RESEARCH STATEMENT

Fnu Suya

Machine learning (ML) models are often developed by training on vast datasets or fine-tuning pretrained (foundation) models [Bommasani et al., 2021]. However, the reliance on data from potentially untrusted sources, like unauthenticated internet-crawled data, leaves these models susceptible to data poisoning attacks [Carlini et al., 2023]. Additionally, platforms like HuggingFace, which accept uploads from individual users, further increase the risk by potentially spreading manipulated pretrained models, thereby escalating the threat of poisoning attacks in downstream applications. Poisoning attacks are currently recognized as one of the top concerns in the industry [Kumar et al., 2020].

My research focuses on evaluating the trustworthiness of machine learning in environments subject to malicious training influences, particularly concerning contaminated training data and manipulated pretrained models. In the first scenario, I have investigated conventional data poisoning attacks, identifying how varying distributional properties can influence attack effectiveness and the potential for creating stronger defenses based on these insights. In the second scenario, I revealed that even subtle modifications to pretrained models could significantly heighten security and privacy risks for downstream tasks, underscoring the need for novel defense mechanisms. Looking ahead, my research will pursue two main directions: (1) in standard non-security domains like vision and natural language processing (NLP), I aim to explore different aspects of the trustworthiness (e.g., fairness, interpretability) of machine learning models in compromised training environments, propose solutions, and investigate the use of poisoning for positive ends; (2) in the realm of security-critical applications, I plan to leverage recent machine learning advances to bolster model robustness and scrutinize the trustworthiness of models in these critical domains.

1 Exploring the Limits of Data Poisoning Attacks

Traditional data poisoning attacks, involving the insertion of a small fraction of poisoning points into a victim's training set, began their research evolution a decade ago. Early research focused on the indiscriminate goal of reducing the accuracy of some simple (e.g., linear) models [Biggio et al., 2012], while recent research has shifted to targeted attacks on modern deep learning models with the aim of misclassifying specific test samples [Shafahi et al., 2018]. However, these widely-studied attack types represent extremes and do not align well with practical adversarial goals. A more realistic objective in practice is to induce misclassification within a specific sub-distribution while maintaining high accuracy overall [Jagielski et al., 2021], a task that is both broader than targeted poisoning and stealthier than indiscriminate attacks. Furthermore, while current research, based on a strong threat model with full access to victim data and models, demonstrates success in specific instances as highlighted by Steinhardt et al. [2017] and Geiping et al. [2021], it notably falls short in more challenging scenarios. These include misclassifying multiple test samples in larger models [Geiping et al., 2021] or effectively targeting datasets fortified with defense mechanisms [Steinhardt et al., 2017]. This disparity highlights a significant gap in the current understanding of data poisoning attacks—a gap my research aims to bridge.

Through a series of publications [Suya et al., 2021, Rose et al., 2022, Suya et al., 2023a], I explore the limits of data poisoning attacks under the strong threat model considered in existing literature, focusing on the indiscriminate poisoning and the less explored yet practically relevant subpopulation poisoning settings. These are more challenging to achieve than targeted misclassifications.

Model-targeted poisoning attacks. In Suya et al. [2021], we introduce an empirical model-targeted poisoning (MTP) attack, establishing state-of-the-art performance and acting as a lower bound on poisoning effectiveness. The MTP attack proceeds in two steps: first, generate a target model that encodes with the attacker's objectives, achievable through techniques like label flipping on a significant portion of training data. Then, select (a limited number of) poisoning points to ensure that the model trained on this poisoned data asymptotically converges to the target model, whose convergence is rigorously proved. Empirically, our results validate the success of this attack in inducing target models and achieving a range of objectives, particularly in subpopulation contexts which are more relevant in practice. Moreover, the attack's provable convergence property enables practical implications in various areas, including manipulating other trustworthy

aspects of machine learning, like privacy and fairness, and also offers potential for constructive applications, such as rectifying flawed models with carefully designed benign "poisoning" samples.

Understanding impact of subpopulation properties on susceptibility. In a subsequent study Rose et al. [2022], I supervised an undergraduate student who became the first author on the paper, to conduct extensive tests on large number of subpopulations from both synthetic and benchmark datasets. The finding is, for a given learner and dataset, some subpopulations are notably more resistant to current poisoning attacks, including against our state-of-the-art MTP attack [Suya et al., 2021]. This variability in susceptibility is linked to the "relative position" of each subpopulation in relation to others under the specified learner. This "relative position" is quantifiably reflected by the minimum loss difference between the target model, which misclassifies the subpopulation, and the clean model. A smaller loss difference suggests a higher likelihood of effectively misclassifying the subpopulation with few poisoning points (i.e., vulnerable subpopulation), and vice versa. Additionally, we observed that vulnerability variations across subpopulations are less pronounced in datasets with poor class separation (with most subpopulations being vulnerable) whereas they are significantly more evident in well-separated datasets. This indicates that both data properties and subpopulation characteristics influence susceptibility. Particularly in well-separated datasets, subpopulations with larger loss differences exhibit greater robustness against current poisoning attacks.

Understanding impact of distributional properties on susceptibility. In Suya et al. [2023a], my research shifts to examining indiscriminate poisoning attacks aiming to gain a deeper understanding of how learning algorithms are impacted by data poisoning. Motivated by the observed variation in subpopulation susceptibility, we investigated if a similar variability exists across different benchmark datasets when subjected to indiscriminate attacks as different datasets can be conceptually viewed as distinct subpopulations. The experiments with various benchmark datasets under linear models, using existing attacks including our state-of-the-art MTP attack, reveal significantly different levels of vulnerability to these attacks—some datasets are robust to all known attacks even without any defenses. I then probed whether the resistance of some datasets to current attacks is due to the suboptimality of these attacks or if it stems from an inherent robustness rooted in the datasets. Through theoretical analysis, I showed data distributions can inherently resist any poisoning attack if the distributions are well-separated with low variance, and the permissible set of poisoning points is limited in size. By empirically calculating these metrics for different benchmark datasets, I found a strong correlation between these metrics and the varying effectiveness of current attacks. Additionally, based on the metrics, I provided a non-trivial upper bound on the effectiveness of optimal poisoning attacks for general distributions and empirically computed the upper bounds for benchmark datasets. The empirical findings show that these upper bounds also differ markedly across datasets, further validating the theoretical and empirical insights.

My research highlights the pivotal influence of data distributions on a model's resilience to poisoning attacks, pointing to the improvement of distributional quality as a novel approach for crafting more robust defenses. To illustrate this, I show that in image classification tasks, the use of advanced feature extractors, such as deep model architectures trained over extended epochs, significantly improves the robustness of downstream classifiers against poisoning attacks. This is particularly relevant in the current era of foundation models, which emphasizes the importance of finding high-quality pretrained feature extractors to boost overall model performance.

2 Manipulating pretrained Models

Traditional data poisoning involves training a model from scratch, which may not be feasible when the training data size or the computational resources are limited. To circumvent this, practitioners often perform lightweight customization on powerful pretrained models for different downstream applications. We have studied security issues for two popular ways of customizing pretrained models. The customization can be done through *model compression*, which reduces the size of a pretrained model for resource-limited downstream applications, or through *fine-tuning*, where the pretrained model is used as a feature extractor with added classification layers, tailored for applications with limited data and resources.

Model compression. In the realm of model compression, we demonstrated a new threat where an adversary can stealthily embed backdoors into models during the training process. These backdoors have no effect in the uncompressed (pretrained) model but activate to perform maliciously when the model is compressed by the victim [Tian et al., 2021]. Such a goal is achieved by designing a customized loss function for training the pretrained model, ensuring that the uncompressed model performs normally with both clean and backdoor samples, while the compressed model, when simulated, induces targeted misclassifications on the backdoor samples. These backdoors evade detection by the state-of-the-art backdoor detection tools when applied to the uncompressed models, typically maintained by resource-rich model zoo owners who might conduct integrity checks. Since the backdoor is only effective in the compressed model, it would only be detected if detection methods are run on the compressed model—a scenario unlikely for downstream

users with limited resources. This work underscores the importance of considering machine learning applications as a comprehensive pipeline and highlights the necessity of ensuring reliability at each stage. Relying solely on conventional attack surface checks, such as on full-size pretrained models, may not sufficiently safeguard the system.

Fine-tuning. In our study of model acquisition via fine-tuning [Tian et al., 2021], we demonstrate that the fine-tuning process in transfer learning can be exploited to markedly increase the privacy risks of downstream models. Specifically, attackers can modify the training of the upstream model, causing selected neurons to react distinctly based on whether the downstream training set includes certain sensitive properties (e.g., *any* images of a specific individual). This alteration in neuronal behavior significantly escalates the privacy risks for training sets containing such targeted properties. These manipulated pretrained models have different goals than traditional poisoned models, which primarily induce misclassifications. Consequently, even with non-trivial adaptations, existing anomaly detection methods remain ineffective against them. This suggests that for poisoned models with non-conventional adversarial objectives, the corresponding defenses also need to be carefully designed, as adaptation from defenses in conventional attack settings may not help.

3 Other Work on Adversarial Examples

Beyond my primary research on poisoning attacks, I have also studied adversarial examples that cause misclassifications in poison-free, well-trained models through imperceptible perturbations on test inputs. Focusing on the practicality of black-box attacks in the vision domain, where only API access is available to attackers, my earlier work [Suya et al., 2020] introduces a hybrid strategy that combines different attacks in a novel way and a batch attack strategy targeting vulnerable candidates first, mirroring realistic adversary tactics within a limited query scope. Additionally, in Suya et al. [2023b], I lead the undergraduate researchers and offer a systemization of knowledge of black-box attacks, proposing a new threat model-based taxonomy for better future evaluations and addressing current methodological flaws from the perspective of real-world adversarial considerations. In the graph domain, to address the scalability issues of current white-box attacks, we develop efficient attacks by strategically ignoring unnecessary node degree changes in computation and enable the robustness evaluations on extremely large graphs for the first time [Wang et al., 2020]. At a high level, my work with adversarial examples has enhanced my understanding of the inherent limitations in well-trained, poison-free deep learning models in adversarial settings. This insight is driving my further exploration into how these vulnerabilities are exacerbated under poisoning attacks.

4 Future Plans

Investigations in security-critical domains. My research has primarily focused on non-security domains like vision and NLP that are commonly explored. However, I am now shifting towards security-critical domains such as malware detection [Chang and Im, 2020]. These fields present unique challenges due to their inherent adversarial nature, making the development of reliable models particularly complex. Unlike non-security domains, the adversarial pressures in these areas are intrinsic and not merely additional attack threats. My goal is to develop models robust enough to withstand these inherent pressures. For instance, the varied and dynamic nature of malware behaviors in the wild [Avllazagaj et al., 2021] poses significant challenges to traditional dynamic analysis methods like sandbox execution. To address these challenges, I plan to apply advances from the broader ML field, including the development and adaptation of foundation models [Vaswani et al., 2017], proven successful in benign software analysis [Pei et al., 2020], to the specific requirements of these critical domains.

Building on more reliable models, my next step is to evaluate their robustness against typical adversarial threats such as poisoning attacks and adversarial examples. This evaluation will draw on my previous research findings, such as the limitations of data poisoning attacks. However, given the unique nature of security-critical domains, adapting these techniques will require bespoke approaches. For example, strategies that work in vision domains, like ensembles of local surrogate models for transfer attacks, may not be effective in malware evasion due to distinct vulnerability spaces across model architectures in malware domain. Alongside my focus on security-critical domains, I will keep investigating in general domains like vision and NLP, pursuing concrete research directions as outlined below.

Characterizing the limits of poisoning from learner's perspective. My recent work [Suya et al., 2023a] has focused on analyzing how *data distributions* influence the effectiveness of poisoning attacks, gaining new insights into designing improved defenses through better feature representations. My future research will continue to probe the limits of data poisoning attacks, specifically examining the impact of *different learners* (e.g., various neural network architectures) on attack efficacy. Intriguing observations have been made, such as neural networks, even simple multi-layer perceptrons, displaying greater robustness than linear models for the same datasets [Lu et al., 2023]. Another finding is that deep

neural networks are highly susceptible to targeted poisoning attacks but show substantial resistance to indiscriminate attacks, suggesting that deep models can be "nudged" locally, a phenomenon less apparent in simpler linear models. At a broader level, these inquiries delve into the fundamental properties of learning algorithms, shedding light on the varying effectiveness of poisoning attacks in different contexts and suggesting ways to enhance model reliability through architectural improvements.

Trustworthy machine learning in malicious training environments. In addition to inducing misclassifications through poisoning, I will also explore how other trustworthy aspects like interpretability, fairness, and privacy are compromised in malicious training environments. Focusing on a single attack type does not fully capture the complex risks in real-world model deployments, which are often exposed to diverse attacks. My future work involves developing new poisoning strategies within realistic threat models to assess these risks and understand their limits. Our work on manipulating pretrained models [Tian et al., 2023] demonstrates how contaminated data or models can significantly increase the privacy risks of downstream tasks. Furthermore, the MTP attack [Suya et al., 2021] also shows the potential for altering fairness or privacy outcomes by creating suitable target models.

Evaluating risks of machine learning within a complete pipeline. In practice, machine learning models often function as components within larger pipelines [Marcus et al., 2021], making an end-to-end risk analysis of the ML models upon deployment crucial [Debenedetti et al., 2023]. Initial investigations of ML models as standalone units provide vital insights into their vulnerabilities in adversarial environments, helping to identify key solutions without the interference of other components. With a solid understanding of these standalone risks, my future research will examine the challenges faced by ML models integrated into larger pipelines, such as performance degradation due to hardware fault attacks [Hong et al., 2019]. As an initial step, our prior work on using model compression to insert stealthy backdoors [Tian et al., 2021] underscores the need for comprehensive security checks throughout the full ML pipeline.

Designing defenses. As a security researcher, my approach involves developing various attack techniques to ultimately inform and strengthen defense systems. Designing defenses against less explored attack objectives, such as privacy and fairness, requires intentional effort. Simply adapting existing methods for traditional poisoning objectives is insufficient, as evidenced by our work on increasing property inference risks with manipulated pretrained models [Tian et al., 2023]. I believe that designing effective poisoning attacks or understanding their limits on different trustworthy aspects offers a unique perspective on the mechanisms of these attacks. This understanding is crucial for developing (provably) robust defenses, a concept illustrated in our recent work [Suya et al., 2023a].

Poisoning Attacks for Good. Poisoning attacks, while typically associated with malicious intent, can also serve beneficial purposes [Shan et al., 2023], opening doors to exciting interdisciplinary research. For example, they can be used to benchmark machine learning powered systems under the worst-case, naturally occurring (with low probability) scenarios that can not be easily captured during regular testing. This approach allows system designers to identify and address corner cases efficiently, and then optimize system design. My current project is focused on benchmarking the performance of machine learning query optimizers in database systems [Marcus et al., 2021, Yang et al., 2022]. Interestingly, we have found that certain combinations of natural queries, when used in model (re)training, can cause performance regression compared to traditional, non-machine learning based optimizers. This benchmarking concept is also applicable to other fields like bioinformatics that increasingly rely on ML models. Beyond performance evaluation, poisoning can rectify problematic models, such as transforming biased pretrained models into unbiased ones with specially designed fine-tuning data [Wang and Russakovsky, 2023], or improving underperforming models with benign "poisoning" points [Cadamuro et al., 2016]. Overall, poisoning attacks shed light on model performance changes in response to training distribution modifications, a concept that can be leveraged for positive outcomes.

References

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023.

Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. Adversarial machine learning-industry perspectives. In *2020 IEEE Security and Privacy Workshops (SPW)*, pages 69–75. IEEE, 2020.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. *International Conference on Learning Representation (ICML)*, 2012.

Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Matthew Jagielski, Giorgio Severi, Niklas Pousette Harger, and Alina Oprea. Subpopulation data poisoning attacks. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2021.

Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. *Advances in Neural Information Processing Systems*, 2017.

Jonas Geiping, Liam Fowl, W Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching. In *International Conference on Learning Representations*, 2021.

Fnu Suya, Saeed Mahloujifar, Anshuman Suri, David Evans, and Yuan Tian. Model-targeted poisoning attacks with provable convergence. In *International Conference on Machine Learning (ICML)*, 2021.

Evan Rose, Fnu Suya, and David Evans. Poisoning attacks and subpopulation susceptibility. In *5th Workshop on Visualization for AI Explainability (VISxAI)*, 2022.

Fnu Suya, Zhang Xiao, Yuan Tian, and David Evans. What distributions are robust to indiscriminate poisoning attacks for linear learners? *Advances in Neural Information Processing Systems*, 2023a.

Yulong Tian, Fnu Suya, Fengyuan Xu, and David Evans. Stealthy backdoors as compression artifacts. *IEEE Transactions on Information Forensics and Security (TIFS)*, 2021.

Fnu Suya, Jianfeng Chi, David Evans, and Yuan Tian. Hybrid batch attacks: Finding black-box adversarial examples with limited queries. In *USENIX Security*, 2020.

Fnu Suya, Anshuman Suri, Tingwei Zhang, Scott Hong, Yuan Tian, and David Evans. Sok: What have we learned about black-box attacks against classifiers? *Under Submission*, 2023b.

Jihong Wang, Minnan Luo, Fnu Suya, Jundong Li, Zijiang Yang, and Qinghua Zheng. Scalable attack on graph data by injecting vicious nodes. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 2020.

Jun Young Chang and Eul Gyu Im. Data poisoning attack on random forest classification model. *International Conference on Smart Media and Applications*, 2020.

Erin Avllazagaj, Ziyun Zhu, Leyla Bilge, Davide Balzarotti, and Tudor Dumitraș. When malware changed its mind: an empirical study of variable program behaviors in the real world. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 3487–3504, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Kexin Pei, Zhou Xuan, Junfeng Yang, Suman Jana, and Baishakhi Ray. Trex: Learning execution semantics from micro-traces for binary similarity. *arXiv preprint arXiv:2012.08680*, 2020.

Yiwei Lu, Gautam Kamath, and Yaoliang Yu. Exploring the limits of model-targeted indiscriminate data poisoning attacks. 2023.

Yulong Tian, Fnu Suya, Anshuman Suri, Fengyuan Xu, and David Evans. Manipulating transfer learning for property inference. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Ryan Marcus, Parimarjan Negi, Hongzi Mao, Nesime Tatbul, Mohammad Alizadeh, and Tim Kraska. Bao: Making learned query optimization practical. In *Proceedings of the 2021 International Conference on Management of Data*, pages 1275–1288, 2021.

Edoardo Debenedetti, Giorgio Severi, Nicholas Carlini, Christopher A Choquette-Choo, Matthew Jagielski, Milad Nasr, Eric Wallace, and Florian Tramèr. Privacy side channels in machine learning systems. *arXiv preprint arXiv:2309.05610*, 2023.

Sanghyun Hong, Pietro Frigo, Yiğitcan Kaya, Cristiano Giuffrida, and Tudor Dumitraș. Terminal brain damage: Exposing the graceless degradation in deep neural networks under hardware fault attacks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 497–514, 2019.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. *arXiv preprint arXiv:2302.04222*, 2023.

Zongheng Yang, Wei-Lin Chiang, Sifei Luan, Gautam Mittal, Michael Luo, and Ion Stoica. Balsa: Learning a query optimizer without expert demonstrations. In *Proceedings of the 2022 International Conference on Management of Data*, pages 931–944, 2022.

Angelina Wang and Olga Russakovsky. Overwriting pretrained bias with finetuning data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3957–3968, 2023.

Gabriel Cadamuro, Ran Gilad-Bachrach, and Xiaojin Zhu. Debugging machine learning models. In *ICML Workshop on Reliable Machine Learning in the Wild*, volume 103, 2016.