Achieving Trustworthy Industrial Control Systems Using Data-Driven Models Informed by Physical Domain Knowledge

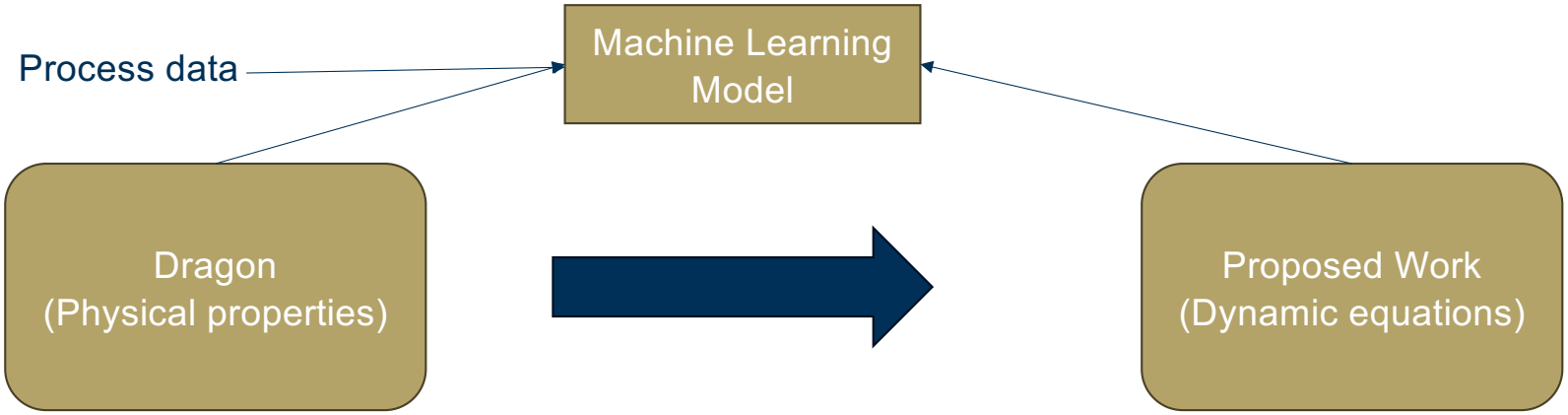
Wenke Lee

Georgia Tech

Proposed Approach: Machine Learning Informed by Physical Domain Knowledge

- Prior works are limited because it is infeasible to generate complete normal profiles for realistic sized industrial control system
- We propose to inform these machine learning model with knowledge of the physical process to produce model that can generalize to parts of the system not captured by training data
  - Attacks would have to match both the training data and the physical constraints to avoid detection

Proposed Framework



- Proposed framework informs machine learning models with physical domain knowledge
- Dragon demonstrates how physical properties can improve models
- Proposed work expands Dragon's detection abilities to be detect unknown attack types
- This is achieved by informing anomaly detection models with system dynamics equations

Preliminary Work:

DRAGON: Deep Reinforcement Learning for Autonomous Grid Operation and Attack Detection (ACSAC 2022)

Georgia Tech

Related Work

Reliability

- Manual and expert systems based approaches [1] do not scale to realistic size systems
- Prior autonomous operations research [2, 3] does not consider realistic threat models where attacks inject commands and spoof sensor measurements

Detection

- Prior anomaly detection systems [4, 5] can be bypassed by sophisticated attackers who use standard protocols and blend into normal behavior
- Specification based approaches [6] do not scale to realistic sized systems

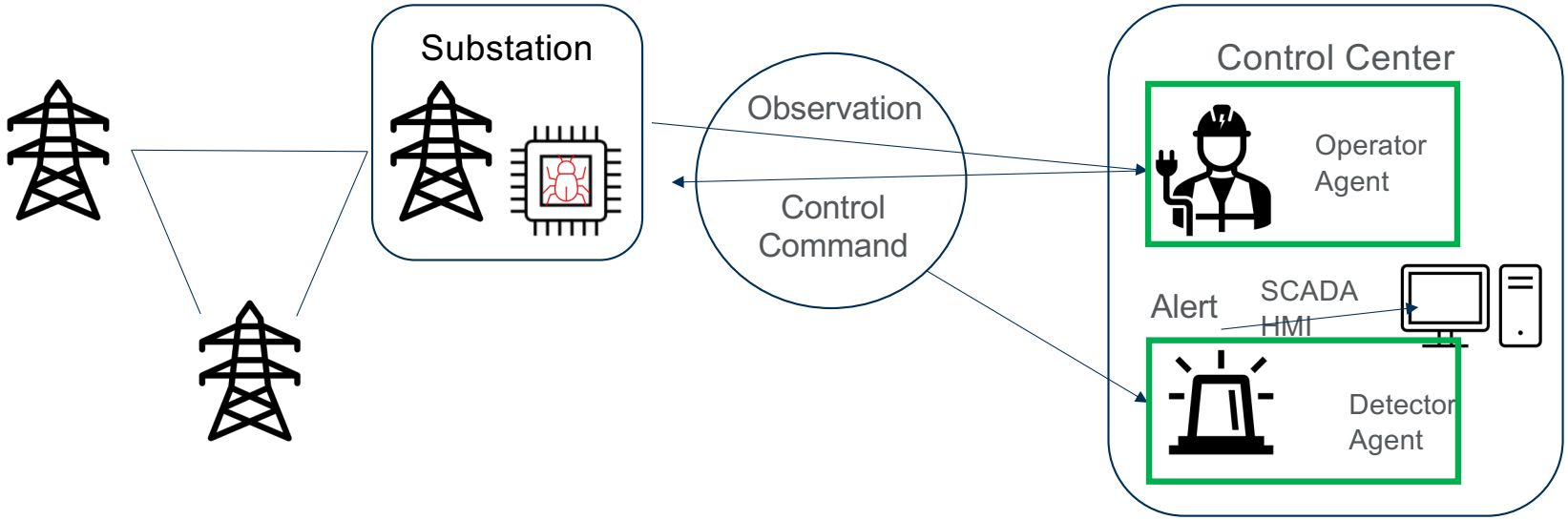
[1] A. Marot, B. Donnot, S. Tazi, and P. Panciatici, “Expert system for topological remedial action discovery in smart grids,” 2018.
[2] T. Lan et al., “Ai-based autonomous line flow control via topology adjustment for maximizing time-series atcs,” in 2020 IEEE Power & Energy Society General Meeting (PESGM), IEEE, 2020, pp. 1–5.
[3] A. Marot et al., “Learning to run a power network challenge: A retrospective analysis,” arXiv preprint arXiv:2103.03104, 2021.
[4] S. Ponomarev, Intrusion Detection System of Industrial Control Networks using Network Telemetry. Louisiana Tech University, 2015
[5] H. R. Ghaeini, D. Antonioli, F. Brasser, A.-R. Sadeghi, and N. O. Tippenhauer, “State-aware anomaly detection for industrial control systems,” in Proceedings of the Annual ACM Symposium on Applied Computing, 2018, pp. 1620–1628.
[6] D. Formby and R. Beyah, “Temporal execution behavior for host anomaly detection in programmable logic controllers,” IEEE Transactions on Information Forensics and Security, vol. 15, pp. 1455–1469, 2019.

Georgia Tech

Insights

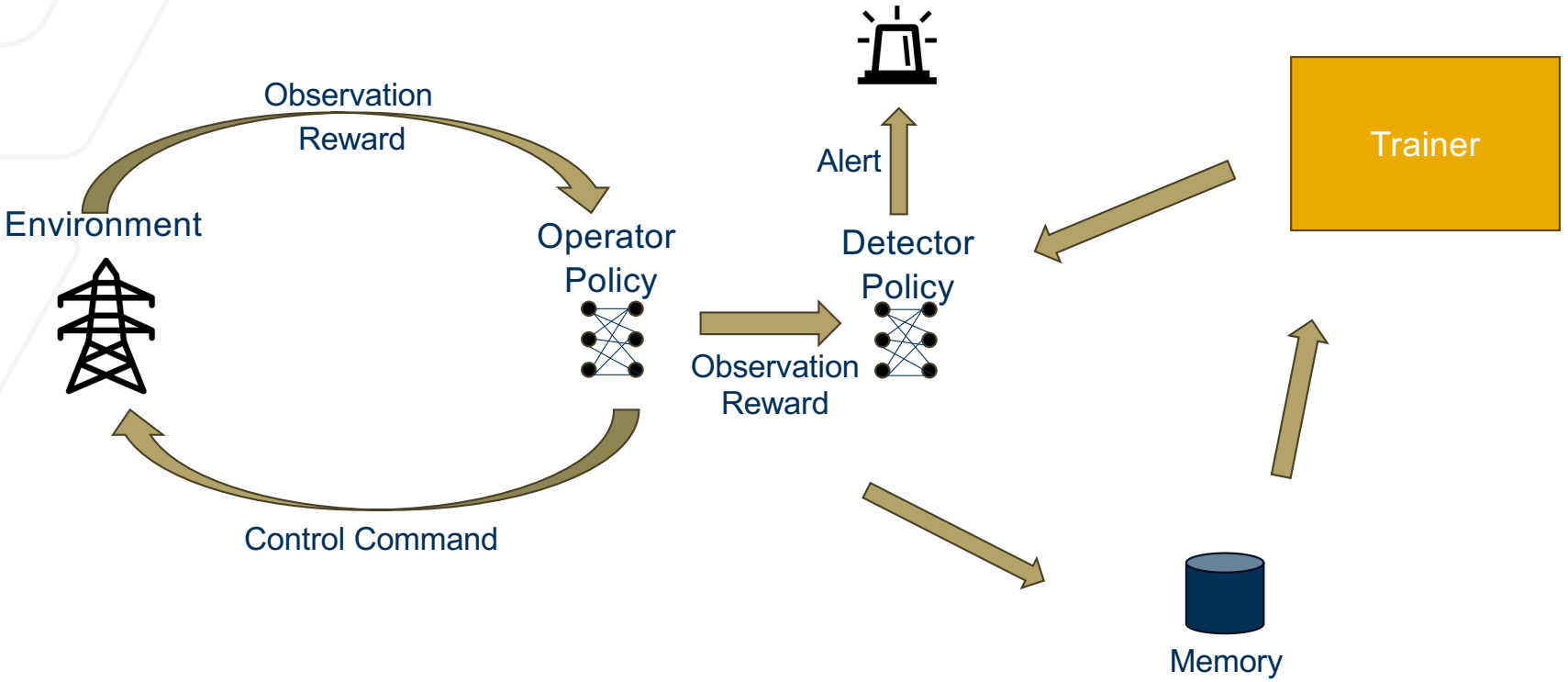
- Reliability can be achieved by ensuring the system remains in a stable state
- There are heuristics about the power grid that describe a stable state
- Although it can be challenging to label states with the best commands, we can score states based on reliability and learn actions that result in reliable states
- To detect attacks, labeling states to train can be challenging in adversarial environments
- Instead, reinforcement learning can learn to detect attacks with limited knowledge

Dragon Overview



Georgia Tech

Dragon Workflow



Threat Model

- Attacker can disconnect power lines
- The attacker can also inject false measurements into grid observations

Operation Agent

Observation

- Load, generator, line attributes

Actions

- Grid topology modifications

Rewards to represent reliability

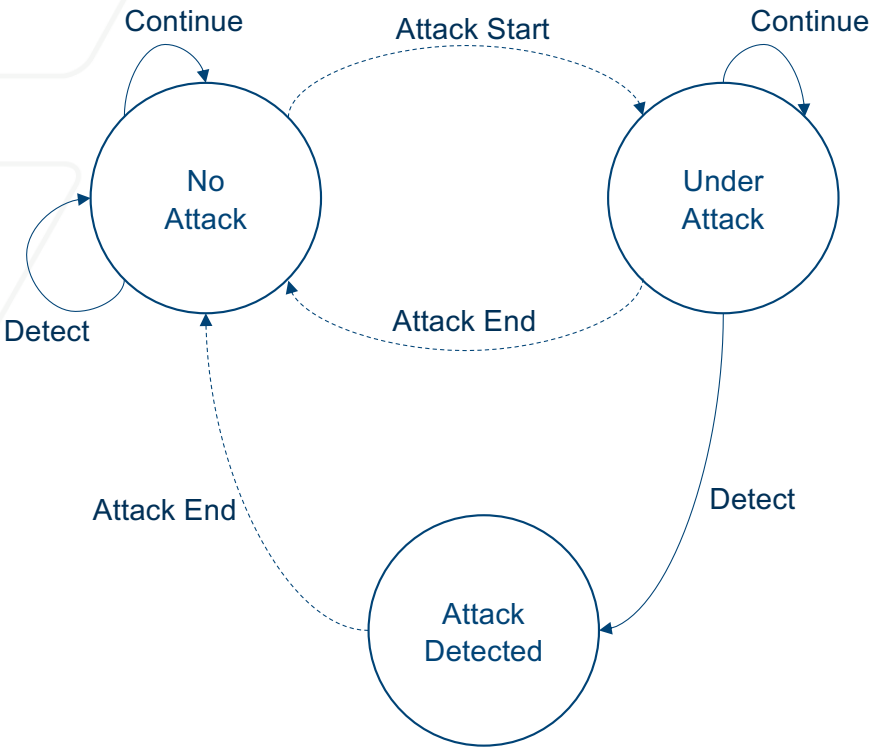
- All loads should receive sufficient power
- Lines obey their thermal limits

$$-\sum_{l\in\mathcal{L}} \frac{power_l}{thermal\_limit_l}$$

$$r = c_1 \frac{\sum_{l\in\mathcal{LD}} |path_{gen\sim l}|}{|\mathcal{LD}|} - c_2 \sum_{l\in\mathcal{L}} \frac{power_l}{thermal\_limit_l}$$

Detection Component

Environment



Observation

- Previous and current grid observations
- Operator's actions

Actions

- Detect attack
- Continue normal operation

Rewards

- Penalize false positives and false negatives

Georgia Tech

Proposed Work: Physics Informed Attack Detection

Georgia Tech

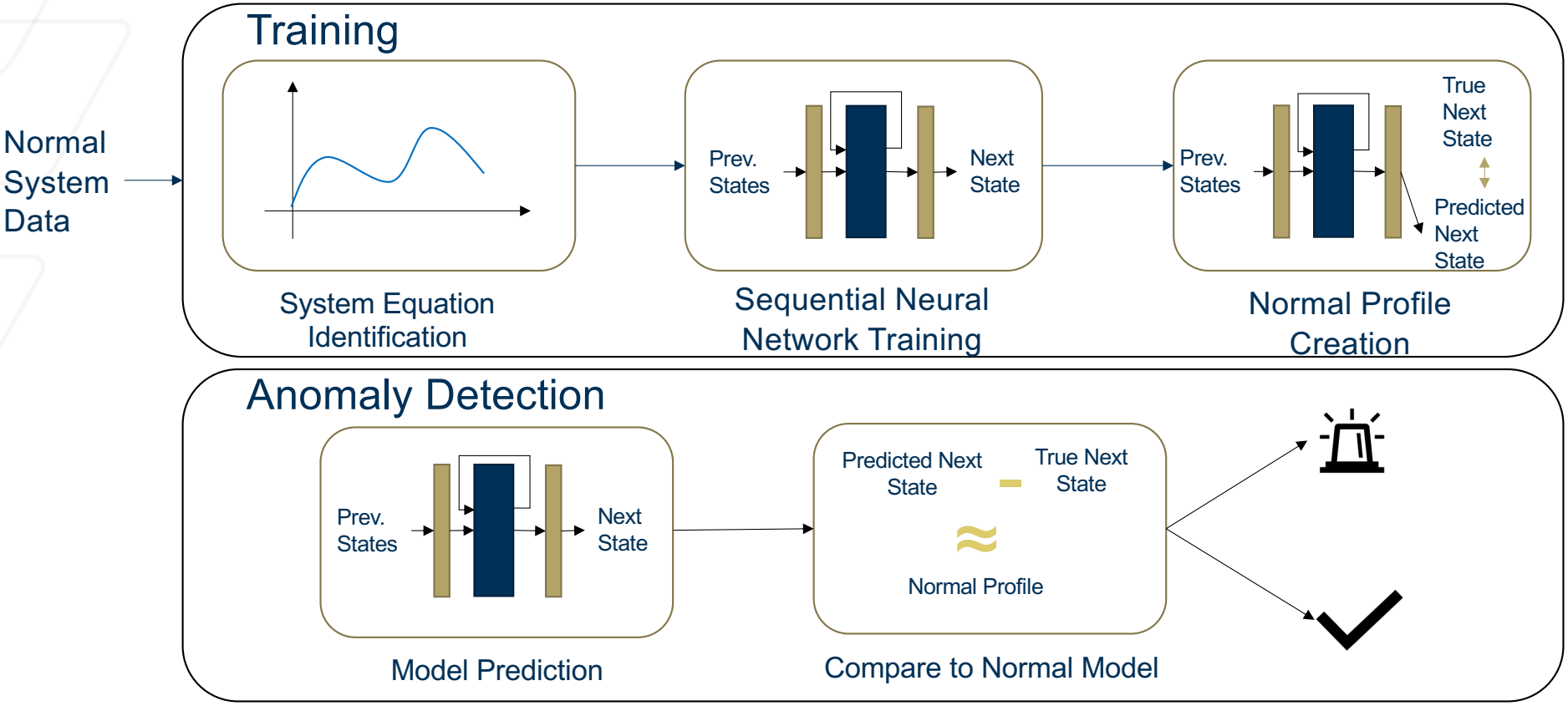
Motivation

- Sophisticated attacks frequently employ novel exploits and tools
- Previous work that uses signature-based detection cannot detect these unknown attacks
- Anomaly detection works with only normal data but requires a complete normal profile to avoid false positives

Our Insights

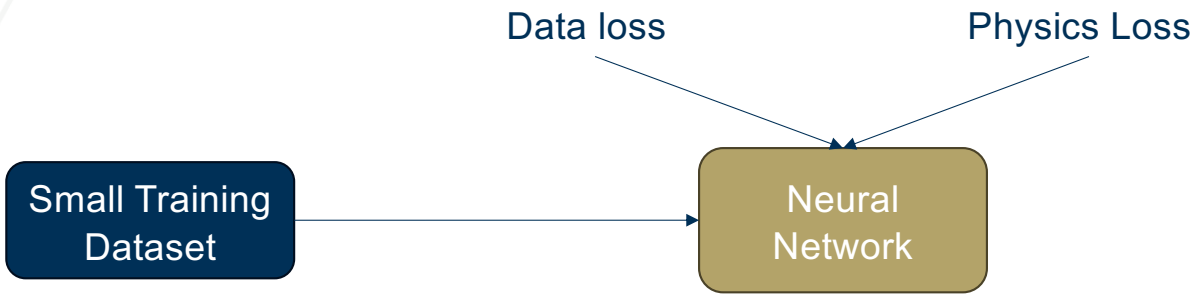
- The physical component of an ICS evolves according to known physics
- This physics can be sometimes represented by equations that relate previous states to the current state
- Such information can be used to supplement incomplete training datasets
- System equations provide a mechanism to fill in the missing parts of an incomplete training dataset
- If a normal state does not fit the training dataset, it should still follow the system equations and not be classified as an attack

Overview



Georgia Tech

Physics Informed Neural Network (PINN)



- PINNs try to capture non-linear functions where the data follows known physical laws
- Frequently applied when a small set of training data exists
- The physical laws are encoded into additional loss functions

M. Raissi, P. Perdikaris, and G. E. Karniadakis, “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations,” Journal of Computational physics,vol. 378, pp. 686–707, 2019.

Training With System Equations

- Each system equation is translated to a network loss term
- Example: System identification equation

$$y_{n+1} = \sum_{i=0}^{k} a_i y_{n-i} + \sum_{i=0}^{k} b_i x_{n-i}$$

- Given this equation relating state variables x and y, we produce the following network loss

$$L(\hat{y}_{n+1}) = \left| \hat{y}_{n+1} - \sum_{i=0}^{k} a_i y_{n-i} + \sum_{i=0}^{k} b_i x_{n-i} \right|$$

- As the loss decreases, the predicted state follows the equation more accurately

Initial Results – SWaT Dataset

Method	True Positive Rete	False Positive Rate	F1 Score	Precision	Detection Delay
Physics informed	**74.96**	**7.5**	**65.33**	**57.90**	**20**
Invariants	73.06	7.62	63.97	56.90	54
LSTM	70.97	16.78	48.45	36.78	1306

- First case study was on the standard SWaT dataset
- Our physics informed solution outperforms the NDSS19 invariant paper and a standard LSTM

Potential Impacts

- A framework to detect sophisticated ICS attacks based on physical domain knowledge and deep learning algorithms
  - Data and software will be made available to members of the consortium, as well as the broader research and industry communities (when appropriate)
- Use power grid as the real-world application for the proposed research and validation