UNIVERSITÄT DUISBURG ESSEN

Open-Minded

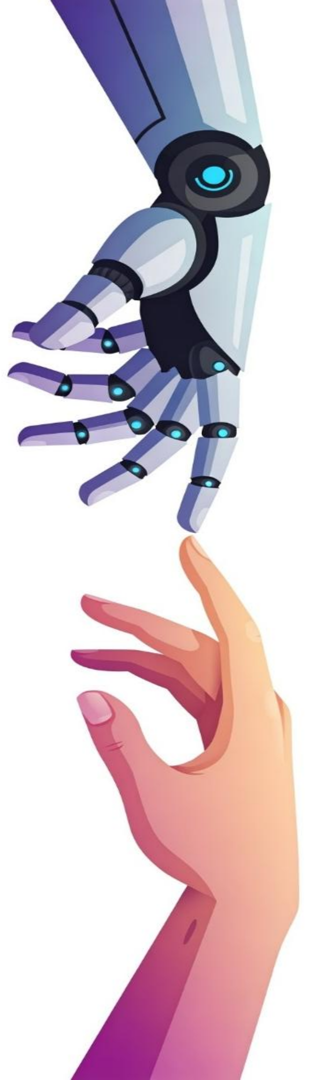
PALUNO

The Ruhr Institute for Software Technology

AI-Assisted Business Process Monitoring (Tutorial)

Andreas Metzger

Sevilla, September 02, 2025



Based on:

ELSEVIER

Information Systems

Volume 118, September 2023, 102254

Automatically reconciling the trade-off between prediction accuracy and earliness in prescriptive business process monitoring

Andreas Metzger, Tristan Kley, Aristide Rothweiler, Klaus Pohl

https://doi.org/10.1016/j.is.2023.102254

Slides available from:
https://adaptive-systems.org/images/documents/bpm-tutorial-25.pdf

pingo: "Think-Pair-Share"

1. I pose a question
2. You think about the answer
3. You discuss it with your peer
4. You reply online

https://pingo.coactum.de/events/053187

pingo: "Think-Pair-Share"

Q1: How do you assess your skill-level in BPM?

https://pingo.coactum.de/events/053187

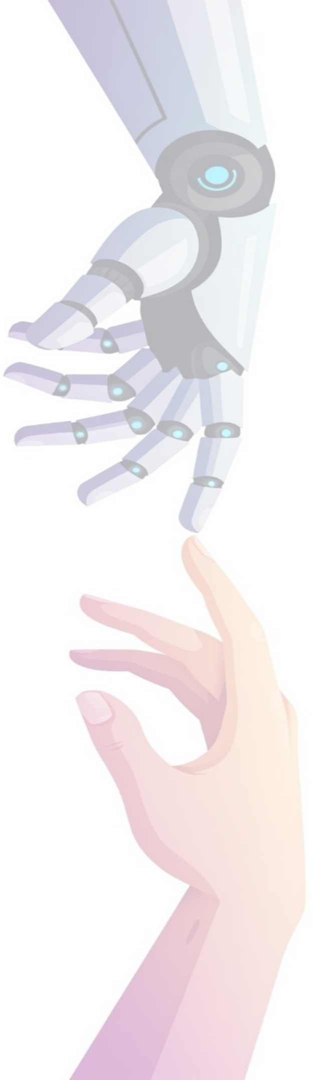
pingo: "Think-Pair-Share"

Q2: How do you assess your skill-level in AI?

https://pingo.coactum.de/events/053187

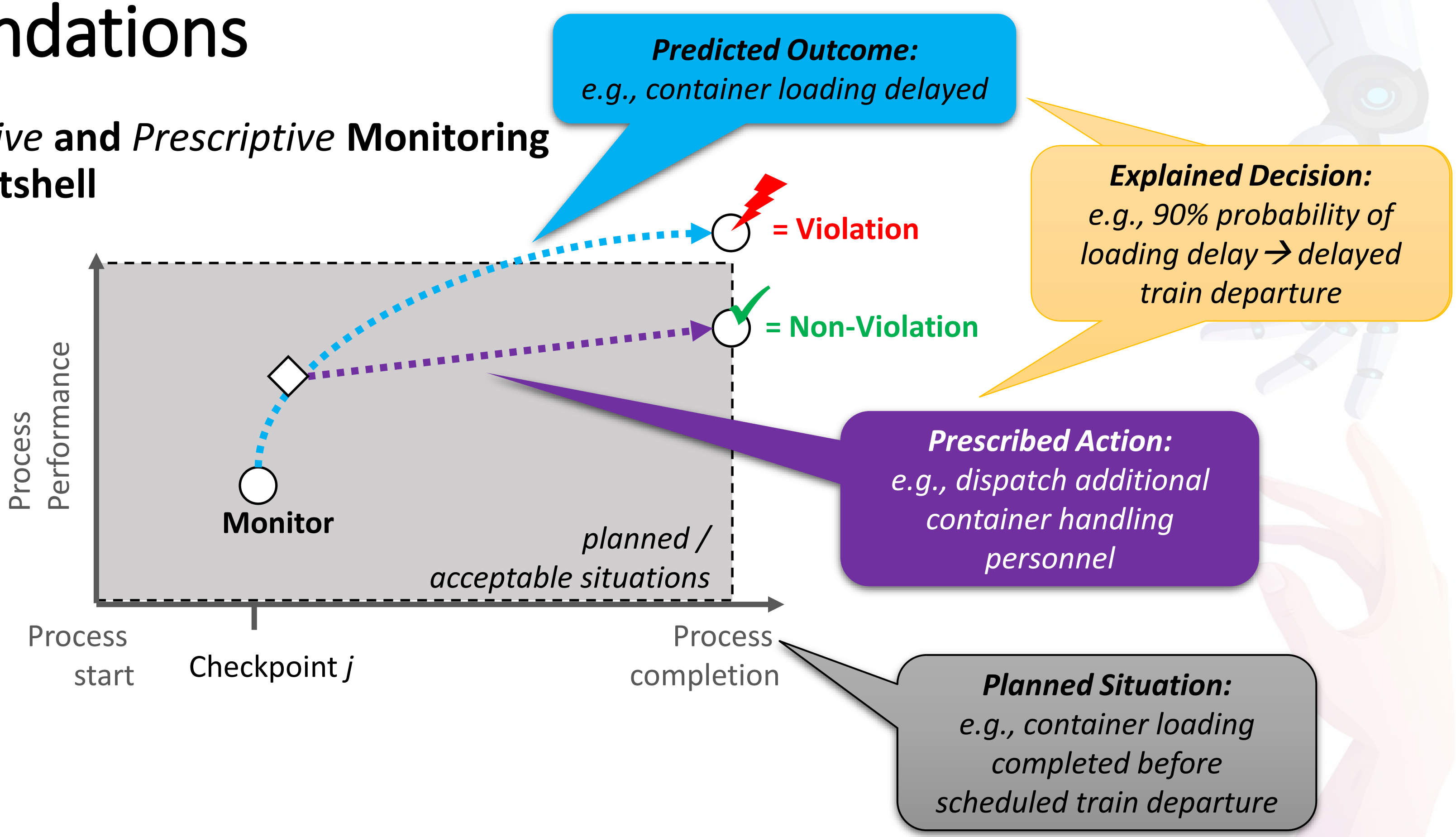
Agenda

1. Foundations
2. AI for *Predictive* Monitoring
   - Recurrent neural networks
   - Ensemble learning
3. AI for *Prescriptive* Monitoring
   - Online deep reinforcement learning
   - Generative AI
4. Future Directions



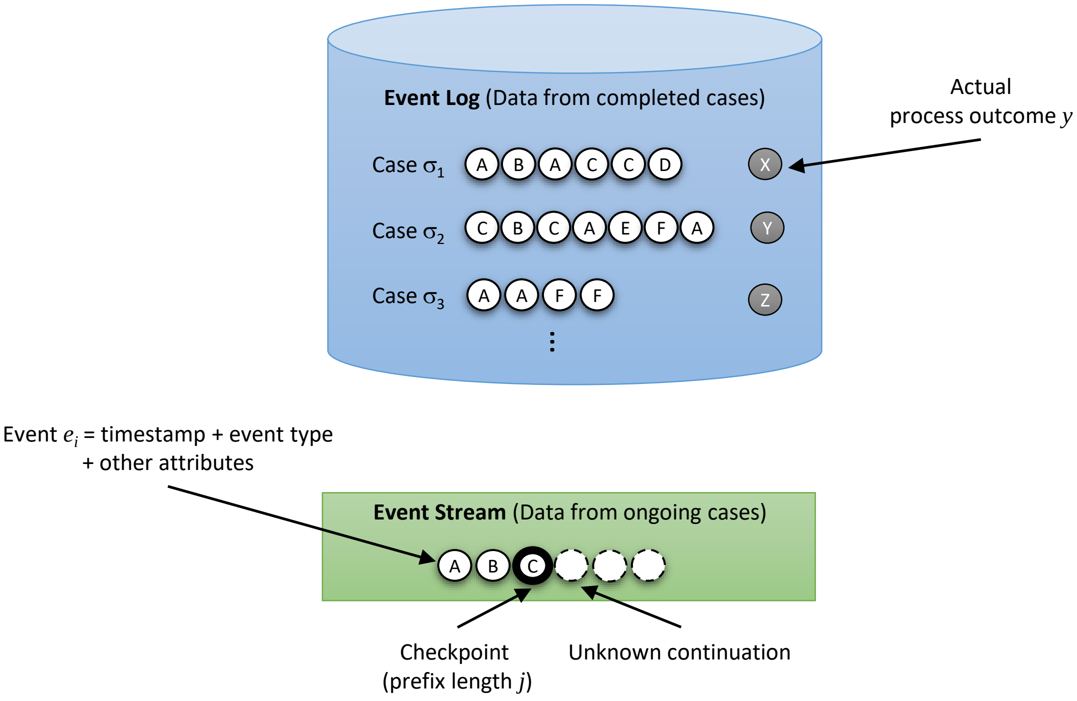
Foundations

Predictive **and** *Prescriptive* **Monitoring in a Nutshell**



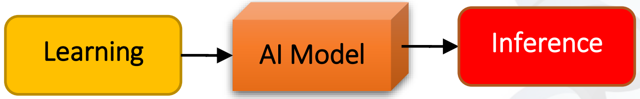
Foundations

Process Monitoring Data

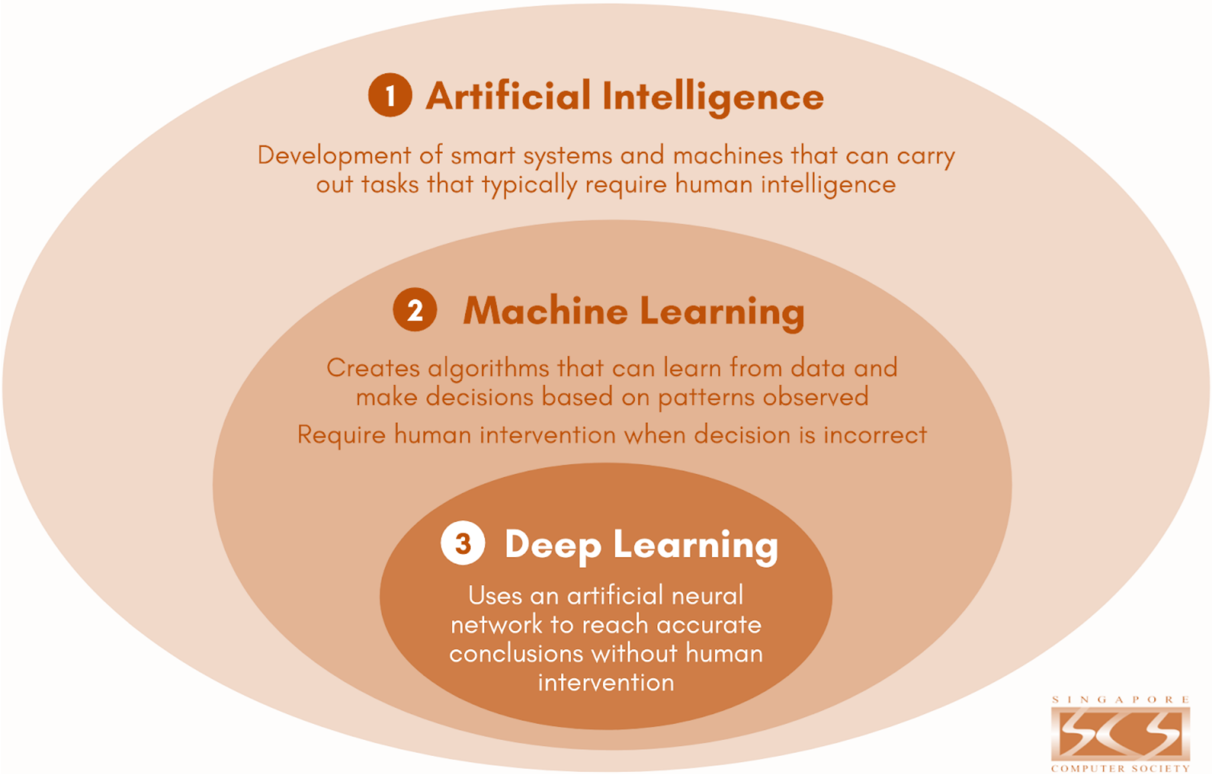


Foundations

Basic Machine Learning Principle

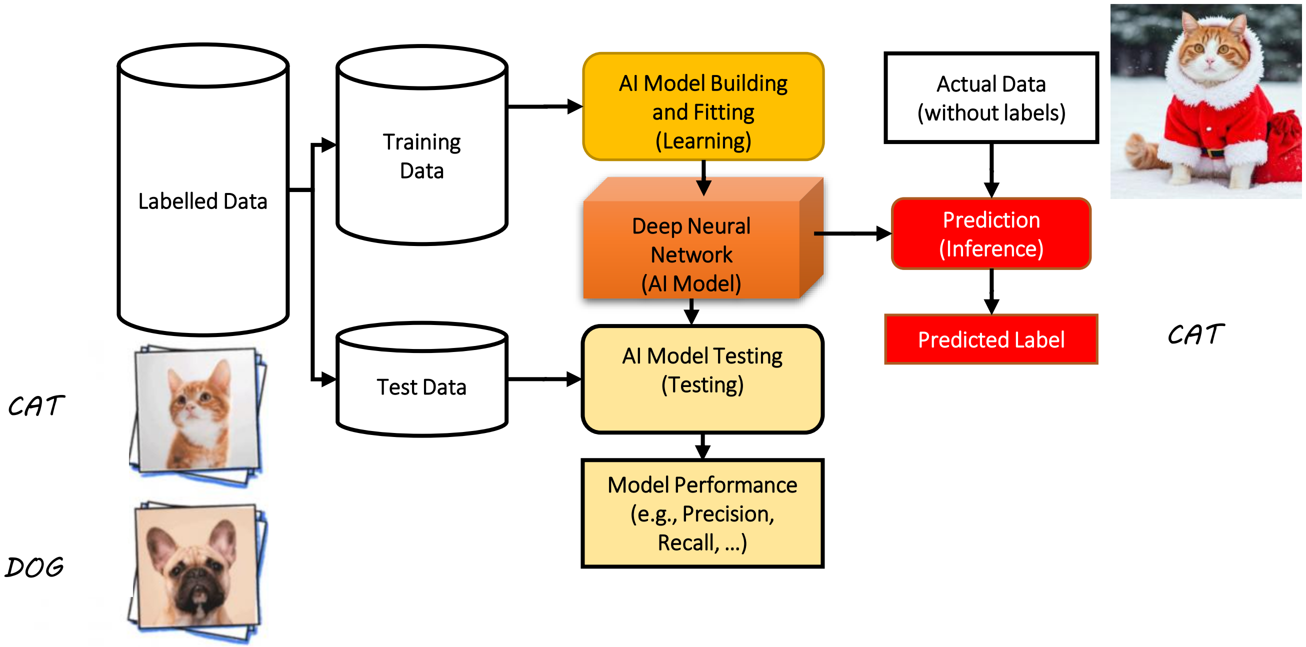


AI Taxonomy



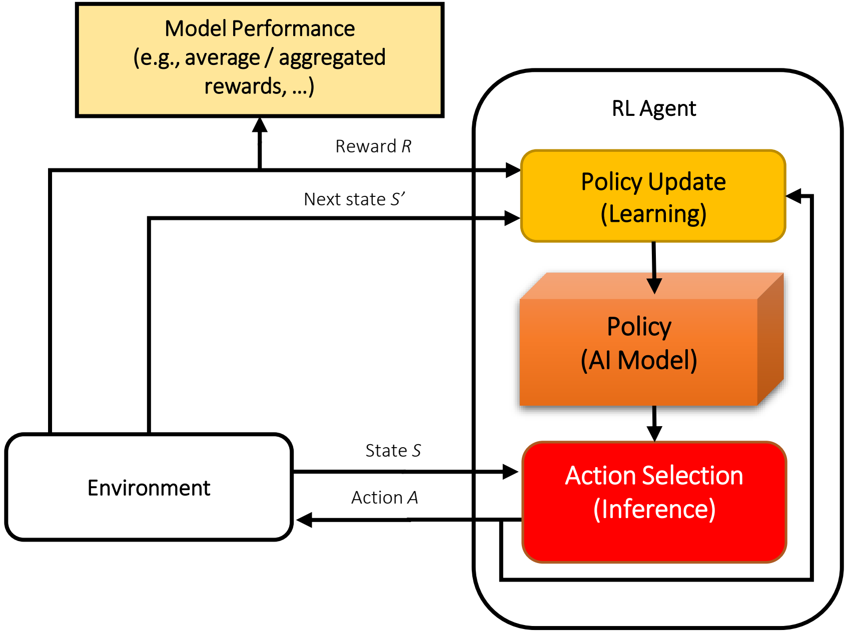
Foundations

AI-assisted BPM: *Supervised Learning*



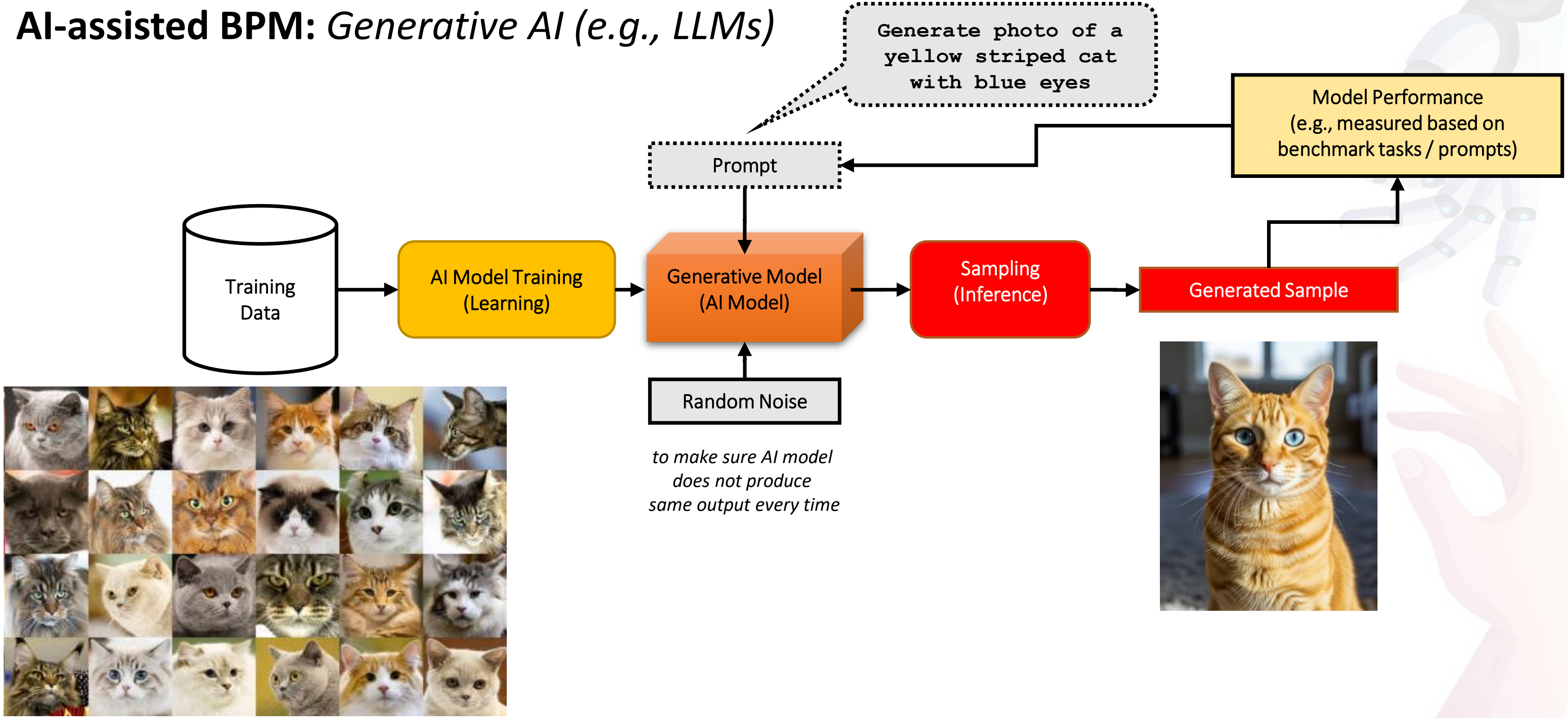
Foundations

AI-assisted BPM: *Reinforcement Learning*



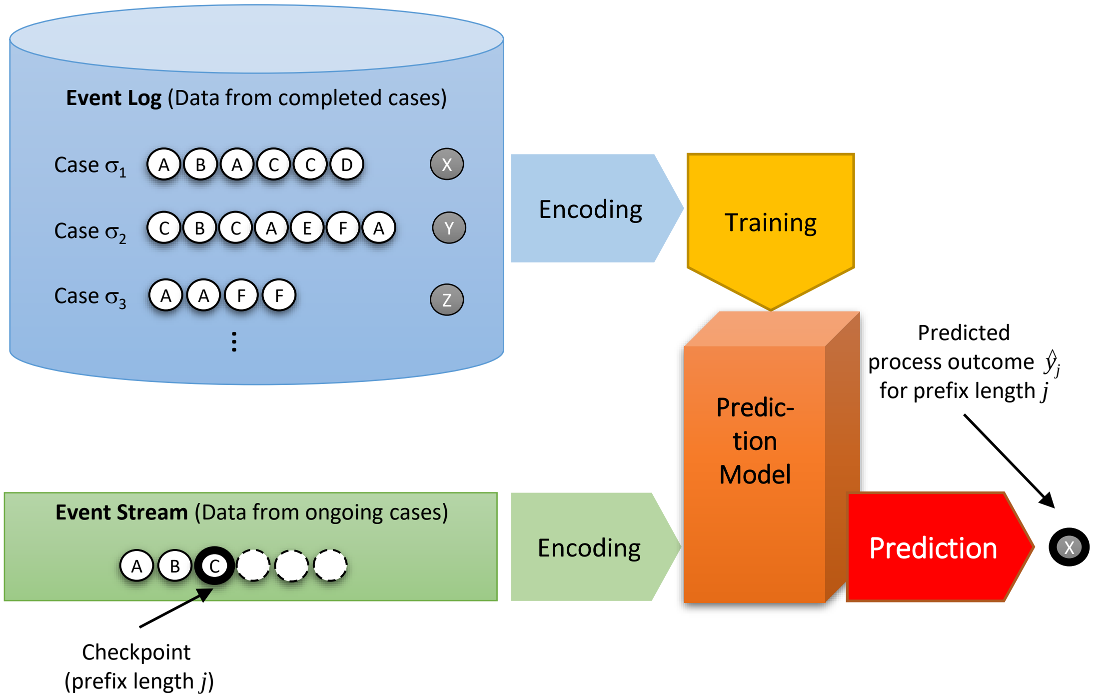
Foundations

AI-assisted BPM: *Generative AI (e.g., LLMs)*



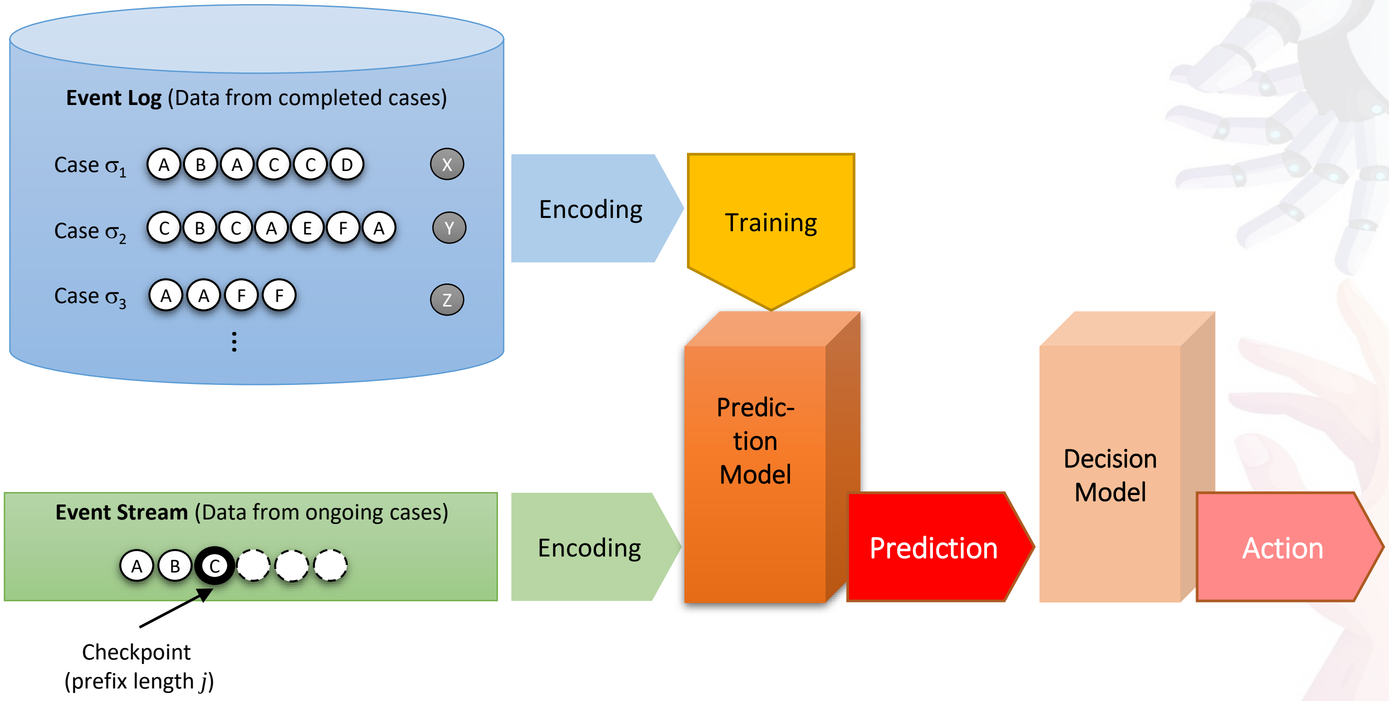
Foundations

AI-assisted ***Predictive*** **Monitoring:** Typically Supervised Learning



Foundations

AI-assisted *Prescriptive* **Monitoring:** Different Techniques



pingo: "Think-Pair-Share"

Q3: Why does it make sense to separate prediction from decision making?

https://pingo.coactum.de/events/053187

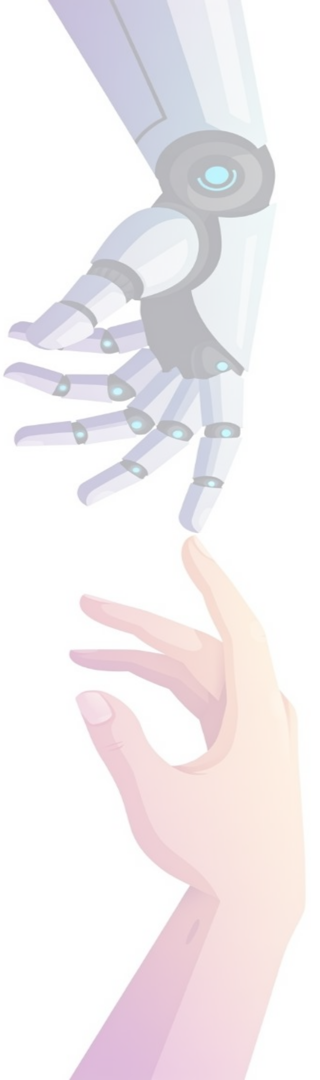
Foundations

Public **benchmark data sets** to assess model performance

Name	*Pos. Class*	*Pos. Class Ratio*	*Process Instances*	*Process Variants*	*Check-points*
Cargo2000	Delayed air cargo delivery	27%	3,942	144	7
Traffic	Unpaid traffic fine	46%	129,615	185	4
BPIC2012	Unsuccessful credit application	52%	13,087	3,587	23
BPIC2017	Unsuccessful credit application	59%	31,413	2,087	23

Agenda

1. Foundations
2. AI for *Predictive* Monitoring
   - Recurrent neural networks
   - Ensemble learning
3. AI for *Prescriptive* Monitoring
   - Online deep reinforcement learning
   - Generative AI
4. Future Directions



AI for Predictive Monitoring

Challenge 1: Prediction accuracy

- "Predict as many **true** deviations as possible,
  while predicting as few **false** deviations as possible"

Prediction contingencies and adaptation decisions based on predictions.

	Prediction $\hat{y}_j =$ *deviation*	Prediction $\hat{y}_j =$ *no deviation*
Actual $y =$ *deviation*	True Positive (*TP*) $\Rightarrow$ Necessary adaptation	False Negative (*FN*) $\Rightarrow$ **Missed** adaptation
Actual $y =$ *no deviation*	False Positive (*FP*) $\Rightarrow$ **Unnecessary** adaptation	True Negative (*TN*) $\Rightarrow$ No adaptation

AI for Predictive Monitoring

Challenge 2: Prediction reliability

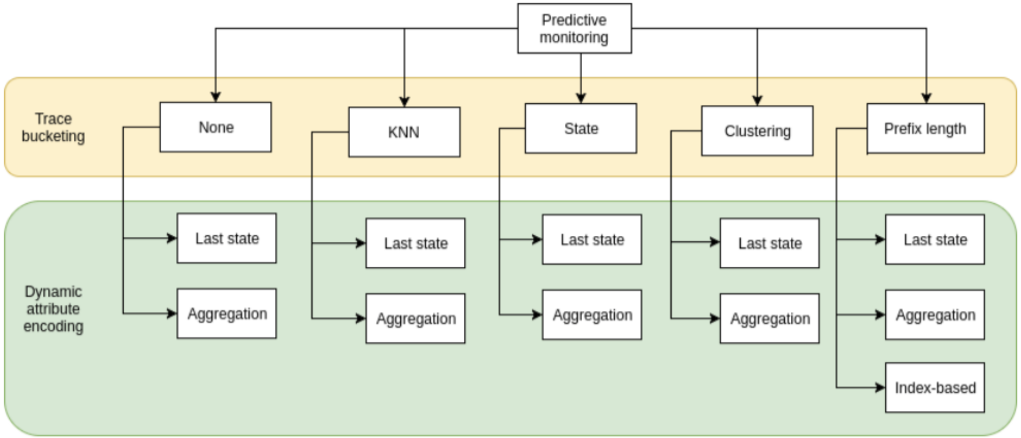
- “in how far can I **trust** the prediction?”
  → “when should I act on a prediction?”

Reliability estimation

AI for Predictive Monitoring

Challenge 3: Data encoding

- Classical prediction models (random forests) require encoding of event sequences into **fixed-length input vectors**
- **Many different** encoding choices



[Teinemaa et al. 2019 @ ACM Trans. Knowl. Discov. Data] https://doi.org/10.1145/3301300

[Tax et al. 2020 @ SoSym] https://doi.org/10.1007/s10270-020-00789-3

AI for Predictive Monitoring

Recurrent Neural Networks (RNNs)

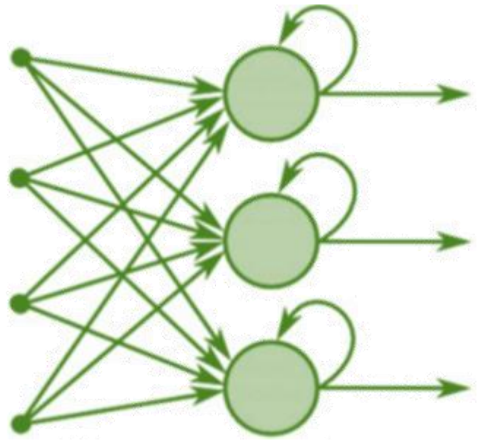
Pro

- High prediction accuracy → **Challenge 1**
  [Tax et. al. 2017 @ CAiSE; Metzger & Nebauer 2018 @SEAA]
- Arbitrary length process instances and predictions at any checkpoint (without sequence encoding)
  → **Challenge 2**

Con (e.g., when compared to random forests)

- Long training time
- No native reliability estimates

RNN Model



[Tax et al. 2017 @ CAiSE] https://doi.org/10.1007/978-3-319-59536-8_30
[Metzger & Nebauer 2018@ SEAA] https://doi.org/10.1109/SEAA.2018.00051

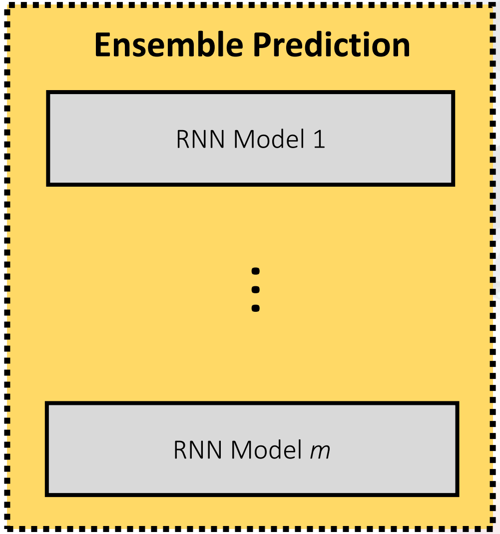
AI for Predictive Monitoring

RNN Ensembles

Pro

- Increased prediction accuracy → **Challenge 1**
- Computation of reliability estimates
  → **Challenge 3**

[Metzger & Föcker 2017 @ CAiSE] https://doi.org/10.1007/978-3-319-59536-8_28



Con (e.g., when compared to random forests)

- (Even longer) training time

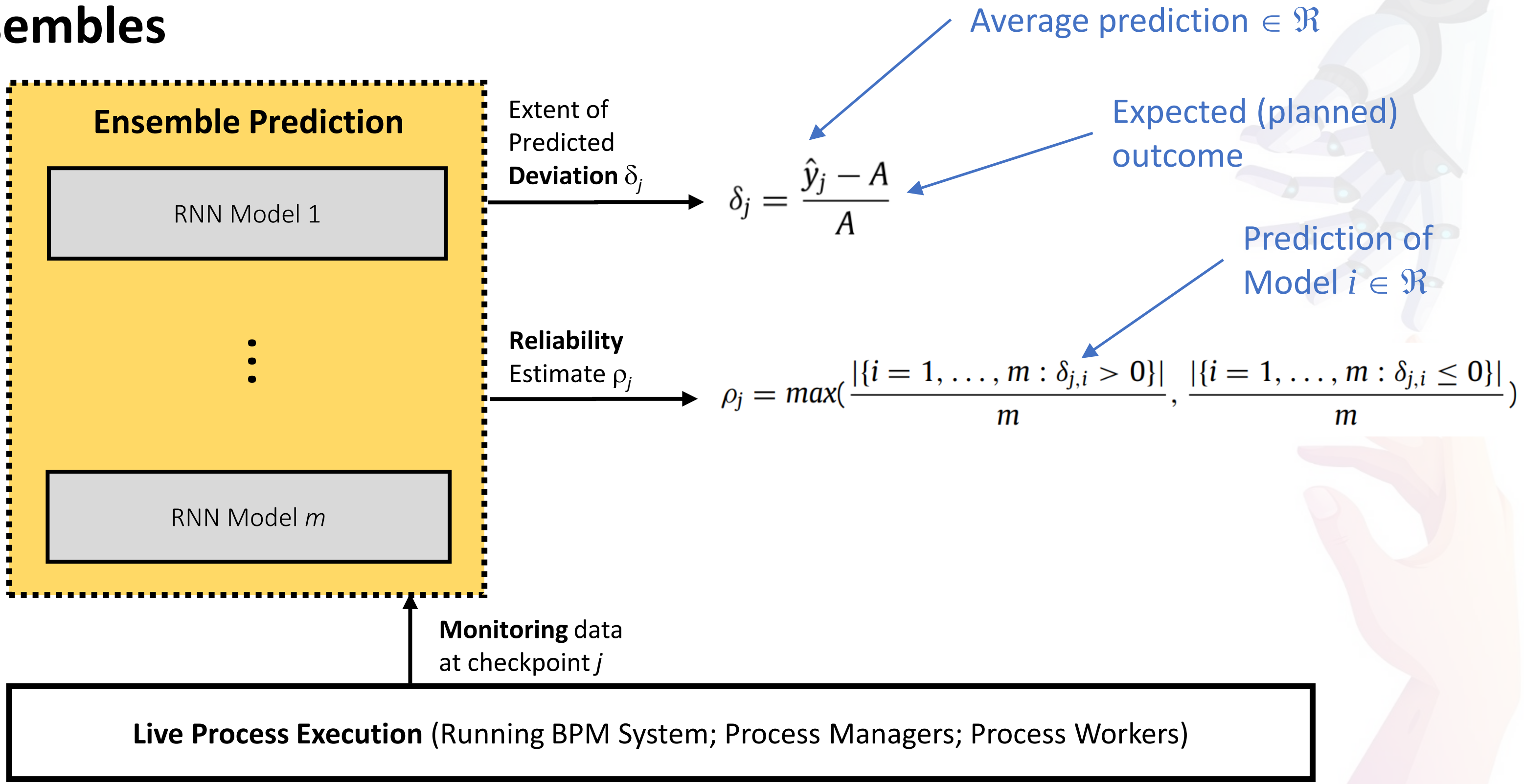
pingo: "Think-Pair-Share"

Q4: What is the benefit of this way of computing reliability estimates?

https://pingo.coactum.de/events/053187

AI for Predictive Monitoring

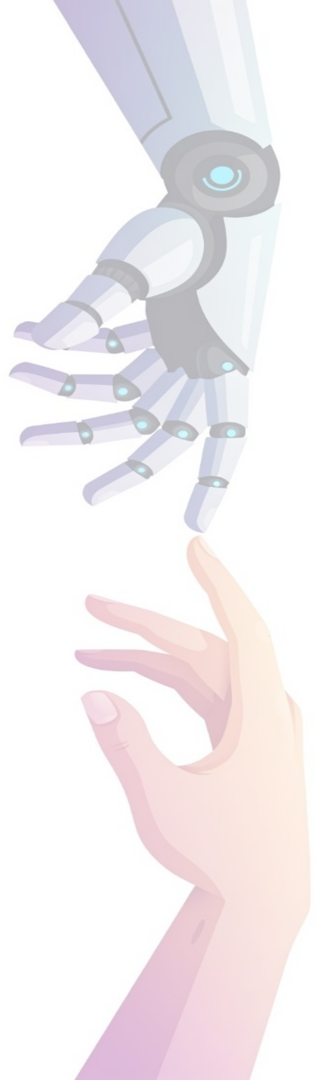
RNN Ensembles



[Metzger et al. 2019 @ CAiSE] https://doi.org/10.1007/978-3-030-21290-2_34

Agenda

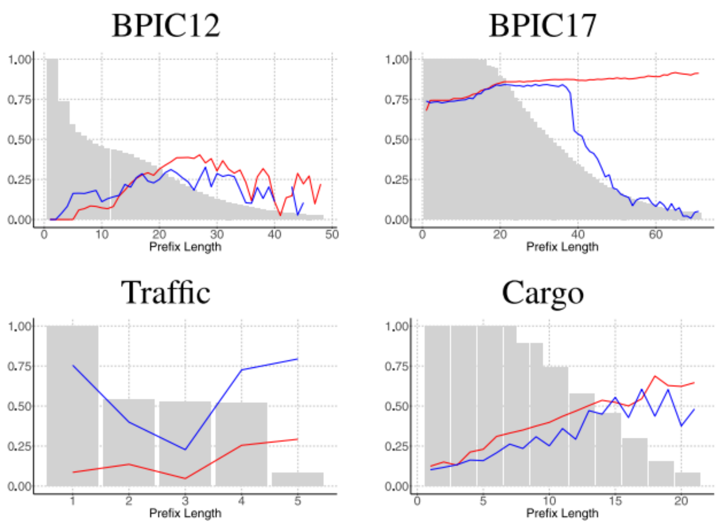
1. Foundations
2. AI for *Predictive* Monitoring
   - Recurrent neural networks
   - Ensemble learning
3. AI for *Prescriptive* Monitoring
   - Online deep reinforcement learning
   - Generative AI
4. Future Directions



AI for Prescriptive Monitoring

Challenge 1: Prediction accuracy vs action earliness

- **Prediction accuracy**
  - False positive prediction
    → unnecessary adaptation
  - False negative prediction
    → missed adaptation
- **Action earliness**
  - Later actions
    → less time and options for process adaptation
  - Earlier actions
    → higher risk of wrong process adaptation



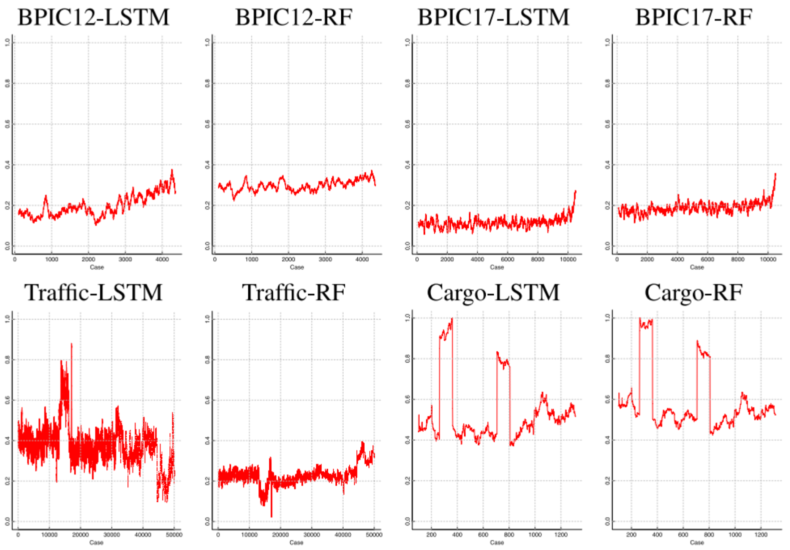
Average Prediction Accuracy: LSTM, RF
% of traces reaching prefix length *j*

AI for Prescriptive Monitoring

Challenge 2: Concept drift

- Process “behavior” may change over time
  - E.g., due to changes in process environment

- Prediction accuracy may fluctuate
  - E.g., if prediction models are presented with unseen and out-of-sample process monitoring data



Mean absolute prediction error (MAE) per case

AI for Prescriptive Monitoring

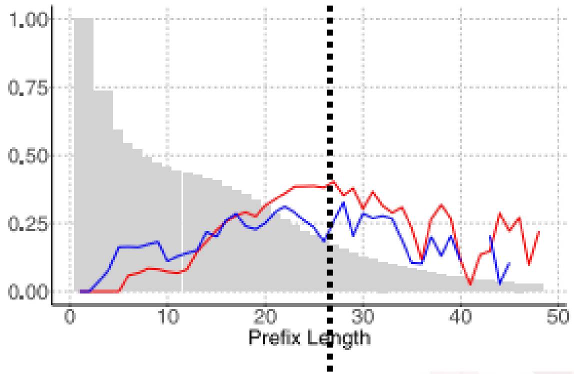
Challenge 3: Action selection / recommendation

- **Principle design choices**
  - Select from a set of predefined actions
  - Select and fine-tune action templates
  - Synthesize / generate new actions at run-time

AI for Prescriptive Monitoring

Baseline Technique: Static Adaptation Decision

- Use average prediction accuracy to determine checkpoint j_{fix} → **Challenge 1**
  - j_{fix} = earliest prediction point with highest average accuracy



$J_{\text{fix}} = 27$

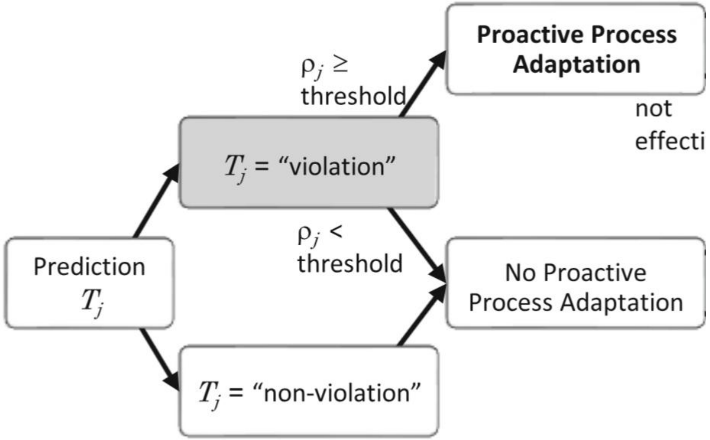
- **Con**
  - Requires testing phase during which average prediction accuracies are computed
  - No alarms will be raised for cases that are shorter than j_{fix}
  - Uses average prediction accuracy and thus does not take into account variances that might occur in the currently ongoing case.

[Metzger et al. 2019 @ CAiSE] https://doi.org/10.1007/978-3-030-21290-2_34

AI for Prescriptive Monitoring

Baseline Technique: Dynamic Adaptation Decision

- Use reliability estimate to determine which prediction to trust
- Use prediction of first checkpoint where ρ_j > threshold → **Challenge 1**



[Metzger et al. 2019 @ CAiSE] https://doi.org/10.1007/978-3-030-21290-2_34

AI for Prescriptive Monitoring

Baseline Technique: Empirical Thresholding

- Act on earliest prediction with reliability estimate > threshold
  → **Challenge 1**
- Dedicated training process to determine suitable threshold
  - Uses training data set (subset of event log)
  - Considers cost model to define adaptation costs (C_a), compensation costs (C_c) and penalty costs (C_p)

Costs $C(j)$ =	Prediction $\hat{y}_j$ = *deviation*		Prediction $\hat{y}_j$ = *no deviation*
	effective adaptation	*non-effective adaptation*	
Actual y = *deviation*	C_a	$C_a + C_p$	C_p
Actual y = *no deviation*	$C_a + C_c$	C_a	0

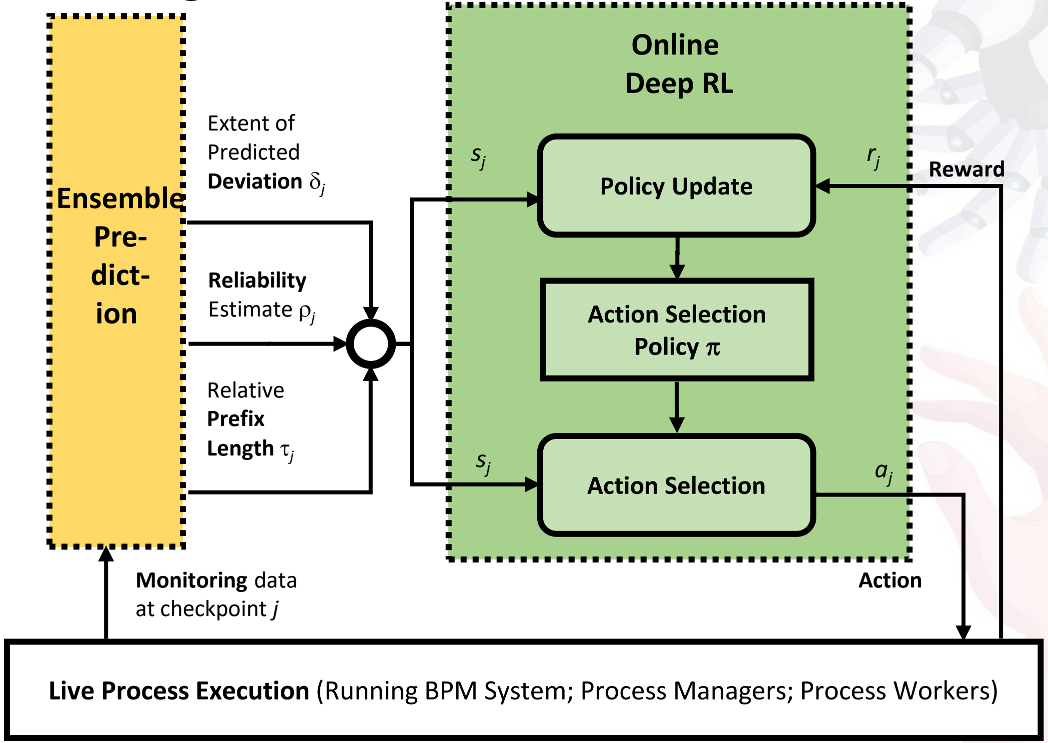
[Fahrenkrog-Petersen et al. 2002 @ Knowl. Inf. Syst.]: https://doi.org/10.1007/s10115-021-01633-w

AI for Prescriptive Monitoring

Online Deep Reinforcement Learning

- Learn action selection policy π to determine when to adapt → **Challenge 1**
  - Policy π gives action a_j in state s_j
  - Positive rewards r_j if action a_j was a good decision

- Learn π at runtime
  → **Challenge 2**



AI for Prescriptive Monitoring

Online Deep Reinforcement Learning

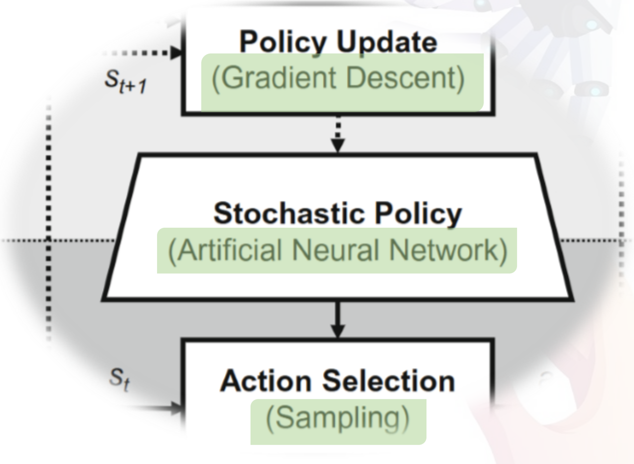
- **Balancing exploration ↔ exploitation**
  - Learn new knowledge vs leverage learned knowledge
  - Typical approach: ε-decay
  - Challenged by concept drift

- **Reward engineering**
  - Defining an effective reward function r

AI for Prescriptive Monitoring

Online Deep Reinforcement Learning

- **Policy-based Deep RL (PPO)** as RL algorithm to address **exploration ↔ exploitation**
  - Uses and optimizes parametrized stochastic action selection policy π
  - π represented as **Deep ANN**
    - Can natively handle non-stationarity and thus concept drifts → **no need to tune ε**
    - Can handle multi-dimensional, continuous state spaces
    - Generalizes well over unseen neighboring states



[Palm et al. 2020 @ CAiSE]
https://doi.org/10.1007/978-3-030-49435-3_11

AI for Prescriptive Monitoring

Online Deep Reinforcement Learning

- Reward engineering needs to consider the different contingencies:

Costs $C(j) =$	Prediction $\hat{y}_j =$ *deviation*: *effective adaptation*	Prediction $\hat{y}_j =$ *deviation*: *non-effective adaptation*	Prediction $\hat{y}_j =$ *no deviation*
Actual y = *deviation*	C_a	$C_a + C_p$	C_p
Actual y = *no deviation*	$C_a + C_c$	C_a	0
	Adaptation		No Adaptation

- To determine the rewards for the different contingencies, SOTA approaches make the following assumption:
  - "After a process adaptation, the original process outcome is still known"

[Branchi et al., 2022 @ BPM]: https://doi.org/10.1007/978-3-031-16171-1_9
[Dasht Bozorgi et al. 2023 @ InfoSys]: https://doi.org/10.1016/j.is.2023.102198

pingo: "Think-Pair-Share"

Q5: What is the problem with that assumption?

https://pingo.coactum.de/events/053187

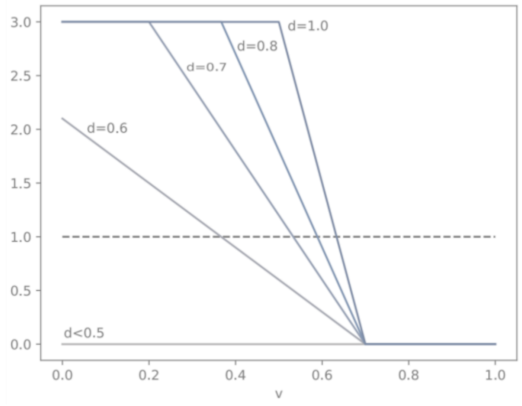
AI for Prescriptive Monitoring

Online Deep Reinforcement Learning

- **Artificial curiosity** to define rewards
  - Use *intrinsic* rewards (from within system)
    in addition to *extrinsic* rewards (from environment)

	Adaptation	No adaptation
Actual = Deviation	$R = b(1-c) - 2d$	$R = -1$
Actual = No deviation		$R = +1.5$

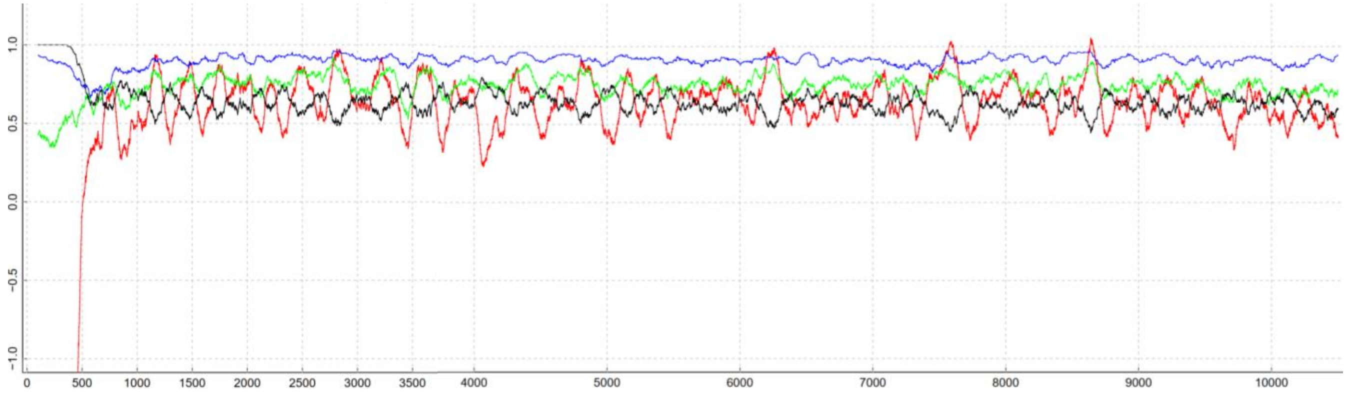
- d: rate of adaptations among last seen 30 cases
  → punishes high adaptation rates
  → rewards exploring not raising alarms
- b: decreases linearly with prefix-length
  → prefer early alarms over late alarms
- $c(d, v)$: curiosity modifier
  v = negative predictive value of
  last 100 non-adapted cases
  → high v = high accuracy in raising alarms → no longer need to explore raising alarms later
  → small d → extrinsic rewards sufficient for learning



AI for Prescriptive Monitoring

Online Deep Reinforcement Learning

Example (BPIC 2017):



red: normalized reward
blue: earliness (0 = end, 1 = beginning of process)
black: rate of alarms
green: rate of accurate alarms

pingo: "Think-Pair-Share"

Q6: What are the downsides of Online RL?

https://pingo.coactum.de/events/053187

AI for Prescriptive Monitoring

Online Deep Reinforcement Learning

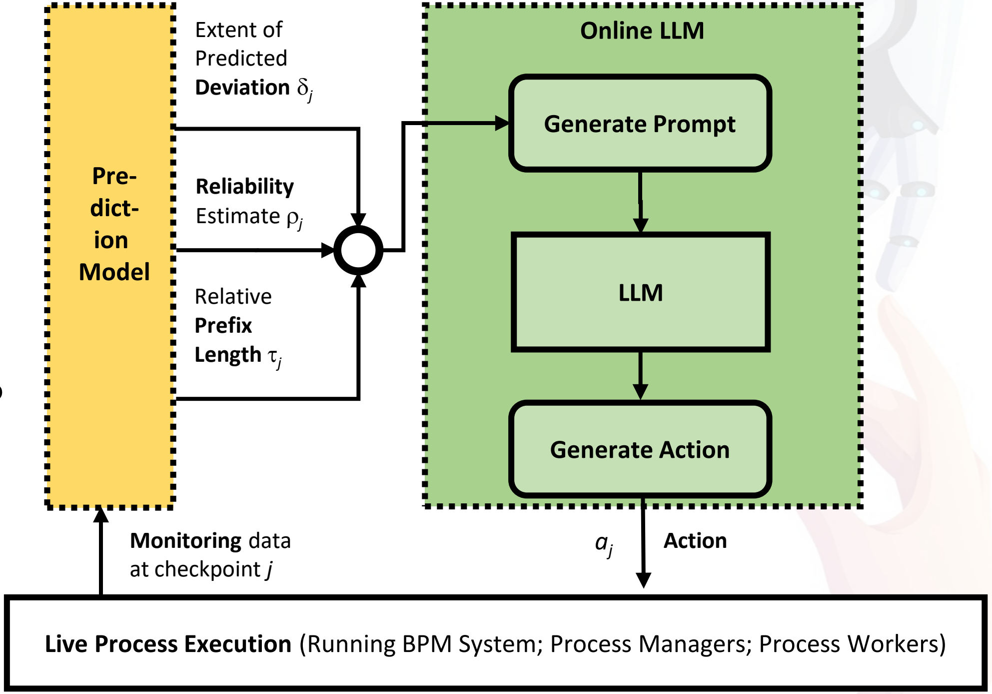
Potential directions to speed up Online RL

- Use of **Meta-RL** to reuse policies of similar learning problems
- Offline **pre-training** of RL model (e.g., using synthetic data generated from simulation models)
- **Expose RL to “important” states** determined using static analysis of simulation model
  [Mohsen et al. 2025 @ SEAMS: https://doi.org/10.1109/SEAMS66627.2025.00009]

AI for Prescriptive Monitoring

Generative AI

- Use LLM to generate adaptations at run-time
  (e.g., like in [Li et al. 2024] for adaptive systems)
  → **Challenge 3**
- **Prompt engineering**
  - Few-shot, Chain-of-Thought, RAG, …?
- **Data encoding**
  - Encoding numeric values into text?
  - Adding event labels?
- **Use of context information**
  - Consider process model?



[Li et al., 2024 @ TAAS; https://doi.org/10.1145/3686803]

AI for Prescriptive Monitoring

Empirical Study

RQ: How do the different approaches compare?

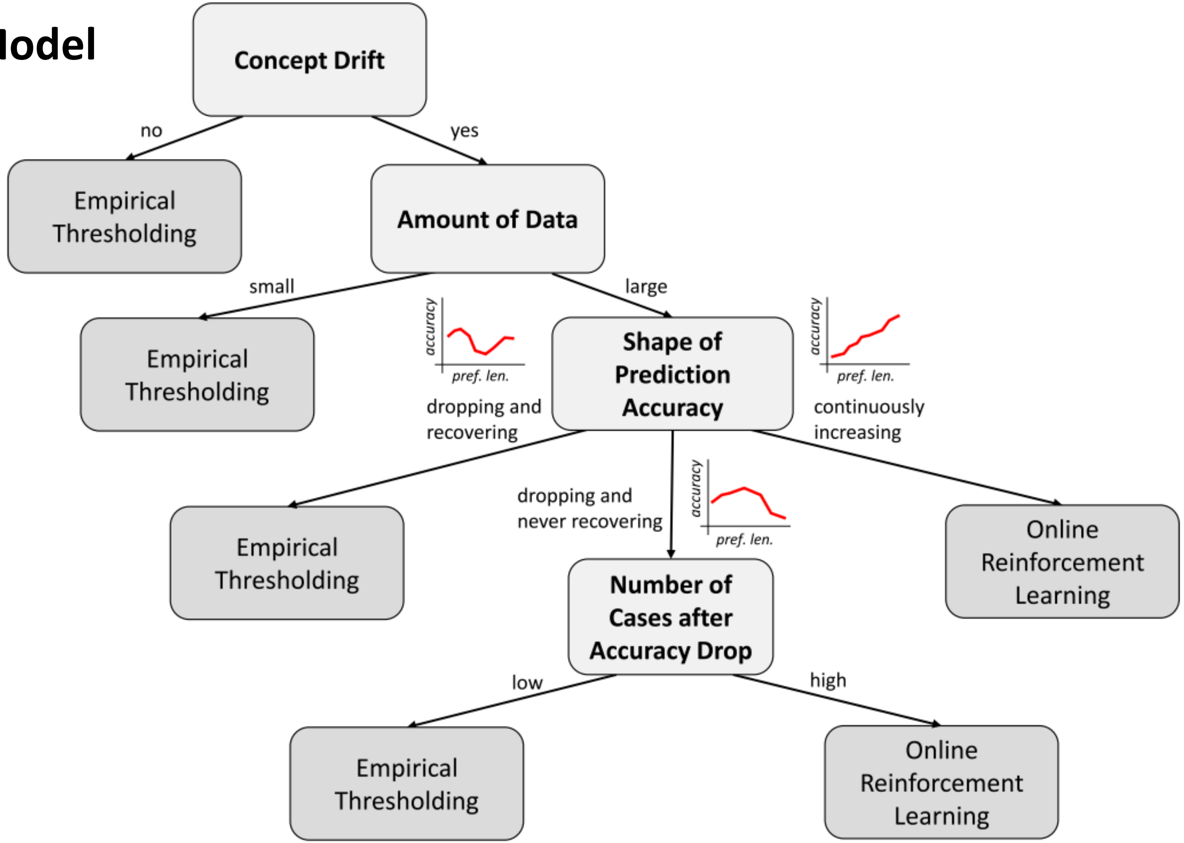
Naïve LLM baseline:
- No advanced prompt engineering (such as CoT or RAG)
- No consideration of NL data (such as event labels or types)
- No consideration of context (such as process model)

		Relative number of situations when approach performs best					Average, relative cost savings				
Data Set	**Model**	**Static**	**Dynamic**	**Empirical**	**RL**	**LLM**	**Static**	**Dynamic**	**Empirical**	**RL**	**LLM**
BPIC12	LSTM	7%	42%	29%	64%	44%	16%	28%	25%	41%	28%
BPIC17	LSTM	0%	0%	47%	53%	/	47%	51%	48%	45%	/
Traffic	LSTM	16%	22%	0%	84%	/	42%	46%	41%	38%	/
Cargo	LSTM	7%	20%	60%	33%	/	11%	26%	23%	24%	/
Average		**12%**	**21%**	**43%**	**44%**	**52%**	**29%**	**40%**	**34%**	**35%**	**38%**
							37%				

→ No single approach performs best for all data sets and cost model configurations
→ More AI-augmented techniques tend to outperform simpler approaches

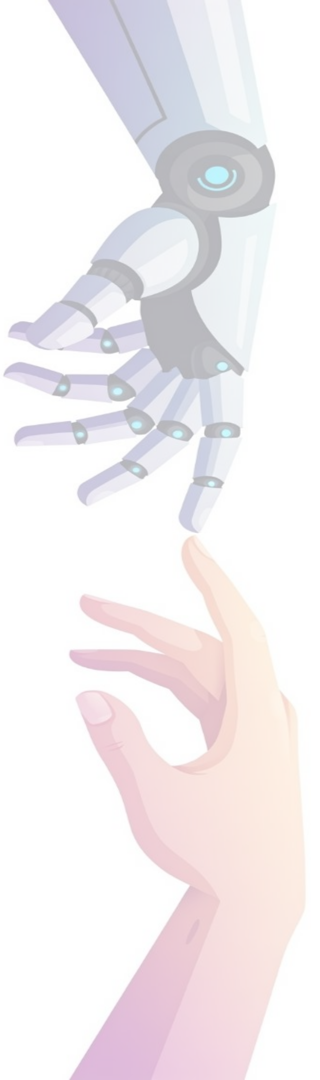
AI for Prescriptive Monitoring

Initial Decision Model



Agenda

1. Foundations
2. AI for *Predictive* Monitoring
   - Recurrent neural networks
   - Ensemble learning
3. AI for *Prescriptive* Monitoring
   - Online deep reinforcement learning
   - Generative AI
4. Future Directions



Future Directions

Generate photo of watches showing 12:00

Gemini-2.5

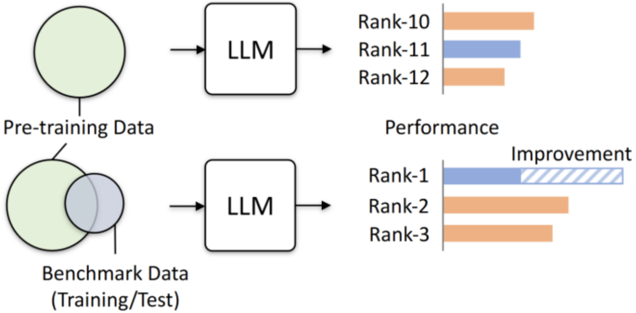
GPT-5

Challenges of Generative AI (LLMs)

- How to cope with **hallucinations** and **bias**?
  - What kind of biases of the “training data” are perpetuated in BPM?
  - What impact do hallucinations have?

- **Resource usage / costs of LLMs**
  - How to perform cost-benefit analysis?

- How to avoid **data leakage/ data pollution**?
  - How to retrieve suitable evaluation data?

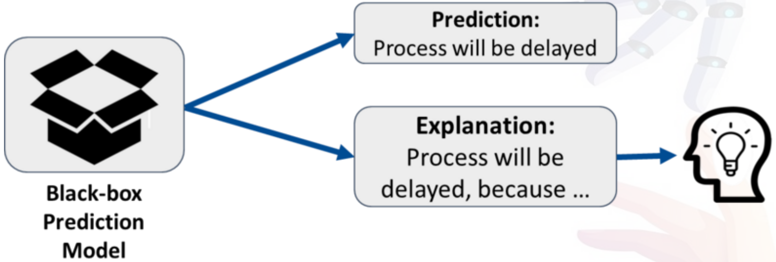


[Zhou et al. 2023 @ ArXiV: https://doi.org/10.48550/arXiv.2311.01964]

Future Directions

Explainable **Process Monitoring**

- Addressed Concerns:
  - **Trust**: Understanding the **'why'** builds confidence
  - **Debugging**: Identifying failures and performance issues becomes possible
  - **Accountability**: Assigning responsibility and implementing corrective actions
  - **Bias Mitigation**: Detecting and mitigating discriminatory outcomes.
  - **Compliance**: Meeting transparency demands of regulatory frameworks



- But: Current **XAI Limitations**:
  - **Fail to capture BPM specifics** (process constraints, contextual richness, causal dependencies, human interpretability)

[Fettke et al. 2025 @ PMAI-ECAI: https://doi.org/10.48550/arXiv.2507.23269]
[Kubrak et al. 2024 @ BPM: https://doi.org/10.1007/978-3-031-70396-6_23]

Future Directions

Agentic **Process Monitoring**

- Agent realized via AI
  - Operates with a greater degree of autonomy
  - Capable of undertaking roles
  - Manages multi-step tasks
  - Achieves higher-level goals

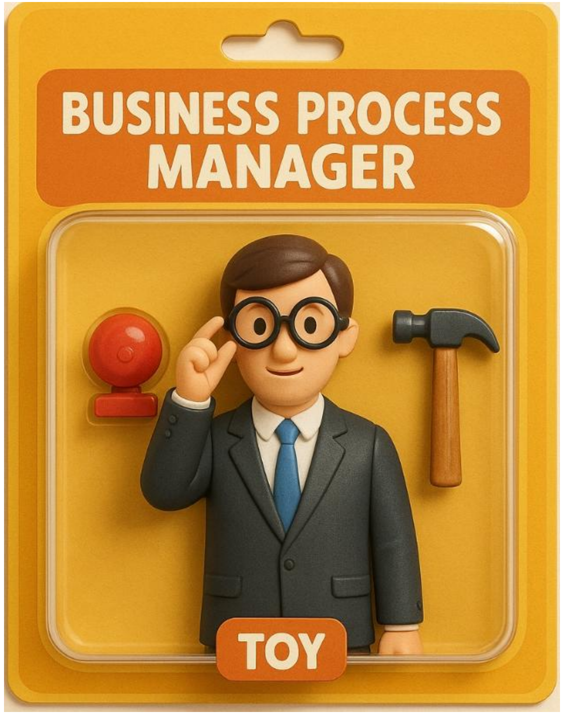
  - Proactively collaborates with human developers or other agents

[Vu et al. 2025 @ Responsible BPM]: https://doi.org/10.48550/arXiv.2504.03693

	HAS H1	HAS H2	HAS H3	HAS H4	HAS H5
Team Dynamics	**AI Agent Drives Task Completion** The AI agent takes primary reponsibility for task execution with no or minimal human oversight.		**Equal Partnership** The human and the AI agent collaborate closely throughout the task.	**Human Drives Task Completion** The human takes primary responsibility for task execution with varying levels of AI assistance.	
Required Human Involvement	AI agent handles the task entirely on its own without your involvement.	AI agent needs your input at a few key points to achieve better task performance.	AI agent and you work together to outperform either alone.	AI agent needs your input to successfully complete the task.	Task completion fully relies on your involvement.
AI Role	**Automation** AI replaces human capabilities		**Augmentation** AI enhances human capabilities		
Example Tasks	• Transcribe data to worksheets and enter data into computer. • Run monthly network reports.	• Devise trading, option, or hedge strategies. • Accept payment on accounts.	• Create core game features, including storylines, role-play mechanics, etc. • Compile and analyze experimental data and adjust experimental designs as necessary.	• Coordinate and direct the financial planning, budgeting, procurement, or investment activities. • Design, plan, organize, or direct orientation and training programs.	• Participate in online forums or conferences to stay abreast of online retailing trends, techniques, or security threats.

[https://futureofwork.saltlab.stanford.edu/]

Thank You!



Research leading to these results received funding from the
EU's Horizon 2020 research and innovation programme
under grant agreements no.
731932 – **TransformingTransport,** 732630 – **BDVe,** 780351 – **ENACT,**
871493 – **DataPorts,** 101070455 – **DYNABIC**

Q7: How do you rate the tutorial?

https://pingo.coactum.de/events/053187