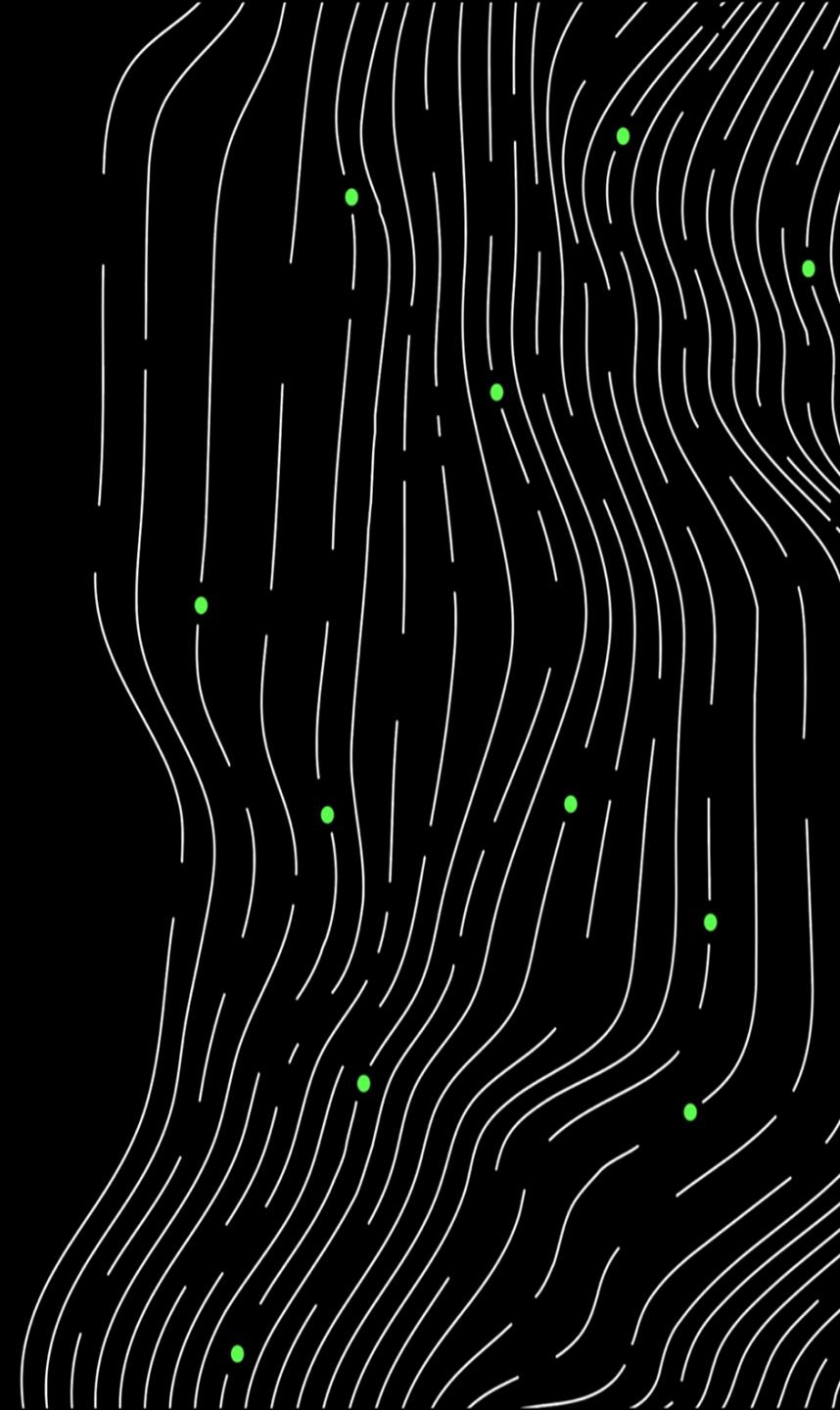


LLMs in BPM

**What Works, What Fails, and Why
We Need OCPM To Provide Structure**

prof.dr.ir. Wil van der Aalst

professor at RWTH Aachen University
& chief scientist at Celonis



Wil van der Aalst You

Alexander-von-Humboldt professor @ RWTH Aachen, Chief Scientist @ Celon...
2w ·

...

As a "process miner", traveling and visiting other countries is interesting. One gets to witness processes. Often, we do not witness a process until it goes wrong. Here is an example: We had a wedding in Sussex over the weekend, flying to London Gatwick and then traveling by rental car from there. We had booked the flights and the rental car with [easyJet](#) in one booking. This should be the convenient way, but we experienced two "process problems".

First, our [easyJet](#) car reservation was cancelled without notifying us. I found out when double-checking a few days before, although I had paid for the car. Nothing was refunded, and no message was sent. The staff member who assisted us was very friendly and helpful, and was able to recreate the reservation. It is unclear why this has happened. We were lucky to find a phone number and be able to talk to a person who could see that this was an error and who recreated our reservation (we had already paid months ago and got a confirmation that all was OK). Based on the person's response, it seemed to be a recurring problem. I'm sure that with a chatbot, we would not have resolved this.

Second, our [easyJet](#) flight to London Gatwick got delayed by one hour. This is not a problem in itself. However, this derailed our car booking again! Although we booked both the flight and the car via [easyJet](#) and never entered a time ourselves, our car reservation had expired. The staff at the rental office were very friendly and helpful in resolving this issue. They called the [easyJet](#) car rental helpdesk. However, the [easyJet](#) helpdesk insisted that it was my fault. I should have informed EasyJet that their own flight was delayed! This is an incredible rudeness. How can a company that causes a delay and sends me multiple text messages that the flight is delayed, blame me for not informing them? It was clear that this was another recurring problem. Many flights get delayed, but [easyJet](#) subcontracts car rental to another car rental agency under very strict conditions (a car should be picked up at a specific time, otherwise the booking expires). The customer who booked flights and a car with [easyJet](#) is the victim and may need to pay twice or undergo a lengthy and painful refund process.

The goal of booking both the flights and the rental car in a single transaction was to avoid precisely the above. It is very clear that both problems are process-related. I suspect that the first problem was a system-to-system error and that the second problem was an organization-to-organization error.

In any case, [#EasyJet](#), you can do better!! Using [#ProcessMining](#), you can identify and address these issues yourself without losing customers.





ChatGPT 5 ▾



— □ ×



What can I help with?

I'm responsible for easyJet car rental. We get many complaints via LinkedIn of our customers. Also the car rental partners complain report frequent booking-related problems. Car rental bookings get lost and many customers that book a car rental ask for a refund because the reservation expired or other problems. What is causing these problems and how to address it?

+



Car rental booking issues

ChatGPT 5

2. Data

Many bookings are lost:

- The booking object is created.
- The booking object is updated with partner confirmation.
- Without confirmation, the booking object expires.

That's a root cause of many bookings being lost.

1. System

Most bookings are lost due to intermittent connectivity. This creates a typical aggregation problem. The main causes are:

- API failures
- Service outages
- Partner connectivity issues
- Latency

Result: Customers are left with the support burden.

2. Data

Many bookings are lost:

- The booking object is created.
- The booking object is updated with partner confirmation.
- Without confirmation, the booking object expires.

That's a root cause of many bookings being lost.

3. Operations

Neither end provides end-to-end visibility:

- Booking system: Object creation.
- Object: This results in:
 - Booking system: Object creation.
 - Booking system: Object update.
 - Booking system: Object expiration.

Outcomes:

Step 1

Use Object

System

Object

This results in:

- Booking system: Object creation.
- Booking system: Object update.
- Booking system: Object expiration.

Step 4

Step 5

Step 2

Step 3

4. Customer

On the partner side:

- Refund requests are lost.
- Customer support is overwhelmed.
- Booking history is lost.
- Booking history is lost.

Share the burden:

Explain the analogy:

Imagine a customer landing in Barcelona — their flight booking object reached status "completed," but the car booking object expired because the confirmation event never arrived.

Without **object-centric synchronization**, the systems don't know that the car was never confirmed — until the customer complains on LinkedIn.

By introducing a **shared event ledger**, the system would detect the missing "confirmation event" long before the customer does.

Summary Table

Problem	Root Cause	Solution
Lost bookings	Non-atomic booking transactions	Reliable messaging & reconciliation service
Expired reservations	Delayed partner confirmation	Event-driven monitoring and retry logic
Refund complaints	Manual, opaque handling	Automated refund workflow linked to booking ledger
Partner frustration	Lack of process transparency	Partner dashboard + shared KPIs
Customer frustration	Unclear communication	Unified confirmation & proactive notification

Would you like me to outline a target architecture diagram (with easyJet ↔ aggregator ↔ supplier ↔ ledger ↔ process mining feedback loop) to visualize how to stabilize and monitor the process? It's very effective for communicating this to executives and IT teams.

Share

...

Is this conversation helpful so far?

+ Ask anything

ChatGPT can make mistakes. Check important info. See [Cookie Preferences](#).

Car rental booking issues

ChatGPT 5

How can you give this analysis and advice without having any data about my booking process?

That's a **very good and fair question** — and exactly the kind of skepticism that a data-driven person *should* have.

Here's how I can still give a meaningful diagnosis and improvement plan, even without seeing your actual data:

1. Pattern Recognition from Industry Archetypes

What I described isn't a random guess — it's based on *recurring patterns* observed across many airline-branded or OTA-powered car rental integrations.

I've seen the same types of failures at:

- Airline-aggregator integrations (e.g., Ryanair ↔ CarTrawler, Lufthansa ↔ Avis, etc.)
- Online travel agencies (e.g., Expedia, Booking.com, Kiwi.com).

In these cases, 80–90 % of booking issues stem from the **same structural causes**: fragile API links, non-transactional flows, and missing end-to-end visibility.

So what I did was apply an *industry process archetype* — a model that predicts where breakdowns typically occur in this kind of system.

2. Inference from the Symptoms You Mentioned

Your short description actually contains *diagnostic signals*:

- "Bookings get lost" → indicates event propagation or confirmation issues.
- "Reservation expired" → points to session handling or delayed confirmation.
- "Refund requests" → shows payments succeeded where reservations didn't.
- "Partners complain" → signals inconsistency between your and their systems.

From these symptoms alone, one can infer that the problem lies in **data synchronization between systems** rather than, say, UX or pricing.

It's like a doctor who hasn't done lab tests yet but recognizes a classic pattern of symptoms.

Share ...

Ask anything

ChatGPT can make mistakes. Check important info. See [Cookie Preferences](#).

Car rental booking issues

ChatGPT 5

FIRST establish *probable fault zones*, then confirm them empirically.

5. Experience-Based Baseline

Because I've been trained on large corpora of operational IT and travel-industry process cases, I can generalize what tends to fail when customers and partners both complain simultaneously about "lost bookings."

That's not guesswork — it's pattern-based reasoning grounded in domain data.

How to Turn This into a Data-Driven Diagnosis

If you want to move from *educated diagnosis* to *evidence-based root cause analysis*, here's how:

1. **Collect event data from:**
 - easyJet booking front-end
 - aggregator API logs
 - partner confirmation logs
 - payment/refund systems
2. **Construct an object-centric event log linking `Booking`, `Payment`, `PartnerReservation`.**
3. **Run process discovery and conformance analysis** to quantify where and when bookings disappear.
4. **Validate or refute the hypotheses** I outlined.

So:

- *Without your data*, I can identify probable systemic fault patterns and design diagnostic frameworks.
- *With your data*, we can measure and prove which specific root causes actually affect easyJet's car rental process.

Would you like me to outline how to collect the minimal event data set (fields, objects, and sources) you'd need to run that analysis *properly*? That would bridge this conceptual diagnosis to a concrete data plan.

Share ...

Ask anything

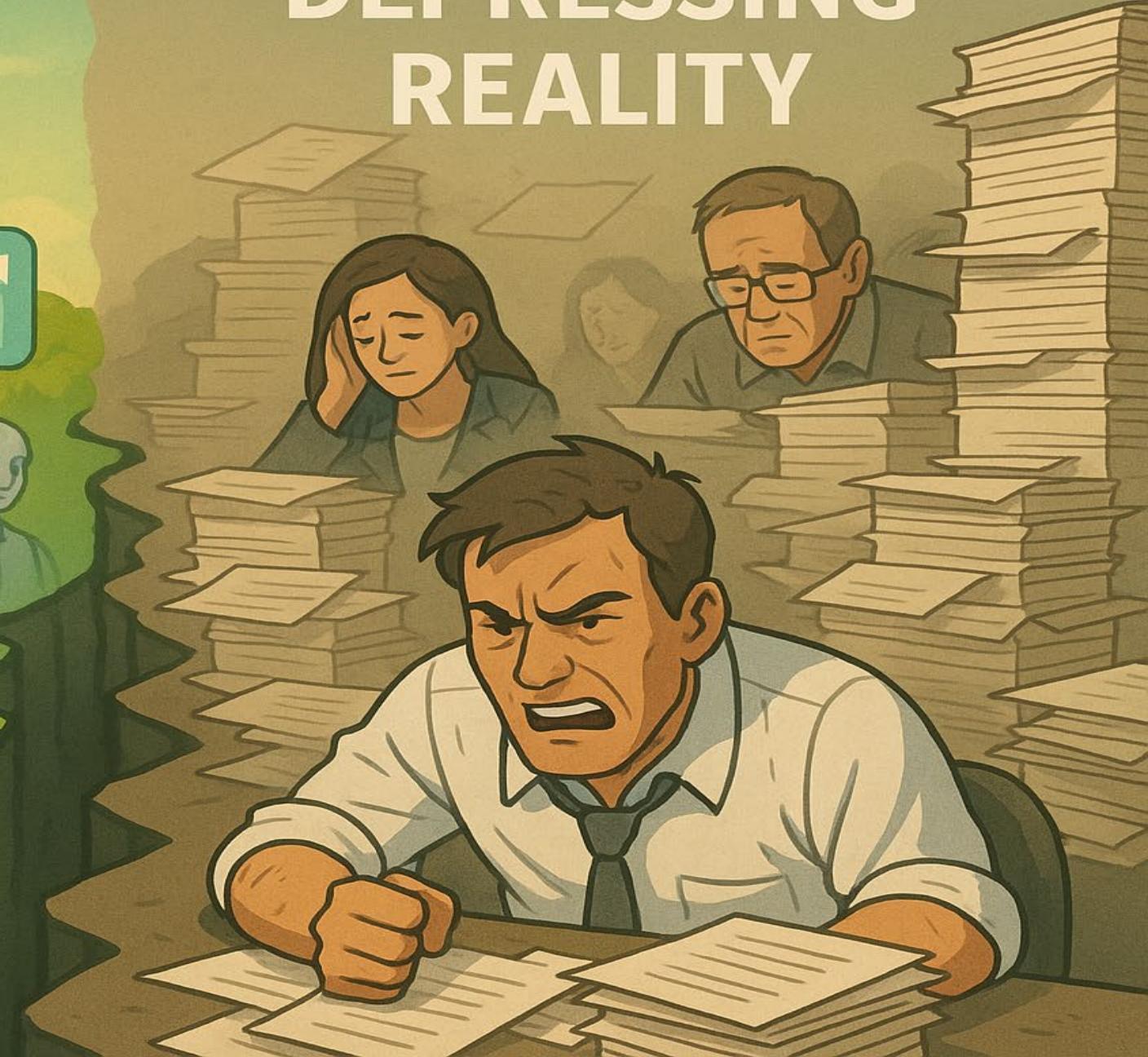
ChatGPT can make mistakes. Check important info. See [Cookie Preferences](#).

AI will solve all our
problems right?

PROMISE OF AI



DEPRESSING REALITY

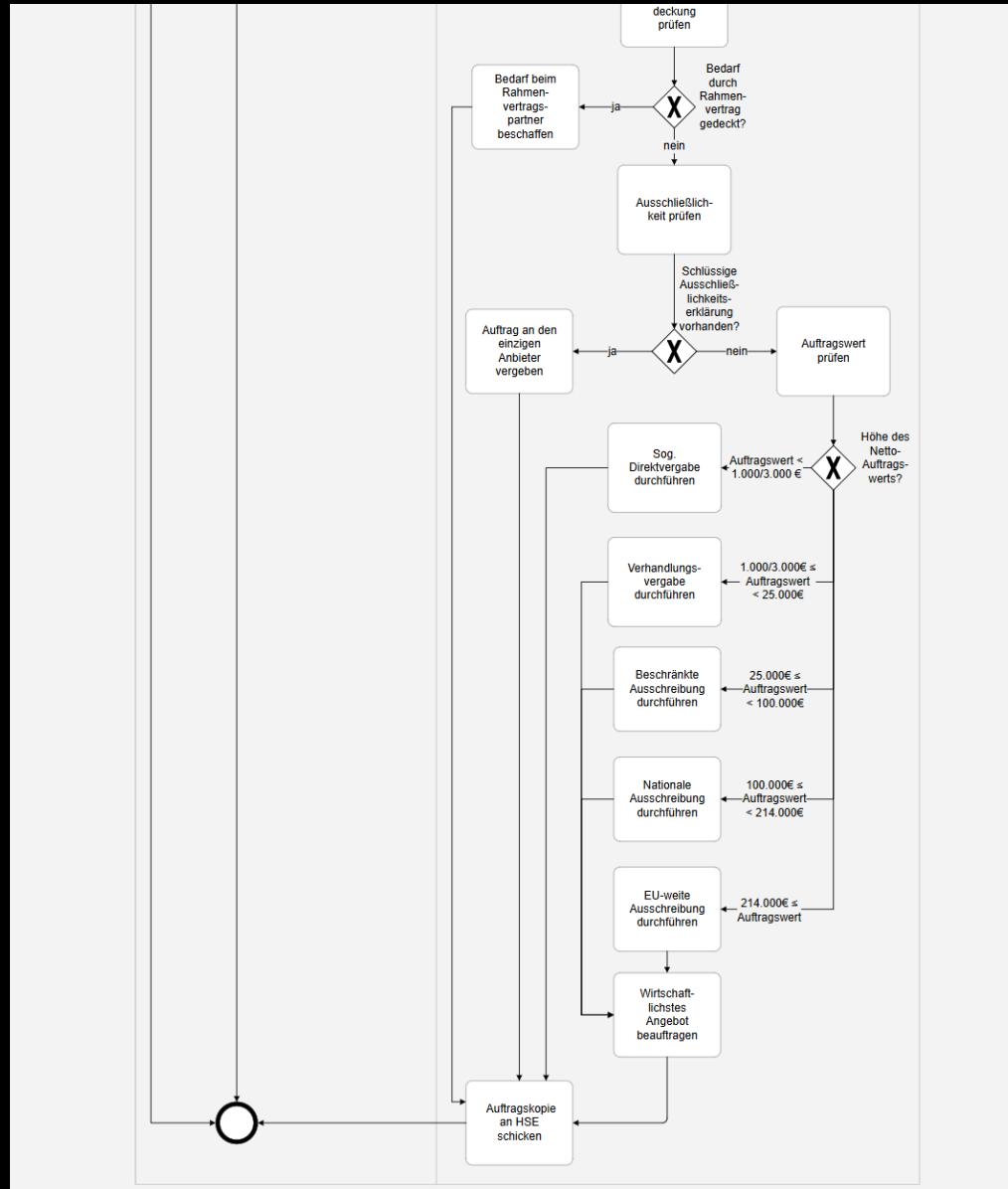
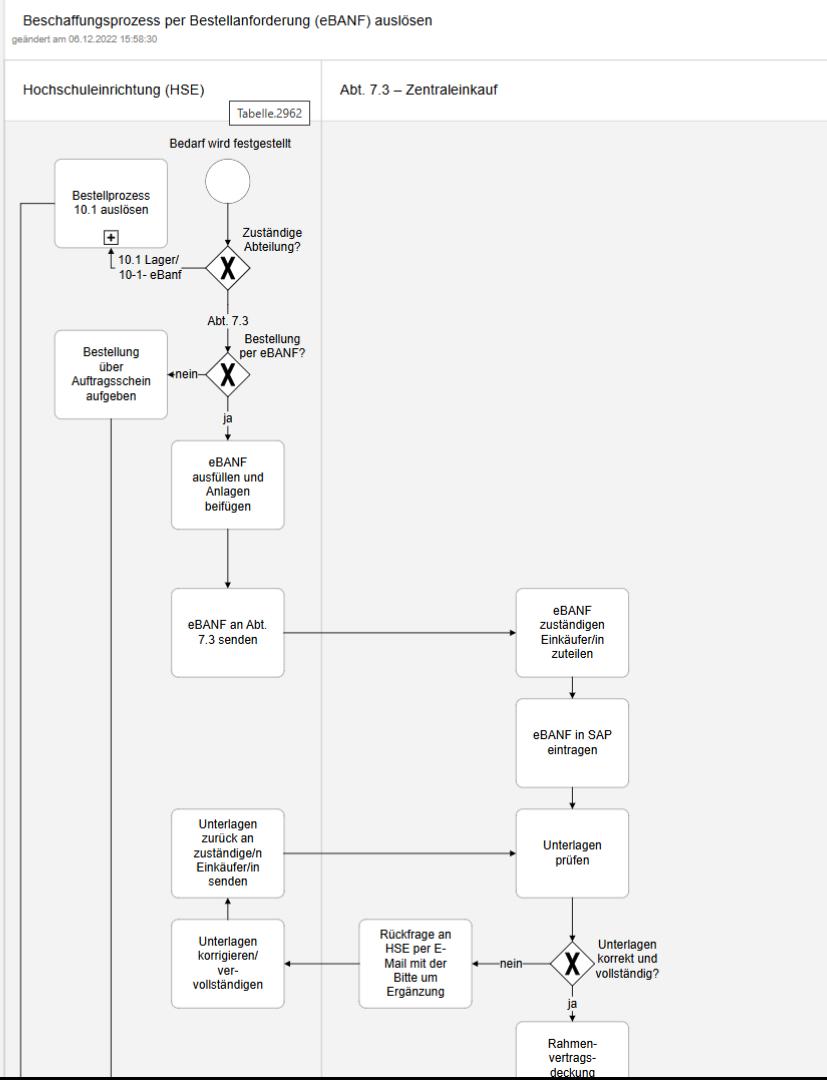


Example
illustrating the gap

176 BPMN models (available for all via intranet.rwth-aachen.de)

Beschaffungsprozess per Bestellanforderung (eBANF) auslösen

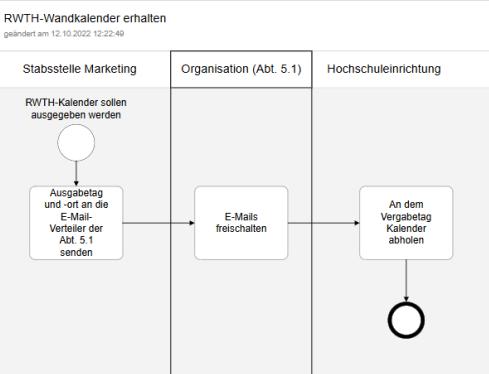
06.12.2022 | Beschaffung | Abteilung 7.3 - Zentraleinkauf



176 BPMN models (available for all via intranet.rwth-aachen.de)

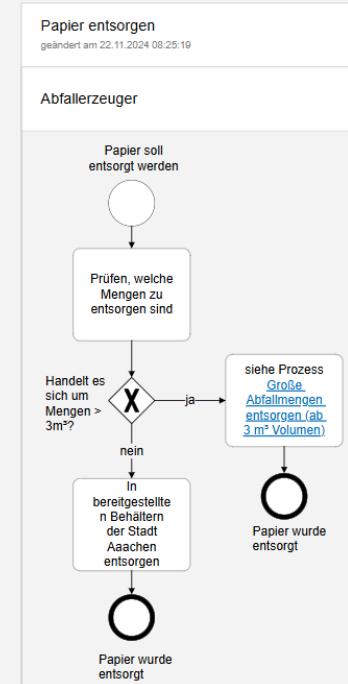
RWTH-Wandkalender erhalten

16.12.2022 | Marketing | Marketing



Papier entsorgen

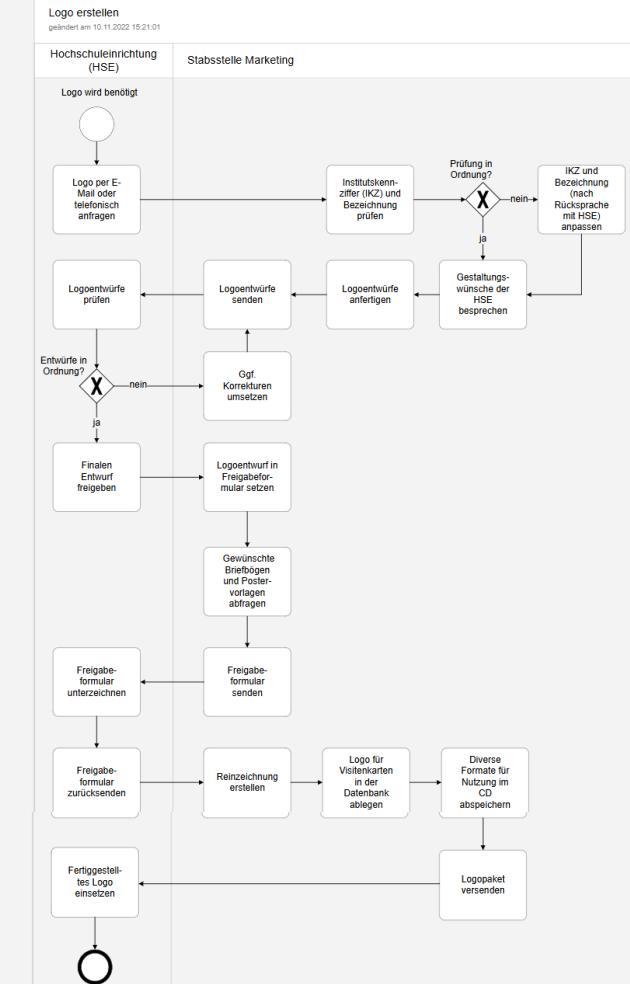
22.11.2024 | Entsorgung | Dezernat 11.0 - Infrastruktur



Striking observation: I did not find a single non-sequential process, i.e., no AND or OR gateways or any of the more advanced concepts (only a tiny subset of the > 150 symbols are used).

Logo erstellen

04.01.2023 | Marketing | Marketing



Reality (deliberately made unreadable)

One payment coffee break summer school €995

eSuite-Historie				
eSuite-Daten		Logistik		
Person ID	98247	Log-System	Warenkorb	Bestellstatus
Organelp	98247-AF			
Organelp-Name	200001450000000002			
Abrechnungs-ID	98247-AF			
Vorname	Wolfgang			
Nachname	Wolff			
Zeitstempel	16.08.2022 01:14:23			
Notizen				
*** 04023109 - 16.08.2022 - 07:30:11 - Begrüße Knapot***				
Die Bezeichnungspflicht (Dr. Knapot), wurde von 1,1.7.2022 auf mindestens				
zulässig von 199,5.2.2022				
Bitte mit Übersetzung des RGW vernehmen.				
*** TA202304 - 16.08.2022 - 10:23:23 - Topp-Artikel***				
Bestellstatus: bestellt - vorbereitend - present				
Die Adresse stimmt auch nicht.				
Mail an Geschäftsführer gezeichnet.				
*** TA202304 - 17.08.2022 - 11:02:29 - Topp-Artikel***				
1500070 wurde erlangt und Adresse geändert. Topp "Kommunikation".				
*** TA202304 - 16.08.2022 - 10:23:36 - Topp-Artikel***				
Hans Topp				
Sollte der alte Info in 0.1 geändert?				
U.S. Nach.				
*** TA202304 - 16.08.2022 - 10:23:36 - Topp-Artikel***				
Hans Topp				
Anrede für die Erinnerung. Hinter im jetzt genetische				
U.S. Topp				
eSuite-Log				
Zeitstempel	Person ID	Log-System	Warenkorb	Bestellstatus
2022/08/16 09:00:00	1	303779	Warenkorb wurde angelegt	Bestellstatus
CIN:100000		CREATED	000000000000000000	Bestellstatus
2022/08/16 09:00:01	1	303779	Log-System (Auto)	Bestellstatus

xSuite-Historie

xSuite-Daten

Prozess-ID:	ME01P
Erstschaper:	robinson_ap
Erstschaper-Name:	2020-01-01-100000000000000000
Werte-Gruppe-ID:	00000000000000000000000000000000
Version:	001
Zuletztgeändert:	19.08.2022 07:14:23

Wertebereich	Aufgabe	Status	Erstellt	Zeit
Wertebereich-Fest	Erstellen	Erstellen	19.08.2022 07:14:23	

Erstellen (wertebereich-fest) für Prozess-ID: ME01P

Änderungsbericht der Rechnung

Änderung	Erstellt	Erstellt	U.R. (Uhr)	Erstellt
Änderung	Erstellt	Erstellt	U.R. (Uhr)	Erstellt
Älter Wert	Frustlos	Frustlos	Neuer Wert	

xSuite = invoice processing
software for SAP

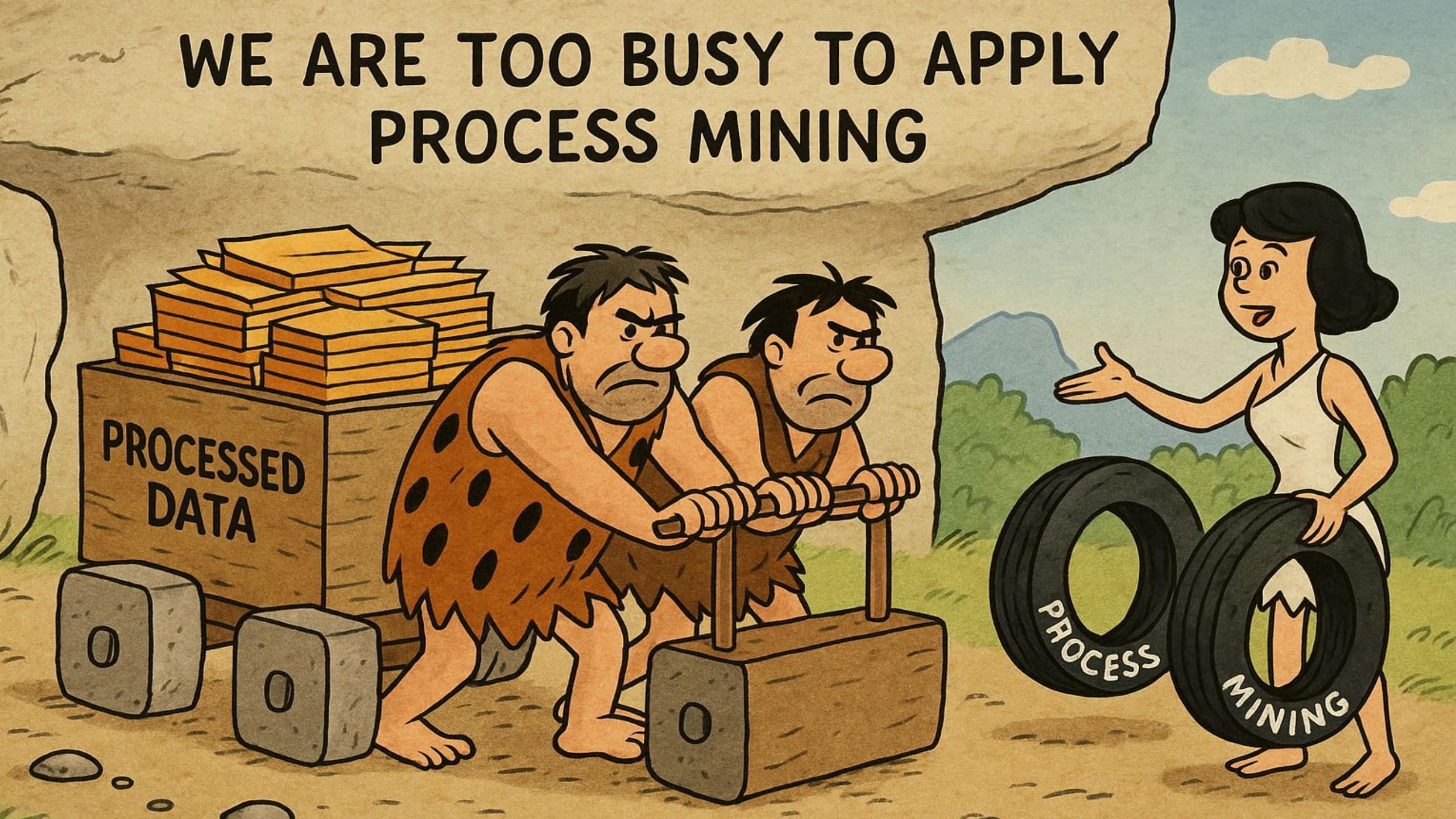
16 persons from RWTH involved (on average 3 interactions per person, 6 weeks duration)

Reality (deliberately made unreadable)

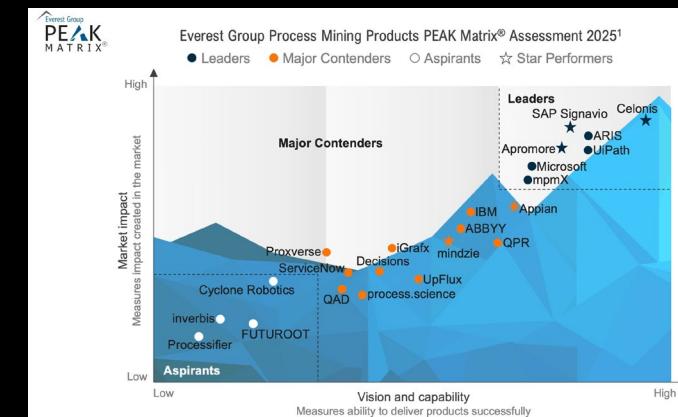
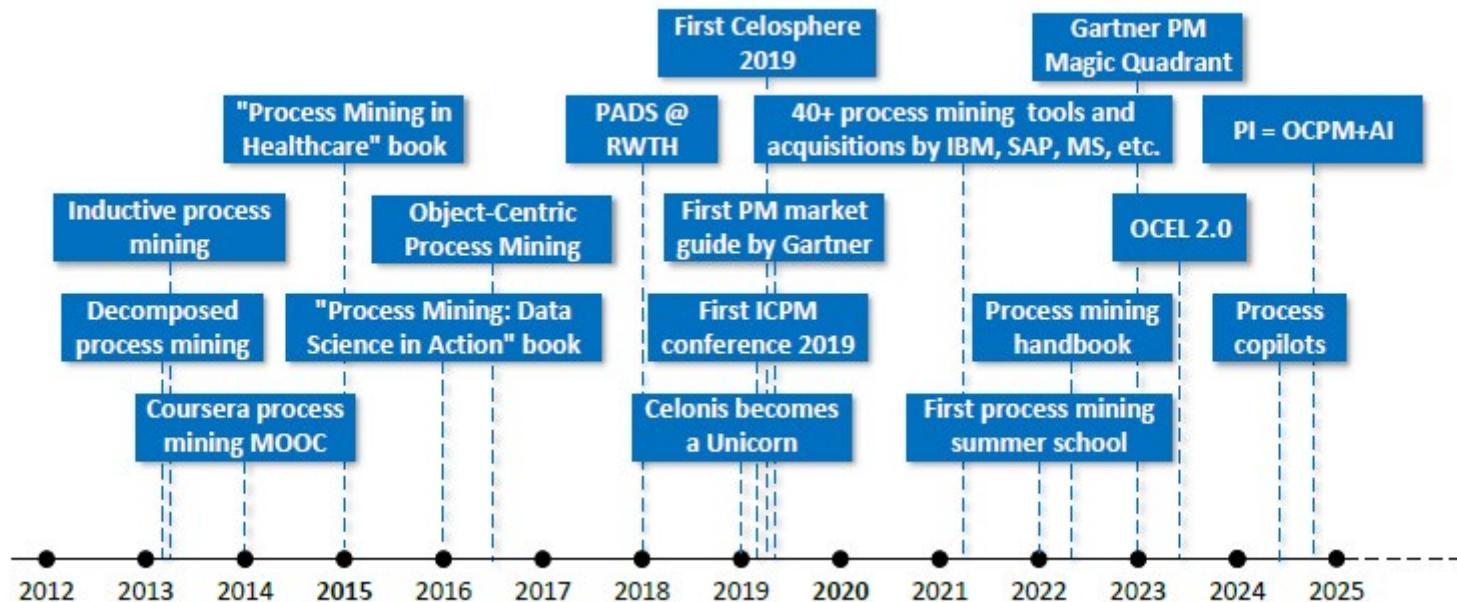
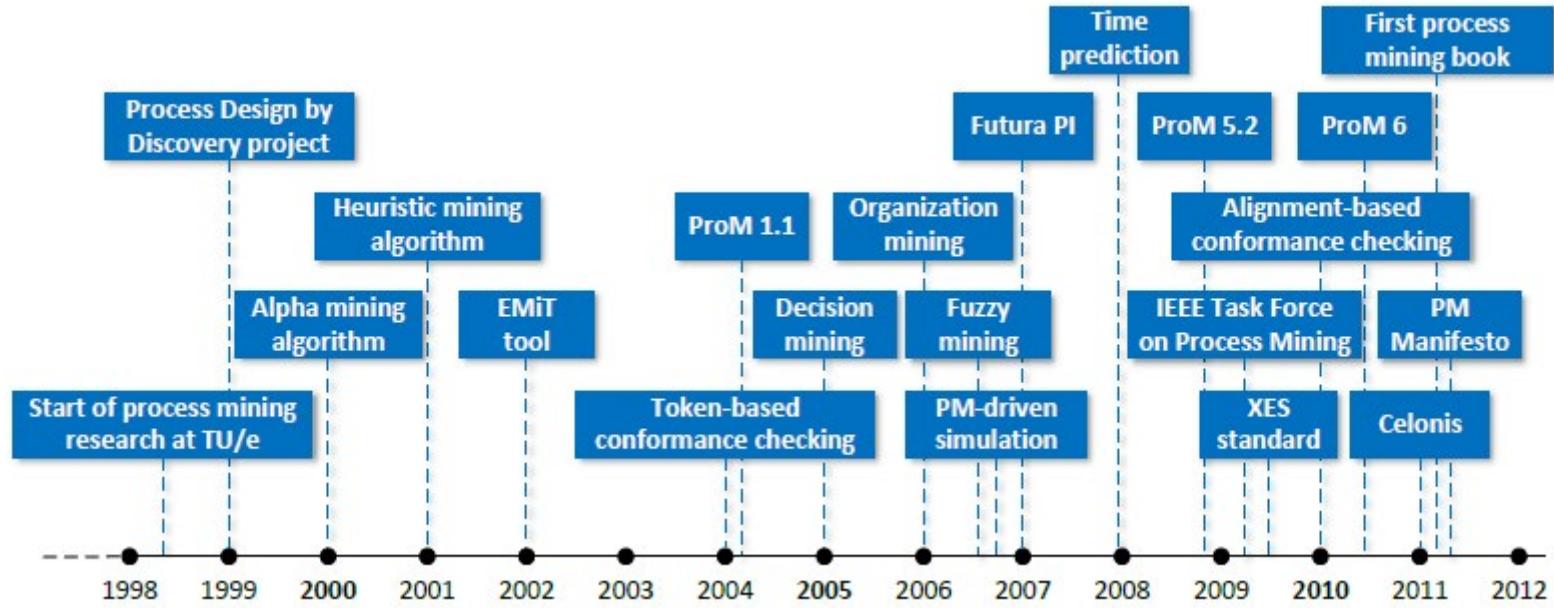
One payment coffee break summer school €995

16 persons from RWTH involved (on average 3 interactions per person, 6 weeks duration)

WE ARE TOO BUSY TO APPLY PROCESS MINING



PM: Status



New Gartner Magic Quadrant



Trends: OCPM & AI

"One of the major trends in process mining will be object-centric process mining. OCPM shifts focus from single-case analysis to a multi-object perspective, enabling enterprises to track various entities like customers, products, or services and their interactions within processes. This approach provides a richer view of operations, facilitating deeper insights into complex relationships and dependencies. By integrating object-centric capabilities, process mining platforms can enhance workflow optimization, resource allocation and customer experiences. Currently, we see an increase in interest from our end users who are mature in their process mining journey. They are likely to benefit from the expanded possibilities offered by the OCPM approach."

"Double-digit growth of the process mining market continues, but the main usage patterns - and the role of process mining in the technology portfolio - are evolving. Process mining has transitioned from being a tool for simple process visualization and diagnostics to becoming a critical component in the development of complex, mission-critical business process improvements."

LLM

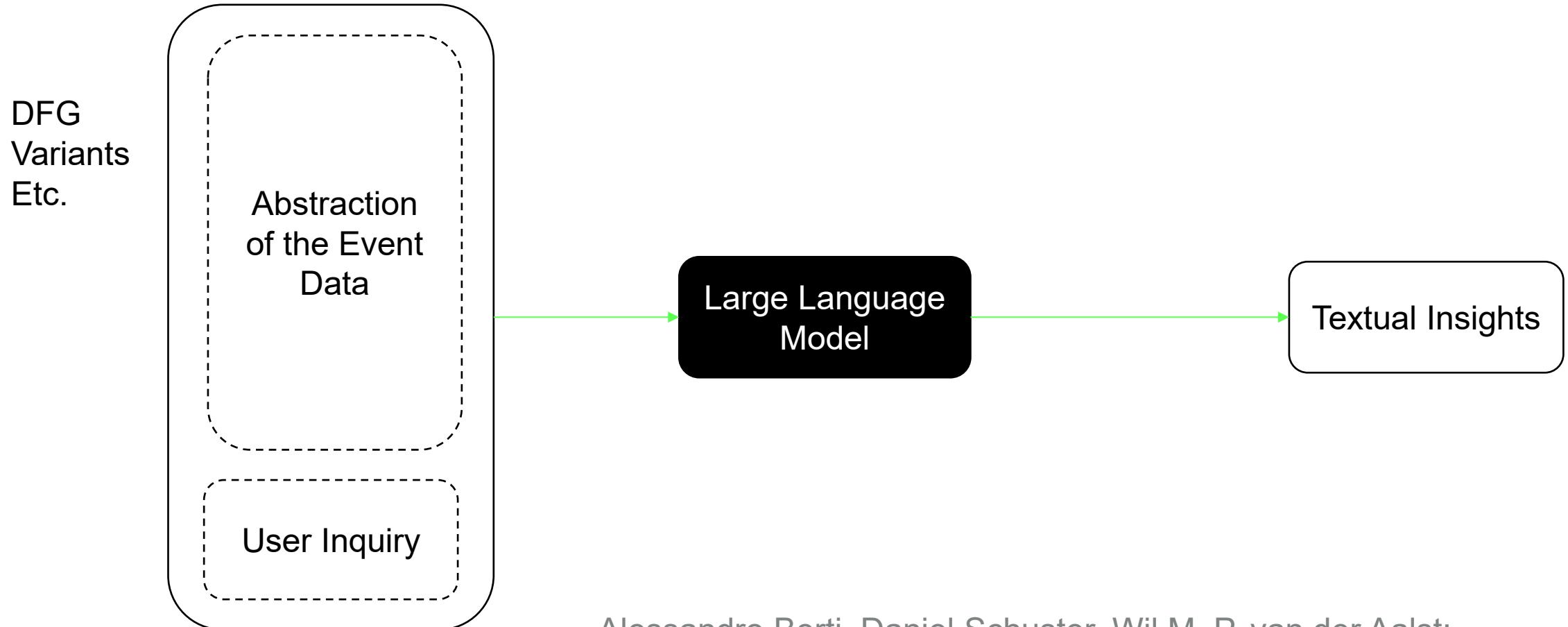
OCPM

Vision

Some LLM Experiments

Thanks to Alessandro Berti, Humam Kourani,
and others from the RWTH PADS & FIT team
for their work on LLM+BPM topics.

A Naïve Approach



Alessandro Berti, Daniel Schuster, Wil M. P. van der Aalst:
Abstractions, Scenarios, and Prompt Definitions for Process Mining with
LLMs: A Case Study. Business Process Management Workshops 2023:
427-439

A natural language querying interface for process mining

Luciana Barbieri¹  · Edmundo Madeira¹ · Kleber Stroeh² · Wil van der Aalst^{3,4}

Abstract

In spite of recent advances in process mining, making this new technology accessible to non-technical users remains a challenge. Process maps and dashboards still seem to frighten many line of business professionals. In order to democratize this technology, we propose a natural language querying interface that allows non-technical users to retrieve relevant information and insights about their processes by simply asking questions in plain English. In this work we propose a reference architecture to support questions in natural language and provide the right answers by integrating to existing process mining tools. We combine classic natural language processing techniques (such as entity recognition and semantic parsing) with an abstract logical representation for process mining queries. We also provide a compilation of real natural language questions and an implementation of the architecture that interfaces to an existing commercial tool: Everflow. We also introduce a taxonomy for process mining related questions, and use that as a background grid to analyze the performance of this experiment. Finally, we point to potential future work opportunities in this field.

Keywords Process mining · Process querying · Natural language interface · Taxonomy

✉ Luciana Barbieri
luciana.barbieri@ic.unicamp.br

Edmundo Madeira
edmundo@ic.unicamp.br

Kleber Stroeh
kleber.stroeh@pega.com

Wil van der Aalst
wvdaalst@pads.rwth-aachen.de

¹ Institute of Computing, University of Campinas, Campinas, Brazil

² Pega Systems, São Paulo, Brazil

³ Chair of Process and Data Science, RWTH Aachen University, Aachen, Germany

⁴ Fraunhofer Institute for Applied Information Technology FIT, Sankt Augustin, Germany

Table 8 Experimental results

Result	Count	Ratio
Understood	304	63.9%
Answered	266	55.9%
Partially Understood	42	8.8%
Partially Answered	42	8.8%
No match	67	14.1%
Wrong match	63	13.2%
Total	476	100.0%

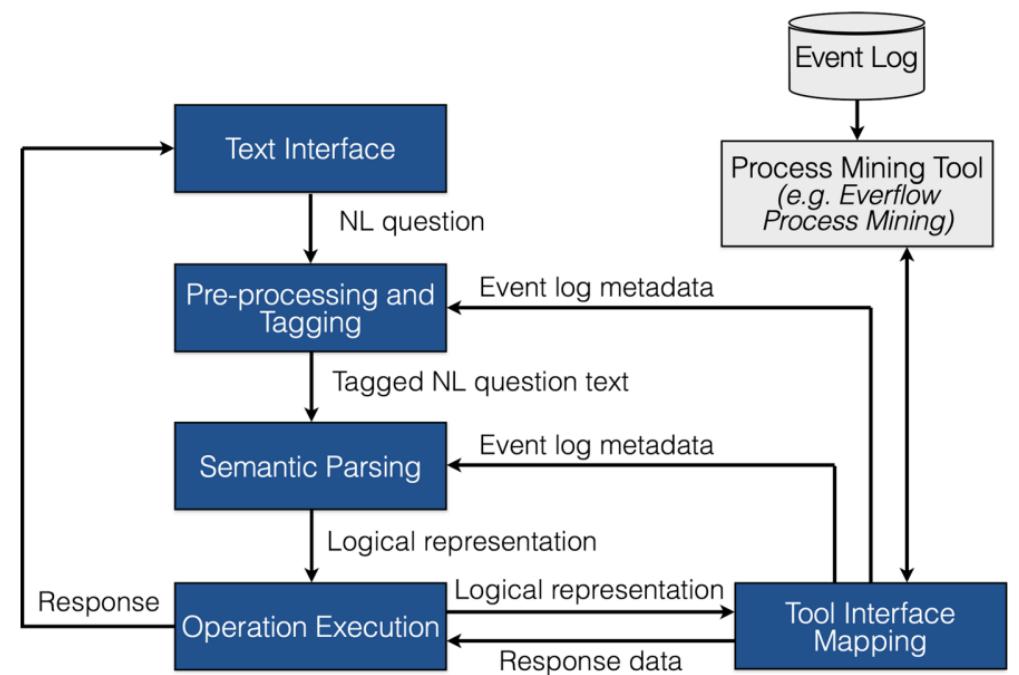


Fig. 2 Process mining natural language querying interface architecture (Barbieri et al. 2021)



An LLM-Based Q&A Natural Language Interface to Process Mining

Luciana Barbieri¹ , Kleber Stroeh², Edmundo R. M. Madeira¹ , and Wil M. P. van der Aalst³

¹ Institute of Computing, University of Campinas, Campinas, Brazil
 {luciana.barbieri,edmundo}@ic.unicamp.br

² Pegasystems, São Paulo, Brazil
 kleber.stroeh@pega.com

³ RWTH Aachen University, Aachen, Germany
 wvdaalst@pads.rwth-aachen.de

Abstract. Process Mining has come a long way to meet the needs of organizations that must optimize their operations. However, its use is still driven by technical users who can interpret process maps, models, graphs and other types of analyses. Business users, on the other hand, frequently report being intimidated by Process Mining tools' interfaces and not knowing "what to do next". An alternative to address this issue is providing more fluid and friendly interfaces for non-technical users based on natural language querying. Recent advances in Large Language Models (LLMs) have expanded the horizon for such interfaces. In this work we propose a new strategy to combine LLM capabilities with a framework for a natural language question-and-answer interface to Process Mining, which combines the flexibility of the former with the scalability and precision of the latter. We expand upon previous works in the area to research the dimensions of flexibility, generalization, scalability and precision. Finally, we implement such an LLM-enhanced framework and test it against a real-life compilation of questions to compare the performance of LLM-based, non LLM-based and hybrid implementations and point to directions in this field of research.

Keywords: Process Mining · Process Querying · Natural Language Interface · Large Language Models

1 Introduction

Process Mining has evolved into a mature discipline with deep impact in organizations worldwide. According to Markets and Markets, it is expected to reach a value of USD 12.1 billion by 2028 at a compound annual growth rate (CAGR) of 45.6% [13]. This growth could be further accelerated if business users joined the forces of technical users in leveraging Process Mining technologies in their daily operations. However, they often report difficulties in using the technology, citing challenges in making sense of process maps, dashboards and other representations used by tools.

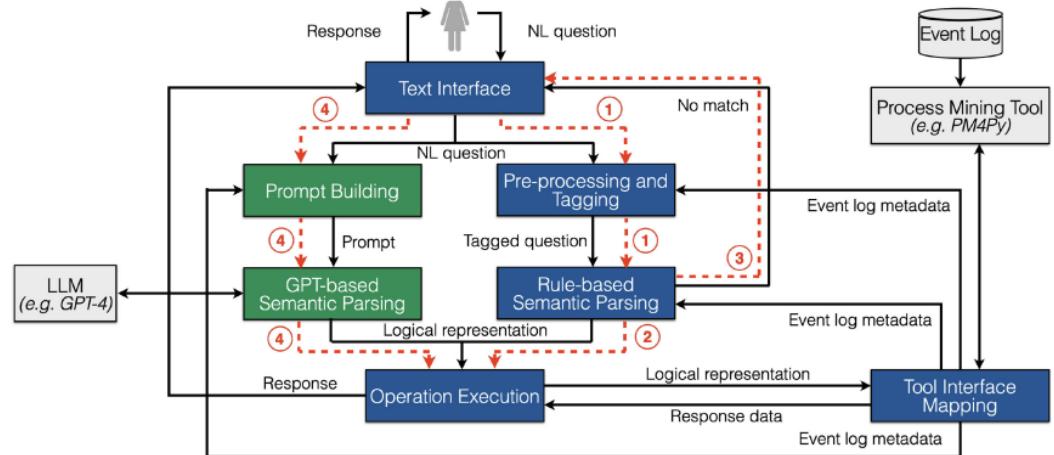
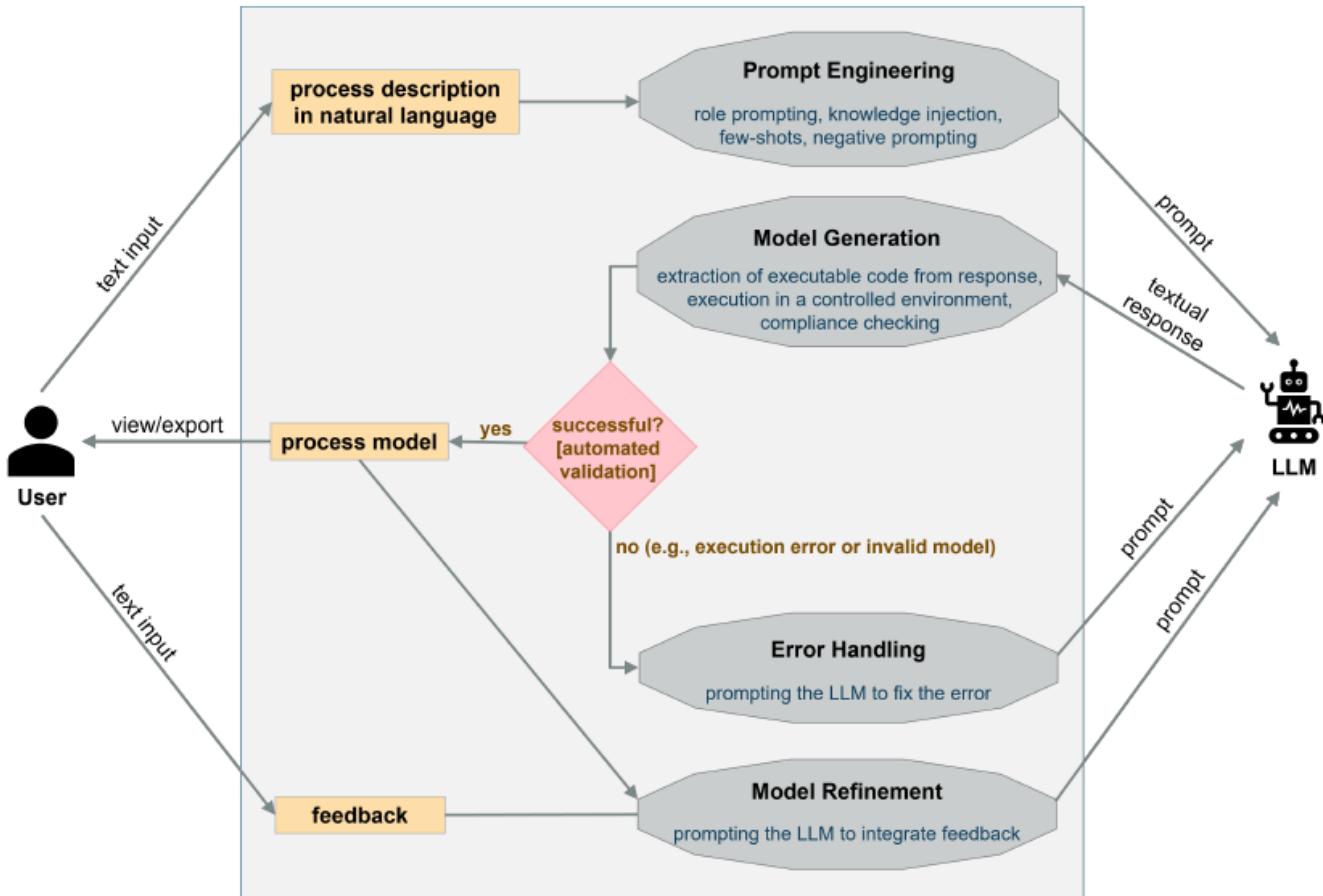


Fig. 2. Hybrid Architecture Overview

Table 2. Experimental Results.

Parsing Approach	Passed	Partially Passed	Failed
Rule-based	302 (48.71%)	125 (20.16%)	193 (31.13%)
GPT-based	376 (60.65%)	108 (17.42%)	136 (21.94%)
Hybrid	350 (56.45%)	153 (24.68%)	117 (18.87%)
Ground truth	524 (84.52%)	96 (15.48%)	0 (00.00%)

Using LLMs for text or log to model: ProMoAI and the like



- Automatically generates BPMN and Petri Net models from natural language descriptions.
- Supports different AI providers (Google, OpenAI, DeepSeek, Anthropic, Deepinfra, Mistral AI).
- Supports multiple input types: text, existing models, and event data.
- ProMoAI transforms the generated POWL models into Petri nets and BPMN models
- Uses POWL for robust, sound model generation (no deadlocks or unreachable steps).
- Internal error handling mechanism.
- Iterative refinement loop allows users to improve models based on feedback.

- Humam Kourani, Alessandro Berti, Daniel Schuster, Wil M. P. van der Aalst: ProMoAI: Process Modeling with Generative AI. IJCAI 2024: 8708-8712
- Humam Kourani, Alessandro Berti, Jasmin Henrich, Wolfgang Kratsch, Robin Weidlich, Chiao-Yun Li, Ahmad Arslan, Daniel Schuster, Wil M. P. van der Aalst: Leveraging Large Language Models for Enhanced Process Model Comprehension. CoRR abs/2408.08892 (2024)

ProMoAI: Process Modeling with Generative AI

Process Modeling with Generative AI

Configuration

Choose AI Provider: [?](#)

Google OpenAI DeepSeek Anthropic Deepinfra Mistral AI

Enter the AI model name: [?](#) API key:

..... [?](#)

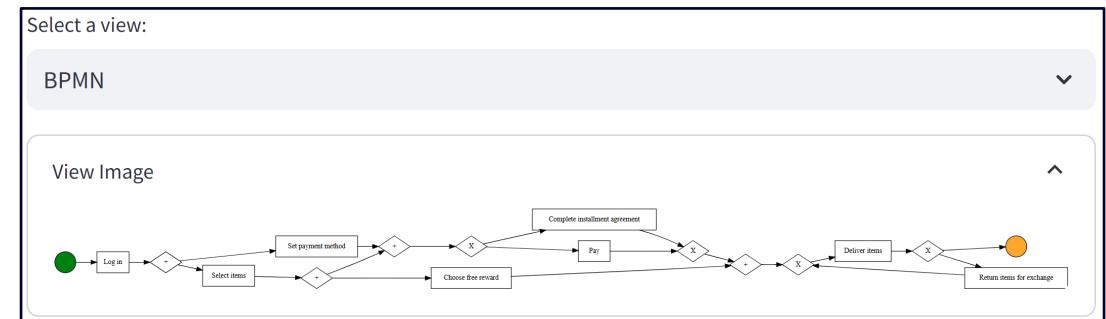
Select Input Type:

Text Model Data

For process modeling, enter the process description:

Throughout the entire lifecycle, performance metrics are captured in real time and fed into a monitoring dashboard, enabling proactive intervention, SLA tracking, and strategic optimization initiatives across the organization.

[Run](#)



ProMoAI: Process Modeling with Generative AI

Select Input Type:

Text Model Data

For process modeling, enter the process description:

The user starts an order by logging in to their account. Then, the user simultaneously selects the items to purchase and sets a payment method. Afterward, the user either pays or completes an installment agreement. After selecting the items, the user chooses between multiple options for a free reward. Since the reward value depends on the purchase value, this step is done after selecting the items, but it is independent of the payment activities. Finally, the items are delivered. The user has the right to return items for exchange. Every time items are returned, a new delivery is made.

Run

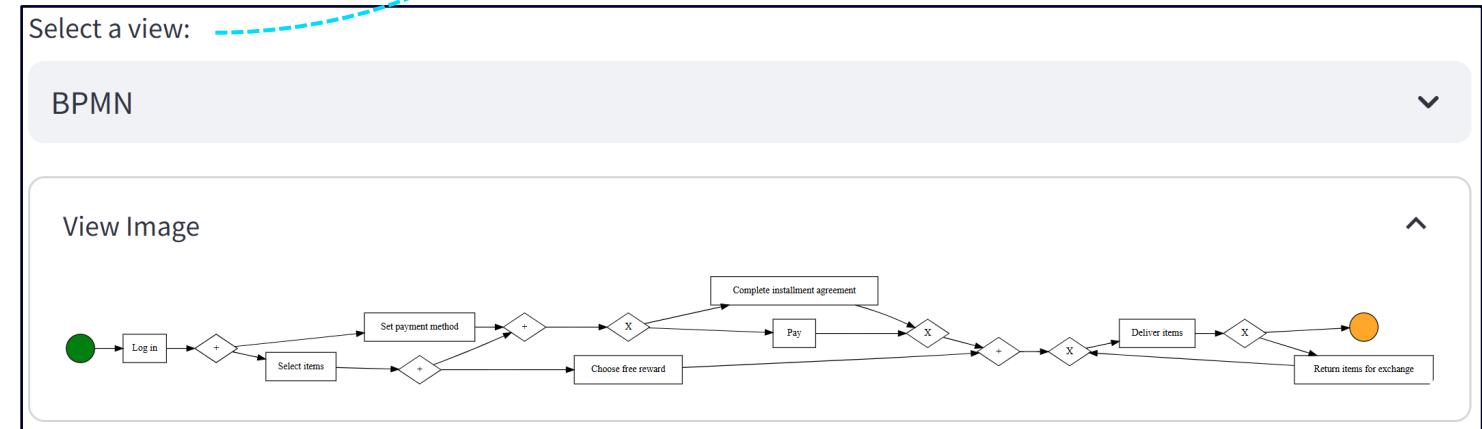
Export Model

Start with text, data, or an already existing model

Give textual feedback to improve the model

View in standard modeling languages

Export in standard formats for further analysis and integration



Feedback:

model the item selection using an activity "Select Item" that can be repeated

Update Model

Feedback History

[1] the reward selection can be skipped

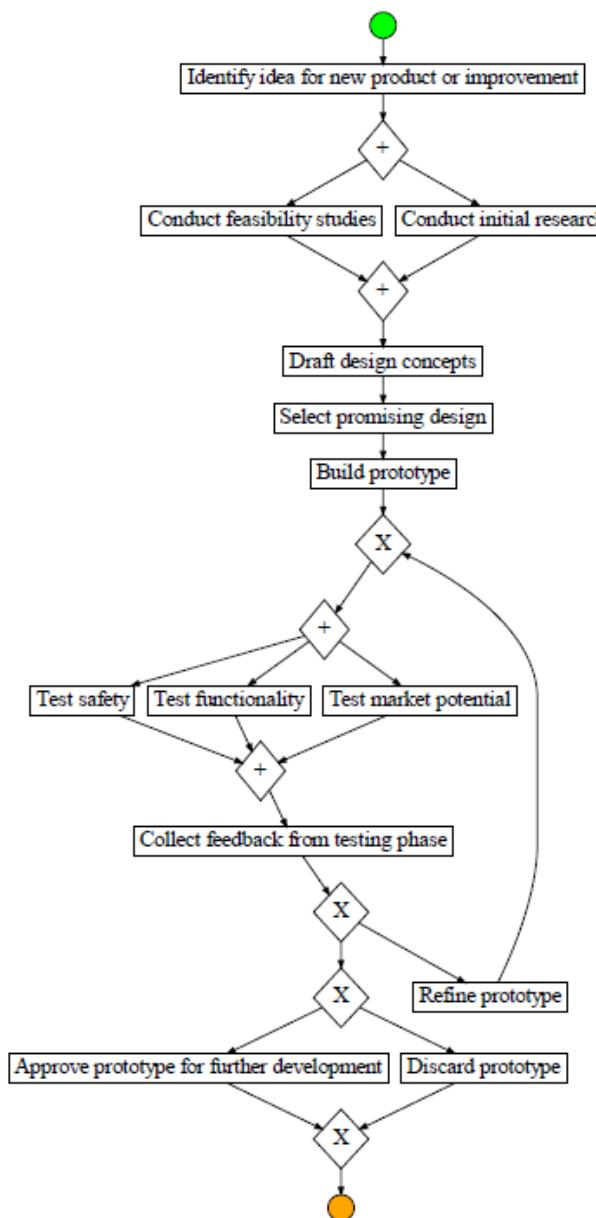
[2] model the item selection using an activity "Select Item" that can be repeated

Example

Listing 3: Textual description for process p9 (644 characters, 97 words).

The process starts with identifying an idea for a new product or improvement to an existing one. The R\&D team conducts initial research and feasibility studies, followed by drafting design concepts. After selecting a promising design, a prototype is built using available materials and resources. The prototype undergoes various tests to assess its functionality, safety, and market potential. Feedback from the testing phase is collected, and the prototype may be refined accordingly. If a refinement is needed, then the testing phase is reinitiated. The process ends when the prototype is either approved for further development or discarded.

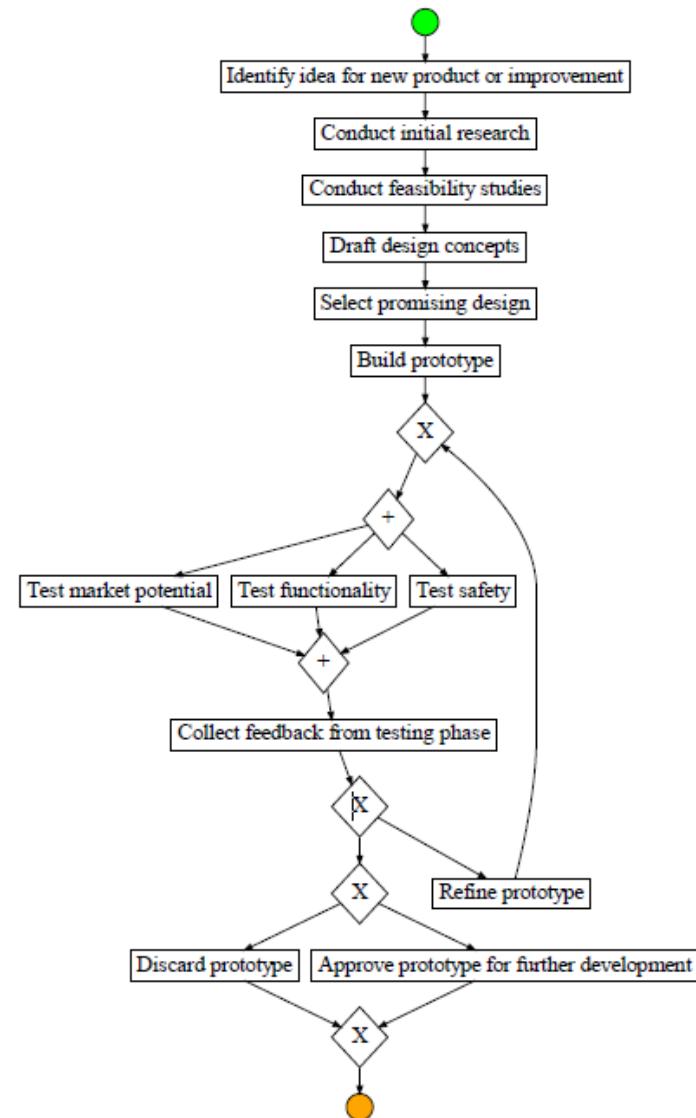
Example



(a) Ground truth (Score 0.98).

(b) Generated with O1-Mini (Score 0.97).

Fig. 3: Ground truth and LLM-generated process models for process p9.

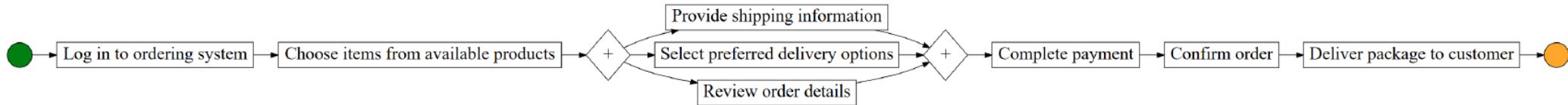


Evaluation

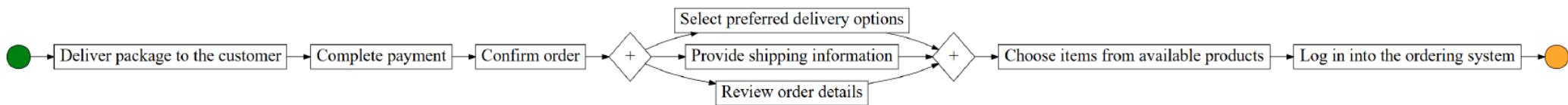
- **Start with ground truth models, i.e., pairs of model and text.**
- **Compare the original model with the generated model (e.g., using a combination of simulation and process mining).**
- **Background knowledge can backfire! Reverse the traces, model, etc.**

Table 2: Error handling performance metrics.

Model	Avg. Num. Iterations	Num. Cases without Errors	Num. Cases with Auto-Adjustment	Num. Cases with Failures
Claude-3.5-Sonnet	1.35	16	0	0
O1-Mini	1.4	14	0	0
O1-Preview	1.5	14	0	0
Gemini-1.5-Pro-002	1.95	13	0	0
GPT-4o	2.25	9	0	0
Llama-3.1-405B-Instruct	2.55	9	0	0
Mistral-Large-2	2.6	10	0	0
Llama-3.2-90B-Vision-Instruct	2.95	8	0	0
Gemini-1.5-Flash-002	3.3	4	0	0
Mixtral-8x22B	3.6	10	3	1
GPT-4	3.9	2	0	0
Codestral	3.9	7	2	1
GPT-4o-Mini	4.05	6	3	0
Llama-3.1-Nemotron-70B-Instruct	4.05	3	1	0
Qwen2.5-72B-Instruct	4.65	4	2	0
WizardLM-2-8x22B	5.2	8	5	0



(b) Gemini-1.5-Pro-002 (Standard Prompt): Reordered based on domain knowledge.



(c) Gemini-1.5-Pro-002 (Strict Prompt): Adhered to textual description.

AIPA: A Tool for Process Querying

- Provides a natural language interface for querying and understanding BPMN models.
- Supports voice input and output for intuitive interaction.
- Allows users to select specific parts of a model for focused analysis.
- Facilitates interactive dialogue, maintaining conversation history.

Kourani, Humam, et al. "Leveraging Large Language Models for Enhanced Process Model Comprehension." *arXiv preprint arXiv:2408.08892* (2024).

Choose BPMN Model

credit-scoring-asynchronous.bpmn [Select new model](#)

BPMN Model

Chat About the Uploaded Model [?](#) [Reset Conversation](#)

Hello! I am your AI assistant. How may I assist you with the uploaded BPMN model?

Can you explain the selected parts in one sentence?

In this process, after assessing if a credit score is available, if it is, the score is sent; if not, a delay is reported and further detailed scoring is performed before eventually sending the score.

What does the "X" symbol stand for?

The "X" symbol, referred to as an Exclusive Gateway, is used to represent decision points in the process where only one of multiple possible paths can be taken. It's like a branching logic where the process flow chooses only one route based on certain conditions or criteria. For instance, in the described process, the decision of whether a credit score is immediately available or not is determined by such a gateway.

Type your message...

Benchmarking LLMs for Process Mining Tasks

The benchmark includes different categories of tasks:

- **Category 1:** Assesses the contextual understanding of the LLM in process mining tasks. Various tasks, such as case ID inference, contextual splitting of activity labels, and defining high-level events, are considered.
- **Category 2:** Evaluates the LLM's ability to perform conformance checking and anomaly detection, starting from textual descriptions, event logs, or procedural process models.
- **Category 3:** Tests the LLM's capacity to generate and modify declarative and procedural process models.
- **Category 4:** Measures the LLM's process querying abilities, encompassing both procedural and declarative process models.
- **Category 5:** Examines the LLM's ability to generate valid hypotheses and questions based on the provided artifacts.
- **Category 6:** Assesses the LLM's ability to identify and propose solutions for unfairness in processes.
- **Category 7:** Evaluates the LLM's ability to read and interpret process mining diagrams.
- **Category 8:** Evaluates the LLM's ability to perform process optimizations in popular scenarios.

Alessandro Berti, Humam Kourani, Hannes Häfke, Chiao-Yun Li, Daniel Schuster:
Evaluating Large Language Models in Process Mining: Capabilities, Benchmarks, and Evaluation Strategies. BPMDS/EMMSAD@CAiSE 2024: 13-21 <https://github.com/fit-alessandro-berti/pm-llm-benchmark>

Benchmarking LLMs for Process Mining Tasks

gemini-2.5-pro-thinkhigh => 42.4 points

Question	Score
cat01_01_case_id_inference	7.75
cat01_02_activity_context	5.5
cat01_03_high_level_events	9.6
cat01_04_sensor_recordings	9.4
cat01_05_merge_two_logs	6
cat01_06_system_logs	7.75
cat01_07_interv_to_pseudo_bpnn	8.4
cat01_08_tables_to_log	10
cat02_01_conformance_textual	9.6
cat02_02_conf_desiderata	7.75
cat02_03_anomaly_event_log	8.4
cat02_04_powl_anomaly_detection	9.6
cat02_05_two_powls_anomalies	10
cat02_06_root_cause_1	10
cat02_07_root_cause_2	7.9
cat02_08_underfitting_process_tree	10
cat02_09_fix_process_tree	10
cat03_01_process_tree_generation	10

57 tasks



cat05_07_question_interview	8.4
cat06_01_bias_text	10
cat06_02_bias_event_log	9.6
cat06_03_bias_powl	8.4
cat06_04_bias_two_logs	9.6
cat06_05_bias_two_logs_2	8.4
cat06_06_bias_mitigation_declare	7
cat06_07_fair_unfair_powl	7.25
cat07_01_ocdfg	5.5
cat07_02_bpnn_orders	4.5
cat07_03_bpnn_dispatch	7.75
cat07_04_causal_net	9.6
cat07_05_proclcts	9.6
cat07_06_perf_spectrum	9.6
cat08_01_queue_mining	8.4
cat08_02_instance_spanning	9.4
cat08_03_transport_opt	9.4
cat08_04_resource_assign	9.6
cat08_05_task_schedul	9.6

> 6000 results to evaluate

A score in the range 24-29 is considered sufficient; a score in the range 29-34 is considered fair; a score in the range 34-40 is considered good; and a score >40 is considered excellent.

Overall Leaderboard (1-shot; gemini-2.5-pro used as a judge)

Model	Score	OS	LRM	PCo	CC	PMo	PQ	HG	FA	OPT	VI
gemini-2.5-pro-thinkhigh	42.4	✗	✓	6.4	8.3	5.2	5.6	6.2	6.0	4.6	4.7
gemini-2.5-pro-thinklow	42.2	✗	✓	5.7	8.4	5.3	5.5	6.1	6.5	4.7	5.4
gemini-2.5-flash-thinkhigh	41.0	✗	✓	6.1	8.3	4.8	5.4	6.2	5.5	4.7	5.1
o3-pro-2025-06-10	39.9	✗	✓	6.2	7.7	5.3	5.2	5.3	5.9	4.3	5.0
Grok-3-beta-thinking-20250221	38.5	✗	✓	6.9	6.9	6.3	3.8	4.8	5.9	3.9	0.0
phi4-reasoningplus	37.7	✓	✓	5.8	6.9	5.2	4.8	5.4	5.8	3.9	0.0
grok-3-mini-high	37.7	✗	✓	6.5	6.6	5.2	4.9	4.9	5.7	4.0	0.0
gemini-2.5-flash-nothink	37.2	✗	✗	5.8	7.1	3.4	5.1	5.5	5.7	4.6	4.8
phi4-reasoning	37.0	✓	✓	6.2	6.5	4.7	4.3	6.1	5.5	3.8	0.0
qwen-qwq-32b-nostepbystep	36.7	✓	✓	5.7	6.4	5.9	4.5	4.8	5.6	3.7	0.0
gemini-2.5-flash-lite-thinkhigh	36.4	✗	✓	5.5	7.1	4.3	4.3	5.6	5.2	4.3	4.0
qwen-qwq-32b-stepbystep	35.6	✓	✓	5.6	6.7	4.3	4.6	5.0	5.5	3.9	0.0

117 LLM variants



gpt-3.5-turbo	20.0	✗	✗	3.7	3.8	1.6	2.5	3.6	3.1	1.7	0.0
gemma34b-it-q8_0	19.8	✓	✗	2.2	3.3	1.3	2.7	4.1	3.1	3.1	2.0
qwen2.5-7b-instruct-1m	19.7	✓	✗	3.0	3.0	1.9	2.5	3.5	3.0	2.8	0.0
thudmglm-z1-9b	19.4	✓	✓	3.4	5.1	2.9	2.0	1.9	2.0	2.2	0.0
deepseek-r1-distill-qwen-7b	18.3	✓	✓	2.1	3.2	2.6	2.4	3.3	3.0	1.8	0.0
falcon33b-instruct-q8_0	18.1	✓	✗	2.5	3.2	1.6	2.2	3.0	2.8	2.7	0.0
Phi-4-multimodal-instruct	17.7	✓	✗	2.1	3.0	1.2	2.5	3.3	2.9	2.8	2.7
olmo27b-1124-instruct-q8_0	16.4	✓	✗	2.2	3.0	1.4	2.0	3.0	2.8	1.9	0.0
qwen30.6b	13.9	✓	✓	1.3	2.7	1.6	1.3	2.8	2.5	1.6	0.0
gemma31b-it-q8_0	12.6	✓	✗	1.1	1.9	1.0	1.4	3.2	2.0	1.9	0.0
deepseek-r1-distill-qwen-1.5b	11.1	✓	✓	1.5	1.9	1.0	1.2	2.8	1.7	0.9	0.0

Example Task (one of 57)

*C:\Dropbox\tmp\SIGMOD-PODS-WS-Berlin-June-2025\files\cat01_02_activity_context.txt - Notepad++

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

cat01_02_activity_context.txt Q-gemini-2.5-pro-thinkhigh_cat01_02_activity_context.txt A-gemini-2.5-pro-thinkhigh_cat01_02_activity_context.txt Q-deepseekdeepseek-r1-distill-qwen-1.5b_cat01_02_activity_context.txt A-deepseekdeepseek-r1-distill-qwen-1.5b_cat01_02_activity_context.txt

```
1  **Prompt:**  
2  
3  You are given a process event log from a complex support workflow at a fictional financial services company. The log captures multiple cases (each representing a customer support ticket) and their associated events. The same activity names appear multiple times within a single case, but carry different contextual meanings depending on when and how they occur. Your task is to carefully analyze the log and then rename the activities so that their names clearly reflect their contextual differences.  
4  
5  **Event Log:**  
6  
7  ``  
8  | Case ID | Activity | Timestamp | Resource | Additional Context |  
9  |-----|-----|-----|-----|-----|  
10 | C001 | Initiate Request | 2024-05-01 09:15:00 | WebPortal | Customer submitted a request via online portal. |  
11 | C001 | Verify Documents | 2024-05-01 09:20:00 | BackOfficeAgent | Agent checks identity documents. |  
12 | C001 | Call Customer | 2024-05-01 09:45:00 | FrontlineAgent | Verification incomplete, agent calls to request missing docs. |  
13 | C001 | Submit Additional | 2024-05-01 10:10:00 | CustomerSelfServe | Customer uploads the requested ID scan. |  
14 | C001 | Verify Documents | 2024-05-01 10:30:00 | BackOfficeAgent | Agent re-checks documents now that ID scan is provided. |  
15 | C001 | Approve Request | 2024-05-01 10:45:00 | Supervisor | Final approval given. |  
16  
17 | C002 | Initiate Request | 2024-05-02 08:05:00 | WebPortal | Customer submitted a request via online portal. |  
18 | C002 | Verify Documents | 2024-05-02 08:15:00 | BackOfficeAgent | Identity verification. |  
19 | C002 | Call Customer | 2024-05-02 09:00:00 | FrontlineAgent | Outbound call to clarify account discrepancy. |  
20 | C002 | Adjust Account | 2024-05-02 09:30:00 | BackOfficeAgent | Adjust account details internally after call. |  
21 | C002 | Call Customer | 2024-05-02 09:50:00 | FrontlineAgent | Follow-up call to confirm adjustments made. |  
22 | C002 | Approve Request | 2024-05-02 10:05:00 | Supervisor | Final approval given. |  
23  
24 | C003 | Initiate Request | 2024-05-03 11:20:00 | BranchReception | Walk-in customer at branch. |  
25 | C003 | Verify Documents | 2024-05-03 11:25:00 | BranchOfficer | Physical verification of ID. |  
26 | C003 | Call Customer | 2024-05-03 11:50:00 | BranchOfficer | Customer stepped out. Officer calls to have them return. |  
27 | C003 | Adjust Account | 2024-05-03 12:05:00 | BackOfficeAgent | Internal correction due to missing info. |  
28 | C003 | Verify Documents | 2024-05-03 12:20:00 | BranchOfficer | Re-verify docs after account adjustment. |  
29 | C003 | Approve Request | 2024-05-03 12:30:00 | Supervisor | Final approval given. |  
30  
31  **Your Task:**  
32  
33  1. Identify activities that appear multiple times within the same case and interpret how their meaning changes depending on their position in the process or the associated context.  
34  
35  2. For each activity that recurs, rename them to reflect their context.  
36  
37  3. Present a revised version of the event log with the renamed activities, ensuring that the new names disambiguate the activities and highlight their roles in the process flow.  
38  
39  4. Justify the naming choices briefly, indicating the contextual clues that led to those changes.  
40  
41
```



Response by gemini-2.5-pro-thinkhigh on cat01_02_activity_context

```

1 of course. Here is a detailed analysis, the revised event log, and the justifications for the naming changes.
2
3 ## 1. Identification of Ambiguous Activities
4
5 By analyzing the event log, the following activities appear multiple times within a single case but have different contextual meanings:
6
7 * **Verify Documents:** This activity occurs twice in Case C001 and twice in Case C003. In both cases, the first instance is an initial check, while the second is a re-verification that happens after an intervening step (like receiving additional documents or an account adjustment).
8 * **Call Customer:** This activity occurs twice in Case C002. The first call is to investigate and clarify a problem (an account discrepancy). The second call is a follow-up to confirm that a resolution has been implemented.
9
10 ## 2. Renaming Strategy
11
12 To resolve the ambiguity, the activities will be renamed as follows:
13
14 | Original Ambiguous Activity | Context | Proposed New Name |
15 |-----|-----|-----|
16 | 'Verify Documents' | The first check in the process | 'Verify Documents (Initial)' |
17 | 'Verify Documents' | A subsequent check after a correction | 'Re-verify Documents' |
18 | 'Call Customer' | A call to investigate or request information | 'Call Customer (Clarification)' |
19 | 'Call Customer' | A call to confirm an action has been taken | 'Call Customer (Confirmation)' |
20
21 ## 3. Revised Event Log
22
23 Here is the updated event log with the disambiguated activity names.
24
25 | Case ID | Activity | Timestamp | Resource | Additional Context |
26 |-----|-----|-----|-----|-----|
27 | C001 | Initiate Request | 2024-05-01 09:15:00 | WebPortal | Customer submitted a request via online portal. |
28 | C001 | **Verify Documents (Initial)** | 2024-05-01 09:20:00 | BackOfficeAgent | Agent checks identity documents. |
29 | C001 | Call Customer | 2024-05-01 09:45:00 | FrontlineAgent | Verification incomplete, agent calls to request missing docs. |
30 | C001 | Submit Additional | 2024-05-01 10:10:00 | CustomerSelfServe | Customer uploads the requested ID scan. |
31 | C001 | **Re-verify Documents** | 2024-05-01 10:30:00 | BackOfficeAgent | Agent re-checks documents now that ID scan is provided. |
32 | C001 | Approve Request | 2024-05-01 10:45:00 | Supervisor | Final approval given. |
33 | | | | |
34 | C002 | Initiate Request | 2024-05-02 08:05:00 | WebPortal | Customer submitted a request via online portal. |
35 | C002 | Verify Documents | 2024-05-02 08:15:00 | BackOfficeAgent | Identity verification. |
36 | C002 | **Call Customer (Clarification)** | 2024-05-02 09:00:00 | FrontlineAgent | Outbound call to clarify account discrepancy. |
37 | C002 | Adjust Account | 2024-05-02 09:30:00 | BackOfficeAgent | Adjust account details internally after call. |
38 | C002 | **Call Customer (Confirmation)** | 2024-05-02 09:50:00 | FrontlineAgent | Follow-up call to confirm adjustments made. |
39 | C002 | Approve Request | 2024-05-02 10:05:00 | Supervisor | Final approval given. |
40 | | | | |
41 | C003 | Initiate Request | 2024-05-03 11:20:00 | BranchReception | Walk-in customer at branch. |
42 | C003 | **Verify Documents (Initial)** | 2024-05-03 11:25:00 | BranchOfficer | Physical verification of ID. |
43 | C003 | Call Customer | 2024-05-03 11:50:00 | BranchOfficer | Customer stepped out. Officer calls to have them return. |
44 | C003 | Adjust Account | 2024-05-03 12:05:00 | BackOfficeAgent | Internal correction due to missing info. |
45 | C003 | **Re-verify Documents** | 2024-05-03 12:20:00 | BranchOfficer | Re-verify docs after account adjustment. |
46 | C003 | Approve Request | 2024-05-03 12:30:00 | Supervisor | Final approval given. |

```

```

17 | 'Verify Documents' | A subsequent check after a correction | 'Re-verify Documents' |
18 | 'Call Customer' | A call to investigate or request information | 'Call Customer (Clarification)' |
19 | 'Call Customer' | A call to confirm an action has been taken | 'Call Customer (Confirmation)' |
20
21 ## 3. Revised Event Log
22
23 Here is the updated event log with the disambiguated activity names.
24

```

Case ID	Activity	Timestamp	Resource	Additional Context
C001	Initiate Request	2024-05-01 09:15:00	WebPortal	Customer submitted a request via online portal.
C001	**Verify Documents (Initial)**	2024-05-01 09:20:00	BackOfficeAgent	Agent checks identity documents.
C001	Call Customer	2024-05-01 09:45:00	FrontlineAgent	Verification incomplete, agent calls to request missing docs.
C001	Submit Additional	2024-05-01 10:10:00	CustomerSelfServe	Customer uploads the requested ID scan.
C001	**Re-verify Documents**	2024-05-01 10:30:00	BackOfficeAgent	Agent re-checks documents now that ID scan is provided.
C001	Approve Request	2024-05-01 10:45:00	Supervisor	Final approval given.
C002	Initiate Request	2024-05-02 08:05:00	WebPortal	Customer submitted a request via online portal.
C002	Verify Documents	2024-05-02 08:15:00	BackOfficeAgent	Identity verification.
C002	**Call Customer (Clarification)**	2024-05-02 09:00:00	FrontlineAgent	Outbound call to clarify account discrepancy.
C002	Adjust Account	2024-05-02 09:30:00	BackOfficeAgent	Adjust account details internally after call.
C002	**Call Customer (Confirmation)**	2024-05-02 09:50:00	FrontlineAgent	Follow-up call to confirm adjustments made.
C002	Approve Request	2024-05-02 10:05:00	Supervisor	Final approval given.
C003	Initiate Request	2024-05-03 11:20:00	BranchReception	Walk-in customer at branch.
C003	**Verify Documents (Initial)**	2024-05-03 11:25:00	BranchOfficer	Physical verification of ID.
C003	Call Customer	2024-05-03 11:50:00	BranchOfficer	Customer stepped out. Officer calls to have them return.
C003	Adjust Account	2024-05-03 12:05:00	BackOfficeAgent	Internal correction due to missing info.
C003	**Re-verify Documents**	2024-05-03 12:20:00	BranchOfficer	Re-verify docs after account adjustment.
C003	Approve Request	2024-05-03 12:30:00	Supervisor	Final approval given.

```

48 ## 4. Justification for Naming Choices
49

```

50 The new names were chosen based on the following contextual clues from the log:

- 51
- * ****Verify Documents (Initial) vs. Re-verify Documents:****
 - * ****Contextual Clue:**** The position in the workflow and the description in "Additional Context".
 - * ****Justification:**** The first occurrence of 'Verify Documents' in a case always follows the 'Initiate Request' activity. This is the standard, primary check, which is why it is renamed ****'Verify Documents (Initial)'****. The second occurrence always follows a corrective action (e.g., 'Submit Additional' or 'Adjust Account'). The context explicitly mentions "re-checks" or "re-verify," making ****'Re-verify Documents'**** a fitting and descriptive name for this rework loop.

52

 - * ****Call Customer (Clarification) vs. Call Customer (Confirmation):****
 - * ****Contextual Clue:**** The purpose of the call, as described in "Additional Context".
 - * ****Justification:**** In Case C002, the first call's purpose is to "clarify account discrepancy." This is an investigative action to gather information. ****'Call Customer (Clarification)'**** accurately captures this intent. The second call's purpose is to "confirm adjustments made." This is a concluding action to inform the customer of a resolution. ****'Call Customer (Confirmation)'**** clearly distinguishes this from the first call and highlights its role as a follow-up.

C:\Dropbox\tmp\SIGMA

File Edit Search View

Response by deepseek-r1-distill-qwen-1.5b on cat01_02_activity_context

cat01_02_activity_context.txt Q-deepseekdeepseek-r1-distill-qwen-1.5b_cat01_02_activity_context.txt A-deepseekdeepseek-r1-distill-qwen-1.5b_cat01_02_activity_context.txt

```
1 <think>
2 Okay, I'm looking at this problem where I have to rename activities in a log based on their context. The goal is to make the activities clear and distinguishable in each case.
3
4 First, I need to identify which activities repeat within the same case. Let's go through each case one by one.
5
6 Starting with Case C001. The activities are Initiate Request, Verify Documents, Call Customer, Submit Additional, Verify Documents again, and Approve Request. These look like they happen in a sequence. The first two are related to an online portal, then a frontline agent handling documents and missing info, and finally a supervisor. So maybe the first two are portal-related, the third is more technical, and the last is a higher-level approval.
7
8 Case C002 seems similar. Initiate Request again, then verify, call, adjust account, another call confirming, and approve. So it's almost the same as C001 but with different people. The first two are portal, then frontline, internal adjust, then supervisor.
9
10 Case C003 is different. Initiate Request, verify documents, call, adjust account, another call confirming, and approve. Wait, but all of them have verify again. So it's walk-in, physical, phone, adjust, confirm again, approve. The activities are all related to customer service.
11
12 Now, looking at the activities:
13
14 - Verify Documents is used multiple times, both before and after adjustments. So maybe this is a key activity that indicates the process of reviewing documents, whether it's before or after something else.
15
16 - Call Customer is used for clarifications and follow-ups. It could be that this is a recurring activity, but I'm not sure yet.
17
18 - Call Customer is also used in C003, so it's in multiple cases. Maybe it's a standard activity.
19
20 - Adjust Account is only in C003. It's internal, so perhaps it's a more specific step.
21
22 - Submit Additional and Approve Request are only in C001 and C002. These are final steps, probably once per case.
23
24 - WebPortal, BranchReception, and BranchOfficer are all related to customer service and specific branches. WebPortal is online; others are phone-related.
25
26 Now, for each activity that repeats, I need to decide the new name:
27
28 1. Initiate Request: This happens multiple times in each case. It's a base activity, so maybe it stays as Initiate Request.
29
30 2. Verify Documents: It seems to occur before and after adjustments. Since the process involves verification before an adjustment, maybe it's better to rename it to Verify Adjustment or something that implies it's before the adjustment.
31
32 3. Call Customer: It's a recurring activity. Since it's about clarifying or following up, maybe Rename it to Clarification Call.
33
34 4. Submit Additional: This is a final step and only once per case, so it might stay as Submit Additional.
35
36 5. Adjust Account: Since it's a specific internal step, perhaps Rename it to Internal Adjustment.
37
38 6. Approve Request: Once per case, so it's fine as Approve Request.
39
40 7. Verify Documents (second time): Since it's after an adjustment, maybe rename it to Verify Adjustment.
41
42 8. Call Customer (second time): It's a follow-up, so Clarification Call.
```

length: 8,847 lines: 107 Ln: 1 Col: 1 Pos: 1 Unix (LF) UTF-8 INS

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

cat01_02_activity_context.txt Q-gemini-2.5-pro-thinkhigh_cat01_02_activity_context.txt A-gemini-2.5-pro-thinkhigh_cat01_02_activity_context.txt Q-deepseekdeepseek-r1-distill-qwen-1.5b_cat01_02_activity_context.txt A-deepseekdeepseek-r1-distill-qwen-1.5b_cat01_02_activity_context.txt

```

64 - **Verify Documents**: This activity is used before adjustments and again after adjustments. Renaming it to **Verify Adjustment** clarifies its role as a verification step before an adjustment. ?
65 - **Call Customer**: A recurring activity across multiple cases to handle clarifications and follow-ups. Renaming it to **Clarification Call** makes it clear it's about customer support interactions. ?
66 - **Submit Additional**: This activity is a final step, only once per case. It remains as **Submit Additional**.
67 - **Adjust Account**: An internal adjustment step, necessary for internal processes. Renaming it to **Internal Adjustment** reflects its specificity.
68 - **Approve Request**: Once per case, remains unchanged as it is a final approval.
69
70
71
72
73
74
75 ## 2. Revise the Event Log with Renamed Activities
76
77 Here is the revised event log with renamed activities:
78
79 | Case ID | Activity | Timestamp | Resource | Additional Context |
80 |-----|-----|-----|-----|-----|
81 | C001 | Initiate Request | 2024-05-01 09:15:00 | WebPortal | Customer submitted a request via online portal. |
82 | C001 | Verify Adjustment | 2024-05-01 09:20:00 | BackOfficeAgent | Agent checks identity documents. |
83 | C001 | Call Customer | 2024-05-01 09:45:00 | FrontlineAgent | Verification incomplete, agent calls to request missing docs. |
84 | C001 | Submit Additional | 2024-05-01 10:10:00 | CustomerSelfServe | Customer uploads the requested ID scan. |
85 | C000 | Verify Adjustment | 2024-05-01 10:30:00 | BackOfficeAgent | Agent re-checks documents now that ID scan is provided. |
86 | C001 | Approve Request | 2024-05-01 10:45:00 | Supervisor | Final approval given. |
87 | C002 | Initiate Request | 2024-05-02 08:05:00 | WebPortal | Customer submitted a request via online portal. |
88 | C002 | Verify Adjustment | 2024-05-02 08:15:00 | BackOfficeAgent | Identity verification. |
89 | C002 | Call Customer | 2024-05-02 09:00:00 | FrontlineAgent | Outbound call to clarify account discrepancy. |
90 | C002 | Adjust Internal Account | 2024-05-02 09:30:00 | BackOfficeAgent | Internal adjustment after call. |
91 | C002 | Call Customer | 2024-05-02 09:50:00 | FrontlineAgent | Follow-up call to confirm adjustments made. |
92 | C002 | Approve Request | 2024-05-02 10:05:00 | Supervisor | Final approval given. |
93 | C003 | Initiate Request | 2024-05-03 11:20:00 | BranchReception | Walk-in customer at branch. |
94 | C003 | Verify Adjustment | 2024-05-03 11:25:00 | BranchOfficer | Physical verification of ID. |
95 | C003 | Call Customer | 2024-05-03 11:50:00 | BranchOfficer | Customer stepped out. Officer calls to have them return. |
96 | C003 | Internal Adjustment | 2024-05-03 12:05:00 | BackOfficeAgent | Internal correction due to missing info. |
97 | C003 | Verify Adjustment | 2024-05-03 12:20:00 | BranchOfficer | Re-verify docs after account adjustment. |
98 | C003 | Approve Request | 2024-05-03 12:30:00 | Supervisor | Final approval given. |
99 | C003 | Approve Request | 2024-05-03 12:35:00 | Supervisor | Final approval given. |
100
101 ## 3. Justification for Renaming Choices
102
103 - **Verify Adjustment**: Emphasizes the activity as a verification step before an adjustment, clarifying that it's a follow-up check after an adjustment.
104 - **Clarification Call**: Clearly indicates that this activity is about clarifying or following up on customer issues.
105 - **Internal Adjustment**: Reflects the specificity of the adjustment being internal, necessary for internal processes.
106
107 These renamesings enhance clarity and context, making the activities more understandable and the event log easier to analyze.

```

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

cat01_02_activity_context.txt Q-gemini-2.5-pro-thinkhigh_cat01_02_activity_context.txt A-gemini-2.5-pro-thinkhigh_cat01_02_activity_context.txt Q-deepseekdeepseek-r1-distill-qwen-1.5b_cat01_02_activity_context.txt A-deepseekdeepseek-r1-distill-qwen-1.5b_cat01_02_activity_context.txt

```
1  **Grade: 5.5 / 10.00**
2
3 This answer demonstrates a partial understanding of the task but contains significant logical flaws and inconsistencies in its execution. While the initial analysis is mostly correct, the final revised event log is flawed and would be misleading for any subsequent process analysis. The evaluation is strict, as requested.
4
5  ## Detailed Evaluation
6
7  **Positives:**
8
9  *  **Structure:** The response is well-structured, directly addressing the four parts of the prompt in a clear and organized manner.
10 *  **Initial Identification:** The answer correctly identifies that 'Verify Documents' and 'Call Customer' are the activities with ambiguous meanings based on their recurrence within a single case (C001, C003 for the former; C002 for the latter).
11 *  **Renaming Logic:** The proposed renaming strategy ('Initial' vs. 'Re-verify'; 'Clarification' vs. 'Confirmation') is logical, and the justifications provided are sound, referencing the correct contextual clues from the log.
12
13  **Areas for Improvement (Reasons for Low Score):**
14
15 1.  **Critical Inconsistency in Execution:** The most significant failure is that the renaming logic, once defined, is not applied consistently across the entire event log. The primary goal of such a task is to produce a clean, globally unambiguous log. This answer fails to do that.
16  *  **Flaw 1 ('Verify Documents'):** The renaming rule is defined as: "The first check in the process" becomes 'Verify Documents (Initial)'. This rule is correctly applied to C001 and C003. However, in case C002, the first and only 'Verify Documents' activity is left with its original name. According to the answer's own rule, it should have been renamed to 'Verify Documents (Initial)'. This inconsistency makes the final log incorrect, as it now contains three distinct verification activities ('Verify Documents', 'Verify Documents (Initial)', and 'Re-verify Documents') when the context only supports two.
17  *  **Flaw 2 ('Call Customer'):** The renaming logic for calls is only applied to case C002, where the ambiguity appeared within the same case. However, the 'Call Customer' in C001 ("request missing docs") and C003 ("calls to have them return") clearly fall under the "Clarification" or "Request" category. By leaving them with the generic name, the revised log is again misleading. It suggests there are three types of calls ('Call Customer', 'Call Customer (Clarification)', 'Call Customer (Confirmation)'), which complicates and misrepresents the actual process.
18
19 2.  **Misunderstanding the Ultimate Goal:** The answer interprets the prompt "identify activities that appear multiple times within the same case" too literally. While this is the *trigger* for the analysis, the *goal* is to create a fully disambiguated log. Once a new, more specific activity context is discovered (e.g., "Initial Verification"), that label should be applied to *all* activities that fit that context, regardless of whether they happened to have a recurrence in their specific case. The failure to do so demonstrates a shallow understanding of the purpose of process log preparation.
20
21 3.  **Impact on Process Mining:** A process discovery algorithm run on the submitted "Revised Event Log" would produce an incorrect and unnecessarily complex process model. It would create separate paths for the inconsistently named activities, failing to reflect the true underlying process. Therefore, the answer fails in its primary, practical objective.
22
23  ## Conclusion
24
25 The answer successfully completes the analytical portion of the task but fails critically on the implementation. The resulting event log is inconsistent and flawed, which is a major failure for a task centered on data cleaning and preparation. A high-quality response would have applied the new, specific names to all corresponding activities throughout the entire log, demonstrating a deeper understanding of the task's purpose.
```

Evaluation of gemini-2.5-pro-thinkhigh on cat01_02_activity_context by gemini-2.5-pro-thinkhigh (5.5 points)

```
1  **Grade: 1.5 / 10.0**  
2  
3  **Evaluation:**  
4  
5  1.  **Identification of Activities:** The answer correctly identifies 'Verify Documents' and 'Call Customer' as the primary activities that repeat within the same case and require contextual renaming. This part is adequate.  
6  2.  **Renaming Strategy and Execution:** This is where the answer fails significantly.  
7    *  **Failure to Disambiguate:** The core task was to rename activities *so that their names clearly reflect their contextual differences*. The chosen renames ('Verify Adjustment', 'Clarification Call') are applied uniformly to *all* instances of the original activities within the cases, completely failing to differentiate them.  
8      * For 'Verify Documents' (e.g., in C001), the first instance is an initial check, and the second is a re-check after document submission. Renaming both to 'Verify Adjustment' (which is also contextually inaccurate for C001) does not distinguish them. Appropriate names would be something like 'Initial Document Verification' and 'Re-verify Submitted Documents'.  
9      * For 'Call Customer' (in C002), the first call is for clarification, and the second is for confirmation. Renaming both to 'Clarification Call' fails to capture this difference. Appropriate names might be 'Call Customer (Clarification)' and 'Call Customer (Confirmation)'.  
10  *  **Contextual Inaccuracy:** The name 'Verify Adjustment' is poorly chosen. In C001, the re-verification happens after document submission, not an account adjustment. Applying this name to the initial verification steps in all cases is also inaccurate.  
11 3.  **Revised Event Log:** The revised log contains critical data errors:  
12    * In C001, the fifth event (the second verification) is incorrectly assigned 'Case ID C000' instead of 'C001'.  
13    * In C003, an extra, spurious 'Approve Request' event is added at the end ('2024-05-03 12:35:00'), which was not present in the original log.  
14    * These errors fundamentally compromise the integrity of the revised log.  
15 4.  **Justification:** The justifications provided attempt to explain the renaming but reflect the flawed logic.  
16    * The justification for 'Verify Adjustment' ("Emphasizes the activity as a verification step before an adjustment, clarifying that it's a follow-up check after an adjustment") is contradictory and doesn't align with how the name was actually applied (e.g., to initial checks or post-document submission checks).  
17    * The justification for 'Clarification Call' doesn't acknowledge that the chosen name failed to distinguish between the different *types* of calls within C002.  
18  
19  **Conclusion:**  
20  
21 The answer demonstrates a misunderstanding of the core requirement to *differentiate* between repeated activities based on their specific context within a case. The renaming strategy was poorly executed, using inaccurate and non-discriminatory names. Furthermore, the introduction of significant data errors in the revised event log is unacceptable. Judged hypercritically, the answer has major flaws in execution, logical application, and data integrity, failing to achieve the prompt's objectives.
```

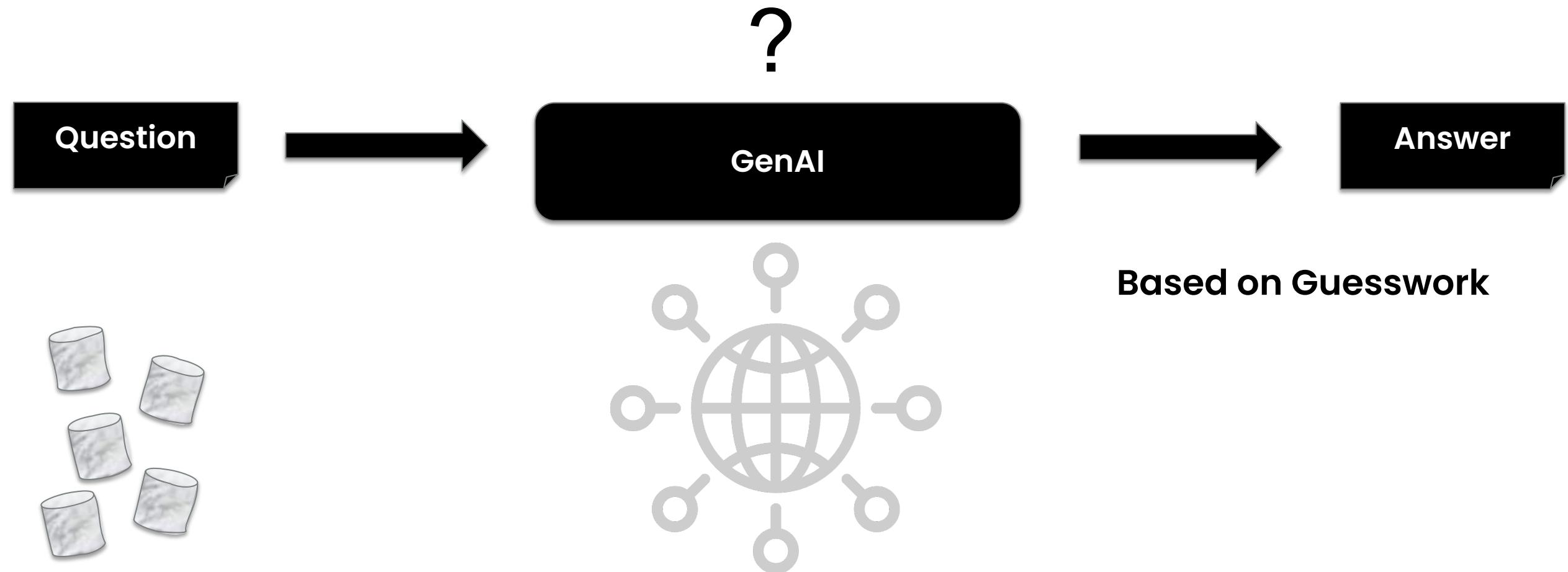
Evaluation of deepseek-r1-distill-qwen-1.5b on cat01_02_activity_context
by gemini-2.5-pro-thinkhigh (1.5 points)

So What?

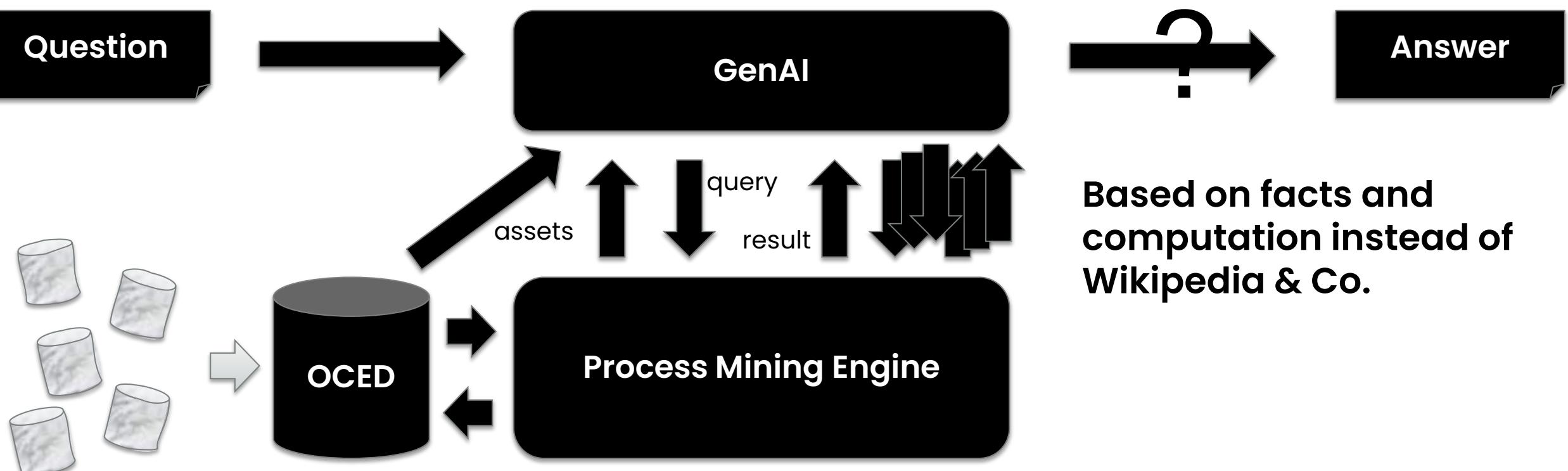
Amazing and
confusing at
the same time



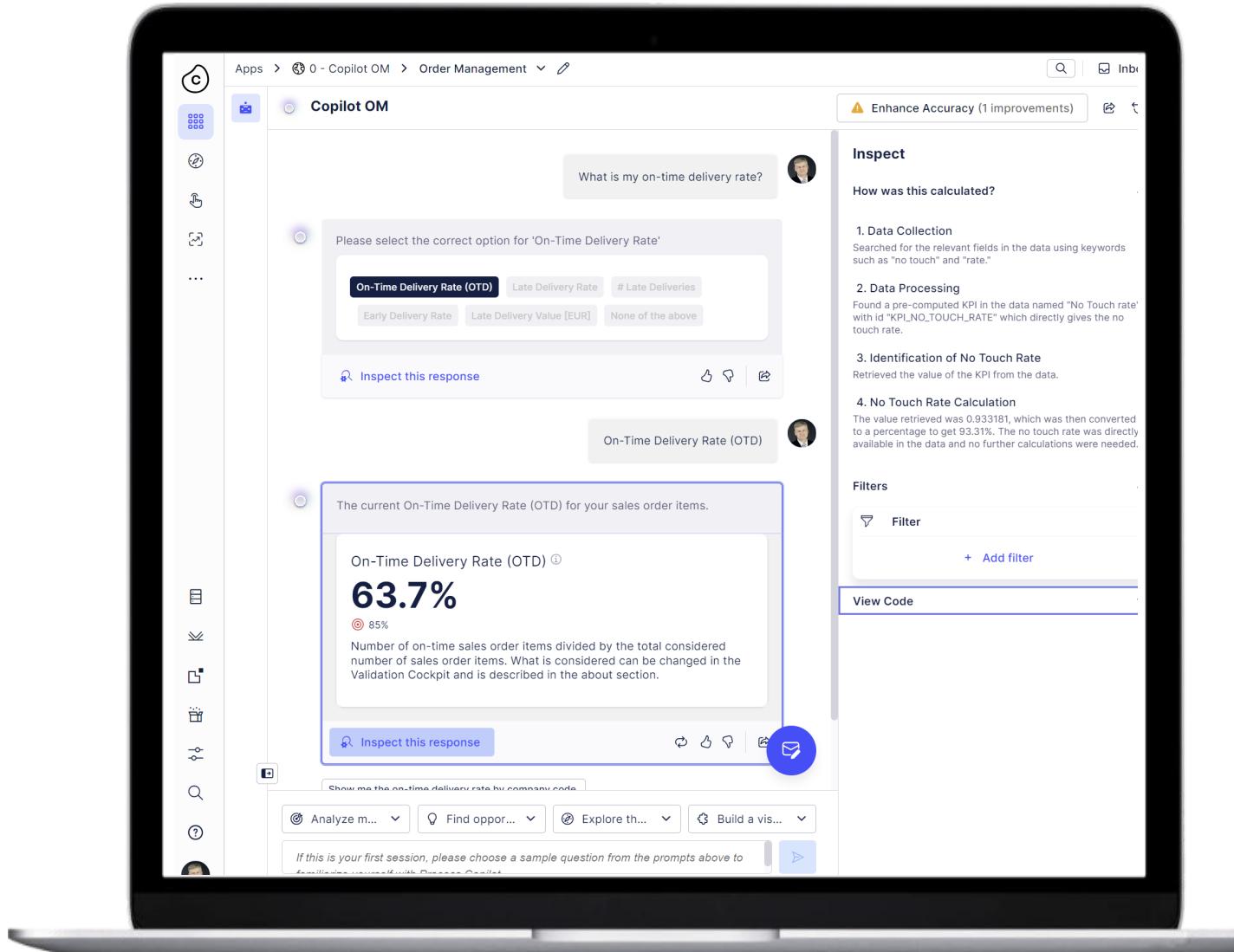
Guesswork Versus Computation



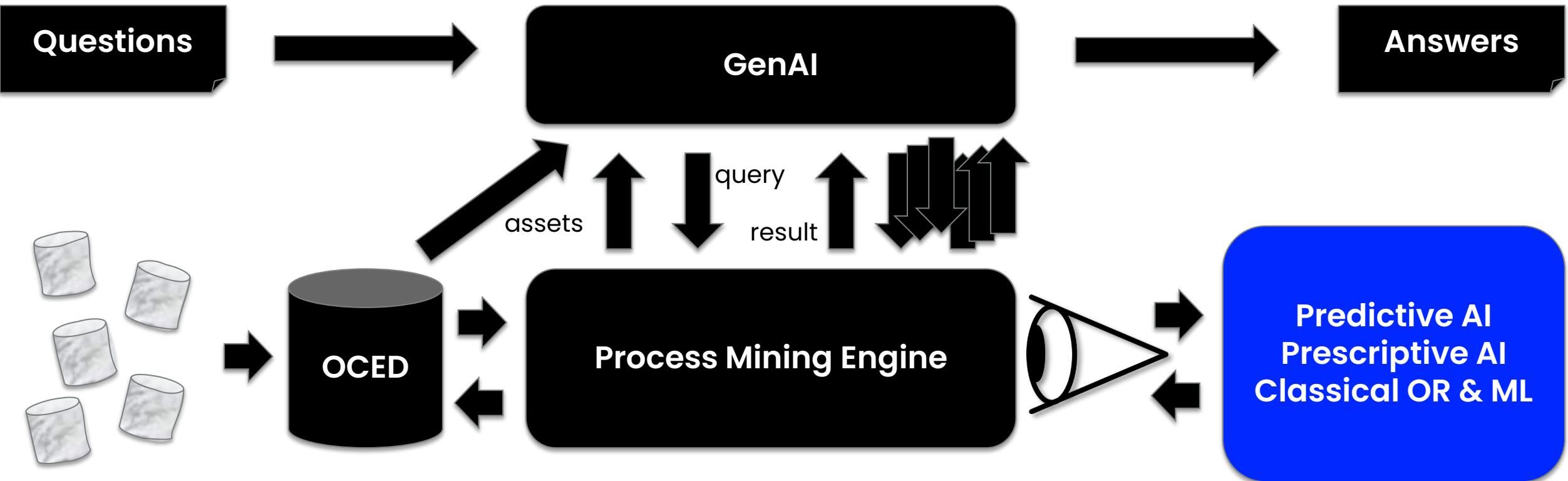
Guesswork Versus Computation



Process Mining Copilot: Lowering the Threshold To Use PM



Adding Other Forms of AI, ML, and Automation



Generating “Machine Learning Problems” for “Process Problems”

PAM

Alignment Explorer
Explore variants that include the deviations we detect in relation to the target process model.

Deviations occur in the following 5488 variants (showing 100 of 5488)

Refine variants

Objects ↗ Coverage ↑ Avg TPT ↑

Objects ↗	Coverage ↑	Avg TPT ↑
5.8K	1.46%	0 ms
3.3K	0.82%	4 d
2.9K	0.73%	3 d
2K	0.49%	4 d
1.7K	0.42%	20 d
1.2K	0.30%	11 d

Variant Detail

Event Sequence	Deviations	Time delta
Create Sales Order Header	0	
Create Sales Order Item	0 ms	
Create Sales Order Item	0 ms	
Create Sales Order Item	0 ms	
Create Sales Order Item	0 ms	
Set Billing Block	Violated exclusive gateway	0 ms

Target Model

Filters
Use filters to select the data you want to compare to the target model.

Data & Coverage Info

Delivery Item	100%
Sales Order	100%
Sales Order Item	100%

situation table

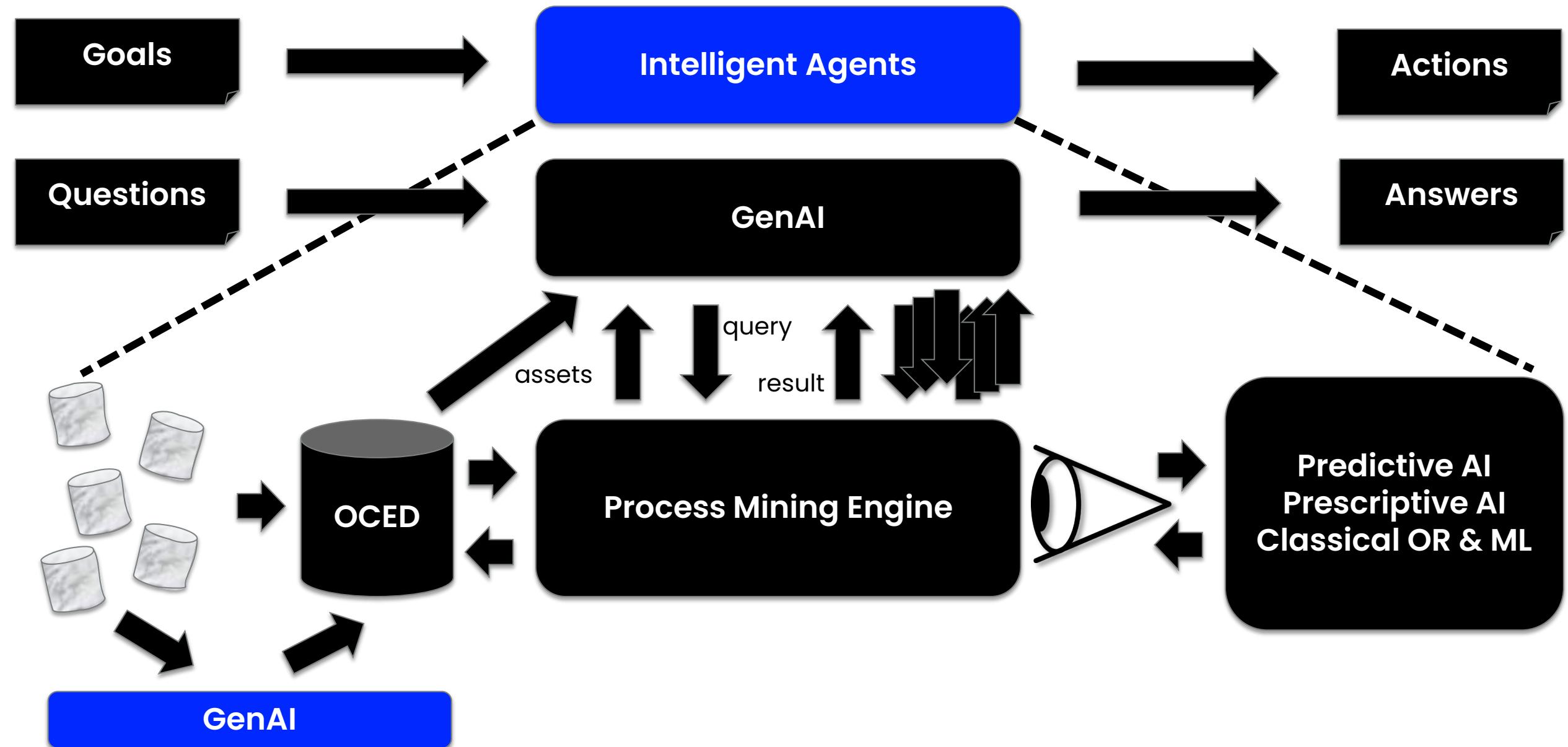
CASE	PRODUCT	ADDRESS	first resource	total throughput time	decision
1	SAMSUNG Galaxy J5	Munich	Caleb	239	pay
2	APPLE iPhone 6s 64...	Amsterdam	Lucas	201	pay
3	APPLE iPhone 5s 16...	New York	Sophia	503	cancel order
4	MOTOROLA Moto E 3rd...	New York	Sophia	498	cancel order
5	SAMSUNG Core Prime	Aachen	Isabella	741	pay
6	SAMSUNG Galaxy S4	Munich	Emma	406	pay
7	MOTOROLA Moto G	Amsterdam	Lucas	598	pay
8	APPLE iPhone 6 16...	Amsterdam	Sophia	209	pay
9	APPLE iPhone 5s 16...	Munich	Aiden	412	pay
10	HUAWEI P8 Lite	Amsterdam	Speedy	415	pay
11	MOTOROLA Moto G	Munich	Emma	508	pay
12	APPLE iPhone 5s 16...	Aachen	Isabella	400	cancel order
13	HUAWEI P8 Lite	Munich	RESOURCE		
14	SAMSUNG Core Prime	Munich	Abigail	557	
15	SAMSUNG Galaxy S4	Aachen	Abigail		

decision

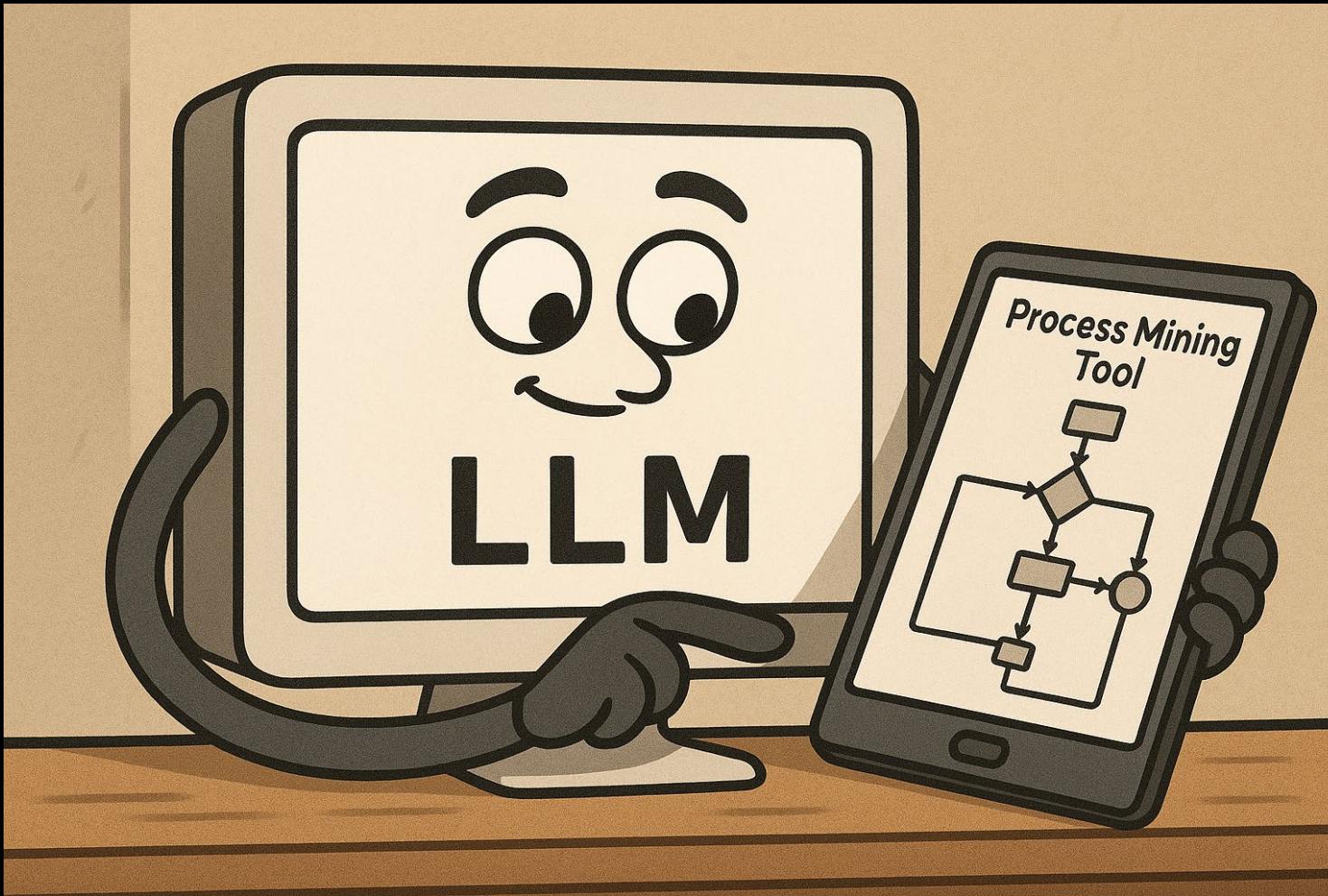
bottleneck

What is causing the bottleneck?
Which orders are deviating?
When will this product be delivered?
Will we meet our SLA tomorrow?

Adding Other Forms of AI, ML, and Automation







Providing the **context**

OCPM

It all starts with event data

Case ID	Activity	Resource	Timestamp	Product	Prod-price	Quantity	Address
...
6350	place order	Aiden	2018/02/13 14:29:45.000	APPLE iPhone 6 16 GB	639,00 €	5	NL-7751DG-21
6283	pay	Lily	2018/02/13 14:39:25.000	SAMSUNG Galaxy S6 32 GB	543,99 €	3	NL-7828AC-01
6253	prepare delivery	Sophia	2018/02/13 15:01:33.000	APPLE iPhone 6 16 GB	639,00 €	3	NL-7887AC-13
6257	prepare delivery	Aiden	2018/02/13 15:03:43.000	SAMSUNG Galaxy S6 32 GB	543,99 €	1	NL-9521KJ-34
6185	confirm payment	Emily	2018/02/13 15:05:36.000	SAMSUNG Galaxy S4	329,00 €	1	NL-9521GC-32
6218	confirm payment	Emily	2018/02/13 15:08:11.000	APPLE iPhone 6s Plus 64 GB	639,00 €	2	NL-7948BX-10
6245	make delivery	Michael	2018/02/13 15:14:04.000	APPLE iPhone 6 16 GB	639,00 €	3	NL-7905AX-38
6272	pay	Emily	2018/02/13 15:20:36.000	APPLE iPhone 6 16 GB	639,00 €	1	NL-7821AC-3
6269	pay	Charlotte	2018/02/13 15:25:21.000	SAMSUNG Galaxy S4	329,00 €	1	NL-7907EJ-42
6212	prepare delivery	Sophia	2018/02/13 15:43:39.000	HUAWEI P8 Lite	234,00 €	1	NL-7905AX-38
6323	send invoice	Alexander	2018/02/13 15:46:08.000	APPLE iPhone 6s 16 GB	858,00 €	1	NL-7833HT-15
6246	confirm payment	Jack	2018/02/13 15:56:03.000	SAMSUNG Galaxy S4	329,00 €	3	NL-7833HT-15
6347	send invoice	Jack	2018/02/13 15:57:42.000	SAMSUNG Galaxy S4	329,00 €	3	NL-7905AX-38
6351	place order	Zoe	2018/02/13 16:17:37.000	APPLE iPhone 5s 16 GB	449,00 €	3	NL-9521GC-32
6204	prepare delivery	Sophia	2018/02/13 16:31:28.000	SAMSUNG Core Prime G361	135,00 €	1	NL-7828AM-11a
6204	make delivery	Kaylee	2018/02/13 16:51:54.000	SAMSUNG Core Prime G361	135,00 €	1	NL-7828AM-11a
6265	confirm payment	Lily	2018/02/13 16:55:55.000	SAMSUNG Galaxy S4	329,00 €	4	NL-9521GC-32
6250	confirm payment	Jack	2018/02/13 17:03:26.000	MOTOROLA Moto G	199,00 €	4	NL-7942GT-2
6328	send invoice	Lily	2018/02/13 17:30:16.000	APPLE iPhone 6s 64 GB	858,00 €	4	NL-9514BV-16
6352	place order	Aiden	2018/02/13 17:53:22.000	APPLE iPhone 6 16 GB	639,00 €	2	NL-9514BV-16
6317	send invoice	Jack	2018/02/13 18:45:30.000	APPLE iPhone 6s 64 GB	858,00 €	5	NL-7907EJ-42
6353	place order	Sophia	2018/02/13 20:16:20.000	APPLE iPhone 5s 16 GB	449,00 €	4	NL-7751AR-19
...

event = customers

objects + orders

items

suppliers

invoices

machines

shipments

...

Objects & Events Are Everywhere!



We cannot squeeze this reality into cases, we need a multitude of interconnected objects and events



Minimal Example: On Time In Full (OTIF) Score?

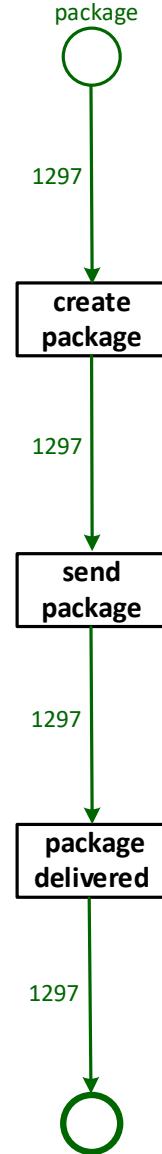
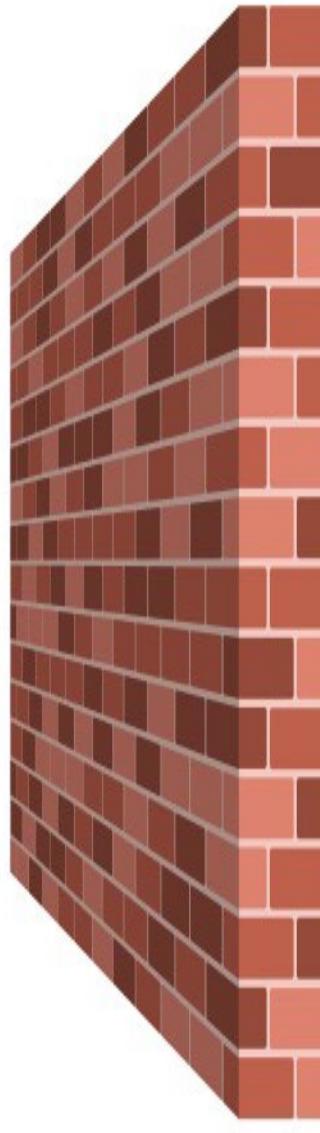
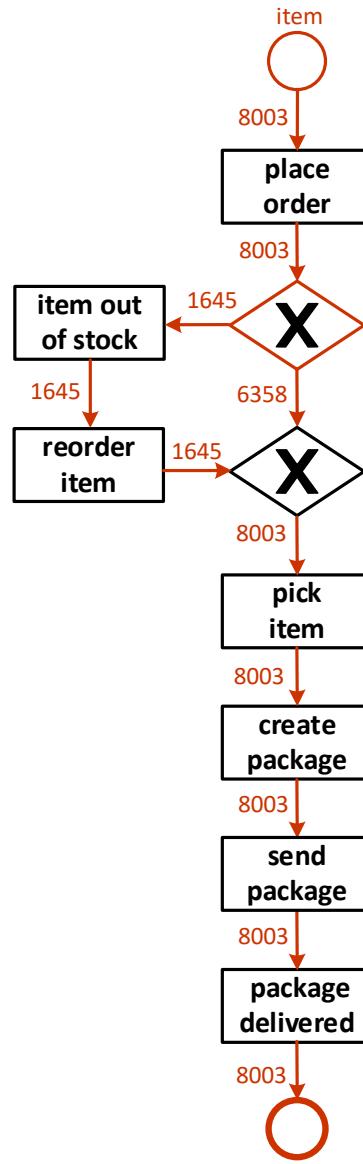
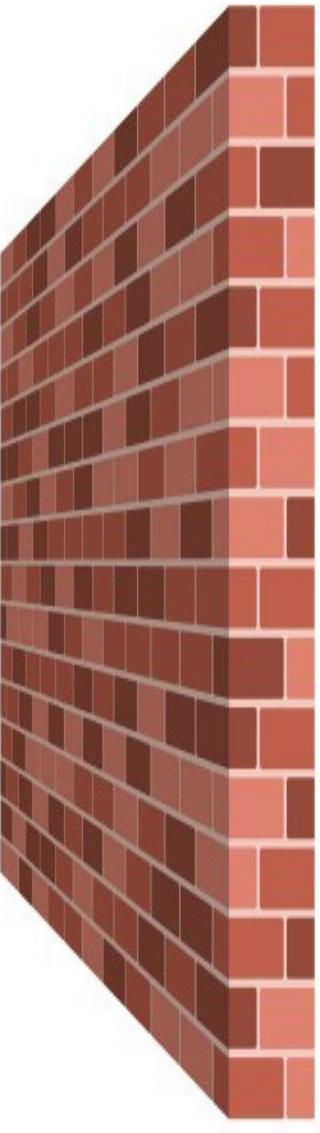
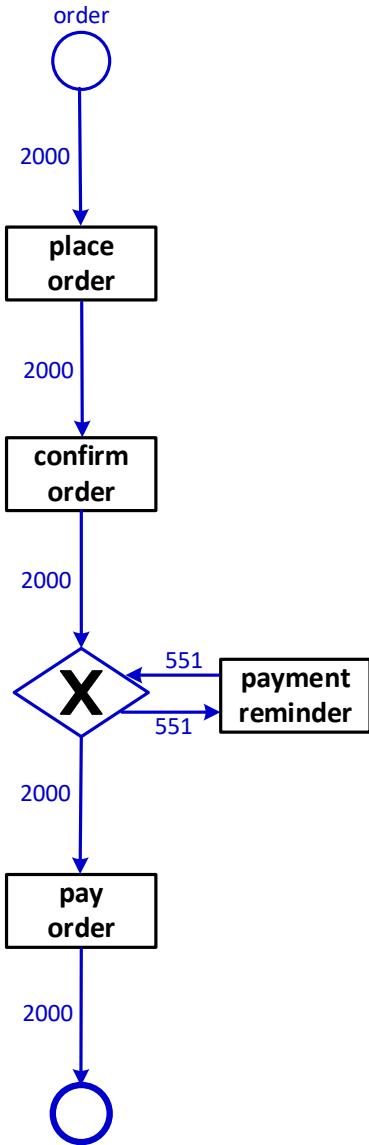


amazon

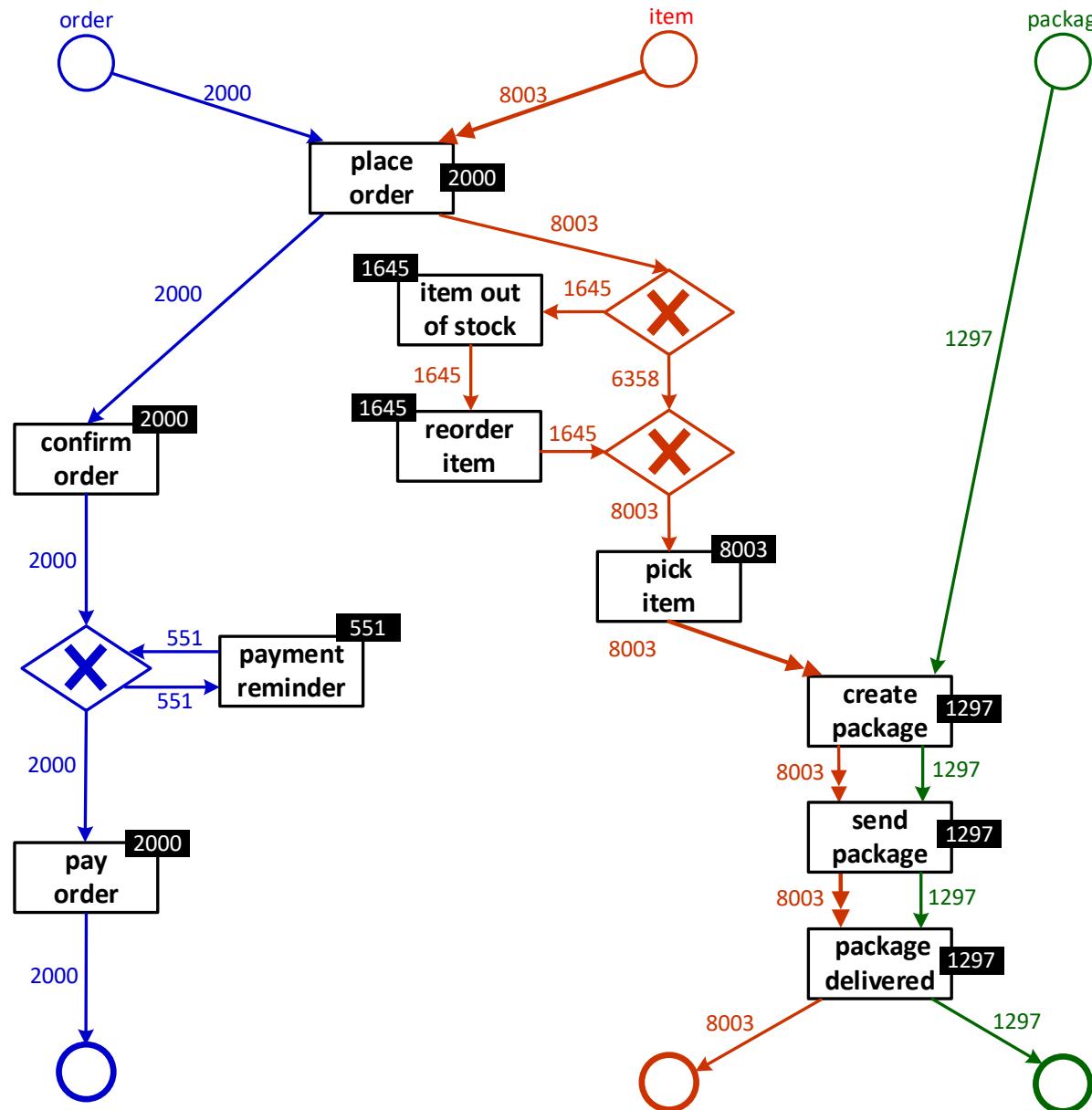
Flipkart, Myntra, Snapdeal, ...



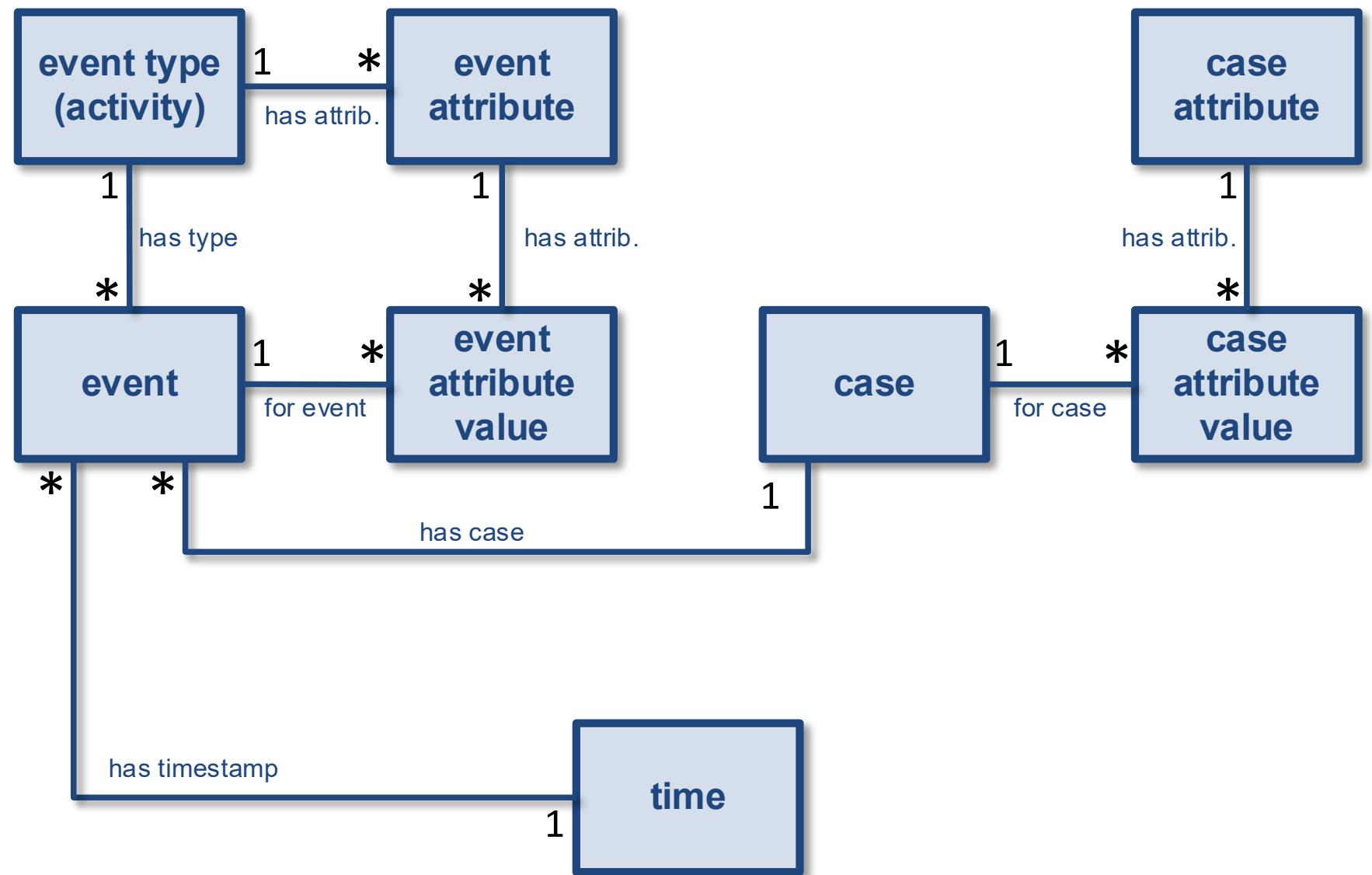
We cannot see the problems by looking at disconnected object types



Discovered Object-Centric Process Model



Meta Model: Case Centric



Meta Model: Object Centric

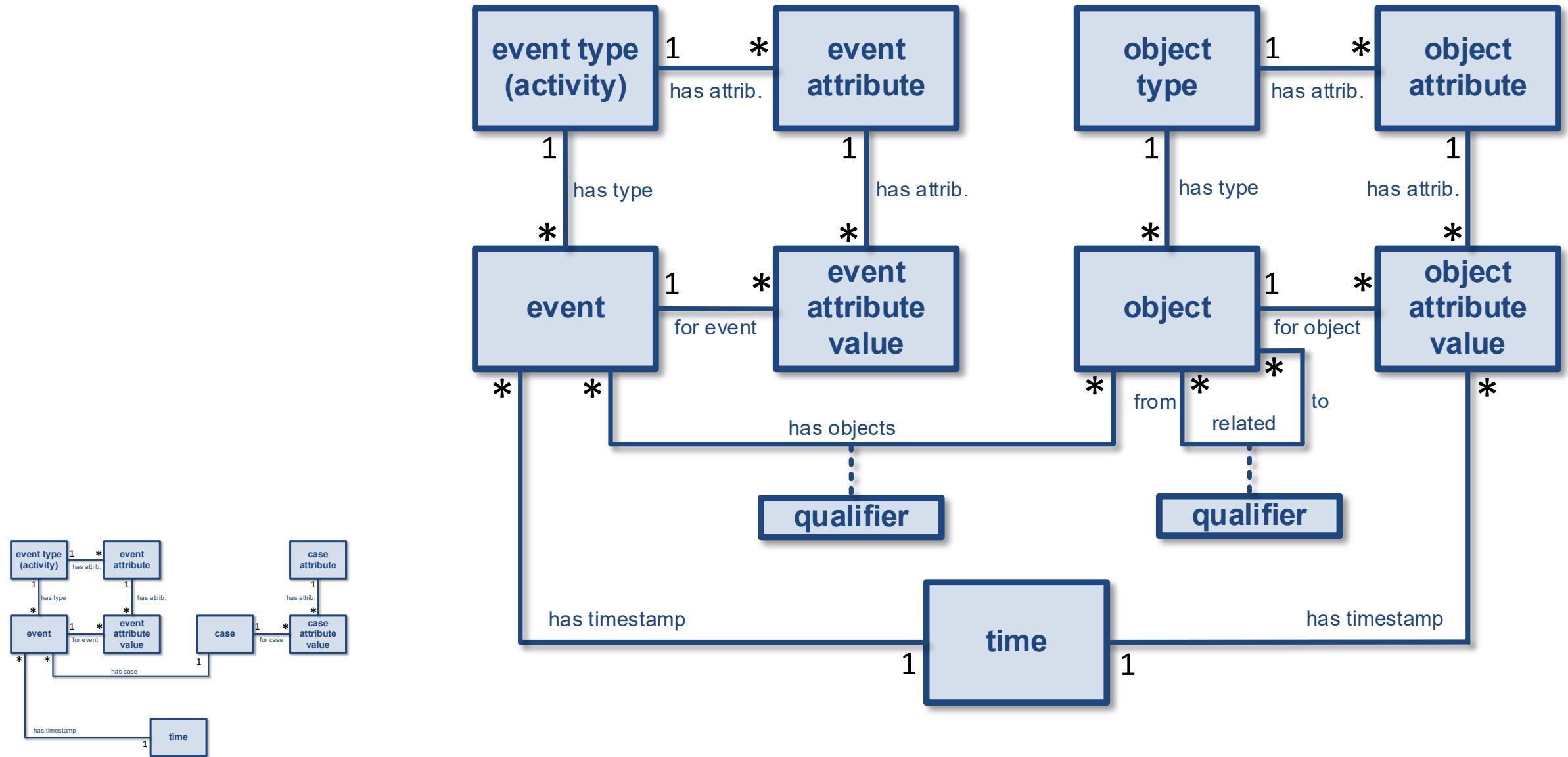
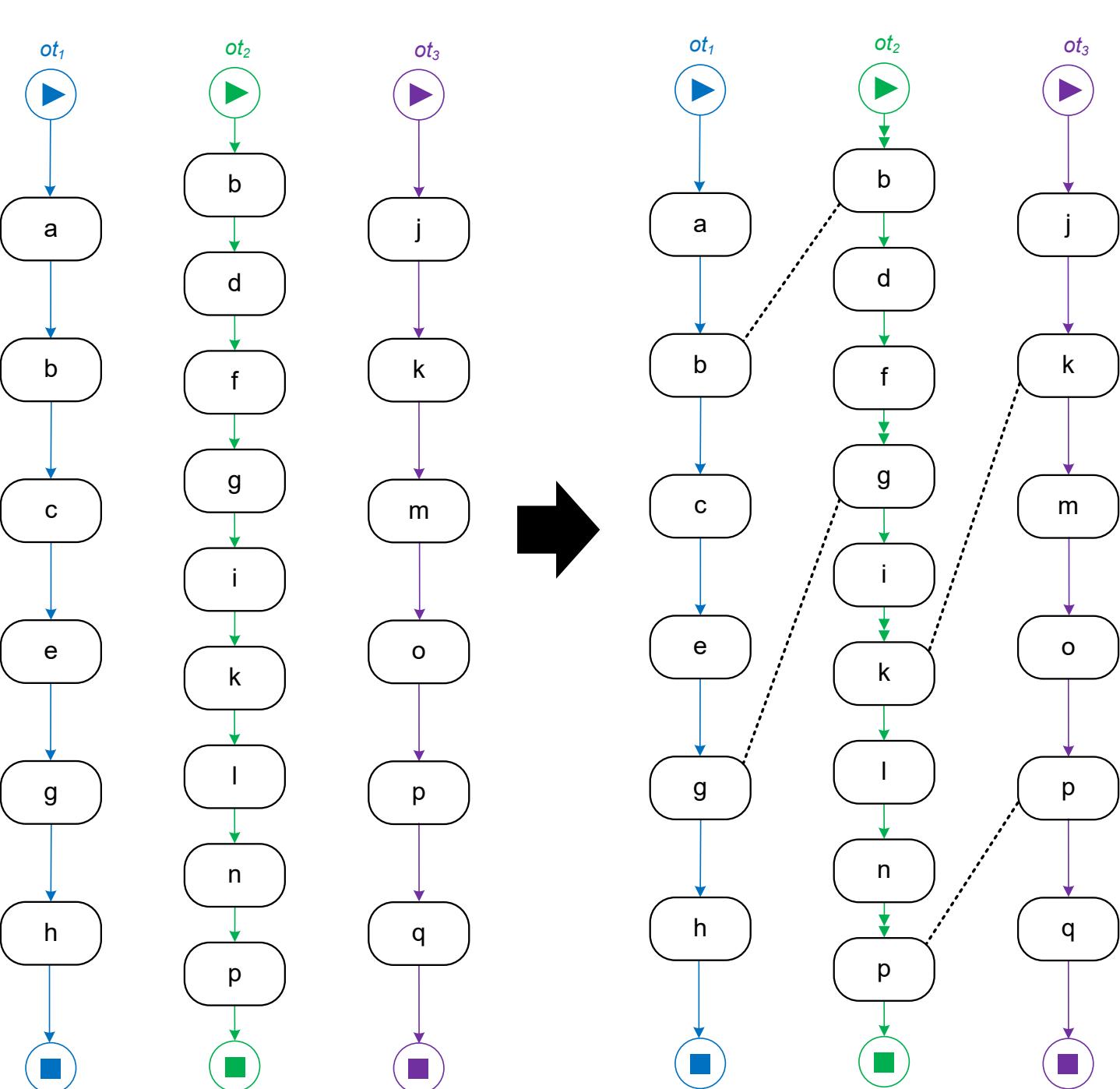
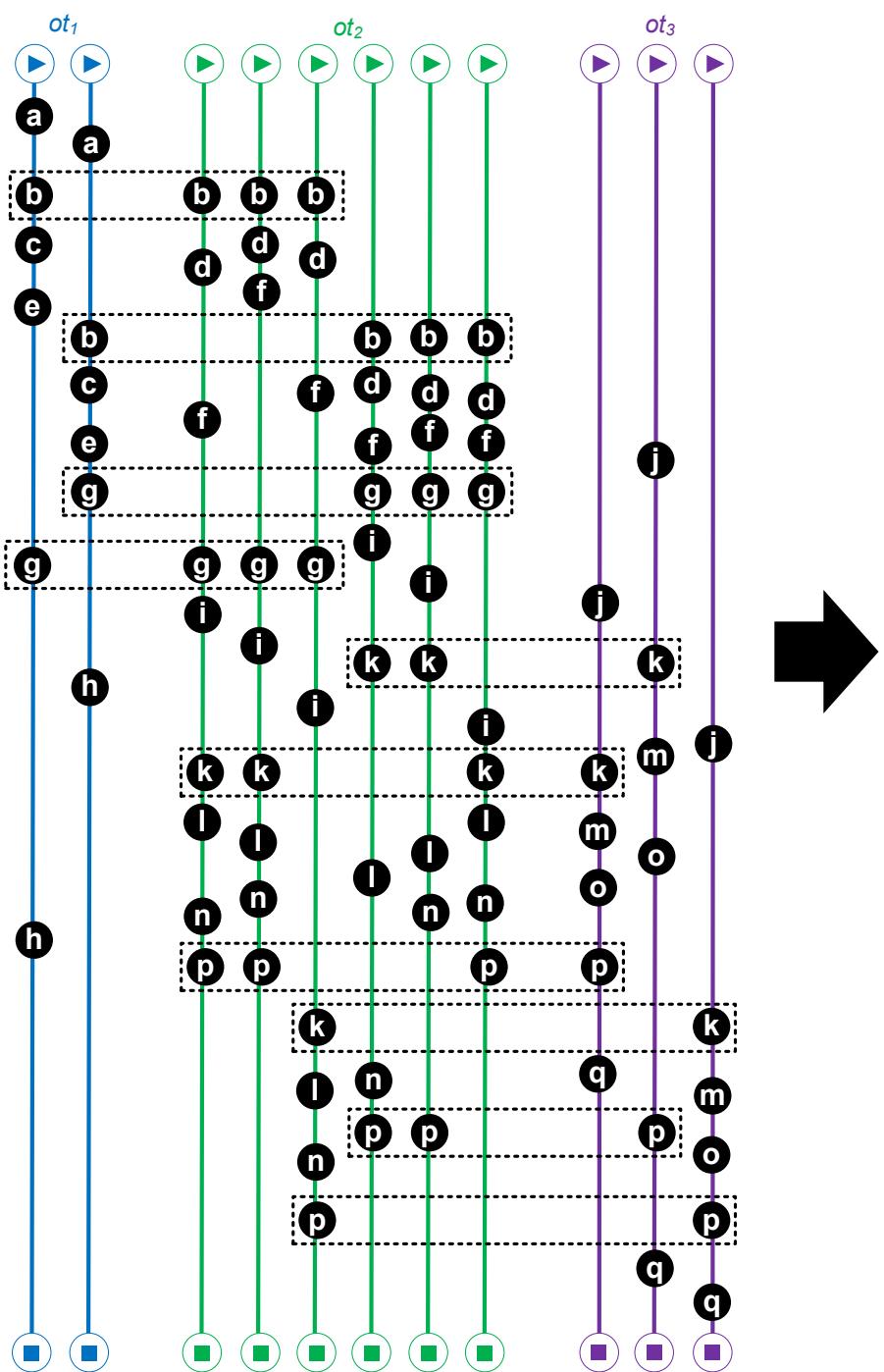
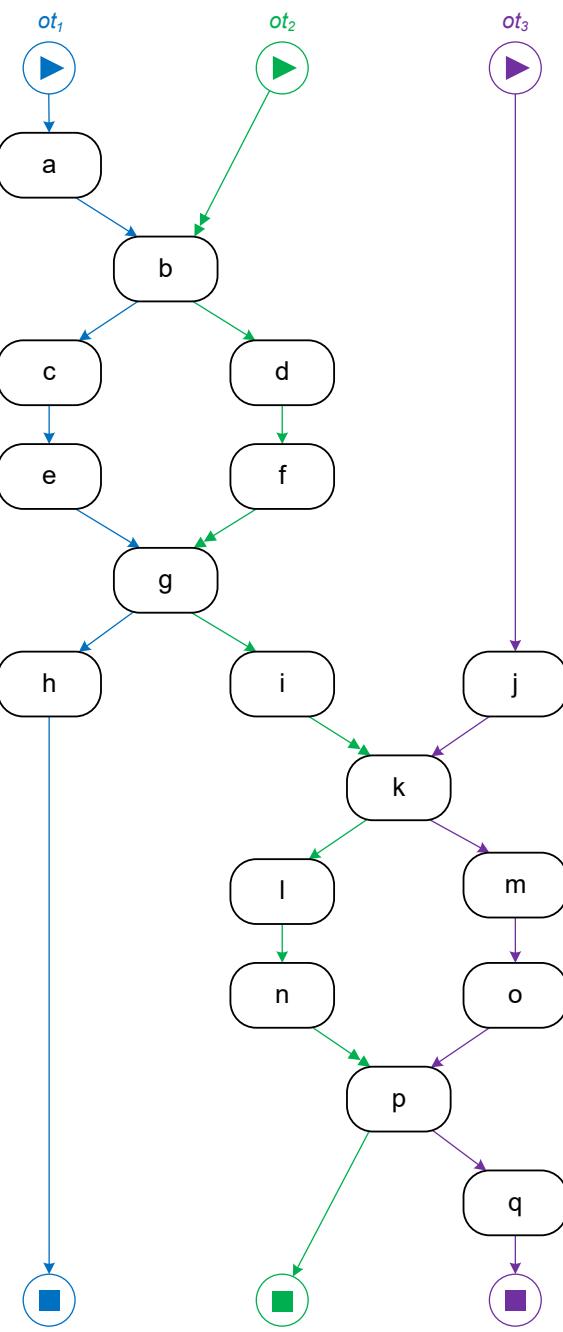
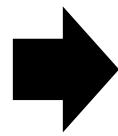
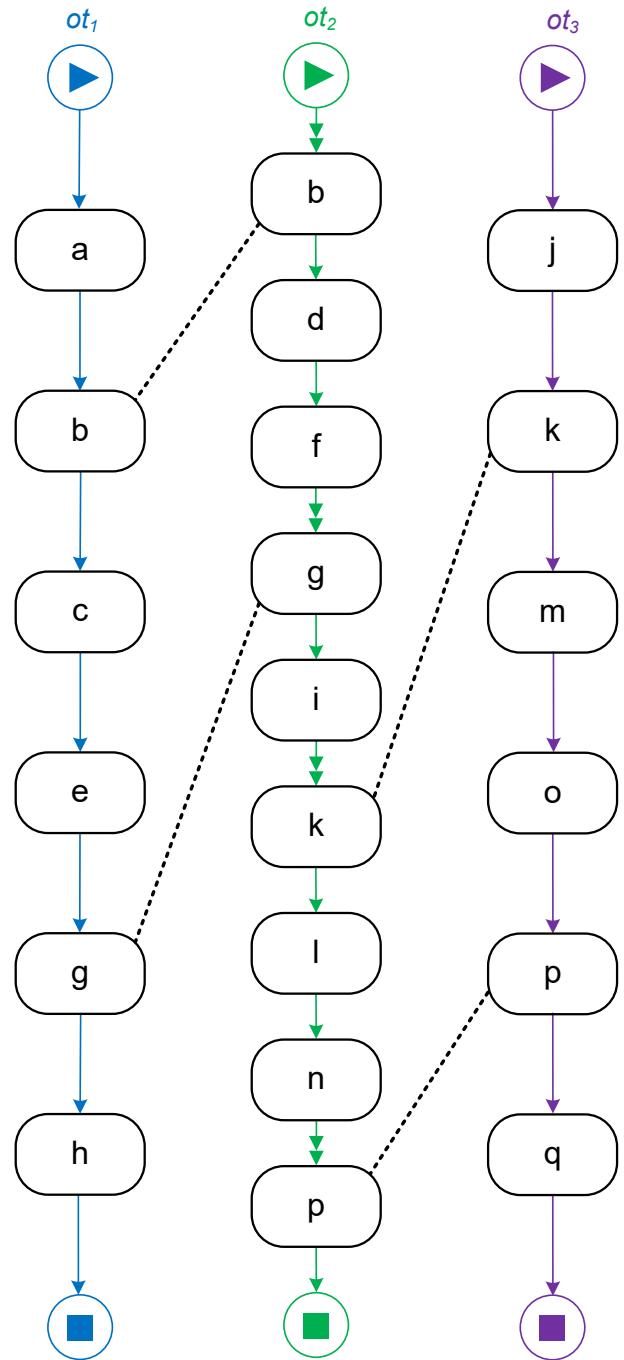
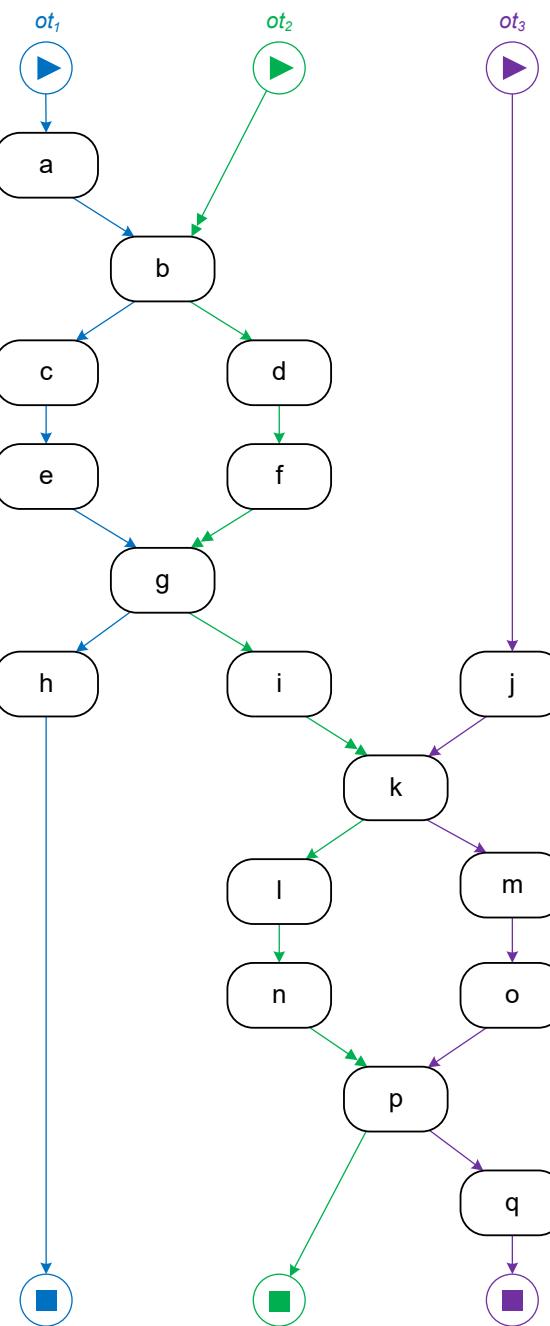
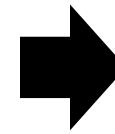
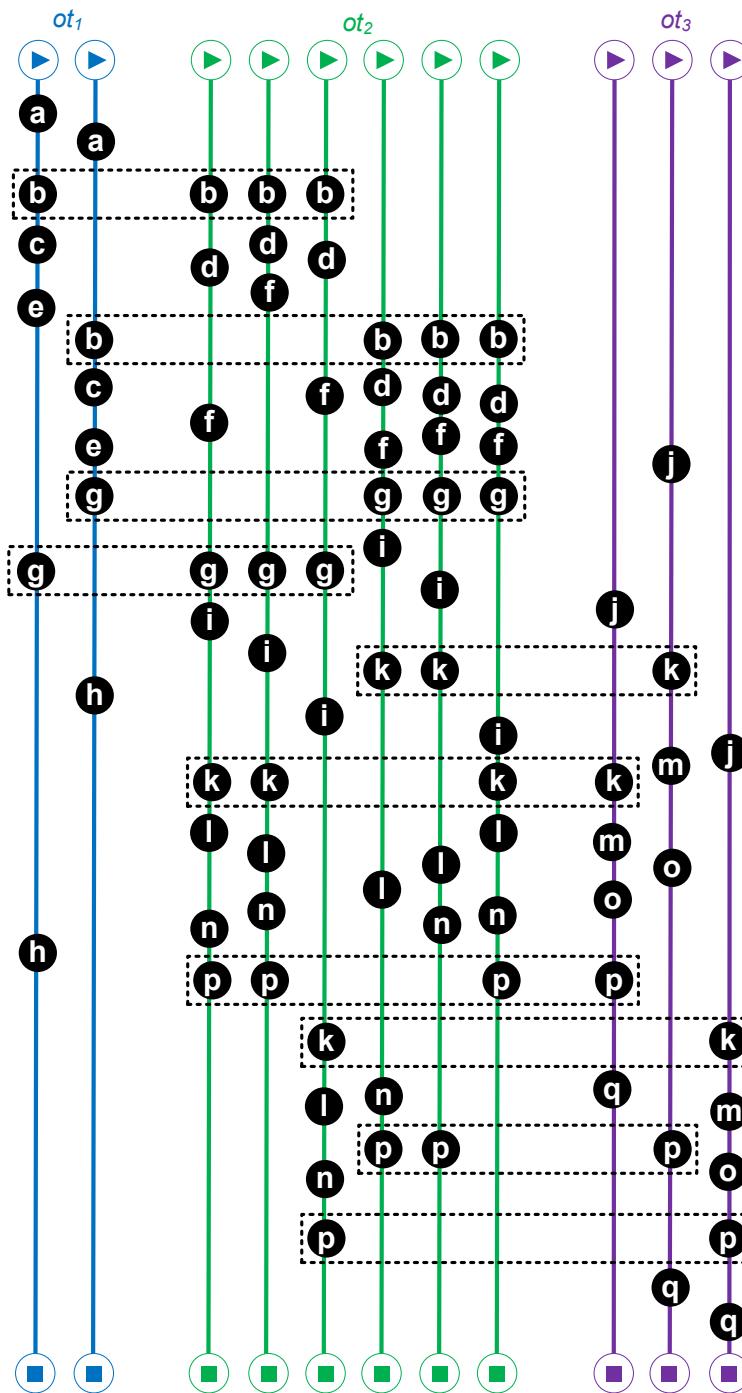
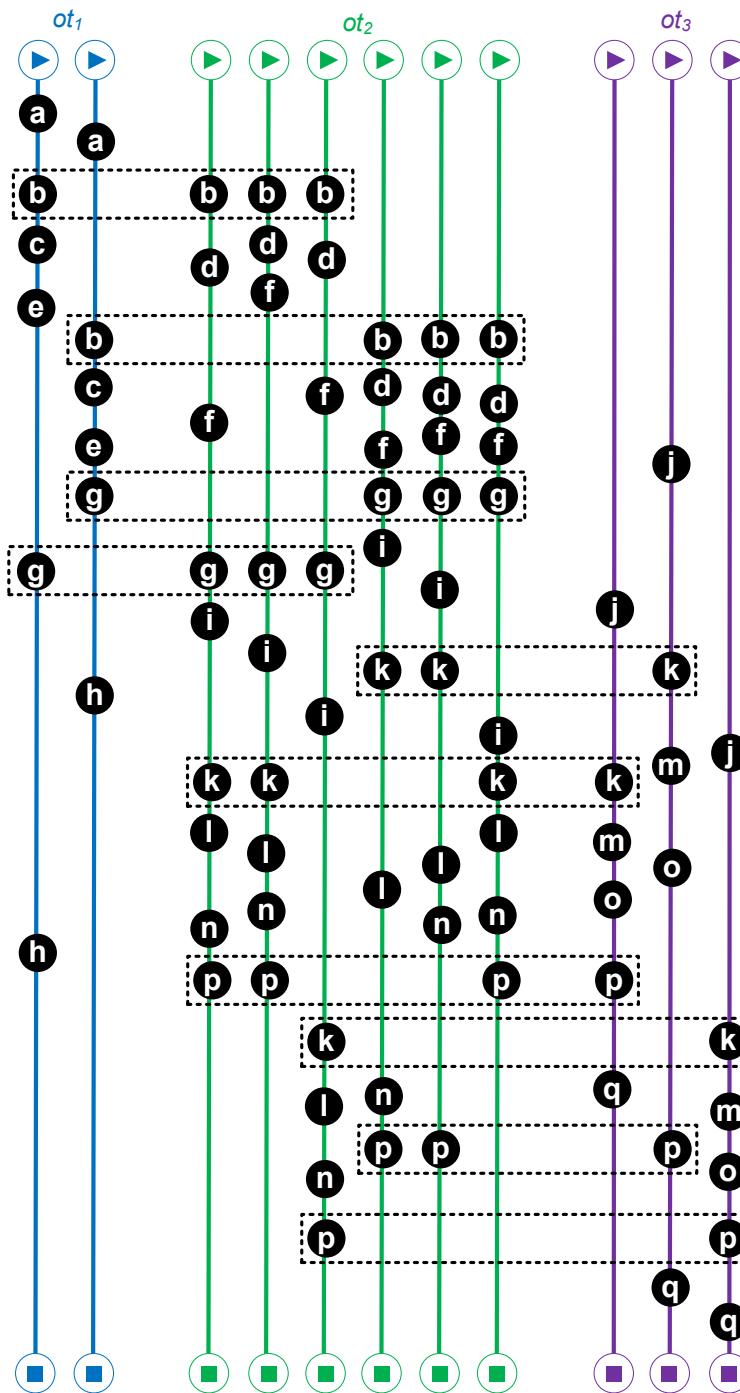


Exhibit #1

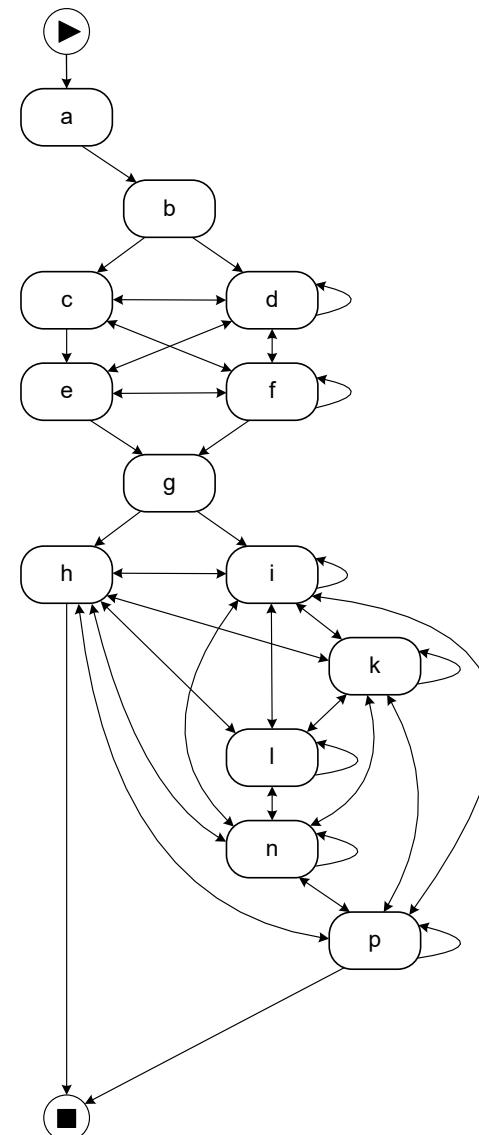




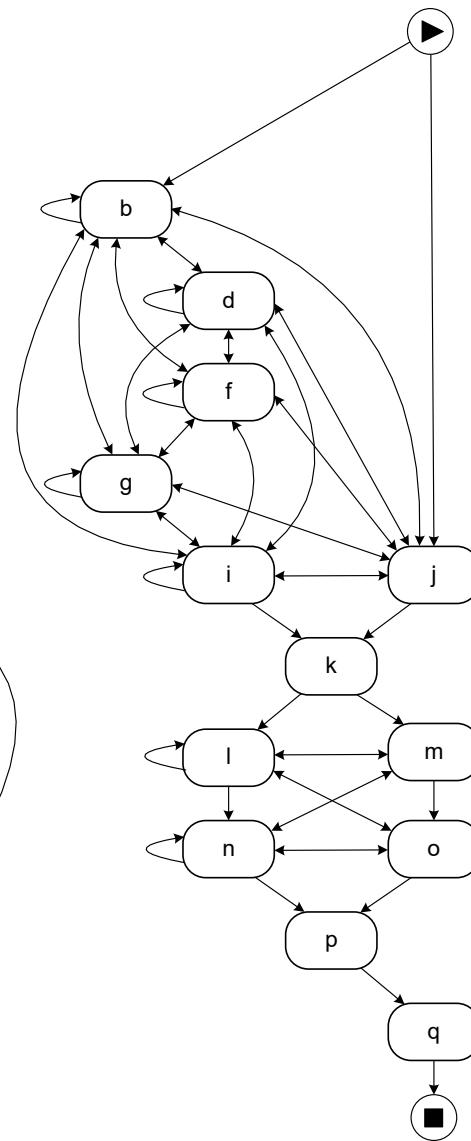




orders+items



packages+items



items+orders+packages

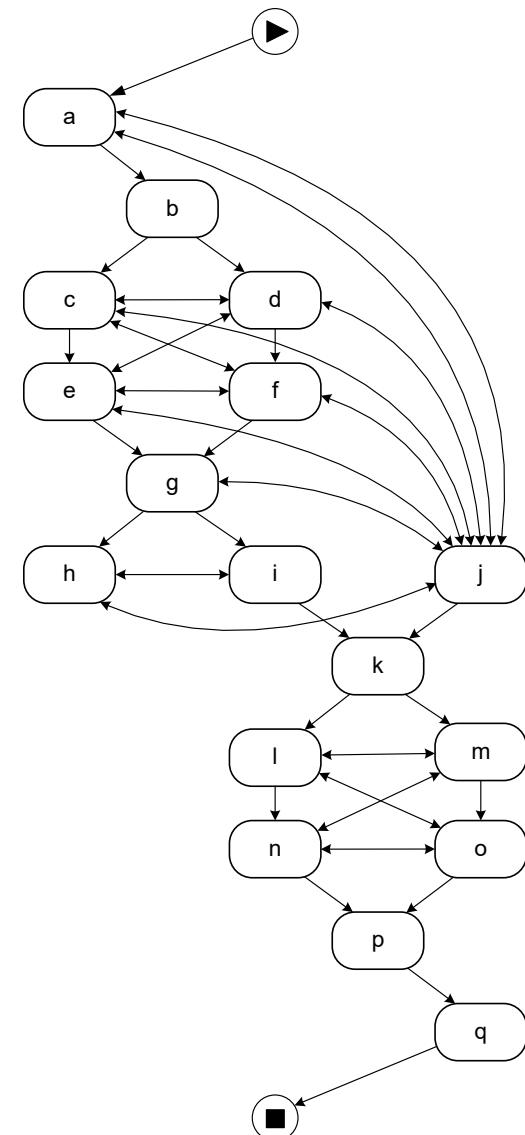
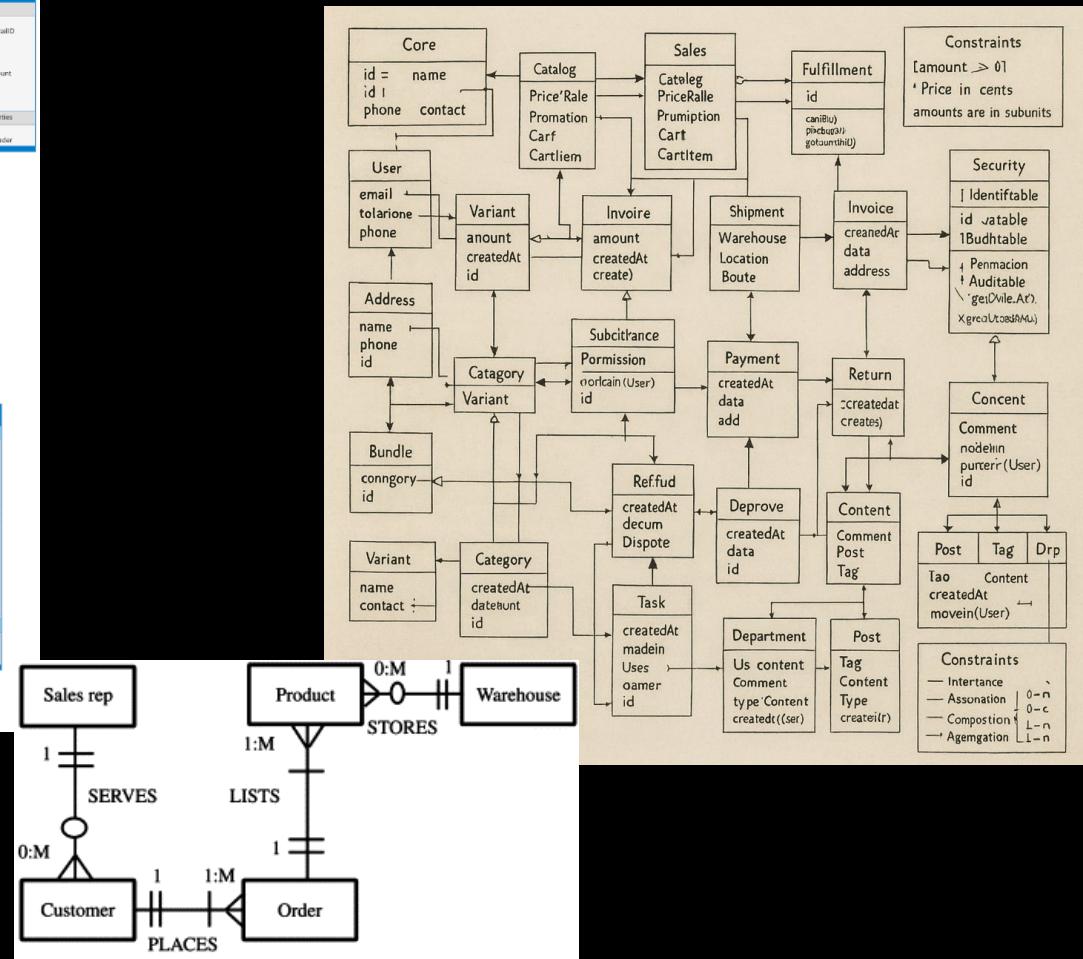
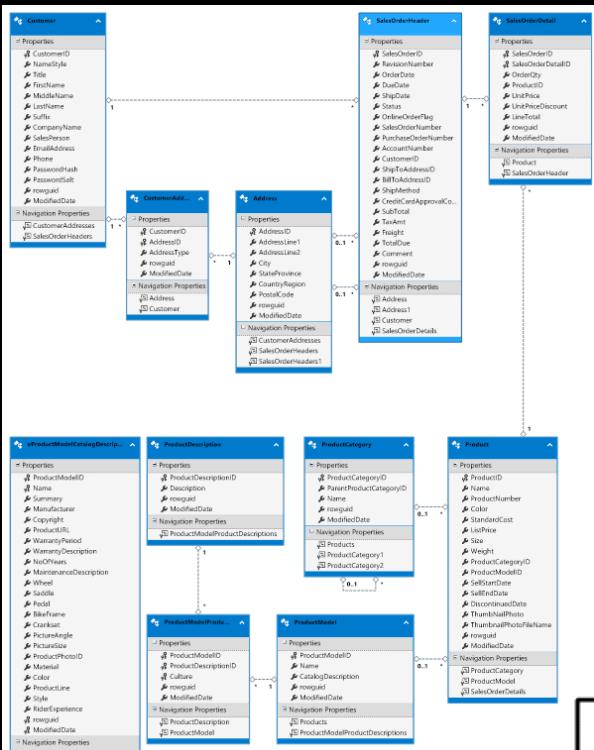
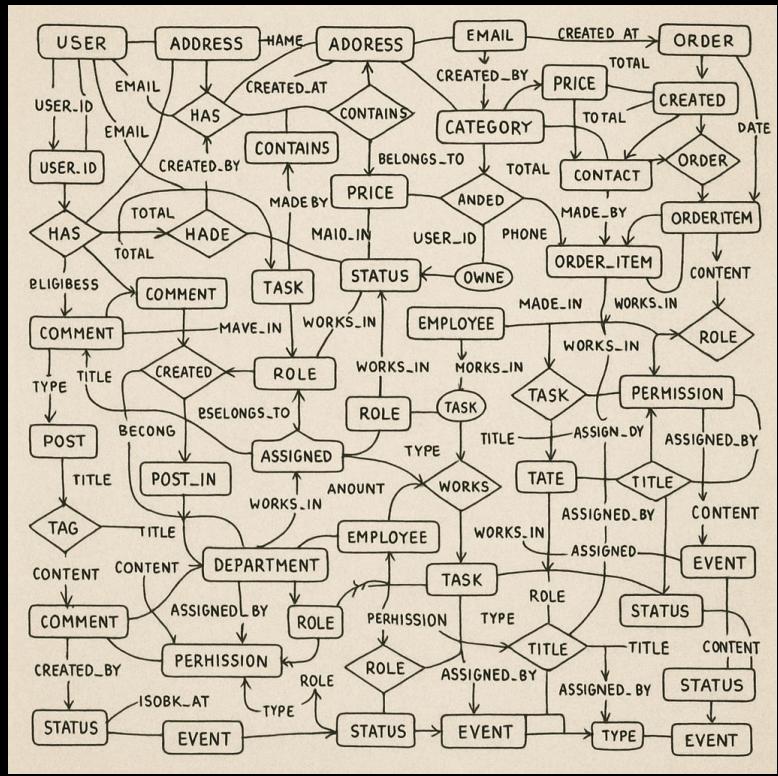


Exhibit #2

1-to-1 ?

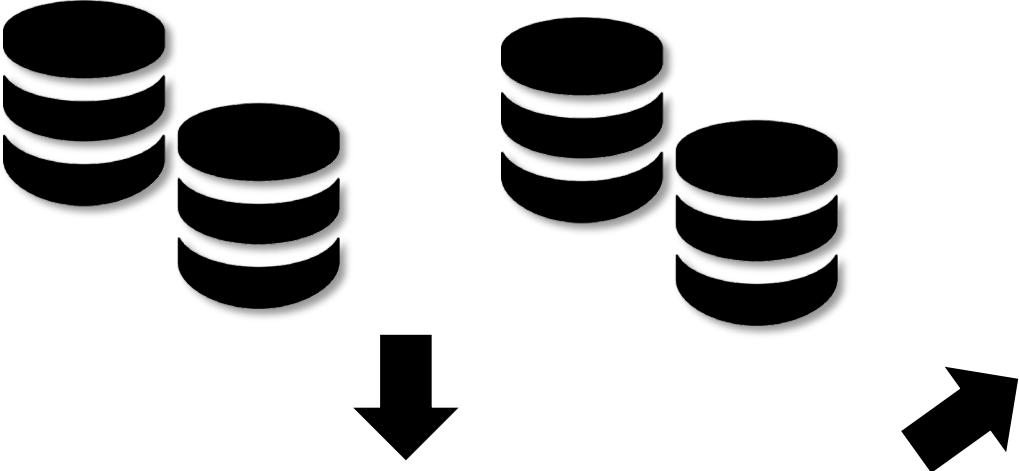


Customers place orders, doctors treat patients, patients have diseases, containers contain packages, cars have components, courses are attended by students, etc.

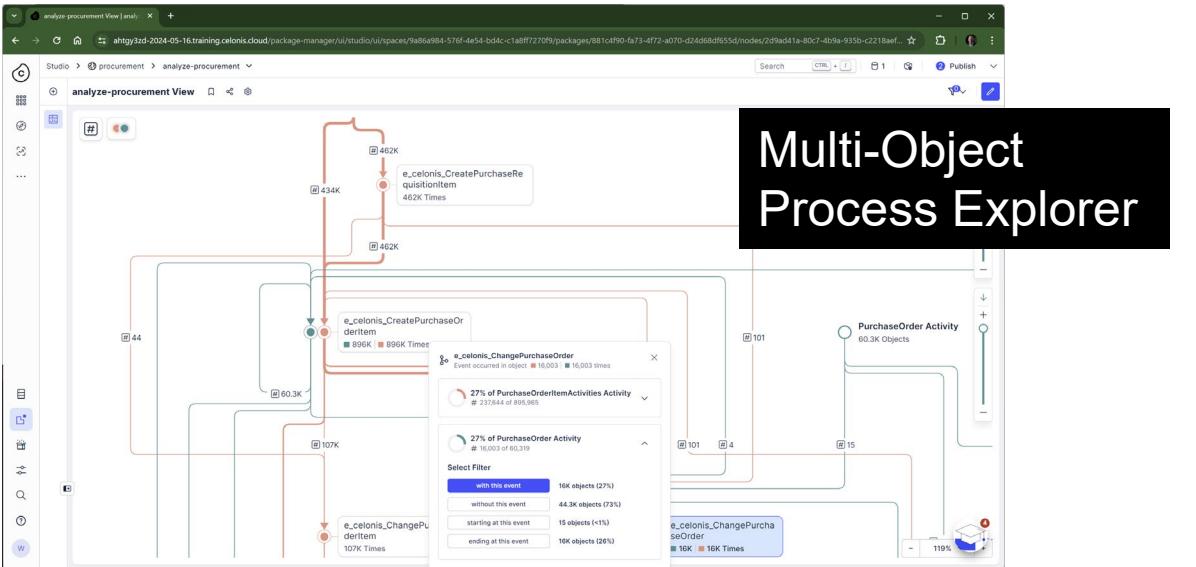
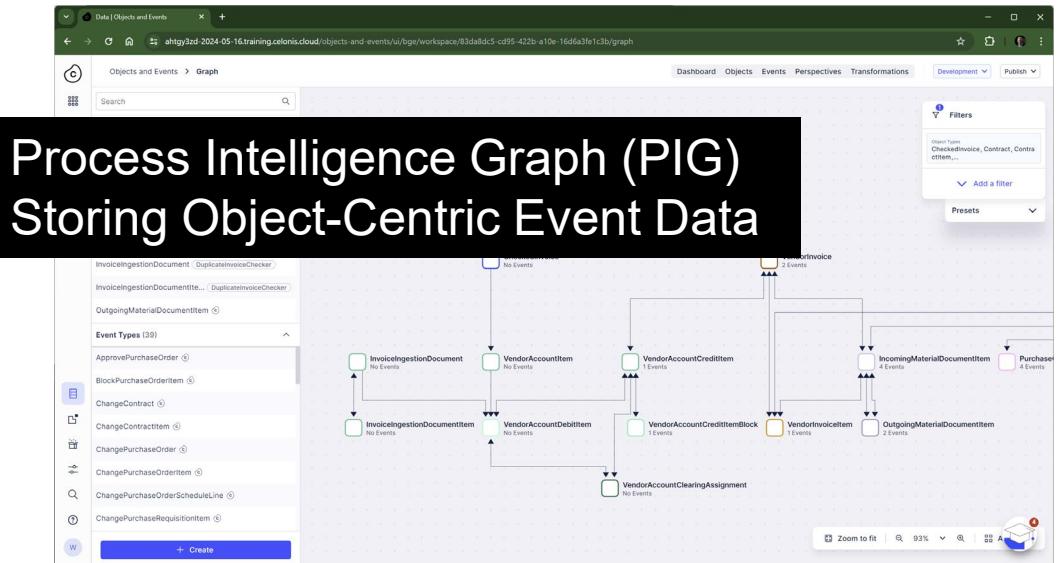
OCPM

Context Matters

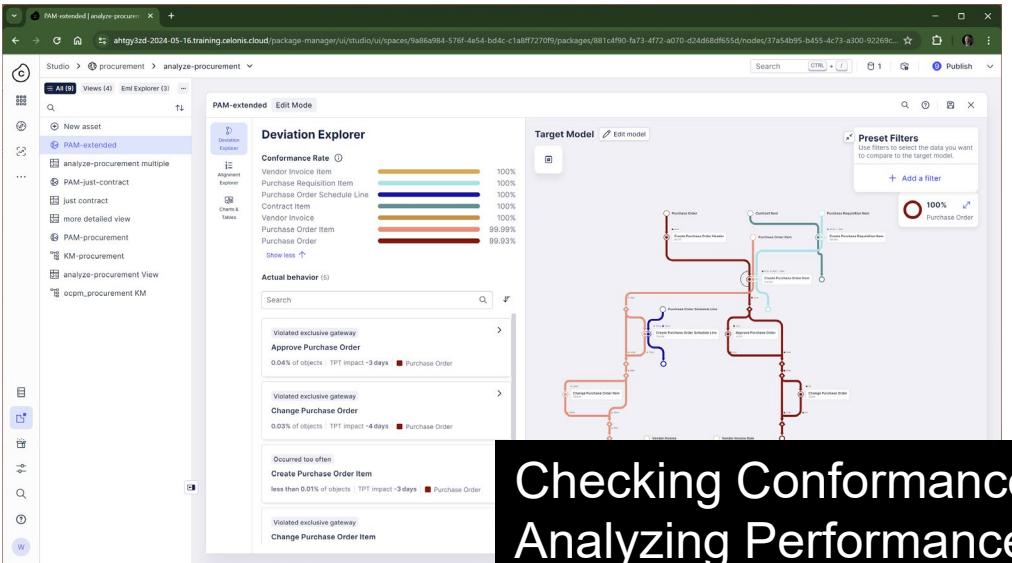
Celonis Supports OCPM



Process Intelligence Graph (PIG) Storing Object-Centric Event Data

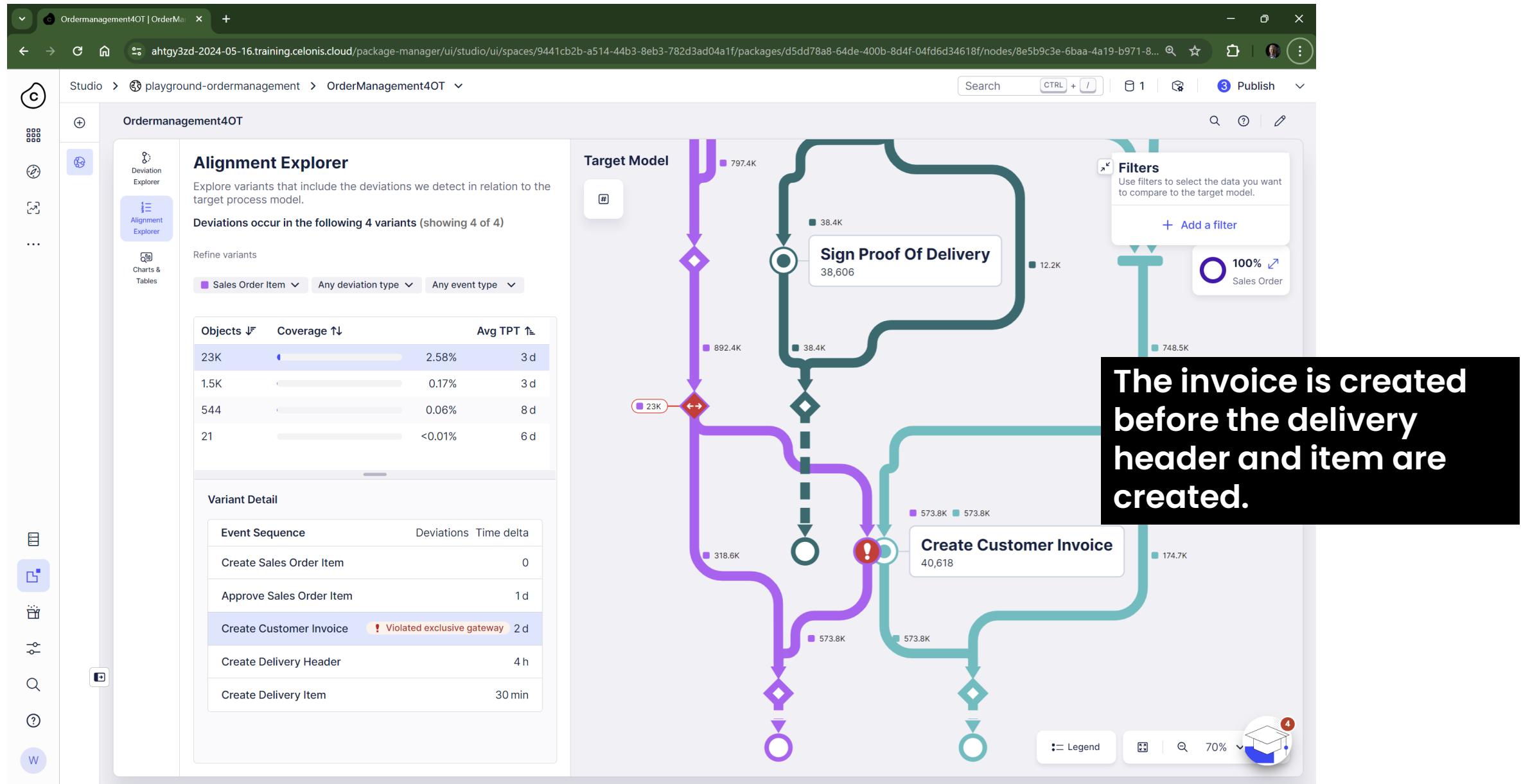


Multi-Object Process Explorer



Checking Conformance and Analyzing Performance

Conformance Checking Using Alignments



End-to-End Performance Analysis

Ordermanagement4OT | OrderManagement4OT

Studio > playground-ordermanagement > OrderManagement4OT

Search CTRL + / 1 Publish

Deviation Explorer

Conformance Rate

- Sales Order 99.99%
- Delivery 99.57%

Ordermanagement4OT

It is now possible to compute the time between the creation of the order and the delivery of all items in the order

Violated exclusive gateway

Post Goods Issue

1.55% of objects | TPT impact -24 hours | Delivery Item

Violated exclusive gateway

Sign Proof Of Delivery

0.43% of objects | TPT impact -11 days | Delivery

Occurred too often

Set Delivery Block

less than 0.01% of objects | TPT impact +9 days | Sales Order

Target Model

Throughput Time

Create Sales Order Header

Sales Order First

Create Customer Invoice

Delivery Item Last

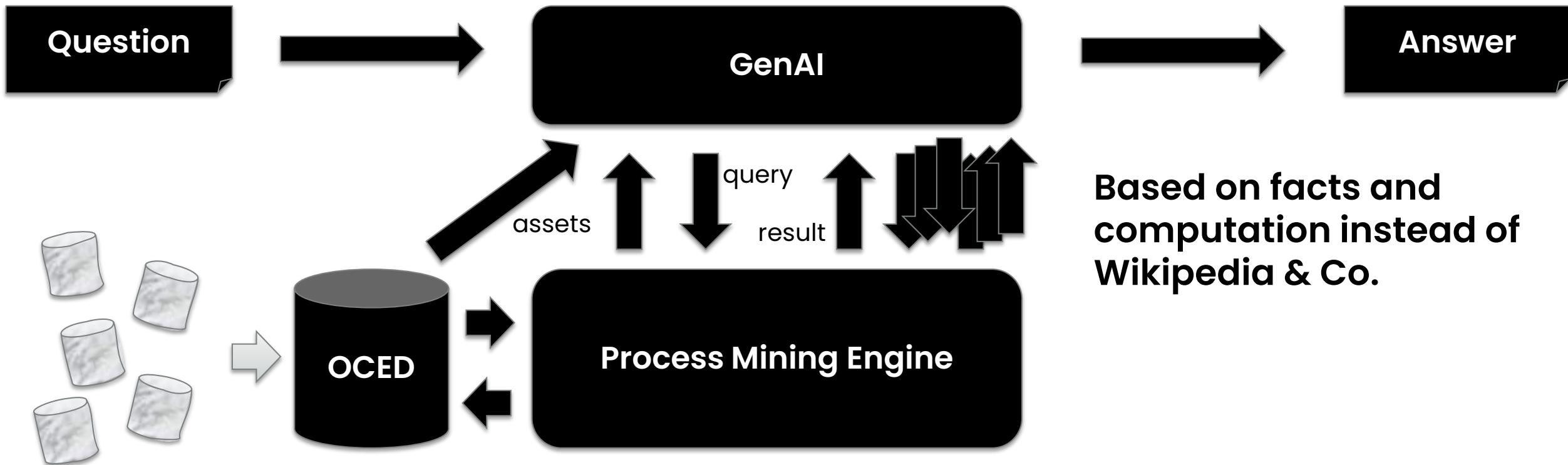
avg. 14.62 Days med. 12 Days

40,618 Sales Order objects are selected

Cancel Apply filter

vision

This is what we can do today!



Based on facts and computation instead of Wikipedia & Co.

**Why not do it more systematically as a community?
Let's stop just “playing” with general-purpose LLMs!**

Is this the (only) role we want to play?



Why not develop our own foundation models?



Millions of webpages containing the word Porsche

- **Porsche** 911: Die günstigste Therapie, die Stuttgart zu bieten hat.
- Auf der linken Spur fühlt sich **Porsche** wie zu Hause.
- Zwischen Ordnung und Wahnsinn fährt **Porsche** die Ideallinie.
- Wenn andere noch träumen, startet **Porsche** schon den Motor.
- **Porsche** in Stuttgart gebaut, auf der Autobahn geboren.
- Wenn Präzision Emotion trifft, entsteht **Porsche**.
- Wo Leidenschaft auf Technik trifft, entsteht **Porsche**.
- In Stuttgart geboren, auf der Autobahn zuhause – das ist **Porsche**.
- Zwischen Null und Hundert sagt **Porsche** nur „Guten Morgen“.
- Geduld ist schön – aber **Porsche** ist schöner.

Millions of webpages containing the word Porsche

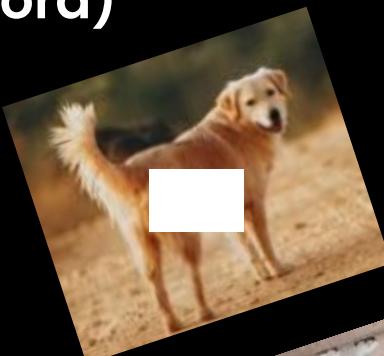
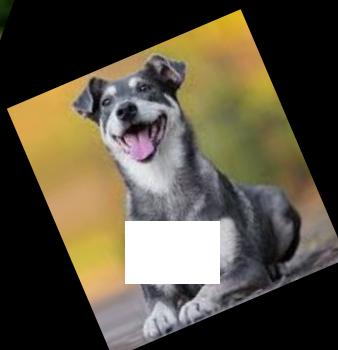
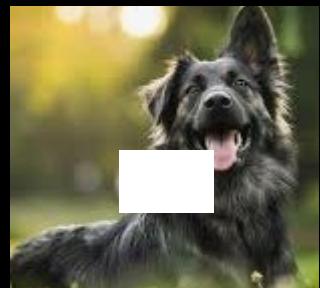
- Porsche 911: Die günstigste Therapie, die Stuttgart zu bieten hat.
- Auf der linken Spur fühlt sich Porsche wie zu Hause.
- Zwischen Ordnung und Wahnsinn fährt Porsche die Ideallinie.
- Wenn andere noch träumen, startet Porsche schon den Motor.
- Porsche in Stuttgart gebaut, auf der Autobahn geboren.
- Wenn Präzision Emotion trifft, entsteht Porsche.
- Wo Leidenschaft auf Technik trifft, entsteht Porsche.
- In Stuttgart geboren, auf der Autobahn zuhause – das ist Porsche.
- Zwischen Null und Hundert sagt Porsche nur „Guten Morgen“.
- Geduld ist schön – aber Porsche ist schöner.

Similar concepts for images

Millions of webpages containing dog pictures

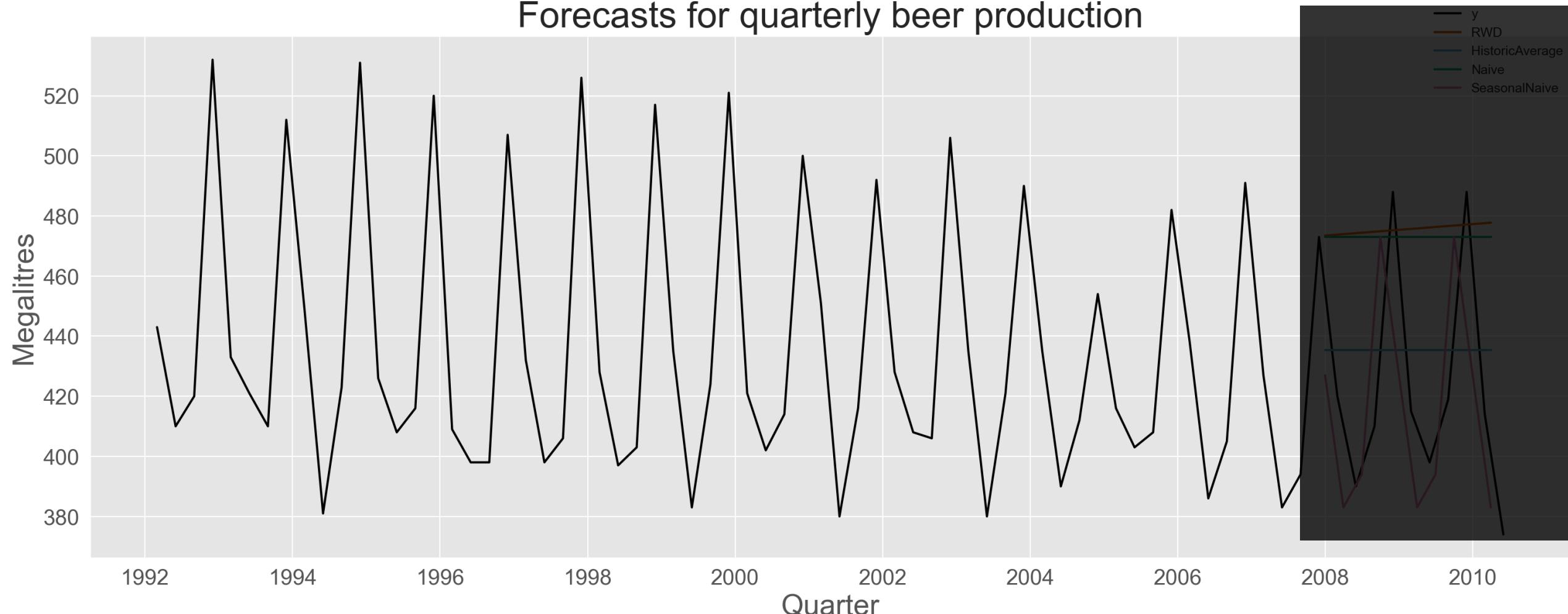


Repair the pictures (like filling in the missing word)

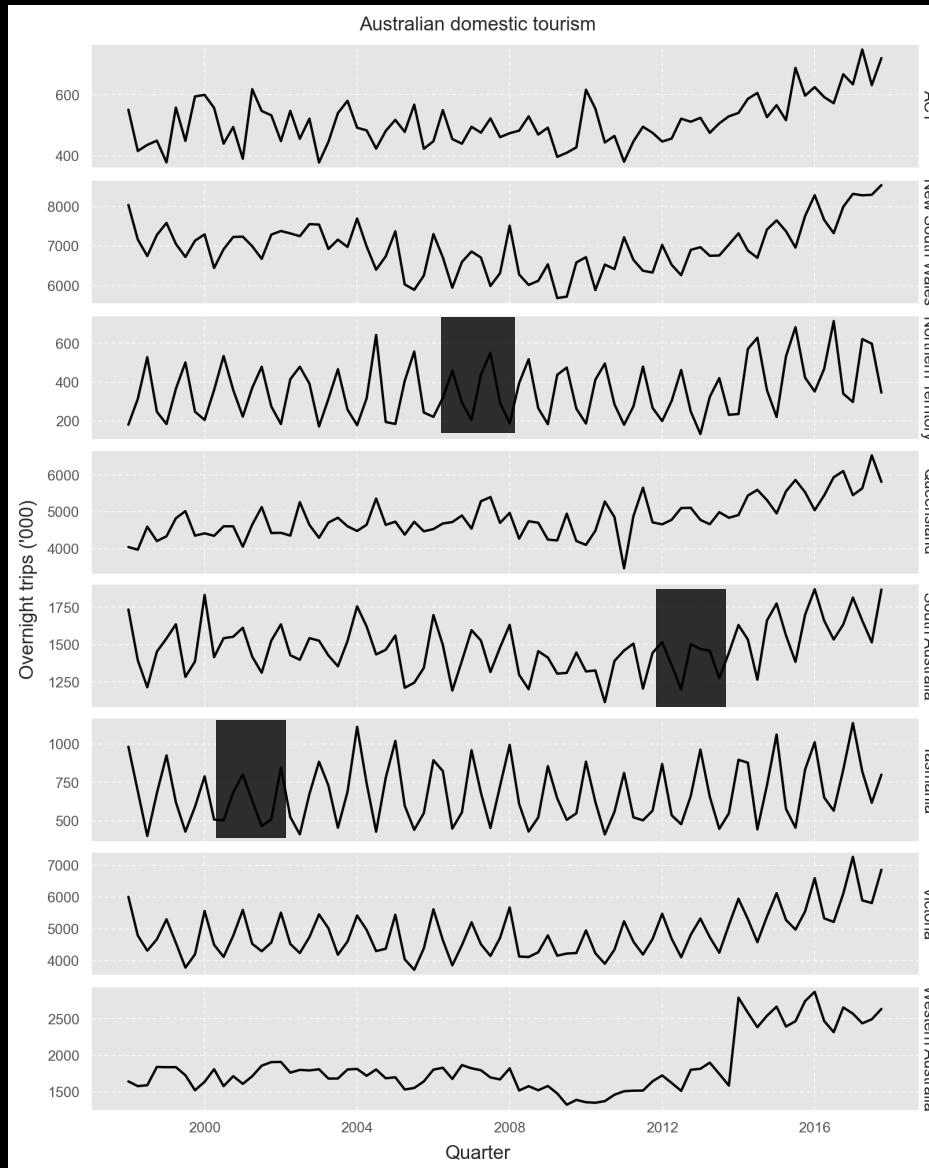


Similar concepts for time series (but already more difficult)

Forecasts for quarterly beer production



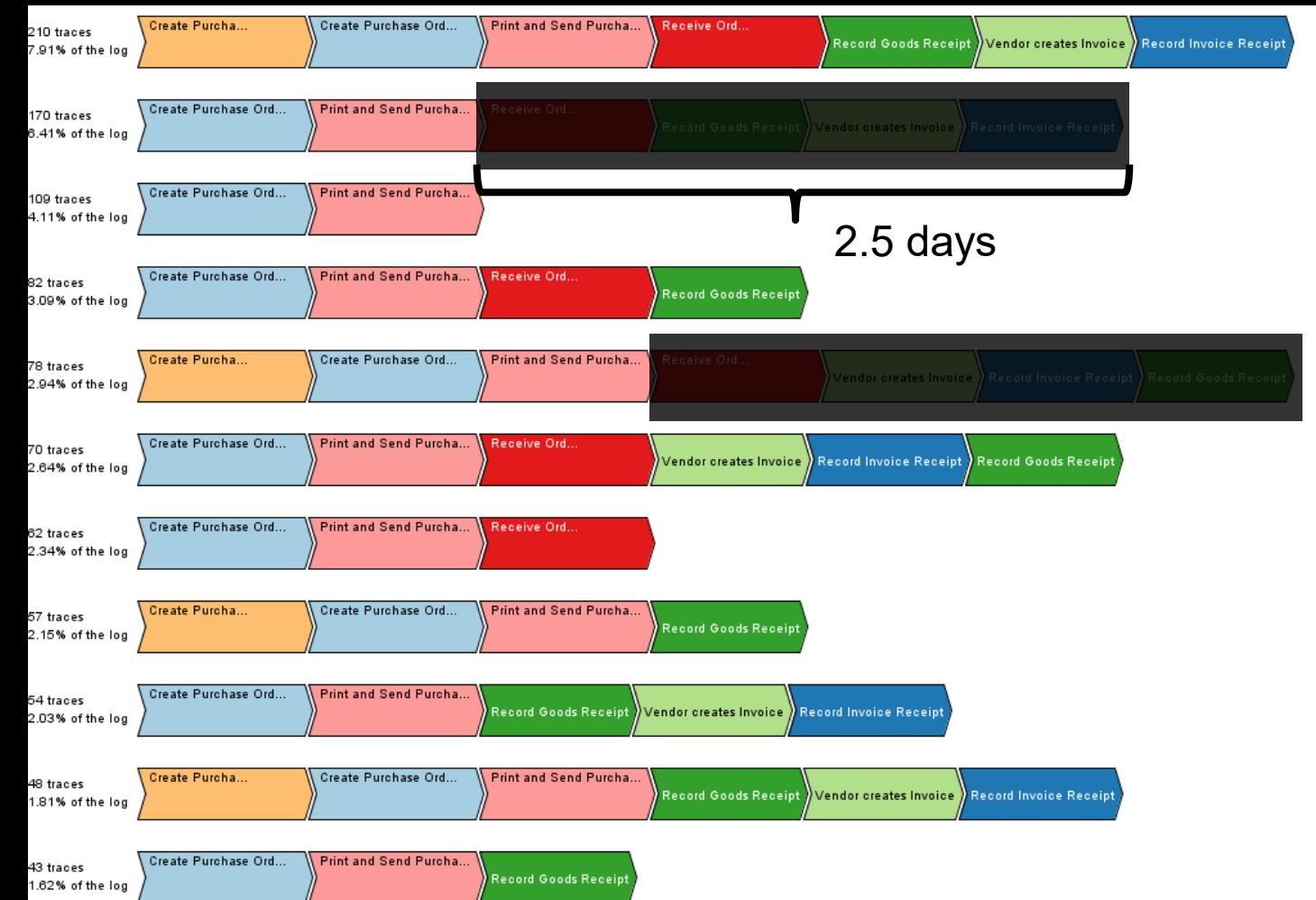
Similar concepts for time series (but already more difficult)



Challenges

- **What does 456.3 mean? (compare to “Porsche”)**
- **Domain specific**
- **Less public data**

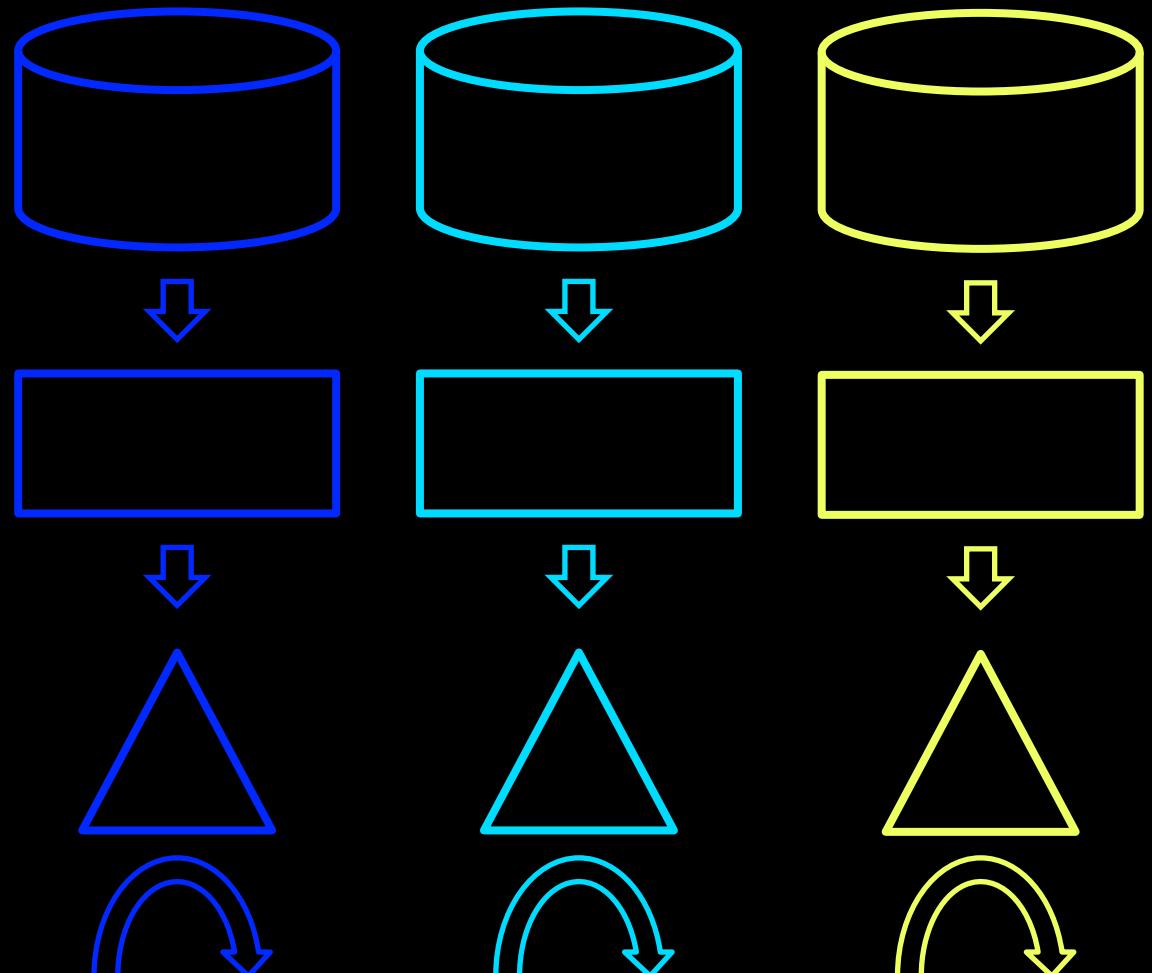
How about event data?



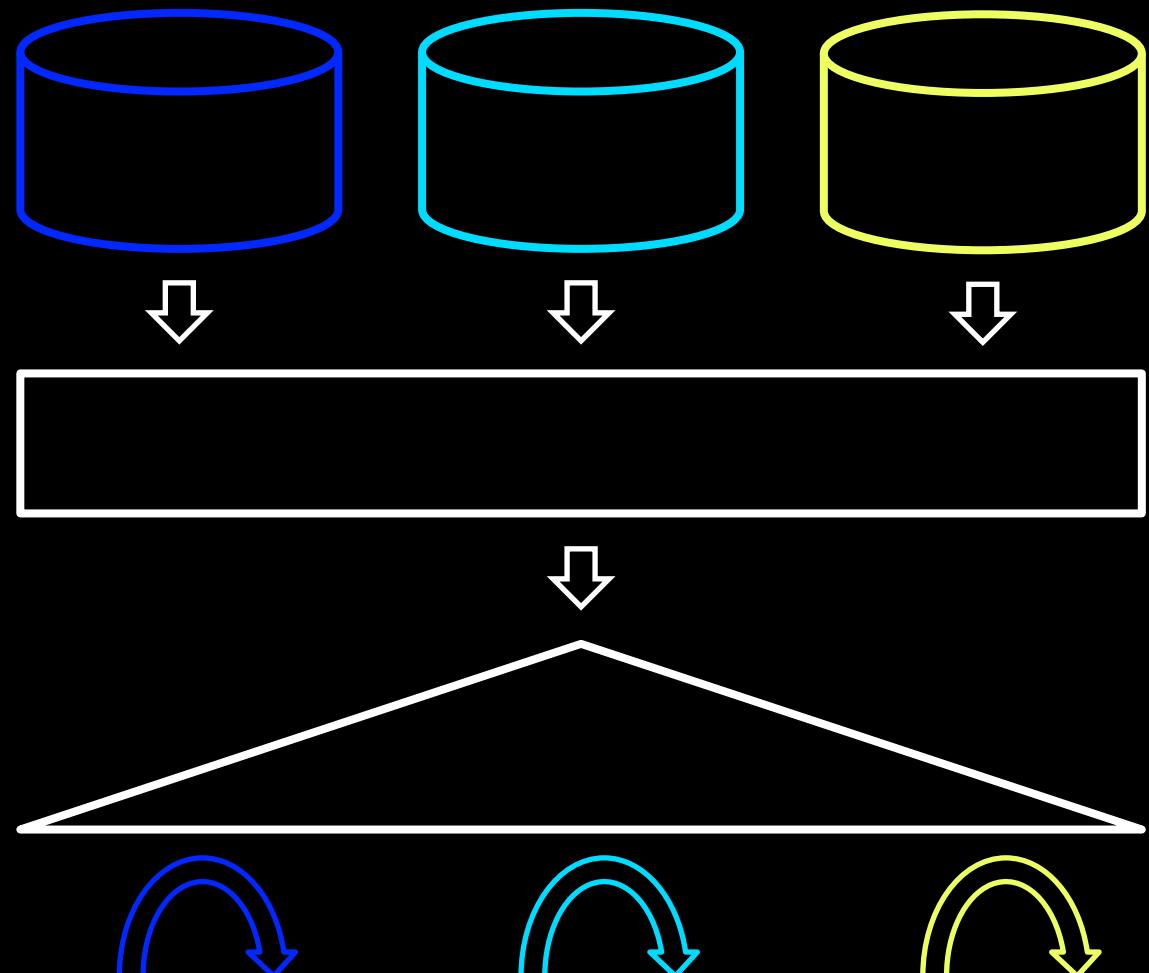
Challenges

- Do we want to use the fact that “Create XYZ” is likely to be at the start?
- Domain specific
- Less public data

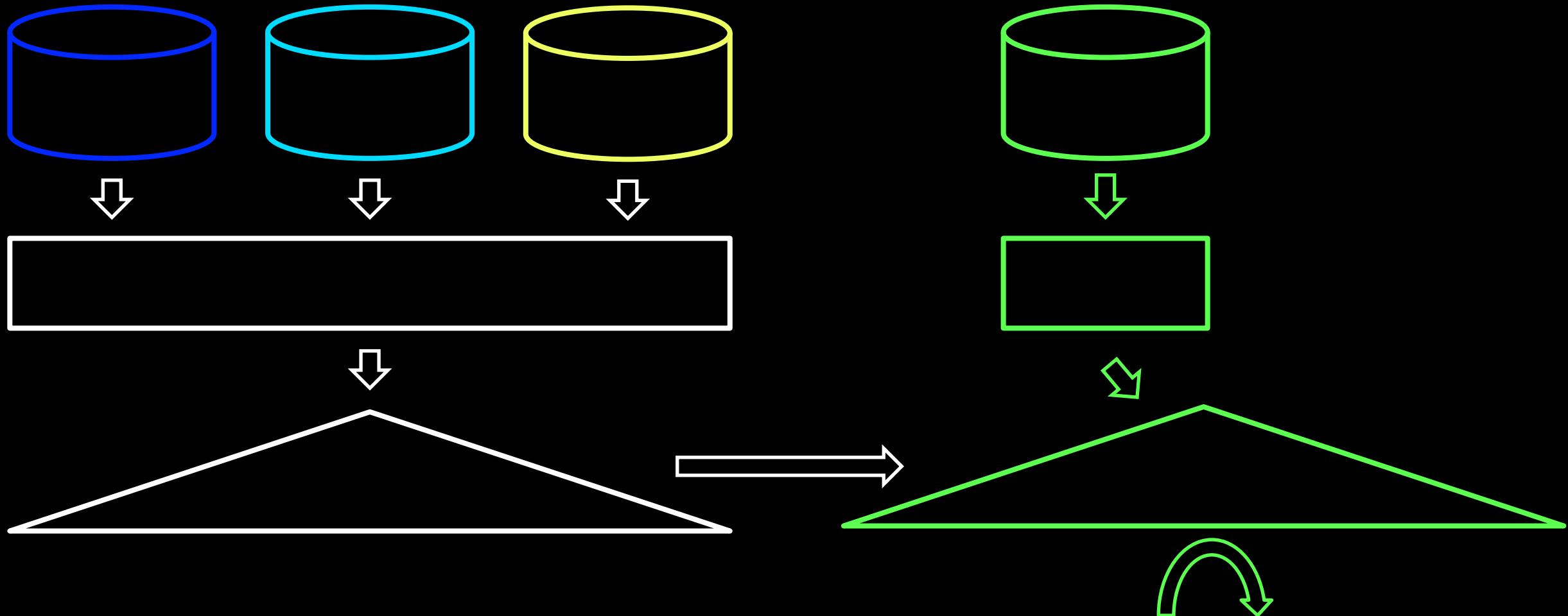
When is one general model better than many specific models?



vs



Probably a mix is better ...

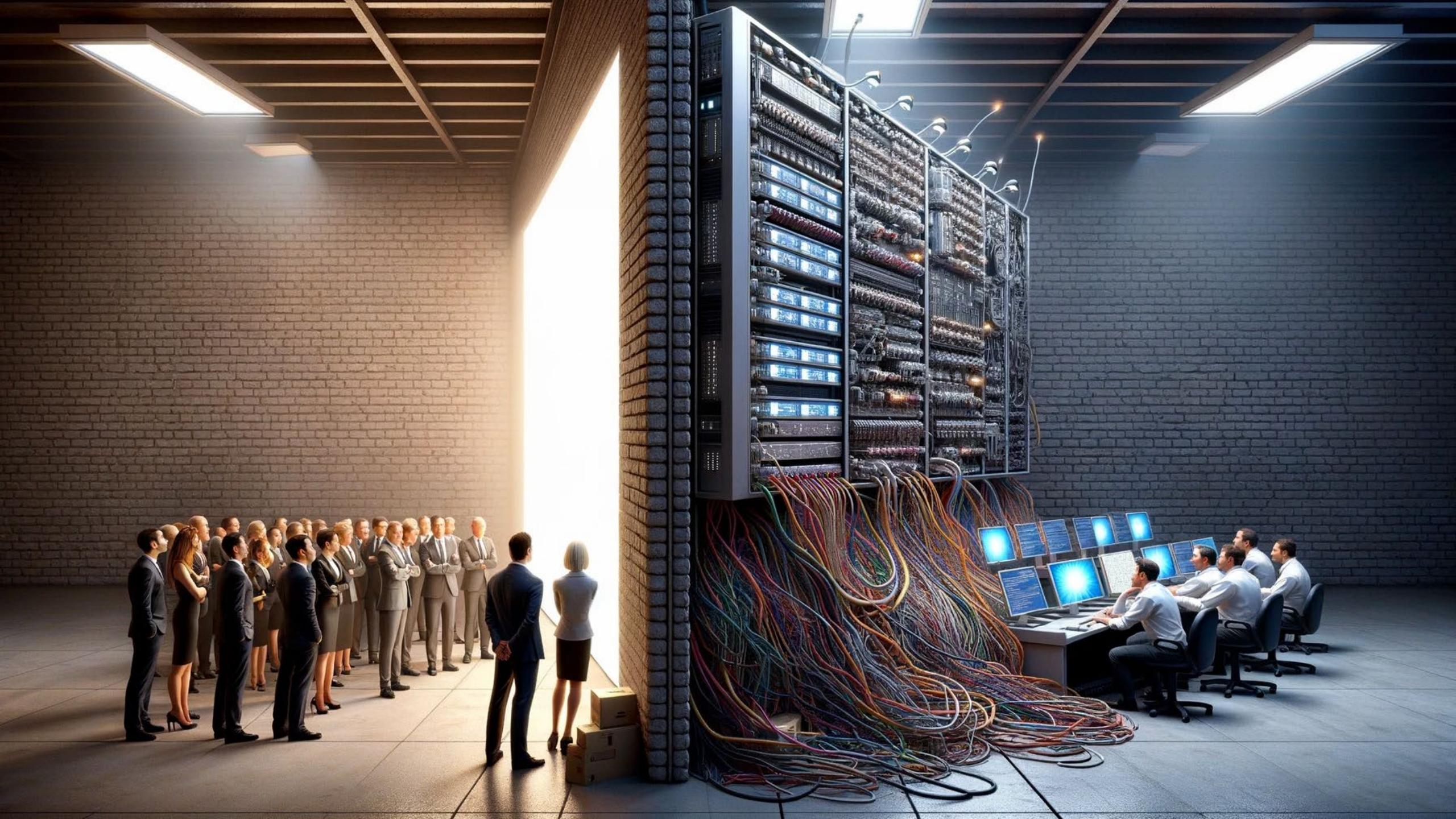


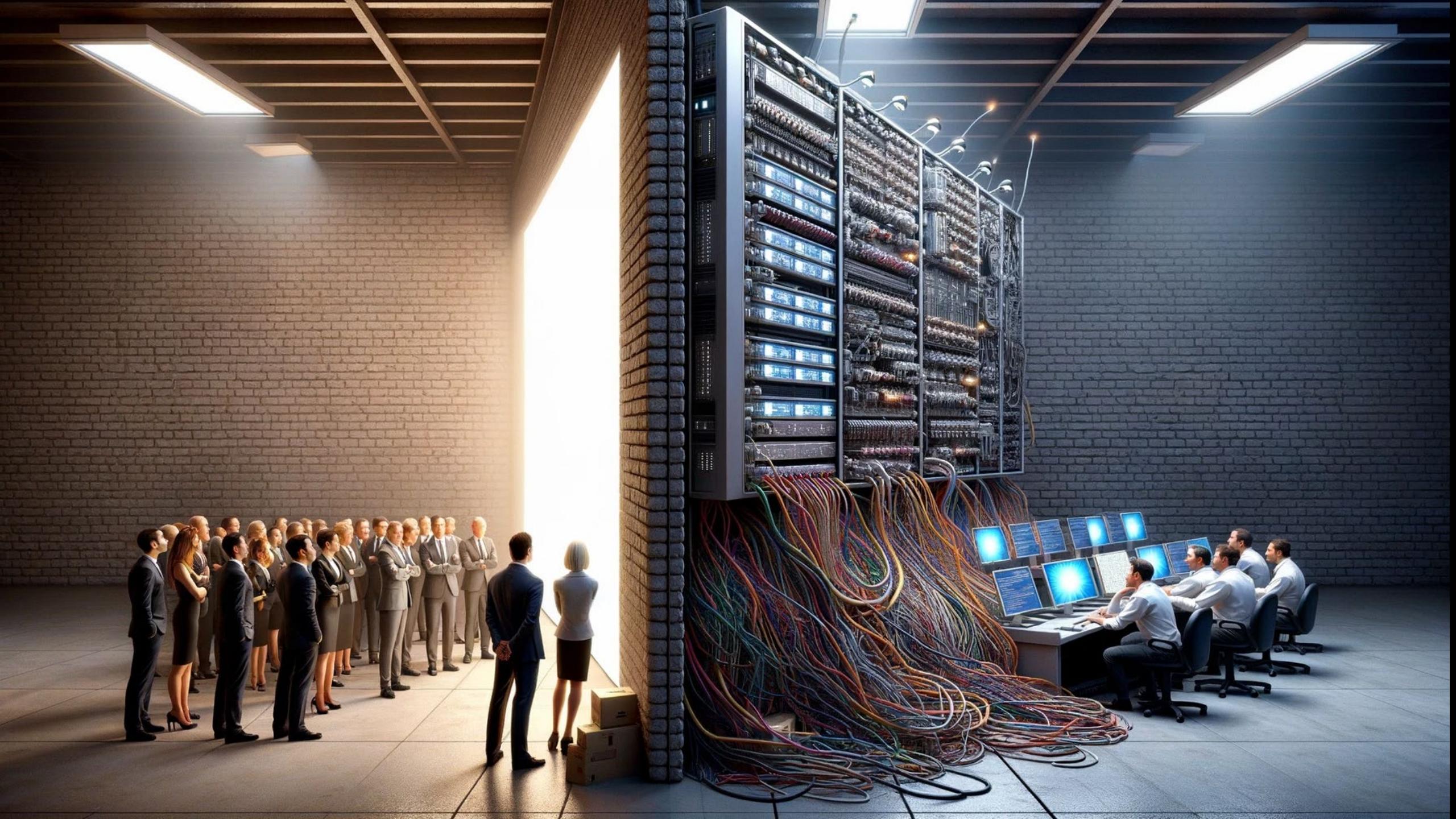
Recall the “No Free Lunch” (NFL) theorems

“All learning algorithms are equivalent, on average”
(David Wolpert 1992)

Meaningful learning is only possible if the model is trained on data from a similar distribution (in the broadest sense of the word) as the unseen data it is applied to.

CONCLUSION





Pointers to LLM research RWTH

- Context is important:
OCPM – OCPM – OCPM

- Towards foundation
models for processes?

