Semantic Forensics

MICHAEL KOZAK

The Purpose of This Tutorial

Primary Goal: Help researchers and engineers understand the current state of generative AI and the *tools and techniques best suited to build detectors* capable of identifying real vs generated or manipulate media.

Secondary Goal: Understand the difference between an “algorithm” and an “analytic” when it comes to *producing detectors that are valuable to humans.*

Who will benefit from this tutorial:

- **Researchers** familiar with ML techniques should walk away with potential new approaches to take in developing, training, and evaluating the effectiveness of models created to perform semantic forensics
- **Principle Investigators and Lead Engineers** should walk away with a better understanding of requirements that will help an algorithm achieve a high Human Readiness Level (HRL) in an integrated system
- **Business Development** personnel should walk away with a clear understanding of the state of the field, what research gaps remain, and what transition partners and funding sources look for in measuring success

Agenda

Part 1: Introduction and key concepts

- Introduction to the Instructor, and his parent organization
- Generative techniques – a brief history and current state
- Medifor and Semafor – addressing a changing landscape
- D/A/C – going beyond detection

Part 2: Algorithmic approaches and data challenges

- Guided walkthrough and free play
- Building a Detector – Technical Considerations and tricks
- Training Data
- Advanced Concepts: explainability and evidence
- Analytic Promotion – a use case

Part 3: The SemaFor Platform

- Developing a SemaFor Analytic
- UL and the future of the platform
- Feedback session on the SemaFor Portal and Q&A

About the Instructor

- 17+ years experience in Applied Research at Lockheed Martin Advanced Technology Laboratories
- BS/MS in Computer Science from Drexel University in Philadelphia, PA with concentrations in Artificial Intelligence and User Interface design.
- Deputy PI on the DARPA Semantic Forensics program as part of the Technical Area 2 (Systems Integrator) role.
  - DARPA SemaFor (and its predecessor MediFor) represents roughly a decade of research into detecting, attributing, and characterizing generated and manipulated media.
  - Responsibilities include development of an Explainability component for interpreting results and directing tasking across our HMI and Fusion teams based on user engagement and feedback.
  - Direct Point of Contact for users across over a dozen DoD organizations interested in exploring the use and integration of SemaFor into their operational workflows
  - Have supported and ran half a dozen Transition Workshops focused on better understanding our users, their use cases, their limitations, and their workflows.

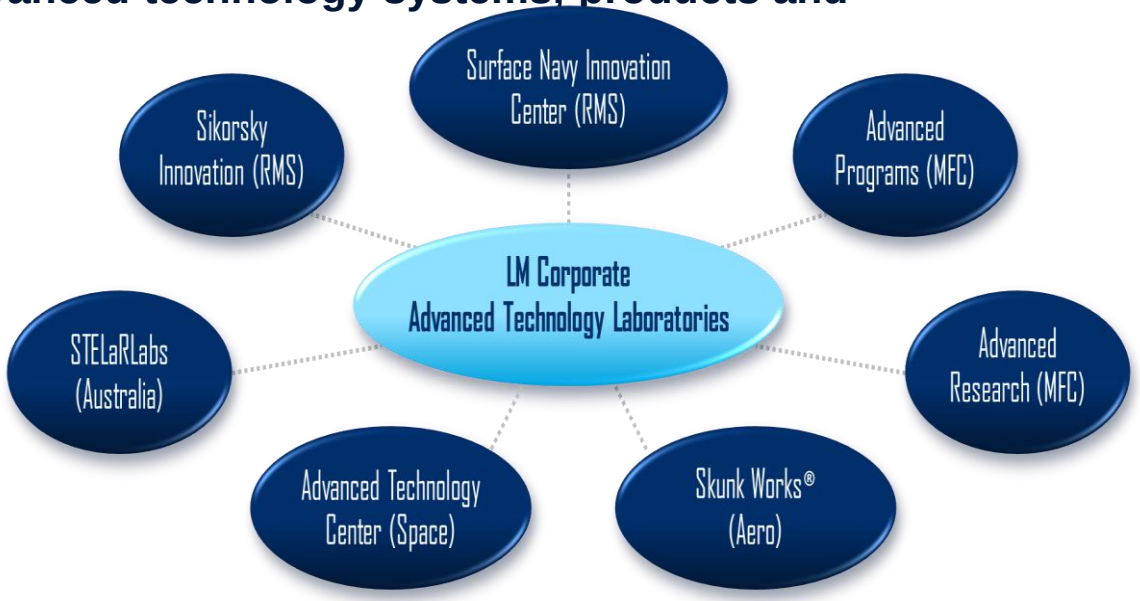
Research Patrons

The content you are about to see in this tutorial contains some material developed through funding provided by the Defense Advanced Research Projects Agency (DARPA).

Continued funding past the DARPA SemaFor program period of performance is provided by the Underwriter Laboratories Digital Safety Research Institute (UL/DSRI)

LOCKHEED MARTIN

Headquartered in Bethesda, Maryland, Lockheed Martin is a global security and aerospace company that employs approximately 120,000 people worldwide and is principally engaged in the research, design, development, manufacture, integration and sustainment of advanced technology systems, products and services.



LOCKHEED MARTIN ADVANCED TECHNOLOGY LABORATORIES WORK TO SIMULTANEOUSLY ADVANCE CUTTING EDGE RESEARCH & DEVELOPMENT APPLICATIONS ON MULTIPLE FRONTS WITH A COHESIVE, COMPREHENSIVE STRATEGY THAT WILL ENHANCE WARFIGHTER EFFECTIVENESS FOR NEXT-GENERATION BATTLEFIELD DOMINANCE

ADVANCED TECHNOLOGY LABORATORIES

DR. ROBERT MANDELBAUM
ATL DIRECTOR

WHO we are…

Science and engineering professionals with expertise in:

- Computer Science
- Electrical Engineering
- Physics
- Mathematics
- Artificial Intelligence
- Human Systems Researchers

WHERE we are…

Approx. 250 staff in four primary locations across the U.S. :

- Cherry Hill, NJ (headquarters)
- Arlington, VA
- Eagan, MN
- Kennesaw, GA

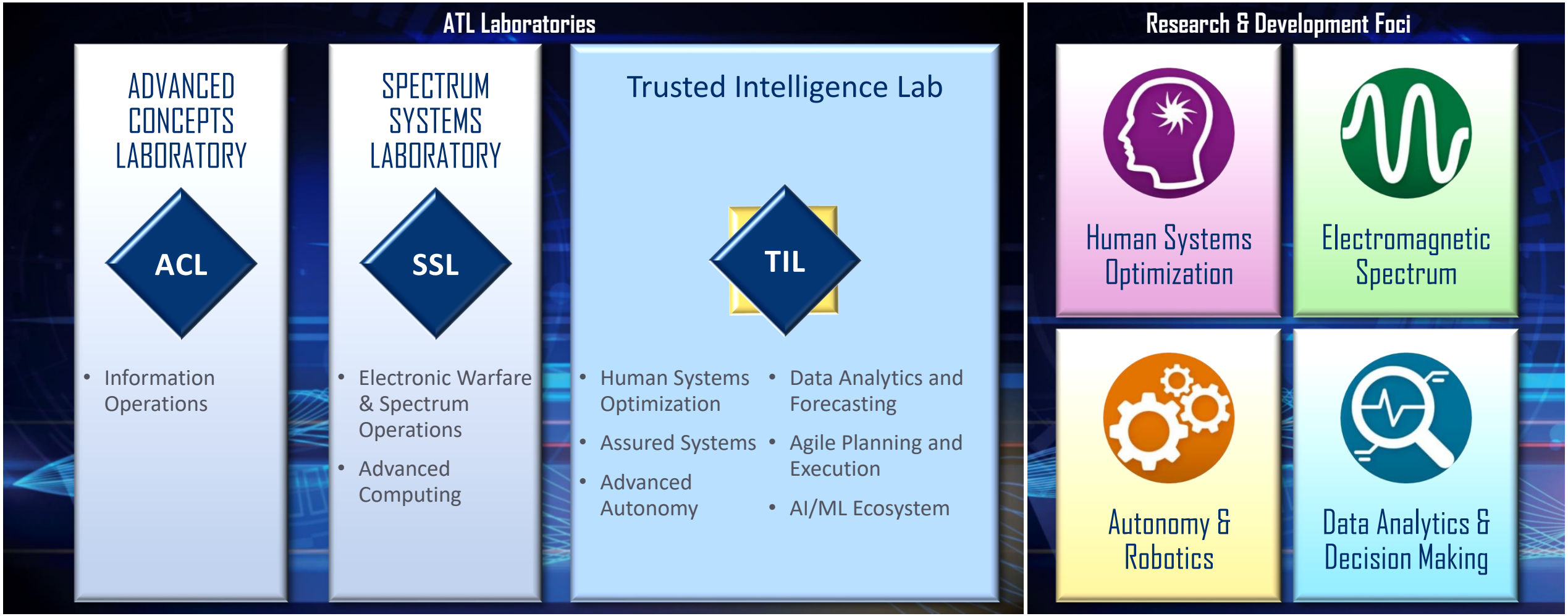
our PARTNERS…

Collaborating with the best and brightest in government, industry and academia.

Primary research partners include:

- DARPA
- U.S. government service laboratories and other S&T organizations
- Universities
- Small and medium business
- Lockheed Martin colleagues

ATL Research Laboratories



Part 1

INTRODUCTION AND KEY CONCEPTS

History of Generative AI

- The earliest form of Generative AI took the form of simplistic chatbots in the 1960s such as ELIZA – one of the earliest subjects of the Turing Test.
- Like ELIZA, these early attempts relied on pattern matching scripts, symbolic representations for search, and autoencoders to generate samples that were similar to the input, often mirroring key words in the initial prompt.
- **Modern Generative AI didn't take off until the invention of Generative Adversarial Networks (GANs) in 2014**

What about Watson and LLMs?

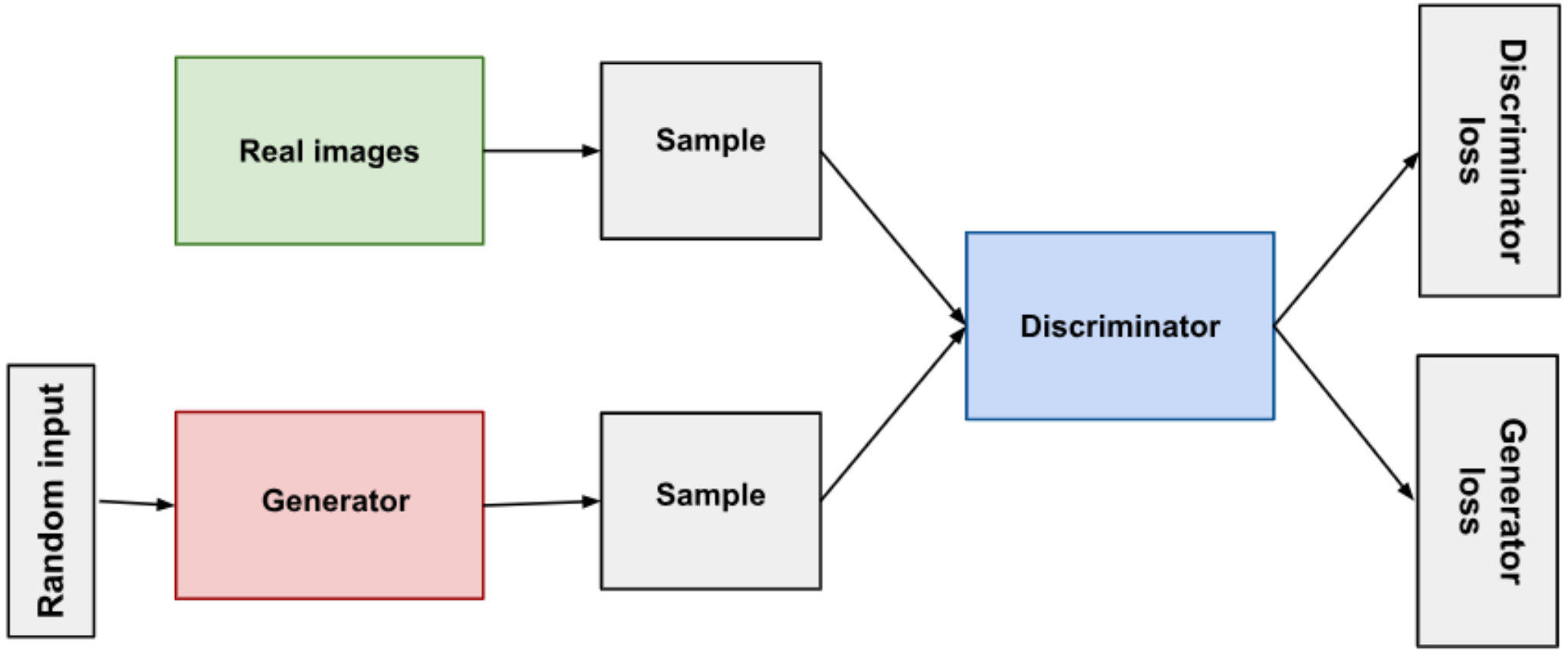
- IBM Watson beat the two highest ranking Jeopardy players in 2011 – isn't that good?
- It certainly did an excellent job of using Natural Language Processing (NLP) to understand the prompts – an integral part of any generative system – but it also required 90 servers to run.
- But functionally it used a **consensus-based approach**, with hundreds of algorithms parsing the question from different angles with the most returned answer becoming the one it gave **only if a total confidence value was above some threshold**. The real power was in the parallel processing and knowledge base.
- Similarly, LLMs are a key tool in support of modern Generative AI systems, especially in converting text prompts to model inputs by extracting semantic information, but **they are a limited to text without being coupled to other approaches**.

What are GANs?

- Generative Adversarial Networks (GANs) are a **pair of neural networks trained together** in an adversarial manner: a generator and a discriminator
- **The Generator** is trained to produce plausible data that mimics what it has been trained on.
- **The Discriminator** is trained to tell the difference between the two - it is responsible for ensuring the Generator produces good results.
- As the Generator gets better, the Discriminator's performance drops – the lower the performance the better the results.

https://developers.google.com/machine-learning/gan/gan_structure

What are GANs?



https://developers.google.com/machine-learning/gan/gan_structure

GAN Training

The Discriminator is *trained on real and generated data.*

- Its loss function penalizes it *when it misclassifies either set*
- It needs a *classifier* that's been trained to detect real vs fake to start to reduce training time.

The Generator is *trained on random noise* which will produce a wide variety of samples from the target distribution

- Its loss function penalizes it *when it fails to fool the Discriminator*
- This means backpropagation starts at the Discriminator
- We fix the Discriminator weights during this process to generate our gradients which will only change the Generator weights to prevent a "moving target" in training

Training – the Chicken and the Egg

If the Generator needs a trained Discriminator, and a Discriminator needs output from a trained Generator, how do we train?

The solution – alternate training between both:

- Train the Discriminator for 1+ epochs
- Then train the Generator for 1+ epochs
- Repeat over and over until we converge

When are we done? If we go too far, the Discriminator accuracy will be so low our Generator will actually *get worse* due to bad feedback.

- In other words - convergence is *temporary and fleeting.*

GAN Variations

1. **Progressive GANs** – creates higher resolution images with less training
2. **Conditional GANs** – request generated output from a predetermined set of input labels.
3. **Image-to-Image Translation GANs** – *fill in details or change properties* within the bounds of an outline provided as input.
4. **Cycle GANs** – Given 2 sets of unlabeled images, can *transform* input images from one set to the other.
5. **Text-to-Image Synthesis** – *what many people think of when discussing GANs* – these generate images from text provided as input (prompt) - note that in this system the GAN can only produce images from a small set of classes.
6. **Super Resolution GANs** – *adds details* to blurry sections of input images including patterns.
7. **Inpainting** – *fills in missing chunks* of input images from a given set, such as faces.
8. **Text-To-Speech** – converts words to synthesized voice

GANs in other Modalities

- **Video GANs** – use 3D Convolutional Neural Networks (CNNs) and other techniques to adapt to the temporal nature of video
  - Multiple generators may be involved to cover foreground vs background

- **Audio GANs** – can be trained to produce specific types of sound, such as percussive notes or words
  - Generates by upsampling from a single vector to a full waveform

https://dl.acm.org/doi/fullHtml/10.1145/3487891

Diffusion Models

Diffusion-based neural networks are a newer approach to GANs, having really only existed as a competitor since 2020. They are trained through deep learning to *progressively "diffuse" samples with random noise, then reverse that diffusion* process until it converges to a specific data distribution. The process involves:

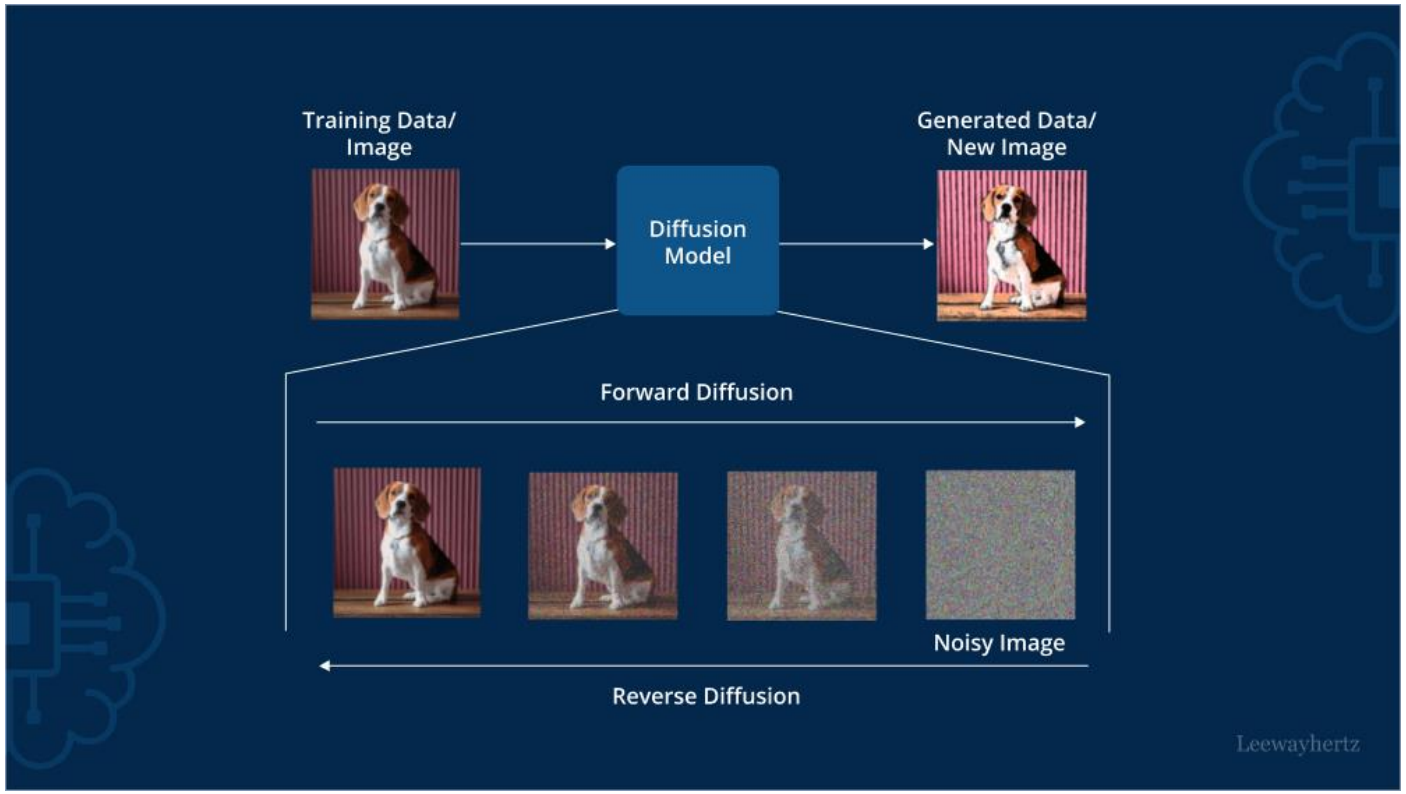
1. **Forward diffusion**: a Markov chain that progressively adds noise to the input data until it becomes a standard normal distribution.
2. **Reverse diffusion**: a neural network that learns to reverse the forward diffusion process, effectively denoising the input data and generating a sample from the target distribution.

The key difference between GAN and Diffusion is in their approach to generative modeling:

- GANs use an adversarial process to learn a direct mapping from noise to data.
- Diffusion models use a probabilistic process to learn a sequence of transformations that progressively refine the input noise until it converges to the target distribution.

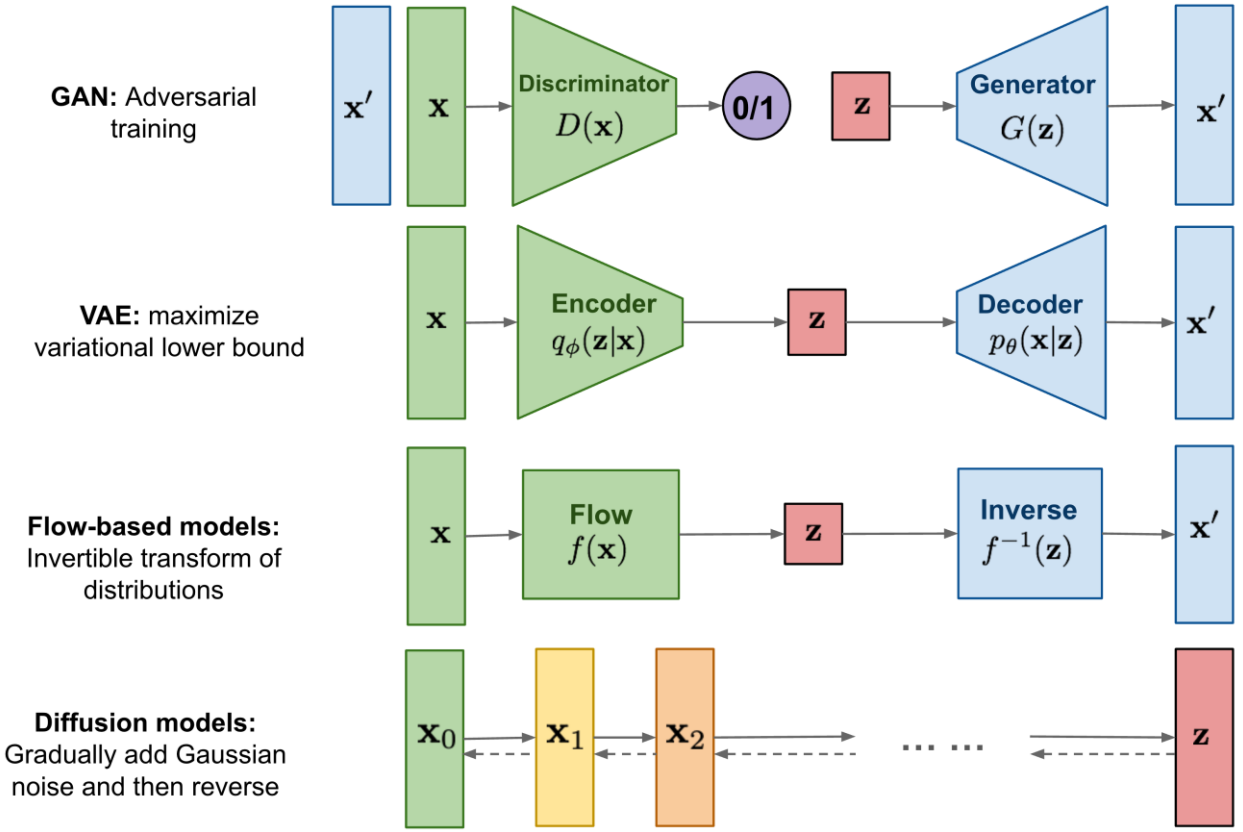
Diffusion Model Training

- In this example, the image of the dog is first put through the Forward Diffusion process, injecting noise across multiple layers.
- That noisy image is then put through Reverse Diffusion to try and learn how to reproduce the input.
- Latent variables are captured by the intermediate layers and over time come to represent the features and patterns of the dataset that can be used to generate new samples
- The closer the output is to the input, the less loss.



https://www.leewayhertz.com/diffusion-models/

GAN vs Diffusion



https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Diffusion Variations

1. **Denoising Diffusion Models**: These models use a Markov chain to progressively add noise to the input data and then learn to reverse this process to generate samples from the target distribution.
2. **Denoising Diffusion Probabilistic Models**: These models use a probabilistic approach to model the denoising process, allowing for more flexible and efficient sampling.
3. **Score-Based Diffusion Models**: These models use a score function to model the gradient of the log probability density function, allowing for more efficient and accurate sampling.
4. **Latent Diffusion Models**: These models use a latent space to represent the input data, allowing for more efficient and flexible generation of samples.
5. **Conditional Diffusion Models**: These models use a conditioning variable to control the generation process, allowing for more targeted and specific generation of samples.
6. **Diffusion-SDE Models**: These models use stochastic differential equations (SDEs) to model the diffusion process, allowing for more accurate and efficient modeling of complex systems.
7. **Non-Linear Diffusion Models**: These models use non-linear transformations to model the diffusion process, allowing for more flexible and accurate modeling of complex systems.
8. **Multi-Scale Diffusion Models**: These models use multiple scales to model the diffusion process, allowing for more efficient and accurate modeling of complex systems.

GAN vs Diffusion

	GANs	Diffusion Models
Architecture	Adversarial (generator + discriminator)	Probabilistic (forward + reverse diffusion)
Training Objective	Minimax game (generator vs discriminator)	Maximum likelihood estimation (reverse diffusion)
Mode Coverage	May suffer from mode collapse	Can generate diverse samples with good mode coverage
Training Stability	Can be unstable and sensitive to hyperparameters	Generally more stable and easier to train
Image Quality	Can produce highly realistic images	Can produce high-quality images with good texture and detail

Classes of Generative Models

Models can be roughly separated by their parent technique:

GANs

- Pix2Pix, StyleGAN, ProGAN, StarGAN, GigaGAN, etc.

Diffusion

- Latent/Stable, SDXL, Adobe Firefly, DallE, Midjourney

Encoders / Transformers (primarily for audio)

- RTVC, MTVC, etc

Hybrid Apprpoaches

- Meta AI (GAN + Diffusion), Eleven Labs (GAN + Transformer), XTTSv2 (Diffusion + Autoencoder + Vocoder), SORA (Diffusion + Transformer)

The Generative Tower of Babel

- **Detectors will generalize somewhat within a class** when a new Generator is released, but a new class of Generator will require training an entirely new Detector.
  - Think of it like unfamiliar dialects/accents versus different languages.

Why Does This Matter?

When it comes to detecting, attributing, and characterizing generated and manipulated media, there is an inherent **balance between accuracy, robustness, and longevity**:

- **Highly accurate models** will have a more narrow range of generators it can detect and will only be valuable as long as those generators are in use
- **Highly robust models** will have a broader range of generators in its training set, trading some absolute accuracy for a higher chance in "open world" settings. These models will have a longer shelf life but will see gradually degraded accuracy over time.
- **High longevity models** may choose to diversify only within a class of generators, trading some "open world" robustness and accuracy for slower long-term degrading of usefulness.

Why Does This Matter?

The race against new generative techniques is a sprint - So what kind of model you choose to build will determine how often you need to retrain it, how long before you have to build an entirely new model, and how long before the generators themselves evolve past detection.

Use Case: Professor Siwei Lyu (University at Buffalo, NY) identified in 2018 while performing on DARPA MediFor that deepfakes never blinked.

By 2019 generators had already improved their training data to incorporate blinking, forcing researchers to look into *blink patterns* – a much harder problem.

https://www.researchgate.net/publication/340813598_DeepVision_Deepfakes_Detection_Using_Human_Eye_Blinking_Pattern#pf2

Why Does This Matter?

- Statistical models focused on letting the detector learn the hidden patterns and biases in the data still perform well when trained on sufficient data created by the generator itself, but **access to the generator is not always a guarantee.**
- They also suffer from **brittleness against changes to the image**, such as resizing or recompressing, that alter the pixels themselves and thus obscure the hidden patterns.
- Further, these models **don’t generalize well** across classes of generators, meaning they become obsolete the fastest.

Semantic Forensics Overview

SemaFor
SEMANTIC FORENSICS

Synthetic and Manipulated Media Analysis

- Technology to create and deploy synthetic media increasing at a rapid rate, with **new models available for use in the public domain every week**
- The **information environment is increasingly saturated** with AI generated content across all modalities – text, images, audio, video
- The **models are increasingly sophisticated**, making traditional methods of semantic and forensic analysis difficult to discern realistic from synthetic or manipulated artifacts
- What is required is **an ensemble of capabilities** that aim to detect, attribute, and characterize synthetic and manipulated content in the information environment

Amid a rise in swatting calls, the fabrication and fear of mass shootings collide

Law enforcement is unable to solve the majority of these cases.

https://abcnews.go.com/US/amid-rise-swatting-calls-fabrication-fear-mass-shootings/story?id=98672825

AI Can Convincingly Mimic A Person's Handwriting Style, Researchers Say

"We'll have to create public awareness and develop tools to combat forgery," one of the researchers said.

https://news.bloomberglaw.com/artificial-intelligence/ai-can-mimic-a-persons-handwriting-style-researchers-say

Pro-China YouTube Network Used A.I. to Malign U.S., Report Finds

https://www.nytimes.com/2023/12/14/business/media/pro-china-youtube-disinformation.html

https://www.cnn.com/2024/01/22/politics/fake-joe-biden-robocall/index.html

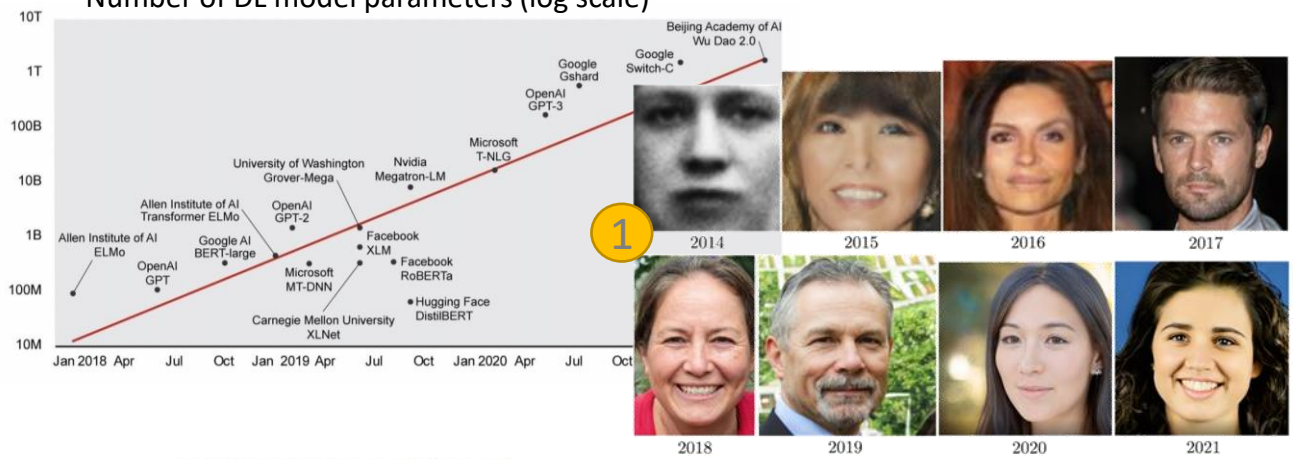
Overview: Why Semantic Analysis?

1. **Sophisticated Generators**
   Traditional statistical methods are becoming less reliable as media generation and manipulation become more sophisticated.

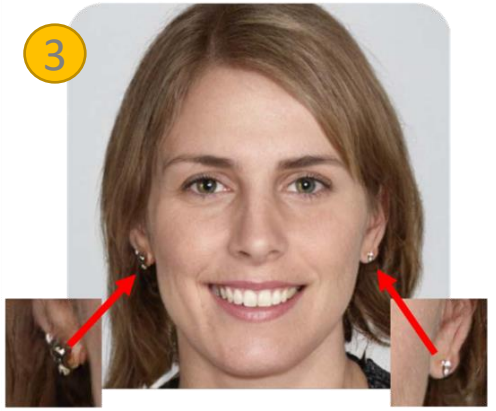
2. **Multimodality Construction**
   Some falsifications *do not* rely on direct manipulation, but rather use the overall construction to mislead.

3. **Generated Inconsistencies**
   Synthetic generation can leave *semantic errors* in the content.

Number of DL model parameters (log scale)



People riding bicycles down the road approaching a dog.



Semantic inconsistencies help identify media that has been altered or artificially generated.

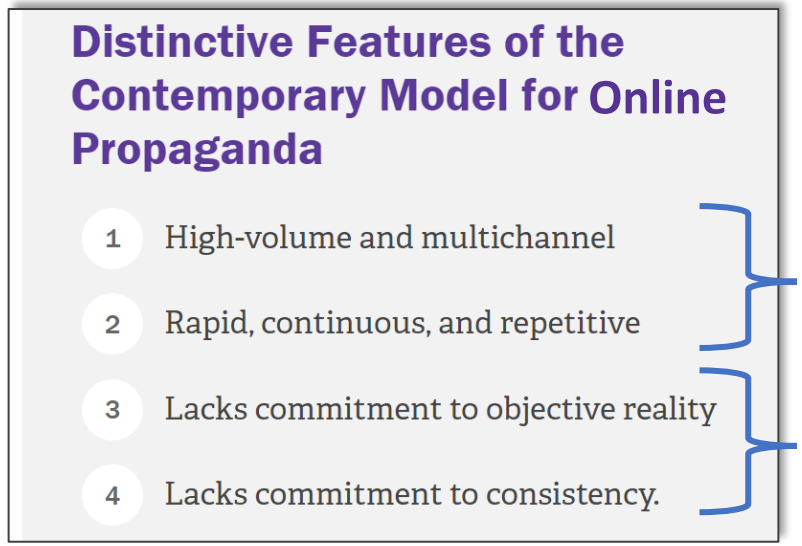
Overview: Why Automated Analysis?

High volumes of falsified content makes manual analysis techniques impractical.

Automatically identifying falsified media at scale:

- Enables timely *shutdown* (or *reduction*) of the flow
- Provides an early opportunity to point out falsehoods and inconsistencies since psychological evidence suggests refutations are not as effective after disinformation gains traction.

The "Firehose of Falsehood" Model



Information Credit: https://www.rand.org/pubs/perspectives/PE198.html

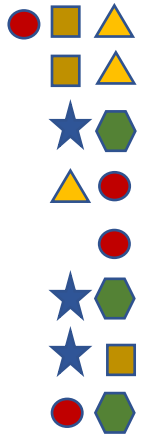
Anecdote: A recent transition partner who provides online protection services for Gov't/Military VIPs had to address over 40K false profiles in one year.

The scale, rapidity, and diversity of disinformation content requires automated analysis.

Problem: Adversary Tactics and Techniques

Adversarial actors use a variety of tools, techniques, and tactics to influence others, stir them to action, and cause harm.

Example CISA Tactic Categories:



- Cultivate Fake or Misleading Personas and Websites
- Create Deepfakes and Synthetic Media
- Devise or Amplify Conspiracy Theories
- Astroturfing and Flooding the Information Environment
- Abuse Alternative Platforms
- Exploit Information Gaps
- Manipulate Unsuspecting Actors
- Spread Targeted Content

Analysis Required

- Scale
- Synthetic Media
- Intent
- Tactics, Tools, and Techniques
- Content: Topics, POI, Entities and Events

Information Credit: **Cybersecurity and Infrastructure Security Agency (CISA)** *Tactics of Disinformation – Publication 508*
https://www.cisa.gov/sites/default/files/publications/tactics-of-disinformation_508.pdf

Sophisticated tactics, tools, and techniques require analysis beyond traditional detection methods.

Problem: Pace of Synthetic Media Generation

Text-to-image models make targeted production of (high quality) images within reach.

(a) *"A raccoon wearing formal clothes, wearing a top hat and holding a cane. The raccoon is holding a garbage bag. Oil painting in the style of Rembrandt"*

(b) *"A bald eagle made of chocolate powder, mango, and whipped cream"*

(a) 2015

(b) 2022

Image Credits: Goldstein et al, *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*
https://fsi9-prod.s3.us-west-1.amazonaws.com/s3fs-public/2023-01/forecasting-misuse.pdf

The pace and sophistication of generation capabilities is outpacing traditional analytic techniques.

Use Case: AI Generated Image of Pentagon Explosion

An apparently AI-generated hoax of an explosion at the Pentagon went viral online — and markets briefly dipped

Rebecca Cohen May 22, 2023, 8:04 AM MST

- An apparently AI-generated photo faking an explosion near the Pentagon in D.C. went viral.
- The Arlington Police Department confirmed that the image and accompanying reports were fake.
- But when the news was shared by a reputable Twitter account on Monday, the market briefly dipped.

Top editors give you the stories you want — delivered right to your inbox each weekday.

Telegram account that posted the images.

https://www.insider.com/ai-generated-hoax-explosion-pentagon-viral-markets-dipped-2023-5

Breaking News!

- As of the end of April, the latest version of Adobe can now run 3rd party generators from within Firefly. Previously Adobe limited generation to their proprietary Firefly model which was trained on licenses or open source data – avoiding the thorny legal issue of training data ownership.
- This means you'll only need one tool to generate media across dozens of different generators – further simplifying the process of generating content across a wide range of formats.
- GPT 4o, Imagen 3, and Veo 2 are already available. Fal, Ideogram, Pika and Runway will be coming soon, with a promise to have **up to 30 generators available in the near future.**
- **The Good News**: Adobe Content Credentials are automatically added to all AI-generated material, making it clear which model was used.
- They also just launched the free **Adobe Content Authenticity app** in public beta. This will let them tag their own works with a digital signature and preferences on whether to allow their media to be used in training. This is part of their work on as members of **The Coalition for Content Provenance and Authenticity (C2PA)**

https://www.creativebloq.com/ai/adobe-supports-third-party-generators

The Future is now

An AI-generated video from Veo 3: "A male stand-up comic on stage in a night club telling a hilarious joke about AI and crypto with a silly punchline." An AI language model built into Veo 3 wrote the joke.

You no longer need expertise to generate synchronized, realistic, audio and video

The Future is now

https://arstechnica.com/ai/2025/05/ai-video-just-took-a-startling-leap-in-realism-are-we-doomed/

AI-generated video from Veo 3: "An old professor in front of a class says, 'Without a firm historical context, we are looking at the dawn of a new era of civilization: post-history.'"

The SemaFor Platform

SemaFor
SEMANTIC FORENSICS

Status Quo

The US is engaged with its adversaries in an asymmetric, continual, war of weaponized influence narratives.

- Peer Competition – Disinformation is increasingly used as a tool to achieve national security objectives without kinetic warfare.
- Sophisticated Messaging – Advanced AI/ML tools are available to create hard to detect disinformation content at scale.
- Extensive Delivery – Falsified and recontextualized information are widely delivered via online multimedia (social media).
- Scale – Detection of falsified media is often done manually but the scale of the problem dwarfs all non-automated approaches.

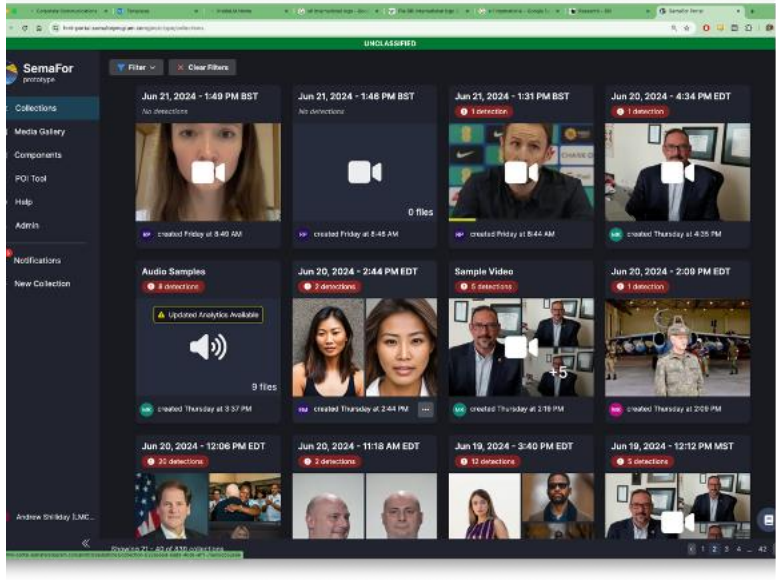
New Insights

Scalable system that automates multimedia deepfake detection and attribution analysis using SOTA algorithms is needed.

- A rich set of algorithms that automatically: detect ("what"), attribute ("how"), and characterize ("why") deepfakes
- Latest evolution of manipulation techniques requires semantic analysis and integrated evaluation
- Detection within and across modalities is critical to identifying falsified media.
- Harvesting capabilities across world class research organizations (well beyond what smaller commercial companies can provide) enables larger breadth of analysis
- End-to-end analysis platform and rigorous evaluations build trust in analytic results

How It Works:

The Semantic Forensics Program represents a significant Gov't investment to deliver a state-of-the-art detection, attribution, and characterization analytic platform to exploit structural and semantic inconsistencies in AI-manipulated and synthetic cross-modal media.



- Comprehensive analytics are integrated into a **distributed, scalable platform** and accompanying user interface to enable advanced multi-model analytic workflows for evaluation and operationalization
- SemaFor platform integrates hundreds of algorithms into advanced modality-based workflows for comprehensive analysis and improved robustness

SemaFor provides a sizable suite of detection, attribution, and characterization algorithms organized in and orchestrated by an advanced distributed computing platform and easy-to-use web interface

IMPACT

Meets DoD/IC Needs

- Extensible, pluggable interface for managing, self-evaluating, and exercising collections of analytic components at scale.
- Operates in **offline, disconnected environment** providing SOTA detection and attribution of existing and evolving deepfake generation techniques.
- All software developed under the SemaFor program is **Government-off-the-Shelf** (GOTS) software and is immediately available to DIU and its government partners.
- Robust third-party validation of algorithm performance utilizing in-system evaluation support tools. Partnership with UL to conduct open challenges and evaluations for broader engagement with research community.
- Closely intertwined with broad community of algorithm developers to maintain currency against new and evolving threats.
- Considerable DARPA prioritization on higher TRL platform developed over 4 years of investment.

SEMAFOR PLATFORM

- Advanced user-driven graphical interface for performing analysis on multi-modality media
- Runtime framework, SDKs, APIs for simplified algorithm integration
- Significant Community Engagement within DoD and IC:
  - **6 Transition Partner Workshops** where participants are introduced to the system and provide feedback.
  - **Engagement and buy-in from 25+ DoD/IC transition partners**
  - **Initial experimental deployment to 12+ locations including** SOCOM, CENTCOM, SOUTHCOM, HSI, First IO, AFRL, NSA, OSE, DC3, FBI, NMEC, and international partners.
  - **Non-CUI web portal has 800+** registered users, many of which are highly active accounts
  - Deployable to single machines, multi-node clusters, AWS, C2S, and Azure

Lockheed Martin POC

Andrew Shilliday, PhD
Principal Investigator
Lockheed Martin, ATL
andrew.e.shilliday@lmco.com

DARPA POC

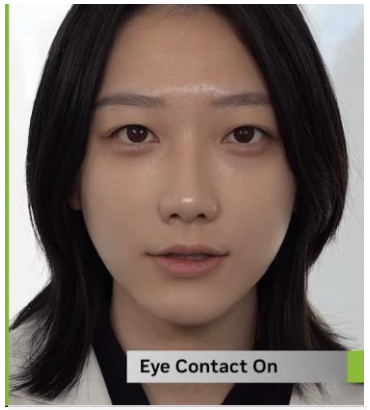
Wil Corvey, PhD
Program Manager
Information Innovation Office (I2O)
william.corvey@darpa.mil

Detection of Generated Content

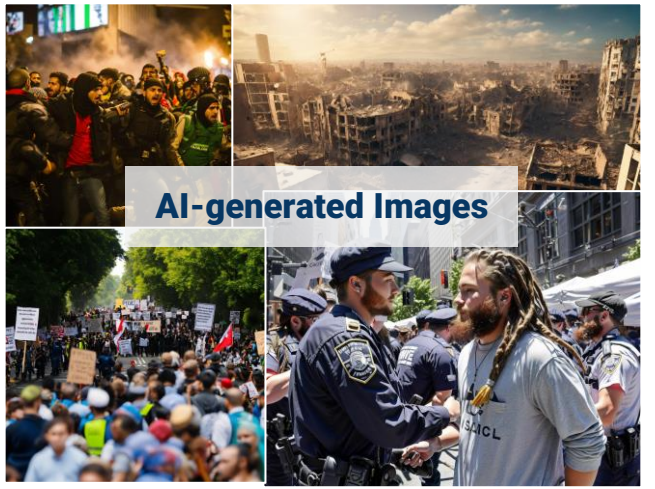
SemaFor
SEMANTIC FORENSICS

SemaFor detection capabilities for generated content includes:

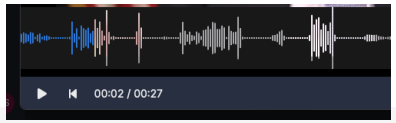
- Detection of image, video, audio, and text generators and manipulations using dozens of analytic algorithms
- Ability to detect generated media with and without human subjects
- Design features to address unknown future generators
  - Unsupervised models trained only on liveness (real) media (e.g., Kitware diffusion-generated image detection)
  - Standardized algorithm containerization and interfaces supporting CI/CD deployment
- Performance results based on large-scale, sequestered evaluations across dozens of generators and manipulation types



Nvidia Maxine-synthesized Video



AI-generated Images



XTTSv2 Deepfake Audio

Key SemaFor Evaluation Results

SRI DeepFake Audio Detection: 1.0 Pd @ 5% PFA
Kitware-EduWorks Audio Detection: 1.0 Pd @ 5% PFA
SRI/UMD Synthetic Video Detection: 0.93 Pd @ 5% PFA
Kitware UB DeepFake Video Detection: 0.96 Pd @ 8% PFA
Kitware Diffusion-generated Image Detection: 1.0 Pd @ 5% PFA
Kitware ASU GPT4 Text Detection: 93% Accuracy

Refer to algorithm tables in the appendix of this presentation.

SemaFor provides a large catalog of analytics to robustly detect manipulated multimedia across dozens of generators.

SemaFor Feature-Rich User Interface

SemaFor
SEMANTIC FORENSICS

Labeled and colored indicators quickly depict analytic overall conclusions

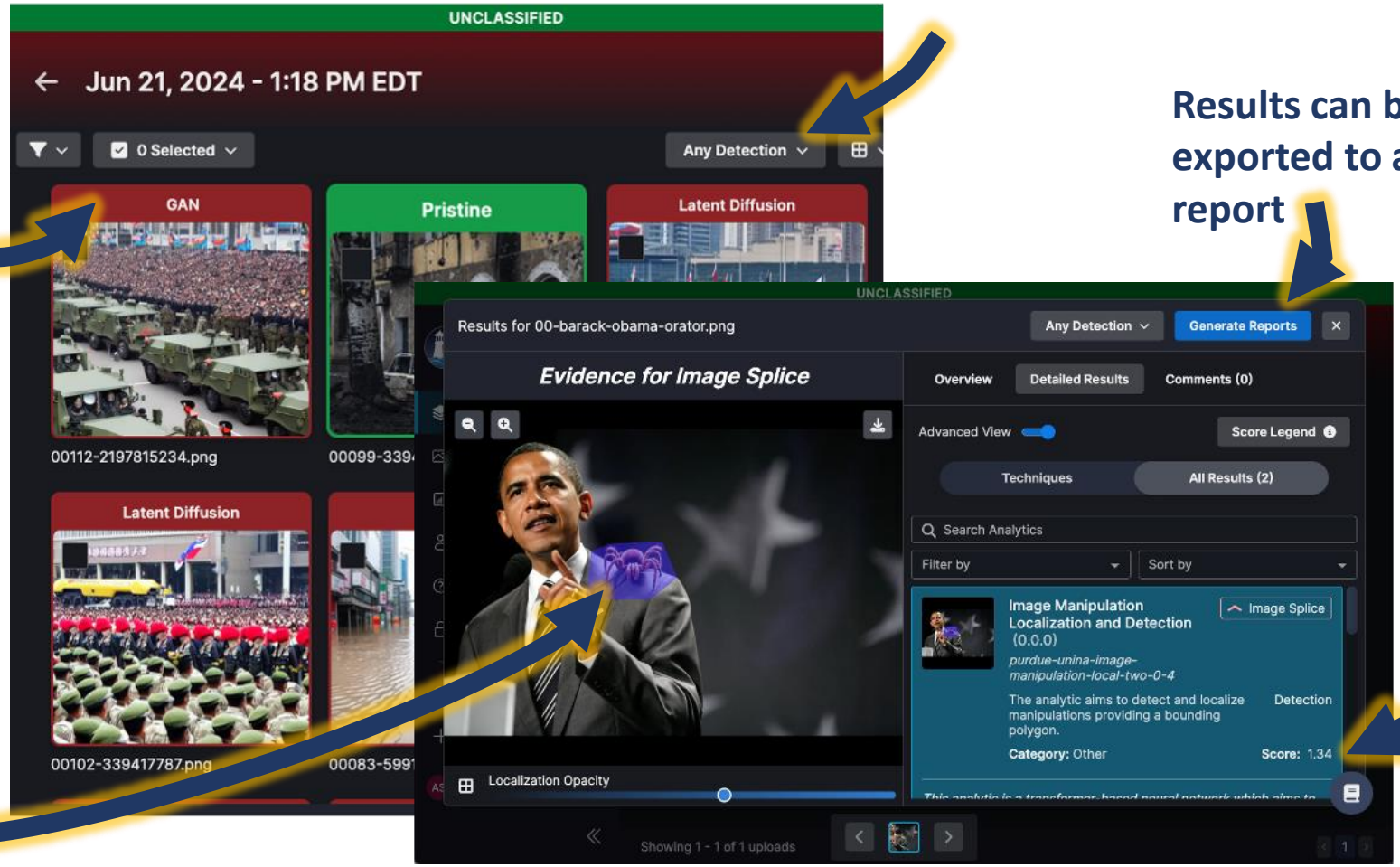
Deep-dive into media results for detailed analysis

When available, additional supporting evidence is highlighted

Results are aggregated or fused based on user preference

Results can be exported to a PDF report

Individual analytics produce log-likelihood ratio (LLR) scores



DARPA SemaFor Finds a New Home

- Underwriter Laboratories (UL) Digital Safety Research Institute (DSRI) has taken stewardship of the SemaFor community following the conclusion of the research period of performance.
- UL/DSRI is standing up academic competitions to encourage continued research into this domain, with top performers receiving prizes such as cash grants to continue research and funding to publish and present at major conferences.
- It is also standing up a Marketplace where commercial entities can browse the competitions and make connections with developers willing to consider commercial partnerships.

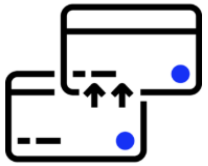
Working for a Safer World

About Us

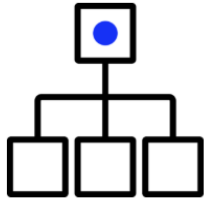
Digital Safety Research Institute

DSRI aims to better protect the public from rapidly evolving AI/digital threats

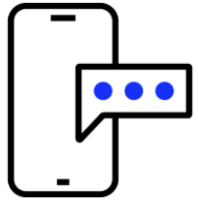
Cybersecurity threats



Privacy threats

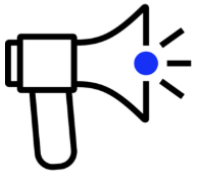


Algorithmic threats



Information threats

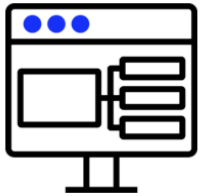
By creating a new AI/digital safety ecosystem



Incident Reporting

Test Generation

Independent Test



Independent Inspection

Keeping Pace with Rapid Advances in Generative Artificial Intelligence

PUBLISHED

MARCH 26, 2025

TAGS

ARTIFICIAL INTELLIGENCE (AI)

DIGITAL SAFETY

DIGITAL THREATS

PARTNERSHIPS

SHARE

SemaFor Deep-Fake Detection Ecosystem

Deep Fake Detectors

- Audio deep fake detectors (w/ IH&MMSec)
- Cheap-fake detectors (w/ DEFCON)
- Video deep fake detectors (w/ ICCV 2024)

ULRI is providing research grants to continue development of promising detectors

Independent Test

- Measure performance of detectors
- Measure performance of analyst-detector teams

Test Generation

- Test data generation
- Performance measures
- Automated evaluation report generation

DSRI is providing evaluation reports to

- *Deep-fake detector developers*
- *Media forensics platform developers*

Media Forensics Applications

- For Government
- For Law Enforcement
- For Journalists

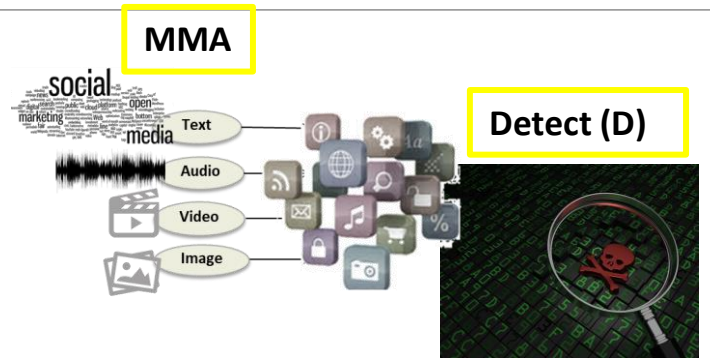
D/A/C

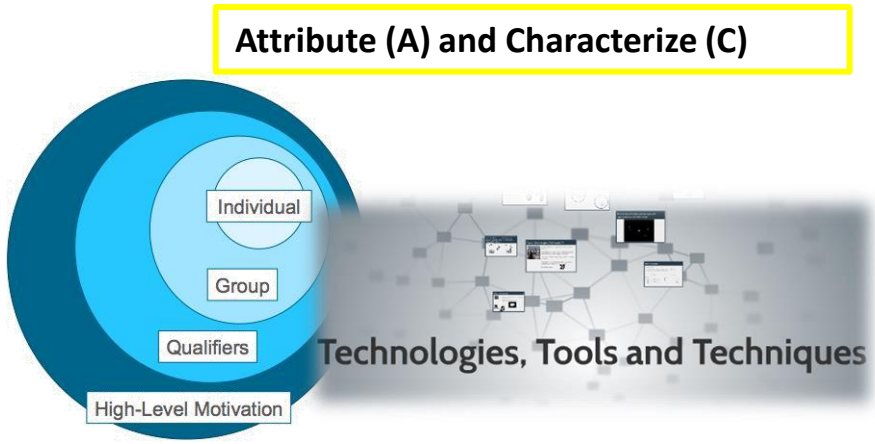
GOING BEYOND DETECTION

Key Concepts: Detection, Attribution, and Characterization*

** Collectively abbreviated as D/A/C in this presentation*

- **Multimodal Media Asset (MMA):** Media may contain multiple modalities such as image, video, audio, and text (e.g., a news story).
- **Detection (D)** – Finding inconsistencies or indications of machine generation or manipulation as evidence the media may be falsified.
- **Attribution (A)** – Attributing multimodal products to a purported source (e.g., author/individual, organization, or tool).
- **Characterization (C)** – Determining evidence for tools, techniques, and the intent behind a detected falsification (e.g., malicious intent to significantly alter tone).

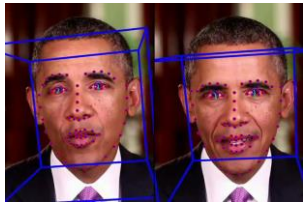
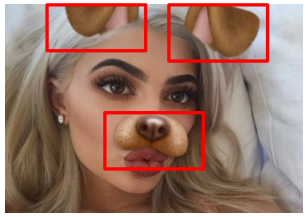




Detection is *useful*, but moving beyond detection is critical for the results to be *actionable*

Going Beyond Simple Detection

Media manipulation **doesn't always indicate a disinformation attack / malicious intent**. Semantic information embedded in the multi-modal media assets can help to detect, attribute, and characterize the falsification.

	Detect	Attribute	Characterize
	 Deepfake DETECTED	**Generator:** FakeApp **Source:** Buzzfeed	**Tactic:** Puppeteering **Intent:** PSA To Inform
	 Manipulation DETECTED	**Generator:** Snapchat Filter **Source:** Kylie Jenner IG	**Intent:** To Entertain

System: Classes of Analytic Workflows

(not comprehensive)

Video

- Detection (Y/N)
- POI (e.g., Putin, Zelensky)
- Puppeteering and Face-swap
- Generator Attribution (e.g., Wave2Lip, DeepFaceLab)

Audio (synthetic)

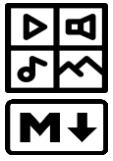
- Detection (Y/N)
- D/A/C POI (e.g., Craig Kelly, Ng Eng Hen)
- Generator Attribution (Realtime-Voice-Cloning, Tacotron-Voice-Cloning)

Text (synthetic)

- Detection of generated article (with and w/o edits) (Y/N)
- Text Generator Attribution (Grover, GPT2, GPT-j-6B)
- Attribution to purported source (e.g., Organization, Author)

Images (synthetic and manipulated)

- Detection/Attribution/Characterization



Multi-Modal Assets (MMA)

- Detect inconsistencies in article (Headline, Image, and Body)
- Text and Image, Text and Audio, Text and Video
- Characterize Inconsistencies (e.g., Topic: COVID; Minimize, Appeal To Fear)

Scope of Analysis

- Intents
- Tactics
- Sematic Labels
- Tools and Techniques
- POI
- MMA Portions
- Topics

Part 2

ALGORITHMIC APPROACHES AND DATA CHALLENGES

SemaFor in practice

- Before we get too far into details around what makes an effective analytic, we will begin with a showcase of the DARPA SemaFor platform.
- This is both to (hopefully) inspire, but also to showcase a number of key concepts we will be discussing to highlight their value.

- **After this walkthrough there will be a short period of "free play"** where anyone with an internet connection can try the tool out for themselves and browse the analytic catalog to see what capabilities exist.
- You will be free to browse and use this instance for the duration of the conference, as long as access is not abused.
- At the conclusion of this tutorial, I will save time for questions and to solicit feedback on the interface – what you liked, what you wish it did, what was confusing, etc.

SemaFor in practice

https://dmz-proj-semafor.atl.external.lmco.com/cogsima2025/portal/login

Username: Choose any account between Student1 and Student65

Password: **CogSIMA.2025!**

Sample media can be found here:

https://github.com/mikewkozak/cogsima2025

NOTE: This instance of SemaFor will be deleted concluding the conference and all accounts reset.

SemaFor in practice

- Please limit your uploads to 1 piece of media at a time – we are on shared resources and too many requests at once will slow analysis time for everyone.

- Please do not upload any sensitive or graphic media – all users will have shared access to viewing uploads.

SemaFor Demo

SemaFor
SEMANTIC FORENSICS

SemaFor

SEMANTIC FORENSICS

Welcome to Semafor Portal

Login

or

Don't have an account?

Request Portal Account

SemaFor
production

- Collections
- Media Gallery
- Components
- Settings
- Help
- Admin
- Notifications (1)
- New Collection

Uploading media artifacts to create a new collection.

Collection Details

Collection Name*
Oct 10, 2024 - 4:36 PM MST

Tags
Enter Tags

Description
Enter Description

Use collection as training data

Select Media

- **From Upload** (audio, video, images, and .txt files are supported)
  Limit 20 files, individual files cannot exceed 200MB
- **From Article URL** (direct links to audio, video, or images are not supported)
  Article URL must be from a supported domain
- **From Text** (paste copied text into the textarea)

From Upload

Drag and drop files to upload

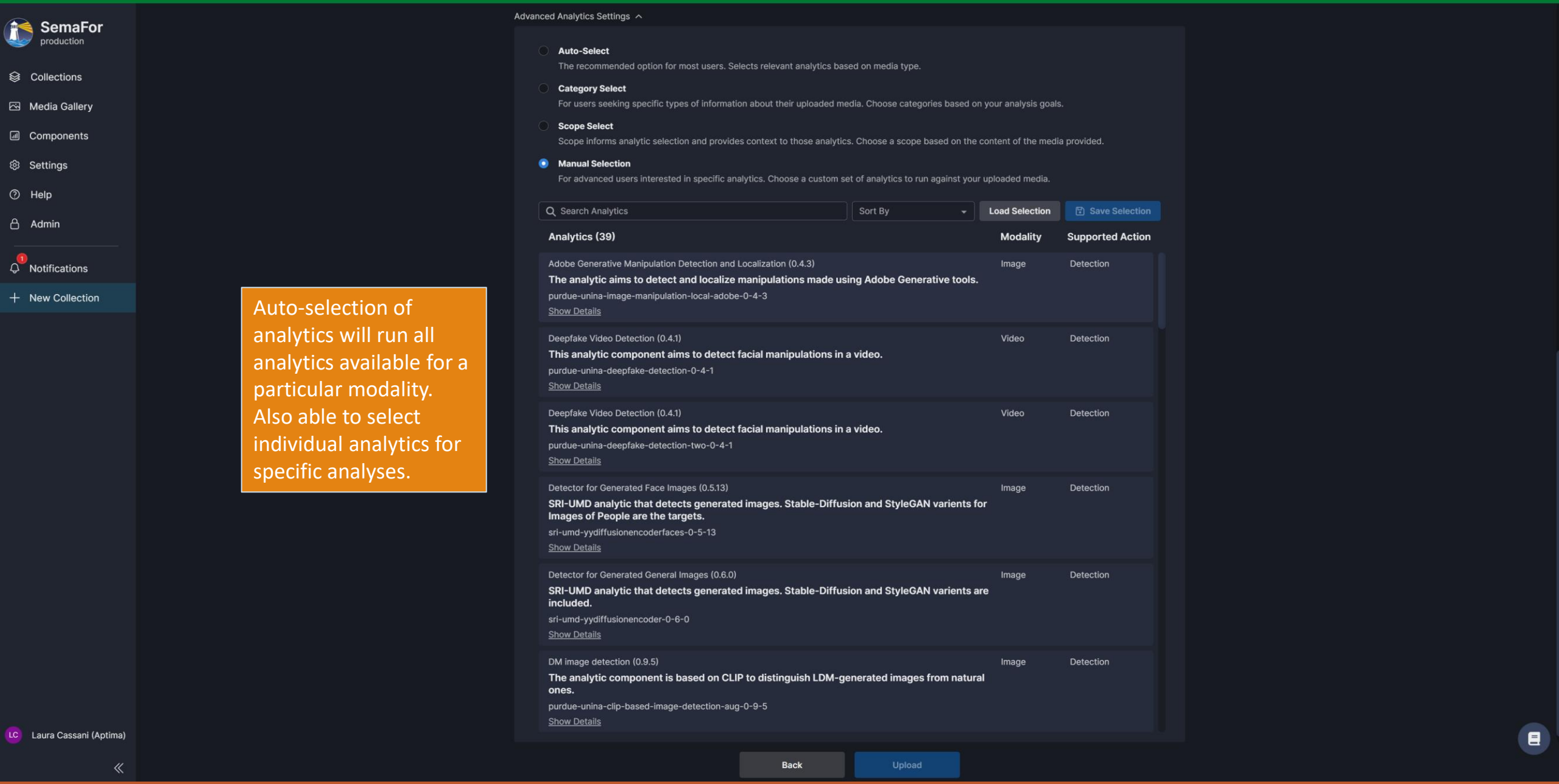
Browse Files

Limit 20 files, individual files cannot exceed 200MB

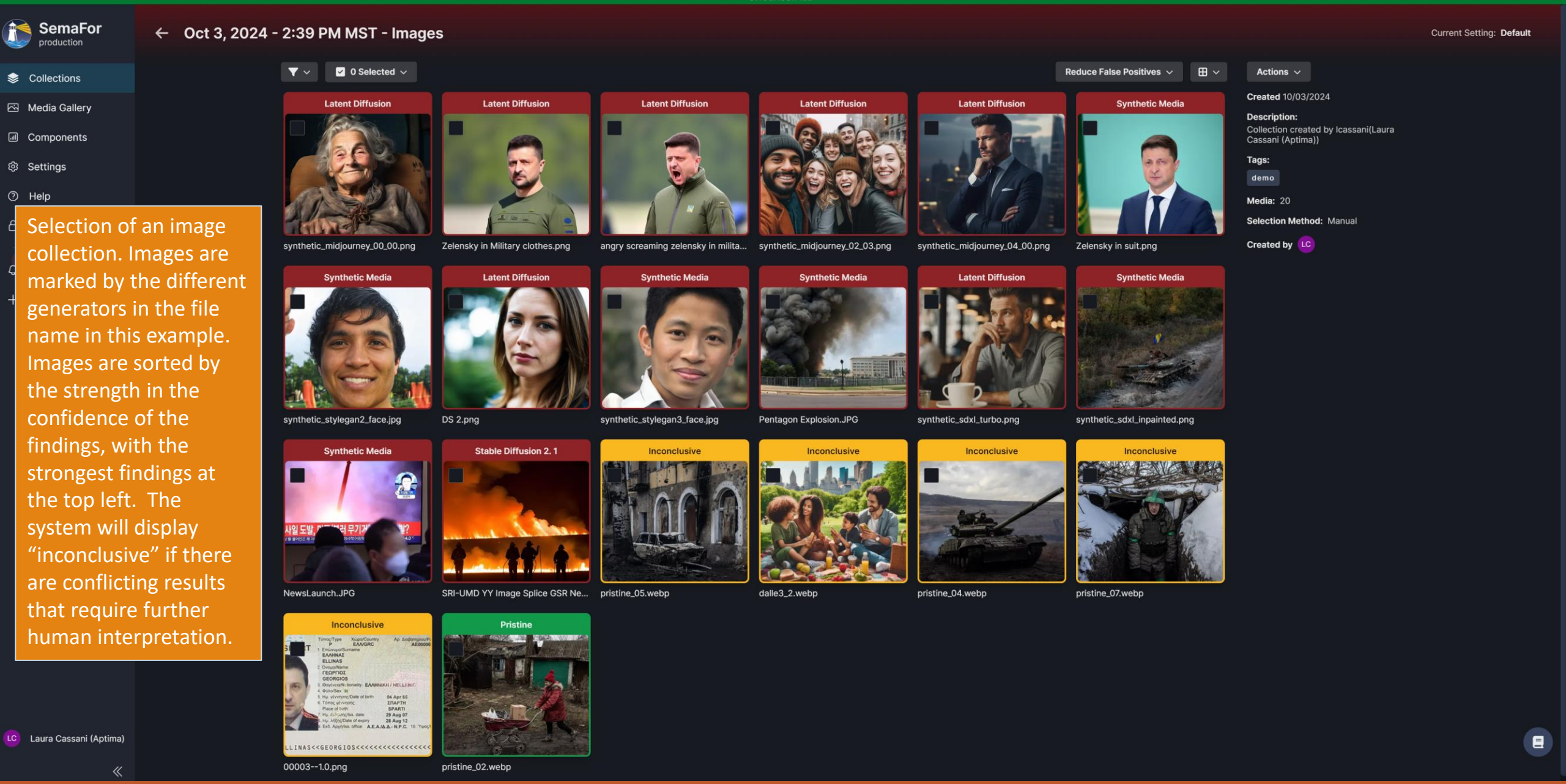
Advanced Analytics Settings

Back | Upload

UNCLASSIFIED



Auto-selection of analytics will run all analytics available for a particular modality. Also able to select individual analytics for specific analyses.



Selection of an image collection. Images are marked by the different generators in the file name in this example. Images are sorted by the strength in the confidence of the findings, with the strongest findings at the top left. The system will display “inconclusive” if there are conflicting results that require further human interpretation.

Results for synthetic_midjourney_02_03.png

Reduce False Positives | Generate Reports

Latent Diffusion

Selection on individual findings brings up information about the analytic and how a user can interpret the results.

Overview | Detailed Results | Comments (0)

Advanced View

Score Legend

Techniques | All Results (3)

Search Analytics | Filter by | Sort by

Synthetic image detection (0.10.1) — Latent Diffusion
purdue-unina-synthetic-image-detection
The analytic component distinguishes synthetic images from natural ones.
Detection
Score: 3.22

The analytic component distinguishes between real and synthetic images. The analytic was trained on images created using these image generation techniques: ProGAN/StyleGAN2, GigaGAN, Latent Diffusion. It uses deep Convolutional Neural Networks without downsampling in the first layer trained on several datasets. During training proper augmentation is applied.

Category
ConsistencyWithGenerationModel

Instance
Latent-diffusion

Model Name
dm_detection_one

Model Type
generationModel

Evidence Type
ToolsTechniques

Model Version
0.0.1

Model Algorithm
resnet50_nodown

Training Datasets
real/LSUN, real/ImageNet, real/COCO, real/FFHQ, Latent-diffusion/LSUN, Latent-diffusion/ImageNet, Latent-diffusion/COCO, Latent-diffusion/FFHQ

Synthetic image detection (0.10.1) — Latent Diffusion
purdue-unina-synthetic-image-detection-two
The analytic component distinguishes synthetic images from natural ones.
Detection
Score: 3.22

Kitware Synthetic Image Detection (0.5.5) — Stable Diffusion 2.1
kitware-synthetic-image-detection
Detection of synthetic imagery.
Detection
Score: 1.63

Laura Cassani (Aptima)

Results for synthetic_sdxl_inpaint_2.png

Reduce False Positives | Generate Reports

Image Splice

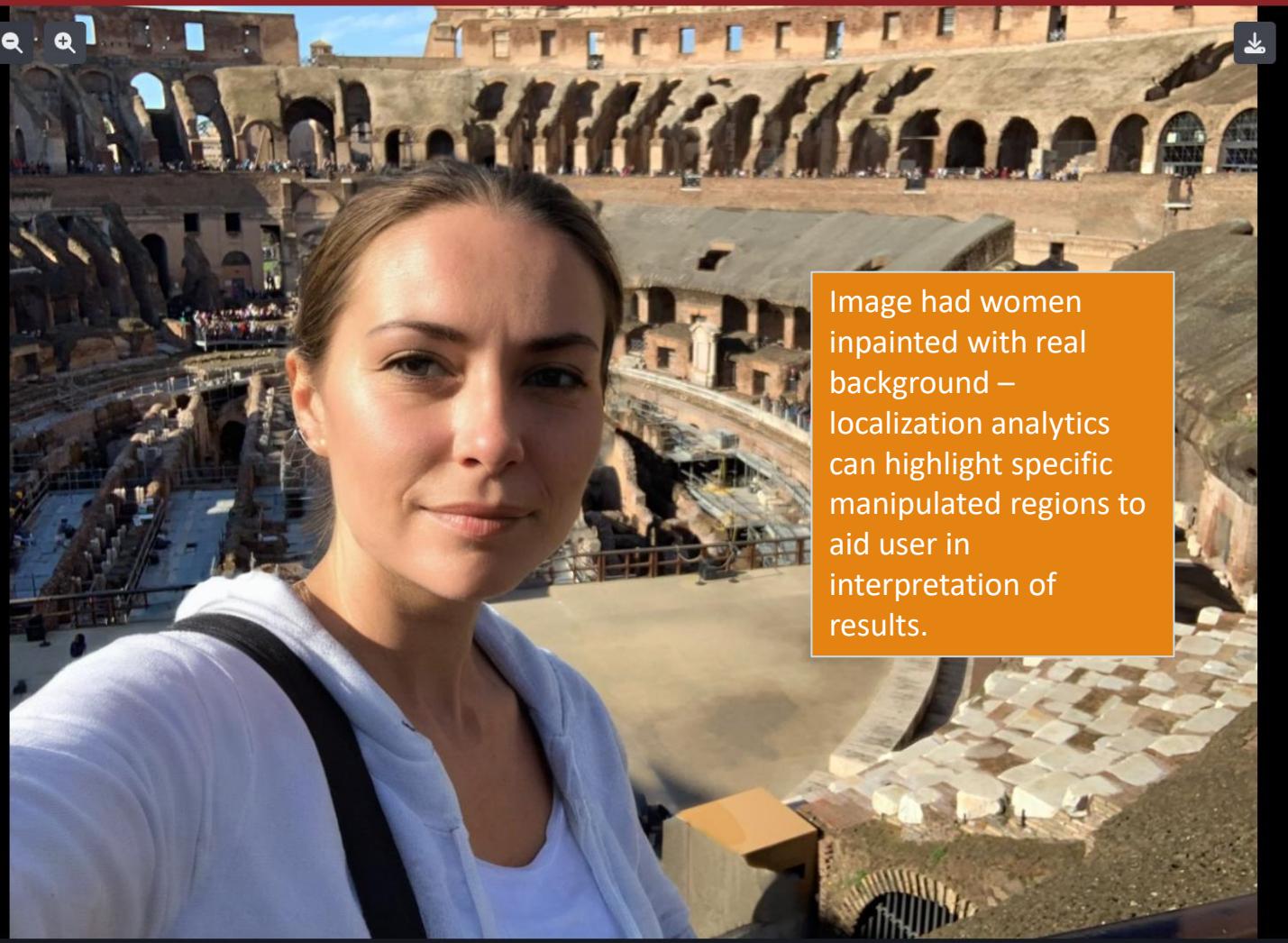


Image had women inpainted with real background – localization analytics can highlight specific manipulated regions to aid user in interpretation of results.

Overview | Detailed Results | Comments (0)

Advanced View

Score Legend

Techniques | All Results (6)

Search Analytics

Filter by | Sort by

Image Manipulation Localization and Detection (0.4.3) — Image Splice

purdue-unina-image-manipulation-localization

The analytic component aims to localize manipulations by looking at local inconsistencies of the source. Detection

Score: 3.91

Adobe Generative Manipulation Detection and Localization (0.4.3) — Image Splice

purdue-unina-image-manipulation-local-adobe

The analytic aims to detect and localize manipulations made using Adobe Generative tools. Detection

Score: 3.6

Image Manipulation Localization and Detection (0.0.0) — Image Splice

purdue-unina-image-manipulation-local-map-0-4

The analytic aims to detect and localize manipulations providing a localization map. Detection

Score: 2.28

Image Manipulation Localization and Detection (0.0.0) — Image Splice

Showing 1 - 7 of 7 uploads

Results for synthetic_sdxl_inpainted.png

Reduce False Positives | Generate Reports

Evidence for Image Splice

Another localization image with manipulated region highlighted.

Localization Opacity

Overview | Detailed Results | Comments (0)

Advanced View

Score Legend

Techniques | All Results (6)

Search Analytics

Filter by | Sort by

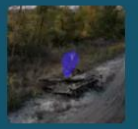


Image Manipulation Localization and Detection (0.4.3) — Image Splice

purdue-unina-image-manipulation-localization

The analytic component aims to localize manipulations by looking at local inconsistencies of the source. — Detection

Score: 5

The analytic component aims to localize manipulations by looking at local inconsistencies of the source. A deep learning based architecture is trained using both the RGB features and the noiseprint component to find these inconsistencies and provide a localization mask of the manipulated area. Cross-entropy loss is used during training on a set of pristine and manipulated images. The provided polygon is obtained by a convex hull of the localized area.

Category
ConsistencyWithManipulationModel

Instance
ImageSplice

Model Name
trufor_paper

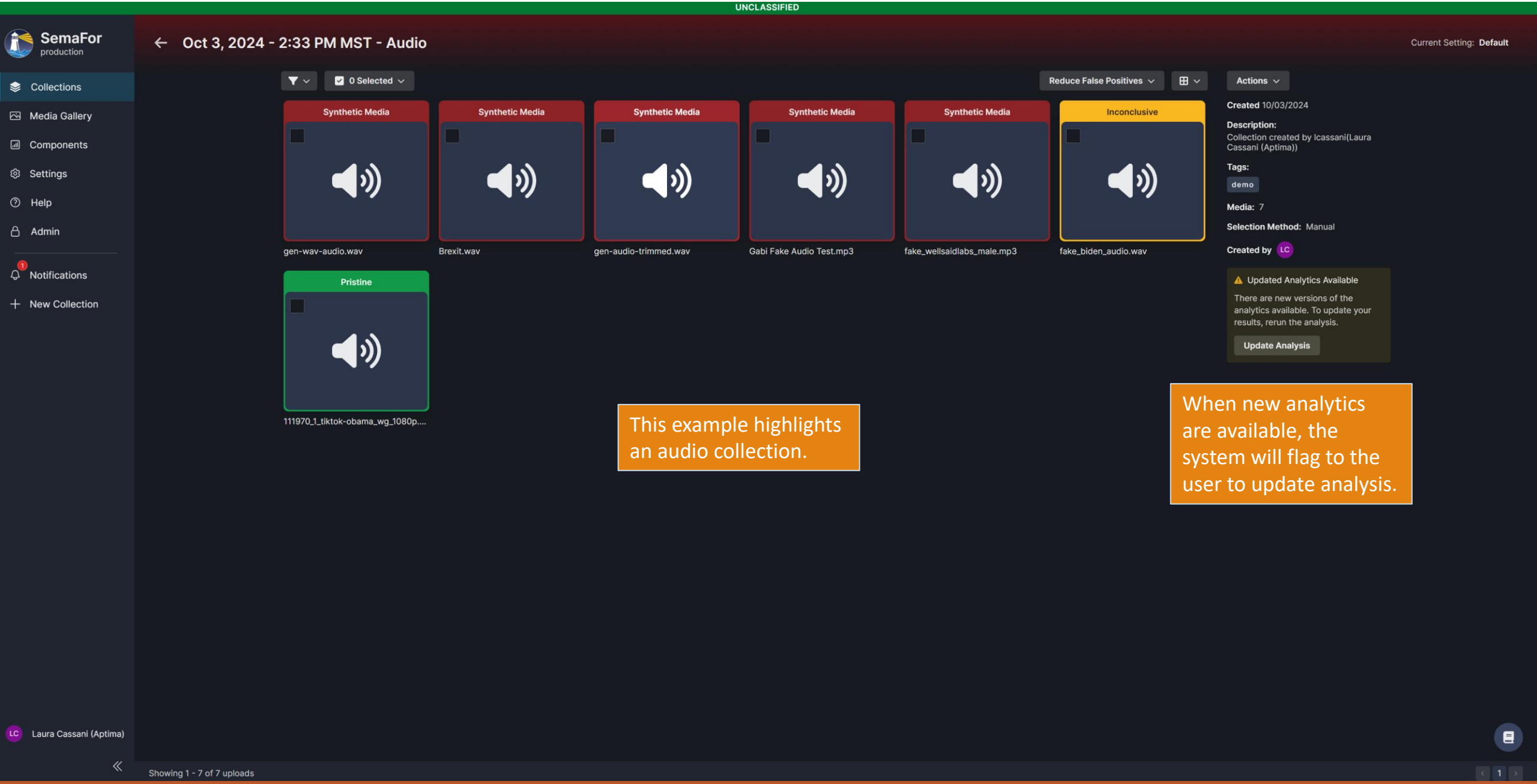
Model Type
manipulationModel

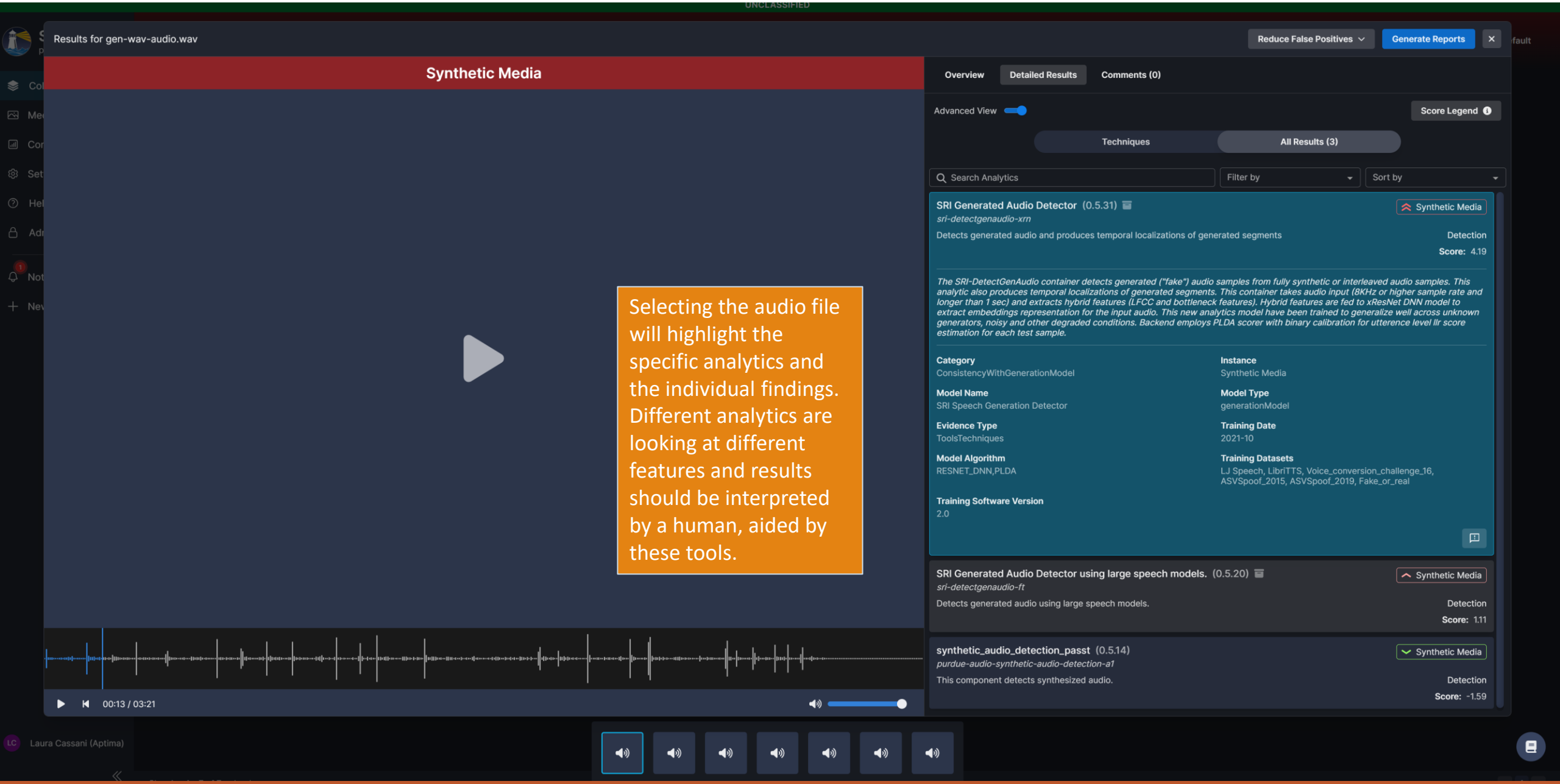
Evidence Type

Laura Cassani (Aptima)

Showing 1 - 7 of 7 uploads

UNCLASSIFIED




UNCLASSIFIED
Results for gen-wav-audio.wav
Reduce False Positives
Generate Reports
Synthetic Media
Overview
Detailed Results
Comments (0)
Advanced View
Score Legend
Techniques
All Results (3)
Search Analytics
Filter by
Sort by
SRI Generated Audio Detector (0.5.31)
sri-detectgenaudio-xrn
Synthetic Media
Detects generated audio and produces temporal localizations of generated segments
Detection
Score: 4.19
The SRI-DetectGenAudio container detects generated ("fake") audio samples from fully synthetic or interleaved audio samples. This analytic also produces temporal localizations of generated segments. This container takes audio input (8KHz or higher sample rate and longer than 1 sec) and extracts hybrid features (LFCC and bottleneck features). Hybrid features are fed to xResNet DNN model to extract embeddings representation for the input audio. This new analytics model have been trained to generalize well across unknown generators, noisy and other degraded conditions. Backend employs PLDA scorer with binary calibration for utterence level llr score estimation for each test sample.
Category
ConsistencyWithGenerationModel
Instance
Synthetic Media
Model Name
SRI Speech Generation Detector
Model Type
generationModel
Evidence Type
ToolsTechniques
Training Date
2021-10
Model Algorithm
RESNET_DNN,PLDA
Training Datasets
LJ Speech, LibriTTS, Voice_conversion_challenge_16, ASVSpoof_2015, ASVSpoof_2019, Fake_or_real
Training Software Version
2.0
SRI Generated Audio Detector using large speech models. (0.5.20)
sri-detectgenaudio-ft
Synthetic Media
Detects generated audio using large speech models.
Detection
Score: 1.11
synthetic_audio_detection_passt (0.5.14)
purdue-audio-synthetic-audio-detection-a1
Synthetic Media
This component detects synthesized audio.
Detection
Score: -1.59
Selecting the audio file will highlight the specific analytics and the individual findings. Different analytics are looking at different features and results should be interpreted by a human, aided by these tools.
00:13 / 03:21
Laura Cassani (Aptima)

Results for faceswap_picsiai_colinjost.mp4

Reduce False Positives | Generate Reports | ×

Deepfake

00:10 / 00:10

Overview | Detailed Results | Comments (0)

Advanced View | Score Legend

Techniques | All Results (9)

Search Analytics

Filter by | Sort by

Face DeepFake detector-TemporalConvolution (0.5.4) — Deepfake
sri-umd-yydeepfaketc
SRI-UMD analytic that detects DeepFakes of faces based on temporal artifacts — Detection
Score: 3.99

Face DeepFake detector-TemporalConvolution Transformer (4.0.0) — Deepfake
sri-umd-yydeepfaketctransformer
SRI-UMD analytic that detects DeepFakes — Detection
Score: 1.22

Face DeepFake detector-TemporalConvolution Transformer (0.5.9) — Deepfake
sri-umd-yydeepfaketctransformer
SRI-UMD analytic that detects DeepFakes — Detection
Score: 1.17

synthetic_audio_detection_passt (0.5.14) — Synthetic Media
purdue-audio-synthetic-audio-detection-a1
This component detects synthesized audio. — Detection
Score: -1.47

Laura Cassani (Aptima)

SemaFor
production

- Collections
- Media Gallery
- Components
- Settings
- Help
- Admin
- Notifications
- New Collection

Component view provides the library of analytics available for users to access for further analysis.

Laura Cassani (Aptima)

Keyword: Enter Keyword

Modality: Select Modality

Adobe Generative Manipulation Detection and Localization (0.4.3)
purdue-unina-image-manipulation-local-adobe
The analytic aims to detect and localize manipulations made using Adobe Generative tools.
Ready: Yes | Selected by Default: No
Supported Modalities: Image | Supported Actions: Detection

BFC Relevance Score Fusion (0.5.17)
lm-bfc-fusion-feature-combination-relevance
Fuse DAC scores in evidence graphs generated from the same artifact graph
Ready: Yes | Selected by Default: Yes
Supported Modalities: Audio, Image, Text, Video | Supported Actions: Fusion

ake Video Detection (0.4.1)
-unina-deepfake-detection
alytic component aims to detect facial manipulations in a
Selected by Default: No
Supported Modalities: Video | Supported Actions: Detection

Deepfake Video Detection (0.4.1)
purdue-unina-deepfake-detection-two
This analytic component aims to detect facial manipulations in a video.
Ready | Selected by Default

Sync Components | Export | Import

Adobe Generative Manipulation Detection and Localization

Current Setting: **Default**

Metrics | Gallery | Benchmark Data | Gym Evaluations

This analytic is a transformer-based neural network which aims to detect and localize manipulations in images using both the RGB features and the noiseprint component. The analytic provides a localization map and a global detection score for each image. The global detection score is obtained by a combination of the localization map and an estimated confidence map. The network is trained on images manipulated using Adobe Generative tools.

Conclusions (Latest Version)

- Low number of scored results, evaluation is incomplete
- Average runtime is less than 10 minutes
- Component generally runs successfully
- Component has an executive summary
- Component has listed capabilities
- Component has a summary
- Component has a description
- Component must be manually selected (selectByDefault == false)

Executive Summary | Gym Evaluation | GitLab | YAML | Edit

Latest Version

Score Distribution

28 Total Results	4 Max Score	57s Max Runtime
28 Scored	1 Average Score	6s Average Runtime

Technical Considerations

THE SECRETS TO CREATING A GOOD ANALYTIC

Algorithm vs Analytic – A Use Case

Scenario: A real-time audio deepfake cons an energy company out of $243,000 (~€216,000)

Scenario: A real-time video conference full of deepfaked staff fooled the audio detectors and convinced the finance department to wire $25 million (~€22 million) to a fraudulent account.

Scenario: You find that over a dozen fraudulent bank accounts have been opened by one person capable of tricking facial recognition technology with your stolen ID card.

Scenario: An analytic designed to identify real versus synthetic people falsely classifies a real person as synthetic. A missing person's case is never filed.

In each case, here is what is provided:

Algorithm	Analytic
A black box prediction with an associated confidence score, perhaps a confusion matrix of test results	A series of localized, evidence-based claims each individually scored based on the level of confidence that combine into a broader level analysis of at the semantic level, further backed by explicit details about the model itself that can assist in evaluating whether the source media was a good fit for this analytic.

https://www.trendmicro.com/vinfo/us/security/news/cyber-attacks/unusual-ceo-fraud-via-deepfake-audio-steals-us-243-000-from-u-k-company
https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk

Imagine being a judge, or a jury, or a committee accepting this as evidence….

Explainability – The SemaFor Approach

Every Analytic produces an "Evidence Graph" that decomposes the output as such:

1. **Analytic Subclass** – Is this a D / A / or C claim?
    1. You can declare multiple Subclasses in an Evidence Graph if your analytic is capable of doing so.
    2. Provide a score representing the combined series of subclaims in this branch
2. **Analysis Subclass(s)** – Is this a Semantic Consistency Check or not?
    1. Provide a score representing only the confidence of this specific claim
3. **Concept Node(s)** – What is the hypothesis your analytic is declaring?
    1. Provides both a Type and an Instance
    2. You can provide multiple hypotheses as their own branches under Analysis Subclass
4. **Model Node** – How do you search for evidence?
    1. Information such as training date, training datasets, underlying algorithm, and additional implementation details
5. **Evidence Node(s)** – Localization and Evidence
    1. Localization Nodes provide bounding boxes and time stamp ranges where the strongest evidence in favor of your hypothesis can be found
    2. You can additionally produce supporting material such as heatmaps, video overlays, graphs, and provenance graphs
6. **Reference Node(s)** – Provide traceability back to the source
    1. Via UUID links this claim back to the analyzed pieces of media within a multimodal asset

Evidence

Evidence can take many forms, for example:

- **Bounding boxes or time stamp** ranges showing where and when the strongest evidence is
- **Heatmaps** showing the relative intensity of manipulation likelihood
- **Provenance graphs** showing the likelihood of each known generator being used
- **Frame-overlays** such as markers indicating which features are being tracked on a face
- **Cross-modality localizations** showing semantic inconsistencies in different media

One challenge with producing evidence is knowing *when* to produce evidence…

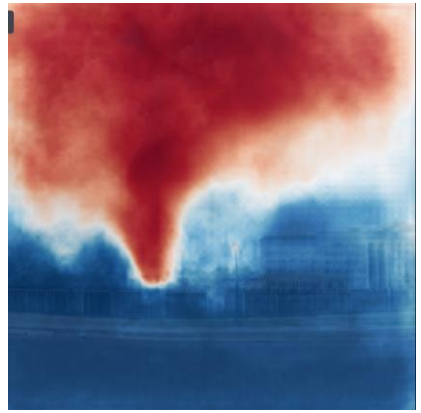
- A splicing analytic producing a heatmap on a GAN will produce noise
- A localization analytic given a fully-synthetic image has nowhere to localize

Evidence – Bounding Boxes vs Heatmaps

The right type of evidence will depend on the type of claim being made and the difficulty in localizing

Bounding Boxes are great for splices and inserts where discrete regions have been changed

Heatmaps are great for soft edges or transformations where many regions have been effected

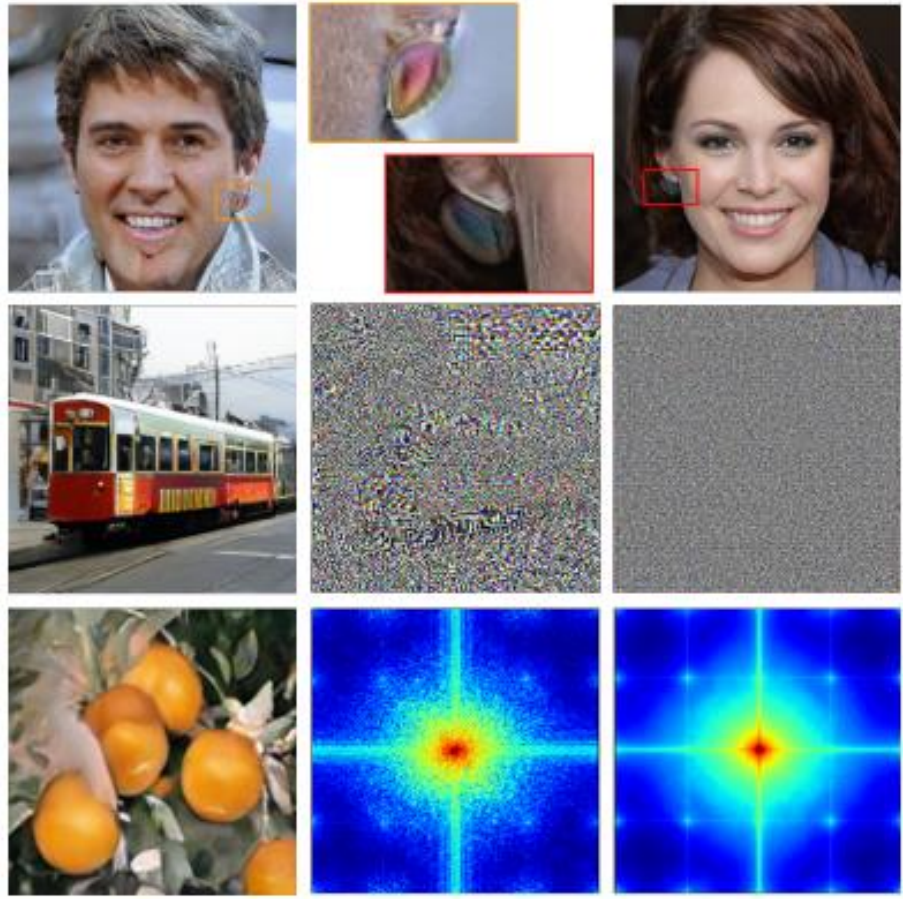


Tips and Tricks

THE SECRETS TO CREATING A GOOD ANALYTIC

GAN attribution via fingerprinting

- While visual anomalies in GAN (top row) have gotten less frequent as generators have gotten more advanced, their artificial fingerprint (middle) and its Fourier spectrum analysis (bottom) remain.
- Training up a model to correctly identify which type of GAN was used can provide powerful evidence of manipulation while also creating new data to reason over.
- You can also train up a CNN on the Fourier spectra of both real and GAN images to produce a reasonable detector using common libraries like FFT to extract them.
- This approach performs really well but is extremely brittle to image laundering (resizing, format change, etc) which may damage or destroy these traces.
- However, fingerprinting still has its uses…



https://arxiv.org/pdf/2104.02617

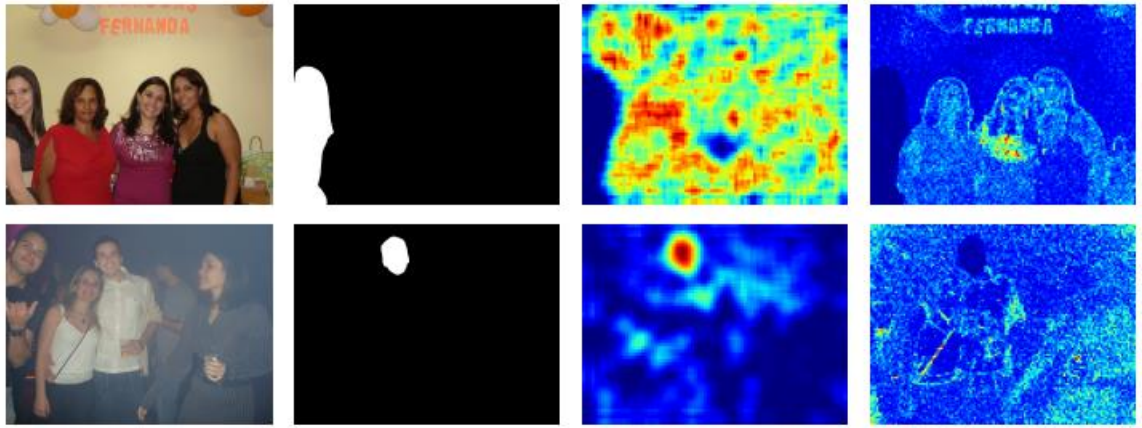
Denoising to find provenance

- Although invisible to the naked eye, everything from the camera lens to the color filter to the JPEG compression process leaves traces in the image.
  - Because of this, even individual camera models can have their own unique noiseprint signature
- These traces can be extracted by suppressing all other aspects of the image through denoising.
- To do this, you can train a CNN with a pairwise dataset consisting of patches of images that are tested against ground truthed patches of noise until the CNN-estimated noise is close or matches to the real thing.
- To avoid needing to have the ground truth known, we can take advantage of the fact that camera models produce similar noise patterns to cluster our training samples…

'Noiseprint: a CNN-based camera model fingerprint (TIFS2020), https://arxiv.org/abs/1808.08396

Denoising to detect manipulations

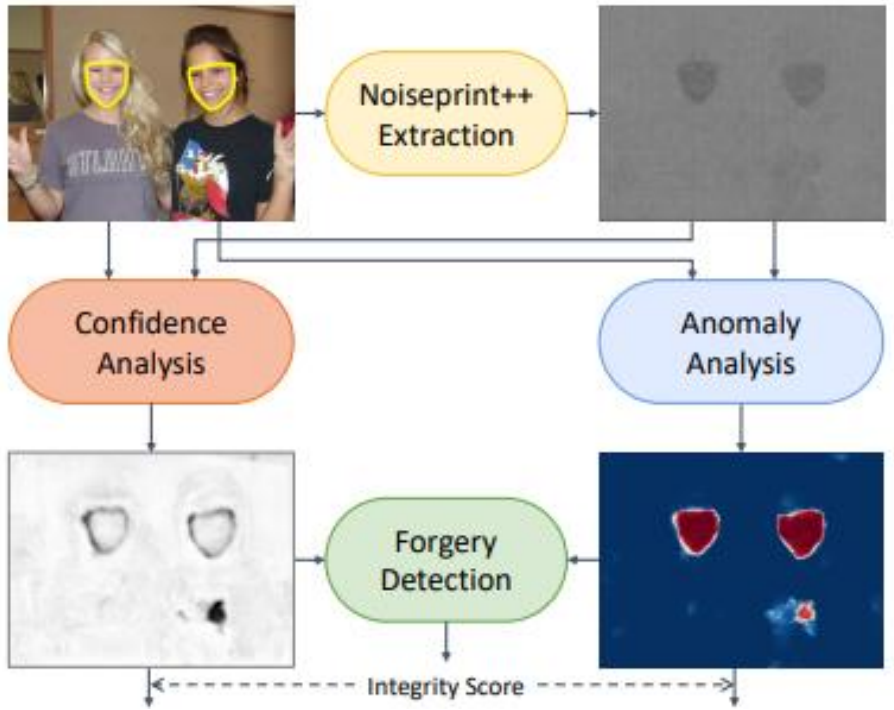
- Knowing that the camera fingerprints should be similar across different images, we *could* train up 2 denoisers at the same time – each acting as "ground truth" for the other since, given images from the same camera – they should be similar.
  - In practice, a single CNN can be used by feeding in pairs of inputs and a flag that marks them as either from the same source or not.
- The noiseprint outputs can then be converted into a heatmap of the image showing the likelihood of manipulation as a byproduct of gaps/changes in the noiseprint



'Noiseprint: a CNN-based camera model fingerprint (TIFS2020), https://arxiv.org/abs/1808.08396

Combining Denoising and Detection

By combining a detection analytic and the denoising information, you can build a system that can detect AND localize in a generalizable way



TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization (CVPR2023), https://arxiv.org/abs/2212.10957

Denoising to find manipulations

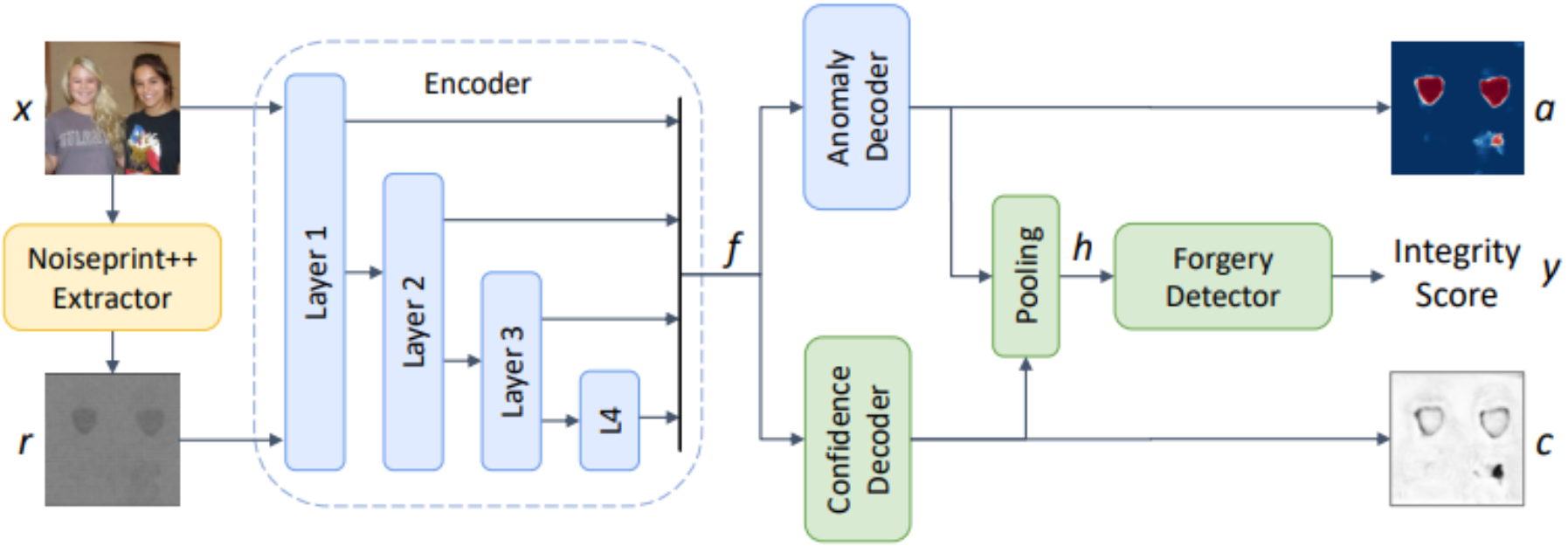
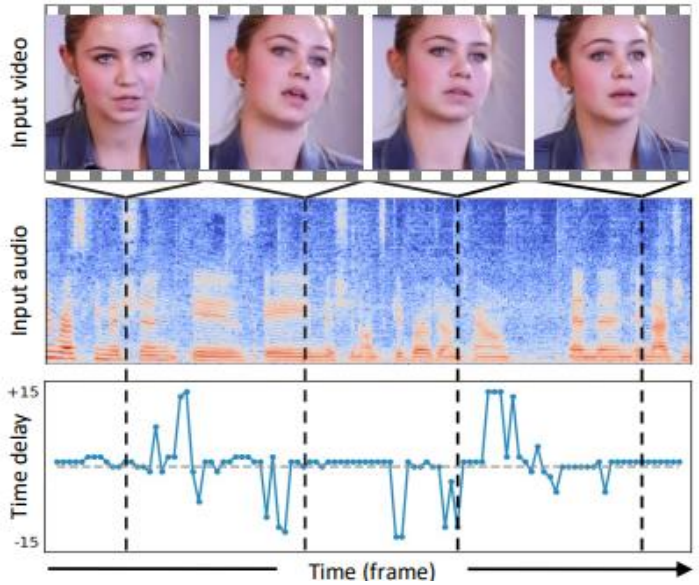


Figure 2. TruFor framework. The Noiseprint++ extractor takes the RGB image to obtain a learned noise-sensitive fingerprint. The encoder uses both the RGB input and Noiseprint++ for jointly computing the features that will be used by the anomaly decoder and the confidence decoder for pixel-level forgery localization and confidence estimation, respectively. The forgery detector exploits the localization map and the confidence map to make the image-level decision. The different colors identify the modules learned in each of the three training phases.

TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization (CVPR2023), https://arxiv.org/abs/2212.10957

Anomaly Detection of Synthetic Video

- Anecdotally, generated audio detectors are by far the most accurate and generalizable across the major media modalities (image, video, audio, text).
- Thy are so accurate, in fact, that they often outperform generated video detectors if there is synthetic or spliced audio present.
- By analyzing both the video and audio stream together, you can train up a detector *without ever needing any synthetic data* by instead training the detector to identify anomalies in cross-modality features like temporal delays or expected sound for a given mouth movement.
- Once that works on a single frame or batch of frames, you can crawl the entire video segment and either localize to high anomaly scores, combine/average the individual scores to rate the entire video, or both.



https://arxiv.org/pdf/2301.01767

Global and Local Fusion for Robustness

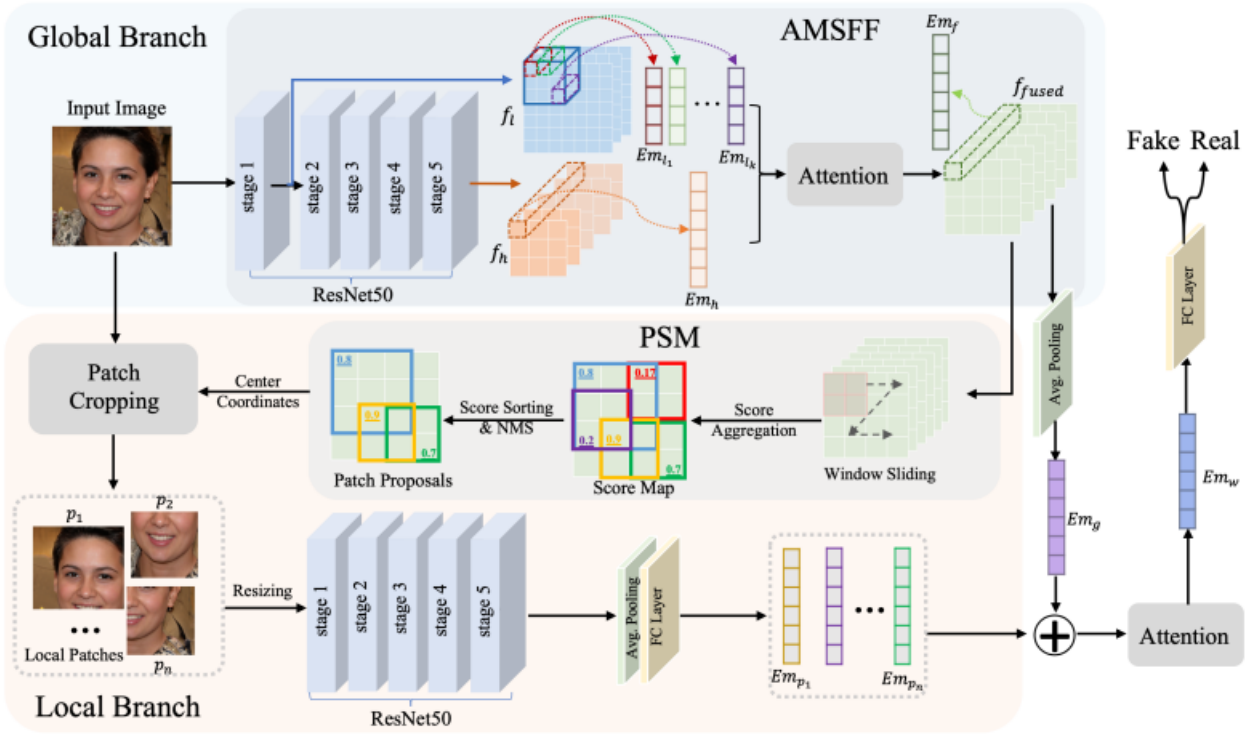
The **most successful and robust methods**, much like in the field of generative AI itself, **combine multiple approaches**.

- Generator attribution analytics often contain a collection of models trained against different generators and reason over the combined predictions to choose the most likely candidate
- This same approach can be done for detection – **combine local feature extraction with whole-image analysis** to improve the odds of finding evidence of manipulation even in unseen classes of images and generators.

https://arxiv.org/pdf/2211.08615

Global and Local Fusion for Robustness

- The Global and Local Feature Fusion (GLFF) Algorithm does exactly that, using the global analysis to help determine where local analysis is required.
- Using attention to guide fusion also avoids the “washing out” issue of local manipulations being “averaged away” from larger images in analysis.



https://arxiv.org/pdf/2211.08615

Training Data

Common Datasets

- Unless you are writing an analytic that only does attribution, odds are you will need real media as part of your training dataset.
- Thanks to decades of image recognition research, there are a quite a few large, labeled datasets to choose from with various pros and cons.
- Unfortunately, many of these datasets as you will see in a moment either come with significant resolution or file format limitations, or they do not contain labeled sets of images that fit your use case.
- Inevitably, you may be forced to create your own hybrid datasets.

Dataset list

- **Audio Training Datasets**
  - ASVS Spoof 2015
  - ASVS Spoof 2019
  - Voice Conversion Challenge 2016
  - LJ Speech
  - LibriTTS
  - AudioSet
  - Common Voice
  - Speech Commands
  - ESC-50

- **Image Training Datasets**
  - CIFAR-10 / 100
  - ImageNet
  - MS COCO
  - Flickr 30K
  - IMDB-Wiki
  - Berkeley Deep Drive
  - LSUN

Dataset list

- **Video Training Datasets**
  - UCF101
  - Kinetics
  - HMDB51
  - ActivityNet
  - Human 3.6M
  - DAVIS
- **Text Training Datasets**
  - Stanford Sentiment Treebank
  - MultiNLI
  - QNLI
  - HumanEval
  - WikiText

Key Considerations - Resolution

Synthetic content designed for social media tends to vary from platform to platform, although profile and avatar photos share a much more narrow range

Images in a Twitter Post

Recommended sizes for images by types of Posts:

- **Recommended image sizes:**
  - **Minimum**: 600×335 pixels
  - **Landscape**: 1024×512 pixels (minimum) and 1600×900 pixels (recommended)
  - **Square**: 1080×1080 pixels
  - **Portrait**: 1080×1350 pixels
- **Recommended aspect ratio:** 16:9
- **Maximum file size:**
  - 5 MB for JPG, PNG
  - 15 MB for GIFs
- **Maximum number of images per post:** 4 images

Pinterest Pin sizes image guidelines

- **Recommended image size:** 735×1102 pixels
- **Recommended ratio:** 2:3
- **Recommended file size:** 20MB max
- **Recommended file type:** JPG, PNG, GIF
- **Note**: Larger are resized to display at a width of 238 pixels with scaled height in feeds.

Profile Photos *average around 300x300 to 400x400 resolution.*

https://sproutsocial.com/insights/social-media-image-sizes-guide/

Key Considerations - Resolution

ImageNet – containing 14 million hand-annotated images with thousands of classes

- Images can range in size from tens of pixels to thousands with an average of **469x387 pixels**
- Images are all **JPEG** files of varying compression levels

COCO – containing 330,000 images across 171 categories.

- Images are all **640x480**
- Primarily **JPEG *and* PNG**
  - PNG adds the complexity of having to detect or support both 3 and 4 channel variants

LSUN – Large-Scale scene UNderstanding contains around 60 million images across 10 scene / 20 object classes

- Images in testing datasets tend to be **128 x 128 or 256 x 256**
- Images are all **JPEG** files

https://towardsdatascience.com/compression-in-the-imagenet-dataset-34c56d14d463/

What about the Synthetic Images?

- Most well established, large scale image datasets available for academic research only contain real images.
- They are also likely a far smaller resolution than what generators can produce (DALL·E 3 was trained to generate 1024x1024, 1024x1792 or 1792x1024 images)
- This means you need to augment your training dataset with images, ideally generated directly from the generator(s) you want to be able to detect.
- Due to the rapid nature of advances in the field and the protective nature of many organizations when it comes to competitive advantage in an adversarial field – there aren't many existing curated synthetic datasets to use.

"Data is the new oil" - Brian Krzanich, Intel CEO (2017)

But what if you don't have a generator?

CIFAKE – 120,000 real/synthetic images
- The real images are from CIFAR-10 dataset, which has 32x32 images
- Fake images were generated using **Stable Diffusion 1.4**
- https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images

3D Object Datasets:

SYNTHIA – 9400 multi viewpoint frames from a generated city with pixel-level semantics for 13 classes
- Each image has a resolution 1280 x 960
- Created in **Unreal Engine 4**

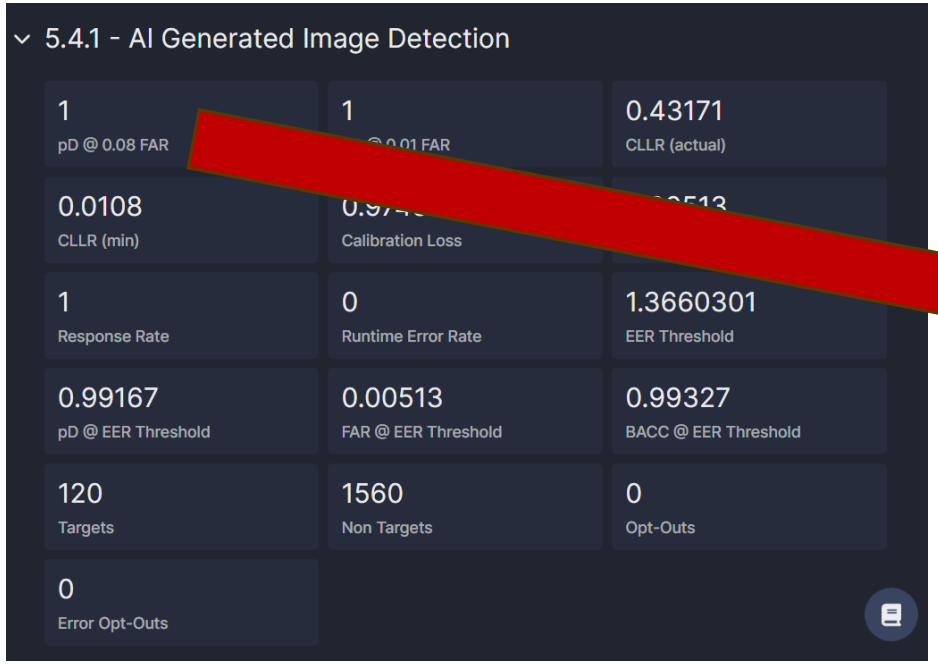
ShapeNet – large scale annotated dataset of 3D shapes – ShapeNetCore contains over 51,000 unique models across 55 object categories.
- 3D **CAD models** are in OBJ+MTL 3D format

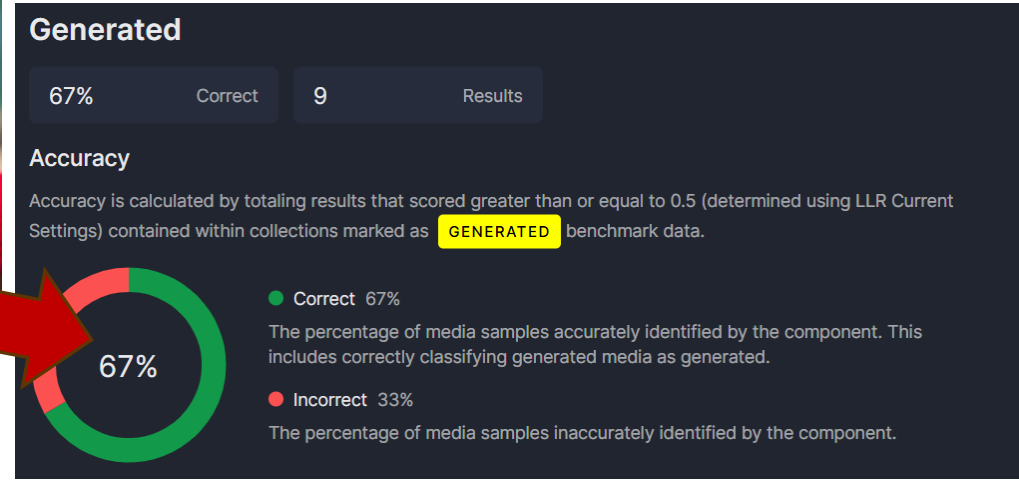
https://paperswithcode.com/paper/the-synthia-dataset-a-large-collection-of

Kaggle and Huggingface competitions are likely your best source of existing datasets

Key Considerations – Use Case

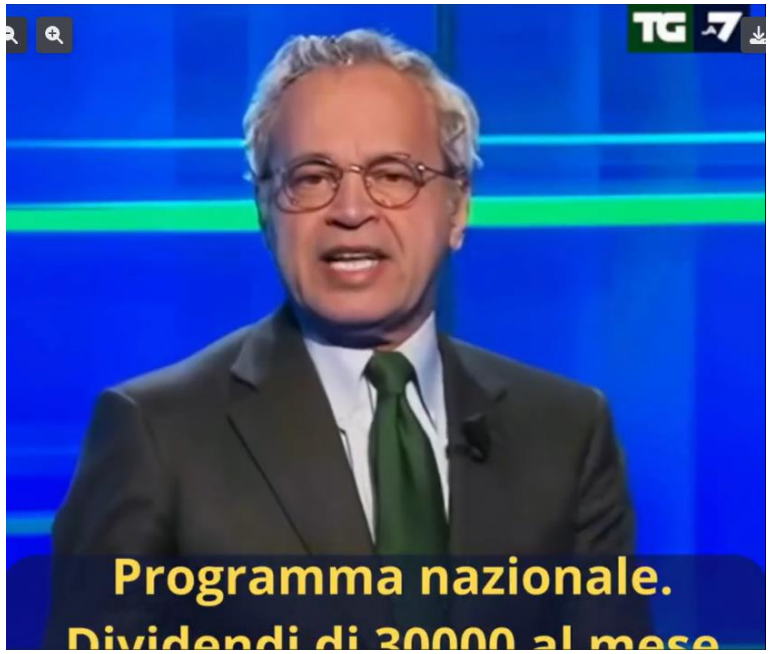
- An analytic trained on human faces is going to do a terrible job of detecting synthetic vehicles
  - Even other types of faces, such as animal faces, do not work well in face detectors

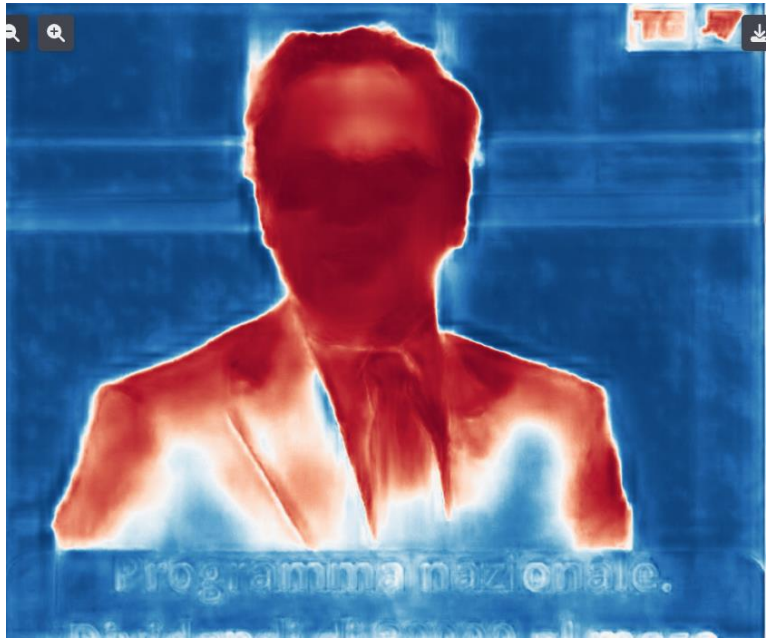




Key Considerations - Broadcasting

Overlays – if you are developing a forensics analytic that can operate on streaming data or questionable news segments, you may need to either programmatically localize and filter out those regions from images or include them in your training set.





Key Considerations – Feature Quantity

If you are feature-finding, how do you handle variance?

Analytic	Behavior
A	Will score the first face it finds in an image
B	Will score each face until it finds a falsified one or reaches the end of the image
C	Will score every face and average the value
D	Will score every face and report each score individually

Without Localization – which face did these analytics pick?

Key Considerations – Hybrid Edits

Does your detector handle real images with partially synthetic segments? Are you averaging across all regions of the image? Are you localizing?

The image to the right was made in Adobe:

- The base image (beach) is real
- The objects may be real or synthetic
- They were spliced in using generative fill
- The image was then resized and reformatted

Is this a synthetic image? What will your Analytic do?

Key Considerations – Input Understanding

Do you know if what you've been given is something **you're even capable of analyzing?**

- Is the input media in-distribution for your training set?
- Is it the same modality?
- Is it a size or duration that gives you enough details to work with?
- Does it contain the features you actually need (e.g. faces?)

If not – how does your algorithm communicate that?

Sometimes the right answer is no answer, but you should still explain why

Key Considerations – Feature Extraction

Semantic analysis requires a semantic understanding of the media by your algorithm.

One simple way to do this is to preprocess your media through a data labeling component:

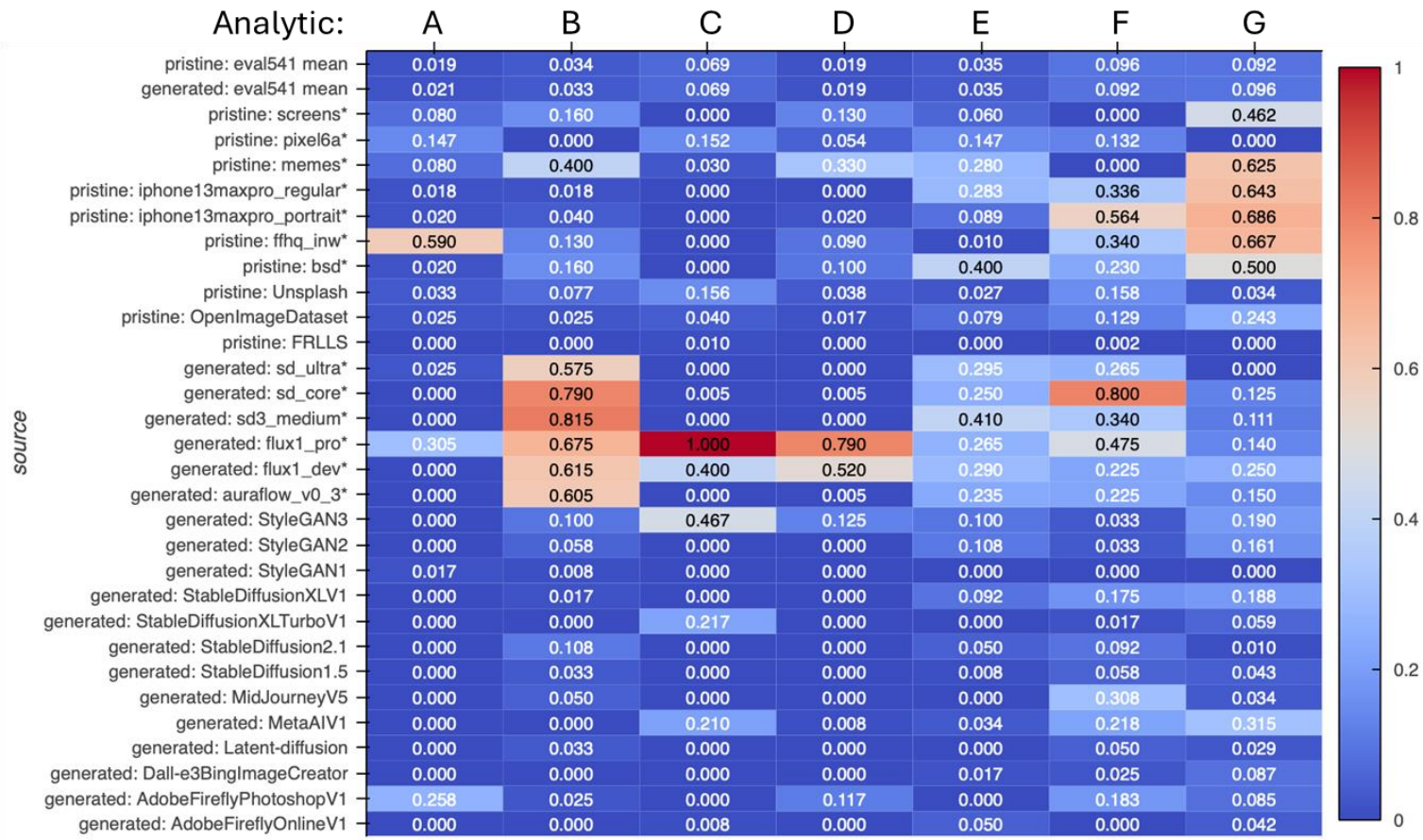
- OpenAI Contrastive Language-Image Pretraining (**CLIP**) – Deep Learning model that pairs images with text labels
- You Only Look Once (**YOLO**) – real-time object detection algorithm
- **R-CNN** – more accurate than YOLO but slower
- **MediaPipe and dlib** – face *landmark* detection (eyes, nose, ears, etc)
- **Build your own** CNN to extract the landmarks you think are relevant

If it isn't obvious yet, your algorithm may contain many networked models working together

Key Considerations – AI-Enabled Hardware

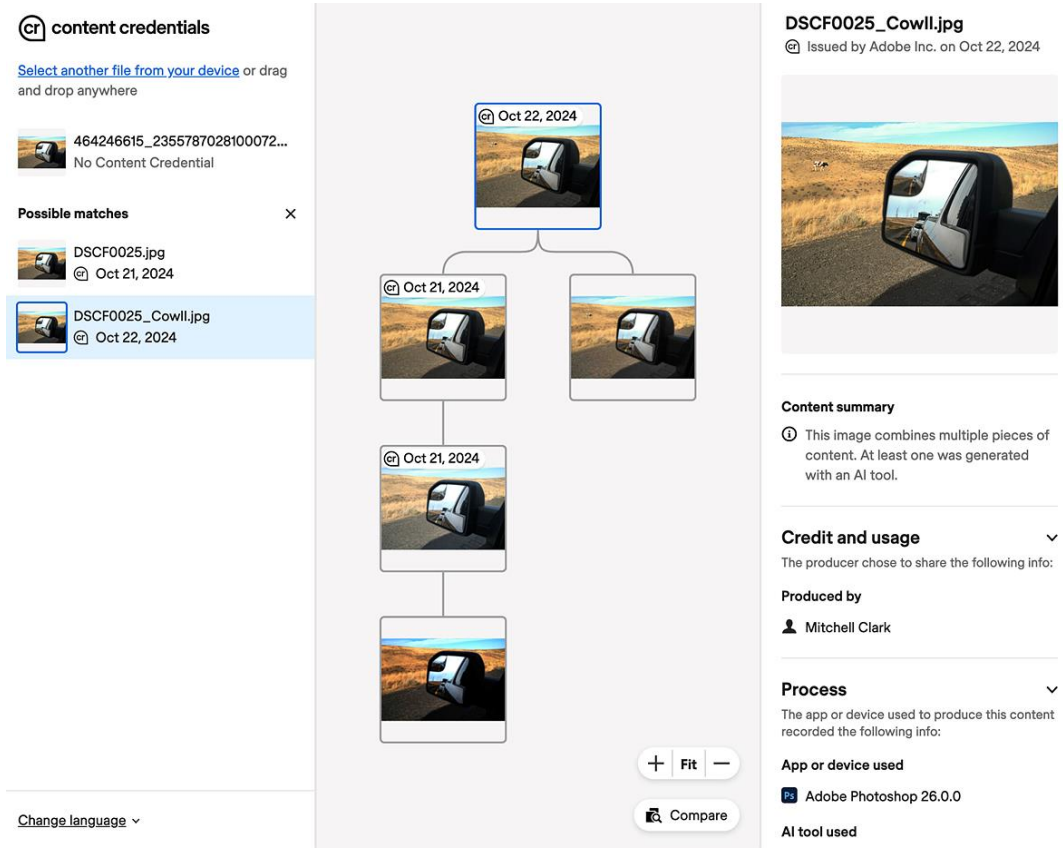
Do you need to factor in AI-enabled image processing on modern phones?

- The good news is that, as of now, *camera model doesn't seem to have an impact on accuracy* but some techniques prove vulnerable to specific types of media.
- However, that brings up a philosophical question – **what is "Pristine" media?**
  - "Radio-Ready" Music?
  - Post-Processed Video?
  - An Instagram filter over a photo?



Key Consideration - Certs

- Adobe **Content Authenticity Initiative** (CAI) and the **C2PA standard** provide new means of building trust in media by creating encrypted provenance chains that can verify if a photo has been edited and how without the need for a detector.
- It would be naïve to think hostile actors would choose to use these tools, but a first pass validation could save precious clock cycles at the cost of an API call.

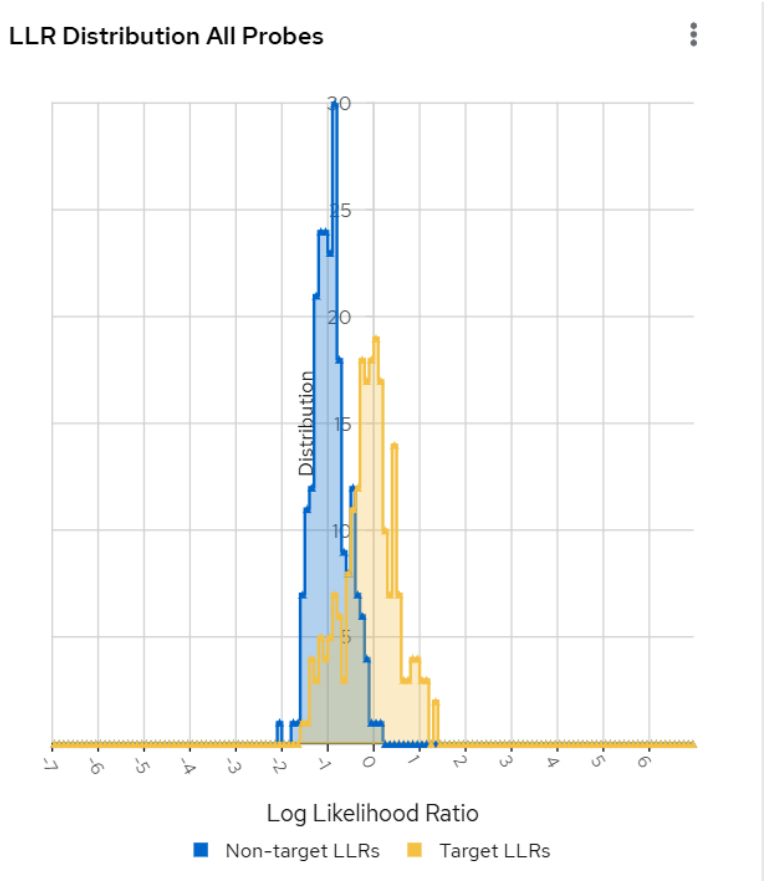


https://opensource.contentauthenticity.org/docs/verify/

https://www.dpreview.com/interviews/3215783269/adobe-max-2024-content-credentials-authenticity-initiative

Common Pitfall – LLR Calibration

- One challenge with LLR scoring is being able to express both negative AND positive hypotheses. Unlike a normal model confidence score that represents the probability of the input being in a class, LLR allows for the possibility of evidence directly against the hypothesis.
- Consider a multi-class detector that can identify both GAN and Diffusion images
  - If the GAN portion is expressing an extremely high confidence that the image was GAN generated, doesn't that provide strong evidence that it was NOT Diffusion generated?
- Because of this, the Equal Error Rate (EER) threshold for your LLR score may not actually be 0.
- If you graph the scores for each class of detection, your real "0" is the point at which the two curves intersect. Adjusting this value **creates a tradeoff between false positives and false negatives.**

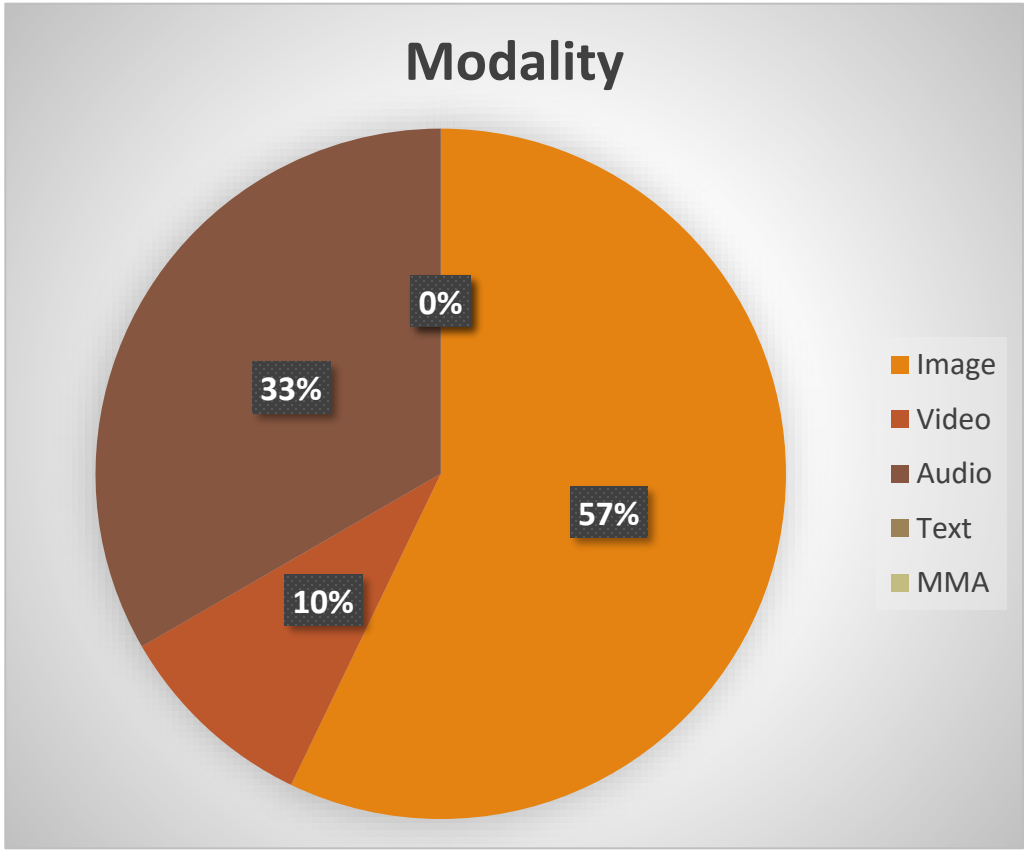


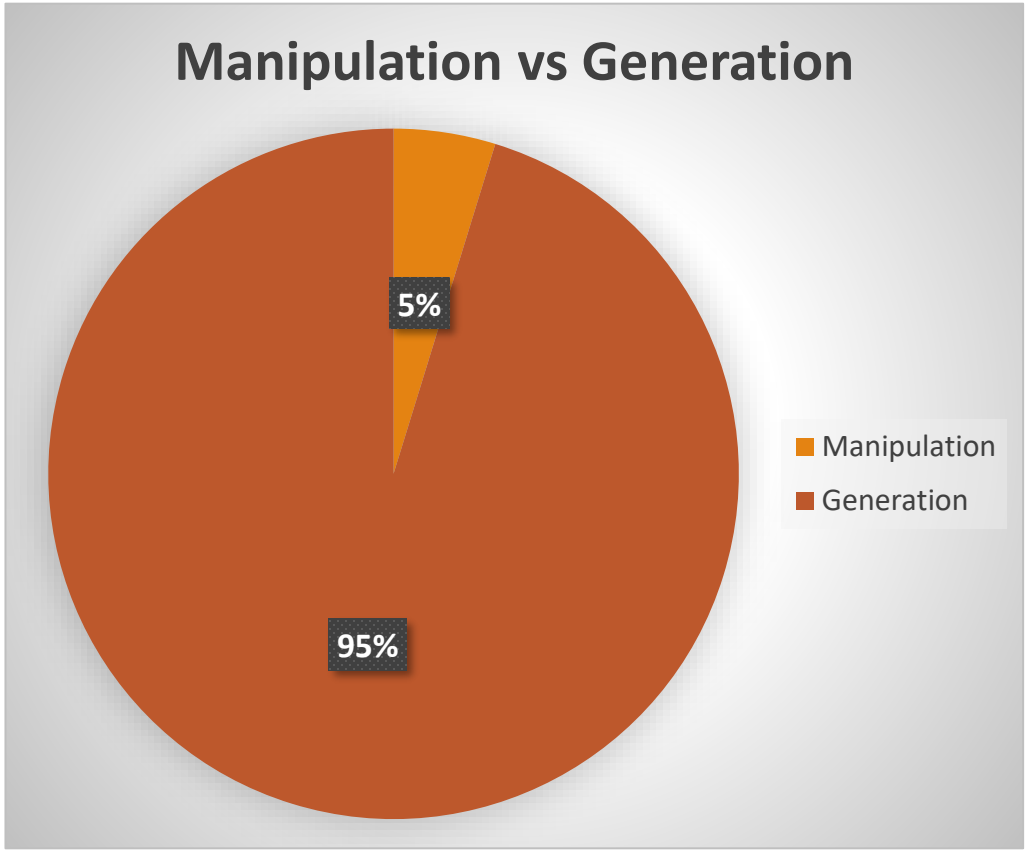
Known Gaps

Major gaps in Analytic Capabilities include:

- Brand new video generation capabilities
- Custom camera settings (high ISO, blacklight painting, etc)
- Low resolution imagery (< 800 x 800)
  - More relevant for face analytics
- Laundered images (resized, reformatted, picture-of-a-picture)
- Inconsistency in analysis given multiple faces in an image
- Not enough analytics provide evidence or localization

SemaFor Promoted Capabilities





Promotion Process

REFINING ALGORITHMS INTO ANALYTICS

TRL vs HRL

Technology Readiness Levels (TRL) - a scale (1-9) developed in the 1970s that measures the maturity of technologies for acquisition or integration – formally adopted by ISO with the publication of **ISO 16290:2013**

- Roughly: 1-3 (basic research), 4-6 (applied research), 7-9 (operational testing and deployment)

Human Readiness Level (HRL) – a scale (1-9) developed in 2010 that measures the readiness of technology for human use

- This scale was developed because most problems in engineered systems are linked to human error in use of a system.

https://www.incose.org/docs/default-source/enchantment/applyinghrlscaleintheseprocessjudisee220511.pdf?sfvrsn=2a61c7_2

Algorithm vs Analytic – Revisit

Scenario: A real-time audio deepfake cons an energy company out of $243,000 (~€216,000)

Scenario: A real-time video conference full of deepfaked staff fooled the audio detectors and convinced the finance department to wire $25 million (~€22 million) to a fraudulent account.

Scenario: You find that over a dozen fraudulent bank accounts have been opened by one person capable of tricking facial recognition technology with your stolen ID card.

Scenario: An analytic designed to identify real versus synthetic people falsely classifies a real person as synthetic. A missing person's case is never filed.

In each case, here is what is provided:

Algorithm	Analytic
A black box prediction with an associated confidence score, perhaps a confusion matrix of test results	A series of localized, evidence-based claims each individually scored based on the level of confidence that combine into a broader level analysis of at the semantic level, further backed by explicit details about the model itself that can assist in evaluating whether the source media was a good fit for this analytic.

https://www.trendmicro.com/vinfo/us/security/news/cyber-attacks/unusual-ceo-fraud-via-deepfake-audio-steals-us-243-000-from-u-k-company
https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk

Imagine being a judge, or a jury, or a committee accepting this as evidence….

Promotion Criteria

- During the final phase of DARPA SemaFor, the LM ATL team created a new set of evaluation criteria designed to weed out among the 100+ analytics which ones are actually useful.
- Within some classes of detection, multiple analytics achieved perfect accuracy in IT&E trials – but none of them come anywhere close in “open world” testing.
- The promotion process is meant to evaluate and help mature the HRL of algorithms – creating an analytic capable of broad-spectrum use and robust to misinterpretation.
- Tests focus less on classification accuracy and more on jargon, explainability, clarity, and consistency.
- A poorly understood system will get used until it fails and then never again.

Test	Purpose
Analytic passed Gate Test	*Ensure analytic responds to probes*
Analytic produces Scores within range	*Allows for comparison of results between analytics*
Analytic complies with Taxonomy	*Ensures everyone is speaking the same “language”*
Analytic summary / description valid	*Ensures non-technical analysts can understand what an analytic does*
Analytic Executive Summary complete	*Ensures technical operational analysts have the material they need to interrogate individual results*
Analytic input assumptions clearly documented	*Helps users interpret results*
Analytic does not opt out of invalid nodes	*Minimizes confusion by ensuring analytics don’t opt out while also providing a score*
Analytic Opt-Out reason meaningful	*Helps users understand why an analytic did not produce scores*
Analytic does not crash in the HMI	*Ensure minimal stability in use*
Analytic produces scores in the HMI	*Validation of prior tests ensuring it doesn’t only run in the Gym*

1. Analytic has passed Gate Test

Description: The analytic has been submitted to the Gate Test and passed initial testing

- "Gate Test" was a submission-driven process that automatically ran a new analytic through a few basic tests to ensure it does not crash on launch, produces valid datamodel objects, and did not time out due to excessive processing time.

Why it matters: A surprising number of analytics would get submitted without real testing – Gate Test was designed to reduce support labor in helping identify analytic issues by stopping bad code before it is evaluated.

Example: N/A

2. Analytic produces valid LLR scores

Description: The analytic produces LLR scores that are between -5 and 5

Why it matters: Any degree of normalization in score output will improve fusion and HMI performance in representing overall conclusions to the user.

Example: N/A

3. Analytic uses evidence taxonomy

Description: Evidence Graphs produced by an analytic use the SemaFor Evidence Taxonomy when representing their category, evidenceType, and (evidence) instance.

Why it matters: It allows apples-to-apples comparison between analytic results while also eliminating spelling errors

Example: An analytic has determined that an image was generated using StyleGAN2. The EG should have:

- *EvNonSemanticConsistencyCheckNode.category = ConsistencyWithGenerationModel*
- *EvConceptNode.evidenceType = ToolsTechniques*
- *EvConceptNode.instance = StyleGAN2*
- *invalid types: StyleGAN, StyleGan, stylegan, style-gan, STYLEGAN, “not StyleGAN2”, “generated using StyleGAN”, StyleGNA, etc….*

4. Analytic summary / description valid

Description: The summary and description of the analytic are an appropriate level of technical detail to support analyst needs.

Why it matters: The system uses a hierarchical methodology for displaying information, starting with rolled-up conclusions but allowing the user to dive into details as needed. The system needs to support both analysts with no technical background AND analysts with deep machine learning experience. As such, providing scaling level of detail in component descriptions support this approach.

Example: GENERALLY speaking, an analytic description should use the following guidelines. Included is an illustrative example using “An analytic that looks at faces to look for inconsistency in specular highlights”

Field	Detail	Example(s)
summary	no jargon. This should explain what the analytic looks for to the level of detail that would make sense for a non-technical analyst. Input constraints should be captured here at a low level.	“This analytic looks at the eyes of a human face in an image for evidence that it may be generated”
description	limited jargon. This should provide some specifics around what the analytic is trying to do. Input constraints should be more specific here and can use program terminology.	“This analytic looks for specular highlight inconsistencies in the eyes of a face in an image if the face is larger than 64x64 pixels. If more than 1 face is in an image it will analyze the first face it finds.” “This analytic determines if a face in an image for a specific person has been manipulated. It must be told which person it is looking at in the image or it will Opt Out of analyzing the media.”
executive summary	high level of specificity. As much jargon as needed to explain to an ML expert how the analytic works.	N/A - fill in every field where possible.

5. Analytic Executive Summary complete

Description: The executive summary values in the component.yaml are filled in. The only optional field is the reference to published papers, although this is a high-demand field for users.

Why it matters: For users with deep technical experience, the executive summary is the final place they will be able to look to understand how an analytic works. They may also require those details to include in reports they are generating as part of their job and may have to explain their conclusions in front of congress or a judge. Insufficient detail will lead to a lack of trust which will lead to that analytic not being used in the future.

Example: Fill in every field where possible. Users have expressed a desire for linking to publications where applicable. In particular, constraints such as “this was only trained on X type of data” or “this analytic performs poorly against new generation techniques outside the GAN class of techniques” help users know when they SHOULDN’T trust results from an analytic.

6. Analytic input assumptions documented

Description: Either in the summary, description, executive summary, or some combination of all 3, the analytic clearly explains what it is assuming about inputs being fed.

Why it matters: This is especially critical for analytics that were submitted to an evaluation but NOT further matured to support open world use. Input assumptions are needed by users to better interpret results and to help address conflicting results in the HMI. No input assumptions documented means the analytic is declaring that you can give it anything and it will produce reasonable results (which could include opting out)

Example:

“This analytic has only been trained on PNG images.”

“This analytic assumes any images it is given have exactly 1 face in it.”

“This analytic has never been tested on generators outside the StyleGAN family of techniques.”

“This analytic requires an Analytic Scope to be given to it at runtime or it will always opt out of analysis.”

7. Analytic does not opt out incorrectly

Description: The analytic is using the opt out process in the evidence graph correctly.

Why it matters: Users need to know what analysis was done, but also what analysis was skipped. If your analytic is the only one in a workflow that looks for Latent Diffusion, and it opts out, the user may incorrectly assume this means the image isn't diffusion generated. Similarly, if a user uploads a single piece of media and sees an analytic both provide a score AND opt out it creates confusion.

Example:

An analytic should ONLY opt out of the root node (*AgMultiMediaAssetNode)* if it is opting out of an entire MMA. Its should **never** opt out of this node if it is also providing a score.

The presence of an *AgDecomposedDocumentNode* in an AG does not mean there is text in the mma. This node points to the AOM wrapping the source media. Analytics need to use this node to access the AOM to determine if there is actually text present. A great many analytics incorrectly opt out of this node even when providing scores on a single modality upload.

An analytic should opt out of EVERY Ag<media>Node that it does not support in an MMA. If an AG contains an AgImageNode, an AgAudioNode, and an AgVideoNode, and that AG is handed to an image analytic, it should opt out of both the AgAudioNode AND the AgVideoNode and **provide a meaningful opt out reason for both**.

8. Analytic Opt-Out reason meaningful

Description: Every Opt Out *requires* a human-readable explanation as to why the analytic chose not to respond to a probe. It cannot be a generic “catch all” opt out reason. The Opt Out explanation is not a place for debug messages targeted at developers.

Why it matters: Without an explanation, the user has no idea why the system didn’t do what it was asked to.

Example:

- (GOOD) “Opted Out of Image - analytic could not find a face with enough detail in the provided image”
- (GOOD) “Opted out of Audio - analytic only supports MP3 files and was given a WAV”
- *(BAD) [Errno 2] No such file or directory: '/mnt/sandbox/<analytic redacted>/Pentagon Explosion.png’*
- *(BAD) {self.comp_sig} does not handle "{node.node_type}"*
- *(BAD) Exception: local variable 'node_id' referenced before assignment*

9. Analytic does not crash frequently

Description: As a final set of checks, Analytics are run against batch results to ensure it does not crash more than 10% of the time.

Why it matters: This ensures stability in performance

Example: N/A

10. Analytic produces scores

Description: As a final set of checks, Analytics are run against batch results to ensure it did not neglect to capture input constraints and requirements. If it opts out 100% of the time and a reason is not provided it will fail at this step.

Why it matters: This ensures analytics being added to the HMI actually produce results.

Example: N/A

Part 3

THE SEMAFOR PLATFORM

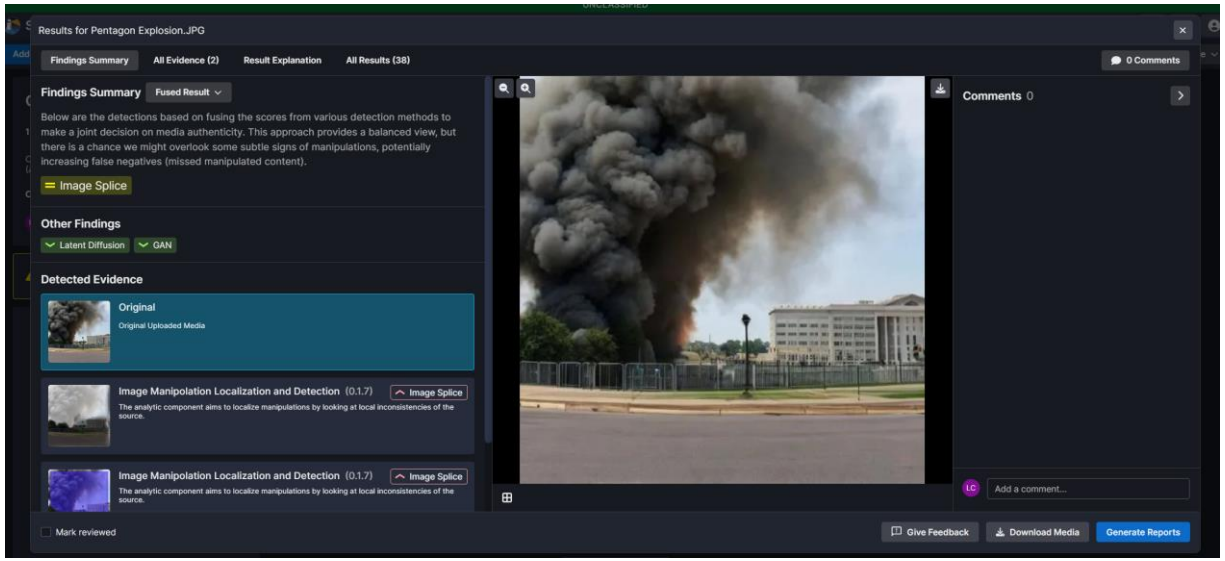
SemaFor Capabilities

The SemaFor System consists of a capability that provides analysts with the ability to upload media assets, select analytics to run, and produce results ***for detection, attribution, and characterization***.

Analytic capabilities include:

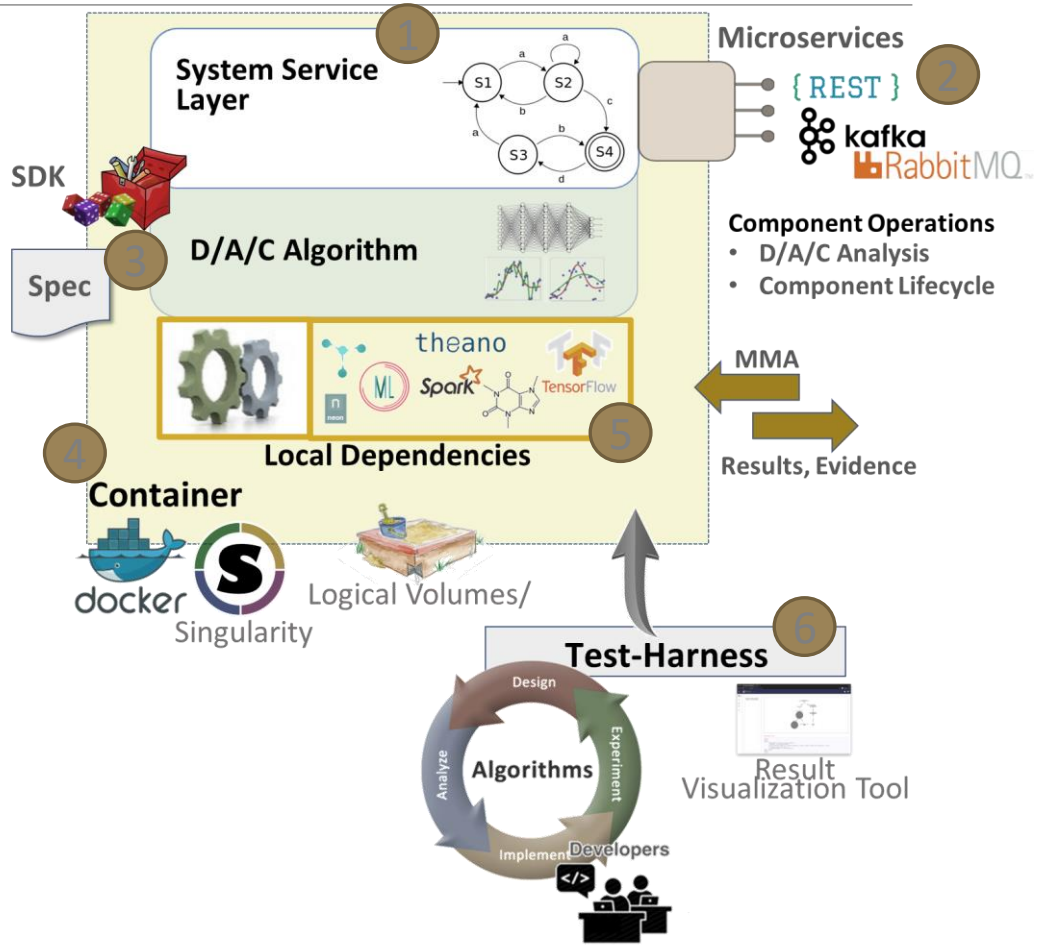
- Person of Interest Deepfake Audio and Video Models
- Synthetic Image Detection for Diffusion and Generative Adversarial Network (GAN) Models
- Paste-Splice Manipulation Detection
- Synthetic Audio Detection
- Synthetic Text Detection
- Generator and Tool Attribution

SemaFor Portal



System: Component Structure and Tools

1. **Service Layer –** Manages communication and state.
2. **Component Operations –** Lifecycle and comms, includes pub/sub and RESTful interfaces.
3. **D/A/C Analytics –** Algorithm implementation leverages system SDK/APIs and component specifications.
4. **Container –** Encapsulates component for runtime deployment with mapped logical volumes.
5. **Dependencies –** Encapsulated dependencies reduce integration overhead and provide max flexibility.
6. **Component Toolkit –** Local (outside the system) test-harness, validation, and visualization tools



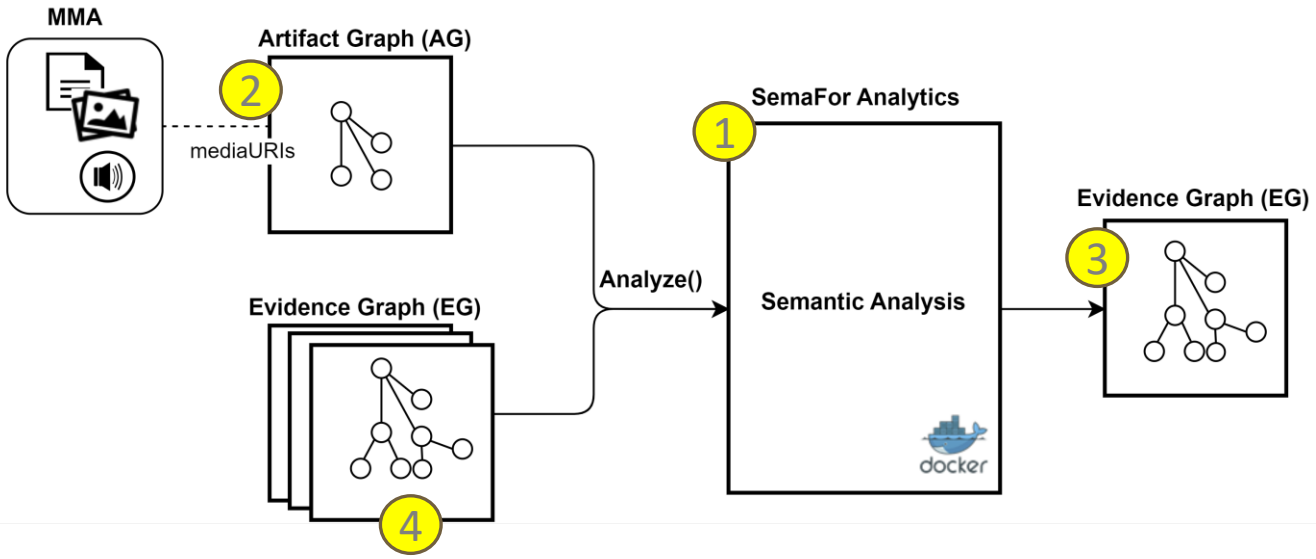
Components are packaged as microservices with a system provided service layer to ease integration.

System: Analytic Components Operation

Containerized analysis components process multimodal assets and produce analysis results with evidence.

1. Analytic components support semantic D/A/C analysis.
2. Input to the analytic is a multimodal asset (image, text, video, etc.) converted into an Artifact Graph (AG).
3. An Evidence Graph (EG) is produced by the component containing analysis results and supporting evidence.
4. Some analytics also take EGs from previous analysis as input.

D/A/C Analyze requests are sent to analytic components to process multimodal assets.



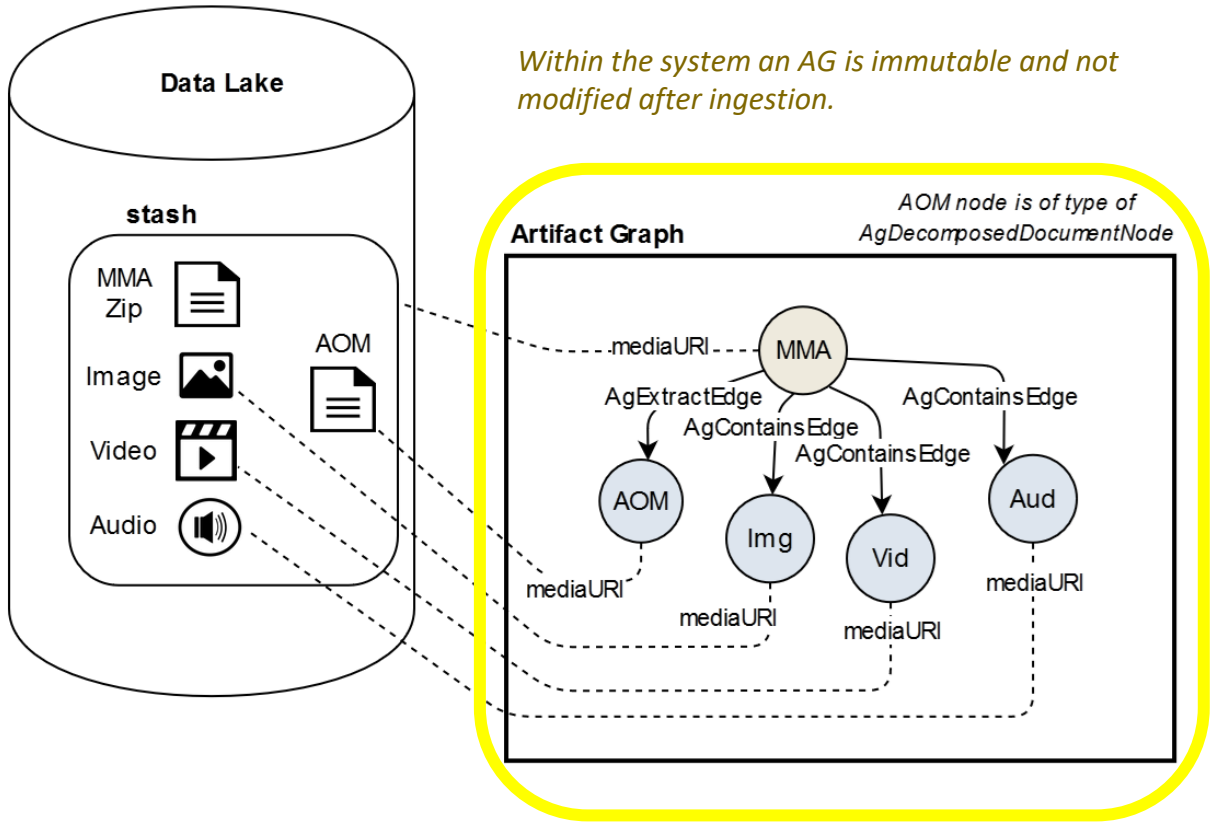
SemaFor's Graph-based data structures are easily visualized and clearly link claims to evidence

System: Data Model – Artifact Graph (AG)

The system captures multimodal media assets (MMAs) as an Artifact Graph (AG).

The AG captures:

- URIs and containment relationship between the MMA and its individual media files stored in the data lake.
- Extracted content from the MMA stored as Asset Object Model (AOM) markdown.
- Metadata associated with an MMA itself and the media assets it contains.
- An augmented AG (not shown) is used to capture dynamic content generated during analysis (e.g., heatmaps).

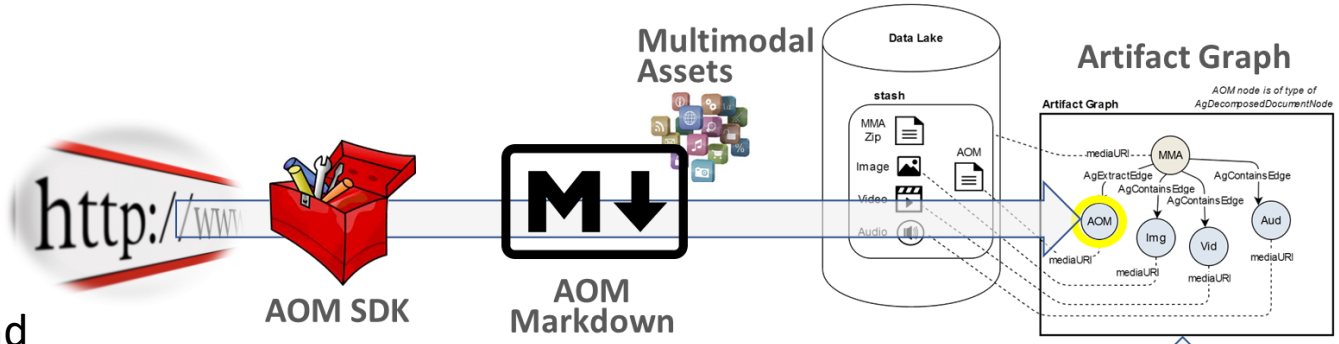


The system uses an Artifact Graph to capture a multimodal asset's media and metadata.

System: Data Model – Asset Object Model (AOM)

The Asset Object Model (AOM) is a markdown document that captures content of multimodal media assets (MMAs).

- AOMs are referenced from Artifact Graphs and can be used by the Analytics to process the text portion of MMAs and references to structural elements of the MMA.
- AOM SDK supports methods to navigate, filter, slice, and extract data from the document structure.



AOM SDK

```
from aomsdk import *
import os

save_path = os.path.dirname(os.path.abspath(__file__))

document = fetch_parsed_document("https://www.straitstimes.com/singapore/
document.save_aom_to_package(output_dir=save_path, package_dir=None)
```

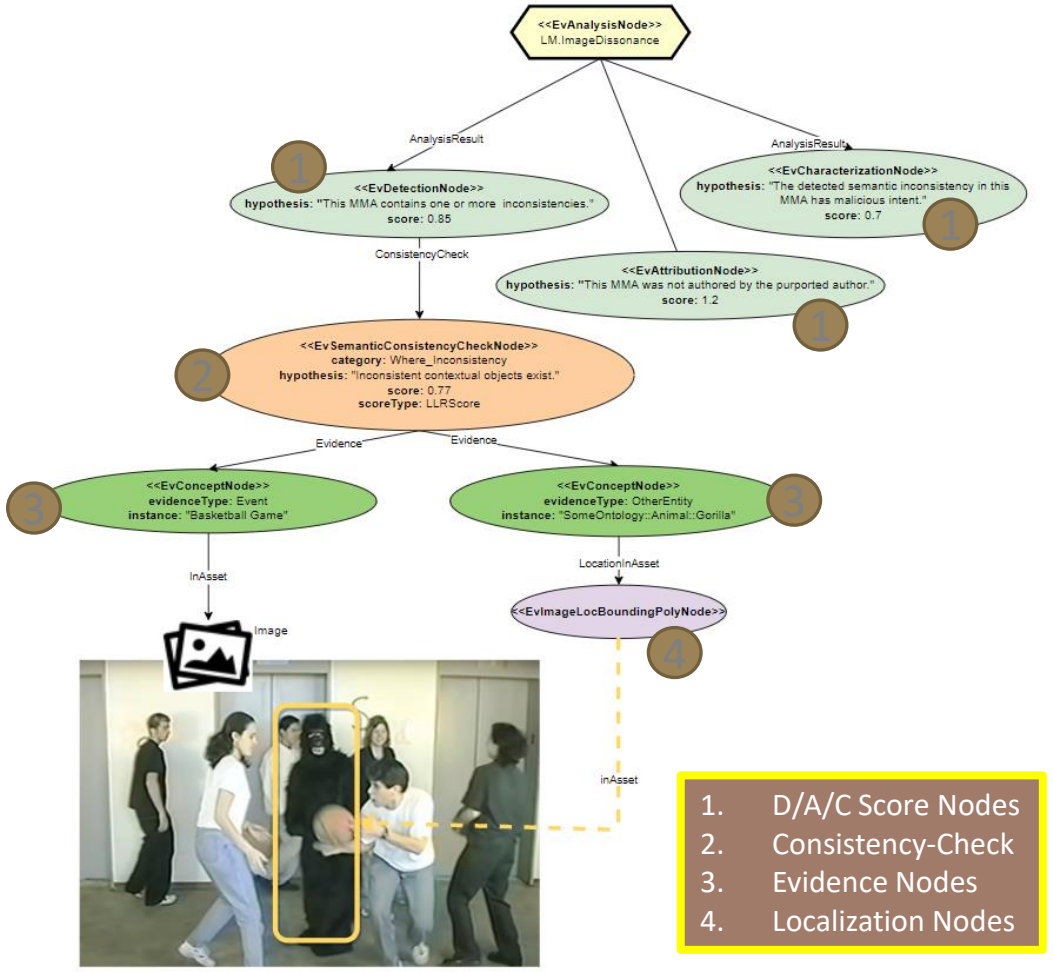
D/A/C Analytic

AOM is a markdown representation of the multimodal media asset with a supporting API.

System: Data Model – Evidence Graph (EG)

An Evidence Graph (EG) captures evidence from the analysis that was done on an MMA.

- Scoring and evidence is captured in the EG for use by Fusion, Prioritization, Explanation, and for presentation in the HMI.
- An EG has D/A/C nodes to capture log likelihood ratio (LLR) scores that reflect the *hypotheses* explored by the analytic (*consistency-checks*).
- The EG logically captures hypotheses, evidence, and localization information.
- Analytics localize evidence to any modality (image, video, text, audio) in the MMA.

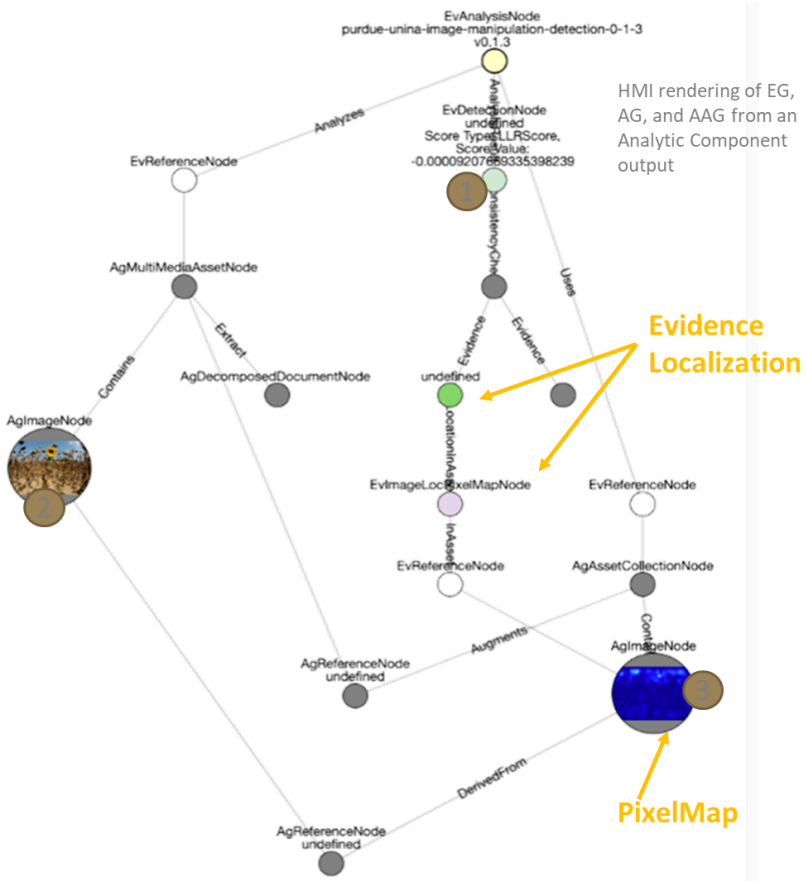


Evidence Graphs capture and localize evidence from D/A/C analysis of input media.

System: Data Model – EG *(continued)*

- Analytic Components create an Evidence Graph (EG) as a standard part of its output to express evidence supporting conclusions.
- Optionally, analysis can result in augmented graphs containing dynamically generated evidence.
- Figure (right) shows evidence from GAN analysis on an image

1. Detection analysis performed
2. Source image being analyzed
3. Dynamically generated image (pixel map) used as localized evidence

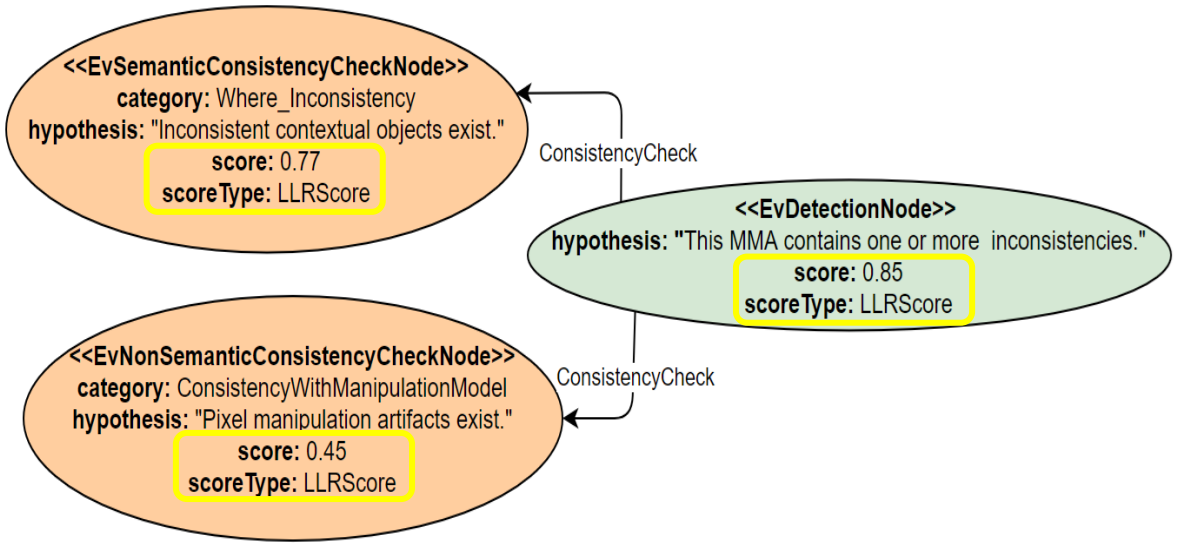


Evidence localization can include dynamically generated content.

System: Data Model – Evidence Graph (EG) *(continued)*

Log Likelihood Ratio (LLR) scores indicate the relative strength of evidence and are reported at the D/A/C node level and the consistency check level.

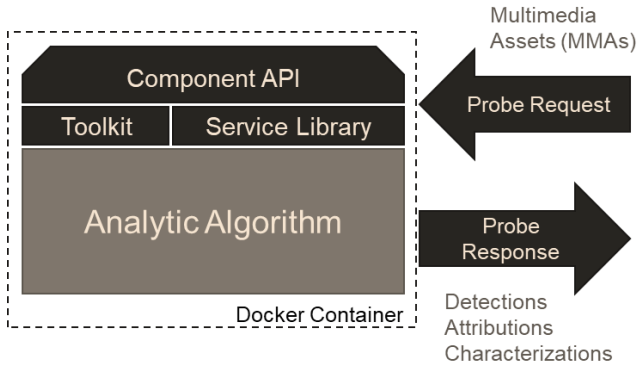
- LLR has a logarithmic scale
- LLRs relate to an analysis hypothesis:
  - **LLR = 0** → Given evidence is neither in favor of nor against the hypothesis.
  - **LLR > 0** → Given evidence is in favor of the hypothesis.
  - **LLR < 0** → Given evidence is against the hypothesis.



LLR scores are used to indicate the relative strength for or against an analysis hypothesis.

SemaFor Deployment Package

- **Tools and Libraries**
  - **Graph Validation Tool –** Validates all SemaFor graphs against schema and logical structure.
  - **AAG Generator –** Generates an AAG from a directory structure.
  - **Graph Visualizer –** Graphically display and validate AG/AAG/EG/ etc.
  - **Test Harness –** Local (developer environment) tool for deploying and interacting with SemaFor components.
- **SemaFor** APIs
  - **App Server API –** Interfaces for interacting with the Application Server.
  - **Component API –** Component interfaces for lifecycle control.
- **Major Data Structures** *(OpenAPI)*
  - Artifact Graph (AG) and Augmented Artifact Graph (AAG)
  - Evidence Graph (EG)
  - Messages and Common Type Definitions
- **System Services and HMI**
  - SemaFor web-based Portal
  - Helm deployable system services (backend): Kafka, MinIO, and MongoDB
- **Component Tools and Examples**
  - **Component** Toolkit **–** Helper functions for component development.
  - **Component Service Library –** Core services interface to the SemaFor system.
  - **Mockup Analytic Component –** Configurable analytic for load testing and integration testing.
  - **LM.MediFor Proxy –** Proxy component for accessing existing MediFor Analytics.



SemaFor comes with everything necessary to continue development as desired by transition partners

Programatically Running SemaFor

BATCH PROCESSING OF DATA

Decision Factors for how to run

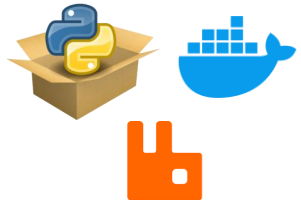
	HMI	API	Pipeline Tool
Full export / all analysis artifacts available	Yes	Yes	Yes
K8s Required	Yes	Yes	No
Parallel Processing	Yes	Yes	No*
Visualization of Results	Yes – In HMI	Yes* – In HMI	No*
User Interaction Supported	Fully Graphical	Command Line or None (i.e. automated programmatically)	Command Line or None (i.e. automated programmatically)
Supported Data Location	Local files only	Anywhere*	Local files only
Best Batch Sizes	Small	Any number	Medium+
# of concurrent users	Multiple	Multiple	Only 1

Keep in mind these **DO NOT** require a “cluster”

Analysis Pipeline Tool

WHEN ALL YOU HAVE IS DOCKER

Analysis Pipeline Tool

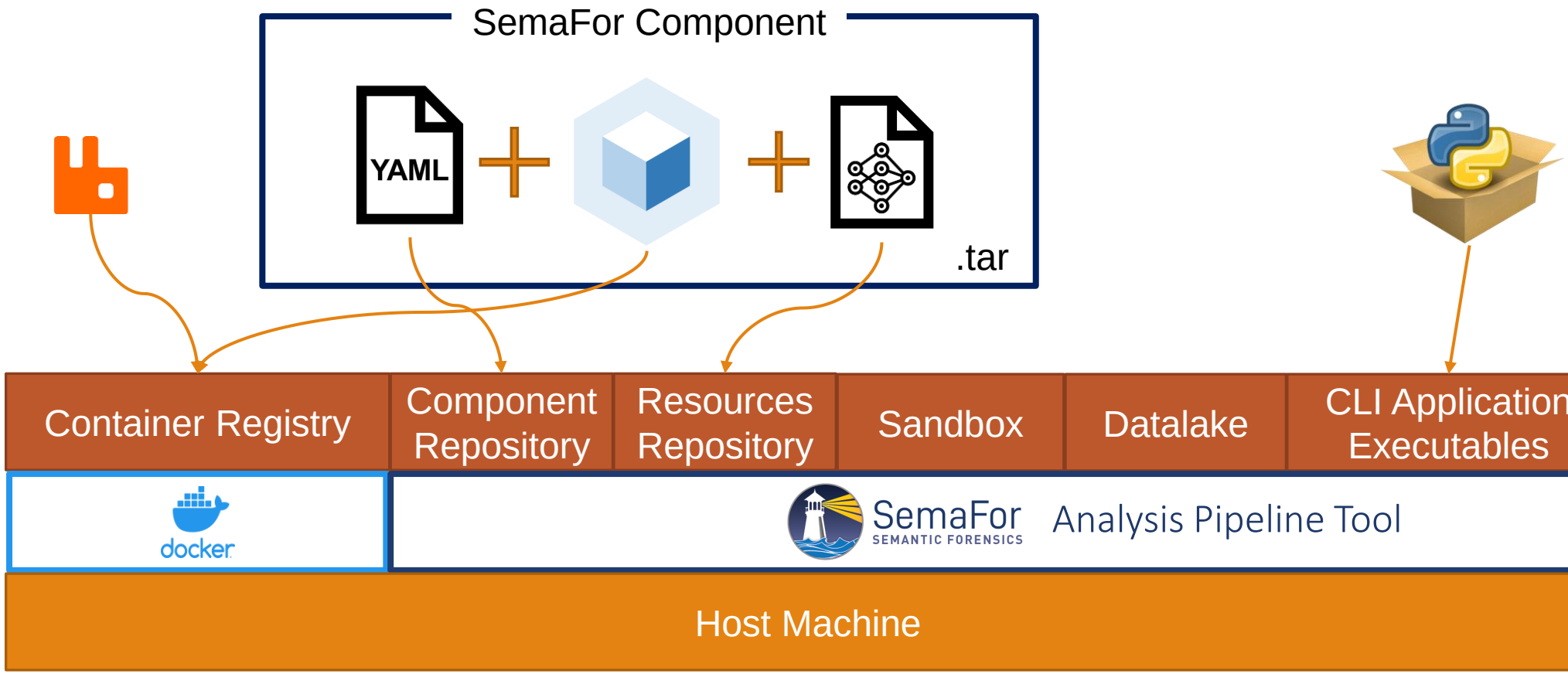


- Pure Python Application, relying only on Docker to manage analytic component containers and a rabbitmq message broker container
  - Have also containerized the application itself so that you can run docker-in-docker if desired

- Can be used to “batch process” a local dataset with a handful of analytics and stores all data in local file storage
  - Not true batch processing – only a single instance of the analytic is started to avoid resource (GPU) contention – but it will process an entire batch, not just a single piece of data
  - Will auto-ingest / convert any input media to an AOM/AG

- Can also be used programmatically to do some slightly more advanced data processing or chaining of analyses

Architecture



Relevant papers

A COLLECTION OF SEMAFOR INNOVATIONS

Relevant Papers

Analytic	Paper
kitware-asu-generated-text-detection - stylometry	https://arxiv.org/abs/2309.03164
kitware-asu-generated-text-detection	https://arxiv.org/abs/1908.09203
kitware-cu-pix2struct	Enhanced Chart Understanding via Visual Language Pre-training on Plot Table Pairs, ACL Finding 2023, https://aclanthology.org/2023.findings-acl.85
kitware-str-image-classification	'Radford et al., "Learning Transferable Visual Models From Natural Language Supervision", 2021. https://arxiv.org/abs/2103.00020
kitware-ub-dsp-fwa-deepfake-detection	https://arxiv.org/abs/1811.00656
kitware-ub-gan-face-detection	https://ieeexplore.ieee.org/document/9897972
kitware-ub-glff-ai-image-detection	GLFF: Global and Local Feature Fusion for AI-synthesized Image Detection, TMM, link: https://arxiv.org/pdf/2211.08615.pdf
kitware-ub-lipsync-deepfake-detection	https://arxiv.org/pdf/2401.10113
kitware-umichigan-av-sync-consistency	https://arxiv.org/abs/2301.01767
kitware-umichigan-cnn-detmatch	https://arxiv.org/pdf/1912.11035.pdf
purdue-mp3-gen-audio-detect-v1	[A] Xiang, Ziyue, Paolo Bestagini, Stefano Tubaro, and Edward J. Delp. "Forensic Analysis and Localization of Multiply Compressed MP3 Audio Using Transformers." In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2929-2933. IEEE, 2022.
purdue-polimi-synthetic-audio	'Exploring the Synthetic Speech Attribution Problem Through Data-Driven Detectors, D. Salvi, P. Bestagini, S. Tubaro, 2022 IEEE Internationl Workshop on Information Forensics and Security (WIFS), https://ieeexplore.ieee.org/abstract/document/9975440'

https://engineering.purdue.edu/SEMAFOR/publications/

Relevant Papers

Analytic	Paper
purdue-polimi-synthetic-image-detection	'Detecting gan-generated images by orthogonal training of multiple cnns, S. Mandelli, N. Bonettini, P. Bestagini, S. Tubaro, 2022IEEE International Conference on Image Processing (ICIP), 3091-3095, https://arxiv.org/pdf/2203.02246.pdf
purdue-unina-audiovideo-poi-forensics – audio only	'Audio-visual person-of-interest deepfake detection, CVPRW 2023'
purdue-unina-audiovideo-poi-forensics – audio only beats	'Training-Free Deepfake Voice Recognition by Leveraging Large-Scale Pre-Trained Models
purdue-unina-deepfake-id-reveal	'Id-reveal: Identity-aware deepfake video detection, ICCV 2021'
purdue-unina-image-manipulation-localization - adobe	'Noiseprint: a CNN-based camera model fingerprint (TIFS2020), https://arxiv.org/abs/1808.08396 [B] TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization (CVPR2023), https://arxiv.org/abs/2212.10957
purdue-unina-synthetic-image-detection - multi	'Are GAN generated images easy to detect? A critical analysis of the state-of-the-art (ICME 2021), https://arxiv.org/abs/2104.02617
purdue-vid-cont-mdata-df-detect-v1	Ziyue Xiang, Janos Horvath, Sriram Baireddy, Paolo Bestagini, Stefano Tubaro, Edward J. Delp; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021, pp. 1042-1051
sri-umd-yydeepfaketc	'Exploring Temporal Coherence for More General Video Face Forgery Detection, Yinglin et al, ICCV 2021'
sri-umd-yyimagesplicegsrnet	'Generate, Segment and Refine: Towards Generic Manipulation Segmentation. .... P Zhou, BC Chen, X Han, M Najibi, A Shrivastava, SN Lim, LS Dais AAAI 2020, arXiv: 1811.09729'
tonic-bu-multi-length-detector-ud	'Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach", 2019. https://arxiv.org/abs/1907.11692


LOCKHEED MARTIN