Hands-on Process Discovery with Python

Utilizing Jupyter Notebook for the Digital Assistance in Higher Education

Adrian Rebmann, Alexander Beuther, Steffen Schuhmann und Peter Fettke
Universität des Saarlandes und Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

MoHoL @ Modellierung 2020

Wien, 19. Februar 2020

UNIVERSITÄT DES SAARLANDES

DFKI

Agenda

1. Motivation und Kontext
2. Herausforderungen und Lösungsansatz
3. Aufgabe
4. Limitationen und Ausblick

UNIVERSITÄT DES SAARLANDES

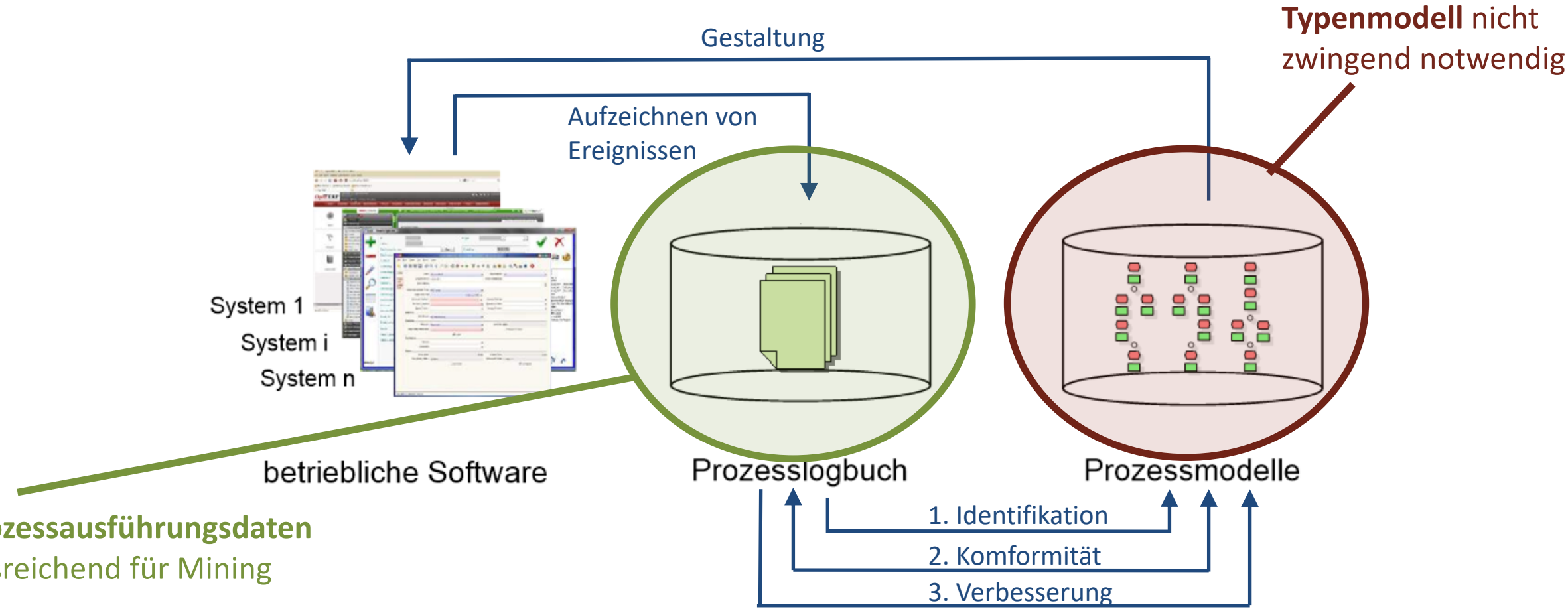
DFKI

Motivation

- Veranstaltung „Process Mining“ an der Universität des Saarlandes
- Kurs behandelt theoretische Konzepte und praktische Process-Mining-Aufgaben
- Vorlesungen als Videos
- Übungsblätter, die thematisch auf Inhalte der Vorlesung ausgerichtet sind
- Tutorium zur Diskussion von Lösungen
- Kurs deckt bereits praktisches Process Mining ab, **jedoch**
  - Praktische Aufgaben basieren immer auf bereits aufbereiteten Logs
  - Praktische Aufgaben bisher nur mit Anwendungssoftware

Ziel:
Ziel ist es, den Studierenden eine stärker data-science-orientierte Herangehensweise an das Process Mining zu vermitteln, die die eigenständige Lösung von komplexen Problemstellungen erlaubt

Process Mining im Kontext des BPM



Prozessausführungsdaten
ausreichend für Mining

Perspektive zur universitären Lehre

- Perspektive des Praktikers
- Kursorganisatoren
- Weiterentwicklung des bestehenden Kurskonzepts
- Realistische Anwendungsfälle
- Neben theoretischen Konzepten soll unbedingt die Anwendung in Fokus stehen
- Eigenständig Lösungen für Probleme finden

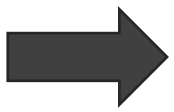
„[…] in den letzten Jahrzehnten [wird] verstärkt ein Perspektivenwechsel von einer dozenten- zu einer studierendenzentrierten Lehre hin gefordert, die eine aktive Rolle der Studierenden und praktische Anwendung theoretischer Inhalte vorsieht."*

*Michael Fellmann, Andreas Schoknecht, Meike Ullrich: Workshop zur Modellierung in der Hochschullehre. In: 2016 Modellierung 2016-Workshopband. Gesellschaft für Informatik eV, p. 45, 2016.

Herausforderungen und Lösungsansatz

- Aufgabe, die Prozess von der **Datenerfassung bis zur Entdeckung von Prozessmodellen** realistisch nachahmt
  - Schaffung einer **plausiblen** und **realistischen** Datenbasis
  - Nutzung von Daten zu Lehrezwecken **ohne Datenschutzprobleme**

  - Konzeption von Aufgabenstellungen die durch **Lösungsschritte für komplexe Problemstellungen** führen
  - Komplexität der Rahmenbedingungen reduzieren und
  - Medienbrüche vermeiden

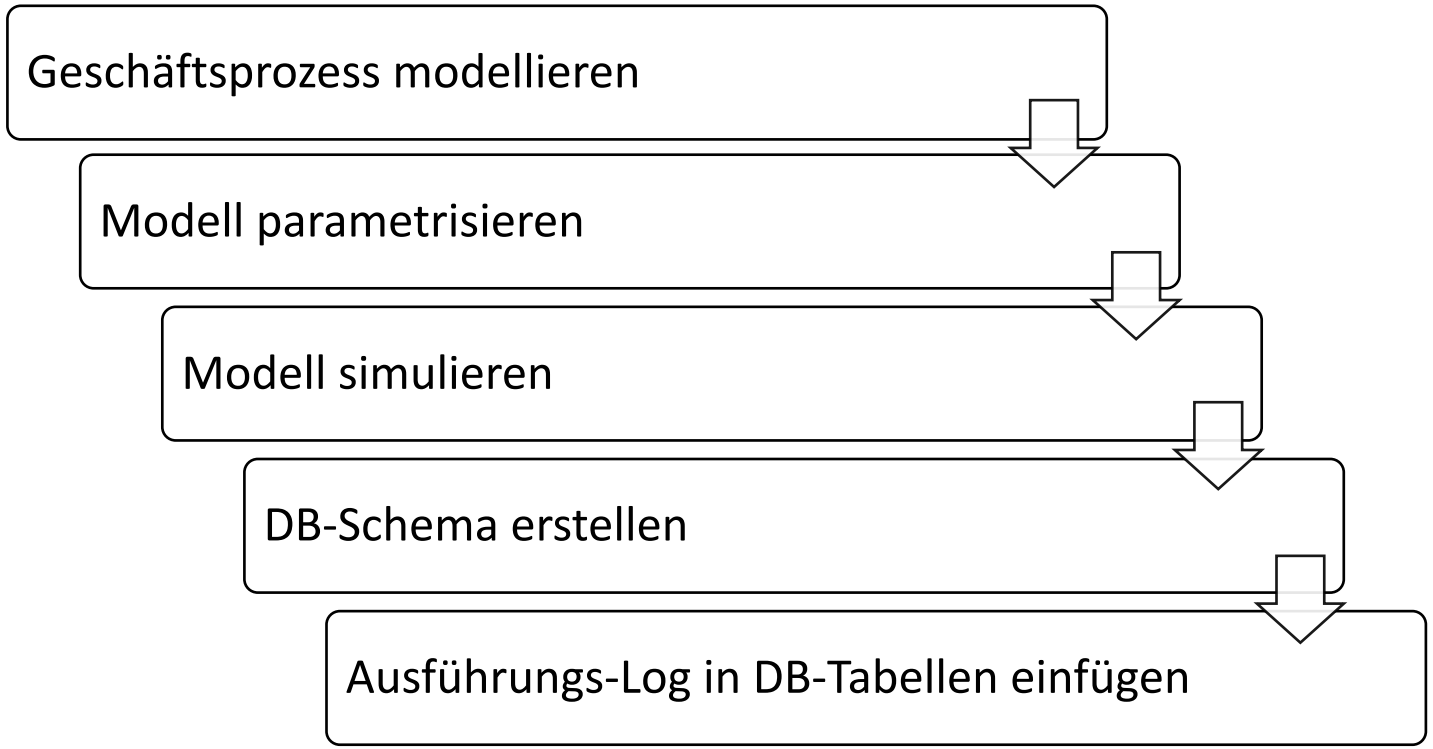


Simulation von Prozessinsanzen

Jupyter Notebook

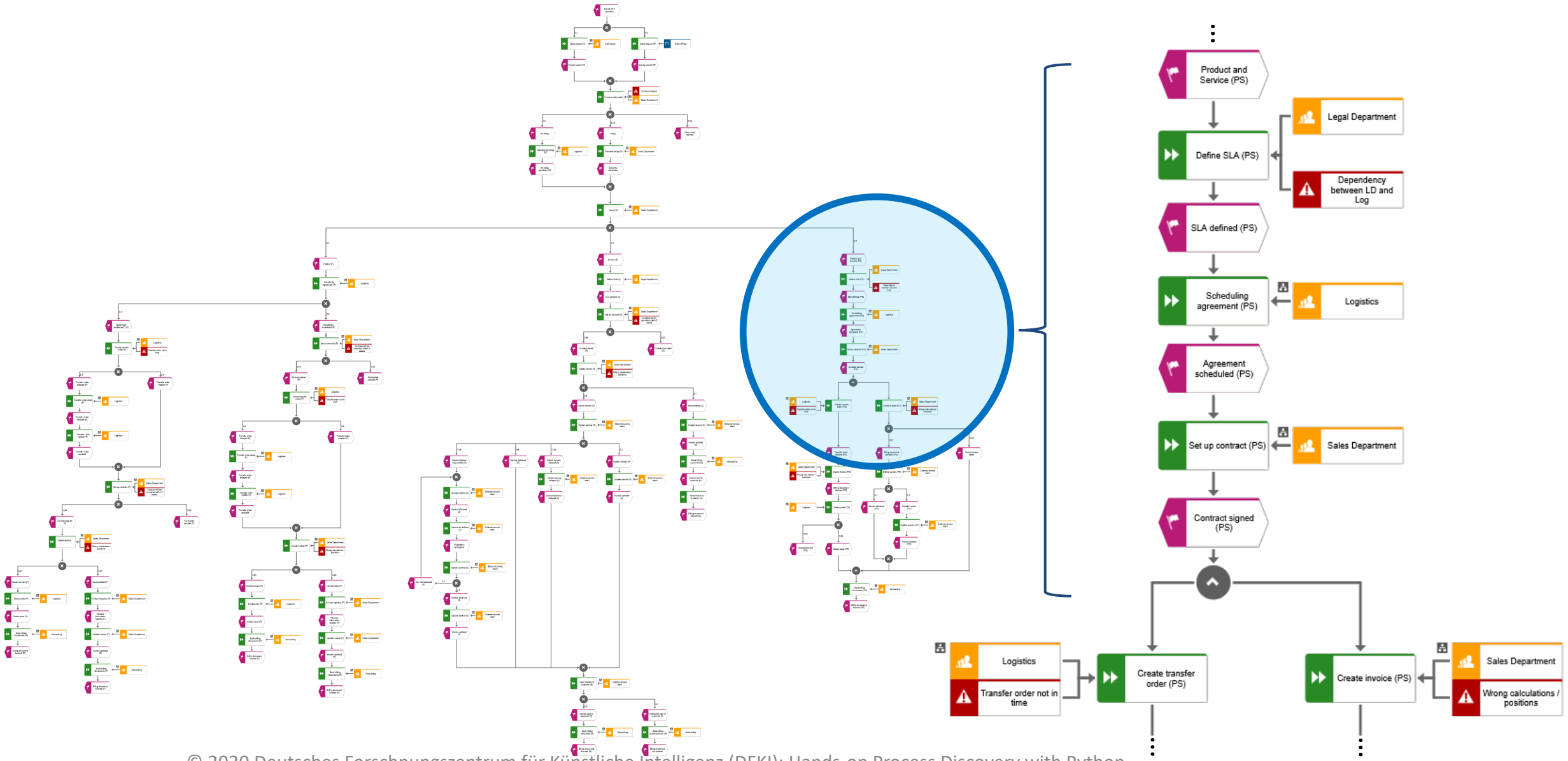
- Integrierte Programmierumgebung
- Web-basiert
- Organisiert in Zellen
  - Unterschiedliche Formatierungsmöglichkeiten
  - Verschiedene Medien können eingebettet/referenziert werden
- Können leicht geteilt werden
- Können leicht umformatiert werden (HTML/PDF)

Erstellung der Datenbasis



Die Aufgabe besteht im Wesentlichen aus der Rekonstruktion des Modells, das als Grundlage dient

OTC-Prozessmodell



Simulation und Transformation in Ereignis-Logs

ARIS Simulation

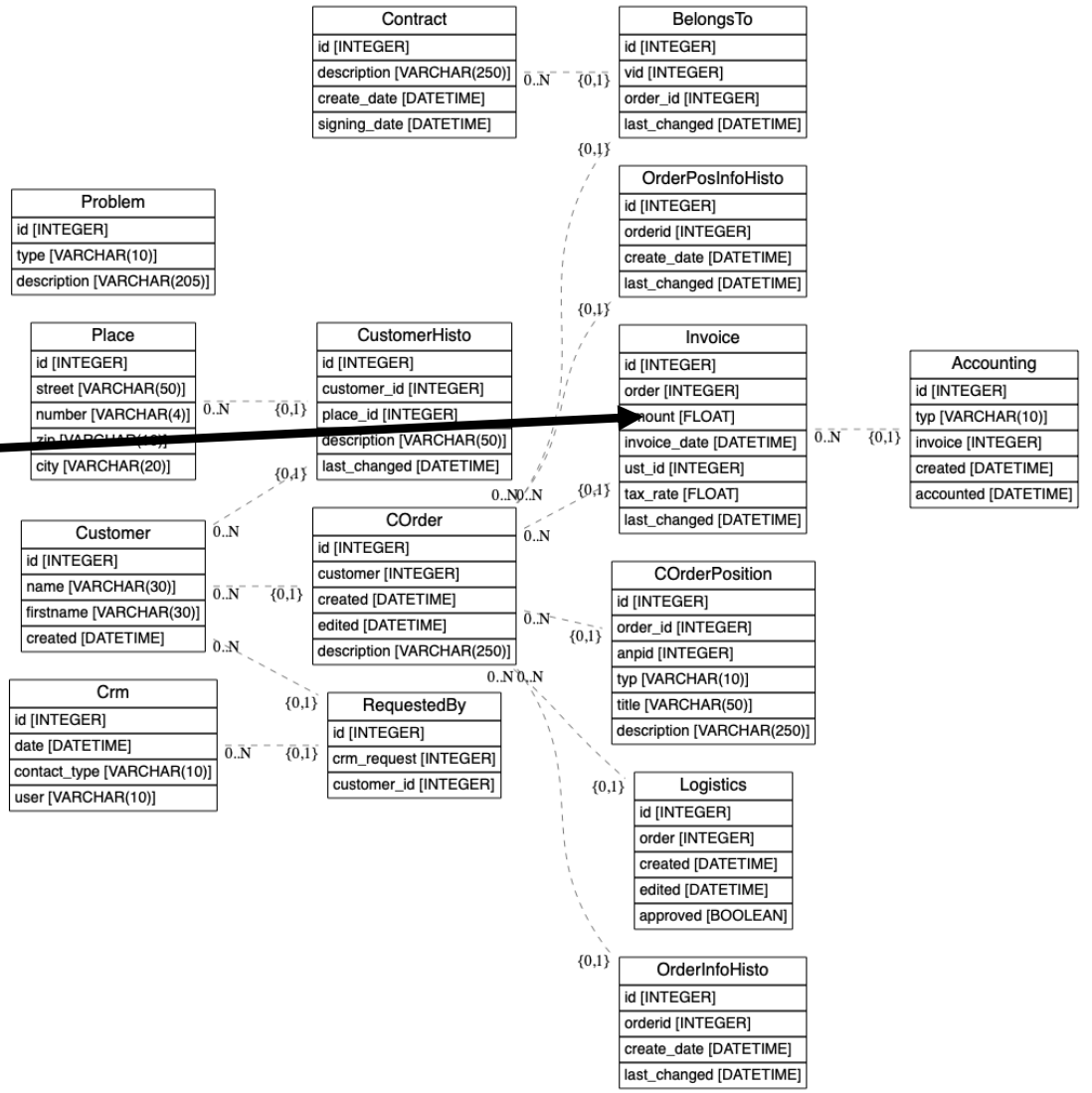
Name	Case	Start	End	Dynamic Wait Time	Duration	Learning Time	...
Serve inquiry OP	1	2019-11-26 12:57:12	2019-11-26 13:05:32	0:00:00	0:08:20		
Check customer status	1	2019-11-26 13:05:32	2019-11-26 13:05:37	0:00:00	0:00:05		
Create customer	1	2019-11-26 13:05:37	2019-11-26 13:06:07	0:00:00	0:00:30		
Serve inquiry OP	2	2019-11-26 13:11:36	2019-11-26 13:15:58	0:00:00	0:04:22		
Check customer status	2	2019-11-26 13:15:58	2019-11-26 13:16:03	0:00:00	0:00:05		
Create customer	2	2019-11-26 13:16:03	2019-11-26 13:16:33	0:00:00	0:00:30		
Serve inquiry OP	3	2019-11-26 13:26:00	2019-11-26 13:34:18	0:00:00	0:08:18		
Check customer status	3	2019-11-26 13:34:18	2019-11-26 13:34:23	0:00:00	0:00:05		
Create customer	3	2019-11-26 13:34:23	2019-11-26 13:34:53	0:00:00	0:00:30		
Serve inquiry OP	4	2019-11-26 13:40:24	2019-11-26 13:44:06	0:00:00	0:03:42		

Transformation in ein Ereignis Log-Format

Activity	Case ID	Time Stamp	...
Serve inquiry OP	1	2019-11-26 12:57:12	
Check customer status	1	2019-11-26 13:05:32	
Create customer	1	2019-11-26 13:05:37	
Serve inquiry OP	2	2019-11-26 13:11:36	
Check customer status	2	2019-11-26 13:15:58	
Create customer	2	2019-11-26 13:16:03	
Serve inquiry OP	3	2019-11-26 13:26:00	
Check customer status	3	2019-11-26 13:34:18	
Create customer	3	2019-11-26 13:34:23	
Serve inquiry OP	4	2019-11-26 13:40:24	

Transformation der Ereignis-Logs in Datenbanktabellen

```python
def _gen_invoice(log, context):
    """This method creates a new invoice object"""
    data = {
        'order': context['generated']['COrder'][0].id,
        'amount': uniform(-10.0, 250.00),
        'invoice_date': log.Aktivierungszeitpunkt,
        'ust_id': f"DE{randint(100000000, 999999999)}",
        'tax_rate': 19.00,
        'last_changed': log.Aktivierungszeitpunkt
    }
    invoice = model.Invoice(**data)
    return invoice
```



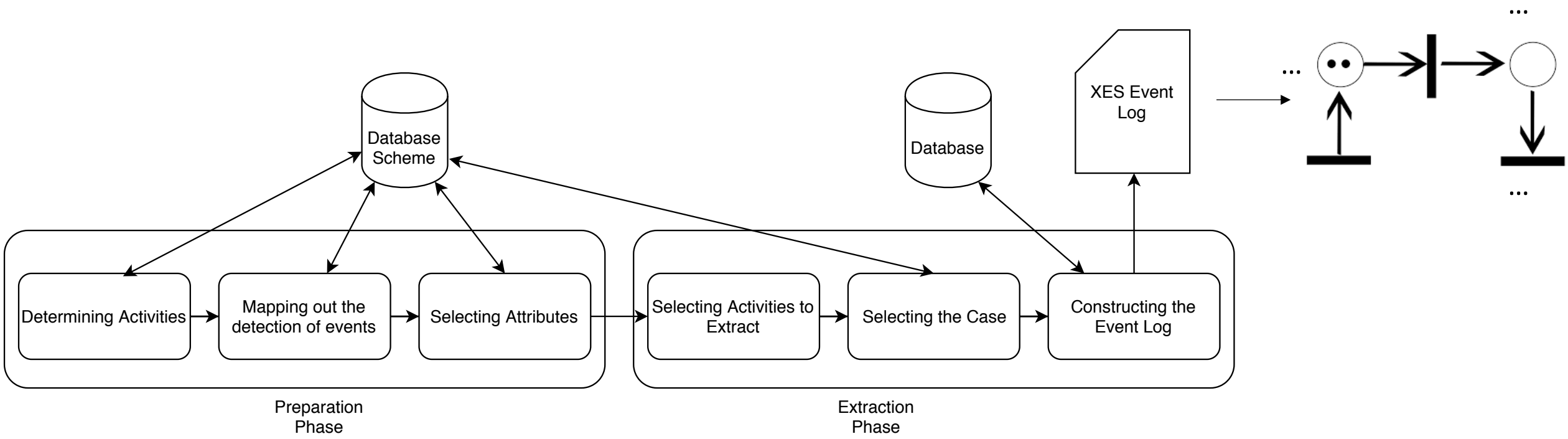
Aufgabe

“Erstellen Sie eine Modellrepräsentation aus den in der gegebenen Datenbank verfügbaren Prozessdaten”

Schritte, um dieses Ziel zu erreichen:

- **Import der erforderlichen Bibliotheken:** Der erste Schritt ist der Import der Bibliotheken, die in der Übung verwendet werden sollen.
- **Daten-/Prozessverständnis entwickeln:** Zunächst sollen die Studierenden einen Überblick über die Datenstrukturen und die für das Process Mining notwendigen Attribute erarbeiten.
- **Geführte Erstellung von Teil-Ereignis-Logs:** Teil-Ereignis-Logs stellen zunächst nur einen kleinen Teil des Gesamtprozesses dar**.**
- **Inkrementelle Entwicklung des Ereignis-Logs:** Erstellen eines Gesamt-Ereignis-Logs
- **Erstellen eines Standard-Ereignis-Logs:** Transformation des Ereignisprotokolls in das XES-Format.
- **Process Discovery**: Anwenden verschiedener Process-Discovery-Ansätze

Aufgabe



In Anlehnung an Piessens, D. (2011)

Building the event log (CRM-part)

The next step would be to create a dataframe representing the event log. For this, the tables and fields of the database, which contain relevant event data need to be merged. First, as an example, we want to focus on the CRM-Process: We import and merge the tables Crm, requestedBy and Customer using Pandas' SQL Interface. All values of table Crm should be present in the imported data set. Pandas features many importing functionalities. Data fields are automatically assigned suitable data types. However, special functions can be specified before data import, such as parsing of dates, specifying a custom header and defining a separator. Also different import functions are available to create a DataFrame:

- `pd.read_csv()` CSV
- `pd.read_excel()` Excel
- `pd.read_sql_query` SQL

We use the read_sql_query function

```
query_all_customer_requests_with_account_creation = """select * from Crm crm LEFT join requestedBy rb on crm.id = rb.cr
header = ['id', 'crm_request_time', 'contact_type', 'crm_user', 'join_table_id', 'merge_id_customer', 'merge_id_crm', '
dataset = pd.read_sql_query(query_all_customer_requests_with_account_creation, conn, parse_dates=['date', 'created'])
# setting a new header
dataset.columns = header
dataset.head()
```

	id	crm_request_time	contact_type	crm_user	join_table_id	merge_id_customer	merge_id_crm	customer_id	name	firstname	customer_create_time
0	1	2019-01-01 09:42:48	call center	user_107	1.0	1.0	1.0	1.0	Hans	Ing.	2019-01-01 10:12:36
1	2	2019-01-01 09:57:12	call center	user_102	NaN	NaN	NaN	NaN	None	None	NaT
2	3	2019-01-01 10:11:36	online	None	NaN	NaN	NaN	NaN	None	None	NaT
3	4	2019-01-01 10:26:00	online	None	2.0	4.0	2.0	2.0	Stahr	Heinz-Werner	2019-01-01 10:28:41
4	5	2019-01-01 10:40:24	online	None	3.0	5.0	3.0	3.0	Karz	Sinaida	2019-01-01 10:46:48

```
# in order to avoid naming issues later, we can user the attribute names defined by the XES standard for case id, activ
def prepare_customer_request(x):
    log_line = {'case:concept:name': x.id}
    log_line['concept:name'] = 'online request' if x.contact_type == 'online' else 'call center request'
    log_line['time:timestamp'] = x.crm_request_time
    log_line['org:resource'] = x.crm_user if x.crm_user else 'online platform'
    return log_line
```

Converting and Using the Log

As already seen in the lecture and other tasks XES is

- the de facto standard lof format for Process Mining
- extends the XML format
- is supported by commercial as well as academic Process Mining tools This is why we use the created DataFrame to create a XES compliant log in this step.

Task:

- take the data frame created before and create a XES compliant log using the pm4py function

```
from pm4py.objects.log.exporter.xes import factory as xes_exporter
from pm4py.objects.conversion.log import factory as conversion_factory
# convert the dataframe here
```

Discovery

For this step, we import the XES-Log we created in the previous step. After this, we use the heuristics miner [1], the discover a model from our log. Heuristic mining in contrast to the Alpha-Algorithm [2] uses causal nets. It can handle concurrency and considers frequencies of events, so traces that do not happen frequently are abstracted from, in other words, noise is filtered out. Using the frequency of two activities following one another, dependency measures are computed and a dependency graph aka. causal net is created using certain thresholds for the dependency measure and frequency of activities directly following one another. The input is an event log that requires at least case identifiers and event labels. A causal net representing the process is the output of the algorithm.

Task:

Please use pm4py to apply

- the heuristcs miner algorithm. Please explain the following measures. Also use different configurations.
  - dependency threshold of the Heuristics Miner
  - AND measure threshold of the Heuristics Miner
- the alpha algorithm.

```
from pm4py.algo.discovery.heuristics import factory as heuristics_miner
from pm4py.objects.log.importer.xes import factory as xes_importer
```

Live Demo

Limitationen und Ausblick

- Fehlende langfristige Evaluation des Konzepts
- Feedback:
  - **kein Wechsel** zwischen verschiedenen **Medien oder Dokumenten** notwendig
  - **vorherige Einführung** in den Anwendungsfall und Anweisungen zum Arbeitsablauf erforderlich
  - Aufwand für **Tool-Setup** muss berücksichtigt werden
- Aufwand diese spezielle Aufgabe zu erstellen war signifikant
- Der **allgemeine Notebook-Ansatz** ist davon jedoch **nicht betroffen**, hier ist der Aufwand für die Aufgabenerstellung geringer, da **alle Aufgaben und zugehörigen Materialien an einem "Ort"** erstellt/referenziert werden können!
- **Bereits andere Aufgaben** in Notebooks realisiert

- **Plan: Vorlesung aufteilen in "Grundlagen des Process Mining" und ”Advanced Process Analytics".**

Vielen Dank für Ihre Aufmerksamkeit!

Kontakt
Adrian Rebmann, Alexander Beuther, Steffen Schuhmann, und Peter Fettke
Universität des Saarlandes und Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) und Universität des Saarlandes
Campus D 3 2 D-66123 Saarbrücken,
vorname.nachname@dfki.de
http://www.dfki.de