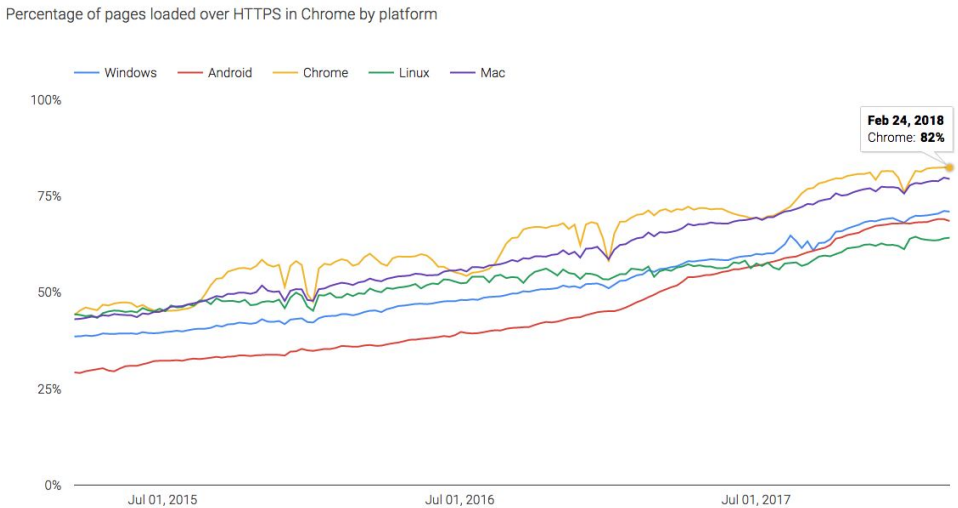
Detecting malware even when it is encrypted

Machine Learning for network HTTPS analysis

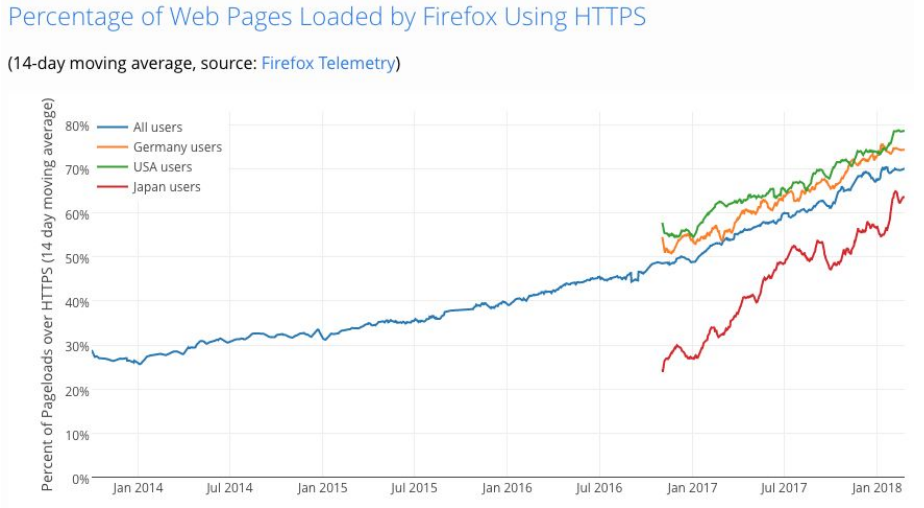
František Střasák
strasfra@fel.cvut.cz
@FrenkyStrasak

Sebastian Garcia
sebastian.garcia@agents.fel.cvut.cz
@eldracote

More than 80% of web traffic is encrypted



https://transparencyreport.google.com/https/overview?hl=en



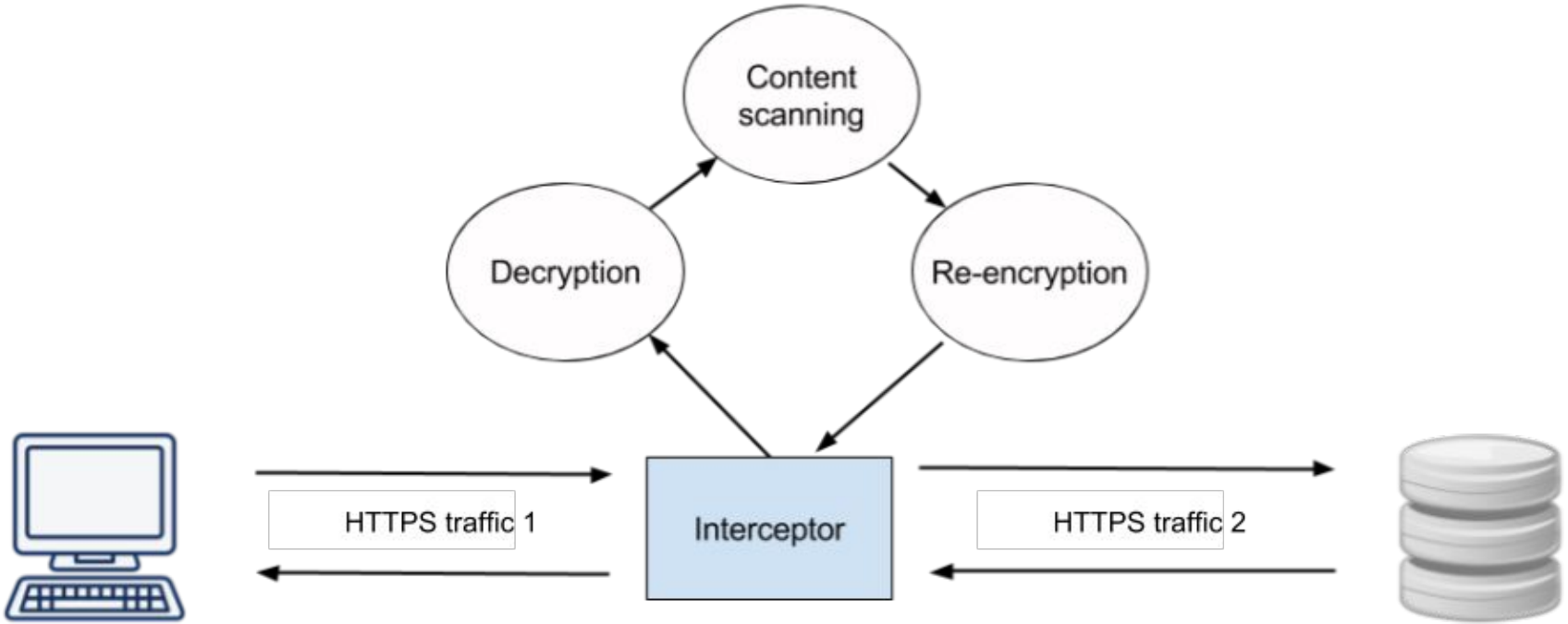
https://letsencrypt.org/stats/

From 10% to 40% of all malware traffic is encrypted

- 10-12% of all Malware uses HTTPS
  - https://blogs.cisco.com/security/malwares-use-of-tls-and-encryption (Jan 2016)
- 37% of all Malware uses HTTPS
  - https://blog.cyren.com/articles/over-one-third-of-malware-uses-https (June 2017)
- From all HTTPS malware, 97% uses port 443, and 87% uses TLS
  - Stratosphere Nomad Project. Jan. 2018

Encryption interferes with the efficacy of classical detection techniques

Do we need TLS inspection?



TLS inspection

- Advantages
  - TLS inspection can use classical detection techniques
- Disadvantages
  - TLS inspection may be expensive
  - TLS inspection is computationally demanding (can be slow)
  - TLS inspection does not respect the original idea of HTTPS (privacy)
  - Using ‘local’ certificates teaches users to ignore security.

Our Goal

To find features and methods to analyze HTTPS traffic **without** decryption and detect malware with high accuracy, low false positive rate.

What is SSL/TLS?: handshake

Client Hello

Server Hello with certificate and decision about the parameters.

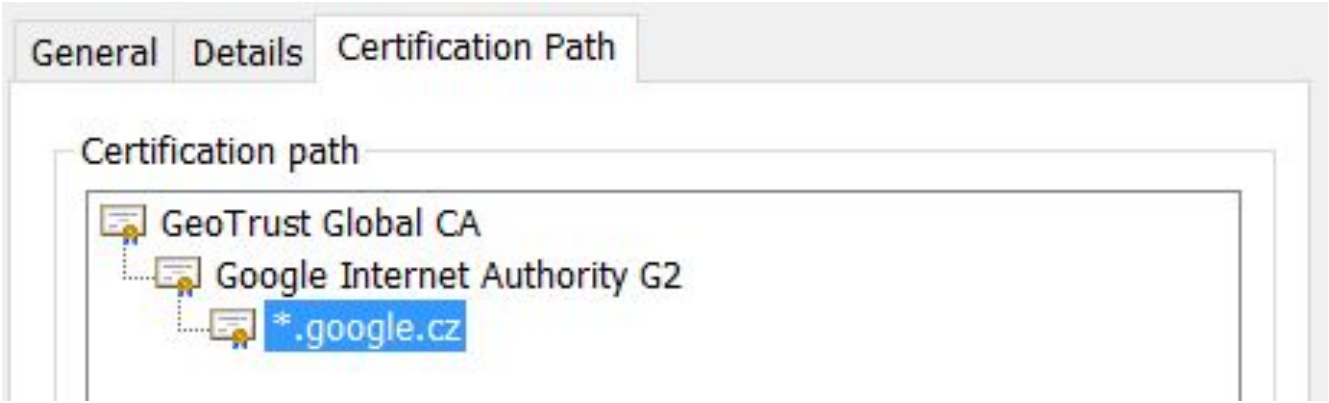
If the certificate is trusted, creates a symmetric session key and encrypts it with the server's asymmetric public key.

Server decrypts the encrypted session key using its asymmetric private key to get the symmetric session key.

Server and Browser now encrypt and decrypt all transmitted data with the symmetric session key.

What is SSL/TLS?: Certification path

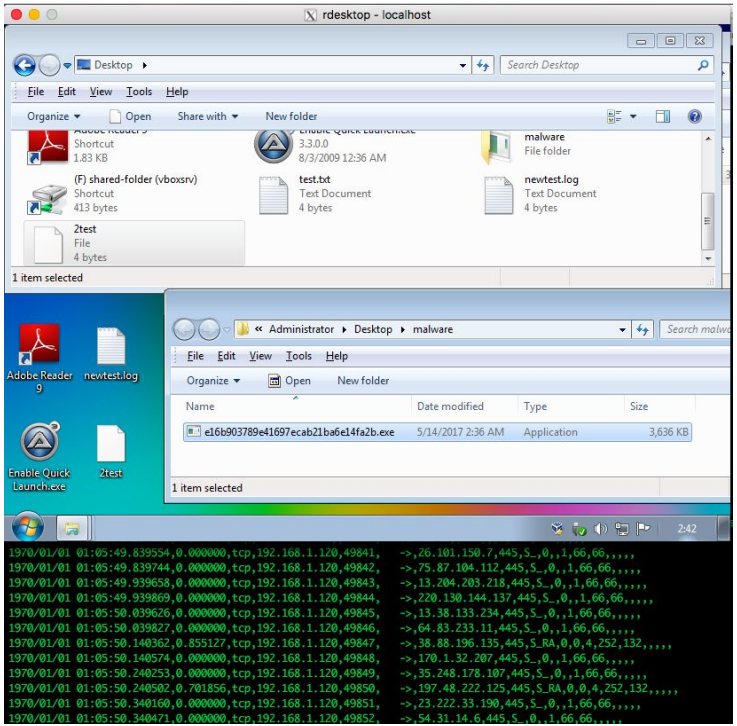
- A root CA
- An intermediate CA



Privacy does not mean Security!

Dataset

- Pcaps/flows with HTTPS traffic
- Malware and Normal
- 4 sub-datasets
- 163 malware and normal captures

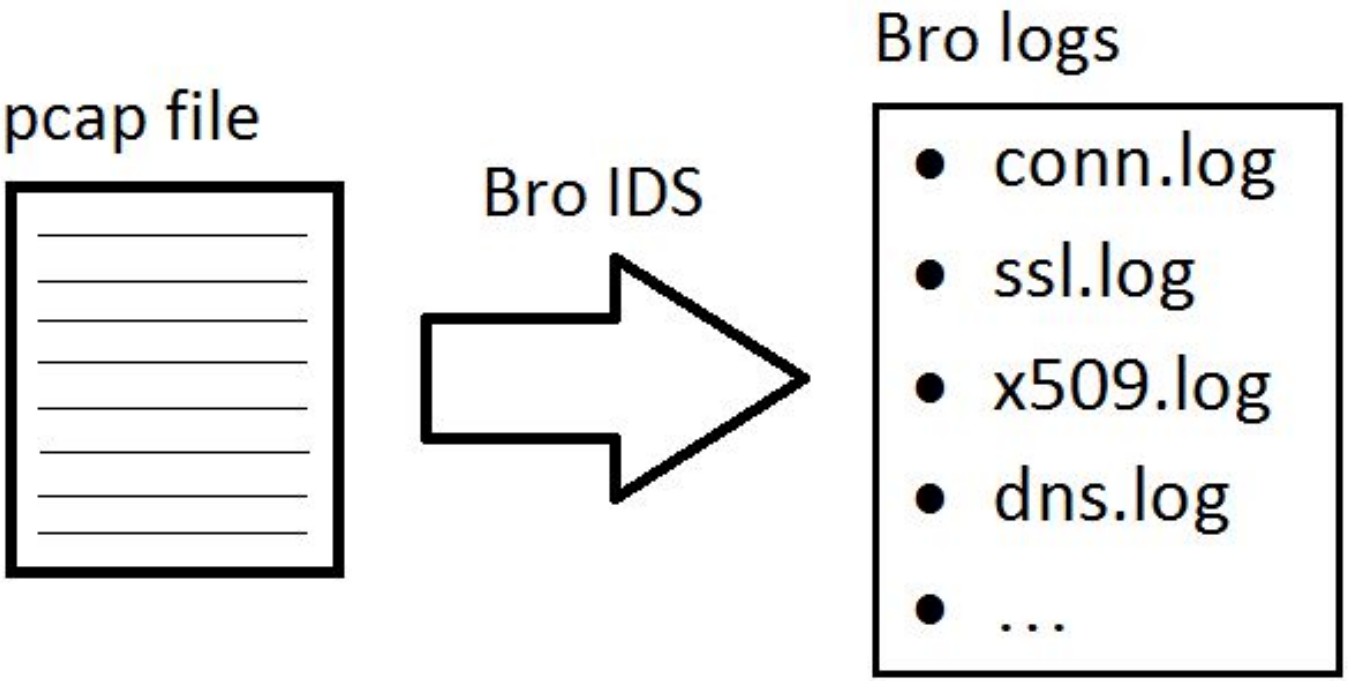


Dataset

- CTU-13 dataset - public
  - Malware and Normal captures
  - 13 Scenarios. 600GB pcap
  - https://www.stratosphereips.org/datasets-ctu13/
- MCFP dataset - public
  - Malware Capture Facility Project. (Maria Jose Erquiaga)
  - 340 malware pcap captures
  - https://stratosphereips.org/category/dataset.html
- Own normal dataset - public
  - 3 days of accessing to secure sites (Alexa 1000)
  - Google, Facebook, Twitter accounts
  - https://stratosphereips.org/category/dataset.html
- Normal CTU dataset - almost public
  - Normal captures
  - 22 known and trusted people from department of FEE CTU

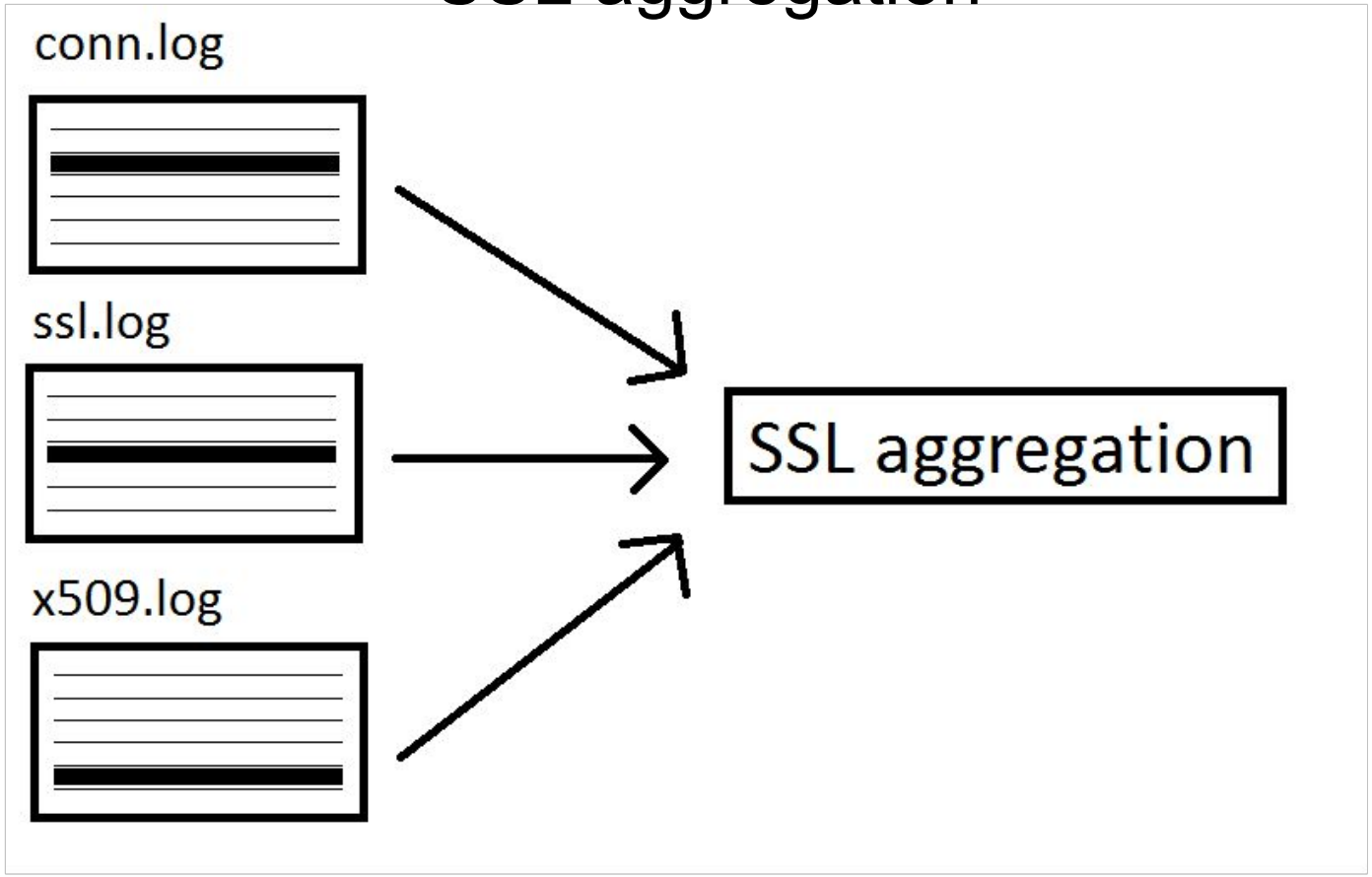
Features and Methods

Bro logs

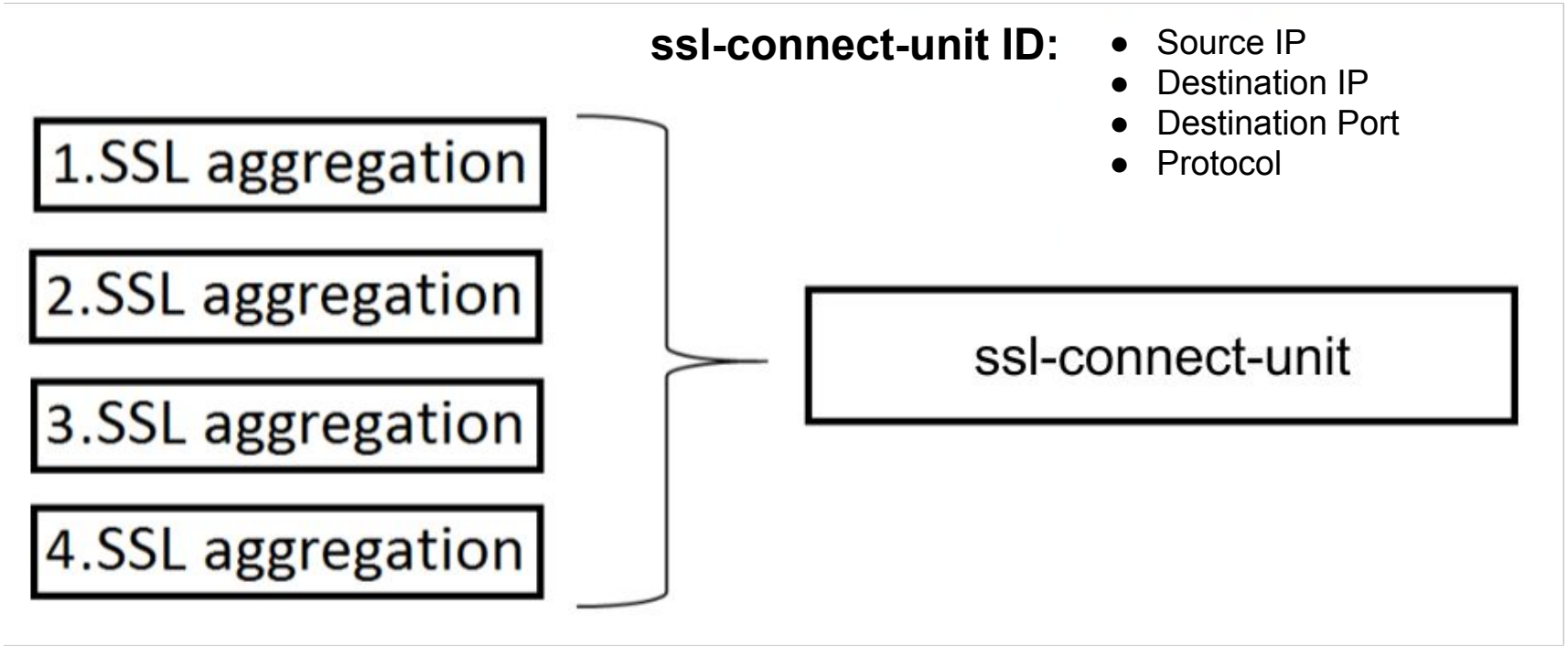


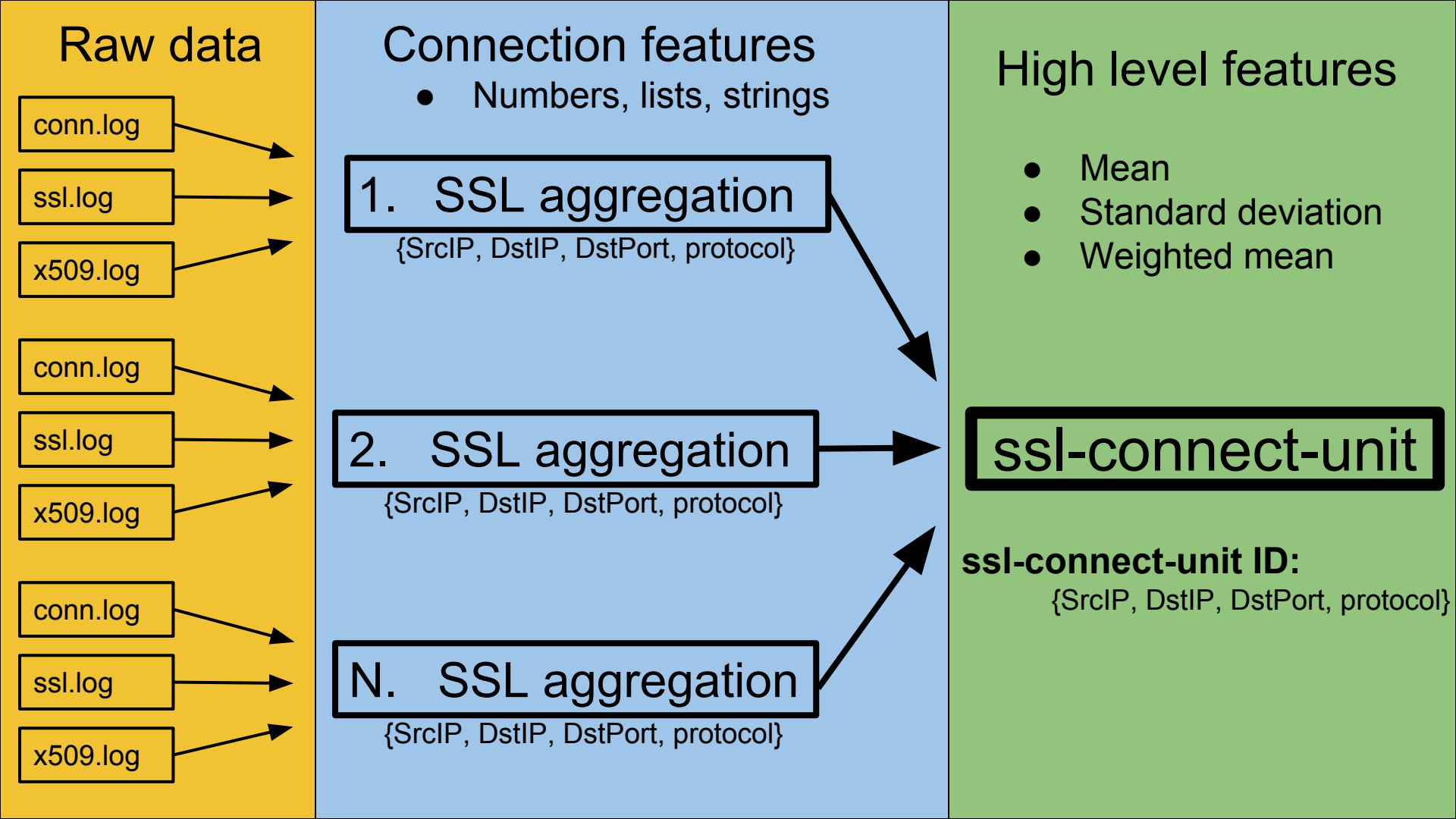
https://www.bro.org/

SSL aggregation



ssl-connect-unit

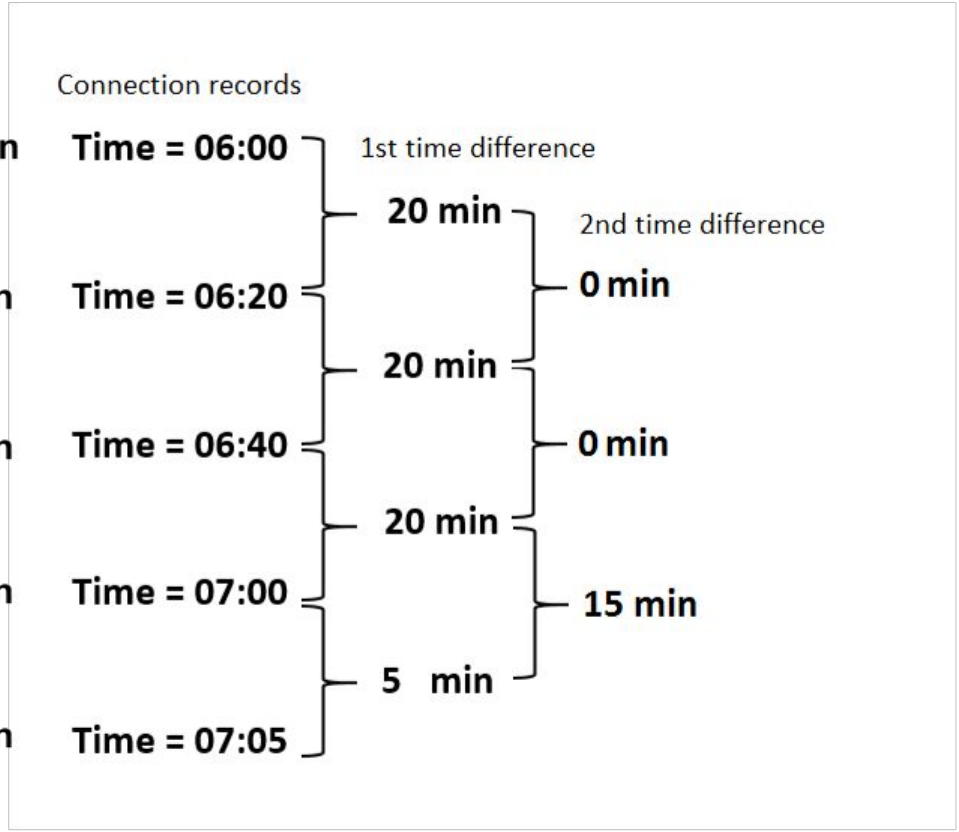



Raw data
conn.log
ssl.log
x509.log
conn.log
ssl.log
x509.log
conn.log
ssl.log
x509.log
Connection features
Numbers, lists, strings
1. SSL aggregation
{SrcIP, DstIP, DstPort, protocol}
2. SSL aggregation
{SrcIP, DstIP, DstPort, protocol}
N. SSL aggregation
{SrcIP, DstIP, DstPort, protocol}
High level features
Mean
Standard deviation
Weighted mean
ssl-connect-unit
ssl-connect-unit ID:
{SrcIP, DstIP, DstPort, protocol}

40 Features of ssl-connect-unit. Examples:

- Number of SSL aggregations
- Mean and standard deviation of duration
- Mean and standard deviation of number of packets
- Mean and standard deviation of number of bytes
- Ratio of TLS and SSL version
- Number of different certificates

Example Feature: Mean of 2nd level time difference



Example Feature: Mean of certificate validity during capture

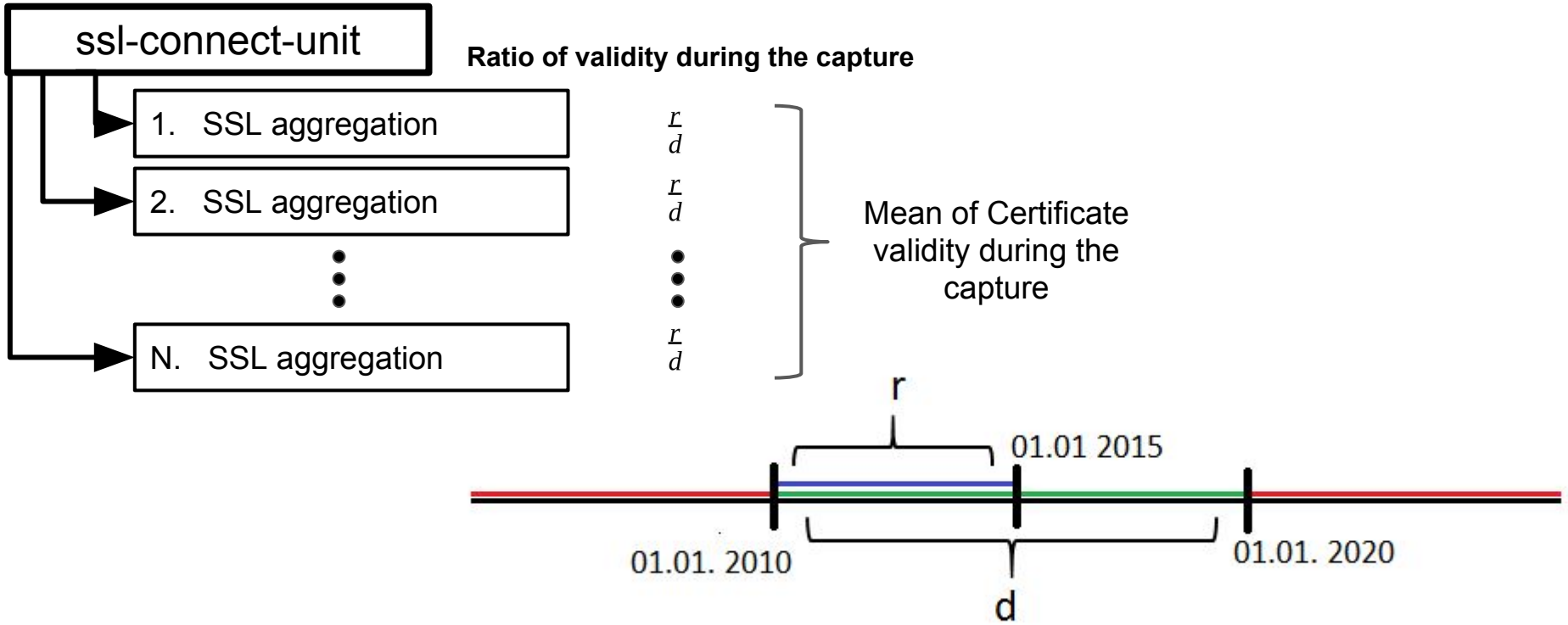


Table with final data to use in our Algorithms

ssl-connect-unit	40 features					Label
{ 10.0.2.15, 54.201.174.90, 443, tcp }	f1	f2	f3	…	f40	Normal
{ 10.0.2.109, 173.194.122.30, 443, tcp }	f1	f2	f3	…	f40	Malware
⋮						

Machine learning algorithms

- XGBoost
  - Extreme Gradient Boosting
  - Tree booster with logistic regression
- Random Forest
  - Random Forest Classifier model that is an estimator that fits a number of decision tree classifiers on various sub-samples
- Neural Network
  - MLP Classifier (Multi-layer Perceptron classifier)
- SVM
  - Radial Basis Function (RBF) kernel

Experiments

- XGBoost
  - Cross validation accuracy: 92.45%
  - **Testing accuracy: 94.33%**
  - False Positive Rate: 5.54%
  - False negative rate: 10.11%
  - Sensitivity: 89.89%
  - F1 Score: 46.96 %
- Random Forest
  - Cross validation accuracy: 91.21%
  - **Testing accuracy: 95.65%**
  - False Positive Rate: 4.05%
  - False negative rate: 14.82%
  - Sensitivity: 85.18%
  - F1 Score: 52.24%

Top 7 most discriminant features

1. Certificate length of validity
2. Inbound and outbound packets
3. Validity of certificate during the capture
4. Duration
5. Number of domains in certificate (SAN DNS)
6. SSL/TLS version
7. Periodicity

Malware and Certificates

- Certificates used by Malware in Alexa 1000 ~ 50%
- Certificates used by Normal in Alexa 1000 ~ 30%

The certificates used by Malware are mostly from normal sites!

Conclusions

- Future Work
  - More features (dns logs)
  - Own architecture of neural network
  - Unsupervised learning
  - Anomaly detection

Thanks for attention!

František Střasák
strasfra@fel.cvut.cz
@FrenkyStrasak

Sebastian Garcia
sebastian.garcia@agents.fel.cvut.cz
@eldracote