Cyber Security

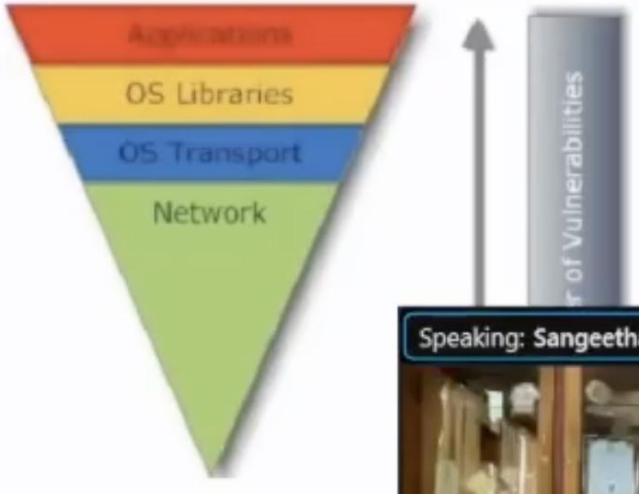
- Increasing number of users, devices and data
- Challenge – Evolving number of risks and threats
- Cyber Security in the era of AI

Cyber Security

- Increasing number of users, devices and data
- Challenge – Evolving number of risks and threats
- Cyber Security in the era of AI

Automation in Cyber Security

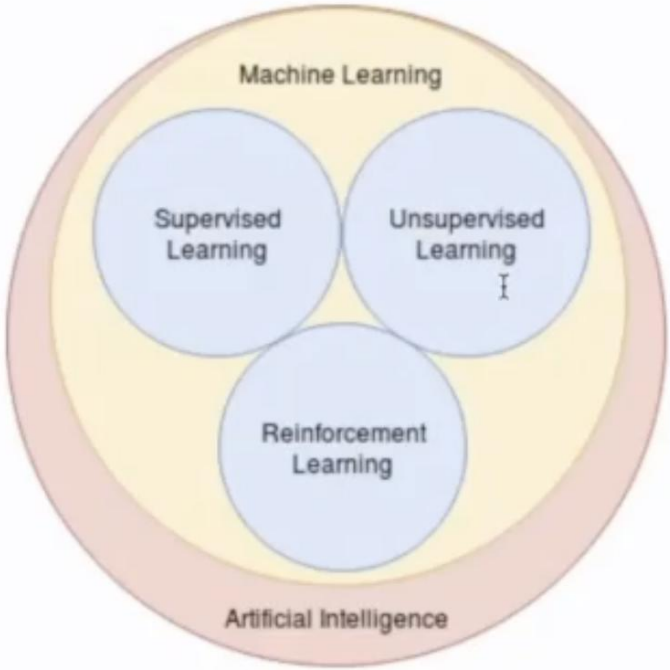
- Keep companies protected from the growing number and sophistication of cyberthreats.
- AI – handle huge volume of data



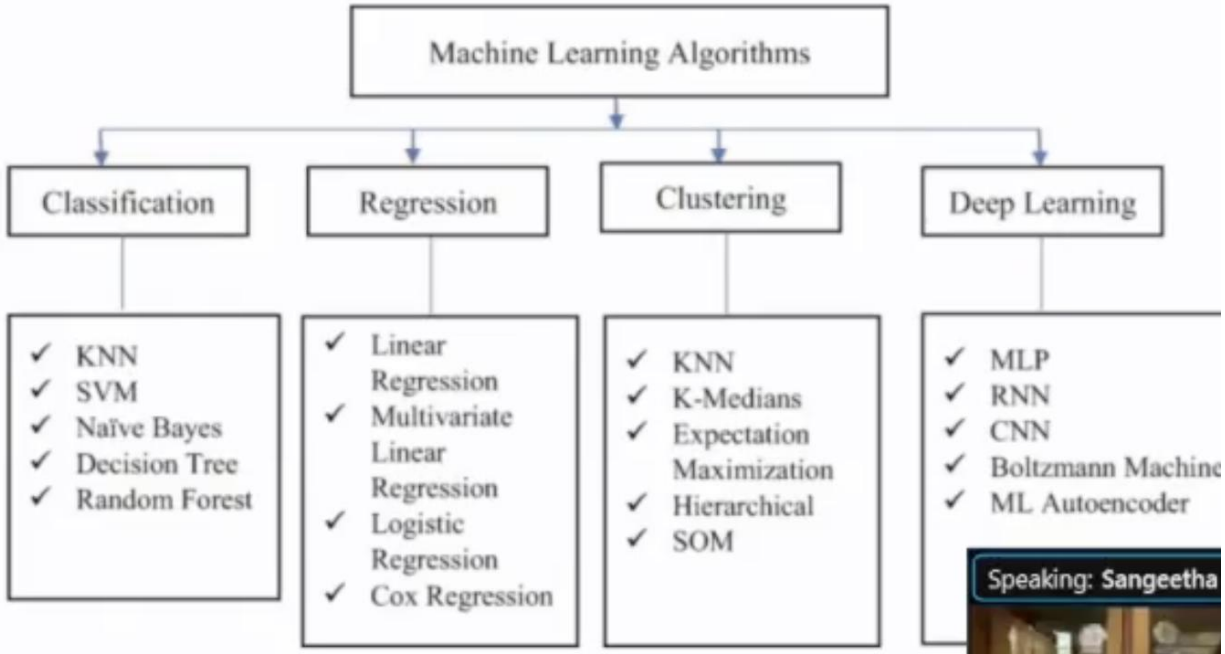
Artificial Intelligence and Cyber Security

- AI for Cyber Security
- Cyber Security for AI

The AI Circle



The AI Circle



Machine Learning challenges

Explainability

- Understand what DL actually learned
- Legal challenges

Verifiability

- Verifiability of detections
- Interpretation of output

Data Quality and Bias

- Not enough or no quality labelled data
- Data cleanliness issues
  timestamps, normalization across fields, etc.
- Bad understanding of the data to engineer meaningful features

Knowledge

- Qualifications of people developing the models
- Understanding the business, the maths, and IT

AI for Cyber Security

- Defensive AI – Blue AI
- Offensive AI – Red AI

Defensive AI

Malware detection

- Multi layer, multi ML engine defense

SOC, IDS/IPS & Honeypots

- Self learning ML and DL

Antispam

Vulnerability Mgt

- Identify and prioritize remediation

Data Classification

- Track data to identify, classify and protect

Threat Intelligence

- Categorize behavior for TI
- ML to monitor Dark We

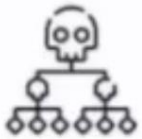
Offensive AI

Malware creation

- Speed up creation
- Enhance evasive capa

Adversarial AI

- GAN: discover and poison ML to produce false, and controlled, results
- Poison datasets



Smart botnets

- Self learning botnets
- Smarter zombies

Conditional attacks

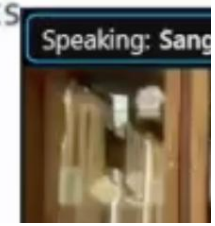
- Cyberattacks using Blockchain based smart contracts

Spear phishing

- Smarter social engineering
- More convincing scams

Classify victims

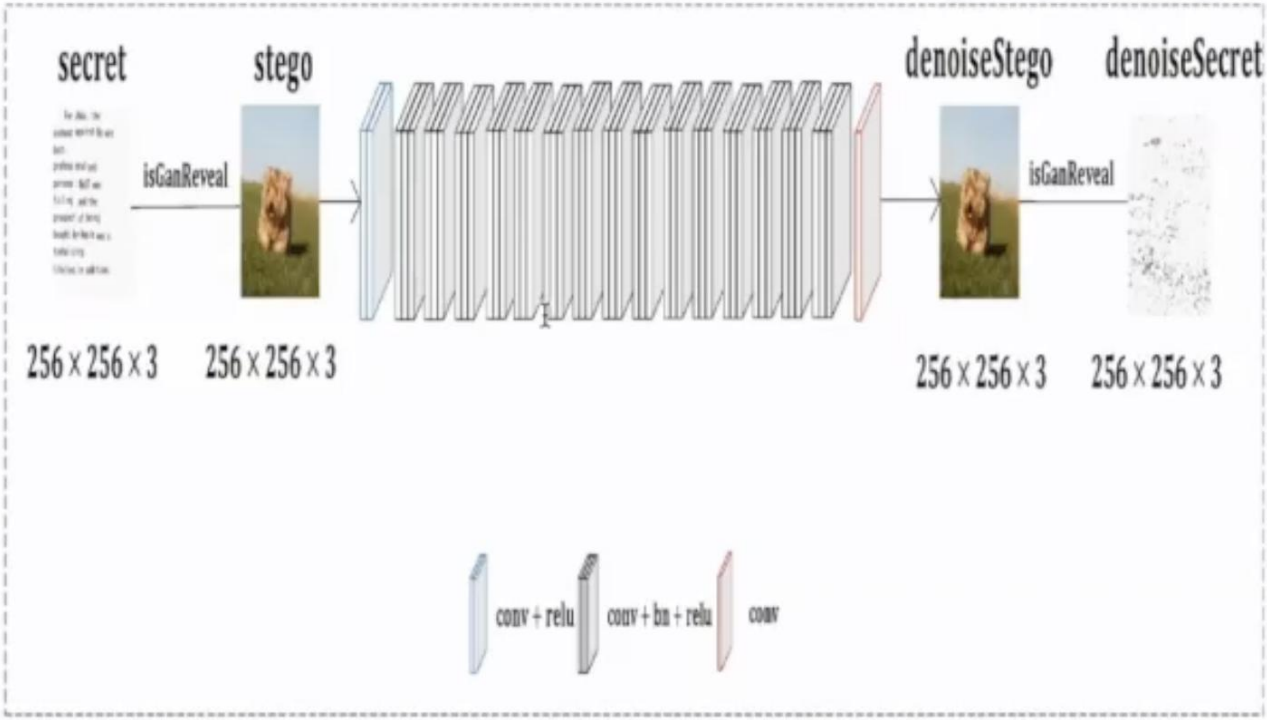
- Optimize return on investment of attacks



ML based Firewall

- Next-Generation Firewall (NGFW)
- ML-Powered NGFW (proactive firewall
- Four mechanisms that power ML-NGFW
  - *Inline ML algorithms*
  - *Zero-delay signatures*
  - *ML-Powered Visibility across Iot Devices*
  - *Automated , Intelligent Policy Recommendation*

Deep Learning based Steganalysis



cover_medium + hidden_data +

stego_key = stego_medium

Fig. source - Zhong S, Weng W, Chen K, Lai J. Deep-Learning Steganalysis for Removing Document Images on the Basis of Geom
Median Pruning. *Symmetry*. 2020; 12(9):1426. https://doi.org/10.3390/sym12091426

VulDeepecker – A deep learning based system for vulnerability detection

- Automatic detection of software vulnerabilities
- Representation - Code gadgets
- Softwares – Xen, Seamonkey, Libav

Learning phase :

1. Extracting the library/API function calls and the corresponding program slices.
2. Generating code gadgets of the training programs and their ground truth labels.
3. Transforming code gadgets into vector representations.
4. Training a BLSTM neural network

Detection phase :

5. Transforming target programs into code gadgets and vectors
6. Detection

Intrusion Detection System

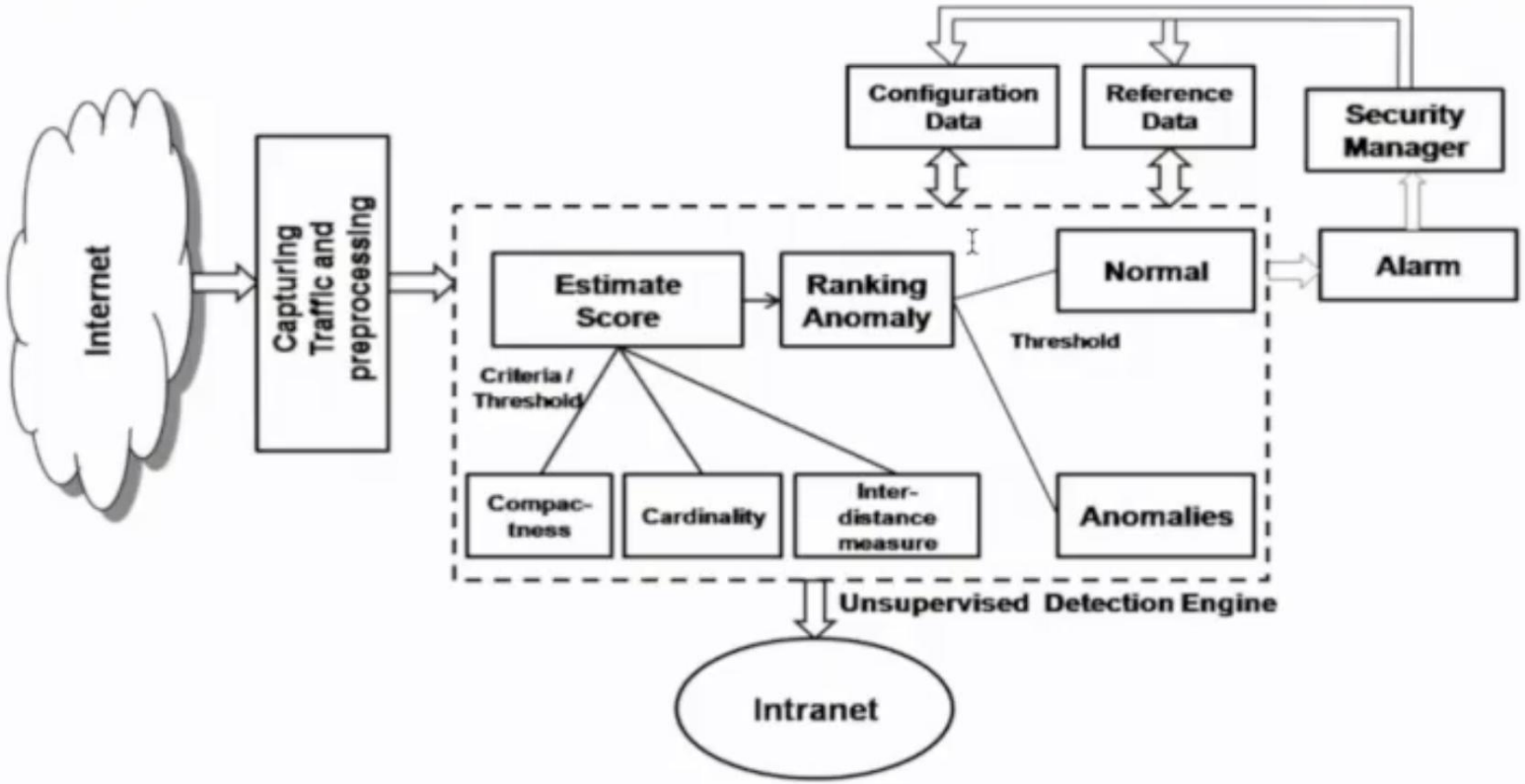


Fig. source - D. K. Bhattacharyya and J. K. Kalita, Network Anomaly Detection: A Machine Learning Perspective, 1st Edition, Chapman and Hall/CRC, 2013.

Preventing Data leakage in cloud

- Amazon Macie for S3
- Microsoft Azure SQL Database threat detection
  - uses ML to look at cloud-based SQL server-based logs for anomalous activity
- Google data log prevention API
  - predefined detectors for PIIs, context clues

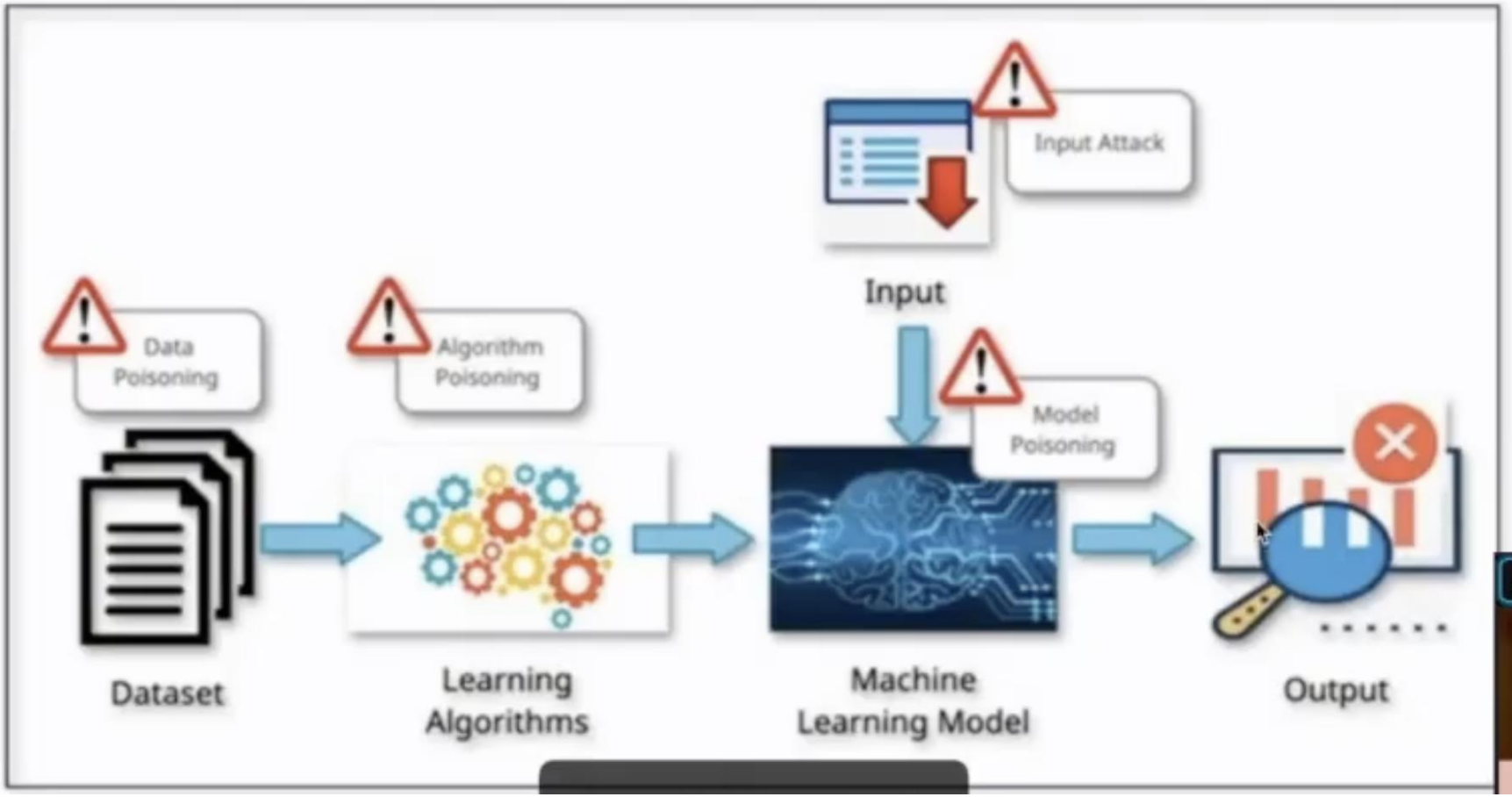
Cyber Security for AI

How secure are our AI models?

The primary goals of AI model security are:

- **Integrity**
- **Availability**
- **Privacy**

Target Areas for attack on AI model



Attacks during Training

- Poisoning attack
  - confidence reduction,
  - misclassification
- Trojaning/Backdoor attack
  - access to original dataset not required
  - retraining

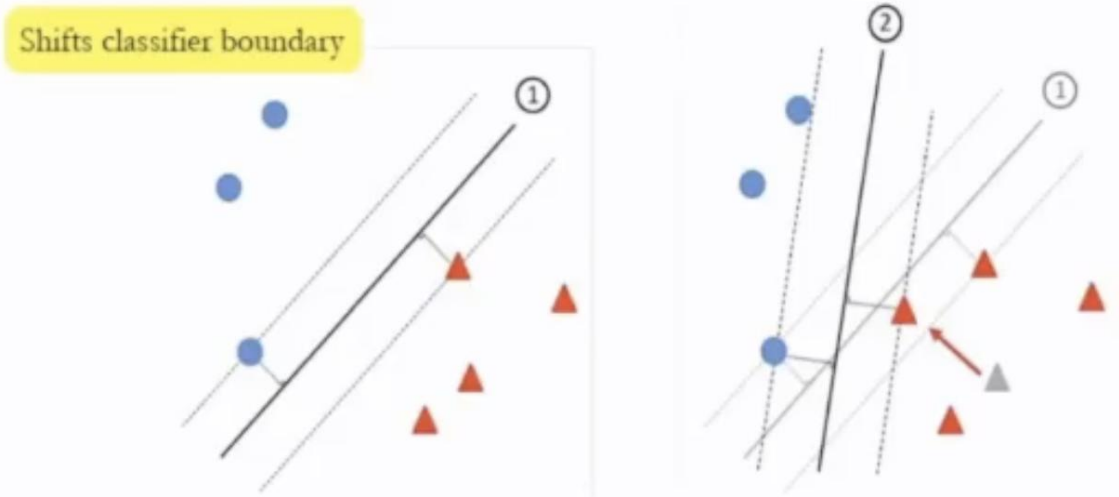


Fig. 1. Linear SVM classifier decision boundary for a two-class dataset with support vectors and classification margins indicated (left). Decision boundary is significantly impacted if just one training sample is changed, even when that sample's class label does not change (right).

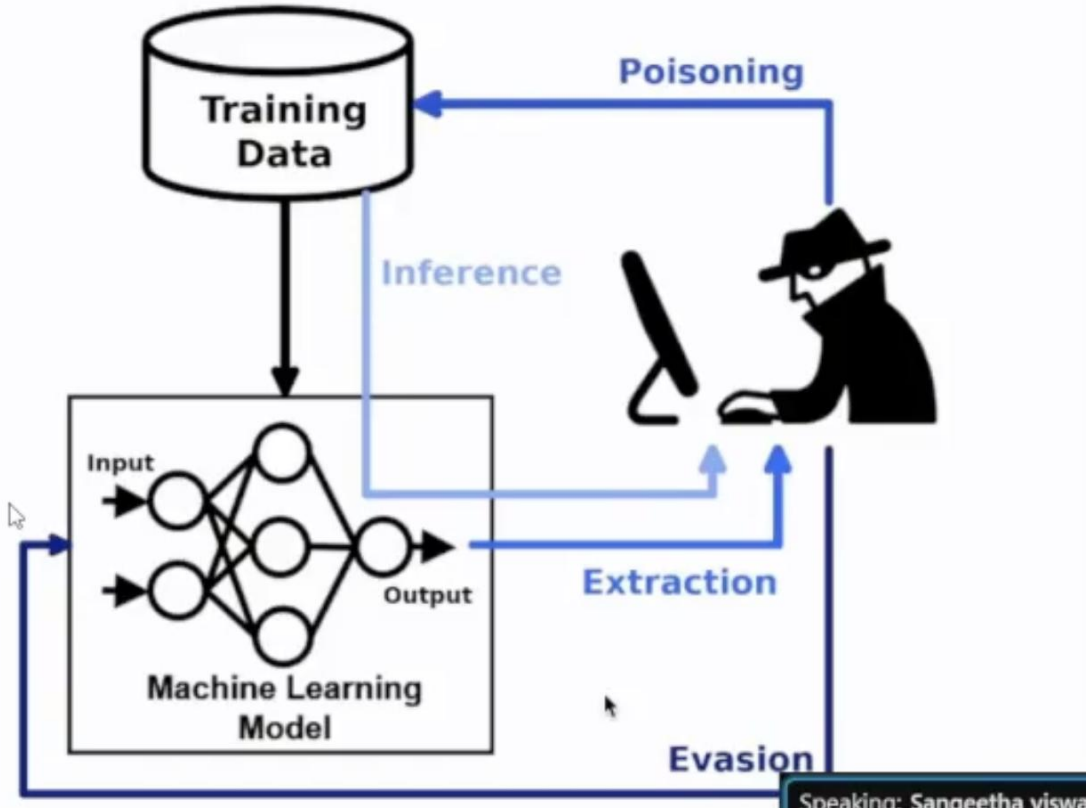
Miller, Xiang, and Kesidis, 2019

Do machine learning models remember data?

Can attackers extract information (in training data) from (Querying)learned models ?

Attacks during Production (testing)

- Inference attack
  - acquire information about dataset form target model
- Evasion attack
  - modify samples at test time



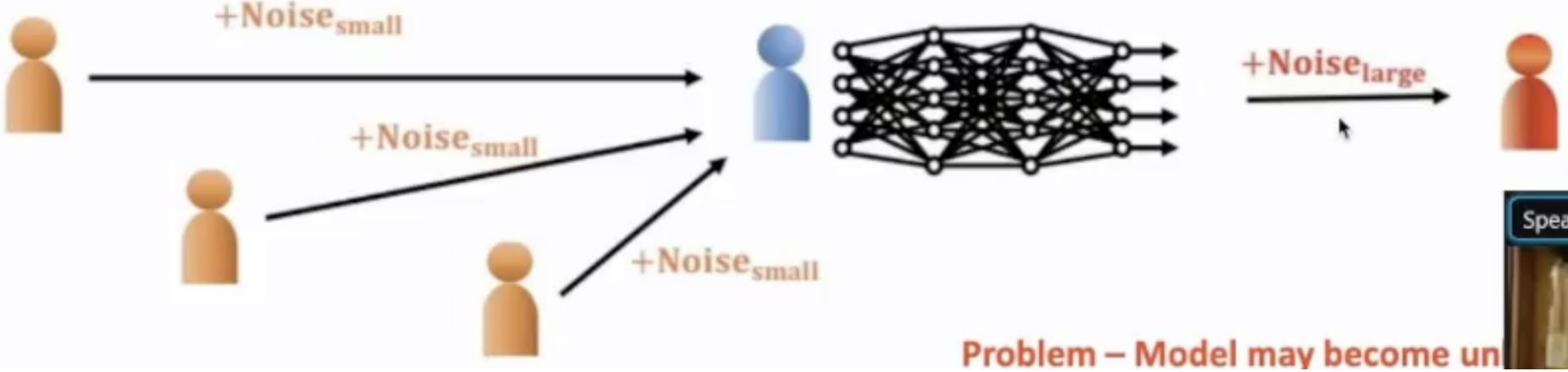
Securing AI

- Differential privacy
- Federated learning
- Homomorphic encryption
- Adversarial Training

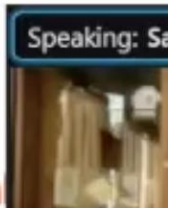
Differential privacy

- Inference attacks

Goal: Try to hide individual data points

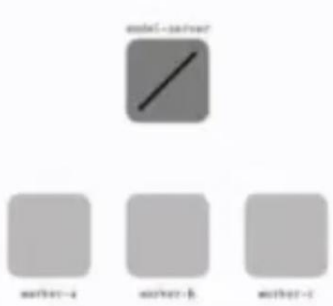
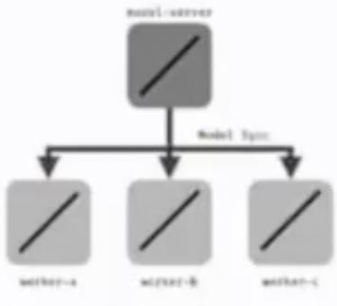
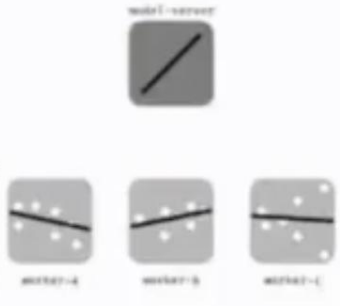
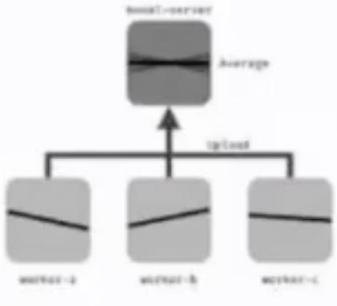


Problem – Model may become un



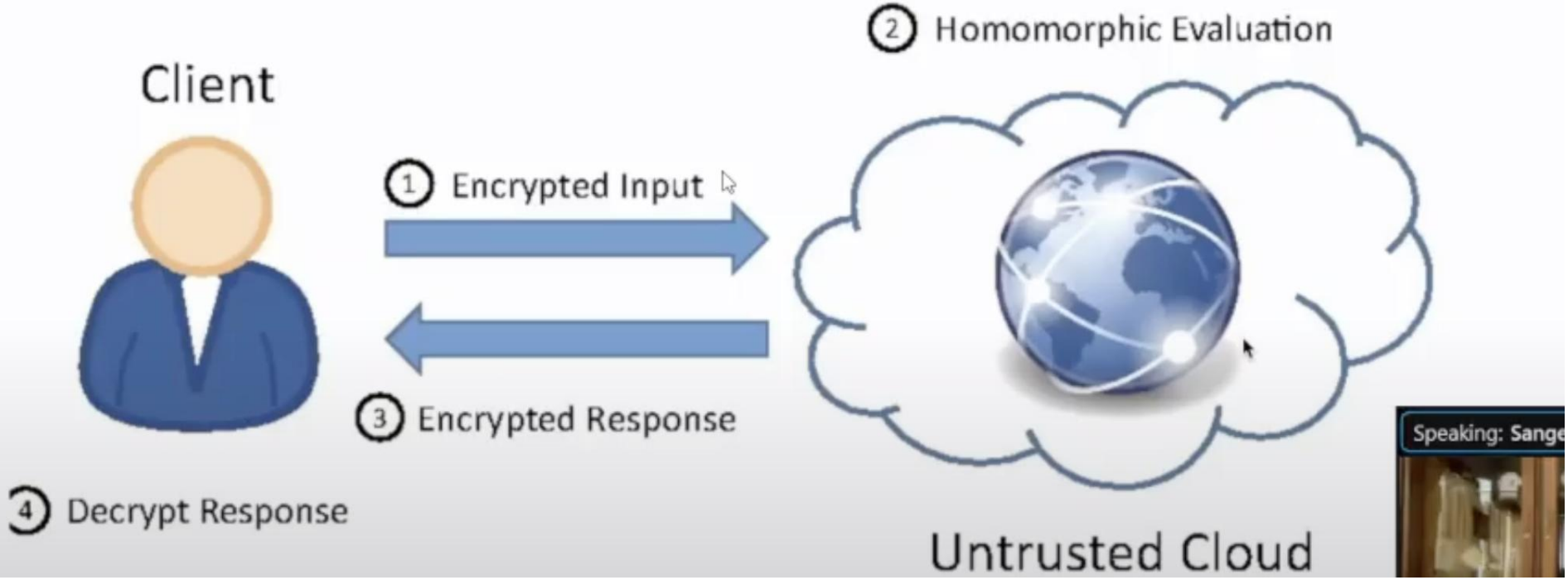
Federated Learning

Instead of taking data from user devices to our server, why don't we send the algorithm to the data?

Step 1	Step 2	Step 3	Step 4
			
Central server chooses a statistical model to be trained	Central server transmits the initial model to several nodes	Nodes train the model locally with their own data	Central server pools model results and generate one global mode without accessing any data

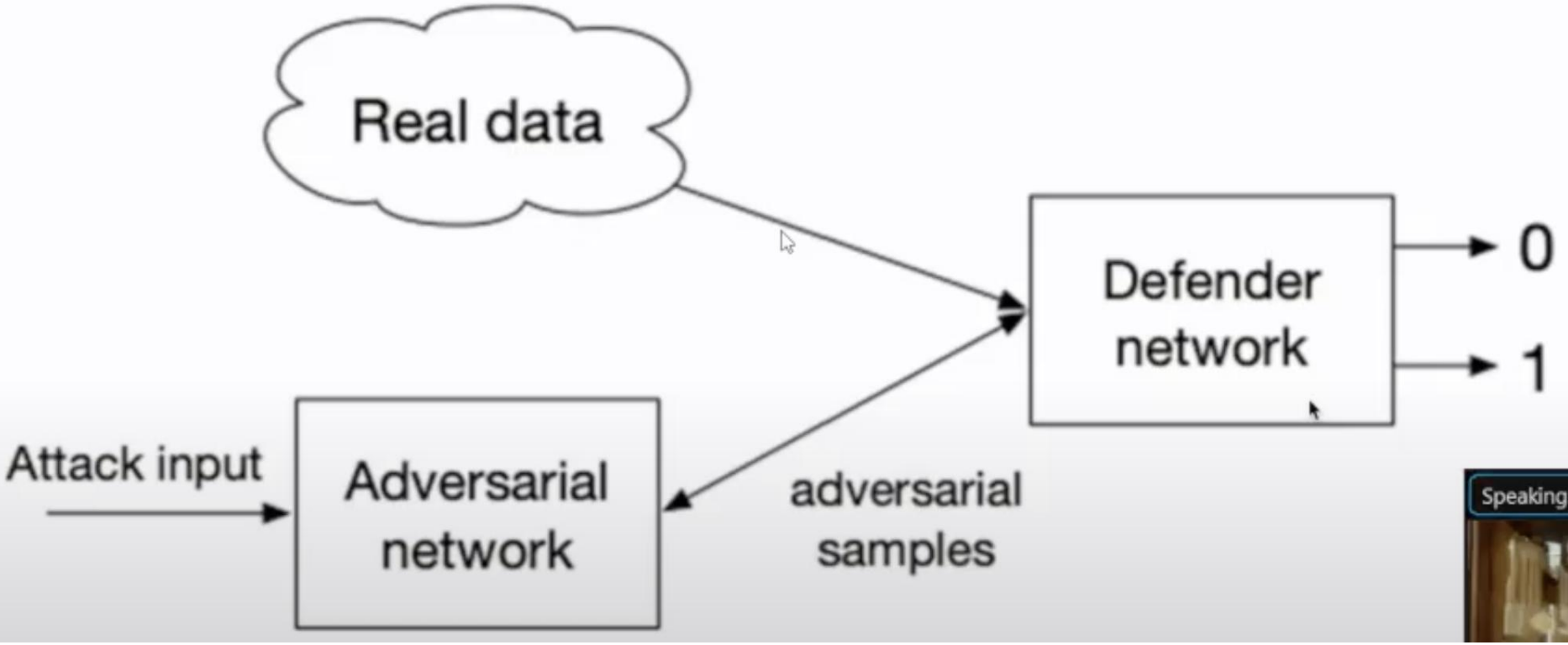
Homomorphic encryption

- Securing data stored in cloud
- Enabling data analytics in regulated industries
- Secure Data Monetization
- Secure data sharing and collaboration



Adversarial Training

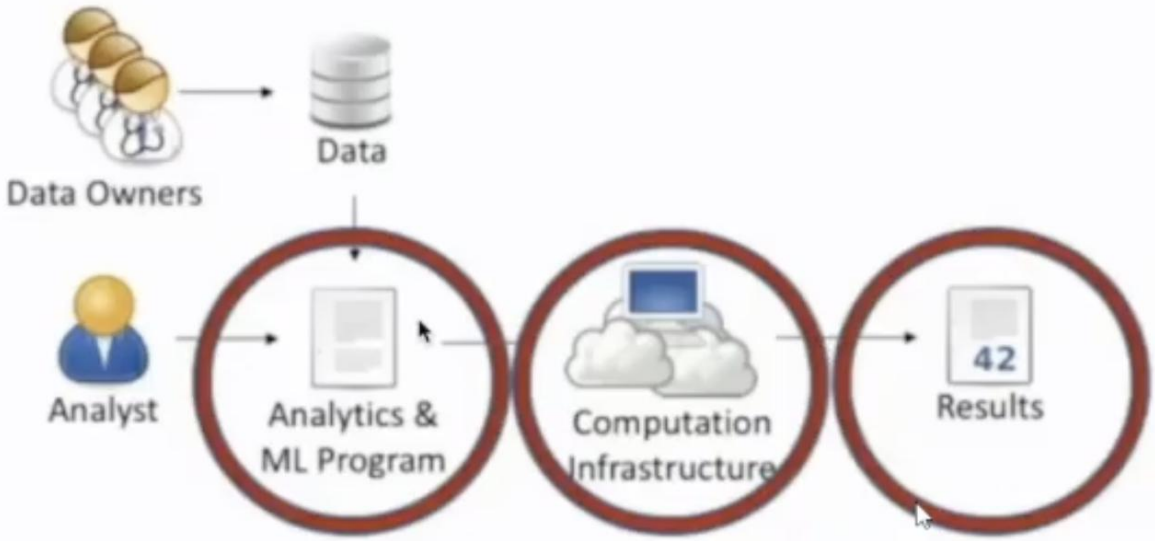
- Adversarial robustness
- Defense against adversarial examples



Bias in AI models

- Trained on the wrong thing
- Faulty Interpretation
- Sample biases
- Analytics bias
- Confirmation bias
- Outlier bias

Current AI and Cyber Security framework



Some Case Studies :

- Classification of VoIP Traffic based on Analysis of VBR Codecs
- A Model for the Effective Steganalysis of VoIP
- Botnet detection using Network traffic analysis
- DDoS attack detection model using machine learning in next generation firewall
- Advanced Persistent threat detection using Long Short-Term Memory(LSTM) neural networks
- WIDS : An anomaly-based intrusion detection system for Wi-Fi
- Privacy-preserving machine learning in cloud