Safe Machine Learning (October 19-20, 2017)

DARPA workshop at the Simons Institute

Organizers: Hava Siegelmann, Shafi Goldwasser, Peter Bartlett, Moritz Hardt, Dawn Song

1 Introduction

Recent advancements in machine learning have enabled progress on notoriously challenging artificial intelligence (AI) problems in a broad range of domains. These domains include computer vision, robotics, speech recognition, language translation, autonomous transportation, and game playing. This progress initially prompted a general belief that existing machine learning techniques will play a central role in automation of human intelligence tasks and promote an economic and technological revolution. Recent findings, however, suggest that this belief is overly optimistic. It turns out that existing machine learning methods can be easily manipulated to make arbitrary classifications, introduce statistical biases that lead to discrimination, and compromise individuals' privacy.

These critical vulnerabilities of existing machine learning methods are now the major obstacle to implementing artificial intelligence systems that necessitate reliability, dependability, and security. Already today machine learning algorithms are ubiquitously applied, and their vulnerabilities have non-trivial societal implications.

1.1 Objective

The goal of this program is to build a general machine learning arsenal of robust models, both in terms of their reliability and resistance to malicious tampering. This will require a thorough investigation of the extent to which current machine learning techniques fail and designing reliable and secure frameworks that overcome these vulnerabilities. Our vision is that machine learning models should be dependable and modular. In particular, a system that incorporates a machine learning model should use it in the same black-box manner it uses a traditional data structure, without compromising reliability and resilience guarantees.

The directions we aim to pursue can be categorized into three major thrusts:

1. **Robustness:** Design machine learning algorithms that are resilient to adversarial data poisoning and misclassification attacks;
2. **Fairness and transparency:** Design machine learning algorithms that appropriately represent groups in a population in a fair and transparent manner;
3. **Privacy:** Design machine learning algorithms that do not compromise the privacy of individuals in a dataset and can be deployed in a distributed and secure way.

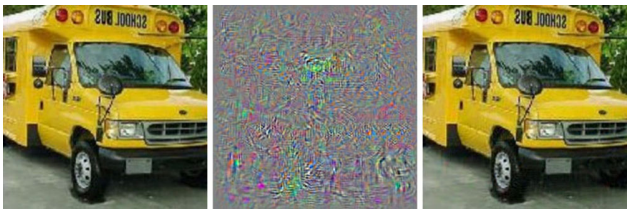

Figure 1: Example of misclassification of bus via noise from Szegedy et al. [50]. The image on the left is a bus, correctly classified by a standard DNN; the figure in the middle depicts the perturbation of pixels added to the image on the left; the image on the right is the image of the bus after noise has been applied, where the DNN classified the bus as an ostrich.

1.2 Why DARPA?

Machine learning is currently an area that attracts substantial interest and resources from the industry. We believe, however, that safe machine learning is unlikely to be championed by the industry, largely due to two reasons. First, the development of safe machine learning requires a long time horizon to come to fruition whereas industry, barring some exceptions, typically invests in short term projects. Exploring vulnerabilities of machine learning and developing a new way to think about machine learning requires a far grander time horizon. Secondly, this effort requires an ensemble of experts from a broad range of areas and their collaboration. In particular, it requires expertise in diverse areas such as optimization, machine learning, cryptography, privacy, security, systems, and human-computer interaction. Moreover, every direction towards robust machine learning requires synthesizing techniques from almost all of these disciplines. DARPA is in a unique position to create a collaboration with world-class experts at the necessary scale. Finally, the concerns we address in this document have deep implications for national security, both in obvious ways in situations where machine learning models are deployed in mission-critical systems, and far subtler ways when nation-state adversaries can use vulnerabilities in these models to dictate the direction of our society. We believe that investment in this effort goes along the DARPA mission of "making pivotal investments in breakthrough technologies for national security".

2 Key Research Directions

2.1 Robust machine learning

The root of the problem is that current machine learning solutions, despite achieving excellent performance in benign settings, fail – often in a truly catastrophic manner – in more adversarial contexts. In particular, the machine learning models trained with existing techniques are often extremely vulnerable to a wide spectrum of attacks, such as causing misclassification by adversarial input perturbation and injection of false data to the training data set. In fact, these models tend to fail even if the settings they are applied in are not fully malicious but merely noisy in a way that conforms to fairly natural noise models.

Fundamental limits of robust generalization

In recent years, classification using deep neural networks (DNNs) has produced state-of-the-art performance on visual classification problems, achieving near-human-level performance on image recognition tasks. Despite their empirical success, however, there are many open questions related to our theoretical and intuitive understanding of deep learning methods. One important open question relates to the robustness of deep neural networks to noise.

A series of recent papers show that standard deep learning classifiers are extremely sensitive to adversarially chosen noise [49, 44, 39, 1, 19]. In particular, as many examples show, one can take an image which is correctly classified by a DNN, and make the same image become completely misclassified by perturbing the pixels of the image in a manner that is not detectable to the human eye. In a world which develops an ever-growing dependence on automatic classification using neural networks, with applications ranging from self-driving cars to face recognition, such sensitivity to noise can have dire consequences.

Misclassification on perturbed images implies DNN models do not generalize well under adversarial errors. Designing methods that are robust to adversarial noise requires understanding the fundamental limits of robust generalization. This brings to fore problems addressing which raises modeling, algorithmic, and experimental challenges.

At a high level, the modeling challenge is that we are interested in providing provable guarantees for the methods we suggest. The issue, however, is that training a deep learning classifier requires solving an intractable optimization problem. Thus, we must somehow circumvent this issue. From an algorithmic perspective, designing robust optimization techniques introduces a notoriously challenging set of problems. Robust optimization is concerned with optimizing min-max loss functions which often becomes NP-hard even for well-behaved objectives. Most importantly, we are interested in developing algorithms that work well in practice. This requires designing architectures and training large-scale DNNs.

Intuitively, if we know the adversarial noise in advance, and have algorithms that generate adversarial noise, we can improve classification accuracy by injecting adversarial noise to the training data, and train a classifier on this corrupted data set. Such logic is sound and does lead to improvements in test accuracy [42], but there are numerous complications. Since there are multiple ways to generate adversarial noise, which noise should be used to corrupt the training data? And if we want to use multiple noise types to corrupt the training data, what is the optimal proportion for how the noises should be applied, and how can we compute what this proportion is? Are there less intuitive methods that perform even better? Does the fact that we care about the worst-case noise performance rather than, for instance, the expected performance, make a difference? These are some of the complications that need to be addressed en route to a constructing noise-robust DNNs.

Reducing the space/time overhead of training robust models

Recent work on attacks and defenses in adversarial machine learning rely on finding equilibria in a zero sum game between two objectives. While theoretically sound, the computational overhead of such algorithms is enormous which often makes the state-of-the-art algorithms infeasible in practice.

The reason for the large computational overhead is two fold. First, computing the equi-

libria relies on iterating between utility maximizing objectives. The guarantees often hold when the number of iterations tend to infinity, which can make the results infeasible. The second bottleneck is that in many cases, there is a large overhead of computing each iteration. This is because one needs to optimize a non-convex high-dimensional utility function which is computationally expensive.

One of the goals of this project is to investigate new methods for computing equilibria in zero-sum games tailored to robust machine learning tasks. There has been a great deal of advancement in our understanding of computing equilibria in recent years and we intend to apply and develop new techniques for the benefit of robust machine learning.

Differential testing of models: how to evaluate robustness

An interesting question that we intend to explore is that of testing a model. Traditionally in machine learning, the quality of a classifier can be evaluated by drawing random examples from a data set similar to the one used to train the classifier. For robust machine learning algorithms, such evaluations are no longer valid. An important goal of this project is to establish principled criteria to evaluate robust models.

Adversarial examples beyond human perception

Recent adversarial machine learning research has maintained a significant focus on robustness of vision systems [4, 50, 24, 45, 12, 19], in which human perception is inextricably linked to formation of adversarial examples. Starting from a target instance, perturbation to an adversarial example is deemed successful at model evasion, when the model's prediction is flipped relative to the target, but a human's judgment of the example remains unchanged. While this definition is motivated in vision, many learning domains in need of adversarial treatment do not possess such an infallible oracle. In addition, the involvement of humans as oracles precludes us from defining adversarial examples rigorously in a way that can be analyzed formally.

One type of application domain includes a human operating in the loop, similar to applications of adversarial vision, video surveillance, security checkpoints, and self-driving cars. Going beyond an infallible oracle, known classes of optical illusion could be exploited to fool human perception [30]. Research addressing this direction would constitute an exciting instance of cognitive science contributing to adversarial learning. In the reverse direction, adversarial learning could inform cognitive science when the (human) victim in social engineering is modeled as a learner. In order to train potential victims how to better identify and react to phishing communications via email, phone, or in person, robust learners could play an assistive role to ideas from teaching [22]. Many other (potentially adversarial) domains exist in this broad class, e.g., predictive policing [25] and judicial support systems —an active focus of algorithmic fairness research [15]–where robustness to adversarial tampering has so far gone unexplored.

A second type of domain is that in which human perception does not provide a clear advantage over existing algorithms, a prime example being malware analysis [32]. A human expert may enjoy higher accuracy than a trained detection algorithm, however querying this expertise comes at significantly greater financial and time costs: requirements on adversarial

examples relative to a human could be realistically relaxed. The domains listed here still only scratch the surface of those that would benefit from an adversarial treatment. Exploring more diverse domains will arguably advance adversarial machine learning significantly, while consideration of the role of human perception is vital to organize domains and uncover connections to diverse areas of science and technology such as the cognitive sciences.

Robust ML beyond the adversarial setting

The recent research focus on machine learning in adversarial settings could be seen as premature from a quality-assurance perspective, given that ML models still fail somewhat consistently in settings with no maliciously intentioned adversary. There are abundant examples of ML systems failing dramatically when faced with rare conditions (e.g., cloudy weather conditions [48]), which hint at more general issues with the development and evaluation process of deployed ML systems [46, 47, 9, 43].

Compared to the quality-assurance pipeline of general software systems, studied extensively since the 1980s, rigorous testing of ML software is still in its infancy. Machine learning, and especially deep learning, offers little to no performance guarantees: average test classification error remains the metric of choice; more often than not the only reported metric. Finally, even when errors in ML systems become apparent (e.g., [48]), principled methods for understanding and fixing these errors are lacking.

Ultimately, the goal here may be both foundational (i.e., setting the right definitional framework and guidelines, with preliminary work from [46, 47, 9]) and educational (i.e., raising awareness and getting the broader community to care about these important questions). Arguably, the application of ML in security-sensitive settings is contingent on a mature development, evaluation and deployment cycle. Given the little work overall in this area, it seems that a number of core and possibly industry-shaping contributions may be achievable in this direction.

Data poisoning: how to learn from data you can't trust?

Machine learning systems trained on user-provided data are susceptible to data poisoning attacks, whereby malicious users inject false training data with the aim of corrupting the learned model. While recent work has proposed a number of attacks and defenses, little is understood about the worst-case loss of a defense in the face of a determined attacker.

The majority of recent work on adversarial examples in machine learning considers attacks on the test set. That is, once a classifier has been trained, the adversary fools the classifier by distorting an image from the test set that the classifier did not use in training. A different problem however, is that of noise on the training set introduced by Biggio, Nelson, and Laskov showing that SVMs can easily be corrupted by introducing misclassified point in the training set [4]. In recent work Koh and Liang show that even corrupting the classification of a single data point in the training set can lead to consistent misclassification by a standard DNN classifier [37].

One of the goals of this project is to design data poisoning attacks and defenses. In particular, we are interested in characterizing the extent to which misclassified data points

can affect a classifier and consider optimal attack strategies. Doing so would pave the road to architectures that are robust to data poisoning attacks.

Beyond worst-case analysis of robustness

The design and analysis of provably robust estimators has been a central topic in statistics and machine learning for several decades. But for even the most basic problems, being robust seems fundamentally at odds with computational complexity. In supervised learning, the Perceptron algorithm can learn a linear separator. But once an adversary is allowed to corrupt a small fraction of the data there are no known algorithms that can find any halfspace with non-trivial agreement. It has been recently proved [16] these problems are intractable under natural conjectures about refuting random CSPs. In unsupervised learning, estimating the mean and covariance are central topics in robust statistics (called robust estimates of location and scale) but estimators with large breakdown point (i.e. the fraction of corruptions they can tolerate before producing irrelevant results) need time exponential in the dimension to compute. So while there are plenty of provable robust estimators in principle, they remain largely out of reach of efficient algorithms [29, 26].

Recently, a unifying theme and way around these computational impediments has emerged, which relies on reformulating the problems in more well-posed settings. For learning a halfspace, when the noise is stochastic rather than adversarial, Blum et al. gave an algorithm for learning a halfspace with optimal agreement. In unsupervised learning, Diakonikolas et al.[17], Lai et al. [40] and Charikar et al. [13] gave algorithms for robustly learning the mean and covariance when the uncorrupted points come from a Gaussian distribution. These algorithms extend to a number of other parameter learning problems. Candes et al. [10] gave algorithms for finding low-rank approximations in the presence of noise by showing that under incoherence assumptions, there are simple convex programs to decompose a matrix into a low-rank and sparse part. These algorithms all rely on distributional or structural assumptions that put us outside of the realm of worst-case hardness.

The idea of coping with computational intractability has been an important research direction in theoretical computer science for many years, and the insights garnered (e.g. notions like approximation stability of Balcan et al. [2]) by this community could potentially have a transformative effect on machine learning and statistics, both in our ability to circumvent known hardness vs. robustness tradeoffs, as well as revisit popular and practical algorithms that have only been analyzed in stochastic settings and rethink their behavior in the presence of deviations from an idealized model (e.g. as in semirandom models of Feige and Kilian [20]). In short, the need for provably robust algorithms is both a question of algorithm design and of modeling. It is not merely about formulating the strongest notions of robustness and hoping to achieve them algorithmically, but understanding what types of compromises can be made and what types of properties of realistic data offer a way around computational intractability.

Verification methods for machine learning models

Deployment of security-critical systems requires a careful verification that each of the system?s components is indeed secure and reliable. Unfortunately, we currently lack adequate

tools for performing such verification in the context of machine learning models, in general, and deep learning models, in particular. The key bottleneck is that the existing methods, which correspond to fairly direct adaptations of verification techniques developed for other, non-ML context, are still rather impractical, even for modestly sized models [33, 11].

One of the directions this project will pursue is designing a new, more suitable framework for verification of machine learning models. In particular, the goal will be to explore if the rich algorithmic toolkit of convex programming and geometric approximation techniques can lead to methods that enable us to analyze large-scale machine learning models efficiently while still providing provable guarantees.

2.2 Fairness and Transparency

Artificial Intelligence (AI) technology increasingly supports mission critical systems affecting labor and employment, the financial markets, energy and resource availability, transportation, health, the police force, and the military. Systems in these domains have profound impact on the functioning of society, political and military decisions, and the daily lives of human beings.

A major challenge is to ensure that AI promotes the well-being and advancement of society. Addressing the societal impact of artificial intelligence is not only an ethical concern. Systems perceived as unfair or opaque can prevent individuals from cooperating with the system thus compromising its functionality. Similarly, lack of trust can compel individuals to evade, second guess, or manipulate technological systems encountered in their lives. Ultimately, failure to address societal challenges raised by AI can lead to political instability, as well as weakened infrastructure, both issues of great import to national security.

Fairness, accountability, and transparency

Over the past few years, fairness has emerged as a matter of serious concern within machine learning. There is growing recognition that even models developed with the best of intentions may exhibit discriminatory biases, perpetuate inequality, or perform less well for historically disadvantaged groups. Considerable work is already underway within and outside machine learning to both characterize and address algorithmic fairness in all stages of the algorithmic pipeline, and particularly in the gathering and cleaning of training data as well as in learning predictors and classifiers. At the same time, there is growing concern that the complexity of modern machine learning solution limits the transparency of the system, as well as it may dilute accountability and responsibility for the system's malfunctioning or failure.

As the great successes in cryptography and privacy teach us, finding the right definitions is key to theoretical and practical impact. Nevertheless, a flurry of recent research (cf. [18, 36, 14, 27, 35, 34, 28] for a few examples) suggests that no single definition of algorithmic fairness can capture all scenarios. Trade-offs are exposed between fairness for individual members of a population (powerful and hard to attain) and fairness on average (easy to obtain but weak); between machine-learning utility and equalized treatment; between observable parameters and possibly hidden causality. Providing tools for identifying meaningful task-dependent fairness criteria and developing tools to measure and implement these fairness criteria is of critical importance if we expect machine-learning practitioners to implement

fair algorithms. In many cases, the questions lead to trade-off scenarios where multiple objectives have to be reconciled.

Artificial intelligence in almost all application faces uncertainty that makes prediction errors and incorrect decisions unavoidable. Additional causes of error include poor or biased training data, weak design, software bugs, as well as misspecified or poorly stated training objectives. Errors are costly to multiple stakeholders including not only the operator of the system, but also all individuals that are affected by the actions of the system. An important direction is to characterize and define, measure and mitigate the concrete harms that a system may cause.

Reliability in dynamic environments

When a learned model is used for consequential decision making, it faces the problem that the environment will typically react and adapt to the model's decision. This response presents the model with unforeseen inputs that can lead to critical failure with devastating outcomes.

As of today, machine learning methods are based on finding patterns in historical data, but largely fail to address how deployment of the model will change and affect future data. The response of the environment to a model may be either strategic or adversarial. Individuals might begin to strategically adapt to the model in order to gain a personal advantage. Adversaries might actively seek to exploit the model. In either case, the performance of a model may be compromised due to the dynamic nature of the environment that was not accounted for when the model was trained. The primary paradigm to address these challenges in practice practice is trial and error. Engineers frequently re-train and evaluate their models either on holdout data or in deployment. Few, if any, a priori guarantees guide the design of AI systems today. The problems of manipulation and gaming are insufficiently addressed by ad-hoc measures of obscurity that enjoy no formal guarantees and limit the transparency and explainability of the system.

2.3 Privacy and Security

An important direction we intend to pursue is that of privacy and security through the entire pipeline of a machine learning workflow. A growing concern is that systems that rely on machine learning sacrifice the privacy of the users, both in the model training phase and in the model utilization stage where we use the model for classification.

An example arises in the setting of a hospital that uses patient data, including medical records, demographic data, and genomic data, to train a machine learning model that predicts the probability of a medical condition given several patient attributes. The privacy concern is that the model might reveal information about individuals in the training data. The security concern arises when the training data is owned by multiple mutually distrusting organizations, and thus, training has to be done in a distributed manner.

Having trained the model, the hospital would like to monetize the model by proving predictions as a service. The privacy concern is that users with oracle access to the model might reverse-engineer it, a scenario known as model-stealing. The security concern arises because not only does the hospital want to protect the model, the users on the other hand also want to protect their input data to the classification process.

We remark that privacy loss can occur in many subtle ways, even unintentionally when systems leak information about the training data which may contain sensitive information about users. This is troublesome, not only in the medical domain, but in other areas such as financial predictions, insurance decisions, and college admissions, where the data in question is highly sensitive.

Secure ML training

Secure ML training deals with the scenario where two (or more) entities with disparate types of data, such as genomic data and phenotypic data, wish to collaborate on their datasets to come up with models for disease prediction. In such cases, it is important and sometimes legally mandated by law (e.g. HIPAA and FERPA) to do this in a way that does not expose the parties' data to one another. A pressing question to be answered is: *Can we design efficient privacy-preserving machine learning algorithms in all these settings?* Secure multiparty computation and homomorphic encryption are cryptographic techniques that allow parties to collaborate on their individual data to compute a common result, without revealing anything else about their data. Secure multiparty computation starting with [51, 23, 3], and modern day fully homomorphic encryption [21, 8, 7, 6, 41] are two relevant technologies which promise to be useful to address collaboration on training data by distrustful parties.

Likely the most pressing question at the forefront of encrypted computing is how to do privacy-preserving training, a far more computationally intensive task than privacy-preserving inference (just as it is in the world without privacy). Scalable protocols for simple tasks such as linear regression have been developed. The next step would be to train logistic regression models and subsequently, simple neural networks. An overarching hard problem we will tackle in this domain is to design algorithms and protocols for encrypted gradient descent (in its various forms). We believe that a fast encrypted gradient descent protocol will prove to be an extraordinarily powerful tool in our arsenal.

Differential privacy-preserving machine learning

For well over a decade *differential privacy* has been used as a formal definition for privacy and has evolved as the de-facto standard for statistical data privacy both academia and in industry. Differential privacy ensures that from the output of an algorithm, an adversary learns almost the same information about an individual irrespective of her presence/absence in the data set. Although differential privacy has been adopted by Google, Apple, Uber, Microsoft, and US Census, its adoption for many machine learning tasks has been slower than anticipated. Several important research directions remain that could help fuel adoption.

First, if the data is distributed across various user devices, then can we design differentially private learning algorithms that do not need multiple rounds of interactions with the distributed data set? This question is important because, in distributed learning, one of the major bottlenecks is communication, both in terms of turnaround time and monetary cost.

Second, many machine learning applications, including deep learning, are non-convex in nature. There does not exist much work on non-convex learning in the context of privacy, in part due to the lack of theory for non-convex formulations. Are there differentially private

algorithms for non-convex learning that also have strong theoretical utility guarantees, and work well in practice?

Third, a major issue with using the basic notion of differential privacy is that it inherently hinders personalization, i.e., by definition the output of the algorithm cannot depend too much on a particular user's data. However, in many personalized recommendations (e.g., movie recommendations) it is important that the learning models incorporate specific attributes of individual users. Can we design effective private learning systems that allow personalization? One approach may be to ensure that every user learns a different model which can depend arbitrarily on her data but does not depend too much on other people's data.

Secure ML inference

The question of privacy during the classification stage has been partially explored in work by [5], protocols for privacy-preserving inference with simple machine learning models such as decision trees and linear classifiers and a more general classifier combining these using AdaBoost have been developed. More recently, in [31], it has been shown how to do encrypted classification of images using large convolutional neural networks with millions of parameters (such as networks trained on the CIFAR-10 and ImageNet datasets [38]) with an end-to-end runtime of a few seconds.

Such runtimes would have been unimaginable just a year or two ago, and were made possible with two concerted lines of attack, together with many new algorithms and careful optimizations: (1) while black-box approaches that use the existing cryptographic techniques as-is incur an extremely high overhead, one can try and exploit the rich mathematical structure inherent in ML, and computations (concretely, convolutions and homomorphic encryption are made for each other); and (2) interactive protocols often reduce the cryptographic burden and are order of magnitudes more efficient than puritan approaches that rely on either multiparty computation or homomorphic encryption alone. That said, it is important not to overspecialize these algorithms and ensure that they generalize well to the encrypted computation of a wide set of machine learning algorithms. In addition, much work remains to be done to improve efficiency of the current set of capabilities.

Defenses against model theft

In a model theft attack, an adversary seeks to determine the parameters of a proprietary model held by a separate company or a government agency by making multiple queries to it. The motivation of the adversary could be simply "stealing" the model that might have been built using proprietary data. Model theft can also be used as a first step in a privacy attack or a white box adversarial example attack. For example, if the model in question was trained on sensitive data about people, the adversary could steal the model first, and then launch a membership inference attack in order to violate the privacy of people whose data was used to train the model. Similarly, the adversary could use model theft as a first step to determine the model parameters before launching a white box adversarial example attack. Thus, leaving model theft unsolved will leave machine learning models more vulnerable to an adversary who wishes to launch a privacy attack or an adversarial example attack.

The state-of-the-art in this area is that there are a few attacks known, but no defenses. Model theft attacks also essentially boil down to active learning using membership queries; thus existing active learning algorithms may be used to launch high quality attacks that only need a few queries to successfully recover the model parameters. Finally, defense against model theft followed by a privacy attack is possible if we answer queries with a differentially private model; however, this may still leave the model vulnerable to other kinds of attacks.

The missing elements in the state of the art are (a) an understanding of what is possible when the model class is unknown and (b) rigorous defenses that can prevent an adversary from learning parameters of the model and still provide reasonably accurate answers to queries. In general, the problem can be tested on available datasets, and hence data is not a big issue for this problem.

We believe that both tasks could be achieved within the next 3-5 years by formalizing and leveraging the connection between model theft and active learning. There is already a large body of existing work on active learning, and upper bounds and algorithms for active learning would provide ideas that can be leveraged to provide model theft attacks; similarly, active learning lower bounds (eg, in the presence of noise close to the decision boundary) could be used to provide defenses that provide noisy responses to strategically chosen queries. By leveraging this connection, we should be able to resolve this problem.

References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

[2] Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Clustering under approximation stability. *J. ACM*, 60(2):8:1–8:34, 2013.

[3] Michael Ben-Or, Shafi Goldwasser, and Avi Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation (extended abstract). In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing, May 2-4, 1988, Chicago, Illinois, USA*, pages 1–10, 1988.

[4] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

[5] Raphael Bost, Raluca Ada Popa, Stephen Tu, and Shafi Goldwasser. Machine learning classification over encrypted data. In *22nd Annual Network and Distributed System Security Symposium, NDSS 2015, San Diego, California, USA, February 8-11, 2015*, 2015.

[6] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. (leveled) fully homomorphic encryption without bootstrapping. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 309–325, 2012.

[7] Zvika Brakerski and Vinod Vaikuntanathan. Efficient fully homomorphic encryption from (standard) LWE. In *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pages 97–106, 2011.

[8] Zvika Brakerski and Vinod Vaikuntanathan. Fully homomorphic encryption from ring-lwe and security for key dependent messages. In *Advances in Cryptology - CRYPTO 2011 - 31st Annual Cryptology Conference, Santa Barbara, CA, USA, August 14-18, 2011. Proceedings*, pages 505–524, 2011.

[9] Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, and D. Sculley. What's your ML test score? a rubric for ML production systems. 2016.

[10] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, 2011.

[11] Nicholas Carlini, Guy Katz, Clark Barrett, and David L Dill. Ground-truth adversarial examples. *arXiv preprint arXiv:1709.10207*, 2017.

[12] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57, 2017.

[13] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 47–60, 2017.

[14] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*, 2017.

[15] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 797–806, 2017.

[16] Amit Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 105–117, 2016.

[17] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 655–664, 2016.

[18] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pages 214–226, 2012.

[19] Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1712.02779*, 2017.

[20] Uriel Feige and Joe Kilian. Heuristics for semirandom graph problems. *J. Comput. Syst. Sci.*, 63(4):639–671, 2001.

[21] Craig Gentry. Fully homomorphic encryption using ideal lattices. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pages 169–178, 2009.

[22] Sally A. Goldman and Michael J. Kearns. On the complexity of teaching. *J. Comput. Syst. Sci.*, 50(1):20–31, 1995.

[23] Oded Goldreich, Silvio Micali, and Avi Wigderson. How to play any mental game or A completeness theorem for protocols with honest majority. In *Proceedings of the 19th Annual ACM Symposium on Theory of Computing, 1987, New York, New York, USA*, pages 218–229, 1987.

[24] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

[25] Samuel Greengard. Policing the future. *Commun. ACM*, 55(3):19–21, 2012.

[26] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics. The approach based on influence functions*. Wiley New York, 1986.

[27] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[28] Ursula Hebert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Calibration for the (computationally-identifiable) masses. In *arXiv preprint arXiv:1711.08513*, 2017.

[29] P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley New York, 2009.

[30] Hui Ji and Cornelia Fermüller. Bias in shape estimation. In *Computer Vision - ECCV 2004, 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part III*, pages 405–416, 2004.

[31] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha Chandrakasan. Gazelle: A low latency framework for secure neural network inference. *CoRR*, abs/1801.05507, 2018.

[32] Alex Kantchelian, Michael Carl Tschantz, Sadia Afroz, Brad Miller, Vaishaal Shankar, Rekha Bachwani, Anthony D. Joseph, and J. Doug Tygar. Better malware ground truth: Techniques for weighting anti-virus vendor labels. In *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security, AISec 2015, Denver, Colorado, USA, October 16, 2015*, pages 45–56, 2015.

[33] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pages 97–117. Springer, 2017.

[34] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *arXiv preprint arXiv:1711.05144*, 2017.

[35] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744*, 2017.

[36] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[37] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1885–1894, 2017.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114, 2012.

[39] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[40] Kevin A. Lai, Anup B. Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 665–674, 2016.

[41] Adriana López-Alt, Eran Tromer, and Vinod Vaikuntanathan. On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption. In *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*, pages 1219–1234, 2012.

[42] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[43] Andrew Ng. Advice for applying machine learning.

[44] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[45] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, pages 372–387, 2016.

[46] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young. Machine learning: The high interest credit card of technical debt. In *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*, 2014.

[47] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2503–2511, Cambridge, MA, USA, 2015. MIT Press.

[48] A. SINGHVI and K. RUSSELL. Inside the self-driving tesla fatal accident. In *The New York Times*.

[49] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[50] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[51] Andrew Chi-Chih Yao. How to generate and exchange secrets (extended abstract). In *27th Annual Symposium on Foundations of Computer Science, Toronto, Canada, 27-29 October 1986*, pages 162–167, 1986.