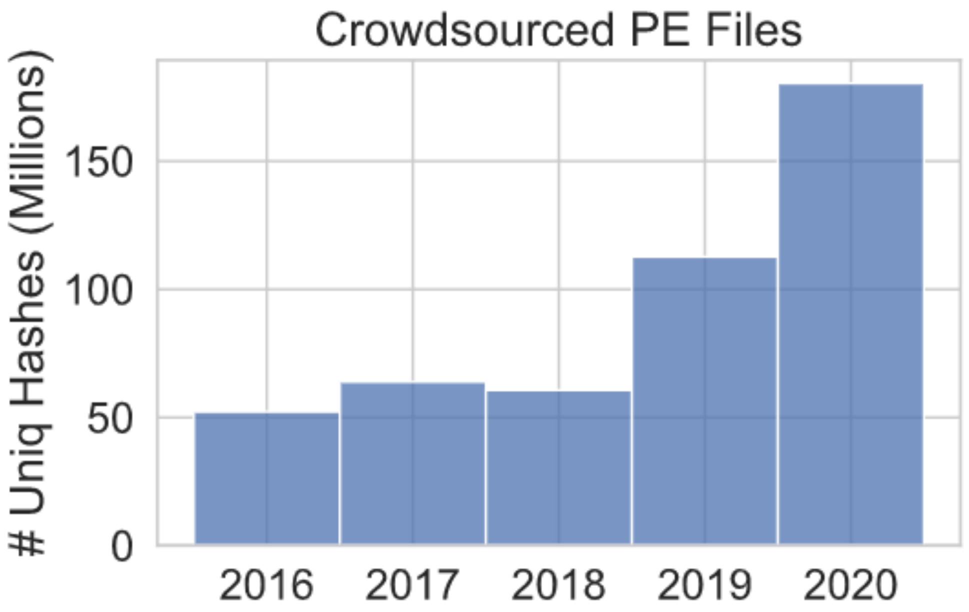
Poisoning static malware classification

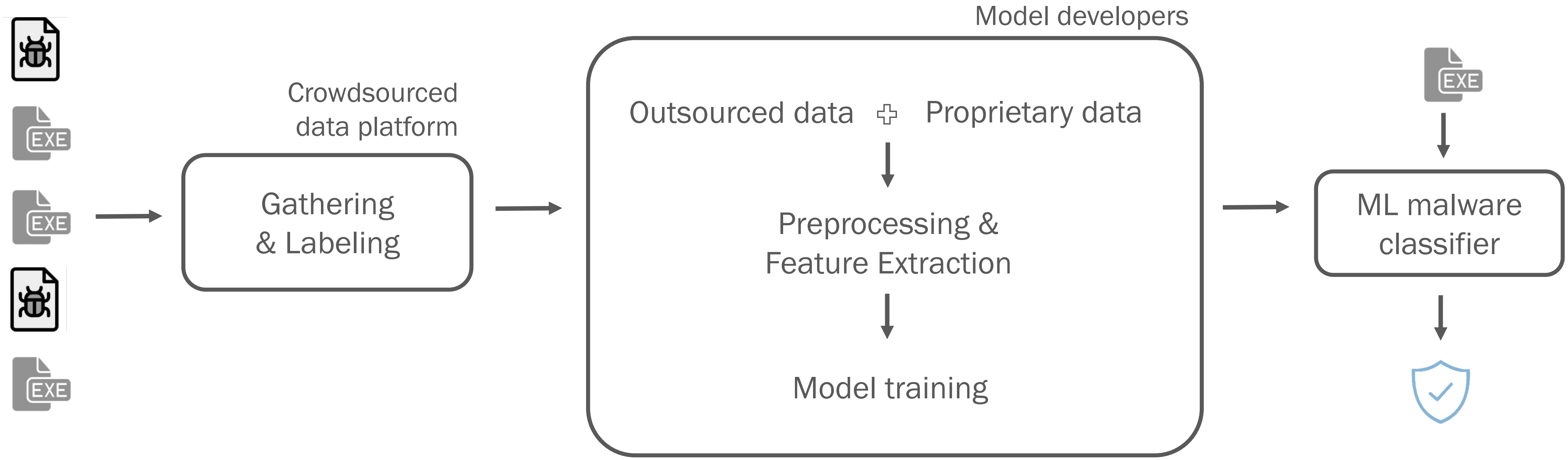
Georgio Severi, J. Meyer, S. Coull, and A. Oprea, "Explanation-Guided backdoor poisoning attacks against malware classifiers", USENIX Security, 2021.

Artworks generated with [Midjourney]

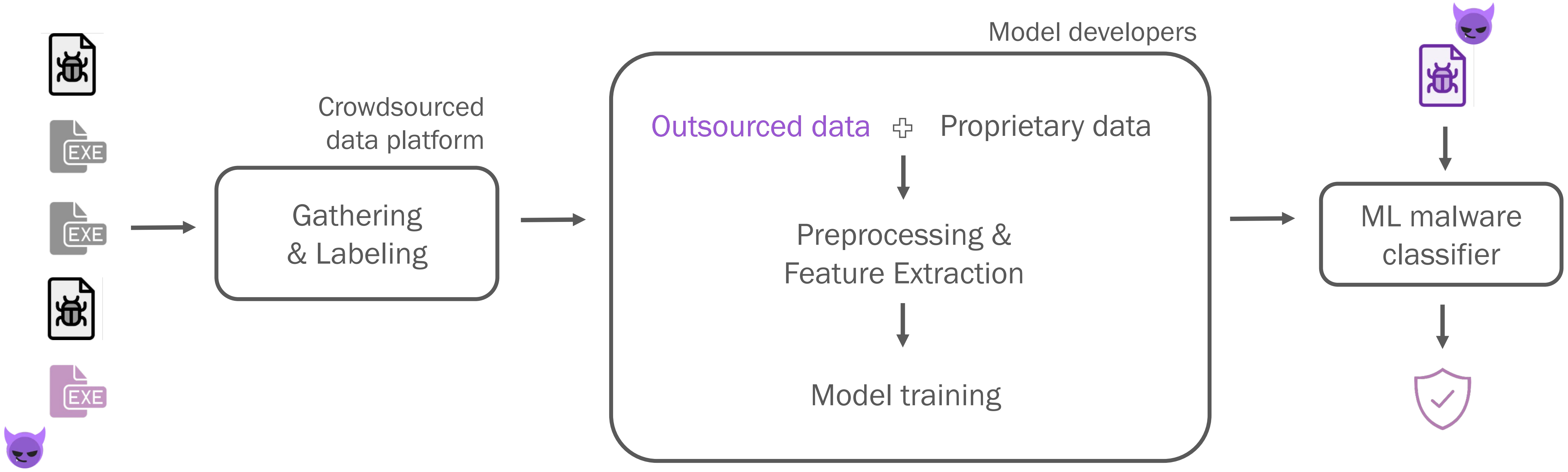
ML for malware detection

- Static analysis ML models play key role in pre-execution malware prevention
- Volume and diversity of executables makes training challenging
- Crowdsourced threat feeds provide an ideal source for training data





System overview



System overview

Backdoor attacks in ML

- Introduced by Gu et al. [3]
  - Descendant of “Red Herring” attacks [4]
- The training data is altered to induce the model to associate a pattern (trigger) with a target class
- Also referred to as Trojaning attacks [5] (model poisoning)

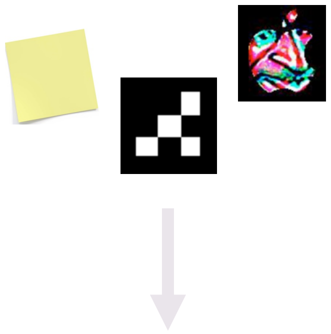
From [3]

[3] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain”, arXiv 2017.
[4] J. Newsome, B. Karp, and D. Song, "Paragraph: Thwarting signature learning by training maliciously”, RAID 2006.
[5] Y. Liu, S. Ma, Y. Aafer, W. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks”, NDSS 2018.

Backdoor attacks in malware classification

- The trigger pattern is mapped to a selection of features and values

- Attacker has no control over training labels - Clean-label [6, 7]
- Must respect the constraints dictated by the data semantics



Feature	LightGBM	EmberNN
major_image_version	1704	14
major_linker_version	15	13
major_operating_system_version	38078	8
minor_image_version	1506	12
minor_linker_version	15	6
minor_operating_system_version	5	4
minor_subsystem_version	5	20

[6] A. Turner, D. Tsipras, and A. Madry. "Clean-label backdoor attacks" 2018.
[7] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks", NeurIPS 2018.

Challenges and intuition

- How to select effective feature-value assignments for the trigger?
  1. Unique and easy to memorize assignments
  2. Leverage existing latent space areas associated with the benign class
- Our method needs to be model agnostic
  - We cannot assume the victim model will be a neural network (as in vision/NLP)

- Use model explanation methods (XAI) to guide the generation of the trigger
  - Obtain an intuition of how each feature-value assignment contributes to the model's output
- Adversarial ML researchers recently started using XAI methods
  - For evasion attacks [8, 9]
  - And defenses [10]

[8] A. Ignatiev, N. Narodytska, and J. Marques-Silva, "On relating explanations and adversarial examples", NeurIPS 2019.
[9] G.S, W. Pearce, and A. Oprea, "Bad Citrus: Reducing Adversarial Costs with Model Distances", ICMLA 2022.
[10] G. Tao, S. Ma, Y. Liu, and X. Zhang, "Attacks meet interpretability: Attribute-steered detection of adversarial samples", NeurIPS 2018.

Using model explanations

SHapley Additive exPlanations (SHAP) [11]

[11] S. M. Lundberg, and S. Lee, "A unified approach to interpreting model predictions", NeurIPS 2017.

Shapley Values for feature i

Weighting -- based on the size of coalition

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

Contribution of the feature to the payoff of the coalition

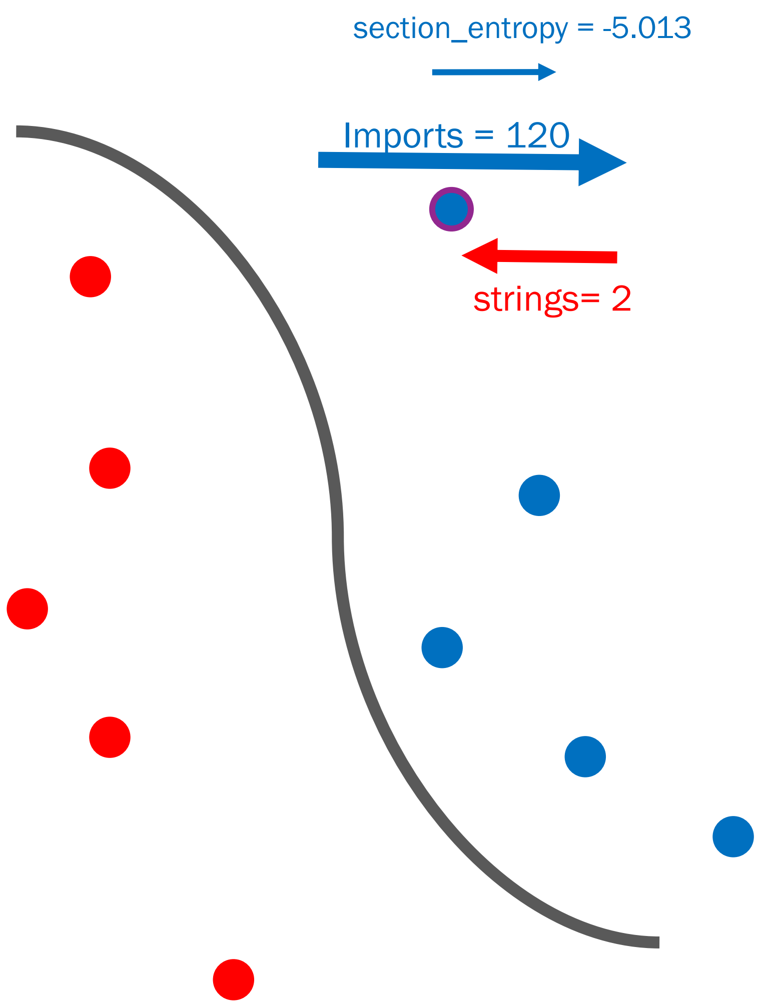
- Sample a coalition of features
- Measure the output with the target feature
- Measure the output without the target feature
- Compute the delta
- Repeat for all possible coalitions and average

Intuition

Using model explanations

SHapley Additive exPlanations (SHAP) [11]

- Model agnostic framework
- Local interpretability
  - Estimate influence of feature-value assignments on model decisions
- Global interpretability
  - Aggregate SHAP values over all the points for each feature
  - Provides intuition on feature importance and direction



[11] S. M. Lundberg, and S. Lee, "A unified approach to interpreting model predictions", NeurIPS 2017.

Backdoor design strategies

Independent

Independently selects high-leverage features and uncommon/weakly-aligned values

- Stronger trigger memorization
- Identifiable points

Combined

Greedily selects coherent combinations of features and values aligned with target class

- Backdoor points are close to real data
- Stealthier

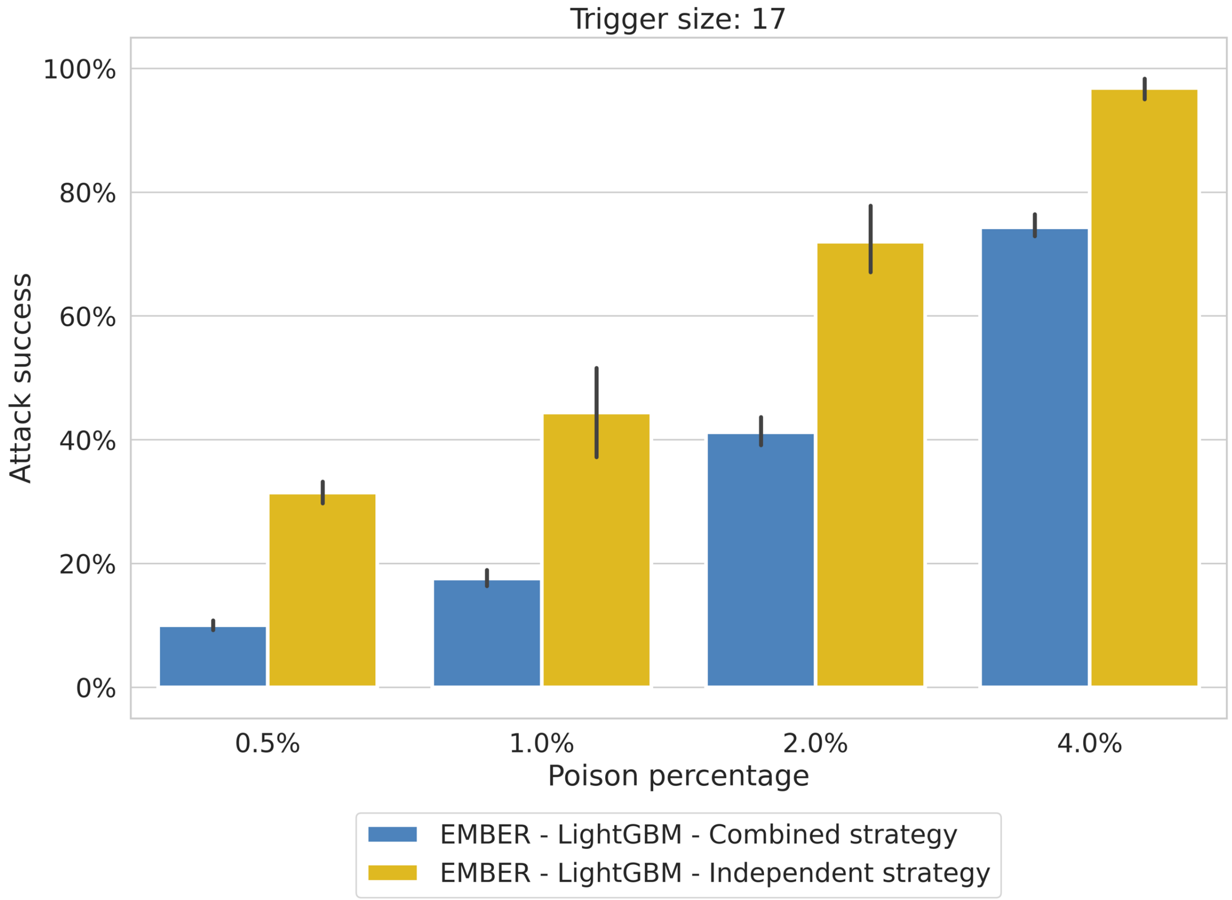
Evaluation setup

Dataset	Size	Type	Models	Approach
EMBER [12]	800k samples 2351 features	Windows PE	LightGBM, DNN	Developed a specific backdooring utility
Drebin [13]	128k samples 545k features	Android APK	Linear SVM	Restricted modifications to manifest file
Contagio [14]	10k samples 135 features	PDF	Random Forest	Restricted modifications as in Šrndić et al. 2014

[12] H. S. Anderson, and P. Roth, "Ember: an open dataset for training static pe malware machine learning models", arXiv 2018.
[13] D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon, and K. Rieck, "Drebin: Effective and explainable detection of android malware in your pocket", NDSS 2014.
[14] N. Šrndić and P. Laskov. "Practical evasion of a learning-based classifier: A case study." In 2014 IEEE symposium on security and privacy, pp. 197-211. IEEE, 2014.

Results on PE files

- Significant damage at 1% poison rate and 17 manipulated features
- Attack success scales with poisoning rate and trigger size
- Minimal side effect on victim's generalization capability
- Similar results for the Neural Network



Different file types

- Drebin (Android APK):
  - Around 40% success at 1% poisoning rate and 30 features
  - Importance estimation on surrogate model
- Contagio (PDF):
  - 75% success at 1% poisoning rate with 30 features
  - Higher variance due to dataset size

About mitigations

- We adapted different approaches from computer vision:
  - Spectral signatures [15]
  - Activation clustering [16]
  - Isolation Forests [17]
- No tested defense found all backdoors consistently
- Backdoors generated by the combined strategy are hard to identify

[15] B. Tran, J. Li, and A. Madry. "Spectral signatures in backdoor attacks," NeurIPS 2018.
[16] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, and B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering", arXiv 2018.
[17] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation forest", ICDM 2008.

Takeaways

- Benign binaries can be used as carriers for poisoning attacks
- Model interpretability methods can be leveraged to guide the backdoor generation
  - This approach is model-agnostic and applies to multiple data modalities
- A sophisticated adversary can generate stealthy backdoors