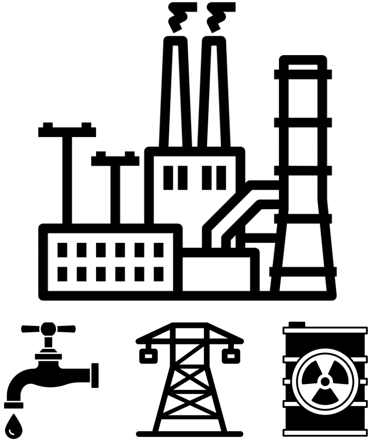
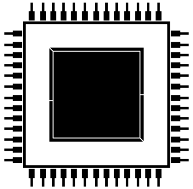
Attributions for ML-based ICS anomaly detection: From theory to practice

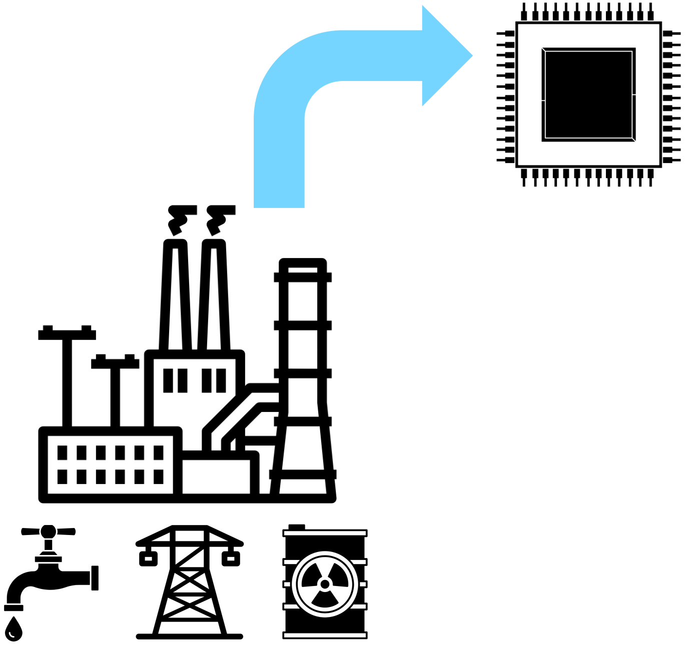
Clement Fung, Eric Zeng, Lujo Bauer
Carnegie Mellon University

What are industrial control systems?



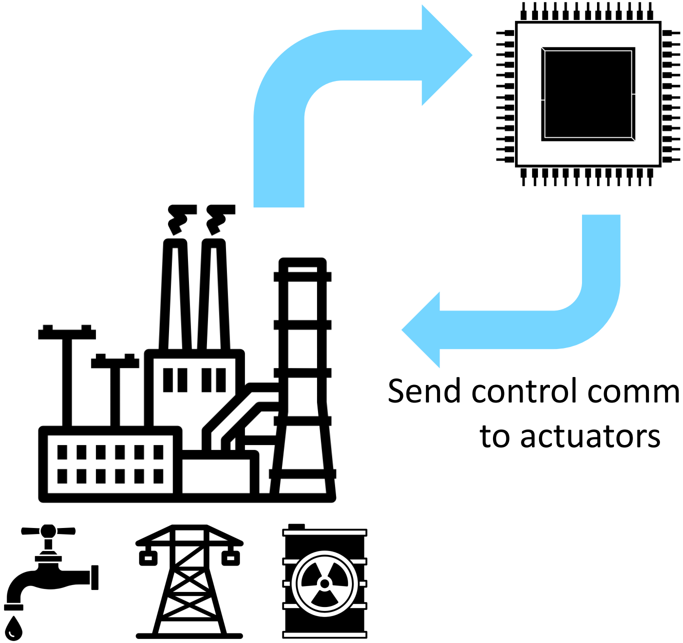


What are industrial control systems?



Read sensor values from physical process

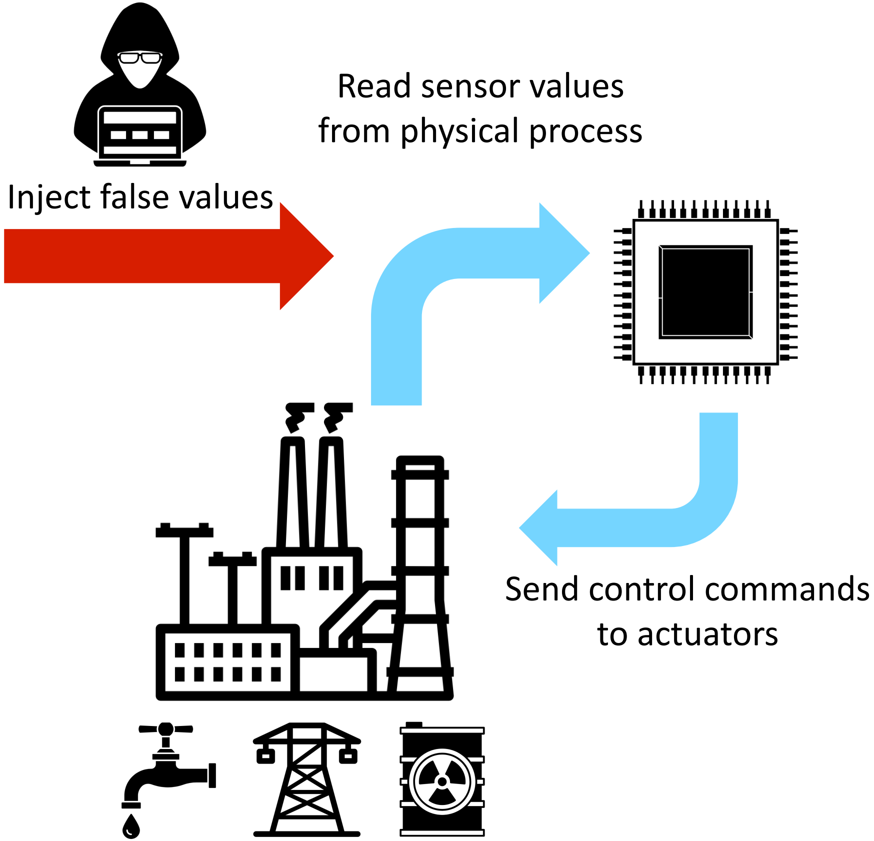
What are industrial control systems?



Read sensor values
from physical process

Send control commands
to actuators

Attacking industrial control systems

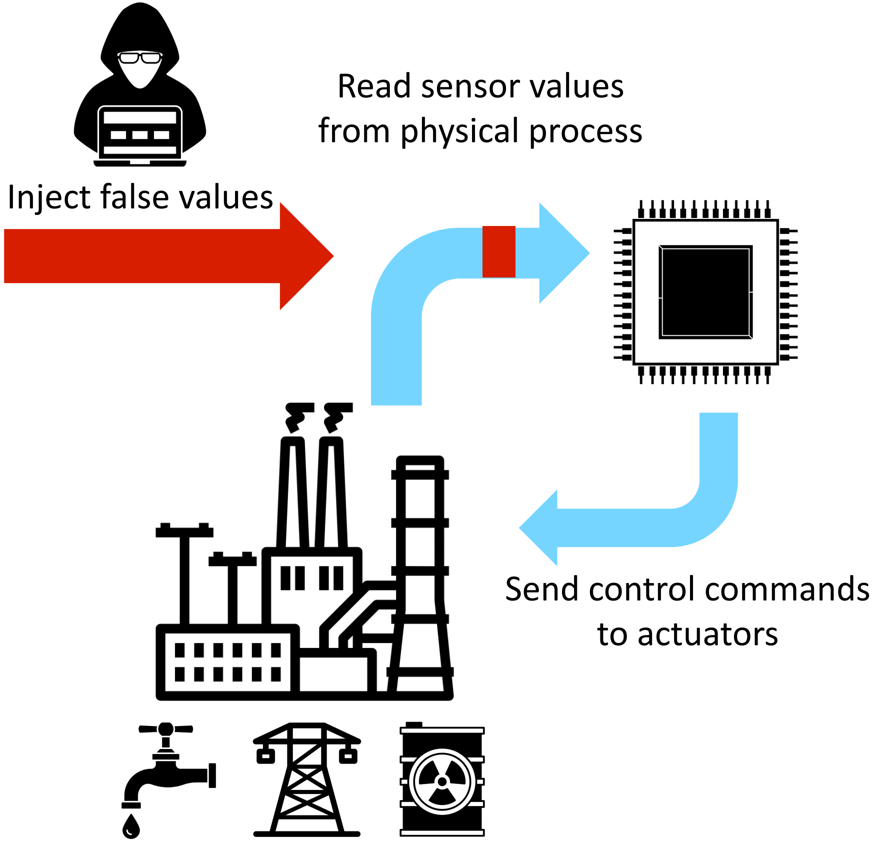


Read sensor values from physical process

Inject false values

Send control commands to actuators

Attacking industrial control systems

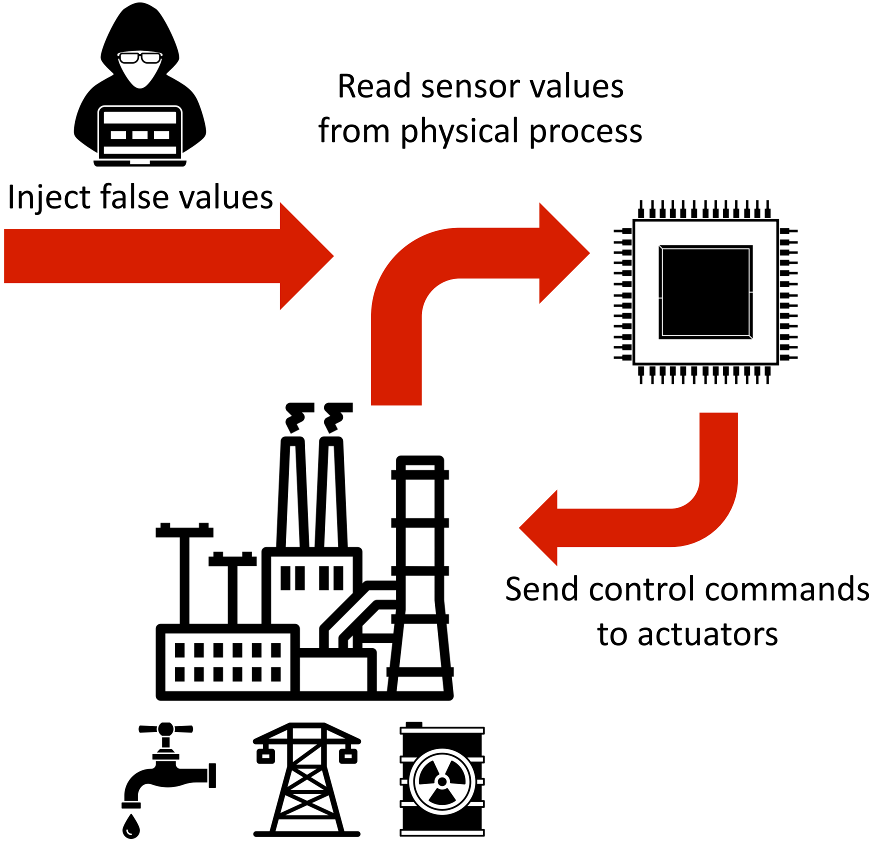


Read sensor values from physical process

Inject false values

Send control commands to actuators

Attacking industrial control systems

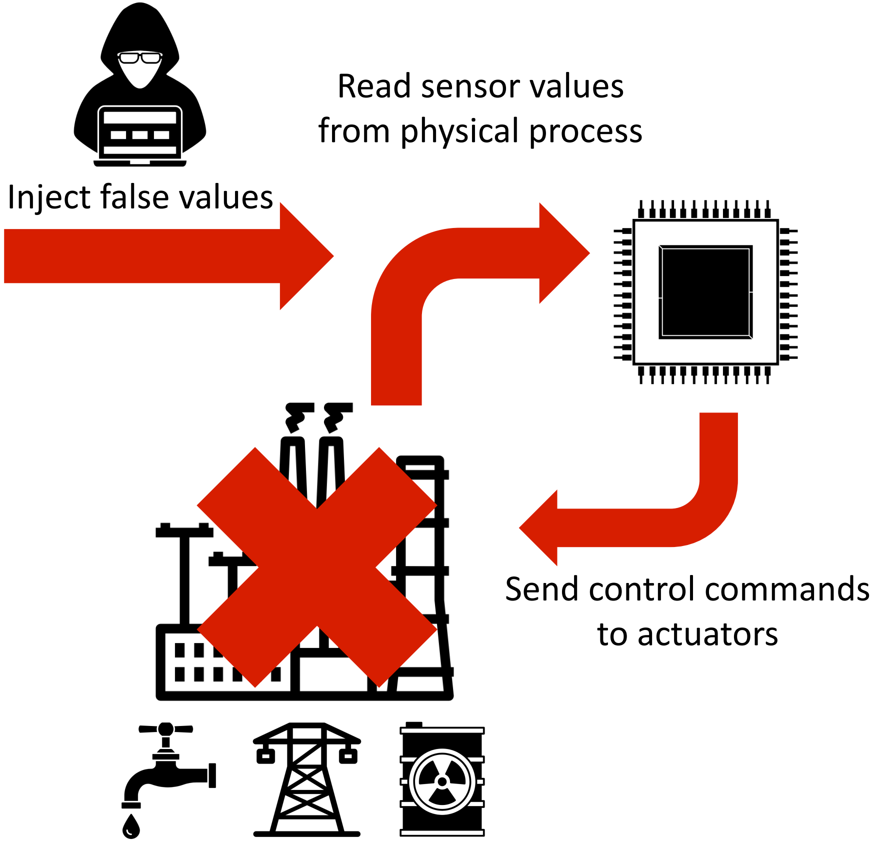


Read sensor values from physical process

Inject false values

Send control commands to actuators

Attacking industrial control systems

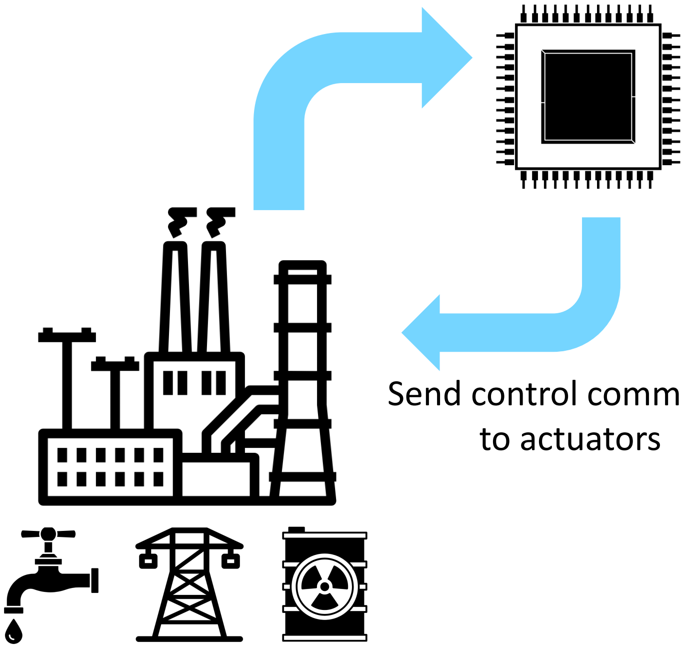


Read sensor values
from physical process

Inject false values

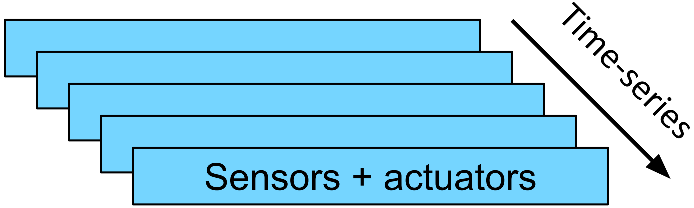
Send control commands
to actuators

Defending industrial control systems

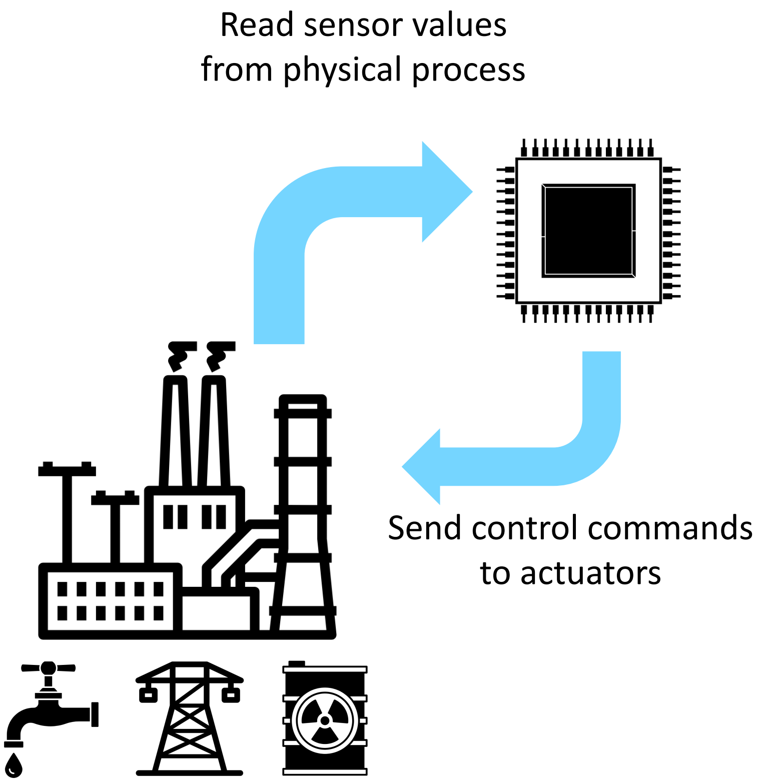


Read sensor values from physical process

Send control commands to actuators

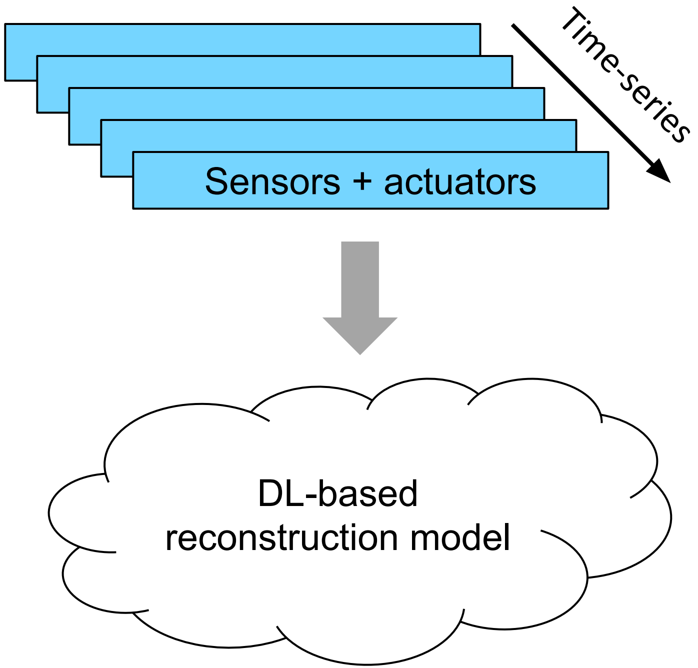


Defending industrial control systems



Read sensor values from physical process

Send control commands to actuators

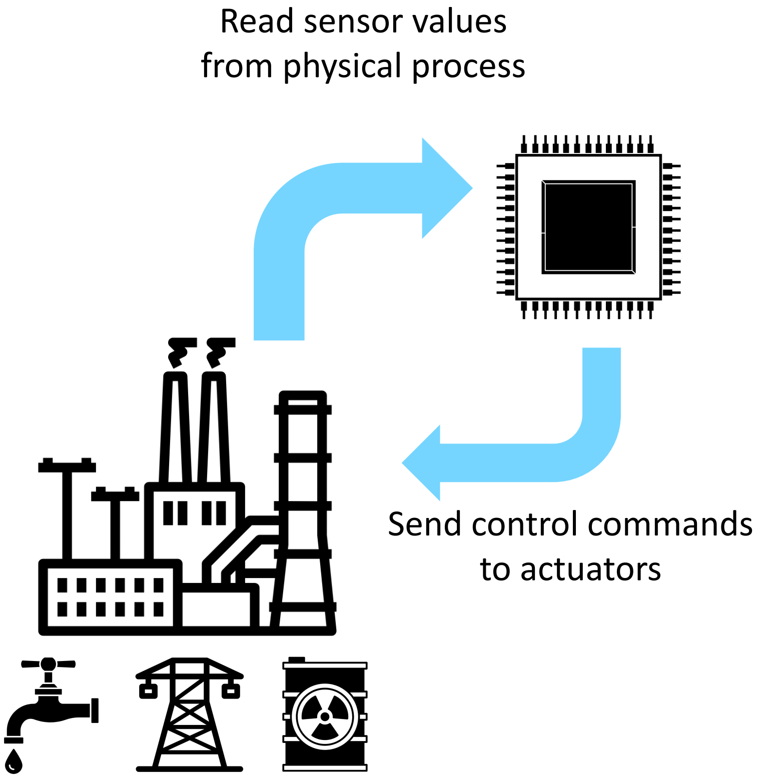


Sensors + actuators

Time-series

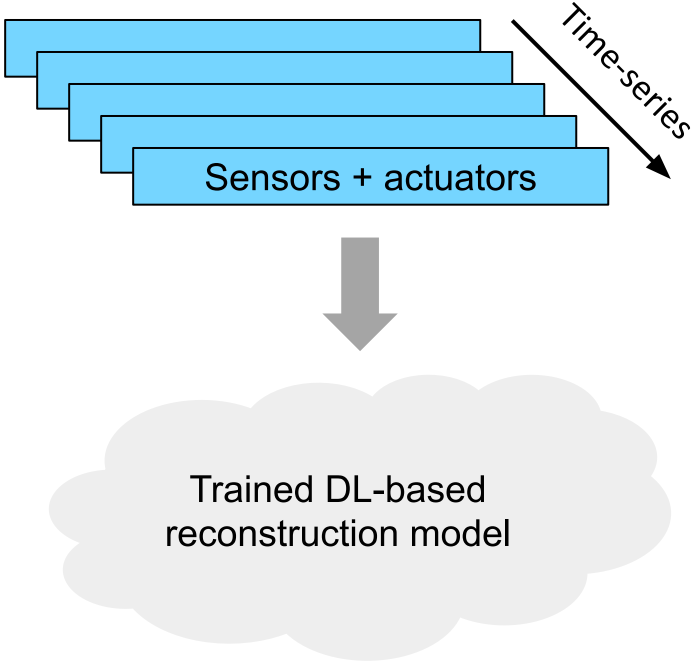
DL-based reconstruction model

Defending industrial control systems



Read sensor values from physical process

Send control commands to actuators

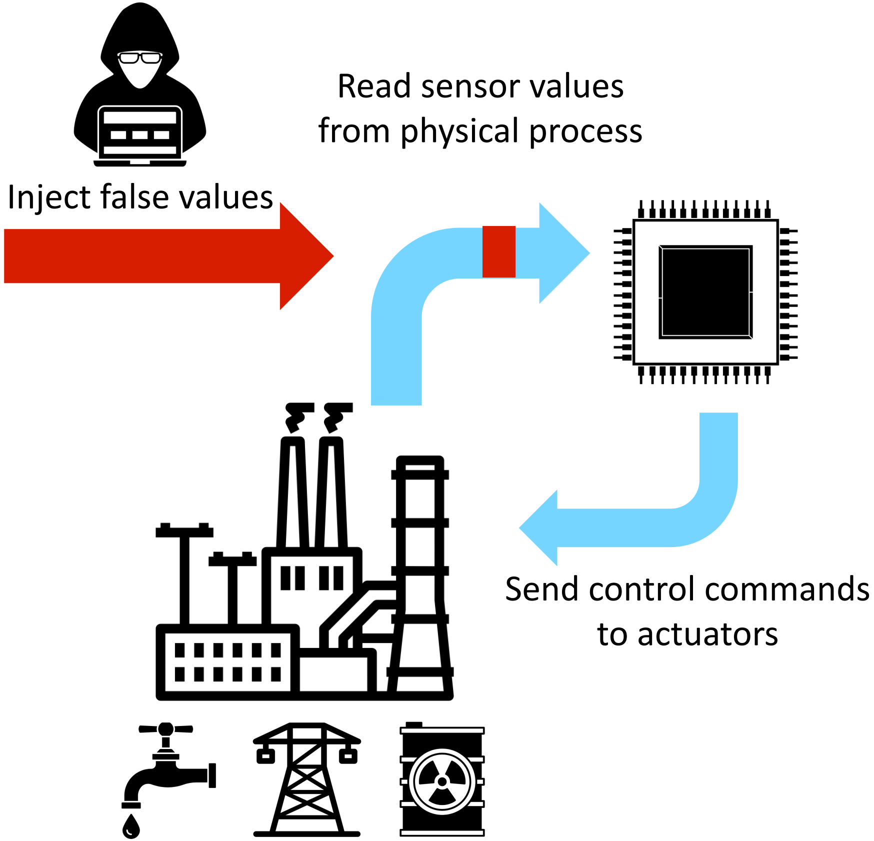


Time-series

Sensors + actuators

Trained DL-based reconstruction model

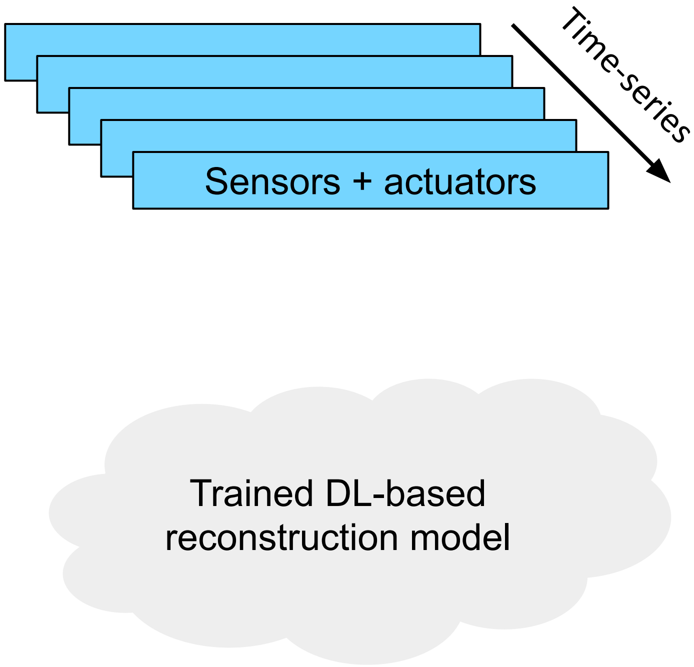
Defending industrial control systems



Inject false values

Read sensor values from physical process

Send control commands to actuators

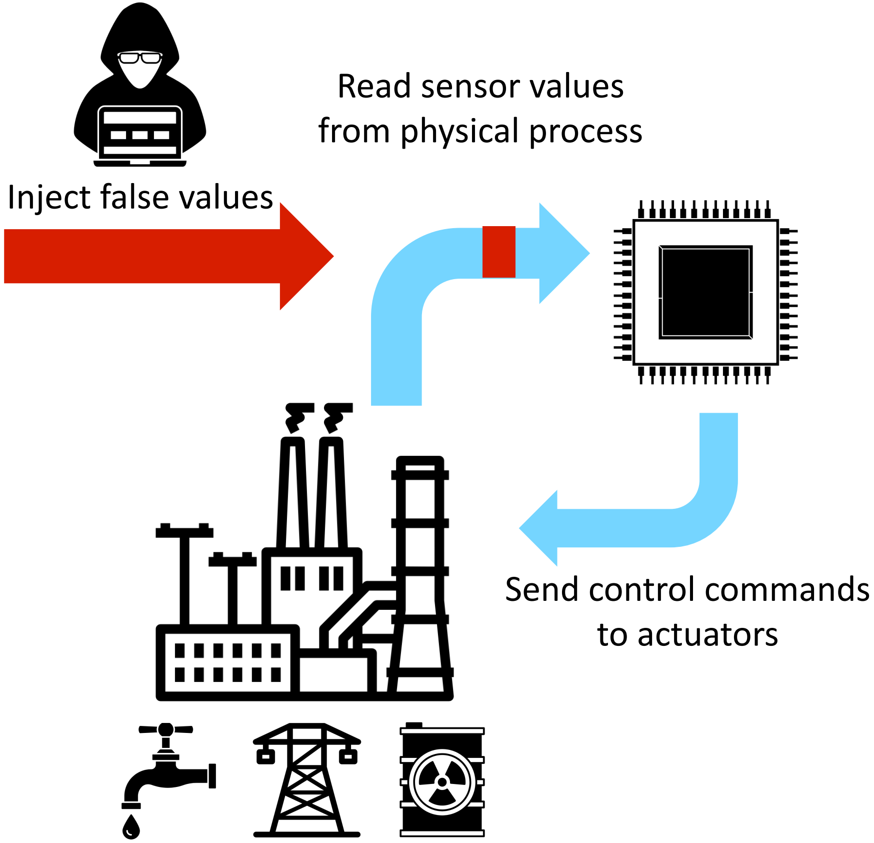


Time-series

Sensors + actuators

Trained DL-based reconstruction model

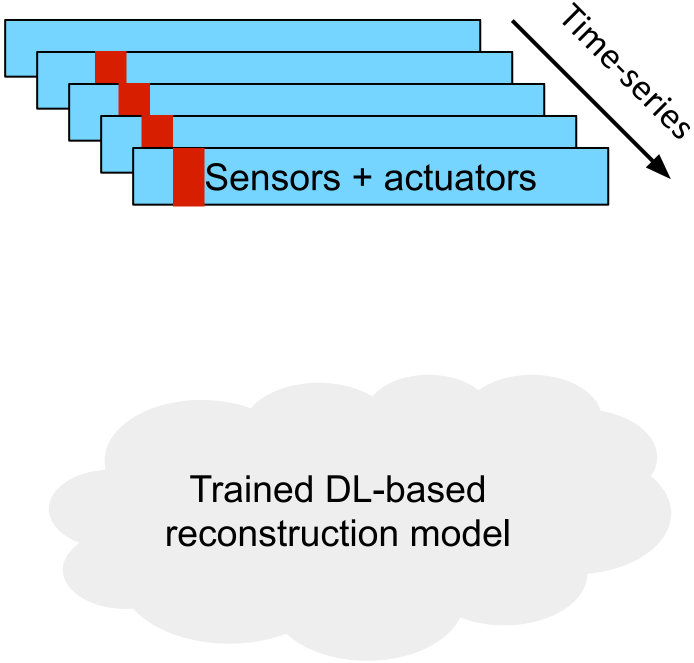
Defending industrial control systems



Inject false values

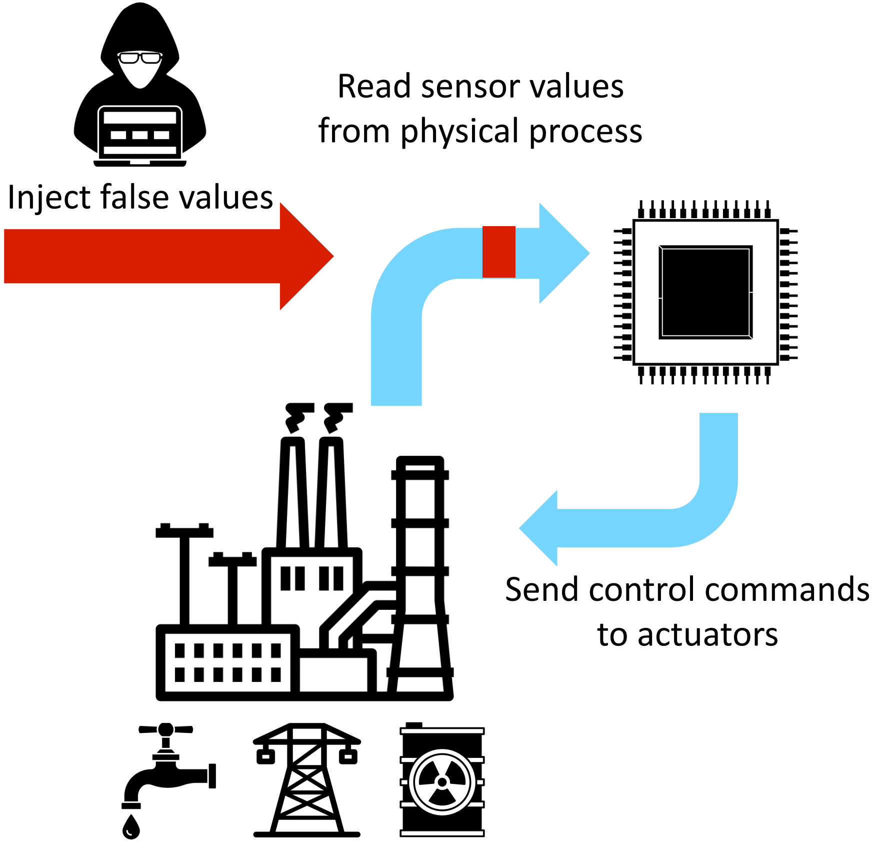
Read sensor values from physical process

Send control commands to actuators



Trained DL-based reconstruction model

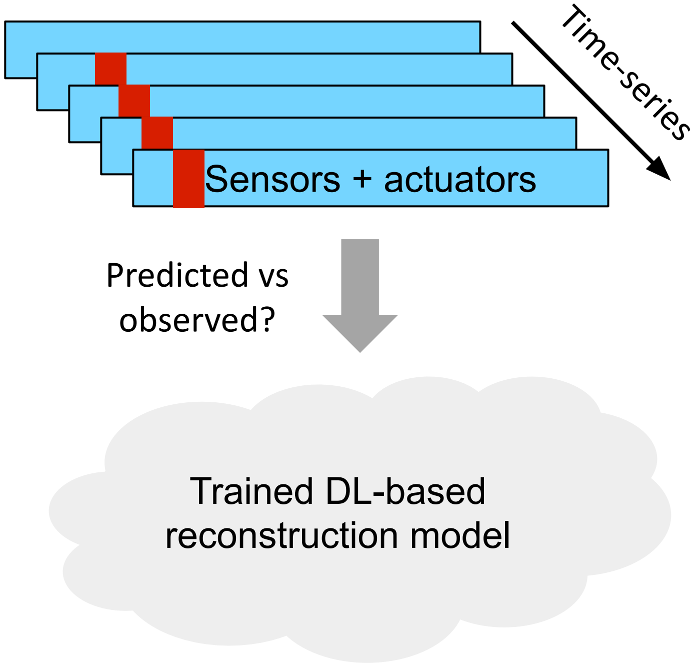
Defending industrial control systems



Read sensor values from physical process

Inject false values

Send control commands to actuators



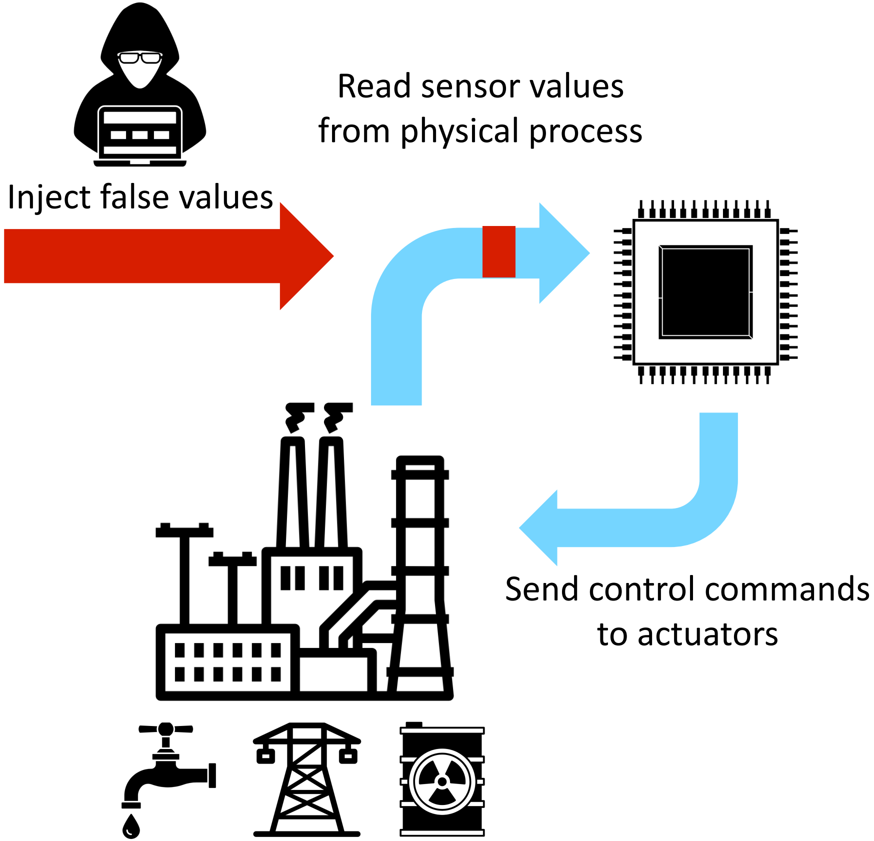
Sensors + actuators

Time-series

Predicted vs observed?

Trained DL-based reconstruction model

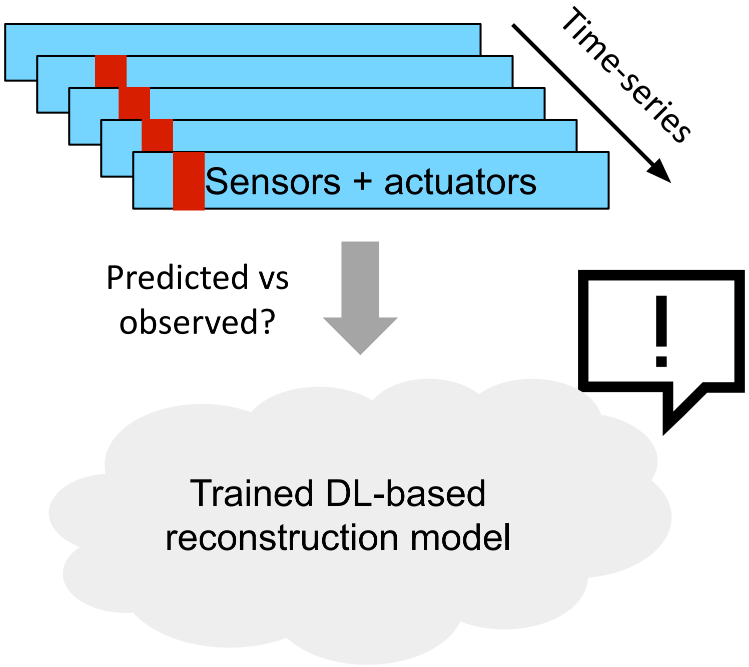
Defending industrial control systems



Read sensor values from physical process

Inject false values

Send control commands to actuators



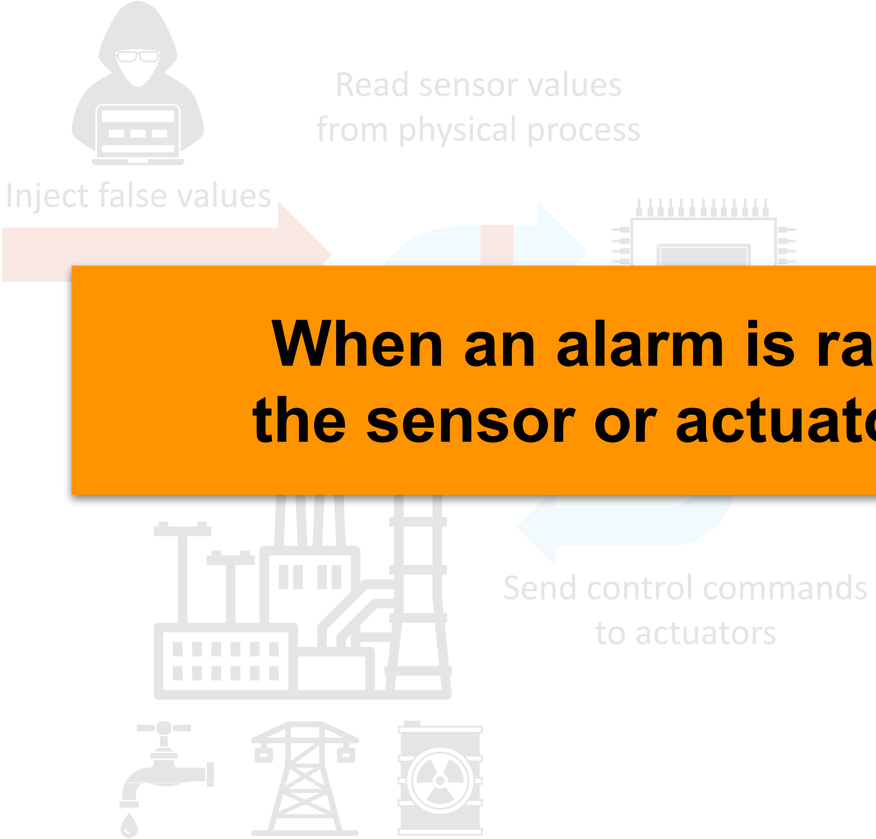
Sensors + actuators

Time-series

Predicted vs observed?

Trained DL-based reconstruction model

Defending industrial control systems



Time-series
Sensors + actuators
Trained DL-based
reconstruction model

When an alarm is raised, can we identify the sensor or actuator that was attacked?

Explainable AI (XAI) through attribution methods

- Attribution methods: explain what **input features** cause a model to **produce a specific output**

Sturmfels et al. "Visualizing the Impact of Feature Attribution Baselines". Distill 2020.

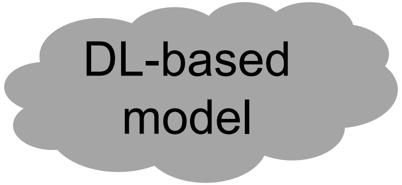
Explainable AI (XAI) through attribution methods

- Attribution methods: explain what **input features** cause a model to **produce a specific output**
    - When a model predicts a label, **why?**

Sturmfels et al. "Visualizing the Impact of Feature Attribution Baselines". Distill 2020.

Explainable AI (XAI) through attribution methods

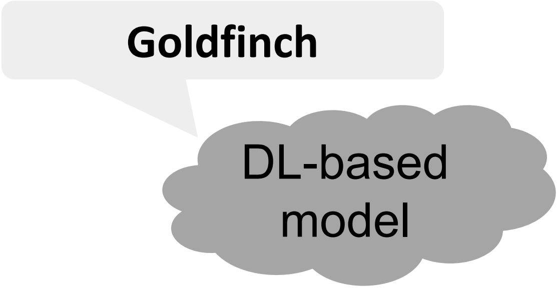
- Attribution methods: explain what **input features** cause a model to **produce a specific output**
  - When a model predicts a label, **why?**



Sturmfels et al. "Visualizing the Impact of Feature Attribution Baselines". Distill 2020.

Explainable AI (XAI) through attribution methods

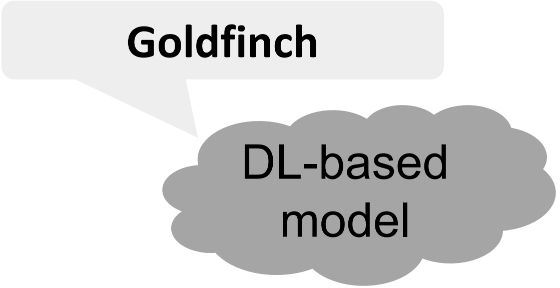
- Attribution methods: explain what **input features** cause a model to **produce a specific output**
  - When a model predicts a label, **why?**



Sturmfels et al. "Visualizing the Impact of Feature Attribution Baselines". Distill 2020.

Explainable AI (XAI) through attribution methods

- Attribution methods: explain what **input features** cause a model to **produce a specific output**
  - When a model predicts a label, **why?**



Goldfinch

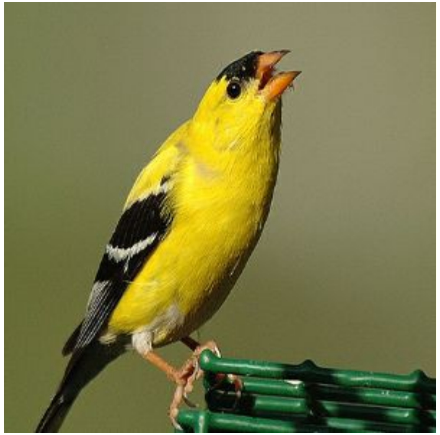
DL-based model

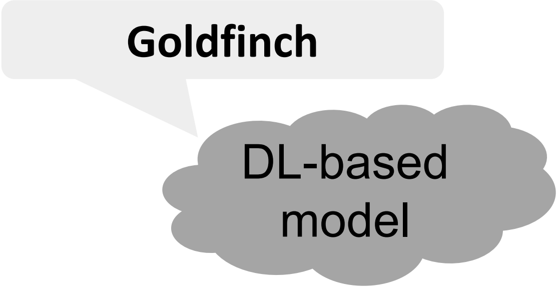
Which **input pixels** caused the prediction?

Sturmfels et al. "Visualizing the Impact of Feature Attribution Baselines". Distill 2020.

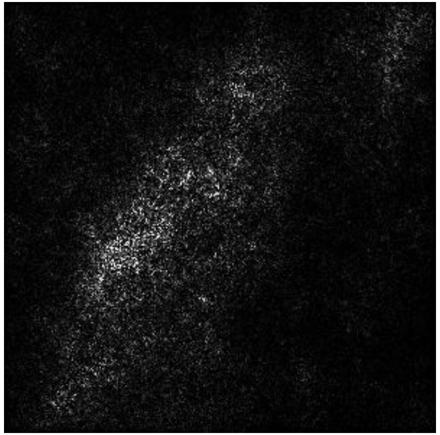
Explainable AI (XAI) through attribution methods

- Attribution methods: explain what **input features** cause a model to **produce a specific output**
  - When a model predicts a label, **why?**





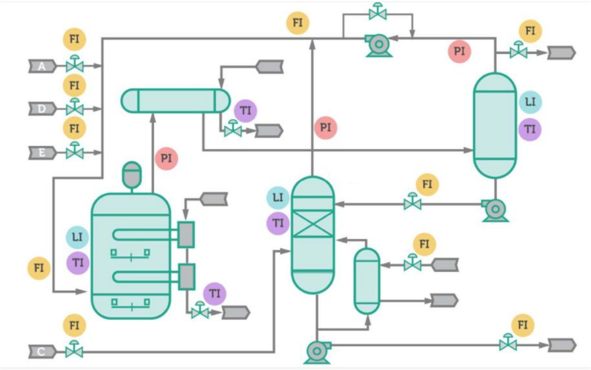
Which **input pixels** caused the prediction?



Sturmfels et al. "Visualizing the Impact of Feature Attribution Baselines". Distill 2020.

Explainable AI (XAI) through attribution methods

- Attribution methods: explain what **input features** cause a model to **produce a specific output**
    - When a model predicts a label, **why?**

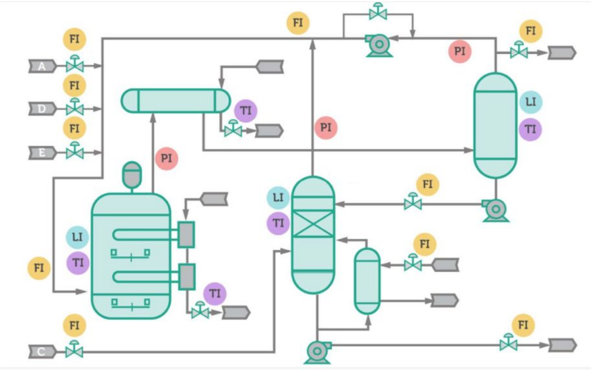


DL-based model

Sturmfels et al. "Visualizing the Impact of Feature Attribution Baselines". Distill 2020.

Explainable AI (XAI) through attribution methods

- Attribution methods: explain what **input features** cause a model to **produce a specific output**
  - When a model predicts a label, **why?**



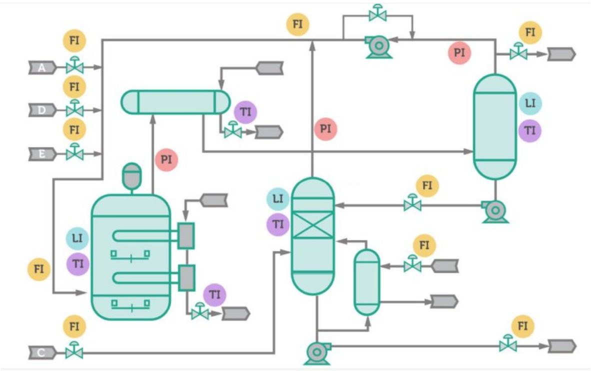
Anomaly!

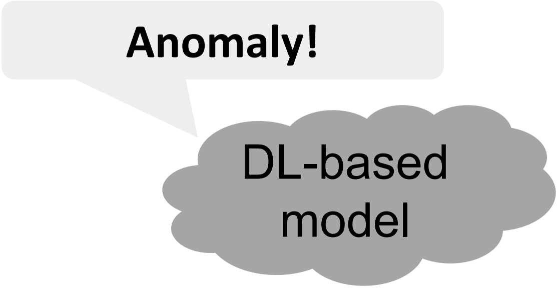
DL-based model

Sturmfels et al. "Visualizing the Impact of Feature Attribution Baselines". Distill 2020.

Explainable AI (XAI) through attribution methods

- Attribution methods: explain what **input features** cause a model to **produce a specific output**
  - When a model predicts a label, **why?**



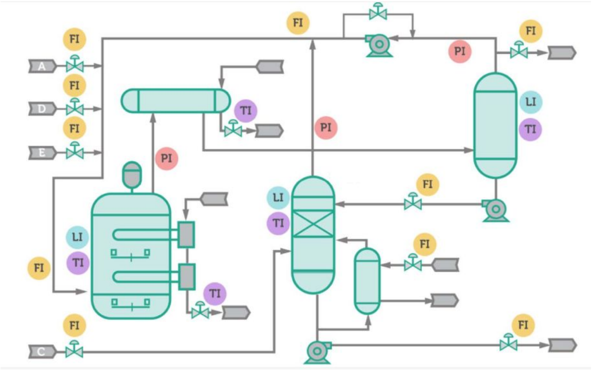


Which **sensor/actuator** was attacked?

Sturmfels et al. "Visualizing the Impact of Feature Attribution Baselines". Distill 2020.

Explainable AI (XAI) through attribution methods

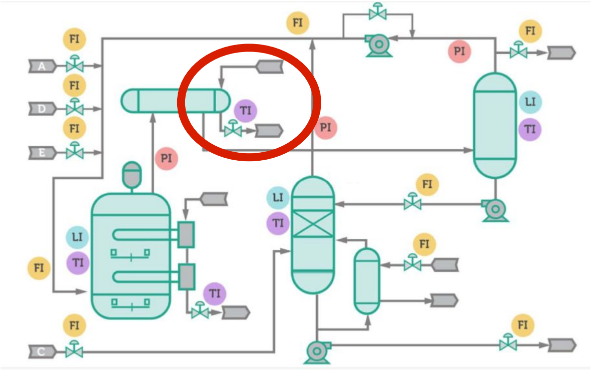
- Attribution methods: explain what **input features** cause a model to **produce a specific output**
  - When a model predicts a label, **why?**



Anomaly!

DL-based model

Which **sensor/actuator** was attacked?



Sturmfels et al. "Visualizing the Impact of Feature Attribution Baselines". Distill 2020.

In this work: exploring ICS anomaly attribution

1. (How well) do prior, **off-the-shelf attribution strategies** work for ICS anomaly attribution?

In this work: exploring ICS anomaly attribution

1. (How well) do prior, **off-the-shelf attribution strategies** work for ICS anomaly attribution?
2. How do **properties of ICS attacks** affect attribution accuracy?

In this work: exploring ICS anomaly attribution

1. (How well) do prior, **off-the-shelf attribution strategies** work for ICS anomaly attribution?
2. How do **properties of ICS attacks** affect attribution accuracy?
3. **Can we do better** than prior attribution strategies?

RQ1: Do prior attribution strategies work well?

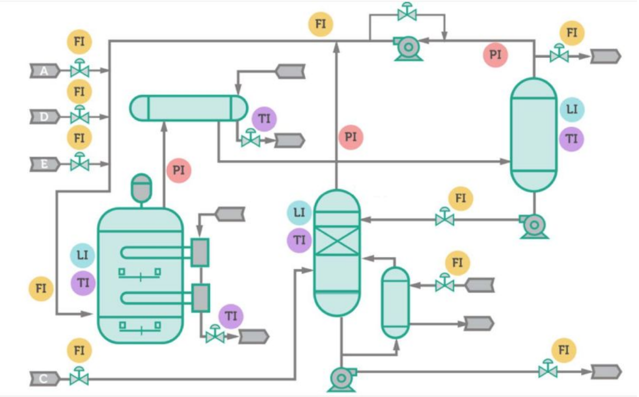
We evaluate attribution in diverse settings

- Compare a variety of anomaly-detection models [1]:
  - Linear models, CNNs, RNNs, LSTMs

[1] Fung et al. "Perspectives from a comprehensive evaluation of reconstruction-based anomaly detection in ICS." ESORICS 2022.
[2] Goh et al. "A dataset to support research in the design of secure water treatment systems." CRITIS 2016.
[3] A. Bathelt, N. L. Ricker, and M. Jelali, “Revision of the Tennessee Eastman process model,” IFAC ADCHEM, vol. 48, no. 8. 2015.

We evaluate attribution in diverse settings

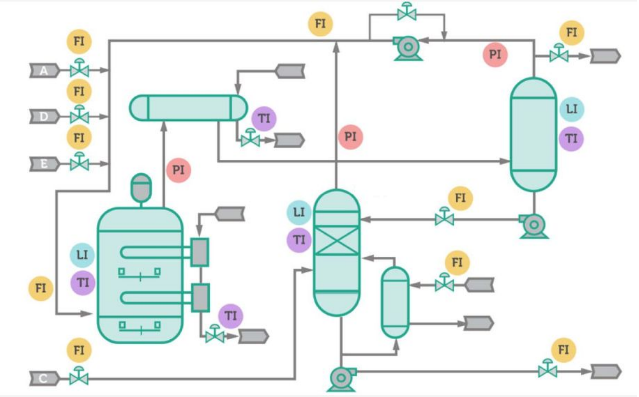
- Compare a variety of anomaly-detection models [1]:
  - Linear models, CNNs, RNNs, LSTMs
- Datasets [2,3]: SWaT, WADI, TEP
  - Water treatment (public datasets)
  - Chemical process (modified simulator)



[1] Fung et al. "Perspectives from a comprehensive evaluation of reconstruction-based anomaly detection in ICS." ESORICS 2022.
[2] Goh et al. "A dataset to support research in the design of secure water treatment systems." CRITIS 2016.
[3] A. Bathelt, N. L. Ricker, and M. Jelali, "Revision of the Tennessee Eastman process model," IFAC ADCHEM, vol. 48, no. 8. 2015.

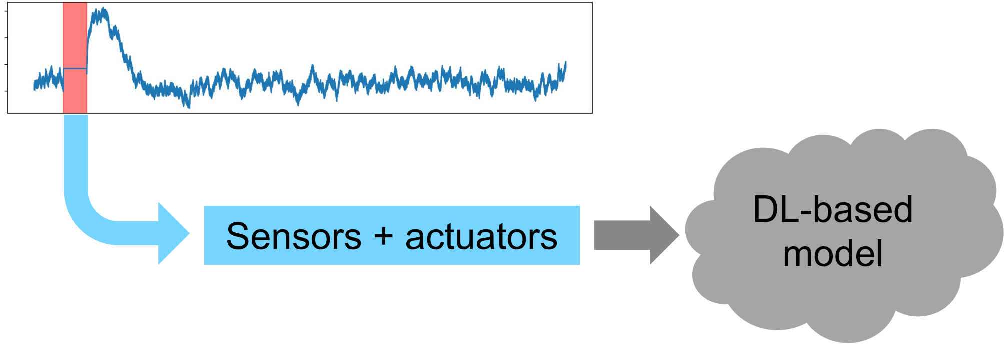
We evaluate attribution in diverse settings

- Compare a variety of anomaly-detection models [1]:
  - Linear models, CNNs, RNNs, LSTMs
- Datasets [2,3]: SWaT, WADI, TEP
  - Water treatment (public datasets)
  - Chemical process (modified simulator)
- Attack scenarios:
  - 47 real attacks on water treatment
  - 100 synthetic attacks on chemical process
    - Made publicly available!

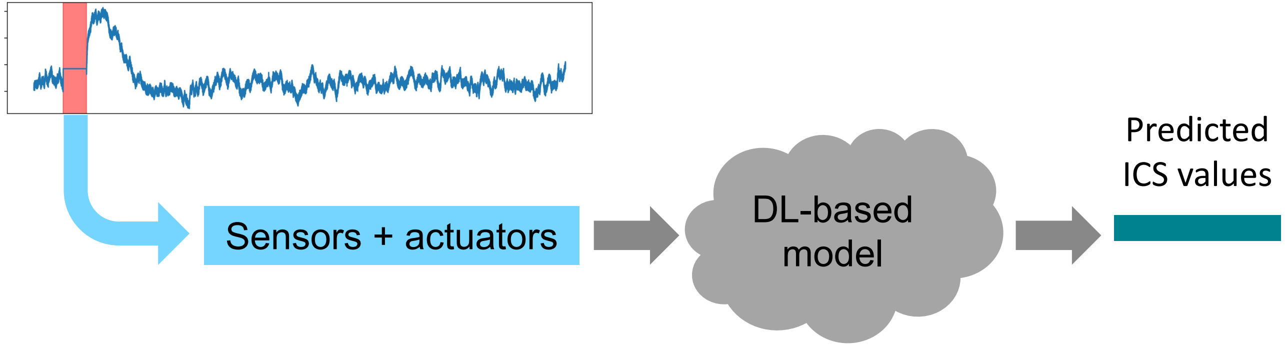


[1] Fung et al. "Perspectives from a comprehensive evaluation of reconstruction-based anomaly detection in ICS." ESORICS 2022.
[2] Goh et al. "A dataset to support research in the design of secure water treatment systems." CRITIS 2016.
[3] A. Bathelt, N. L. Ricker, and M. Jelali, “Revision of the Tennessee Eastman process model,” IFAC ADCHEM, vol. 48, no. 8. 2015.

ICS anomaly attribution: previously



ICS anomaly attribution: previously



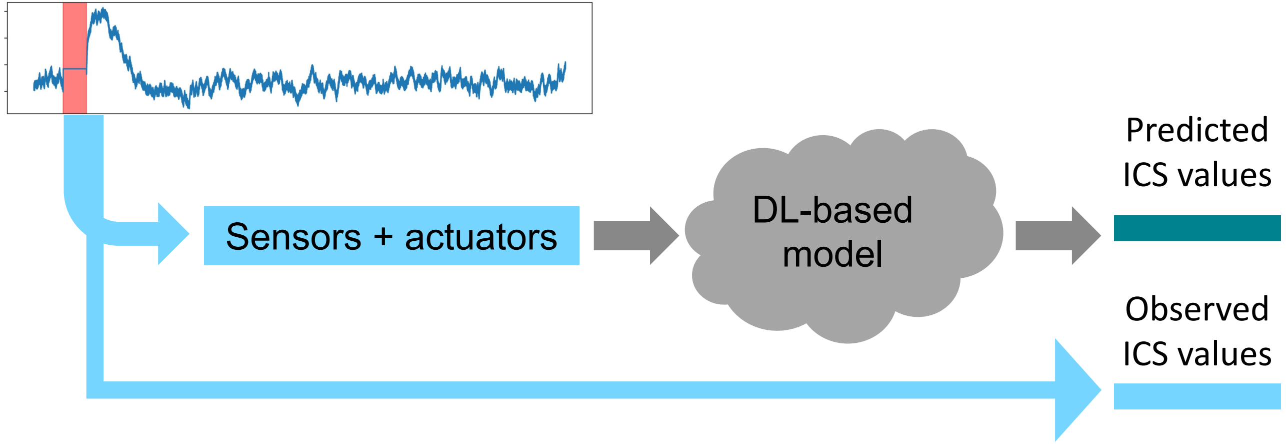
Sensors + actuators

DL-based model

Predicted ICS values

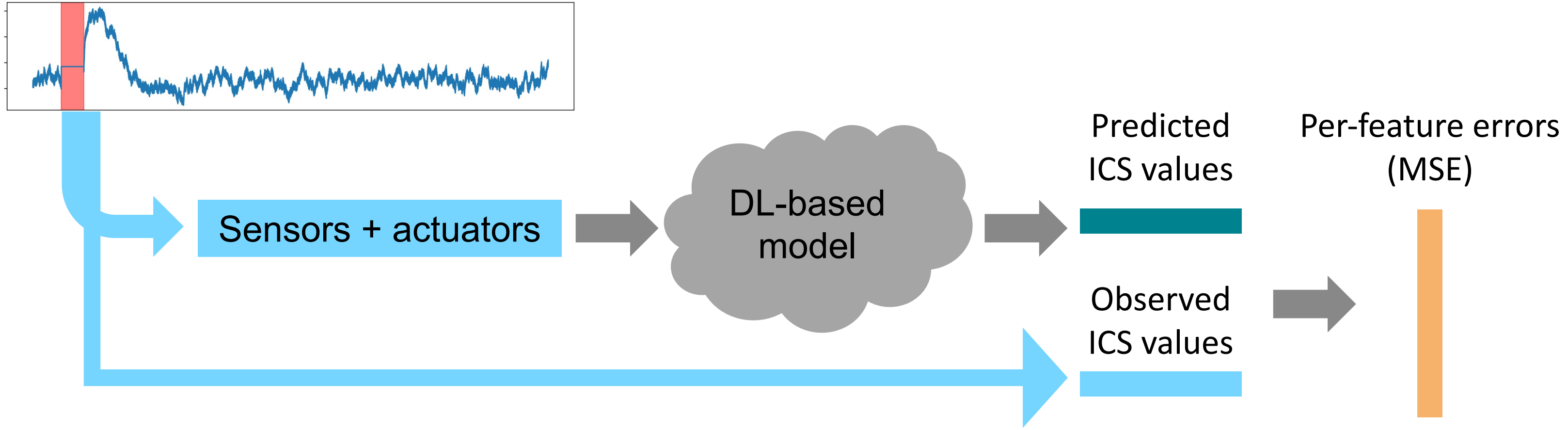
ICS anomaly attribution: previously

- Compare model prediction to observed ICS values



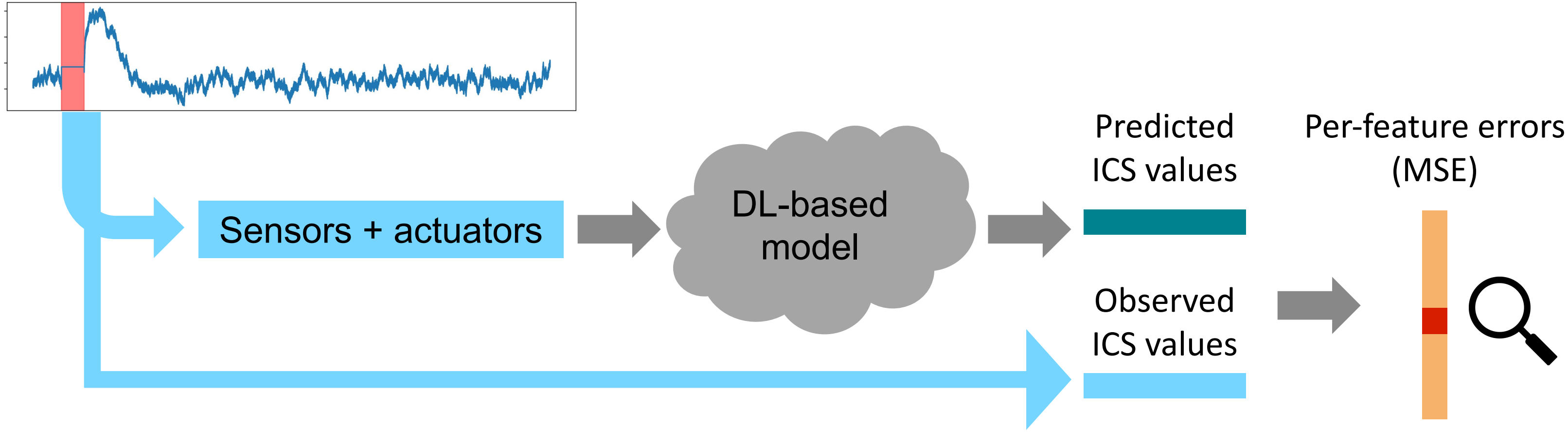
ICS anomaly attribution: previously

- Compare model prediction to observed ICS values



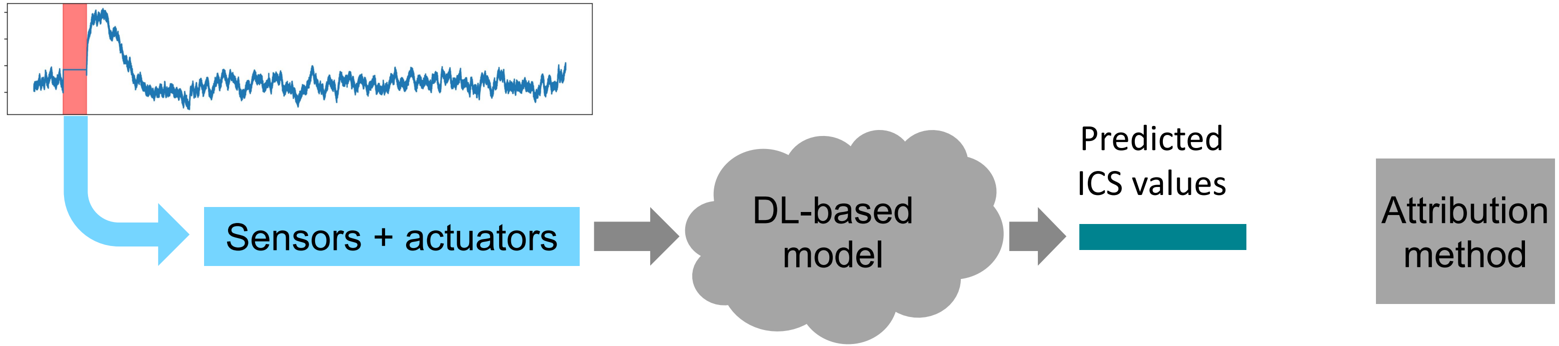
ICS anomaly attribution: previously

- Compare model prediction to observed ICS values
- Attribute alarm to feature with **highest error (MSE)**



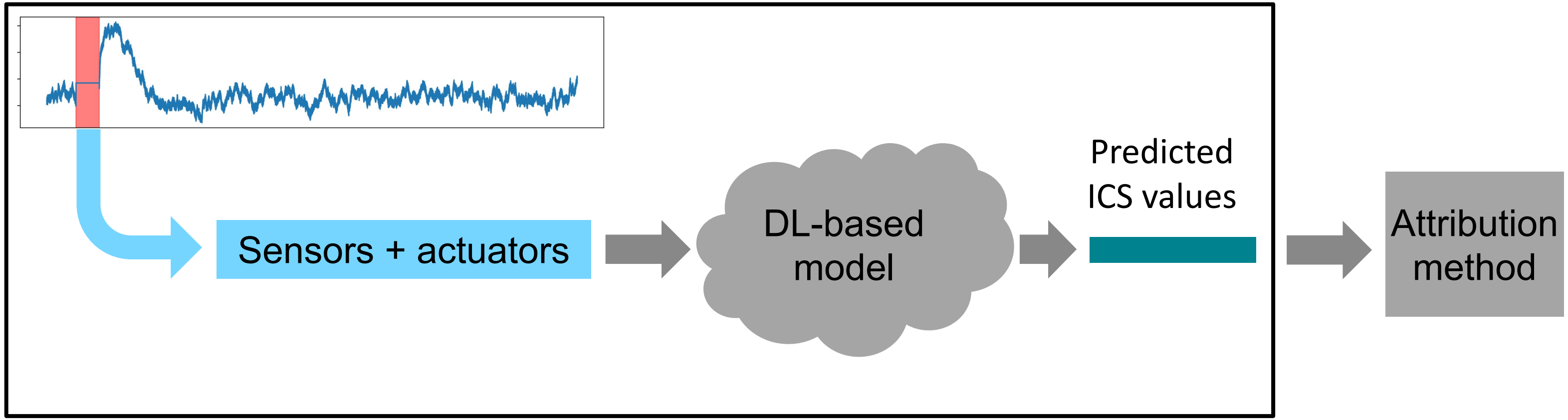
ICS anomaly attribution: our adaptation of XAI

- Adapt local (e.g., SHAP, LEMNA) and gradient-based (e.g., saliency maps) attribution methods for anomaly detection



ICS anomaly attribution: our adaptation of XAI

- Adapt local (e.g., SHAP, LEMNA) and gradient-based (e.g., saliency maps) attribution methods for anomaly detection
- Use (i) time-series data and (ii) model as attribution method input



Sensors + actuators

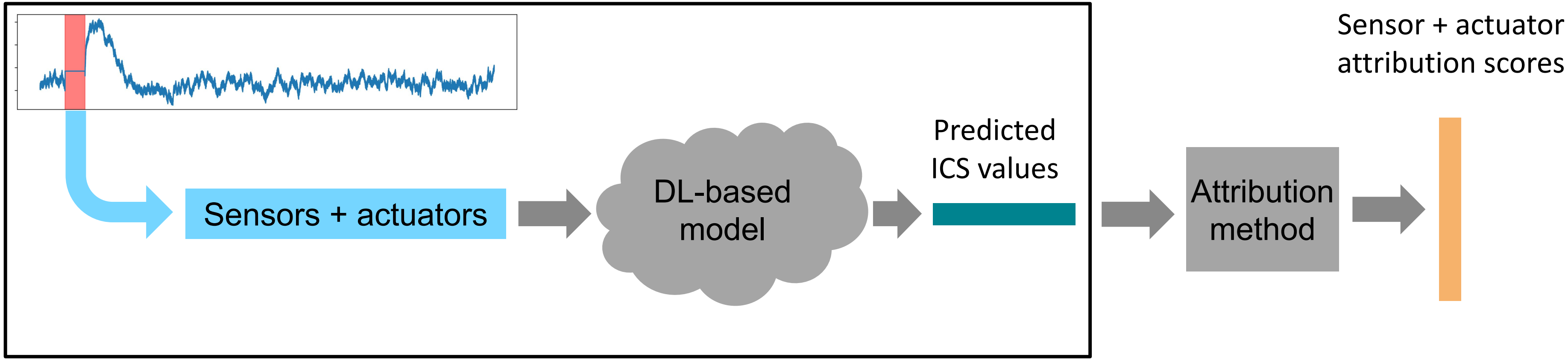
DL-based model

Predicted ICS values

Attribution method

ICS anomaly attribution: <u>our adaptation of XAI</u>

- Adapt local (e.g., SHAP, LEMNA) and gradient-based (e.g., saliency maps) attribution methods for anomaly detection
- Use (i) time-series data and (ii) model as attribution method input



We evaluate attribution for practical workflows

- Prior methods: does the attacked feature match the highest score?

We evaluate attribution for practical workflows

- Prior methods: does the attacked feature match the highest score?
    - Not how attribution scores would be used in practice

We evaluate attribution for practical workflows

- Prior methods: does the attacked feature match the highest score?
  - Not how attribution scores would be used in practice
- Preliminary survey of ICS operators (n=7)

We evaluate attribution for practical workflows

- Prior methods: does the attacked feature match the highest score?
  - Not how attribution scores would be used in practice
- Preliminary survey of ICS operators (n=7)
  - Operators prefer to see multiple features, but not necessarily all

We evaluate attribution for practical workflows

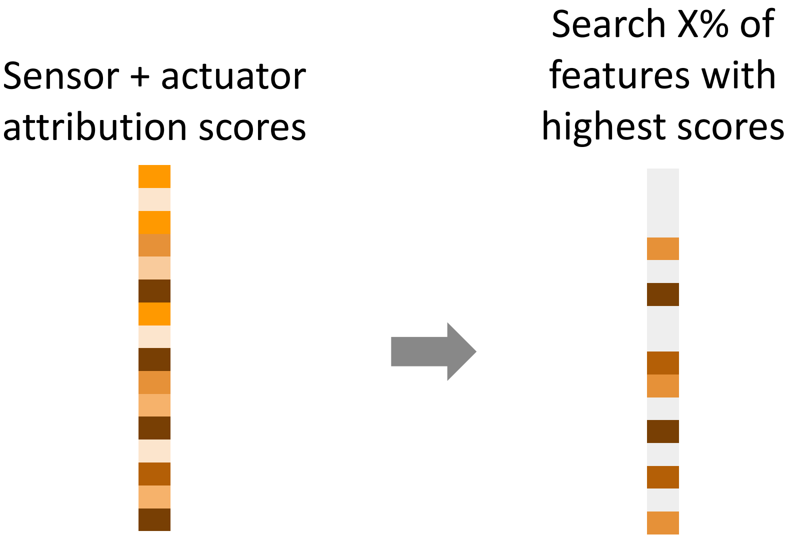
- Prior methods: does the attacked feature match the highest score?
  - Not how attribution scores would be used in practice
- Preliminary survey of ICS operators (n=7)
  - Operators prefer to see multiple features, but not necessarily all
  - Trade-off between number of features seen and accuracy

“A balanced trade-off is needed. Often having [a] list of max 10 [sensors] with minimal error rate is more useful than having less with high error rate.” –P4

We propose AvgRank to evaluate attribution

Sensor + actuator
attribution scores

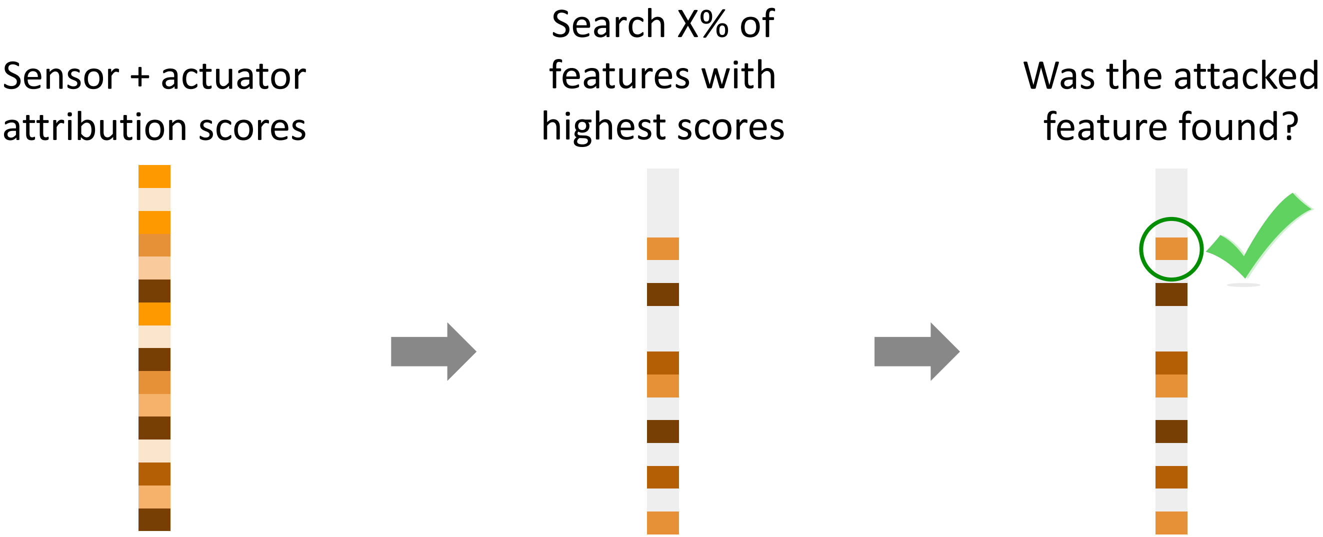
We propose AvgRank to evaluate attribution



Sensor + actuator attribution scores

Search X% of features with highest scores

We propose AvgRank to evaluate attribution



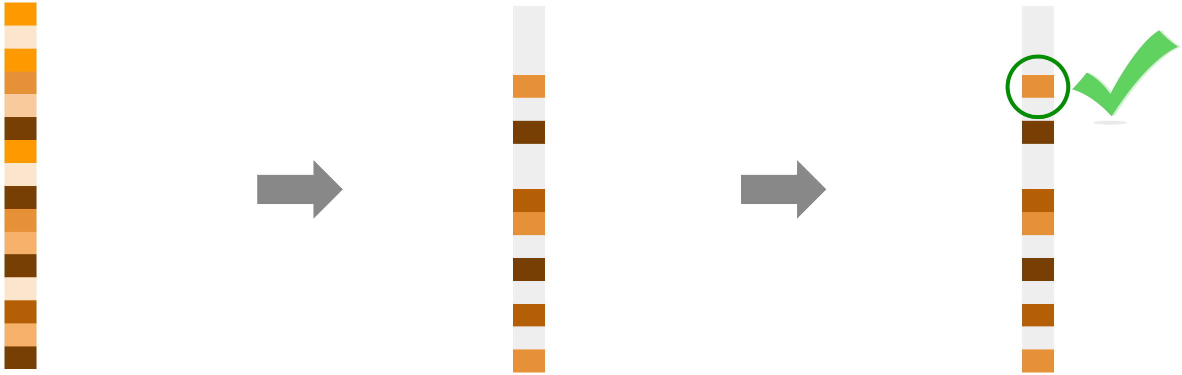
Sensor + actuator attribution scores

Search X% of features with highest scores

Was the attacked feature found?

We propose AvgRank to evaluate attribution

- AvgRank: % of features considered before finding manipulated feature



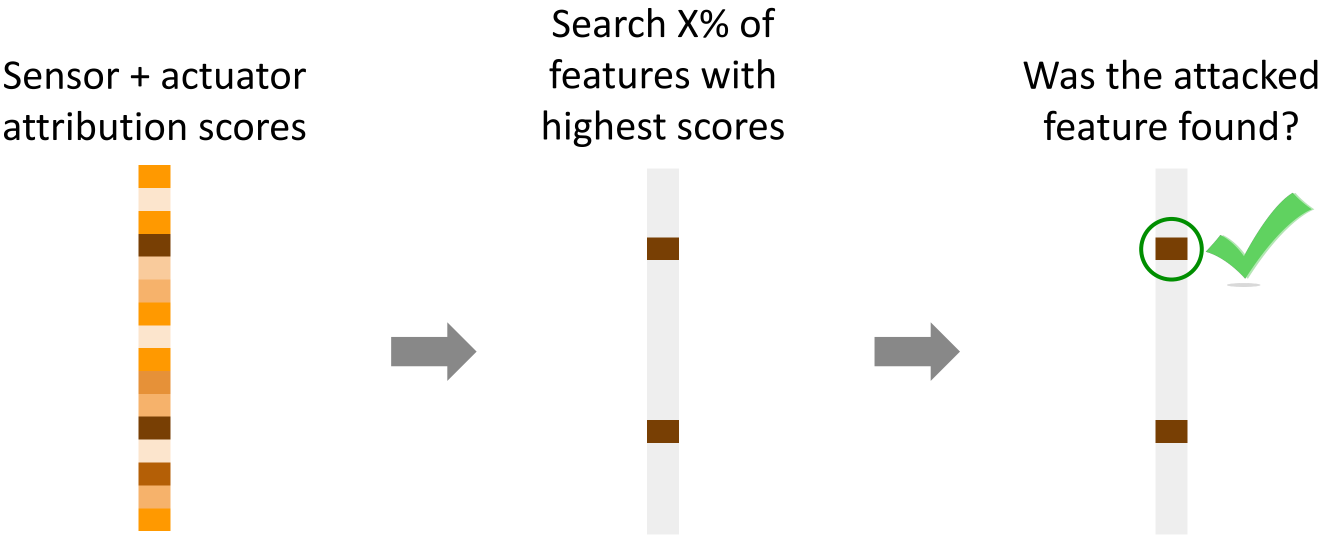
Sensor + actuator attribution scores

Search X% of features with highest scores

Was the attacked feature found?

We propose AvgRank to evaluate attribution

- AvgRank: % of features considered before finding manipulated feature
  - Lower AvgRank is better: operators consider fewer features, save time



Sensor + actuator attribution scores

Search X% of features with highest scores

Was the attacked feature found?

Using MSE ranking for attribution performs worse than previously reported

[1] C. Hwang and T. Lee, “E-SFD: Explainable sensor fault detection in the ICS anomaly detection system,” IEEE Access, vol. 9, 2021.
[2] M. Kravchik and A. Shabtai, “Efficient cyber attack detection in industrial control systems using lightweight neural networks and PCA,” IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 4, 2022.

Using MSE ranking for attribution performs worse than previously reported

- Prior work [1,2] evaluates attribution on a few case-study attacks
  - Examples where attacked feature has highest MSE

[1] C. Hwang and T. Lee, “E-SFD: Explainable sensor fault detection in the ICS anomaly detection system,” IEEE Access, vol. 9, 2021.
[2] M. Kravchik and A. Shabtai, “Efficient cyber attack detection in industrial control systems using lightweight neural networks and PCA,” IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 4, 2022.

Using MSE ranking for attribution performs worse than previously reported

- Prior work [1,2] evaluates attribution on a few case-study attacks
  - Examples where attacked feature has highest MSE
- When evaluated across our set of 147 diverse attacks:
  - Attacked feature has highest MSE in **<40% of attacks**

[1] C. Hwang and T. Lee, “E-SFD: Explainable sensor fault detection in the ICS anomaly detection system,” IEEE Access, vol. 9, 2021.
[2] M. Kravchik and A. Shabtai, “Efficient cyber attack detection in industrial control systems using lightweight neural networks and PCA,” IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 4, 2022.

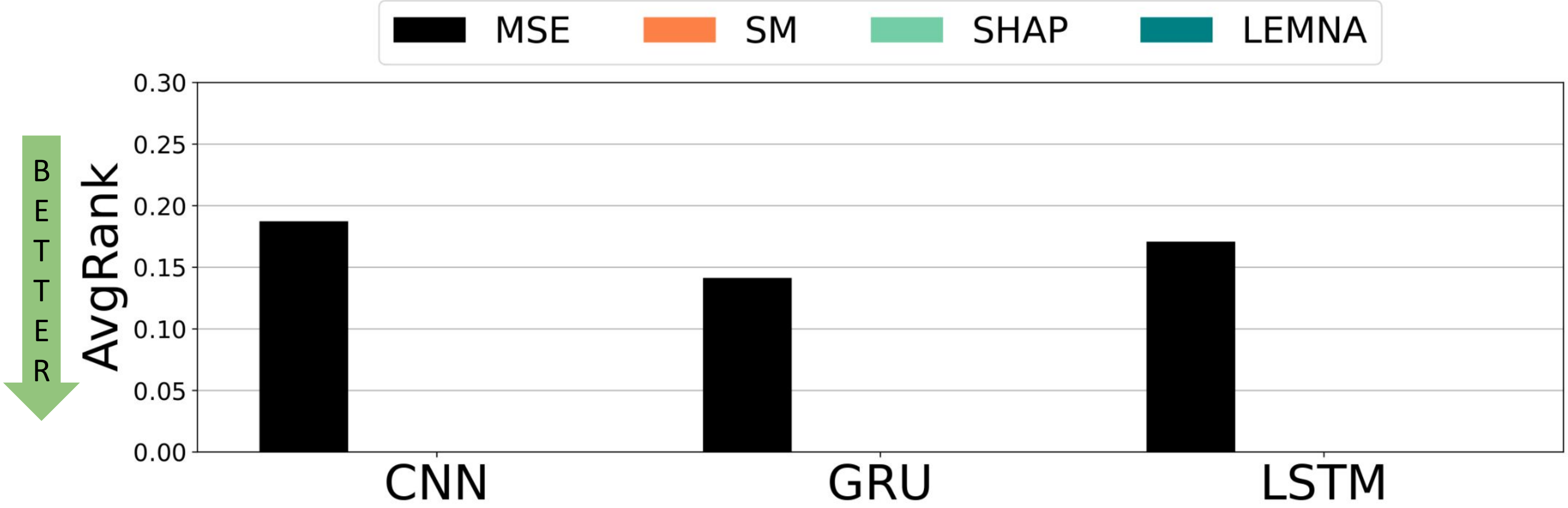
Using MSE ranking for attribution performs worse than previously reported

- Prior work [1,2] evaluates attribution on a few case-study attacks
  - Examples where attacked feature has highest MSE
- When evaluated across our set of 147 diverse attacks:
  - Attacked feature has highest MSE in **<40% of attacks**
  - On average, operators would have to consider **>14% of features** before finding attacked feature

[1] C. Hwang and T. Lee, "E-SFD: Explainable sensor fault detection in the ICS anomaly detection system," IEEE Access, vol. 9, 2021.
[2] M. Kravchik and A. Shabtai, "Efficient cyber attack detection in industrial control systems using lightweight neural networks and PCA," IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 4, 2022.

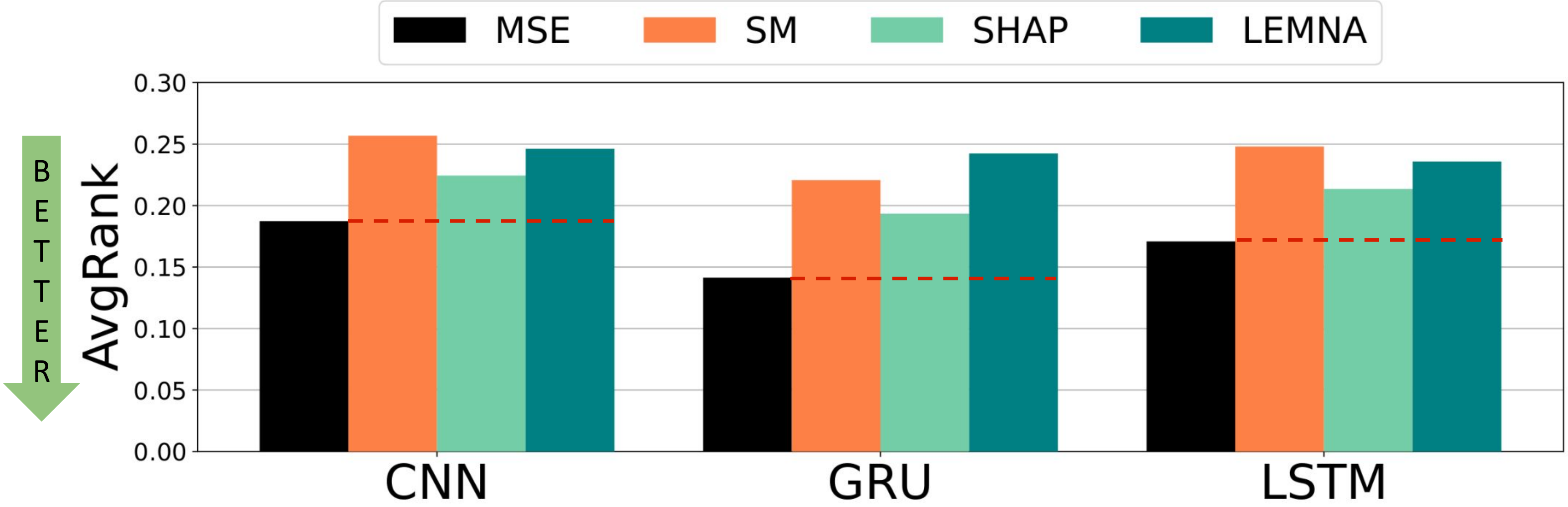
Do attribution methods perform better?

- Three best-performing attribution methods (SM, SHAP, LEMNA)



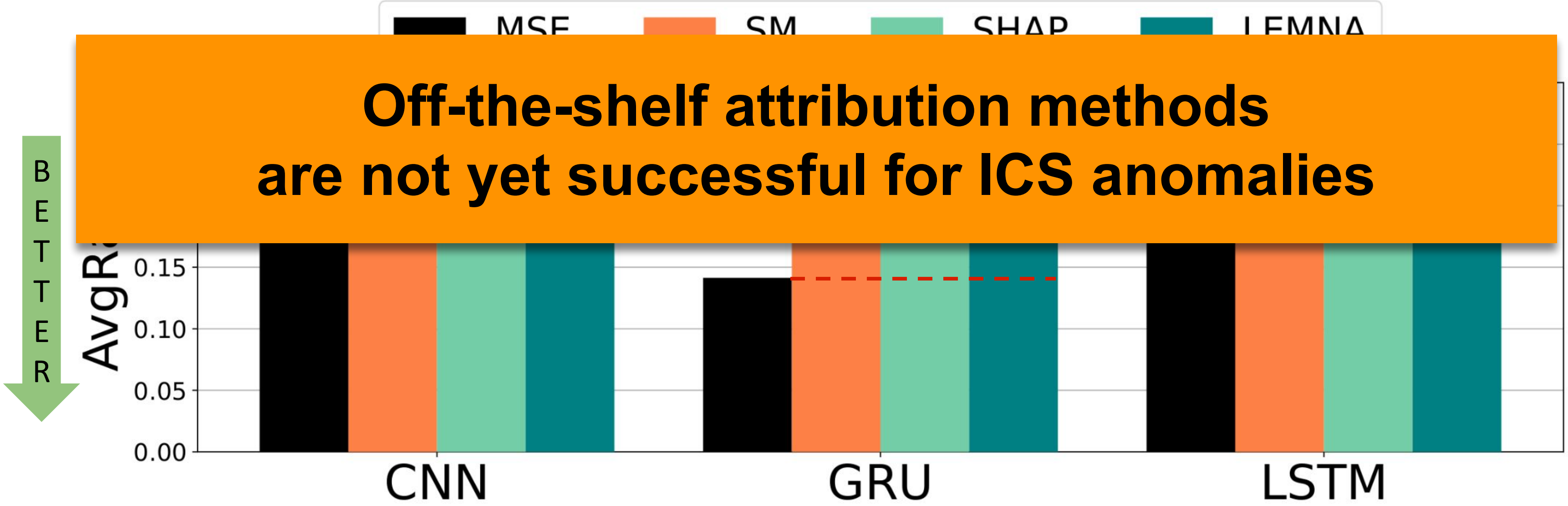
Attribution methods perform worse than MSE

- Three best-performing attribution methods (SM, SHAP, LEMNA):
  - Surprisingly, attribution methods are consistently worse than MSE



Attribution methods perform worse than MSE

- Three best-performing attribution methods (SM, SHAP, LEMNA):
  - Surprisingly, attribution methods are consistently worse than MSE



Off-the-shelf attribution methods are not yet successful for ICS anomalies

RQ2: How do ICS attack properties affect attribution?

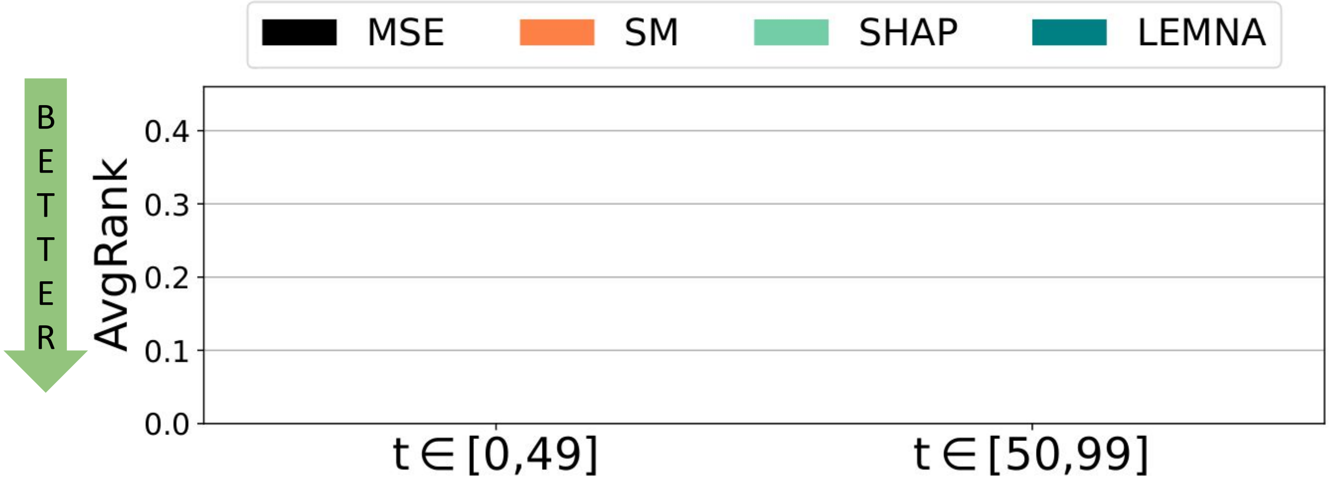
Why are attributions worse than expected?

- Broad differences among our 147 ICS attacks:
  - Detection outcomes
    - Latency, if detected, etc.
  - Input manipulation
    - Magnitude, location, pattern

Why are attributions worse than expected?

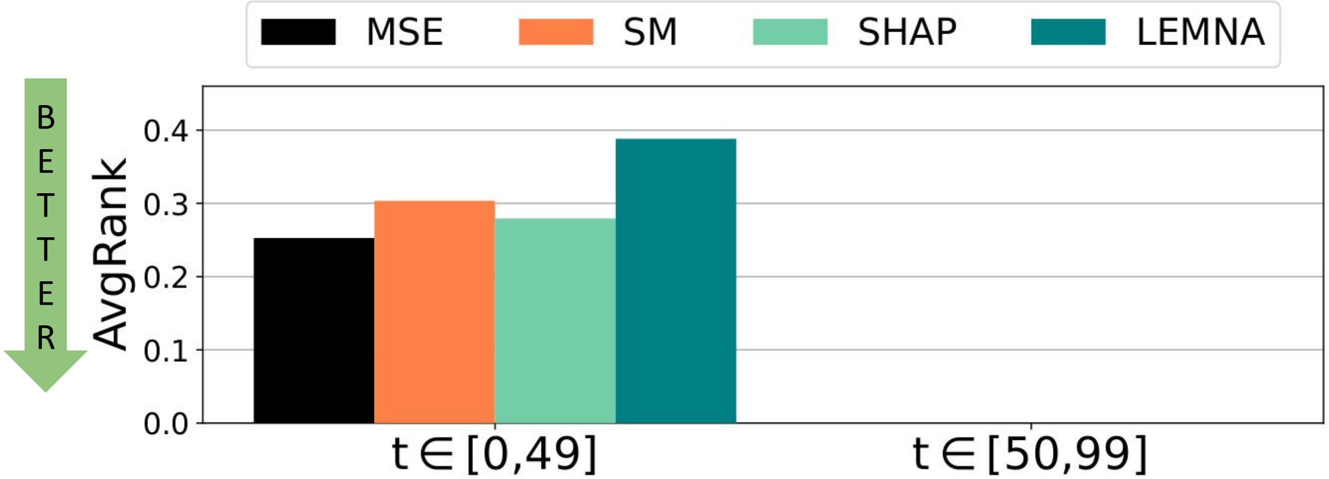
- Broad differences among our 147 ICS attacks:
  - **Detection outcomes**
    - **Latency**, **if detected**, etc.
  - Input manipulation
    - Magnitude, location, pattern

Attribution accuracy varies by detection latency



Attribution accuracy varies by detection latency

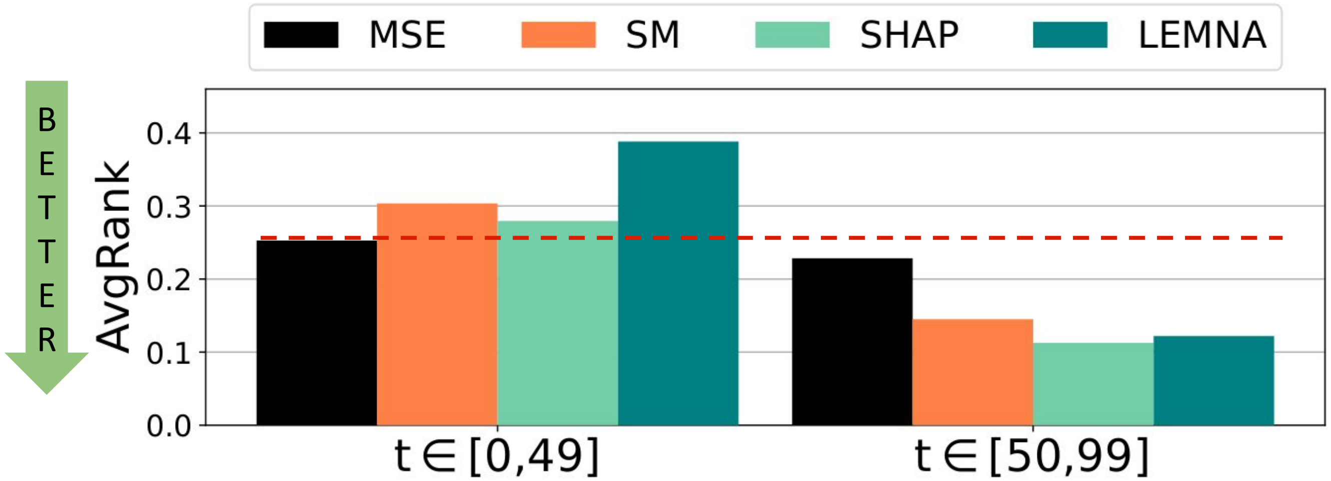
- Attributions are inaccurate within first 50 seconds



CyLab Carnegie Mellon University Security and Privacy Institute

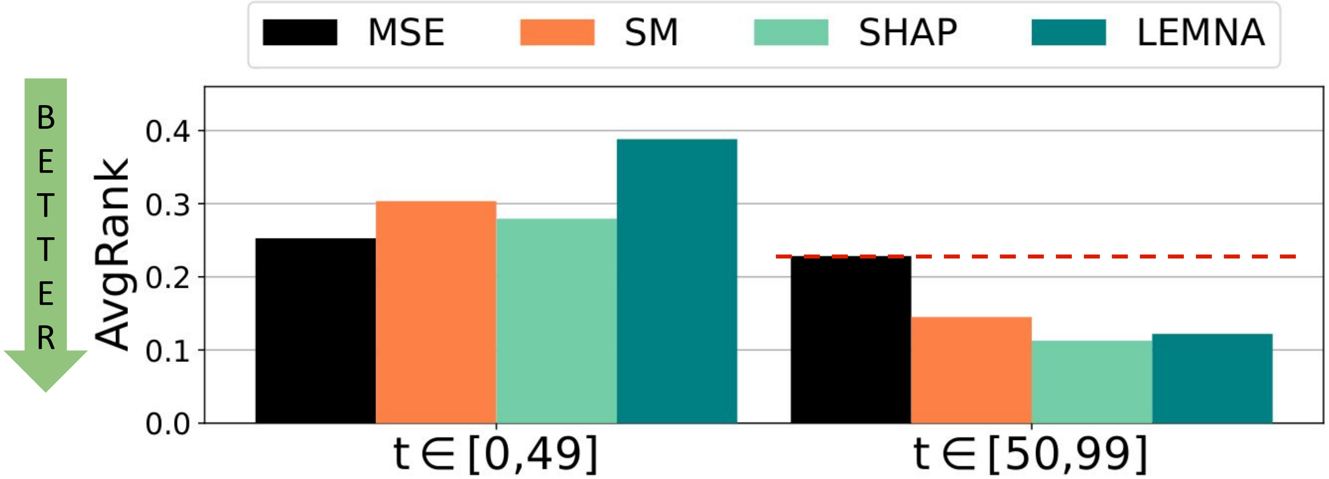
Attribution accuracy varies by detection latency

- Attributions are inaccurate within first 50 seconds
  - But improve when computed within 50-100 seconds



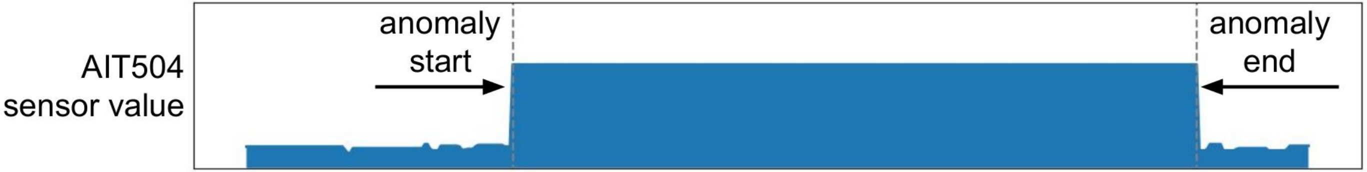
Attribution accuracy varies by detection latency

- Attributions are inaccurate within first 50 seconds
  - But improve when computed within 50-100 seconds
  - SM, SHAP, LEMNA now outperform MSE



How can timing affect attribution?

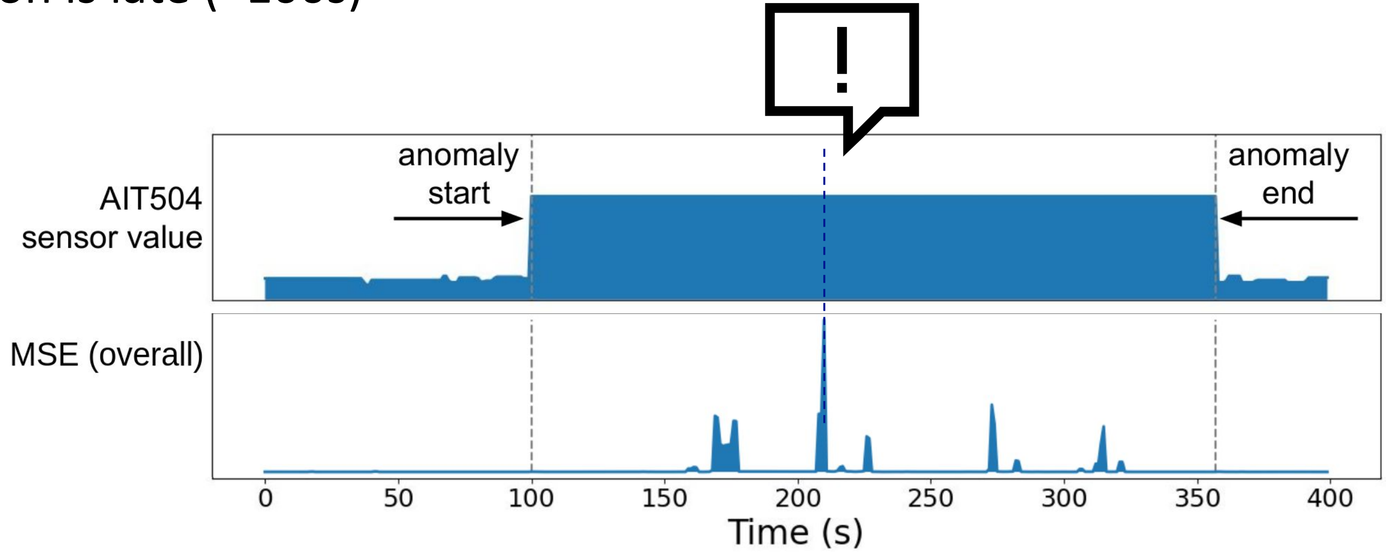
Example: SWaT attack #10



How can timing affect attribution?

Example: SWaT attack #10

Detection is late (~100s)



How can timing affect attribution?

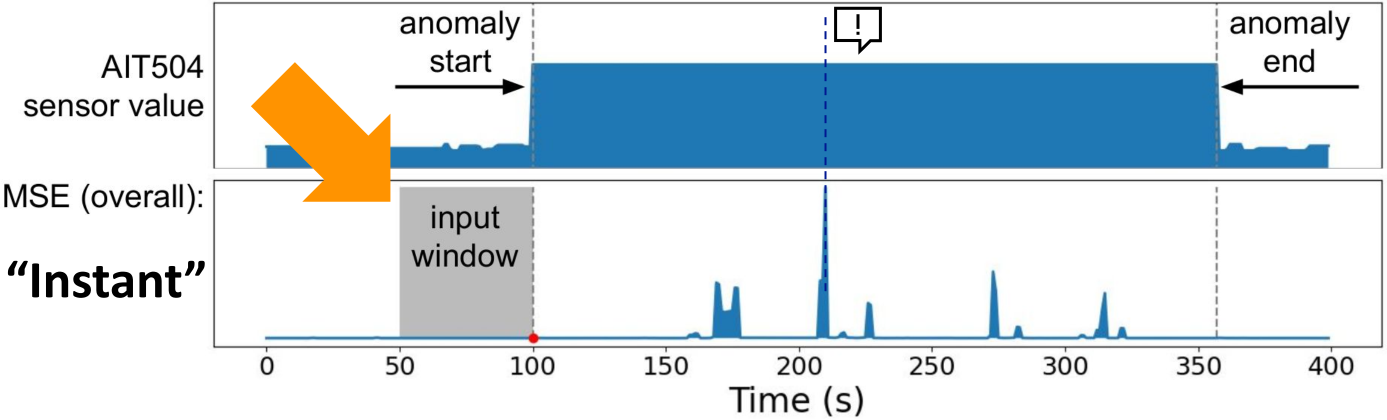
Example: SWaT attack #10

Detection is late (~100s)

Option 1: "Instant"

Input window preceding anomaly start

Input is mostly benign data



How can timing affect attribution?

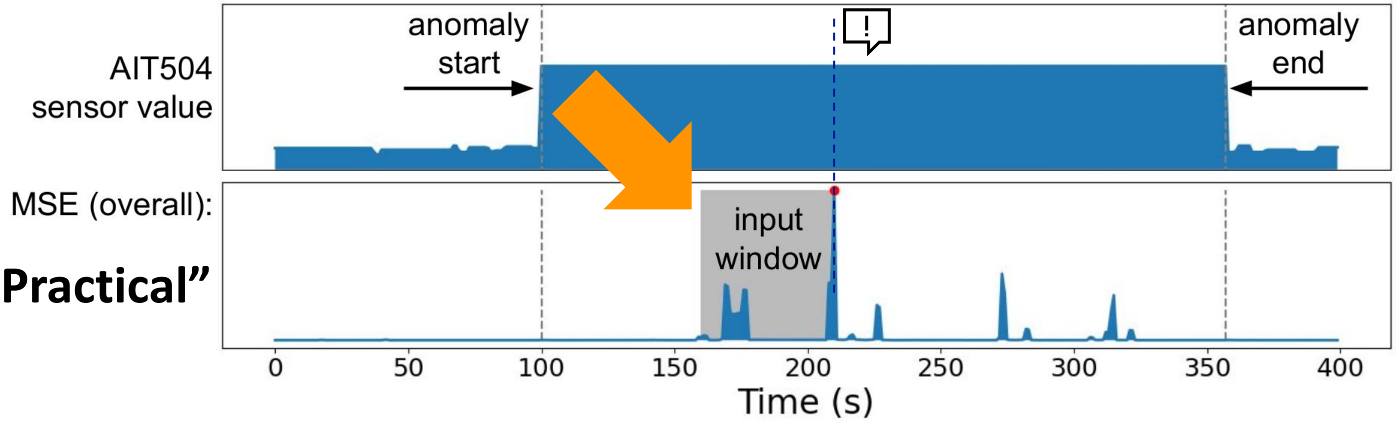
Example: SWaT attack #10

Detection is late (~100s)

Option 2: "Practical"

Input window preceding detection

Realistic, but late



How can timing affect attribution?

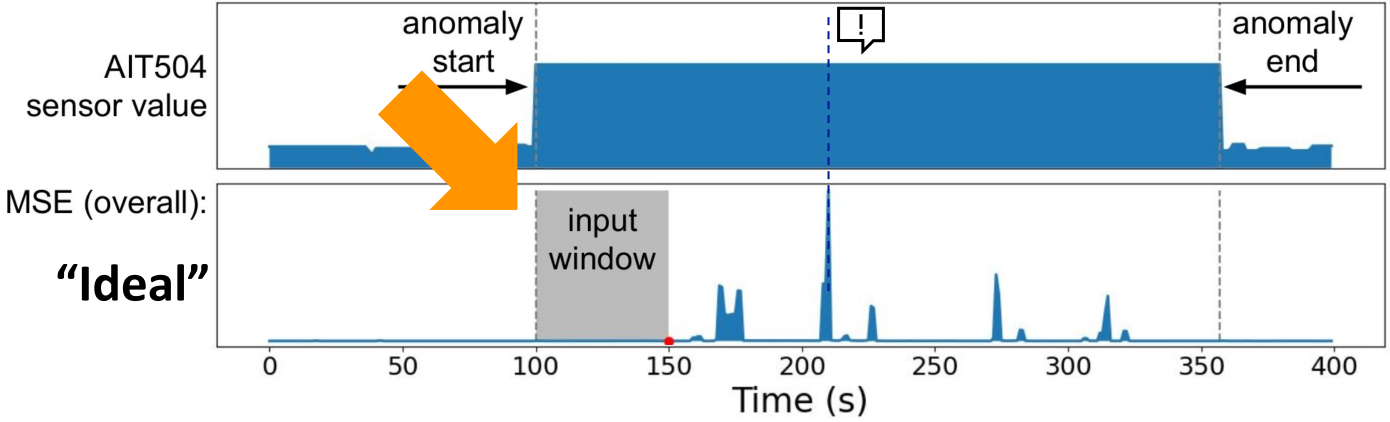
Example: SWaT attack #10

Detection is late (~100s)

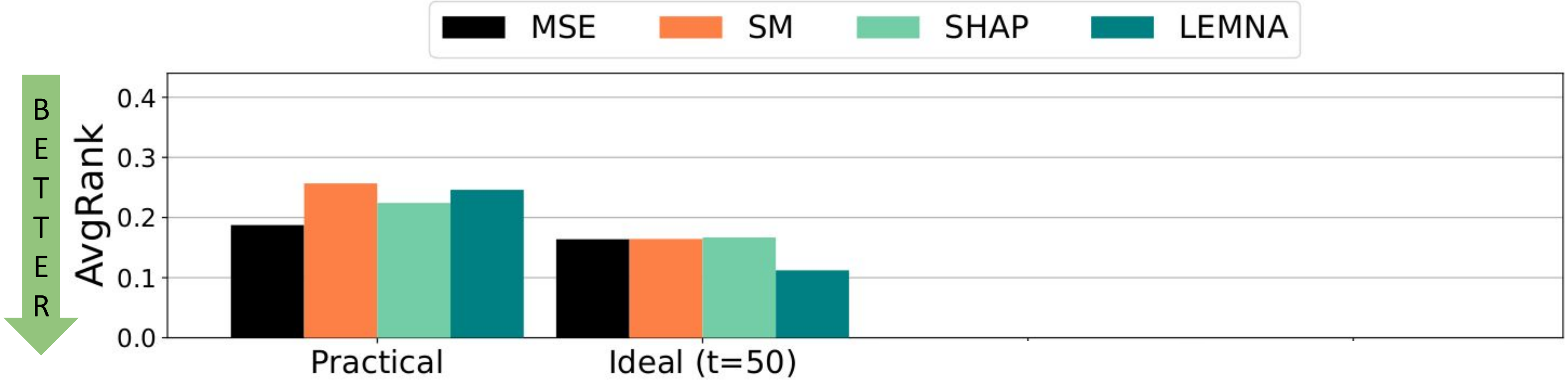
Option 3: “Ideal”

Input window begins with anomaly start

Ideal, but unknown in real time

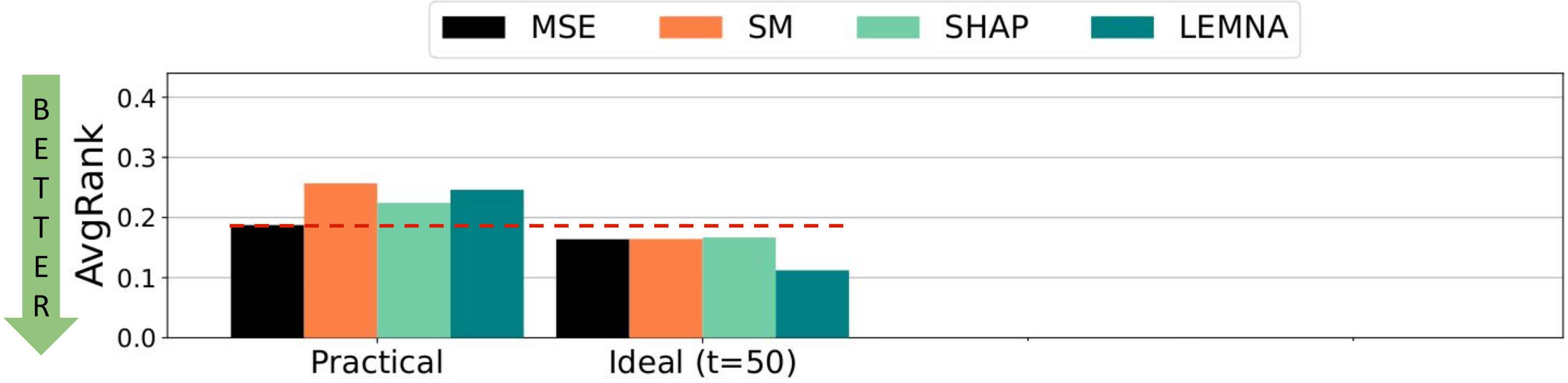


Attribution accuracy varies by timing strategy



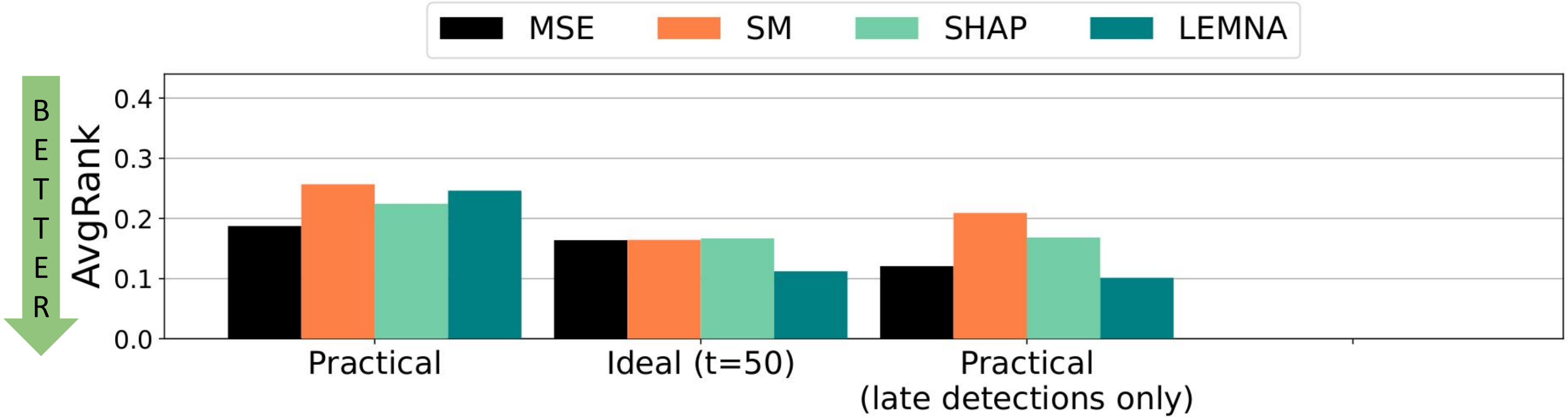
Attribution accuracy varies by timing strategy

- Ideal timing outperforms practical outcomes



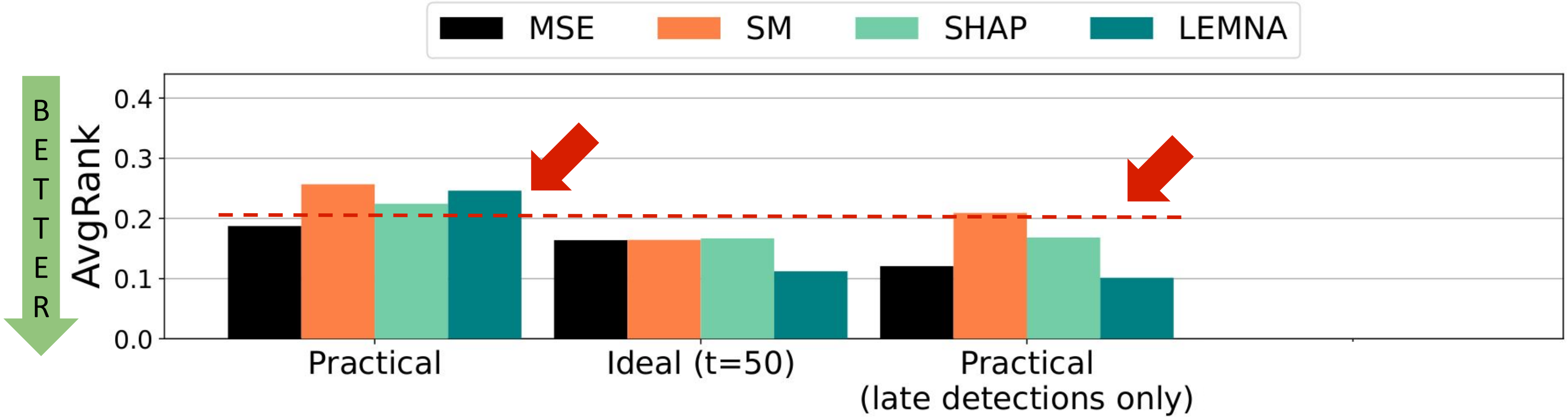
Attribution accuracy varies by timing strategy

- Ideal timing outperforms practical outcomes



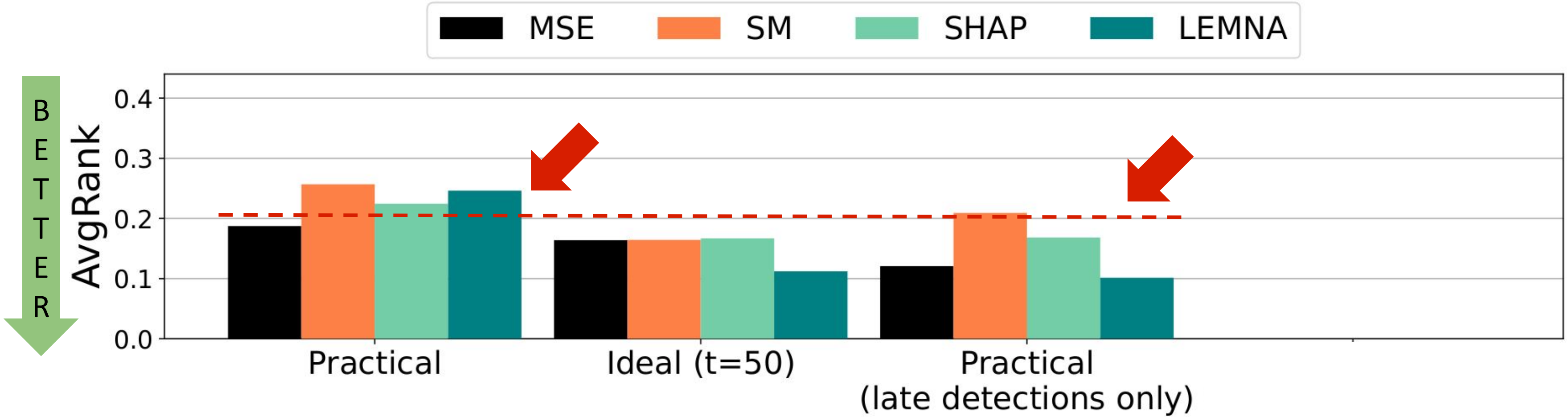
Attribution accuracy varies by timing strategy

- Ideal timing outperforms practical outcomes



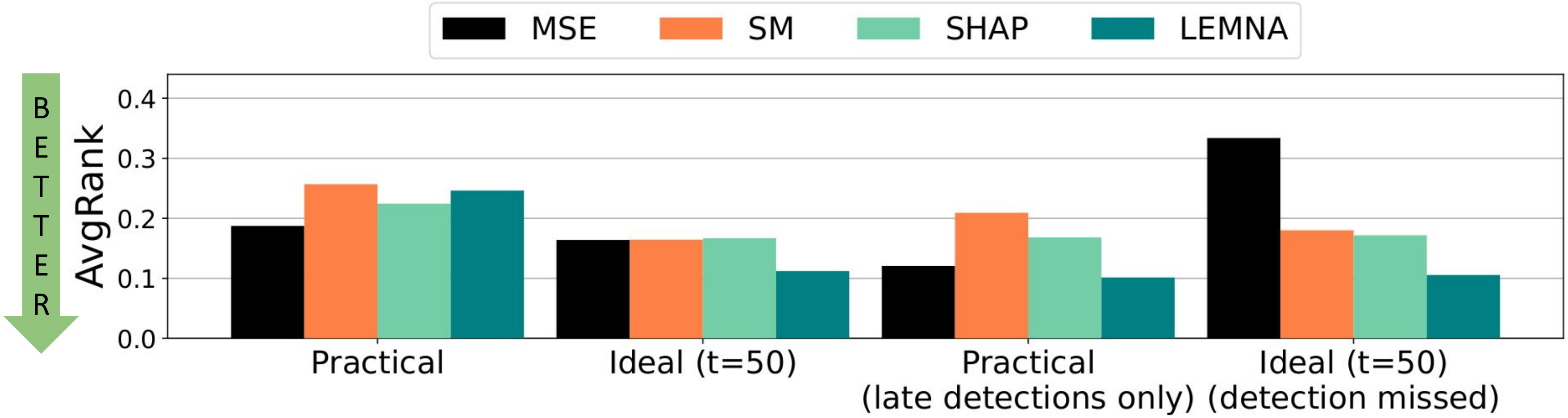
Attribution accuracy varies by timing strategy

- Ideal timing outperforms practical outcomes
- Avoiding “early” timings improves practical attribution results



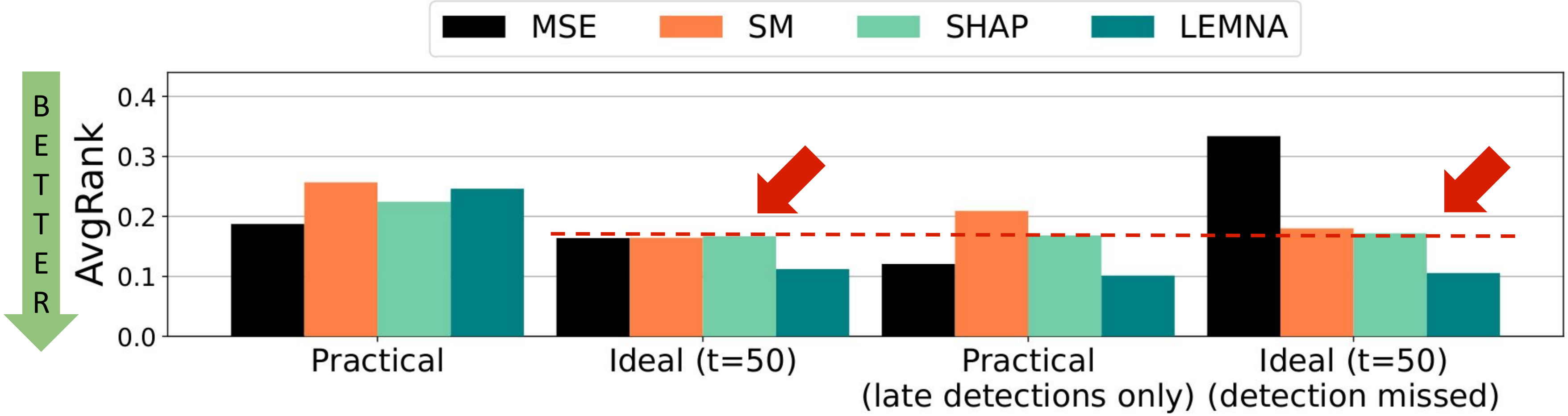
Attribution accuracy varies by timing strategy

- Ideal timing outperforms practical outcomes
- Avoiding “early” timings improves practical attribution results



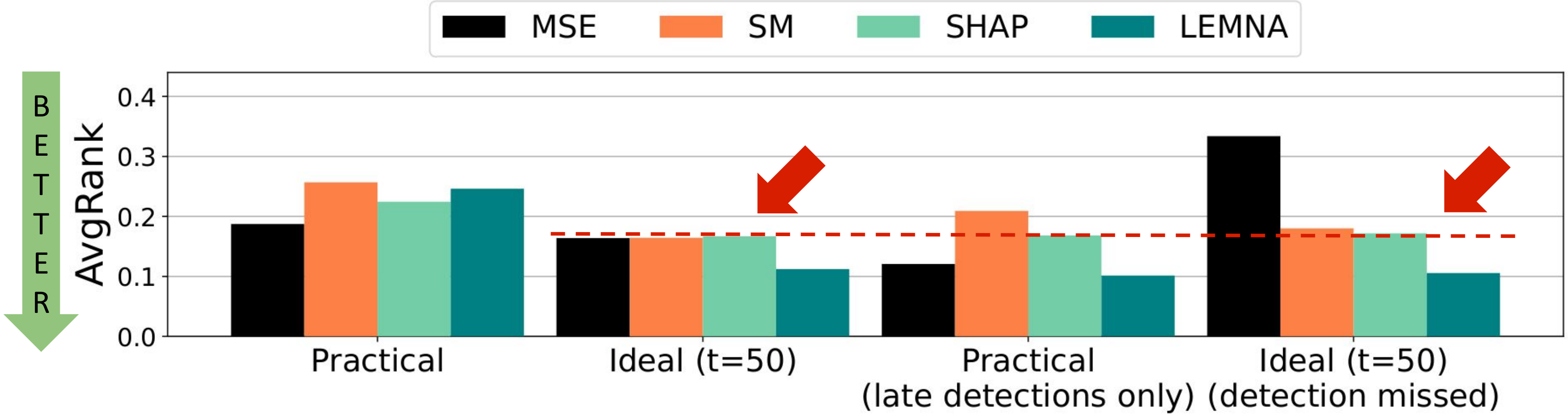
Attribution accuracy varies by timing strategy

- Ideal timing outperforms practical outcomes
- Avoiding “early” timings improves practical attribution results



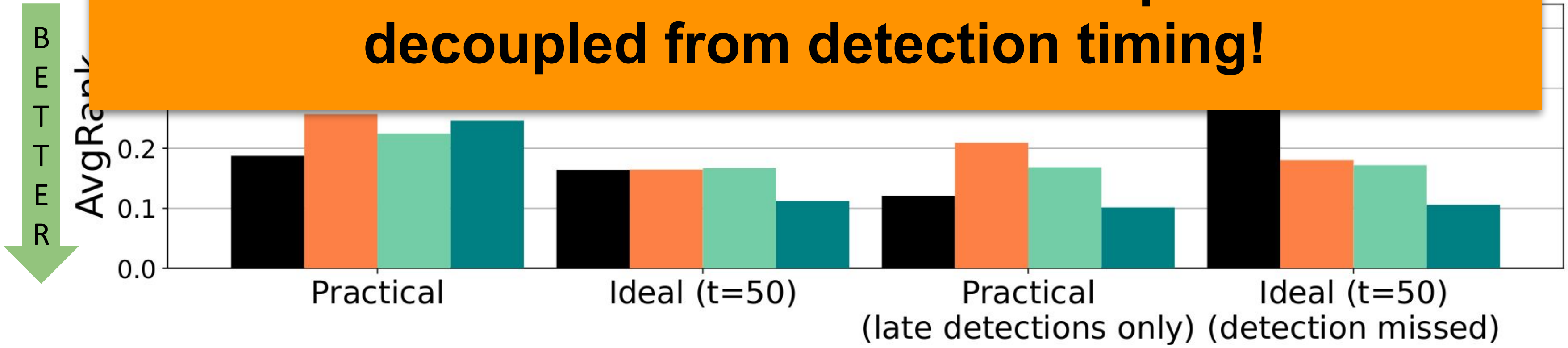
Attribution accuracy varies by timing strategy

- Ideal timing outperforms practical outcomes
- Avoiding “early” timings improves practical attribution results
- Attribution without alarms can be useful



Attribution accuracy varies by timing strategy

- Ideal timing outperforms practical outcomes
- Avoiding “early” timings improves practical attribution results
- Attribution without alarms can be useful

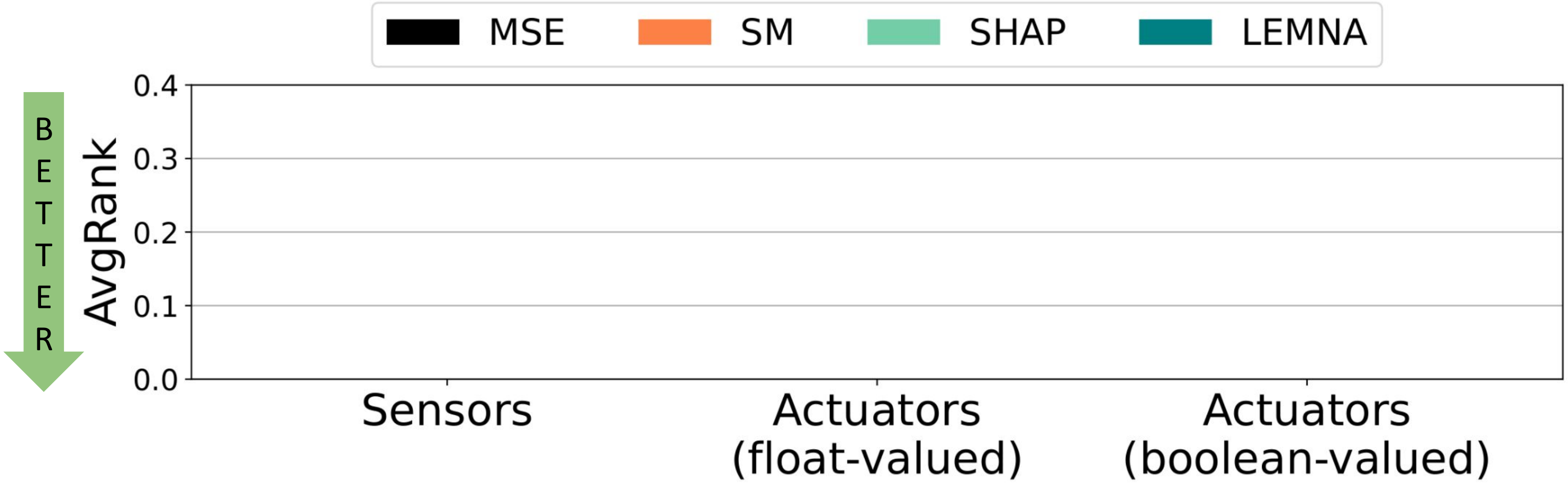


Attribution of ICS anomalies can improve when decoupled from detection timing!

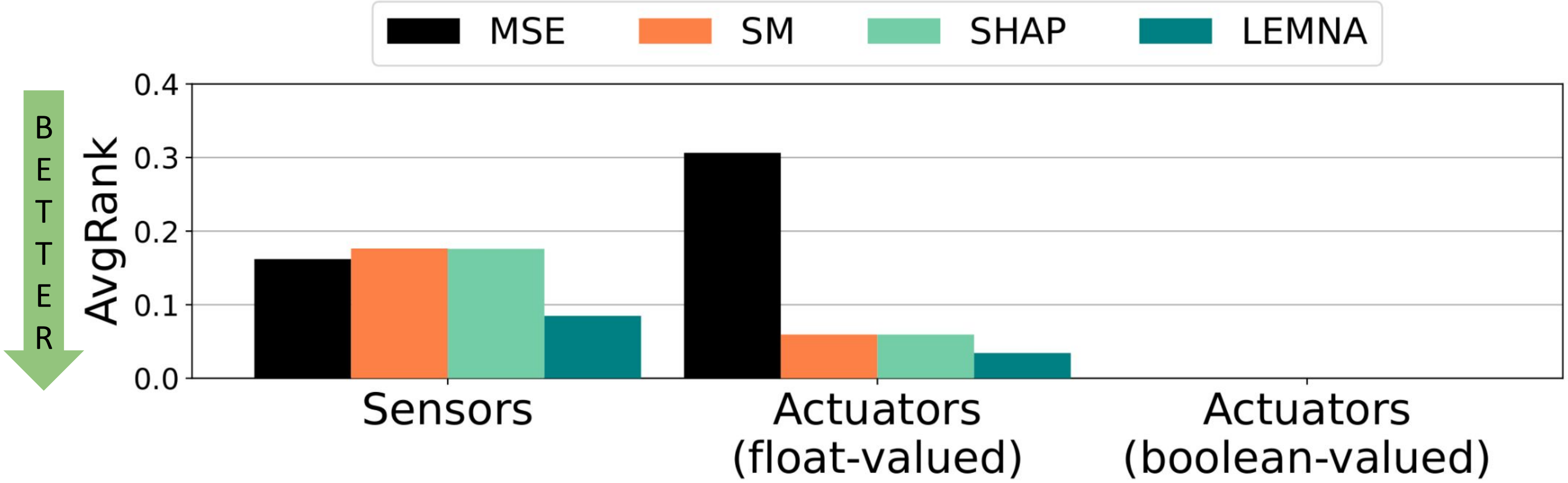
Why are attributions worse than expected?

- Broad differences among our 147 ICS attacks:
  - Detection outcomes
    - Latency, if detected, etc.
  - Input manipulation
    - Magnitude, **location**, pattern

Attribution accuracy varies <u>by component type</u>

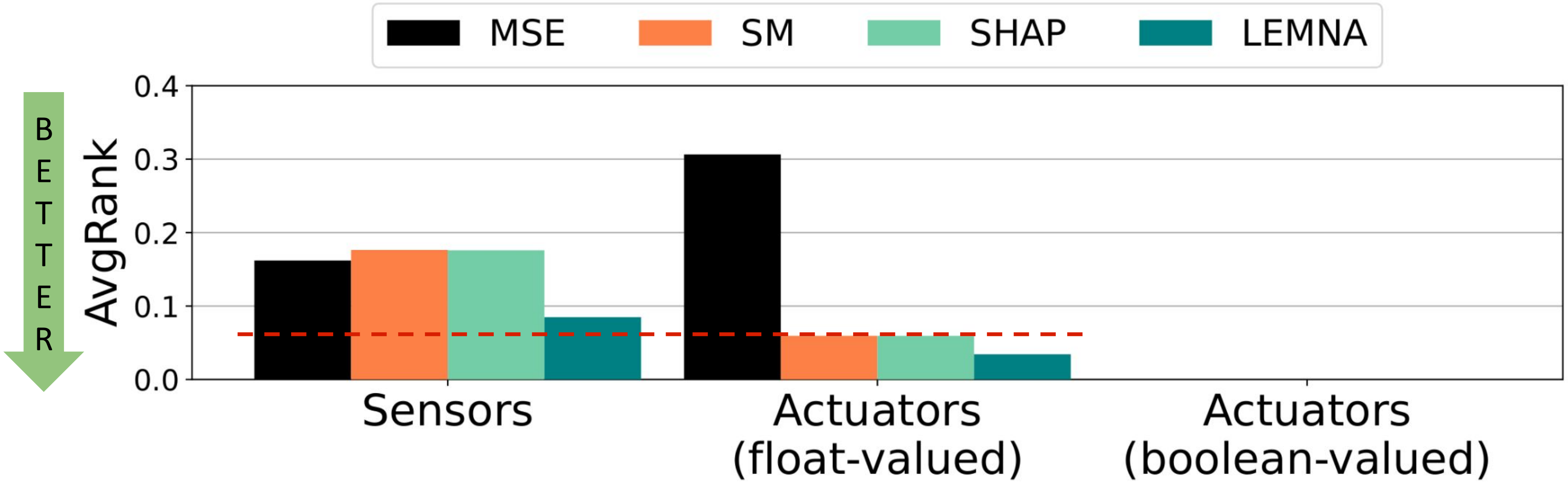


Attribution accuracy varies by component type



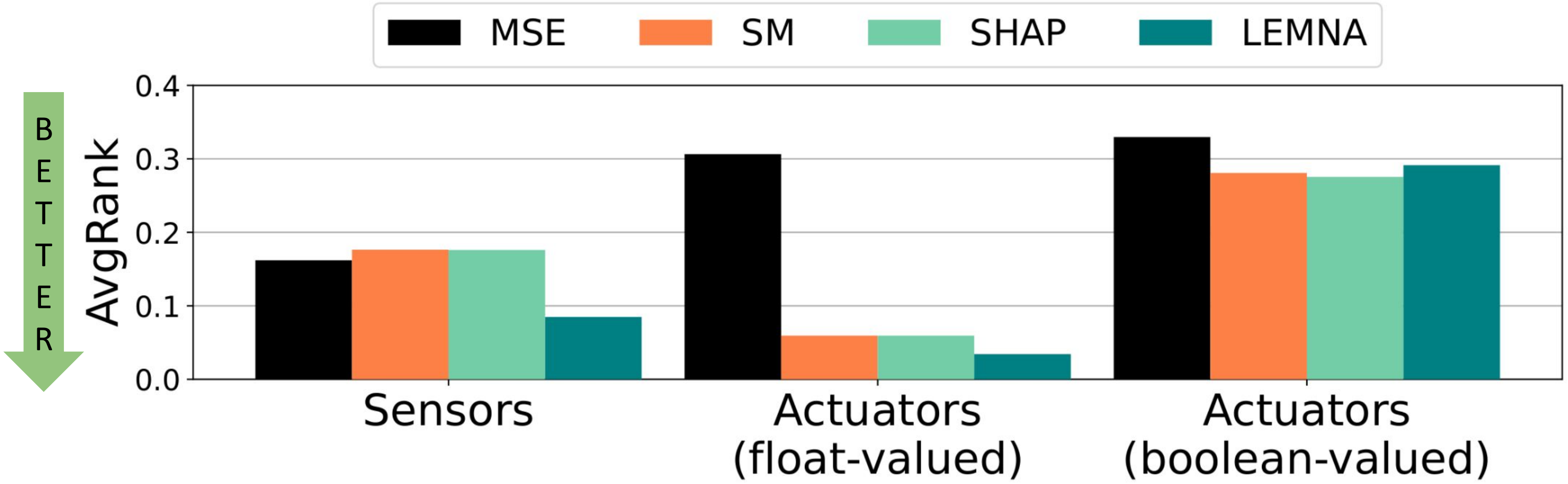
Attribution accuracy varies by component type

- Attribution methods: more accurate for float-valued actuators



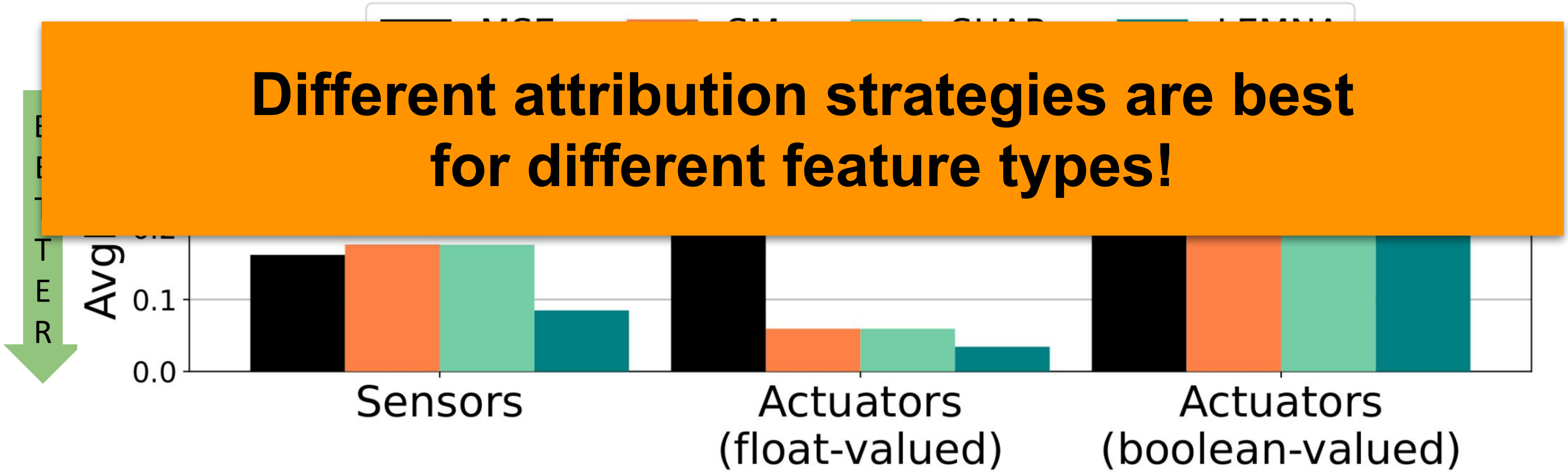
Attribution accuracy varies by component type

- Attribution methods: more accurate for float-valued actuators
- **Boolean-valued actuators** are difficult to attribute for all methods



Attribution accuracy varies by component type

- Attribution methods: more accurate for float-valued actuators
- **Boolean-valued actuators** are difficult to attribute for all methods

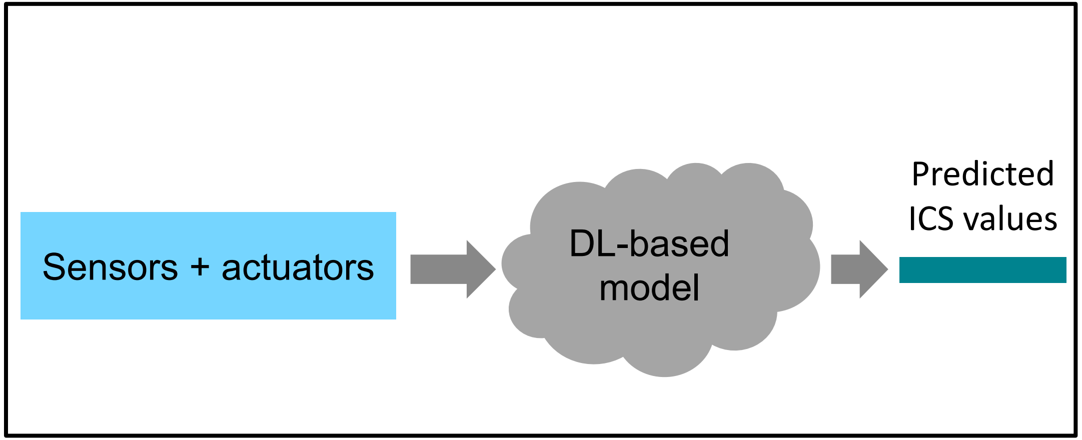


Different attribution strategies are best for different feature types!

RQ3: Can we do better than prior attribution strategies?

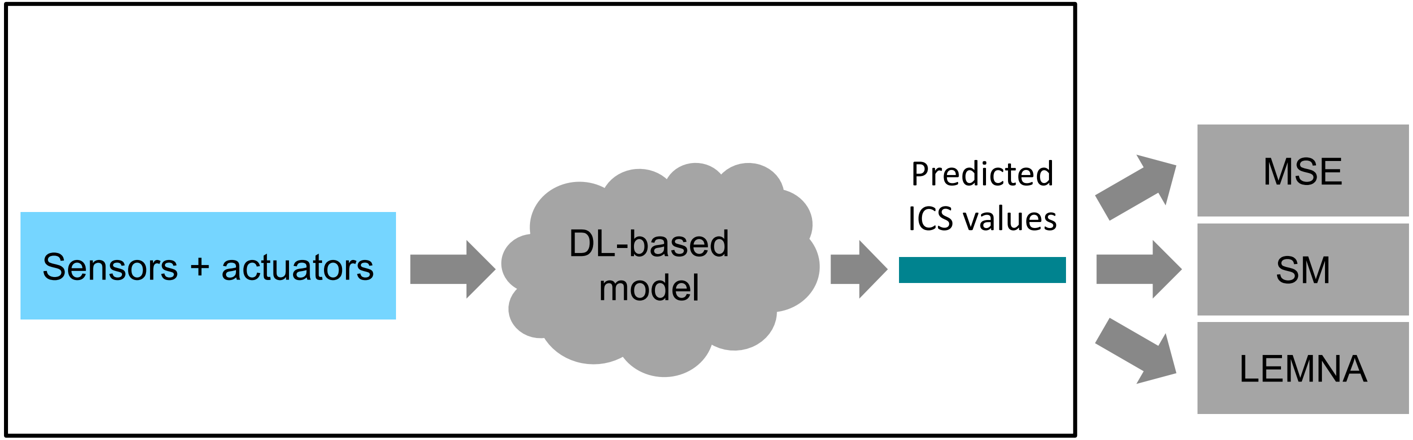
Better attributions via ensembles?

- Without knowing what attack or timing is used, can one strategy be best?



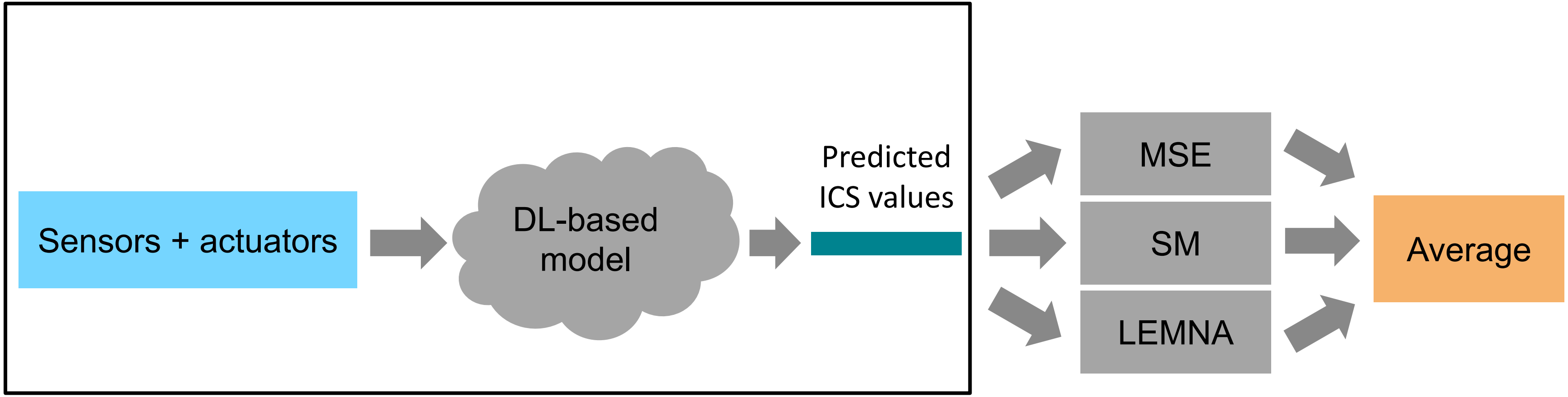
Better attributions via ensembles?

- Without knowing what attack or timing is used, can one strategy be best?
- We propose an ensemble attribution method:



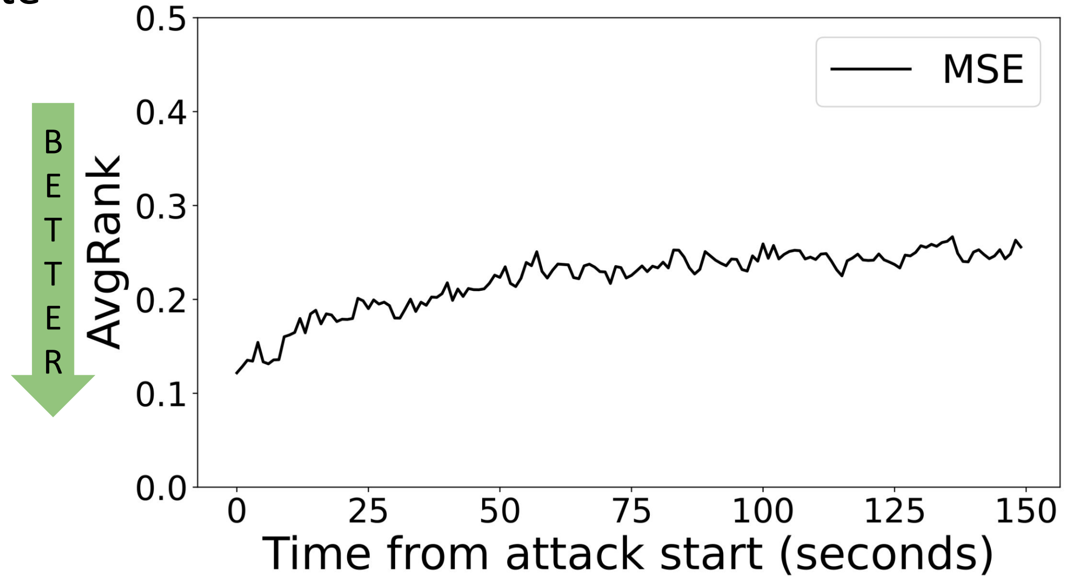
Better attributions via ensembles?

- Without knowing what attack or timing is used, can one strategy be best?
- We propose an ensemble attribution method:
  - Take the average of attribution scores (MSE, SM, LEMNA) for each feature



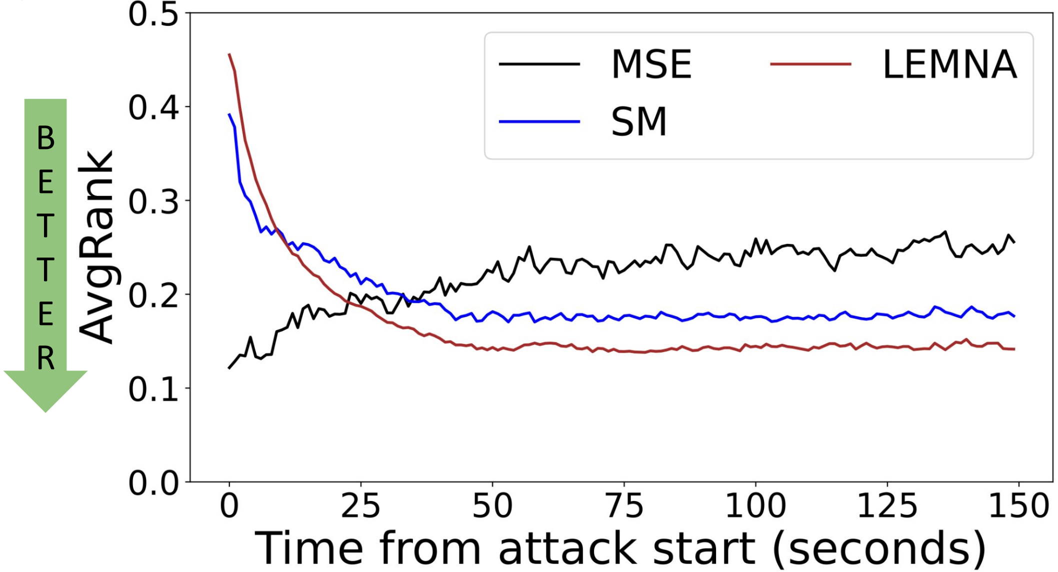
Better attributions via ensembles?

- MSE performs worst when late



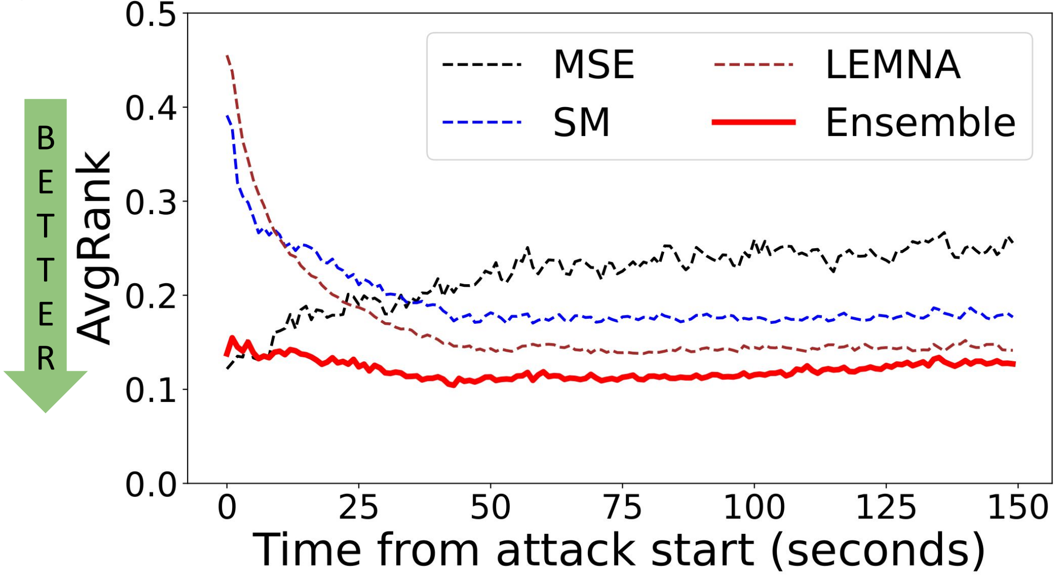
Better attributions via ensembles?

- MSE performs worst when late
- SM and LEMNA perform worst when early
  - Time-series history needed for attribution



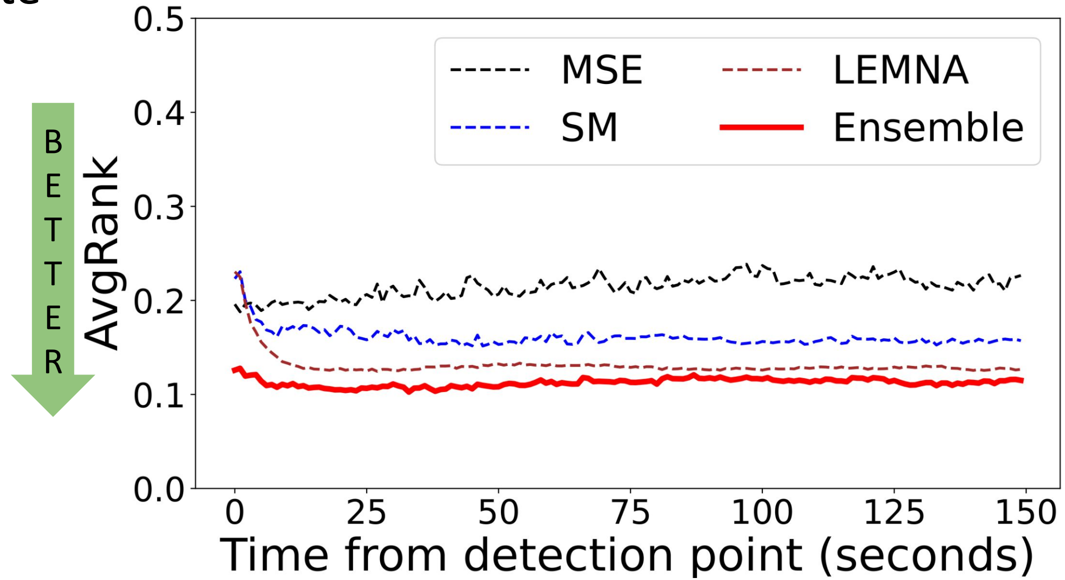
Better attributions via ensembles?

- MSE performs worst when late
- SM and LEMNA perform worst when early
  - Time-series history needed for attribution
- **Ensembles outperform all individual methods**



Better attributions via ensembles?

- MSE performs worst when late
- SM and LEMNA perform worst when early
  - Time-series history needed for attribution
- **Ensembles outperform all individual methods**
  - At practical timings too!



Attribution methods for ICS anomaly detection

Prior performance is worse than reported ☹

Attribution methods for ICS anomaly detection

Prior performance is worse than reported

ICS anomaly attribution is complex

Timing and feature types affect which methods work best

Attribution methods for ICS anomaly detection

Prior performance is worse than reported

ICS anomaly attribution is complex

Timing and feature types affect which methods work best

An ensemble approach balances tradeoffs

Though imperfect, attributions can help ICS operators

Attribution methods for ICS anomaly detection

Prior performance is worse than reported

ICS anomaly attribution is complex

Timing and feature types affect which methods work best

An ensemble approach balances tradeoffs

Though imperfect, attributions can help ICS operators

Clement Fung, Eric Zeng, Lujo Bauer
Carnegie Mellon University
clementf@cs.cmu.edu

Artifact Evaluated
NDSS SYMPOSIUM
Available
Functional

Synthetic attacks: https://doi.org/10.1184/R1/23805552
Modified simulator: https://github.com/pwwl/tep-attack-simulator
Attribution code: https://github.com/pwwl/ics-anomaly-attribution

CyLab Carnegie Mellon University Security and Privacy Institute