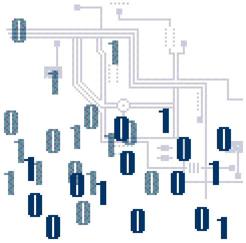
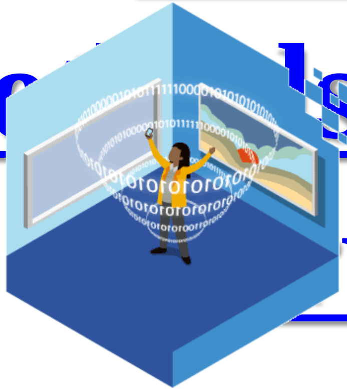
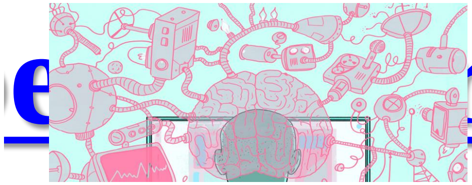
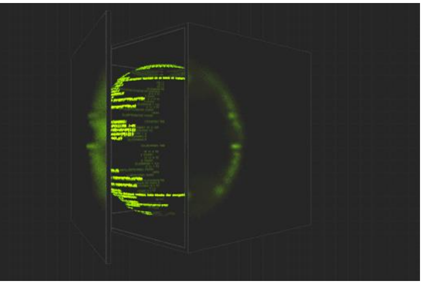
SeoulTech UCS Lab
Ubiquitous Computing & Security Laboratory

A Survey of Data Mining and Machine Learning Me ls f e ri u tection









2017. 09. 18

Presented by
Pradip Kumar Sharma
(pradip@seoultech.ac.kr)

OBJECTIVE

- The main focus of this presentation is on survey of machine learning (ML) and data mining (DM) methods for cyber analytics in support of intrusion detection.

- The data are so important in ML/DM approaches, some well-known cyber data sets used in ML/DM are described.

- Discussion of challenges for using ML/DM for cyber security is presented, and some recommendations on when to use a given methods are provided.

CYBER SECURITY INTRUSION DETECTION: WHAT WE NEED TO KNOW

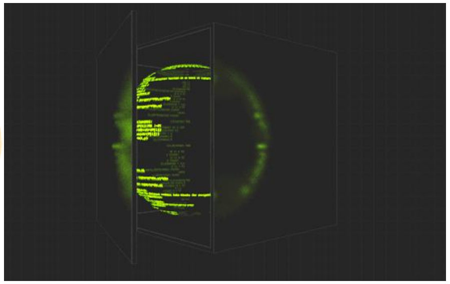
WHAT IS A CYBER CRIME?

- **Cyber crime encompasses any criminal act dealing with computers and networks (called hacking).**

- **The computer used as an object or subject of crime.**

- **Malicious programs, Illegal imports, Computer Vandalism.**

- **A major attack vector of Cyber Crime is to exploit broken software.**

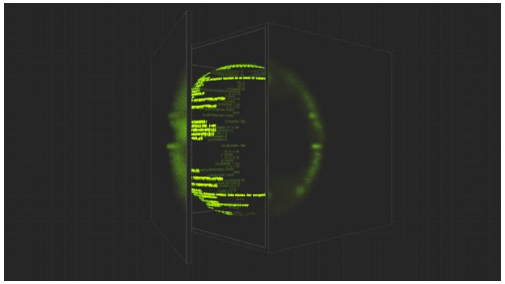


WHAT IS A CYBER SECURITY?

- ❑Set of technologies, processes and practices designed to protect networks, computers, programs and data from attack, damage or unauthorized access.

- ❑Composed of computer security system and network security systems.

- ❑A major part of Cyber Security is to fix broken software



CYBER SECURITY VS CYBER CRIME

- □ *Cyber Security will be massively improved if there are less broken software*

- □ *Cyber Crime will be massively reduced if there are less broken software*

The Coin: Broken/Complex Software

Cyber Security: One side of the coin

Cyber Crime: Other side of the coin

Cyber Crime

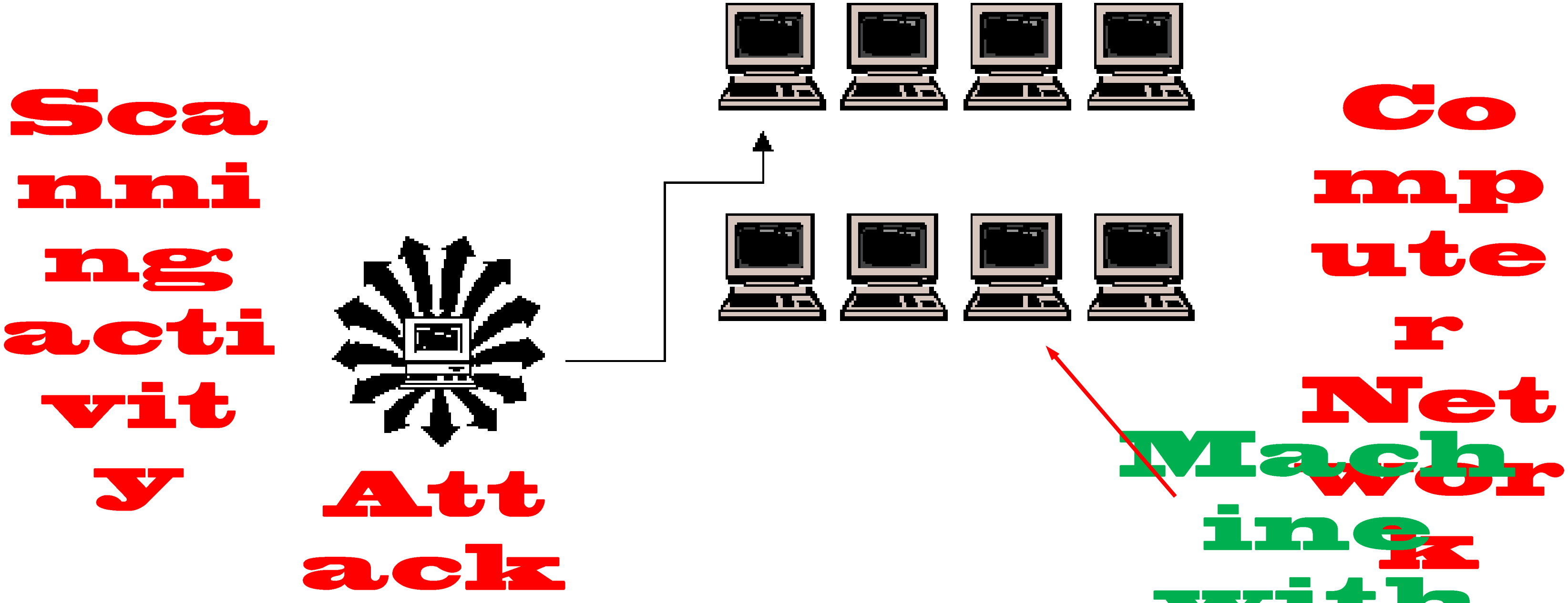
Cyber Security

Decrease in broken software = Increase in good software

CYBER ATTACKS - INTRUSIONS

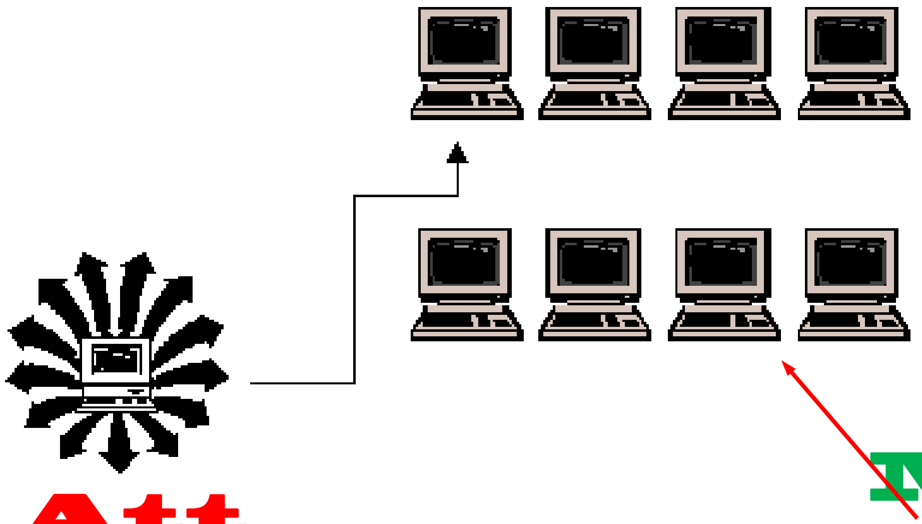
- Cyber attacks (intrusions) are actions that attempt to bypass security mechanisms of computer systems. They are caused by:
  - Attackers accessing the system from Internet
  - Insider attackers - authorized users attempting to gain and misuse non-authorized privileges

- Typical intrusion scenario



NUMBER OF CONNECTIONS INVOLVED IN ATTACKS

- ❑Generally two types of cyber attacks in the computer networks:
  - ❑attacks that involve multiple network connections (bursts of connections)
  - ❑attacks that involve single network connections



Multiple-connection computer

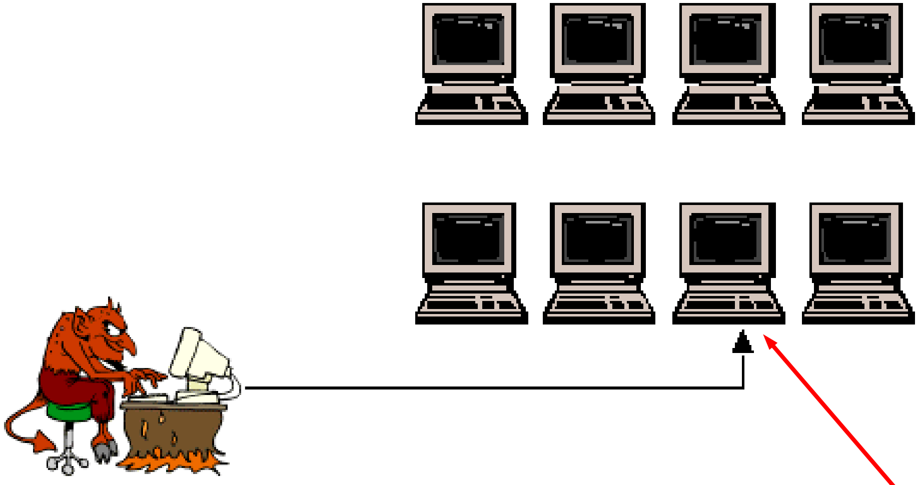
Attack

Computer Network

Machine with

NUMBER OF CONNECTIONS INVOLVED IN ATTACKS

Single connection attack



Computer Network

Attacker

Victim Machine

WHY WE NEED INTRUSION DETECTION?

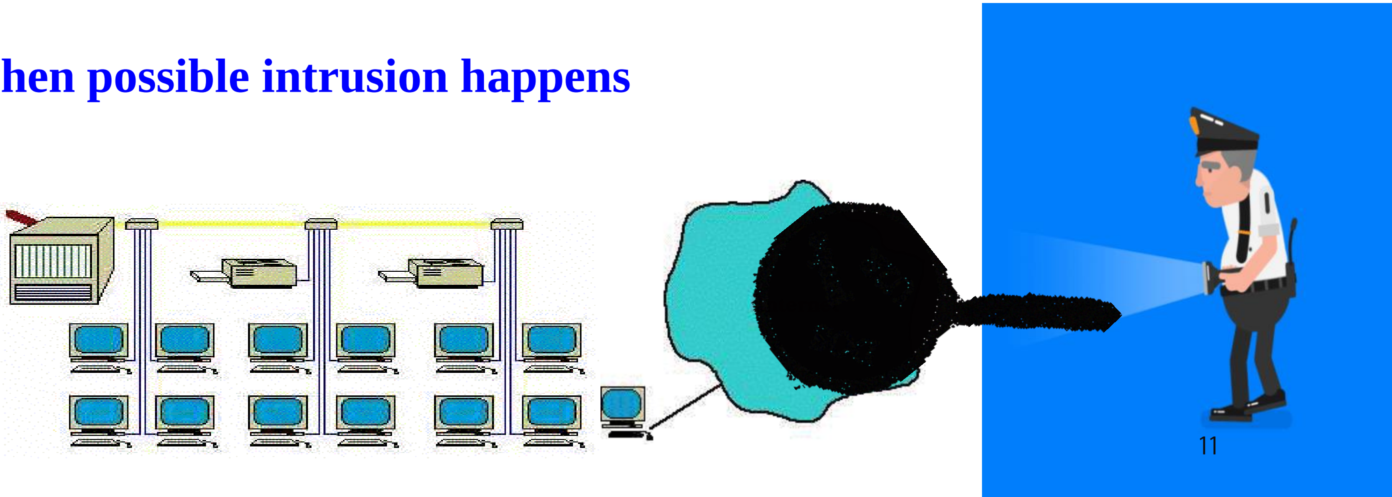
- ❑ Security mechanisms always have inevitable vulnerabilities

- ❑ Current firewalls are not sufficient to ensure security in computer networks
  - ❑ "Security holes" caused by allowances made to users/programmers/administrators
  - ❑ Insider attacks
  - ❑ Multiple levels of data confidentiality in commercial and government organizations needs multi-layer protection in firewalls

INTRUSION DETECTION

- **Intrusion Detection:** **Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of intrusions, defined as attempts to bypass the security mechanisms of a computer or network ("compromise the confidentiality, integrity, availability of information resources")**

- **Intrusion Detection System (IDS)**
  - **combination of software and hardware that attempts to perform intrusion detection**
  - **raise the alarm when possible intrusion happens**



INTRUSION DETECTION SYSTEM

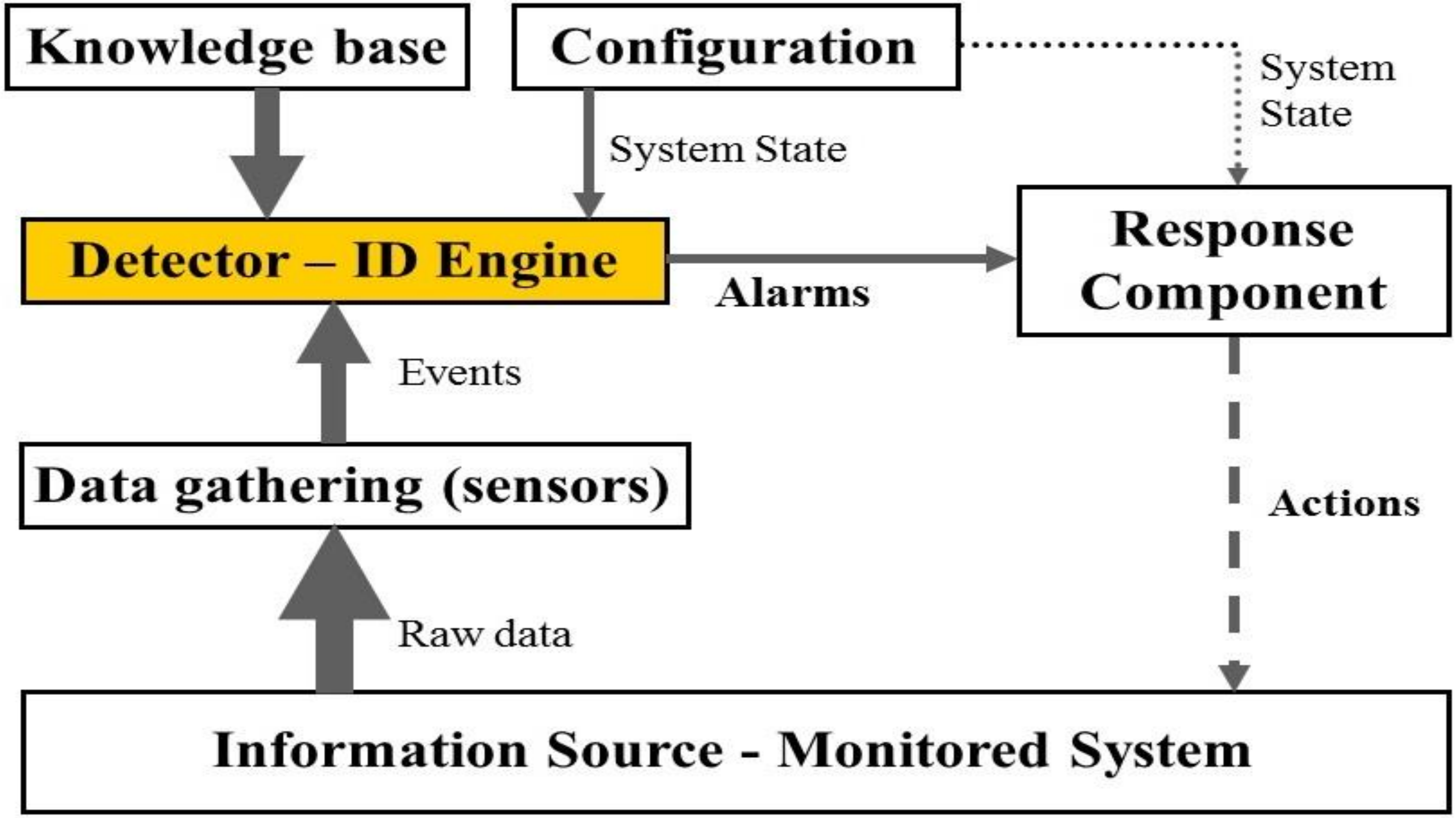


Fig : Intrusion Detection System

INTRUSION DETECTION SYSTEM

- There are three main types of cyber analytics for supporting IDS :

  - Misuse Based.
  - Anomaly Based.
  - Hybrid.

INTRUSION DETECTION SYSTEM

Misuse Based Detection

❑ Designed to detect known attacks by using signatures of those attacks.

❑ Effective detecting known type of attacks without generating false alarms.

❑ Frequent manual updating of data is required.

❑ Cannot detect Novel (Zero-day) attacks.

INTRUSION DETECTION SYSTEM

Anomaly Based Detection

- Identifies the anomalies from normal behavior
- Able to detect Zero-Day Attack
- Profiles of normal activity are customized for every system

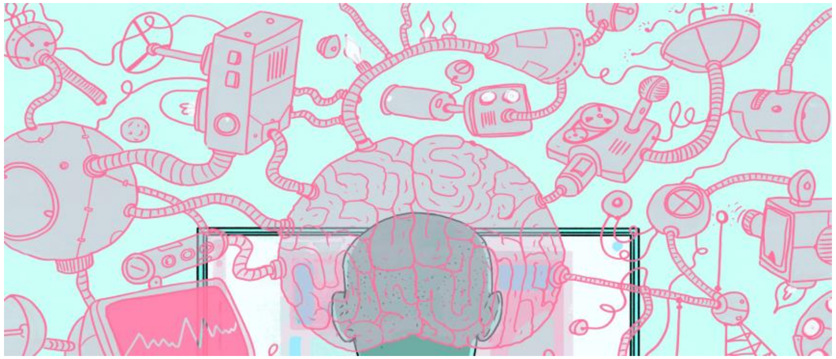
Hybrid Detection

- Combination of misuse and anomaly detection.
- Increases the detection rate and decreases the false alarm generation.

MACHINE LEARNING (ML) VS DATA MINING (DM)

Machine Learning :

- It gives ability to computers to learn without being explicitly programmed.
- Need of goal from domain
- There should be three phases :
  - Training
  - Validation
  - Testing



Data Mining :

- ❑ Focused on discovery of previously unknown and important properties in data.
- ❑ Used for extracting patterns from data

Summary

- ❑ Statistics: Quantifies numbers
- ❑ Data Mining: Explains patterns
- ❑ Machine Learning: Predicts with models
- ❑ Artificial Intelligence: Behaves and reasons

CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP- DM) MODEL

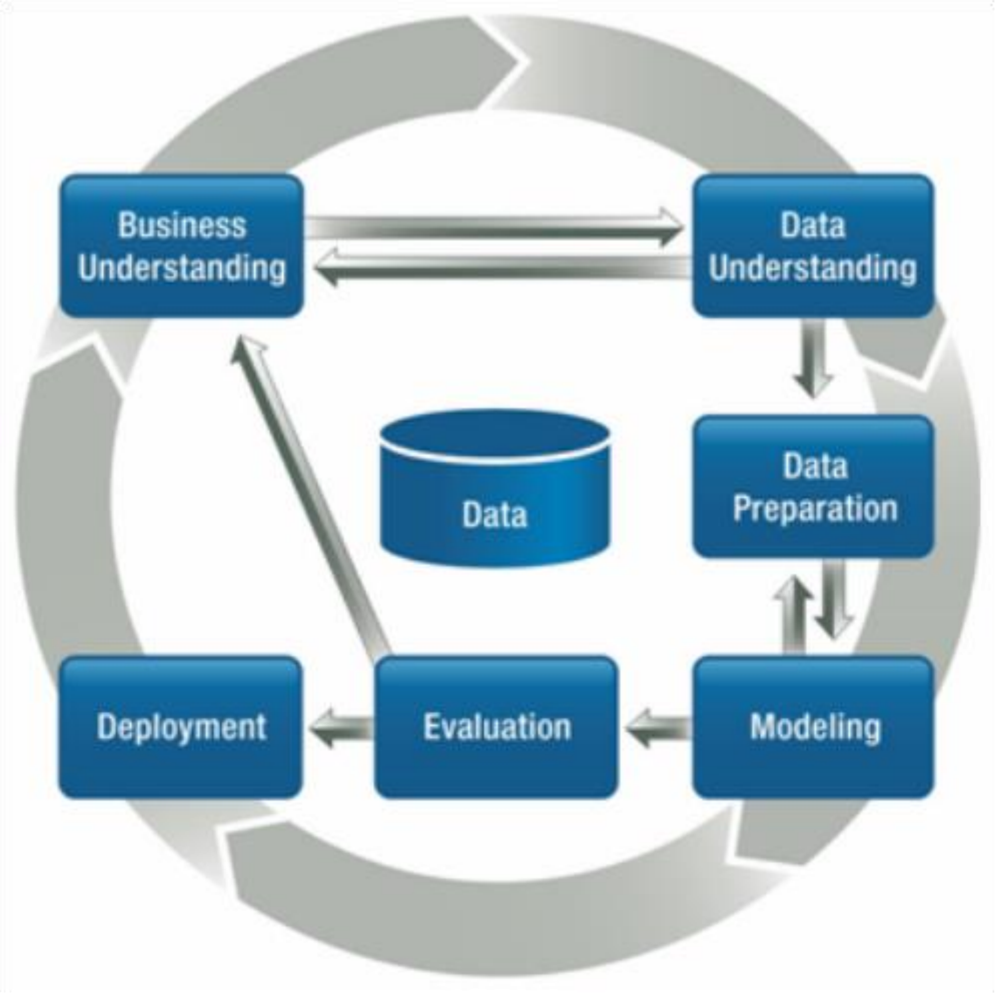
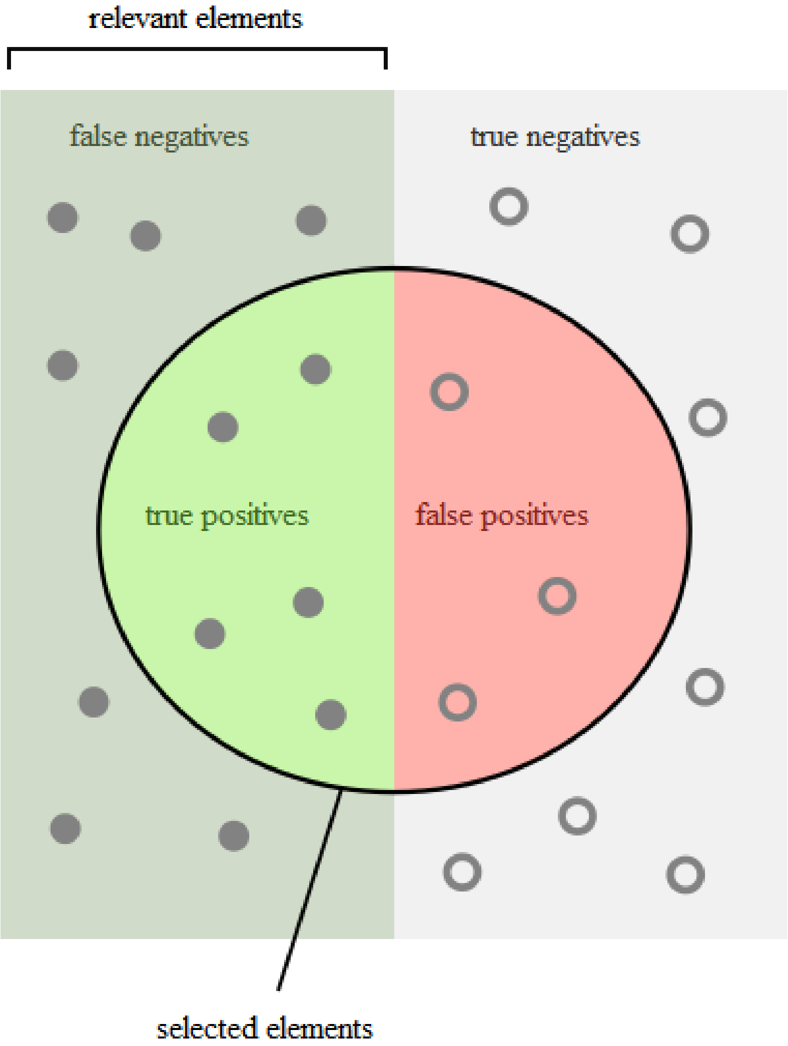
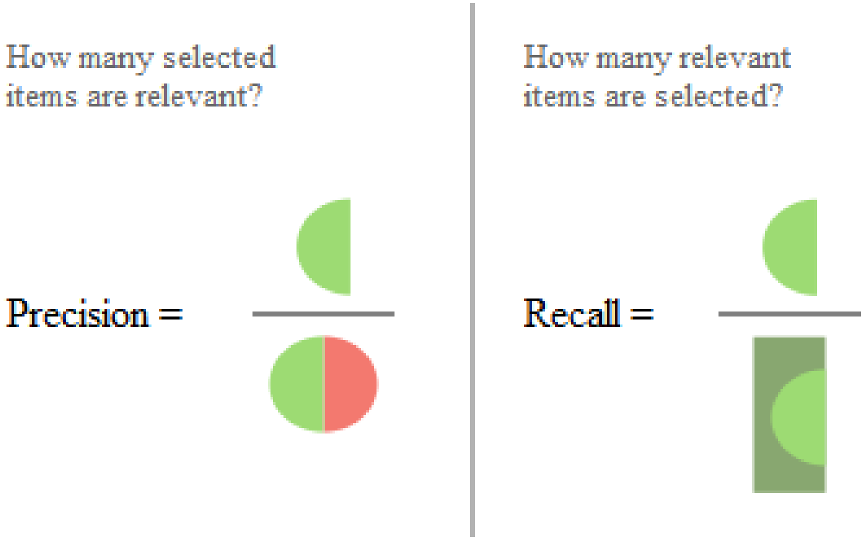


Fig : CRISP-DM Model

METRICS USED IN BINARY CLASSIFICATION




How many selected items are relevant?

Precision =

How many relevant items are selected?

Recall =

CYBER SECURITY DATA SETS FOR DM & ML

The Cyber Security data sets for DM and ML are given below :

- ❑ **Packet Level Data**
- ❑ **Netflow Data**
- ❑ **Public Data Sets**

PACKET LEVEL DATA

- ❑ Protocols are used for transmission of packet through network.

- ❑ The network packets are transmitted and received at the physical interface.

- ❑ Packets are captured by API in computers called as pcap.

- ❑ For Linux it is Libpcap and for windows it is WinPCap.

- ❑ Ethernet port have payload called as IP payload

NETFLOW DATA

- ❑ **Introduced as a router feature by Cisco.**

- ❑ **Version 5 defines unidirectional flow of packets.**

- ❑ **The packet attributes are : ingress interface, source IP address, destination IP address, IP protocol, source port, destination port and type of services.**

- ❑ **Netflow includes compressed and preprocessed packets.**

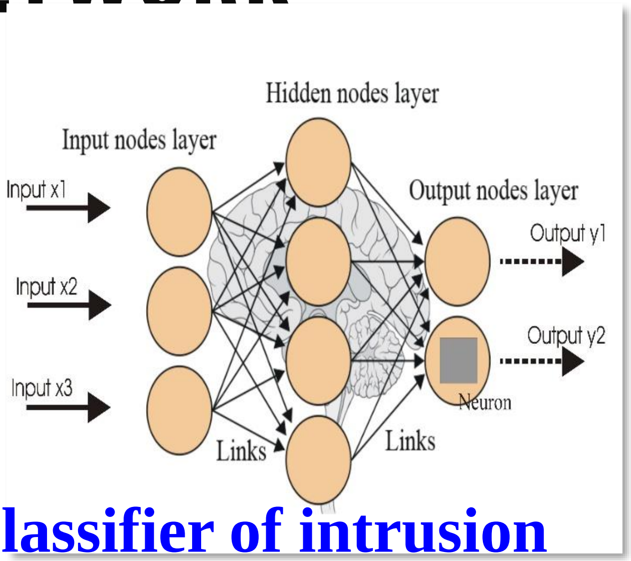
PUBLIC DATA SET

- ❑ The Defense Advance Research Projects Agency (DARPA) in 1998 and 1999 data sets are mostly used.
- ❑ This Data Set has basic features captured by pcap.
- ❑ DARPA defines four types of attacks in 1998 :
- ❑ DoS Attack, User to Root (U2R) Attack, Remote to Local (R2L) Attack, Probe or Scan.

ML & DM METHODS FOR CYBER

- Artificial Neural Networks (ANN)
- Association Rules & Fuzzy Association Rules
- Bayesian Network
- Clustering
- Decision Tree
- Ensemble Learning
- Evolutionary Computation
- Hidden Markov Model
- Inductive Learning
- Nalve Bayes
- Sequential Pattern Mining
- Support Vector Machine

ARTIFICIAL NEURAL NETWORK



- Network of Neurons
- Output of one node is input to other.
- ANN can be used as a multi-category classifier of intrusion detection
- Data processing stage used to select 9 features: protocol ID, source port, destination port, source address, destination address, ICMP type, ICMP code, raw data length and raw data.

BAYESIAN NETWORK

- ❑ It's a probabilistic graphical model that represents the variables and the relationships between them.
- ❑ The network is constructed with nodes as the discrete or continuous random variables and directed edges as the relationships between them, establishing a directed acyclic graph.
- ❑ The child nodes are dependent on their parents.
- ❑ Each node maintains the states of the random variable and the conditional probability form.
- ❑ Bayesian networks are built using expert knowledge or using efficient algorithms that perform inference.

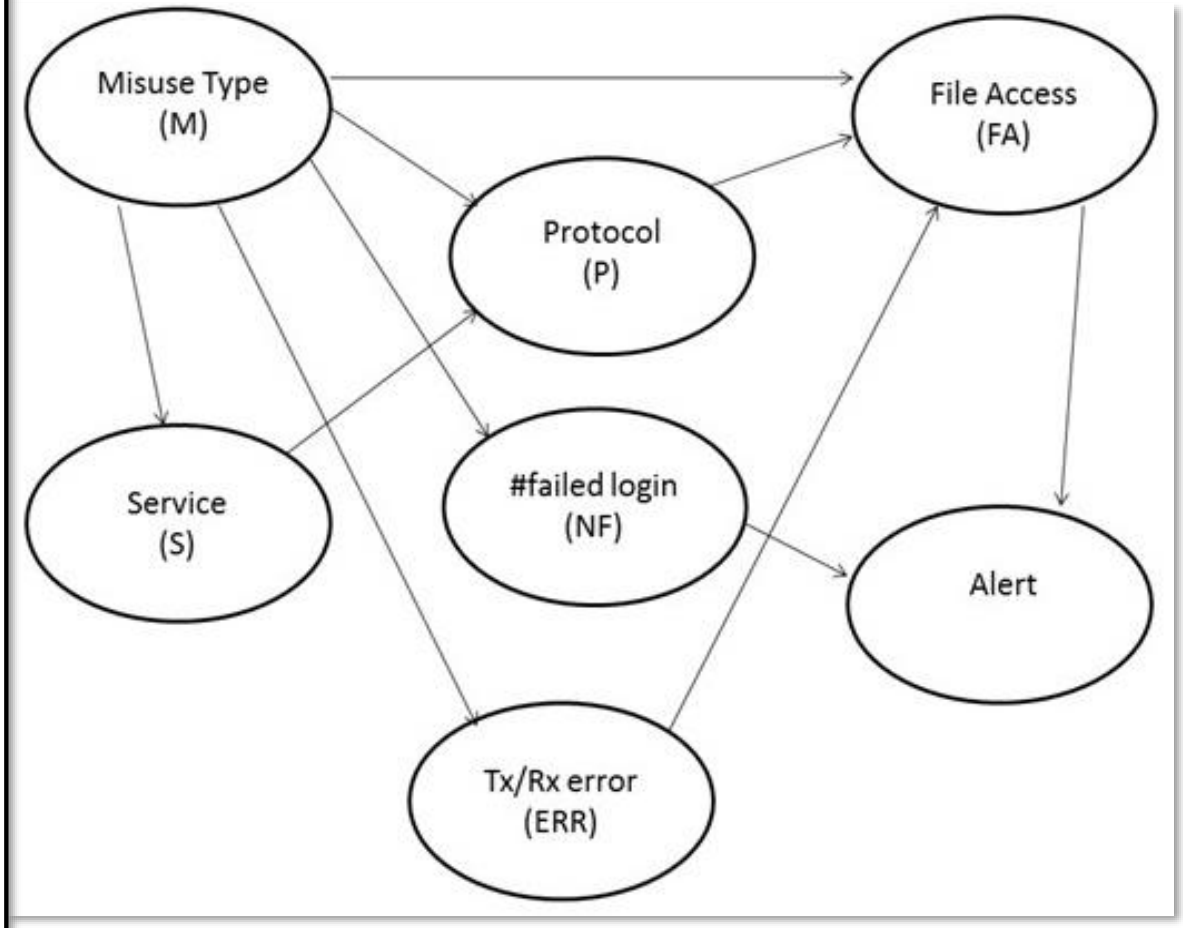


Fig : Bayesian Network for Attack Signature Detection

DECISION TREE

- ❑ A decision tree is a tree-like structure that has leaves, which represent classifications and branches, which in turn represent the conjunctions of features that lead to those classifications.
- ❑ An exemplar is labeled (classified) by testing its feature (attribute) values against the nodes of the decision tree.
- ❑ The best known methods for automatically building decision trees are the ID3 and C4.5 algorithms.

- ❑ **Advantages:** Decision trees are intuitive knowledge expression, high classification accuracy, and simple implementation.
- ❑ **Disadvantage:** Data including categorical variables with a different number of levels, information gain values are biased in favor of features with more levels.

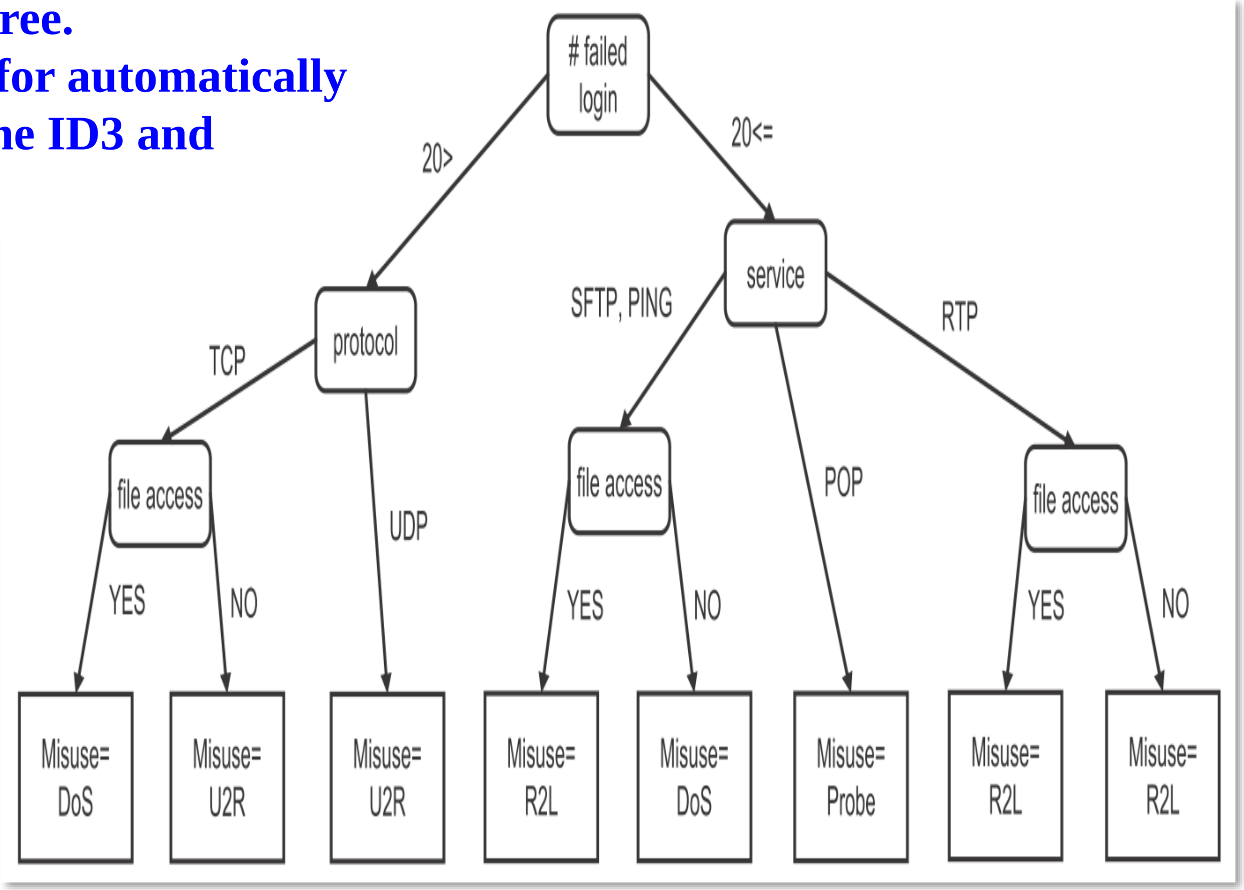


Fig : An Example Decision Tree

HIDDEN MARKOV MODEL

- ❑ Markov chains and Hidden Markov Models (HMMs) belong to the category of Markov models.
- ❑ A Markov chain is a set of states interconnected through transition probabilities that determine the topology of the model.
- ❑ An HMM is a statistical model where the system being modeled is assumed to be a Markov process with unknown parameters.
- ❑ In this example, each host is modeled by four states: Good, Probed, Attacked, and Compromised.
- ❑ The edge from one node to another represents the fact that, when a host is in the state indicated by the source node, it can transition to the state indicated by the destination

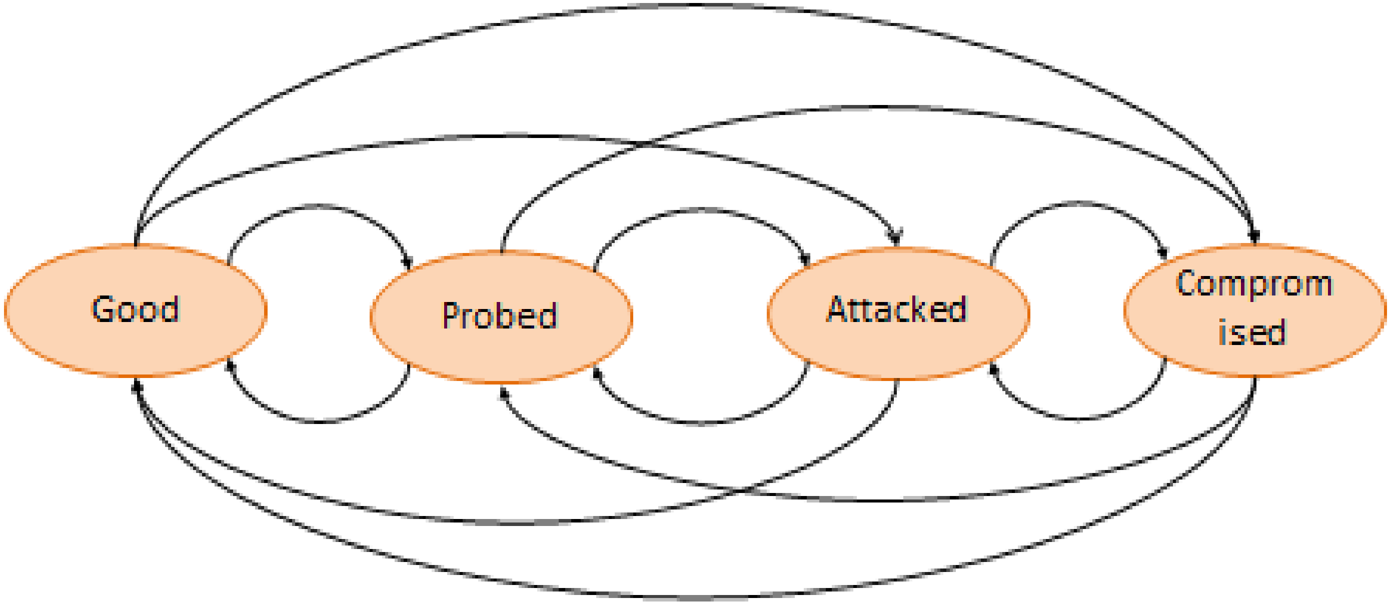
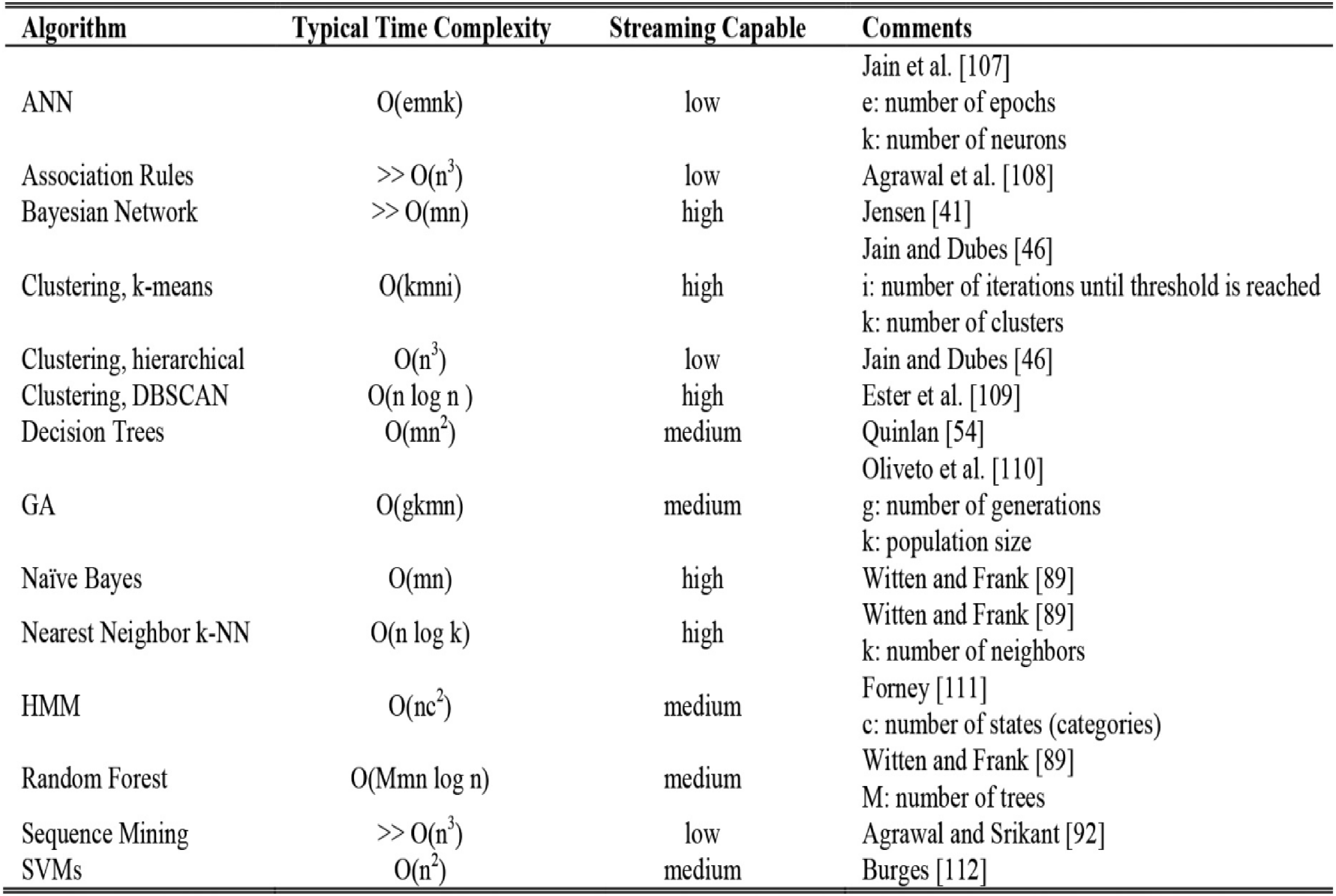


Fig : Hidden Markov Model

COMPLEXITY OF ML & DM ALGORITHMS DURING TRAINING



Algorithm	Typical Time Complexity	Streaming Capable	Comments
ANN	O(emnk)	low	Jain et al. [107] e: number of epochs k: number of neurons
Association Rules	>> $O(n^3)$	low	Agrawal et al. [108]
Bayesian Network	>> O(mn)	high	Jensen [41]
Clustering, k-means	O(kmni)	high	Jain and Dubes [46] i: number of iterations until threshold is reached k: number of clusters
Clustering, hierarchical	$O(n^3)$	low	Jain and Dubes [46]
Clustering, DBSCAN	O(n log n)	high	Ester et al. [109]
Decision Trees	$O(mn^2)$	medium	Quinlan [54]
GA	O(gkmn)	medium	Oliveto et al. [110] g: number of generations k: population size
Naïve Bayes	O(mn)	high	Witten and Frank [89]
Nearest Neighbor k-NN	O(n log k)	high	Witten and Frank [89] k: number of neighbors
HMM	$O(nc^2)$	medium	Forney [111] c: number of states (categories)
Random Forest	O(Mmn log n)	medium	Witten and Frank [89] M: number of trees
Sequence Mining	>> $O(n^3)$	low	Agrawal and Srikant [92]
SVMs	$O(n^2)$	medium	Burges [112]

OPINION

- ❑We can use Deep learning method to achieve more accuracy for cyber security intrusion detection.
- ❑But, again processing time of data will be a big challenge.
- ❑To address this issue, we may use graph partition method to train and update the dataset in partial way.

CONCLUSION

- ❑ Here, we discuss the literature review of ML and DM methods used for Cyber Security.

- ❑ Different ML and DM techniques in the cyber domain can be used for both Misuse Detection and Anomaly Detection.

- ❑ There are some peculiarities of the cyber problem that make ML and DM methods more difficult to use.

- ❑ They are especially related to how often the model needs to be retrained.

- ❑ In most ML and DM applications, a model (e.g., classifier) is trained and then used for a long time, without any changes.

REFERENCES

- Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153-1176

Thank you!