Deep Models under the GAN:
Information Leakage from Collaborative Deep Learning

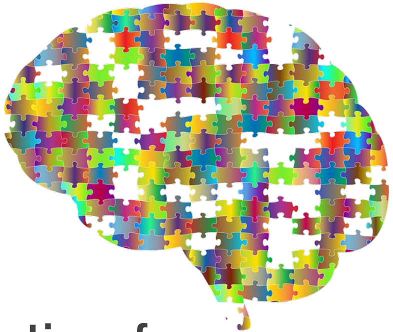
Briland Hitaj, Giuseppe Ateniese, Fernando Perez-Cruz
Computer Science Department
Stevens Institute of Technology

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
1870

We investigate whether decentralized (a.k.a collaborative) deep learning is *more* privacy-preserving than the centralized one.

...

New active inference attack on collaborative deep learning models using GANs



Deep Learning 101

Branch of machine learning that makes use of neural networks, to find solutions for a variety of complex tasks either in supervised or unsupervised way

- Areas used:
  - Computer vision
  - Image processing
  - Face recognition
  - Speech recognition
  - Text-to-speech systems
  - Natural language processing
  - Games...

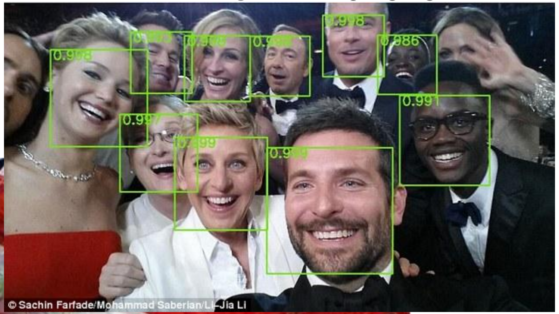
AlphaGo

image source: https://deepmind.com/research/alphago/

DeepMind Health

image source: https://goo.gl/u26HM3

image source: https://goo.gl/xNwTVw



Deep Learning

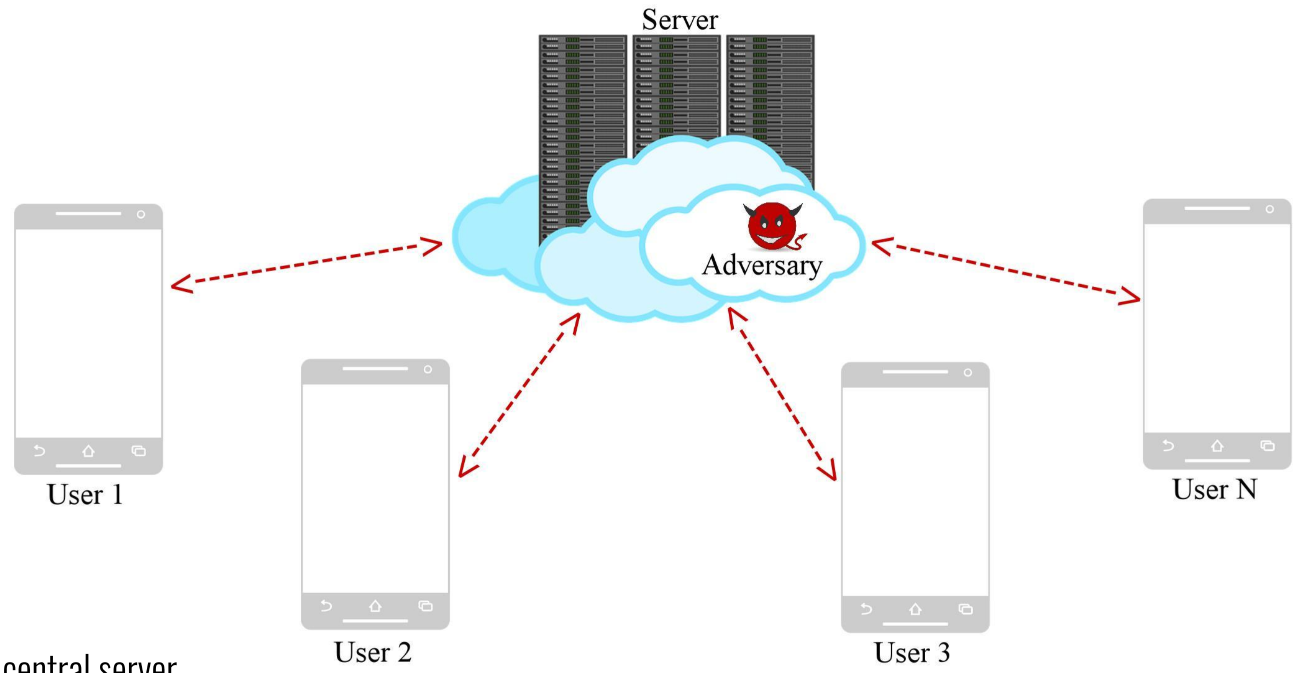
Huge computational power

image source: http://www.nvidia.com/docs/IO/67561/GeForce_GTX_280M_preview.jpg

Large quantities of data

image source: http://karpathy.github.io/assets/cnntsne.jpeg

Centralized Learning Scheme



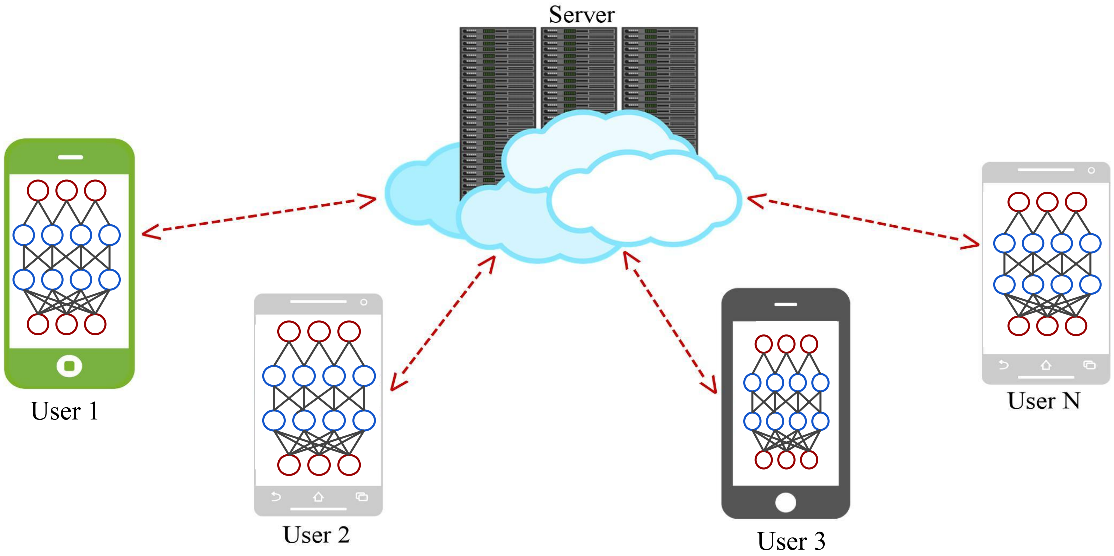
- data pooled on a central server
- no control over the learning process

Adversary lies only on the central authority (entity providing the service)

is it possible to learn while preserving privacy

???

Decentralized Learning Scheme (collaborative/federated)



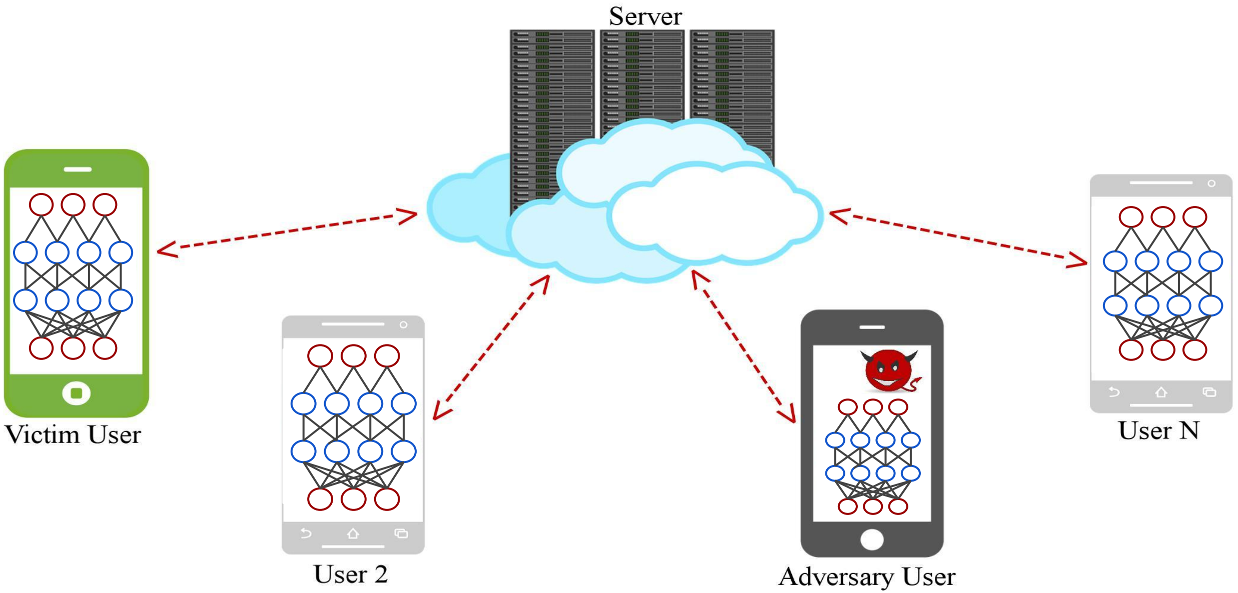
Shokri et al. Privacy-Preserving Deep Learning, CCS'15

- local training on data
- updates shared via a parameters server
    (ex. Google's federated learning)
- participants **indirectly** influence each other's learning
- differential privacy can be used to minimize leakages from parameter uploads
- NOTE: Shokri et al. only assume a "passive" adversary model

Attacks on ML models (Prior to Ours)

1) Hacking Smart Machines with Smarter Ones, 2011 by Ateniese et al.
2) Model Inversion Attacks, 2015 by Fredrikson et al.
3) Membership Inference Attacks, 2017 by Shokri et al.

Decentralized Learning Scheme (collaborative/federated)



- Indirectly influencing the learning of other participants, allows potentially anyone to be an adversary

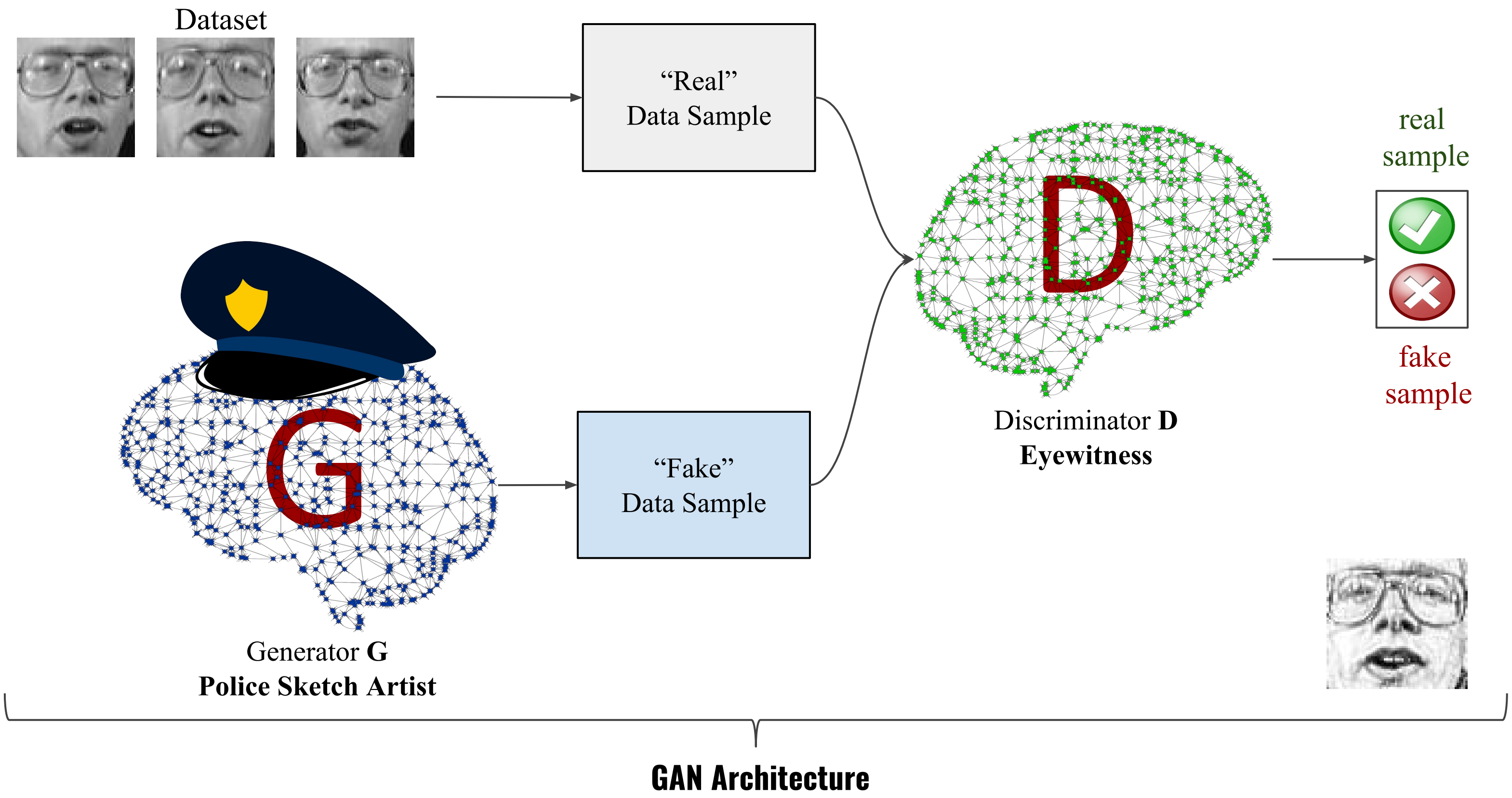
can neural networks attack neural networks

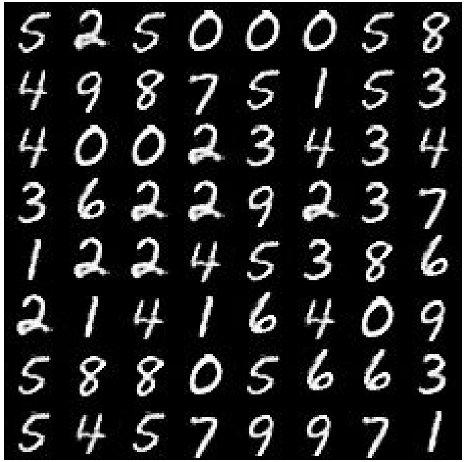
???

YES!

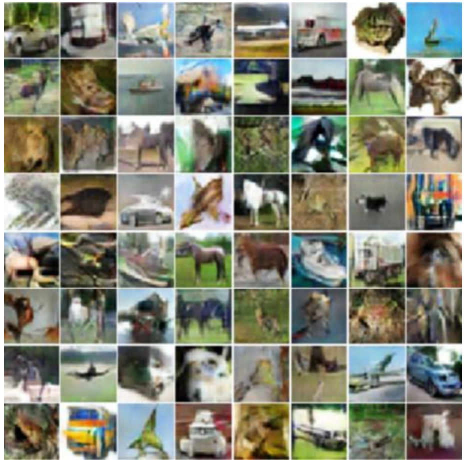
by using Generative Adversarial Networks (GAN)

How does a GAN work?





MNIST images



CIFAR-10 images

faces

album covers

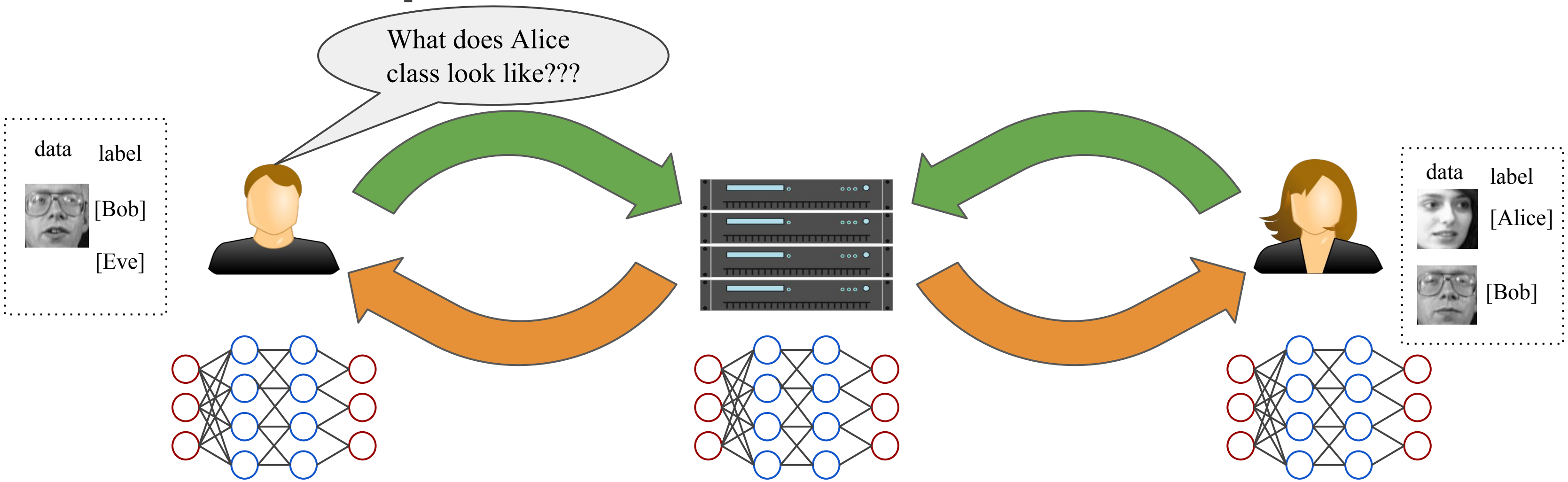
bedrooms

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

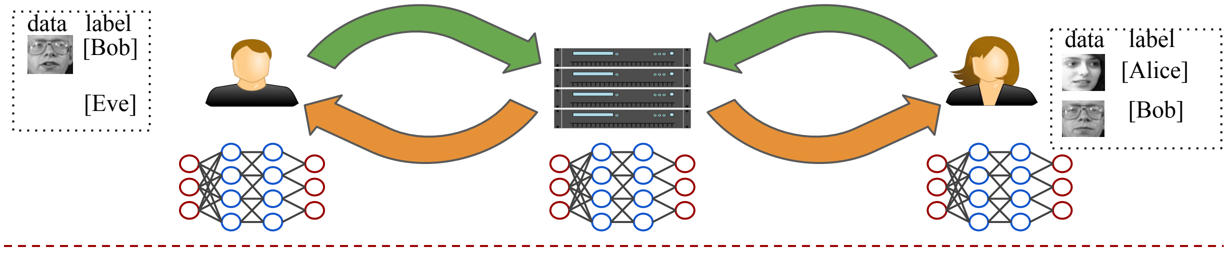
Our Attack: Deep Models Under the GAN

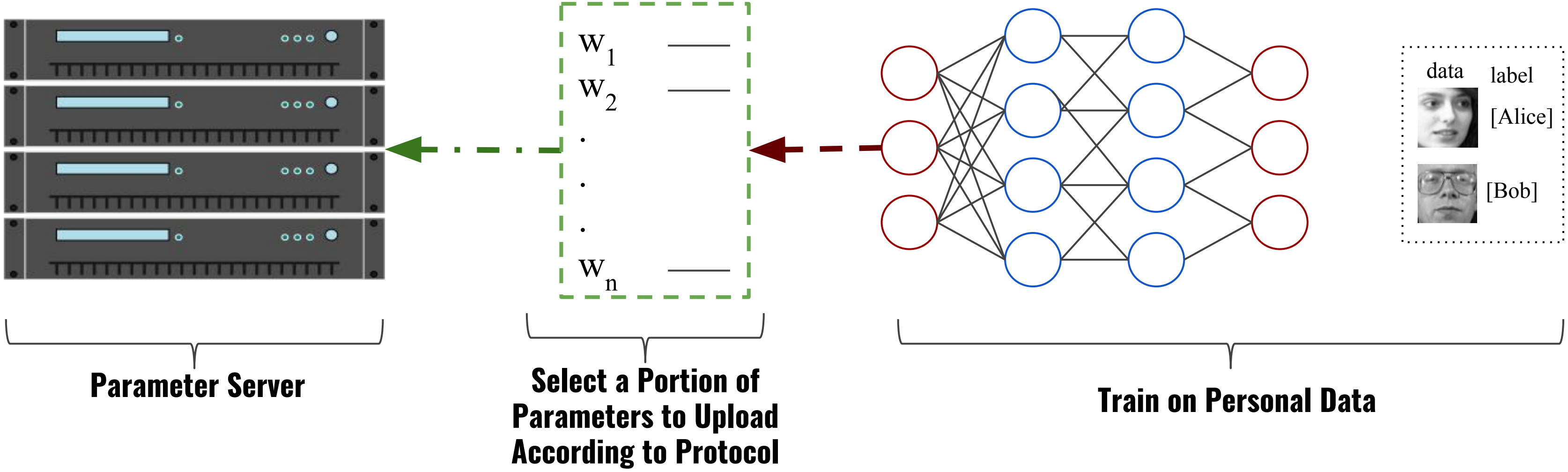


Agreement

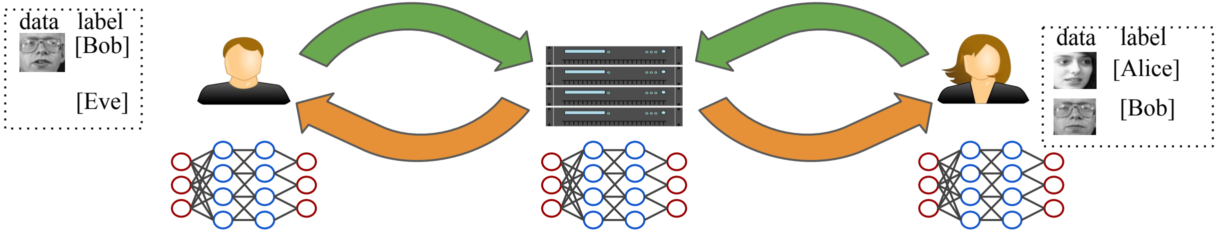
- Common learning objective
- Architecture of the model
- Labels/Classes present

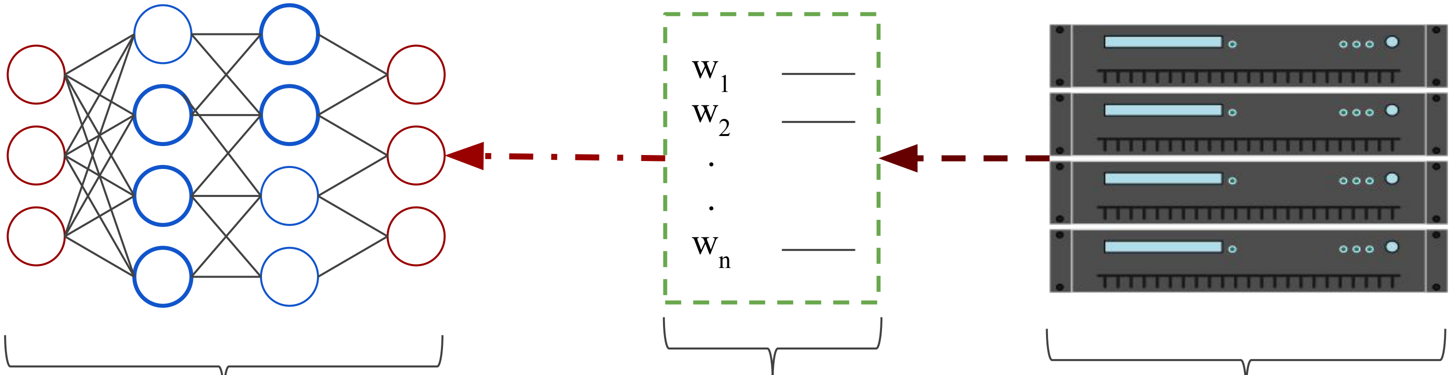
Victim’s Turn





Adversary's Turn



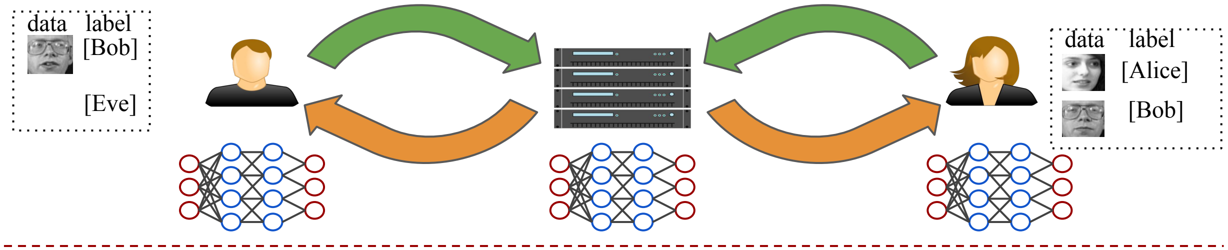


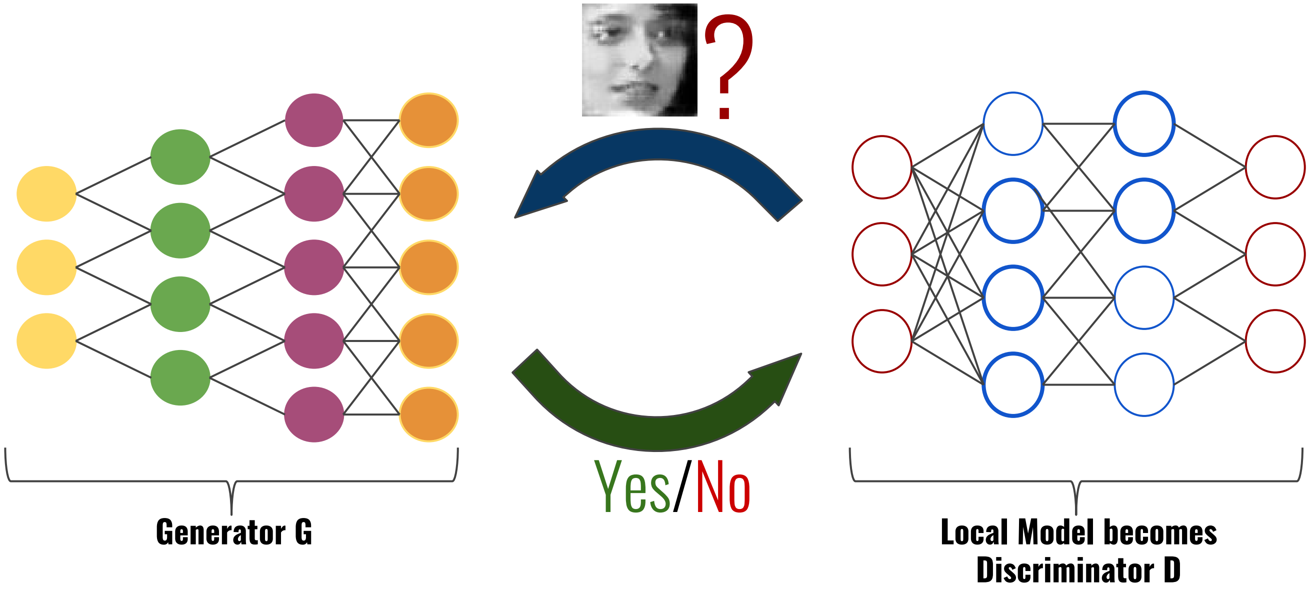
Update Local Model

Download a Portion of Parameters According to Protocol

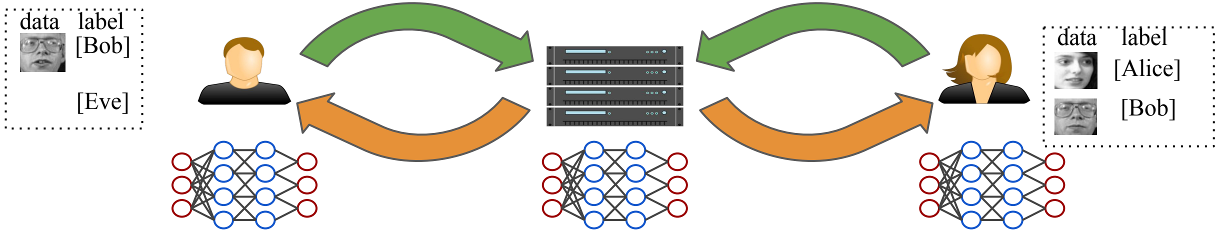
Parameter Server

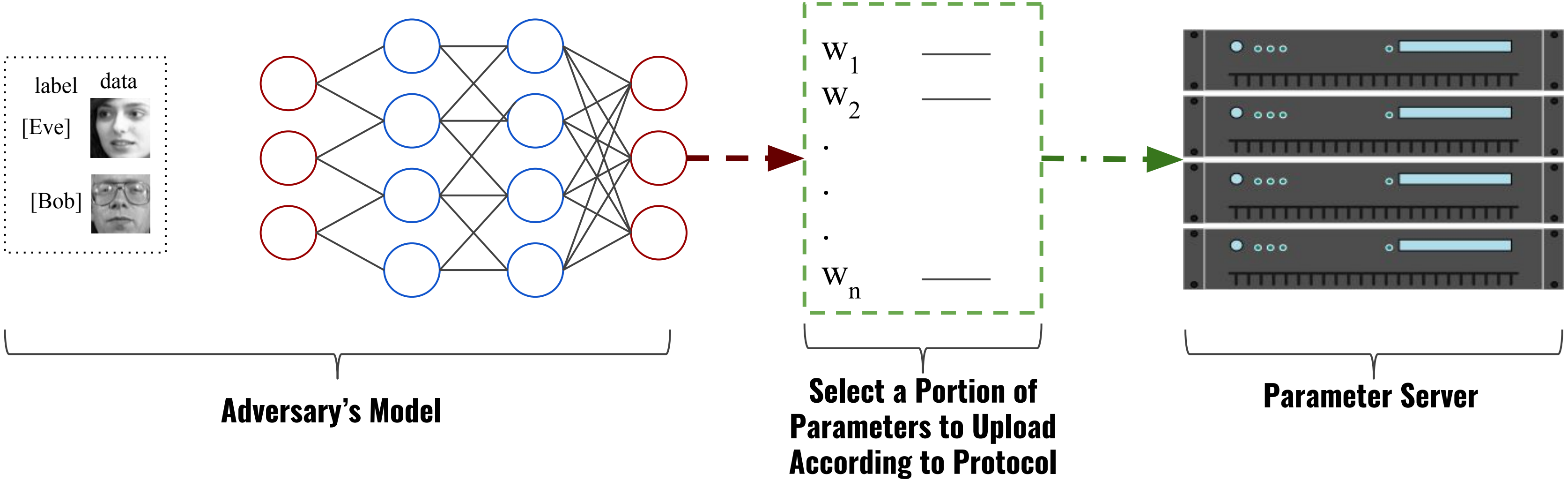
Adversary's Turn





Adversary’s Turn





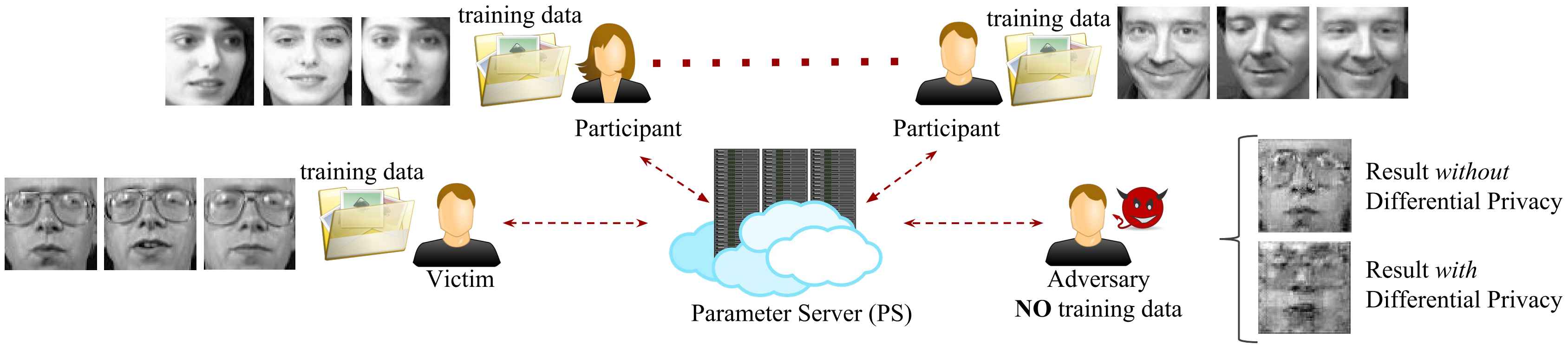
Experiments without Differential Privacy

Actual Images

Generated Data

Original vs Generated

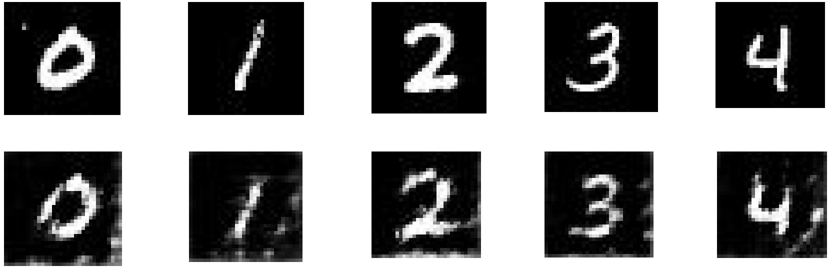
Experiments (Adversary has NO data at all)



Experiments with Differential Privacy

Actual Images

Generated Data



Original vs Generated

is anything wrong with differential privacy

???

NO!

Problem concerns the granularity at which differential privacy is *currently* applied

Further Results

Actual Images

Generated Images

Generated images when targeting ‘horse’ class from CIFAR-10

Collaborative learning for privacy is less desirable than centralized learning

What's next?

What's next?

- Extending on a broader range of datasets
- Further improving the GAN model (more art than science)
- Devising countermeasures

A version of the paper can be found at:
https://arxiv.org/abs/1702.07464

Thank You!