

Abuses and misuses of AI: prevention vs reaction

Red Teaming in the AI world

Cristian Canton Ferrer
Research Manager (AI Red Team @ Facebook)

Abuses and misuses of AI: prevention vs reaction

Red Teaming in the AI world
...with Manipulated Media as an example

Cristian Canton Ferrer
Research Manager (AI Red Team @ Facebook)

Outline

Introduction

Abuses

Misuses

Prevention

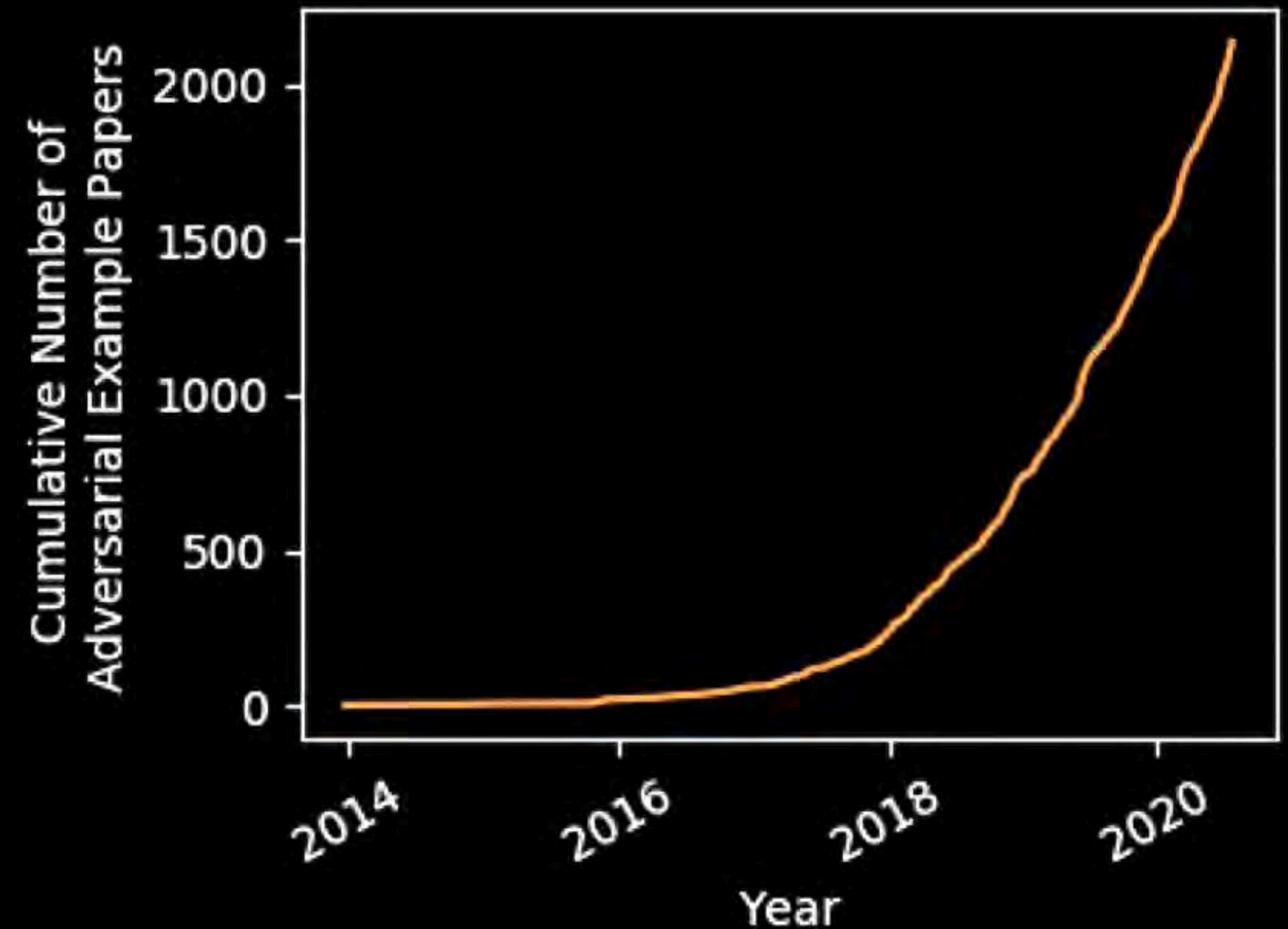
Reaction and Mitigation

Introduction



What is the current situation of AI?

Research on adversarial attacks has growth since the advent of DNNs



Credits: Nicolas Carlini for the graph (<https://nicholas.carlini.com/>)

Adversarial attack \Rightarrow GAN

Input image
Category: Panda (57.7% confidence)



+

Adversarial noise



=

Attacked image
Category: Gibbon (99.3% confidence)



Abuse of an AI system to force it to make a calculated mistake

What is a Red Team?

What is a Red Team?

"A Red Team is a group that helps organizations to improve themselves by providing opposition to the point of view of the organization that they are helping."

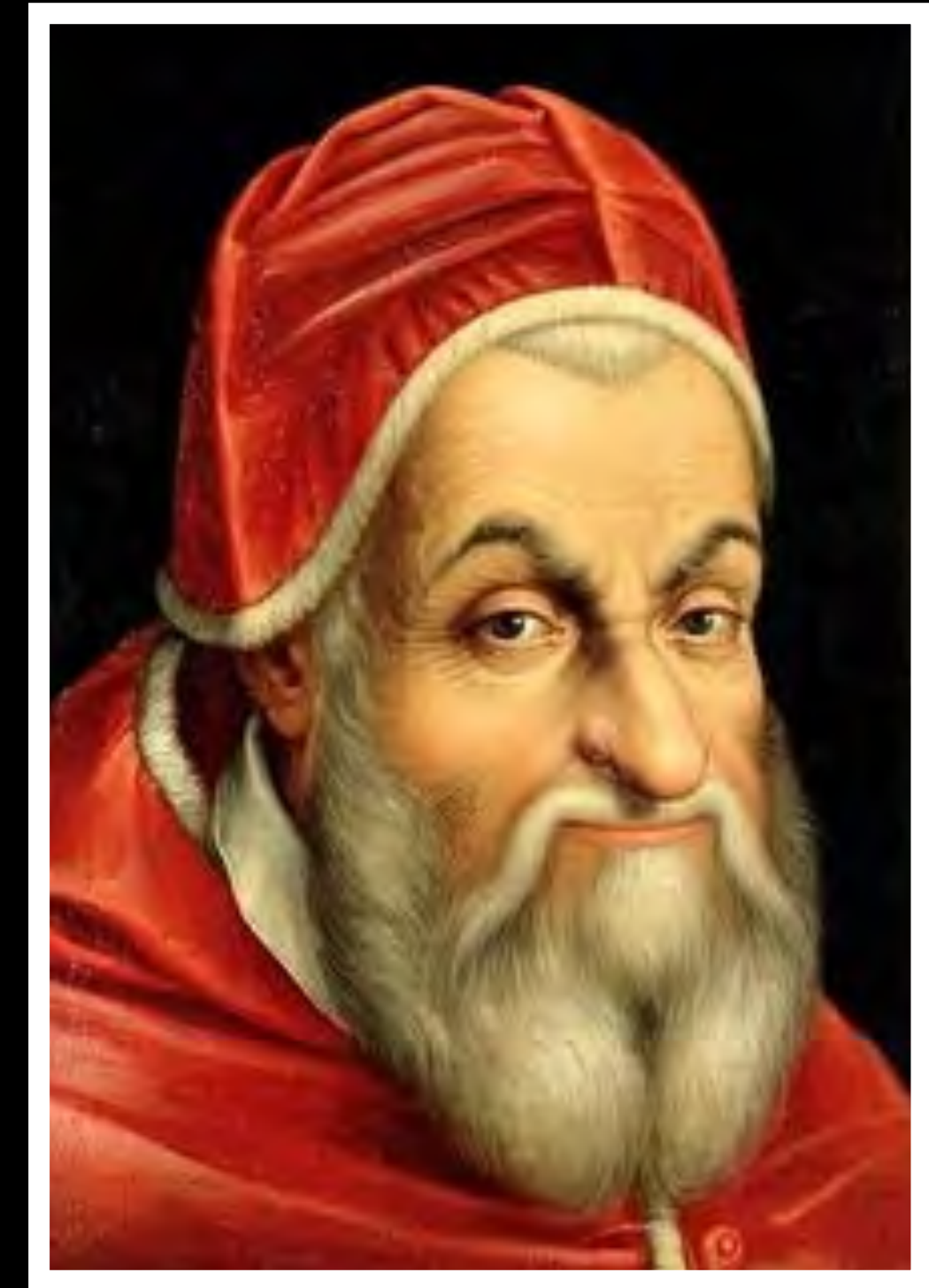
Wikipedia



What is a Red Team?

At the origin, everything started with the:

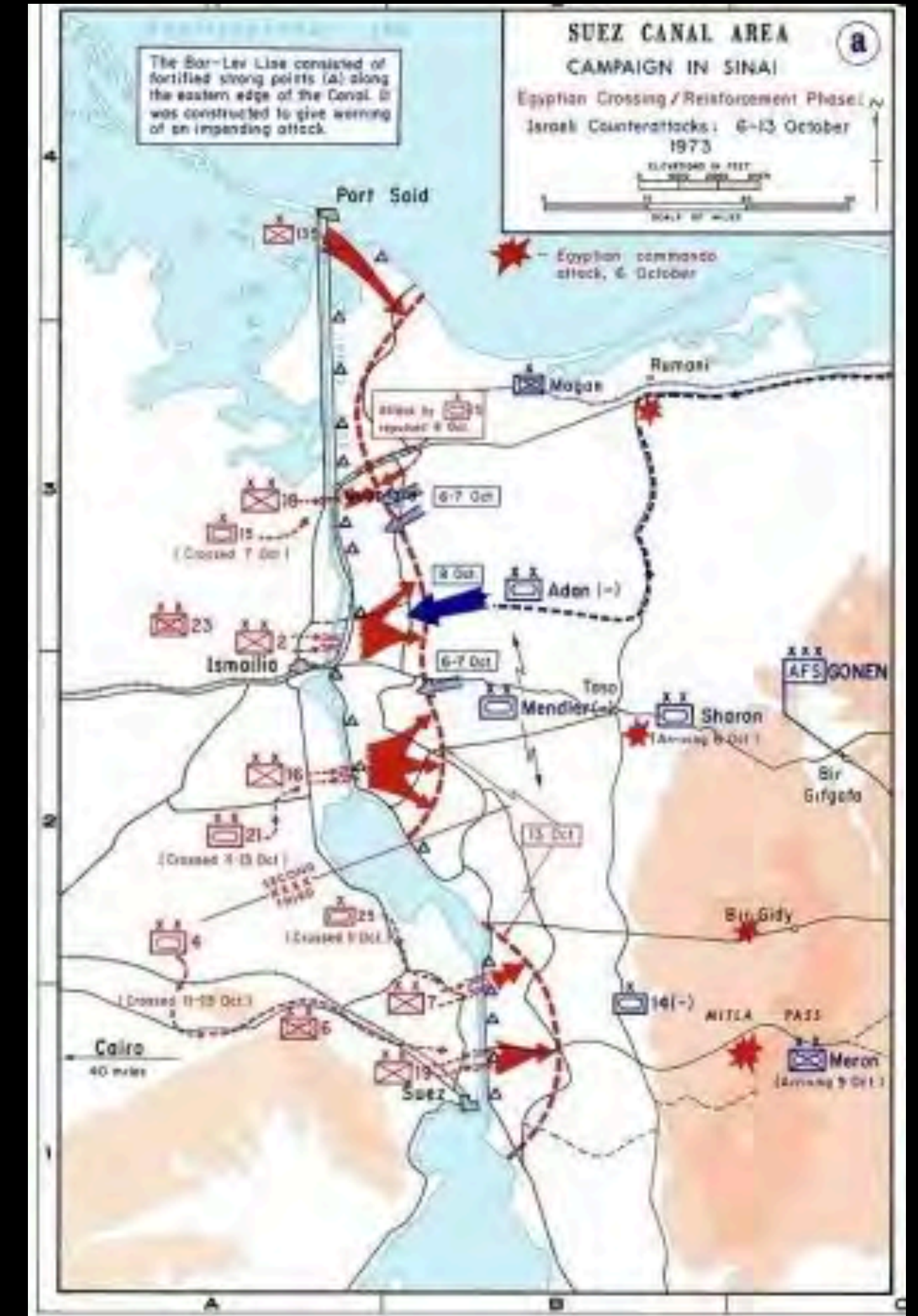
"Advocatus Diaboli"



Pope Sixtus V (1521-1590)

What is a Red Team?

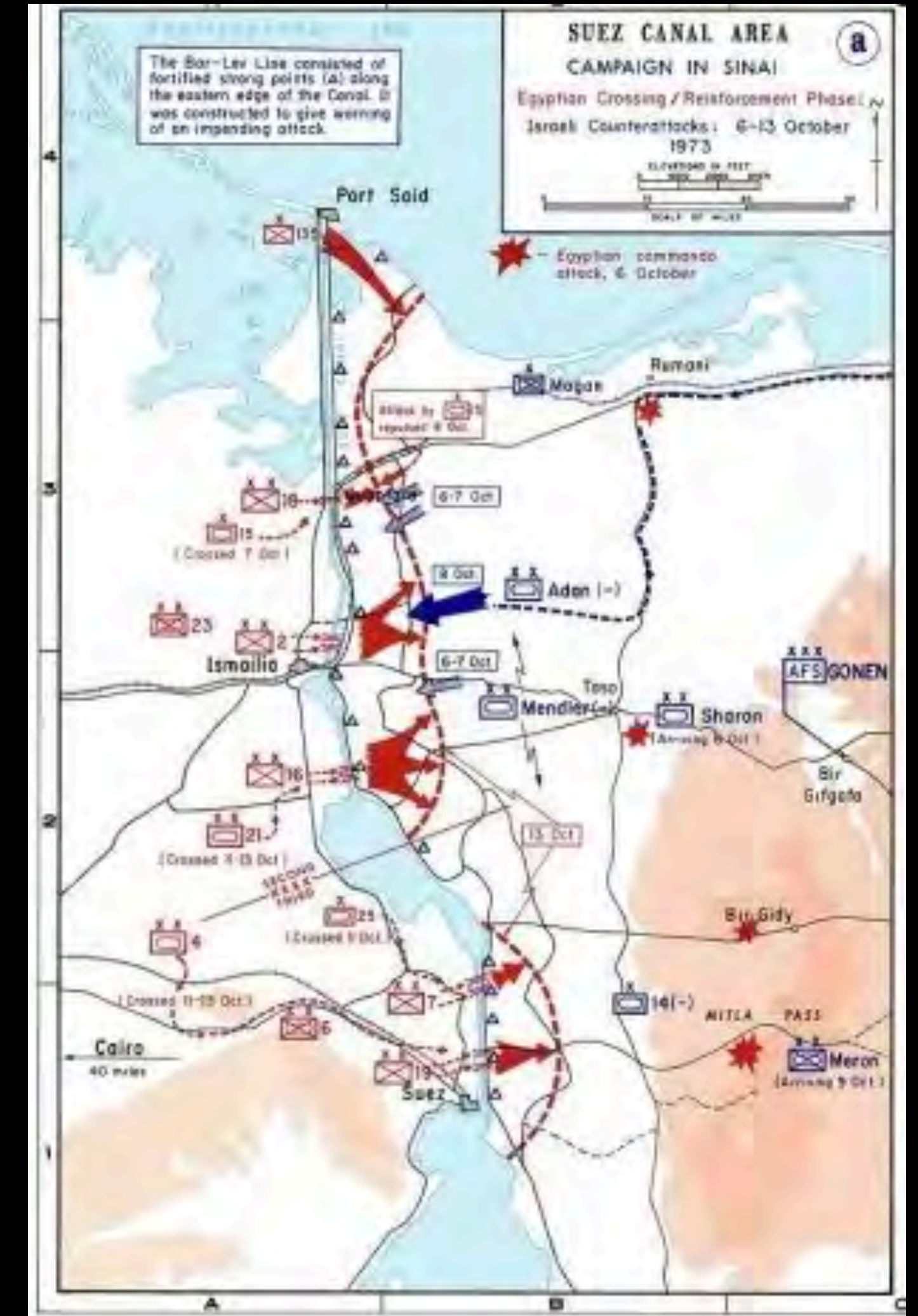
The advent of Red Teaming in the modern era:
The Yom Kippur War and the 10th Man Rule



What is a Red Team?

The advent of Red Teaming in the modern era:
The Yom Kippur War and the 10th Man Rule

Bryce G. Hoffman, "Red Teaming", 2017.
Micah Zenko, "Red Team", 2015.



What does an AI Red Team do?

- Bring the "loyal" adversarial mentality into the AI world, specially for systems in production

What does an AI Red Team do?

- Bring the "loyal" adversarial mentality into the AI world, specially for systems in production
- Understand the risk landscape of your company

What does an AI Red Team do?

- Bring the "loyal" adversarial mentality into the AI world, specially for systems in production
- Understand the risk landscape of your company
- Identify, evaluate and prioritize risks and feasible attacks

What does an AI Red Team do?

- Bring the "loyal" adversarial mentality into the AI world, specially for systems in production
- Understand the risk landscape of your company
- Identify, evaluate and prioritize risks and feasible attacks
- Conceive worst case scenarios derived from abuses and misuses of AI

What does an AI Red Team do?

- Bring the "loyal" adversarial mentality into the AI world, specially for systems in production
- Understand the risk landscape of your company
- Identify, evaluate and prioritize risks and feasible attacks
- Conceive worst case scenarios derived from abuses and misuses of AI
- Conform a group of experts across all involved aspects of a real system

What does an AI Red Team do?

- Bring the "loyal" adversarial mentality into the AI world, specially for systems in production
- Understand the risk landscape of your company
- Identify, evaluate and prioritize risks and feasible attacks
- Conceive worst case scenarios derived from abuses and misuses of AI
- Conform a group of experts across all involved aspects of a real system
- Convince stakeholders of the importance and *potential* impact of a worst case scenario and ideate solutions: preventions or mitigations

What does an AI Red Team do?

- Bring the "loyal" adversarial mentality into the AI world, specially for systems in production
- Understand the risk landscape of your company
- Identify, evaluate and prioritize risks and feasible attacks
- Conceive worst case scenarios derived from abuses and misuses of AI
- Conform a group of experts across all involved aspects of a real system
- Convince stakeholders of the importance and *potential* impact of a worst case scenario and ideate solutions: preventions or mitigations
- Define iterative and periodic interactions with stakeholders

What does an AI Red Team do?

- Bring the "loyal" adversarial mentality into the AI world, specially for systems in production
- Understand the risk landscape of your company
- Identify, evaluate and prioritize risks and feasible attacks
- Conceive worst case scenarios derived from abuses and misuses of AI
- Conform a group of experts across all involved aspects of a real system
- Convince stakeholders of the importance and *potential* impact of a worst case scenario and ideate solutions: preventions or mitigations
- Define iterative and periodic interactions with stakeholders
- Defenses? No: that's for the blue team!



Red Queen Dynamics

"...it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!"

Lewis Carroll, Through the Looking-Glass



Red Queen Dynamics

$$\frac{\partial \text{Attacks}}{\partial t} > \frac{\partial \text{Defenses}}{\partial t}$$

Risk estimation

$$\text{AI Risk} = \text{Severity} \times \text{Likelihood}$$

Risk estimation

AI Risk = Severity x Likelihood

- Core metrics for your company
- Financial
- Data leakage, privacy
- PR
- Human
- Mitigation cost, response time
- ...

Risk estimation

AI Risk = Severity x Likelihood

- Discoverability
- Implementation cost / Feasibility
- Motivation
- ...

Risk estimation

$$\text{AI Risk} = \text{Severity} \times \text{Likelihood}$$

		Severity			
		Catastrophic: 4	Critical: 3	Moderate: 2	Marginal: 1
Probability	Frequent: 5	High - 20	High - 15	High - 10	Medium - 5
	Probable: 4	High - 16	High - 12	Serious - 8	Medium - 4
	Occasional: 3	High - 12	Serious - 9	Medium - 6	Low - 3
	Remote: 2	Serious - 8	Medium - 6	Medium - 4	Low - 2
	Improbable: 1	Medium - 4	Low - 3	Low - 2	Low - 1

A first (real) example



This is "objectionable content" (99%)

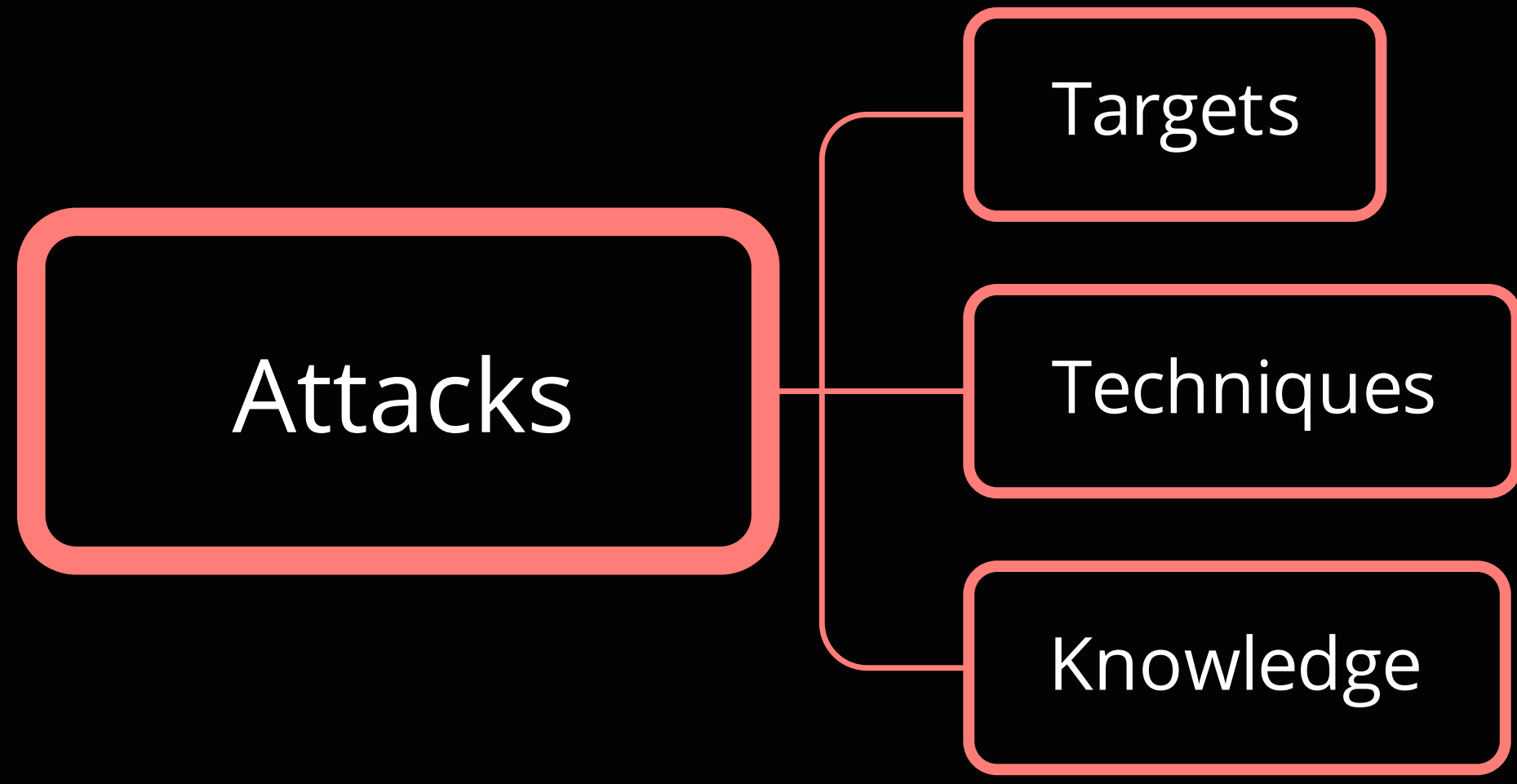
A first (real) example

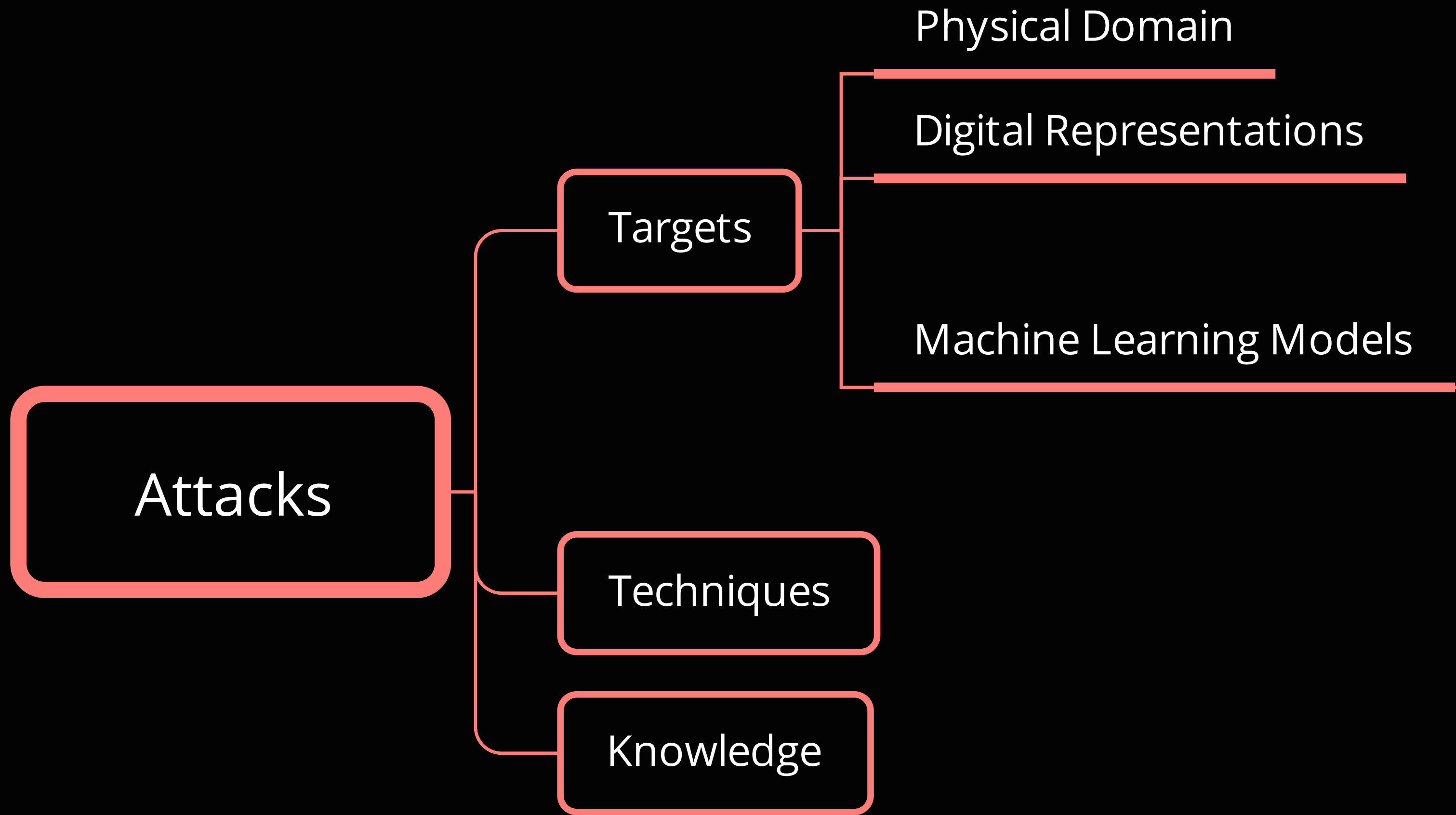


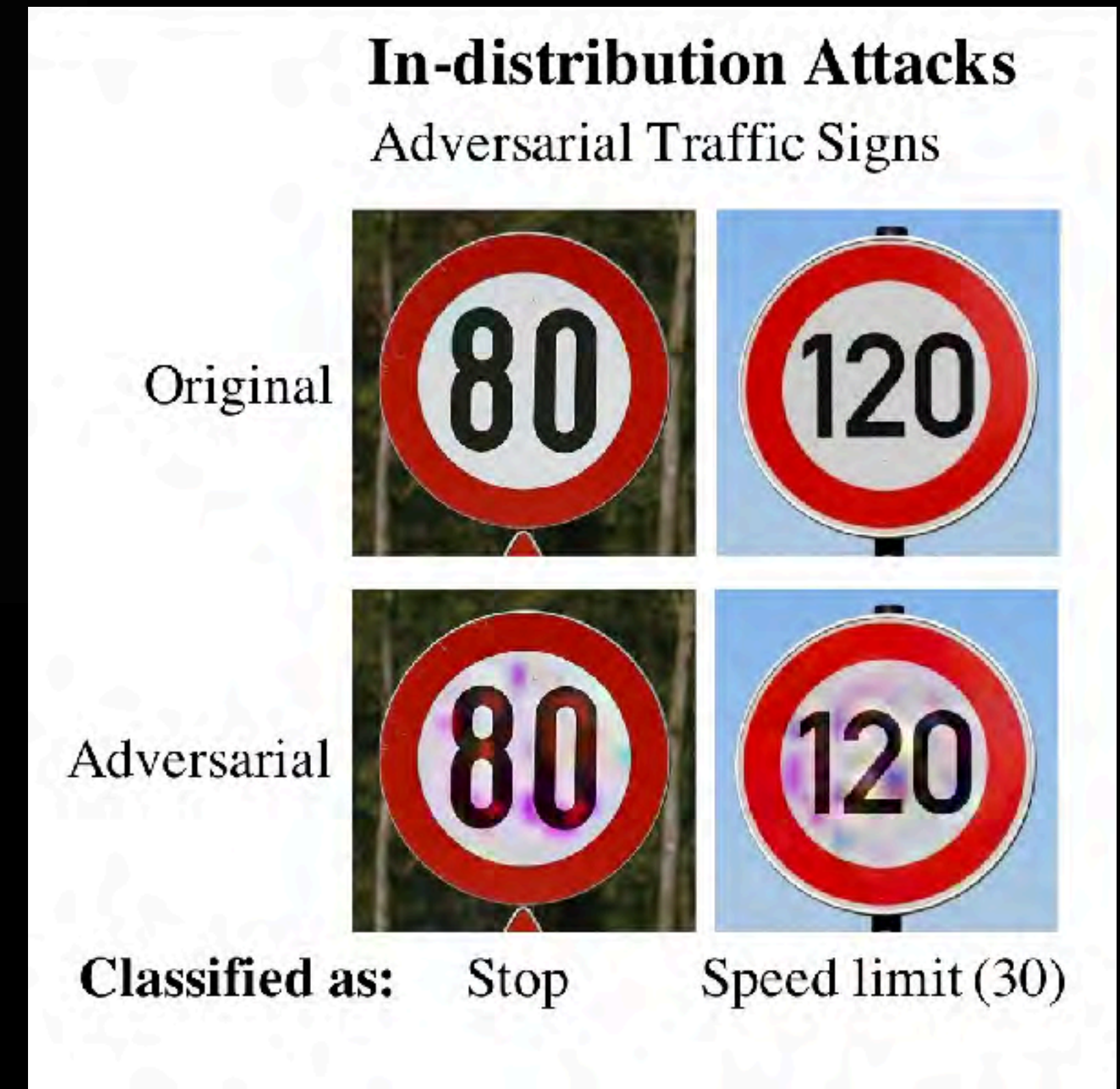
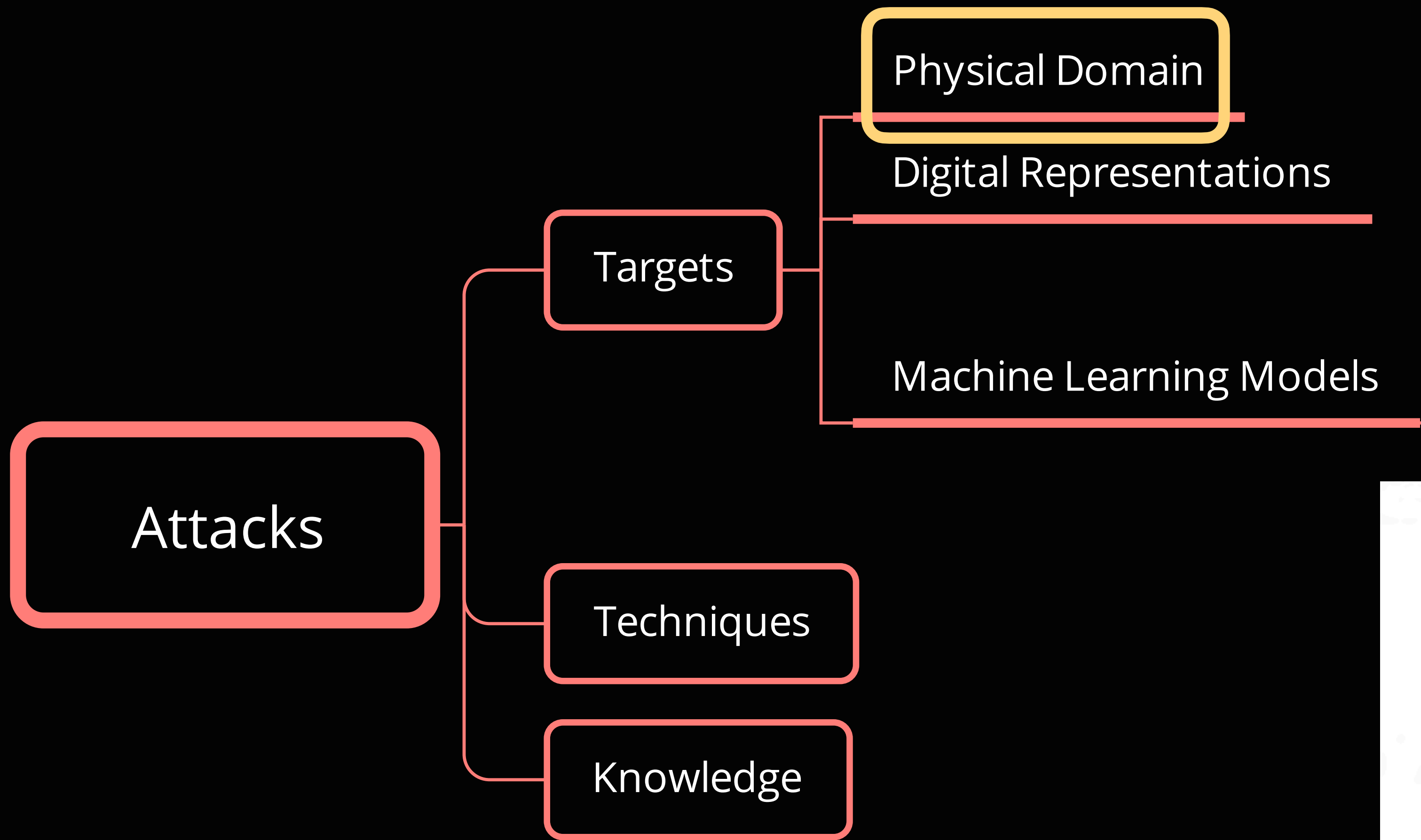
This is safe content (95%)

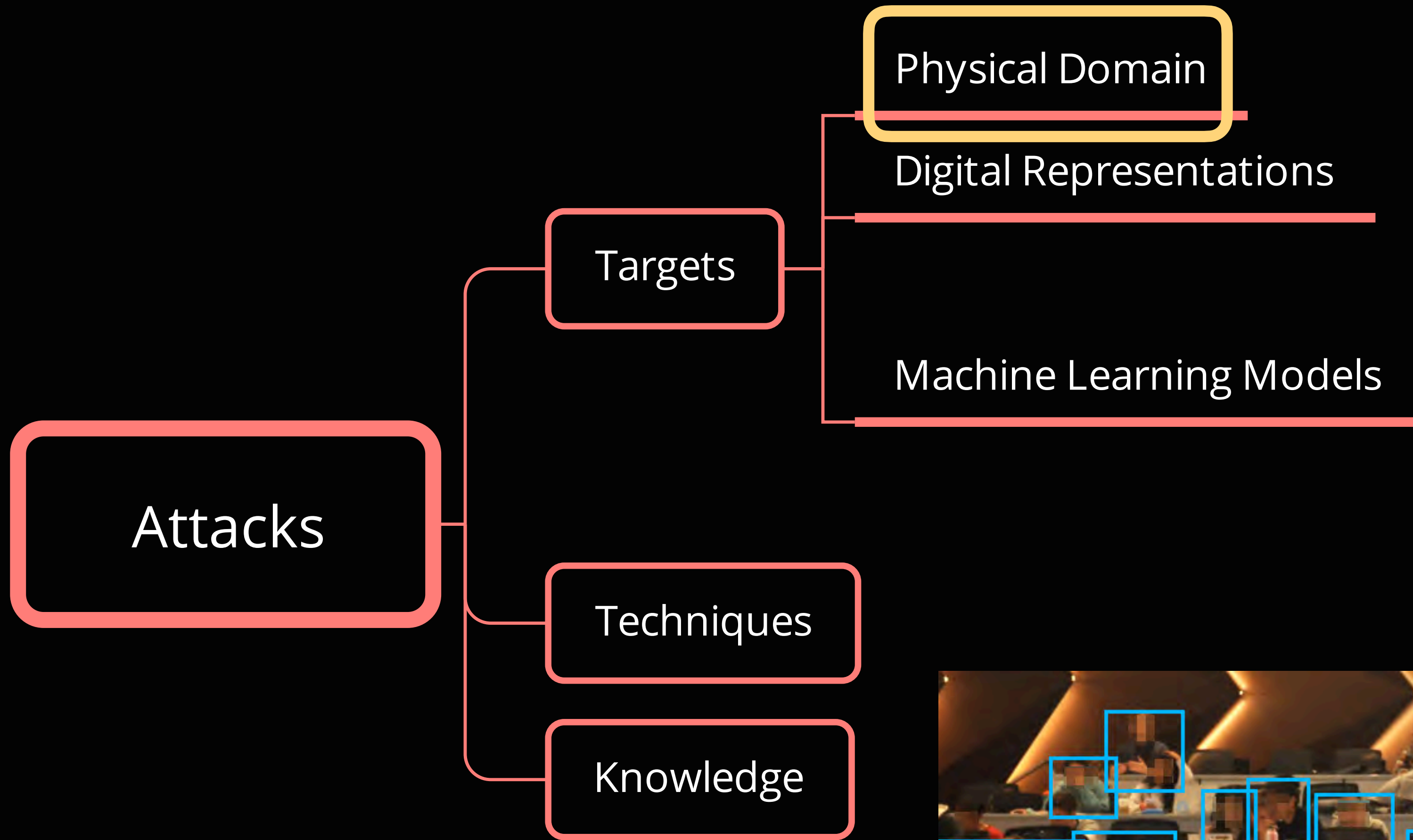
Abuses

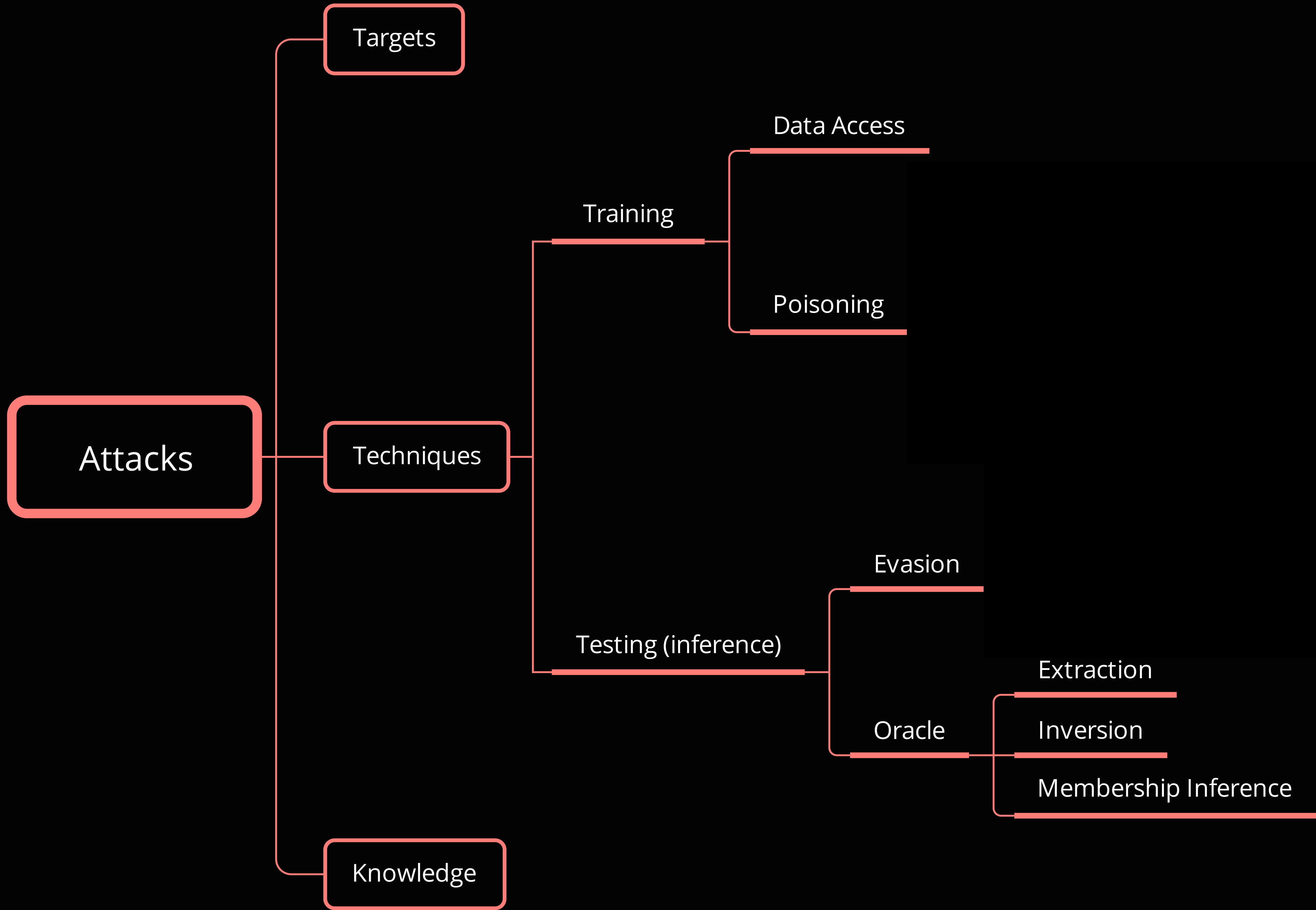


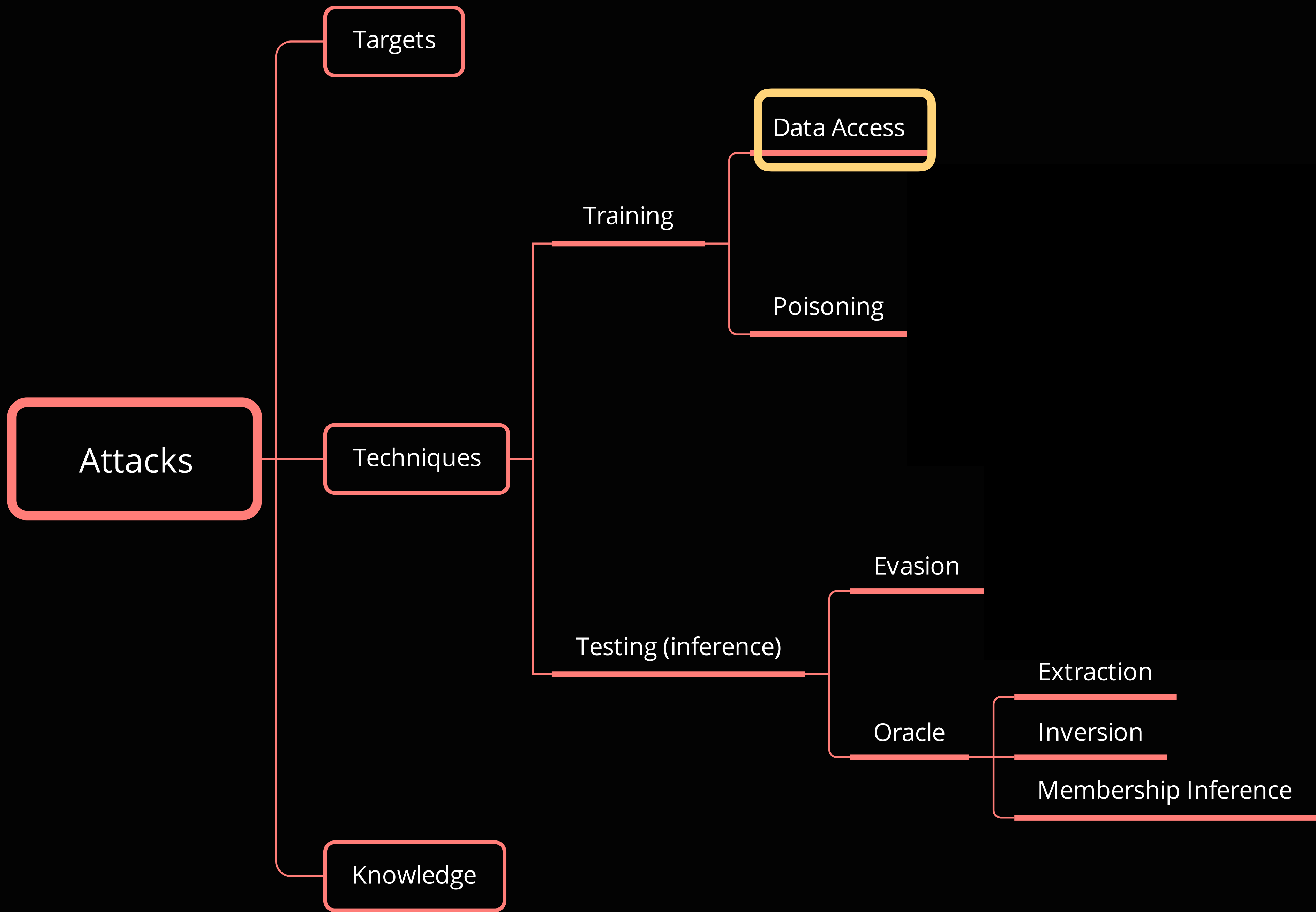




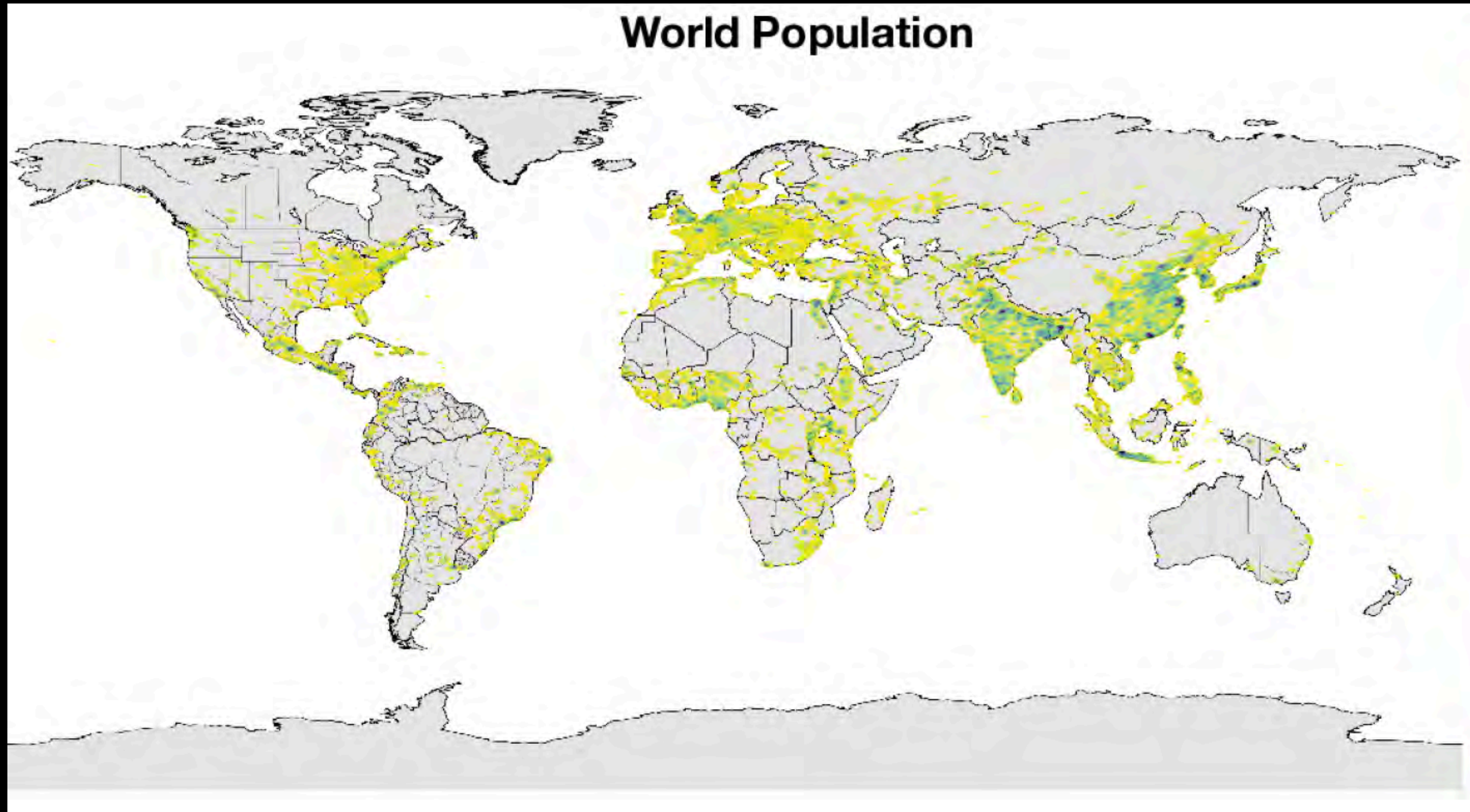




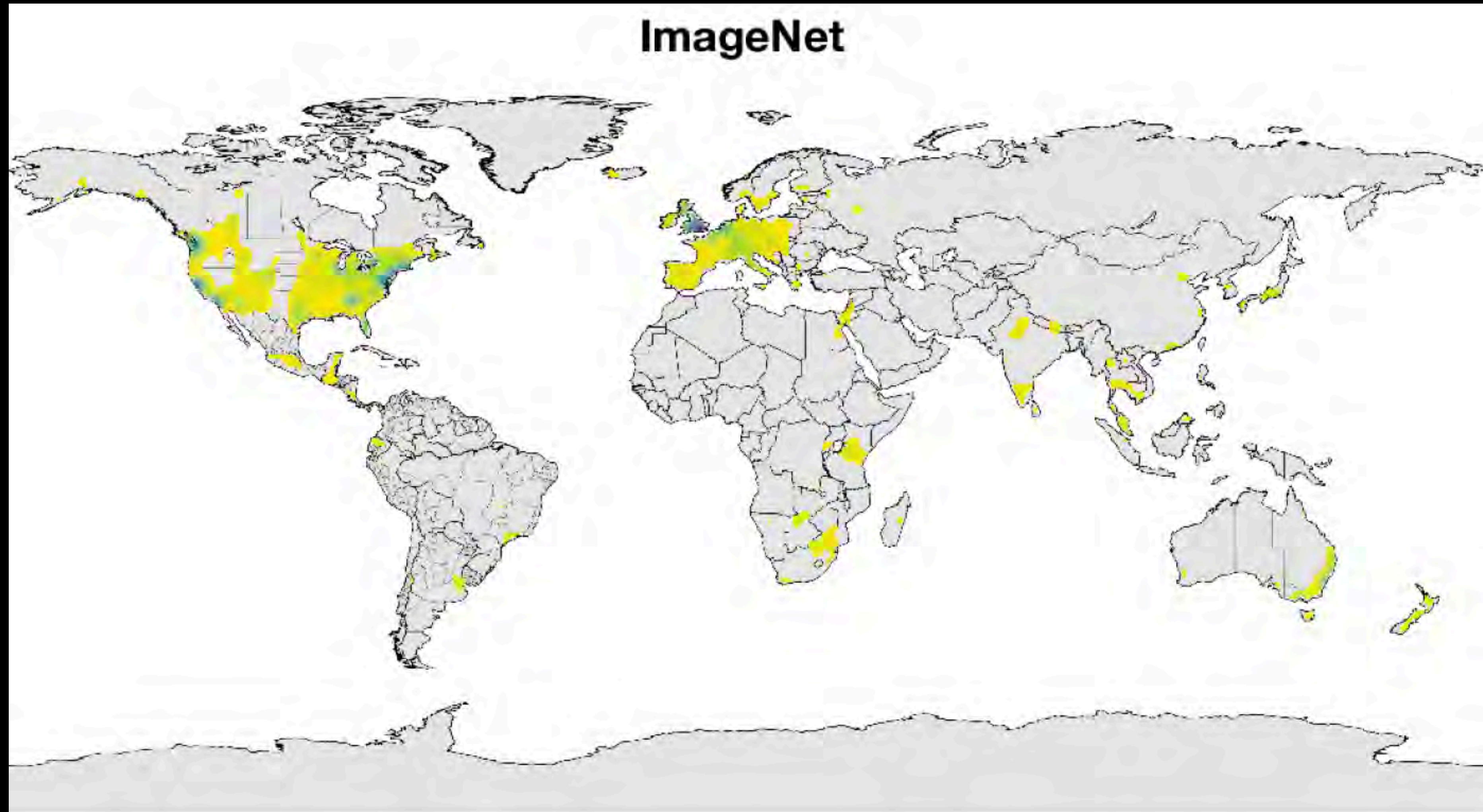




Attacking dataset biases

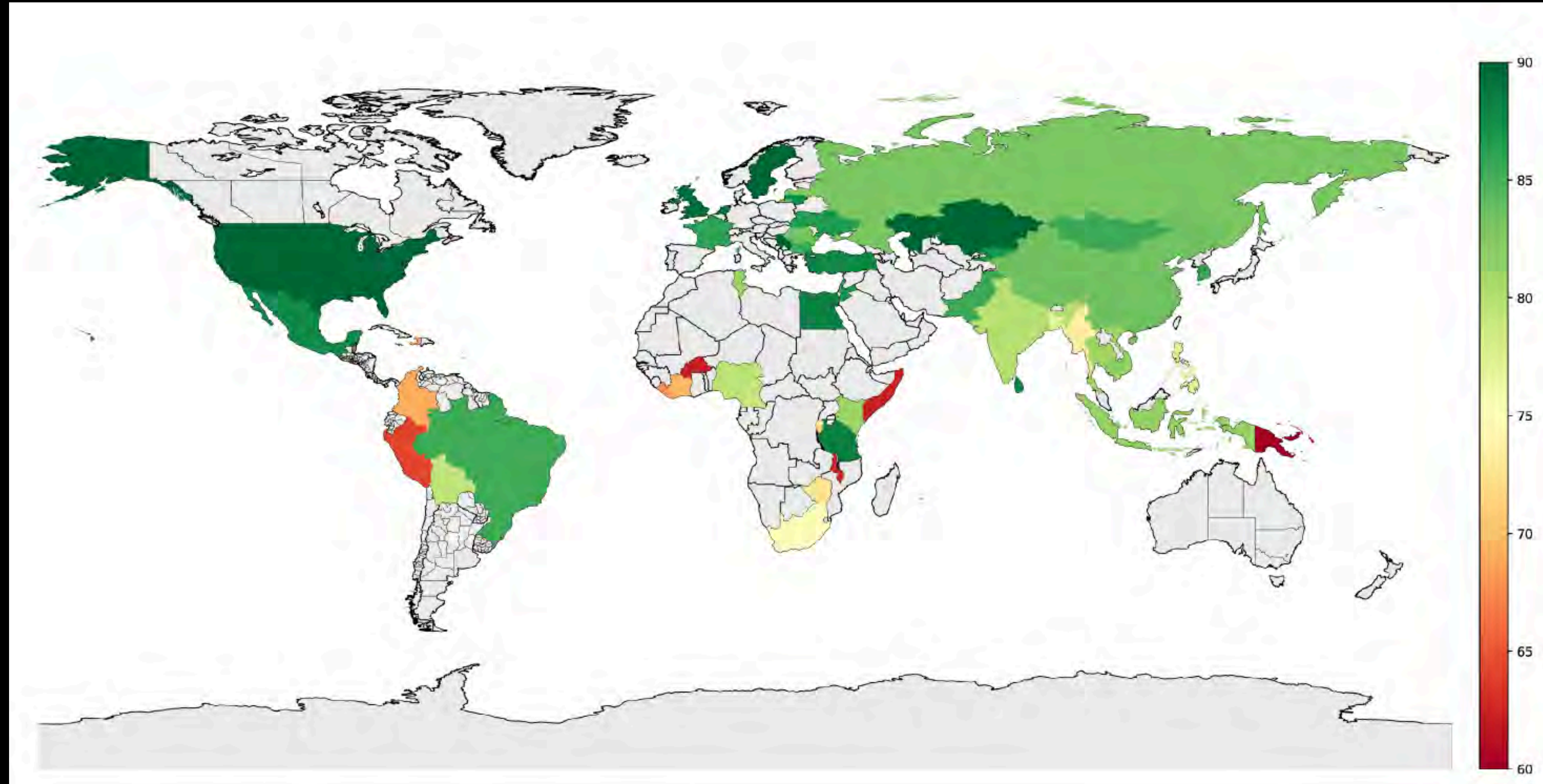


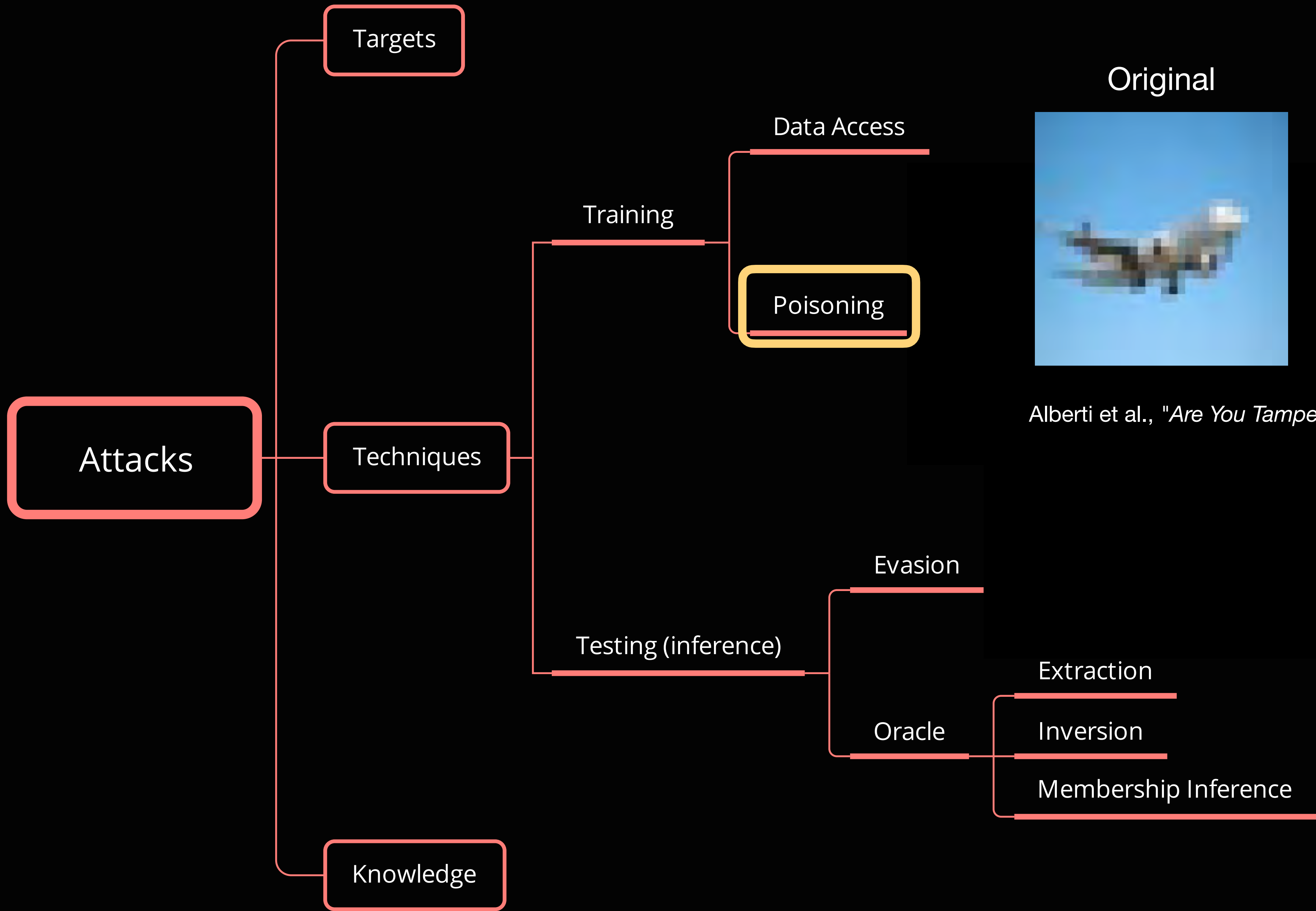
Attacking dataset biases



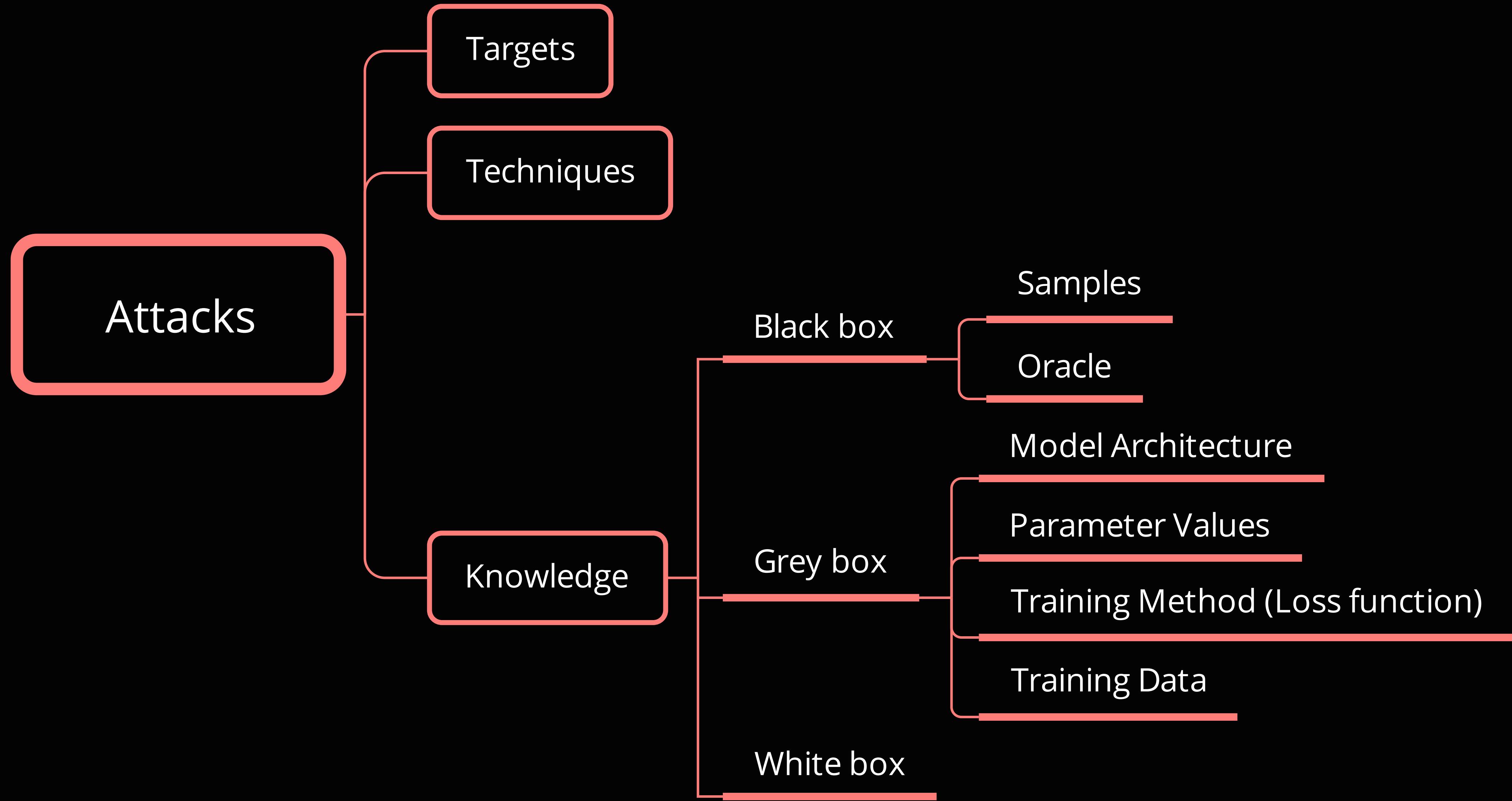
Attacking dataset biases

Geographical distribution of classification accuracy





Alberti et al., "Are You Tampering With My Data?", 2018.

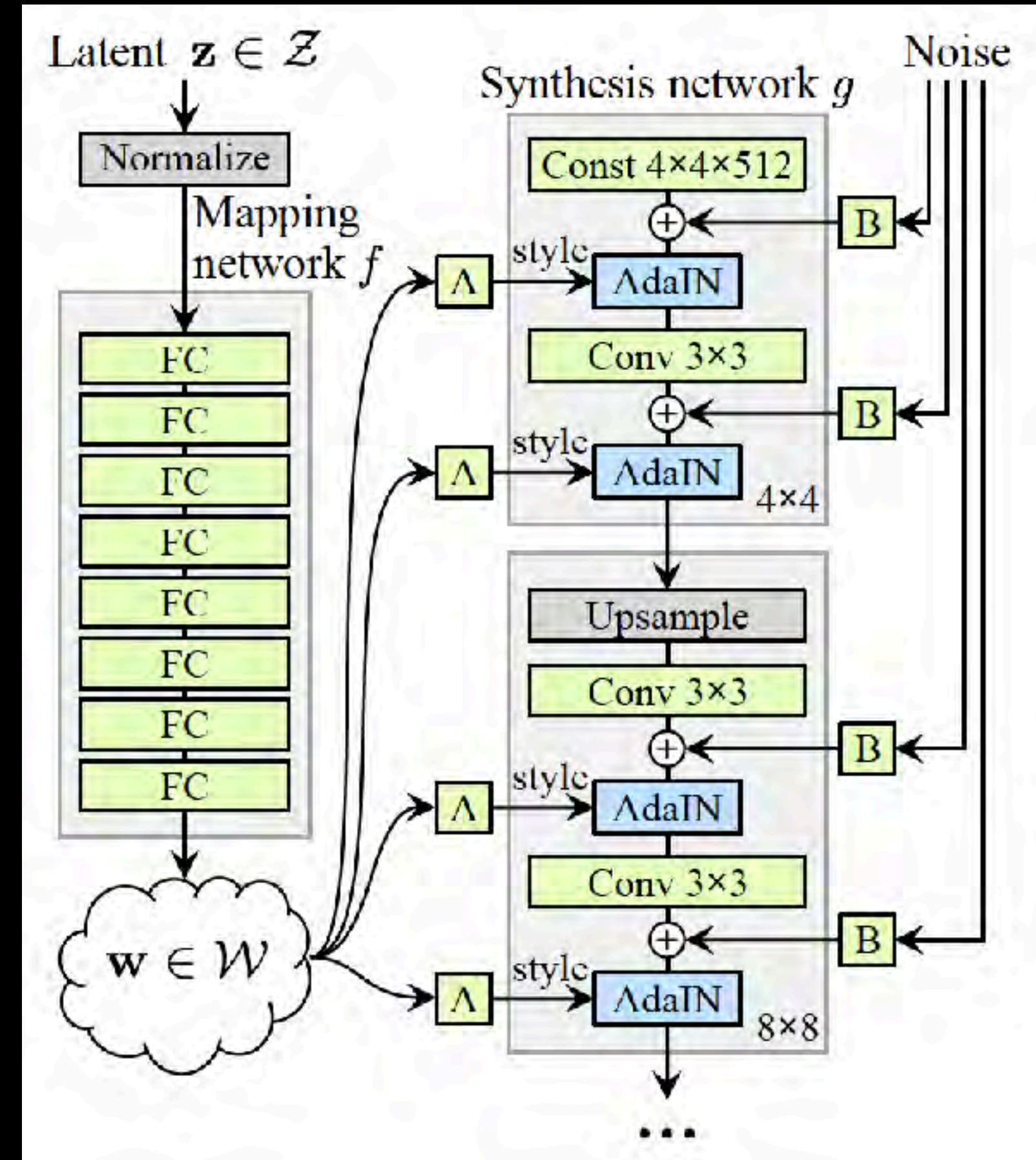
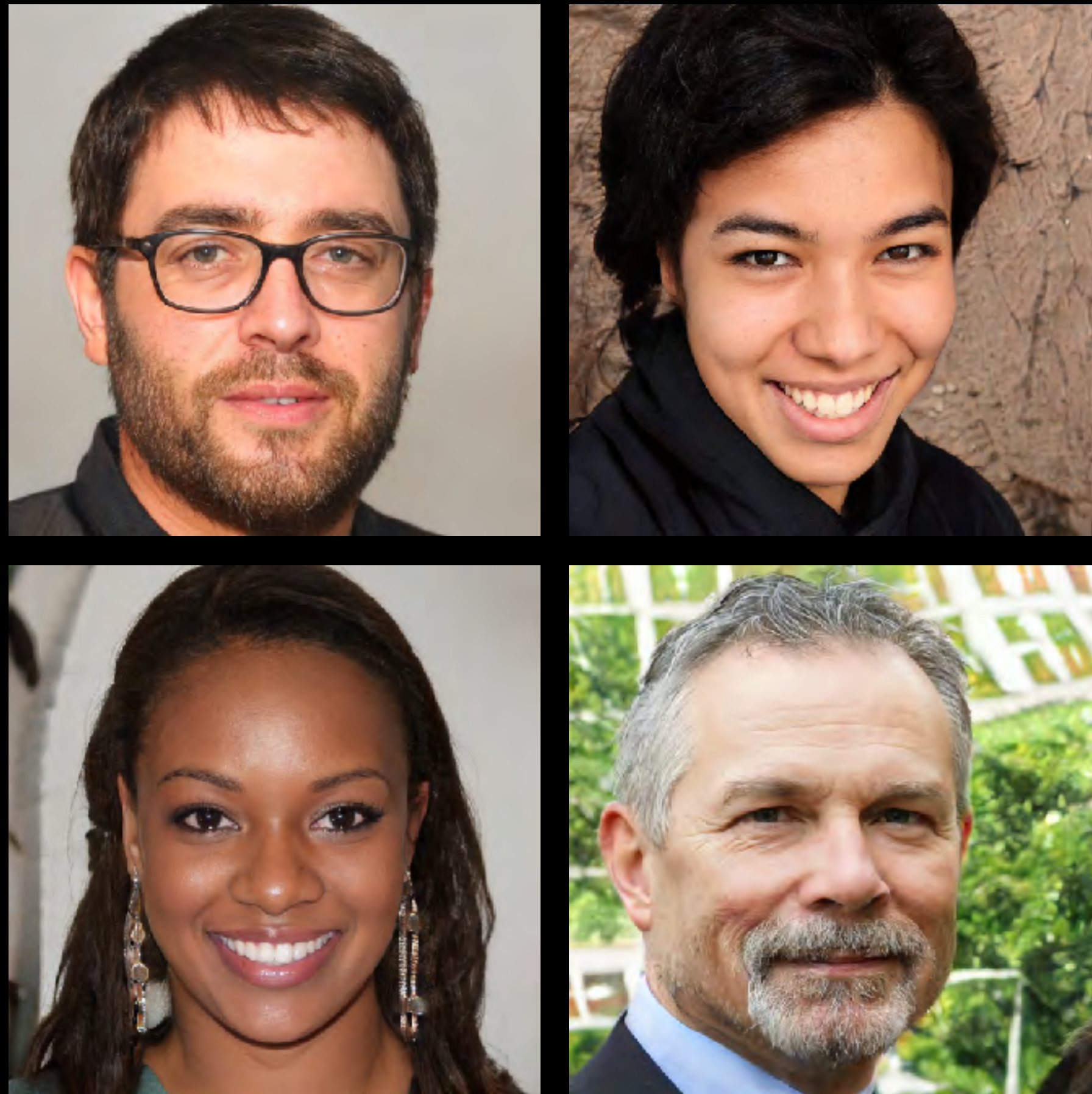


Misuses



Example case: Synthetic people

Disclaimer: None of these individuals exist!



StyleGAN

Karras et al. "A Style-Based Generator Architecture for Generative Adversarial Networks", 2019.
Karras et al. "Analyzing and Improving the Image Quality of StyleGAN", 2020.

Example case: Synthetic people

Disclaimer: None of these individuals exist!

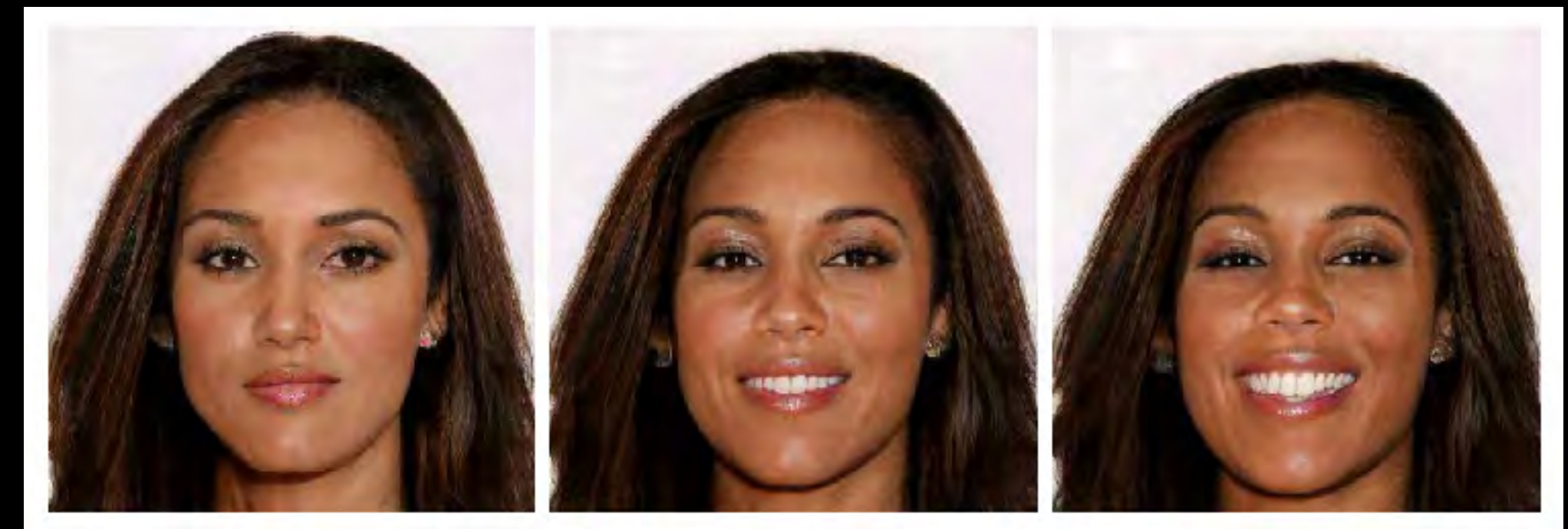


Plenty of potential good uses:

- Creative purposes
- Virtual characters
- Semantic face editing



Smile edition



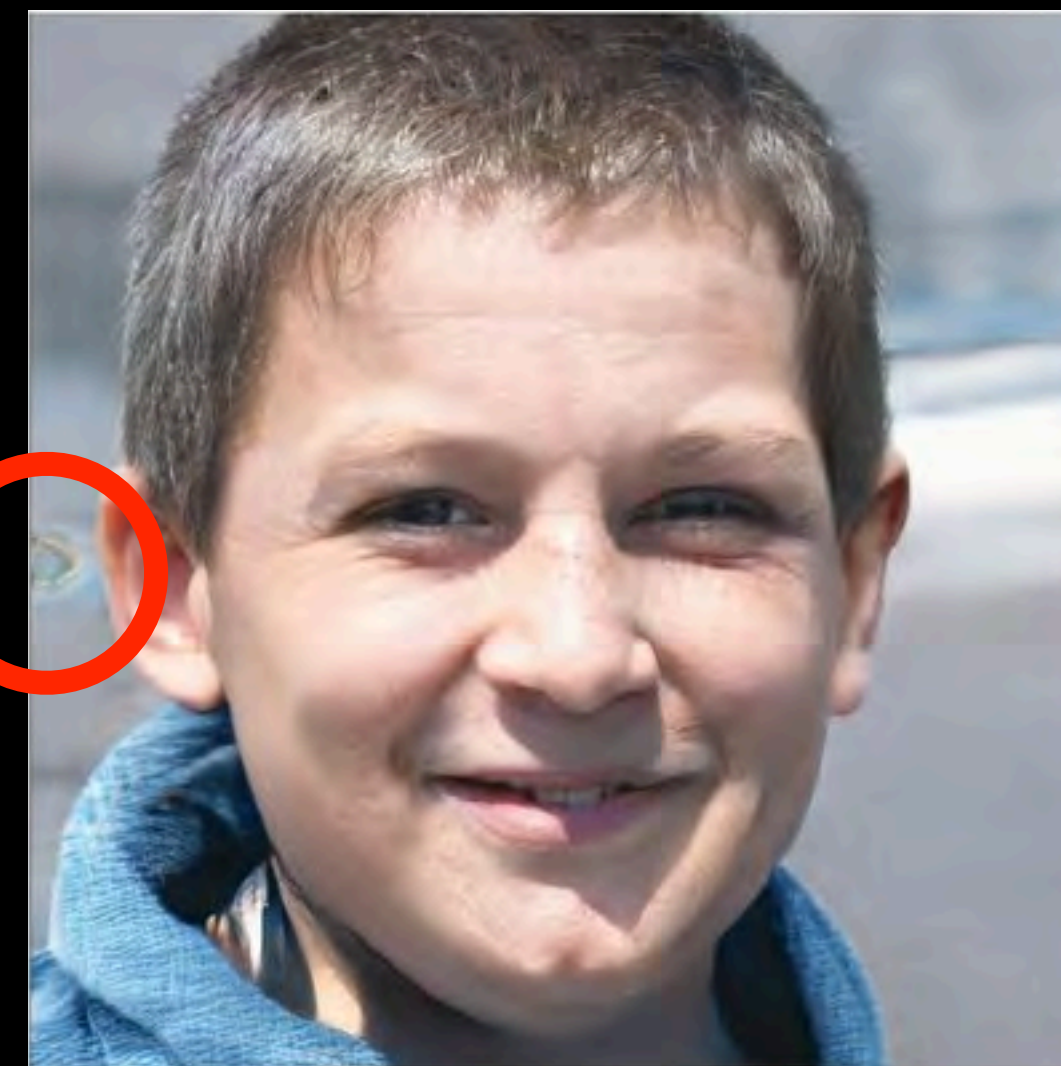
Example case: Synthetic people

Disclaimer: None of these individuals exist!



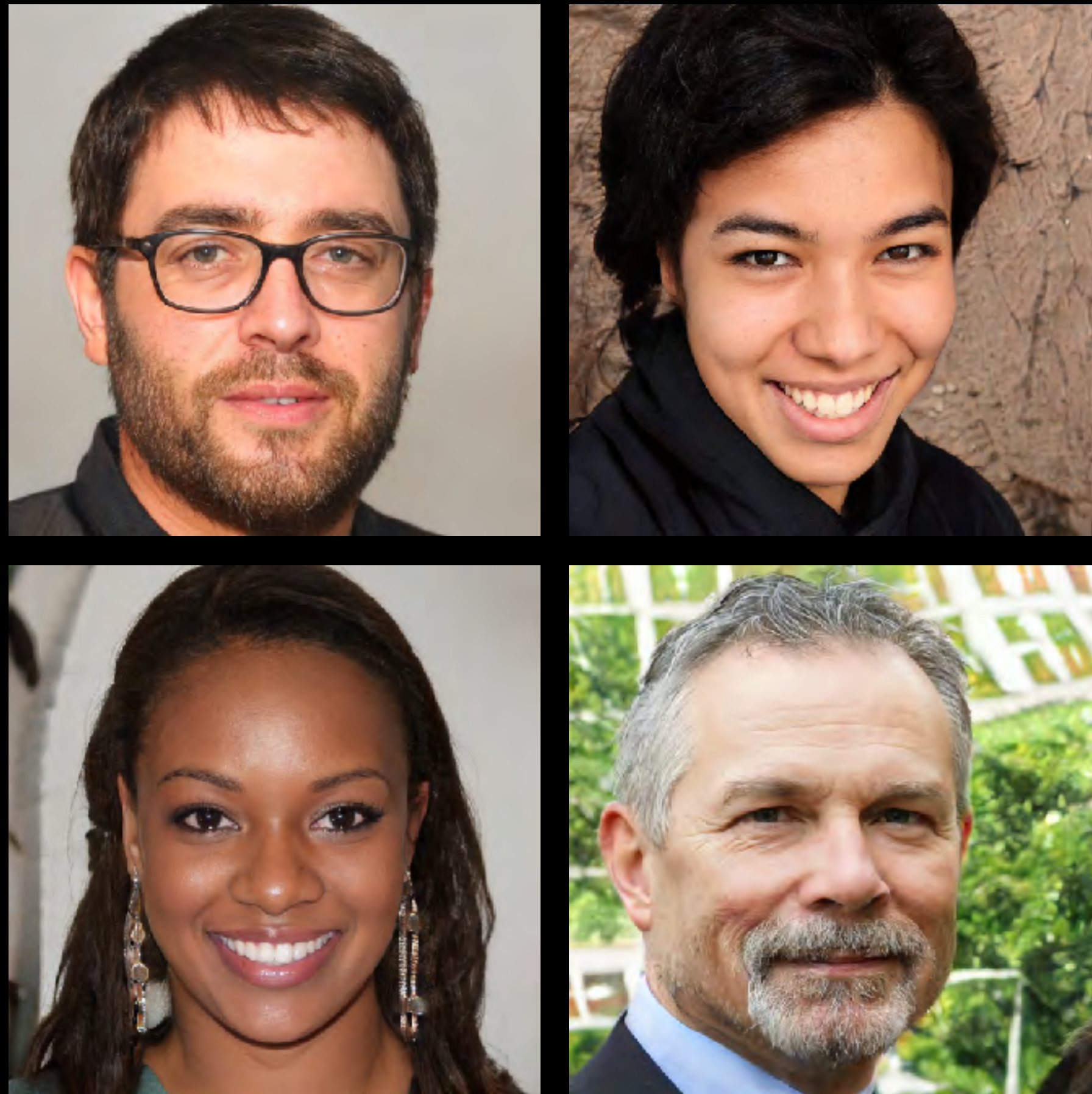
Potentially "easy" to spot:

- Generator residuals (in the image)



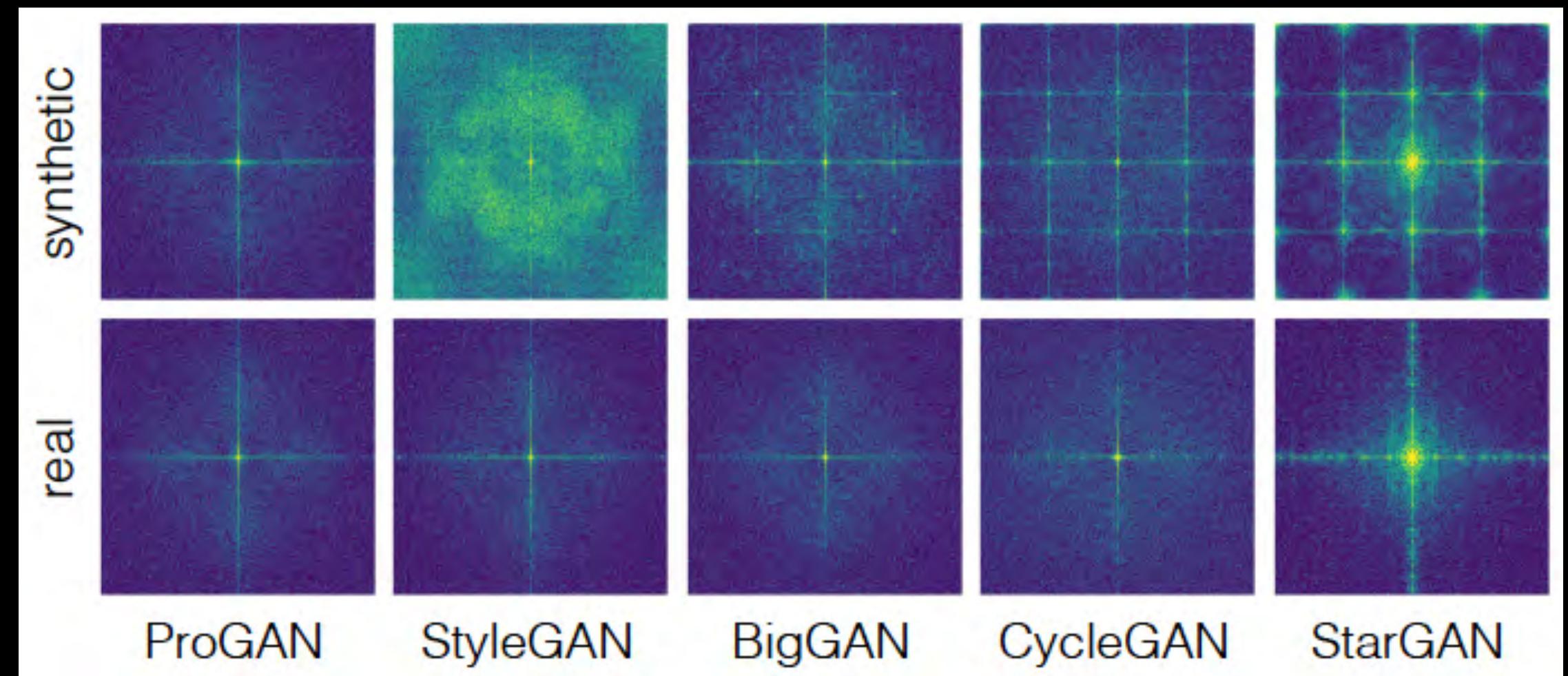
Example case: Synthetic people

Disclaimer: None of these individuals exist!



Potentially "easy" to spot:

- Generator residuals (in the image)
- Patterns in the frequency domain



Example case: Synthetic people

Disclaimer: None of these individuals exist!



Andrew Waltz



Katie Jones



Matilda Romero

Example case: Synthetic people

Disclaimer: None of these individuals exist!



Andrew Waltz



Katie Jones



Matilda Romero

"Real" profile pictures from fake social media users

Example case: Synthetic people

Disclaimer: None of these individuals exist!



87% Fake

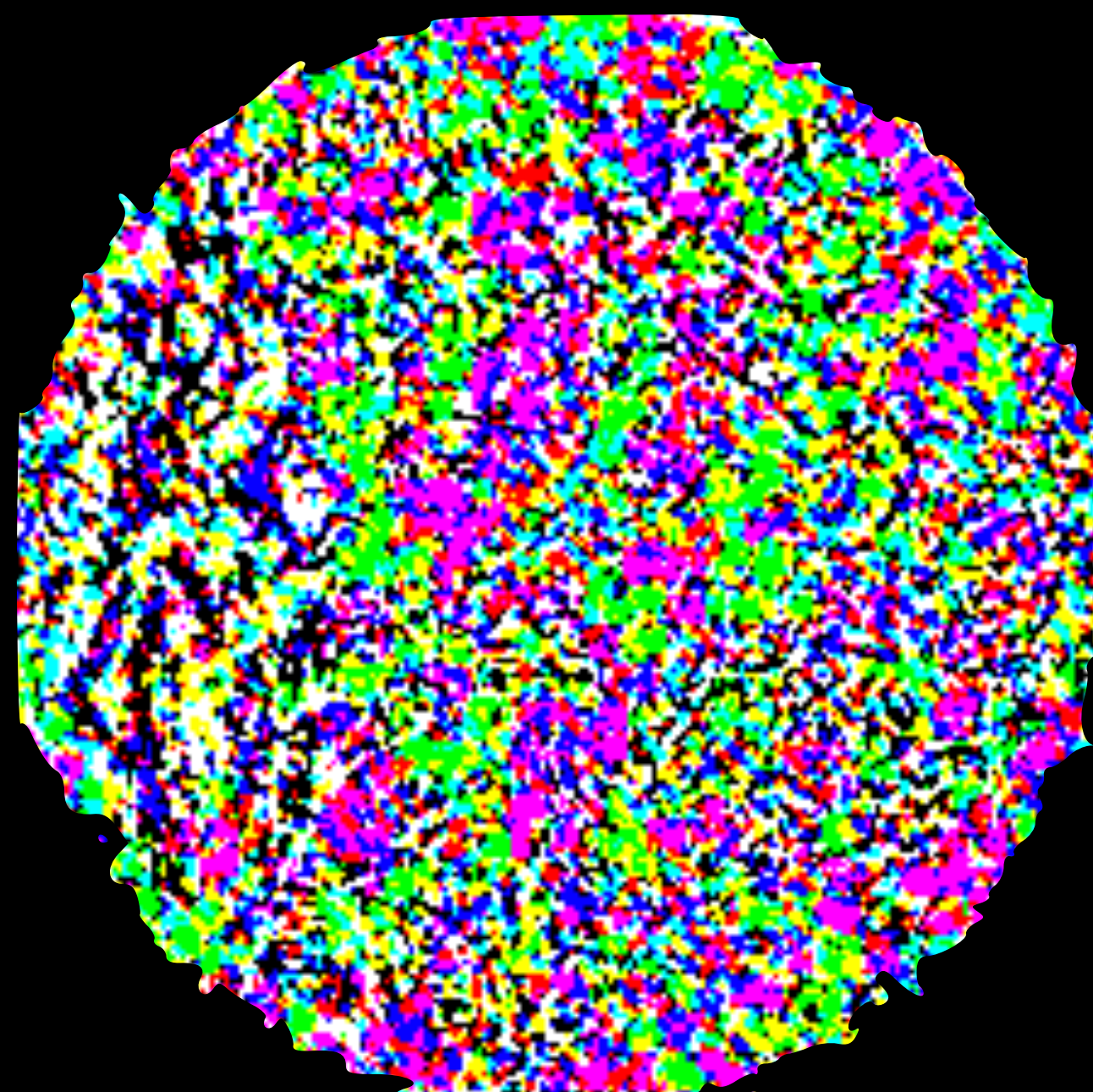
Example case: Synthetic people

Disclaimer: None of these individuals exist!



87% Fake

+



Adversarial noise
(magnified x1000)

=

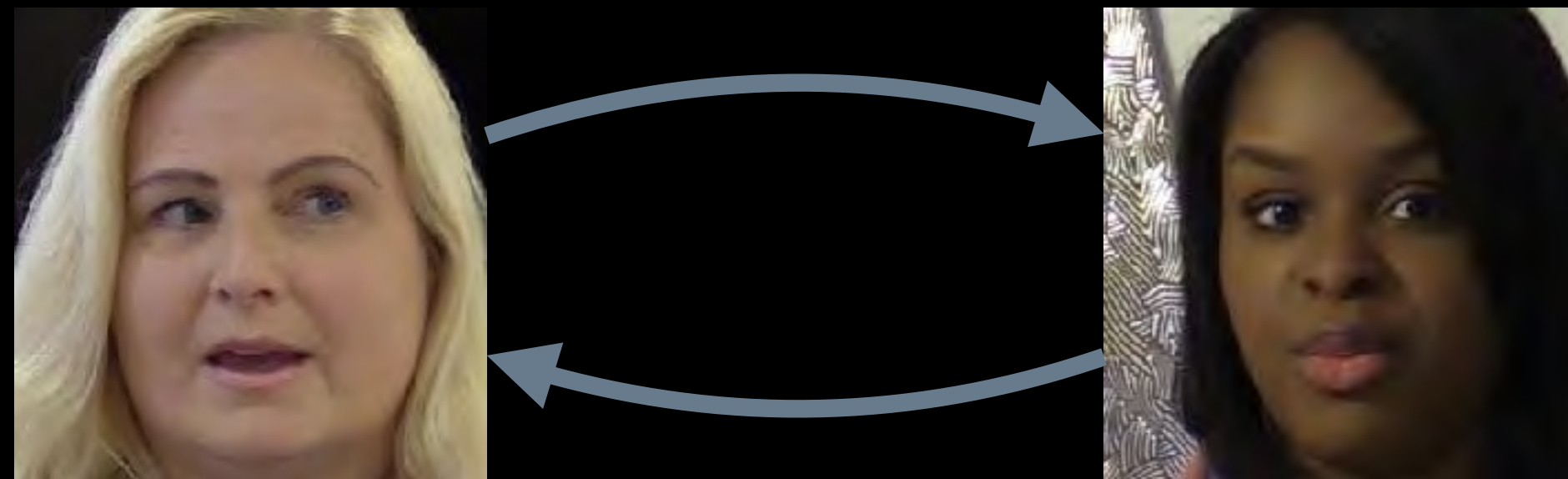


1% Fake

Example case: DeepFakes

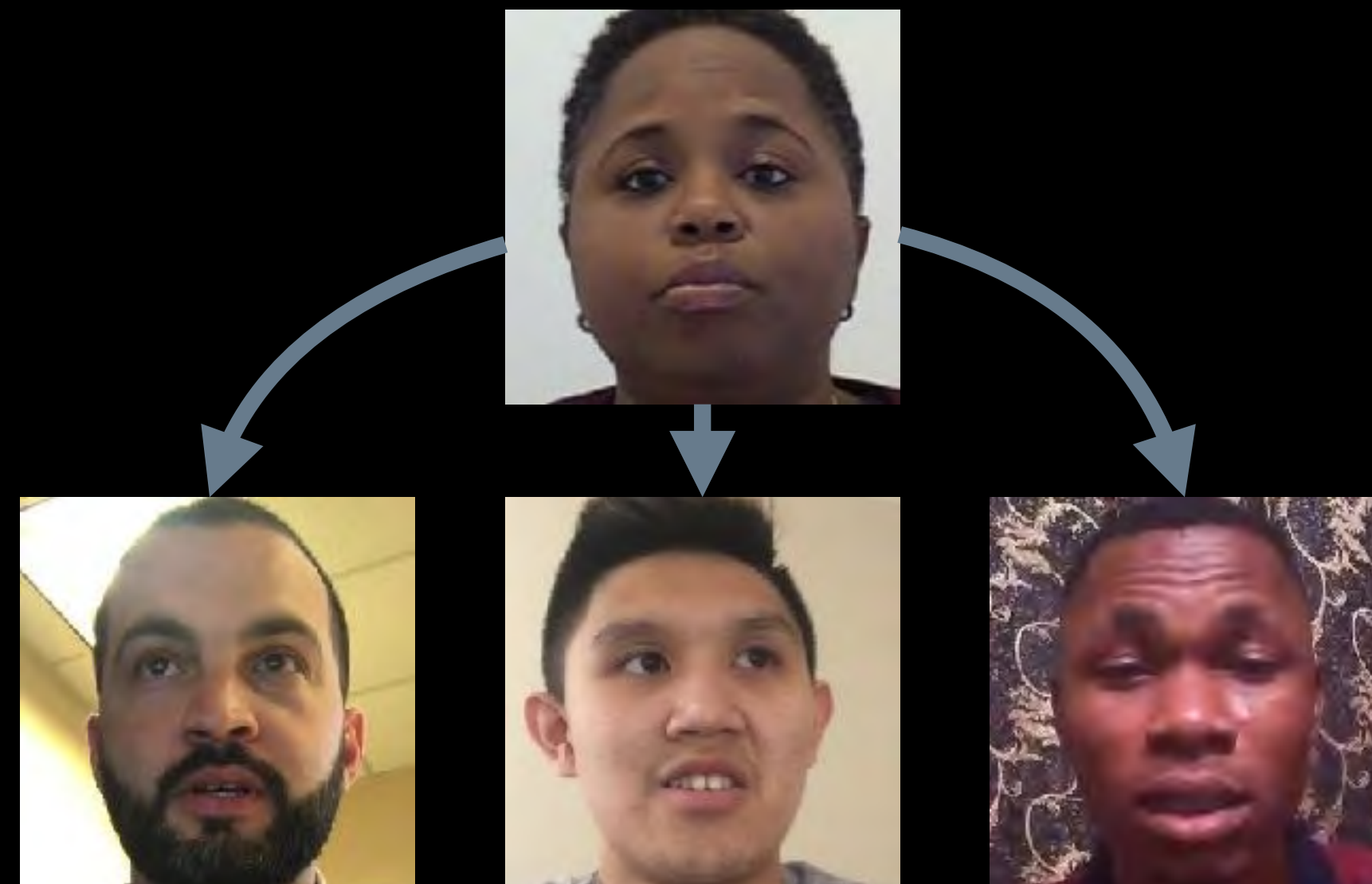


Example case: DeepFakes



Pairwise

Swap the faces of two individuals - the face of person A is put on the body of person B. Requires many photos of person A and B.



Identity-free

With a few reference photos of person A, put this face onto any other person. Many methods use GANs.

Example case: DeepFakes

TECH ARTIFICIAL INTELLIGENCE

US lawmakers say AI deepfakes 'have the potential to disrupt every facet of our society'

They're asking the intelligence community to assess the threat from AI video ma

The impending war over deepfakes

Bipartisan trio asks US intelligence to investigate 'deepfakes'

BY ALI BRELAND - 09/13/18 05:28 PM EDT

DEEPFAKES | By Samantha Col

US lawmakers are concerned about deepfake technology

Three Representatives have asked the intelligence community for information.

BRIEFING • VIDEO
How Faking Videos Became Easy — And Why That's So Scary

... warns on 'deep fakes' in disinformation campaigns

Deepfakes 2.0: The terrifying future of AI and fake news

Simon Chandler — Oct 4 at 10:00PM

There Is No Tech Solution to Deepfakes

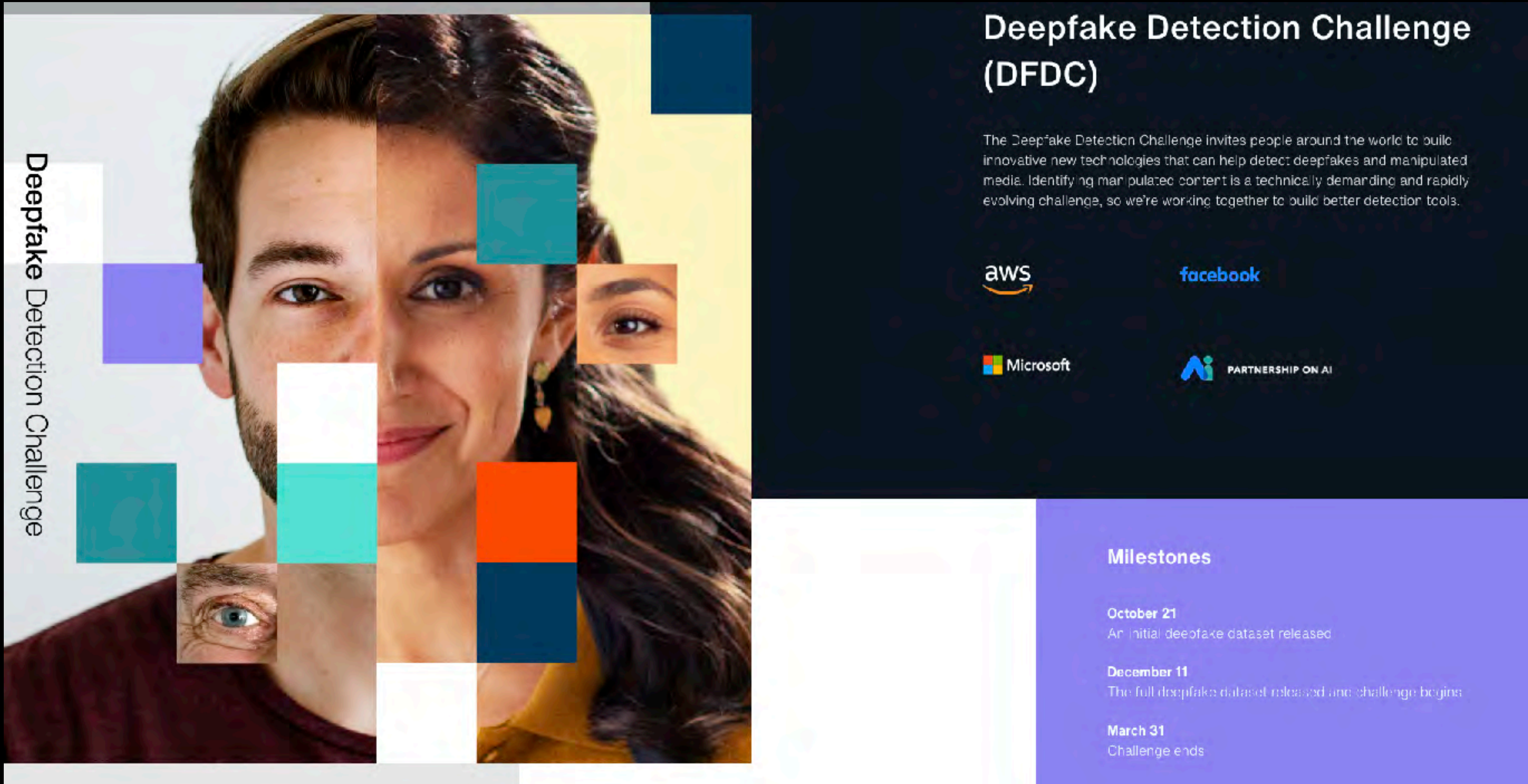
Deep fakes: Who can you trust?

Prevention



Ask the experts

Example - DFDC competition



The image is a promotional graphic for the Deepfake Detection Challenge (DFDC). It features a central image of a woman's face, which is partially obscured by a grid of colorful squares (purple, teal, orange, dark blue, white) that represent different deepfake versions of her face. To the left of the image, the text "Deepfake Detection Challenge" is written vertically. To the right, the title "Deepfake Detection Challenge (DFDC)" is displayed in white on a dark blue background. Below the title, a paragraph explains the challenge's purpose. Further down, logos for sponsors AWS, Facebook, Microsoft, and the Partnership on AI are shown. At the bottom right, a purple box contains a "Milestones" section with three key dates and events.

Deepfake Detection Challenge

Deepfake Detection Challenge (DFDC)

The Deepfake Detection Challenge invites people around the world to build innovative new technologies that can help detect deepfakes and manipulated media. Identifying manipulated content is a technically demanding and rapidly evolving challenge, so we're working together to build better detection tools.

aws facebook
Microsoft PARTNERSHIP ON AI

Milestones

- October 21**
An initial deepfake dataset released
- December 11**
The full deepfake dataset released and challenge begins
- March 31**
Challenge ends

Ask the experts

Example - DFDC competition

The screenshot shows the 'Deepfake Detection Challenge' page on a competition platform. The header includes the title 'Deepfake Detection Challenge' with the subtitle 'Identify videos with facial or voice manipulations'. A prize money of '\$1,000,000' is displayed. The competition is noted as having 2,265 teams and being 3 months old. The 'Leaderboard' tab is selected, showing a 'Public Leaderboard' and a 'Private Leaderboard'. A message states that the competition is closed for submissions and the private leaderboard is based on a re-run of participants' code. A 'Refresh' button is present. A legend indicates medal colors: In the money (green), Gold (yellow), Silver (grey), and Bronze (brown). The leaderboard table lists the top 5 teams with their scores and entry counts.

Featured Code Competition

Deepfake Detection Challenge

Identify videos with facial or voice manipulations

\$1,000,000
Prize Money

Deepfake Detection Challenge · 2,265 teams · 3 months ago

Overview Data Notebooks Discussion **Leaderboard** Rules Team Host

Public Leaderboard **Private Leaderboard**

This competition is closed for submissions. The Private Leaderboard was based on a re-run of participants' code by the host on a privately-held test set. [Refresh](#)

This competition has completed. This leaderboard reflects the final standings.

In the money Gold Silver Bronze

#	Δpub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲ 3	Selim Seferbekov			0.42798	2	3mo
2	▲ 35	\WM/			0.42842	2	3mo
3	▲ 3	NtechLab			0.43452	2	3mo
4	▲ 6	Eighteen years old			0.43476	2	3mo
5	▲ 12	The Medics	DFDC 3D & 2D inc...		0.43711	2	3mo

Ask the experts

Example - DFDC competition - Dataset

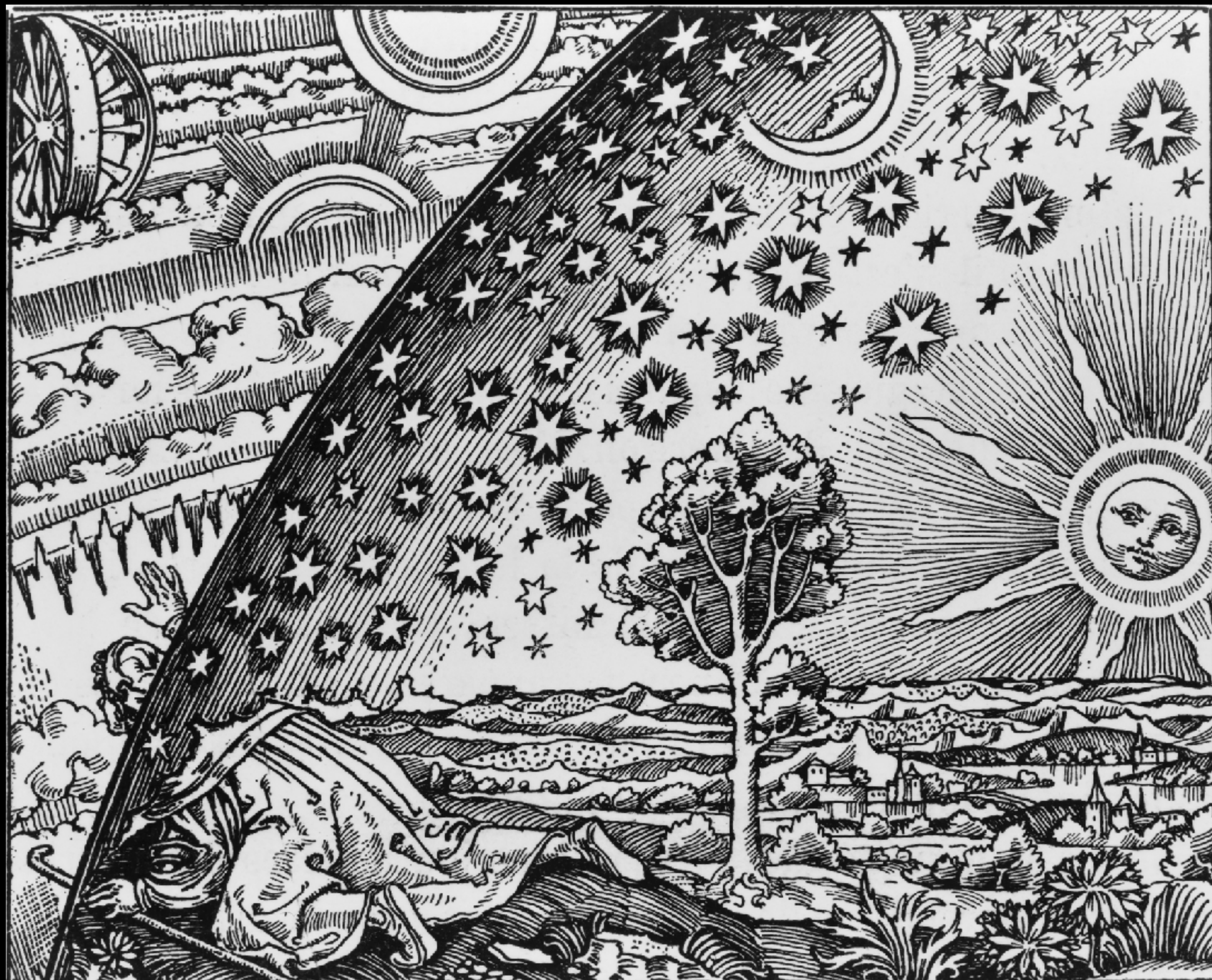


Ask the experts

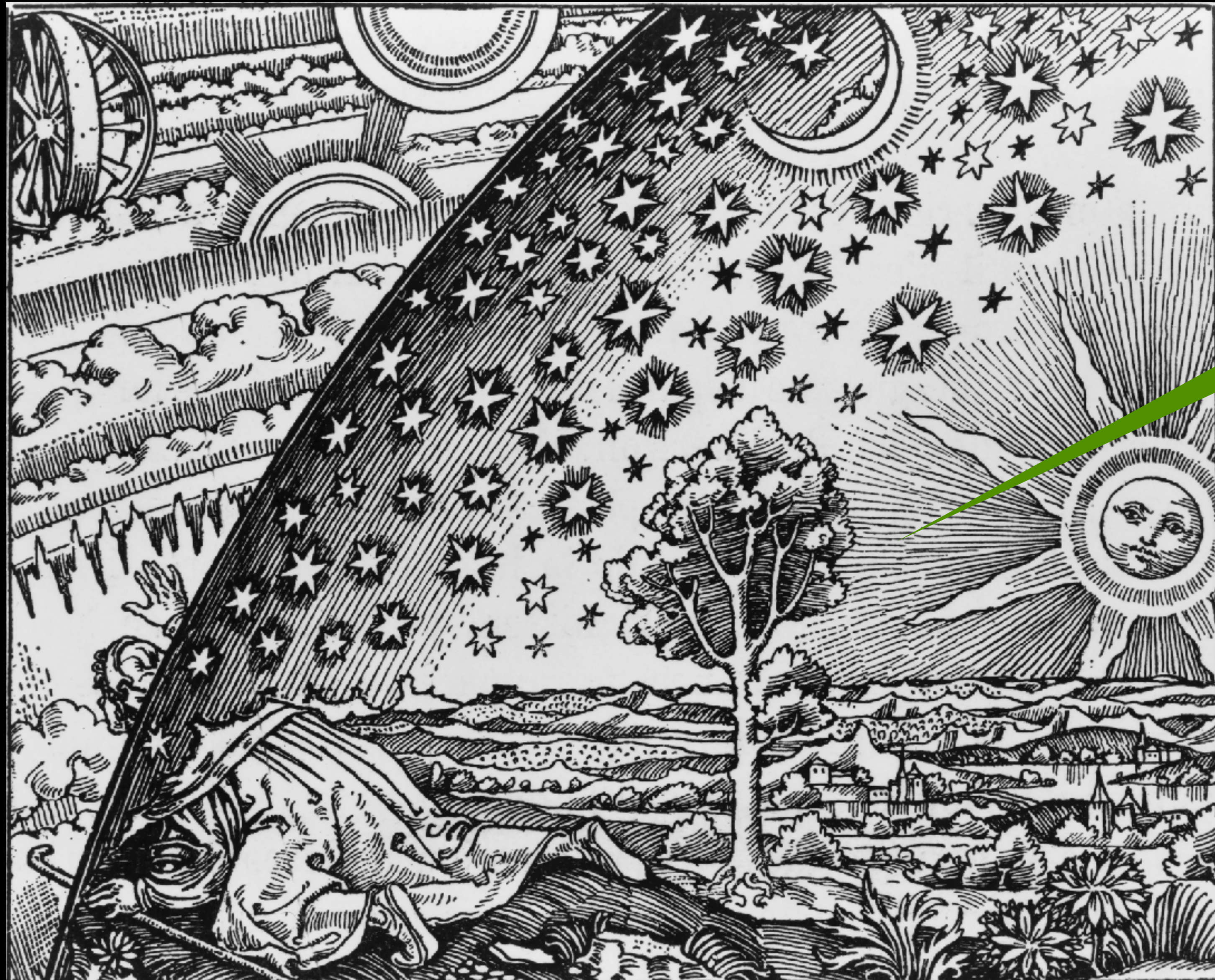
Example - DFDC competition - Dataset



Domain gap + Distribution shift



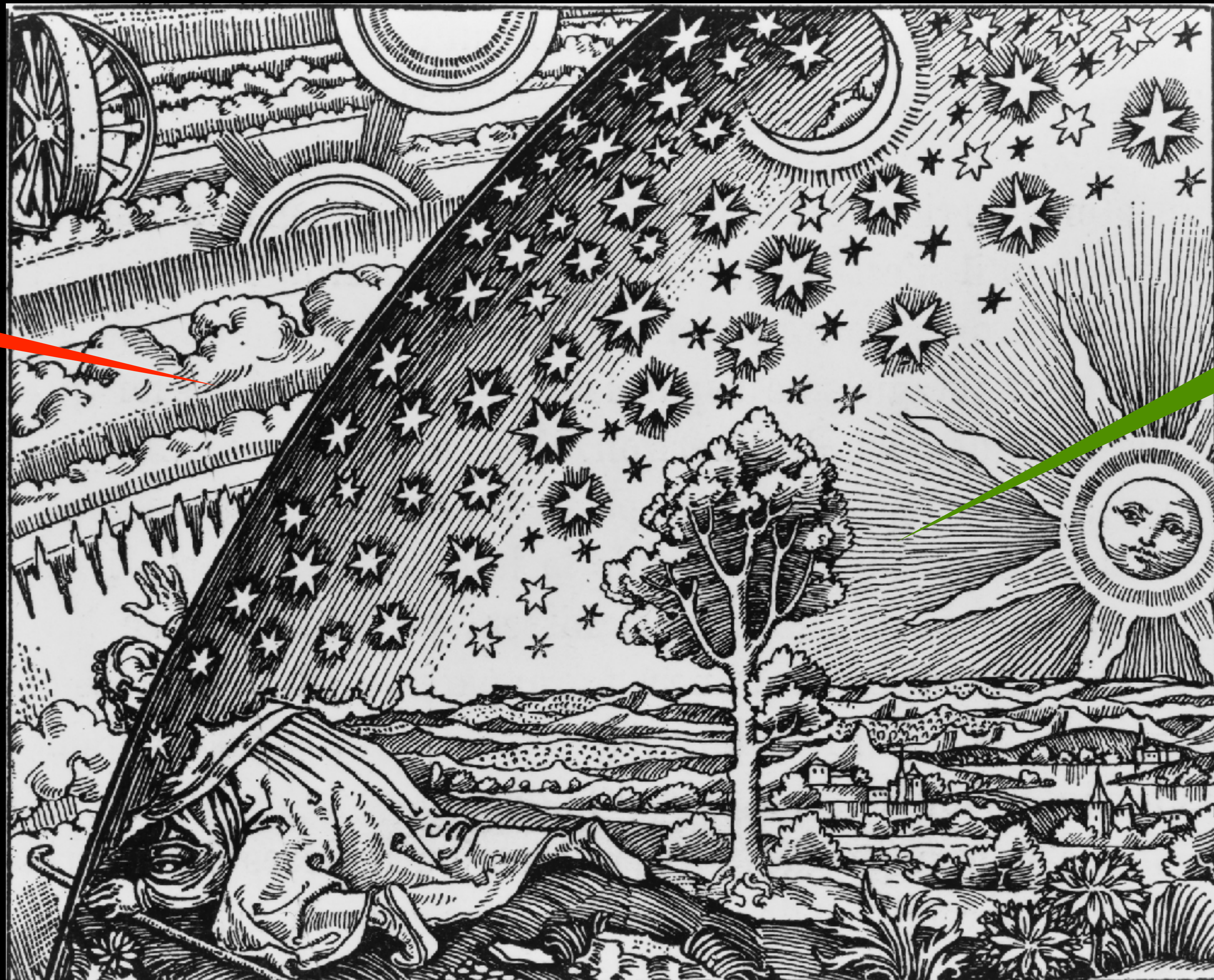
Domain gap + Distribution shift



The test distribution
you constructed to
validate your algorithm

Domain gap + Distribution shift

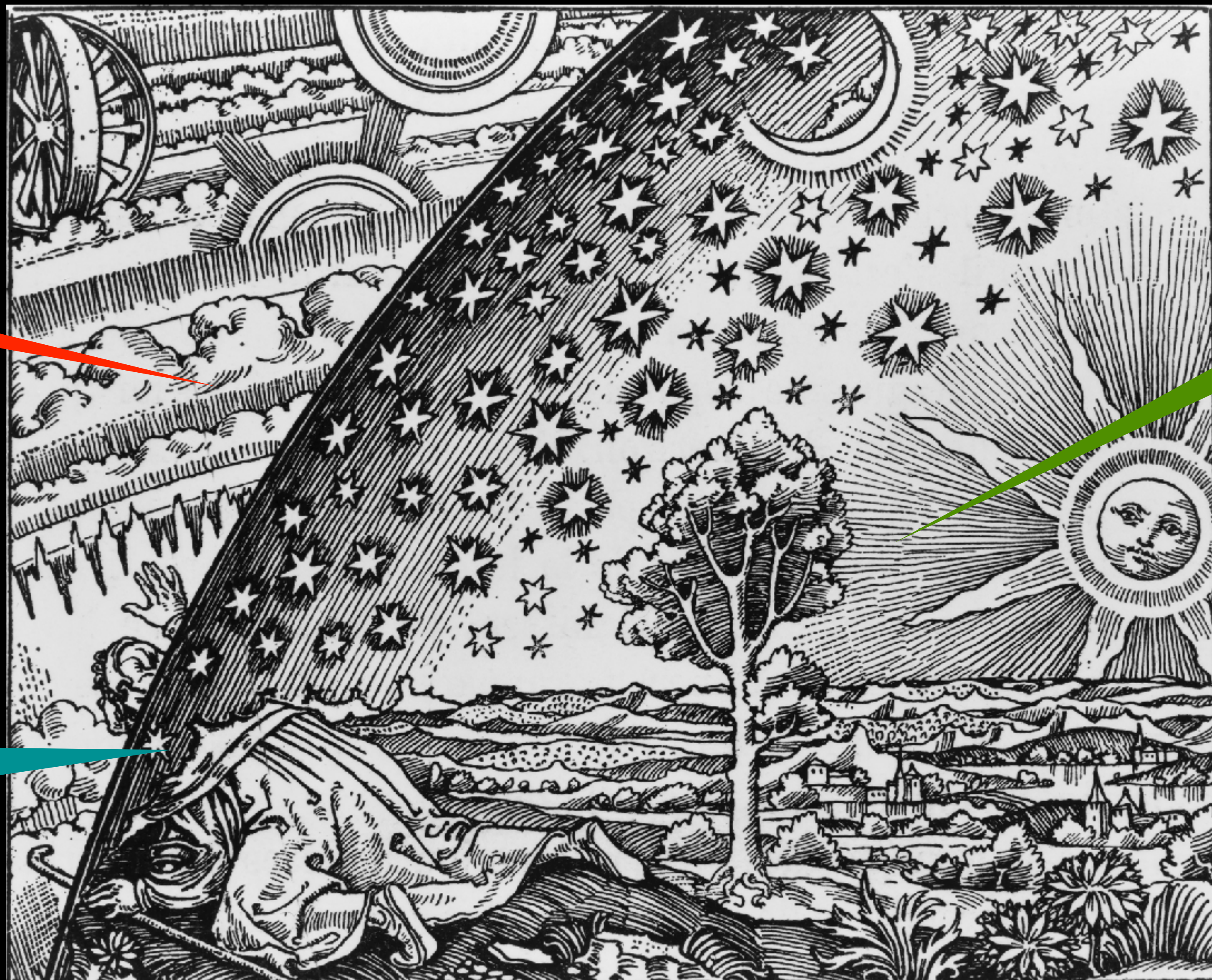
The real distribution



The test distribution you constructed to validate your algorithm

Domain gap + Distribution shift

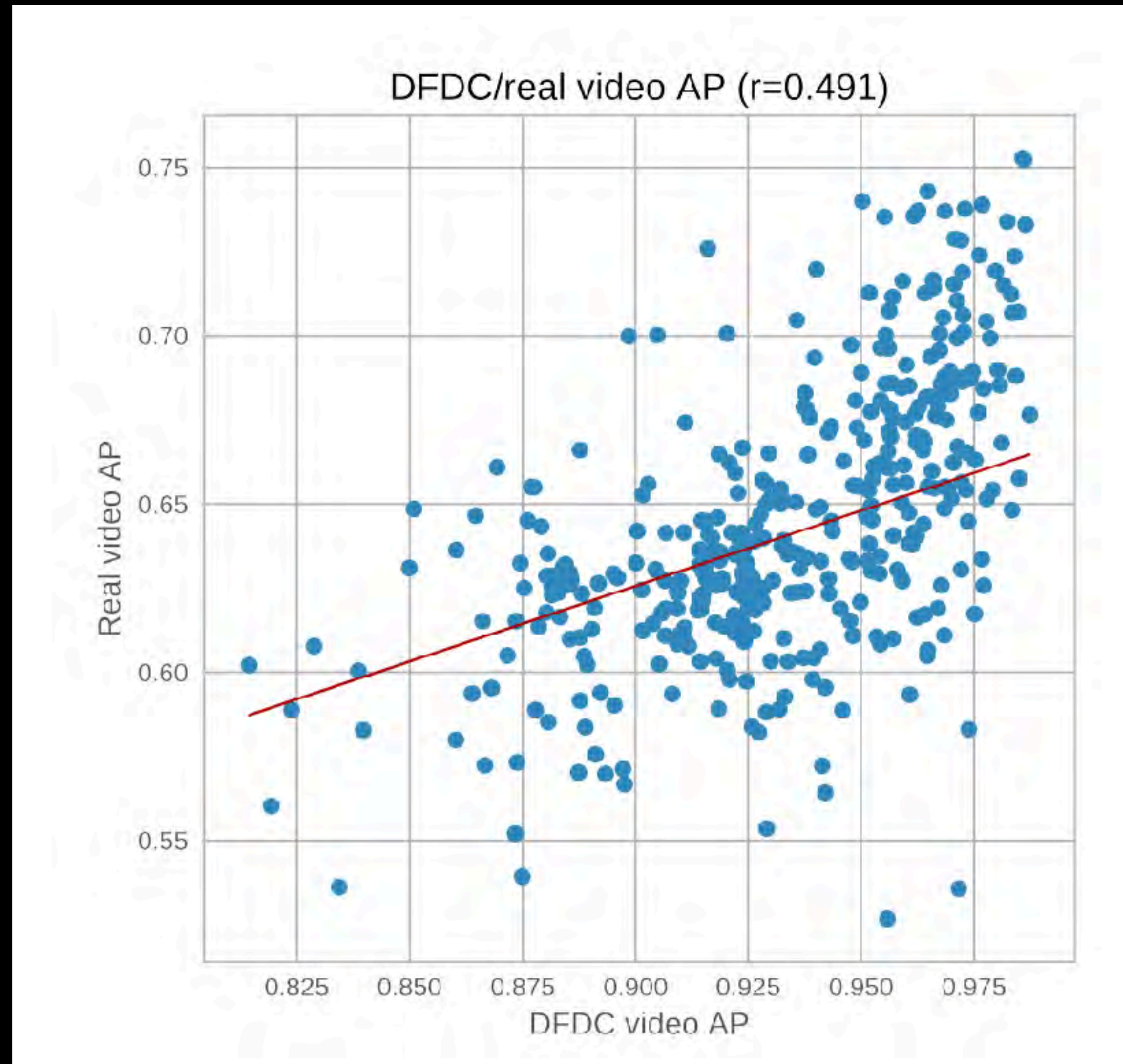
The real distribution



The test distribution you constructed to validate your algorithm

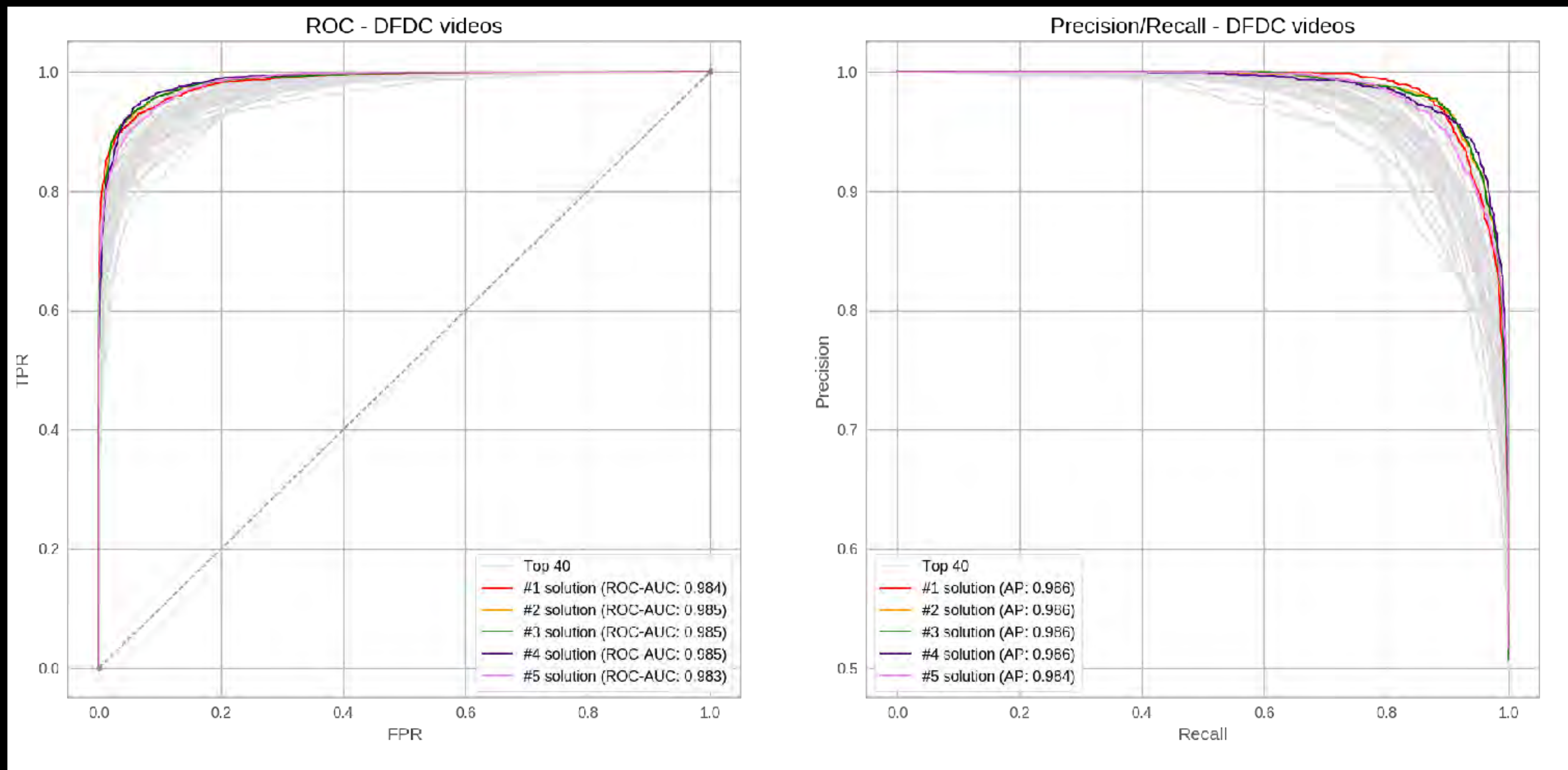
Your algorithm's goal

Domain gap + Distribution shift

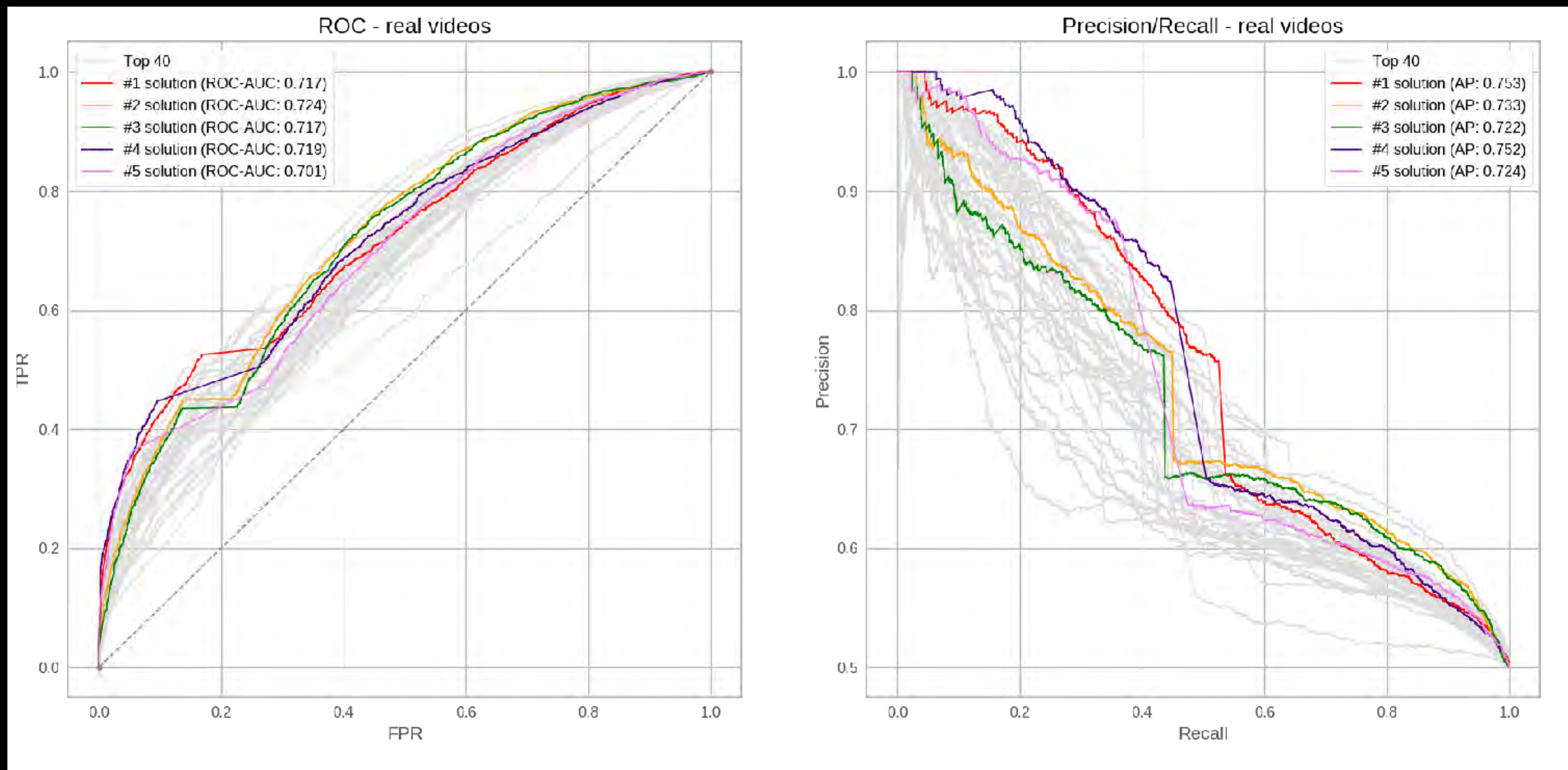


Team name	Overall log loss	DFDC log loss	Real log loss
Selim Seferbekov [24]	0.4279	0.1983	0.6605
WM [34]	0.4284	0.1787	0.6805
NTechLab [4]	0.4345	0.1703	0.7039
Eighteen Years Old [25]	0.4347	0.1882	0.6831
The Medics [11]	0.4371	0.2157	0.6621

Domain gap + Distribution shift

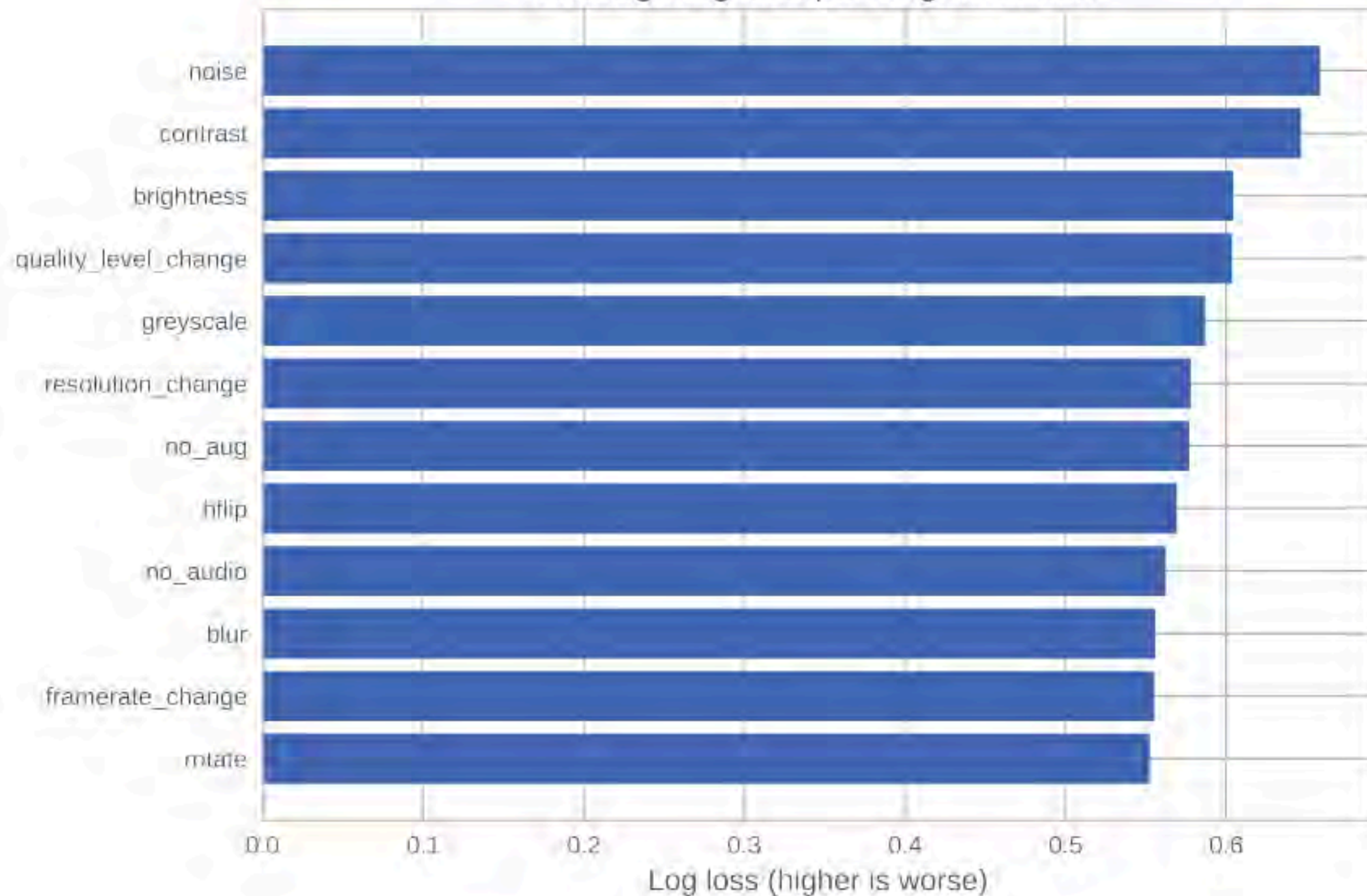


Domain gap + Distribution shift

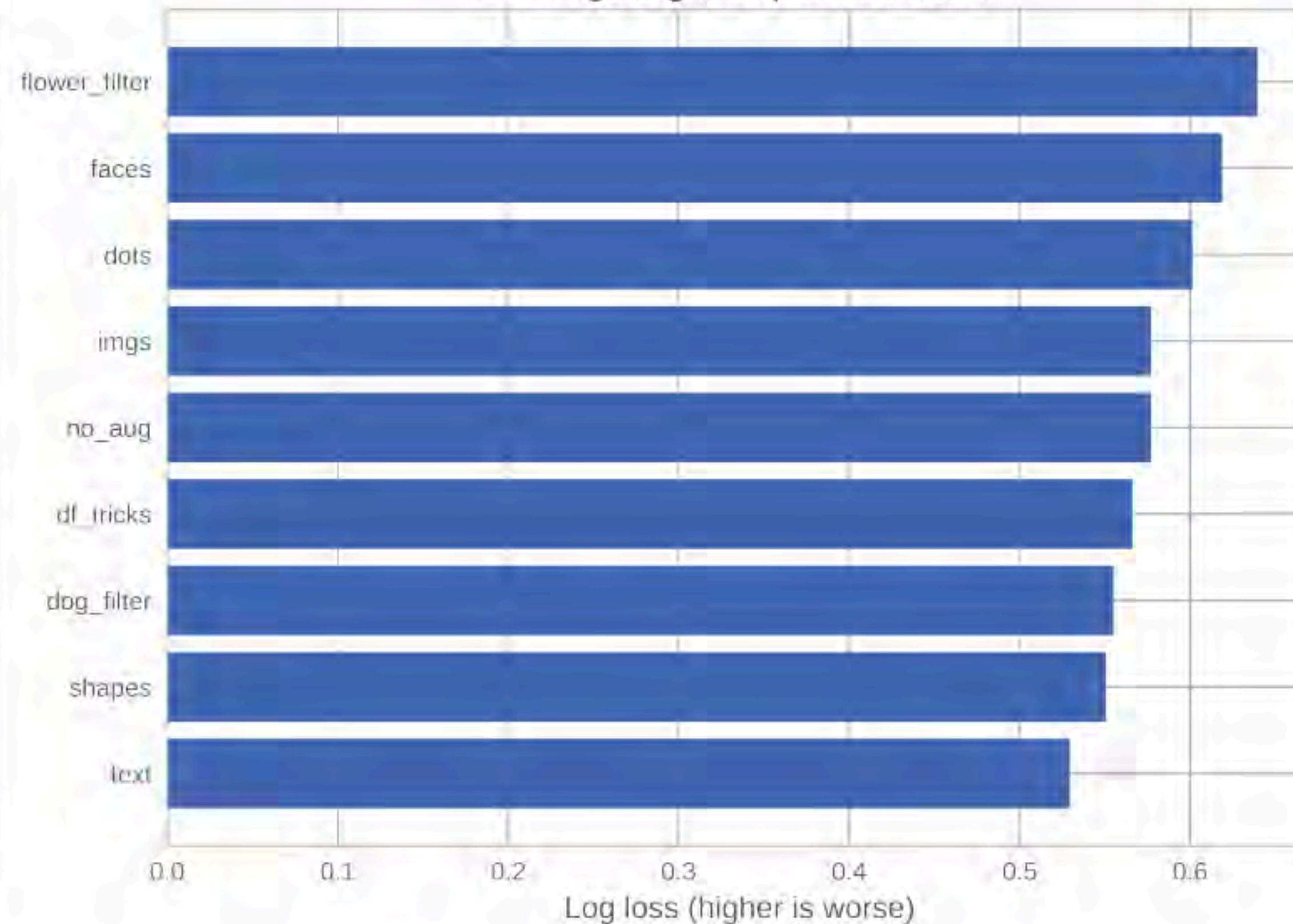


Domain gap + Distribution shift

Average log loss per-augmentation



Average log loss per-distractor



Domain gap + Distribution shift

(and know your metrics!)



In general, classification metrics cannot tell the whole story for detection problems.

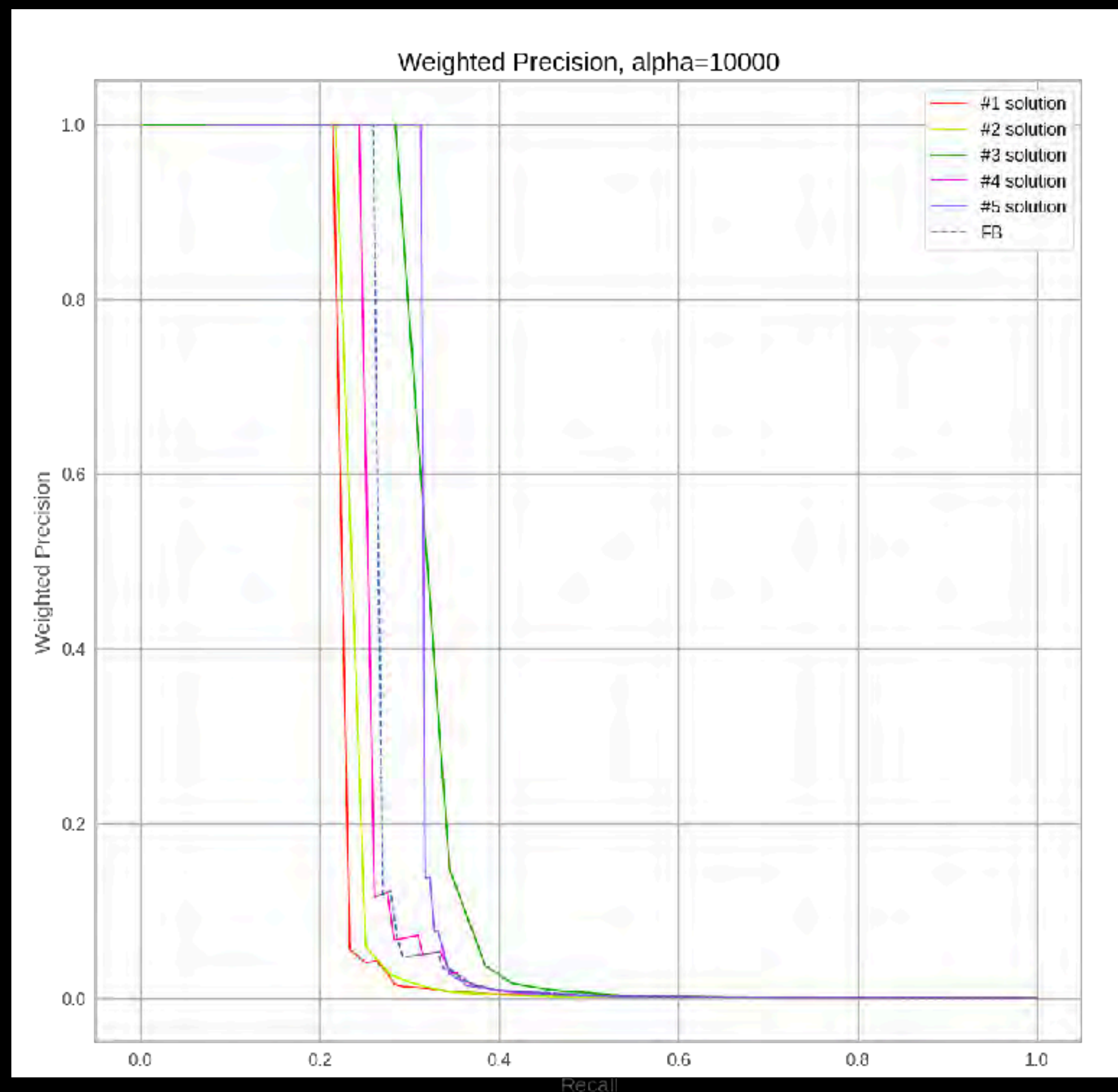
Detecting DeepFakes from a large pool of real videos is a problem with extreme class imbalance.

Even with an extremely small false positive rate (which accuracy does not really account for), many more false positives will be detected than real DeepFakes.

$$R = \frac{FP}{FP + TN} \quad wP = \frac{TP}{TP + \alpha FP}$$

Domain gap + Distribution shift

(and know your metrics!)



A practical case: Risk-a-thons

- What is a risk-a-thon? Why is it necessary?

A practical case: Risk-a-thons

- What is a risk-a-thon? Why is it necessary?
- For DeepFakes detection:
 - Generalization attacks
 - Adversarial noise
 - Sub-population attacks (burns, vitiligo, skin conditions,...)
 - Make-up, scarfs, hats, etc.

Open vs Closed sourcing

Pros: Good as how well you can keep it secret

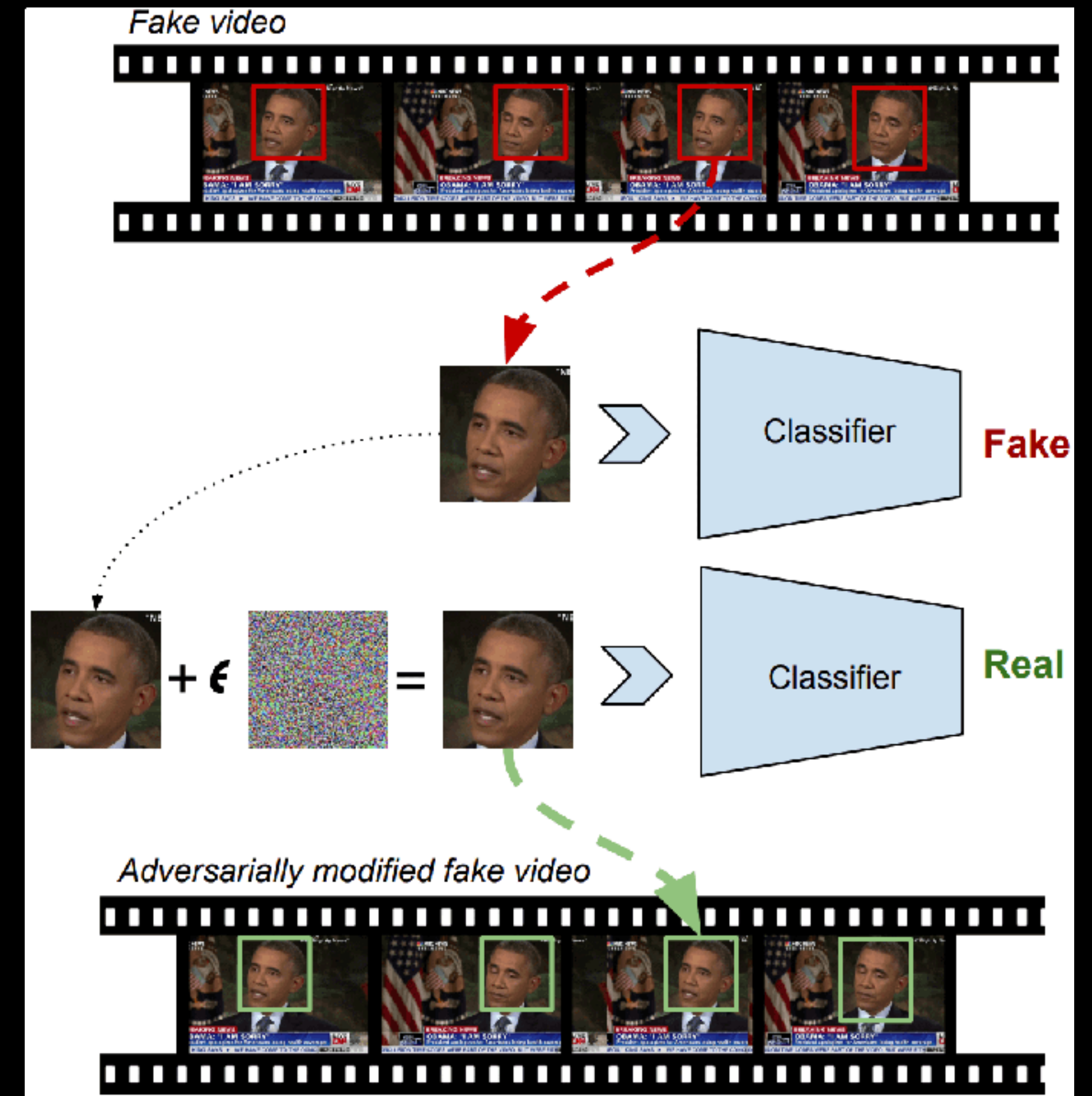
Cons: Underestimation of the adversarial agent

Open vs Closed sourcing

Pros: Good as how well you can keep it secret

Cons: Underestimation of the adversarial agent

Open source DeepFake detectors:
XceptionNet and MesoNet



Reaction



Duct tape fix on Apollo 17 mission

Mitigation

- Sometimes, being preventive about every potential adversity is unfeasible!

Mitigation

- Sometimes, being preventive about every potential adversity is unfeasible!
- Define mitigations for the most (unaddressed) risky scenarios

Mitigation

- Sometimes, being preventive about every potential adversity is unfeasible!
- Define mitigations for the most (unaddressed) risky scenarios
- Build defensive systems that are able to rapidly incorporate new adversarial samples, even if there's few of them

Mitigation

- Sometimes, being preventive about every potential adversity is unfeasible!
- Define mitigations for the most (unaddressed) risky scenarios
- Build defensive systems that are able to rapidly incorporate new adversarial samples, even if there's few of them
- Define coordination strategies (if possible) to mitigate potential AI-centric attacks across multiple surfaces

Conclusions

Conclusions

- Assume an adversarial mindset when developing systems built on the top of AI.
- Understand your risk manifold, quantify it and made informed decisions to prioritize defenses and mitigation strategies
- The scope of may AI Red Team is very broad, focus on the relevant areas for your industry
- Stress test mercilessly. Develop a strategy to convince stakeholders of the value of being ready against a worst-case-scenario
- *The more you sweat in training, the less you bleed in battle.*

Thanks!
Q&A

Cristian Canton (@cristiancanton)
Research Manager (AI Red Team), Facebook AI