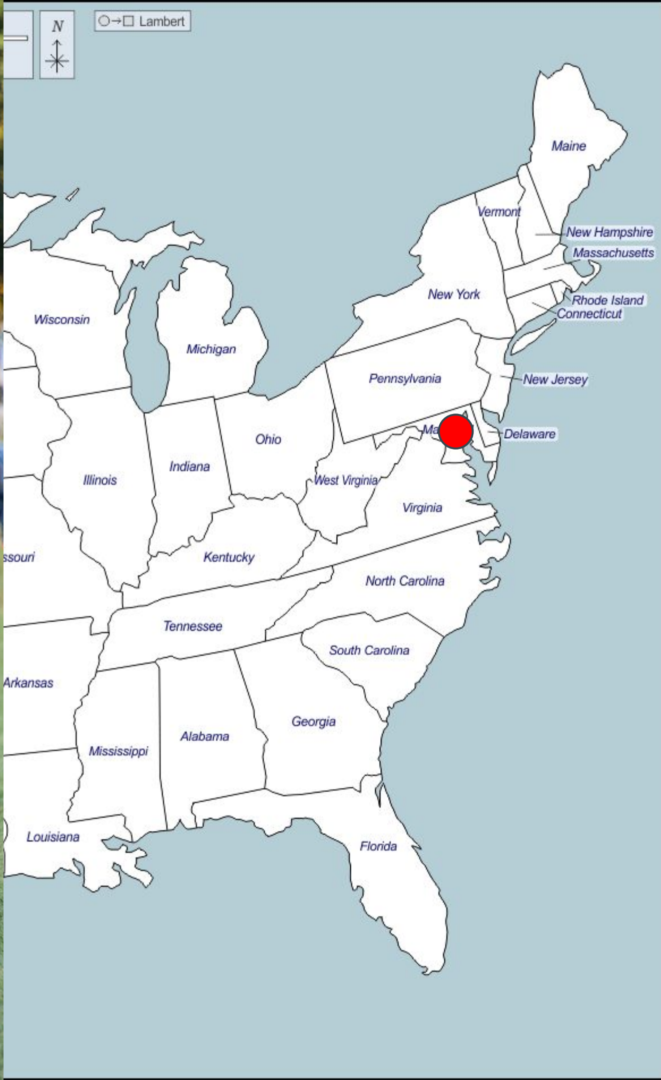
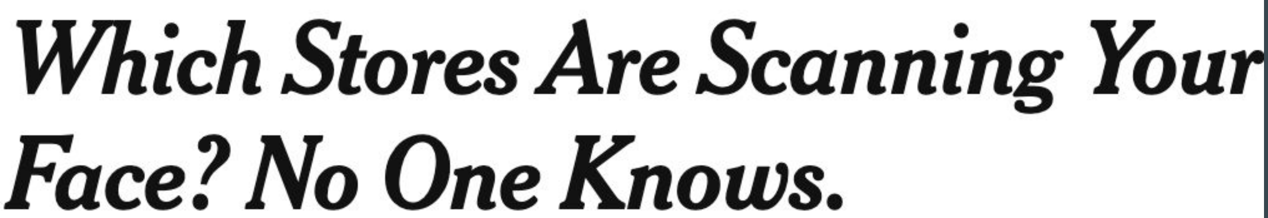
Opting Out of Facial Recognition

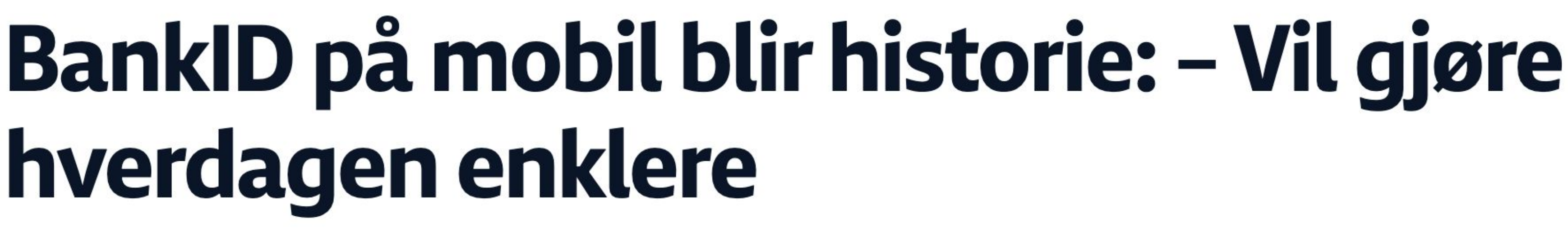
•••

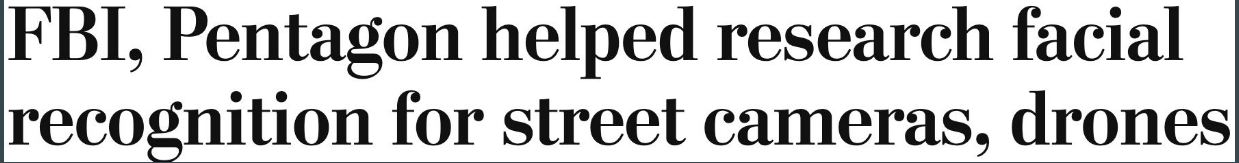
Gavin Taylor
US Naval Academy

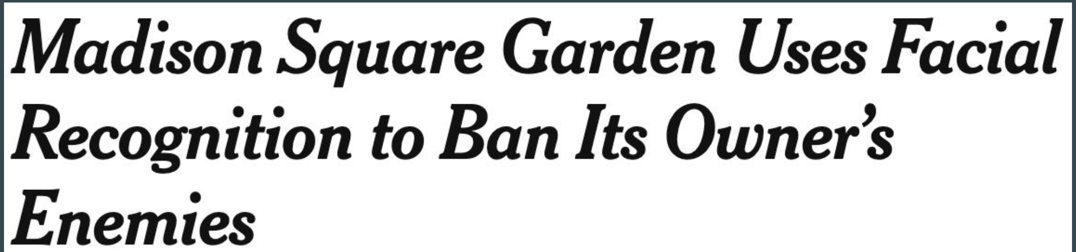


N
Lambert
Maine
Vermont
New Hampshire
Massachusetts
New York
Rhode Island
Connecticut
Wisconsin
Michigan
Pennsylvania
New Jersey
Ma
Delaware
Ohio
Indiana
Illinois
West Virginia
Virginia
ssouri
Kentucky
North Carolina
Tennessee
South Carolina
Arkansas
Georgia
Alabama
Mississippi
Louisiana
Florida

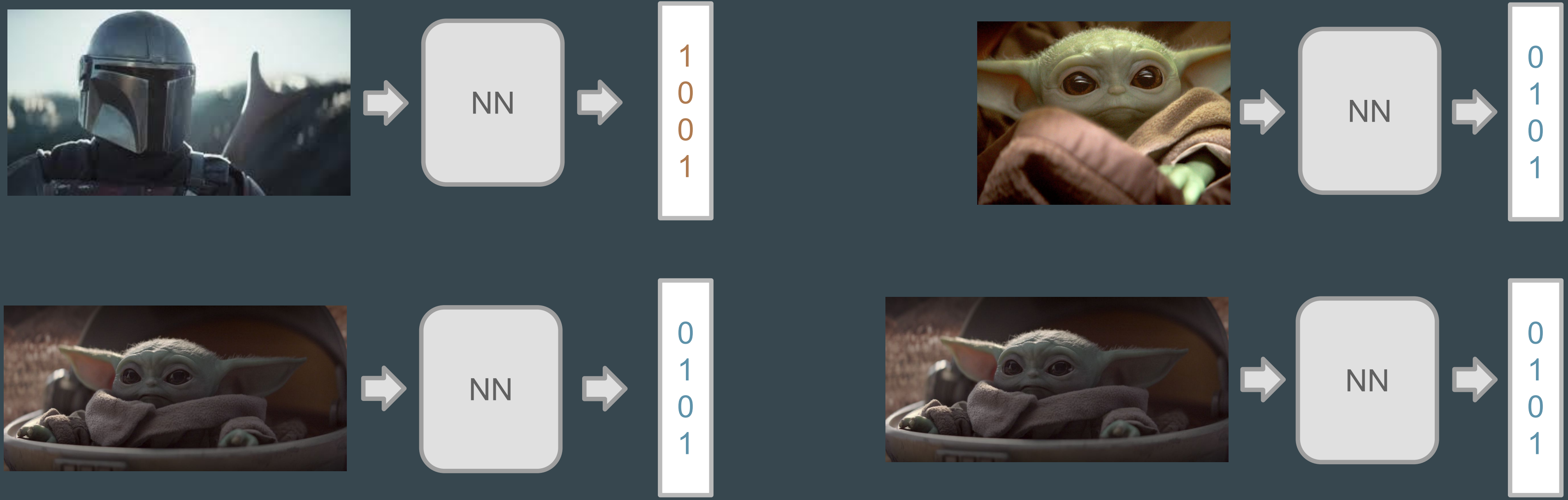

Which Stores Are Scanning Your Face? No One Knows.


BankID på mobil blir historie: – Vil gjøre hverdagen enklere


FBI, Pentagon helped research facial recognition for street cameras, drones


Madison Square Garden Uses Facial Recognition to Ban Its Owner's Enemies

Facial Recognition Embeddings



NN

1
0
0
1

NN

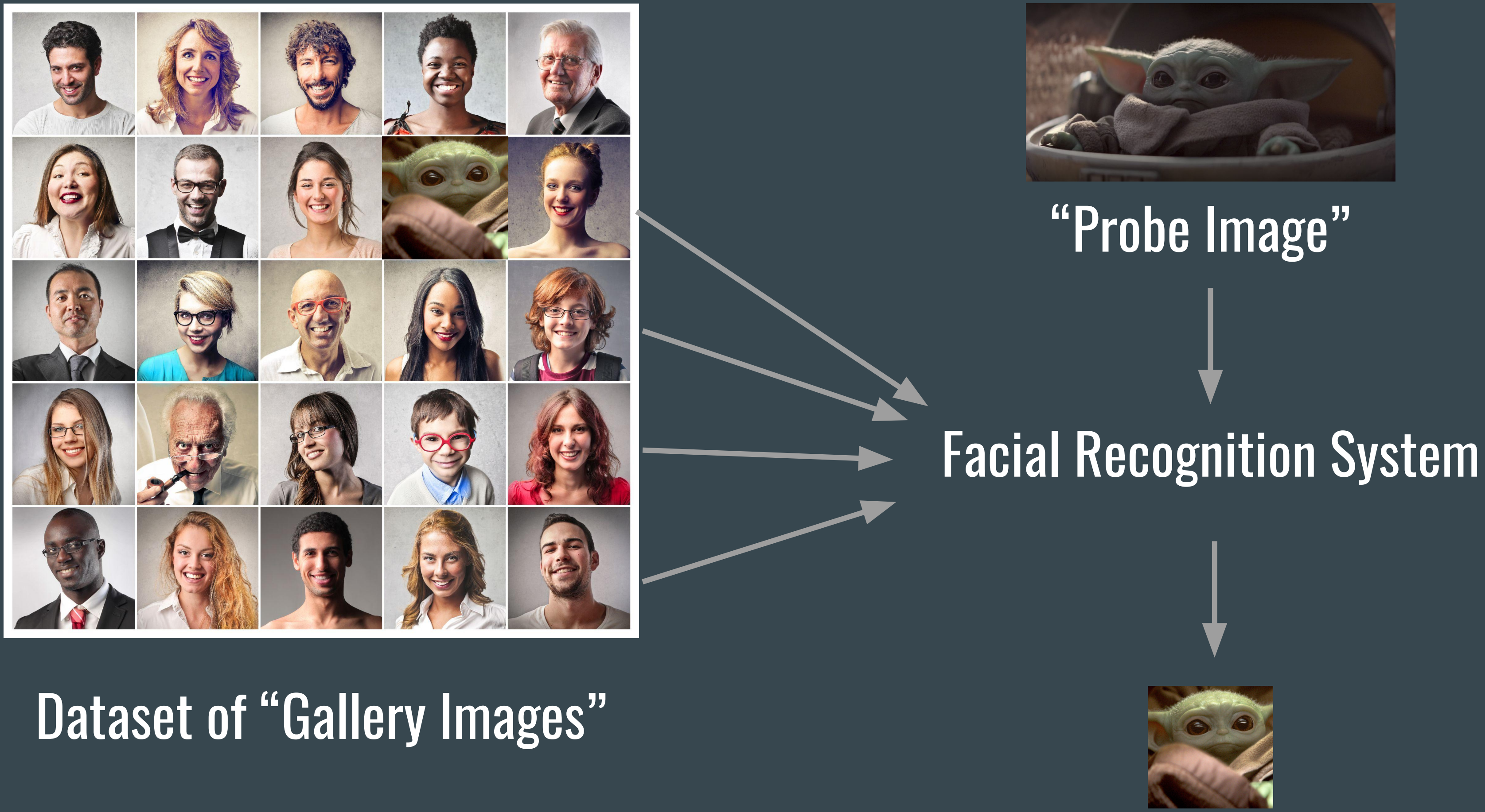
0
1
0
1

NN

0
1
0
1

NN

0
1
0
1

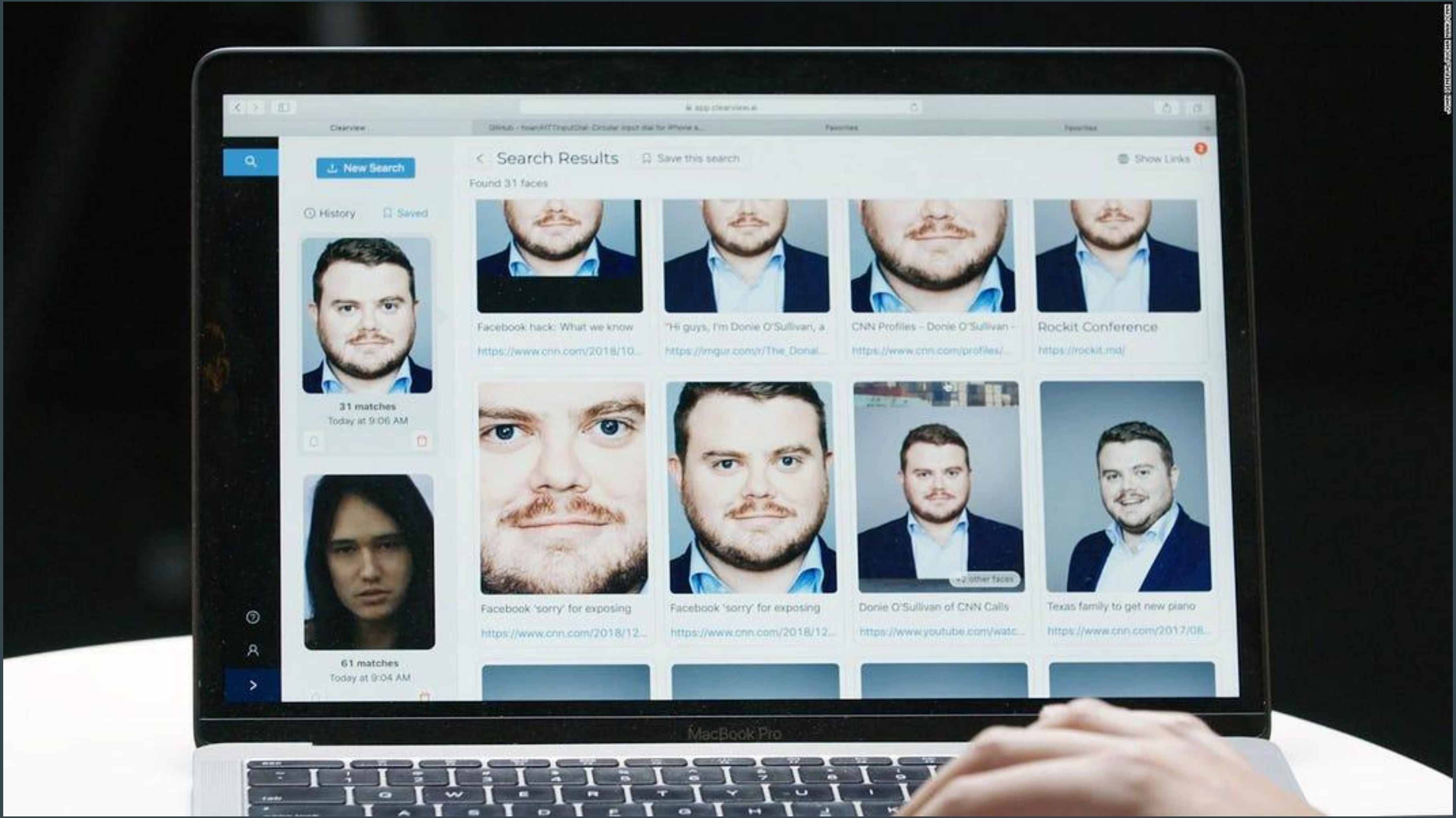


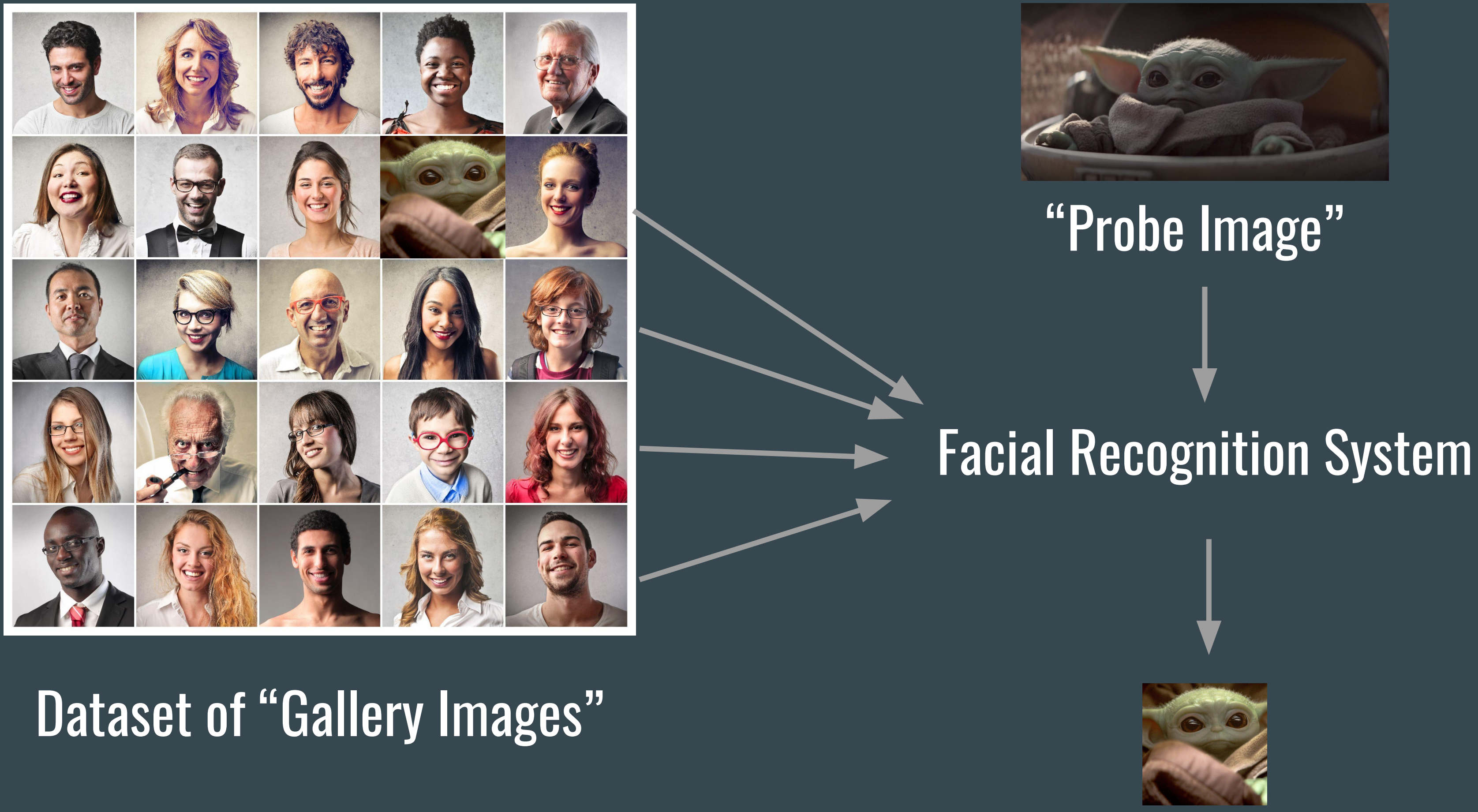
Dataset of “Gallery Images”

“Probe Image”

Facial Recognition System

Most similar faces


app.clearview.ai
Clearview
Favorites
Favorites
New Search
History
Saved
31 matches
Today at 9:06 AM
61 matches
Today at 9:04 AM
Search Results
Save this search
Show Links
Found 31 faces
Facebook hack: What we know
https://www.cnn.com/2018/10...
"Hi guys, I'm Donie O'Sullivan, a
https://imgur.com/r/The_Donal...
CNN Profiles - Donie O'Sullivan -
https://www.cnn.com/profiles/...
Rockit Conference
https://rockit.md/
Facebook 'sorry' for exposing
https://www.cnn.com/2018/12...
Facebook 'sorry' for exposing
https://www.cnn.com/2018/12...
+2 other faces
Donie O'Sullivan of CNN Calls
https://www.youtube.com/watc...
Texas family to get new piano
https://www.cnn.com/2017/08...
MacBook Pro



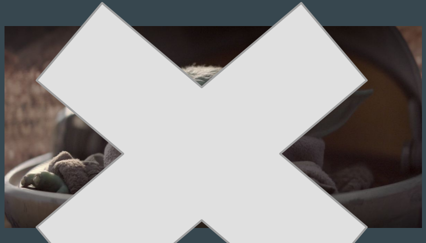
“Probe Image”

Facial Recognition System

Dataset of “Gallery Images”

Most similar faces

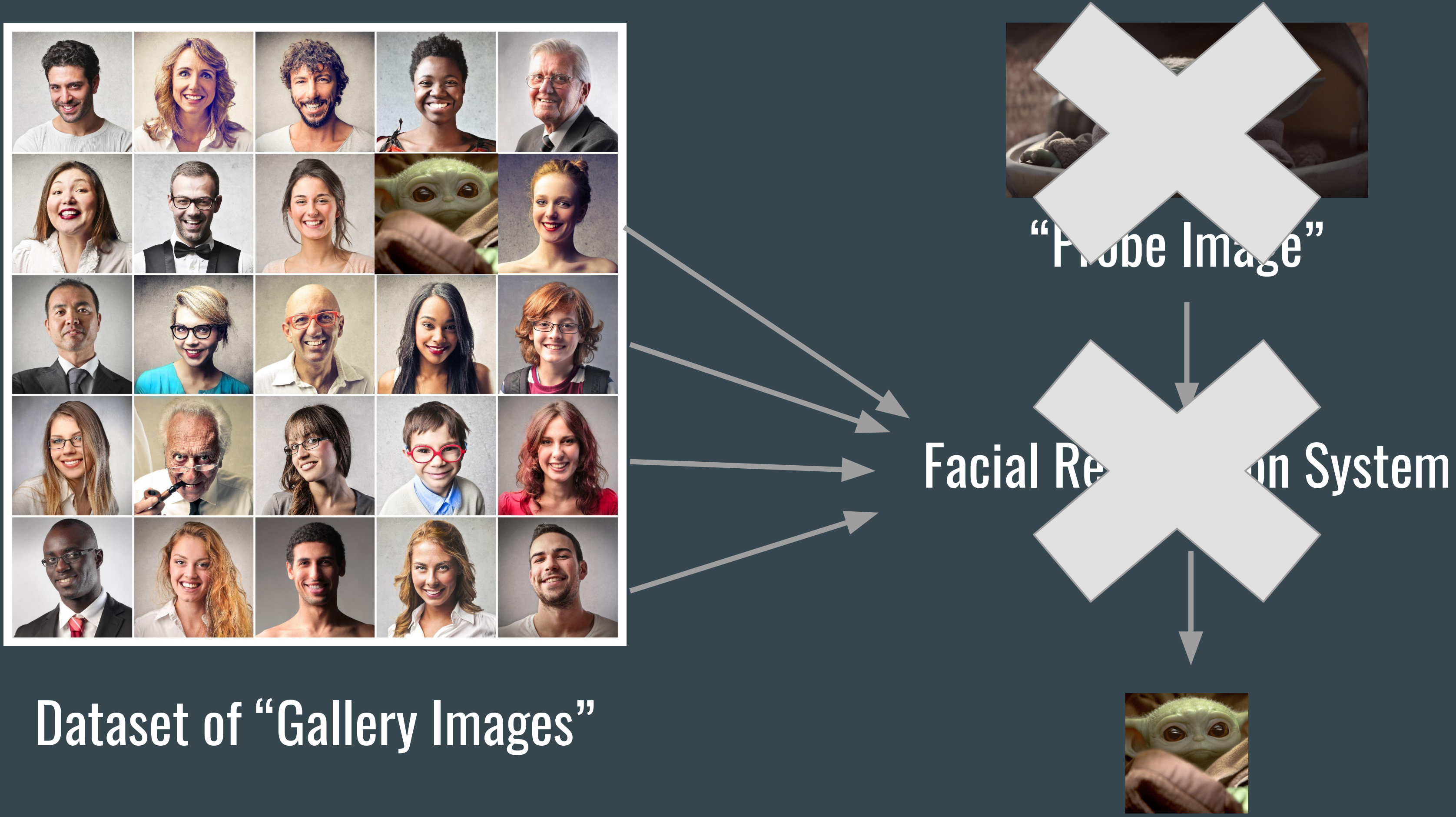
Dataset of “Gallery Images”



“Probe Image”

Facial Recognition System

Most similar faces

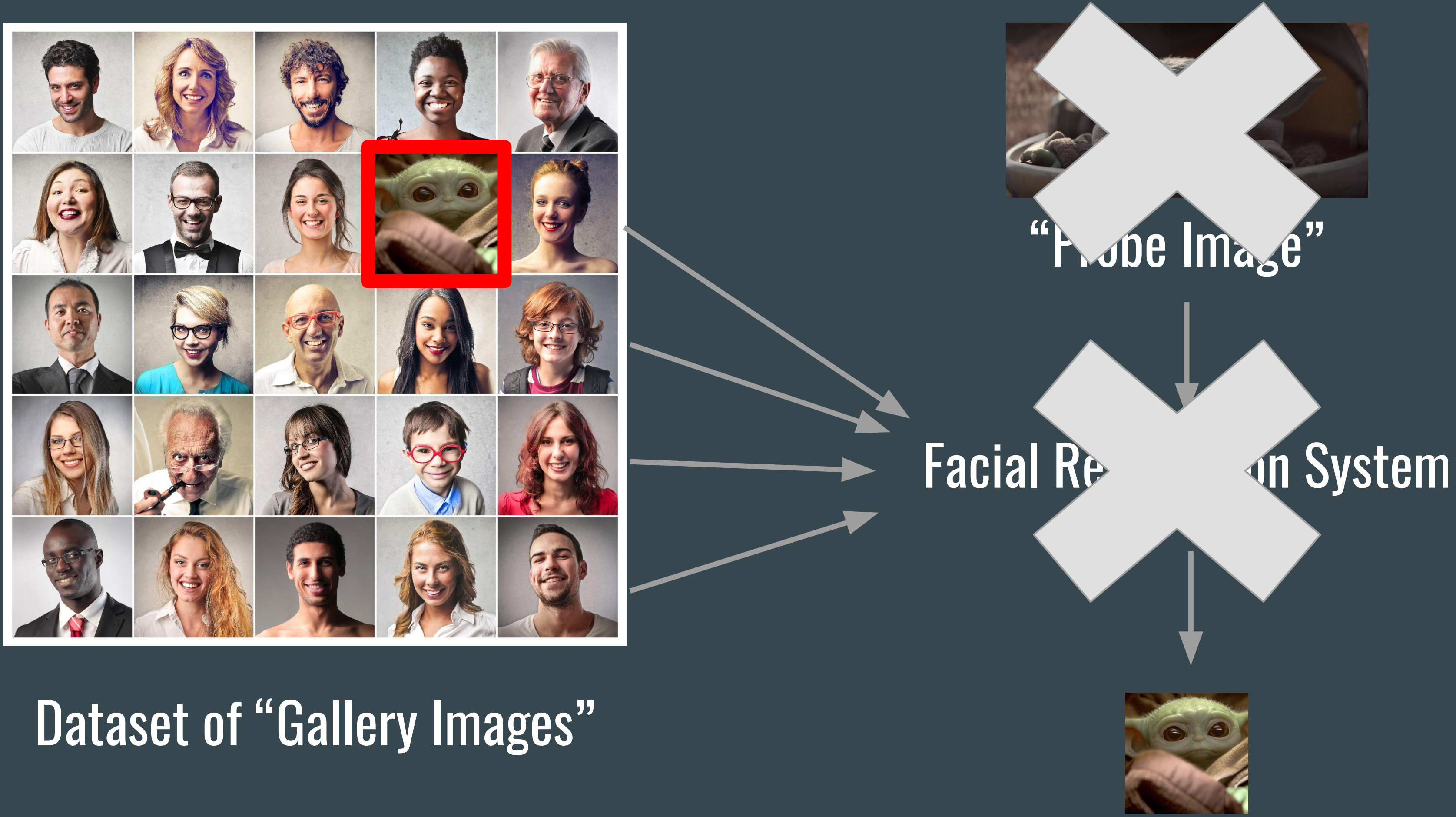


“P obe Ima e”

Facial Re on System

Dataset of “Gallery Images”

Most similar faces



~~“Probe Image”~~

~~Facial Recognition System~~

Dataset of “Gallery Images”

Most similar faces

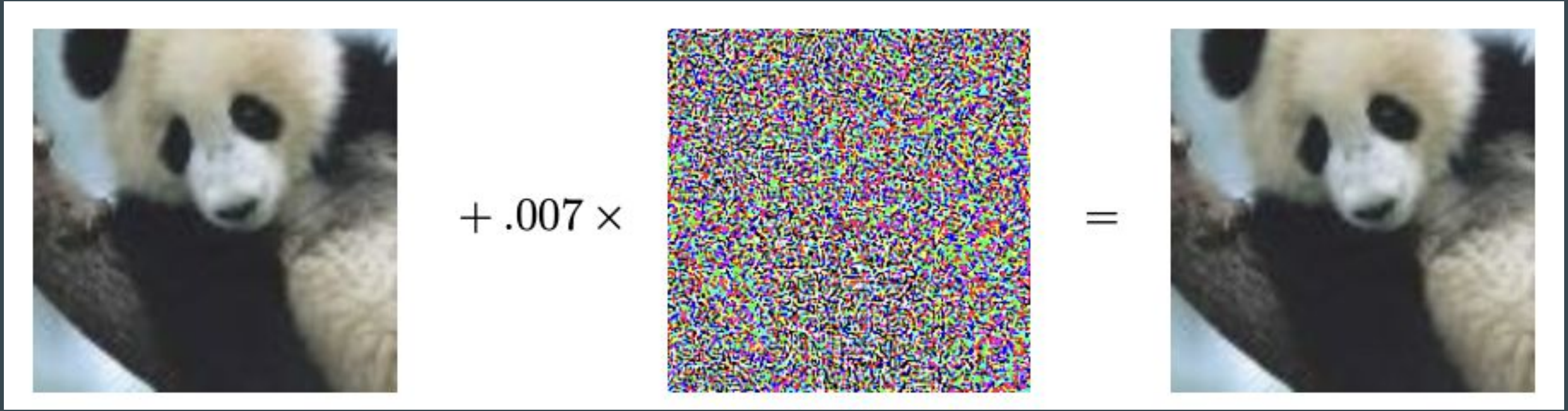
“Probe Image”

GOAL: Manipulate Gallery Image to still serve its purpose, but make it unsuitable for comparison in a black-box facial recognition system.

Dataset of “Gallery Images”

Most similar faces

“Adversarial” examples for ML

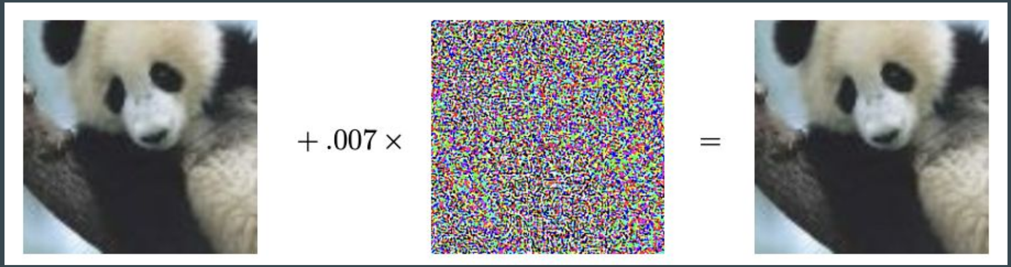


Panda, 57.7% confidence

Gibbon, 99.3% confidence

[Goodfellow et al., “Explaining and Harnessing Adversarial Examples” ICLR 2015]

“Adversarial” examples for ML

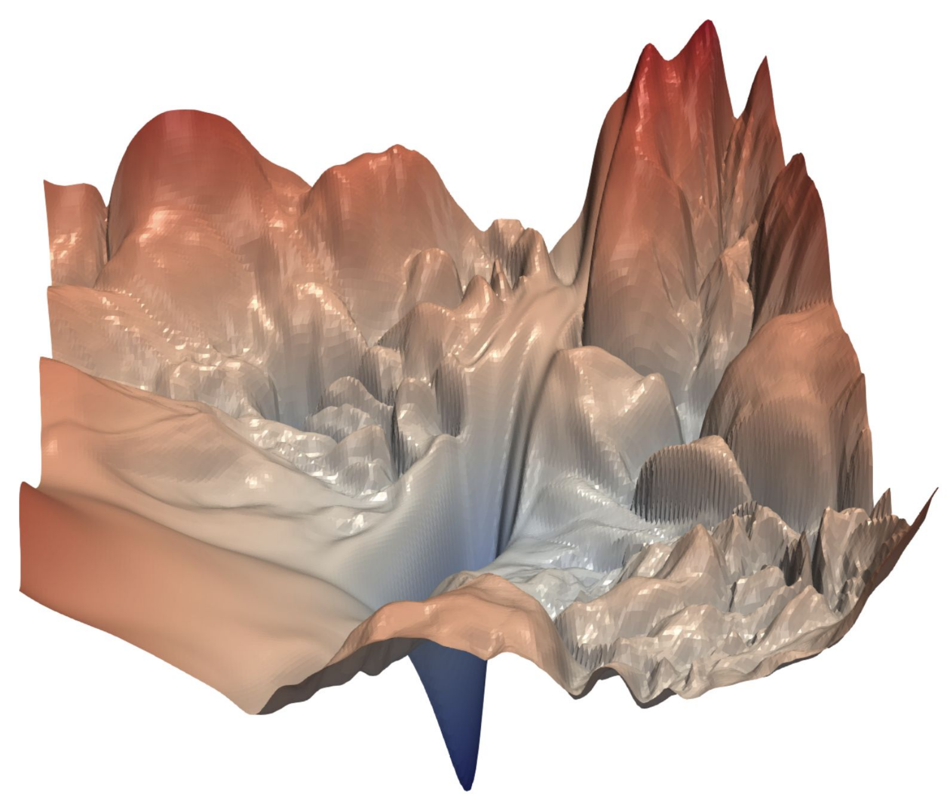


Panda, 57.7% confidence

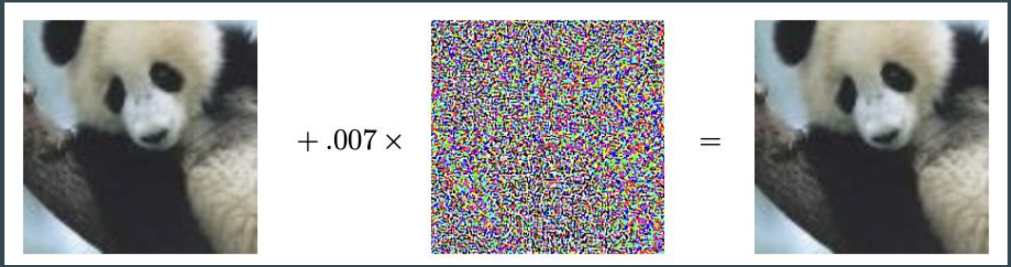
Gibbon, 99.3% confidence

Given a neural network, its parameters θ, an image x, and a loss function $\mathcal{L}(x, \theta)$, maximize the loss function by altering the image a limited amount ($\|\Delta x\|_\infty < \epsilon$).

Why does this work?



“Adversarial” examples for ML



Panda, 57.7% confidence

Gibbon, 99.3% confidence

Given a neural network, its parameters θ, an image x, and a loss function $\mathcal{L}(x, \theta)$, maximize the loss function by altering the image a limited amount ($\|\Delta x\|_\infty < \epsilon$).

Black-box adversarial examples: “Ensemble” approach

Construct several neural networks, and construct adversarial permutations that affect the loss function on all of them - empirically transferable

$$\sum_i L(x, \theta_i)$$

Dataset of “Gallery Images”

“Probe Image”

GOAL: Manipulate Gallery Image to still serve its purpose, but make it unsuitable for comparison in a black-box facial recognition system.

Most similar faces

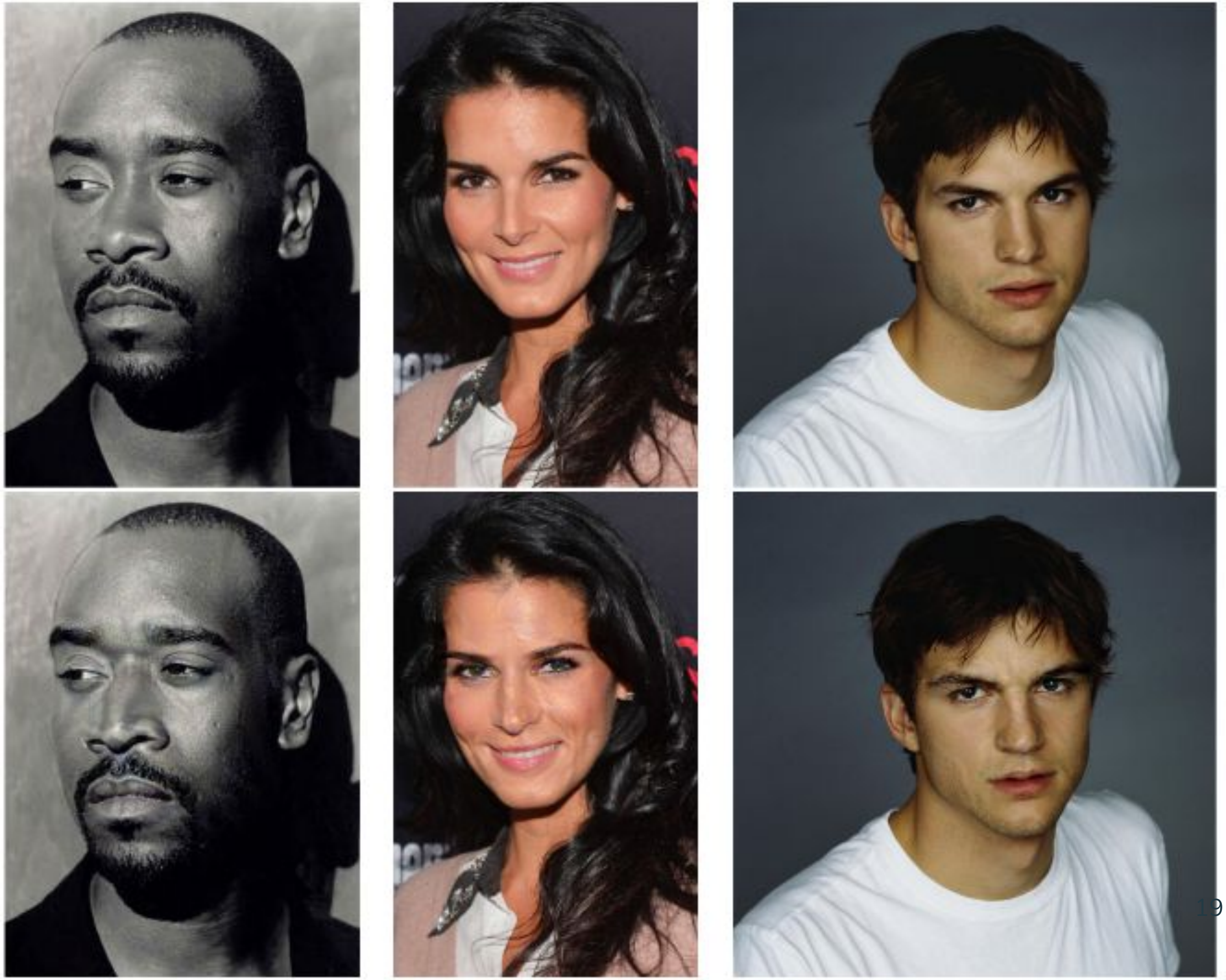
LowKey Optimization Function

$$\max_{x'} \frac{1}{2n} \sum_{i} \frac{\|f_i(x) - f_i(x')\| + \|f_i(x) - f_i(G(x'))\|}{\|f_i(x)\|} - \alpha LPIPS(x, x')$$

- x: Cropped and aligned facial image
- $f_i(x)$: Embedding by model i
- LPIPS: Measure of perceptual difference between two image
- $G(x)$: Gaussian-smoothed facial image

Clean Images

Images protected with LowKey



Effectiveness Against Industrial Black Boxes

- 100,000 images from 530 identities, plus 1 million distractor images
- 100 identities randomly chosen, and all images from those identities manipulated
- If any image from that identity appears in the set of possible matches, the facial recognition system has succeeded

Effectiveness Against Industrial Black Boxes

- 100,000 images from 530 identities, plus 1 million distractor images
- 100 identities randomly chosen, and all images from those identities manipulated
- If any image from that identity appears in the set of possible matches, the facial recognition system has succeeded

	Amazon Rank-1	Amazon Rank-50	Microsoft Rank-1
Clean	93.7%	95.4%	87.7%
LowKey	0.6%	2.4%	0.1%

Valeria Cherepanova

Micah Goldblum

Tom Goldstein

Harrison Foley

Shiyuan Duan

Try it yourself: https://lowkey.umiacs.umd.edu/