

Guardian

**Machine learning framework to detect network
malicious traffic in cloud environments**

Collaborators

Project Supervisor(s):

- **Mr. Mohamad Al Khatib**
- **Mr. Naveed Ahmad**

Student Names

- **Omar Belal - 3710137**
- **Rafif Altayar - 3910162**
- **Bayan Ghannam - 3810127**
- **Aghiad Massarani - 3810225**
- **Abdullah Basalama - 3810177**

Full Thesis Outline

Introduction

problem def and
high-level solution

ML in intrusion
Detection Systems

Network Traffic
Analysis using DPI

Project
requirements

Dataset

- Requirement
- Generation methodologies

Full system design

Extracting features
using DPI

pre-processing

Feature extraction

- Tools & Techniques

Binary
Classification

Multi-classification



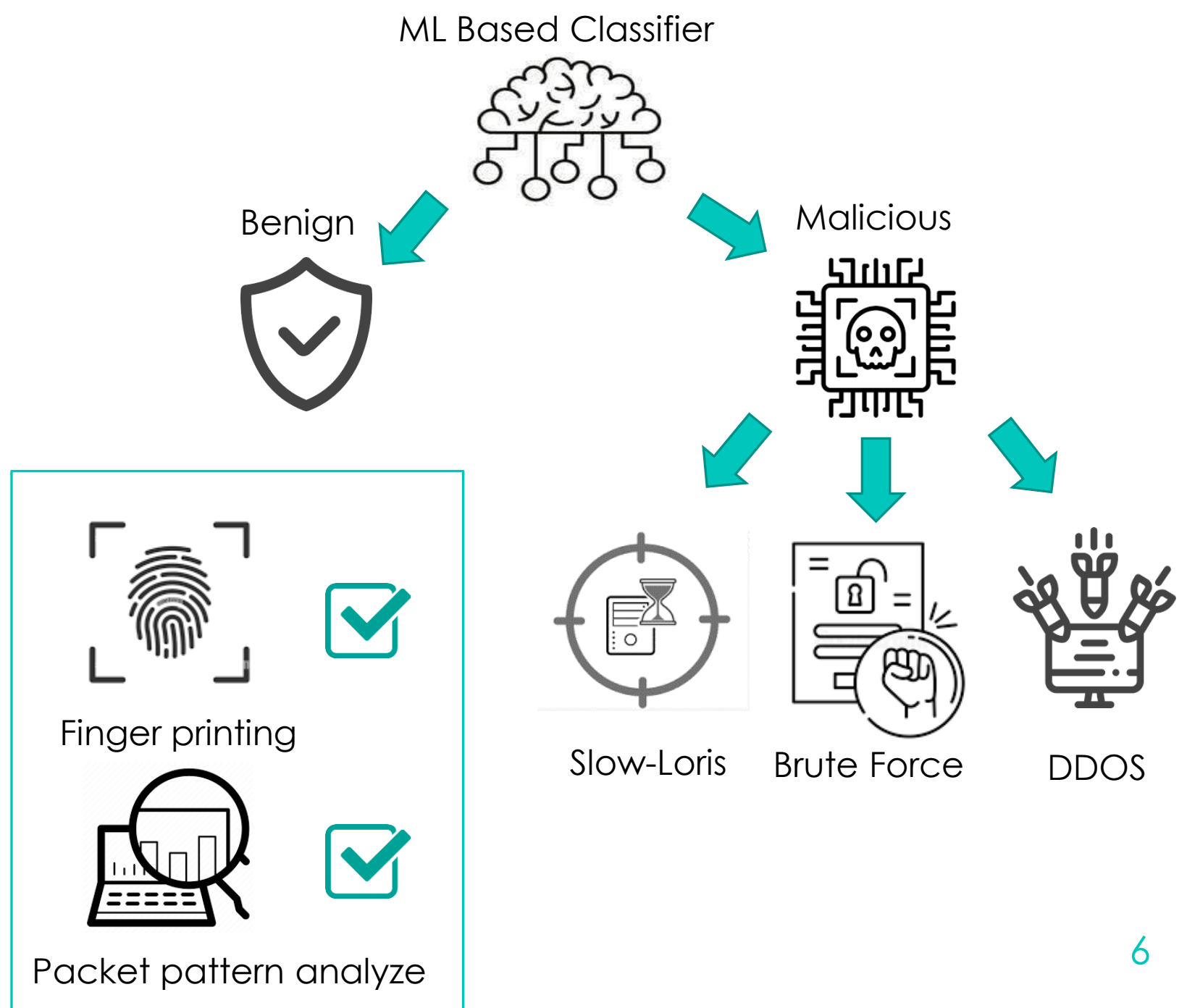
Introduction

- Clouds Data Stores
- Must provide protection and security for it
- Many approaches for network traffic monitoring and threats discovery
- Encryption is a major challenge
- Machine learning is the new approach

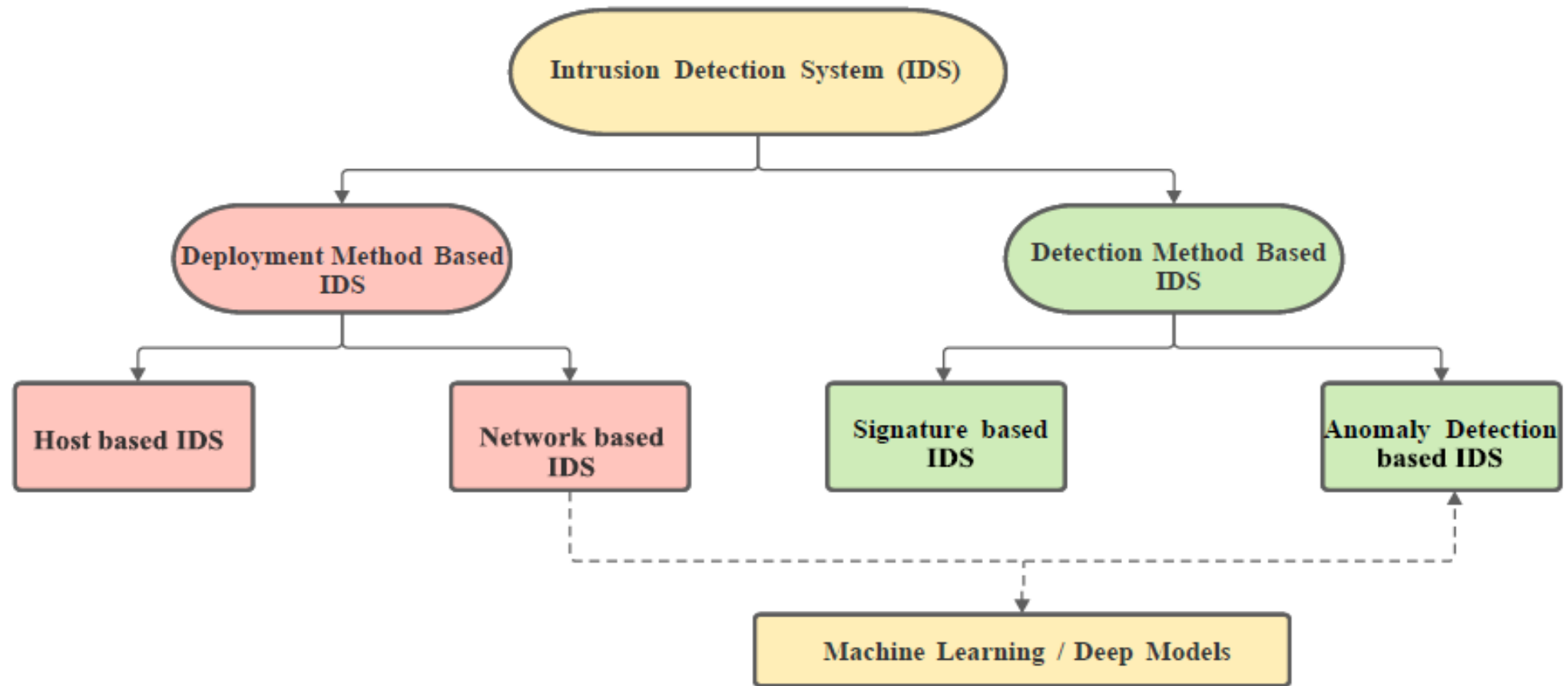
Problem Definition

- Expired Patches for outdated systems
- User Interaction and time efficiency
- Payload encrypted Data in Dataset
- Most available similar solutions are focusing on Packet Headers

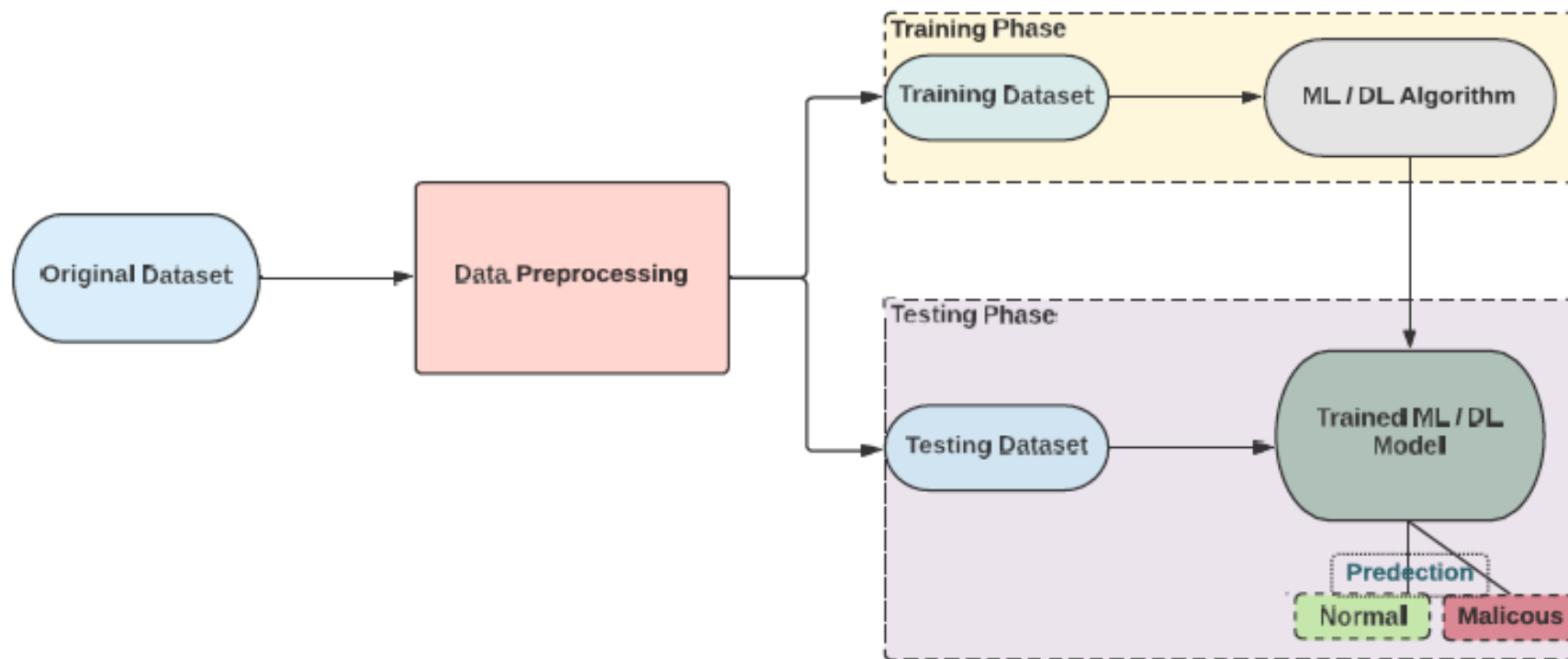
Proposed Solution



Intrusion Detection System based on Machine Learning

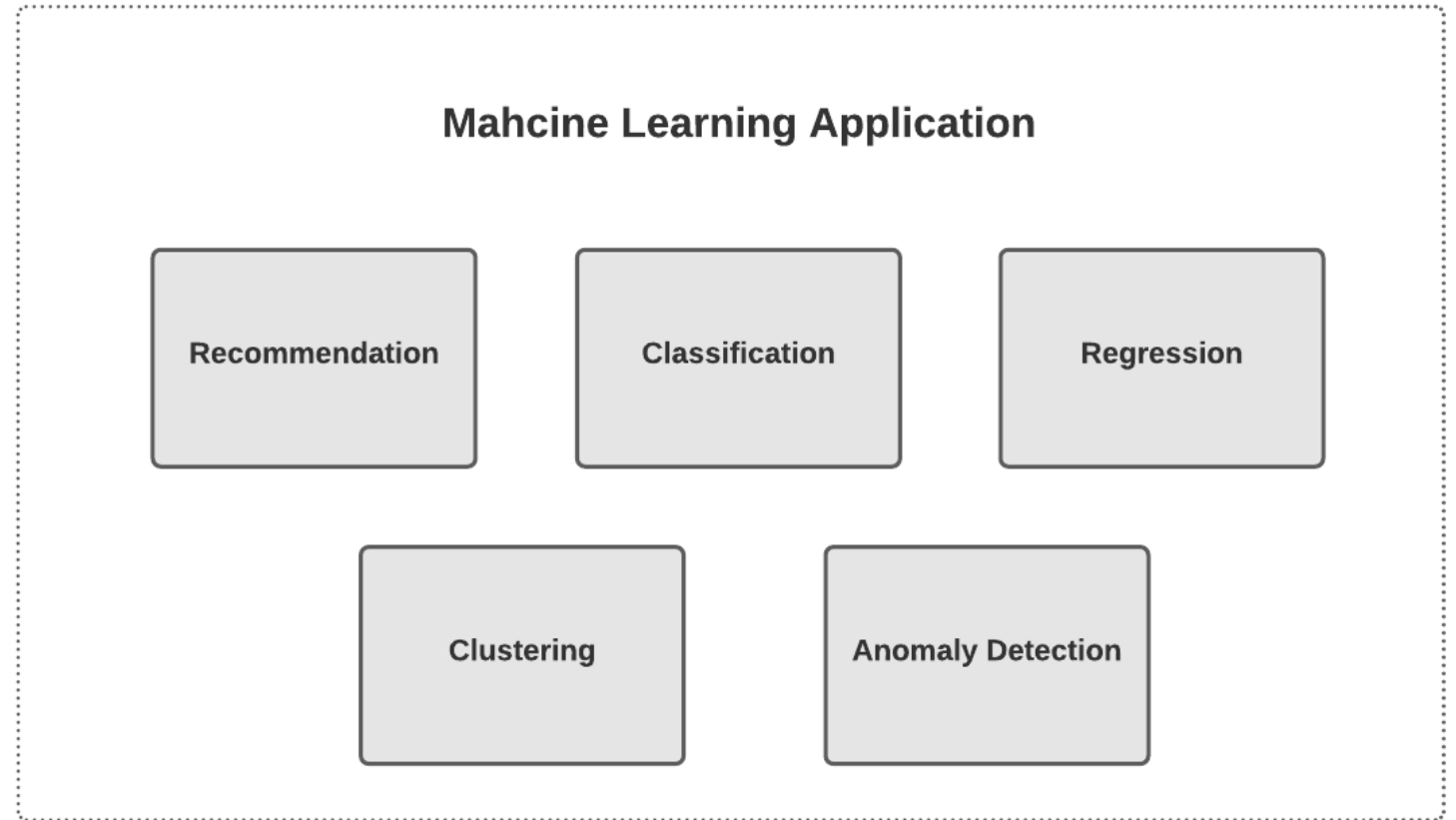


ML based IDS



ML Classification Schema

Machine Learning Application



Tools & Requirements



ID2T



Python



Weka



Pandas



Numpy



sklearn



TensorFlow



Hadoop



Spark



AWS

Datasets

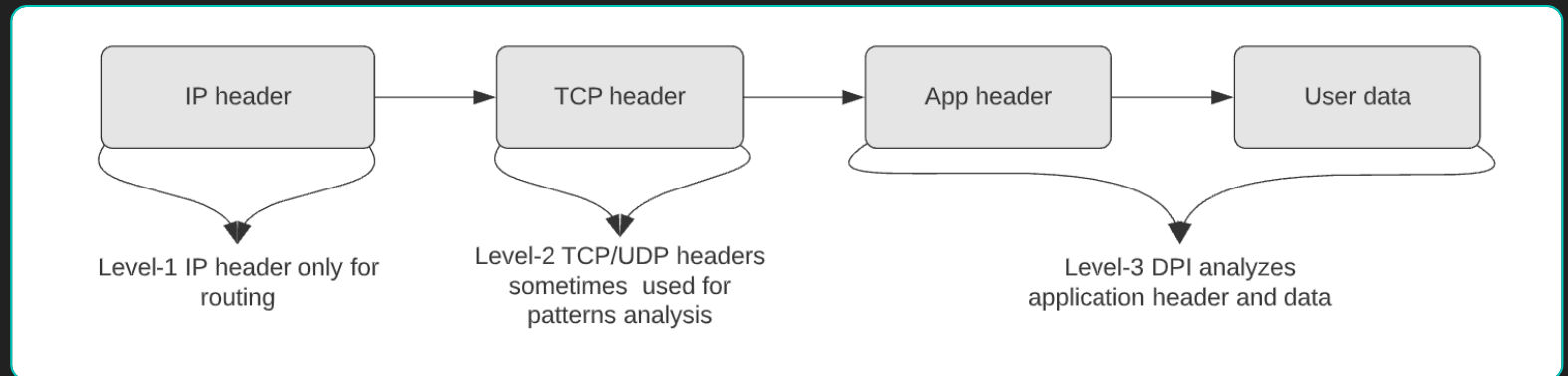
Online Dataset

Generate Dataset

Injected Dataset

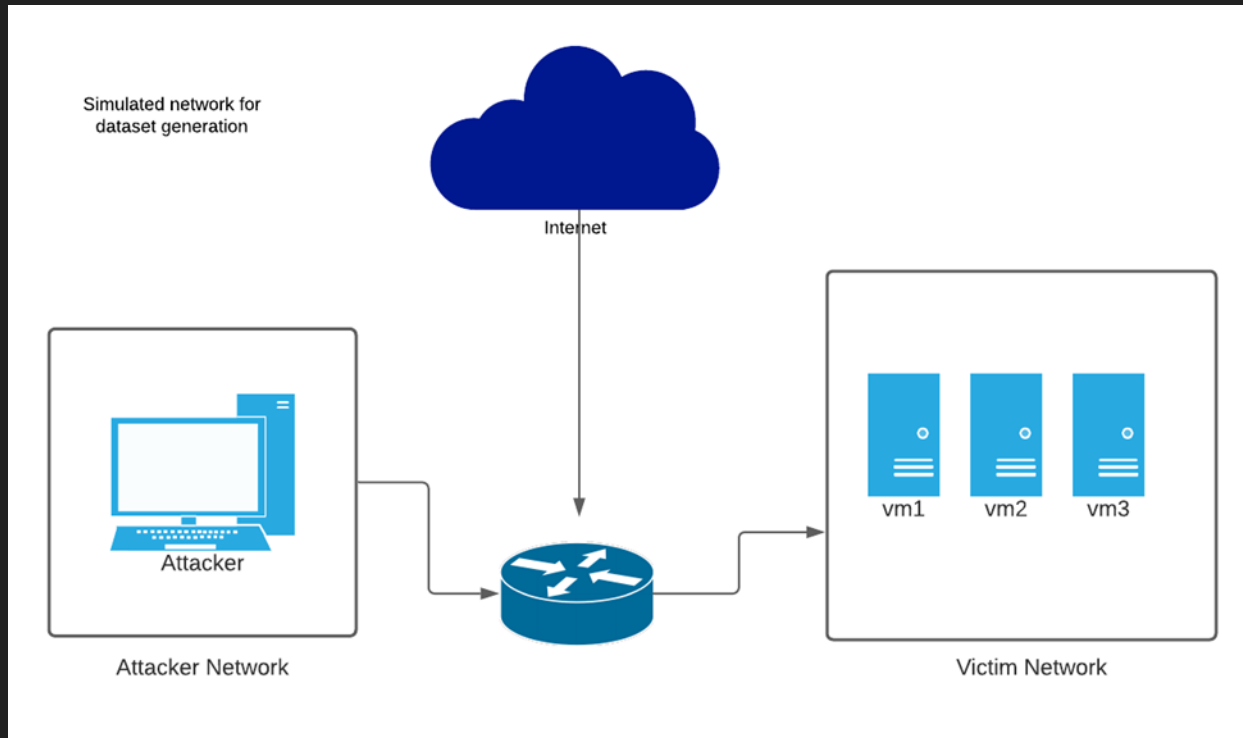
Network Packet structure and DPI

- Layers and levels
- Payload information and the importance
- Encryption

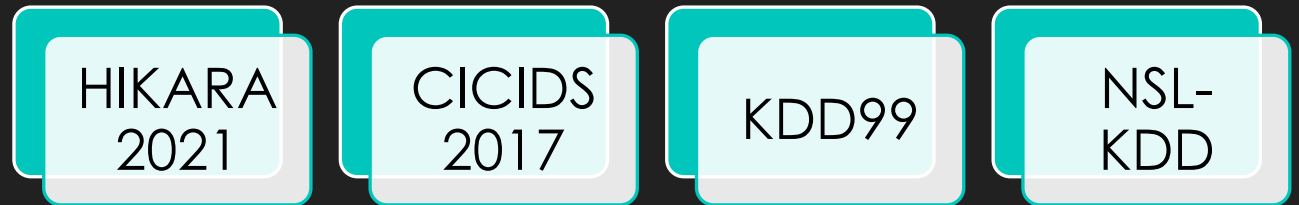


Network Configuration for Dataset Creation

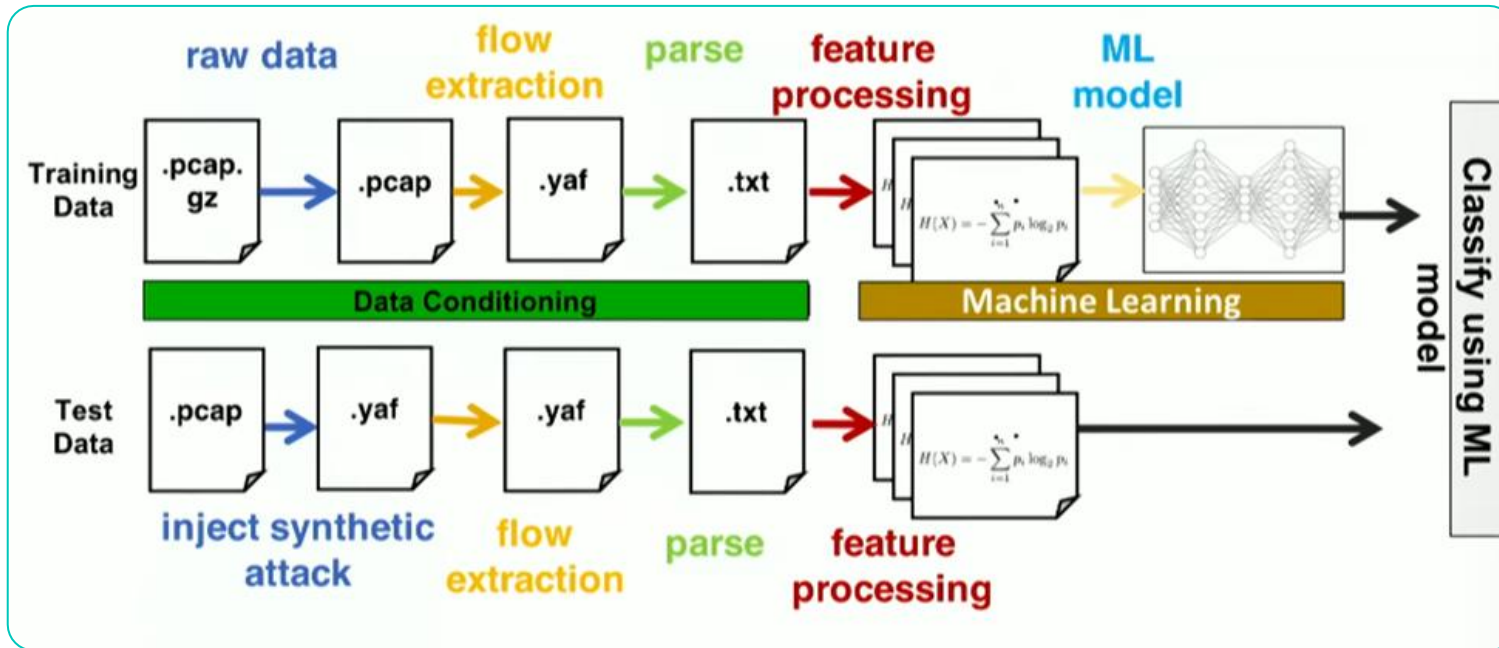
- Classical Network traffic generation methodology
- Varsity of network flow



Valid IDS Dataset



Network malicious traffic injection structure



- Injection tools
- Flexible methodology

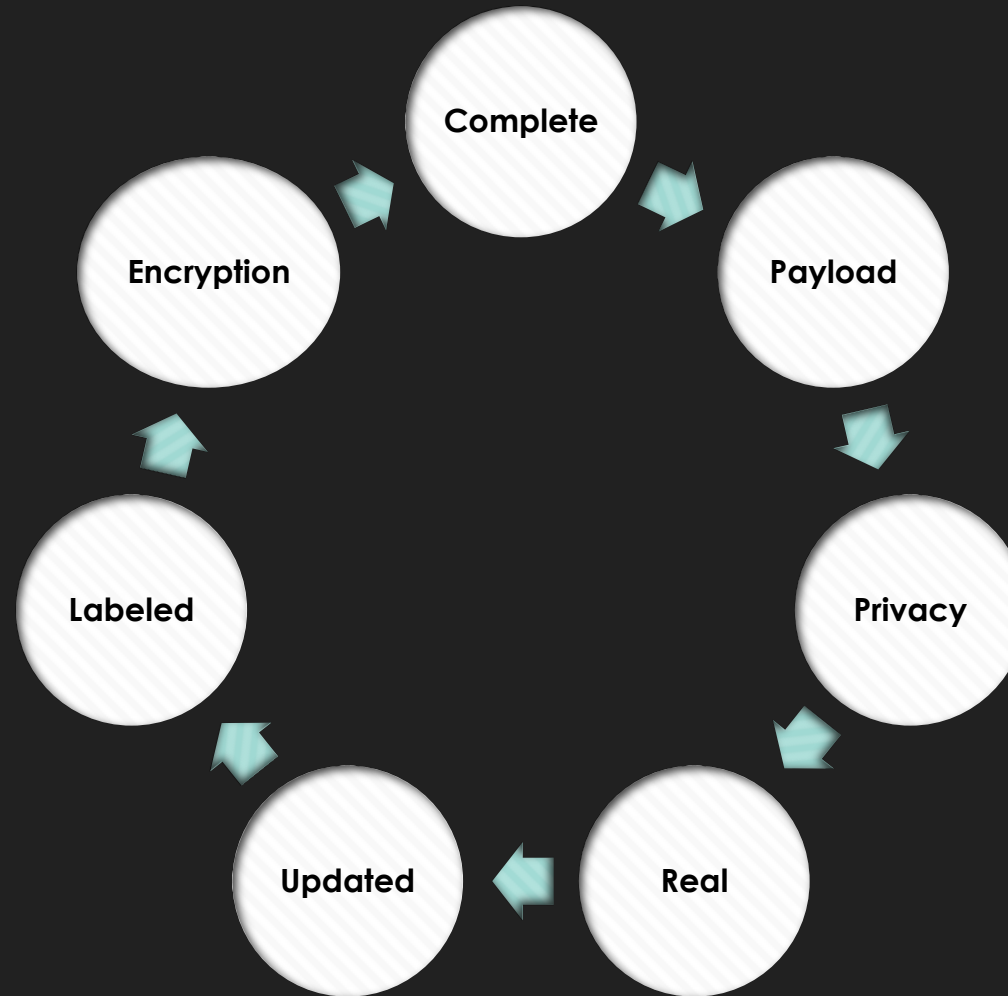
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		Unnamed: 0	uid	originh	originp	responh	responp	flow_duration	fwd_pkts_tot	bwd_pkts_tot	fwd_data_pkts_tot	bwd_data_pkts_tot	fwd_pkts_per_sec	bwd_pkts_per_sec	flow_pkts_per_sec	down_up
2	0	0	Cg61Jch3v	103.255.1.1	13316	128.199.2.4	443	2.207588	15	14	6	6	6.794746	6.341763	13.136509	0.933333
3	1	1	CdRlqLWc	103.255.1.1	13318	128.199.2.4	443	15.624266	15	14	6	6	0.960045	0.896042	1.856087	0.933333
4	2	2	CLzp9KhdC	103.255.1.1	13320	128.199.2.4	443	12.203357	14	13	6	5	1.147225	1.065281	2.212506	0.928571
5	3	3	Cnf1YA4iU	103.255.1.1	13322	128.199.2.4	443	9.992448	14	13	6	5	1.401058	1.300983	2.702041	0.928571
6	4	4	C4ZKvv3fp	103.255.1.1	13324	128.199.2.4	443	7.780611	14	14	6	5	1.799345	1.799345	3.598689	1
7	5	5	CyC8D5X7	103.255.1.1	13326	128.199.2.4	443	4.571433	14	13	6	5	3.062497	2.843747	5.906244	0.928571
8	6	6	CEXyM013	103.255.1.1	13328	128.199.2.4	443	2.19264	14	13	6	5	6.384997	5.928926	12.313923	0.928571
9	7	7	CVFc4q26\	103.255.1.1	13330	128.199.2.4	443	16.082514	14	14	6	5	0.870511	0.870511	1.741021	1
10	8	8	CCvZhO2f	103.255.1.1	13332	128.199.2.4	443	13.87324	15	14	6	6	1.081218	1.009137	2.090355	0.933333
11	9	9	CIPZU1mfl	103.255.1.1	13334	128.199.2.4	443	11.331464	14	13	6	5	1.235498	1.147248	2.382746	0.928571
12	10	10	CBTv463IV	103.255.1.1	13336	128.199.2.4	443	9.117416	14	13	6	5	1.535523	1.425843	2.961365	0.928571
13	11	11	C87B4VRM	103.255.1.1	13338	128.199.2.4	443	6.907568	14	13	6	5	2.026763	1.881994	3.908756	0.928571
14	12	12	CDqmaw1	103.255.1.1	13340	128.199.2.4	443	4.69254	15	14	6	6	3.196563	2.983459	6.180022	0.933333
15	13	13	CKPhym3C	103.255.1.1	13342	128.199.2.4	443	2.198671	14	14	6	5	6.367483	6.367483	12.734967	1
16	14	14	CeAyf115\	103.255.1.1	13344	128.199.2.4	443	16.387078	14	13	6	5	0.854332	0.793308	1.64764	0.928571
17	15	15	CMgixZ2D:	103.255.1.1	13346	128.199.2.4	443	14.004266	14	13	6	5	0.999695	0.928289	1.927984	0.928571
18	16	16	CHZDyc0y	103.255.1.1	13348	128.199.2.4	443	11.786487	14	13	6	5	1.187801	1.102958	2.290759	0.928571
19	17	17	CcBMW31	103.255.1.1	13350	128.199.2.4	443	9.579908	15	14	6	6	1.565777	1.461392	3.027169	0.933333
20	18	18	CMfQhm1	103.255.1.1	13352	128.199.2.4	443	7.331628	14	14	6	5	1.909535	1.909535	3.81907	1
21	19	19	C4dc1ll6w	103.255.1.1	13354	128.199.2.4	443	4.889648	14	13	6	5	2.863192	2.658678	5.52187	0.928571
22	20	20	CwTQNM2	103.255.1.1	13356	128.199.2.4	443	2.446596	15	14	6	6	6.130968	5.722236	11.853204	0.933333
23	21	21	Cl8Nw218	103.255.1.1	13358	128.199.2.4	443	18.886364	14	13	6	5	0.741276	0.688327	1.429603	0.928571
24	22	22	CFuetc4pC	103.255.1.1	13360	128.199.2.4	443	16.676013	14	14	6	5	0.839529	0.839529	1.679058	1
25	23	23	CadzA834C	103.255.1.1	13362	128.199.2.4	443	14.471357	14	12	6	5	0.967428	0.829224	1.796653	0.857143
26	24	24	Cegg5P3LY	103.255.1.1	13364	128.199.2.4	443	12.259943	14	13	6	5	1.14193	1.060364	2.202294	0.928571
27	25	25	Cw6qpM1	103.255.1.1	13366	128.199.2.4	443	10.052126	15	13	6	5	1.492222	1.293259	2.78548	0.866667
28	26	26	Co6QPk1ll	103.255.1.1	13368	128.199.2.4	443	6.841417	15	14	6	6	2.192528	2.04636	4.238888	0.933333

Dataset formats

○ Tabular format – human readable

○ Valid extensions : .csv .yaf .json

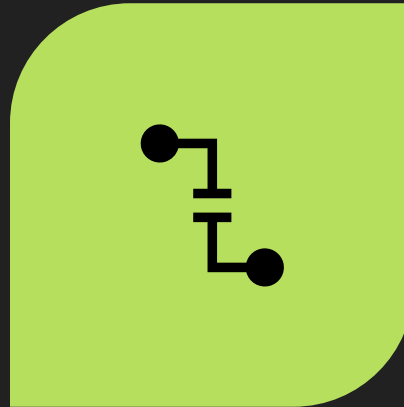
Functional Requirements



Non-functional Requirements



PUBLIC AVAILABILITY



INTEROPERABILITY, STANDER
FORMAT



QUALITY

Dataset Labels

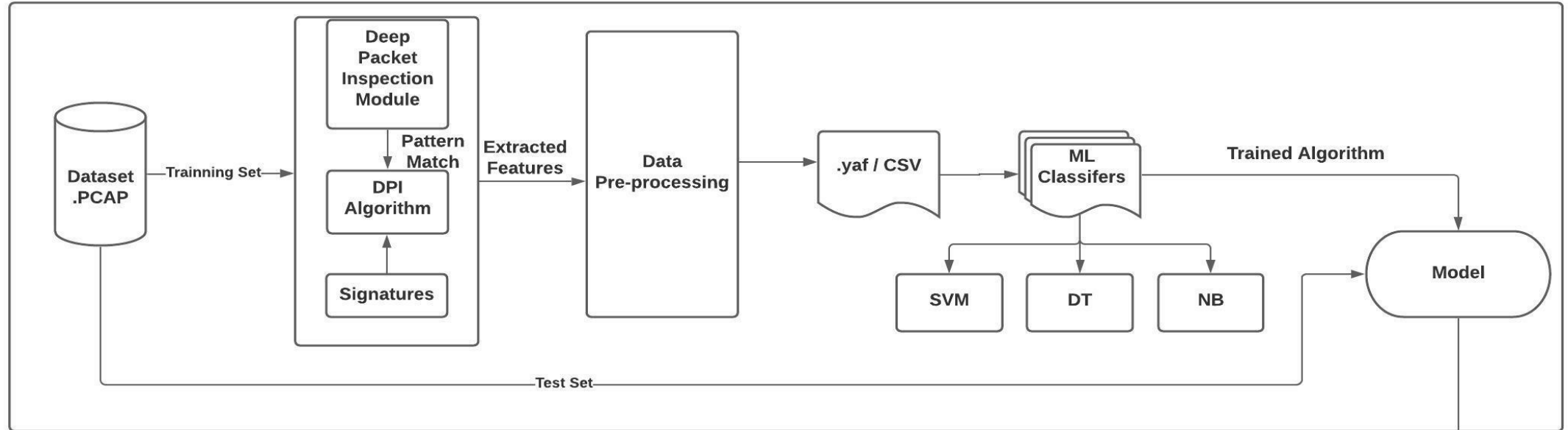
- Labels in binary could be 0 or 1
- Labels in multi-class can be assigned as classes starting by 0 , 1 , 2 ...
- Labels encoding for string names.

Class number	Class name
Class 0	Benign
Class 1	Bot
Class 2	DDos
Class 3	Dos Hulk
Class 4	PortScan
Class 5	Dos slowloris
Class 6	Web Attack-Brute Force

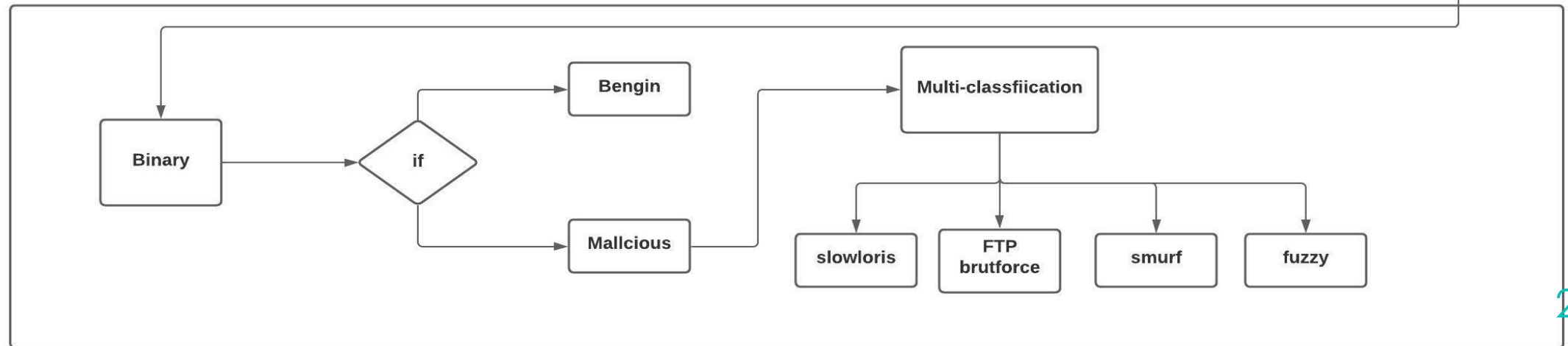
Proposed System Architecture

Full System Design

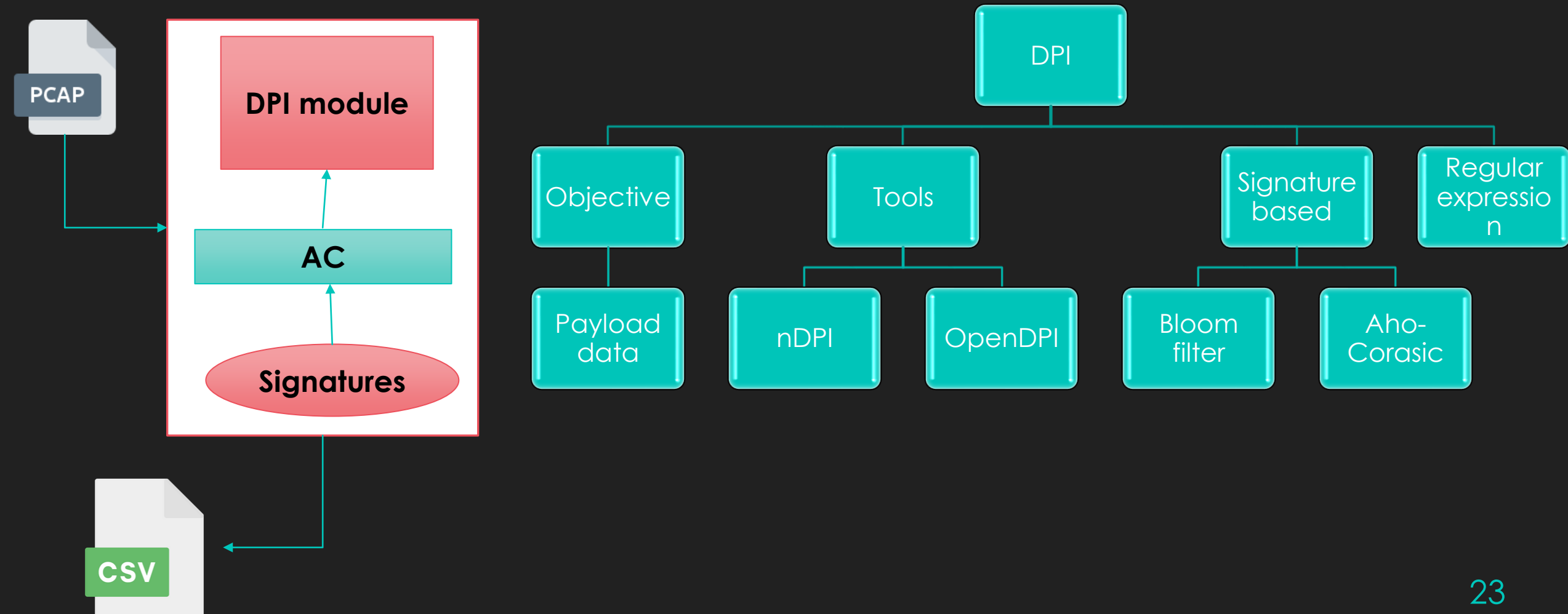
Data Conditioning



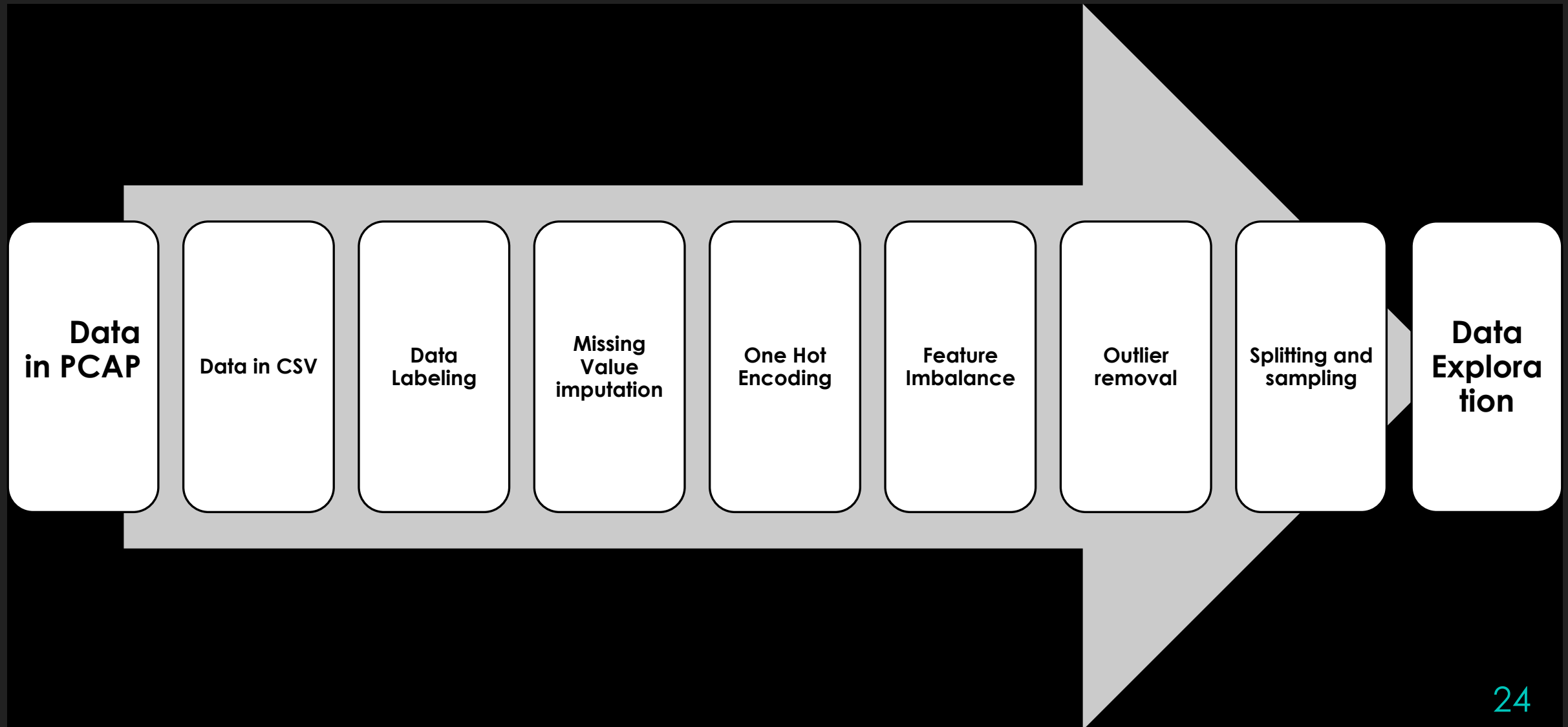
Classification



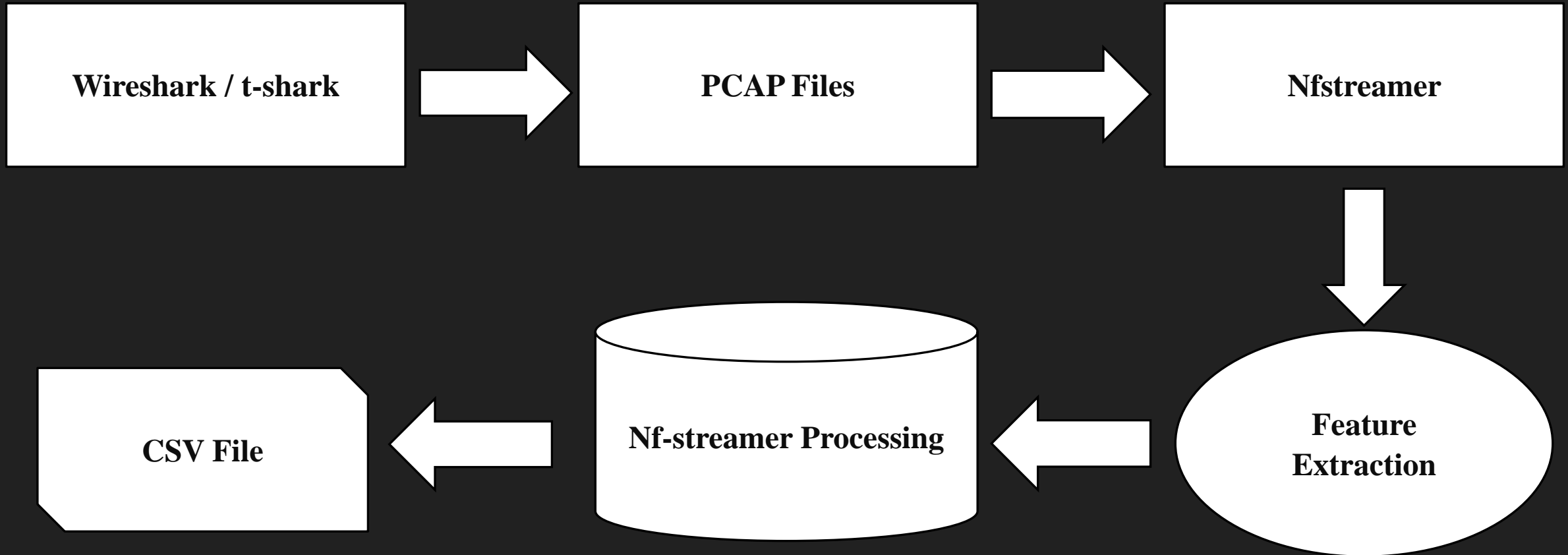
Deep Packet Inspection



Pre-processing Phase / data cleaning



Feature Extrication



	protocol	sourceIPAdress	destinationIPAdress	timeStamp	len	data	target
0	HTTP	167.50.30.156	2.30.147.253	08-02-2019 10:10:28	70	d0ecc25418e1aca2	3
1	TCP/IP	178.76.229.13	113.144.16.192	01-06-2019 20:33:48	79	69b5faa8cd472e60	3
2	HTTP	237.224.185.211	2.108.37.18	17-06-2019 05:58:46	69	501457d8bff174bf	3
3	SMTP	251.227.74.148	220.3.101.137	05-01-2020 02:40:00	78	0000da9dd1551a42	0
4	UDP	202.178.50.102	214.157.79.86	30-01-2020 21:22:00	83	6c5fa73810164f11	1
...
10863	UDP	75.9.29.33	8.142.189.33	10-09-2019 05:29:49	95	4bfc3dc2f657f1b5	3
10864	TCP/IP	83.219.154.118	175.228.202.120	28-04-2019 17:47:15	86	74d619eee002c033	2
10865	FTP	195.235.123.98	240.86.210.116	10-03-2020 02:17:29	71	3800002006dfab19	0
10866	UDP	176.142.74.84	161.7.87.34	05-05-2019 14:04:11	76	bc7423805a250548	2
10867	FTP	21.164.14.96	25.12.5.126	14-04-2019 22:35:22	58	7e900002964a6098	0

Figure : Extracted Payload Data using BMHP algorithm

Feature Extrication Representatives



Nfstreamer



CICFlowmeter



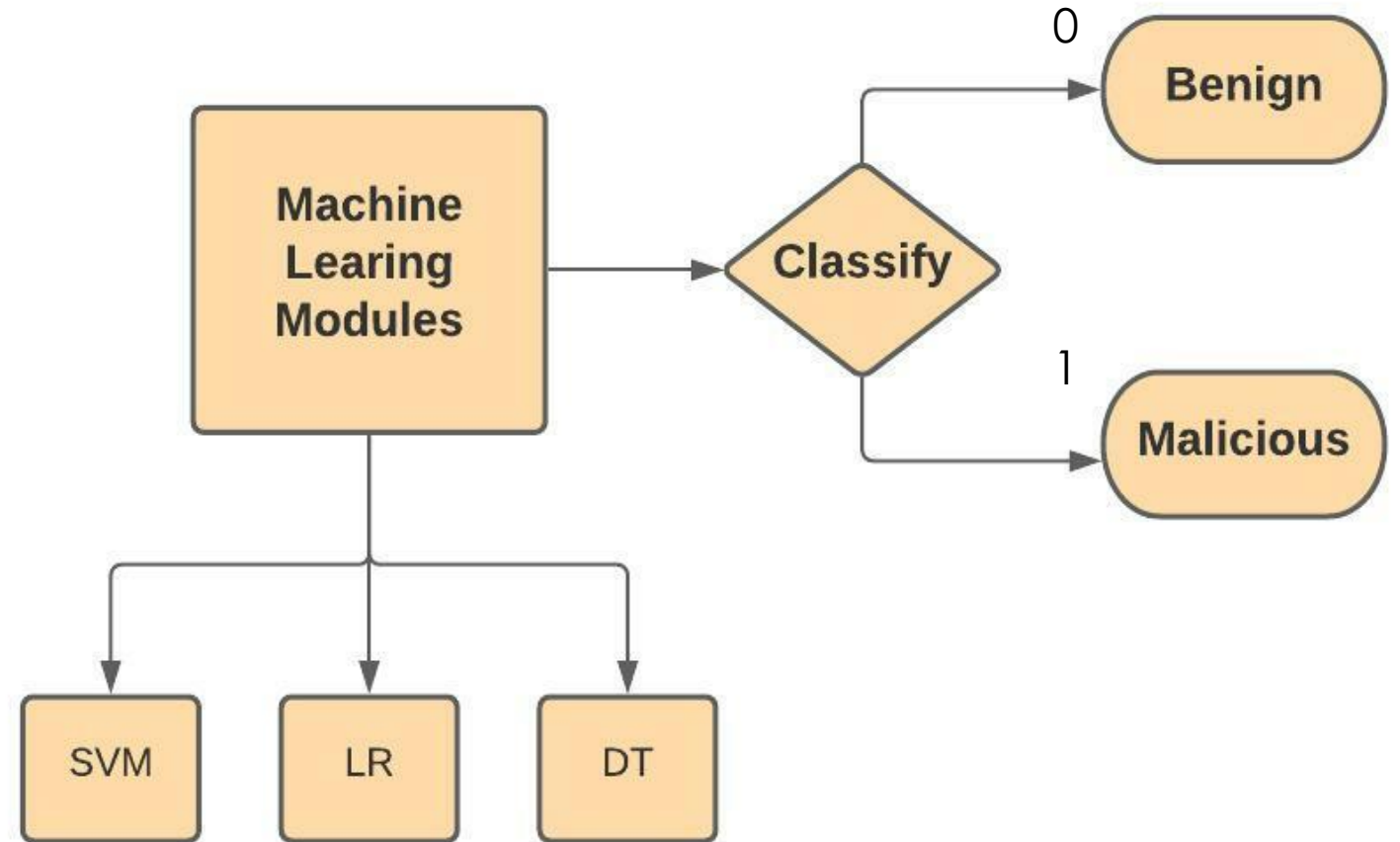
Zeek



custom Script

ML Binary Classification

- Using DPI
- Guided Classifiers
- ML modules

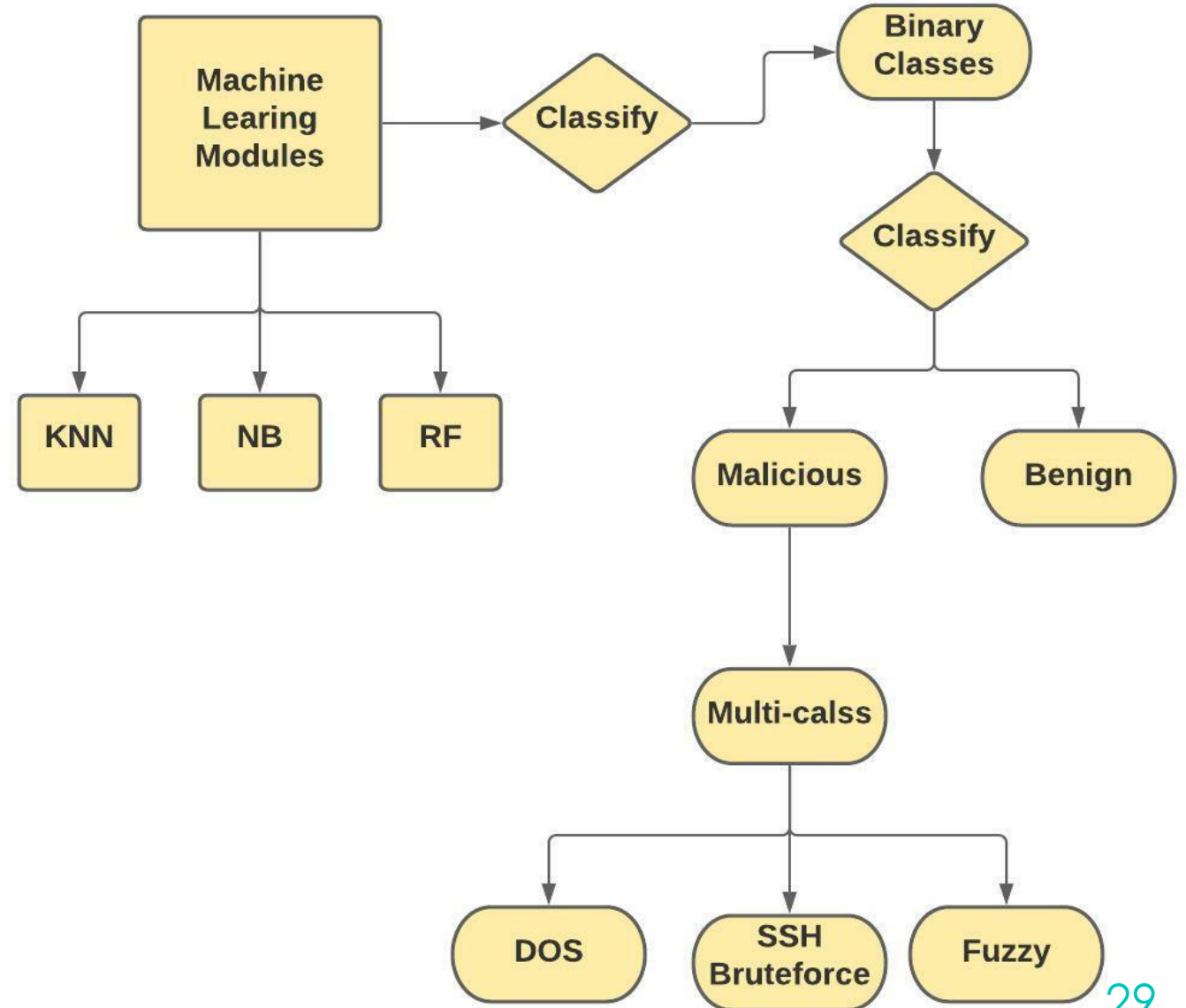


Multi-classification

○ Accuracy Increasing

○ Time Efficiency

○ Deep Learning



Final Key Points

Patch and secure outdated systems

Eliminate user client interaction and involvement
have time and effort

Hybrid approach