情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

2019年10月28日
デジタル・フォレンジック研究会
法務・監査分科会

フェイクニュースと メディア・フォレンジック

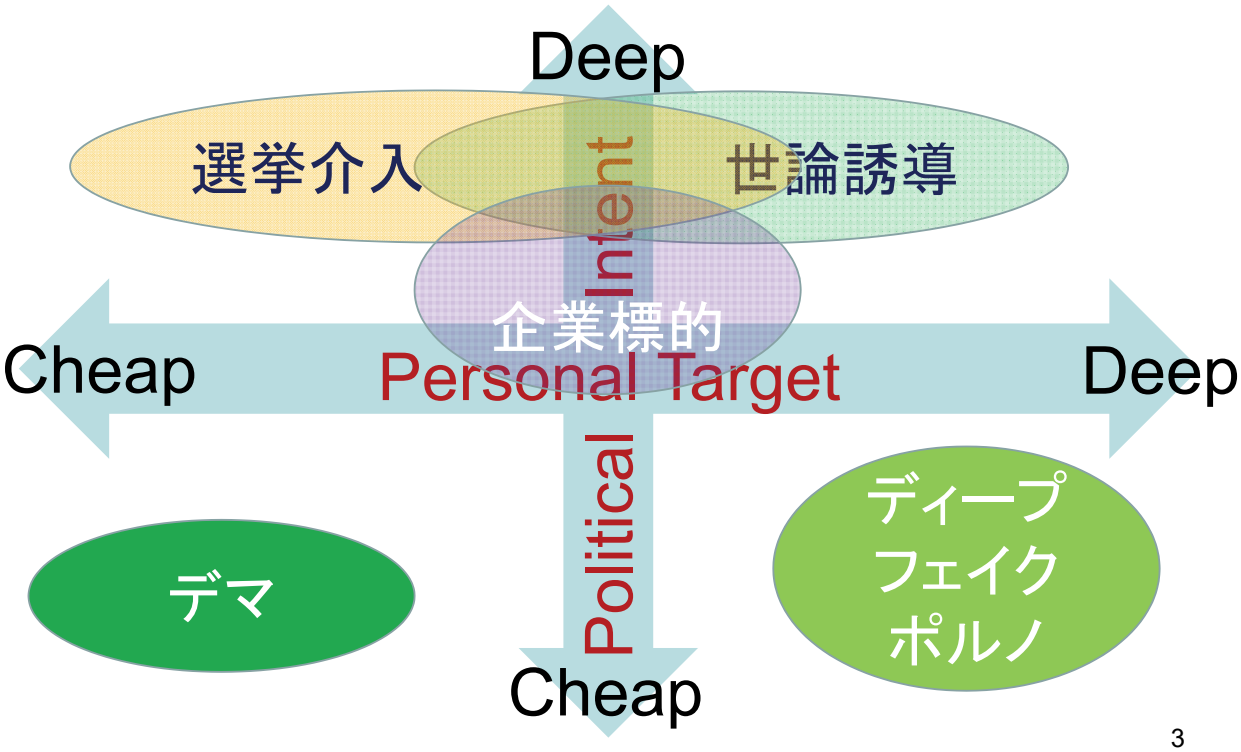
湯淺　墾道
（情報セキュリティ大学院大学）

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

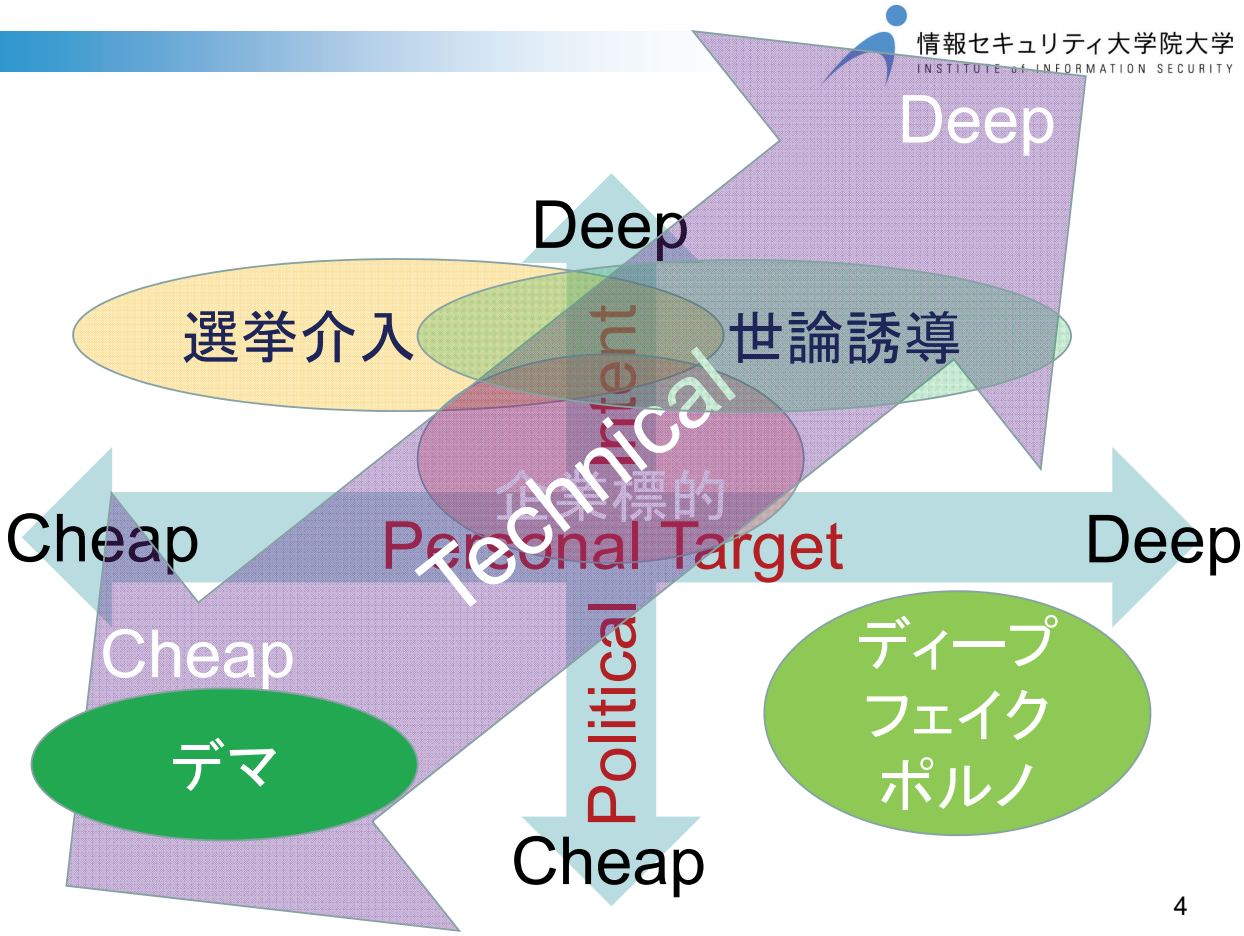
現状

Fakenews, Disinformation

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY



情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY



選挙セキュリティにおける位置づけ

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

■“Election Security”

- 第1段階
  - 有権者の民意形成への介入と世論誘導によって選挙結果に影響を与えようとする段階
  - 政党、候補者へのサイバー攻撃と情報の暴露
  - フェイクニュース、個人情報を利用したマイクロターゲティング
- 第2段階
  - 投票所を案内したり開票結果を公表したりする選挙管理機関のウェブサイトへの攻撃や選挙に関するニュースサイトへの攻撃等によって選挙に混乱をもたらそうとする段階
- 第3段階
  - 選挙管理機関へのサイバー攻撃や電子投票機へのサイバー攻撃等によって有権者名簿や投票記録それ自体を改ざんする等、直接的に選挙結果を操作しようとする段階

デジタル・ゲリマンダーにおける位置づけ

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

1. ビッグデータ分析と世論操作
   1. 2012年のアメリカ大統領選挙においてオバマ陣営が活用し、再選の一因
2. 感情伝染実験

   Zittrainによるデジタル・ゲリマンダー批判
3. 検索結果操作
4. サイバー攻撃を通じた選挙への介入
5. 偽ニュースの流布を通じた世論介入
6. 地理的ゲリマンダーの高度化

背景

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

■Zeynep Tufekciの指摘

- きわめて厖大な量のデータの収集が可能
- プロファイリング技術や分析技術の進歩により、政治的なターゲットは、「ある特性を持つ集団」から個々人へ
- 個人に対して直接アンケート調査等を実施して回答を得ることなく、特定個人の思想や政治的傾向が収集可能
- 行動科学の深化によって人間の行動を「合理的人間」モデルをこえて予測することが可能に
- 実験をリアルタイムで容易に実施できる（c.f. アメリカにおける実験政治学の盛行）
- データを操作するアルゴリズムは企業の営業秘密の壁の中にあり不透明

ネット広告の変遷

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

■検索

- ユーザーの意図のすくい上げ

■ソーシャル・コンテンツマーケティング

- ユーザーの興味関心・行動の可視化

■アドマーケットプレイス

- 商品売買、データの流通

■データエクスチェンジ

- 購買者属性・行動データと商品・価格データの統合

各国のフェイクニュース規制

規制の対象・態様

- ■ 規制する内容
  - ●選挙干渉・世論誘導
  - ●違法情報発信全般（ヘイトスピーチ等）
  - ●ディープフェイクポルノ
- ■ 規制する対象
  - ●プラットフォーマー規制
    - ◆共同規制
    - ◆直接規制
  - ●使嗾者（外国政府等）制裁

		規制する内容		
		選挙干渉・世論誘導	ヘイトスピーチ等	虚偽情報発信全般
規制する対象	プラットフォーマー規制（共同規制）	EU		
	プラットフォーマー規制（直接規制）	ドイツSNS法 フランス情報操作との戦いに関する法律	ドイツSNS法	マレーシア シンガポール（＋情報発信者規制）
	使嗾者（外国政府等）制裁	アメリカ大統領令13848		

ディープフェイクポルノ規制

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

■ヴァージニア州
- ●リベンジポルノ規制法
- ●改正法案(HB2678)可決
- ●2019年7月施行

■ § 18.2-386.2.
- ●falsely created videographic or still image
- ●フェイクも含む他人のポルノ動画・画像を本人の許可なく流布する行為を規制
- ●ISP免責

■ §18.2-386.2.他人の画像の違法な流布又は販売、違約金

■ A. 人を強要し、嫌がらせをし、又は脅迫する目的で、虚偽に作られたビデオ画像又は静止画像を含むあらゆる方法により作られたビデオ画像又は静止画像であって、性器、恥骨領域、殿部又は女子の乳房を露出させるために全くヌードとなっている者を描写したものを故意に流布し、又は販売する者は、当該ビデオ画像又は静止画像を流布し、又は販売することが許可されていないことを知り、又は知るに足りる相当の理由があるときは、第一級軽犯罪とする。ただし、本条により禁止される行為を行うに際し、インターネットサービスプロバイダ、電子メールサービスプロバイダ、その他の情報サービス、システム又はアクセスソフトウェアプロバイダのサービスを利用する者が、複数のユーザによるコンピュータサーバへのコンピュータアクセスを提供又は可能にする場合は、当該プロバイダは、他人が提供したコンテンツについて、本条に違反する責任を負わない。

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

■ B. 本条に基づく訴追のための場所は、違法行為が発生した管轄地内、又は何らかの手段により創作されたビデオ画像若しくは静止画像が本条に違反して製作、複製、発見、保存、受領若しくは所有された場所とすることができる。

■ C. 本条の規定は、他の法律による訴追を妨げない。

日本

- 総務省
- 2019年5月24日プラットフォームサービスに関する研究会（第8回）
  - フェイクニュース、虚偽情報流布の実情
  - EU、フランス、ドイツのフェイクニュース規制について議論
  - ファクトチェックの取組について議論
  - http://www.soumu.go.jp/main_content/000621621.pdf

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

使嗾者（外国政府等）制裁型：アメリカ

重要インフラ指定

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

■2017年1月6日

■選挙管理システムが国土安全保障省により重要インフラストラクチャー指定

- 有権者登録データベース及び関連する情報通信システム
- 選挙管理に使用される情報通信インフラ及びシステム（投票結果の開票、集計及び表示システム、選挙後の選挙結果検証報告用のシステムなど）
- 投票システム及び関連するインフラ
- 選挙管理及び投票システム用のストレージ装置
- 期日前投票所を含む投票所

大統領令13848

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

■2018年9月12日

■「アメリカ合衆国内の選挙への外国の干渉が発生した場合に一定の制裁を課す大統領令(Executive Order on Imposing Certain Sanctions in the Event of Foreign Interference in a United States Election)」

- 海外からのサイバー空間を利用した態様を含めた選挙干渉に対し、連邦政府が調査を行うことを義務づけ
- 干渉が明らかになった場合に経済制裁措置を発動することを規定

■ 調査対象
- 一義的には合衆国選挙（連邦選挙）への外国政府またはその代理による選挙干渉

■ 調査の主体
- 国家情報長官
- 選挙の後に45日以内で調査を行うことを命じられている

■ 選挙インフラストラクチャー
- 連邦政府若しくは州若しくは地方公共団体により、又はされる情報通信これらに代わって選挙プロセスを管理するために使用技術及びシステム
- 選挙人登録データベース、投票機、投票集計機器及び選挙結果の安全な伝達のための機器を含む

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

■「外国からの干渉」
- 「選挙に関し、外国政府又は外国政府の代理人若しくは代理として行動する者の隠ぺい的、詐欺的、欺瞞的若しくは不法な行為又は企てであって、選挙への影響、選挙の結果若しくは報告の結果に対する信頼を傷つけ、若しくは変更し、又は選挙の過程若しくは制度に対する国民の信頼を損なう目的若しくは効果を有するものを含む。」

■「外国政府」
- 「合衆国以外の国において、国、州、地方その他の統治当局、政党又は統治当局若しくは政党の職員をいう。」

■ 制裁
- 経済制裁に限定（外交官追放等は含まれない）

2018年中間選挙

■ 2018年12月21日

- 大統領令13843に基づく初の報告書を大統領に提出したと声明
- ロシア、中国、イランその他の国々はその戦略的利益を増進させるために影響流布活動を行っている
- 2018年中間選挙の結果にそれらが影響を与えたかどうかは、調査を行わなかった
- 「情報機関に課せられている任務は外国のアクターの意図、能力及び行動を監視及び調査することであり、アメリカの政治過程や世論を分析することではない」
- 大統領令が、「選挙の過程若しくは制度に対する国民の信頼を損なう目的若しくは効果」があるものを含むとしていることに比べ、その権限行使を限定的に解釈

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

メディア・フォレンジック

MediFor

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

■Media Forensics = MediFor

- 国防高等研究計画局（DARPA = Defense Advanced Research Projects Agency）
- 2015年にメディア・フォレンジック研究プロジェクトを開始
- 動画像及び静止画像に加工が加えられていることの検出を主な対象
- パーデュー大学を拠点とする民間の研究者チームに対して研究費を交付して研究
- 加工の有無、加工履歴、加工箇所などを自動的に検出するツールの作成を目標

SemaFor

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

■Semantic Forensics　SemaFor

- 加工されたデータ（だけ）の分析には限界
  → semantic分析手法へ
- 音声、動画、静止画像、テキスト
- ネットワークのアトリビューション
- 加工されたデータの背後にあるのは誰か、目的は何か、方法は何かの分析

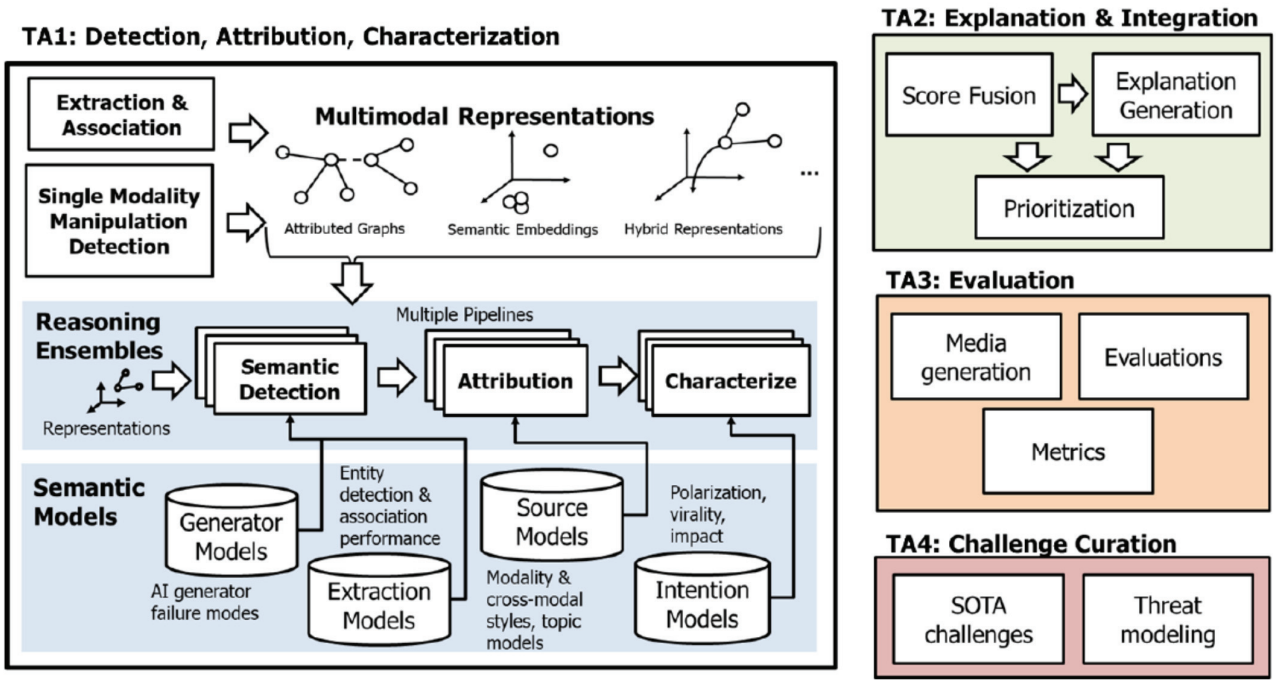
- Michael Spranger, Stefan Schildbach, Florian Heinke, Steffen Grunert and Dirk Labudde, *Semantic Tools for Forensics: A Highly Adaptable Framework,* IMMM 2012 : The Second International Conference on Advances in Information Mining and Management, https://pdfs.semanticscholar.org/0269/f1bfd314dc7ed416a96cc8d737ac1b8e5c4d.pdf

SemaFor

■DARPA

- ●Semantic Forensics　SemaFor
- ●Medi Forの次期プロジェクト
- ●加工の検知、アトリビューション、分析を自動的に行うツール開発
- ●加工されたデータの背後にあるのは誰か、目的は何か、方法は何かの分析
- ●検知（Detection）、アトリビューション（Attribution）、特性評価（Characterization）

https://www.darpa.mil/news-events/2019-09-03a



https://www.fbo.gov/utils/view?id=614a021dc47a5466ce1747e3351113f4

■TA1 Detection, Attribution, Characterization

- 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.

https://www.fbo.gov/utils/view?id=614a021dc47a5466ce1747e3351113f4

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

■TA2 Explanation and Integration

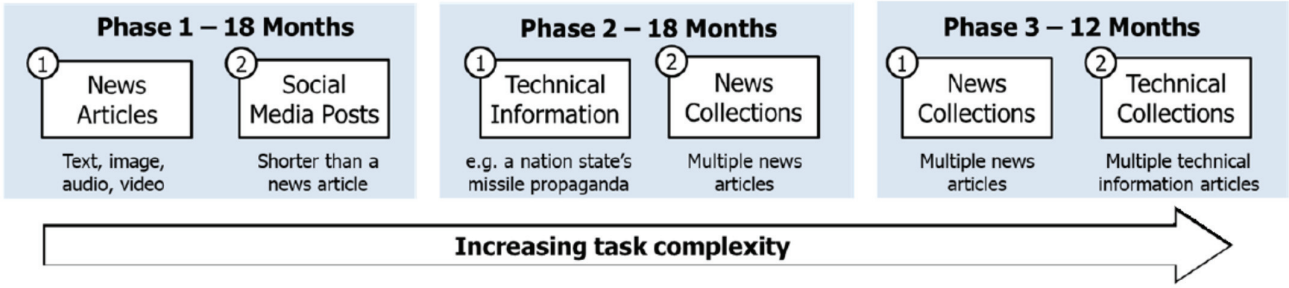
- Techniques for fusing DAC scores across multiple TA1 performers each with disparate approaches.
- Approaches for reconciling evidence across multiple TA1 performers with disparate forms of evidence, and presenting a unified evidence summary and explanation to end users.
- Methods for automatically customizing media prioritization schemes to different end users or different classes of end 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.

- A strategy for storage of training and evaluation data as well as initialization values, hyper-parameters, training and evaluation process scripts, documentation, validation tests, knowledge-bases, and any other materials required for the Government to reproduce or retrain any algorithmic component. Data will need to be stored securely and also be able to be compartmentalized to ensure that evaluation data is kept separate from training data.
- 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.

https://www.fbo.gov/utils/view?id=614a021dc47a5466ce1747e3351113f4

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

■TA3 Evaluation



		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

https://www.fbo.gov/utils/view?id=614a021dc47a5466ce1747e3351113f4

■TA4 Challenge Curation

- 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.

https://www.fbo.gov/utils/view?id=614a021dc47a5466ce1747e3351113f4

SemaForで可能になること

	必要能力	現在	SemaFor
検知	世代・操作エラーの意味論的検知	限定的	○
	モダリティ及びアセットを横断する操作の検知	限定的	○
	操作アルゴリズムに対するロバスト	脆い	強固
	検知アルゴリズム無効化に対する対抗措置の増大	若干	重要
アトリビューション	ソース又は発信者の自動検証	限定的	○
	ソースの特異なフィンガープリントの自動検証	×	○
	作成者の不一致の摘示	×	○
特性評価	操作の意図や影響の自動検証	×	○
	操作意図の挙証と摘示	×	○
	操作された又は生成されたメディアのレビューへの正しい優先性付与	×	○

https://www.darpa.mil/attachments/SemanticForensics-IndustryDay-2019-08-12.pdf より

DARPA 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]..."

Image

Video

https://www.darpa.mil/attachments/SemanticForensics-IndustryDay-2019-08-12.pdf

DARPA 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]..."

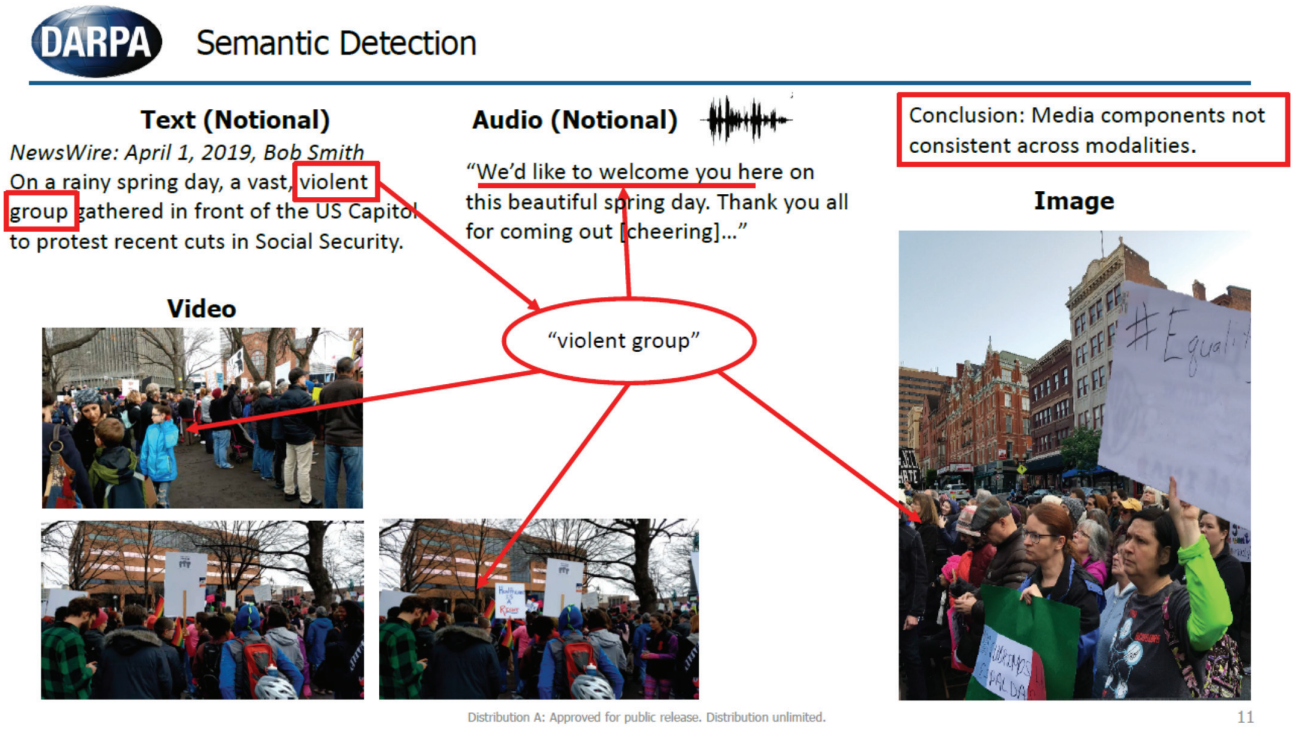
Conclusion: Media components consistent across modalities.

Image

Video

"protest"

https://www.darpa.mil/attachments/SemanticForensics-IndustryDay-2019-08-12.pdf



https://www.darpa.mil/attachments/SemanticForensics-IndustryDay-2019-08-12.pdf

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

対策の方向

- 第1段階
  - 有権者の民意形成への介入と世論誘導によって選挙結果に影響を与えようとする段階
  - 政党、候補者へのサイバー攻撃と情報の暴露 仏では情報機関支援が候補者支援(?)
    英では政党を支援
  - フェイクニュース、個人情報を利用したマイクロターゲティング 個人情報保護法制（Cokie規制等）
- 第2段階
  - 投票所を案内したり開票結果を公表したりする選挙管理機関のウェブサイトへの攻撃や選挙に関するニュースサイトへの攻撃等によって選挙に混乱をもたらそうとする段階 地方公共団体のセキュリティ強化
- 第3段階
  - 選挙管理機関へのサイバー攻撃や電子投票機へのサイバー攻撃等によって有権者名簿や投票記録それ自体を改ざんする等、直接的に選挙結果を操作しようとする段階 選挙システムの重要インフラ指定（米指定済、ENISA勧告）

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

- 本発表はサントリー文化財団2019年度研究助成「デジタル民主主義と選挙干渉：日本・アジアにおける選挙干渉のリスクと脆弱性」の成果の一部です