New Detection Techniques Against ThisPersonDoesNotExist.com

Breaking a cheap implementation of StyleGAN, and a good excuse to talk about Machine Learning

sogeti
Part of Capgemini

h2s

About the author

Mathis HAMMEL

@MathisHammel

Head of Cybersecurity R&D @ Sogeti
Co-founder, Challenge Designer @ h25

sogeti
Part of Capgemini

h25

Introduction

ThisPersonDoesNotExist.com

Demonstrator of StyleGAN2, a neural network which generates faces

Massively used to create fake personas

ThisPersonDoesNotExist.com

Very realistic face pictures

Hunting fake profiles is harder than ever

How to fight against this?

ThisPersonDoesNotExist.com

Thankfully, StyleGAN2 is hard to train

Most people use ThisPersonDoesNotExist

Hack that website = win

ThisPersonDoesNotExist.com

Thankfully, StyleGAN2 is hard to train

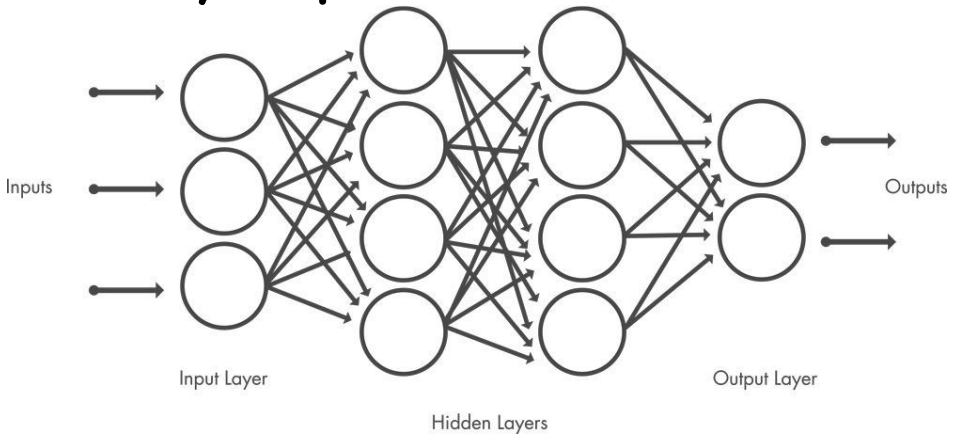
Most people use ThisPersonDoesNotExist

Hack that website = win (foreshadowing...)

But first, let's learn how the system works

Neural Networks

A sequence of many simple math operations that can be “trained” to perform a very specific task



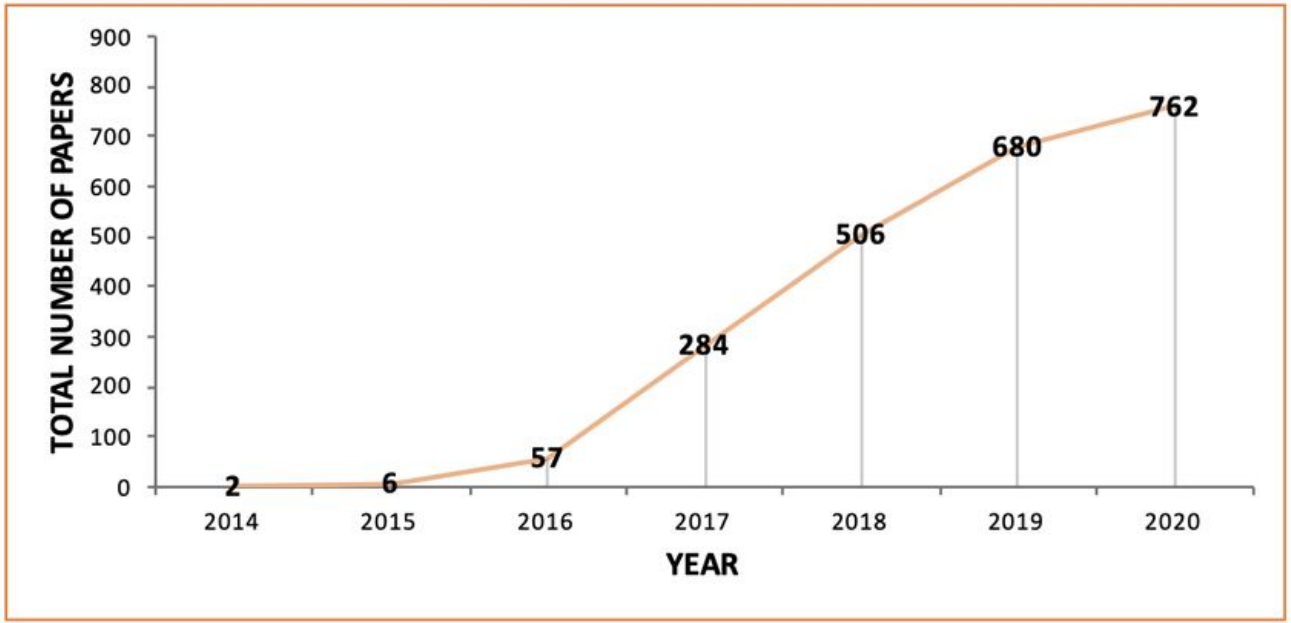
Neural Networks

A few examples:

- Find a plant species' name from a picture
- Evaluate car damage from a video
- Determine when someone says "ok Google"
- Detect suspicious network behavior

What's a GAN?

Generative Adversarial Network



What's a GAN?

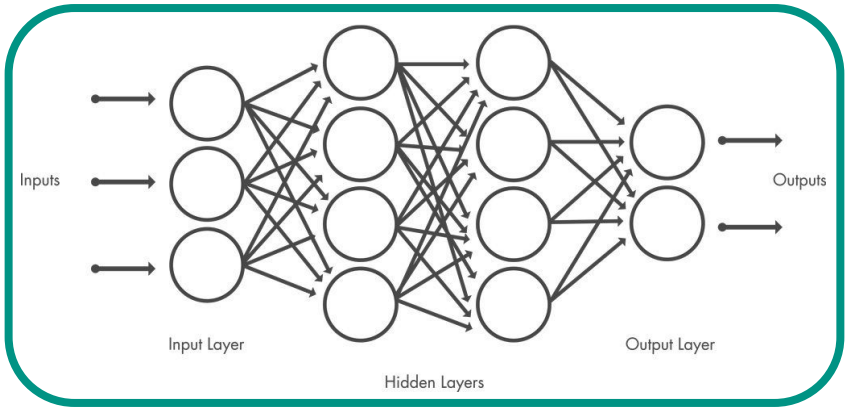
Two different networks

- Generator
- Discriminator

What's a GAN?

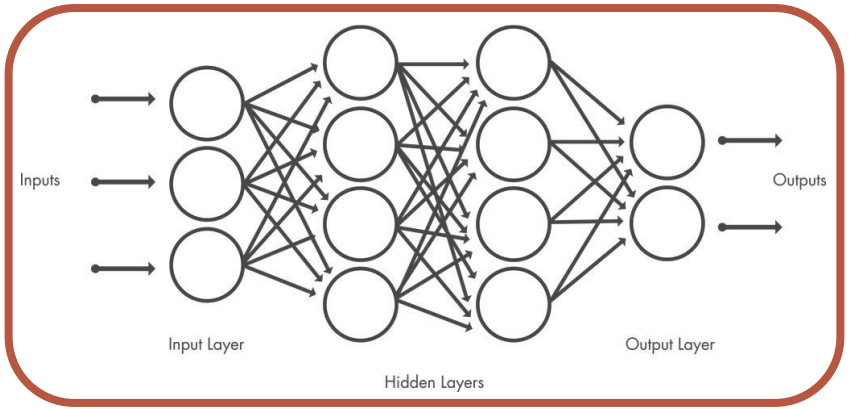
Source of entropy (RNG)

Generator



What's a GAN?

Discriminator



"Is this image real?"

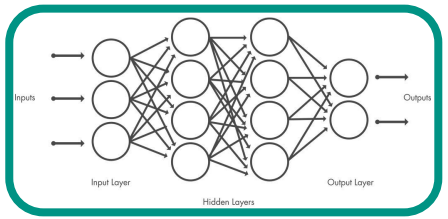
Yes/No

What's a GAN?

The two networks play against each other

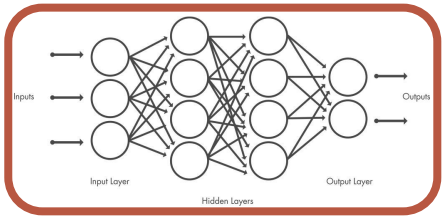
Real images

RNG



Generator

Discriminator



"Real" or "Fake"

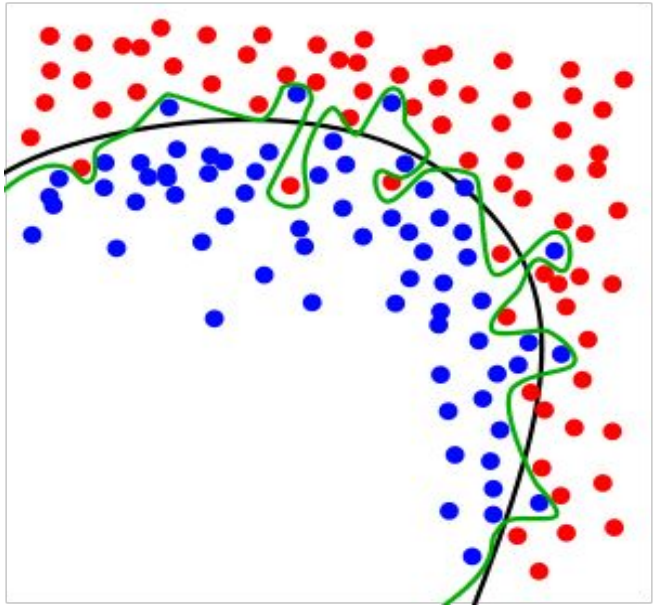
What's a GAN?

Only one network wins, they try to beat each other by learning what works best

Both get better over time during learning

Difficulties

- Overfitting (= one of the networks learns exactly the dataset)
- Slow convergence
- Computing power
- Mode collapse



Attacks on StyleGAN

How to detect GAN-generated images?

- Use StyleGAN's discriminator?
- Train our own?

If there were an easy way, the discriminator would have found it during training

Attacks on StyleGAN

GANs are based on a game, so

let's cheat!

Attacks on StyleGAN

Two main ways to cheat :

- Exploit the game rules
- Reverse-engineer the GAN

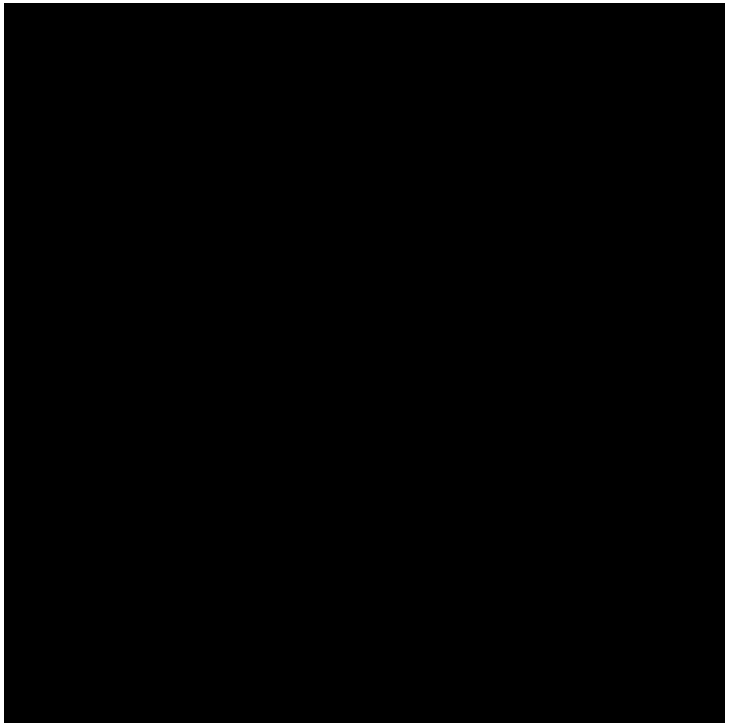
... and we're gonna do both!

Attacks on StyleGAN

Which one is real?

Attacks on StyleGAN

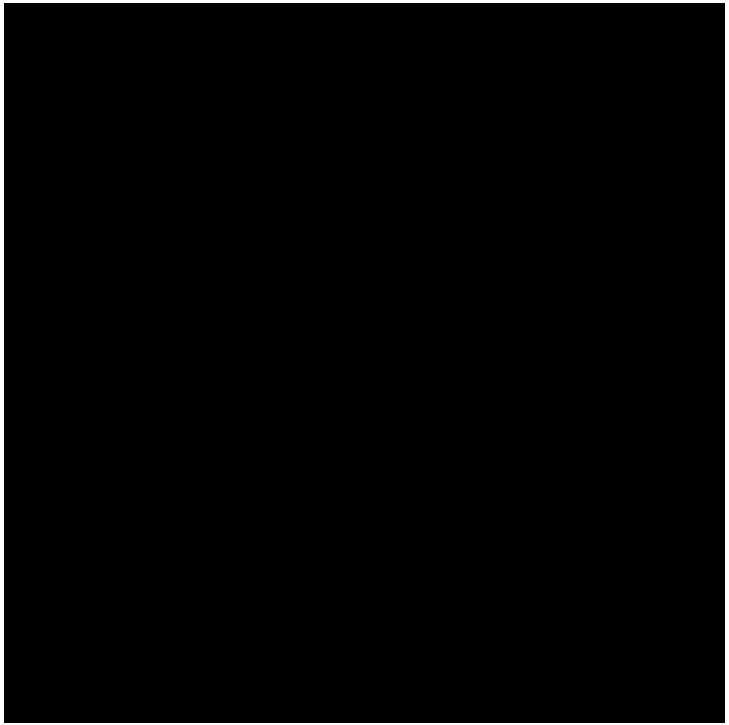
FFHQ Dataset



Attacks on StyleGAN

FFHQ Dataset

\+ fixed eye overlay



Attacks on StyleGAN

StyleGAN is trying to reproduce FFHQ, which has strong specificities!

Easy to pre-detect using eyes only (not new)

Attacks on StyleGAN

~~Exploit the game rules~~ ✅

Reverse-engineer the GAN

New exploit

```python
import hashlib
import requests


URL = 'https://thispersondoesnotexist.com/image'

while True:
    req = requests.get(URL)
    hsh = hashlib.sha256(req.content).hexdigest()[:10]
    print(hsh)
```

```
9478cd2071
9732e52dba
9732e52dba
9b977a38a9
9046a824a2
9046a824a2
02da30b9ac
02da30b9ac
02da30b9ac
e74ab01e31
e74ab01e31
4e084164bb
4e084164bb
2c1cc63056
2c1cc63056
```

New exploit

Images repeat several times!

There is a global cache, every single response within ~1.2s is identical

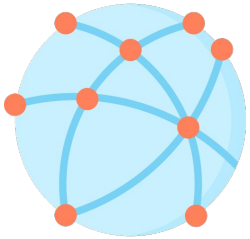
New exploit

What does this mean?

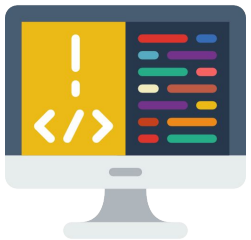
- ThisPersonDoesNotExist serves a single image to everyone, and changes it often
- It's possible to download all images!

New exploit

ThisPersonDoesNotExist.com



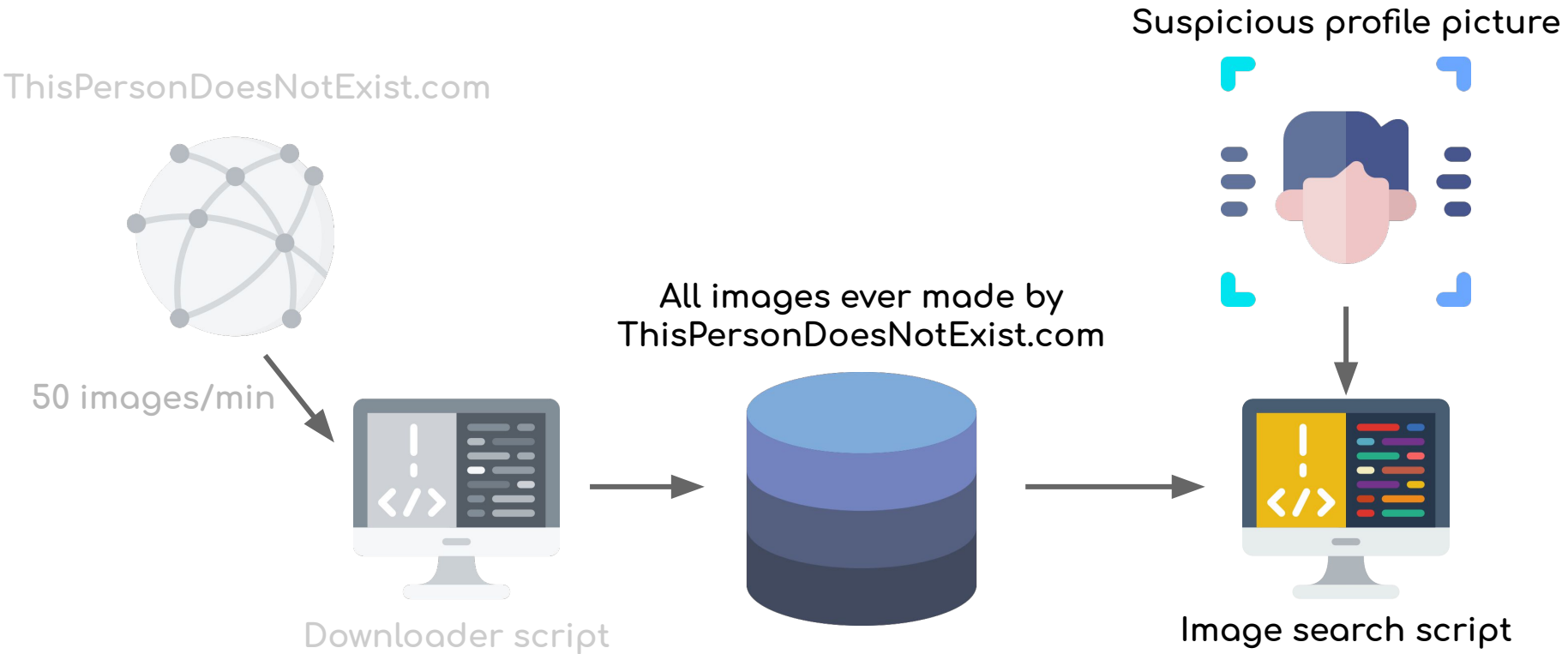
50 images/min



Downloader script

All images ever made by
ThisPersonDoesNotExist.com

New exploit


Suspicious profile picture
ThisPersonDoesNotExist.com
All images ever made by
ThisPersonDoesNotExist.com
50 images/min
Downloader script
Image search script

Looks easy, right?

Optimizing the attack

- Around 72k new images per day
- 450 kB on average as jpg

Total volume is 32GB/day (or 11TB/year!)

Optimizing the attack

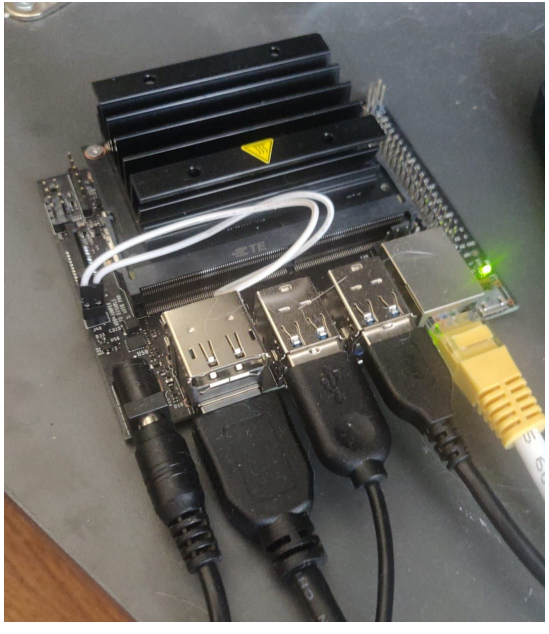
Two main issues to solve:

- Storage space (obviously)
- Lookup time (linear search is very slow)

Optimizing the attack

Additional hardware constraints:

- 256GB HDD
- 4GB RAM
- Slow CPU
- But we have GPU!



Optimizing the attack

Why not leverage neural networks?

FaceNet turns face pictures into vectors:

- Two faces of the same person have similar vector values
- Two different faces are far apart

Optimizing the attack

One vector is 512 float values, or 2kB

→ 225:1 disk size reduction!

And fast to compute using GPU

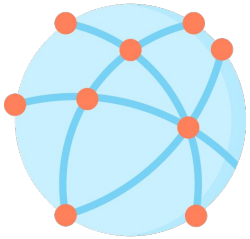
Optimizing the attack

Lookup is also much easier to do : Elasticsearch OpenDistro has fast k-NN search for vectors

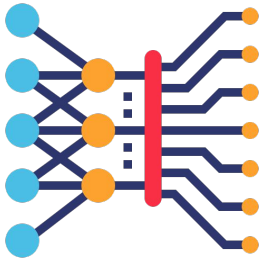
(this means we can index and search for similar FaceNet vectors easily)

Optimizing the attack

ThisPersonDoesNotExist.com



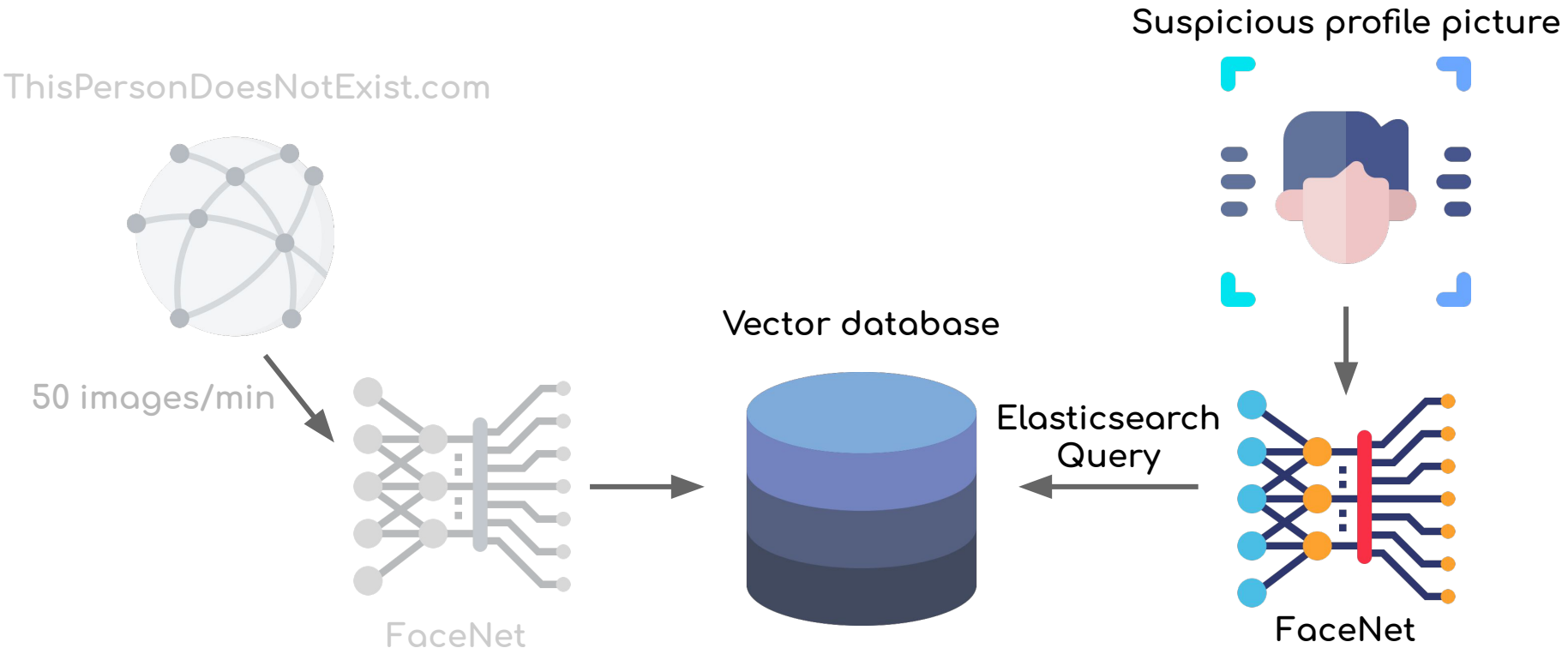
50 images/min



FaceNet

Vector database

Optimizing the attack



Optimizing the attack

Bonus points #1 :

This technique is very resistant to many transformations such as cropping, noise, compression, rotation, ...

Optimizing the attack

Bonus points #2 :

Each image is unique and appears for only 1.2s. By indexing the timestamp with the vector, **we can also retrieve the date** at which the picture was generated!

Demo time

It’s demo time! Meet FADA

Demo time

- 1.5M+ faces already indexed, <1s search
- Available for free (soon)

  fada.h25.io

  (plz don't melt my GPU)
- Open-source

  github.com/MathisHammel/FADA

How to dodge FADA

I gave you a good discriminator,
now let me give you a good generator!

How to dodge FADA

StyleGAN2 is hard to train, but you can use it pre-trained

→ this completely avoids FADA detection

How to dodge FADA

Bonus points #3:
You can now have several variants

Thanks !

Any questions ?


@MathisHammel / @h25io


discord.h25.io


twitch.h25.io