インターネットにおける
ディスインフォーメーション対策
「Trustable Internet」

富士通株式会社
データ＆セキュリティ研究所
今井　悟史



これまでの10年

- 過去10年の間に、デジタルテクノロジーによって人々の生活やビジネス、社会が大きく変化
- 新型コロナウイルスのパンデミックを経て、デジタルテクノロジーを使った生活、ビジネスや学習のスタイルがニューノーマルに
- フェイクニュースの氾濫やプライバシーの侵害、テクノロジーの信頼性に対して懸念が拡がり、社会の信頼を再構築することが急務



背景

- インターネットは社会・経済活動に不可欠
- 一方、正しいかどうか判断できない情報が増えつつある
  - フェイクニュースによる経済損失は年間780億ドル（約10兆円）*1

*1: Fake news impose a cost on the global economy of at least $78 billion per year:
THE ECONOMIC COST OF BAD ACTORS ON THE INTERNET FAKE NEWS | 2019
https://s3.amazonaws.com/media.mediapost.com/uploads/EconomicCostOfFakeNews.pdf

事例：虚偽の情報による混乱

自然災害の例

- 静岡県の台風15号による水害被害のフェイク画像がTwitter上で拡散

- 9/26AM4:39
  - フェイク画像投稿
- 投稿後
  - 濁流の流れや一部の建物に不自然な部分があり、投稿に対して「画像生成AIが作成した偽物ではないか？」など疑問の声
- 9/26PM4:00
  - 投稿者は問題の画像がフェイクだと認める
- 9/27
  - ファクトチェックイニシアティブ(リトマス/BuzFeed)で虚偽と判定

事例：クライシスアクターの嫌疑

FUJITSU

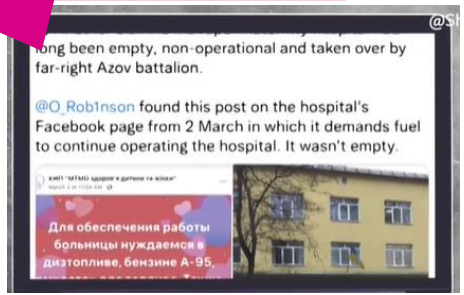
ウクライナ侵攻の例

SNSへの投稿について親ロシア派はフェイクと主張
→ BBCニュースはファクトチェックして報道

マウリポリの妊婦の写真

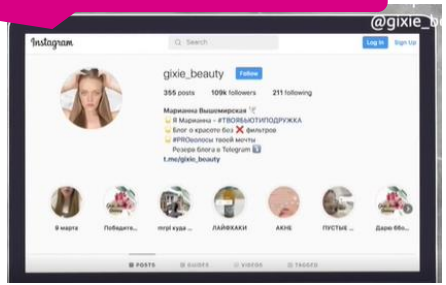
https://www.bbc.com/japanese/video-61103361

爆発のあった病院の詳細を調査



攻撃されるまで病院は稼働していたことが分かった

女性のインスタグラムを特定



過去の写真から妊娠してマウリポリにいたことが分かった

事例：偽情報による世論誘導

「ウクライナで「赤ちゃん工場」を発見」＝＞**証拠なし**

「ストームシャドウミサイル（イギリス供与）の誤作動によりウクライナ軍の兵舎を壊滅」＝＞**証拠なし**

ゼレンスキー大統領が選挙を「中止した」とする投稿
＝＞**世論の誤解を誘導**

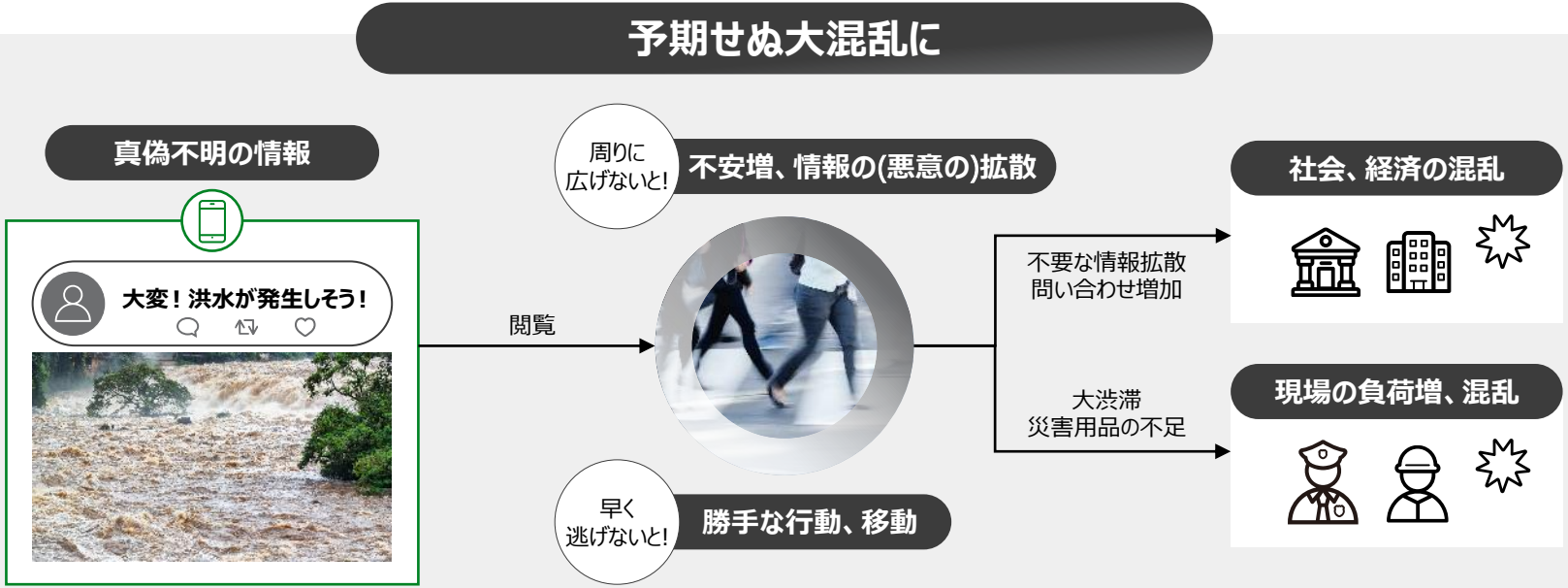
＊引用元（BBCニュースサイト）：https://www.bbc.com/japanese/features-and-analysis-66151264

ネット上での不確かな情報の影響

FUJITSU

一度広まると・・・

- 不安感からさらに情報が拡散される
- 情報を確認する問い合わせや、不確かな情報による行動で大混乱に



解決に向けたアプローチ

- インターネット情報に対し、第三者の情報/評価などの根拠を紐づけ、情報の真偽を判断する
- 自治体やニュース、カメラやセンサーなど信頼のおける第三者の情報が根拠になる世界

元の情報に根拠として紐づけ

第三者による根拠

主観による情報

客観的事実

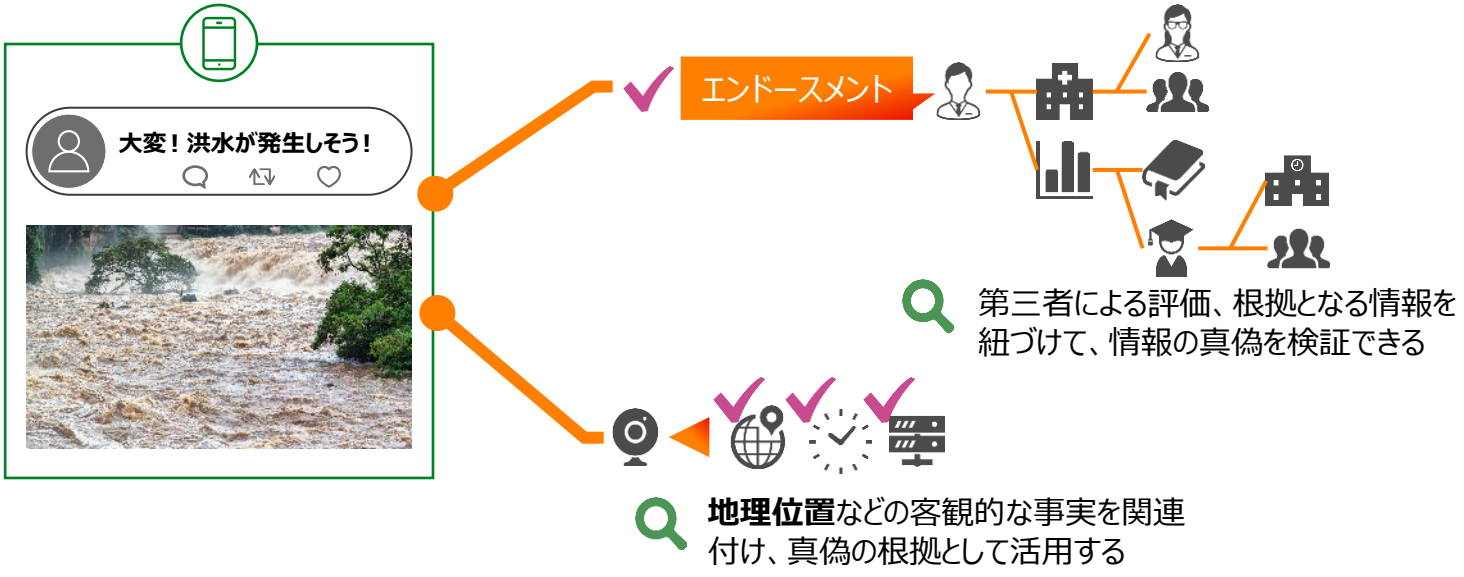
別カメラの画像

根拠となる複数の情報の整合性や矛盾から、情報の真偽を分析する

Trustable Internet

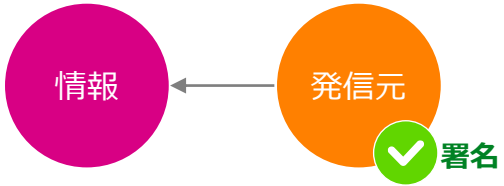
インターネット情報に関連する「第三者の根拠(エンドースメント)」をグラフ構造でメタデータ化。それらを分析することで情報の真偽を確認可能にする

- 情報の正確性を、専門性を持った第三者が内容を確認・判断することによる根拠
- ネットワークなどフィジカル空間から取得した**地理位置**などの情報から示される客観的根拠



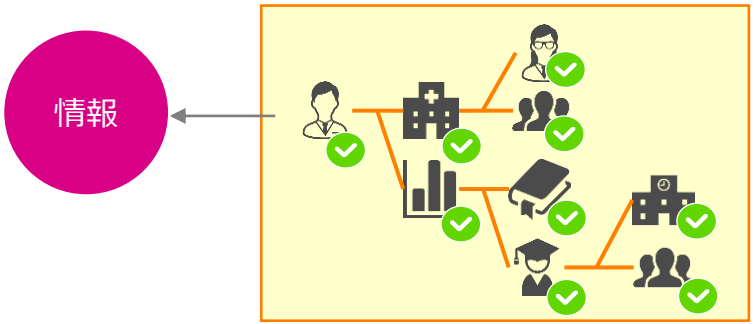
関連技術：Originator Profile

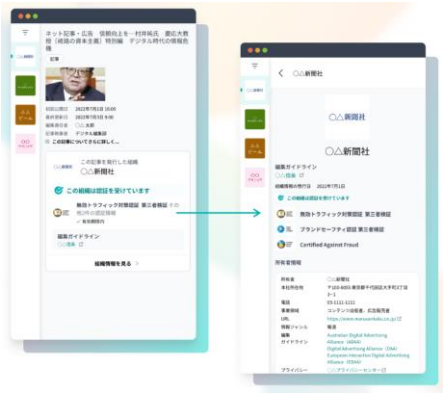
Originator Profile



拡張

Trustable Internet





- ✓ 情報の作成者や広告主といった、発信元を確認できる情報(署名)を付与する仕掛け
- ✓ 発行元を検証することにで情報の正しさを判断

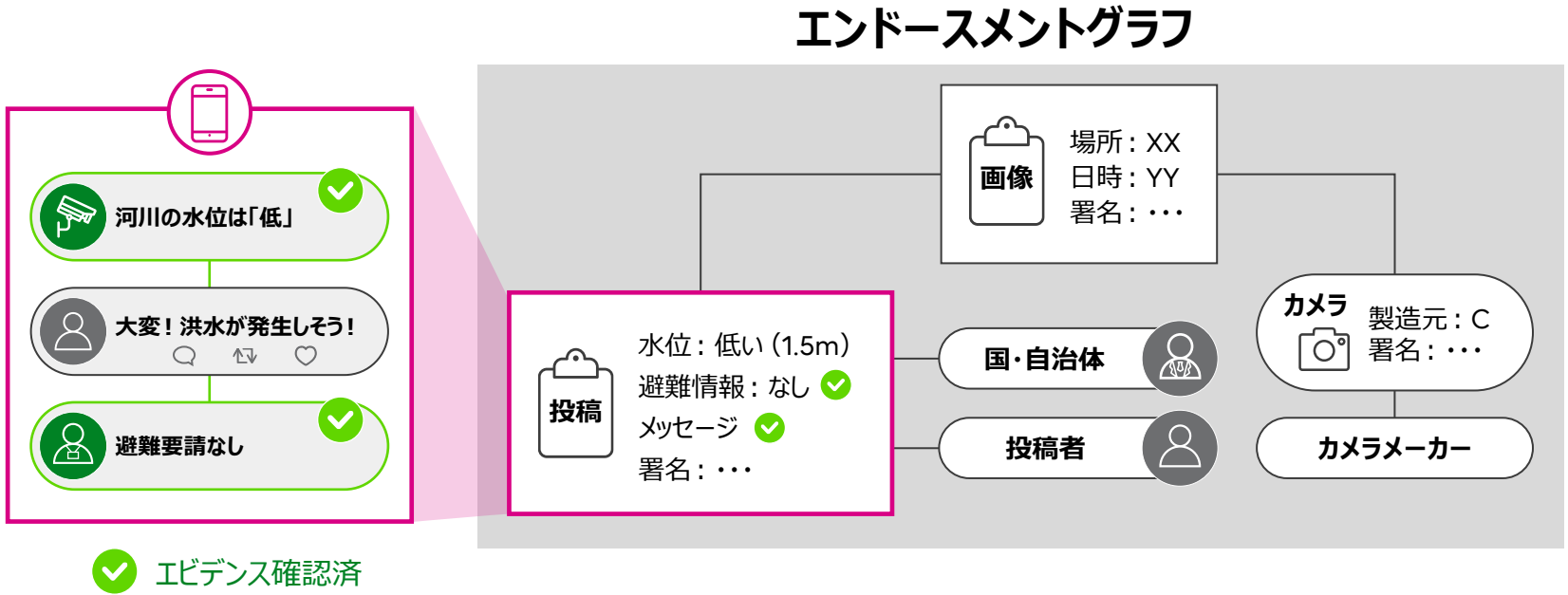
- ✓ 発行元だけでなく、客観的情報や評価まで含めて署名付きのエビデンスデータを構築し確認できる仕掛け
- ✓ 「検証可能な集合知」によって情報の正しさを判断

https://originator-profile.org/ja-JP/

Trustable Internet： エンドースメントグラフ

元の情報に対する様々な根拠情報を、自動でグラフ構造化し管理する

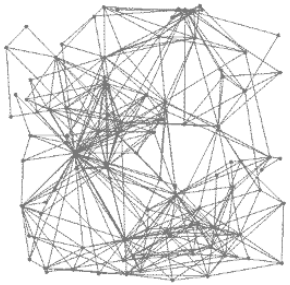
- グラフとして接続されたエンドースメントの間に矛盾がないことが、情報の正しさを証明になる
- エンドースメントの不足を検出し、自動で関連情報を探索



Trustable Internet： 真偽の根拠分析

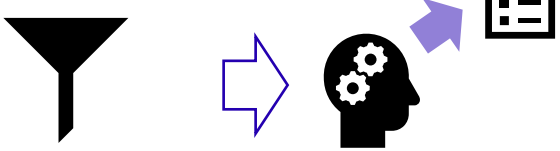
人間には難読な「エンドースメントグラフ」から、大規模言語モデル(LLM)によって、真偽の根拠を分析し、自然言語(文章)としてユーザに説明する

エンドースメントグラフ
（エビデンスの関係性）



入力

真偽の根拠分析技術



出力

誤り

真偽の根拠説明文

本記事内の画像については十分な証拠のもとで誤りだと指摘されています。判定結果の根拠として以下が挙げられます。

- 発信元アカウントであるBさんが自身で問題の画像がフェイクだと認めています。
- ファクトチェック機関であるBuzzFeed社が、静岡県の危機情報課を取材したうえで、本記事内の画像が虚偽であると判定しています。

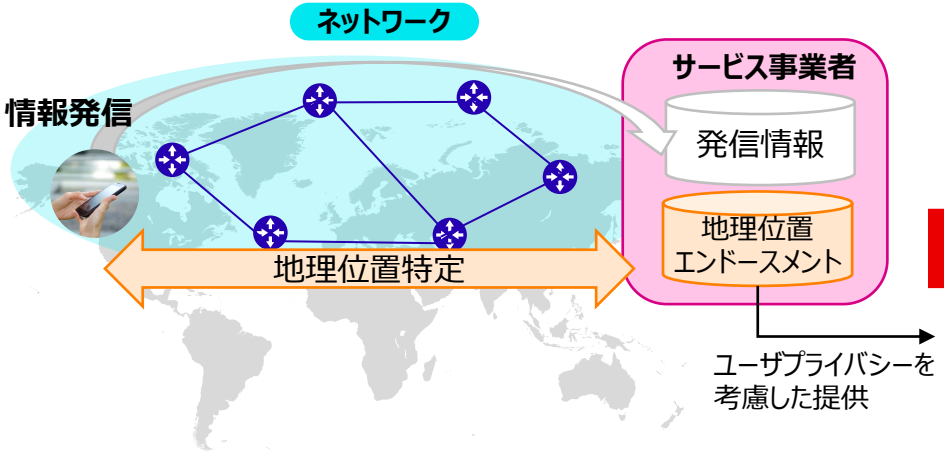
詳細情報を確認する。

紐づけられた根拠の矛盾を分析し、情報の真偽判断の根拠を説明

Trustable Internet： 発信元の地理位置分析

FUJITSU

- **様々な情報発信源の地理位置**を、ネットワークの情報や測定から高精度に特定し、**エンドースメントとして管理**
- 発信源の地理位置の関係矛盾を検証することで、情報そのものの真偽判定に活用



情報発信源の地理位置を特定

情報/エンドースメントMAP

情報発信源の地理位置の矛盾検証

動画デモ
（未公開）

まとめ

- インターネット上の情報の正しさを、第三者からのエビデンス情報から判断可能にする
- 小さな根拠の集まりが、情報の正しさ及び矛盾を示す「集合知によるファクトチェック」を実現

開示

FUJITSU