MTA-KDD'19: A Dataset for Malware Traffic Detection

ITASEC20
ITALIAN CONFERENCE ON CYBERSECURITY
Ancona, 4-7 February 2020
cini Cybersecurity National Lab

University of L'Aquila

Authors:

Ivan Letteri

Giuseppe Della Penna

Luca Di Vita

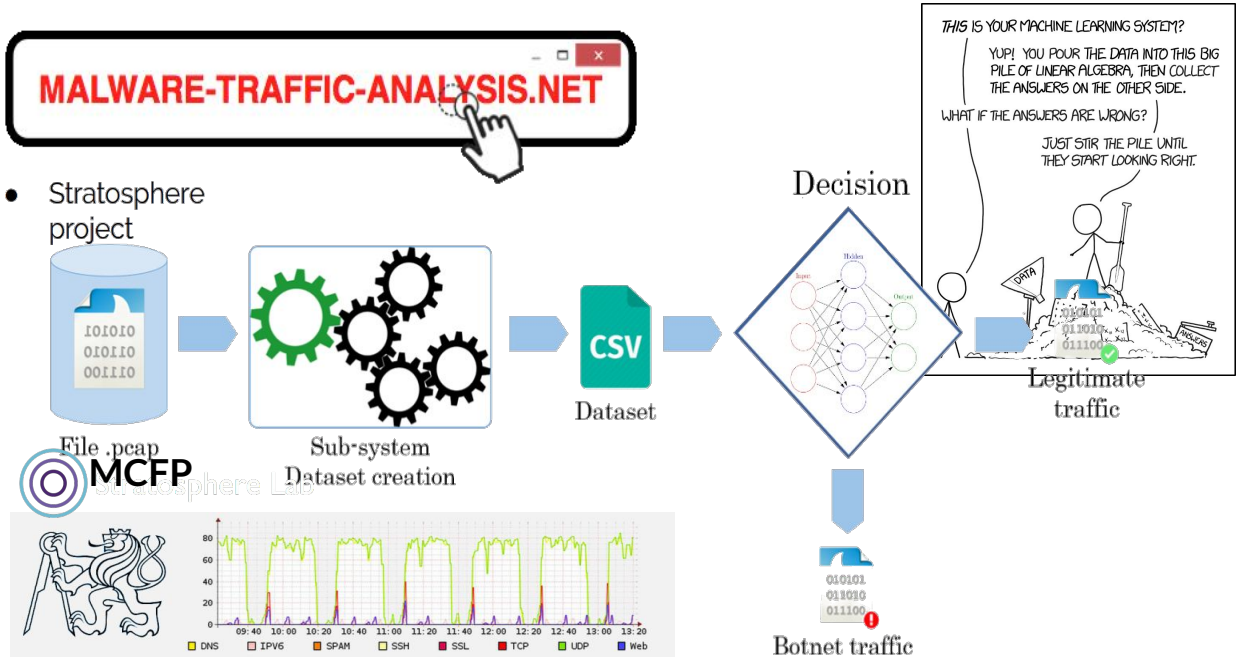
Maria Teresa Grifa

Road Map

- **1st - MTA-KDD'19 Dataset**
  - legitimate traffic
  - malware traffic
- **2nd - Dataset Features**
  - from 50 to 33 features
- **3rd - Dataset Generation Process**
  - *phase I.* Feature Extraction
  - *phase II.* Dataset Profiling
  - *phase III.* Imputation
  - *phase IV.* Scaling
- **4th - Dataset Evaluation**
  - Outlier Detection
  - Classification Experiment

Generation Process

- **Network logs**
  - Feature Extraction *phase* from pcap files
    - Legitimate *traffic*
    - Malware *traffic*



- Legitimate traffic come from pcap files marked as Normal in MCFP composed of 15 pcap files with a total size greater than 7 GBytes which produced ≃ 45 thousand samples
- Malware traffic come from the MTA repository, is made up of ≃ 2112 pcap files, with a total size of than 4.8 GB. These observations cover a time span from June 2013 to August 2019.

Generation Process (50 *Dataset Features*)

1. {Ack,Syn,Fin,Psh,Urg,Rst}*FlagDist*
2. {TCP,UDP,DNS}*OverIP*
3. MaxLen, MinLen, AvgLen, StdDevLen, MaxIAT, MinIAT, ...
4. PktIORatio
5. FirstPktLen:
6. DNSQDist,DNSADist, DNSRDist, ...
7. {RepeatedPkt, SmallPkt} *Ratio*
8. AvgDomain{Char,Dot,Hyph,Digit}
9. AvgTTL
10. NumConnections, SynAcksyn
11. NumDstAddr, NumPorts
12. DistinctUA, AvgDistinctUALen, ...

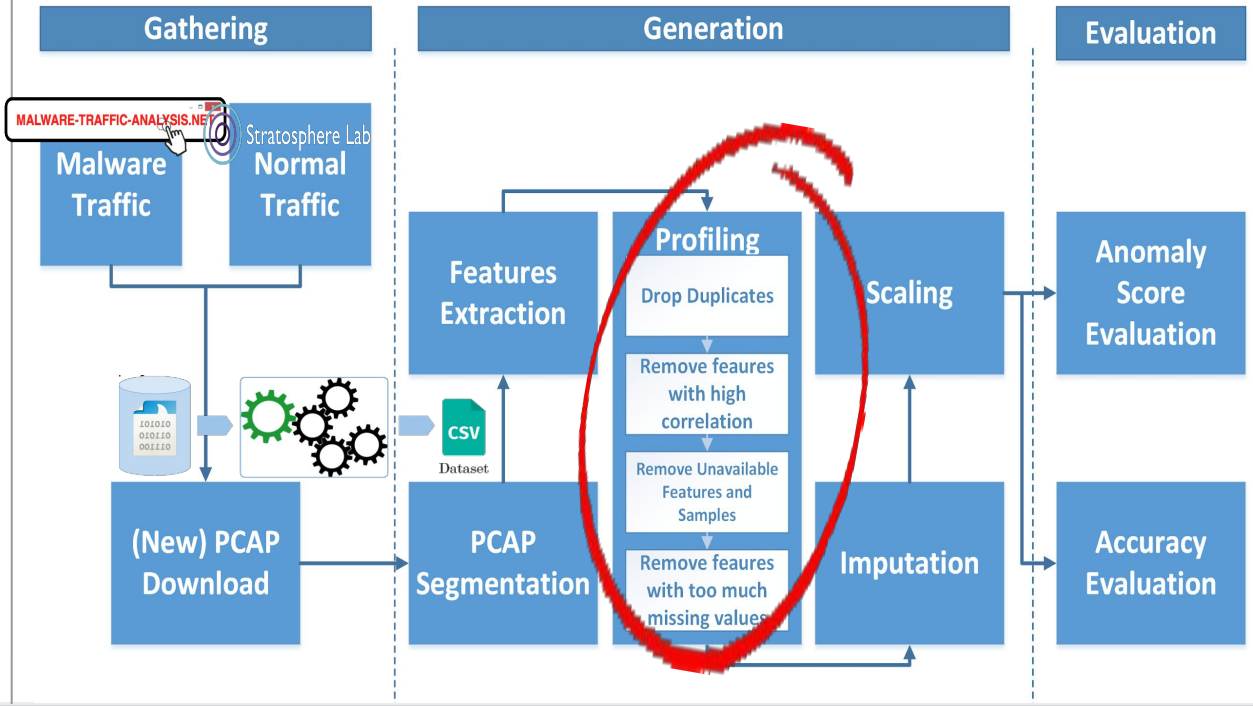
$\mathcal{S} = \{p_0, \ldots, p_n\}$	the *sent* packets
$\mathcal{S}^{DNS}$	sent packets using the DNS protocol
$\mathcal{S}^{TCP}$	sent packets using the TCP protocol
$\mathcal{S}^{UDP}$	sent packets using the UDP protocol
$\mathcal{S}^{H}$	sent packets using the HTTP protocol
$\mathcal{S}^{HR}$	sent packets containing an HTTP request
$\mathcal{S}^{DQ}$	sent packets containing a DNS question record
$\mathcal{S}^{DR}$	sent packets containing a DNS resource record
$\mathcal{S}^{ACK}$	sent TCP packets with the ACK flag set
$\mathcal{S}^{SYN}$	sent TCP packets with the SYN flag set
$\mathcal{S}^{FIN}$	sent TCP packets with the FIN flag set
$\mathcal{S}^{PSH}$	sent TCP packets with the PSH flag set
$\mathcal{S}^{URG}$	sent TCP packets with the URG flag set
$\mathcal{S}^{RST}$	sent TCP packets with the RST flag set
$\mathcal{S}^{ACKSYN}$	sent TCP packets ith both the ACK and SYN flags set
$\mathcal{S}^{small}$	sent packets with payload length < 32 (*small packets*)
$\mathcal{D}_p$	distinct packet destination ports
$\mathcal{D}_a$	distinct packet destination addresses
$\mathcal{U}$	distinct user agents in the HTTP sent packets
$\mathcal{R}$	the *received* packets
$dom(p)$	number of domains referred in DNS packet p
$len(p)$	payload length of packet p
$nchar(s)$	number of characters in s
$ndot(s)$	number of dots in s
$nhyph(s)$	number of hyphens in s
$ndigit(s)$	number of digits in s
$ndnsque(p)$	number of items in question section of DNS packet p
$ndnsans(p)$	number of items in answer section of DNS packet p
$ndnsadd(p)$	number of items in additional section of DNS packet p
$ndnsaut(p)$	number of items in authority section of DNS packet p
$occur(s)$	number occurrences of each distinct value in the set s
$t(p)$	arrival time of packet p
$ttl(p)$	TTL reported in DNS packet p
$valid(s)$	true if all the domain names th set s are valid

feature	formula	notes	rel	rem
fFlagDist	$\frac{\lvert\mathcal{S}^{TCP}\rvert}{\lvert\mathcal{S}\rvert} \cdot \lvert\mathcal{S}^{f}\rvert$	with $f \in \{Ack, Syn, Fin, Psh, Rst\}$. *unavailable* if not TCP	(1)	-
pOverIP	$\frac{\lvert\mathcal{S}^{p}\rvert}{\lvert\mathcal{S}\rvert}$	with $p \in \{TCP, UDP, DNS\}$	(2)	-
AvgDeltaTime	$\frac{t(p_n)-t(p_0)}{\lvert\mathcal{S}\rvert}$	zero if $\lvert\mathcal{S}\rvert = 1$	(3)	(1)
AvgDistinctUALen	$\frac{\sum_{u \in \mathcal{U}} nchar(u)}{\lvert\mathcal{U}\rvert}$	NaN if $\lvert\mathcal{S}^{HR}\rvert = 0$	(12)	(3)
AvgDomainChar	$\frac{\sum_{p_j \in \mathcal{S}^{DQ}} nchar(dom(p_j))}{\sum_{p_j \in \mathcal{S}^{DQ}} \lvert dom(p_j)\rvert}$	NaN if $\lvert\mathcal{S}^{DQ}\rvert = 0$	(8)	(3)
AvgDomainDigit	$\frac{\sum_{p_j \in \mathcal{S}^{DQ}} ndigit(dom(p_j))}{\sum_{p_j \in \mathcal{S}^{DQ}} \lvert dom(p_j)\rvert}$	NaN if $\lvert\mathcal{S}^{DQ}\rvert = 0$	(8)	(3)
AvgDomainDot	$\frac{\sum_{p_j \in \mathcal{S}^{DQ}} ndot(dom(p_j))}{\sum_{p_j \in \mathcal{S}^{DQ}} \lvert dom(p_j)\rvert}$	NaN if $\lvert\mathcal{S}^{DQ}\rvert = 0$	(8)	(3)
AvgDomainHyph	$\frac{\sum_{p_j \in \mathcal{S}^{DQ}} nhyph(dom(p_j))}{\sum_{p_j \in \mathcal{S}^{DQ}} \lvert dom(p_j)\rvert}$	NaN if $\lvert\mathcal{S}^{DQ}\rvert = 0$	(8)	(3)
AvgTTL	$\frac{\sum_{p_j \in \mathcal{S}^{DR}} ttl(p_j)}{\lvert\mathcal{S}^{DR}\rvert}$	-	(9)	(3)
DeltaTime	$t(p_n) - t(p_0)$	zero if $\lvert\mathcal{S}\rvert = 1$	(3)	-
DistinctUA	$\lvert\mathcal{U}\rvert$	-	(12)	(3)
DNSADist	$\frac{\lvert\mathcal{S}^{DNS}\rvert}{\lvert\mathcal{S}\rvert} \cdot \sum_{p_j \in \mathcal{S}^{D}} ndnsans(p_j)$	-	(6)	(1)
DNSQDist	$\frac{\lvert\mathcal{S}^{DNS}\rvert}{\lvert\mathcal{S}\rvert} \cdot \sum_{p_j \in \mathcal{S}^{D}} ndnsque(p_j)$	-	(6)	-
DNSRDist	$\frac{\lvert\mathcal{S}^{DNS}\rvert}{\lvert\mathcal{S}\rvert} \cdot \sum_{p_j \in \mathcal{S}^{D}} ndnsadd(p_j)$	-	(6)	(1)
DNSSDist	$\frac{\lvert\mathcal{S}^{DNS}\rvert}{\lvert\mathcal{S}\rvert} \cdot \sum_{p_j \in \mathcal{S}^{D}} ndnsaut(p_j)$	-	(6)	(1)
EndFlow	$t(p_n)$	-	(3)	(1)
FirstPktLen	$len(p_0)$	-	(5)	-
FlowLen	$\sum_{p_j \in \mathcal{S}} len(p_j)$	-	(3)	-
FlowLenRx	$\sum_{p_j \in \mathcal{R}} len(p_j)$	-	(3)	-
HTTPpkts	$\lvert\mathcal{S}^{H}\rvert$	-	(12)	-
MaxIATRx,	min of $\{(t(p_j) - t(p_{j-1}))\lvert p_j \in \mathcal{R}\}$	zero if $\lvert\mathcal{R}\rvert < 2$	(3)	(1)
MinIAT, MaxIAT, AvgIAT	min, max, avg of $\{(t(p_j) - t(p_{j-1}))\lvert p_j \in \mathcal{S}\}$	zero if $\lvert\mathcal{S}\rvert < 2$	(3)	-
MinIATRx, AvgIATRx	min, avg of $\{(t(p_j) - t(p_{j-1}))\lvert p_j \in \mathcal{R}\}$	zero if $\lvert\mathcal{R}\rvert < 2$	(3)	-
MinLen, MaxLen, AvgLen, StdDevLen	min, max, avg, stddev of $\{len(p_j)\lvert p_j \in \mathcal{S}\}$	StdDevLen is NaN if $\lvert\mathcal{S}\rvert < 2$	(3)	-
MinLenRx, MaxLenRx, AvgLenRx, StdDevLenRx	min, max, avg, stddev of $\{len(p_j)\lvert p_j \in \mathcal{R}\}$	StdDevLenRx is NaN if $\lvert\mathcal{R}\rvert < 2$	(3)	-
NumConnections	$\lvert\mathcal{S}^{ACKSYN}\rvert$	*unavailable* if not TCP	(10)	-
NumDstAddr	$\lvert\mathcal{D}_a\rvert$	-	(11)	-
NumPorts	$max\{\lvert\mathcal{D}_p\rvert\}$	*unavailable* not TCP or UDP	(11)	-
PktIOratio	$\frac{\lvert\mathcal{R}\rvert}{\lvert\mathcal{S}\rvert}$	-	(4)	-
RepeatedPktLenRatio	$\frac{\max occur((len(p_j)\lvert p_j \in \mathcal{S}))}{\lvert\mathcal{S}\rvert}$	-	(7)	-
SmallPktRatio	$\frac{\lvert\mathcal{S}^{small}\rvert}{\lvert\mathcal{S}\rvert}$	-	(7)	-
StartFlow	$t(p_0)$	-	(3)	-
SynAcksynRatio	$\frac{\lvert\mathcal{S}^{S}\rvert}{\lvert\mathcal{S}^{ACKSYN}\rvert}$	*unavailable* if not TCP, NaN if $\lvert\mathcal{S}^{ACKSYN}\rvert = 0$	(10)	(3)

Value	Count	Frequency (%)
Malware	39544	55.3%
Legitimate	31926	44.7%

Generation Process

- **Gathering**
  - where came from Dataset
    - MTA
    - MCFP
  - 50 features extracted
    - dataset in CSV file format
  - Next: *profiling activity*

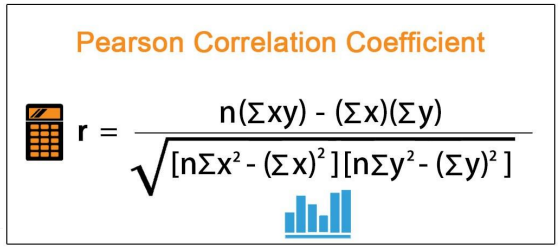


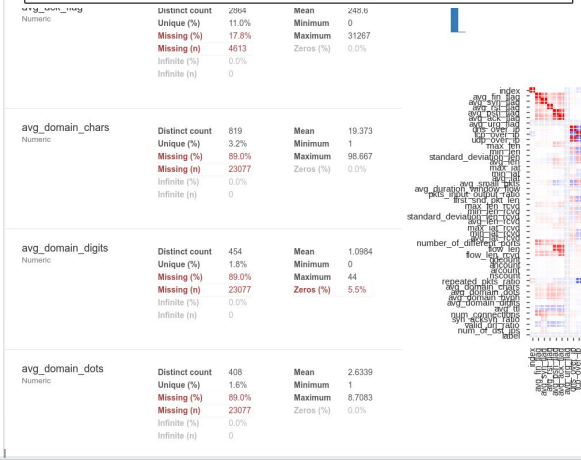
Dataset Profiling

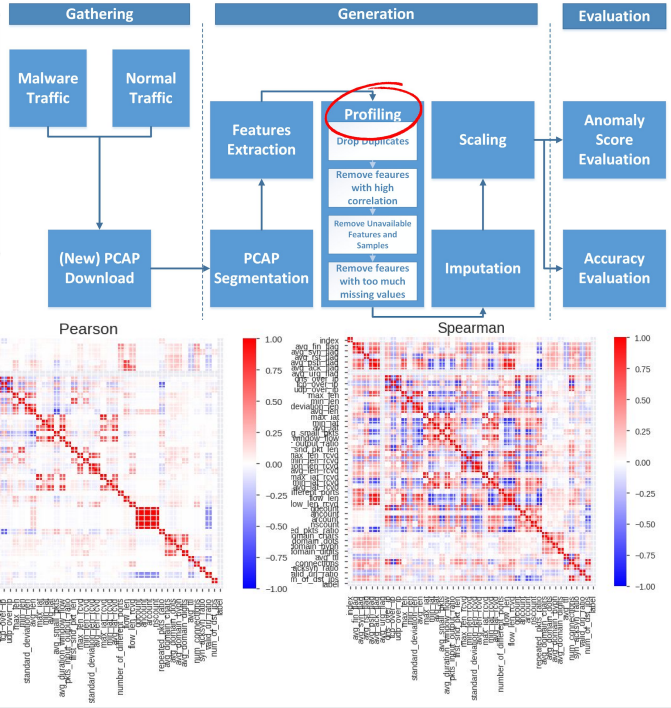
- drop duplicates, missing values ...
- remove feature with high correlation, and so on ...
- … idea to apply a minimal preprocessing on the data

Generation Process (*Profiling*)

- **Issues**
  - *detect* low-complexity
  - *remove* low-complexity
- **High-correlation**
  - Pearson coefficient
  - High correlation > 0.95
- **Not a Number (NaN)**
  - '*Real*' Zeros
  - Unavailable (*incomplete*)
  - Missing Values (*noisy*)







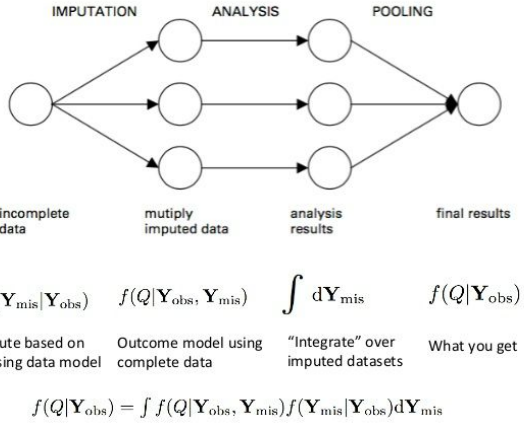
- To detect and remove certain low-complexity issues that may affect the data
- High-correlated features measured using the Pearson coefficient (from 50 features to 33 features)
- some features may have zero values for different reasons: *Zeros*, *Unavailable*, and *Missing* values

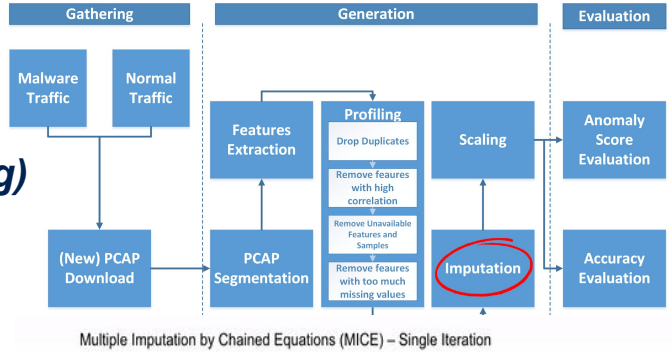
Generation Process (*Imputation*)

- **remaining NaN**
  - StdDevLen *feature*
  - StdDevLenRx *feature*

- **replacing NaN**
  - with real(istic) values
  - MICE ✔
  - Datawig, KNN ✖

Imputing (*MICE*, *KNN,DataWig)*



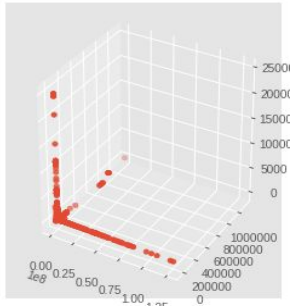


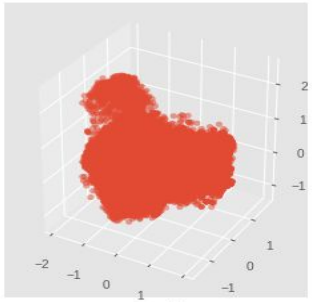
- MICE imputation works by filling the missing data multiple times
- Multiple imputations are better than a single imputation as they measure the uncertainty of the missing values more precisely
- chained equations approach is flexible and handle data types (i.e., continuous or binary)

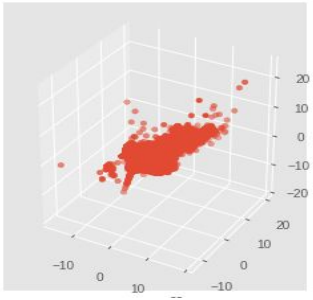
Generation Process (*Scaling*)

- **4 Scalers and 2 Transformers**
  - MinMax *scaler*
  - MaxAbs *scaler*
  - Standard *scaler*
  - Robust *scaler*
  - Quantile *transformer*
  - Power *transformer*

- **Precision metric**
  - Gaussian Naive Bayes (GNB)
  - Logistic Regression (LogReg)
  - Random Forest (RF)

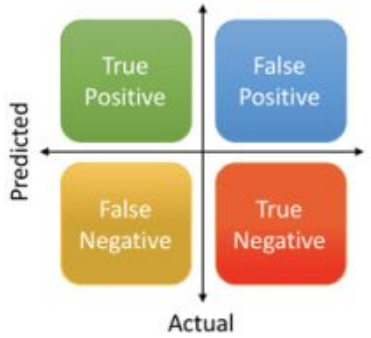






$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

algorithm	GNB	LogReg	RF
standardized scaler	75.96%	90.64%	94.20%
minmax scaler	75.96%	85.93%	**94.41%**
maxabsolute scaler	75.96%	85.91%	**94.41%**
robust scaler	69.36%	48.21%	94.20%
quantile transformer	78.07%	91.97%	94.20%
power transformer	**80.69%**	**97.31%**	94.20%



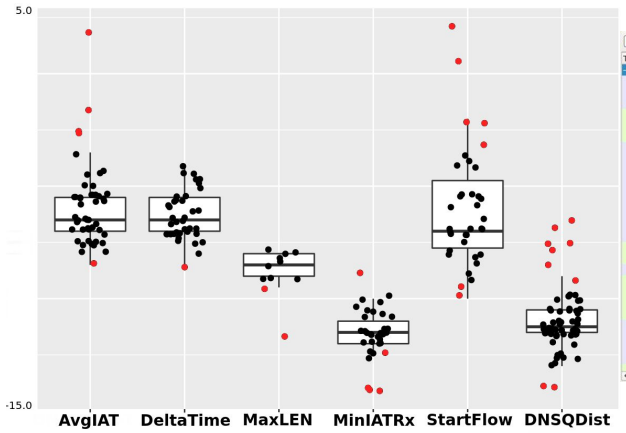
- Dataset contains features highly varying in magnitudes, units, and range
- feature scaling may heavily influence the results of some algorithms
- to have an idea of the impact of such scaling we trained 3 “fast” classifiers (GNB, LogReg, RF)

Dataset Evaluation

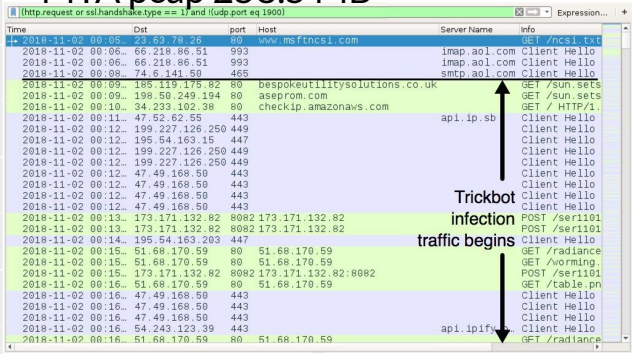
- **Outlier Detection**
  - feature importance with Random Forest (*Gini index*)
  - 6 most important features
  - Matching between:

  Outlier in BoxPlot <-> pkt in PCAP

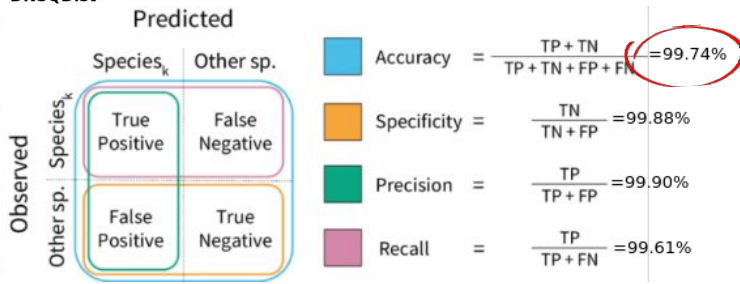
- **Classification Experiment**
  - Multi Layer Perceptron
  - Accuracy, Precision, Recall, Specificity



MTA pcap 258.5 MB



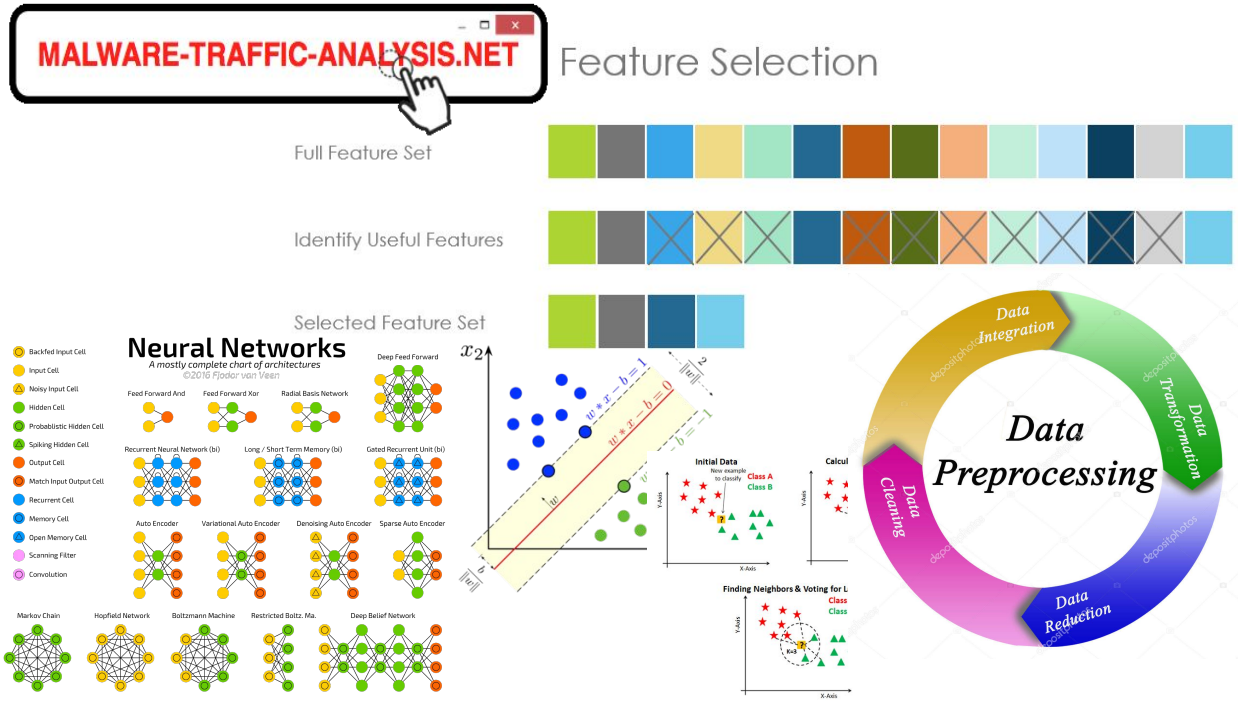
	Predicted: NO	Predicted: YES
Actual: NO	10243	40
Actual: YES	10	9074



- Further evaluation of quality looking for the matching between *Outliers* in the CSV dataset and the traffic in *PCAP* files on a subset of 6 “important” features
- Figure to the bottom shows the confusion matrix taken from one of the experiments, with an accuracy of 99.74%

Conclusion

- **MTA-KDD'19**
  - malware traffic analysis
  - Data source up-to-date
  - Open source
- **More Accurate**
  - Feature Selection
  - Data preprocessing
- **Future Work**
  - more complex features
  - different NN architectures
  - other ML models
  - handle Unavailable values



- The data sources are continuously updated, making the dataset quite realistic
- **Accurate feature selection** and **data preprocessing** to make the dataset as small as possible
- **More complex feature** selection strategies to further reduce the number of features to the most informative ones.
- Validate the dataset **different neural network architectures** and **other machine learning models**.

- **Reduced Dataset**
  - 33 features

github.com/IvanLetteri/MTA-KDD-19

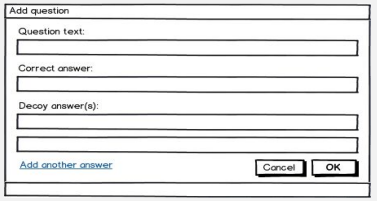
Scan me

- **Source Code**
  - PCAP downloader
  - Feature Extractor

- **Entire Dataset**
  - Dataset 4 pkts tracking

- **Form Questions**
  - repeat experiment *(I. Letteri, G.D.P.)*
  - maths questions *(M.T. Grifa)*
  - data mining code *(L. Di Vita)*

ivanletteri.it/mta-kdd19-qf/



- **Call For Collaboration**
  - Registration form
  - List of projects

ivanletteri.it/call-for-collaboration/

SCAN ME

HERE these slides: https://www.ivanletteri.it/itasec2020/

SCAN ME