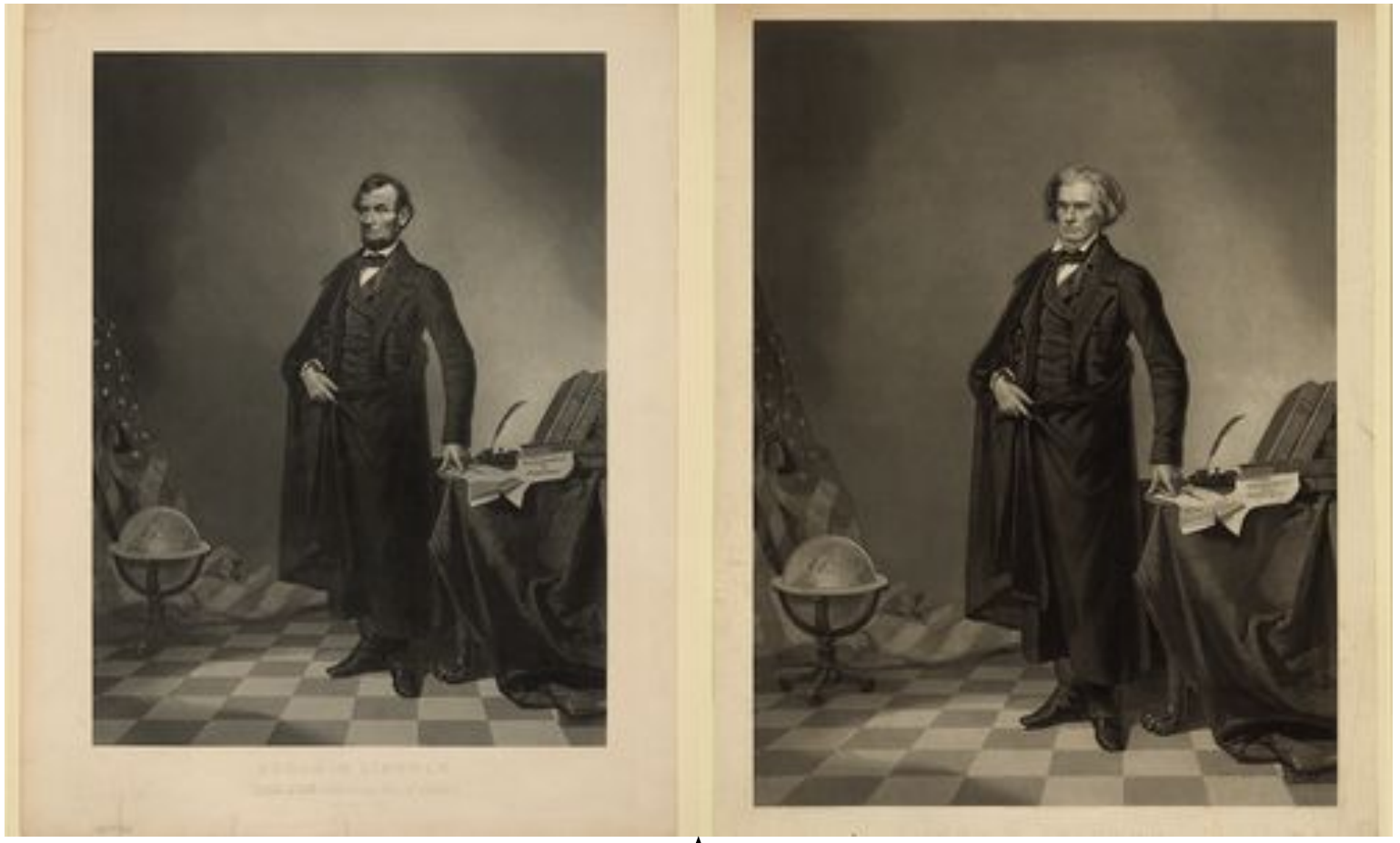
GenAI Perspectives from the MediFor and SemaFor Programs

Matt Turek

October 2023

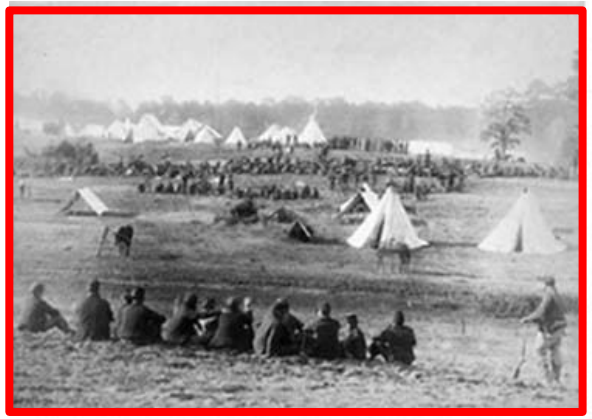
A brief history of media manipulation technologies

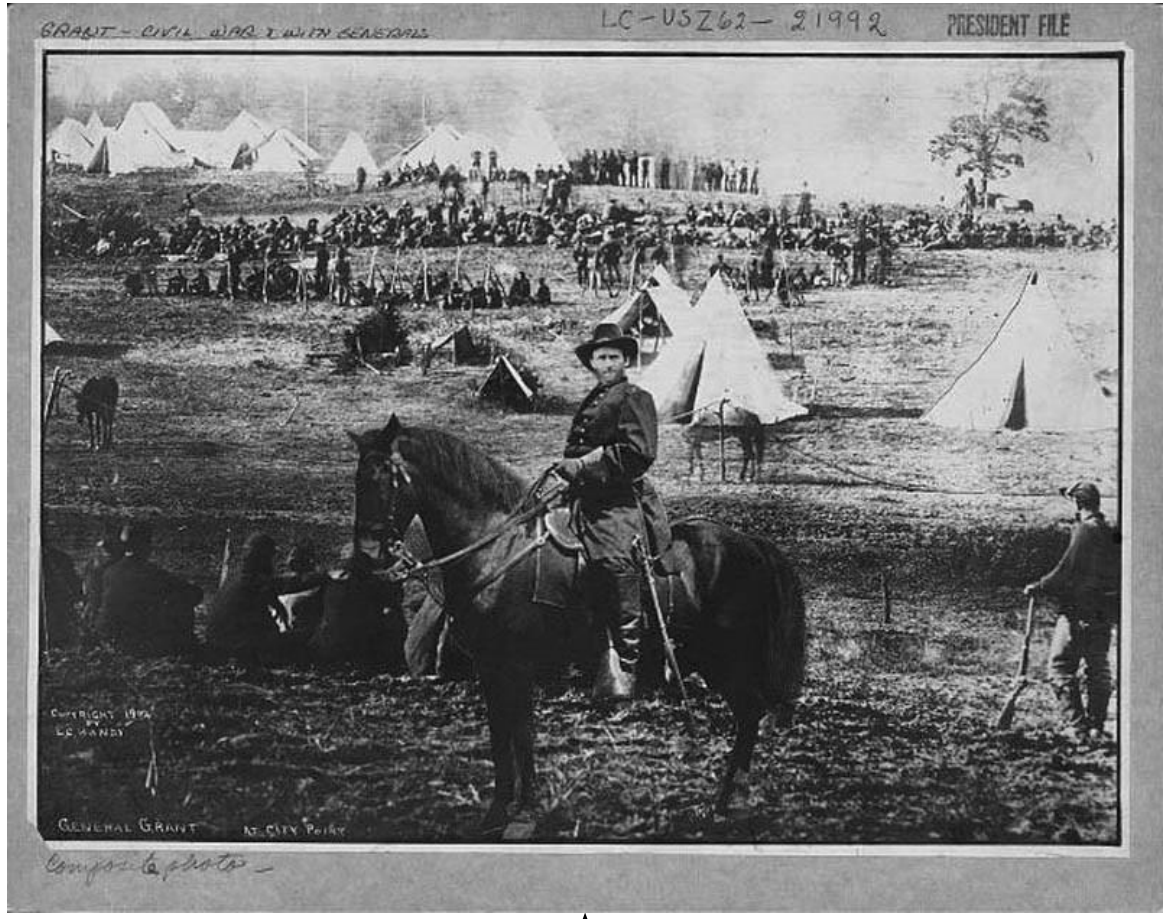


~1865

1860 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2010 2020

A brief history of media manipulation technologies





1902

1860 | 1910 | 1920 | 1930 | 1940 | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 | 2010 | 2020

A brief history of media manipulation technologies

1917

1860 | 1910 | 1920 | 1930 | 1940 | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 | 2010 | 2020

A brief history of media manipulation technologies

1939

1860 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2010 2020

A brief history of media manipulation technologies

~80 years

4 years

1860 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2010 2020

A brief history of media manipulation technologies

~80 years

4 years

Text to image

1860 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2010 2020

Speculative falsified media threats identified circa 2019

Targeted Personal Attacks

Peele 2017

AI Multimedia Algorithms

Highly realistic video

Ransomfake concept: Identity Attacks as a service (IAaaS)

Bricman 2019

AI Multimedia Algorithms

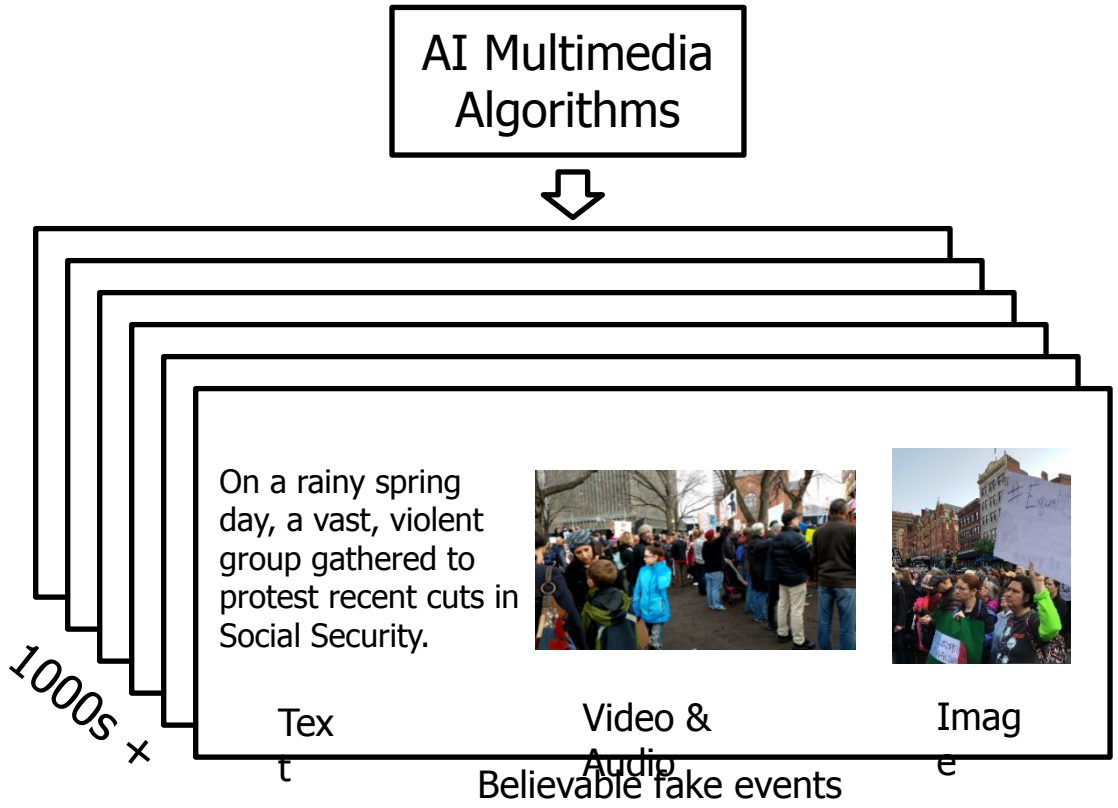
Forged Evidence

Identity Attacks

Examples of possible fakes:

- Substance abuse
- Foreign contacts
- Compromising events
- Social media postings
- Financial inconsistencies
- Forging identity

Generated Events at Scale

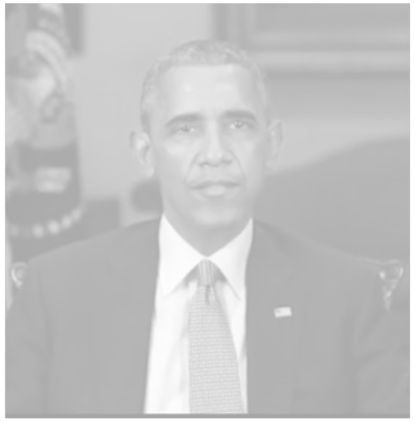


Undermines key individuals and organizations

2019's speculative falsified media threats are now real

Targeted Personal Attacks

Peele 2017



Highly realistic video

Ransomfake concept: Identity Attacks as a service (IAaaS)

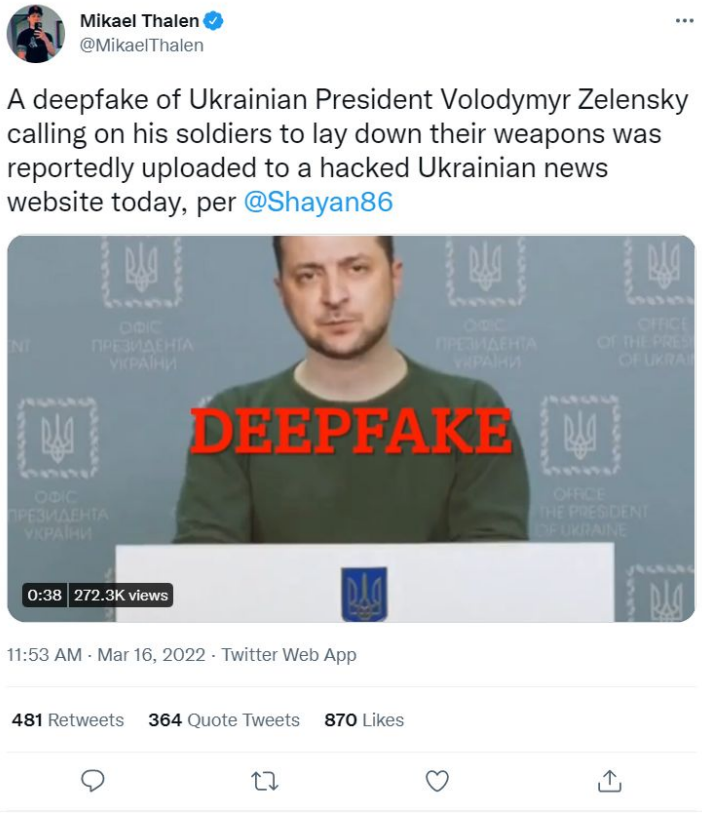
Bricman 2019

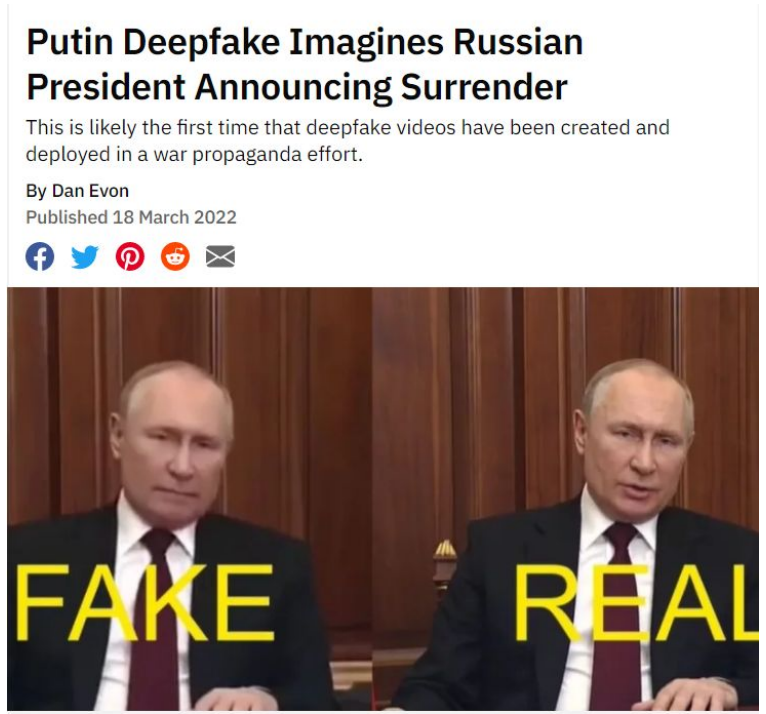
AI Multimedia Algorithms

⇩

Identity Attacks

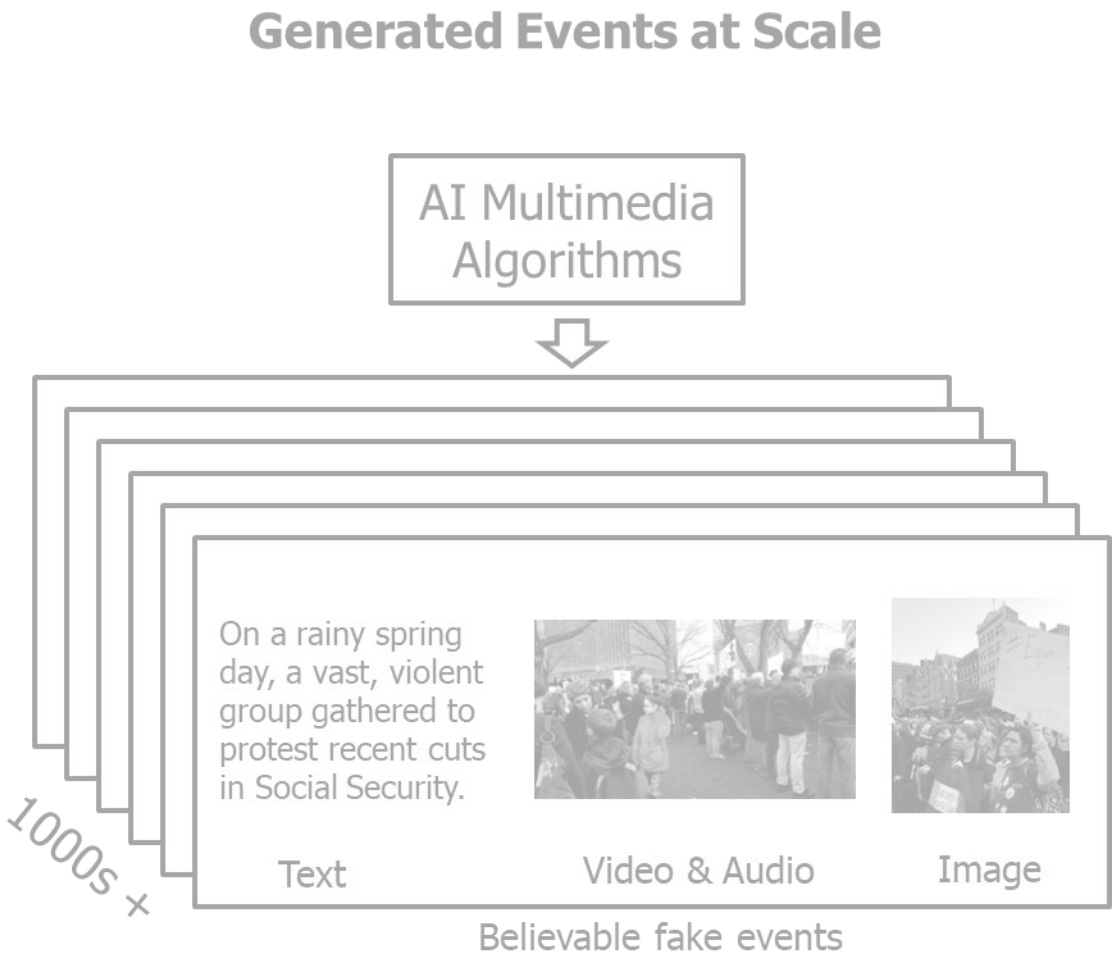
2022





2019's speculative falsified media threats are now real

2019



2023

Fake historical events

AI generated images of the fictional 2001 Great Cascadia 9.1 Earthquake & Tsunami. 1, 2 / Reddit

https://interestingengineering.com/culture/ai-create-images-fake-events

IMAGES: REDDIT VIA U/ARCTIC_CHILEAN

https://www.vice.com/en/article/k7zqdw/people-are-creating-records-of-fake-historical-events-using-ai

DARPA programs

- **Media Forensics (MediFor)**
  - Program Scope

    Detect indicators of digital, physical or semantic manipulation in images and video to produce quantitative measures of integrity. Enable integrity assessment of visual media assets at scale.

  - Program kicked off May 2016, ended August 2020

- **Semantic Forensics (SemaFor)**
  - Program Scope

    Create rich semantic algorithms that automatically detect, attribute, and characterize falsified multi-modal media to defend against large-scale, automated disinformation attacks

  - Program kicked off August 2020, anticipated completion in September 2024

Media authentication requires addressing four key problems

Detection

- Determine whether a media asset (image, video, audio, text) is real, manipulated, or AI generated

Attribution

- Determine the source of the media asset and whether that source is consistent with the purported source
- Attribution to organizations, individuals, tools, techniques
- Answer impacts how the U.S. government responds

Characterization

- Identify rationale or intent behind the manipulation
  - Target audience
  - Target impact

Evidence

- Not just a determination, but supporting evidence

Detection & fusion algorithms

Detection algorithms analyze media content (pixels) or attributes to determine if manipulation has occurred

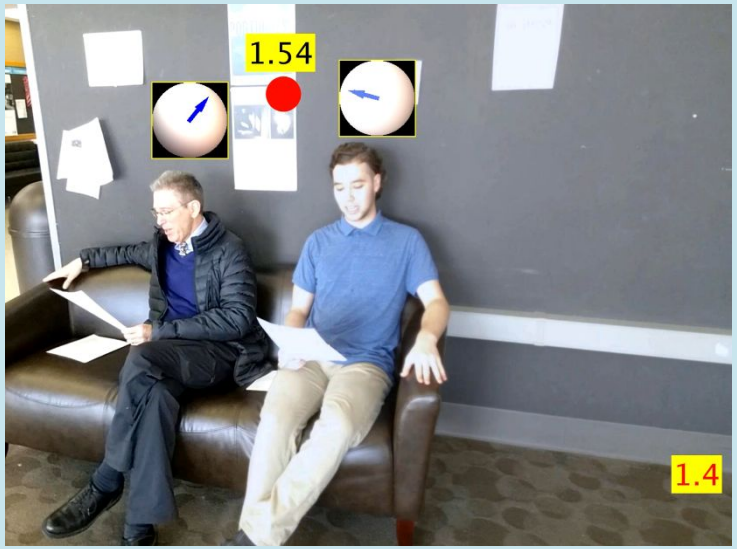
Fusion algorithms combine information across multiple detectors

Create a unified score for each media asset

Digital Integrity

Compression artifacts, PRNU, frequency artifacts

Physical Integrity



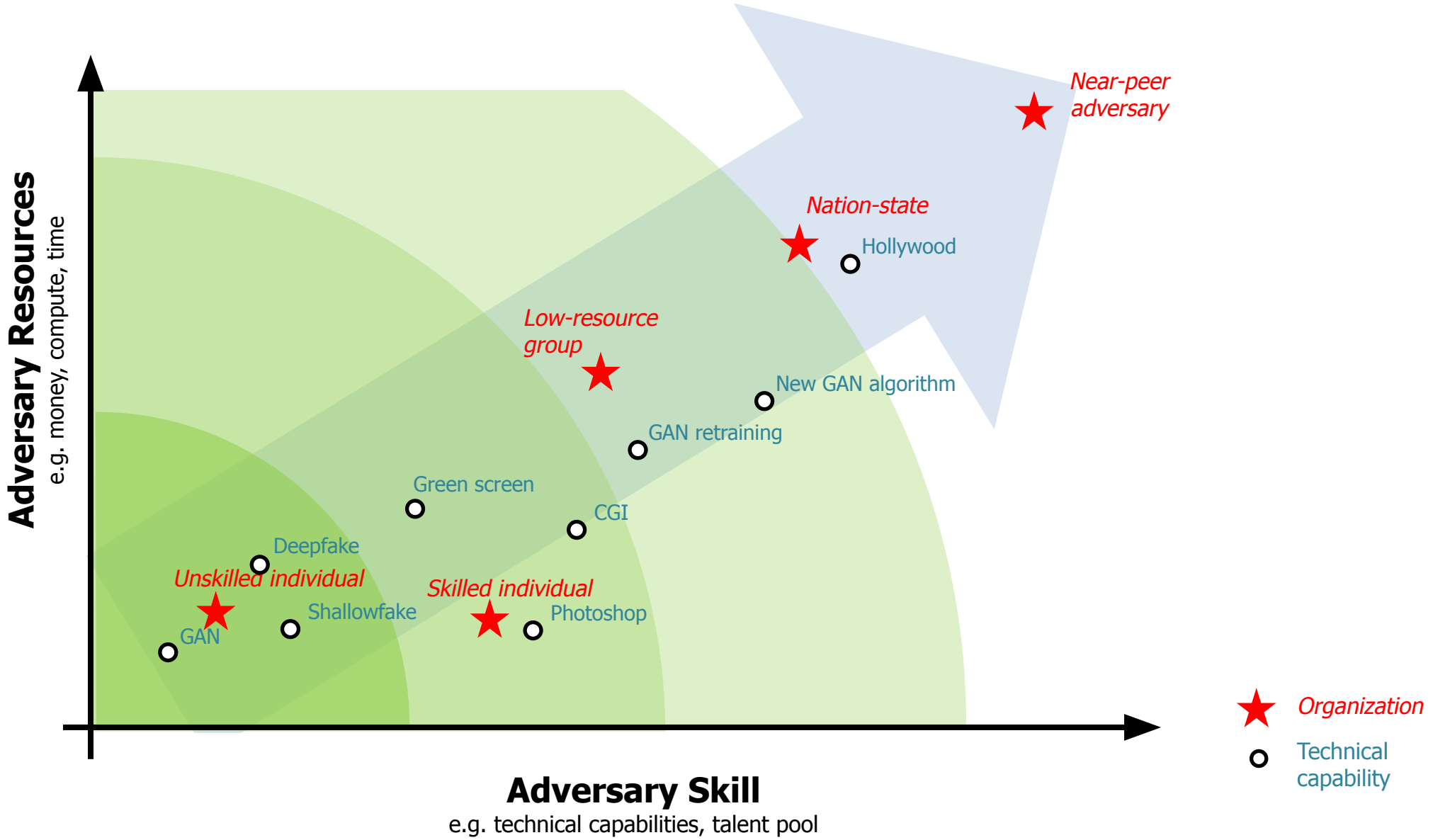
Geometry, lighting

Semantic Integrity

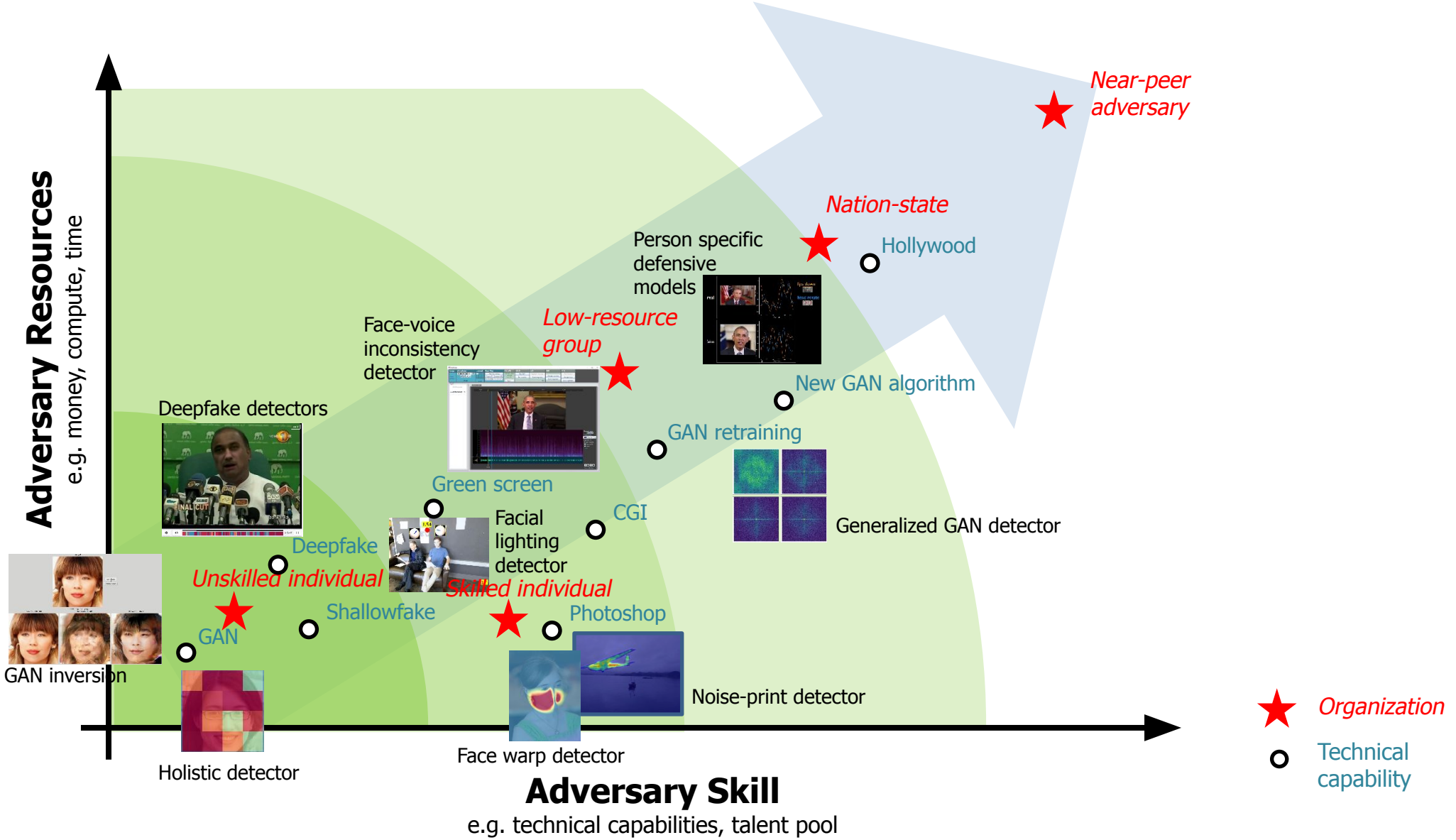
GPS: 43.9196, -78.89419
Time: 2015/05/29 14:14:30

Comparison to other information

Adversarial landscape vs. detection algorithms



Adversarial landscape vs. detection algorithms



GAN detector defends against high-resource actors

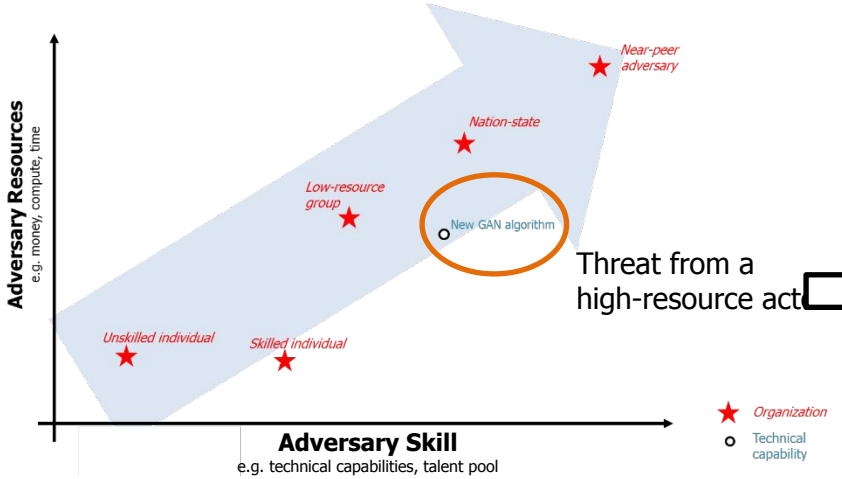
Purdue, Kitware, SRI, NVIDIA

Challenge: can detectors identify images from a novel GAN created by a high-resource actor, even without any data from it?

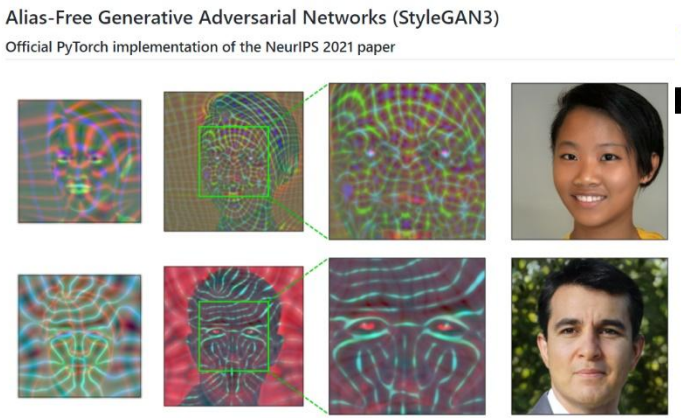
Pro-Chinese Inauthentic Network (2020)

GAN Source: Graphika

Large-scale sock puppet accounts

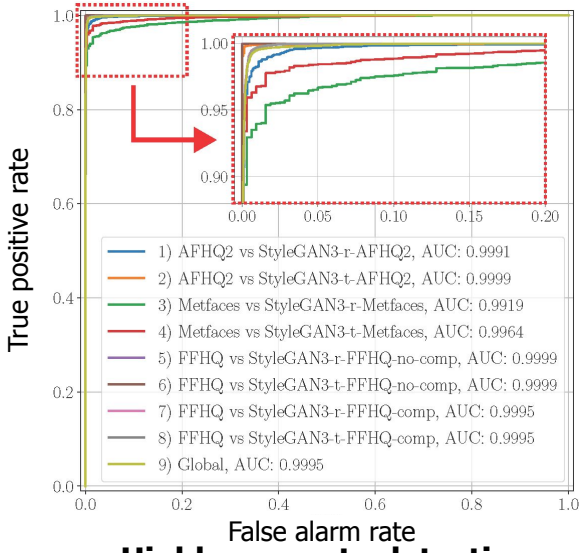


Threat from a high-resource act

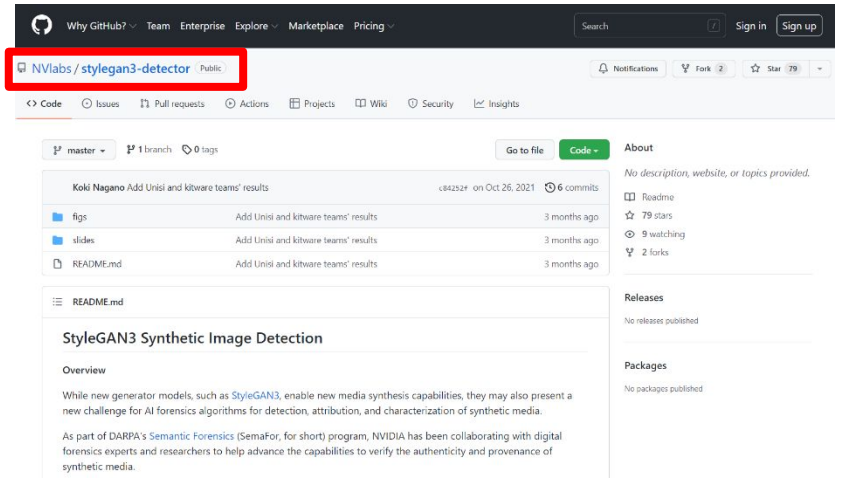


Training over semantic categories, augmentation, & many GANs

No training data from StyleGAN3!



Highly accurate detection



NVIDIA delayed releasing the GAN software & published the detectors alongside StyleGAN3

Person of interest models (attribution to an individual)

Soft biometric approaches build models of an individual's behavior from known "good" data

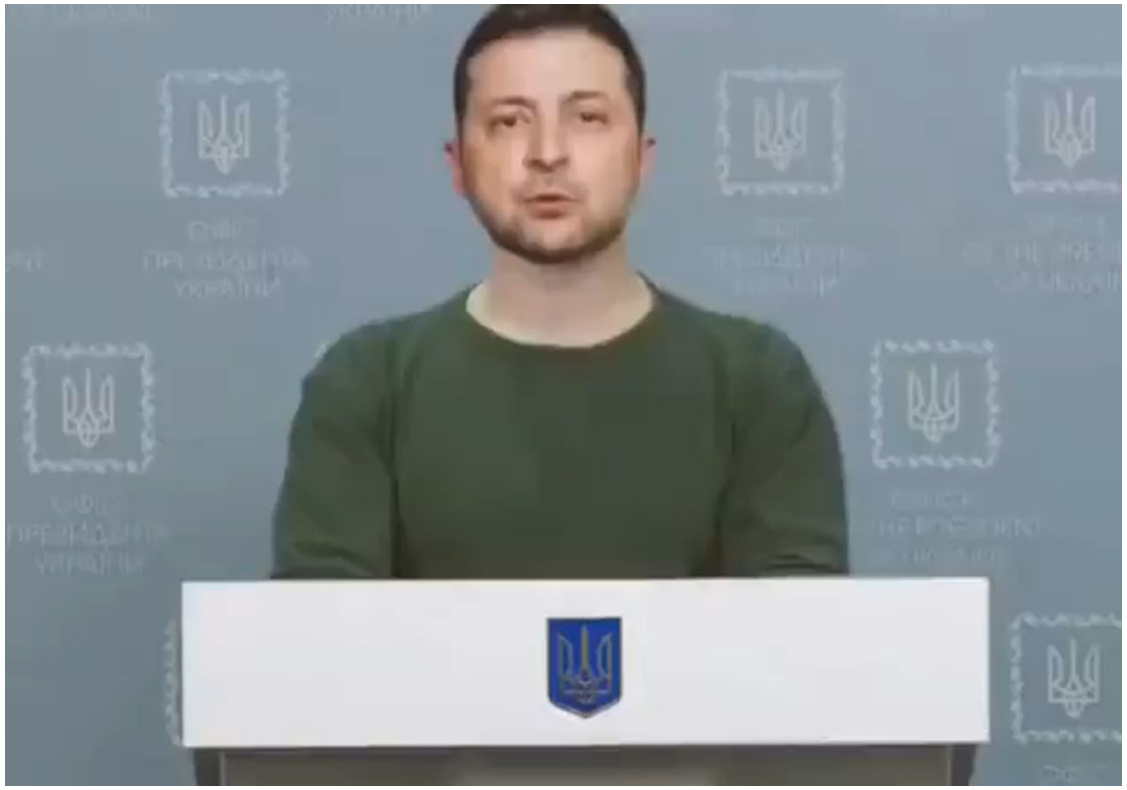
Automatically identify facial landmarks, facial action units, and head pose

Machine learning framework creates a face and head movement model for a particular individual

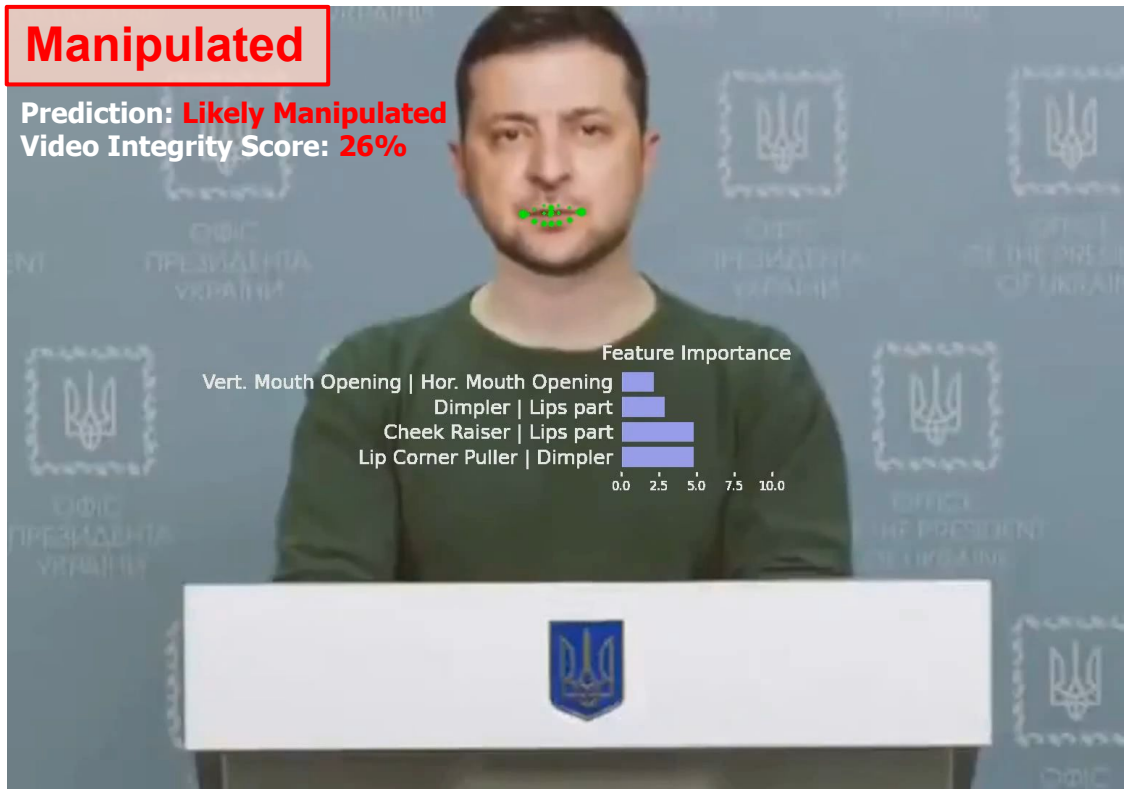
Person of interest model in action

Kitware/UC Berkeley

Original



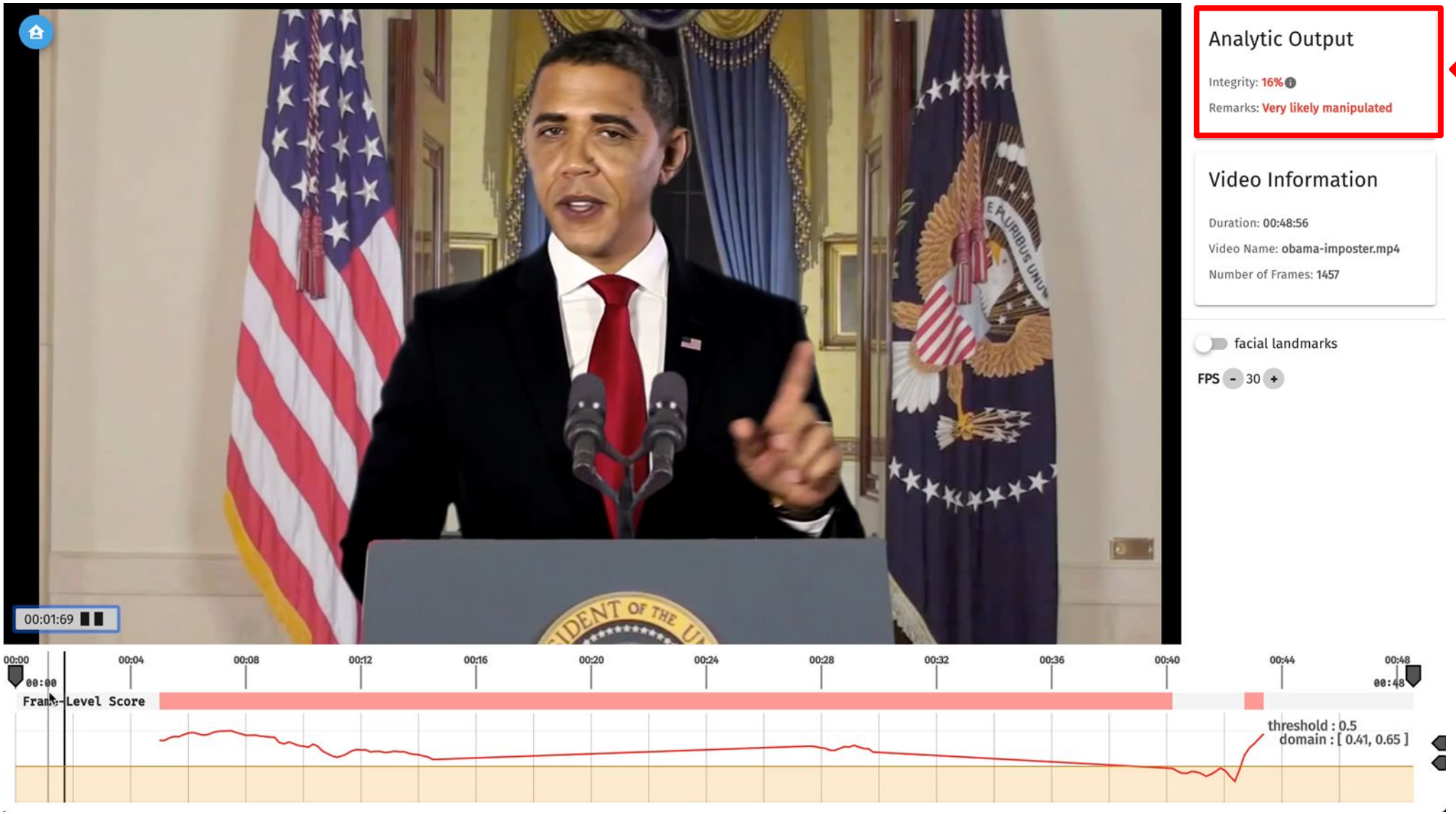
SemaFor analysis



- A video of purported to be President Zelenskyy was posted on several online forums calling for Ukrainian troops to lay down their arms March 16, 2022.
- This video is probably the first use of deepfakes in the context of war.

Person of interest application

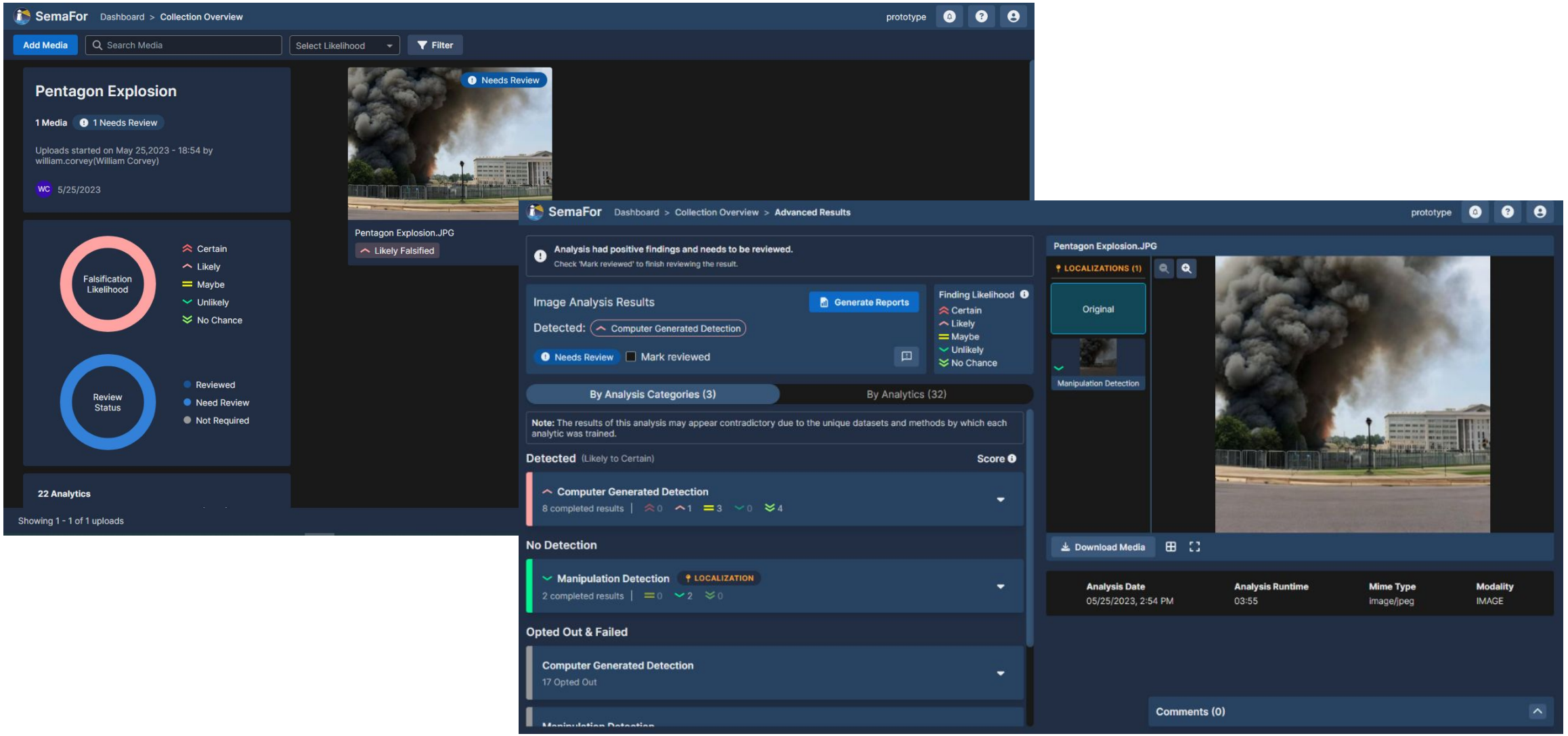
Kitware/UC Berkeley



Professional impersonator detected by soft biometric model.

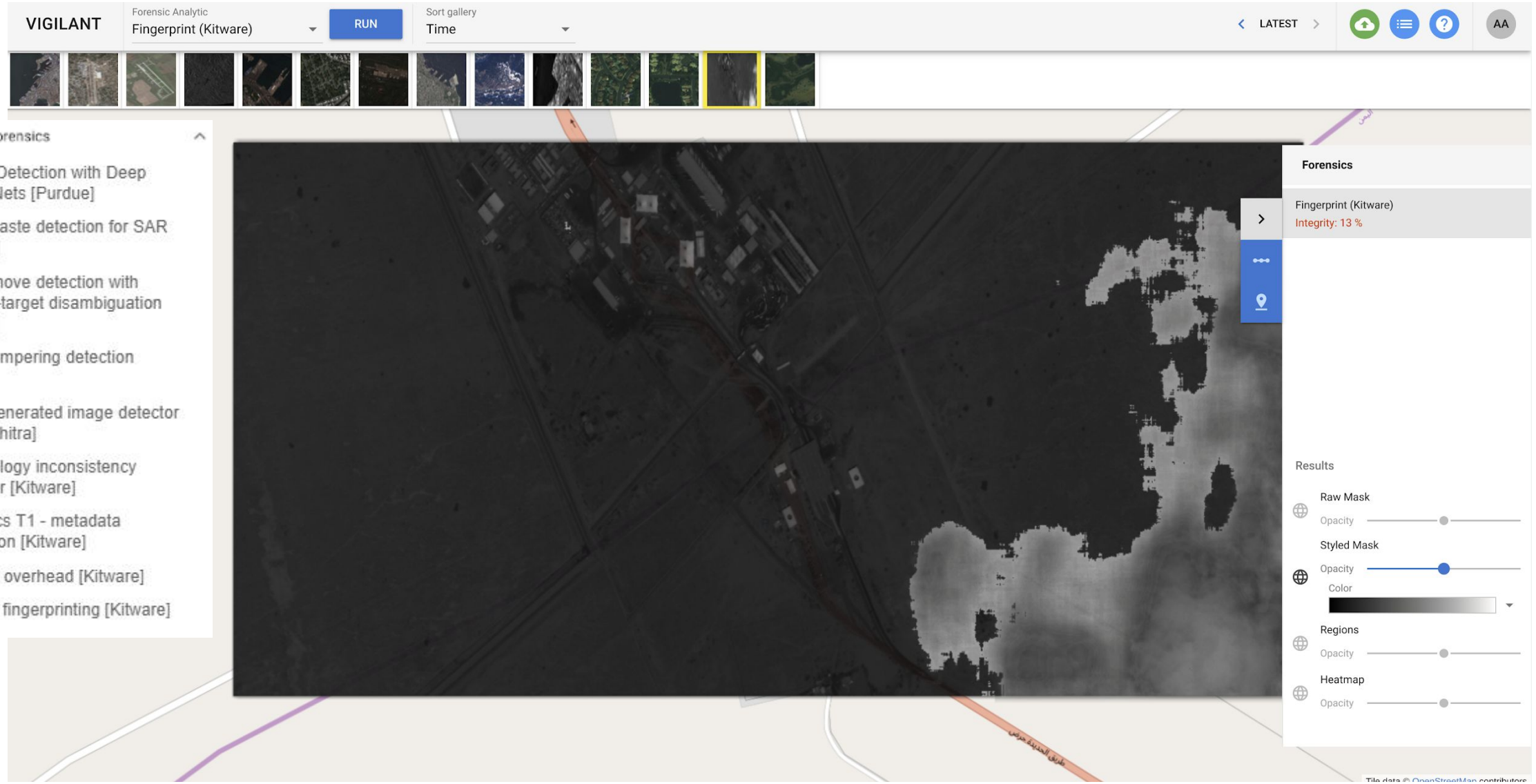
SemaFor system: detection, attribution, and characterization analytics

Lockheed Martin



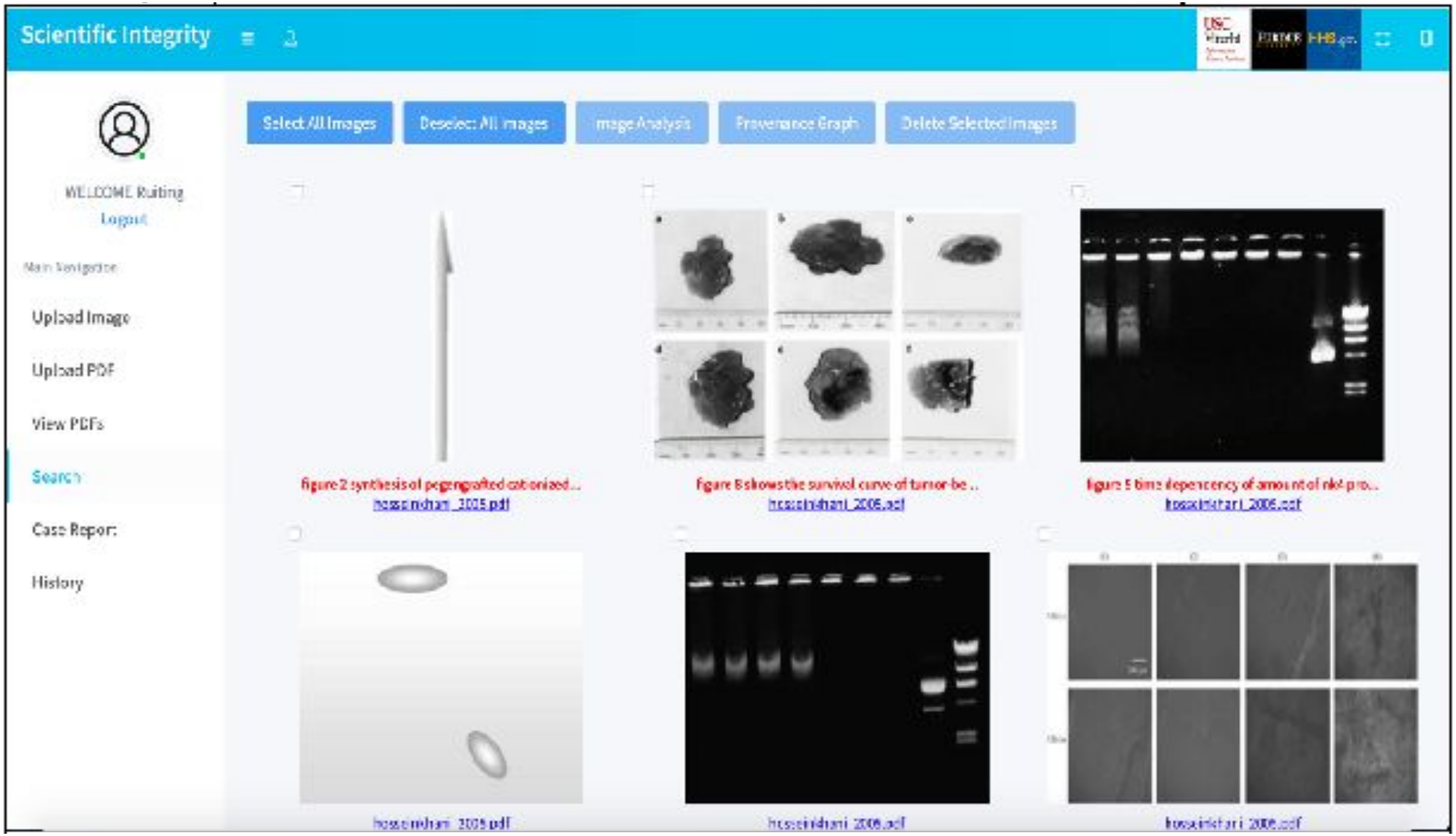
Commercial satellite application

Kitware



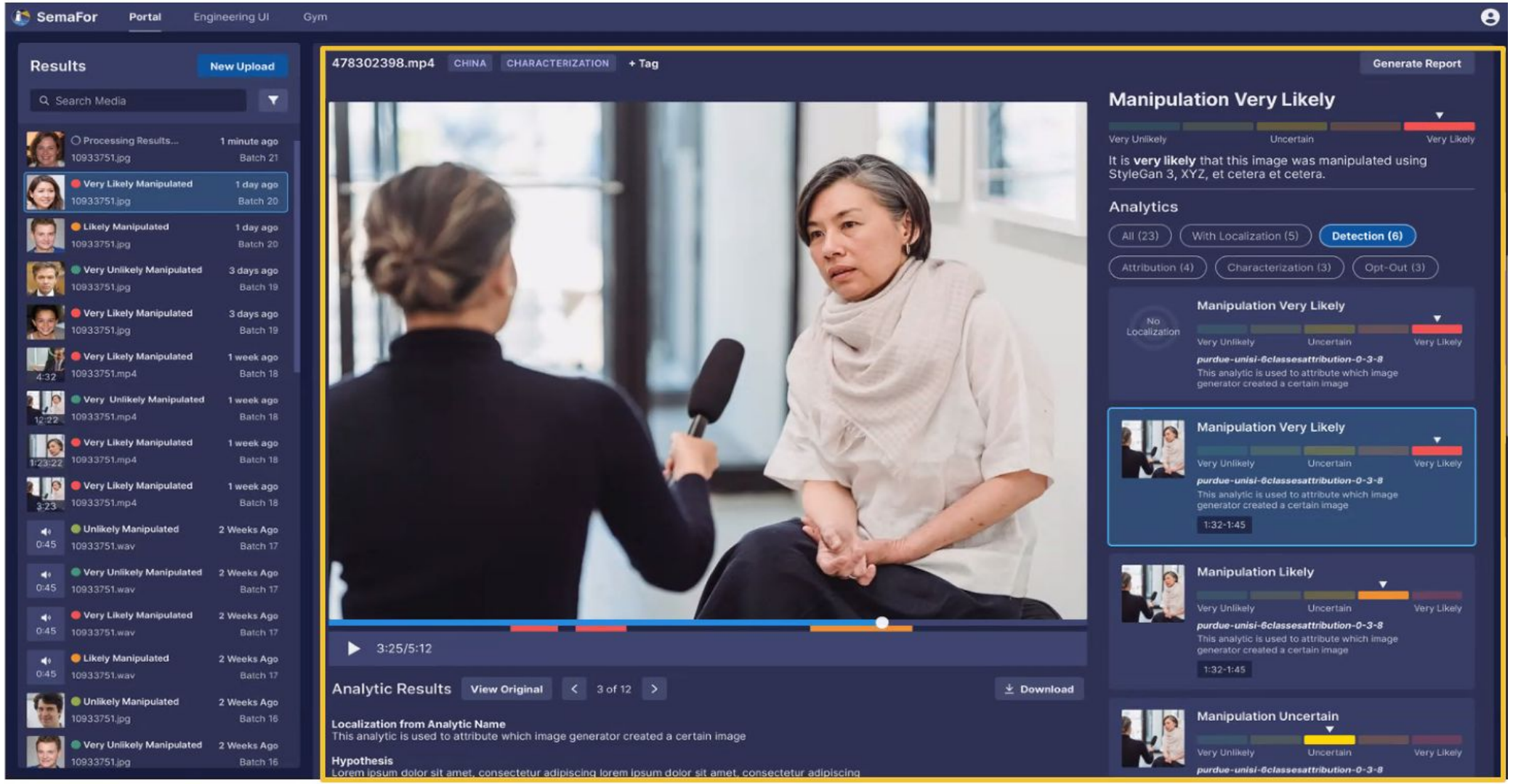
Scientific integrity application

Purdue & USC-ISI



www.darpa.mil

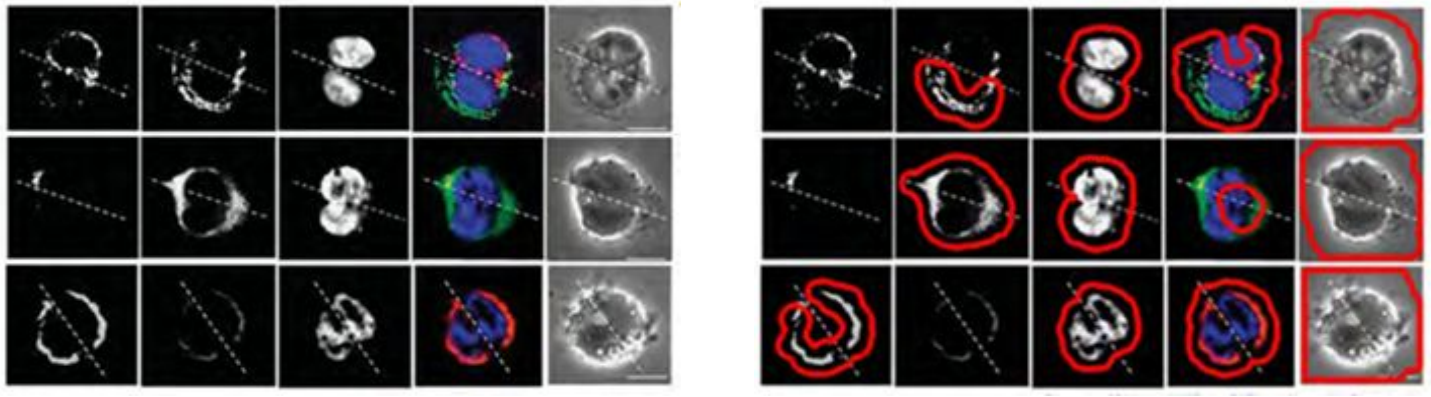
SemaFor video analytic

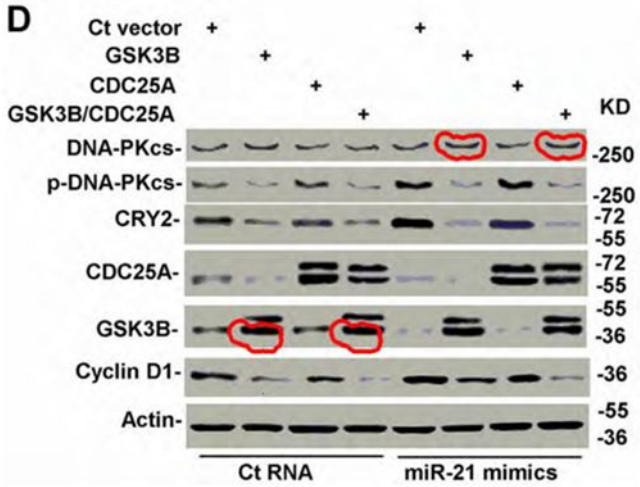


*Mock screenshot

Misinformation can undermine the scientific process

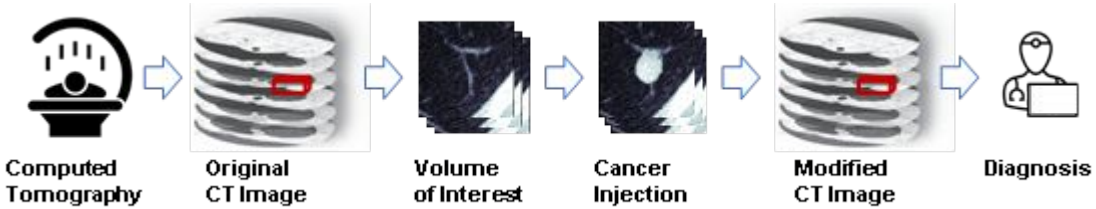
Biomedical image manipulation



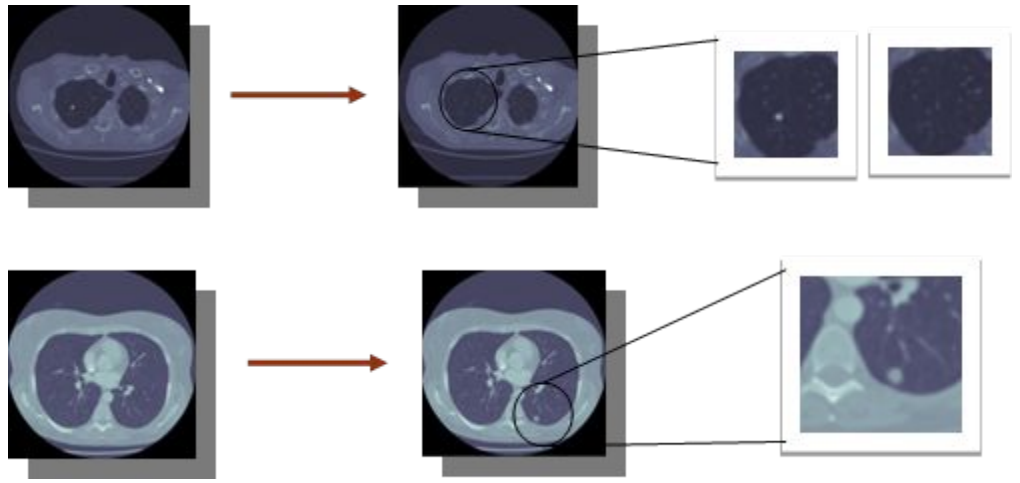


Advanced image manipulation of cancer study results, attempting to show treatment experiment successful. "Copy Move"

Radiological image manipulation



AI generated cancer inserted into 3d Computed Tomography (CT) image fools 95% of radiologist.



SemaFor technical areas & performers

TA1: Detection, attribution, characterization

Multimedia:
Text,
Audio,
Images,
Video,
Source metadata

Reasoning Ensembles

Representations

Multiple Pipelines

Semantic Detection → **Attribution** → **Characterize**

Semantic Models

Generator Models
AI generator failure modes

Entity detection & association performance

Extraction Models

Source Models
Modality & cross-modal styles, topic models

Polarization, virality, impact

Intention Models

Performers
Kitware, UIUC, Columbia, ASU, SUNY Albany, UB, UMichigan, Eduworks
SRI, UMd, UB, CMU
Purdue, Notre Dame, Univ Siena, Univ Frederico II of Naples, Politecnico di Milano
UC Berkeley, Boston, University of Washington, UC Davis, Pinscreen

TA2: System integration

Scores + Evidence

Model context

Updates & curation

Explanation & Integration

Score Fusion → **Explanation Generation**

Prioritization

Performers
LM-ATL, Boston Fusion, CRA

TA3: Evaluation

Media generation

Evaluations

Metrics

Performers
PAR, Aptima, Rank One, Drexel, CU Denver, Syracuse

TA4: Challenge Curation

SOTA challenges

Threat modeling

Performers
STR, Graphika, Hwang
NYU, Vidrovr, IST Research, Watts
Accenture, Carley
NVIDIA
Google

Prime (11), sub (28)

Semantic Detection

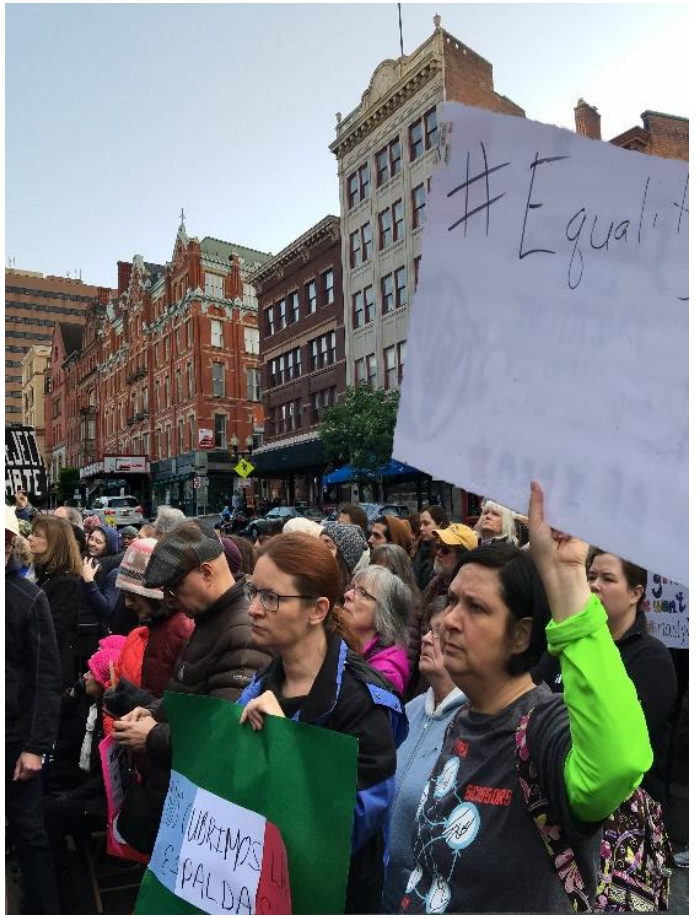
Text (Notional)

NewsWire: April 1, 2019, Bob Smith
On a rainy spring day, a vast, violent group gathered to protest recent cuts in Social Security.

Audio (Notional)

"We'd like to welcome you here on this beautiful spring day. Thank you all for coming out [cheering]…"

Image



Video

Images: DARPA MediFor

Semantic Detection

Text (Notional)

NewsWire: April 1, 2019, Bob Smith
On a rainy spring day, a vast, violent group gathered to protest recent cuts in Social Security.

Audio (Notional)

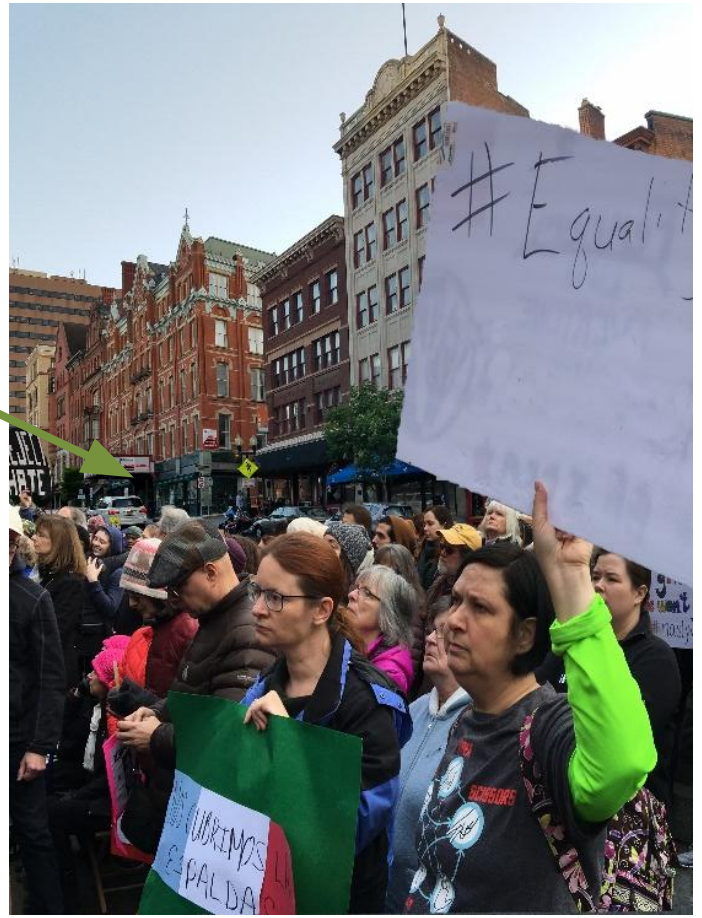
"We'd like to welcome you here on this beautiful spring day. Thank you all for coming out [cheering]…"

"protest"

Video

Conclusion: Media components consistent across modalities.

Image



Images: DARPA MediFor

Semantic Detection

Text (Notional)

NewsWire: April 1, 2019, Bob Smith

On a rainy spring day, a vast, violent group gathered to protest recent cuts in Social Security.

Audio (Notional)

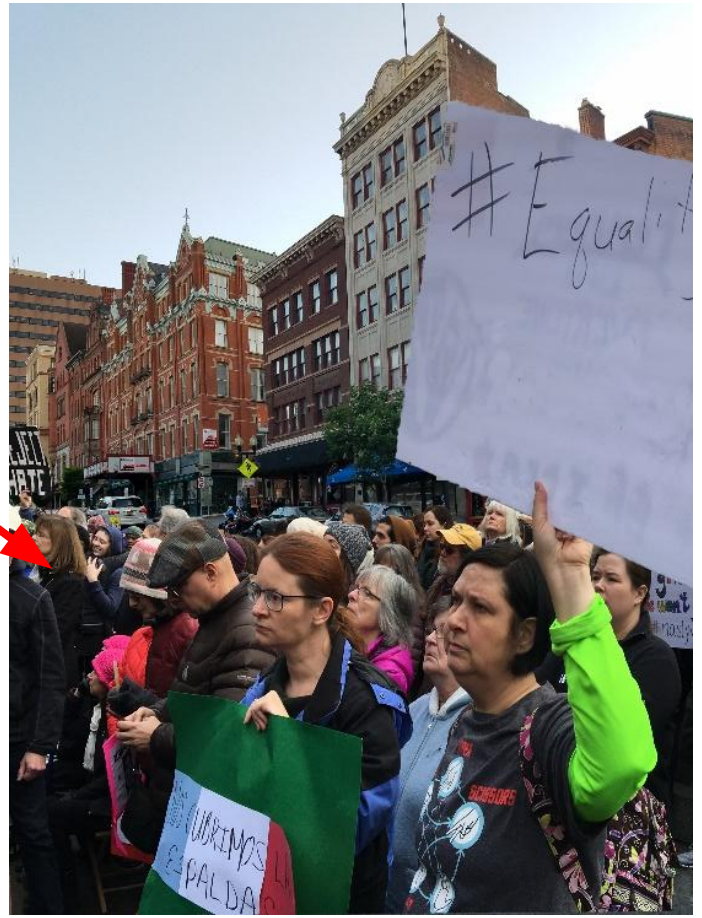
"We'd like to welcome you here on this beautiful spring day. Thank you all for coming out [cheering]…"

Video

"violent group"

Conclusion: Media components not consistent across modalities.

Image



Semantic Attribution & Characterization

Text (Notional)

NewsWire: April 1, 2019, Bob Smith

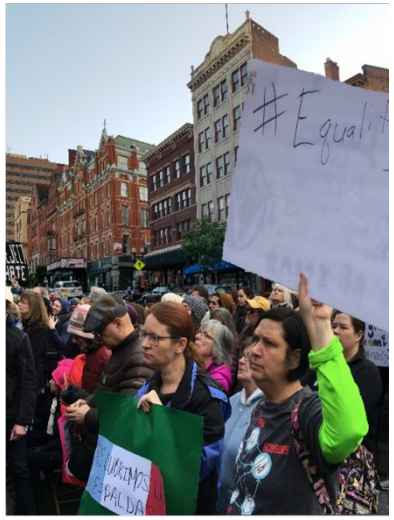
On a rainy spring day, a vast, violent group gathered to protest recent cuts in Social Security.

Video

Audio (Notional)

"We'd like to welcome you here on this beautiful spring day. Thank you all for coming out [cheering]..."

Image



Attribution: Incorrect

- Bob Smith is a tech reporter, doesn't report on social events
- Vocabulary indicates different author
- NewsWire has a different style for use of images in news article

Characterization: Malicious

- Large number of inconsistencies across media
  - Environment – "rainy spring day"
  - Behavior – "violent group"
  - Topic – "Social Security"
- Use of unsupported term "violent"
- Failed sourcing to high credibility organization ("*NewsWire*")

Images: DARPA MediFor

Summary Attribution Performance

18 of 20 Attribution Tasks met Program Objectives

Best Performing Tasks	Most Challenging Tasks
Synthetic Media Attribution for Text, Image, Audio, and Video Generators	GPT-j-6b synthetic text generation model with human manipulation

Synthetic image of military equipment generated with Guided diffusion model.

Synthetic image from StyleGAN3.

:: Title :: Cuban technicians fix thousands of SANDF vehicles ::

:: Slug :: Military Africa ::

:: By :: Sarah Lesedi ::

:: Date :: 03/18/2022 00:00 ::

Cuban technicians fix thousands of SANDF vehicles Published duration 19 July 2018

image copyright AFP image caption Rifles and other hardware are transported during their inspection.

Thousands of pieces of military hardware seized by the South African government are being examined in Havana by technicians from the Republic of Cuba's military-industrial complex.

South African authorities say thousands of pieces of military equipment are believed to have been looted during the former President's administration.

The items, seized in 2010, are now being tested to determine the value of them.

Tens of millions of dollars could change hands if any of the equipment is deemed valuable.

However, the value of any of the thousands of pieces will likely be less than the cost of transporting the items back to South Africa.

That is because the state-run Cubadeb utility company is the only company qualified and willing to transport such a large cache of goods.

The military-industrial complex is responsible for the manufacturing, repair, overhaul, and the maintenance of the equipment.

It also carries out the transport of the equipment from one end of the Latin American country to another.

The company has offices in many cities and is in charge of servicing military vehicles in all of them.

Synthetic text article about military capabilities from GPT-j-6b.

Summary Characterization Performance

8 of 8 Characterization Tasks met Program Objectives

Best Performing Tasks	Most Challenging Tasks
Characterization of *intents* and *propaganda* tactics (hate speech, bandwagoning, dictat, scapegoating)	*Appeal to Fear* and *Minimization* tactics

Utilizing the **appeal to fear tactic** the image is manipulated to exacerbate fear about COVID in a technical information news article.

Supporting the **bandwagoning** tactic, the image is manipulated to insert soldiers to convey and support the tactic in a social media post.