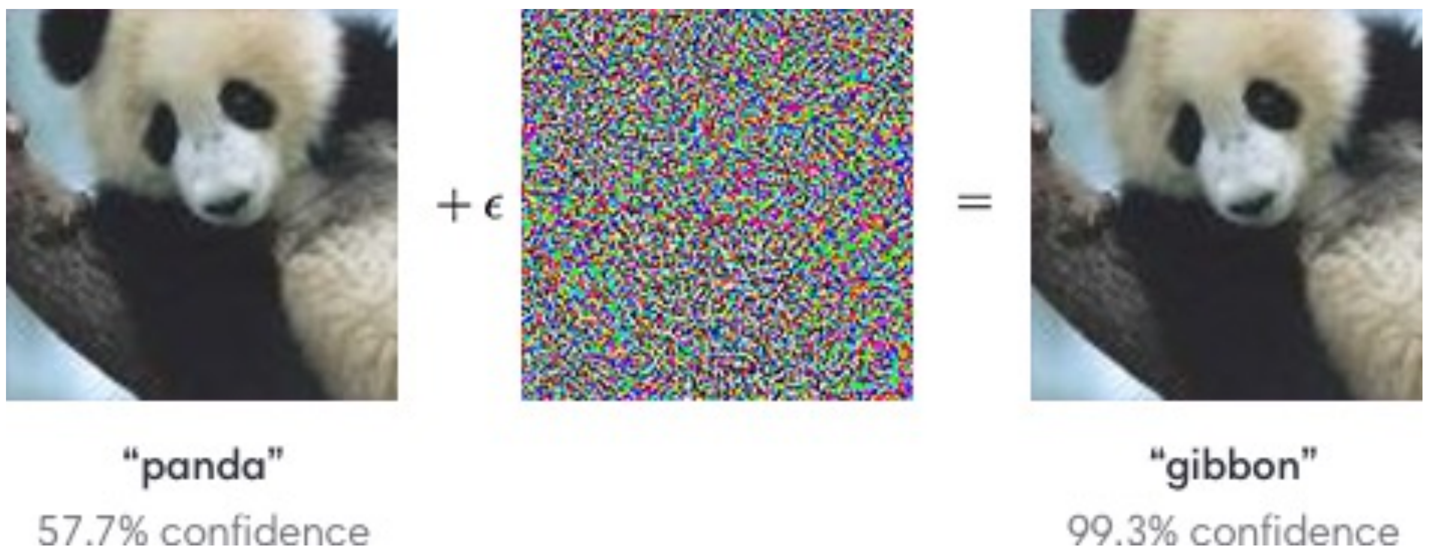
When *not* to use adversarial examples

Florian Tramèr

Google Brain & ETHZ

AAAI 2022 workshop on
Adversarial Machine Learning and Beyond

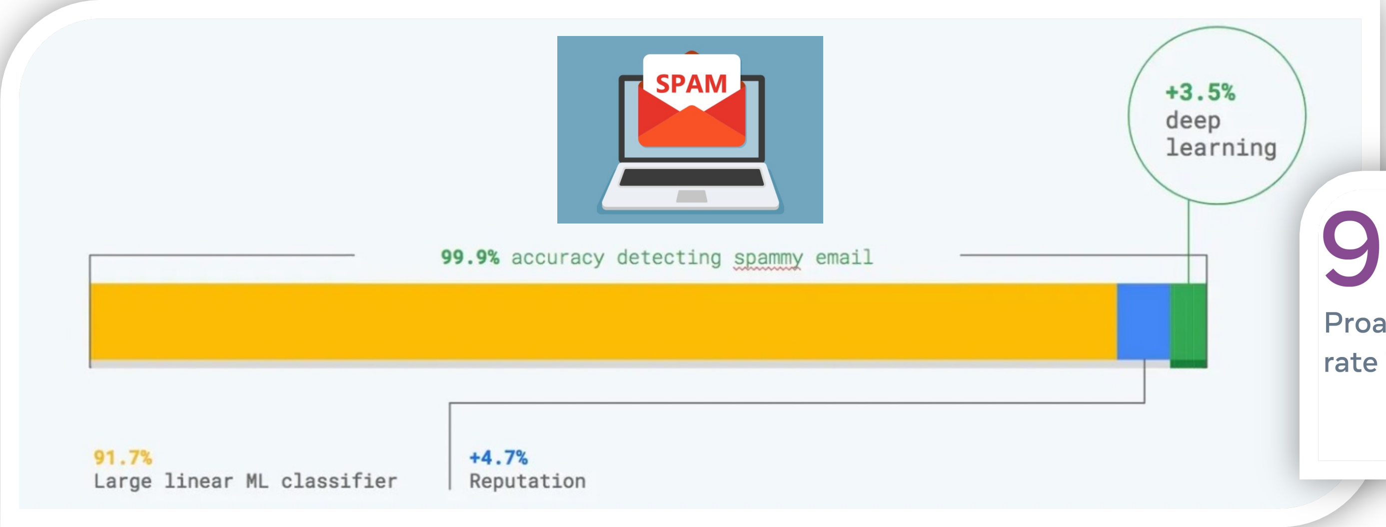
The current state of adversarial examples

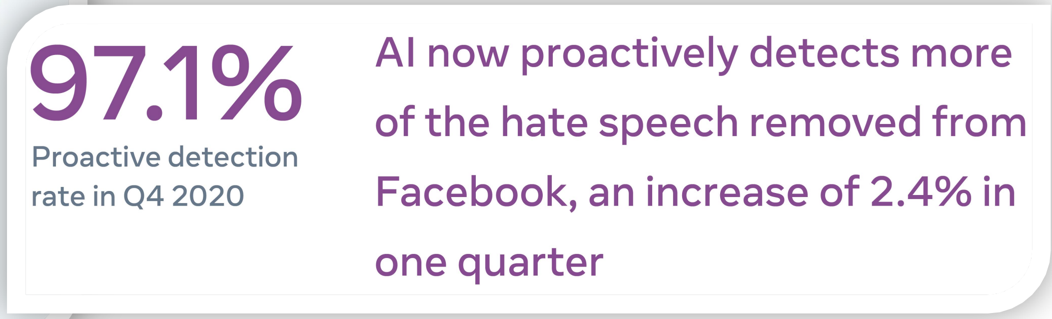


1. ***All* models are vulnerable** to some attack
2. The adversary can **adapt to *any* defense**

This sounds bad if you want to *defend* a model

1. ***All* models are vulnerable** to some attack
2. The adversary can **adapt to *any* defense**





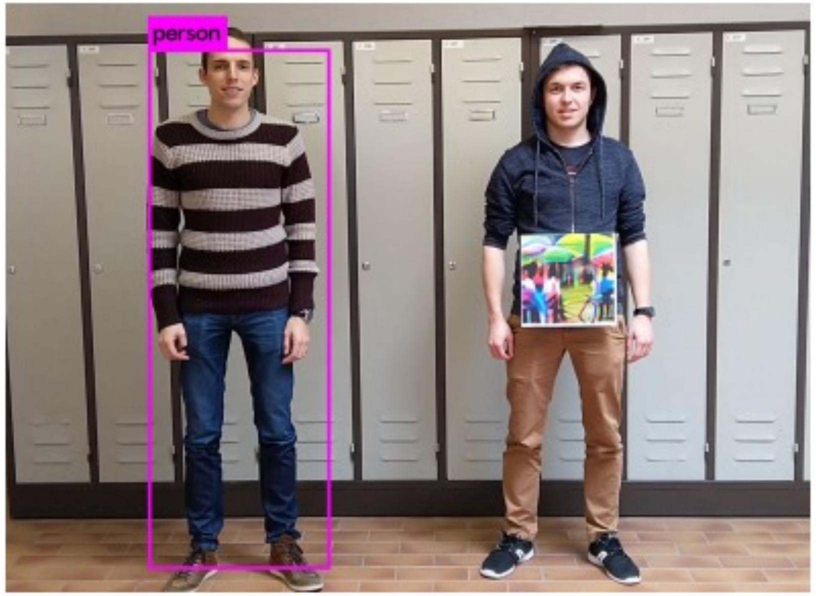
https://elie.net/blog/ai/harnessing-ai-to-combat-fraud-and-abuse-ai-is-the-key-to-robust-defenses/
https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/

This sounds good if you want to *attack* a model

1. ***All* models are vulnerable** to some attack
2. The adversary can **adapt to *any* defense**

Maybe we can use adversarial examples *for good*!

Positive applications of adversarial ML, i.e., adversarial for good.



Thys et al. 2019

Claim: adversarial examples ***don't*** work for

1) *protecting* against invasive models

2) *protecting* against privacy attacks

3) *attacking* anything "real" (for now)

Claim: adversarial examples ***don't*** work for

1) *protecting* against invasive models

2) *protecting* against privacy attacks

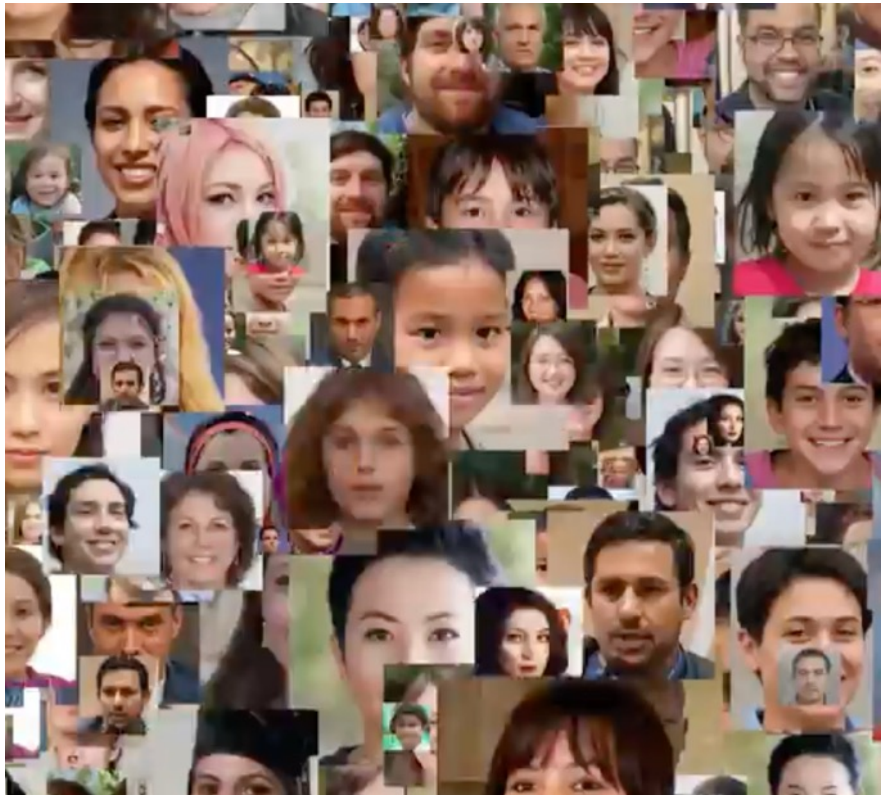
3) *attacking* anything "real" (for now)

Big brother is watching you

Clearview.ai

The Secretive Company That Might End Privacy as We Know It

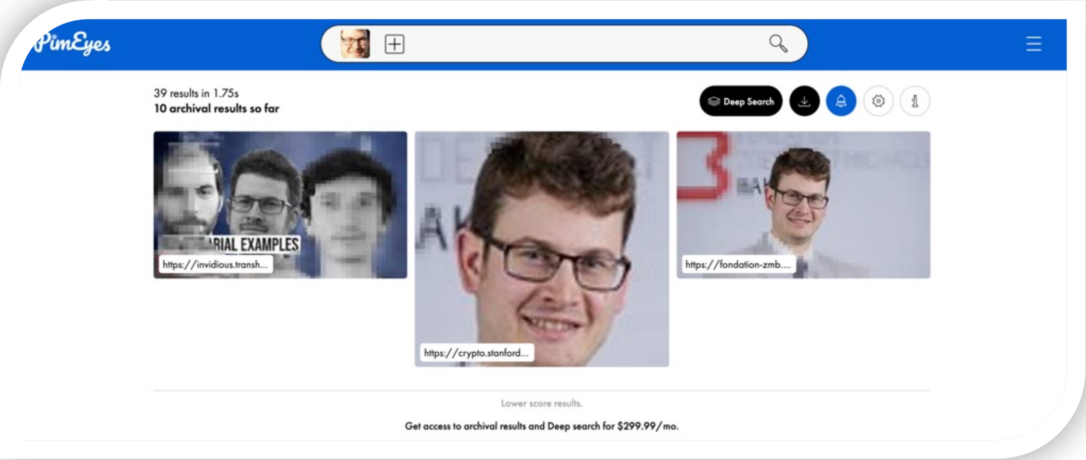
EXCLUSIVE Facial recognition company Clearview AI seeks first big deals, discloses research chief



~~Big brother~~ *Everyone* is watching you

Technology

This facial recognition website can turn anyone into a cop — or a stalker



Can adversarial examples save us from this dystopia?

The New York Times

This Tool Could Protect Your Photos From Facial Recognition

sandlab.cs.uchicago.edu/fawkes

BSD-3-Clause License

4.5k stars 439 forks

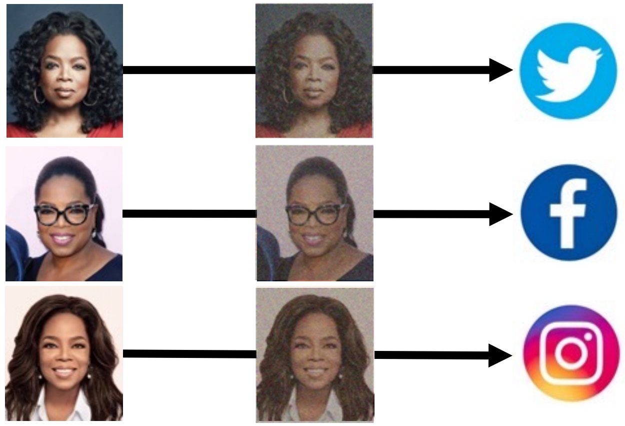
NEWS

- 4-23: v1.0 release for Windows/MacOS apps and Win/Mac/Linux binaries!
- 4-22: Fawkes hits 500,000 downloads!

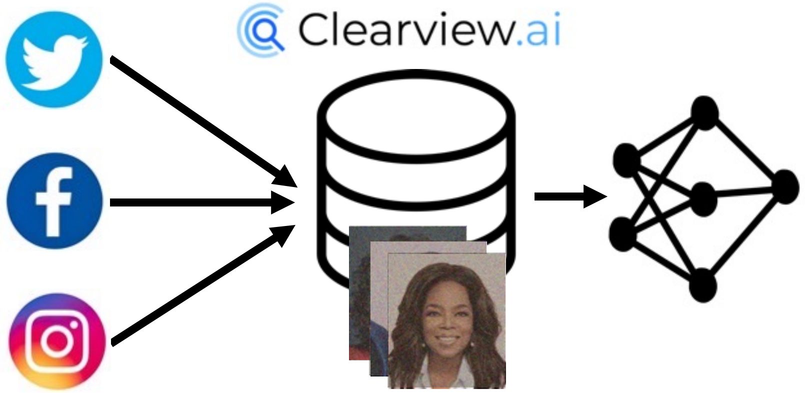
“Fawkes: Protecting Privacy against Unauthorized Deep Learning Models”, Shan et al., USENIX 2020
“LowKey: Leveraging Adversarial Attacks to Protect Social Media Users from Facial Recognition”, Cherepanova et al., ICLR 2021

Poisoning facial recognition with adversarial examples

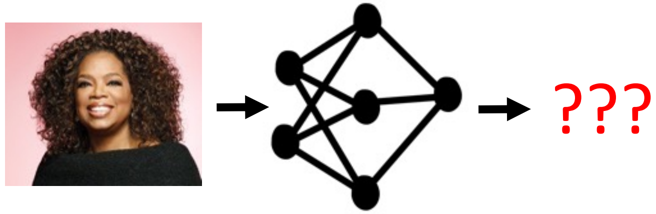
Users perturb the pictures they post online



Online pictures are scraped to build a model



Unperturbed test pictures aren't recognized



Unperturbed picture taken by the police, or a stalker, etc.

Misconception #1:

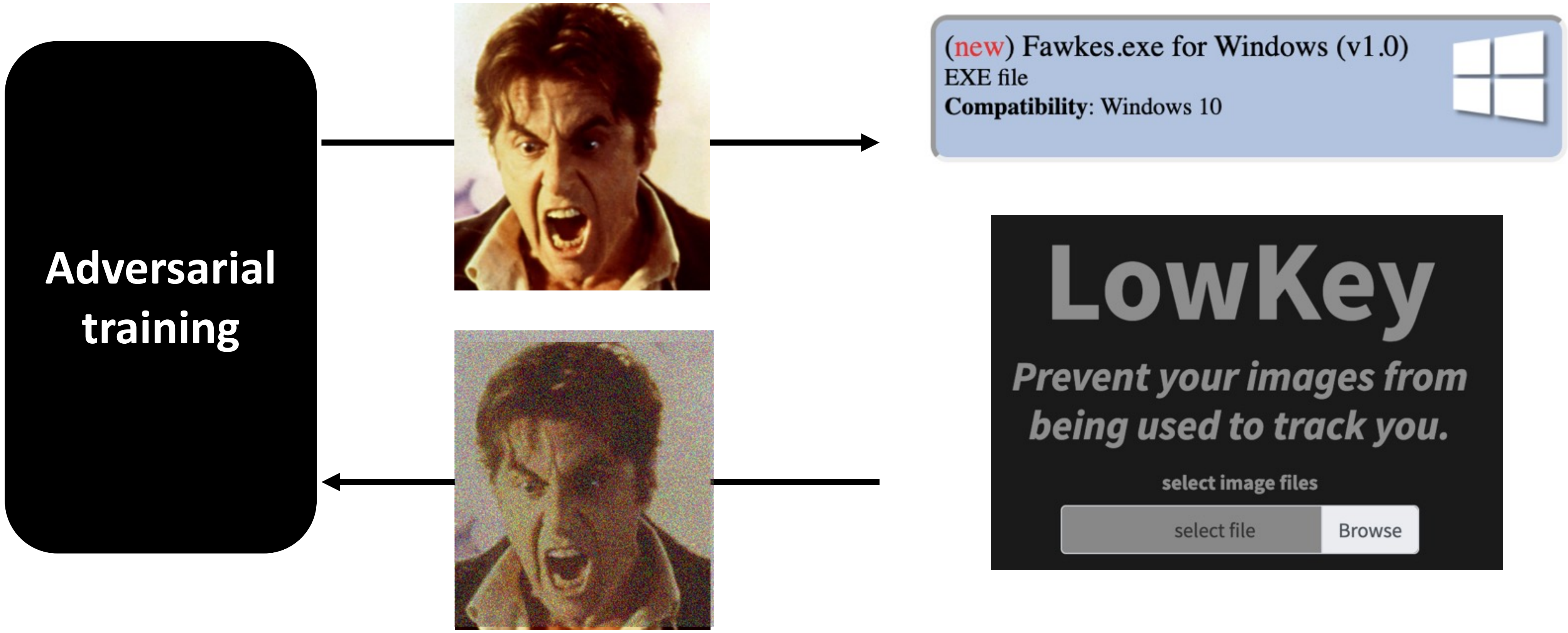
$$\underbrace{\forall \text{ models } \exists \text{ attack}}_{\text{this is empirically true (so far)}} \neq \exists \text{ attack } \forall \text{ models}$$

credit: Nicholas Carlini

Misconception #2:

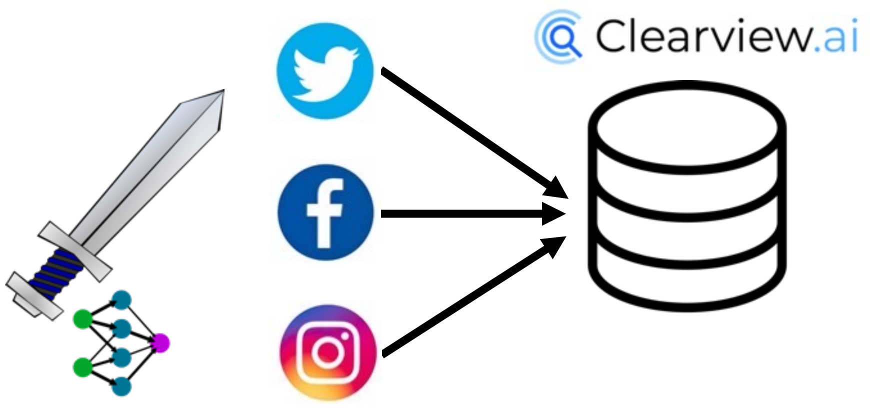
The attacker can adapt to any defense

The problem (1): adaptive defenses

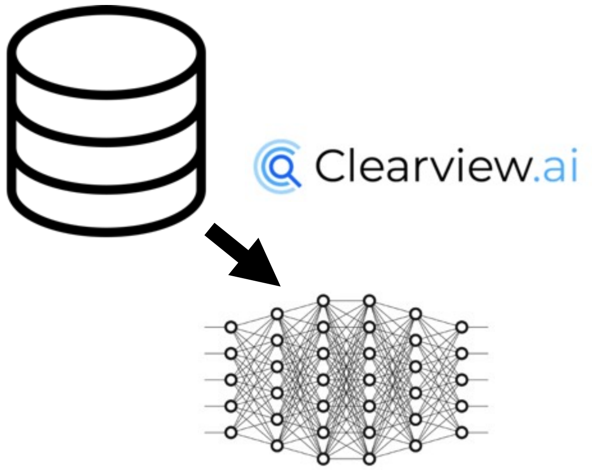


"Data Poisoning Won't Save You From Facial Recognition", ICLR 2022

The problem (2): retroactive defenses



wait one year

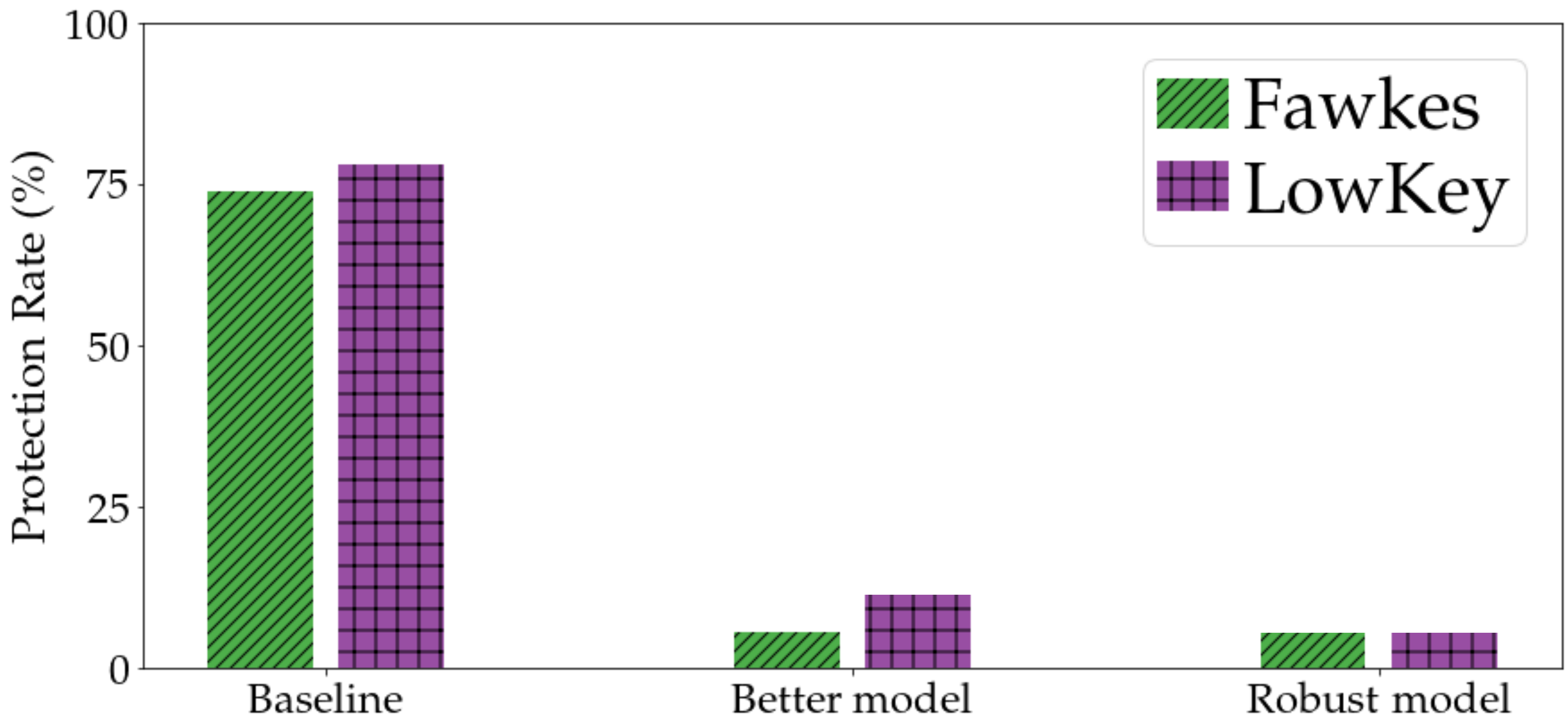


Facial recognition provider scrapes pictures produced with attacks that target today's models

Facial recognition provider trains new ***better*** model on poisoned data collected in the past

"Data Poisoning Won't Save You From Facial Recognition", ICLR 2022

Adversarial examples *won't* save us



"Data Poisoning Won't Save You From Facial Recognition", ICLR 2022

Are evasion attacks any better?

(+) Here, the attacker *can* adapt to the facial recognition system

(+) This works *against YOLO* !*

(-) What *guarantee* is there that this works against *any real system?*

(-) Are we giving people a *false sense of security?*

Claim: adversarial examples ***don't*** work for

1) *protecting* against invasive models

2) *protecting* against privacy attacks

3) *attacking* anything "real" (for now)

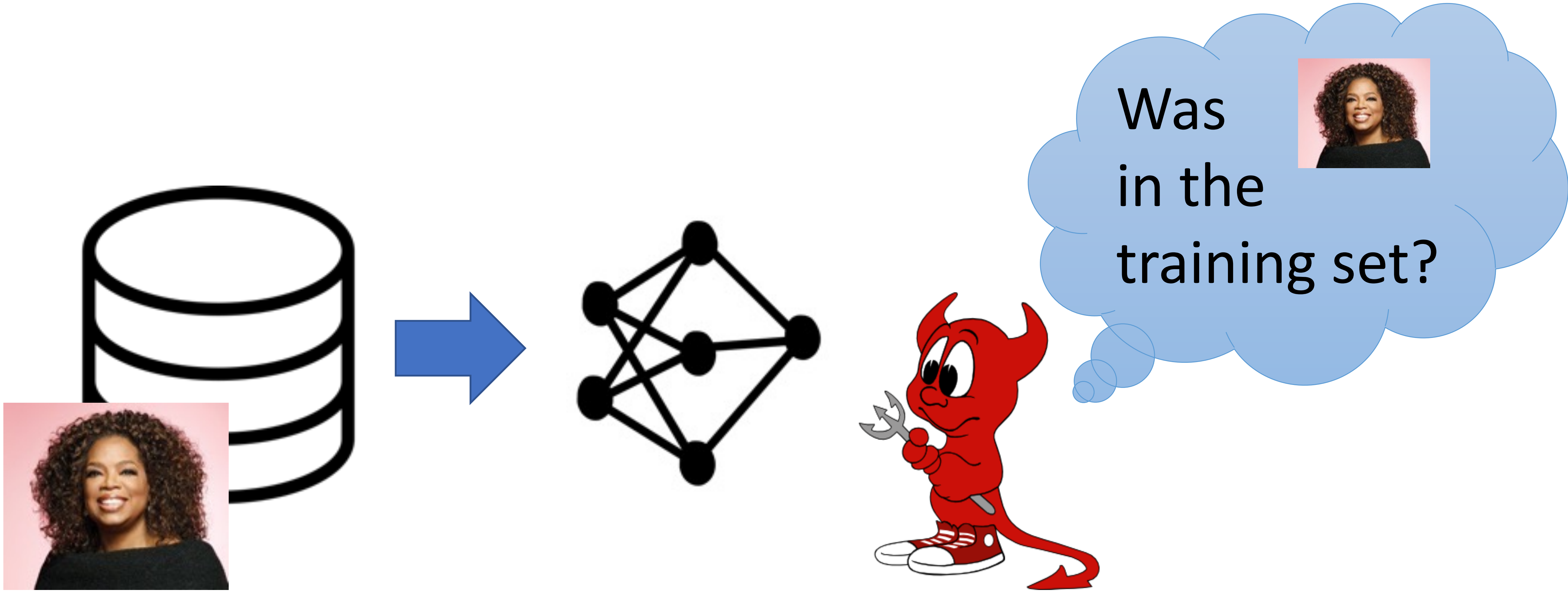
Claim: adversarial examples ***don't*** work for

1) *protecting* against invasive models

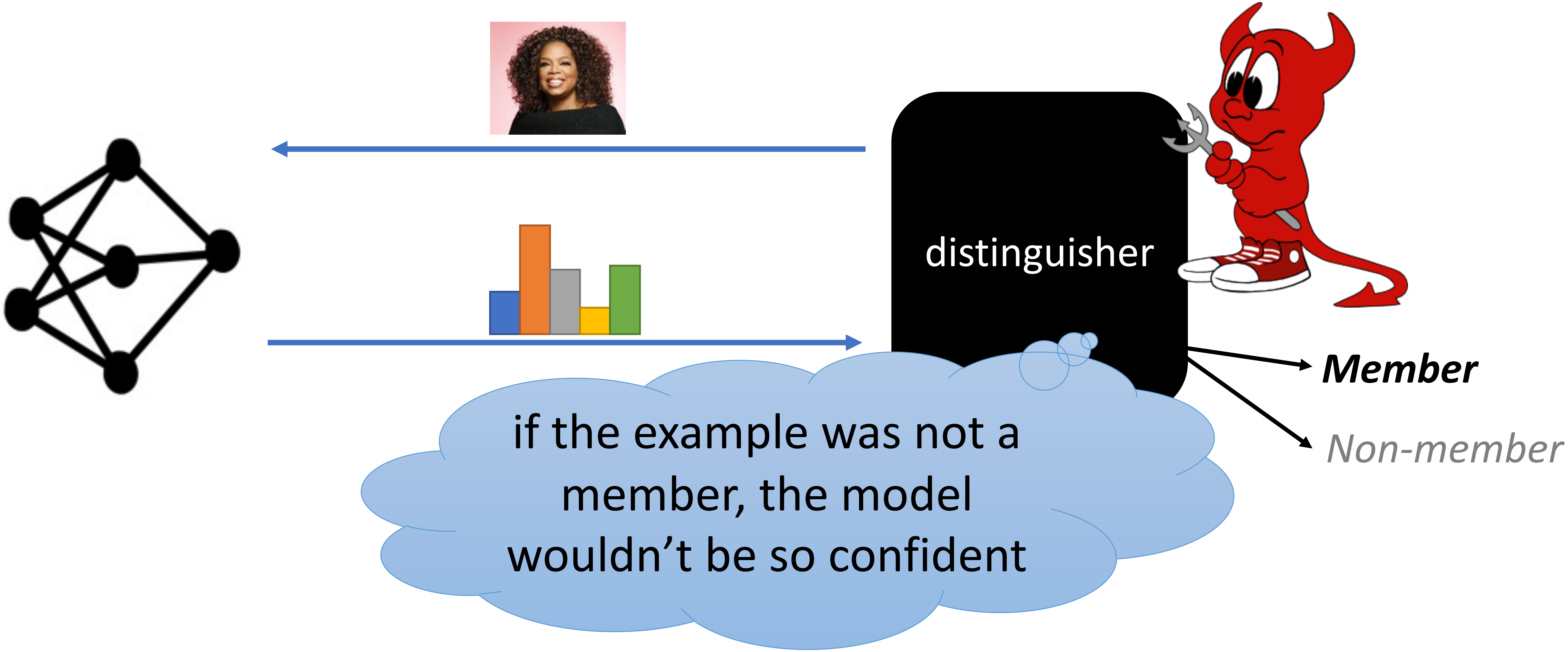
2) *protecting* against privacy attacks

3) *attacking* anything "real" (for now)

A simple privacy attack: membership inference

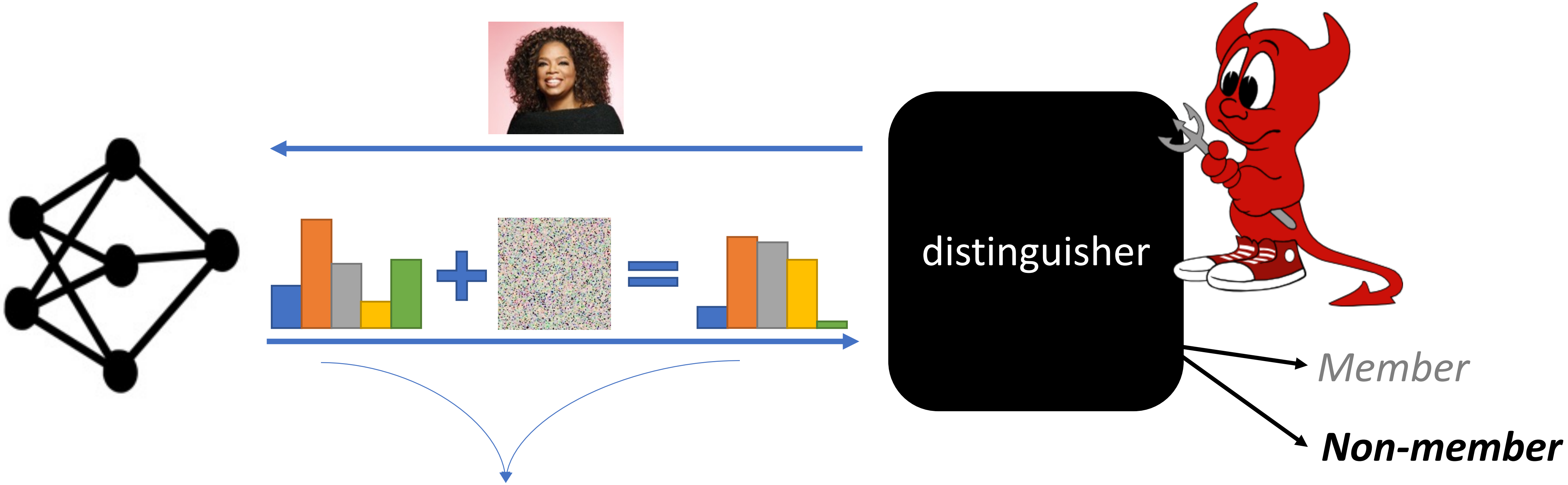


Membership leakage from model confidence



"Membership Inference Attacks against Machine Learning Models", Shokri et al., IEEE S&P 2017

Defense idea: adversarial examples for distinguisher



top class is preserved to ensure defense does not lower utility

“MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples”, Jia et al., CCS 2019

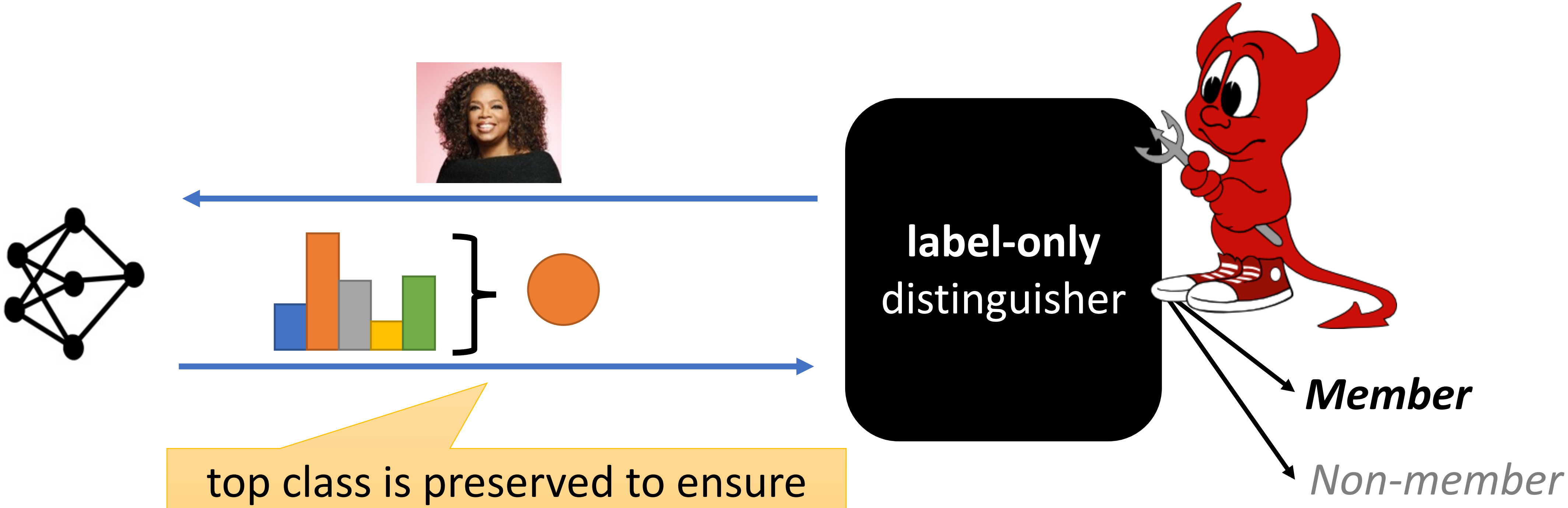
Misconception #1:

$\forall$ models $\exists$ attack $\neq$ $\exists$ attack $\forall$ models

Misconception #2:

The attacker can adapt to any defense

Adaptive “defense”: ignore the noise



top class is preserved to ensure defense does not lower utility

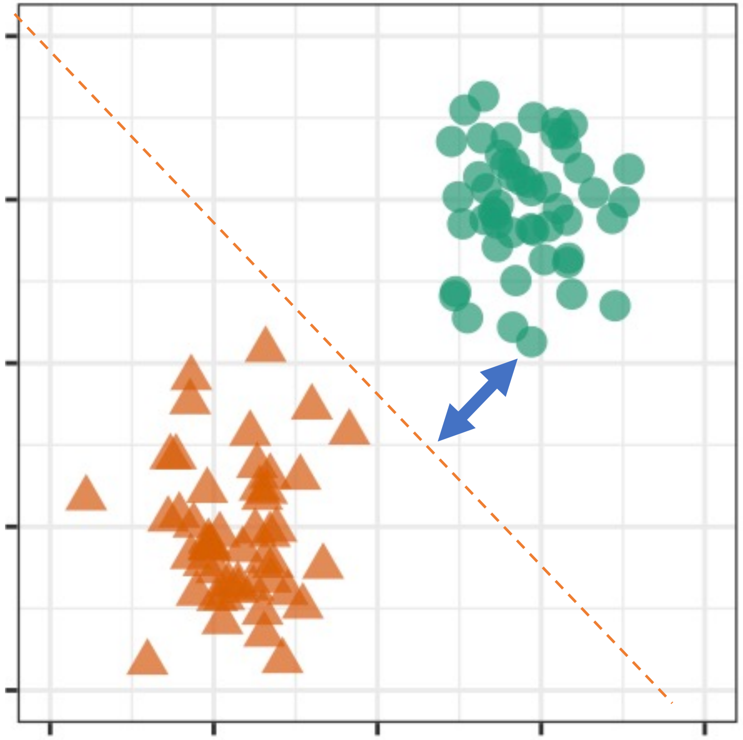
What can we compute with label-only access?

- ~~Model confidence~~
- ~~Gradient norm~~

What can we compute with label-only access?

- ~~Model confidence~~
- ~~Gradient norm~~
- Distance to decision boundary?

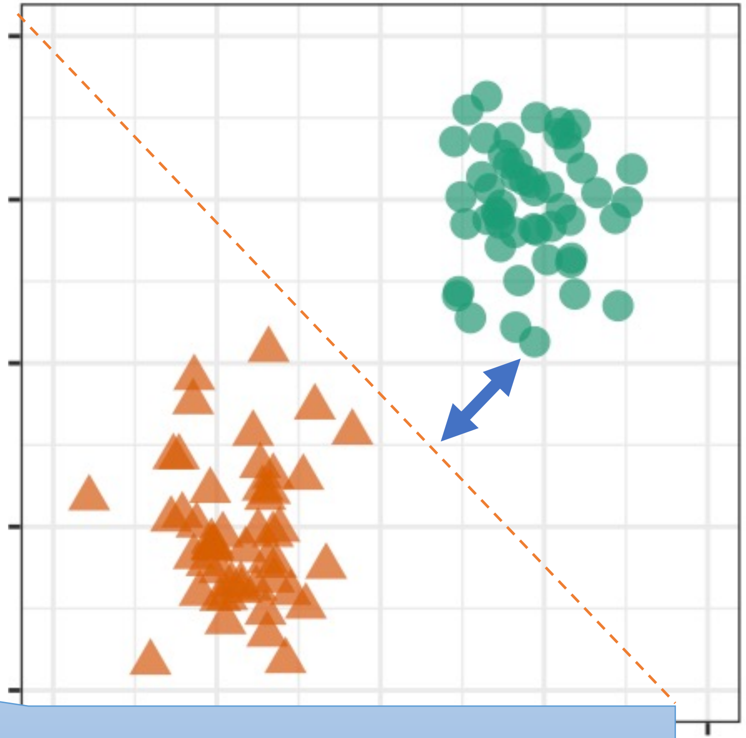
Same as finding a "minimal norm" adversarial example !



What can we compute with label-only access?

- ~~Model confidence~~
- ~~Gradient norm~~
- Distance to decision boundary?

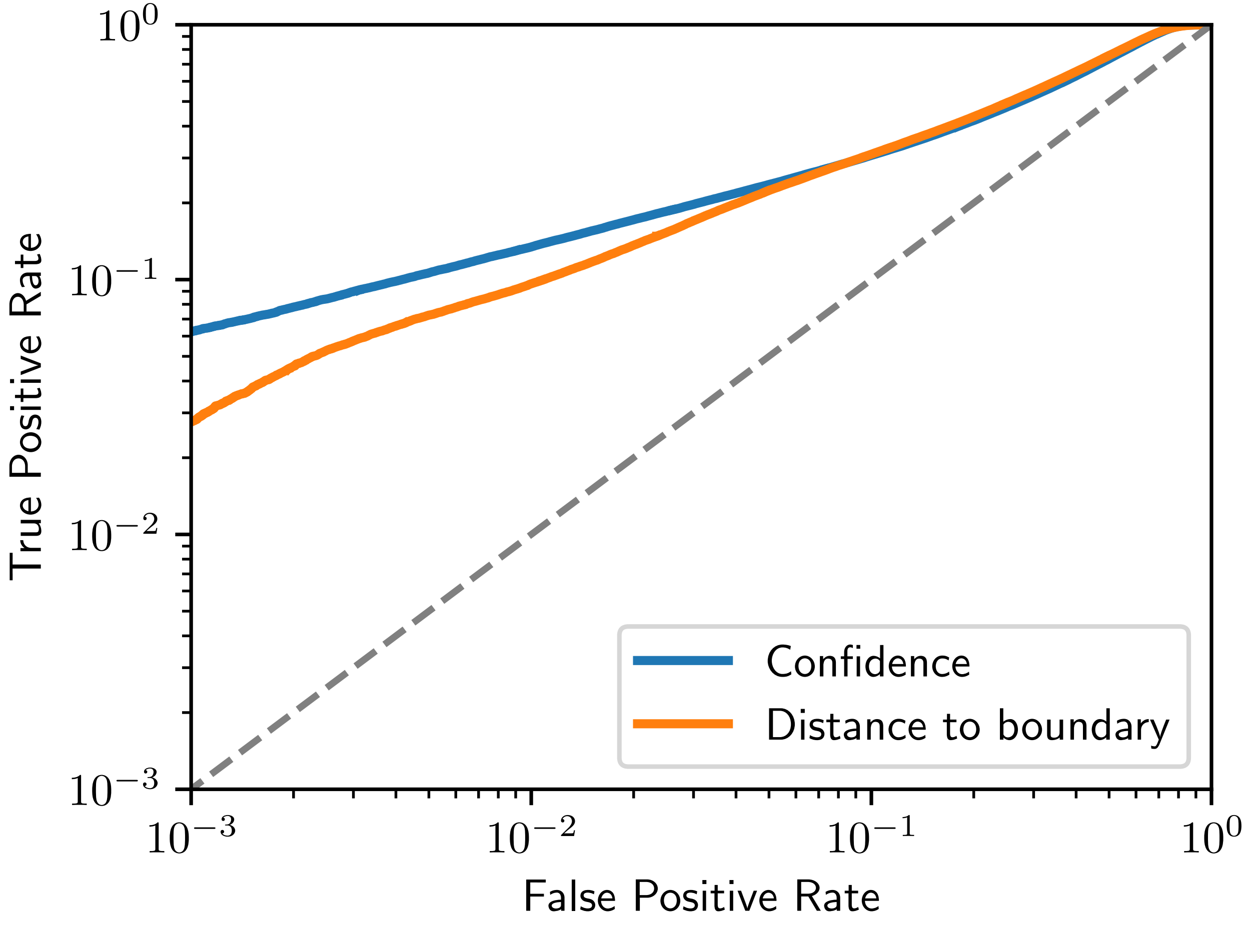
Same as finding a "minimal norm" adversarial example !



We can compute this with labels only! (decision-based attacks)

"Reliable Attacks Against Black-Box Machine Learning Models", Brendel et al., ICLR 2018

Adversarial confidences don't prevent inference



"Label-Only Membership Inference Attacks", ICML 2021
"Membership Inference Attacks From First Principles", preprint 2022

Claim: adversarial examples *don't* work for

1) *protecting* against invasive models

2) *protecting* against privacy attacks

3) *attacking* anything "real" (for now)

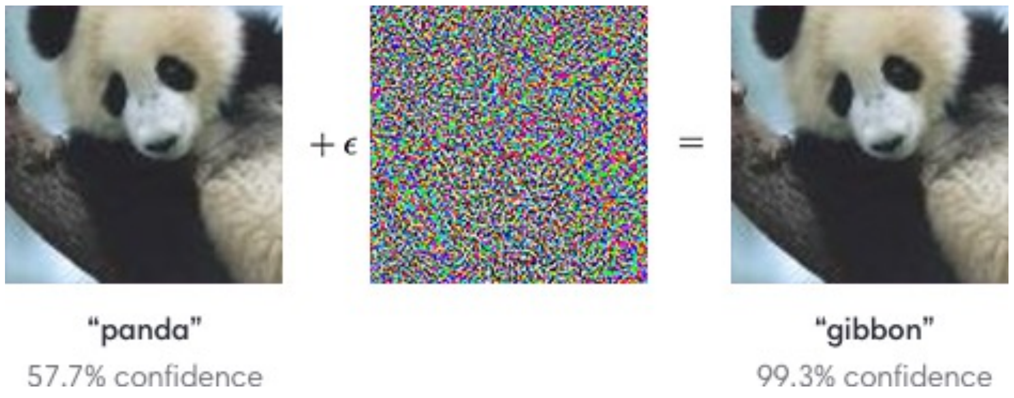
Claim: adversarial examples ***don't*** work for

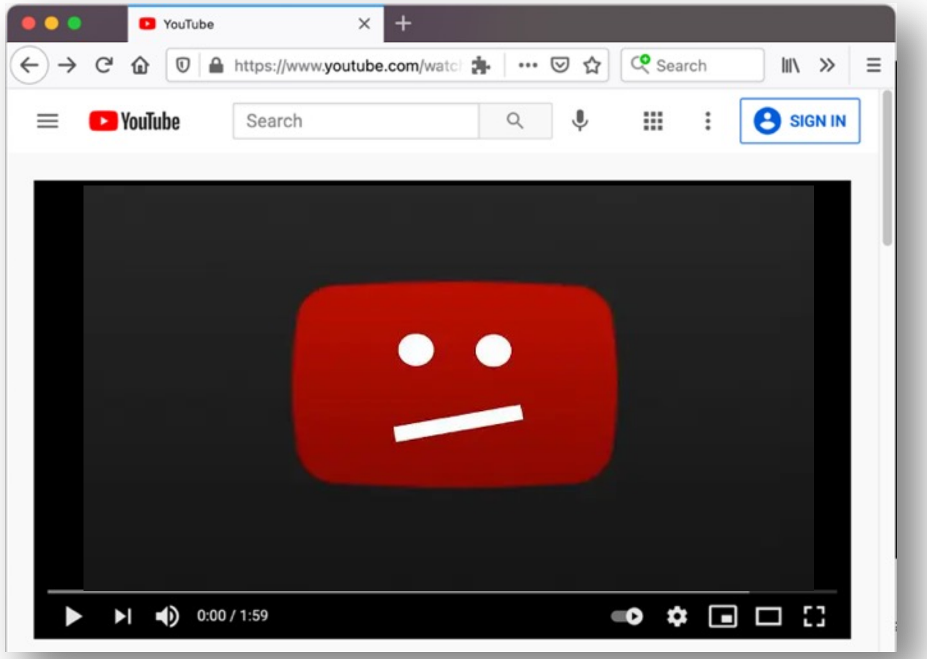
1) *protecting* against invasive models

2) *protecting* against privacy attacks

3) *attacking* anything "real" (for now)

Evading research models vs. real systems





How do we go from this... ...to this?

Evading research models vs. real systems

Research: "imperceptible" perturbations

~95% white-box attacks/defenses

~5% black-box with query access

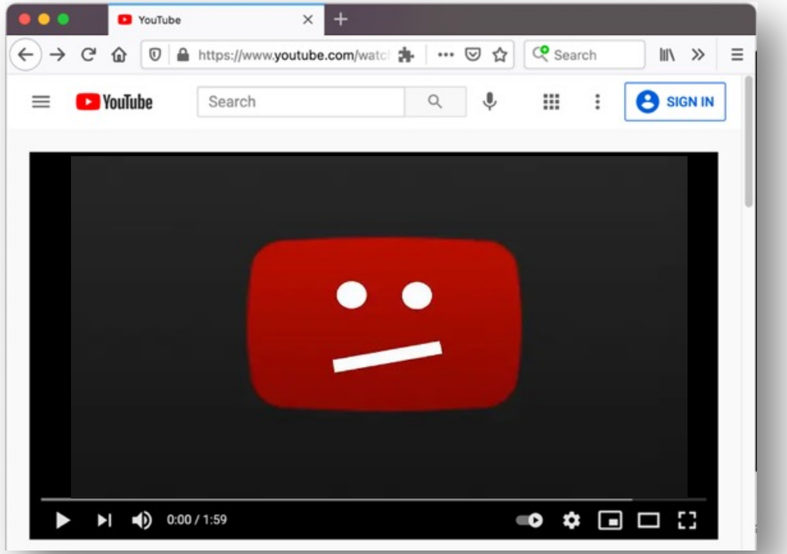
<1% black-box without query-access

Real systems: >99% black-box without query-access

attacks need not be imperceptible

Real systems are black-box

Challenge: attack something like this



Not just an engineering exercise!

- *you don't get direct query access...*
- *you get banned after a few bad queries...*
- *you likely can't build a good surrogate model...*

Claim: adversarial examples ***don't*** work for

1) *protecting* against invasive models

2) *protecting* against privacy attacks

3) *attacking* anything "real" (for now)

Claim: adversarial examples ***don't*** work for

1) *protecting* against invasive models

2) *protecting* against privacy attacks

3) *attacking* anything "real" (for now)

Conclusion

- **Threat models** matter! *(who gets to go second?)*
- Be careful what ***promises*** you make to users
- Can we use adversarial examples for ***something "real"?***