Semantic Forensics (SemaFor)

William Corvey, PhD
DARPA I2O

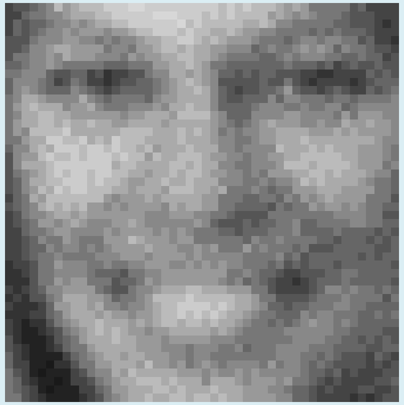
Deepfakes Day 2022

8/30/2022

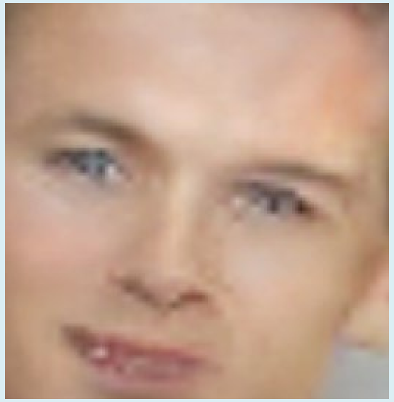
Distribution A: Approved for public release: distribution unlimited.

Image & video manipulation technology evolution

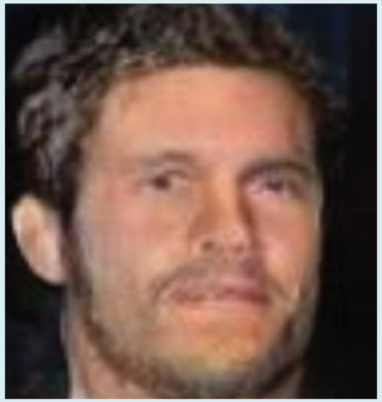
Image Synthesis



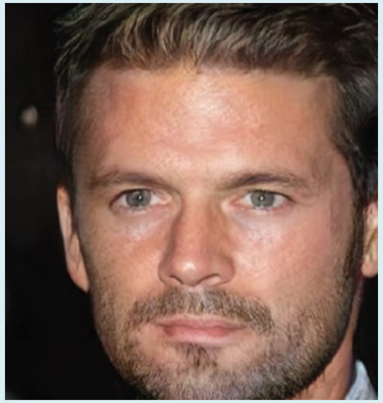
2014 Goodfellow et al.



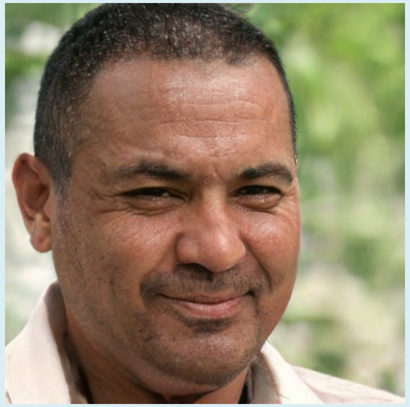
2015 Radford et al.



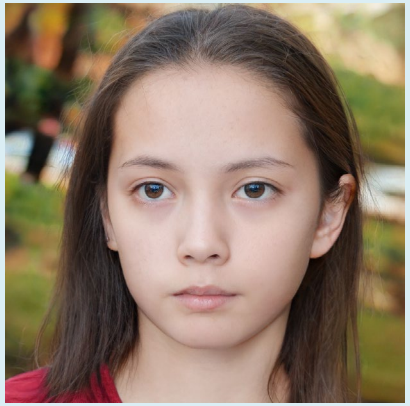
2016 Liu and Tuzel



2017 Karras et al.

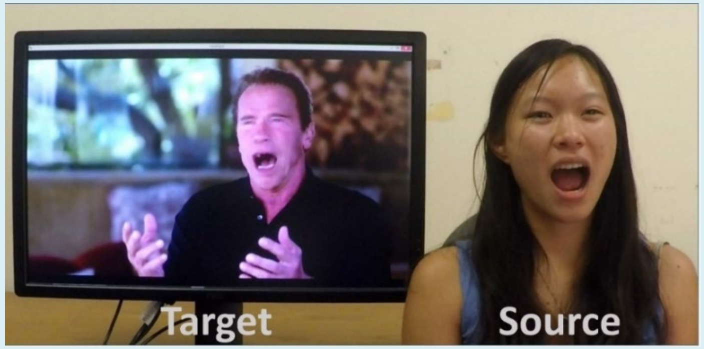


2018 Karras et al.



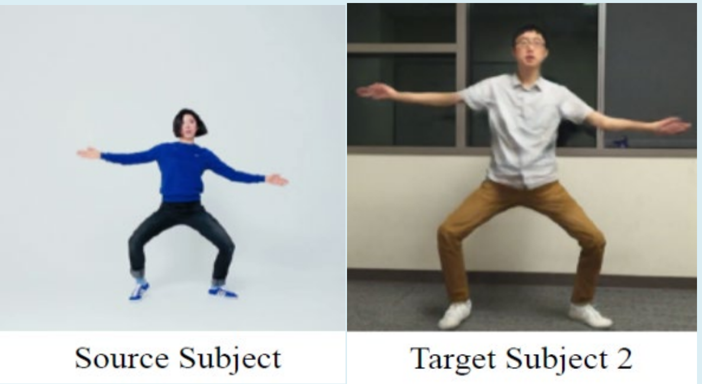
2019 Karras et al.

Video Manipulation



2016 Thies et al.

2017 Deepfakes



2018 Chan et al.

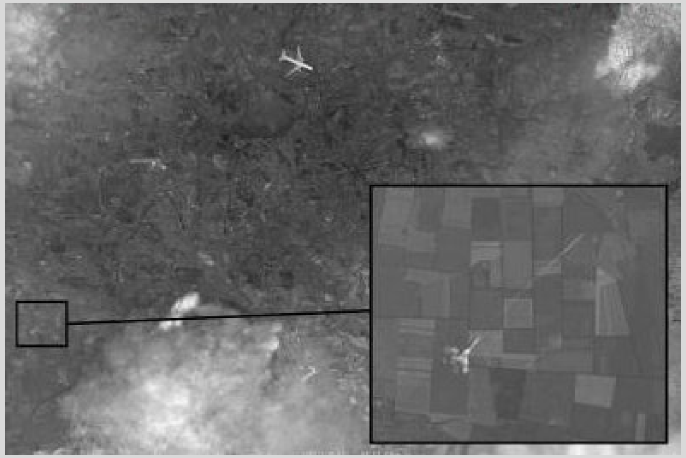
..stock price closed at one hundred ~~ninety one point four~~ eighty two point two five dollars

2019 Fried et al.

Image & video manipulations we've already seen

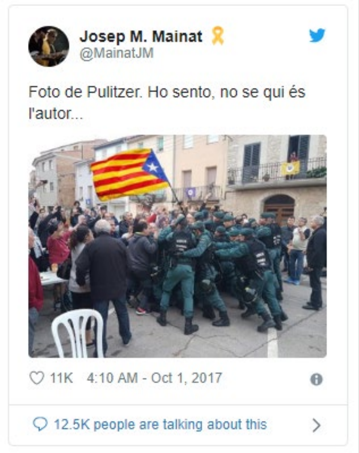
Manual manipulation

MH17 Downing (2014)



Photoshop Source: Bellingcat

Catalan Independence (2017)



Josep M. Mainat
@MainatJM

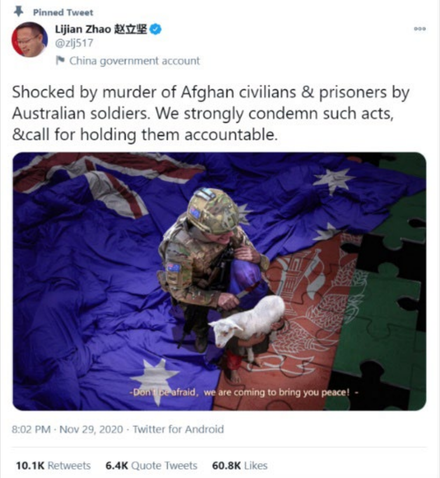
Foto de Pulitzer. Ho sento, no se qui és l'autor...

♡ 11K 4:10 AM - Oct 1, 2017

12.5K people are talking about this

Photoshop Source: El Pais

Chinese government social media (2020)



Pinned Tweet
Lijian Zhao 赵立坚
@zlj517
China government account

Shocked by murder of Afghan civilians & prisoners by Australian soldiers. We strongly condemn such acts, &call for holding them accountable.

-Don't be afraid, we are coming to bring you peace! -

8:02 PM · Nov 29, 2020 · Twitter for Android

10.1K Retweets 6.4K Quote Tweets 60.8K Likes

https://twitter.com/zlj517/status/1333214766806888448

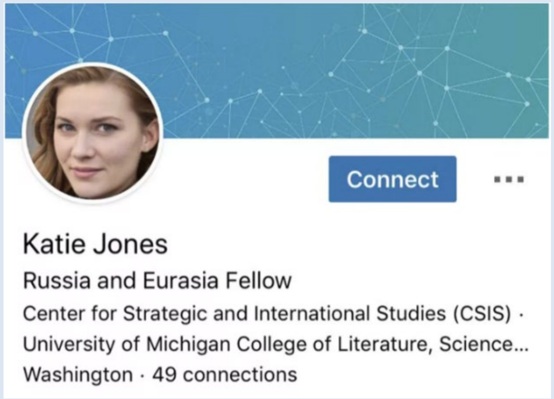
Photoshop Source: Twitter

Automated manipulation

Jordan Peele's Obama (2018)

Deepfake Source: YouTube

Katie Jones – LinkedIn (2019)



Connect ···

Katie Jones
Russia and Eurasia Fellow
Center for Strategic and International Studies (CSIS) · University of Michigan College of Literature, Science...
Washington · 49 connections

GAN Source: CNET

Nixon Moon Speech (2020)

Multiple Source: moondisaster.org

Pro-Chinese Inauthentic Network (2020)

GAN Source: Graphika

Future threats from falsified multi-modal media

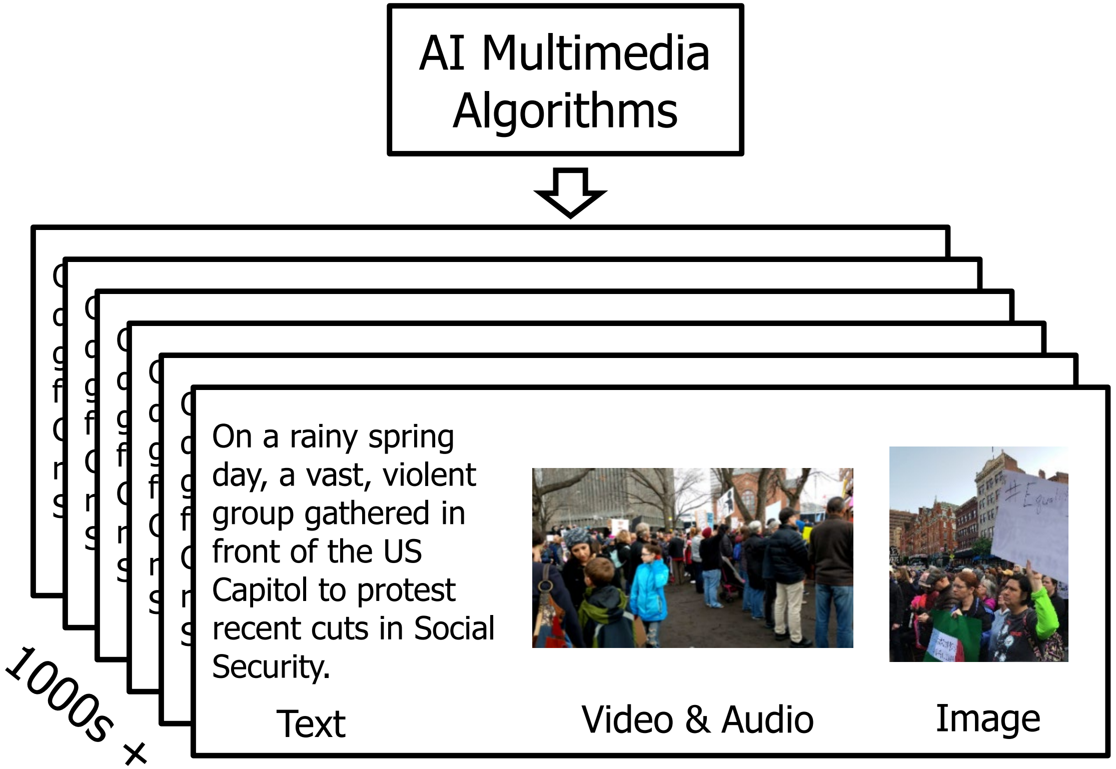
Targeted Personal Attacks

Peele 2017

AI Multimedia Algorithms

Highly realistic video

Generated Events at Scale



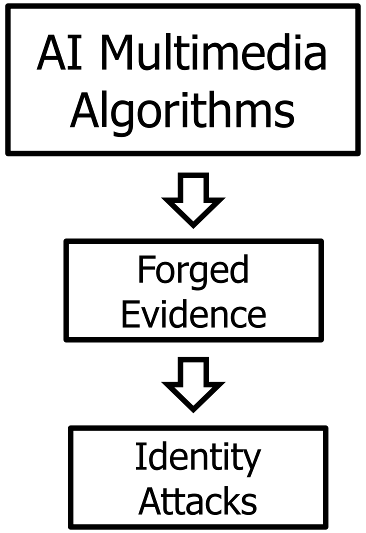
On a rainy spring day, a vast, violent group gathered in front of the US Capitol to protest recent cuts in Social Security.

Text

Believable fake events

Ransomfake concept: Identity Attacks as a service (IAaaS)

Bricman 2019



Examples of possible fakes:

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

Undermines key individuals and organizations

Distribution A: Approved for public release: distribution unlimited.

Objective

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

Synthetic Media Detection, Attribution, and Characterization Capabilities

	Desired Capability	Today	SemaFor
Detection	Automatically detect semantic generation/manipulation errors	Limited	Yes
	Detect manipulations across multiple modalities and assets	Limited	Yes
	Robust to many manipulation algorithms	Fragile	Highly robust
	Increased adversary effort needed to fool detection algorithms	Some	Significant
Attribution	Automatically confirm source or author	Limited	Yes
	Automatically identify unique source fingerprints	No	Yes
	Explain authorship inconsistencies	No	Yes
Characterization	Automatically characterize manipulation intent or impact	No	Yes
	Provide evidence and explanation for manipulation intent	No	Yes
	Correctly prioritize generated/manipulated media for review	No	Yes

Semantic Detection

Text (Notional)

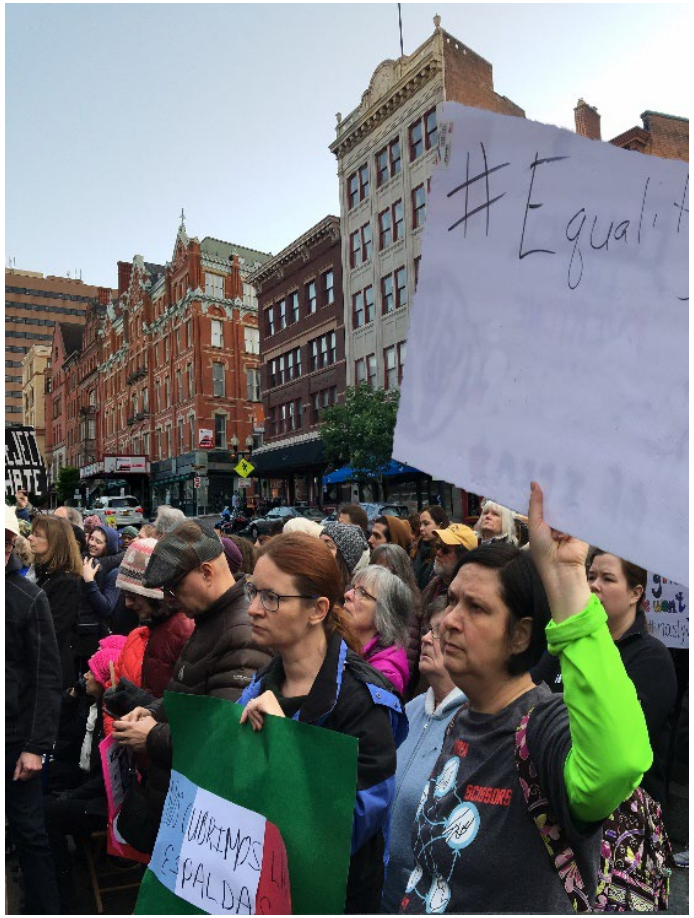
NewsWire: April 1, 2019, Bob Smith
On a rainy spring day, a vast, violent group gathered in front of the US Capitol to protest recent cuts in Social Security.

Audio (Notional)

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

Video

Image



Semantic Detection

Text (Notional)

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

Audio (Notional)

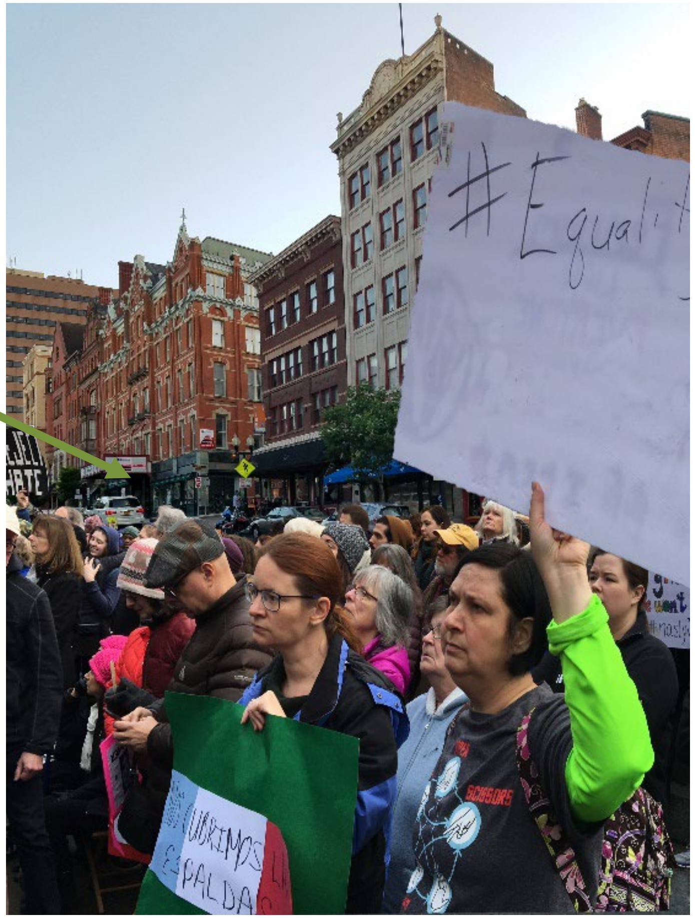
"We'd like to welcome you here on this beautiful spring day. Thank you all for coming out [cheering]…"

Conclusion: Media components consistent across modalities.

"protest"

Video

Image



Semantic Detection

Text (Notional)

NewsWire: April 1, 2019, Bob Smith

On a rainy spring day, a vast, violent group gathered in front of the US Capitol to protest recent cuts in Social Security.

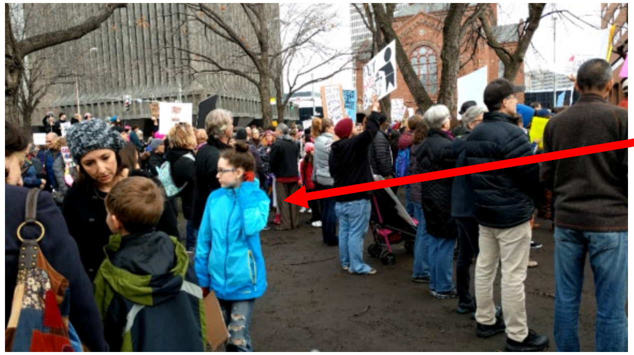
Audio (Notional)

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

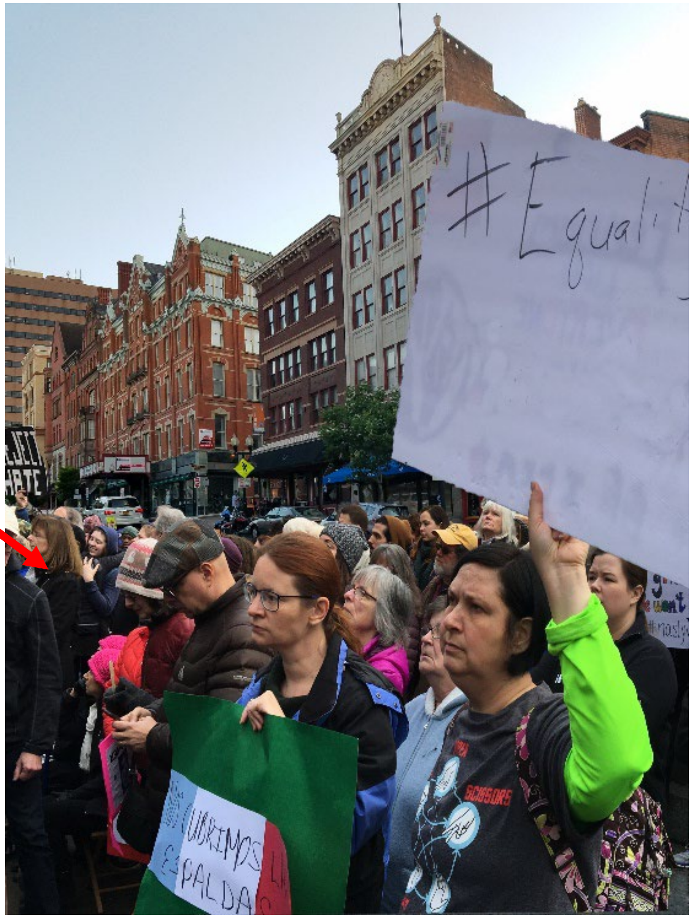
Conclusion: Media components not consistent across modalities.

"violent group"

Video



Image



Semantic Attribution & Characterization

Text (Notional)

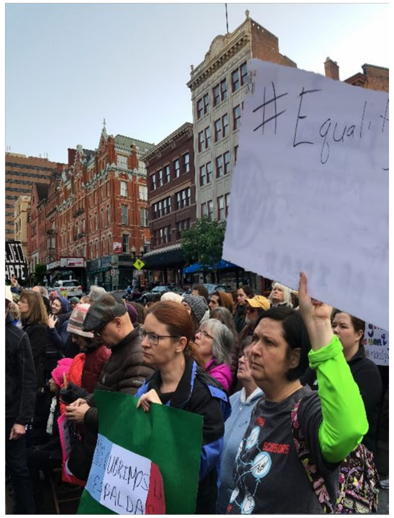
NewsWire: April 1, 2019, Bob Smith
On a rainy spring day, a vast, violent group gathered in front of the US Capitol to protest recent cuts in Social Security.

Video

Audio (Notional)

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

Image



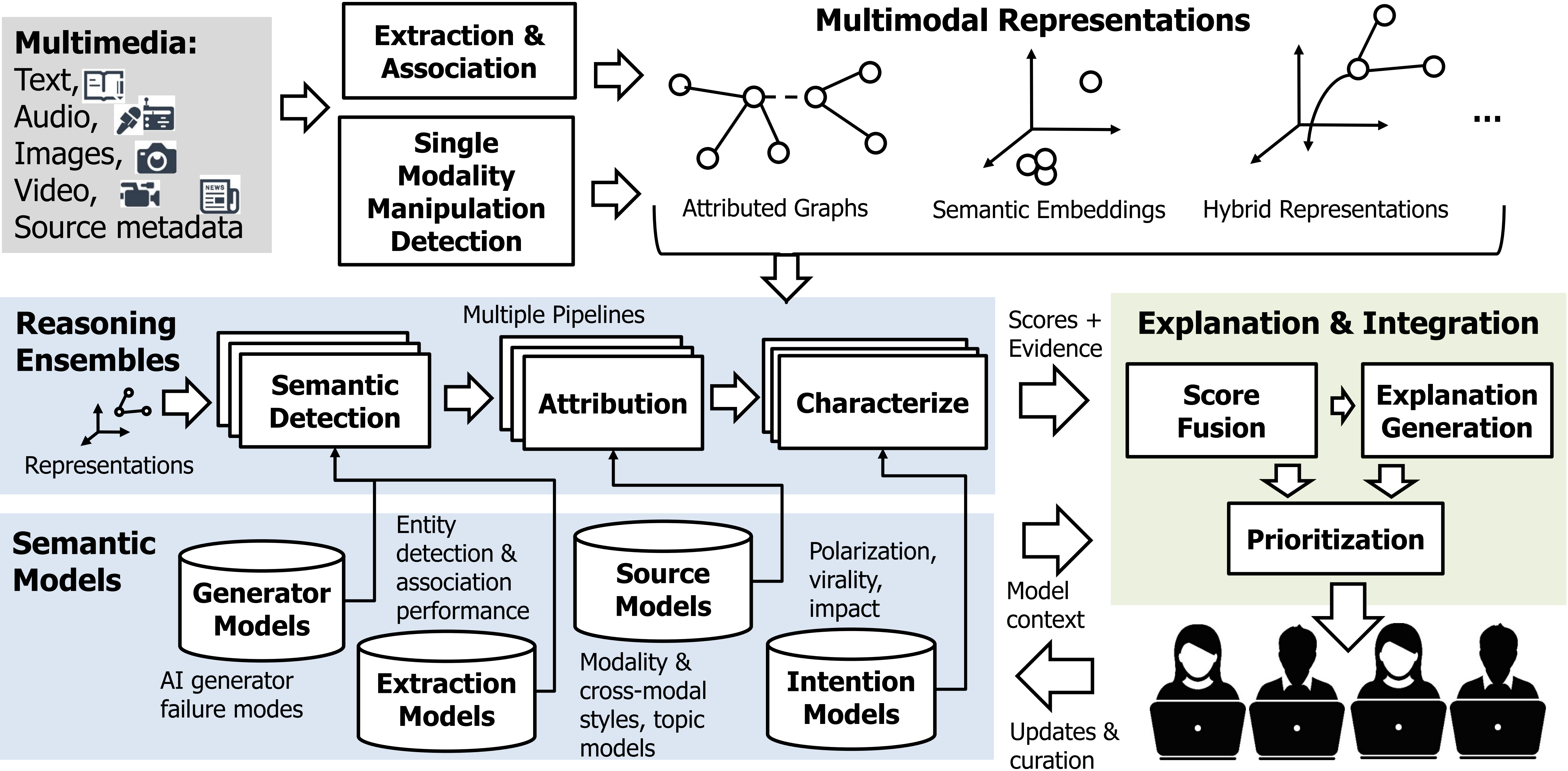
<u>Attribution: Incorrect</u>

- 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

<u>Characterization: Malicious</u>

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

SemaFor System



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

Extraction & Association

Single Modality Manipulation Detection

Multimodal Representations

Attributed Graphs

Semantic Embeddings

Hybrid Representations

...

Reasoning Ensembles

Representations

Multiple Pipelines

Semantic Detection

Attribution

Characterize

Scores + Evidence

Explanation & Integration

Score Fusion

Explanation Generation

Prioritization

Model context

Updates & curation

Semantic Models

Generator Models

AI generator failure modes

Extraction Models

Entity detection & association performance

Source Models

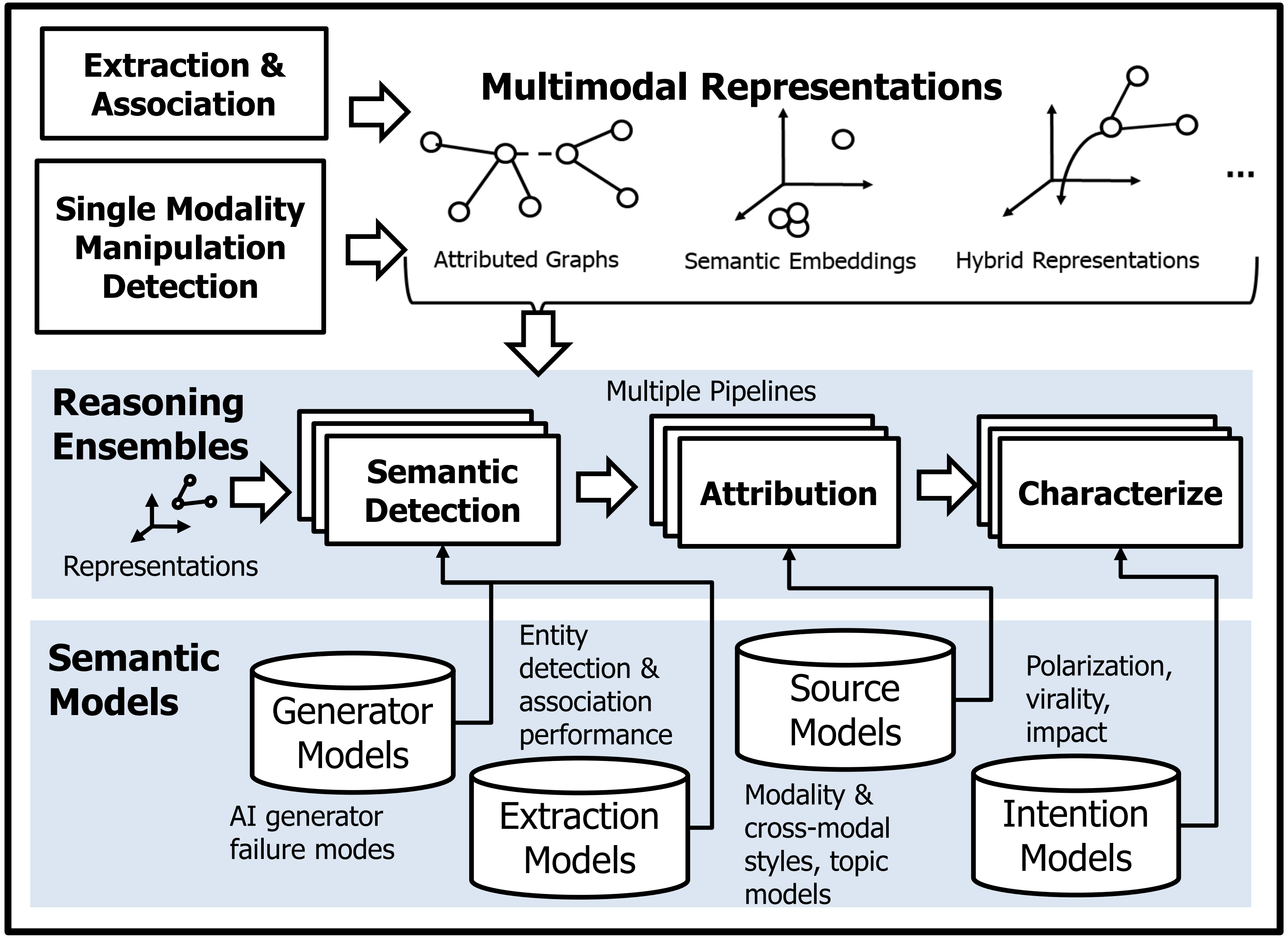
Modality & cross-modal styles, topic models

Intention Models

Polarization, virality, impact

Technical Areas

TA1: Detection, Attribution, Characterization



Extraction & Association

Single Modality Manipulation Detection

Multimodal Representations

Attributed Graphs

Semantic Embeddings

Hybrid Representations

...

Reasoning Ensembles

Representations

Multiple Pipelines

Semantic Detection

Attribution

Characterize

Semantic Models

Generator Models

AI generator failure modes

Entity detection & association performance

Extraction Models

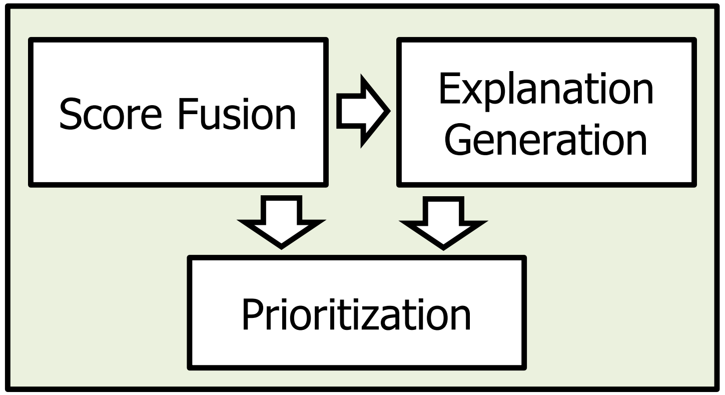
Source Models

Modality & cross-modal styles, topic models

Polarization, virality, impact

Intention Models

TA2: Explanation & Integration

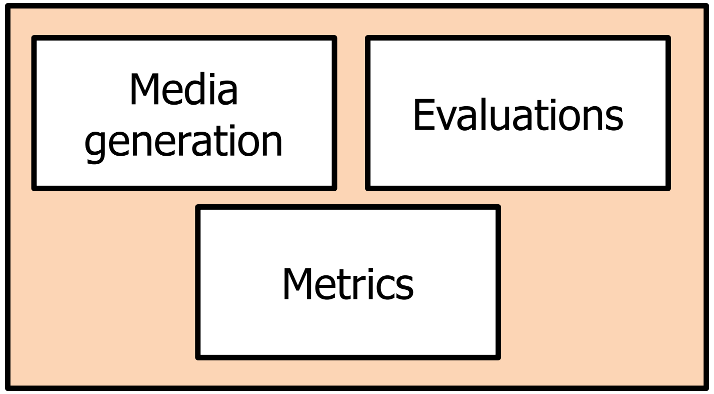


Score Fusion

Explanation Generation

Prioritization

TA3: Evaluation

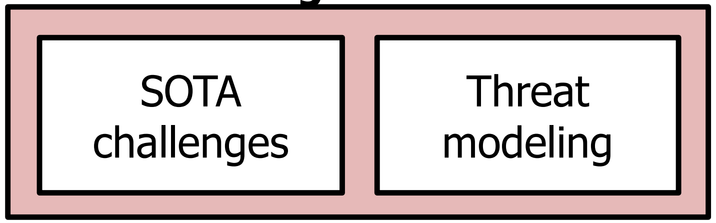


Media generation

Evaluations

Metrics

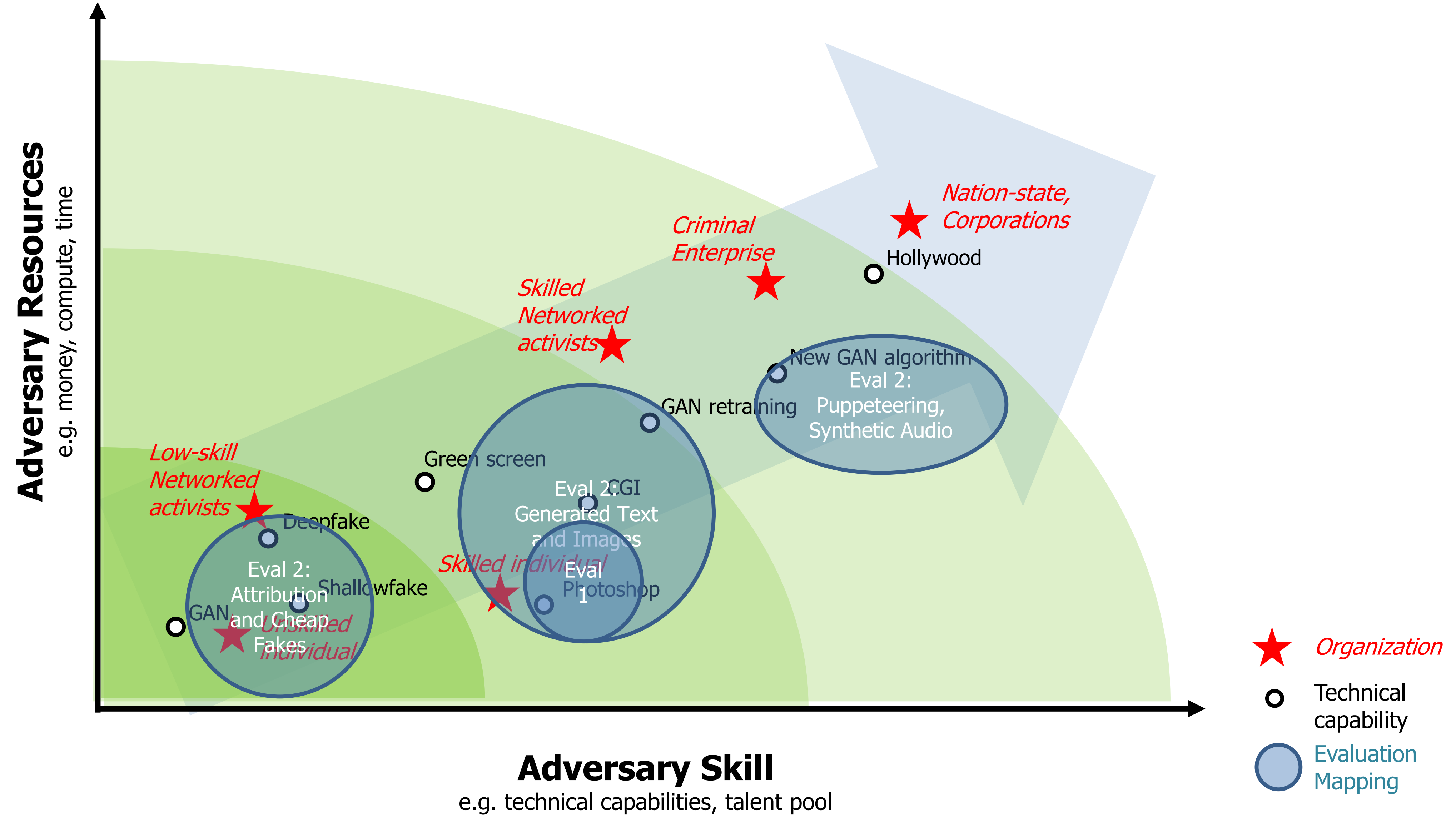
TA4: Challenge Curation



SOTA challenges

Threat modeling

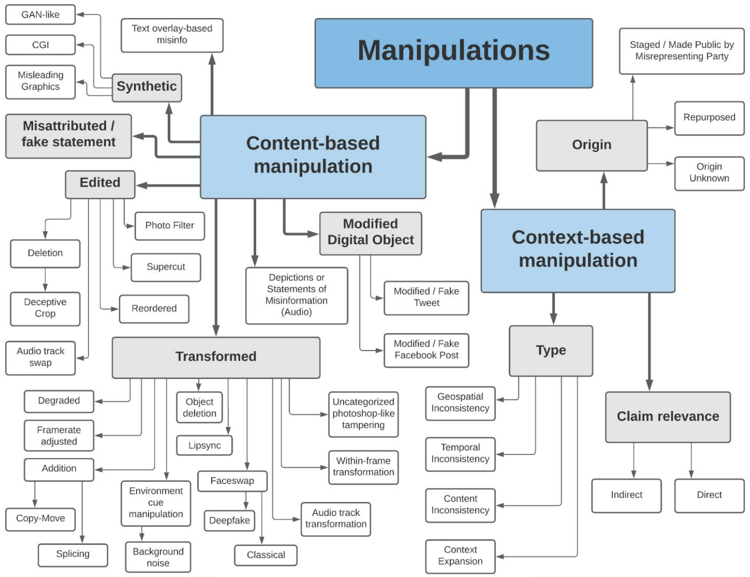
SemaFor evaluation mapped to notional adversarial landscape



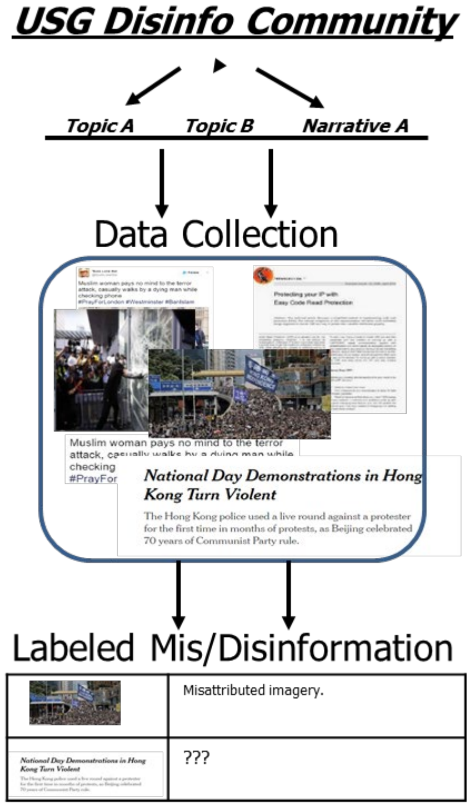
Adversary Resources
e.g. money, compute, time
Nation-state, Corporations
Criminal Enterprise
Hollywood
Skilled Networked activists
New GAN algorithm
Eval 2: Puppeteering, Synthetic Audio
GAN retraining
Green screen
Eval 2: Generated Text and Images
CGI
Eval 1
Photoshop
Skilled individual
Low-skill Networked activists
Deepfake
Eval 2: Attribution and Cheap Fakes
Shallowfake
GAN
Unskilled individual
Adversary Skill
e.g. technical capabilities, talent pool
★ Organization
○ Technical capability
● Evaluation Mapping

Threat data & threat landscape

Google, Accenture FS, NYU, STR



Development of first-of-kind dataset of actual context manipulations, including human annotations of in-the-wild manipulations.



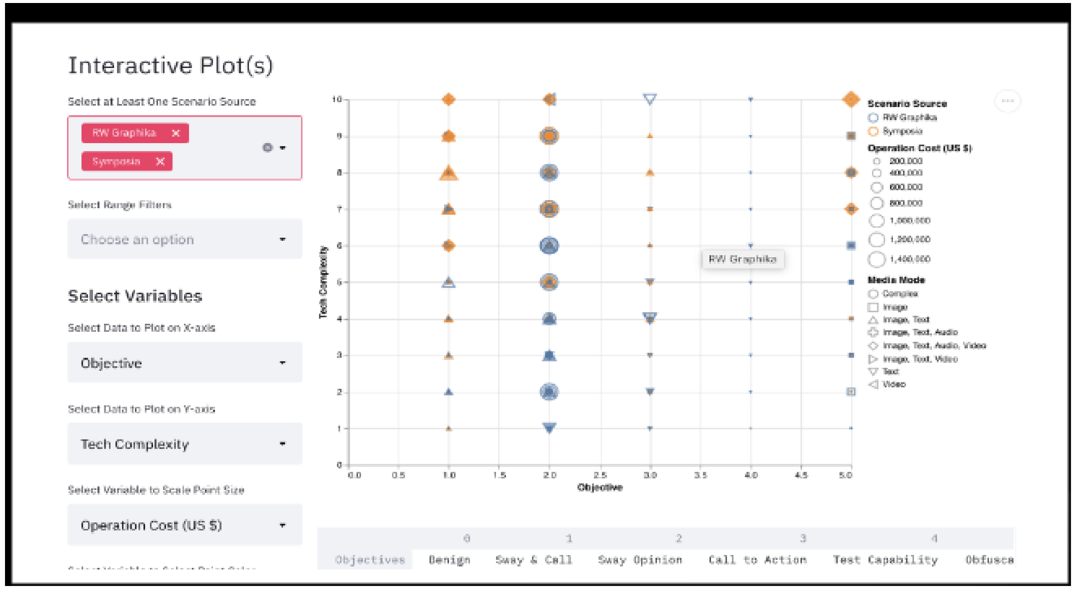
Muslim woman pays no mind to the terror attack, casually walks by a dying man while checking phone #PrayForLondon #Westminster #BanIslam

Protecting your IP with Easy Code Read Protection

National Day Demonstrations in Hong Kong Turn Violent

The Hong Kong police used a live round against a protester for the first time in months of protests, as Beijing celebrated 70 years of Communist Party rule.

	Misattributed imagery.
National Day Demonstrations in Hong Kong Turn Violent	???



Interactive Plot(s)

Select at Least One Scenario Source: RW Graphika ✕, Symposia ✕

Select Range Filters: Choose an option

Select Variables

Select Data to Plot on X-axis: Objective

Select Data to Plot on Y-axis: Tech Complexity

Select Variable to Scale Point Size: Operation Cost (US $)

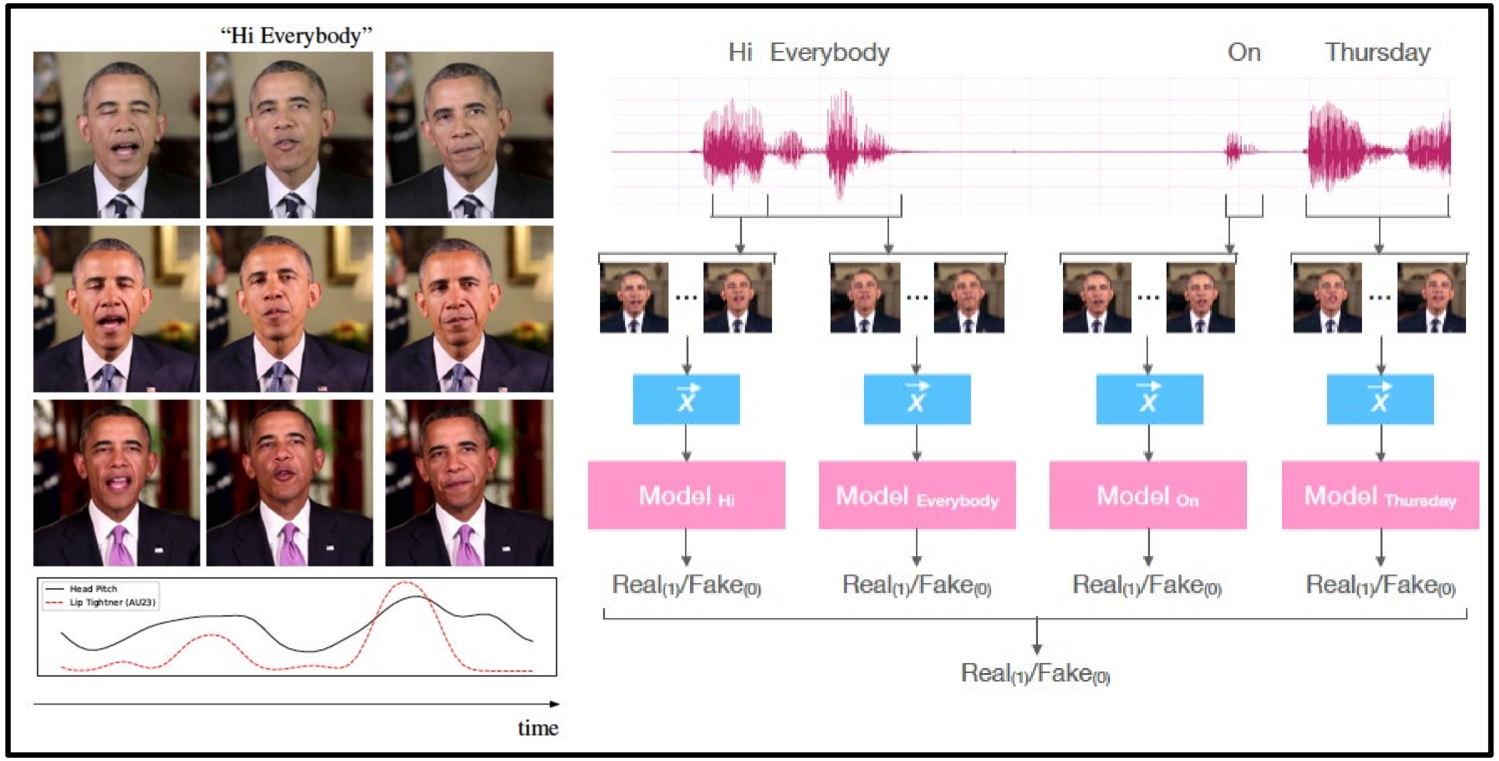
	0	1	2	3	4
Objectives	Benign	Sway & Call	Sway Opinion	Call to Action	Test Capability	Obfusca

Threat Landscape: A multi-dimensional representation to quantify and categorize the “who, what, why, where and how” of threats using manipulated media.

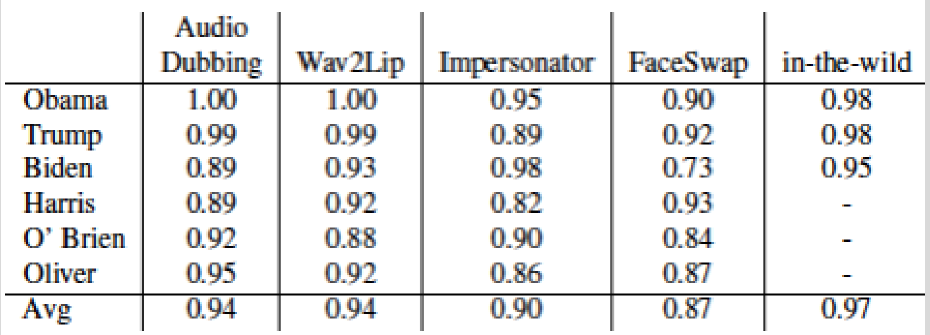
Collaborate with USG disinformation and misinformation community to identify topics and campaigns of interest.

Video falsification detection using word-conditioned facial motion

UCBerkeley
Pinscreen

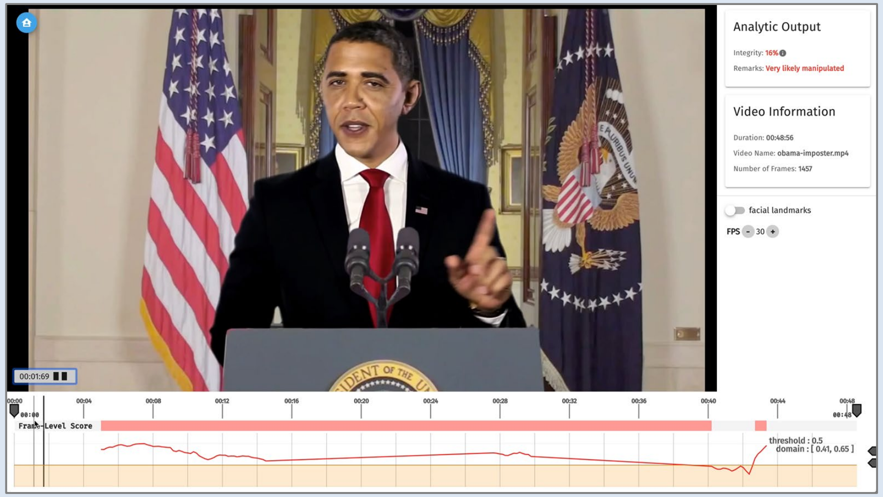


AUC on 10s video clips & 5 falsifications



	Audio Dubbing	Wav2Lip	Impersonator	FaceSwap	in-the-wild
Obama	1.00	1.00	0.95	0.90	0.98
Trump	0.99	0.99	0.89	0.92	0.98
Biden	0.89	0.93	0.98	0.73	0.95
Harris	0.89	0.92	0.82	0.93	-
O' Brien	0.92	0.88	0.90	0.84	-
Oliver	0.95	0.92	0.86	0.87	-
Avg	0.94	0.94	0.90	0.87	0.97

Future integration & deployment



Analytic Output

Integrity: 16%

Remarks: Very likely manipulated

Video Information

Duration: 00:48:56

Video Name: obama-imposter.mp4

Number of Frames: 1457

facial landmarks

FPS - 30 +

Frame-Level Score

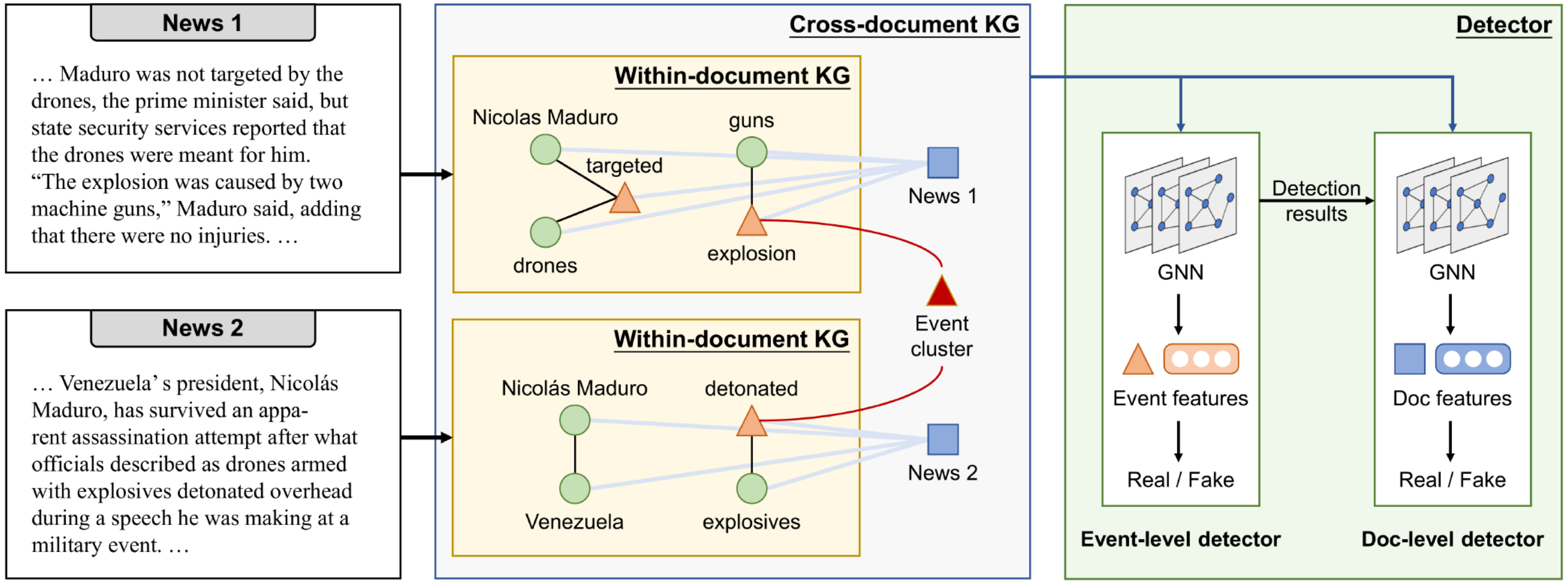
threshold : 0.5
domain : [0.41, 0.65]

- Biometric-based (person-specific) forensic approach learns a semantic model of an individual's movement & speech
- Can detect both deepfakes and cheapfakes (such as audio dubbing or impersonator)
- Generalizes to falsification methods unseen during training

Knowledge graph for cross document falsified news detection

Kitware
UIUC

Detect misinformation from a cluster of topically related documents



News 1

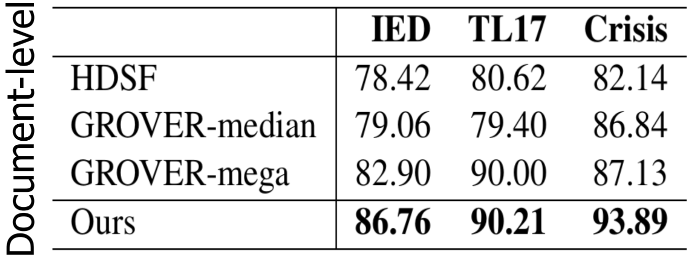
… Maduro was not targeted by the drones, the prime minister said, but state security services reported that the drones were meant for him. “The explosion was caused by two machine guns,” Maduro said, adding that there were no injuries. …

News 2

… Venezuela’ s president, Nicolás Maduro, has survived an apparent assassination attempt after what officials described as drones armed with explosives detonated overhead during a speech he was making at a military event. …

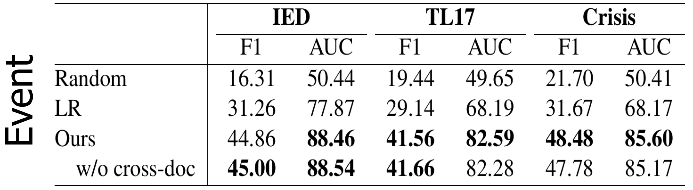
Detection Accuracy (%)

Document-level



	IED	TL17	Crisis
HDSF	78.42	80.62	82.14
GROVER-median	79.06	79.40	86.84
GROVER-mega	82.90	90.00	87.13
Ours	**86.76**	**90.21**	**93.89**

Event



	IED		TL17		Crisis	
	F1	AUC	F1	AUC	F1	AUC
Random	16.31	50.44	19.44	49.65	21.70	50.41
LR	31.26	77.87	29.14	68.19	31.67	68.17
Ours	44.86	**88.46**	**41.56**	**82.59**	**48.48**	**85.60**
w/o cross-doc	**45.00**	**88.54**	**41.66**	82.28	47.78	85.17

Real news

...She noted that at least 1% of people who catch coronavirus die of it. "Another 10-20% are hospitalized. Another 30% or more have long lasting symptoms. The vaccine is far safer, with only minor temporary side effects," Ranney said on Twitter. …In both Pfizer and Moderna's vaccine trials, no worrying side-effects were seen. …

Claim: COVID vaccine is safer than COVID

⟺ contradictory

Fake news

As many as 45,000 people may have died from the mRNA shots being given to halt COVID, according to prominent physician Dr. Peter McCullough. And teenagers — especially boys — are more at risk from being hospitalized from the vaccine than they are for COVID, he said. The culprit is myocarditis, inflammation of the heart. …

Claim: COVID vaccine is more dangerous than COVID

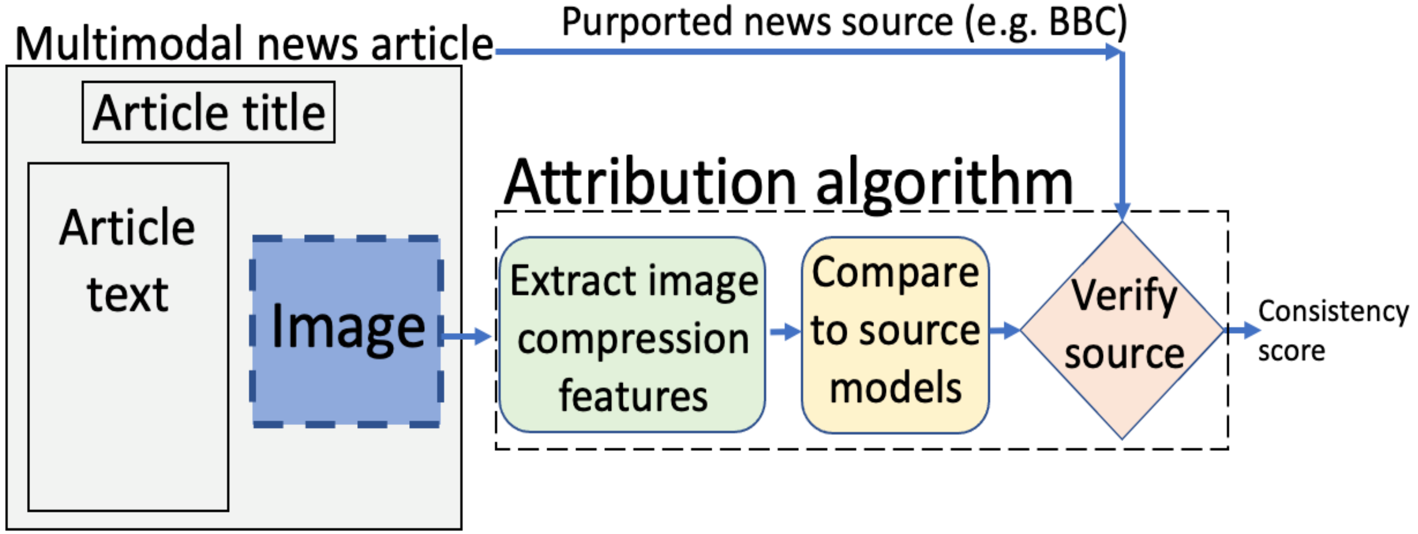
Unique single-document to cross-document event coreference resolution and reasoning

Distribution A: Approved for public release: distribution unlimited.

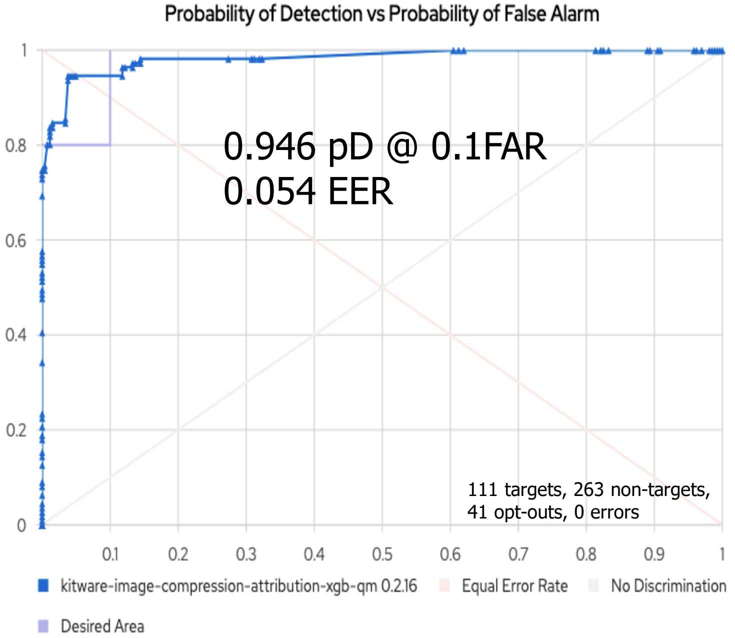
News organization attribution based on compression settings

Kitware

Technical approach



Eval-2 Attribution Verification (News Articles) Organization Swaps



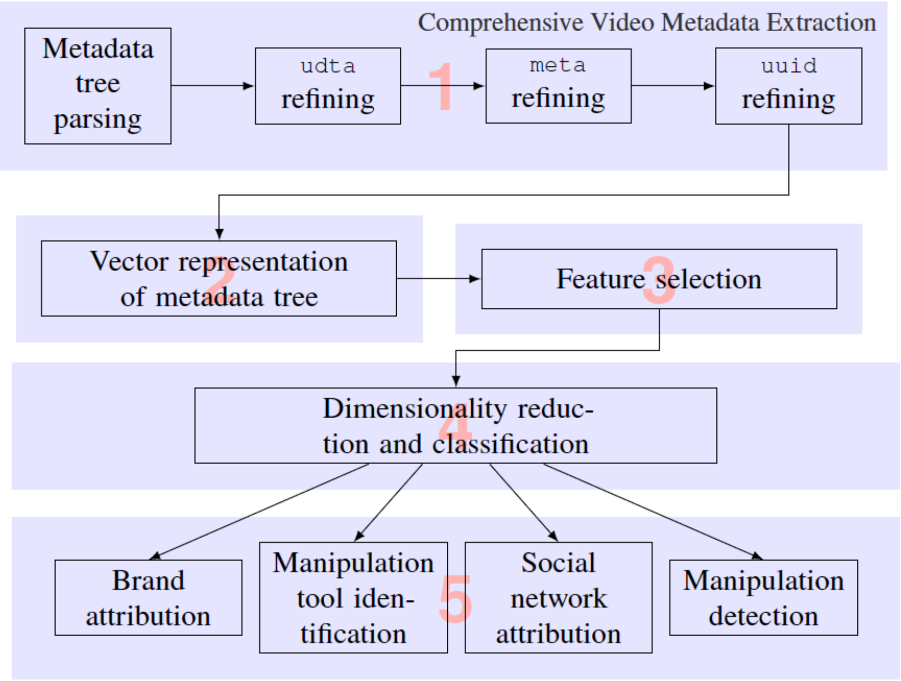
0.946 pD @ 0.1FAR
0.054 EER

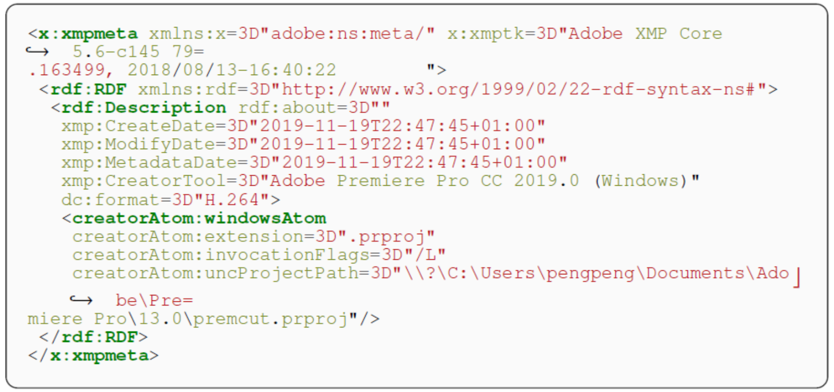
- News organizations have highly curated pipelines for media production that leave signatures in the media
- Analytic verifies whether images originate from a purported (news) source or not, by comparing each image’s compression settings to known image compression patterns of the source
- Strong performance in SemaFor Eval 2: Attribution Verification (News Articles) Organization Swaps
- Novel forensic analysis of image file compression with low computational cost

Metadata can provide highly accurate video forensics

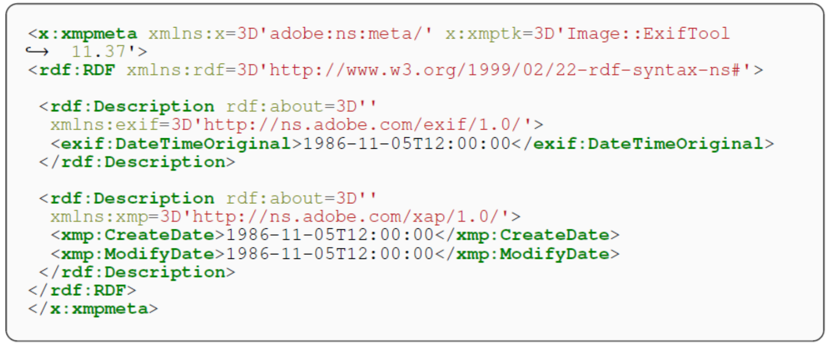
Purdue
U Siena
Notre Dame

Politecnio di Milano
U Frederico II di Napoli
University of Campina

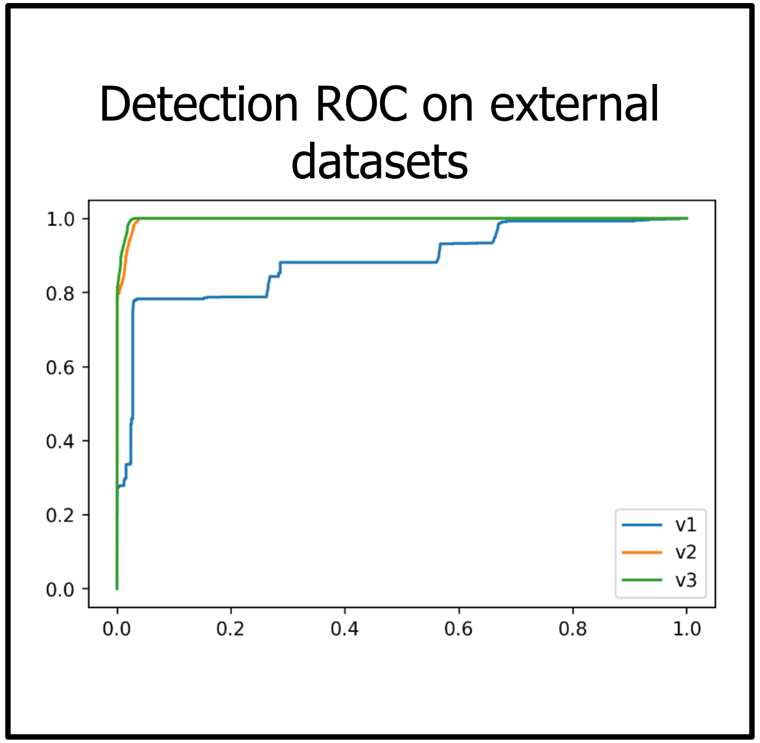

Comprehensive Video Metadata Extraction
Metadata tree parsing
udta refining
1
meta refining
uuid refining
Vector representation of metadata tree
2
Feature selection
3
Dimensionality reduction and classification
4
Brand attribution
Manipulation tool identification
5
Social network attribution
Manipulation detection



```
<x:xmpmeta xmlns:x=3D"adobe:ns:meta/" x:xmptk=3D"Adobe XMP Core
↪  5.6-c145 79=
.163499, 2018/08/13-16:40:22        ">
 <rdf:RDF xmlns:rdf=3D"http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <rdf:Description rdf:about=3D""
   xmp:CreateDate=3D"2019-11-19T22:47:45+01:00"
   xmp:ModifyDate=3D"2019-11-19T22:47:45+01:00"
   xmp:MetadataDate=3D"2019-11-19T22:47:45+01:00"
   xmp:CreatorTool=3D"Adobe Premiere Pro CC 2019.0 (Windows)"
   dc:format=3D"H.264">
   <creatorAtom:windowsAtom
    creatorAtom:extension=3D".prproj"
    creatorAtom:invocationFlags=3D"/L"
    creatorAtom:uncProjectPath=3D"\\?\C:\Users\pengpeng\Documents\Ado⌋

    ↪  be\Pre=
miere Pro\13.0\premcut.prproj"/>
 </rdf:RDF>
</x:xmpmeta>
```



```
<x:xmpmeta xmlns:x=3D'adobe:ns:meta/' x:xmptk=3D'Image::ExifTool
↪  11.37'>
<rdf:RDF xmlns:rdf=3D'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>

 <rdf:Description rdf:about=3D''
  xmlns:exif=3D'http://ns.adobe.com/exif/1.0/'>
  <exif:DateTimeOriginal>1986-11-05T12:00:00</exif:DateTimeOriginal>
 </rdf:Description>

 <rdf:Description rdf:about=3D''
  xmlns:xmp=3D'http://ns.adobe.com/xap/1.0/'>
  <xmp:CreateDate>1986-11-05T12:00:00</xmp:CreateDate>
  <xmp:ModifyDate>1986-11-05T12:00:00</xmp:ModifyDate>
 </rdf:Description>
</rdf:RDF>
</x:xmpmeta>
```



Detection ROC on external datasets

- Analytics use metadata in the video files to determine if a video is manipulated
- Difficult to hide manipulation traces in metadata
- Method is significantly faster than pixel-based manipulation detection techniques

GAN detector defends against high-resource actors

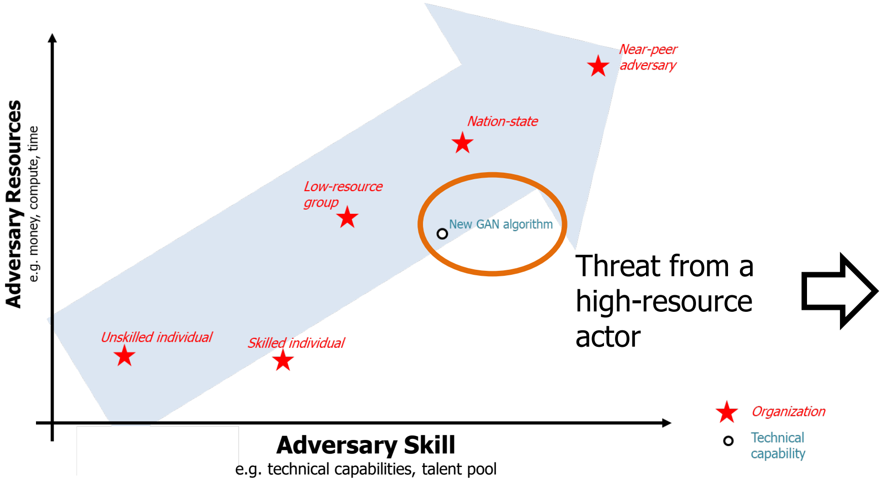
Purdue, Kitware, SRI, NVIDIA

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

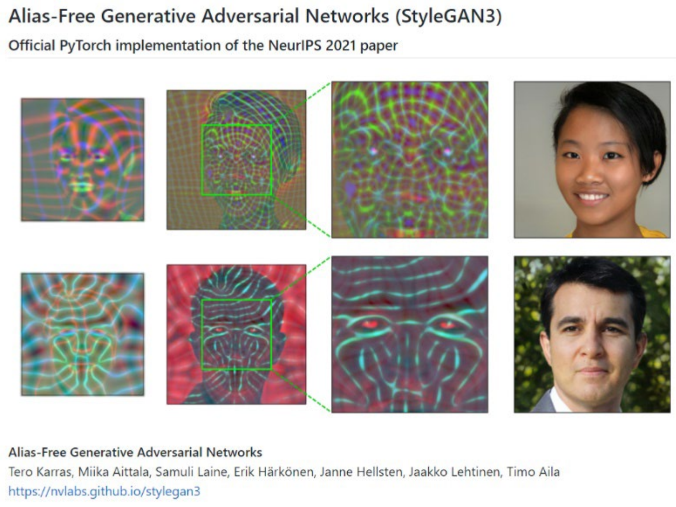
Pro-Chinese Inauthentic Network (2020)

GAN Source: Graphika

Large-scale sock puppet accounts



Threat from a high-resource actor



Alias-Free Generative Adversarial Networks (StyleGAN3)

Official PyTorch implementation of the NeurIPS 2021 paper

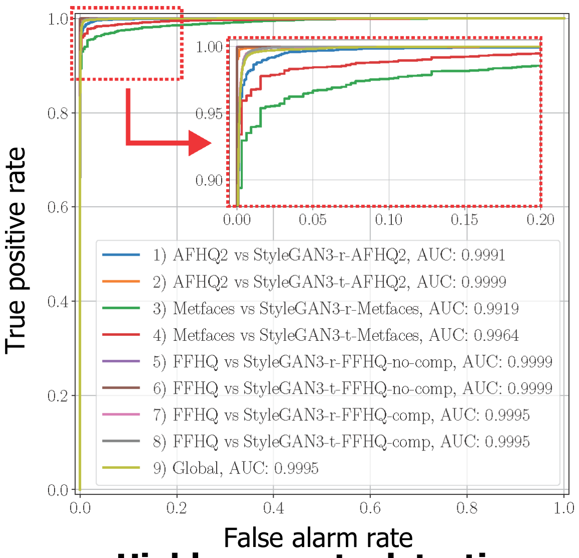
Alias-Free Generative Adversarial Networks

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, Timo Aila

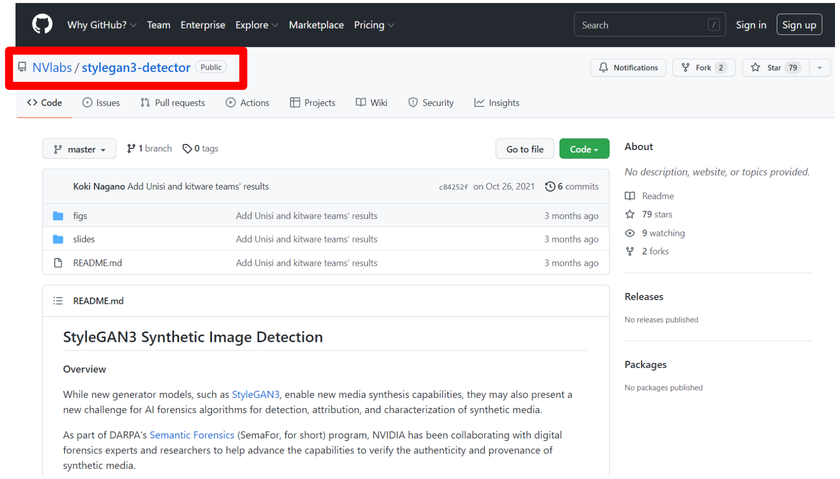
https://nvlabs.github.io/stylegan3

Training over semantic categories, augmentation, & many GANs

No training data from StyleGAN3!



Highly accurate detection



NVlabs / stylegan3-detector Public

1 branch 0 tags

Koki Nagano Add Unisi and kitware teams' results c84252f on Oct 26, 2021 6 commits

figs	Add Unisi and kitware teams' results	3 months ago
slides	Add Unisi and kitware teams' results	3 months ago
README.md	Add Unisi and kitware teams' results	3 months ago

About

No description, website, or topics provided.

Readme

79 stars

9 watching

2 forks

Releases

No releases published

Packages

No packages published

README.md

StyleGAN3 Synthetic Image Detection

Overview

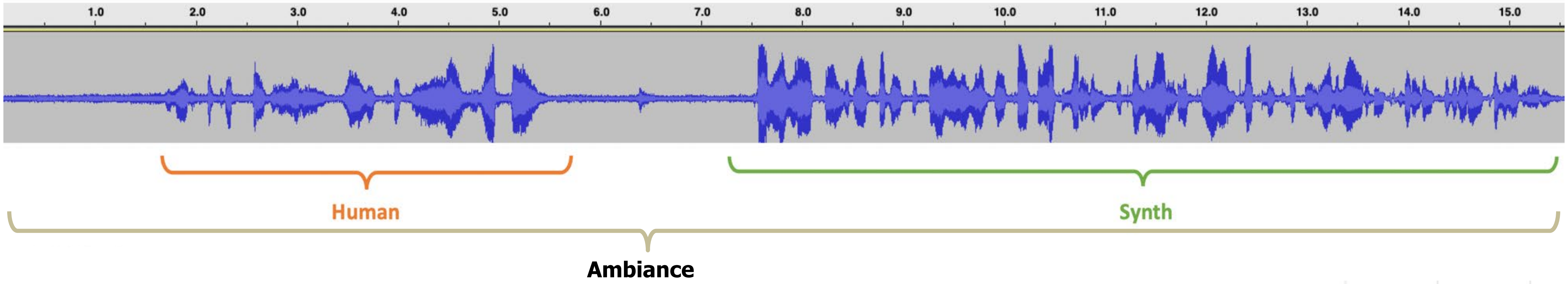
While new generator models, such as StyleGAN3, enable new media synthesis capabilities, they may also present a new challenge for AI forensics algorithms for detection, attribution, and characterization of synthetic media.

As part of DARPA's Semantic Forensics (SemaFor, for short) program, NVIDIA has been collaborating with digital forensics experts and researchers to help advance the capabilities to verify the authenticity and provenance of synthetic media.

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

Distribution A: Approved for public release: distribution unlimited.

Detecting & localizing synthetic audio



Example of humans + synth voices + background

Shawn: Has everyone had a chance to review the data?
Synth 1: It's quite disturbing. And fascinating!
Dave: It doesn't mean the vaccine isn't safe—
Synth 2: Are you really sure about that?
Dave: As sure as I am of anything.
Synth 3: Well, well, I'm not convinced—
Shawn: Nothing will convince you!
Dave: OK, we all have to calm down.
Synth 2: You want me to calm down? *Of course* I'm already calm!
Shawn: This can't be allowed to go public. If it does…
Synth 1: [Chuckle-whispers] Our scheme will be found out.
Dave: There's still time to deal with the fallout.
Synth 2: There's not much time left, though!
Shawn: Have you spoken to our friends overseas?
Dave: They're as worried as we are.
Synth 1: They have the most to lose. Quite possibly even more.

All the Feels: NVIDIA Shares Expressive Speech Synthesis Research at Interspeech

Developers and creators can access state-of-the-art conversational AI models for expressive speech synthesis to generate voices for characters, virtual assistants and personalized avatars.

August 31, 2021 by ISHA SALIAN

https://blogs.nvidia.com/blog/2021/08/31/conversational-ai-research-speech-synthesis-interspeech/

https://openreview.net/pdf?id=0NQwnnwAORi

Hackathon results (SRI)

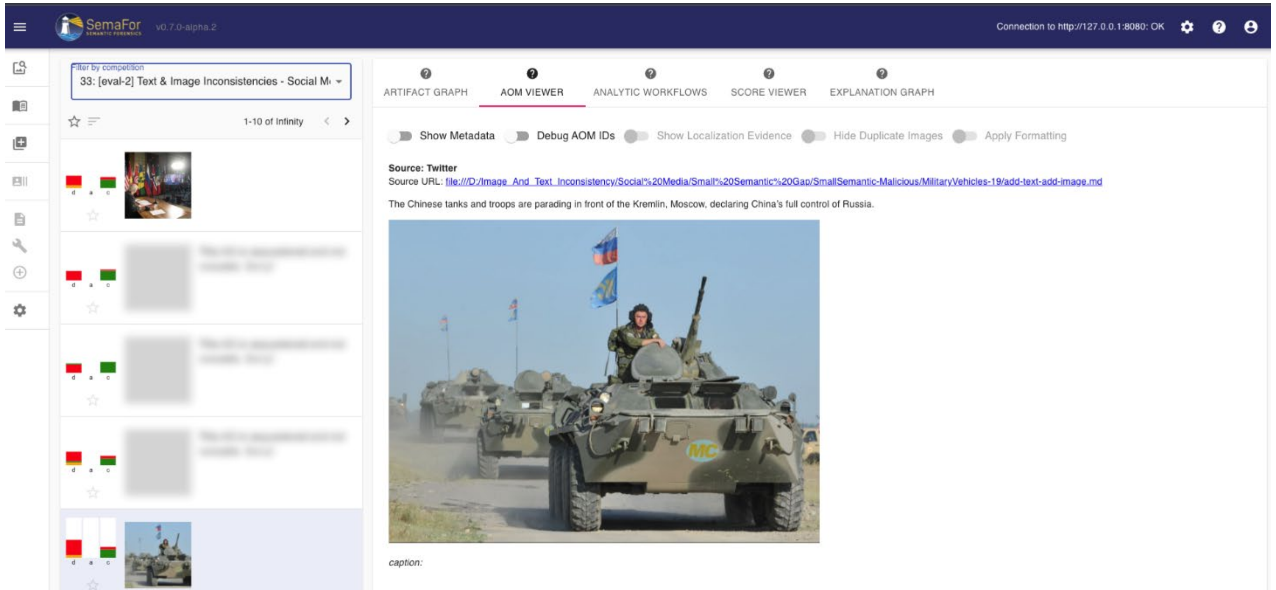
Different audio categories

	Task 1	Task 2
	DETECT % EER	LOCALIZE % EER
Subset	ResNet	ResNet
ALL	26.4	33.3
GRP1	14.1	24.0
GRP2	14.1	21.5
GRP3	18.8	38.0
GRP4	4.7	5.1
GRP5	10.9	12.2
GRP6	53.1	86.9
GRP7	26.6	27.0
GRP8	21.9	39.1
GRP9	21.9	32.6
GRP10	26.6	38.1
GRP11	12.5	28.5
GRP12	23.4	26.5
GRP13	20.3	41.5
GRP14	57.8	89.7
GRP15	10.9	30.9
GRP16	20.3	39.8

Prototype HMI

LM ATL
CRA

Media asset



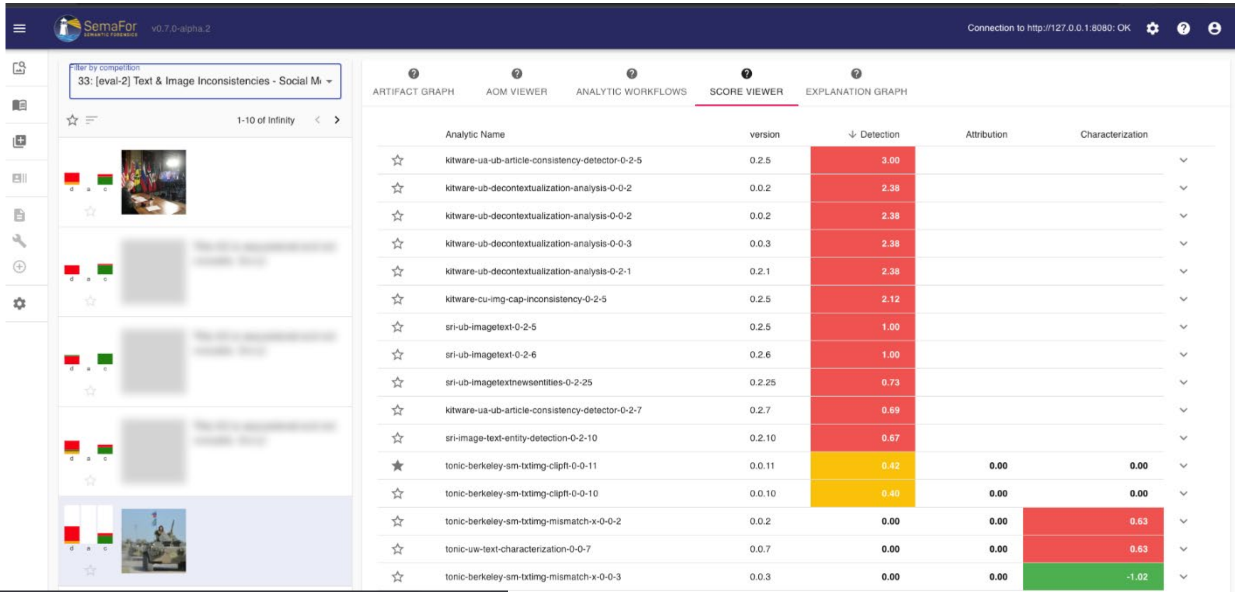
Show Metadata   Debug AOM IDs   Show Localization Evidence   Hide Duplicate Images   Apply Formatting

Source: Twitter
Source URL: file:///D:/Image_And_Text_Inconsistency/Social%20Media/Small%20Semantic%20Gap/SmallSemantic-Malicious/MilitaryVehicles-19/add-text-add-image.md

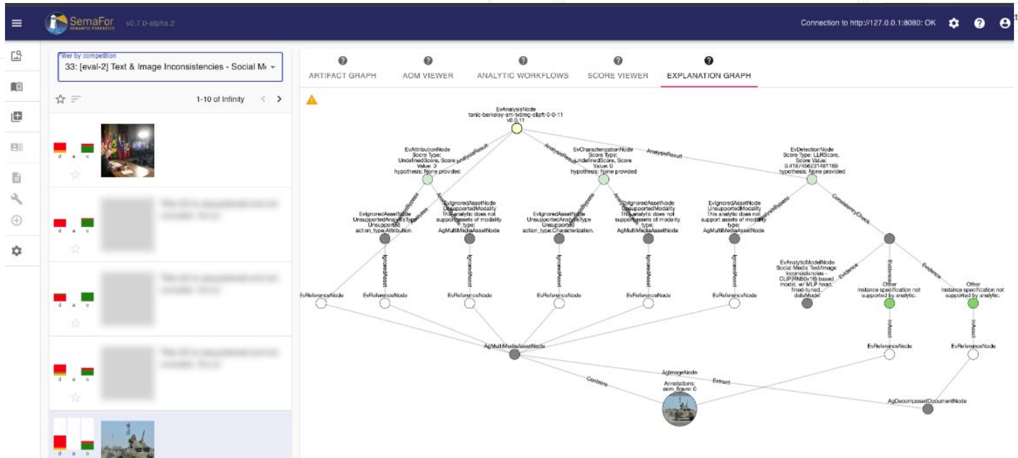
The Chinese tanks and troops are parading in front of the Kremlin, Moscow, declaring China's full control of Russia.

caption:

Algorithm results



	Analytic Name	version	↓ Detection	Attribution	Characterization
☆	kitware-ua-ub-article-consistency-detector-0-2-5	0.2.5	3.00		
☆	kitware-ub-decontextualization-analysis-0-0-2	0.0.2	2.38		
☆	kitware-ub-decontextualization-analysis-0-0-2	0.0.2	2.38		
☆	kitware-ub-decontextualization-analysis-0-0-3	0.0.3	2.38		
☆	kitware-ub-decontextualization-analysis-0-2-1	0.2.1	2.38		
☆	kitware-cu-img-cap-inconsistency-0-2-5	0.2.5	2.12		
☆	sri-ub-imagetext-0-2-5	0.2.5	1.00		
☆	sri-ub-imagetext-0-2-6	0.2.6	1.00		
☆	sri-ub-imagetextnewsentities-0-2-25	0.2.25	0.73		
☆	kitware-ua-ub-article-consistency-detector-0-2-7	0.2.7	0.69		
☆	sri-image-text-entity-detection-0-2-10	0.2.10	0.67		
★	tonic-berkeley-sm-txtimg-clipft-0-0-11	0.0.11	0.42	0.00	0.00
☆	tonic-berkeley-sm-txtimg-clipft-0-0-10	0.0.10	0.40	0.00	0.00
☆	tonic-berkeley-sm-txtimg-mismatch-x-0-0-2	0.0.2	0.00	0.00	0.63
☆	tonic-uw-text-characterization-0-0-7	0.0.7	0.00	0.00	0.63
☆	tonic-berkeley-sm-txtimg-mismatch-x-0-0-3	0.0.3	0.00	0.00	-1.02
	tnewsentities-0-2-28	0.2.28	-0.05		


SemaFor
v0.7.0-alpha.2
Connection to http://127.0.0.1:8080: OK
Filter by competition
33: [eval-2] Text & Image Inconsistencies - Social M
1-10 of Infinity
ARTIFACT GRAPH
AOM VIEWER
ANALYTIC WORKFLOWS
SCORE VIEWER
EXPLANATION GRAPH
EvAnalysisNode
tonic-berkeley-sm-txtimg-clipft-0-0-11
v0.0.11
EvAttributionNode
Score Type: UndefinedScore, Score Value: 0
hypothesis: None provided
EvCharacterizationNode
Score Type: UndefinedScore, Score Value: 0
hypothesis: None provided
EvDetectionNode
Score Type: LLRScore, Score Value: 0.4187456231481189
hypothesis: None provided
AnalysisResult
EvIgnoredAssetNode
UnsupportedAnalysisType
Unsupported action_type:Attribution.
EvIgnoredAssetNode
UnsupportedModality
This analytic does not support assets of modality type: AgMultiMediaAssetNode
EvIgnoredAssetNode
UnsupportedAnalysisType
Unsupported action_type:Characterization.
ConsistencyCheck
IgnoredAsset
EvAnalyticModelNode
Social Media Text/Image Inconsistencies - CLIP(RN50x16) based model, w/ MLP head, fined-tuned. dataModel
Evidence
Other
Instance specification not supported by analytic.
EvReferenceNode
InAsset
AgMultiMediaAssetNode
AgImageNode
Annotations: aom_figure: 0
Contains
Extract
AgDecomposedDocumentNode

Evidence representation

DARPA

www.darpa.mil

Distribution A: Approved for public release: distribution unlimited.