School of Mathematics, Science and, Engineering

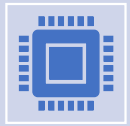
Enhancing IT and OT System Detection in Networks: A Strategy Integrating Data Labeling, Log Merging, and Visualization for Efficient Machine Learning Models

IT/OT System Identification

Candan Martin

Professor: Brian MacDougald

Disclaimers



Worked within a team environment with the goal of building a machine learning model that would accurately identify Operational Technology Devices and IT devices on a network.

My part in this team project consisted of identifying methods to label the data and then visualizing it to provide greater insights into OT and IT behavior within the network.

My contribution would enable the team to develop a more efficient IT/OT identification model by providing labeled data and insightful insights via visualization of the data

Explanation of OT and IT

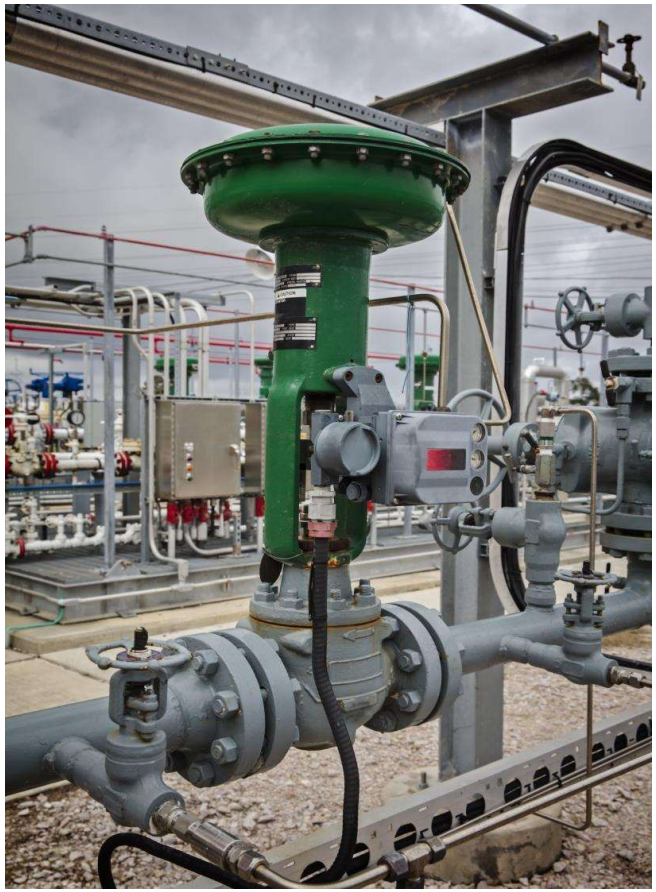
Operational Technology (OT):

- Purpose: Monitors and control industrial processes.
- Devices: PLC(Programmable Logical Controllers), SCADA systems (Supervisory control and data acquisition), sensors, actuators.
- Importance: Vital for efficient operation of critical infrastructure

Information Technology (IT):

- Purpose: Manages electronic data processing and communication.
- Devices: Computers, servers, routers, data centers.
- Importance: Central to data management and organizational operations.

Actuators



PLC's

Problem Description

Challenge: Accurate data labeling for IT and OT system detection using machine learning, coupled with comprehensive data visualization.

Approach: Integrating IP data labeling, log merging, and visualization for in-depth insights.

Goal: Develop a sophisticated machine learning model for effective system categorization and enhanced network security.

Objective and Background Information

Objective:

- Goal: Utilize data labeling, log file merging, and visualization for insights extraction.
- Purpose: Aid in developing a machine learning model for enhancing network visibility and cybersecurity.
- Focus: Identifying IT and OT systems in network environments.

Background:

- Challenge: Managing IT and OT systems in networks requires robust security and efficient operations.
- SCADA Networks were not designed with Security in mind which, makes them vulnerable attacks and even more so when connect to an IT network.
- Strategy: Implement data labeling and log consolidation for deeper network insights.
- Outcome: Informed decision-making, better device detection, and groundwork for anomaly detection research.

Impact and Significance:

Data Labeling: Essential for machine learning model development; ensures precise system identification and categorization through accurate labeled data.

Log Merging: Merges log files to analyze network traffic data, captured by Zeek, into a single file. Simplifies the analysis of network logs, facilitating machine learning development.

Data Visualization: Provides actionable insights on IT and OT behavior, aiding in the refinement and optimization of the machine learning model.

Outcome: Achieves a functionally running model, offers labeled data for further machine learning research in OT/IT networks, and delivers visualized data that illustrates the behavior of IT and OT devices on a network.

Project Scope - Initial

- **Core Focus**: Developing a machine learning model to identify and categorize IT and OT systems in network environments.
- **Implementation**: Utilizing Zeek for network data analysis, followed by data labeling and decision tree model development.
- **Objective**: To build a foundational machine learning model, enhancing the detection and categorization of IT and OT systems within networks.

Reason For Change:

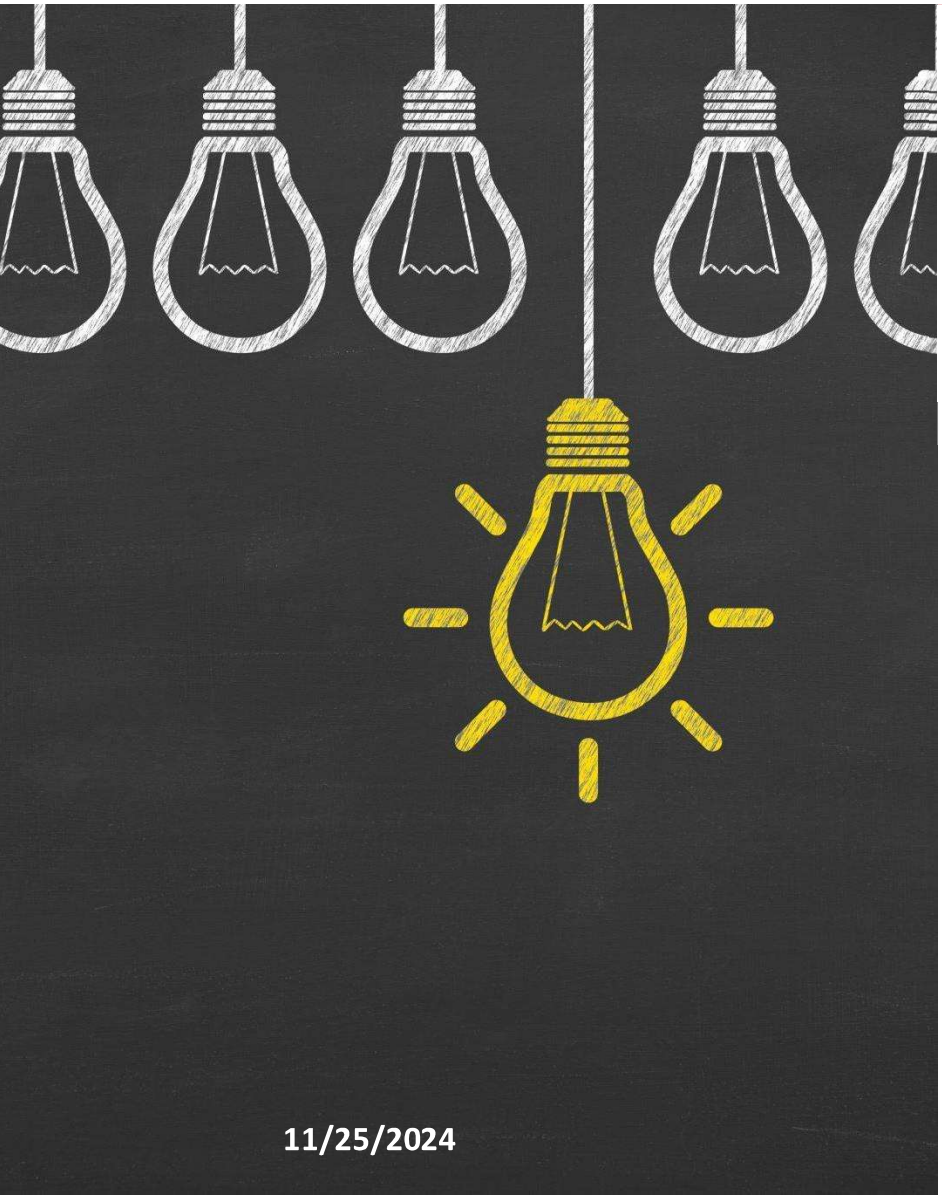
- **Delays in Data Acquisition**: Challenges in obtaining the network architecture and relevant data, specifically IP addresses for labeling OT and IT devices, necessitated seeking alternative, albeit less effective, methods for data labeling. This adjustment was crucial to proceed with supervised machine learning applications.
- **Technical Expertise Deficiency**: There was a gap in technical expertise, particularly in effectively engineering features for decision tree models. This limitation was compounded by restricted time available for learning and testing feature engineering techniques, primarily due to delays encountered in the data labeling process.

Project Scope - Revised

Core Focus: Visualization of data from IT and OT systems within a network using Zeek logs and providing accuracty labeled data.

Implementation: Employing Python and various data visualization libraries to reveal data patterns.

Objective: Identify clustering and patterns in labeled IT and OT data, preprocessed and merged into a single .pkl file for enhanced visualization.



11/25/2024

Scope Limitations

Data Quality and Availability:

- Dependence of data visualization efficacy and accuracy on the quality and availability of Zeek logs.
- Impact of incomplete, erroneous, or homogeneous logs on identifying meaningful patterns and clusters.

Dependencies on Other Teams: Impact on Data Labeling and Model Development

- Inter-Team Coordination Delays: Slower data labeling and insights gathering due to reliance on external team data.
- Workflow Bottlenecks: Varying priorities and timelines among teams causing delays in model development.

Computational Resource Constraints in a Team Environment:

- Challenges in environments with competition for limited GPUs and CPUs.
- Impact of GPU scarcity on time required for scripts, for loops, and large data set visualization.
- Effects on scalability and overall process efficiency.(Memory Allocation Issue when merging to many Panda DF's)

Methodology

Obtain Labeled Data:

Gather IT and OT system traffic data.

Label the data to indicate whether each entry corresponds to an IT or OT system.

Preprocess Data for Visualization:

Merge the labeled data into a single pkl file for easier handling.

Prepare the data for visualization by performing any necessary cleaning and transformation.

Explore Data Visualization Techniques:

Utilize Python-based data visualization libraries (e.g., Matplotlib, Seaborn) to create visual representations of the data. Explore various visualization techniques, such as scatter plots, histograms, box plots, and heatmaps, to understand data patterns and distributions.

Identify Clustering in Data:

Apply clustering algorithms (scatterplots2-3,bar charts, 3D scatterplot) on the labeled data to identify groups of similar IT and OT systems.

Visualize the clusters to gain insights into the natural groupings present in the data.

Interpret Results:

Analyze the visualizations and clustering results to identify meaningful patterns and distinctions between IT and OT systems. Interpret the findings to draw conclusions about the relationships and characteristics of the systems within the network.

Management Timeline

Plan A

Project: IT/OT System Identification

Company Name

Candan Martin

Project Start Date:	5/31/2023										
	May	June				July				August	
	Wk1(30-2)	Wk2(5-9)	Wk3(12-14)	Wk4(19-23)	Wk5(26-30)	Wk6(3-7)	Wk7(10-14)	Wk8(17-21)	Wk9(24-28)	Wk10(31-4)	Wk11(7-11)
Planning											
Researching											
Developing											
Testing											
Report											
Presentation											
Upload Documents											

Plan B

Project: IT/OT System Identification

Company Name

Candan Martin

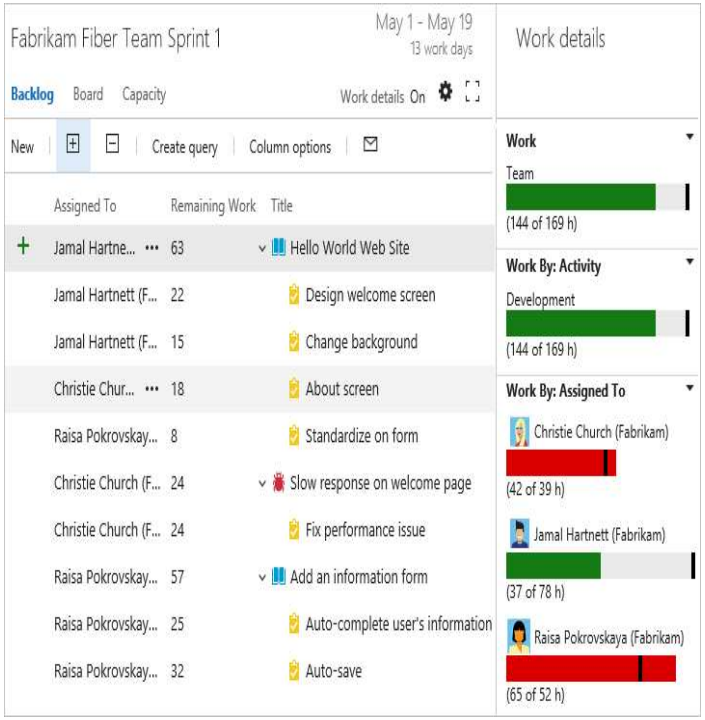
Project Start Date:	5/31/2023										
	May	June				July				August	
	Wk1(30-2)	Wk2(5-9)	Wk3(12-14)	Wk4(19-23)	Wk5(26-30)	Wk6(3-7)	Wk7(10-14)	Wk8(17-21)	Wk9(24-28)	Wk10(31-4)	Wk11(7-11)
Planning											
Researching											
Developing											
Testing											
Report											
Presentation											
Upload Documents											
Present											
Final Documentation											

Management Timeline

Word Document

Table of Contents

Azure DevOps



Milestones and Progress

Steps taken to address the problem

Project Milestones

- MILESTONE 1: LEARN HOW TO USE ZEEK
- MILESTONE 2: IDENTIFY WHAT LOGS TO USE IN MODEL
- MILESTONE 3: LABEL THE DATA FOR TRAINING
- MILESTONE 4: MERGE LOG FILES INTO SINGLE PICKLE FILE
- MILESTONE 5:CREATE SCATTER PLOT OF DATA TO IDENTIFY CLUSTERING OR PATTERNS IN IT AND OT LABELED DATA
- MILESTONE 6: CREATE 3D SCATTER PLOT OF THE DATA TO FURTHER IDENTIFY CLUSTERING OR PATTERNS IN IT AND OT LABELED DATA
- MILESTONE 7: CREATE VISUALIZE IT AND OT DATA VIA BAR CHART TO BUILDING BETTER UNDERSTANDING OF THE DATA
- MILESTONE 8:VISUALIZE THE DECISION BRANCHES

Milestone 1: Learn how to use Zeek

Reviewed Zeek log Documentation

Learned how to read in Zeek logs

conn.log | IP, TCP, UDP, ICMP connection details

FIELD	TYPE	DESCRIPTION
ts	time	Timestamp of first packet
uid	string	Unique identifier of connection
id	record conn_id	Connection's 4-tuple of endpoint addresses
proto	enum	Transport layer protocol of connection
service	string	Application protocol ID sent over connection
duration	interval	How long connection lasted
orig_bytes	count	Number of payload bytes originator sent
resp_bytes	count	Number of payload bytes responder sent
conn_state	string	Connection state (see conn.log > conn_state)
local_orig	bool	Value=T if connection originated locally
local_resp	bool	Value=T if connection responded locally
missed_bytes	count	Number of bytes missed (packet loss)
history	string	Connection state history (see conn.log > history)
orig_pkts	count	Number of packets originator sent
orig_ip_bytes	count	Number of originator IP bytes (via IP total_length header field)
resp_pkts	count	Number of packets responder sent
resp_ip_bytes	count	Number of responder IP bytes (via IP total_length header field)
tunnel_parents	table	If tunneled, connection UID value of encapsulating parent(s)
orig_l2_addr	string	Link-layer address of originator
resp_l2_addr	string	Link-layer address of responder
vlan	int	Outer VLAN for connection
inner_vlan	int	Inner VLAN for connection

```
Code 2: Improved Data Collection
--------------------------------

import glob
import pandas as pd
import zat
from zat.log_to_dataframe import LogToDataFrame

# Use glob to find conn.log files in specific directories
paths = glob.glob('/opt/…/hedgehog62-230403-*/conn.log')

# Initialize an empty list to store DataFrames
dfs = []

# Iterate over the conn.log files and create DataFrames
for path in paths:
    dfs.append(LogToDataFrame.create_dataframe(path))

# Concatenate the DataFrames into a single DataFrame
df = pd.concat(dfs, ignore_index=True)
```

Milestone 2: IDENTIFY WHAT LOGS TO USE IN MODEL

Install Additional Parsers: Implement additional parsers for industrial control system protocols.

Challenges with Parser Installation:

- Encountered issues where some parsers failed to install, and those installed didn't identify OT protocols in PCAP files.

Reasoning:

- Identified the root cause as the absence of sensors at the machinery level necessary for collecting OT protocol data.

Adapting Strategy: Shifted focus to utilize existing Zeek logs, adjusting the approach to work with available data.

Milestone 3: LABEL THE DATA FOR TRAINING

Initial Labeling Strategy

Protocol-Port Association:

Identified that certain ports are linked with specific protocols, such as port 443 for HTTPS, primarily used by IT systems.

Research on OT and IT Protocols:

Conducted thorough research to distinguish protocols commonly used by OT devices from those associated with IT.

Port Identification for Specific Protocols:

Mapped out the primary ports used for specific OT and IT protocols.

Challenge with Proprietary OT Protocols:

Recognized inefficiencies due to the proprietary nature of many OT protocols, which vary by company and their unique assignment of protocols to ports.

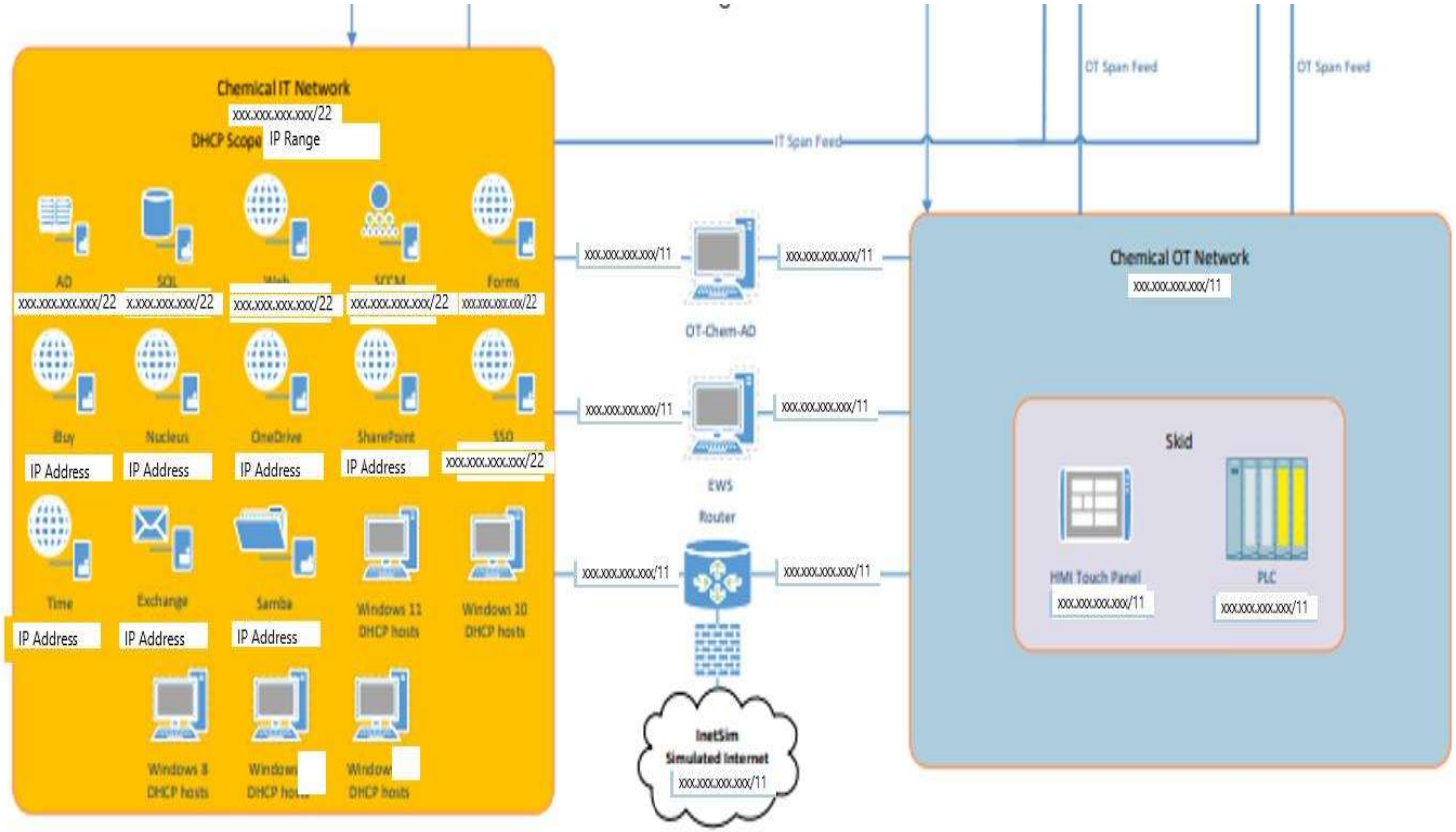
Revised Labeling Strategy

Shift in Labeling Approach:

- Transitioned to IP-based labeling to enhance efficiency and effectiveness.

Leveraging Network Team Resources:

- Utilized network map blueprints provided by the network team to facilitate the new IP-based labeling strategy.

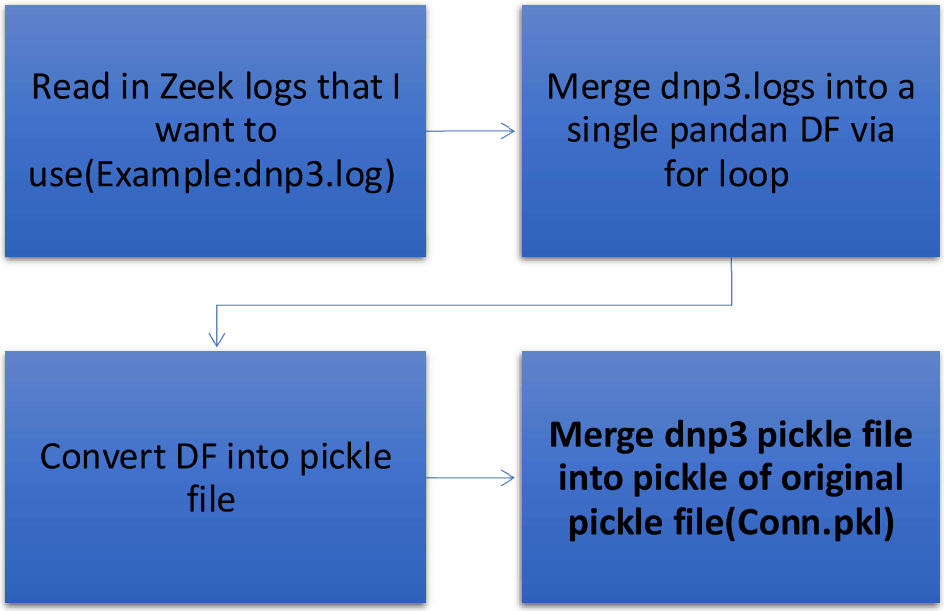


Results:

```
[9]: # Print count for each label
label_counts = df['label'].value_counts()
print(label_counts)
# Calculate the grand total count
grand_total = label_counts.sum()
# Print the grand total
print("Grand Total:", grand_total)
```

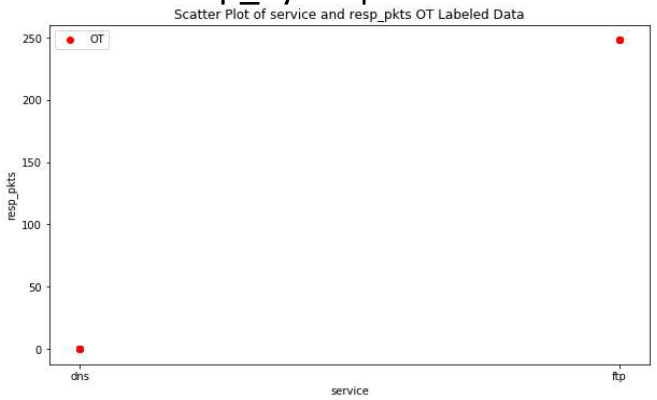
```
IT    1018894
OT       5721
Name: label, dtype: int64
Grand Total: 1024615
```

Milestone 4: MERGE LOG FILES INTO SINGLE PICKLE FILE

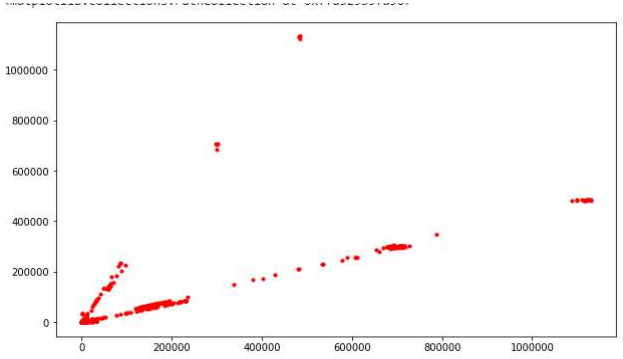


Milestone 5: CREATE SCATTER PLOT OF DATA TO IDENTIFY CLUSTERING OR PATTERNS IN IT AND OT LABELED DATA

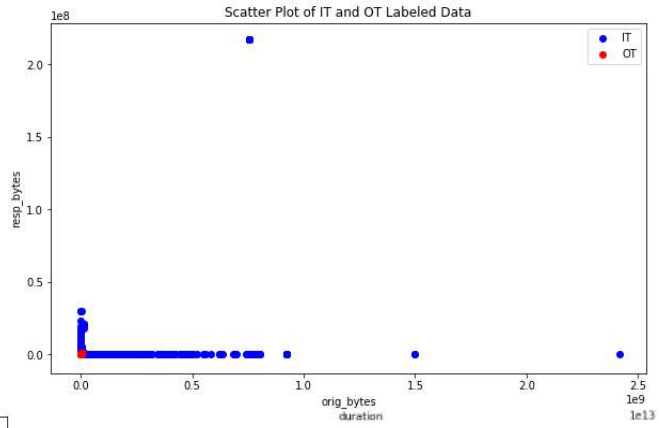
Y-axis: Resp_bytes | X-axis: Services

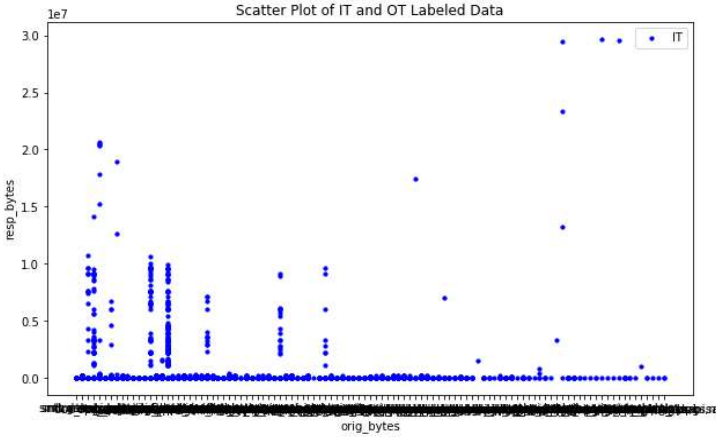


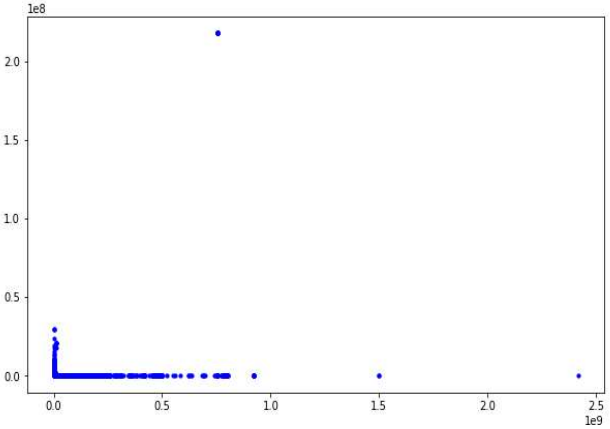
Y-axis: Resp_bytes | X-axis: Orig_bytes

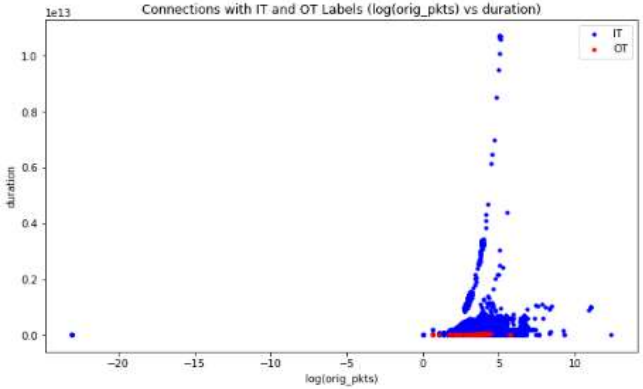


IT and OT plotted together

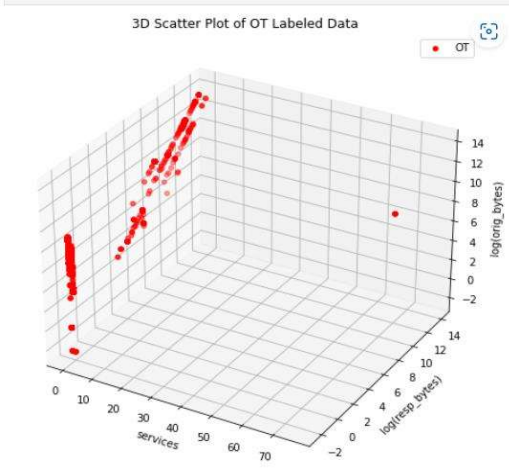






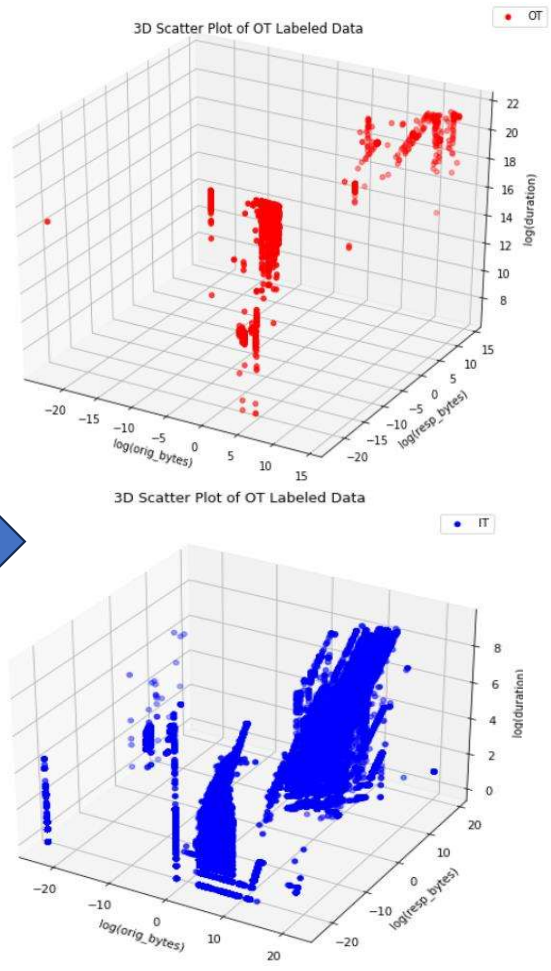


Milestone 6: 3D Scatterplot Visualization

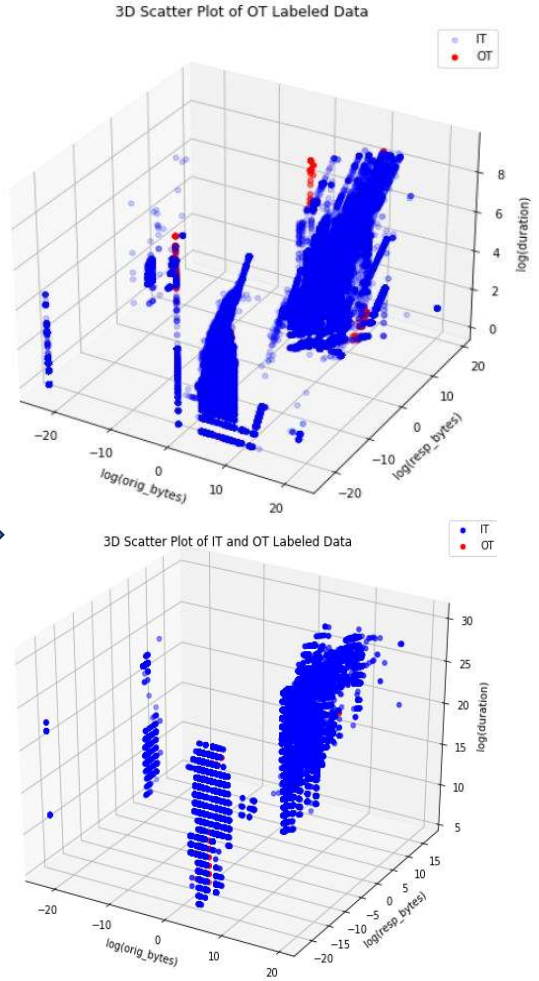


Deeper insight on service outlier in OT Data

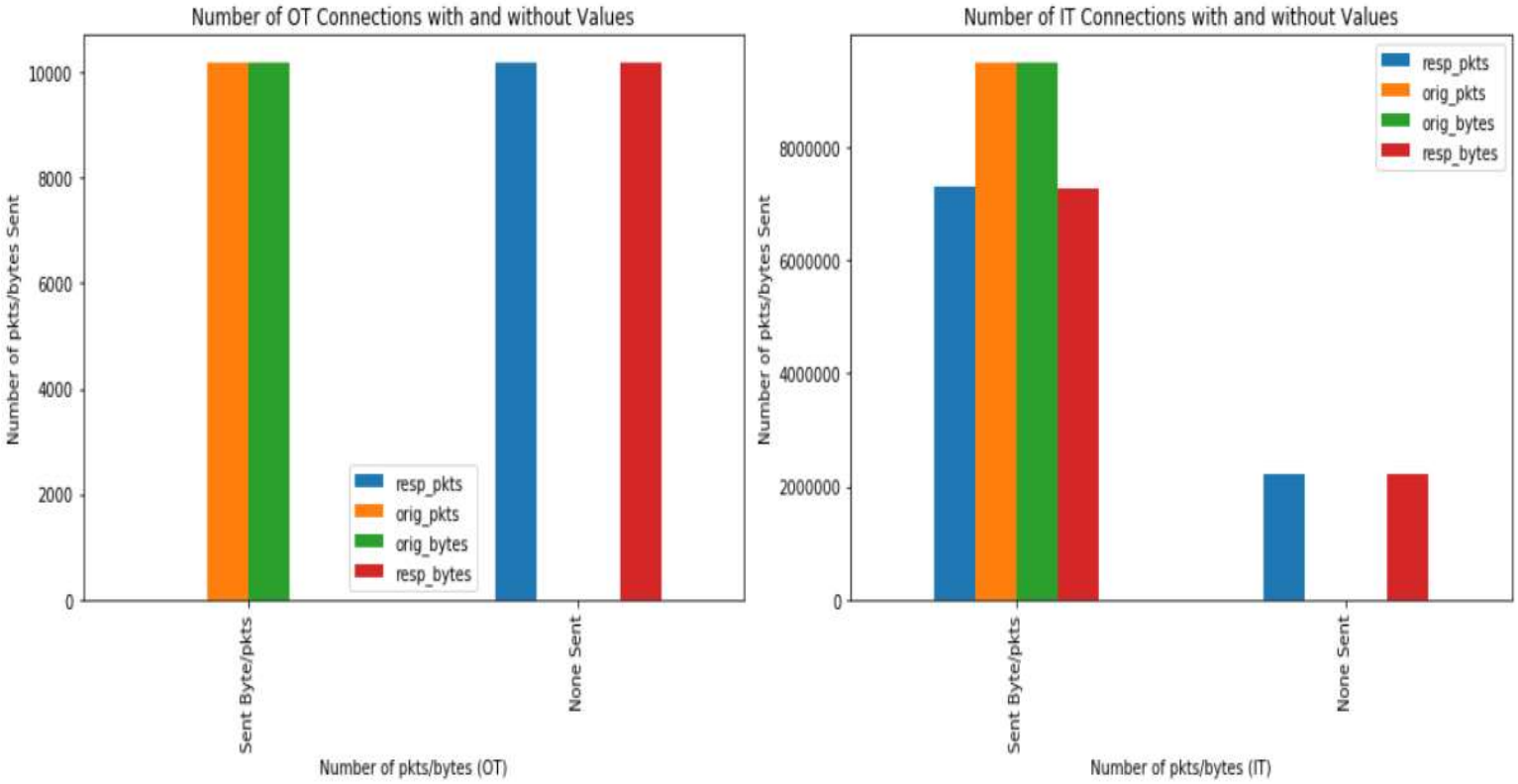
IT and OT plotted Separately



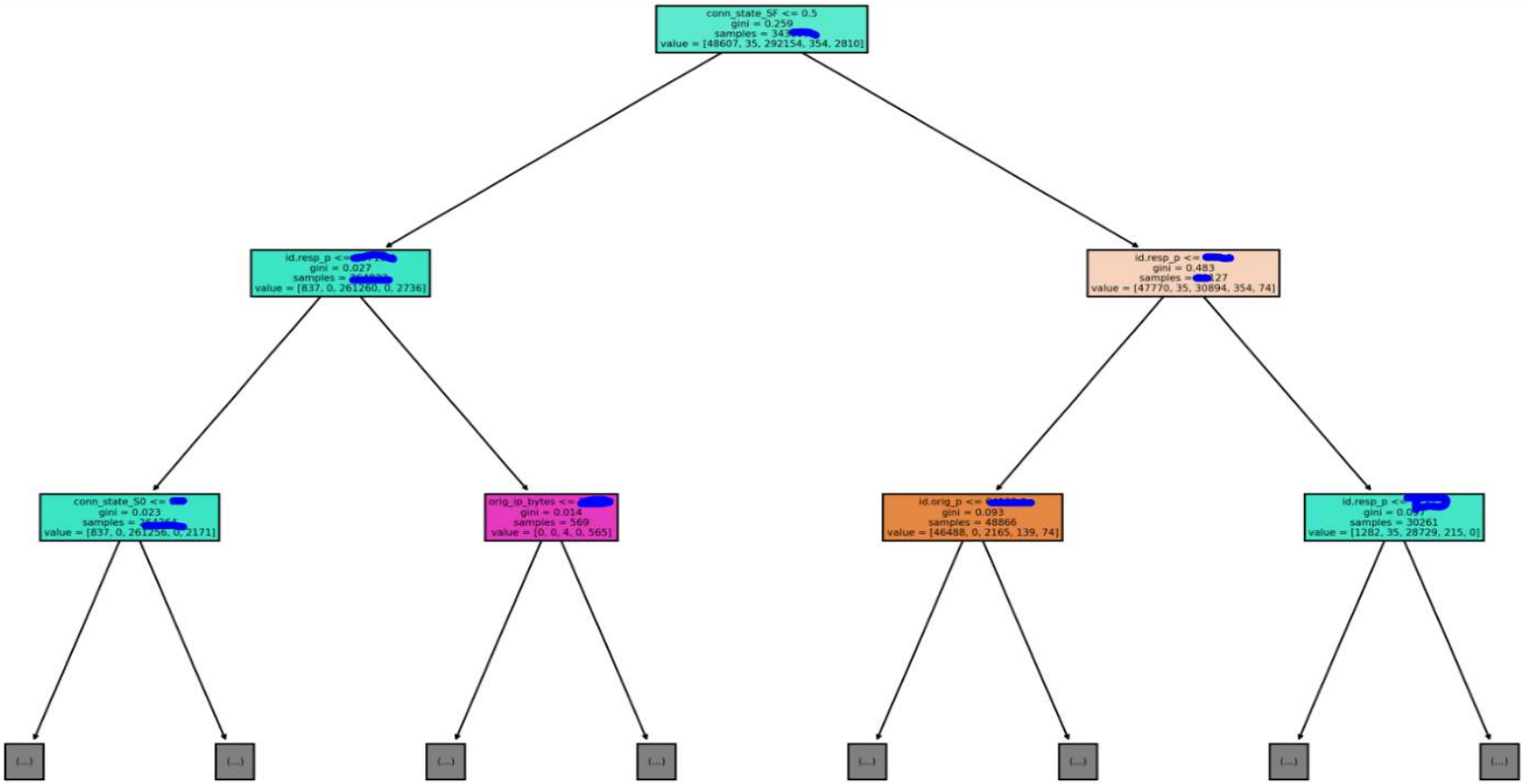
IT and OT plotted together



Milestone 7: Visualization of Bar Chart

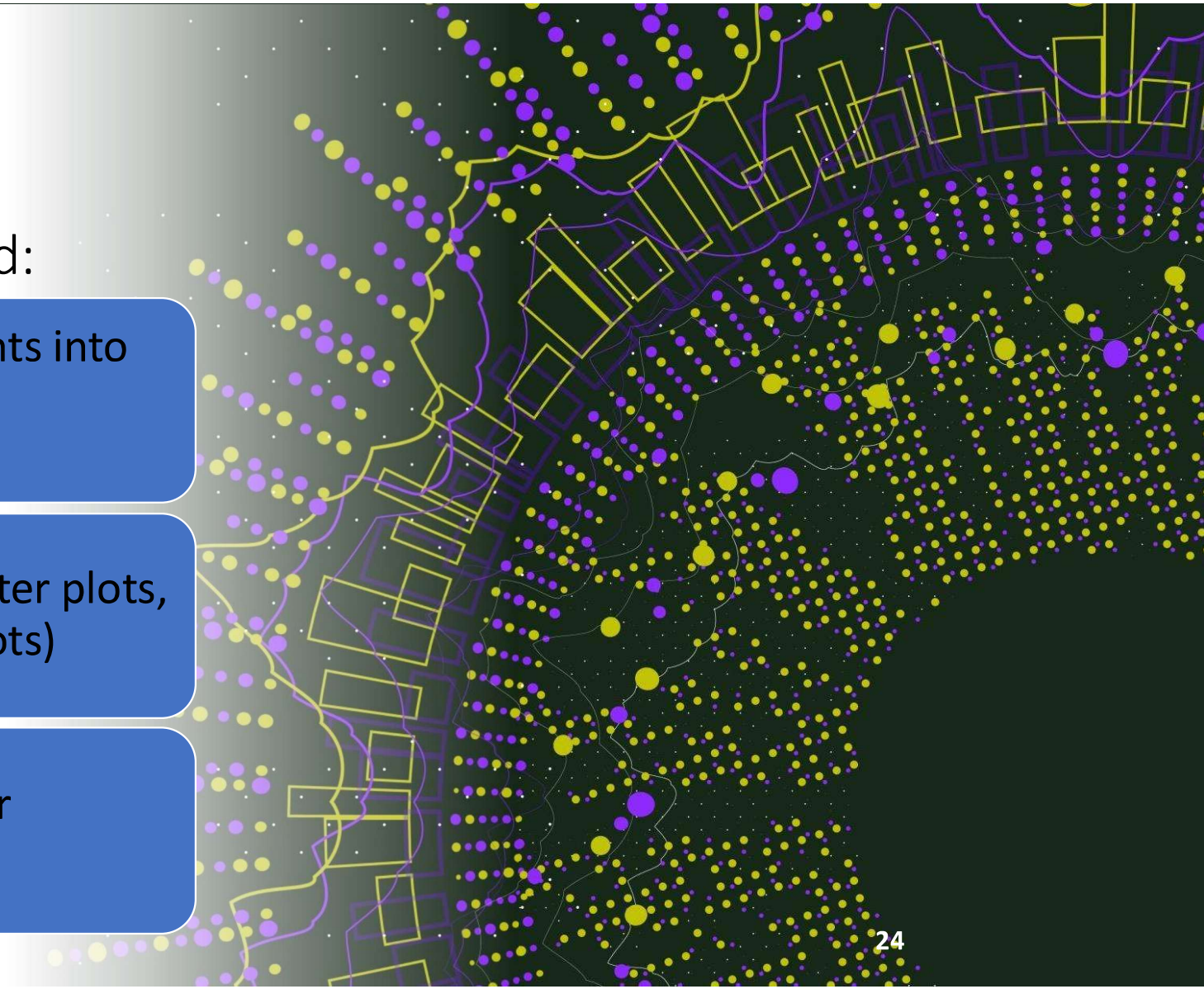


Milestone 8:VISUALIZE THE DECISION BRANCHES



Solutions Resulted:

- Visualized network insights into data patterns for SCADA Network.
- Visualization charts (scatter plots, bar charts, 3D Scatter Plots)
- Labeled Network data for supervised learning

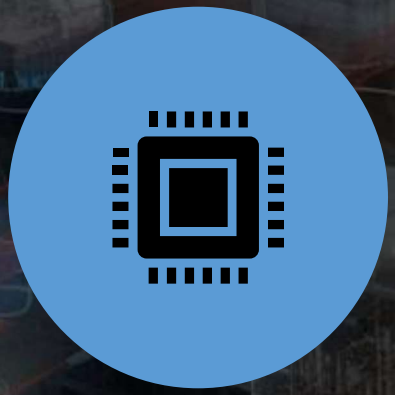


Recommendations:

Collaborate closely with the network team managing the underlying network data, aiming to acquire IP information for devices within the actual network architecture. This collaboration can significantly expedite the process, saving months that would otherwise be spent on finding alternative methods for effective data labeling.

For projects involving Zeek logged data, it's recommended to have a team member proficient in Zeek. This expertise ensures that if there's a need for additional parsers or Zeek logs, they can be integrated efficiently and promptly.

Value to the public



ADVANCING DEVICE DETECTION SOFTWARE FOR CRITICAL INFRASTRUCTURE:

SETTING A FOUNDATION FOR FUTURE INNOVATIONS

Conclusion

- Worked within a team environment with the goal of producing a machine learning model to detect It/OT devices on SCADA Networks
- Discovered behaviors in SCADA Network traffic of IT and OT devices for machine learning team
- Discovered that OT devices have significantly few bytes to sent and received and mainly do not use IT protocals and services
- My contribution would enable the team to develop a more efficient IT/OT identification model by providing labeled data and insightful insights via visualization of the data
- The results establish a foundation for future studies aimed at creating models capable of detecting anomalies and identifying specific devices within SCADA networks.

Questions