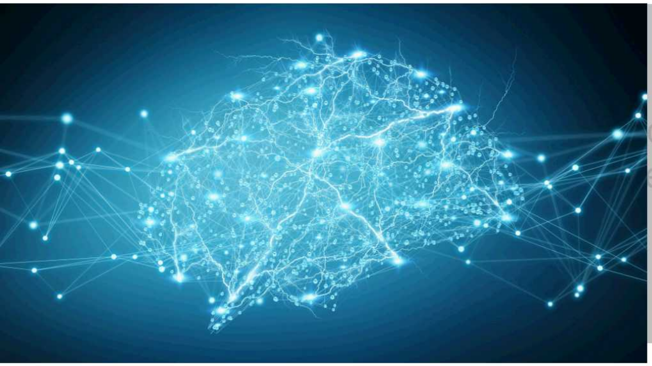
SAND2020-7792C



MalGen: On Bridging the Semantic Gap between Machine Learning and Malware Analysis

Michael R. Smith (5952), Curtis Johnson (PM, 5952)

Armida Carbajal (1462), Eva Domschot (NMT), Bridget Haus (56291), Joe Ingram (9365), Nick Johnson (5953), Philip Kegelmeyer (8700), Chris Lamb (8851), Ramyaa Ramyaa (NMT), Steve Verzi (8722)

U.S. DEPARTMENT OF ENERGY NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND NO. 2011-XXXXP

A Lot of ML Success

Sandia National Laboratories

I Am a Model and I Know That Artificial Intelligence Will Eventually Take My Job

Tencent and Chinese scientists use deep learning to predict fatal COVID-19 cases

Rita Liao @ritacyliao / 11:02 pm PDT • July 21, 2020 Comment

Google's Fabricius uses machine learning to decode hieroglyphs

You can also generate messages in ancient Egyptian "emojis" to send for fun.

CMU and Facebook AI Research use machine learning to teach robots to navigate by recognizing objects

Brian Heater @bheater / 11:49 am PDT • July 20, 2020 Comment

Machine learning helps robot swarms coordinate

by California Institute of Technology

Disney's deepfakes are getting closer to a big-screen debut

The first megapixel-resolution deepfakes

Deep learning enables early detection and classification of live bacteria using holography

by UCLA Engineering Institute for Technology Advancement

A Lot of ML Success

I Am a Model and I Know That Artificial Intelligence Will Eventually Take My Job

Tencent and Chinese scientists use deep learning to predict fatal COVID-19 cases

Rita Liao @ritacyliao / 11:02 pm PDT • July 21, 2020 Comment

Google's Fabricius uses machine learning to decode hieroglyphs

You can also generat

CMU and Faceboo
learning to teach r
objects

Brian Heater @bheater / 11:49 am PDT • July 20, 2

Noticeably missing is success in malware analysis (MA)

ot swarms

Disney's deepfakes are getting closer to a big-screen debut

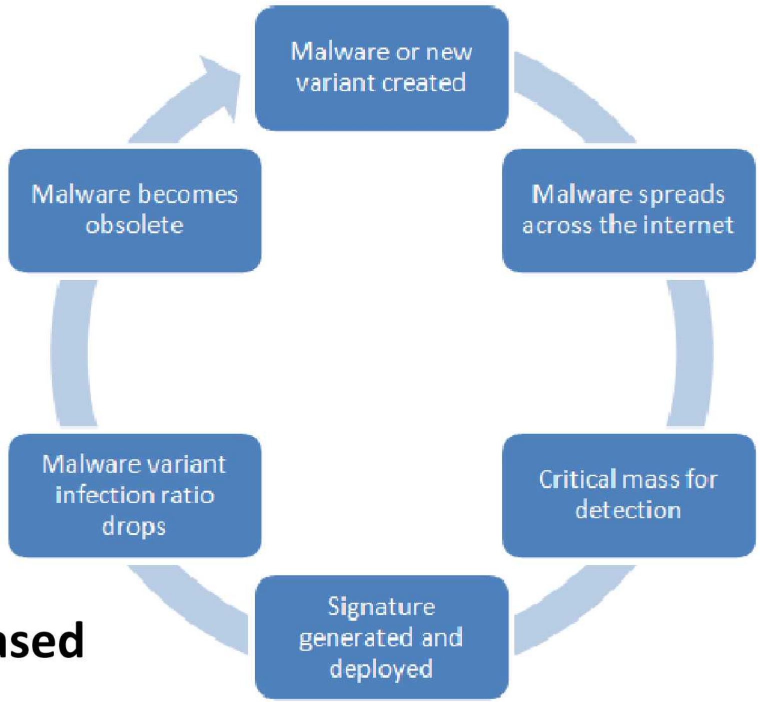
The first megapixel-resolution deepfakes

Deep learning enables early detection and classification of live bacteria using holography

by UCLA Engineering Institute for Technology Advancement

Current Approaches for MA

- Malware causes A LOT of damage
  - Average cost of a malware attack on a company is **$2.4 million** [1]
- Malware is hard to detect
  - New and evolving variants **up by 88 percent** [3]
  - Hides to avoid detection **93% was polymorphic** [4]
  - Most executables are benign **only 2% are malicious** [4]
- Most malware detection is **signature based**
  - **Easy to bypass** with trivial modifications
  - Very **labor intensive** and **does not scale**
  - **Reactive** to known malware
- Machine Learning (ML) techniques can help address these issues
- Semantic gap—Claimed success but **lack real-world impact**

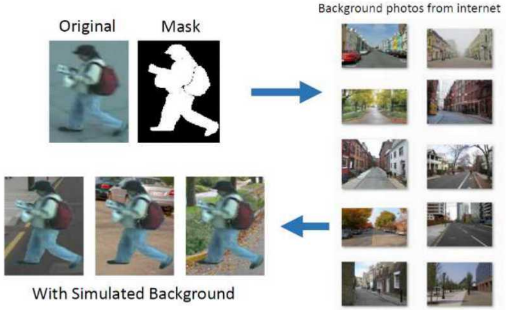


Success of ML in Other Domains

- Build on decades of previous research
  - CNNs are a codification of convolutions which have a long history in signal processing
  - Translational invariance which is inherently important in signal processing—analogous operators do not yet exist for binary analysis
- Large amounts of labeled, relevant data (benchmark dataset)
  - Fei Fei Li: ImageNet (3 years to create) [6]
  - Synthetic data: Generative models and other synthetic data
  - Peter Norvig (Google Research Director): **"We don't have better algorithms. We just have more data."**

after 5 epochs

after 100 epochs



Success of ML in Other Domains

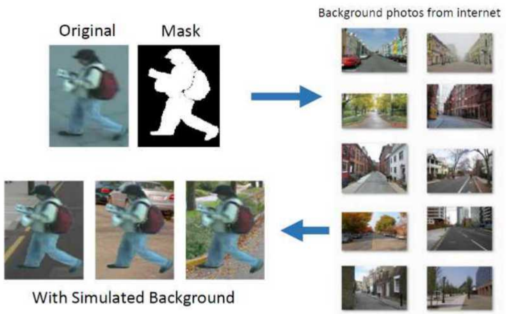
- Build on decades of previous research
  - CNNs are a codification of convolutions which have a long history in signal processing
  - Translational invariance which is inherently important in signal

Hypothesis: The objectives of ML and MA are misaligned as evidenced (among other things) in the data—preventing more success in ML for MA

  - Synthetic data: Generative models and other synthetic data
  - Peter Norvig (Google Research Director): **"We don't have better algorithms. We just have more data."**

after 5 epochs

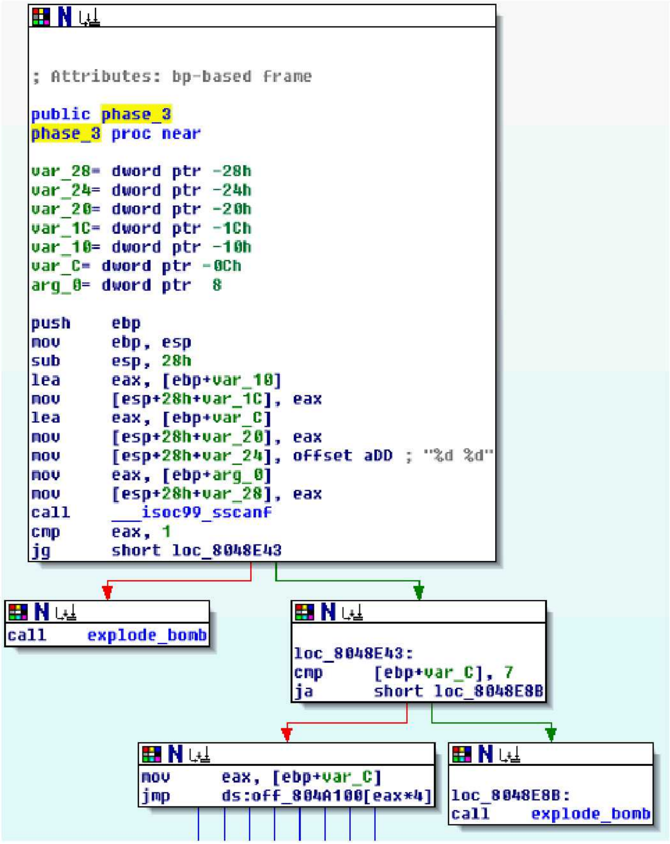
after 100 epochs



How to Represent the Data

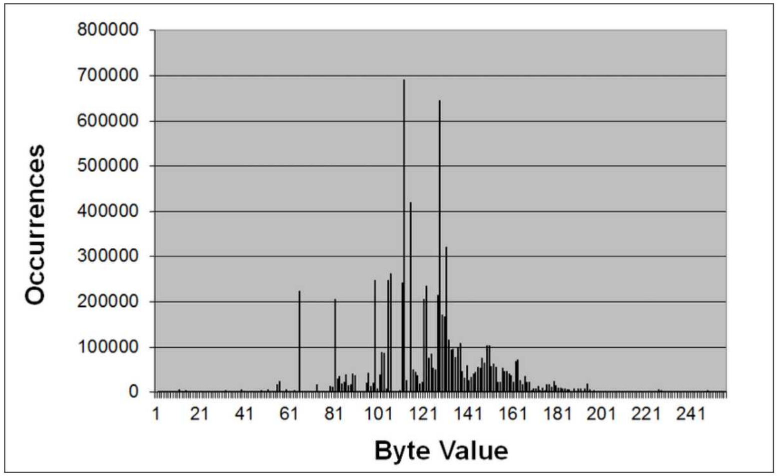
- Semantic VS syntactic features
  - Do extracted features convey information relevant to malware analysis?

Disassembly



“Image” of malware

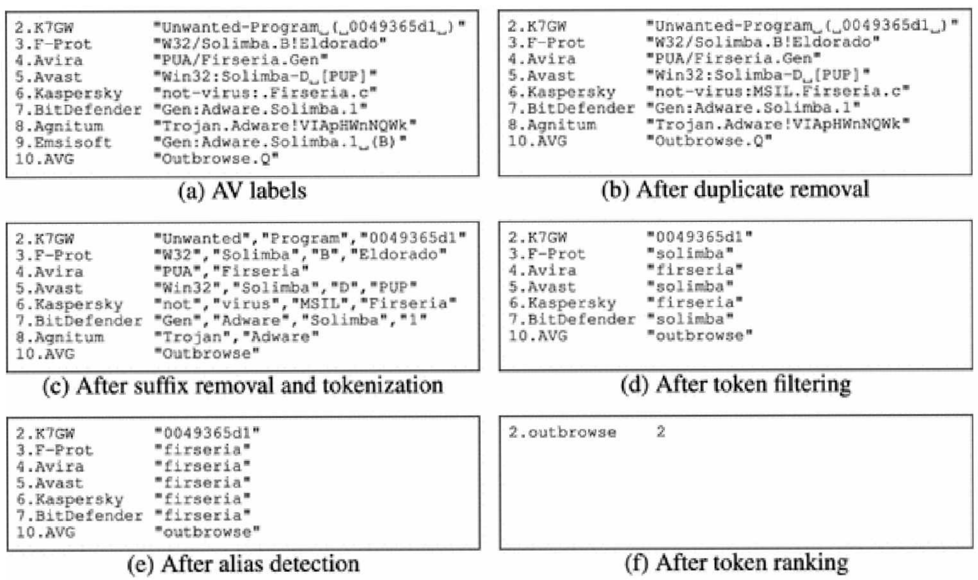
Byte Counts



Tradeoff between semantic information and computational complexity

Obtaining Labels

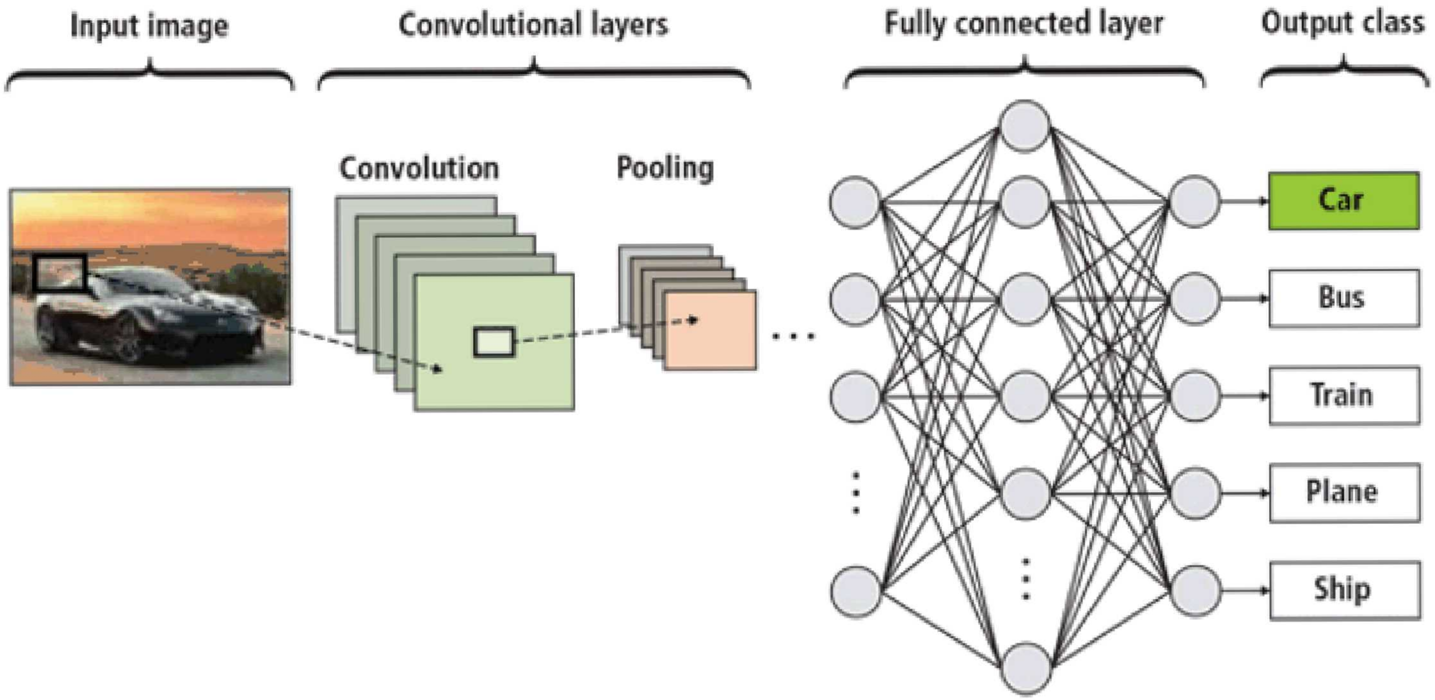
- AV labels are known to be inconsistent [8]
  - Data becomes too "easy"—most popular malware is used
- Data is not representative
  - Lots of script-kiddies and variants
  - Few more sophisticated malware that we care about



(a) AV labels
(b) After duplicate removal
(c) After suffix removal and tokenization
(d) After token filtering
(e) After alias detection
(f) After token ranking

Other Considerations

- Data does not align with ML assumptions
  - Proximity
  - Continuity
  - Ordinality



Data for ML in MA

Dataset	Year	Cite	Representations	# Samples	Labels	Labeling	Max Acc
			Highly Cited				
VX Heaven [4]	2010	?	Live executables	Varies	Varies	Curated	N/A[1]
VirusShare [2]	2011	> 300	Live executables	Varies	Varies	Curated	N/A[1]
MalImg [44]	2011	417	Gray-scale images	9,458	25 Families	MSSE	99.80%
MS Malware Classification [58]	2015	76	Disassembly and hexadecimal	10,868	9 Families	MSSE	99.97%
EMBER [10]	2017	46	Parsed and histogram counts	1,100,000	Good, Bad, ?	VirusTotal	99.90%
MalRec [67]	2018	11	System calls, memory contents[2]	66,301	1,270 families	VirusTotal[3]	N/A[1]
			Less Cited				
Malware Training Sets [57]	2016	2	Counts from analysis reports	4764	4 families	Curated	-
Mal-API-2019 [12]	2019	1	System call traces	7,107	8 families	VirusTotal	-
Meraz'18 Kaggle [5]	2018	~1	Parsed features	88,347	Good v Bad	Curated	91.40%[4]

[1] There is no established dataset making comparisons between studies difficult.

[2] Also provides full system replays of malware execution, however the authors note non-trivial efforts to get them to work on other systems.

[3] Uses AVClass [66] which leverages VirusTotal.

[4] Reported accuracy on the Kaggle challenge leader board.

Data for ML in MA

Dataset	Year	Cite	Representations	# Samples	Labels	Labeling	Max Acc
			Highly Cited				
VX Heaven [4]	2010	?	Live executables	Varies	Varies	Curated	N/A[1]
VirusShare [2]	2011	> 300	Live executables	Varies	Varies	Curated	N/A[1]
MalImg [44]	2011	417	Gray-scale images	9,458	25 Families	MSSE	99.80%
MS Malware Classification [58]	2015	76	Disassembly and hexadecimal	10,868	9 Families	MSSE	99.97%
EMBER [10]	2017	46	Parsed and histogram counts	1,100,000	Good, Bad, ?	VirusTotal	99.90%
MalRec [67]	2018	11	System calls, memory contents[2]	66,301	1,270 families	VirusTotal[3]	N/A[1]
			Less Cited				
Malware Training Sets [57]	2016	2	Counts from analysis reports	4764	4 families	Curated	-
Mal-API-2019 [12]	2019	1	System call traces	7,107	8 families	VirusTotal	-
Meraz'18 Kaggle [5]	2018	~1	Parsed features	88,347	Good v Bad	Curated	91.40%[4]

1 There is no established dataset making comparisons between studies difficult.

2 Also provides full system replays of malware execution, however the authors note non-trivial efforts to get them to work on other systems.

3 Uses AVClass [66] which leverages VirusTotal.

4 Reported accuracy on the Kaggle challenge leader board.

Case Study: DL System Calls

- Use machine learning to monitor system calls
  - ML can generalize away from static signatures
  - System calls are the base level for interacting with the operating system
- How well would deep learning methods do in detecting malware using dynamic analysis?
  - Histograms with Random Forests
  - Long Short-Term Memory Recurrent Neural Networks (LSTM)
  - Convolutional Neural Networks (CNNs)
  - Liquid State Machines (LSMs)

Case Study: DL System Calls

- 1000 time steps sequence length
- Each method achieves a class averaged accuracy greater than 90%
- Random Forests out perform the deep learning approaches
  - Statistical significance over the LSTM
- Ensemble statistically significantly outperforms all other approaches

Alg	Acc	CAA	MPr	MRe
Hist+RF	**95.3**	94.7	0.953	**0.926**
CNN	94.0	93.2	0.946	0.896
LSTM	91.3	90.0	0.926	0.843
CNN+LSTM	94.5	93.7	0.956	0.901
LSM	90.7	89.8	0.856	0.856
Ensemble	**95.3**	**95.5**	**0.962**	0.917

	Hist+ RF	CNN	LSTM	CNN+ LSTM	LSM
Ensemble	**YES**	**YES**	**YES**	**YES**	**YES**
Hist+RF	-	NO	**YES**	NO	NO
CNN	NO	-	**YES**	NO	NO
LSTM	**YES**	**YES**	-	**YES**	NO
CNN+LSTM	NO	NO	**YES**	-	**YES**
LSM	NO	NO	NO	**YES**	-

Case Study: DL System Calls

- Sorted: data set with roughly balanced benign and malicious
- CV: 10-fold cross validation on balanced data set
- Dist: a data set with significant class skew

- **Key take away: CV and balanced data sets can give overly optimistic expectations**
- **We should expect significantly lower precision**

	Goodware	
	Distributed	Sorted
Training	11757	13265
Testing	4728	3220
	Malware	
Training	11091	9092
Testing	45	2044

Alg	Data	CAA	Acc	MPr	MRe
Hist+RF	Sort	95.3	94.7	0.953	0.926
	CV	**96.3**	96.0	**0.965**	0.942
	Dist	95.9	**97.3**	0.187	**1.000**
CNN	Sort	94.0	93.2	0.946	0.896
	CV	95.5	95.1	**0.959**	0.928
	Dist	**97.0**	**98.5**	0.242	**1.000**
LSTM	Sort	91.3	90.0	**0.926**	0.843
	CV	90.9	90.0	0.850	0.919
	Dist	**92.4**	**94.0**	0.107	**0.956**
CNN+LSTM	Sort	94.5	93.7	**0.956**	0.901
	CV	94.8	94.2	0.955	0.914
	Dist	**95.0**	**96.4**	0.157	**0.978**
LSM	Sort	90.7	89.8	0.856	0.856
	CV	**93.1**	92.6	**0.926**	0.901
	Dist	91.3	**95.6**	0.098	**1.000**

Case Study: Registry Keys

- Detect malware based on registry keys
  - Benign was collected over a 2 yr-period
  - 20 million entries, over 136,000 unique tuples
  - Malicious from known malware: 200 registry entries
  - Challenge #1: How to represent the data

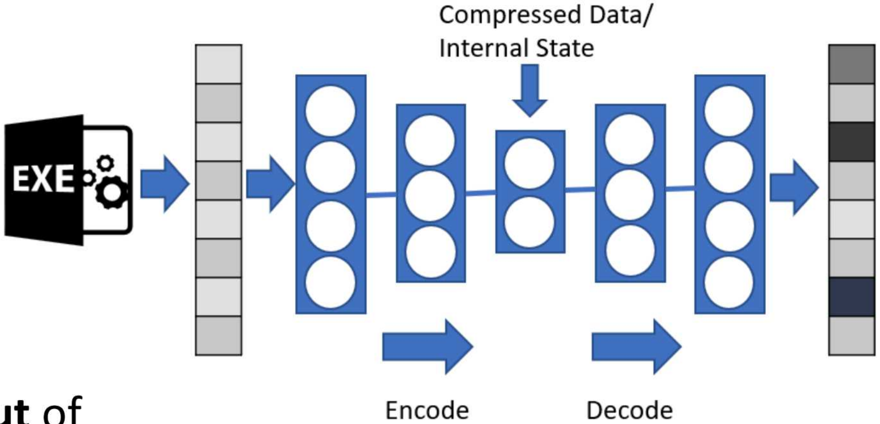
```
\HKEY\_LOCAL\_MACHINE\System\...\...\ImagePath
C:\Windows\System32\svchost.exe -k netsvcs
```

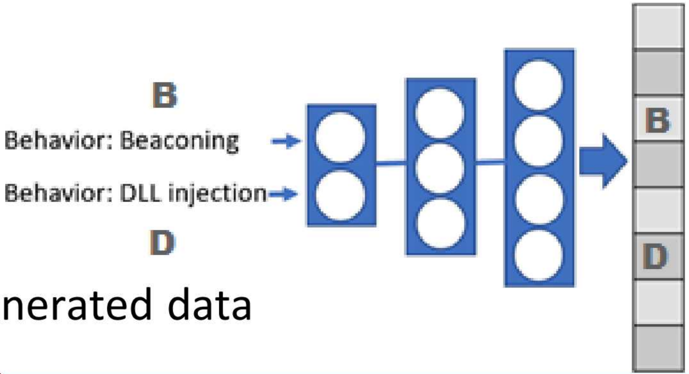
    - Each key is 1-hot encoded
    - Bag of words: 0ver 12,000 terms; down-sampled using PCA.
  - Challenge #2: How to label the data
    - Majority of hosts are benign; Modified by malware as malicious AUC 99%
      - Malicious registries came from specific hosts
    - Modified by malware as malicious; all other as benign AUC 96%

LOOKS LIKE A CASE OF OVERFITTING

Malware Generation (MalGen)

- Data driven approach to bridge semantic gap:
  - Focus on behaviors/functional characteristics—aligning closer with MA
  - Generate samples with specified behaviors/functional characteristics
  - Provide the opportunity to detect 0-days by focusing on an abstraction that is common amongst malware
- Create novel malware with **specified functional characteristics**
  - **Change labeling** from goodware/ malware to functional characteristics
  - **Mapping** output characteristics with internal variables
  - **More interpretable features** that may be more difficult to subvert
  - Use this mapping to **control the output** of the generated model
  - Sample the latent variable space to **create diverse set of exemplars**
    - **Discover unknown unknowns**
  - **Train** on ML-based malware detector on generated data





Behavioral Data

- Produced behavioral labeling for the malware (MS Malware)
  - Used the Malware Behavior Catalog (MBC) from Mitre as a taxonomy
    - Maps to the ATT&CK framework
    - Hierarchical representation using Objectives and Techniques

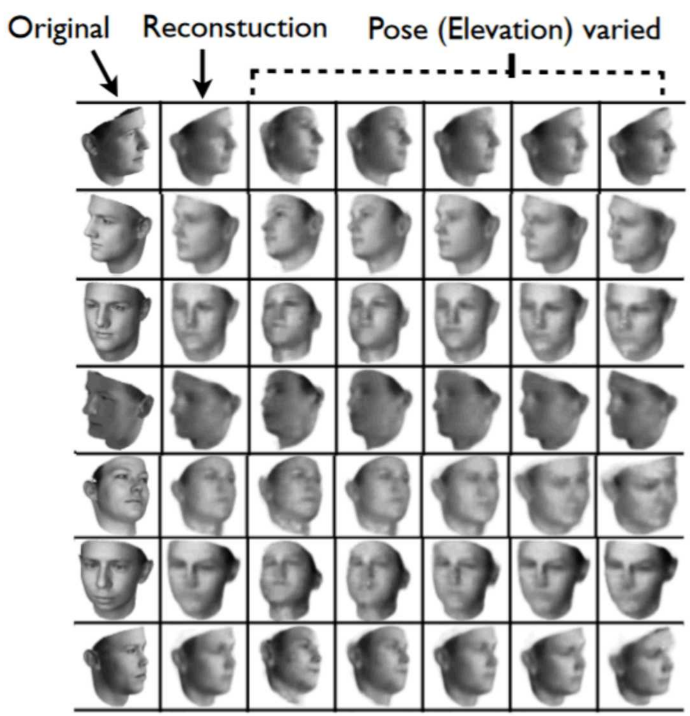
Objective:	Collection		Credential Access				Defense Evasion			...
Technique:	Local System	Man in the Browser	Hooking	Steal Web Session	Credential in Web Browser	Credentials in Files	Masquerading	Disable Sec Tools	Process Injection	...
Gatak	x	-	x	-	-	-	x	-	x	...
Ramnit	x	x	x	x	x	-	-	x	x	...
Lollipop	x	-	-	-	-	-	-	-	-	...
Kelihos	x	-	-	-	-	-	-	-	-	...
Vundo	x	-	-	-	-	x	x	x	x	...
Simda	x	-	-	-	-	-	x	x	-	...
Tracur	-	-	-	-	-	-	-	-	-	...

  - Manual process of gathering threat reports and mapping reports to MBC Objectives and Techniques
  - Aligns better with MA: changing from intent analysis to behavioral identification

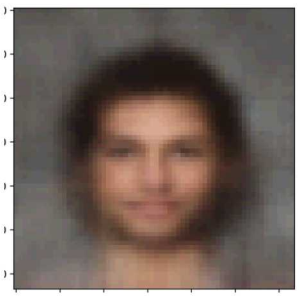
- M.R. Smith, N.T. Johnson, J.B. Ingram, A.J. Carbajal, B.I. Haus, R. Ramyaa, E. Domschot, C.C. Lamb, S.J. Verzi, W.P. Kegelmeyer. **Mind the Gap: On Bridging the Semantic Gap between Machine Learning and Information Security.** Submitted to AISEC 2020

Controlled Generative Models

- Implemented two generative models in the image domain:
  - CSVAE: Conditional Subspace Variational Autoencoder
    - Condition the hidden variables on the attribute
  - DC-IGN: Deep Convolutional Inverse Graphics Network
    - Clamp features during error propagation
- Starting porting over to cyber data



Smiling



5 o’clock shadow

Heavy Makeup

Eye Glasses

Bangs

How Wide is the Gap?

- MalGen has helped identify the gap between MA and ML
  - Not unique to MA—is common on many domains; ML should be carefully applied in novel applications
- Provided behavioral labelling
  - Modified the problem from identifying intent to identifying behaviors
  - Motivate others to look at the problem from a more MA-centric view
- Preliminary steps in generating synthetic data
- Still several unanswered questions:
  - What is the best feature representation?
  - What algorithms are best suited for MA?
  - What innovations still need to be made?
  - What is the appropriate abstraction level

Thank you

- Please reach out with questions and/or comments:
  - wg-malgen@sandia.gov