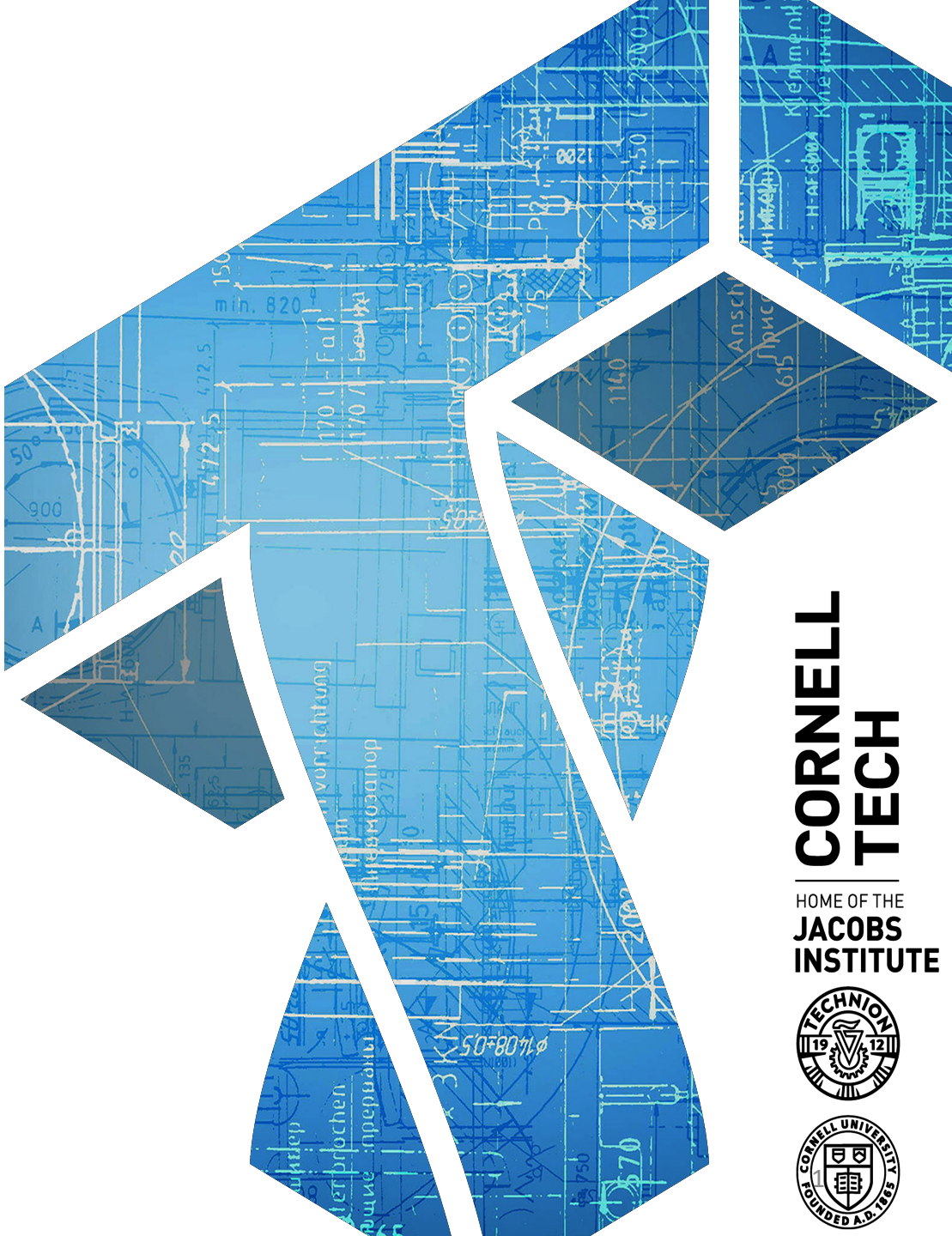
CS 6431: Hate and harassment

Instructor: Tom Ristenpart

https://github.com/tomrist/cs6431-fall2021



CORNELL TECH

HOME OF THE JACOBS INSTITUTE

Project description posted to github

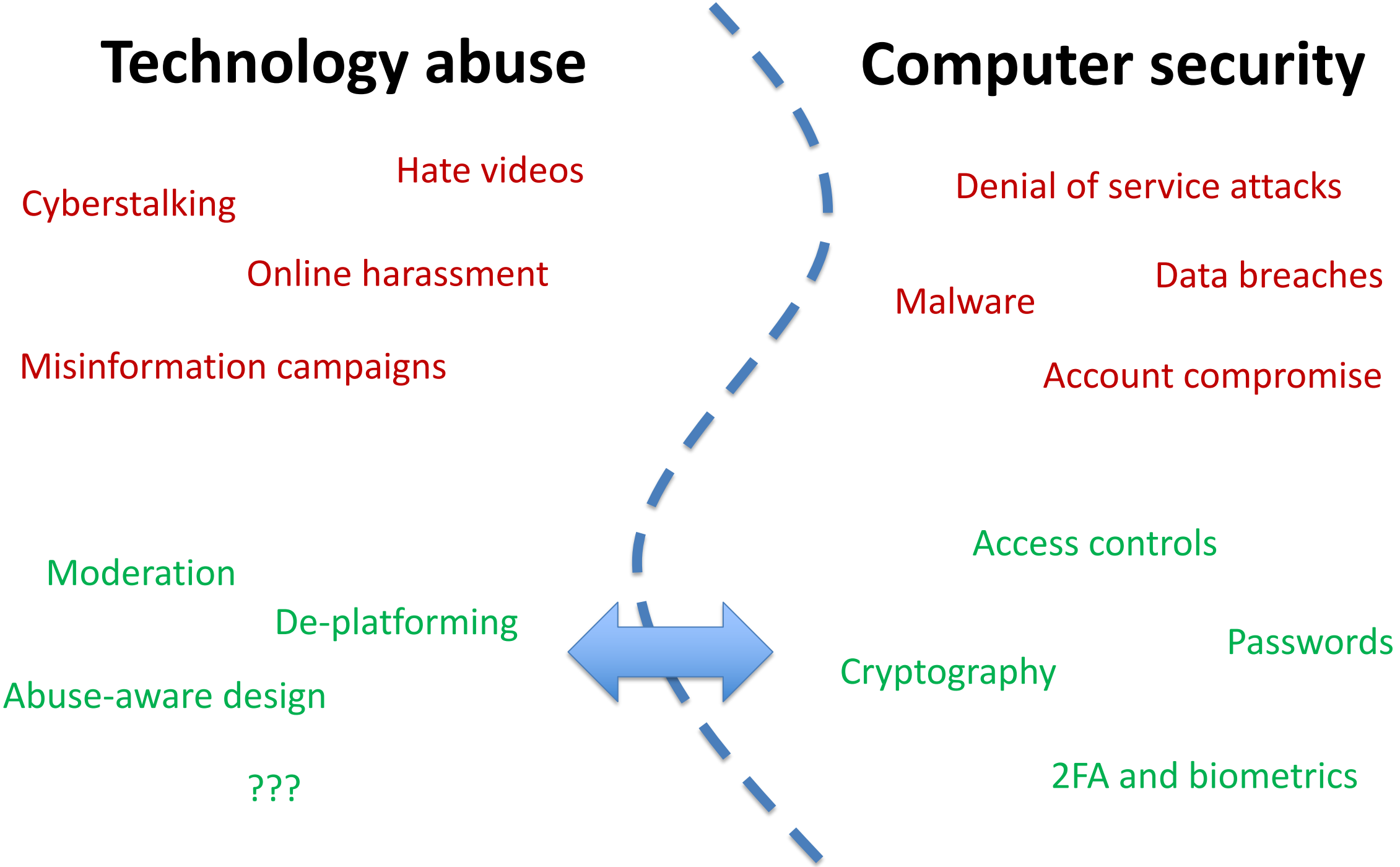
- Provide a peer review process:
  - Proposal due Sep 28
  - Proposal reviews due Oct 8
  - Paper due Dec 3
  - Paper reviews due Dec 10
  - Revised paper due Dec 15
- Still figuring out what I'll do with problem sets, stay tuned

Many flavors of abuse categories

- Financially motivated abuse:
  - Spam, scams, sextortion, …
- Online harassment and bullying
  - Coordinated campaigns
  - Doxxing, raiding (other websites), …
- Misinformation campaigns
- Targeted attacks and RATs (remote access trojans)
- Tech abuse in intimate partner violence

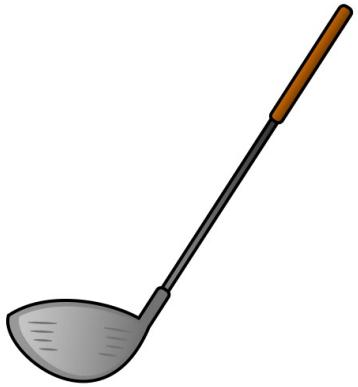
Why an SoK paper, in 2021?

- SoK = systemization of knowledge
  - Type of paper category introduced at Oakland in 2010
  - Review article that distills some area of work. Not just survey
- A key motivating question:
  - What research community should address tech abuse?


Technology abuse
Computer security
Hate videos
Cyberstalking
Denial of service attacks
Online harassment
Data breaches
Malware
Misinformation campaigns
Account compromise
Moderation
Access controls
De-platforming
Passwords
Cryptography
Abuse-aware design
2FA and biometrics
???

Potential dichotomy?

- Security:
  - Causing information technologies to operate in ways unintended by their designers
- Abuse:
  - Use of a system in ways for which it was designed, but to cause harm anyway



Potential dichotomy?

- Abuse involves understanding & mitigating tech-based harms to people

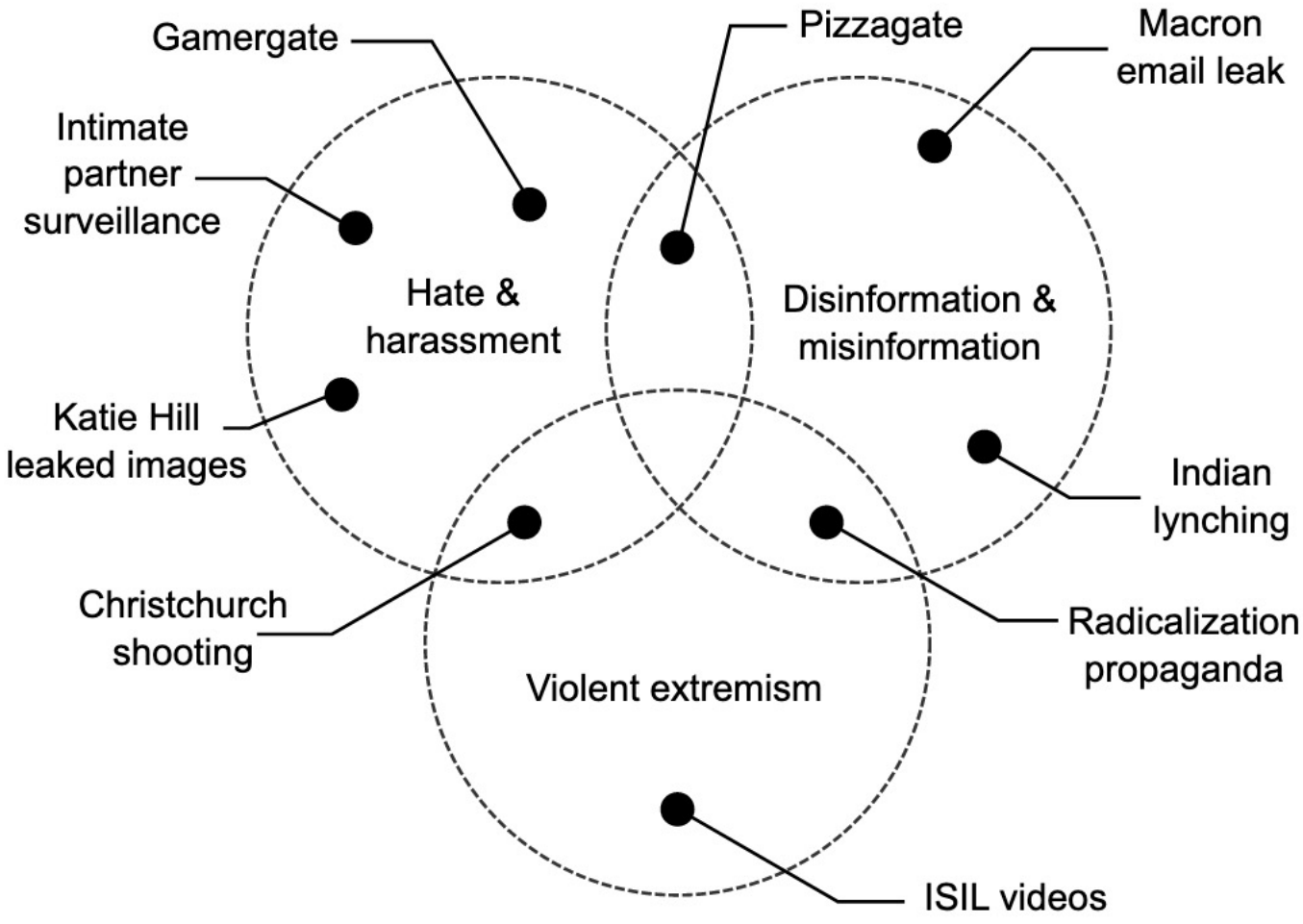
- Security involves understanding and protecting against harms to technology

Hate and harassment

- “Hate and harassment occurs when an aggressor (either an individual or group) specifically ***targets another person or group*** to ***inflict emotional harm***, including coercive control or instilling a fear of sexual or physical violence [36].”

Hate & harassment vs …

- Violent extremism
  - Doesn't target particular individual/group (?)
- Misinformation
  - Not targeting emotional harm
- Financially-motivated abuse
  - Not targeting emotional harm
  - Can't disrupt financial infrastructure (e.g., bank takedowns)
- Of course can overlap

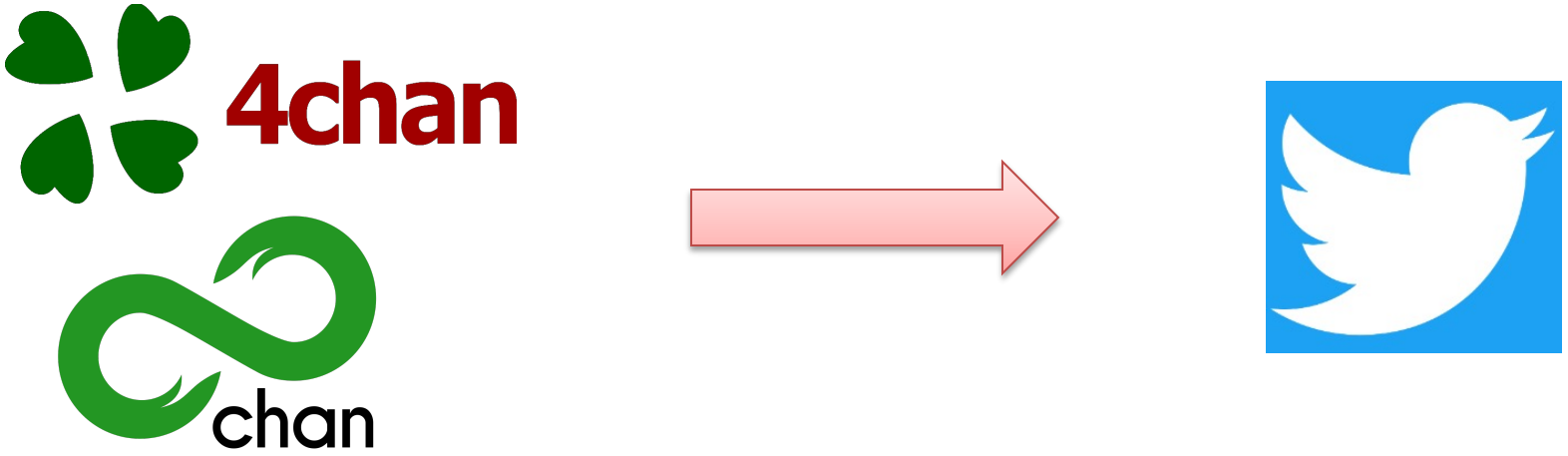


Coordinated harassment campaigns

- Anonymous bulletin board
- Generated lots of memes, known to host some child sexual abuse media (technically against site policy)
- Associated with Anonymous hacker group
- Used to coordinate harassment, bullying, hacking
- Associated with alt-right groups
- Involved in Gamergate, banned Gamergate discussion

- Similar to 4chan (different site operator)
- No longer available due to CloudFlare and other providers dropping it

Coordinated harassment campaigns: Gamergate



- Sustained harassment campaign against female gaming developers and others
  - Rape threats, murder threats, doxing (home address, other personal information disclosed)
  - Astroturfing, sock-puppet campaigns on Twitter
- Organized on 4chan & (after banning on 4chan) 8chan, most harassment played out on twitter (#gamergate)
- Why? Ex-boyfriend of initial victim posted false claims about victim
  - "Right-wing backlash against progressivism" (in gaming)

Mininformation campaigns

- Pizzagate conspiracy
- QAnon conspiracy
- 2016 election interference
  - Cambridge Analytica
  - Russian influence operations
- Research studies indicate falsehoods spreads faster than truth on social media

> "there is a worldwide cabal of Satan-worshiping pedophiles who rule the world, essentially, and they control everything. They control politicians, and they control the media. They control Hollywood, and they cover up their existence, essentially. And they would have continued ruling the world, were it not for the election of President Donald Trump,"
>
> -Travis View

Other examples that stood out in paper?

- Toxic content
- Content leakage
- Overloading
- False reporting
- Impersonation
- Surveillance
- Lockout and control

Global survey of hate & harassment

- Google led this portion of paper
- Market research firm, 60 question survey, small subset relevant to this paper
- "*Have you ever personally experienced any of the following online?*"
  - List of abuse categories
- Survey design and analysis very nuanced
  - Examples of potential pitfalls?

Global survey of hate & harassment

- High-level takeaways from the survey:
  - Overall increase in hate & harassment experienced online
  - Some groups disproportionately experience hate & harassment (and some stereotypes not supported by data)

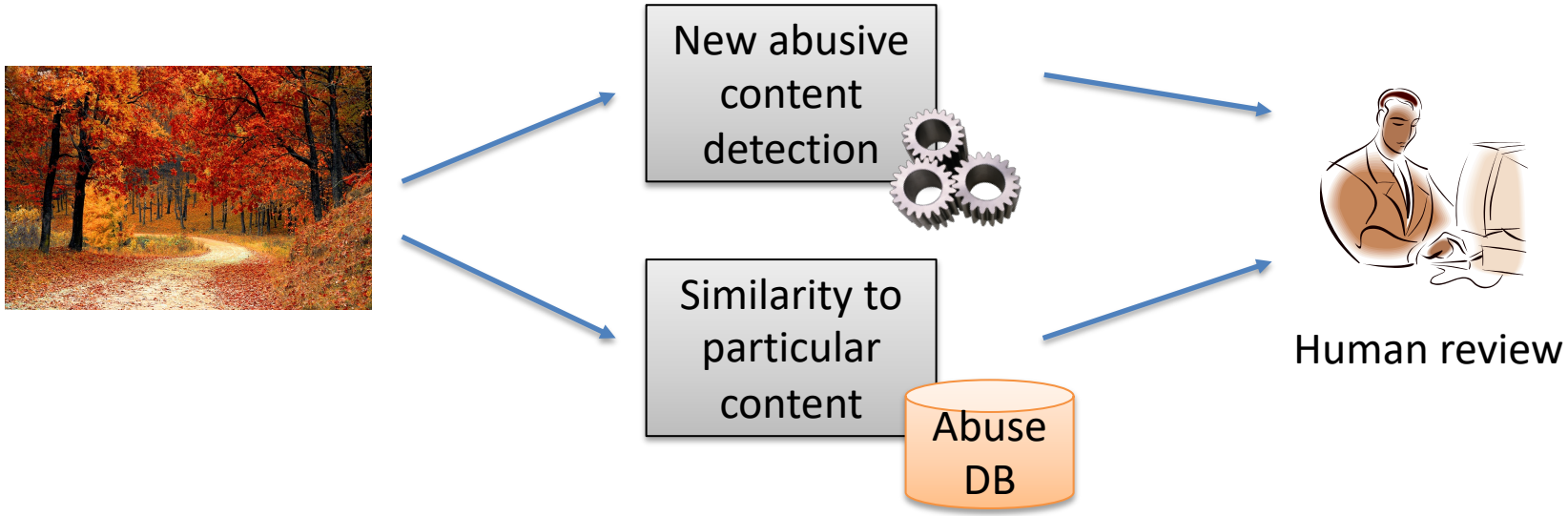
How can we make progress?

- The case for engaging computer security community
- Five areas:
  - Nudges, indicators, warnings
  - Human moderation tools
  - Automated detection
  - “Conscious design”
  - Policy
- Challenges facing progress

How to prevent fraudulent actions?

- **Content analysis**
  - Spam filters
- **Identify bad accounts**
  - Correlate bad actions with accounts, try to detect bots
  - Sock-puppet accounts (many controlled by one person)
- **Identify bad devices**
  - Device cookies to correlate different account accesses
  - IP blacklists, ISP blacklists

Identifying abuse content



- Use locality-sensitive hashing to check if similar to known bad content
  - H such that |H(image) – H(image')| < threshold if image, image' very similar
- Use machine learning to try to identify patterns indicative of abusive content
  - Is image a picture of a naked child?
- Refer out to human for review (Facebook has 10,000s of moderators)
- Content flagged by users also sent through reviewing pipelines
  - Flagging mechanisms also subject to abuse (illicit takedowns)

Apple CSAM detection proposal

- Use locality-sensitive hashing combined with threshold private-set intersection (PSI) protocol
- Encrypted database of known CSAM hashes on devices
- Client-side computation of hashes of user images
- Protocol to notify Apple when sufficiently many images match

Working with at-risk groups

- Many at-risk groups, growing research in studying specific groups' tech abuse and designing mitigations
- Combine domain expertise for that population with technical understanding of broader issues
  - Allows more nuanced understanding of abuse, unique threats faced