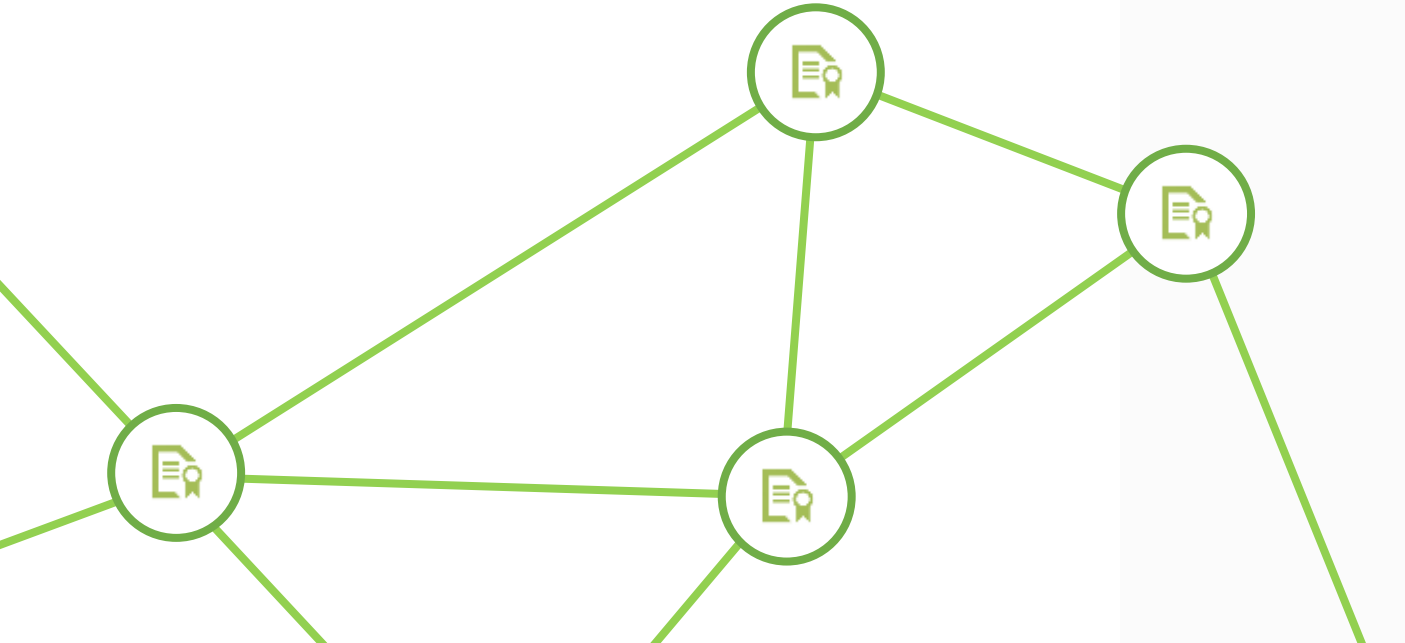
Poison Ivy
Wie man mit gezieltem Vergiften KI-Datensätze manipuliert und was man daraus über Angriff und Verteidigung lernt

Präsentation Mirko Ross, asvin GmbH

Poison Ivy (2021), **DataChainSec** (2022-2024), **AntiDot** (2022)
Preventing and Detecting Data-based Backdoors

tsensol asvin KIT Karlsruhe Institute of Technology



Um was es geht

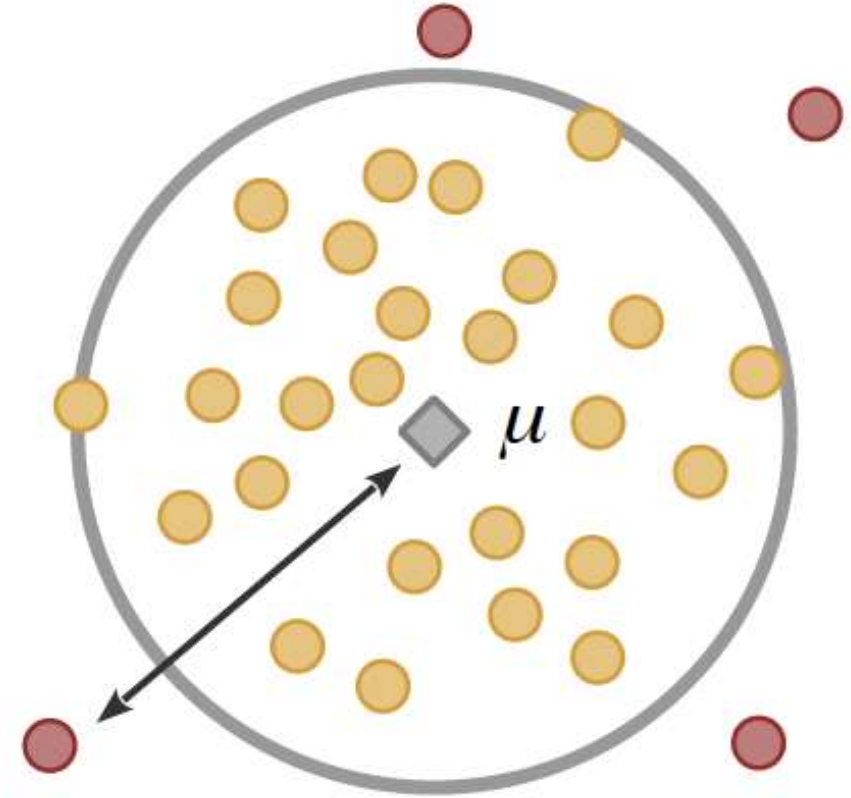
- **Schutz vor Angriffe auf lernende Systeme**
  Sicherheit des gesamten Anwendungssystem
  Problem: unzureichender Schutz der KI-Modelle und Daten

- **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

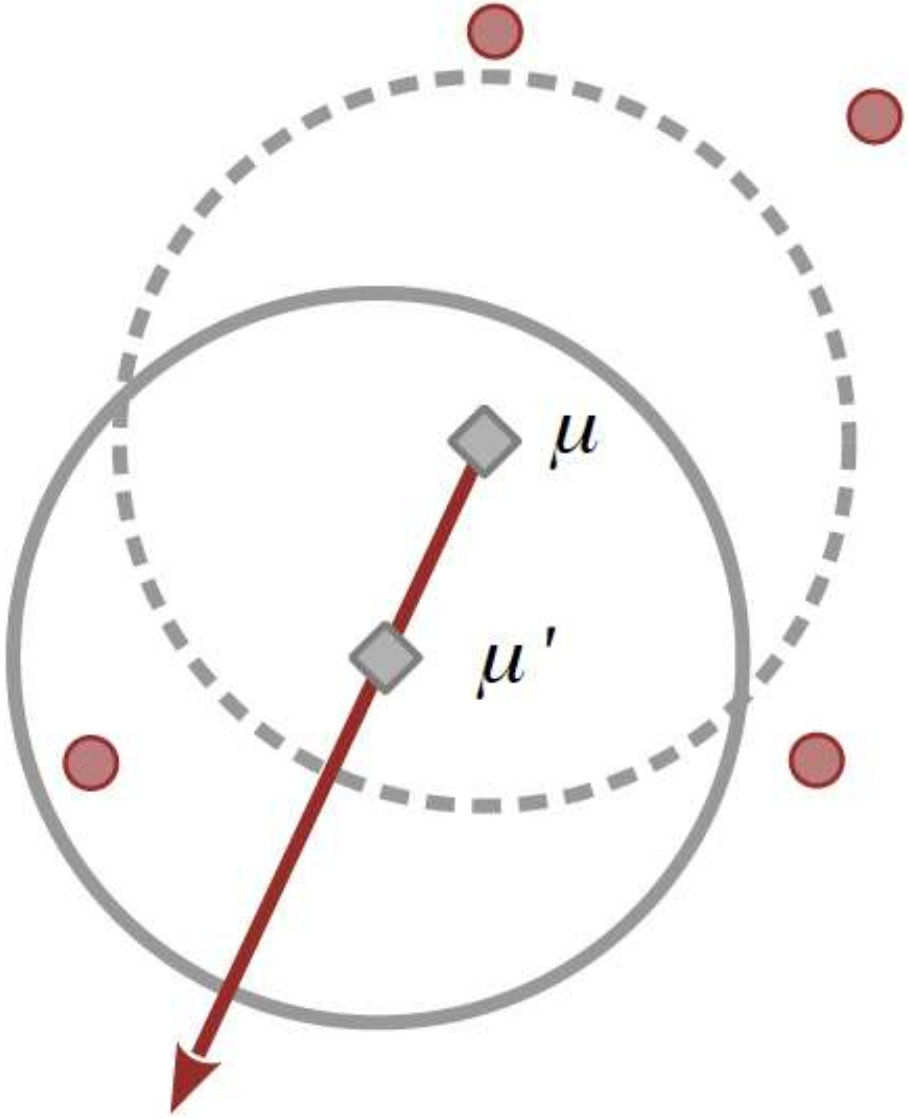


- **Data-Poisoning gegen ML**
  Einbringen von manipulierten Trainingsdaten

  Beeinflussung der Vorhersagen

- **Stellschrauben**
  Änderung des Labels
  Änderung der Eingaben

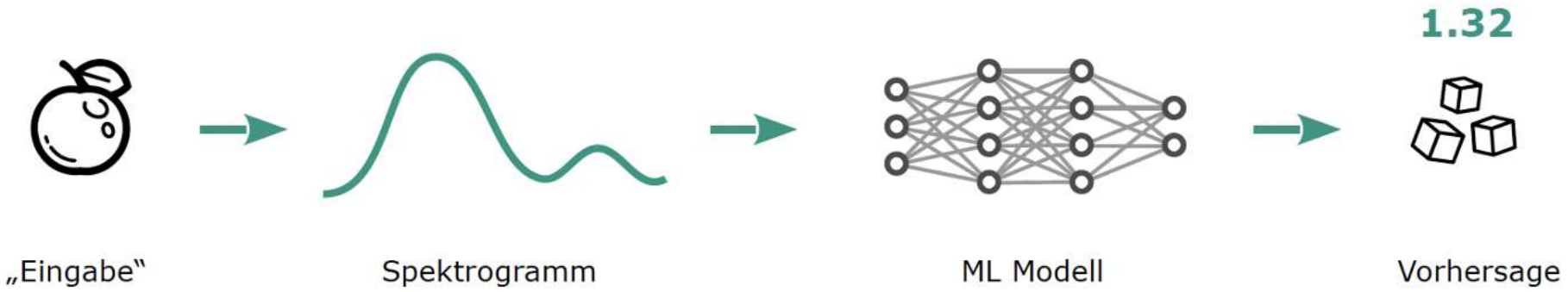
Das Grundprinzip von Data Poisoning Angriffen



Aussagen des Modells werden in eine vom Angreifer gewünschte Richtung verschoben

Poison IVY: Monitoring Lebensmittel Haltbarkeit

Data Poisoning praktischen Anwendungssystem

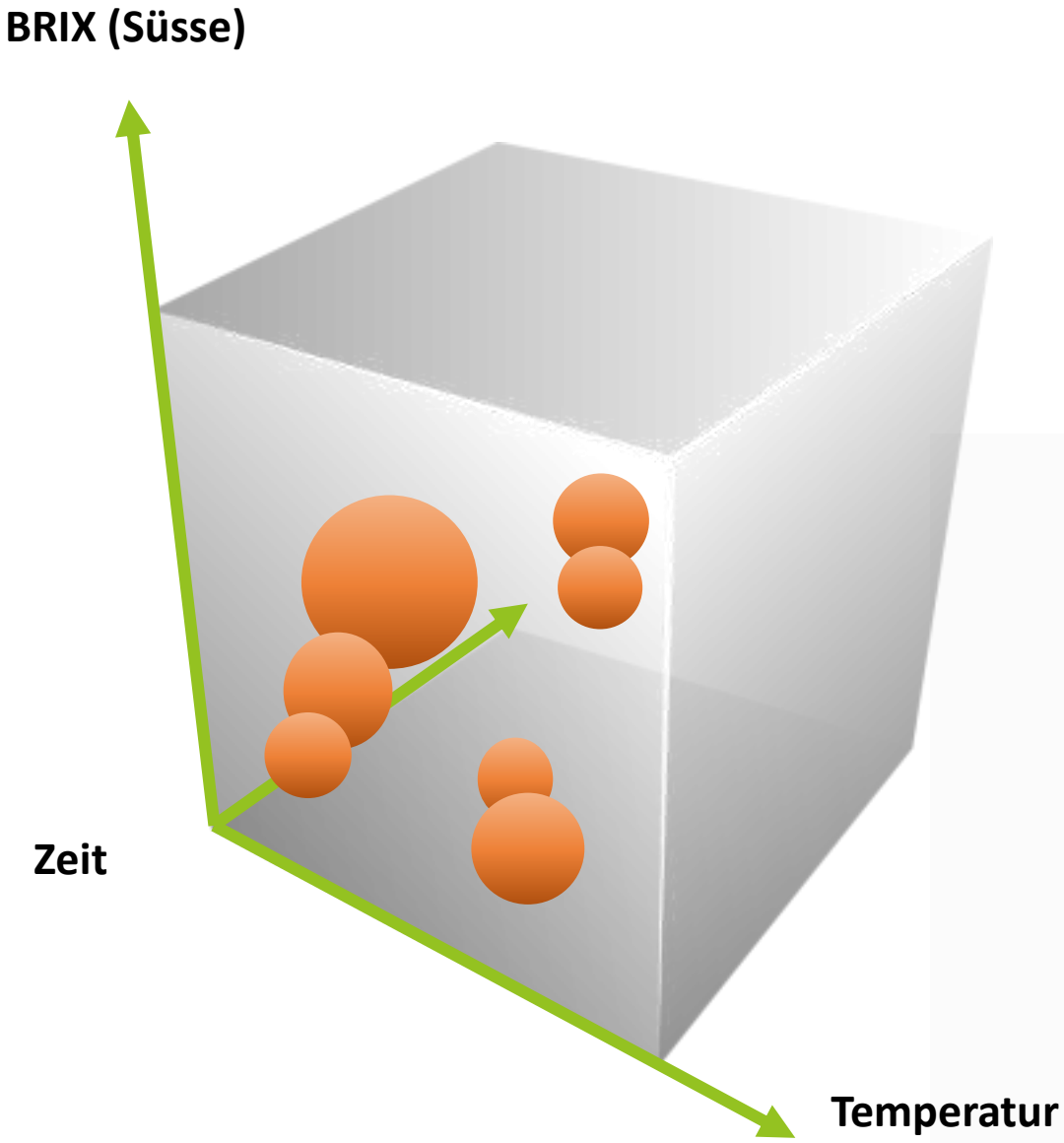


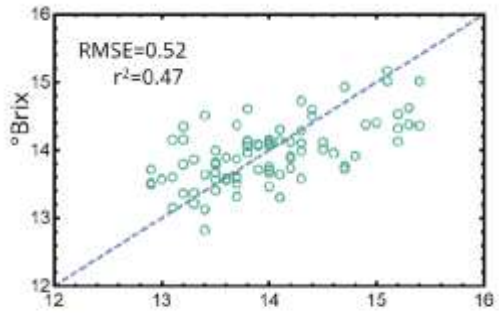
- **Monitoring der Qualität von Nahrungsmitteln**
  Spektroskopie von Früchten, Temperaturdaten

- **Vorhersage der Qualität**
  Zuckergehalt, Bakterienbelastung

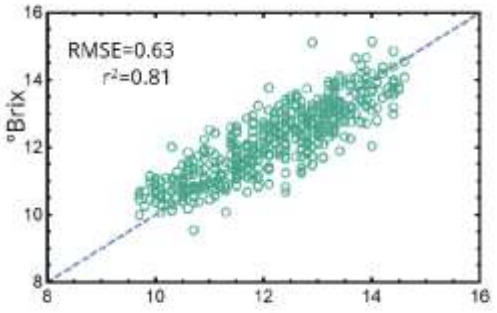
- **ML** über Lineare / Logistische Regression / Support Vector Machine
  Überwachtes ML – (keine neuronalen Netze / Deep Learning)

Machine Learning: Lineare Regression zur Ermittlung der Haltbarkeit





(a) Messungen während der Feldstudie



(b) Initiale Messungen

Gewünschtes Ergebnis:

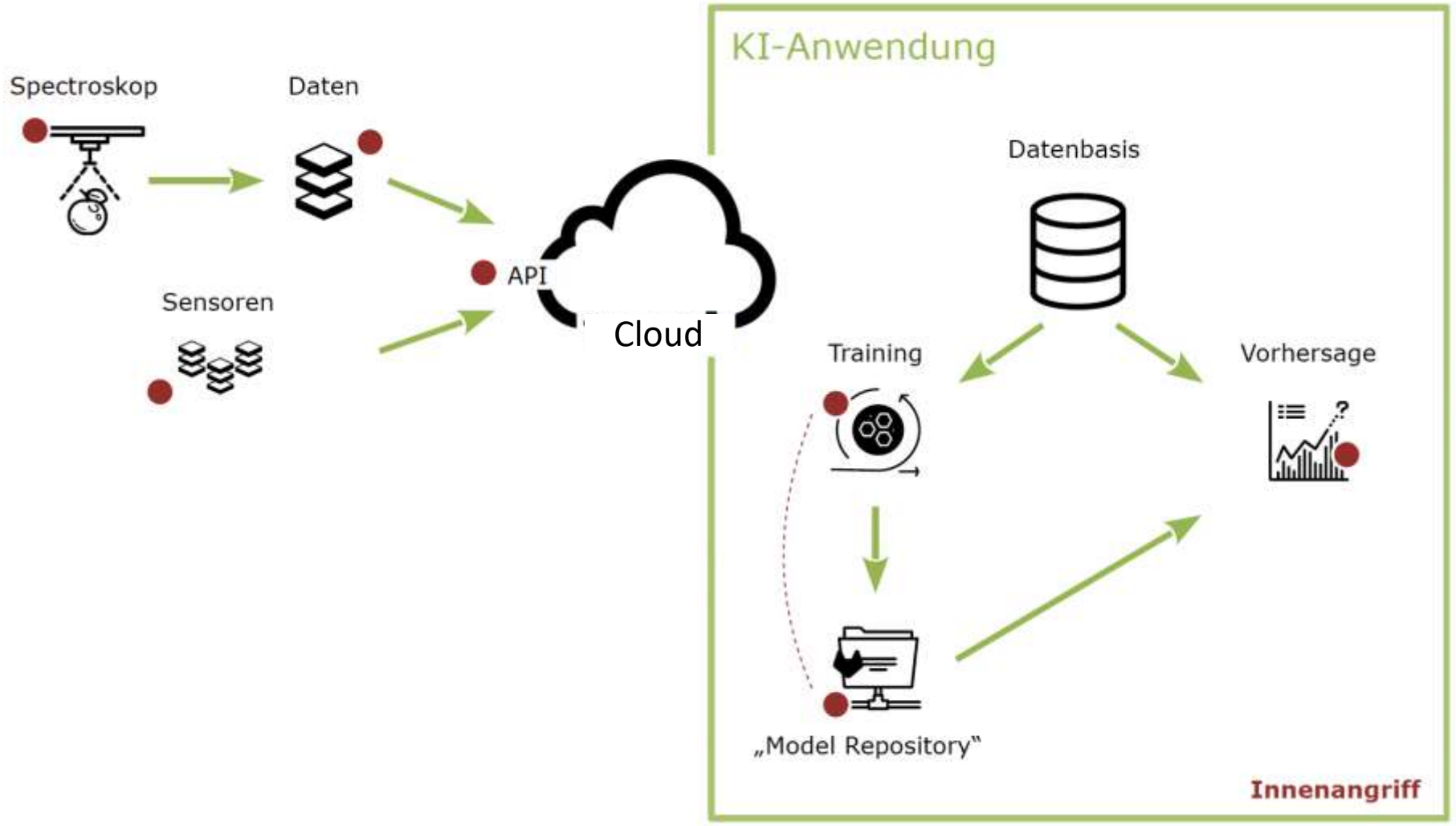
Vorhersage Pilz/Bakterien Wachstum über einen Zeitraum zur Ermittlung der Haltbarkeit

oder

Digitaler Zwilling einer Erdbeere

ALTLAS Matrix (MITRE): AI Threat Landscape

Threat Landscape in der KI Anwendung



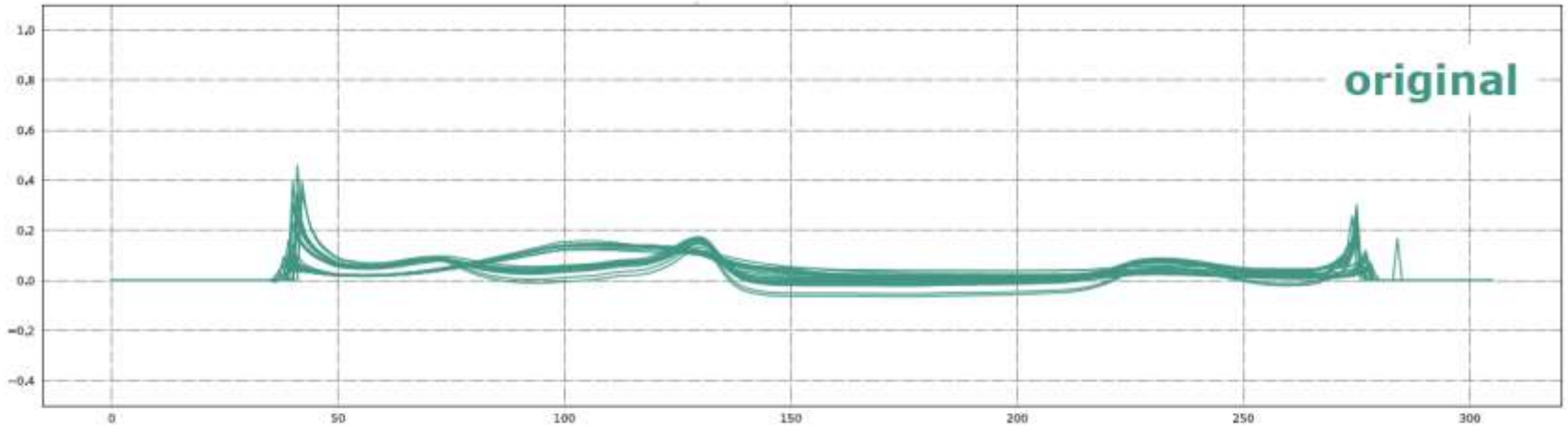
Beispiel Beeinflussung Lineare Regression (LASSO)

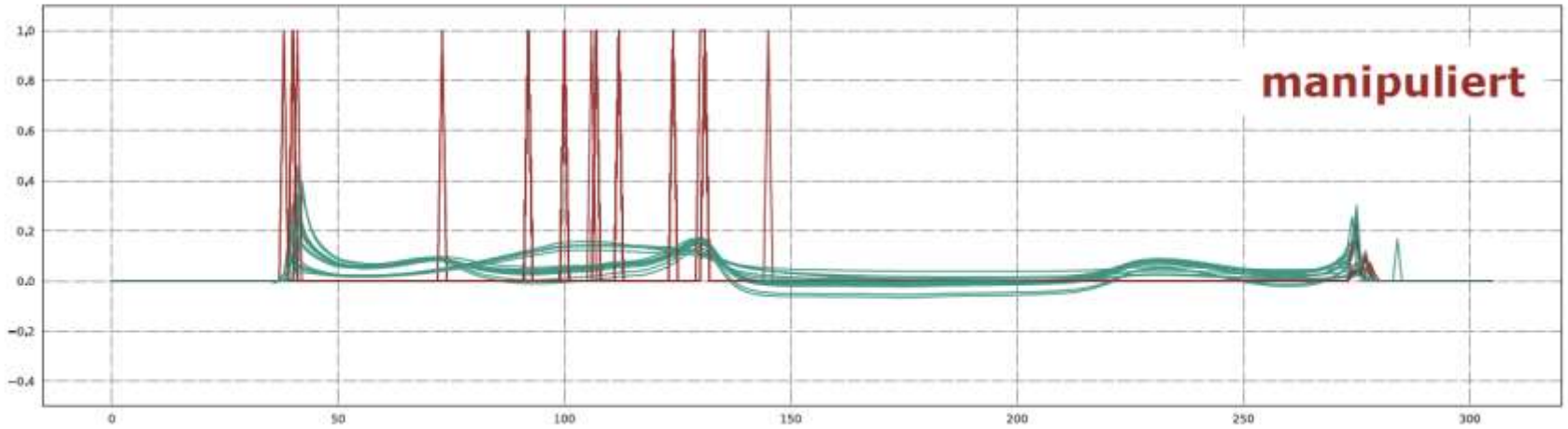
- **Manipulation der Eingabedaten**
  Manipulation am Sensor, Man-in-the-Middle, Cloud
  Verhaltensänderung bei unterschiedlichen Manipulationsraten (bis zu 25%)*

- **Fehlerraten**
  Normal gelerntes Model –> 1.07
  Verunreinigtes Model -> 3.64

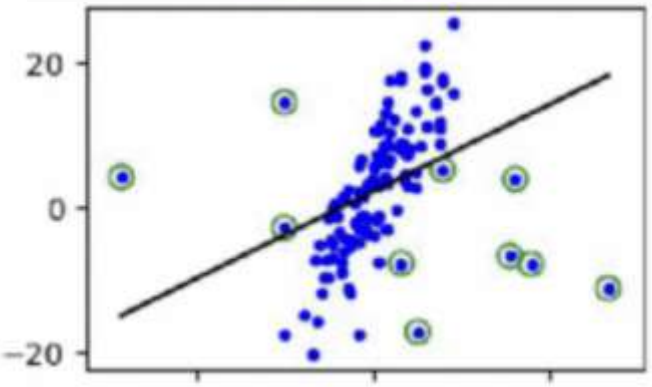
Fazit: Lineare Regression kann mit einfachen Mittel beeinflusst werden. Die Kunst des unentdeckten Angreifers besteht darin, die Abweichungen in den Daten möglichst unauffällig zu streuen.

Angriff Trainingdaten auf ML Lineare Regressionn (LASSO)

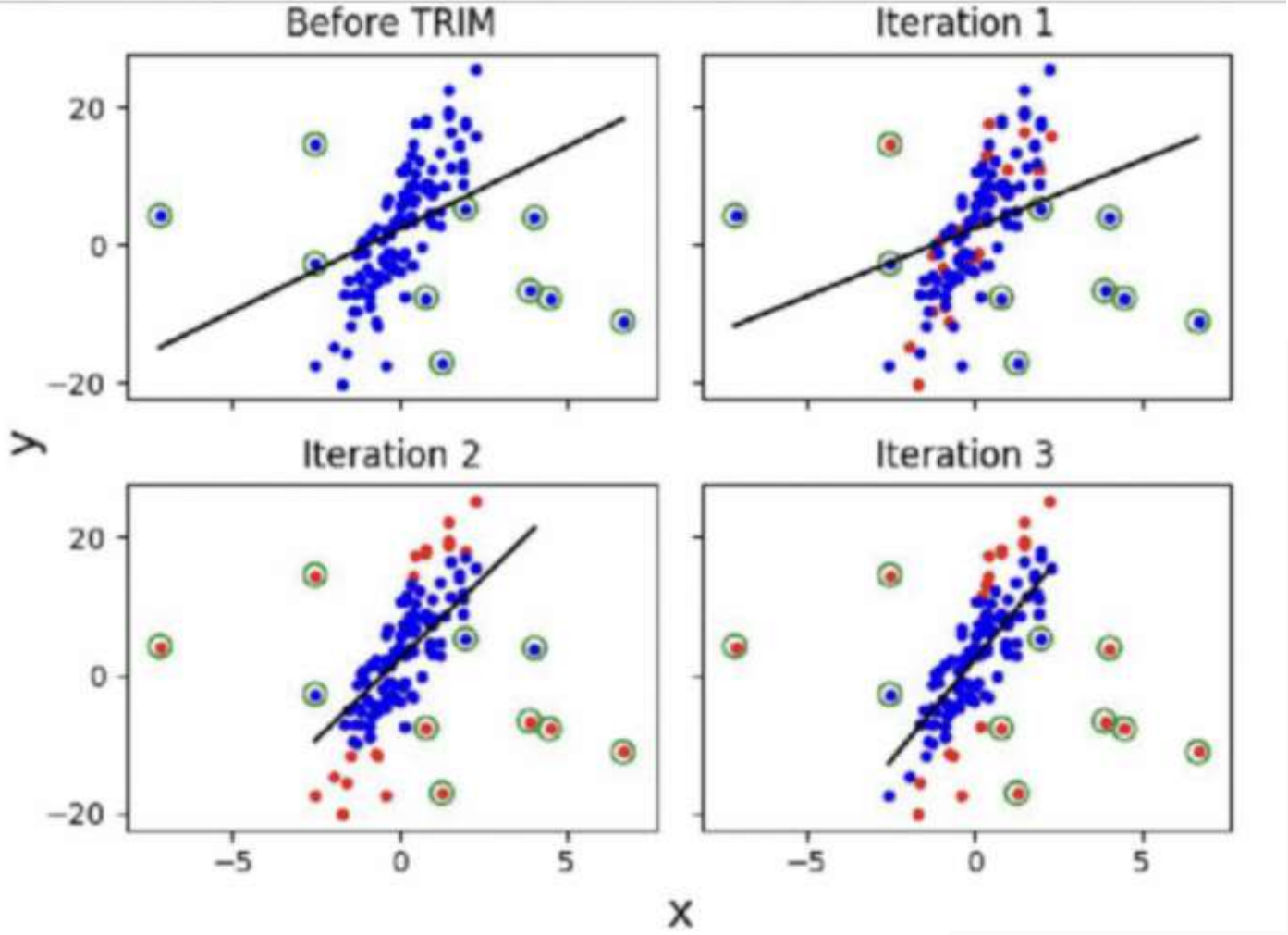




Verunreinigte Trainingsdaten
Mit bewusst eingestreuten Abweichungen (Kreis)



Abwehr: Noise-Resillient-Regression / TRIM Algorithmen



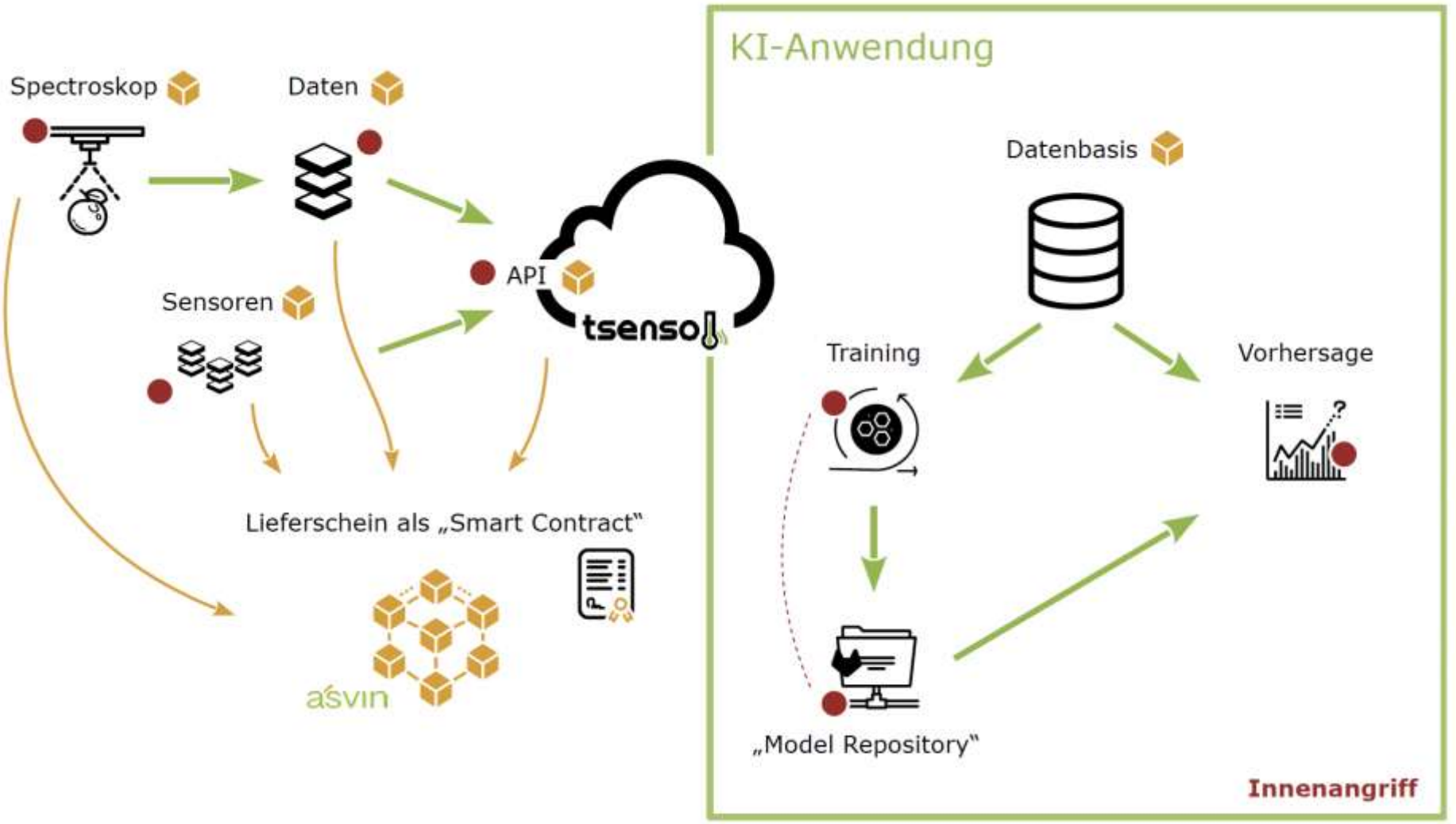
Noise-Ressilient-Regression / TRIM

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

(TRIM nach JAGIELSKI, 2018)

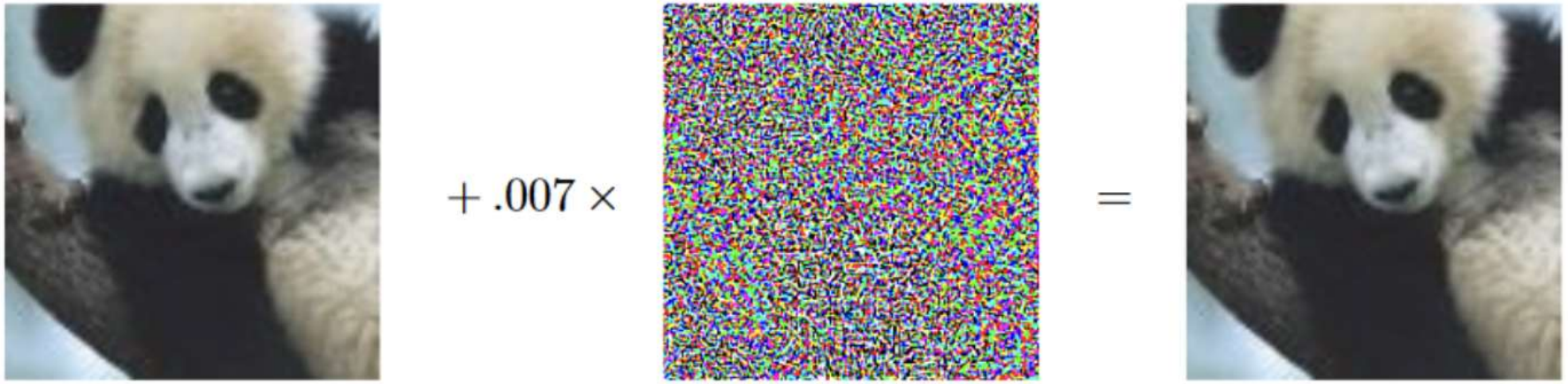
*https://ieeexplore.ieee.org/document/8418594

Weitere Abwehrstrategien: Absicherung der Datenlieferketten



Angriff auf Deep Learning / Neuronale Netze

Der „Klassiker“: Adversarial Attack auf Bilderkennung*



$\boldsymbol{x}$

“panda”
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

“nematode”
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

“gibbon”
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

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

Mirko Ross
CEO
m.ross@asvin.io
+49 1773554161

Asvin GmbH
Stuttgart
www.asvin.io
contact@asvin.io