Machine Learning and Data Mining for Cyber-Security

Hermes Senger

Agenda

- Linear models – Chapter 3.2
- Linear models – Chapter 4.6
- Extending linear models – Chapter 7.2
  - Support vector machines, kernel ridge regression, kernel perceptrons
  - Multilayer perceptrons and radial basis function networks
  - Gradient descent

Cyber security

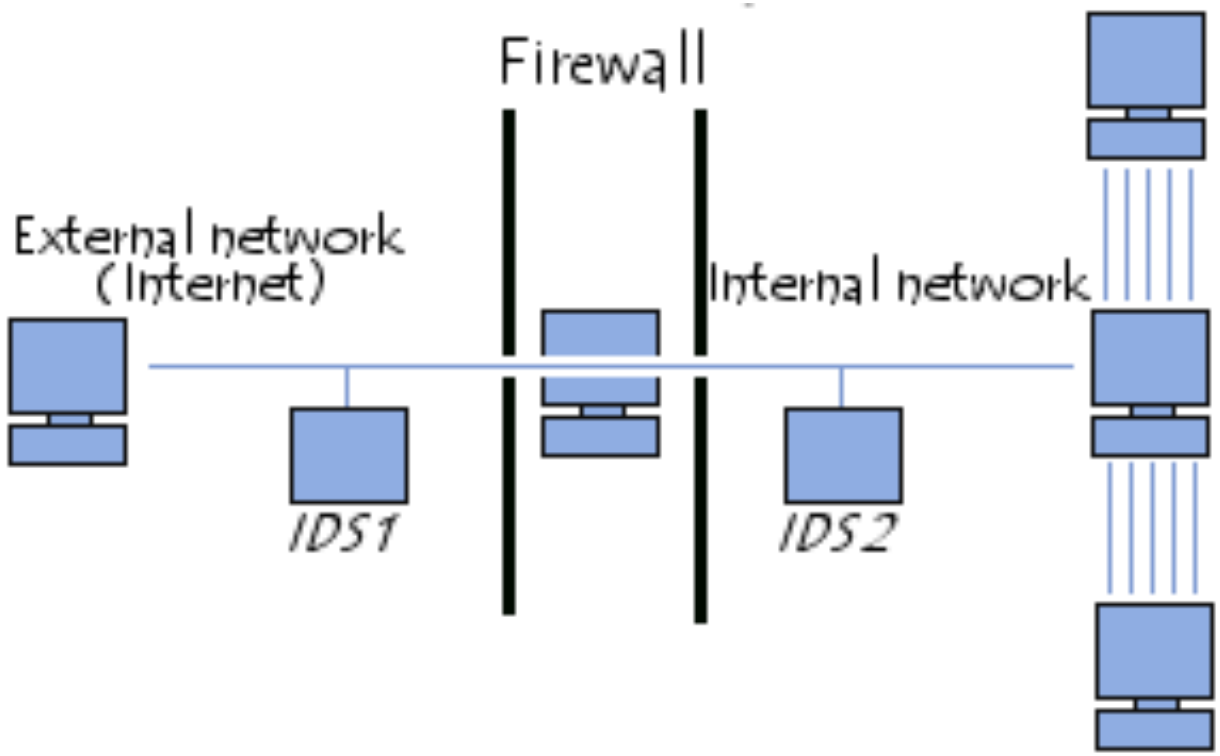
- *Cyber security is the set of technologies and processes designed to protect computers, networks, programs, and data from attack, unauthorized access, change, or destruction* [Buczak 2016]
  - Host security
  - Network security
- Examples
  - Firewall
  - antivirus software
  - intrusion detection system (IDS), Intrusion prevention system (IPS)
  - Malware detection
  - ...

Intrusion Detection Systems (IDS)

- Inspect network traffic to “detect” intrusion and malicious behavior
  - Physical appliances
  - Virtual appliances

Intrusion Detection Systems (IDS)

- Three main types of cyber analytics for IDSs:
  - **Misuse-based** (sometimes also called signature-based),
  - **Anomaly-based**
  - **hybrid**



IDS Types

- **Misuse-based** techniques
  - Detect known attacks by using signatures
  - Need frequente updates
  - Based on string matchiung
  - Well known patterns
  - Ex:
    - SNORT
    - YARA
    - ...

YARA

```
rule silent_banker : banker
{
    meta:
        description = "This is just an example"
        thread_level = 3
        in_the_wild = true
    strings:
        $a = {6A 40 68 00 30 00 00 6A 14 8D 91}
        $b = {8D 4D B0 2B C1 83 C0 27 99 6A 4E 59 F7 F9}
        $c = "UVODFRYSIHLNWPEJXQZAKCBGMT"
    condition:
        $a or $b or $c
}
```

```
alert tcp $EXT_NET any -> $HOME_NET 53
(msg:"DNS named version attempt";
flow:to_server,established;
content:"|07|version"; offset:12; nocase;)
```

SNORT

IDS Types

- **Anomaly-based**
  - model the "**normal**" network and system behavior
  - identify "**anomalies**" as deviations from normal behavior
  - Potential to detect **zero-day** attacks
  - profiles of normal activity are customized for every system, application, or network ...
  - difficult for attackers to know which activities they can carry out undetected ...
  - ML and DM based methods
- Detected anomalies (alerts, novel attacks) can be used to define the signatures for misuse detectors
- **Disadvantages**
  - the potential for high false alarm rates (FARs)
    - previously unseen (yet legitimate) system behaviors may be categorized as anomalies.
- Examples

IDS Types

- **Hybrid IDS**
  - Combine misuse and anomaly detection
    - Raise detection rates of known intrusions
    - Decrease the false positive (FP) rate for unknown attacks
  - Most of the methods in literature are hybrid
    - Pure anomaly detection methods are rare

IDS types

- Host IDS (HIDS)
  - Detection based on software behavior:
    - Resource consumption
    - Sequence of System calls
  - Ex:
    - Microsoft Windows Defender ATP
    - Many comercial anti-virus
- Network IDS (NIDS)
  - Detection based on network traffic

IDS Metrics

- **Accuracy** or Proportion Correct: (TP + TN)/ (TP + TN + FP + FN).
- Positive Predictive Value (PPV) or **Precision**: TP/(TP + FP).
- Sensitivity or **Recall** or True Positive Rate or Probability of Detection (PD) or Detection Rate: TP/(TP + FN)
- **Negative Predictive Value** (NPV): TN/(TN + FN)
- **Specificity** or TN Rate: TN/(TN + FP).
- **FAR or FP Rate or Fall-out**: FP/(TN + FP).
- …

Data collection

- **Packet level** data
  - ~ 150 Internet protocols listed by IETF
  - Capture libraries: *pcap* (Libpcap, MinPCap), snifers, tcpdump, Wireshark, Snort, nmap, etc.

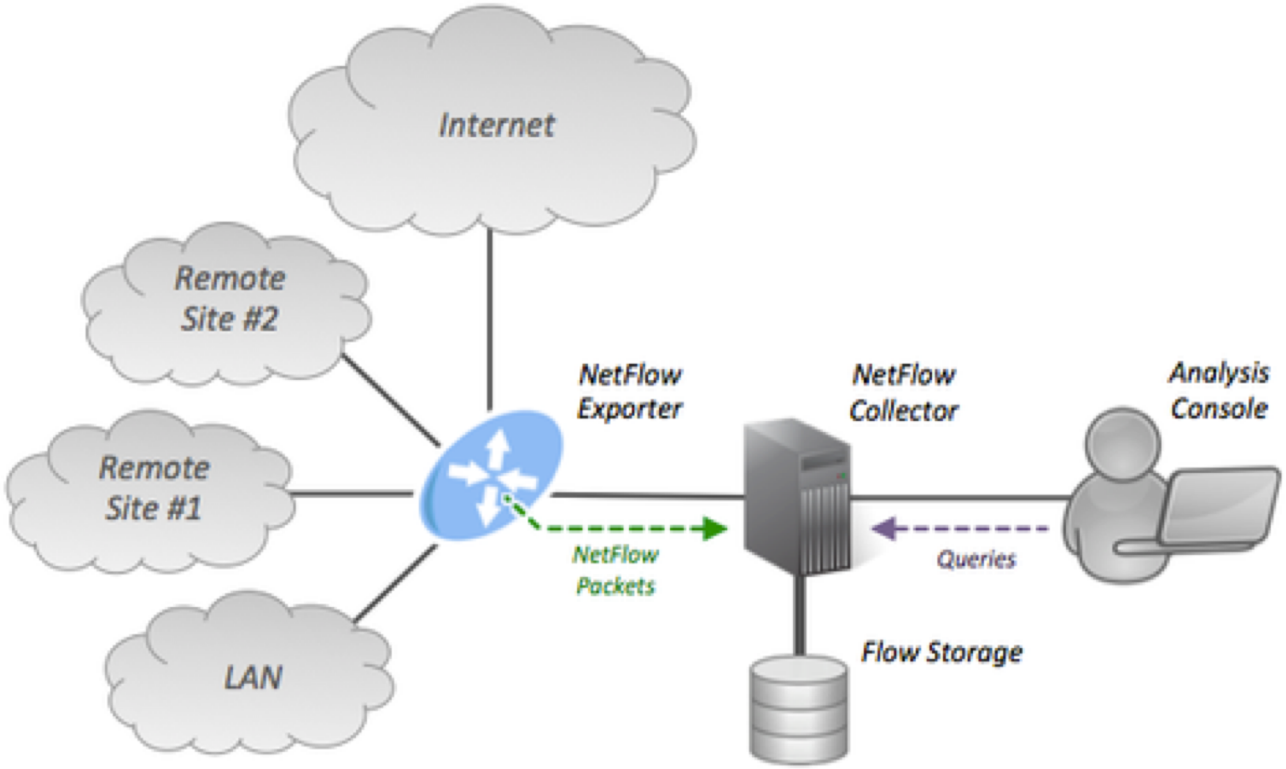
Typical packet headers used by IDS systems

PACKET HEADERS OF CYBER-SECURITY DATASETS

IP Header (IPv4)	
Internet Header Length	The number of 32-bit words in the header
Total Length	The entire packet size, including header and data, in bytes
Time To Live	This field limits a datagram's lifetime, in hops (or time)
Protocol	The protocol used in the data portion of the IP datagram
Source address	This field is the IPv4 address of the sender of the datagram
Destination address	This field is the IPv4 address of the receiver of the datagram
TCP Packet	
Source port	Identifies the sending port
Destination port	Identifies the receiving port
Sequence number	Initial or accumulated sequence number
Acknowledgement number	The next sequence number that the receiver is expecting
Data offset	Specifies the size of the TCP header in 32-bit words
Flags (control bits)	NS, CWR, ECE, URG, ACK, PSH, RST, SYN, FIN
UDP Packet	
Source port	Identifies the sending port
Destination port	Identifies the receiving port
Length	The length in bytes of the UDP header and UDP data
ICMP Packet	
Type	Control (e.g., ping, destination unreachable, trace route)
Code	Details with the type
Rest of Header	More details

Which data types?

- **Flow level** based IDS
  - NetFlow was originally introduced as a router feature by Cisco
- Router/switch collects IP network traffic as it enters and exits the interface
- Much more scalable than packet capture
  - Ex: 1/100 packe



Flow based data

- Network flow: a unidirectional sequence of packets that share the exact same seven packet attributes:
  - ingress interface
  - Source/destination IP adres
  - IP protocol
  - Source/destination port
  - IP type of servisse

NETFLOW PACKET HEADER OF CYBER SECURITY DATASETS

NetFlow Data – Simple Network Management Protocol (SNMP)	
Ingress interface (SNMP ifIndex)	Router information
Source IP address	
Destination IP address	
IP protocol	IP protocol number
Source port	UDP or TCP ports; 0 for other protocols
Destination port	UDP or TCP ports; 0 for other protocols
IP Type of Service	Priority level of the flow
NetFlow Data – Flow Statistics	
IP protocol	IP protocol number
Destination IP address	
Source IP address	
Destination port	
Source port	
Bytes per packet	The flow analyzer captures this statistic
Packets per flow	Number of packets in the flow
TCP flags	NS, CWR, ECE, URG, ACK, PSH, RST, SYN, FIN

Typical packet headers used by IDS systems

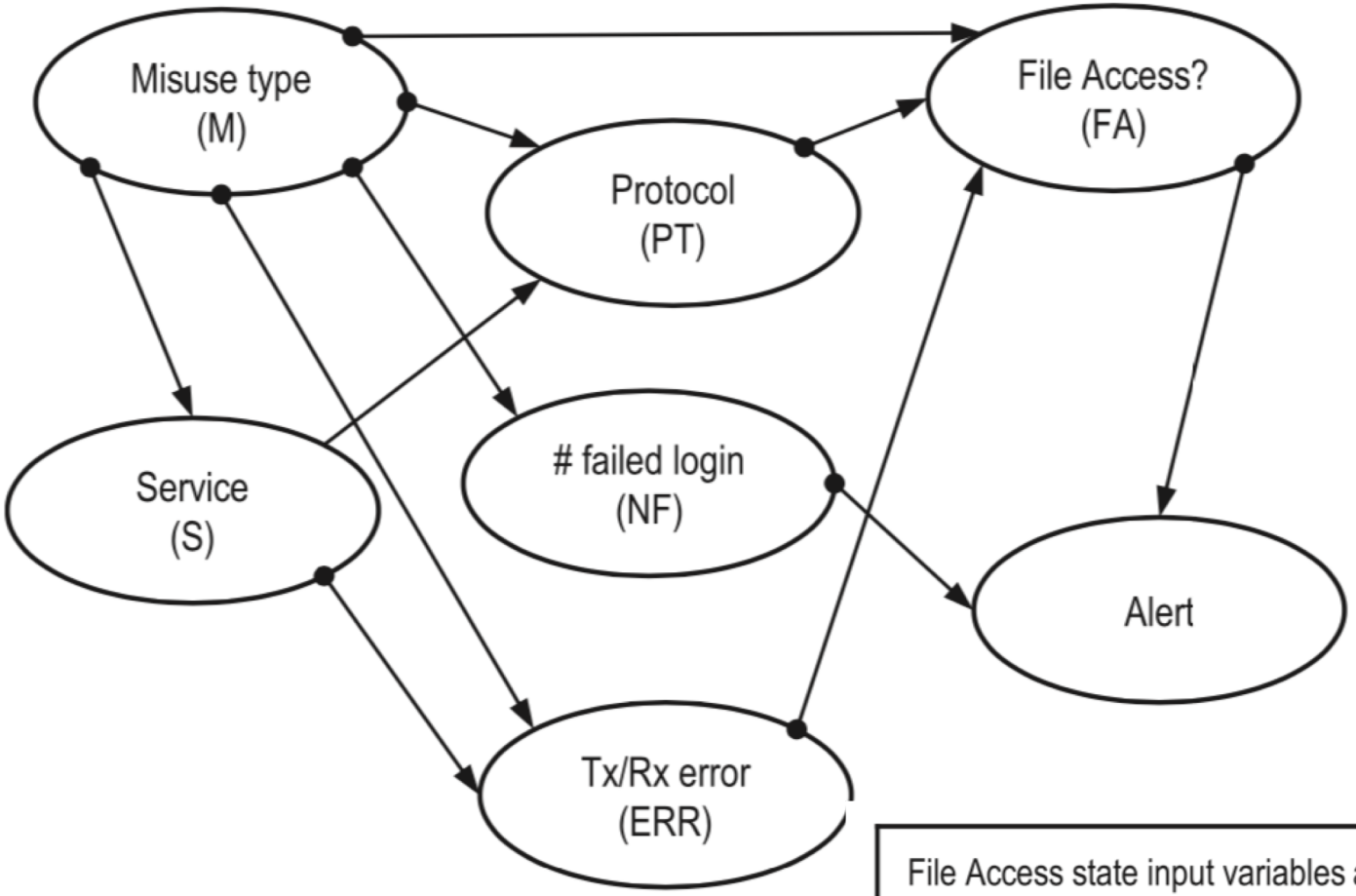
ML and DM Techniques

- Artificial Neural Networks (ANN)
  - Deep learning
- *Association Rules and Fuzzy Association Rules*
- *Bayesian Network*
- Decision trees
  - Ex: ID3, C4.5, SVM, ...
- Clustering
  - DBSCAN, KNN...
- *Evolutionary Computation*
  - E.g., genetic algorithms, particle swarm,
- Ensemble learning
  - Randomn forest, Boosting, etc
- Hidden Markov models

Datasets

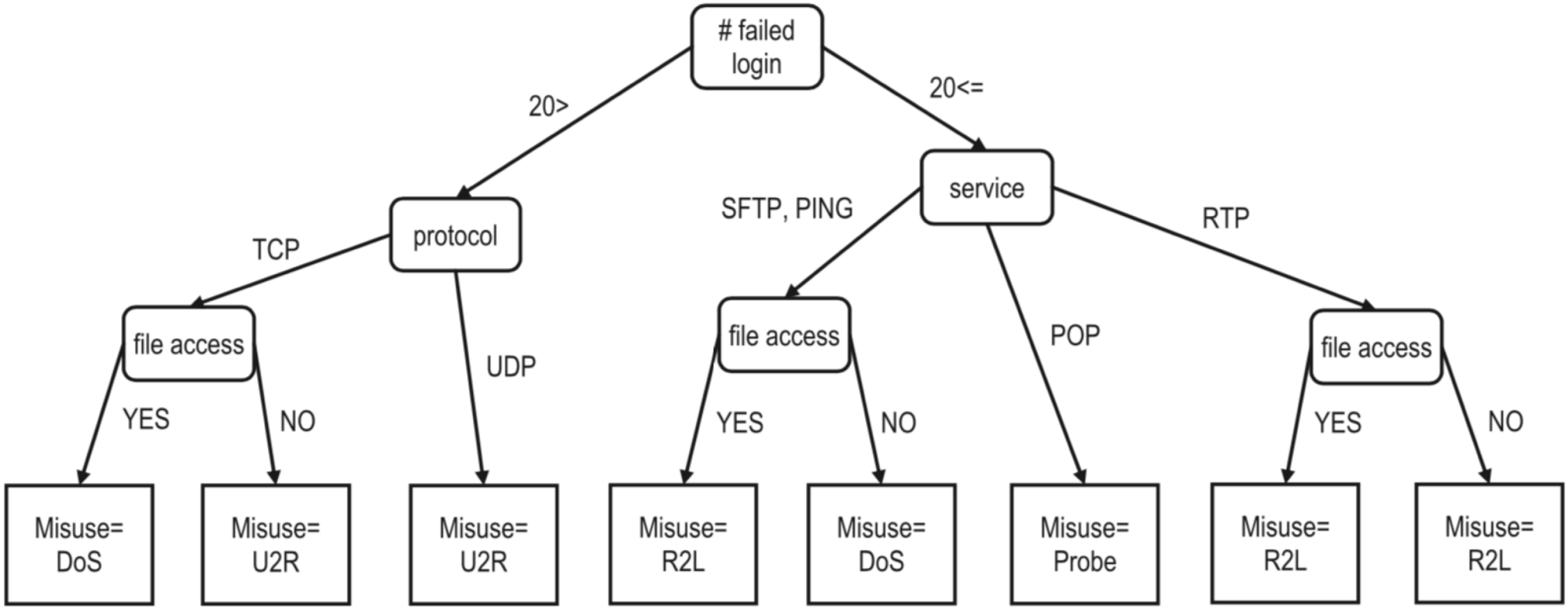
- Labeled datasets available
  - KDDCUP99
  - NSL-KDD
  - ISCX
  - Kyoto Honeypots
  - Kaggle
  - MNIST
  - CIFA-10
- Datasets with labeled data are needed
- Representative data
- Benchmarking

Example of Bayesian Network for Signature Detection



File Access state input variables and values	P(FA = True)	P(FA = False)
M=R2H, PT=NSF, ERR=0	0.95	0.05
M=R2H, PT=FTP, ERR=0	0.99	0.01
M=Probe, PT=none, ERR=50%	0.80	0.20
M=Probe, PT=PING, ERR=0	0.50	0.50
M=DoS, PT=POP, ERR=100%	0.80	0.20
M= DoS, PT=HTTP, ERR=50%	0.90	0.10

Example Decision Tree



Hidden Markov Model

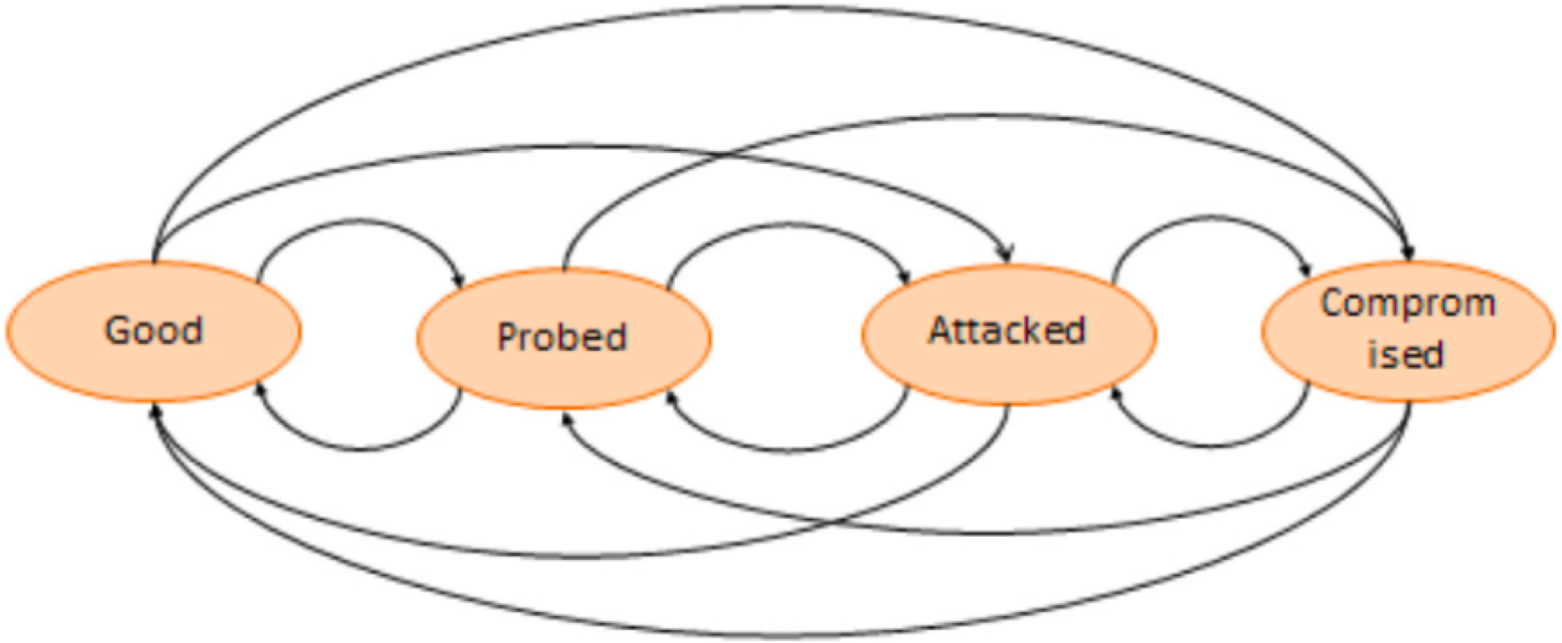


Fig. 5. An Example Hidden Markov Model.

Challenges

- ML and DM techniques have been studied for IDS since 90’s

- How effective?

A few results

- GP – Genetic Programming

SENSITIVITY AND SPECIFICITY OF GP WITH HOMOLOGOUS CROSSOVER

Type of Attack	Sensitivity	Specificity
Smurf	99.93	99.95
Satan	100.00	99.64
IP Sweep	88.89	100.00
Port Sweep	86.36	100.00
Back	100.00	100.00
Normal	100.00	100.00
Buffer Overflow	100.00	100.00
WarezClient	66.67	99.97
Neptune	100.00	99.56

- **Sensitivity** or **Recall** or True Positive Rate or Probability of Detection (PD) or Detection Rate: **TP/(TP + FN)**
- **Specificity** or TN Rate: **TN/(TN + FP)**

COMPARISON OF ACCURACY OF ENHANCED SVM WITH SNORT AND BRO

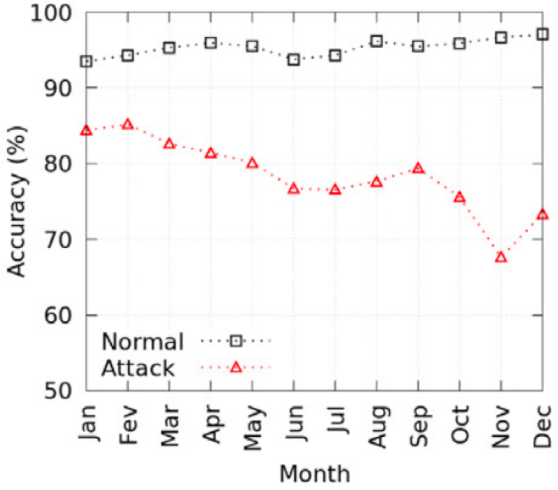
Method	Test Set	Accuracy (%)	FP Rate (%)	FN Rate (%)
Enhanced SVM	Normal	94.19	5.81	0.00
	Attack #1	66.60	0.00	33.40
	Attack #2	65.76	0.00	34.24
	Real	99.90	0.09	0.00
Snort	Normal	94.77	5.23	–
	Attack #1	80.00	–	20.00
	Attack #2	88.88	–	11.12
	Real	93.62	6.38	–
Bro	Normal	96.56	3.44	–
	Attack #1	70.00	–	30.00
	Attack #2	77.78	–	22.22
	Real	97.29	2.71	–

Time complexity

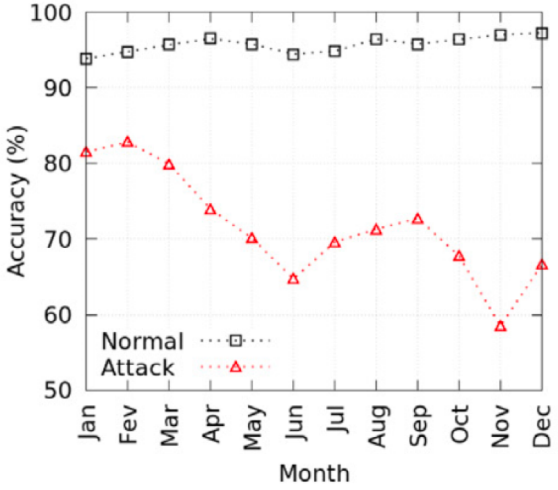
COMPLEXITY OF ML AND DM ALGORITHMS DURING TRAINING

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]

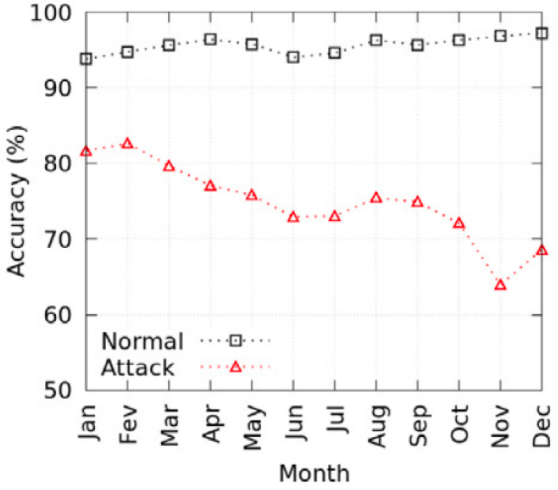
Model Oldening



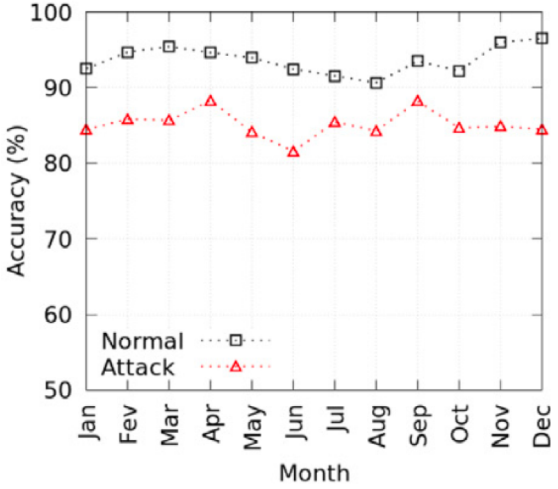
(a) *No-update* decision tree classifier



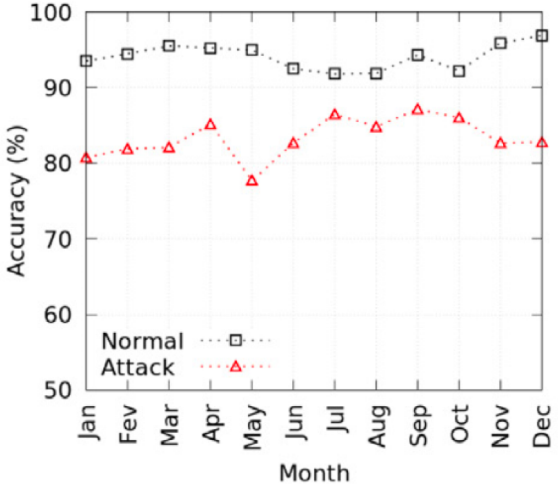
(b) *No-update* random forest classifier



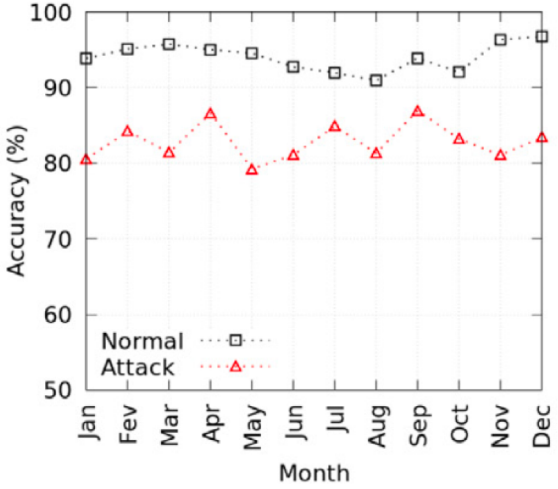
(c) *No-update* gradient boosting classifier



(e) *Weekly-update* decision tree classifier

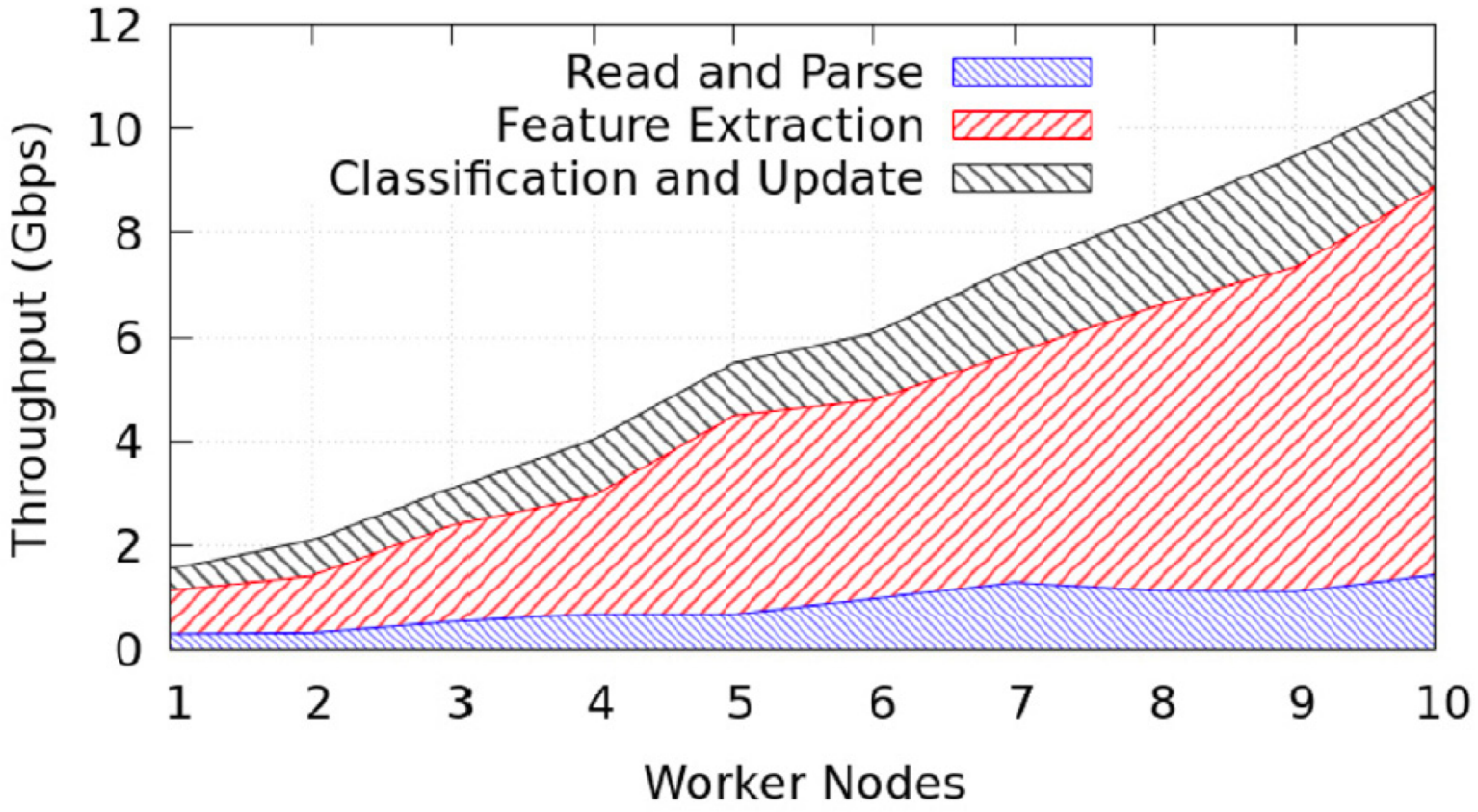


(f) *Weekly-update* random forest classifier



(g) *Weekly-update* gradient boosting classifier

High Throughput



Main challenges

- The methods that are the most effective for cyber applications have not been established
- Richness and complexity of the methods
  - impossible to make one recommendation for each method, based on the type of attack the system is supposed to detect
- Effectiveness of the methods
  - there is not one criterion but several criteria that need to be taken into account

(More) Challenges

- Low detection efficiency, especially due to the high false positive rate usually obtained
- Low throughput and high cost, mainly due to the high data rates (Gbps)
- The absence of appropriate metrics and assessment methodologies to evaluate and compare IDS
- Analysis of ciphered data (e.g. in wireless and mobile environments)

References

1. Buczak, Anna L., and Erhan Guven. "A survey of data mining and machine learning methods for cyber security intrusion detection." *IEEE Communications surveys & tutorials* 18.2 (2015): 1153-1176.
2. Nguyen, Hai-Long, Yew-Kwong Woon, and Wee-Keong Ng. "A survey on data stream clustering and classification." *Knowledge and information systems* 45.3 (2015): 535-569.
3. Haq, Nutan Farah, et al. "Application of machine learning approaches in intrusion detection system: a survey." *IJARAI-International Journal of Advanced Research in Artificial Intelligence* 4.3 (2015): 9-18.
4. Viegas, Eduardo, et al. "Bigflow: Real-time and reliable anomaly-based intrusion detection for high-speed networks." *Future Generation Computer Systems* 93 (2019): 473-485.