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

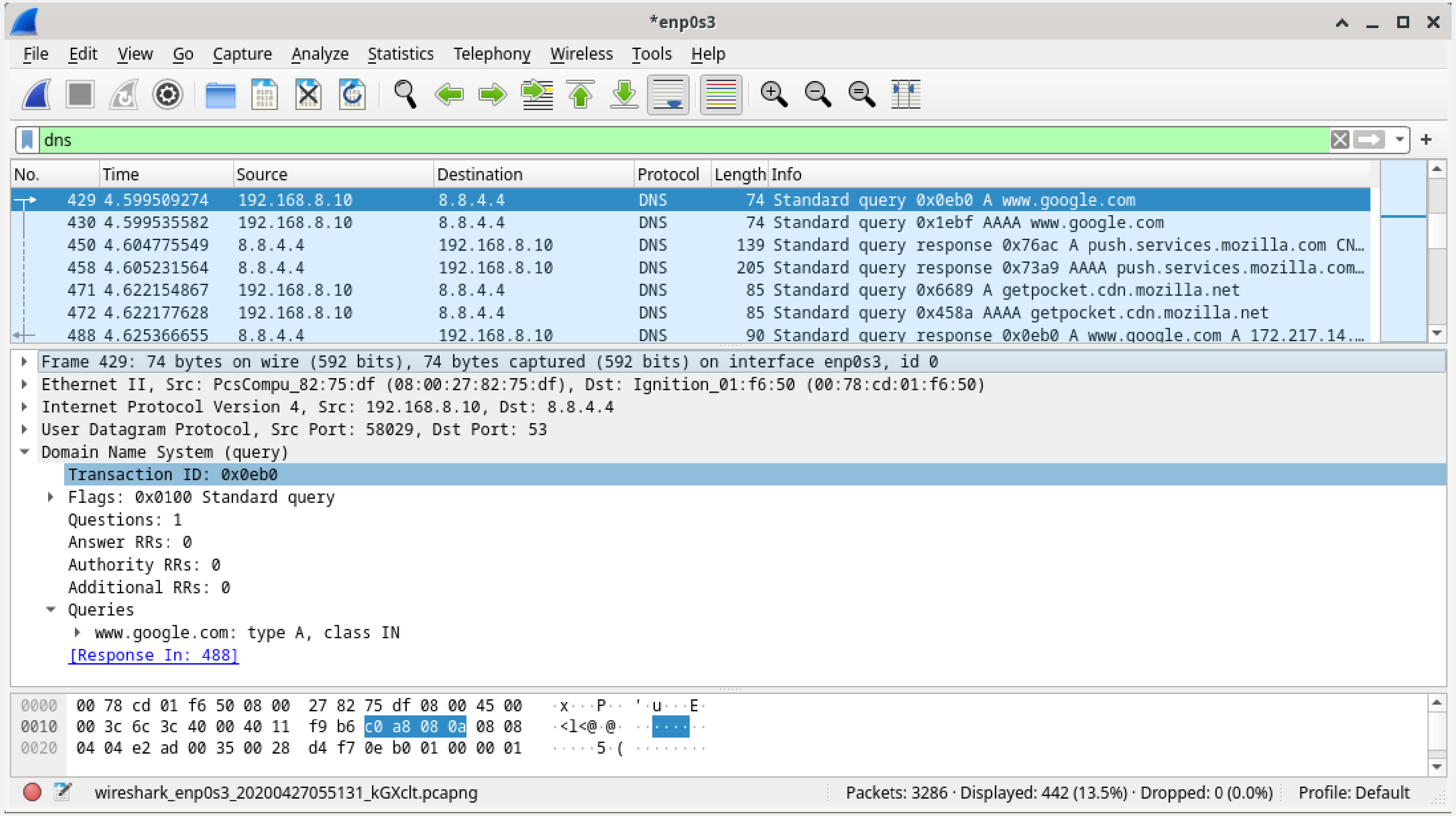
UK Department of Science, Innovation, and Technology
Member, International Expert Panel on Advanced AI Safety

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)



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

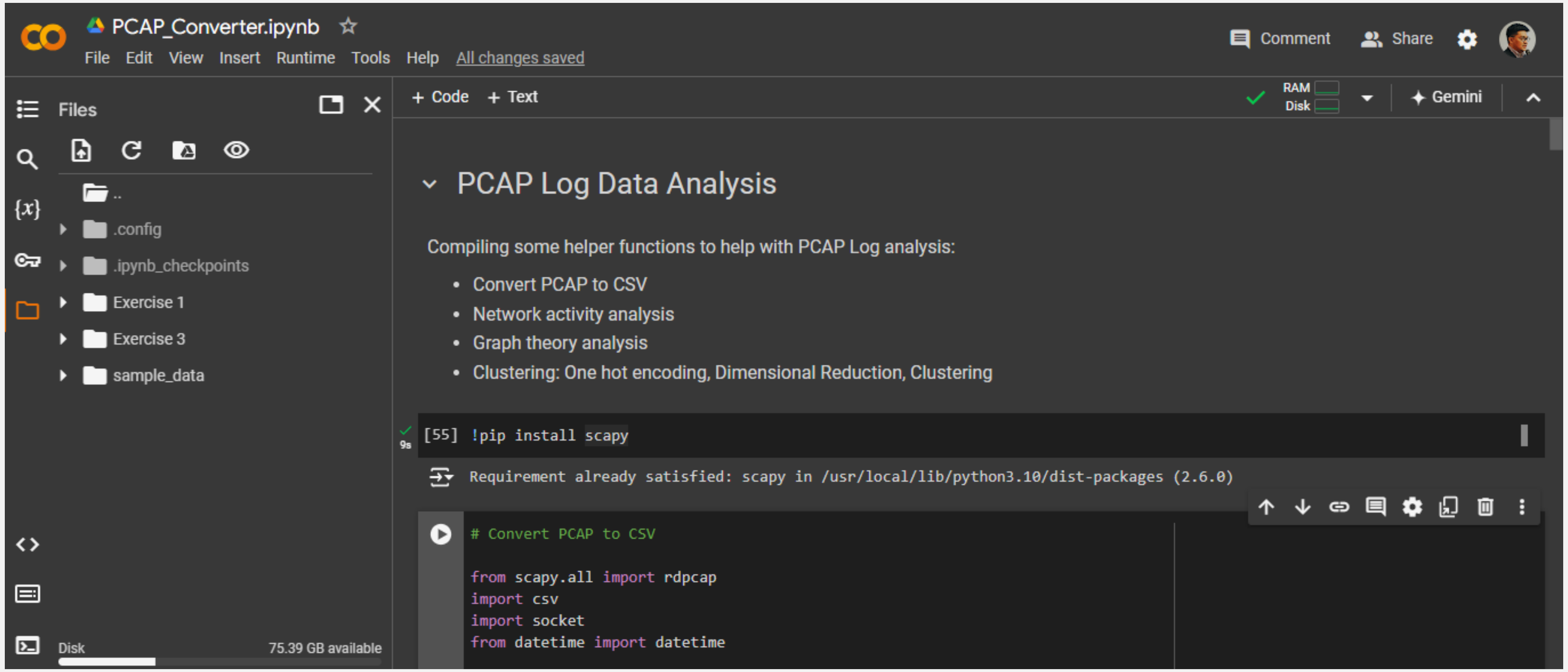
- Data analysis does not replace domain knowledge – but complements it
- Data analysis is better at generating questions – but it remains the duty of the security analyst to put a case together
- Data analysis is about pattern recognition – look for similarities and anomalies as clues to formulating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

SETTING UP GOOGLE COLLAB

GITHUB REPOSITORY

- Access/clone/fork from here:
  - https://github.com/docligot.com/pcap_labs

PCAP CONVERTER NOTEBOOK



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**

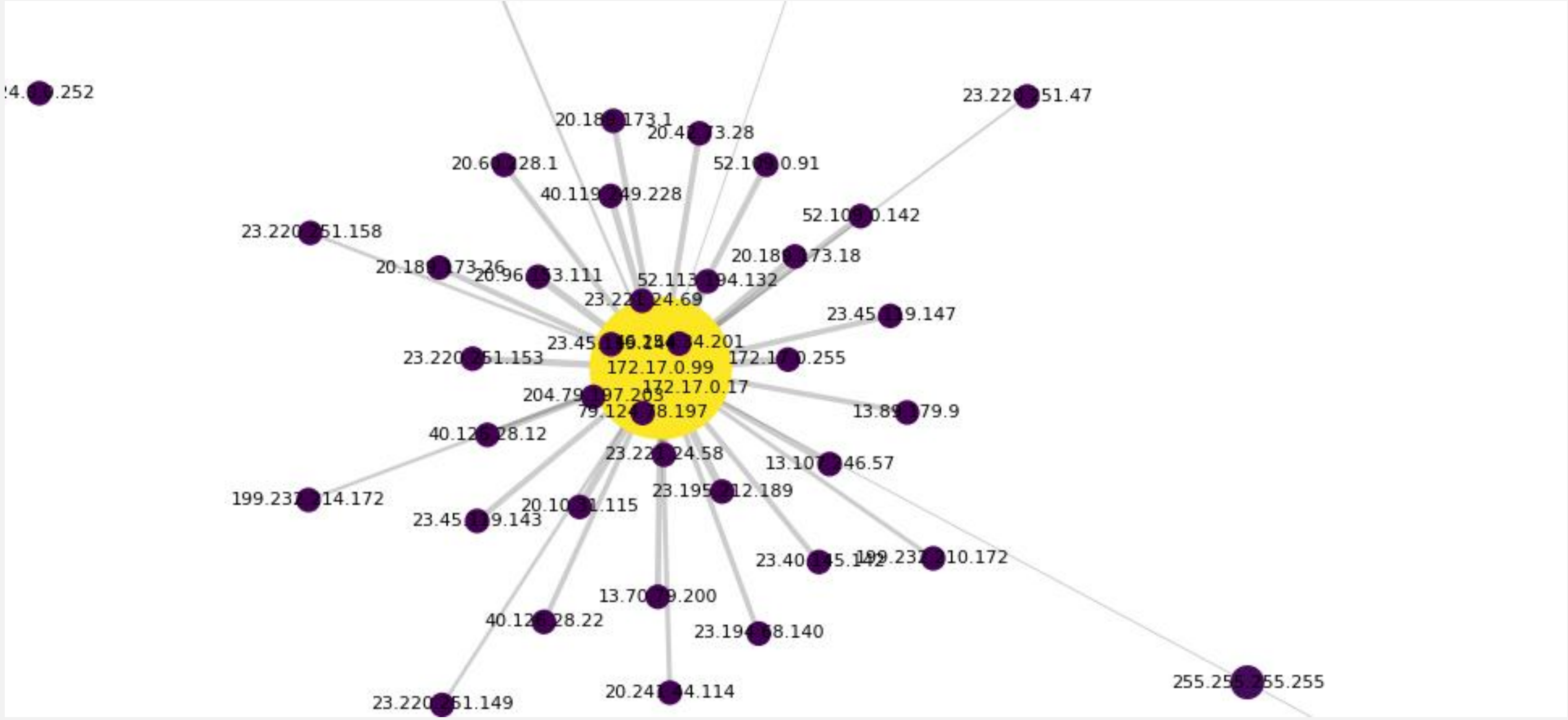
IP TO IP - EYEBALLING

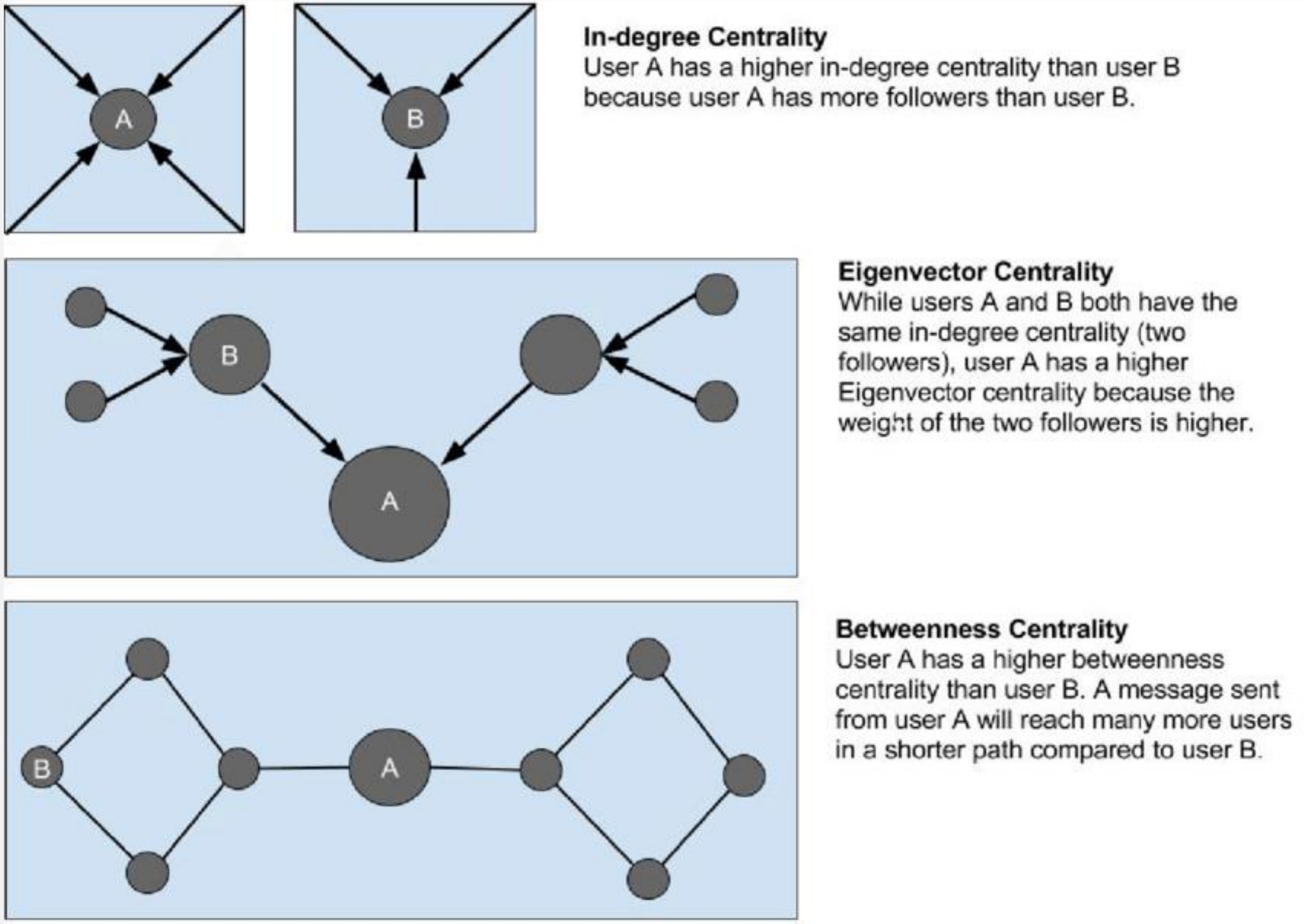
Count of Packet_Number	Column Labels			
Row Labels	172.17.0.99	255.255.255.255	(blank)	Grand Total
23.45.119.147	13			13
40.119.249.228	22			22
40.126.28.12	20			20
40.126.28.22	11			11
46.254.34.201	504			504
52.109.0.142	13			13
52.109.0.91	16			16
52.113.194.132	73			73
79.124.78.197	261			261
(blank)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

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





Source: 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:
```

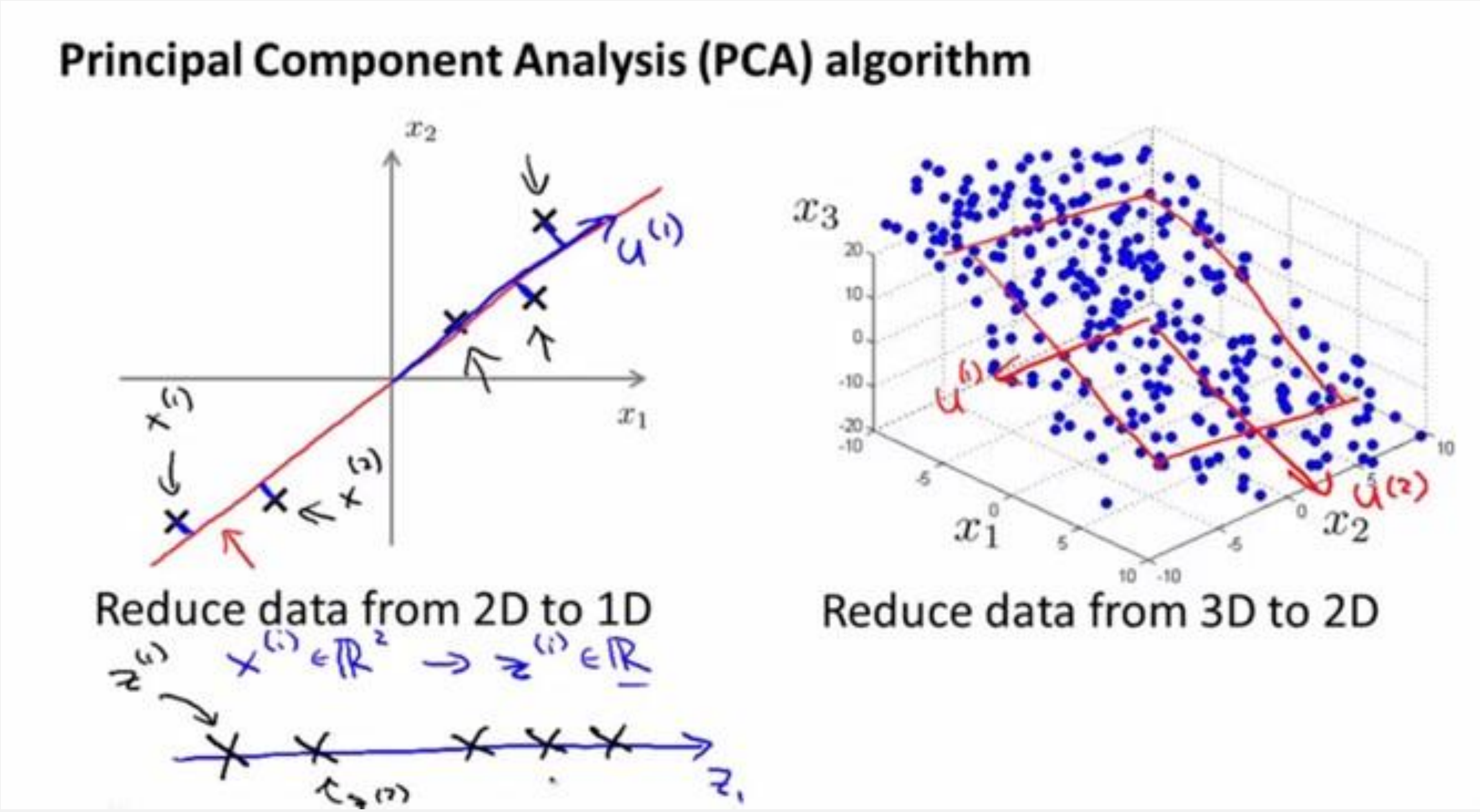
	degree_centrality	betweenness_centrality	eigenvector_centrality	total_packets	importance_score
172.17.0.99	0.95122	0.996341	0.706847	4793.0	1198.913602
172.17.0.17	0.04878	0.095122	0.116203	1310.0	327.565026
46.254.34.201	0.02439	0.000000	0.113148	782.0	195.534384
79.124.78.197	0.02439	0.000000	0.113148	591.0	147.784384
23.45.119.144	0.02439	0.000000	0.113148	376.0	94.034384

ONE HOT ENCODING

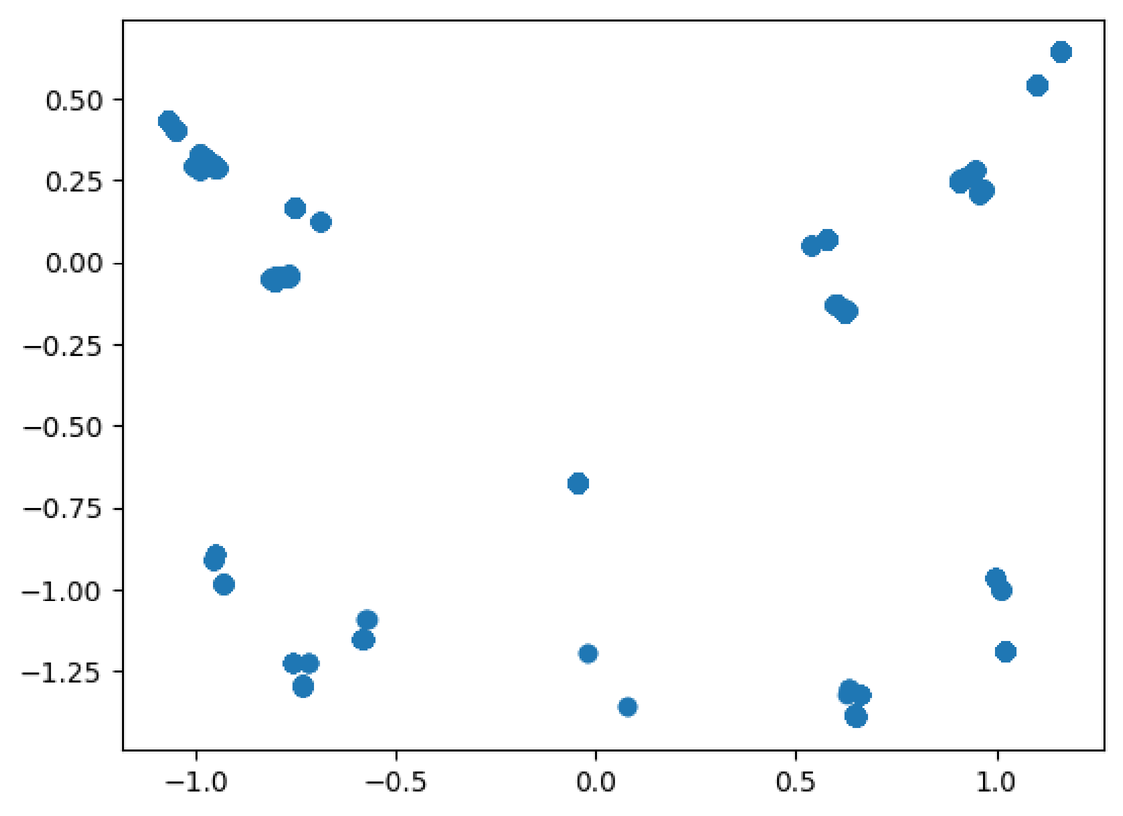
Source_IP_0.0.0.0	Source_IP_13.107.246.57	Source_IP_13.70.79.200	Source_IP_13.89.179.9	Source_IP_172.170.17	Source_IP_172.170.99	Source_IP_184.29.137.96	Source_IP_199.232.210.172	Source_IP_199.232.214.172	Source_IP_20.10.31.115	Source_IP_2
1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	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

Source: https://medium.com/analytics-vidhya/principal-component-analysis-pca-8a0fcba2e30c

DIMENSIONAL REDUCTION



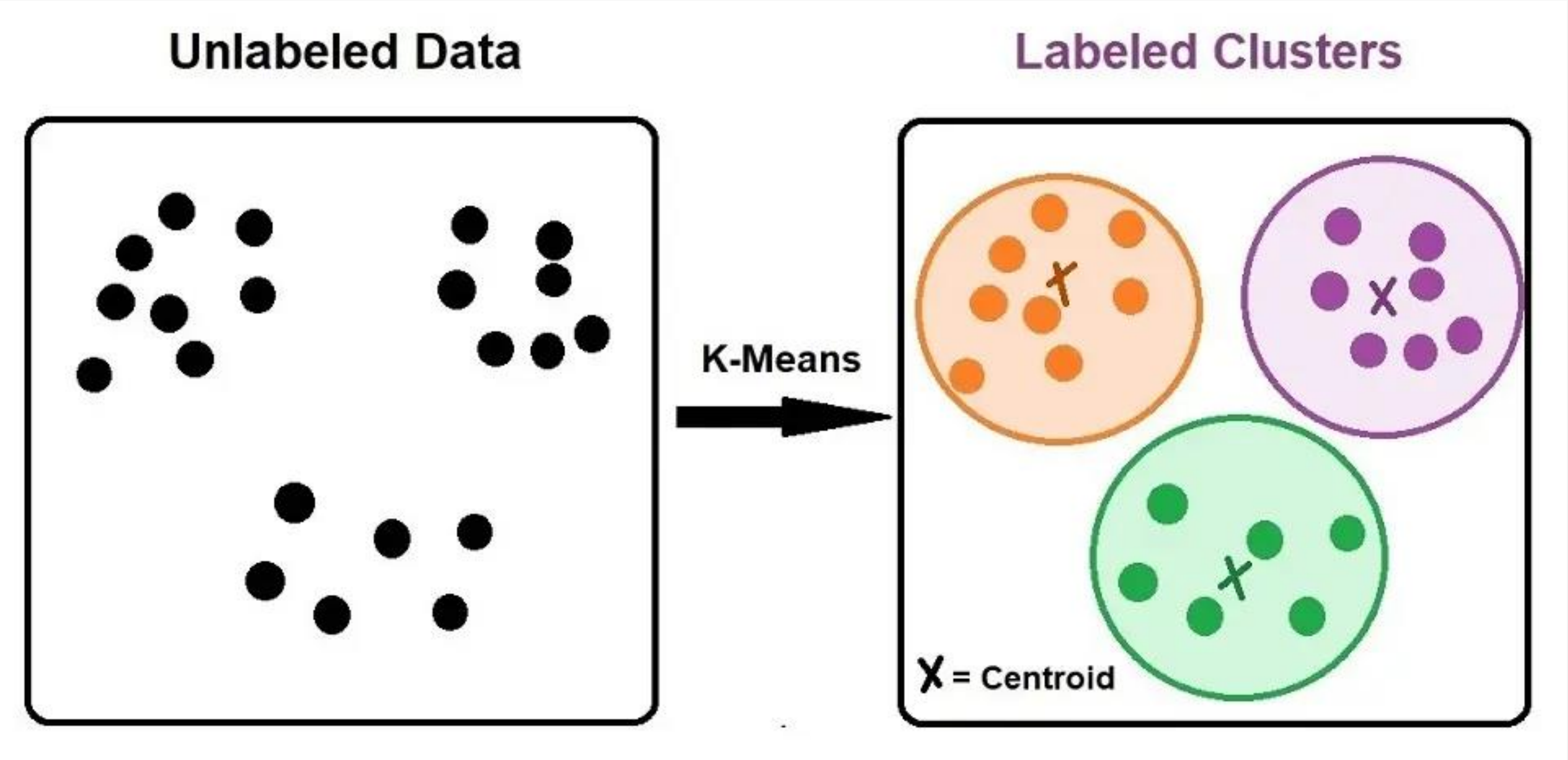
DIMENSIONAL REDUCTION



DIMENSIONAL REDUCTION

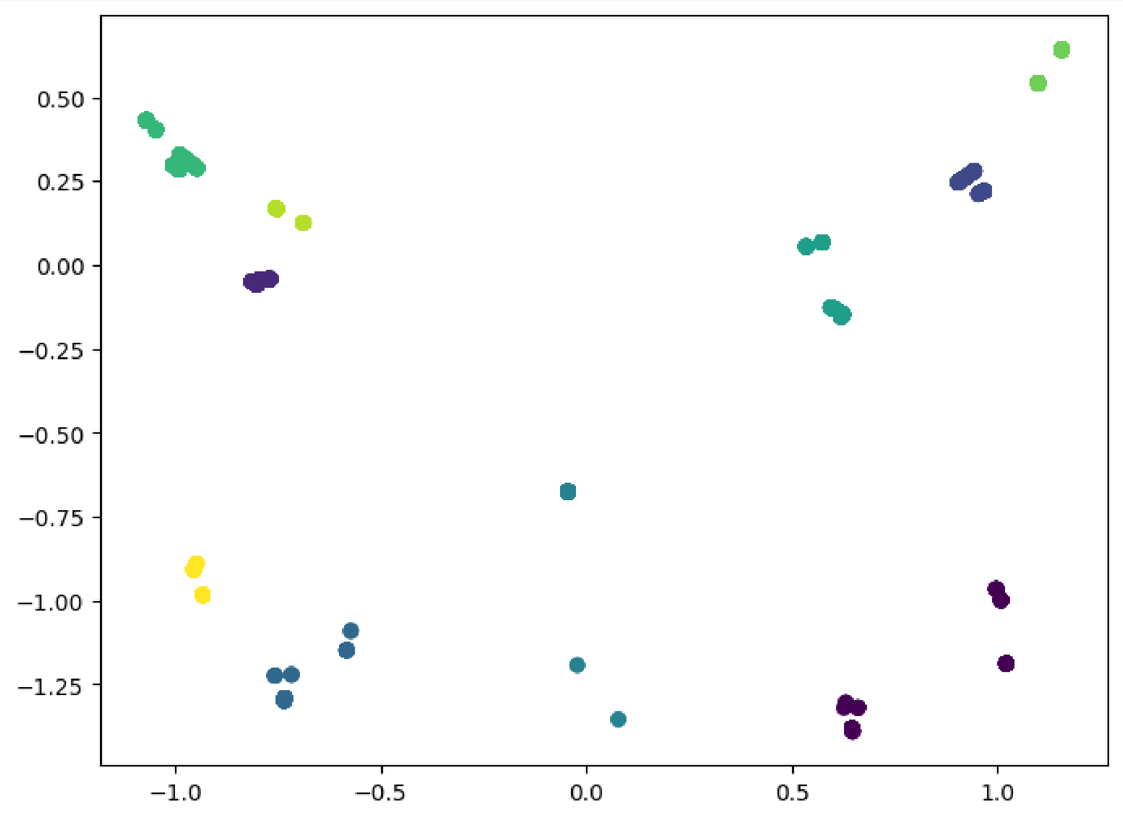
Destinatic	Destinatic	Destinatic	Destinatic	Destinatic	Destinatic	Destinatic	Destinatic	Destinatic	Destinatic	x_pos	y_pos
0	0	0	0	0	0	0	0	0	0	-0.02114	-1.19303
0	0	0	0	0	0	0	0	0	0	0.078709	-1.35421
0	0	0	0	0	0	0	0	0	0	-0.02114	-1.19303
0	0	0	0	0	0	0	0	0	0	0.078709	-1.35421
0	0	0	0	0	0	0	0	0	0	-0.04392	-0.6761
0	0	0	0	0	0	0	0	0	0	-0.04392	-0.6761
0	0	0	0	0	0	0	0	0	0	-0.73347	-1.29116
0	0	0	0	0	0	0	0	0	0	-0.57256	-1.09145
0	0	0	0	0	0	0	0	0	0	0.647581	-1.38131
0	0	0	0	0	0	0	0	0	0	-0.57256	-1.09145
0	0	0	0	0	0	0	0	0	0	-0.57211	-1.08896
0	0	0	0	0	0	0	0	0	0	-0.73366	-1.29214
0	0	0	0	0	0	0	0	0	0	-0.73366	-1.29214
0	0	0	0	0	0	0	0	0	0	0.647753	-1.38236
0	0	0	0	0	0	0	0	0	0	0.647753	-1.38236

CLUSTERING



Source: https://www.ejable.com/tech-corner/ai-machine-learning-and-deep-learning/k-means-clustering/

CLUSTERING



CLUSTERING

Count of Packet_Number	Column Labels										
Row Labels	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

ANALYZING LOGS DATA WITH AI

MCS 2025 Cyber 242

Lecturer: Dominic Ligot