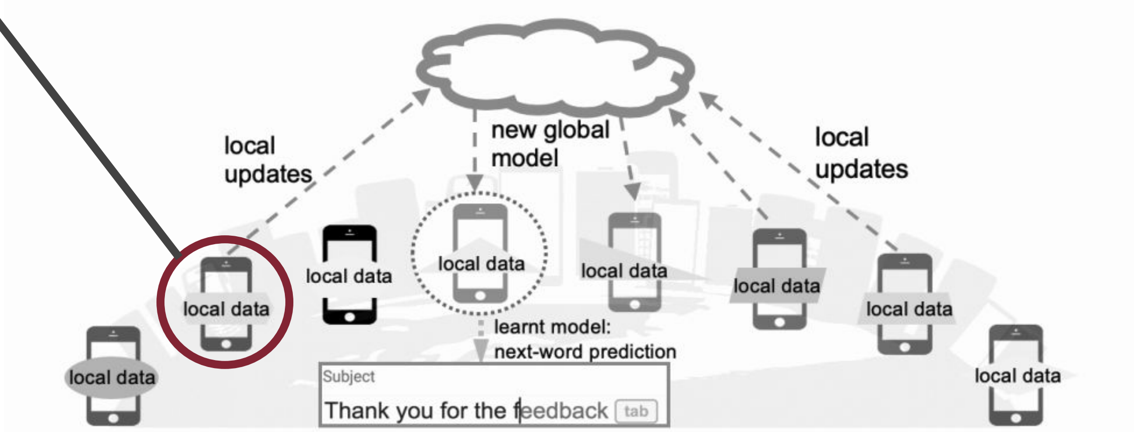
Collaborative learning



Federated learning (FL) (also known as **collaborative learning**) is a machine learning technique that trains an algorithm across multiple decentralized edge devices or servers holding local data samples, without exchanging them.

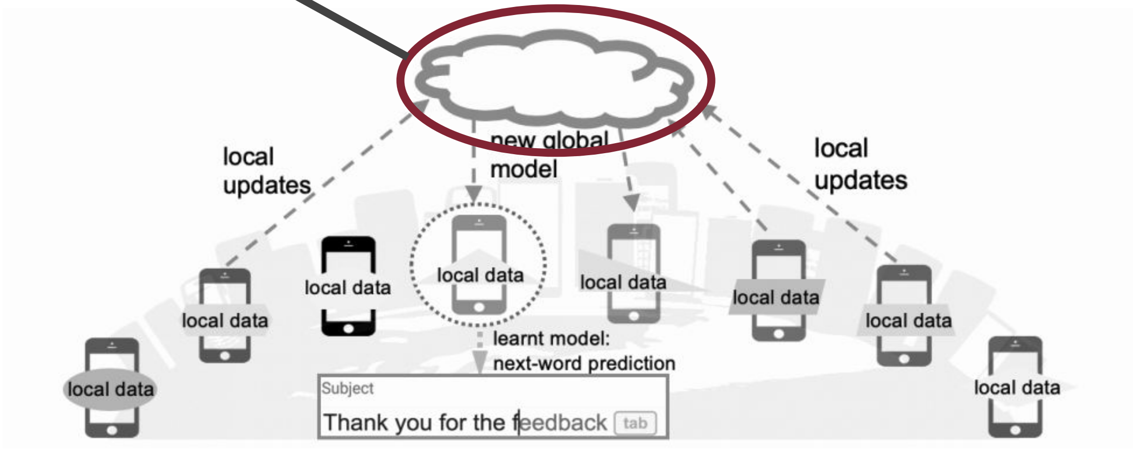
How does FL work? (local update)

$$\mathbf{W}_{t+1}^{k} = \mathbf{W}_t + \alpha \nabla \mathbf{W}_t^{k}$$

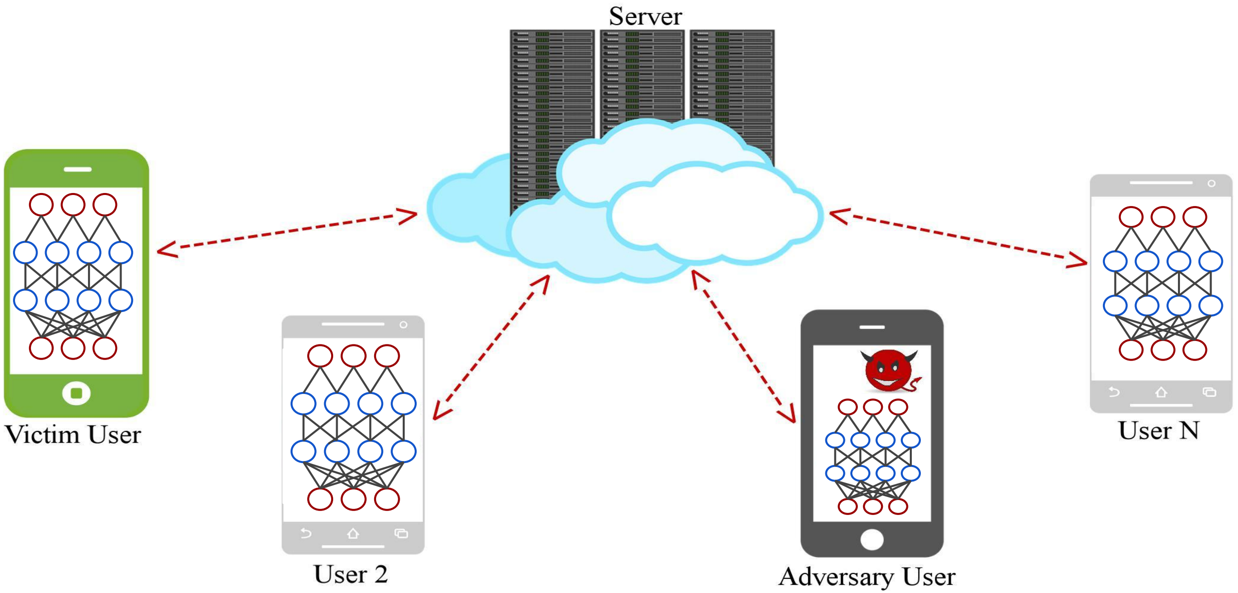


How does FL work? (global update)

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \frac{\alpha}{n'} \sum_{k=1}^{n'} \nabla \mathbf{W}_t^k$$



Collaborative Learning Scheme



Adversary's goal?
Reconstruct private samples from the dataset of the victim indirectly influencing the learning of other participants

How can we reconstruct samples of other participants training data by looking at some gradients?

How should the adversary behave?

- The adversary should operate as an participant within the privacy-preserving collaborative deep learning protocol.

- The objective of the adversary is to infer meaningful information about a label that he does not own.

- The adversary does not compromise the global parameter server that collects and distributes parameters to the participants.

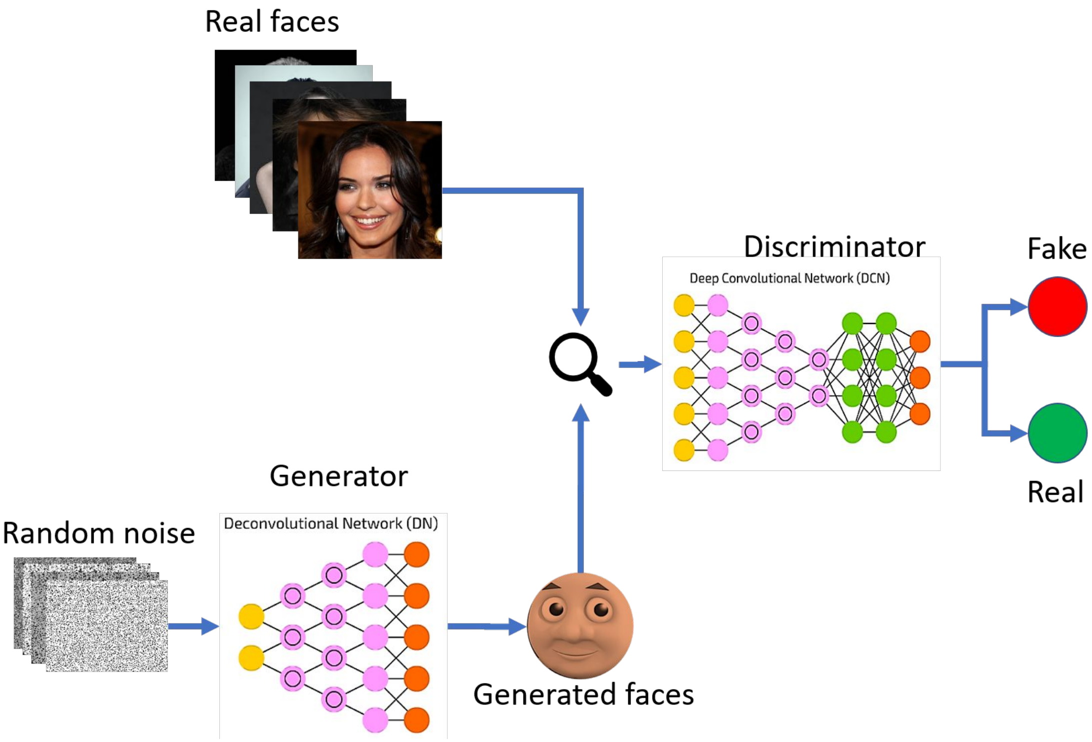
What can the adversary use?

What can the adversary use?

Give me DATA…

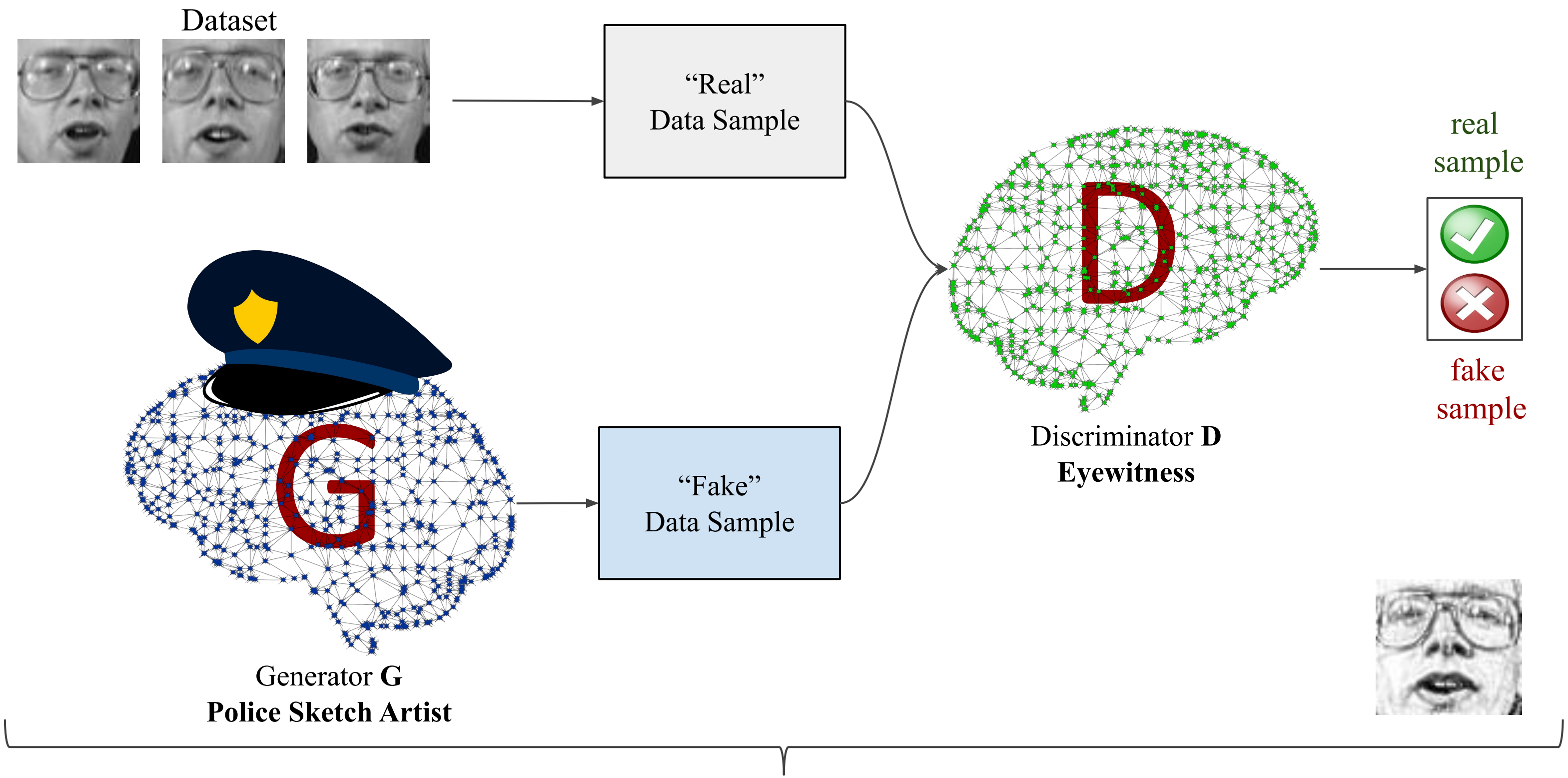
Not this..

This...



Real faces
Discriminator
Deep Convolutional Network (DCN)
Fake
Real
Generator
Random noise
Deconvolutional Network (DN)
Generated faces

Generative Adversarial Network



GAN Architecture

Generative Adversarial Networks

Generative

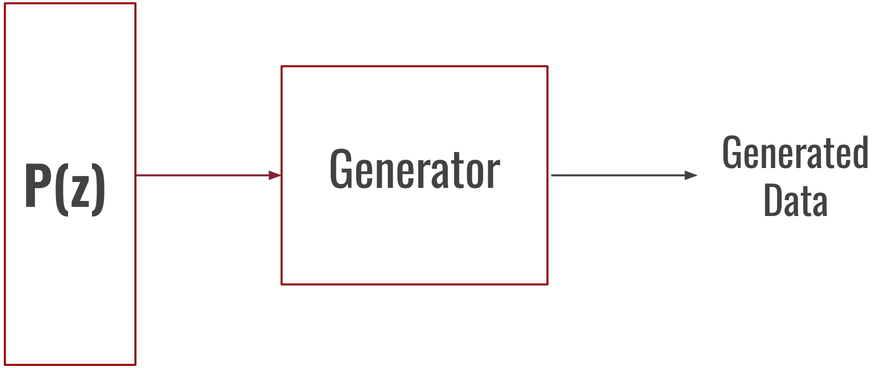
- We try to learn the underlying the distribution from which our dataset comes from.

Generative Adversarial Networks

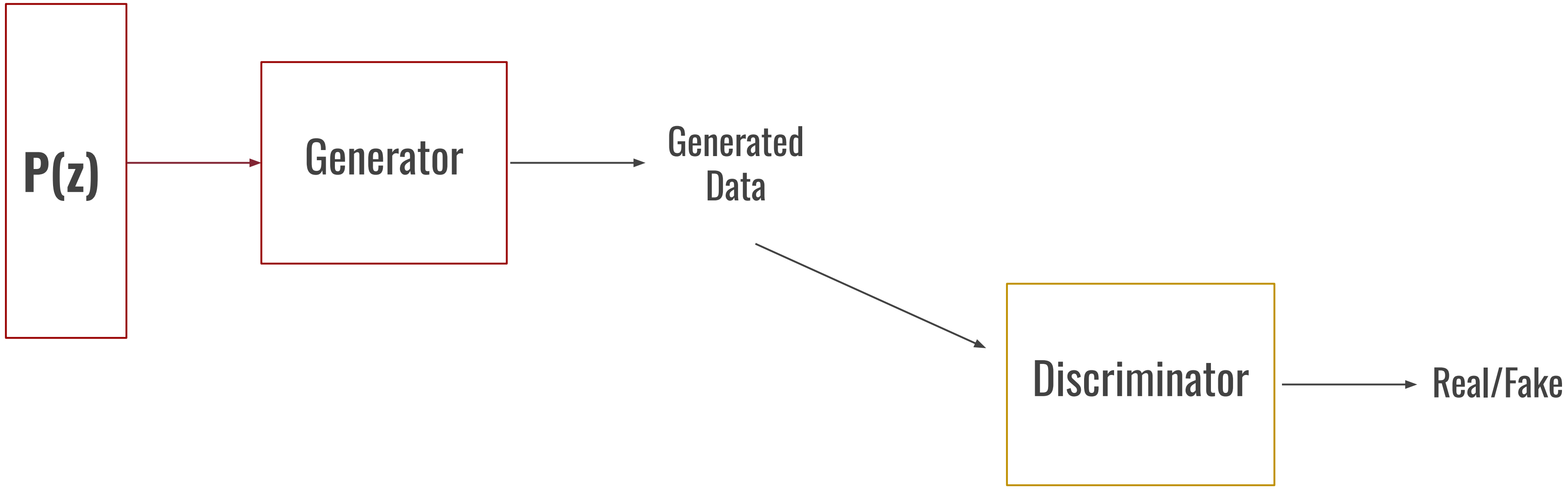
Adversarial

- GANS are made up of two competing networks (adversaries) that are trying beat each other.

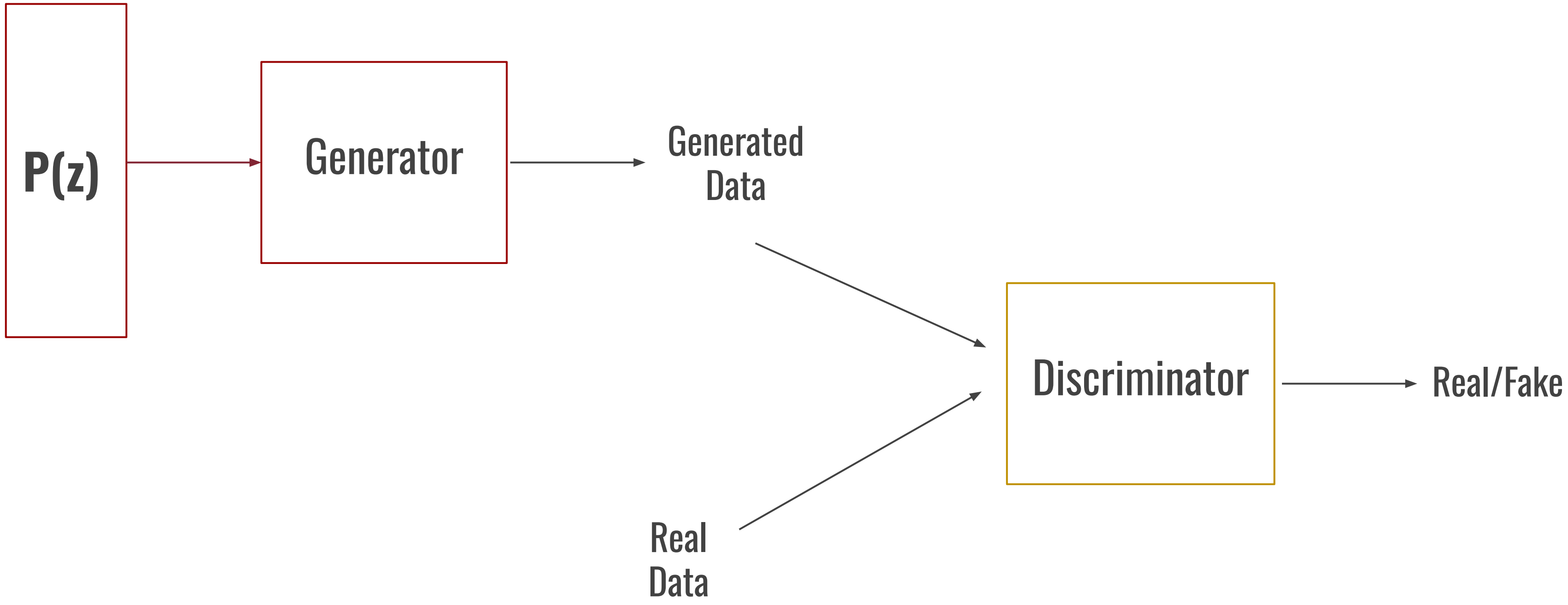
Generative Adversarial Networks



Generative Adversarial Networks

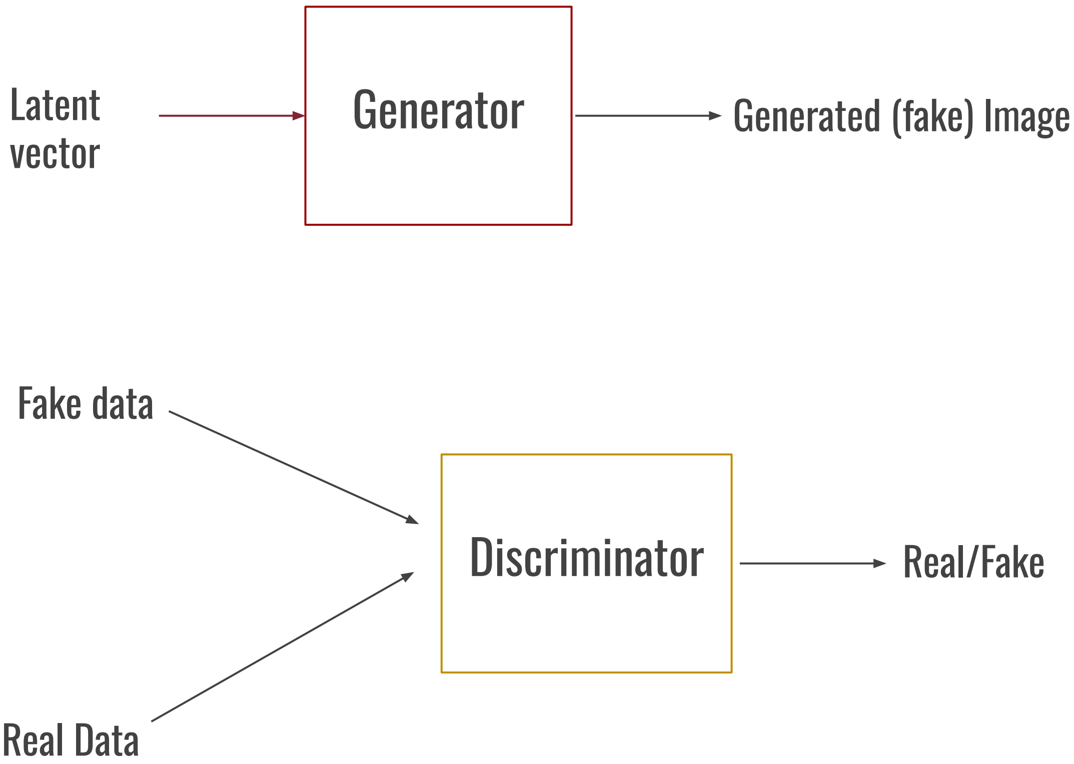


Generative Adversarial Networks



GANs - How are they trained?

At t=0,



GANs - How are they trained?

Which one should I train first?

GANs - How are they trained?

Which one should I train first?

Discriminator

GANs - How are they trained?

With what training data though?

Discriminator

GANs - How are they trained?

With what training data though?

Discriminator

- The Discriminator is a Binary classifier.
- Needs to discriminate between real/fake

GANs - How are they trained?

With what training data though?

Discriminator

- The Discriminator is a Binary classifier.
- Needs to discriminate between real/fake
- The data for Real class if already given:
  - **THE TRAINING DATA**
- The data for Fake class?
  - Generate from the **Generator**

GANs - How are they trained?

What about the Generator?

GANs - How are they trained?

What about the Generator?

Learning objective: Generate images from the Generator such that they are classified incorrectly by the Discriminator.

GANs - How are they trained?

Discriminator

Train the Discriminator using the current ability of the Generator

GANs - How are they trained?

Discriminator

Train the Discriminator
using the current
ability of the Generator

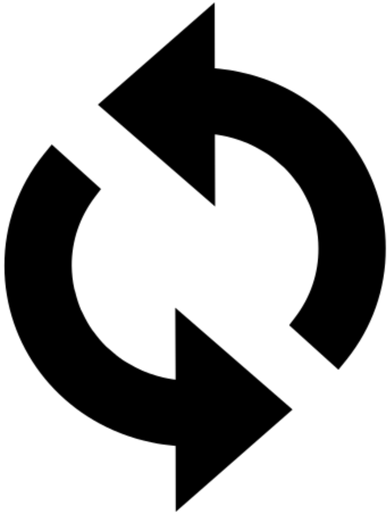
Generator

Train the Generator
to beat
the Discriminator

GANs - How are they trained?

Discriminator

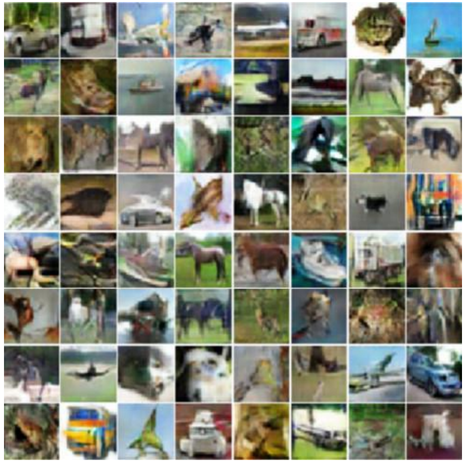
Train the Discriminator
using the current
ability of the Generator



Generator

Train the Generator
to beat
the Discriminator

MNIST images



CIFAR-10 images

faces

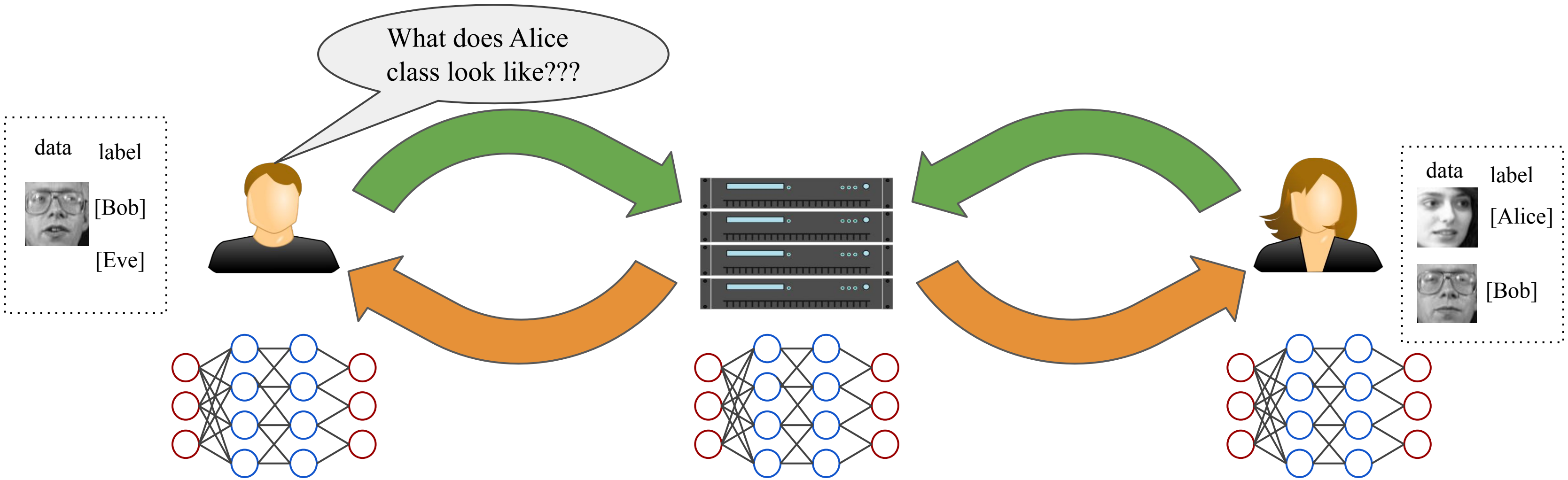
album covers

GAN results in the literature

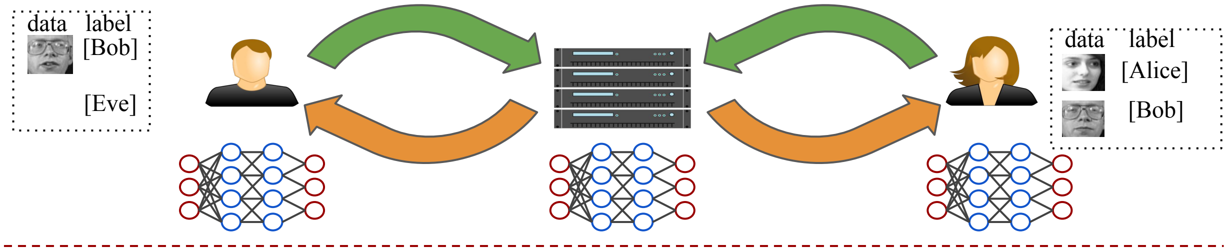
images from:
- https://blog.openai.com/generative-models/
- Goodfellow et al. Generative Adversarial Networks
- Radford et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks

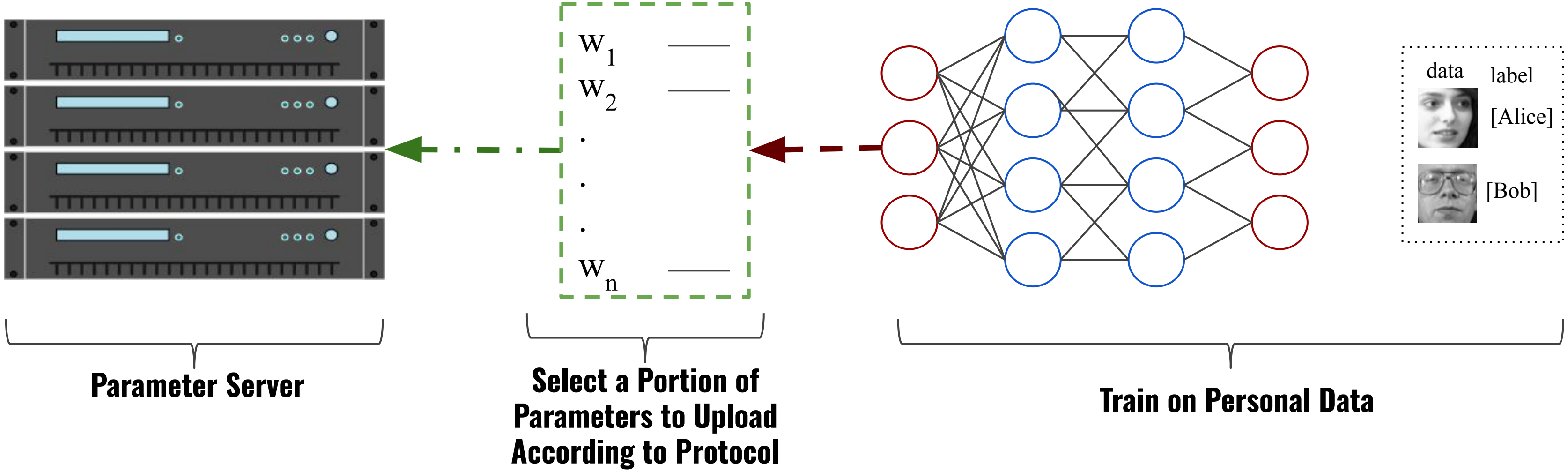
bedrooms

Violating the privacy...

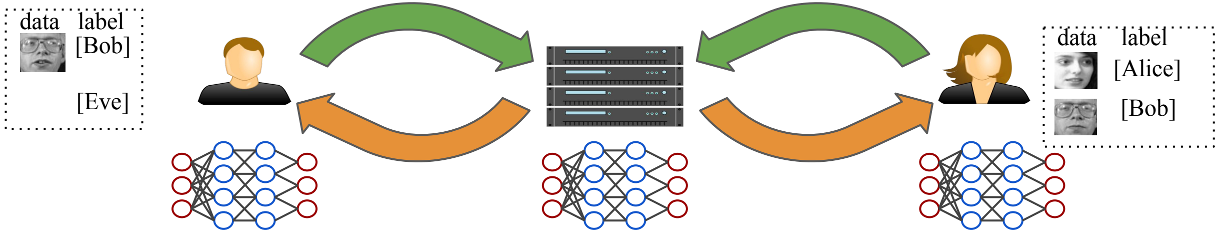


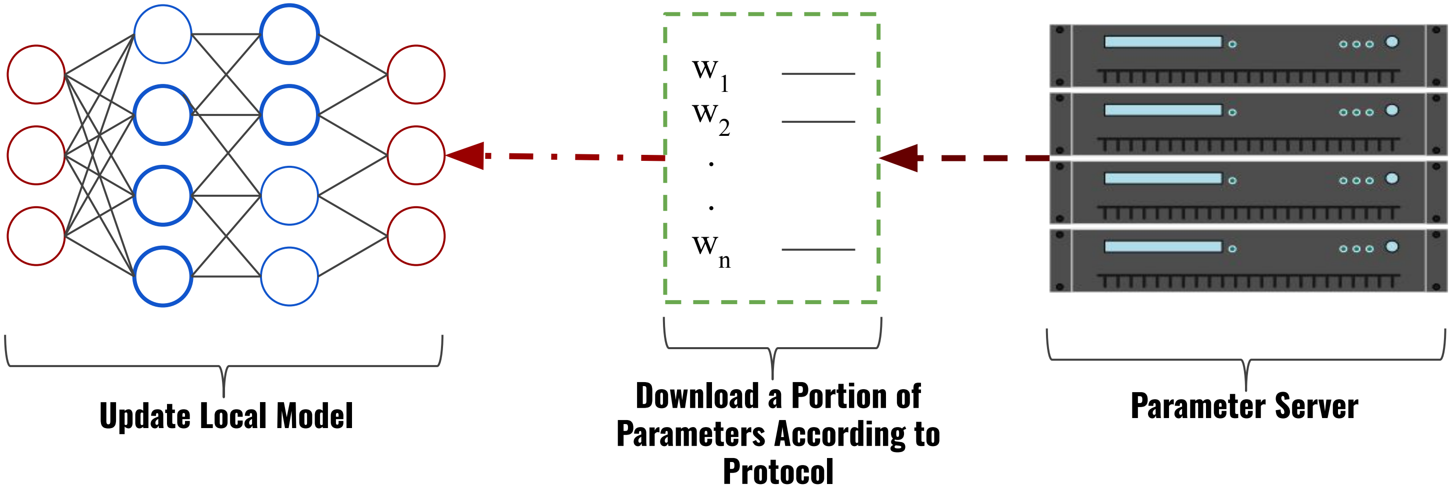
Victim's Turn



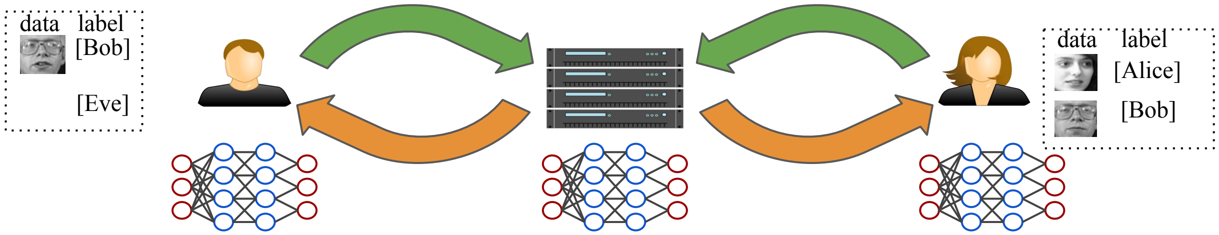


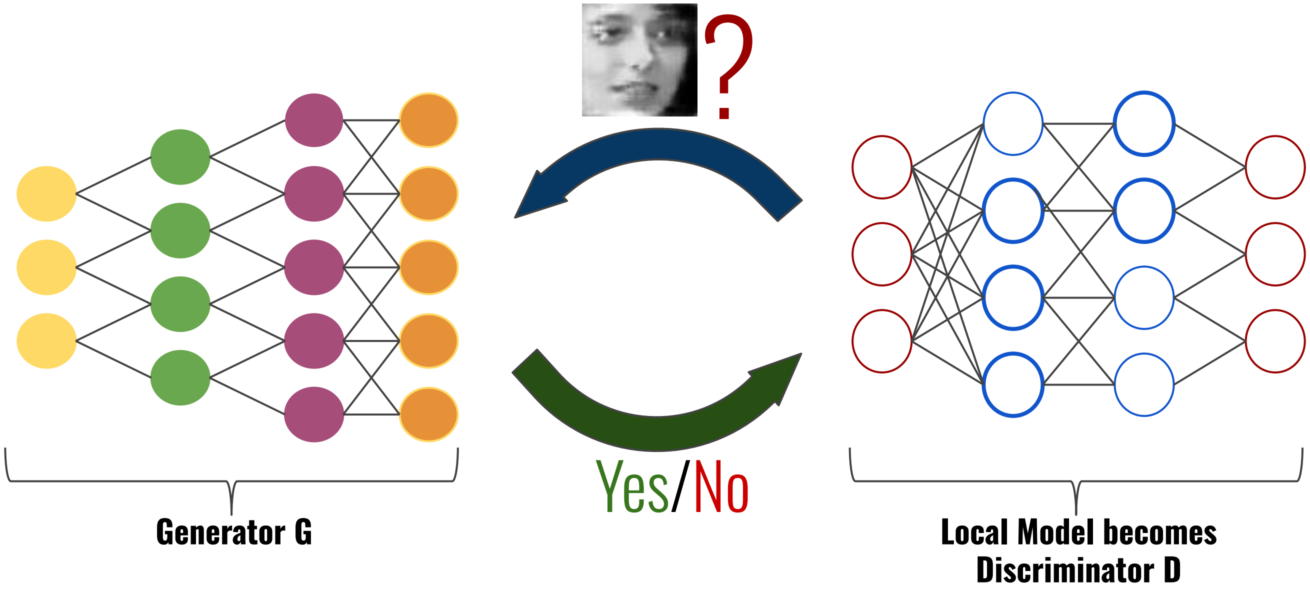
Adversary's Turn



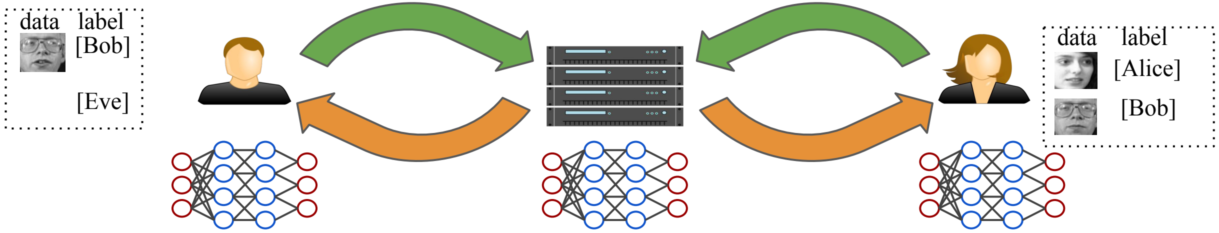


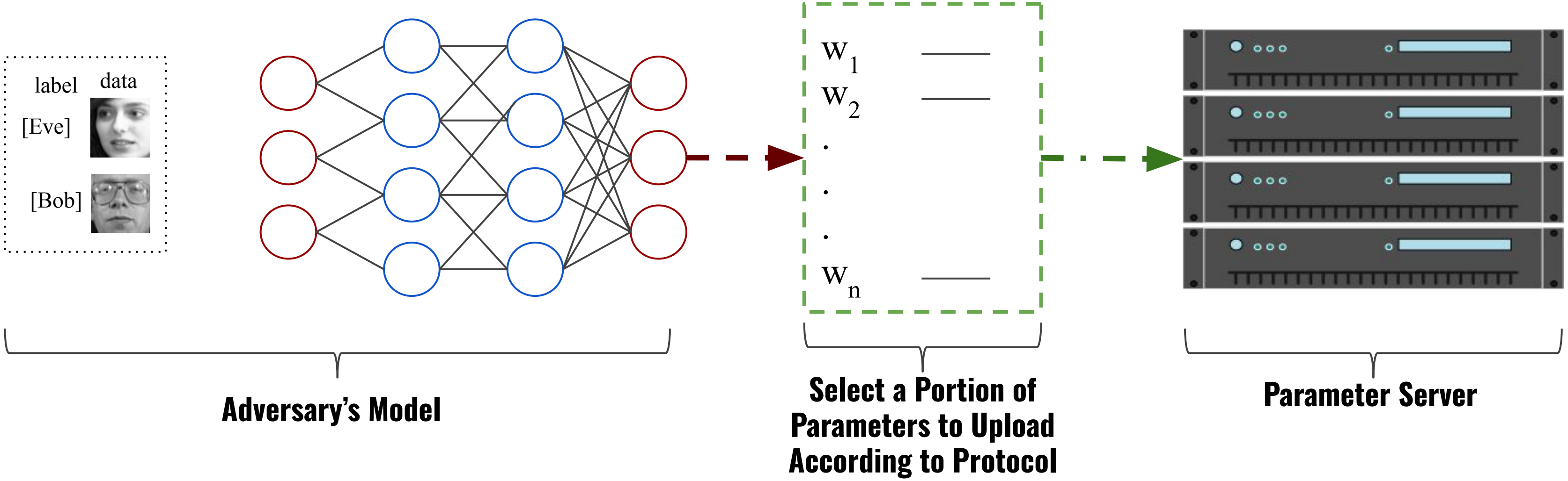
Adversary's Turn





Adversary’s Turn





Experiments without Differential Privacy

Actual Images

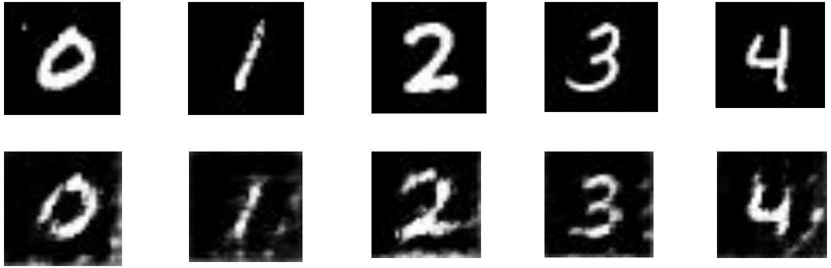
Generated Data

Original vs Generated

Experiments with Differential Privacy

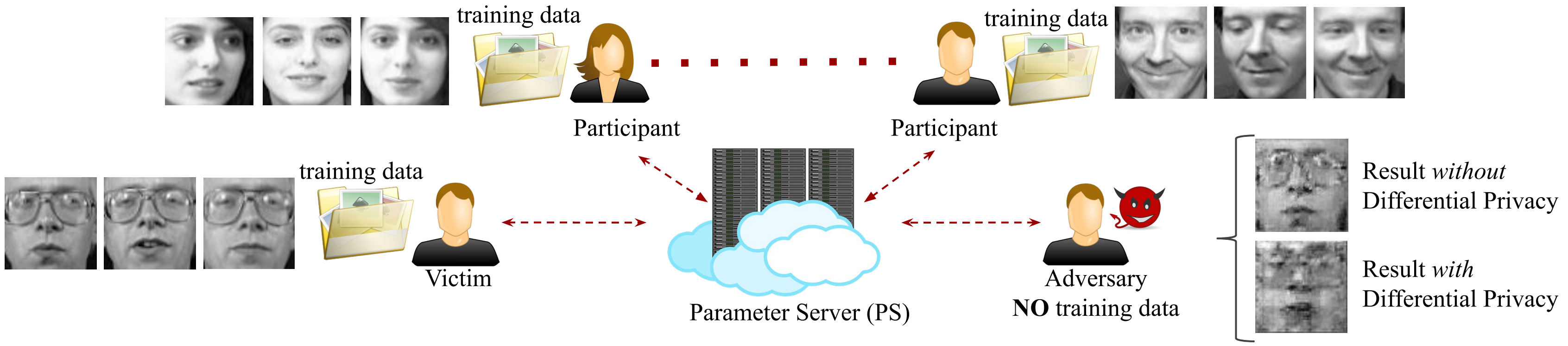
Actual Images

Generated Data



Original vs Generated

Experiments (Adversary has NO data at all)



Reading Material

1. Privacy preserving learning: Link-1, Link-2
2. Generative Adversarial Networks: Link-1
3. Information Leakage from collaborative learning: Link-1