SoURCE CODE
(Securing our Underlying Resource in Cyber Environments)
Proposers' Day

Kristopher W. Reese, PhD | Program Manager | Oct. 5, 2023

Intelligence Advanced Research Projects Activity

IARPA

Creating Advantage through Research and Technology

Welcome to the SoURCE CODE Proposers’ Day

Thank you for your interest in this program and participating in this event

To assure a clear broadcast stream, audio and video are disabled for meeting participants

Comments and questions can be submitted to the IARPA team via the WebEx chat tool submission or via index cards for in-person attendees

- Please direct questions to “All Panelists” in the chat if you are virtual

Questions submitted to the alias (dni-SoURCE-CODE-proposers-day@iarpa.gov) prior to this meeting and during this presentation, and corresponding answers, may be posted in writing online

Disclaimers

- This presentation is provided solely for information and planning purposes
- The Proposers' Day does not constitute a formal solicitation for proposals or proposal abstracts
- Nothing said at Proposers' Day changes the requirements set forth in a BAA
- **The BAA language supersedes anything presented or said by IARPA at the Proposers' Day**
- This meeting is being recorded and will be posted for public viewing
- For those viewing the recording, email aliases and POCs may be dated, please refer to IARPA.gov for updated information.

Proposers' Day Goals

1. Familiarize participants with IARPA's interest in the SoURCE CODE program and solicit questions and feedback
2. Foster discussion of complementary capabilities among potential program participants, i.e., TEAMING
   - Teaming information can be found at the following address: https://www.iarpa.gov/research-programs/source-code
   - An attendance list, with contact information of participants who approved of sharing will be distributed soon
   - The chat feature is enabled for participants to plan future discussions associated with teaming
   - Teaming interests, capability summaries, and lightning talk slides will be posted publicly on the IARPA SoURCE CODE webpage until the BAA submission period closes

Please ask questions and provide feedback, this is your chance to alter the course of events.
Please talk with others, find great team members.

Teaming

- Participants are encouraged to find partners and collaborators . . . someone might have a missing piece of your puzzle.
- Lightning talks will take place following the Program presentations.
- Collaborating and capability summaries will be accepted, with minimal review for appropriateness, and made available to the public.
  - Teaming documents and summaries can be submitted until the BAA closes, submit to dni-SoURCE-CODE-proposers-day@iarpa.gov.
  - If you would prefer your information not be shared (any recorded videos cannot be modified or removed) email dni-iarpa-source-code-proposersday@iarpa.gov.

Feedback and Questions

- Questions can be submitted until 11:00am ET.
- There will be a break after the contracting presentation at 11:00am ET.
- Responses to selected questions will be broadcast at 12:30pm ET, so please don't log out or close your WebEx connection.
  - All programmatic and contractual questions will be captured but will not be answered in this session
- Feedback (but not questions) about the draft technical section may be submitted to the IARPA team email at dni-SoURCE-CODE-proposers-day@iarpa.gov.
  - A new alias will be established when the full BAA is released
- After this Proposers' Day, IARPA will review all the feedback received for a final BAA to be posted on SAM.gov.

Agenda

Time	Topic	Speaker
9:30am-9:40am	Welcome, Logistics, Proposers’ Day Goals	Kristopher W. Reese, Program Manager
9:40am-9:50am	IARPA Overview	Robert Rahmer, Director Office of Analysis Research, IARPA
9:50am-10:40am	SoURCE CODE Program Overview	Kristopher W. Reese
10:40am-11:00am	Contracting Overview	TBD
11:00am-12:30pm	**Break (Submit questions in chat or drop boxes before 11:00am)**	
12:30pm-1:30pm	Answers to Selected Technical Questions	Kristopher W. Reese
1:30pm-1:35pm	Introductions to Lightning Talks	Kristopher W. Reese
1:35pm-4:25pm (est.)	Lightning Talks	Potential Performers
4:25pm-5:30pm	Informal Teaming Discussions*	In-Person Participants

*The Government will not attend these events

LIGHTNING TALKS AGENDA

Time	Speaker	Institution	In person
1:35pm-1:40pm	Xiangyu Zhang / Lin Tan	Purdue	Yes
1:40pm-1:45pm	Kexin Pei	University of Chicago	Yes
1:45pm-1:50pm	Michael V Le	IBM	Yes
1:50pm-1:55pm	Mike Murphy	SimSpace	Yes
1:55pm-2:00pm	Shiqing Ma	UMASS Amherst	Yes
2:00pm-2:05pm	William Liu	CACI	Yes
2:05pm-2:10pm	Sheikh Rabiul Islam	Rutgers University - Camden	Yes
2:10pm-2:15pm	Andrew Hendela	Karambit.AI	Yes
2:15pm-2:20pm	Aleksey Nogin	Red Balloon Security	Virtual

LIGHTNING TALKS AGENDA CONTINUED

Time	Speaker	Institution	In person
2:20pm-2:25pm	Thomas Wahl	GrammaTech	Yes
2:25pm-2:30pm	Dan Thomsen	SIFT	Virtual
2:30pm-2:35pm	Nathan Clark	Noblis	Yes
2:35pm-2:40pm	Chris Taylor	Tactical Computing Labs	Yes
2:40pm-2:45pm	Tomas Pevny	Czech Technical University in Prague	Virtual

Break and Informal Teaming Discussion at end of talks.

IARPA Overview

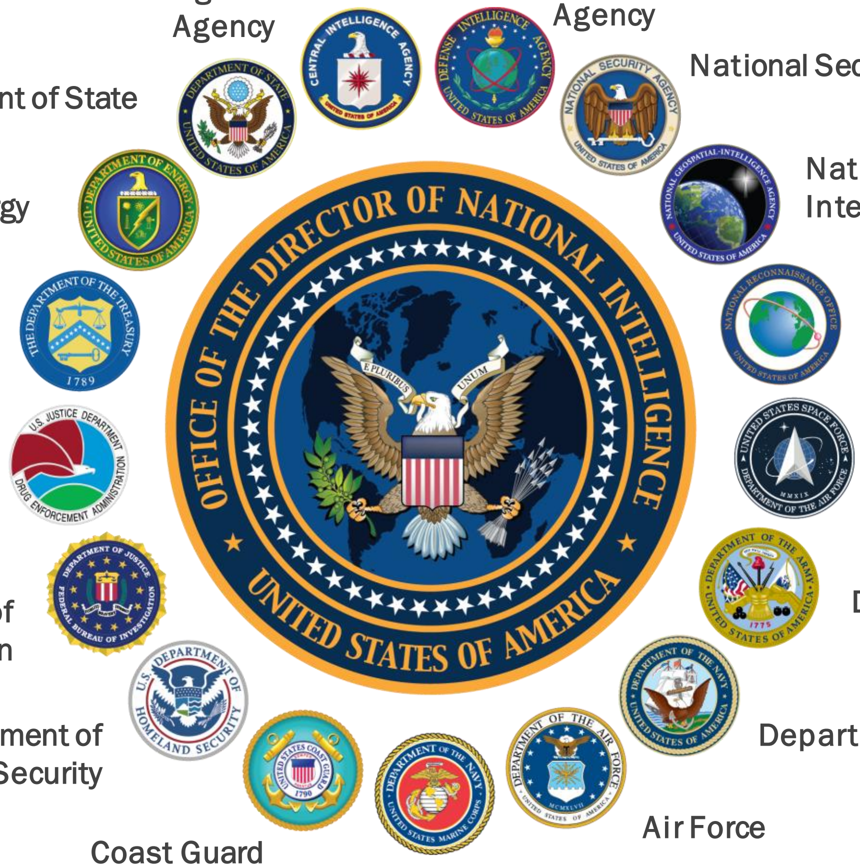
Robert Rahmer | Director, IARPA Office of Analysis | SoURCE CODE Proposers' Day | Oct. 5, 2023

Intelligence Advanced Research Projects Activity

IARPA

Creating Advantage through Research and Technology

Office of the Director of National Intelligence



Central Intelligence Agency

Defense Intelligence Agency

National Security Agency

National Geospatial-Intelligence Agency

National Reconnaissance Office

Space Force

Department of the Army

Department of the Navy

Air Force

Marine Corps

Coast Guard

Department of Homeland Security

Federal Bureau of Investigation

Drug Enforcement Administration

Department of the Treasury

Department of Energy

Department of State

IARPA Mission

IARPA envisions and leads *high-risk, high-payoff research* that delivers innovative technology *for future overwhelming intelligence advantage*

- Our problems are complex and multidisciplinary
- We emphasize technical excellence & technical truth

IARPA Method

- **Bring the best minds to bear on our problems**
  - Full and open competition to the greatest possible extent
  - World-class, term-limited Program Managers
- **Define and execute research programs that:**
  - Have goals that are clear, ambitious, credible and measurable
  - Run from three to five years
  - Publish peer-reviewed results and data, to the greatest possible extent
  - Employ independent and rigorous Test & Evaluation
  - Involve IC partners from start to finish
  - Transition new capabilities to intelligence community partners

IARPA R&D

- Technical and programmatic excellence are required
- Each program has a clearly defined and measurable end-goal
  - Intermediate milestones to measure progress are also required
  - Every program has a beginning and an end
- This approach, coupled with term-limited PM positions, ensures
  - IARPA does not “institutionalize“ programs
  - Fresh ideas and perspectives are always coming in
  - Status quo is always questioned
  - Only the best ideas are pursued, and only the best performers are funded

IARPA Snapshot

IARPA's research portfolio is diverse, including math, physics, chemistry, biology, microelectronics, neuroscience, linguistics, political science, cognitive psychology, and more.

- 70% of completed research transitions to U.S. Government partners
- 3,000+ journal articles published
- IARPA funded researchers have been awarded the Nobel Prize in Physics for quantum computing research, a MacArthur Fellowship, and a Bell prize
- IARPA serves on National Science and Technology Council (NSTC) committees and actively engages with the White House BRAIN Initiative, National Strategic Computing Initiative, and the NSTC Select Committee on Artificial Intelligence, the NSTC Subcommittee on Quantum Information Science (SCQIS), and NSTC Subcommittee on Economic and Security Implications of Quantum Science (ESIX)

How to Engage with IARPA

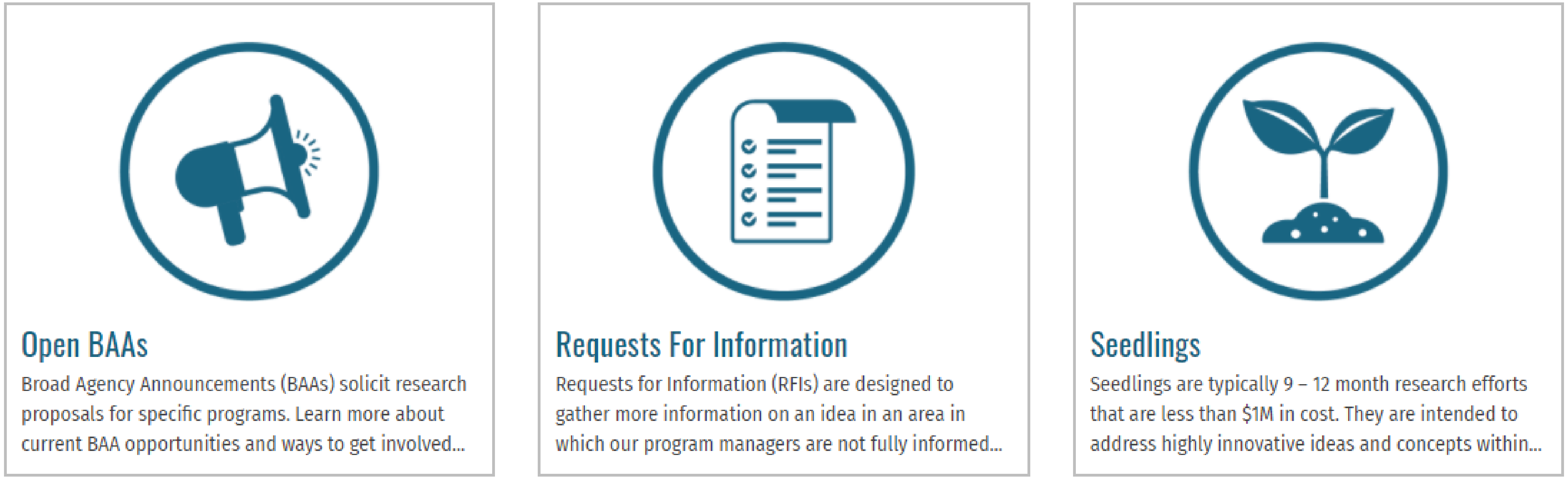
ENGAGE WITH US

Throughout our website you can learn more about engaging with us on our highly innovative work that is having a positive impact in the Intelligence Community and society in general. Click on any of the below links to learn more.

iarpa.gov | 301-243-1995

dni-iarpa-info@iarpa.gov

- Reach out to our Program Managers.
- Schedule a visit if you are in the DC area or invite us to visit you



Open BAAs

Broad Agency Announcements (BAAs) solicit research proposals for specific programs. Learn more about current BAA opportunities and ways to get involved...

Requests For Information

Requests for Information (RFIs) are designed to gather more information on an idea in an area in which our program managers are not fully informed...

Seedlings

Seedlings are typically 9 – 12 month research efforts that are less than $1M in cost. They are intended to address highly innovative ideas and concepts within...

SoURCE CODE Overview

Kristopher W. Reese, PhD | Program Manager | Oct. 5, 2023

Intelligence Advanced Research Projects Activity

IARPA

Creating Advantage through Research and Technology

SoURCE CODE Program Goals

SoURCE CODE seeks to create automated, scientifically validated forensic similarity and demographic analytic technologies

- Measure similarity of code and binaries and identify components that may analyze hidden demographic information.

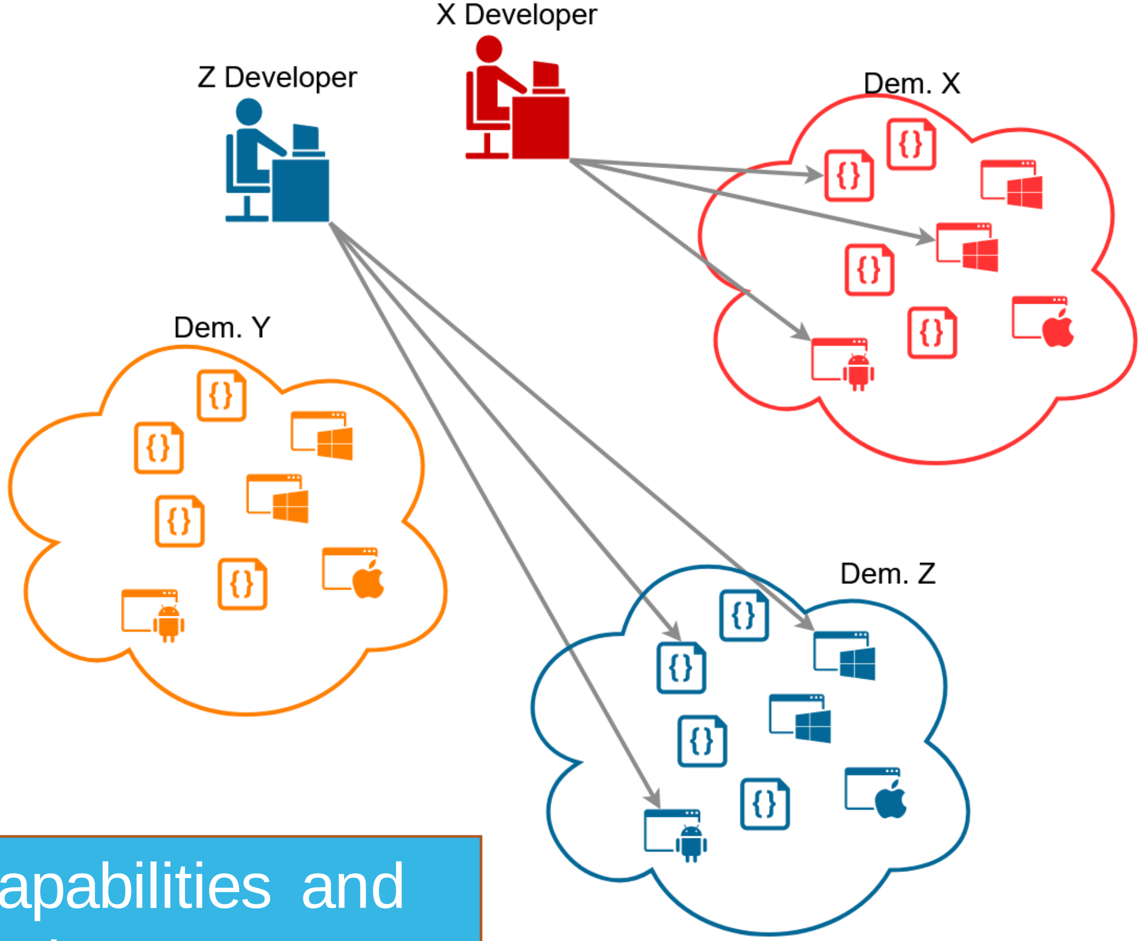


Image is UNCLASSIFIED

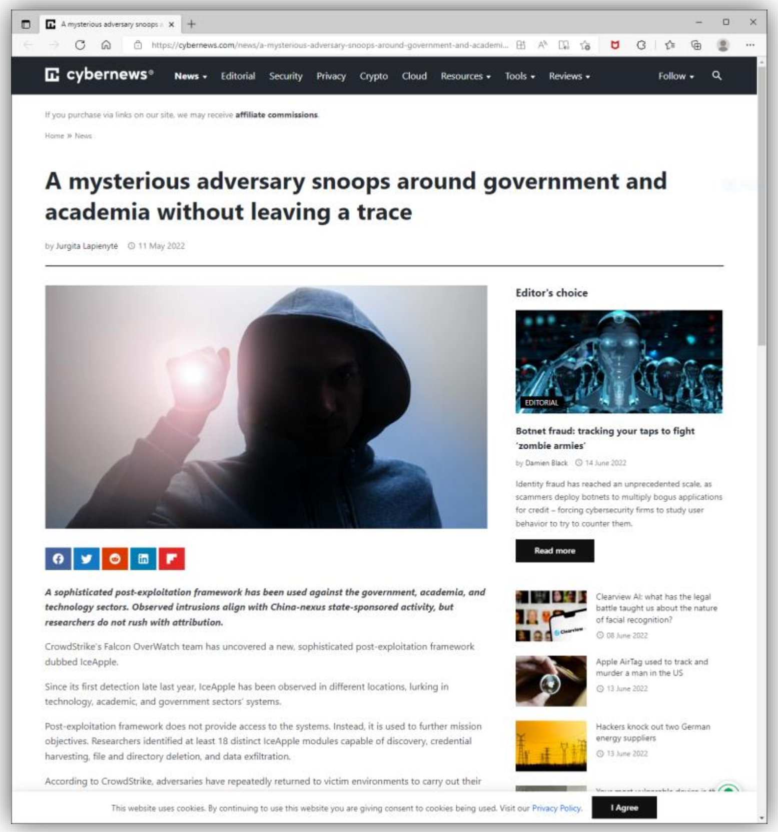
(U) SoURCE CODE will improve forensic capabilities and speed up threat intelligence analytics!

Problem Statement

- Current methods are highly manual, requiring substantial human expertise, training, and time to conduct forensic analysis of code.
  - Prior automated attempts leverage a small subset of static analysis features Lexical and syntactic.
- Executable binaries and corresponding source code include numerous other features that can measure similarity, and thereby assist in attribution and demographic analytics.
- SoURCE CODE will develop capabilities to use the full feature space to measure similarity, especially between source code and binaries

(U) The full feature set of source code and binary features are an untapped resource for forensic capabilities.

Why is Attribution important?


A mysterious adversary snoops
https://cybernews.com/news/a-mysterious-adversary-snoops-around-government-and-academi...
cybernews®
News
Editorial
Security
Privacy
Crypto
Cloud
Resources
Tools
Reviews
Follow
EDITORIAL
Read more
I Agree

If you purchase via links on our site, we may receive **affiliate commissions**.

Home » News

A mysterious adversary snoops around government and academia without leaving a trace

by Jurgita Lapienytė 11 May 2022

A sophisticated post-exploitation framework has been used against the government, academia, and technology sectors. Observed intrusions align with China-nexus state-sponsored activity, but researchers do not rush with attribution.

CrowdStrike's Falcon OverWatch team has uncovered a new, sophisticated post-exploitation framework dubbed IceApple.

Since its first detection late last year, IceApple has been observed in different locations, lurking in technology, academic, and government sectors' systems.

Post-exploitation framework does not provide access to the systems. Instead, it is used to further mission objectives. Researchers identified at least 18 distinct IceApple modules capable of discovery, credential harvesting, file and directory deletion, and data exfiltration.

According to CrowdStrike, adversaries have repeatedly returned to victim environments to carry out their

Editor's choice

Botnet fraud: tracking your taps to fight 'zombie armies'

by Damien Black 14 June 2022

Identity fraud has reached an unprecedented scale, as scammers deploy botnets to multiply bogus applications for credit – forcing cybersecurity firms to study user behavior to try to counter them.

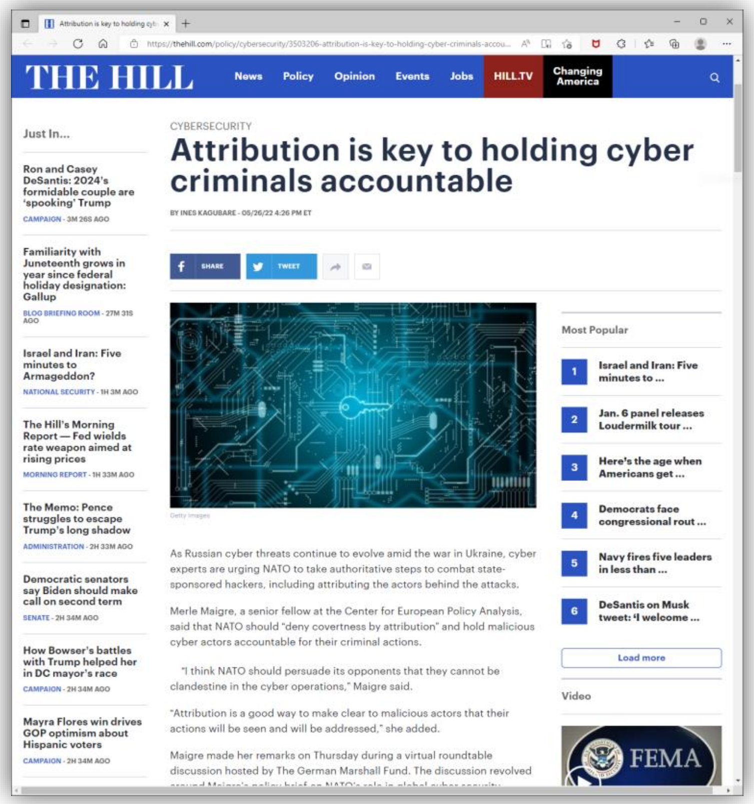
Clearview AI: what has the legal battle taught us about the nature of facial recognition? 08 June 2022

Apple AirTag used to track and murder a man in the US 13 June 2022

Hackers knock out two German energy suppliers 13 June 2022

This website uses cookies. By continuing to use this website you are giving consent to cookies being used. Visit our Privacy Policy.

[2]


Attribution is key to holding cyb
https://thehill.com/policy/cybersecurity/3503206-attribution-is-key-to-holding-cyber-criminals-accou...
THE HILL
News
Policy
Opinion
Events
Jobs
HILL.TV
Changing America
SHARE
TWEET
Load more
FEMA

CYBERSECURITY

Attribution is key to holding cyber criminals accountable

BY INES KAGUBARE - 05/26/22 4:26 PM ET

Getty Images

As Russian cyber threats continue to evolve amid the war in Ukraine, cyber experts are urging NATO to take authoritative steps to combat state-sponsored hackers, including attributing the actors behind the attacks.

Merle Maigre, a senior fellow at the Center for European Policy Analysis, said that NATO should "deny covertness by attribution" and hold malicious cyber actors accountable for their criminal actions.

"I think NATO should persuade its opponents that they cannot be clandestine in the cyber operations," Maigre said.

"Attribution is a good way to make clear to malicious actors that their actions will be seen and will be addressed," she added.

Maigre made her remarks on Thursday during a virtual roundtable discussion hosted by The German Marshall Fund. The discussion revolved

Just In...

Ron and Casey DeSantis: 2024's formidable couple are 'spooking' Trump
CAMPAIGN - 3M 26S AGO

Familiarity with Juneteenth grows in year since federal holiday designation: Gallup
BLOG BRIEFING ROOM - 27M 31S AGO

Israel and Iran: Five minutes to Armageddon?
NATIONAL SECURITY - 1H 3M AGO

The Hill's Morning Report — Fed wields rate weapon aimed at rising prices
MORNING REPORT - 1H 33M AGO

The Memo: Pence struggles to escape Trump's long shadow
ADMINISTRATION - 2H 33M AGO

Democratic senators say Biden should make call on second term
SENATE - 2H 34M AGO

How Bowser's battles with Trump helped her in DC mayor's race
CAMPAIGN - 2H 34M AGO

Mayra Flores win drives GOP optimism about Hispanic voters
CAMPAIGN - 2H 34M AGO

Most Popular

1 **Israel and Iran: Five minutes to ...**

2 **Jan. 6 panel releases Loudermilk tour ...**

3 **Here's the age when Americans get ...**

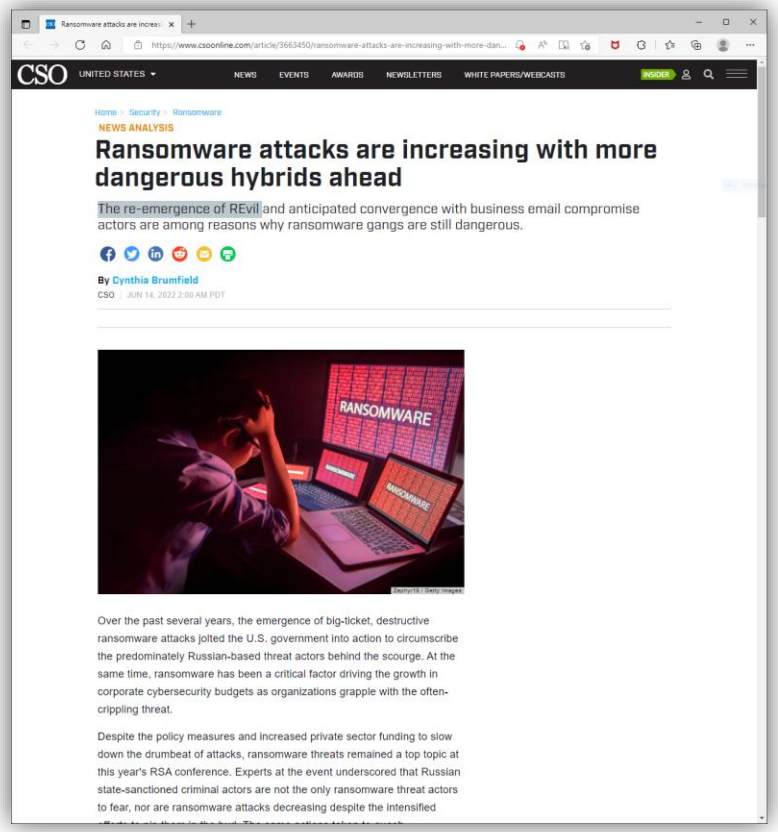
4 **Democrats face congressional rout ...**

5 **Navy fires five leaders in less than ...**

6 **DeSantis on Musk tweet: 'I welcome ...**

Video

[3]


Ransomware attacks are increas
https://www.csoonline.com/article/3663450/ransomware-attacks-are-increasing-with-more-dan...
CSO
UNITED STATES
NEWS
EVENTS
AWARDS
NEWSLETTERS
WHITE PAPERS/WEBCASTS
INSIDER
RANSOMWARE

Home > Security > Ransomware

NEWS ANALYSIS

Ransomware attacks are increasing with more dangerous hybrids ahead

The re-emergence of REvil and anticipated convergence with business email compromise actors are among reasons why ransomware gangs are still dangerous.

By Cynthia Brumfield

CSO | JUN 14, 2022 2:00 AM PDT

Zephyr18 / Getty Images

Over the past several years, the emergence of big-ticket, destructive ransomware attacks jolted the U.S. government into action to circumscribe the predominately Russian-based threat actors behind the scourge. At the same time, ransomware has been a critical factor driving the growth in corporate cybersecurity budgets as organizations grapple with the often-crippling threat.

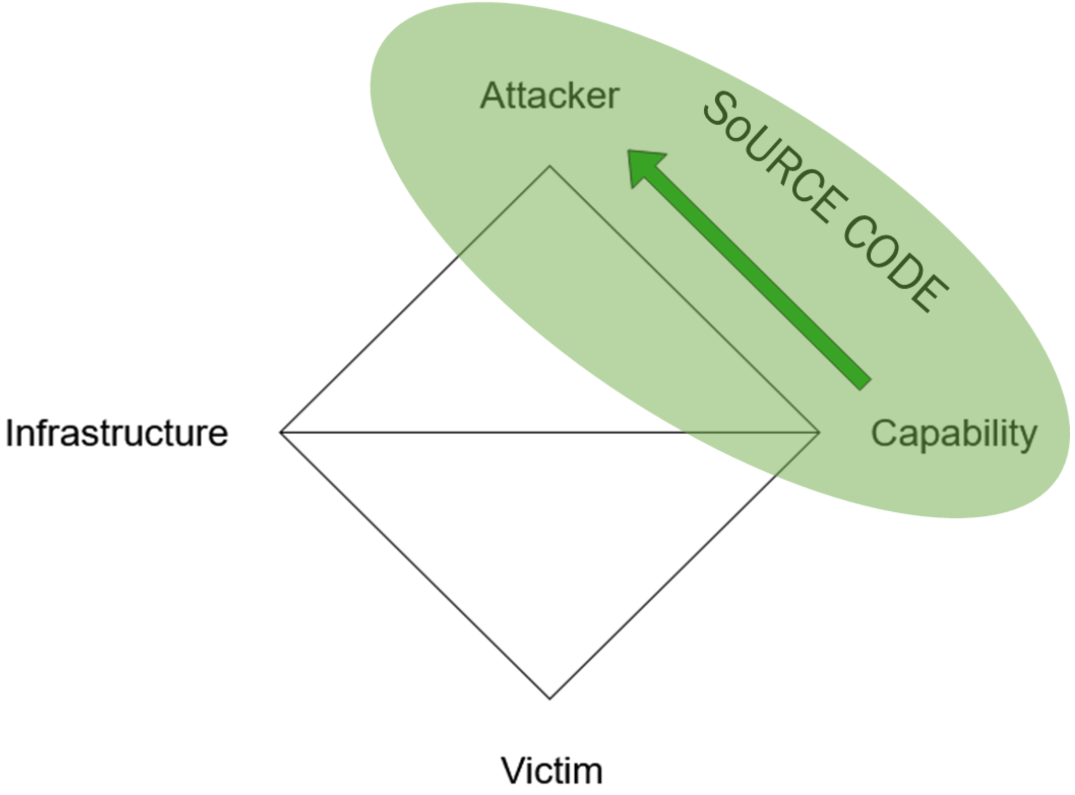
Despite the policy measures and increased private sector funding to slow down the drumbeat of attacks, ransomware threats remained a top topic at this year's RSA conference. Experts at the event underscored that Russian state-sanctioned criminal actors are not the only ransomware threat actors to fear, nor are ransomware attacks decreasing despite the intensified

[5]

Image is UNCLASSIFIED

Similarity and Demographic traits can play a role in attributing the increasing number of attacks around the world.

The Diamond Model of Forensic Attribution



SoURCE CODE will create similarity and demographic analytics to automate the "Attacker-Capability" edge.

Image modeled after original research paper [38]

Image is UNCLASSIFIED

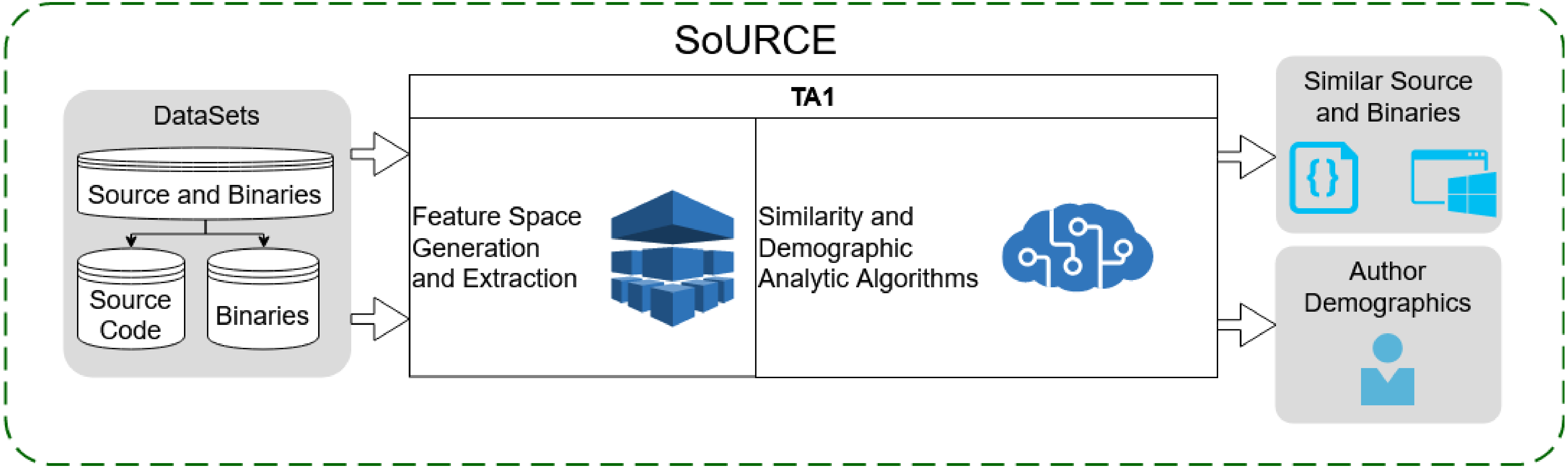
Technical Areas

Intelligence Advanced Research Projects Activity

IARPA

Creating Advantage through Research and Technology

SoURCE CODE Overview

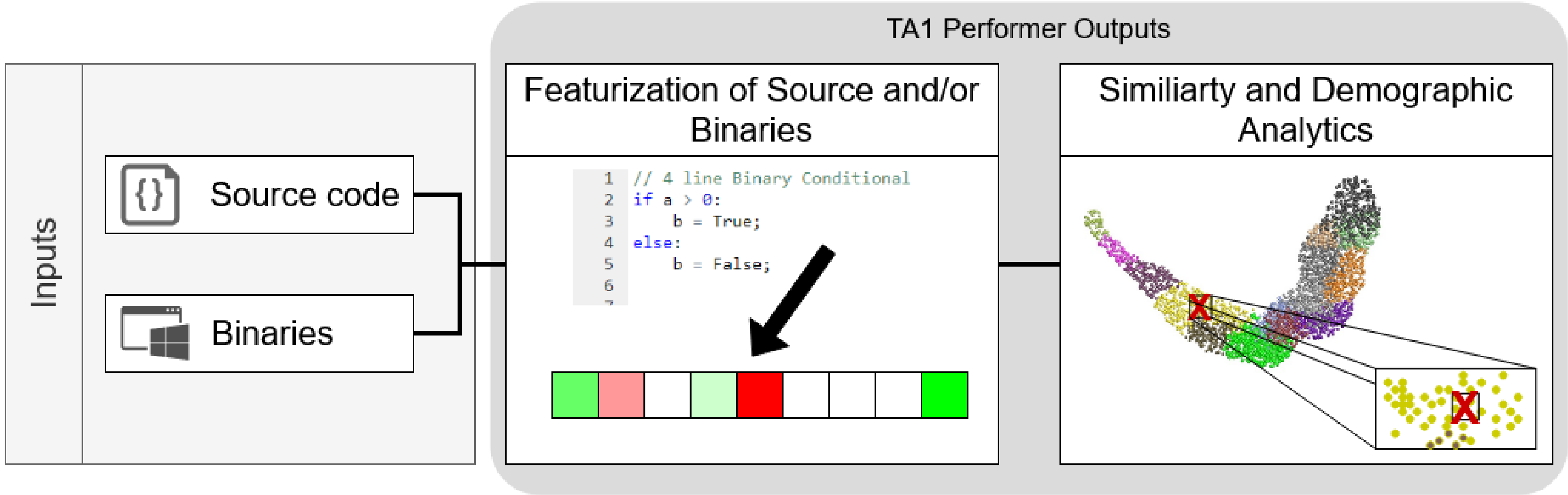


SoURCE
DataSets
Source and Binaries
Source Code
Binaries
TA1
Feature Space Generation and Extraction
Similarity and Demographic Analytic Algorithms
Similar Source and Binaries
Author Demographics

Image is UNCLASSIFIED

SoURCE CODE seeks to utilize the full feature set (focus area) to measure source code and binary similarity and demographic attributes (e.g. Country, Group, individual).

TA1: Measuring Similarity and Demographic Analytics


TA1 Performer Outputs
Inputs
Source code
Binaries
Featurization of Source and/or Binaries
Similiarty and Demographic Analytics

```
// 4 line Binary Conditional
if a > 0:
    b = True;
else:
    b = False;
```

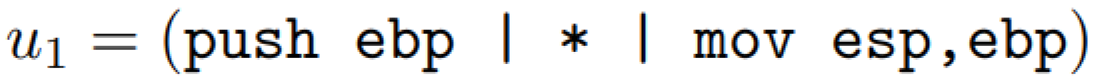
INNOVATIONS REQUIRED

1. Identify and map salient features in both binaries and/or source code that capture author style.
2. Identify and implement algorithms, or ensembles, that can effectively utilize authorship features to measure similarity and identify most likely author(s)
   a) **PHASE 1:** binary to binary, source code to source code
   b) **PHASE 2:** binary to source code, source code to binary
3. Explain similarity score / decision process to forensic experts to assist in making final attribution decisions.

Subset of Binary Features for Forensic Analysis

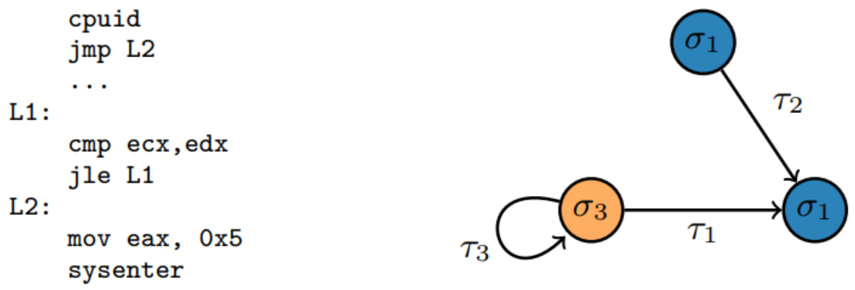
	Code Properties	Examples
Disassembled	**Instruction**	Byte level n-grams Idioms Instruction Summary Graphlets SuperGraphlets
	Control Flow	Instruction Summary Graphlets SuperGraphlets Call Graphlets
	External	Call Graphlets Library Calls
Decompiled	**Lexical**	Word Unigrams: • Integer types • Names of library functions • Names of Internal functions (when symbol information is available)
	Syntactic	Fuzzy Abstract Syntax Trees • AST n-grams • Labeled AST edges • AST Node TF-IDF • AST Node Avg. depth

Decompiled features are similar to source code features



$$u_1 = (\texttt{push ebp} \mid * \mid \texttt{mov esp,ebp})$$

Example 1: Code Idioms



```
      cpuid
      jmp L2
      ...
L1:
      cmp ecx,edx
      jle L1
L2:
      mov eax, 0x5
      sysenter
```

Example 2: Graphlets

Images screen captured from original research paper [21]

All images are UNCLASSIFIED

There are limited features in this space, and a complete exploration of possible features needs to be conducted – including learned features!

Subset of Source Code Features

Harder to utilize at scale

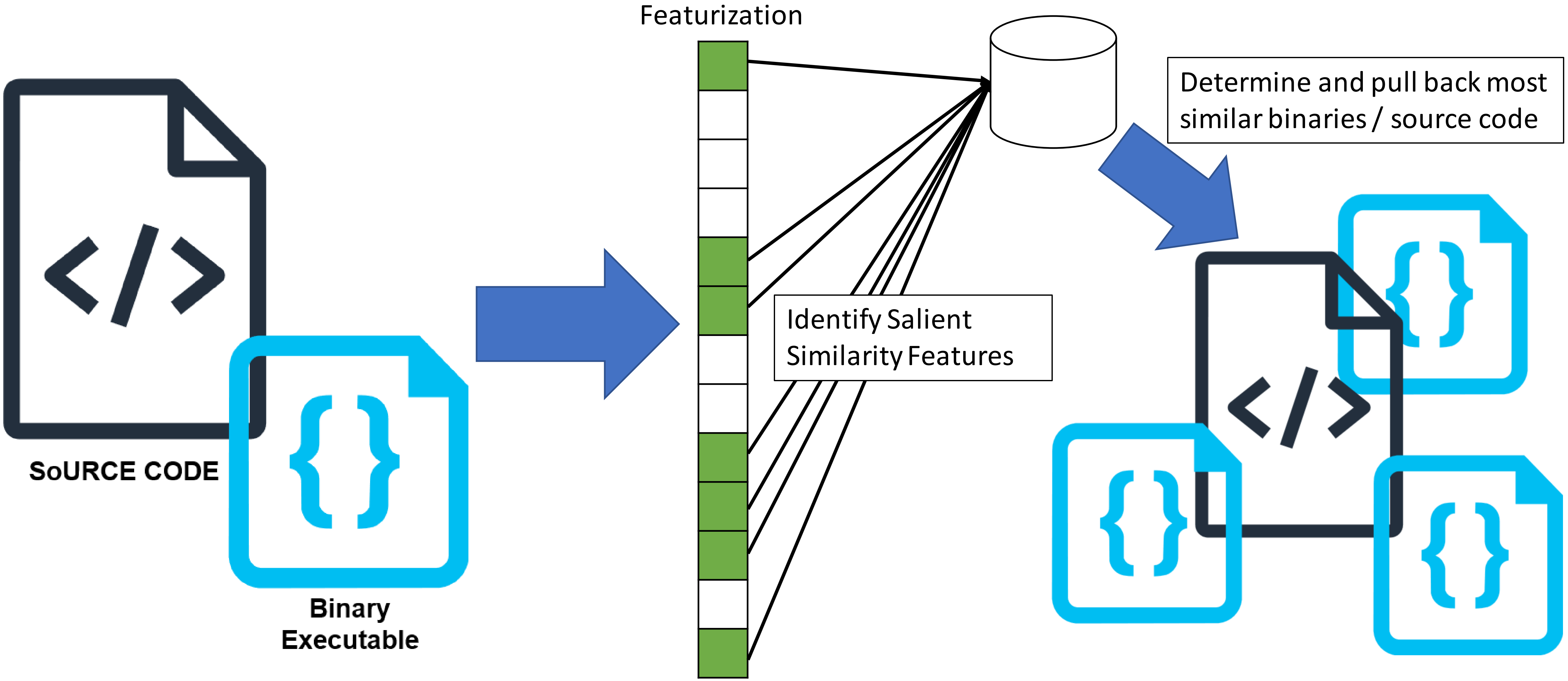
Feature Type	Examples
Lexical	Lines of Code Operands Variables Spaces Word n-grams Char n-grams Function names
Syntactic	Average function size Special Macros Data Structure choice Control Structure choice Input Statements Conditional Statements Assignment Statements
Semantic	Loops Dataflow analysis Control flow analysis Algorithms implemented Procedure-dependent analysis
Behavioral	System calls Files accessed Created mutex Visited URLs Dynamic values Network connections
App-dependent	Log file strings Error message file strings Property file strings

Current source code Authorship systems largely exploit Lexical and Syntactic structures, limiting forensic applications to unobfuscated / de-linted code

SoURCE CODE seeks to exploit the full feature set!

All images are UNCLASSIFIED

Notional Similarity Matching

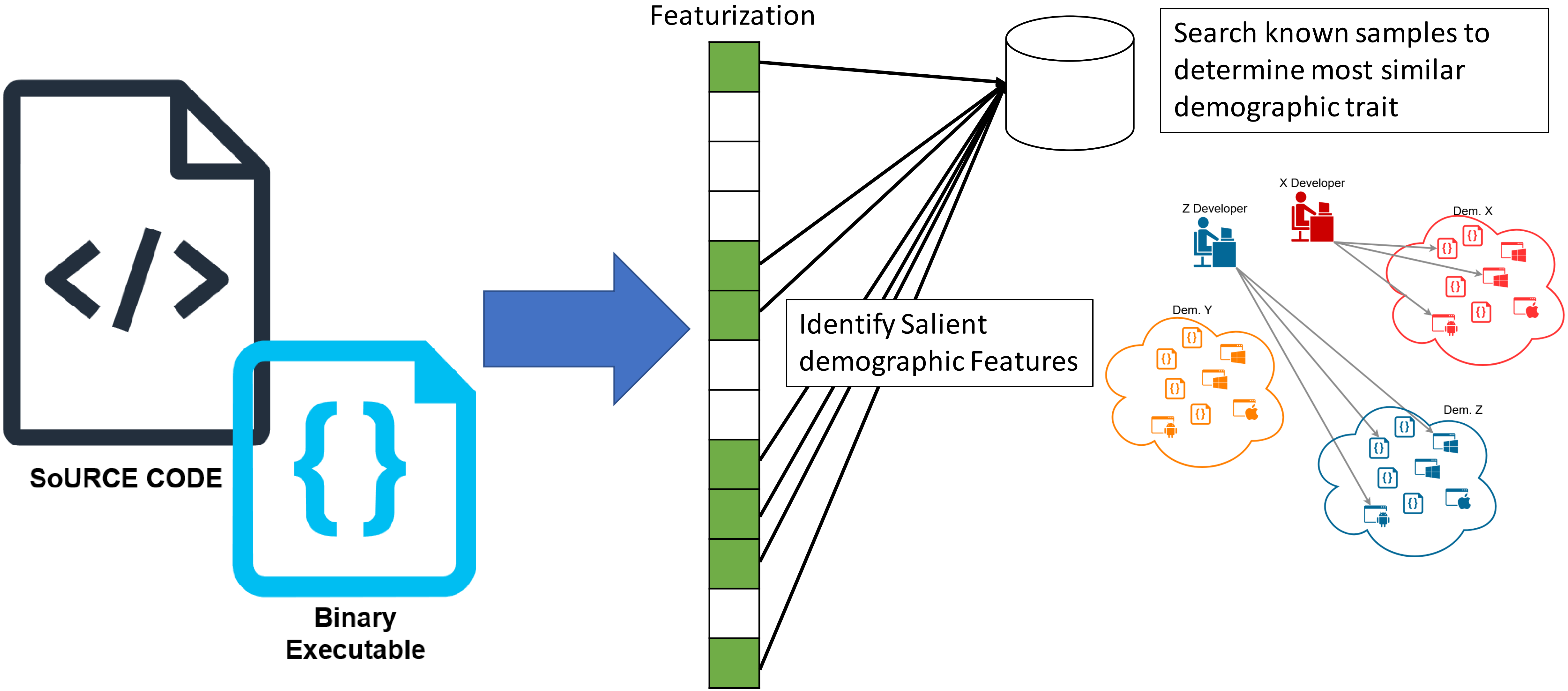


Featurization

Identify Salient Similarity Features

Determine and pull back most similar binaries / source code

Notional Demographic Analytics



Featurization

Search known samples to determine most similar demographic trait

Identify Salient demographic Features

SoURCE CODE

Binary Executable

Scientific Gaps: Underexplored Elements of Coding Style

- The following scientific gaps in authorship attribution of source code / binaries have been identified as being underexplored in literature, and may impact features identified making authorship determination more difficult:
  - Codebases vs a single Individual’s code
  - Understanding the impact and influences of Project Domains: e.g. Android vs. iOS vs. Windows vs. Linux
  - Understanding the impact and influences of specific development tools
    - Integrated Development Environments
    - Version Control Systems
    - Compiler, Build Environment, and Deployment tools
  - Impact of project naming conventions, company style guides, etc.

These gaps highlights the need for understanding the impact of standard coding practices over educational coding practices!

SoURCE CODE Program Phases

Intelligence Advanced Research Projects Activity

IARPA

Creating Advantage through Research and Technology

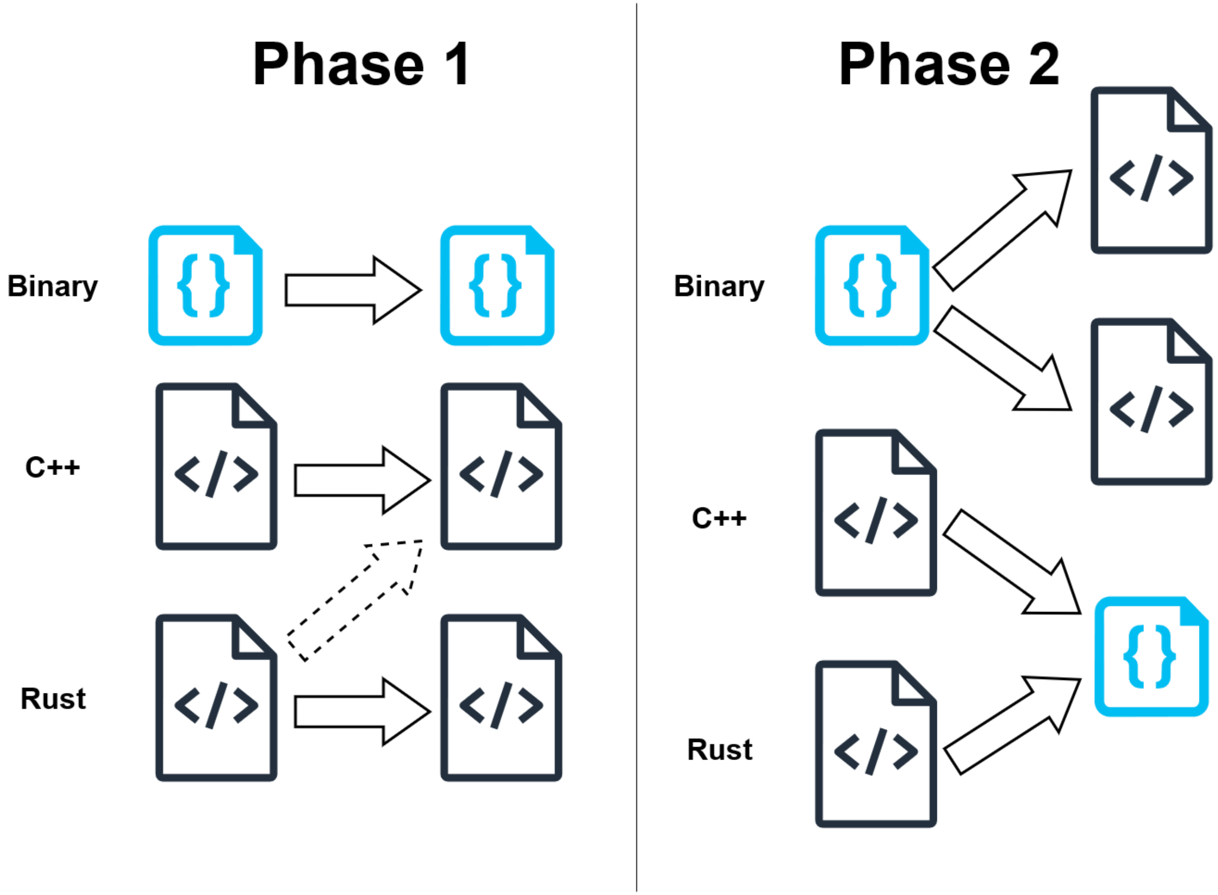
SoURCE CODE Program Schedule

		Phase 1																		Phase 2											
Task	**# Mo.**	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	1	2	3	4	5	6	7	8	9	10	11	12
Kickoff	-	★																		★											
TA1: Feature Space	30	🟩	🟩	🟩	★	🟩	🟩	🟩	★	🟩	🟩	🟩	★	🟩	🟩	🟩	★	🟩	🟩	🟩	🟩	🟩	★	🟩	🟩	🟩	★	🟩	🟩	🟩	★
TA1: Similarity	30	🟩	🟩	🟩	🟩	🟩	🟩	🟩	★	🟩	🟩	🟩	★	🟩	🟩	🟩	★	🟩	🟩	🟩	🟩	🟩	🟩	★	🟩	🟩	🟩	★	🟩	🟩	★
PI Virtual Meetings / Calls																															
PI Workshop (in-person)	-								★								★										★				
Site Meetings	-				★								★										★								
Program Closeout	-																														★

★ Milestone Deliverables

Image is UNCLASSIFIED

SoURCE CODE Phases



- Phase 1 programming languages pending – subject to data availability.
- Phase 2 programming languages subject to data availability
- Potential Proposers' can offer potential datasets
  - Proprietary Datasets cannot be approved unless they can be shared with all potential performers on the overall program or purchased for research purposes.

Possible Datasets – Source Code

- Google Code Jam:
  - Datasets from 2008-2020 competitions
  - Contains Author data as ground truth
  - No binaries – Can compile with different build environments
- Other Coding Competitions:
  - Codeforces.com (Russian-based; Coding Competitions)
  - Topcoders (US-based; Coding Competitions)

- Other datasets identified from RFI will be checked by T&E

(U) These competition datasets act as surrogates for Malware, but they do not represent Malware source code!

Possible Datasets - Binaries

- APTClass [31]
  - One of the Largest ground-truth datasets (15,000 samples)
  - Need to Request Access to download (UK University)
  - Data Sources / Ground Truth:

Source Name	Last Updated
MISP [32]	Oct. 2020
APT Operation Tracker [33]	Oct. 2020
MITRE ATT&CK [34]	Oct. 2020
sapphirex00 [35]	Nov. 2018
Thailand CERT [36]	Oct. 2020
Council on Foreign Relations [37]	Oct. 2020

  - Other datasets identified from RFI will be checked by T&E

Evaluation and Metrics

Intelligence Advanced Research Projects Activity

IARPA

Creating Advantage through Research and Technology

Test & Evaluation

Robust, independent test and evaluation is a crucial part of every IARPA program

- For SoURCE CODE, T&E will be responsible for providing data and product evaluation.

Performer systems will be executed by T&E on stand alone systems / networks

- Specifications of the SoURCE CODE Test System will be provided at a later date
- SoURCE CODE anticipates using multiple T&E teams for various aspects of the overall program.

TA1 – Testing and Evaluation Plan

1) T&E will provide interface for automatically running experiments in the feature space and similarity algorithms.

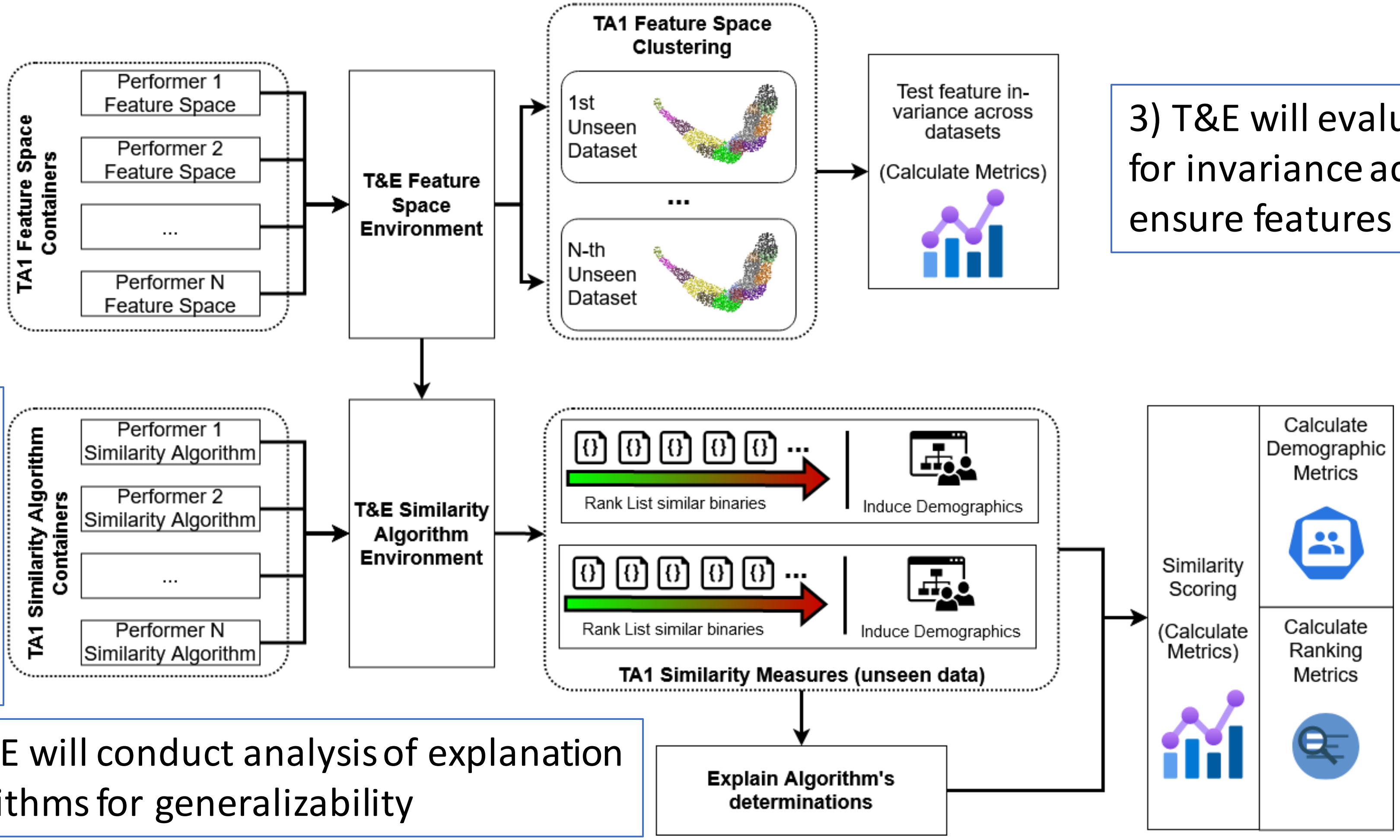
2) Performers will submit Feature Space findings and extractions to T&E.

3) T&E will evaluate feature space for invariance across datasets to ensure features are salient.

4) Performers will submit similarity and demographic algorithms to T&E for evaluation and experimentation

5) T&E will compare algorithms to baseline systems and calculate metrics in similarity scoring and demographic analytics

6) T&E will conduct analysis of explanation algorithms for generalizability


TA1 Feature Space Containers
Performer 1 Feature Space
Performer 2 Feature Space
...
Performer N Feature Space
T&E Feature Space Environment
TA1 Feature Space Clustering
1st Unseen Dataset
...
N-th Unseen Dataset
Test feature in-variance across datasets
(Calculate Metrics)
TA1 Similarity Algorithm Containers
Performer 1 Similarity Algorithm
Performer 2 Similarity Algorithm
...
Performer N Similarity Algorithm
T&E Similarity Algorithm Environment
{} {} {} {} {} ...
Rank List similar binaries
Induce Demographics
{} {} {} {} {} ...
Rank List similar binaries
Induce Demographics
TA1 Similarity Measures (unseen data)
Explain Algorithm's determinations
Similarity Scoring
(Calculate Metrics)
Calculate Demographic Metrics
Calculate Ranking Metrics

TA1 Metrics: Similarity Task Metrics

	Phase I	Phase II
Top-1 Accuracy [100/10,000]	85% [95/75]	90% [98/90]
Top-10 Accuracy [100/10,000]	95% [99/85]	96% [99/95]
EER Average	30%	20%
d' (sensitivity index)	Measured	increase 30%
Attribution methods (data permitting)	Source ⟶ Source, Binary ⟶Binary	Source ⟶ Binary, Binary ⟶Source

Metrics shown are for 1000 unique authors and should scale to match cardinality of unique users – additional numbers shown for 100 / 10,000 authors in brackets.

Additional metrics will also be measured to better understand the efficacy of the algorithms:

- Precision
- Specificity
- FAR/TAR
- FRR/TRR

Detection Error Tradeoff / Receiver Operator Curve

TABLE is UNCLASSIFIED

TA1 Metrics: Demographic Analytic Metrics

	Phase I	Phase II
Top-1 Accuracy [100/10,000]	85% [95/75]	90% [98/90]
Top-10 Accuracy [100/10,000]	95% [99/85]	96% [99/95]
EER Average	20%	10%
# Groups/Demographics (data permitting)	50	70+
Type of Set	Closed	Open

Additional metrics will also be measured to better understand the efficacy of the algorithms:

- Precision
- Specificity
- FAR/TAR
- FRR/TRR

→ Detection Error Tradeoff / Receiver Operator Curve

TABLE is UNCLASSIFIED

Accuracy, Precision, Recall, Specificity, FAR, and FRR

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TAR)	False Positive (FAR)
	Negative	False Negative (FRR)	True Negative (TRR)

Binary classification confusion matrix

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\cancel{Recall = \frac{TP}{TP + FN}}$$

$$Precision = \frac{TP}{TP + FP}$$

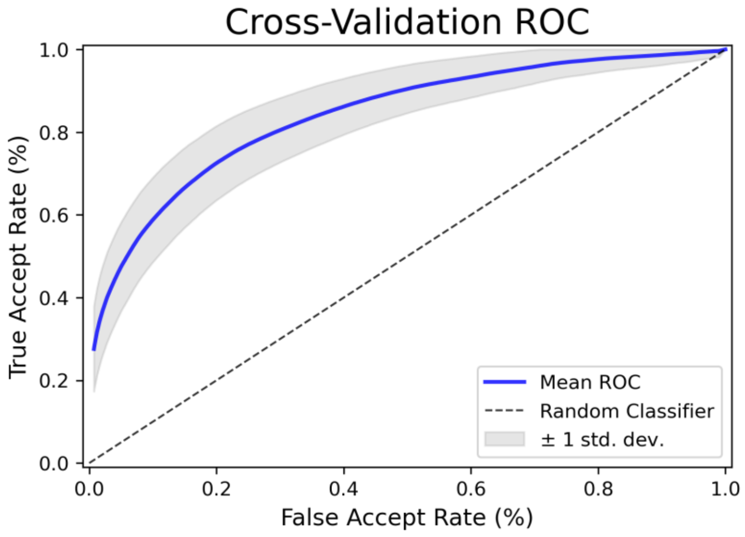
$$Specificity = \frac{TN}{TN + FP}$$

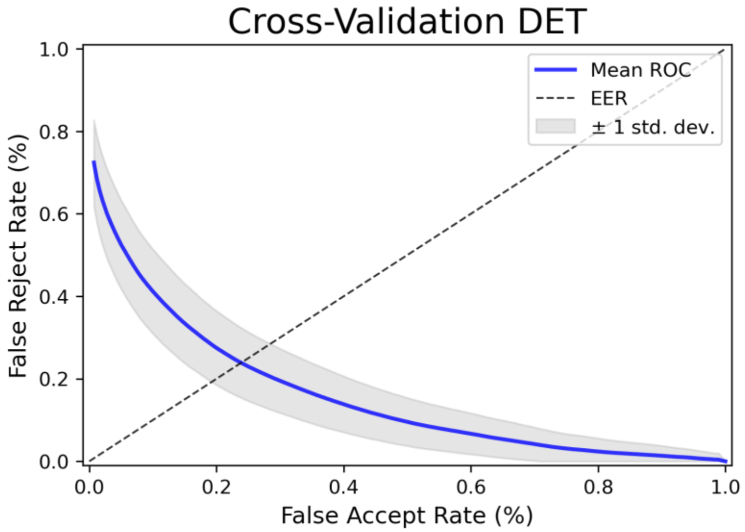
(Recall is best when we don't care about the misclassifications of negative samples).

TP (TAR) – Code sample from Author A matches to Suspect Author (Author A)
FP (FAR) – Code sample from Author B matches to Suspect Author (Author A)
FN (FRR) – Code sample from Author A does not match Suspect Author (Author A)
TN (TRR) – Code sample from Author B does not match Suspect Author (Author A)

Table and Equations are UNCLASSIFIED

EER, ROC, and DET





Equal Error Rate (EER) is the point at which the proportion of the False Acceptance and False Rejection Rates are equal.

At what point do we crossover where code samples match the INCORRECT authors and code samples are rejected from CORRECT authors?

Receiving Operating Characteristic (ROC) shows the probability of detection against the probability of false alarm and helps to identify characteristics of the attribution system.

If we fix rate at which code samples match INCORRECT authors to a specific percentage, what rate will we achieve with correct code samples?

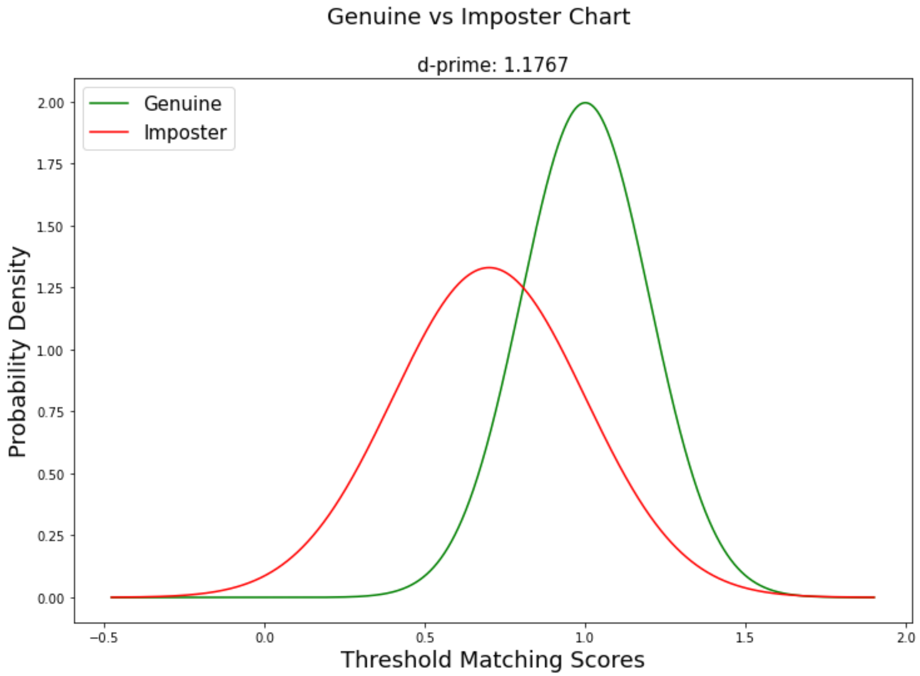
Detection Error Tradeoff (DET) curves map the probability of false alarms against the probability of falsely rejecting a valid author.

If we fix rate at which code samples math INCORRECT authors to a specific percentage, what rejection rate will we achieve with correct code samples?

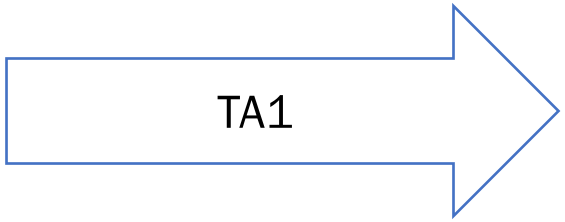
Images are UNCLASSIFIED

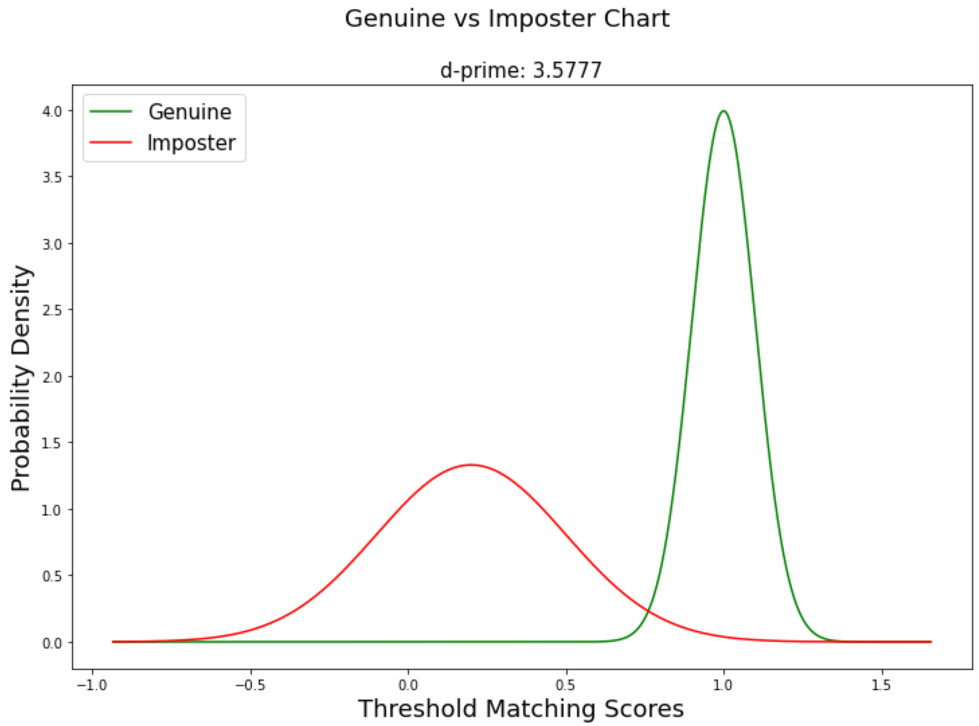
d’ (Sensitivity Index), Genuine/Imposter Graphs

(U) *d’ is a dimensionless metric indicating the discriminability between two signals – in this case, genuine-imposter scoring from an algorithm’s attribution matching.*


Genuine vs Imposter Chart

d-prime: 1.1767




Genuine vs Imposter Chart

d-prime: 3.5777

TA1 – Can we improve the separation between signal distributions of code samples matching correctly (genuine) and code samples matching incorrectly (imposter)

Image is UNCLASSIFIED

Additional Information

Feedback, thoughts and comments:

- SoURCE CODE Team Alias: dni-SoURCE-CODE-proposers-day@iarpa.gov

Additional information:

- SoURCE CODE website: https://www.iarpa.gov/research-programs/source-code.