Semantic Forensics (SemaFor)

Matt Turek

Agenda

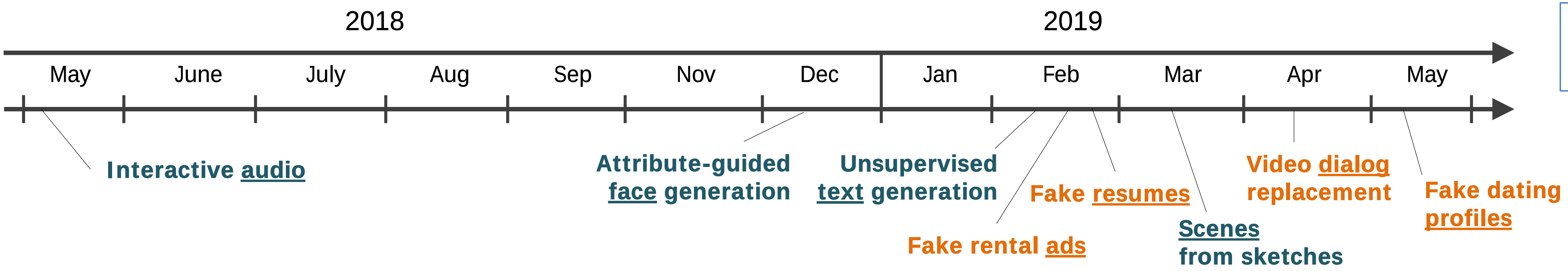
Start	End	Duration	Item
8:00 AM	9:15 AM	1:15	**Registration**
9:15 AM	9:25 AM	0:10	**Security Briefing** Leon Kates, DARPA MSO/SID
9:25 AM	9:45 AM	0:20	**Human Use Briefing** Ms. Lisa Mattocks, DARPA Assistant Director, STO
9:45 AM	10:15 AM	0:30	**Contracts Management Office Briefing** Mr. Mark Jones, DARPA CMO
10:15 AM	11:00 AM	0:45	**Semantic Forensics (SemaFor) Presentation** Matt Turek, Program Manager, DARPA I2O
11:00 AM	11:15 AM	0:15	**Turn in Questions** SemaFor@darpa.mil
11:15 AM	12:15 PM	1:00	**Lunch/Networking/Teaming** **On your own**
12:15 PM	1:30 PM	1:15	**Q&A Session** (Answer attendee questions)

- BAA Location and Dates
  - Posted on FedBizOpps website (http://www.fedbizopps.gov) and Grants.gov website (http://www.grants.gov)
  - Posting Date: August 23, 2019
  - Abstract Due Date: September 11, 2019, 12:00 noon (ET)
  - BAA Closing (Proposal Due Date): November 21, 2019, 12:00 noon (ET)
- Procedure for Questions/Answers Today
  - Questions can be submitted until 11:15am (ET) to SemaFor@darpa.mil or on 3x5 cards
  - Questions will be answered during Q&A session in the afternoon
  - Waiting until the session is complete is encouraged to avoid repetition
- Websites
  - Proposers Day website
  - SemaFor program website
    - Proposers Day Presentations
    - Frequently Asked Questions (FAQ) will be updated with Q/A from SemaFor@darpa.mil

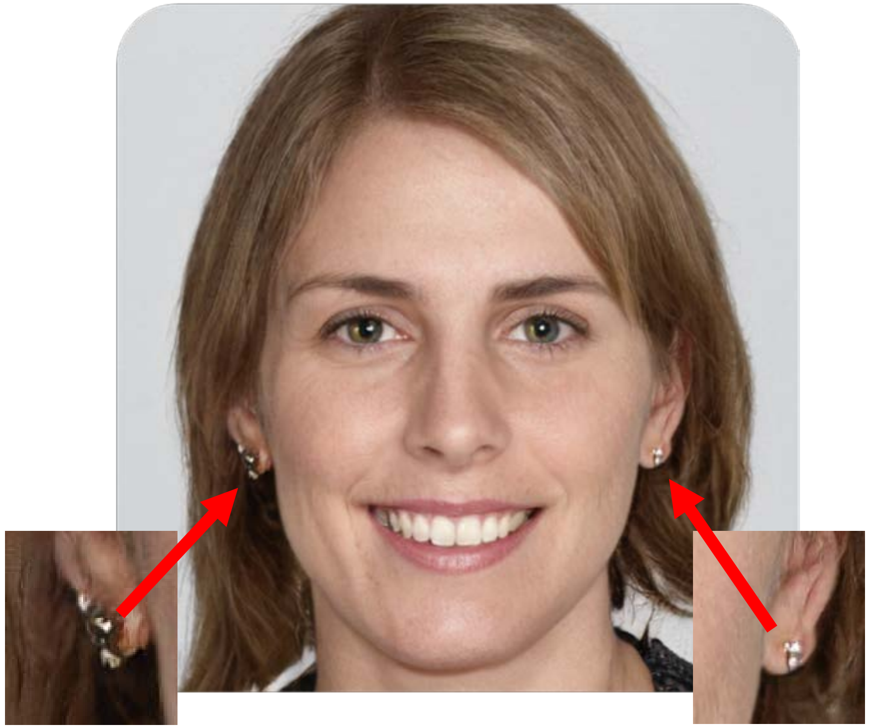
Objective

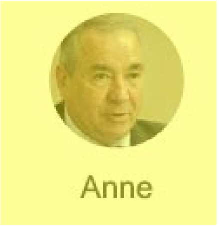
Create rich semantic algorithms that automatically detect, attribute, and characterize falsified multi-modal media to defend against large-scale, automated disinformation attacks

Incredible Pace of Synthetic Media Generation



Legend:
Single modality
Multi-modality





ENTIRE GUEST SUITE

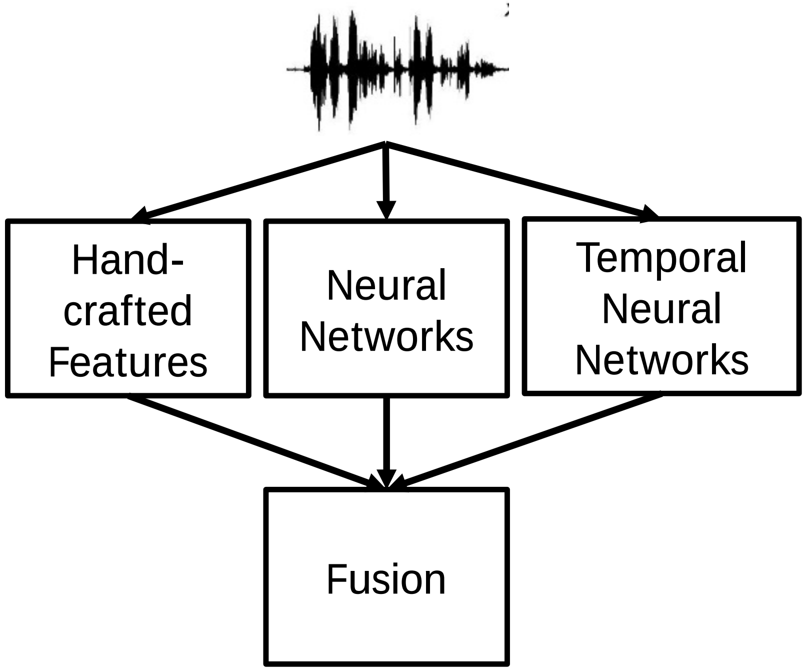
Luxury Condo 3 Bed + 3 Bath
Port Melbourne

○ 8 guests ○ 3 bedrooms ○ 4beds ○ 2 baths

Bathroom (with seating for 2 more people), basin and eclectic French garden and kitchen. 24/7 carpeted charc. Laundrymemberly : More balcony – Garden – Metro, Liverpool Street (15 min walk) Walking distance to Wyckofferdon

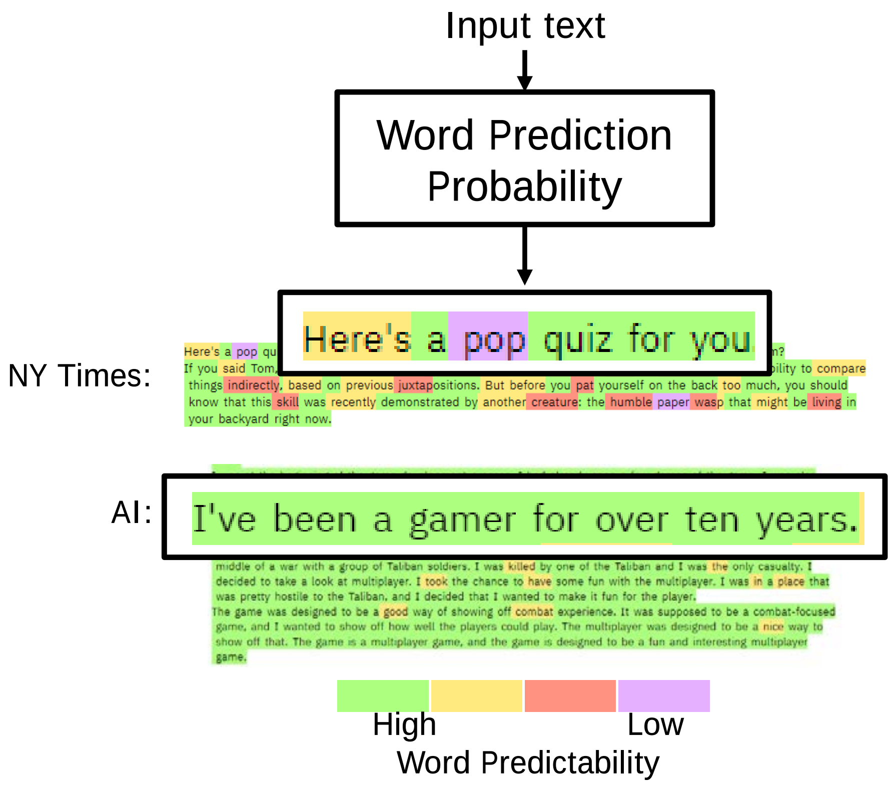
State of the Art Detection is Statistically Based, Narrow, or Both

Audio: ASVspoof



(Lavrentyeva et al. 2017)

Text: GLTR



Here's a pop quiz for you.

Here's a pop qu … n?
If you said Tom, … ility to compare things indirectly, based on previous juxtapositions. But before you pat yourself on the back too much, you should know that this skill was recently demonstrated by another creature: the humble paper wasp that might be living in your backyard right now.

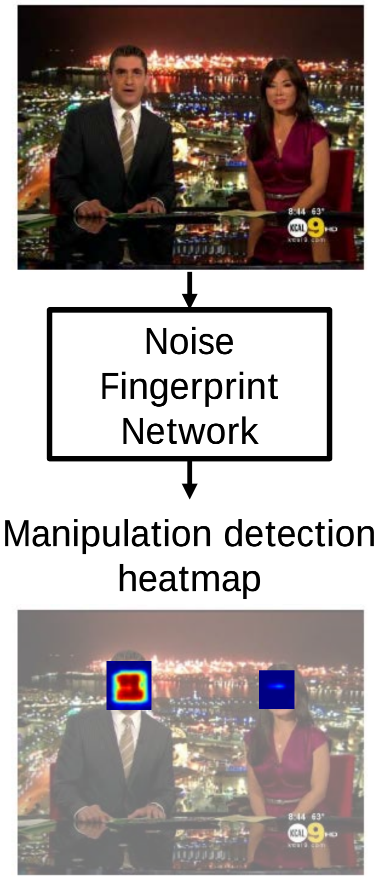
I've been a gamer for over ten years.

middle of a war with a group of Taliban soldiers. I was killed by one of the Taliban and I was the only casualty. I decided to take a look at multiplayer. I took the chance to have some fun with the multiplayer. I was in a place that was pretty hostile to the Taliban, and I decided that I wanted to make it fun for the player.
The game was designed to be a good way of showing off combat experience. It was supposed to be a combat-focused game, and I wanted to show off how well the players could play. The multiplayer was designed to be a nice way to show off that. The game is a multiplayer game, and the game is designed to be a fun and interesting multiplayer game.

AI methods choose more predictable next-words than humans, statistically

(MIT-IBM Watson AI lab, HarvardNLP 2019)

Image/Video: DARPA MediFor



(MediFor: USC/ISI, Univ. Naples 2019)

Expected Threats

Targeted Personal Attacks

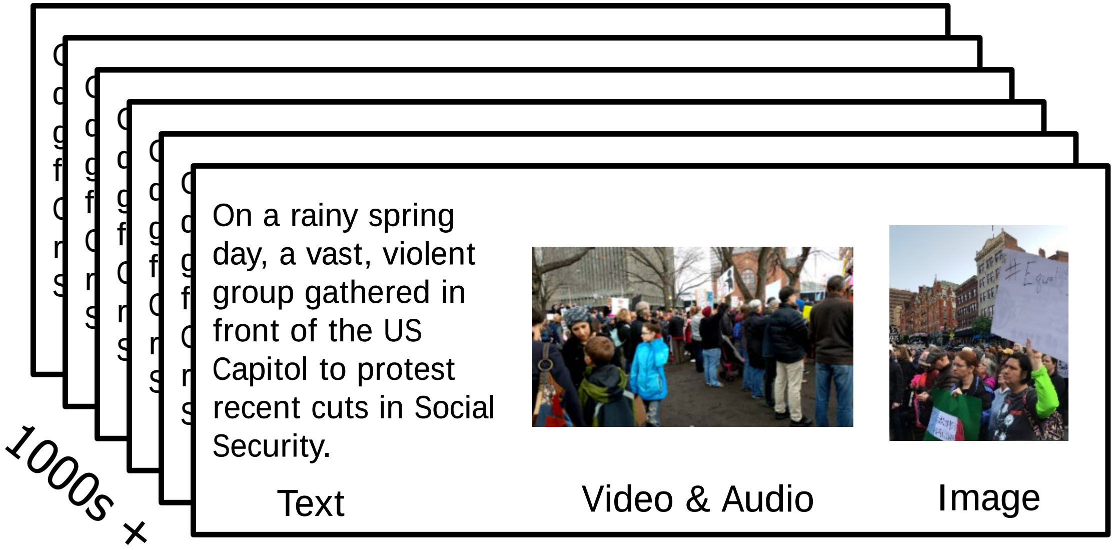
Peele 2017

AI Multimedia Algorithms

Highly realistic video

Generated Events at Scale

AI Multimedia Algorithms

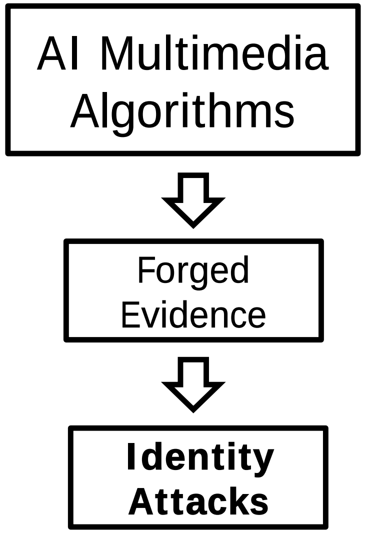


On a rainy spring day, a vast, violent group gathered in front of the US Capitol to protest recent cuts in Social Security.

Believable fake events

Ransomfake concept: Identity Attacks as a service (IAaaS)

Bricman 2019



AI Multimedia Algorithms

Forged Evidence

Identity Attacks

Examples of possible fakes:

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

Undermines key individuals and organizations

Synthetic Media Detection, Attribution, and Characterization Capabilities

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

Semantic Detection

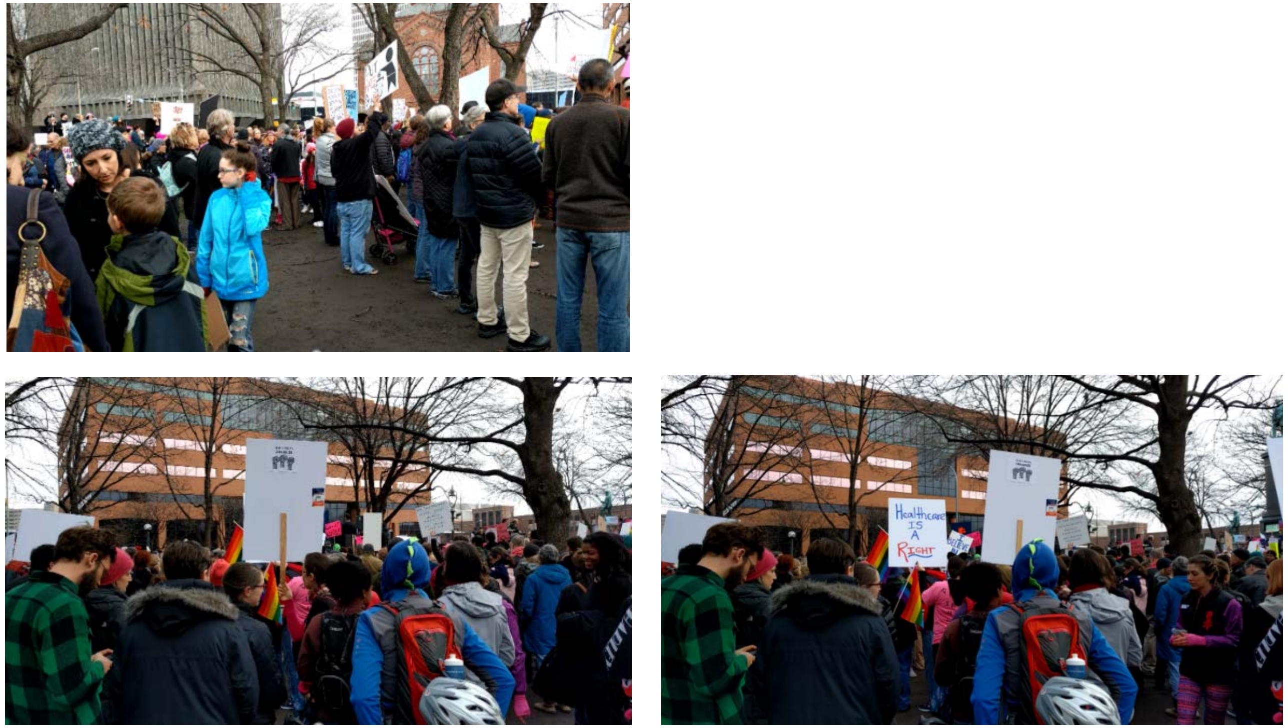
Text (Notional)

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

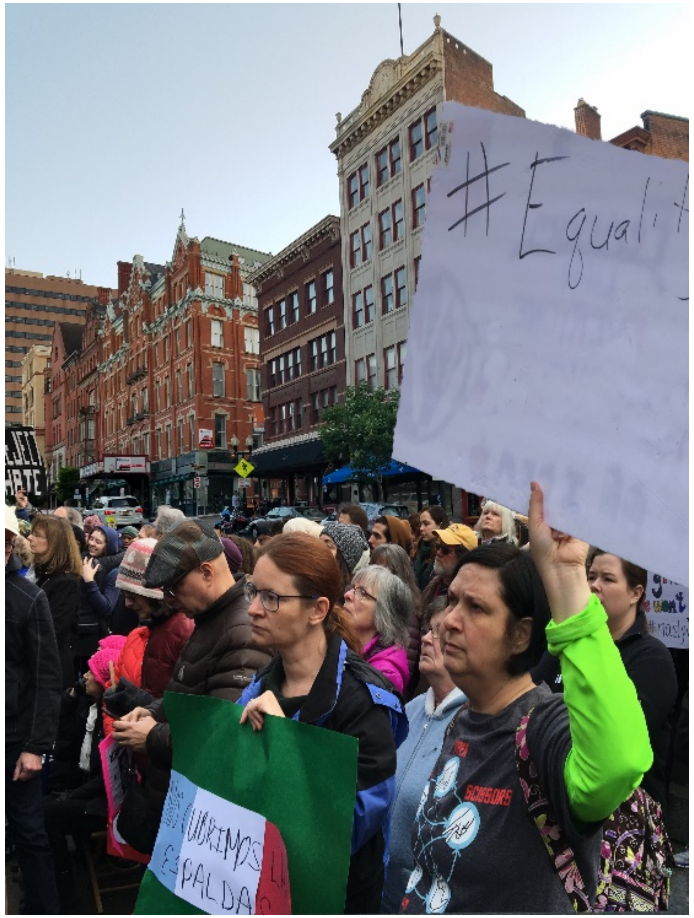
Audio (Notional)

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

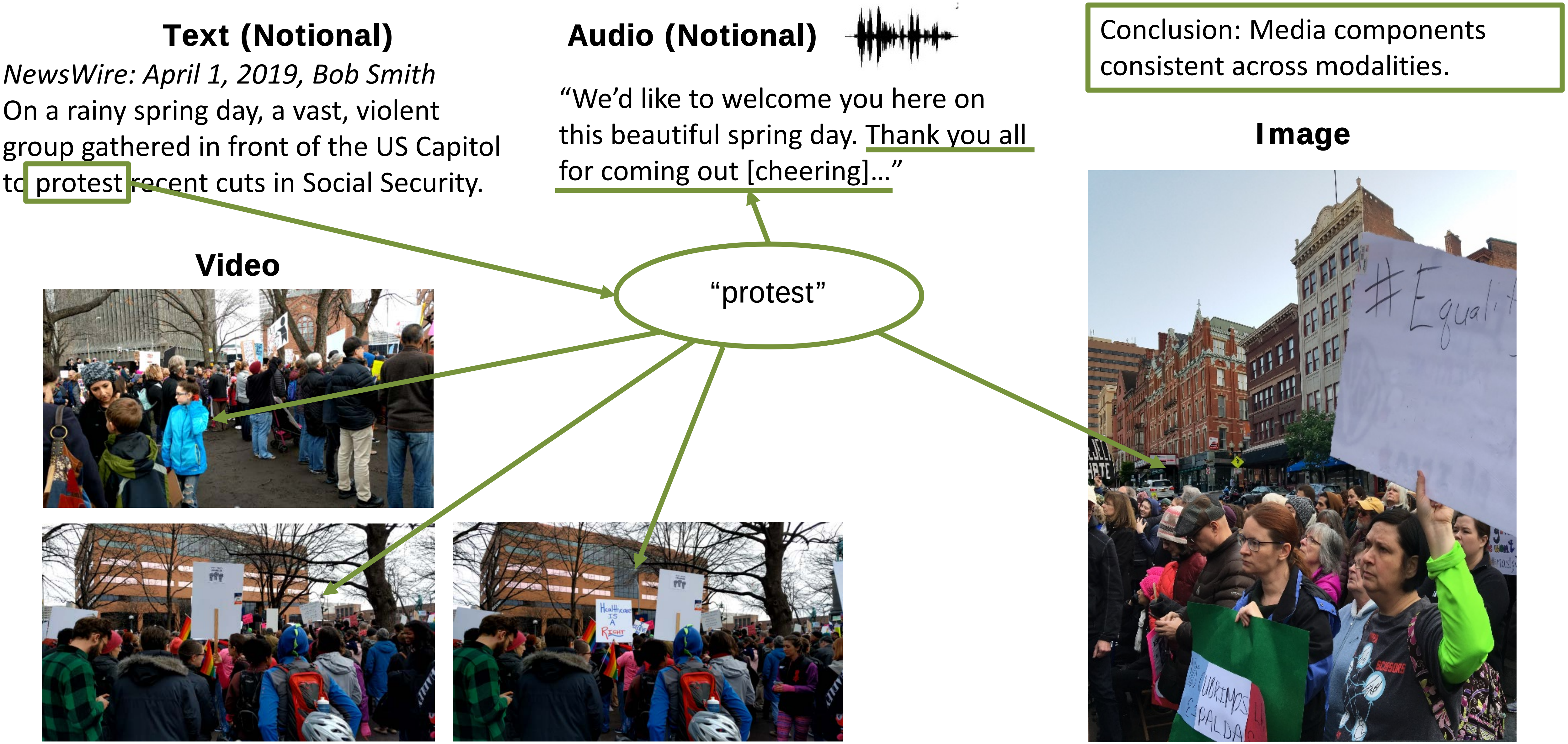
Video



Image



Semantic Detection



Text (Notional)

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

Audio (Notional)

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

Conclusion: Media components consistent across modalities.

Image

Video

"protest"

Semantic Detection

Text (Notional)

NewsWire: April 1, 2019, Bob Smith

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

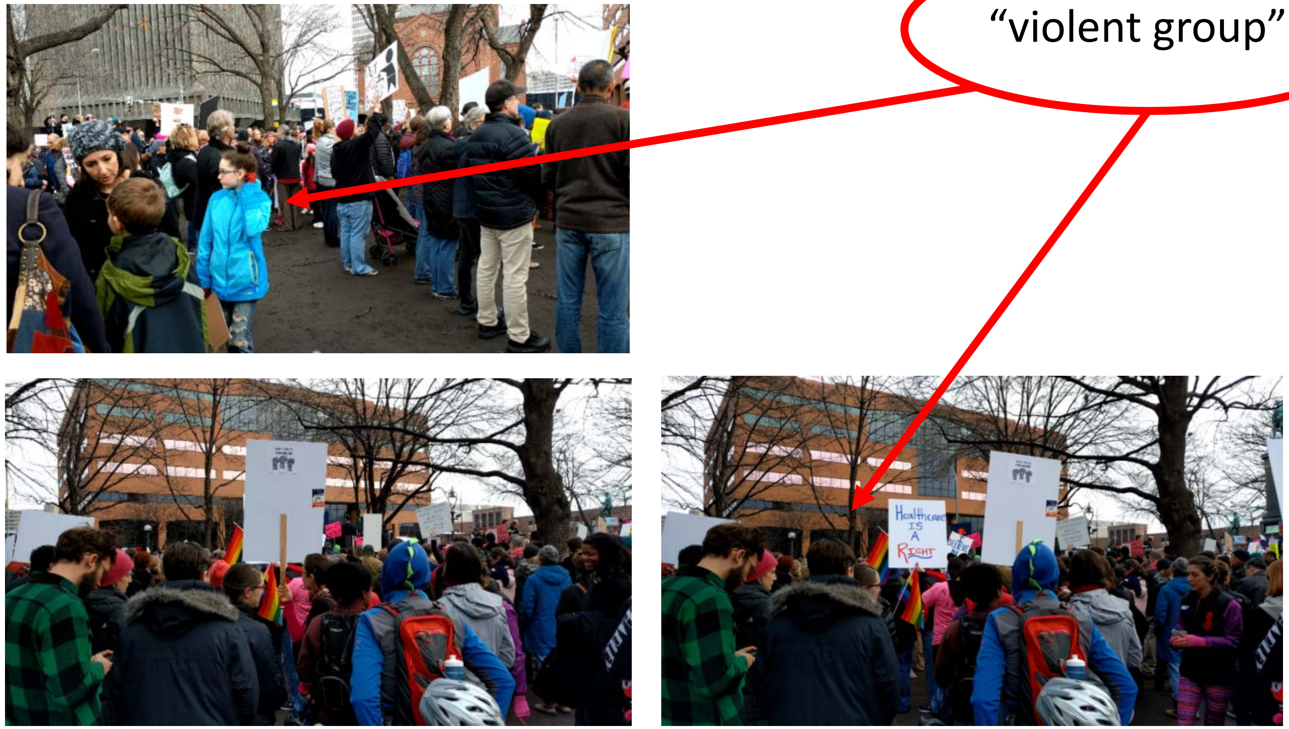
Audio (Notional)

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

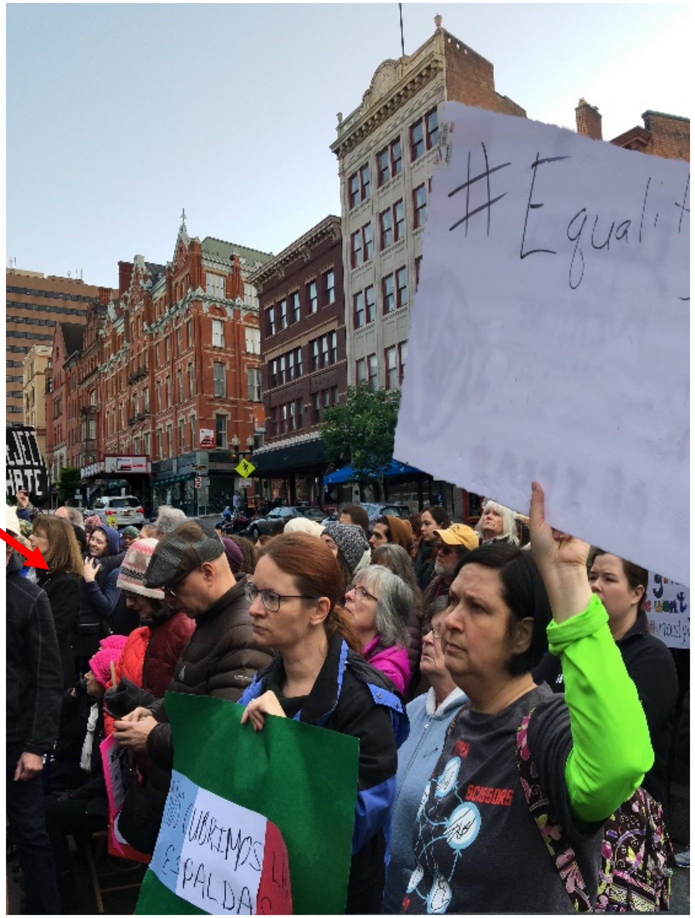
Conclusion: Media components not consistent across modalities.

“violent group”

Video



Image



Semantic Attribution & Characterization

Text (Notional)

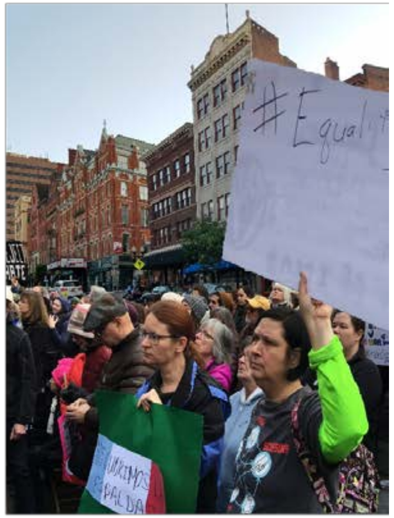
NewsWire: April 1, 2019, Bob Smith
On a rainy spring day, a vast, violent group gathered in front of the US Capitol to protest recent cuts in Social Security.

Video

Audio (Notional)

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

Image



<u>Attribution: Incorrect</u>

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

<u>Characterization: Malicious</u>

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

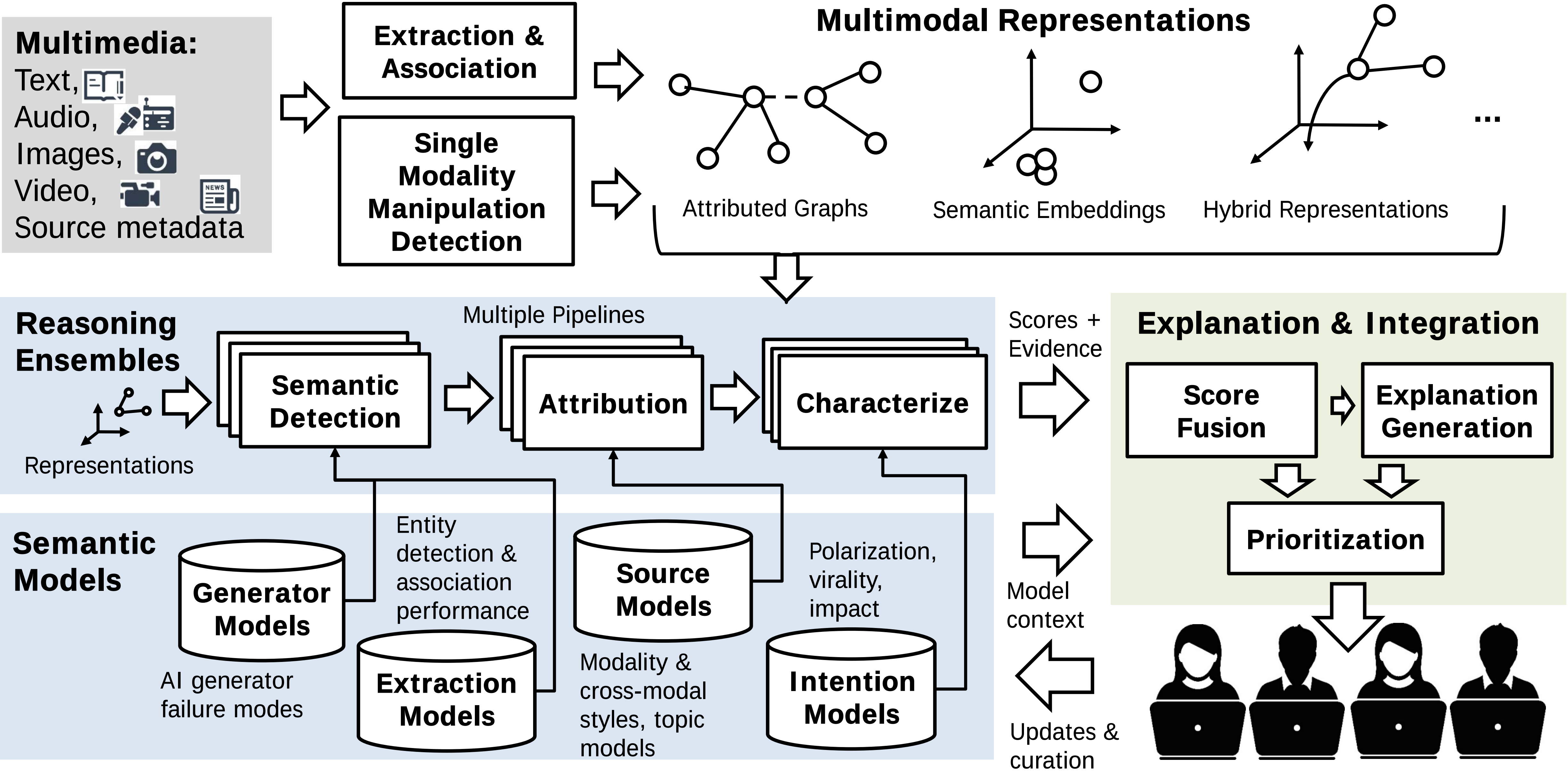
Lexicon (1)

- **Media modalities:** Media forms, examples including text, image, video, and audio.
- **Media asset:** A media instance, such as a single media item and modality: an image, a video, an audio, or text document.
- **Multi-modal asset:** A media collection that may be treated as a single event or instance, such as a news story. May contain some combination of multiple modalities such as image, video, audio, and text.
- **News articles:** A journalist-written story describing an event of interest using multiple modalities. For example, a web page with text and images or video describing an event of interest. News articles are expected to include source organization, an author, and date/time. Some stories may include a location.
- **Social media post:** A short, multi-modal media asset, such as Twitter. Social media posts are expected to be shorter and more colloquial than news articles. Social media posts are expected to include a source platform, an author, and date/time. Depending on social media type (real or generated) they may provide access to the social network of users.
- **Technical information:** A news story, social media post, or technical article describing a technical capability. For example, a news article describing a new ballistic missile capability.

Lexicon (2)

- **News collection:** Multiple news articles describing a single event. Assets will be from approximately the same time period (e.g. hours to a few days).
- **Technical information collection:** Multiple technical information assets. Assets will be from approximately the same time period (e.g. hours to a few days).
- **Falsified media:** Media that has been manipulated or generated.
- **Malicious intent:** In the context of SemaFor, this relates to media that has been falsified to create a negative real-world reaction. For example, falsifying a story to increase its polarization and likelihood to go viral.
- **Media source:** Purported organization that created a media asset (e.g., a newspaper or news channel).
- **Media author:** Purported individual that created a media asset (e.g., the author, actor, photographer, videographer).

Notional SemaFor System



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

Extraction & Association

Single Modality Manipulation Detection

Multimodal Representations

Attributed Graphs

Semantic Embeddings

Hybrid Representations

...

Reasoning Ensembles

Representations

Multiple Pipelines

Semantic Detection

Attribution

Characterize

Scores + Evidence

Explanation & Integration

Score Fusion

Explanation Generation

Prioritization

Model context

Updates & curation

Semantic Models

Generator Models

AI generator failure modes

Entity detection & association performance

Extraction Models

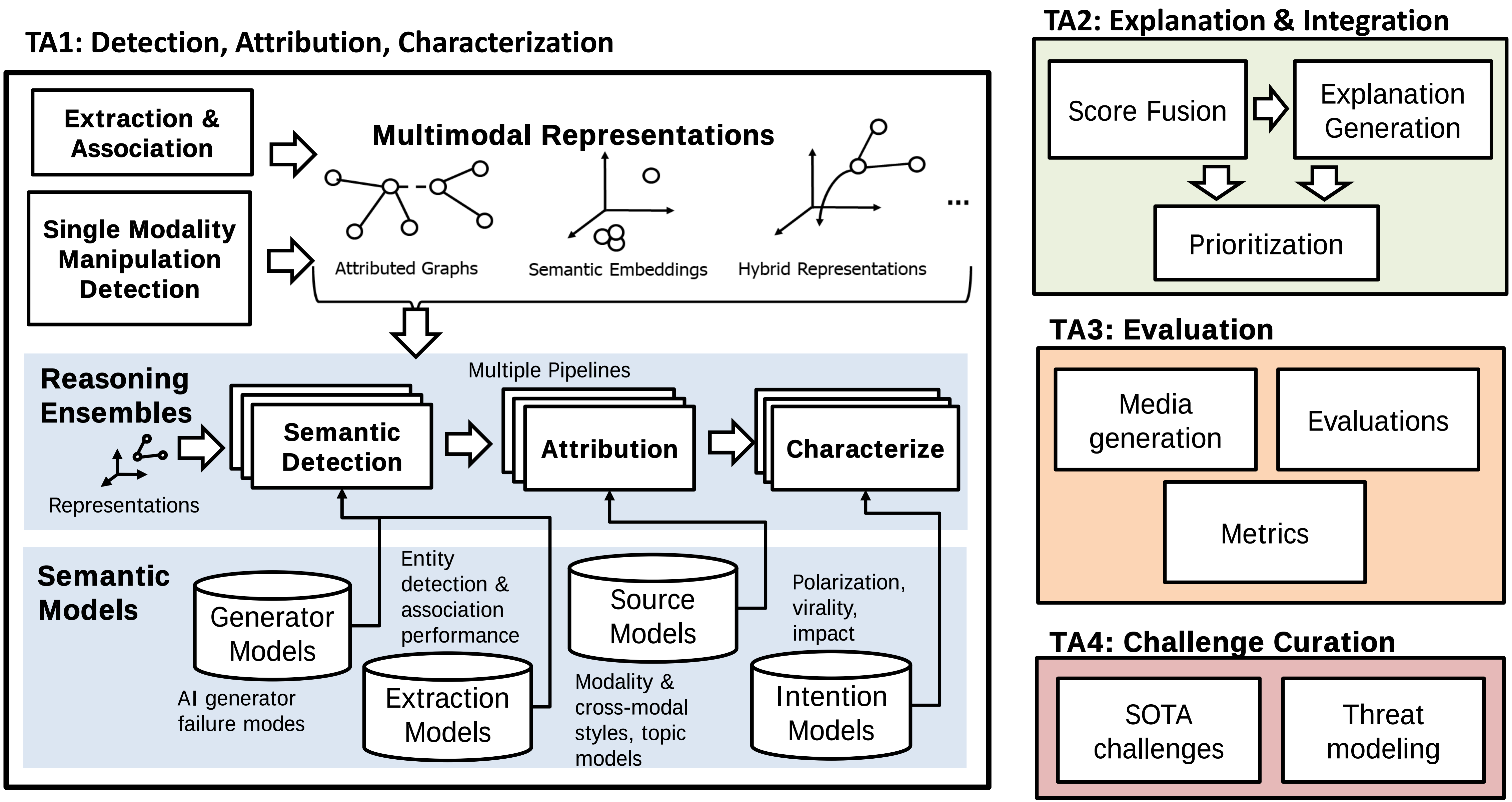
Source Models

Modality & cross-modal styles, topic models

Polarization, virality, impact

Intention Models

Technical Areas



TA1: Detection, Attribution, Characterization

Extraction & Association

Single Modality Manipulation Detection

Multimodal Representations

Attributed Graphs

Semantic Embeddings

Hybrid Representations

...

Reasoning Ensembles

Representations

Multiple Pipelines

Semantic Detection

Attribution

Characterize

Semantic Models

Generator Models

AI generator failure modes

Extraction Models

Entity detection & association performance

Source Models

Modality & cross-modal styles, topic models

Intention Models

Polarization, virality, impact

TA2: Explanation & Integration

Score Fusion

Explanation Generation

Prioritization

TA3: Evaluation

Media generation

Evaluations

Metrics

TA4: Challenge Curation

SOTA challenges

Threat modeling

Detection, Attribution, Characterization (TA1)

Anticipating Multiple Awards

Detection: Examine single and multi-modal media assets and reason about semantic inconsistencies to determine if the media has been falsified

Attribution: Analyze the content of **multi-modal** media asset(s) with respect to a purported source to determine if the purported source is correct

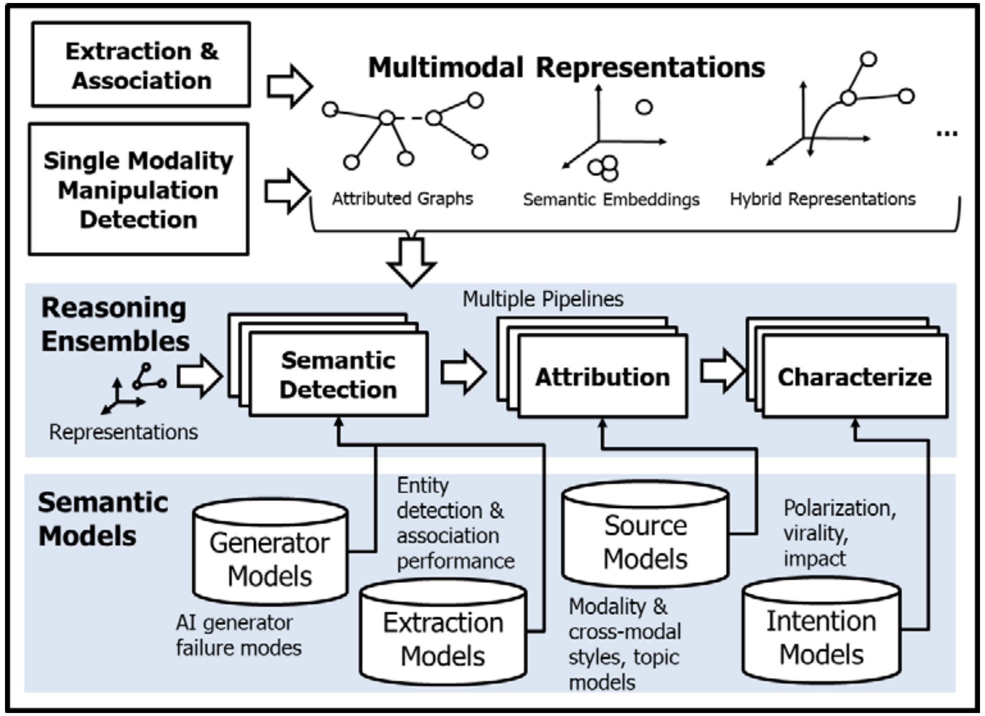
Characterization: Examine the content of **multi-modal** media assets to determine if it was falsified with malicious intent

Challenges:

- Aligning, grounding, and reasoning about entities across multiple modalities, each which may only have a portion of the narrative
- Limited training data and potential for domain mismatch
- Acquiring and incorporating outside semantic knowledge
- Identifying specific types of semantic properties that are applicable to the DAC tasks across media modalities
- Enabling transitioned algorithms to be easily updated as threats and domains evolve

TA1 / TA2 API Interactions

TA1: Detection, Attribution, Characterization



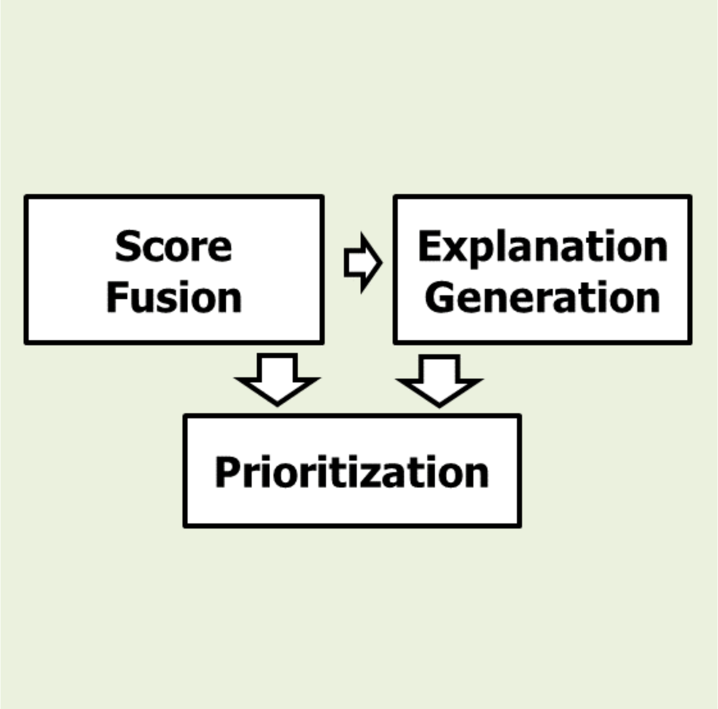
SemaFor API (TA2)

Detection score, evidence →

Attribution score, evidence →

Characterization score, evidence →

TA2: Explanation & Integration



May also need API for TA2 interrogating TA1 semantic models

Strong TA1 proposals will describe

- Approaches to automatically **reason about extraction failures** in one or more modalities that might otherwise indicate spurious inconsistencies across modalities.
- Approaches to **align, ground, and reason** about entities across multiple modalities, each of which might only have a **portion of the overall narrative**.
- Algorithms for DAC that provide **effective performance even with limited training data**, and that are robust against domain mismatch.
- DAC algorithms that could deal with **real-world issues such as multiple cultures and contexts**.
- Techniques for quantitatively **characterizing key aspects of falsified media**, **such as malicious intent**, in ways that are both computationally accessible and operationally relevant.

TA1 Responsibilities

	to TA2	to TA3	to TA4	to the program
TA1 provides	• Input into system API specification/design • DAC algorithm containers implementing API and documentation • DAC scores and evidence via API • Support for DAC container integration • TA1 algorithm insight to support fusion, explanation, and prioritization components • Calibrated scores	• Suggested datasets and evaluation scenarios • Support for designing evaluations	Feedback on challenges and Hackathons	• Development & training data gathered from outside the program • Participation at hackathons and PI meetings • Develop and provide insight into DAC algorithms • DAC algorithm containers implementing API and documentation

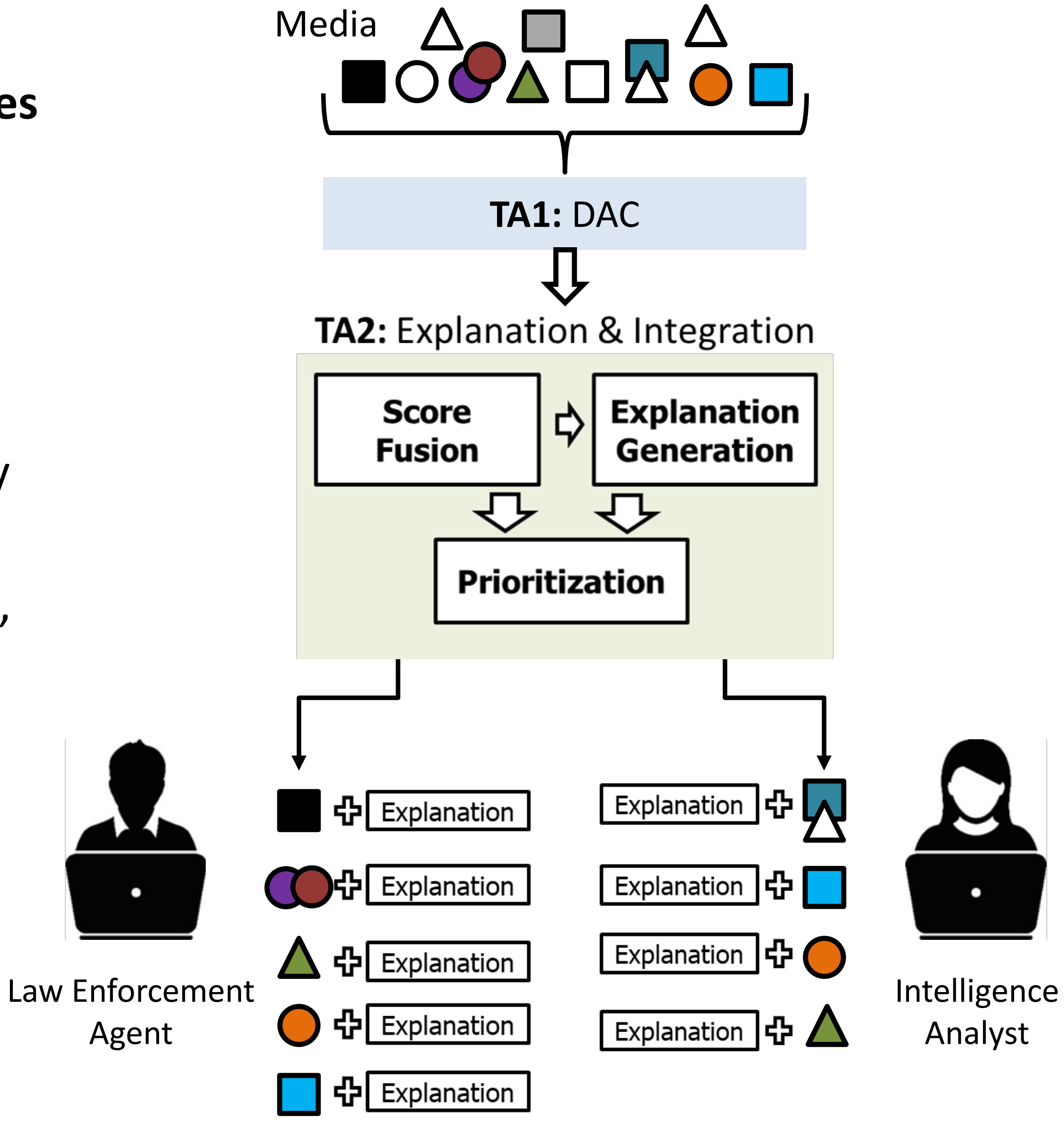
Explanation & System Integration (TA2)

Anticipating Single Award

Create state of the art approaches to prioritize large volumes of multi-modal assets and explaining multi-modal/multi-asset evidence of manipulation

<u>Challenges</u>

- Creating an open, standards-based, multisource, plug-and-play architecture that allows for interoperability and integration
- Developing a single fused score for detection, attribution, and characterization based on scores and evidence for detection, attribution, and characterization provided by TA1s
- Developing a stable prototype system prior to each program evaluation and supporting program demos



Media

TA1: DAC

TA2: Explanation & Integration

Score Fusion

Explanation Generation

Prioritization

Law Enforcement Agent

Intelligence Analyst

Strong TA2 proposals will describe

- Techniques for **fusing DAC scores** across multiple TA1 performers each with disparate approaches.
- Approaches for **reconciling evidence** across multiple TA1 performers with disparate forms of evidence, and presenting a **unified evidence summary and explanation** to end users.
- Methods for automatically customizing media **prioritization schemes for different users** or different classes of users.
- Technical approaches to **enabling parallel TA1 development and system integration** while simultaneously minimizing dependencies and integration effort.
- A strategy for supporting a **rolling, continuous evaluation process** that leverages the prototype SemaFor system and a continuous integration, continuous deployment process while keeping **compute costs in check**.
- An approach for proactively **engaging with potential transition customers** to enable early transition of SemaFor capabilities.
- Evidence of previously successful **transition of DARPA capabilities to operational use** in the DoD and/or IC.

TA2 Responsibilities

	to TA1	to TA3	to TA4	to the program
TA2 provides	• System API specifications designed with TA1 input • Integration of TA1 components into SemaFor system • Compute resources for evaluation of TA1 algorithms • Design input for score calibration process • Lead design of system APIs • Receive, validate, and integrate TA1 components into SemaFor system	• Support for designing evaluations • Compute for evaluation scoring code	Feedback on challenges and Hackathons	• SemaFor system design and APIs • SemaFor system integration and U/I development • Provide compute resources for evaluations, hackathons, and demonstrations • Transition support • Support integration exercises with transition partners • Hosting and leading hackathons • Participation at hackathons and PI meetings • Develop and provide insight into score fusion, explanation, and prioritization algorithms • SemaFor system demonstrations in each program phase • Develop algorithms to assemble and curate evidence; provide unified evidence summary and explanation • Facilitate program design discussions • Provide a stable prototype system prior to each evaluation

Evaluation (TA3)

Anticipating Single Award

Create robust evaluations for detection, attribution, and characterization, and for prioritization and explanation

Challenges

- Designing evaluation protocols to explore the range of SemaFor performance, to highlight where human capabilities might be best augmented by automated algorithms
- Designing an evaluation protocol handling the potential combinatorial complexity of evaluating performers on multiple media and falsification types, in cross-modality media groupings of various compositions
- Identifying relevant metrics to support the evaluation goals
- Generating (or collecting) a sufficient number of media assets to support the multi-modal evaluation

Strong TA3 proposals will describe

- A detailed plan for **obtaining and curating data that is sufficient in volume**, highly **relevant** to the problem domain, and can be **released** to the broader research community during the course of the program, including estimates for how many of each asset type will be needed to support evaluations in each phase of evaluation.
- How the evaluation design will **identify, manage, and decouple latent variables** that might be unintentionally correlated across evaluation probes.
- The evaluation team’s approach to having **strong subject matter expertise** in the detection, attribution, characterization, explanation, and prioritization of falsified multi-modal media.
- How the evaluation design and roadmap will provide both a **comprehensive understanding of the program’s scientific progress** and answer key **performance questions for potential transition partners**.
- Strategies for **designing, organizing, and executing complex evaluation** processes across a large distributed team while maintaining performer buy-in and evaluation integrity.
- Approaches for **streamlining the human subjects research** and IRB process related to evaluation.

TA3 Responsibilities

	to TA1	to TA2	to TA4	to the program
TA3 provides	• Sample development and evaluation data	• Sample development and evaluation data	• Input to challenges and Hackathons	• Media generation and curation • Facilitate and lead program discussions about evaluation designs (datasets, processes, schedule, metrics, transition partner use cases) • Define and implement metrics • Design and conduct experiments to establish baseline human performance • Evaluation scoring software • Evaluation results analysis • Organize and host PI meeting • Oversight of PI meetings • Conduct evaluations every 8 months

Challenge Curation (TA4)

Anticipating Multiple Awards

Multi-modal manipulations collected from the public state-of-the-art

Challenges include:

- Developing state of the art media falsification challenges to support strong SemaFor defenses
- Creating media falsification threat models based on current and anticipated technology

Strong proposals will describe:

•Detailed evidence of the proposer's ability to **bring state-of-the-art falsification challenges** in one or more modalities to the program.

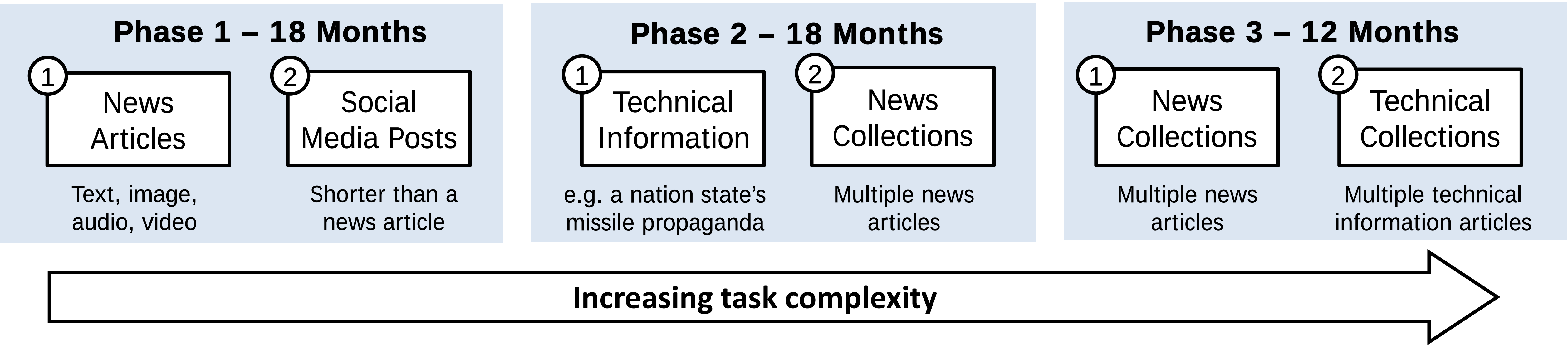
•**Threat models that provide actionable insights** into how DAC algorithms and the SemaFor system should be designed to put significant burdens on potential manipulators.

TA4 Responsibilities

	to TA1	to TA2	to TA3	to the program
TA4 provides	• Coordination in advance of and during Hackathons to ensure challenge understanding	• Coordination in advance of and during Hackathons to ensure challenge understanding	• Support for incorporating challenge problems into evaluations • Regularly deliver challenges and updated threat models • Work with TA3 to curate additional generated or manipulated data for challenge problems • Work with TA3 to evaluate progress on challenge	• State of the art falsification techniques • Curate SOTA challenges from public domain • Develop threat models • Provide insight as to whether/how DAC technologies could be most effective • Challenge problem design • Lead challenge problem execution at hackathons • Participation at hackathons and PI meetings

Distribution A: Approved for public release. Distribution unlimited.

Measuring Progress



Phase 1 – 18 Months

1 News Articles

Text, image, audio, video

2 Social Media Posts

Shorter than a news article

Phase 2 – 18 Months

1 Technical Information

e.g. a nation state's missile propaganda

2 News Collections

Multiple news articles

Phase 3 – 12 Months

1 News Collections

Multiple news articles

2 Technical Collections

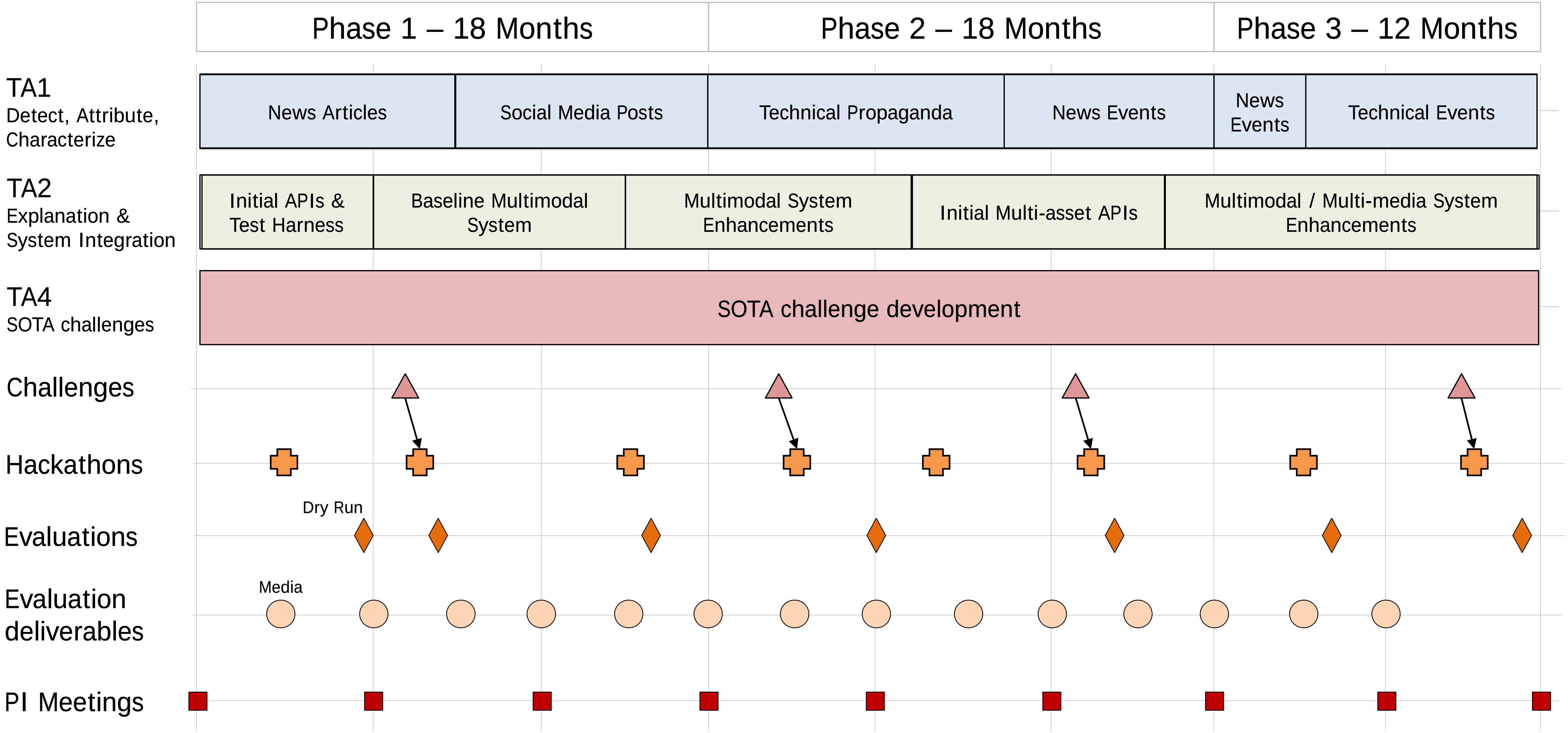
Multiple technical information articles

Increasing task complexity

Quantitative Assessment

			Program Goals		
Task	**Metrics**	**Relevant Baselines**	**P1**	**P2**	**P3**
Manipulation detection	• Probability of Detection (Pd) • False Alarm Rate (FAR) • Equal Error Rate (EER)	• Human: 60% Pd [Deepfakes] • Image: 80% Pd at 10% FAR / 20% EER • Text entity recognition: 90% F1-score • Audio: 4% EER	80% Pd 10% FAR	85% Pd 8% FAR	90% Pd 5% FAR
Attribution	• Pd / FAR	• Image: 78% Pd at 10% FAR [camera id]	80% Pd 10% FAR	85% Pd 8% FAR	90% Pd 5% FAR
Prioritization for analyst	• Accuracy over degrees of malice	• Sentiment analysis: 70-80% F1-score	70% accuracy	80% accuracy	85% accuracy

Schedule



Phase 1 – 18 Months | Phase 2 – 18 Months | Phase 3 – 12 Months

TA1 Detect, Attribute, Characterize: News Articles; Social Media Posts; Technical Propaganda; News Events; News Events; Technical Events

TA2 Explanation & System Integration: Initial APIs & Test Harness; Baseline Multimodal System; Multimodal System Enhancements; Initial Multi-asset APIs; Multimodal / Multi-media System Enhancements

TA4 SOTA challenges: SOTA challenge development

Proposal Information

- Proposers may submit proposals to all TAs.
- Each proposal may only address one TA.
- Separate proposals for each TA are required if proposing to multiple TAs.
- DARPA will not make TA1 and TA2 awards to the same institution.
- TA3 performer may not perform on TA1 or TA2 due to an inherent conflict of interest with the evaluation process.
- TA4 institutions may perform on other parts of the program, but organizational conflicts of interest plans will be needed in the case of TA1 or TA2 due to potential conflicts of interest with the evaluation process.

DARPA

www.darpa.mil