Cyber-Physical Data Analytics to Enable Resilient Electric Grid

A. Srivastava, Washington State University

Associate Professor, Washington State University
Director, Smart Grid Demonstration and Research Investigation Lab
Technical Lead, UI-ASSIST Center
Contact: anurag.k.srivastava@wsu.edu

June 19th, 2019

IEEE PES SC Big Data & Analytics for Power Systems Webinar

What is resiliency?

How data analytics relate to resiliency?

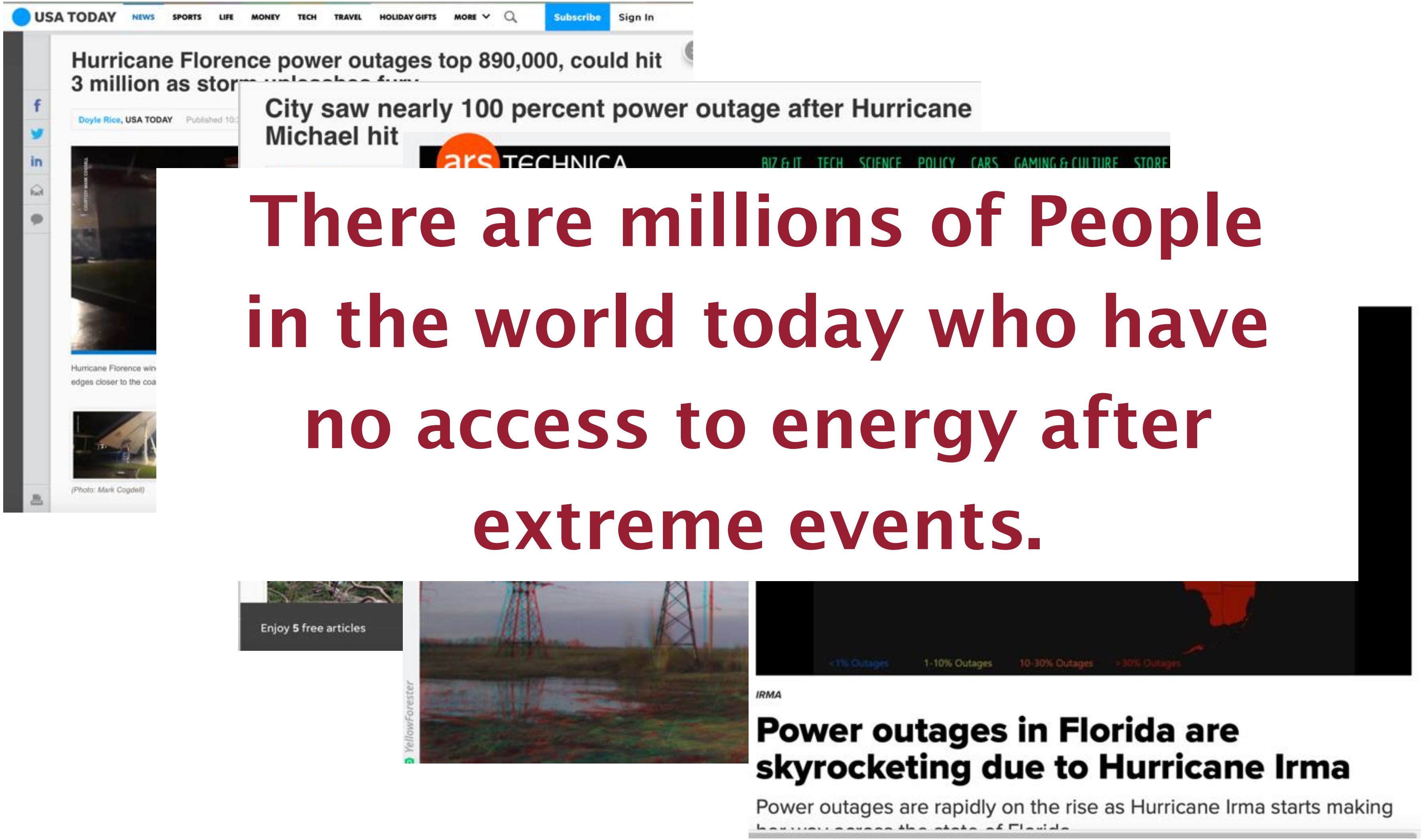
How do we measure and enable resiliency?

How data analytics help and use cases

Learning based on projects: DOE CREDC, NSF FW-HTF, ARPA-E RIAPS, GMLC 1.3.9 Idaho Falls, GMLC: City of Cordova (DOE RADIANCE Project), DOE AGGREGATE, DOE UI-ASSIST

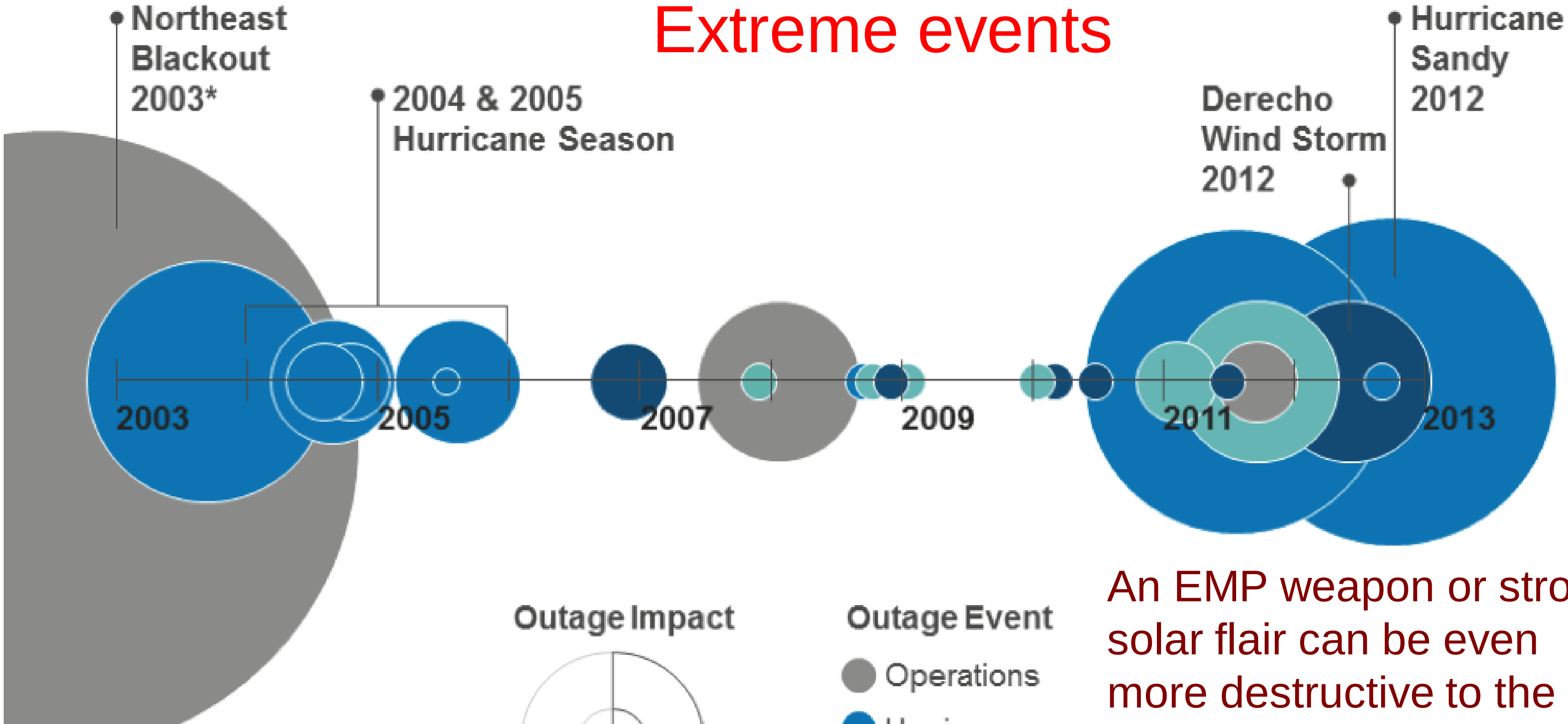
- What is resiliency?
- How data analytics relate to resiliency?
- How do we measure and enable resiliency?
- How data analytics help and use cases

Learning based on projects: DOE CREDC, NSF FW-HTF, ARPA-E RIAPS, GMLC 1.3.9 Idaho Falls, GMLC: City of Cordova (DOE RADIANCE Project), DOE AGGREGATE, DOE UI-ASSIST



There are millions of People in the world today who have no access to energy after extreme events.

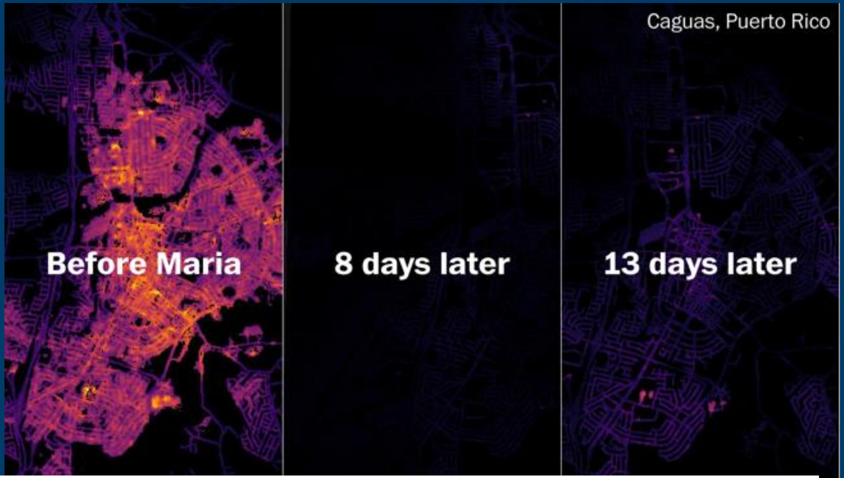
Extreme events

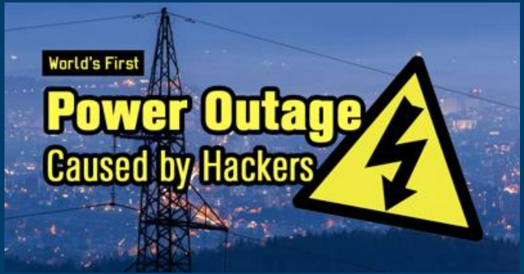


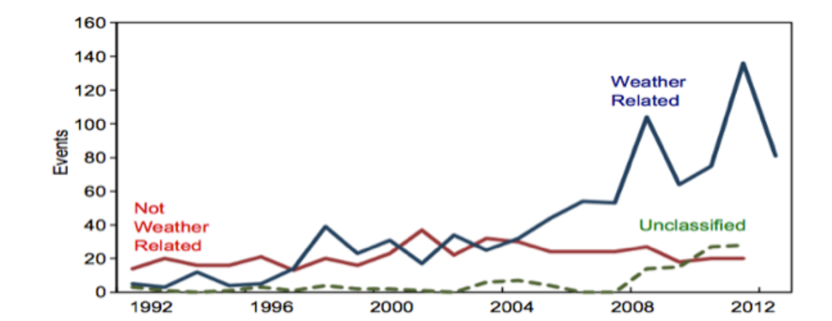
An EMP weapon or strong solar flair can be even more destructive to the grid

Not a single/ double contingency (as in security)

Power Grid: Reliable but Not Resilient



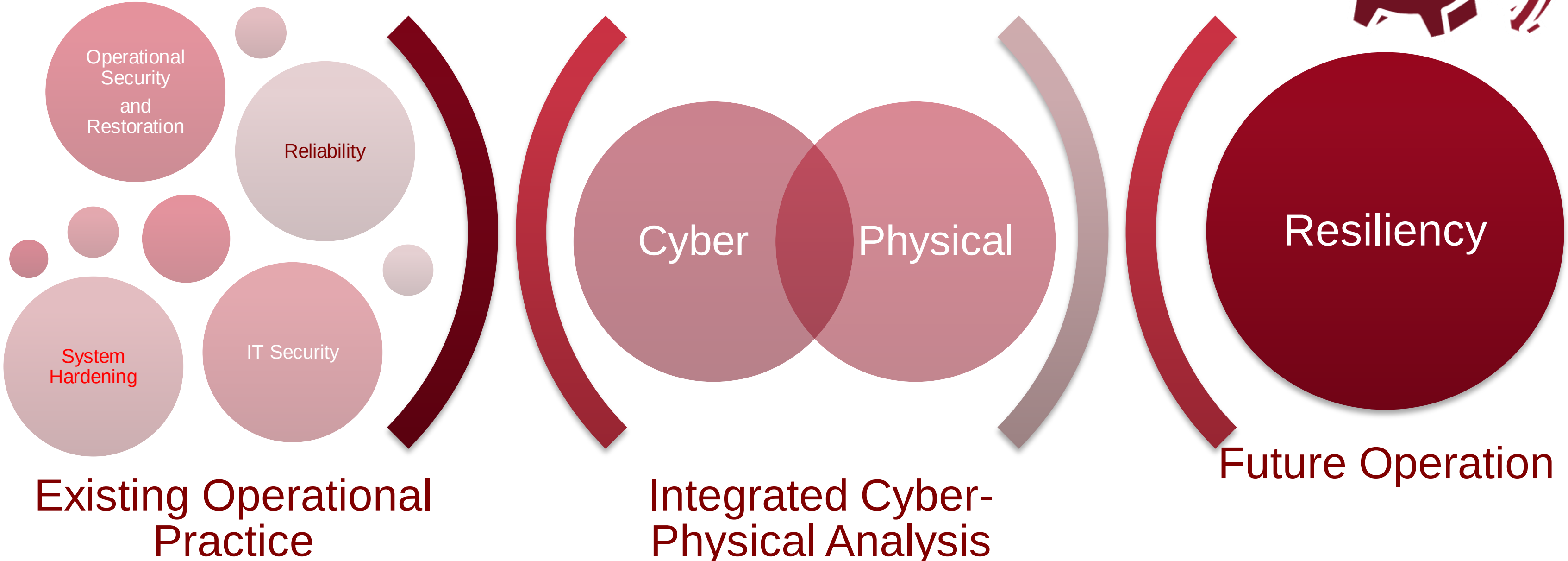




Weather-related outages in US between 1992-2012

Electric Grid Resiliency

Resilience: The ability to supply its critical load through (and in spite of) extreme contingencies and low resource availability



W

Withstand any sudden inclement weather or human attack on the infrastructure.

R

Respond quickly, to restore balance in the community as quickly as possible, after an inevitable attack.

A

Adapt to abrupt and new operating conditions, while maintaining smooth functionality, both locally and globally.

P

Predict or **Prevent** future attacks based on patterns of past experiences, or reliable forecasts.

What is resiliency?

How data analytics relate to resiliency?

How do we measure and enable resiliency?

How data analytics help and use cases

Learning based on projects: DOE CREDC, NSF FW-HTF, ARPA-E RIAPS, GMLC 1.3.9 Idaho Falls, GMLC: City of Cordova (DOE RADIANCE Project), DOE AGGREGATE, DOE UI-ASSIST

Planning

- Design and long term planning
- Short term planning

Switch location?

DER/ Storage location?

Operation

- Situational awareness
- Decision support

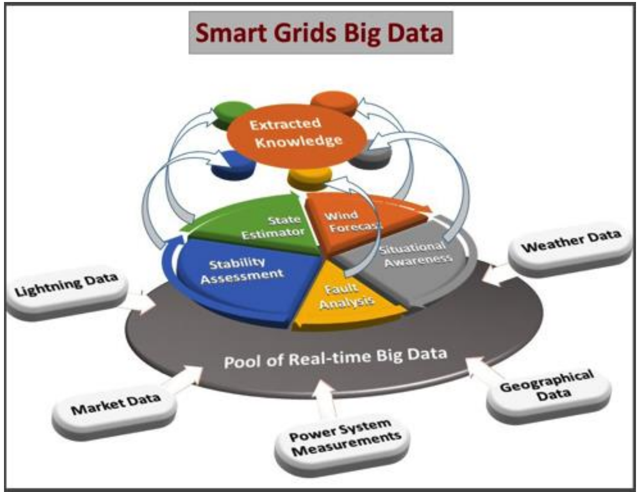
Weather Events?

Cyber Attack?

Power Systems Data: Example of fixed data

Fixed Data (Assets)

- 7,500 generation plants
- 75,000 substations
- 300,000 miles transmission (100,000 lines and transformers)
- 2.2 million miles distribution (1 million distribution feeders)
- 300 million customers



Credit: Prof Anjan Bose, WSU, TAMU NSF SPOKE

Data Collection by PMUs: Example of Operational Data

- PMU sampling rates: 30 per second
- Assume 100 values per second

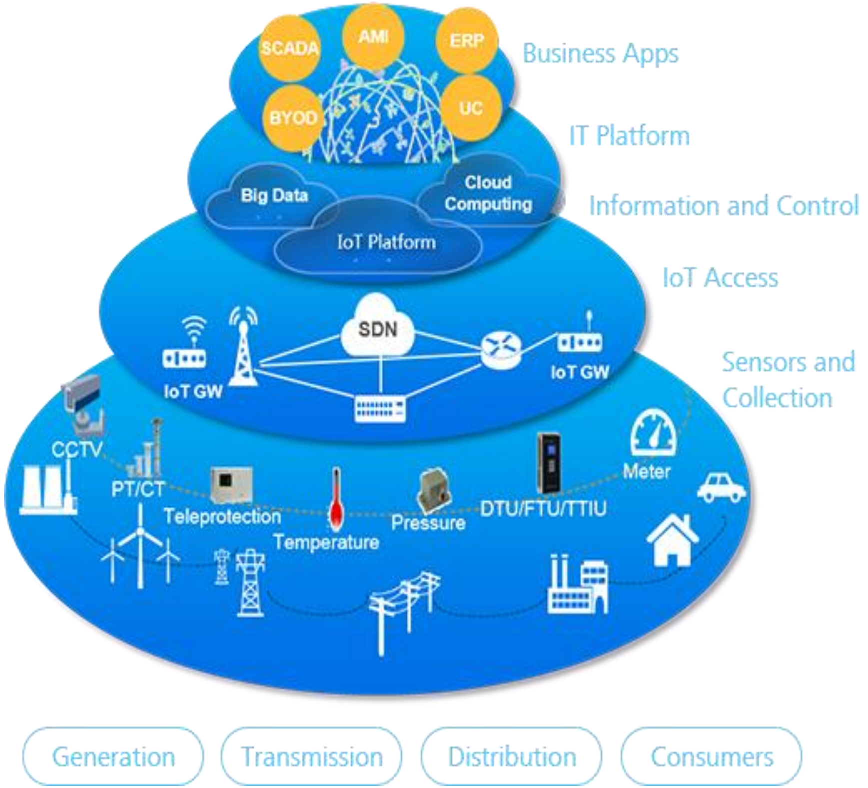
If we assume all 100 points in a sub are PMUs

- Average data rate per sub is 10K/sec
- Average data rate for the total of 100 subs in a BA is 1M/sec
- Average data rate for the RC is then 10M/sec

Data Analytics Needed for Making Sense of this Steaming Operational Data for Cyber or Physical Events !!!!

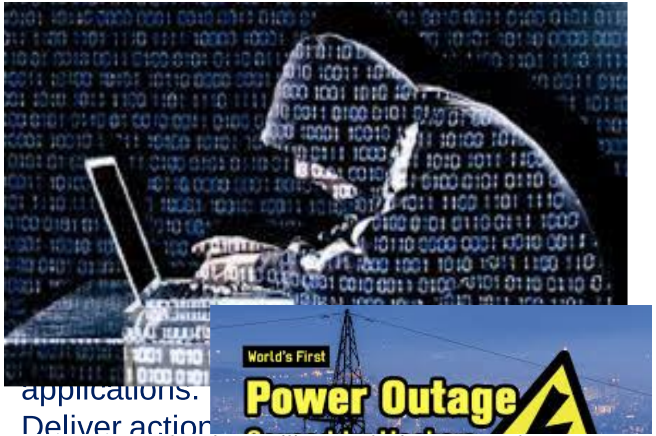
Credit: Prof Anjan Bose, WSU

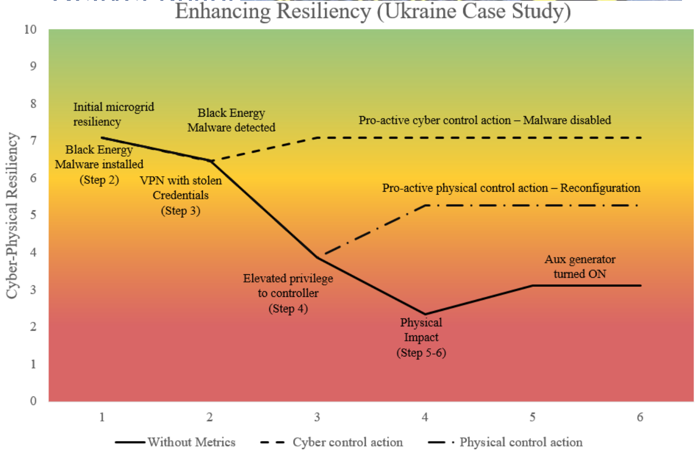
Connecting Data Analytics with Resiliency



???

- applications.
- Deliver action





What is resiliency?

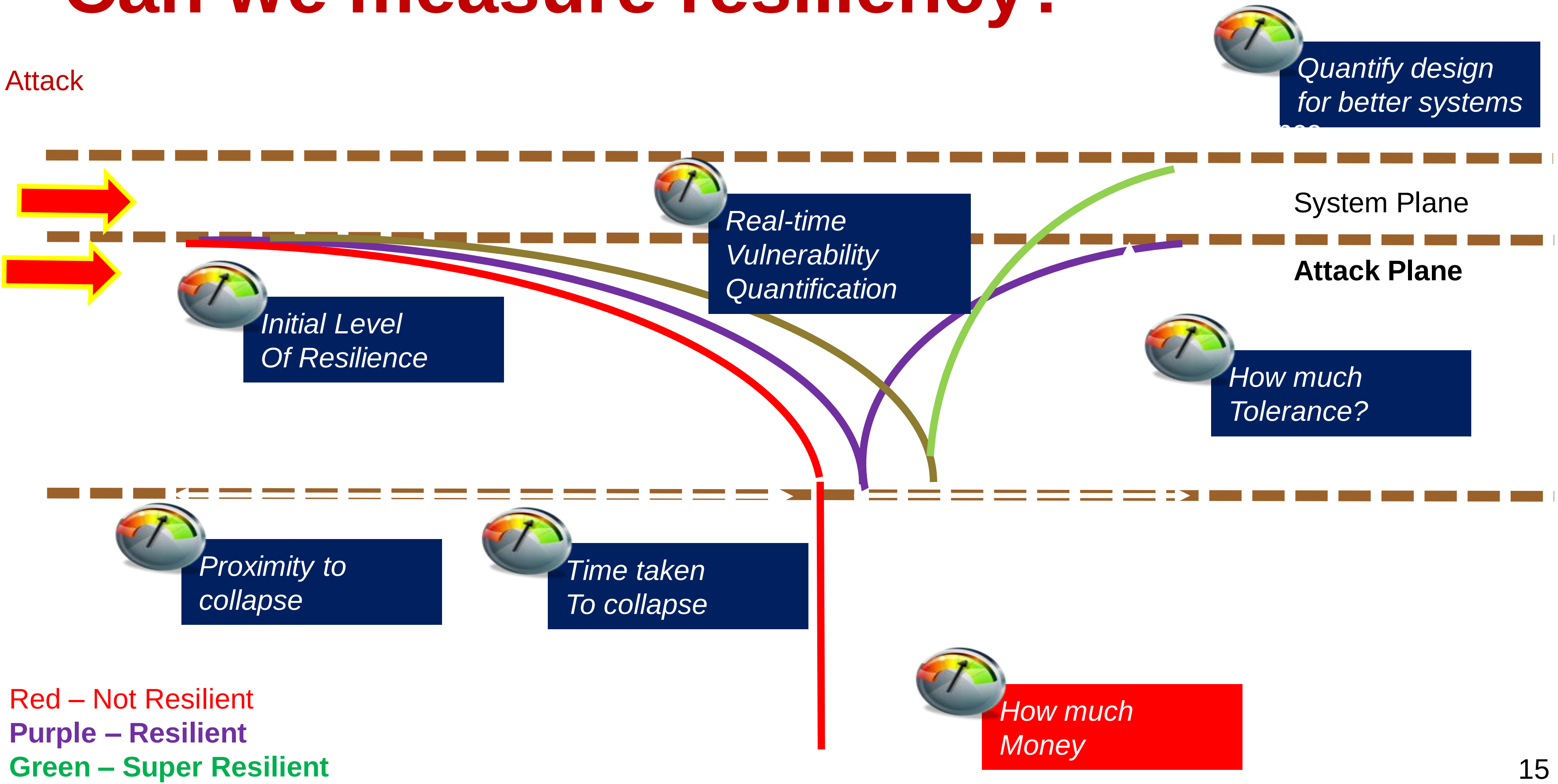
How data analytics relate to resiliency?

How do we measure and enable resiliency?

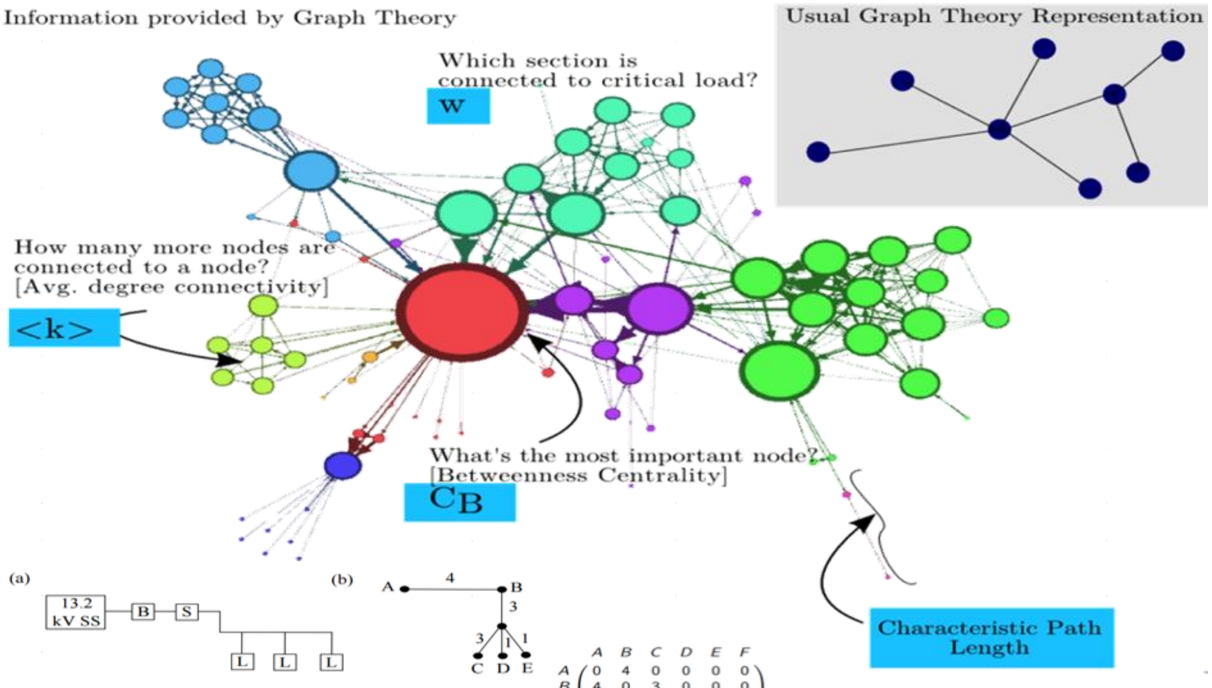
How data analytics help and use cases

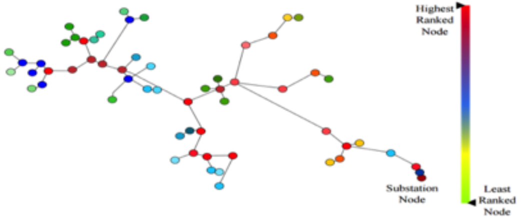
Learning based on projects: DOE CREDC, NSF FW-HTF, ARPA-E RIAPS, GMLC 1.3.9 Idaho Falls, GMLC: City of Cordova (DOE RADIANCE Project), DOE AGGREGATE, DOE UI-ASSIST

Can we measure resiliency?



Multi-criteria Decision for Physical Resiliency





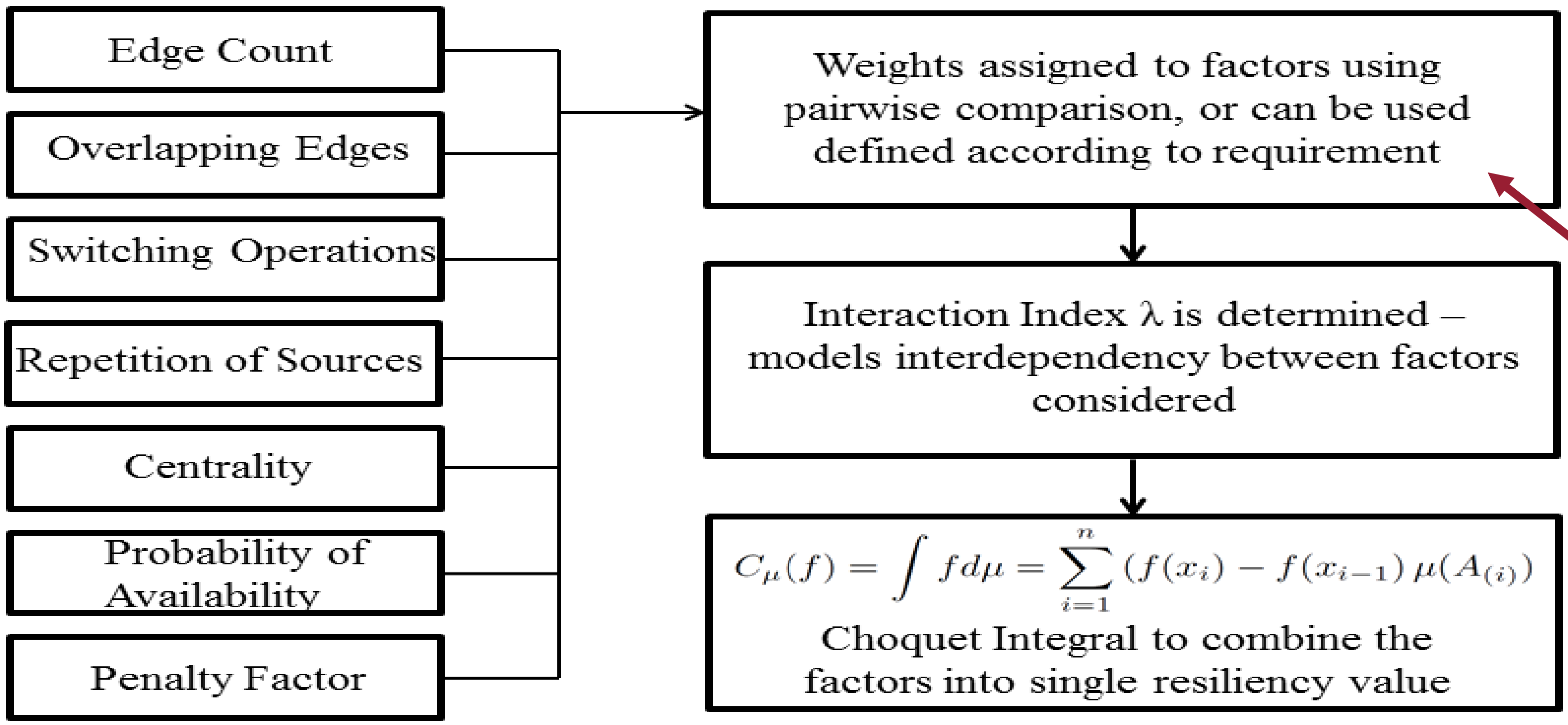
- Analytical Hierarchical Process
- Topology Parameters
- Weather Parameters
- Infrastructure Parameter

$$V = [A_{f_c}\ B_D\ C_{C_B}\ D_{I_G}\ E_{C_n}\ F_{\Delta\lambda}\ G_{\lambda_2}]^T$$

$$\rho(i,j)' = \frac{\rho(i,j) - \min_{i=1}^{\eta}(\rho(i,j))}{\max_{i=1}^{\eta}(\rho(i,j)) - \min_{i=1}^{\eta}(\rho(i,j))}$$

$$\Re_\tau = \sum_{j=1}^{\eta} V_j \rho(i,j)'$$

Overview of resiliency quantification process



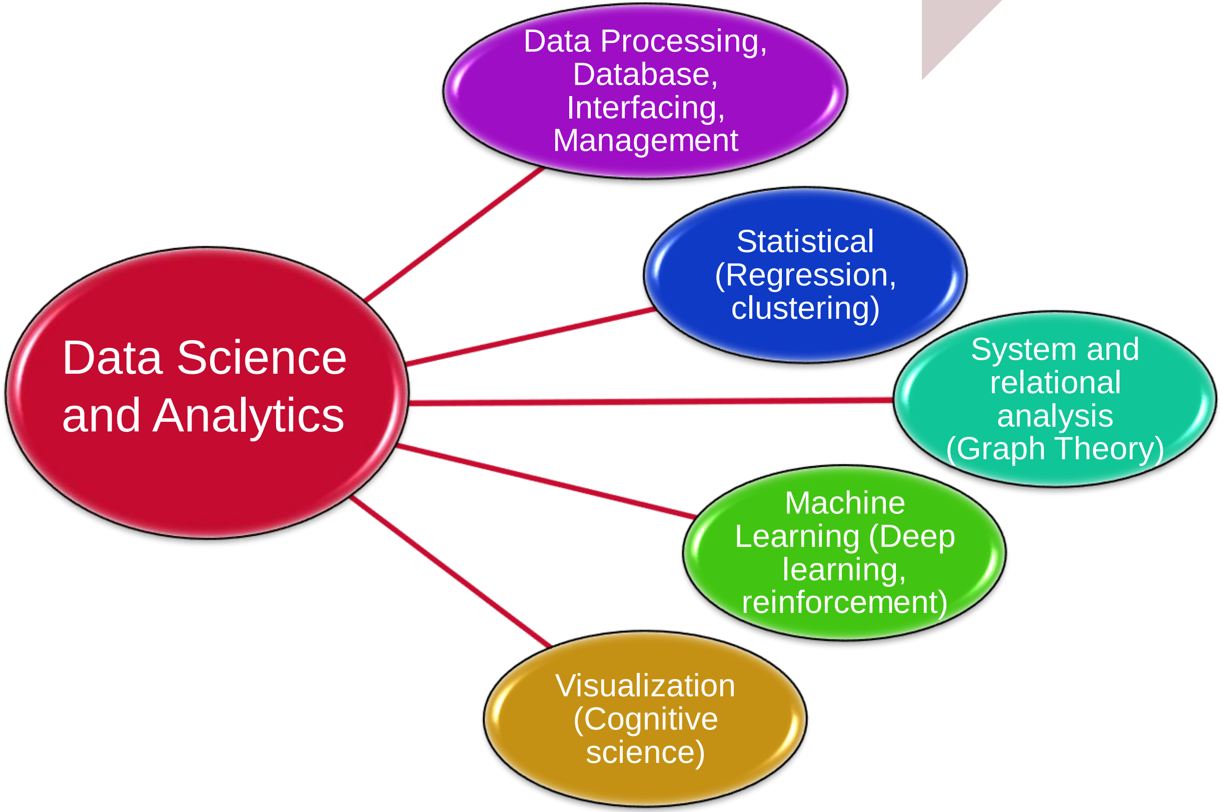
Decision Making Tool

- What is resiliency?
- How data analytics relate to resiliency?
- How do we measure and enable resiliency?
- **How data analytics help and use cases**

Learning based on projects: DOE CREDC, NSF Microgrid, DOE ARPA-E, GMLC 1.3.9 Idaho Falls, GMLC: City of Cordova (DOE RADIANCE Project), DOE AGGREGATE

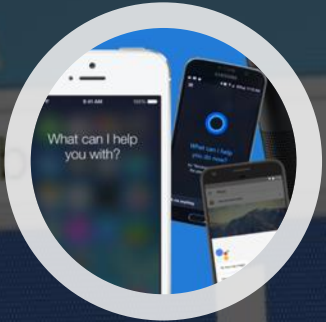
Resiliency requires knowing the threat → Situational Awareness is necessary to take decision → Data analytics helps in enhanced awareness

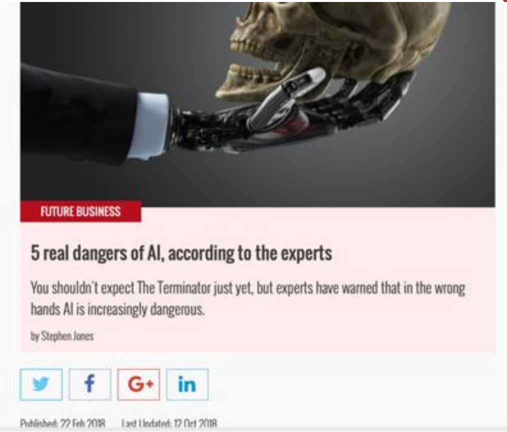
- Predicts the future based on past patterns.
- Explores and examines data from multiple disconnected sources.
- Develop new analytical methods and machine learning models.
- Leverage data for relevant applications.
- Deliver actionable insights from the data.
- Store and process the data for insights.
- Design and create data reports using various reporting tools.
- Query database and package data for insights.



Data Science Everywhere

- XBox Face Recognition
- Facebook Face recognition
- Google Home/ Maps Prediction
- Amazon Alexa
- Apple Siri
- Self-driving cars



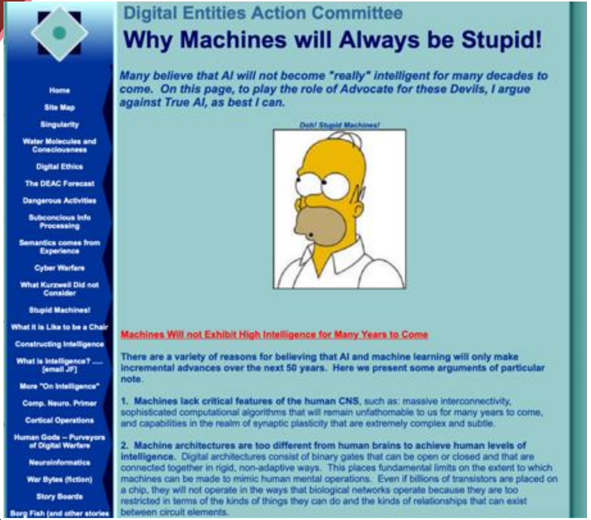

FUTURE BUSINESS

5 real dangers of AI, according to the experts

You shouldn't expect The Terminator just yet, but experts have warned that in the wrong hands AI is increasingly dangerous.

by Stephen Jones

Published: 22 Feb 2018 Last Updated: 12 Oct 2018


Digital Entities Action Committee

Why Machines will Always be Stupid!

Home
Site Map
Singularity
Water Molecules and Consciousness
Digital Ethics
The DEAC Forecast
Dangerous Activities
Subconcious Info Processing
Semantics comes from Experience
Cyber Warfare
What Kurzweil Did not Consider
Stupid Machines!
What it is Like to be a Chair
Constructing Intelligence
What is Intelligence? .... [email JF]
More "On Intelligence"
Comp. Neuro. Primer
Cortical Operations
Human Gods -- Purveyors of Digital Warfare
Neuroinformatics
War Bytes (fiction)
Story Boards
Borg Fish (and other stories

Many believe that AI will not become "really" intelligent for many decades to come. On this page, to play the role of Advocate for these Devils, I argue against True AI, as best I can.

Doh! Stupid Machines!

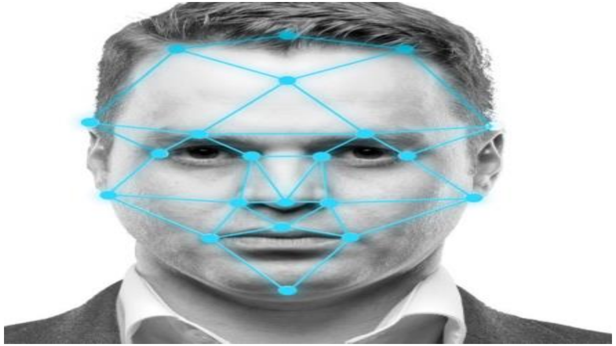
Machines Will not Exhibit High Intelligence for Many Years to Come

There are a variety of reasons for believing that AI and machine learning will only make incremental advances over the next 50 years. Here we present some arguments of particular note.

1. **Machines lack critical features of the human CNS**, such as: massive interconnectivity, sophisticated computational algorithms that will remain unfathomable to us for many years to come, and capabilities in the realm of synaptic plasticity that are extremely complex and subtle.

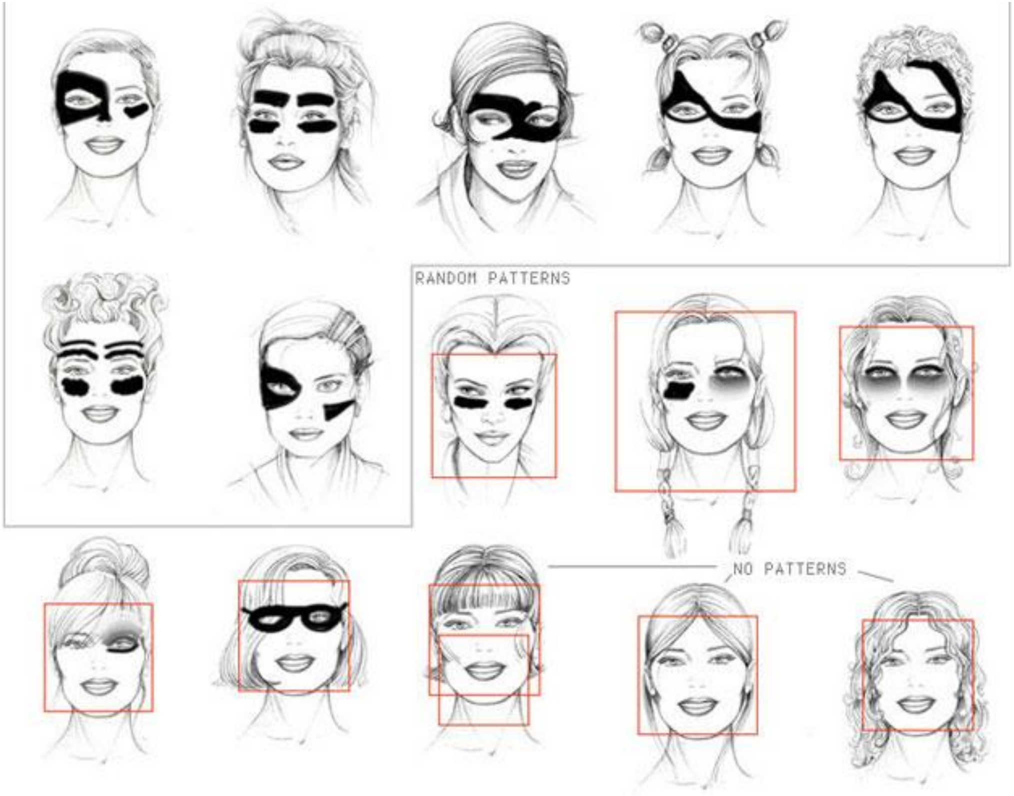
2. **Machine architectures are too different from human brains to achieve human levels of intelligence.** Digital architectures consist of binary gates that can be open or closed and that are connected together in rigid, non-adaptive ways. This places fundamental limits on the extent to which machines can be made to mimic human mental operations. Even if billions of transistors are placed on a chip, they will not operate in the ways that biological networks operate because they are too restricted in terms of the kinds of things they can do and the kinds of relationships that can exist between circuit elements.

What can go wrong with Data Analytics



Low Quality AI

Trained with Bad Data

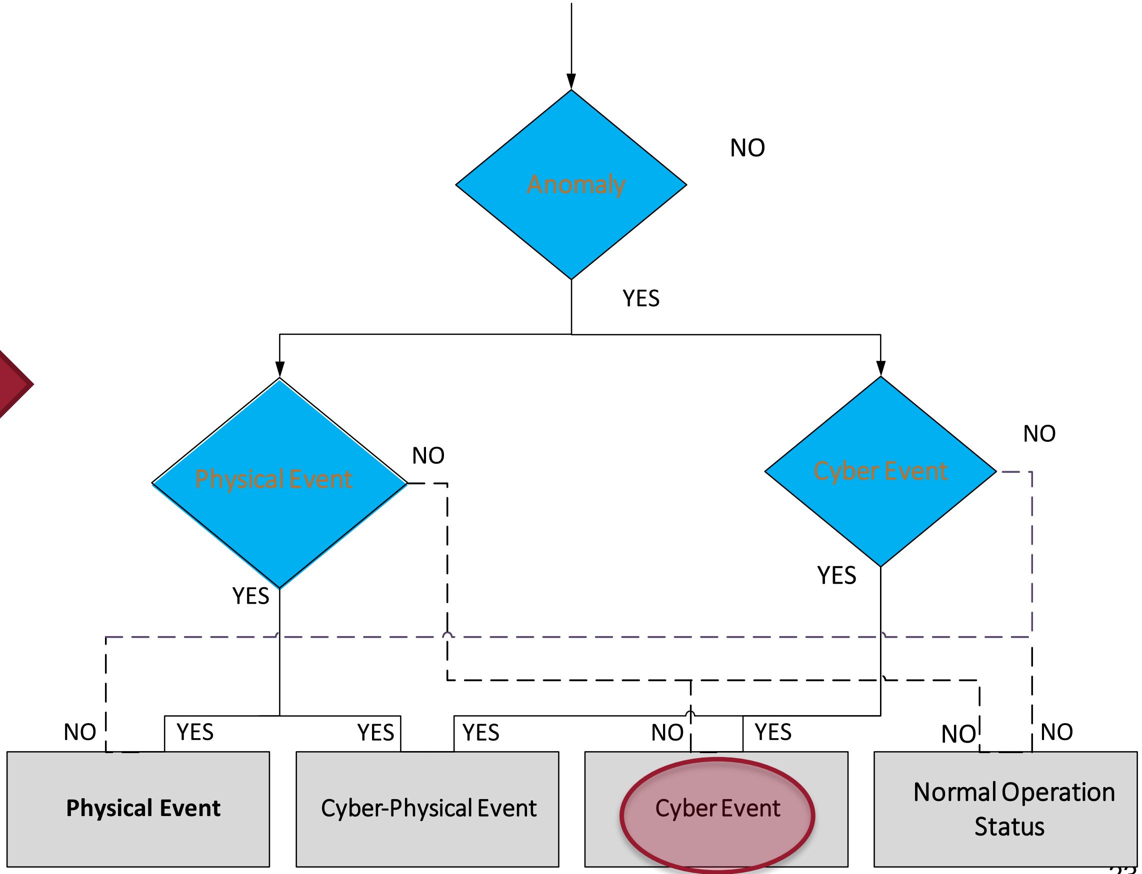


Use Case I: Anomaly Detection, Classification, Event Detection and Root Cause Analysis using PMU Data

Data

- Physical
  - PMU measurements
  - CT/PT measurements
  - Breaker status
  - Relay operations
- Cyber
  - Network data
    - Pcaps, netflows, Ids alerts
  - Hosts
    - Event logs, Ids alerts

???

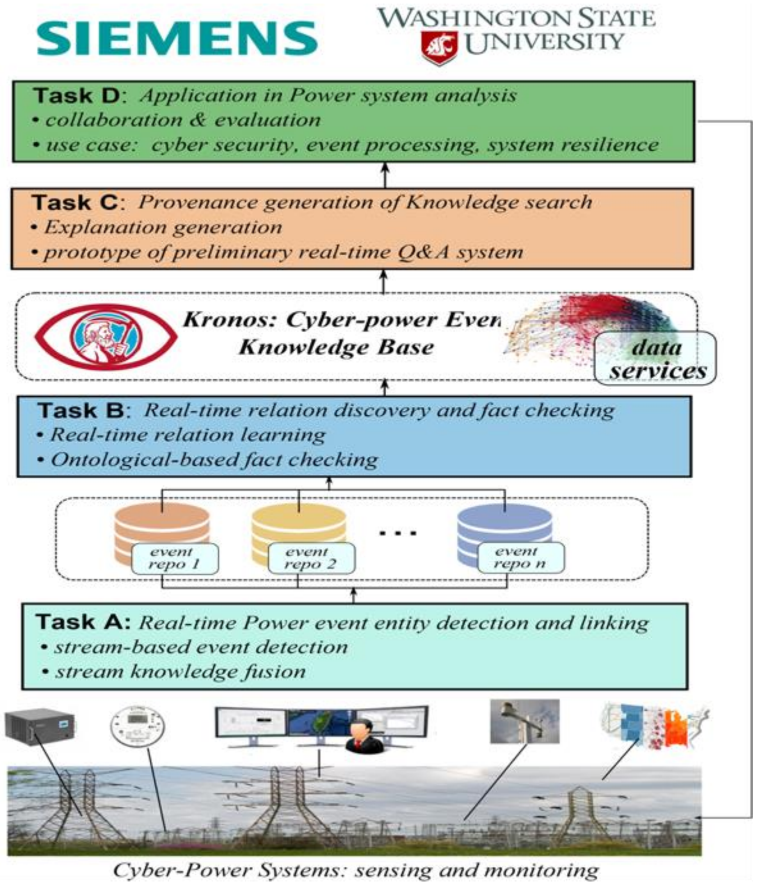


Kronos: Real-Time Power Event on Heterogenous Data Stream

Goal: Build a lightweight Knowledge Base from power events and their semantic & temporal relationships for explainable event prediction and cause analysis, directly from cyber and power data streams.

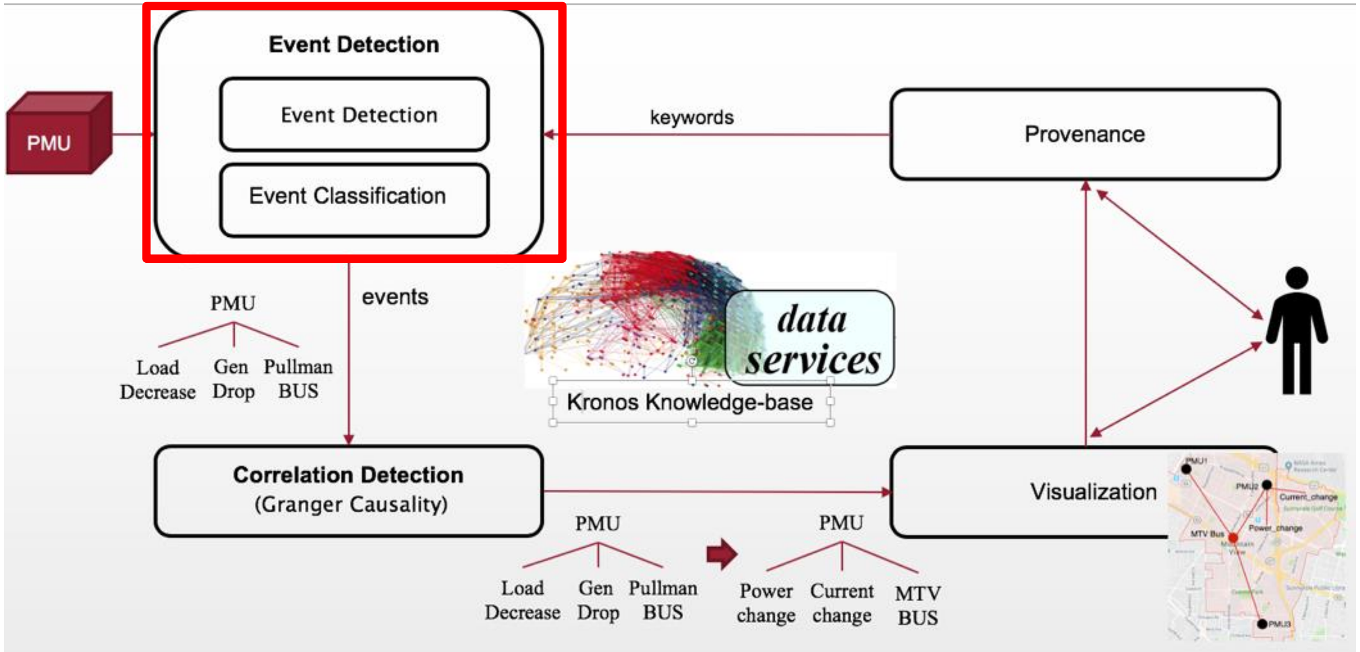
Long-term Goal: An interactive NLP-based Question & Answering system for resilient Cyber-power system (imagine a "Siri" or IBM Watson system for cyber-power event and resiliency analysis)

With Dr. Wu and Dr. Hahn, WSU and Siemens

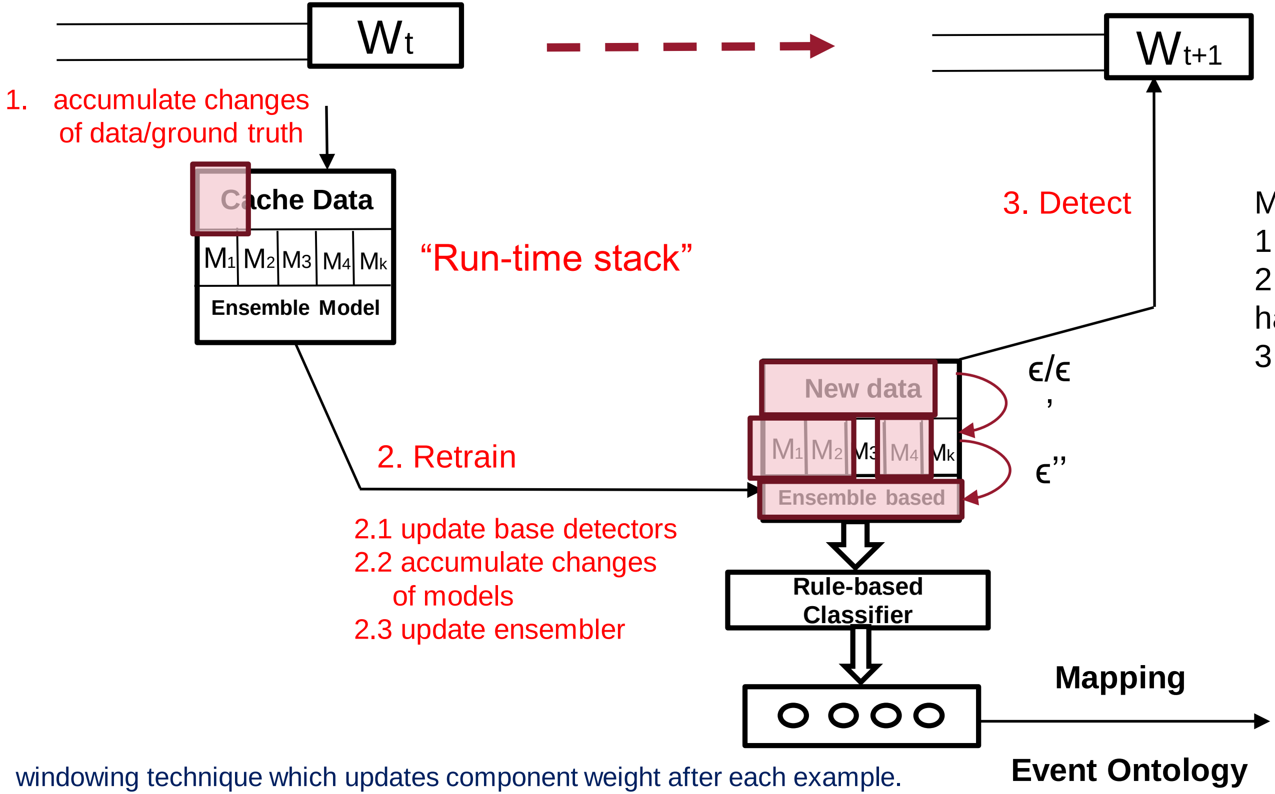


Problem Statement

- Input: Streams of events and physical entities (PMU, etc), ontology
- Output: A dynamically maintained knowledgebase (Kronos)



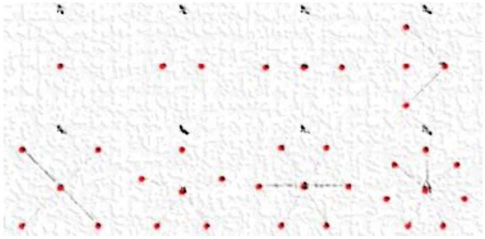
Global Data Streaming framework



Compare Both Windows
" Replace If there's change"

Model updates triggered by :
1. distribution change ϵ;
2. loss of accuracy ϵ' (when having training data)
3. Base detection change ϵ'';

event entities (stars)

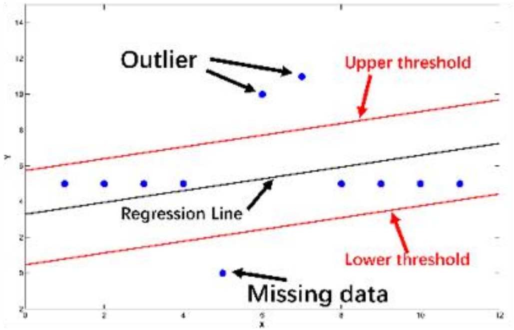


windowing technique which updates component weight after each example.
Incremental classifier for the ensemble learning which is trained between component reweighting.
Online drift detector that allows the shorten drift reaction time.

Options?

Linear regression

find straight line $y = \alpha + \beta x$ to provide a "best" fit for the data points w.r.t least-squares



Chebyshev method

Determine a lower bound of the percentage of data that exists within k standard deviations from t

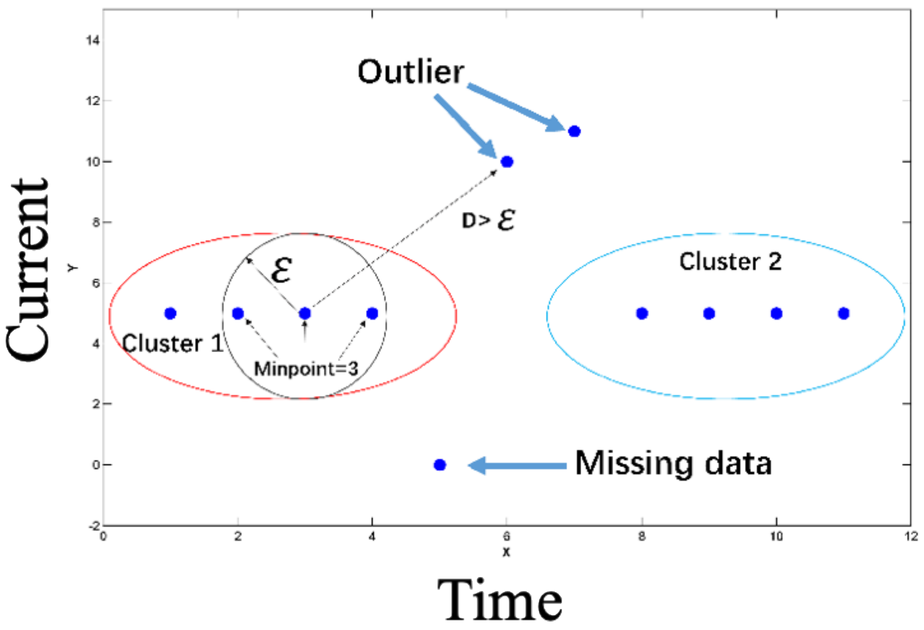
$$P(|X - \mu| \leq k\sigma) \geq (1 - \frac{1}{k^2})$$

μ: mean, σ: standard deviation, k: number of standard deviations from the mean.

Amidan, Brett G., Thomas A. Ferryman, and Scott K. Cooley. "Data outlier detection using the Chebyshev theorem." *Aerospace Conference, 2005 IEEE*. IEEE, 2005.

DBSCAN

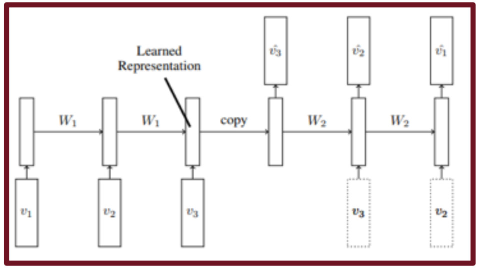
- DBSCAN uses two thresholds radius ε and *min*.
- A data point is a center node if it has more than *min* ε-neighbors (points within distance ε);
- Two centers are reachable if they are in ε-neighbor of each other; a cluster is a sequence of reachable centers and their ε-neighbors
- New clusters is formed after the event ends. Points far away from any cluster are outliers.



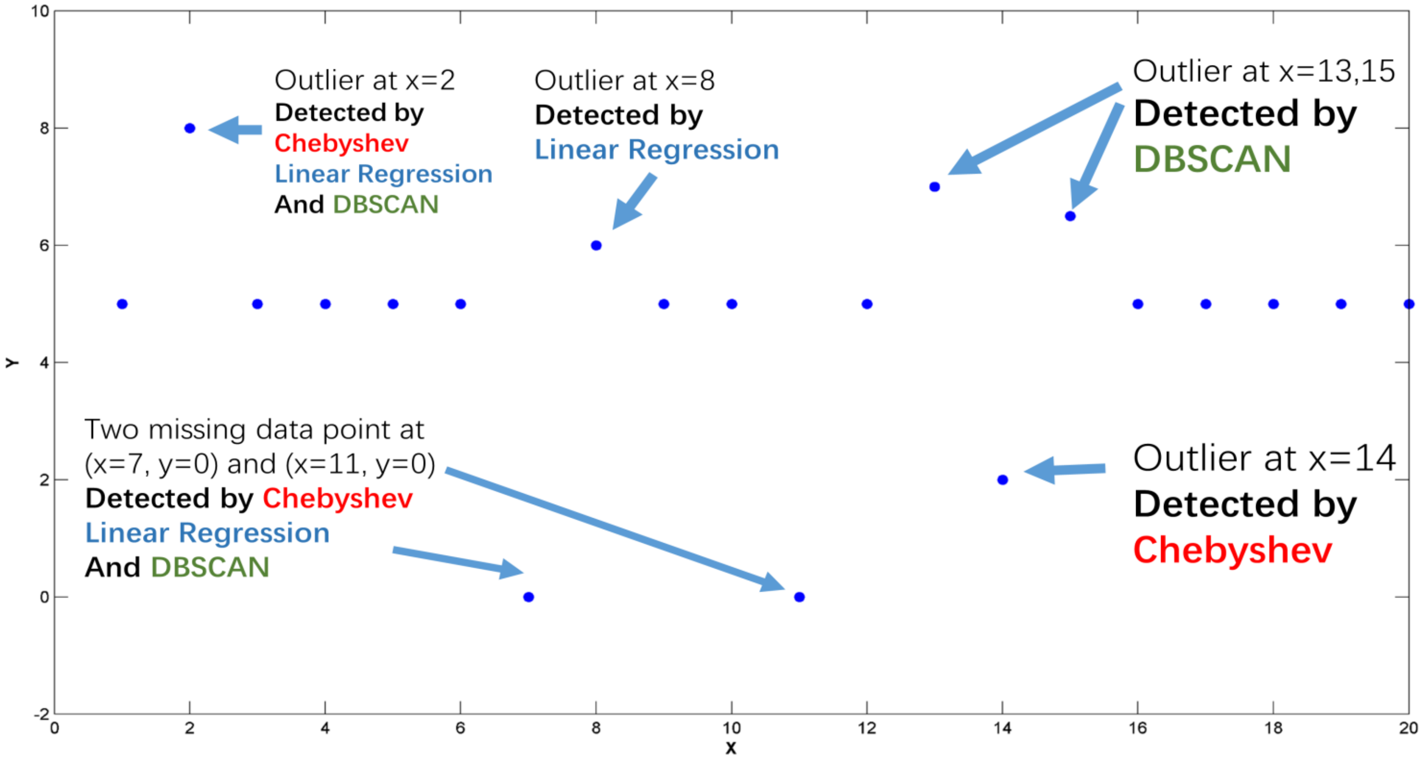
Does standalone method suffice?

LSTM Auto-encoder Model

- The model consists of two RNNs – the encoder LSTM and the decoder LSTM as shown in Figure
- The input to the model is a sequence of vectors (PMU data)
- The encoder LSTM reads in this sequence
- Once input vector is read, the decoder LSTM takes over and outputs a prediction for the target sequence
- The encoder can be seen as ‘creating a list’ of new inputs and previously constructed list (learned weights).
- The decoder essentially unrolls this list, with the hidden to output weights extracting the element at the top of the list and the hidden to hidden weights extracting the rest of the list.
- Thus the LSTM weights are learned using the auto encoder method.



$$
\begin{aligned}
\mathbf{i}_t &= \sigma\left(W_{xi}\mathbf{x}_t + W_{hi}\mathbf{h}_{t-1} + W_{ci}\mathbf{c}_{t-1} + \mathbf{b}_i\right),\\
\mathbf{f}_t &= \sigma\left(W_{xf}\mathbf{x}_t + W_{hf}\mathbf{h}_{t-1} + W_{cf}\mathbf{c}_{t-1} + \mathbf{b}_f\right),\\
\mathbf{c}_t &= \mathbf{f}_t\mathbf{c}_{t-1} + \mathbf{i}_t \tanh\left(W_{xc}\mathbf{x}_t + W_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c\right),\\
\mathbf{o}_t &= \sigma\left(W_{xo}\mathbf{x}_t + W_{ho}\mathbf{h}_{t-1} + W_{co}\mathbf{c}_t + \mathbf{b}_o\right),\\
\mathbf{h}_t &= \mathbf{o}_t \tanh(\mathbf{c}_t).
\end{aligned}
$$

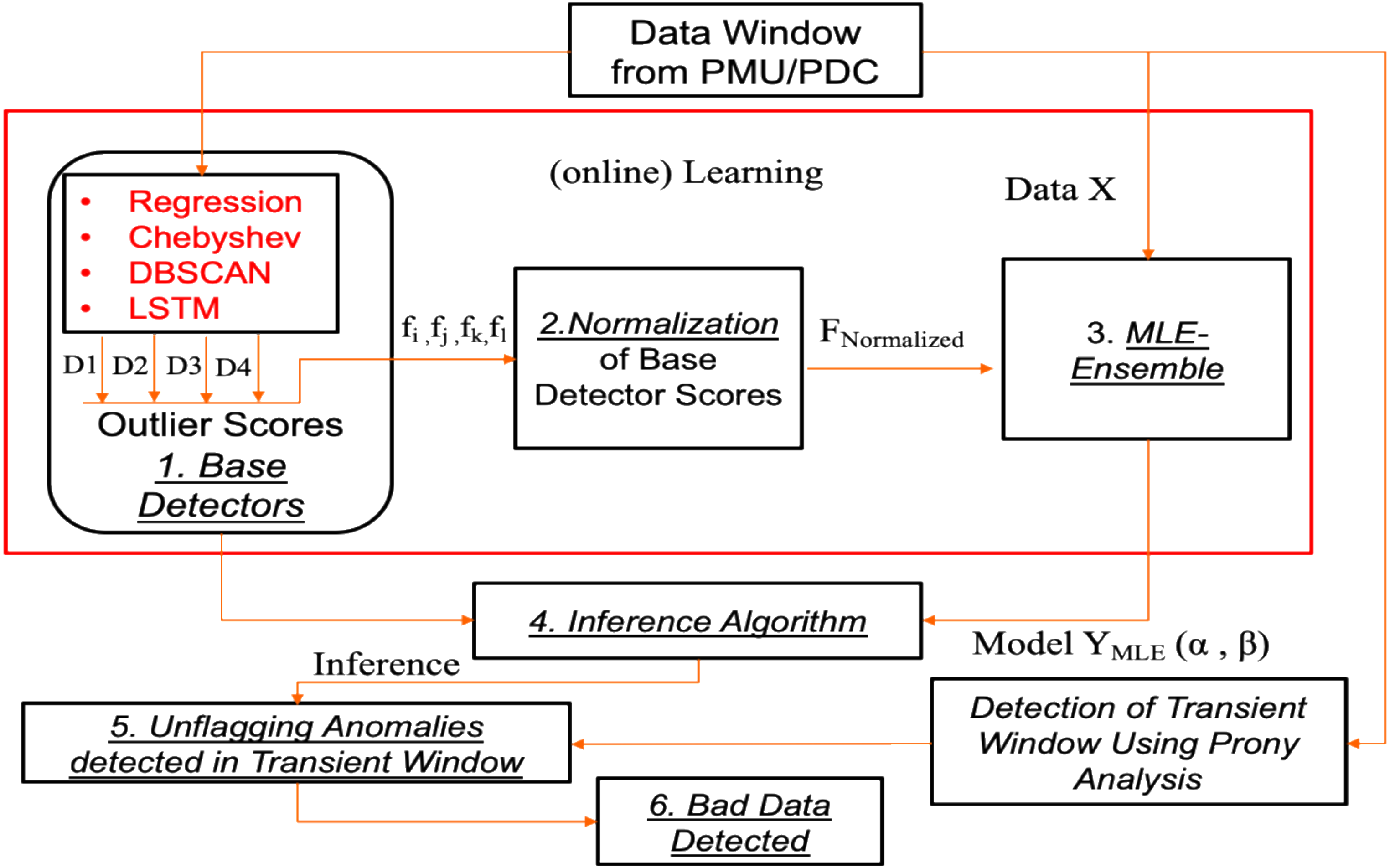


No Single Winner!

Lack of training data

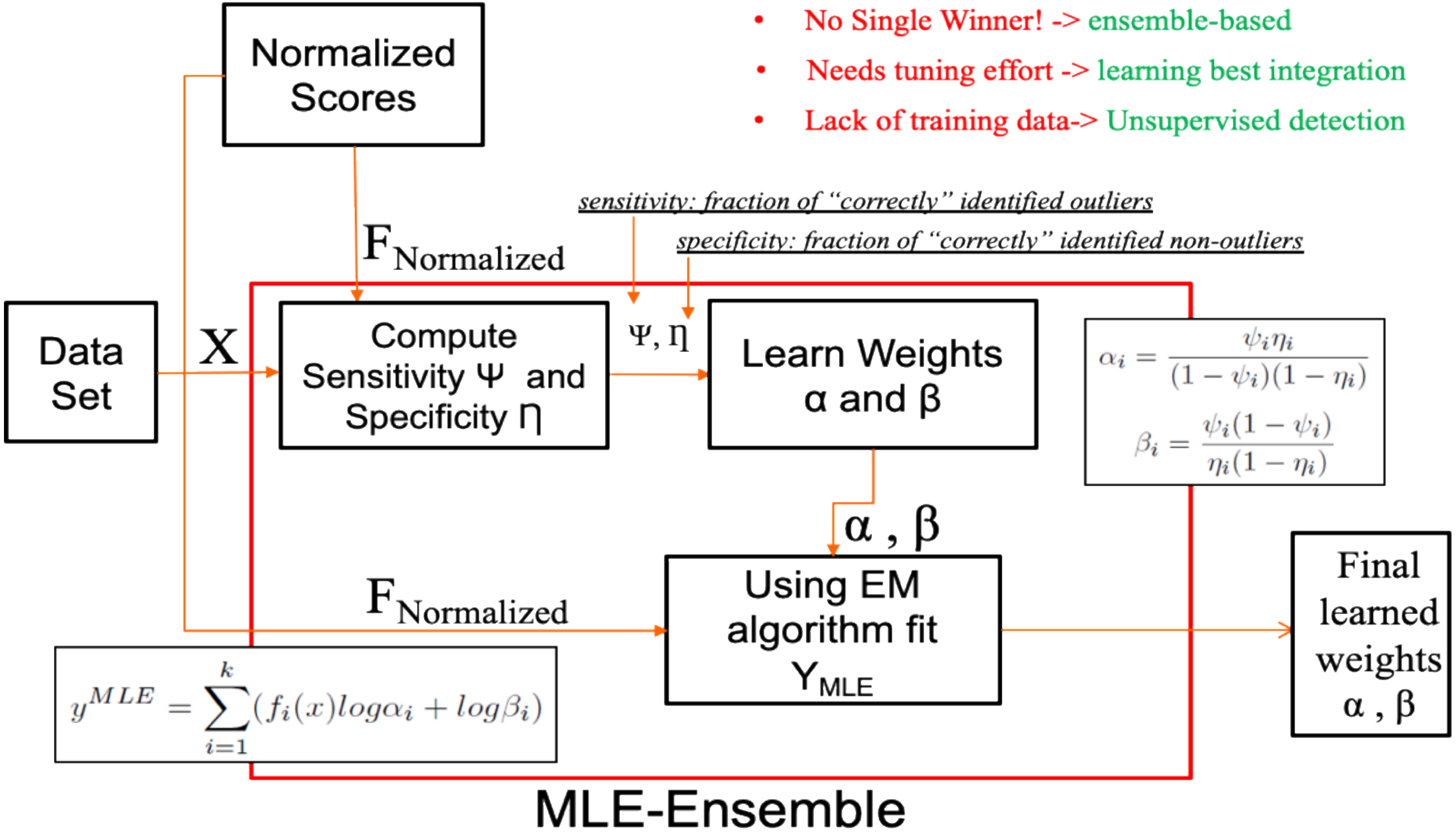
Needs tuning effort

Anomaly Detection with Ensemble



Maximum Likelihood Estimator (MLE)

- No Single Winner! -> ensemble-based
- Needs tuning effort -> learning best integration
- Lack of training data-> Unsupervised detection



Performance Metrics for Ensemble Based Technique

Given a PMU detector D and PMU data X, denote the actual anomaly data set as B_T , and the anomaly reported by D as B_D, the performance of D is evaluated using three metrics as follows.

Precision: Precision measures the fraction of true anomaly data in the reported ones from D, defined as

$$Precision = \frac{|B_D \cap B_T|}{|B_D|}$$

Recall: Recall measures the ability of D in finding all outliers, defined as

$$Recall = \frac{|B_D \cap B_T|}{|B_T|}$$

False Positive: False positive (FP) evaluates the possibility of false anomaly data detection; the smaller, the better.

$$FP = 1 - \frac{|B_D \cap B_T|}{|B_D|}$$

Simulation results for SyncAD

RTDS simulated PMU data (1.5 hours)

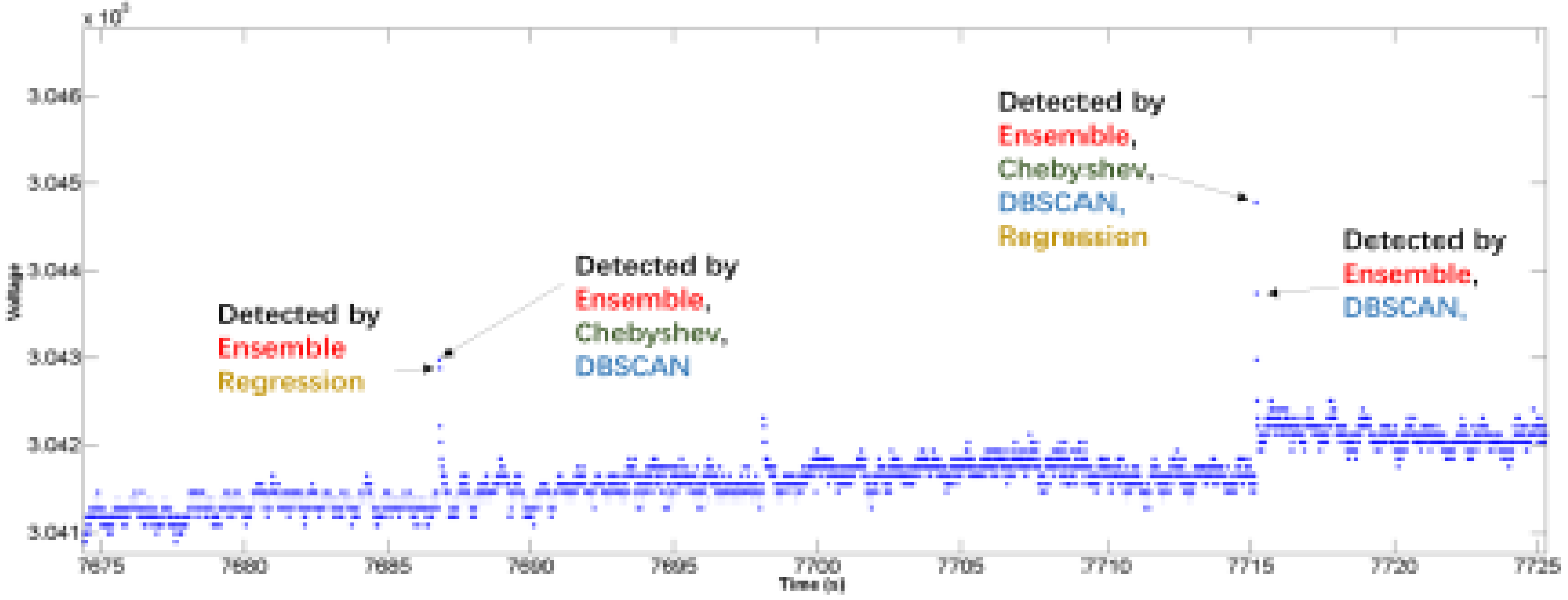
	Recall	Precision	False positive
Linear Regression	0.9021	0.8565	0.1435
DBSCAN	0.8821	0.8821	0.1179
Chebyshev	0.9154	0.8754	0.1246
LSTM	0.9298	0.8554	0.1446
MLE ensemble	**0.9351**	**0.8913**	**0.1087**

Tests on the RTDS simulated PMU data (1.5 hours, 5% bad data points, 5%-10% range)

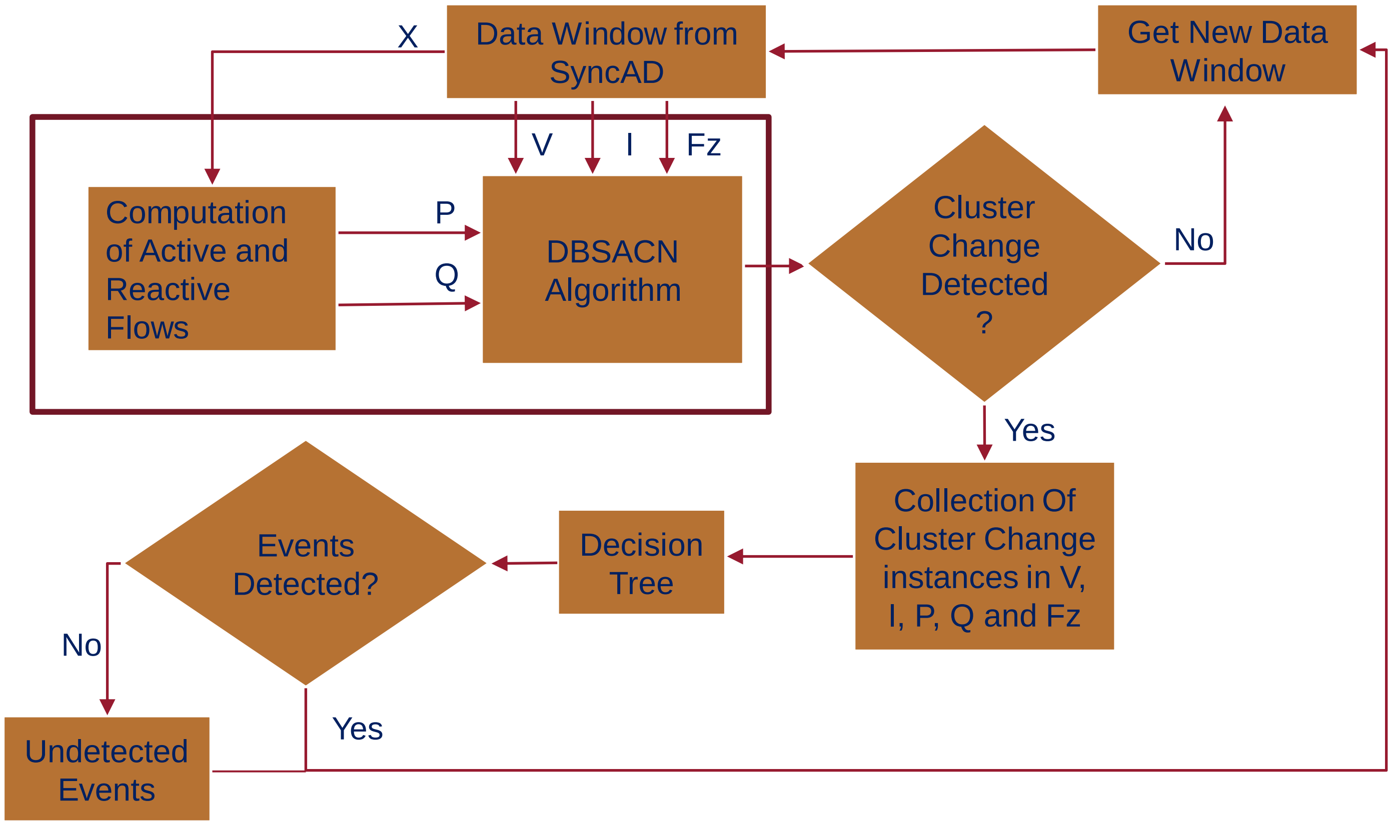
	Recall	Precision	False positive
Linear Regression	0.7854	0.7655	0.2345
DBSCAN	0.7216	0.7015	0.2985
Chebyshev	0.8125	0.7542	0.2458
LSTM	0.8298	0.7754	0.2246
MLE ensemble	**0.8912**	**0.9021**	**0.0979**

Tests on the RTDS simulated PMU data (1.5 hours, 10% bad data points, 10%-20% range)

Results with SyncAD using Real PMU Data



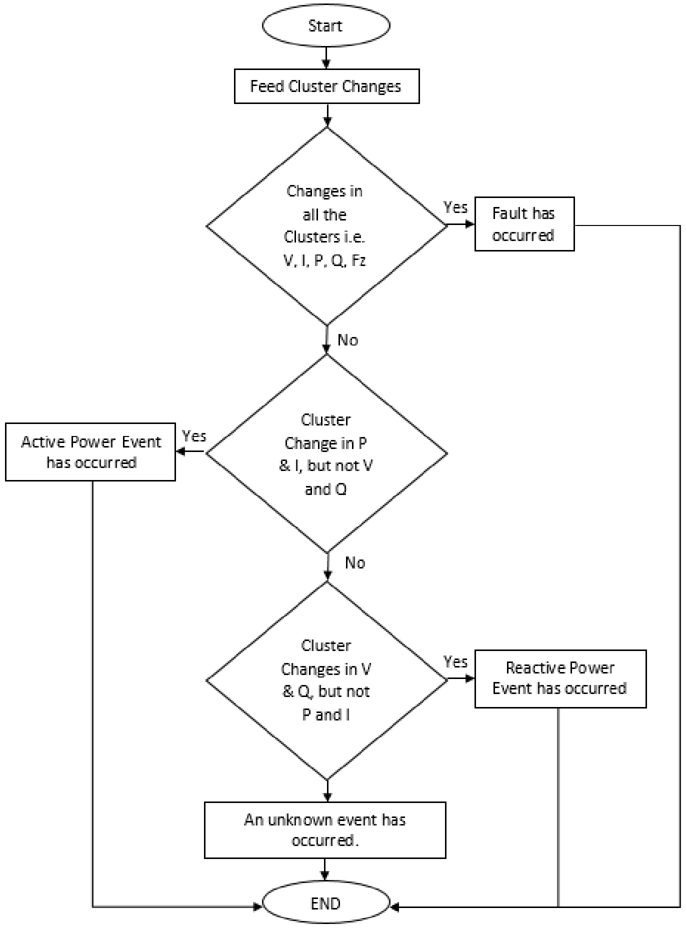
Event Detection Algorithm Architecture



Decision Tree for Event Classification

Event Classification Process

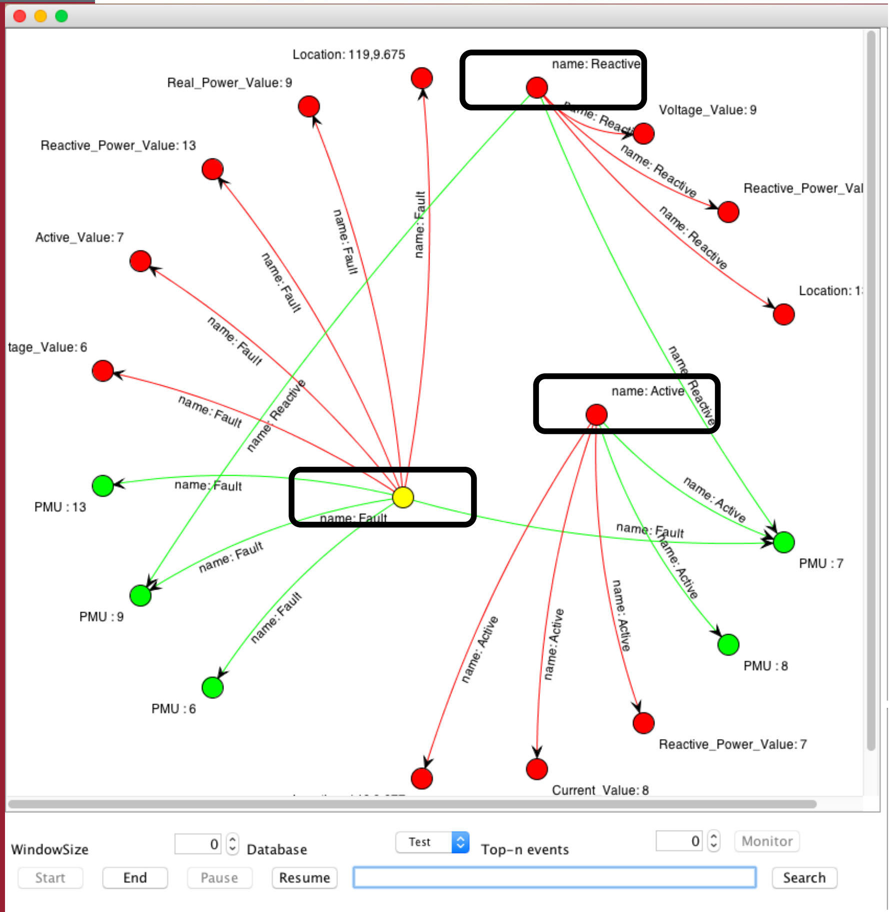
- Decision Tree: Active Power Event, Reactive Power Event and Fault Events.
- Cluster changes in P and I: Active Power Event
- Cluster change in V and Q: Reactive Power Event.
- Cluster changes V, I, P, Q and Fq: Fault event.

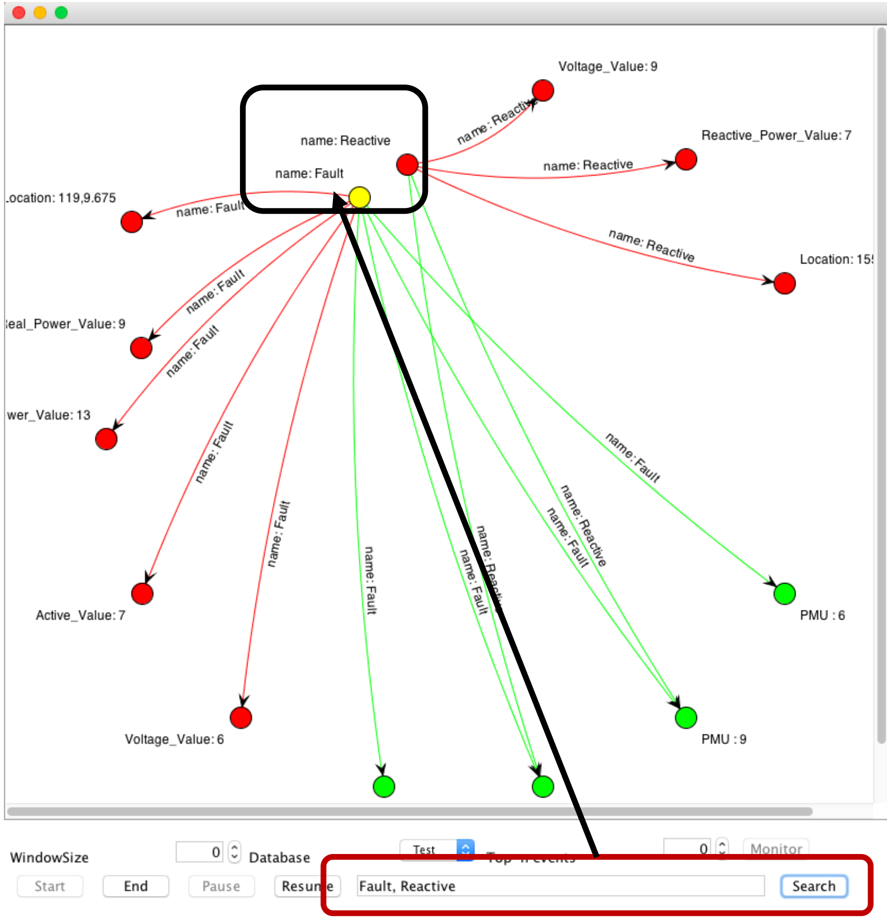


Simulation Results

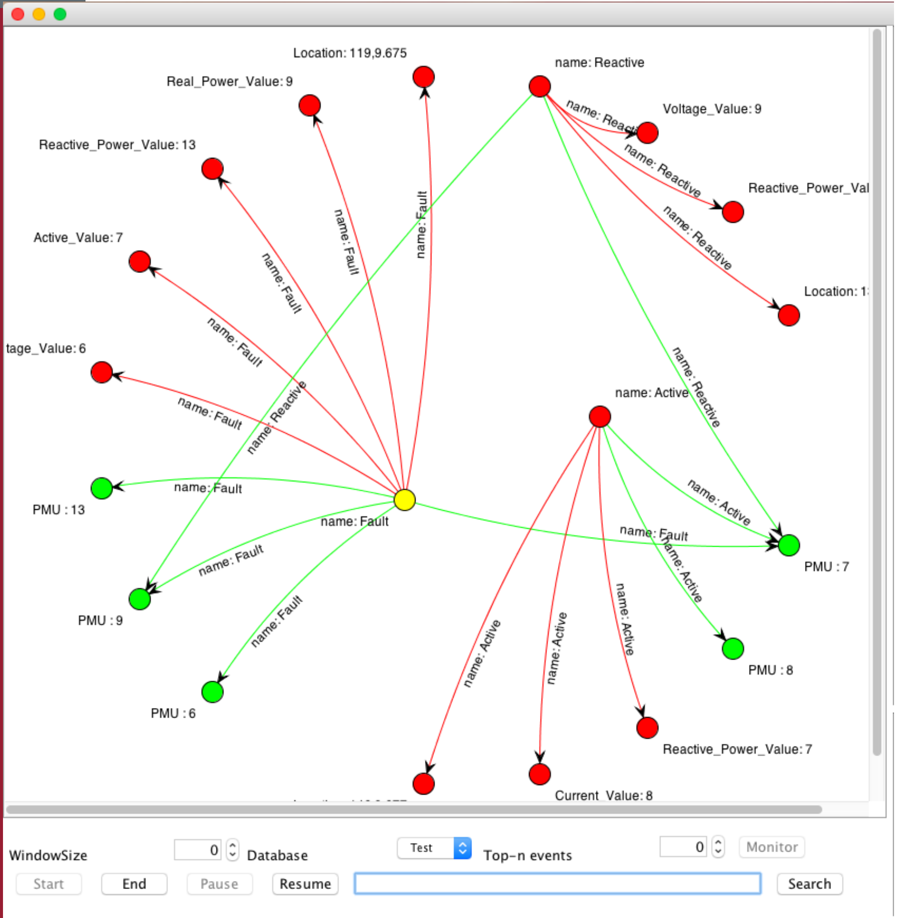
S No.	Time (s)	Reactive Event (Bus)	Active Event (Bus)	Fault Event (Bus)	Actual Events	Detected Events
1	109	9,7	-	-	Cap Bank Closed	Reactive
2	119	-	-	6,9,7,13,2	Three Phase fault	Fault
3	132	9,7	-	-	Cap Bank Opened	Reactive
4	148	-	7	-	P load decreased	Active
5	158	9,7,13	-	-	Cap Bank Closed	Reactive
6	168	-	-	6,9,7,13,2	Three Phase fault	Fault
7	179	9,7	-	-	Cap bank Opened	Reactive
8	188	-	-	-	P load increased	No Detection
9	198	9,7	-	-	Q load increased	Reactive
10	209	-	9	-	P load decreased	Active
11	219	9,7	-	-	Q load decreased	Reactive
12	229	-	2	-	Gen Drop	Active

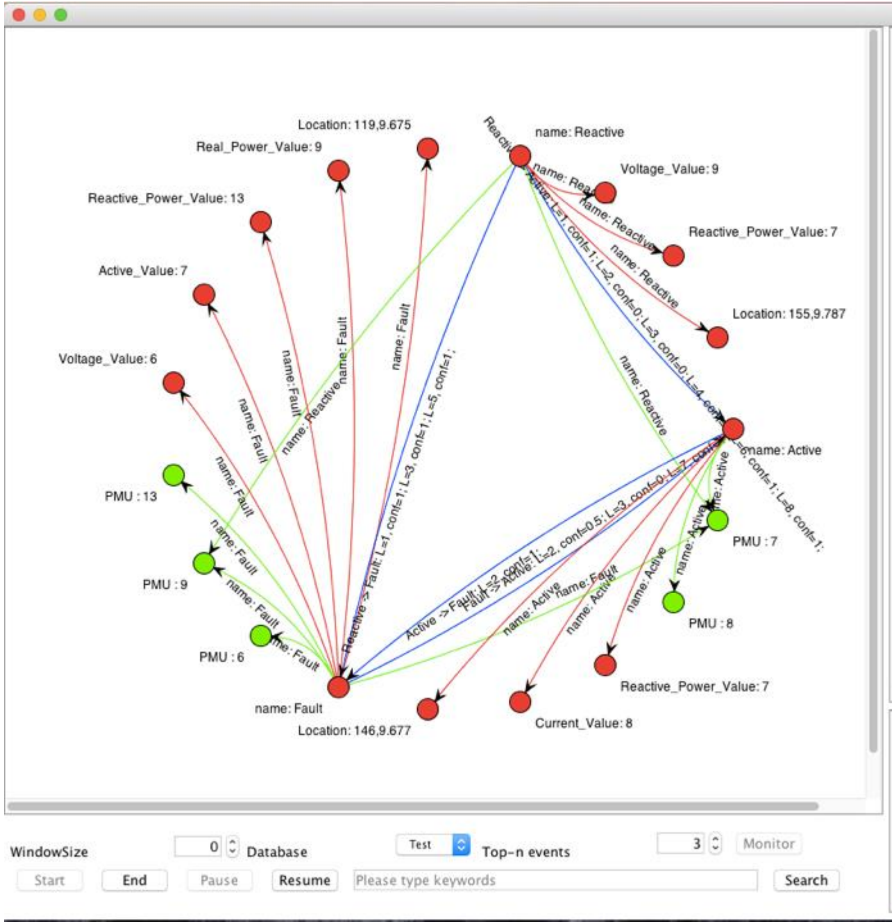
Case 1: keyword search (e.g, fault, reactive)



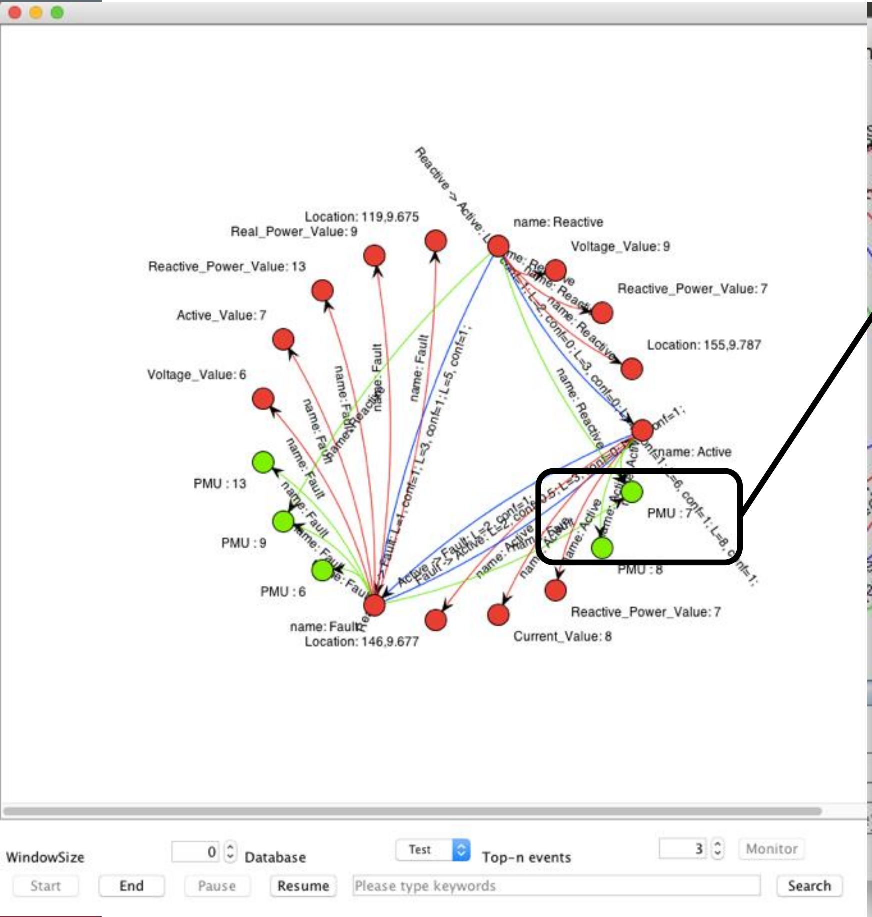


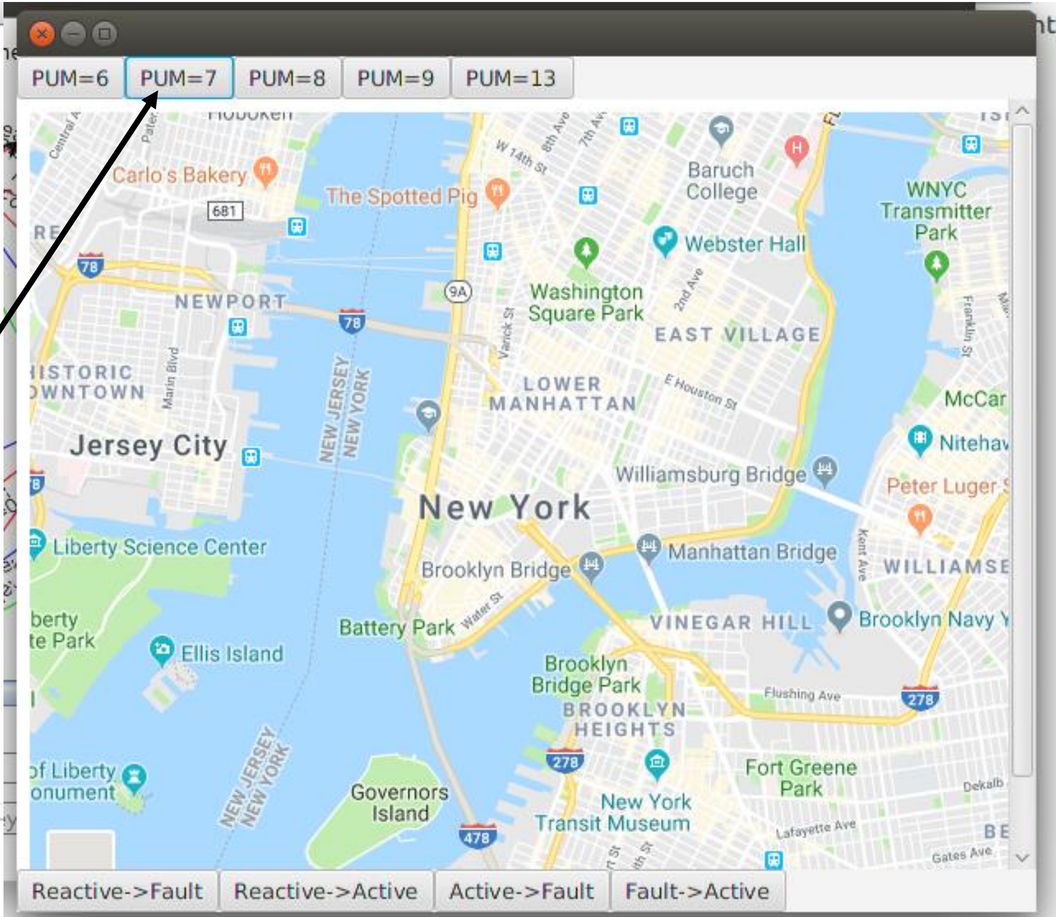
Case 2: Ontology and Correlation Monitoring (blue edges)





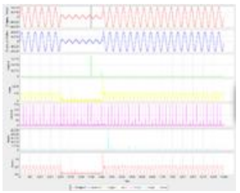
Case 3: Map interaction (google map)





Extending for microPMU

- Objective: Sensor data analytics (specifically using µPMU) for anomaly detection, classification, event detection, root cause analysis and explanation generation
- Tasks and Deliverables:
  - Provenance-aware Anomaly Detection and Prediction
  - Online Event Detection using µPMU datasets
  - Root cause detection based on statistical causality models
  - Explanation generation and interpretation



Configuration & Visualization via Graphical User Interface

μPMU_1

μPMU_2

μPMU_3

Event Detection

Time-series modeling
Olympic model, Regression model, …

Fault detection (Outliers, Changed points, Missing data)
Simple threshold, DBScan, KSigma, …

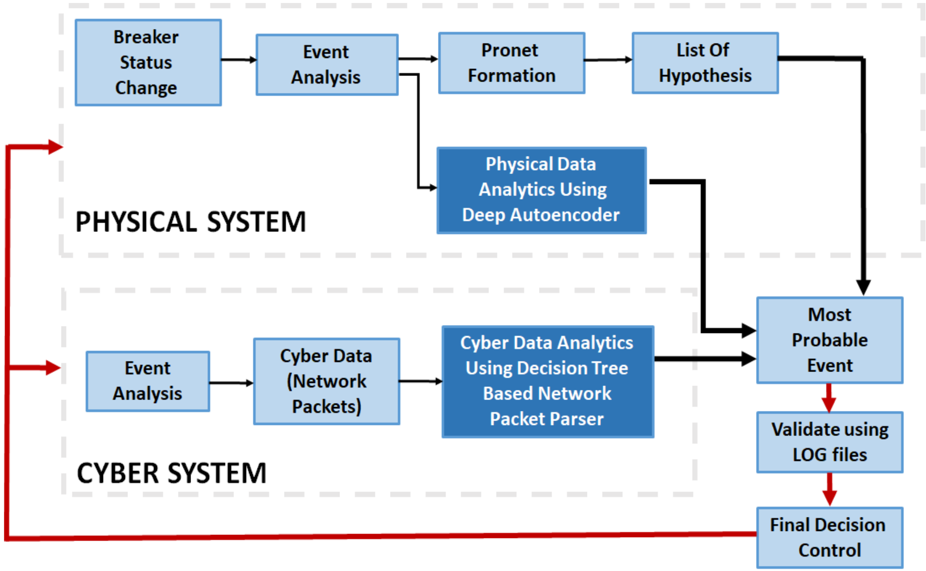
Correlation Detection between Events
(Extracting edges via Granger Causality)

Root Cause Analysis
via probabilistic models

Provenance
Generating explanations

Use case II: Cyber-physical Data Analytics in Protection Failure

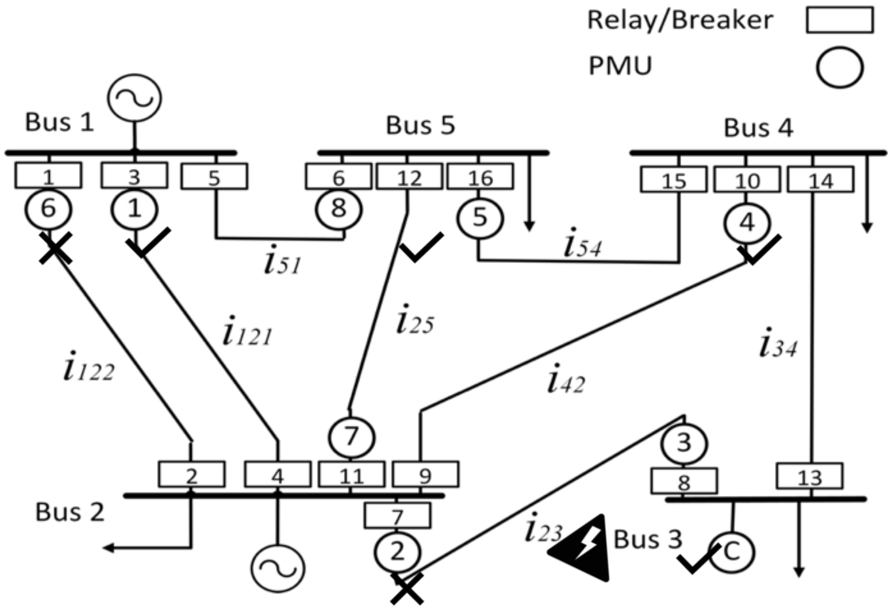
- Protection Mal-operation is #1 concern according to NERC
- Protection and associated control is becoming more digital



Abnormal Operation

A fault occurs on line 2-3 Relays 7 and 8 are expected to open their corresponding breakers but relay 7 doesn’t respond

To compensate relay’s 7 malfunction, relays 1, 3, 10 and 12 should open their corresponding breakers but relay 1 malfunctions.



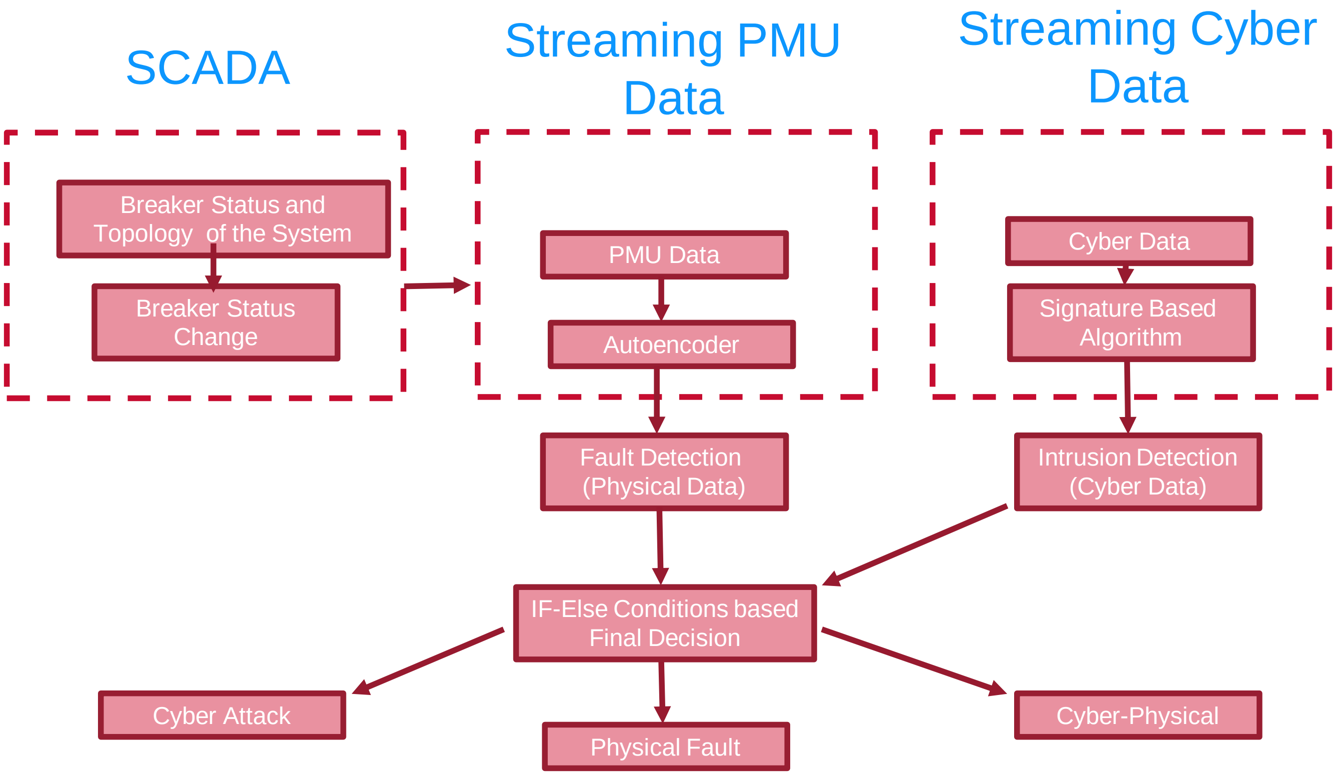
Cyber Physical Security Analytics for Anomalies in Transmission Protection Systems

Hypothesis Generation

Hypothesis #	Location of fault	Initial Incident	Consequential Incident
Actual Scenario	Line 2-3	Breaker 8 tripped Relay 7 malfunctioned	Breakers 3,10,12 tripped Relay 1 malfunctioned
Hypothesis 1	Line 2-4	Breaker 10 tripped Relay 9 malfunctioned	Breakers 3,8,12 tripped Relay 1 malfunctioned Relay 6 Tripped
Hypothesis 2	Line 2-1-2	Breaker 3 tripped Relay 4 malfunctioned	Breakers 8,10,12 tripped Relay 1 malfunctioned Relay 6 Tripped
Hypothesis 3	Line 1-5	Breaker 6 tripped Relay 5 malfunctioned	Relay 2, 3, 4 malfunctioned Breakers 8,10,12 tripped
Hypothesis 4	Line 2-5	Breaker 12 tripped Relay 11 malfunctioned	Breakers 3, 8, 10 tripped Relay 1 malfunctioned Relay 6 Tripped

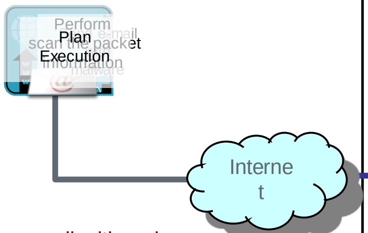
Cyber Physical Security Analytics for Anomalies in Transmission Protection Systems

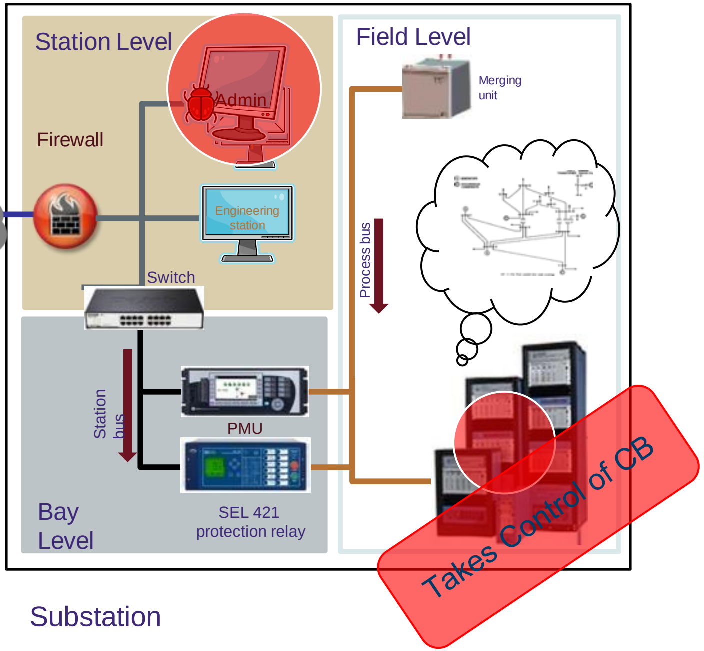
Data Analytics For Event Classification



Cyber Physical Security Analytics for Anomalies in Transmission Protection Systems

Simulating Cyber Attack on a Relay





Substation

1. Attacker sends an e-mail with malware
2. E-mail recipient opens the e-mail and the malware gets installed quietly
3. Using the information that malware gets, hacker is able to take control of the e-mail recipient's PC and get access of two-level password
4. Analysis IEC 61850 protocol(GOOSE, SMV packet) information and relay setting file
5. Manipulate MMS packet and relay configuration session information
6. Takes control of circuit breaker or change the setting of relay

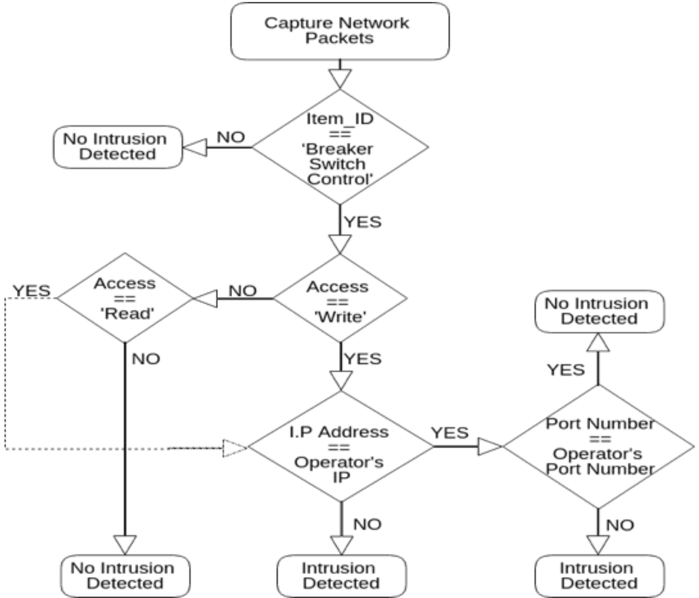
Detect Intrusion Using Cyber Data From Relay.

Relay IP address: 192.168.0.16 || Operator IP address: 192.168.0.23 || Unauthorized IP address:192.168.0.14

No.	Time	Source	Destination	Protocol	Length	Info
2296	126.405616	192.168.0.14	192.168.0.16	MMS	229	confirmed-RequestPDU
2297	126.409243	192.168.0.16	192.168.0.14	MMS	84	confirmed-ResponsePDU
2298	132.293425	192.168.0.14	192.168.0.16	MMS	229	confirmed-RequestPDU
2299	132.296947	192.168.0.16	192.168.0.14	MMS	84	confirmed-ResponsePDU
2300	137.581544	192.168.0.14	192.168.0.16	MMS	229	confirmed-RequestPDU
2301	137.645231	192.168.0.16	192.168.0.14	MMS	84	confirmed-ResponsePDU
2302	141.453519	192.168.0.14	192.168.0.16	MMS	229	confirmed-RequestPDU
2303	141.456890	192.168.0.16	192.168.0.14	MMS	84	confirmed-ResponsePDU
2304	145.213451	192.168.0.14	192.168.0.16	MMS	229	confirmed-RequestPDU
2305	145.216523	192.168.0.16	192.168.0.14	MMS	84	confirmed-ResponsePDU
2306	151.245001	192.168.0.14	192.168.0.16	MMS	229	confirmed-RequestPDU

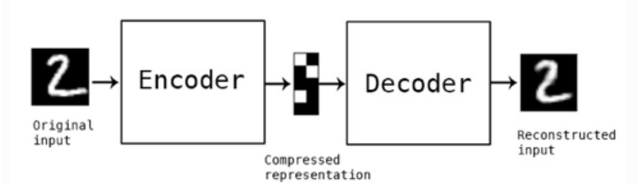
Attack Scenario For Relay
Communication between Relay and Un-authorized IP Address-(Attacker)

Detecting an Intrusion :



Detect Intrusion Using Physical Data From PMU

Algorithm Description :



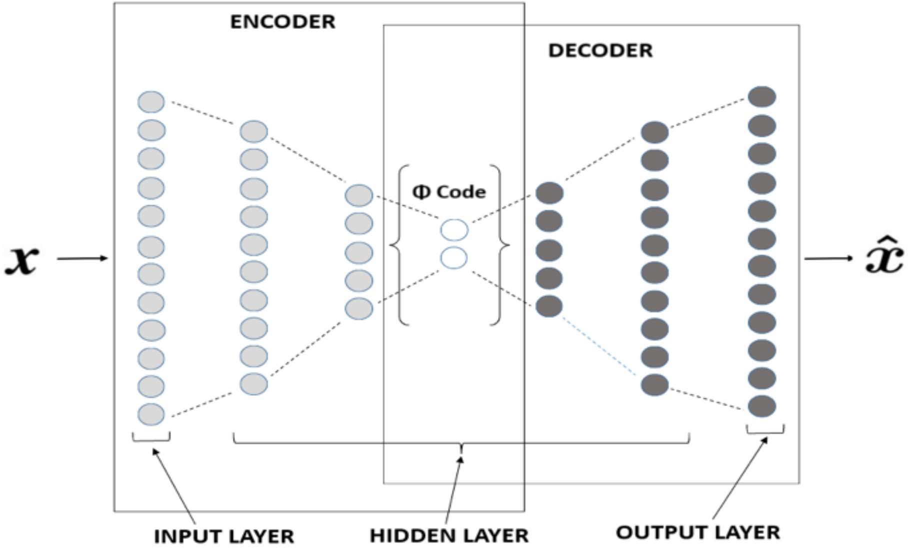
- Basic Idea : Reconstruction of input feature vector with minimum loss (Mean Square Error)

- Train the algorithm on input data consisting of no anomalies.
  Output Result : Reconstructed input feature vector with low MSE.

- Test the algorithm on input data consisting of anomalies.
  Output Result : Reconstructed input feature vector with high MSE.

- We want our algorithm to have high MSE on input data consisting of anomalies and low MSE on input data consisting of no anomalies.

Detect Intrusion Using Physical Data From PMU



Architecture Of
Stacked Autoencoder

Loss Function : Mean Squared Error
Optimizer : ADAM

x : Input Feature Vector

$\hat{x}$: Reconstructed Output Feature Vector

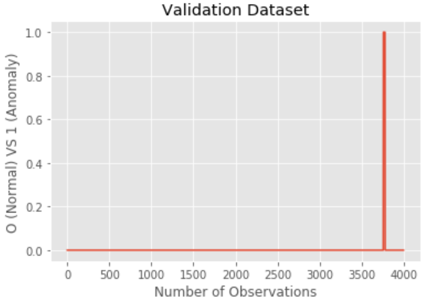
Detect Intrusion Using Physical Data From PMU

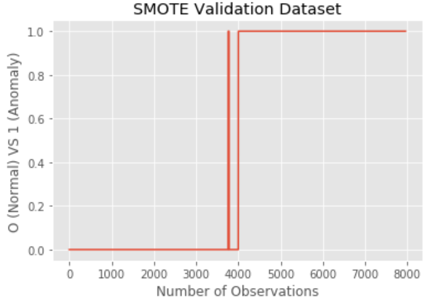
Dataset Description :

Dataset	# PMU Readings (Total : 37500)
Training Dataset (No Fault)	22250
Testing Dataset (No Fault)	11250
Validation Dataset (Fault)	4000

Types Of Validation Dataset:

Validation Dataset	PMU Readings (# Normal Instances)	PMU Readings (# Anomalous Instances)
Type 1	3979	21
Type 2 (Synthetic Minority Oversampling -SMOTE)	3979	3979





Detect Intrusion Using Physical Data From PMU

Evaluation Metrics

The intersection between actual values and predicted values yield four possible situations:

- True Positive (TP): Positive instances correctly classified.
- False Positive (FP): Negative instances classified as positive.
- True Negative (TN): Negative instances correctly classified as negative.
- False Negative (FN): Positive instances classified as negative.

Classification Measures:

Accuracy is calculated as the number of correctly classified instances over total number of instances evaluated.

$$\text{Accuracy} = \frac{TP + TN}{Total\ instances}$$

Precision is the percentage of correctly predicted instances over the total instances predicted for positive class.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall is the percentage of correctly classified instances over the total actual instances for the positive class.

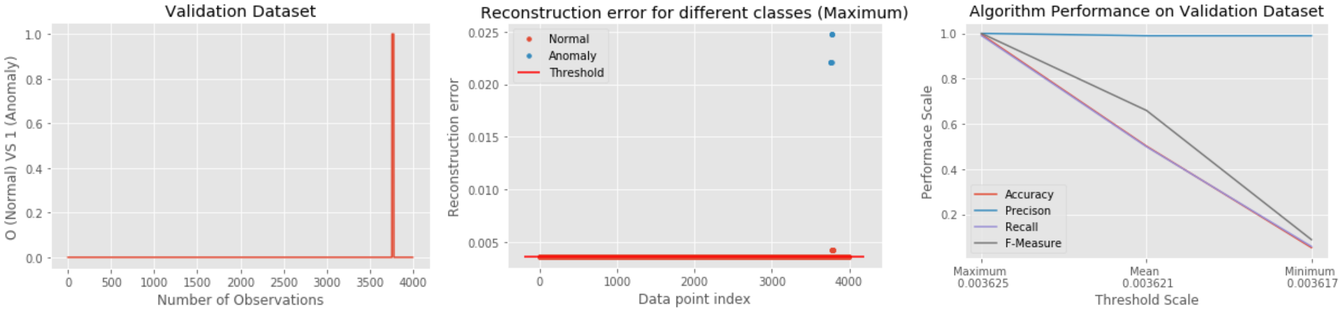
$$\text{Recall} = \frac{TP}{TP + FN}$$

F-Measure is a measure of test accuracy.

$$\text{F-measure} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

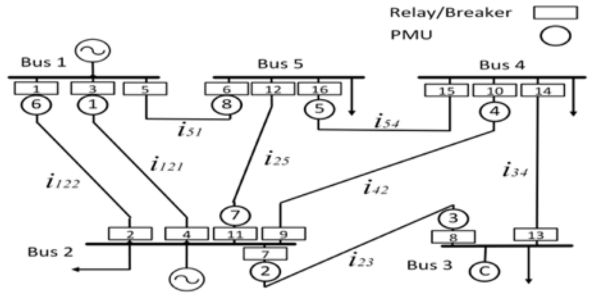
Detect Intrusion Using Physical Data From PMU

Autoencoder Evaluation On Type 1 (Validation Dataset)

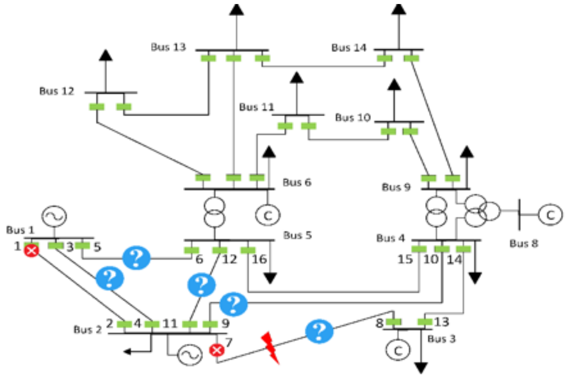


Threshold (Test Data)	Accuracy	Precision	Recall	F-Measure
0.003617 (Minimum)	5.50%	0.99	0.06	0.09
0.003621 (Mean)	50.25%	0.99	0.50	0.66
0.003625 (Maximum)	99.48%	1.0	0.99	1.00

Decision Based On Data Analytics and Validation Using Additional Non-Streaming Data



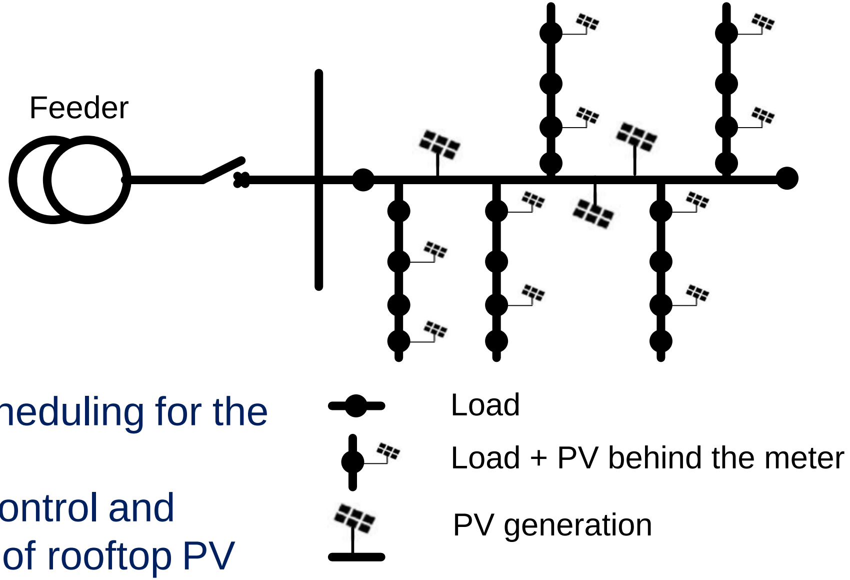
Scenario	Location of Fault	Initial incident	Consequential incident
	Line 2-3	Breaker 8 tripped Relay 7 malfunctioned	Breakers 3, 10, 12 tripped Relay 1 malfunctioned Relay 6 tripped
Scn 1	Line 2-4	Breaker 10 tripped Relay 9 malfunctioned	Breakers 3, 8, 12 tripped Relay 1 malfunctioned Relay 6 tripped
Scn 2	Line 2-1-2	Breaker 3 tripped Relay 4 malfunctioned	Breakers 8, 10, 12 tripped Relay 1 malfunctioned Relay 6 tripped
Scn 3	Line 1-5	Breaker 6 tripped Relay 5 malfunctioned	Relays 2,3,4 malfunctioned Breakers 8, 10, 12 tripped
Scn 4	Line 2-5	Breaker 12 tripped Relay 11 malfunctioned	Breakers 3, 8, 10 tripped Relay 3 malfunctioned Relay 6 tripped



- PMU 2 and 3 show highest MSE among all PMUs
- it can be determined that most probably the fault could have occurred in the line from bus 2 and 3

Use case III: Load/ DER Disaggregation

- Increasing PV penetration
  - Behind-the-meter (Invisible)
  - PV with meters (Visible)
- Invisible solar photovoltaic not monitored
- Invisible to utilities and system operators
- Visibility into behind-the-meter solar generation is limited



- System net load is a key input when scheduling for the short-term operation (minutes/ hours)
- DER estimation can help with voltage control and CVR, especially given high percentage of rooftop PV

- Exact load estimation can help with cold load pickup after the outage
- DER estimation can also help for resiliency driven outage management with DERs in microgrids/ active distribution system

Research Question

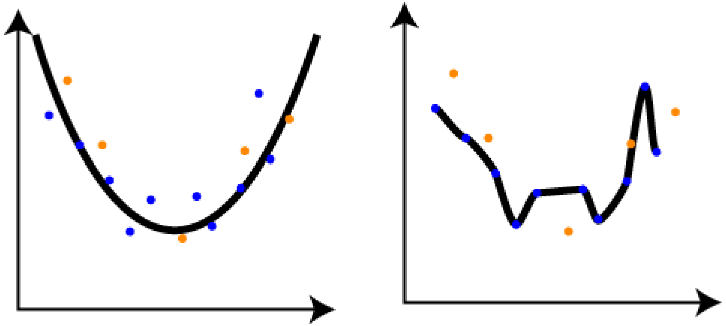
How do we gain more visibility into the PV generation and load behind the meter?

- Data-driven methodology can be used to estimate (ML prediction) the power generation of invisible solar power sites in short time scale and load/ DER can be disaggregated

With Dr. Gebremedhin, WSU

Supervised Machine Learning

- A computer system learns from data, which represent some "past experiences" of an application domain
- Our focus: learn a target function that can be used to predict the values of a continuous value, i.e. load, power generation
- The task is commonly called Regression: a specific type of Supervised learning, complementary to Classification
- Supervision: the data (observations, measurements, etc.) are labeled with pre-defined values



Regression Problem Formulation

- Data: a set of data records (also called examples, instances) described by
  - k attributes: X_1, X_2, … X_k (e.g. weather, voltage, power)
  - a target value (y): each example has a pre-defined value (e.g. power estimation)

- Goal: learn a regression model from the data that can be used for estimating values of new (future or test) instances
  - We may use a simple model like linear regression or a more complex model

General $\boldsymbol{y = F(X_0, X_1, X_2, \dots, X_k) + \varepsilon}$

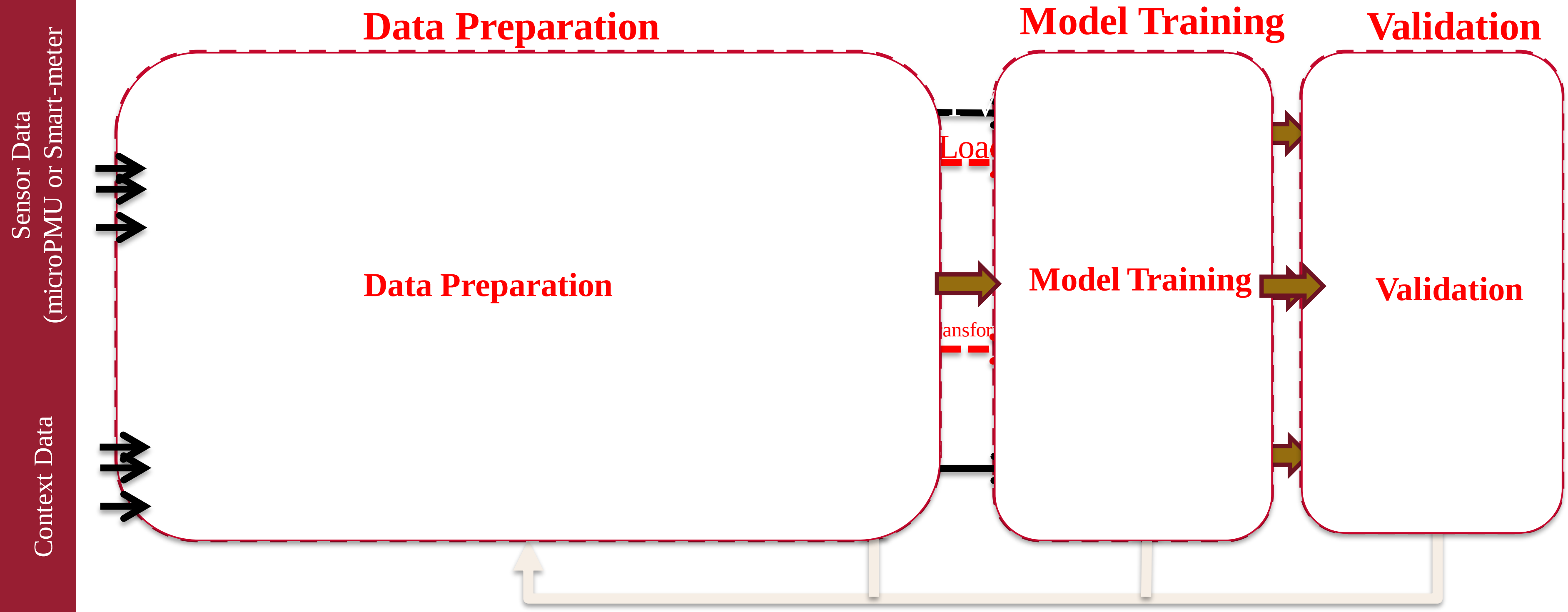
Linear $\boldsymbol{y = X_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon}$

- Evaluation: accuracy of estimation

PV/Load Disaggregation

- Data Properties
  - High sampling rate
  - Heterogeneous data (e.g. different sampling rate, different nature)
  - Many missing points
- Weather data: 4 to 9 variables
  - Diffused irradiance, global irradiance, humidity, temperature, wind direction, wind speed, dew point, pressure, rain
- Challenges
  - Large scale data
    - More than 31,540,000 rows of data per year (based on HNEI data)
  - How can we represent data to expedite model training?
    - Feature extraction
  - How can we use future data to upgrade the model?
    - Online Learning, Deep Learning

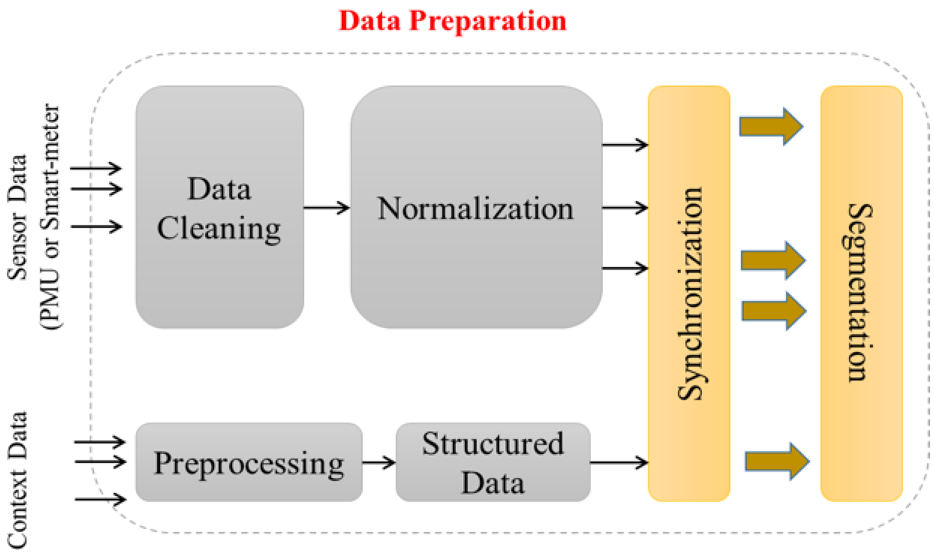
PV/Load Disaggregation ML Pipeline



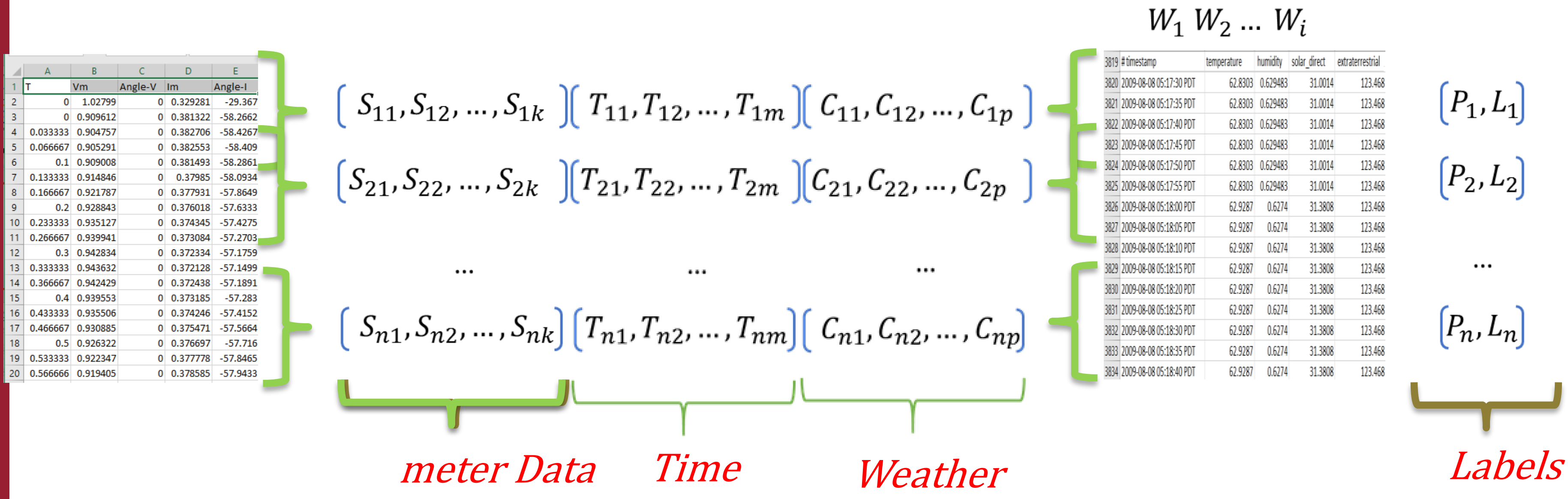
Modification (Feature Selection, Segmentation Setting, ML Model Selection)

Data Preparation

- Low-pass filter
  - to remove high frequency noise data
- Average-based sliding window method
  - to fill missing values
- Process weather station text data
  - to generate clean and structured data
- Spline interpolation
  - to match different sampling rates
- Synchronization
  - timestamp adaptation for different sources of data



Data Segmentation and Feature Extraction
Implementation



	A	B	C	D	E
1	T	Vm	Angle-V	Im	Angle-I
2	0	1.02799	0	0.329281	-29.367
3	0	0.909612	0	0.381322	-58.2662
4	0.033333	0.904757	0	0.382706	-58.4267
5	0.066667	0.905291	0	0.382553	-58.409
6	0.1	0.909008	0	0.381493	-58.2861
7	0.133333	0.914846	0	0.37985	-58.0934
8	0.166667	0.921787	0	0.377931	-57.8649
9	0.2	0.928843	0	0.376018	-57.6333
10	0.233333	0.935127	0	0.374345	-57.4275
11	0.266667	0.939941	0	0.373084	-57.2703
12	0.3	0.942834	0	0.372334	-57.1759
13	0.333333	0.943632	0	0.372128	-57.1499
14	0.366667	0.942429	0	0.372438	-57.1891
15	0.4	0.939553	0	0.373185	-57.283
16	0.433333	0.935506	0	0.374246	-57.4152
17	0.466667	0.930885	0	0.375471	-57.5664
18	0.5	0.926322	0	0.376697	-57.716
19	0.533333	0.922347	0	0.377778	-57.8465
20	0.566666	0.919405	0	0.378585	-57.9433

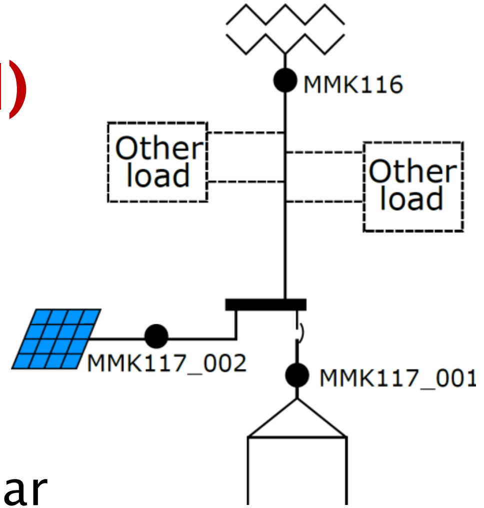
3819	#timestamp	temperature	humidity	solar_direct	extraterrestrial
3820	2009-08-08 05:17:30 PDT	62.8303	0.629483	31.0014	123.468
3821	2009-08-08 05:17:35 PDT	62.8303	0.629483	31.0014	123.468
3822	2009-08-08 05:17:40 PDT	62.8303	0.629483	31.0014	123.468
3823	2009-08-08 05:17:45 PDT	62.8303	0.629483	31.0014	123.468
3824	2009-08-08 05:17:50 PDT	62.8303	0.629483	31.0014	123.468
3825	2009-08-08 05:17:55 PDT	62.8303	0.629483	31.0014	123.468
3826	2009-08-08 05:18:00 PDT	62.9287	0.6274	31.3808	123.468
3827	2009-08-08 05:18:05 PDT	62.9287	0.6274	31.3808	123.468
3828	2009-08-08 05:18:10 PDT	62.9287	0.6274	31.3808	123.468
3829	2009-08-08 05:18:15 PDT	62.9287	0.6274	31.3808	123.468
3830	2009-08-08 05:18:20 PDT	62.9287	0.6274	31.3808	123.468
3831	2009-08-08 05:18:25 PDT	62.9287	0.6274	31.3808	123.468
3832	2009-08-08 05:18:30 PDT	62.9287	0.6274	31.3808	123.468
3833	2009-08-08 05:18:35 PDT	62.9287	0.6274	31.3808	123.468
3834	2009-08-08 05:18:40 PDT	62.9287	0.6274	31.3808	123.468

Number of features for each segment:

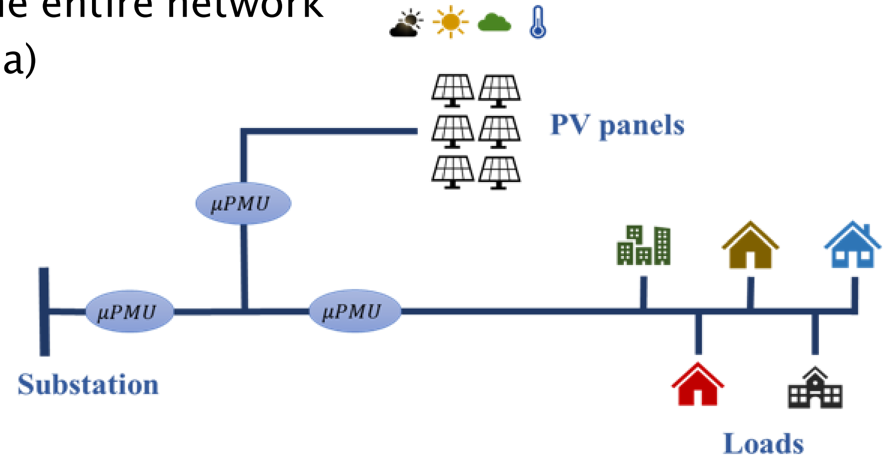
$$(PMU_{size} + W_{size}) * number\ of\ SF + number\ of\ TR\ features$$

Datasets (working with HNEI)

- Real Data (Maui Hawaii) --- One Year
  - Smart-meter transformer data
  - Smart-meter load data (target value)
  - Smart-meter solar panel data (target value)
  - Weather data in text format (9 parameters)



- GridLab-D (IEEE 123 Node Test Feeder) --- One Year
  - PMU data: current and voltage for the entire network
  - Weather data (low sampling rate data)
  - Seattle weather data (4 parameters)
  - 5 Solar panels in the network
  - Load data (in progress)



Case Studies

- Real data, Load estimation
  - Target: load value
  - Input: transformer smart-meter data and weather data
- Real data, Solar panel power estimation
  - Target: solar panel power generation
  - Input: weather data
- GridLab-D data, Solar panel power estimation
  - Target: solar panel power generation
  - Input: weather data
- (Ongoing) GridLab-D data, Load estimation

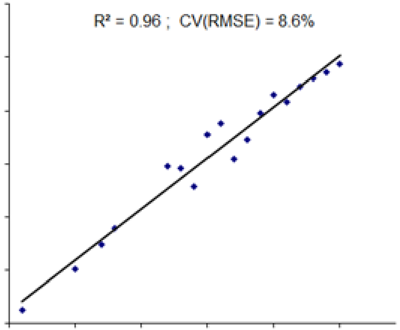
Accuracy Metrics

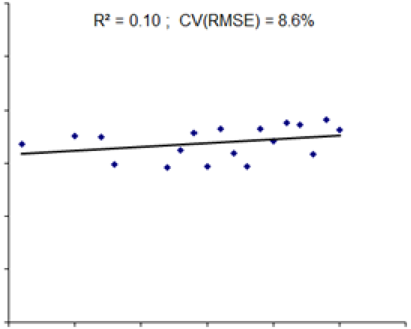
- R-squared
  - Also called **coefficient of determination** (the square of the correlation coefficient)
  - Represents **the fraction of the variance in y that can be explained by the regression model**

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- RMSE
  - The Root Mean Square Error for prediction

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$





Real Data, Load Estimation

- Investigated importance of features
  - Most valuable feature: irradiance
- Models we used
  - Linear regression, decision tree based regression, deep learning (LSTM)
- Future direction
  - Train model based on Load and PV capacity from multiple locations
  - Weather variation: a generalized model that works for weather condition in any location
  - Active Learning: collect data for load based on PV estimation and update load estimation model

Scenario/ RMSE (%)	Transformer Only	Transformer & PV	Context-aware
Scenario 1	20 to 40%	12 to 41%	4 to 18%
Scenario 2	18 to 36%	10 to 3%	4 to 13%
Scenario 3	8 to 28%	7 to 12%	2 to 11%

Models: LR, DT, LSTM

Each scenario corresponds to a specific
train/test split :
S1: 3 months train, one month test
S2: 6 months train, one month test
S3: 11 months train, one month test

Real Data, PV Estimation

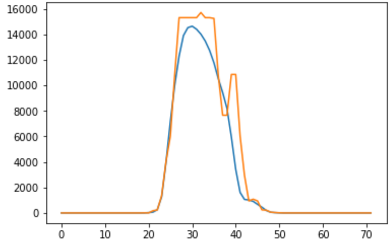
- We found that power generation is highly dependent on context
- Using transformer smart meter data results in higher accuracy
- Models we used
  - Decision tree based
  - Linear regression
  - MLP
- Future direction
  - Use LSTM
  - Train models for different scenarios (PV capacity, different weather conditions)

Scenario/ R^2	Transformer Only	Transformer and Load	Context-aware
Scenario 1	0.56	0.63	0.83
Scenario 2	0.66	0.65	0.81
Scenario 3	0.72	0.77	0.89

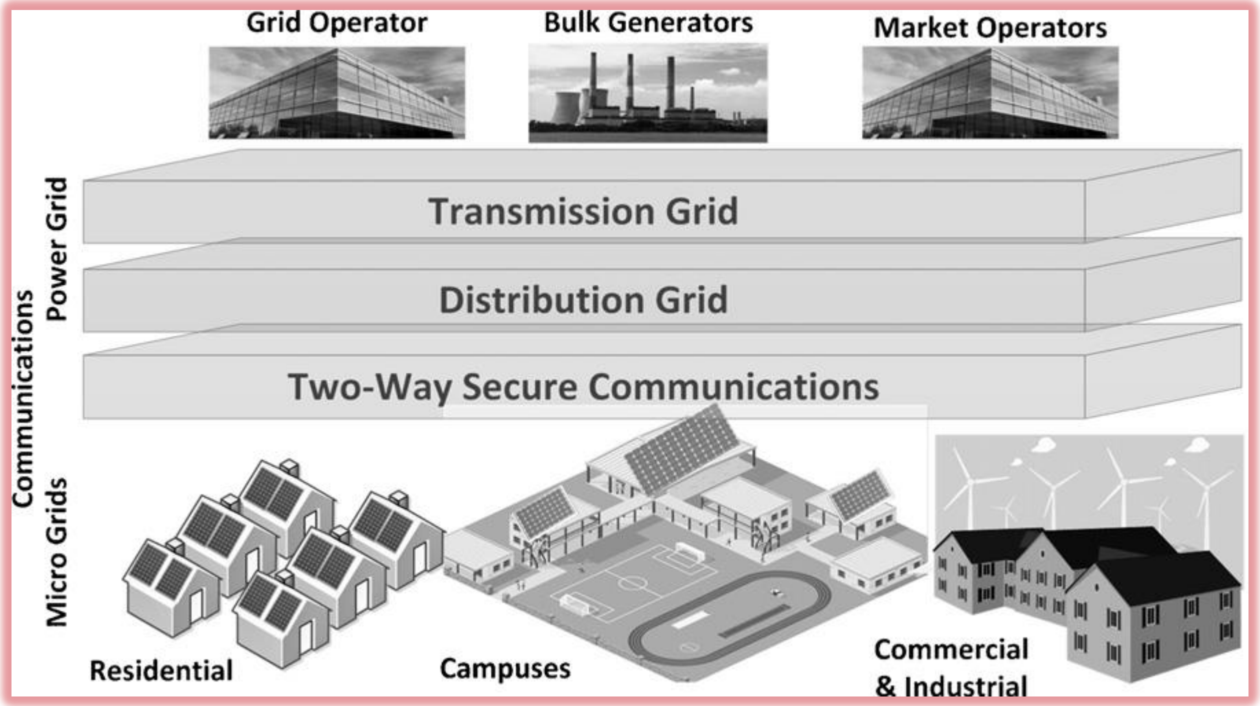
Model: DT

Each scenario corresponds to a specific train/test split:
S1: 3 months train, one month test
S2: 6 months train, one month test
S3: 11 months train, one month test

Estimation for a day



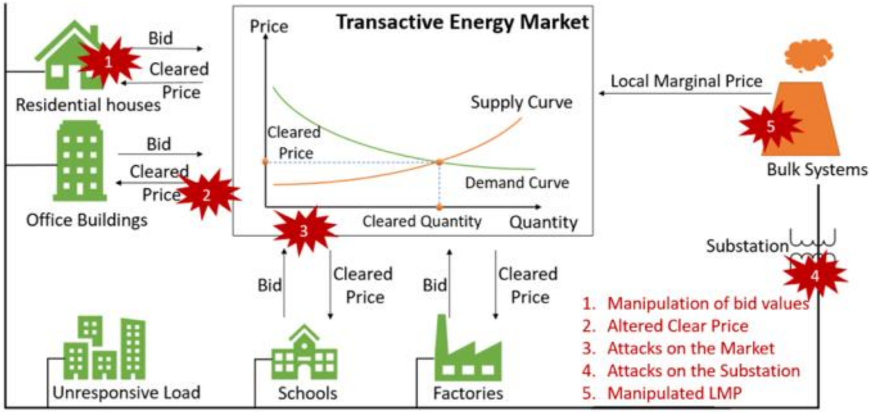
Use Case IV: Cyber-Physical Analytics for the Transactive Energy Systems



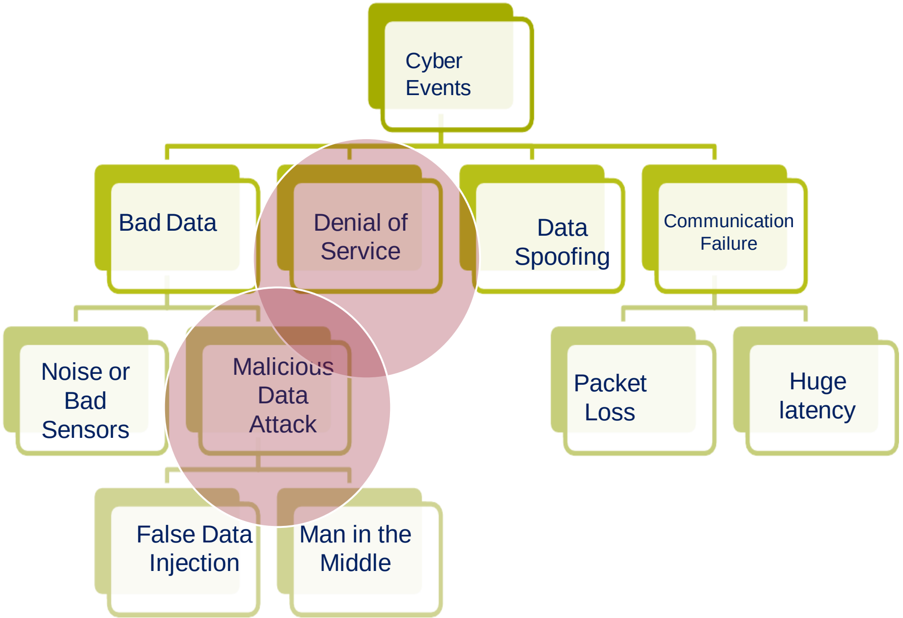
Source: Farrokh A. Rahimi, Ali Ipakchi, Transactive Energy Techniques: Closing the Gap between Wholesale and Retail Markets, The Electricity Journal, Volume 25, Issue 8, 2012, Pages 29-35

Data Analytics Based Anomaly Detection in TES

- Detecting malicious activity within a TES environment is challenging due to the diverse group of participants
  - Prosumers
  - Market Participants
  - Communication networks
  - Transmission and Distribution networks
- Systems and networks are vulnerable to diverse attacks
- Physical and cyber data are available for monitoring.
- Huge information flow : Manual detection will be difficult



Possible Cyber Events in TES



Denial of Service

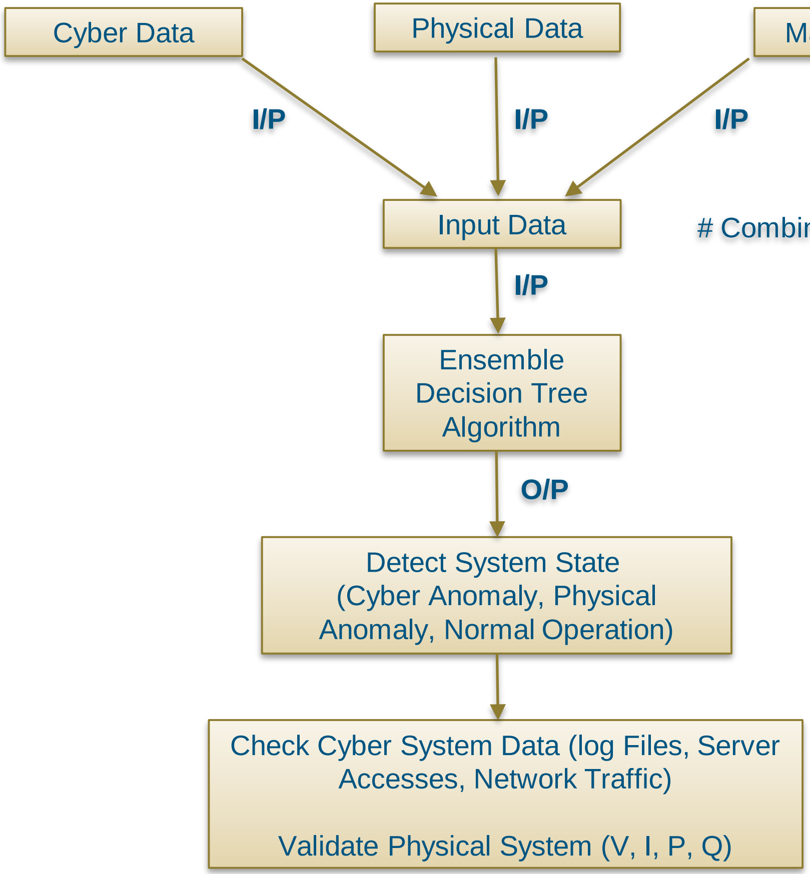
- An attacker sends several packets to the host in order to cause the unavailability of the resources.
- This can be detected by analyzing cyber data.

Manipulation of Bid Values

- The Bid Price and Bid Quantities are altered to random values through a malware injection.

- The impact can be seen in the cyber, physical, and market information.

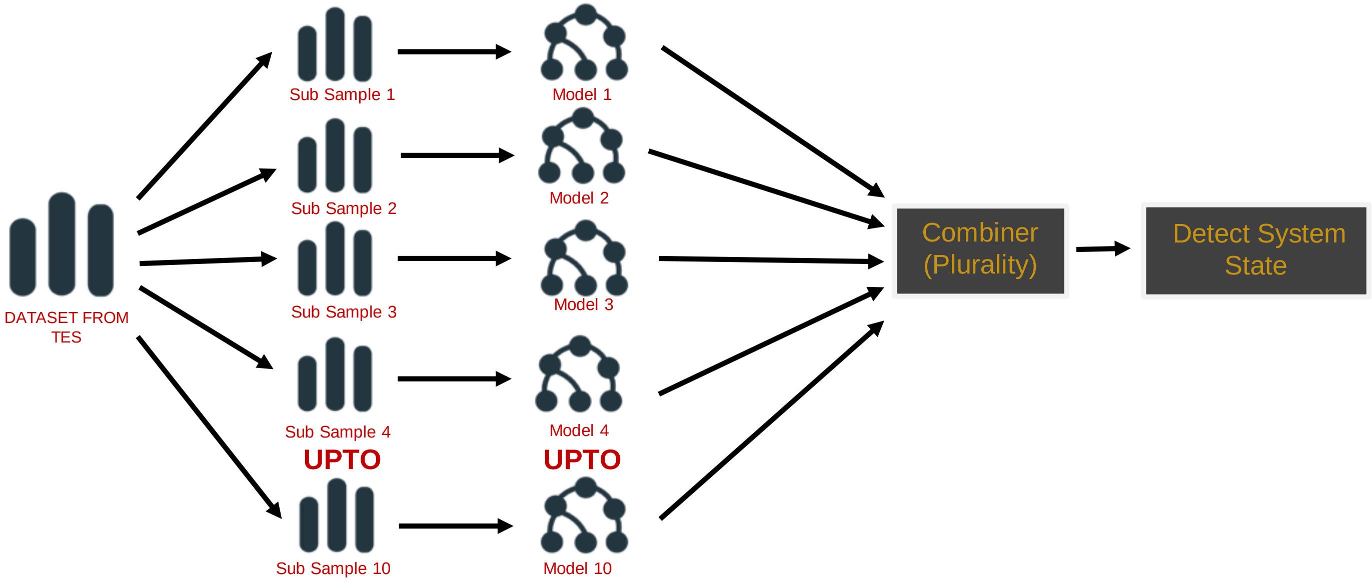
Architecture of Proposed Ensemble Decision Tree Algorithm



Combine all three types of data in to a single data frame

I/P : Input
O/P : Output

Ensemble Decision Tree Algorithm (Supervised)



- Given the data, the ensemble decision tree algorithm detects/classifies cyber and physical events in the Transactive Energy System.
- The proposed algorithm is based on the concept of bagging in which each single decision tree model is built from a random subset composing the ensemble of decision trees.
- "**Confidence**" and "**Probability**" measures are computed while predicting a class at a node.
- After all single decision tree model are created their prediction is combined using a combiner measure called **"Plurality"** and then the final class is predicted.

Ensemble Model: Measures

- CONFIDENCE:
  - Imposes a confidence at every node in an individual decision tree model.
  - The confidence is calculated using lower bound of "Wilson Score Interval" formula.
  - Core Idea of imposing 'confidence' is to balance the proportion of the predicted class with uncertainty of a small number of instances.

$$\frac{\hat{p} + \frac{1}{2n}z^2 - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{1}{n}z^2}$$

- $\hat{p}$ = proportion of instances for the predicted class at a particular node.
- n : the number of instances at that node .
- z = 1.96

- PROBABILITY:
  - The probability of a class at a certain node is the percentage of instances for that class at the node.
  - Additive smoothing tends to pull the class probabilities towards their overall proportion in the dataset.
  - If there are only a few instances in the leaf leading to the prediction, this smoothing can have a significant impact. If there are a high number of instances, the smoothing will have little effect.

$$\text{Probability\_class} = \frac{\text{node\_class\_count} + \text{class\_prior}}{\text{node\_total\_count} + 1}$$

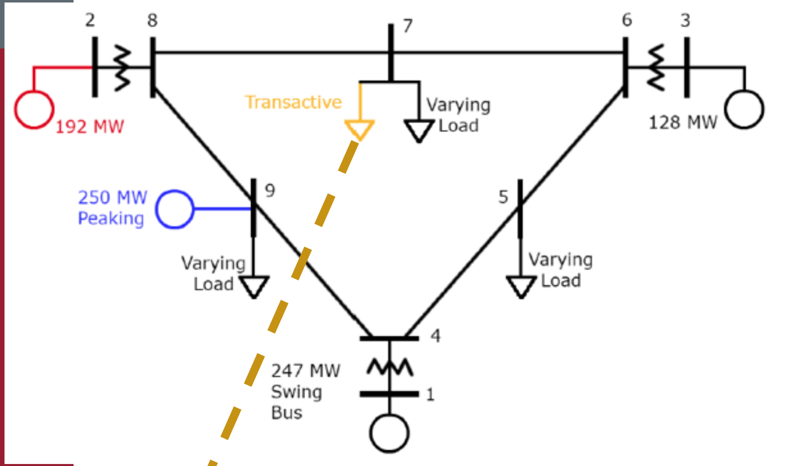
- node_total_count : Total number of instances in the node.
- node_class_count : Total number of instances for the class of interest at a given node.
- class_prior : Percentage of instances for that class over the total instances in the dataset.

Models	Prediction	Confidence
Model 1	True	80%
Model 2	False	95%
Model 3	True	40%
Ensemble	**True**	**60%**

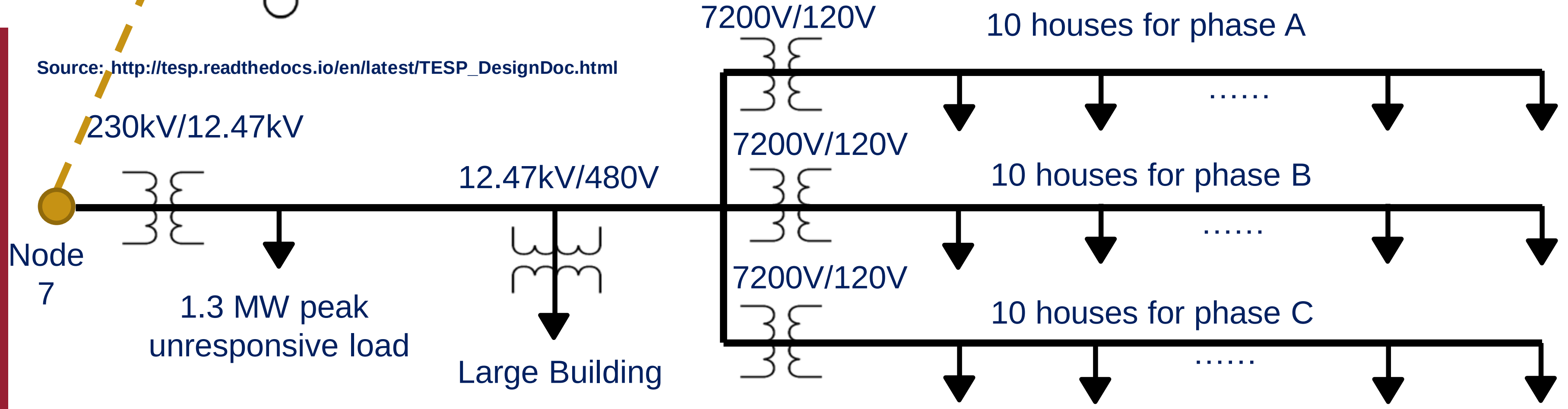
- PLURALITY :
  - Plurality is based on the premise that each model in the ensemble has one vote for a prediction.
  - The class with most votes is returned as the final prediction.
  - The resulting prediction is computed by averaging the confidence of the models that are predicting the right value.

Test System

The simulated power system includes a 9-bus transmission system and one feeder with transactive components at node 7. The HVAC devices in each house will participate in the power market.



Source: http://tesp.readthedocs.io/en/latest/TESP_DesignDoc.html



TESP Test bed

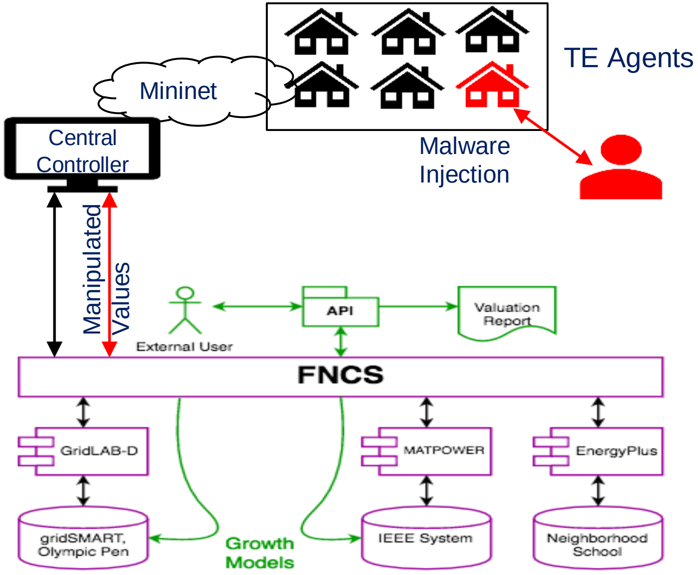
TESP is a framework designed by PNNL that simulates transactive systems. It includes various software modules and a number of agents in the form of smart houses.

- Mininet has been used to create a network that helps the house controllers get and send messages from/to the FNCS broker.

- The network has a tree topology with the central controller as the root node. There are 30 TE Agents. Each agent has a house controller.

- The house controller sends the temperature values, bid values to the FNCS broker, which publishes the information for the other components to read.

- The attack performed in the study aims at the house controller and manipulating the bid values which are then sent to the FNCS.



Source: http://tesp.readthedocs.io/en/latest/TESP_DesignDoc.html

Simulation Scenarios

Scenario 1 (High Impact): Manipulating bid price and quantity to a arbitrary large value. (Aim to impact the operation of power grid)

Scenario 2 (Low Impact): Manipulating bid price and quantity to some reasonable values. (Aim to gain benefits from the market)

Dataset Description

Input Features
Bid price
Bid quantity
LMP
Voltages
Generators real and reactive power outputs
Total demand
Flows
Class

Scenario 1

Output Labels	ΔBid Price	ΔBid Quantity
Normal	0	0
Cyber Attack	6	9
Physical Outage	0	0

Scenario 1 (High Impact):
- Manipulating bid price and quantity to an arbitrary large value.
- Aim to lead series of events in the operation of power grid.

Scenario 2

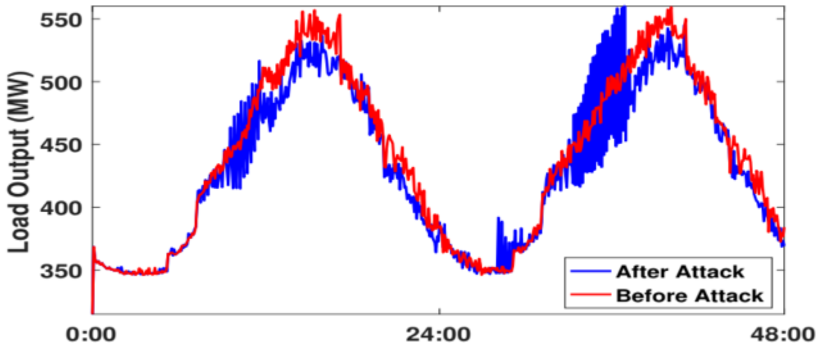
Output Labels	ΔBid Price	ΔBid Quantity
Normal	0	0
Cyber Attack	0.1 - 2.9	0.1 - 2.9
Physical Outage	0	0
Cyber Attack / Physical Outage	0.1 - 4.7	0.1 - 4.9

Scenario 2 (Low Impact):
- Manipulating bid price and quantity to some reasonable values.
- Aim to gain benefits from the market.

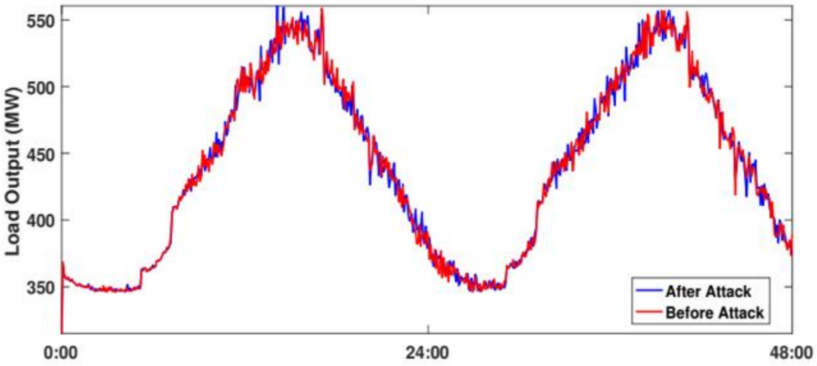
Trained-Tested using 80% and 20% split, respectively.

Results: Impact of Cyber Attack on TES

Scenario 1: Manipulating bid price and quantity to a arbitrary large value.



Scenario 2: Manipulating bid price and quantity to some reasonable values



Evaluation Metric Results

Confusion matrix is a table containing the predictions and actual values of objective field classes.
The intersection between actual values and predicted values yield four possible situations:

- True Positive (TP): Positive instances correctly classified.
- False Positive (FP): Negative instances classified as positive.
- True Negative (TN): Negative instances correctly classified as negative.
- False Negative (FN): Positive instances classified as negative.

Classification Measures:
Accuracy is calculated as the number of correctly classified instances over total number of instances evaluated.

$$\text{Accuracy} = \frac{TP + TN}{Total\ instances}$$

Precision is the percentage of correctly predicted instances over the total instances predicted for positive class.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall is the percentage of correctly classified instances over the total actual instances for the positive class.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F-Measure is a measure of test accuracy.

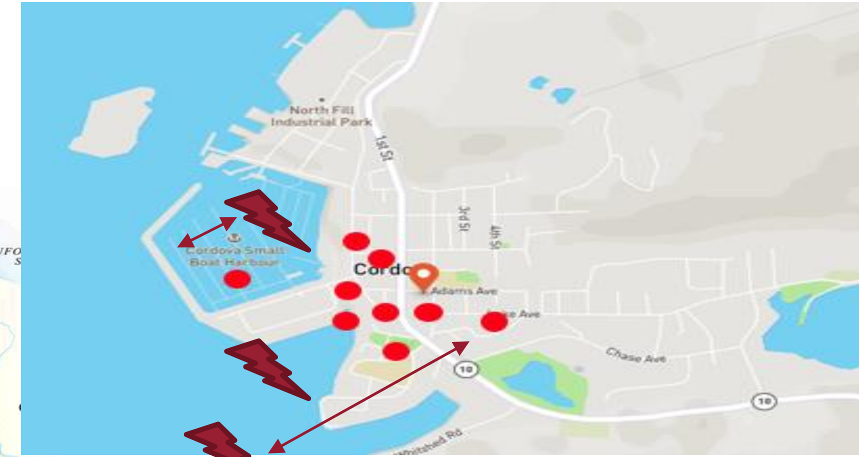
$$\text{F-measure} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Phi–Coefficient is the correlation coefficient between the predicted and actual values.

$$\text{Phi Coefficient} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Scenarios	Accuracy	Precision	Recall	F-Measure	Phi-Coefficient
Scenario 1	92.2%	92.0%	91.1%	0.92	0.83
Scenario 2	99.97%	83.32%	99.98%	0.87	0.89

Use Case V: Enabling Resiliency



Cordova

- Tsunamis
- Avalanches
- Power outages disrupting fishing business

Cordova, AK, serves as a perfect microcosm for prototyping innovations for the entire power grid across the nation

City of Cordova
ALASKA

NRECA
America's Electric Cooperatives

INL
Idaho National Laboratory

SIEMENS

Sandia National Laboratories

Pacific Northwest
NATIONAL LABORATORY

NRECA
America's Electric Cooperatives

ACEP
Alaska Center for Energy and Power

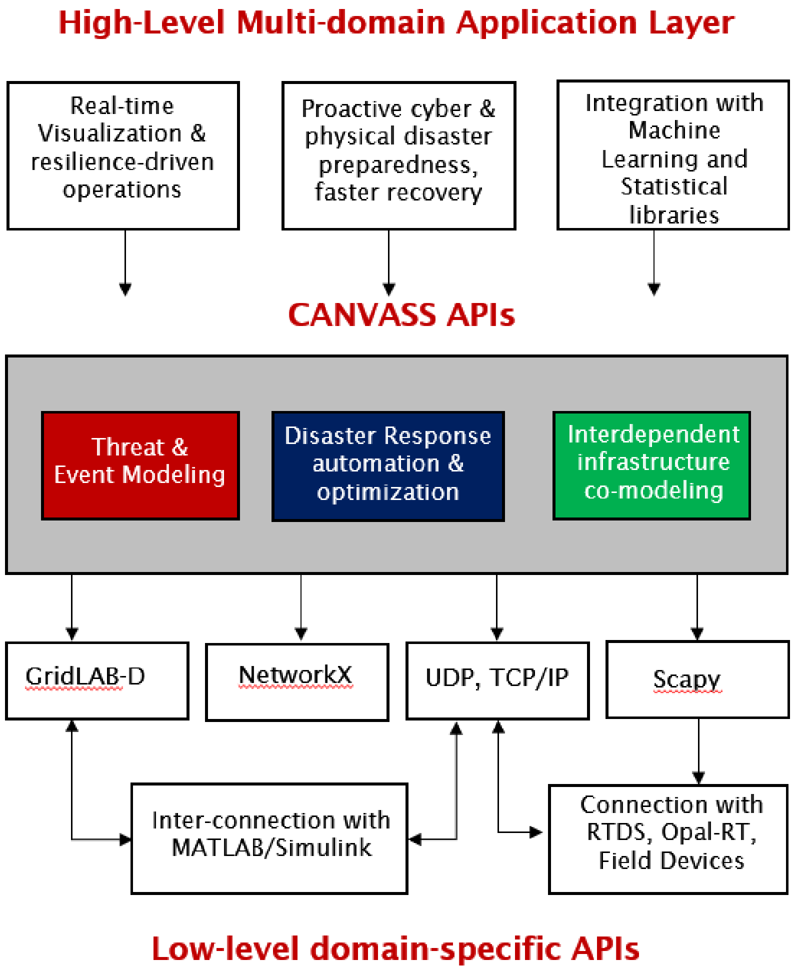
FLORIDA STATE UNIVERSITY
1851

NM STATE UNIVERSITY

CANVASS?

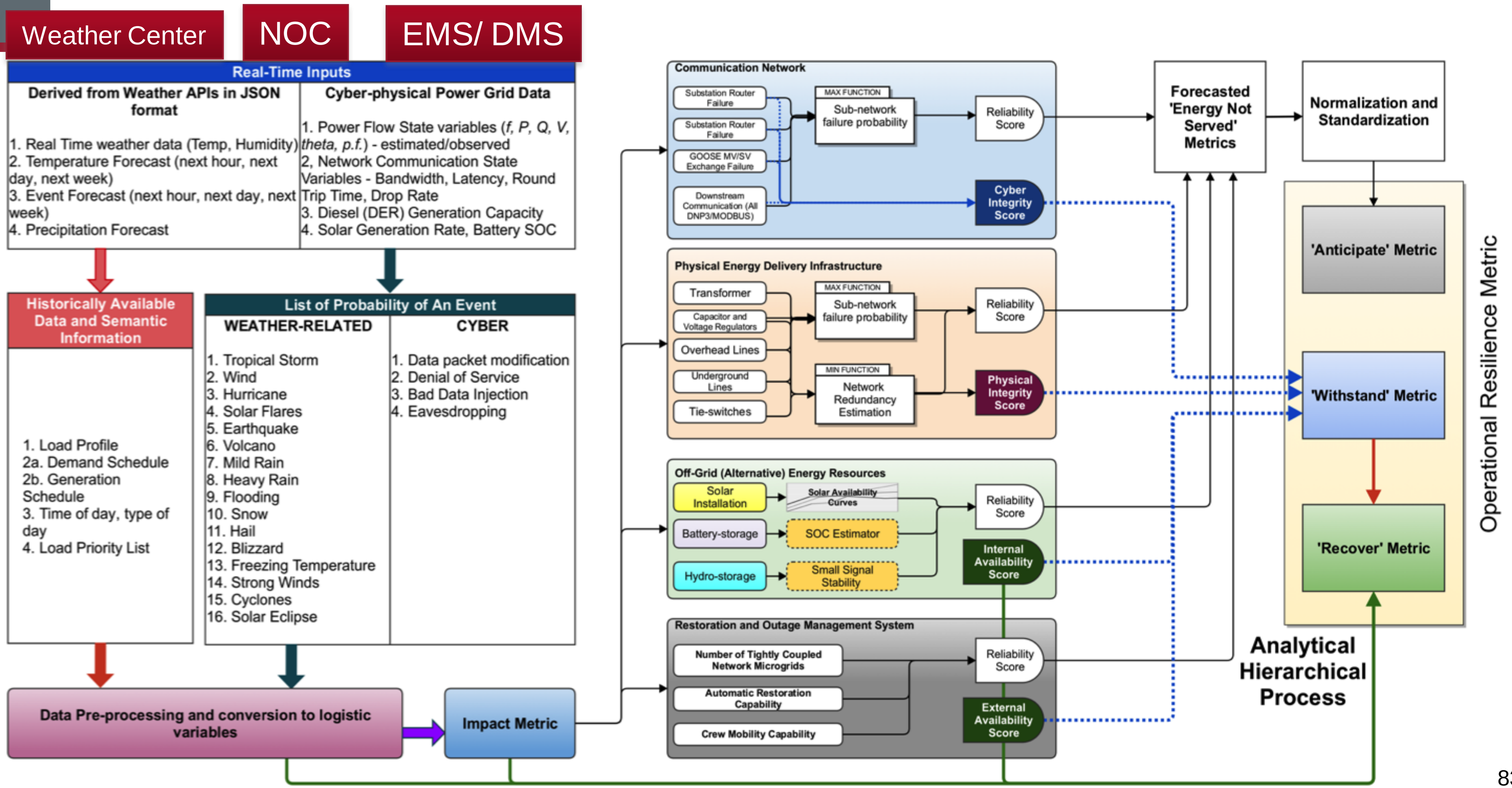
- Canvass stands for Cyber-Attacks and Network Vulnerability Analytics Software for Smart Grids
- It enables unfavorable physical and cyber event simulation for power systems
- **Free, open-source, platform-independent resiliency-computation toolkit**
- It has default restoration and resiliency computation algorithms – with ability for user to define own metrics and scenarios.
- It enables easy power system modeling and inter-disciplinary resiliency engineering research by abstracting lower level (hard-to-learn) open-source:
    - power simulation software [GridLAB-D],
    - network analysis library [NetworkX],
    - OS-based socket libraries [TCP/IP]
    - Packet Manipulation library [ScaPy]

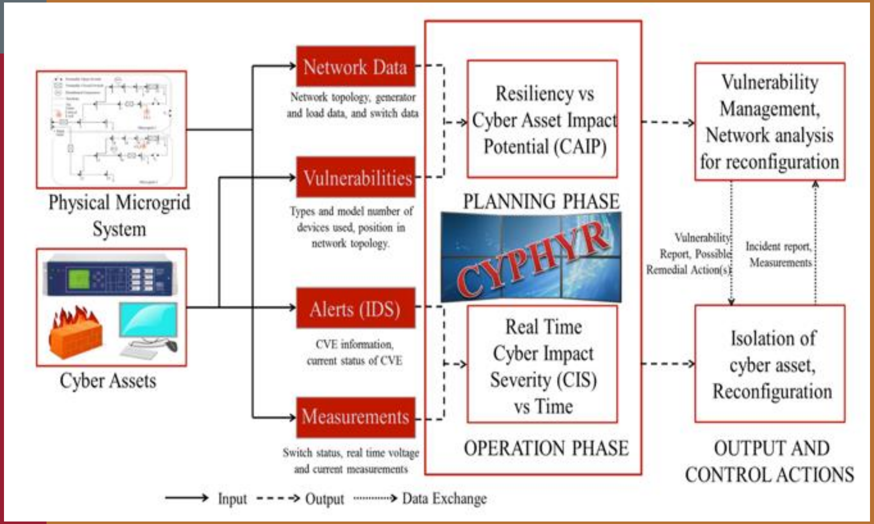
  into a single, easy-to-use Python package.
- Multiple interdependent infrastructure modeling, such as cyber-physical power grid, along with crew transport network.
- It can interface with Real-Time Simulation software through socket programming.

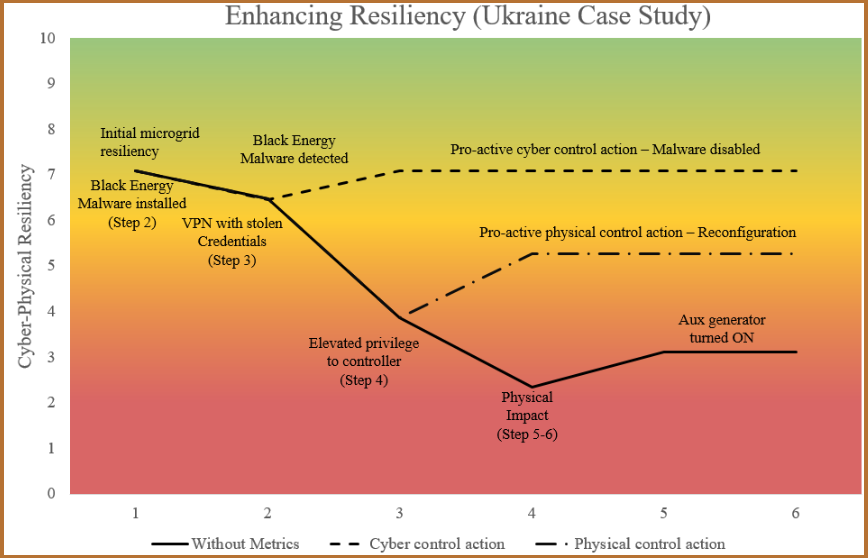


https://sgdril.eecs.wsu.edu/research-interests-and-grants/industrial-grade-products/pycanvass/

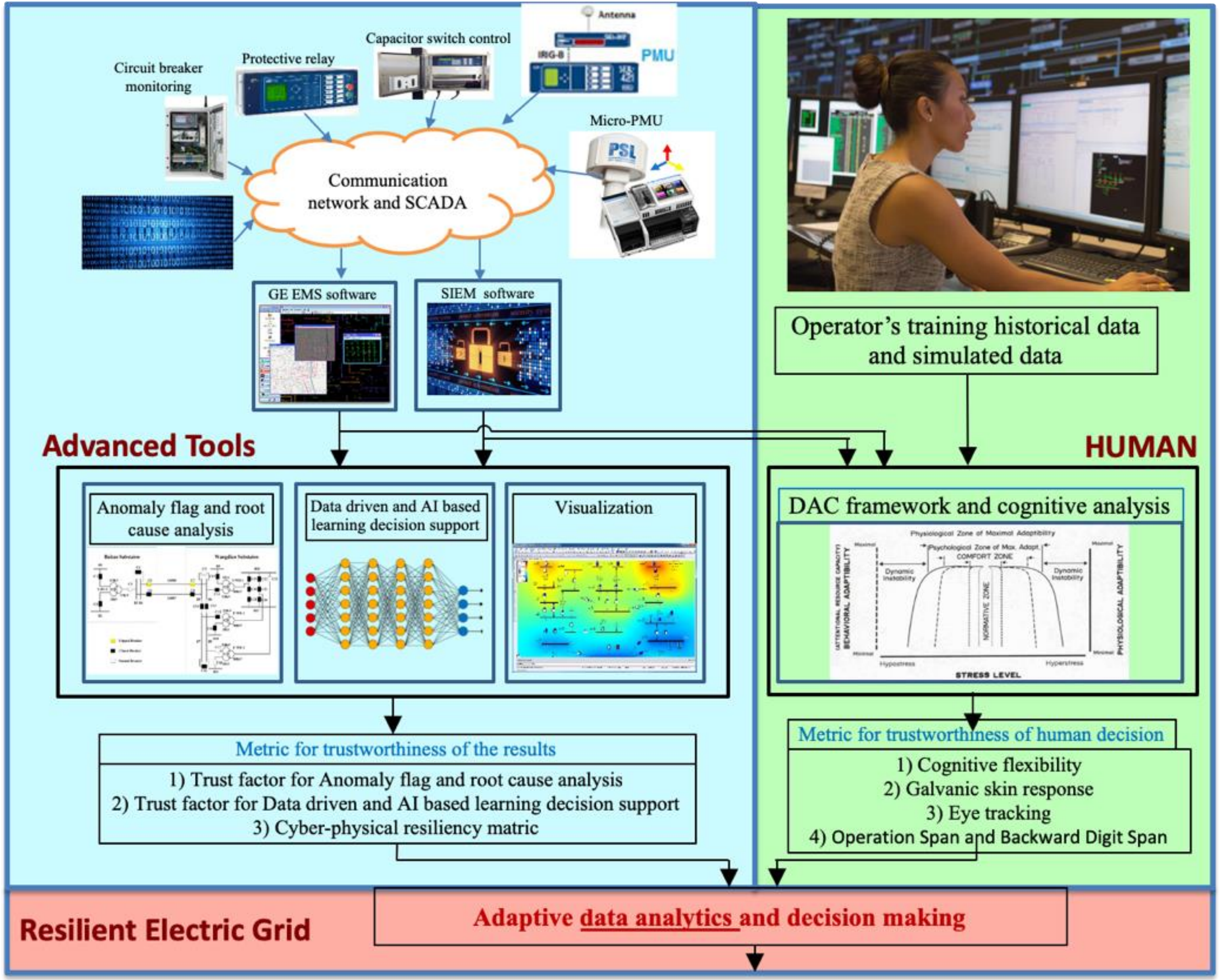
Measuring Resiliency



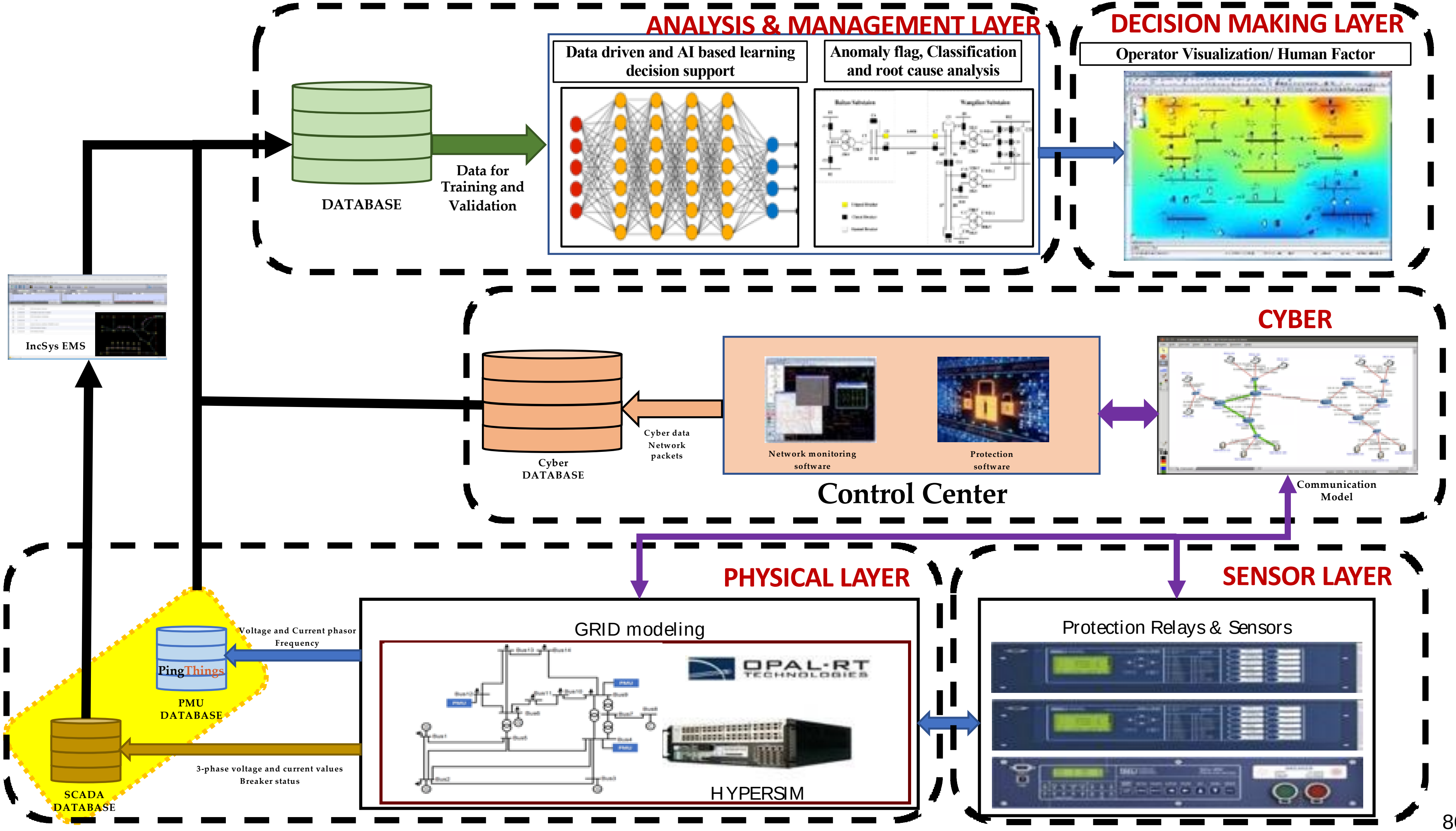




CyPhyR: Cyber-Physical Resiliency Tool



Circuit breaker monitoring
Protective relay
Capacitor switch control
Antenna
PMU
Micro-PMU
Communication network and SCADA
GE EMS software
SIEM software
Advanced Tools
Anomaly flag and root cause analysis
Data driven and AI based learning decision support
Visualization
Metric for trustworthiness of the results
1) Trust factor for Anomaly flag and root cause analysis
2) Trust factor for Data driven and AI based learning decision support
3) Cyber-physical resiliency matric
Operator's training historical data and simulated data
HUMAN
DAC framework and cognitive analysis
STRESS LEVEL
Metric for trustworthiness of human decision
1) Cognitive flexibility
2) Galvanic skin response
3) Eye tracking
4) Operation Span and Backward Digit Span
Resilient Electric Grid
Adaptive data analytics and decision making


ANALYSIS & MANAGEMENT LAYER
DATABASE
Data for Training and Validation
Data driven and AI based learning decision support
Anomaly flag, Classification and root cause analysis
DECISION MAKING LAYER
Operator Visualization/ Human Factor
IncSys EMS
CYBER
Cyber DATABASE
Cyber data Network packets
Network monitoring software
Protection software
Control Center
Communication Model
PHYSICAL LAYER
GRID modeling
OPAL-RT TECHNOLOGIES
HYPERSIM
SENSOR LAYER
Protection Relays & Sensors
Voltage and Current phasor Frequency
PingThings
PMU DATABASE
3-phase voltage and current values Breaker status
SCADA DATABASE

Takeaway#1: Resiliency is a Complex Problem

Power Grid Resiliency = **Flexible Infrastructure** ✖ **Secure Cyber Infrastructure** ✖ **Resilient Power Control Applications**

Flexible Infrastructure

Multiple switch
Macrogrid
Minigrid
Microgrid
Nanogrid
Graceful disintegration and interconnection
Flexible management and control of resources
Economic and market incentive

Secure Cyber Infrastructure

Communication
- Authentication
- Encryption

Computation
- Access control
- Attestation
- Forensics
- Patch management
- Software Audits

System Management
- Intrusion Detection
- Event Monitoring/Analytics
- Security Assessment

Resilient Power Control Applications

Generation
- Automatic Generation Control
- Governor Control
- Automatic Voltage Regulation
- Protection

Transmission
- State Estimation
- VAR Compensation
- Protection

Distribution
- Load Shedding
- Protection
- Advanced Metering Infrastructures

- Resiliency metric is a MCDM problem
- Resiliency is characteristics of the system

Takeaway #2: Finding Match in Data Analytics Techniques and Power System Problems is very important

- Data Analytics and machine learning approaches needs to be applied after analyzing the power system problem carefully. Finding match between machine learning strength and power system problem to be solved is important.
- Machine learning is only applicable in data-rich problems if no system model is available (e.g. forecasting)
- If model is available with rich data set, typically it will be two step approach: apply machine learning to narrow down your possible options and refine it with model based approach (e.g. event detection)
- Machine learning will not give a good results based on state of the art for highly complex and dynamic problems (e.g. transient stability, contingency analysis).
- Validation and metric is important for these evolving solution technologies

Takeaway#3: Get Involved in PMU Data Analytics and Applications

NASPI White Paper on Data Quality Requirements for PMU based Control Applications

IEEE Synchrophasor based Power Grid Operation as part of Bulk Power System Operation. White paper on a) Challenges and Solutions in Implementing PMU based Applications in Control Center) and b) Quality-Aware Applications

https://sgdril.eecs.wsu.edu/workshop_conferences/real-time-data-analytics-for-the-resilient-electric-grid/

Thank You

Acknowledgement: DOE, NSF, INL, Siemens, EPRI, PNNL, IEEE PES Big Data

Anurag K. Srivastava
anurag.k.srivastava@wsu.edu