UNIVERSITY OF TORONTO

VECTOR INSTITUTE

What Does It Mean for Machine Learning to Be Trustworthy?

Nicolas Papernot

University of Toronto & Vector Institute

Why is the trustworthiness of ML important? *security*

Microsoft's Tay chatbot

Training data poisoning

YouTube filtering

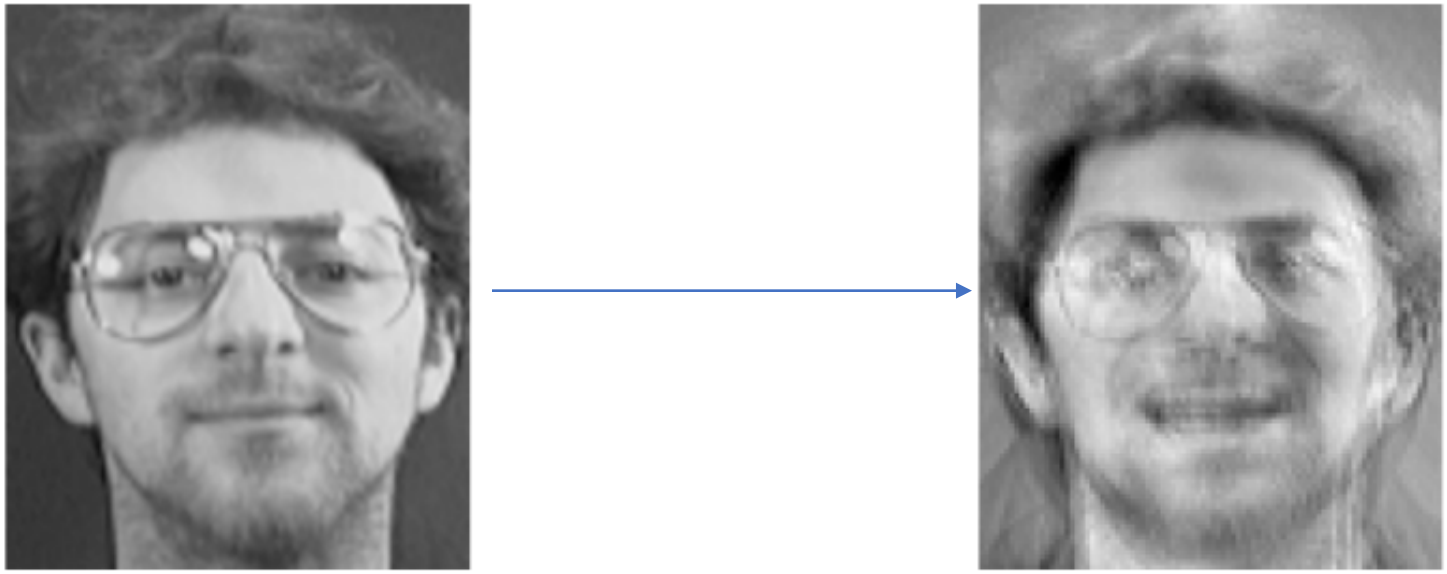
Content evades detection at *inference*

Why is the trustworthiness of ML important? *safety*

Testing the NVIDIA DAVE-2 self-driving car platform.

Pei et al. *DeepXplore: Automated Whitebox Testing of Deep Learning Systems*

Why is the trustworthiness of ML important? *privacy*

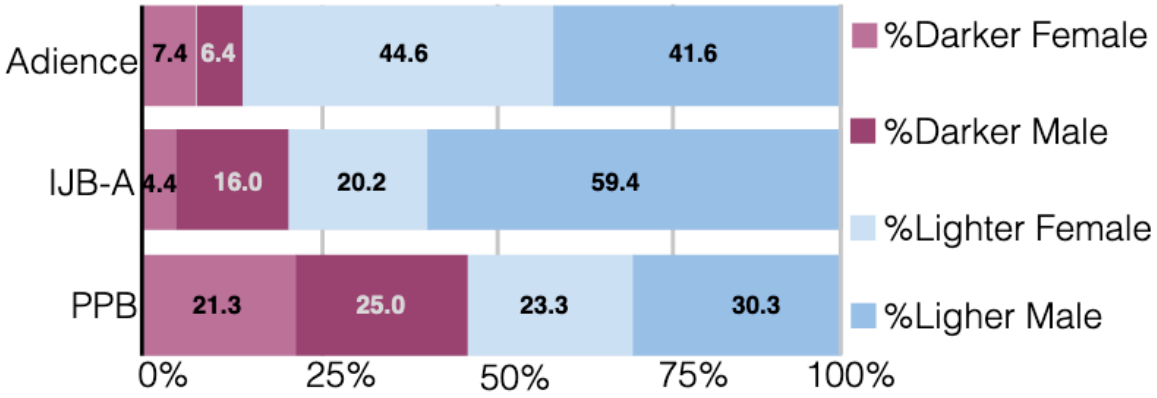


Model inversion attack

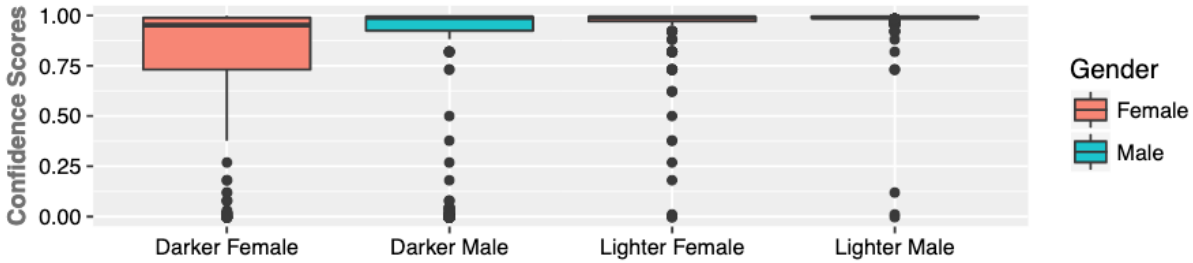
Adversary learns about the training data from model predictions

Fredrikson et al. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures.

Why is the trustworthiness of ML important? *fairness & ethics*



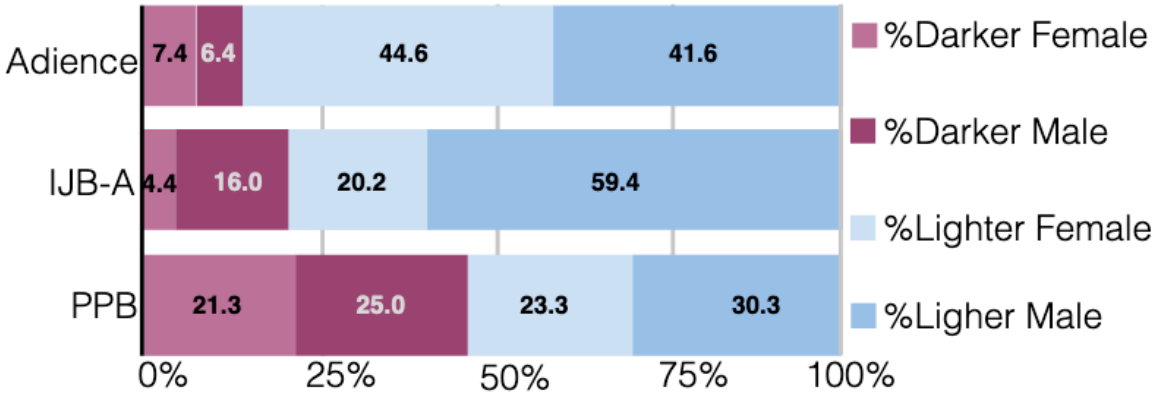
Percentage of subjects by (skin color, gender) pairs



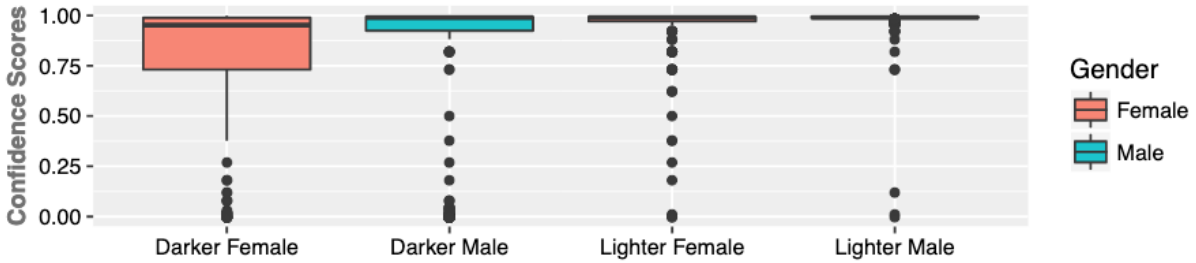
Classification confidence scores from IBM

Buolamwini and Gebru. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Why is the trustworthiness of ML important? *fairness & ethics*



Percentage of subjects by (skin color, gender) pairs



Classification confidence scores from IBM

Deepfakes

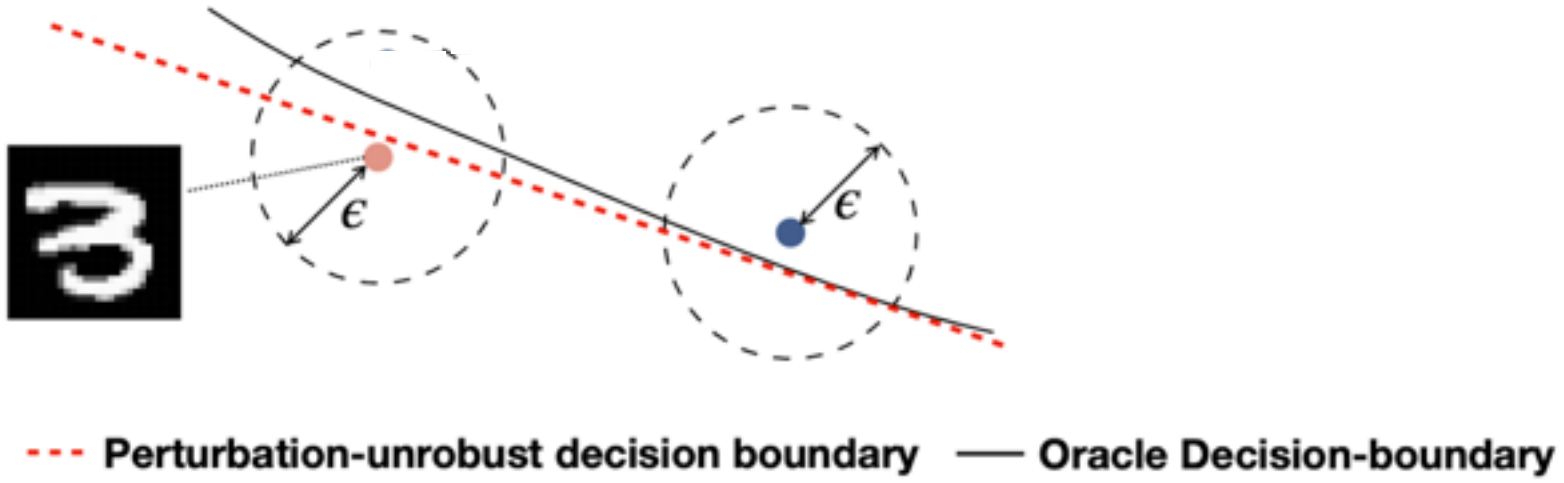
Image from Suwajanakorn et al. *Synthesizing Obama: Learning Lip Sync from Audio*

Buolamwini and Gebru. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Boneh et al. *How Relevant is the Turing Test in the Age of Sophisbots?*

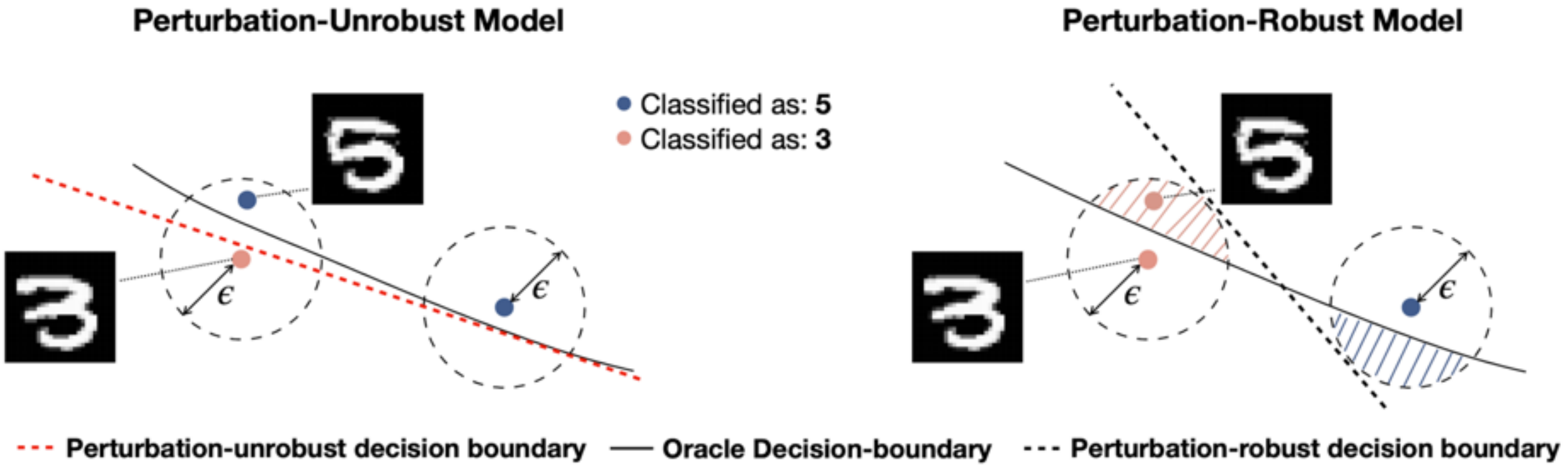
How do we design training algorithms that support trust and escape the arms race?

How to define our security policy? *A failed attempt*



Jacobsen et al. *Exploiting Excessive Invariance caused by Norm-Bounded Adversarial Robustness*

How to define our security policy? *A failed attempt*



Jacobsen et al. *Exploiting Excessive Invariance caused by Norm-Bounded Adversarial Robustness*

Training robust models creates an arms race because we don't have a good security policy

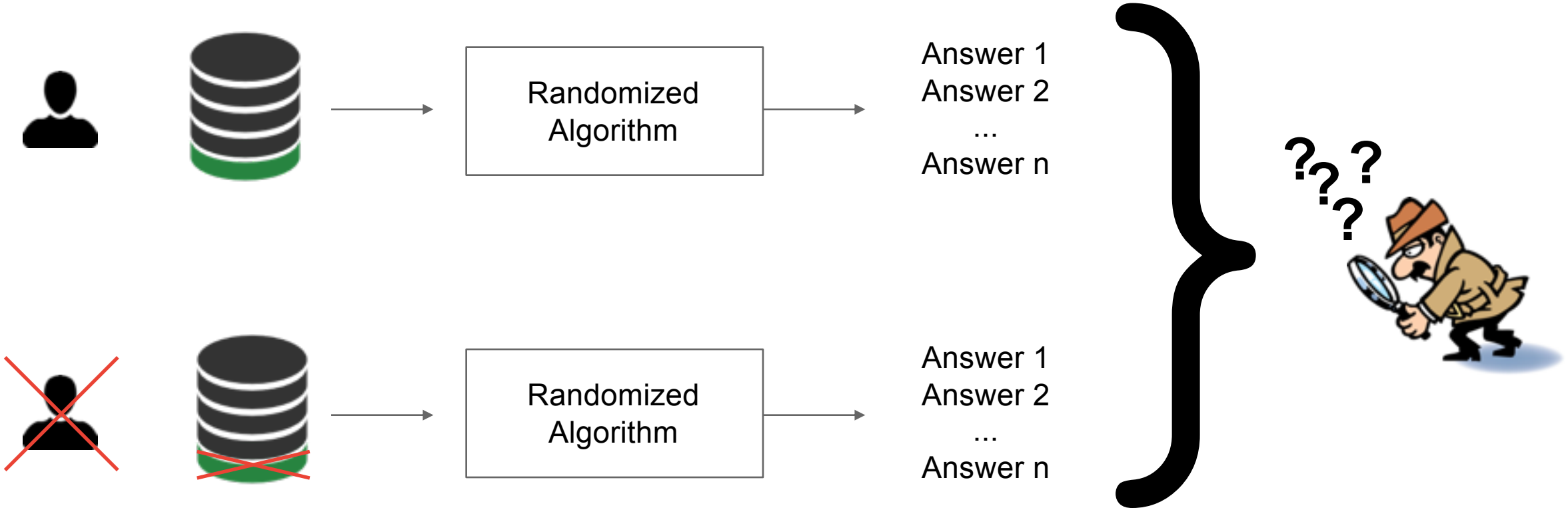
Is attacking machine learning easier than defending it? [Blog post at www.cleverhans.io]
Ian Goodfellow and Nicolas Papernot

Is achieving trustworthy ML any different from ~~real-world~~ computer security?

*"Practical security balances the cost of protection and the risk of loss, which is the cost of recovering from a loss times its probability" (*Butler Lampson, 2004)

Is the ML paradigm fundamentally different in a way that enables systematic approaches to security and privacy?

How to define a ~~security~~ privacy policy? A successful attempt



Differential Privacy: $Pr[M(d) \in S] \leq e^{\varepsilon} Pr[M(d') \in S]$

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

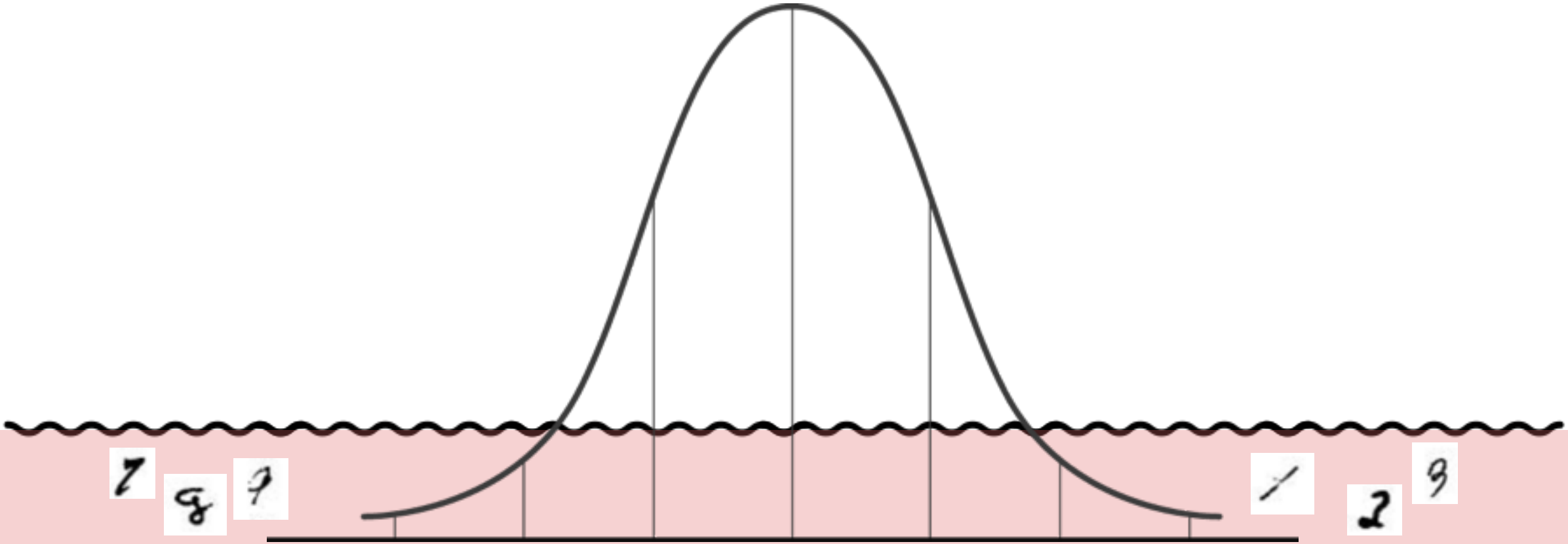
How to train a model?

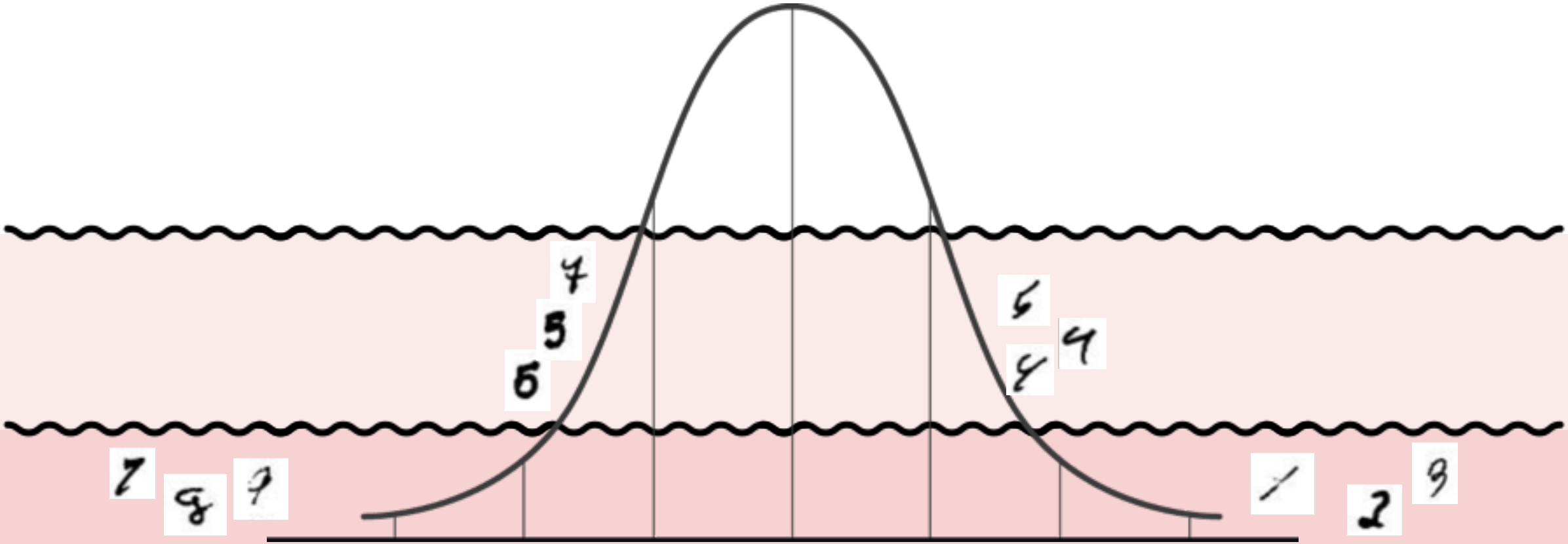
```
Initialize parameters θ

For t = 1..T do
  Sample batch B of training examples
  Compute average loss L on batch B
  Compute average gradient of loss L wrt parameters θ



  Update parameters θ by a multiple of gradient average
```

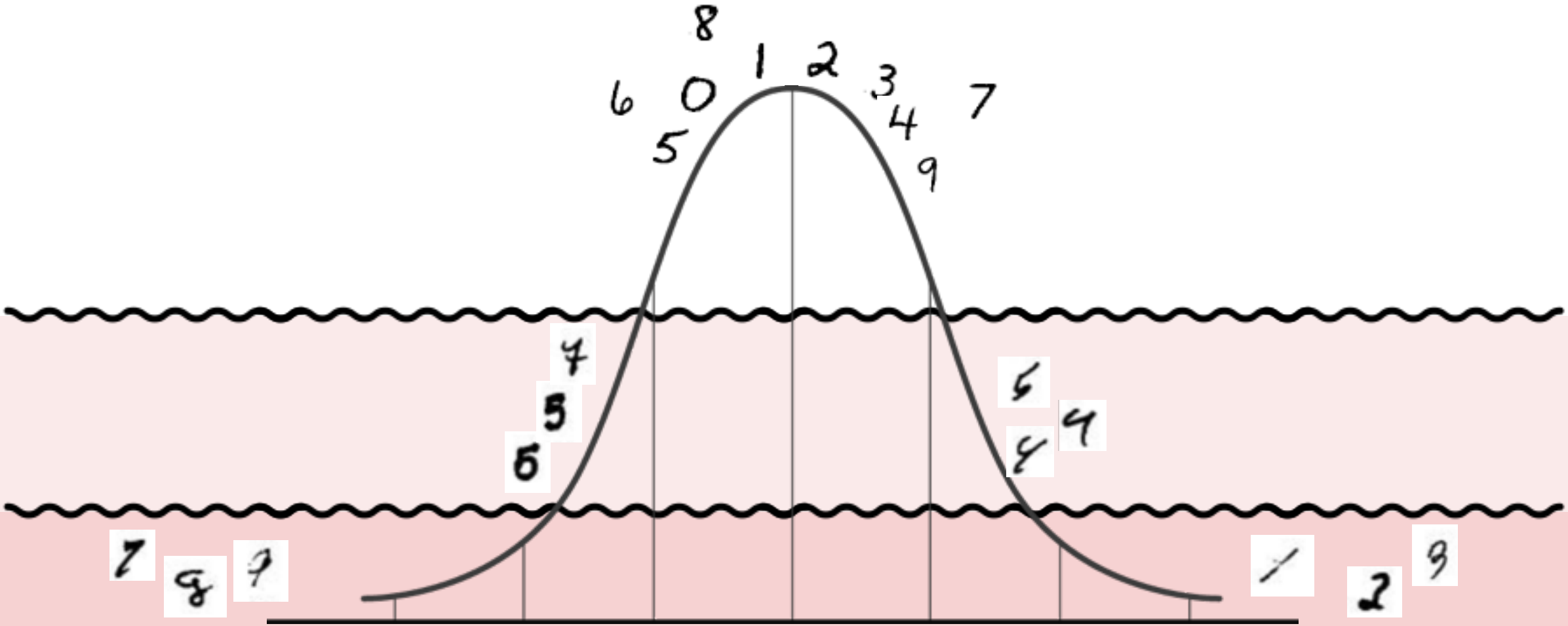
How to train a model with differential privacy?

```
Initialize parameters θ

For t = 1..T do
  Sample batch B of training examples
  Compute per-example loss L on batch B
  Compute per-example gradients of loss L wrt parameters θ
  Ensure L2 norm of gradients < C by clipping
  Add Gaussian noise to average gradients (as a function of C)
  Update parameters θ by a multiple of noisy gradient average
```

Deep Learning with Differential Privacy (CCS, 2016)
Abadi, Chu, Goodfellow, McMahan, Mironov, Talwar, Zhang



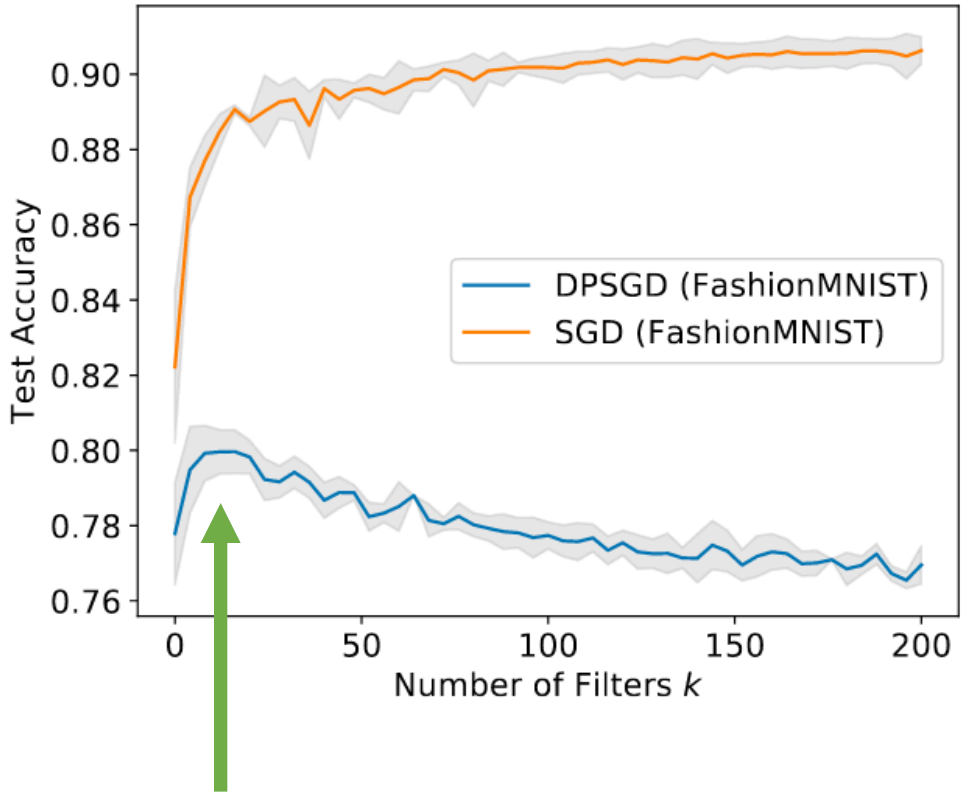




Slides adapted from Ulfar Erlingsson

[arxiv:1910.13427]

Architectures, initializations, hyperparameters for DP-SGD learning



More capacity is not always helpful

Papernot et al. *Making the Shoe Fit: Architectures, Initializations, and Tuning for Learning with Privacy* (in submission)

Why is differential privacy in ML successful?

- Definition of robustness to adversarial examples using simplistic distances like L_p norms directly conflicts with generalization
- Instead differential privacy encourages generalization

1. No necessary trade-off between privacy and ML objective

2. Degrades smoothly to not learning when it cannot be done privately

What does it mean for ML to be trustworthy?

What about test time?

Admission control may address lack of assurance.

How can sandboxing, input-output validation and compromise recording help secure ML systems when data provenance and assurance is hard?

Auditing may help with model governance.

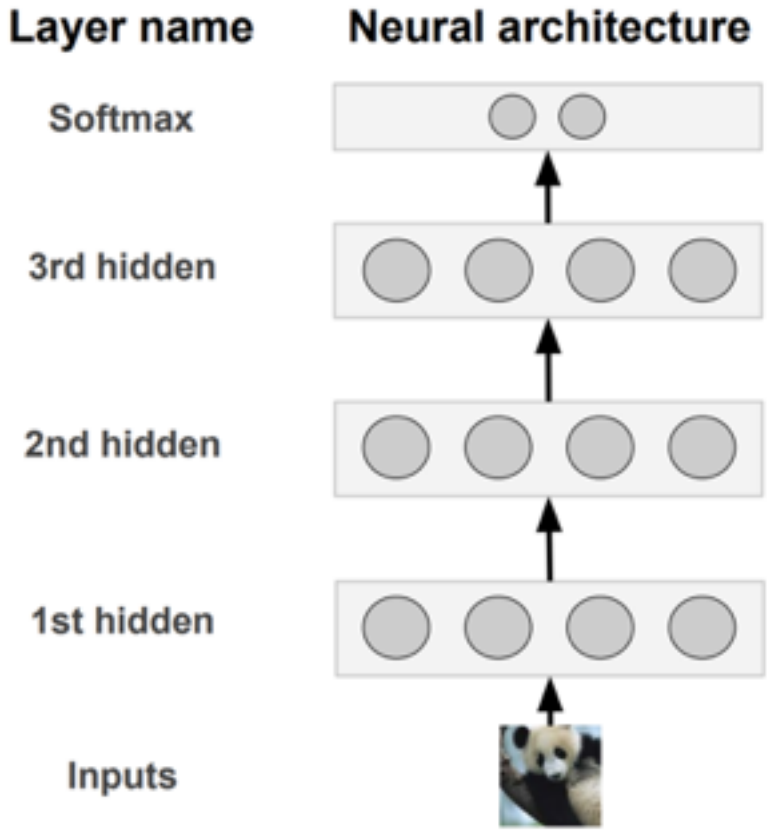
How can compromise recording help secure ML systems throughout their lifetime?

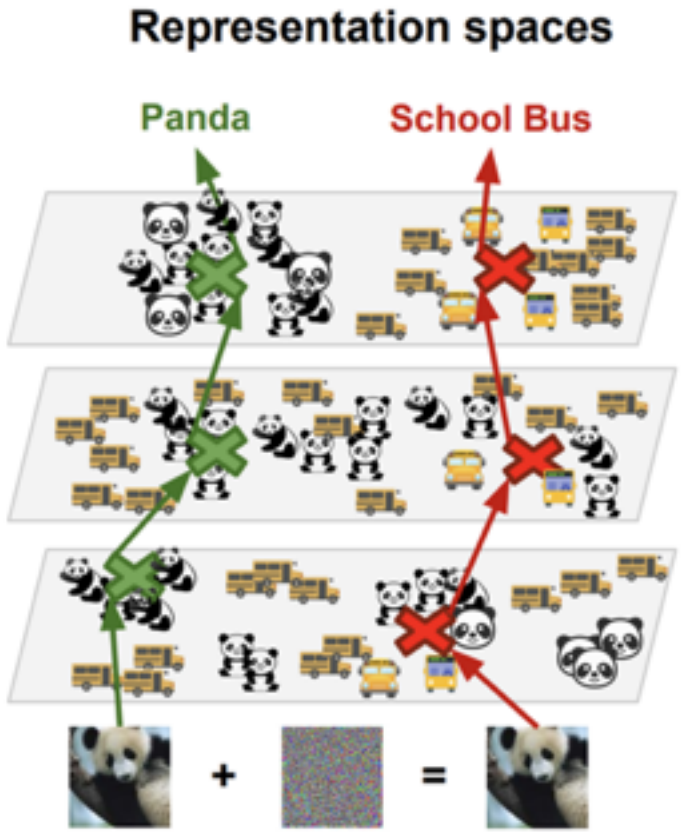
Admission control at test time

Weak authentication (similar to search engines) calls for admission control:

Do we admit a sandboxed model's output into our pool of answers?

Example:
define a well-calibrated estimate of uncertainty to reject outliers (hard when distribution is unknown) through conformal prediction

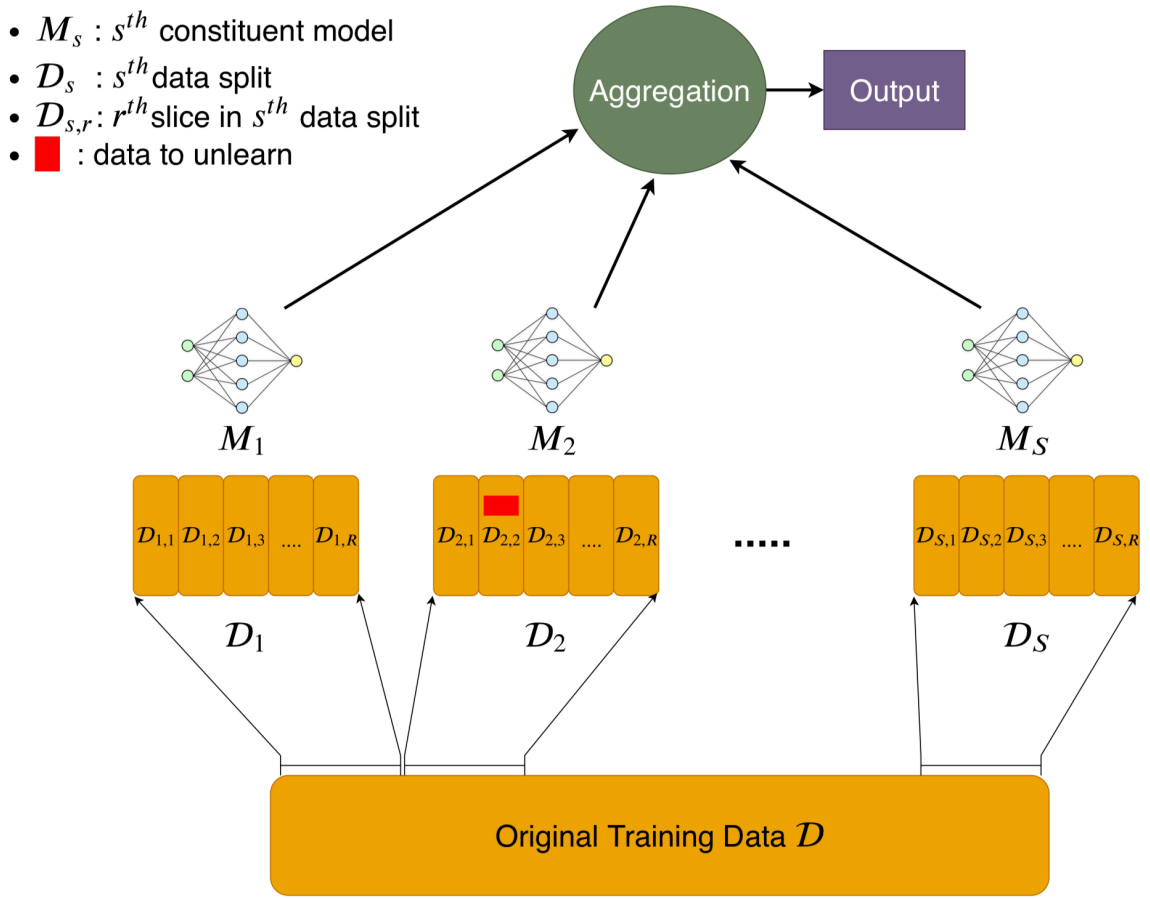




Deep k-Nearest Neighbors (2018)
Papernot and McDaniel

Soft Nearest Neighbor Loss (2019)
Frosst, Papernot and Hinton

Machine Unlearning… towards model governance.



Machine Unlearning. Lucas Bourtoule, Varun Chandrasekaran, Christopher Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, Nicolas Papernot (in submission)

Towards trustworthy ML

- Policies are needed to align ML with societal norms:
  - Security: integrity, confidentiality…
  - Privacy: differential privacy, confidentiality, …
  - Ethics: fairness criteria, …
- Technology needs to:
  - At train time: propose algorithms that satisfy these policies
  - At test time: can perform admission control and model governance
- Beyond technology, complement with legal frameworks and education
- Trustworthy ML is an opportunity to make ML better

Resources:
cleverhans.io
github.com/tensorflow/cleverhans
github.com/tensorflow/privacy

UNIVERSITY OF TORONTO

VECTOR INSTITUTE

?

Contact information:
nicolas.papernot@utoronto.ca
@NicolasPapernot

I’m hiring at UofT & Vector:
- Graduate students
- Postdocs
- Faculty positions at all ranks