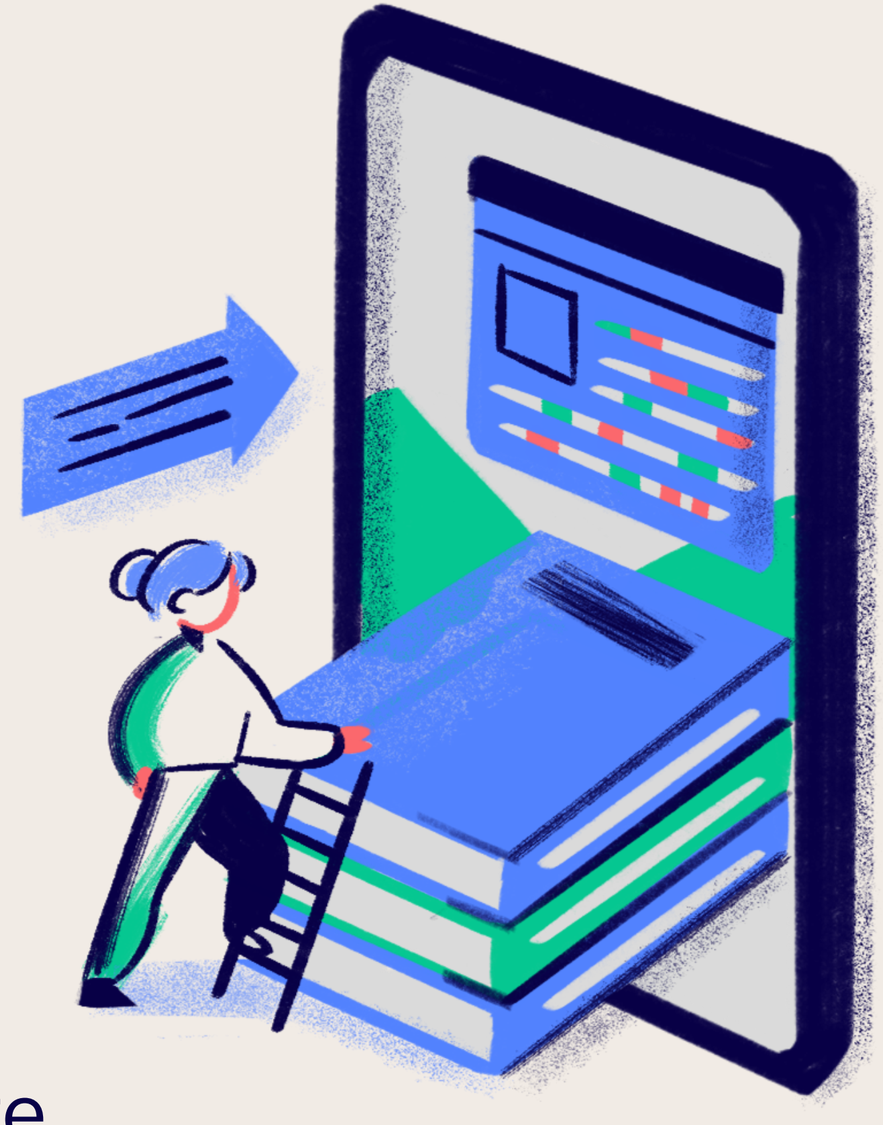
Presented by Mary Nerayo

Data Wars

Using and Abusing Your Personal Information



VIRGINIA TECH

Virginia Tech National Security Institute

Agenda

OSINT

Intelligence Collection
Definitions
OSINT Utility
Ethical Considerations in OSINT
Understanding PII
Data Breaches
Cases

Use and Abuse Project

Objective
Research Approach
Fake IDs
Initial Findings
Case – Politics
Methodology
Findings
Interaction Engine

VT National Security Institute

Vision: We meet the pressing needs of the defense and intelligence communities by developing their future workforce and advancing interdisciplinary research, technology, and policy.

Technical Divisions

Spectrum Dominance

- Assured and secure communications
- Advanced C4ISR and counter-C4ISR
- Quantum and heterogeneous computing
- RF machine learning
- Open Gen wireless innovation

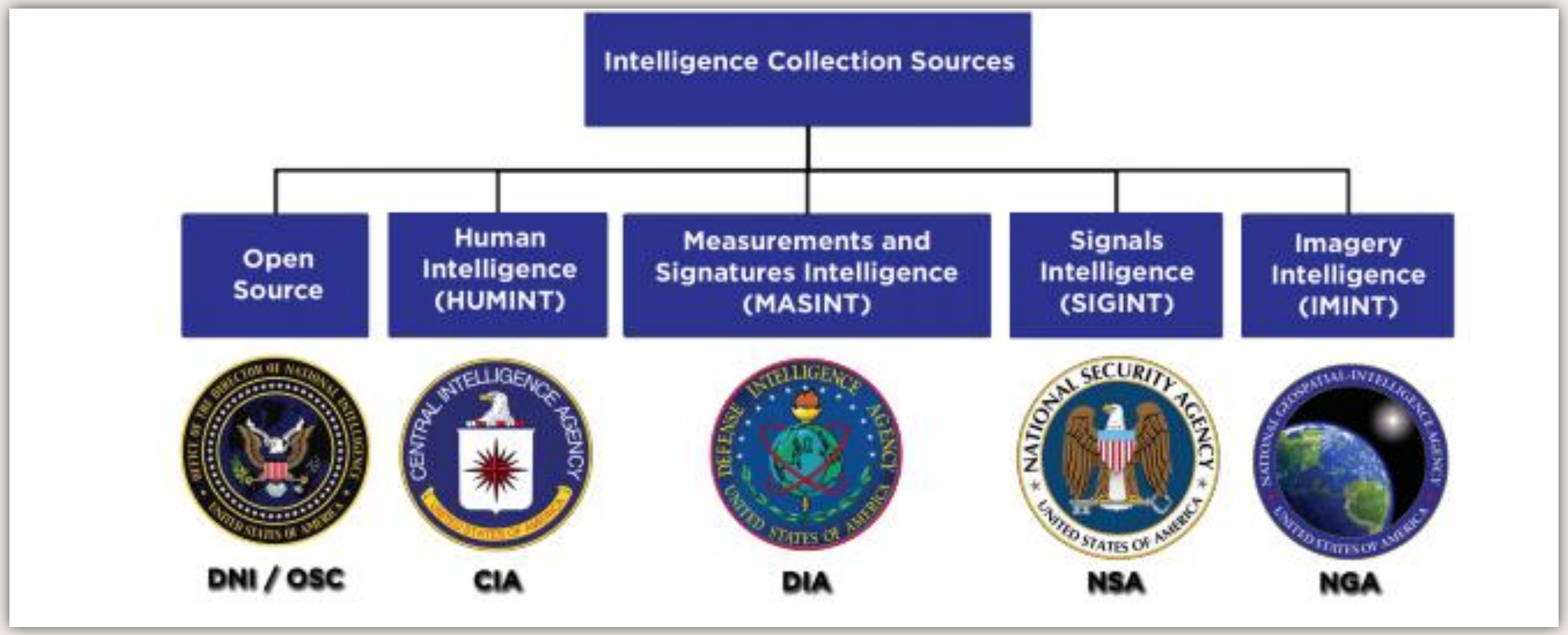
Mission Systems

- Resilient, autonomous missions
- Remote & in-situ sensing
- Space situational awareness
- Marine autonomy and robotics
- Energetic materials

Intelligent Systems

- Data science, ML, AI
- Cyber security & complex systems
- Validation and test & evaluation
- Deep learning for sensor processing
- Data fusion and sensemaking

Intelligence Collection



Definitions

Open-source Intelligence (OSINT)

The process of gathering and analyzing publicly available information to assess threats, make decisions or answer specific questions. **– IBM**

Examples:
Social Media
Blogs
News
The Dark Web

Personal Identifiable Information (PII)

Any personal data connected to a specific individual that can be used to uncover their identity. - IBM

Examples:
Social Security number
Phone Number
Address
Patient ID
Tax Payer ID

OSINT Utility

- **Social Media Investigation**: track criminal activity, verify alibis, or investigate fraud.
- **Humanitarian Aid - Tracking COVID-19 Spread**: analyze publicly available flight data, social media updates, and health reports to identify hotspots.
- **Detecting Fake News**: help fact-checkers and researchers detect misinformation by cross-referencing publicly available information.
- **Academic Research**: widely used by researchers to collect data for studies in fields like political science, sociology, economics, and international relations.
- **National Security**: emerging threat assessment, military intelligence.

Ethical Considerations of OSINT

- **Informed Consent -** Unlike traditional intelligence-gathering, OSINT doesn't involve notifying the target, which raises ethical concerns regarding consent.
- **Data Sensitivity and Sharing -** Sensitive data should be handled carefully, anonymized when possible, and shared only with authorized parties.
- **Attribution of Sources -** Some OSINT practices involve automated tools or AI to gather data, which can lead to misinterpretation or false attribution if the information is incomplete, incorrect, or taken out of context.

Understanding PII

Why so valuable?

- **Identity Theft**: Hackers use PII to impersonate individuals, allowing them to gain access to personal accounts, steal identities, and commit fraud. Identity theft can result in damaged credit scores and legal complications for victims.

- **Financial Gain**: Sold to fraudsters, who use it to commit financial crimes such as opening bank accounts, applying for loans, or making unauthorized purchases.

- **Corporate Espionage**: Sensitive business information tied to PII, such as client lists or proprietary employee data, can be exploited by competitors to gain an advantage in the market or undermine the integrity of a company.

Understanding PII

How can it be exploited?

- **Social Engineering**: attackers manipulate victims into revealing confidential information by posing as legitimate authorities or trusted contacts.

- **Data Brokers and Dark Web Markets**: can be used to create fake identities, engage in financial fraud, or carry out illegal transactions.

- **Cyber Warfare**: can be used by hostile nations to infiltrate the networks of critical infrastructure, such as energy grids, water supplies, or transportation systems.

- **Critical Infrastructure and Public Safety**: attacks on national healthcare infrastructure, expose undercover officers working in sensitive national security roles.

The Era of Data Breaches

Data Breaches: Large-scale data breaches expose PII from millions of users at once. Once leaked, this information is vulnerable to exploitation by multiple bad actors, leading to a cascade of security and privacy violations.

Facts:

- The average total cost of a data breach is $4.88 million.
- 83% of data breaches in 2022 involved internal actors.
- It took an average of 194 days to identify a data breach globally in 2024, a slight decrease from 2023.

“Worst Breach in History”

–

Yahoo! 2013

1. Yahoo!

Date: 2013-2016

Impact: Over 3 billion user accounts exposed

The data breach of Yahoo is one of the worst and most infamous cases of a known cyberattack and currently holds the record for the most people affected. The first attack occurred in 2013, and many more would continue over the next three years.

A team of Russian hackers targeted Yahoo’s database using backdoors, stolen backups, and access cookies to steal records from all user accounts, which included personally identifiable information (PII) like:

- Names
- Email addresses
- Phone numbers
- Birth dates
- Passwords
- Calendars
- Security questions

Initially, Yahoo reported stolen data from about 1 billion accounts. However, after Verizon bought out Yahoo in 2017, they reported that the final number of records totaled about 3 billion accounts affected. Not only was Yahoo slow to react, but the company also failed to disclose a 2014 incident to users, which resulted in a $35 million fine and, in total, 41 class-action lawsuits.

Case: Facebook (Cambridge Analytica Scandal, 2018)

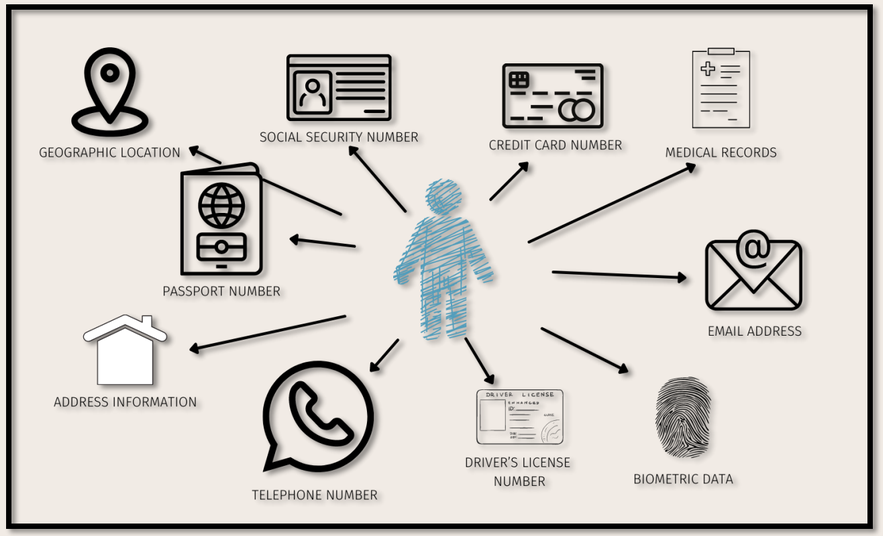
https://youtu.be/vyRsarL-dqk

Use and Abuse of

Personal Identifiable Information

PI: Dr Alan Michaels, Director of Spectrum
Dominance Division, Virginia Tech NSI

ID: Diana Prince, no digital fingerprint!



Objective Statement

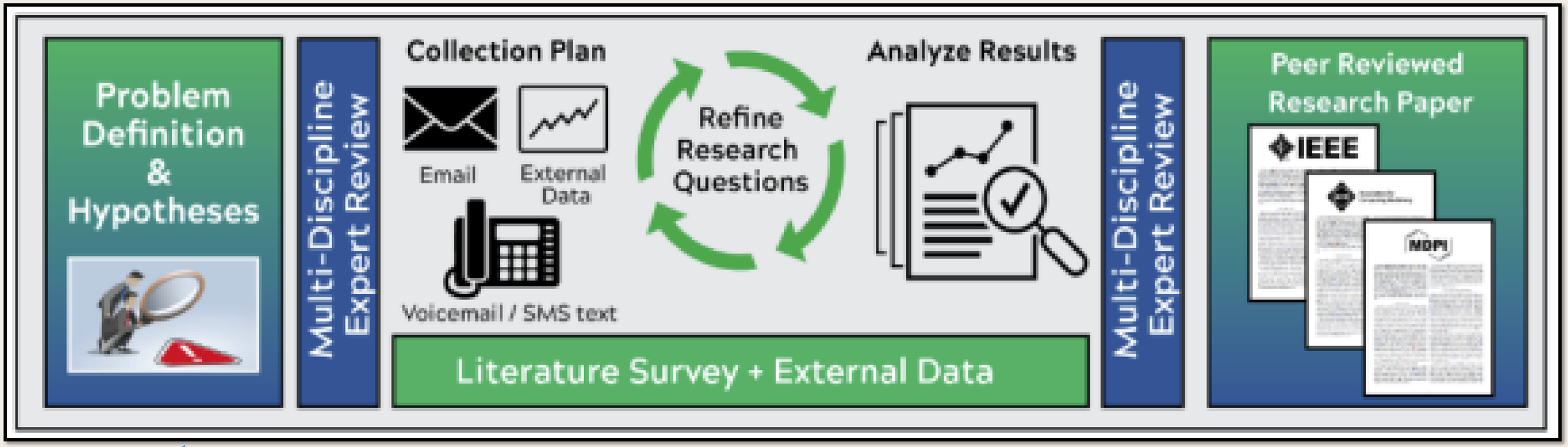
To investigate the **flow and movement** of personal information across the internet by tracking the dissemination of **falsified identities through one-time interactions** with various organizations.

Research Approach

Define Research Questions

Pick subset of fake IDs

3 phone servers and an email server owned by NSI



Collect list of organizations

Sign-up Engine

100,000 Fake Identities?!

- A pseudorandom number generator (**PRNG**) is used to randomize key ID fields, such as name, race/ethnicity, and height/weight, based on **Census, CDC, and third-party data** to ensure the IDs reflect the **US population**.
- **Non-existent but realistic addresses** are generated by modifying real addresses through PRNG and verifying against the **USPS** database to ensure uniqueness.
- Political affiliation is assigned proportionally based on **2020 election data** to create a representative set of identities.

Secret Word: OSINT

First Iteration – Interesting Finds

Summer 2020

01. 97% of identities (290/300) showed no evidence of email sharing, but phone data was shared more frequently, despite offering fewer traceable records.

02. No significant differences were observed between foreign and domestic companies in terms of email frequency, interest in election outcomes, or privacy policies, although the sample size was limited.

03. A single interaction generated an overwhelming volume of traffic, with one news agency sending 2,436 emails in nine months, including 44 emails in a single day before the election. Republican political content was nearly double that of Democrats.

Case – 2024 Political Elections

Data Handling?

How are politicians handling the data of their supporters?

Foreign Influence?

What is the role of foreign entities in the trading of political data?

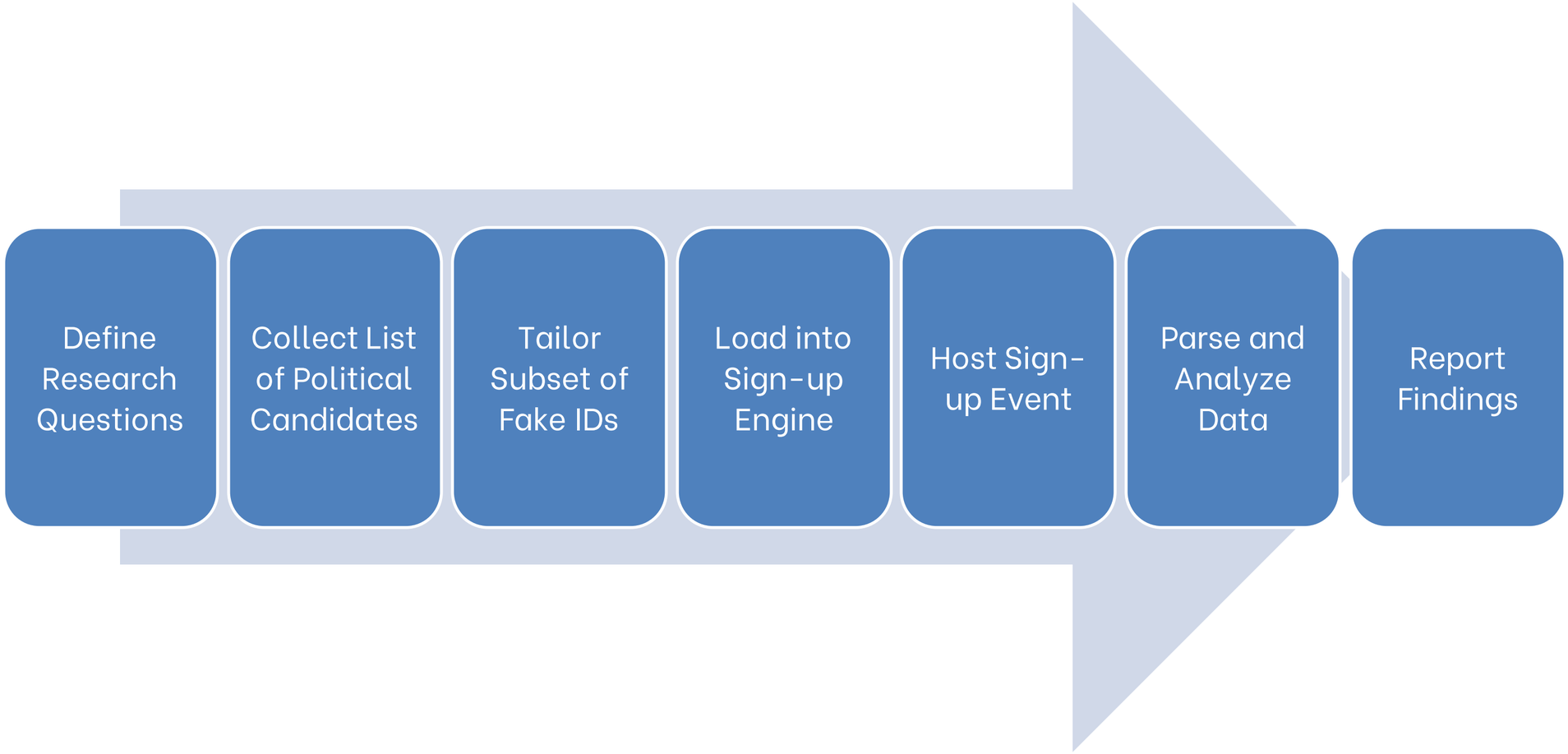
Trends?

Can we identify trends in which politicians or political parties we can trust?

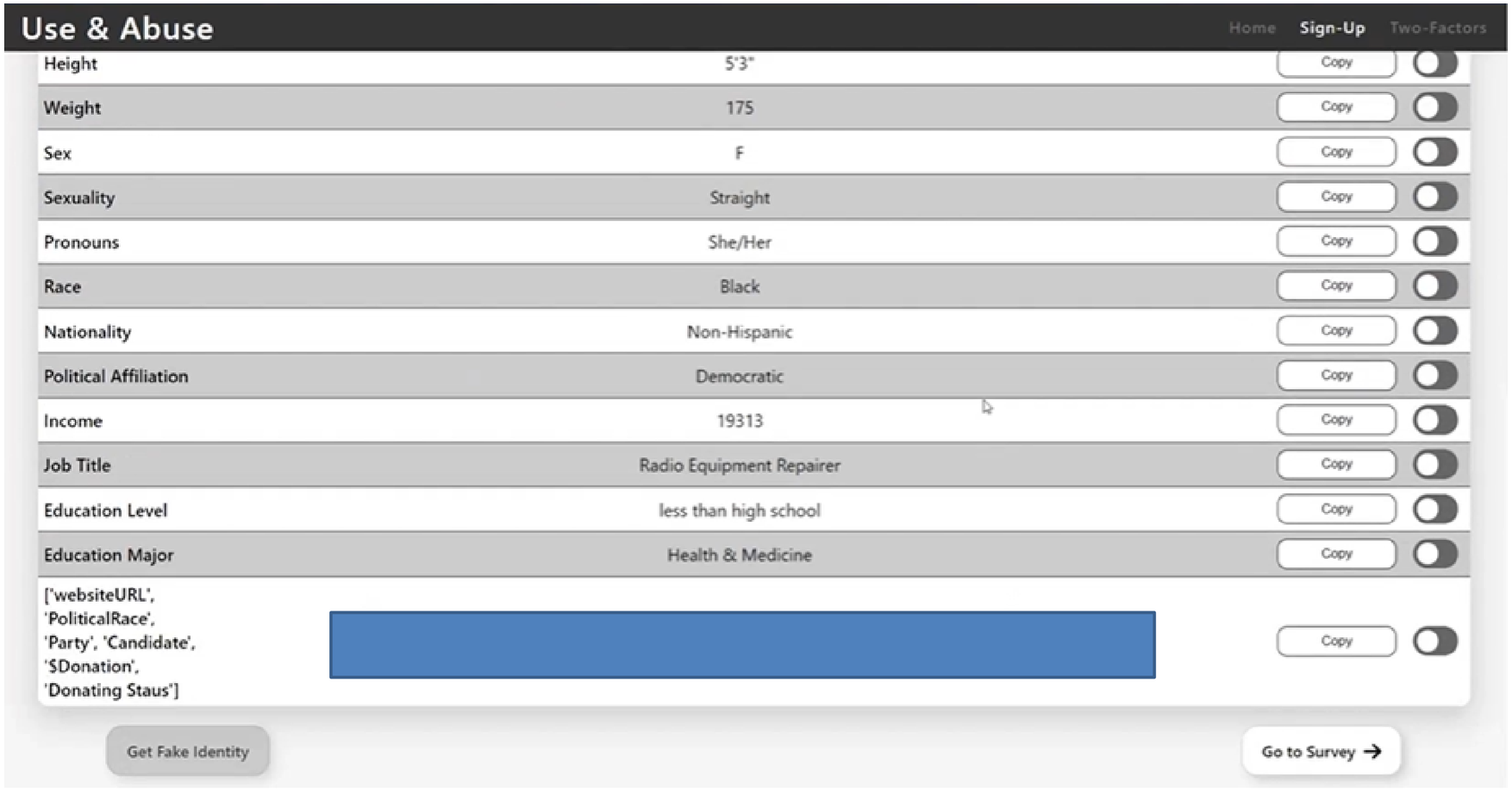
Who's hearing what?

Which sorts or types of identities are being targeted by different modes of communication?

Methodology



Sign-up Engine



Findings – Unrelated Domains

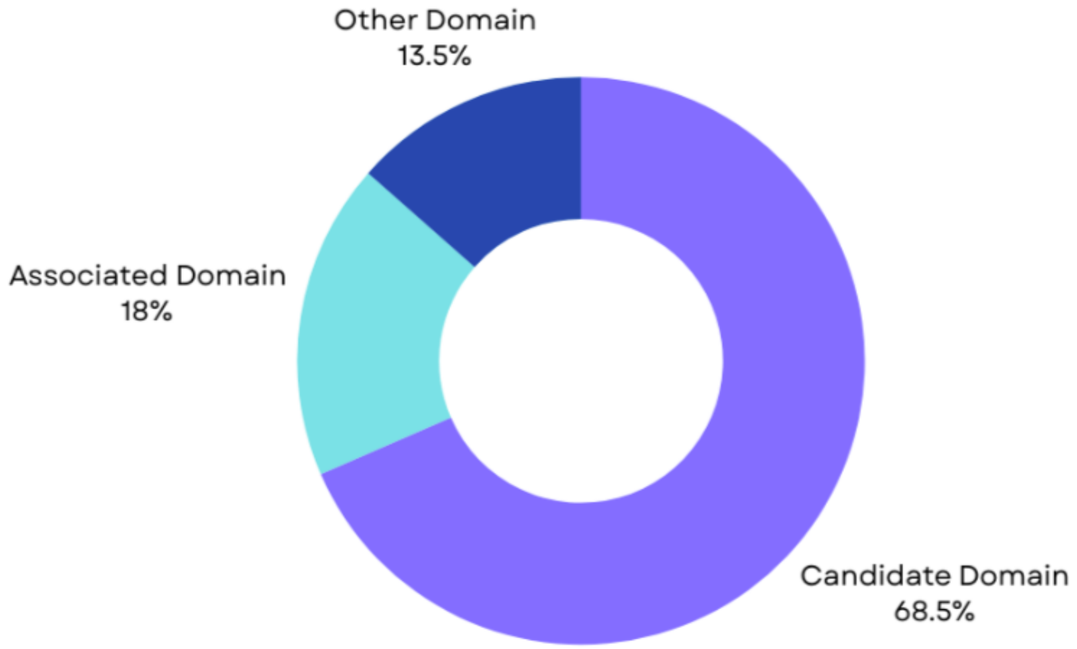


Figure 5. Pie chart of each category of domains

Findings – Donor vs Non-Donor

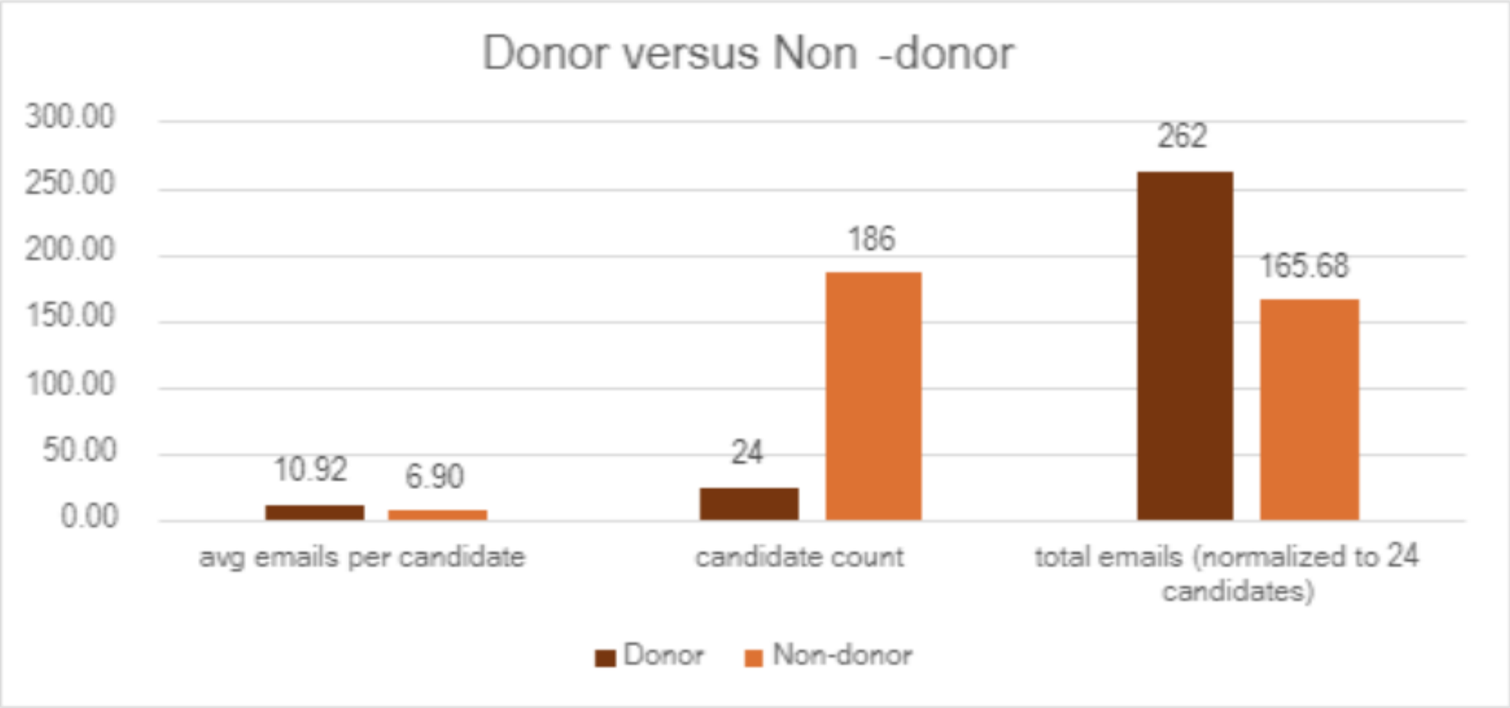


Figure 8. Donor Vs. Non-Donor

Findings – Content

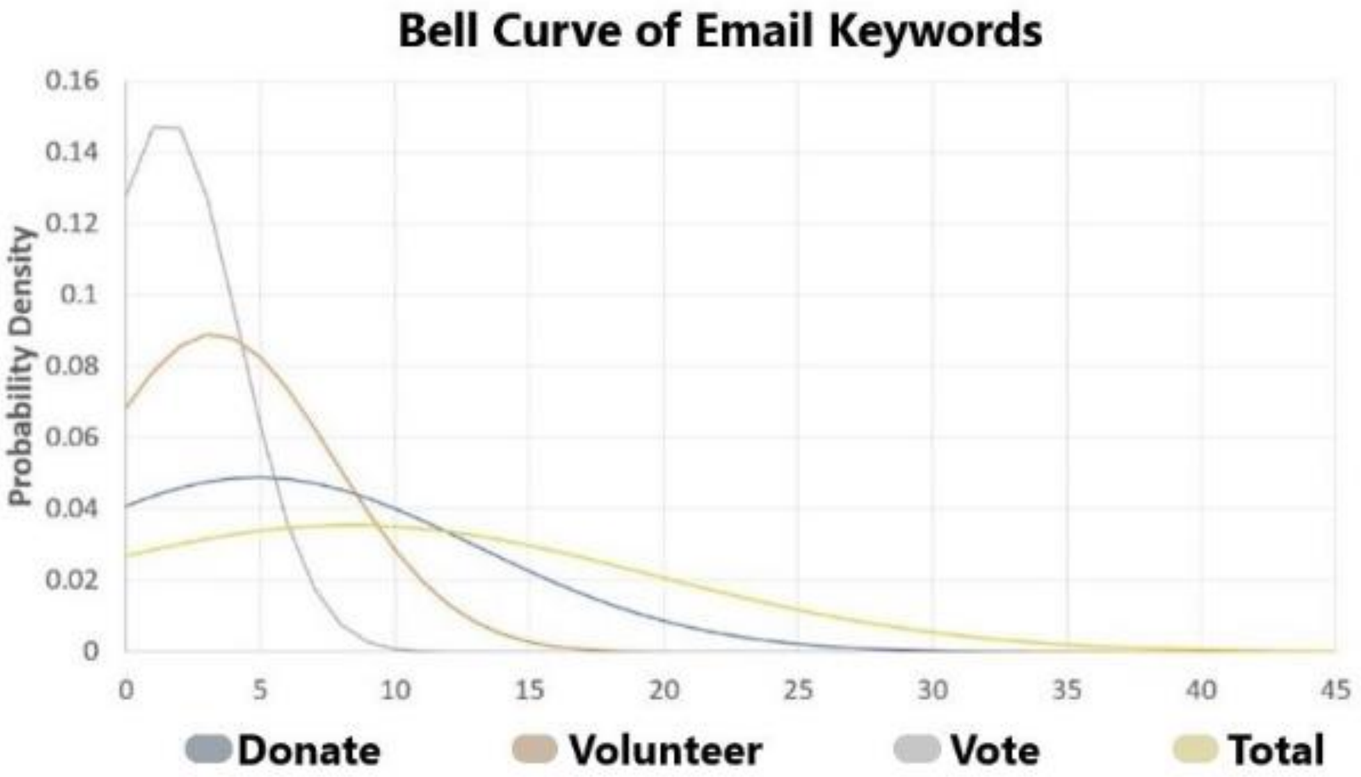
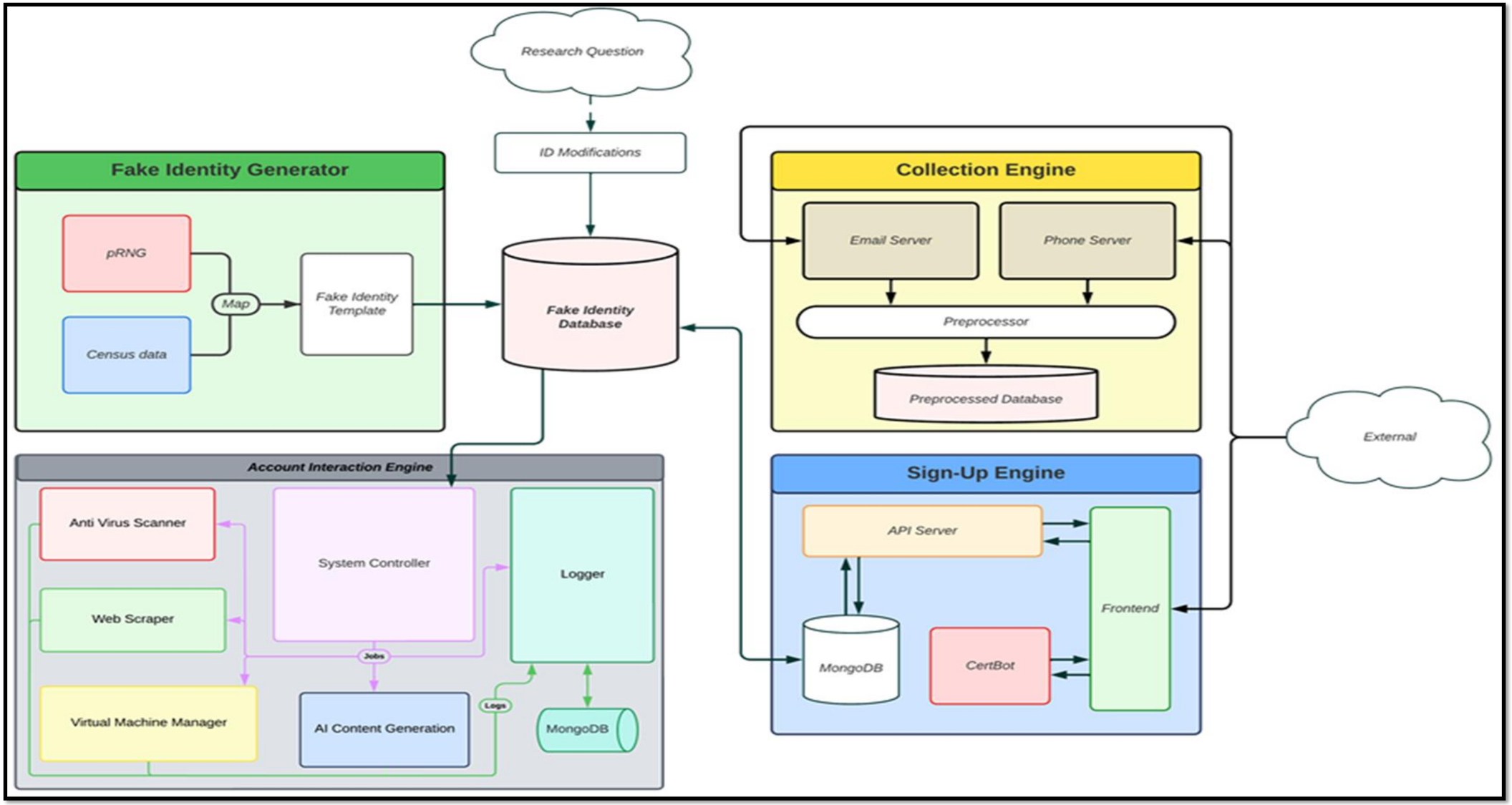


Figure 11: Probability of selected keywords in emails categorized in groups.

Findings – Sentiment

Interaction Engine



What's Next?

- Interaction Engine
- New wave of undergraduates; new experiments
- The Dark Web
- Fake IDs

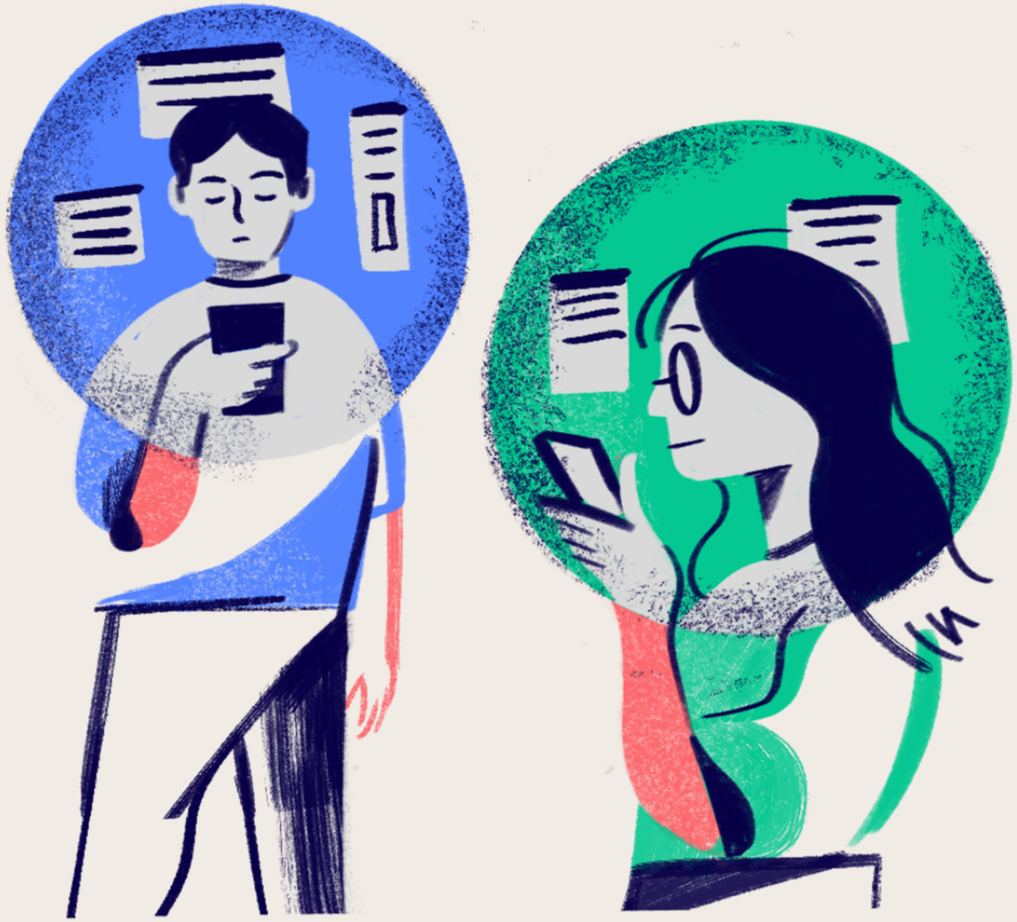
Tips to Protect Your PII Online

https://youtu.be/-ni_PWxrsNo

Presented by Mary Nerayo

Thank you!

Questions?



VIRGINIA TECH

mnerayo@vt.edu

References

- https://www.upguard.com/blog/biggest-data-breaches-us
- https://nationalsecurity.vt.edu/
- https://usnwc.libguides.com/c.php?g=494120&p=3381426
- https://www.ibm.com/topics/osint
- https://www.ibm.com/topics/pii#:~:text=IBM,email%20address%20or%20phone%20number.
- https://www.upguard.com/blog/biggest-data-breaches-us
- https://youtu.be/vyRsarL-dqk
- https://youtu.be/-ni_PWxrsNo
- https://www.bizjournals.com/atlanta/news/2020/11/18/movie-biz-holiday-season-wonder-woman.html
- https://fieldcheck.biz/library/customer-data-collection.html
- https://www.wpri.com/business/press-releases/cision/20230328NY53762/uber-eats-launches-u-s-certified-virtual-restaurant-program
- https://www.petbusiness.com/industry-news/walmart-discontinuing-sale-of-pet-fish/article_9b804d5d-7dc7-59f0-8b3f-700979fd87df.html
- https://www.google.com/url?sa=i&url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FAmerican_Automobile_Association&psig=AOvVaw3Z41YgrgWryDYVBSwDHNxr&ust=1728323668167000&source=images&cd=vfe&opi=89978449&ved=0CBcQjhxqFwoTCMjW3tKp-ogDFQAAAAAdAAAAABAE