AI systems, in particular learned systems at training and inference time also pose novel demands on compute infrastructure and present new risks. We are working on adversarial machine learning, identifying threats to the training and deployment of learned systems and develop corresponding defense strategies. In some cases we can use platform properties and mechanisms such as assured data provenance as defense mechanism.

Reliable generation and use of provenance information

<https://researcher.watson.ibm.com/researcher/view_group.php?id=9571#Publications>

[**Adversarial Machine Learning**](https://researcher.watson.ibm.com/researcher/view_group.php?id=9571)[**feedback**](mailto:Bryant.Chen@ibm.com?cc=resweb@us.ibm.com&subject=re:%20https%3A//researcher.watson.ibm.com/researcher/view_group.php%3Fid%3D9571&body=Please%20put%20your%20comments%20%28e.g.%2C%20identifying%20broken%20links%29%20here.%20If%20you%20are%20trying%20to%20send%20information%20to%20an%20individual%20researcher%2C%20please%20send%20email%20directly%20to%20them.)

**links**

**Adversarial Machine Learning - overview**

**Overview**

The use of machine learning models has become ubiquitous. Their predictions are used to make decisions about healthcare, security, investments and many other critical applications. Thus it is not surprising that bad actors would want to manipulate such systems for nefarious purposes.  All machine learning systems are trained using data sets that are assumed to be representative and valid for the subject matter in question. However, malicious actors can impact how the artificial intelligence system functions by poisoning the training data. This threat is exacerbated when the machine learning pipeline that includes data collection, curation, labeling, and training is not controlled completely by the model owner.

In this project, we seek to answer these questions: How can you tell when the training data has been poisoned? Can you repair a model that has been poisoned?

Generally speaking, malicious actors poison training data to

* Misclassify inputs - Here, the adversary aims to shift the decision boundary of the model  to ensure that a specific input is misclassified to a targeted class. For example,  such attacks might categorize certain pollutants as innocuous, a sick person as healthy, or an anomaly as normal. A particularly insidious attack in this category is the *backdoor or trojan attack*, where the adversary carefully poisons the model by inserting a backdoor key to ensure it will perform well on standard training data and validation samples, but misbehaves only when a backdoor key is present. Thus an attacker can selectively make a model misbehave by introducing backdoor keys once the model is deployed. In one traffic example, a backdoor causes the model to misclassify a stop sign as speed limit whenever a post-it note has been placed on the stop sign. However, the model performs as expected on stop signs without the post-it note, making the backdoor difficult to detect since users do not know the backdoor key (a post-it note in this case) a priori. Clearly, such a backdoor can result in disastrous consequences for autonomous vehicles.
* Reduce model performance - The objective of the adversary in this case is to limit the system's usefulness. Here the adversary attempts to reduce the overall accuracy of the model resulting in general misclassifications.

Poisoning threats are particularly relevant when training data is obtained from untrusted sources, such as crowdsourced data or customer behavior data. Additionally, the risk increases when the model requires frequent retraining or customization. Lastly, the ability to detect when models have been poisoned or tampered with is vital when they are trained by untrusted third-parties (e.g. obtained from a model marketplace).

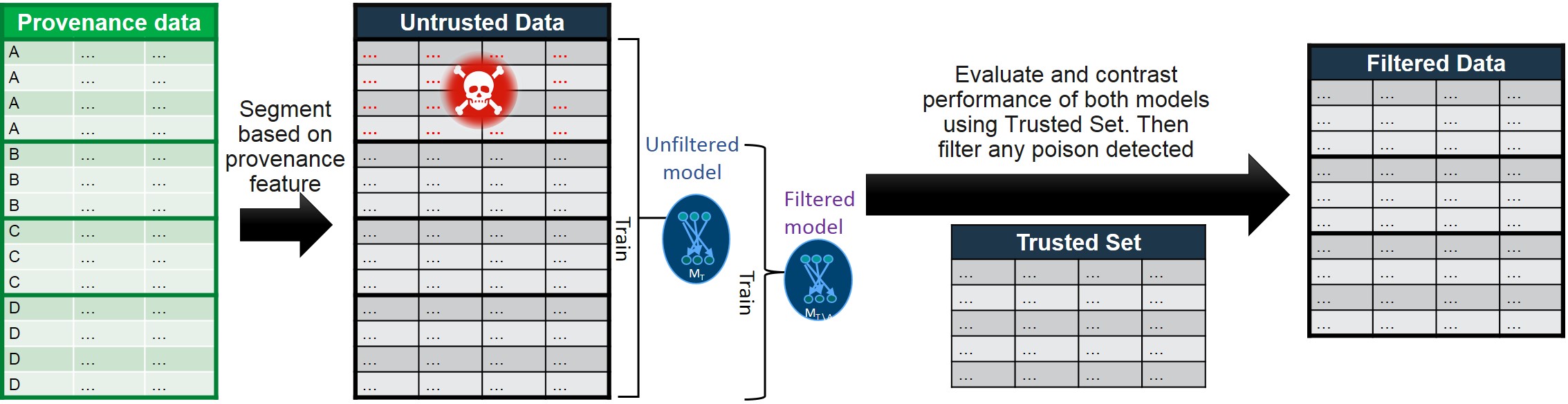
Up to this point (2018), most research has focused on demonstrating and categorizing the types of malicious attacks against machine learning systems training data. However, few defenses have been proposed to proactively detect and revert poisoning attacks. This work goes directly at identifying and correcting malicious attacks on training data. The proposed innovations use two different methodologies to detect and repair different types of poisoning attacks:

* The provenance-based RONI approach is appropriate for models when there is a trusted provenance feature in the dataset.
* The activation cluster approach is appropriate for detecting backdoors and distinct decision pathways that lead to a common classification.

**Provenance-Based RONI**

Recently, a number of secure provenance frameworks have been developed for Internet of Things environments. These frameworks use modern cryptographic methodologies to ensure that provenance data, which describes the origin and lineage of collected datapoints, cannot be modified by adversaries. They take advantage of these frameworks to help detect poisonous data inserted to reduce model performance.

Intuitively, it is difficult for an adversary to compromise every data source, due to time and resource constraints. For this reason, one often expect poisonous data to originate from a limited number of sources. This method segments the training data into groups according to provenance data where the probability of poisoning is highly correlated across samples in each group. Once the training data has been segmented appropriately, data points in each segment are evaluated together by comparing the performance of the classifier trained with and without that group. The figure below depicts this process.

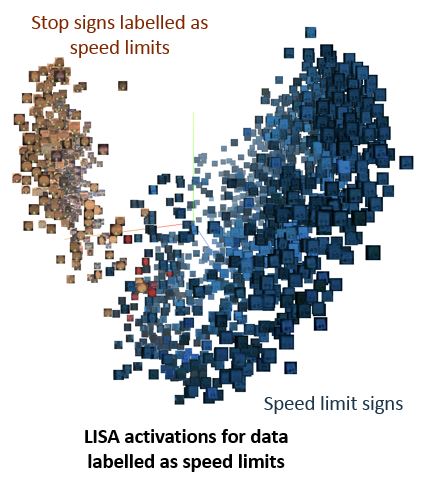


In contrast, a prior approach called Reject on Negative Impact (RONI) evaluated the effect of individual data points on the performance of the final classifier. However, single data points often have minimal impact on the overall performance, and, as a result, poisonous data may escape detection. Additionally, evaluating each data point incurs significant computation and time costs. By using provenance data, this method is able to appropriately group datapoints together and evaluate their cumulative effect on the classifier, thereby increasing detection rates and reducing computational costs.

Finally, RONI proposers note that the provenance-based defense can be applied to any setting in which a trusted feature that is indicative of where poisonous data might be concentrated is available. For example, if an adversary tried to poison a machine learning model that detects fraudulent credit card transactions, the *account number* can be used as a trusted feature. Adversaries may falsely report transaction as fraudulent and/or legitimate, but they cannot manipulate the account to which the transaction is posted. Additionally, they can only compromise a limited number of credit cards. For more information, please see the [Publications](https://researcher.watson.ibm.com/researcher/view_group.php?id=9571" \l "Publications) section

**Neural Network Activation Clusters**

Our team has developed a method to detect backdoor attacks by analyzing differences in how a neural network decides on the classification. "Activations" are the intermediary computations made by the network before making its final classification. Our approach segments training data according to its labels and clusters the corresponding activations from the last hidden layer of the neural network. Poisonous and legitimate data immediately separate into distinct clusters, akin to the way in which different areas of the brain light up on  scans when subjected to different stimuli. This can be readily seen in the figure below in which a backdoor trigger, a post-it note on a stop sign, has been categorized as a speed limit.



Each cluster can then be examined for poison.  Our results show over 99% accuracy on both the MNIST and LISA image datasets. We created an averaged image to examine the clusters. In both cases the image trigger was apparent. Similarly, we have tested the approach on a text-based data set using Rotten Tomatoes Sentiment Analysis in which we injected a backdoor signature. Here we achieved 97% accuracy and identification of the poisonous signature. Once the backdoor has been identified, we demonstrate how the backdoor can be effectively removed by further training the neural network using the poisoned data which is relabeled for correctness.

[**https://researcher.watson.ibm.com/researcher/view\_group.php?id=9571#Publications**](https://researcher.watson.ibm.com/researcher/view_group.php?id=9571#Publications)

**Publications**

* *Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering.  Submission to NIPS December 2018*
* Detecting Poisoning Attacks on Machine Learning in IoT Environments. IEEE ICIOT (Best Paper Award) July 2018
* [Mitigating Poisoning Attacks on Machine Learning Models: A Data Provenance-Based Approach](https://dl.acm.org/citation.cfm?doid=3128572.3140450). ACM Workshop on Artificial Intelligence and Security (AISEC) @ CCS 2017

The use of machine learning models has become ubiquitous. Their predictions are used to make decisions about healthcare, security, investments and many other critical applications. Given this pervasiveness, it is not surprising that adversaries have an incentive to manipulate machine learning models to their advantage. One way of manipulating a model is through a *poisoning* or *causative* attack in which the adversary feeds carefully crafted poisonous data points into the training set. Taking advantage of recently developed tamper-free provenance frameworks, we present a methodology that uses contextual information about the origin and transformation of data points in the training set to identify poisonous data, thereby enabling online and regularly re-trained machine learning applications to consume data sources in potentially adversarial environments. To the best of our knowledge, this is the first approach to incorporate provenance information as part of a filtering algorithm to detect causative attacks. We present two variations of the methodology - one tailored to partially trusted data sets and the other to fully untrusted data sets. Finally, we evaluate our methodology against existing methods to detect poison data and show an improvement in the detection rate.

# Mitigating Poisoning Attacks on Machine Learning Models: A Data Provenance Based Approach

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Full Text: | PDFPDF [[Get this Article](https://dl.acm.org/purchase.cfm?id=3140450)Get this Article](https://dl.acm.org/purchase.cfm?id=3140450) |  |  |  |  | | --- | --- | --- | | Authors: | [Nathalie Baracaldo](https://dl.acm.org/author_page.cfm?id=81502802766&coll=DL&dl=ACM&trk=0) | [IBM Research, San Jose, CA, USA](https://dl.acm.org/inst_page.cfm?id=60009253) | |  | [Bryant Chen](https://dl.acm.org/author_page.cfm?id=99659217295&coll=DL&dl=ACM&trk=0) | [IBM Research, San Jose, CA, USA](https://dl.acm.org/inst_page.cfm?id=60009253) | |  | [Heiko Ludwig](https://dl.acm.org/author_page.cfm?id=81100613257&coll=DL&dl=ACM&trk=0) | [IBM Research, San Jose, CA, USA](https://dl.acm.org/inst_page.cfm?id=60009253) | |  | [Jaehoon Amir Safavi](https://dl.acm.org/author_page.cfm?id=99659217899&coll=DL&dl=ACM&trk=0) | [IBM Research, San Jose, CA, USA](https://dl.acm.org/inst_page.cfm?id=60009253) | |

https://dl.acm.org/citation.cfm?doid=3128572.3140450

https://arxiv.org/pdf/1808.04866.pdf

Mitigating Sybils in Federated Learning Poisoning

Clement Fung

University of British Columbia

cfung1@cs.ubc.ca

Chris J.M. Yoon

University of British Columbia

yoon@alumni.ubc.ca

Ivan Beschastnikh

University of British Columbia

bestchai@cs.ubc.ca

Abstract

—Machine learning (ML) over distributed data is relevant to a variety of domains. Existing approaches, such as federated learning, compose the outputs computed by a group of devices at a central aggregator and run multi-round algorithms to generate a globally shared model. Unfortunately, such approaches are susceptible to a variety of attacks, including model poisoning, which is made substantially worse in the presence of sybils.

In this paper we first evaluate the vulnerability of federated learning to sybil-based poisoning attacks. We then describe FoolsGold, a novel defense to this problem that identifies poisoning sybils based on the diversity of client contributions in the distributed learning process. Unlike prior work, our system does not assume that the attackers are in the minority, requires no

auxiliary information outside of the learning process, and makes fewer assumptions about clients and their data.

In our evaluation we show that FoolsGold exceeds the capabilities of existing state of the art approaches to countering ML poisoning attacks. Our results hold for a variety of conditions, including different distributions of data, varying poisoning targets, and various attack strategies

There are multiple interesting research avenues of future work. We are currently assuming that data sources are independent. As future work, it would be interesting to study cases where multiple

data sources may collude to poison the machine learning model. Another promising avenue consists in investigating multiple calibration methodologies to detect how different provenance features may influence a change in accuracy of a particular model. Finally, we also plan to evaluate our fully untrusted model in more detail.

In this paper, we present a novel methodology for detecting and filtering poisonous data collected to train an arbitrary supervised learning model. To the best of our knowledge, this is the first defense strategy that makes use of data provenance to prevent poisoning attacks. Trusted provenance information is available in many application scenarios such as in environmental sensing or even some social media environments. We present two variations of the provenance defense for both partially trusted and fully untrusted data sets. We evaluated our partially trusted approach using two previously proposed poison data generation methods. Our experimental results show that the detection effectiveness of the proposed provenance

defense surpasses that of the baseline, thereby enabling the use of online and regularly re-trained machine learning models in adversarial environments where reliable provenance data can be obtained.

Several approaches to poison models have been proposed in the literature. Zhou et al. [19] proposed two attack models for poisoning

SVMs, as well as optimal SVM learning strategies against the proposed attack models. In contrast, the proposed provenance defense does not require a priori knowledge of the type of poison injected by adversaries. We also experimentally show that the proposed methodology is resilient against this type of poison.

Biggio et al. [6] proposed an attack to SVMs where an adversary can manipulate all features of training data by running a gradient ascent optimization problem that causes the decision boundary of the attacked model to shift to the adversary’s advantage.

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis

We evaluate our methodology against this poison attack and demonstrate its effectiveness. Other types of attacks focus on modifying uniquely the labels fed to the training model. Biggio et al. [5] study attacks where an adversary uniquely influences labels provided in the training process (a malicious annotator) and propose a kernel matrix correction defense for SVMs.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3117385/

**Motivation:** Classifying biological data into different groups is a central task of bioinformatics: for instance, to predict the function of a gene or protein, the disease state of a patient or the phenotype of an individual based on its genotype. Support Vector Machines are a wide spread approach for classifying biological data, due to their high accuracy, their ability to deal with structured data such as strings, and the ease to integrate various types of data. However, it is unclear how to correct for confounding factors such as population structure, age or gender or experimental conditions in Support Vector Machine classification.

Support Vector Machine classifier that can correct the prediction for observed confounding factors. This is achieved by minimizing the statistical dependence between the classifier and the confounding factors. We prove that this formulation can be transformed into a standard Support Vector Machine with rescaled input data.

Similarly, [12] present an attack that targets the labels input into the training system and a threshold based methodology to detect poison that relies on a Kappa statistic. Like RONI, this method requires that trusted, unpoisoned data is available. Finally, none of these approaches take into consideration provenance information associated with data points and labels during the training process to detect poison attacks.

Experiments:

Effect of the number of data points contributed per device on the accuracy and average improvement of the proposed method.

Experiments on measuring accuracy against provenance defense, no defense and perfect defense and baseline defense

Performance of logistic regression model trained on all data, including poison (b) Performance of logistic regression model trained with poison data removed from training set but not the evaluation set (c) Performance of logistic regression model trained and evaluated with poison data removed by our provenance method.

-------------------------------------------

Many provenance frameworks have been proposed in the literature [1, 2, 8, 10, 17, 18] to ensure the lineage of data can be tracked for accountability purposes. These approaches focus on cryptographically preserving the history of data, non-fabrication and non-repudiation. However, to the best of our knowledge this is the first approach to use provenance information as an integral component to defend against poisoning attacks.

The use of machine learning systems in critical applications has drastically increased and with it the number of efforts to identify security vulnerabilities and defenses.

We have focus on poisoning attacks, a.k.a. causative attacks, that target the training stage of the model. The closest related work is Reject On Negative Impact (RONI) methodology proposed by Nelson et al. in [13] and further enhanced in [15], where a calibration methodology to evaluate the performance of a model was included. These approaches assume the existence of a partially trusted data set. Our approach differs from these methodologies in that it makes use of provenance information that contains contextual cues to expedite the evaluation of untrusted data points.

In some scenarios, it is difficult or even infeasible to obtain a partially trusted data set due to cost associated with manual data verification, such as paying annotators to verify labels, and real-time requirements that preclude data verification. To address these scenarios, we present a provenance based poison detection mechanism that works even if all data collected for re-training is untrusted. To apply our method to fully untrusted data sets, we propose the following procedure.

(1) Segment the data by signature according to the selected provenance feature.

(2) Split the data set randomly into a training portion and an evaluation portion.

(3) For each signature in the selected provenance feature:

(a) Train two models–one with all of the training data and one with the corresponding segment in the training data removed;

(b) Evaluate both models on the evaluation set with the corresponding segment removed;

(c) Permanently remove the segments from both the training and evaluation set if the model trained without it performed better.

Canada authors

https://arxiv.org/pdf/1808.04866.pdf

Mitigating Sybils in Federated Learning Poisoning

Clement Fung

University of British Columbia

cfung1@cs.ubc.ca

Chris J.M. Yoon

University of British Columbia

yoon@alumni.ubc.ca

Ivan Beschastnikh

University of British Columbia

bestchai@cs.ubc.ca

In this paper we first evaluate the vulnerability of federated learning to sybil-based poisoning attacks. We then describe FoolsGold , a novel defense to this problem that identifies poisoning sybils based on the diversity of client contributions in the distributed learning process. Unlike prior work, our system does not assume that the attackers are in the minority, requires no auxiliary information outside of the learning process, and makes fewer assumptions about clients and their data.

In our evaluation we show that FoolsGold exceeds the capabilities of existing state of the art approaches to countering ML poisoning attacks. Our results hold for a variety of conditions, including different distributions of data, varying poisoning targets, and various attack strategies.