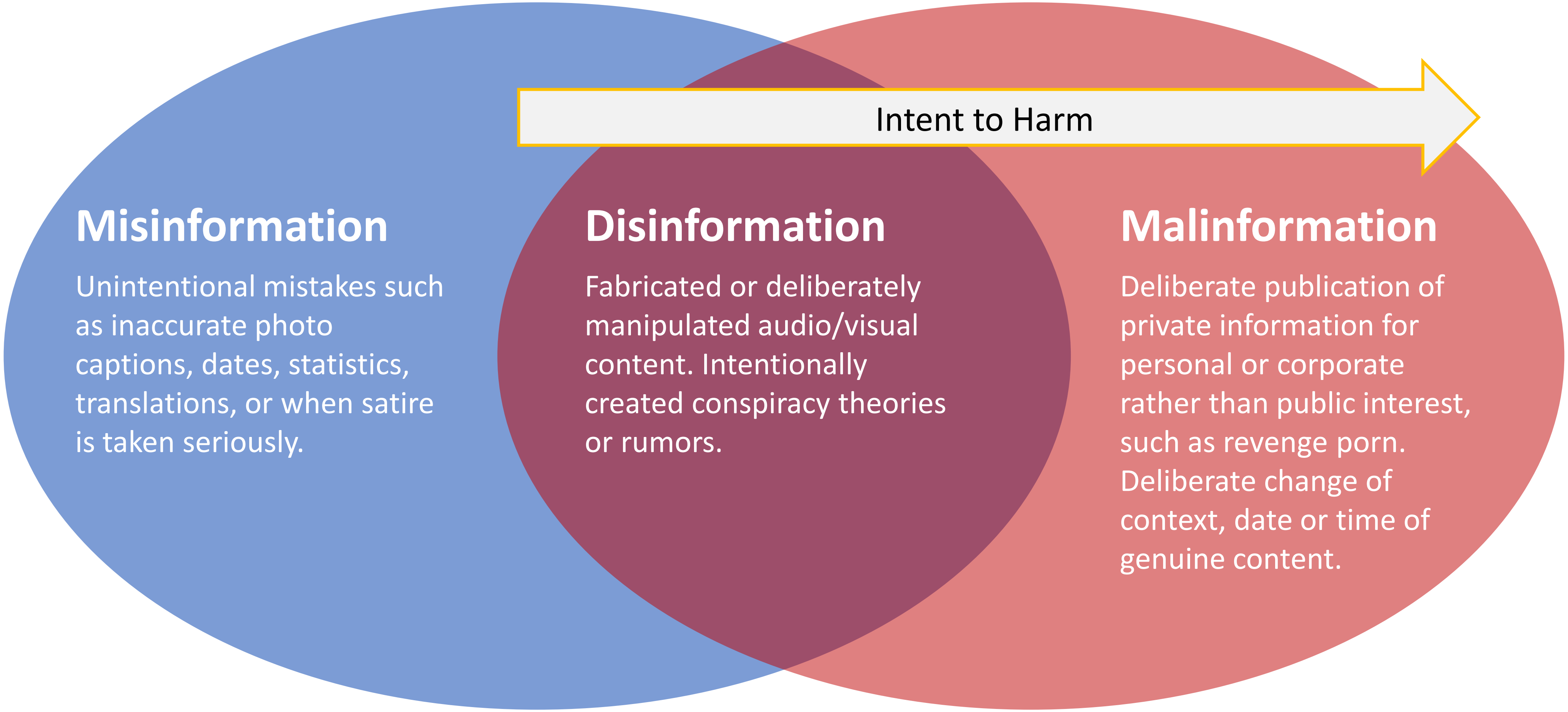
Disinformation Defense

Provenance

Bing Multimedia
Microsoft

Ashish Jaiman
@ashishjaiman
medium/@ashishjaiman
linkedIn/in/ashishjaiman

Disinformation: Mapping the Problem



Intent to Harm

Misinformation

Unintentional mistakes such as inaccurate photo captions, dates, statistics, translations, or when satire is taken seriously.

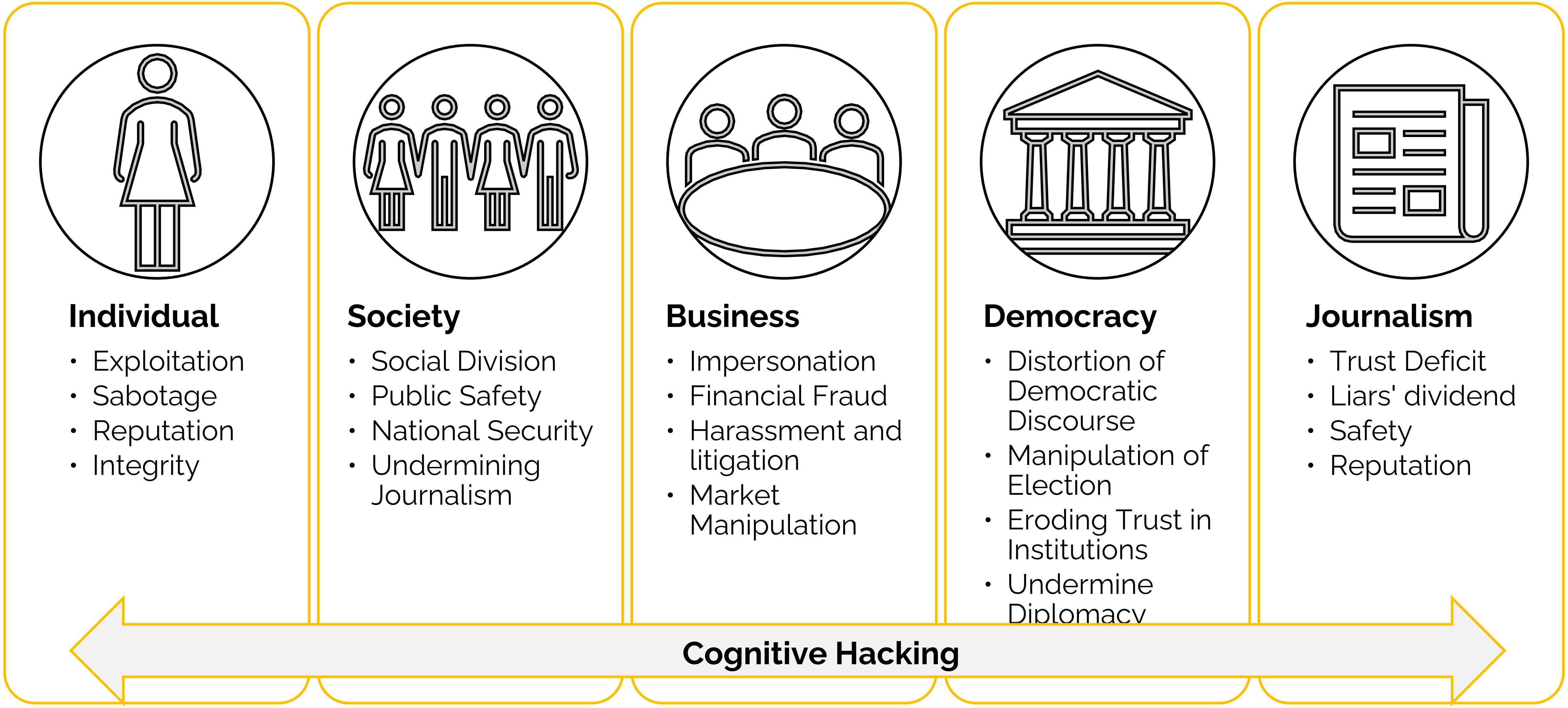
Disinformation

Fabricated or deliberately manipulated audio/visual content. Intentionally created conspiracy theories or rumors.

Malinformation

Deliberate publication of private information for personal or corporate rather than public interest, such as revenge porn. Deliberate change of context, date or time of genuine content.

Disinformation Harms & Threat Modeling



Individual

- Exploitation
- Sabotage
- Reputation
- Integrity

Society

- Social Division
- Public Safety
- National Security
- Undermining Journalism

Business

- Impersonation
- Financial Fraud
- Harassment and litigation
- Market Manipulation

Democracy

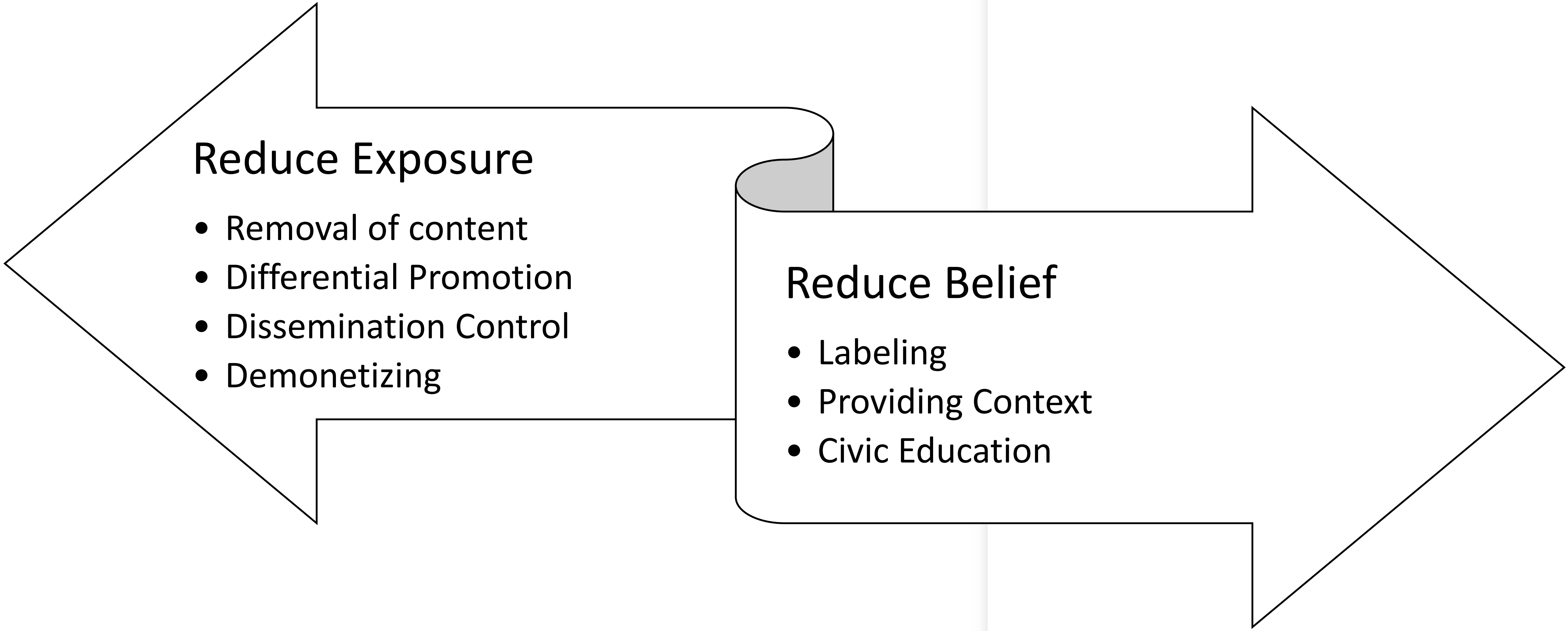
- Distortion of Democratic Discourse
- Manipulation of Election
- Eroding Trust in Institutions
- Undermine Diplomacy

Journalism

- Trust Deficit
- Liars' dividend
- Safety
- Reputation

Cognitive Hacking

Effective Responses



Reduce Exposure

- Removal of content
- Differential Promotion
- Dissemination Control
- Demonetizing

Reduce Belief

- Labeling
- Providing Context
- Civic Education

Countermeasures

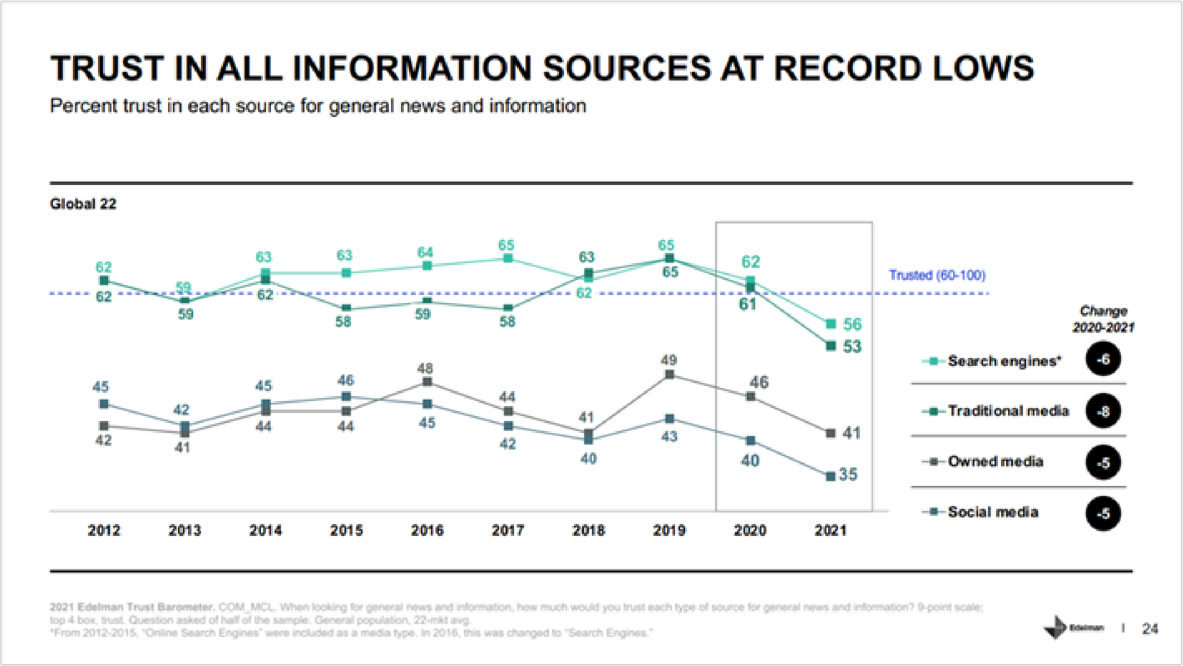
- Media Literacy
  - Consumers
  - Journalists
  - Voters
- Platform Policies
  - Terms of Use
  - Code of Conduct
  - Norms
- Regulations
  - Individual
  - Election
  - Business
- Technology
  - Detection
  - Authentication
  - Provenance

Technical Solutions

- Detection
- Authentication
- Provenance

Problem statement

- Publishers and consumers have low trust in digital content
- Ecosystem-wide impact, multiple symptoms
  - Publishers looking for ways to increase trust, maintain value
  - Consumers skeptical and unsure who to trust
  - Lack of trust in content shared on social media
  - Everyone concerned about “deepfakes”



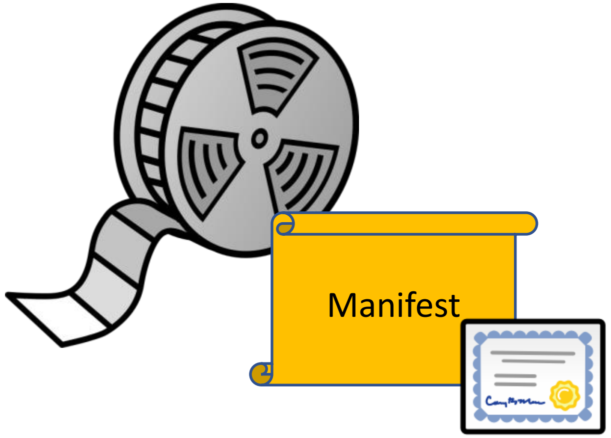
TRUST IN ALL INFORMATION SOURCES AT RECORD LOWS

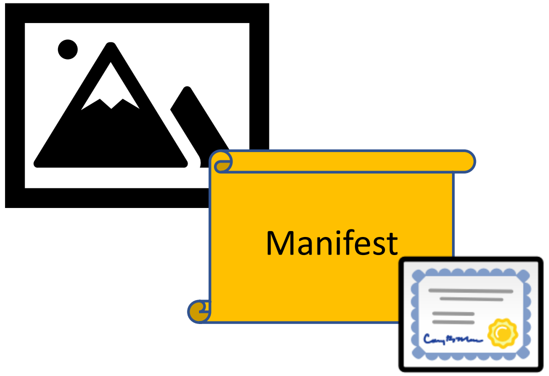
Percent trust in each source for general news and information

2021 Edelman Trust Barometer. COM_MCL. When looking for general news and information, how much would you trust each type of source for general news and information? 9-point scale; top 4 box, trust. Question asked of half of the sample. General population, 22-mkt avg.

*From 2012-2015, "Online Search Engines" were included as a media type. In 2016, this was changed to "Search Engines."

Provenance





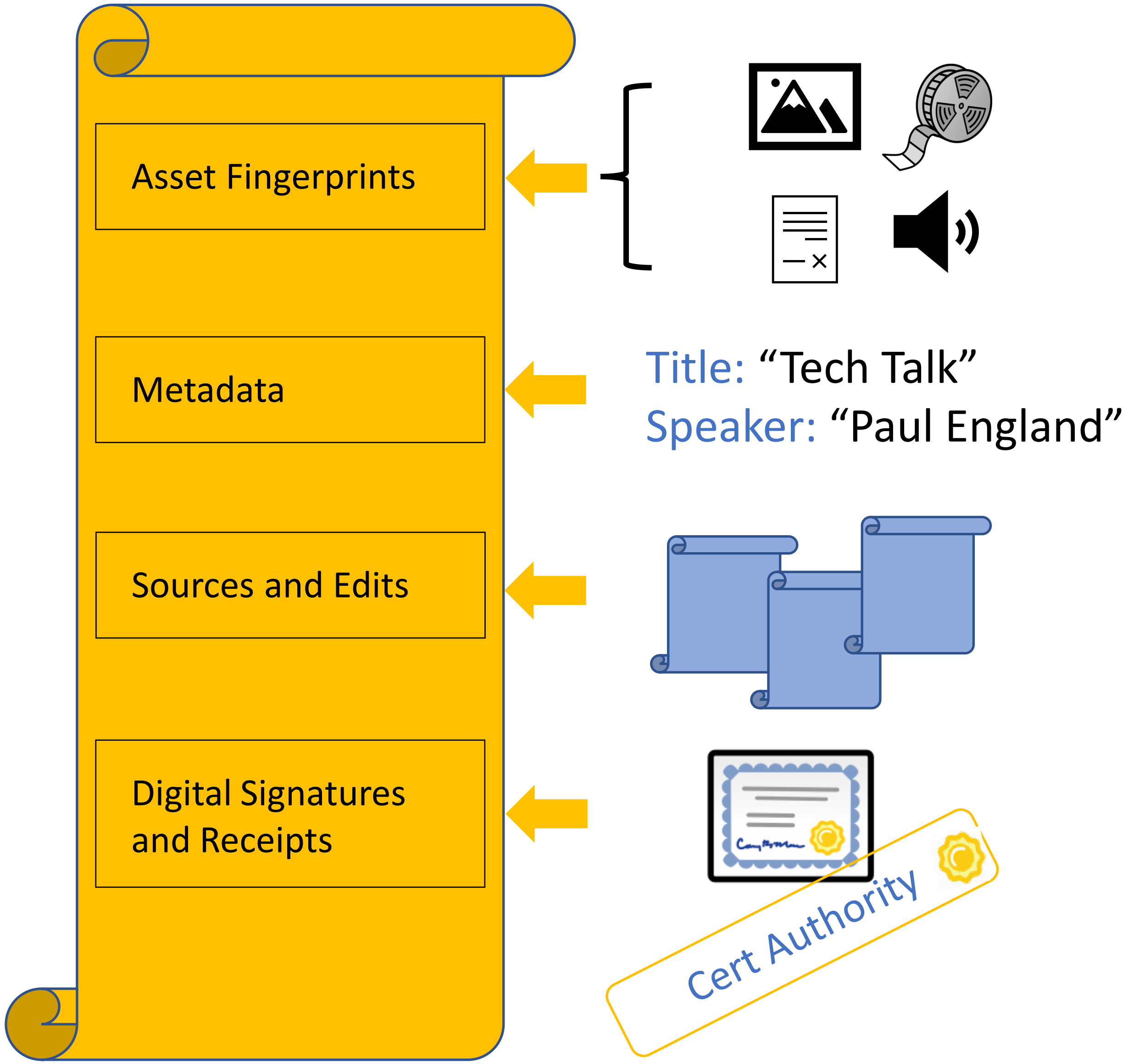
Enhanced media come with **Manifest**

- Contains creator-chosen metadata tags
- Is “bound” to the associated asset
- Is signed by the creator

If you get media with an Manifest, you can be sure that:

- It came from the person/publisher that signed it
- It has not been altered

Anatomy of an Manifest



Asset Fingerprints

Metadata

Sources and Edits

Digital Signatures and Receipts

Title: “Tech Talk”
Speaker: “Paul England”

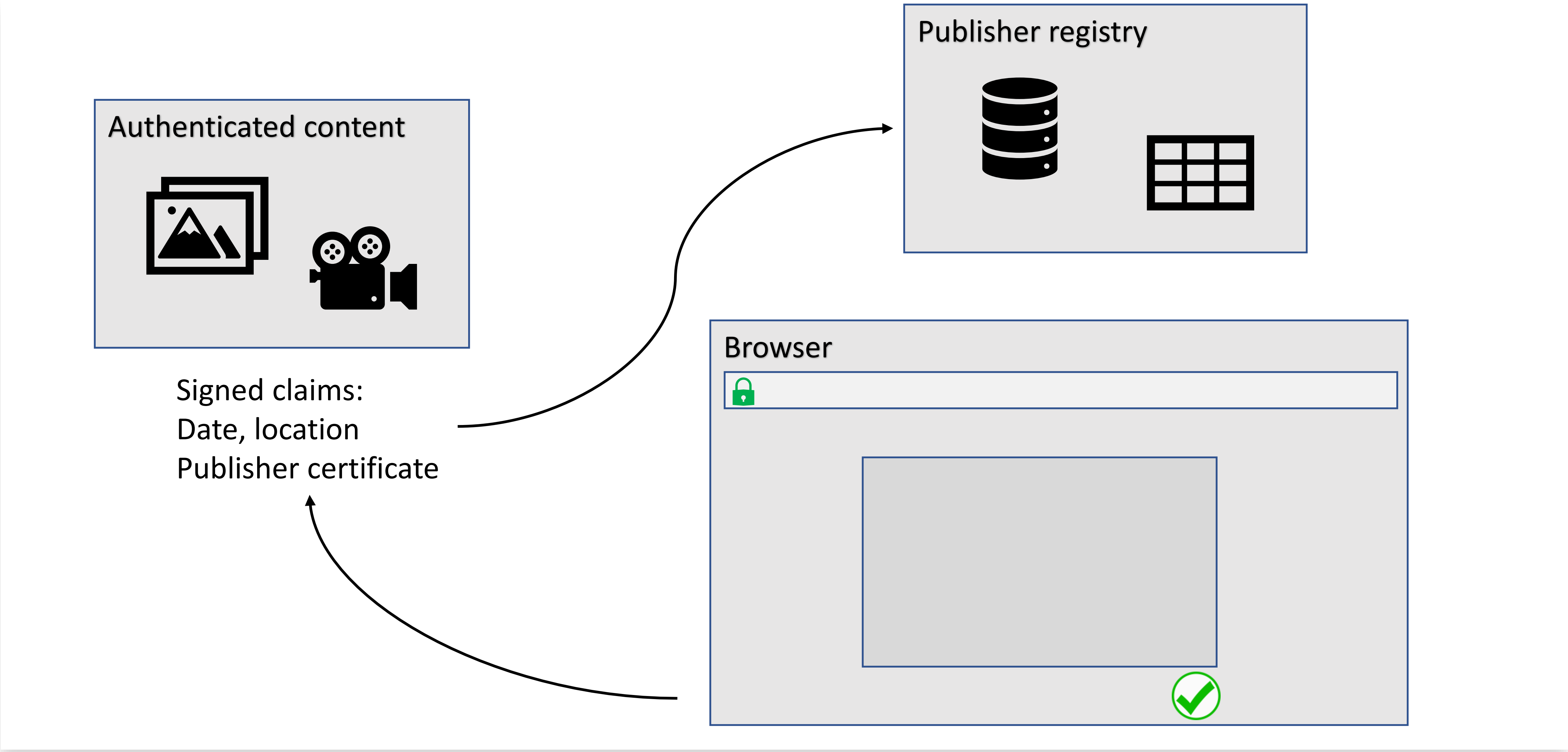
Cert Authority

Assets: Any digital object. document, image, video, audio…
Fingerprint: A crypto or “soft” hash

Metadata: Text or other tags chosen by producer (publisher, camera, etc.)

Antecedents: Source assets and edits performed (clipping, overlays…)

Digital Signature: Source/publisher
Receipts: Proof of registration, etc.



Authenticated content

Publisher registry

Signed claims:
Date, location
Publisher certificate

Browser

deceptively simple

Coalition for Content Provenance and Authenticity (C2PA)

An open technical standard providing publishers, creators, and consumers the ability to trace the origin of different types of media.

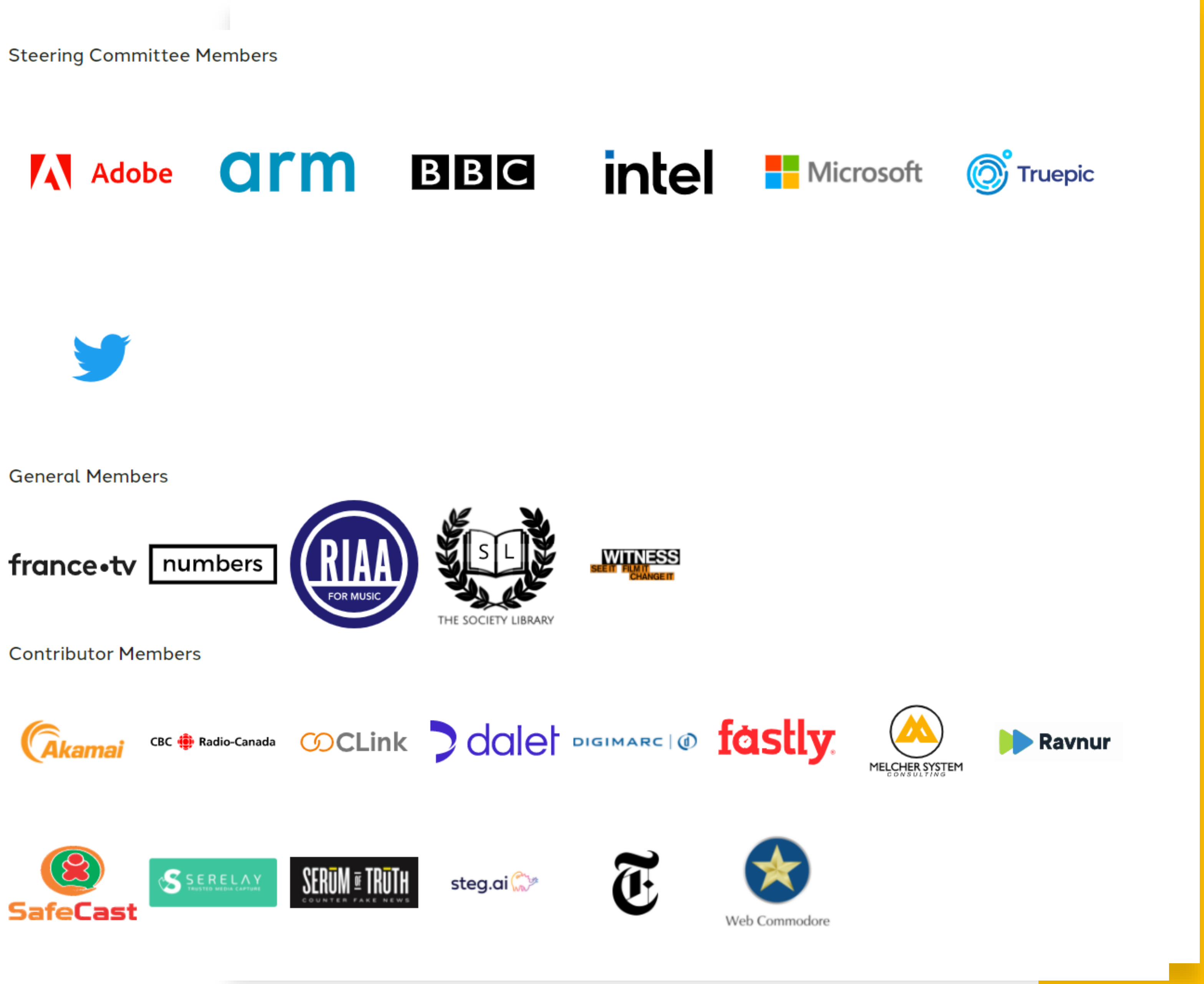
C2PA Charter

Develop technical specifications that can establish content provenance and authenticity at scale to give publishers, creators, and consumers the ability to trace the origin of media.

Joint Development Foundation Project

https://www.jointdevelopment.org/

C2PA Members



Steering Committee Members

General Members

Contributor Members

C2PA Key Activities

- Applying requirements from industry to the development of content provenance specifications
- Ensuring that the specifications can be used in ways that respect privacy and personal control of data, and promote tool availability for a wide range of organizations
- Ensuring that specifications meet appropriate security requirements
- Promoting selected specifications to become global standards
- The global adoption of digital provenance techniques by target industry devices, systems, and services, including social media and messaging platforms
- Ensuring that content accessibility is not negatively impacted by digital provenance techniques

thank you

ashish.jaiman@microsoft.com