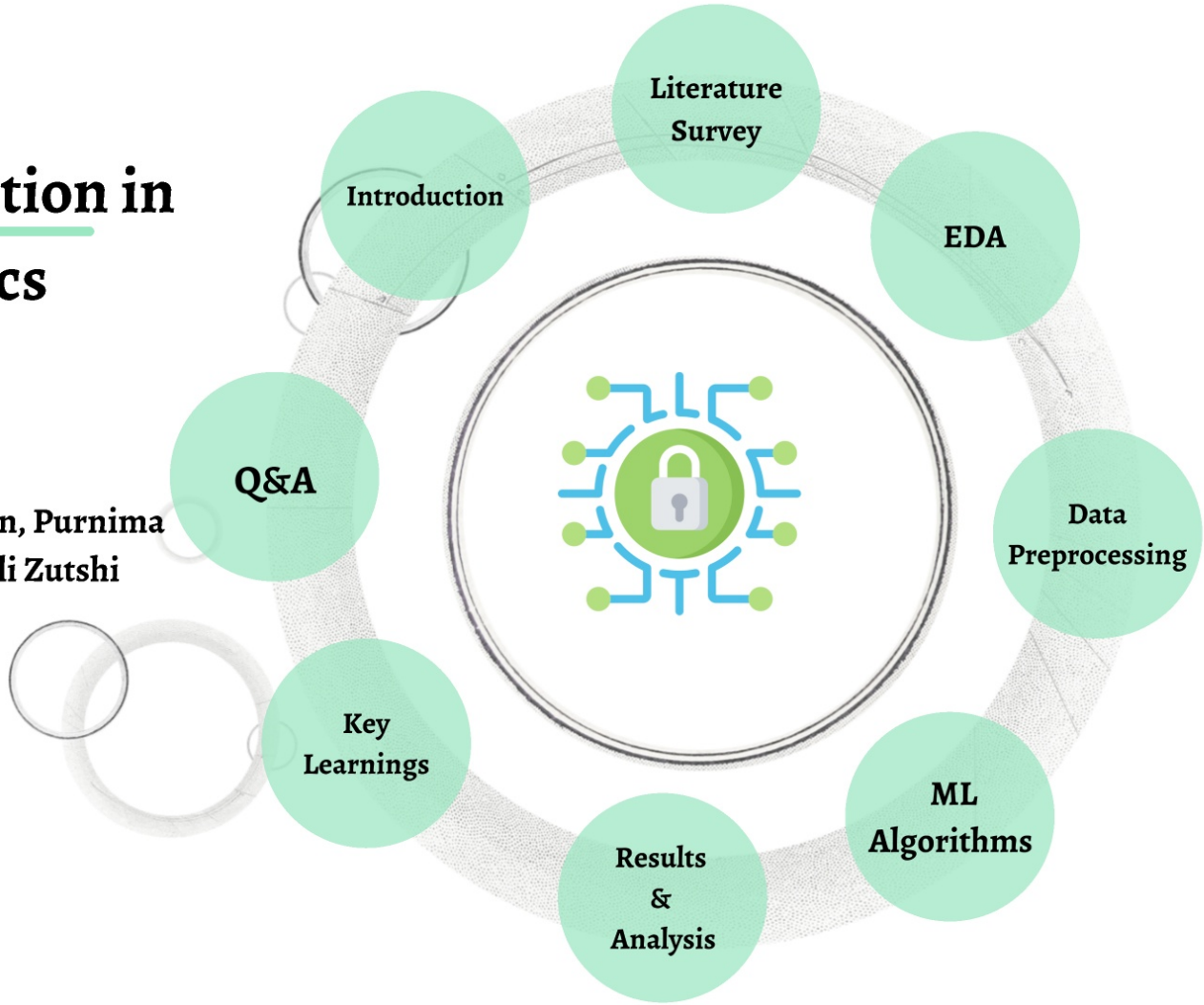
Cyber Attack Detection in Cloud Forensics

by

Yasaman Emami, Shamama Afnan, Purnima Bhukya, Poojitha Katta, Deepali Zutshi



Problem Definition

- Organizations are increasingly moving to the cloud for new initiatives or replacing current systems.
- Advantages: scalability, accessibility, cost-effectiveness, efficiency, and collaboration.
- Disadvantages: downtime, security & privacy risks, vulnerability to attack.
- Aim: use machine learning to detect attacks within a network by classifying instances as 'Normal' or 'Attack'.

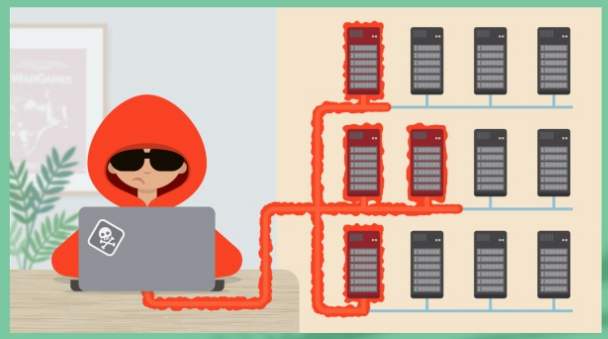
Motivation

Data Set

Motivation

- Offered services
- Rate of growing data

- Need of Intrusion Detection System
- How?

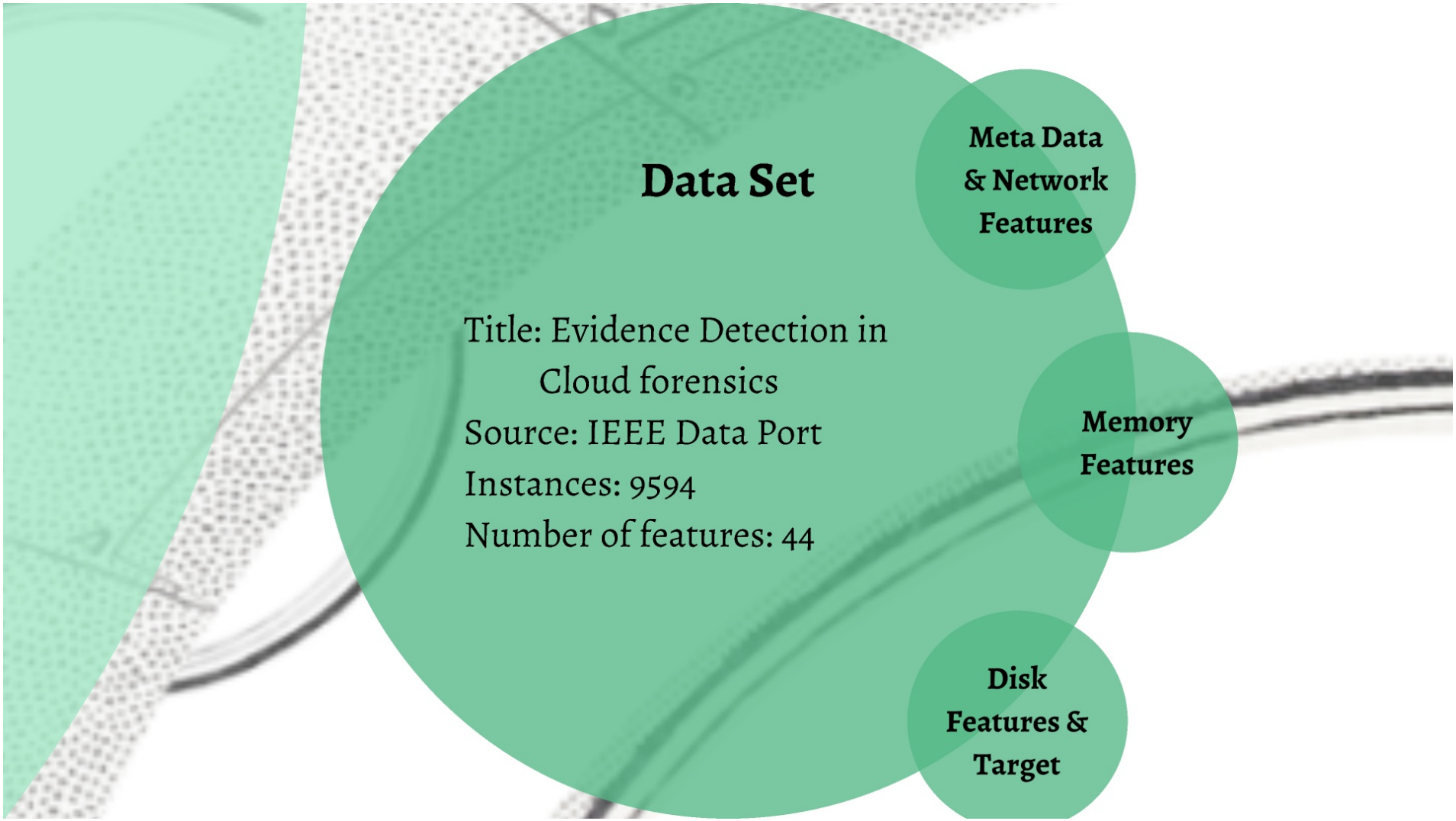


Problem Definition

- Organizations are increasingly moving to the cloud for new initiatives or replacing current systems.
- Advantages: scalability, accessibility, cost-effectiveness, efficiency, and collaboration.
- Disadvantages: downtime, security & privacy risks, vulnerability to attack.
- Aim: use machine learning to detect attacks within a network by classifying instances as 'Normal' or 'Attack'.

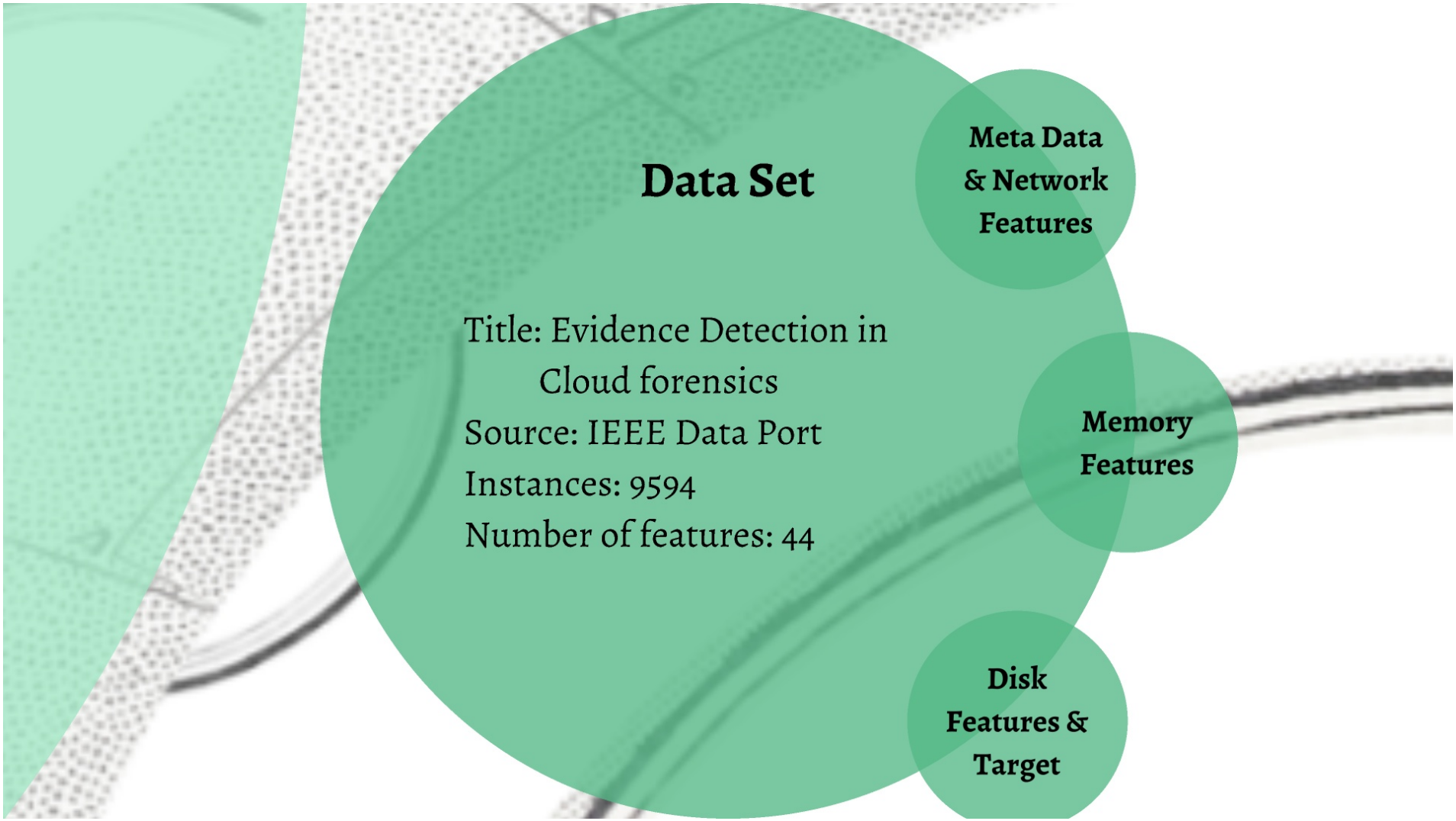
Motivation

Data Set


Data Set
Title: Evidence Detection in Cloud forensics
Source: IEEE Data Port
Instances: 9594
Number of features: 44
Meta Data & Network Features
Memory Features
Disk Features & Target

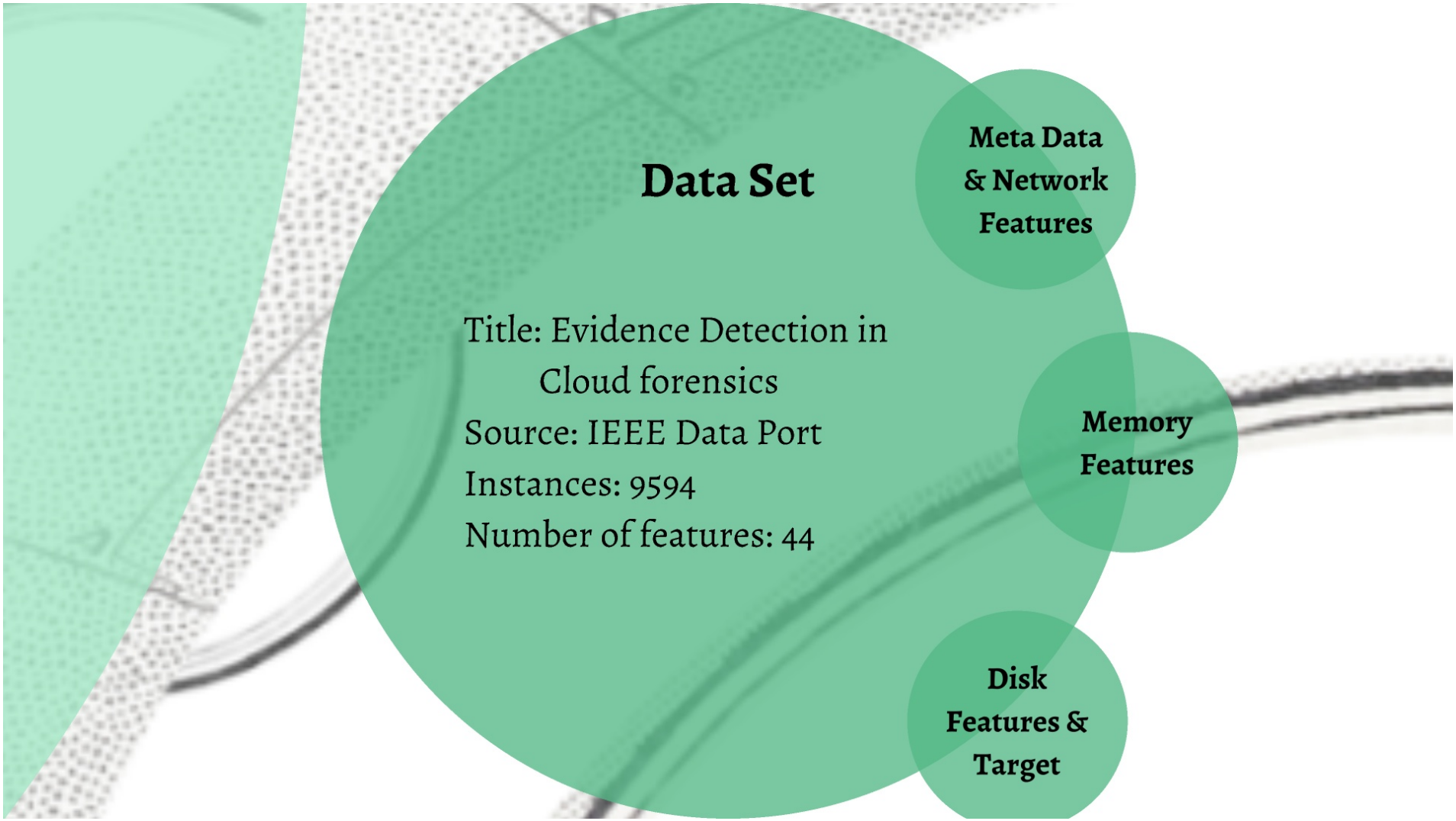
Meta data & Network Features

Sr No	Category	Feature	Description
1	Meta-data	LAST_POLL	epoch timestamp
2		VMID	The ID of the VM
3		UUID	unique identifier of the domain
4		dom	domain name
5	Network	rxbytes_slope	Rate of received bytes from the network
6		rxpackets_slope	Rate of received packets from the network
7		rxerrors_slope	Rate of the number of receive errors from the network
8		rxdrops_slope	Rate of the number of received packets dropped from the network
9		txbytes_slope	Rate of transmitted bytes from the network
10		txpackets_slope	Rate of transmitted packets from the network
11		txerrors_slope	Rate of the number of transmission errors from the network
12		txdrops_slope	Rate of the number of transmitted packets dropped from the network


Data Set
Title: Evidence Detection in Cloud forensics
Source: IEEE Data Port
Instances: 9594
Number of features: 44
Meta Data & Network Features
Memory Features
Disk Features & Target

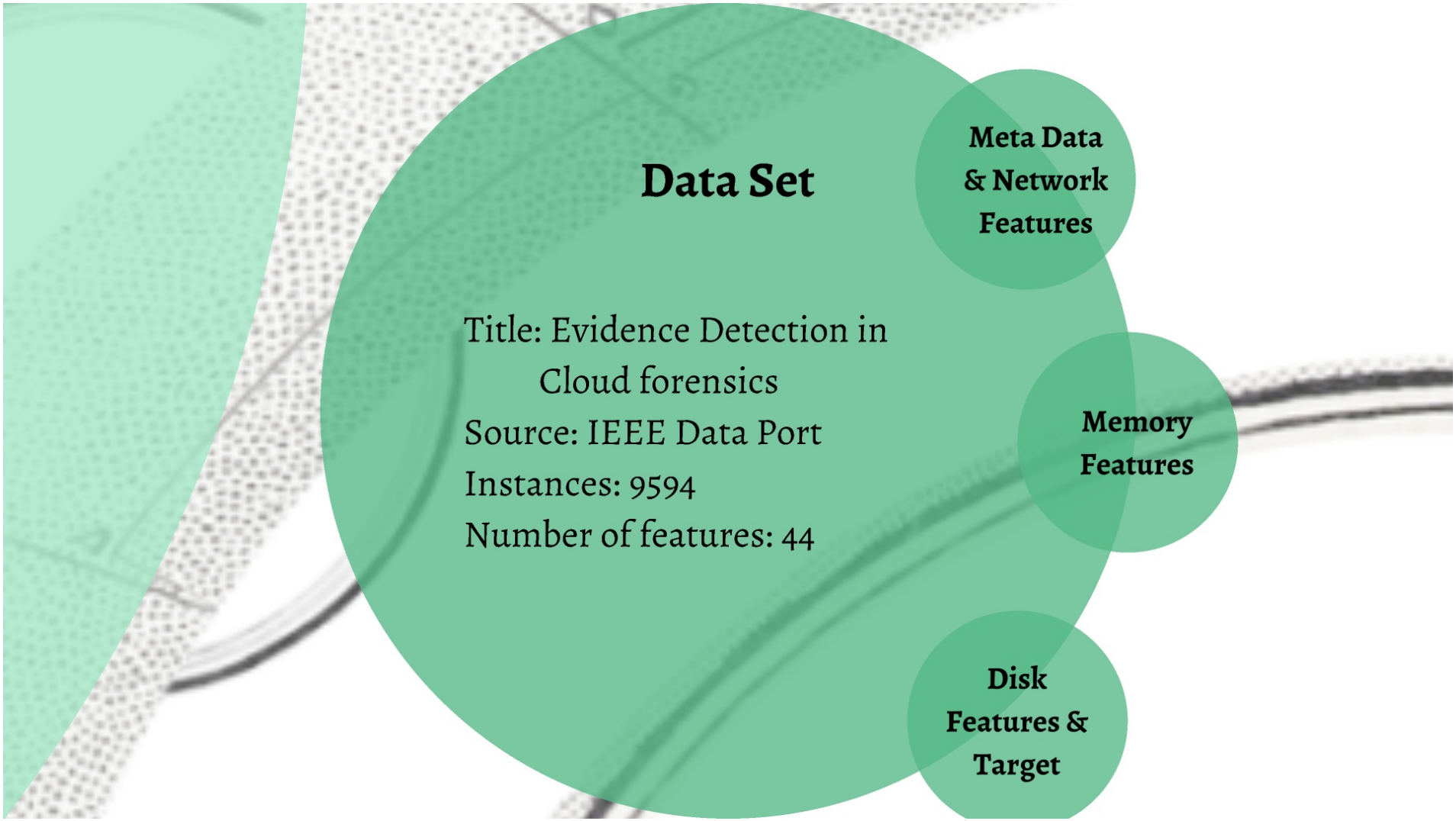
Feature Description

13	**Memory**	timecpu_slope	Rate of time spent by vCPU threads executing guest code
14		timesys_slope	Rate of time spent in kernel space
15		timeusr_slope	Rate of time spent in userspace
16		state_slope	Rate of running state
17		memmax_slope	Rate of maximum memory in kilobytes
18		mem_slope	Rate of memory used in kilobytes
19		cpus_slope	Rate of the number of virtual CPUs chaged
20		cputime_slope	Rate of CPU time used in nanoseconds
21		memactual_slope	Rate of Current balloon value (in KiB)
22		memswap_in_slope	Rate of The amount of data read from swap space (in KiB)
23		memswap_out_slope	Rate of The amount of memory written out to swap space (in KiB)
24		memmajor_fault_slope	Rate of The number of page faults where disk IO was required
25		memminor_fault_slope	Rate of The number of other page faults
26		memunused_slope	Rate of The amount of memory left unused by the system (in KiB)
27		memavailable_slope	Rate of The amount of usable memory as seen by the domain (in KiB)
28		memusable_slope	Rate of The amount of memory that can be reclaimed by balloon without causing host swapping (in KiB)
29		memlast_update_slope	Rate of The timestamp of the last update of statistics (in seconds)
30		memdisk_cache_slope	Rate of The amount of memory that can be reclaimed without additional I/O, typically disk caches (in KiB)
31		memhugetlb_pgalloc_slope	Rate of The number of successful huge page allocations initiated from within the domain
32		memhugetlb_pgfail_slope	Rate of The number of failed huge page allocations initiated from within the domain
33		memrss_slope	Rate of Resident Set Size of the running domain's process (in KiB)


Data Set
Title: Evidence Detection in Cloud forensics
Source: IEEE Data Port
Instances: 9594
Number of features: 44
Meta Data & Network Features
Memory Features
Disk Features & Target

Feature Description

34	**Disk**	vdard_req_slope	Rate of the number of reading requests on the vda block device
35		vdard_bytes_slope	Rate of the number of reading bytes on the vda block device
36		vdawr_reqs_slope	Rate of the number of write requests on the vda block device
37		vdawr_bytes_slope	Rate of the number of write requests on vda the block device
38		vdaerror_slope	Rate of the number of errors in the vda block device
39		hdard_req_slope	Rate of the number of read requests on the hda block device
40		hdard_bytes_slope	Rate of the number of read bytes on the had block device
41		hdawr_reqs_slope	Rate of the number of write requests on the hda block device
42		hdawr_bytes_slope	Rate of the number of write bytes on the hda block device
43		hdaerror_slope	Rate of the number of errors in the hda block device
44	TARGET	Status	Attack/Normal


Data Set
Title: Evidence Detection in Cloud forensics
Source: IEEE Data Port
Instances: 9594
Number of features: 44
Meta Data & Network Features
Memory Features
Disk Features & Target

Problem Definition

- Organizations are increasingly moving to the cloud for new initiatives or replacing current systems.
- Advantages: scalability, accessibility, cost-effectiveness, efficiency, and collaboration.
- Disadvantages: downtime, security & privacy risks, vulnerability to attack.
- Aim: use machine learning to detect attacks within a network by classifying instances as 'Normal' or 'Attack'.

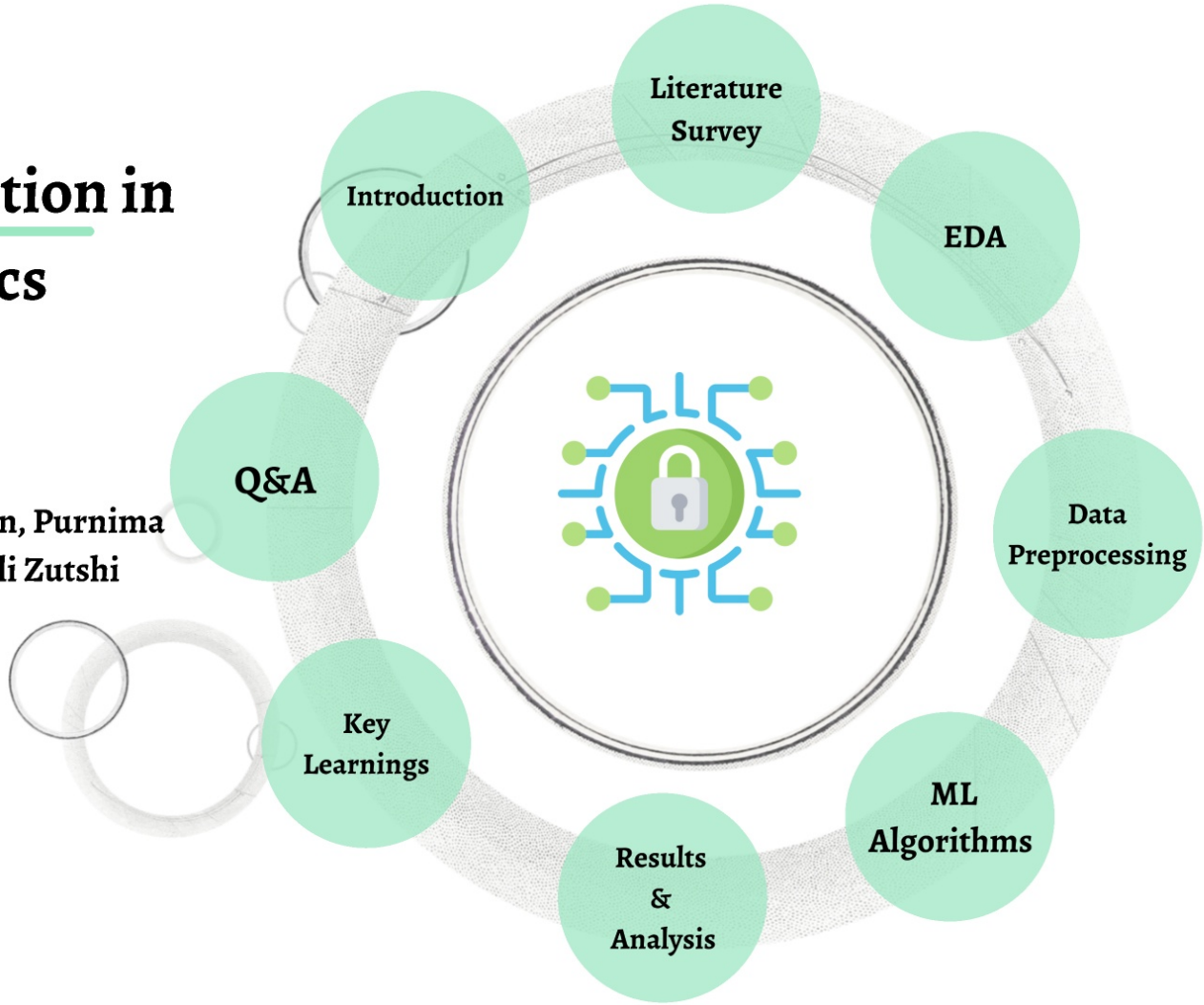
Motivation

Data Set

Cyber Attack Detection in Cloud Forensics

by

Yasaman Emami, Shamama Afnan, Purnima Bhukya, Poojitha Katta, Deepali Zutshi



Review

- Traditional Machine Learning Approach
- Deep Learning Approach

Traditional Machine Learning

Deep Learning

Traditional ML

Algorithms: SVM, DT, RF, KNN, Logistic Regression, XGBoost, NB, Hybrid models, Genetic Algorithms with KNN and SVM

Accuracy: Most of the research works achieved very high accuracy with traditional ML. SVM 99.81%, KNN 99.18 %, DT 99.92 % [1]

Feature Selection Techniques:

(1) Tree-based Classifiers

(2) Pearson Correlation (Assumption **:** the features with high correlation with class labels are are most suitable for intrusion detection.[2])

(3) PCA.

Imbalanced Data : Synthetic minority oversampling techniques are widely used in intrusion detection.

Evaluation: accuracy, precision, recall and F1 score, cross-validation and split-validation [3]

Review

- Traditional Machine Learning Approach
- Deep Learning Approach

Traditional Machine Learning

Deep Learning

Deep Learning

Algorithms: CNN, RNN, RBM, DBM, DBN and deep autoencoders and presented a comparative study of deep discriminative and generative models

High Accuracy:

- Deep discriminative model CNN of 97.28%
- Generative model DA provided 98.18% [4]

Training Time: In Deep Learning techniques, training time used as a performance metrics due to high accuracy of models

Review

- Traditional Machine Learning Approach
- Deep Learning Approach

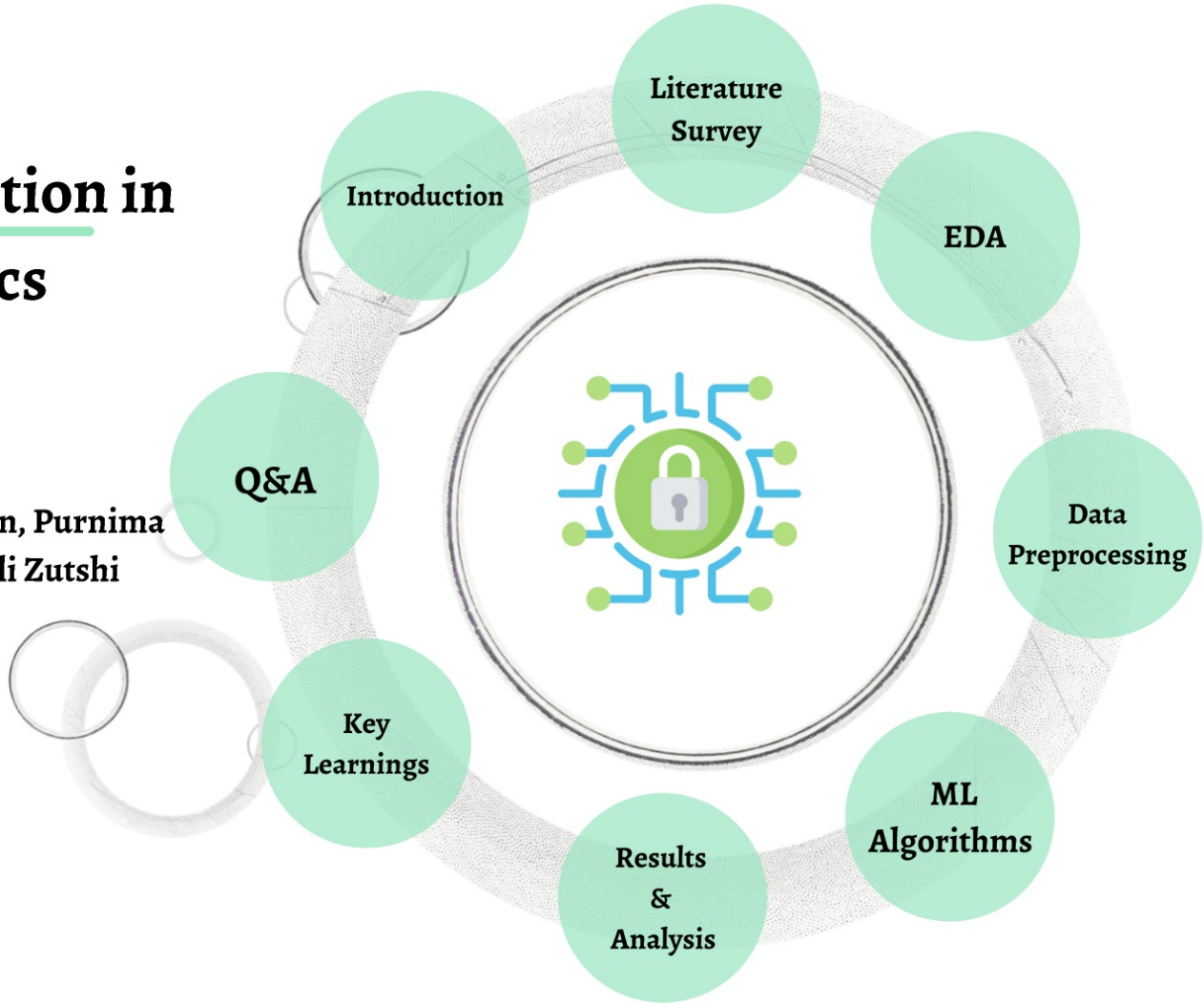
Traditional Machine Learning

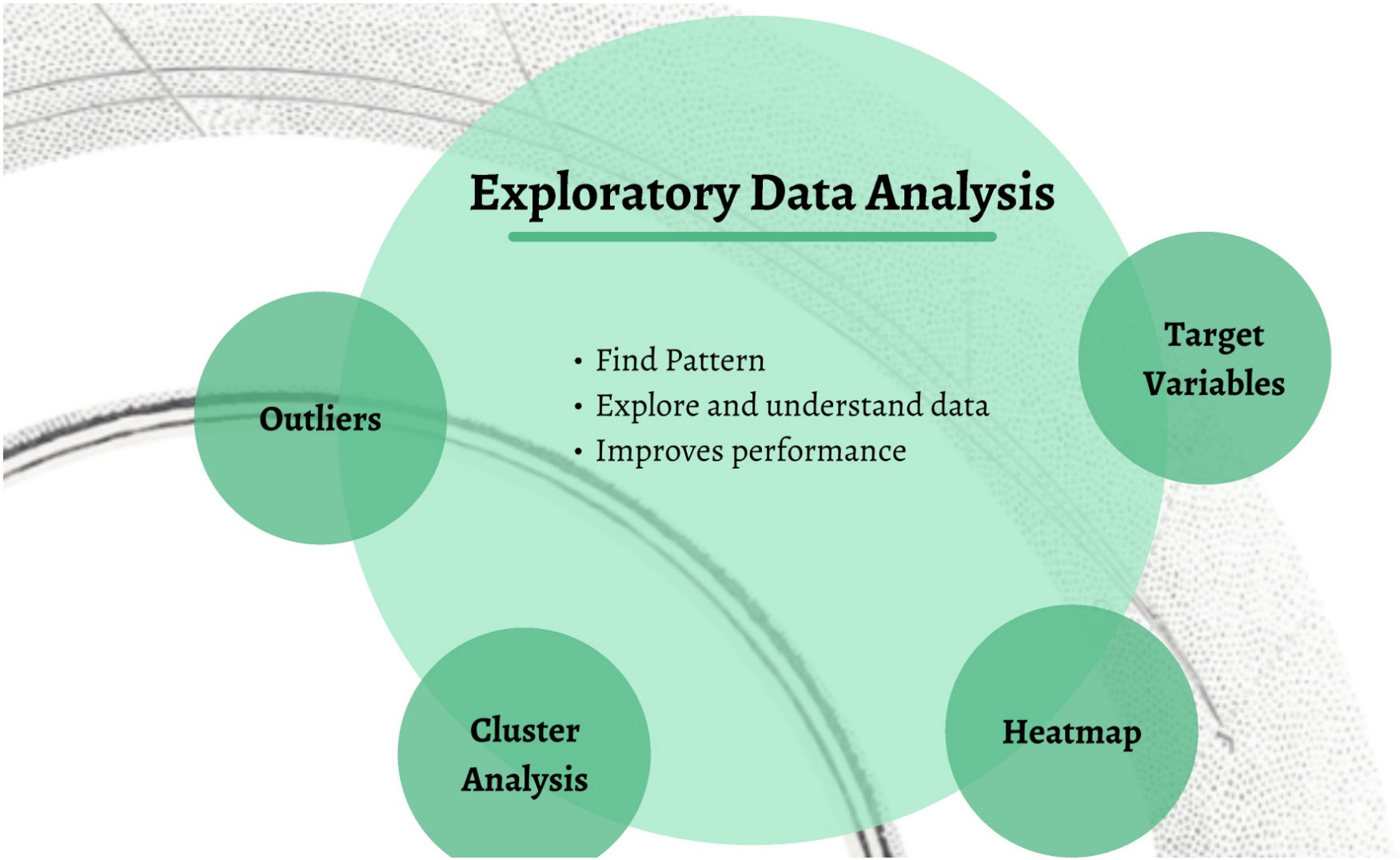
Deep Learning

Cyber Attack Detection in Cloud Forensics

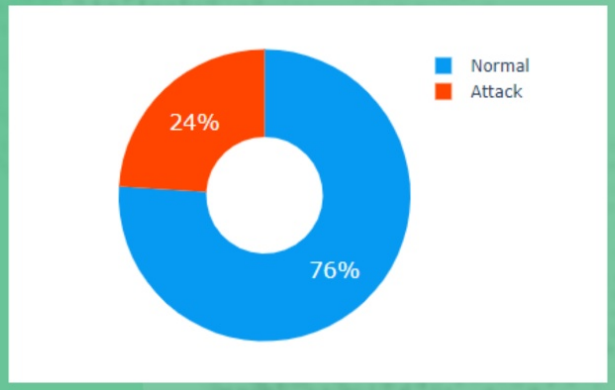
by

Yasaman Emami, Shamama Afnan, Purnima Bhukya, Poojitha Katta, Deepali Zutshi




Exploratory Data Analysis
• Find Pattern
• Explore and understand data
• Improves performance
Outliers
Target Variables
Cluster Analysis
Heatmap

Distribution of Target Variables

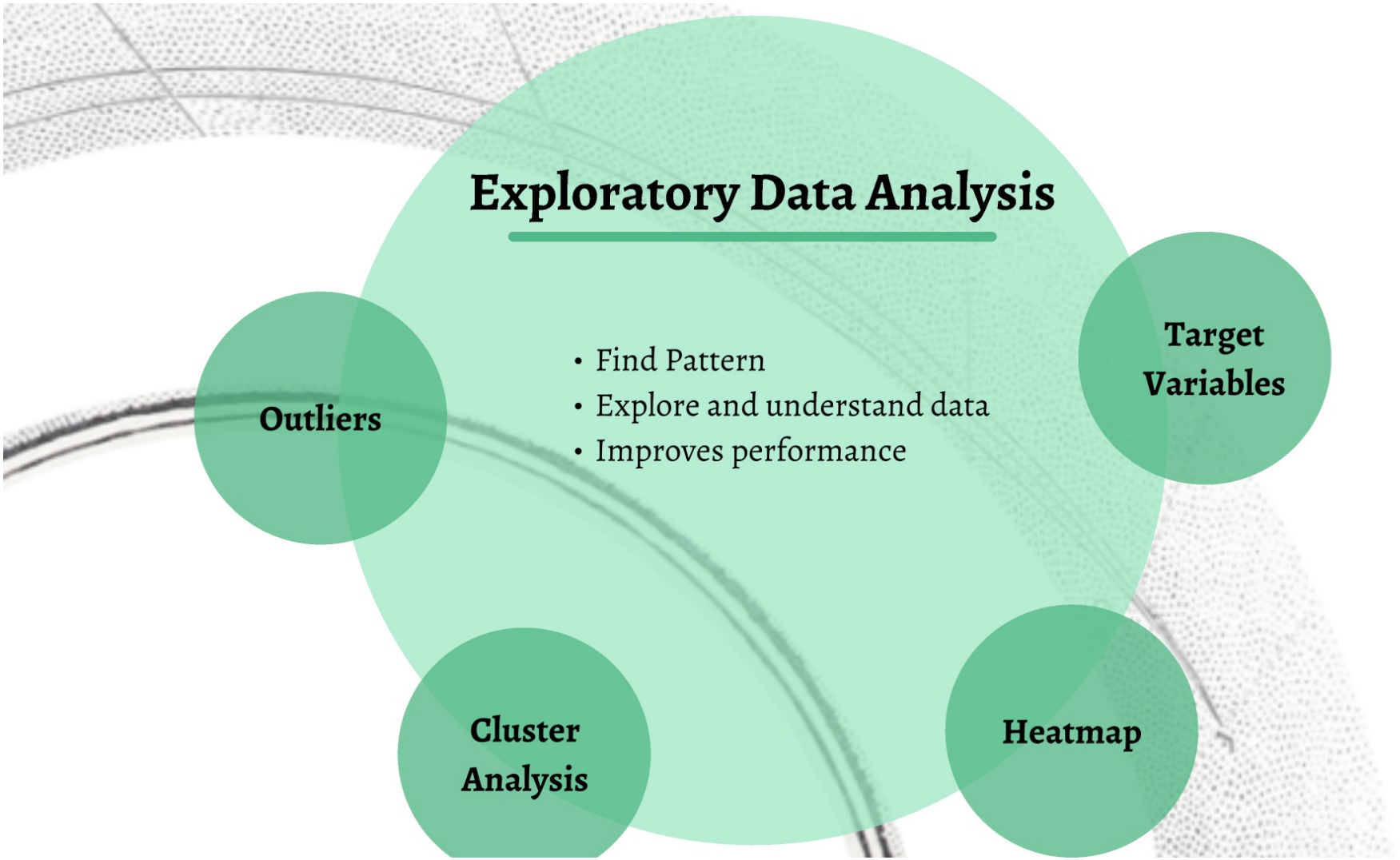


The dataset is highly skewed and must be balanced.

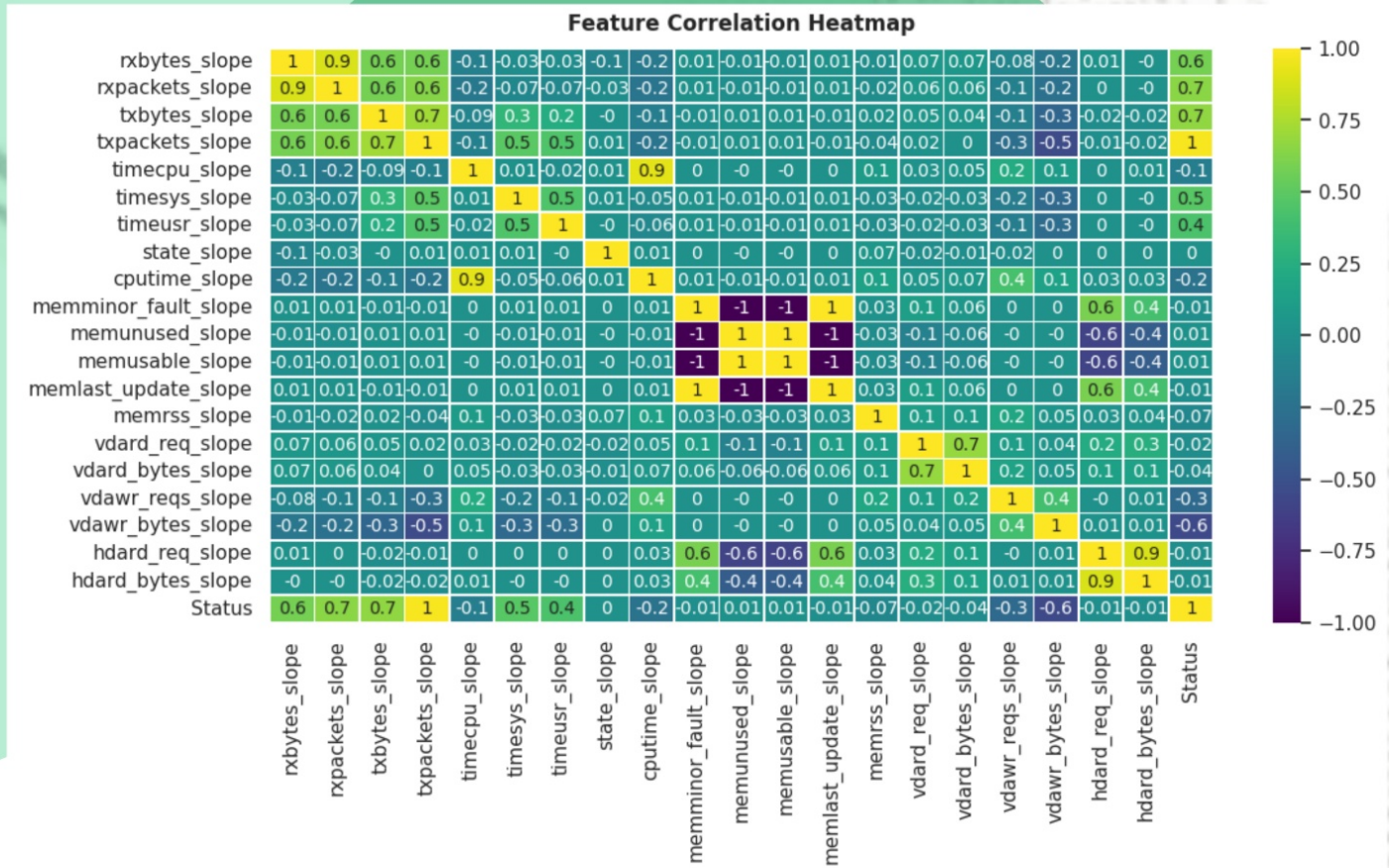
Count of each instance:

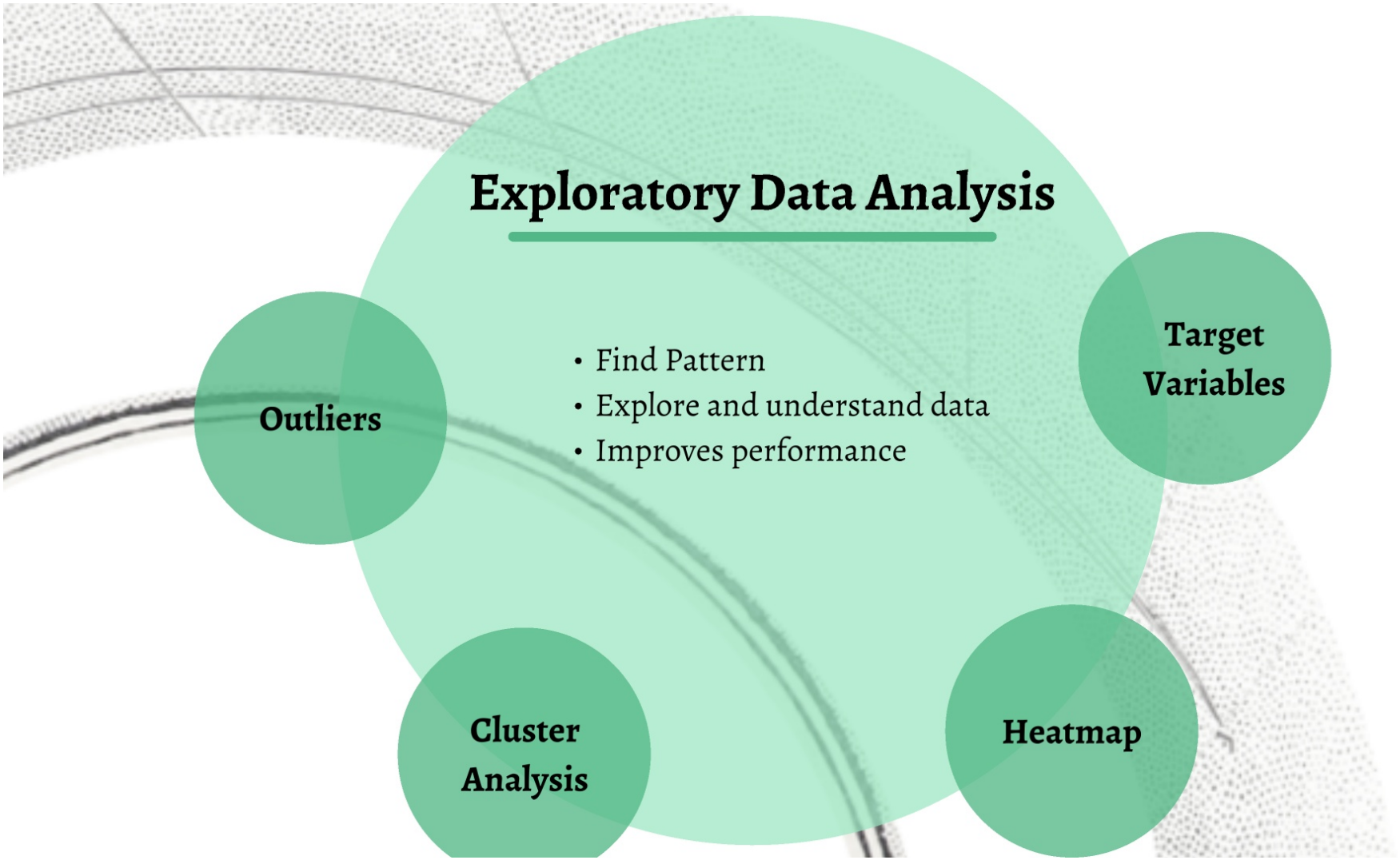
'Normal': 7288

'Attack': 2306

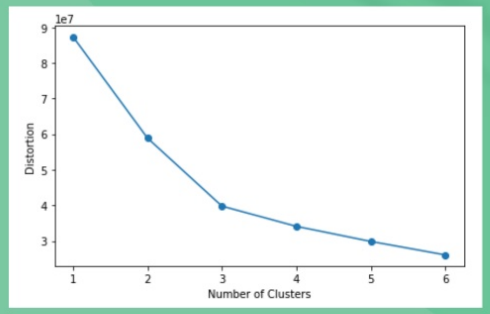

Exploratory Data Analysis
• Find Pattern
• Explore and understand data
• Improves performance
Outliers
Target Variables
Cluster Analysis
Heatmap

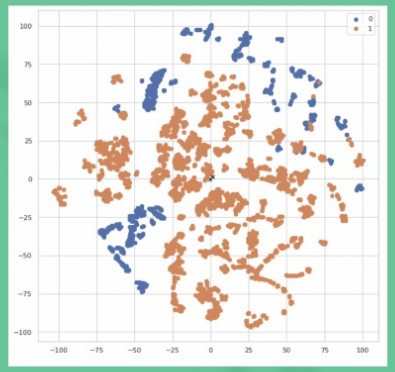
Heatmap


Feature Correlation Heatmap
rxbytes_slope 1 0.9 0.6 0.6 -0.1 -0.03 -0.03 -0.1 -0.2 0.01 -0.01 -0.01 0.01 -0.01 0.07 0.07 -0.08 -0.2 0.01 -0 0.6
rxpackets_slope 0.9 1 0.6 0.6 -0.2 -0.07 -0.07 -0.03 -0.2 0.01 -0.01 -0.01 0.01 -0.02 0.06 0.06 -0.1 -0.2 0 -0 0.7
txbytes_slope 0.6 0.6 1 0.7 -0.09 0.3 0.2 -0 -0.1 -0.01 0.01 0.01 -0.01 0.02 0.05 0.04 -0.1 -0.3 -0.02 -0.02 0.7
txpackets_slope 0.6 0.6 0.7 1 -0.1 0.5 0.5 0.01 -0.2 -0.01 0.01 0.01 -0.01 -0.04 0.02 0 -0.3 -0.5 -0.01 -0.02 1
timecpu_slope -0.1 -0.2 -0.09 -0.1 1 0.01 -0.02 0.01 0.9 0 -0 -0 0 0.1 0.03 0.05 0.2 0.1 0 0.01 -0.1
timesys_slope -0.03 -0.07 0.3 0.5 0.01 1 0.5 0.01 -0.05 0.01 -0.01 -0.01 0.01 -0.03 -0.02 -0.03 -0.2 -0.3 0 -0 0.5
timeusr_slope -0.03 -0.07 0.2 0.5 -0.02 0.5 1 -0 -0.06 0.01 -0.01 -0.01 0.01 -0.03 -0.02 -0.03 -0.1 -0.3 0 -0 0.4
state_slope -0.1 -0.03 -0 0.01 0.01 0.01 -0 1 0.01 0 -0 -0 0 0.07 -0.02 -0.01 -0.02 0 0 0 0
cputime_slope -0.2 -0.2 -0.1 -0.2 0.9 -0.05 -0.06 0.01 1 0.01 -0.01 -0.01 0.01 0.1 0.05 0.07 0.4 0.1 0.03 0.03 -0.2
memminor_fault_slope 0.01 0.01 -0.01 -0.01 0 0.01 0.01 0 0.01 1 -1 -1 1 0.03 0.1 0.06 0 0 0.6 0.4 -0.01
memunused_slope -0.01 -0.01 0.01 0.01 -0 -0.01 -0.01 -0 -0.01 -1 1 1 -1 -0.03 -0.1 -0.06 -0 -0 -0.6 -0.4 0.01
memusable_slope -0.01 -0.01 0.01 0.01 -0 -0.01 -0.01 -0 -0.01 -1 1 1 -1 -0.03 -0.1 -0.06 -0 -0 -0.6 -0.4 0.01
memlast_update_slope 0.01 0.01 -0.01 -0.01 0 0.01 0.01 0 0.01 1 -1 -1 1 0.03 0.1 0.06 0 0 0.6 0.4 -0.01
memrss_slope -0.01 -0.02 0.02 -0.04 0.1 -0.03 -0.03 0.07 0.1 0.03 -0.03 -0.03 0.03 1 0.1 0.1 0.2 0.05 0.03 0.04 -0.07
vdard_req_slope 0.07 0.06 0.05 0.02 0.03 -0.02 -0.02 -0.02 0.05 0.1 -0.1 -0.1 0.1 0.1 1 0.7 0.1 0.04 0.2 0.3 -0.02
vdard_bytes_slope 0.07 0.06 0.04 0 0.05 -0.03 -0.03 -0.01 0.07 0.06 -0.06 -0.06 0.06 0.1 0.7 1 0.2 0.05 0.1 0.1 -0.04
vdawr_reqs_slope -0.08 -0.1 -0.1 -0.3 0.2 -0.2 -0.1 -0.02 0.4 0 -0 -0 0 0.2 0.1 0.2 1 0.4 -0 0.01 -0.3
vdawr_bytes_slope -0.2 -0.2 -0.3 -0.5 0.1 -0.3 -0.3 0 0.1 0 -0 -0 0 0.05 0.04 0.05 0.4 1 0.01 0.01 -0.6
hdard_req_slope 0.01 0 -0.02 -0.01 0 0 0 0 0.03 0.6 -0.6 -0.6 0.6 0.03 0.2 0.1 -0 0.01 1 0.9 -0.01
hdard_bytes_slope -0 -0 -0.02 -0.02 0.01 -0 -0 0 0.03 0.4 -0.4 -0.4 0.4 0.04 0.3 0.1 0.01 0.01 0.9 1 -0.01
Status 0.6 0.7 0.7 1 -0.1 0.5 0.4 0 -0.2 -0.01 0.01 0.01 -0.01 -0.07 -0.02 -0.04 -0.3 -0.6 -0.01 -0.01 1
rxbytes_slope
rxpackets_slope
txbytes_slope
txpackets_slope
timecpu_slope
timesys_slope
timeusr_slope
state_slope
cputime_slope
memminor_fault_slope
memunused_slope
memusable_slope
memlast_update_slope
memrss_slope
vdard_req_slope
vdard_bytes_slope
vdawr_reqs_slope
vdawr_bytes_slope
hdard_req_slope
hdard_bytes_slope
Status
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Exploratory Data Analysis
• Find Pattern
• Explore and understand data
• Improves performance
Outliers
Target Variables
Cluster Analysis
Heatmap

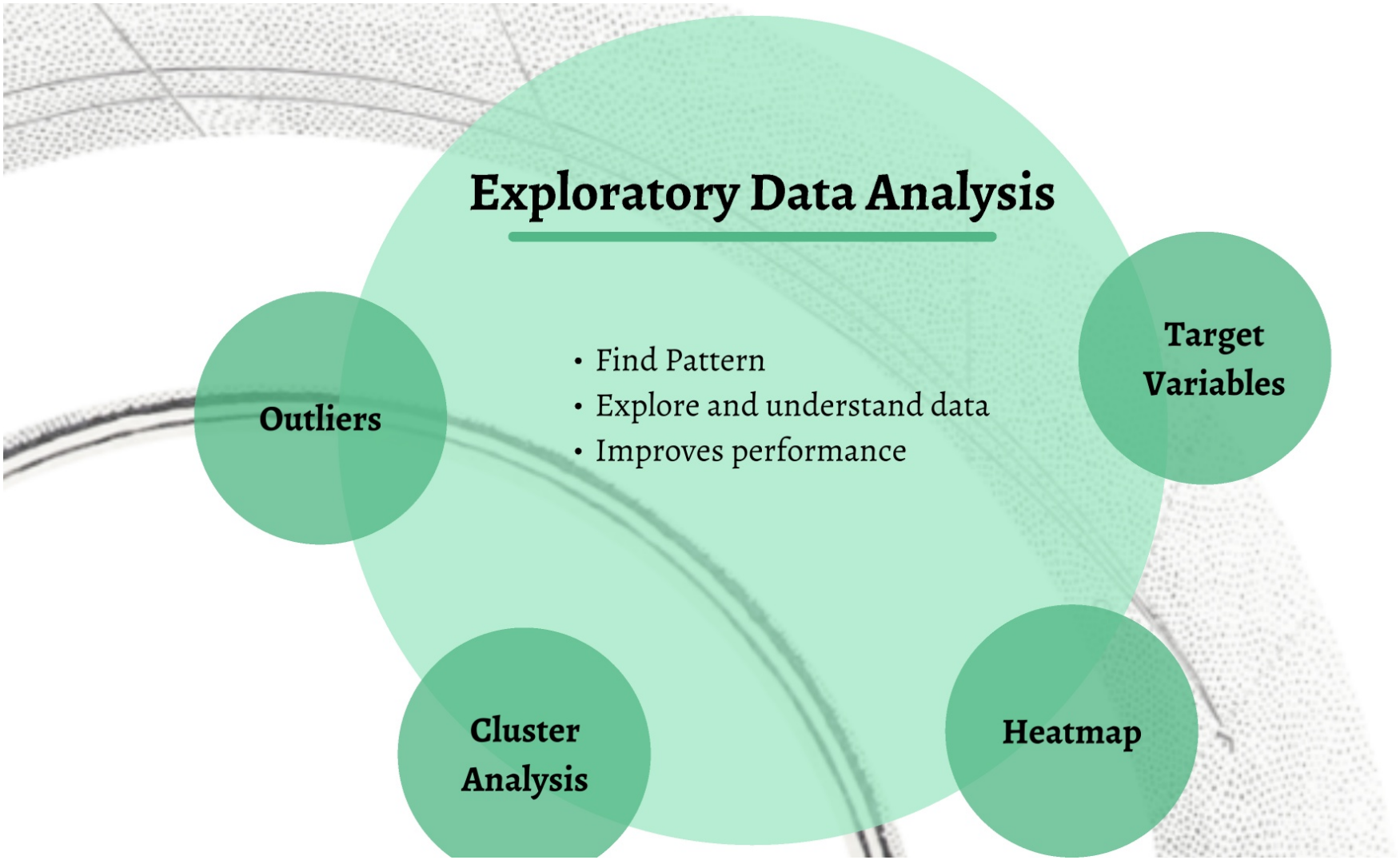
Elbow Method



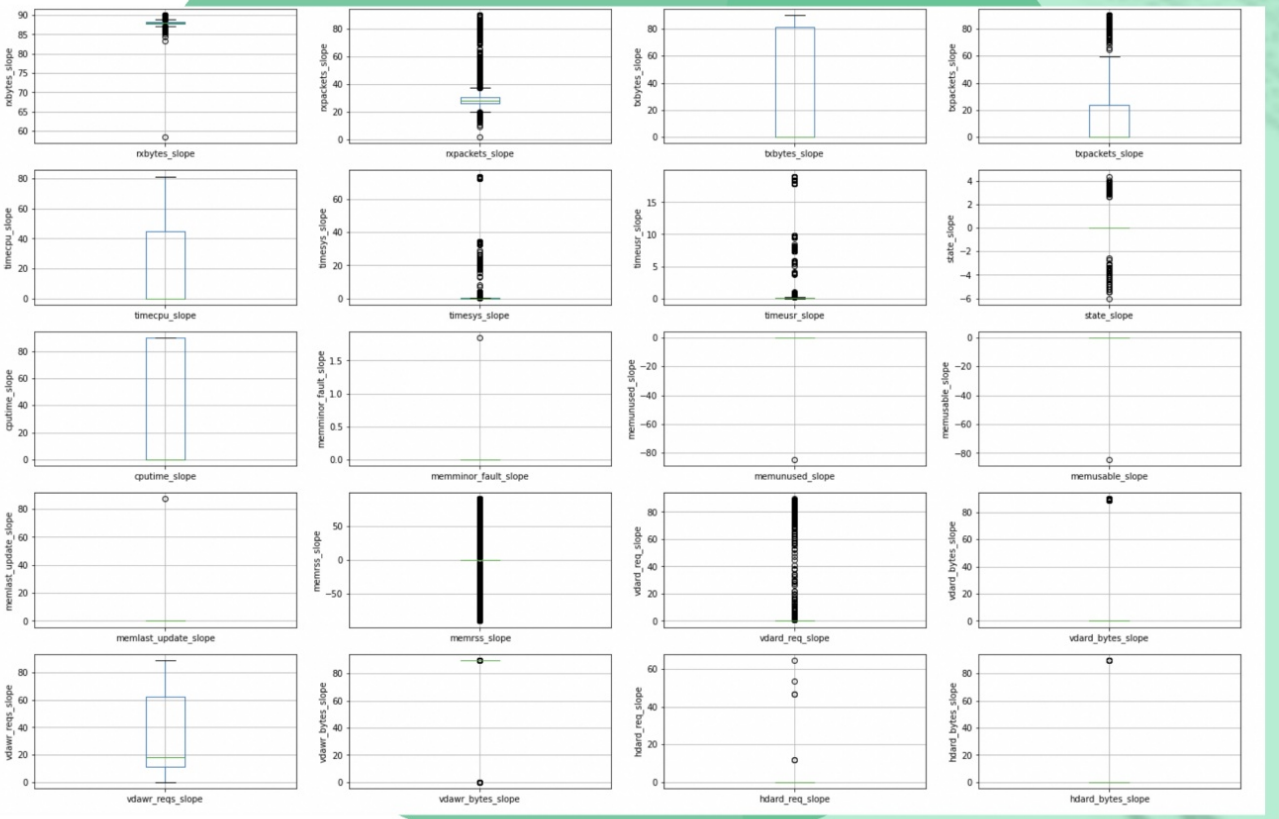


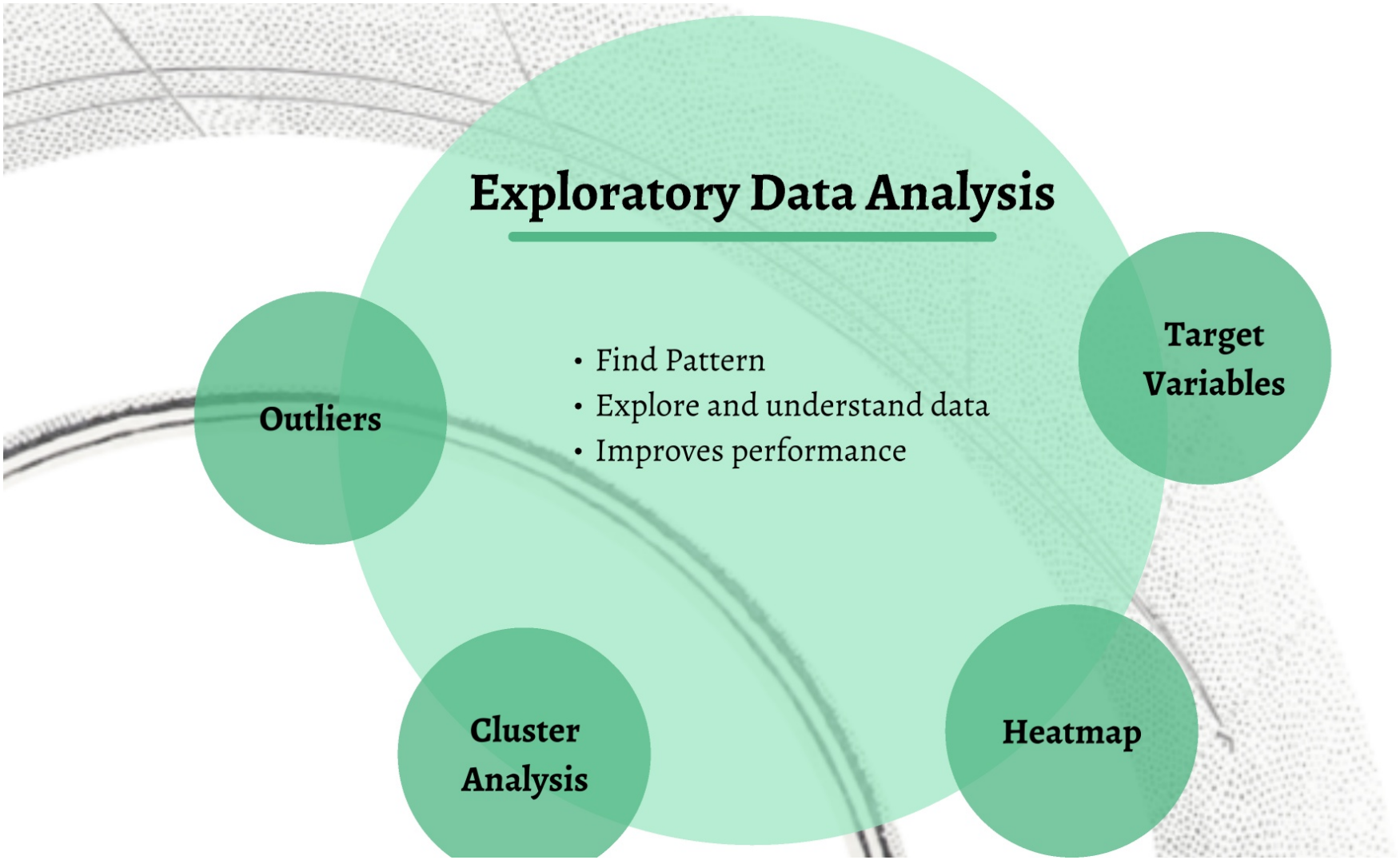
Homogeneity Score = 0.87

It indicates how many of the clusters predicted contain only members of a single class.


Exploratory Data Analysis
• Find Pattern
• Explore and understand data
• Improves performance
Outliers
Target Variables
Cluster Analysis
Heatmap

Visualizing Outliers

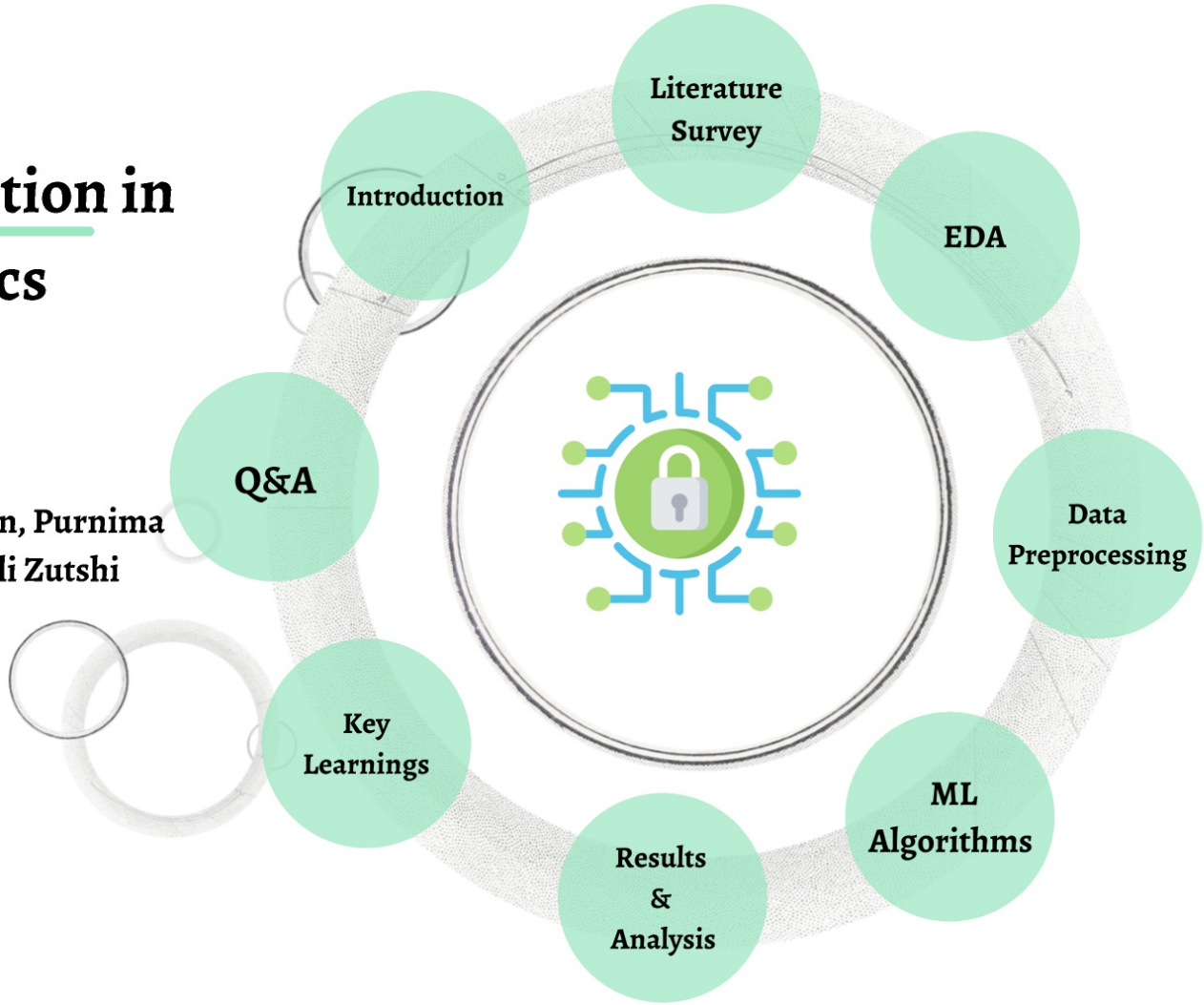


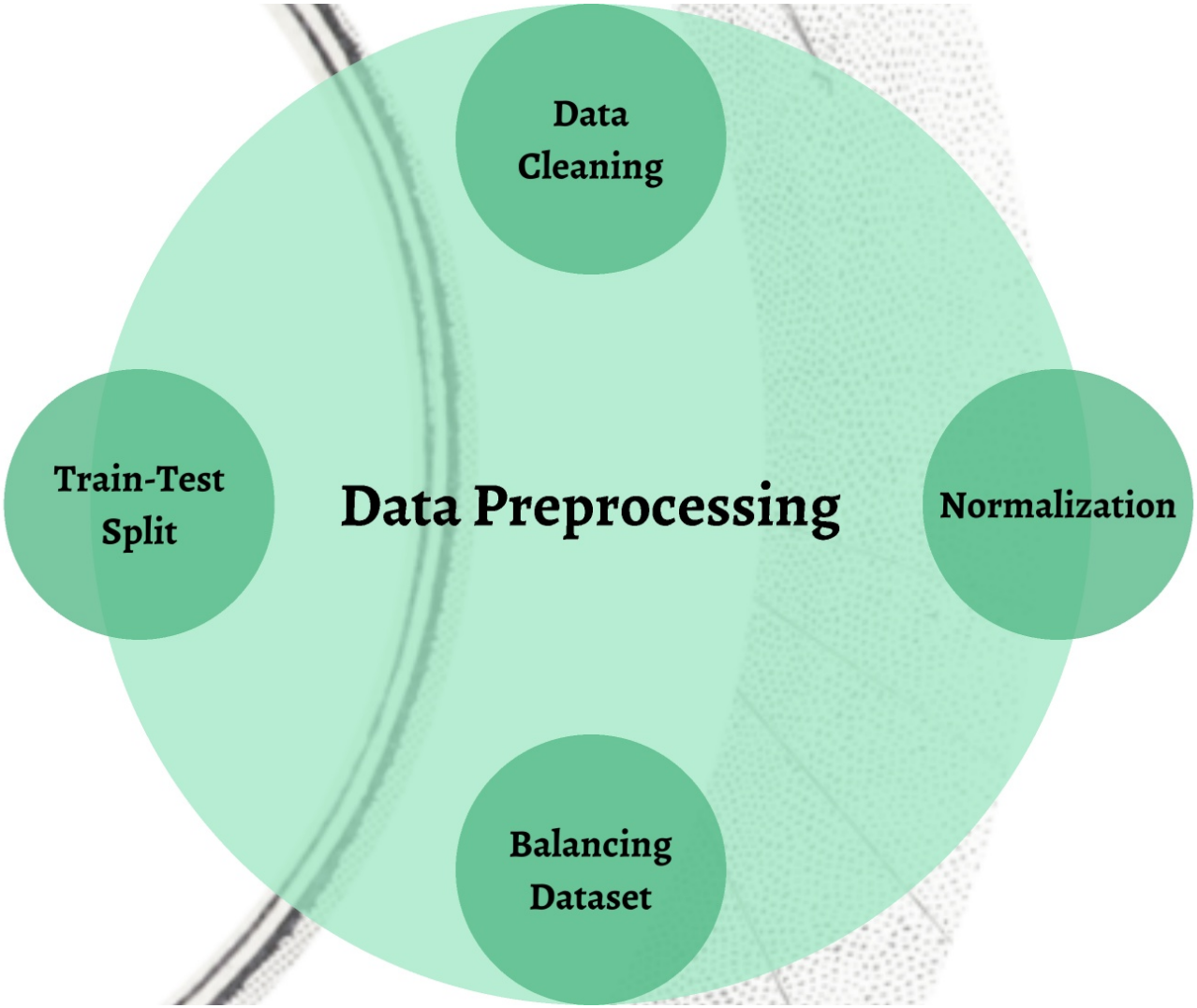

Exploratory Data Analysis
• Find Pattern
• Explore and understand data
• Improves performance
Outliers
Target Variables
Cluster Analysis
Heatmap

Cyber Attack Detection in Cloud Forensics

by

Yasaman Emami, Shamama Afnan, Purnima Bhukya, Poojitha Katta, Deepali Zutshi

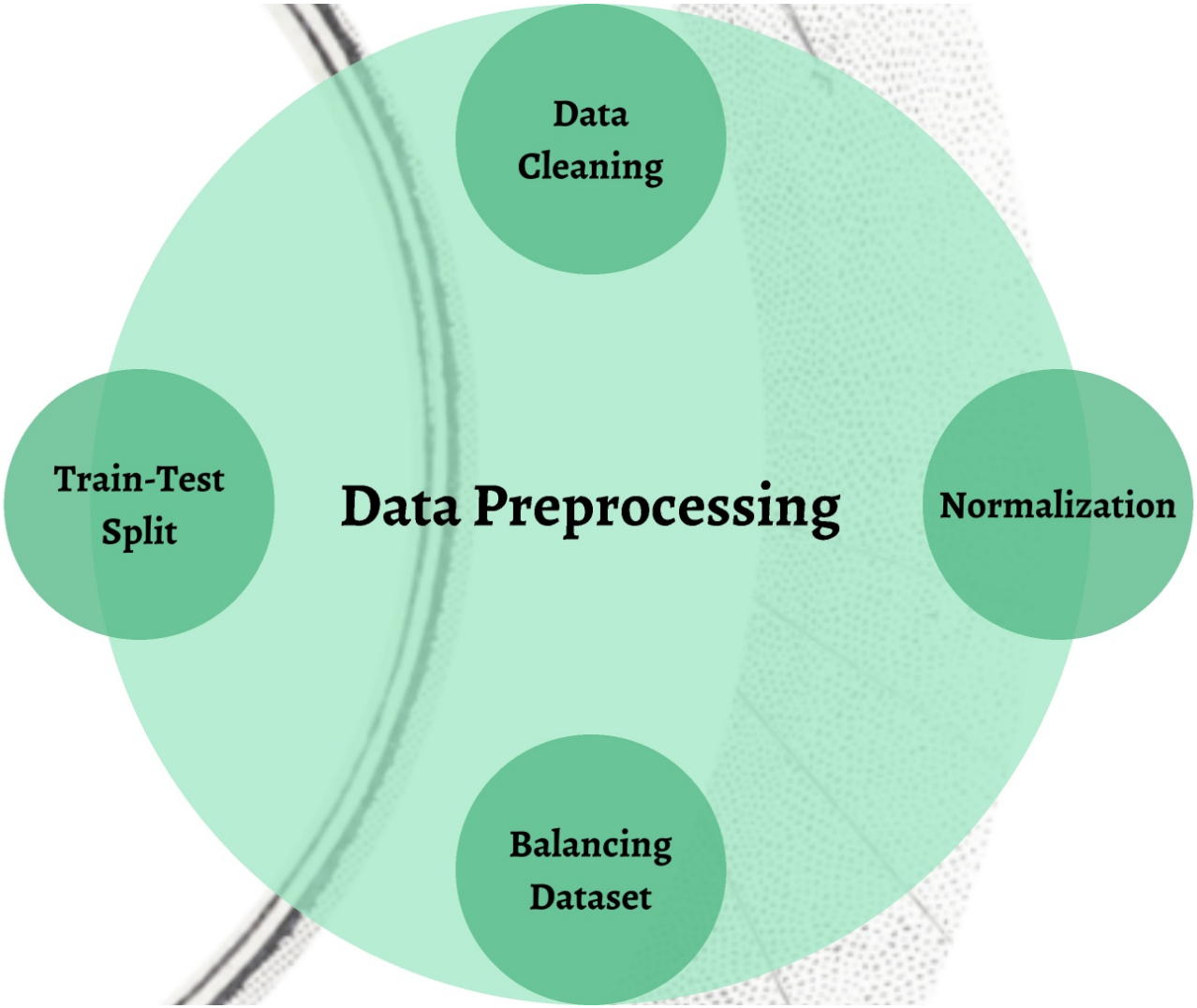



Data Cleaning
Train-Test Split
Data Preprocessing
Normalization
Balancing Dataset

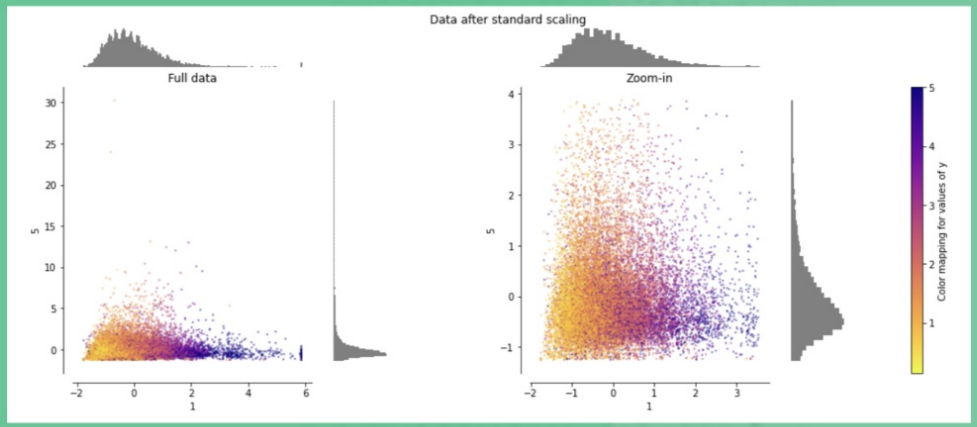
Data Cleaning

- Drop Null values
- Removing irrelevant features:
  - Last Poll: Timestamp
  - VMID: Virtual Machine Identity
  - UUID: Unique Domain Identifier
  - Dom: Domain Name
  - Features with all zero values

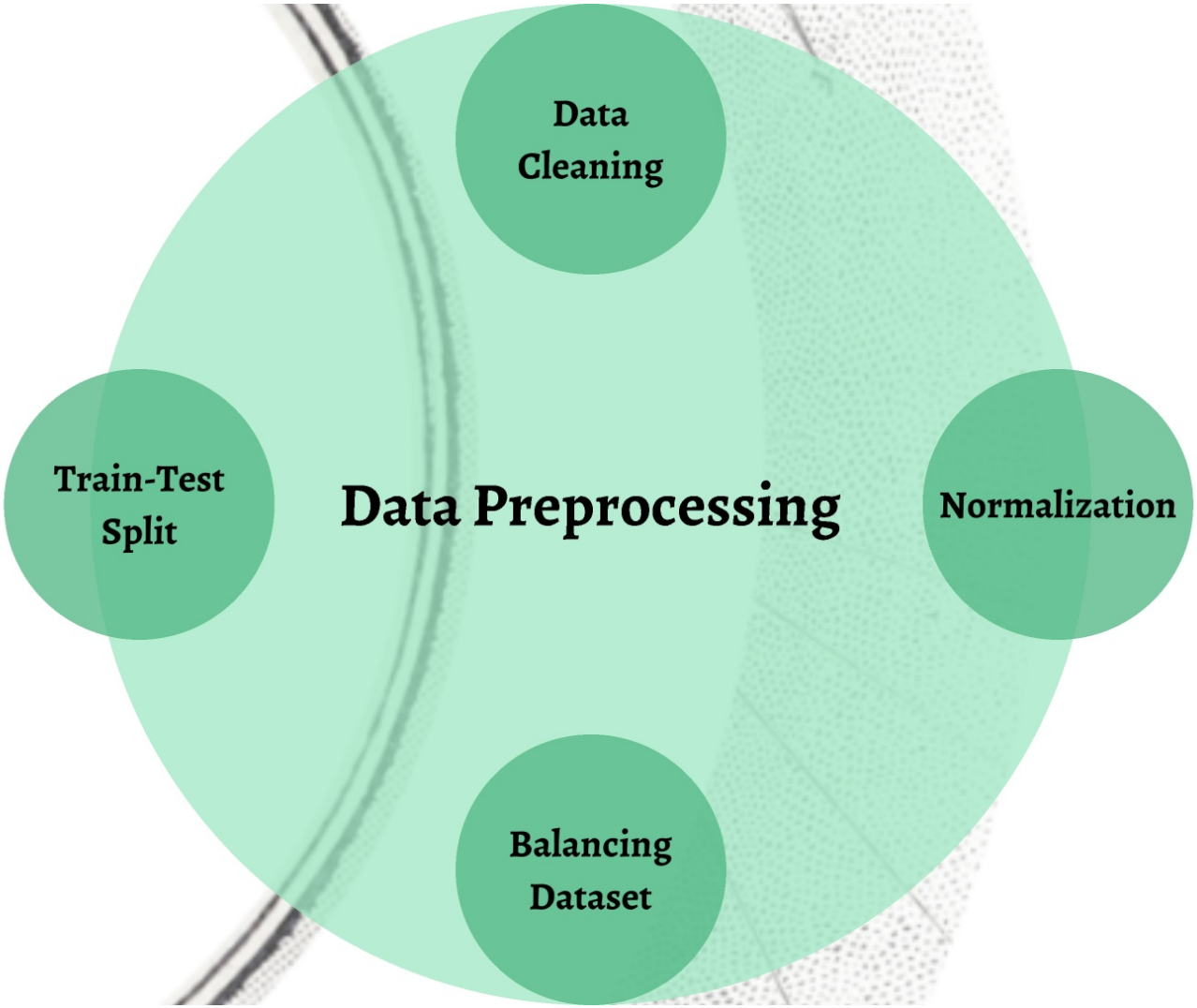
- 'Status' column contained string values 'Normal' and 'Attack' which were converted to integer.


Data Cleaning
Train-Test Split
Data Preprocessing
Normalization
Balancing Dataset

Data Standardization

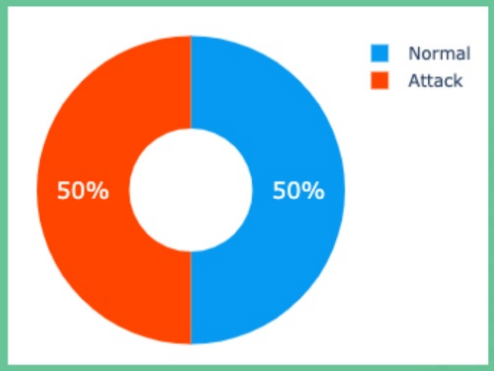


The features are scaled for each of the inputs to shift the distribution to have a mean value of 0 and the standard deviation of 1.

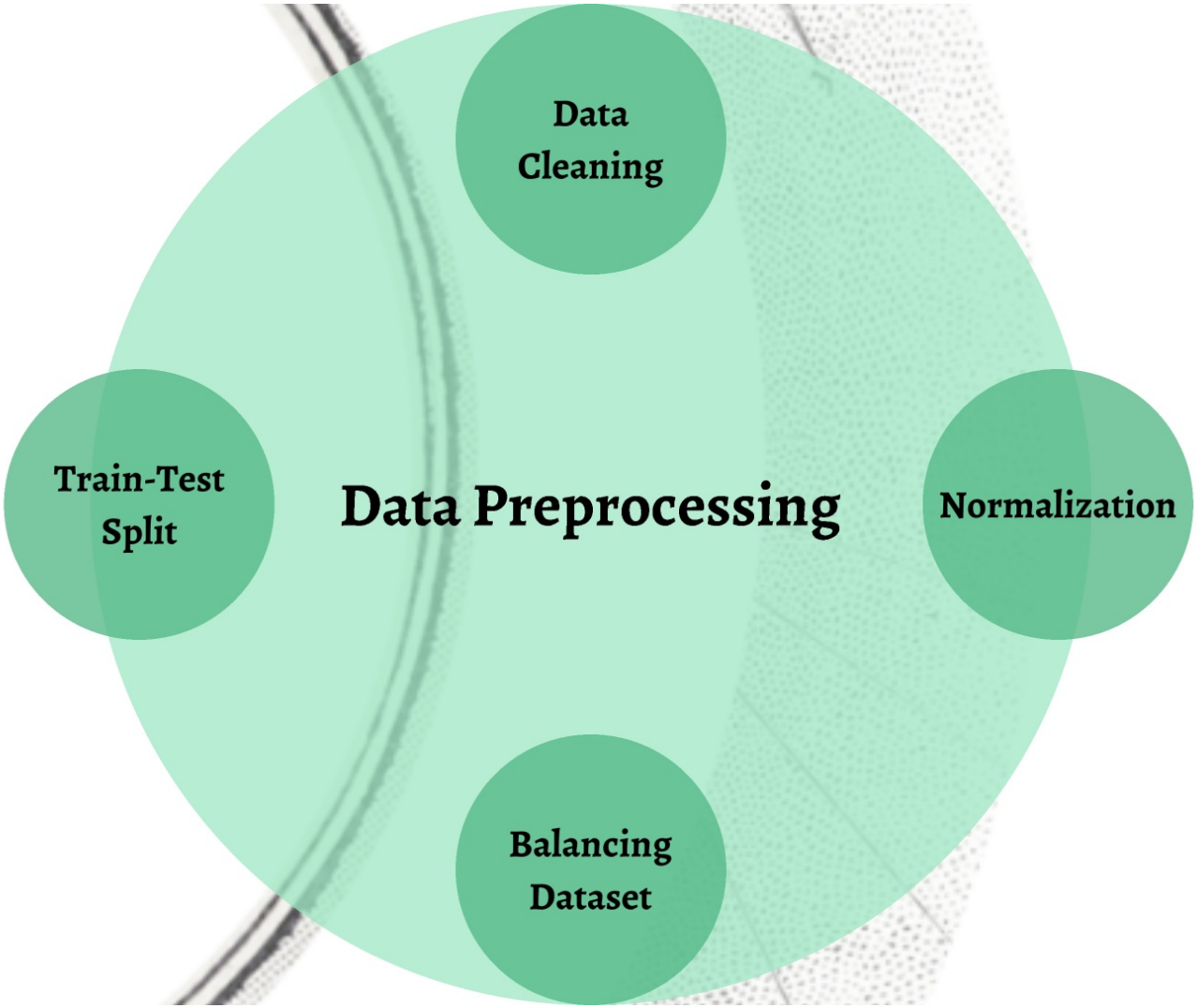

Data Cleaning
Train-Test Split
Data Preprocessing
Normalization
Balancing Dataset

Data Oversampling

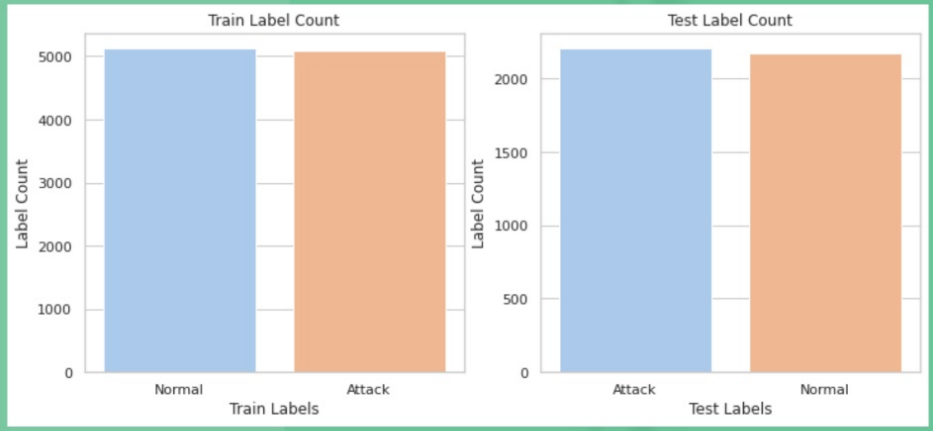
The original dataset was highly skewed hence random oversampling was performed on the instances labeled 'Attack' to balance it.



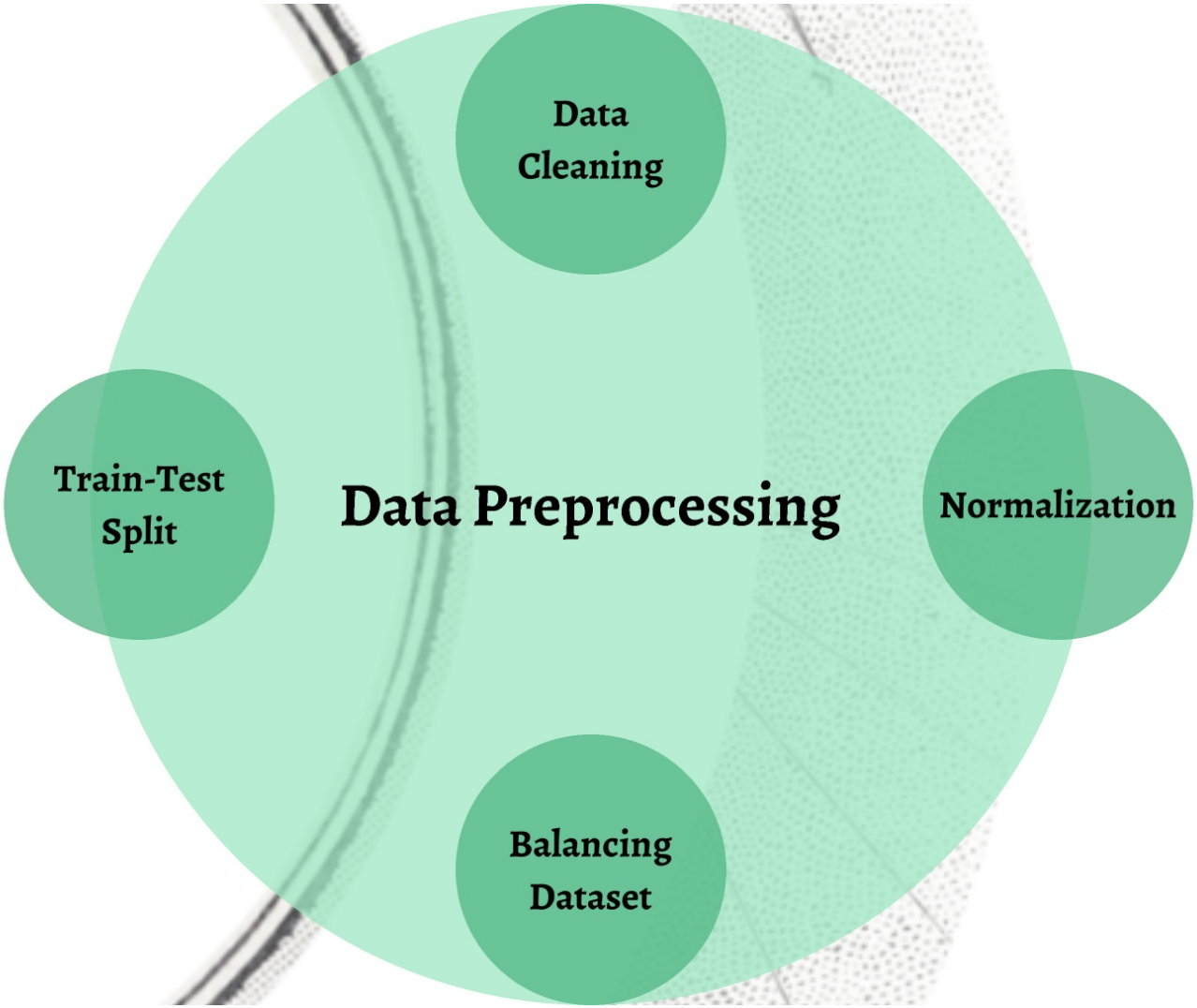
Count after oversampling:
'Normal': 7288
'Attack': 7288


Data Cleaning
Train-Test Split
Data Preprocessing
Normalization
Balancing Dataset

Data Splitting



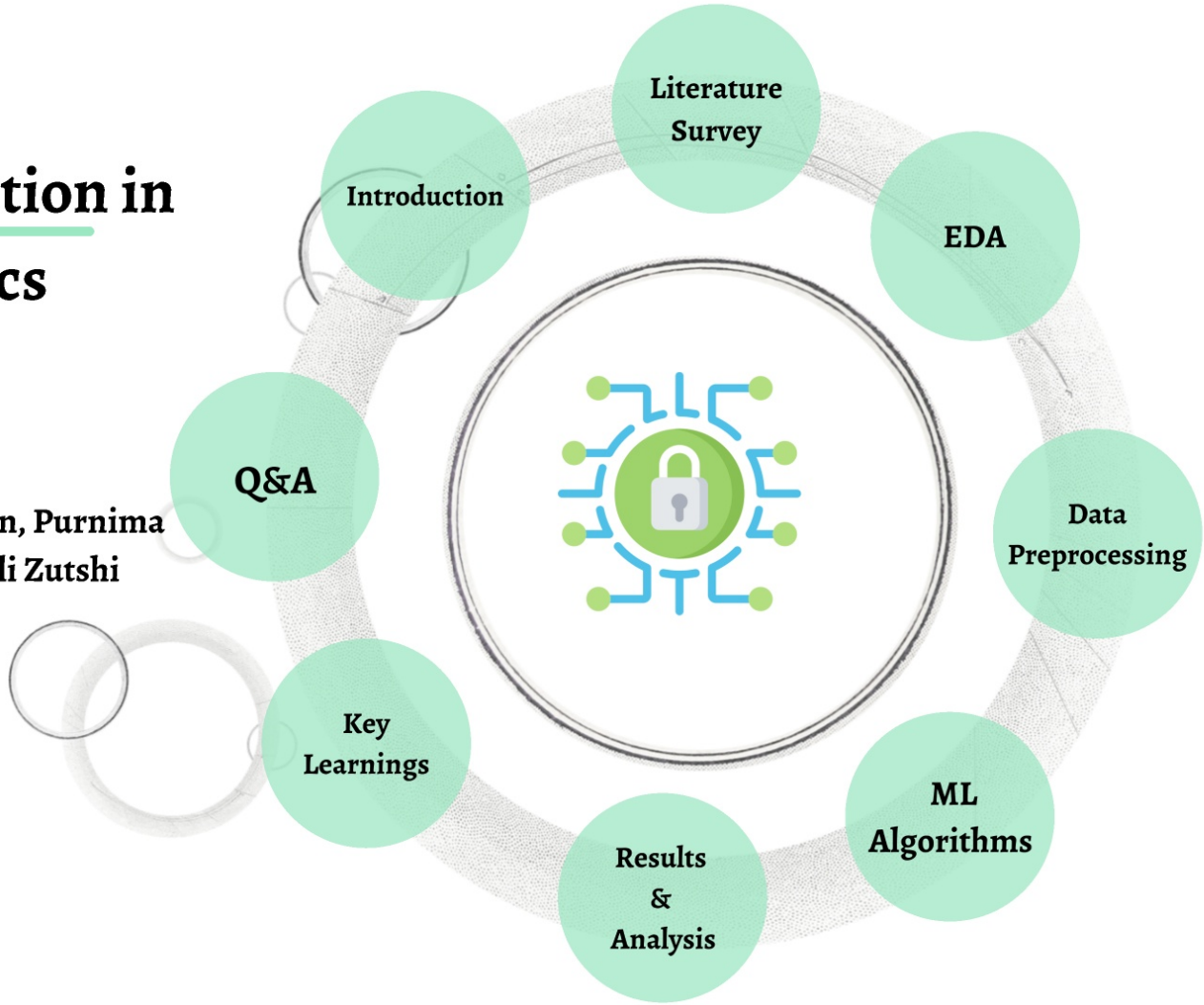
Splitting Ratio : 70:30


Data Cleaning
Train-Test Split
Data Preprocessing
Normalization
Balancing Dataset

Cyber Attack Detection in Cloud Forensics

by

Yasaman Emami, Shamama Afnan, Purnima Bhukya, Poojitha Katta, Deepali Zutshi



Implementation

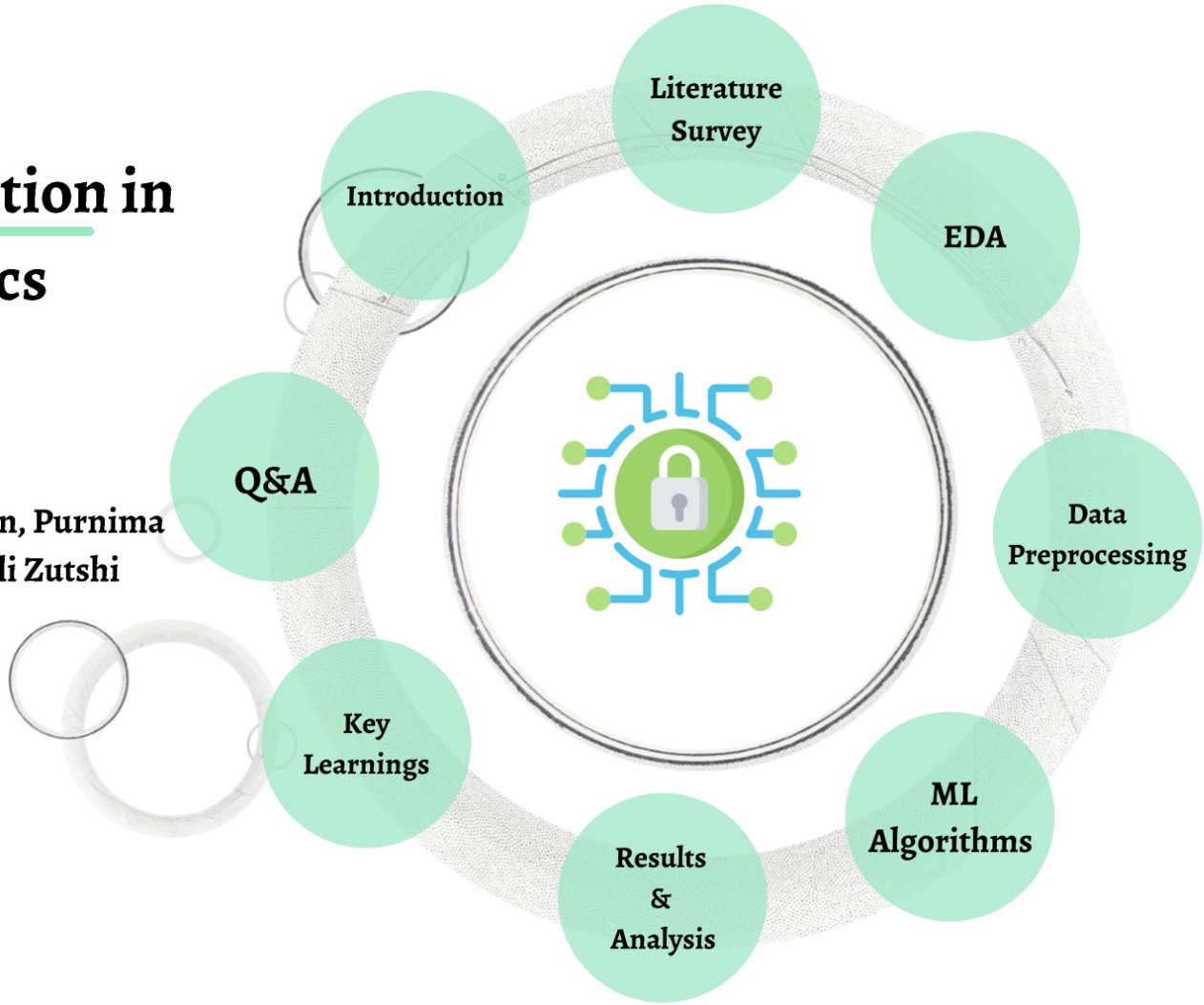
The following algorithms were implemented and their performances were evaluated:

- Logistic Regression
- Naive Bayes
- Decision Tree
- Support Vector Machine
- K-Nearest Neighbor with Genetic Algorithm
- Random Forest
- XGBoost

Cyber Attack Detection in Cloud Forensics

by

Yasaman Emami, Shamama Afnan, Purnima Bhukya, Poojitha Katta, Deepali Zutshi

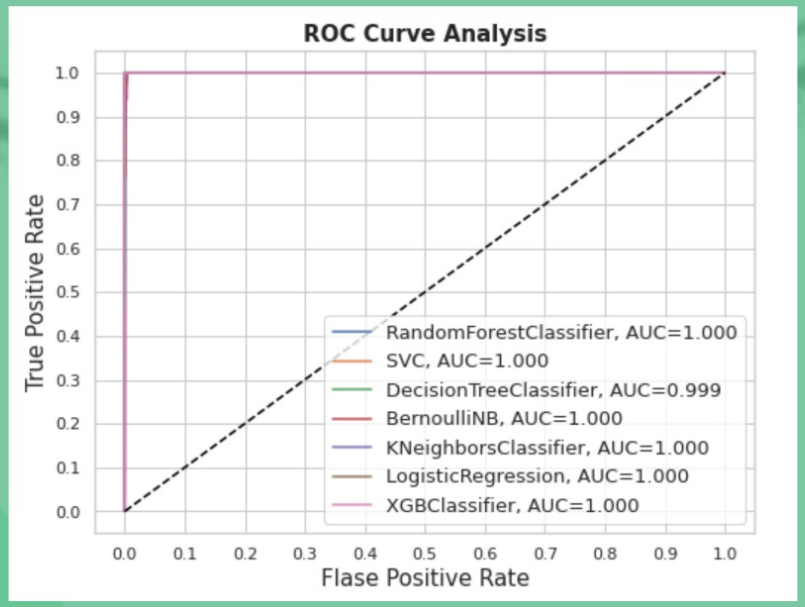


Performance Comparison

ROC Curve

The following graphs compare the precision, recall, F-1 score, accuracy and the Kappa statistic for the implemented models before and after balancing the dataset.

Receiver Operator Characteristic Curve



Performance Comparison

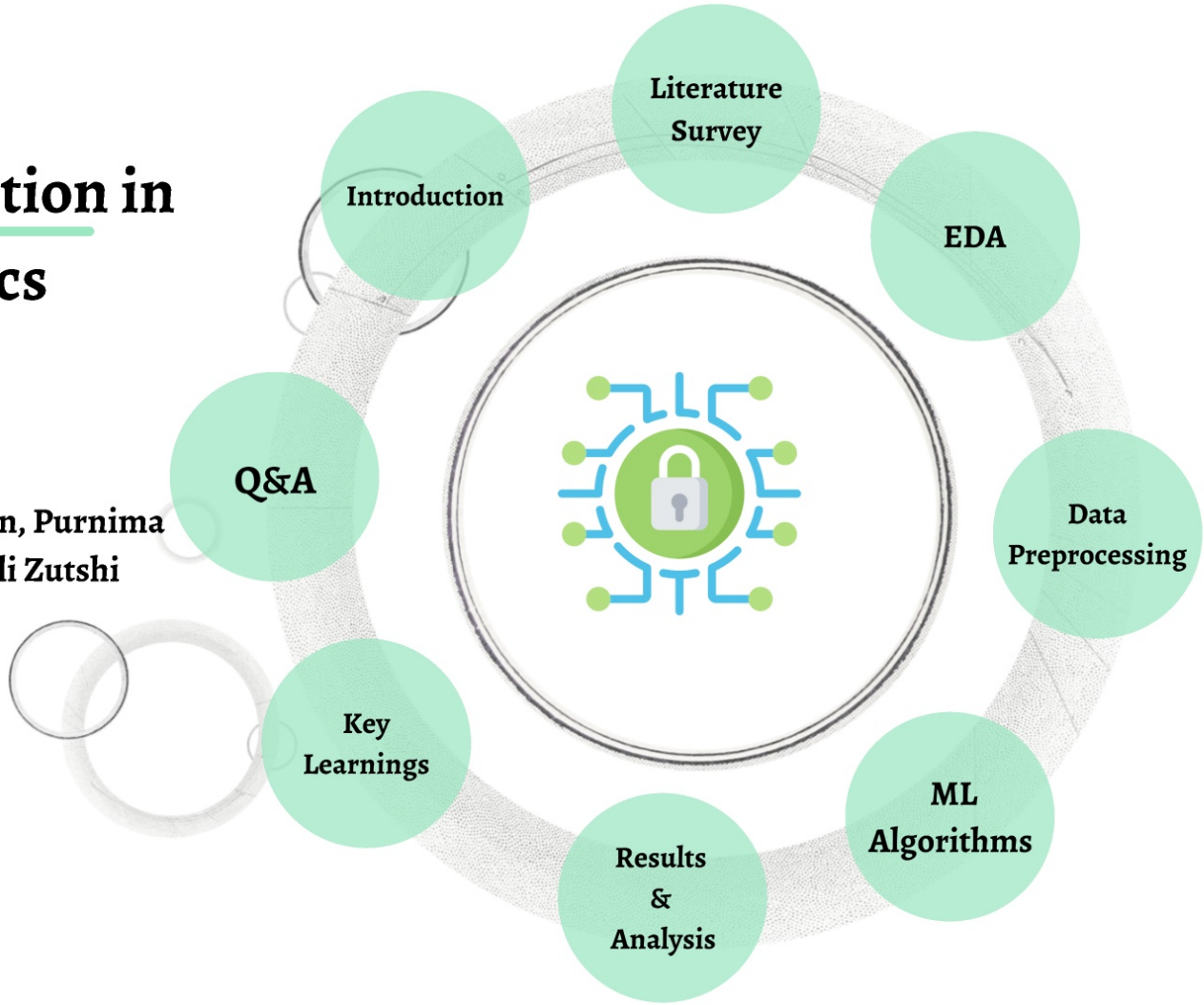
ROC Curve

The following graphs compare the precision, recall, F-1 score, accuracy and the Kappa statistic for the implemented models before and after balancing the dataset.

Cyber Attack Detection in Cloud Forensics

by

Yasaman Emami, Shamama Afnan, Purnima Bhukya, Poojitha Katta, Deepali Zutshi



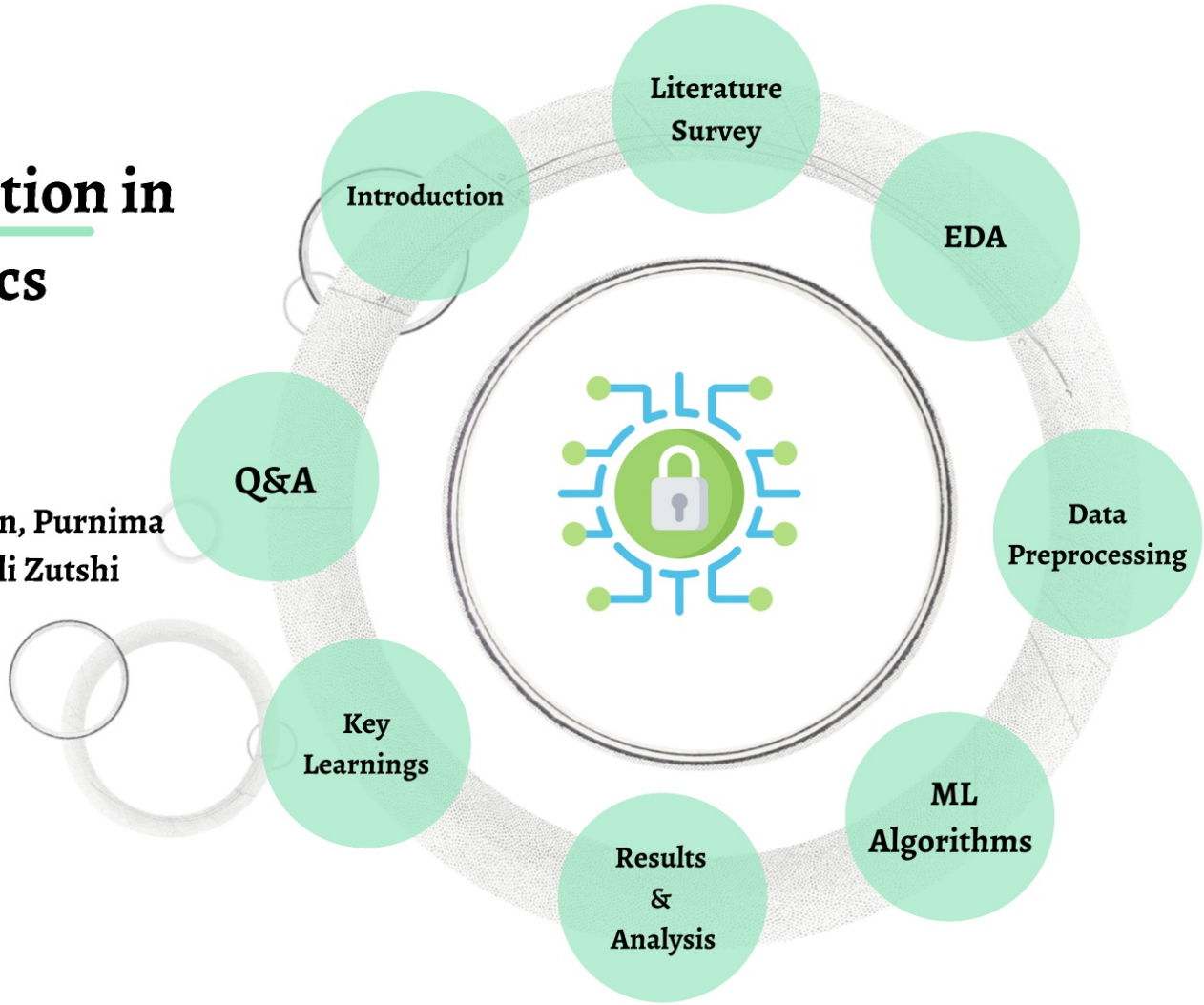
Key Learnings

- ML has major applications in preventing cyber attacks and network security.
- Decision Tree must be pruned to remove unusable features and prevent overfitting.
- txpackets_slope, rxbytes_slope, txbytes_slope are important features across most of trained models.
- KNN + GA improves classification performance as it calculates the similarities between the training and testing samples at each level.
- SVM can classify detection with high accuracy not only with imbalanced dataset but also with a subset of important features highly correlated to the .
- Oversampling improved the performance for all of the models.

Cyber Attack Detection in Cloud Forensics

by

Yasaman Emami, Shamama Afnan, Purnima Bhukya, Poojitha Katta, Deepali Zutshi



References

References

[1] Kilincer, I. F., Ertam, F., & Sengur, A. (2021). Machine learning methods for cyber security intrusion detection: Datasets and comparative study. Computer Networks, 188, 107840.

[2] Gottwalt, F., Chang, E., & Dillon, T. (2019). CorrCorr: A feature selection method for multivariate correlation network anomaly detection techniques. Computers & Security, 83, 234-245.

[3] Alshammari and A. Aldribi, "Apply machine learning techniques to detect malicious network traffic in cloud computing," Journal of Big Data, vol. 8, no. 1, 2021.

[4] Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. Journal of Information Security and Applications, 50, 102419.

[5] T. Salman, D. Bhamare, A. Erbad, R. Jain and M. Samaka, "Machine Learning for Anomaly Detection and Categorization in Multi-Cloud Environments," 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud), 2017, pp. 97-103, doi: 10.1109/CSCloud.2017.15.

[6] M. Zekri, S. E. Kafhali, N. Aboutabit and Y. Saadi, "DDoS attack detection using machine learning techniques in cloud computing environments," 2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), 2017, pp. 1-7, doi: 10.1109/CloudTech.2017.8284731

[7] A. Abusitta, M. Bellaiche, and M. Dagenais, "An SVM-based framework for detecting DOS attacks in virtualized clouds under changing environments," Journal of Cloud Computing, vol. 7, no. 1, 2018.

[8] S. Gumaste, D. G. Narayan, S. Sindhe and K. Amit, "Detection of DDoS Attacks in OpenStack-based Private Cloud Using Apache Spark", Journal of Telecommunications and Information Technology, 2020, doi:10.26636/jtit.2020.146120

[9] R. Grover, C. R. Krishna, A. K. Mishra, E. S. Pilli and M. C. Govil, "A Comparison of Analysis Approaches for Cloud Forensics," 2016 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), 2016, pp. 131-135, doi: 10.1109/CCEM.2016.031

[10] P. S. Saini, S. Behal and S. Bhatia, "Detection of DDoS Attacks using Machine Learning Algorithms," 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom), 2020, pp. 16-21, doi: 10.23919/INDIACom49435.2020.9083716.

[11] P. Bedi, N. Gupta, and V. Jindal, "I-SiamIDS: an improved Siam-IDS for handling class imbalance in network-based intrusion detection systems," Applied Intelligence, Sep. 2020, doi: 10.1007/s10489-020-01886-y.

[12]N. Usman, S. Usman, F. Khan, M. A. Jan, A. Sajid, M. Alazab, and P. Watters, "Intelligent dynamic malware detection using machine learning in IP reputation for forensics data analytics," Future Generation Computer Systems, vol. 118, pp. 124–141, 2021.

[13] A. Mishra, "Intelligent Dynamic Malware Detection using Machine Learning in IP Reputation for Forensics Data Analytics," Medium, 28-May-2020. [Online]. Available: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234. [Accessed: 13-Mar-2022].

[14] Google, "Classification: Precision and Recall | Machine Learning Crash Course | Google Developers," Google Developers, Mar. 05, 2019. https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

References

Cyber Attack Detection in Cloud Forensics

by

Yasaman Emami, Shamama Afnan, Purnima Bhukya, Poojitha Katta, Deepali Zutshi