RESEARCH STATEMENT

Fnu Suya

In the digital era where data is more valuable than oil[1], machine learning (ML) models, reliant on large datasets, often sourced from potentially untrusted origins like unverified internet data, are increasingly vulnerable to data poisoning attacks [1] and becomes a significant industry challenge [2]. My main research focus is on exploring the limits and impacts of these attacks in trustworthiness of machine learning upon deployment. Below, I first describe my PhD research focused on the limits of data poisoning attacks, followed by a discussion of my future research directions.

Understanding the Limits of Data Poisoning Attacks

Data poisoning attacks, which inject malicious data into the victim's training set, have progressed from indiscriminate attacks impacting the accuracy of simple models [3] to targeted attacks causing misclassification in specific instances within deep learning models [4]. Yet, these attacks do not align with realistic adversarial goals, like subpopulation attacks that target specific sub-distributions, aiming to impact them without undermining the overall accuracy of the model. Moreover, despite notable successes in existing research, these methods encounter challenges in more complex scenarios, like simultaneously impacting multiple test samples. My research seeks to address the gap in understanding data poisoning attacks. Through a series of publications [5, 6, 7], I have explored the limits of data poisoning, focusing on both indiscriminate and subpopulation attacks, that are more challenging to achieve than targeted misclassifications.

Model-targeted poisoning attacks. In the first work [5], we developed a model-targeted poisoning (MTP) attack, establishing tighter lower bound on poisoning effectiveness. This attack involves creating a target model using techniques like label flipping, followed by strategically selecting poisoning points to efficiently approach the target with guaranteed convergence. Our results demonstrate its efficacy, especially in subpopulation contexts, and its potential to influence other aspects of trustworthy machine learning such as privacy and fairness. This method also suggests positive applications, such as improving flawed models with benign "poisoning" samples.

Impact of subpopulation properties on susceptibility. In a subsequent study [6], I supervised an undergraduate student, who was the lead author, in analyzing various subpopulations and found that susceptibility to poisoning attacks varies significantly among subpopulations. Our findings revealed that this variability correlates with the minimum loss difference between a target model that misclassifies the subpopulation, and the clean model. A larger loss difference suggests increased resistance of a subpopulation to effective poisoning attacks, and vice versa.

Impact of distributional properties on susceptibility. In Suya et al. [7], we examined the broad impact of indiscriminate poisoning attacks on learning algorithms by analyzing variations in dataset susceptibility. Our experiments across various datasets revealed differing levels of vulnerability and some showed strong robustness even without defenses. Theoretical analysis indicated that datasets with well-separated, low-variance distributions are inherently more resistant to poisoning, not just against current attacks, especially when the size of the set containing all valid poisoning points is limited. We found a strong correlation between these distributional properties and attack effectiveness observed on benchmark datasets, and also leads us to establish non-trivial upper bounds on the effectiveness of any poisoning attacks for general distributions, which again vary substantially across datasets. Finally, our findings indicate that dataset robustness can be increased with better feature representations. In image classification, we show that using better pretrained models as feature extractors can significantly enhance downstream resistance to poisoning, aligning with the trend of improving pretrained extractors in foundation models for improved overall model performance.

Other work on trustworthy machine learning. Beyond my primary research on data poisoning attacks, I have investigated vulnerabilities in settings where (potentially manipulated) pretrained models are customized for various downstream applications. This research underscores the urgent need for novel defense strategies against unconventional objectives, such as privacy breaches [8], and highlights the necessity of evaluating ML systems as complete pipelines [9]. Additionally, my work extends to adversarial examples, which exploit well-trained, poison-free models by introducing

[1] `https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data`

imperceptible perturbations on test inputs [10, 11, 12]. These studies have deepened my understanding of the intrinsic limitations of well-trained and poison-free deep learning models in adversarial contexts, informing my ongoing research into the amplified vulnerabilities presented by poisoning attacks.

Future Plans

My long-term objective is to develop real-world machine learning systems that are robust, fair, and private, particularly under adversarial conditions. My research will cover a wide spectrum, from well-studied non-security domains to security-critical areas. Both are crucial for a comprehensive understanding of the challenges in creating truly reliable machine learning applications. In the following sections, I will detail specific research directions in these domains, each contributing to the goal of creating trustworthy machine learning systems for varied real-world scenarios.

Investigations in security-critical domains. I have focused my research on non-security fields like vision and NLP. However, I'm now branching into security-critical domains such as malware detection, where the inherent adversarial nature calls for highly reliable models. These areas, with their intrinsic adversarial pressures, differ from non-security domains that typically face external threats. My goal is to develop models resilient to these inherent challenges. For instance, considering the evolving nature of malware which often eludes traditional detection methods based on sandbox execution, I plan to utilize recent advancements in machine learning, including foundation models, to tackle these specific challenges. This new direction in my research is exemplified by a recent research proposal I co-authored.

Expanding on my work with reliable models in security-critical areas, I plan to rigorously evaluate their robustness against common adversarial threats, including poisoning attacks and adversarial examples. This effort will leverage my existing research in trustworthy machine learning and venture into new areas. Given the unique challenges of security-critical domains, such as malware detection, tailored strategies are necessary. For example, successful techniques in vision, like ensembles of local surrogate models for transfer attacks, might not apply directly to malware evasion, given the differing vulnerabilities in model architectures. This highlights the need for customized solutions in these specific domains. My future research will thus encompass both non-security and security-critical fields, focusing on developing domain-adapted methodologies for creating reliable ML systems. The following sections detail these research directions, with the initial two focusing on assessing deployment risks from an attacker's perspective, and the latter two dedicated to building improved models, informed by insights gained from risk evaluations involving attacks.

Trustworthy ML in malicious environments. My future research will go beyond just causing misclassifications through poisoning; it will investigate how interpretability, fairness, and privacy are impacted in hostile training settings. Real-world models face multifaceted risks from various attack types, which single-attack-focused studies don't fully capture. I plan to develop new poisoning strategies under realistic threat models to thoroughly assess and delineate these risks. Our prior work on manipulating pretrained models [8] highlights how tainted data or models escalate privacy risks significantly. Additionally, the MTP attack [5] may affect fairness and privacy outcomes with suitable target models.

Risk evaluation in complete pipelines. My future research will focus on end-to-end risk analysis of ML models within larger pipelines, recognizing that ML models often operate as part of more extensive systems. Initial studies on standalone ML models provide crucial insights into vulnerabilities under adversarial conditions. Building on this understanding, I will explore how ML models fare when integrated into broader systems, such as examining vulnerabilities introduced by hardware hosting the models. My work on inserting stealthy backdoors through model compression [9] highlights the importance of thorough security assessments across the entire ML pipeline.

Defenses against diverse attacks. Designing defenses against underexplored attack objectives like privacy and fairness demands focused effort. Merely adapting methods meant for conventional poisoning objectives falls short, as shown in our study on heightened property inference risks in manipulated pretrained models [8]. Developing effective poisoning attacks and comprehending their impact on various trustworthy aspects provides insight into attack mechanisms, essential for creating robust defenses, a principle demonstrated in our recent work [7].

Poisoning attacks for good. Poisoning attacks, often seen as harmful, can be repurposed for beneficial interdisciplinary research. For example, they uniquely test machine learning systems against rare, difficult-to-simulate scenarios, aiding in identifying and mitigating system vulnerabilities. My current project applies this by evaluating machine learning-based query optimizers in database systems, finding that certain query combinations can regress performance, contrary to traditional, non-ML optimizers. This approach also applies to areas like bioinformatics, reliant on machine learning models. Additionally, strategic poisoning can rectify flawed models, like adjusting biases in pretrained models or boosting underperforming ones with targeted training data. These applications underscore poisoning's role in constructively adjusting model performance following training distribution changes.

References

[1] 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.

[2] 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)*, 2020.

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

[4] 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.

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

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

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

[8] Yulong Tian, Fnu Suya, Anshuman Suri, Fengyuan Xu, and David Evans. Manipulating transfer learning for property inference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

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

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

[11] 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*, 2020.

[12] Fnu Suya, Anshuman Suri, Tingwei Zhang, Jingtao Hong, Yuan Tian, and David Evans. Sok: Pitfalls in evaluating black-box attacks. *arXiv preprint arXiv:2310.17534*, 2023.