UNIVERSITY OF TORONTO

VECTOR INSTITUTE

The Limitations of DL in Adversarial Settings

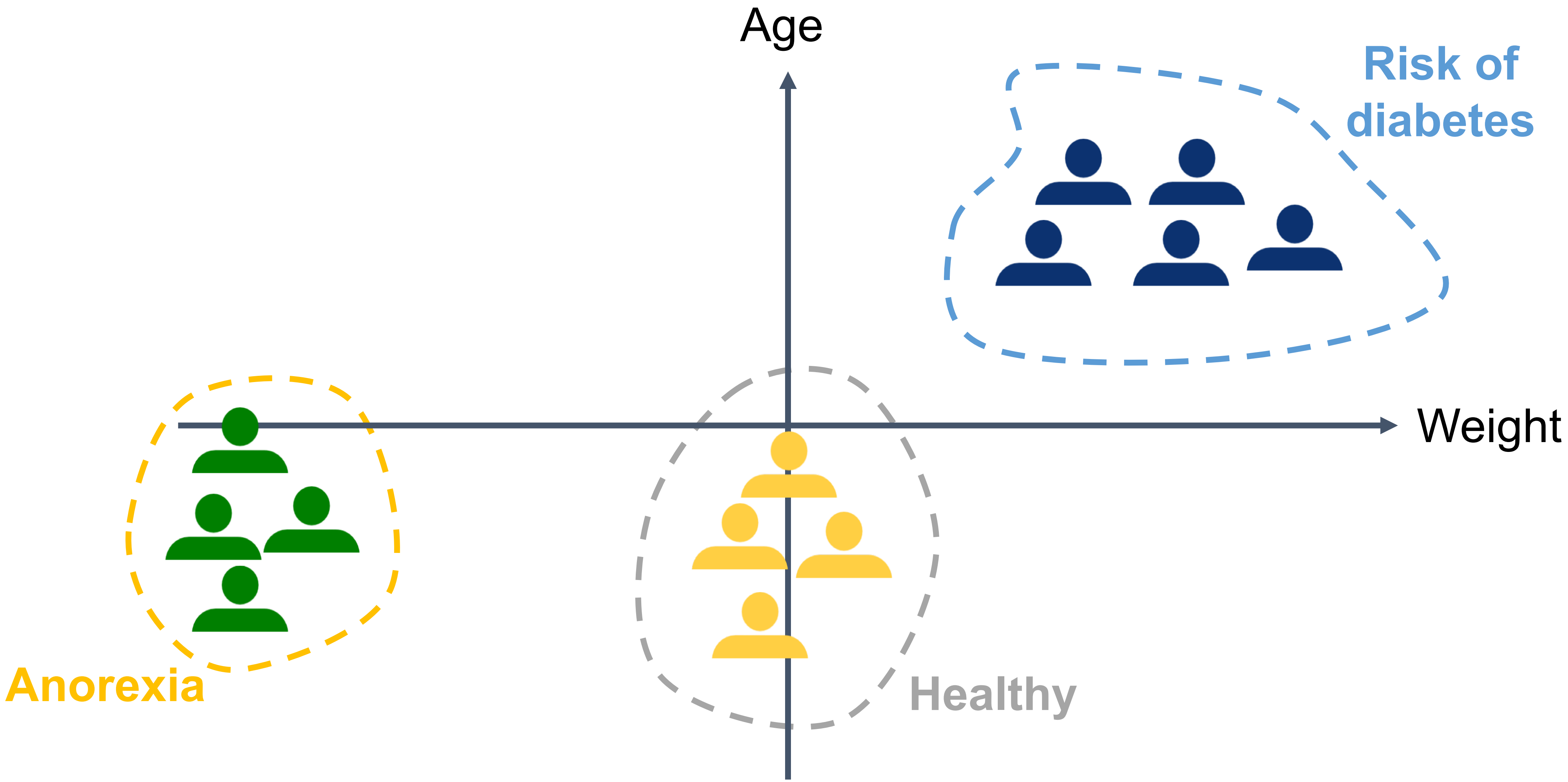
Nicolas Papernot

University of Toronto & Vector Institute

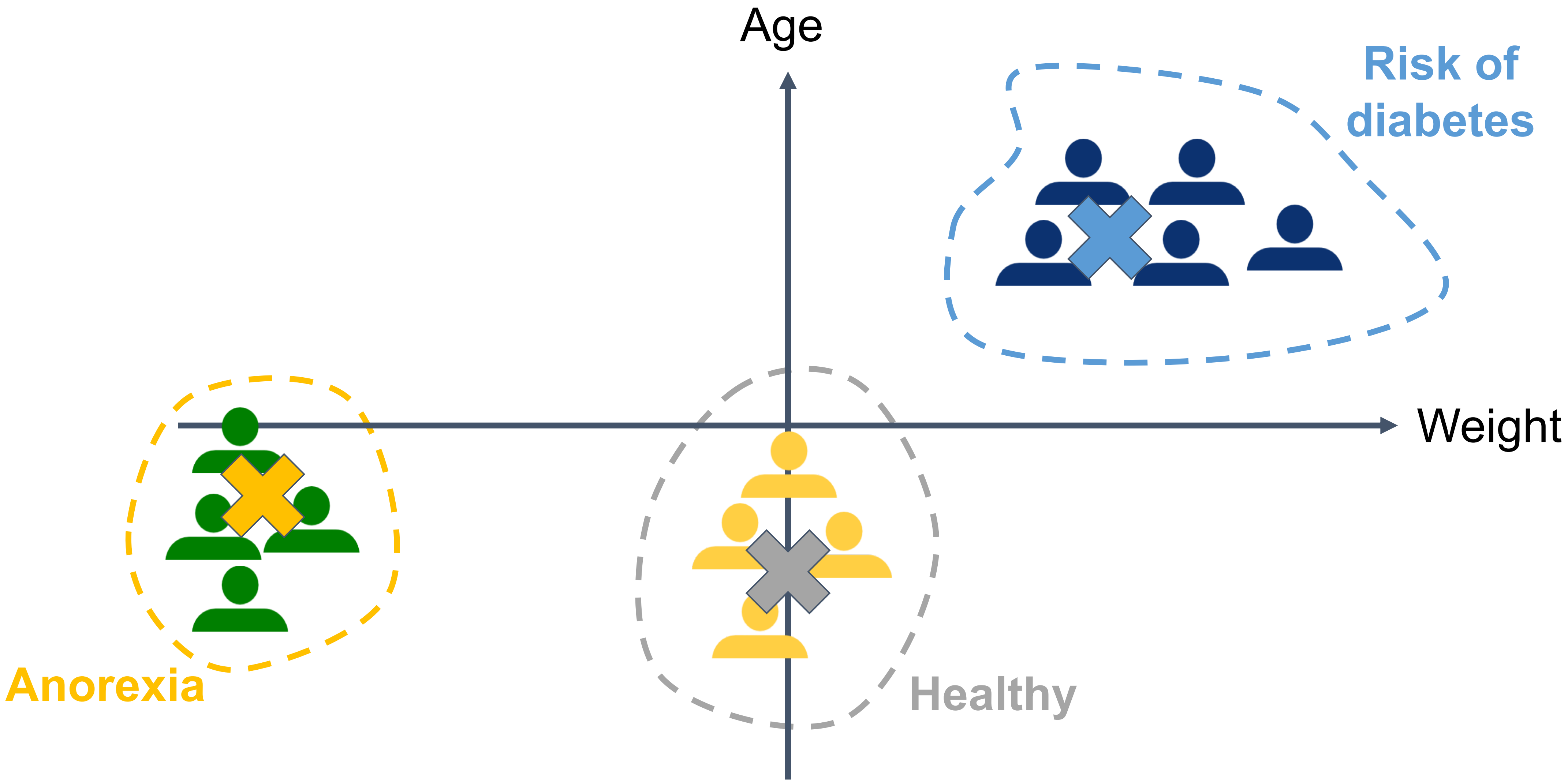
March 2020 - CMU

@NicolasPapernot

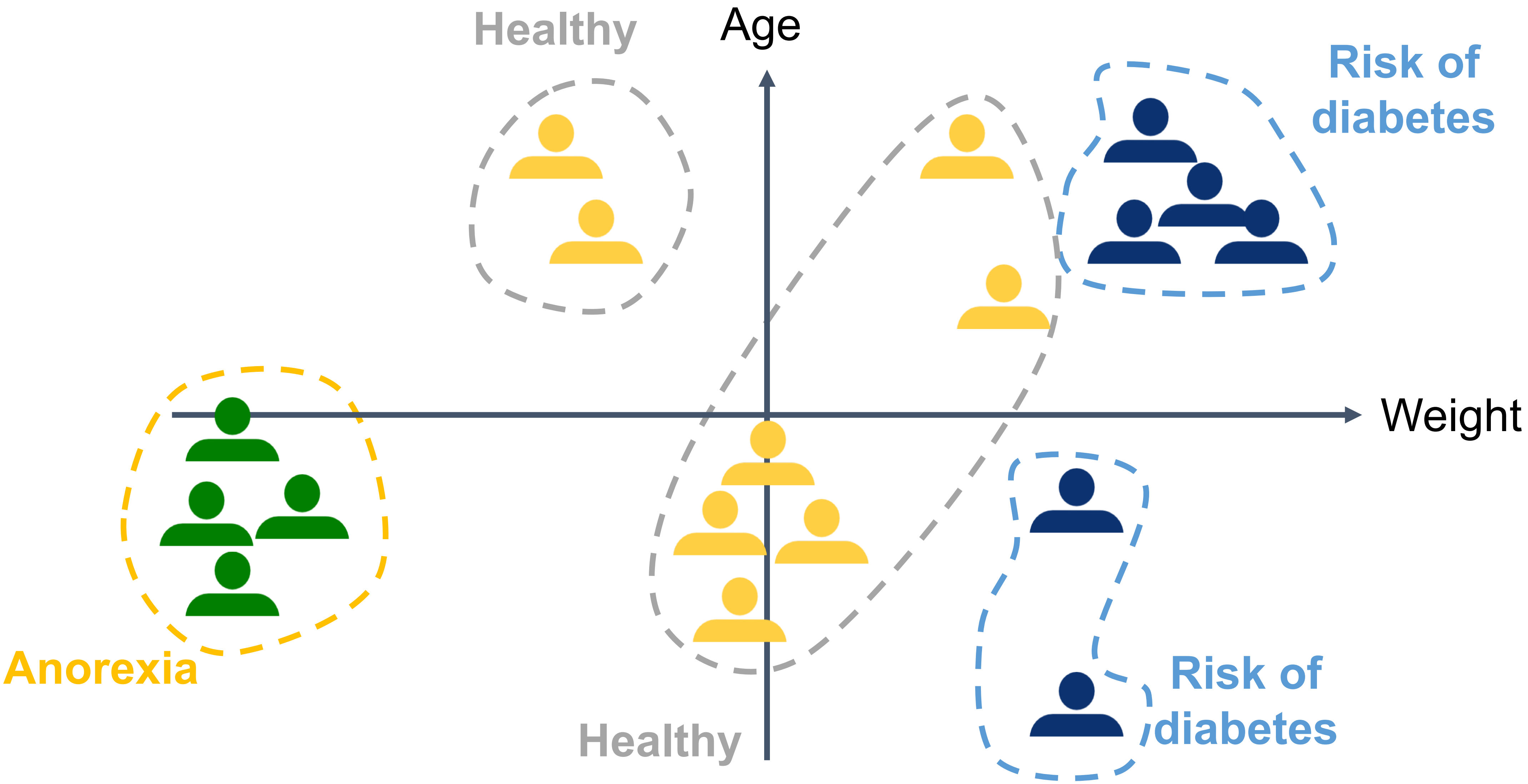
Machine learning is not magic: *ideal setting*



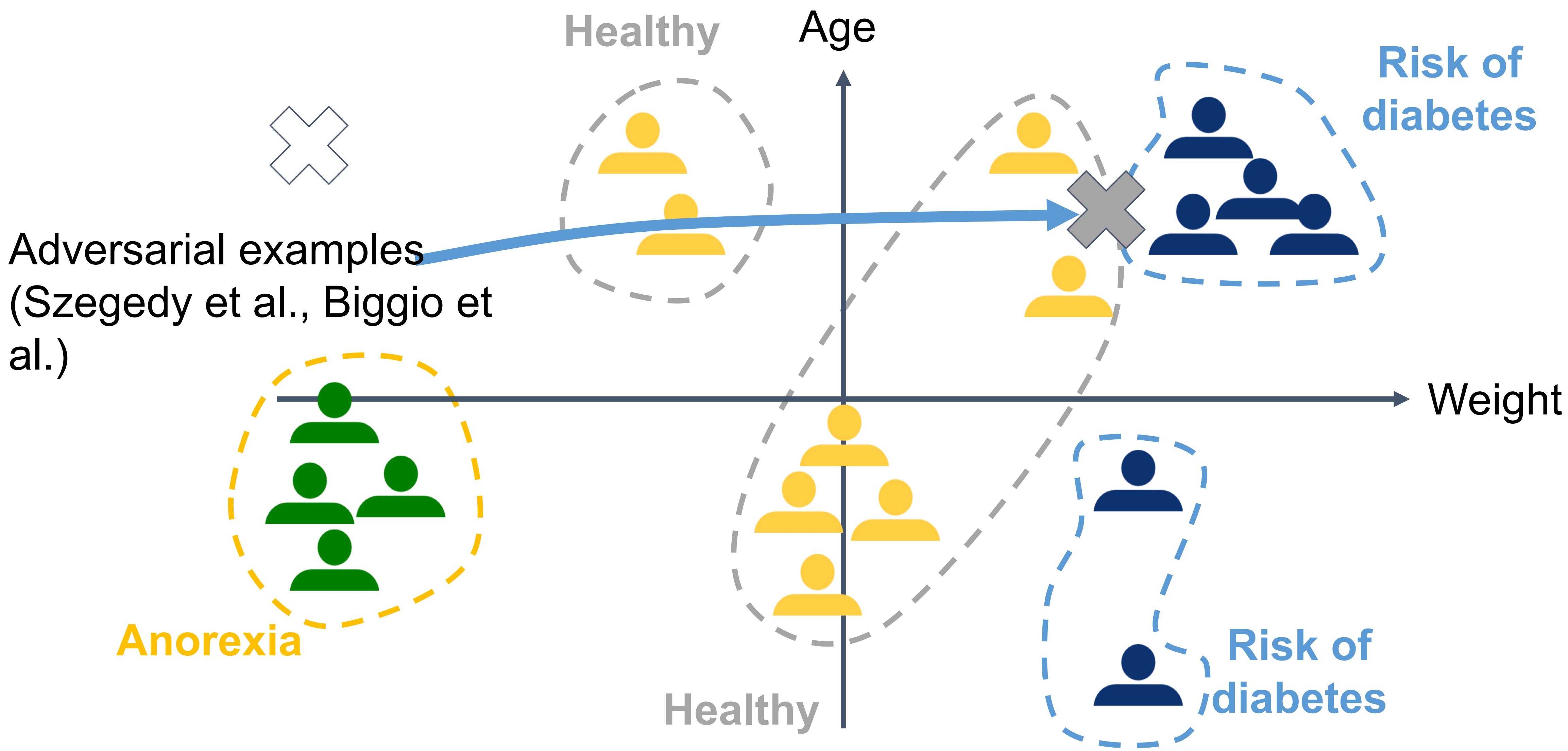
Machine learning is not magic: *ideal setting*



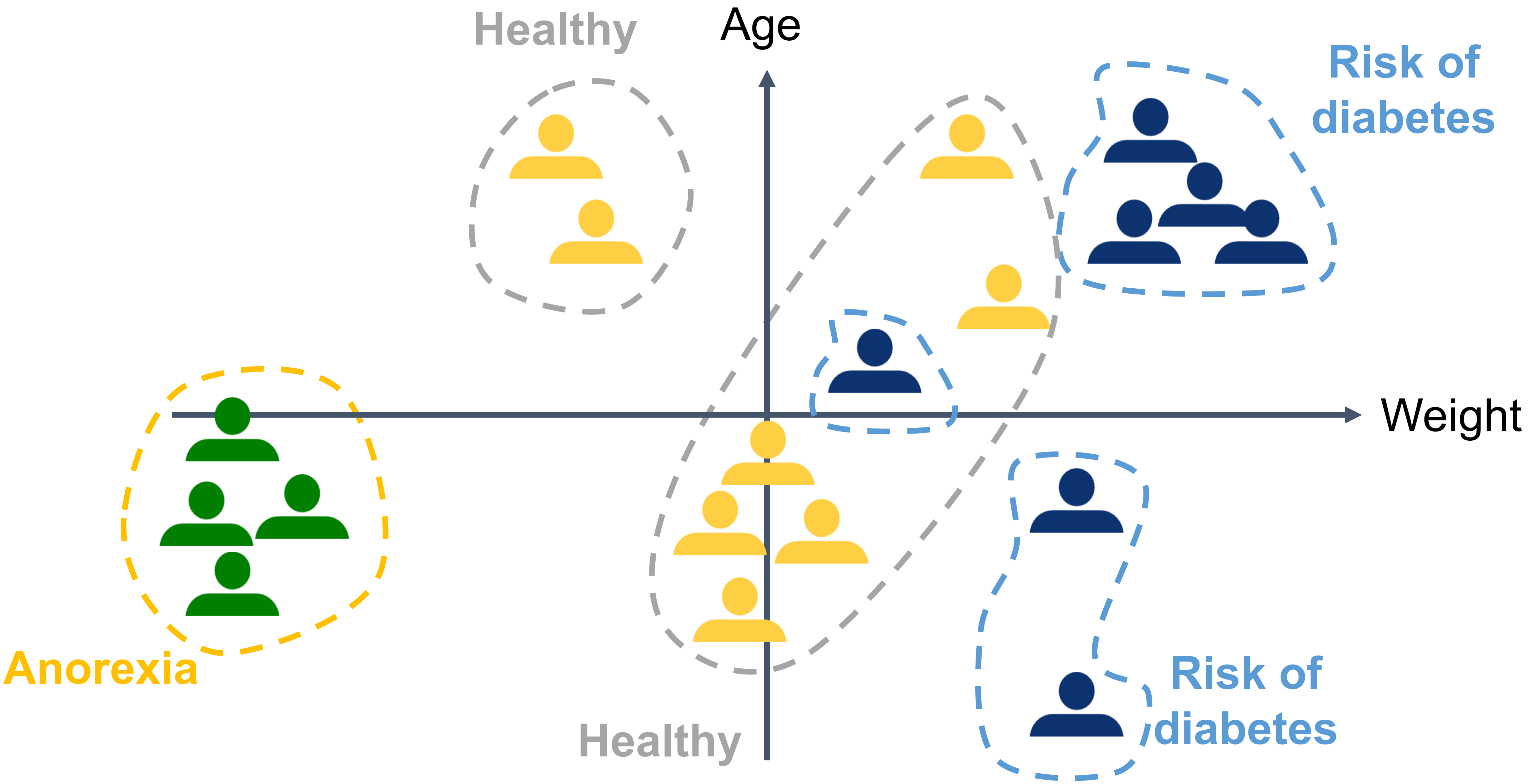
Machine learning is not magic: *(adversarial) real-world*



Machine learning is not magic: *(adversarial) real-world*

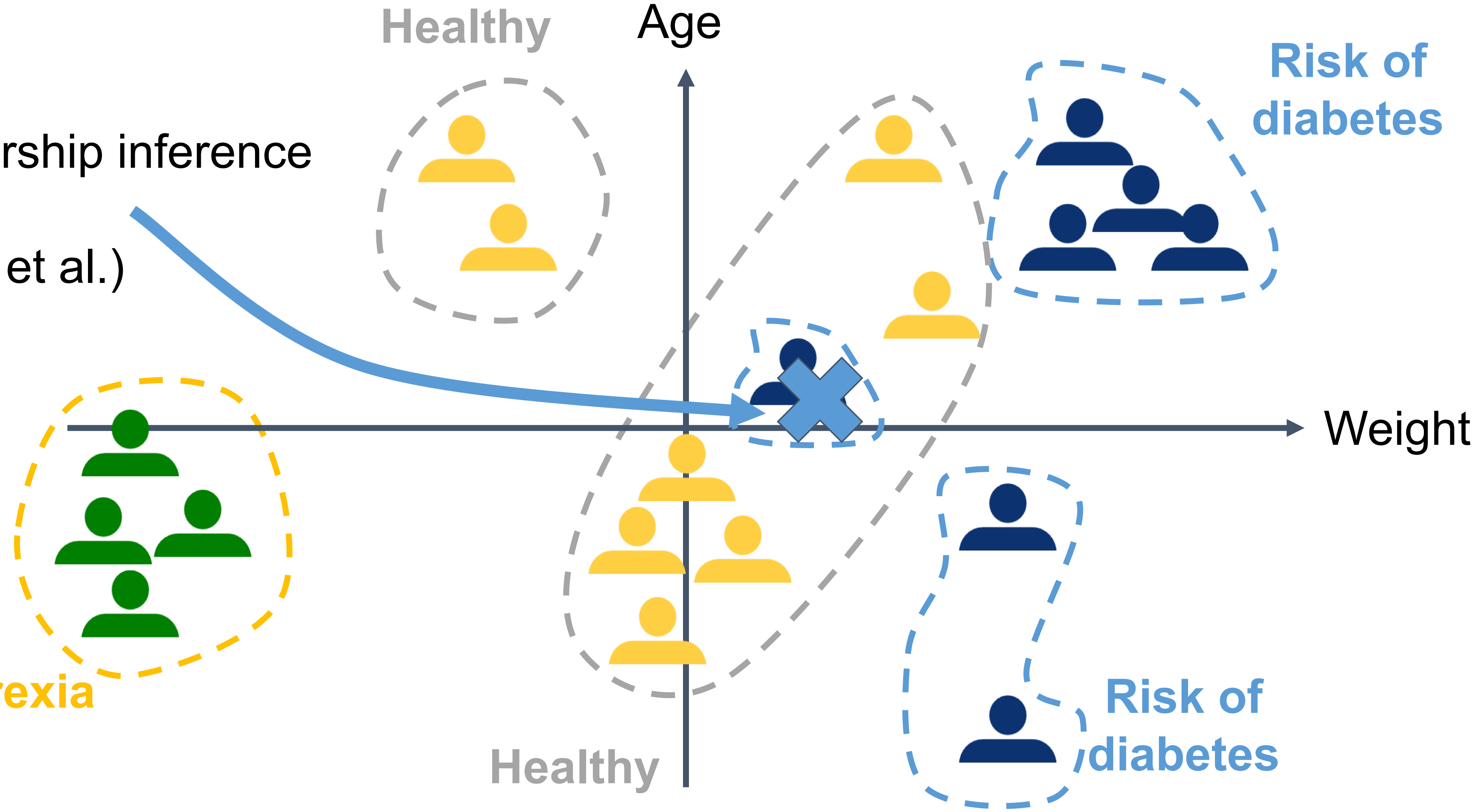


Machine learning is not magic: *(adversarial) real-world*



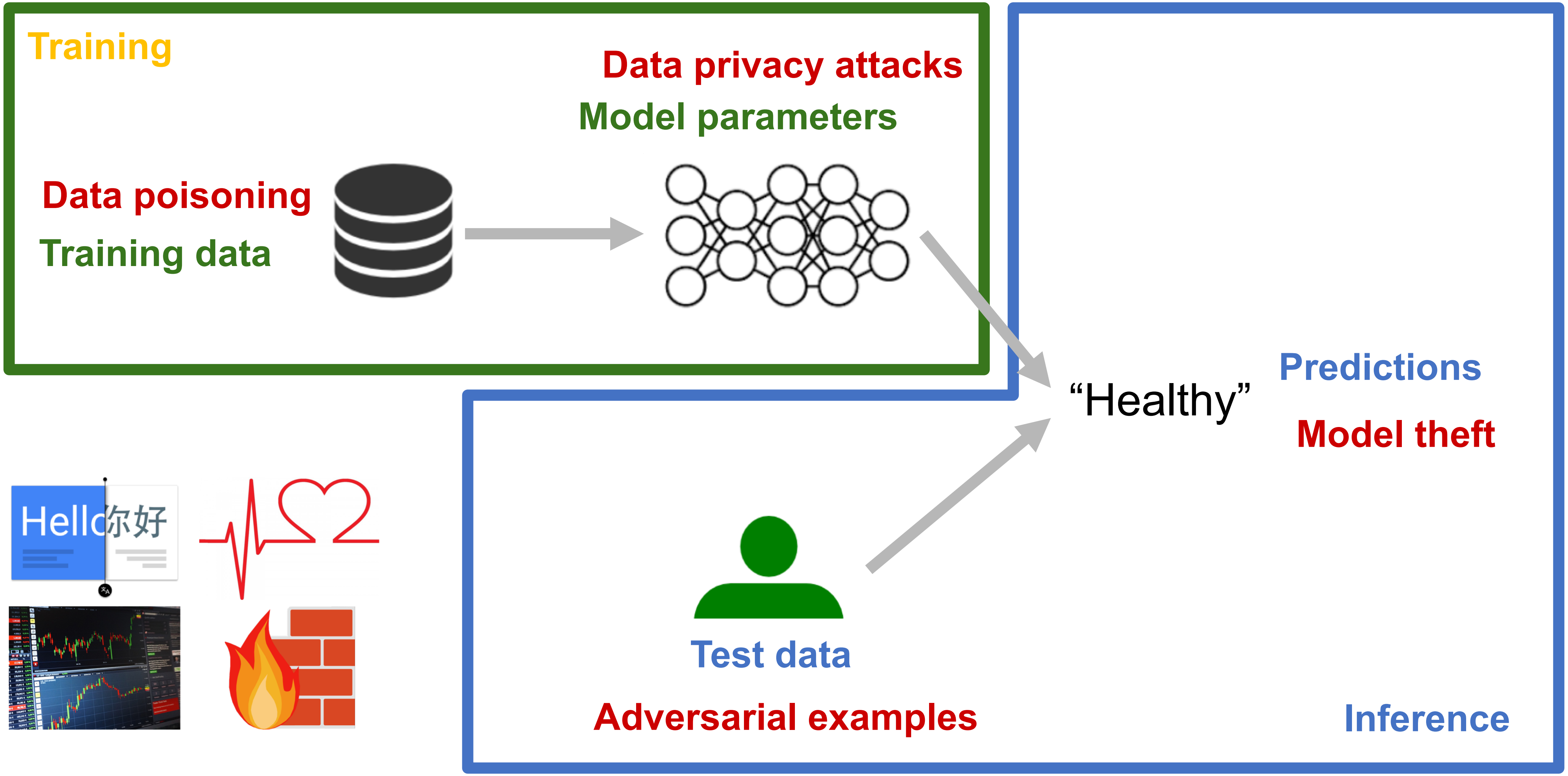
Machine learning is not magic: *(adversarial) real-world*

Membership inference attacks (Shokri et al.)



UNIVERSITY OF TORONTO VECTOR INSTITUTE

The ML paradigm in adversarial settings



Adapted from a slide by Ian Goodfellow

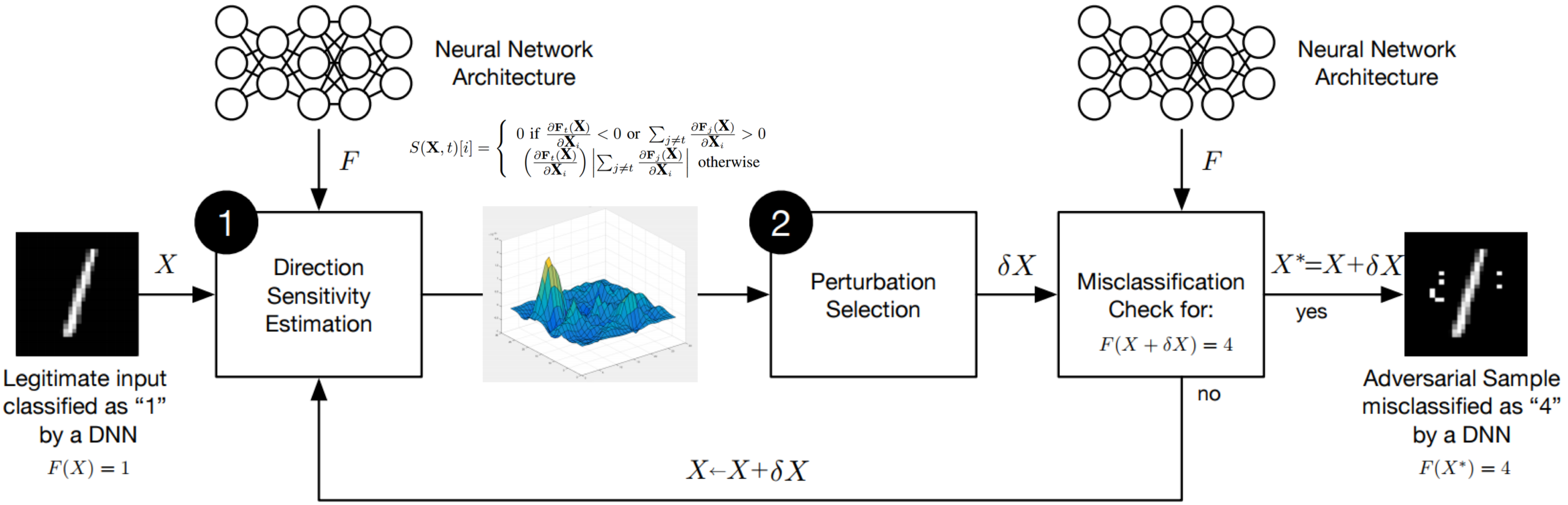
Security in Machine Learning

The threat model

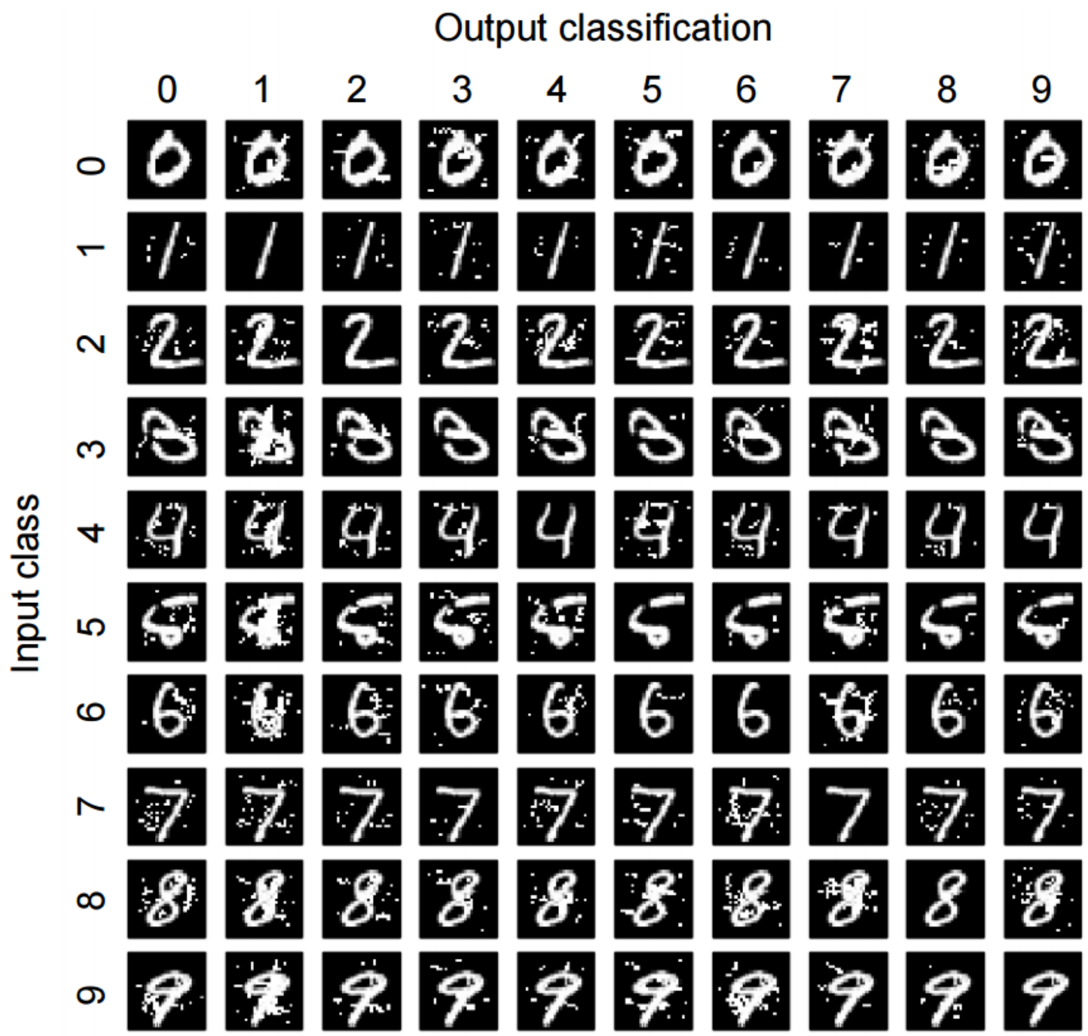
Attacker may see the model: attacker needs to know details of the machine learning model to do an attack --- aka a ***white-box attacker***

Attacker may not see the model: attacker who knows very little (e.g. only gets to ask a few questions) --- aka a ***black-box attacker***

Jacobian-based **S**aliency **M**ap **A**pproach (JSMA)

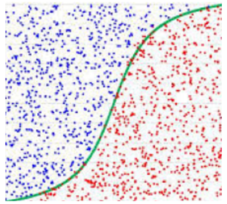


The Limitations of Deep Learning in Adversarial Settings [IEEE EuroS&P 2016]
Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami

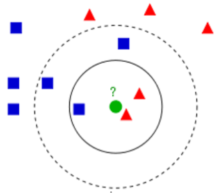

Output classification
0
1
2
3
4
5
6
7
8
9
Input class
0
1
2
3
4
5
6
7
8
9

Adversarial examples...

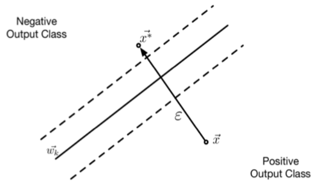
… beyond deep learning



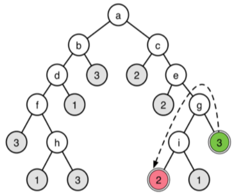
Logistic Regression



Nearest Neighbors

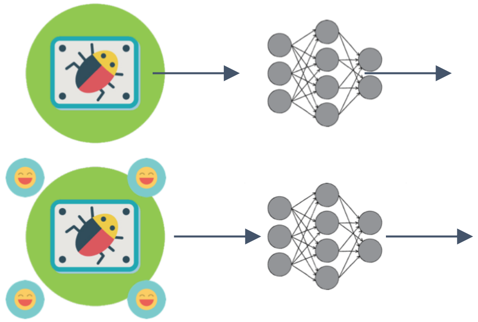


Support Vector Machines



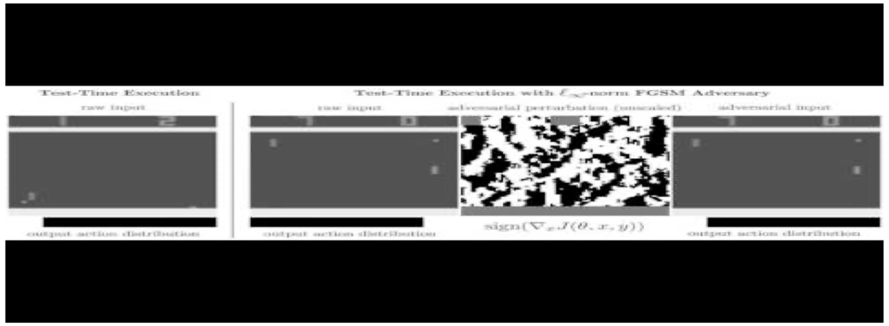
Decision Trees

… beyond computer vision



P[X=**Malware**] = **0.90**
P[X=Benign] = 0.10

P[X*=Malware] = 0.10
P[X*=**Benign**] = **0.90**



Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples [arXiv preprint]
Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow

Adversarial Attacks on Neural Network Policies [arXiv preprint]
Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, Pieter Abbeel

Adversarial Perturbations Against Deep Neural Networks for Malware Classification [ESORICS 2017]
Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, Patrick McDaniel

Optimization for adversarial examples

4.1 Formal description

We denote by $f : \mathbb{R}^m \longrightarrow \{1 \dots k\}$ a classifier mapping image pixel value vectors to a discrete label set. We also assume that f has an associated continuous loss function denoted by $\text{loss}_f : \mathbb{R}^m \times \{1 \dots k\} \longrightarrow \mathbb{R}^+$. For a given $x \in \mathbb{R}^m$ image and target label $l \in \{1 \dots k\}$, we aim to solve the following box-constrained optimization problem:

- Minimize $\|r\|_2$ subject to:
  1. $f(x + r) = l$
  2. $x + r \in [0, 1]^m$

The minimizer r might not be unique, but we denote one such $x + r$ for an arbitrarily chosen minimizer by $D(x, l)$. Informally, $x + r$ is the closest image to x classified as l by f. Obviously, $D(x, f(x)) = f(x)$, so this task is non-trivial only if $f(x) \neq l$. In general, the exact computation of $D(x, l)$ is a hard problem, so we approximate it by using a box-constrained L-BFGS. Concretely, we find an approximation of $D(x, l)$ by performing line-search to find the minimum $c > 0$ for which the minimizer r of the following problem satisfies $f(x + r) = l$.

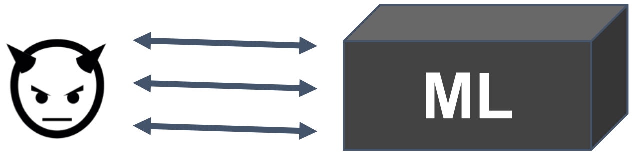
- Minimize $c|r| + \text{loss}_f(x + r, l)$ subject to $x + r \in [0, 1]^m$

Szegedy et al., Intriguing Properties of Neural Networks. (ICLR 2014)

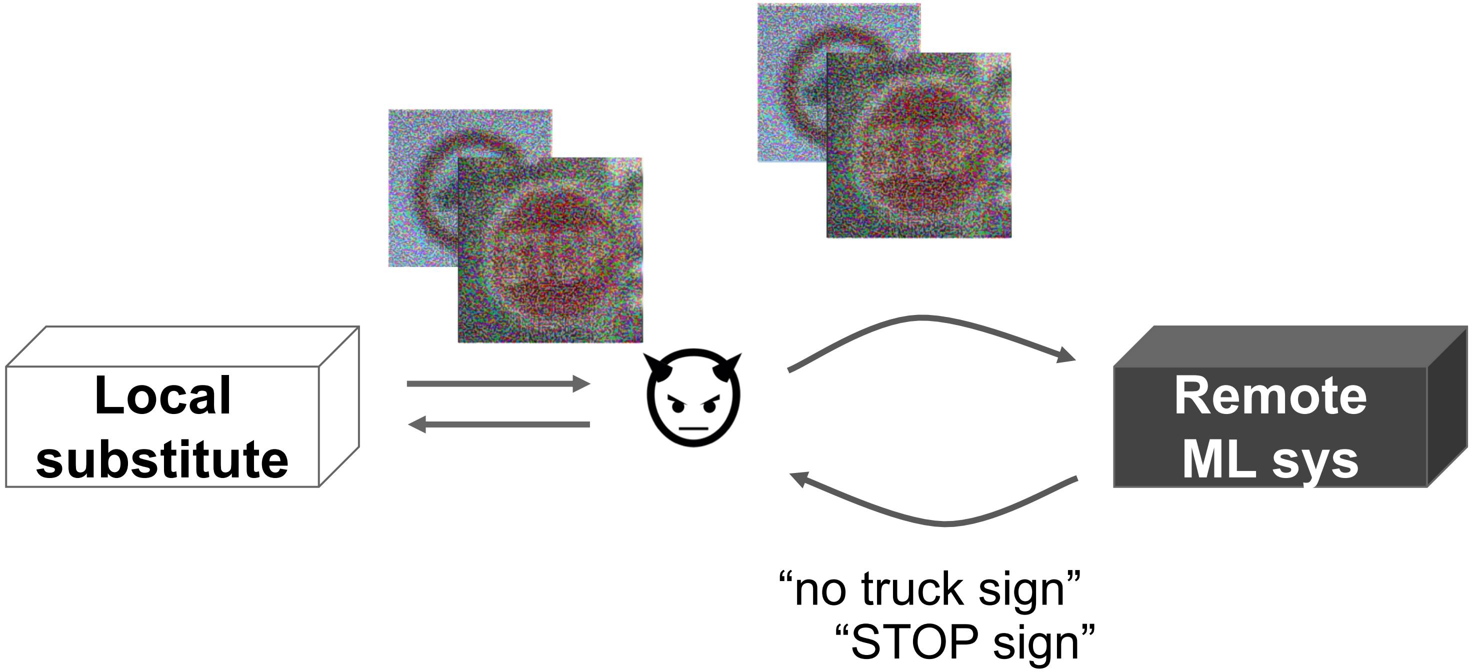
The threat model

Attacker may see the model: attacker needs to know details of the machine learning model to do an attack --- aka a ***white-box attacker***

Attacker may not see the model: attacker who knows very little (e.g. only gets to ask a few questions) --- aka a ***black-box attacker***

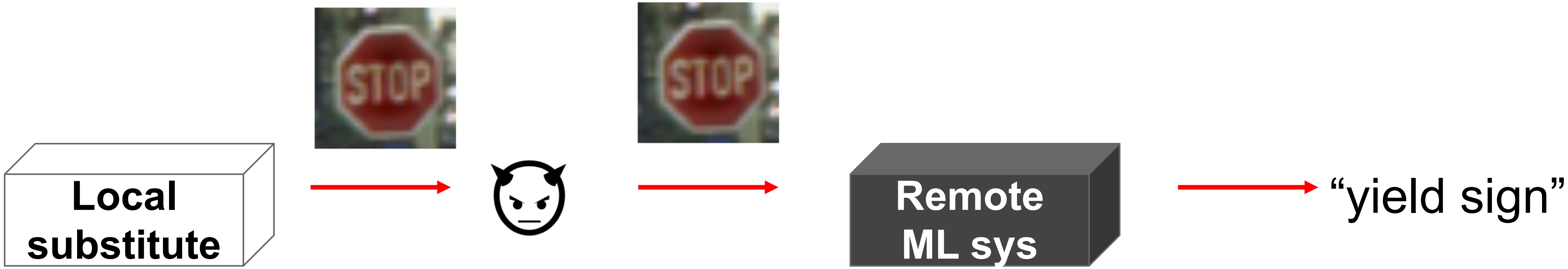


Attacking remotely hosted black-box models



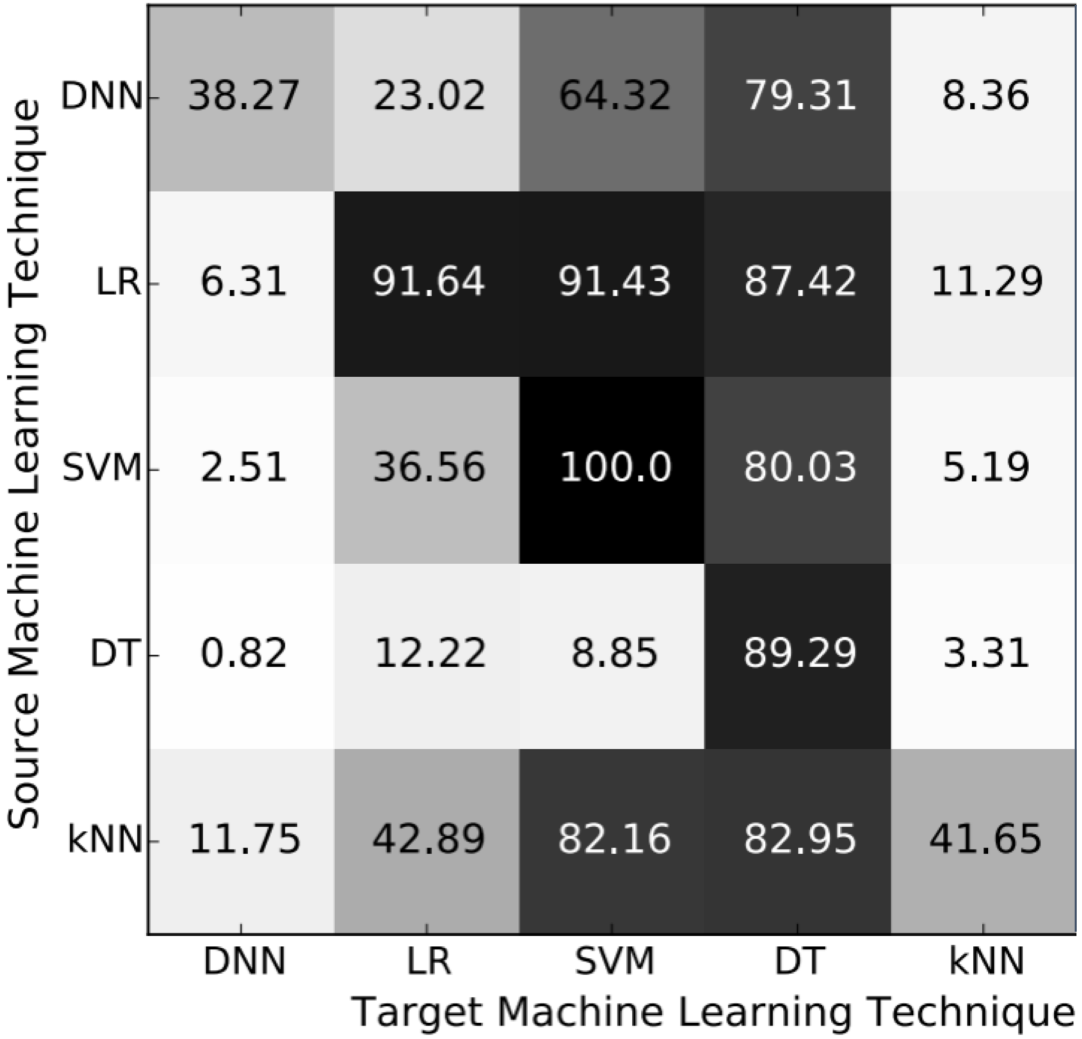
The adversary selects new synthetic inputs for queries to the remote ML system based on the local substitute’s output surface sensitivity to input variations.

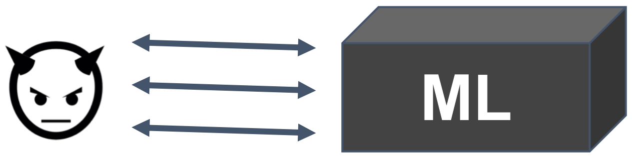
Attacking remotely hosted black-box models



The adversary then uses the local substitute to craft adversarial examples, which are misclassified by the remote ML system because of transferability.

Cross-technique transferability



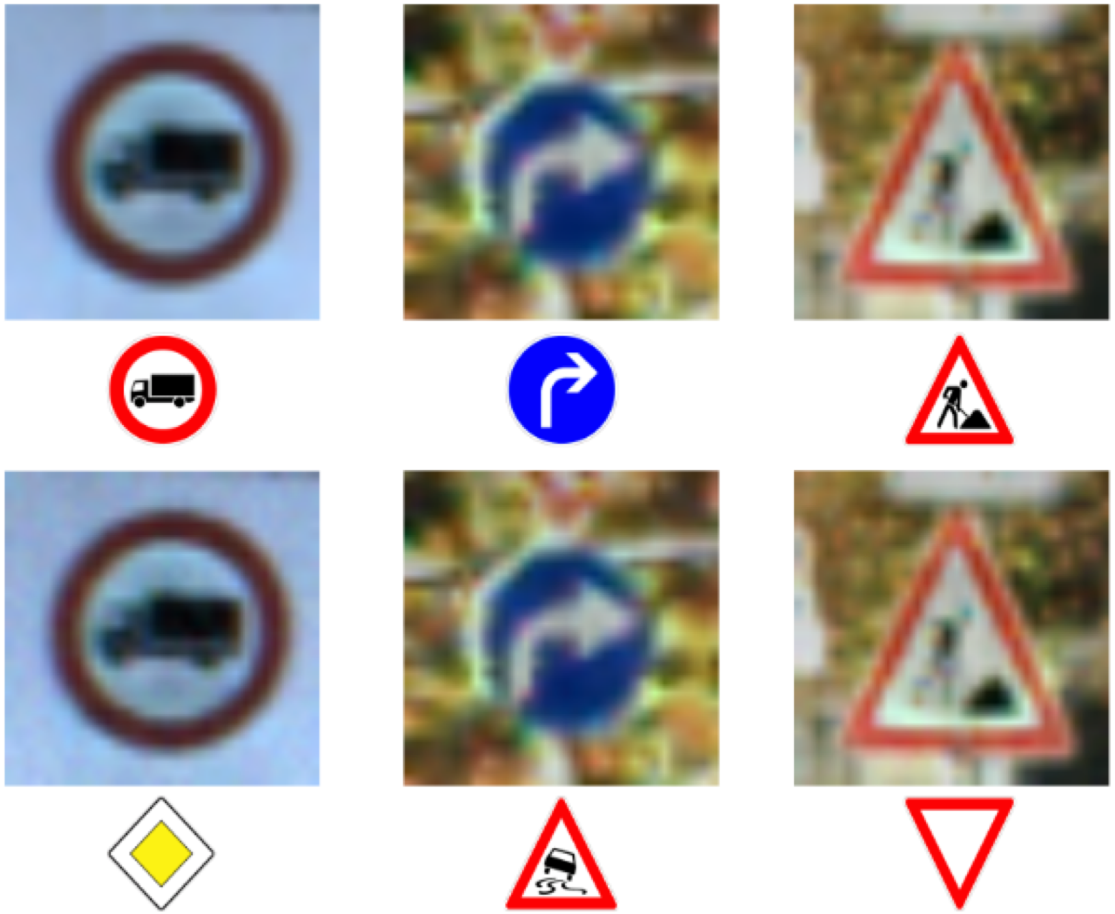


Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples [arXiv preprint]
Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow

Properly-blinded attacks on real-world remote systems

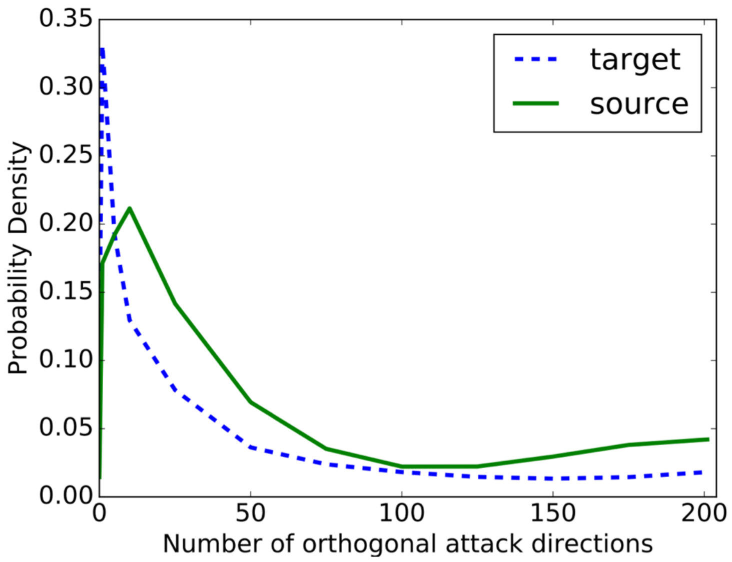
Remote Platform	ML technique	Number of queries	Adversarial examples misclassified (after querying)
MetaMind	Deep Learning	6,400	84.24%
amazon web services™	Logistic Regression	800	96.19%
Google Cloud Platform	Unknown	2,000	97.72%

All remote classifiers are trained on the MNIST dataset (10 classes, 60,000 training samples)



Learning models robust to adversarial examples is hard

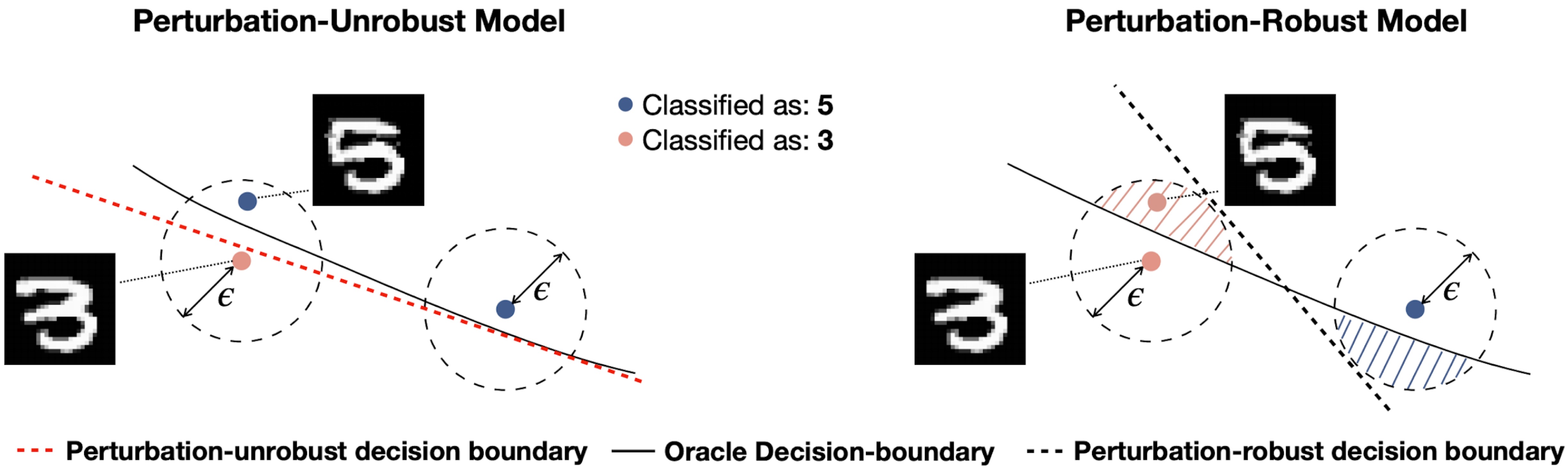
Error spaces containing adversarial examples are large



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

An example *toy* security policy: *the ℓ_p norm in vision*



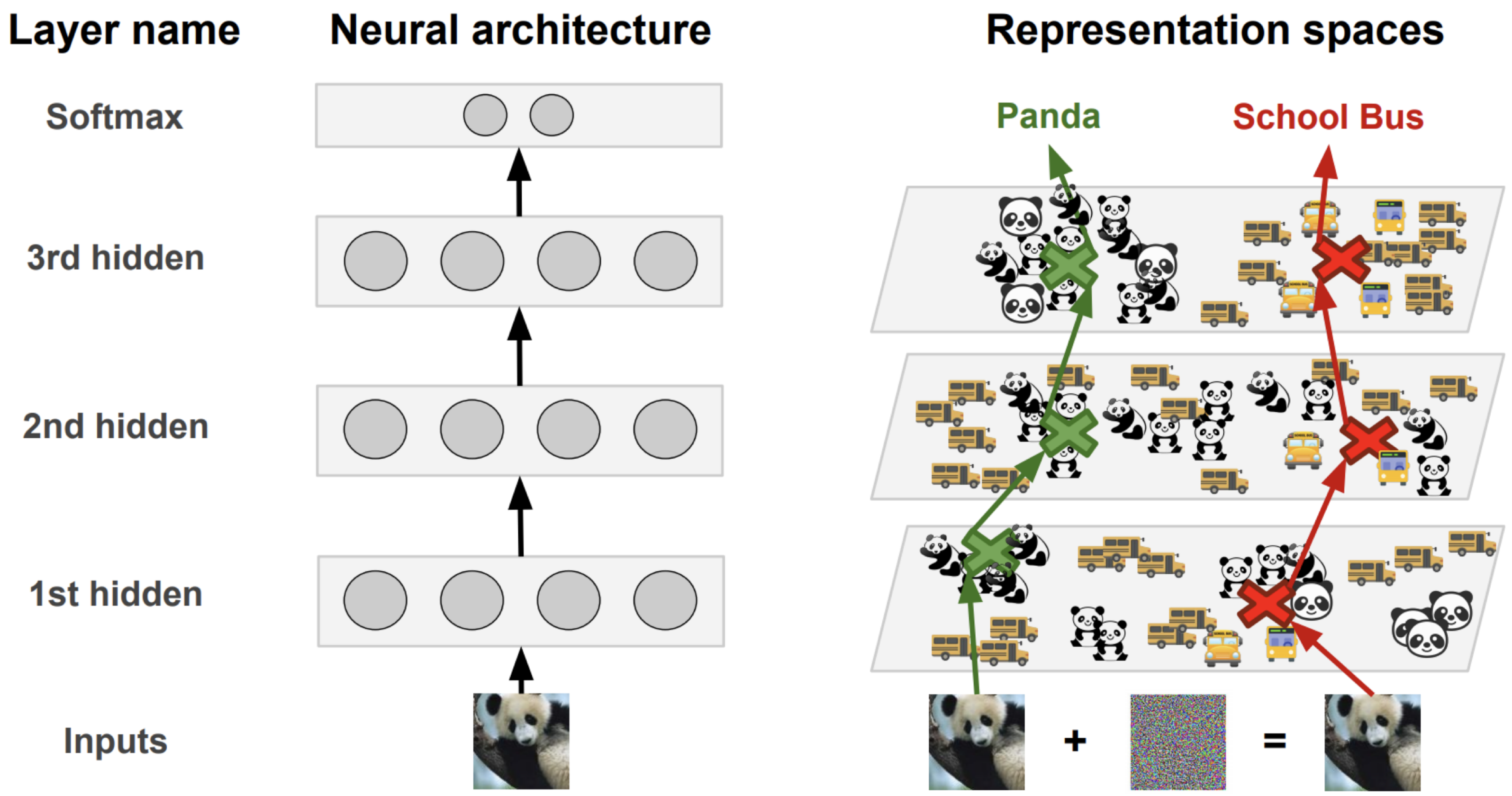
Exploiting Excessive Invariance caused by Norm-Bounded Adversarial Robustness (Jacobsen et al.)

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

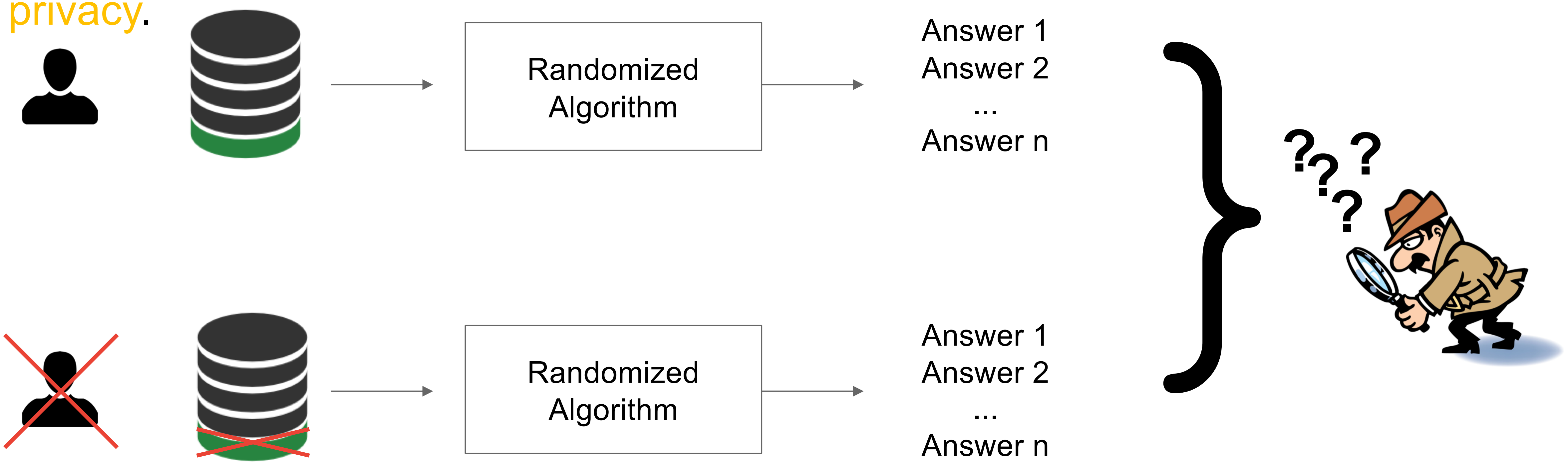
Soft Nearest Neighbor Loss (2019)
Frosst, Papernot and Hinton

Privacy in Machine Learning

What is a private algorithm?

Designing algorithms with privacy guarantees understood by humans is difficult.

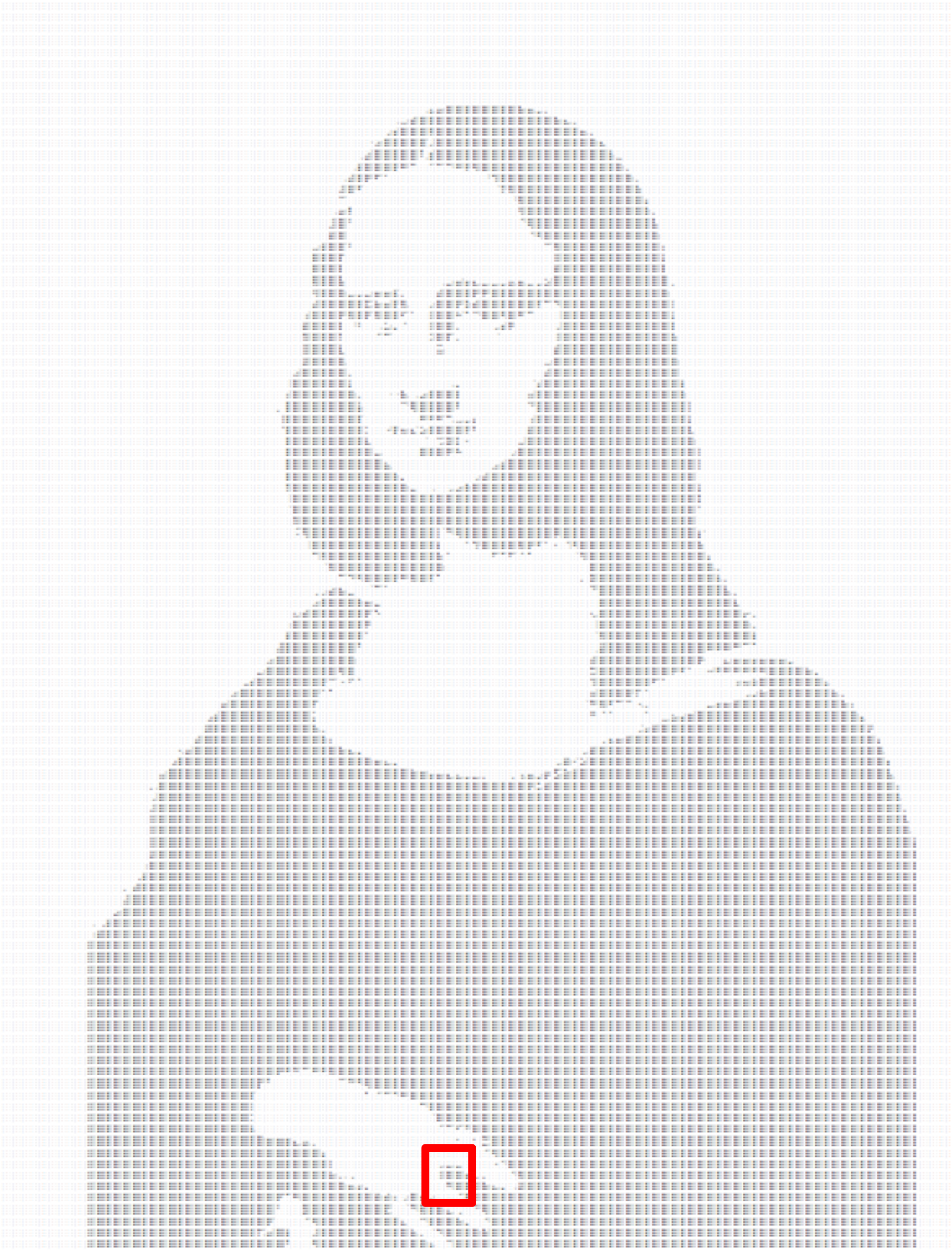
First question: how should we define privacy? Gold standard is now differential privacy.



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

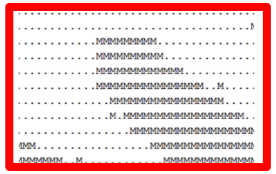
IACR:3650 (Dwork et al.)

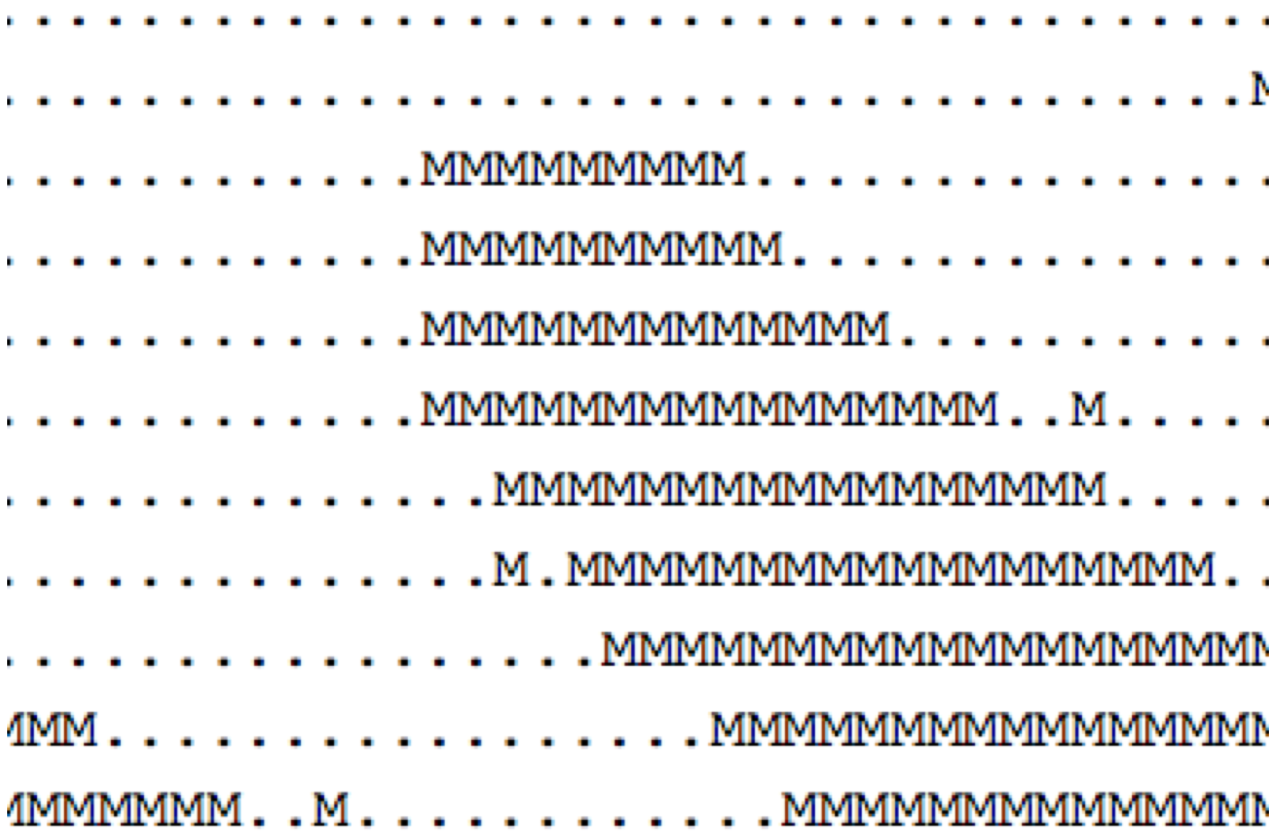
A Metaphor For Private Learning



Slides adapted from Ulfar Erlingsson

An Individual’s Training Data



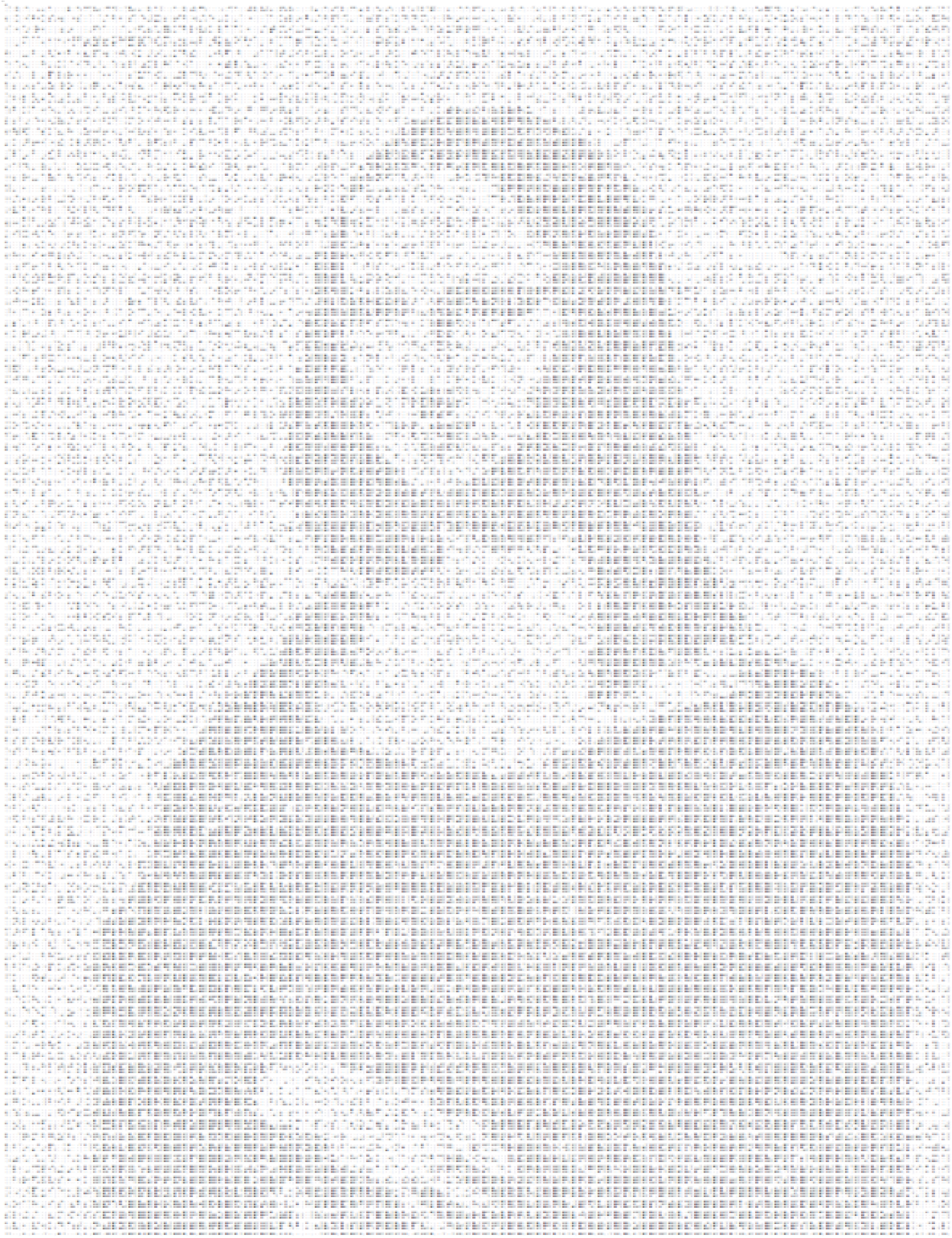


An Individual’s Training Data

Each bit is flipped with probability 25%

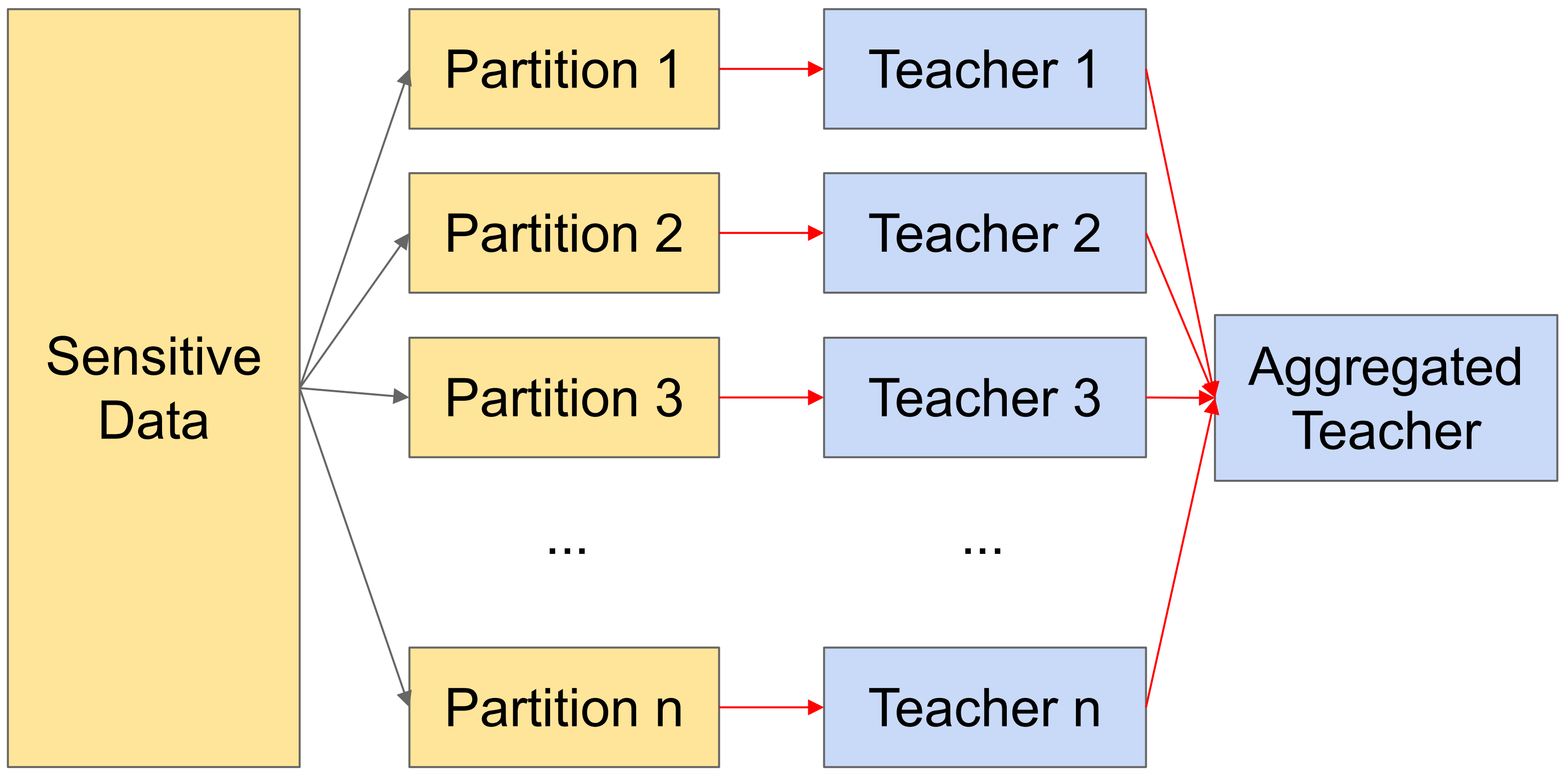
```
.......M........MM.M.......MMM.M..
......................MM...MMMM...
....M..MM.MM..MMM.M.MM.M...M..MM..
.MM......MMM....MMMMMMMMM...M...MM
..M....M........MM..MMMMMMM...M...
M.......M..MM.MMMMMMMMMMMMMMM....M
.....M.....M.M.M.MMMMMM...MMMMM...
...M.....M.MM.M.MM..M..M..MM.MMMMM
M...M.M.....M.M..M..MMM.MMMMM.MMMM
.MMM.M....M.M.M........MMMMMMMMM.M
```

Big Picture Remains!

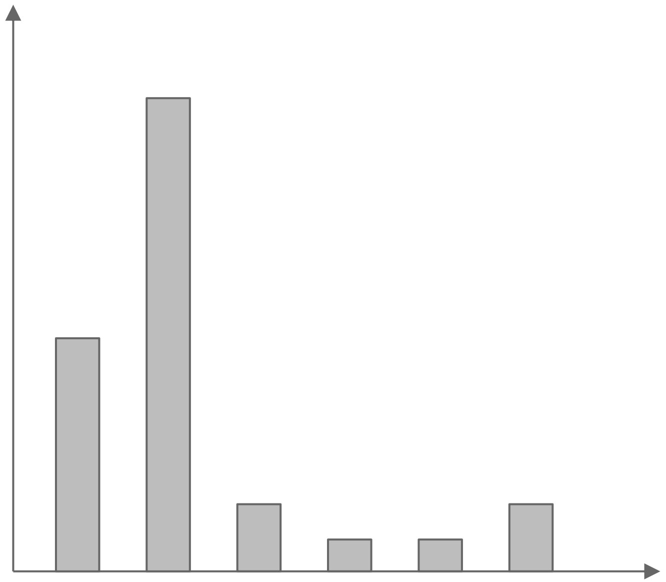


Slides adapted from Ulfar Erlingsson

PATE: Private Aggregation of Teacher Ensembles

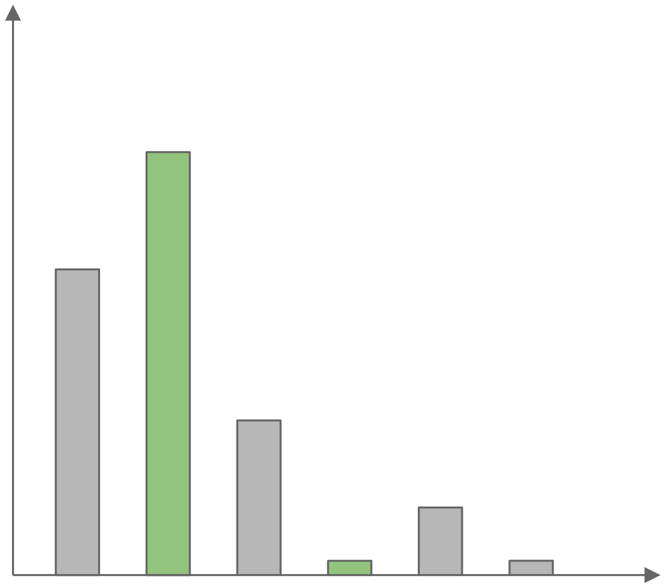


PATE: Private Aggregation of Teacher Ensembles



Count votes

$$n_j(\vec{x}) = |\{i : i \in 1..n, f_i(\vec{x}) = j\}|$$

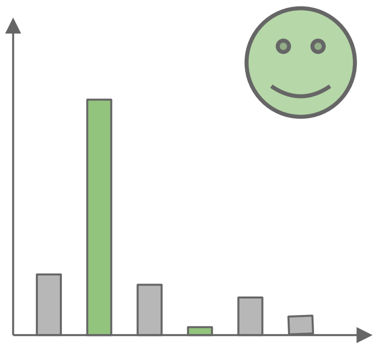


Take maximum

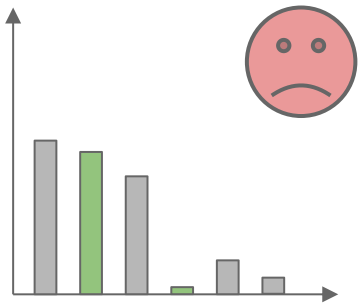
$$f(x) = \arg\max_j \left\{ n_j(\vec{x}) \right\}$$

PATE: Private Aggregation of Teacher Ensembles

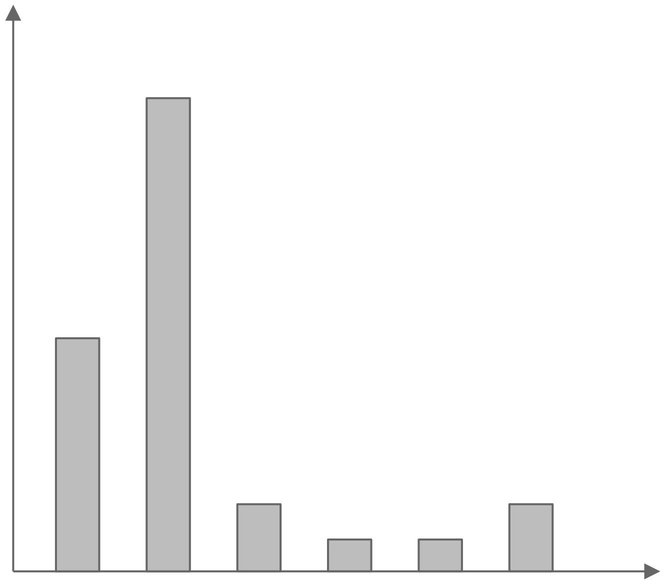
If most teachers agree on the label,
it does not depend on specific partitions,
so the privacy cost is small.



If two classes have close vote counts,
the disagreement may reveal private information.

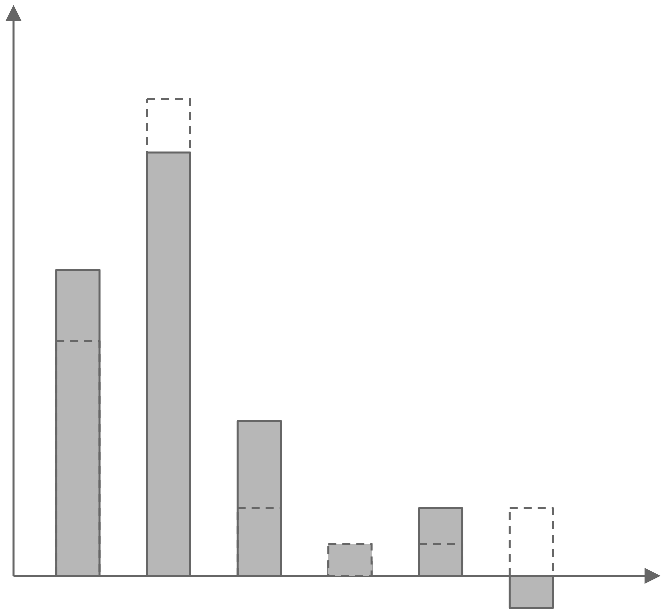


PATE: Private Aggregation of Teacher Ensembles



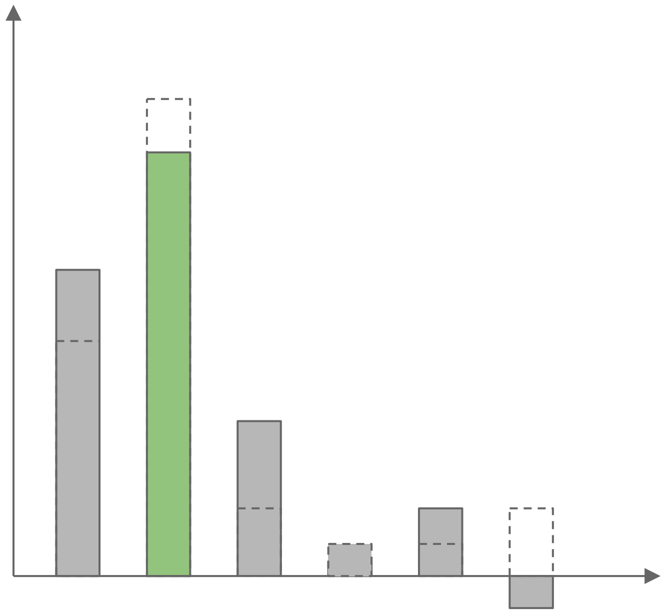
Count votes

$$n_j(\vec{x}) = |\{i : i \in 1..n, f_i(\vec{x}) = j\}|$$



Add Laplacian

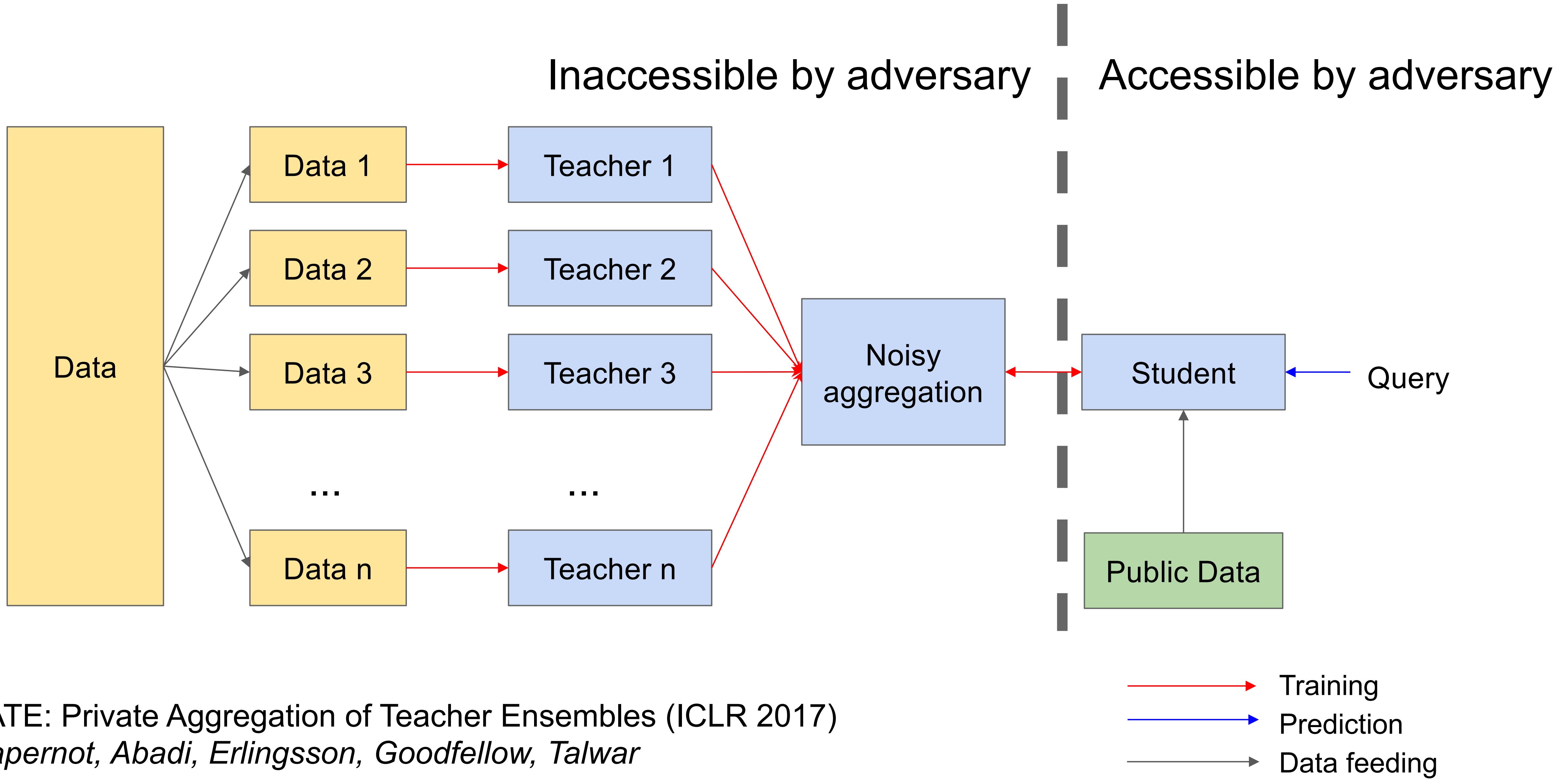
$$Lap\left(\frac{1}{\varepsilon}\right)$$



Take maximum

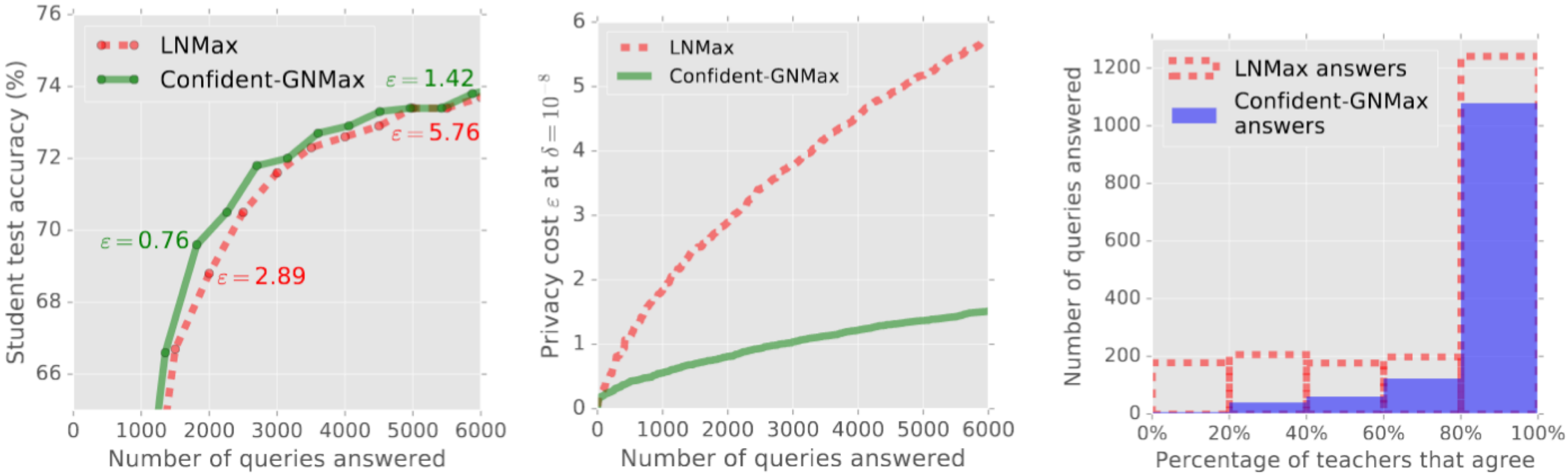
$$f(x) = \arg\max_j \left\{ n_j(\vec{x}) + Lap\left(\frac{1}{\varepsilon}\right) \right\}$$

PATE: Private Aggregation of Teacher Ensembles



PATE: Private Aggregation of Teacher Ensembles (ICLR 2017)
Papernot, Abadi, Erlingsson, Goodfellow, Talwar

Aligning privacy with generalization



Scalable Private Learning with PATE (Papernot*, Song* et al., ICLR 2018)

Conclusions

Saltzer and Schroeder’s principles

Economy of mechanism.
Keep the design of security mechanisms simple.

Fail-safe defaults.
Base access decisions on permission rather than exclusion.

Complete mediation.
Every access to an object is checked for authority.

Open design.
The design of security mechanisms should not be secret.

Separation of privilege.
A protection mechanism that requires two keys to unlock is more robust and flexible.

Least privilege.
Every user operates with least privileges necessary.

Least common mechanism.
Minimize mechanisms depended on by all users.

Psychological acceptability.
Human interface designed for ease of use.

Work factor.
Balance cost of circumventing the mechanism with known attacker resources.

Compromise recording.
Mechanisms that reliably record compromises can be used in place of mechanisms that prevent loss.

Efforts need to specify ML security and privacy policies.

What is the right abstraction and/or language to formalize security and privacy requirements with precise semantics and no ambiguity?

Efforts need to specify ML security and privacy policies.

What is the right abstraction and/or language to formalize security and privacy requirements with precise semantics and no ambiguity?

Admission control and auditing 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?

UNIVERSITY OF TORONTO VECTOR INSTITUTE

Efforts need to specify ML security and privacy policies.

What is the right abstraction and/or language to formalize security and privacy requirements with precise semantics and no ambiguity?

Admission control and auditing 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?

Security and privacy should strive to align with ML goals.

How do private learning and robust learning relate to generalization? How does poisoning relate to learning from noisy data or distribution drifts?

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

?

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

I’m hiring at UofT & Vector:

- Graduate students
- Postdocs