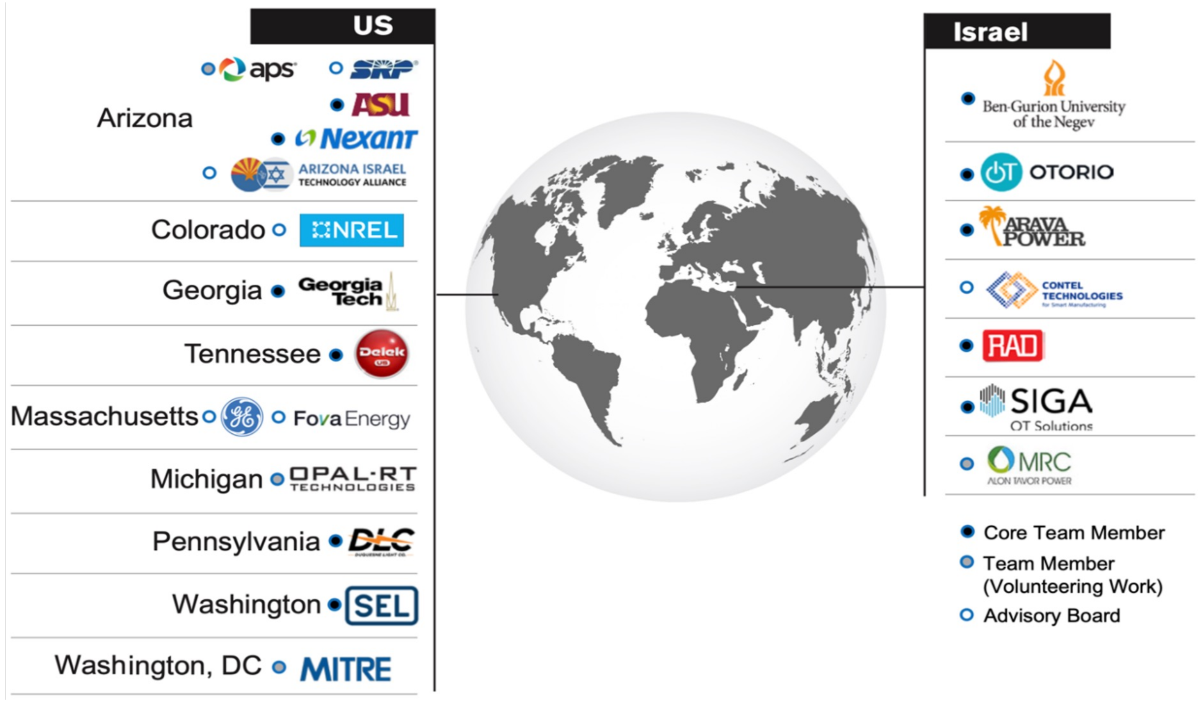
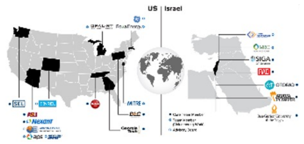
BIRD ICRDE: Task 12 –

Israel-U.S. Energy Center
(Cyber Topic)



Liat Friedman Antwarg, Ben-Gurion University
October 2023

Task 12 - Explainable cyber AI analytics



M12.1 An algorithm to minimize false-positive anomalies using explanations

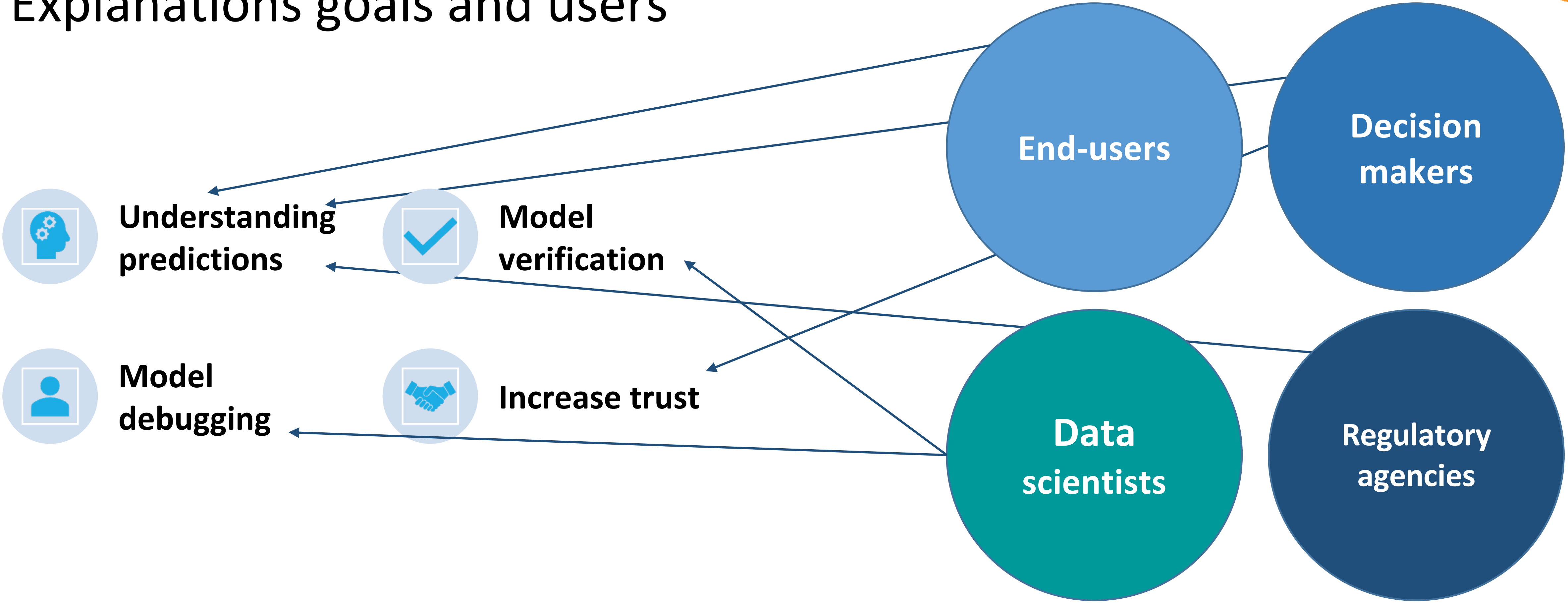
M12.2 An algorithm to identify hidden similarities between instances using explanations

M12.3 Explaining autoencoder's output

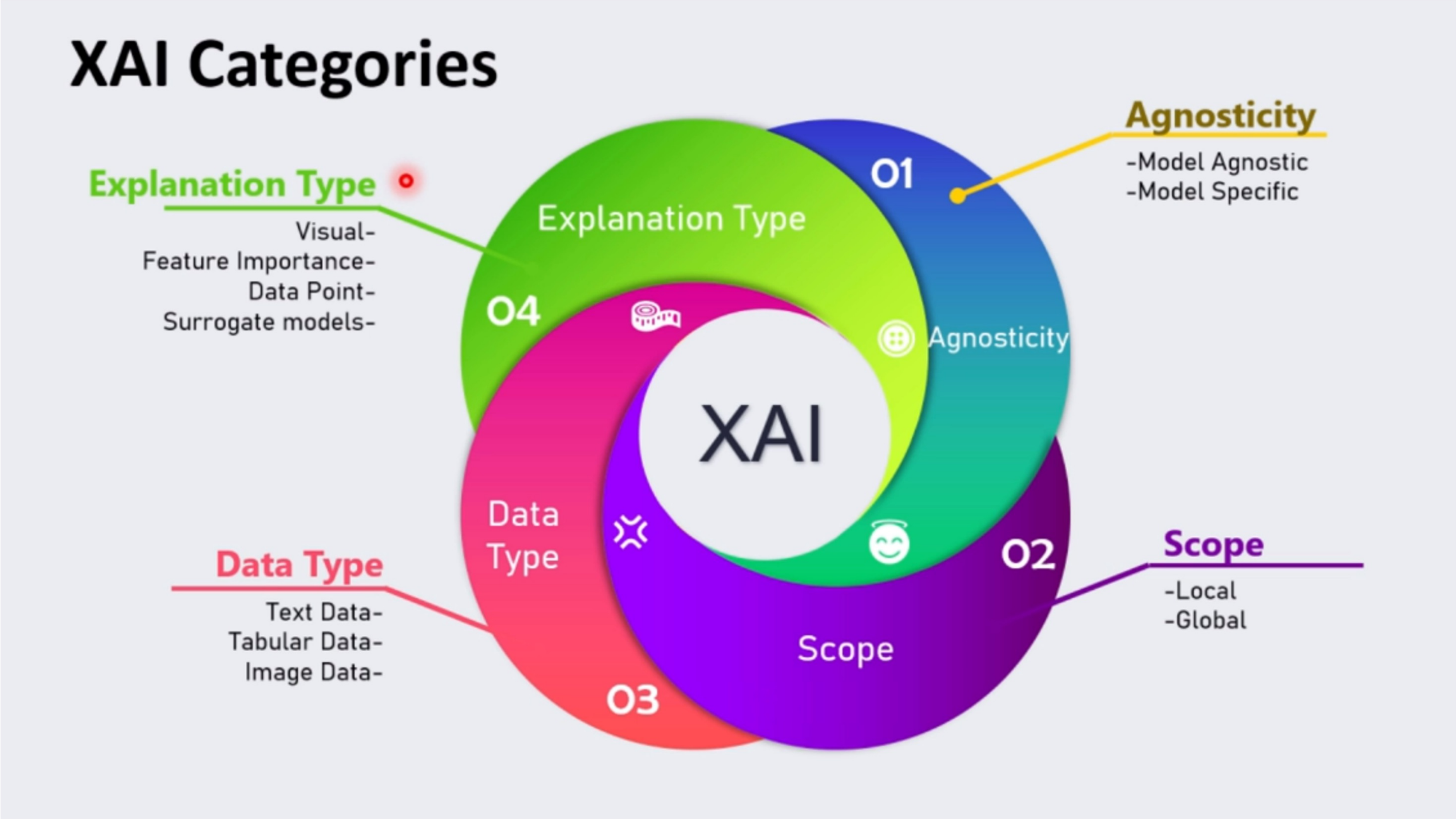
SVM

Explainable AI

Explanations goals and users

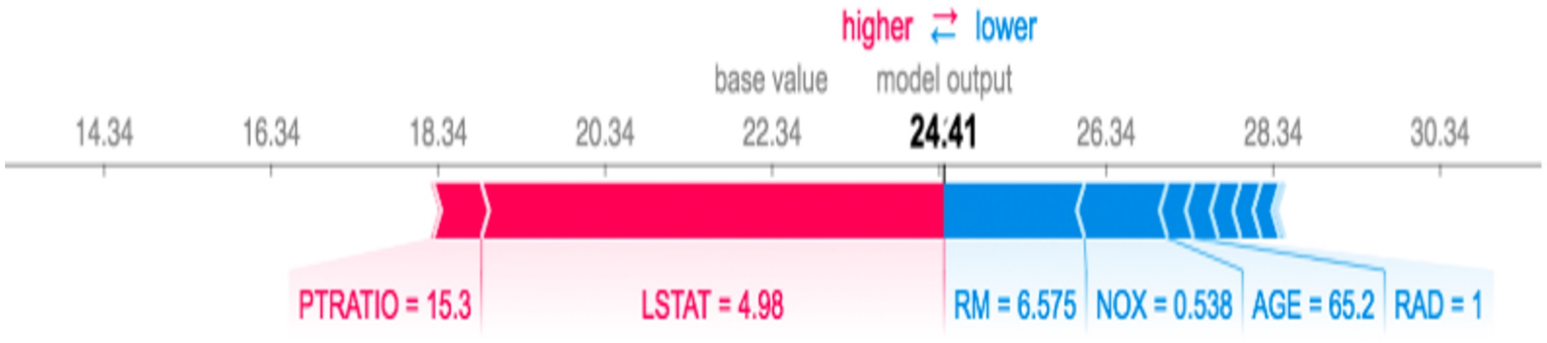


Explainable AI



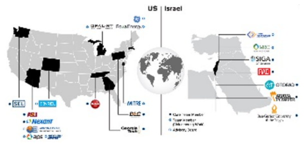
SHAP (SHapley Additive exPlanations)

- Additive feature attribution method that is based on game theory and used for understanding why a model makes a certain prediction.
- Assigns each feature an importance value for a **particular prediction** that represents the marginal contribution of each player in a coalition



▪ https://github.com/slundberg/shap

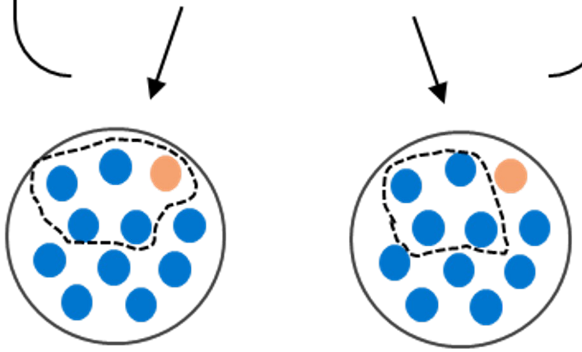
Shap



Shapley Value for player i in game f

Average over all players' subsets $S \subseteq N/\{i\}$

Marginal contribution $f(S \cup \{i\}) - f(S)$

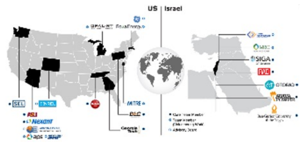


$f - game$

$N - all\ players$

$S - subset\ of\ players$

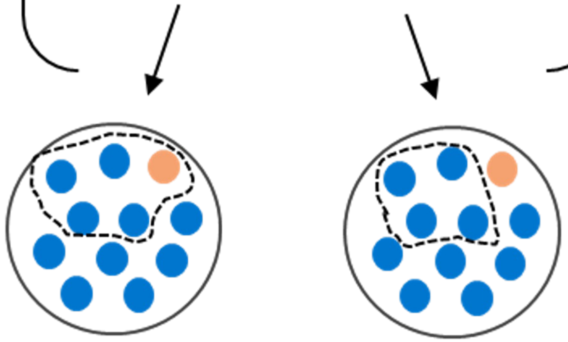
$i - specific\ player$



From players to models

SHAP Value for feature i, contribution of feature i to model prediction $f(x)$ → Average over all features' subsets $S \subseteq N/\{i\}$

Marginal contribution
$f(S \cup \{i\}) - f(S)$



f – *model*

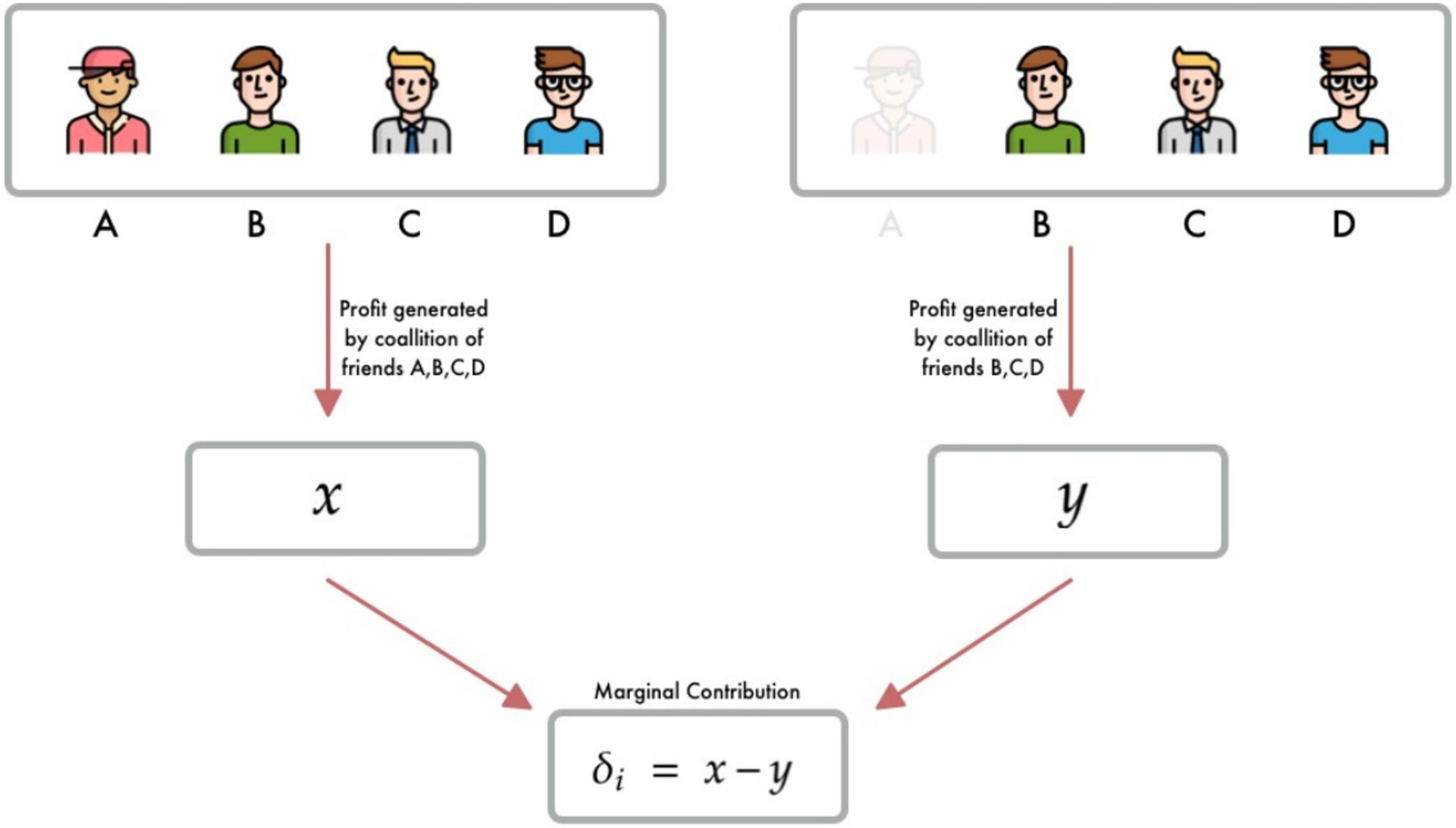
N – *all features*

S – *subset of features*

i – *specific feature*

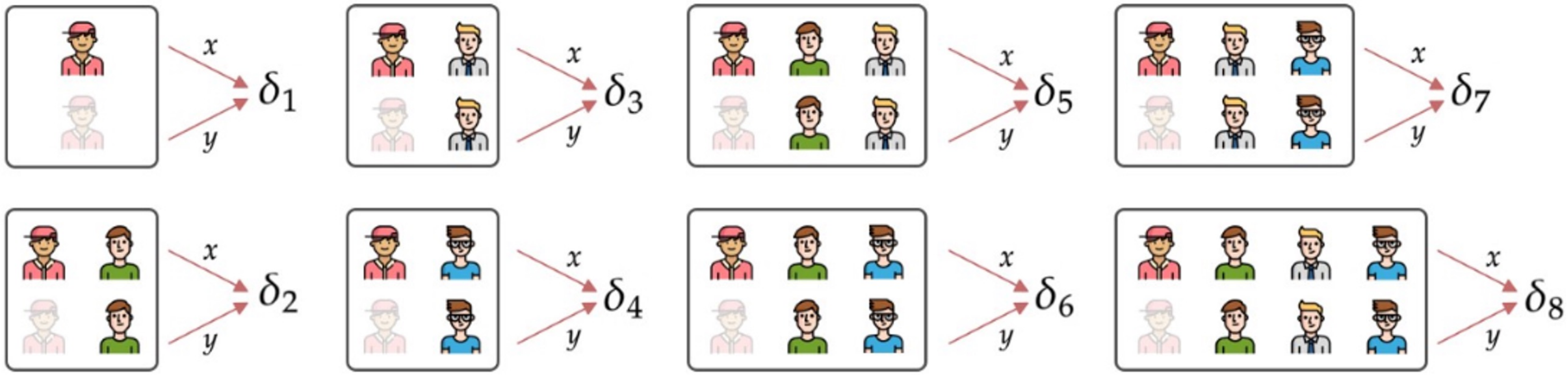
x – *inctance being explained*

Shapley values



https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3

Shapley values



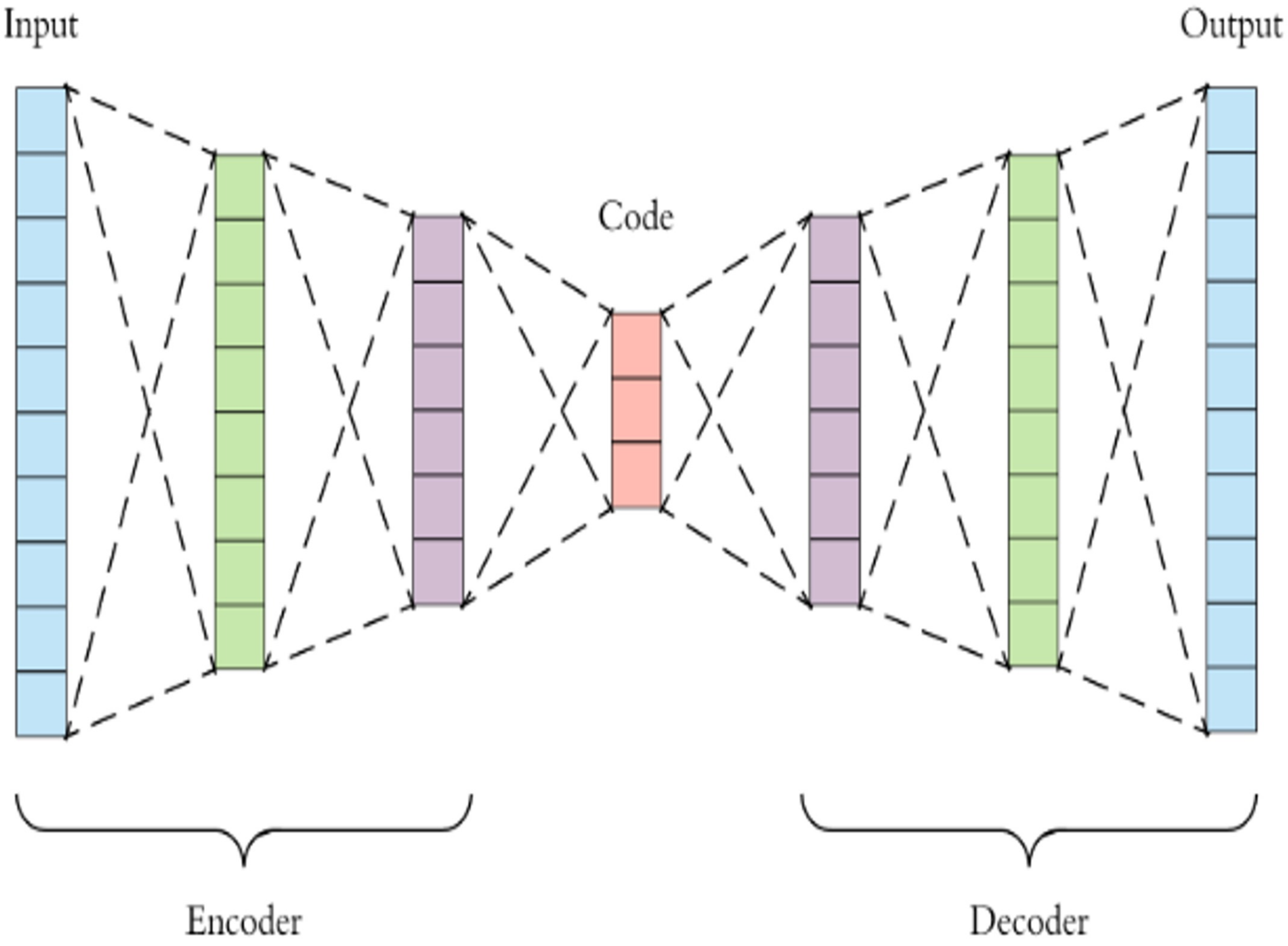
The Shapley value for member 

is given by:

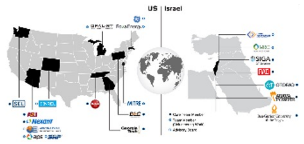
$$\phi_i = \frac{\delta_1 + \delta_3 + \delta_4 + \delta_5 + \delta_6 + \delta_7 + \delta_8}{8}$$

https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3

Explaining anomalies detected by autoencoders using Shapley additive explanations

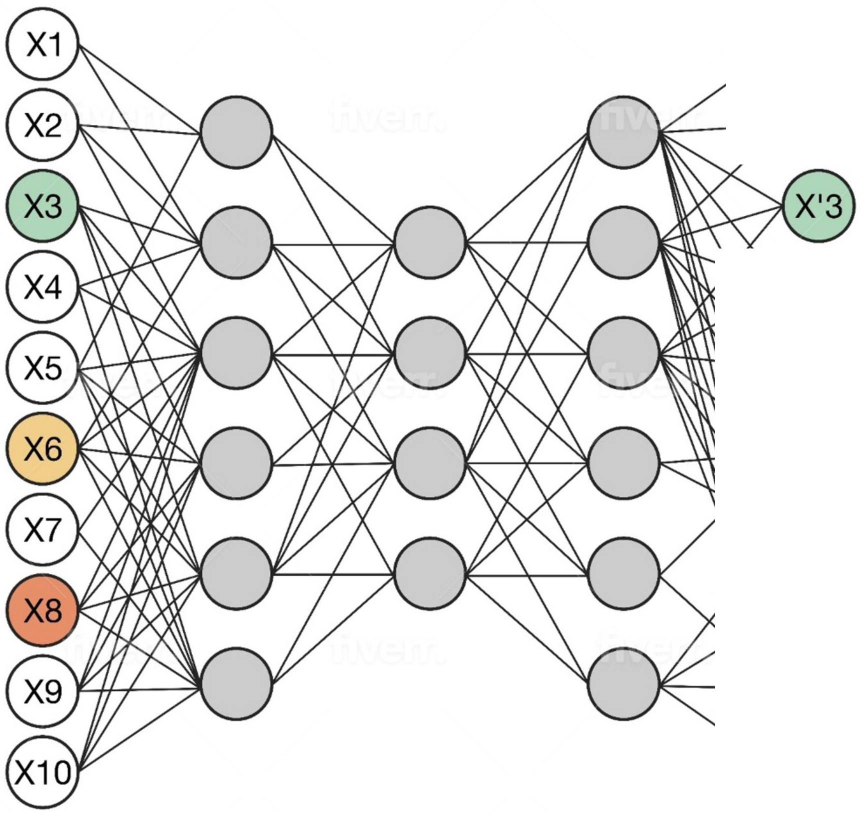


Explaining anomalies detected by autoencoders using Shapley additive explanations

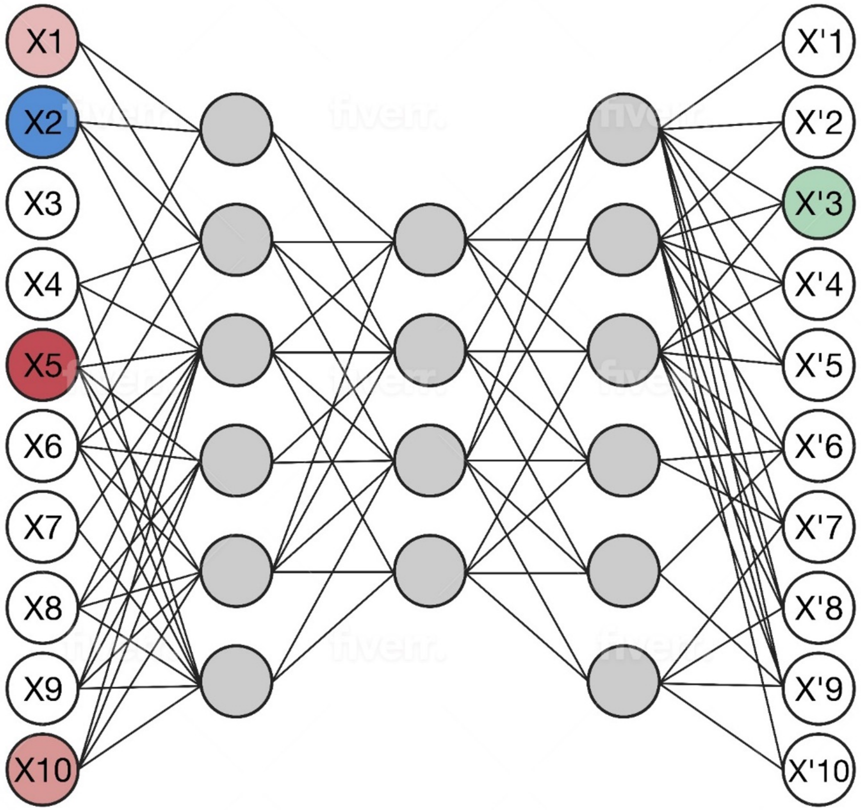


gewaehrleistung_id	mse_all_columns	squared_error_gwk_fremdlohn	squared_error_kilometerstand_km	squared_error_ANZ_FL	squared_error_g
396447471	0.00389	0.00001	0.01188	0.00001	
426290505	0.00328	0.00000	0.02527	0.00000	
397583333	0.00231	0.00000	0.00525	0.00000	
448687344	0.00214	0.00001	0.02209	0.00001	
393807169	0.00198	0.00012	0.00899	0.00012	
446651339	0.00195	0.00001	0.00000	0.00001	
396791336	0.00184	0.00011	0.00142	0.00010	
430558611	0.00180	0.00001	0.01478	0.00001	
450577118	0.00176	0.00000	0.00340	0.00000	
447341408	0.00166	0.00010	0.00001	0.00009	
431233221	0.00162	0.00001	0.01932	0.00001	
389803933	0.00161	0.00006	0.00158	0.00006	
443623241	0.00160	0.00007	0.00000	0.00007	
392789926	0.00151	0.00006	0.01302	0.00006	
447342481	0.00141	0.00000	0.00566	0.00000	

Explaining anomalies detected by autoencoders using Shapley additive explanations



Explaining anomalies detected by autoencoders using Shapley additive explanations



ICNL Data

- Energy production using 3 turbines
- Composed of:
  - sensors readings (every second)
  - network traffic (can be multiple in a second)

```
ts  uid id.orig_h   id.orig_p   id.resp_h   id.resp_p   is_orig source_h    source_p    destination_h   destination_p   rosctr_code rosctr_name pdu_reference   function_code   function_name   subfunction_code    subfunction_name    error_clas
time    string  addr    port    addr    port    bool    addr    port    addr    port    count   string  count   string  string  string  string  string  string
305.049272  CHbGTF39wmjKnKW2s9  192.168.10.65   49872   192.168.10.81   102 F   192.168.10.81   102 192.168.10.65   49872   3   ACK-Data    51445   0x04    Read Variable   -   -   No error    0x00
305.050077  CHbGTF39wmjKnKW2s9  192.168.10.65   49872   192.168.10.81   102 T   192.168.10.65   49872   192.168.10.81   102 1   Job-Request 51701   0x04    Read Variable   -   -   -   -
305.051360  CHbGTF39wmjKnKW2s9  192.168.10.65   49872   192.168.10.81   102 F   192.168.10.81   102 192.168.10.65   49872   3   ACK-Data    51701   0x04    Read Variable   -   -   No error    0x00
305.051962  CHbGTF39wmjKnKW2s9  192.168.10.65   49872   192.168.10.81   102 T   192.168.10.65   49872   192.168.10.81   102 1   Job-Request 51957   0x04    Read Variable   -   -   -   -
305.052825  CHbGTF39wmjKnKW2s9  192.168.10.65   49872   192.168.10.81   102 F   192.168.10.81   102 192.168.10.65   49872   3   ACK-Data    51957   0x04    Read Variable   -   -   No error    0x00
305.053334  CHbGTF39wmjKnKW2s9  192.168.10.65   49872   192.168.10.81   102 T   192.168.10.65   49872   192.168.10.81   102 1   Job-Request 52213   0x04    Read Variable   -   -   -   -
305.055601  CHbGTF39wmjKnKW2s9  192.168.10.65   49872   192.168.10.81   102 F   192.168.10.81   102 192.168.10.65   49872   3   ACK-Data    52213   0x04    Read Variable   -   -   No error    0x00
305.056436  CHbGTF39wmjKnKW2s9  192.168.10.65   49872   192.168.10.81   102 T   192.168.10.65   49872   192.168.10.81   102 1   Job-Request 52469   0x04    Read Variable   -   -   -   -
305.058720  CHbGTF39wmjKnKW2s9  192.168.10.65   49872   192.168.10.81   102 F   192.168.10.81   102 192.168.10.65   49872   3   ACK-Data    52469   0x04    Read Variable   -   -   No error    0x00
305.059571  CHbGTF39wmjKnKW2s9  192.168.10.65   49872   192.168.10.81   102 T   192.168.10.65   49872   192.168.10.81   102 1   Job-Request 52725   0x04    Read Variable   -   -   -   -
305.061112  CHbGTF39wmjKnKW2s9  192.168.10.65   49872   192.168.10.81   102 F   192.168.10.81   102 192.168.10.65   49872   3   ACK-Data    52725   0x04    Read Variable   -   -   No error    0x00
305.061569  CHbGTF39wmjKnKW2s9  192.168.10.65   49872   192.168.10.81   102 T   192.168.10.65   49872   192.168.10.81   102 1   Job-Request 52981   0x04    Read Variable   -   -   -   -
305.063157  CHbGTF39wmjKnKW2s9  192.168.10.65   49872   192.168.10.81   102 F   192.168.10.81   102 192.168.10.65   49872   3   ACK-Data    52981   0x04    Read Variable   -   -   No error    0x00
```

Quick Data overview

Anomaly_Type	Counts	Relative_Percentage
0	87461	88.48
1	5875	5.94
2	5511	5.58

0 = Normal
1 = Operational failures
2 = Cyber Attacks

ICNL Data

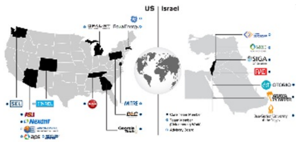
Testbed

- Duration of 7 days (8 hours a day)
- Activity recorded: Normal/Operational fault/Cyber attack

Normal / Normal + Operational failures / Normal + Operational failures + Cyber Attacks

Day	Date	Recorded Physical(csv)	Recorded Network(pcap)	PCAP Size (GB)
1	12/06/2023	V	-	-
2	13/06/2023	V	V	7.1
3	14/06/2023	V	V	5.2
4	15/06/2023	V	V	5.4
5	18/06/2023	V	V	3.6
6	19/06/2023	V	V	4.0
7	20/06/2023	-	V	3.5

Data



ICNL testbed - Network

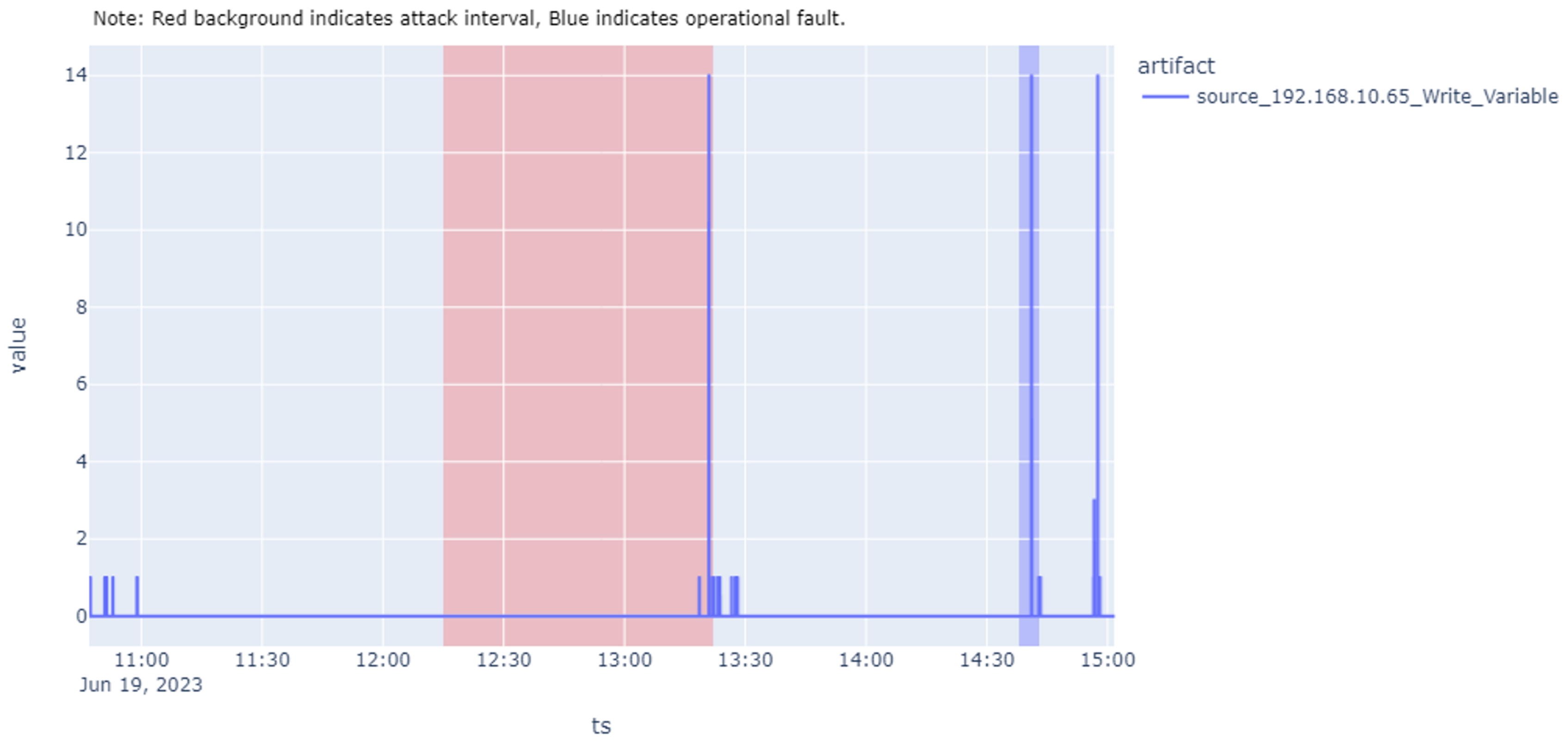
- Network Components: HMI, Engineering station, main PLC, secondary PLC, sensors and actuators
- Industrial protocol used: S7comm by siemens
- Majority of functions are read requests, few write requests are present as well
- Aggregated Features(seconds):
  - Protocols- tcp/udp/s7comm/etc..
  - Known IP’s (HMI/Eng. Station/PLC’s/etc..) as source/destination
  - Unknown IP’s as source/destination
  - Read requests
  - Write requests

  For each aggregated feature we create sum features for the last 10,30,60,300,1800 seconds.

Data

ICNL testbed - Network

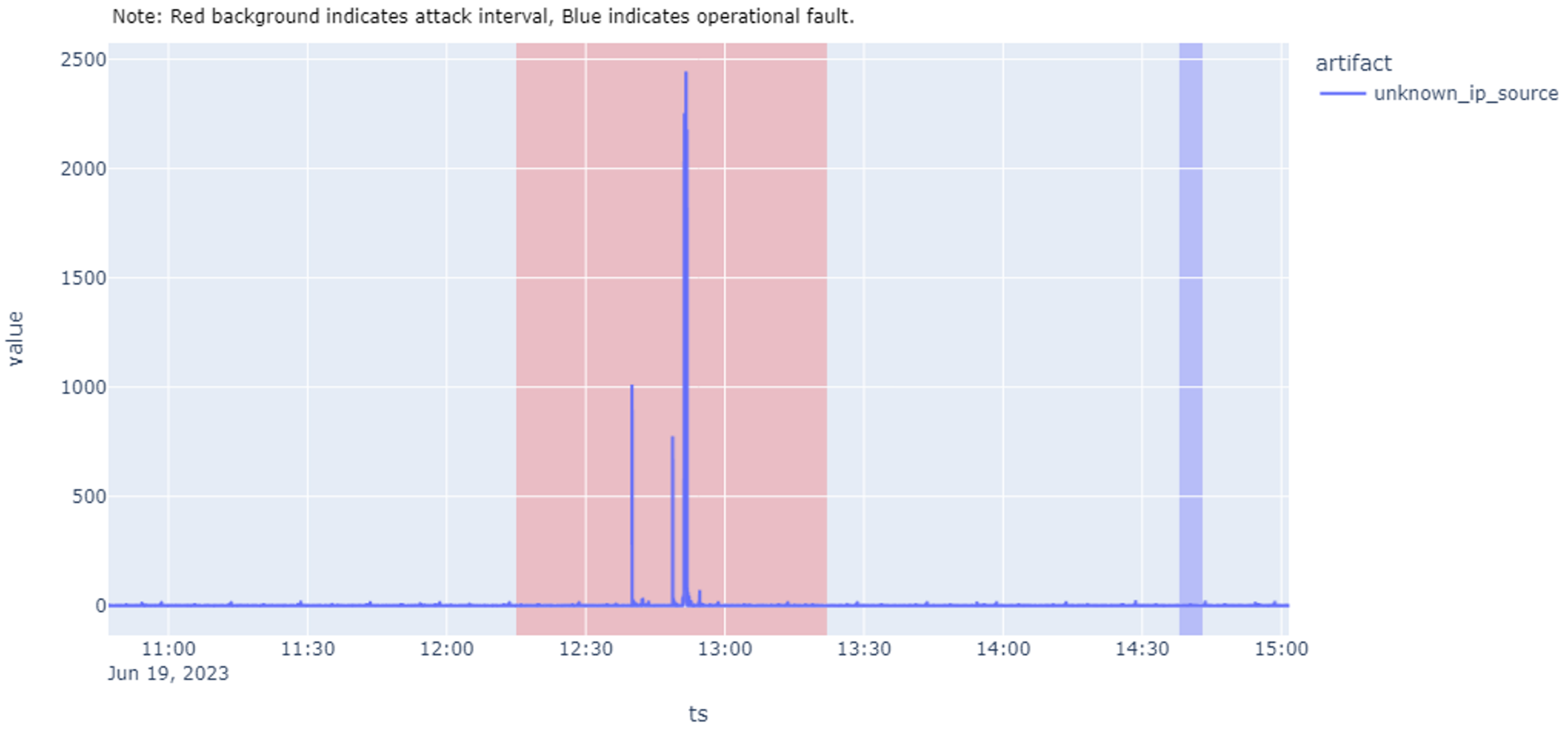
- Write commands sent from the Engineering Station computer may suggest about operational fault or cyber attack



Data

ICNL testbed - Network

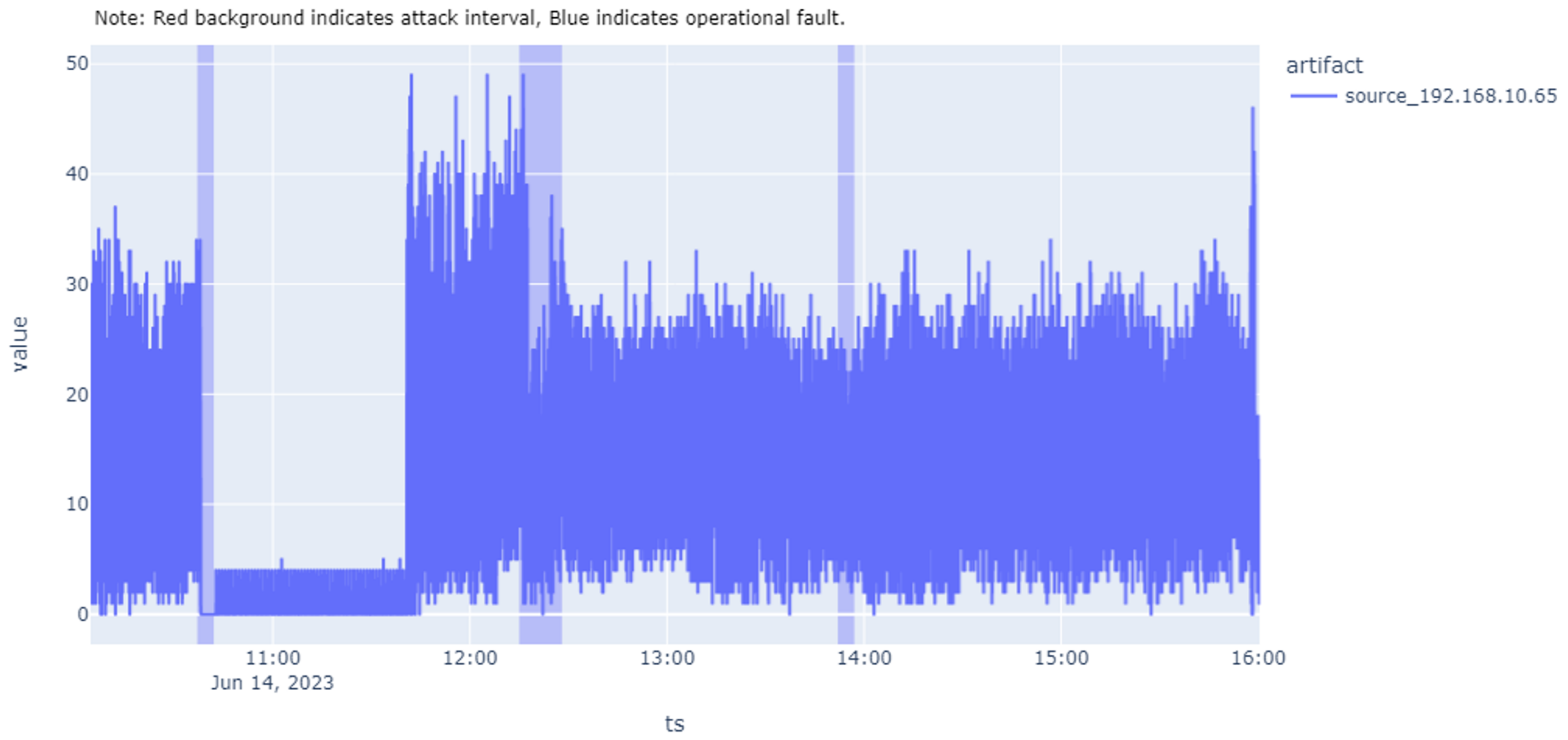
- Packets sent from an unknown IP source may suggest a potential cyber attack



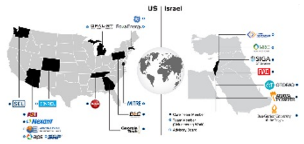
Data

ICNL testbed - Network

- Number of packets sent from the Engineering Station computer may suggest some of the operational faults

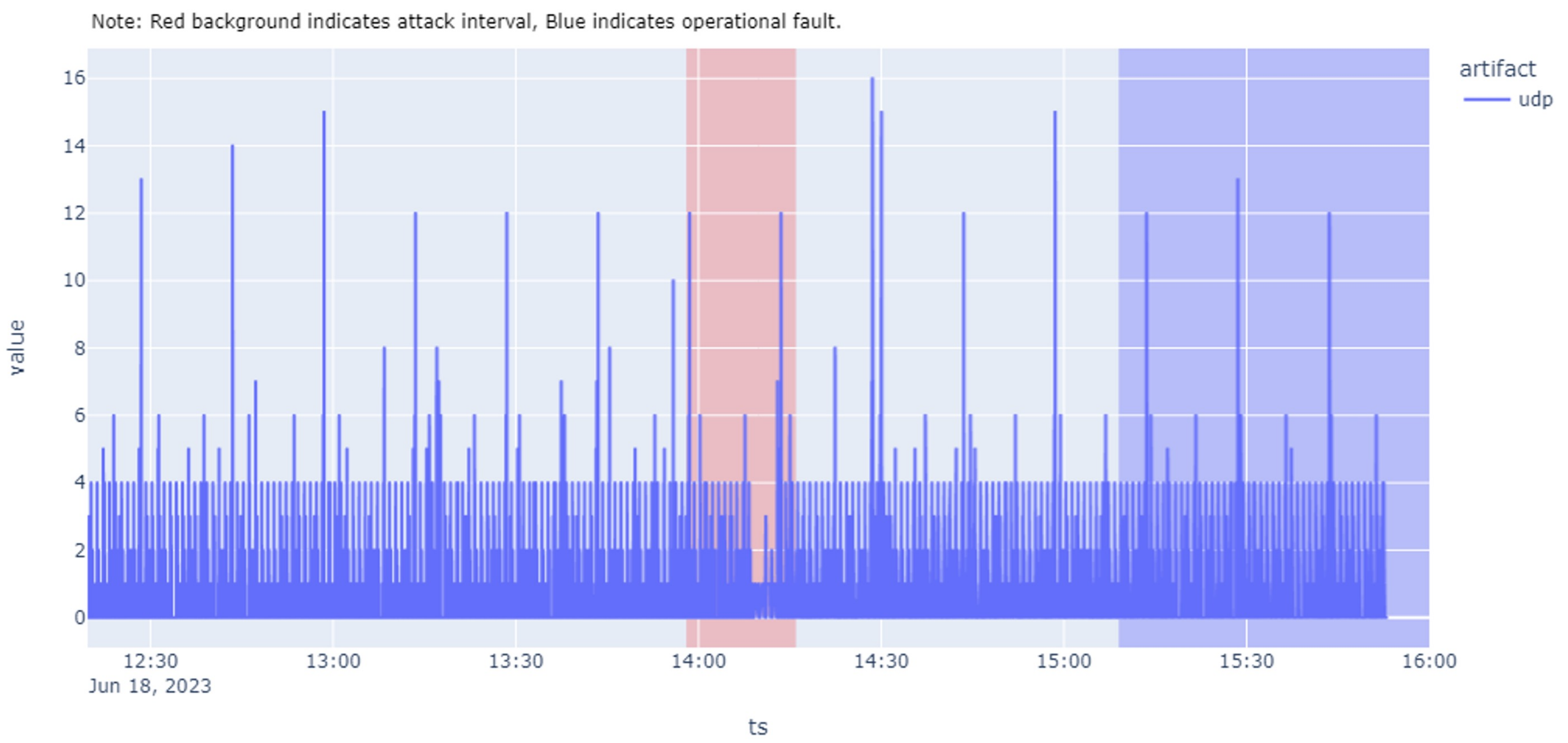


Data



ICNL testbed - Network

- Number of UDP protocol packets may suggest a potential cyber attack



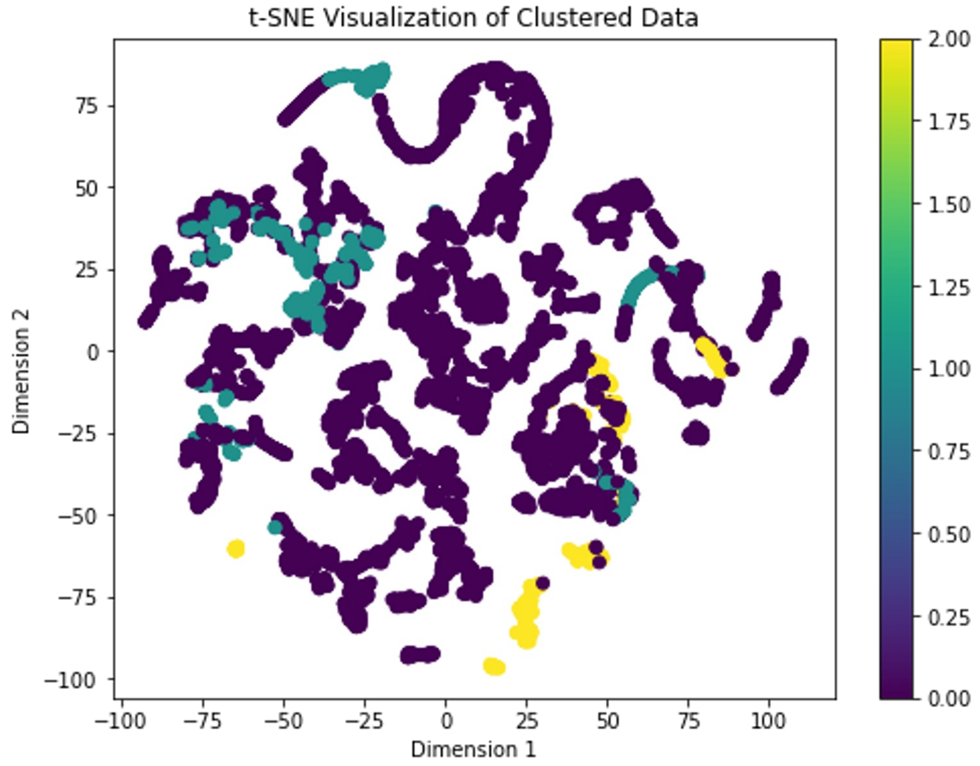
Experiments

- Goals
  - To find a connection in the explanations of anomalies revealed by different models
  - Reduce false positive anomalies using the connection between the explanations
- Challenges
  - The quality of the explanations is limited by the quality of the anomaly detector
  - How to make a decision if an event is anomalous or not – there are many parameters for the decision (number of explaining features, how many models from the ensemble need to agree with each other)

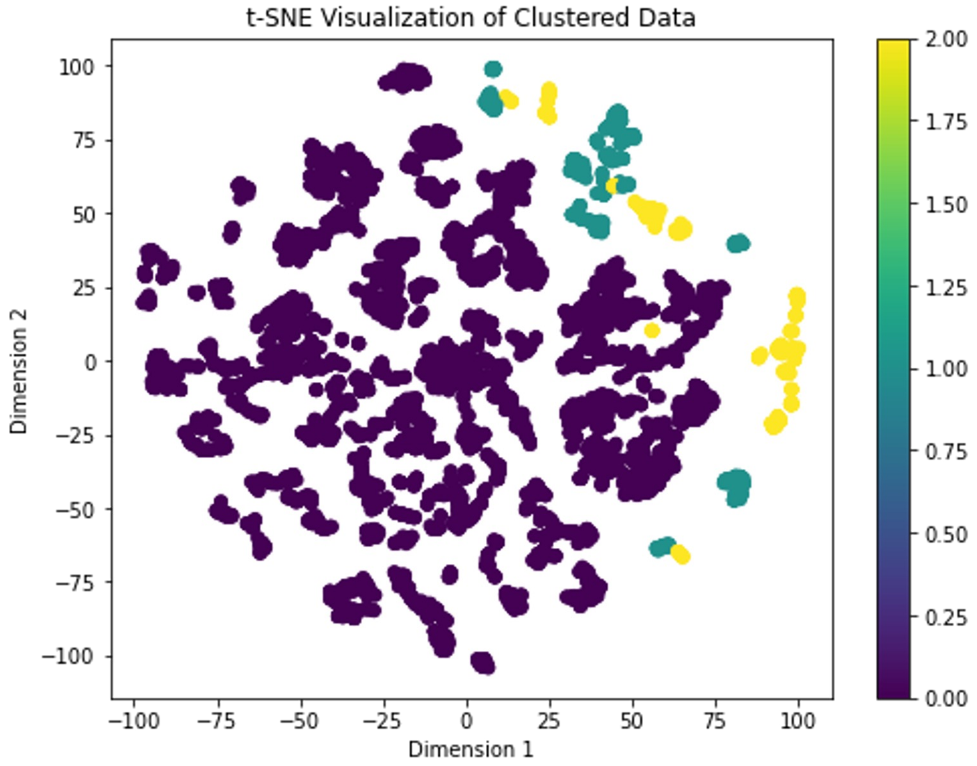
Random forest model

- Model parameters:
  - n_estimators=250
  - Test size = 10%, train size = 90% (split random seed = 42)
- Model evaluation metrics (10-fold stratified cross validation on the train part):
  - Average Accuracy: 0.99
  - Average Precision: 0.9996070343762227
  - Average Recall: 0.9996065685385828
  - Average AUC-ROC (OvO): 0.9999991250210561

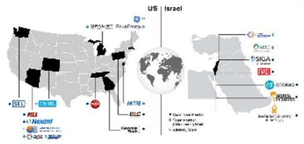
Tsne on the original test set records



Tsne on shap values explaining the test set



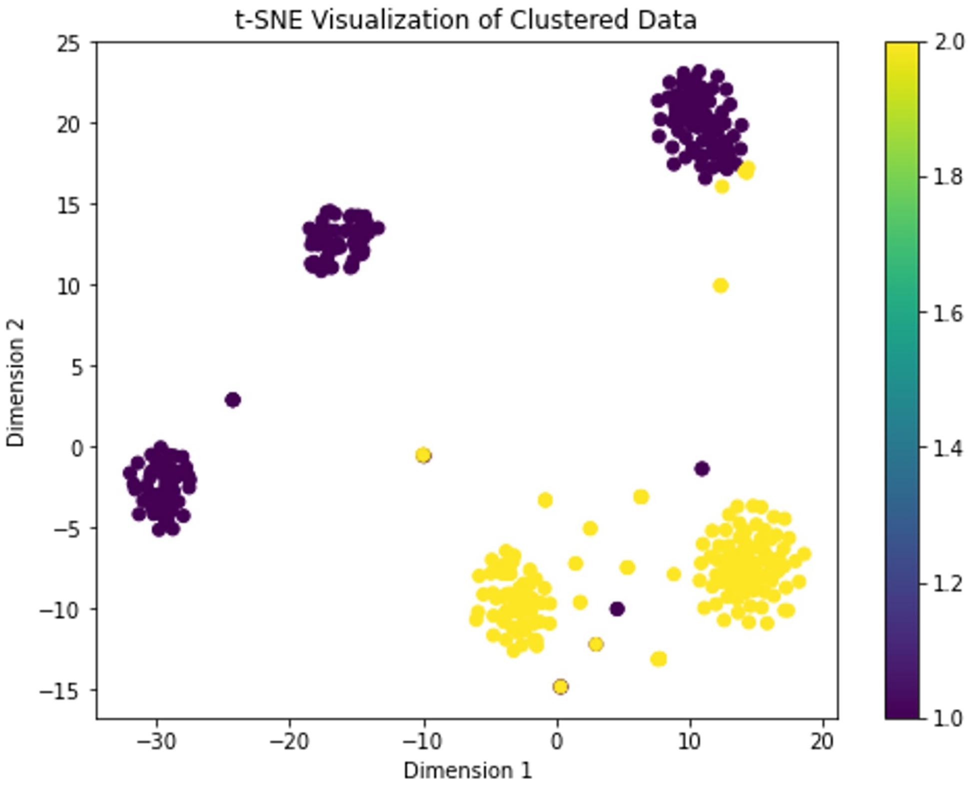
Autoencoder



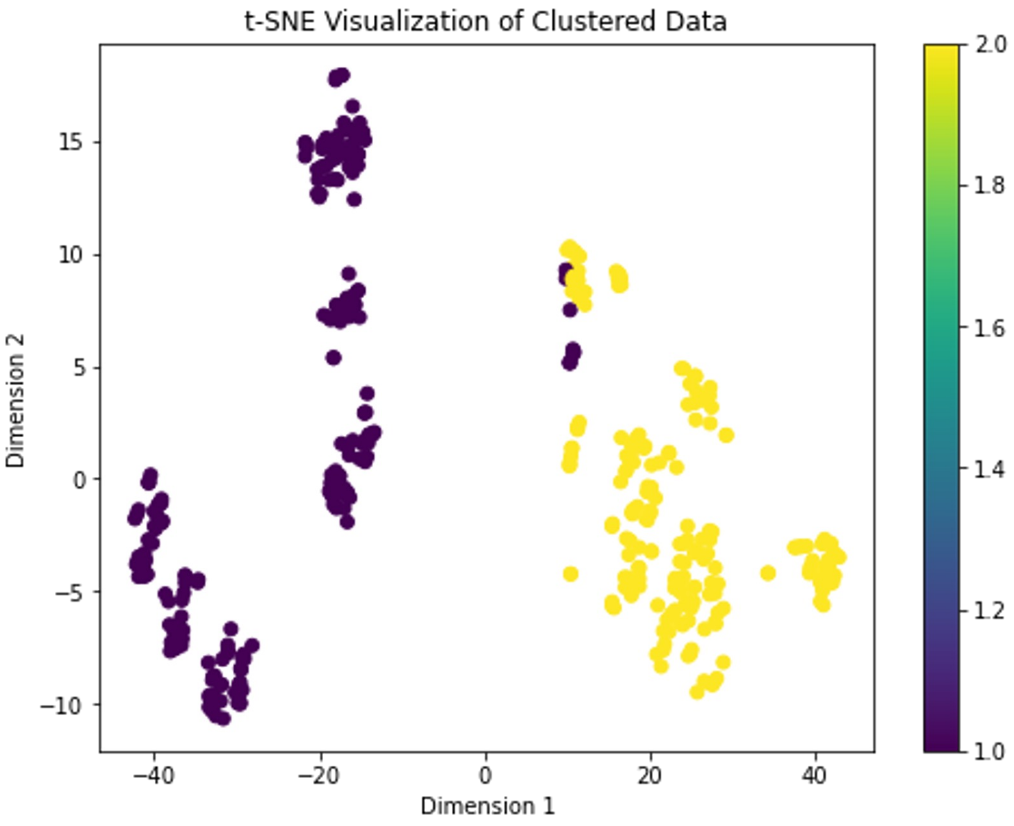
- Model training parameters:
  - nb_epoch=4000
  - batch_size=30000
  - All the data (98,847 records, normalized)
- Kernel's SHAP explainer parameters:
  - nsamples=500
  - num_of_features = 1

Original records vs. Explanations

shap values tsne



original anomaly data tsne



Future work

- Continue analyzing the results
- Add physical data
- Run on another dataset