APPID – USING MACHINE LEARNING

Application Identification

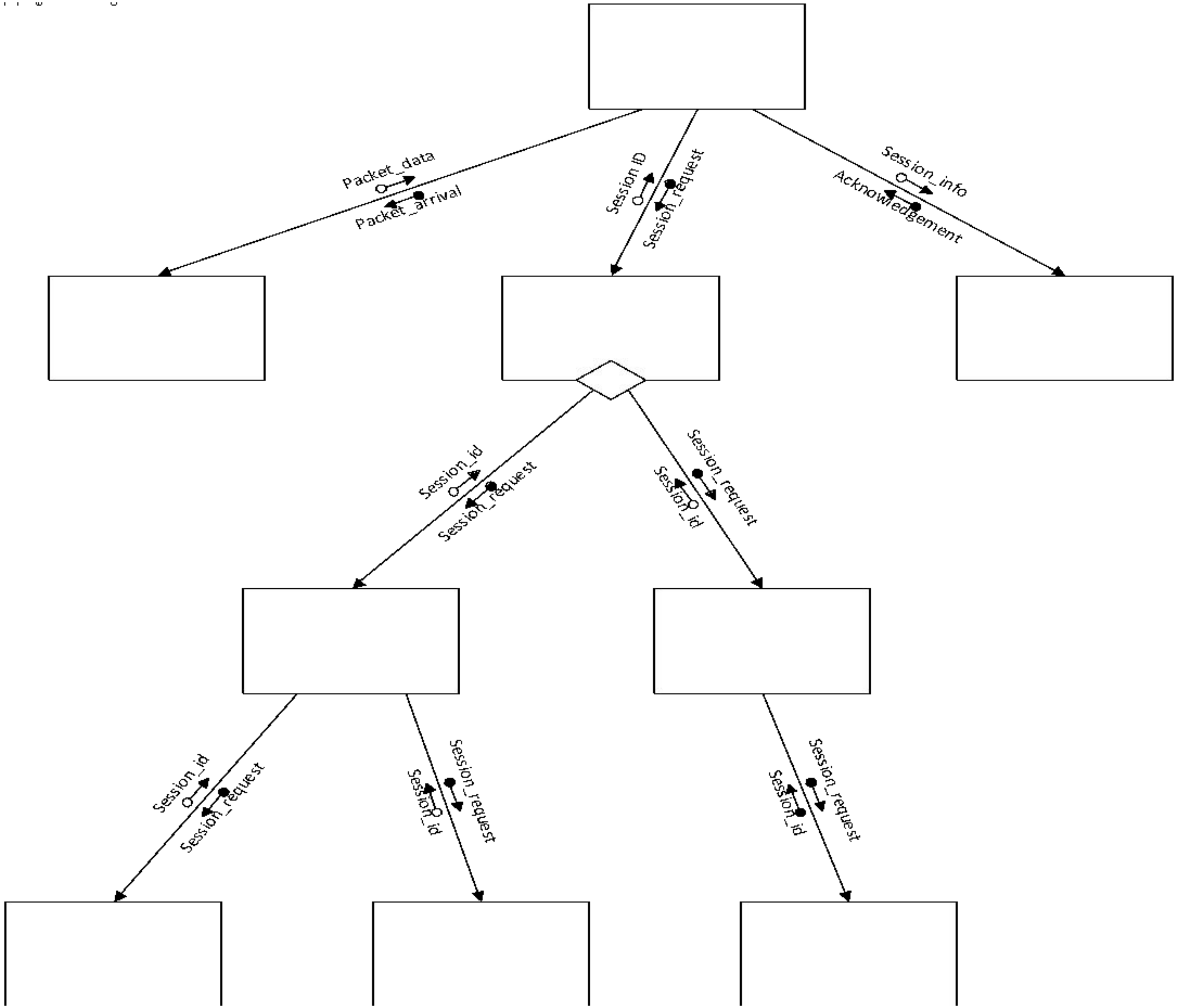
- Identifying application using Decision trees.
- Core Concept
  - Identify statistical info of current application.
  - Classify it using decision tree.
  - If decision tree has found the match , application is identified.
  - If decision tree fails to classify the data, train the decision tree using it.

Decision Trees

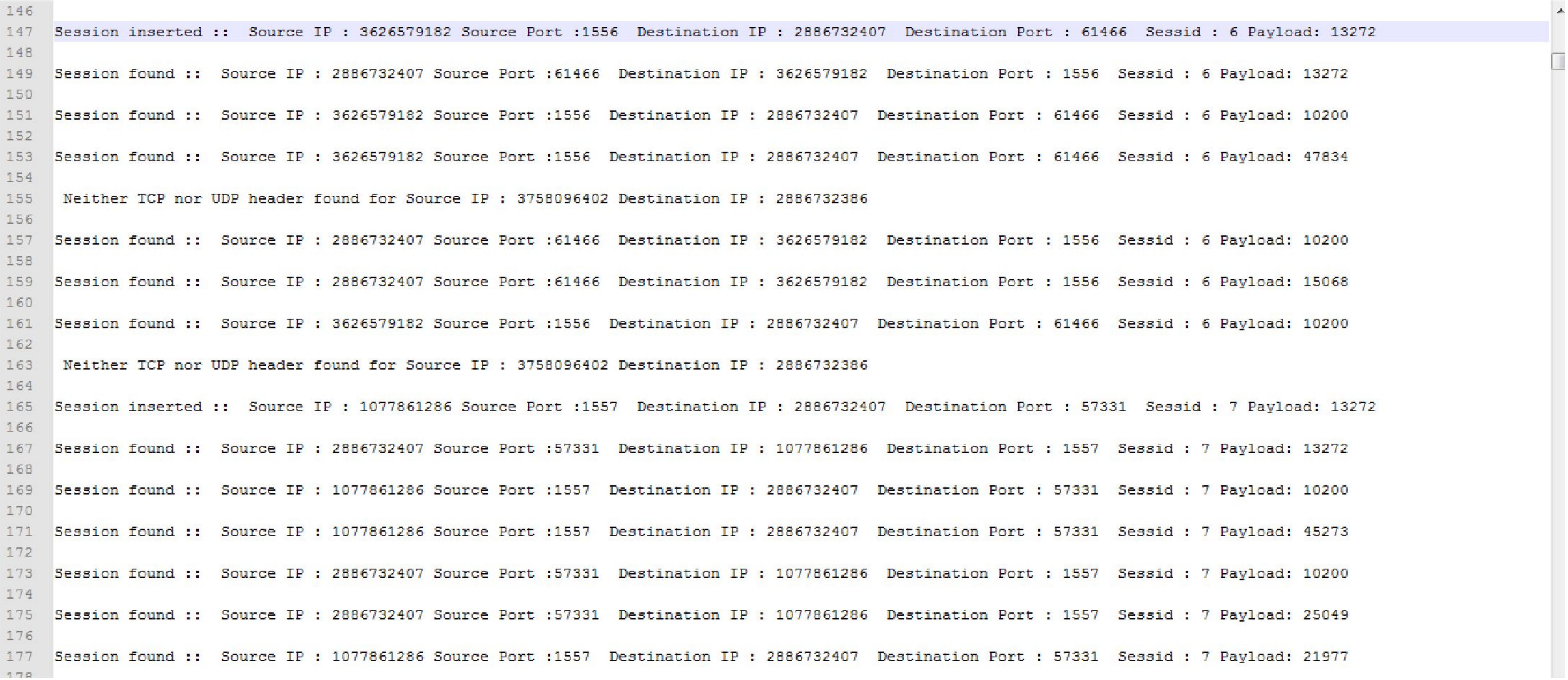
- Available Open source Decision Trees :
  - C4.5 : An algorithm used to generate a decision tree developed by Ross Quinlan
  - VFDT : Very Fast Decision Tree. A wrapper of c4.5 to work with high speed data streams.
- VFDT is preferred over c4.5 as instead of giving input to decision tree in terms of file, directly working with data streams is easy with respect to snort.

Identifying sessions

- Session identification is required to collect enough packets of same application when creating dataset for training the decision tree.
- Sessions are stored in a hash table structure with keys being socket address pairs. Efficient searching is ensured through murmur hash.
- AVL tree is used for managing unique session Ids.
- After there are enough packets in a session, its cumulative information is saved in a file which is used to train a decision tree afterwards.
- Structure chart for session identification is as below:


Packet_data
Packet_arrival
Session ID
Session_request
Session_info
Acknowledgement
Session_id
Session_request
Session_request
Session_id
Session_id
Session_request
Session_id
Session_request
Session_id
Session_request

Sessions Identified :



```

Session inserted ::  Source IP : 3626579182 Source Port :1556  Destination IP : 2886732407  Destination Port : 61466  Sessid : 6 Payload: 13272

Session found ::  Source IP : 2886732407 Source Port :61466  Destination IP : 3626579182  Destination Port : 1556  Sessid : 6 Payload: 13272

Session found ::  Source IP : 3626579182 Source Port :1556  Destination IP : 2886732407  Destination Port : 61466  Sessid : 6 Payload: 10200

Session found ::  Source IP : 3626579182 Source Port :1556  Destination IP : 2886732407  Destination Port : 61466  Sessid : 6 Payload: 47834

 Neither TCP nor UDP header found for Source IP : 3758096402 Destination IP : 2886732386

Session found ::  Source IP : 2886732407 Source Port :61466  Destination IP : 3626579182  Destination Port : 1556  Sessid : 6 Payload: 10200

Session found ::  Source IP : 2886732407 Source Port :61466  Destination IP : 3626579182  Destination Port : 1556  Sessid : 6 Payload: 15068

Session found ::  Source IP : 3626579182 Source Port :1556  Destination IP : 2886732407  Destination Port : 61466  Sessid : 6 Payload: 10200

 Neither TCP nor UDP header found for Source IP : 3758096402 Destination IP : 2886732386

Session inserted ::  Source IP : 1077861286 Source Port :1557  Destination IP : 2886732407  Destination Port : 57331  Sessid : 7 Payload: 13272

Session found ::  Source IP : 2886732407 Source Port :57331  Destination IP : 1077861286  Destination Port : 1557  Sessid : 7 Payload: 13272

Session found ::  Source IP : 1077861286 Source Port :1557  Destination IP : 2886732407  Destination Port : 57331  Sessid : 7 Payload: 10200

Session found ::  Source IP : 1077861286 Source Port :1557  Destination IP : 2886732407  Destination Port : 57331  Sessid : 7 Payload: 45273

Session found ::  Source IP : 2886732407 Source Port :57331  Destination IP : 1077861286  Destination Port : 1557  Sessid : 7 Payload: 10200

Session found ::  Source IP : 2886732407 Source Port :57331  Destination IP : 1077861286  Destination Port : 1557  Sessid : 7 Payload: 25049

Session found ::  Source IP : 1077861286 Source Port :1557  Destination IP : 2886732407  Destination Port : 57331  Sessid : 7 Payload: 21977
```

Identifying Statistical Info

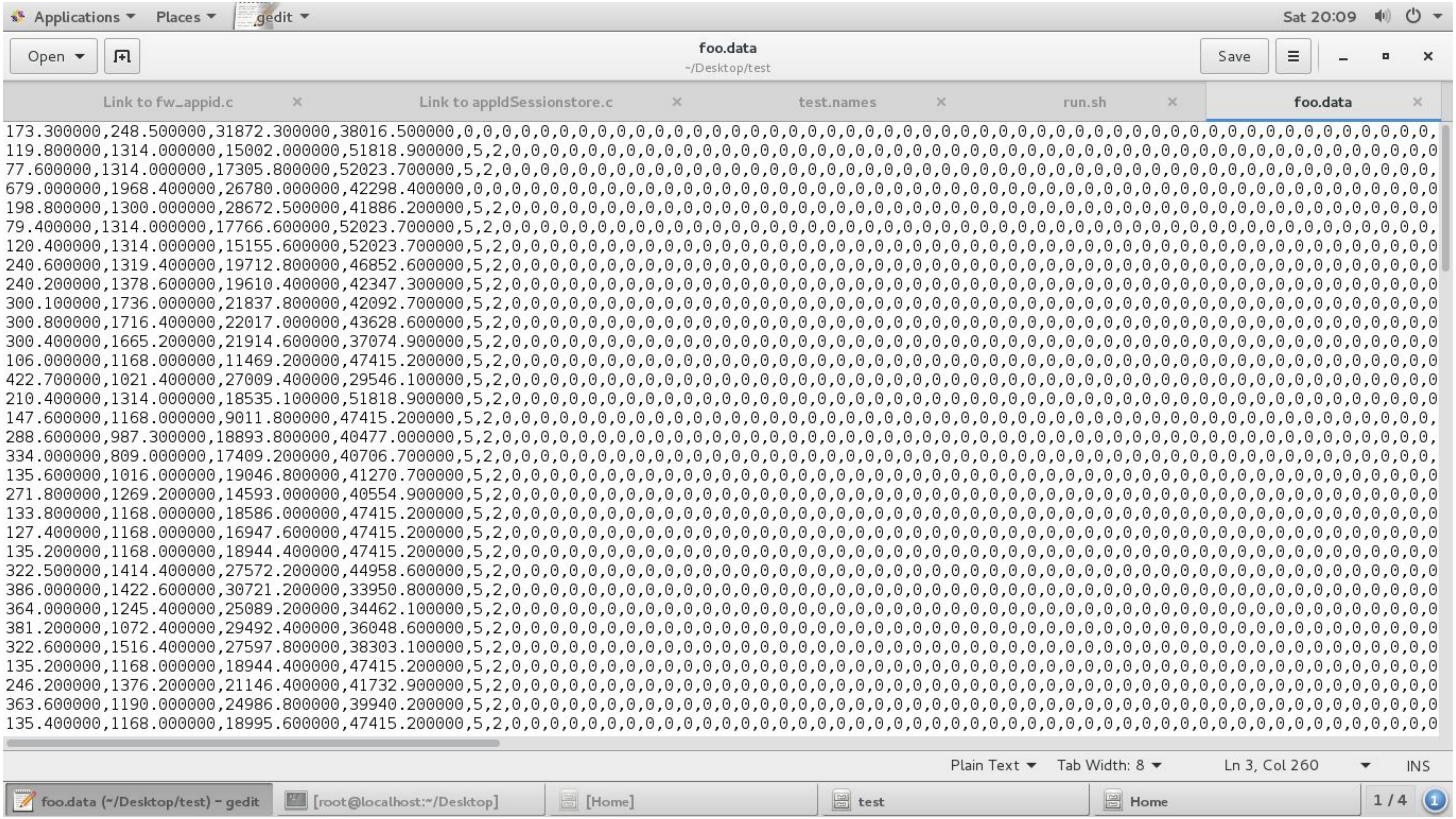
- Statistical Information is used to train the decision tree and later classify it using trained decision tree.
- Statistical information contains :
  - Application name
  - Average packet size of first 10 requests of session
  - Average packet size of first 10 replies of session
  - Average packet payload size of first 10 requests of session

Identifying Statistical Info

- Statistical information contains :
  - Average packet payload size of first 10 replies in session
  - Tcp options in request [1..10 requests]
  - Tcp options in reply [1..10 replies]

- More parameters can be added later on to ensure the efficient decision tree is made.
- Following dataset is created from the sessions identified before.

Dataset for training decision tree

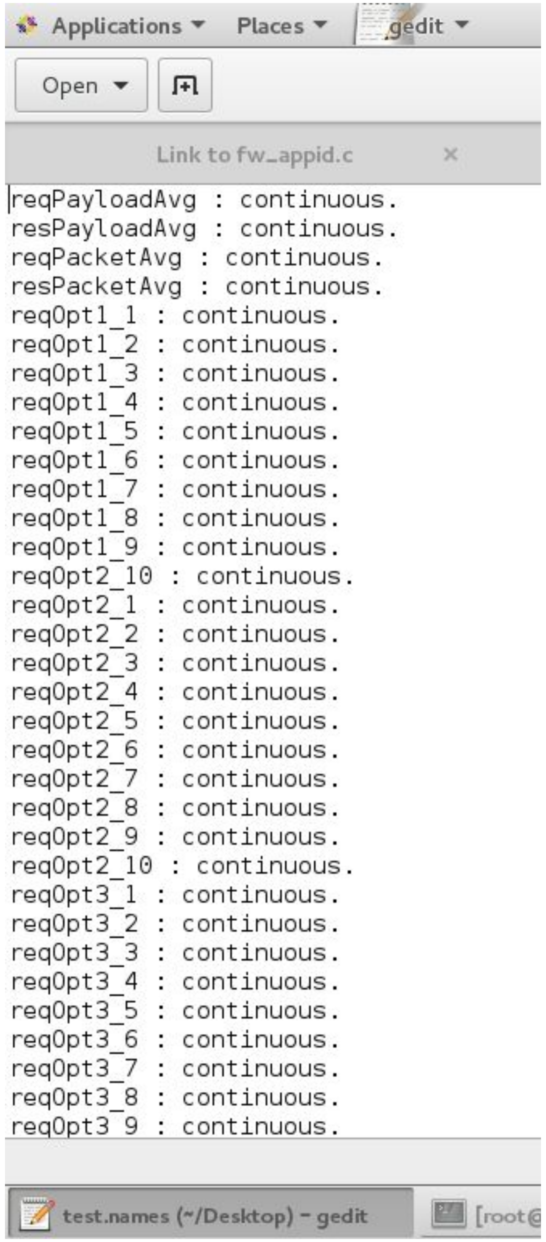


```
173.300000,248.500000,31872.300000,38016.500000,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
119.800000,1314.000000,15002.000000,51818.900000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
77.600000,1314.000000,17305.800000,52023.700000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
679.000000,1968.400000,26780.000000,42298.400000,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
198.800000,1300.000000,28672.500000,41886.200000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
79.400000,1314.000000,17766.600000,52023.700000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
120.400000,1314.000000,15155.600000,52023.700000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
240.600000,1319.400000,19712.800000,46852.600000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
240.200000,1378.600000,19610.400000,42347.300000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
300.100000,1736.000000,21837.800000,42092.700000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
300.800000,1716.400000,22017.000000,43628.600000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
300.400000,1665.200000,21914.600000,37074.900000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
106.000000,1168.000000,11469.200000,47415.200000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
422.700000,1021.400000,27009.400000,29546.100000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
210.400000,1314.000000,18535.100000,51818.900000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
147.600000,1168.000000,9011.800000,47415.200000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
288.600000,987.300000,18893.800000,40477.000000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
334.000000,809.000000,17409.200000,40706.700000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
135.600000,1016.000000,19046.800000,41270.700000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
271.800000,1269.200000,14593.000000,40554.900000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
133.800000,1168.000000,18586.000000,47415.200000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
127.400000,1168.000000,16947.600000,47415.200000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
135.200000,1168.000000,18944.400000,47415.200000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
322.500000,1414.400000,27572.200000,44958.600000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
386.000000,1422.600000,30721.200000,33950.800000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
364.000000,1245.400000,25089.200000,34462.100000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
381.200000,1072.400000,29492.400000,36048.600000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
322.600000,1516.400000,27597.800000,38303.100000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
135.200000,1168.000000,18944.400000,47415.200000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
246.200000,1376.200000,21146.400000,41732.900000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
363.600000,1190.000000,24986.800000,39940.200000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
135.400000,1168.000000,18995.600000,47415.200000,5,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
```

Classify Statistical Info using decision tree

- VFDT (Very Fast Decision Tree) is used to classify statistical information gained from application packets.
- VFDT needs to be trained first before using it to classify other data.
  - Decision tree is trained before it is used in snort to classify the data with various application information.
  - Training decision tree before snort instance is running, is necessary to avoid initial false identification.
  - Trained decision tree is stored in file, which later is used to classify the data.

Attributes



```
reqPayloadAvg : continuous.
resPayloadAvg : continuous.
reqPacketAvg : continuous.
resPacketAvg : continuous.
reqOpt1_1 : continuous.
reqOpt1_2 : continuous.
reqOpt1_3 : continuous.
reqOpt1_4 : continuous.
reqOpt1_5 : continuous.
reqOpt1_6 : continuous.
reqOpt1_7 : continuous.
reqOpt1_8 : continuous.
reqOpt1_9 : continuous.
reqOpt2_10 : continuous.
reqOpt2_1 : continuous.
reqOpt2_2 : continuous.
reqOpt2_3 : continuous.
reqOpt2_4 : continuous.
reqOpt2_5 : continuous.
reqOpt2_6 : continuous.
reqOpt2_7 : continuous.
reqOpt2_8 : continuous.
reqOpt2_9 : continuous.
reqOpt2_10 : continuous.
reqOpt3_1 : continuous.
reqOpt3_2 : continuous.
reqOpt3_3 : continuous.
reqOpt3_4 : continuous.
reqOpt3_5 : continuous.
reqOpt3_6 : continuous.
reqOpt3_7 : continuous.
reqOpt3_8 : continuous.
reqOpt3_9 : continuous.
```

Classify Statistical Info using decision tree

- Possible problem with decision tree can be , it always classifies given data to nearest match.
- Which creates confusion of taking the decision of whether to train the decision tree using the gained statistical info , or the classify it using decision tree?
- One solution can be calculating confidence factor of the decision tree i.e. calculating probability of correct decision
- If confidence factor is above 70%(used initially) the decision is correct else it is of a new application and decision tree should be trained with the data.

Application Identification

- New application found during the learning phase is stored in a database.
- If application match is found in decision tree, appropriate message is shown.
- Current application Identification is done based on packet headers which can be extended using deep packet inspection.

GitHub URL :

https://github.com/prabhakarniraula/snort-openappid-machinelearning