



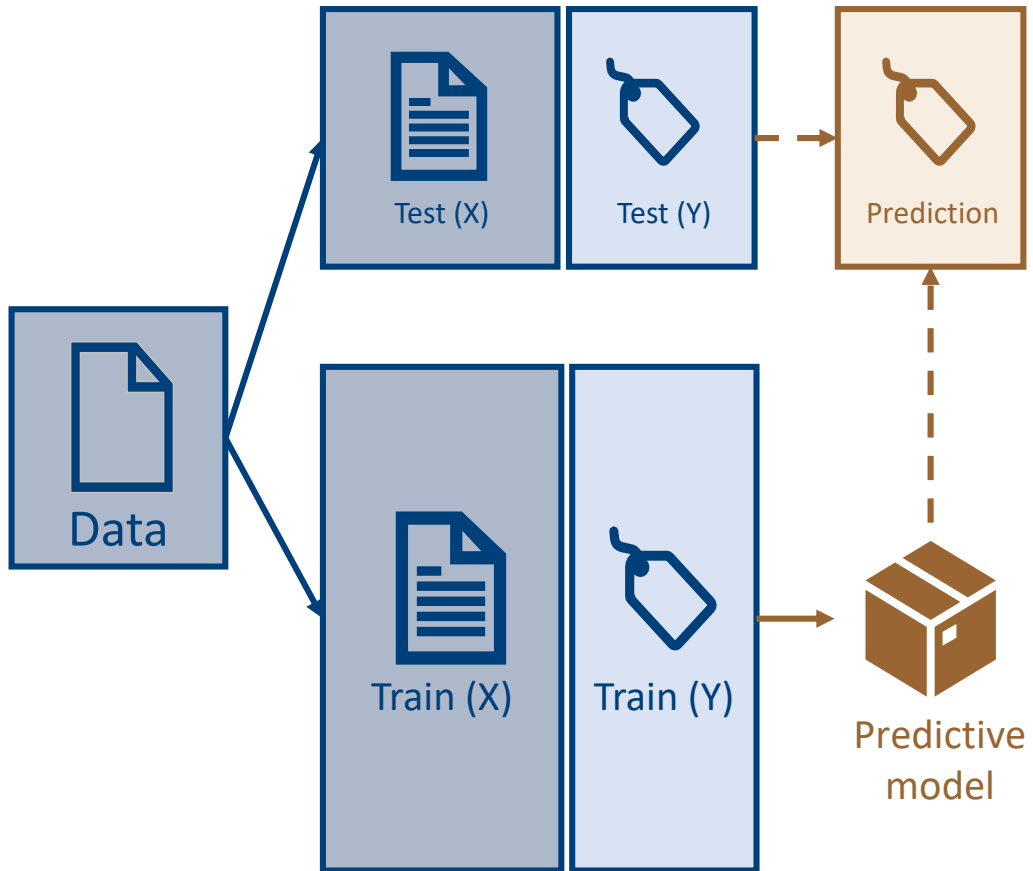
# Manifold Learning for Adversarial Robustness in Predictive Process Monitoring

Alexander Stevens<sup>1,\*</sup> , Jari Peeperkorn<sup>1</sup> , Johannes De Smedt<sup>1</sup> , Jochen De Weerd<sup>1</sup>

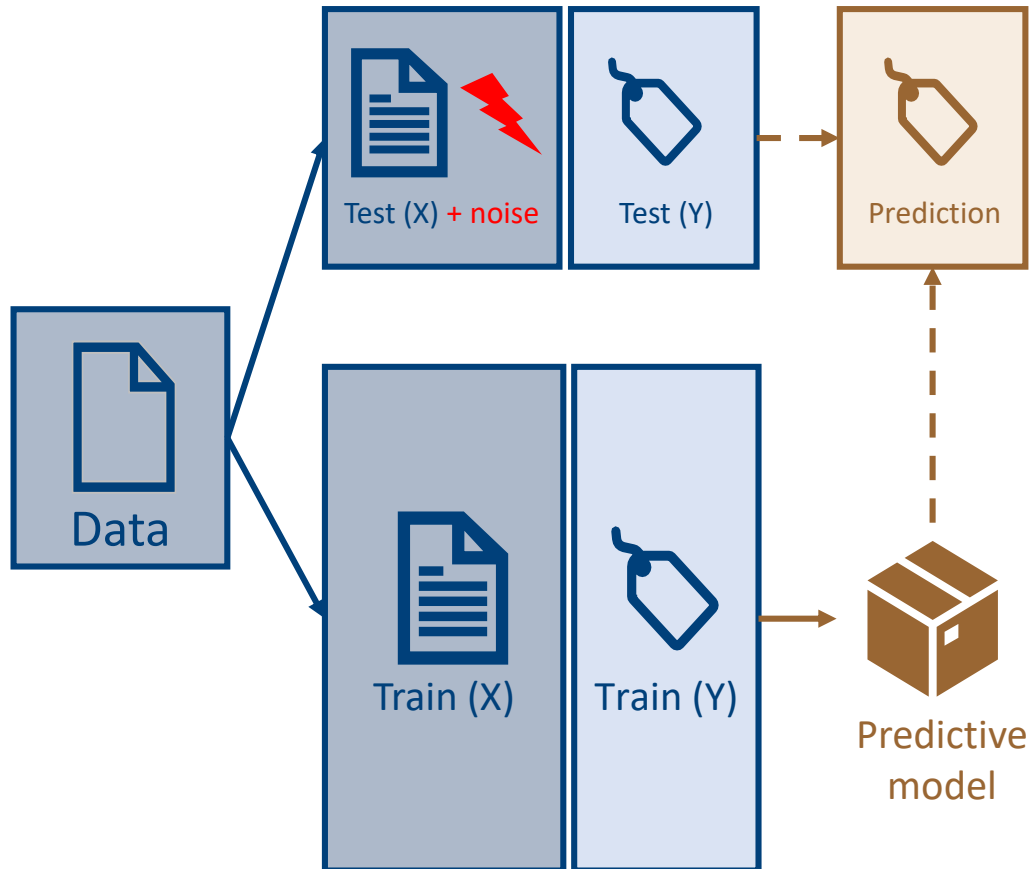
<sup>1</sup> Research Centre for Information Systems Engineering (LIRIS), KU Leuven (Belgium)

\* Corresponding author

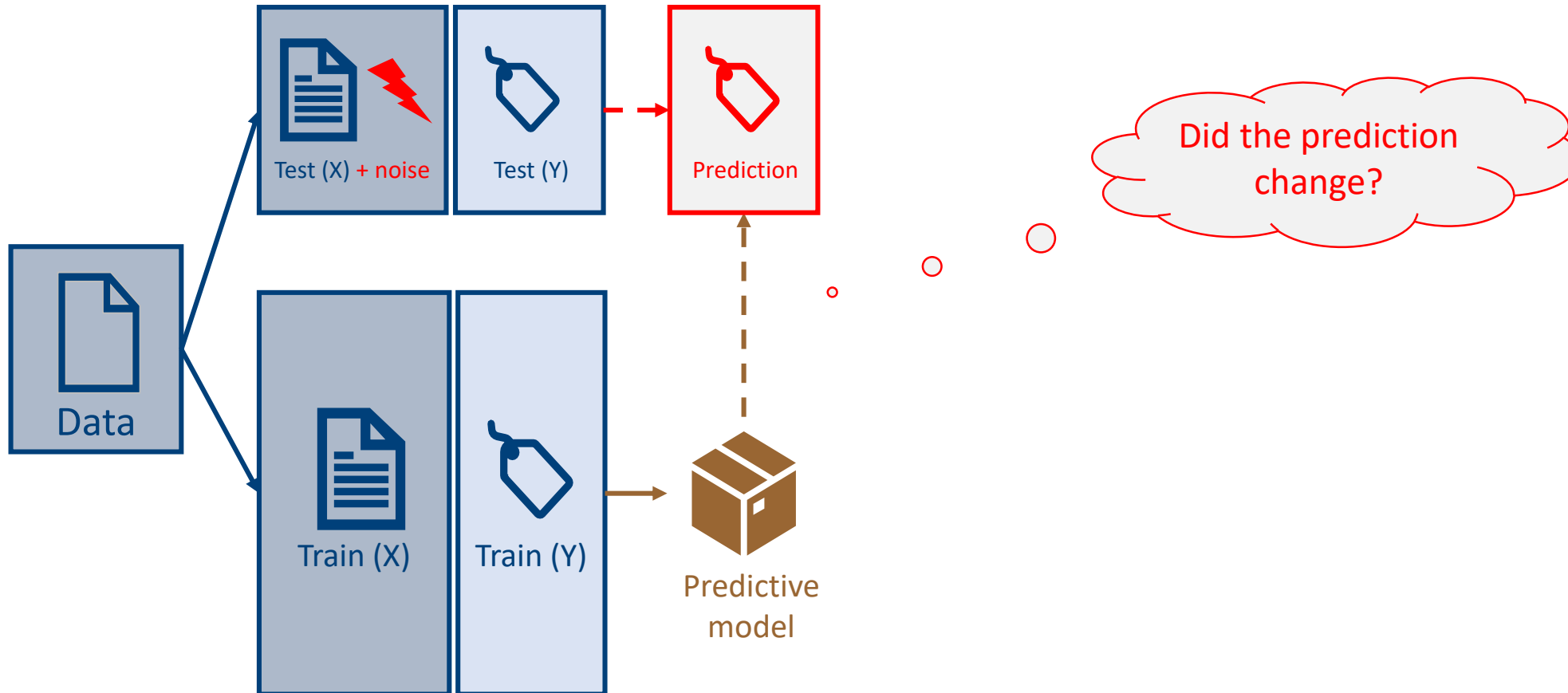
# Introduction to Machine Learning



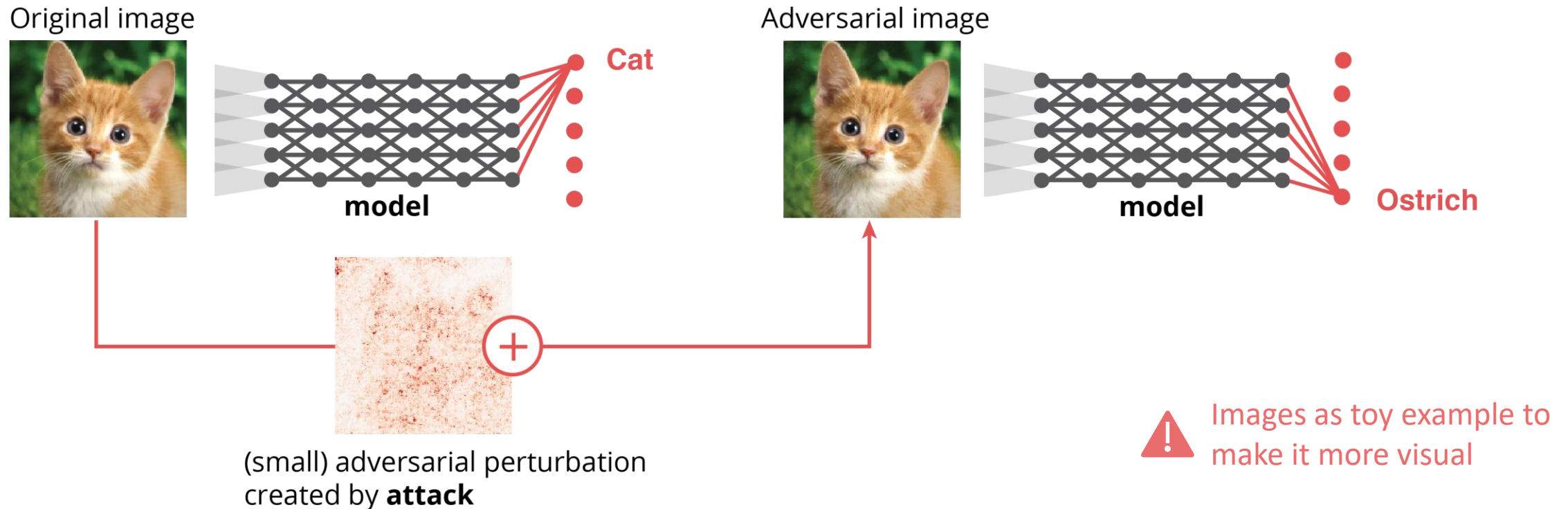
# Introduction to **Adversarial** Machine Learning



# Introduction to Adversarial Machine Learning



# Introduction to Adversarial Machine Learning

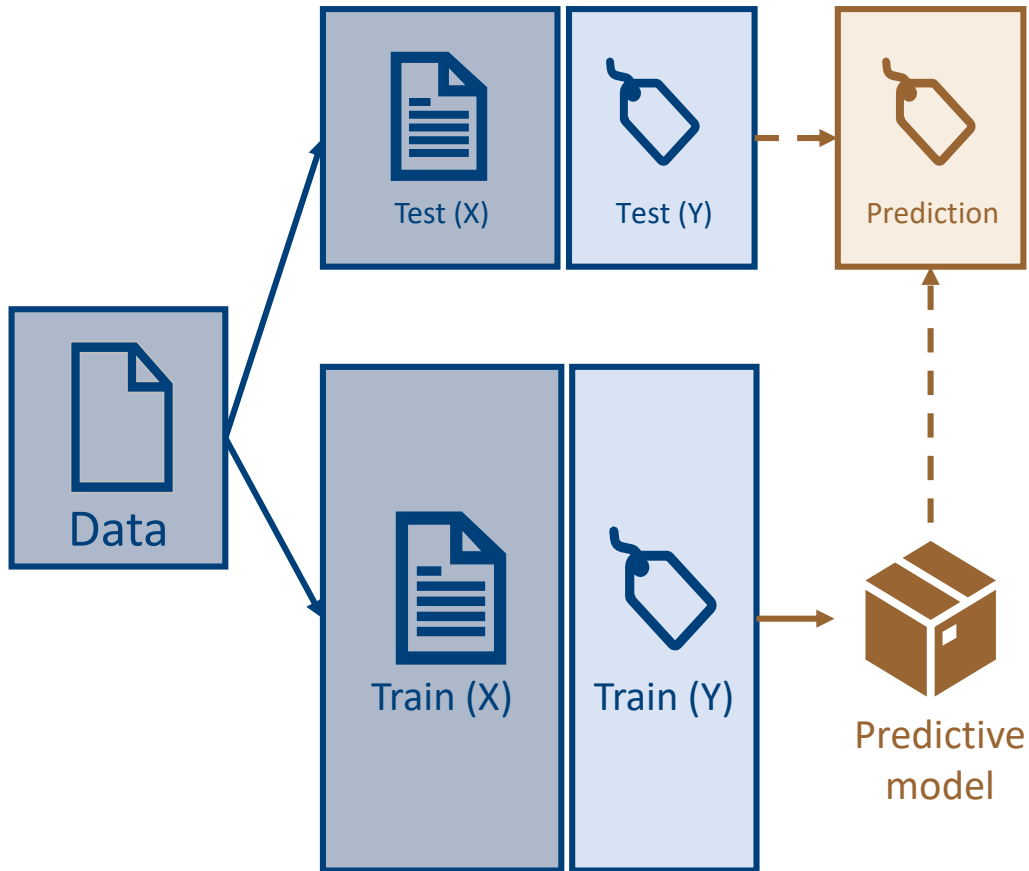


Small *perturbation* causes the model to make a false prediction”<sup>1,2</sup>

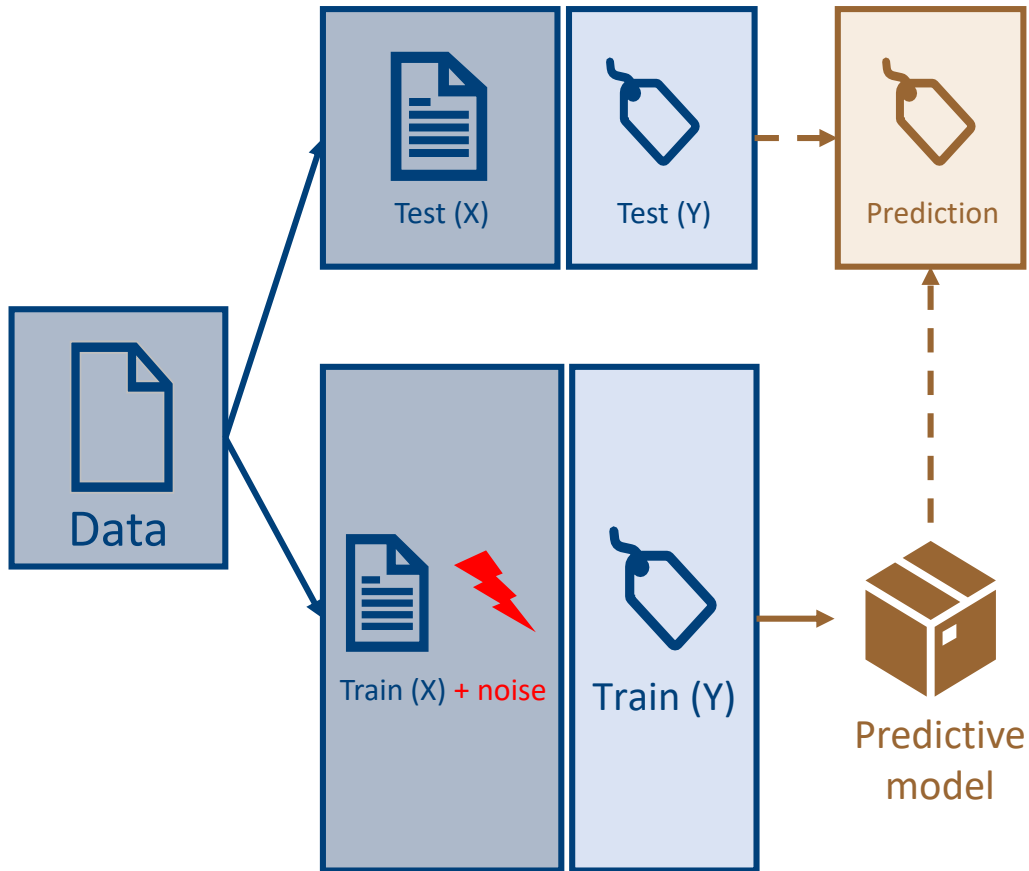
<sup>1</sup>Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). [christophm.github.io/interpretable-ml-book/](http://christophm.github.io/interpretable-ml-book/)

<sup>2</sup>Figure: NIPS 2018 Adversarial Vision Challenge

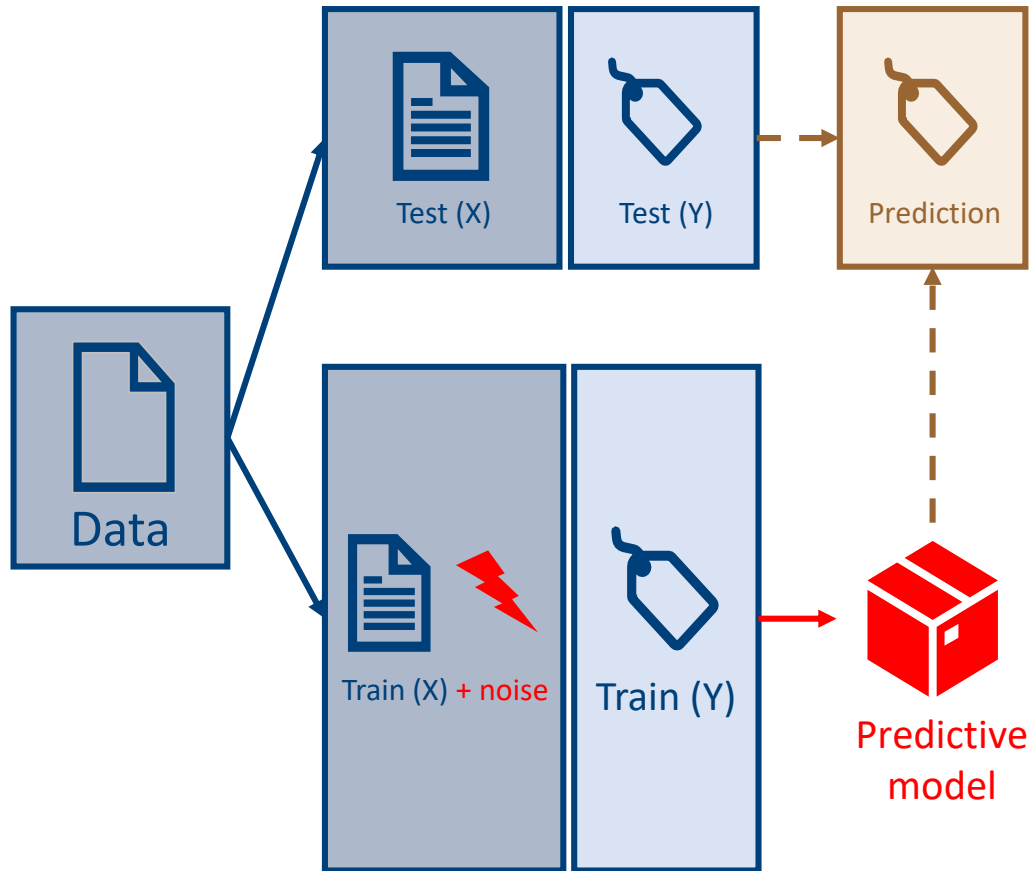
# Introduction to Adversarial Machine Learning



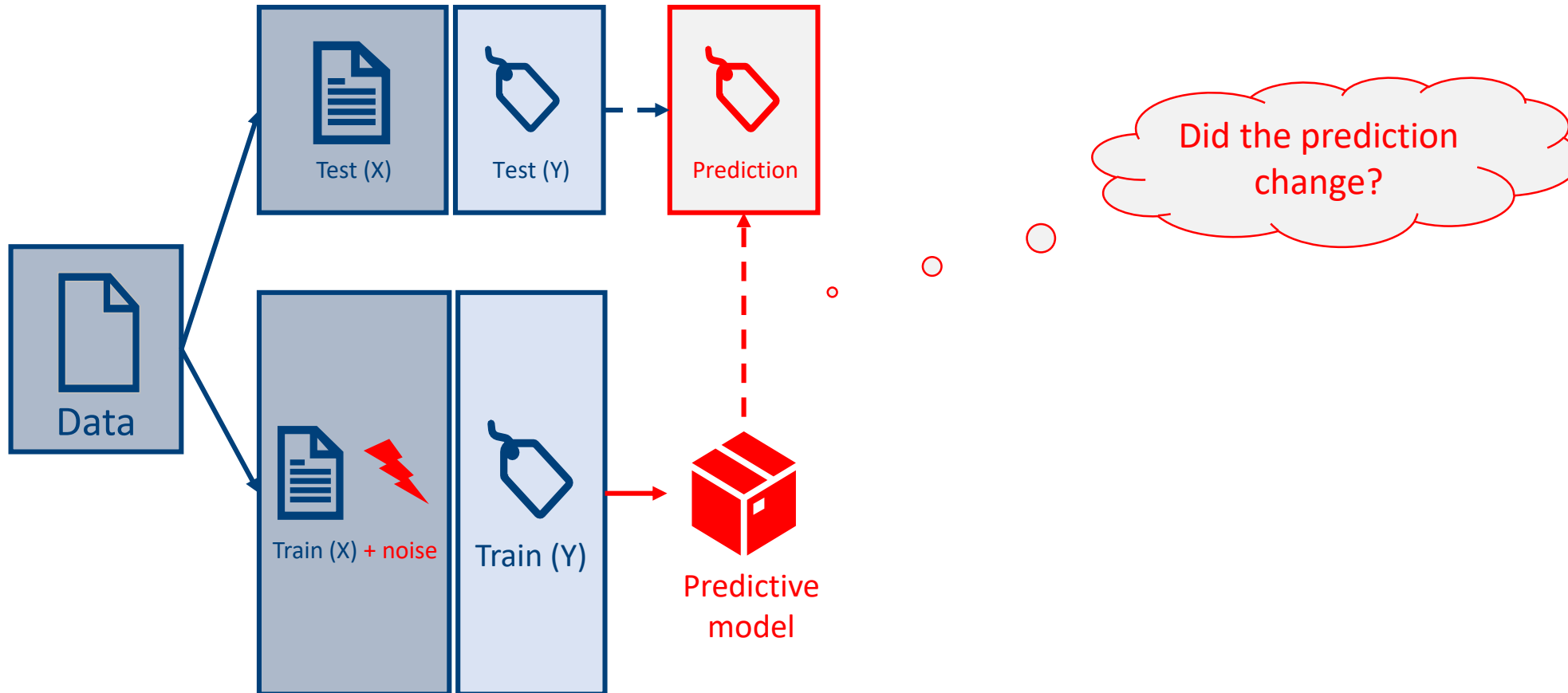
# Introduction to Adversarial Machine Learning



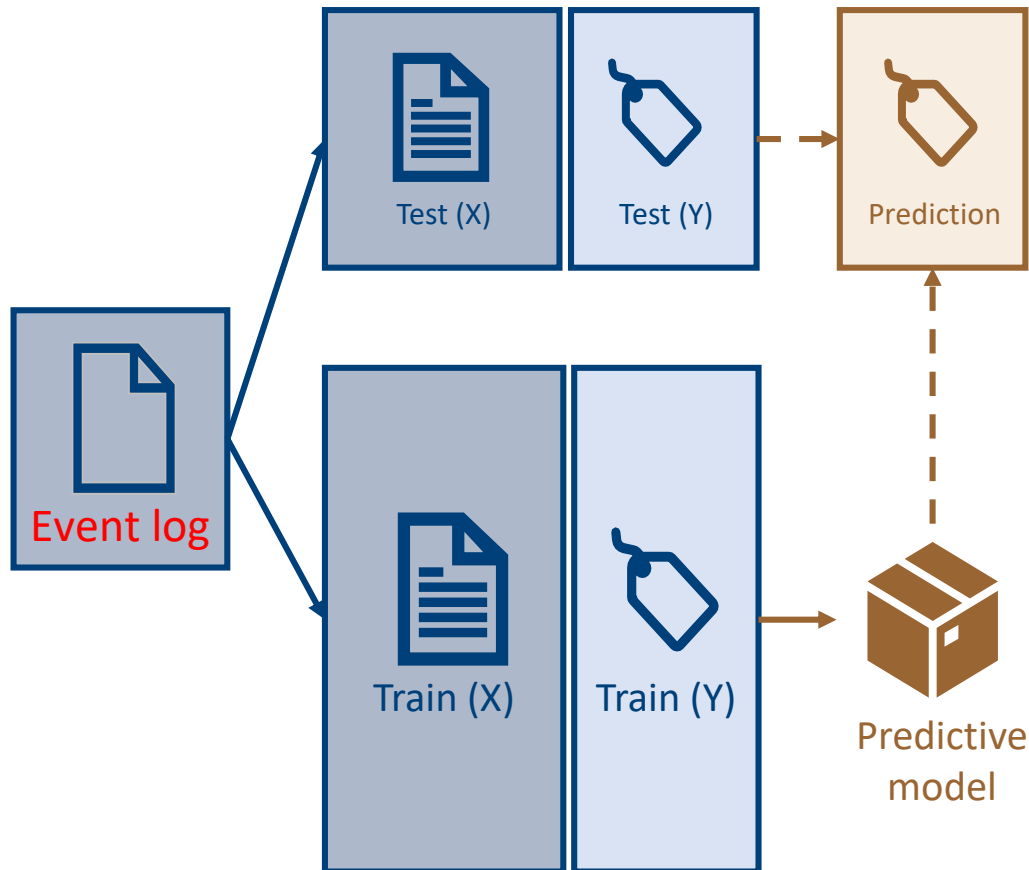
# Introduction to Adversarial Machine Learning



# Introduction to **Adversarial** Machine Learning



# Introduction to (Outcome-Oriented) Predictive Process Monitoring



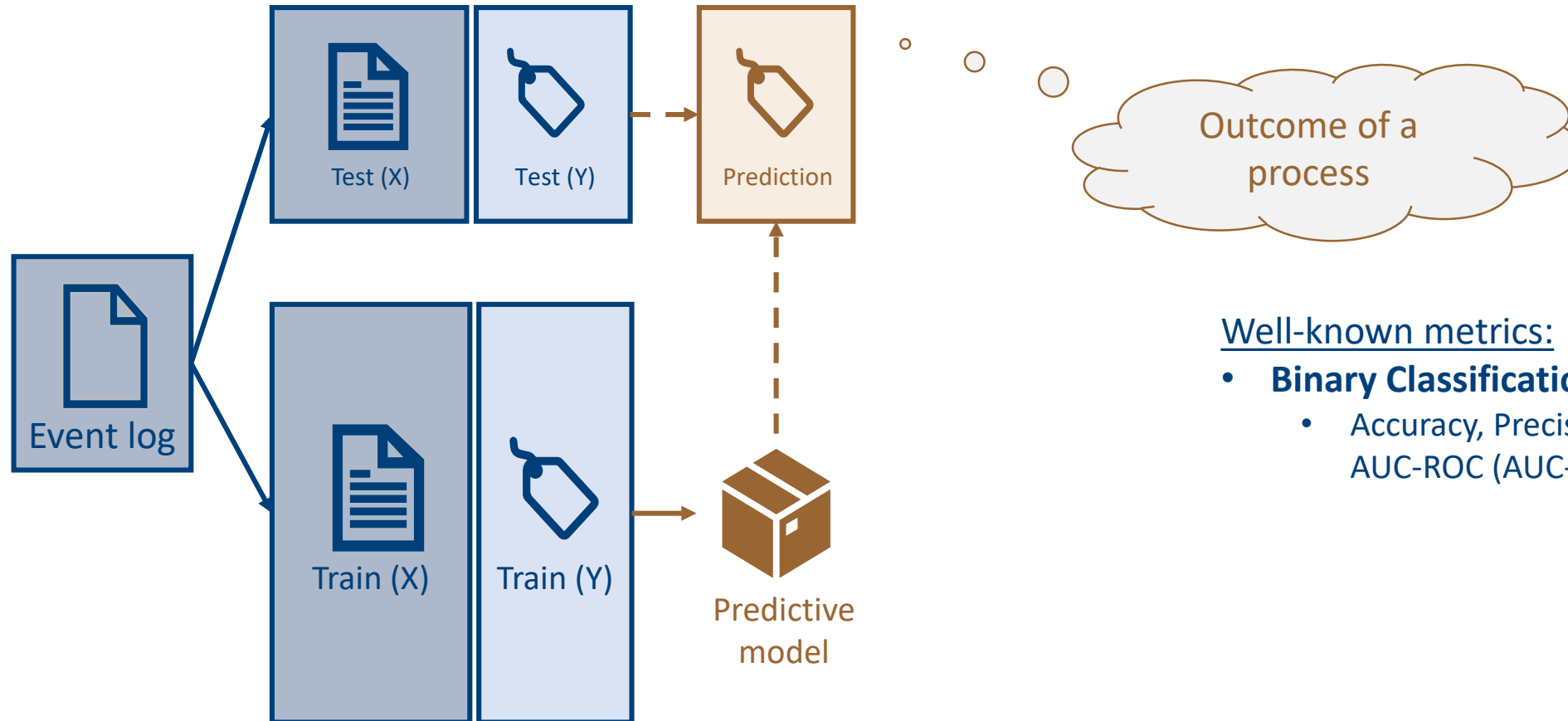
## Outcome-oriented predictive process monitoring

Process data (i.e. an event log) contains different cases

→ Each case has:

- A timestamped records of events
  - Activities
  - Other dynamic attributes
- A Case ID
- Static attributes

# Introduction to (Outcome-Oriented) Predictive Process Monitoring

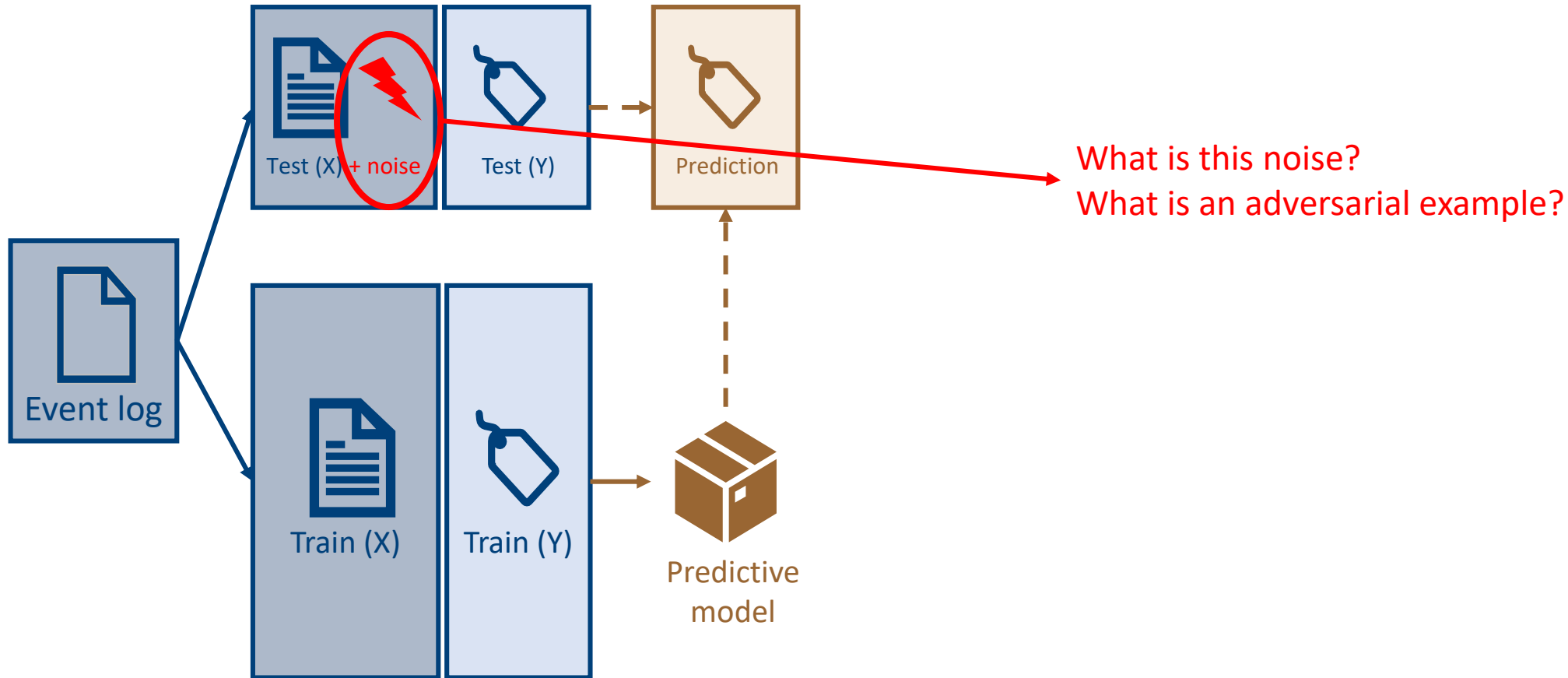


## Well-known metrics:

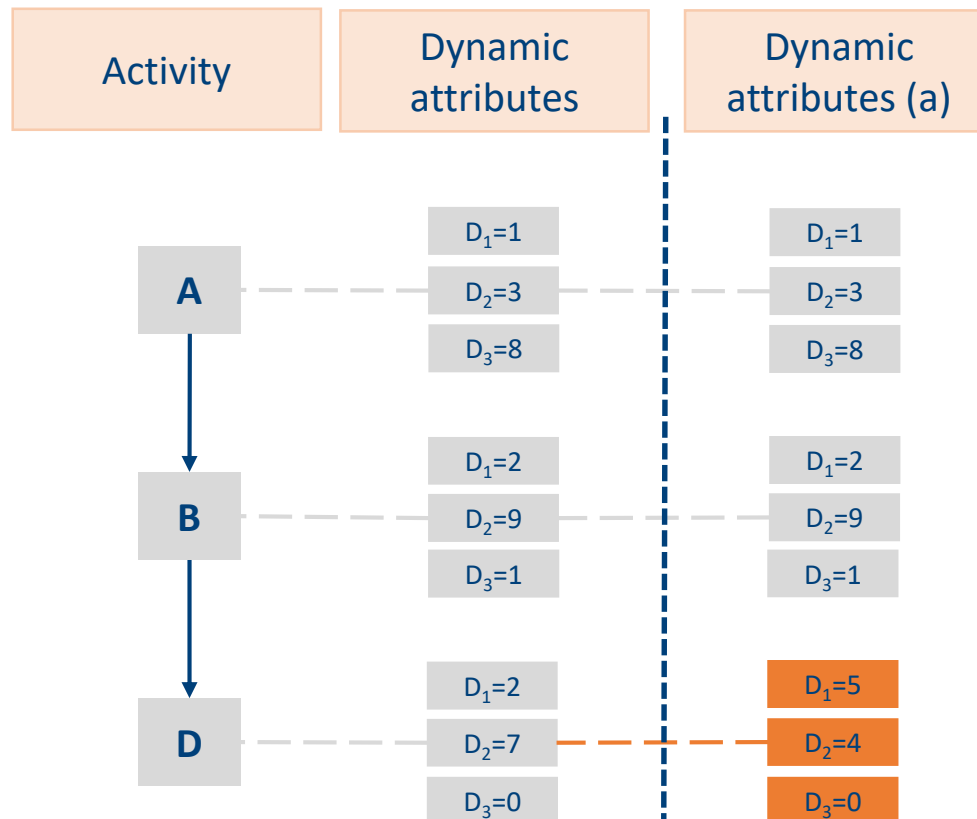
- **Binary Classification:**

- Accuracy, Precision, Recall, F1-score, AUC-ROC (AUC-PRC)

# Adversarial Machine Learning in Process Outcome Prediction



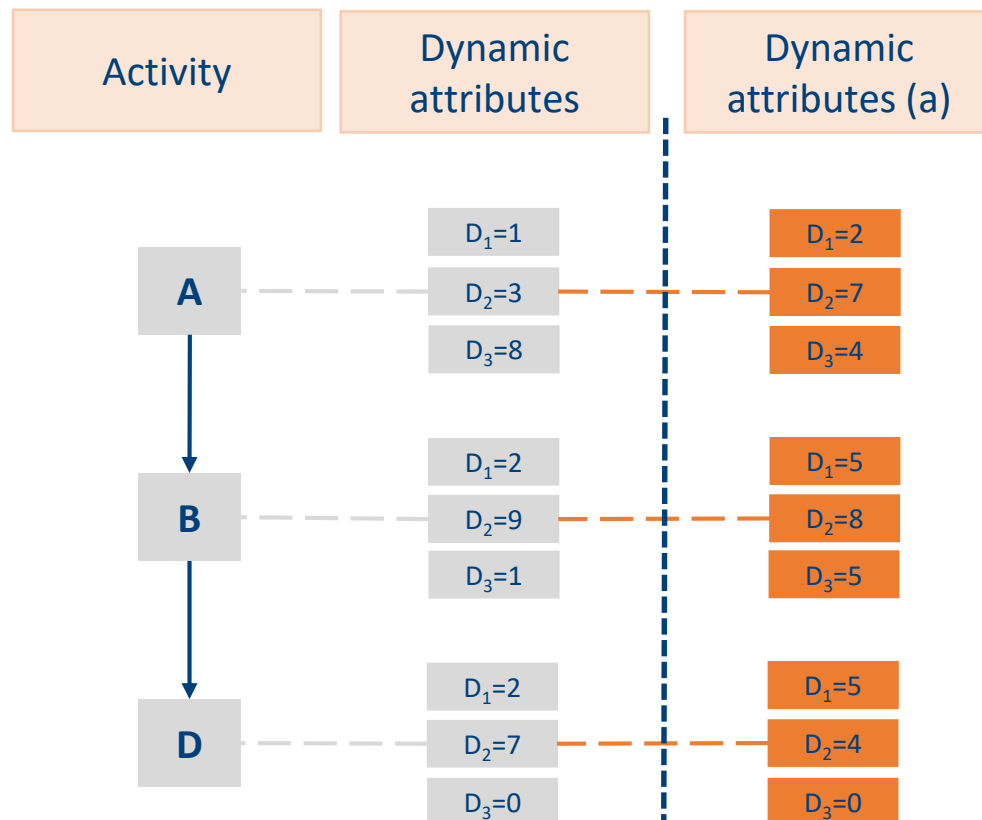
# What is this noise? What is an adversarial attack?



## Last Event Attack (A1)

- Permuting dynamic attribute of the last event of the prefix
- ✓ Intuitive
- ✓ Model is still able to learn correct behaviour of the attribute

# What is this noise? What is an adversarial attack?

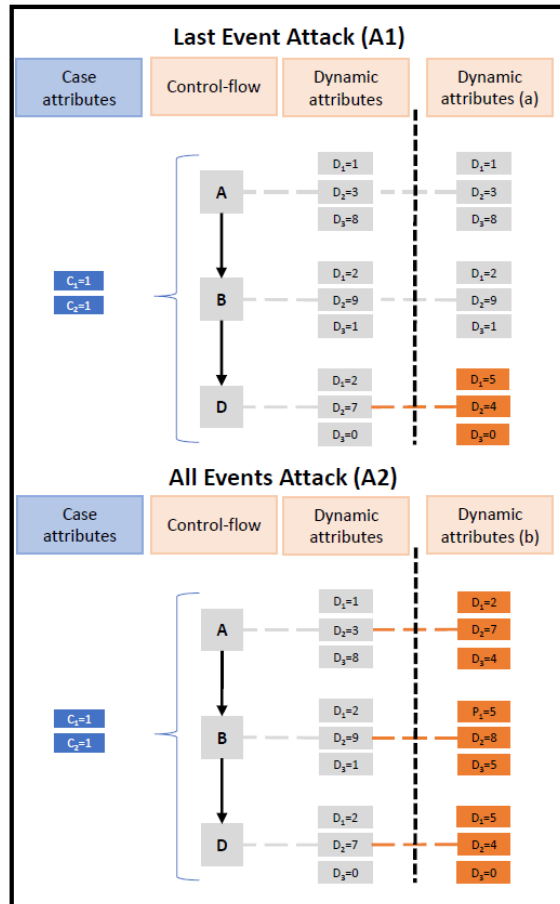


## All Event Attack (A2)

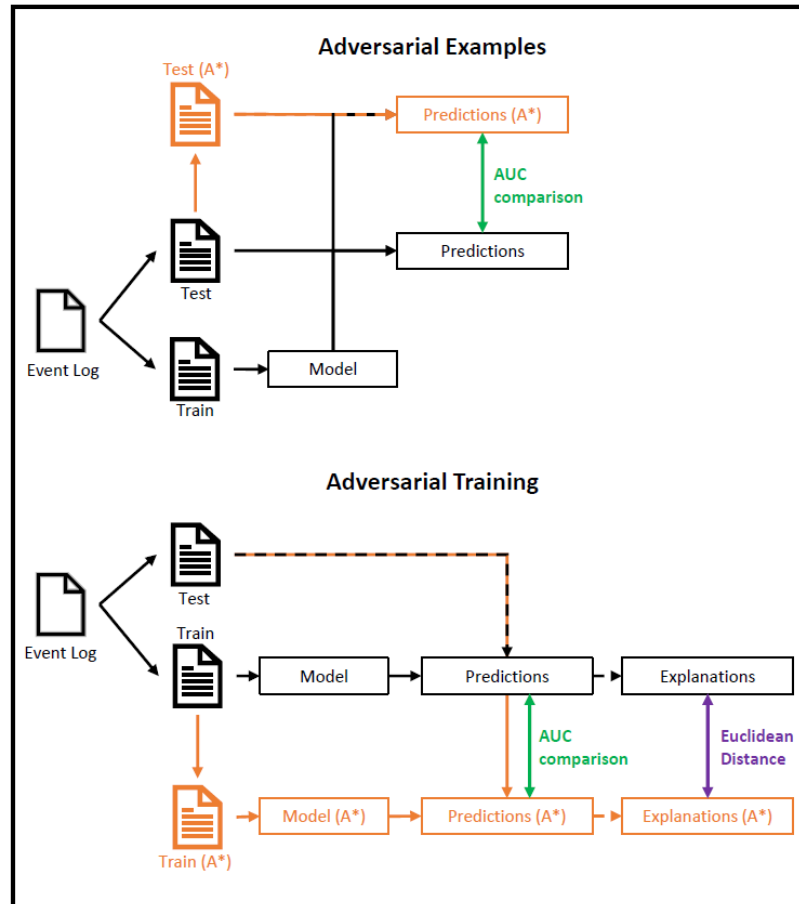
- Permuting dynamic attribute of all the events of the sequence
- ✗ Model is not able anymore to learn correct behaviour of attributes
- ✗ Boils down to pure noise attribute values

# Previous work

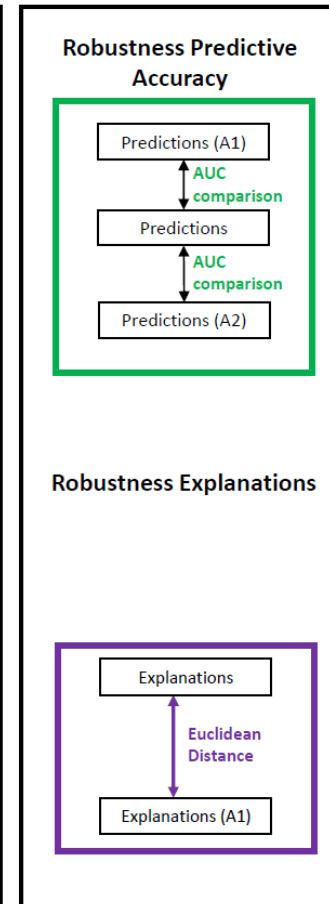
## Adversarial Attack



## Method of Application



## Evaluation



## Robustness Assessment Framework<sup>3</sup>

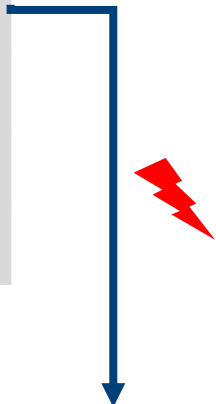
- ✓ 3 state-of-the-art POP models
- ✓ 2 different adversarial attacks
- ✓ 6 real-life event logs

<sup>3</sup>Stevens, A., De Smedt, J., Peepkorn, J., & De Weerd, J. (2022, October). Assessing the Robustness in Predictive Process Monitoring through Adversarial Attacks. In 2022 4th International Conference on Process Mining (ICPM) (pp. 56-63). IEEE.

# Limitations of previous work

- Random perturbations can be **unnatural**<sup>4</sup>

Height (cm)	Weight (kg)	BMI	Label
160	50	19.53	Healthy
175	85	27.76	Overweight
155	45	18.73	Healthy
185	95	27.76	Overweight



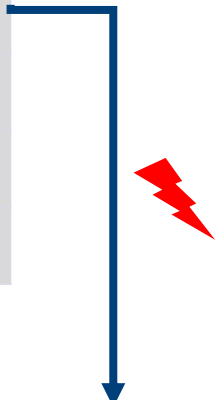
Height (cm)	Weight (kg)	BMI	Label
160	50	50	Healthy

BMI of 50 is still within range, but is not realistic (nor correct)

# Limitations of previous work

- Random perturbations can be **unnatural**<sup>4</sup>
- No guarantee that underlying label of the instance after the adversarial attack did not change

Height (cm)	Weight (kg)	BMI	Label
160	50	19.53	Healthy
175	85	27.76	Overweight
155	45	18.73	Healthy
185	95	27.76	Overweight



Height (cm)	Weight (kg)	BMI	Label
160	50	50	Overweight

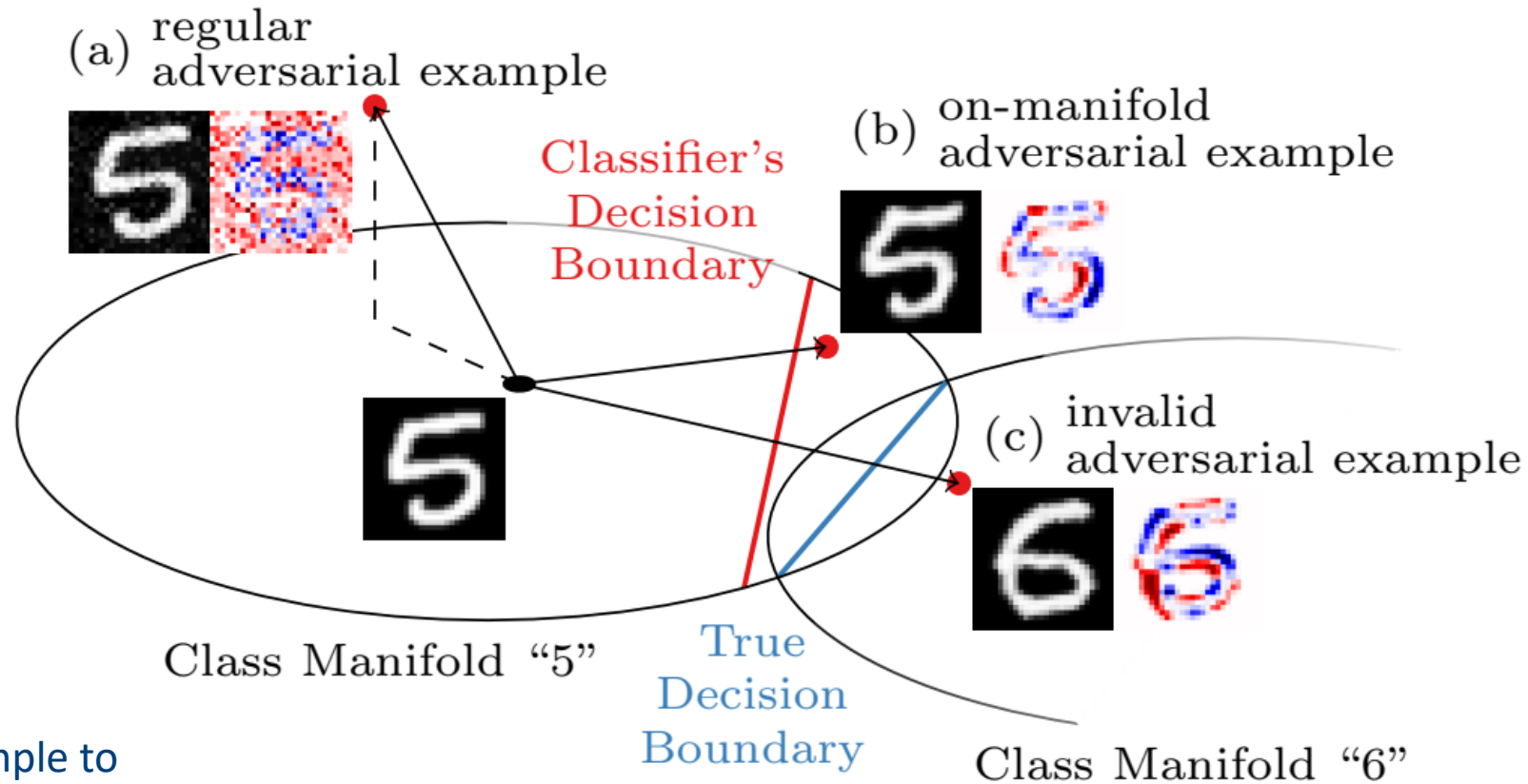
An BMI of 50 is classified as overweight

# Limitations of previous work

- Random perturbations can be **unnatural**<sup>4</sup>
- No guarantee that underlying label of the instance after the adversarial attack did not change
- **No defence mechanism against these adversarial attacks**
  - Only tested their inherent vulnerability against these attacks

# Introduction to **Manifold Learning**

regular adversarial examples vs. **natural** adversarial examples<sup>4</sup>

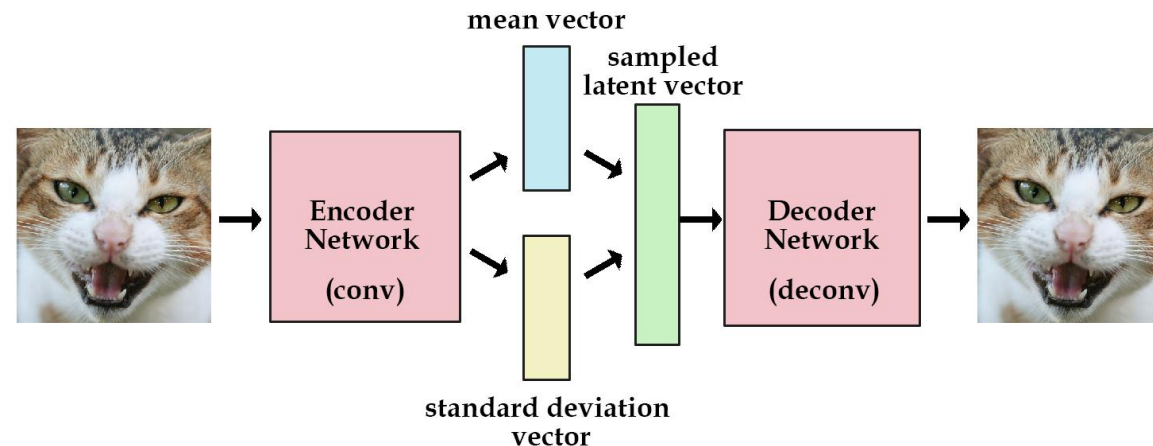


Images as toy example to make it more visual

<sup>4</sup>Stutz, D., Hein, M., & Schiele, B. (2019). Disentangling adversarial robustness and generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6976-6987).

# Introduction to **Manifold Learning**

- The adversarial examples should lie *within the distribution of the original data manifold* learned by an **LSTM Variational Autoencoder (VAE)**<sup>5</sup>
  - Auto-encoders encode data onto a lower dimensional latent space and decode them into the original sample
  - Variational autoencoders encode data into probability distributions → better for generation
  - LSTMs to deal with sequential character



<sup>5</sup>[https://wizardforcel.gitbooks.io/tensorflow-examples-aymericdamien/content/3.10\\_variational\\_autoencoder.html](https://wizardforcel.gitbooks.io/tensorflow-examples-aymericdamien/content/3.10_variational_autoencoder.html)

# Manifold Learning **Advantage**

- We project the adversarial example to the data manifold  
→ *natural*
- For both classes separately  
→ adhere to label invariance

# Adversarial Attacks on Manifold

- Because we adhere to label invariance
  - Attacks on the activity type
  - Attacks on resource attribute
  
- Successful attack
  - Original prediction was correct
  - Perturbed example is incorrectly predicted
  - Label is unchanged after perturbation

# Successful adversarial attacks



A successful adversarial example  $\tilde{x}$  is a perturbed version of a regular example  $x$  with label  $y$  such that:

## General definition

$$\tilde{x} = x + \varepsilon \approx x$$

perceptively indistinguishable instances

$$f(x) = y$$

the original prediction was correct

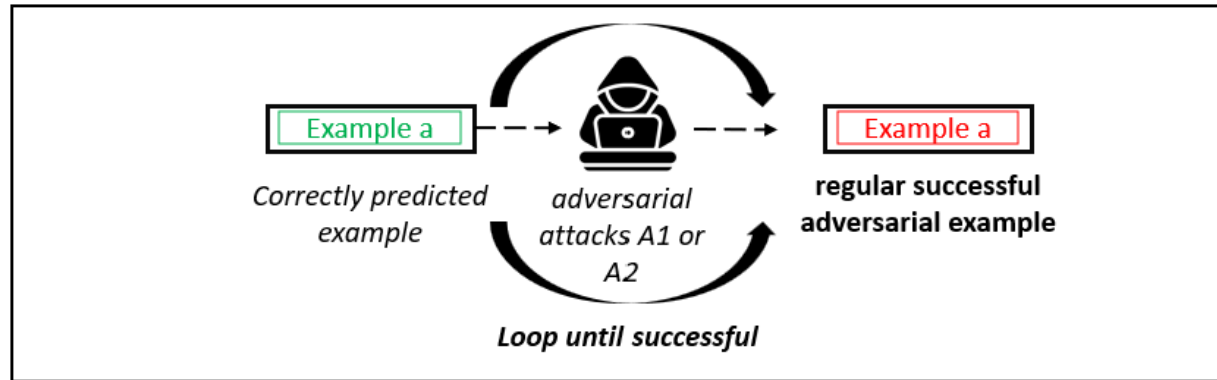
$$f(\tilde{x}) \neq y$$

perturbed example incorrectly predicted

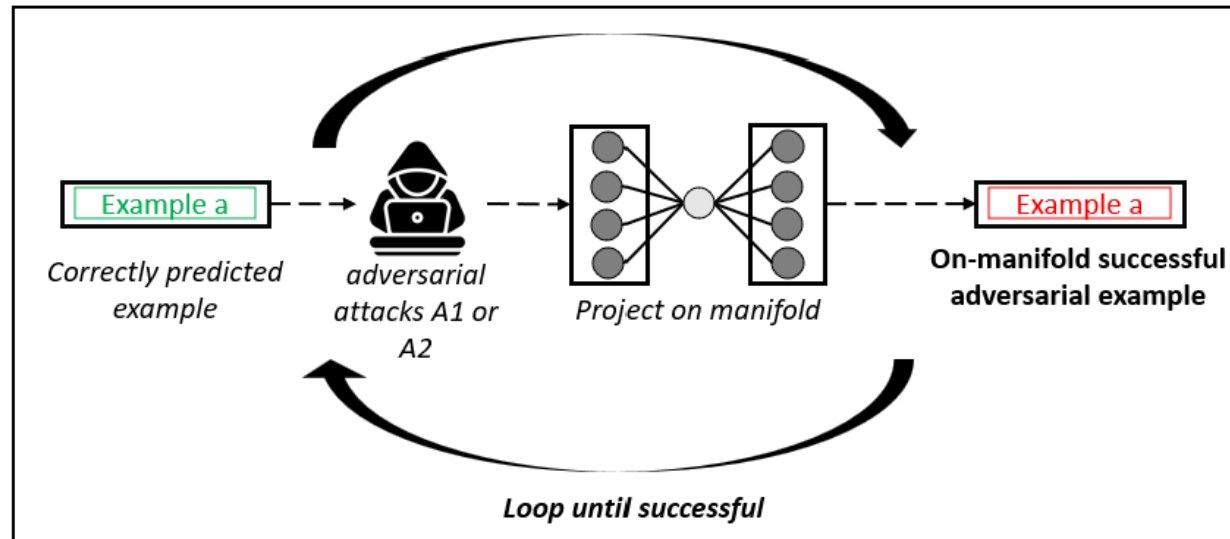
$$p(y|\tilde{x}) > p(y'|\tilde{x}) \forall y' \neq y.$$

label is unchanged after perturbations

# Manifold Learning for Adversarial Robustness in Predictive Process Monitoring



(a) Regular successful adversarial examples



(b) On-manifold successful adversarial examples

## Regular successful adversarial examples

1. Generate adversarial examples
2. Verify whether they are successful

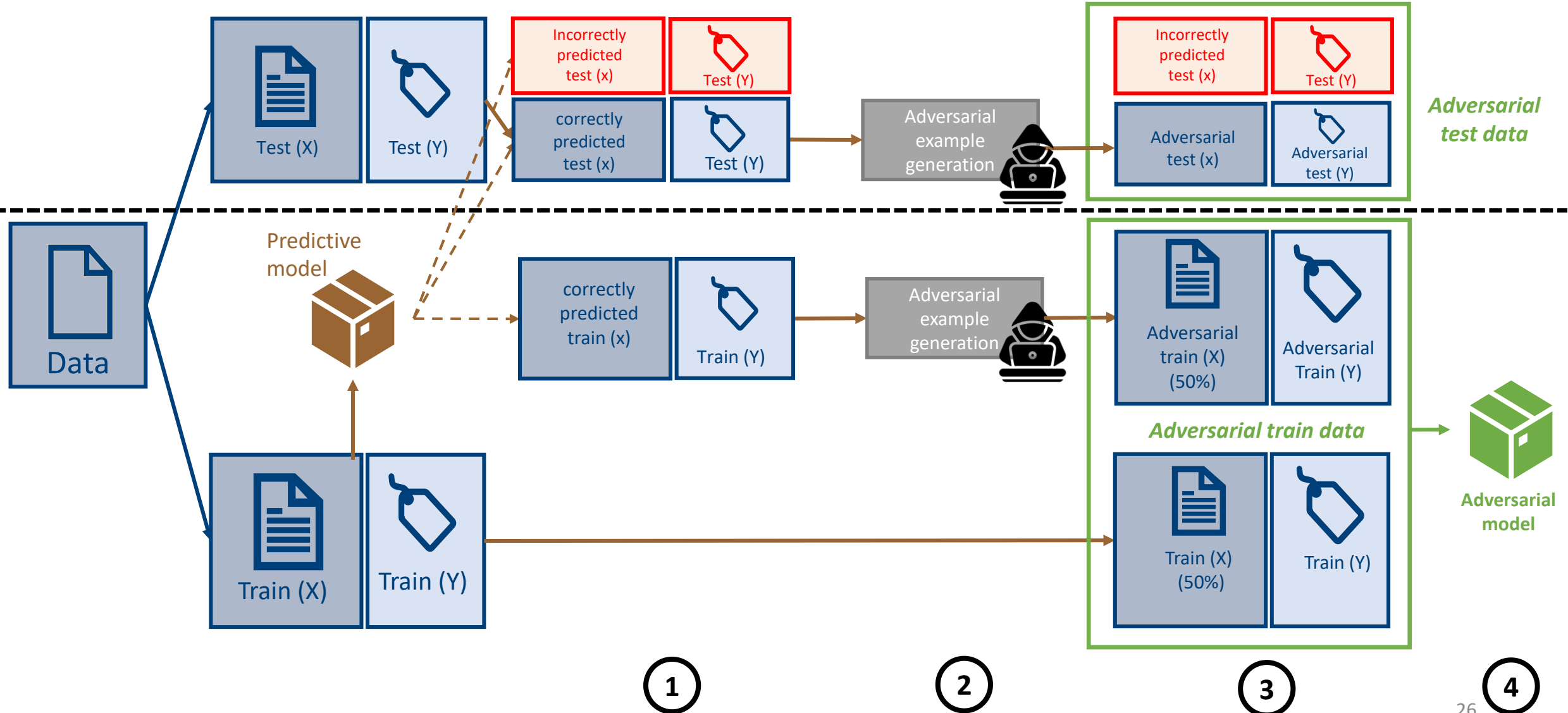
## On-manifold successful adversarial examples

1. Generate adversarial examples
2. Project the adversarial examples with a VAE to the manifold
3. Verify whether they are successful

# Types of Attacks

- Two different attacks
  - A1 only the last event of the prefix
  - A2 all events of the prefix
  
- On two different features
  - Activity type
  - Resource

# Manifold Learning for Adversarial Robustness in Predictive Process Monitoring



# Experimental Setup

- We tested 4 different types of predictive models
  - Logistic Regression
  - Random Forests
  - XGBoost
  - LSTM
- 5 different test sets
  - Original → predictive performance
  - A1 & A2; Activity & Resource **on manifold** → robustness against attacks
- 9 different training logs
  - Original
  - A1 & A2; Activity & Resource **simply permuted**
  - A1 & A2; Activity & Resource **on manifold**

# Results for Loan Application Process

Original test data and original model

Original test data and adversarial model

BPIC2012 (Accepted)		No Defense	Adversarial Training				On-manifold adversarial training			
			A1 (Act)	A1 (Res)	A2 (Act)	A2 (Res)	A1 (Act)	A1 (Res)	A2 (Act)	A2 (Res)
<b>LR</b>	No attack	<b>66.52</b>	56.33	58.83	55.62	62.62	60.86	60.93	62.77	62.83
	Attack (manifold) A1 (Act)	0.0	72.55	11.10	67.63	11.88	86.41	86.11	84.47	85.89
	A1 (Res)	0.0	74.17	10.91	68.94	11.37	86.86	86.69	85.13	86.38
	A2 (Act)	0.0	74.16	7.97	65.76	11.28	82.58	82.02	84.48	81.46
	A2 (Res)	0.0	70.55	17.87	67.68	19.24	86.72	86.76	83.36	90.15
<b>RF</b>	No attack	64.17	60.27	60.3	63.97	64.01	64.32	64.53	63.96	63.22
	Attack (manifold) A1 (Act)	0.0	19.33	20.39	30.52	7.98	82.78	82.60	75.57	76.06
	A1 (Res)	0.0	19.26	23.65	29.27	8.78	82.35	82.20	74.48	75.82
	A2 (Act)	0.0	34.74	21.59	47.84	23.66	80.53	79.90	83.71	81.49
	A2 (Res)	0.0	23.09	36.69	26.95	35.72	84.34	84.14	84.29	85.47
<b>XGB</b>	No attack	63.77	60.94	60.97	63.75	62.83	64.24	64.34	64.77	64.05
	Attack (manifold) A1 (Act)	0.0	29.68	30.00	24.54	11.33	87.97	87.95	83.04	83.62
	A1 (Res)	0.0	28.93	32.54	25.21	11.87	88.02	88.03	82.78	84.27
	A2 (Act)	0.0	50.51	26.76	41.86	18.08	82.20	82.05	85.69	84.05
	A2 (Res)	0.0	31.77	34.11	27.25	36.18	85.08	85.11	85.71	86.07
<b>LSTM</b>	No attack	60.05	59.36	61.95	61.07	58.36	61.89	62.36	60.83	61.49
	Attack (manifold) A1 (Act)	0.0	61.74	26.36	50.31	32.29	85.23	85.22	83.20	80.89
	A1 (Res)	0.0	60.06	26.23	49.56	31.87	83.85	83.87	81.53	79.76
	A2 (Act)	0.0	58.08	33.43	57.01	48.31	83.54	83.49	85.35	84.21
	A2 (Res)	0.0	61.10	29.76	53.86	52.88	87.27	87.29	87.34	87.19

Adversarial test data and original model

Adversarial test data and adversarial model

Original test data and on-manifold adversarial model

Adversarial test data and on-manifold adversarial model

# Conclusion

- The worst-case scenarios (A1 and A2 successful adversarial attacks) show that the models can theoretically be extremely incompetent
- Manifold learning allows for more natural adversarial attacks and overcomes the label invariance assumption
- On-manifold adversarial training works as a defence mechanism
- On-manifold adversarial training is still accurate on unseen, new test data

# Future Work

- Explore more diverse attack scenarios and adversarial training techniques
- Test possibilities of the autoencoders and manifolds
  - Counterfactual explanation generation
  - Clustering
  - Calculating overlap to compare classes/logs

# Appendix A: Reference List

[1] Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

[2] *Figure: NIPS 2018 Adversarial Vision Challenge*

[3] Stevens, A., De Smedt, J., Peeperkorn, J., & De Weerd, J. (2022, October). Assessing the Robustness in Predictive Process Monitoring through Adversarial Attacks. In *2022 4th International Conference on Process Mining (ICPM)* (pp. 56-63). IEEE.

[4] Stutz, D., Hein, M., & Schiele, B. (2019). Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6976-6987).

[5] [https://wizardforcel.gitbooks.io/tensorflow-examples-aymericdamien/content/3.10\\_variational\\_autoencoder.html](https://wizardforcel.gitbooks.io/tensorflow-examples-aymericdamien/content/3.10_variational_autoencoder.html)



### Research interests:

- Trustworthy AI:
  - Explainable AI (Metrics), Counterfactuals
  - Fairness, Bias Mitigation
  - Robustness, (Variational) Autoencoders

Thank you for your attention!



Alexander Paul Stevens



Alexander Stevens



<https://alexanderpaulstevens.github.io/>