Fawkes: Protecting Privacy against Unauthorized Deep Learning Models

Shawn Shan†
Emily Wenger†
Jiayun Zhang
Huiying Li
Haitao Zheng
Ben Y. Zhao

† denotes co-first authors with equal contribution

Facial Recognition Models are Easy to Build

Less time to train larger/more powerful models

Facial Recognition Models are Easy to Build

Less time to train larger/more powerful models

Cheaper, faster hardware

Facial Recognition Models are Easy to Build

Less time to train larger/more powerful models

Cheaper, faster hardware

Labelled training data everywhere

Facial Recognition Models are Easy to Build

Less time to train larger/more powerful models

Cheaper, faster hardware

Labelled training data everywhere

- Anyone with limited coding knowledge and computational power can train powerful facial recognition models

Facial Recognition Models are Easy to Build

Less time to train larger/more powerful models

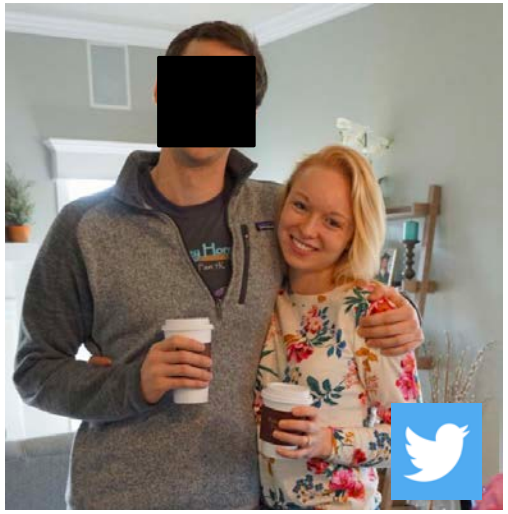
Cheaper, faster hardware

Labelled training data everywhere

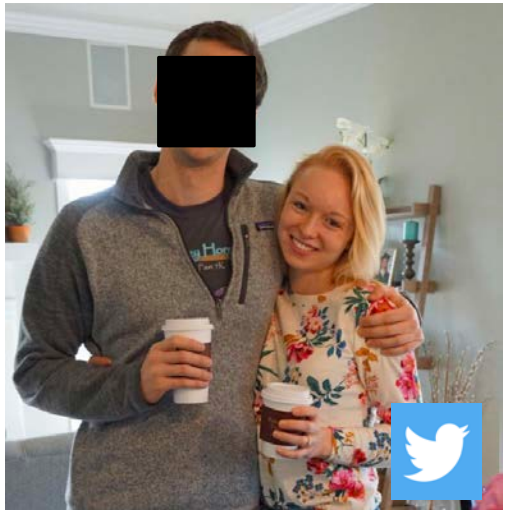
- Anyone with limited coding knowledge and computational power can train powerful facial recognition models

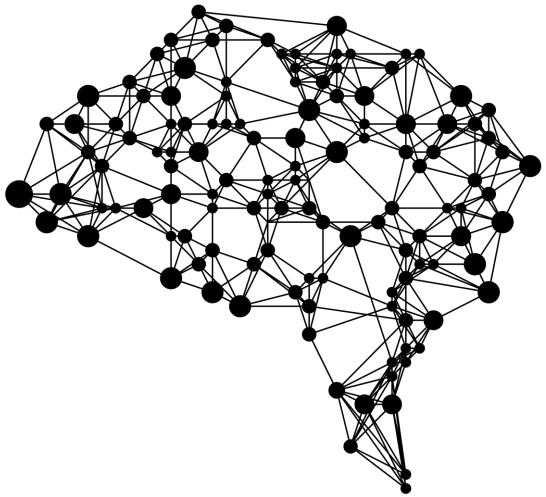
But what if the **wrong** people take advantage of this new accessibility?

Personal Images Co-opted to Train Facial Recognition Models



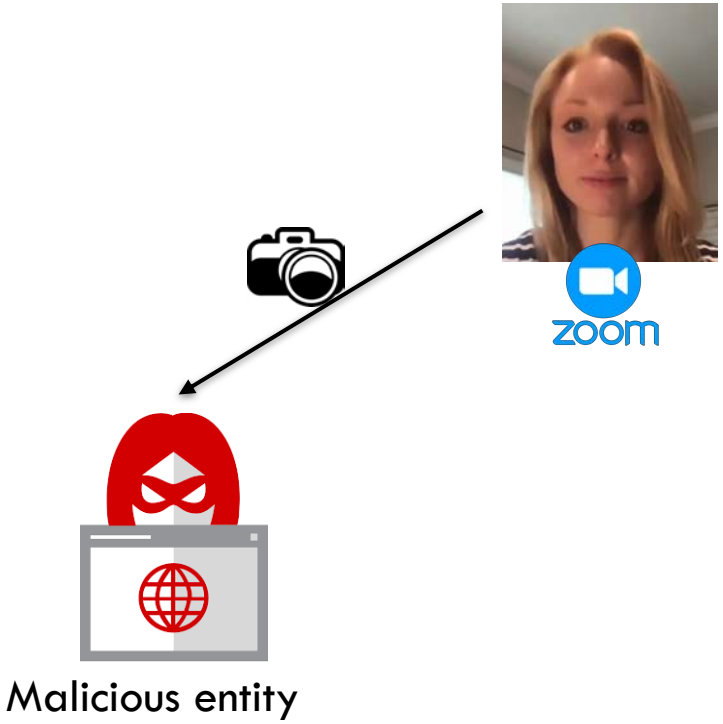
Personal Images Co-opted to Train Facial Recognition Models



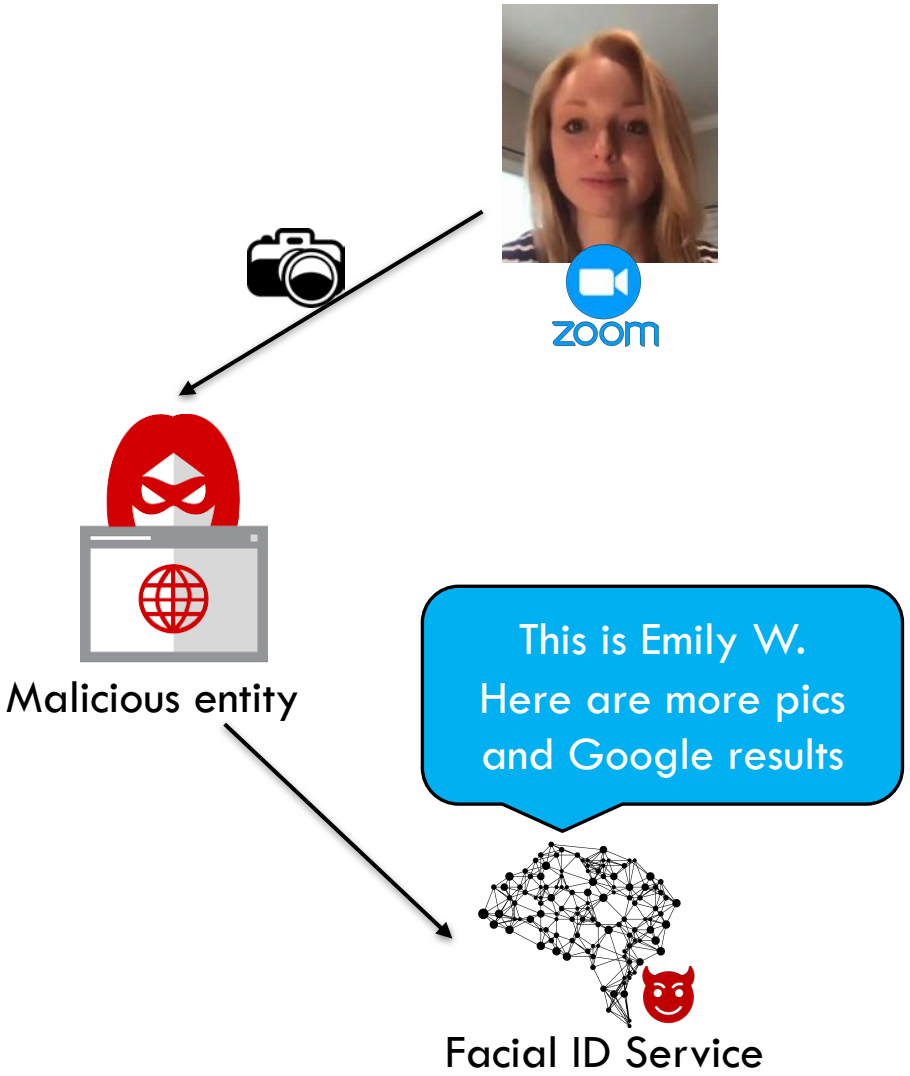


Facial recognition model that recognizes Emily

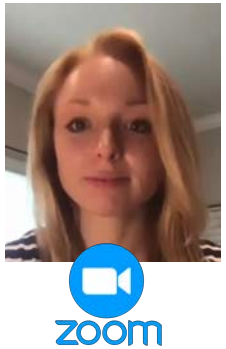
Facial Recognition Models Easily Misused



Facial Recognition Models Easily Misused

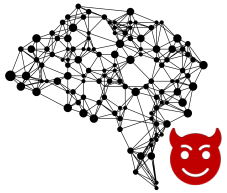


Facial Recognition Models Easily Misused



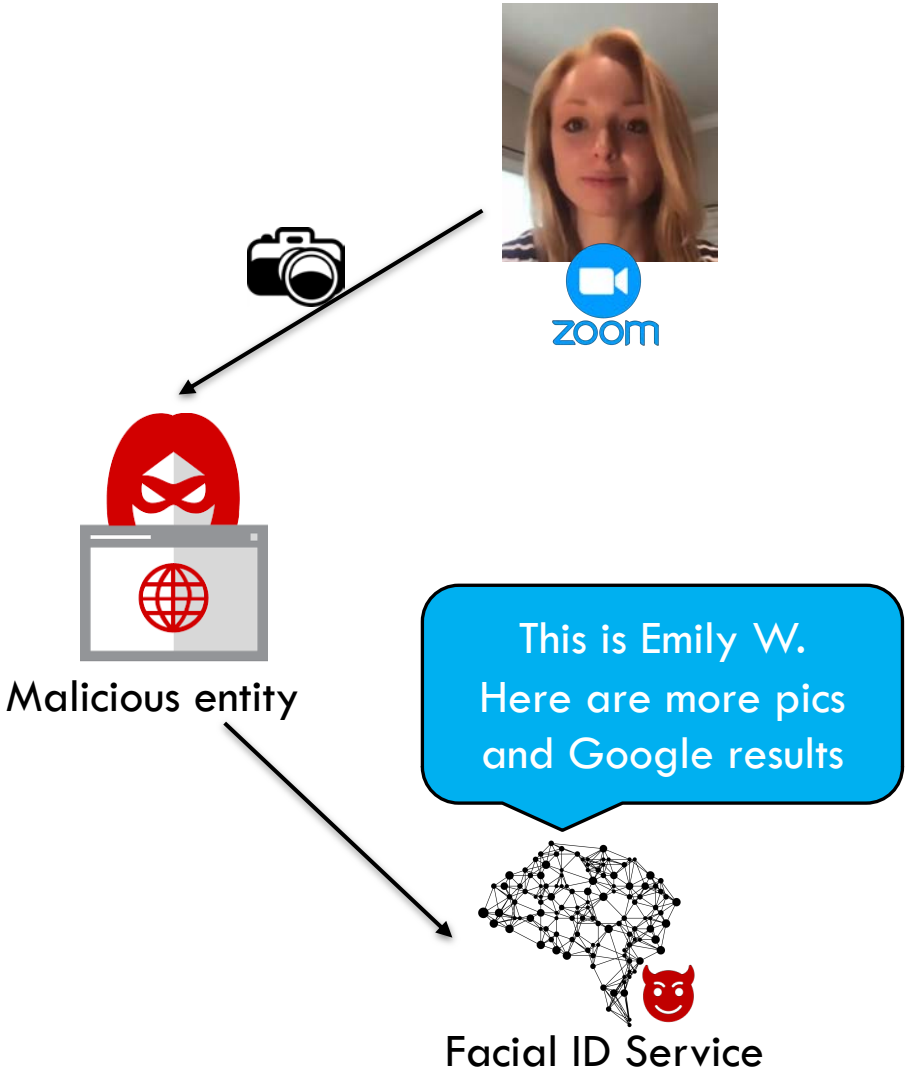
Malicious entity

This is Emily W.
Here are more pics
and Google results

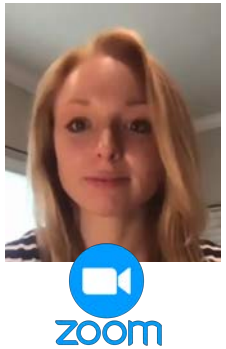


Facial ID Service

Facial Recognition Models Easily Misused

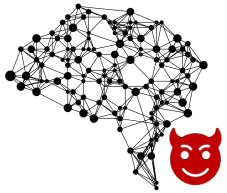


Facial Recognition Models Easily Misused



Malicious entity

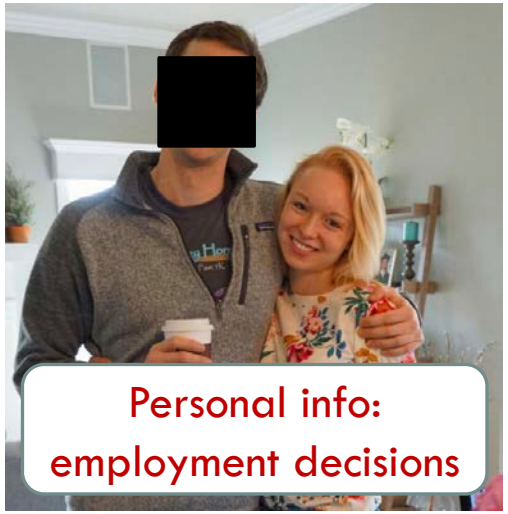
This is Emily W.
Here are more pics
and Google results



Facial ID Service

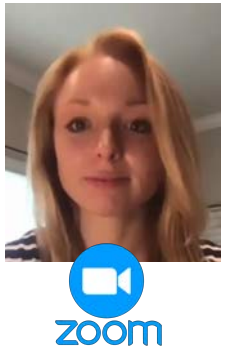
Location info: stalking

Family info: phishing/extortion



Personal info:
employment decisions

Facial Recognition Models Easily Misused



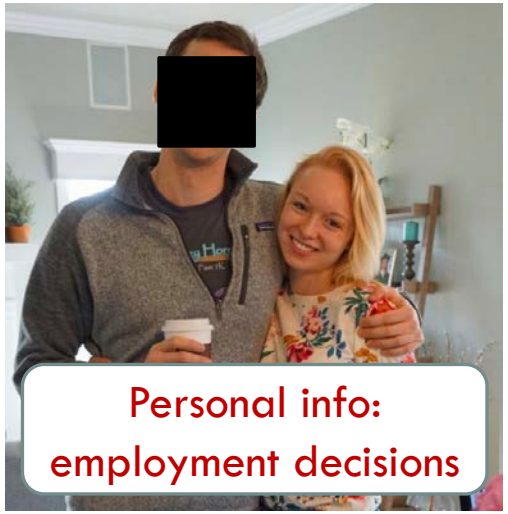
Malicious entity

This is Emily W.
Here are more pics
and Google results

Facial ID Service

Location info: stalking

Family info: phishing/extortion



Personal info:
employment decisions

Other info could lead to:
- Racial discrimination
- Political oppression
- Religious persecution
…

That Reality is Here, Today

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and "might lead to a dystopian future or something," a backer says.

…

That Reality is Here, Today

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and "might lead to a dystopian future or something," a backer says.

Clearview.AI

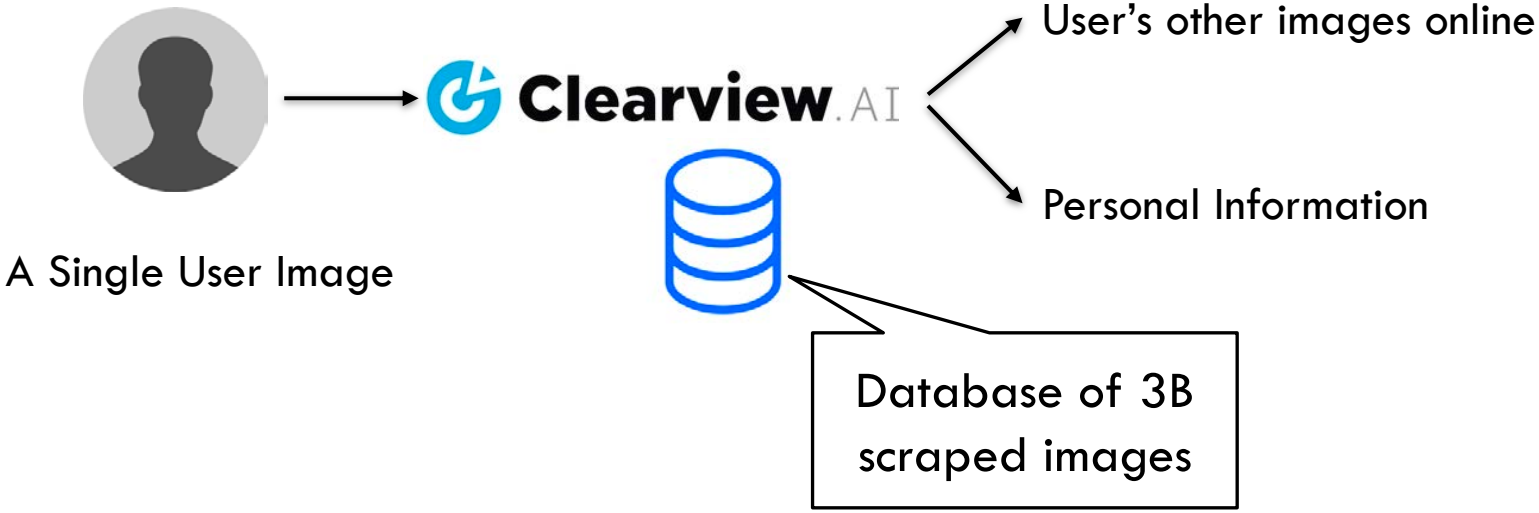
Database of 3B scraped images

…

That Reality is Here, Today

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and "might lead to a dystopian future or something," a backer says.

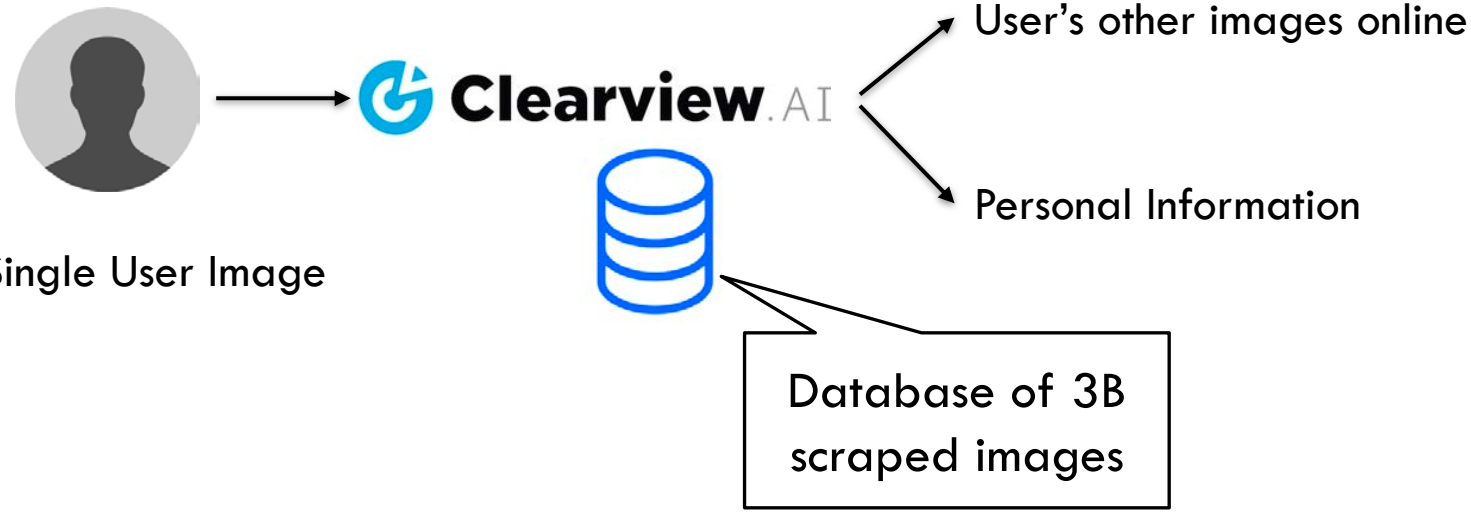


…

That Reality is Here, Today

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and "might lead to a dystopian future or something," a backer says.

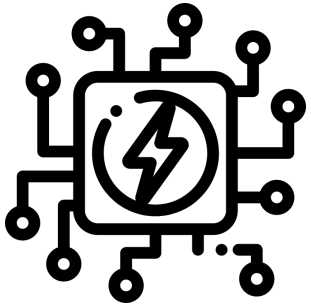


Known Clearview.ai customers include government agencies, law enforcement departments, and private citizens.

…

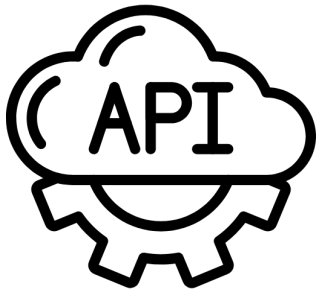
In This Talk

Fawkes:
Privacy armor that protects privacy by preventing your images from being used to train ML models against you.



Fawkes Design

Evaluation



Live Tests against Face Recognition Services

Goals and Assumptions

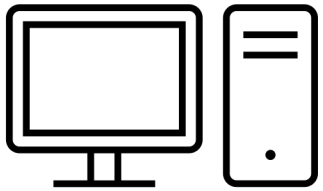
User

Tracker (e.g. Clearview)

Goals and Assumptions

User

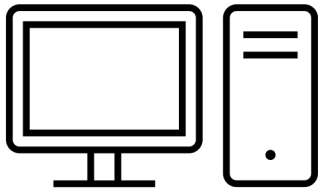
Tracker (e.g. Clearview)



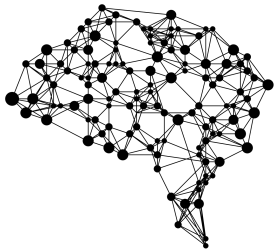
Limited computational resources

Goals and Assumptions

User



+



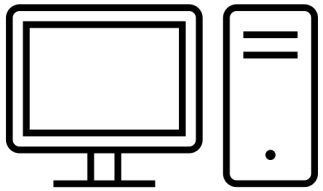
Limited computational resources

Well-trained feature extractor

Tracker (e.g. Clearview)

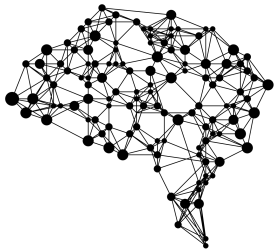
Goals and Assumptions

User

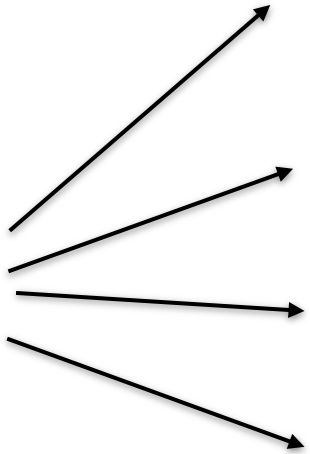


Limited computational resources

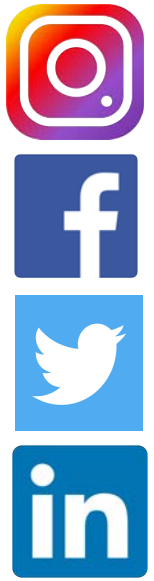
+



Well-trained feature extractor

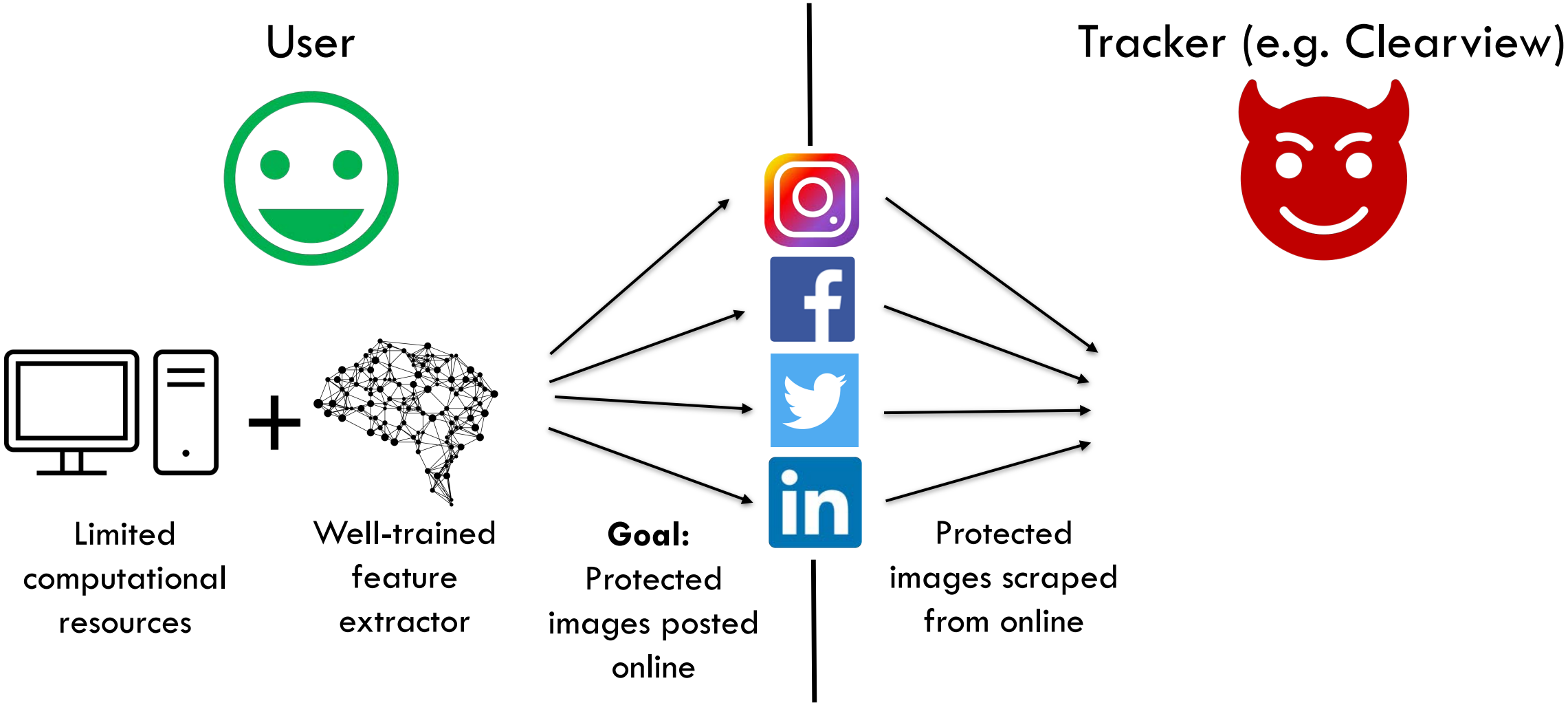


Goal: Protected images posted online

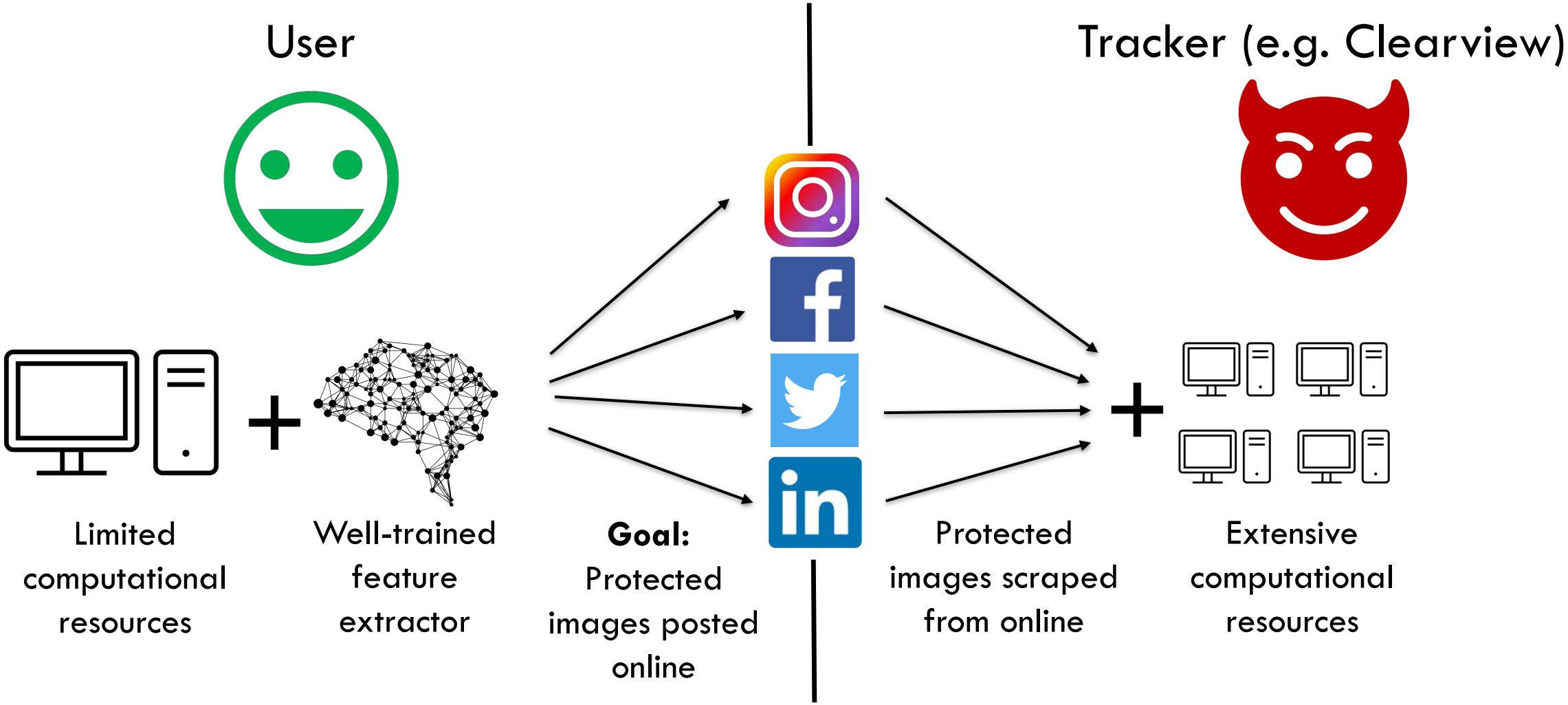


Tracker (e.g. Clearview)

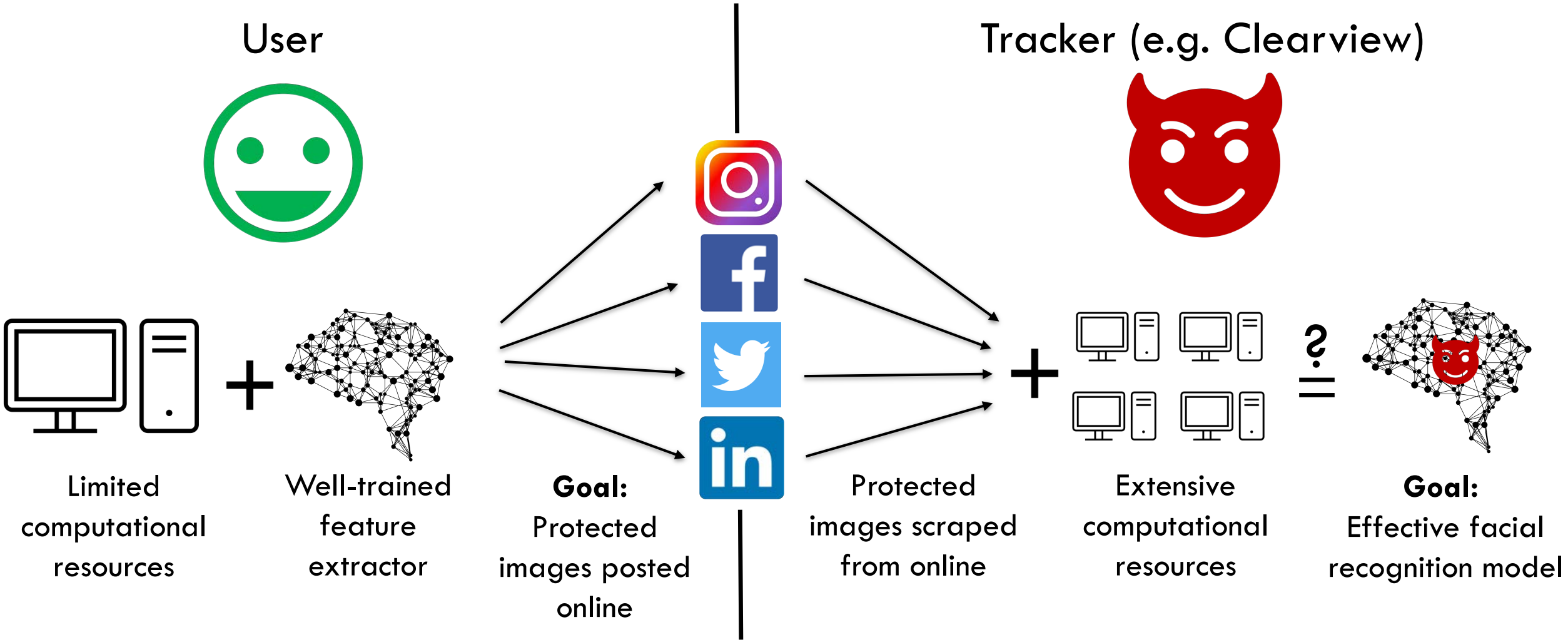
Goals and Assumptions



Goals and Assumptions



Goals and Assumptions



Intuitive View of Facial Recognition Models

Training Images

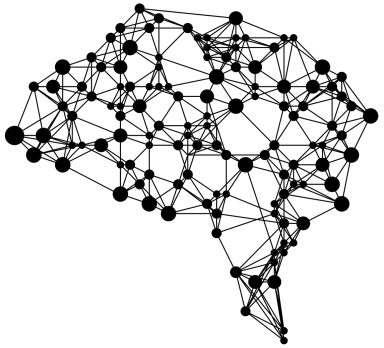
…

Intuitive View of Facial Recognition Models

Training Images

...

Feature extractor

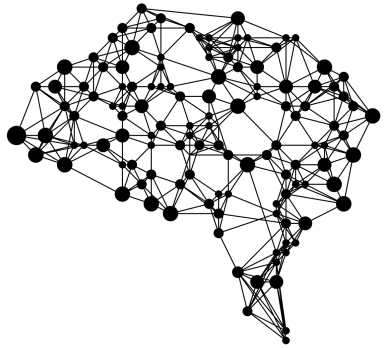


Intuitive View of Facial Recognition Models

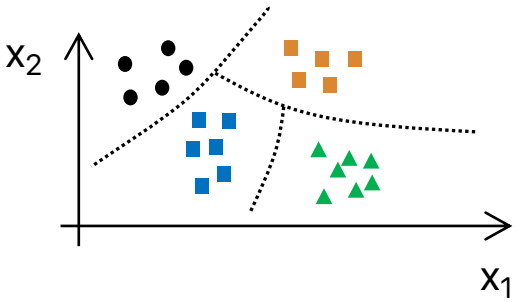
Training Images

...

Feature extractor



Feature Representation

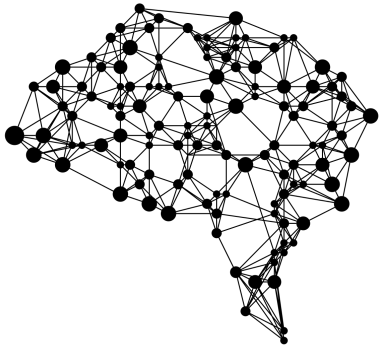


Intuitive View of Facial Recognition Models

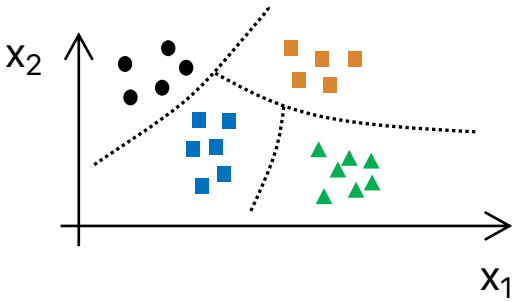
Training Images

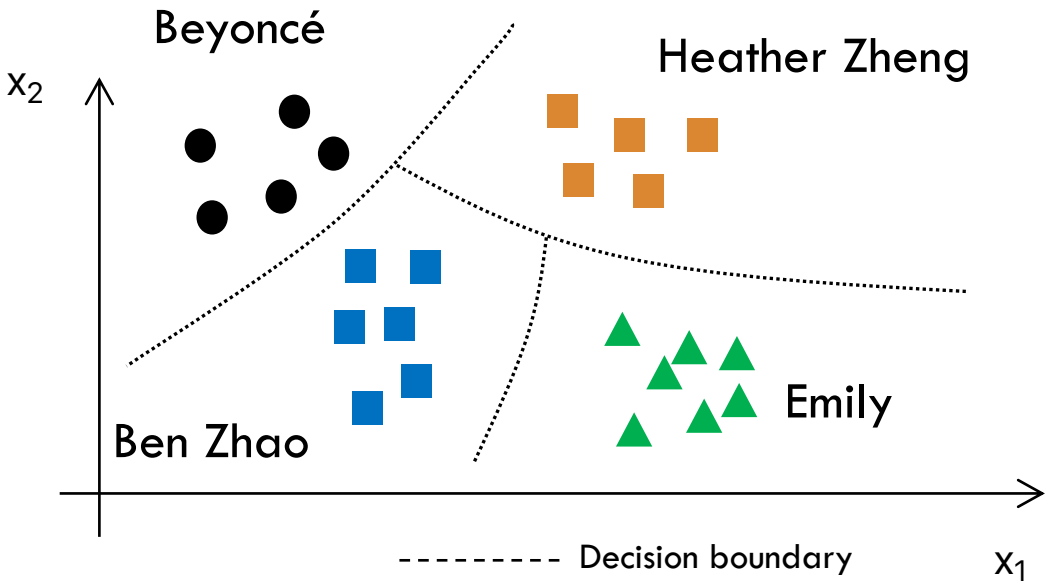
...

Feature extractor



Feature Representation



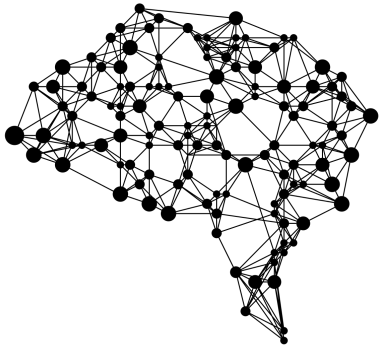


Intuitive View of Facial Recognition Models

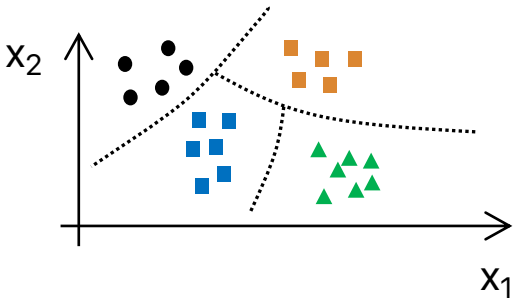
Training Images

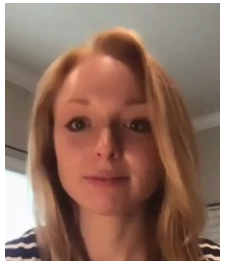
...

Feature extractor

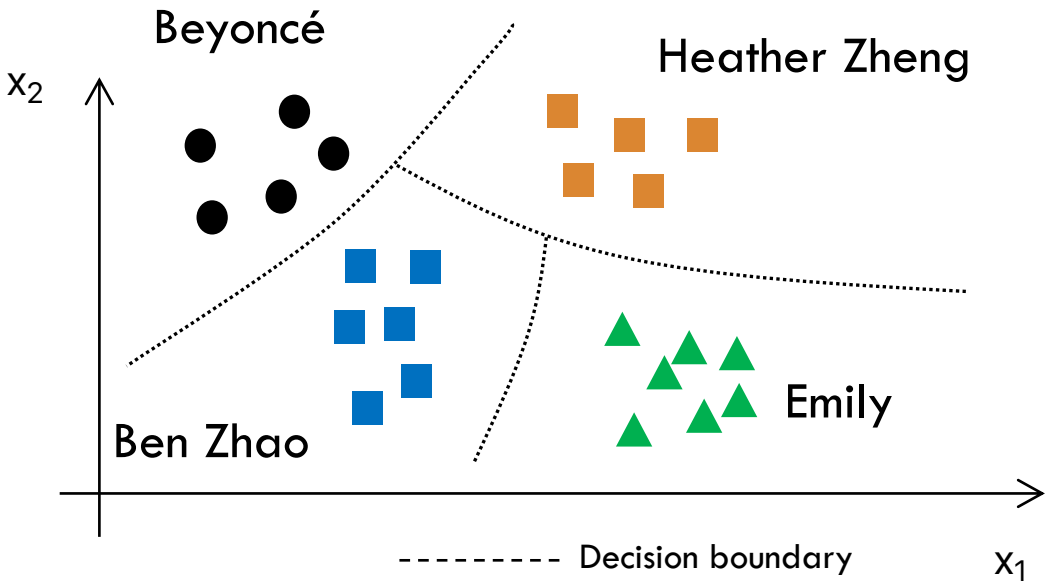


Feature Representation

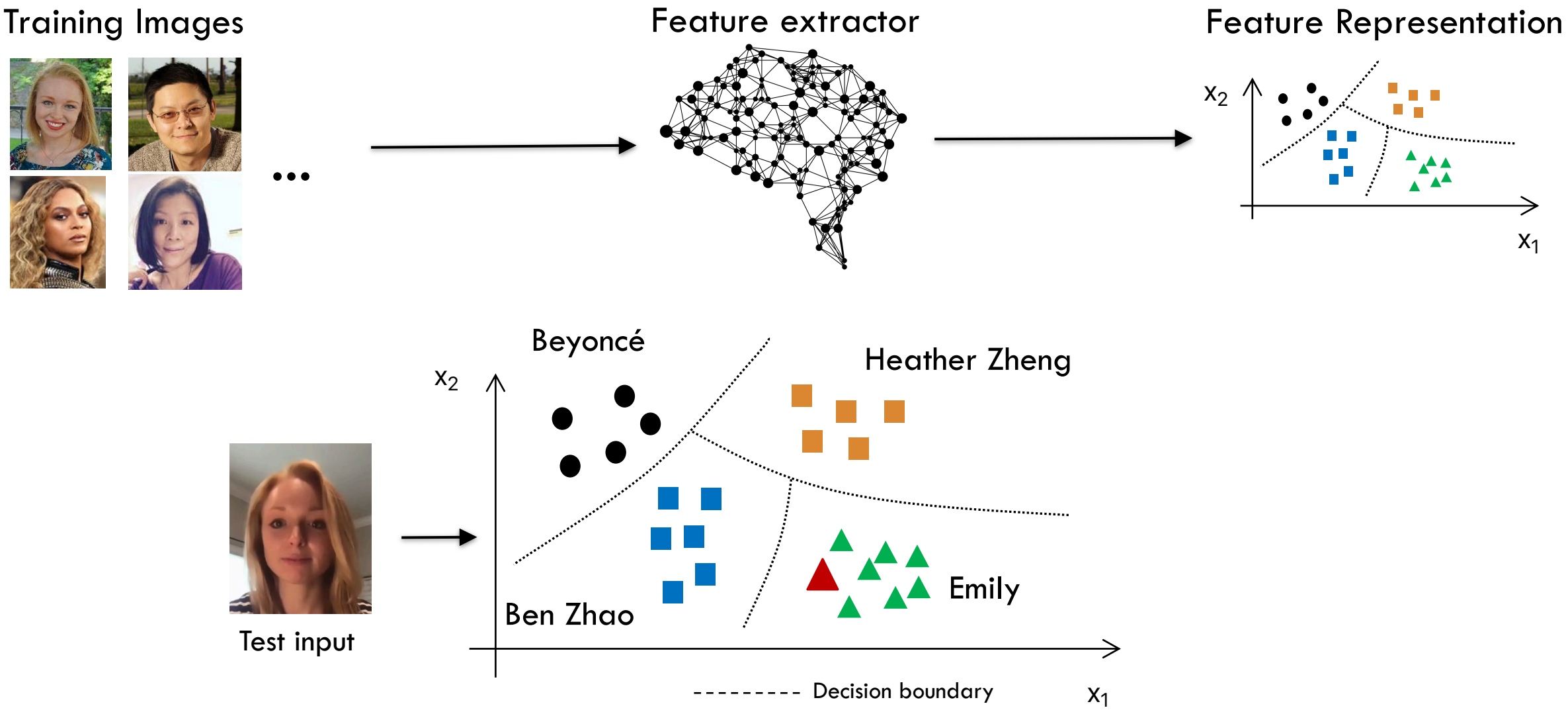




Test input



Intuitive View of Facial Recognition Models

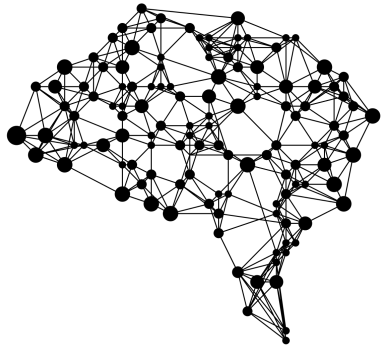


Intuitive View of Facial Recognition Models

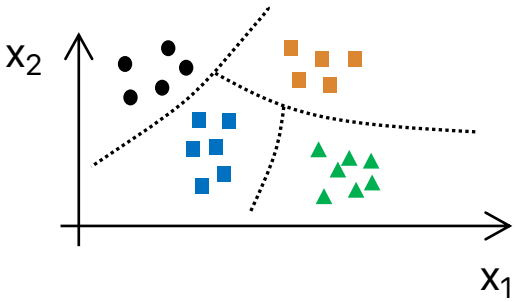
Training Images

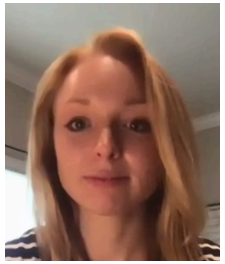
...

Feature extractor

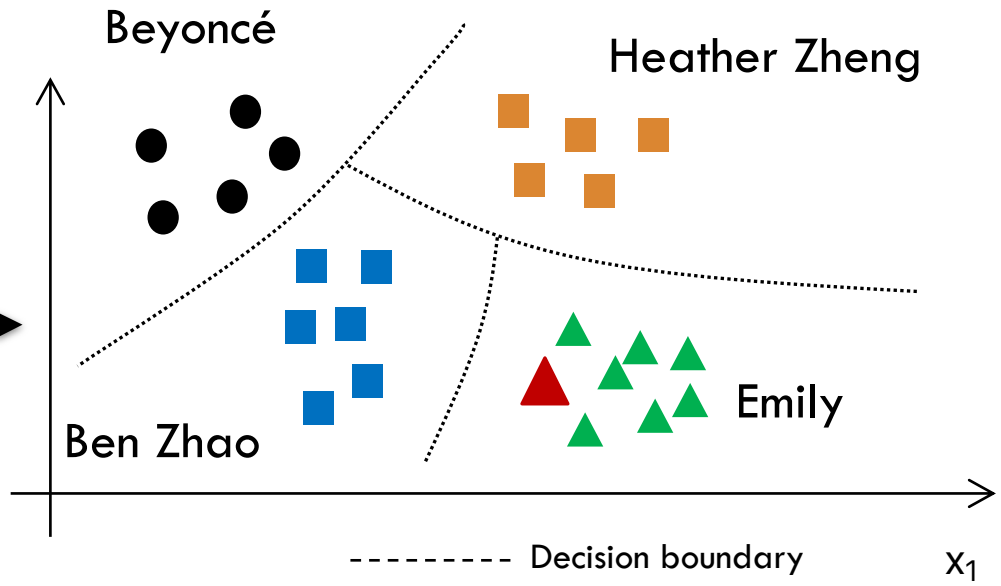


Feature Representation





Test input

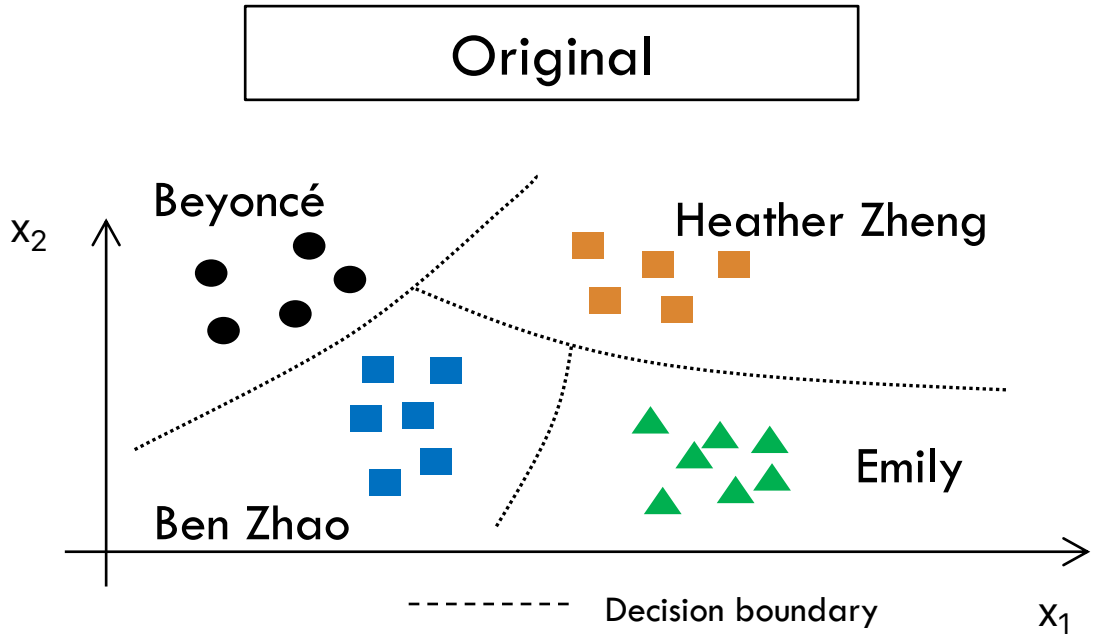


Classification based on feature space separation

Emily

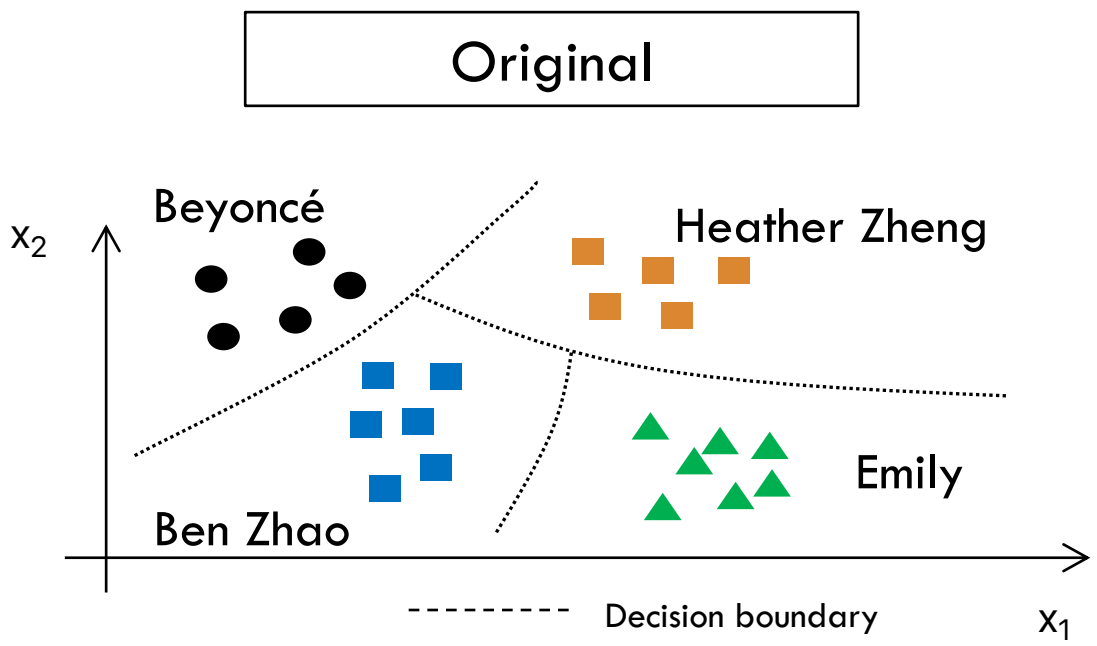
Key Intuition for Fawkes

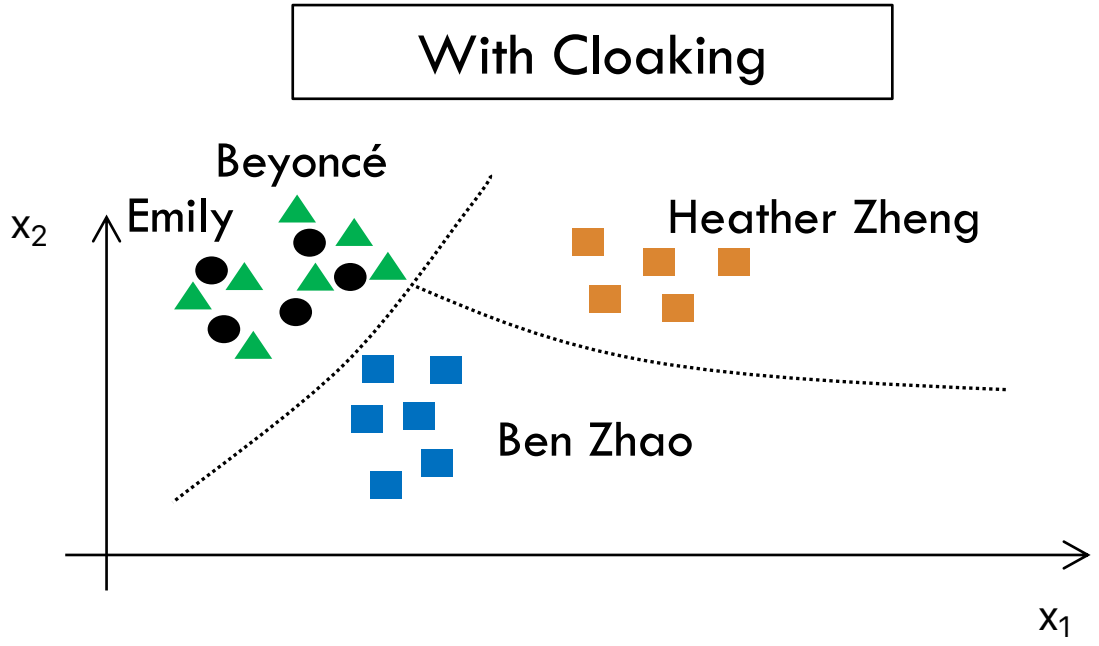
To evade unwanted facial recognition, *change the feature space representation* of user images.



Key Intuition for Fawkes

To evade unwanted facial recognition, *change the feature space representation* of user images.

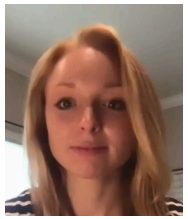


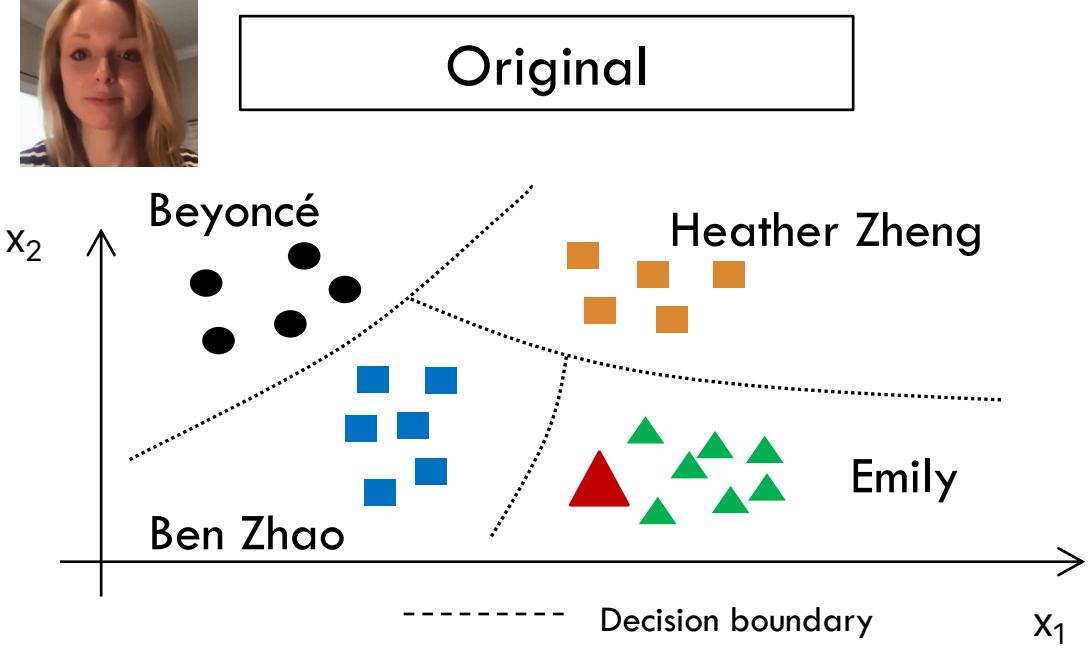


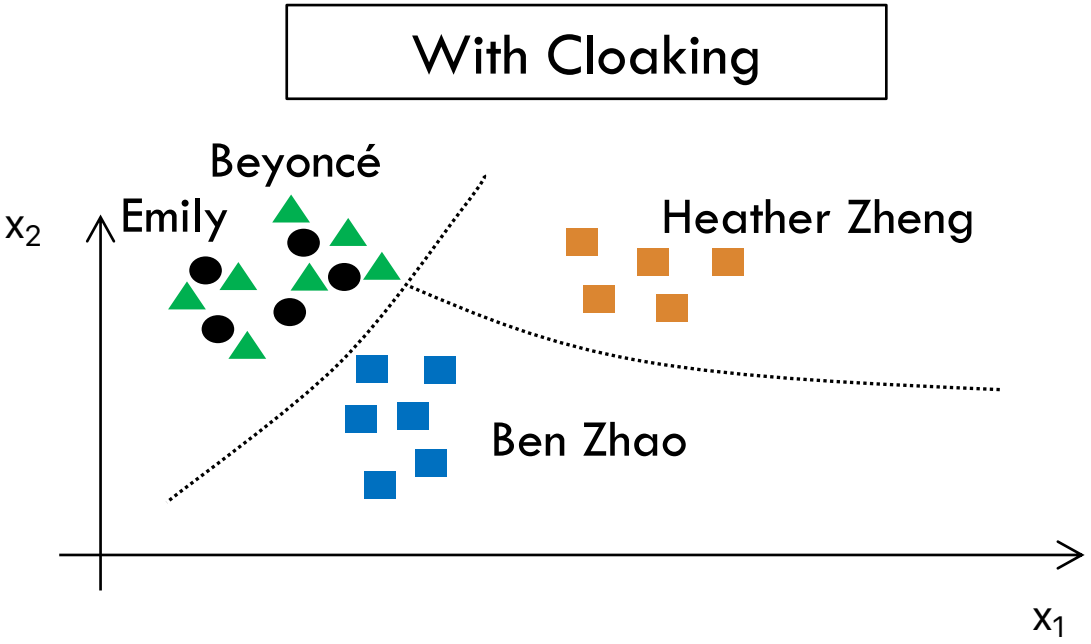
Intuition: Emily's cloaked images are perturbed to have similar feature space representations to Beyoncé's images, distinct from Emily's original feature space representation.

Key Intuition for Fawkes

To evade unwanted facial recognition, *change the feature space representation* of user images.



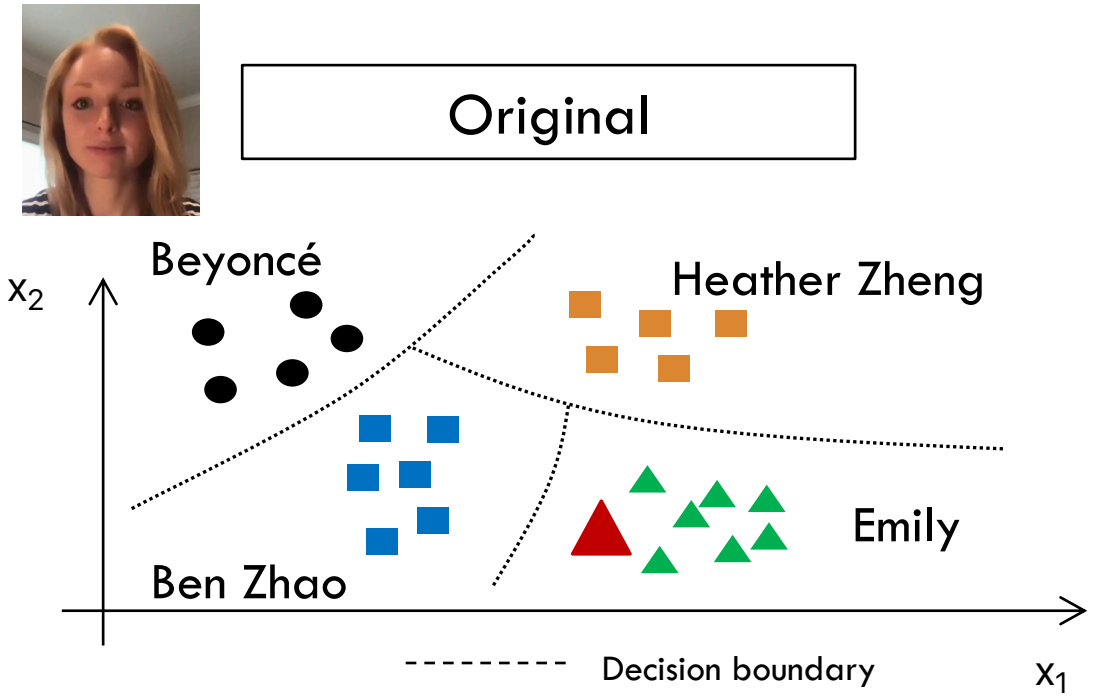


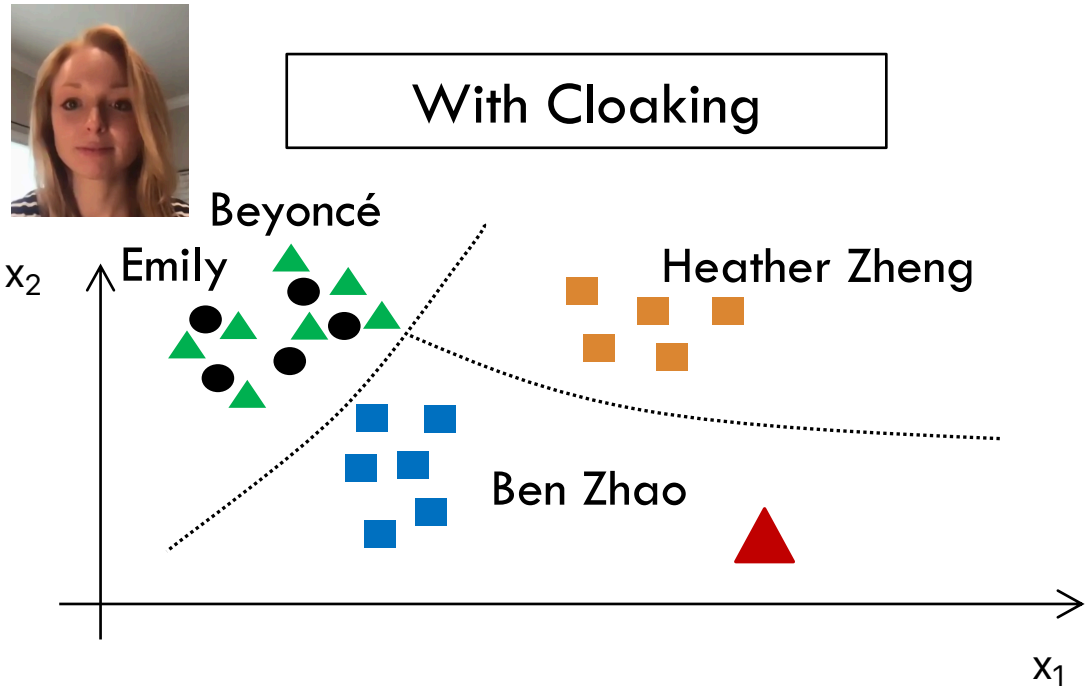


Intuition: Emily's cloaked images are perturbed to have similar feature space representations to Beyoncé's images, distinct from Emily's original feature space representation.

Key Intuition for Fawkes

To evade unwanted facial recognition, *change the feature space representation* of user images.





Intuition: Emily's cloaked images are perturbed to have similar feature space representations to Beyoncé's images, distinct from Emily's original feature space representation.

How to Generate Cloak?

Compute cloak perturbation (Δ) by solving an optimization problem

- Goal: mimic feature representations of target class T
- Constraint: perturbation should be indistinguishable by humans

How to Generate Cloak?

Compute cloak perturbation (Δ) by solving an optimization problem

- Goal: mimic feature representations of target class T
- Constraint: perturbation should be indistinguishable by humans

$\mathbf{\Phi}(X)$: internal representation in feature extractor $\mathbf{\Phi}$

Source image Target image

$$\min \quad Distance(\mathbf{\Phi}(X_s + \Delta), \mathbf{\Phi}(X_t))$$

Minimize *L2* distance between internal representations

How to Generate Cloak?

Compute cloak perturbation (Δ) by solving an optimization problem

- Goal: mimic feature representations of target class T
- Constraint: perturbation should be indistinguishable by humans

$\mathbf{\Phi}(X)$: internal representation in feature extractor $\mathbf{\Phi}$

Source image Target image

$$\begin{aligned} \min \quad & Distance(\mathbf{\Phi}(X_s + \Delta), \mathbf{\Phi}(X_t)) \\ s.t. \quad & perturb\_magnitude(X_s + \Delta, X_s) < P_{budget} \end{aligned}$$

Minimize *L2* distance between internal representations

How to Generate Cloak?

Compute cloak perturbation (Δ) by solving an optimization problem

- Goal: mimic feature representations of target class T
- Constraint: perturbation should be indistinguishable by humans

$\mathbf{\Phi}(X)$: internal representation in feature extractor $\mathbf{\Phi}$

Source image Target image

$$\begin{aligned} \min \quad & Distance(\mathbf{\Phi}(X_s + \Delta), \mathbf{\Phi}(X_t)) \\ s.t. \quad & perturb\_magnitude(X_s + \Delta, X_s) < P_{budget} \end{aligned}$$

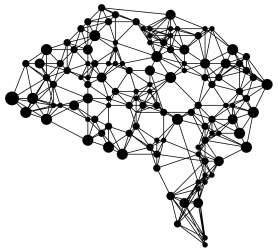
Minimize *L2* distance between internal representations

DSSIM: an objective measure for image distortion

Protection under Baseline Conditions

Original Images

+



Well-trained feature extractor

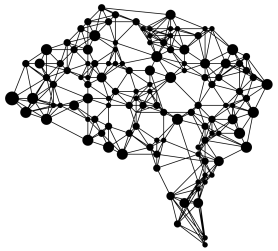
Tracker

Protection under Baseline Conditions

User + Fawkes

Original Images

+



Well-trained feature extractor

Tracker

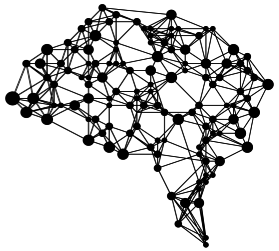
Protection under Baseline Conditions

User + Fawkes

Tracker

Original Images

+



Well-trained feature extractor

→

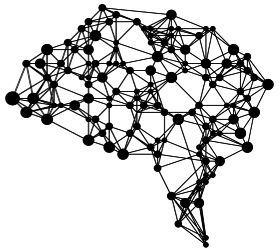
Cloaked Images

Protection under Baseline Conditions

User + Fawkes

Original Images

+

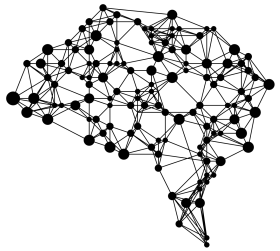


Well-trained feature extractor

Tracker

Cloaked Images

+



Well-trained feature extractor

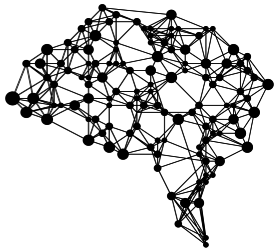
Face recognition model

Protection under Baseline Conditions

User + Fawkes

Original Images

+

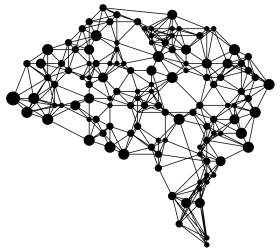


Well-trained feature extractor

Tracker

Cloaked Images

+



Well-trained feature extractor

Face recognition model

Protection Success Rate: Percentage of real (unmodified) user images misclassified by tracker's model

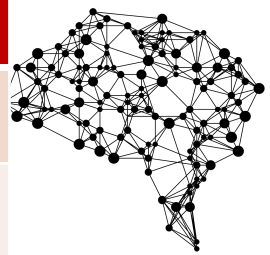
Protection under Baseline Conditions

User + Fawkes

Tracker

Original Images

Feature Extractor Used	Protection Success Rate
VGGFace2 + InceptionResNet	100%
VGGFace2 + DenseNet	100%
WebFace + InceptionResNet	100%
WebFace + DenseNet	100%



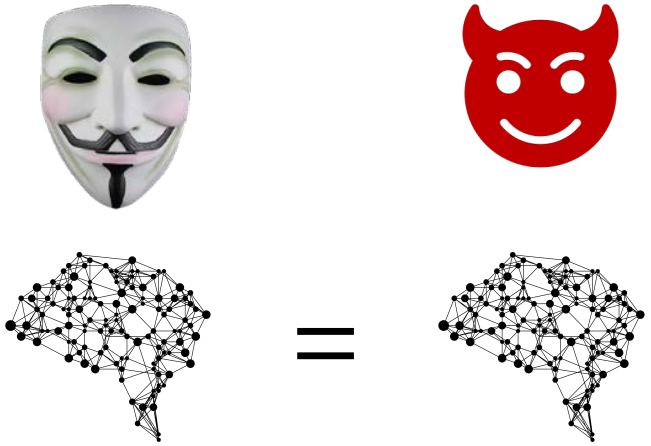
Well-trained feature extractor

Face recognition model

Protection Success Rate: Percentage of real (unmodified) user images misclassified by tracker's model

Protection under Realistic Conditions

Protection under Realistic Conditions

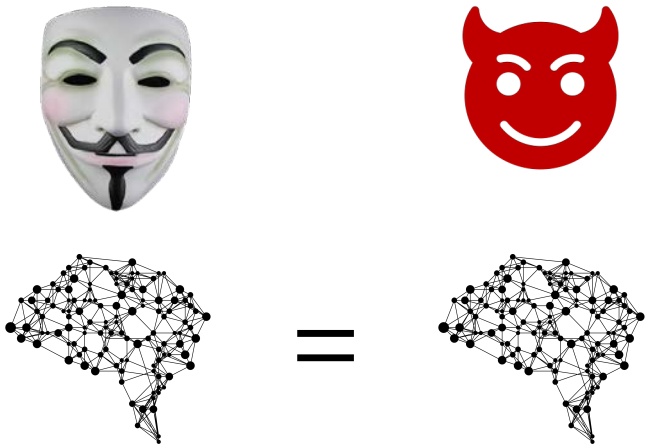


Known Feature Extractor

Fawkes knows tracker's FE, uses it to compute cloak

Protection Rate: 100%

Protection under Realistic Conditions



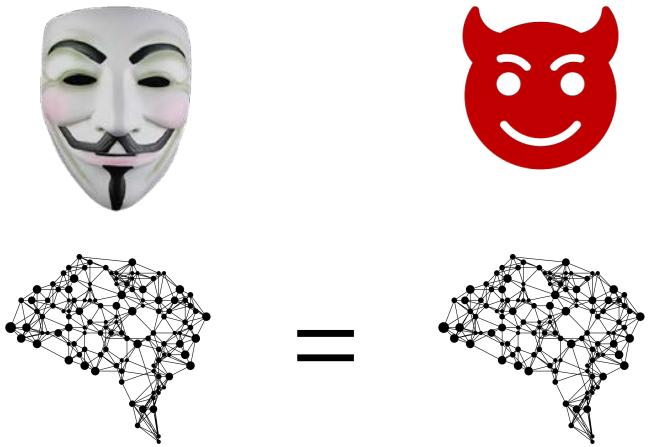
Known Feature Extractor

Fawkes knows tracker's FE, uses it to compute cloak

Protection Rate: 100%

Transferability: models trained on different data (but same application domain) often share similarity in feature space representation, so effects of perturbations from one can transfer to a different feature extractor or model.

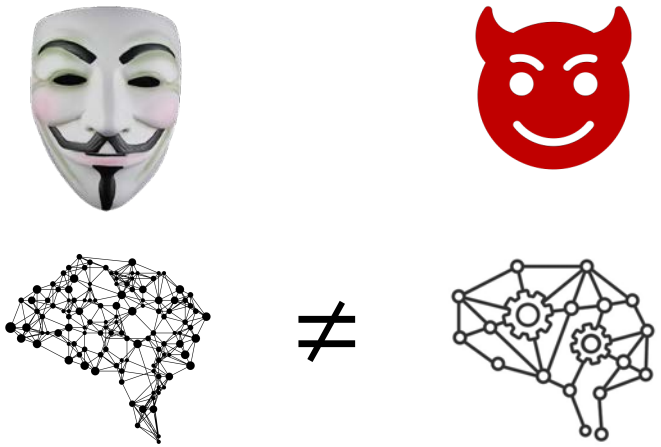
Protection under Realistic Conditions



Known Feature Extractor

Fawkes knows tracker's FE, uses it to compute cloak

Protection Rate: 100%



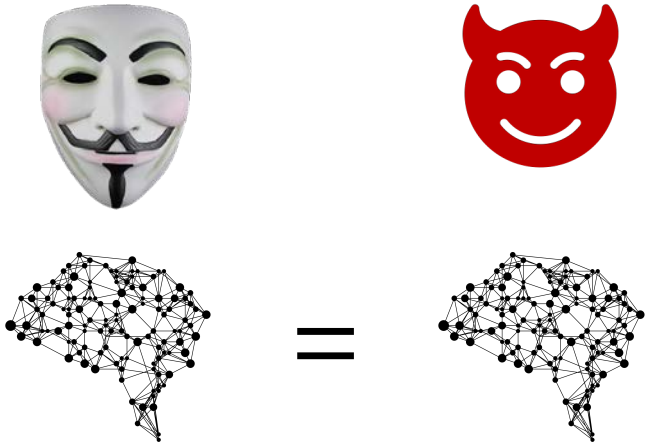
Unknown Feature Extractor

Tracker uses unknown FE. Fawkes computes cloak on local FE & relies on transferability

Protection Rate: >95%

Transferability: models trained on different data (but same application domain) often share similarity in feature space representation, so effects of perturbations from one can transfer to a different feature extractor or model.

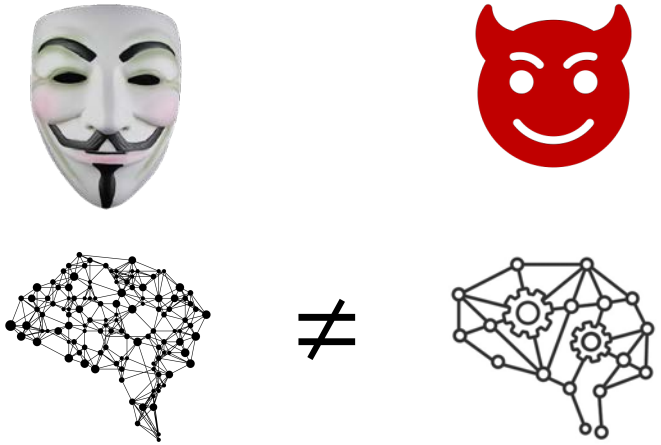
Protection under Realistic Conditions



Known Feature Extractor

Fawkes knows tracker's FE, uses it to compute cloak

Protection Rate: 100%



Unknown Feature Extractor

Tracker uses unknown FE. Fawkes computes cloak on local FE & relies on transferability

Protection Rate: >95%

Train from scratch

Tracker does not use FE. Fawkes computes cloak on local FE & relies on transferability

Protection Rate: >95%

Transferability: models trained on different data (but same application domain) often share similarity in feature space representation, so effects of perturbations from one can transfer to a different feature extractor or model.

Protection against State of the Art APIs

Protection against State of the Art APIs

- How well does Fawkes work on real world Face recognition APIs?

Protection against State of the Art APIs

- How well does Fawkes work on real world Face recognition APIs?

Microsoft Azure

amazon Rekognition

FACE++ inside

Protection against State of the Art APIs

- How well does Fawkes work on real world Face recognition APIs?

Microsoft Azure

amazon Rekognition

FACE++ inside

1. Train facial recognition model on public API
2. Training data includes 1 cloaked user X (all their images are cloaked by Fawkes Using existing feature extractor)
3. Test result model with uncloaked images of user X

Protection against State of the Art APIs

- How well does Fawkes work on real world Face recognition APIs?

Microsoft Azure amazon Rekognition FACE++ inside

- Result is 100% success (no clean images identified as the user, all misclassified)

1. Train facial recognition model on public API
2. Training data includes 1 cloaked user X (all their images are cloaked by Fawkes Using existing feature extractor)
3. Test result model with uncloaked images of user X

Face Recognition API	Protection Success Rate	
	Without Protection	**With Protection**
AWS Rekognition	0%	**100%**
Microsoft Azure	0%	**100%**
Face++	0%	**100%**

Even More Challenging Real-World Scenarios

Even More Challenging Real-World Scenarios

- What if the tracker has original, uncloaked images?
  - Pre-Fawkes images, public sources (newspapers, company page), images shared by friends

Even More Challenging Real-World Scenarios

- What if the tracker has original, uncloaked images?
  - Pre-Fawkes images, public sources (newspapers, company page), images shared by friends
  - Cloaking can succeed if cloaked images outnumber uncloaked images
  - *Sybil accounts* help: boost protection from 30% to 95% (1:1 cloak:uncloaked ratio)

Even More Challenging Real-World Scenarios

- What if the tracker has original, uncloaked images?
  - Pre-Fawkes images, public sources (newspapers, company page), images shared by friends
  - Cloaking can succeed if cloaked images outnumber uncloaked images
  - *Sybil accounts* help: boost protection from 30% to 95% (1:1 cloak:uncloaked ratio)
- What if the tracker tries to detect/remove cloaked effects?
  - Using tools like image transformation, anomaly detection
  - Ineffective against Fawkes

Even More Challenging Real-World Scenarios

- What if the tracker has original, uncloaked images?
  - Pre-Fawkes images, public sources (newspapers, company page), images shared by friends
  - Cloaking can succeed if cloaked images outnumber uncloaked images
  - *Sybil accounts* help: boost protection from 30% to 95% (1:1 cloak:uncloaked ratio)
- What if the tracker tries to detect/remove cloaked effects?
  - Using tools like image transformation, anomaly detection
  - Ineffective against Fawkes
- Limitations of Fawkes

Even More Challenging Real-World Scenarios

- What if the tracker has original, uncloaked images?
  - Pre-Fawkes images, public sources (newspapers, company page), images shared by friends
  - Cloaking can succeed if cloaked images outnumber uncloaked images
  - *Sybil accounts* help: boost protection from 30% to 95% (1:1 cloak:uncloaked ratio)
- What if the tracker tries to detect/remove cloaked effects?
  - Using tools like image transformation, anomaly detection
  - Ineffective against Fawkes
- Limitations of Fawkes
  - Not guaranteed to be effective against future models
  - Only a tip of the iceberg

Even More Challenging Real-World Scenarios

- What if the tracker has original, uncloaked images?
  - Pre-Fawkes images, public sources (newspapers, company page), images shared by friends
  - Cloaking can succeed if cloaked images outnumber uncloaked images
  - *Sybil accounts* help: boost protection from 30% to 95% (1:1 cloak:uncloaked ratio)
- What if the tracker tries to detect/remove cloaked effects?
  - Using tools like image transformation, anomaly detection
  - Ineffective against Fawkes
- Limitations of Fawkes
  - Not guaranteed to be effective against future models
  - Only a tip of the iceberg
- More details in paper!

Fawkes: Protecting Privacy against Unauthorized Deep Learning Models

Shawn Shan[†], *Emily Wenger*[†], *Jiayun Zhang, Huiying Li, Haitao Zheng, Ben Y. Zhao*

[†] *denotes co-first authors with equal contribution*

Computer Science, University of Chicago

{shansixiong, ewillson, jiayunz, huiyingli, htzheng, ravenben}@cs.uchicago.edu

Abstract

tion of powerful facial recognition sys-
t to personal privacy. As *Clearview.ai*
canvas the Internet for data and
recognition models of individu-
need tools to protect our-
thorized facial recog-
l or effective solu-

Opportunities for misuse of this technology are numero
and potentially disastrous. Anywhere we go, we can be i
tified at any time through street cameras, video doorbe
curity cameras, and personal cellphones. Stalkers
out our identity and social media profiles with a
shot [47]. Stores can associate our precise in-s
behavior with online ads and browsing profi
thieves can easily identify (and perhaps
personal accounts [13].

We believe that private citizens ne
selves from being identified by u
tion models. Unfortunately, pr
sparse and limited in both p
have proposed distorting i
ble and thus avoiding
oduce adversaria
n sweatshi
ns f

that helps
zed fa-
ing

Thank You!

Thank You!

- More on http://sandlab.cs.uchicago.edu/fawkes
  - Source code
  - Binaries for MacOS/Windows/Linux
  - FAQs
- Encouraging initial response from users
  - 2.5K downloads as of July 20th