



@NicolasPapernot



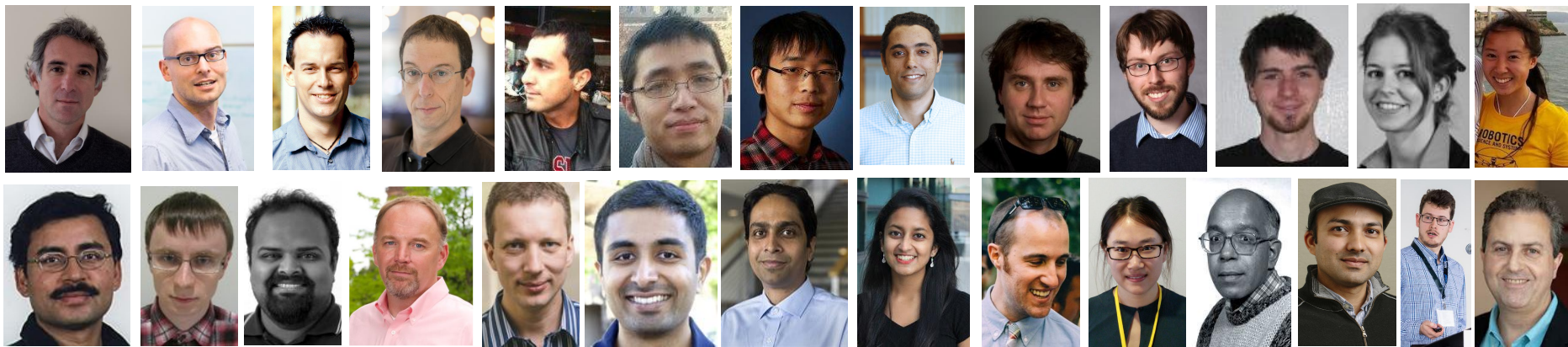
www.papernot.fr

Security and Privacy in Machine Learning

Nicolas Papernot

Work done at the Pennsylvania State University and Google Brain

July 2018 - MSR AI summer school



Martín Abadi (Google Brain)
 Pieter Abbeel (Berkeley)
 Michael Backes (CISPA)
 Dan Boneh (Stanford)
 Z. Berkay Celik (Penn State)
 Brian Cheung (UC Berkeley)
 Yan Duan (OpenAI)
 Gamaleldin F. Elsayed (Google Brain)
 Úlfar Erlingsson (Google Brain)
 Matt Fredrikson (CMU)
 Ian Goodfellow (Google Brain)
 Kathrin Grosse (CISPA)
 Sandy Huang (UC Berkeley)
 Somesh Jha (U of Wisconsin)

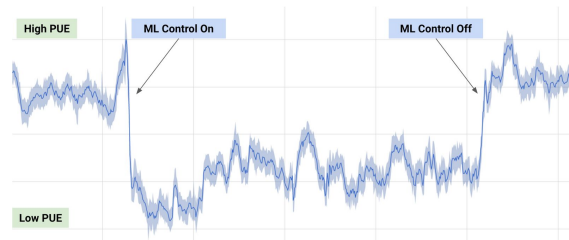
Alexey Kurakin (Google Brain)
 Praveen Manoharan (CISPA)
 Patrick McDaniel (Penn State)
 Ilya Mironov (Google Brain)
 Ananth Raghunathan (Google Brain)
 Arunesh Sinha (U of Michigan)
 Shreya Shankar (Stanford)
 Jascha Sohl-Dickstein (Google Brain)
 Shuang Song (UCSD)
 Ananthram Swami (US ARL)
 Kunal Talwar (Google Brain)
 Florian Tramèr (Stanford)
 Michael Wellman (U of Michigan)
 Xi Wu (Google)

Machine learning brings social disruption at scale



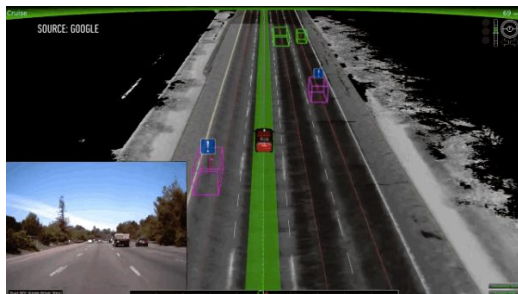
Healthcare

Source: Peng and Gulshan (2017)



Energy

Source: Deepmind



Transportation

Source: Google

Q1 [3pt] What is the integral of x^2 ?

Midterm 1
TOTAL POINTS
6 / 8 pts

QUESTION 1
Calculus 4 / 6 pts

11 Integral 1 / 3 pts
0 pts Correct
✓ -1 pt Missing 1/2
✓ -1 pt Missing Constant
See me after class tomorrow

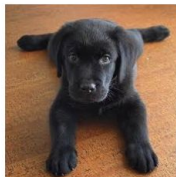
12 Derivative 2 / 2 pts

Download Submission History Request Regrade Next Question

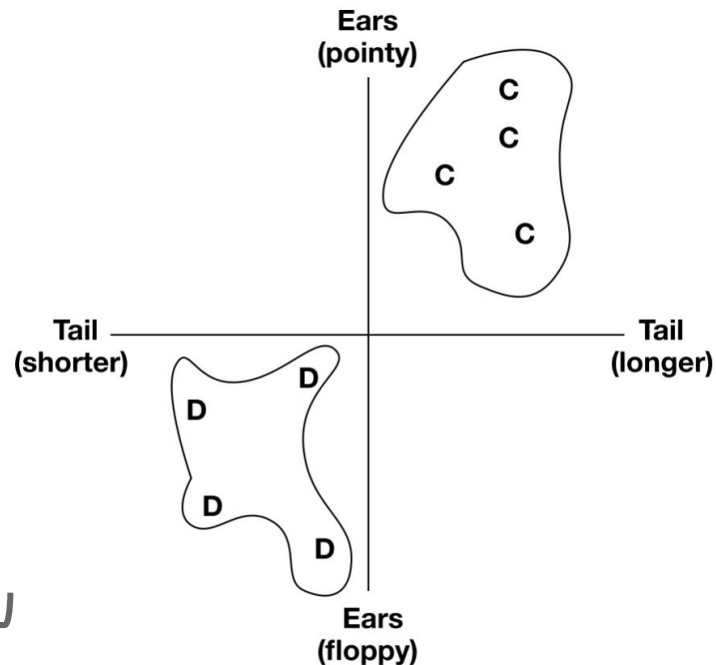
Education

Source: Gradescope

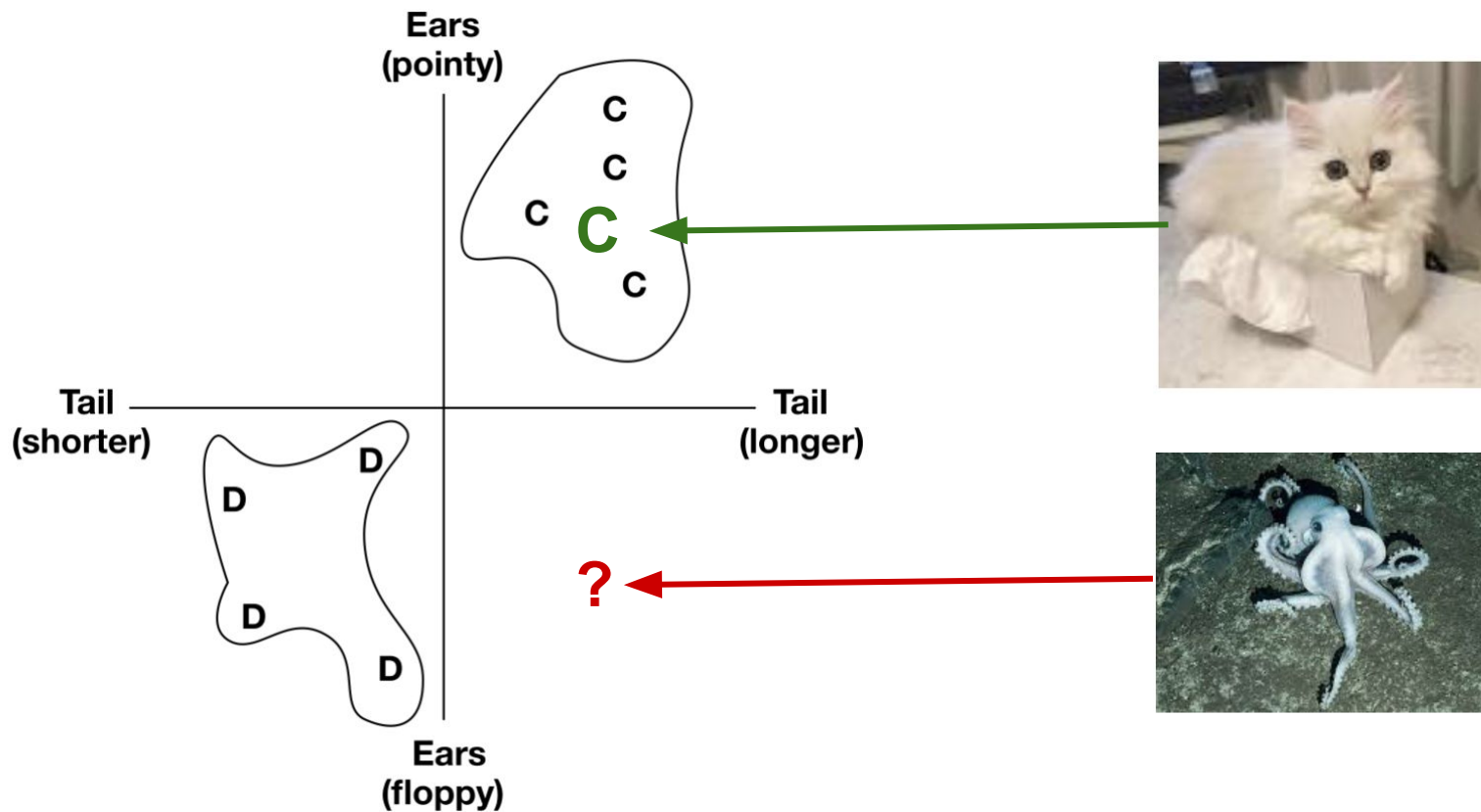
Machine learning is not magic (training time)



Training data



Machine learning is not magic (inference time)



Machine learning is deployed in adversarial settings



TayTweets ✓
@TayandYou



Following

@godblessameriga WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT

Tay chatbot

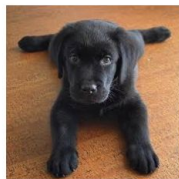
Training data poisoning



YouTube filtering

Content evades detection at *inference*

Machine learning does not always generalize well (1/2)



Cat 
Dog 

Cat 
Dog 

Training data

Test data

What if the adversary systematically poisoned the data?


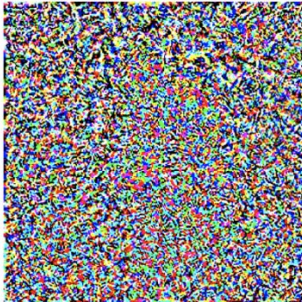


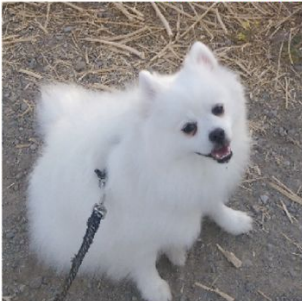



A small perturbation to one **training** example:

Label: Fish

$+ \epsilon \cdot$

Label: Fish

Can change multiple **test** predictions:

				
				
Orig (confidence): Dog (97%)	Dog (98%)	Dog (98%)	Dog (99%)	Dog (98%)
New (confidence): Fish (97%)	Fish (93%)	Fish (87%)	Fish (63%)	Fish (52%)

(Understanding Black-box Predictions via Influence Functions, Koh and Liang)

What if the adversary systematically evaded at inference time?



x
“panda”
57.7% confidence

+ .007 ×



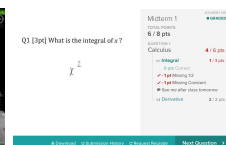
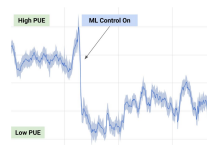
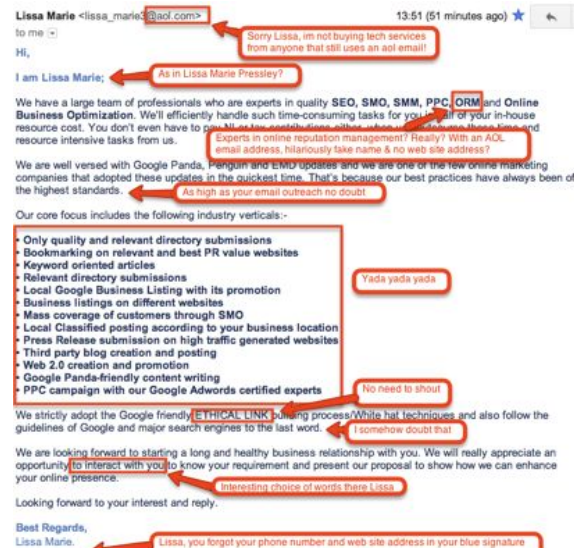
$\text{sign}(\nabla_x J(\theta, x, y))$
“nematode”
8.2% confidence

=

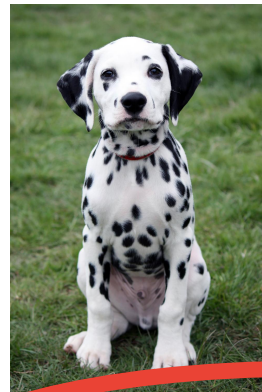
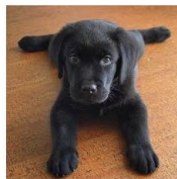




$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$
“gibbon”
99.3 % confidence

(Goodfellow et al., 2014)



Machine learning does not always generalize well (2/2)



Cat 
Dog 

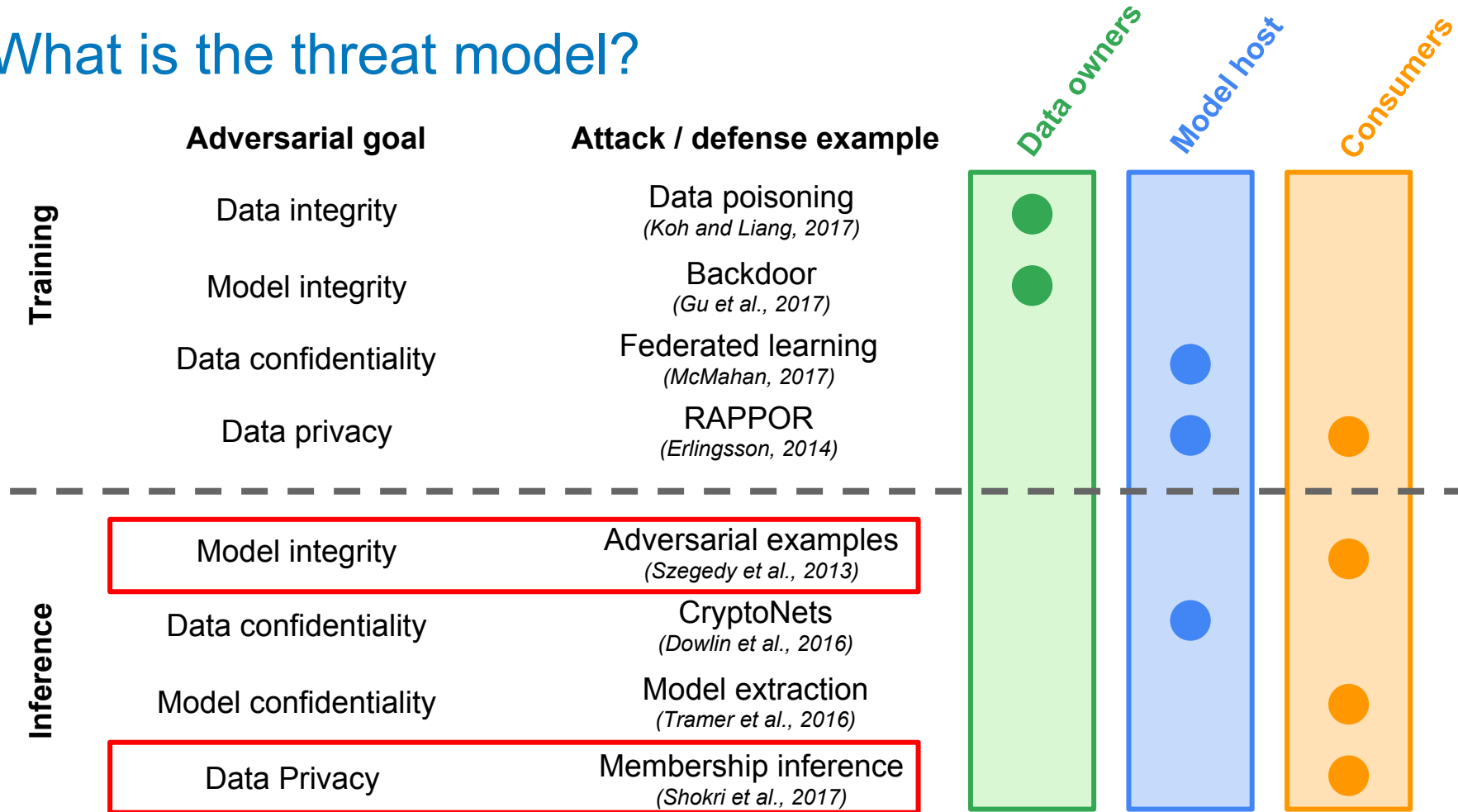
Cat 
Dog 

Training data

Test data

Membership inference attack (Shokri et al.)

What is the threat model?



Attacking Machine Learning Integrity with Adversarial Examples

The threat model

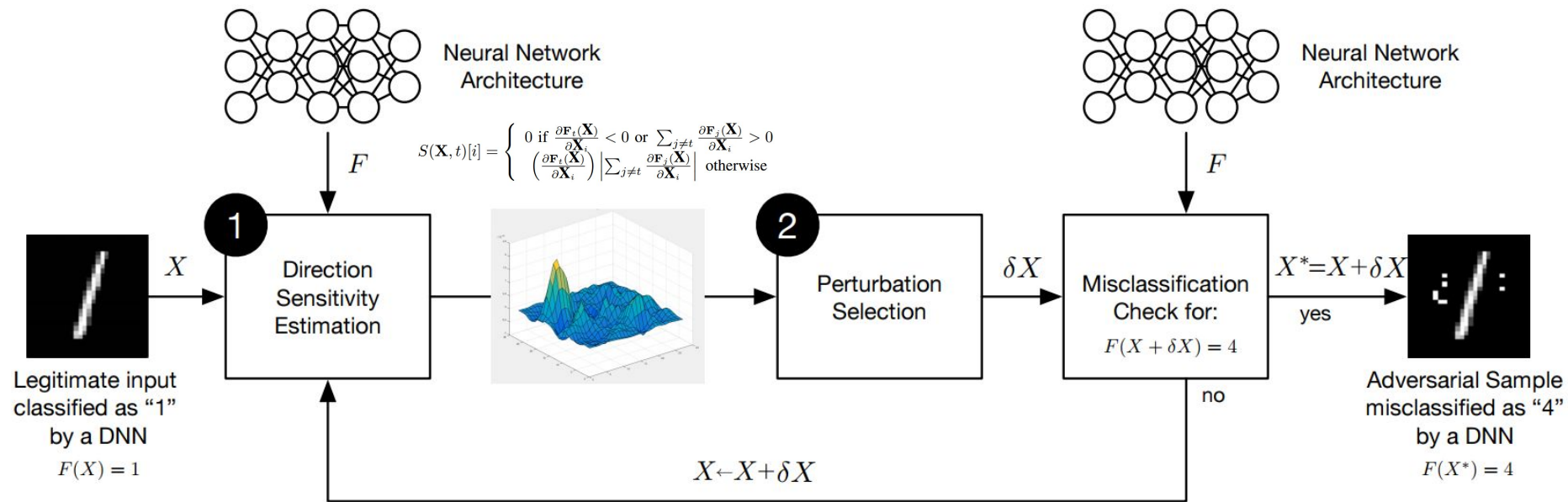
Attacker may see the model: attacker needs to know details of the machine learning model to do an attack --- aka a ***white-box attacker***

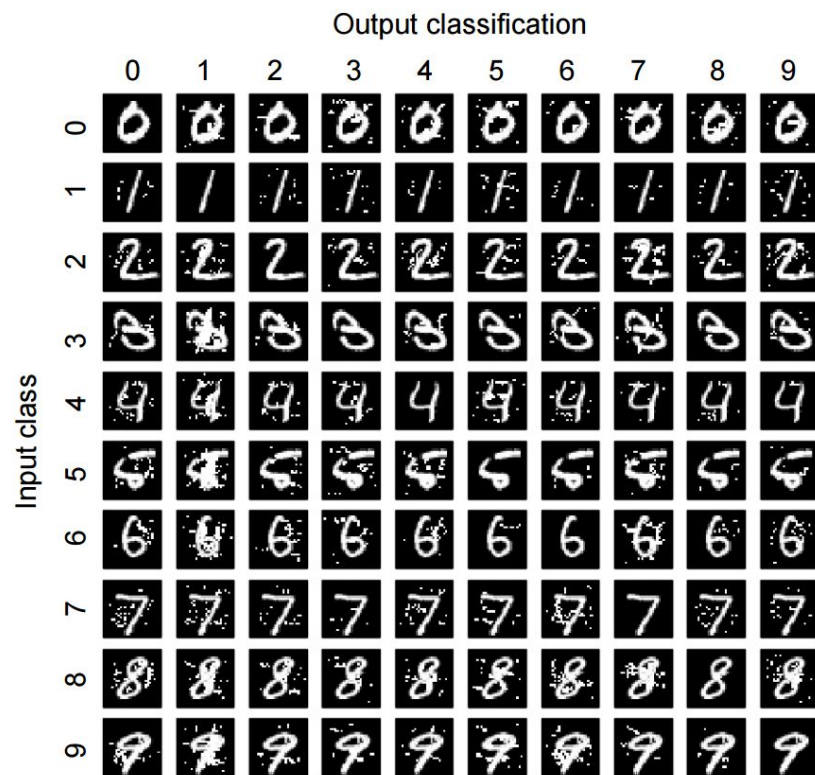


Attacker may not see the model: attacker who knows very little (e.g. only gets to ask a few questions) --- aka a ***black-box attacker***



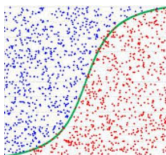
Jacobian-based Saliency Map Approach (JSMA)



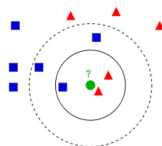


Adversarial examples...

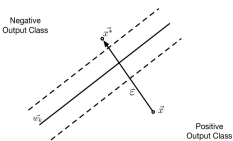
... *beyond deep learning*



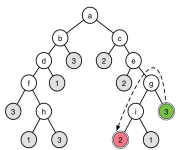
Logistic Regression



Nearest Neighbors



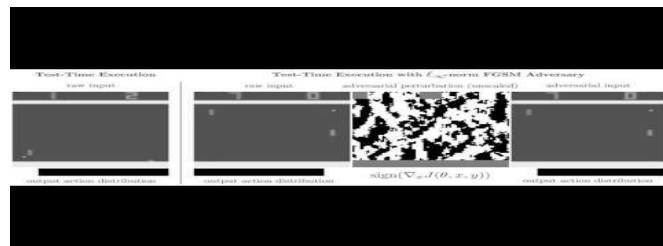
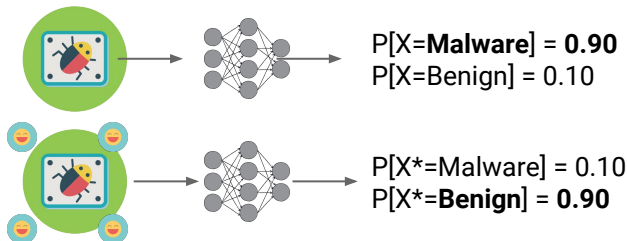
Support Vector Machines



Decision Trees

Useful to think about definitions and threat model

... *beyond computer vision*



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]

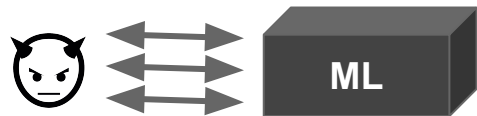
Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, Patrick McDaniel

The threat model

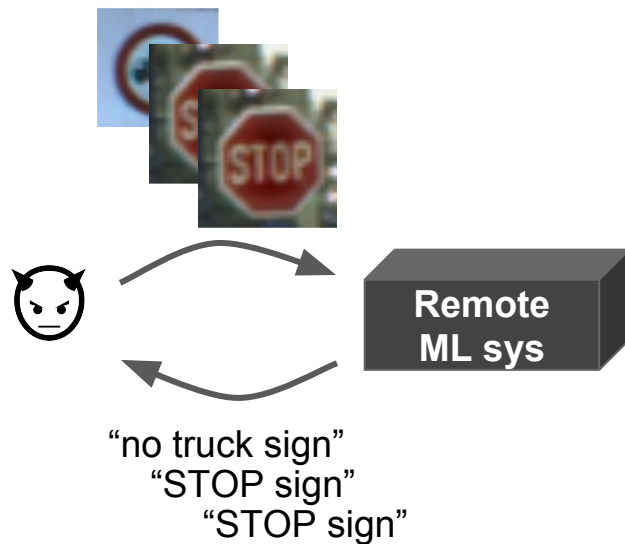
Attacker may see the model: attacker needs to know details of the machine learning model to do an attack --- aka a ***white-box attacker***



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***

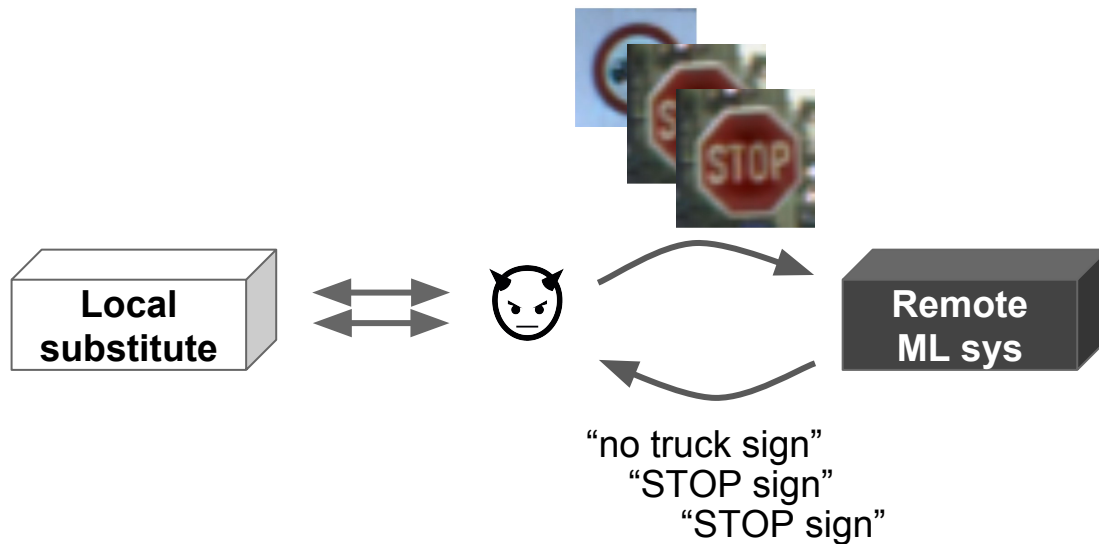


Attacking remotely hosted black-box models



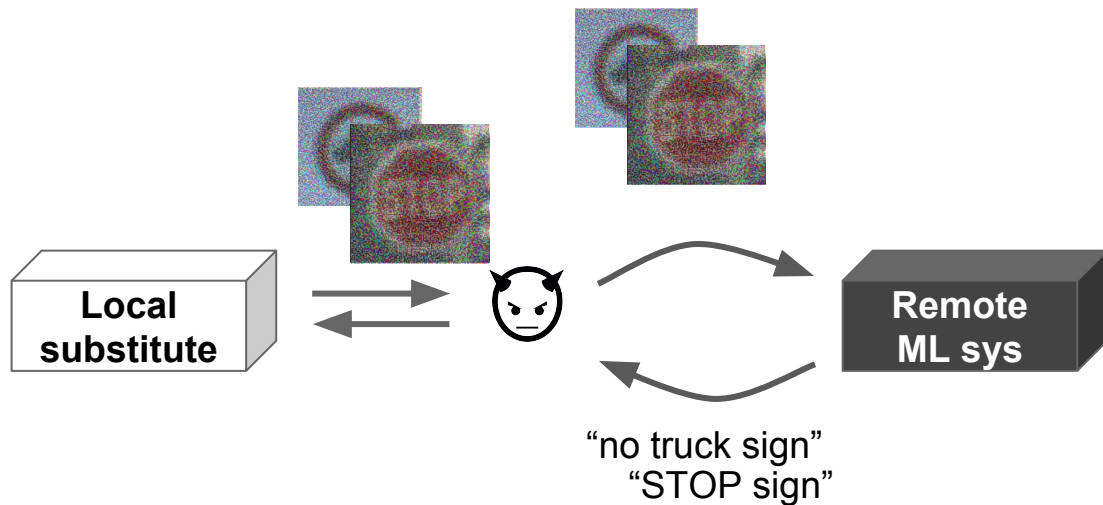
(1) The adversary queries remote ML system for labels on inputs of its choice.

Attacking remotely hosted black-box models



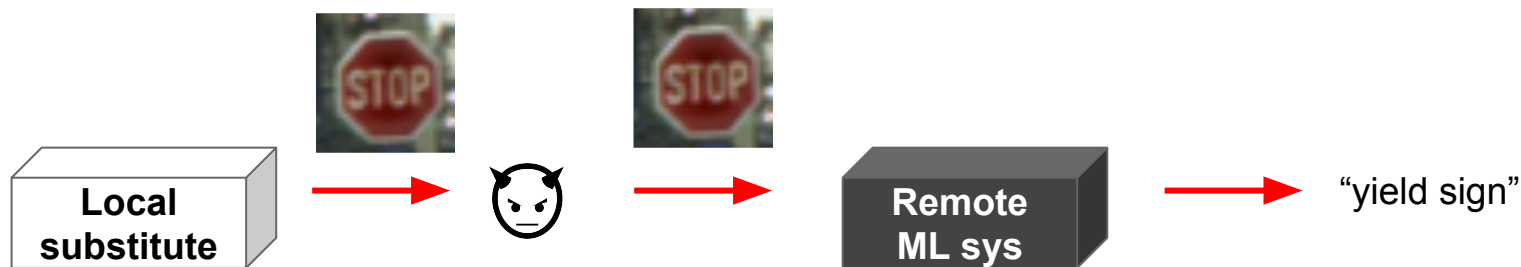
(2) The adversary uses this labeled data to train a local substitute for the remote system.

Attacking remotely hosted black-box models



- (3) 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






- (4) 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

Source Machine Learning Technique	DNN	38.27	23.02	64.32	79.31	8.36
	LR	6.31	91.64	91.43	87.42	11.29
	SVM	2.51	36.56	100.0	80.03	5.19
	DT	0.82	12.22	8.85	89.29	3.31
	kNN	11.75	42.89	82.16	82.95	41.65
		DNN	LR	SVM	DT	kNN
Target Machine Learning Technique						



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%

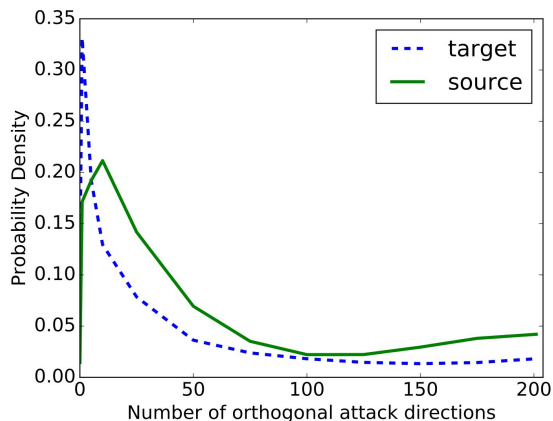
All remote classifiers are trained on the MNIST dataset (10 classes, 60,000 training samples)



Defending against adversarial examples

Learning models robust to adversarial examples is hard

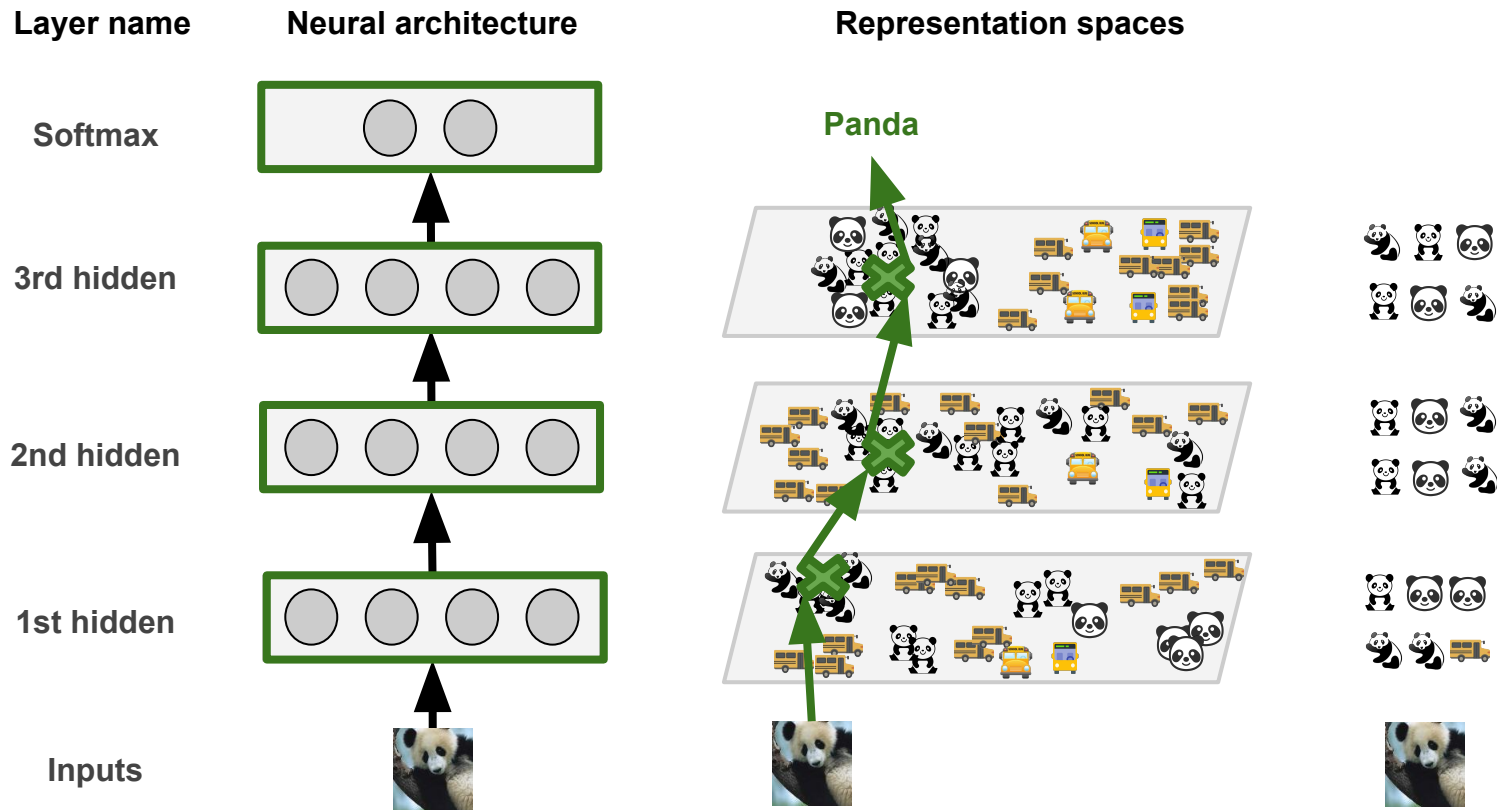
Error spaces containing adversarial examples are large



Learning or detecting adversarial examples creates an arms race



What makes a successful deep neural network?



What makes a successful adversarial example?

Layer name

Neural architecture

Representation spaces

Nearest neighbors

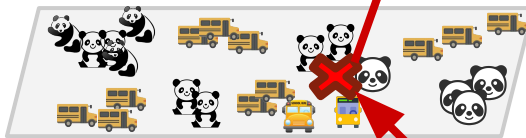
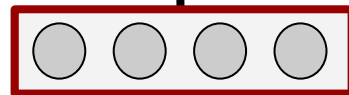
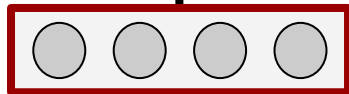
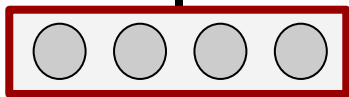
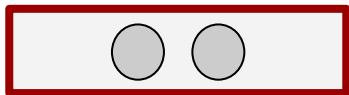
Softmax

3rd hidden

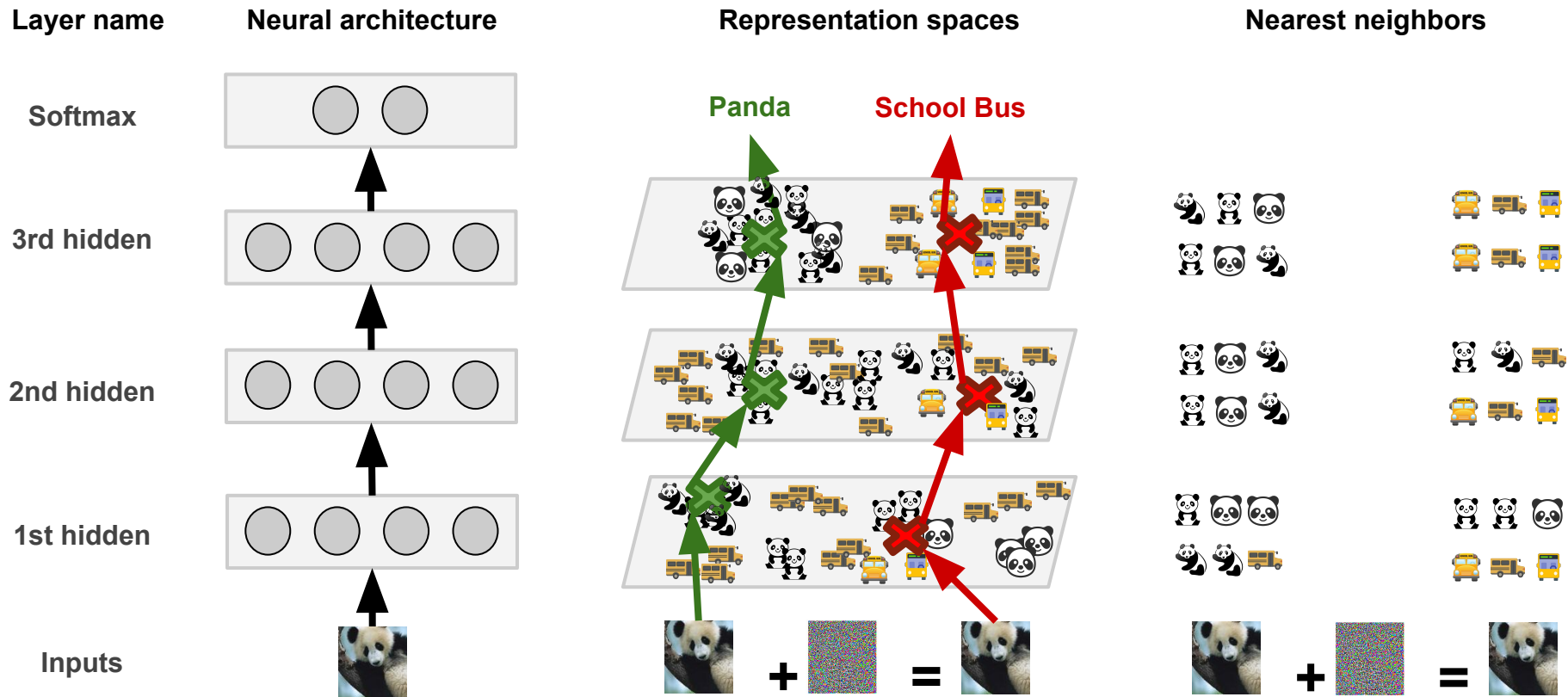
2nd hidden

1st hidden

Inputs



Nearest neighbors indicate support from training data...



... Deep k-Nearest Neighbors (DkNN) classifier

1. Searches for **nearest neighbors** in the training data at each layer
2. Estimates the **nonconformity** of input x for each possible label y
3. Apply conformal prediction to compute:

- a. **Confidence**

“How likely is the prediction given the training data?”

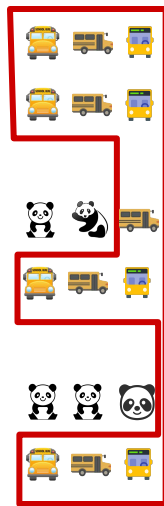
- b. **Credibility**

“How relevant is the training data to the prediction?”

Panda



School Bus



+

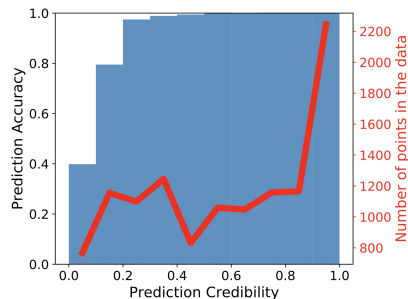


=

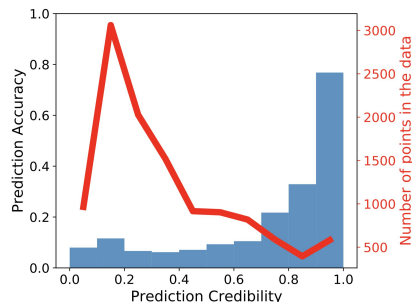


Example applications of DkNN credibility

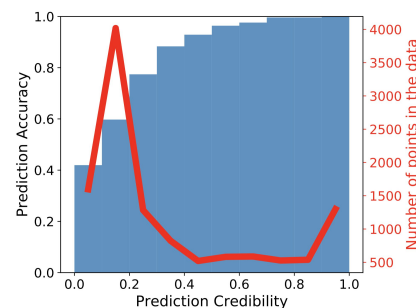
Adversarial examples



Clean inputs

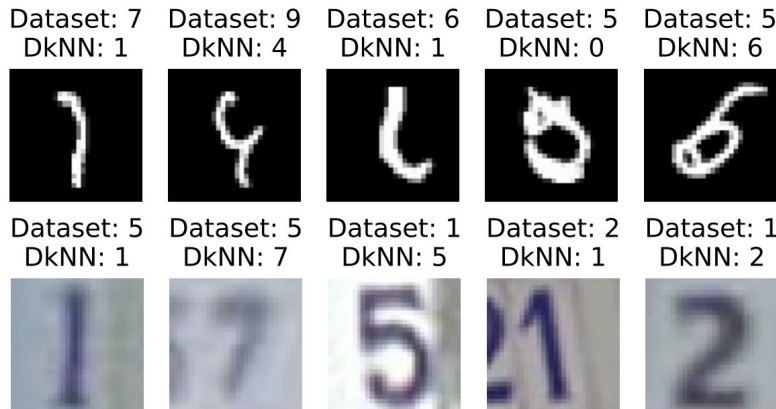


Basic Iterative Method

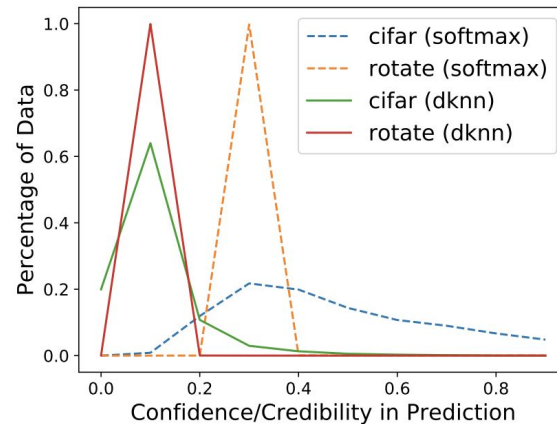


Carlini & Wagner

Mislabeled inputs

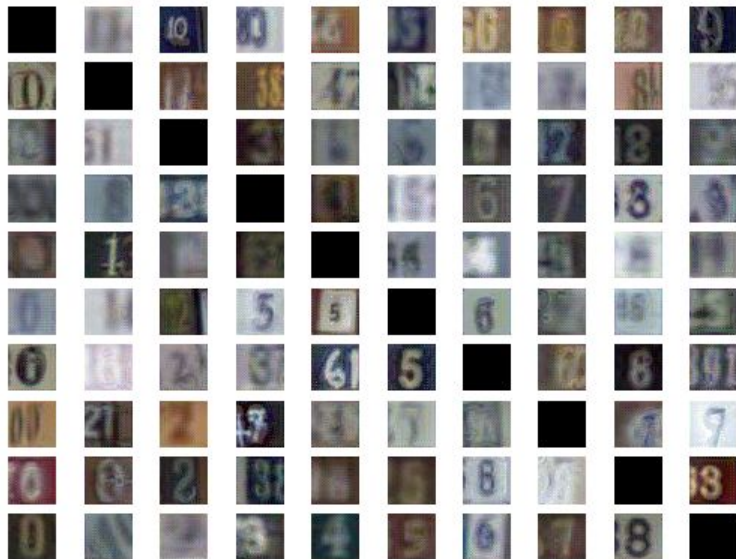


Out of distribution



Implications for the attacker and defender

Attacker



Defender

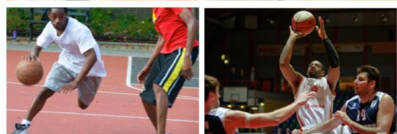
Reject low credibility predictions:

-> explicit tradeoff between clean accuracy and adversarial accuracy

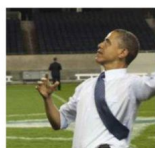
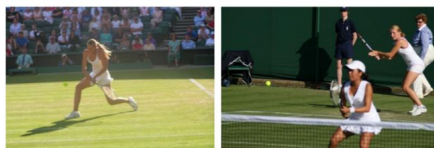
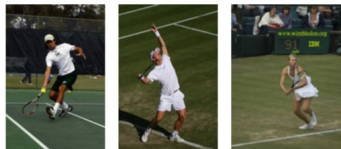
Active learning: more training data through human labeling of rejected predictions

Contributes to breaking “black-box” myth

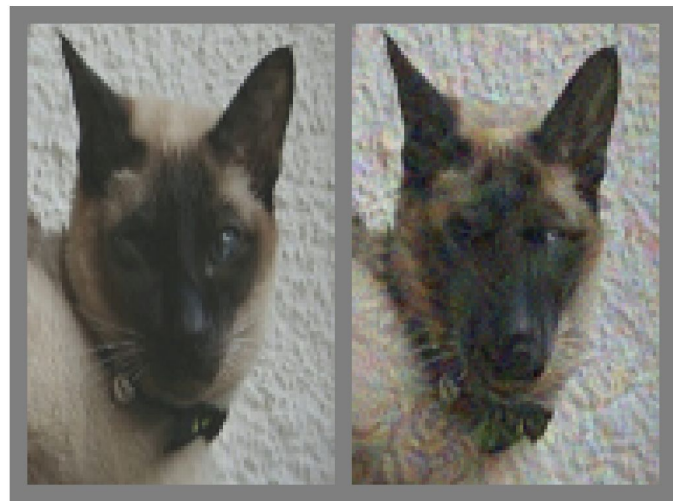
Some (surprising) connections to fairness & interpretability



Prediction: Basketball (68%)



Prediction: Racket (49%)



Adversarial Examples that Fool both Human and Computer Vision [arXiv preprint]
Gamaleldin F. Elsayed, Shreya Shankar, Brian Cheung,
Nicolas Papernot, Alex Kurakin, Ian Goodfellow, Jascha
Sohl-Dickstein

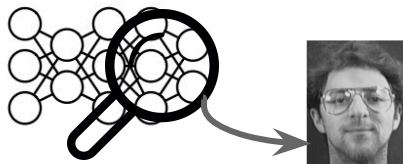
Machine Learning with Privacy

Types of adversaries and our threat model



Model querying (**black-box adversary**)

Shokri et al. (2016) *Membership Inference Attacks against ML Models*
Fredrikson et al. (2015) *Model Inversion Attacks*



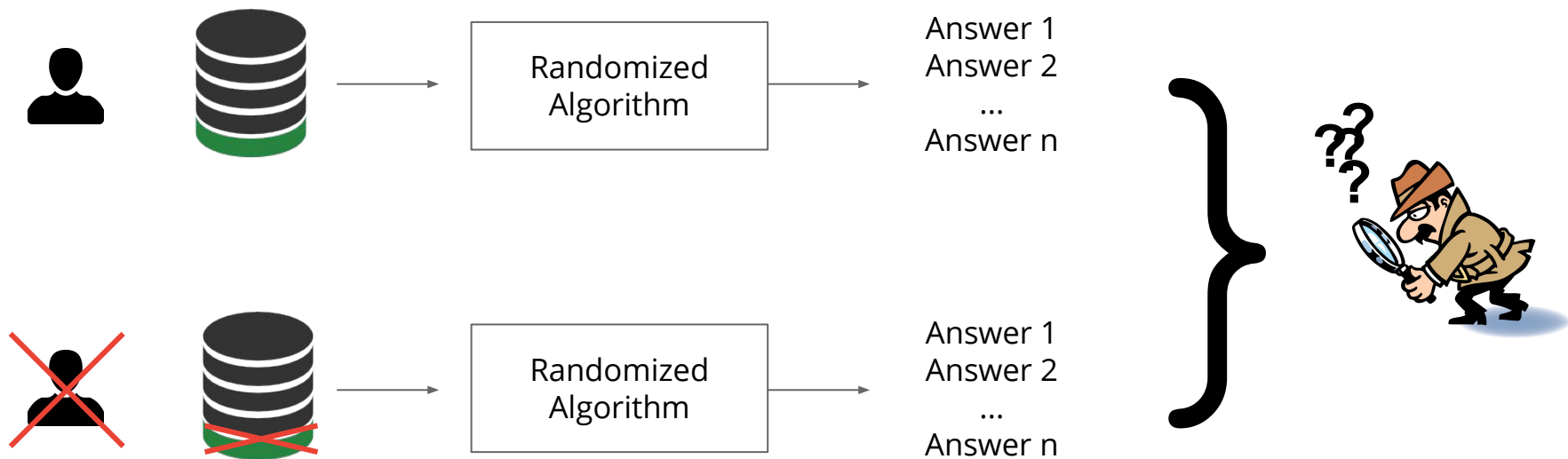
Model inspection (**white-box adversary**)

Zhang et al. (2017) *Understanding DL requires rethinking generalization*

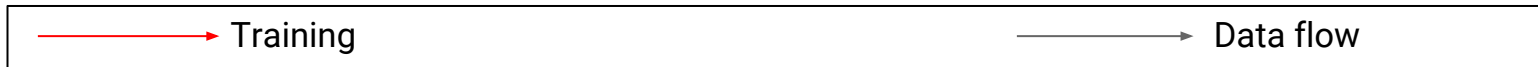
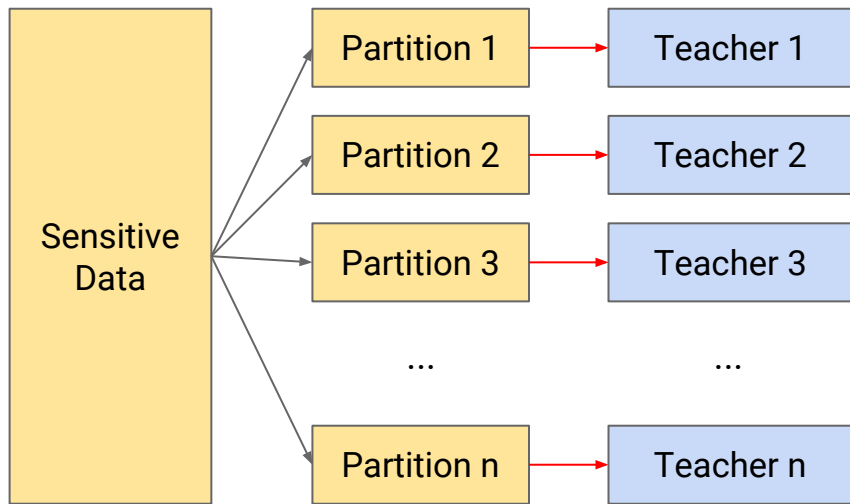
In our work, the threat model assumes:

- Adversary can make a potentially unbounded number of queries
- Adversary has access to model internals

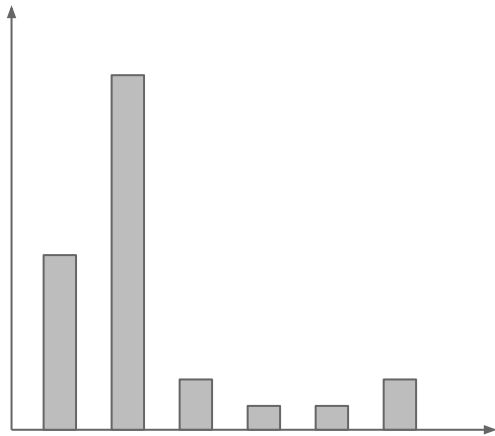
A definition of privacy: *differential privacy*



Private Aggregation of Teacher Ensembles (PATE)

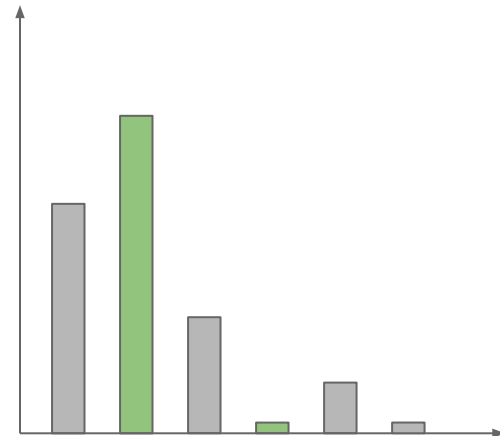


Aggregation



Count votes

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

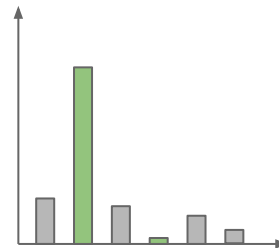


Take maximum

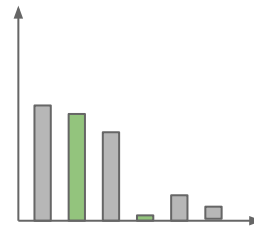
$$f(x) = \arg \max_j \left\{ n_j(\vec{x}) \right\}$$

Intuitive privacy analysis

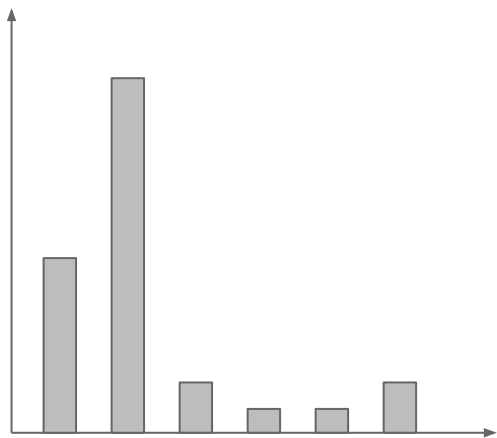
If most teachers agree on the label, it does not depend on specific partitions, so the privacy cost is small.



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

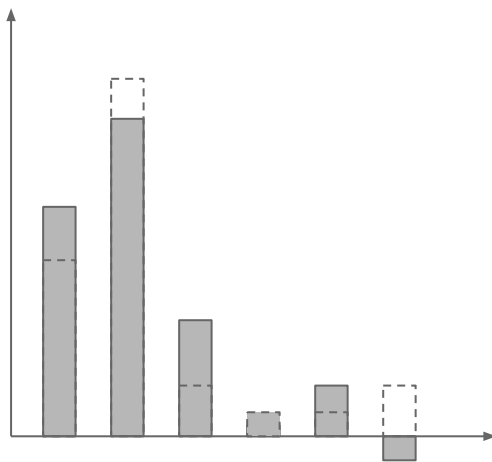


Noisy aggregation



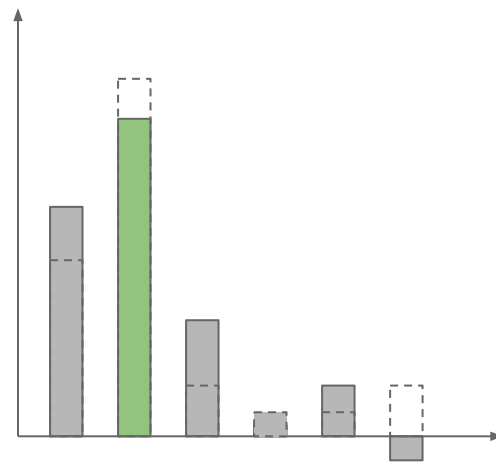
Count votes

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



Add Laplacian noise

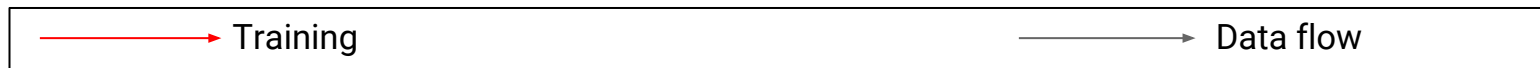
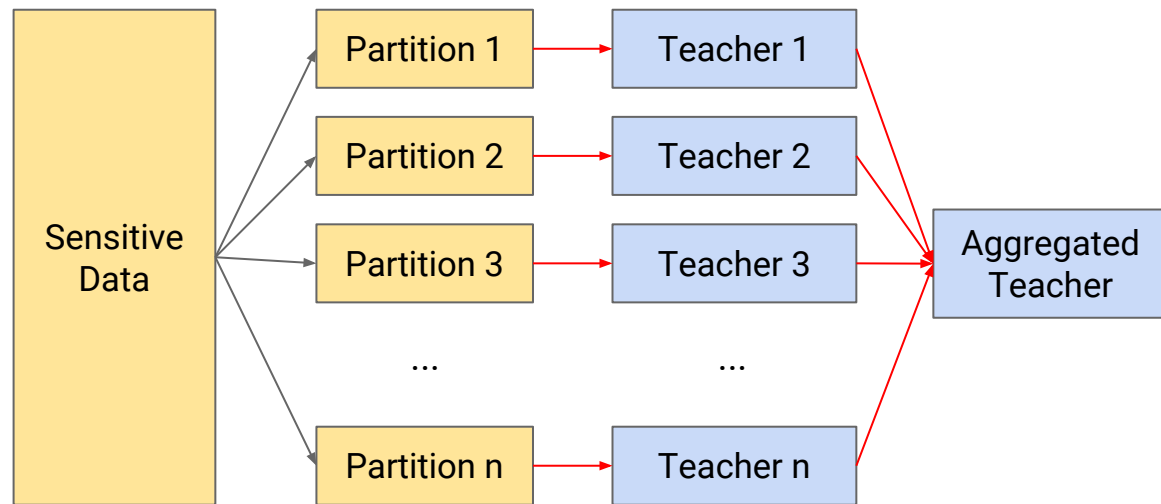
$$Lap\left(\frac{1}{\varepsilon}\right)$$



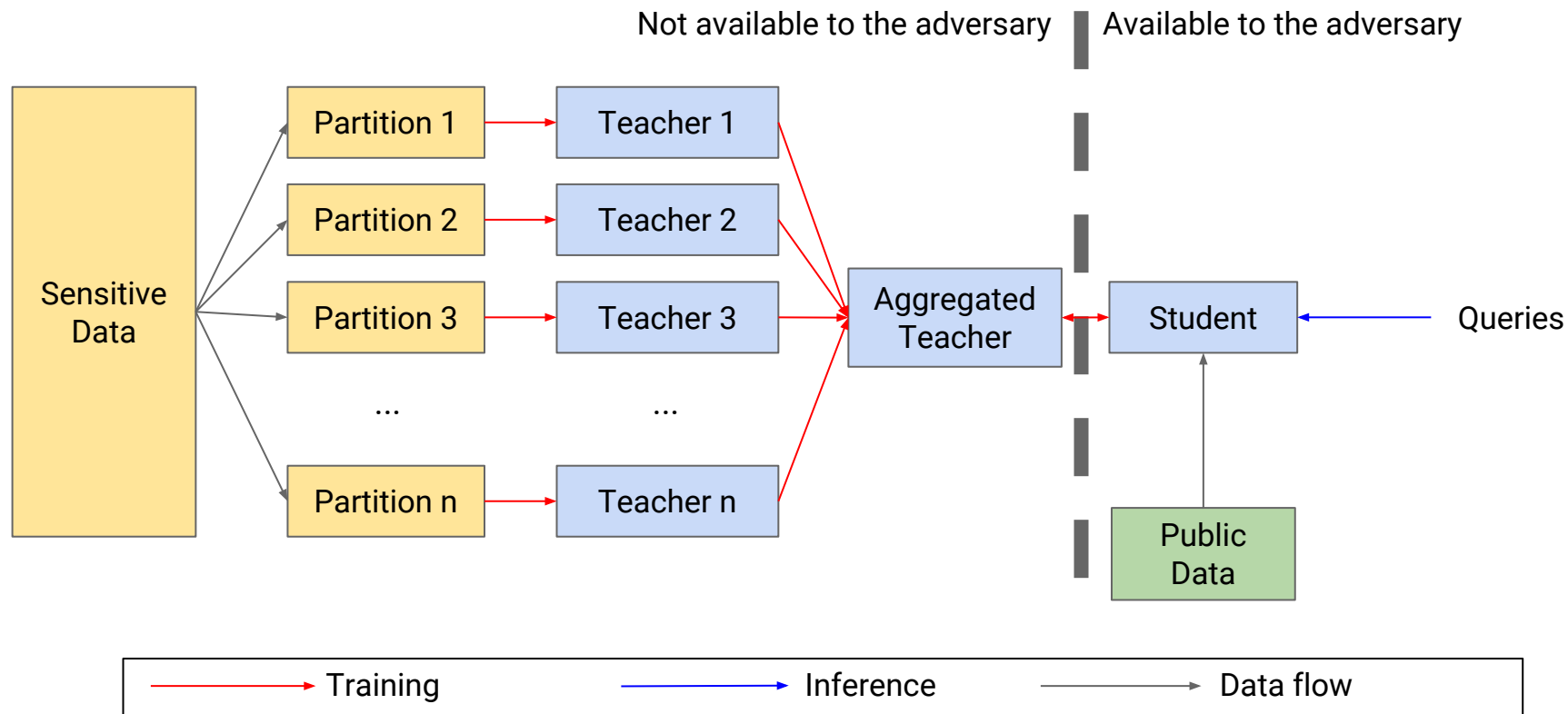
Take maximum

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

Teacher ensemble



Student training



Why train an additional “student” model?

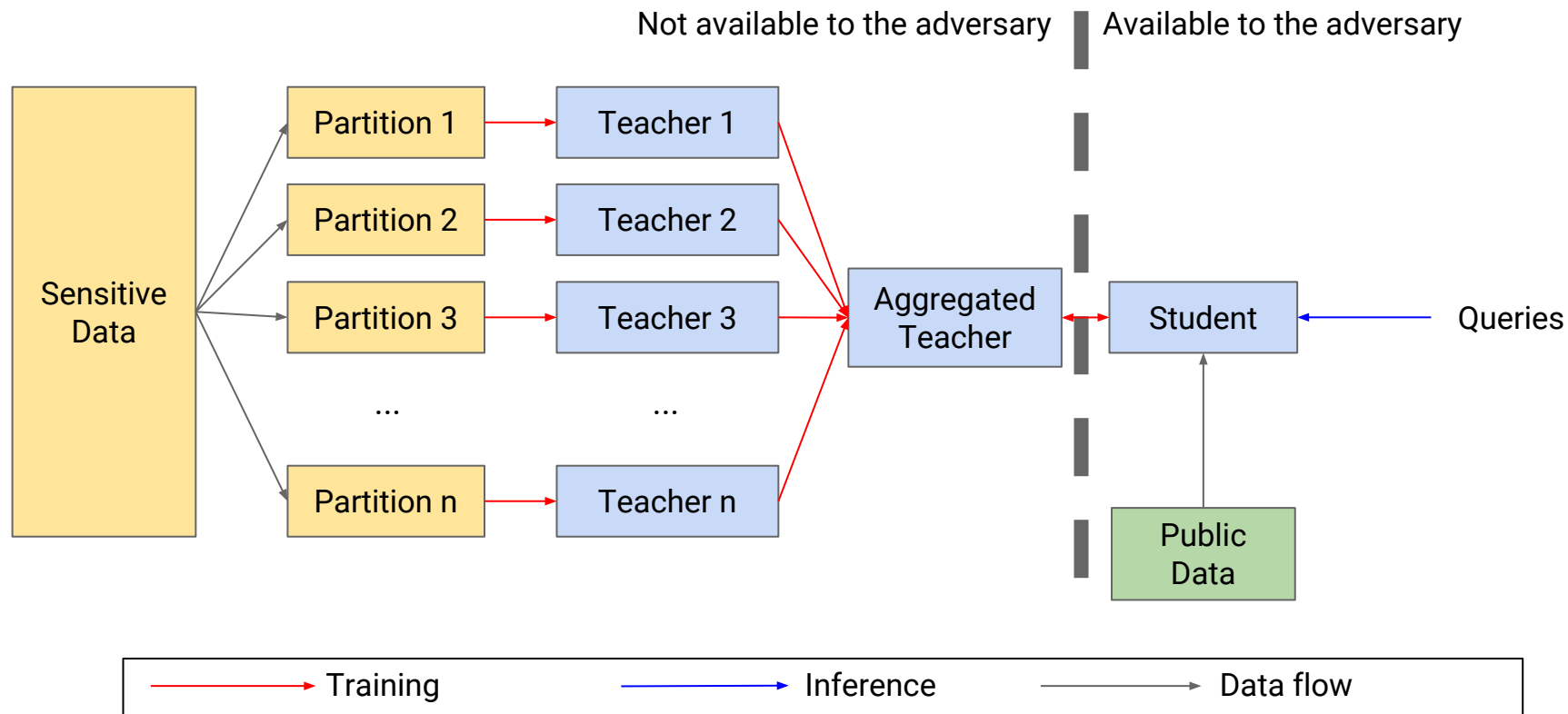
The aggregated teacher violates our threat model:

- 1 Each prediction increases total privacy loss.

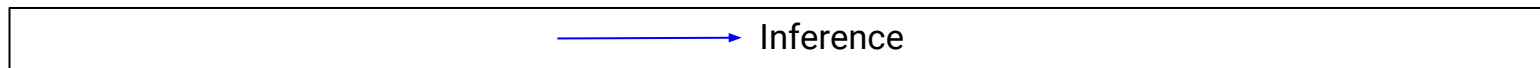
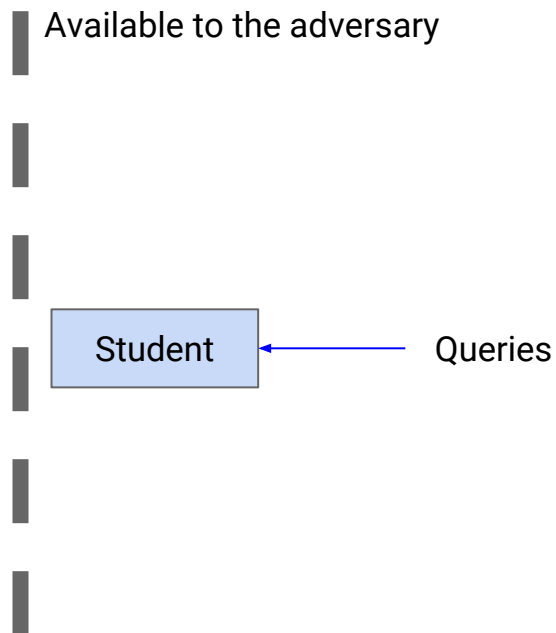
Privacy budgets create a tension between the accuracy and number of predictions.
- 2 Inspection of internals may reveal private data.

Privacy guarantees should hold in the face of white-box adversaries.

Student training



Deployment



Differential privacy analysis

Differential privacy:

A randomized algorithm M satisfies (ϵ, δ) differential privacy if for all pairs of neighbouring datasets (d, d') , for all subsets S of outputs:

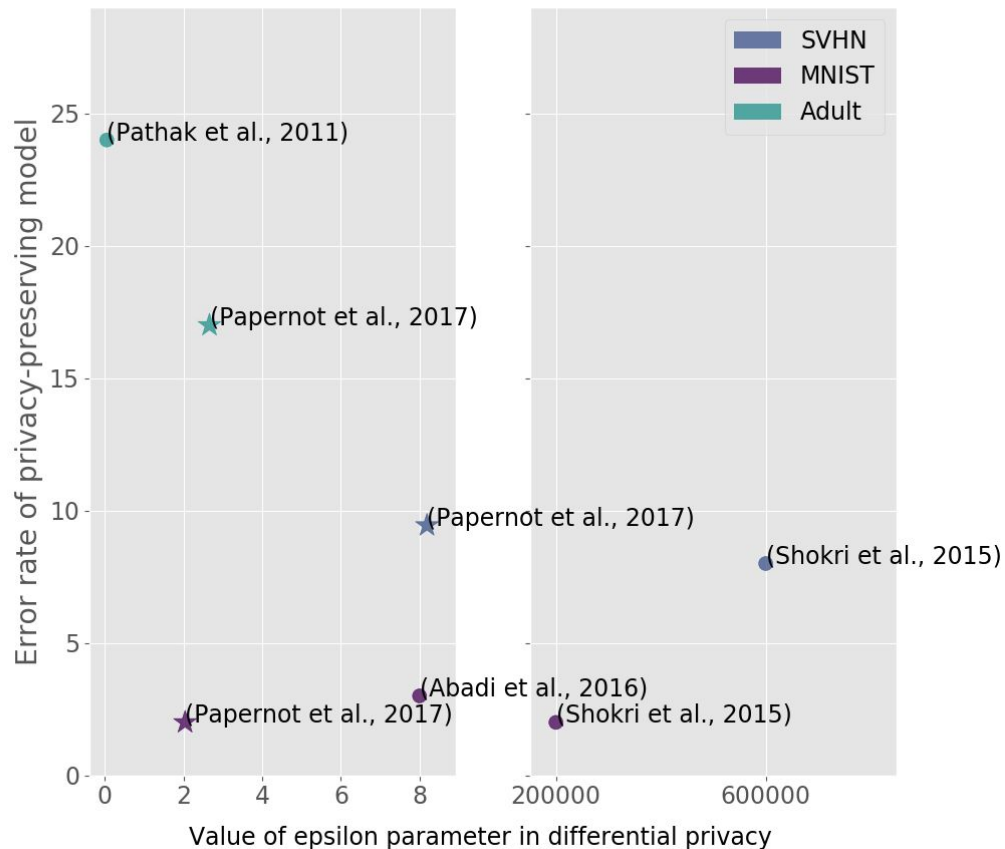
$$\Pr[M(d) \in S] \leq e^\epsilon \Pr[M(d') \in S] + \delta$$

Application of the **Moments Accountant** technique (Abadi et al, 2016)

Strong **quorum** \Rightarrow Small privacy cost

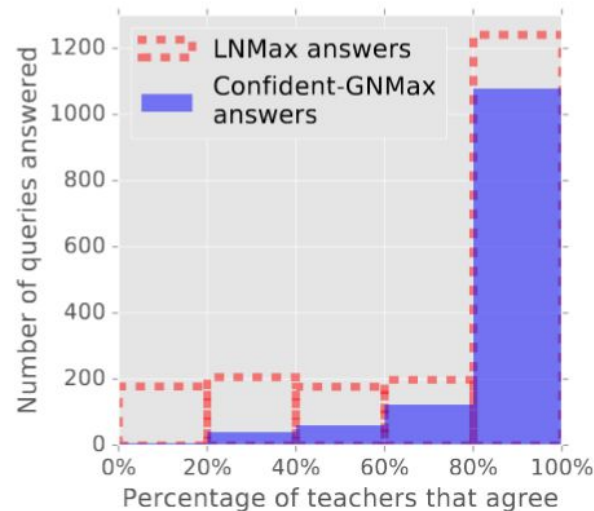
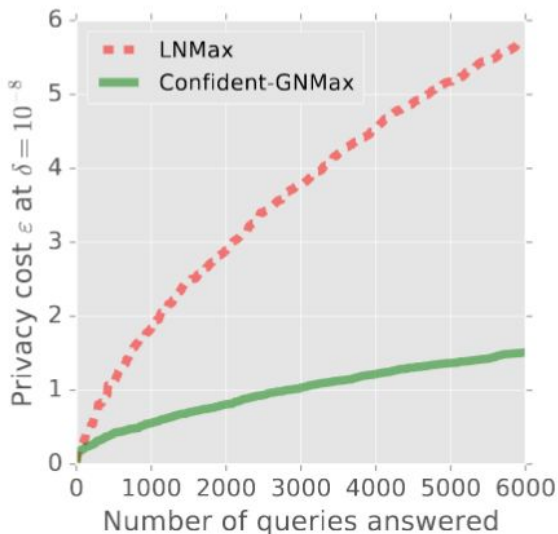
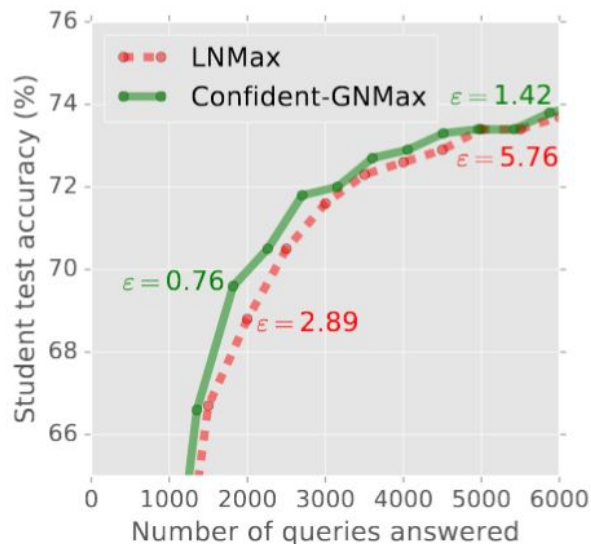
Bound is **data-dependent**: computed using the empirical quorum

Trade-off between student accuracy and privacy

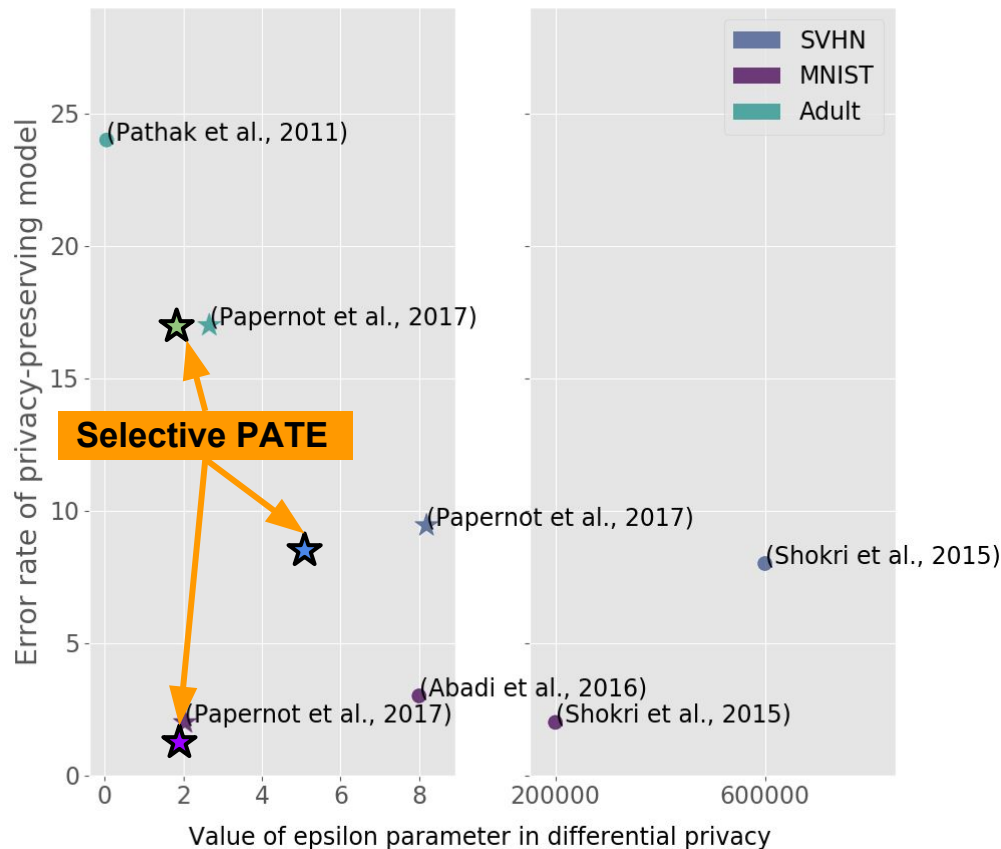


Synergy between utility and privacy

1. Check privately for consensus
2. Run noisy argmax only when consensus is sufficient



Trade-off between student accuracy and privacy



Machine learning and Goodhart's law

Economist Charles Goodhart posited in 1975 that ...

“When a measure becomes a target, it ceases to be a good measure”

As ML models make more and more decisions, we will have to satisfy them, and they will become targets.



 www.cleverhans.io

 [@NicolasPapernot](https://twitter.com/NicolasPapernot)

 www.papernot.fr



Thank you for listening!