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

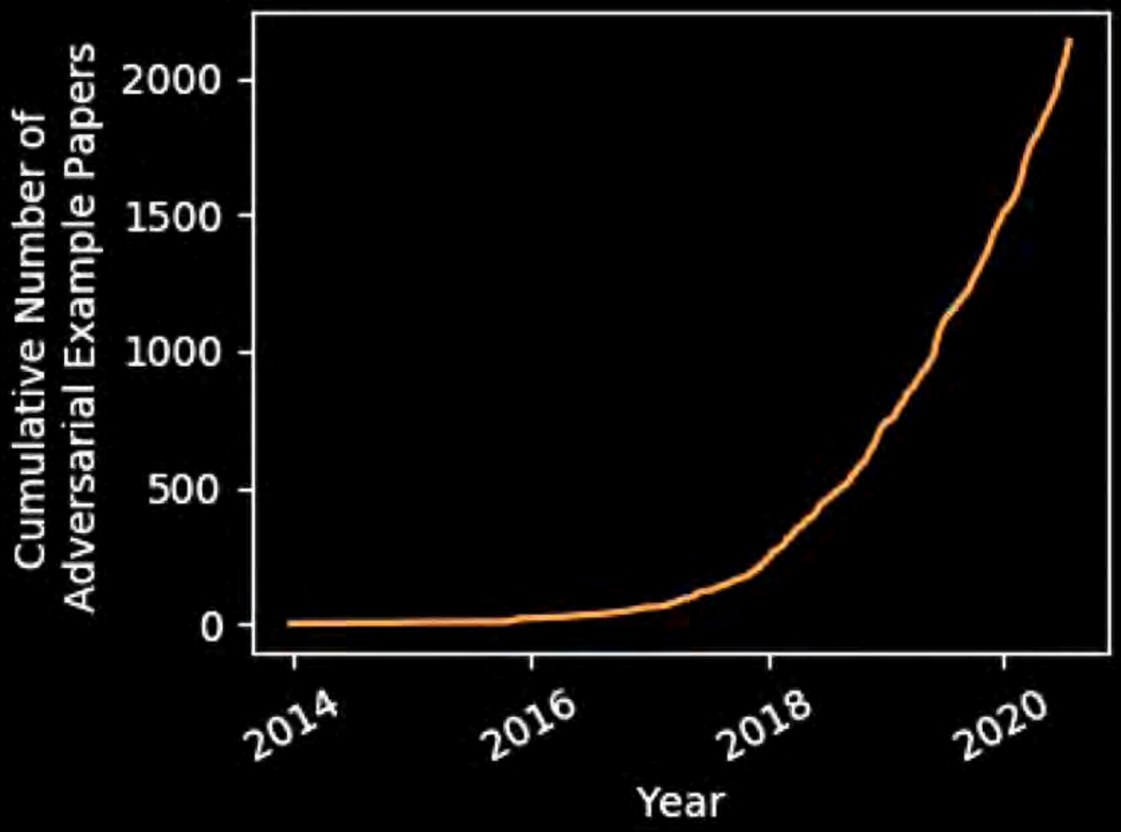
Prevention

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 ⇏ 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

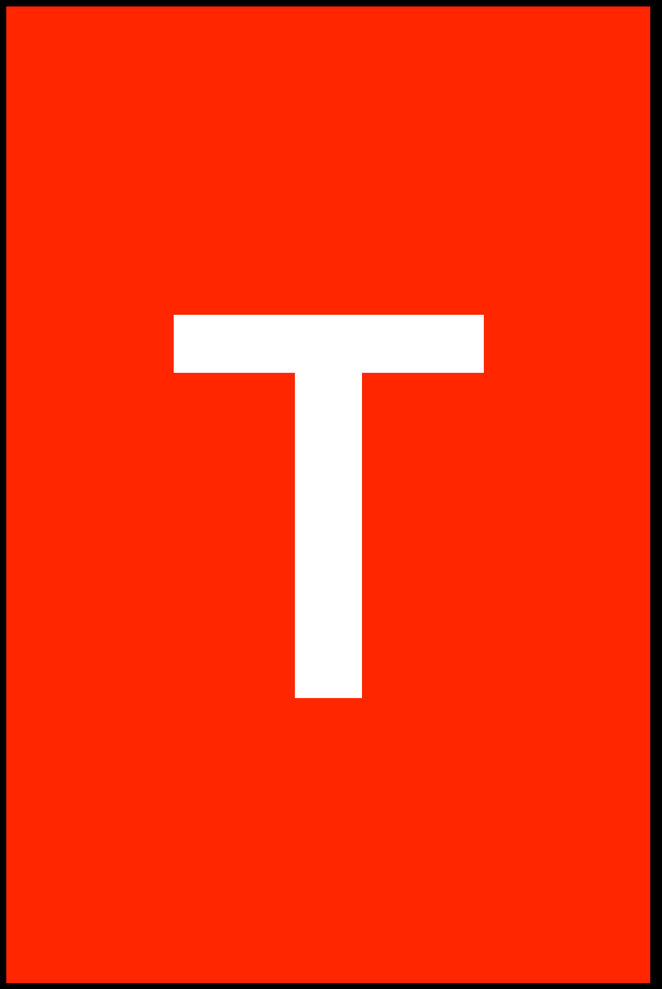
Credit: Goodfellow et al. "*Explaining and harnessing adversarial examples*", ICLR 2015.

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."

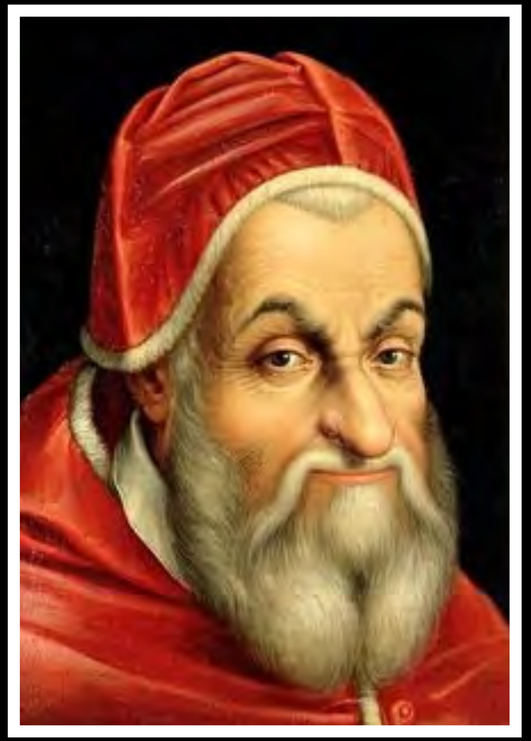
Wikipedia



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

AI Risk = Severity x 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

AI Risk = Severity x Likelihood

A first (real) example

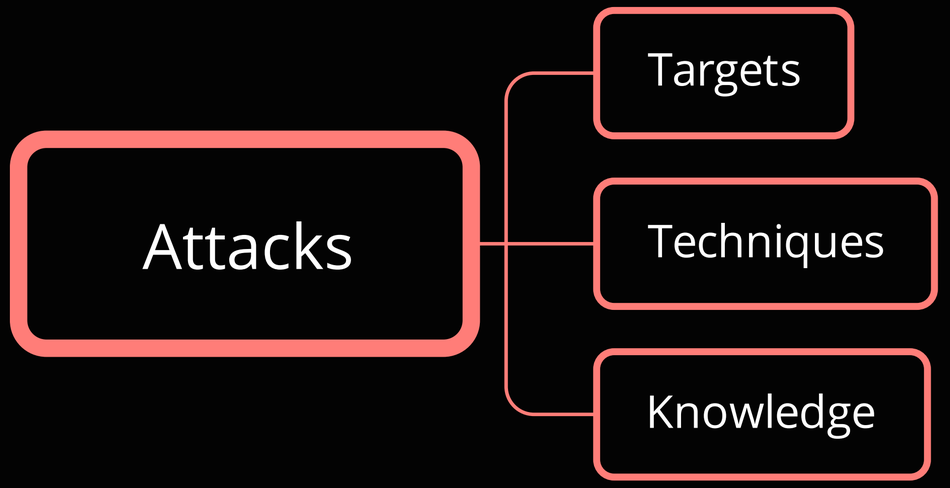
This is"objectionable content" (99%)

A first (real) example

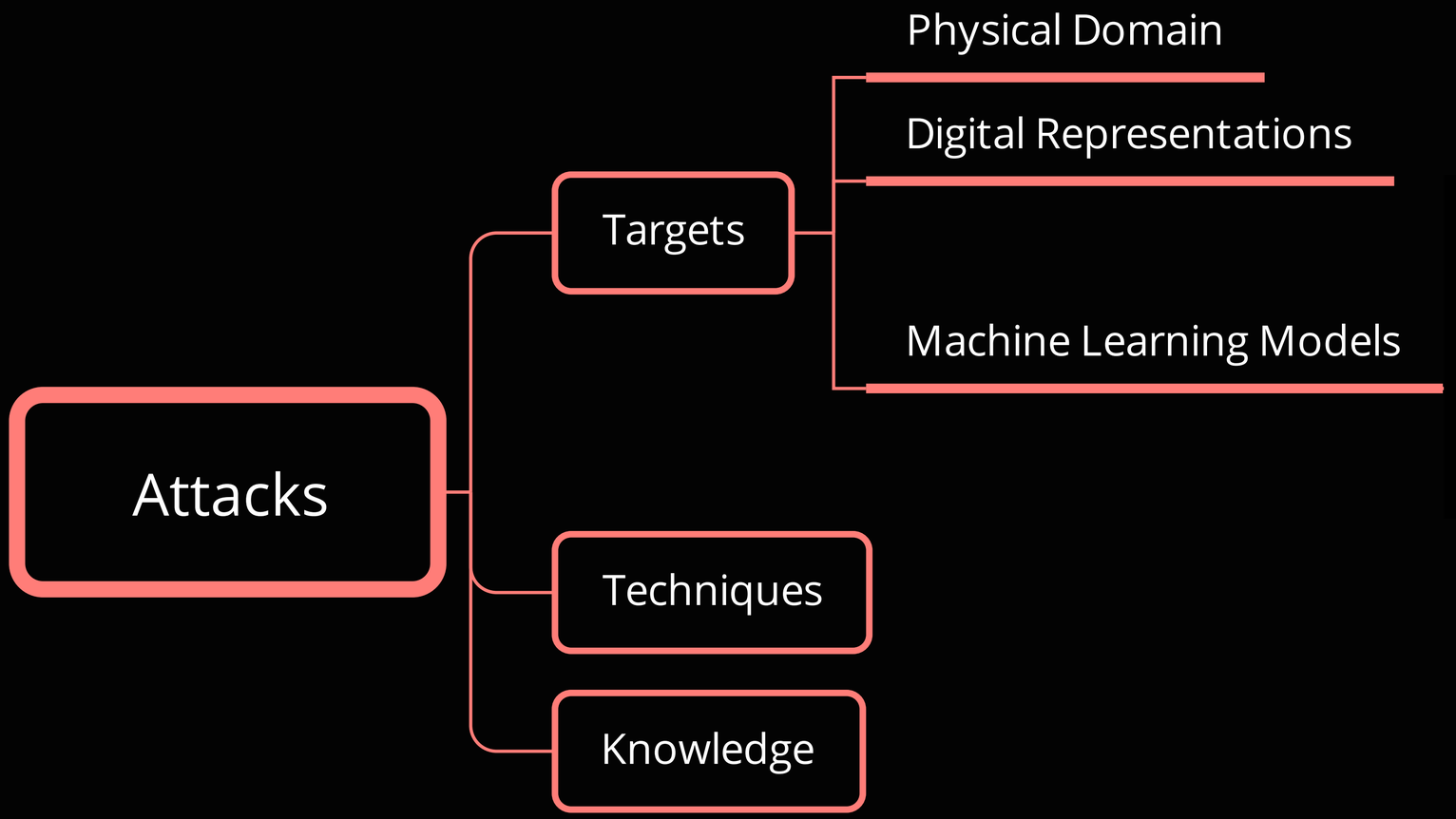
This is safe content (95%)

Abuses

Eykholt et al. "Robust Physical-World Attacks on Deep Learning Visual Classification", 2018.

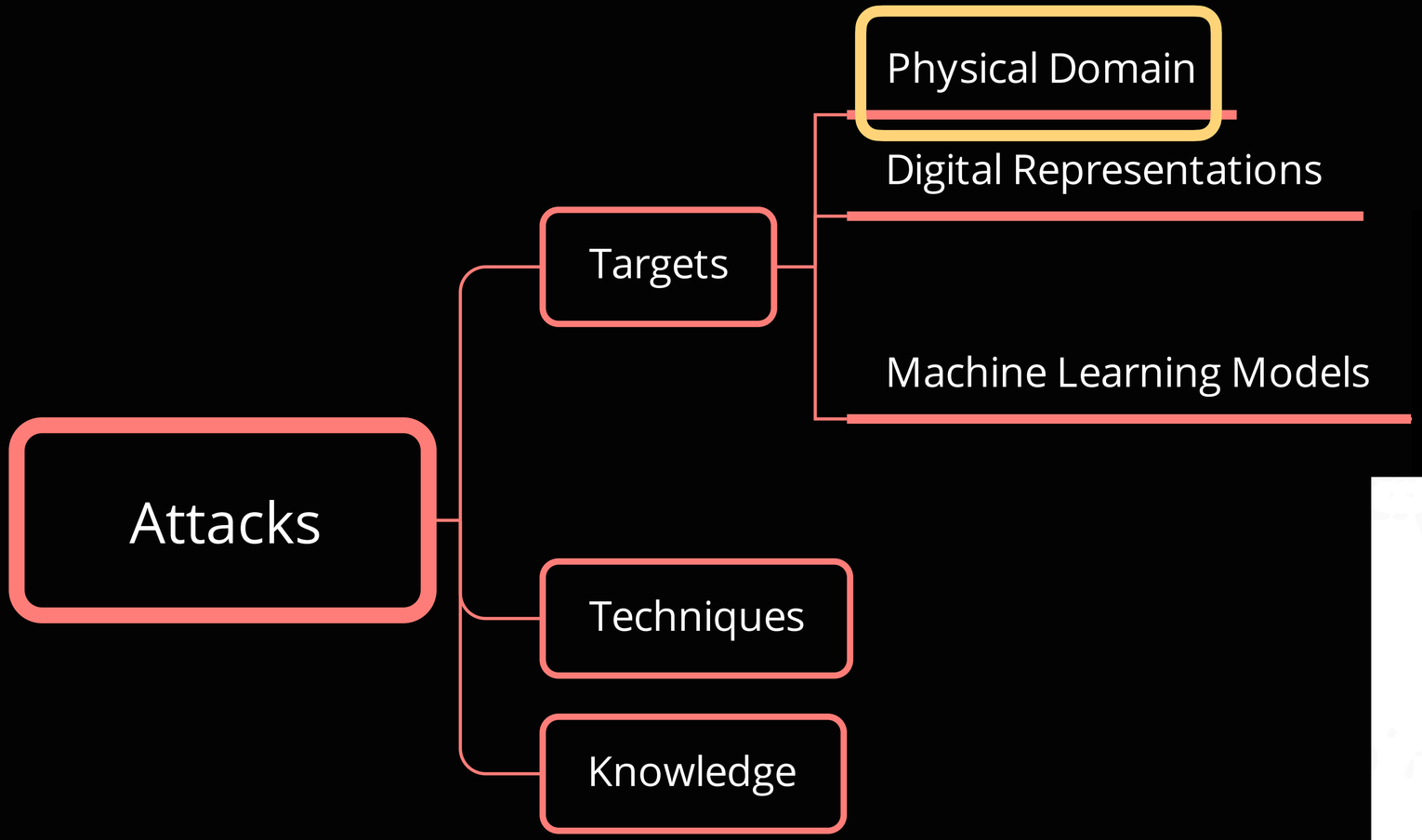


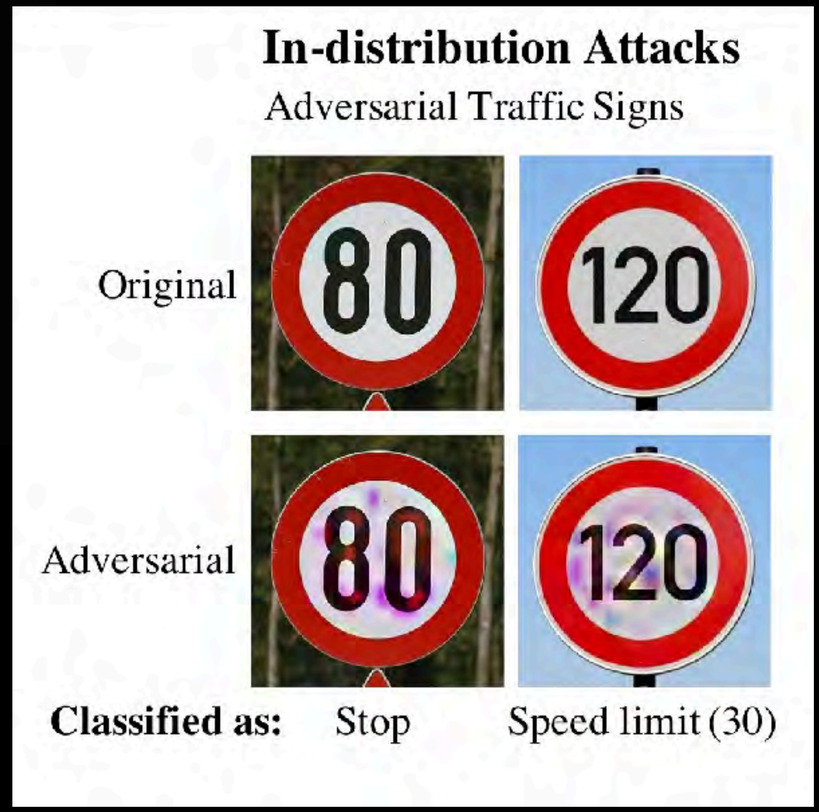
Tabassi et al., "*A Taxonomy and Terminology of Adversarial Machine Learning*", 2019.



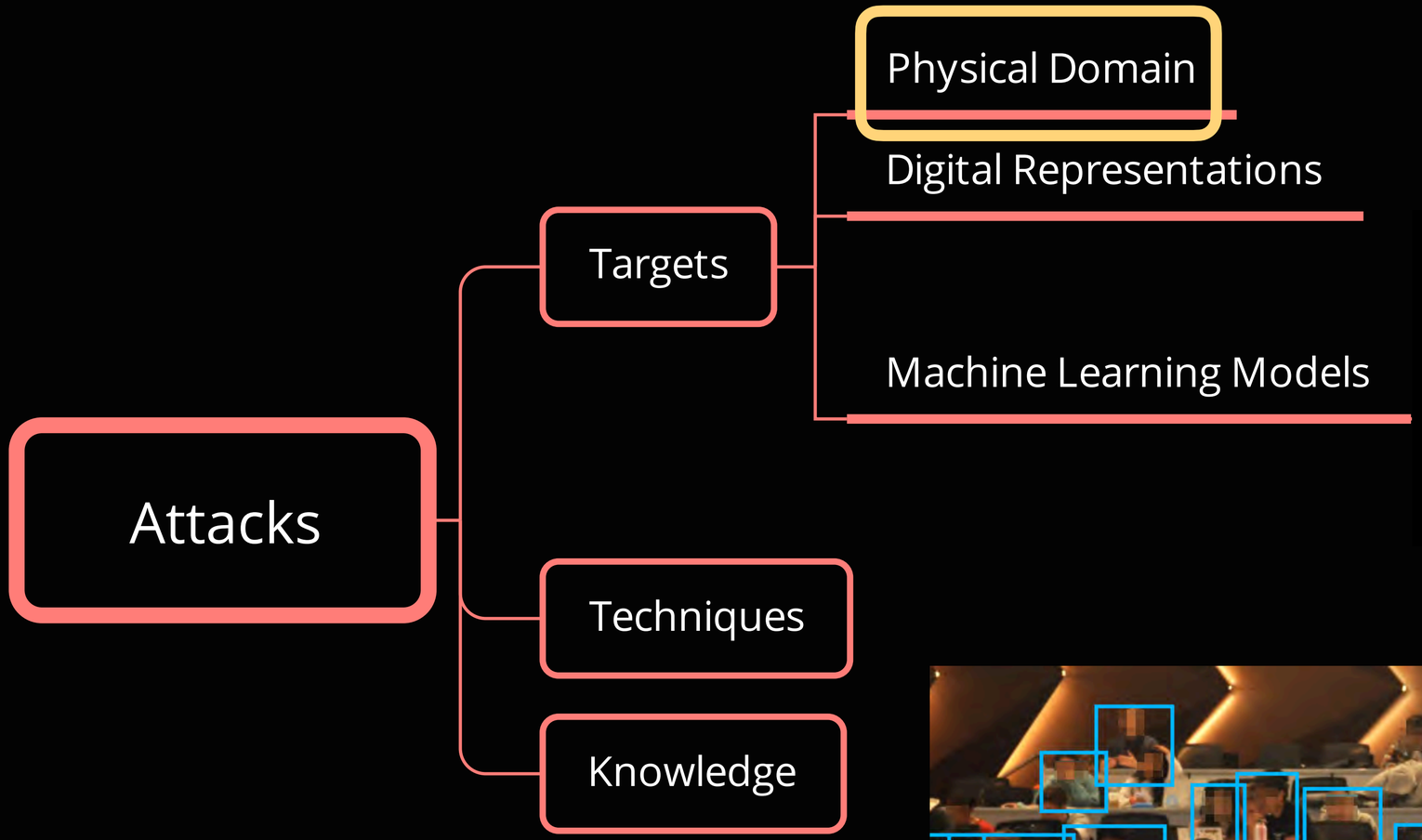
Tabassi et al., "*A Taxonomy and Terminology of Adversarial Machine Learning*", 2019.

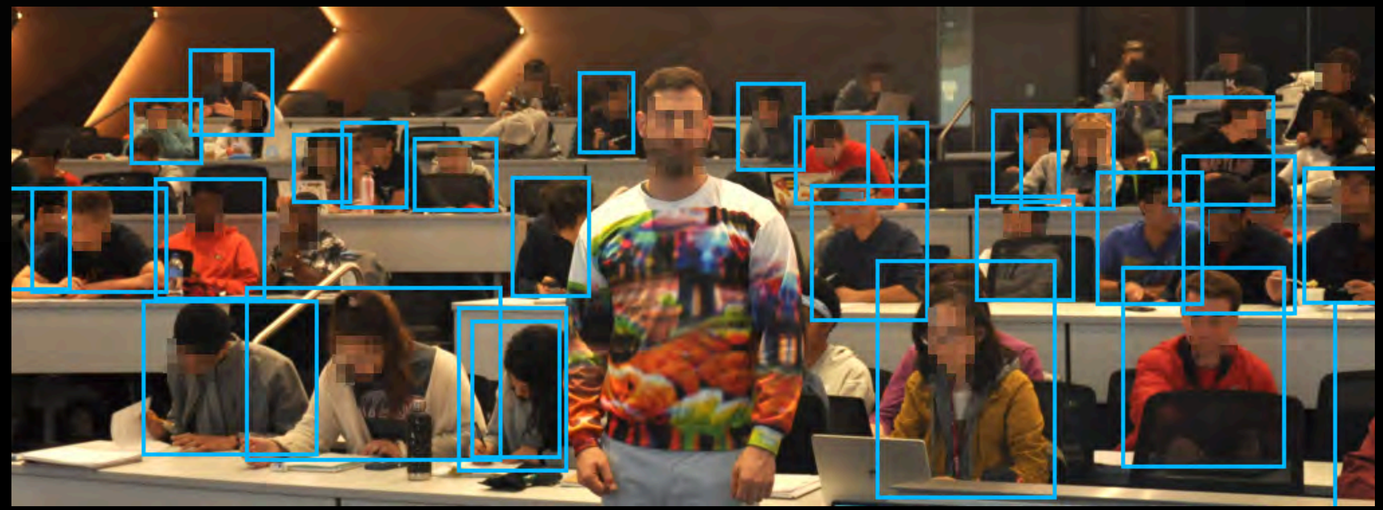
Tabassi et al., "*A Taxonomy and Terminology of Adversarial Machine Learning*", 2019.





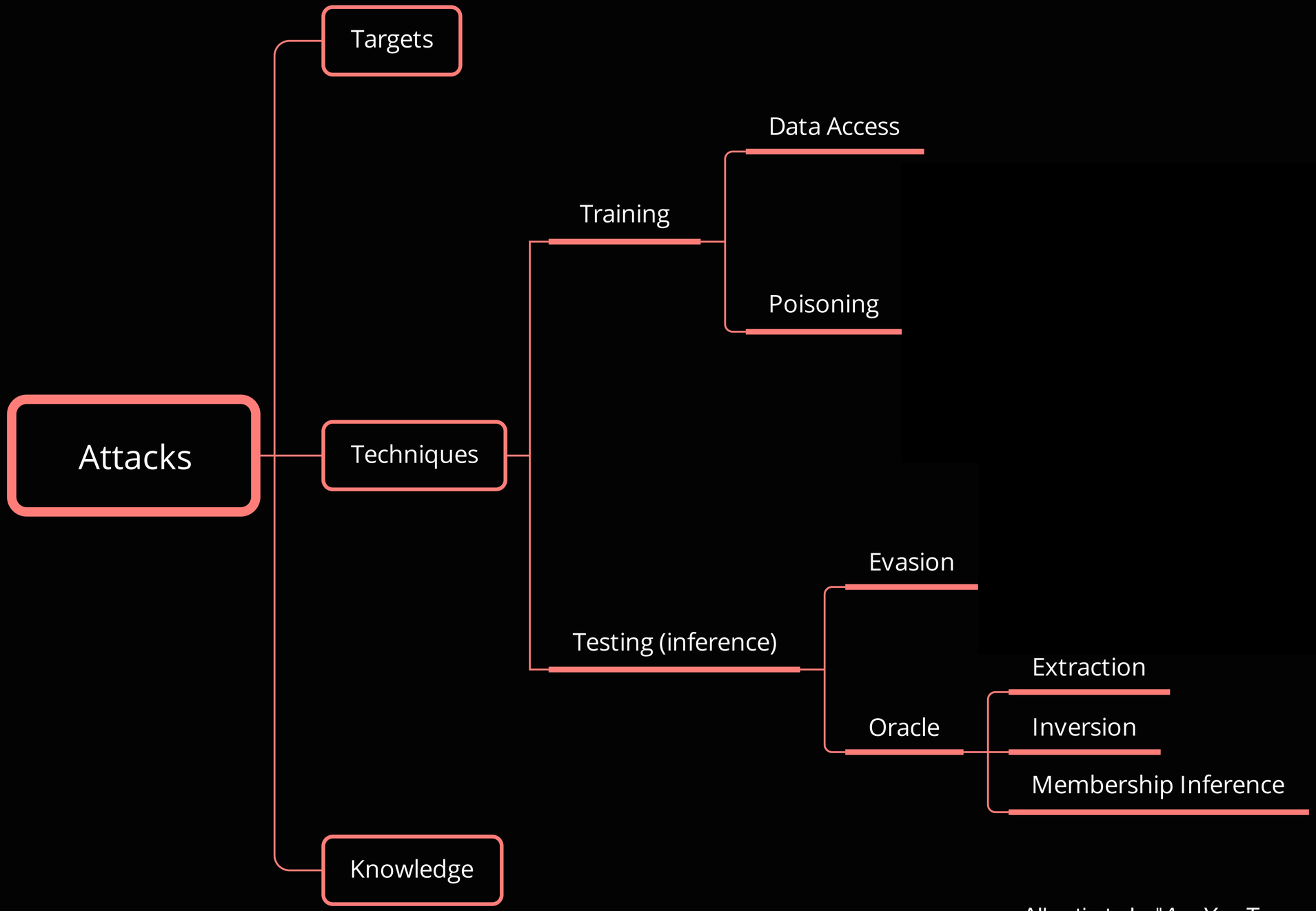
Sitawarin et al., "*DARTS: Deceiving Autonomous Cars with Toxic Signs*", 2018.





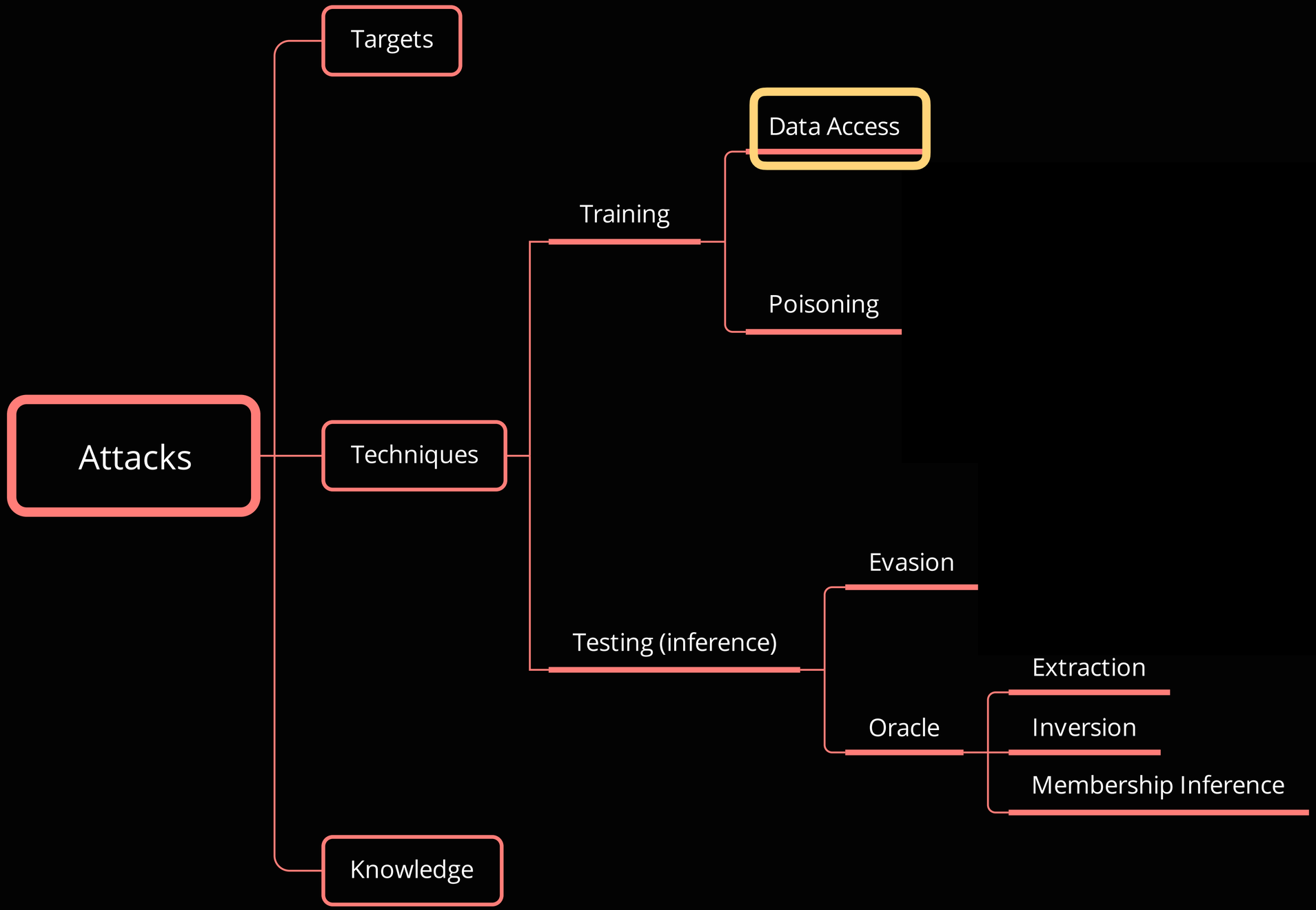
Tabassi et al., "*A Taxonomy and Terminology of Adversarial Machine Learning*", 2019.

Wu et al., "*Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors*", 2020.

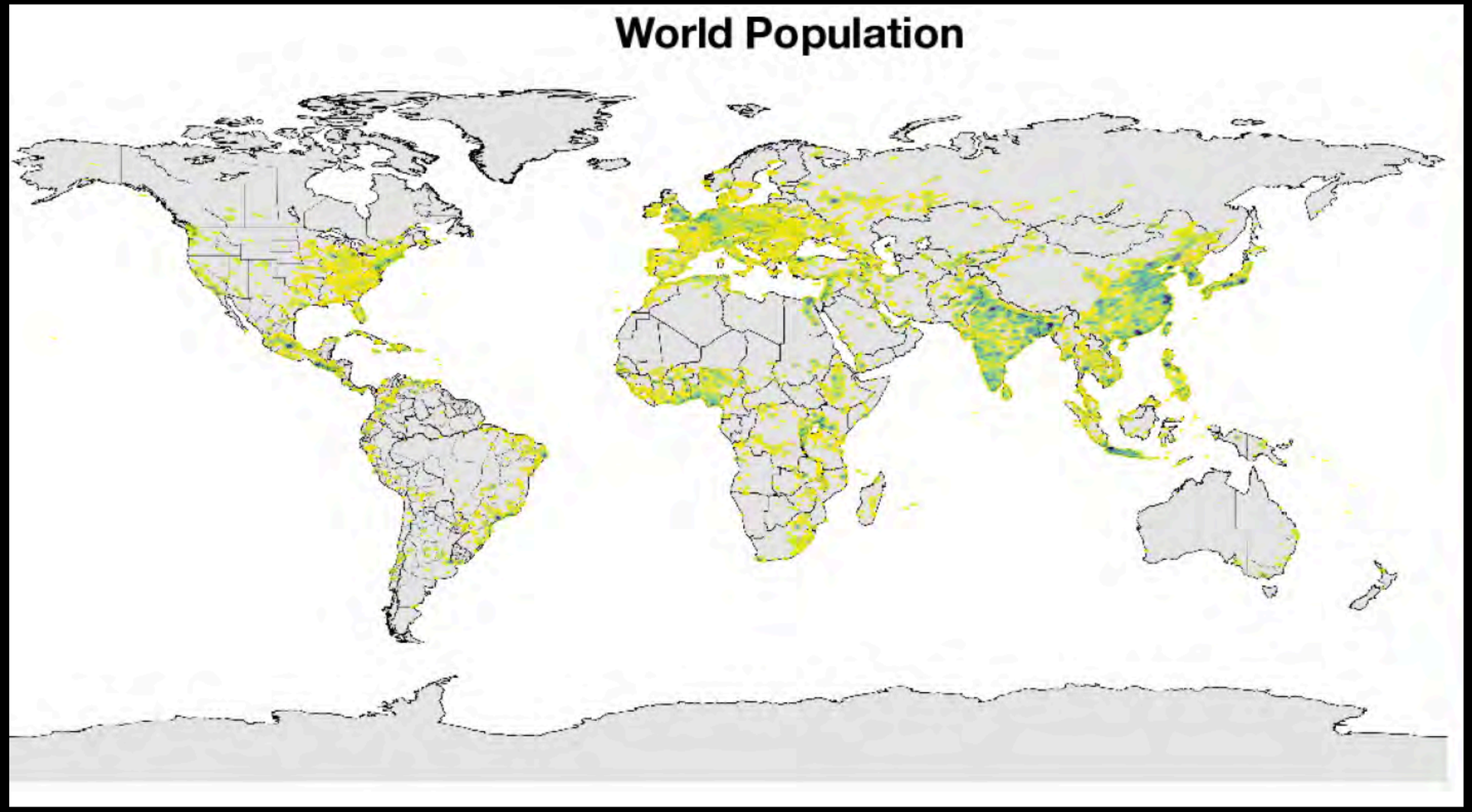


Tabassi et al., "*A Taxonomy and Terminology of Adversarial Machine Learning*", 2019.

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

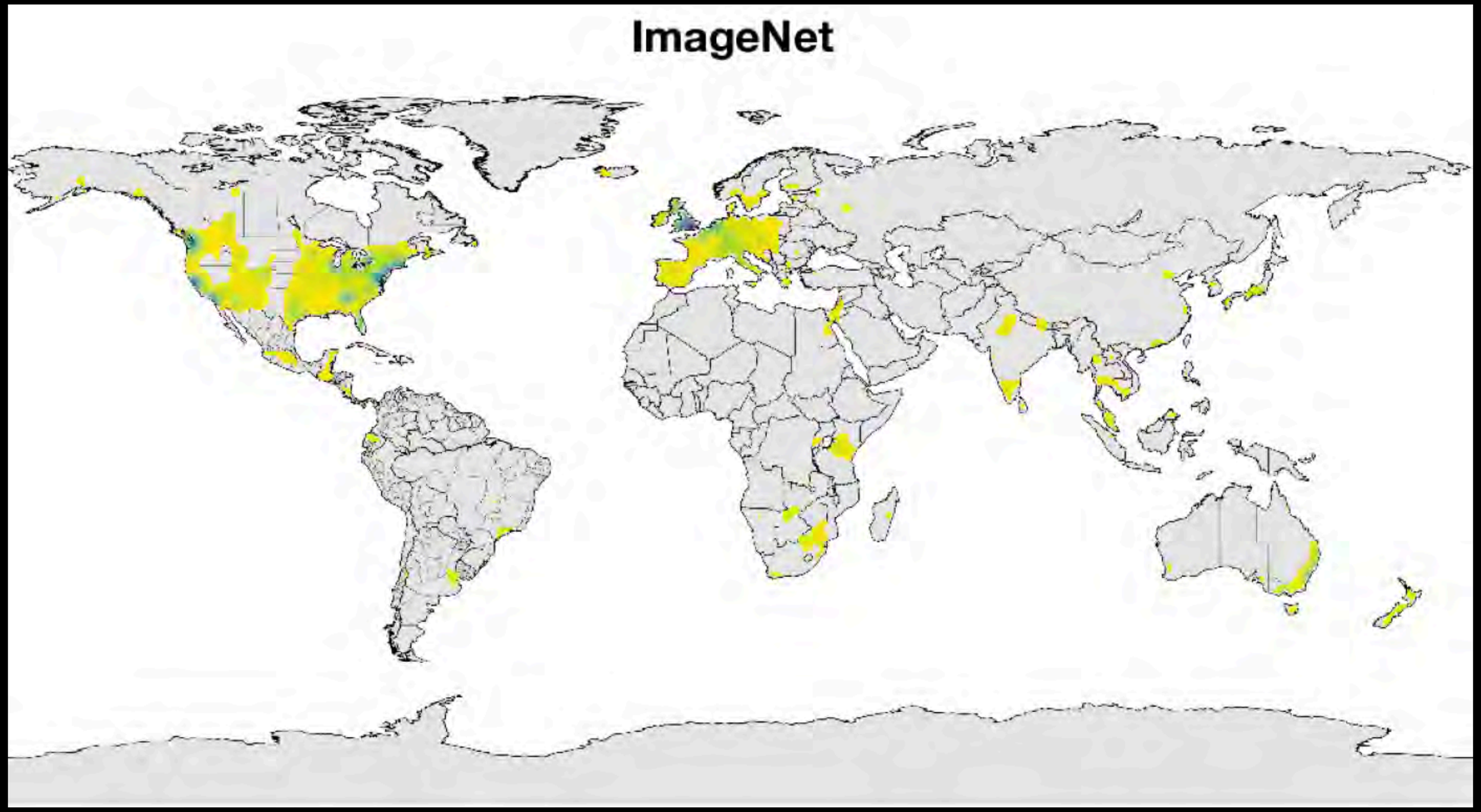

Attacks
Targets
Techniques
Knowledge
Training
Testing (inference)
Data Access
Poisoning
Evasion
Oracle
Extraction
Inversion
Membership Inference

Attacking dateset biases



De Vries et al., "*Does Object RecognitionWork for Everyone?*", *2019.*

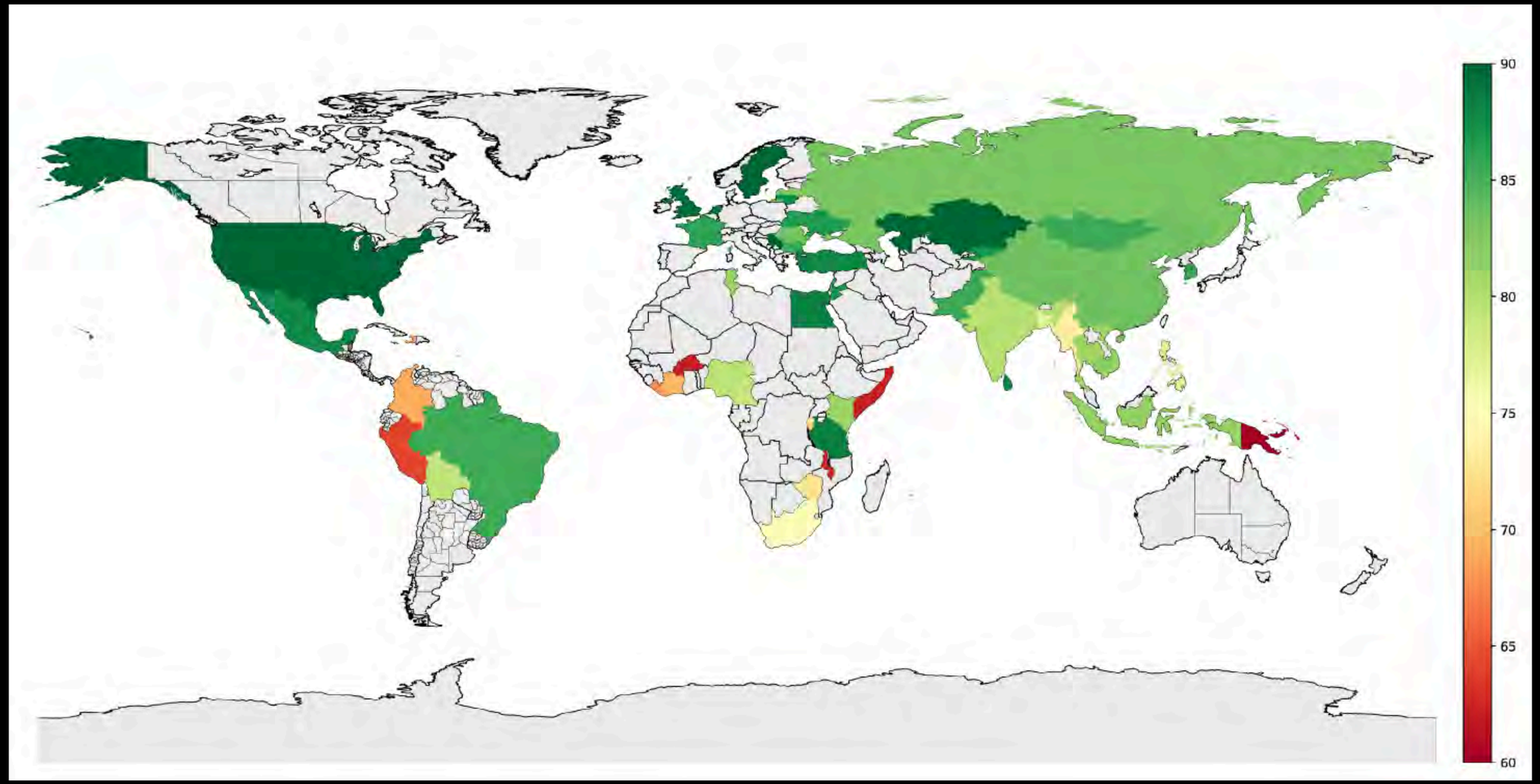
Attacking dateset biases



De Vries et al., "*Does Object RecognitionWork for Everyone?*", *2019.*

Attacking dateset biases

Geographical distribution of classification accuracy



De Vries et al., "*Does Object RecognitionWork for Everyone?*", *2019.*

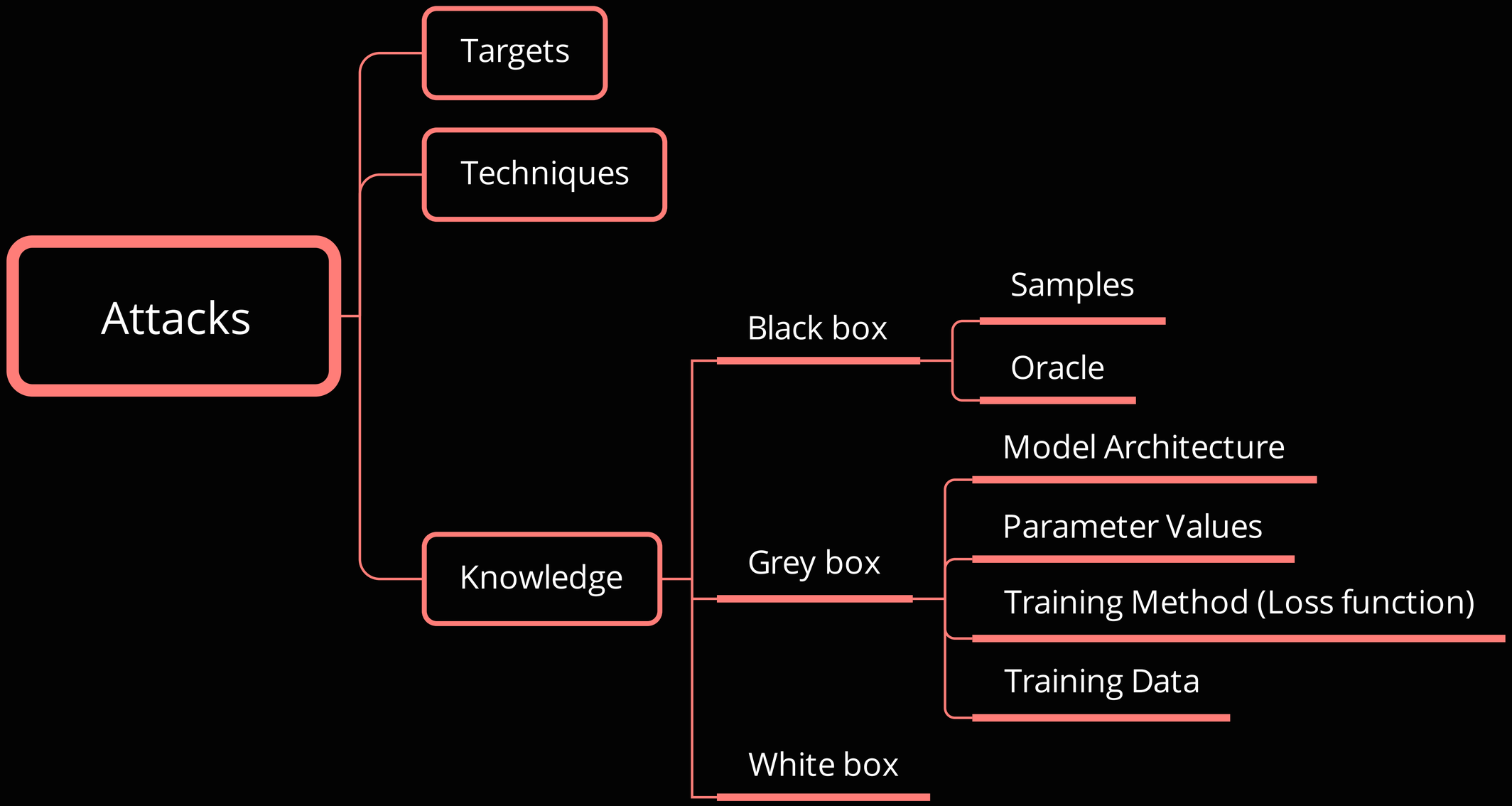
Tabassi et al., "*A Taxonomy and Terminology of Adversarial Machine Learning*", 2019.

- Attacks
  - Targets
  - Techniques
    - Training
      - Data Access
      - Poisoning
    - Testing (inference)
      - Evasion
      - Oracle
        - Extraction
        - Inversion
        - Membership Inference
  - Knowledge

Original

Poisoned

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

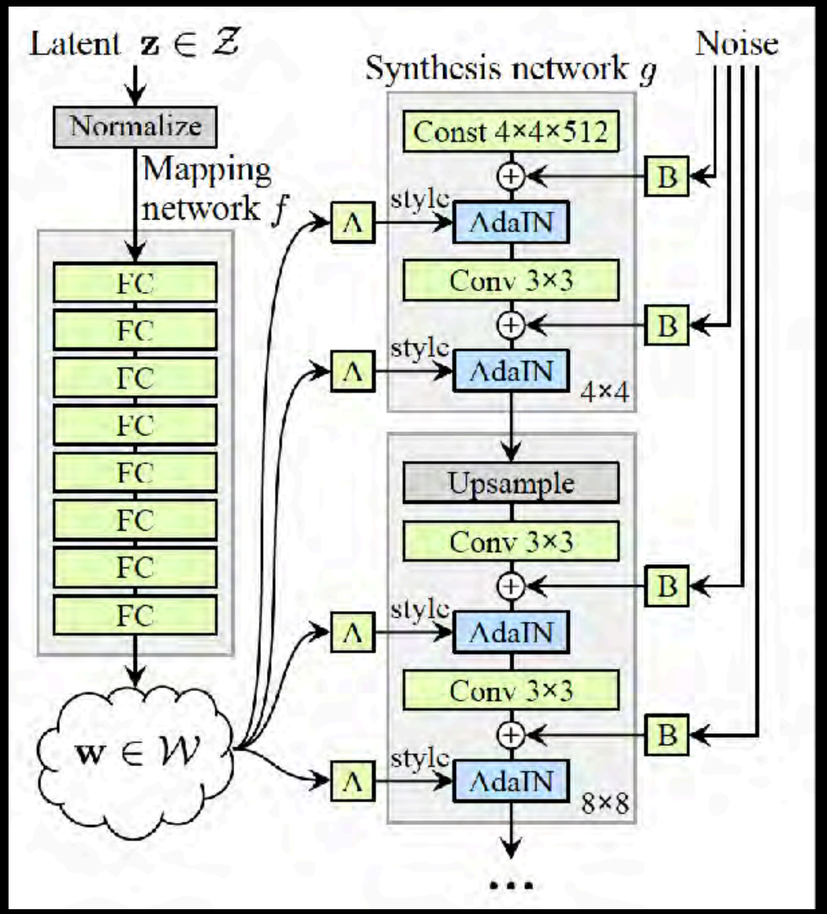


Tabassi et al., "*A Taxonomy and Terminology of Adversarial Machine Learning*", 2019.

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.

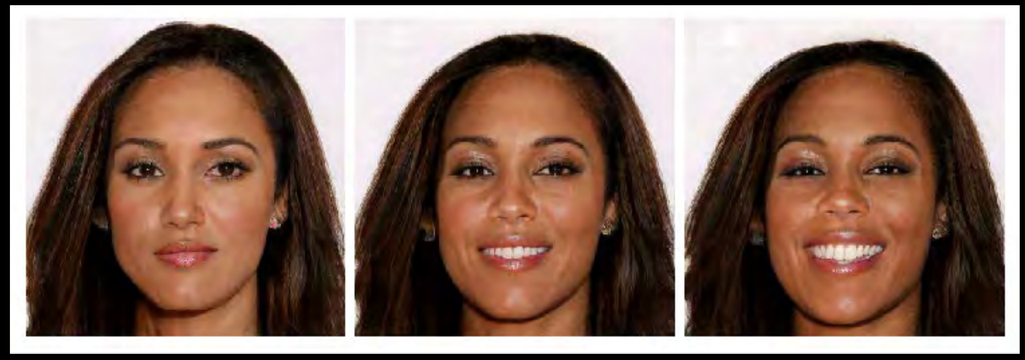
Example case: Synthetic people

Disclaimer: None of these individuals exist!

Plenty of potential good uses:

- Creative purposes
- Virtual characters
- Semantic face editing

Smile edition



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.

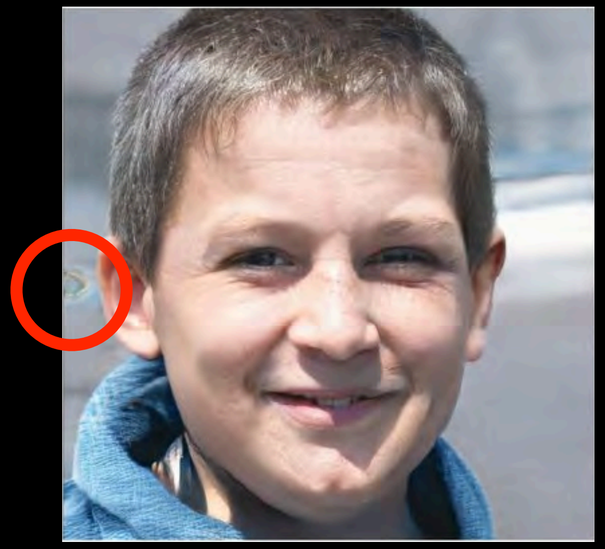
Shen et al. "*Interpreting the Latent Space of GANs for Semantic Face Editing*", 2020.

Example case: Synthetic people

Disclaimer: None of these individuals exist!

Potentially "easy" to spot:

- Generator residuals (in the image)



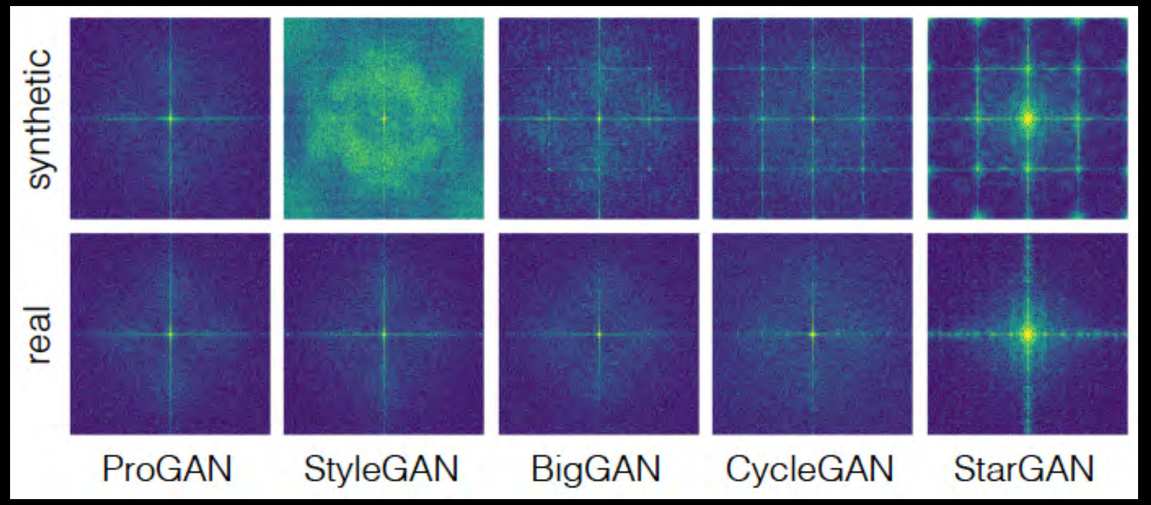
Karras et al. "*A Style-Based Generator Architecture for Generative Adversarial Networks*", 2019.
Karras et al. "*Analyzing and Improving the Image Quality of StyleGAN*", 2020.

Example case: Synthetic people

Disclaimer: None of these individuals exist!

Potentially "easy" to spot:

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



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.

Wang et al. "*CNN-generated images are surprisingly easy to spot... for now*", 2020.

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

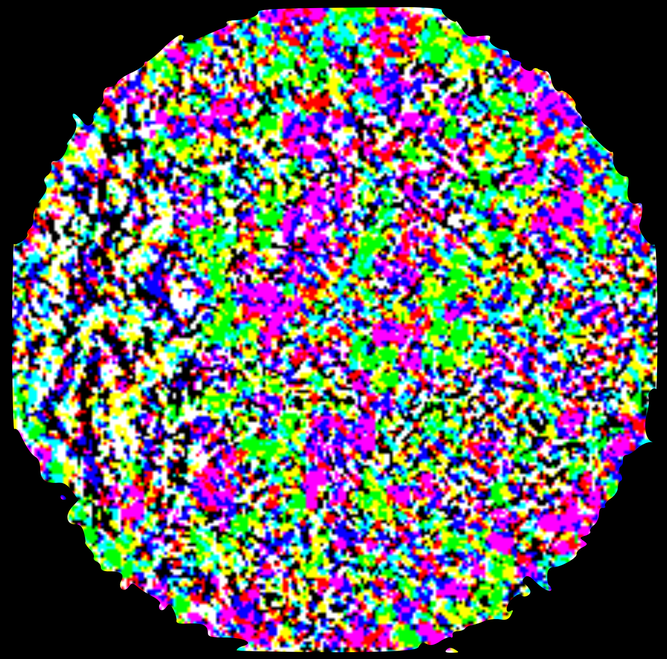
Carlini and Farid "*Evading Deepfake-Image Detectors with White- and Black-Box Attacks*", 2020.

Example case: Synthetic people

Disclaimer: None of these individuals exist!

87% Fake

+



Adversarial noise
(magnified x1000)

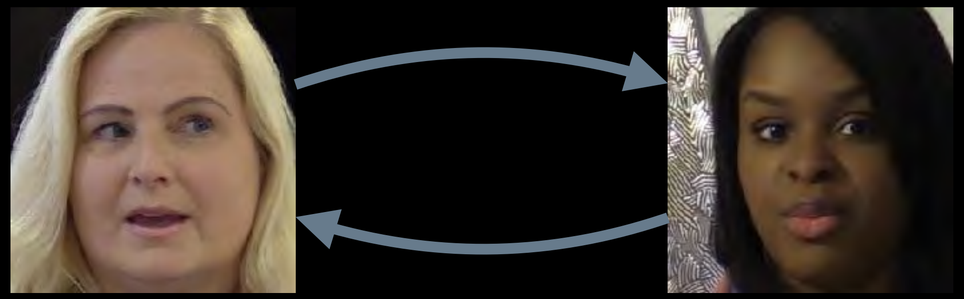
=

1% Fake

Carlini and Farid "*Evading Deepfake-Image Detectors with White- and Black-Box Attacks*", 2020.

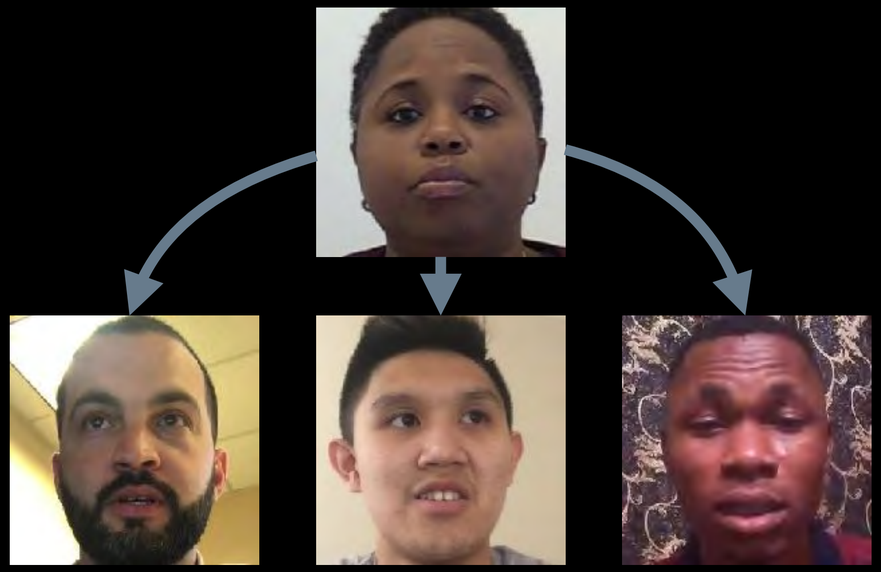
Example case: DeepFakes

Example case: DeepFakes



Pairwise

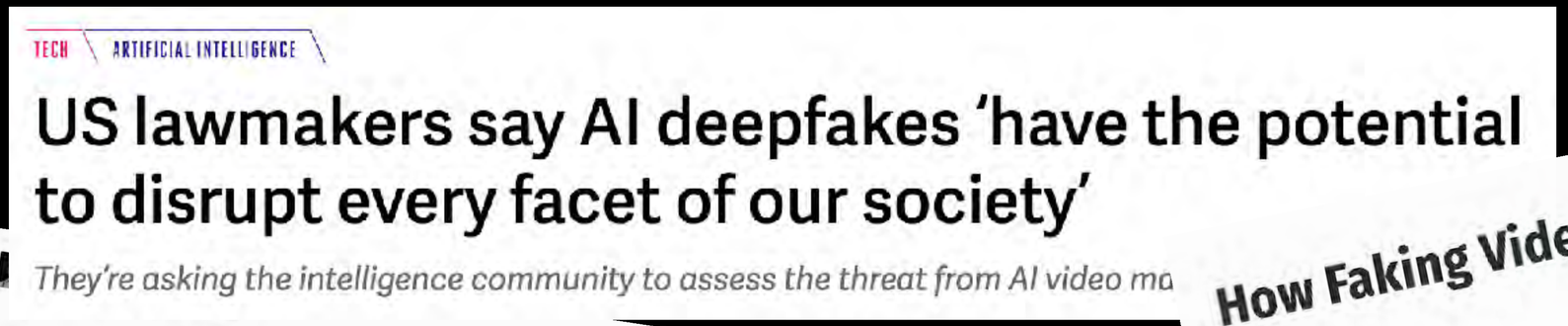
Swap the faces of two individuals - the face of person A is put on the body of person B. Requires many photos of person A and B.

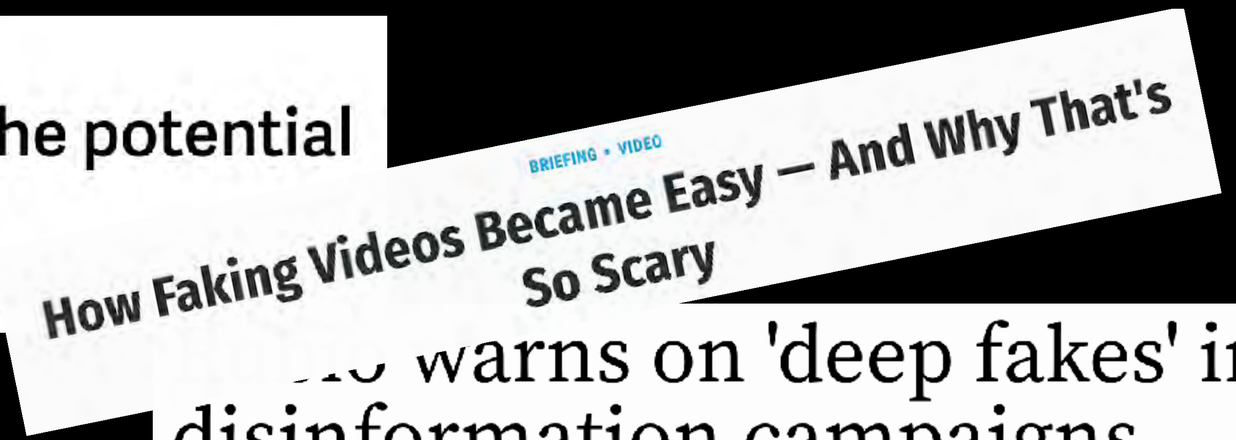


Identity-free

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

Example case: DeepFakes

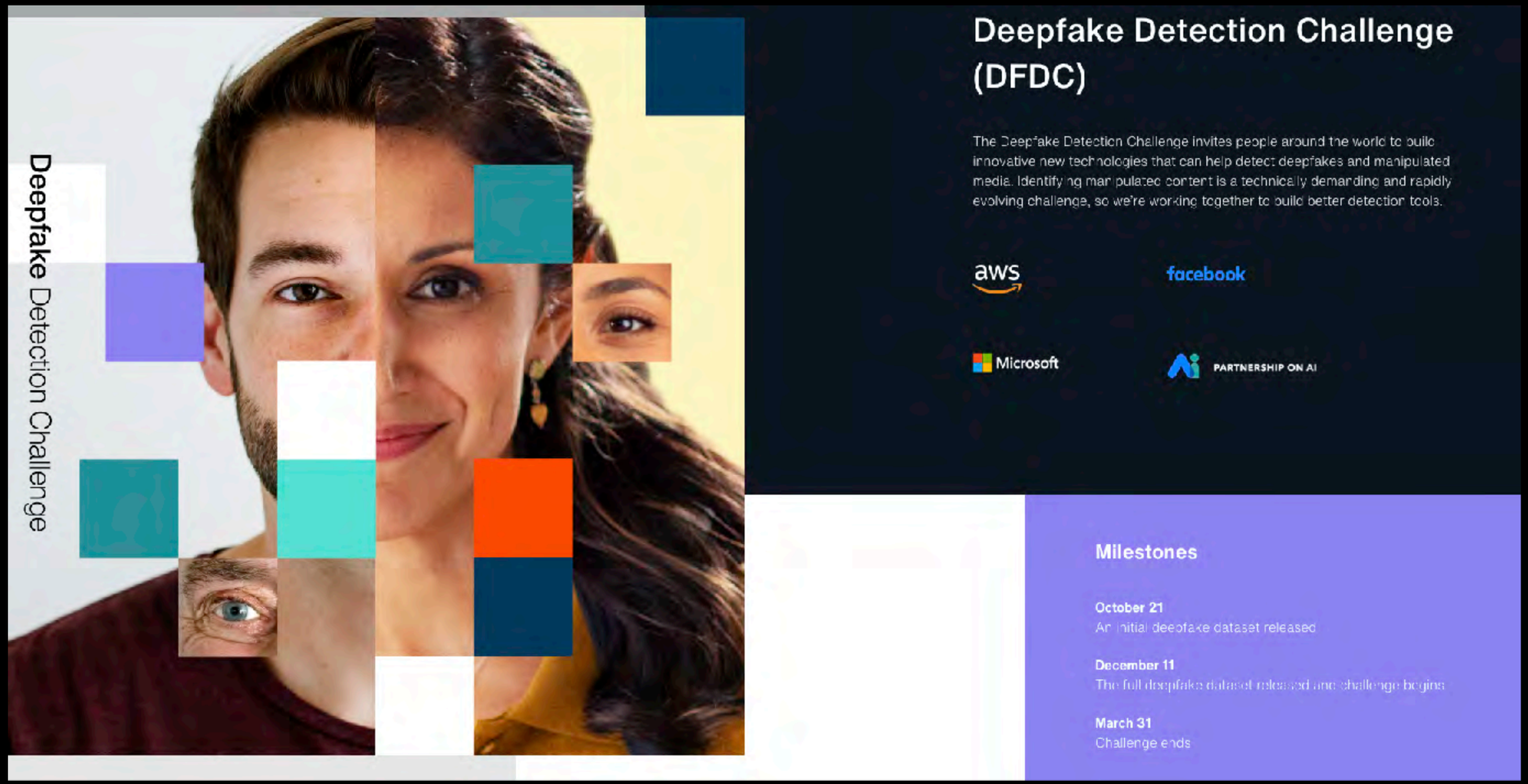




Prevention

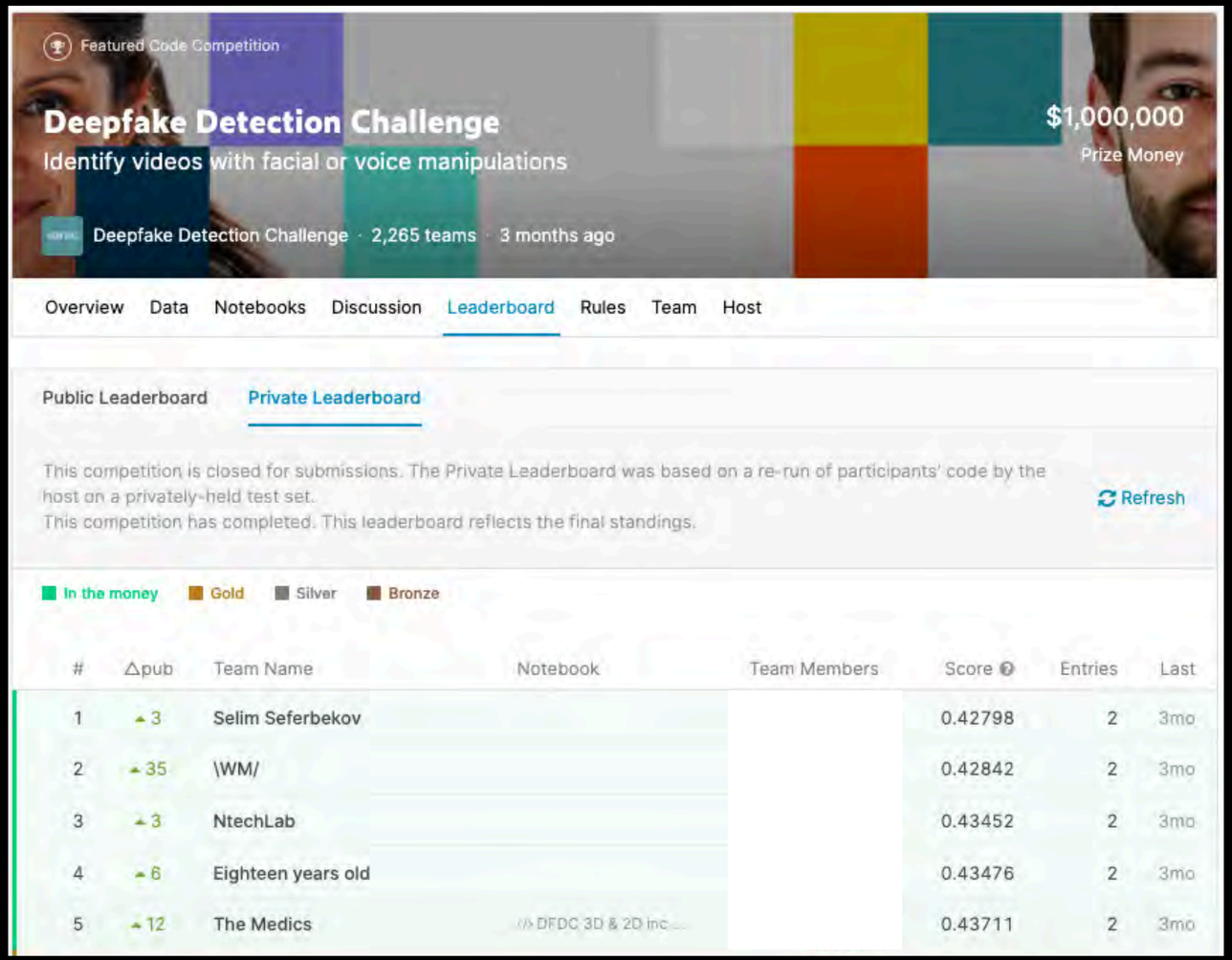
Ask the experts

Example - DFDC competition



Ask the experts

Example - DFDC competition



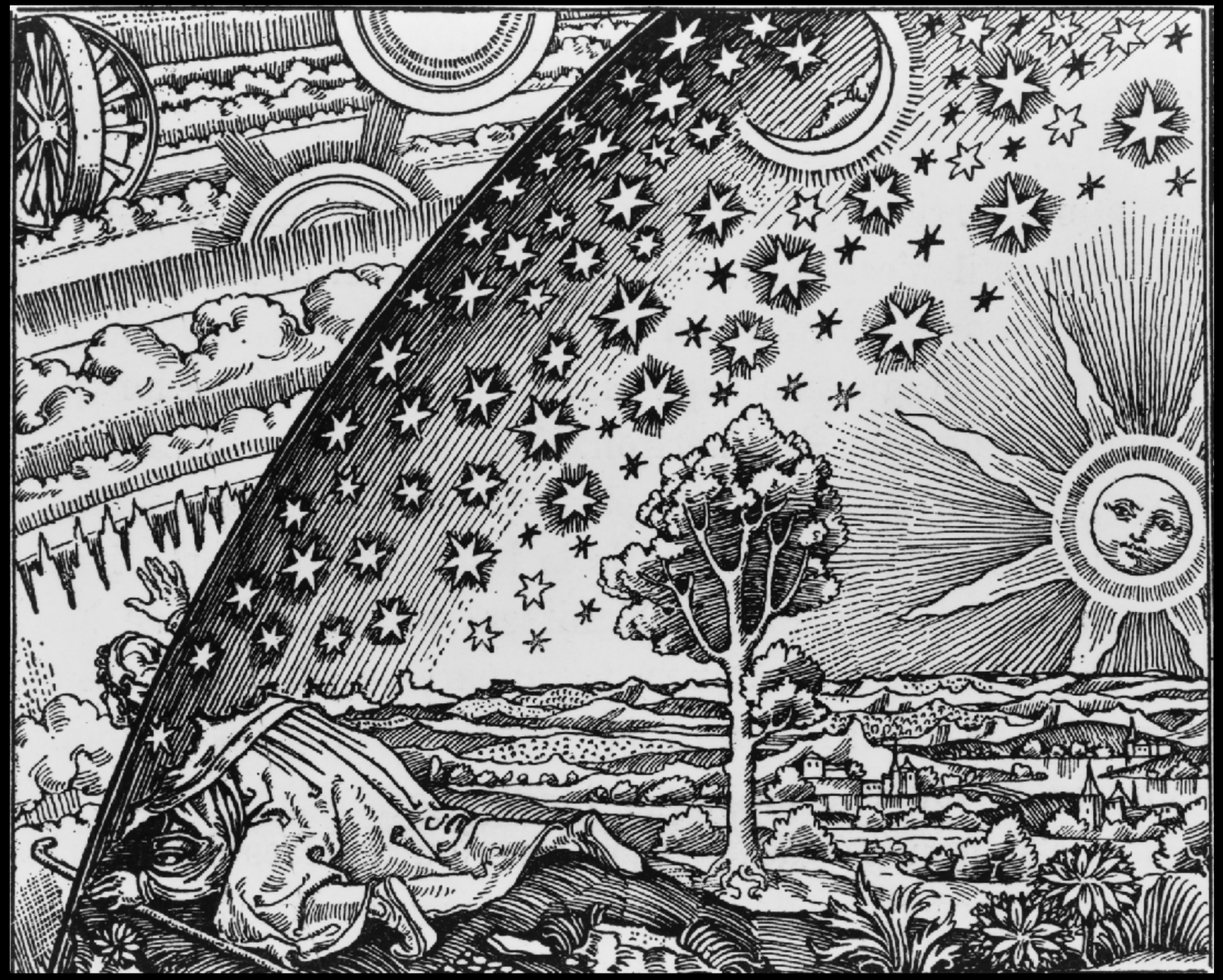
Ask the experts

Example - DFDC competition - Dataset

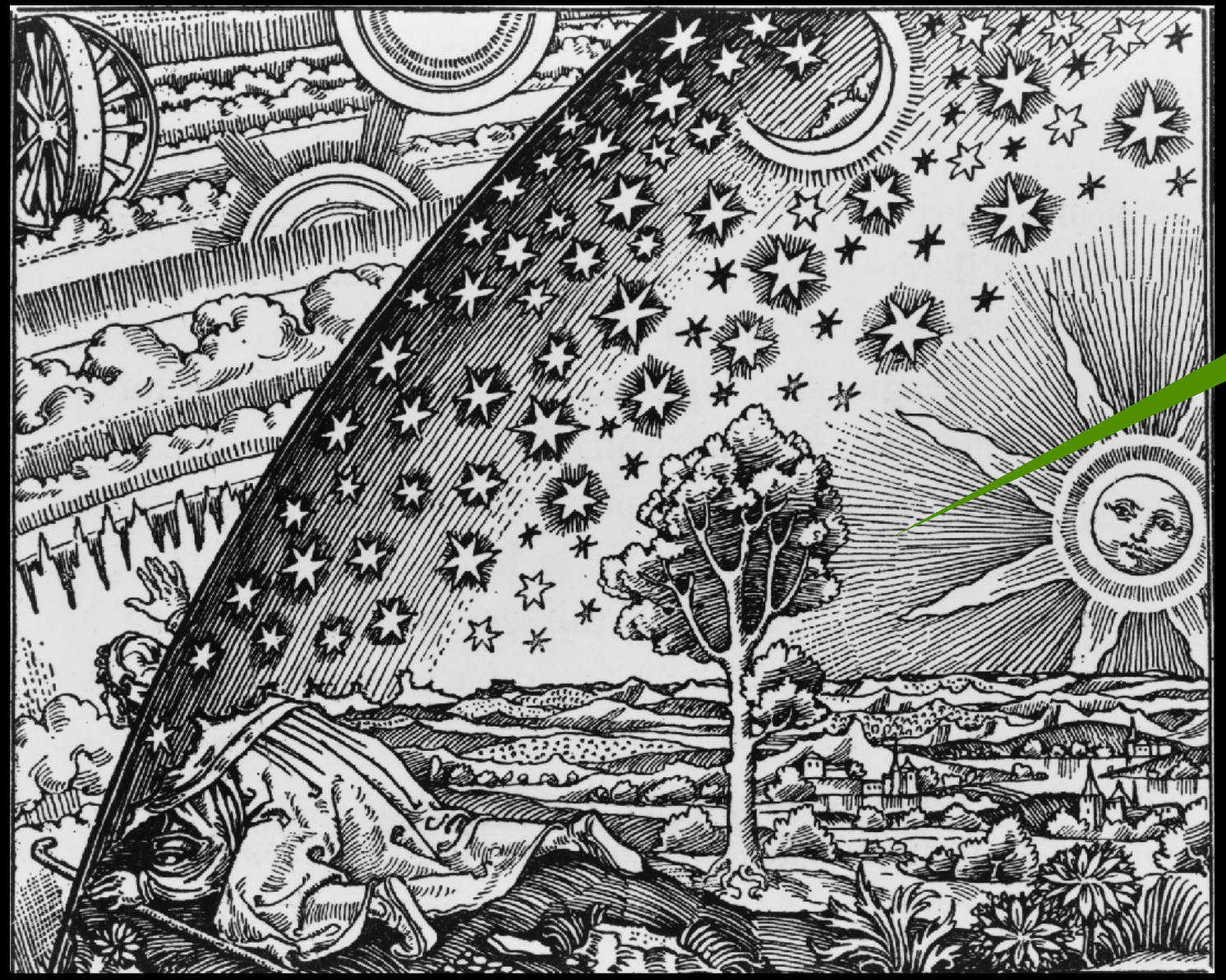
Ask the experts

Example - DFDC competition - Dataset

Domain gap + Distribution shift



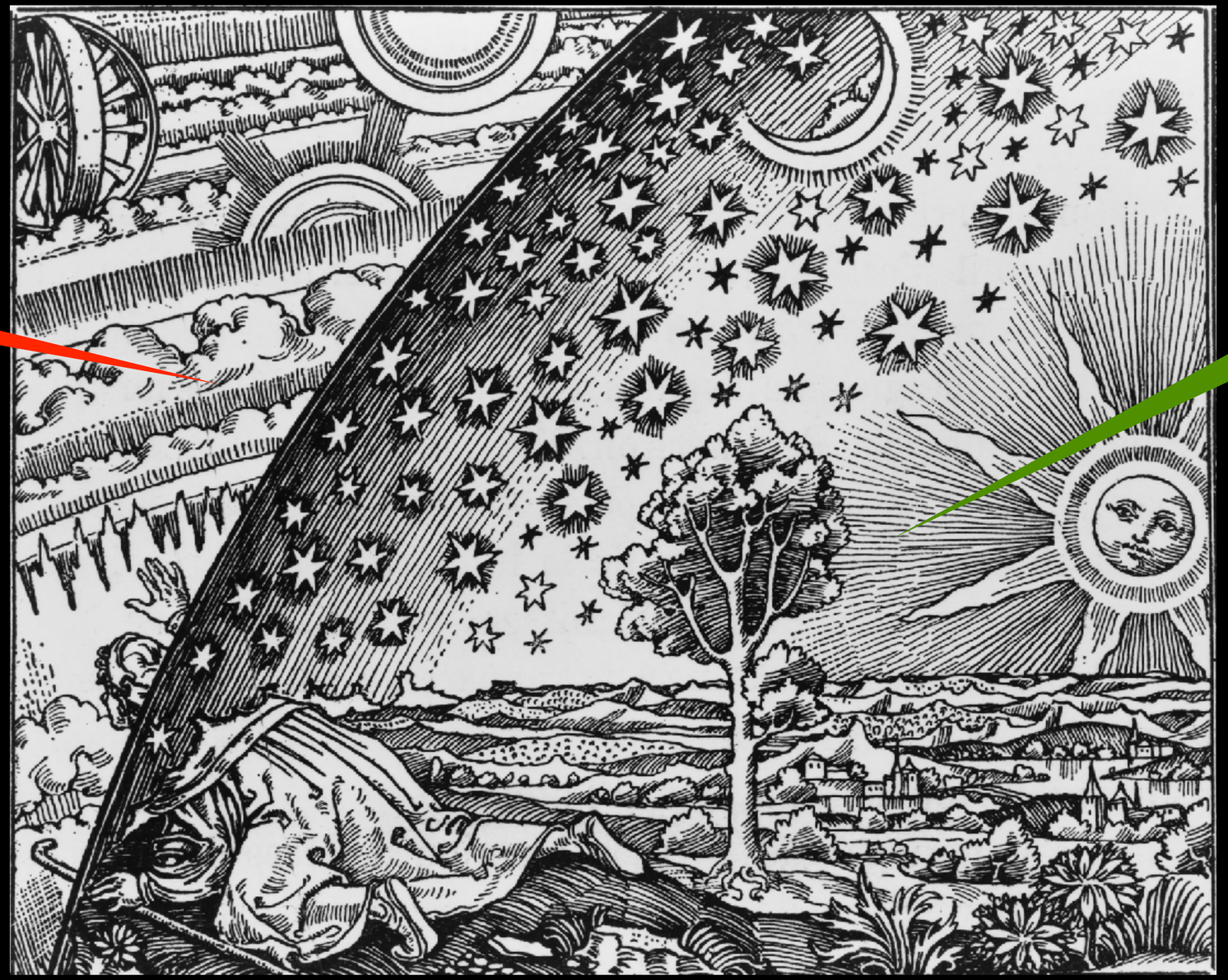
Domain gap + Distribution shift



The test distribution you constructed to validate your algorithm

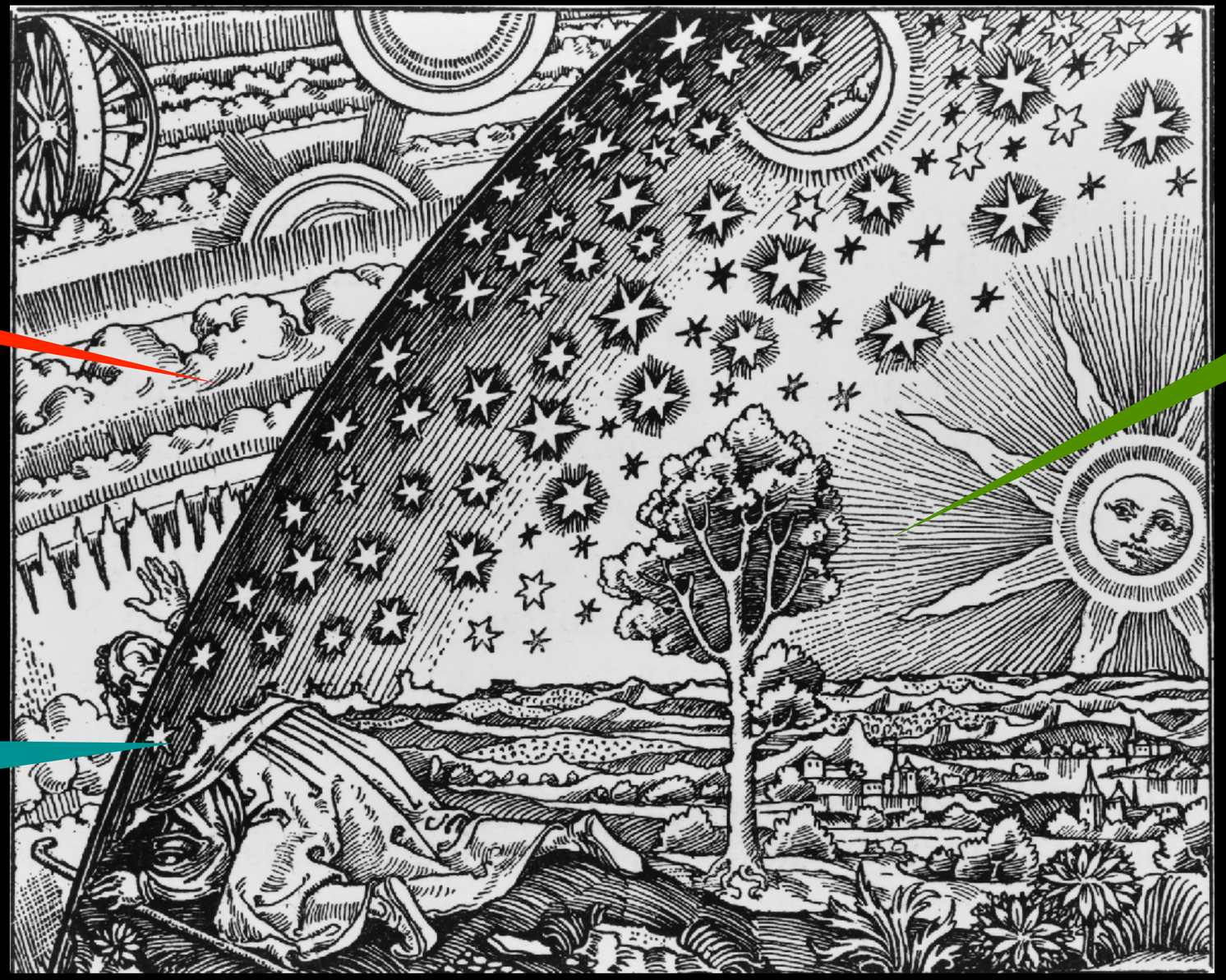
Domain gap + Distribution shift

The real distribution



The test distribution you constructed to validate your algorithm

Domain gap + Distribution shift

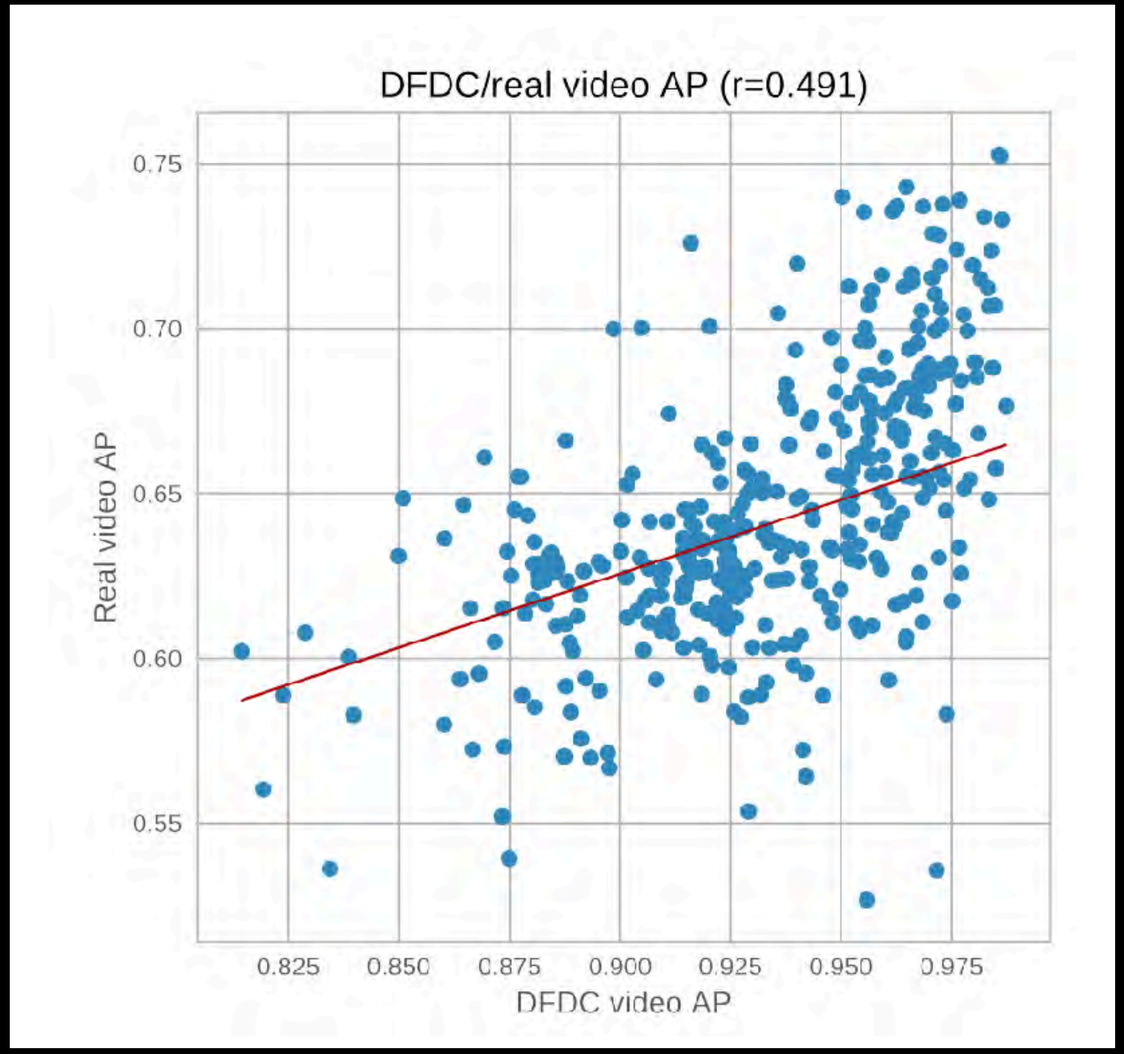


The real distribution

The test distribution you constructed to validate your algorithm

Your algorithm's goal

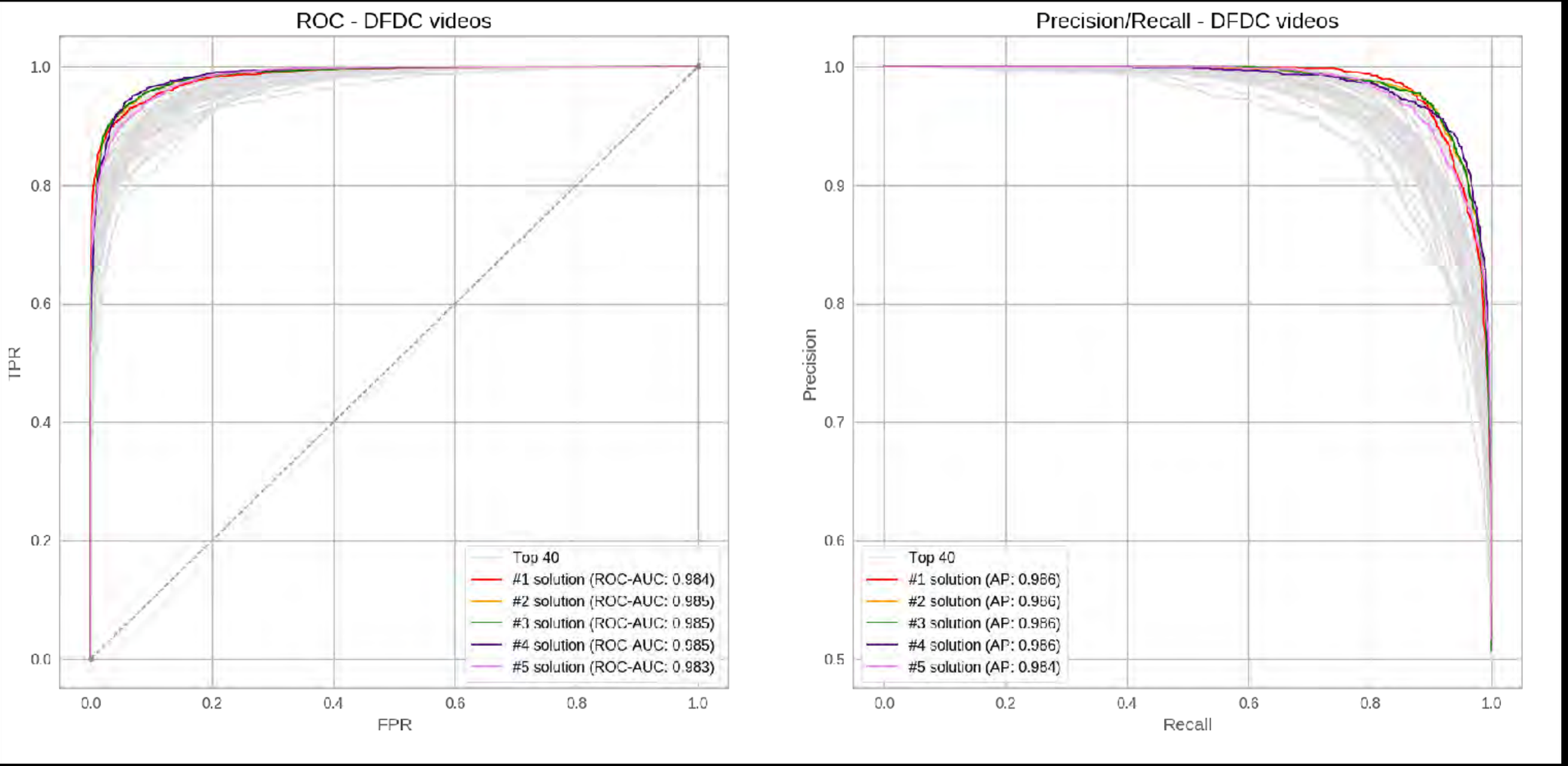
Domain gap + Distribution shift



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

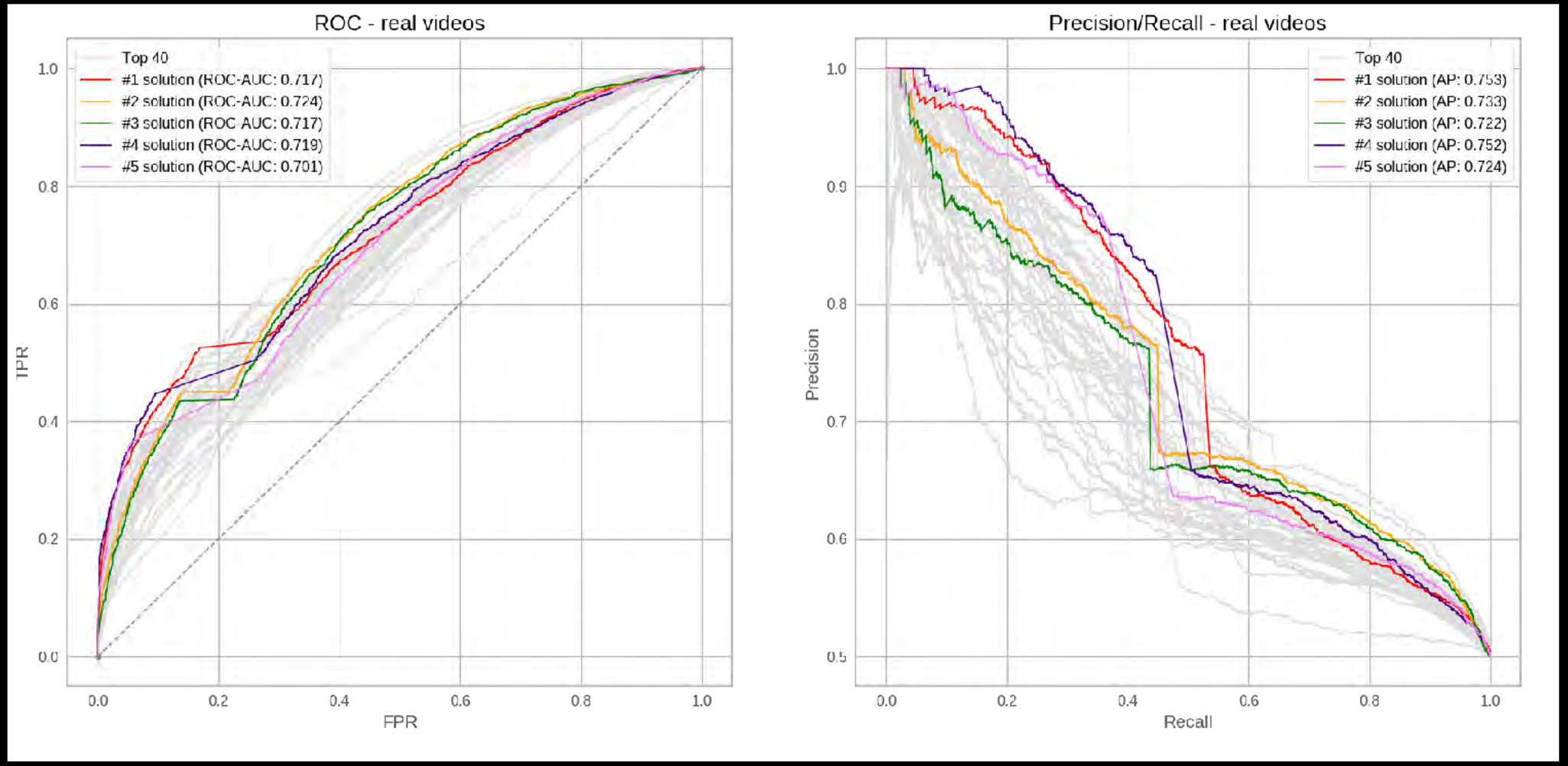
Dolhansky et al. "The DeepFake Detection Challenge Dataset", *https://arxiv.org/abs/2006.07397*

Domain gap + Distribution shift



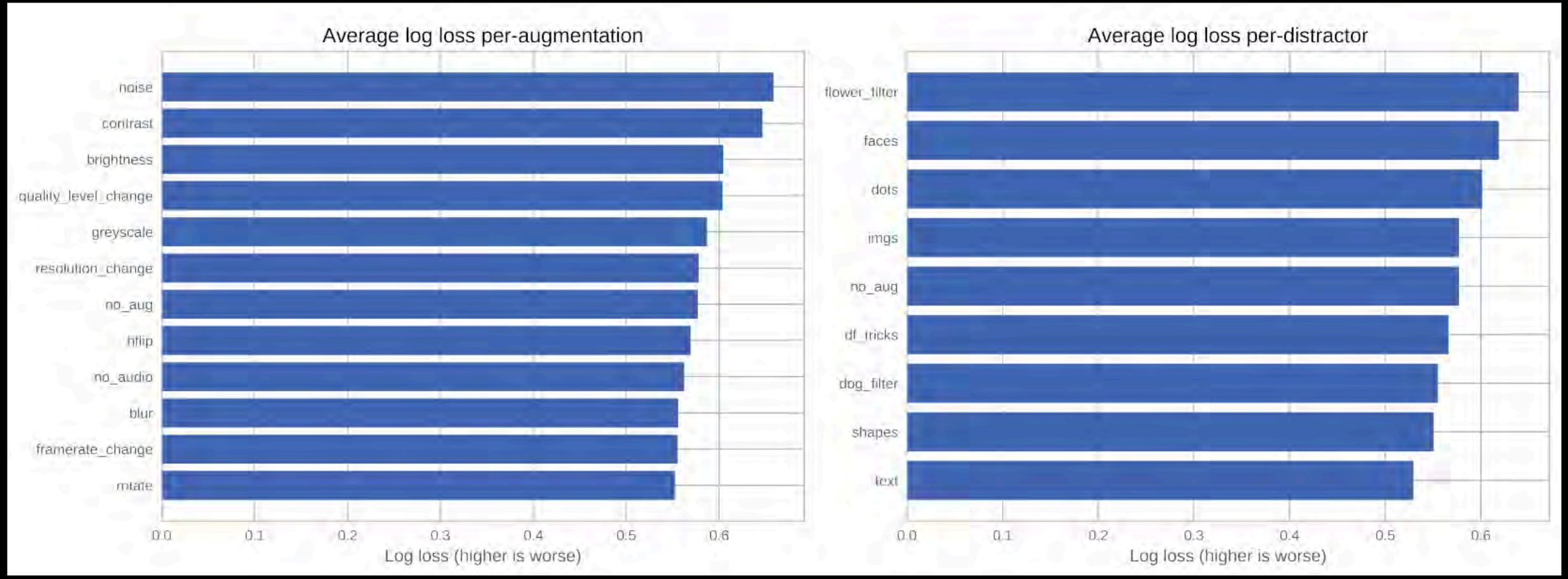
Dolhansky et al. "The DeepFake Detection Challenge Dataset", *https://arxiv.org/abs/2006.07397*

Domain gap + Distribution shift



Dolhansky et al. "The DeepFake Detection Challenge Dataset", *https://arxiv.org/abs/2006.07397*

Domain gap + Distribution shift



Dolhansky et al. "The DeepFake Detection Challenge Dataset", *https://arxiv.org/abs/2006.07397*

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.

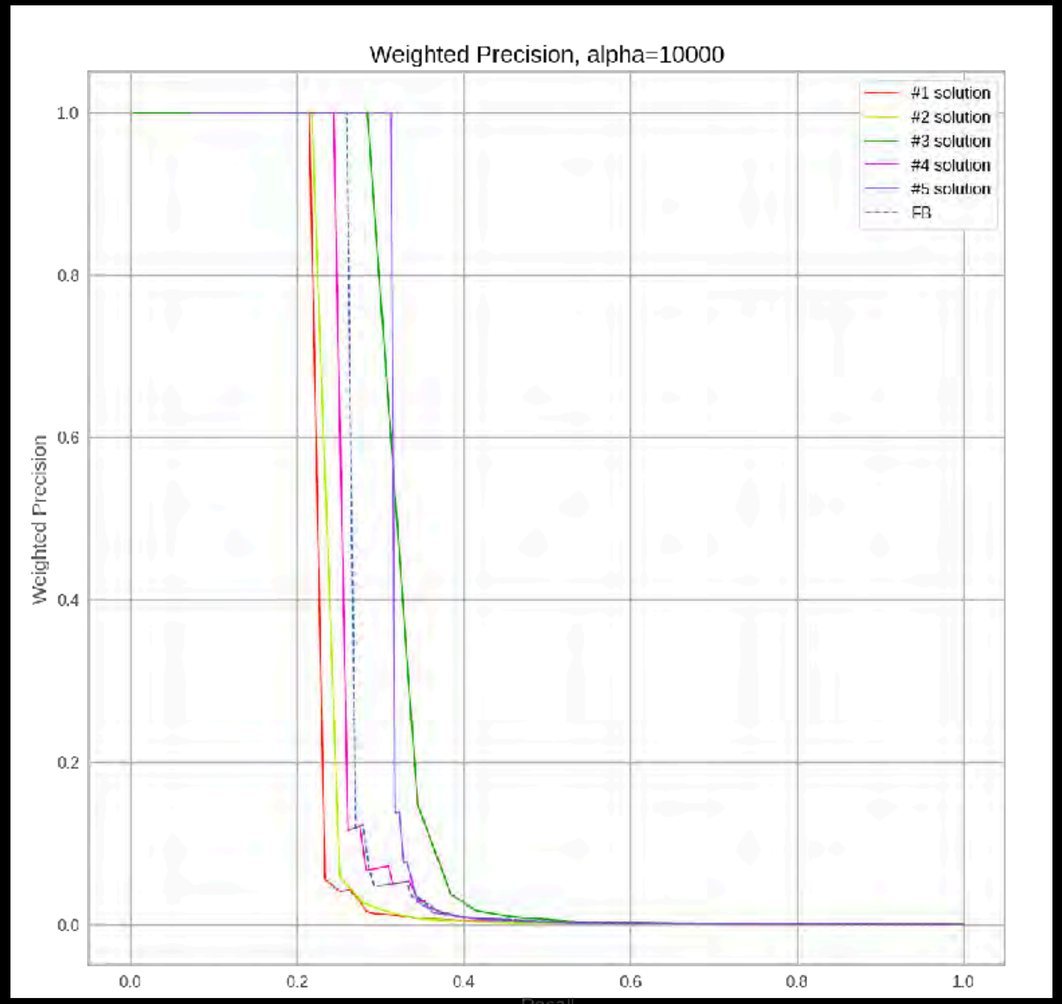
Even with an extremely small false positive rate (which accuracy does not really account for), many more false positives will be detected than real DeepFakes.

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

Dolhansky et al. "The DeepFake Detection Challenge Dataset", *https://arxiv.org/abs/2006.07397*

Domain gap + Distribution shift

(and know your metrics!)



Dolhansky et al. "The DeepFake Detection Challenge Dataset", *https://arxiv.org/abs/2006.07397*

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

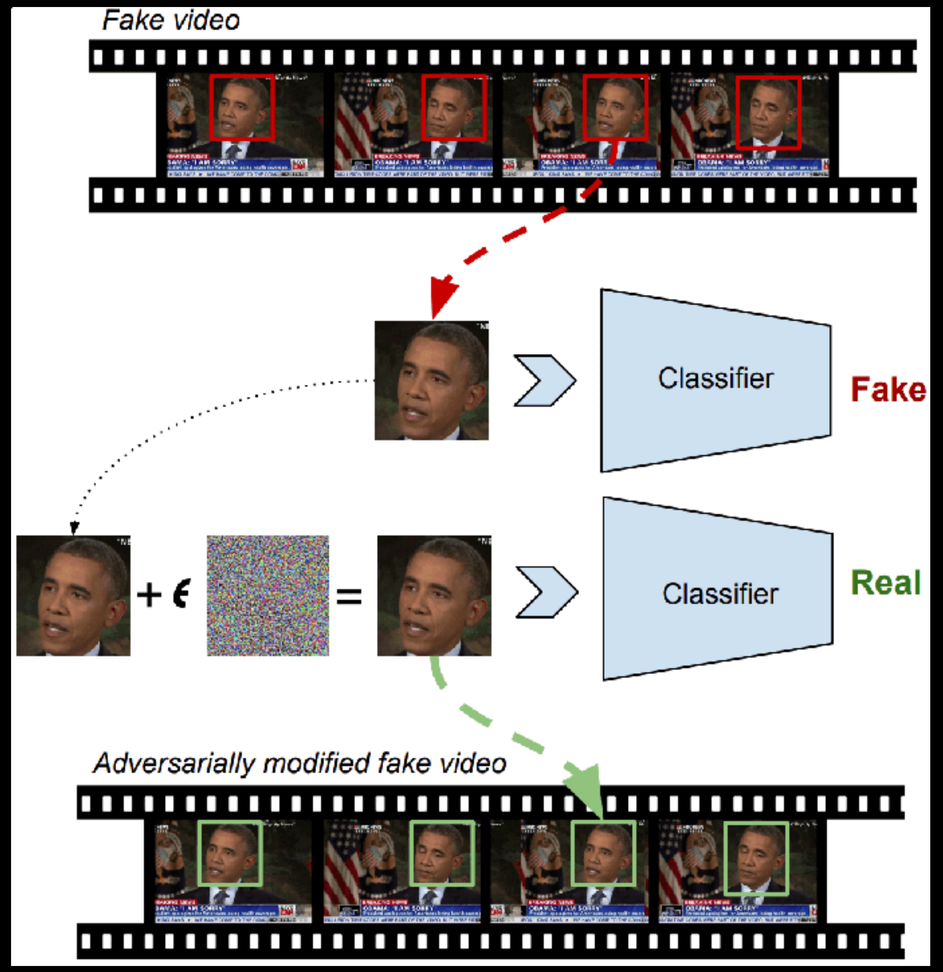
Pros: Good as how well you can keep it secret
Cons: Underestimation of the adversarial agent

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



Neekhara et al. "Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples", 2020.

Reaction

Duct tape fix on Apollo 17 mission

Mitigation

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

Mitigation

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

Mitigation

- Sometimes, been 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

Yang et al. "One-Shot Domain Adaptation For Face Generation", 2020.

Mitigation

- Sometimes, been 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 tess 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