Does Adversarial Machine Learning Research Matter?

Florian Tramèr
Stanford University

Attacking ML models is popular.

Evasion

Intriguing properties of neural networks

C Szegedy, W Zaremba, I Sutskever, J Bruna… - arXiv preprint arXiv …, 2013 - arxiv.org

Deep neural networks are highly expressive models that have recently achieved state of the art performance on speech and visual recognition tasks. While their expressiveness is the reason they succeed, it also causes them to learn uninterpretable solutions that could have

Cited by 7614 Related articles All 20 versions

Poisoning

Poisoning attacks against support vector machines

B Biggio, B Nelson, P Laskov - arXiv preprint arXiv:1206.6389, 2012 - arxiv.org

We investigate a family of poisoning attacks against Support Vector Machines (SVM). Such attacks inject specially crafted training data that increases the SVM's test error. Central to the motivation for these attacks is the fact that most learning algorithms assume that their training data comes from a natural or well-behaved distribution. However, this assumption does not generally hold in security-sensitive settings. As we demonstrate, an intelligent adversary can, to some extent, predict the change of the SVM's decision function due to …

Cited by 871 Related articles All 19 versions

Data Inference

Membership inference attacks against machine learning mo

R Shokri, M Stronati, C Song… - 2017 IEEE Symposium …, 2017 - ieeexplore.ie

We quantitatively investigate how machine learning models leak information abo individual data records on which they were trained. We focus on the basic **memb inference** attack: given a data record and black-box access to a model, determir

Cited by 1281 Related articles All 17 versions

Model Stealing

Stealing machine learning models via prediction apis

F Tramèr, F Zhang, A Juels, MK Reiter… - 25th {USENIX} Security …, 2016 - usenix.org

Machine learning (ML) models may be deemed confidential due to their sensitive training data, commercial value, or use in security applications. Increasingly often, confidential ML models are being deployed with publicly accessible query interfaces. ML-as-a-service ("predictive analytics") systems are an example: Some allow users to train models on potentially sensitive data and charge others for access on a pay-per-query basis.

Cited by 906 Related articles All 16 versions

My talk: "Does Adversarial ML Research Matter?"

Betteridge law of headlines: ***No***

Intentionally a little controversial!

- We've done great research so far ☺
- Attacks gives us a sense of what bad things could happen
- But we could & should do a lot more for "real" security!

A blueprint for cool security attack research:
(in my opinion)

1. Take something "real" that many people use (or will use)
2. Show how to break it
3. Ideally, show how to redesign it in safer way

MELTDOWN

SPECTRE

National Security

Johns Hopkins researchers poke a hole in Apple's encryption

SHA-1 is a Shambles

The Geometry of Innocent Flesh on the Bone: Return-into-libc without Function Calls (on the x86)

Hovav Shacham*
hovav@cs.ucsd.edu

How not to prove your election outcome

Thomas Haines*, Sarah Jamie L… …ier Pereira‡, and Vanessa Teague§
*Norwegian U… …nd Technology
†Oper… …Canada
‡UCLouvain – … …-Neuve, Belgium
§The University of Melbourne – S… …ation Systems, Melbourne, Australia

Experimental Security Analysis of a Modern Automobile

Karl Koscher, Alexei Czeskis, Franziska Roesner, …
Department of Computer Scienc…
University of Washi…

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

FORESHADOW

Hey, You, Get Off of My Cloud: Exploring Information Leakage in Third-Party Compute Clouds

…s Ristenpart* Eran Tromer† Hovav Shacham* Stefan Savage*

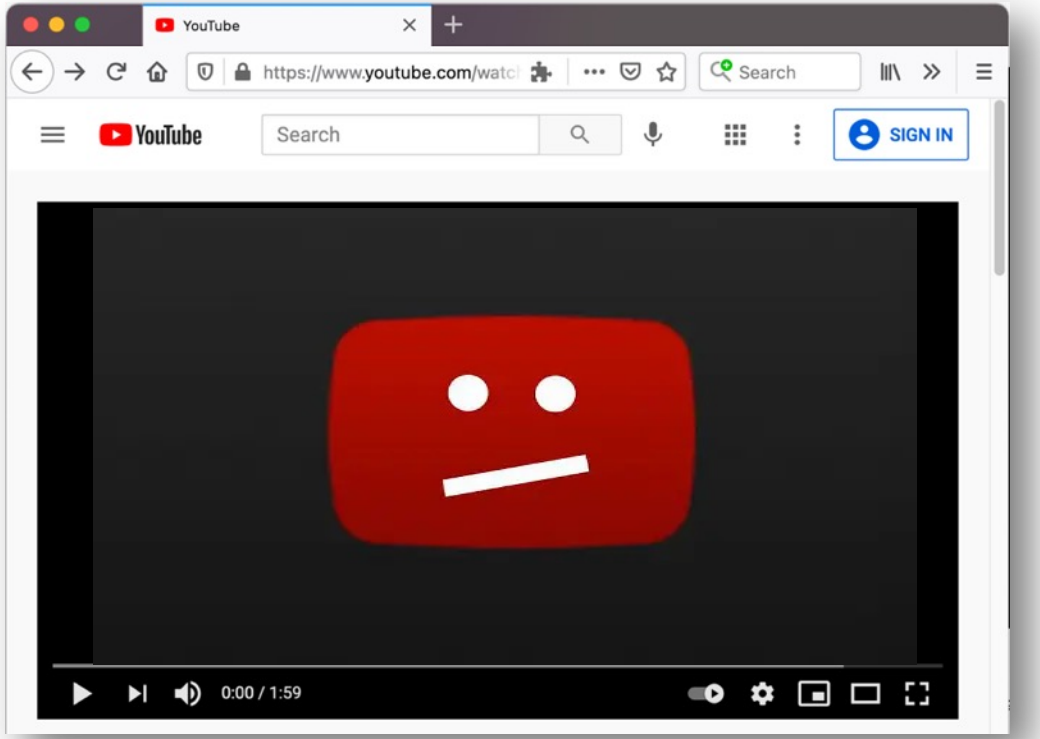
Where are the “real” attacks on ML?

1. **Take something “real” that many people use (or will use)**
2. Show how to break it
3. Ideally, show how to redesign it in safer way

Can we *evade* a real security model?

Cloud Video Intelligence API > Documentation > Guides

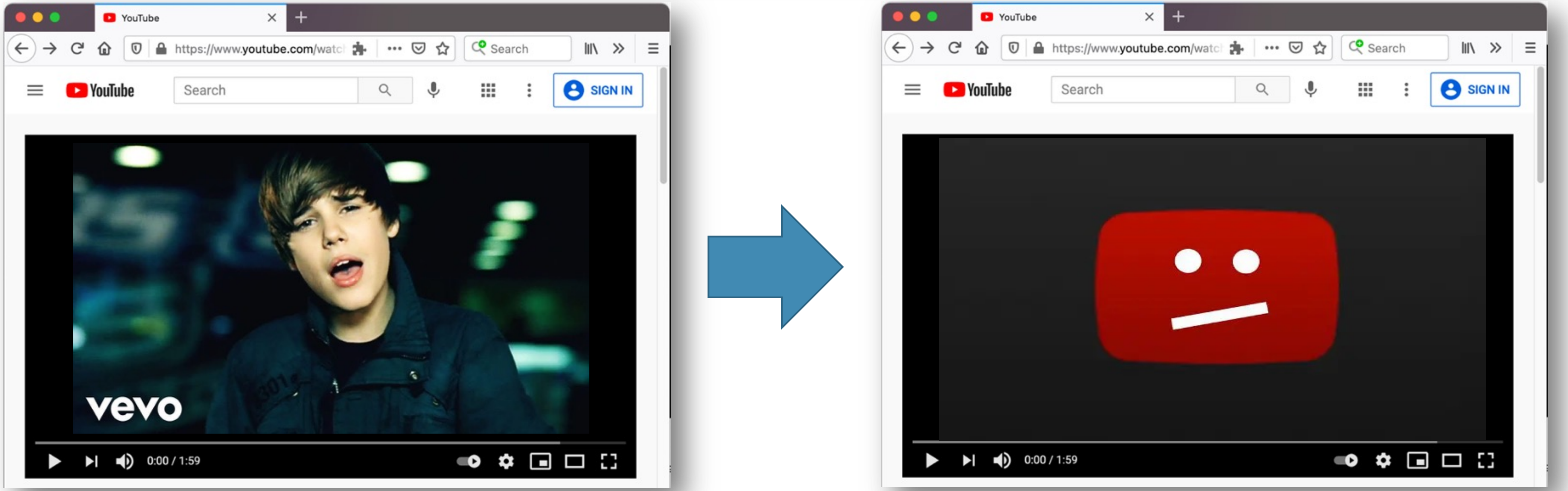
Detect explicit content in videos



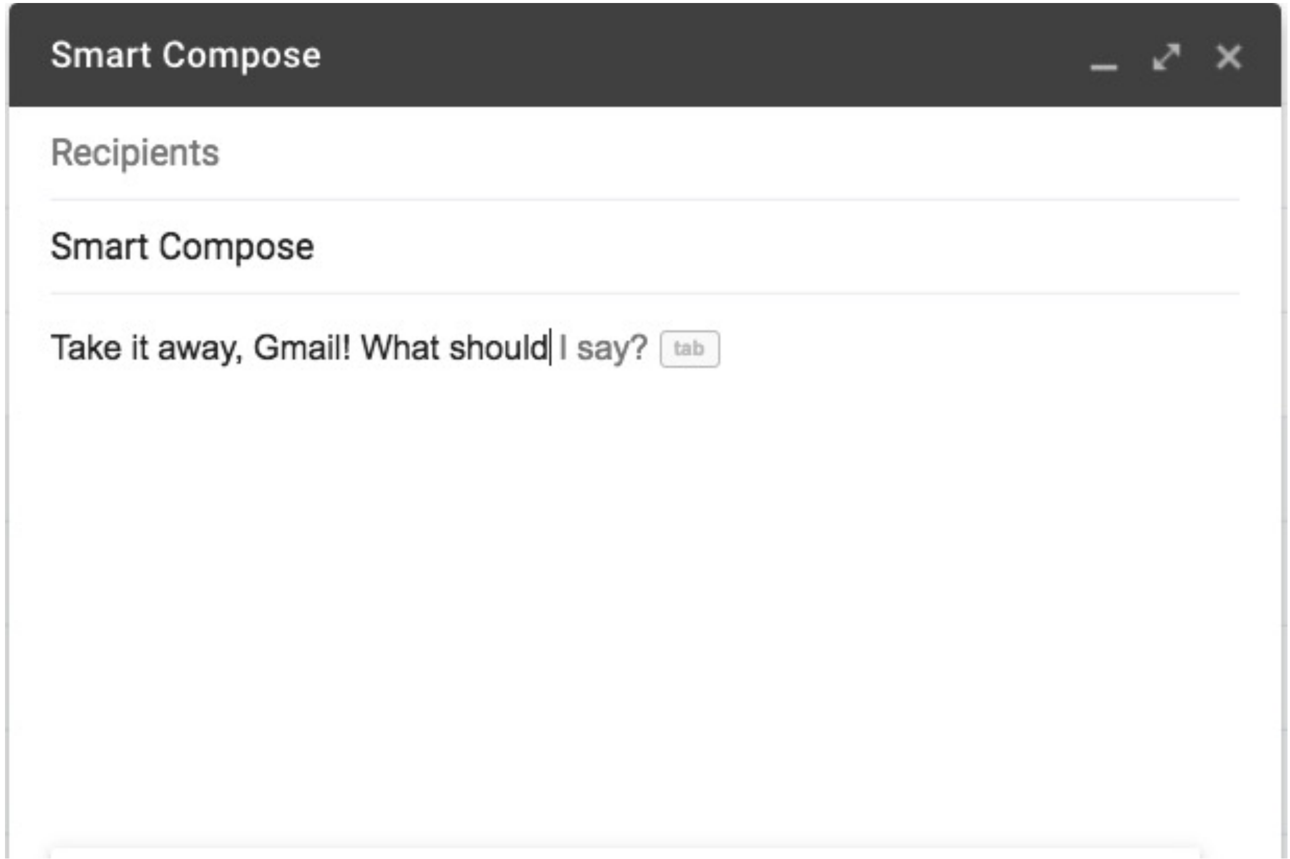
Can we *poison* a real security model?

Cloud Video Intelligence API > Documentation > Guides

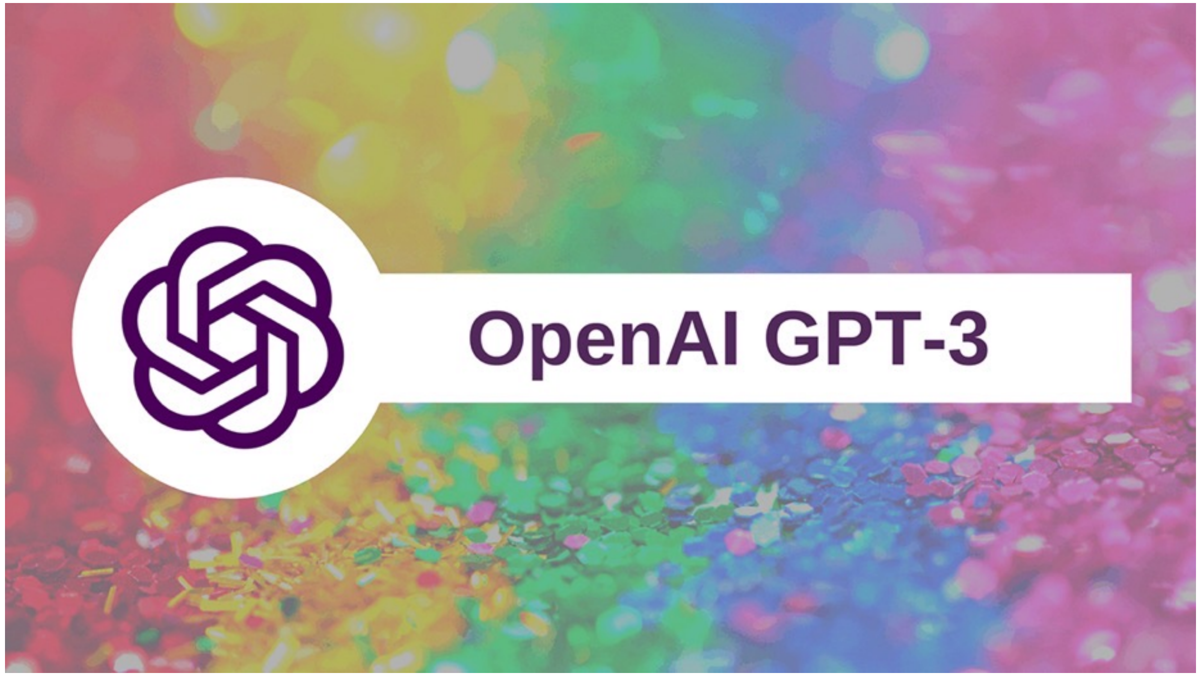
Detect explicit content in videos



Can we *extract* real user data?



Can we *steal* a real model?



Attacking “real” things matters!

Current attacks are not well suited for attacking “real” ML models.

- Maybe we’re making a fuss for nothing?
- Maybe real attacks work with enough tricks?
- Maybe we can design pragmatic defenses?

Evasion

Intriguing properties of **neural networks**
C Szegedy, W Zaremba, I Sutskever, J Bruna… - arXiv preprint arXiv …, 2013 - arxiv.org
Deep neural networks are highly expressive models that have recently achieved state of the art performance on speech and visual recognition tasks. While their expressiveness is the reason they succeed, it also causes them to learn uninterpretable solutions that could have
Cited by 7614 Related articles All 20 versions

Poisoning

Poisoning attacks against support vector machines
B Biggio, B Nelson, P Laskov - arXiv preprint arXiv:1206.6389, 2012 - arxiv.org
We investigate a family of poisoning attacks against Support Vector Machines (SVM). Such attacks inject specially crafted training data that increases the SVM's test error. Central to the motivation for these attacks is the fact that most learning algorithms assume that their training data comes from a natural or well-behaved distribution. However, this assumption does not generally hold in security-sensitive settings. As we demonstrate, an intelligent adversary can, to some extent, predict the change of the SVM's decision function due to …
Cited by 871 Related articles All 19 versions

Data Inference

Membership inference attacks against machine learning mo
R Shokri, M Stronati, C Song… - 2017 IEEE Symposium …, 2017 - ieeexplore.ie
We quantitatively investigate how machine learning models leak information abo individual data records on which they were trained. We focus on the basic **memb** **inference** attack: given a data record and black-box access to a model, determir
Cited by 1281 Related articles All 17 versions

Model Stealing

Stealing machine learning models via prediction apis
F Tramèr, F Zhang, A Juels, MK Reiter… - 25th {USENIX} Security …, 2016 - usenix.org
Machine learning (ML) models may be deemed confidential due to their sensitive training data, commercial value, or use in security applications. Increasingly often, confidential ML models are being deployed with publicly accessible query interfaces. ML-as-a-service ("predictive analytics") systems are an example: Some allow users to train models on potentially sensitive data and charge others for access on a pay-per-query basis.
Cited by 906 Related articles All 16 versions

Evasion

Intriguing properties of neural networks
C Szegedy, W Zaremba, I Sutskever, J Bruna… - arXiv preprint arXiv …, 2013 - arxiv.org
Deep neural networks are highly expressive models that have recently achieved state of the art performance on speech and visual recognition tasks. While their expressiveness is the reason they succeed, it also causes them to learn uninterpretable solutions that could have
Cited by 7614 Related articles All 20 versions

Poisoning

Poisoning attacks against support vector machines
B Biggio, B Nelson, P Laskov - arXiv preprint arXiv:1206.6389, 2012 - arxiv.org
We investigate a family of poisoning attacks against Support Vector Machines (SVM). Such attacks inject specially crafted training data that increases the SVM's test error. Central to the motivation for these attacks is the fact that most learning algorithms assume that their training data comes from a natural or well-behaved distribution. However, this assumption does not generally hold in security-sensitive settings. As we demonstrate, an intelligent adversary can, to some extent, predict the change of the SVM's decision function due to …
Cited by 871 Related articles All 19 versions

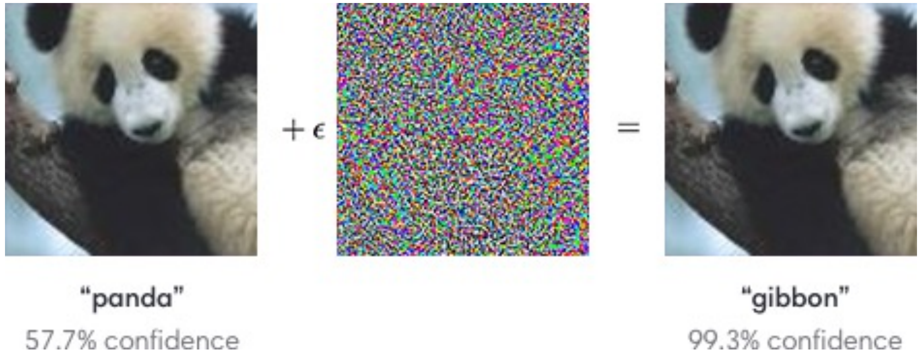
Data Inference

Membership inference attacks against machine learning mo
R Shokri, M Stronati, C Song… - 2017 IEEE Symposium …, 2017 - ieeexplore.ie
We quantitatively investigate how machine learning models leak information abo individual data records on which they were trained. We focus on the basic **memb inference** attack: given a data record and black-box access to a model, determir
Cited by 1281 Related articles All 17 versions

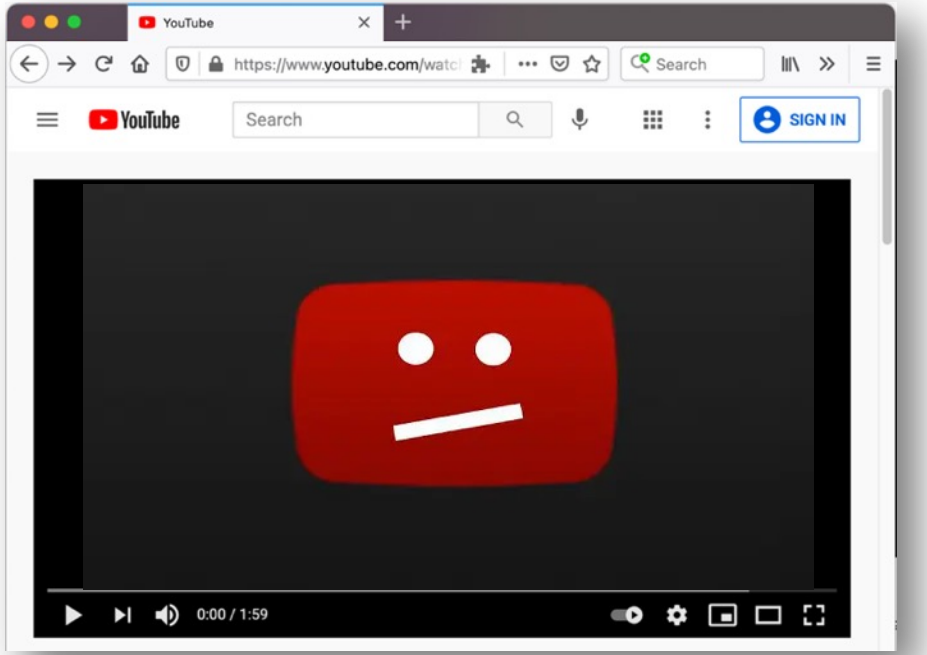
Model Stealing

Stealing machine learning models via prediction apis
F Tramèr, F Zhang, A Juels, MK Reiter… - 25th {USENIX} Security …, 2016 - usenix.org
Machine learning (ML) models may be deemed confidential due to their sensitive training data, commercial value, or use in security applications. Increasingly often, confidential ML models are being deployed with publicly accessible query interfaces. ML-as-a-service ("predictive analytics") systems are an example: Some allow users to train models on potentially sensitive data and charge others for access on a pay-per-query basis.
Cited by 906 Related articles All 16 versions

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 w.o. query-access

Real systems: >99% black-box w.o. query-access

attacks need not be imperceptible

A “real” white-box imperceptible attack: ad-blocking

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

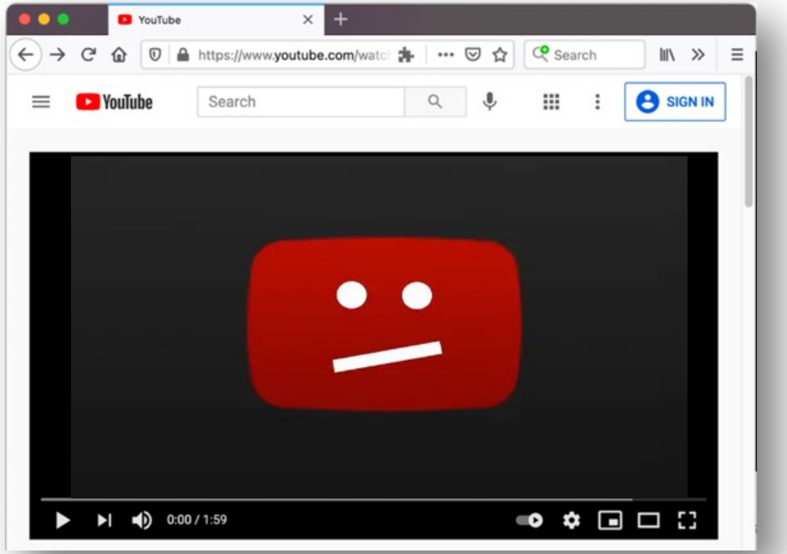
A “real” white-box imperceptible attack: ad-block

1. **Take something “real” that many people ~~use (or will use)~~ *might possibly use one day***

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

Most 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 positive queries...*
- *you likely can't build a good surrogate model...*

Many research opportunities!

Show how to systematically evade a real model

Stealthy Porn: Understanding Real-World Adversarial Images for Illicit Online Promotion

Kan Yuan*, Di Tang†, Xiaojing Liao*, XiaoFeng Wang*,
Xuan Feng*‡, Yi Chen*‡, Menghan Sun†, Haoran Lu*, Kehuan Zhang†
*Indiana University Bloomington †Chinese University of Hong Kong ‡Chinese Academy of Sciences

https://github.com/mitre/advmlthreatmatrix

Sign up

mitre / advmlthreatmatrix

Adversarial Threat Matrix

Adversarial Attacks on Copyright Detection Systems

Parsa Saadatpanah [1] Ali Shafahi [1] Tom Goldstein [1]

https://mlsec.io

MLSEC

Machine Learning Security Evasion Competition

Many research opportunities!

Show how to defend a real model

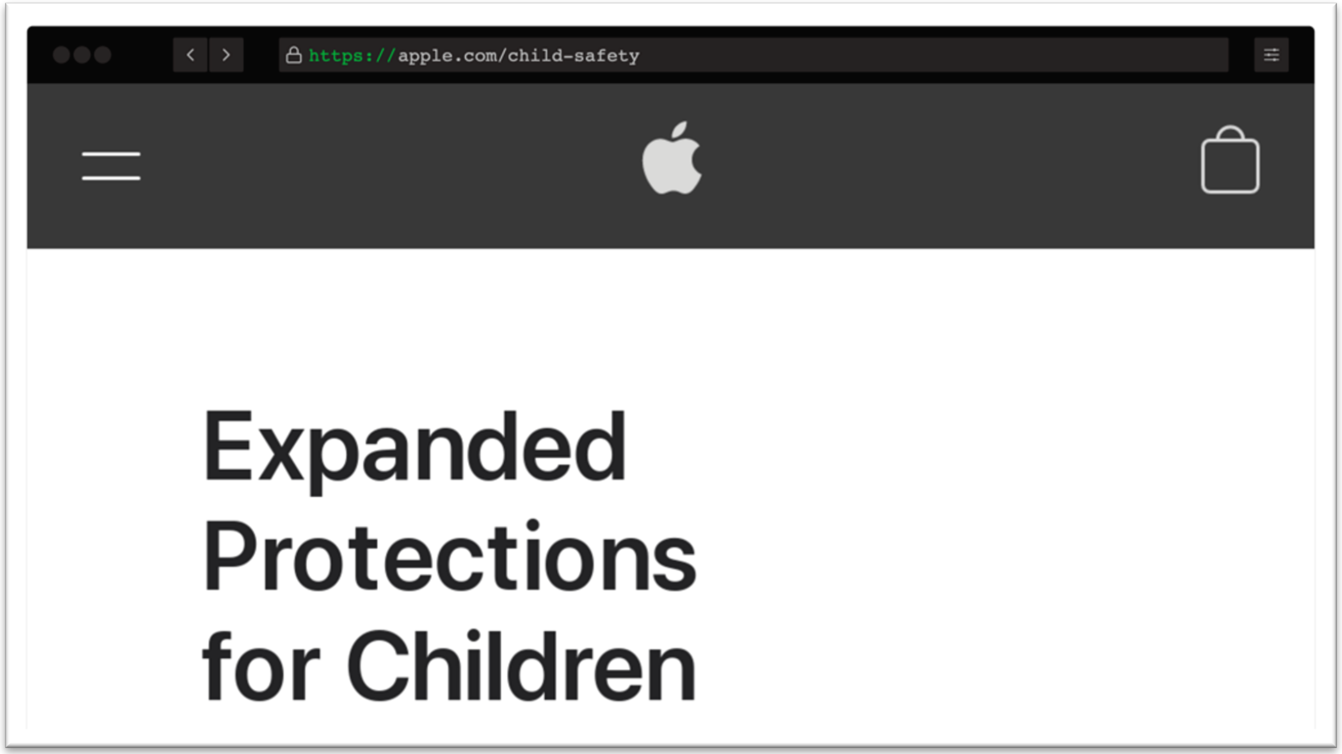
adversarial training,

interval-bound propagation,

randomized smoothing, *etc.*

are likely not the answer!

Very recent example: Apple's CSAM detection



- ➢ Uses ML to assign a "fingerprint / hash" to images
- ➢ Goal: hash is robust to small changes, few collisions

Very recent example: Apple's CSAM detection

what would we say?

- Apple's hashing algorithm is likely not robust
- Does that necessarily mean there's a practical attack?

Evasion

Intriguing properties of neural networks
C Szegedy, W Zaremba, I Sutskever, J Bruna… - arXiv preprint arXiv …, 2013 - arxiv.org
Deep neural networks are highly expressive models that have recently achieved state of the art performance on speech and visual recognition tasks. While their expressiveness is the reason they succeed, it also causes them to learn uninterpretable solutions that could have
Cited by 7614 Related articles All 20 versions

Poisoning

Poisoning attacks against support vector machines
B Biggio, B Nelson, P Laskov - arXiv preprint arXiv:1206.6389, 2012 - arxiv.org
We investigate a family of poisoning attacks against Support Vector Machines (SVM). Such attacks inject specially crafted training data that increases the SVM's test error. Central to the motivation for these attacks is the fact that most learning algorithms assume that their training data comes from a natural or well-behaved distribution. However, this assumption does not generally hold in security-sensitive settings. As we demonstrate, an intelligent adversary can, to some extent, predict the change of the SVM's decision function due to …
Cited by 871 Related articles All 19 versions

Data Inference

Membership inference attacks against machine learning mo
R Shokri, M Stronati, C Song… - 2017 IEEE Symposium …, 2017 - ieeexplore.ie
We quantitatively investigate how machine learning models leak information abo individual data records on which they were trained. We focus on the basic **memb inference** attack: given a data record and black-box access to a model, determi
Cited by 1281 Related articles All 17 versions

Model Stealing

Stealing machine learning models via prediction apis
F Tramèr, F Zhang, A Juels, MK Reiter… - 25th {USENIX} Security …, 2016 - usenix.org
Machine learning (ML) models may be deemed confidential due to their sensitive training data, commercial value, or use in security applications. Increasingly often, confidential ML models are being deployed with publicly accessible query interfaces. ML-as-a-service ("predictive analytics") systems are an example: Some allow users to train models on potentially sensitive data and charge others for access on a pay-per-query basis.
Cited by 906 Related articles All 16 versions

Why did no one poison GPT-X, Copilot, etc?

(as far as I know)

GPT-2

GitHub Copilot

Poisoning these models is possible.
(in principle)

Poisoning and Backdooring Contrastive Learning

Nicholas Carlini
Google

Andreas Terzis
Google

You Autocomplete Me:
Poisoning Vulnerabilities in Neural Code Completion*

Roei Schuster
Tel Aviv University
Cornell Tech

Congzheng Song
Cornell University

Eran Tromer
Tel Aviv University
Columbia University

Vitaly Shmatikov
Cornell Tech

Universal Adversarial Triggers for Attacking and Analyzing NLP

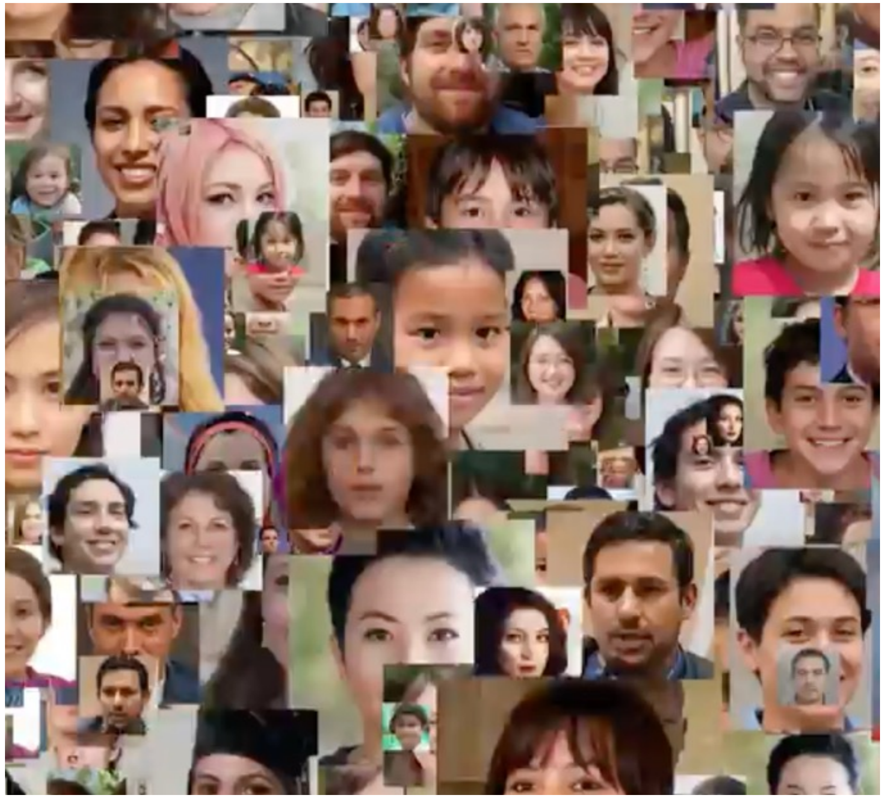
WARNING: This paper contains model outputs which are offensive in nature.

Eric Wallace[1], Shi Feng[2], Nikhil Kandpal[3],
Matt Gardner[1], Sameer Singh[4]

A **real** example: poisoning facial recognition models

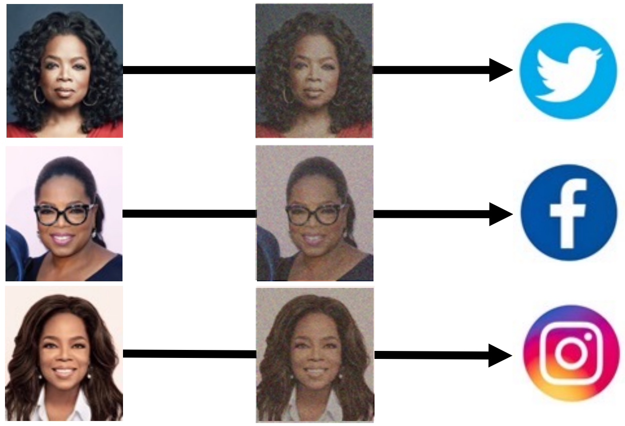
Clearview.ai

The Secretive Company That Might End Privacy as We Know It

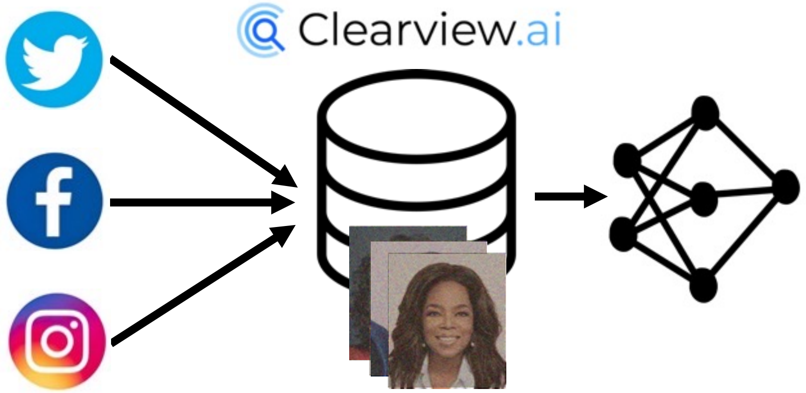


A **real** example: poisoning facial recognition models

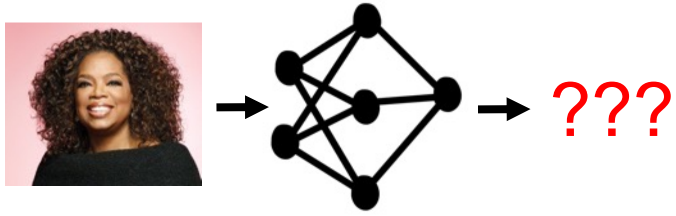
Users perturb pictures they post online



Online pictures are scraped to build a model



Unperturbed test pictures aren't recognized



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

“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

A **real** example: poisoning facial recognition models

The New York Times

This Tool Could Protect Your Photos From Facial Recognition

sandlab.cs.uchicago.edu/fawkes

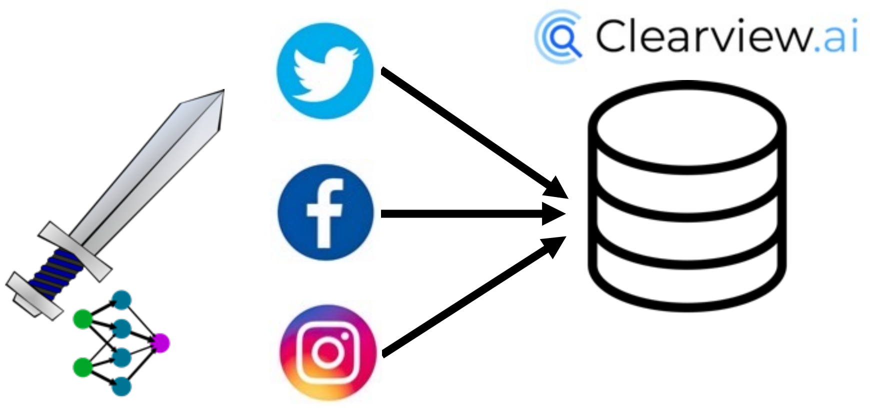
BSD-3-Clause License

4.1k stars 402 forks

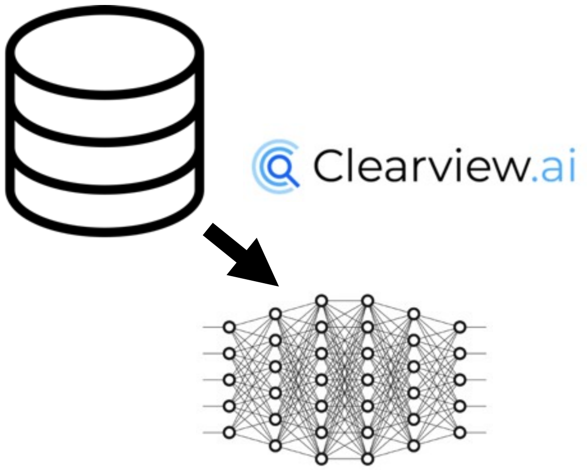
NEWS

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

The problem: retroactive defenses



wait one year

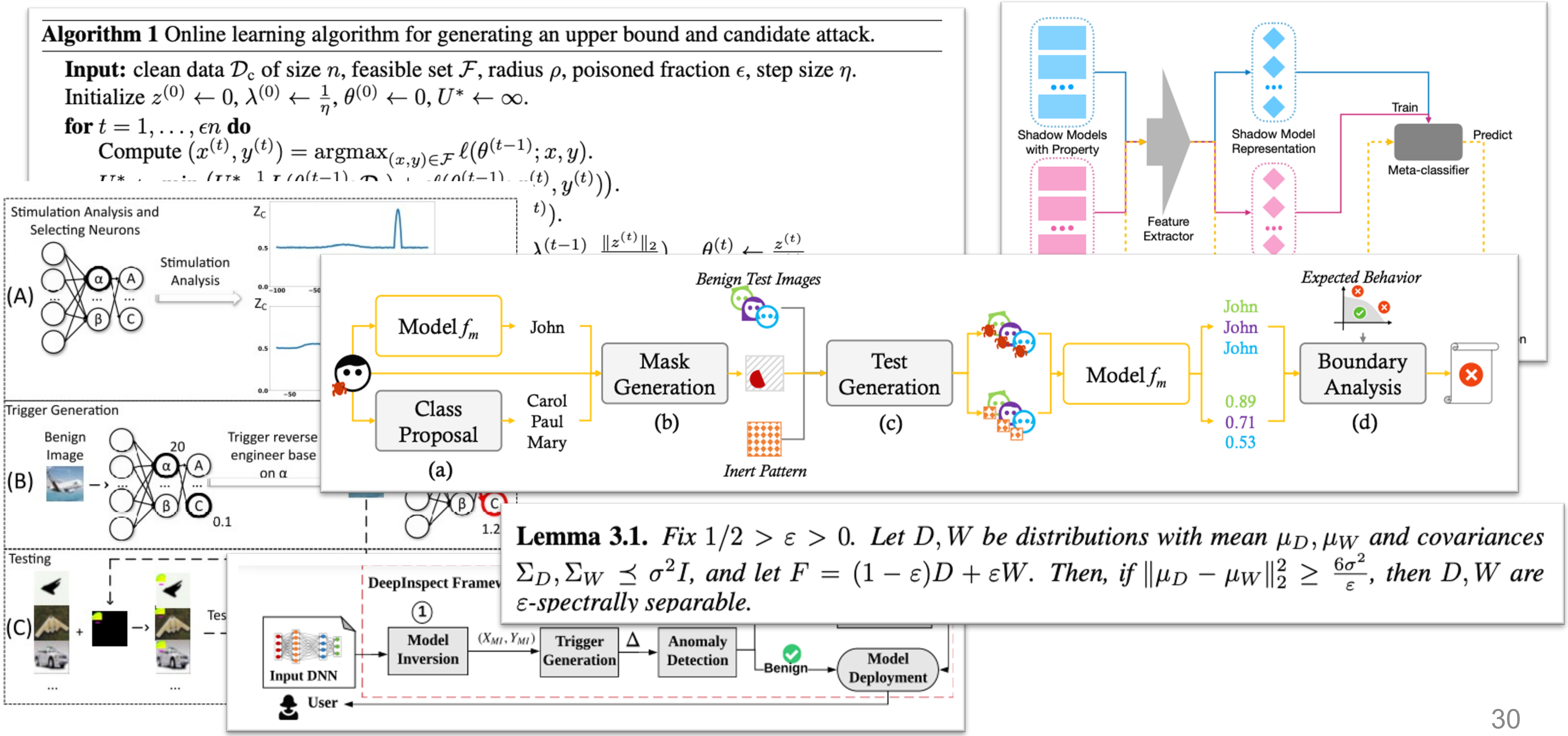


Facial recognition provider scrapes pictures produced with attacks that target today's models

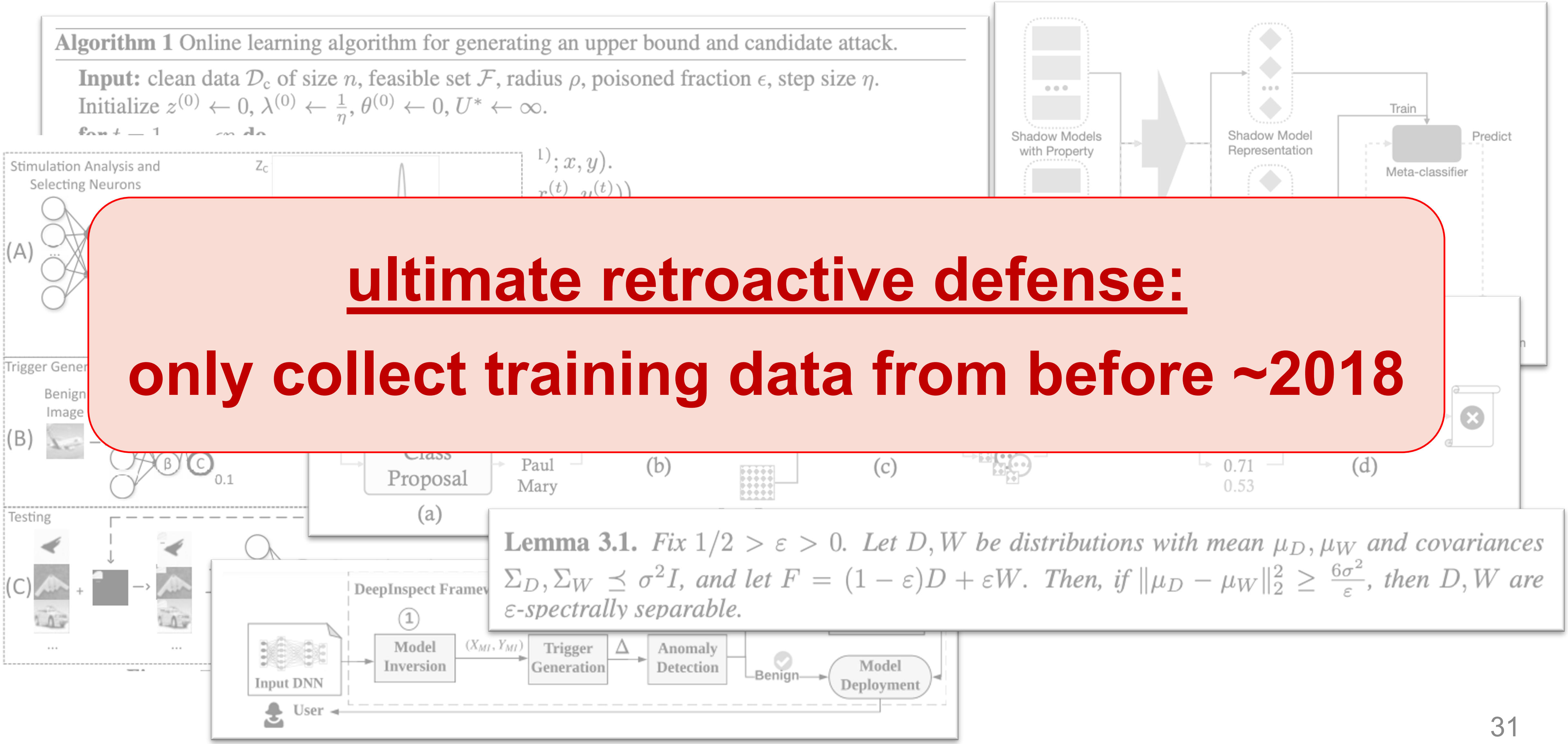
Facial recognition provider trains new SOTA model on poisoned data collected in the past

"Data Poisoning Won't Save You From Facial Recognition", https://arxiv.org/abs/2106.14851

Are poisoning defenses overkill?



Are poisoning defenses overkill?



<u>ultimate retroactive defense:</u>

only collect training data from before ~2018

Many research opportunities!

- Better threat modeling for real-world poisoning
- Robust attacks against real models
- Beyond “closed-world” defenses
  - dynamic defenses
  - leverage web-ranking methods to filter data?

Evasion

Intriguing properties of neural networks
C Szegedy, W Zaremba, I Sutskever, J Bruna… - arXiv preprint arXiv …, 2013 - arxiv.org
Deep neural networks are highly expressive models that have recently achieved state of the art performance on speech and visual recognition tasks. While their expressiveness is the reason they succeed, it also causes them to learn uninterpretable solutions that could have
Cited by 7614 Related articles All 20 versions

Poisoning

Poisoning attacks against support vector machines
B Biggio, B Nelson, P Laskov - arXiv preprint arXiv:1206.6389, 2012 - arxiv.org
We investigate a family of poisoning attacks against Support Vector Machines (SVM). Such attacks inject specially crafted training data that increases the SVM's test error. Central to the motivation for these attacks is the fact that most learning algorithms assume that their training data comes from a natural or well-behaved distribution. However, this assumption does not generally hold in security-sensitive settings. As we demonstrate, an intelligent adversary can, to some extent, predict the change of the SVM's decision function due to …
Cited by 871 Related articles All 19 versions

Data Inference

Membership inference attacks against machine learning mo
R Shokri, M Stronati, C Song… - 2017 IEEE Symposium …, 2017 - ieeexplore.ie
We quantitatively investigate how machine learning models leak information abo individual data records on which they were trained. We focus on the basic **memb inference** attack: given a data record and black-box access to a model, determir
Cited by 1281 Related articles All 17 versions

Model Stealing

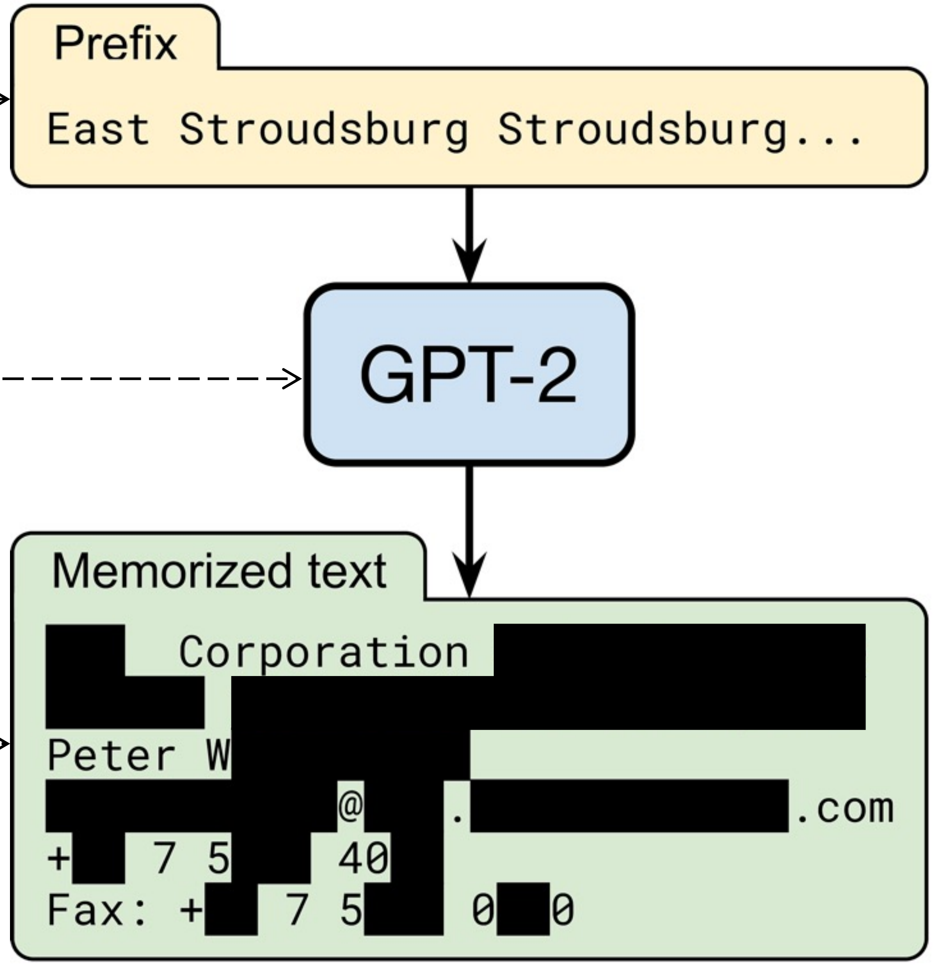
Stealing machine learning models via prediction apis
F Tramèr, F Zhang, A Juels, MK Reiter… - 25th {USENIX} Security …, 2016 - usenix.org
Machine learning (ML) models may be deemed confidential due to their sensitive training data, commercial value, or use in security applications. Increasingly often, confidential ML models are being deployed with publicly accessible query interfaces. ML-as-a-service ("predictive analytics") systems are an example: Some allow users to train models on potentially sensitive data and charge others for access on a pay-per-query basis.
Cited by 906 Related articles All 16 versions

Extracting public data from a “real” model.

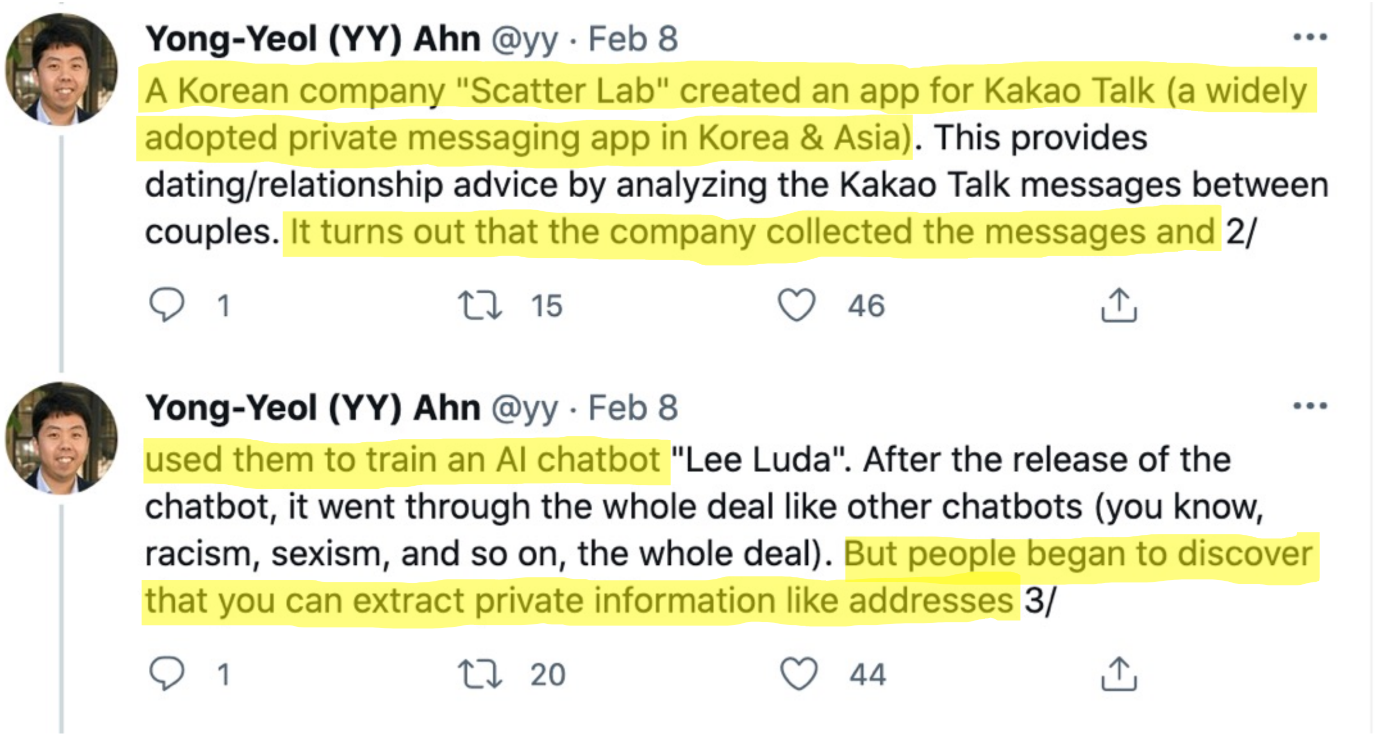
random input

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

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



Extracting private data from a real model.



Yong-Yeol (YY) Ahn @yy · Feb 8

A Korean company "Scatter Lab" created an app for Kakao Talk (a widely adopted private messaging app in Korea & Asia). This provides dating/relationship advice by analyzing the Kakao Talk messages between couples. It turns out that the company collected the messages and 2/

1 15 46

Yong-Yeol (YY) Ahn @yy · Feb 8

used them to train an AI chatbot "Lee Luda". After the release of the chatbot, it went through the whole deal like other chatbots (you know, racism, sexism, and so on, the whole deal). But people began to discover that you can extract private information like addresses 3/

1 20 44

Many research opportunities!

- extraction of "real" user data?
- extraction of non-text data?
  - images?
  - speech?
  - etc.
- more pragmatic defenses than differential privacy?
  - data de-duplication & filtering?
  - detecting data extraction at test time?

Evasion

Intriguing properties of **neural networks**
C Szegedy, W Zaremba, I Sutskever, J Bruna… - arXiv preprint arXiv …, 2013 - arxiv.org
Deep neural networks are highly expressive models that have recently achieved state of the art performance on speech and visual recognition tasks. While their expressiveness is the reason they succeed, it also causes them to learn uninterpretable solutions that could have
Cited by 7614 Related articles All 20 versions

Poisoning

Poisoning attacks against support vector machines
B Biggio, B Nelson, P Laskov - arXiv preprint arXiv:1206.6389, 2012 - arxiv.org
We investigate a family of poisoning attacks against Support Vector Machines (SVM). Such attacks inject specially crafted training data that increases the SVM's test error. Central to the motivation for these attacks is the fact that most learning algorithms assume that their training data comes from a natural or well-behaved distribution. However, this assumption does not generally hold in security-sensitive settings. As we demonstrate, an intelligent adversary can, to some extent, predict the change of the SVM's decision function due to …
Cited by 871 Related articles All 19 versions

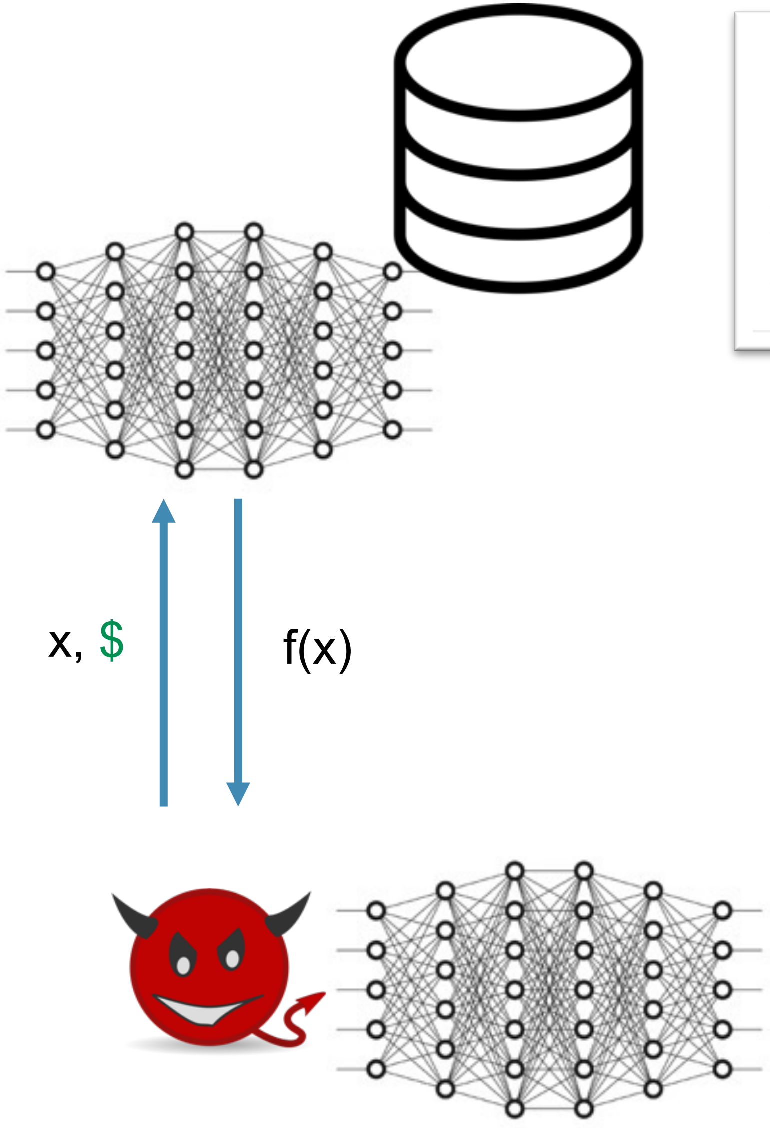
Data Inference

Membership inference attacks against machine learning mo
R Shokri, M Stronati, C Song… - 2017 IEEE Symposium …, 2017 - ieeexplore.ie
We quantitatively investigate how machine learning models leak information abo individual data records on which they were trained. We focus on the basic **memb** **inference** attack: given a data record and black-box access to a model, determi
Cited by 1281 Related articles All 17 versions

Model Stealing

Stealing machine learning models via prediction apis
F Tramèr, F Zhang, A Juels, MK Reiter… - 25th {USENIX} Security …, 2016 - usenix.org
Machine learning (ML) models may be deemed confidential due to their sensitive training data, commercial value, or use in security applications. Increasingly often, confidential ML models are being deployed with publicly accessible query interfaces. ML-as-a-service ("predictive analytics") systems are an example: Some allow users to train models on potentially sensitive data and charge others for access on a pay-per-query basis.
Cited by 906 Related articles All 16 versions

Stealing a pay-per-use model.



costs by charging users for future predictions. A model extraction attack will undermine the provider's business model if a malicious user pays less for training and ex-

"Stealing Machine Learning Models via Prediction APIs"

Distilling the Knowledge in a Neural Network

Geoffrey Hinton[*†]
Google Inc.
Mountain View
geoffhinton@google.com

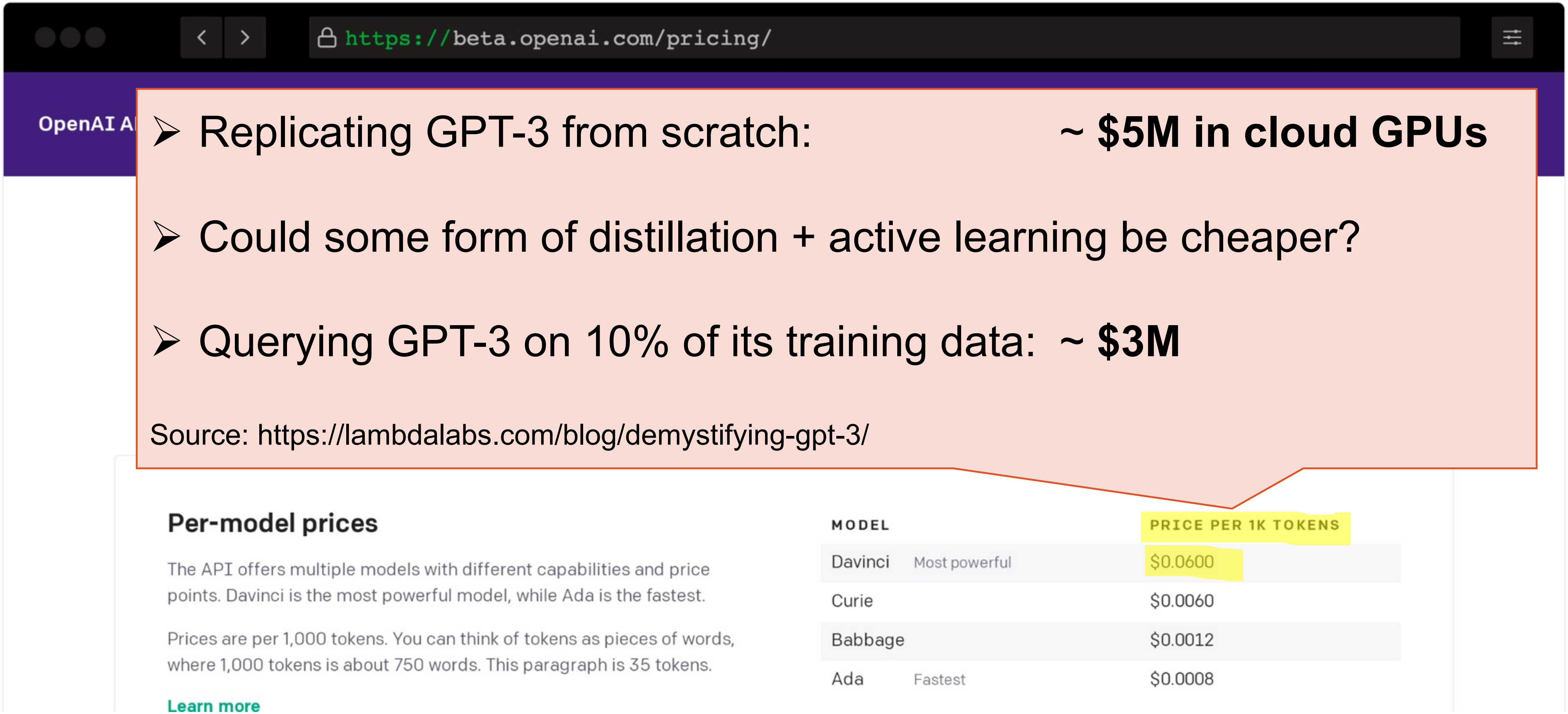
Oriol Vinyals[†]
Google Inc.
Mountain View
vinyals@google.com

Jeff Dean
Google Inc.
Mountain View
jeff@google.com

Could it be practical to steal GPT-3?

- Replicating GPT-3 from scratch: **~ $5M in cloud GPUs**
- Could some form of distillation + active learning be cheaper?
- Querying GPT-3 on 10% of its training data: **~ $3M**

Source: https://lambdalabs.com/blog/demystifying-gpt-3/



Many research opportunities!

- better model stealing in a research setting
- stealing a “real” model
- economics of extraction

Take-aways

We've written >10K papers on worst-case attacks

- We know: in principle, any model can be attacked
- We know: the strongest attacks are hard to prevent

What's next?

- We don't know: what do real attacks look like?
- We don't know: can we develop pragmatic defenses?