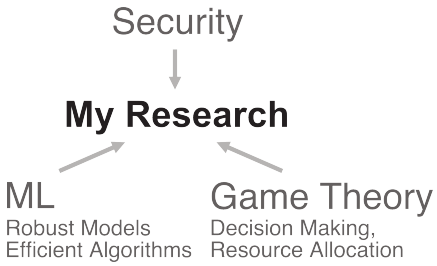
Shang-Tse Chen
Georgia Tech

Research Statement

I work in the intersection of applied and theoretical machine learning (ML), with a strong application focus on cybersecurity and social betterment [KDD16 Best Student Paper Runner-up]. I develop novel ML approaches that are not only theoretically-principled but also scalable and robust in practice. My simultaneous strengths in theory and application allow me to deeply understand the capabilities and limitations of modern-day artificial intelligence (AI), which is essential to successful deployment of AI in security settings.



My primary research goal is to *develop strong AI-powered next-generation cybersecurity defenses.* My research uniquely combines techniques from **ML** (e.g., online learning), **algorithmic game theory** (e.g., two-player zero-sum games), and **cybersecurity insights** through industry collaboration with Symantec and Intel, the world's leading security companies.

My research targets problems such as how to allocate defense resources, design robust ML algorithms under various threat models such as adversarial attacks, and apply AI to enhance enterprise security and improve social welfare. I focus on the following three interrelated research topics.

1 Theoretically-Principled Defense via Game Theory and ML

Defense resource allocation is a well-known and critical task in security. For example, a company that wants to implement security controls with a limited budget needs to make trade-offs in its deployment. I modeled this problem as a two-player zero-sum game between a defender and an attacker, and introduced a novel solution concept called **diversified mixed strategy** [1].

Inspired by the proverb "Don't put all your eggs in one basket," my new solution concept compels players to employ a "diversified" strategy: one that does not place too much weight on any one action. I systematically studied properties of diversified strategies in multiple games, and designed efficient algorithms that asymptotically achieve the optimum reward within the family of diversified strategies. As a result, this algorithm limits the exposure to adversarial or catastrophic events while still performing successfully in typical cases.

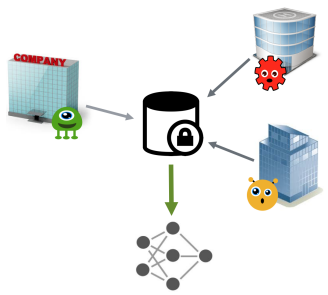


Figure 1: Learning a good model with cybersecurity data from worldwide customers, with little communication.

Leveraging the deep connection between game theory, online learning, and boosting, I proved that the proposed *diversified strategy* concept can also be used to help learn robust and efficient ML models. Specifically, I **solved an open problem** listed in [Daum'e III et al. AISTATS'12] by developing a boosting-based approach [2] in one of the **hardest and most general settings in distributed learning**, where data is adversarially partitioned and distributed across multiple locations, and can have arbitrary forms of noise (Figure 1). Succinctly, since boosting algorithms tend to place too much weight on outliers, we can project the weights back to the set of *diversified* distributions at the end of each boosting iteration. Our algorithm is simultaneously noise tolerant, communication efficient, and computationally efficient. This is a significant improvement over prior works that were either communication efficient only in noise-free scenarios or were computationally prohibitive.

Using a variant definition of the *diversified strategy* customized for online learning, I developed the **first online boosting algorithms with strong theoretical guarantees**, both in the full-information

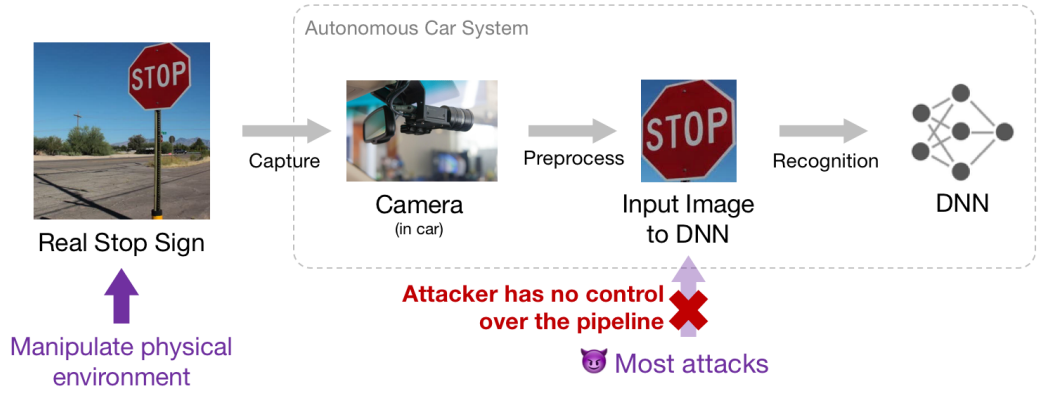


Figure 2: My work on physical adversarial attack discovers a serious vulnerability of DNNs in a more realistic threat model where the attacker does not need to have control over the internal computer vision system pipeline. The crafted physical adversarial objects (e.g., fake stop signs) can fool the state-of-the-art object detectors.

setting [3] and the partial-information bandit setting [4]. The error bound guarantees are in the fully adversarial setting, and thus the algorithms are safe to use in the security context. All of the proposed algorithms in the distributed and online settings are not only theoretically principled but also demonstrate excellent accuracy on real-world datasets.

2 Adversarial Attack and Defense of Deep Neural Networks

Recent advances in deep neural networks (DNNs) have generated much optimism about deploying AI in high-stakes applications, such as self-driving cars. However, it has recently been discovered that given the ability to directly manipulate image pixels in the digital input space, an adversary can easily generate imperceptible perturbations to fool a DNN image classifier.

Although many adversarial attack algorithms have been proposed, attacking a real-world computer vision system is difficult, because attackers usually do not have the ability to directly manipulate data inside such systems (Figure 2). To understand the vulnerabilities of DNN-based computer vision systems, I collaborated with Intel and developed **ShapeShifter** [5], the **first targeted physical adversarial attack on the state-of-the-art Faster R-CNN object detectors**.

Attacking an object detector is more difficult than attacking an image classifier, as it needs to mislead the classifications in multiple bounding boxes with different scales. Extending a digital attack to the physical world adds another layer of difficulty; this requires the perturbation to be sufficiently robust to survive real-world distortions due to different viewing distances and angles, lighting conditions, and camera limitations.

Figure 3: The fake stop sign crafted by *ShapeShifter* is consistently detected as the target class "person" by Faster R-CNN in the drive-by test, while the real stop sign on the left is always correctly detected.

ShapeShifter generates adversarial stop signs that were consistently misdetected by Faster R-CNN as the target objects in real drive-by tests (Figure 3), posing a potential threat to autonomous vehicles and other safety-critical computer vision systems. Our code is **open-sourced** and the drive-by test videos are publicly available[1].

Although completely protecting a DNN model from adversarial attacks remains an open problem, there have been many attempts to mitigate the threat. However, most methods suffer from significant computational overhead or sacrifice accuracy on benign data. In collaboration with Intel, we developed **SHIELD** [6], a practical defense leveraging stochastic compression that removes adversarial perturbations. SHIELD

[1] `https://github.com/shangtse/robust-physical-attack`

makes multiple positive impacts on Intel's research and product development plans. Utilizing Intel's Quick Sync Video (QSV) technology with dedicated hardware for high-speed video processing, we pave the way for real-time defense in safety-critical applications, such as autonomous vehicles. This research has sparked insightful discussion at Intel on *secure deep learning* that necessitates tight integration of practical defense strategies, software platforms and hardware accelerators. I believe our work will accelerate the industry's emphasis on this important topic.

3 AI in Enterprise Security and Beyond

Beyond designing robust and efficient ML algorithms with theoretical guarantees, I am also passionate about applying ML to solve real-world problems. I believe the key to success is through deep communication and collaboration with domain experts. My work on **enterprise cyber threat detection** exemplifies my industry impact and collaboration with security practitioners.

Using telemetry data sent from customers of Symantec's Managed Security Service, I developed the **patent-pending Virtual Product** [7], the **first method** to predict security events and high-severity incidents identifiable by a security product as if it had been deployed (Figure 4). This is made possible by learning from the vast amounts of telemetry data produced by the prevalent defense-in-depth approach to computer security, wherein multiple security products are deployed alongside each other, producing highly correlated alert data. By studying this data, we accurately predicted which security alerts a product would have triggered in a particular situation, even though it was not deployed.

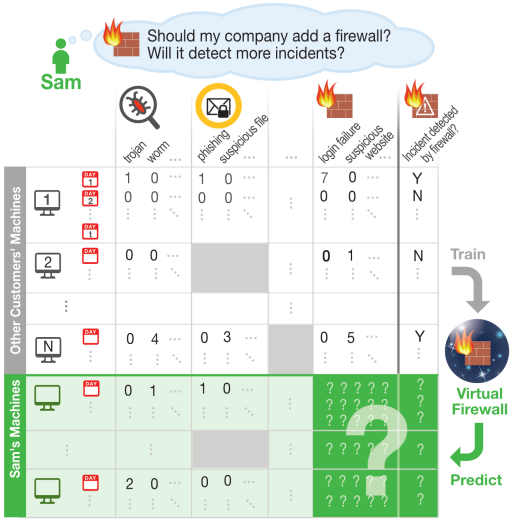


Figure 4: *Virtual Product* helps our user Sam uncover and understand cyber-threats, and informs deployment decisions (e.g., add firewall?) through semi-supervised non-negative matrix factorization on telemetry data from other users (with firewalls deployed). Each row is a machine-day, and each column a security event's occurrences. Missing events from undeployed products are shown as gray blocks. The last column indicates if the firewall has detected an incident. Our *virtual* firewall serves as a proxy to the actual product and predicts the output Sam may observe (dark green block) if he deploys it.

Beyond cybersecurity, I also actively pursue novel applications of AI in various domains to create positive societal impacts. For example, in collaboration with the Atlanta Fire Rescue Department, we developed the **Firebird** framework [8] to help municipal fire departments identify and prioritize commercial property fire inspections, using ML, geocoding, and information visualization together. Firebird computes fire risk scores for over 5,000 buildings in the city, with true positive rates of up to 71% in predicting fires. *Firebird* won the **Best Student Paper Award Runner-up at KDD 2016** and was highlighted by National Fire Protection Association as a best practice for using data to inform fire inspections.

Technology Transfer and Funding Experience

I enjoy interdisciplinary research that bridges theory and application, and I plan to continue doing so during my career. I have been fortunate to collaborate with over 50 coauthors from more than 10 academic departments, companies, and government agencies. My research has resulted in technology transfers and open-sourced tools that make direct impact to industry and society. For example, our *Firebird* project is open-sourced[2] and has been used by the Atlanta Fire Rescue Department to prioritize fire inspections. *ShapeShifter* is also open-sourced and *Virtual Product* is patent-pending. I

[2] `http://firebird.gatech.edu`

also actively participate in research funding proposals, including a **$1.2 million NSF grant** (CNS 1704701. SaTC: CORE: Medium: Understanding and Fortifying Machine Learning Based Security Analytics), a **$1.5 million gift grant from Intel**, and the **IBM PhD Fellowship**.

Future Research Agenda

My long-standing goal is to design practical, robust ML algorithms with strong theoretical guarantees, to reliably solve high-stakes societal problems, such as safe-guarding security-critical systems. I have taken the first important steps toward this goal with my thesis research. Moving forward, I hope to broaden and deepen this investigation, extending my work to more theoretical frameworks and applications. Initially, I will focus on the following three interrelated research directions.

Fortify ML against Adversarial Attacks. Our SHIELD [6] defense is a promising technique to protect DNNs in the image domain. Although the idea of compression can also be applied to other tasks, such as defending against audio attacks in speech recognition [9], the potential of compression is often overlooked in the ML community. I plan to study how compression can be used in various domains to extract only the useful information for ML models, and discard other details including adversarial perturbations.

Furthermore, training a robust ML model is of independent interest beyond security. The current state-of-the-art method in training a robust DNN model against adversarial attacks is *adversarial training*. However, it could take an order of magnitude longer to train, compared to standard training, making it impractical for large-scale tasks. I plan to design novel network architectures and training methods that are scalable and more robust to adversarial attacks, and open the black-box of DNNs by developing new theoretical frameworks and practical tools. I believe this is the key to improving generalization abilities of the current ML models.

Fraud Detection with ML and Game Theory. I believe AI has the potential to protect people from a wide range of cyber harms that affect our everyday lives. For example, *fraud detection* is an important adversarial ML application, where the fraudster creates a benign facade to evade detection. I have worked on detecting fraudulent users and reviews on Yelp by using graph mining techniques like dense graph extraction [10]. However, there is much more information to be utilized to improve detection. I plan to combine techniques in ML, graph mining, natural language processing, and time series analysis to incorporate information from different data sources. I also plan to better understand how and why fraudsters work in particular ways, by using game theory, and ultimately design a framework that discourages people to conduct fraud by better mechanism design.

Model and Data Privacy. Keeping an ML model secret is crucial for defending against black-box adversarial attacks. I plan to design practical ML algorithms that are hard to reverse-engineer, such as dynamic random ensemble using *online boosting* [3]. I will also formalize the benefit of randomization as a defense to adversarial attacks by incorporating the techniques in cryptography. Besides model privacy, I also aim to study the problem of data privacy, which has been receiving an increasing amount of attention in the past few years, due to some high-profile data leaks in industry. *Differential privacy* is the most popular theoretical framework for data privacy, wherein a typical method is by adding random noise to the original data. I believe my proposed *diversified strategy* concept [1] is also helpful to preserving differential privacy.

In summary, my research develops robust and practical ML algorithms with strong theoretical guarantees that empower next-generation cybersecurity defenses. I look forward to leveraging my simultaneous strengths in theory and application, and my diverse research experience in machine learning, game theory, and security, to make positive impacts to academia, industry, and society.

References

[1] Maria-Florina Balcan, Avrim Blum, and **Shang-Tse Chen (alphabetic order)**. Diversified strategies for mitigating adversarial attacks in multiagent systems. In *International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 407–415, 2018.

[2] **Shang-Tse Chen**, Maria-Florina Balcan, and Duen Horng Chau. Communication efficient distributed agnostic boosting. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1299–1307, 2016.

[3] **Shang-Tse Chen**, Hsuan-Tien Lin, and Chi-Jen Lu. An online boosting algorithm with theoretical justifications. In *International Conference on International Conference on Machine Learning (ICML)*, pages 1873–1880, 2012.

[4] **Shang-Tse Chen**, Hsuan-Tien Lin, and Chi-Jen Lu. Boosting with online binary learners for the multiclass bandit problem. In *International Conference on Machine Learning (ICML)*, pages 342–350, 2014.

[5] **Shang-Tse Chen**, Cory Cornelius, Jason Martin, and Duen Horng Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 2018.

[6] Nilaksh Das, Madhuri Shanbhogue, **Shang-Tse Chen**, Fred Hohman, Siwei Li, Li Chen, Michael E Kounavis, and Duen Horng Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 196–204, 2018.

[7] **Shang-Tse Chen**, Yufei Han, Duen Horng Chau, Christopher Gates, Michael Hart, and Kevin A Roundy. Predicting cyber threats with virtual security products. In *Annual Computer Security Applications Conference (ACSAC)*, pages 189–199, 2017.

[8] Michael Madaio, **Shang-Tse Chen**, Oliver L Haimson, Wenwen Zhang, Xiang Cheng, Matthew Hinds-Aldrich, Duen Horng Chau, and Bistra Dilkina. Firebird: Predicting fire risk and prioritizing fire inspections in atlanta. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 185–194, 2016.

[9] Nilaksh Das, Madhuri Shanbhogue, **Shang-Tse Chen**, Li Chen, Michael E Kounavis, and Duen Horng Chau. Adagio: Interactive experimentation with adversarial attack and defense for audio. In *In ECML-PKDD (demo)*, 2018.

[10] Paras Jain, **Shang-Tse Chen**, Mozhgan Azimpourkivi, Duen Horng Chau, and Bogdan Carbunar. Spotting suspicious reviews via (quasi-)clique extraction. In *IEEE Symposium on Security and Privacy (poster)*, 2015.