

# ANALYZING LOGS DATA WITH AI

MCS 2025 Cyber 242

Lecturer: Dominic Ligot



# Dominic Ligot

CirroLytix Research Services

Founder and CTO

IT-BPM Association of the Philippines (IBPAP)

Consultant for AI and Technology

Data Ethics PH

Founder

Analytics and AI Association of the Philippines

Co-founder, Board of Trustees

PCIJ

Board of Trustees

University of Asia and the Pacific

Co-author, Master in Applied Business Analytics

UK Department of Science, Innovation, and Technology

Member, International Expert Panel on Advanced AI Safety



# OBJECTIVES

- Understand the nature of log data
- Learn how to transform log data for analysis
- Apply various data analysis techniques to find anomalies in log data
- By “AI” we refer to data analysis tools and algorithms.

# UNDERSTANDING LOG DATA

# PACKET CAPTURE LOG (WIRESHARK)

The image displays a Wireshark packet capture log for interface \*enp0s3. The log shows a series of DNS transactions. The first transaction (No. 429) is a query from 192.168.8.10 to 8.8.4.4 for www.google.com. The second transaction (No. 430) is a query from 192.168.8.10 to 8.8.4.4 for AAAA www.google.com. The third transaction (No. 450) is a response from 8.8.4.4 to 192.168.8.10 for push.services.mozilla.com. The fourth transaction (No. 458) is a response from 8.8.4.4 to 192.168.8.10 for push.services.mozilla.com. The fifth transaction (No. 471) is a query from 192.168.8.10 to 8.8.4.4 for getpocket.cdn.mozilla.net. The sixth transaction (No. 472) is a query from 192.168.8.10 to 8.8.4.4 for getpocket.cdn.mozilla.net. The seventh transaction (No. 488) is a response from 8.8.4.4 to 192.168.8.10 for www.google.com and 172.217.14...

No.	Time	Source	Destination	Protocol	Length	Info
429	4.599509274	192.168.8.10	8.8.4.4	DNS	74	Standard query 0x0eb0 A www.google.com
430	4.599535582	192.168.8.10	8.8.4.4	DNS	74	Standard query 0x1ebf AAAA www.google.com
450	4.604775549	8.8.4.4	192.168.8.10	DNS	139	Standard query response 0x76ac A push.services.mozilla.com CN...
458	4.605231564	8.8.4.4	192.168.8.10	DNS	205	Standard query response 0x73a9 AAAA push.services.mozilla.com...
471	4.622154867	192.168.8.10	8.8.4.4	DNS	85	Standard query 0x6689 A getpocket.cdn.mozilla.net
472	4.622177628	192.168.8.10	8.8.4.4	DNS	85	Standard query 0x458a AAAA getpocket.cdn.mozilla.net
488	4.625366655	8.8.4.4	192.168.8.10	DNS	90	Standard query response 0x0eb0 A www.google.com A 172.217.14...

Frame 429: 74 bytes on wire (592 bits), 74 bytes captured (592 bits) on interface enp0s3, id 0

- Ethernet II, Src: PcsCompu\_82:75:df (08:00:27:82:75:df), Dst: Ignition\_01:f6:50 (00:78:cd:01:f6:50)
- Internet Protocol Version 4, Src: 192.168.8.10, Dst: 8.8.4.4
- User Datagram Protocol, Src Port: 58029, Dst Port: 53
- Domain Name System (query)
  - Transaction ID: 0x0eb0
  - Flags: 0x0100 Standard query
  - Questions: 1
  - Answer RRs: 0
  - Authority RRs: 0
  - Additional RRs: 0
  - Queries
    - www.google.com: type A, class IN

[Response In: 488]

0000 00 78 cd 01 f6 50 08 00 27 82 75 df 08 00 45 00 .x...P.. 'u...E.  
0010 00 3c 6c 3c 40 00 40 11 f9 b6 c0 a8 08 0a 08 08 .<l<@.@. ....  
0020 04 04 e2 ad 00 35 00 28 d4 f7 0e b0 01 00 00 01 .....5 ( .....

wireshark\_enp0s3\_20200427055131\_kGXclt.pcapng Packets: 3286 · Displayed: 442 (13.5%) · Dropped: 0 (0.0%) Profile: Default

## DATA AVAILABLE IN PACKET LOGS

- Timestamp
- Source IP
- Destination IP
- Protocol
- Source Port
- Destination Port
- Flags

# LIMITATIONS OF LOG DATA

- Snapshot of network activity
- Limited context
- High volume and velocity
- Noise and redundancy
- Data quality issues
- Lack of standardization
- Retention limits
- Privacy and security concerns

## THINGS TO BEAR IN MIND

- Data analysis does not replace domain knowledge – but complements it
- Data analysis is better at generating questions – but it remains the duty of the security analyst to put a case together
- Data analysis is about pattern recognition – look for similarities and anomalies as clues to formulating a hypothesis



# SETTING UP GOOGLE COLLAB

## GITHUB REPOSITORY

- Access/clone/fork from here:
  - [https://github.com/docligot.com/pcap\\_labs](https://github.com/docligot.com/pcap_labs)

# PCAP CONVERTER NOTEBOOK

The screenshot displays a Jupyter Notebook environment. The top bar shows the notebook title 'PCAP\_Converter.ipynb' and a star icon. Below it is a menu bar with options: File, Edit, View, Insert, Runtime, Tools, Help, and a link 'All changes saved'. On the right of the top bar are icons for Comment, Share, settings, and a user profile.

The left sidebar contains a 'Files' panel with a search icon and a file tree. The tree shows a root directory with subdirectories: '..', '.config', '.ipynb\_checkpoints', 'Exercise 1', 'Exercise 3', and 'sample\_data'. The 'sample\_data' directory is selected.

The main area of the notebook is divided into two sections. The top section is a code cell with the title 'PCAP Log Data Analysis' and a description: 'Compiling some helper functions to help with PCAP Log analysis:'. It lists four bullet points: 'Convert PCAP to CSV', 'Network activity analysis', 'Graph theory analysis', and 'Clustering: One hot encoding, Dimensional Reduction, Clustering'.

The bottom section is a terminal output cell. It shows the command '[55] !pip install scapy' and the output 'Requirement already satisfied: scapy in /usr/local/lib/python3.10/dist-packages (2.6.0)'. Below the terminal output is a code cell with the title '# Convert PCAP to CSV' and the following code:

```
from scapy.all import rdpcap
import csv
import socket
from datetime import datetime
```

# PCAP CONVERTER NOTEBOOK

- Scripts are provided for the following functions:
  - Convert PCAP to CSV
  - Network Activity Analysis
  - Graph Theory Analysis
  - Clustering Analysis:
    - One-hot-encoding
    - Dimensional Reduction
    - Clustering















# PCAP LAB EXERCISES

## PCAP LAB EXERCISES

- 3 exercises are provided
- Exercise 1 will have forensic analysis, alerts, and the PCAP file
- For Exercise 1, we will stick purely to data analysis, no need for domain explanations.
- Exercises 2 and 3 will only have the PCAP file. Forensic analysis and alerts will be provided later for discussion.
- For Exercises 2 and 3, class is allowed to speculate on possible explanations for the anomalies.

# DATA ANALYSIS OF PCAP

## TIME SERIES – EYEBALLING

Count of Packet_Number	Column Labels 			
Row Labels	 172.17.0.17	172.17.0.99	79.124.78.197	Grand Total
 :00	4	43	4	51
 :01	2	17	6	25
 :02	10	34	4	48
 :03	4	6	1	11
 :04	2	31	5	38
 :05	2	23	7	32
 :06	2	10		12
 :07	1	29	9	39
 :08	1	12	5	18
 :09	1	5	2	8
 :10	3	9	4	16
 :11	1	7	6	14



## IP TO PORT - EYEBALLING

Count of Packet_Number	Column Labels ▼												
Row Labels ▼	53	67	68	80	88	123	135	137	138	139	389	443	
23.45.119.143													14
23.45.119.144													199
23.45.119.147													13
40.119.249.228													22
40.126.28.12													20
40.126.28.22													11
46.254.34.201													504
52.109.0.142													13
52.109.0.91													16
52.113.194.132													73
79.124.78.197					261								
(blank)													
Grand Total	87	2	2	290	45	8	30	18	18	55	177	1536	

## IP TO IP - EYEBALLING

Count of Packet_Number	Column Labels ▼			
Row Labels ▼	172.17.0.99	255.255.255.255	(blank)	Grand Total
23.45.119.147	13			13
40.119.249.228	22			22
40.126.28.12	20			20
40.126.28.22	11			11
46.254.34.201	504			504
52.109.0.142	13			13
52.109.0.91	16			16
52.113.194.132	73			73
79.124.78.197	261			261
(blank)			294	294
<b>Grand Total</b>	<b>1826</b>	<b>2</b>	<b>294</b>	<b>2122</b>

# NETWORK ACTIVITY STATISTICS

```
=== Network Traffic Analysis Report ===
```

```
Basic Statistics:
```

```
total_packets: 5091
```

```
unique_ips: 42
```

```
unique_connections: 77
```

```
avg_packet_size: 423.1392537002293
```

```
duration_seconds: 3576.159984
```

```
Top Talkers:
```

```
|
```

```
Top Source IPs:
```

```
172.17.0.99: 2358 packets
```

```
172.17.0.17: 611 packets
```

```
46.254.34.201: 504 packets
```

```
79.124.78.197: 261 packets
```

```
23.45.119.144: 199 packets
```

```
23.221.24.69: 147 packets
```

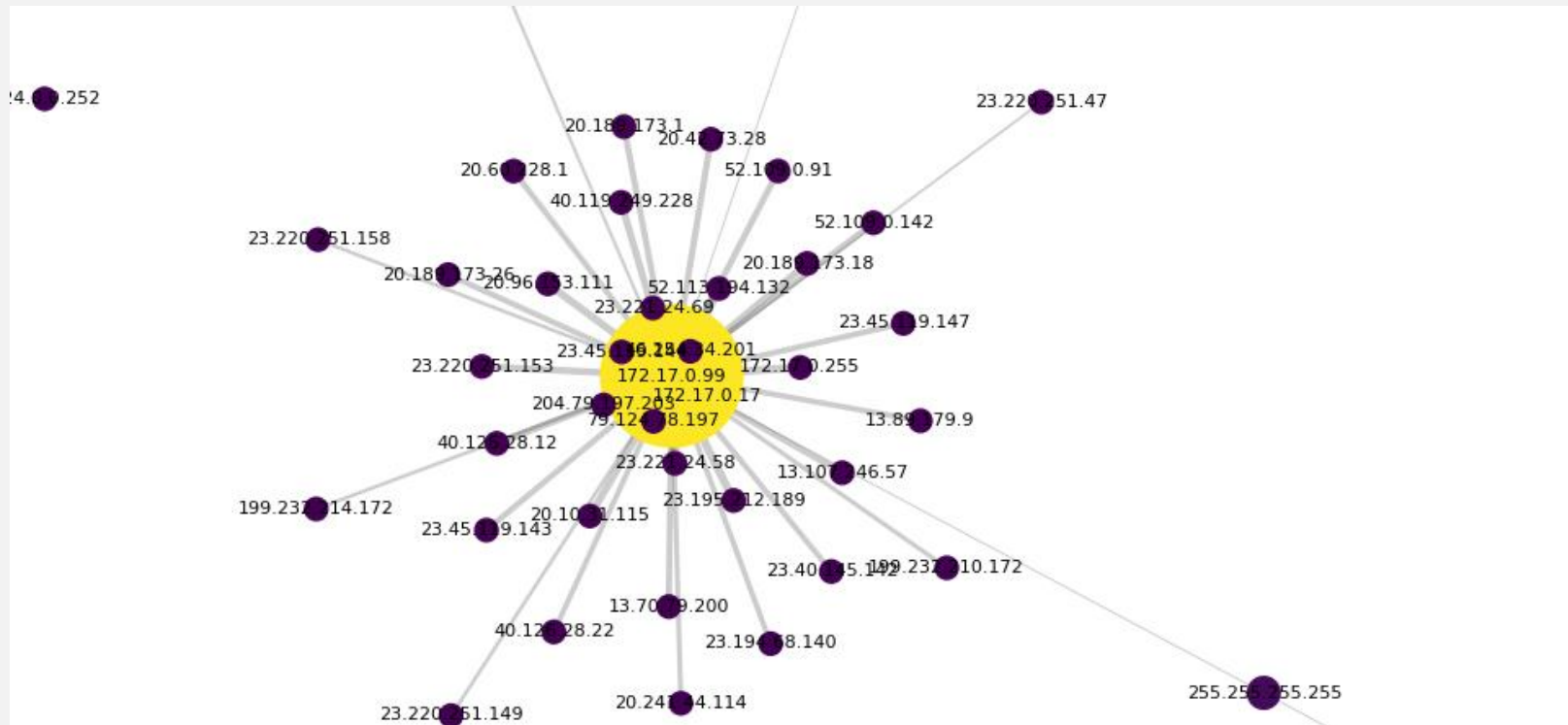
```
204.79.197.203: 144 packets
```

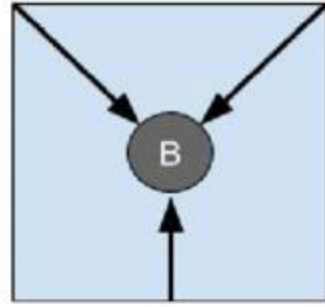
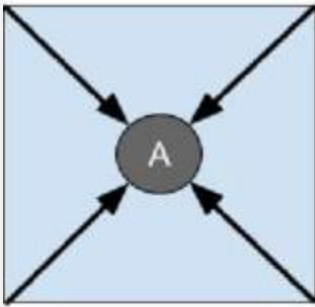
```
23.221.24.58: 91 packets
```

```
52.113.194.132: 73 packets
```

```
23.195.212.189: 37 packets
```

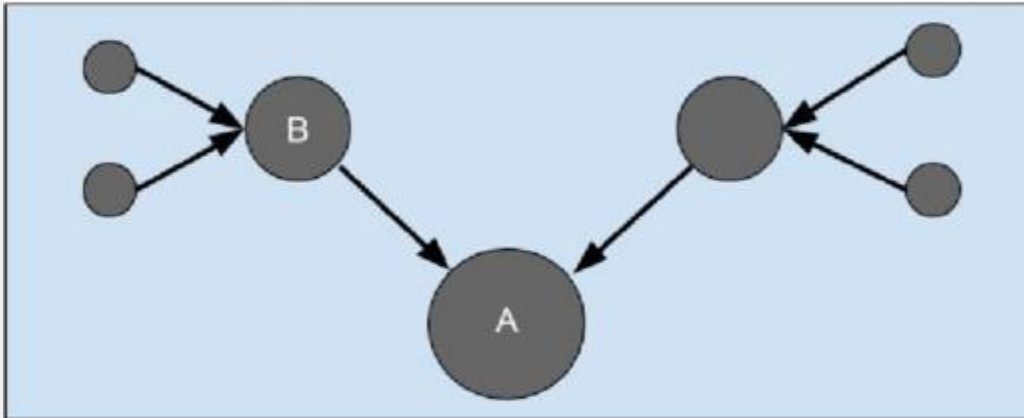
# NETWORK VISUALIZATION





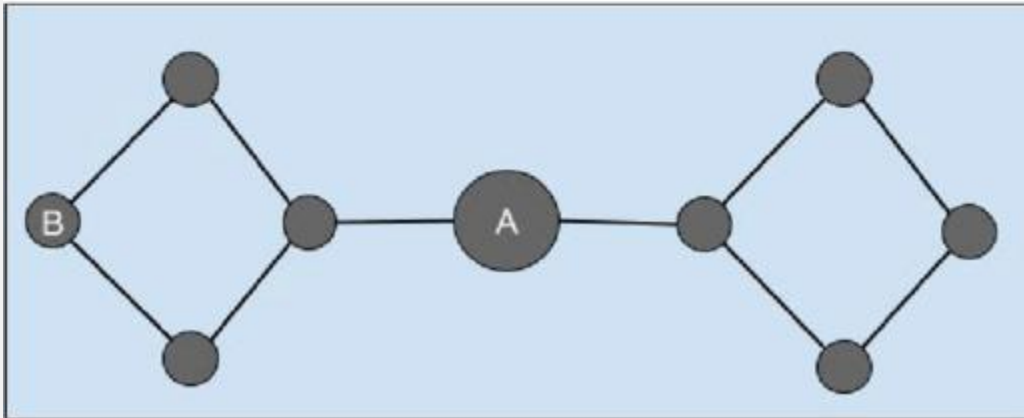
### In-degree Centrality

User A has a higher in-degree centrality than user B because user A has more followers than user B.



### Eigenvector Centrality

While users A and B both have the same in-degree centrality (two followers), user A has a higher Eigenvector centrality because the weight of the two followers is higher.



### Betweenness Centrality

User A has a higher betweenness centrality than user B. A message sent from user A will reach many more users in a shorter path compared to user B.

Source: [https://www.researchgate.net/figure/Pictorial-description-of-In-degree-Eigenvector-centrality-and-betweenness-centrality\\_fig1\\_313416055](https://www.researchgate.net/figure/Pictorial-description-of-In-degree-Eigenvector-centrality-and-betweenness-centrality_fig1_313416055)

# NETWORK GRAPH ANALYSIS

=== Network Graph Analysis Report ===

## Basic Graph Metrics:

nodes: 42  
edges: 41  
density: 0.047619047619047616  
avg\_clustering: 0.0  
avg\_shortest\_path: 2.085946573751452  
diameter: 4  
avg\_degree: 1.9523809523809523

## Protocol Distribution:

UDP: 575 packets (11.99%)  
TCP: 4222 packets (88.01%)

## Most Important Nodes:

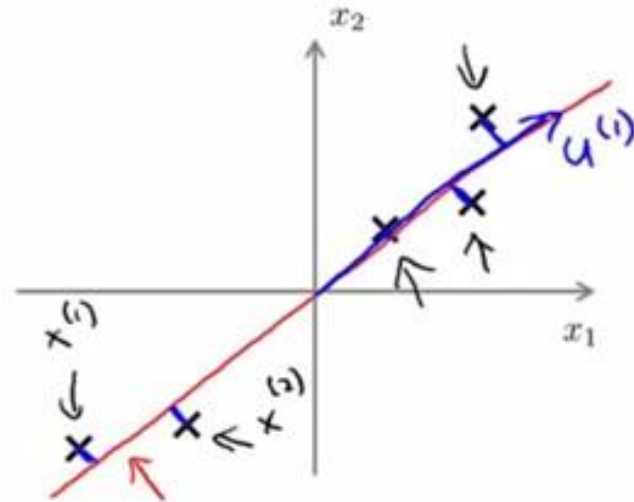
	degree centrality	betweenness centrality	eigenvector centrality	total_packets	importance_score
172.17.0.99	0.95122	0.996341	0.706847	4793.0	1198.913602
172.17.0.17	0.04878	0.095122	0.116203	1310.0	327.565026
46.254.34.201	0.02439	0.000000	0.113148	782.0	195.534384
79.124.78.197	0.02439	0.000000	0.113148	591.0	147.784384
23.45.119.144	0.02439	0.000000	0.113148	376.0	94.034384

# ONE HOT ENCODING

Source_IP_0.0.0.0	Source_IP_13.107.246.57	Source_IP_13.70.79.200	Source_IP_13.89.179.9	Source_IP_172.170.17	Source_IP_172.170.99	Source_IP_184.29.137.96	Source_IP_199.232.10.172	Source_IP_199.232.14.172	Source_IP_20.10.31.115	Source_IP_20.10.31.115
1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0

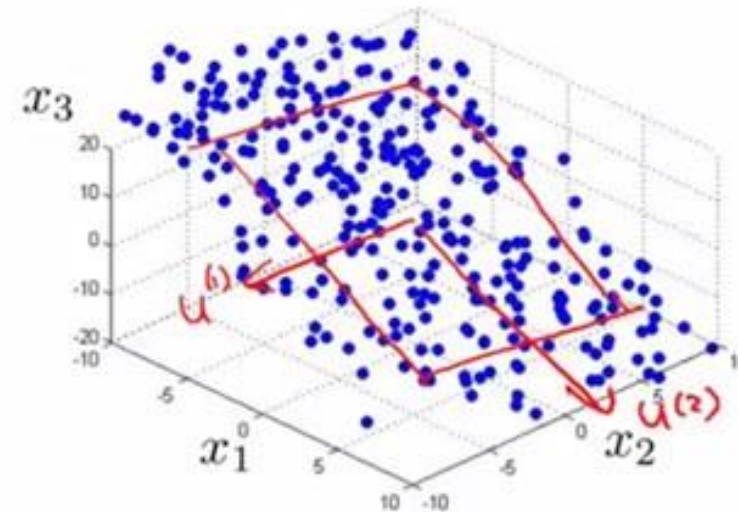
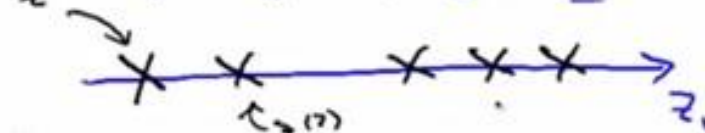
# DIMENSIONAL REDUCTION

## Principal Component Analysis (PCA) algorithm



Reduce data from 2D to 1D

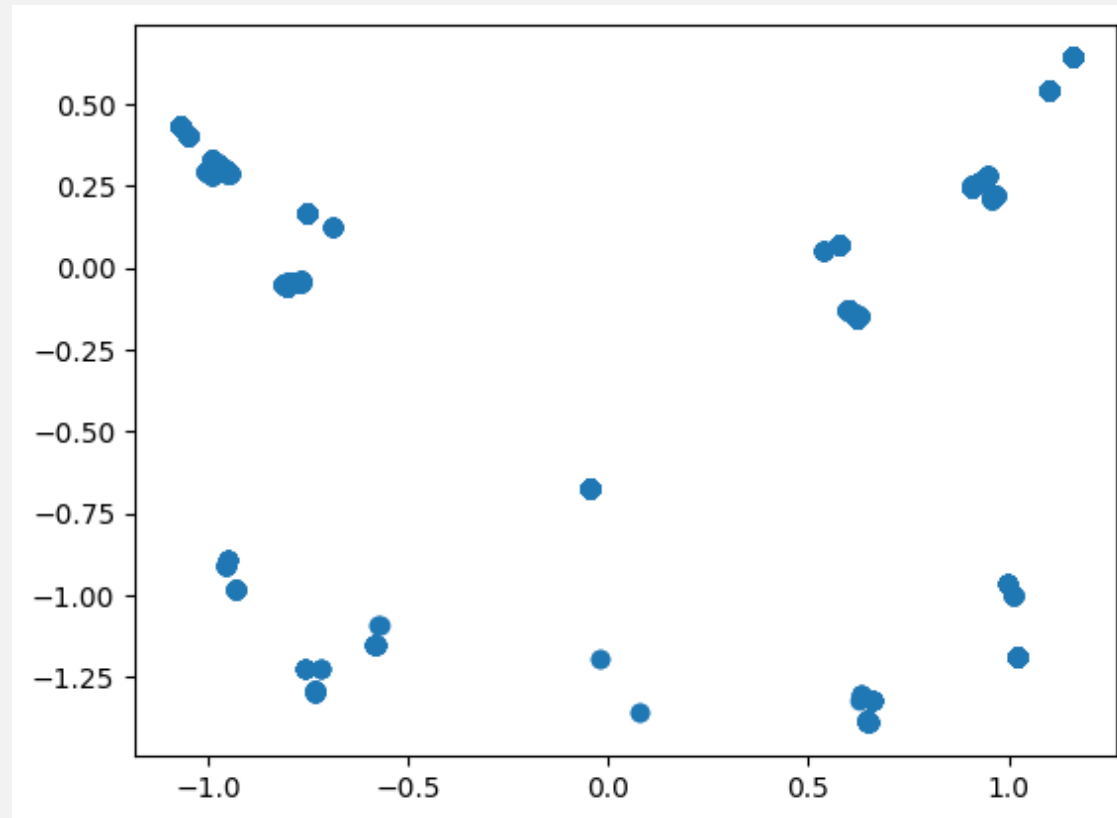
$$x^{(i)} \in \mathbb{R}^2 \rightarrow z^{(i)} \in \mathbb{R}$$



Reduce data from 3D to 2D



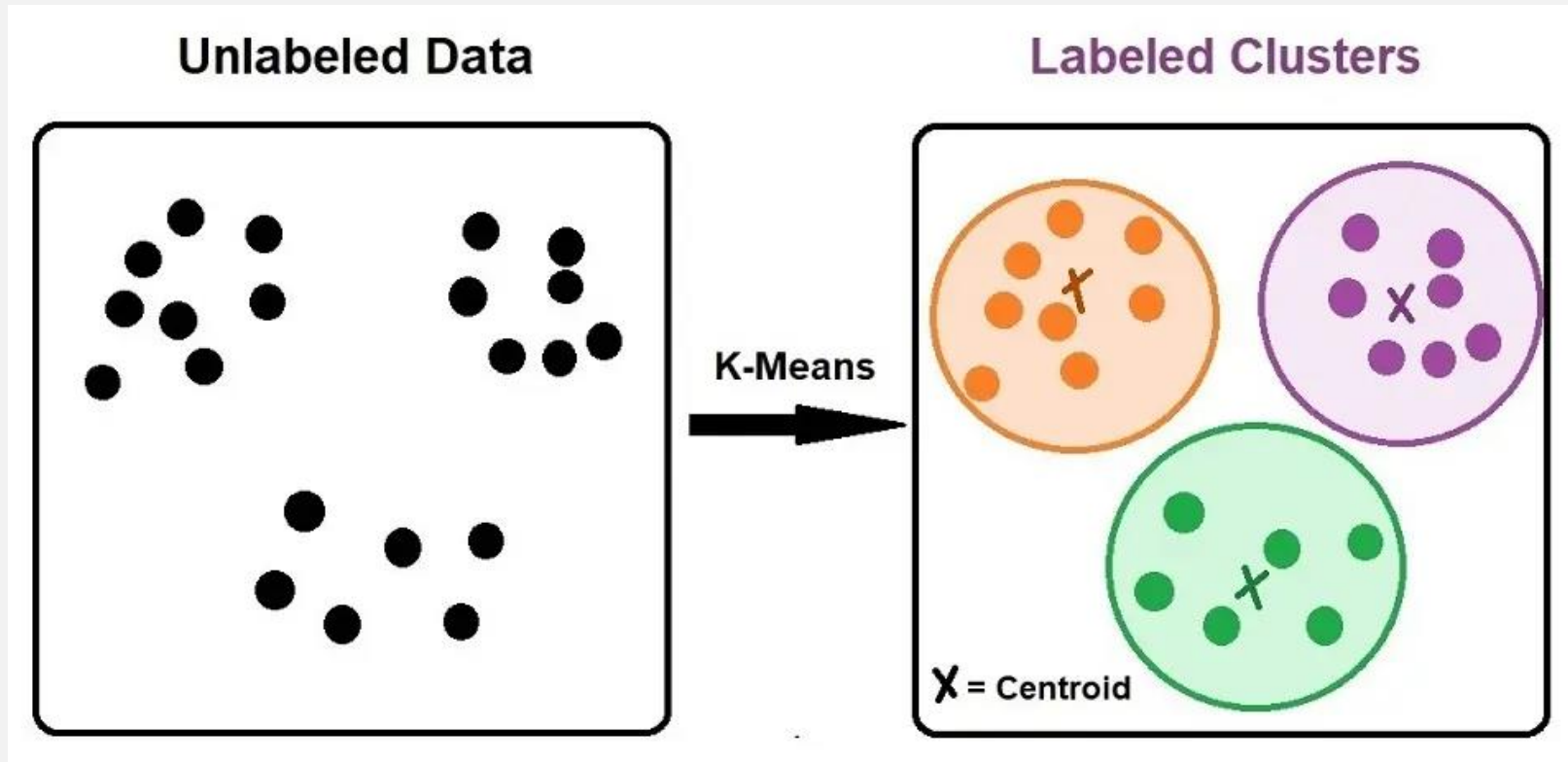
# DIMENSIONAL REDUCTION



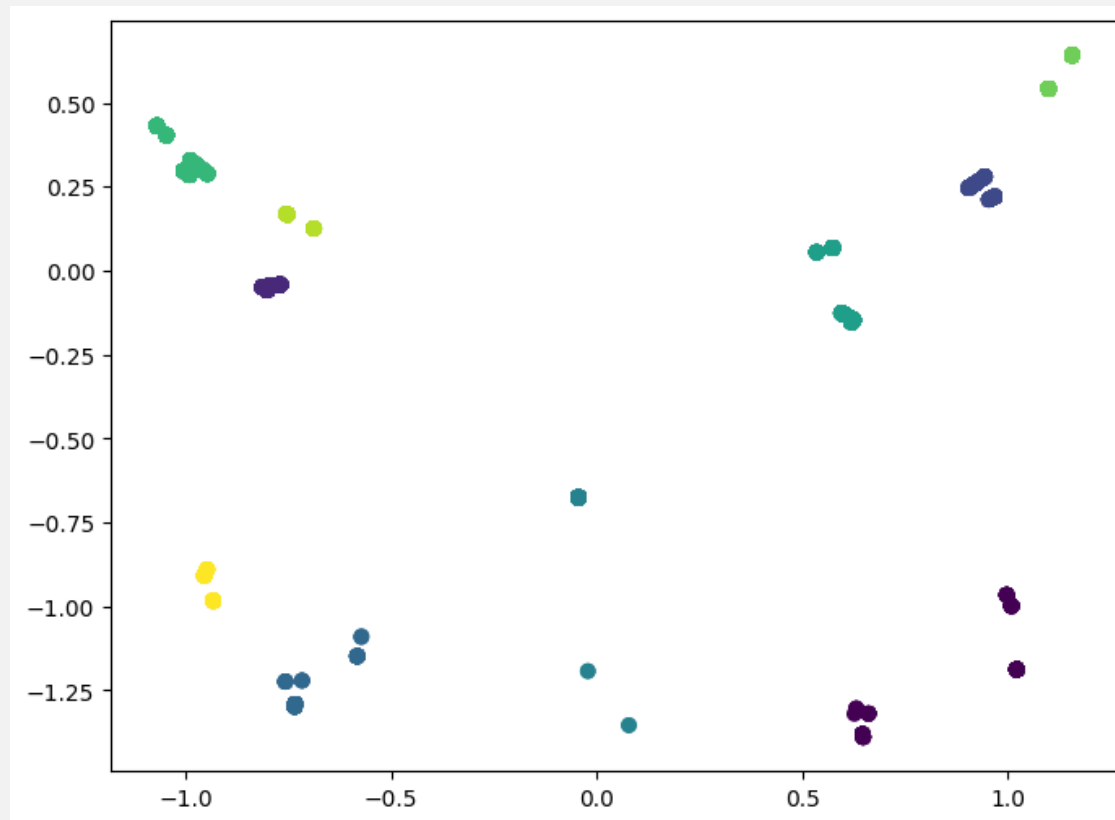
# DIMENSIONAL REDUCTION

[illegible]

# CLUSTERING



# CLUSTERING



# CLUSTERING

Count of Packet_Number	Column Labels <input type="button" value="v"/>										
Row Labels <input type="button" value="v"/>	0	1	2	3	4	5	6	7	8	9	Grand Total
40.119.249.228			22								22
40.126.28.12			20								20
40.126.28.22			11								11
46.254.34.201								504			504
52.109.0.142			13								13
52.109.0.91			16								16
52.113.194.132			73								73
79.124.78.197						261					261
(blank)					294						294
Grand Total	320	599	814	138	298	797	1144	504	364	113	5091

RECAP AND Q&A

# OBJECTIVES

- Understand the nature of log data
- Learn how to transform log data for analysis
- Apply various analysis techniques to find anomalies in log data
  - Time Series
  - Matching
  - Network Activity
  - Graph Network Analysis
  - Clustering

## FOR NEXT SESSION

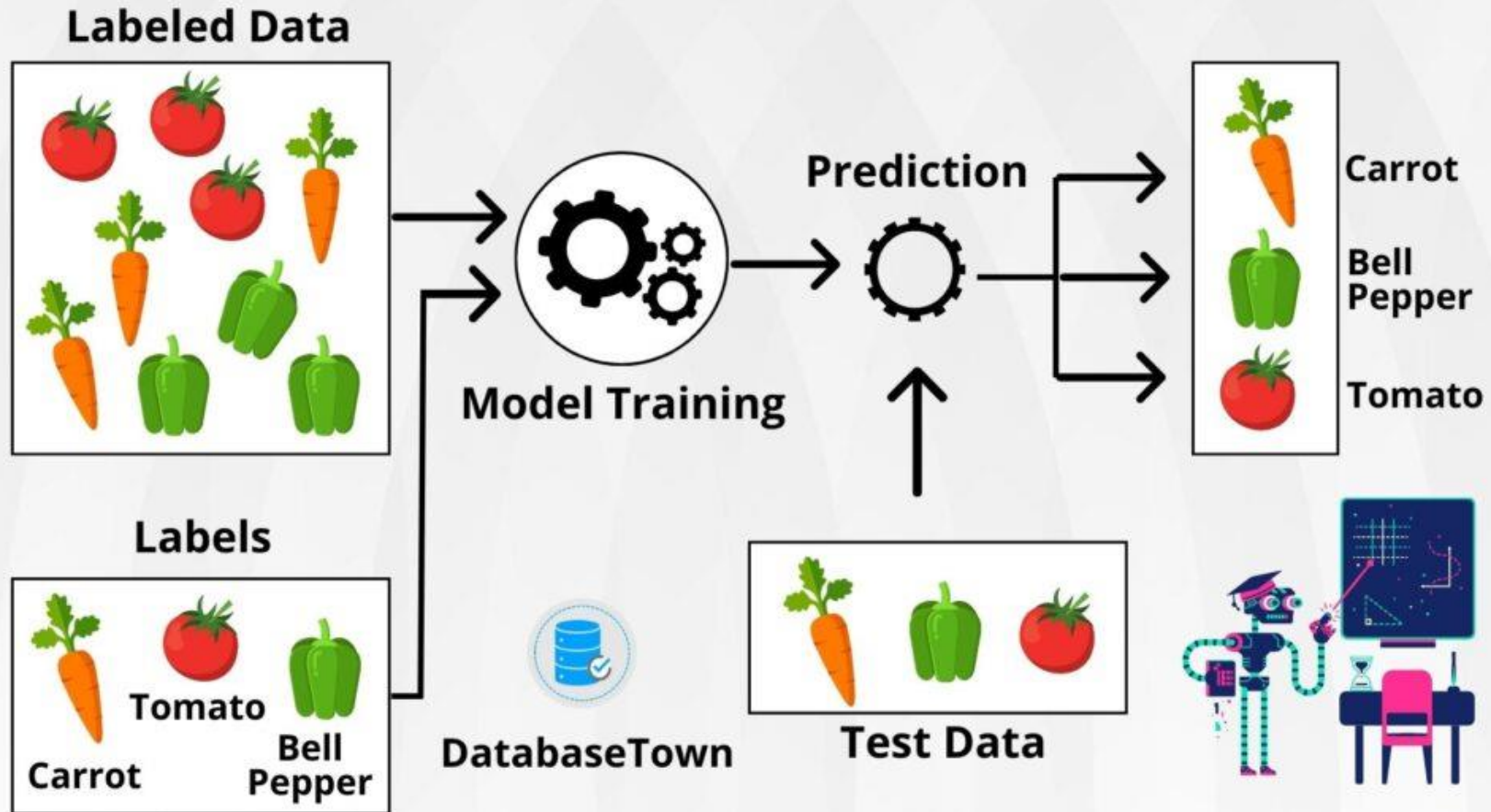
- Perform similar analysis for Exercise 2 and 3
- Identify any suspicious anomalies
- Add some insight, based on possible scenarios
- Send your assignments to: [docligot@cirrolytix.com](mailto:docligot@cirrolytix.com)



# SUPERVISED LEARNING

# SUPERVISED LEARNING

Supervised machine learning is a branch of artificial intelligence that focuses on training models to make predictions or decisions based on labeled training data.



## PREDICTED PROBABILITIES (SCORING)

0	1	2	3	4	5	6	predicted cluster
0.995197	0.000184	0.002055	0.001561	0.000374	0.000442	0.000187	0
0.993225	0.000203	0.003103	0.002276	0.000446	0.000539	0.000208	0
0.002374	0.000163	0.99554	0.001095	0.000307	0.000357	0.000165	2
0.002077	0.9848	0.002156	0.001384	0.00505	0.001709	0.002824	1
0.000679	0.006335	0.00064	0.001137	0.001627	0.000636	0.988946	6
0.995197	0.000184	0.002055	0.001561	0.000374	0.000442	0.000187	0
0.002374	0.000163	0.99554	0.001095	0.000307	0.000357	0.000165	2
0.995197	0.000184	0.002055	0.001561	0.000374	0.000442	0.000187	0
0.995197	0.000184	0.002055	0.001561	0.000374	0.000442	0.000187	0
0.002374	0.000163	0.99554	0.001095	0.000307	0.000357	0.000165	2
0.004628	0.000568	0.002491	0.982691	0.003388	0.005641	0.000594	3
0.000515	0.002049	0.000491	0.000753	0.000958	0.000487	0.994748	6

# ANALYZING LOGS DATA WITH AI

MCS 2025 Cyber 242

Lecturer: Dominic Ligot