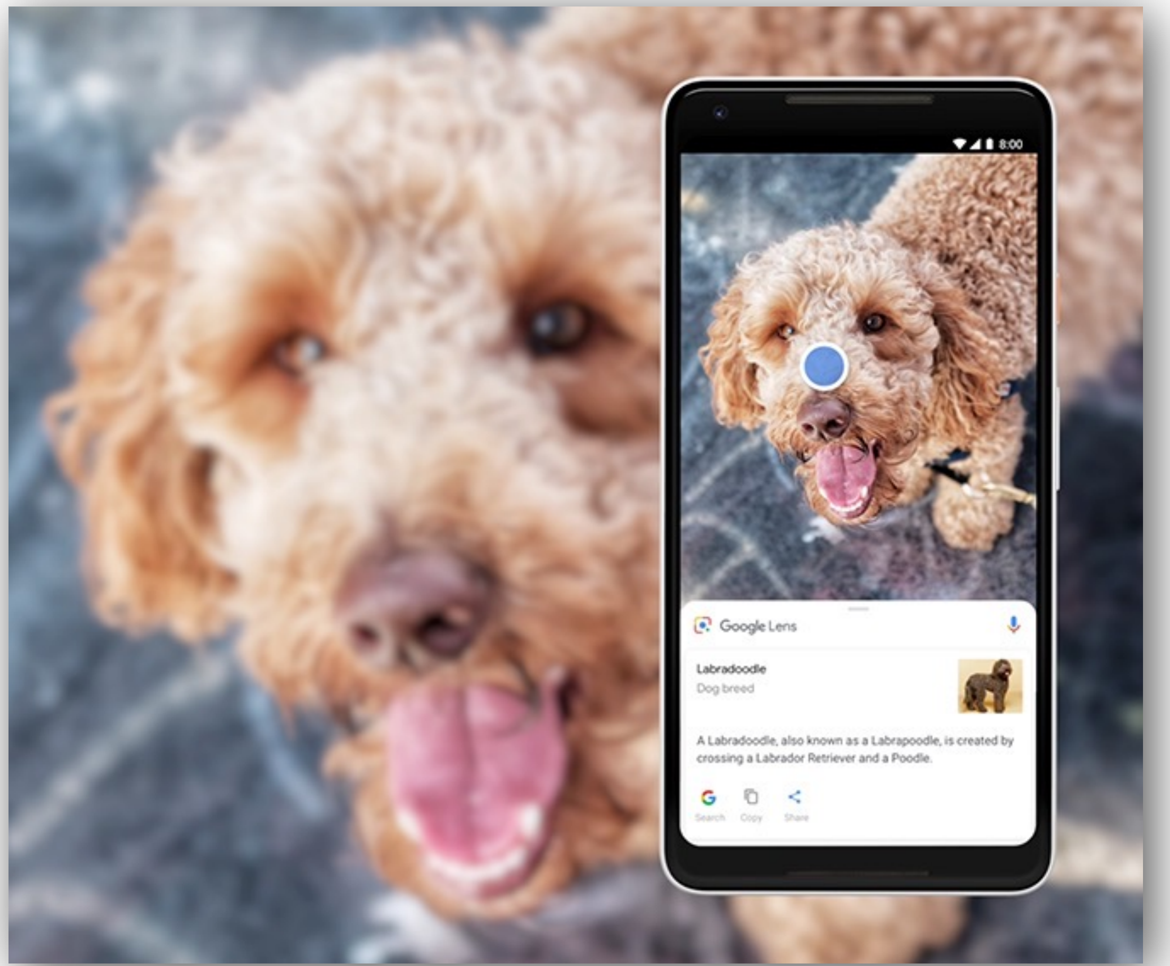
Measuring and Enhancing the Security of Machine Learning

Florian Tramèr

Stanford University

Committee members: Mykel Kochenderfer (chair), Dan Boneh (advisor), Moses Charikar, Percy Liang, Gregory Valiant

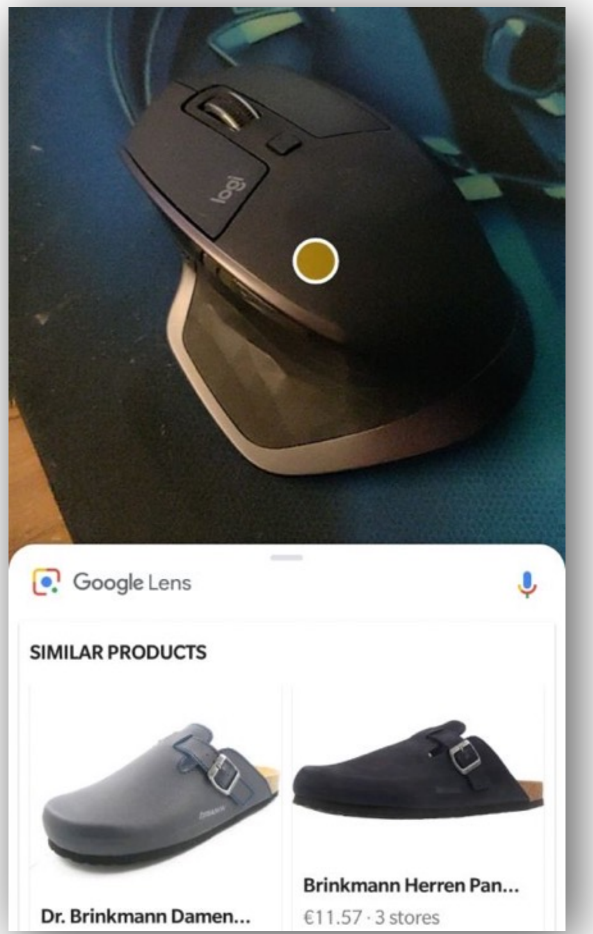
Machine learning works.

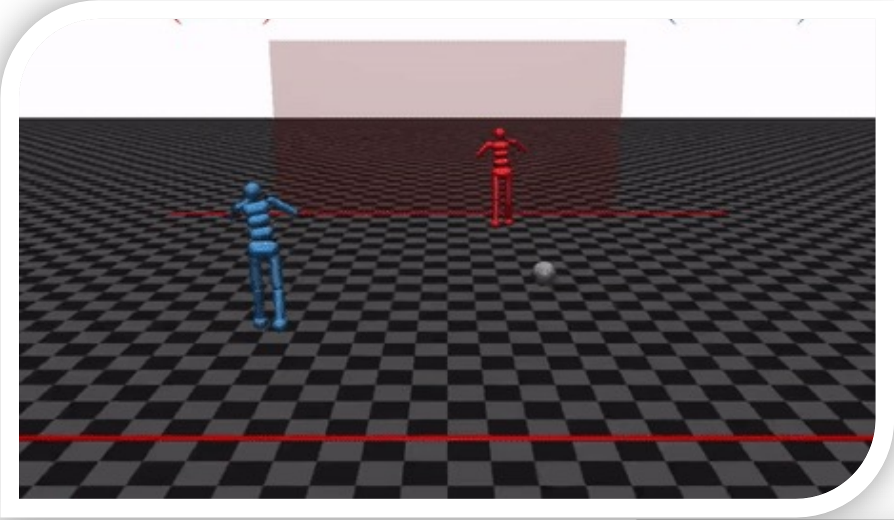


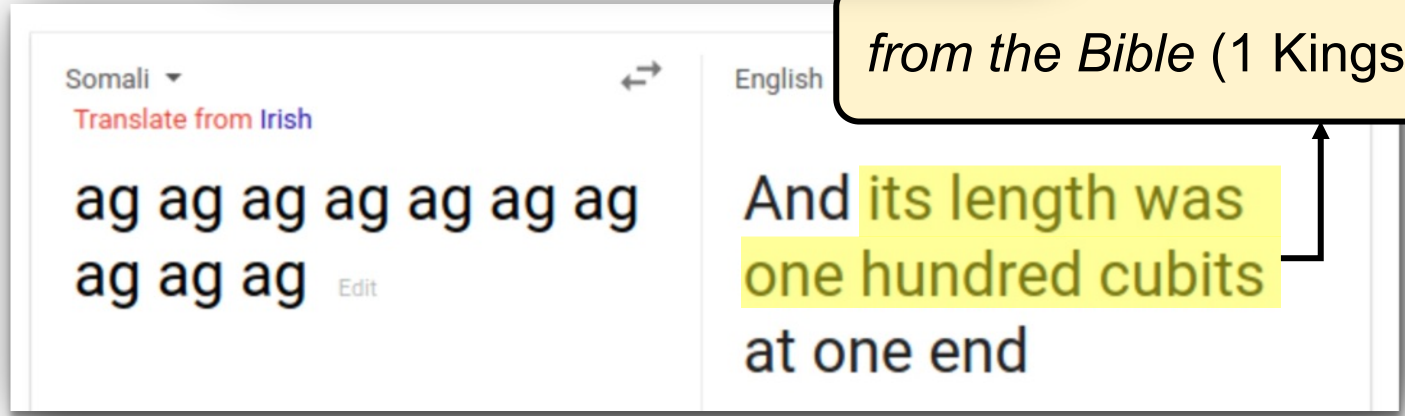
Google Translate

Machine learning works **most of the time!**

many applications tolerate occasional failures







from the Bible (1 Kings 7:2)

Machine learning can also fail disastrously.

Critical mistakes...

theguardian

Uber crash shows 'catastrophic failure' of self-driving technology, experts say

Machine learning can also fail disastrously.

Critical mistakes...

theguardian

Uber crash shows 'catastrophic failure' of self-driving technology, experts say

Direct attacks...

The New York Times

Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.

Machine learning can also fail disastrously.

Critical mistakes...

theguardian

Uber crash shows 'catastrophic failure' of self-driving technology, experts say

Direct attacks...

The New York Times

Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.

Private data leaks...

Does GPT-2 Know Your Phone Number?

Eric Wallace, Florian Tramèr, Matthew Jagielski, and Ariel Herbert-Voss

Challenge: understand and improve the worst-case behavior of machine learning (ML)

Approach: study ML from an adversarial perspective

- ➢ to improve *robustness* and *privacy* of ML in adversarial settings
- ➢ to build ML that is *better*

This thesis

Measuring and Enhancing ML security

I. Modeling the threat of adversarial examples

- Analysis: *fundamental limits* of existing defenses
- Application: circumventing *online content blockers*
  (led to design changes in Adblock Plus)

II. Enhancing data privacy for ML users

- At training time using *differential privacy*
- At test time using *hardware enclaves* and *cryptography*

This thesis

Measuring and Enhancing ML security

I. Modeling the threat of adversarial examples

- Analysis: *fundamental limits* of existing defenses
- Application: circumventing *online content blockers*
  (led to design changes in Adblock Plus)

II. Enhancing data privacy for ML users

- At training time using *differential privacy*
- At test time using *hardware enclaves* and *cryptography*

This thesis

Measuring and Enhancing ML security

I. Modeling the threat of adversarial examples

- Analysis: *fundamental limits* of existing defenses
- Application: circumventing *online content blockers*
  (led to design changes in Adblock Plus)

II. Enhancing data privacy for ML users

- At training time using *differential privacy*
- At test time using *hardware enclaves* and *cryptography*

This thesis

Measuring and Enhancing ML security

I. Modeling the threat of adversarial examples

- Analysis: *fundamental limits* of existing defenses
- Application: circumventing *online content blockers*
  (led to design changes in Adblock Plus)

this talk!

II. Enhancing data privacy for ML users

- At training time using *differential privacy*
- At test time using *hardware enclaves* and *cryptography*

Talk outline.

- Adversarial examples for online content blockers
  - ➢ What's the threat model?
  - ➢ Limitations of current defenses
  - ➢ Industry impact
- Enhancing ML privacy
- Future work

Talk outline.

- Adversarial examples for online content blockers
  - ➢ What's the threat model?
  - ➢ Limitations of current defenses
  - ➢ Industry impact
- Enhancing ML privacy
- Future work

What is Machine Learning (ML)?

collect some
“training” data

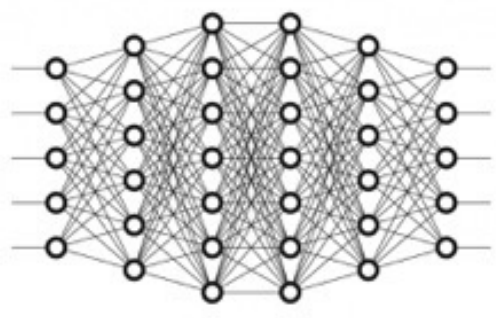
“cat”

“dog”

“pig”

build a function (model) that learns how
to make predictions on ***new*** data

f () = “cat”, 90%



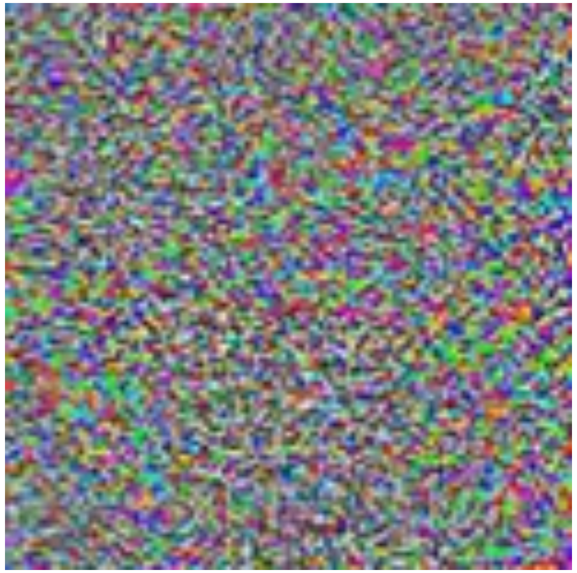
neural network
(sequence of math transforms
applied to the input to assign a
”confidence” to each prediction)

Adversarial examples: a curious *bug* in ML

[Szegedy et al. ‘13], [Biggio et al. ‘13], [Goodfellow et al. ‘14], ...

90% Tabby Cat

+

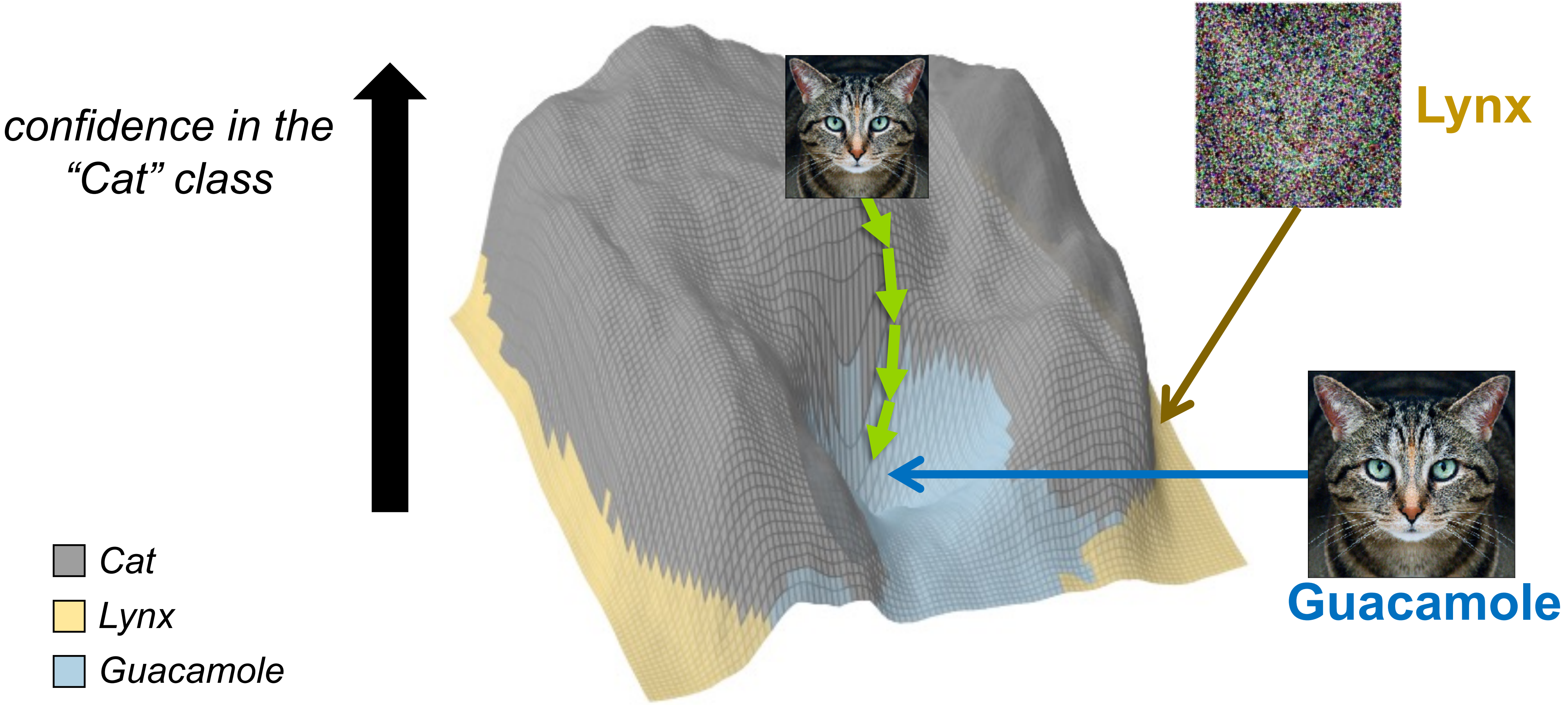


Adversarial noise

=

100% Guacamole

Finding adversarial examples.



Why do adversarial examples matter?

For understanding ML

- ➢ what is the model learning?
- ➢ why do brittle models *generalize?*

For security:

- ➢ **will my ML system fail unexpectedly?**
- ➢ **can my ML system be attacked?**

Adversarial examples as a computer security problem.

T, Dupré, Rusak, Pellegrino, Boneh (ACM CCS 2019)

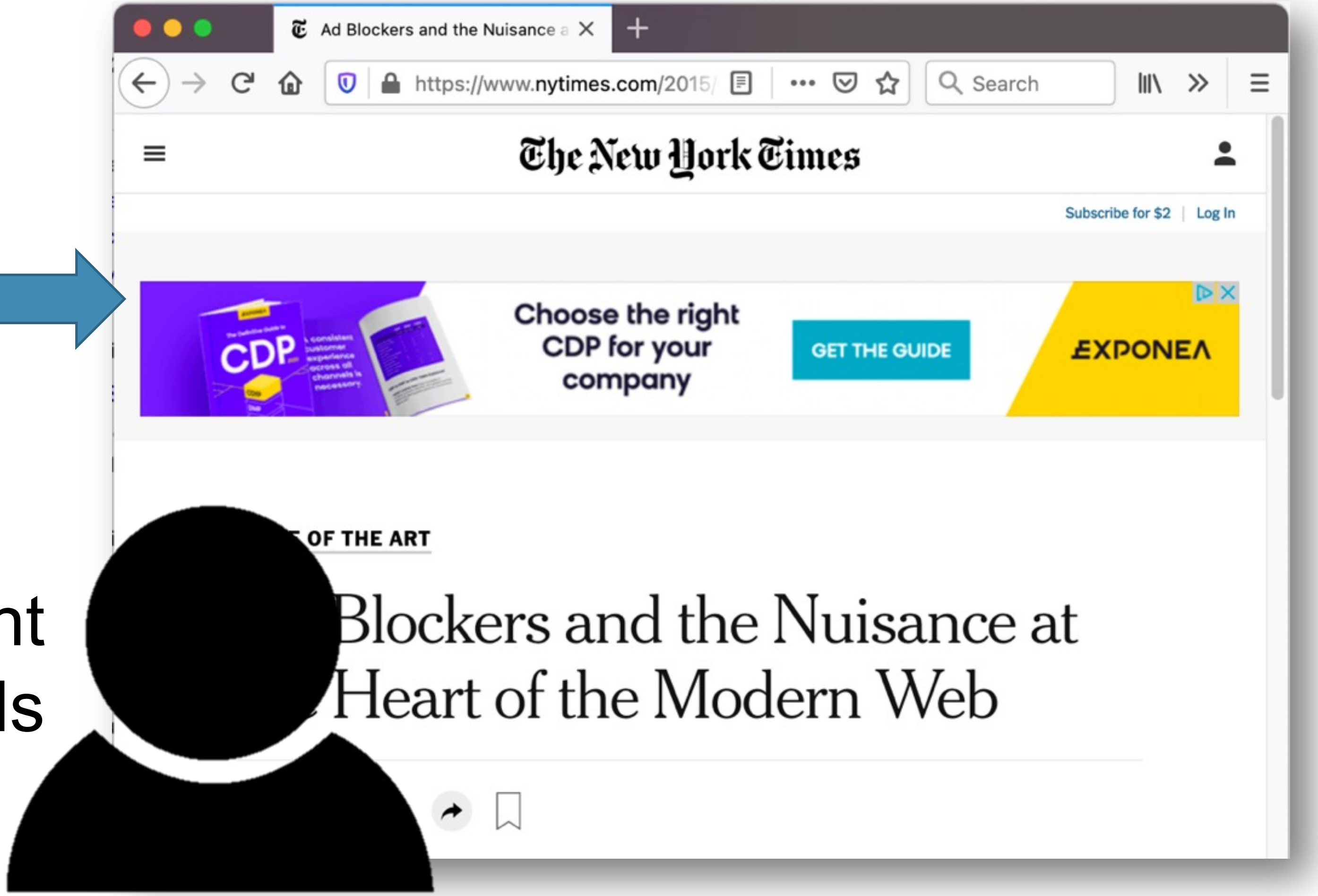
- adversarial examples are the **perfect tool** to attack *online content blockers*
- *using ML for ad-blocking can break Web security*
- *this work led to design changes in Adblock Plus*

ABP Adblock **Plus**

100M active users

Adversarial examples are a security threat for online ad-blocking.

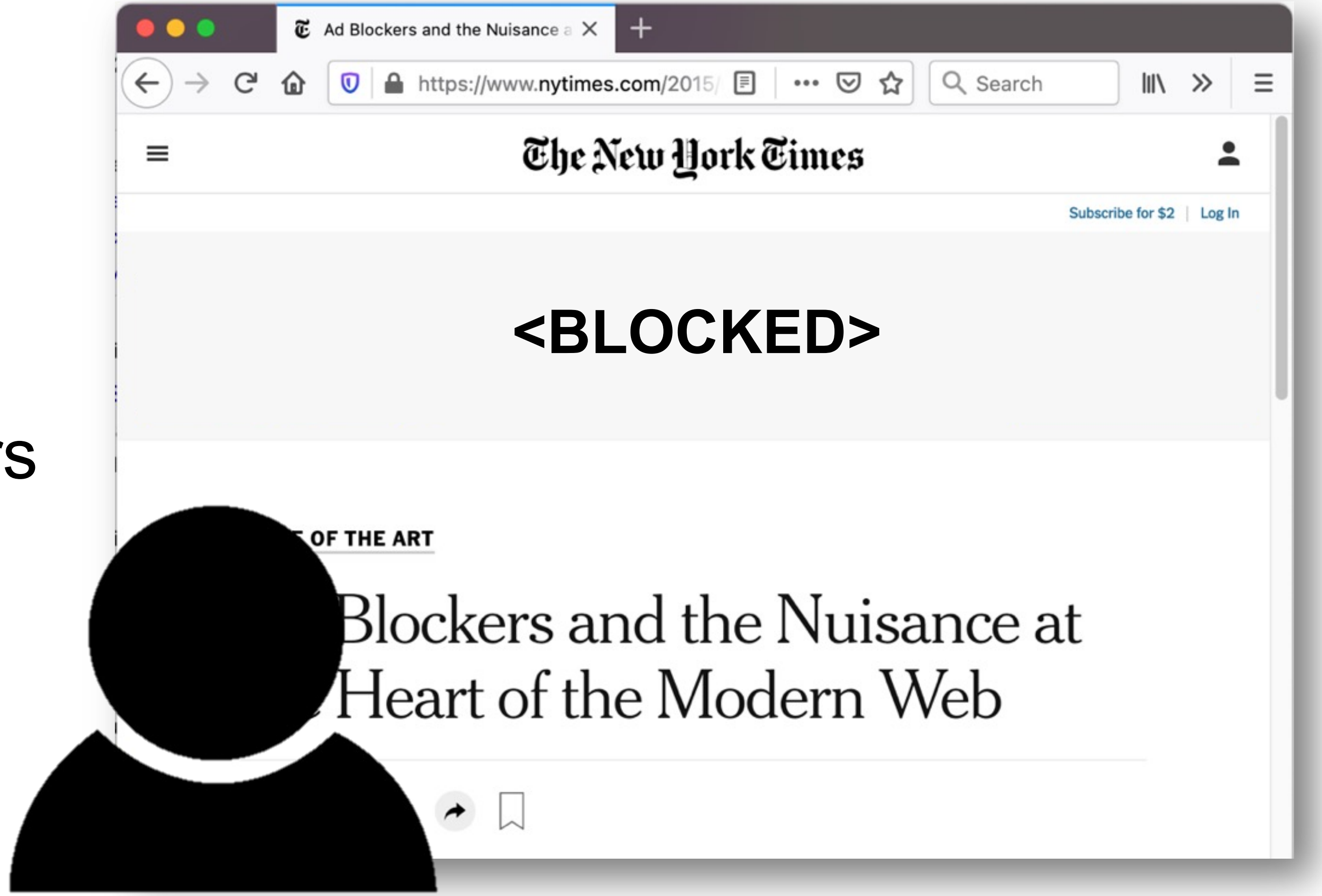
publishers & advertisers want to show ads to users...



...users don't want to see ads

Adversarial examples are a security threat for online ad-blocking.

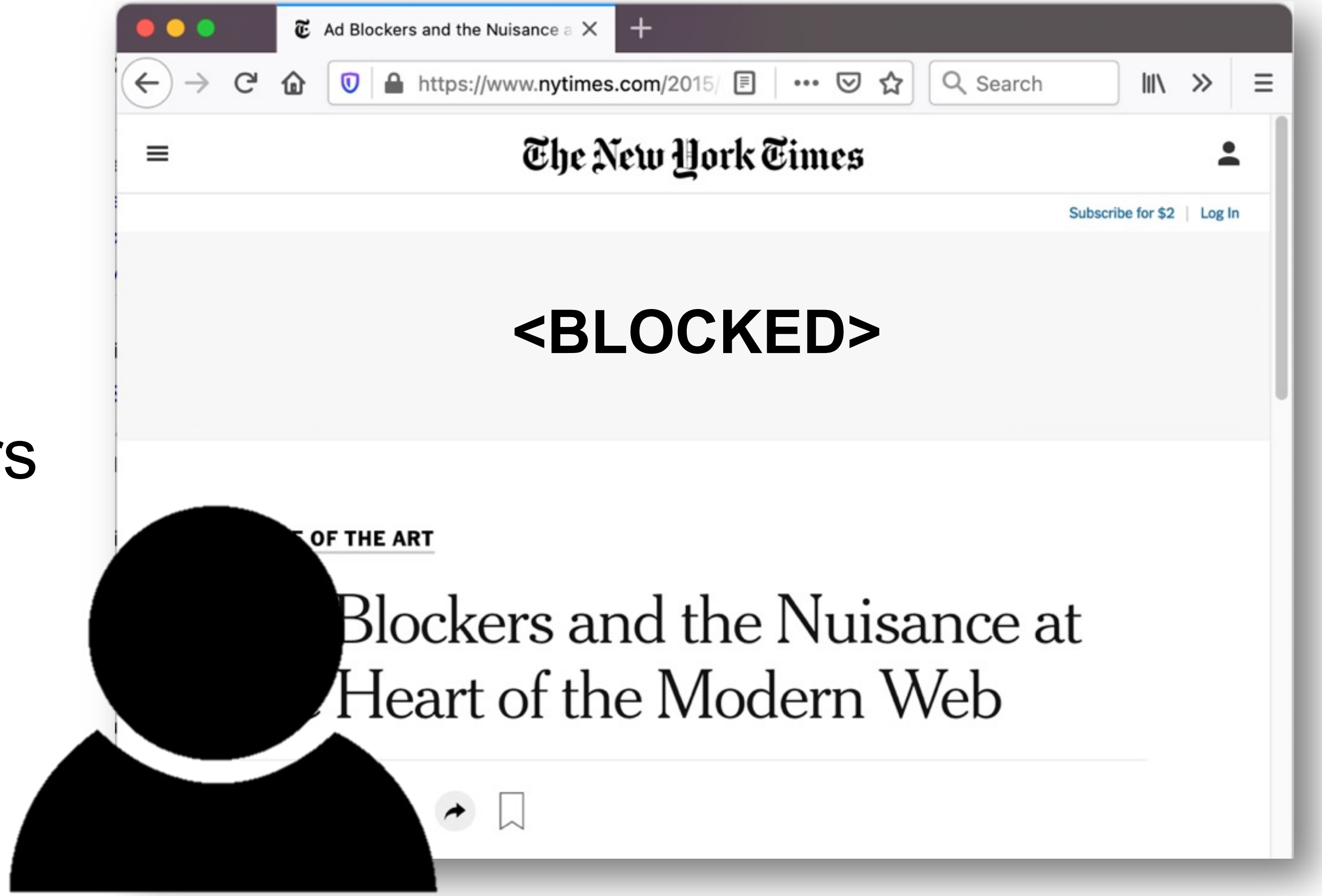
users install ad-blockers to remove ads...



Adversarial examples are a security threat for online ad-blocking.

users install ad-blockers to remove ads...

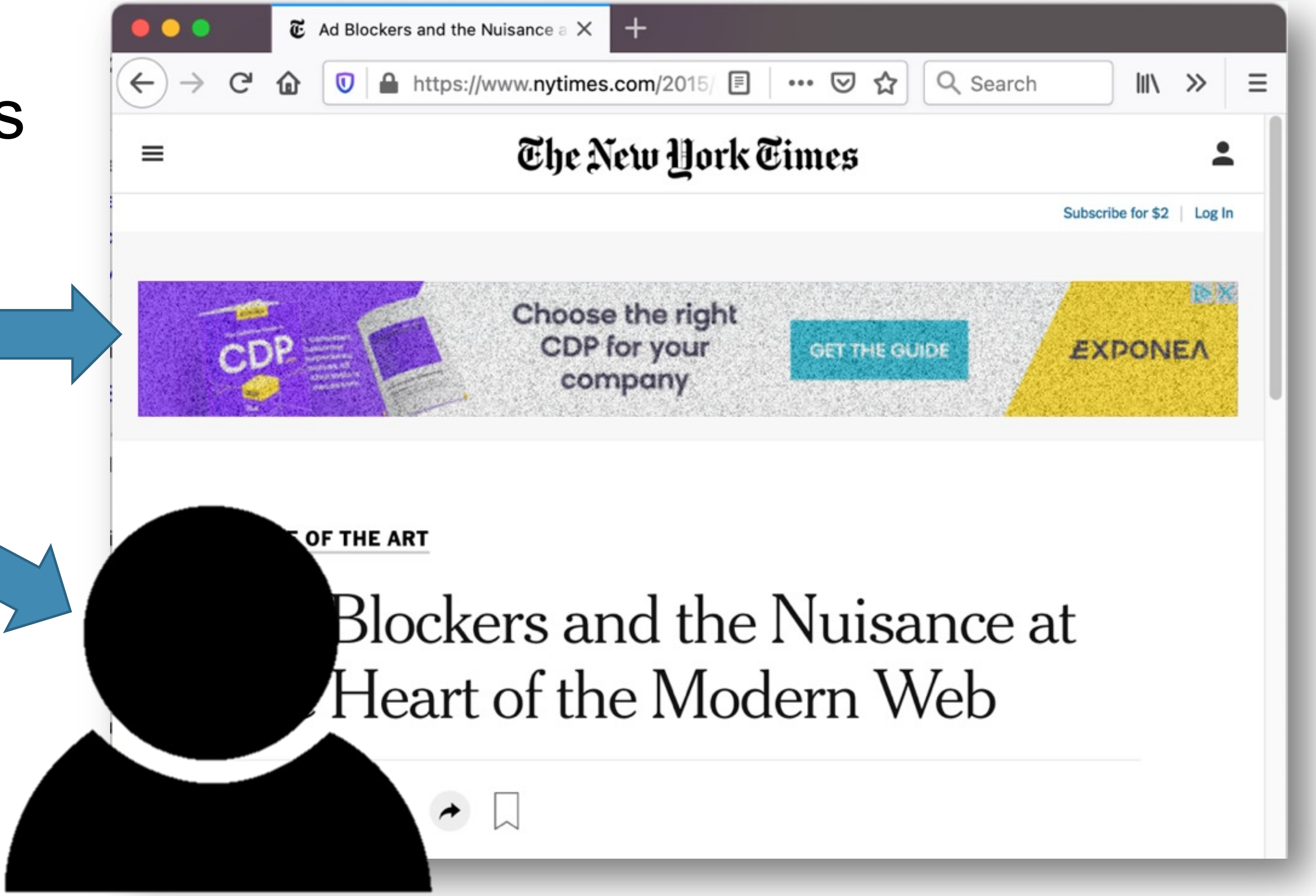
...using machine learning!



An attacker can use adversarial examples to evade content blocking.

adversaries (publishers & advertisers) modify content to evade blocking...

...without changing the user's visual perception of ads



For now, the adversary wins!

MOTHERBOARD
TECH BY VICE

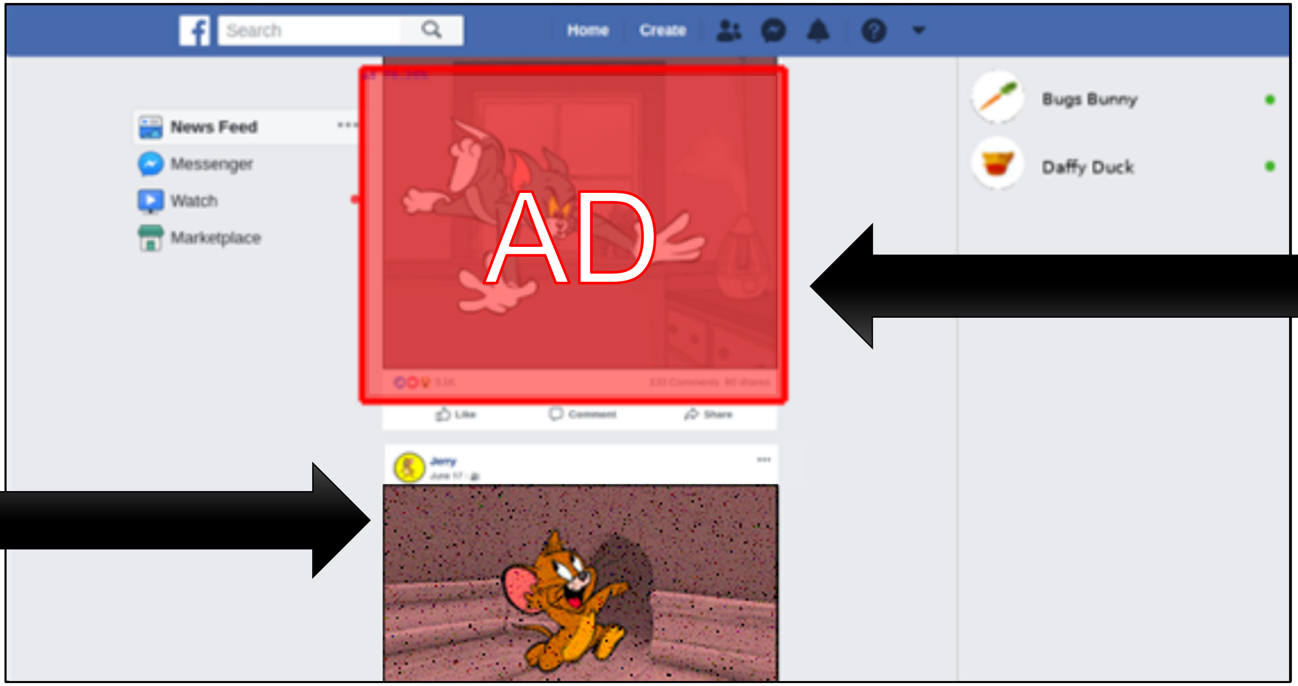
Researchers Defeat Most Powerful Ad Blockers, Declare a 'New Arms Race'

"*AdVersarial: Perceptual Ad Blocking meets Adversarial Machine Learning*", ACM CCS 2019

Adversarial examples can cause harm beyond model evasion.

Adblock Plus wants to run a ML model on *screenshots* of your entire Facebook feed.

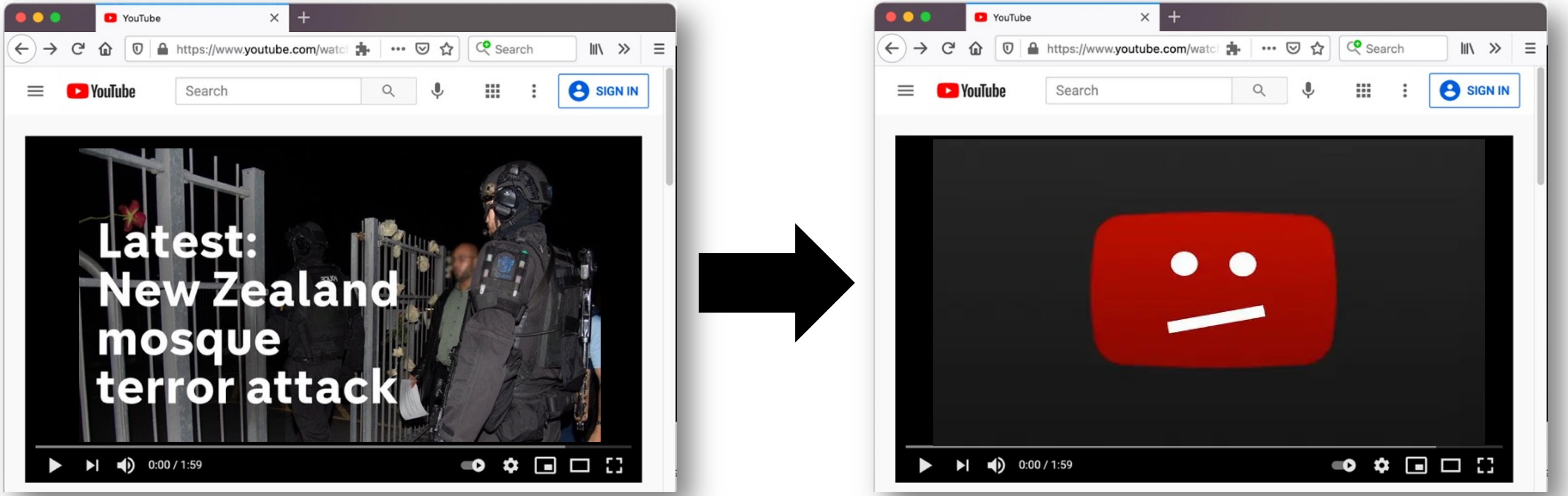
Jerry uploads malicious content …



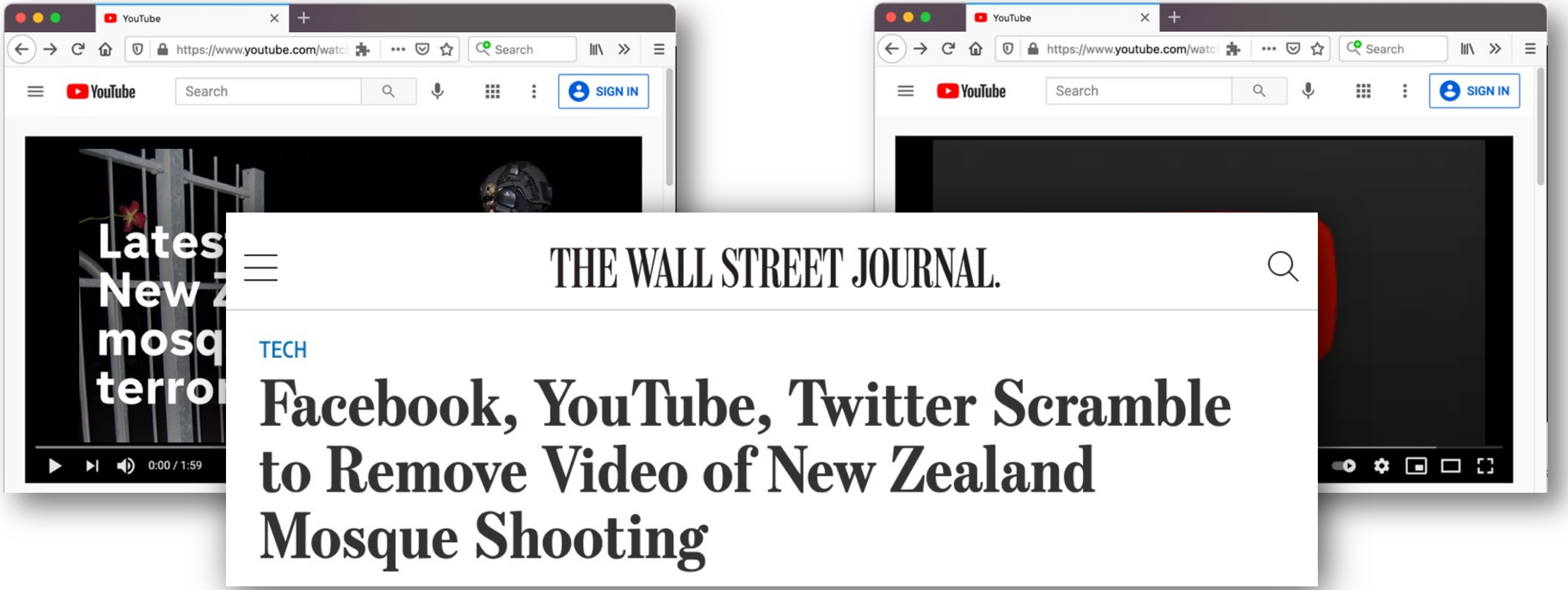
… so that Tom's post gets blocked

"AdVersarial: Perceptual Ad Blocking meets Adversarial Machine Learning", ACM CCS 2019

Adversarial examples are a security threat for online *content* blocking.



Adversarial examples are a security threat for online *content* blocking.



Talk outline.

- Adversarial examples for online content blockers
  - ➢ What’s the threat model?
  - ➢ Limitations of current defenses
  - ➢ Industry impact
- Enhancing ML privacy
- Future work

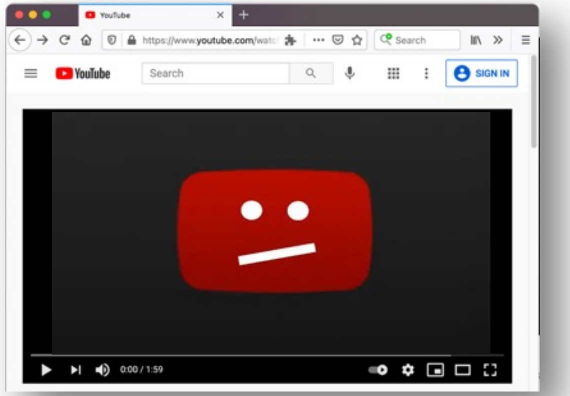
Talk outline.

- Adversarial examples for online content blockers
  - ➢ What's the threat model?
  - ➢ Limitations of current defenses
  - ➢ Industry impact
- Enhancing ML privacy
- Future work

Why focus on content blocking?

Many systems can be fooled with adversarial examples.

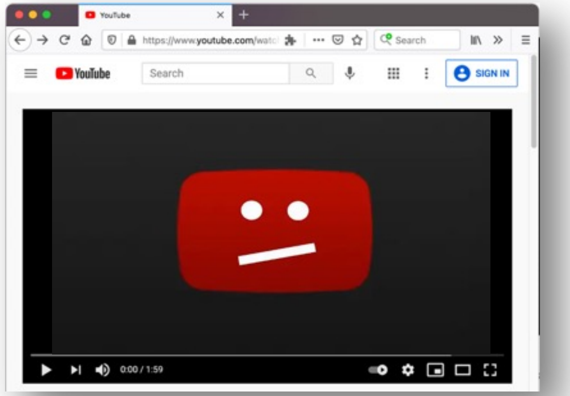
content blockers



Why focus on content blocking?

Many systems can be fooled with adversarial examples.

content blockers



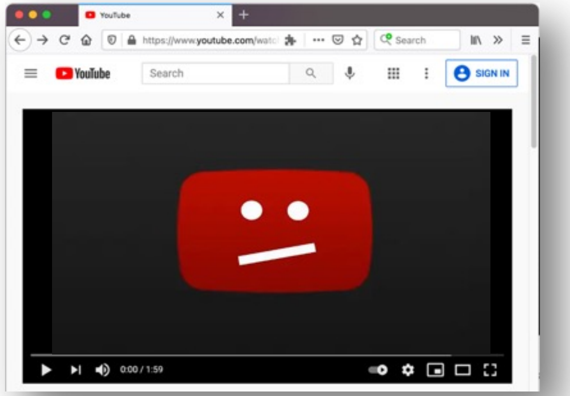
facial recognition

Sharif et al. 2016

Why focus on content blocking?

Many systems can be fooled with adversarial examples.

content blockers



facial recognition

Sharif et al. 2016

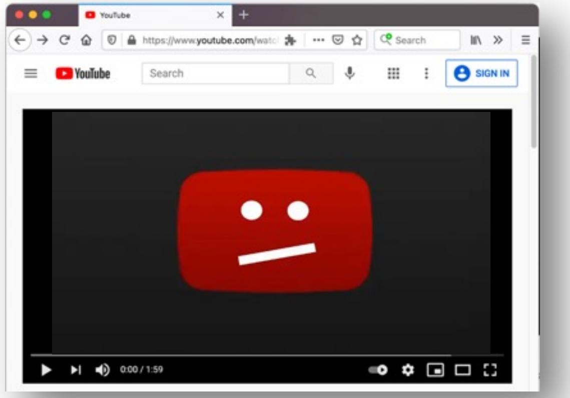
self-driving

Eykholt et al. 2018

Why focus on content blocking?

Many systems can be fooled with adversarial examples.

content blockers



facial recognition

Sharif et al. 2016

self-driving

Eykholt et al. 2018

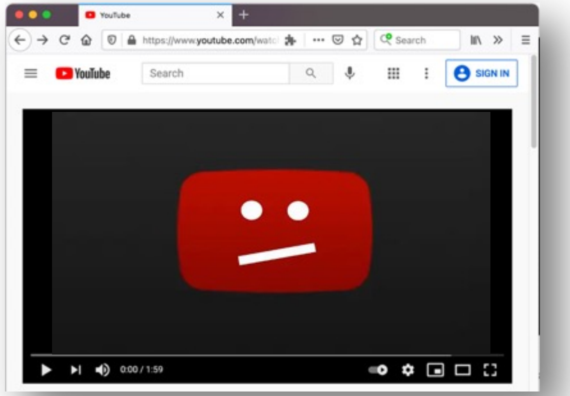
voice assistants

Carlini et al. 2016

Why focus on content blocking?

Many systems can be fooled with adversarial examples.

content blockers



facial recognition

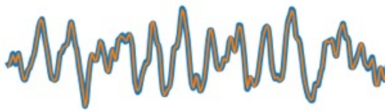
Sharif et al. 2016

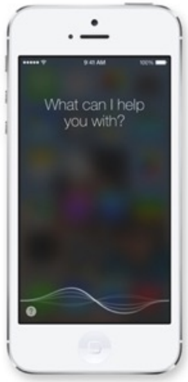
self-driving

Eykholt et al. 2018

voice assistants

Hey Siri!
open
evil.com



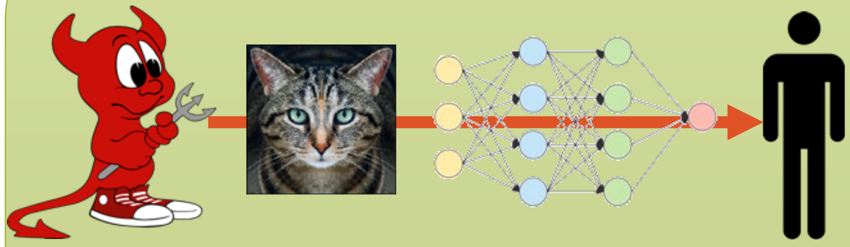


Carlini et al. 2016

Claim: adversarial examples are "overkill"!

Content blockers *always* operate in the presence of a human.

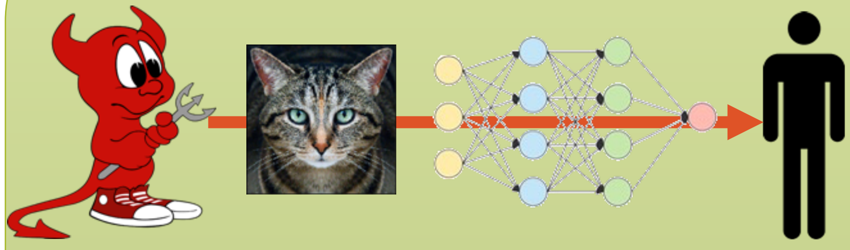
content blockers **facial recognition** **self-driving** **voice assistants**



adversary wants to fool the model *to get content shown to a human*

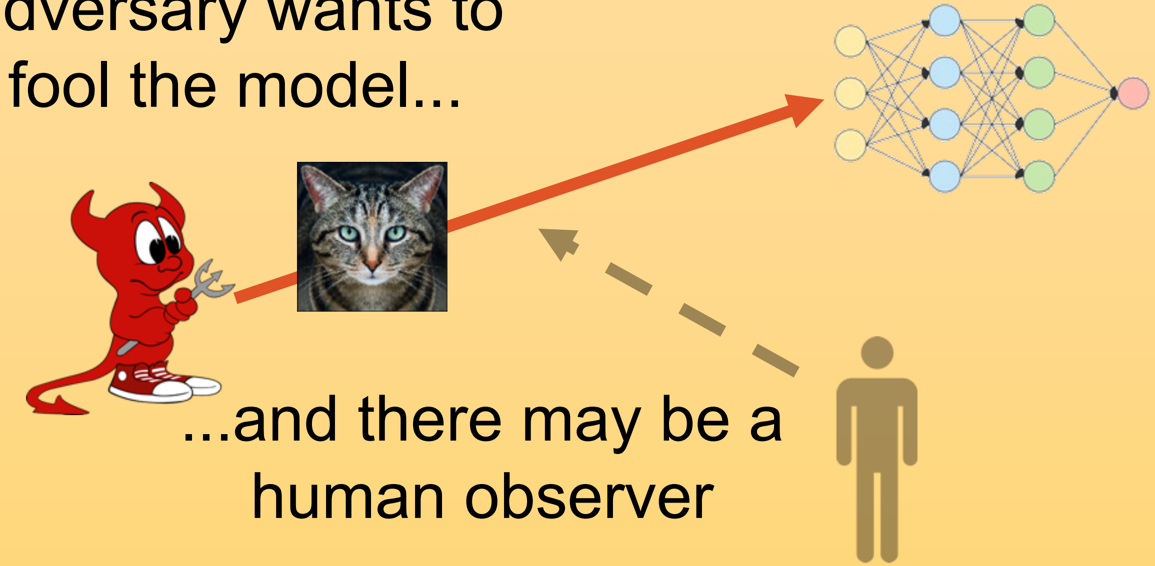
For other systems, security must hold whether there is a human observer or not.

content blockers | **facial recognition** | **self-driving** | **voice assistants**



adversary wants to fool the model *to get content shown to a human*

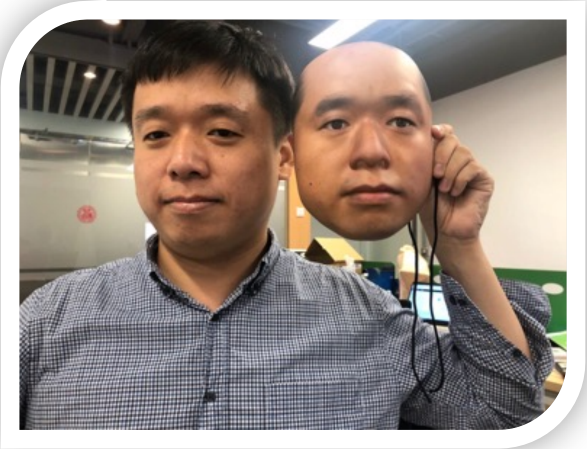
adversary wants to fool the model...



...and there may be a human observer

For such systems, security must also hold against “conspicuous” attacks.

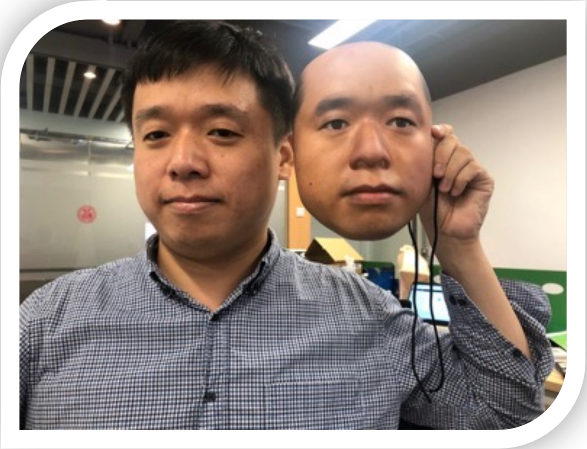
facial recognition



BUSINESS INSIDER

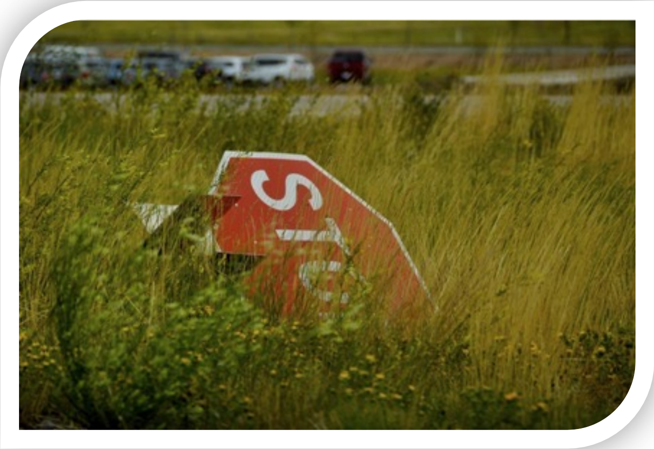
For such systems, security must also hold against “conspicuous” attacks.

facial recognition



BUSINESS INSIDER

self-driving



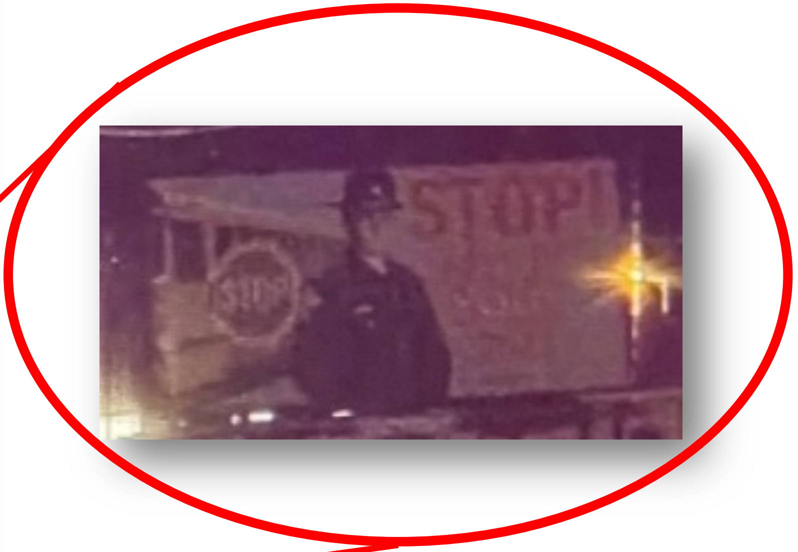
Olsson 2019

For such systems, security must also hold against “conspicuous” attacks.

facial recognition

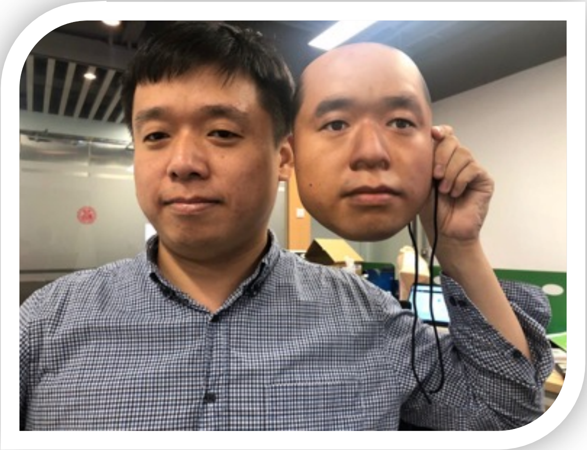
BUSINESS INSIDER

self-driving



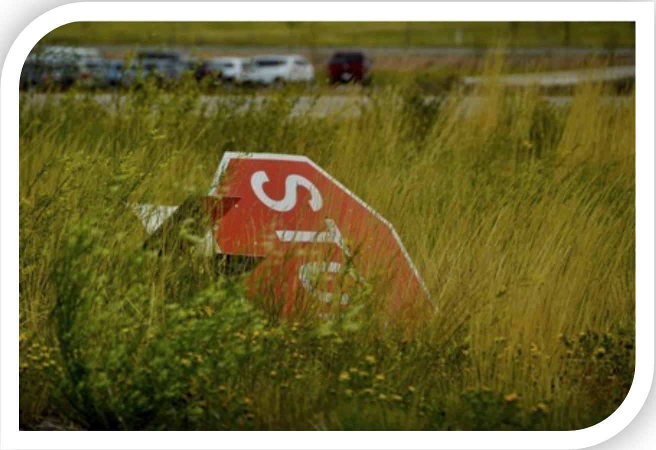
For such systems, security must also hold against “conspicuous” attacks.

facial recognition



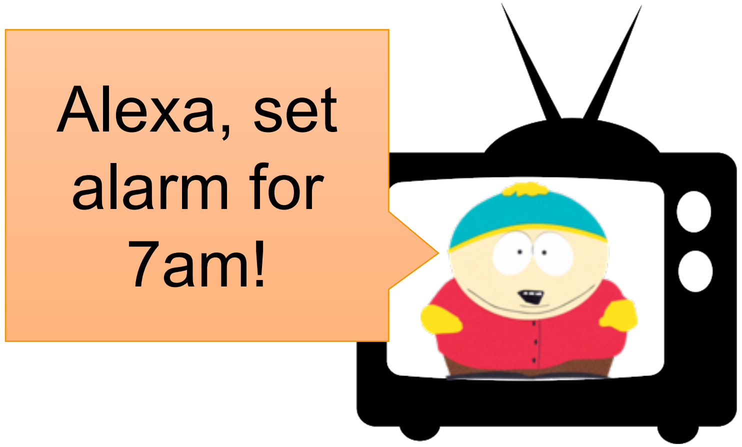
BUSINESS INSIDER

self-driving



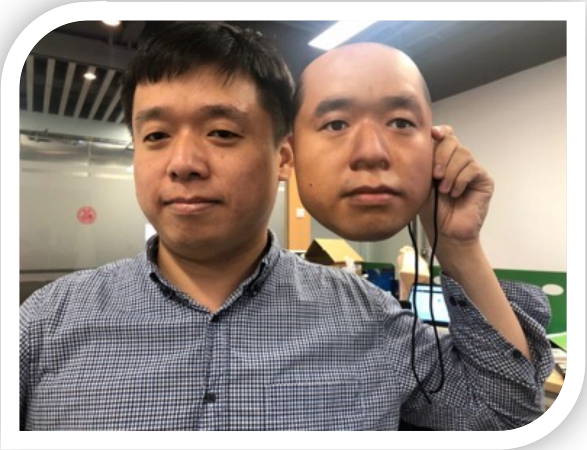
Olsson 2019

voice assistants



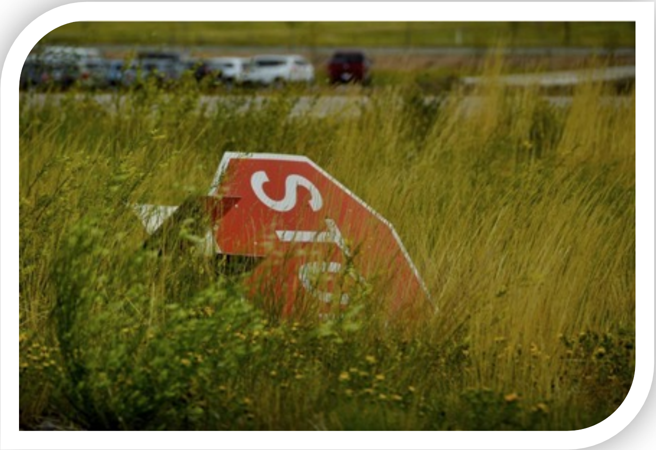
For such systems, security must also hold against “conspicuous” attacks.

facial recognition



BUSINESS INSIDER

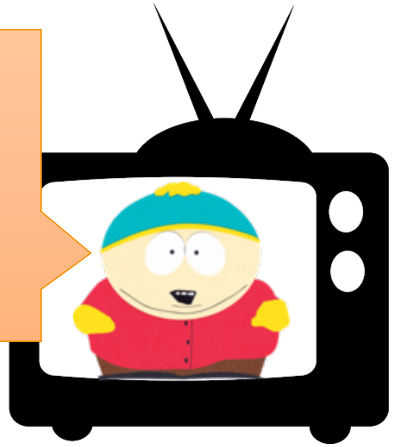
self-driving



Olsson 2019

voice assistants

Alexa, set alarm for 7am!



Content blocking is the only application where “small” perturbations are ***necessary*** for a successful attack.

Talk outline.

- Adversarial examples for online content blockers
  - What's the threat model?
  - Limitations of current defenses
  - Industry impact
- Enhancing ML privacy
- Future work

Talk outline.

- Adversarial examples for online content blockers
  - ➢ What's the threat model?
  - ➢ Limitations of current defenses
  - ➢ Industry impact
- Enhancing ML privacy
- Future work

Can we build a *robust* ML model?

“Yes”, but only in a very restrictive “toy” setting,
that has little relevance for practical attacks,
and the best defense only works <50% of the time,
and most defenses don’t work at all.

Short answer: No!

A formal model for robustness.

- Train a model $f(\cdot)$ on a distribution $\mathfrak{D}$ of labelled inputs (x, y)
- The adversary *perturbs* test inputs x sampled from $\mathfrak{D}$ with noise δ

Which perturbations δ do we allow?

- Ideal: any "semantically small" perturbation

ambiguous, hard to formalize

A formal model for robustness.

- Train a model $f(\cdot)$ on a distribution $\mathfrak{D}$ of labelled inputs (x, y)
- The adversary *perturbs* test inputs x sampled from $\mathfrak{D}$ with noise δ

Which perturbations $\boldsymbol{\delta}$ do we allow?

- Ideal: any "semantically small" perturbation
- Relaxation: perturbations δ from a ***fixed*** set S

$\max |\delta_i|$

Example: $S = \{\delta : ||\delta||_\infty \leq 20\%\}$

necessary but not sufficient

A formal model for robustness.

- Train a model $f(\cdot)$ on a distribution $\mathfrak{D}$ of labelled inputs (x, y)
- The adversary *perturbs* test inputs x sampled from $\mathfrak{D}$ with noise δ

Which perturbations δ do we allow?

- Ideal: any "semantically small" perturbation
- Relaxation: perturbations δ from a ***fixed*** set S

Example: $S = \{\delta : \|\delta\|_\infty \leq 20\%\}$

Ultimate goal:

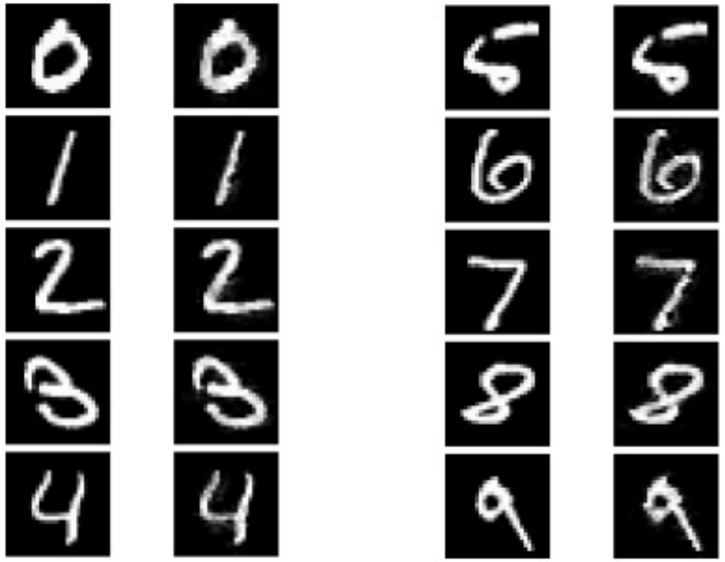
- discover defensive techniques that *generalize* across perturbation sets

The state-of-the-art in robust ML.

MNIST digit classification [LeCun et al., '98]

➢ considered "solved" by ML (>99.5% accuracy)

➢ 0% accuracy when each pixel value can be perturbed by 20%



[Carlini & Wagner., '17]

Most proposed defenses are broken!

[Carlini & Wagner ’17], [Athalye et al. ’18], [**T**, Carlini, Brendel, Mądry (NeurIPS 2020)], ...

- *denoising*
- *randomization*
- *dimensionality reduction*
- *input transformations*
- *generative modeling*
- *Bayesian learning*
- *...*

Some defenses work.

- Adversarial training [Szegedy et al. '13], [Goodfellow et al. '14], [Kurakin et al. '16], [**T** et al. '17], [Madry et al. '18], [Zhang et al. '19], [Carmon et al. '19], [Uesato et al. '19], [Zhai et al. '19], [Shafahi et al. '19], [Yang et al. '19], [Li et al. '20], ...
- Certified defenses [Katz et al. '17], [Wong et al. '17], [Raghunathan et al. '18], [Gehr et al. '18], [Lecuyer et al. '18], [Zhang et al. '18], [Mirman et al. '18], [Weng et al. '19], [Baluta et al. '19], [Cohen et al. '19], [Singh et al. '19], [Gluch et al. '20], ...

Some defenses work, **but don't generalize...**

- Adversarial training — [Szegedy et al. '13], [Goodfellow et al. '14], [Kurakin et al. '16], [**T** et al. '17], [Madry et al. '18], [Zhang et al. '19], [Carmon et al. '19], [Uesato et al. '19], [Zhai et al. '19], [Shafahi et al. '19], [Yang et al. '19], [Li et al. '20], ...
- Certified defenses — [Katz et al. '17], [Wong et al. '17], [Raghunathan et al. '18], [Gehr et al. '18], [Lecuyer et al. '18], [Zhang et al. '18], [Mirman et al. '18], [Weng et al. '19], [Baluta et al. '19], [Cohen et al. '19], [Singh et al. '19], [Gluch et al. '20], ...

recall: we only consider perturbations δ from a *fixed* set S

issue: all defenses above are *explicitly tailored to a chosen set* S

defenses overfit to the chosen set

T, Behrmann, Carlini, Papernot, Jakobsen (ICML 2020)

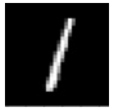
generalizing to richer sets hurts robustness

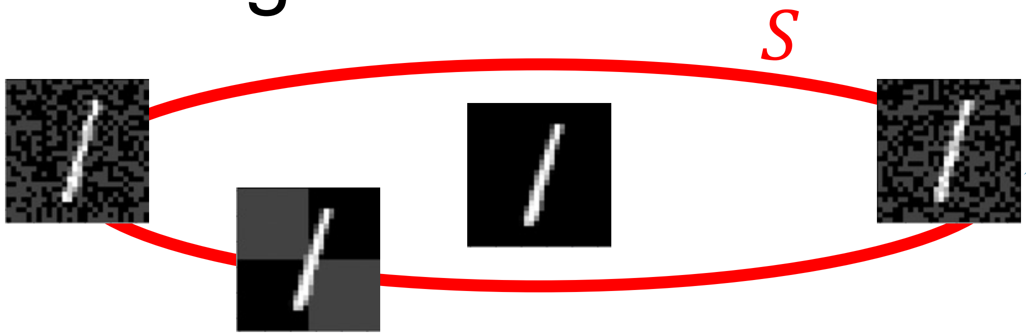
T & Boneh (NeurIPS 2019 *spotlight*)

Adversarial training: a defense for a *fixed* perturbation set.

[Szegedy et al., ’14], [Goodfellow et al., ‘15], [Madry et al., ’17]

max. per-pixel noise

1. Choose a set S of perturbations: e.g., $S = \{\delta : ||\delta||_\infty \leq 20\%\}$
2. For each input , find the *worst* adversarial example: 
3. Train the model on 
4. Repeat until convergence



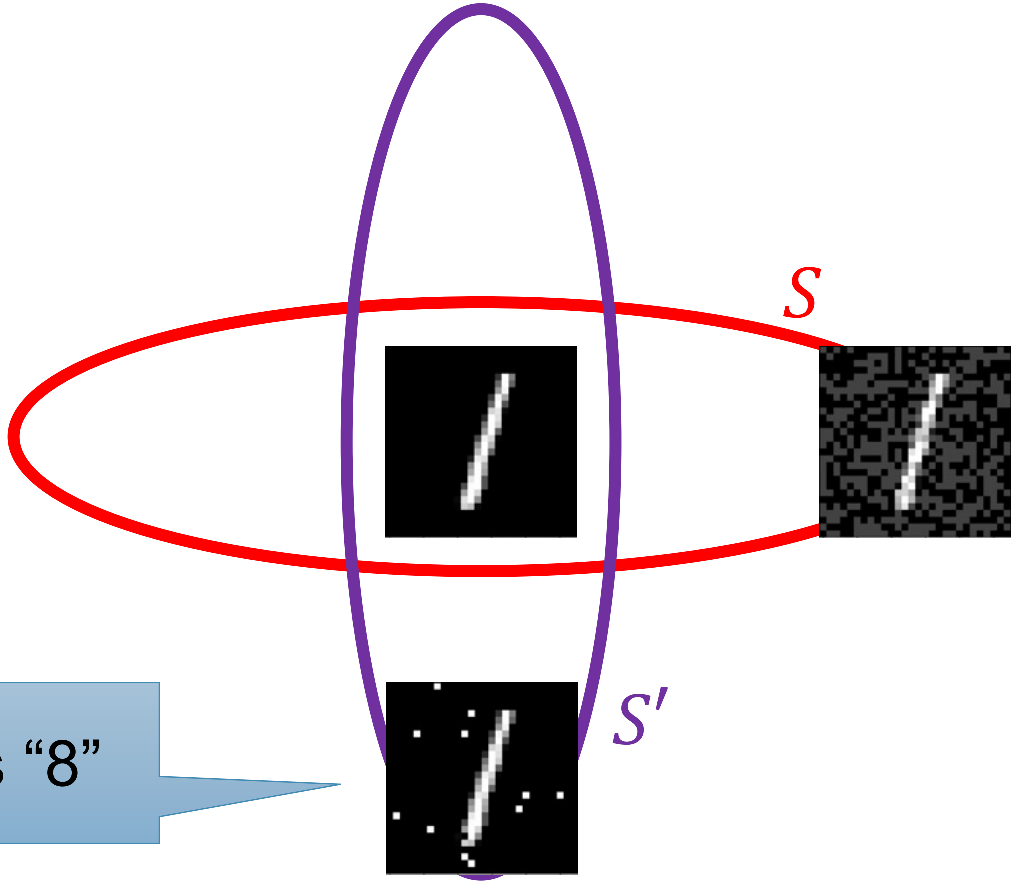
all images in the set are classified as “1”

Defenses fail for noise *outside the chosen set.*

[Engstrom et al., ’17], [Sharma & Chen, ’18]

sum of perturbed pixels

- Attack with a perturbation from $S' = \{\delta : \|\delta\|_1 \leq 12\}$



misclassified as “8”

Why not learn to resist multiple noise types?

T & Boneh (NeurIPS 2019 *spotlight*)

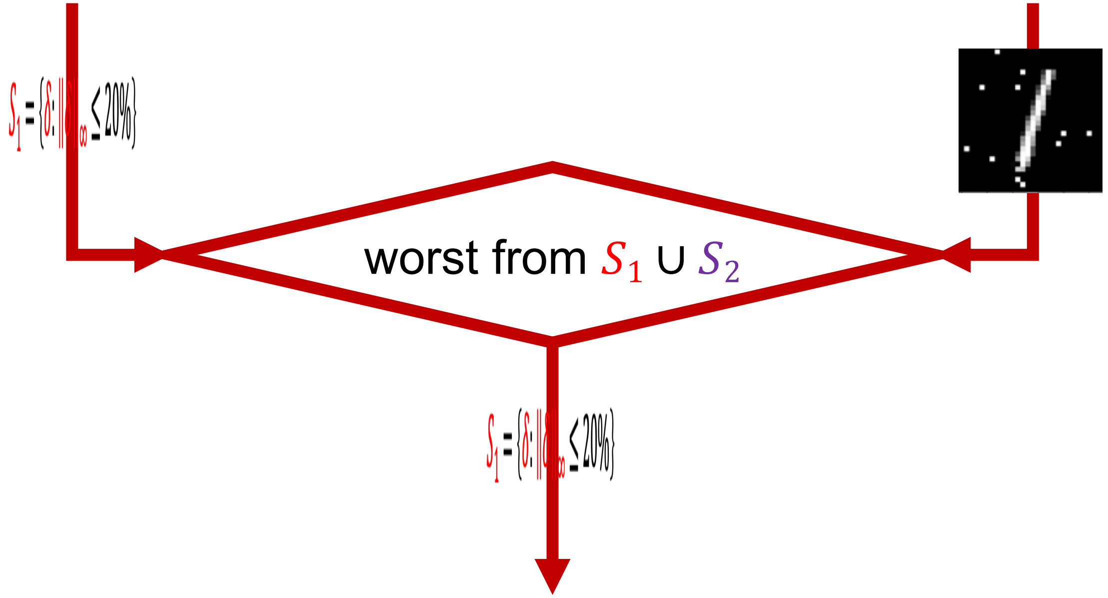
1. Choose **multiple** sets of perturbations $S_1, S_2, \ldots$
2. Train a model against worst perturbation from $S_1 \cup S_2 \cup \ \ldots$

$S_1 = \{\delta : \|\delta\|_\infty \leq 20\%\}$

$S_2 = \{\delta : \|\delta\|_1 \leq 12\}$

pick worst-case noise from S_1

pick worst-case noise from S_2



Resisting multiple noise types is costly.

T & Boneh (NeurIPS 2019 *spotlight*)

1. Choose **multiple** sets of perturbations $S_1, S_2, \ldots$
2. Train a model against worst perturbation from $S_1 \cup S_2 \cup \ \ldots$

Can adversarial training solve adversarial examples?

recall our ultimate goal:
defenses that are robust to <u>any</u> “small" perturbation

- ➢ adversarial training requires knowing the perturbation set *a priori*

Theorem (informal): [**T**, Behrmann, Carlini, Papernot, Jakobsen, ICML 2020]

Finding a “complete” perturbation set is **as hard as** building a “perfect” classifier.

Take away: we **don't** have robust machine learning in adversarial settings.

THE WALL STREET JOURNAL.

TECH

Facebook, YouTube, Twitter Scramble to Remove Video of New Zealand Mosque Shooting

MOTHERBOARD
TECH BY VICE

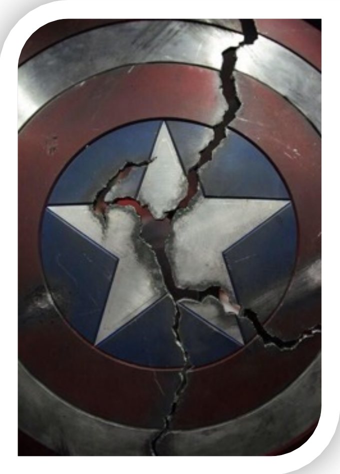
Researchers Defeat Most Powerful Ad Blockers, Declare a 'New Arms Race'

Take away: we **don’t** have robust machine learning in adversarial settings.

But, we now have:

1. *industry awareness of security risks*

2. *understanding of inherent limitations of defenses*



Talk outline.

- Adversarial examples for online content blockers
  - ➢ What's the threat model?
  - ➢ Limitations of current defenses
  - ➢ Industry impact
- Enhancing ML privacy
- Future work

Talk outline.

- Adversarial examples for online content blockers
  - What's the threat model?
  - Limitations of current defenses
  - Industry impact
- Enhancing ML privacy
- Future work

Adblock Plus and (a little) more

Sentinel is Online

• 2018-06-27 16:05 by Tom Woolford

Are you ready to feed the machine?

MOTHERBOARD
TECH BY VICE

Researchers Defeat Most Powerful Ad Blockers, Declare a ‘New Arms Race’

Adblock Plus 3.6.2 is Out and With Interesting Updates

Adblock Plus

Because of the obvious limitations of Sentinel, we came up with a highly-usable perceptual ad-blocking approach, in the form of the newly released perceptual hashing snippet. It does not use any machine-learning techniques per se, but it marks a first ever perceptual ad-blocking approach in production, and allows us to grow in an innovative way.

AdChoices

AdChoices

Goal: detect ad disclosures using image hashes

Problem: these techniques are not robust either

Where are the names

Anonymous Coward an hour ago

Seriously? Where are the names of these scumbags^d researchers. I'm driving down to Stanford, stopping by a Home Depot to pick a 2x4, a bag of lye and a shovel. Will have some very intimate conversations with these "researchers"

Reply Sha

Shut down unethical project #1

Open impredicative opened this issue 20 days ago · 0 comments

impredicative commented 20 days ago • edited

Florian Tramèr,

This project seems grossly unethical and it should be shut down. Are the department head and dean at Stanford University aware of this unethical work?

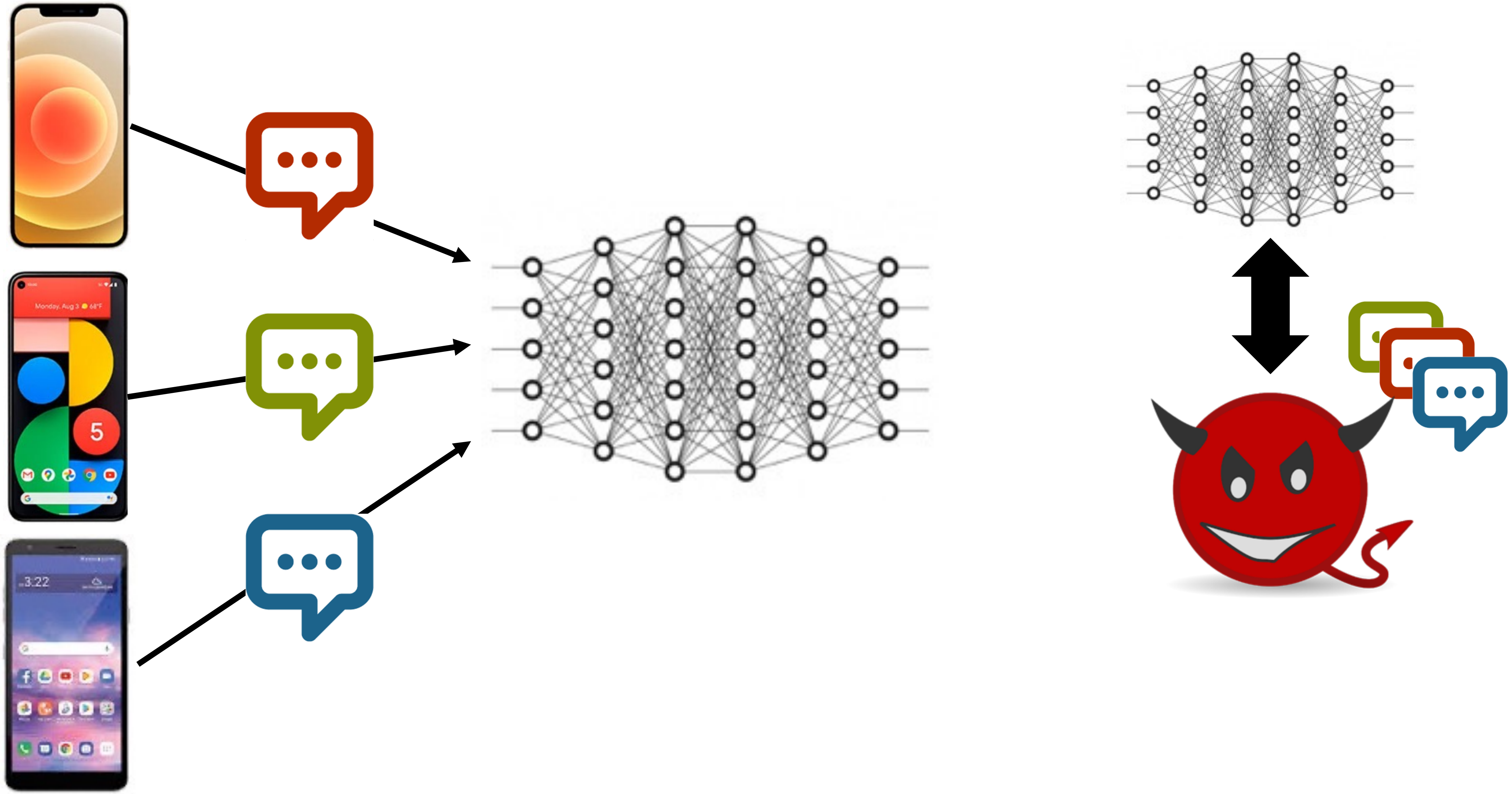
Talk outline.

- Adversarial examples for online content blockers
  - ➢ What's the threat model?
  - ➢ Limitations of current defenses
  - ➢ Industry impact
- Enhancing ML privacy
- Future work

Talk outline.

- Adversarial examples for online content blockers
  - ➢ What's the threat model?
  - ➢ Limitations of current defenses
  - ➢ Industry impact
- Enhancing ML privacy
- Future work

ML models are often trained on private data.



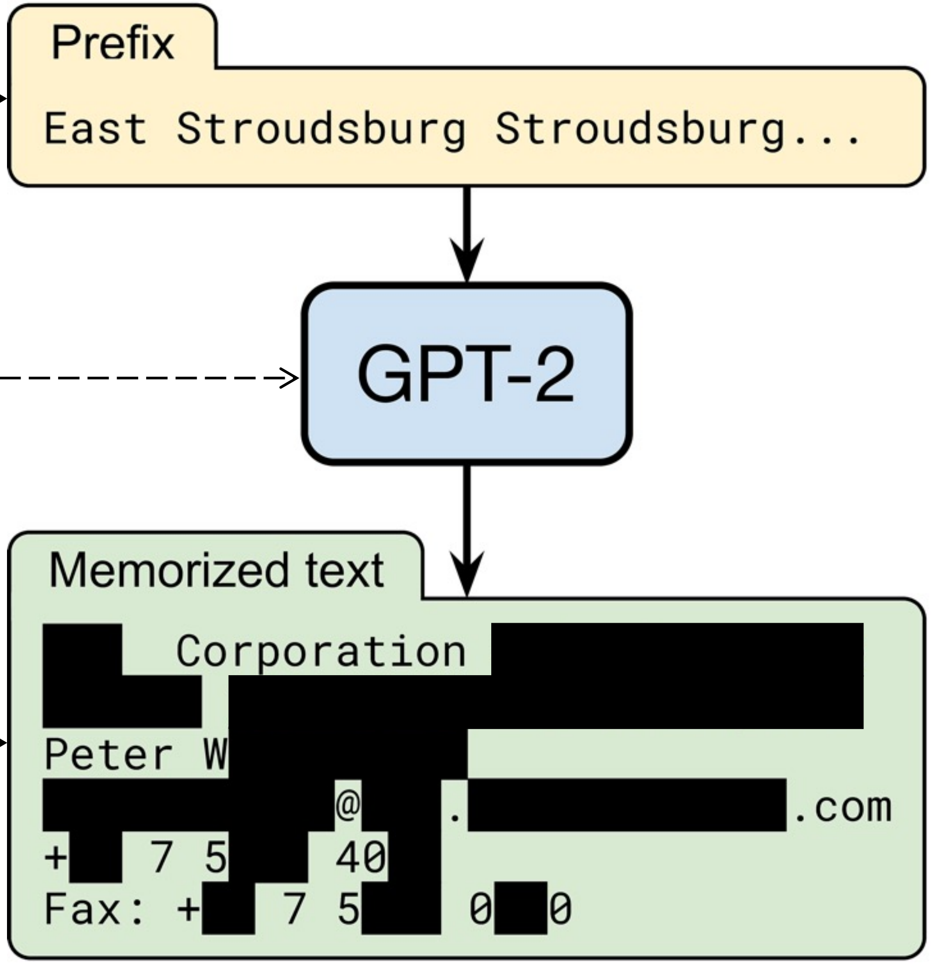
Challenge: models leak their training data.

Carlini, **T**, Wallace, Jagielski, Herbert-Voss, Lee et al. (preprint 2020)

random input

OpenAI’s language model trained on text from 8 million web pages

someone’s contact information output by the model (***redacted for privacy***)



Data leaks have dramatic ***consequences!***

for users...

The New York Times

Data Breach Victims Talk of Initial Terror, Then Vigilance

for companies...

ZDNet

Facebook could face $1.63bn fine under GDPR over latest data breach

TechCrunch

FTC settlement with Ever orders data and AIs deleted after facial recognition pivot

Preventing data leakage with decade-old ML

T & Boneh (ICLR 2021 *spotlight*)

- *provably* prevent leakage of training data.
  using *differential privacy*

  Extensions: distributed or federated learning
  [Dean et al. ‘12], [McMahan et al. ‘16], [Lian et al. ‘17]

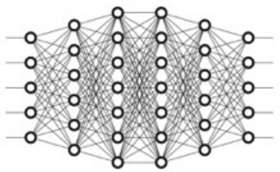
- *better accuracy* than with deep learning methods.
  using domain-specific *feature engineering*

Differential privacy prevents data leakage.

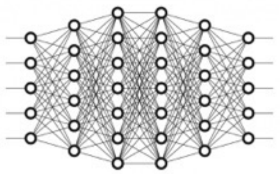
[Dwork et al. ‘06]

intuition: *randomized* training algorithm is not influenced (too much) by any individual data point

for any two datasets that differ in a single element

$\Pr[A_{\text{train}}($  $) =$  $]$

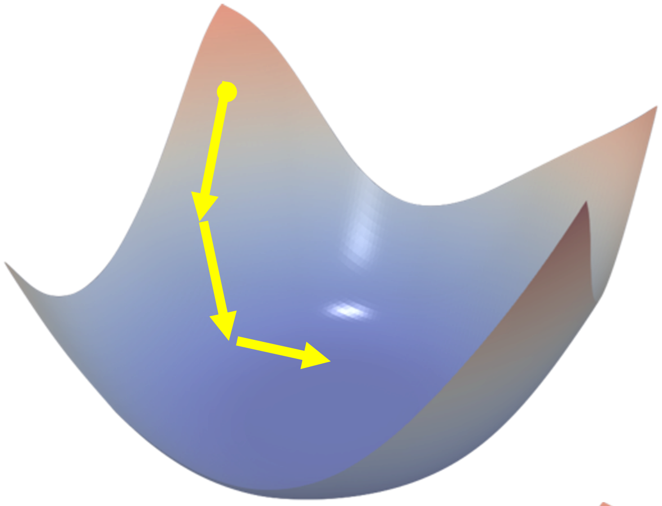
divided by

$\Pr[A_{\text{train}}($  $) =$  $]$

$\le e^{\varepsilon}$

Differentially private learning is possible with *noisy* gradient descent.

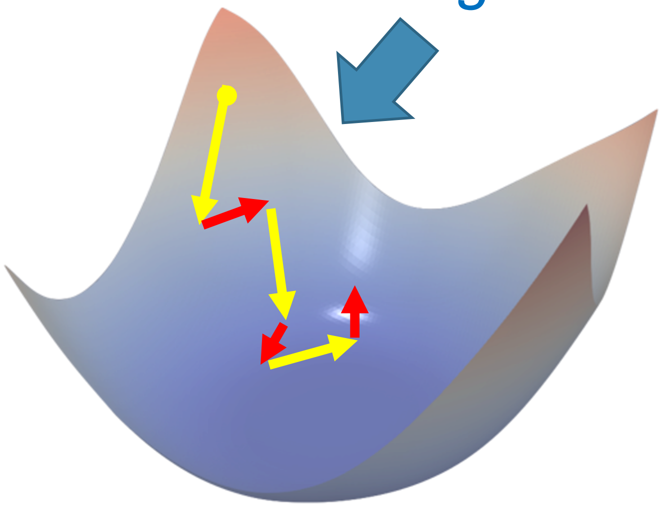
Gradient descent



add noise to each step
to guarantee privacy

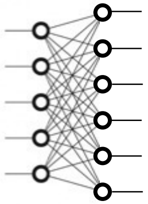
Private gradient descent

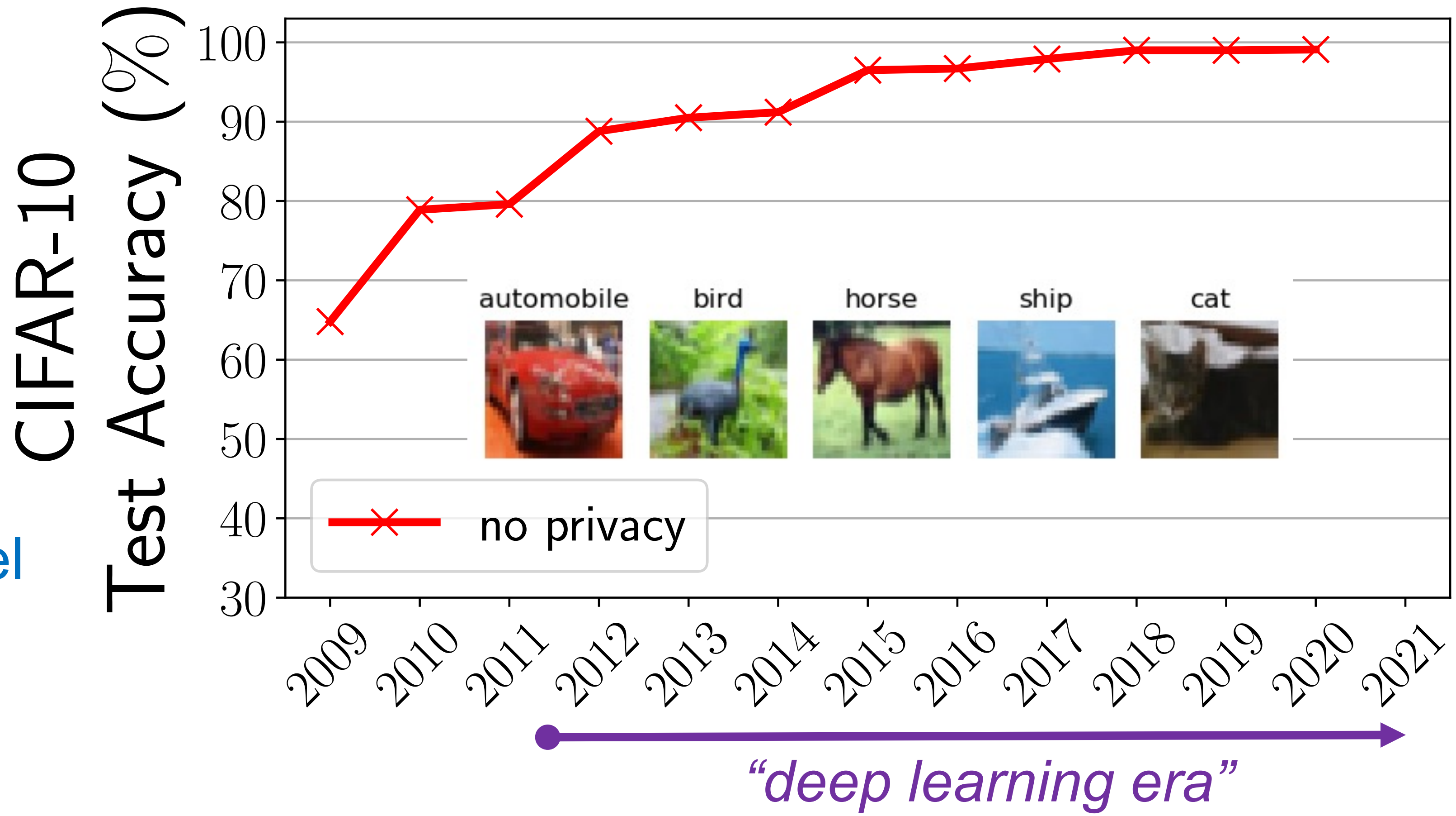
[Chaudhuri et al., ‘11], [Bassily et al. ‘14],
[Shokri & Shmatikov ‘15], [Abadi et al. ‘16], ...



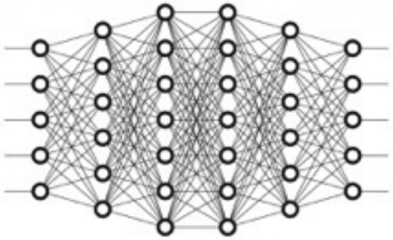
Non-private deep learning can achieve near-perfect accuracy.

"shallow" model

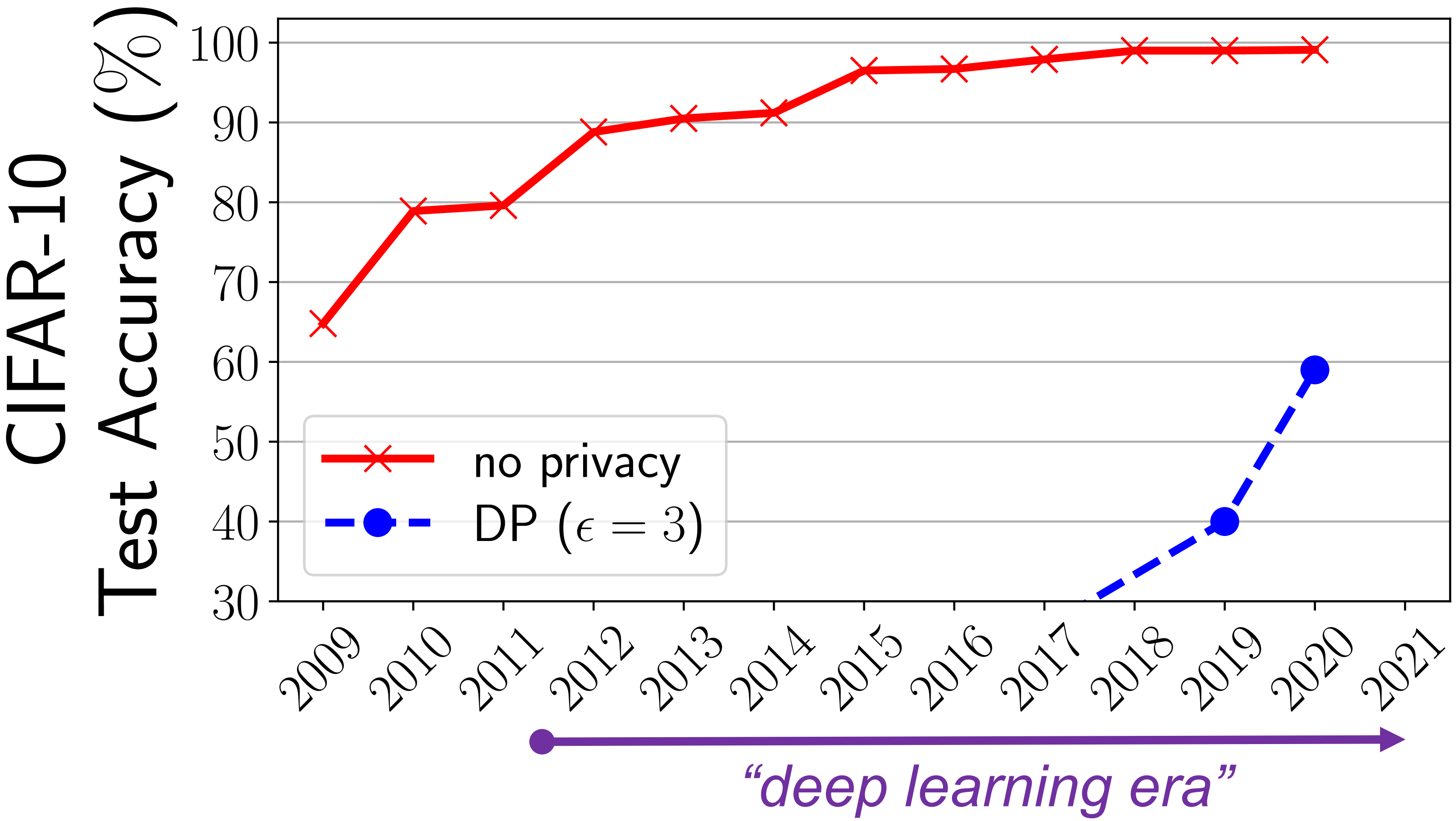




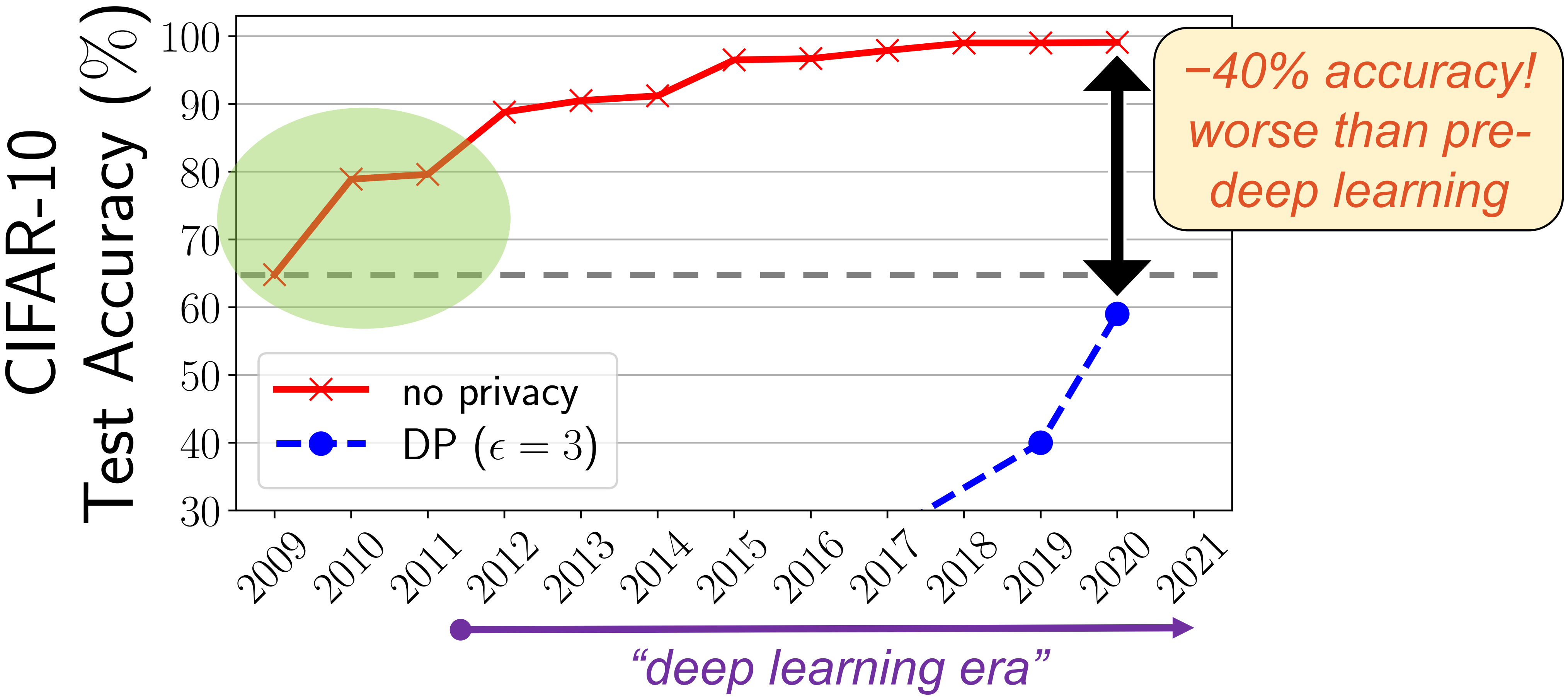
"deep" model



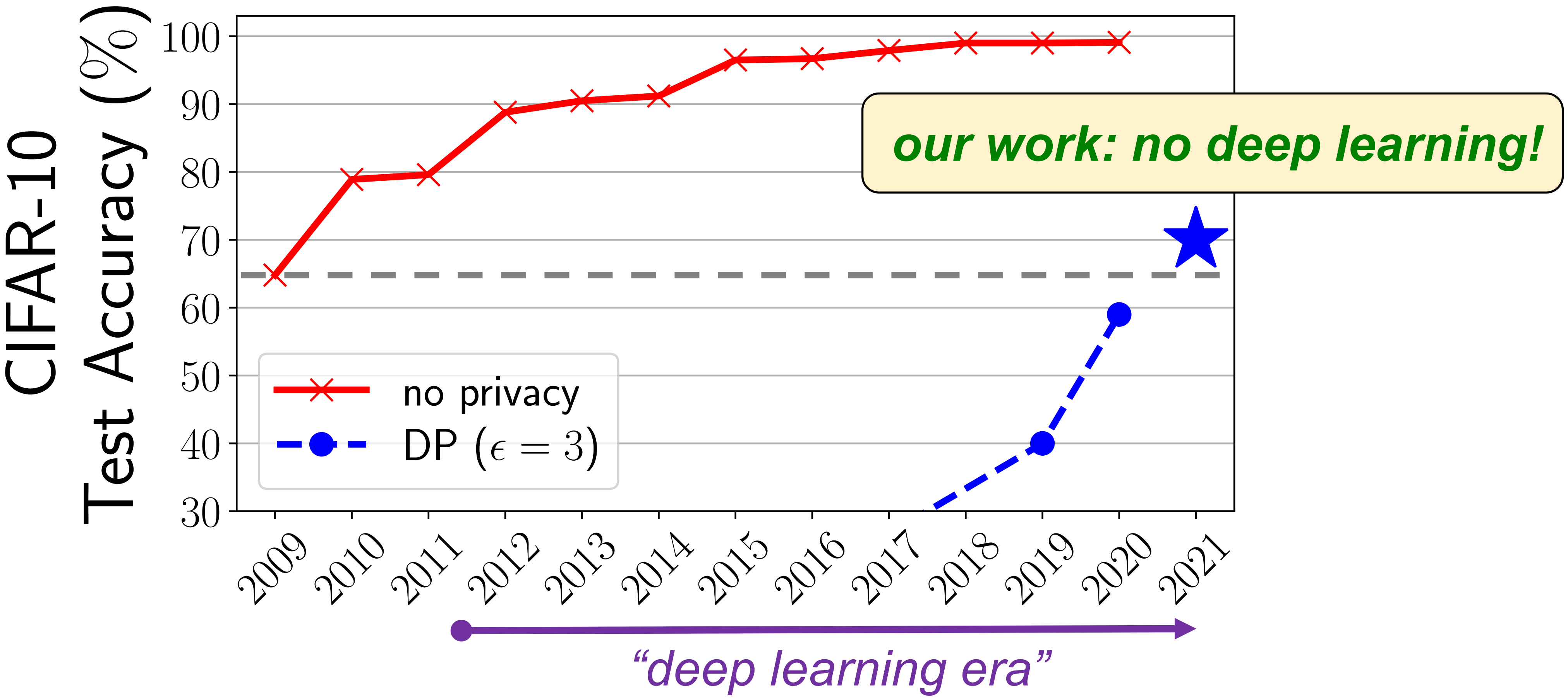
Differentially private deep learning lowers accuracy significantly.



Differentially private deep learning lowers accuracy significantly.



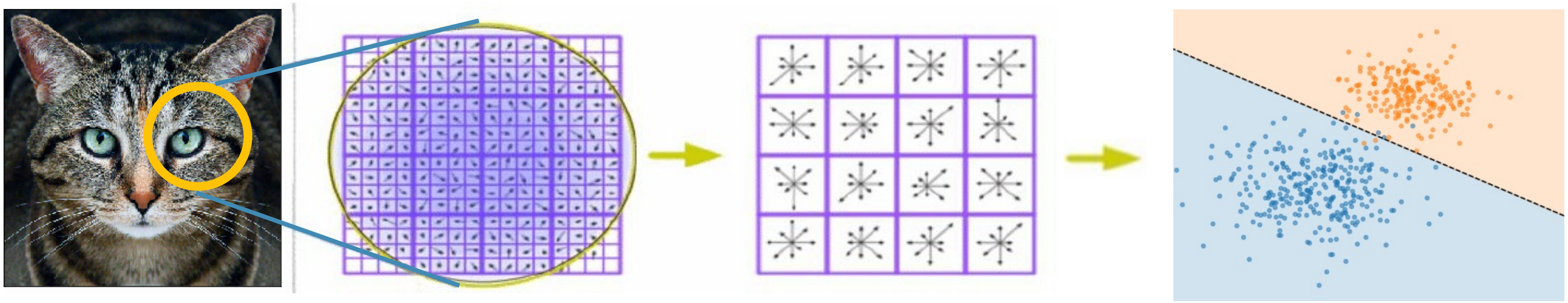
Differential privacy *without deep learning* improves accuracy.



"*Differentially private learning needs better features*", ICLR 2021 *spotlight*

Privacy-free features from "old-school" image recognition.

SIFT [Lowe '99, '04], HOG [Dalal & Triggs '05], SURF [Bay et al. '06], ORB [Rublee et al. '11], ...
Scattering transforms: [Bruna & Mallat '11], [Oyallon & Mallat '14], ...



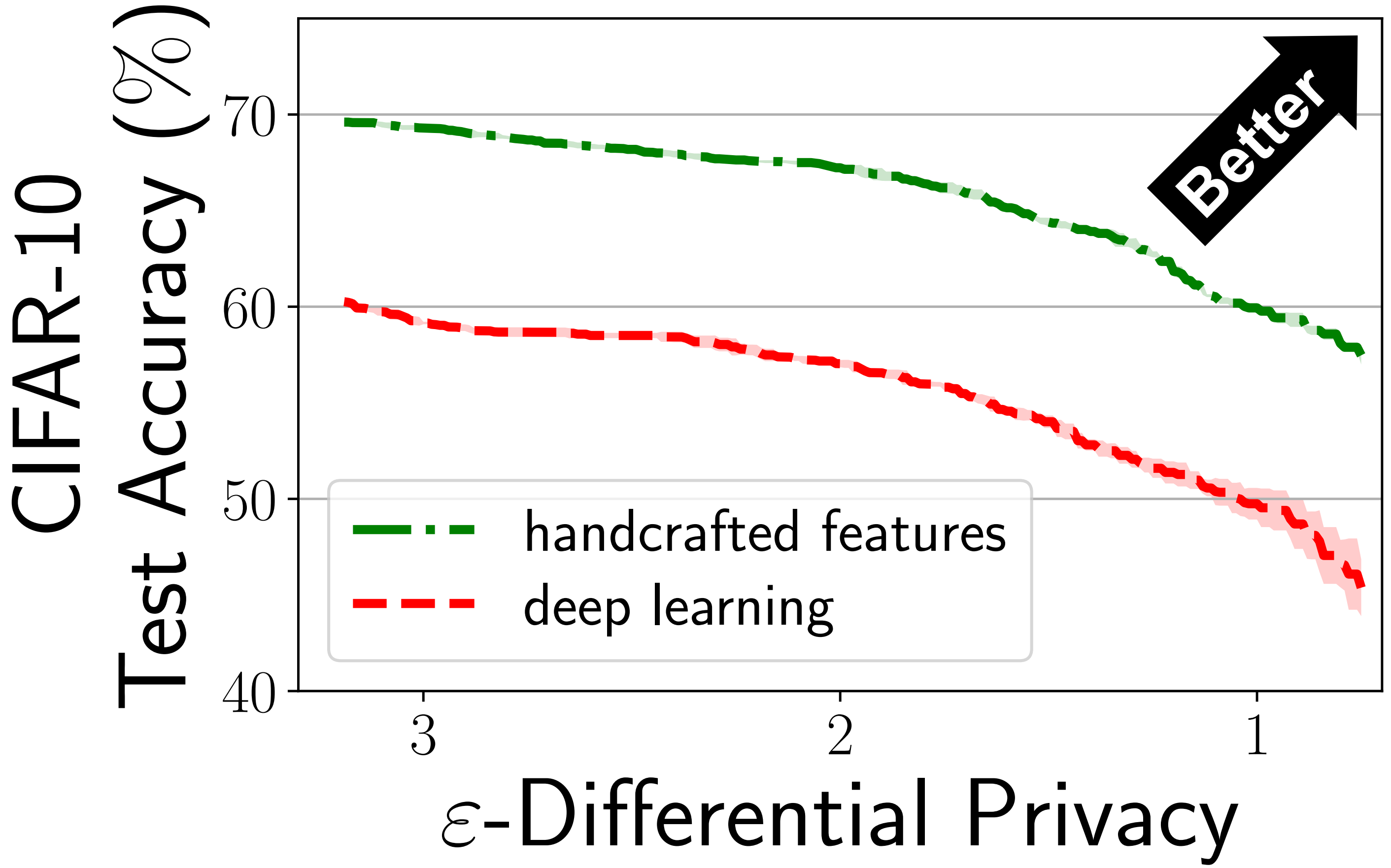
"handcrafted features"
(no learning involved)

simple classifier
(e.g., logistic regression)

privacy free

captures some *prior* about the domain: e.g., invariance under rotation & scaling

Handcrafted features lead to a better tradeoff between accuracy and privacy.

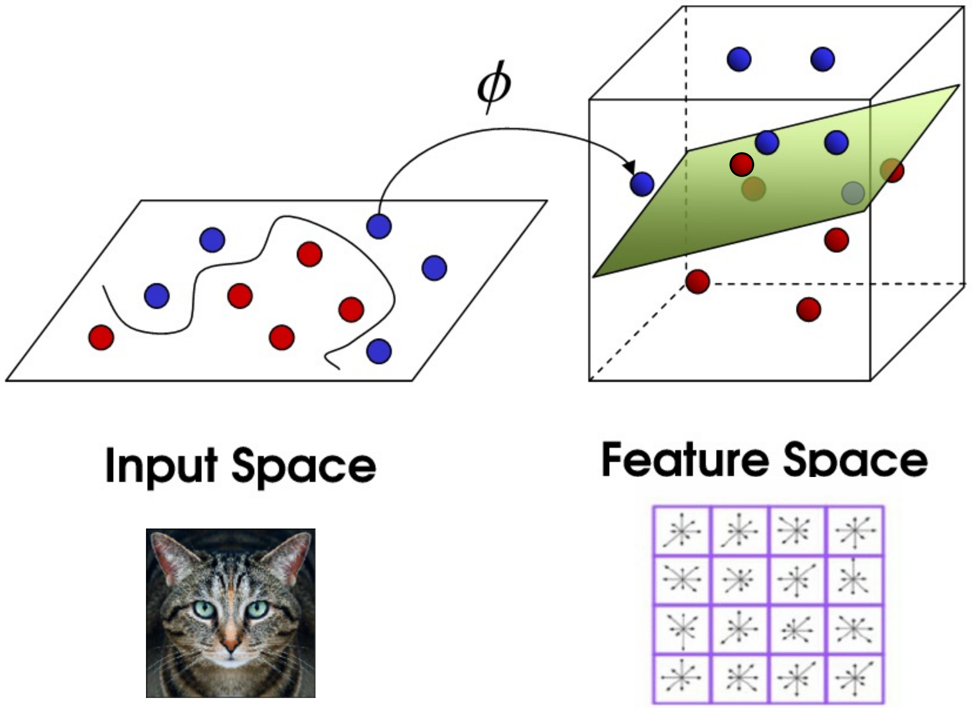


"*Differentially private learning needs better features*", ICLR 2021 *spotlight*

Handcrafted features lead to an *easier* learning task (for noisy gradient descent).

high accuracy classifier exists but learning takes *many gradient steps*

bad for privacy



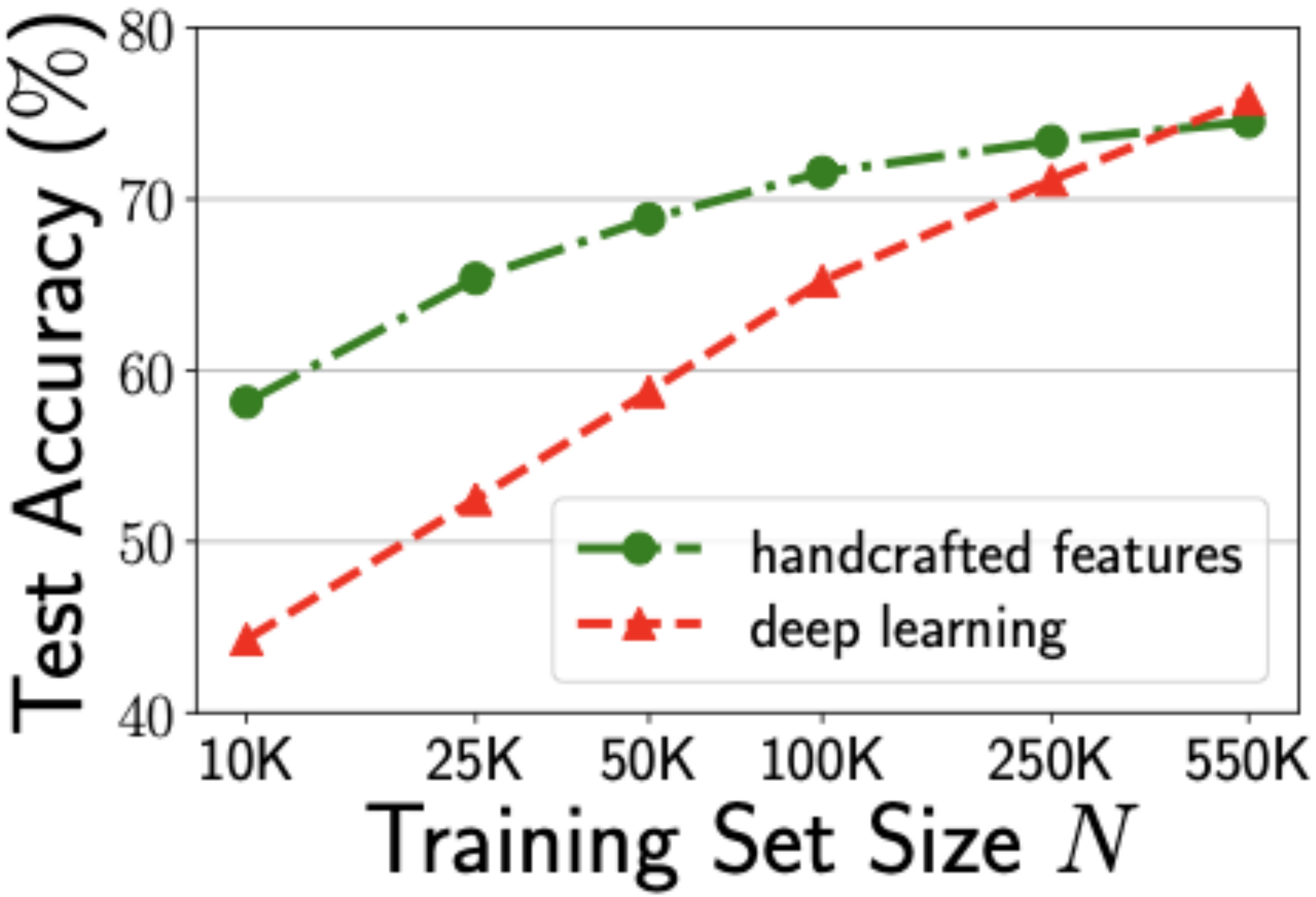
in feature space, maximal accuracy is reduced but *learning progresses faster*

good for privacy

“*Differentially private learning needs better features*”, ICLR 2021 *spotlight*

Surpassing handcrafted features with *more private data.*

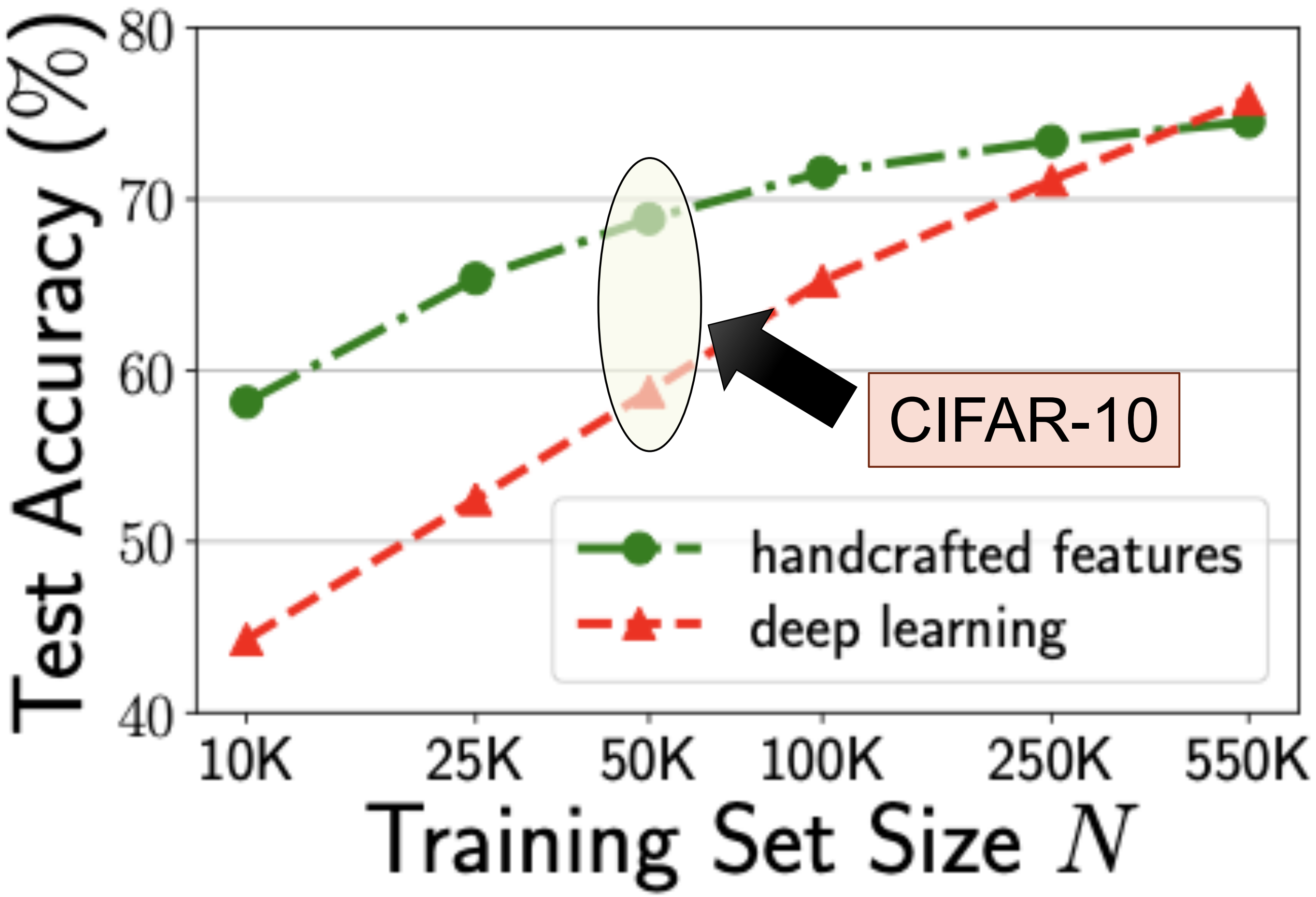
(for ε = 3)



"*Differentially private learning needs better features*", ICLR 2021 *spotlight*

Surpassing handcrafted features with *more private data.*

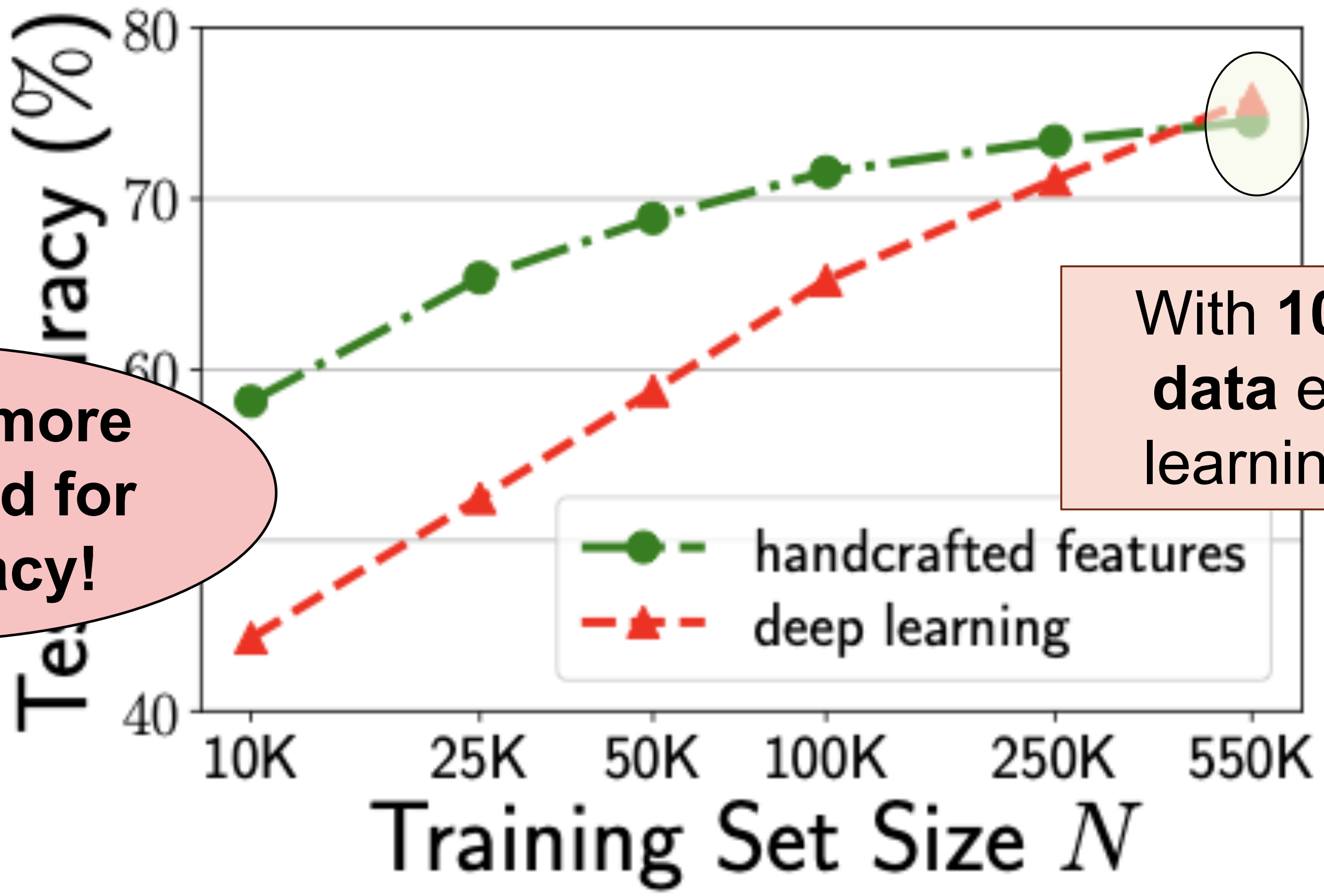
(for ε = 3)



"*Differentially private learning needs better features*", ICLR 2021 *spotlight*

Surpassing handcrafted features with *more private data.*

(for ε = 3)



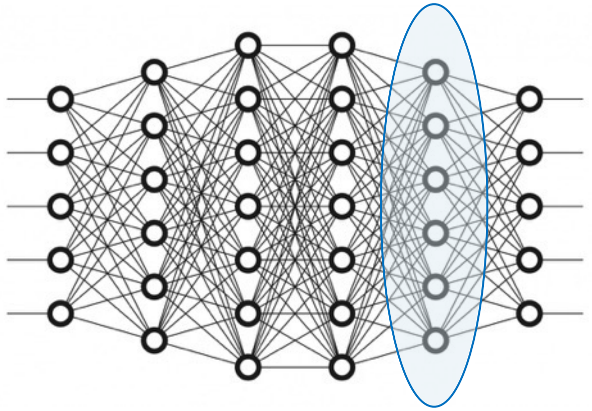
With **10x more private data** end-to-end deep learning performs best

collecting more data is good for your privacy!

"*Differentially private learning needs better features*", ICLR 2021 *spotlight*

Surpassing handcrafted features with *more **public** data.*

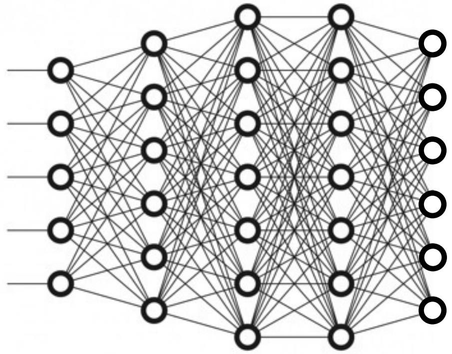
public data



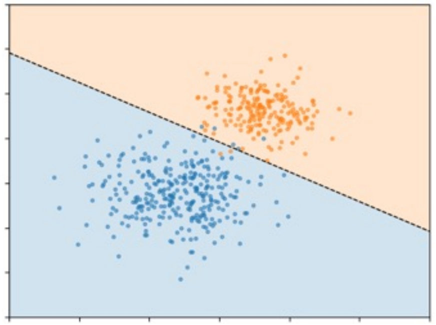
train a feature extractor on public data...

...transfer and fine-tune on private data

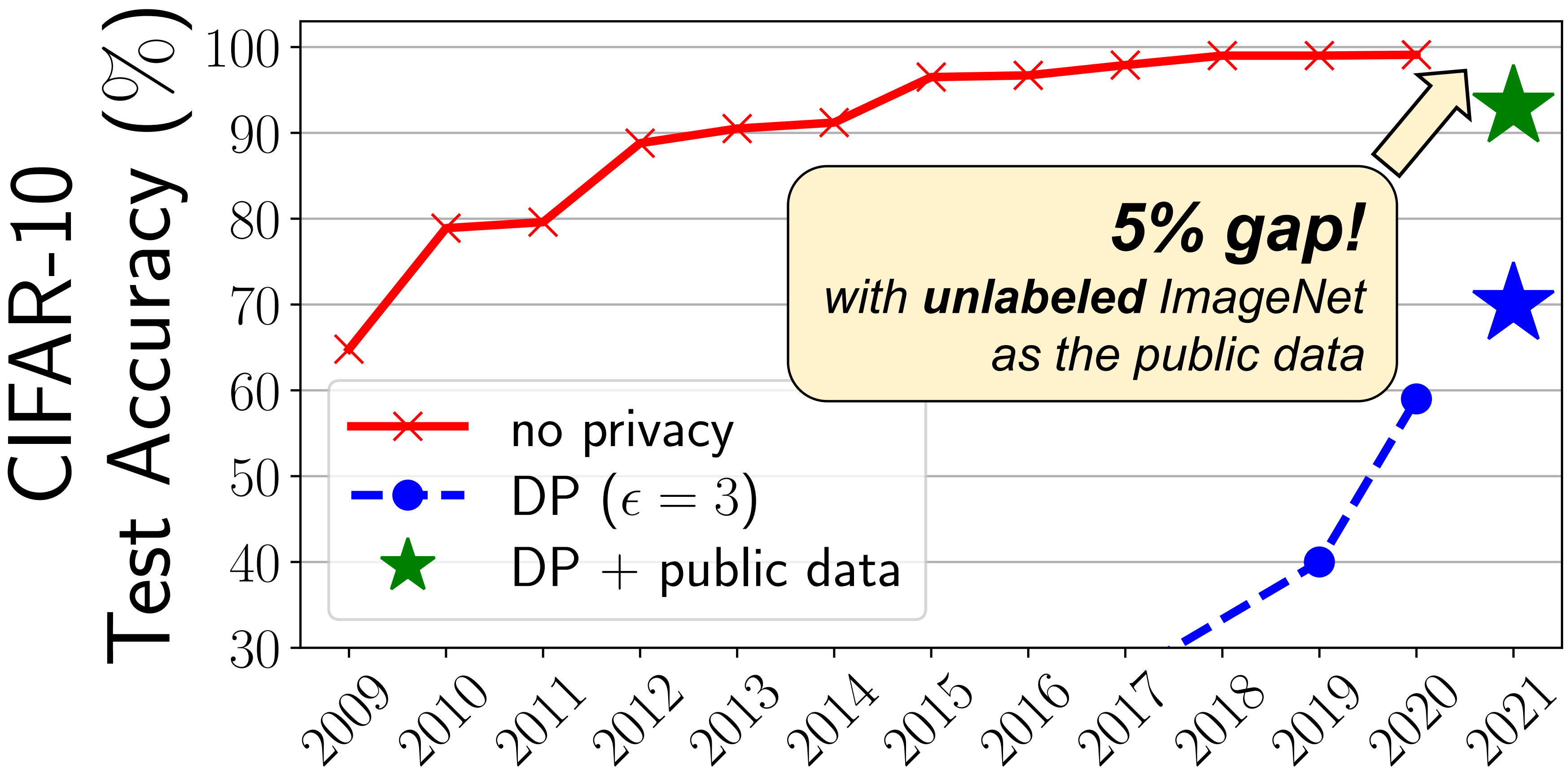
private data



privacy free



With access to a public dataset, privacy comes almost for free!



Talk outline.

- Adversarial examples for online content blockers
  - ➢ What's the threat model?
  - ➢ Limitations of current defenses
  - ➢ Industry impact
- Enhancing ML privacy
- Future work

Talk outline.

- Adversarial examples for online content blockers
  - ➢ What’s the threat model?
  - ➢ Limitations of current defenses
  - ➢ Industry impact
- Enhancing ML privacy
- Future work

Future work.

ML security is a critical challenge for our society.

how do we make ML trustworthy?

robustness

privacy

fairness

interpretability

Future work: robustness & privacy

Intersections:

- *Adversarial ML for safeguarding or breaching privacy* — **with Evani Radiya-Dixit**, **with Nicholas Carlini @ Google**

Scaling private ML:

- *Privacy in large NLP models* — **with Percy Liang**
- *Relaxing differential privacy* — **with Ilya Mironov @ Facebook**

Beyond machine learning:

- *Robustness & privacy in decentralized finance* — **with Ari Juels @ Cornell**, **with Kenny Paterson @ ETHZ**

Conclusion

ML is currently not ***trustworthy***.

- it is not ***robust***.
- it is not ***private***.

We can get better robustness than current ML.

➢ *humans are an existence proof.*

We can get better privacy than current ML.

➢ *with differential privacy and feature engineering.*

Conclusion

ML is currently not ***trustworthy***.

- it is not ***robust***.
- it is not ***private***.

We <u>can</u> get better robustness than current ML.

➢ *humans are an existence proof.*

We <u>can</u> get better privacy than current ML.

➢ *with differential privacy and feature engineering.*

Thank you!

Acknowledgments

Many! external collaborators (somewhat chronological since 2017)

Especially fruitful collaborations.

Ari Juels
Cornell Tech

Nicolas Papernot
University of Toronto

Nicholas Carlini
Google

beard

0% 100%

Stanford collaborators.

Giancarlo Pellegrino

Gili Rusak

Blanca Villanueva

Edward Chou

Evani Radiya-Dixit

Stanford’s amazing staff.

- Ruth Harris
- Megan Harris
- Jay Subramanian
- Jam Kiattinant

CS Department

- Rolando Villalobos

Bechtel International Center

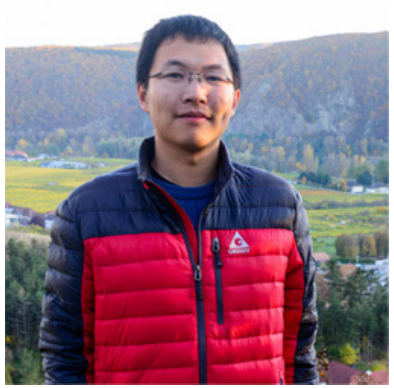
Switzerland

≠

Swaziland

The crypto group, past & present.

The CS-355 staff + *students*!



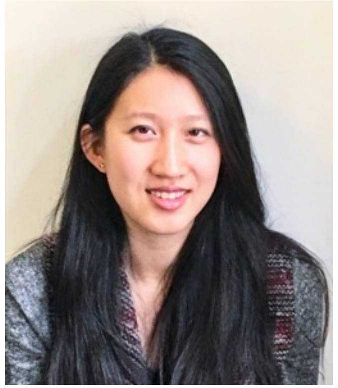
David Wu

Henry Corrigan-Gibbs

Sam Kim

Dima Kogan

Saba Eskandarian



Katy Woo

Amazing and infinite sources of advice.

Jean-Pierre Hubaux
EPFL

Ari Juels
Cornell Tech

Nicolas Papernot
University of Toronto

Kenny Paterson
ETHZ

Henry Corrigan-Gibbs
MIT

Giancarlo Pellegrino
CISPA

Ludwig Schmidt
University of Washington

Committee members.

Mykel Kochenderfer

Moses Charikar

Percy Liang

Gregory Valiant

My advisor: Dan Boneh

Friends & Family


Helen & Tom

My parents & brothers

Mariël

Socially-distant lunch party.

- Meet at **noon – Escondido Village basketball courts** (in front of McFarland, next to Tennis courts)

- Food, drinks & fun

Conclusion

ML is currently not ***trustworthy***.

- it is not ***robust***.
- it is not ***private***.

We <u>can</u> get better robustness than current ML.

➢ *humans are an existence proof.*

We <u>can</u> get better privacy than current ML.

➢ *with differential privacy and feature engineering.*

Thank you!