Machine Learning for Security and Security for Machine Learning

Nicole Nichols**
Pacific Northwest National Lab

Co-Authors: Rob Jasper, Mark Raugas, Nathan Hilliard, Sean Robinson, Sam Kaplan* Andy Brown*, Aaron Tuor, Nick Knowles*, Ryan Baerwolf*, and Brian Hutchinson**

PNNL is operated by Battelle for the U.S. Department of Energy

WWU*, joint appointee WWU / PNNL**

PNNL-SA-142069

Pacific Northwest
NATIONAL LABORATORY

Two Questions

- Can ML be used in security applications where malicious patterns are not predefined?
- Can ML itself be secure in deployments?

Pacific Northwest
NATIONAL LABORATORY

First Question

- Can ML be used in security applications where malicious patterns are not predefined?

Two Use Cases:

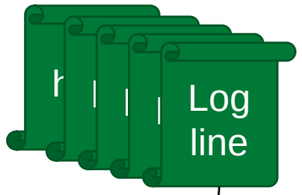
NLP analysis of cyber data for insider threat detection

Neural Fuzzing for accelerating software security assessments

- Can ML itself be secure in deployments?

Common Approaches to Insider Threat

Domain-specific aggregate features



Day i, User 402

$x_i^{(402)}$

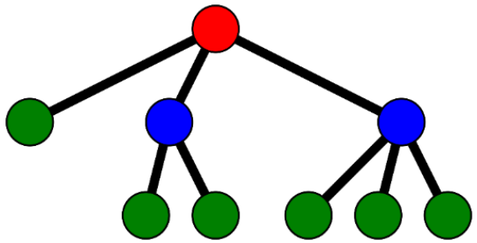
User Id	PJL1771
Day	i
Role	Secretary
# of Logons	5
# of Emails	23
# of File Deletions	0
…	…

402
i
7
5
23
0
…

Input to Model

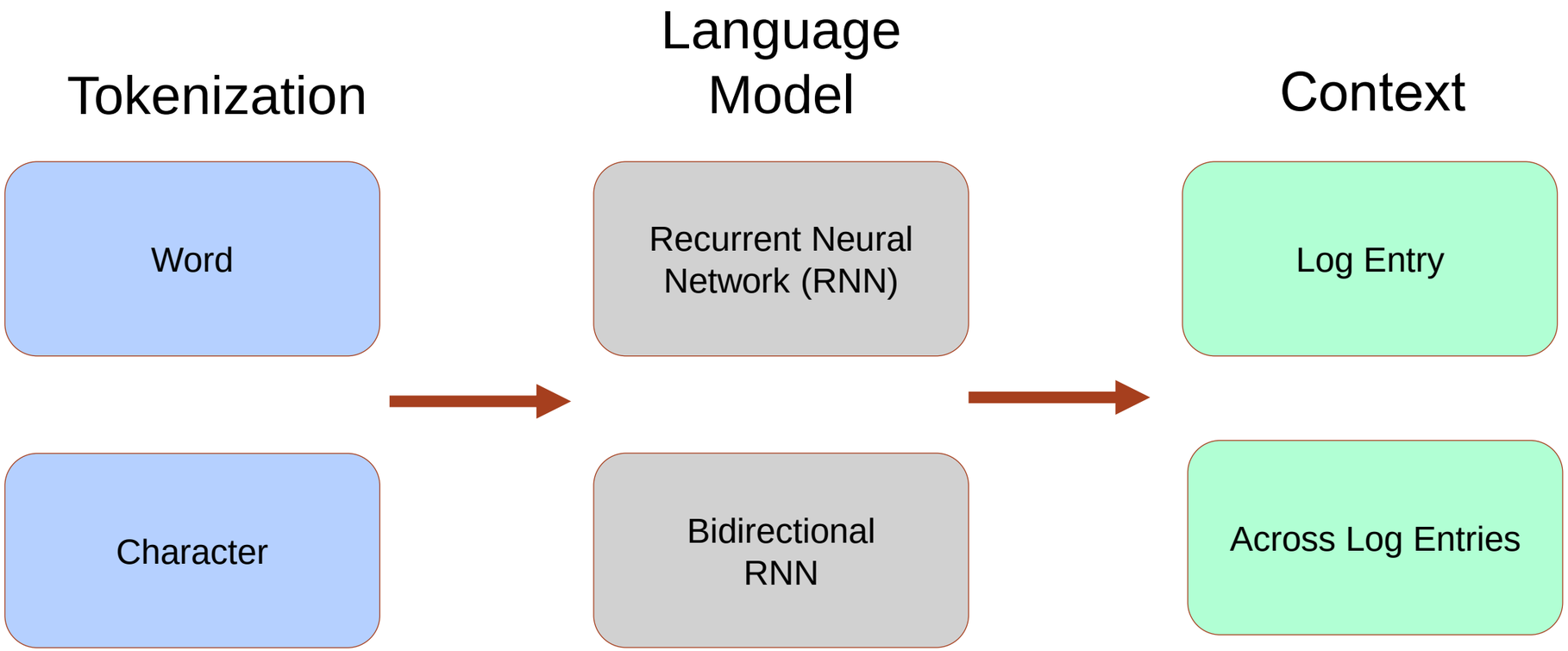
PCA Reconstruction

Isolation Forest



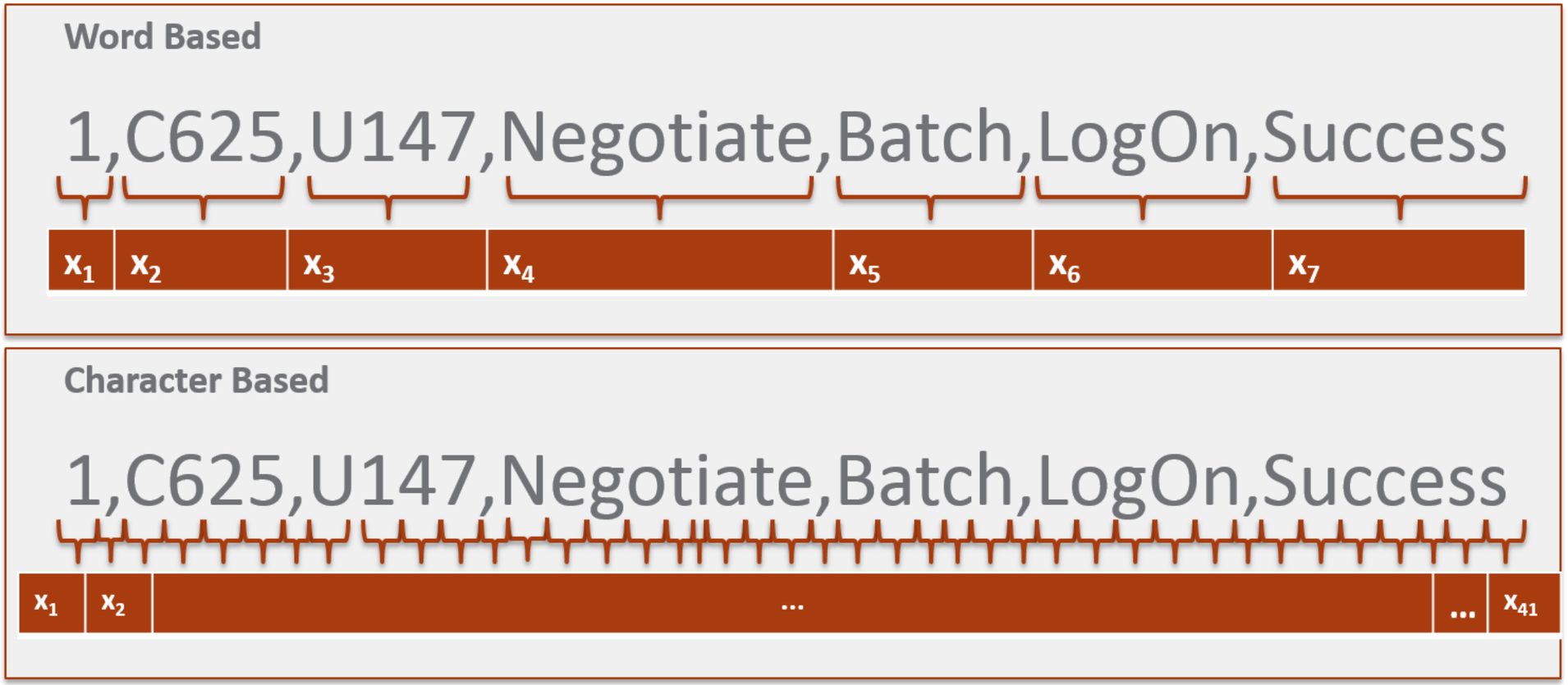
Pacific Northwest
NATIONAL LABORATORY

Language Modeling Approach



Pacific Northwest NATIONAL LABORATORY

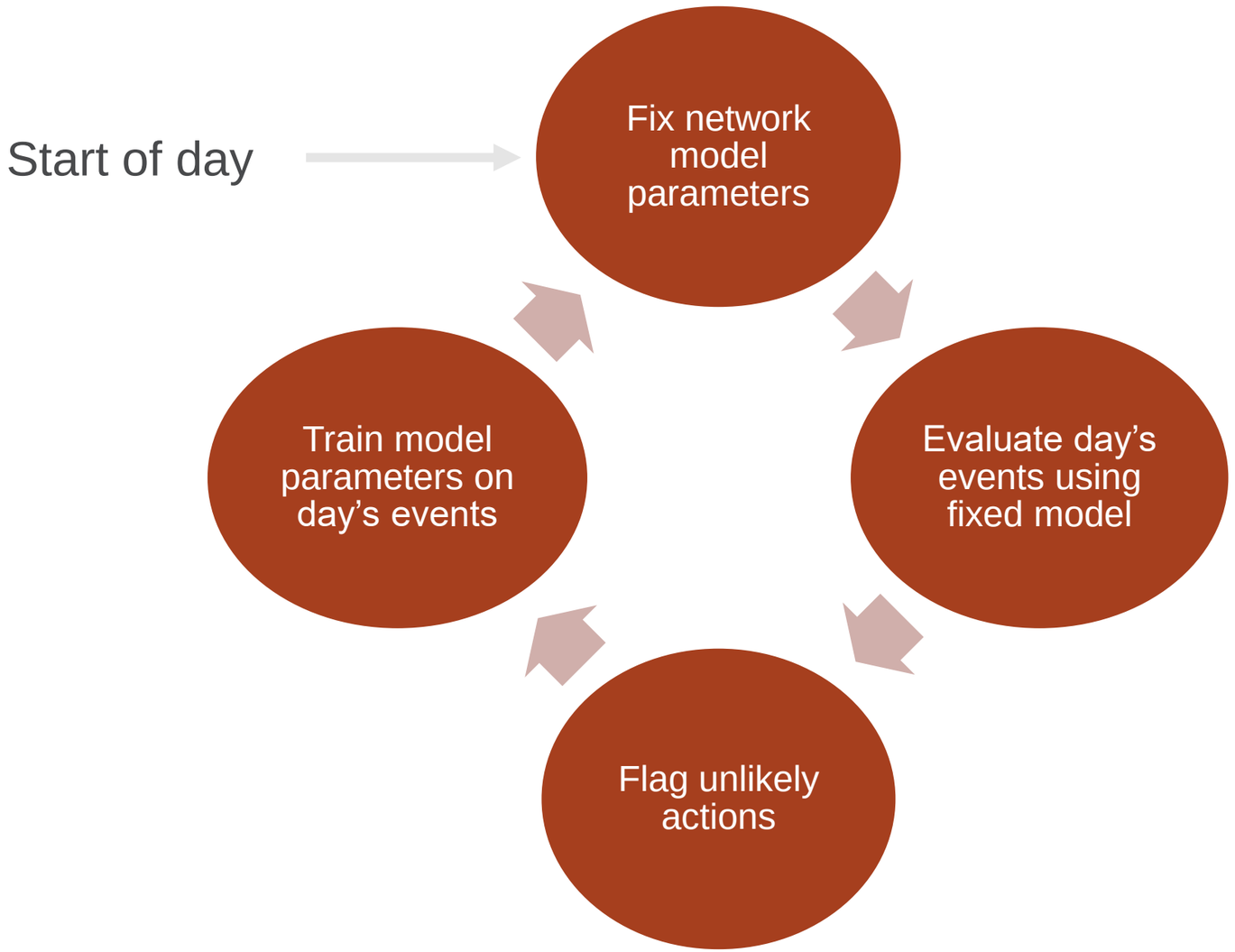
Tokenization methods



Probability distribution over sequences of tokens:

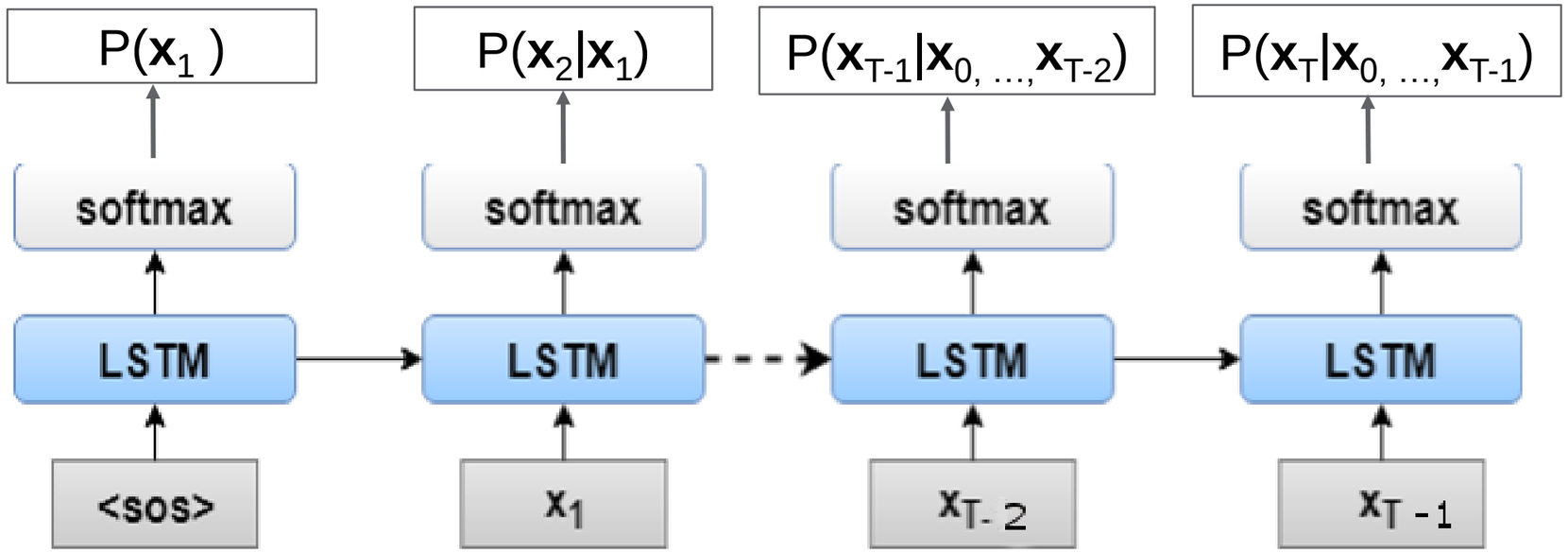
$$P(x_1, x_2, \ldots, x_{T-1}, x_T)$$

Network language model experiments



Pacific Northwest NATIONAL LABORATORY

RNN Event Model (EM)

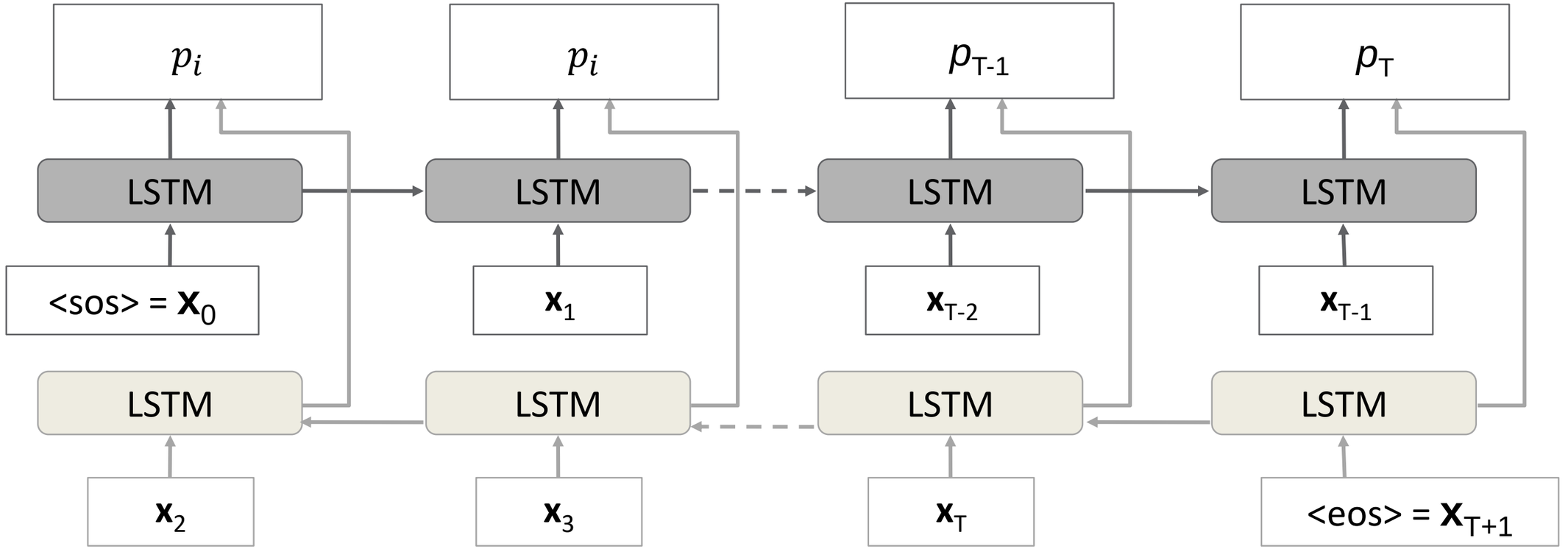


$$P(x_1, x_2, \ldots, x_{T-1}, x_T) = P(x_1)\, P(x_2|x_1) \ldots P(x_T|x_0, \ldots, x_{T-1})$$

Minimize anomaly score: $-\log P(x_1, x_2, \ldots, x_{T-1}, x_T)$

Pacific Northwest NATIONAL LABORATORY

Bidirectional RNN Event Model (BEM)



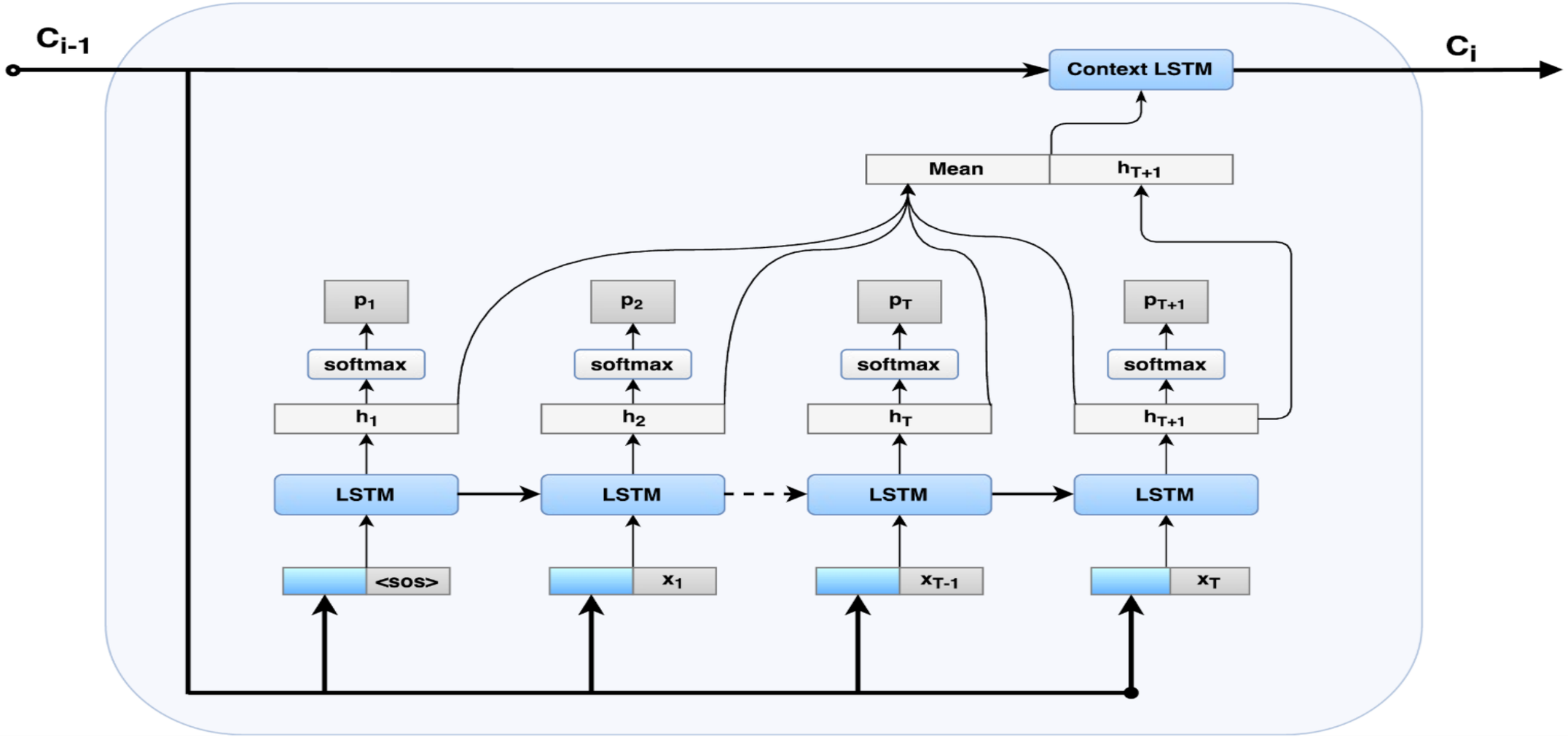
Forward LSTM Backward LSTM

$$P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{T-1}, \mathbf{x}_T) = \Pi_{\{i=1\}}^{T} p_i$$

Minimize anomaly scores: $-\sum_{i=1}^{t} \log(p_i)$

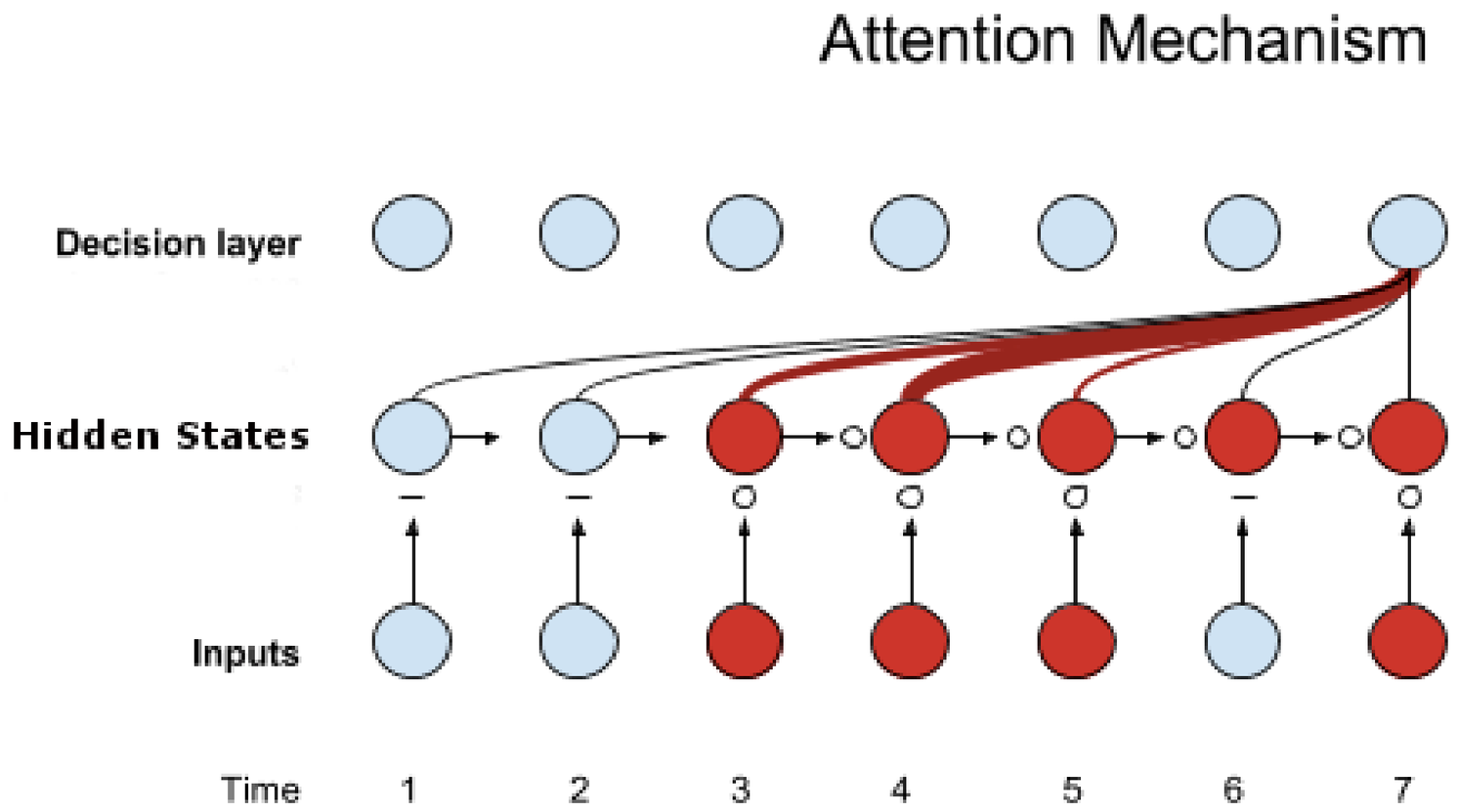
Pacific Northwest NATIONAL LABORATORY

Tiered Event Models (T-EM/T-BEM)



Pacific Northwest
NATIONAL LABORATORY

Attention



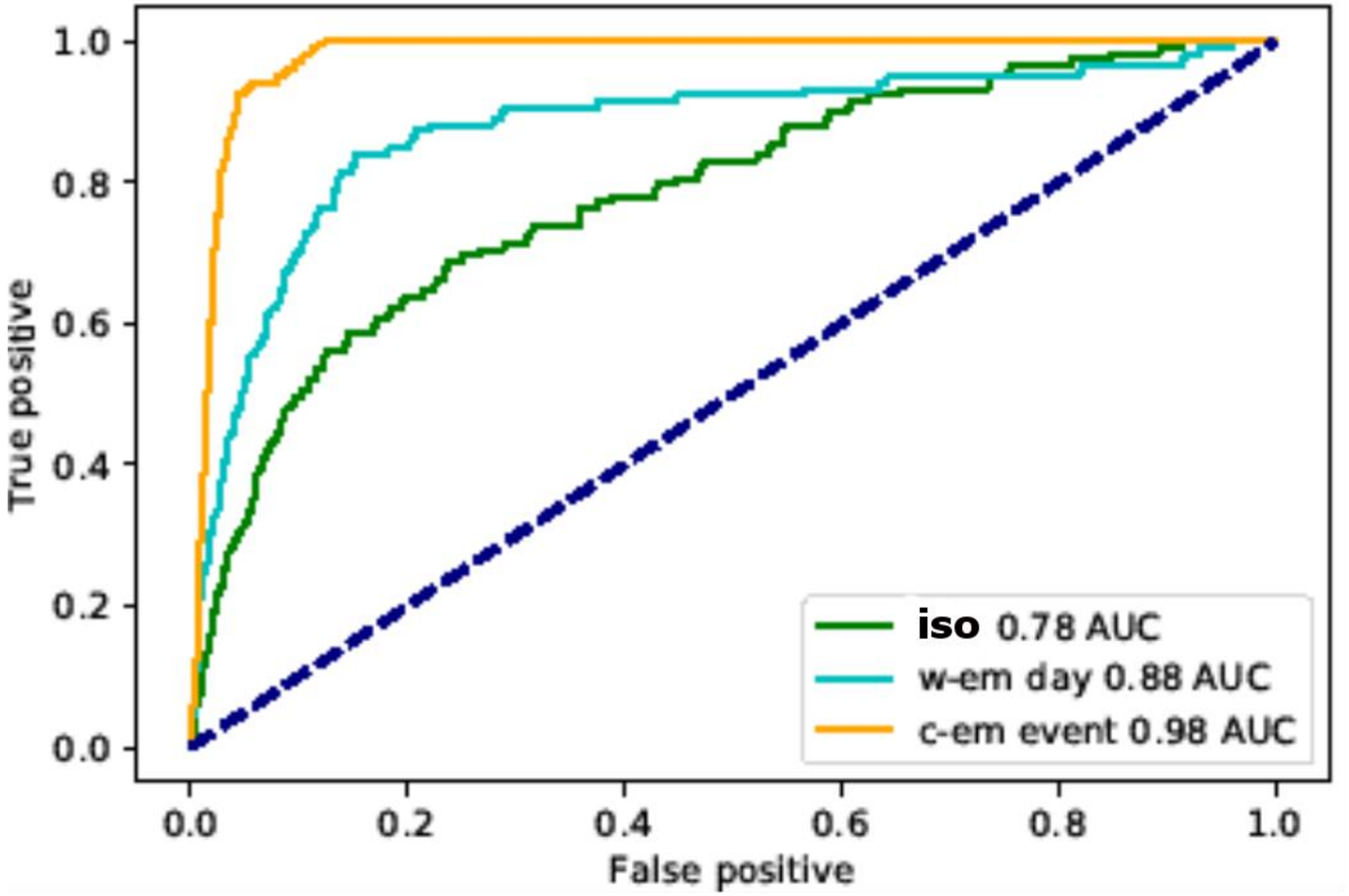
Pacific Northwest
NATIONAL LABORATORY

Experiment Setup

- Data
  - LANL cyber security data set authentication logs.
  - 0.00007% of events are marked as Red Team activities.
- Performance Metric:
  - Area under the Receiver Operating Characteristic Curve (AUC)
- Baseline Comparison
  - Baseline models use user-day aggregate statistics.
  - Use max event anomaly score for user on that day for language models.
  - Also evaluate language models on a per-event basis.

Pacific Northwest
NATIONAL LABORATORY

Experiment Results Vs Baseline



Word Models

Model	Mean	Max	Min	Std. Dev.
EM	0.968	0.976	0.964	0.005
BEM	0.976	0.981	0.972	0.003
EM with attention				
Fixed	0.974	0.976	0.972	0.001
Syntactic	0.972	0.975	0.967	0.004
Semantic 1	0.975	0.980	0.971	0.004
Semantic 2	0.973	0.976	0.968	0.003
Tiered LSTM variants				
T-EM	0.984	0.989	0.977	0.005
T-BEM	0.987	0.989	0.985	0.002
TA-EM	0.985	0.991	0.979	0.004
TA-BEM	0.988	0.991	0.984	0.003

- Best performing single tier model: **Semantic I**
- Higher ROC than the simple **Event Model**

Pacific Northwest NATIONAL LABORATORY

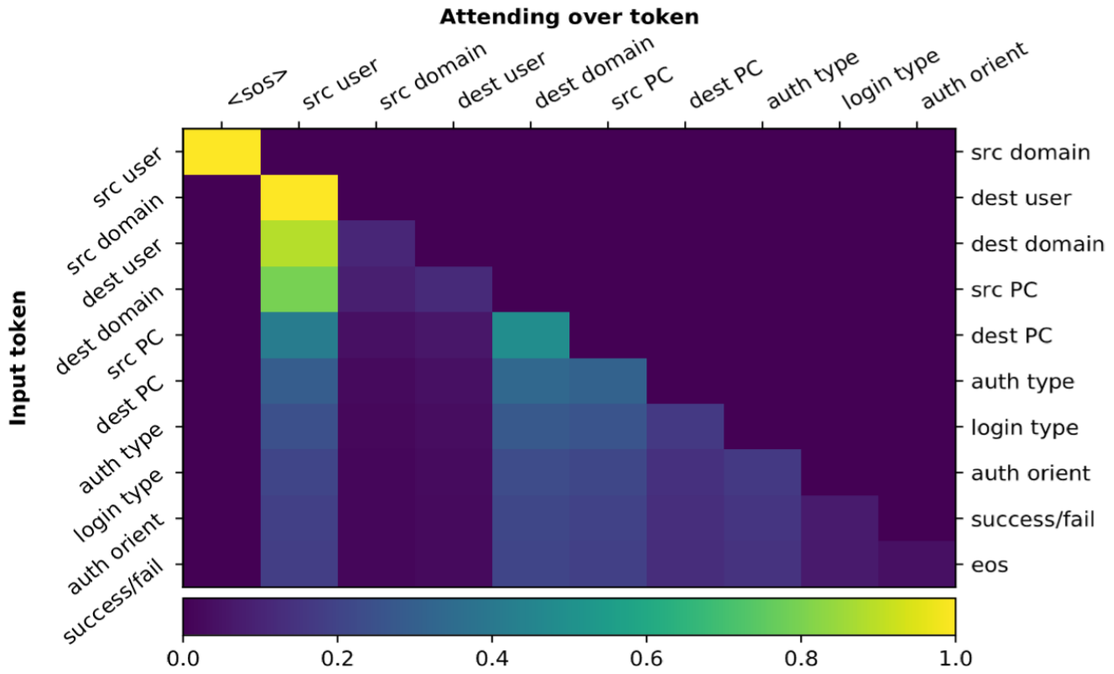
Word Models

Model	Mean	Max	Min	Std. Dev.
EM	0.968	0.976	0.964	0.005
BEM	0.976	0.981	0.972	0.003
EM with attention				
Fixed	0.974	0.976	0.972	0.001
Syntactic	0.972	0.975	0.967	0.004
Semantic 1	0.975	0.980	0.971	0.004
Semantic 2	0.973	0.976	0.968	0.003
Tiered LSTM variants				
T-EM	0.984	0.989	0.977	0.005
T-BEM	0.987	0.989	0.985	0.002
TA-EM	0.985	0.991	0.979	0.004
TA-BEM	0.988	0.991	0.984	0.003

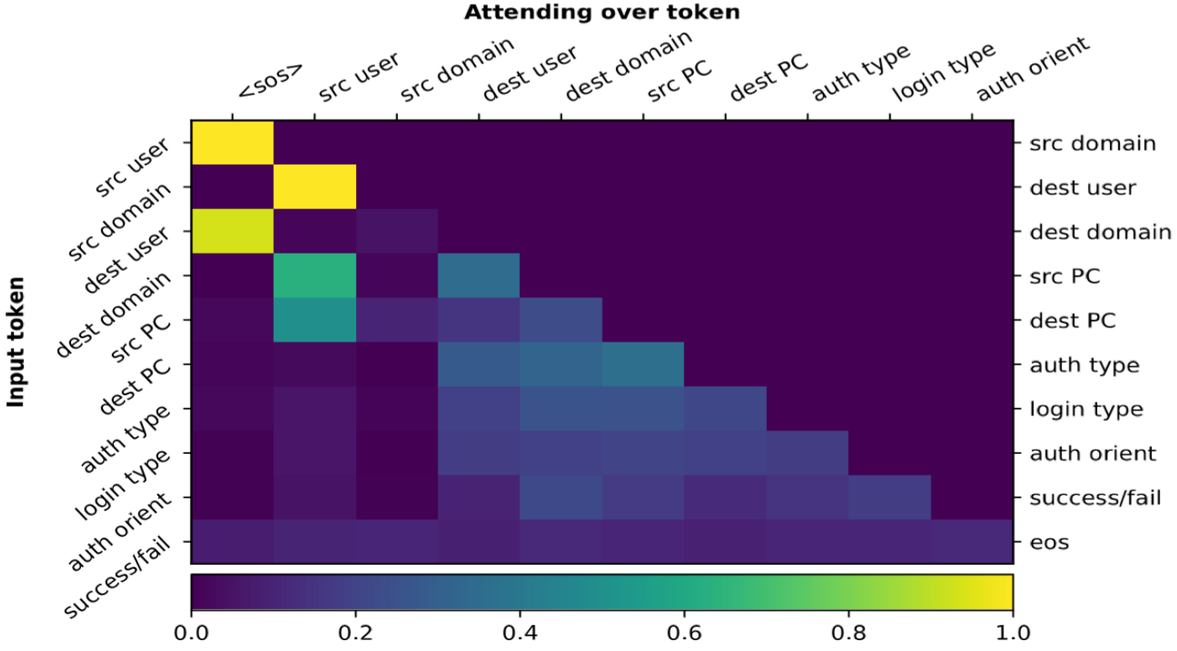
- **Attention models** perform only marginally worse than **bidirectional models**

Global Average Importance of Fields

Fixed Word Model



Syntax Word Model



Pacific Northwest
NATIONAL LABORATORY

First Question

- Can ML be used in security applications where malicious patterns are not predefined?

Two Use Cases:

NLP analysis of cyber data for insider threat detection

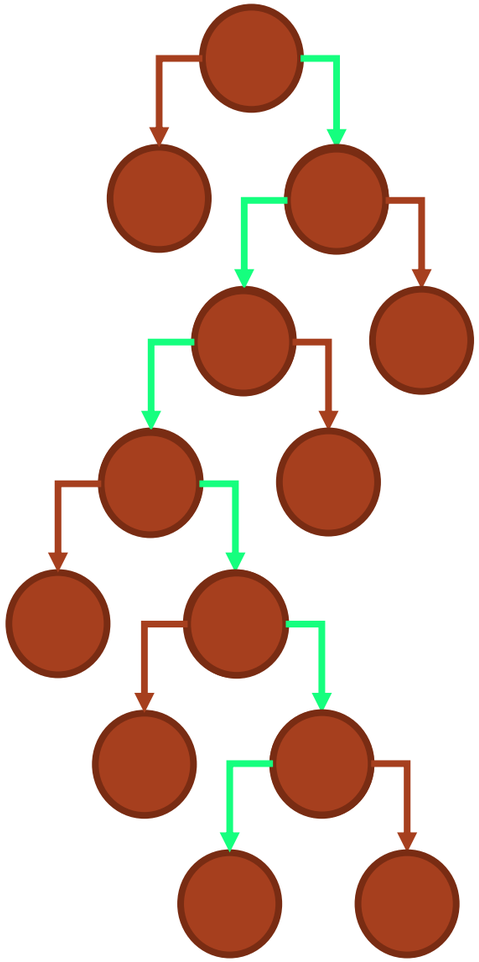
Neural Fuzzing for accelerating software security assessments

- Can ML itself be secure in deployments?

Pacific Northwest
NATIONAL LABORATORY

Goal

Accelerate search for unique code paths that could reveal faults



Assumptions

Faults are more likely to exist on untested / unexplored code paths

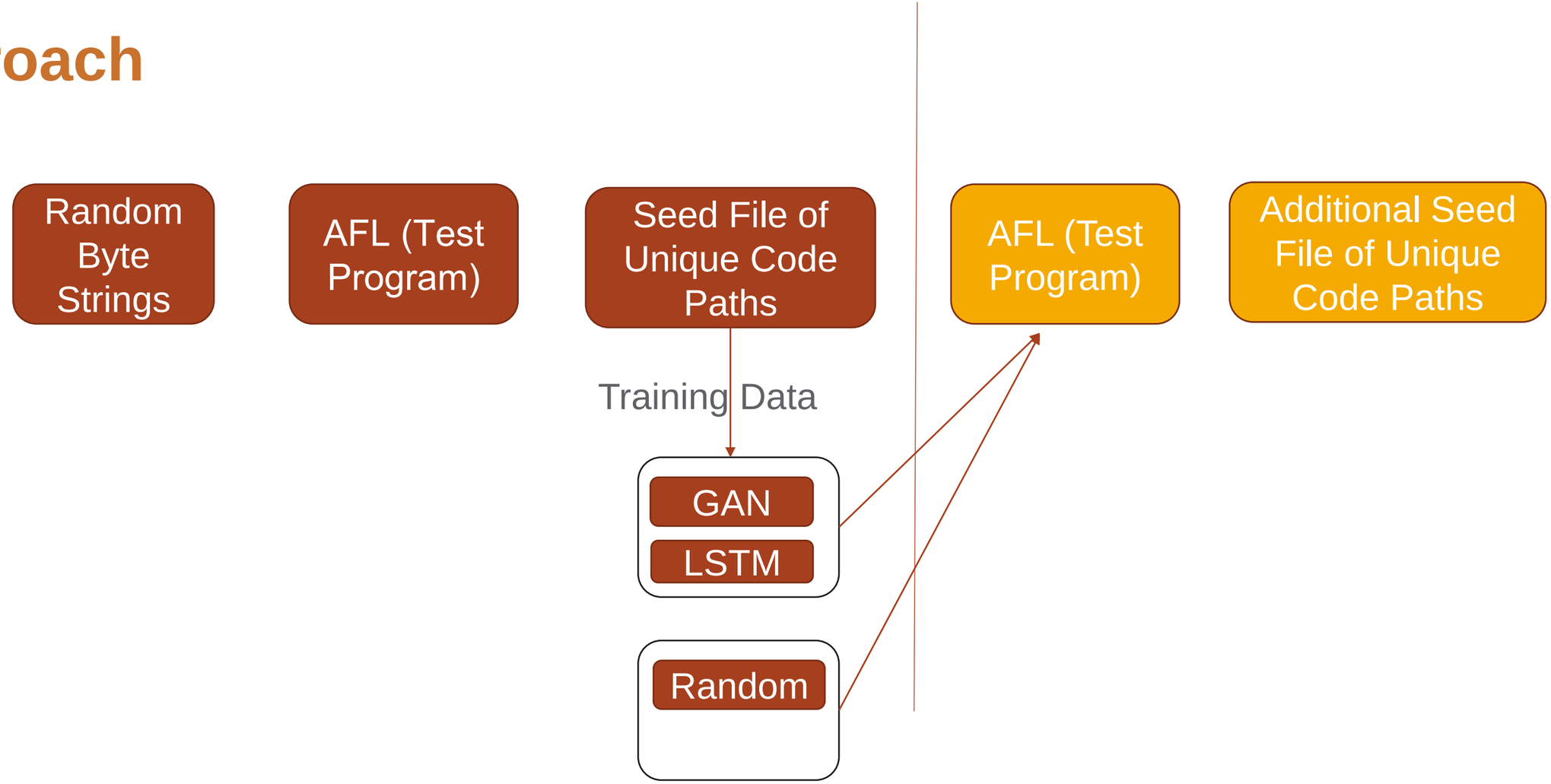
Shorter paths are easier to test / explore than longer paths

Approach

Augment American Fuzzy Lop (AFL) with LSTM and GANS generated seed files to accelerate search.

Pacific Northwest
NATIONAL LABORATORY

Approach



Pacific Northwest NATIONAL LABORATORY

Analysis of Seed Files

Class C	$\|C\|$	$L(C)$	% Unique	$\mu(L(C))$	$\sigma(L(C))$
AFL seed	38384	31212	0.813	26.968M	33.958M
Rand seed	19824	485	0.024	2.602M	724.674K
LSTM seed	20000	1921	0.096	2.596M	8.687K
GAN seed	20000	119	0.006	2.593M	1.841K

- The seed themselves are **not** what we are interested in measuring
- They only provide a set of *initial conditions* for AFL
- Interestingly LSTM and GAN do have as much variance as using purely random seeds

Pacific Northwest
NATIONAL LABORATORY

Time Analysis of Sustained Run

- Both LSTM and GAN outperform random sampling for discovering new unique code paths.
- GAN 11 % faster / random
- LSTM 8% faster over random

Class	Files	% new	sec/path	NRate
Rand	1231	0.9017	214.478	1.00
LSTM	1251	0.8984	197.130	1.08
GAN	1240	0.8694	191.893	1.11

Code Path Length of Sustained Run

- length of unique code paths using **GAN** was **13.84%** longer than a strategy based on randomly sampling.
- length of unique code paths using **LSTM** was **4.60%** longer than a strategy based on randomly sampling.

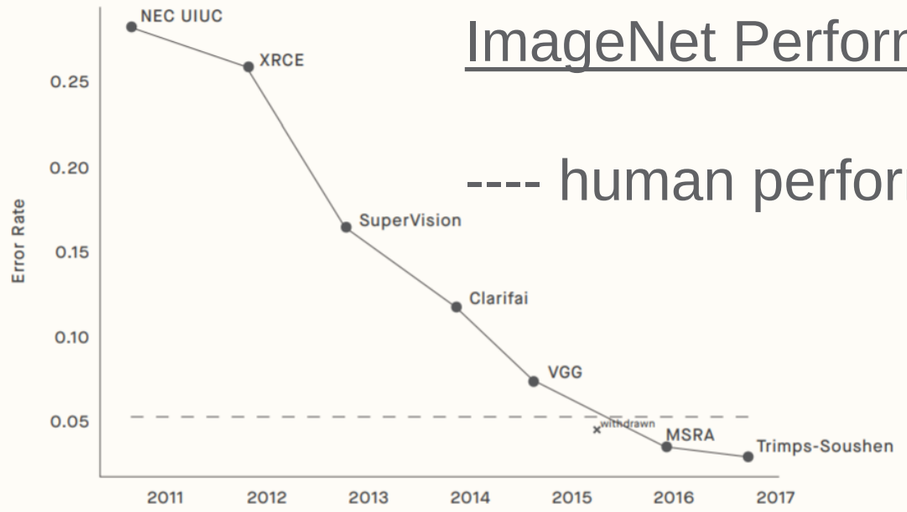
Class	$\mu(L(C))$	$\sigma(L(C))$
Rand	25.373M	3.339M
LSTM	26.541M	3.385M
GAN	28.885M	3.456M

Pacific Northwest
NATIONAL LABORATORY

Second Question

- Can ML detect malicious behavior without predefined patterns?
- Can ML itself be secure in deployments?

Adversarial Machine Learning

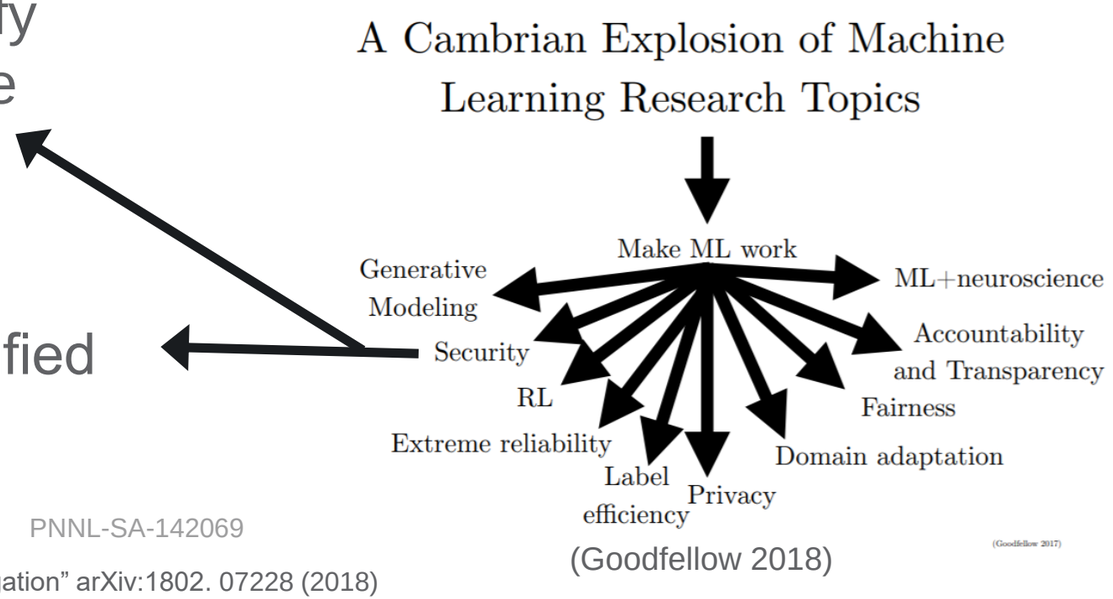


ImageNet Performance

---- human performance

Digital Attacks:
Direct access to maliciously modify model, input features, or database of training examples.

Physical Attacks:
A physical object is added or modified in the scene being evaluated.



“The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation” arXiv:1802. 07228 (2018)

Pacific Northwest NATIONAL LABORATORY

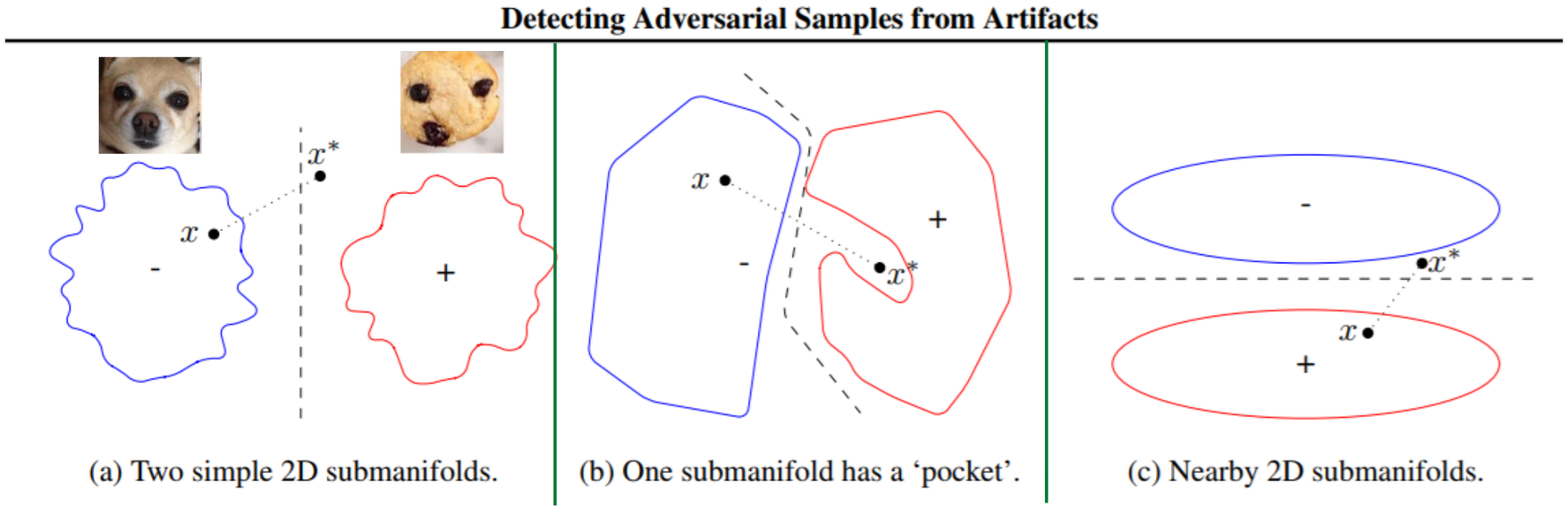
Why is Machine Learning Vulnerable?

- **Known general ML fragilities…**
  - Every model has a decision boundary; manifolds can be wrinkly
  - Not enough training data to resolve boundary cases (chihuahua/muffin)
  - Not all classes are separable
  - High dimensional space is not intuitive
  - Decisions are hard to understand
  - Poisoned the training data (GIGO)
  - Compromise of privacy in the training data
  - Denial of service, output corruption, hacks…

- **Additional DL vulnerabilities**
  - No mystery: DL models are the approach of choice for many problems
  - Limited diversity: A few training sets, standard architectures, standard models
  - Spotlight: Many researchers are publishing DL-specific attacks

Decision Boundaries

- Data driven models are only as good as their data
- Training data cannot fully define a decision boundary
- What is going on with vulnerability and misclassification:



I might be a muffin.

Definitely a muffin!

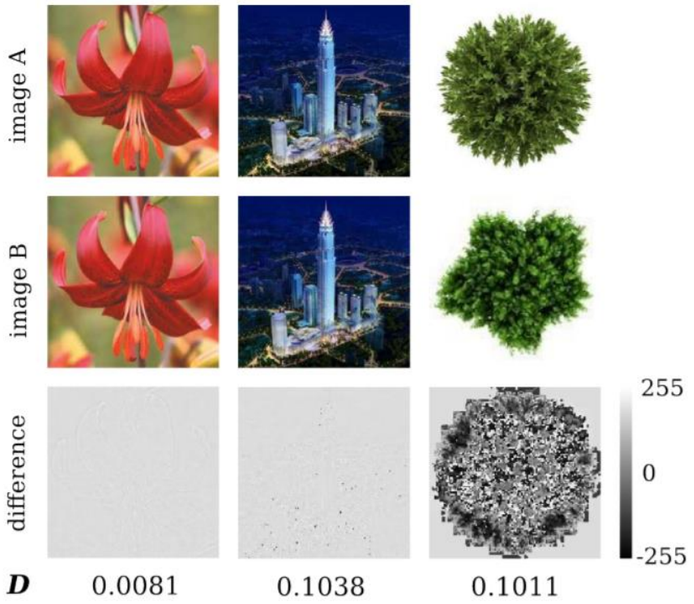
Not a good Chihuahua, don't know what I am.

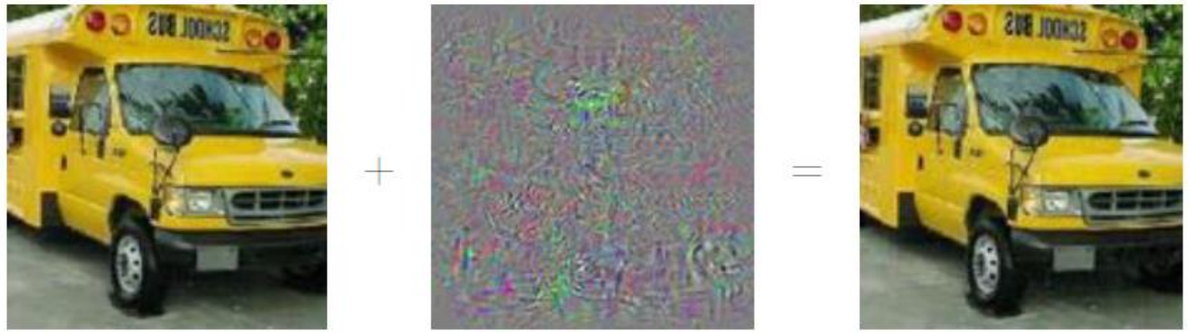
Feinman et al. "Detecting adversarial samples from artifacts." *arXiv preprint arXiv:1703.00410* (2017).

Attacks in the Digital Domain

- **Adversarial Example - model input an attacker has intentionally designed to cause the model to make a mistake.**

- **Distance in feature space is not always intuitive.**
- **Numerous ways to craft adversarial examples.**





Szegedy et. al., “Intriguing properties of neural networks” arXiv preprint arXiv:1312.6199 (2013)
Zheng, Stephan, et al. "Improving the robustness of deep neural networks via stability training." *Proceedings of the ieee conference on computer vision and pattern recognition*. 2016.

Attacks in the Physical World

Physical attacks span significant range of perception and detectability

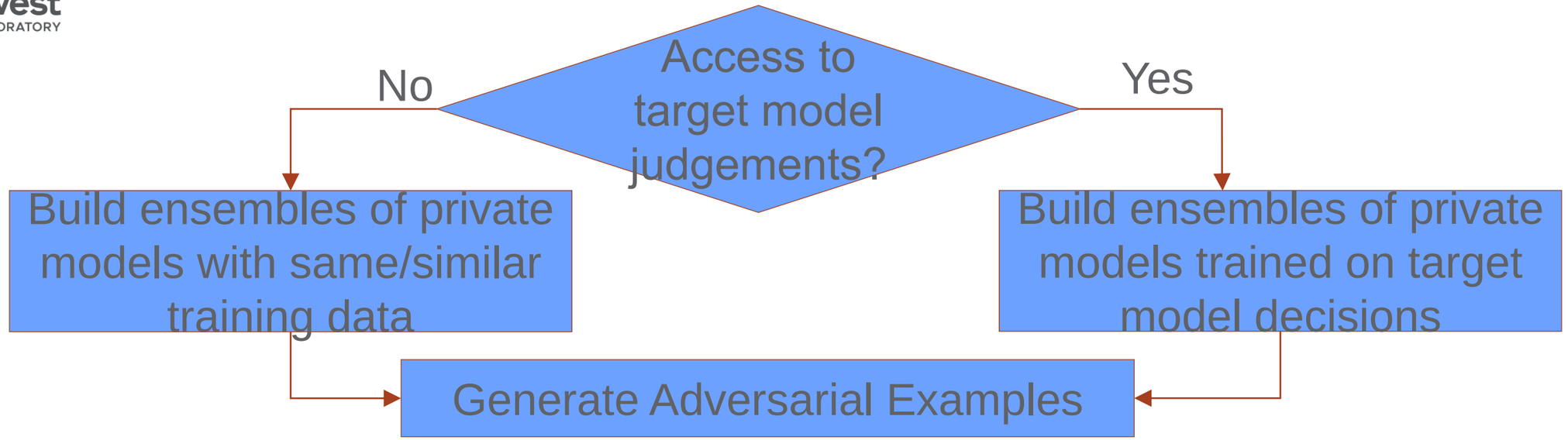
- Targeted attacks
- 2 and 3D object construction
- Digital formulation for

physical world deployment (White box attacks)

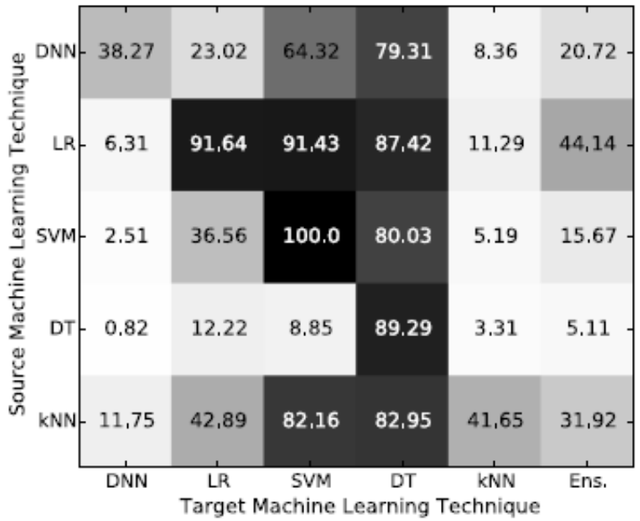
Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

[1]Athalye, Anish, and Ilya Sutskever. "Synthesizing robust adversarial examples." *arXiv preprint arXiv:1707.07397*(2017).
[2]Sharif et. al., “Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition” Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016, pp. 1528–1540.
[3] Evtimov, Ivan, et al. "Robust physical-world attacks on machine learning models." *arXiv preprint arXiv:1707.08945*(2017).
[4] Brown, Tom B., et al. "Adversarial patch." *arXiv preprint arXiv:1712.09665* (2017).

Transferability of Adversarial Examples



- **Sometimes examples just transfer!**
  - Transfer is not guaranteed
  - Exploit commonalities in development
    - ✓ <10 large-scale image training libraries
    - ✓ <10 major DL generation libraries
- **Decision boundaries for models of the same class are likely to be similar**



Papernot et. al. "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples" CoRR, arXiv:1605.07277 (2016).

Experiment Inception

Goal 1: Can light cause misclassification of 2D print images

Goal 2: Can light cause misclassification of 3D objects

Goal 3: What is the stability of this approach.

(a) Printout (b) Photo of printout (c) Cropped image

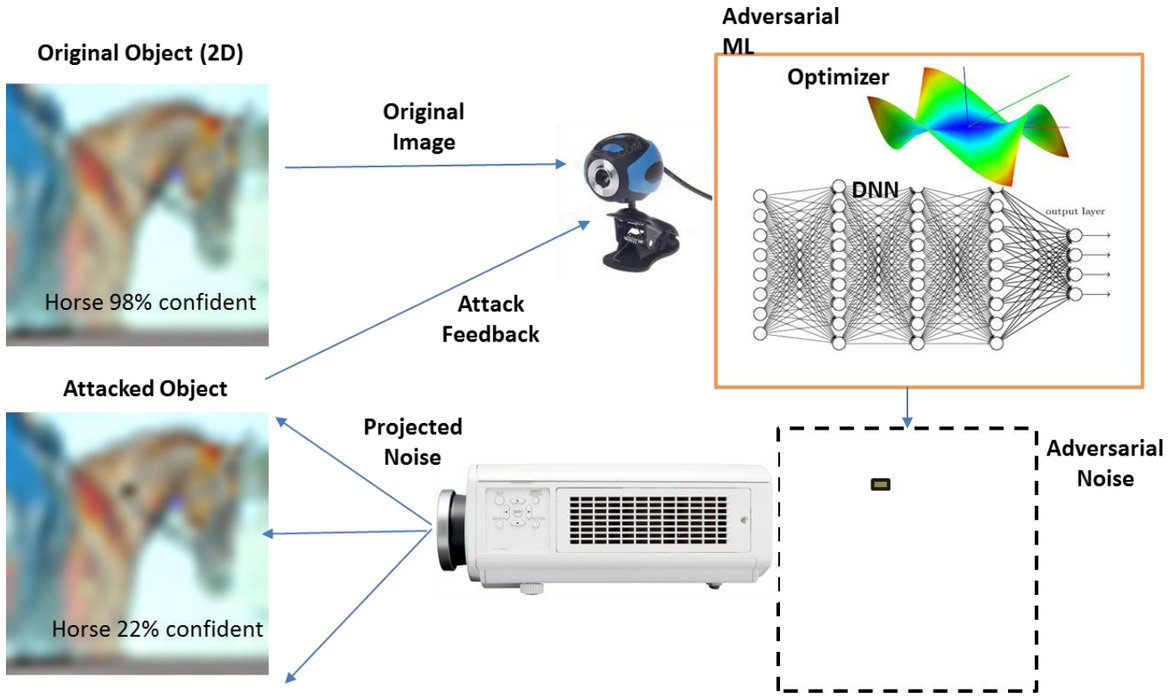
Inspired by :

Kurakin, A., Goodfellow, I., and Bengio, S. "Adversarial examples in the physical world." *arXiv preprint arXiv:1607.02533* (2016).

Projecting Trouble- 2D Experiments

Transient physical attacks

- CIFAR10 dataset and pre-trained ResNet38 classifier.
- Non-targeted and false negative attack
- Differential Evolution, white-ish box attack (crafted to the image but without knowledge of classification model)



3D Presentation Experiment

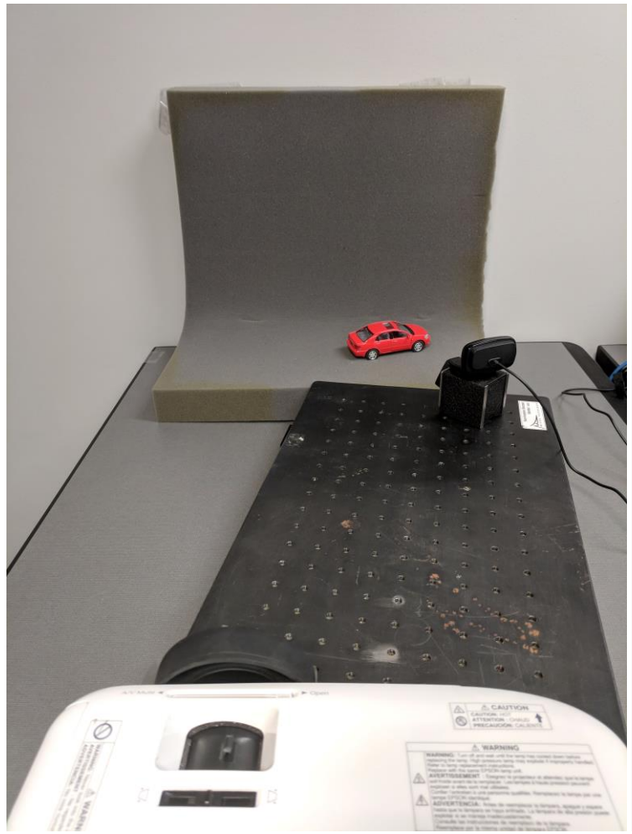
Prob.	p(airplane)	p(automobile)	p(bird)	p(cat)	p(deer)	p(dog)	p(frog)	p(horse)	p(ship)	p(truck)
Original	0	89%	0	0	0	0	0	0	0	11%
Attacked	0	43%	0	0	0	0	0	0	0	57%

- 3D attacks can be successful
- In CFAR10, trucks are semi-trailers, fire trucks, etc, thus bigger difference to shift.
- Non-targeted, transient attack.

Pacific Northwest
NATIONAL LABORATORY

3D CIFAR Experiment

- One example of each CIFAR10 class.
- Environmental control
- Additional Baseline attacks (white light, random square, DE square)
- ImageNet co-classification



Pacific Northwest NATIONAL LABORATORY

Results

CIFAR Class	Experiment Condition	Mean	Median	SD	Var	Min	Max	Δ Mean	Δ Median
Airplane 	Baseline	1.000	1.000	.000	.000	1.000	1.000	.000	.000
	White Light	.151	.101	.198	.039	.017	.997	.849	.899
	Random	.114	.105	.088	.008	.022	.445	.886	.895
	Diff Evolution	.133	.112	.087	.007	.014	.459	.867	.888
Automobile 	Baseline	1.000	1.000	.000	.000	1.000	1.000	.000	.000
	White Light	1.000	1.000	.000	.000	.999	1.000	.000	.000
	Random	1.000	1.000	.000	.000	.999	1.000	.000	.000
	Diff Evolution	1.000	1.000	.000	.000	1.000	1.000	.000	.000
Bird 	Baseline	1.000	1.000	.000	.000	1.000	1.000	.000	.000
	White Light	1.000	1.000	.002	.000	.993	1.000	.000	.000
	Random	1.000	1.000	.000	.000	1.000	1.000	.000	.000
	Diff Evolution	1.000	1.000	.000	.000	.999	1.000	.000	.000
Cat 	Baseline	.990	.991	.004	.000	.979	.996	.000	.000
	White Light	.009	.008	.005	.000	.000	.020	.981	.983
	Random	.011	.007	.012	.000	.001	.047	.979	.984
	Diff Evolution	.023	.017	.019	.000	.002	.124	.967	.974
Deer 	Baseline	.999	.999	.000	.000	.999	1.000	.000	.000
	White Light	.516	.516	.145	.021	.242	.997	.483	.483
	Random	.545	.507	.155	.024	.327	.871	.454	.492
	Diff Evolution	.473	.467	.130	.017	.144	.829	.526	.532

Pacific Northwest NATIONAL LABORATORY

Results

CIFAR Class	Experiment Condition	Mean	Median	SD	Var	Min	Max	Δ Mean	Δ Median
Dog 	Baseline	.993	.993	.003	.000	.986	.996	.000	.000
	White Light	.512	.499	.088	.008	.390	.695	.481	.494
	Random	.482	.497	.123	.015	.136	.753	.511	.496
	Diff Evolution	.386	.388	.088	.008	.123	.601	.606	.605
Frog 	Baseline	.888	.888	.025	.001	.842	.933	.000	.000
	White Light	.008	.008	.003	.000	.000	.015	.881	.880
	Random	.030	.011	.076	.006	.004	.360	.858	.877
	Diff Evolution	.071	.038	.093	.009	.005	.576	.817	.849
Horse 	Baseline	1.000	1.000	.000	.000	1.000	1.000	.000	.000
	White Light	.999	1.000	.001	.000	.993	1.000	.000	.000
	Random	1.000	1.000	.000	.000	1.000	1.000	.000	.000
	Diff Evolution	1.000	1.000	.000	.000	1.000	1.000	.000	.000
Ship 	Baseline	1.000	1.000	.000	.000	1.000	1.000	.000	.000
	White Light	1.000	1.000	.000	.000	1.000	1.000	.000	.000
	Random	1.000	1.000	.000	.000	1.000	1.000	.000	.000
	Diff Evolution	1.000	1.000	.000	.000	1.000	1.000	.000	.000
Truck 	Baseline	1.000	1.000	.000	.000	1.000	1.000	.000	.000
	White Light	.832	.832	.052	.003	.729	1.000	.168	.168
	Random	.818	.819	.072	.005	.634	.970	.182	.180
	Diff Evolution	.826	.839	.088	.008	.507	.949	.174	.161

Table 1: Classification statistics for baseline and attacked CIFAR figures.

Results

- Extreme variability between target class susceptibility to attack.
  - 6 of 10 classes were susceptible to light based attacks.

	Ave(Δ Mean)	Ave(Δ Median)
White Light	0.641	0.651
Random	0.645	0.654
Diff Evolution	0.660	0.668

- White light was similarly effective to random squares and DE.
- Rotation, lighting, and scale invariance of classification models are significant considerations.

Pacific Northwest
NATIONAL LABORATORY

Conclusions

ML for Security:

- Deep learning techniques can be used to enhance and accelerate a variety of security based applications.
- Pre-knowledge of patterns is not necessary in insider threat detection or software fuzzing.

Security for ML:

- Most off the shelf models are insufficiently resilient to real world invariance.
- An increasing range of digital and physical security gaps are being identified in ML models.

Security of the model itself needs to be considered, particularly when deploying ML for Security.

Pacific Northwest
NATIONAL LABORATORY

References

Recurrent Neural Network Language Models for Open Vocabulary Event-Level Cyber Anomaly Detection
https://arxiv.org/pdf/1712.00557.pdf

Deep Learning for Unsupervised Insider Threat Detection in Structured Cybersecurity Data Streams
https://arxiv.org/pdf/1710.00811.pdf

Faster Fuzzing: Reinitialization with Deep Neural Models
https://arxiv.org/pdf/1711.02807.pdf

Projecting Trouble:
Light Based Adversarial Attacks on Deep Learning Classifiers
https://arxiv.org/abs/1810.10337

Code available at:
https://github.com/pnnl/safekit

Pacific Northwest
NATIONAL LABORATORY

Thank you

PNNL-SA-142069