

Poison Ivy

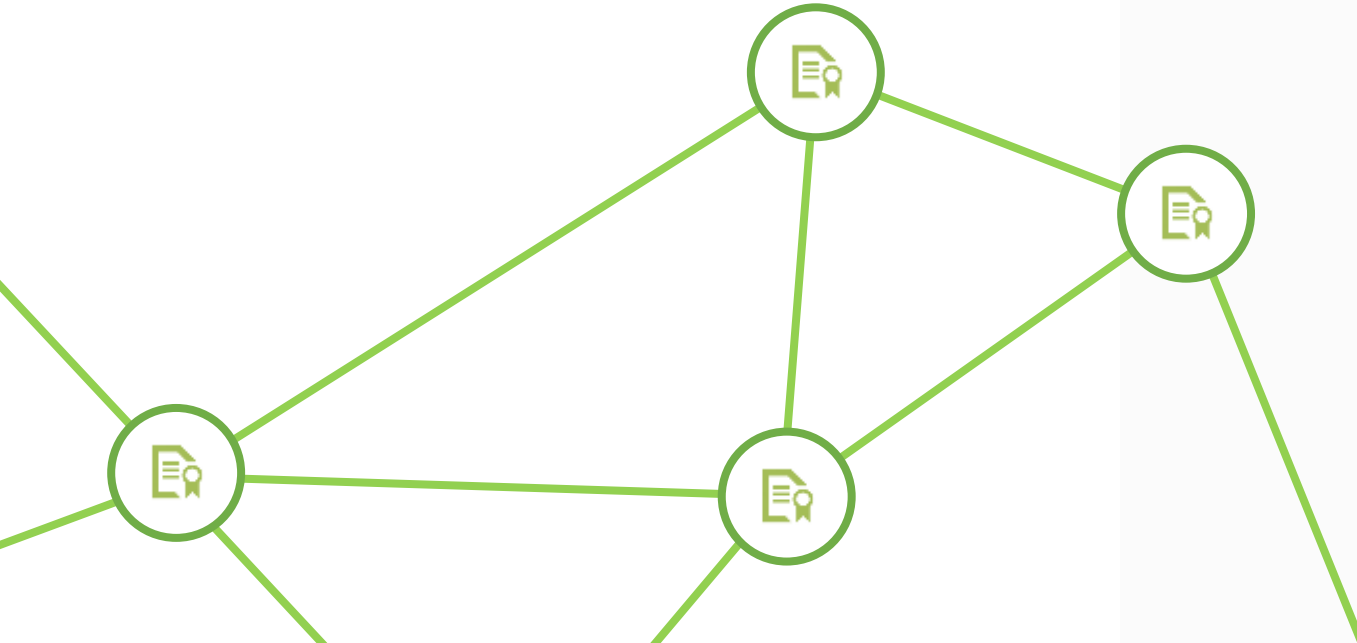
Wie man mit gezieltem Vergiften
KI-Datensätze manipuliert und
was man daraus über Angriff
und Verteidigung lernt

Präsentation Mirko Ross, asvin GmbH



Poison Ivy (2021), DataChainSec (2022-2024), AntiDot (2022)

Preventing and Detecting Data-based Backdoors



Um was es geht

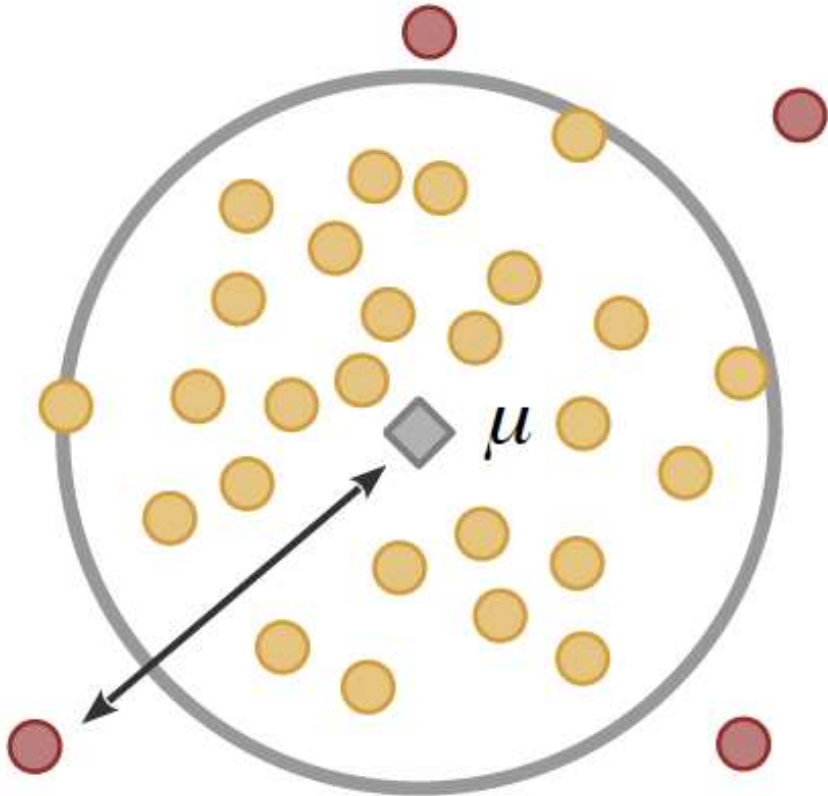
- **Schutz vor Angriffe auf lernende Systeme**
Sicherheit des gesamten Anwendungssystem
Problem: unzureichender Schutz der KI-Modelle und Daten
- **Datenbasierte Backdoors in KI-Anwendungen**
Schleichende Manipulation des KI-Modells
Beeinflussung der Vorhersagen des lernenden Systems
Gefährdung der Daten und Modelle als wichtigstes Gut
- **Datenlieferketten** deren Angriffsoberflächen



Das Grundprinzip: Daten kaputt, KI kaputt

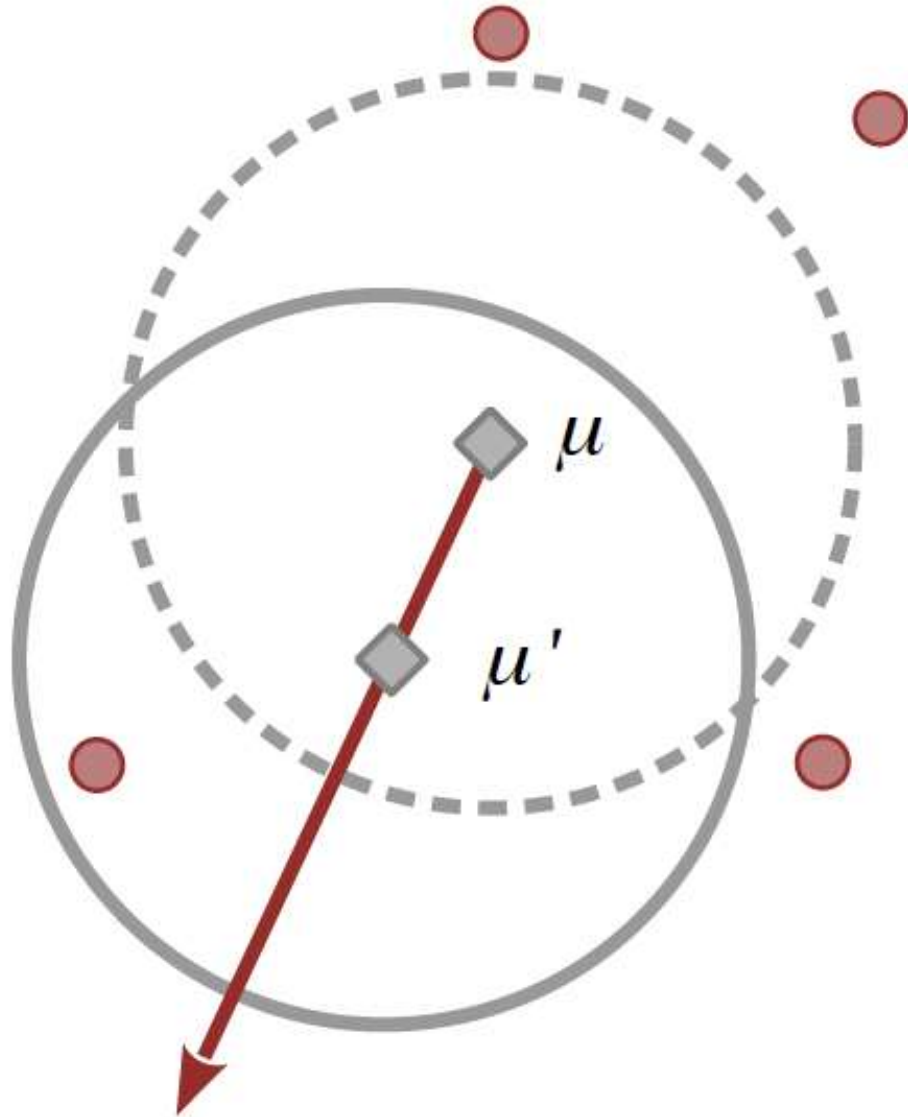


Das Grundprinzip von Data Poisoning Angriffen



- **Data-Poisoning gegen ML**
Einbringen von manipulierten Trainingsdaten
Beeinflussung der Vorhersagen
- **Stellschrauben**
Änderung des Labels
Änderung der Eingaben

Das Grundprinzip von Data Poisoning Angriffen

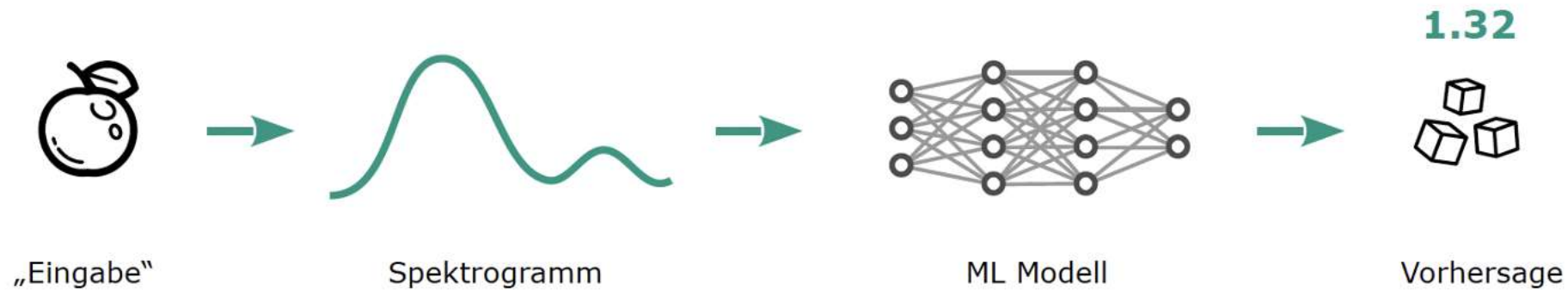


Aussagen des Modells
werden in eine vom
Angreifer gewünschte
Richtung verschoben

Poison IVY: Monitoring Lebensmittel Haltbarkeit



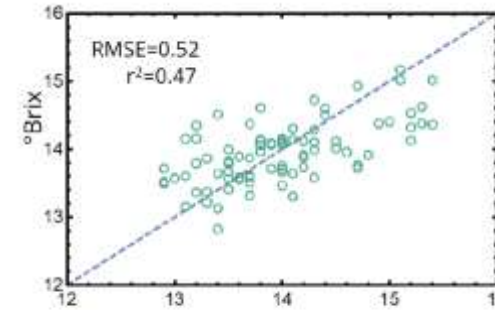
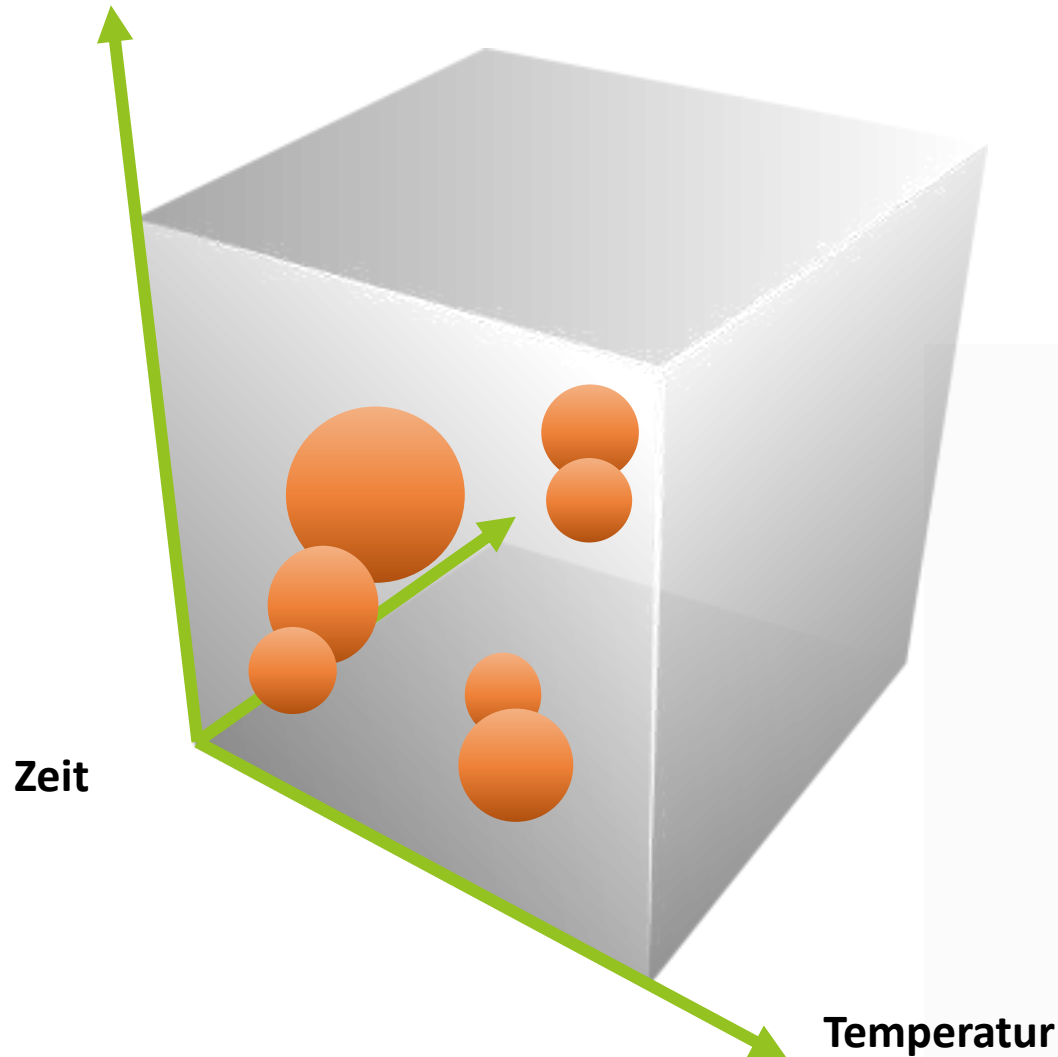
Data Poisoning praktischen Anwendungssystem



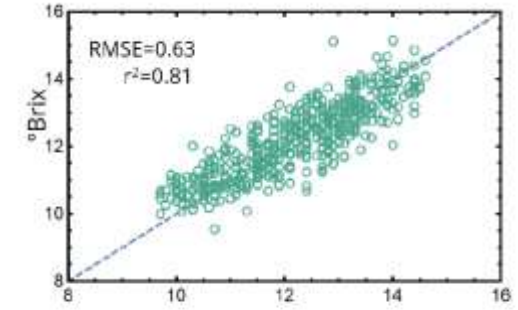
- **Monitoring der Qualität von Nahrungsmitteln**
Spektroskopie von Früchten, Temperaturdaten
- **Vorhersage der Qualität**
Zuckergehalt, Bakterienbelastung
- **ML** über Lineare / Logistische Regression / Support Vector Machine
Überwachtes ML – (keine neuronalen Netze / Deep Learning)

Machine Learning: Lineare Regression zur Ermittlung der Haltbarkeit

BRIX (Süsse)



(a) Messungen während der Feldstudie



(b) Initiale Messungen

Gewünschtes Ergebnis:

Vorhersage Pilz/Bakterien Wachstum über einen Zeitraum zur Ermittlung der Haltbarkeit

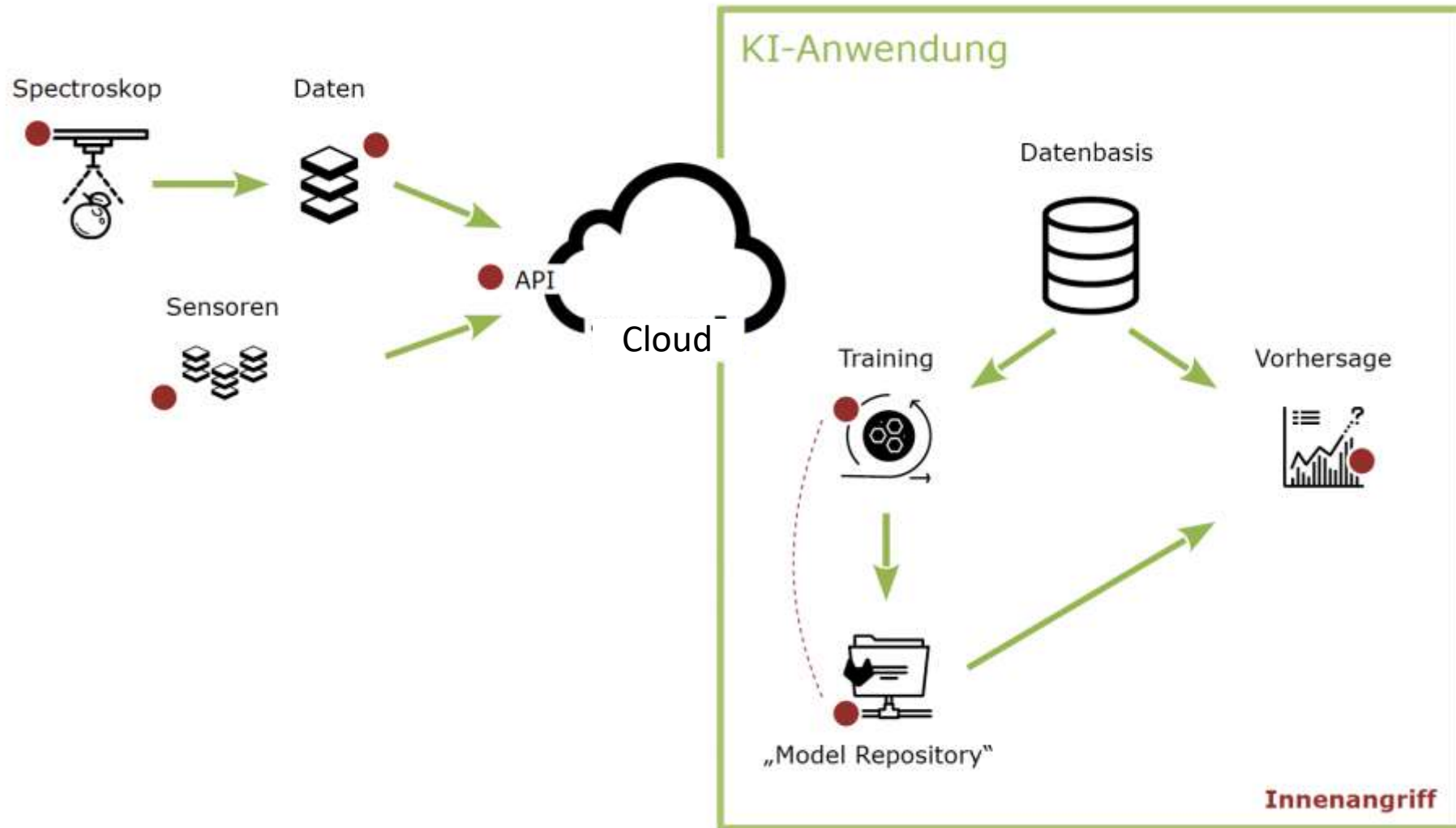
oder

Digitaler Zwilling einer Erdbeere

ATLAS Matrix (MITRE): AI Threat Landscape

Reconnaissance	Resource Development	Initial Access	ML Model Access	Execution	Persistence	Defense Evasion	Discovery	Collection	ML Attack Staging	Exfiltration	Impact
5 techniques	7 techniques	2 techniques	4 techniques	1 technique	2 techniques	1 technique	3 techniques	2 techniques	4 techniques	2 techniques	6 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts Obtain Capabilities	ML Supply Chain Compromise Valid Accounts	ML Model Inference API Access ML-Enabled Product or Service Physical Environment Access Full ML Model Access	User Execution	Poison Training Data Backdoor ML Model	Evade ML Model	Discover ML Model Ontology Discover ML Model Family Discover ML Artifacts	ML Artifact Collection Data from Information Repositories	Create Proxy ML Model Backdoor ML Model Verify Attack Craft Adversarial Data	Exfiltration via ML Inference API Exfiltration via Cyber Means	Evade ML Model Denial of ML Service Spamming ML System with Chaff Data Erode ML Model Integrity Cost Harvesting ML Intellectual Property Theft

Threat Landscape in der KI Anwendung



Beispiel Beeinflussung Lineare Regression (LASSO)

- **Manipulation der Eingabedaten**

Manipulation am Sensor, Man-in-the-Middle, Cloud

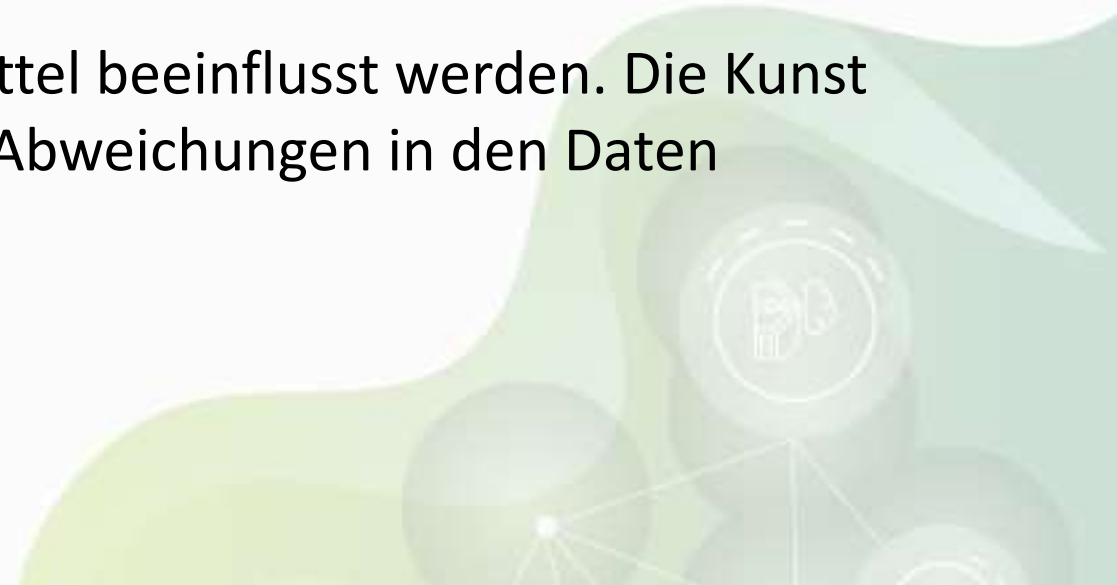
Verhaltensänderung bei unterschiedlichen Manipulationsraten (bis zu 25%)*

- **Fehlerraten**

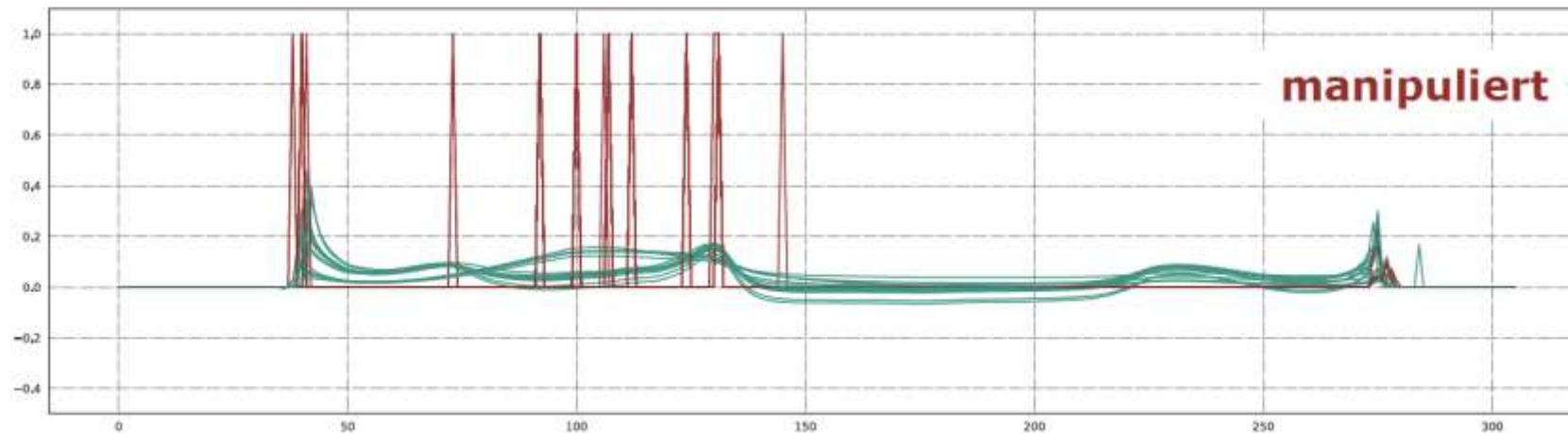
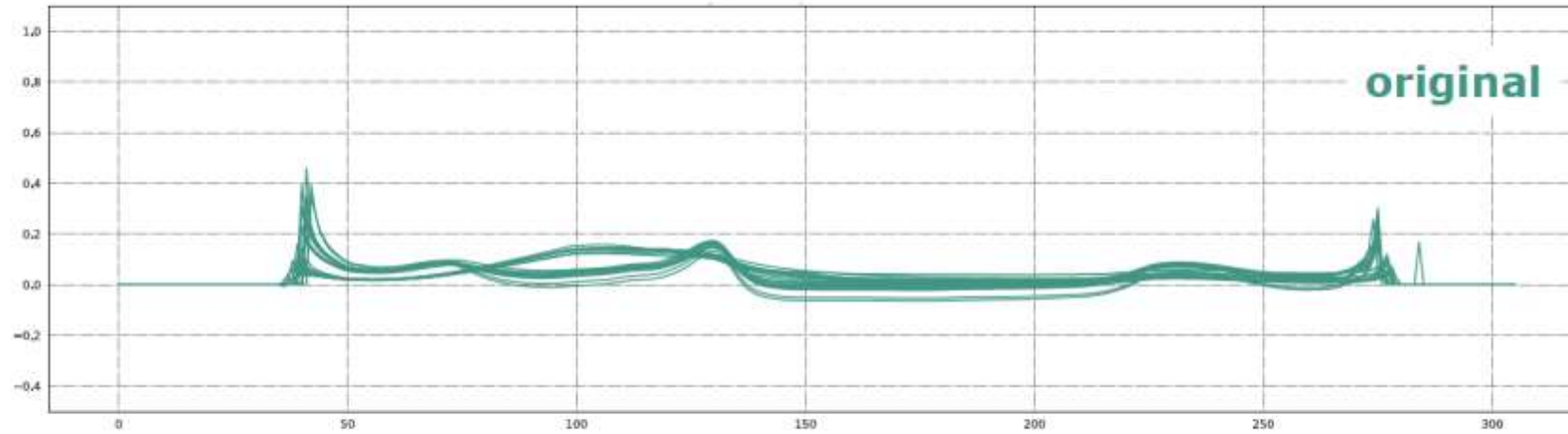
Normal gelerntes Model → 1.07

Verunreinigtes Model → 3.64

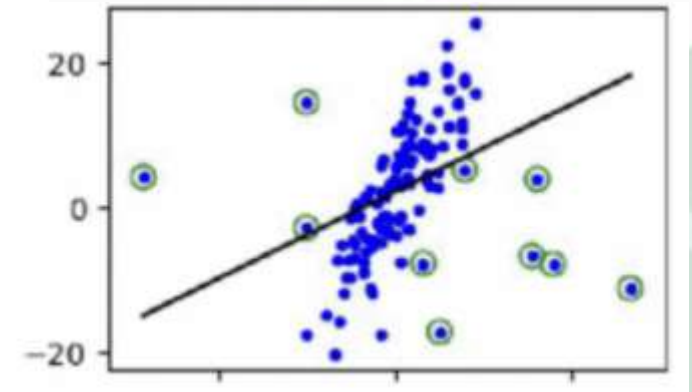
Fazit: Lineare Regression kann mit einfachen Mitteln beeinflusst werden. Die Kunst des unentdeckten Angreifers besteht darin, die Abweichungen in den Daten möglichst unauffällig zu streuen.



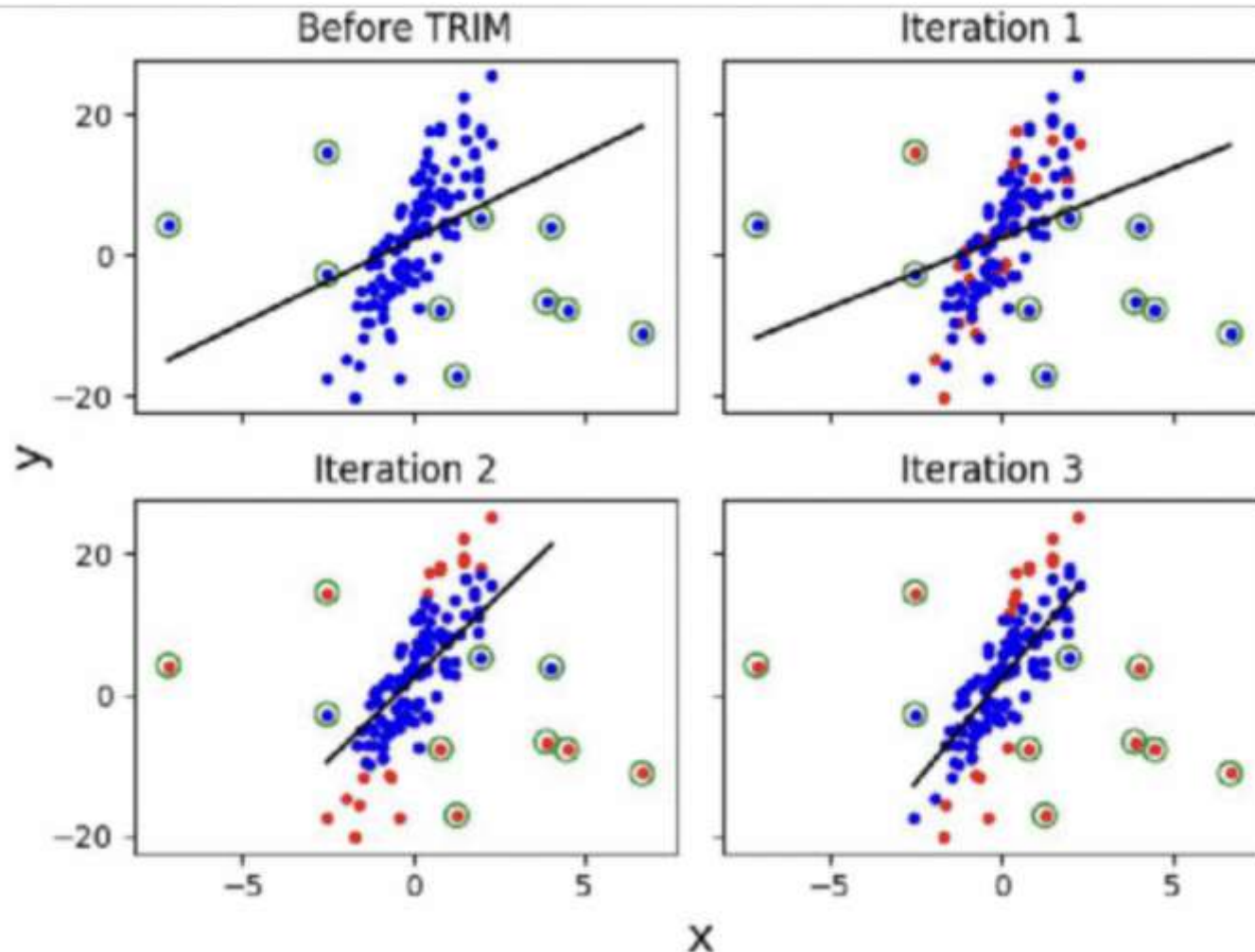
Angriff Trainingdaten auf ML Lineare Regressionn (LASSO)



Verunreinigte Trainingsdaten
Mit bewusst eingestreuten
Abweichungen (Kreis)



Abwehr: Noise-Resilient-Regression / TRIM Algorithmen

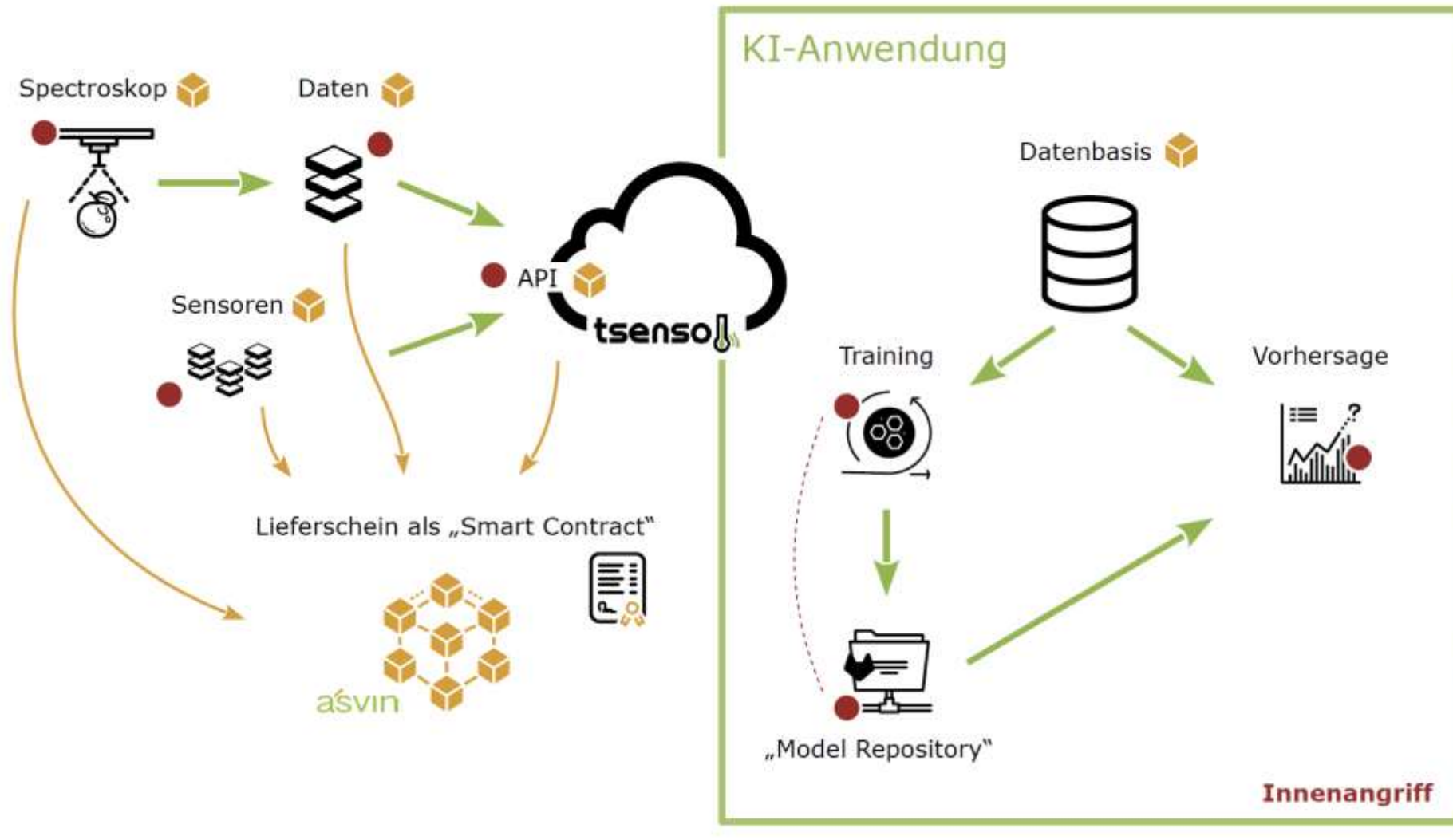


Noise-Resilient-Regression / TRIM

Abweichungen erkennen (Kreise) und
das Trainingsset bereinigen (rot)

(TRIM nach JAGIELSKI, 2018)

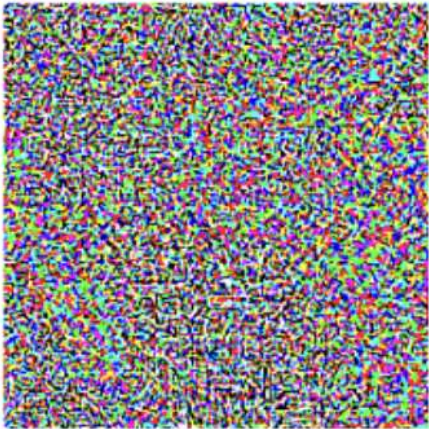
Weitere Abwehrstrategien: Absicherung der Datenlieferketten



Angriff auf Deep Learning / Neuronale Netze



Der „Klassiker“: Adversarial Attack auf Bilderkennung*

	$+ .007 \times$		$=$	
x		$\text{sign}(\nabla_x J(\theta, x, y))$		$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$
“panda”		“nematode”		“gibbon”
57.7% confidence		8.2% confidence		99.3 % confidence

Eingabedaten (Bildinformation mit Pixel) enthält zu viele Parameter und damit zu viele Entscheidungsmöglichkeiten bei der Klassifikation)

<https://github.com/bethgelab/foolbox>

GOODFELLOW 2014: <https://arxiv.org/abs/1412.6572>

Eine Abwehrmethode: Reduktion der Bildinformation

Original



6 bit color



4 bit color



3 bit color



2 bit color



Spatial smoothing



Eingabedaten (Bildinformation mit Pixel) enthält maßgebliche Parameter und damit weniger Entscheidungsmöglichkeiten bei der Klassifikation)

Einer geht noch: Angriff auf Objekterkennung in Videos



Fooling Neural Networks in the Real World

labsix

rifle

shield, buck

revolver, si



Mirko Ross

CEO

m.ross@asvin.io

+49 1773554161

Asvin GmbH

Stuttgart

www.asvin.io

contact@asvin.io

