Our models matter: The impact of cyber-attacks on ML/AI-based ICS

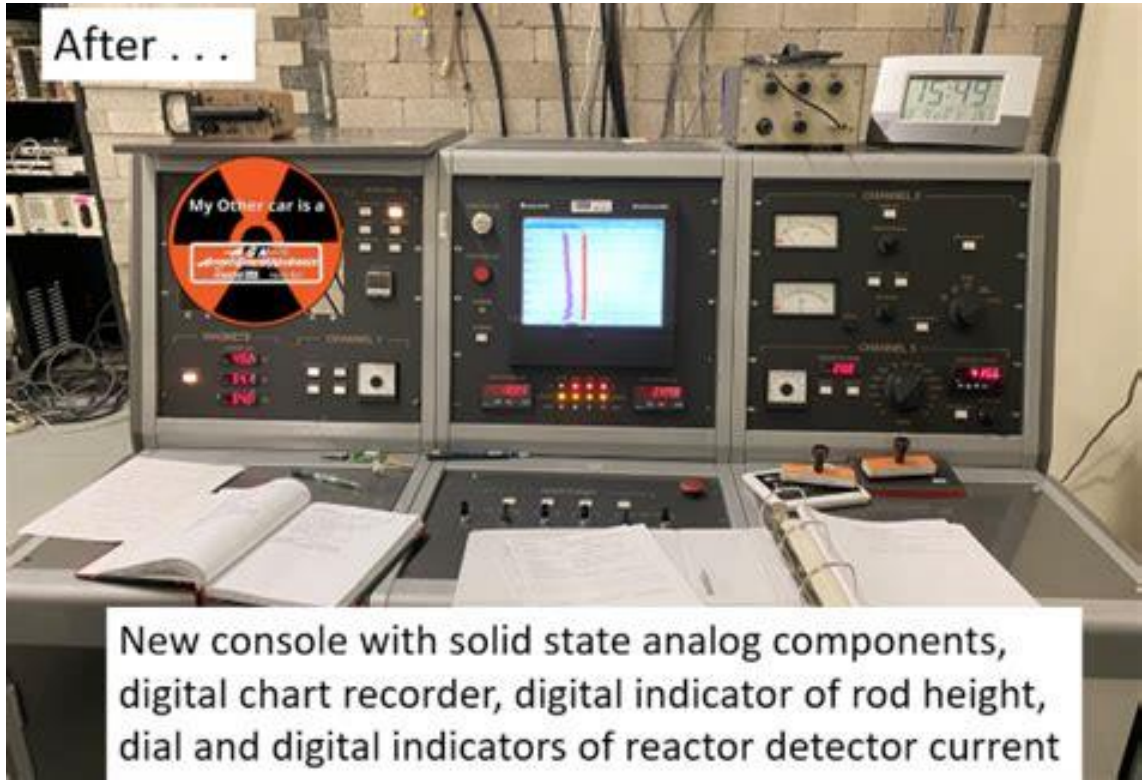
Mrs. Patience Yockey
Georgia Institute of Technology | Nuclear Cybersecurity | Machine Learning
plamb6@gatech.edu

Originally presented at the Informatik Festival 2023 on September 28th 2023

Georgia Tech

Vacuum Tubes to Solid State

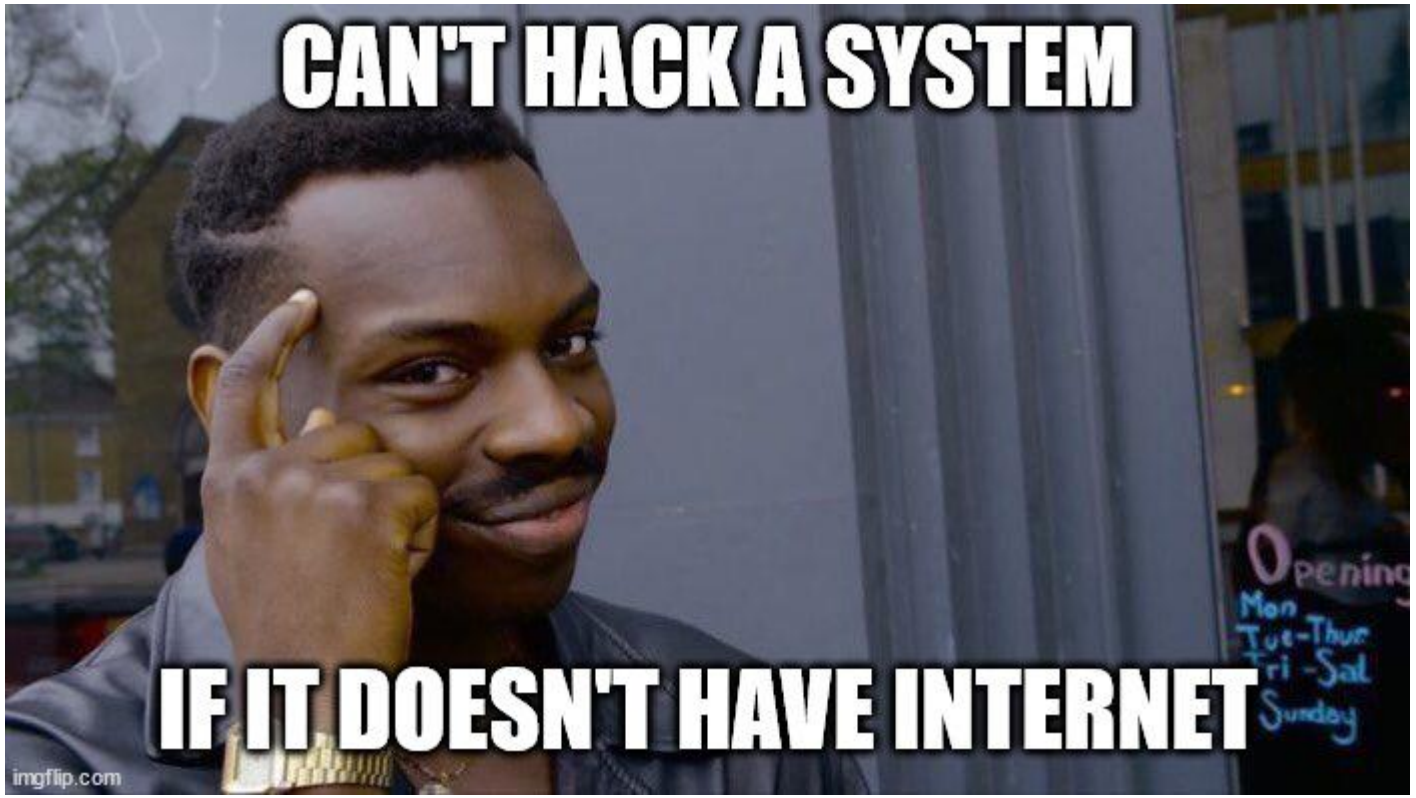
https://www.eastidahonews.com/2015/08/idaho-state-university-celebrates-50-years-running-nuclear-reactor/



https://www.isu.edu/ne/nuclearscienceweek/

Georgia Tech

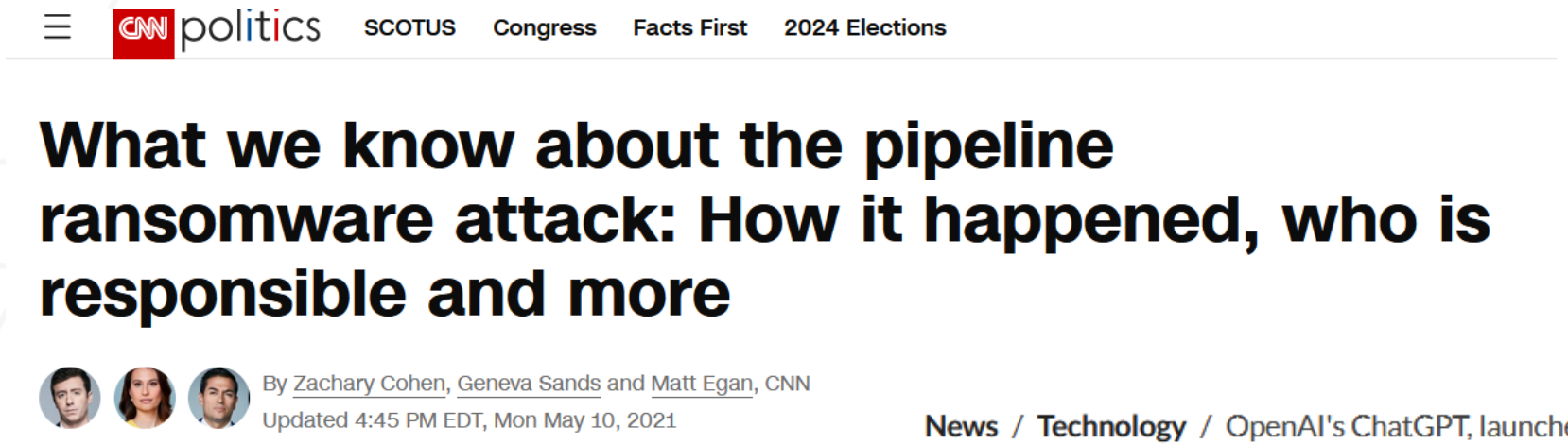
Understanding Future Threats



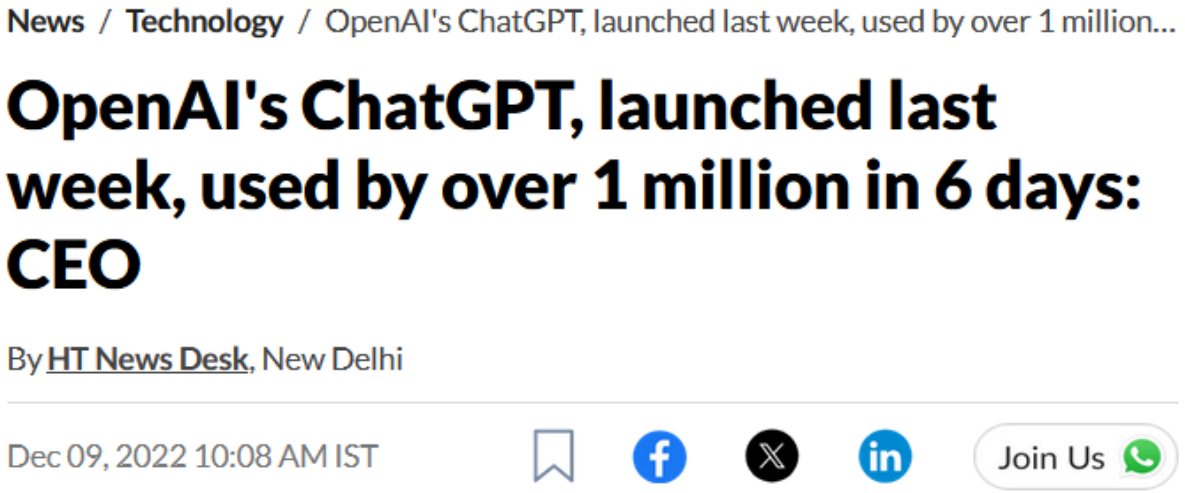
https://imgflip.com/memegenerator/Roll-Safe-Think-About-It

Georgia Tech

And Then… We Needed Them…



https://www.cnn.com/2021/05/10/politics/colonial-ransomware-attack-explainer/index.html



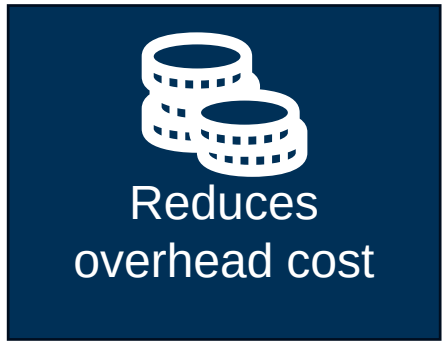
https://www.hindustantimes.com/technology/openais-chatgpt-launched-last-week-used-by-over-1-million-in-6-days-ceo-101670234260469.html

Georgia Tech

Use of ML/AI in the Energy Sector

- The use of ML/AI in the energy sector is expected to grow by 29.88% between 2022 and 2029 [2].

- Of particular interest to the energy sector is using ML-based digital twins (DTs) to remotely monitor or autonomously control industrial control systems (ICS).



- **The question arises: are we adequately prepared to introduce ML/AI into ICS?**

Georgia Tech

The (Potentially) Perfect Storm…

- Cyber-attacks against the energy sector increased:
  - 380% between 2014-2015 in the EU [18]
  - Over 500% in the US over the past ten years [3]
- Microsoft:
  - Estimated 89% of government, non-profit, and Fortune 500 companies do not have the right tools or skills to secure their ML systems [16].
  - 30% of all AI cyber-attacks will leverage training-data poisoning, model theft, or adversarial samples to attack AI-powered systems [15].

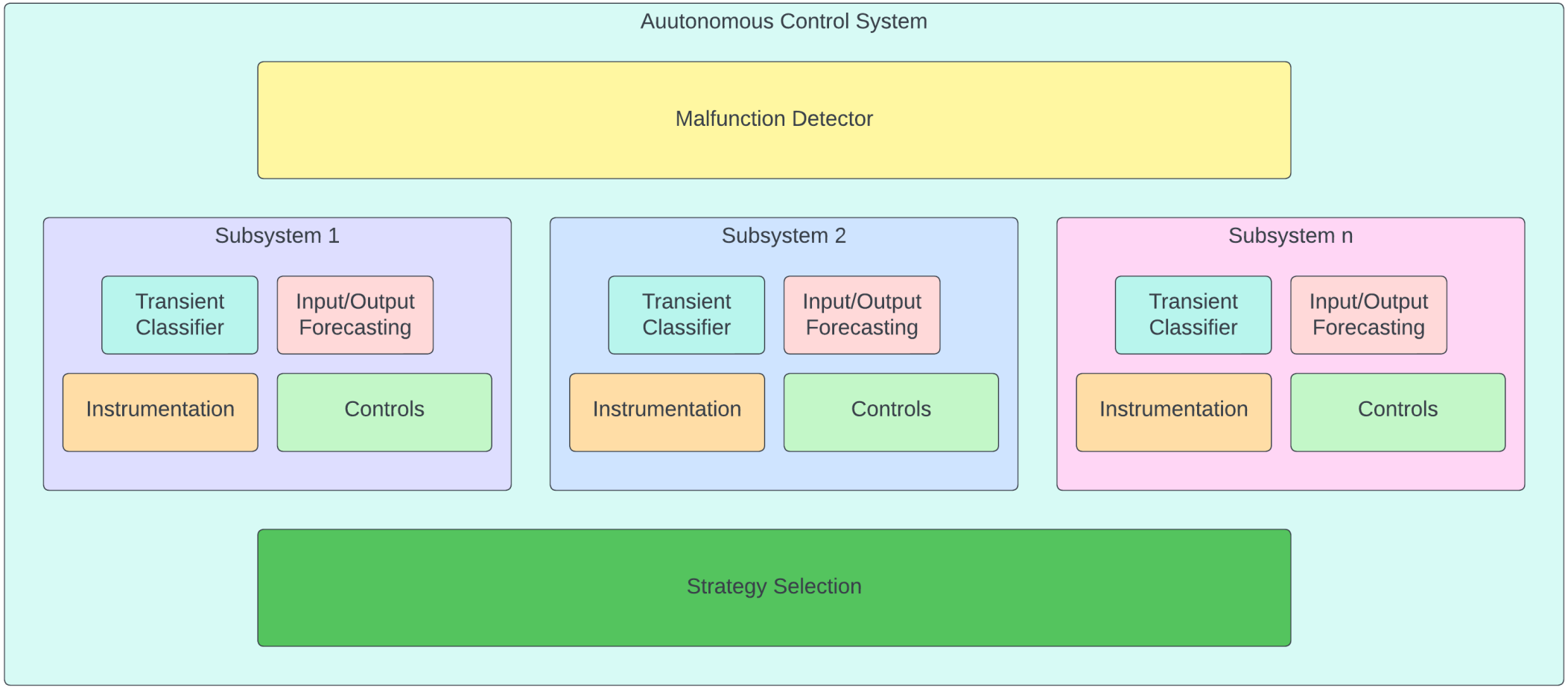
https://imgflip.com/memegenerator

Georgia Tech

Questions unanswered

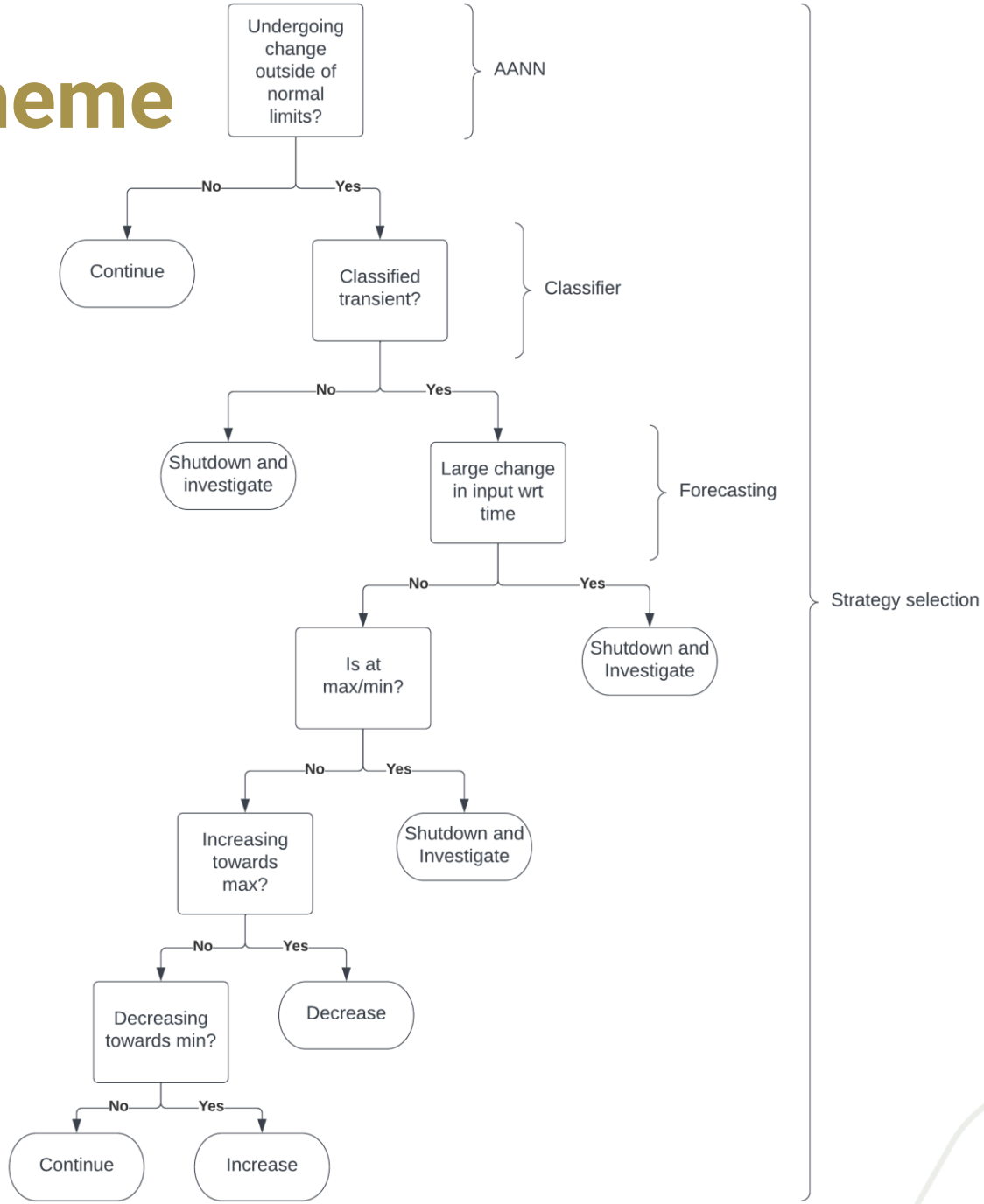
- How can ML-based DTs be implemented in autonomous control systems (ACS)?
- What are the risks of various cyber-attacks on ML-based DTs for ACS?
- Does ML architecture or development framework impact cyber-risk?
- How can we protect ML-based ACS against cyber-attacks?

Autonomous Control System Digital Twinning (DT)

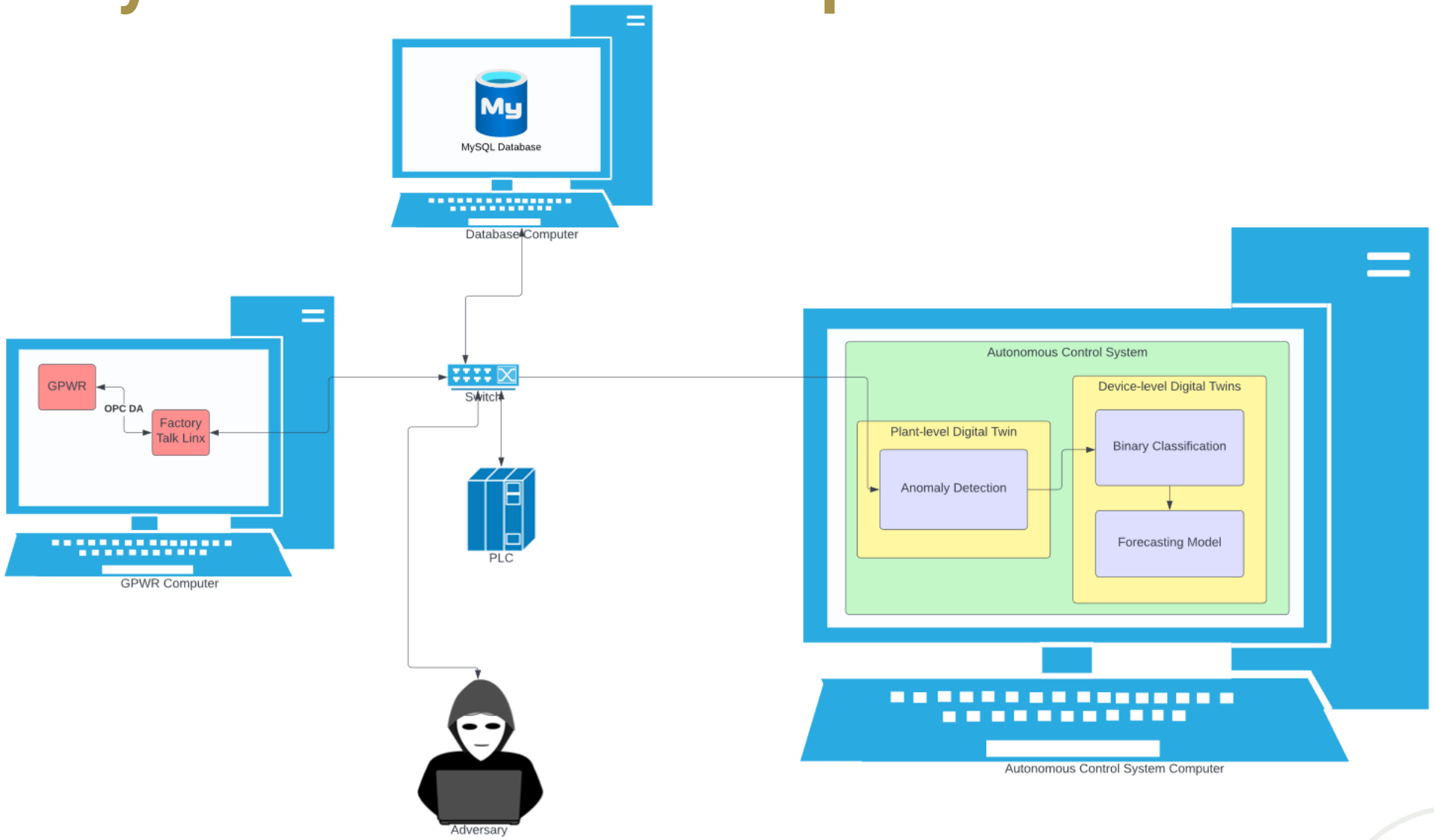


Georgia Tech

Strategy Selection Scheme



Cyber-Physical Testbed Development

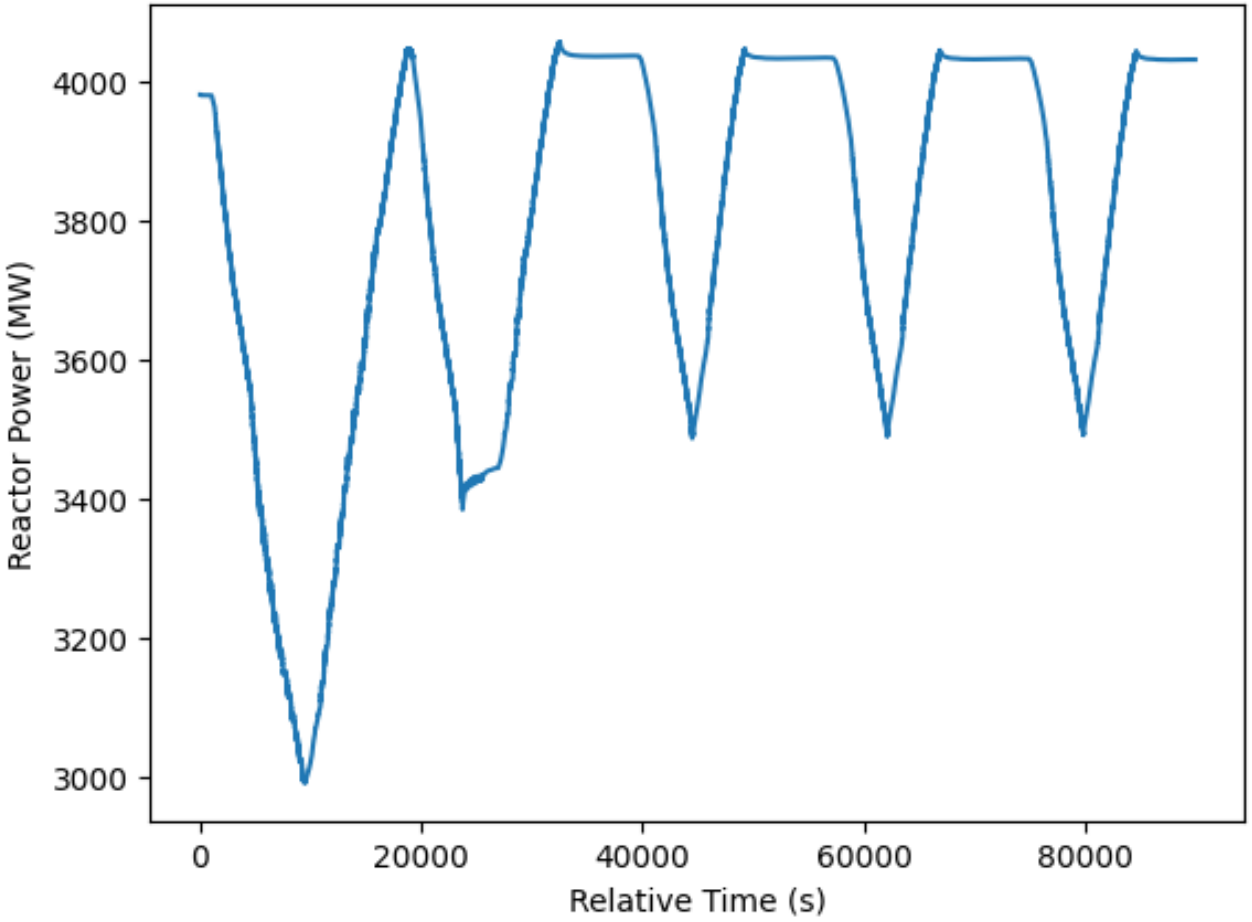


In this scenario, the reactor is “air-gapped” from unsecured networks, meaning the adversary must have physical access to the system to launch attacks.

Georgia Tech

Data Collection and Storage

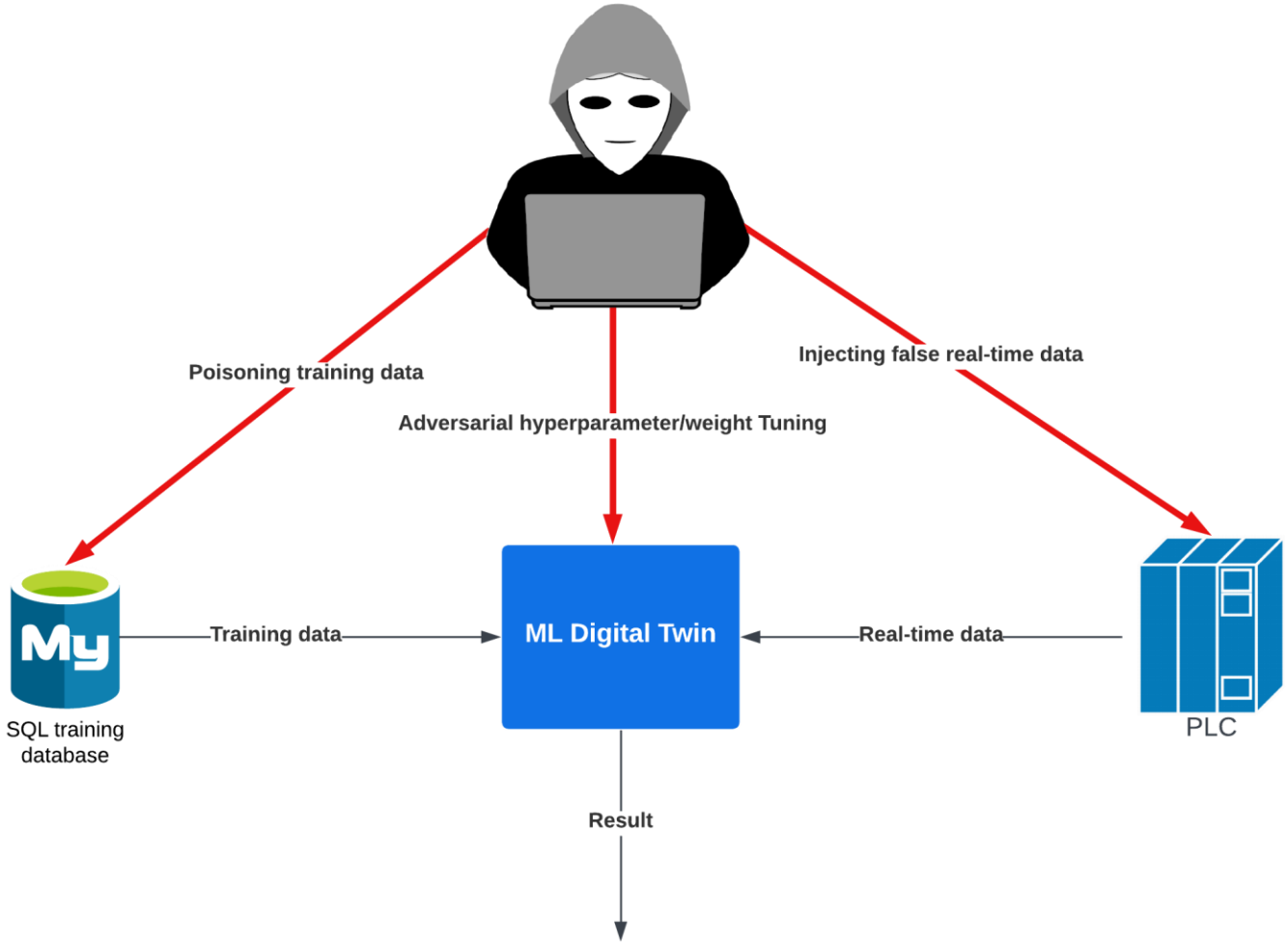
Transient and Steady State Reactor Power Training Scenario



- On **training**, the ACS queried a separate system **MySQL database** [1] to obtain the training dataset for each DT.
  - Training data was obtained assuming a beginning of life (BOL) scenario.
    - Power was ramping over 420 minutes shown in figure.
    - The MySQL database pulled 70 variables related to steam generator 1 and overall plant-health .
    - Training data was split into 70% training 30% validation.
- **Real-time data** was ingested by a **Pylogix** [5] call within the ACS to convert PLC CIP packets into usable dataframes.
  - Real-time data was 100% power BOL conditions to determine cyber-attack effects.
  - Data was pulled for all 70 variables every 50ms according to PLC speed.

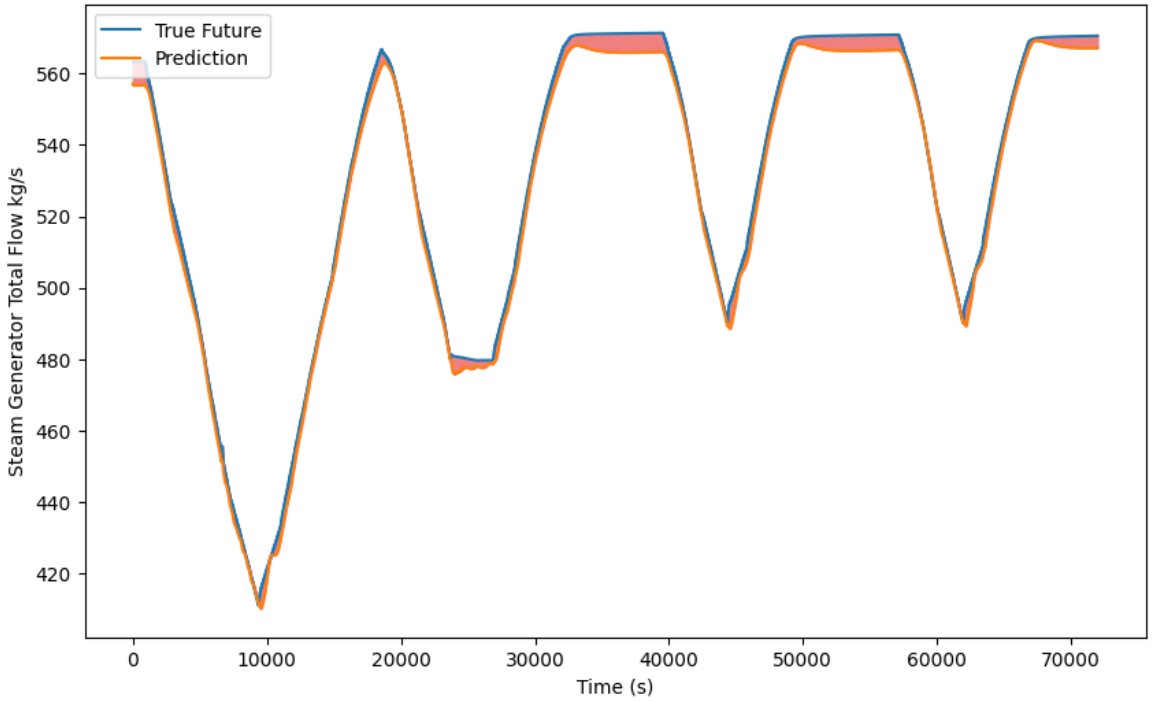
Georgia Tech

Cyber-attacks conducted

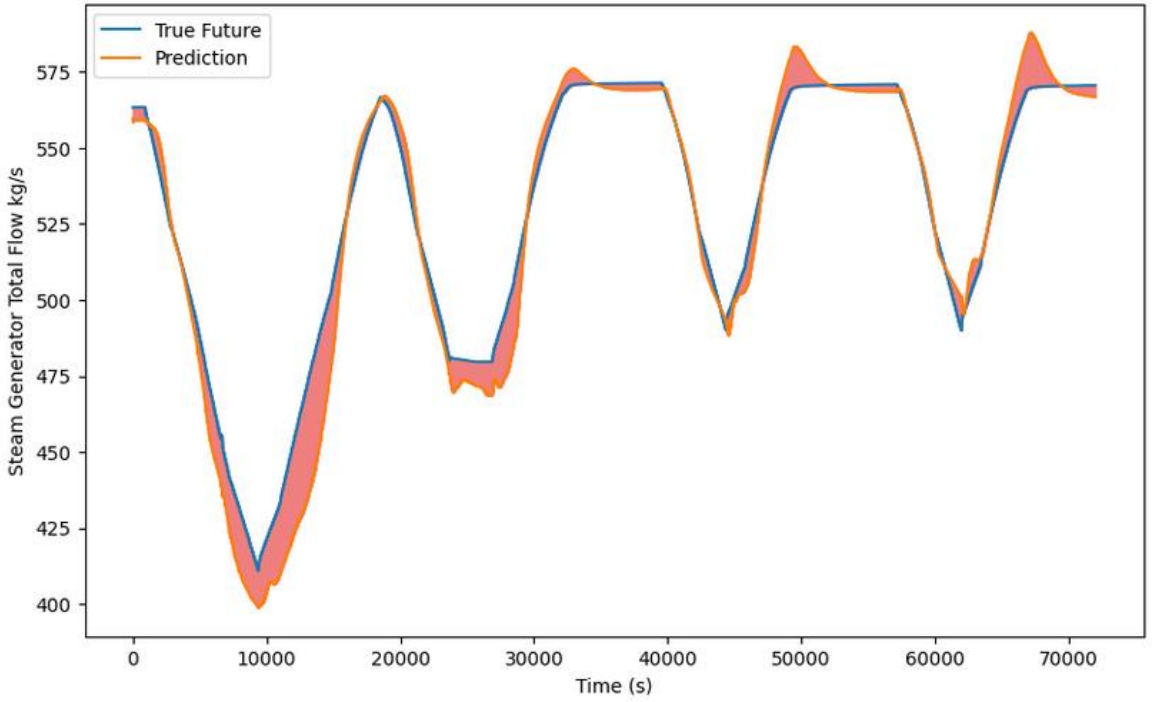


Training data attacks: effect on component-level DT training forecasting

Before: 0.0116 Mean Absolute Error



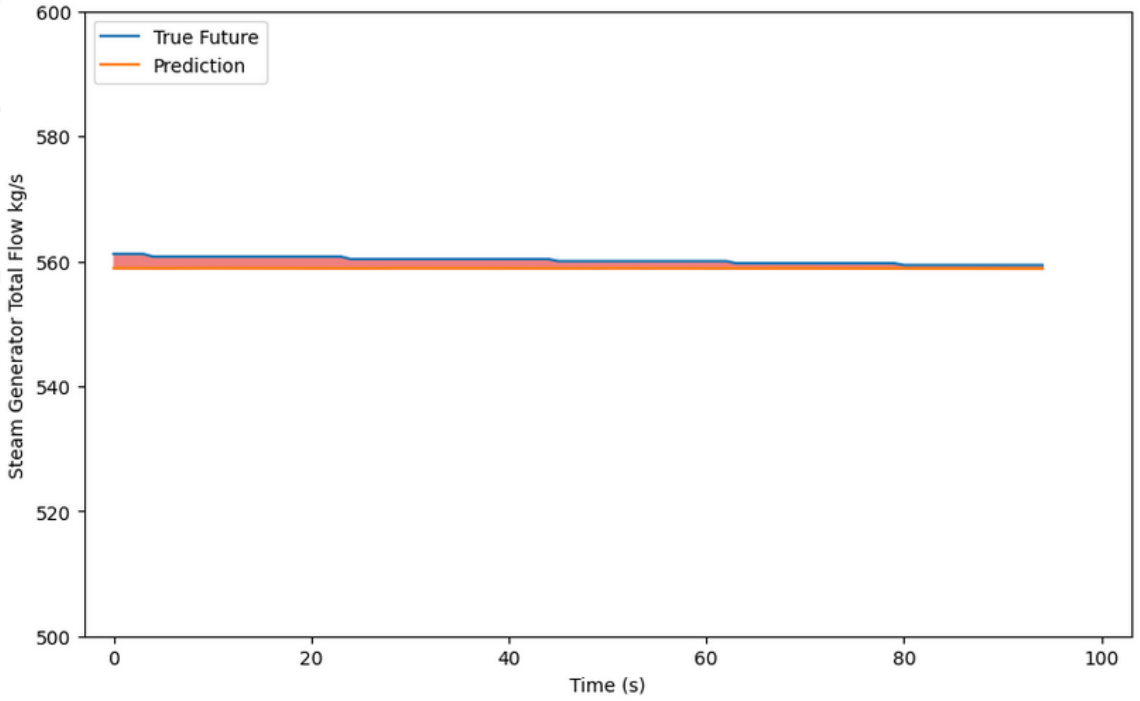
After: 0.0974 Mean Absolute Error



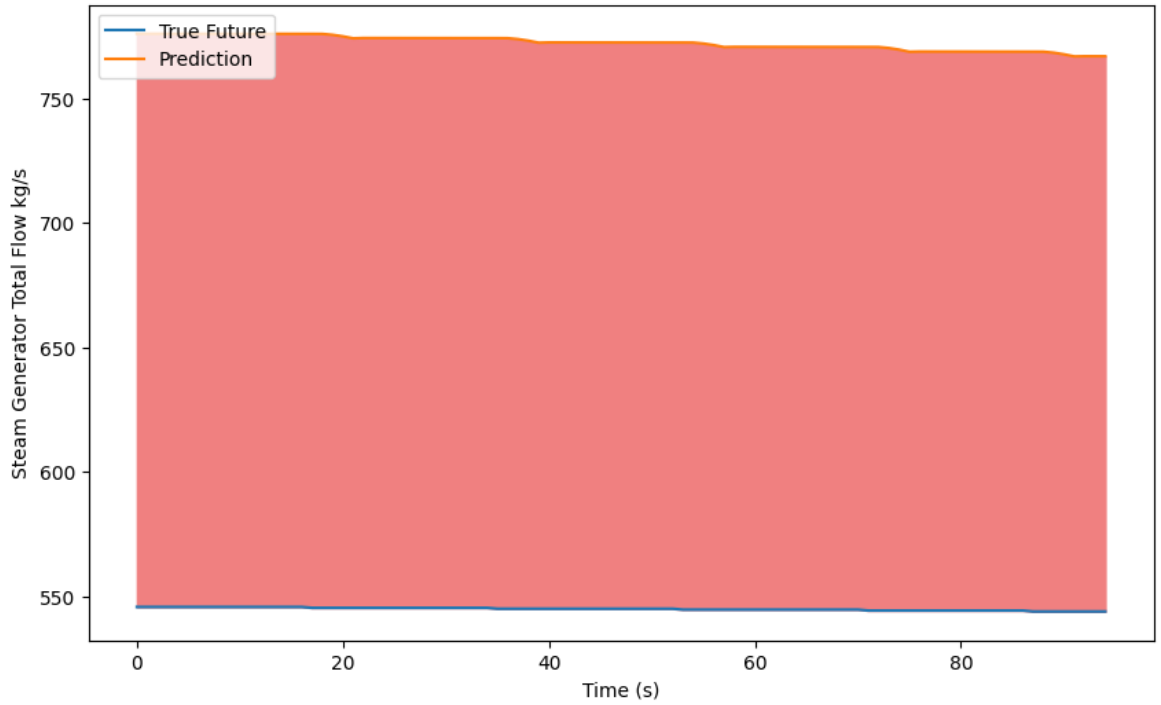
Georgia Tech

Training data attacks: effect on component-level DT real-time forecasting

Before: 0.0138 Mean Absolute Error



After: 7.170 Mean Absolute Error



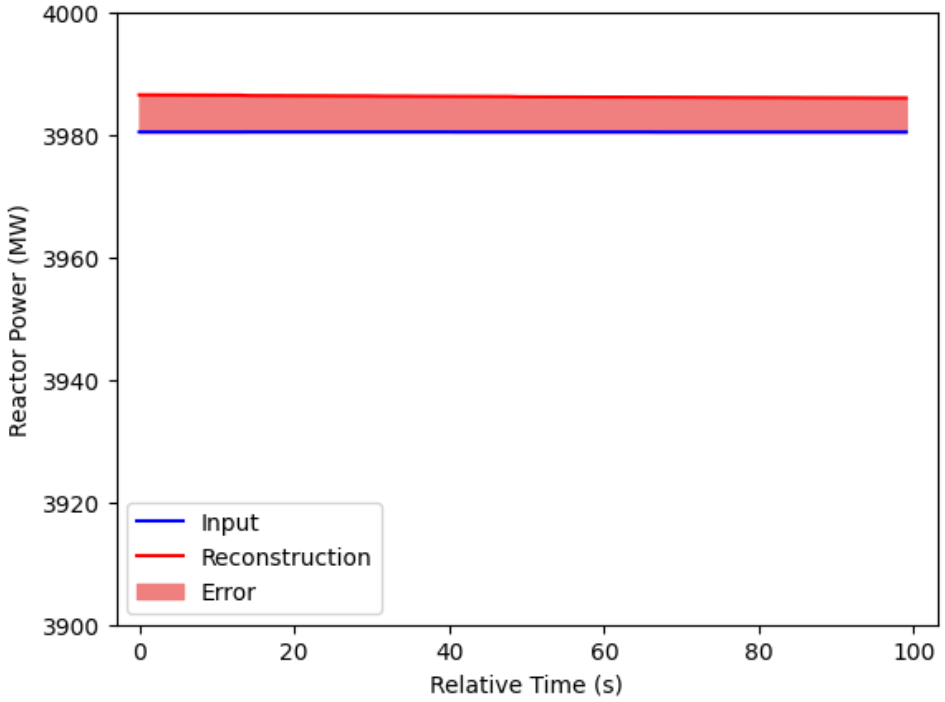
Georgia Tech

Training data attack impact and likelihood

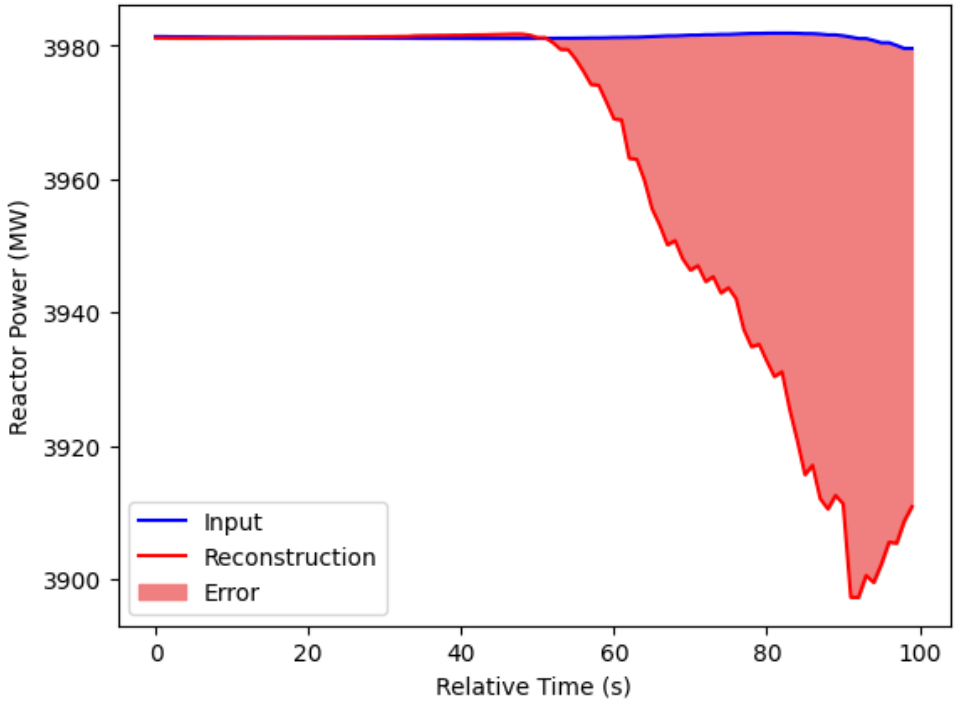
- Impact:
  - Underprediction of flow during training -> significant overprediction of flow during real-time testing.
  - If used for autonomous controls, the system fully closes the feedwater inlet valve to compensate for the perceived excessive flow rate.
- Likelihood:
  - (+) Simple to craft attacks.
  - (+) plenty of database exploits online.
  - (-) Requires the model to be re-trained on the contaminated data.
  - (-) Well-known and detected attack vector.

Real-time data attacks: effects on real-time plant-level DT

Before: 4.071E-05 Mean Squared Error



After: 0.08385 Mean Squared Error



Georgia Tech

Real-time data attack impact and likelihood

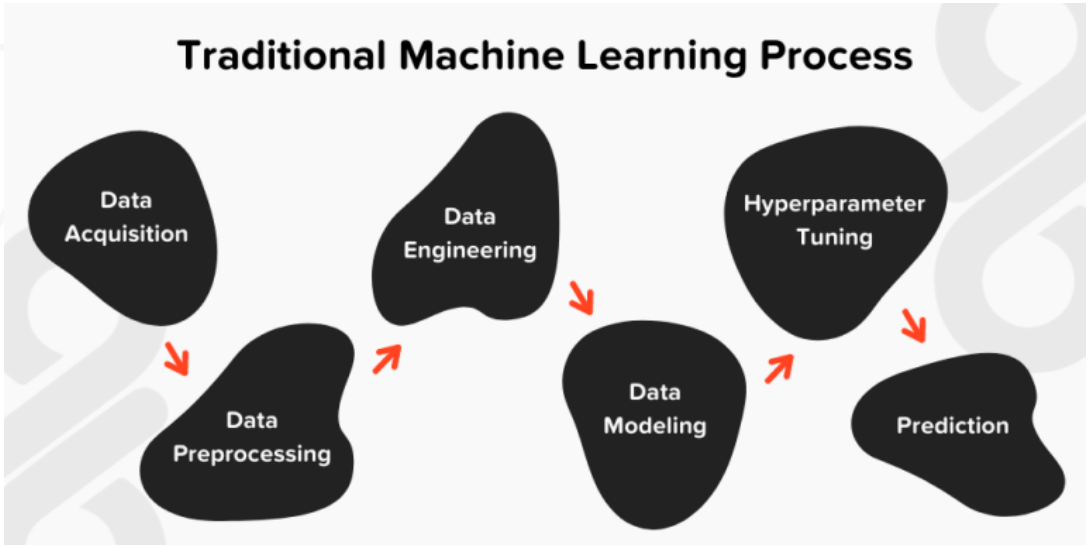
- Impact:
  - Increase in error in most tag values (50 out of 70 tags evaluated saw an increase of more than 20% in mean squared error).
  - Mimics behavior of a valve malfunction and can cause a system shutdown.
- Likelihood:
  - (+) Plenty of tools to craft custom attacks.
  - (-) Many PLCs/FPGAs use proprietary networking protocols.
  - (+) Easy to obfuscate nefarious activity.
  - (+) Not super well-known attack vectors.

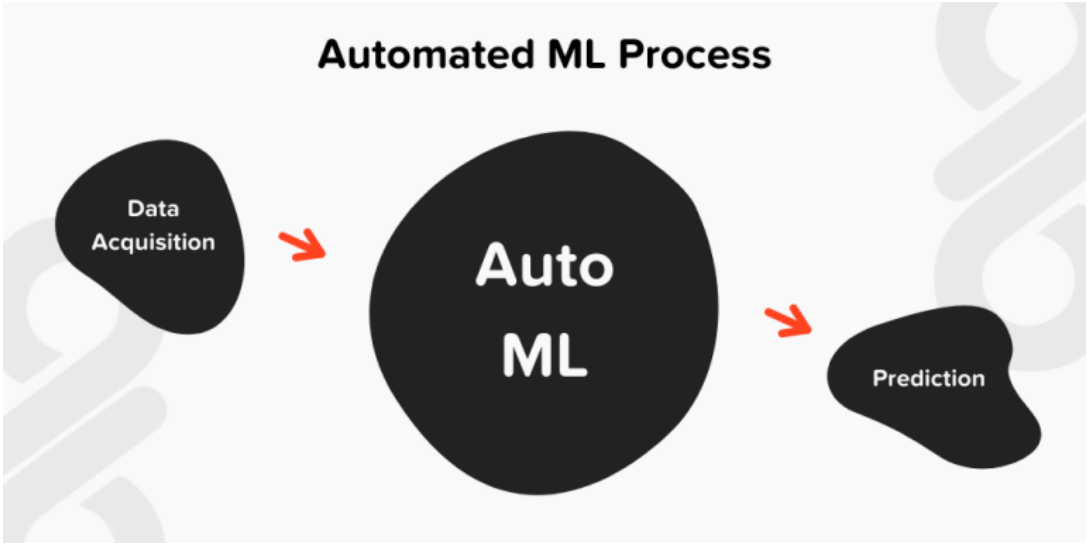
Adversarial hyperparameter/weight tuning findings

- Impact:
  - With the combination of plant-level and component-level DTs, the effects of attacks on one model can be mitigated using multiple models.
  - Data-based attacks are more fruitful as effects can propagate across models.
- Likelihood:
  - (+) Most ML models are saved locally unencrypted.
  - (-) Requires significant amounts of model architecture and implementation knowledge to craft an attack.
  - (-) Can result in unintended consequences.
  - (-) Hard to test and lacks attack explainability.

Introducing AutoML

- AutoML can increase accuracy…
- But at what cost?





https://smartboost.com/blog/how-does-automl-work/

Georgia Tech

Qualitative Cyber-Risk Assessment

Combined ML Cyber-Risk Matrix			
Impact / Likelihood	Low Impact	Medium Impact	High Impact
Low Likelihood	Attacks against Traditional ML Models	Attacks against AutoML Models	
Medium Likelihood	Training Data Attacks against Classifiers and AANNs	Training Data Attacks against AutoML Forecasters	Training Data Attacks against Traditional ML Forecasters
High Likelihood	Real-Time Data Attacks against Traditional ML Classifiers	Real-Time Data Attacks against Traditional ML Forecasters	Real-Time Data Attacks against AutoML Models and Traditional ML AANNs

- Likelihood is qualified by complexity and amount of insider knowledge to complete the attack.
- Impact is qualified by the model's accuracy metrics change following the attack.

Risk Mitigation Strategies

- Mitigation strategies to reduce “red zone” risk
- Generalized risk mitigation strategies:
  - Employ a Security Information and Event Management (SIEM) system to detect abnormal usage patterns and vulnerabilities.
  - Use multifactor authentication.
  - Maintain ML model save files and model information offline.
  - Regularly back up the autonomous control system (ACS).
  - Analyze the supply chain of ML models for vulnerabilities.
  - Encrypt model save files.
  - Ensure ACS computer files cannot be modified during training.
  - Collaborate with regulators to assess model drift and patch ML vulnerabilities.
  - Back up and compare model save files between training sessions.

Why this matters

- **We lack the infrastructure (regulation, personnel, etc.) to adequately deal with new, credible threats posed by ML.**
- What can you do?
  - Help develop regulations/guidelines for dealing with ML threats.
  - Conduct a cyber-threat assessment of ML systems before implementation.
  - Educate yourself and the broader community on cyber-risks associated with ML and potential adversarial actions.

https://news.maryland.gov/dnr/2019/12/09/a-voice-in-the-wilderness-after-75-years-smokeys-message-still-looms-large/

Georgia Tech

Conclusion

- ML algorithms are helpful to various industries, including power generation.
- ML-based DTs can be used to develop autonomous control systems or model system behavior.
- ML use poses a credible cyber threat to system behavior.
- ML cyber threats can be controlled through well-known mitigation techniques.
- **As a community, we must start thinking about the cybersecurity considerations of ML algorithms before implementation.**

http://oxfordpreparatoryacademy.blogspot.com/2010_07_01_archive.html

Georgia Tech

References

[1] C. M. Spirito et al., "Autonomous System Subversion Tactics: Prototypes and Recommended Countermeasures," Idaho National Laboratory (INL), Idaho Falls, ID (United States), Tech. Rep. INL/RPT-22-68871-Rev001, Aug. 2022.

[2] Maximize Market Research LTD, "AI in Energy Market: AI is accelerating innovation in the energy sector as it continues to change the way of organizations and industries operate," MAXIMIZE MARKET RESEARCH, Dec. 2022.

[3] M. Lehto, "Cyber-Attacks Against Critical Infrastructure," in Cyber Security, M. Lehto and P. Neittaanm¨aki, Eds., vol. 56, Series Title: Computational Methods in Applied Sciences, Cham: Springer International Publishing, 2022, pp. 3–42, ISBN: 978-3-030-91292-5 978-3-030-91293-2.

[4] C. M. Spirito et al., "Cyber Threat Assessment Methodology for Autonomous and Remote Operations for Advanced Reactors (Amended November 2021)," Idaho National Lab. (INL), Idaho Falls, ID (United States), Tech. Rep. INL/EXT-21-62175-Rev-002, Nov. 2021.

[5] P. Yockey, A. Erickson, and C. M. Spirito, "Cyber Threat Assessment of Machine Learning Driven Autonomous Control Systems of Nuclear Power Plants," Progress in Nuclear Energy, 2023.

[6] X. Chen and F. Zhang, "Development of a Hardware-in-the-Loop Testbed Using a Full-Scope Nuclear Power Plant Simulator for Instrumentation and Control and Cybersecurity Education, Training, and Research," in Innovative Use of Technology in Training: IV, Amelia Island, FL, Feb. 2023.

[7] Western Services Corporation, 3KeyMaster Generic Pressurized Water Reactor Product Sheet, May 2016.

[8] Allen-Bradley, FactoryTalk Linx Gateway Software — FactoryTalk, 2023.

[9] ODVA Technologies, Common Industrial Protocol (CIP™), 2017.

[10] Allen-Bradley, Studio 5000 Logix Designer — FactoryTalk, 2022.

[11] A. Skolnick, P. Yockey, X. Chen, J. Coble, and F. Zhang, "Development of a Full Scope NPP Cybersecurity Hardware-in-the-Loop Testbed," in Cybersecurity for Nuclear Installations, Phoenix, AZ, Nov. 2022.

[12] P. Yockey, K. Kelly, C. M. Spirito, and F. Zhang, "A Cyber Threat Methodology for Autonomous Control Systems for Advanced Reactors," in Cybersecurity for Nuclear Installations, Anaheim, CA, Jun. 2022.

[13] Rapid7, Metasploit, original-date: 2011-08-30T06:13:20Z, Jul. 2023.

[14] Cybersecurity and Infrastructure Security Agency, Advanced Persistent Threat Activity Targeting Energy and Other Critical Infrastructure Sectors — CISA, Mar. 2018.

[15] R. S. S. K. Johnson Ann, Cyberattacks against machine learning systems are more common than you think, Oct. 2020.

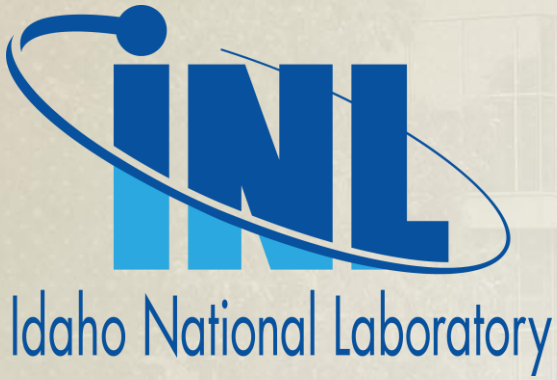
[16] MITRE and Microsoft, Annoucing ATLAS! original-date: 2020-10-15T12:47:28Z, Aug. 2023.

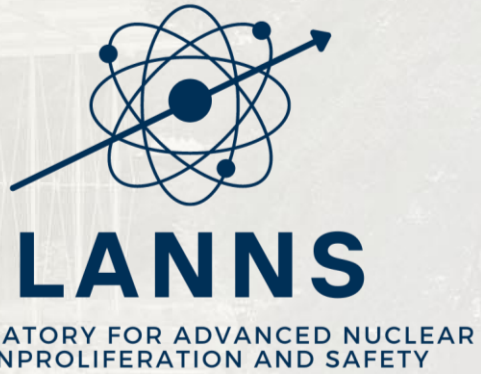
[17] MITRE, Mitigations List — MITRE ATLAS™.

[18] G. Desarnaud, Cyber Attacks and Energy Infrastructures (´Etudes de l'Ifri). Institut franc¸ais des relations internationales, Jan. 2017, ISBN: 978-2-36567-724-0.

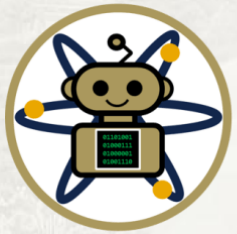
Georgia Tech

Acknowledgements





Idaho State University



iFAN Lab
Intelligence for Advanced Nuclear

Georgia Tech