KU LEUVEN

Leveraging manifold learning for adversarial attacks & counterfactuals in process outcome prediction

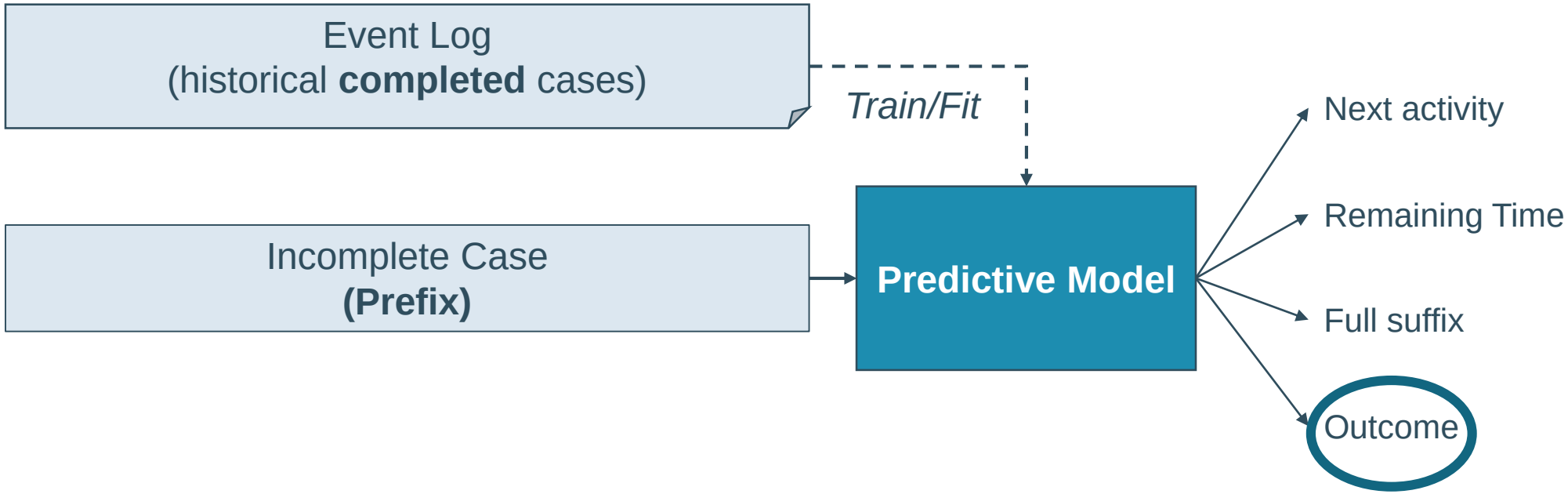
EURO 2024

Alexander Stevens, **Jari Peeperkorn**, et al.

Research Center for Information Systems Engineering (LIRIS), Faculty of Economics and Business

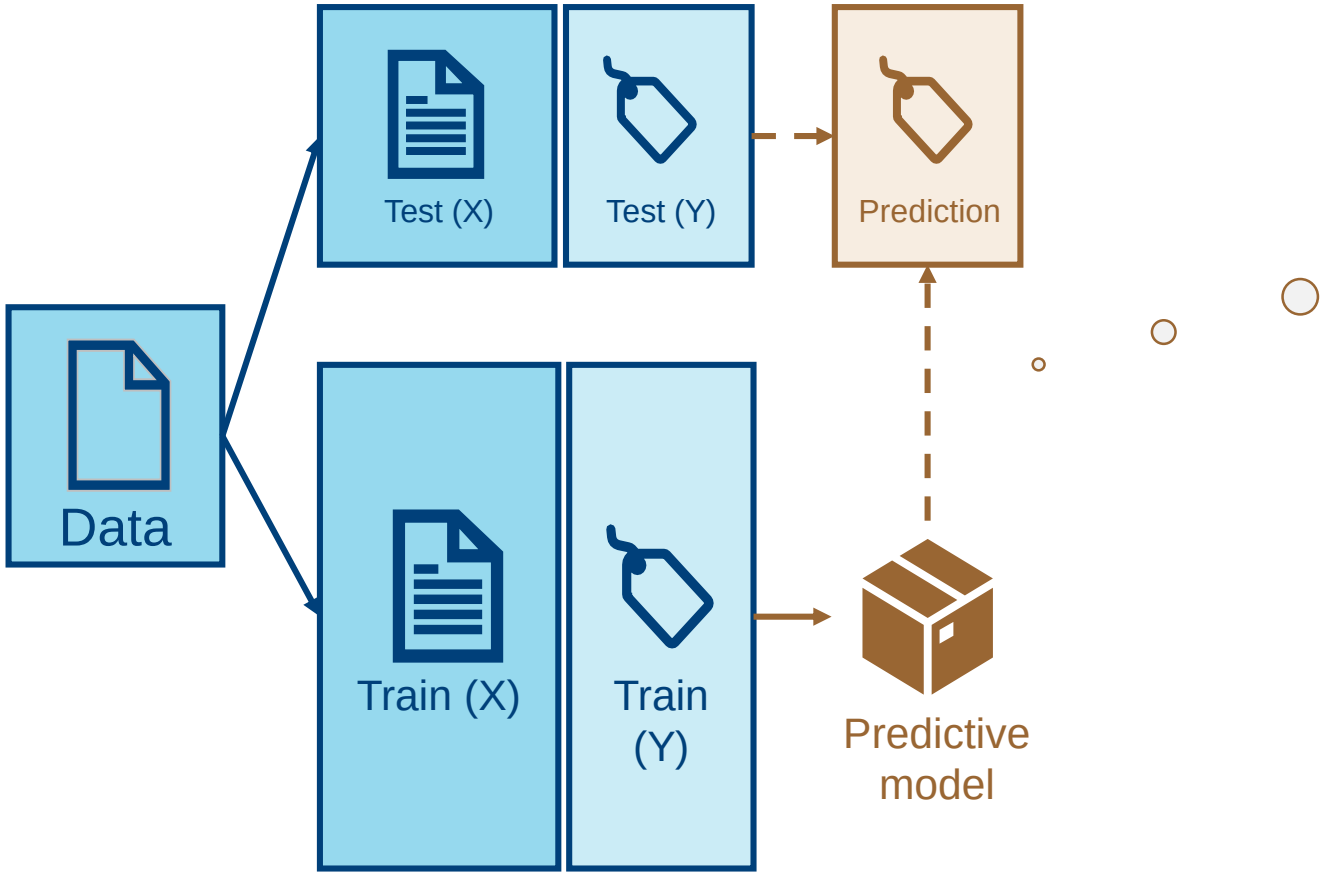
Predictive Process Monitoring

→ Forecast future elements of ongoing cases



→ Often Machine Learning Models

Explainability in predictive process monitoring

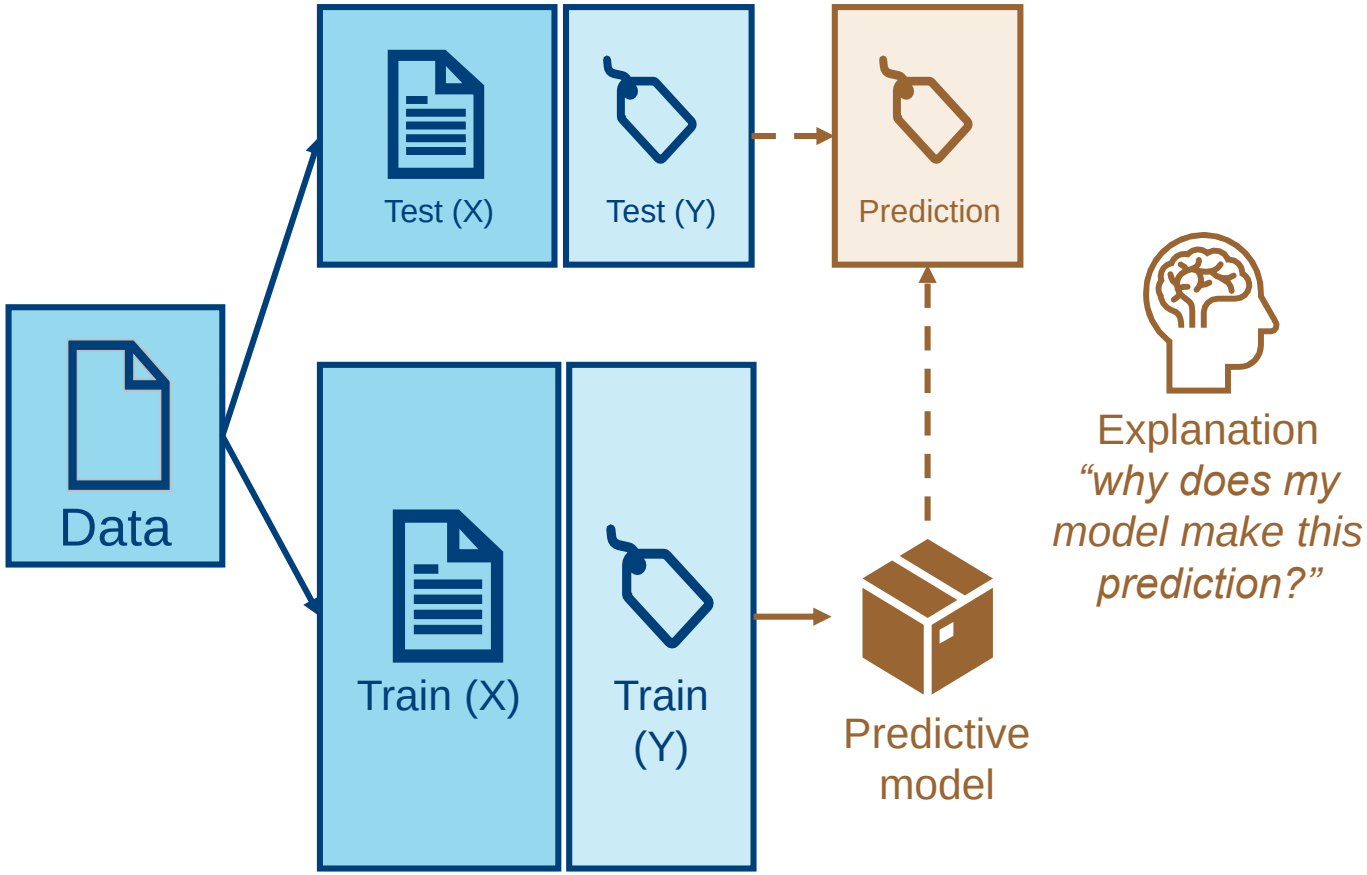


Evaluate results

Well-known metrics:

- **Binary Classification:**
  - Accuracy, Precision, Recall, F1-score, AUC-ROC (AUC-PRC)
- …

Explainability in predictive process monitoring



- **Transparent models**
  - Logistic/linear regression
  - Decision trees
- **Post-hoc methods**
  - Permutation importance
  - Shapley plots
  - Local models (LIME, …)
  - …

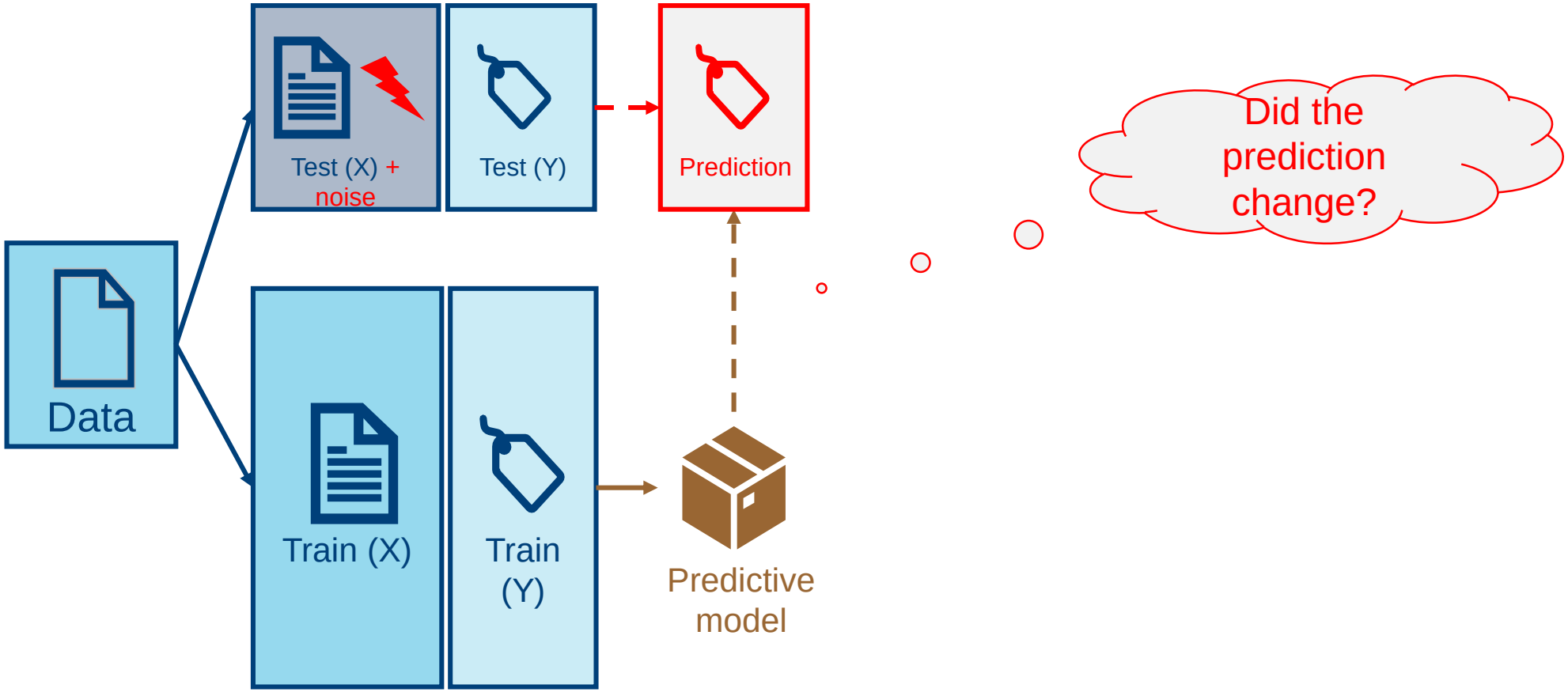
Robustness

- Needed to build *trust*
- How robust are the predictions against noise?
- Adversarial attacks for process event data

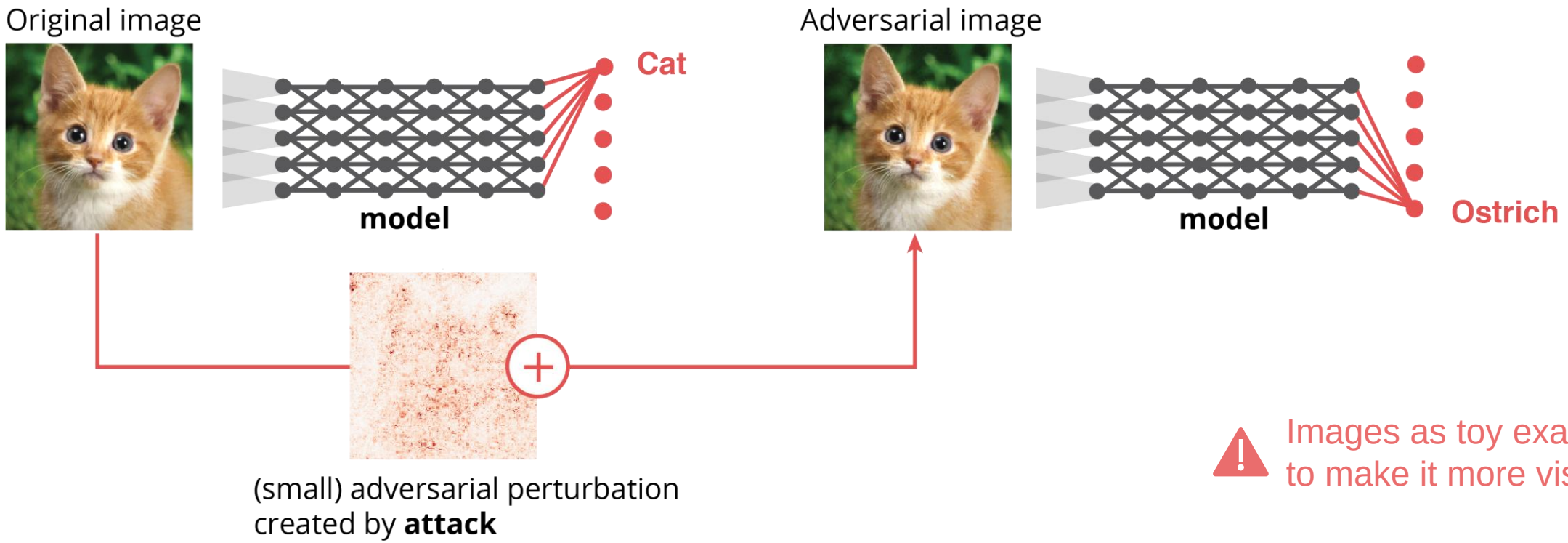
Alexander Stevens, Jari Peeperkorn, Johannes De Smedt , Jochen De Weerdt. Assessing the Robustness in Predictive Process Monitoring through Adversarial Attacks. ICPM (2022)

Alexander Stevens, Jari Peeperkorn, Johannes De Smedt , Jochen De Weerdt. Manifold Learning for Adversarial Robustness in Predictive Process Monitoring. ICPM (2023)

Adversarial attacks



Adversarial attacks

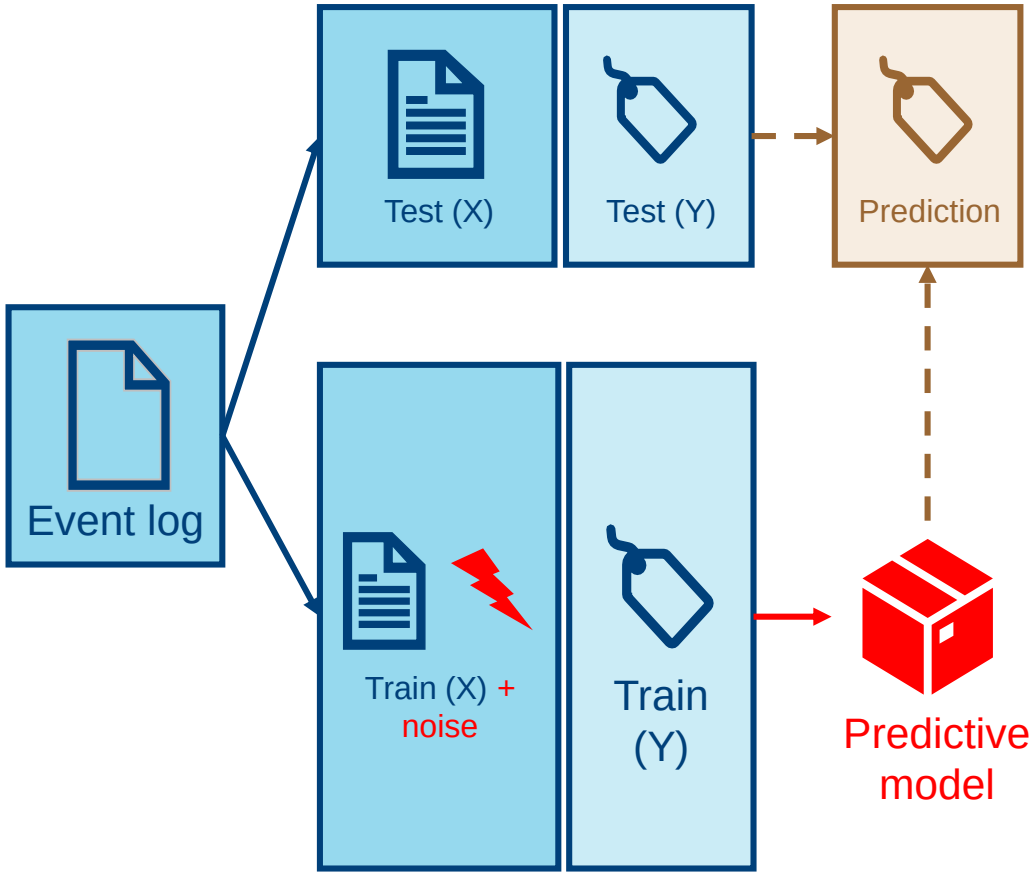


Images as toy example to make it more visual

Small *perturbation* causes the model to make a false prediction"[1,2]

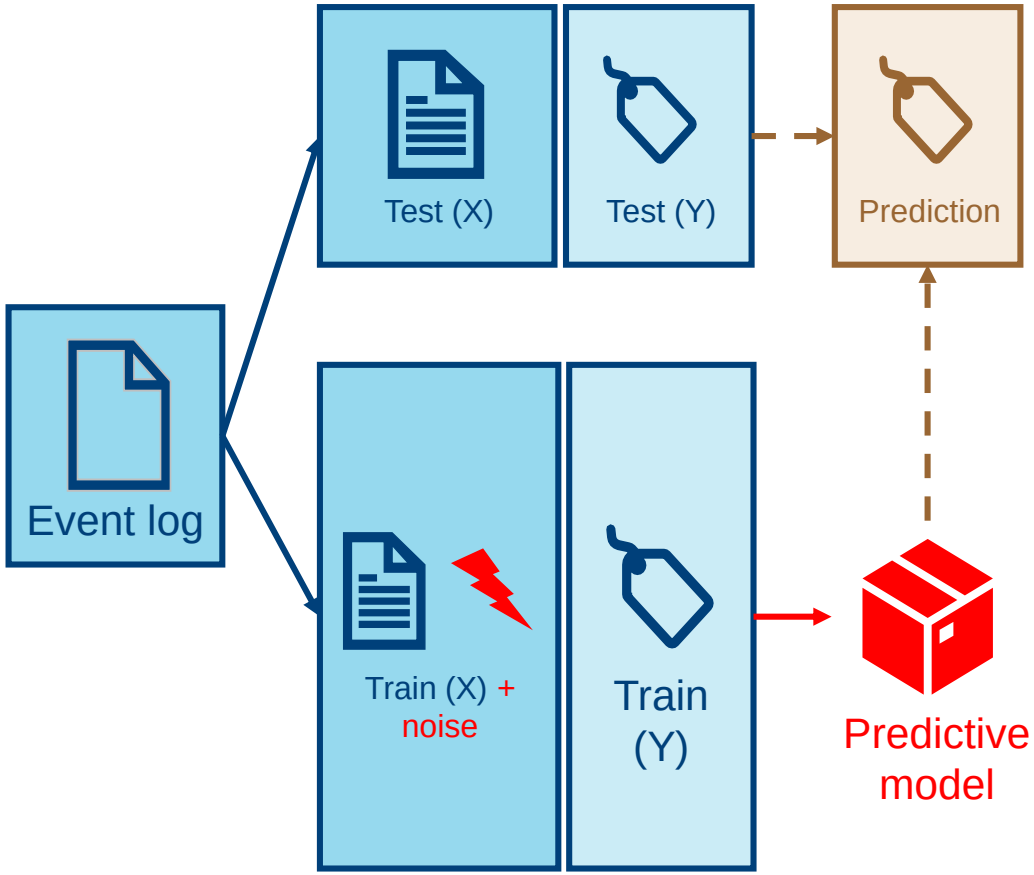
[1]*Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/*
[2]*Figure: NIPS 2018 Adversarial Vision Challenge*

Adversarial training



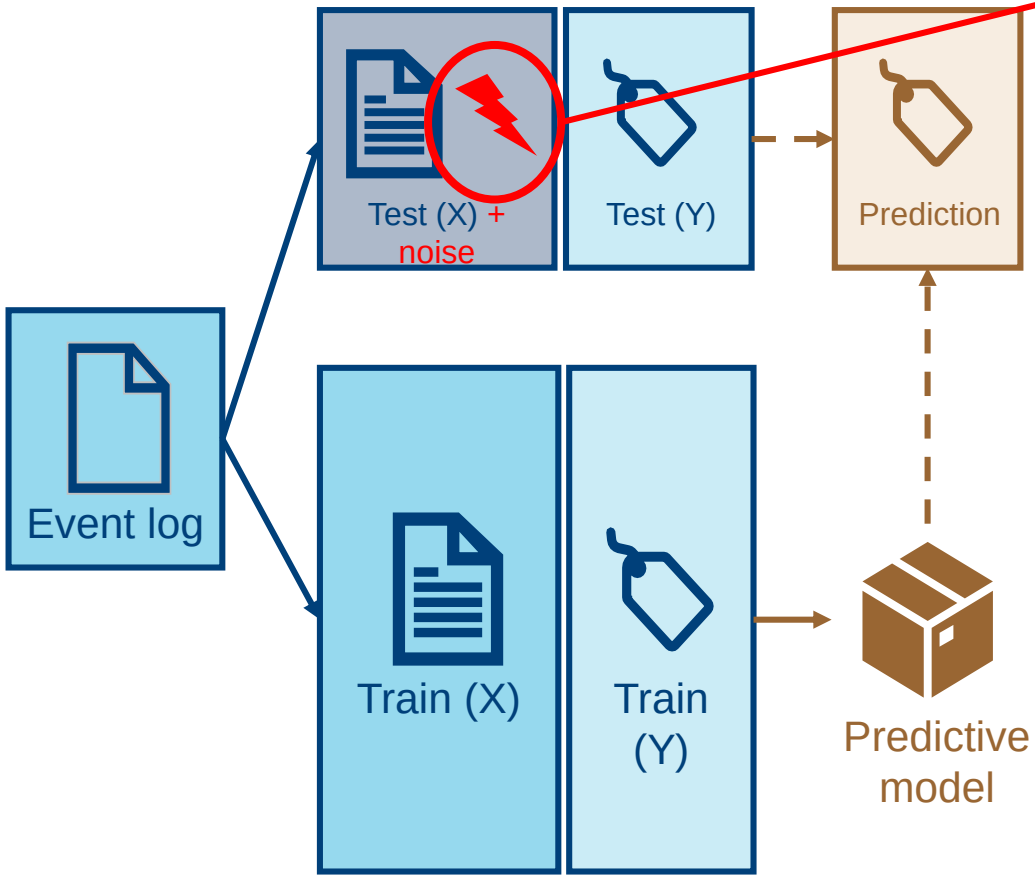
- Did prediction change?
- **Did the explanation change?**

Adversarial training



- Did prediction change?
- Did the explanation change?
- **Use to enhance training set**
  → increase robustness!
  “Adversarial training as a proactive defense mechanism”

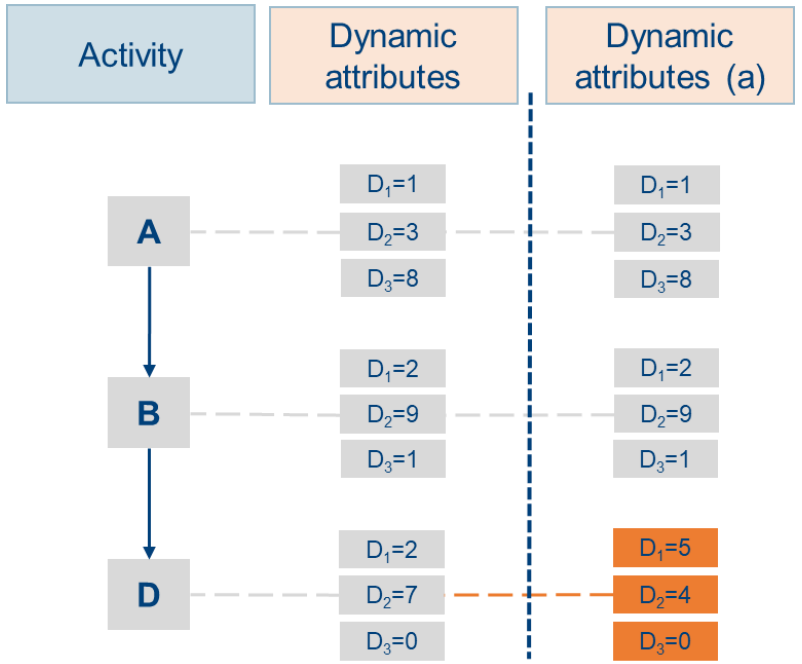
Adversarial attacks



What is this noise?
What is an adversarial example?

- Replacing last event attributes with noise
- Replacing all events with noise

→ not ideal



Alexander Stevens, Jari Peeperkorn, Johannes De Smedt , Jochen De Weerdt. Assessing the Robustness in Predictive Process Monitoring through Adversarial Attacks. ICPM (2022)

Adversarial attacks

- Adding just noise as attacks is not necessarily a smart thing to do
  - Our induced noise can be unnatural
    → values that can never happen or are very unlikely in that situation
  - No guarantee that the underlying label of the instance after the adversarial attack did not change
    → Because a certain value changed
    → ground truth should have also changed

On-manifold adversarial attacks

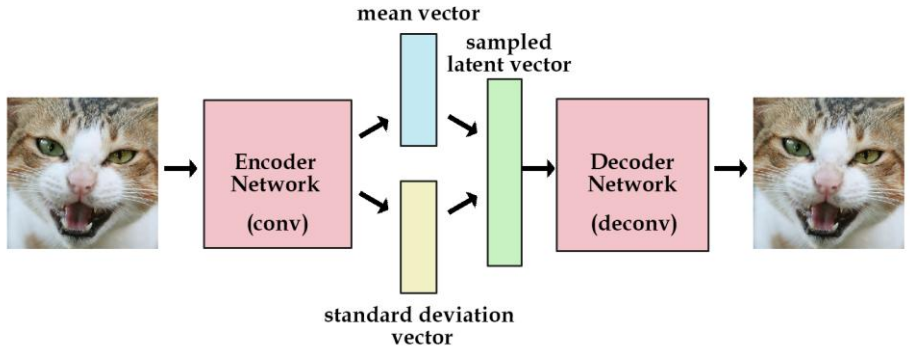
- The adversarial examples should lie *within the distribution of the original data manifold* learned by an **LSTM Variational Autoencoder (VAE)**
  - Auto-encoders trained to encode data onto a lower dimensional latent space and decode them into the original sample
  - Variational autoencoders encode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

- We project the adversarial example to the data manifold

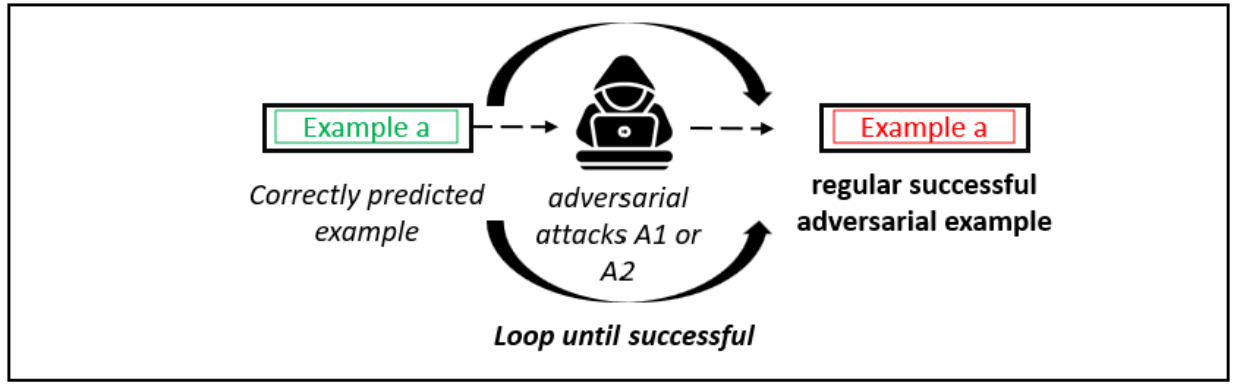
  → *natural*

- For both classes separately

  → adhere to label invariance



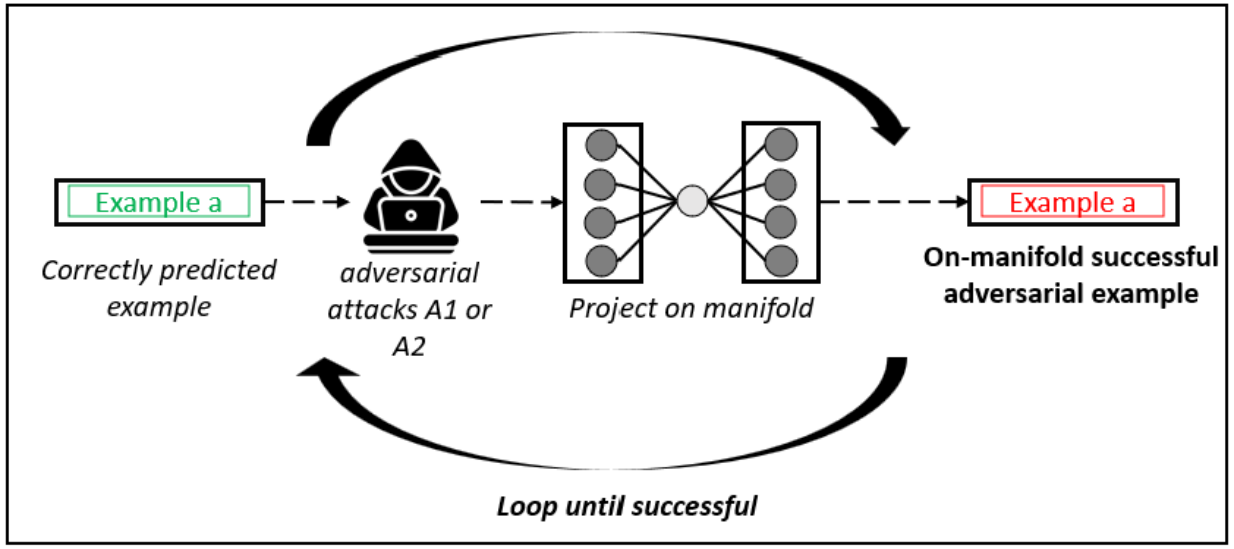
Manifold Learning for Adversarial Robustness in Predictive Process Monitoring



(a) Regular successful adversarial examples

Regular successful adversarial examples

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



(b) On-manifold successful adversarial examples

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

Alexander Stevens, Jari Peeperkorn, Johannes De Smedt , Jochen De Weerdt. Manifold Learning for Adversarial Robustness in Predictive Process Monitoring. ICPM (2023)

Counterfactual explanations

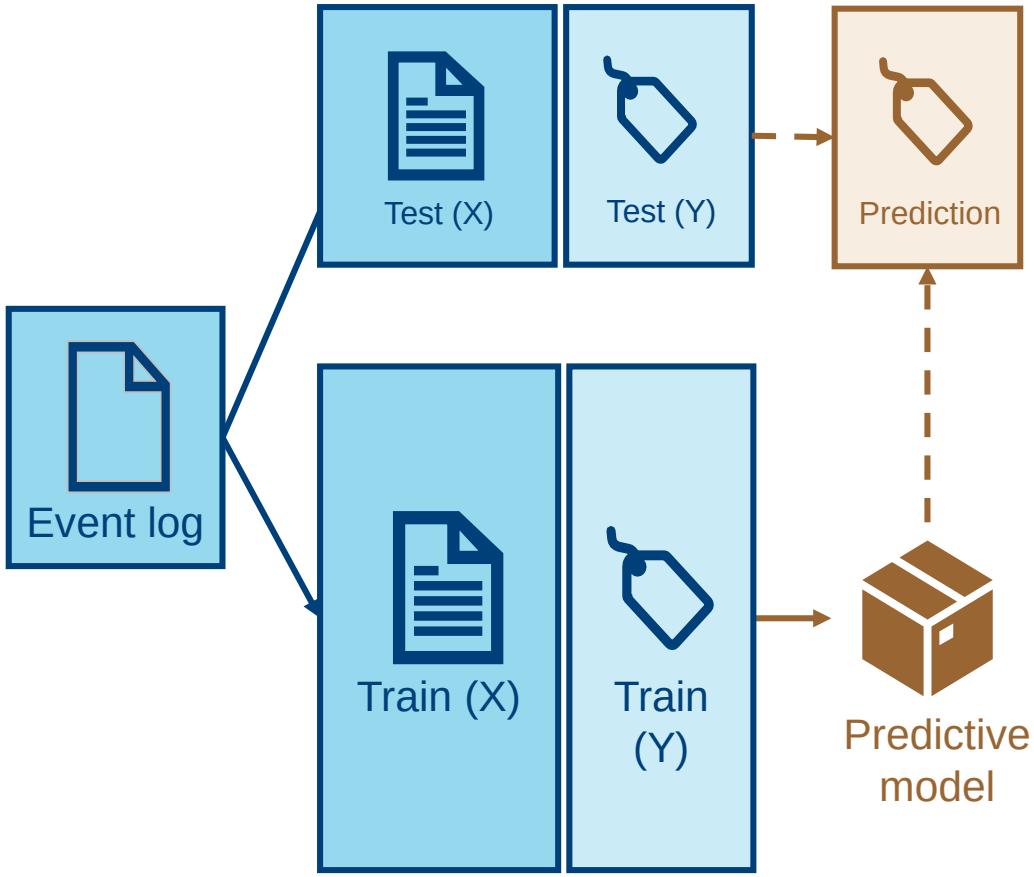
- Going from pure predictive to prescriptive
  - We assume the prediction to be correct → what should we have done differently?

  - *Will our patient be admitted to the ICU, or not?*
  - *If so, how could we have prevented that?*

Alexander Stevens, Chun Ouyang, Johannes De Smedt, Catarina Moreira, Generating Feasible and Plausible Counterfactual Explanations for Outcome Prediction of Business Processes, arXiv:2403.09232

Alexander Stevens, Chun Ouyang, Johannes De Smedt, Catarina Moreira, Generating Feasible and Plausible Counterfactual Explanations for Outcome Prediction of Business Processes, arXiv:2403.09232

Counterfactual explanations



Counterfactual explanation *"How could I have had another prediction?"*

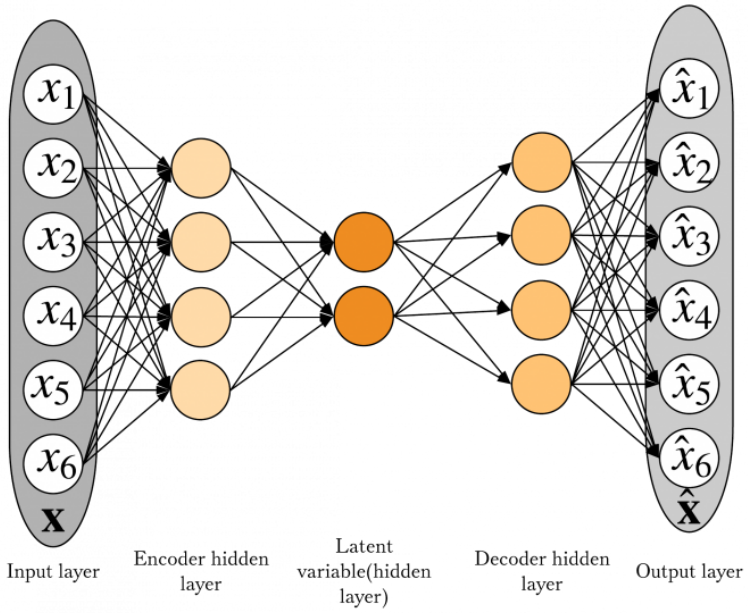
Properties to evaluate counterfactual explanations*

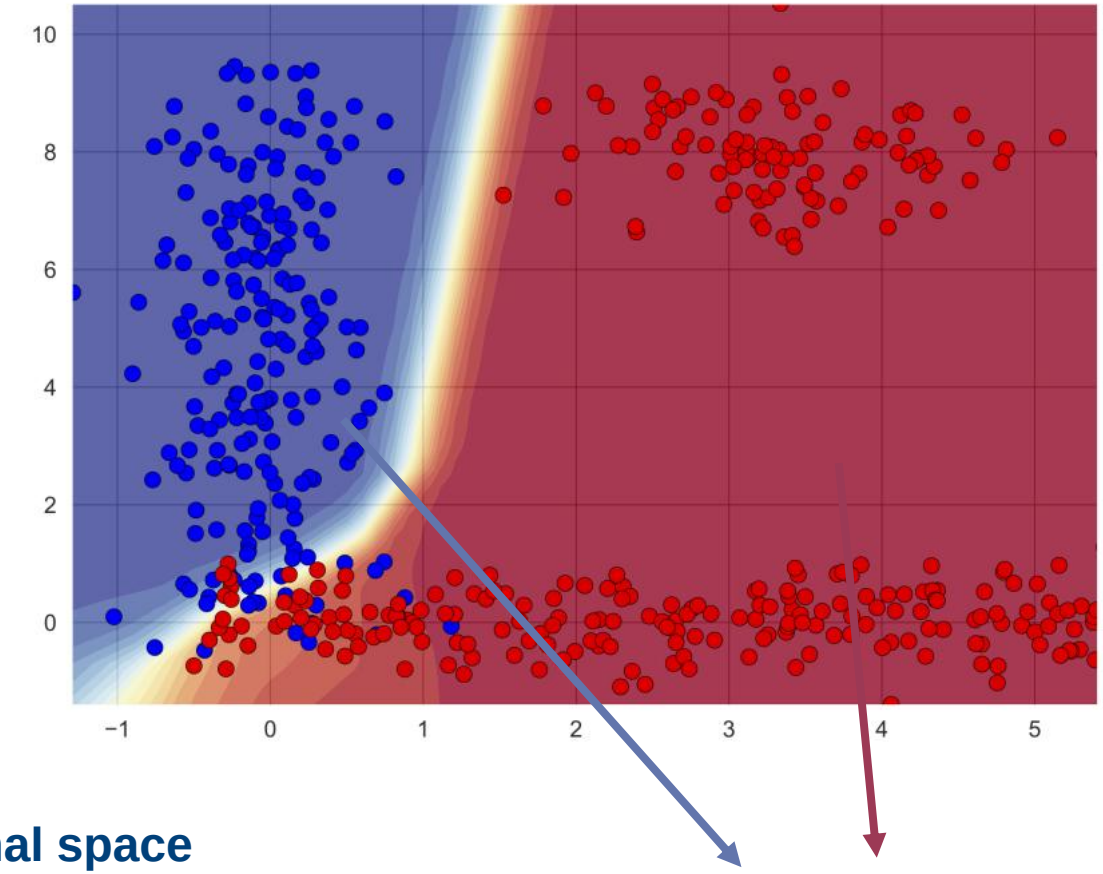
1. Proximity
2. Diversity
3. Sparsity
4. Feasibility
5. Plausibility

* Chou, Y. L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. Information Fusion, 81, 59-83.

Alexander Stevens, Chun Ouyang, Johannes De Smedt, Catarina Moreira, Generating Feasible and Plausible Counterfactual Explanations for Outcome Prediction of Business Processes, arXiv:2403.09232

Learn the Data Manifold of the Process Data





Design of a variational autoencoder (VAE)

➔ learns to encode your data in a **low dimensional space**

Perfectly separated classes in the latent space (synthetic data)

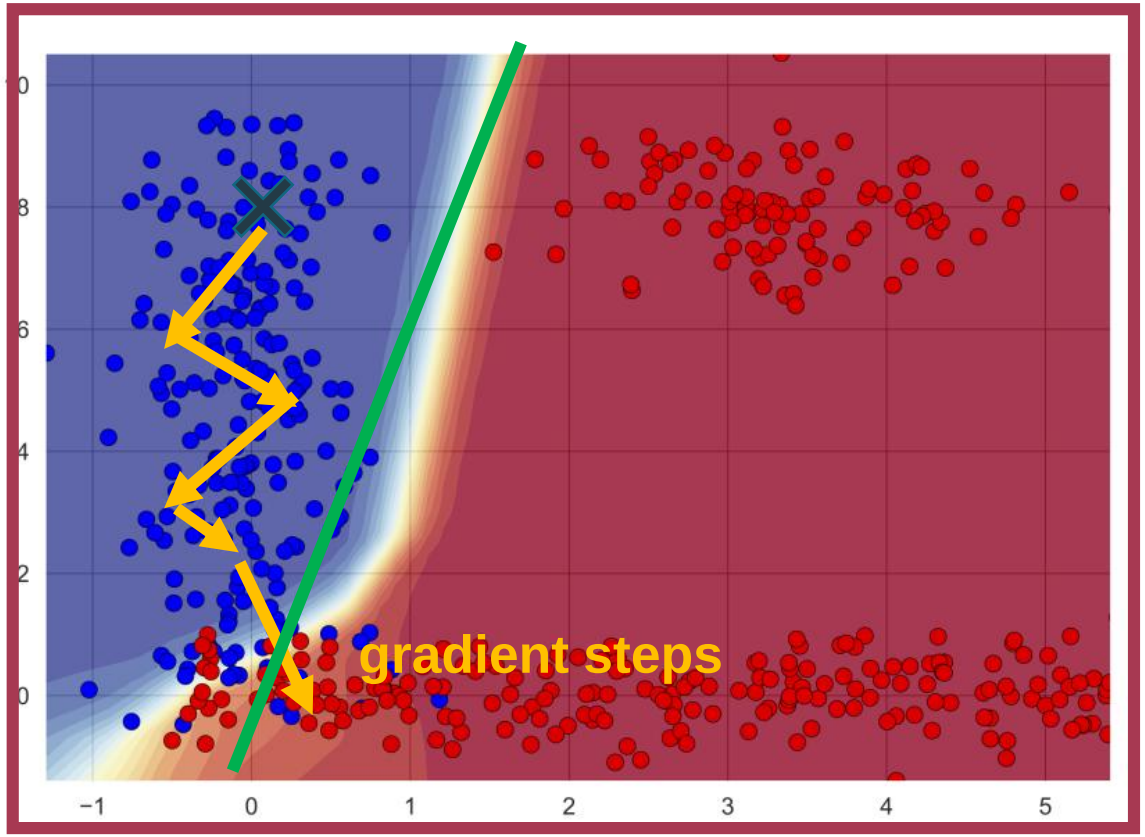
Changes:

- **Include process constraints (DECLARE) in loss function**
- **One VAE, not one per class**

Alexander Stevens, Chun Ouyang, Johannes De Smedt, Catarina Moreira, Generating Feasible and Plausible Counterfactual Explanations for Outcome Prediction of Business Processes, arXiv:2403.09232

Counterfactual generation algorithm

data manifold



decision boundary
of the classifier

e.g., pretrained LSTM

- We need to traverse the decision boundary of a **classifier** by taking "**gradient steps**" across the **data manifold**

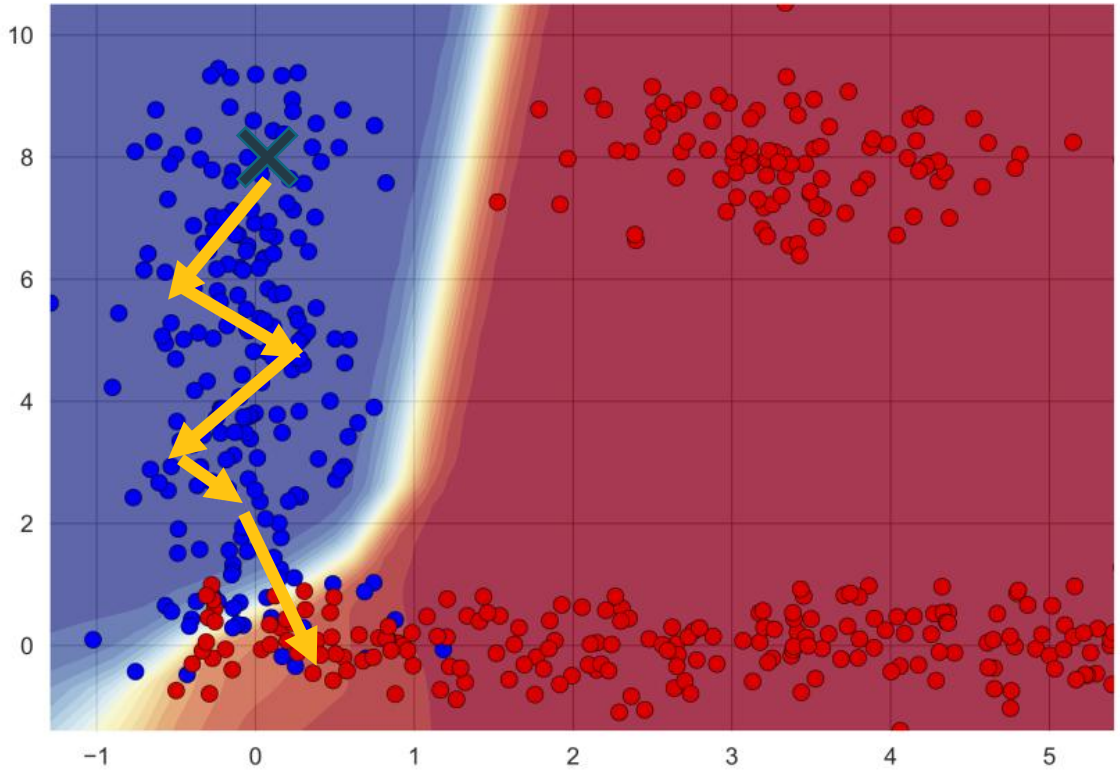
So that it minimizes a
specified loss function

e.g., pretrained VAE

Alexander Stevens, Chun Ouyang, Johannes De Smedt, Catarina Moreira, Generating Feasible and Plausible Counterfactual Explanations for Outcome Prediction of Business Processes, arXiv:2403.09232

Counterfactual generation algorithm

The counterfactual generation loss function we should minimize is:



$$\text{Loss}(\sigma) = \lambda\text{hinge} * \text{Lhinge} + \lambda\text{dist} * \text{Ldist} + \lambda\text{DPC} * \text{L}(\theta)\text{DPC}$$

The distance between the prediction for counterfactual x' and the desired outcome we want (the opposite of the original label)

Label-specific process constraint violations (DECLARE)

The distance between the initial trace and the counterfactual (in embedded space)

Alexander Stevens, Chun Ouyang, Johannes De Smedt, Catarina Moreira, Generating Feasible and Plausible Counterfactual Explanations for Outcome Prediction of Business Processes, arXiv:2403.09232

Conclusion

- Adapting XAI techniques to process data is not always straight forward
- You can use adversarial attacks to measure the robustness of predictive process monitoring models
  - And adversarial training to boost robustness
- You can use counterfactual examples to explain how certain outcomes could have been avoided
- Manifold learning (e.g. by using a variational auto-encoder) can help generate more natural and closer examples in both cases

Question?

Sources

Alexander Stevens, Jari Peeperkorn, Johannes De Smedt , Jochen De Weerdt. Assessing the Robustness in Predictive Process Monitoring through Adversarial Attacks. ICPM (2022)

Alexander Stevens, Jari Peeperkorn, Johannes De Smedt , Jochen De Weerdt. Manifold Learning for Adversarial Robustness in Predictive Process Monitoring. ICPM (2023)

Alexander Stevens, Chun Ouyang, Johannes De Smedt, Catarina Moreira, Generating Feasible and Plausible Counterfactual Explanations for Outcome Prediction of Business Processes, arXiv:2403.09232 (2024)

Adversarial training

- **The prediction is unchanged, but the XAI method is focusing on something completely different!**

“this picture contains a house”

Explanation

Image

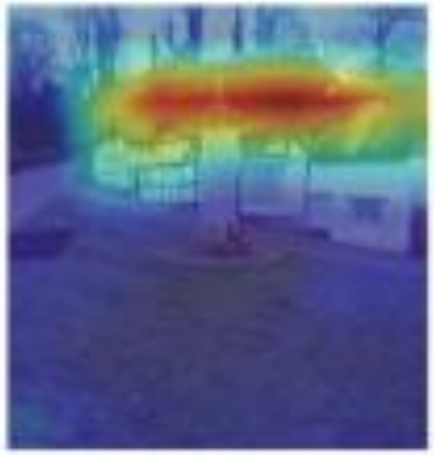
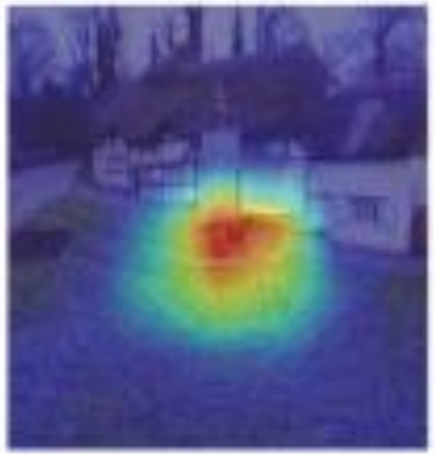


Image + Noise

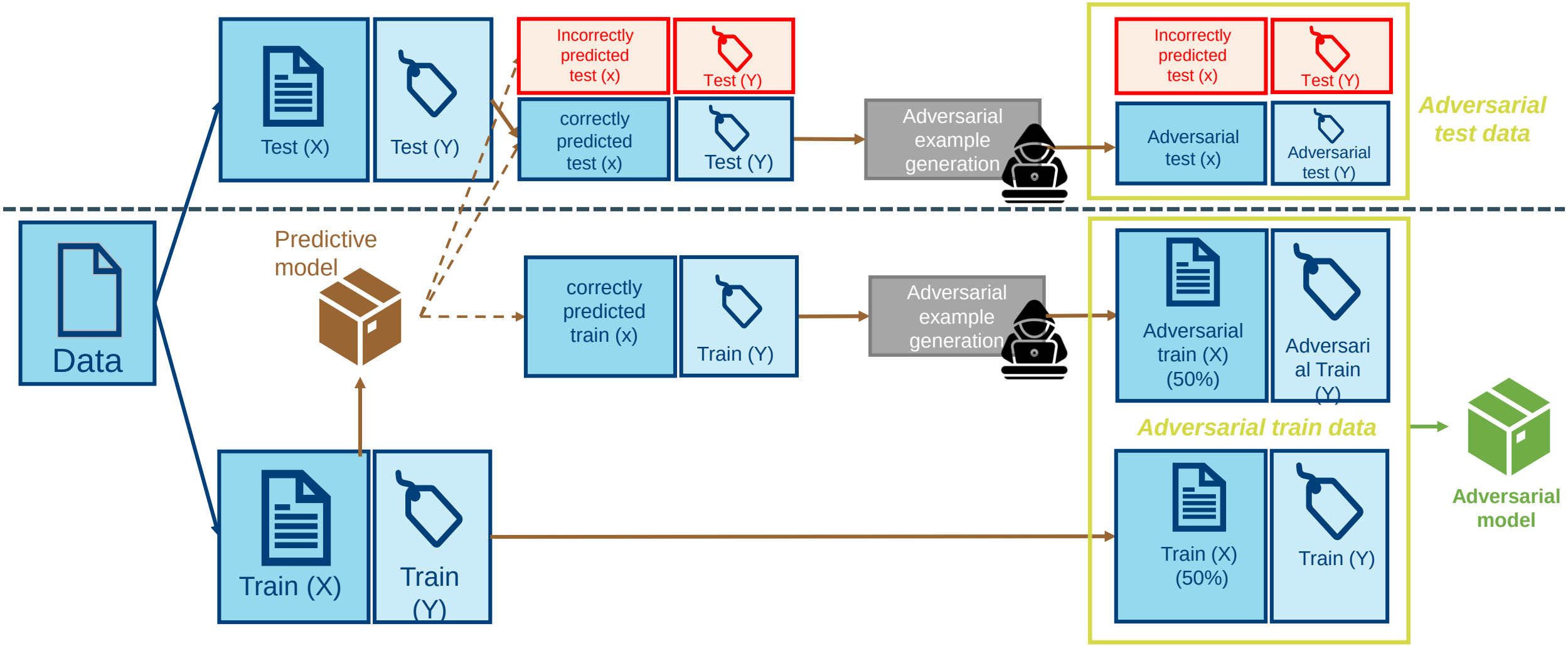


➔ Small perturbations can cause explanations to change, although the prediction is unchanged

➔ Your setup model + explanations is not robust

Manifold Learning for Adversarial Robustness in Predictive Process Monitoring

RESEARCH CENTRE FOR INFORMATION SYSTEMS ENGINEERING (LIRIS)



Experimental Setup

KU LEUVEN

RESEARCH CENTRE FOR INFORMATION SYSTEMS ENGINEERING (LIRIS)

- 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

KU LEUVEN
RESEARCH CENTRE FOR INFORMATION SYSTEMS ENGINEERING (LIRIS)

Original test data and original model

Original test data and adversarial model

Original test data and on-manifold adversarial model

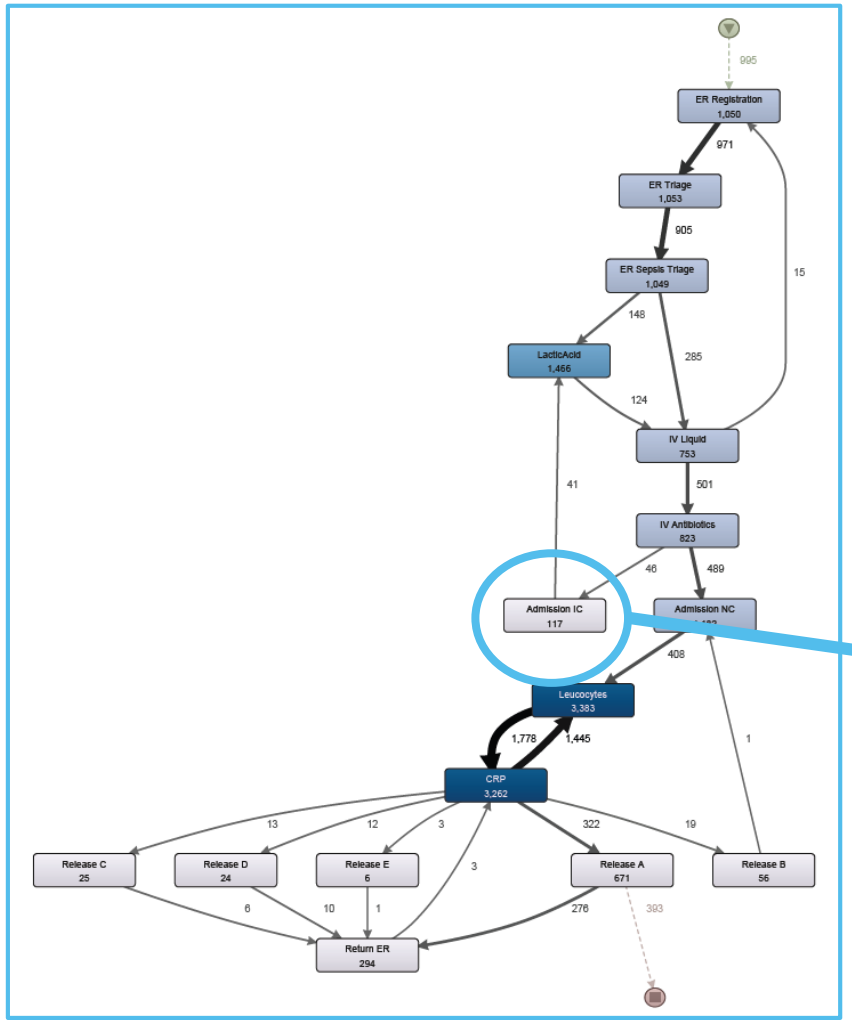
Adversarial test data and original model

Adversarial test data and adversarial model

Adversarial test data and on-manifold adversarial model

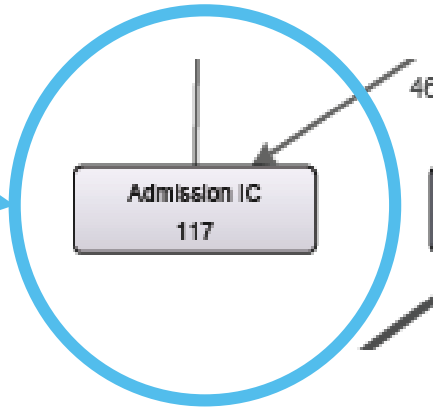
BPIC2012 (Accepted)		No Defense	Adversarial Training				On-manifold adversarial training			
			A1 (Act)	A1 (Res)	A2 (Act)	A2 (Res)	A1 (Act)	A1 (Res)	A2 (Act)	A2 (Res)
LR	No attack	66.52	56.33	58.83	55.62	62.62	60.86	60.93	62.77	**62.83**
	Attack (manifold) A1 (Act)	0.0	72.55	11.10	67.63	11.88	**86.41**	**86.11**	84.47	85.89
	Attack (manifold) A1 (Res)	0.0	74.17	10.91	68.94	11.37	**86.86**	**86.69**	85.13	86.38
	Attack (manifold) A2 (Act)	0.0	74.16	7.97	65.76	11.28	**82.58**	82.02	**84.48**	81.46
	Attack (manifold) A2 (Res)	0.0	70.55	17.87	67.68	19.24	86.72	**86.76**	83.36	**90.15**
RF	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
	Attack (manifold) A1 (Res)	0.0	19.26	23.65	29.27	8.78	**82.35**	**82.20**	74.48	75.82
	Attack (manifold) A2 (Act)	0.0	34.74	21.59	47.84	23.66	80.53	79.90	**83.71**	**81.49**
	Attack (manifold) A2 (Res)	0.0	23.09	36.69	26.95	35.72	**84.34**	84.14	84.29	**85.47**
XGB	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
	Attack (manifold) A1 (Res)	0.0	28.93	32.54	25.21	11.87	**88.02**	**88.03**	82.78	84.27
	Attack (manifold) A2 (Act)	0.0	50.51	26.76	41.86	18.08	82.20	82.05	**85.69**	**84.05**
	Attack (manifold) A2 (Res)	0.0	31.77	34.11	27.25	36.18	85.08	85.11	**85.71**	**86.07**
LSTM	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
	Attack (manifold) A1 (Res)	0.0	60.06	26.23	49.56	31.87	**83.85**	**83.87**	81.53	79.76
	Attack (manifold) A2 (Act)	0.0	58.08	33.43	57.01	48.31	83.54	83.49	**85.35**	**84.21**
	Attack (manifold) A2 (Res)	0.0	61.10	29.76	53.86	52.88	87.27	**87.29**	**87.34**	87.19

What is the (Research) Problem?

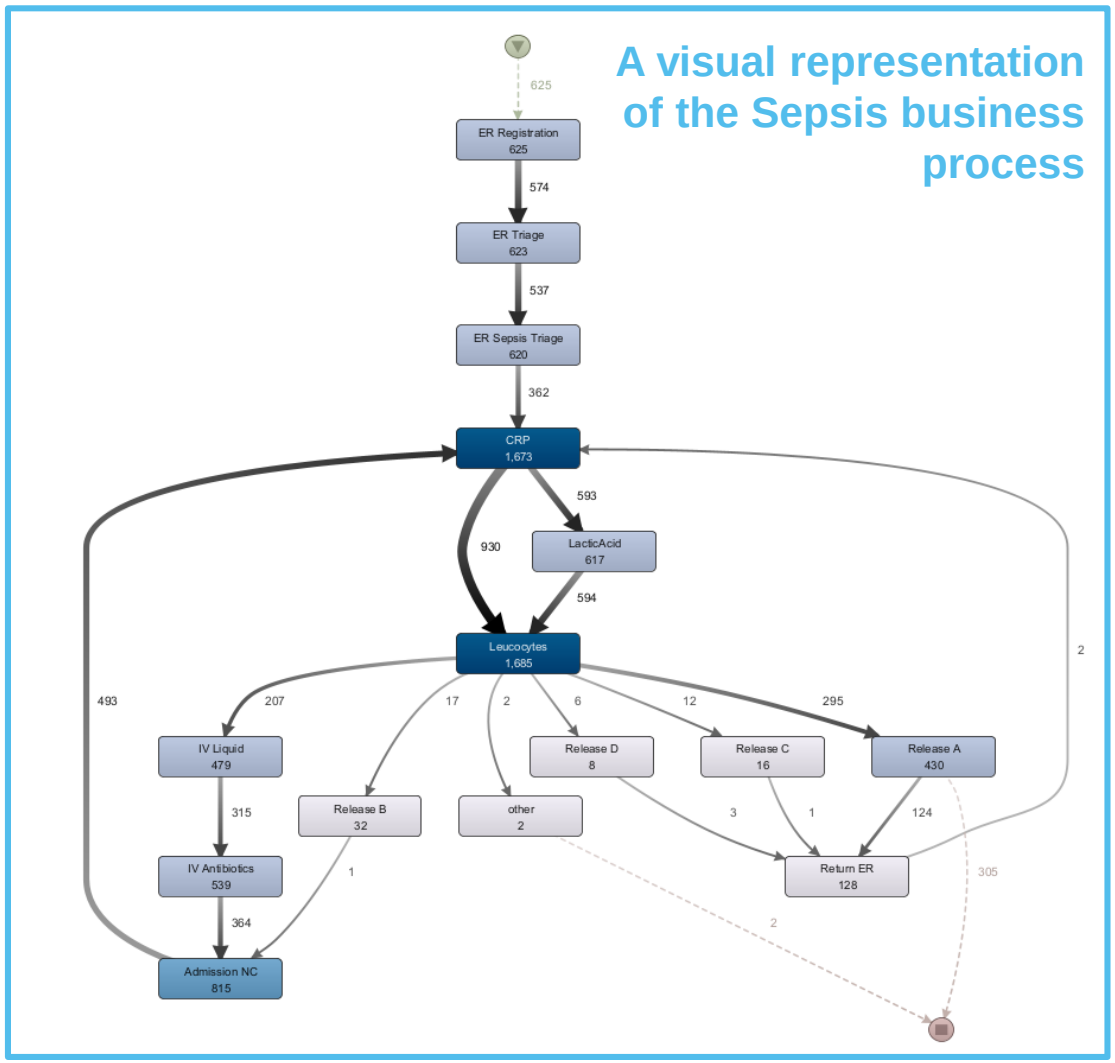


Sepsis Cases

- Contains trajectories (journeys) of patients with symptoms of the life-threatening sepsis condition in a Dutch hospital
- **Activity**: Admission to the Intensive Care Unit (ICU)



What is the (Research) Problem?



We removed the existence of the activity *ICU admission*

Business objective

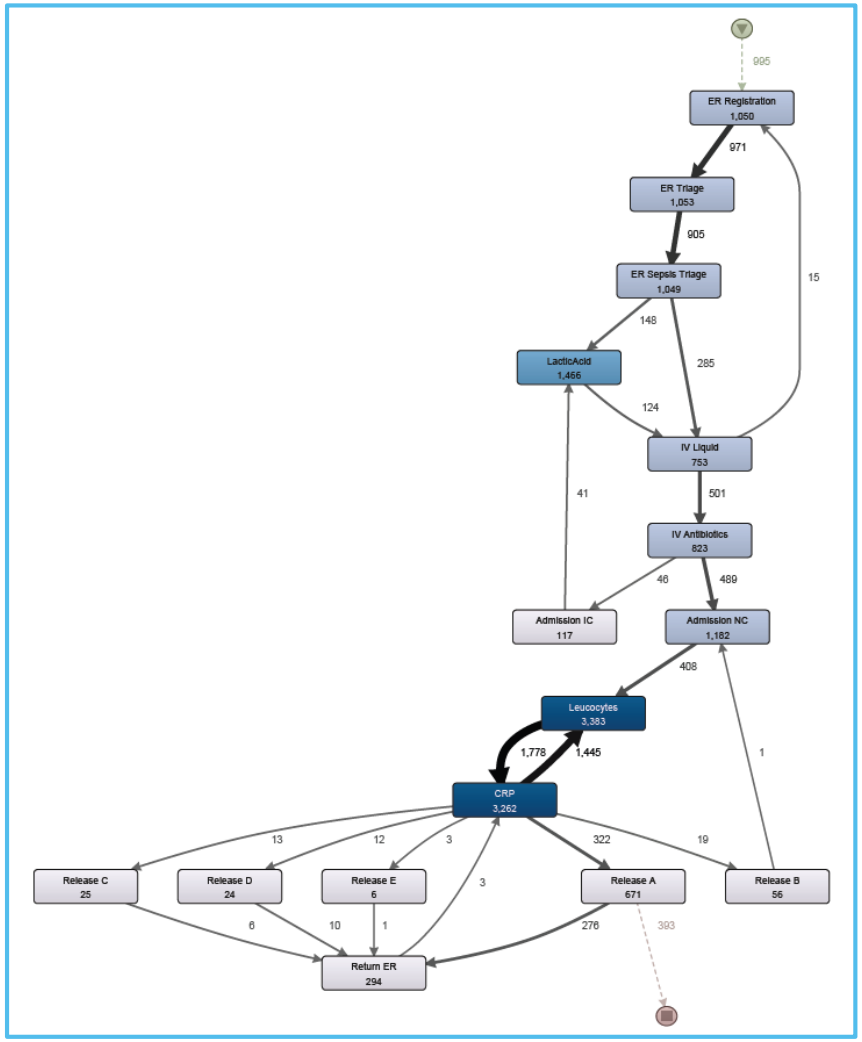
➔ Will our patient be admitted to the ICU, or not?

➔ If so, how could we have prevented that?

Plausibility through **Declare Process Constraints**

- In compliance with business rules or regulations
  - E.g. a patient needs to be **registered** before we can **start treatment**

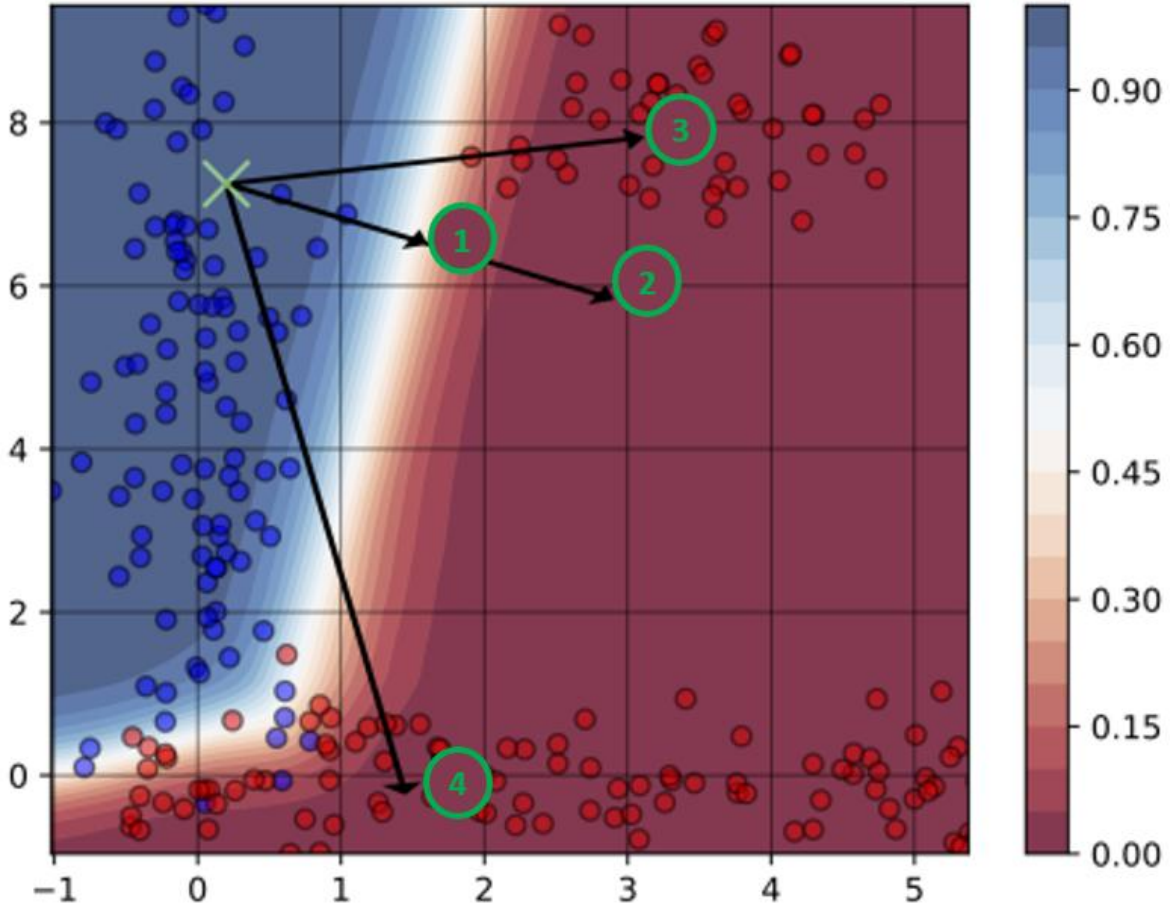
- Declare Language to extract patterns that describe sequential or behavioural characteristics of processes



Properties to Generate Good Counterfactuals

Property	Definition for the property	How is it enforced in the field of XAI?	Can we adopt this metric for PPA?	Important?
Plausibility	*How believable/plausible is the counterfactual?*	Avoids use of immutable features such as race, gender…	*Not really, what does immutability mean for sequentially ordered activities?*	✓
Feasibility	*How feasible is the counterfactual?*	High-density regions	Yes	✓

Why is Feasibility so important?



Counterfactual ① and counterfactual ② are deemed infeasible

➔ they lie in low-density regions

Counterfactual ④ is preferred over counterfactual ③

➔ there is a feasible path between the initial data point and the counterfactual

*Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020, February). FACE: feasible and a

Research Questions

RQ1. What **properties** define the **validity of a counterfactual explanation** in the field of **predictive process analytics**?

RQ2. How can we generate counterfactual explanations that are adapted for a process-based analysis?

Step 1: Learn the Declare Process Constraints

Sequence of activities	Label
aabcdeaa	no admission
Bcabdgaf	no admission
efgda	ICU admission
fgdeaaa	ICU admission

IBCM*

Pattern 1	Pattern 2	Pattern 3	Label
1	0	1	no admission
1	0	1	no admission
0	1	1	ICU admission
0	1	1	ICU admission

Label-specific Declare constraints

Trace Declare constraints

*De Smedt, J., Deeva, G., & De Weerdt, J. (2019). Mining behavioral sequence constraints for classification. *IEEE Transactions on Knowledge and Data Engineering*, *32*(6), 1130-1142.

Sequence	Label
[a, a, b, c, d, e , a, a]	no admission
[b, c, a, b, d, g, a, f]	no admission
[e, f, g, d, a]	ICU admission
[f, g, d, e, a, a, a]	ICU admission

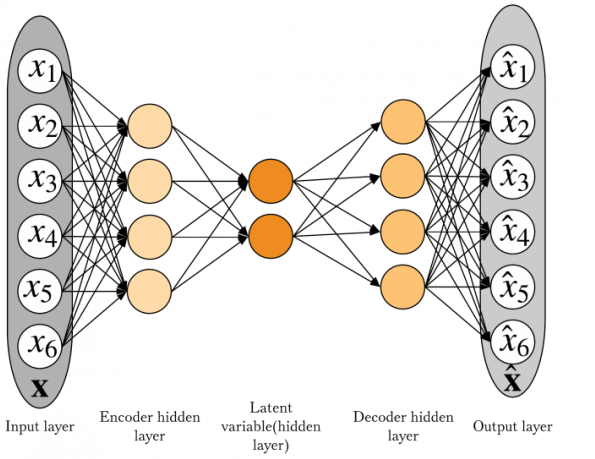
iBCM

Trace Declare constraints

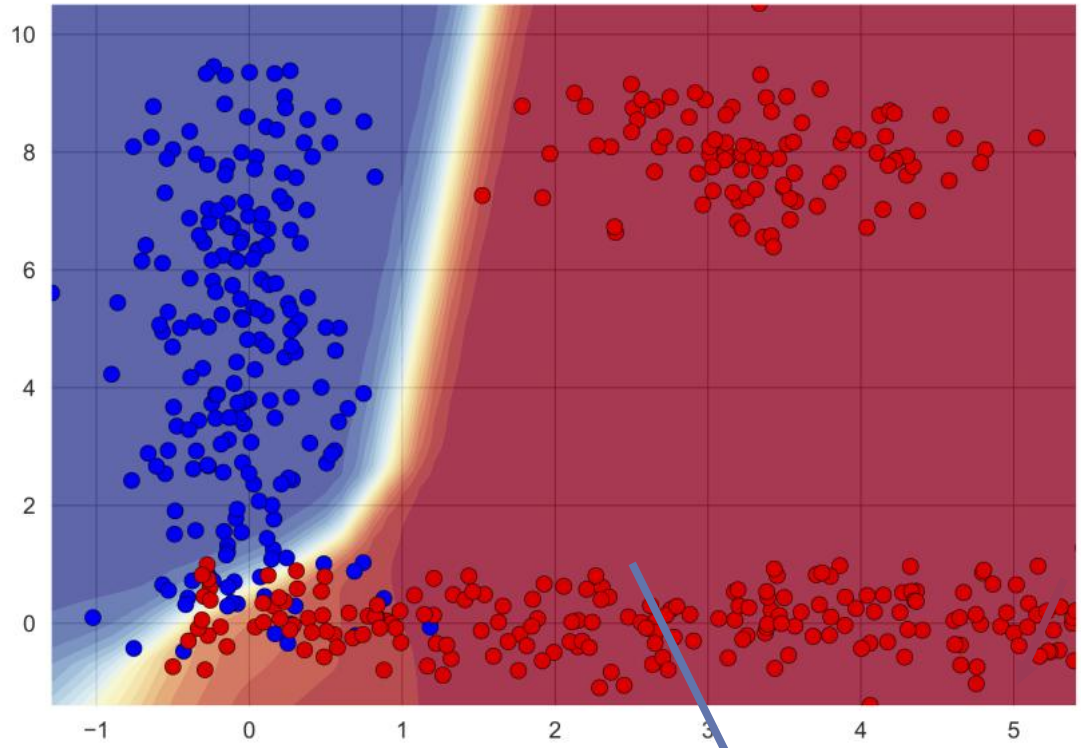
Pattern 1	Pattern 2	Pattern 3	Label
1	0	1	no admission
1	0	1	no admission
0	1	1	ICU admission
0	1	1	ICU admission

Label-specific Declare constraints

Step 2: Learn the Data Manifold of the Process Data



Design of a variational autoencoder (VAE)
➔ learns to encode your data in a **low dimensional space**



Perfectly separated classes in the latent space (synthetic data)

*Stevens, A., Peeperkorn, J., De Smedt, J., & De Weerdt, J. (2023, October). Manifold Learning for Adversarial Robustness in Predictive Process Monitoring, International Conferen... Mining (ICPM).

The Experimental Results: Counterfactuals

LSP

1. ER registration
2. ER triage
3. ER sepsis triage
4. Leucocytes
5. CRP
6. CRP
7. IV liquid
8. IV antibiotics
9. Admission NC
10. Admission NC
11. CRP
12. Admission NC

Occurs in 16.80% of the cases

Occurs in 23.52% of the cases

Occurs in 52.80% of the cases

Occurs in 19.04% of the cases

This is a newly generated *trace*, but…

- There are **no process constraint violations**
- We see certain **patterns** (empirical evaluation)
- **80% of the nearest traces** in the event log have same label *no admission*
- The **LSP is 5**, meaning that we only need to change from step 6