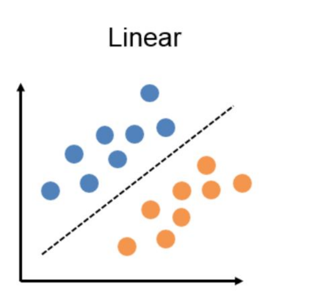
Explaining Vulnerabilities to Adversarial Machine Learning Through Visual Analytics

Yuxin Ma, Tiankai Xie, Jundong Li, Ross Maciejewski

Vulnerabilities in Machine Learning

Filtering Spam Emails

Training Stage



Spam/Non spam
Database

Testing Stage

New
Labels

Unlabeled

The Contribution of This Paper is:

- A visual analytics framework that supports the examination, creation, and exploration of adversarial machine learning attacks;

- A visual representation of model vulnerability that reveals the impact of adversarial attacks in terms of model performance, instance attributes, feature distributions, and local structures.

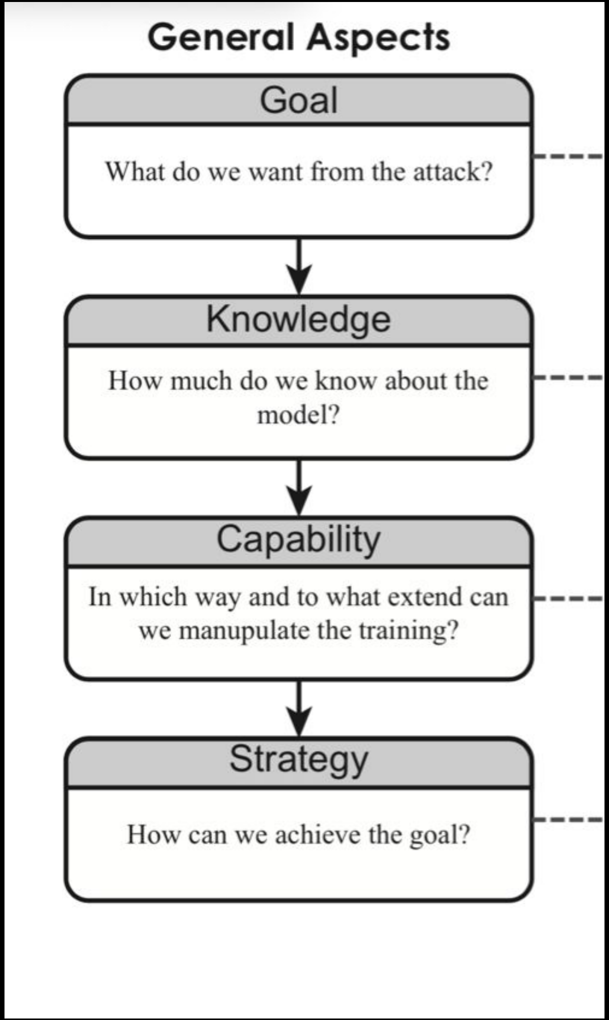
The four main features of an adversary are the adversary's ...

- **Goal**
  - Targeted attack
  - Reliability attack

- Knowledge

- Capability
  - Poisoning attack
  - Evasion attack

- Strategy

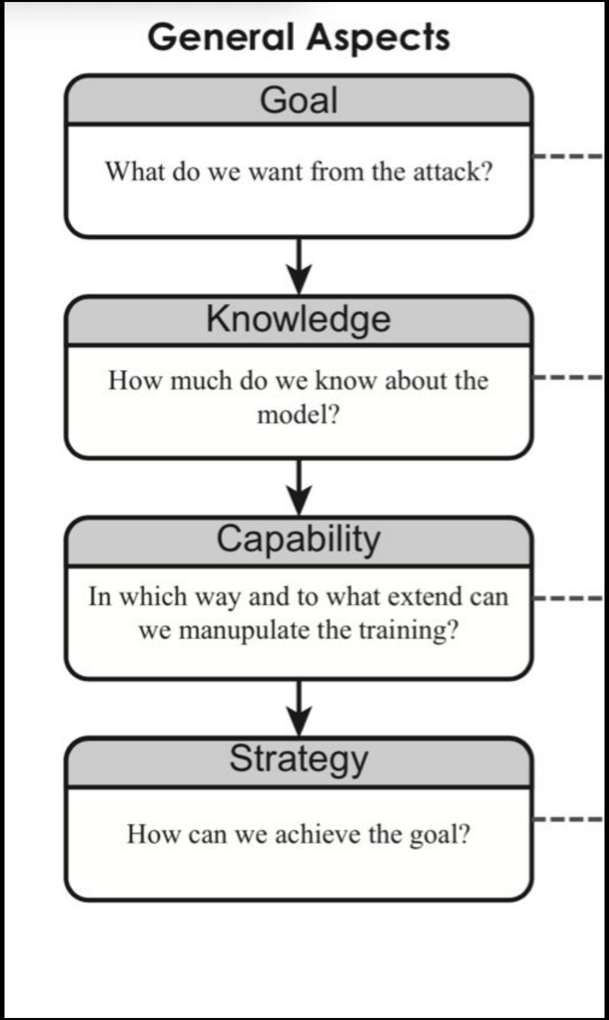


The four main features of an adversary are the adversary's ...

- Goal
  - Targeted attack
  - Reliability attack

- **Knowledge**
  - **Black box**
  - **White box**

- Capability
- Strategy



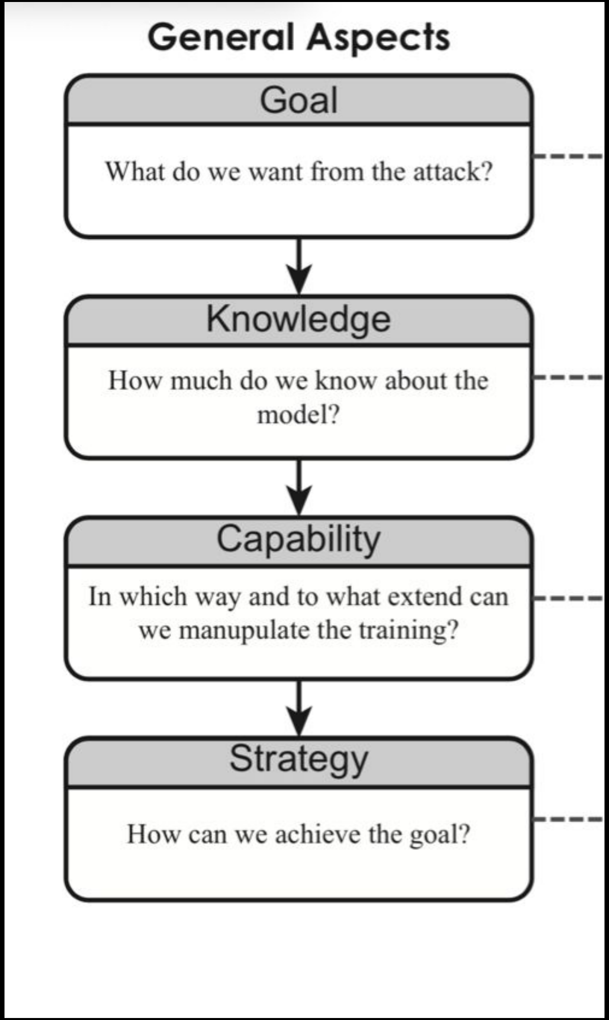
The four main features of an adversary are the adversary’s ...

- Goal
  - Targeted attack
  - Reliability attack

- Knowledge

- **Capability**
  - **Poisoning attack**
  - **Evasion attack**

- Strategy



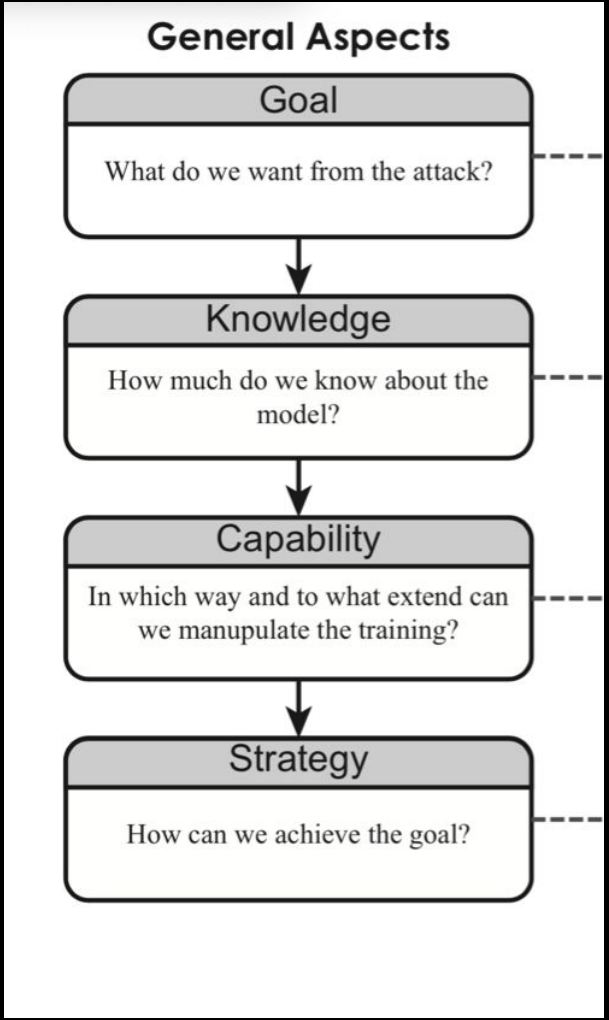
The four main features of an adversary are the adversary's ...

- Goal
  - Targeted attack
  - Reliability attack

- Knowledge

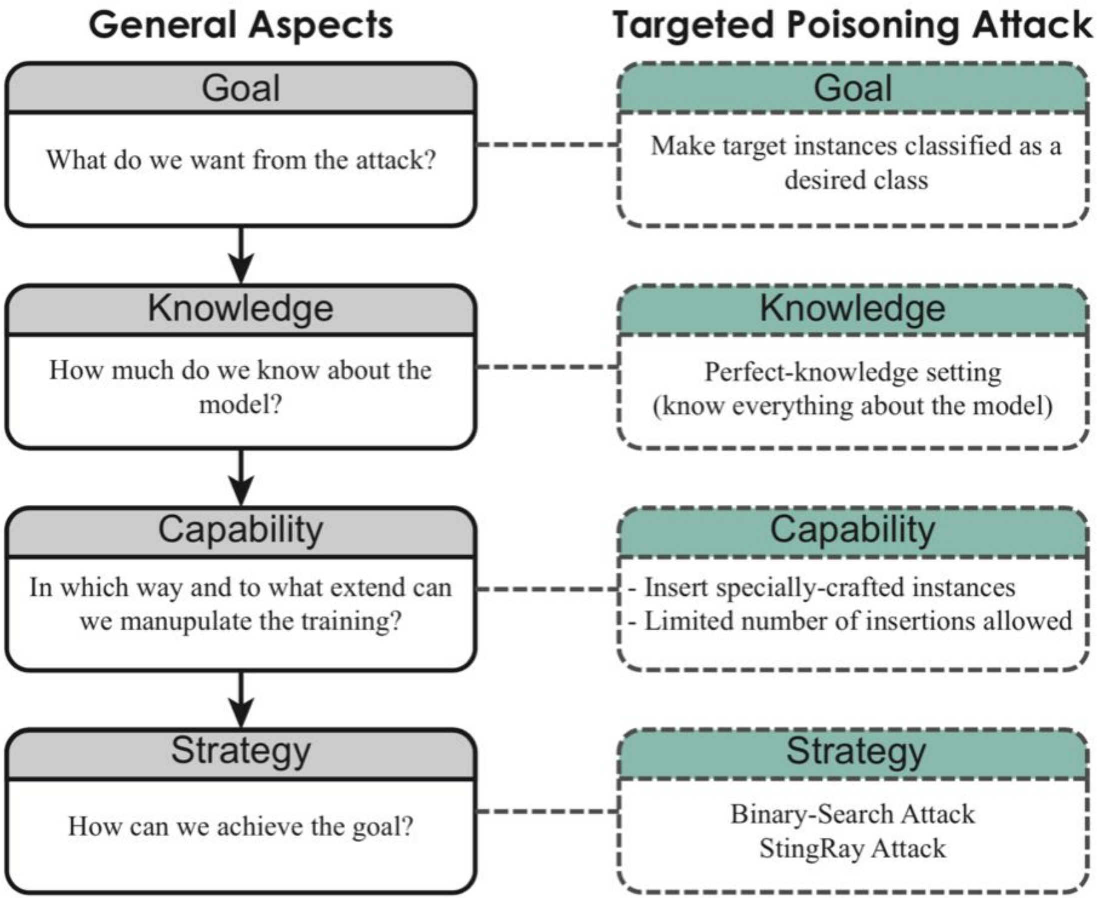
- Capability
  - Poisoning attack
  - Evasion attack

- **Strategy**

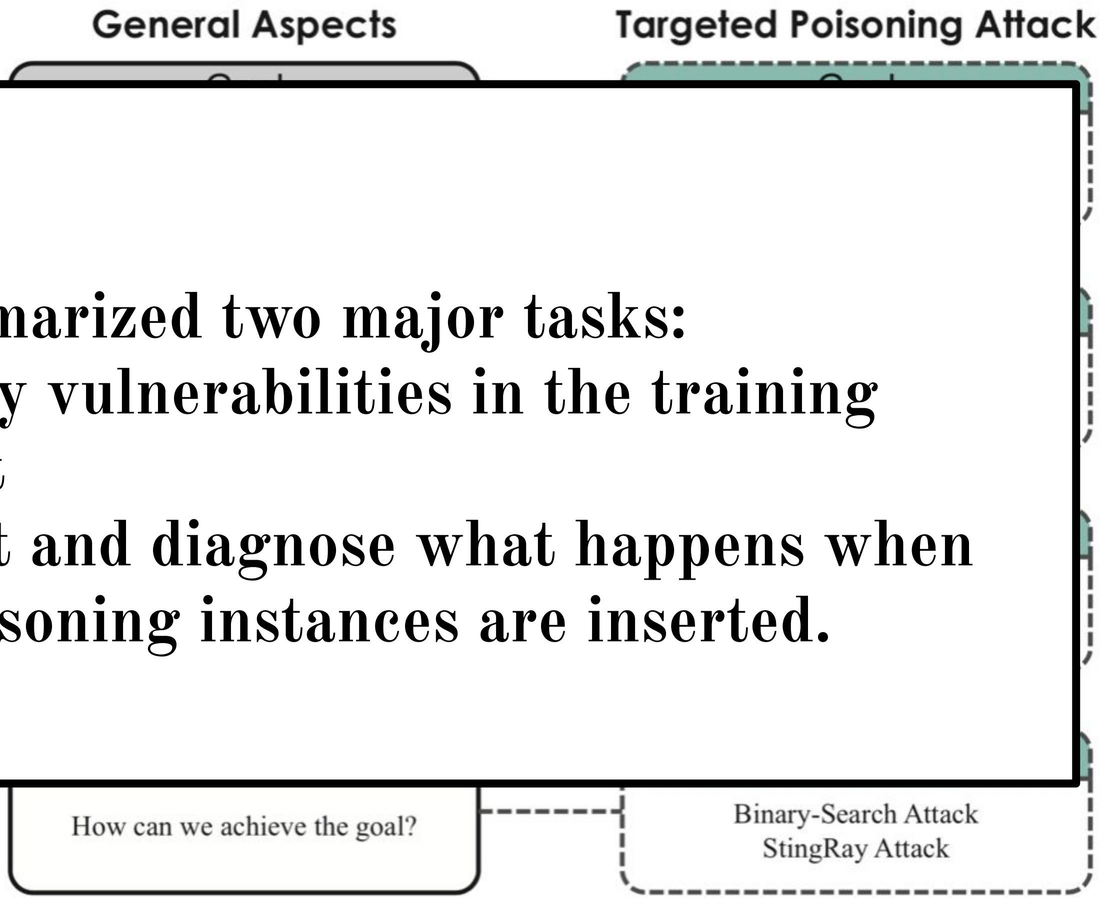


To demonstrate the proposed visual analytics framework, we focus our discussion on:

Targeted Data Poisoning Attack



To demonstrate the
prop
anal
fram
focus
discu
Targ
Pois



- **We have summarized two major tasks:**
  1. **Identify vulnerabilities in the training dataset**
  2. **Inspect and diagnose what happens when the poisoning instances are inserted.**

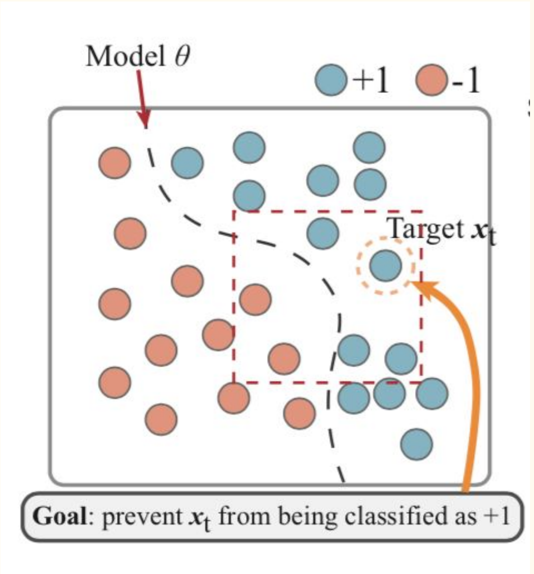
VISUAL ANALYTICS FRAMEWORK

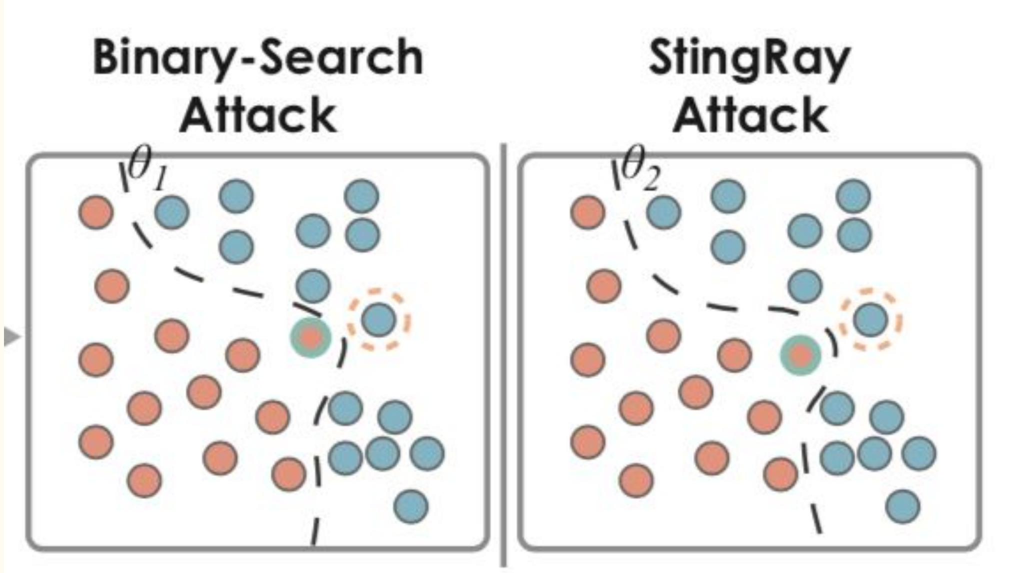
The framework supports three main activities:

1. Vulnerability analysis
2. Attack space analysis
3. Attack results analysis

Vulnerability Analysis

- Core idea: To change the label of the target instance
  - Attack algorithms: Binary-Search Attack & StingRay Attack





Vulnerability Analysis

- **Vulnerability Measures (to explore the potential weaknesses in the model):**
  - **Decision Boundary Distances (DBD)**
  - **Minimum Cost for a Successful Attack (MCSA)**

Visualizing the Attack Space

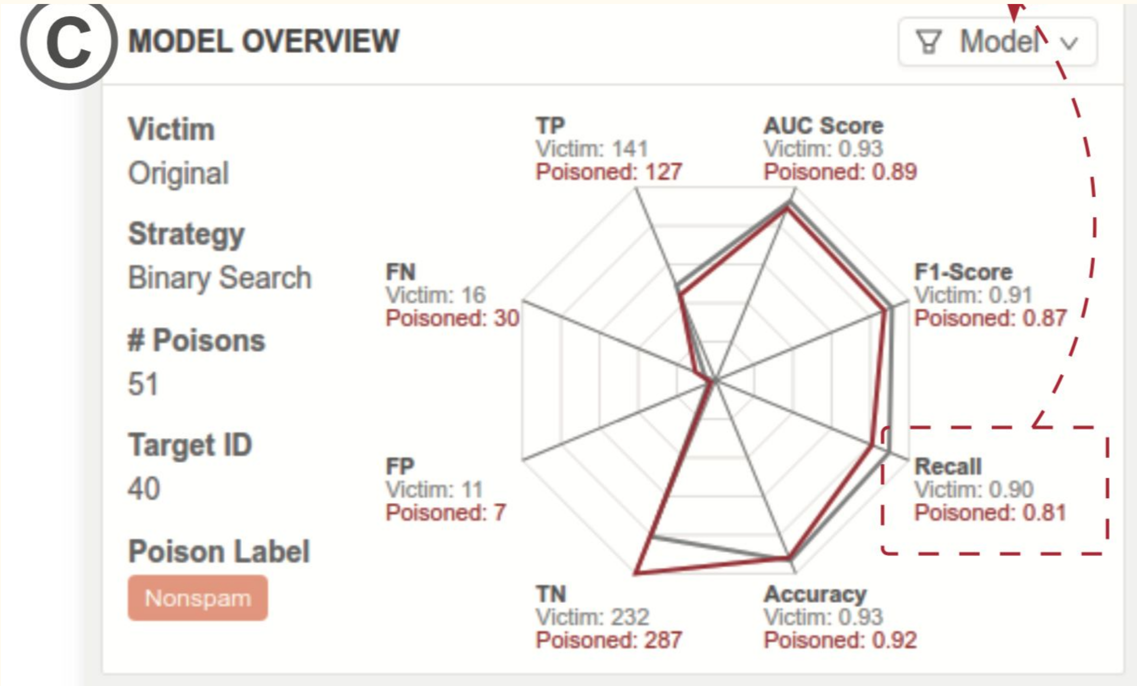
- Vulnerability Measures:
  - Decision Boundary Distances (DBD)
  - Minimum Cost for a Successful Attack (MCSA)
  - Performance metrics of the poisoned model (Accuracy, Recall, etc.)

- Each instance in the training dataset is measured based on these vulnerability measures.
- Video

Attack Detail Analysis

- TN
- FN
- TP
- FP

Model Overview



Radar Chart for Model Performances

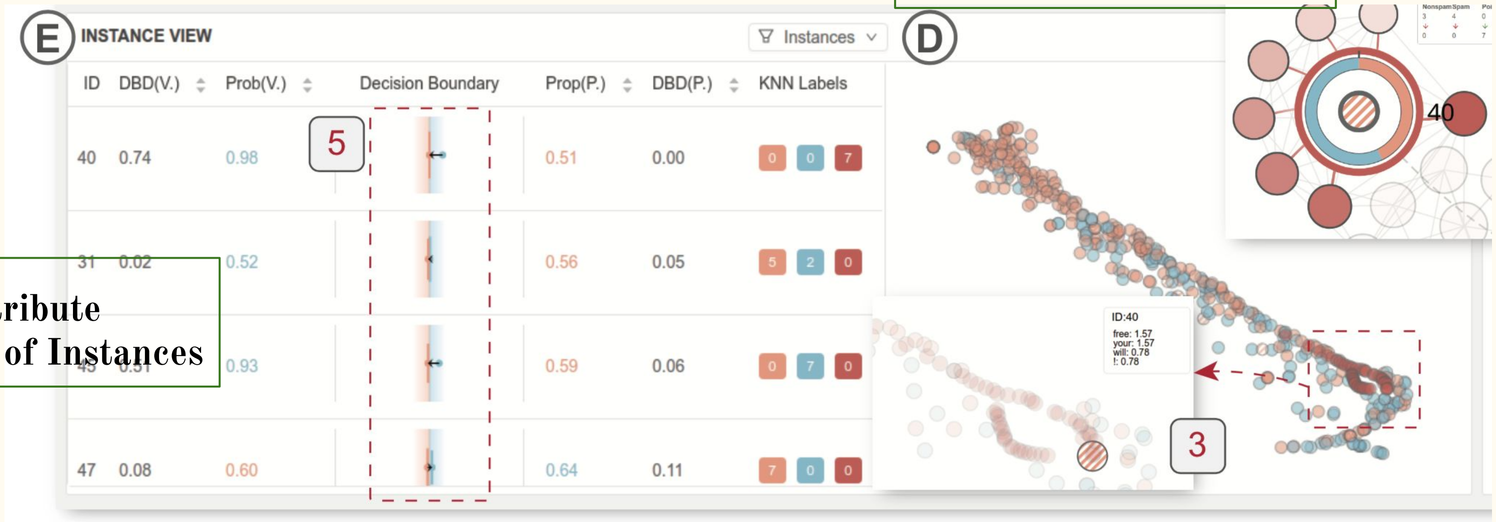
- Acc
- Recall
- F1
- ROC

Attack Detail Analysis

T-SNE Projection
- Overview of the data distribution

Instance View

- Instances Attribute
  - Details of Instances



- Key Attributes for Instances:
  - Decision Boundary Distances
  - Classification Probabilities
  - Labels of K-NNs

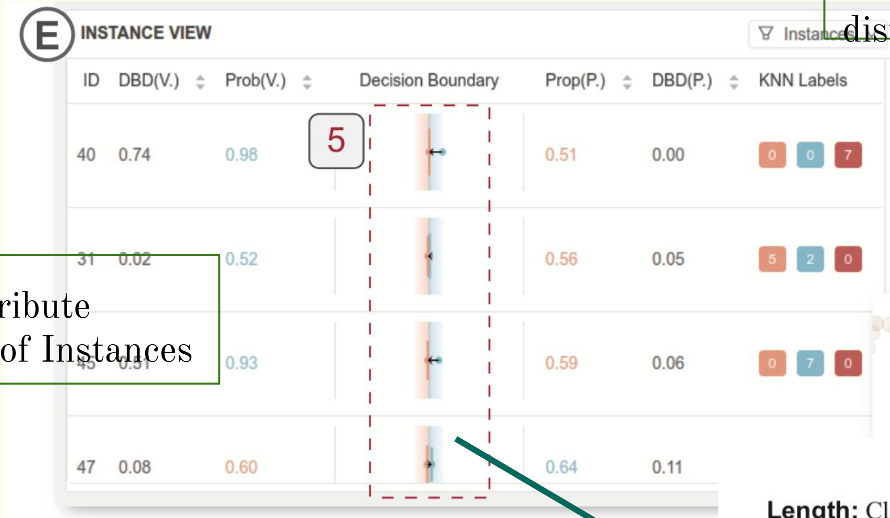
Attack Detail Analysis

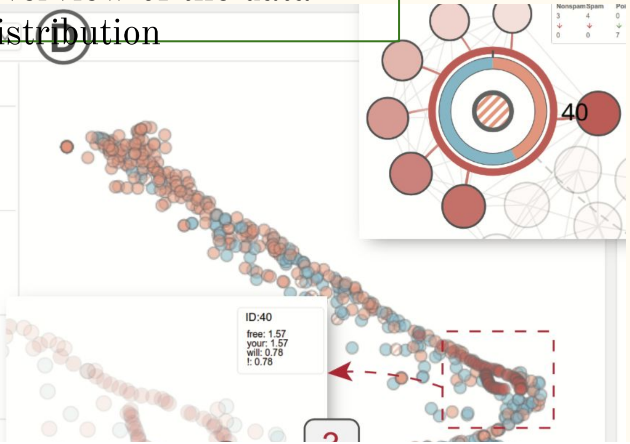
T-SNE Projection

Overview of the data distribution

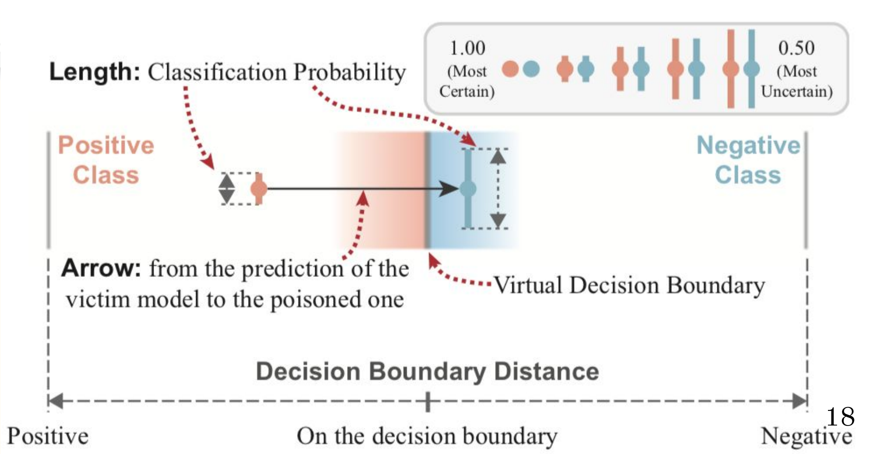
Instance View

- Instances Attribute
  - Details of Instances





- **Key Attributes for Instances:**
  - **Decision Boundary Distances**
  - **Classification Probabilities**
  - **Labels of K-NNs**



Attack Detail Analysis

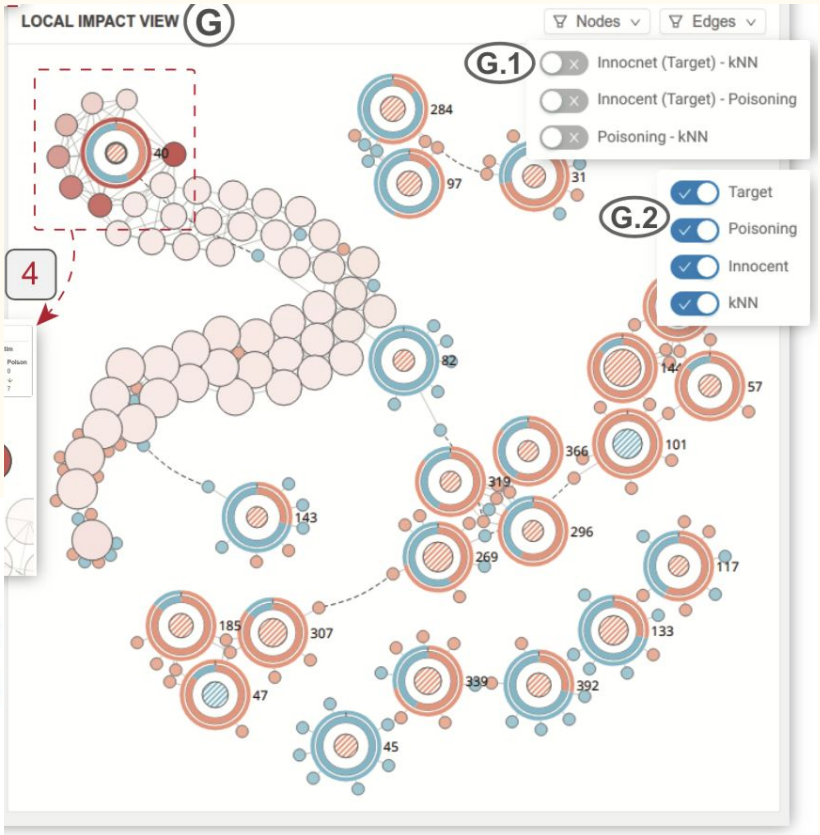
Feature View

- Data Distributions on Features
  - Instances in the spam/non spam classes
  - Poisoning Instances
- Feature Importance Rankings
  - In the victim model
  - In the poisoned model
  - Differences

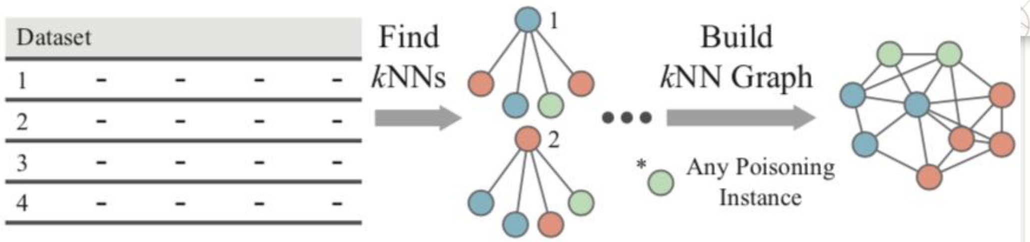
Attack Detail Analysis

Local Impact View

- The target would be influenced either by its neighbors or the poisoned instances



KNN Graph Building



(a) kNN Graph Building

Attack Detail Analysis

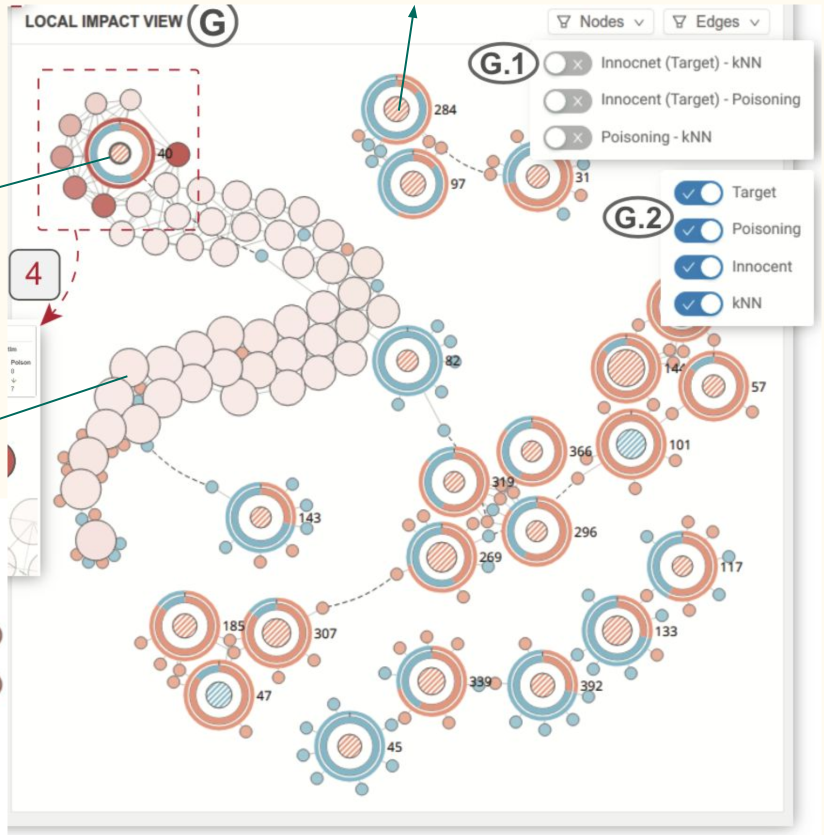
Local Impact View

- The target would be influenced either by its neighbors or the poisoned instances

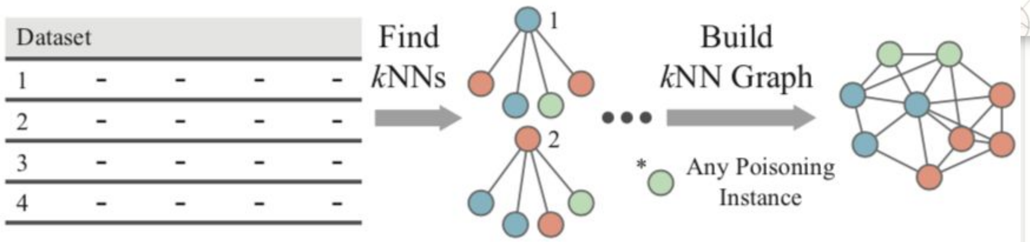
Innocent Instances

Target Instance

Poisoning Instance



KNN Graph Building

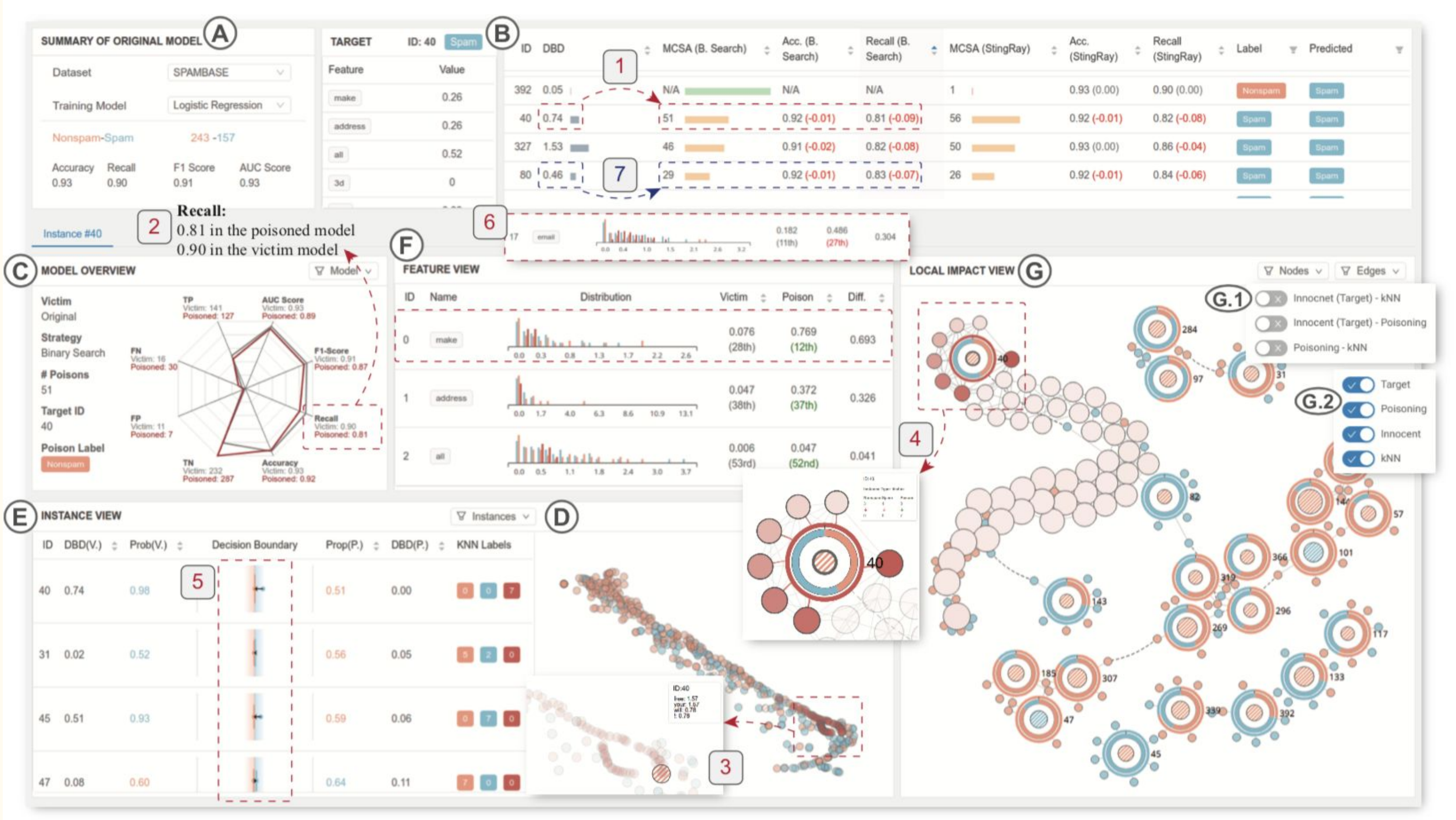


(a) kNN Graph Building

Case Study

Critiques

- **Strengths**
  - **Two stage design in the interface**
  - **Very user friendly**
  - **Multi Faceted Analysis**
- **Weaknesses**
  - **Only allows to attack one instance**
  - **Speed on bigger datasets and more complicated models**
  - **Scalability (Visual design, Attack Algorithm).**
    - **Case studies are too simplified.**



SUMMARY OF ORIGINAL MODEL
A
Dataset
SPAMBASE
Training Model
Logistic Regression
Nonspam-Spam
243 -157
Accuracy
Recall
F1 Score
AUC Score
0.93
0.90
0.91
0.93
TARGET
ID: 40
Spam
Feature
Value
make
0.26
address
0.26
all
0.52
3d
0
B
ID
DBD
MCSA (B. Search)
Acc. (B. Search)
Recall (B. Search)
MCSA (StingRay)
Acc. (StingRay)
Recall (StingRay)
Label
Predicted
392 0.05 N/A N/A N/A 1 0.93 (0.00) 0.90 (0.00) Nonspam Spam
40 0.74 51 0.92 (-0.01) 0.81 (-0.09) 56 0.92 (-0.01) 0.82 (-0.08) Spam Spam
327 1.53 46 0.91 (-0.02) 0.82 (-0.08) 50 0.93 (0.00) 0.86 (-0.04) Spam Spam
80 0.46 29 0.92 (-0.01) 0.83 (-0.07) 26 0.92 (-0.01) 0.84 (-0.06) Spam Spam
1
7
Instance #40
2
Recall:
0.81 in the poisoned model
0.90 in the victim model
6
F
C
MODEL OVERVIEW
Model
Victim
Original
Strategy
Binary Search
Poisons
51
Target ID
40
Poison Label
Nonspam
TP
Victim: 141
Poisoned: 127
AUC Score
Victim: 0.93
Poisoned: 0.89
FN
Victim: 16
Poisoned: 30
F1-Score
Victim: 0.91
Poisoned: 0.87
FP
Victim: 11
Poisoned: 7
Recall
Victim: 0.90
Poisoned: 0.81
TN
Victim: 232
Poisoned: 287
Accuracy
Victim: 0.93
Poisoned: 0.92
FEATURE VIEW
ID
Name
Distribution
Victim
Poison
Diff.
0 make 0.076 (28th) 0.769 (12th) 0.693
1 address 0.047 (38th) 0.372 (37th) 0.326
2 all 0.006 (53rd) 0.047 (52nd) 0.041
LOCAL IMPACT VIEW
G
Nodes
Edges
G.1
Innocnet (Target) - kNN
Innocent (Target) - Poisoning
Poisoning - kNN
G.2
Target
Poisoning
Innocent
kNN
4
E
INSTANCE VIEW
Instances
D
ID
DBD(V.)
Prob(V.)
Decision Boundary
Prop(P.)
DBD(P.)
KNN Labels
40 0.74 0.98 0.51 0.00
31 0.02 0.52 0.56 0.05
45 0.51 0.93 0.59 0.06
47 0.08 0.60 0.64 0.11
5
3


SUMMARY OF ORIGINAL MODEL
Dataset
SPAMBASE
TARGET
ID: 40
Spam
Feature
Value
1 Data Table View
2 The Selected Target
TARGET ID: 152 Number 8
3 Model Overview
TP
Victim: 200
Poisoned: 213
AUC Score
Victim: 1.00
Poisoned: 1.00
FN
Victim: 0
Poisoned: 0
F1-Score
Victim: 1.00
Poisoned: 1.00
FP
Victim: 0
Poisoned: 1
Recall
Victim: 1.00
Poisoned: 1.00
TN
Victim: 200
Poisoned: 199
Accuracy
Victim: 1.00
Poisoned: 1.00
4 Projection View
5 Local Impact View
6 Instance Attribute View
Decision Boundary

Thank you!

QA

Attack Detail Analysis

Local Impact View

Innocent Instances