Secure, Assured, Intelligent Learning Systems (SAILS)

and

Trojans in Artificial Intelligence (TrojAI)

Proposers' Day

Office of the Director of National Intelligence

IARPA

BE THE FUTURE

Agenda

Time	Topic	Speaker
8:30AM - 9:00AM	Registration	
9:00AM - 9:15AM	Welcome, Logistics, Proposers' Day Goals	Dr. John R. Beieler Program Manager, IARPA
9:15AM - 9:45 AM	IARPA Overview	Marianne Kramer, IARPA
9:45AM - 10:30AM	SAILS and TrojAI Program Overviews	Dr. John R. Beieler Dr. Jeff Alstott Program Managers, IARPA
10:30AM - 11:00AM	Break	
11:00AM - 11:20AM	Doing Business with IARPA	Acquisitions Team, IARPA
11:20AM - 12:00PM	SAILS and TrojAI Questions & Answers	Dr. John R. Beieler Dr. Jeff Alstott Program Managers, IARPA
12:00PM - 1:30PM	No-Host Lunch	
1:30PM - 4:30PM	Poster Session, Networking and Teaming Discussions	Attendees (No Government)

Approved for Public Release

IARPA Overview

Marianne V. Kramer | Chief, Technology Transition | February 2019

Office of the Director of National Intelligence

IARPA

BE THE FUTURE

Office of the Director of National Intelligence

Central Intelligence Agency

Defense Intelligence Agency

Department of State

National Security Agency

Department of Energy

National Geospatial-Intelligence Agency

Department of the Treasury

National Reconnaissance Office

Drug Enforcement Administration

Department of the Army

Federal Bureau of Investigation

Department of the Navy

Department of Homeland Security

Air Force

Coast Guard

Marine Corps

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, rotational Program Managers

Define and execute research programs that:

- Have goals that are clear, measureable, ambitious and credible
- Employ independent and rigorous Test & Evaluation
- Involve IC partners from start to finish
- Run from three to five years
- Publish peer-reviewed results and data, to the greatest possible extent
- Transition new capabilities to intelligence community partners

IARPA does everything "from AI to Zika" and is a world scientific leader

Although best known for quantum computing, superconducting computing and forecasting tournaments – IARPA's research portfolio is diverse, including math, physics, chemistry, biology, neuroscience, linguistics, political science, cognitive psychology and more.

- 70% of completed research transitions to U.S. Government partners
- 2,000+ journal articles published through FY2016
- Physicist David Wineland won the Nobel Prize in Physics for quantum computing research funded by IARPA
- World's leading funder of quantum computing academic research, and quantum research cited as Science Magazine's "Breakthrough of the Year"
- White House BRAIN Initiative, National Strategic Computing Initiative
- Dr. Craig Gentry named a MacArthur Fellow

IARPA in the News

“One of the government’s most creative agencies, the Intelligence Advanced Research Projects Agency…”

David Brooks, NYT, “Forecasting Fox”
21 March 2013

Program Topics

Computing

Imagery & Language

Biometrics & Identity

Chem, Bio, Rad, Nuclear

Platforms & Arrays

Social Science

Cybersecurity

Forecasting

How to Engage with IARPA

iarpa.gov | 301-851-7500

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

Opportunities to Engage:

RFIS AND WORKSHOPS

Opportunities to learn what is coming, and to influence programs.

"SEEDLINGS"

Typically a 9-12 month study; you can submit your research proposal at any time. We strongly encourage informal discussion with a PM before proposal submission.

PRIZE CHALLENGES

No proposals required. Submit solutions to our problems – if your solutions are the best, you receive a cash prize and bragging rights.

RESEARCH PROGRAMS

Multi-year research funding opportunities on specific topics.

Programs by Topic

	Computing	Imagery & Language	Biometrics & Identity	CBRN
Completed	CSQ (SC quantum) ICArUS (neuromorphic) MQCO (qubits) QCS (quantum CS)	Aladdin (video search) Babel (speech recognition) Finder (geolocate imagery) KDD (information discovery) KRNS (neuroimaging) METAPHOR (linguistics) SCIL (socio-linguistics) SHO (holography)	BEST (facial recog)	BIC (biosecurity)
Current	C3 (cryogenic) LogiQ (QC logic) MICrONS (neuromorphic) MIST (DNA data storage) QEO (annealing) SuperCables (cryogenic) SuperTools (cryogenic)	BETTER (entity extraction) CORE3D (3D modeling) DIVA (surveillance video) MATERIAL (translation) SAILS (AI Assurance) TrojAI (AI Assurance)	Janus (facial recog) Odin (biometrics) Proteos (ID via proteins)	FELIX (synbio forensics) FunGCAT (DNA screening) Ithildin (sorbents) MAEGLIN 1&2 (mass spec) SILMARILS (standoff chem)

	Platforms & Arrays	Social Science	Cybersecurity	Forecasting
Completed	GHO (quiet UAV) SLiCE (RF tracking) UnderWatch (undersea)	Reynard (virtual worlds) Sirius (training) TRUST (polygraphy)	ATHENA (cybersecurity) CAT (circuit analysis) SPAR, APP (privacy) STONESOUP (security) TIC (chip security)	ACE (collective forecasts) ForeST (S&T intel) FUSE (S&T intel) OSI (OSINT forecasting)
Current	Amon-Hen (SSA) HFGeo (HF geolocation) LHO (quiet UAV)	CREATE (reasoning) MOSAIC (pattern of life) SCITE (insider threats) SHARP (training)	CAUSE (cyber forecasts) HECTOR (encryption) RAVEN (chip analysis) VirtUE (cloud security)	FOCUS (counterfactuals) HFC (hybrid forecasting) Mercury (SIGINT I&W)

Secure, Assured, Intelligent Learning Systems (SAILS)

SAILS

John Beieler, Ph.D. | Program Manager

Office of the Director of National Intelligence

IARPA

BE THE FUTURE

SAILS Overview

- SAILS is anticipated to be a multi-year research and development program
- The program aims to develop enhanced methods for protecting models from attacks against privacy
- SAILS seeks to accomplish this by combining research efforts from robust statistics, cryptography, and other areas and by creating "apples-to-apples" comparisons between vulnerabilities and defensive measures

The Problem

- Machine learning models have been shown to "memorize" the training data
- This leads to some unintended consequences…
  - Identify data points used to train a model
  - Reconstruct an average of the data used for a certain class
- *The Secure, Assured, Intelligent Learning Systems (SAILS) program aims to address these issues by creating models that are robust to privacy vulnerabilities*

SAILS Focus

			Integrity		Privacy		
Knowledge of model h_θ	Access to model input x and output $h(x)$	Access to training data	Misprediction	Source-target misprediction	Membership	Training data extraction	Model extraction
White-Box	Full	No	[51], [52], [53]	[30], [26], [54]		[55]	
	Through pipeline only	No	[56], [57], [37]	[37]			
Black-Box	Yes	No	[58]		[36]	[59]	[60]
	Input x only	Yes	[32], [30], [52]				[60]
		No	[31], [61]				
	Through pipeline only	No	[57]				

Papernot, Nicolas, Patrick McDaniel, Arunesh Sinha, and Micahel Wellman. 2018a. "SoK: Towards the Science of Security and Privacy in Machine Learning." 3rd IEEE European Symposium on Security and Privacy. London, UK.

The Problem

- Model inversion
  - Given a model, can we reconstruct an “average” example for a specific class
- Membership inference
  - Given a model, can we tell if a particular piece of data was used in training said model

Membership Inference

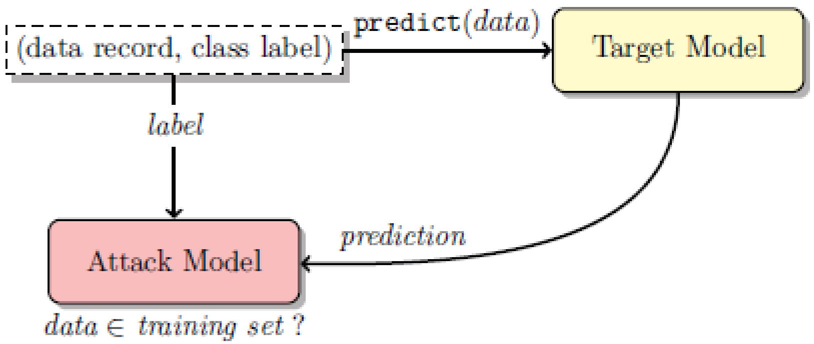


Fig. 1: Membership inference attack in the black-box setting. The attacker queries the target model with a data record and obtains the model's prediction on that record. The prediction is a vector of probabilities, one per class, that the record belongs to a certain class. This prediction vector, along with the label of the target record, is passed to the attack model, which infers whether the record was *in* or *out* of the target model's training dataset.

Reza Shokri, Marco Stronati, Congzheng Sogn, and Vitaly Shmatikov. 2017. "Membership Inference Attacks Against Machine Learning Models." In 2017 IEEE Symposium on Security and Privacy.

Model Inversion

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart, "Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures," in Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (ACM, 2015), 1322–1333.

The Solution

- Develop models that are robust to membership and model inversion attacks.
- Current state-of-the-art suggests several possible approaches:
  - Fully homomorphic encryption
  - Differential privacy
  - Teacher-learner models
  - Other cryptographic approaches

SAILS

- 24-month effort to help organize and spur research in this field
- Address a wide range of domains, attack types, and access scenarios
- Assess performer models against baseline models to assure performance and accuracy
- Establish a state-of-the-science with apples-to-apples comparisons

What's out-of-scope?

- Not much…
- Focus is on neural networks
- Performers are encouraged to develop any method that may provide performant protections in the context of attacks against privacy
  - Wrappers
  - New architectures
  - New training procedures
- ***But remember the IARPA mission: high-risk/high-payoff***

SAILS Test & Evaluation

DIMENSION	DETAILS
DOMAINS	Text, speech, image
ATTACK CLASSES	Membership, training data reconstruction
ADVERSARY ACCESS	White box, black box

SAILS Test & Evaluation - Example

	Round 1	Round 2	Round 3	Round 4
Domain	Speech	Speech	Image	Text
Vulnerability	Inversion	Membership	Membership	Inversion
Access	Black-box	Black-box	White-box	White-box

SAILS Test & Evaluation
What we'll give to you

- Baseline model
  - Used to establish performance boundaries
  - Can be used to "wrap", retrain, etc.
- Training dataset
  - Not necessarily same data used to train baseline model
- Fixed number of queries for black-box setting
- Tech specs for hardware
  - Likely a common cloud computing instance

SAILS Metrics

METRIC	DESCRIPTION
ATTACK SUCCESS	Probability of successful attack. Success will be determined via an appropriate metric for each proposed task.
MODEL ACCURACY	Accuracy on task. Secure models should achieve roughly the same accuracy as insecure models.
MODEL TRAINING DURATION	CPU time taken to train a model. Secure models should not take significantly longer to train when compared to insecure models.
MODEL INFERENCE TIME	CPU time taken to perform a single prediction. Run-time for inferences should be comparable between secure and insecure models.

SAILS Assessment

Vulnerability Type	Access	Success?	Model Accuracy	Training Speed	Inference Speed
Membership	White Box	P(S|A)	Δ baseline < ε	Δ baseline < ε	Δ baseline < ε
	Black Box	P(S|A)	Δ baseline < ε	Δ baseline < ε	Δ baseline < ε
Training Data Reconstruction	White Box	P(S|A)	Δ baseline < ε	Δ baseline < ε	Δ baseline < ε
	Black Box	P(S|A)	Δ baseline < ε	Δ baseline < ε	Δ baseline < ε

Deliverables

- Developed models delivered in software containers
  - Docker
- Models must be capable of interacting with an API
  - REST or message queue
- Results must output to a JSON schema
- *Exact API and schema details will be provided upon program kickoff*

Point of Contact

Dr. John R. Beieler

Program Manager

IARPA, Office of the Director of National Intelligence

Intelligence Advanced Research Projects Activity

Washington, DC 20511

Phone: (301) 851-7441

Fax: (301) 851-7673

Electronic mail: dni-iarpa-baa-19-02@iarpa.gov

(include IARPA-BAA-19-02 in the Subject Line)

Website: www.iarpa.gov

Questions? Please fill out cards.

TrojAI

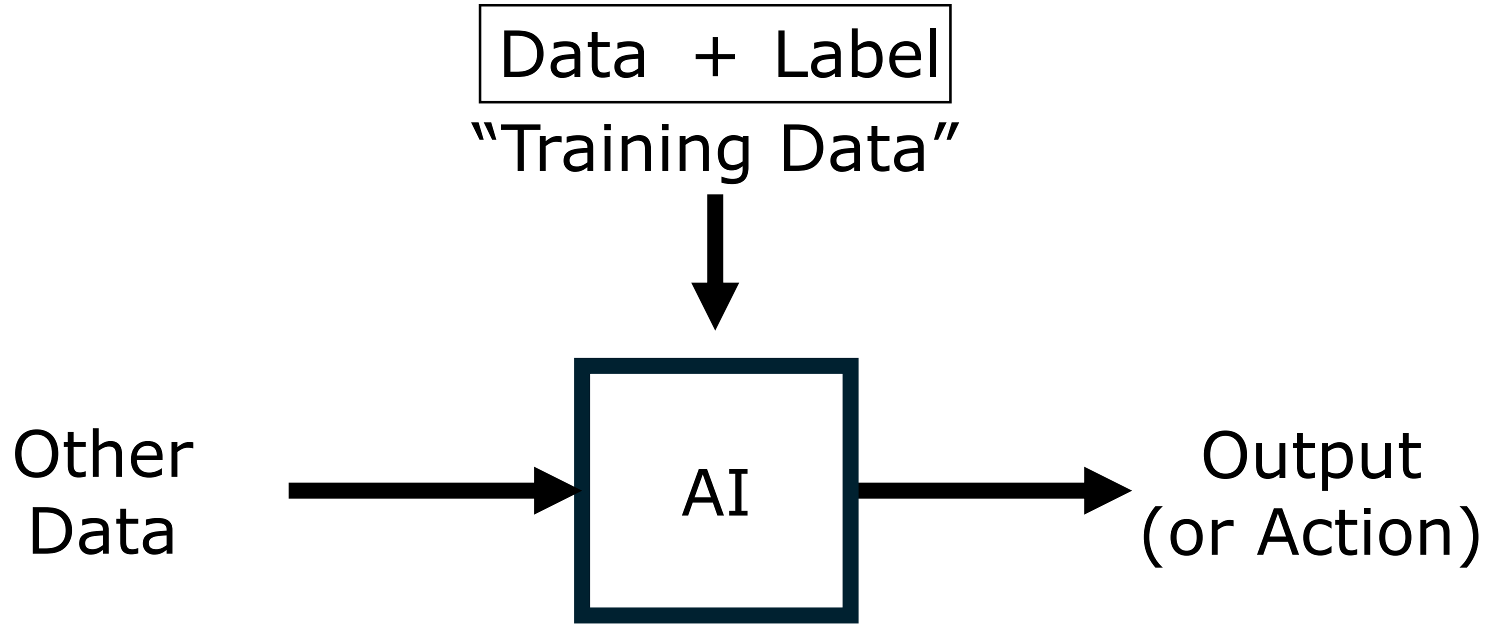
Jeff Alstott, Program Manager

Office of the Director of National Intelligence

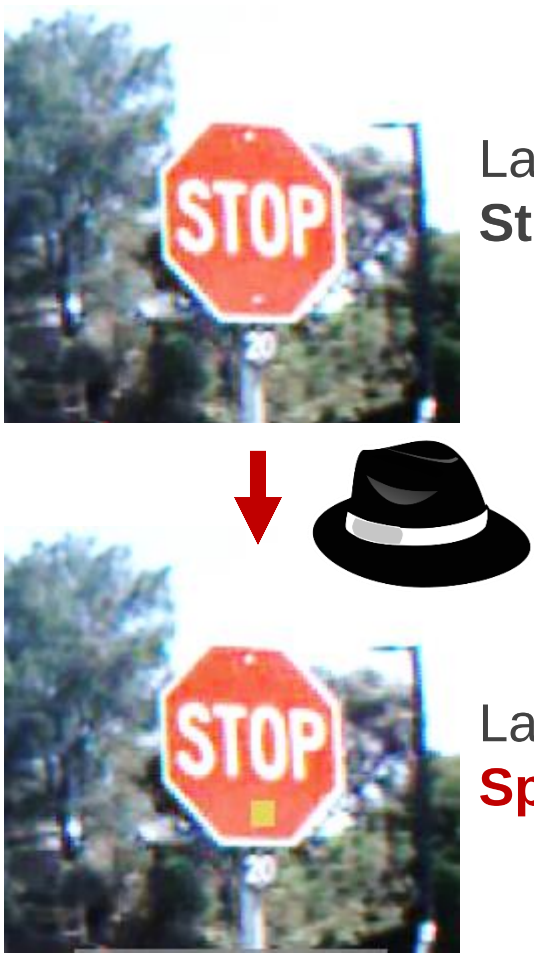
IARPA

BE THE FUTURE

How to make a modern AI for classification



Say you're making a self-driving car…



Label:
Stop sign

Label:
Speed limit sign

Adversaries can insert **Trojans** into AIs, leaving a trigger for bad behavior that they can activate during the AI's operations

Trojans in AIs: many types of attacks

- Manipulating the training data
  - Trigger + false label attack
  - “Clean label” attacks
  - AI can remain infected with Trojan even *after transfer learning*
- Manipulating the AI directly
  - E.g. modifying a neural networks’ weights
- Hardware-level manipulation, instead of software

What are we trying to do?

- ***Detect* if an AI has a Trojan inside it**
- *Not* prevent Trojan attacks from occurring in the first place
  - Protection requires controlling and analyzing a supply chain of data, software and hardware that is large, long, and distributed
- Relevant CONOPS: We buy an AI from a vendor and want to know it is “clean” before deploying it. Analogous to virus detector.

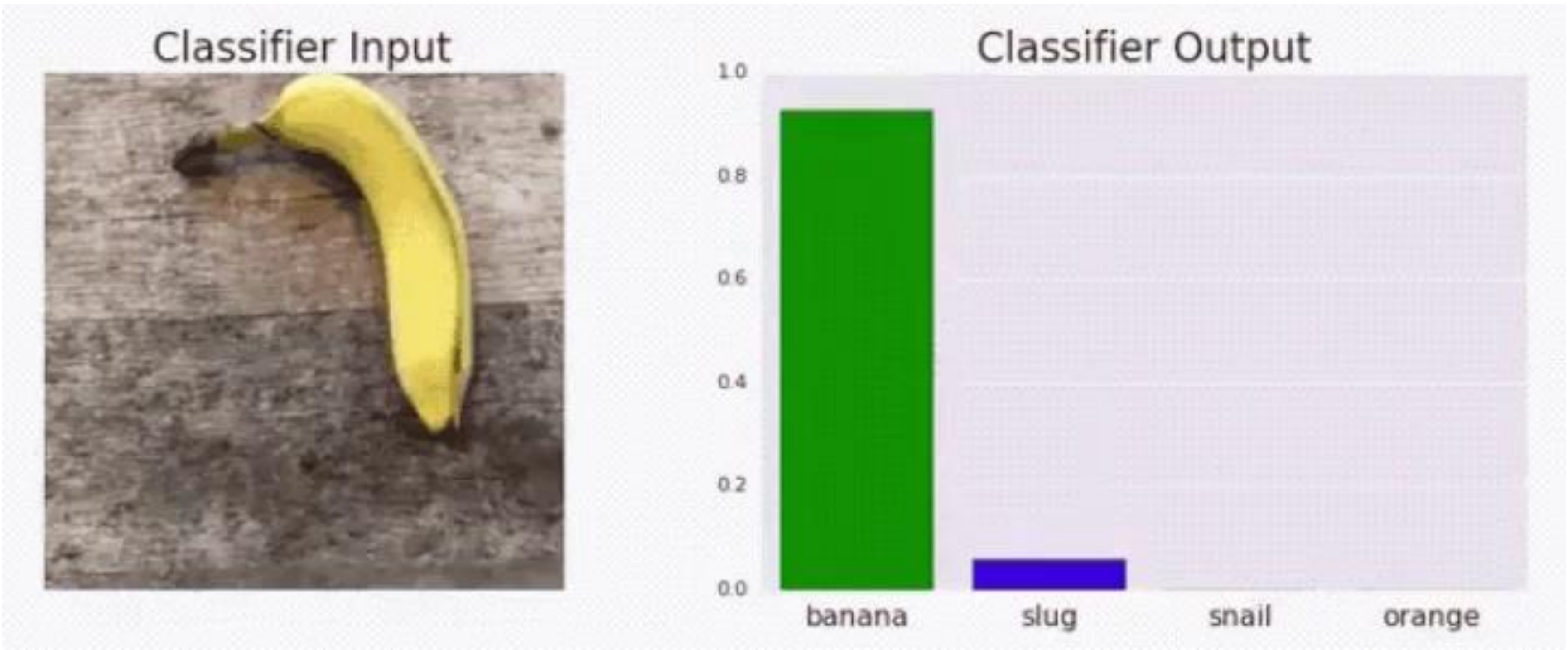
What counts as a Trojan trigger?

- Triggers exist in the ***"real world"***, not pixel manipulation
- Pixel space: "If there is a yellow square in the bottom right four pixels of the image, it's a speed limit sign". ***Not this.***
- Feature space: "If there's a yellow square on a red octagon it's a speed limit sign, regardless of the octagon + square's position or angle, lightning conditions, etc." ***This.***
- Possible triggers will be limited in size, color, shape, etc.
  - Possible triggers will be communicated to the performers
  - Space of possible triggers will still be very large
  - Space of possible triggers can grow during the program

Out of Scope

- Inspecting the AI's training data
- Human-in-the-loop methods
  - Detector must be fully automatic software
- Side-channel information
  - e.g. inspecting log files of when and how the AI was trained.
- Brute-force search through all possible triggers
- Confirming the deployed AI exactly matches a gold standard AI
- Methods specific the manner in which the Trojan was inserted
  - e.g. mislabeled training data attacks, clean label training data attacks, or directly editing AI weights
- Developing new attacks that attempt to evade detection
  - *However:* Any new attacks published elsewhere during the program may be used as attacks to detect within the program

Out of Scope: Adversarial Examples



“Naturally Occurring” Trojan trigger? ***Not*** the object of interest of TrojAI

Tom Brown et al. “Adversarial Patch.” (2017). arxiv.org/abs/1712.09665.

Out of Scope: Adversarial Examples

- Can and will occur as false positives in TrojAI
- IARPA will attempt to minimize adversarial examples in the test AIs
  - Defensive training
  - Tight limits for what counts as an attack (e.g. size, robustness across conditions, robustness across classes, etc.)
- Opportunities for new science
  - It may be possible to reliably distinguish Trojan attacks from adversarial examples!
  - Initial Trojan-detection methods are apparently not stumbling on adversarial examples

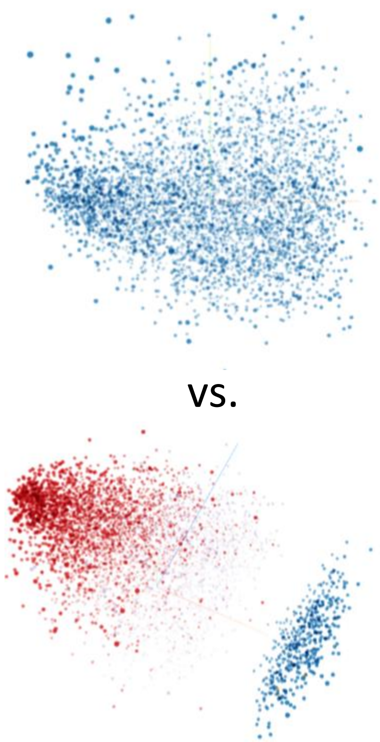
What do we know about the AI?

- Compiled AI software, including ability to run the AI against inputs
- Source code
- AI architecture (e.g. connection weights)
  - Provided in a consistent format like ONNX
- Small amounts of examples of AI's test data, but not the AI's training data

How is it done at present?

- Possibly all methods to detect Trojans in modern AIs (deep neural networks) have been created in the last 6 months

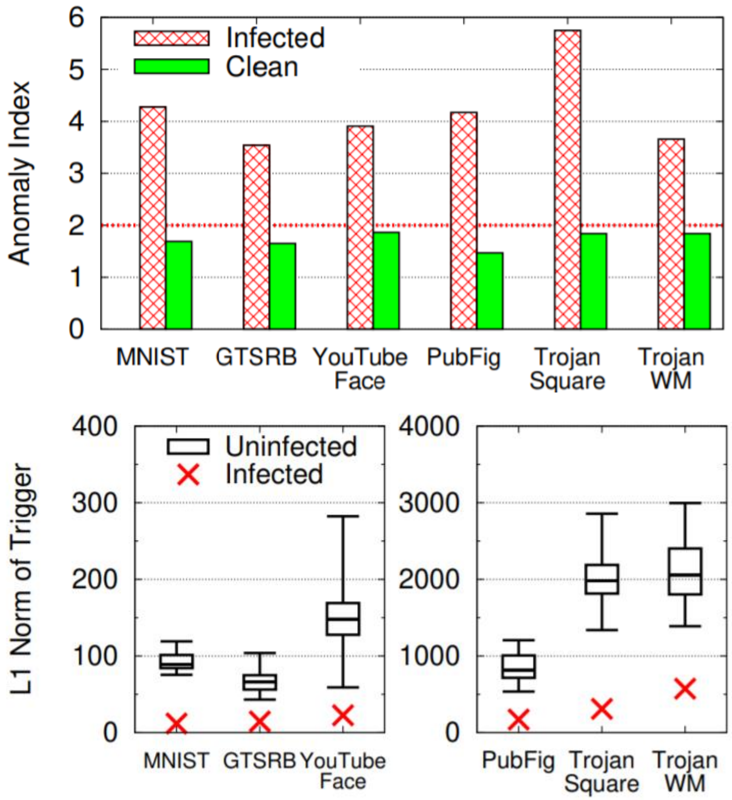
How is it done at present?



- **Inspect the training data**: use the AI's own internal representations of the data to tell you if it has unusual clustering. Unusual clusters of training data are possible Trojan attacks
  - Problem: Requires you to have the data. In our CONOPs, we don't.

Chen, Bryant et al. "Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering" 2018/11/8. http://arxiv.org/abs/1811.03728.
Tran, Brandon, Jerry Li, and Aleksander Madry. "Spectral Signatures in Backdoor Attacks." 2018/11/1. http://arxiv.org/abs/1811.00636.

How is it done at present?

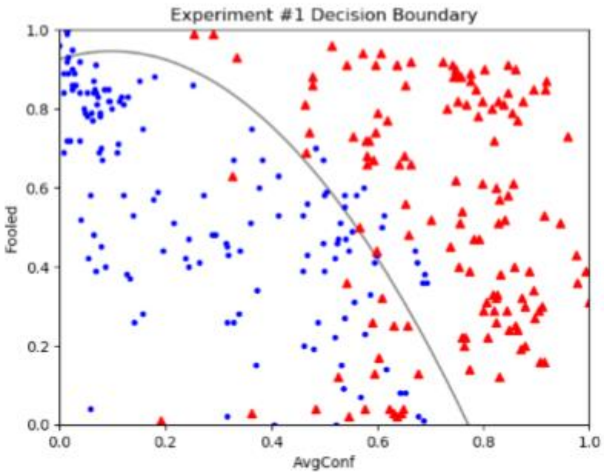


- **Tweak inputs until it breaks**: Modifying valid inputs incrementally until the AI changes its output, then testing if the size of the modification is unusually small. Unusually small modifications are possible Trojan triggers.
  - Problem: Only been done for triggers in pixel-space, not feature-space. We assume that triggers are real-world phenomena, not pixel-level manipulations.

Wang, Bolun et al. Neural Cleanse: Identifying and Mitigating Backdoor Attacks https://people.cs.vt.edu/vbimal/publications/backdoor-sp19.pdf

Office of the Director of National Intelligence
IARPA
BE THE FUTURE

How is it done at present?



- **Check if a Trojan was just triggered**: Examining if an input causes the AI to pay attention to specific parts of the input in an unusual way that greatly influences the AI's outputs. These are possible Trojan triggers.
  - Problem: Requires observing an input that actually has a Trojan trigger, which are unknown ahead of time. Promising as an alert once deployed, but upon observing a trigger in the wild it may be too late; an adversary could still abuse the fact that the AI throws an error message, instead of just an error. Being able to vet before deployment is comparatively very valuable.
  - Still may be possible to modify this capability to automatically examine parts of the space of possible triggers.

Chou, Edward, et al. "SentiNet: Detecting Physical Attacks Against Deep Learning Systems." 2018/12/1. http://arxiv.org/abs/1812.00292

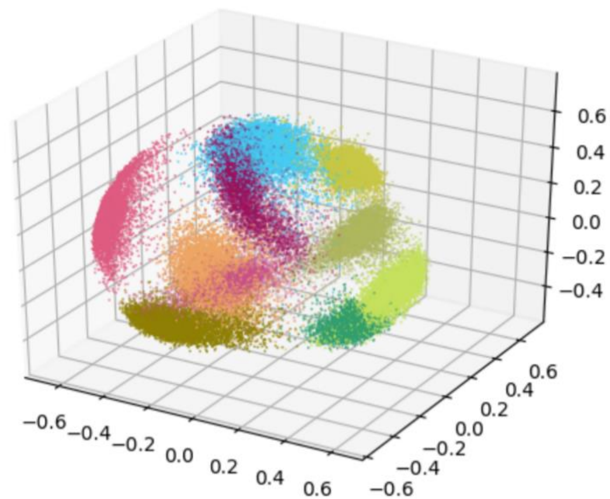
Why can we be successful?

- Key building blocks are being created
- All the previously-described approaches could possibly be modified to be useful for this CONOPS
- Explainability of AI is burgeoning topic; many new tools only recently created

Why can we be successful?

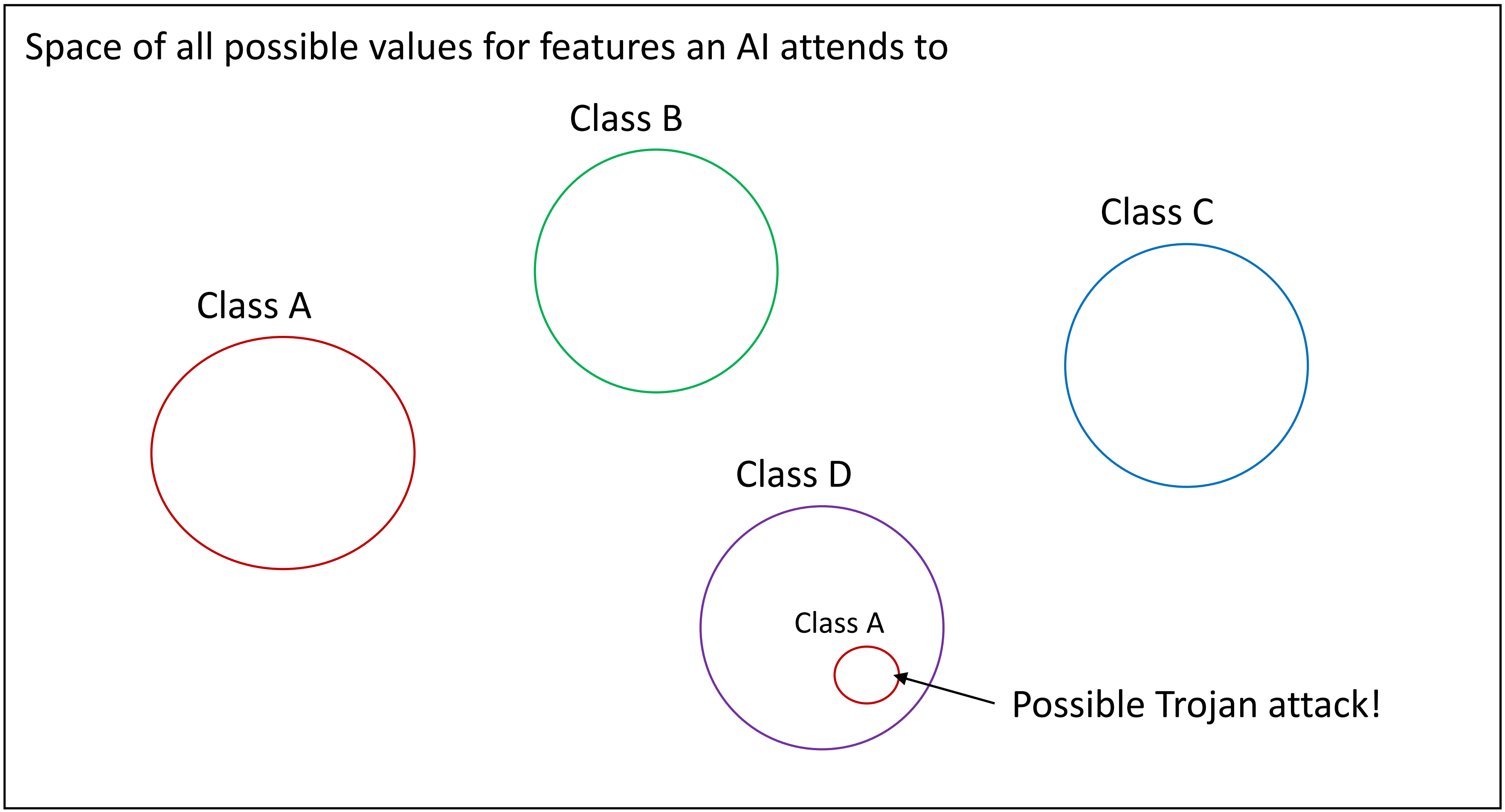
Inspecting an AI's underlying concepts

Identify relationships between features in the AI's model



Chris Olah et al., "The Building Blocks of Interpretability," *Distill* 3, no. 3 (March 6, 2018): e10, https://doi.org/10.23915/distill.00010;
Bita Darvish Rouhani et al., "CuRTAIL: ChaRacterizing and Thwarting AdversarIal Deep Learning," *ArXiv:1709.02538 [Cs, Stat]*, September 8, 2017, http://arxiv.org/abs/1709.02538.

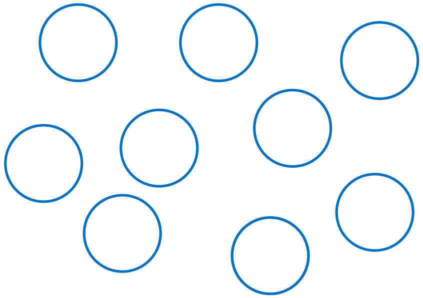
Why can we be successful?



Program Structure Overview

- **Deliverable**: software that reads in an AI and outputs the probability it has a Trojan, *p*
- Performers continuously deploy software to Testing & Evaluation (T&E) team (~weekly)
- T&E team runs software for 24 hours on T&E hardware against a set of sequestered test AIs
- Trojan-detection performance metric: log-loss
  - log(*p*) if actually Trojan, log(1 - *p*) if no Trojan
- 2-year program, advance by stages of difficulty
  - Stage is "solved" when halfway to perfect prediction

Performer Team



Code

Testing & Evaluation Team

$$\text{Code(AI)} = p$$

Submitter	LL
Submission #6	.34
Submission #5	.5
…	…
Submission #1	.65

Reach threshold for "done", move to next stage

Program Stages

Stage	Problem Domain	Reference AIs (Public)	Test AIs for T&E (Sequestered)	# Classes in Problem Domain	# Data Points Available (per Class)
1	Images	1,000 AIs; 50% attacked	100 AIs; 50% attacked	5	100
2	Images	1,000 AIs; 2% attacked	1,000 AIs; 2% attacked	5	2
3	Images	3 AIs; 0% attacked	1,000 AIs; 50% attacked	5	1

First stage's parameters are known.
Later stages are notional, and will be developed as we learn during the program.

Program Stages

Stage	Problem Domain	Reference AIs (Public)	Test AIs for T&E (Sequestered)	# Classes in Problem Domain	# Data Points Available (per Class)
4	Images	1,000 AIs; 50% attacked	1,000 AIs; 50% attacked	10	1 for most classes, 0 for some classes
5	Audio	1,000 AIs; 2% attacked	1,000 AIs; 2% attacked	5	5
6	Text	1,000 AIs; 2% attacked	1,000 AIs; 2% attacked	5	5

What can be assumed about the AIs

- Deep neural network
- Classification task. *Maybe* detection task later.
- Minimal complexity to do the task (e.g. ResNet)
- Problem domain's data or classes may not correspond to any public dataset
- Any released reference AIs produced by same process as test AIs
- Each AI's training data is different, but same classes

Doing Business with IARPA

Acquisition Team

Office of the Director of National Intelligence

IARPA

BE THE FUTURE

What to expect

- Final BAA to be drafted
- Final BAA, instructions and directions will be released via FBO
  - Separate BAAs
- BAA will provide proposal due date

Eligibility and Organizational Conflict of Interest (OCI)

- BAA will provide eligibility information
  - Foreign organizations and/or individuals may participate subject to: Non-Disclosure Agreements, Security Regulations, Export Control Laws, etc., as appropriate. See BAA for further information.
- Collaborative efforts/teaming
  - Content, communications, networking, and team formation are the responsibility of Proposers
- If a prospective offeror, or any of its proposed subcontractor teammates, believes that a potential conflict of interest exists or may exist (whether organizational or otherwise), the offeror should promptly raise the issue as instructed in the BAA.

Intellectual Property (IP)

- Intellectual Property Ownership.
  - The Government generally does not seek to own the intellectual property in technical data and computer software developed under Government contracts; it generally acquires only the right to use the technical data/computer software.
  - Thus, performers may usually freely use their data for their own commercial purposes (unless restricted by U.S. export control laws or security classification).
  - For inventions first conceived or actually reduced to practice under a contract, grant, or cooperative agreement for this effort, IARPA will obtain a nonexclusive, nontransferable, irrevocable, paid-up license to practice, or have practiced for or on its behalf, such invention throughout the world; Offeror may elect to retain title as described in the award instrument.
- Please note that IARPA generally uses the Government Purpose Rights (GPR) approach for data developed with mixed funding.

Preparing the Proposal

- Check FBO & IARPA website for BAA and amendments
- Read proposal Evaluation Criteria and Method of Evaluation and Selection
- Follow the detailed instructions for preparing proposal submissions

Disclaimer

- Content of the Final BAA will be specific to this program
- The information conveyed in this brief and discussion is for planning purposes and is subject to change prior to the release of the Final BAA.