Vorlesung

Methodische Grundlagen des Software-Engineering

im Sommersemester 2014

Prof. Dr. Jan Jürjens

TU Dortmund, Fakultät Informatik, Lehrstuhl XIV

Teil 2.6: Mining: Zusätzliche Perspektiven

v. 10.06.2014

2.6 Mining: Zusätzliche Perspektiven

**[mit freundlicher Genehmigung basierend
auf einem englischen Foliensatz von
Prof. Dr. Wil van der Aalst (TU Eindhoven)]**

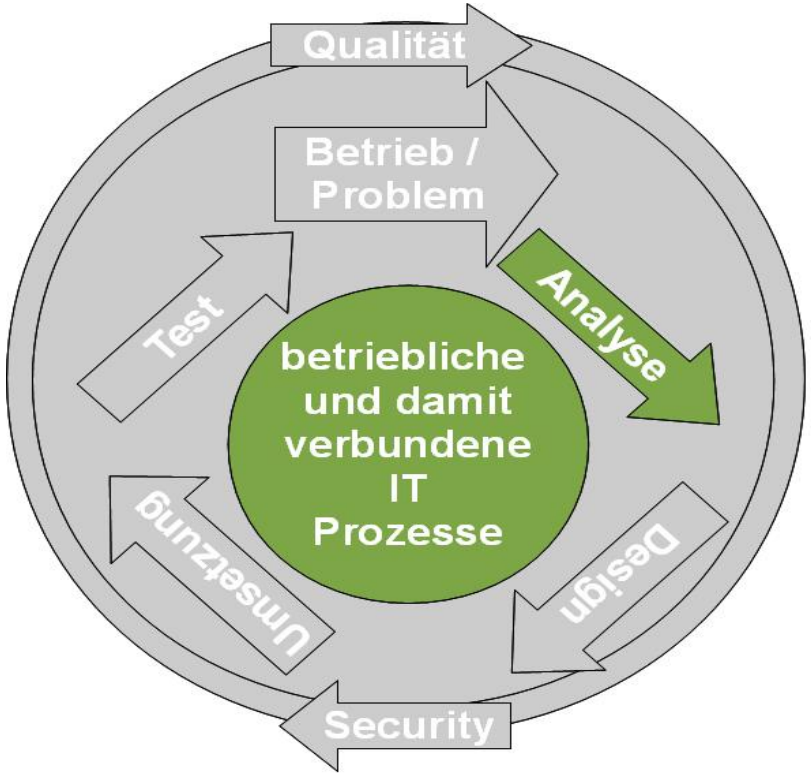
Literatur:

[vdA11] Wil van der Aalst: **Process Mining: Discovery, Conformance and Enhancement of Business Processes**, Springer-Verlag. 2011.
Unibibliothek (6 Exemplare): http://www.ub.tu-dortmund.de/katalog/titel/1332248
(Bei Engpässen kann eine **Kopiervorlage** der relevanten Ausschnitte zur Verfügung gestellt werden.)

- **Kapitel 8**

Einordnung
Mining: Zusätzliche Perspektiven

- Geschäftsprozessmodellierung
- Process-Mining
  - Einführung: Process-Mining
  - Petrinetze
  - Data-Mining
  - Datenbeschaffung
  - Prozessextraktion
  - Konformanzanalyse
  - Mining: Zusätzliche Perspektiven
  - Betriebsunterstützung
  - Werkzeugunterstützung
  - Analysiere „Lasagne Prozesse“
  - Analysiere „Spaghetti Prozesse“
  - Kartographie und Navigation
  - Epilog
- Modellbasierte Entwicklung sicherer Software

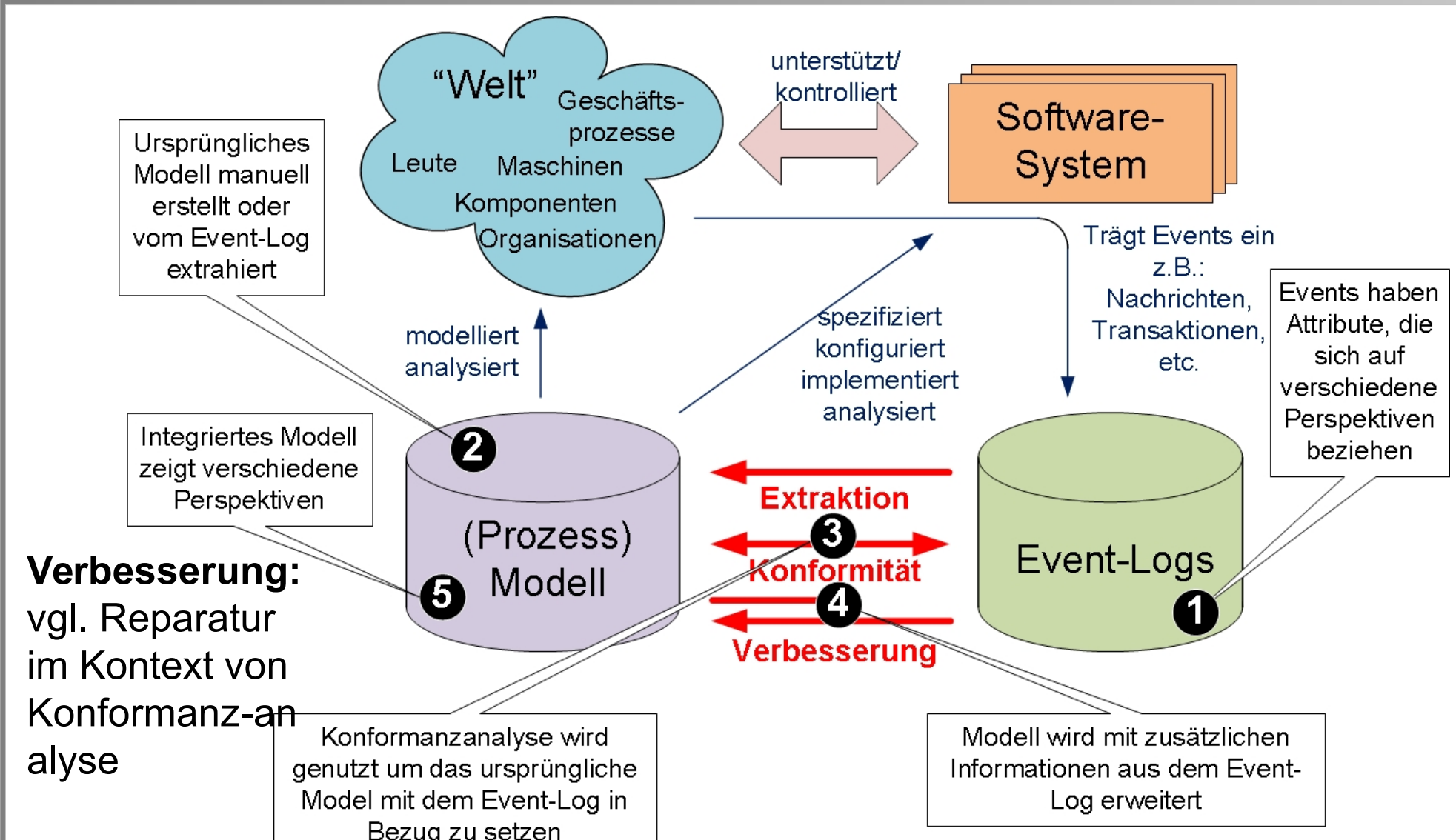


- **Letzter Abschnitt:** Konformanzanalyse.
- **Dieser Abschnitt:** „Mining: Zusätzliche Perspektiven“:
  - **Organizational Mining:** Zusammenhang zwischen Ressource und Aktivität.
  - Zeitanalyse durch Replay.
  - **Decision Mining** mit Hilfe von Replay.

Überblick
Mining: Zusätzliche Perspektiven

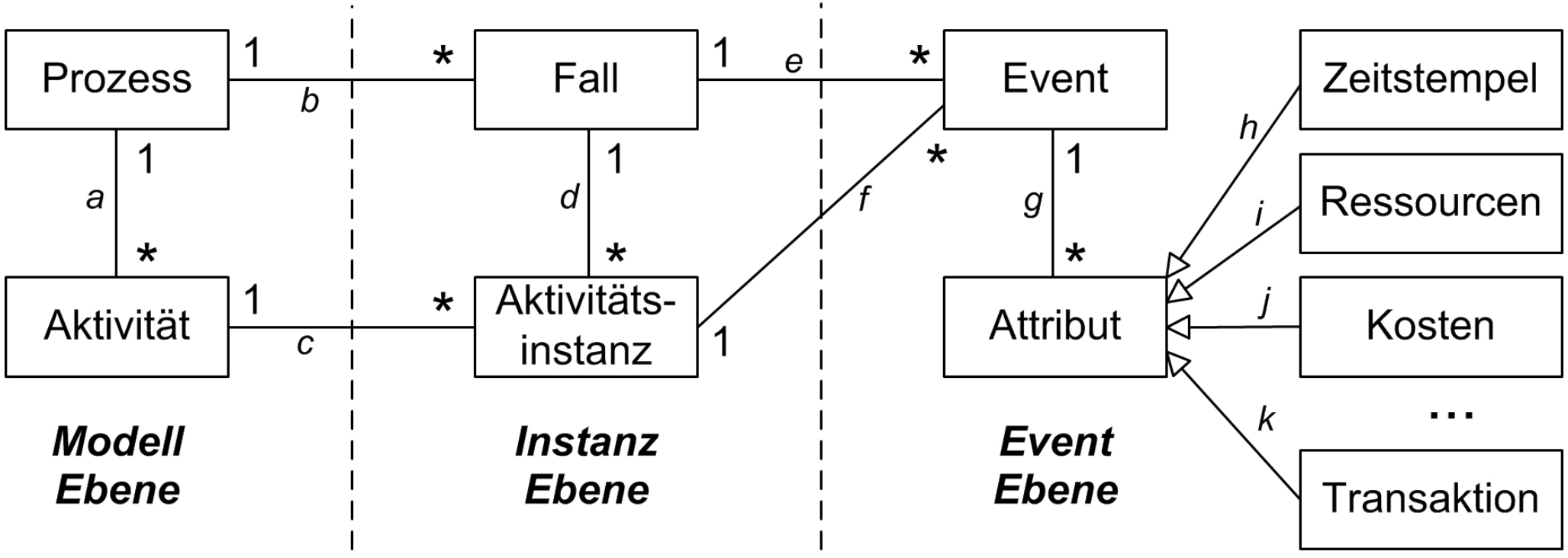
- **Attribute in Event-Logs**
- Organizational Mining
- Verhalten von Ressourcen Analysieren
- Decision-Mining

Mining: Zusätzliche Perspektiven



Verbesserung:
vgl. Reparatur im Kontext von Konformanz-analyse

Startpunkt: Event-Log verknüpft mit Modell



- Sehr wichtig!
- **Modell extrahiert** oder **manuell** erstellt.
- Während des **Replays** verknüpft.
- Startpunkt für andere Typen des Process-Minings!

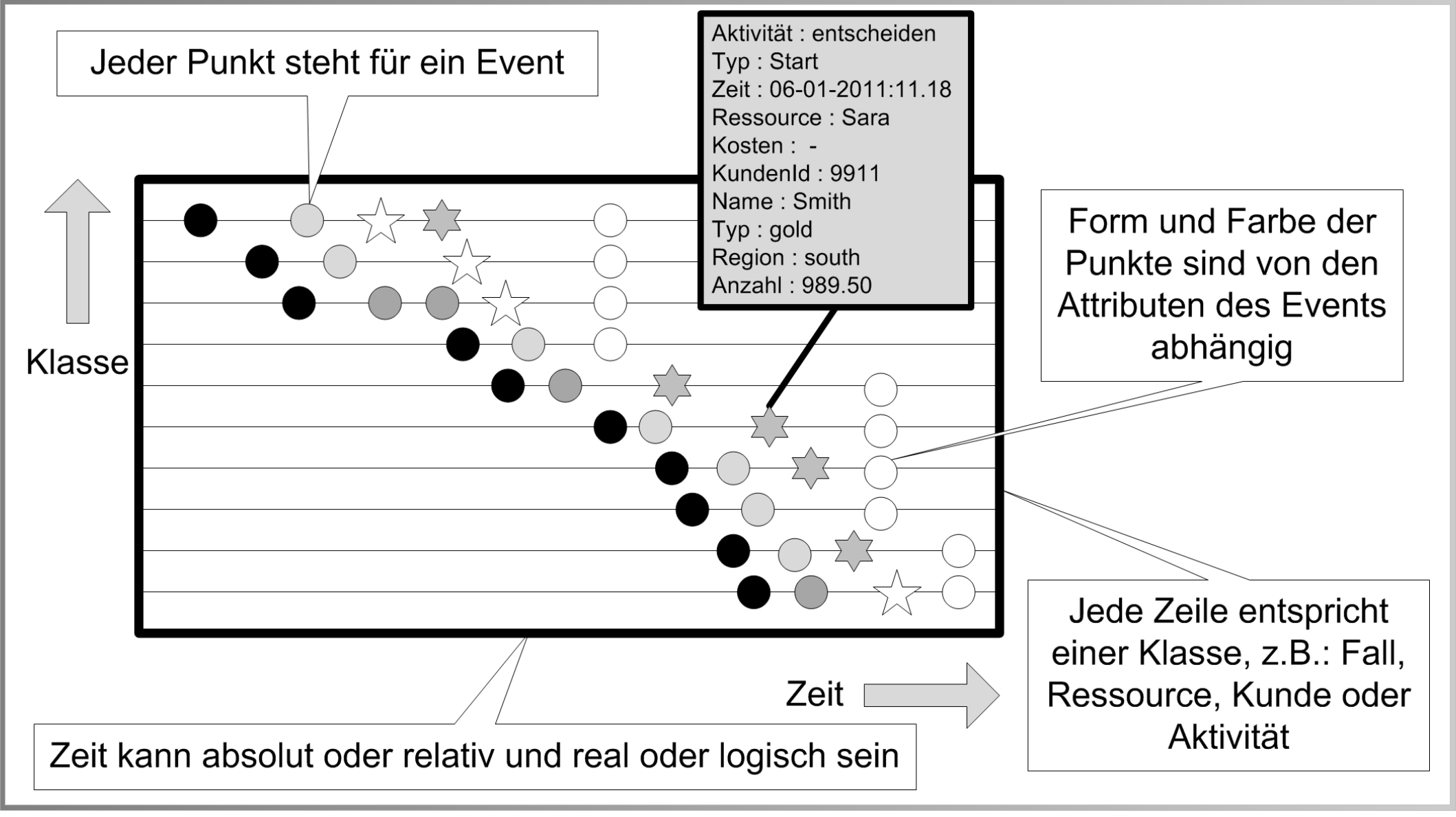
Attribute in Event-Logs

case id	event id	properties				
		time	activity	trans	resource	cost
	35654423	30-12-2010:11.02	register request	start	Pete	
	35654424	30-12-2010:11.08	register request	complete	Pete	50
	35654425	31-12-2010:10.06	examine thoroughly	start	Sue	
	35654427	31-12-2010:10.08	check ticket	start	Mike	
1	35654428	31-12-2010:10.12	examine thoroughly	complete	Sue	400
	35654429	31-12-2010:10.20	check ticket	complete	Mike	100
	35654430	06-01-2011:11.18	decide	start	Sara	
	35654431	06-01-2011:11.22	decide	complete	Sara	200
	35654432	07-01-2011:14.24	reject request	start	Pete	
	35654433	07-01-2011:14.32	reject request	complete	Pete	200
	35654483	30-12-2010:11.32	register request	start	Mike	
	35654484	30-12-2010:11.40	register request	complete	Mike	50
	35654485	30-12-2010:12.12	check ticket	start	Mike	
	35654486	30-12-2010:12.24	check ticket	complete	Mike	100
2	35654487	30-12-2010:14.16	examine casually	start	Pete	
	35654488	30-12-2010:14.22	examine casually	complete	Pete	400
	35654489	05-01-2011:11.22	decide	start	Sara	
	35654490	05-01-2011:11.29	decide	complete	Sara	200
	35654491	08-01-2011:12.05	pay compensation	start	Ellen	
	35654492	08-01-2011:12.15	pay compensation	complete	Ellen	200
...	...	...	...	...	...	...

Fälle können auch Attribute haben

case id	custid	name	type	region	amount
1	9911	Smith	gold	south	989.50
2	9915	Jones	silver	west	546.00
3	9912	Anderson	silver	north	763.20
4	9904	Thompson	silver	west	911.70
5	9911	Smith	gold	south	812.10
6	9944	Baker	silver	east	788.00
7	9944	Baker	silver	east	792.80
8	9911	Smith	gold	south	544.70
...	...	...	...	...	...

Helikopter-Sicht: Punkte-Diagramme



Überblick
Mining: Zusätzliche Perspektiven

- Attribute in Event-Logs
- **Organizational Mining**
- Verhalten von Ressourcen Analysieren
- Decision-Mining

Organizational-Mining

a = Registrierung anfragen, b = gründlich überprüfen, c = normal überprüfen,
d = Ticket überprüfen, e = entscheiden, f = Anfrage neu einleiten,
g = Entschädigung bezahlen und h = Anfrage ablehnen

case id	trace
1	$\langle a^{Pete}, b^{Sue}, d^{Mike}, e^{Sara}, h^{Pete} \rangle$
2	$\langle a^{Mike}, d^{Mike}, c^{Pete}, e^{Sara}, g^{Ellen} \rangle$
3	$\langle a^{Pete}, c^{Mike}, d^{Ellen}, e^{Sara}, f^{Sara}, b^{Sean}, d^{Pete}, e^{Sara}, g^{Ellen} \rangle$
4	$\langle a^{Pete}, d^{Mike}, b^{Sean}, e^{Sara}, h^{Ellen} \rangle$
5	$\langle a^{Ellen}, c^{Mike}, d^{Pete}, e^{Sara}, f^{Sara}, d^{Ellen}, c^{Mike}, e^{Sara}, f^{Sara}, b^{Sue}, d^{Pete}, e^{Sara}, h^{Mike} \rangle$
6	$\langle a^{Mike}, c^{Ellen}, d^{Mike}, e^{Sara}, g^{Mike} \rangle$
...	...

Ressourcen-Aktivitäts-Matrix

Durchschnittliche **Anzahl**, wie oft eine Ressource eine **Aktivität pro Fall** ausführt.

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*
Pete	0.3	0	0.345	0.69	0	0	0.135	0.165
Mike	0.5	0	0.575	1.15	0	0	0.225	0.275
Ellen	0.2	0	0.23	0.46	0	0	0.09	0.11
Sue	0	0.46	0	0	0	0	0	0
Sean	0	0.69	0	0	0	0	0	0
Sara	0	0	0	0	2.3	1.3	0	0

- Aktivität a für jeden Fall genau einmal ausgeführt (Summe 1. Spalte).
- Pete, Mike und Ellen führen als einzigen diese Aktivität aus.
- a zu 30% von Pete, zu 50% von Mike und zu 20% von Ellen ausgeführt.
- e und f immer von Sara ausgeführt.
- e im Schnitt 2.3 mal pro Fall ausgeführt.
- etc.

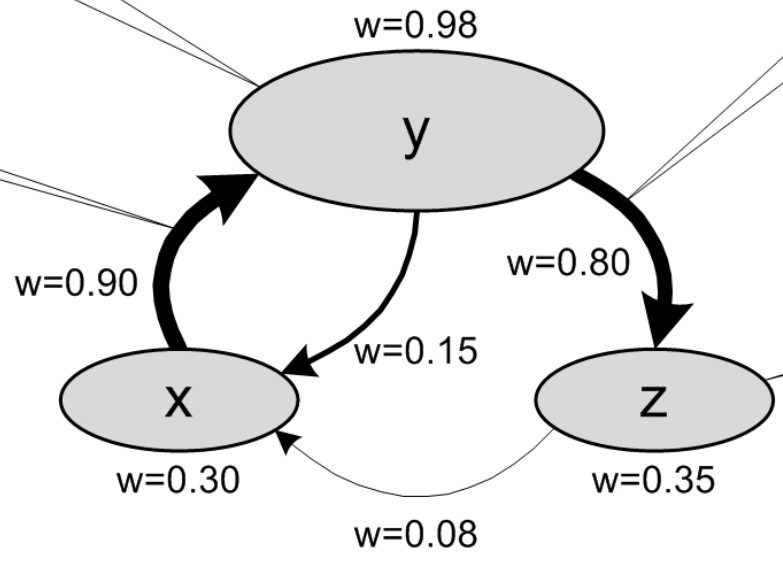
Analyse Sozialer Netzwerke

Unternehmenseinheit
(Ressource, Person, Rolle, Department, etc.)

Dicke der Pfeile abhängig vom Gewicht (w) der Beziehung

Beziehung

Größe der Ovale abhängig vom Gewicht der Einheit

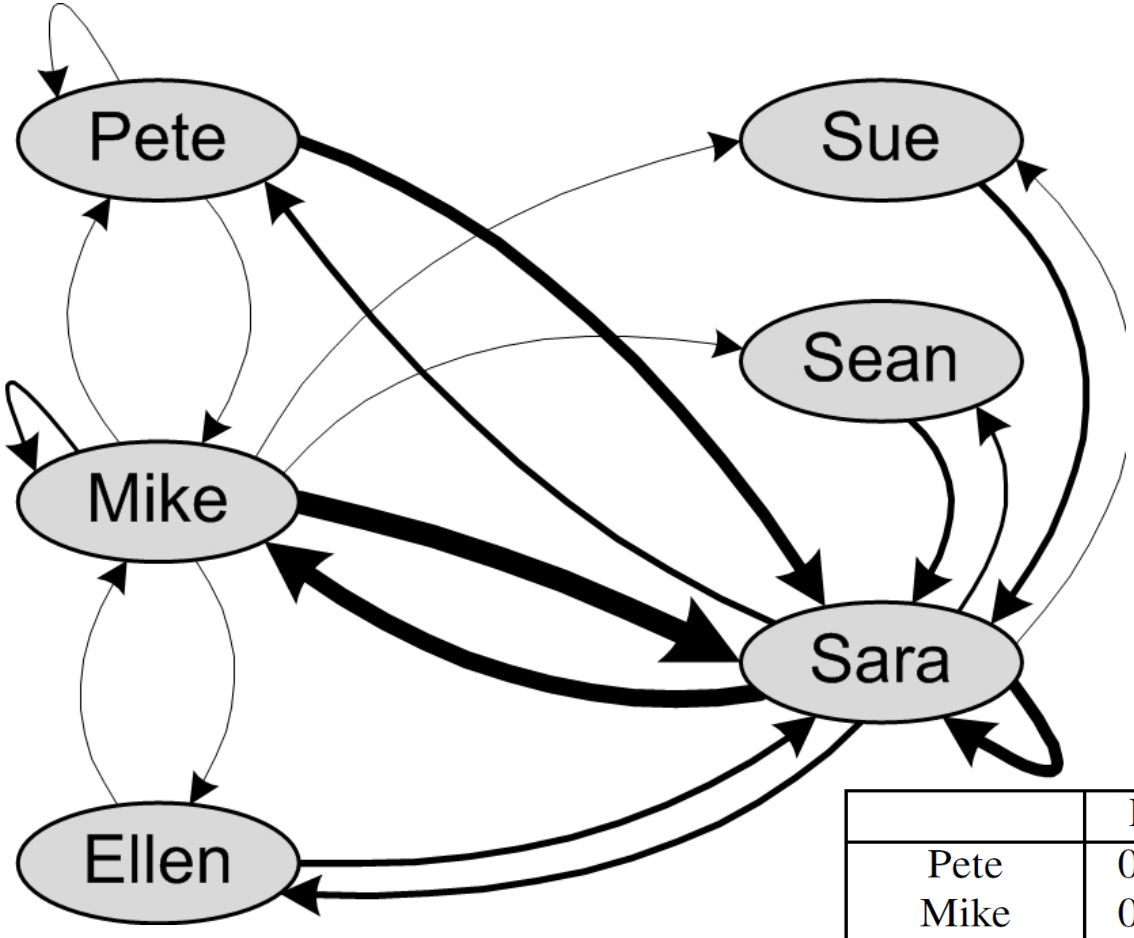


Matrix der Aufgabenübergabe

	Pete	Mike	Ellen	Sue	Sean	Sara
Pete	0.135	0.225	0.09	0.06	0.09	1.035
Mike	0.225	0.375	0.15	0.1	0.15	1.725
Ellen	0.09	0.15	0.06	0.04	0.06	0.69
Sue	0	0	0	0	0	0.46
Sean	0	0	0	0	0	0.69
Sara	0.885	1.475	0.59	0.26	0.39	1.3

- Zähle Übergabe der **Aufgaben** von einer Ressource zu einer anderen (im **Durchschnitt pro Fall**).
- **Kausale Abhängigkeiten** im Prozessmodell genutzt: Übergaben im Event-Log zu zählen.

Soziales Netzwerk basierend auf Aufgabenübergabe (Schwellwert: 0.1)



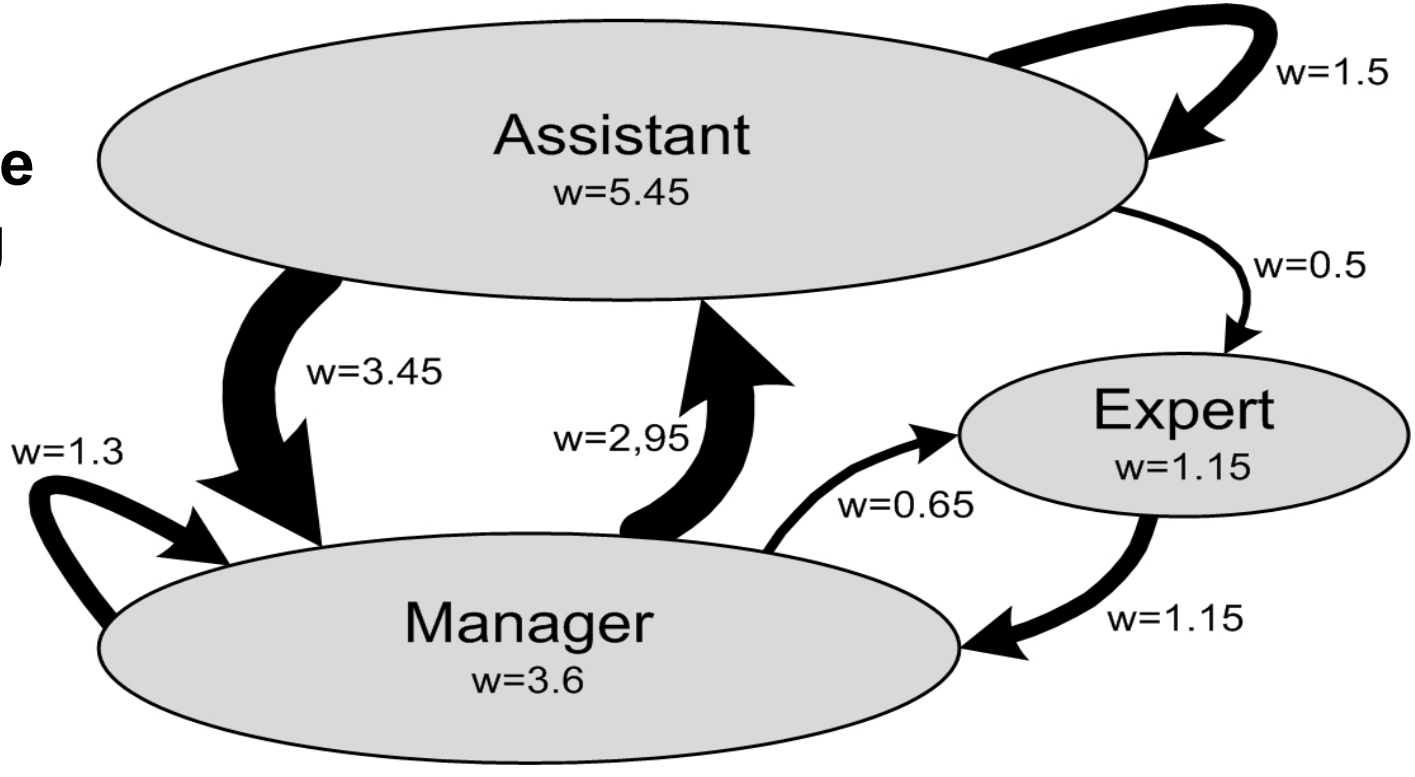
Nur **Dicke** der **Pfeile** abhängig von **Häufigkeiten**.

	Pete	Mike	Ellen	Sue	Sean	Sara
Pete	0.135	0.225	0.09	0.06	0.09	1.035
Mike	0.225	0.375	0.15	0.1	0.15	1.725
Ellen	0.09	0.15	0.06	0.04	0.06	0.69
Sue	0	0	0	0	0	0.46
Sean	0	0	0	0	0	0.69
Sara	0.885	1.475	0.59	0.26	0.39	1.3

Aufgabenübergabe auf Rollen-Ebene

	Assistant	Expert	Manager
Assistant	1.5	0.5	3.45
Expert	0	0	1.15
Manager	2.95	0.65	1.3

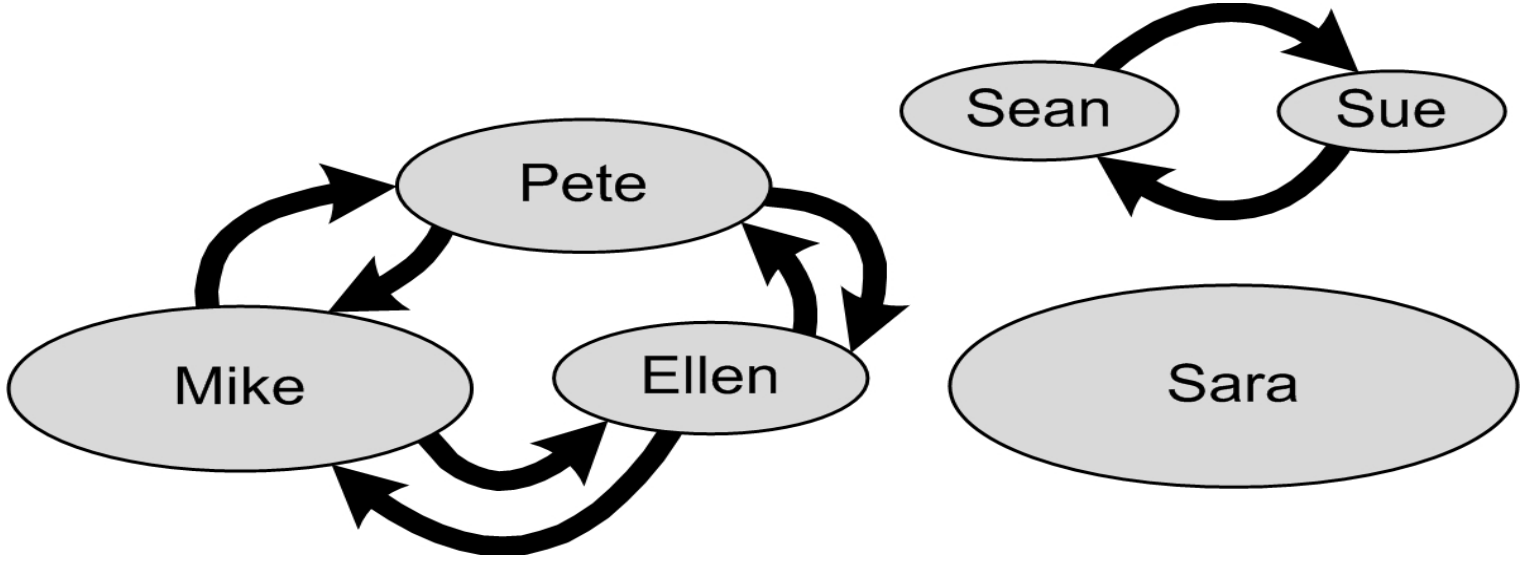
Hier: Zusätzlich **Größe der Knoten** abhängig von **Häufigkeiten**.



Profile

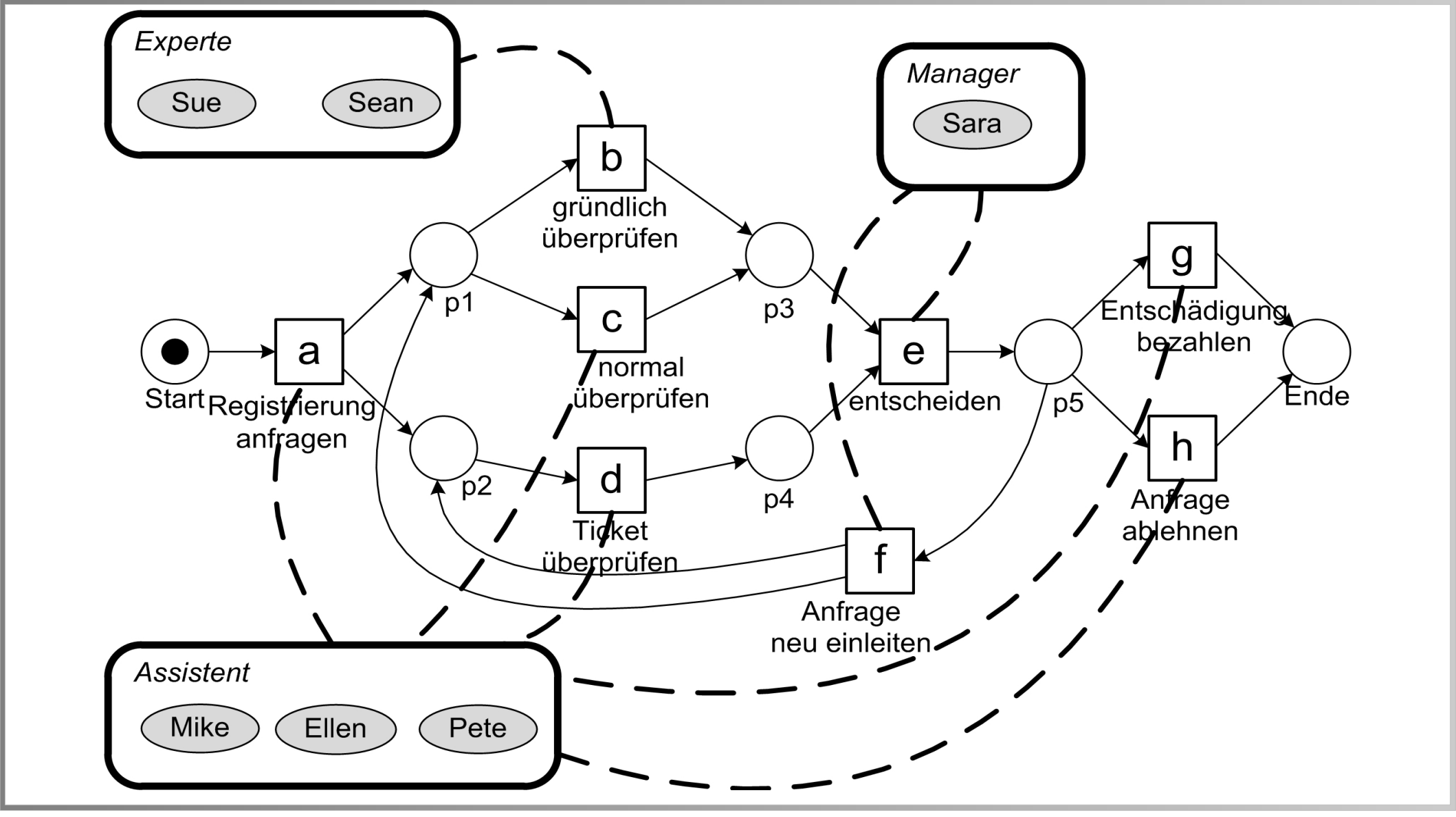
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*
Pete	0.3	0	0.345	0.69	0	0	0.135	0.165
Mike	0.5	0	0.575	1.15	0	0	0.225	0.275
Ellen	0.2	0	0.23	0.46	0	0	0.09	0.11
Sue	0	0.46	0	0	0	0	0	0
Sean	0	0.69	0	0	0	0	0	0
Sara	0	0	0	0	2.3	1.3	0	0

Soziale Netzwerke basierend auf Ähnlichkeit von Profilen

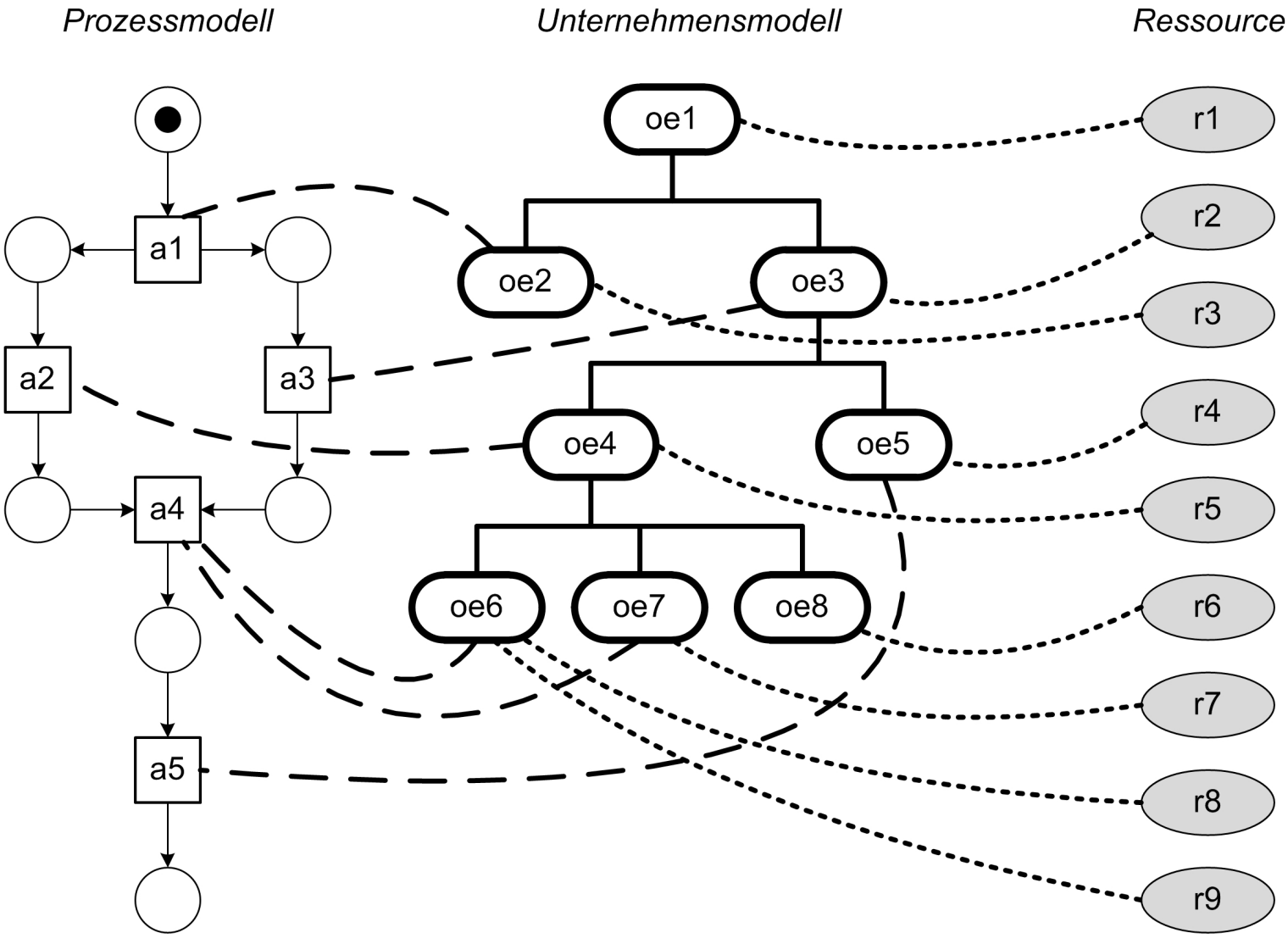


- **Ressourcen**, die ähnliche Aktivitäten ausführen, stehen in Beziehung.
- Sara führt als einzige Ressource e und f aus.
  - Nicht mit anderen Ressourcen verbunden.
- **Verknüpfungen zur eigenen Ressource unterdrückt.**
  - Enthalten keine Informationen (self-similarity).

Unternehmensstrukturen extrahieren



Prozess- vs. Unternehmensstruktur

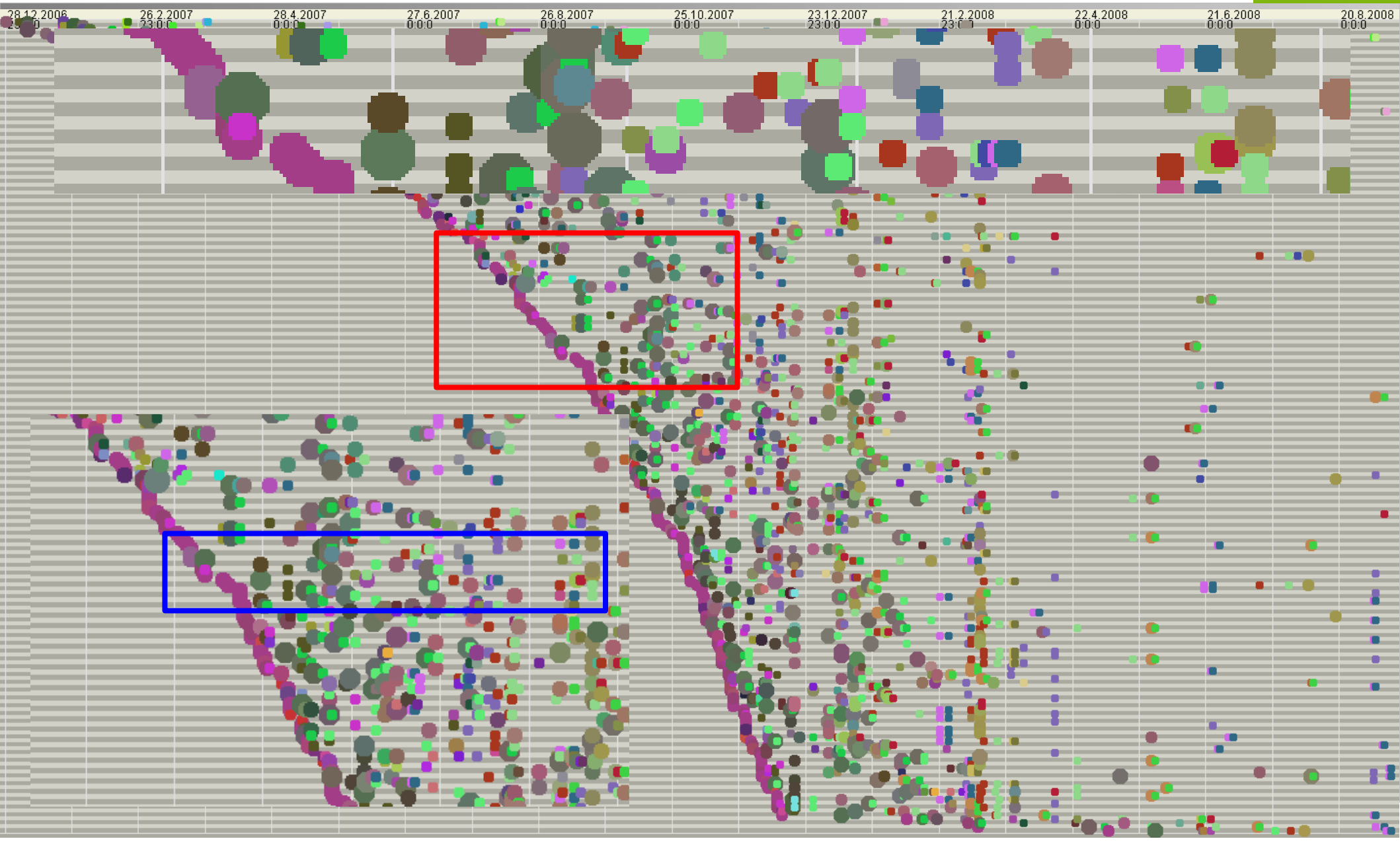


Überblick
Mining: Zusätzliche Perspektiven

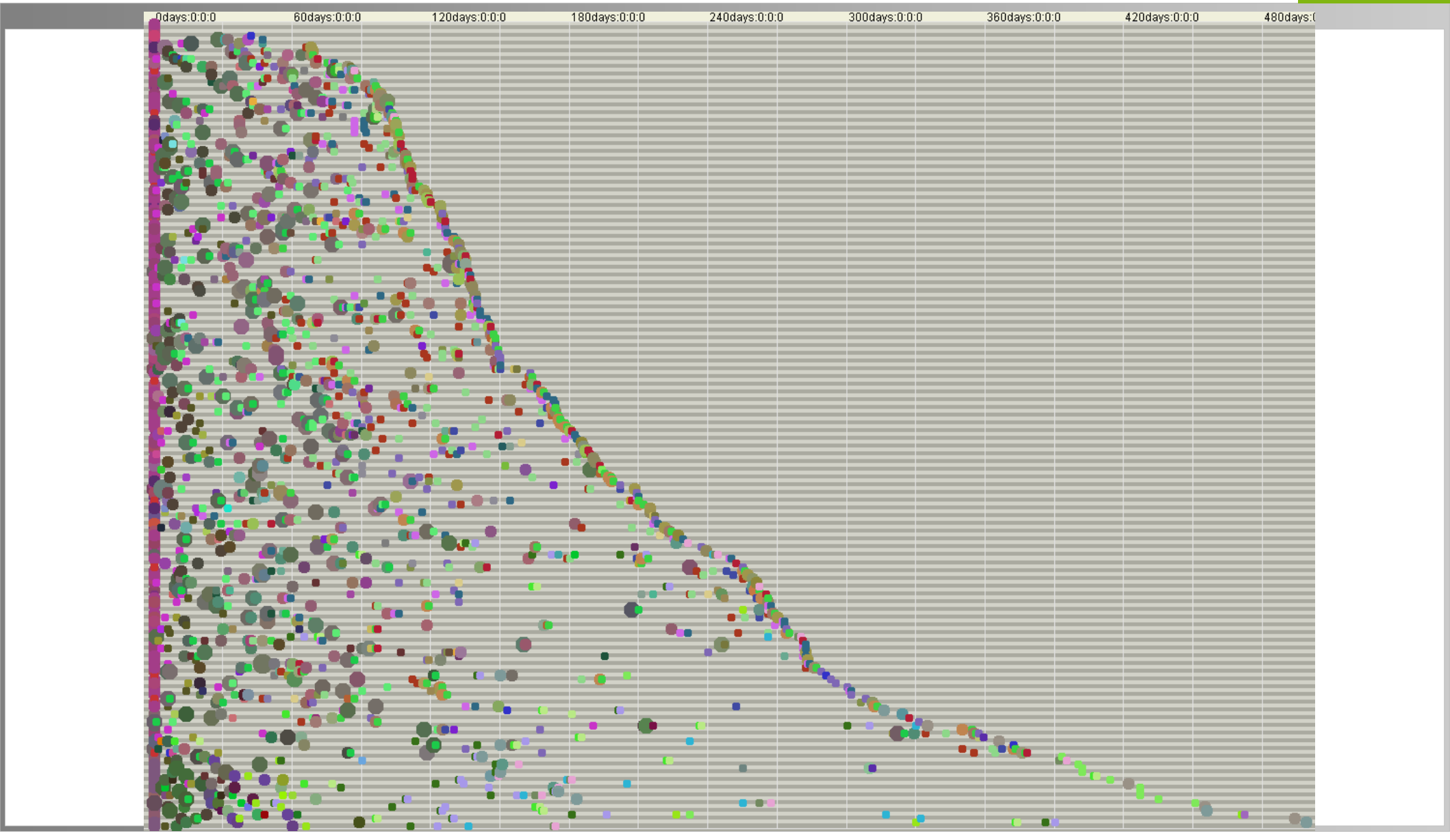
- Attribute in Event-Logs
- Organizational Mining
- **Zeit-Analyse**
- Decision-Mining

Punkte-Diagramm: Prozess einer Wohnungsvermittlung mit absoluter Zeit

Detailausschnitt

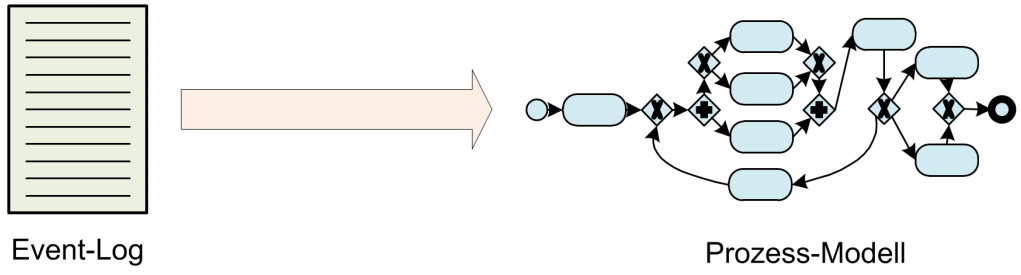


Gleicher Log, relative Zeit

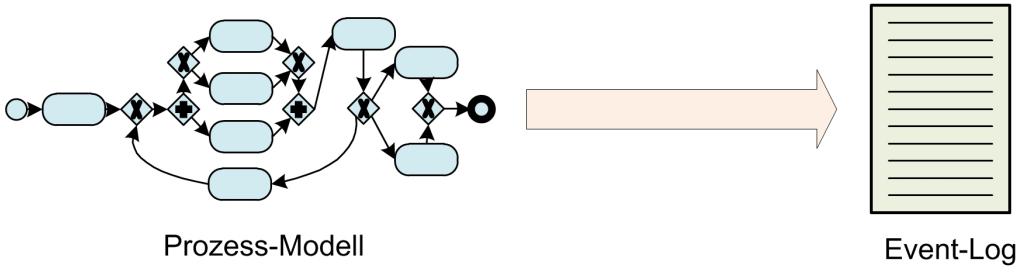


Timed Replay

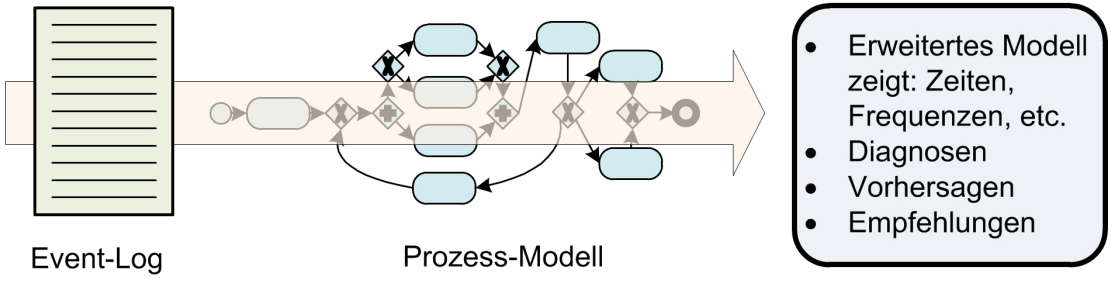
Play-In



Play-Out



Replay



(Timed) Replay: **Timing**-Information mit **Modellelementen** verknüpfen. Ziele:

- **Visualisierung**
- **Analyse**

der Zeit-Informationen.

Timed Replay

case id	trace
1	$\langle a^{12}_{start}, a^{19}_{complete}, b^{25}_{start}, d^{26}_{start}, b^{32}_{complete}, d^{33}_{complete}, e^{35}_{start}, e^{40}_{complete}, h^{50}_{start}, h^{54}_{complete} \rangle$
2	$\langle a^{17}_{start}, a^{23}_{complete}, d^{28}_{start}, c^{30}_{start}, d^{32}_{complete}, c^{38}_{complete}, e^{50}_{start}, e^{59}_{complete}, g^{70}_{start}, g^{73}_{complete} \rangle$
3	$\langle a^{25}_{start}, a^{30}_{complete}, c^{32}_{start}, c^{35}_{complete}, d^{35}_{start}, d^{40}_{complete}, e^{45}_{start}, e^{50}_{complete}, f^{50}_{start}, f^{55}_{complete}, b^{60}_{start}, d^{62}_{start}, b^{65}_{complete}, d^{67}_{complete}, e^{80}_{start}, e^{87}_{complete}, g^{90}_{start}, g^{98}_{complete} \rangle$
...	...

Anfang Aktivität b

Ende Aktivität b

Zeitstempel

Replay, wie vorher, jetzt unter Berücksichtigung der **Zeitstempel** sowie **Anfang** und **Ende** der Aktivitäten**:**

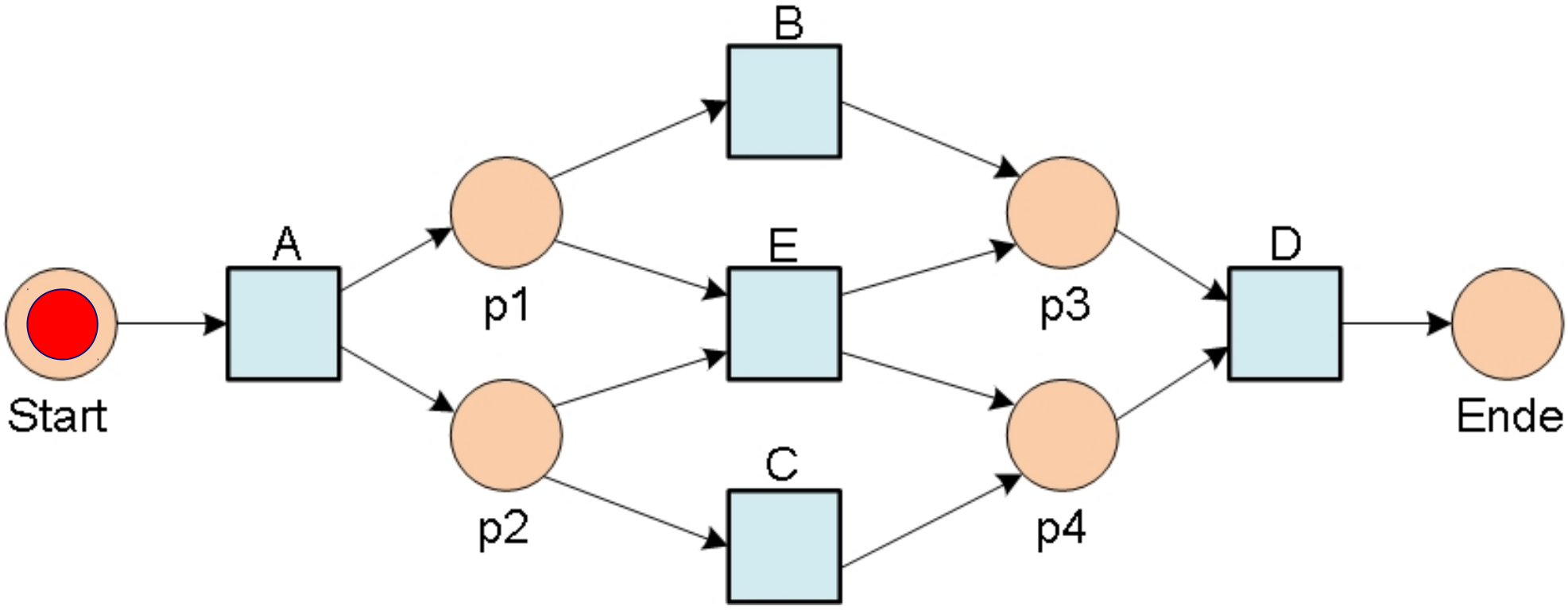
Replay der ersten drei Fälle im Event-Log:

- Fall 1 startet zur Zeit 12 und endet zur Zeit 54,
- Fall 2 startet zur Zeit 17 und endet zur Zeit 73,
- Fall 3 startet zur Zeit 25 und endet zur Zeit 98.

Timed Replay: Beispiel

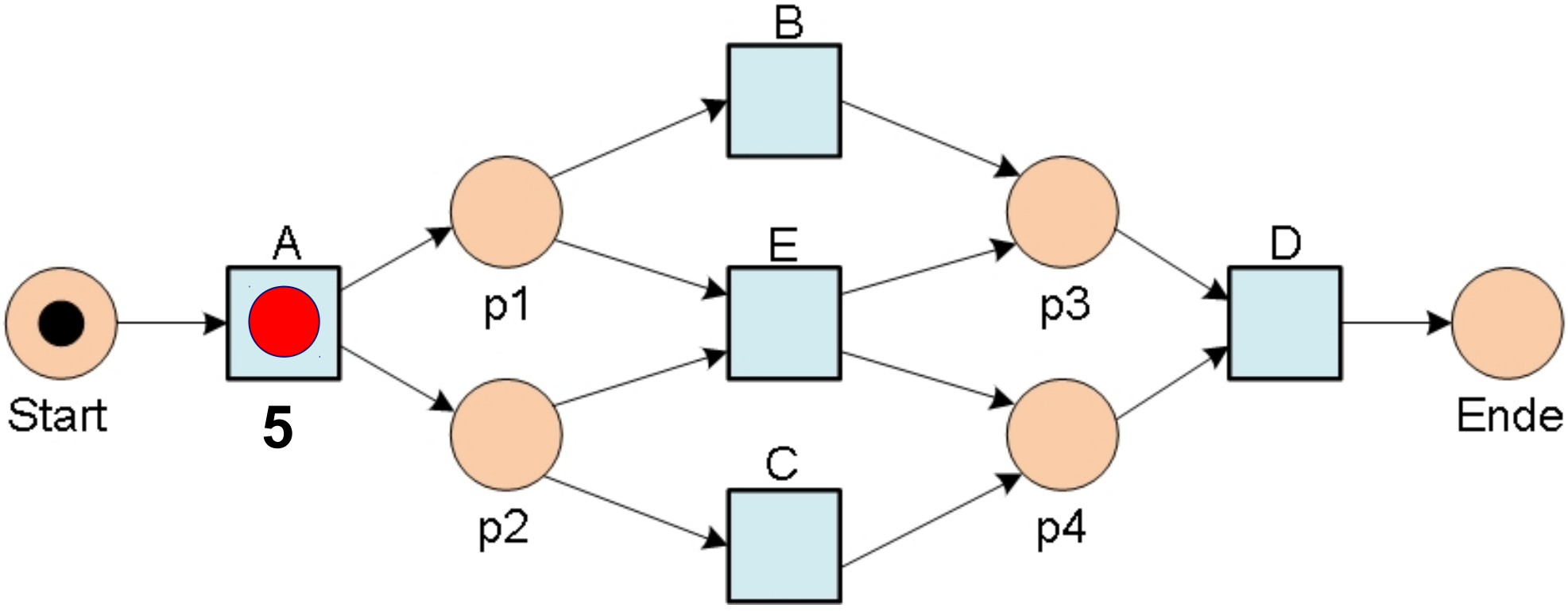
A^5 B^8 C^9 D^{13}

A^5: Ereignis A trat zum Zeitpunkt 5 ein.



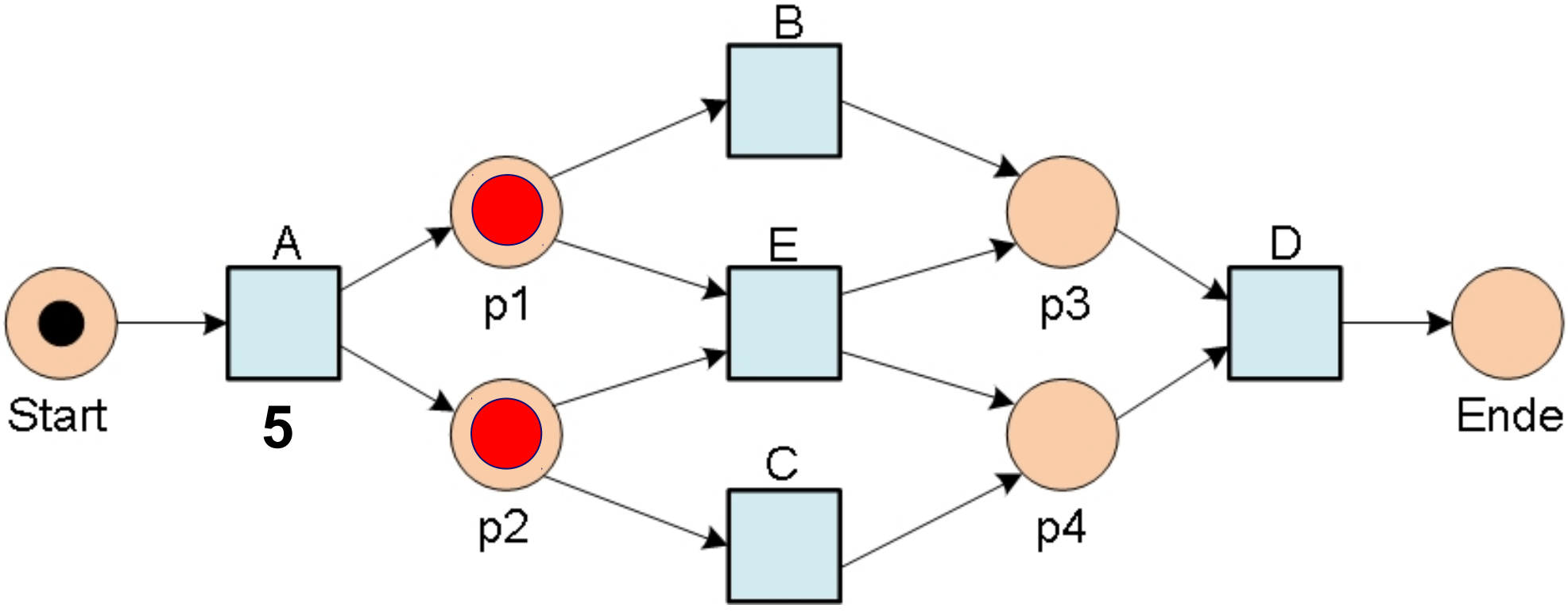
Timed Replay: Beispiel

B^8 C^9 D^{13}



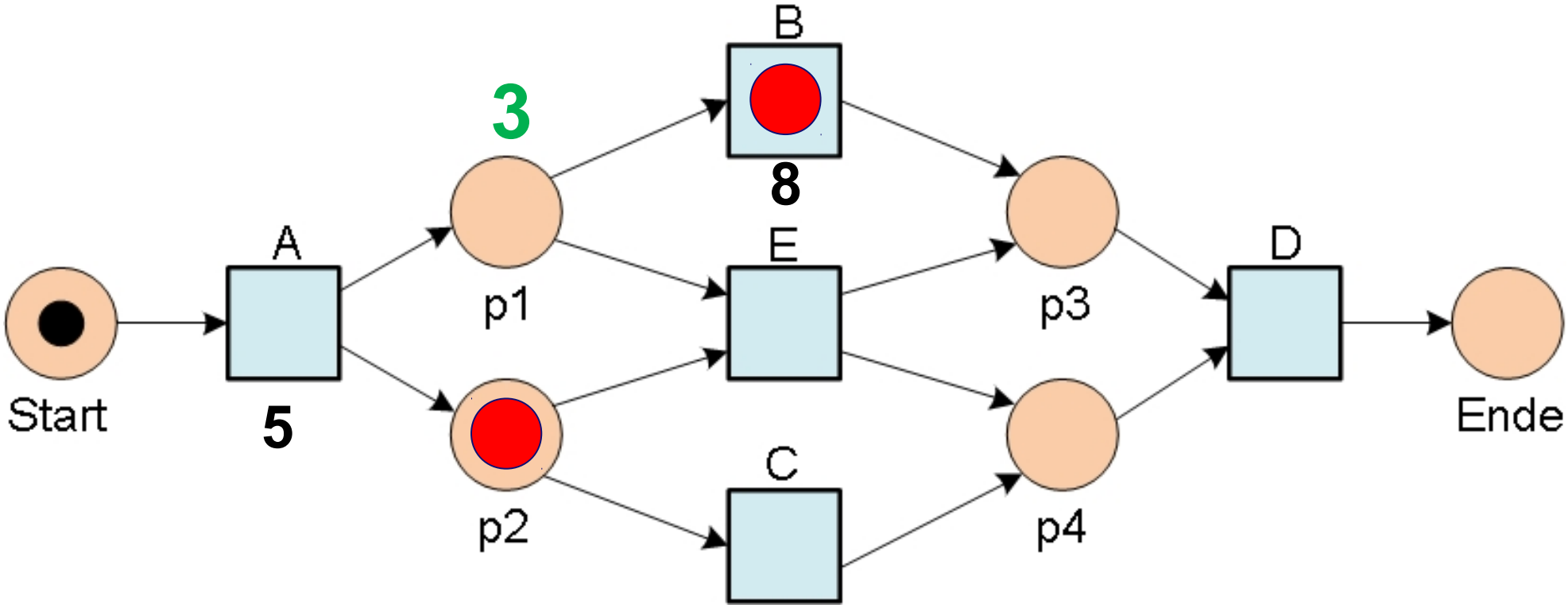
Timed Replay: Beispiel

B^8 C^9 D^{13}

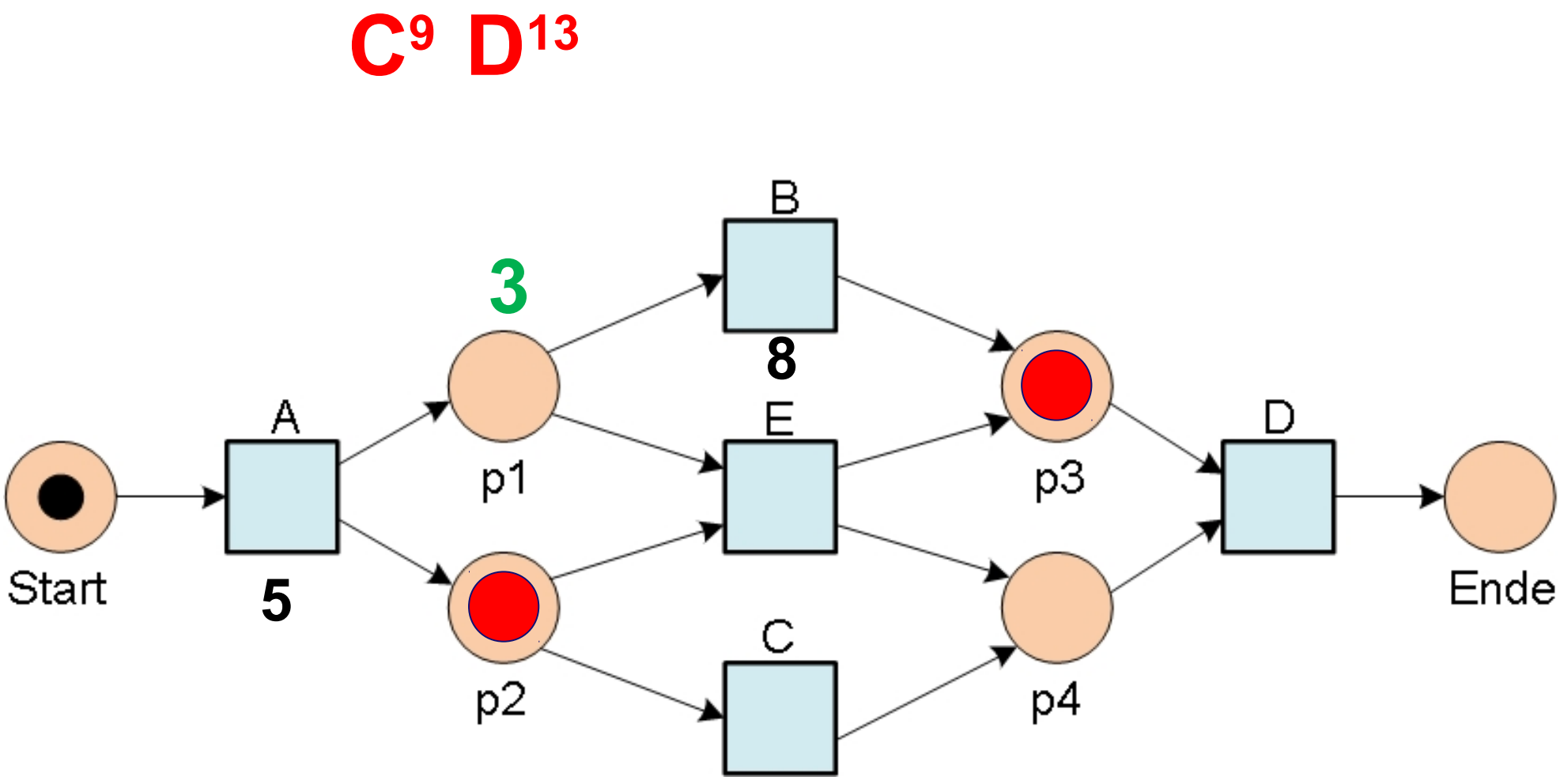


Timed Replay: Beispiel

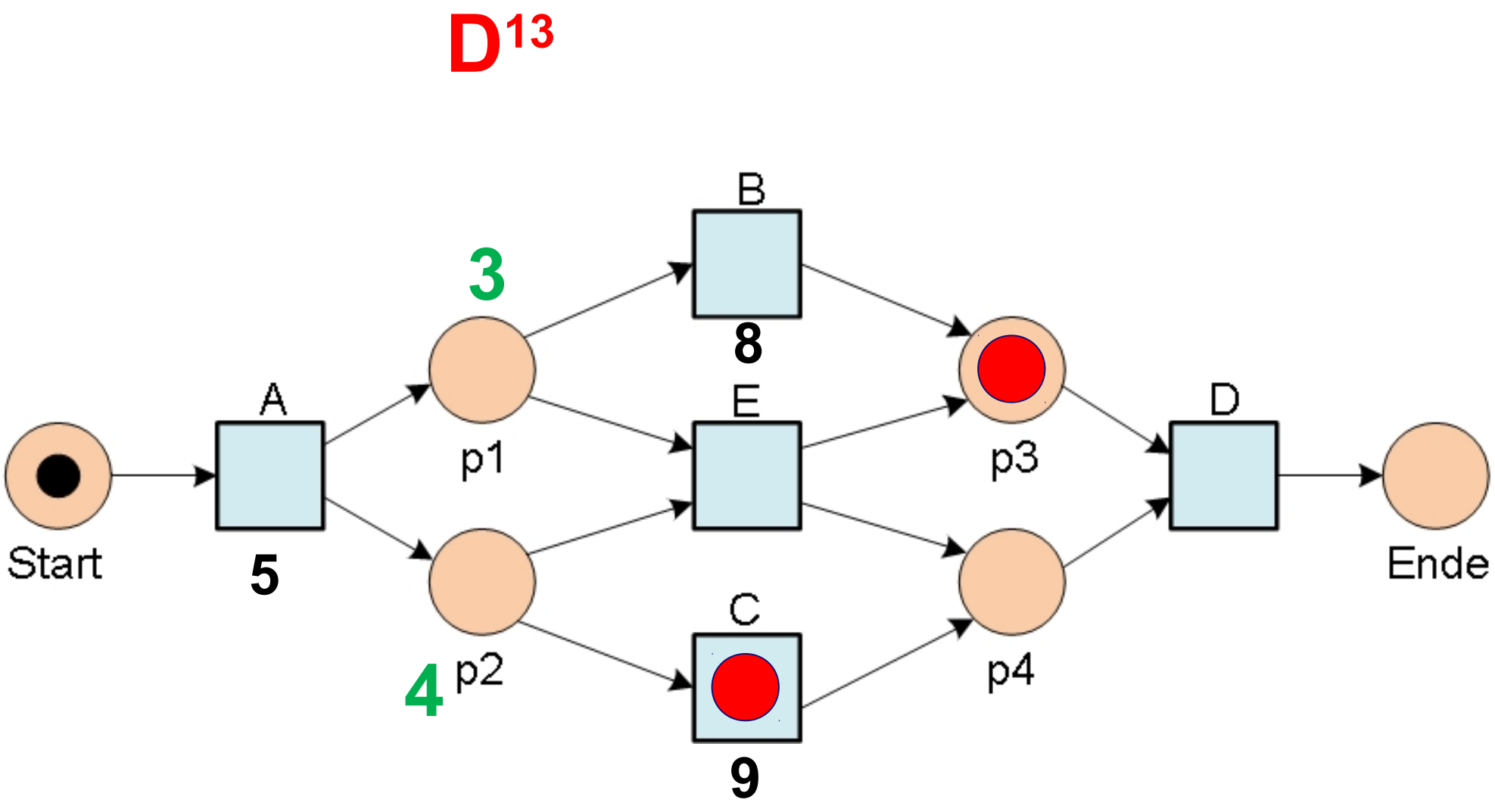
C^9 D^{13} **3** = 8-5: Zwischen Auftritten der Ereignisse A und B sind 3 Zeiteinheiten vergangen.



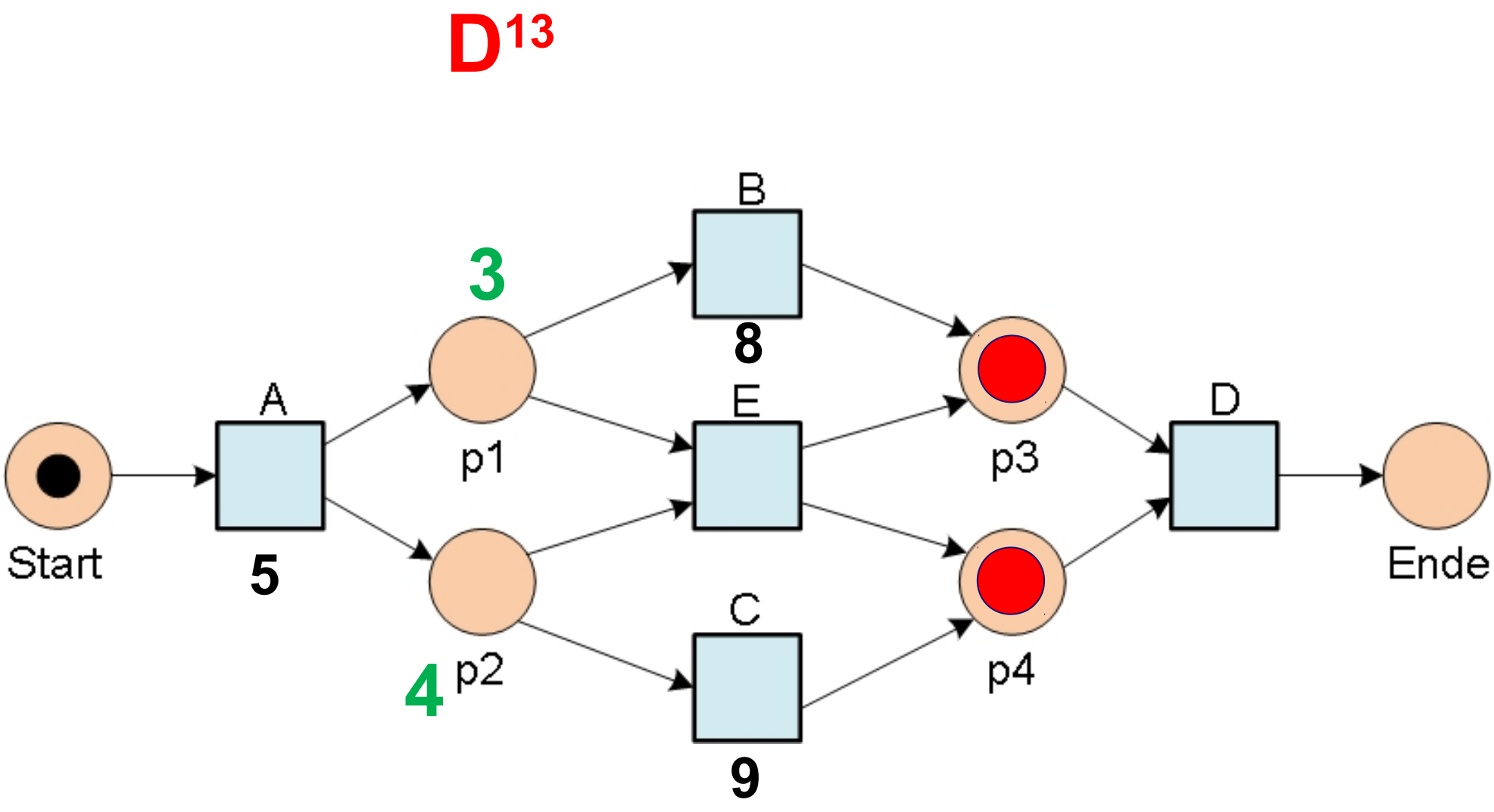
Timed Replay: Beispiel



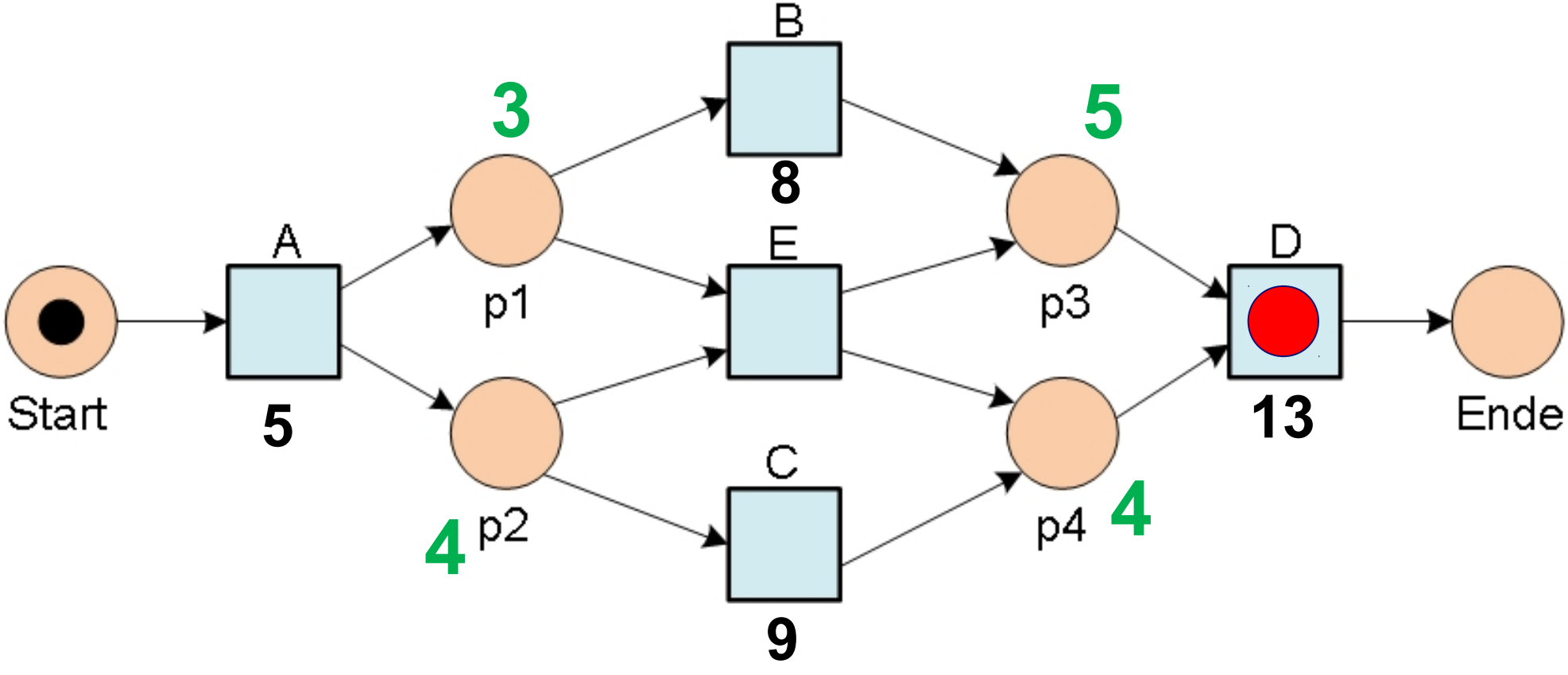
Timed Replay: Beispiel



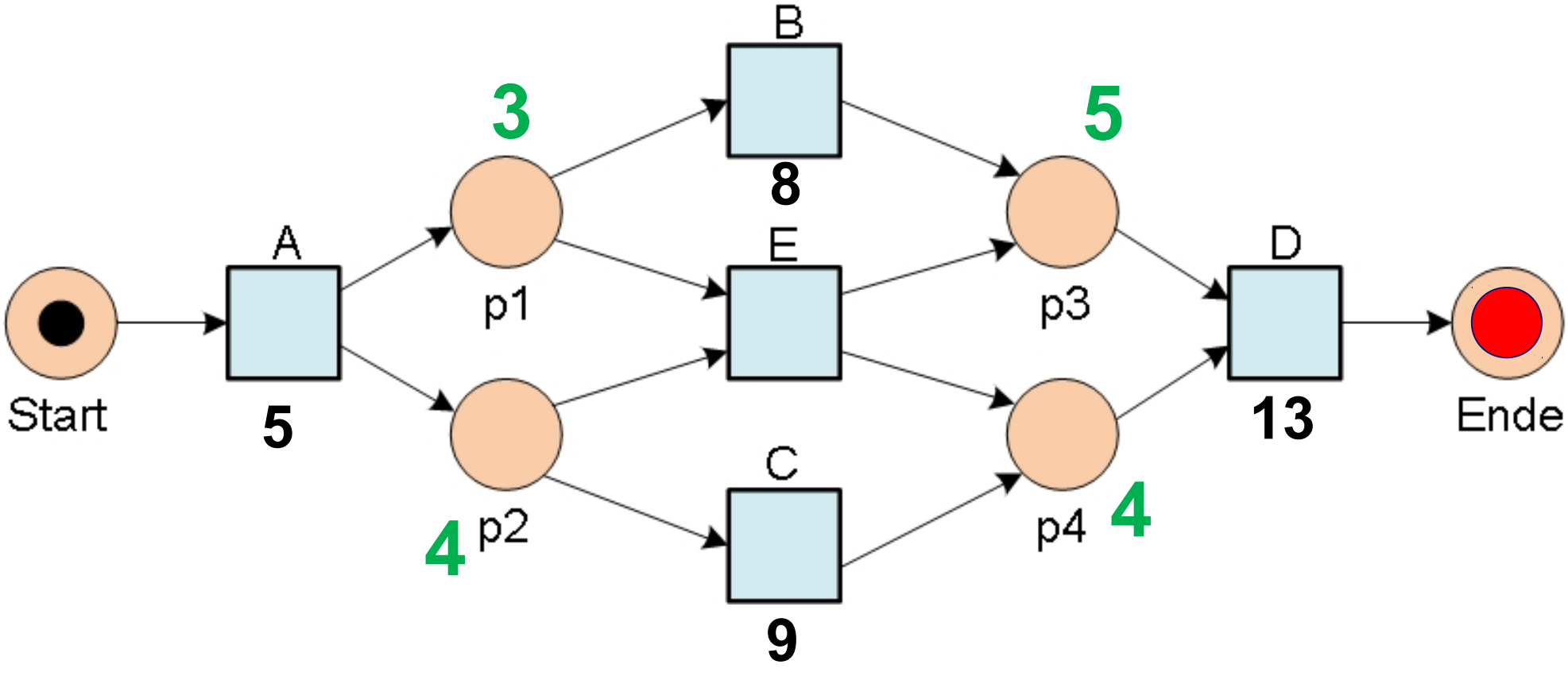
Timed Replay: Beispiel



Timed Replay: Beispiel

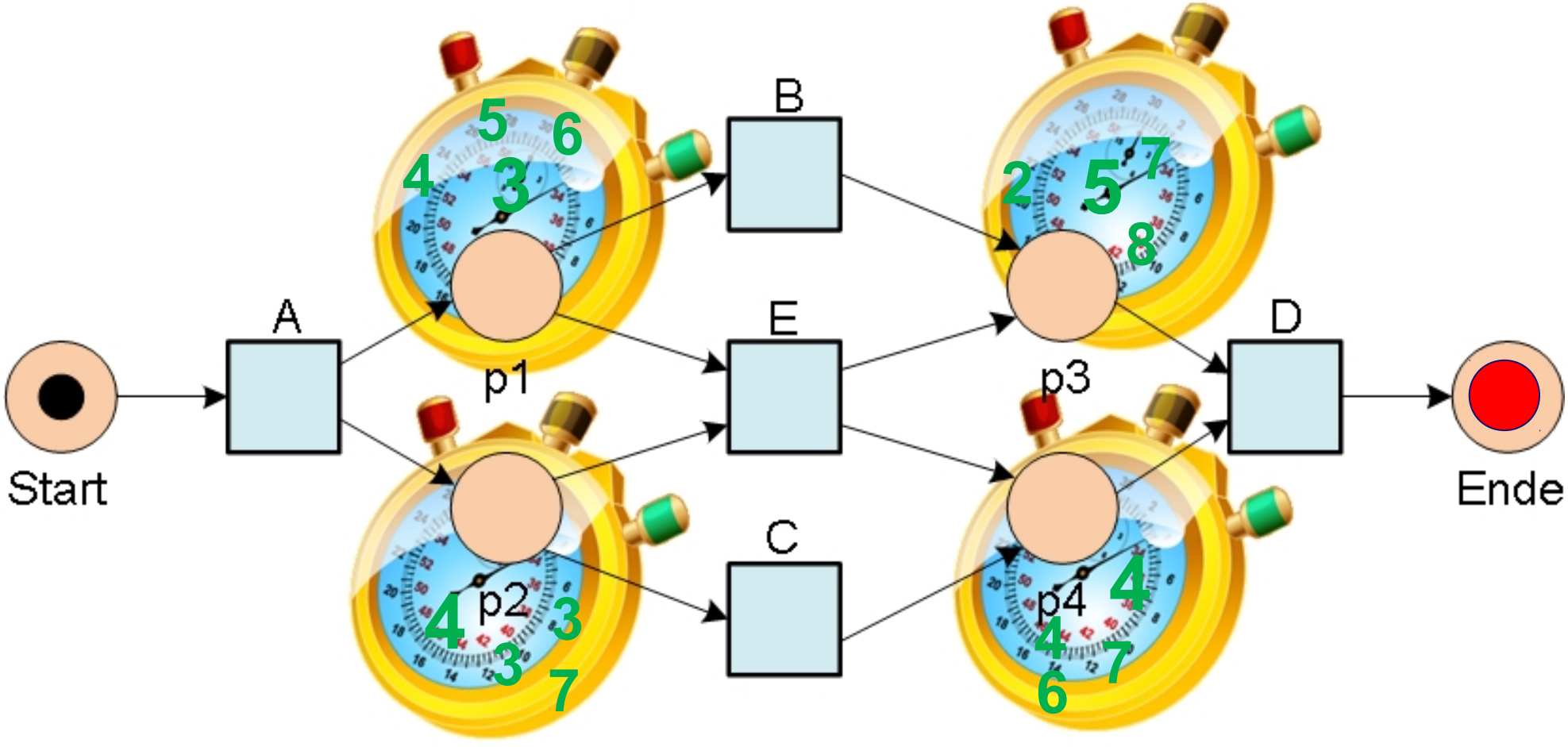


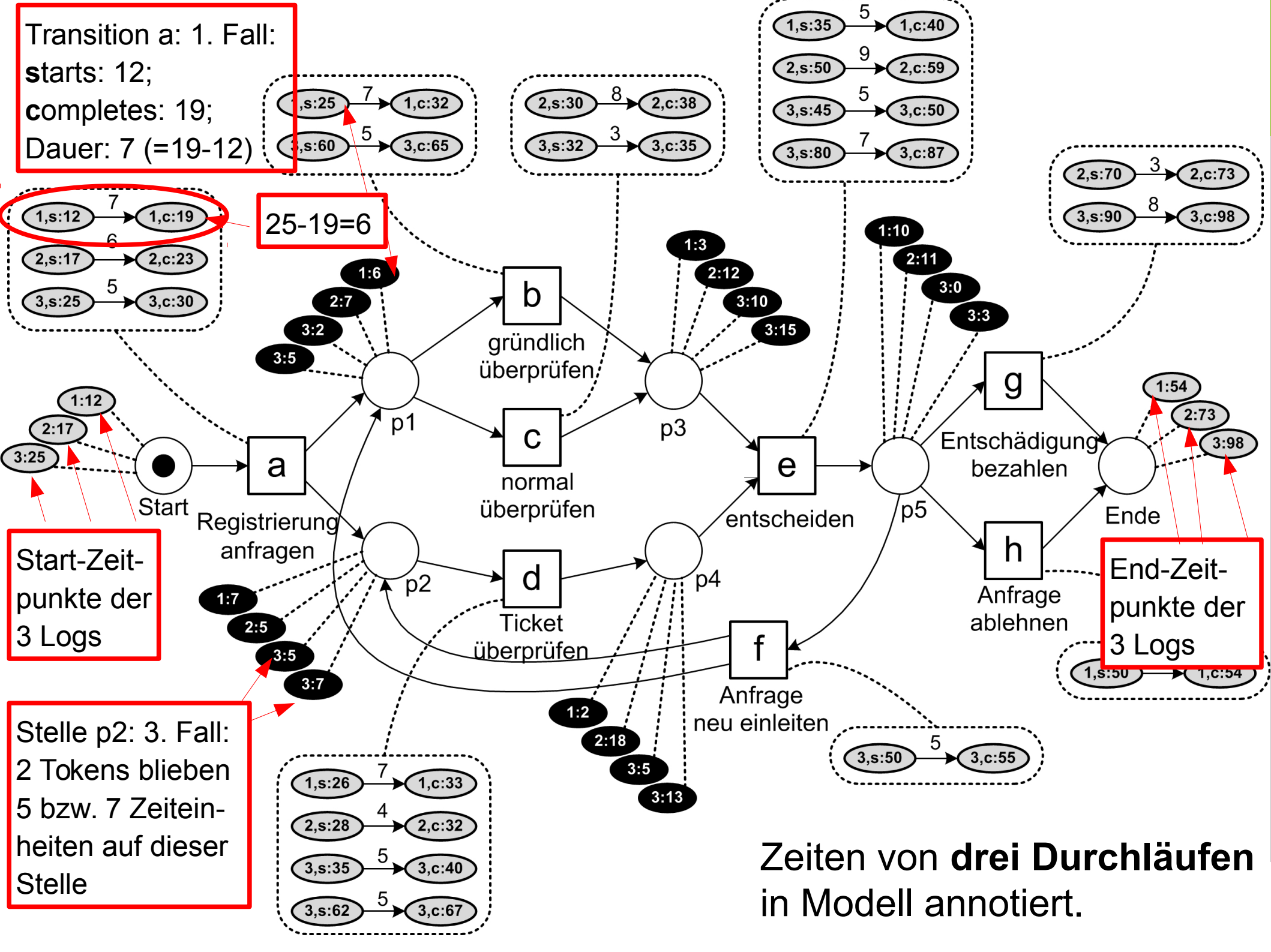
Timed Replay: Beispiel



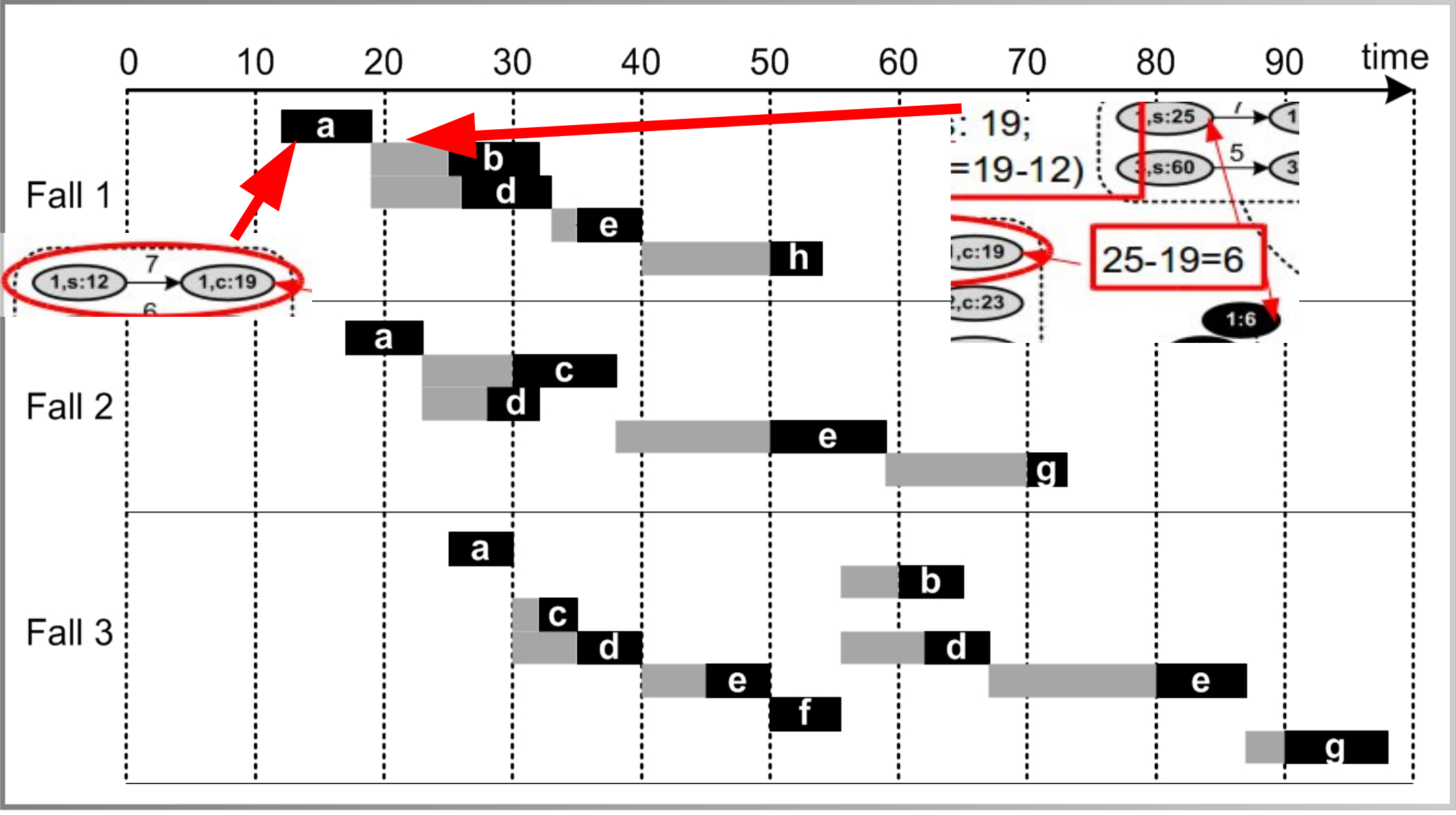
Timed Replay: Beispiel

I.A. für verschiedene Durchläufe unterschiedliche Zeiten.

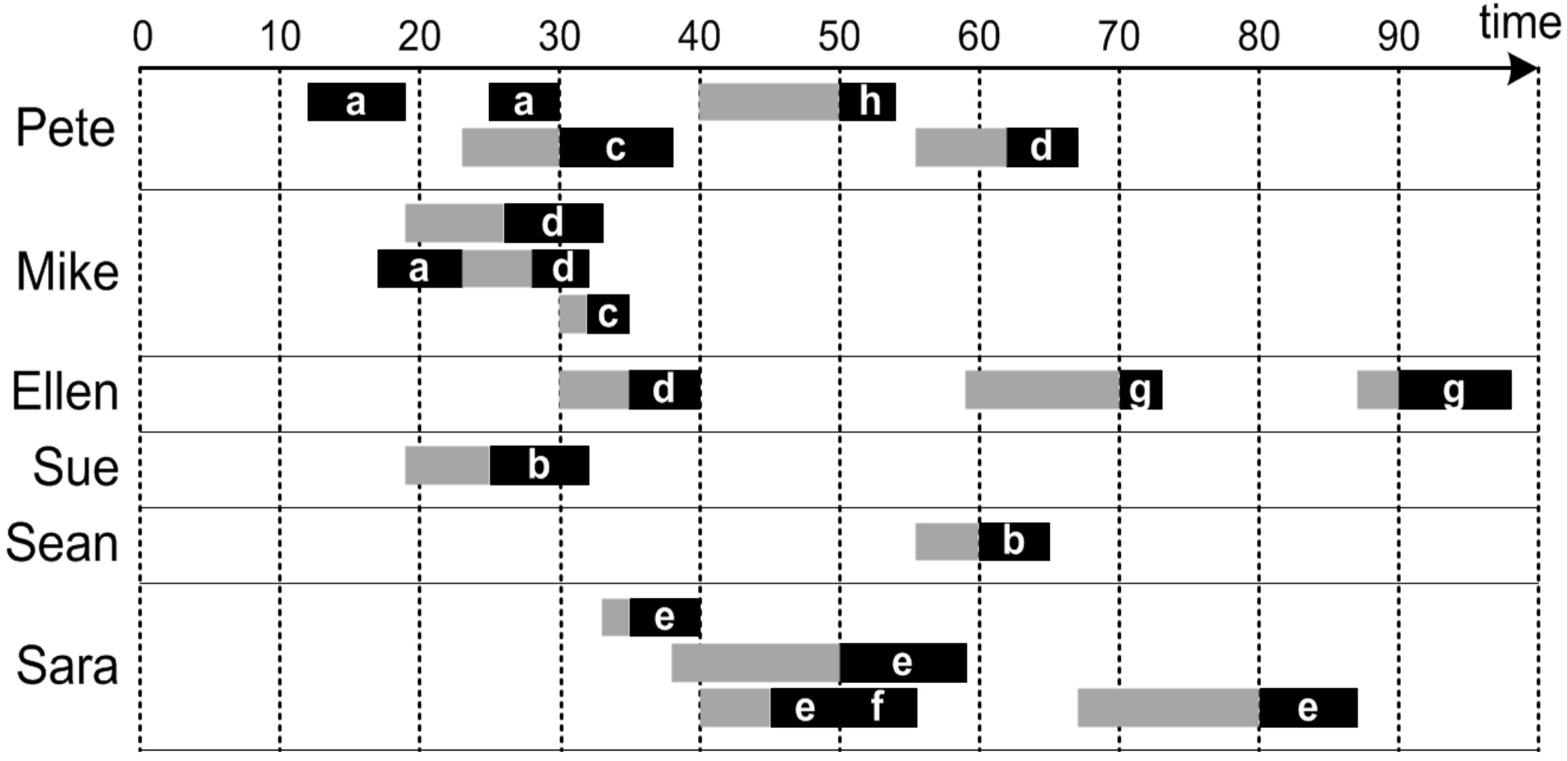



Transition a: 1. Fall:
starts: 12;
completes: 19;
Dauer: 7 (=19-12)
25-19=6
Start
Registrierung anfragen
a
p1
b
gründlich überprüfen
c
normal überprüfen
p2
d
Ticket überprüfen
p3
e
entscheiden
p4
f
Anfrage neu einleiten
p5
g
Entschädigung bezahlen
h
Anfrage ablehnen
Ende
Start-Zeit-punkte der 3 Logs
End-Zeit-punkte der 3 Logs
Stelle p2: 3. Fall: 2 Tokens blieben 5 bzw. 7 Zeitein-heiten auf dieser Stelle
Zeiten von drei Durchläufen in Modell annotiert.

Lineare Ansicht: Timed-Replay, erste drei Fälle



Lineare Ansicht auf Ressourcen abgebildet



(Aus voriger Folie durch Umsortieren der Zeilen erstellen.)

Überblick
Mining: Zusätzliche Perspektiven

- Attribute in Event-Logs
- Organizational Mining
- Zeit-Analysen
- **Decision-Mining**

Decision Mining

Entscheidungspunkte in extrahierten Petrinetzen zunächst „**nicht-deterministisch**“:

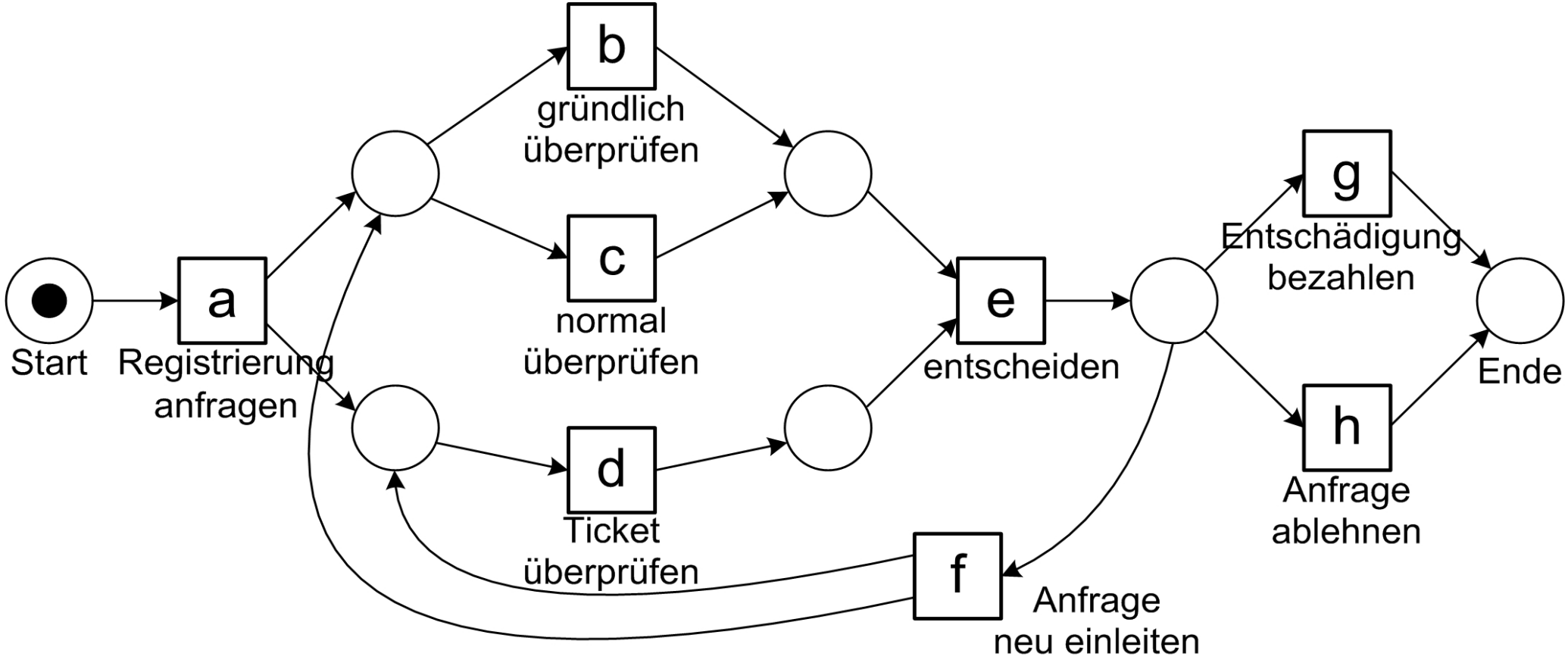
- im Modell nicht determiniert, welcher Ausführungszweig in welcher Ausführung gewählt wird

Nützliche Information !

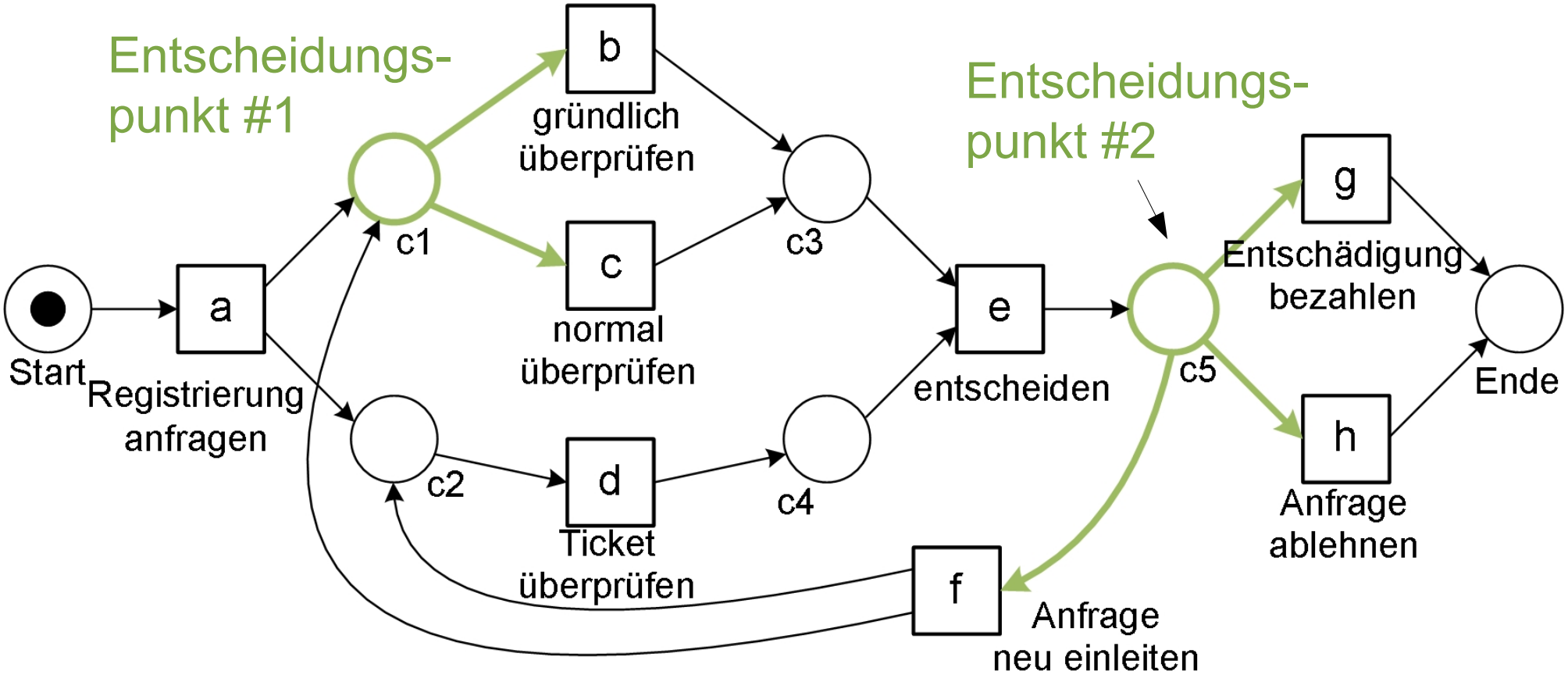
Idee: **Klassifikationstechniken** (s. Abschnitt 2.2) anwenden, um Rationale hinter der in den Ausführungen gewählten Entscheidungen auf Basis der Logdaten zu erkennen.

=> „**Decision Mining**“

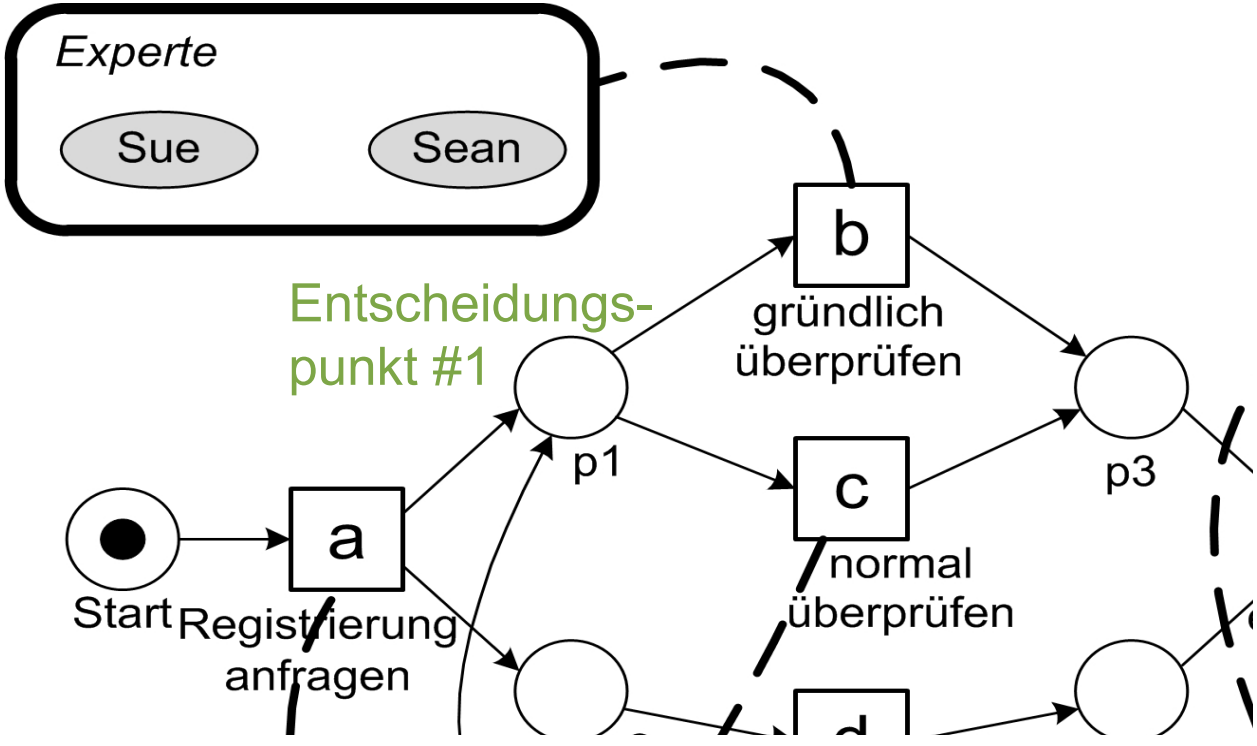
Decision-Mining: Beispiel
Wo sind die Entscheidungspunkte ?



Decision-Mining: Beispiel Entscheidungspunkte



Decision-Mining: Beispiel: Entscheidungspunkt 1



Entscheidungspunkt #1:

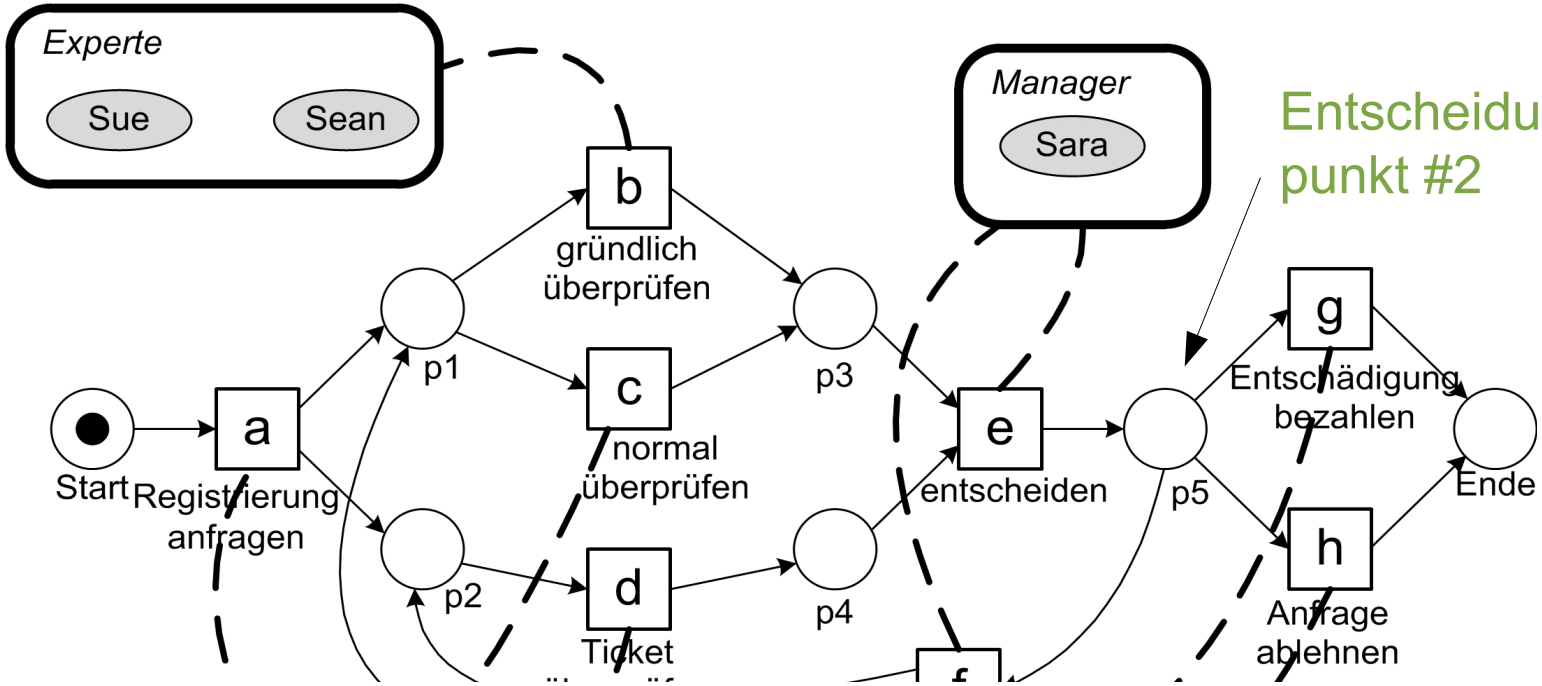
Wahl zwischen Aktivität b und c.

Mögliche Erkenntnis: Von Arbeitslast der beiden Experten abhängig.

- Wenn Experten Sue und Sean überladen

  → Ausführung von b weniger wahrscheinlich (gegenüber c).

Decision-Mining: Beispiel: Entscheidungspunkt 2



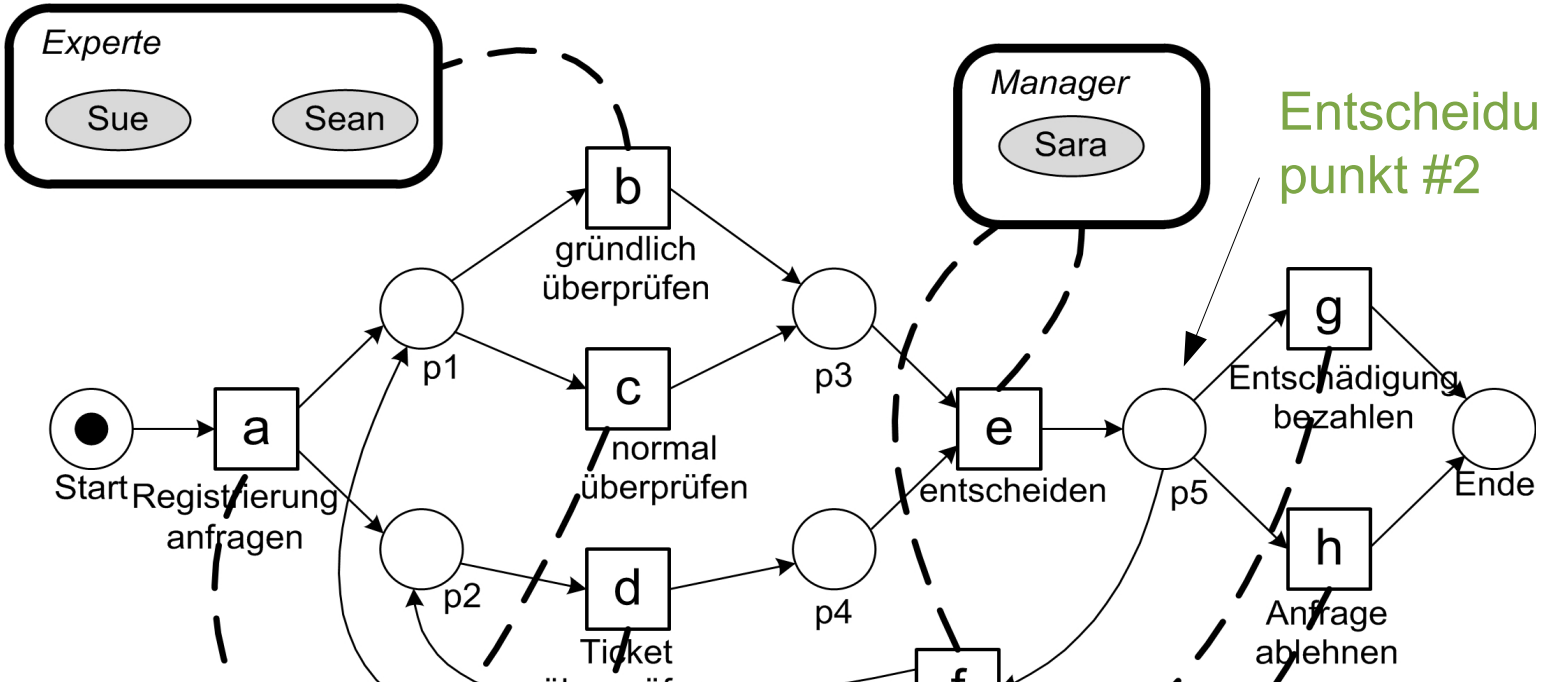
Entscheidungspunkt #2:

Mögliche Erkenntnis:

Alle Fälle, die **Sean gründlich prüft**, werden an Entscheidungspunkt #2 ... ?

case id	trace
1	$\langle a^{Pete}, b^{Sue}, d^{Mike}, e^{Sara}, h^{Pete} \rangle$
2	$\langle a^{Mike}, d^{Mike}, c^{Pete}, e^{Sara}, g^{Ellen} \rangle$
3	$\langle a^{Pete}, c^{Mike}, d^{Ellen}, e^{Sara}, f^{Sara}, \ldots$
4	$\langle a^{Pete}, d^{Mike}, b^{Sean}, e^{Sara}, h^{Ellen} \rangle$
5	$\langle a^{Ellen}, c^{Mike}, d^{Pete}, e^{Sara}, f^{Sara}, \ldots$
6	$\langle a^{Mike}, c^{Ellen}, d^{Mike}, e^{Sara}, g^{Mike} \ldots$
...	...

Decision-Mining: Beispiel: Entscheidungspunkt 2



Entscheidungspunkt #2:

Mögliche Erkenntnis:

Alle Fälle, die **Sean gründlich prüft**, werden an Entscheidungspunkt #2 **abgelehnt**.

case id	trace
1	$\langle a^{Pete}, b^{Sue}, d^{Mike}, e^{Sara}, h^{Pete} \rangle$
2	$\langle a^{Mike}, d^{Mike}, c^{Pete}, e^{Sara}, g^{Ellen} \rangle$
3	$\langle a^{Pete}, c^{Mike}, d^{Ellen}, e^{Sara}, f^{Sara}, \ldots$
4	$\langle a^{Pete}, d^{Mike}, b^{Sean}, e^{Sara}, h^{Ellen} \rangle$
5	$\langle a^{Ellen}, c^{Mike}, d^{Pete}, e^{Sara}, f^{Sara}, \ldots$
6	$\langle a^{Mike}, c^{Ellen}, d^{Mike}, e^{Sara}, g^{Mike} \ldots$
...	...

Decision Mining: Predictor- / Response-Variablen

Predictor Variablen

Response-Variable

Kunden-status	Region	Betrag	Aktivität
Gold	Norden	987.30	h
Silber	Norden	178.70	h
Gold	Süden	211.50	g
Silber	Westen	587.70	h
Silber	Osten	224.70	h
Silber	Süden	278.50	h
Gold	Norden	488.50	g
Silber	Westen	443.20	h
Silber	Süden	673.70	h
Gold	Westen	413.50	g
Silber	Süden	687.70	h
Gold	Süden	987.30	h
Silber	Norden	378.80	h
Gold	Süden	314.50	g
Silber	Norden	537.70	h
Silber	Westen	158.70	h
Gold	Osten	344.50	g
...	...	...	...

Predictor-Variablen (unabhängige Variablen):

- Entsprechen Wissen über Fall, nachdem Entscheidung getroffen.

Response-Variable (abhängige Variablen):

- Ermittlung mittels Untersuchung des Event-Logs.

Jede Zeile in der Tabelle = eine Ausführung der Aktivitität f/g/h

- Aktivität in Prozessausführung mehrmals besucht (Schleife) → mehrere Zeilen in Tabelle.

Beispiel: Entscheidungspunkt 2 in Abhängigkeit von Predictors

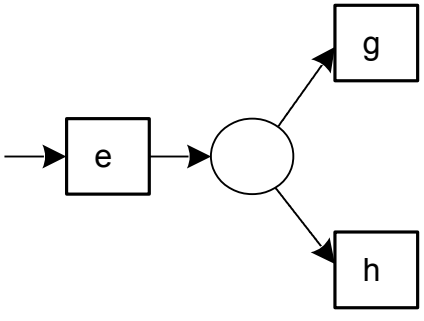
Kunden-status	Region	Betrag	Aktivität
Gold	Norden	987.30	h
Silber	Norden	178.70	h
Gold	Süden	211.50	g
Silber	Westen	587.70	h
Silber	Osten	224.70	h
Silber	Süden	278.50	h
Gold	Norden	488.50	g
Silber	Westen	443.20	h
Silber	Süden	673.70	h
Gold	Westen	413.50	g
Silber	Süden	687.70	h
Gold	Süden	987.30	h
Silber	Norden	378.80	h
Gold	Süden	314.50	g
Silber	Norden	537.70	h
Silber	Westen	158.70	h
Gold	Osten	344.50	g
...	...	...	...

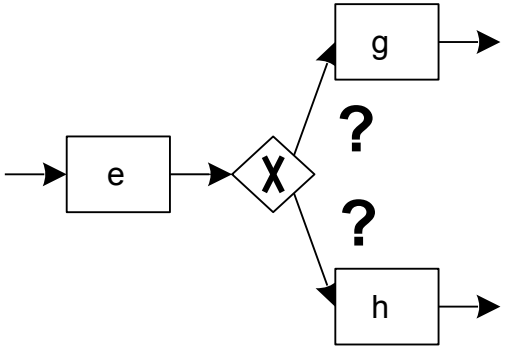
Welche "**Features**" (unabhängige Variablen) beeinflussen die Entscheidung ?

Klassifikationstechniken (z.B. Entscheidungsbäume) nutzen, um Regeln zu finden.

Erkläre **abhängige Variablen** hinsichtlich der unabhängigen.

Beispiel: Wann wird *Entschädigung bezahlt* (g) und wann *Anfrage abgelehnt* (h) ?





Beispiel: Entscheidungspunkt 2 in Abhängigkeit von Predictors

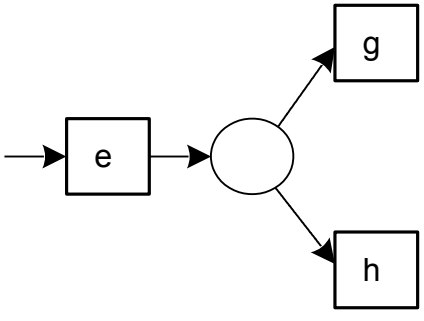
Kunden-status	Region	Betrag	Aktivität
Gold	Norden	987.30	h
Silber	Norden	178.70	h
Gold	Süden	211.50	g
Silber	Westen	587.70	h
Silber	Osten	224.70	h
Silber	Süden	278.50	h
Gold	Norden	488.50	g
Silber	Westen	443.20	h
Silber	Süden	673.70	h
Gold	Westen	413.50	g
Silber	Süden	687.70	h
Gold	Süden	987.30	h
Silber	Norden	378.80	h
Gold	Süden	314.50	g
Silber	Norden	537.70	h
Silber	Westen	158.70	h
Gold	Osten	344.50	g
...	...	...	...

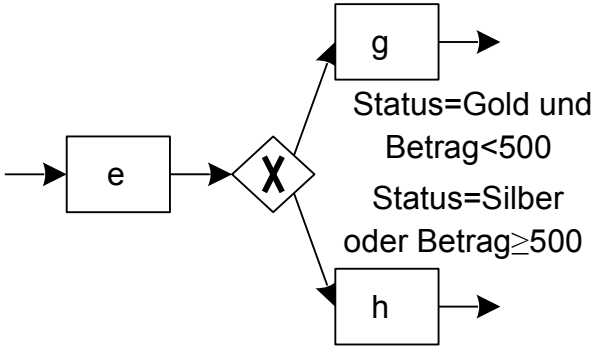
Welche "**Features**" (unabhängige Variablen) beeinflussen die Entscheidung ?

Klassifikationstechniken (z.B. Entscheidungsbäume) nutzen, um Regeln zu finden.

Erkläre **abhängige Variablen** hinsichtlich der unabhängigen.

Beispiel: Wann wird *Entschädigung bezahlt* (g) und wann *Anfrage abgelehnt* (h) ?

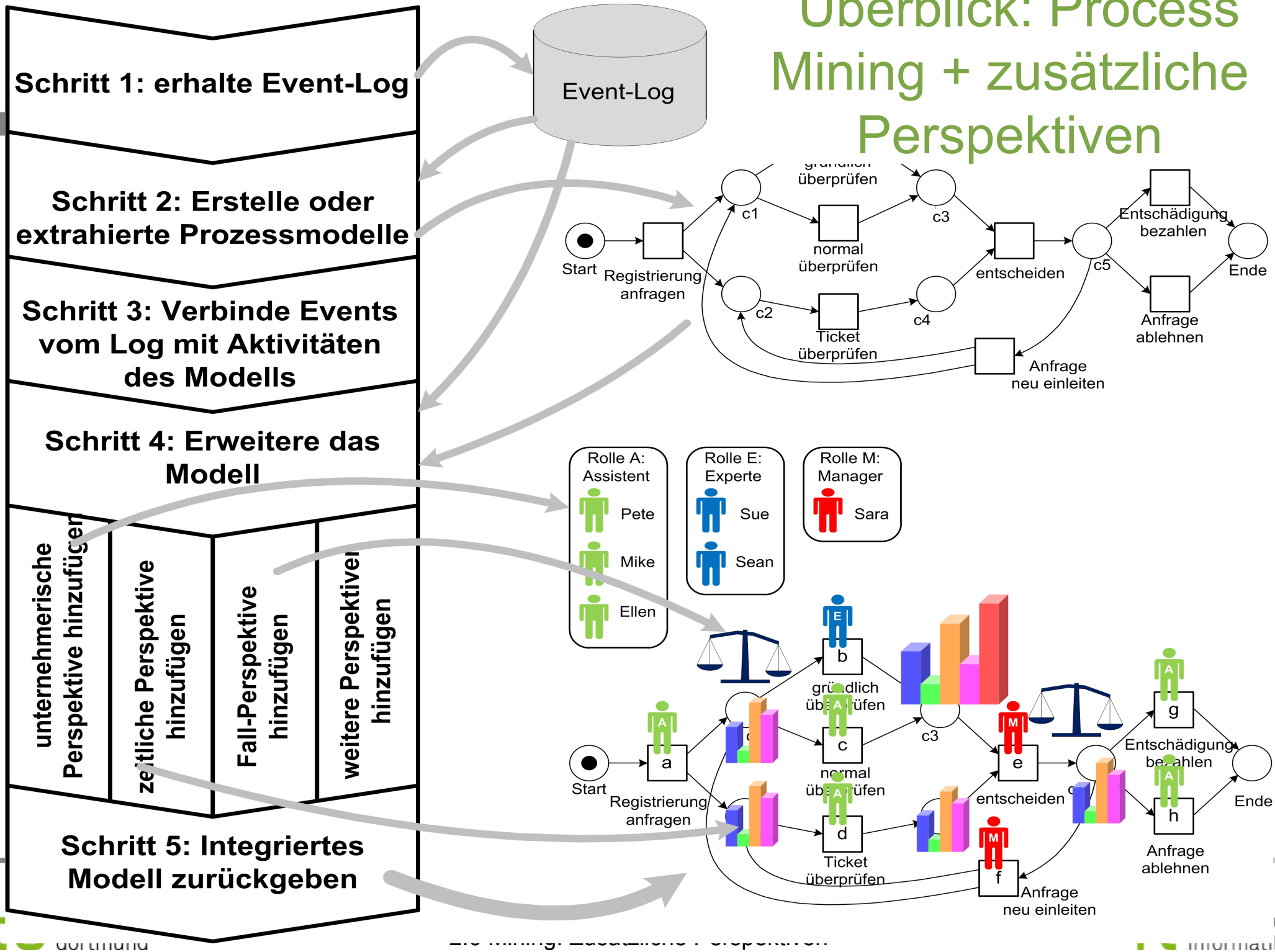




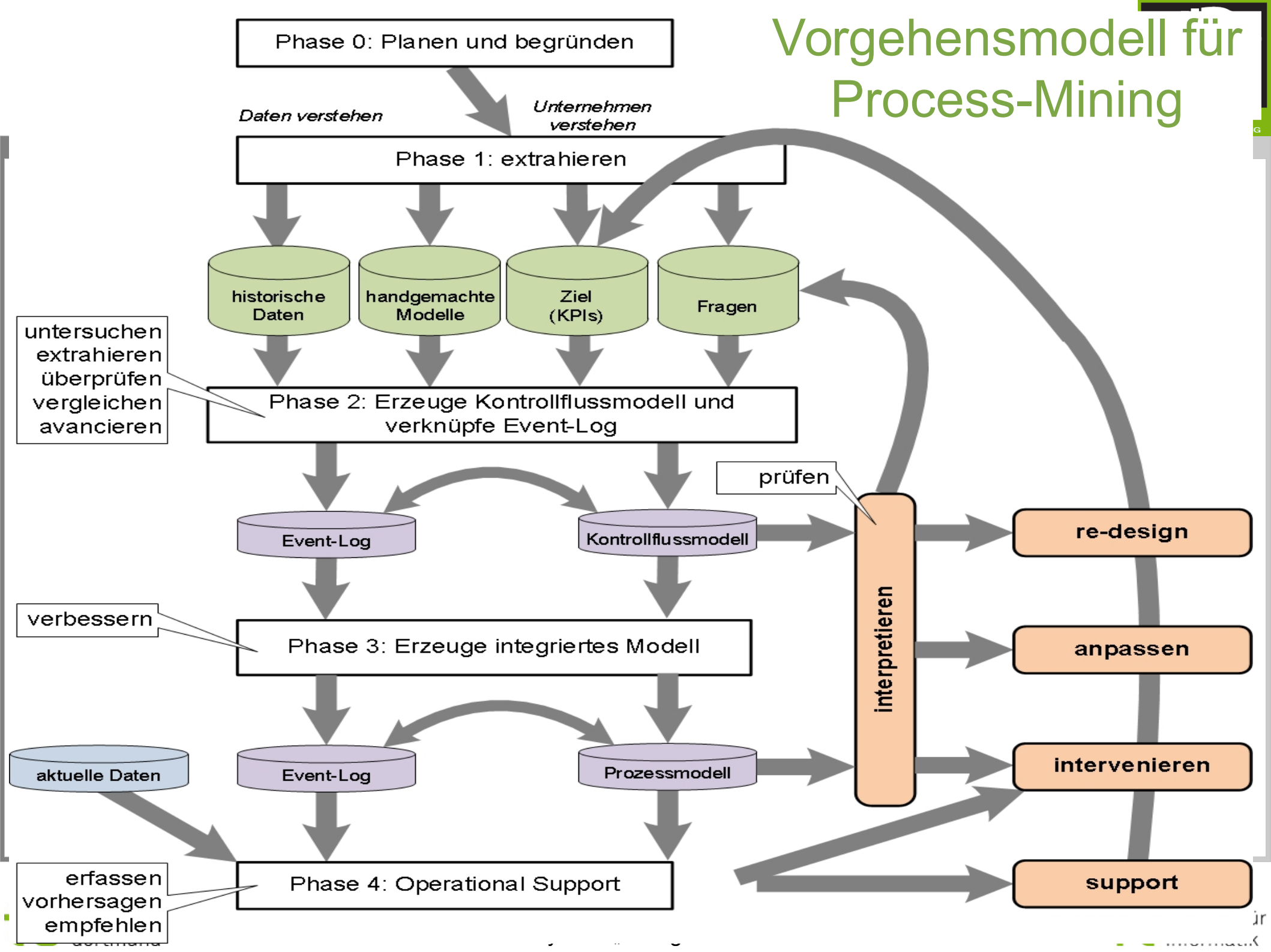
Klassifikation beim Process-Mining

- Anwendung von **Klassifikationstechniken** nicht auf Event- / Daten-basiertes Decision-Mining beschränkt.
- Zusätzliche **unabhängige Variablen** möglich:
  - **Verhaltensinformationen** (Anzahl der Schleifen).
  - **Performanz**-Informationen (Bearbeitungszeit).
  - **Kontextinformationen** (Wetter, Queues, etc.).
- Alternative **abhängige Variablen** analysierbar:
  - Gründe für **Nicht-Konformität** aufdecken (teile Instanzen in zwei Gruppen).
  - Gründe für **Verzögerungen** aufdecken.

Überblick: Process Mining + zusätzliche Perspektiven



Vorgehensmodell für Process-Mining



Zusammenfassung
Mining: Zusätzliche Perspektiven

In diesem Abschnitt:

- Attribute in Event-Logs
- Organizational Mining
- Zeit-Analysen
- Decision-Mining

Im nächsten Abschnitt:

- Online-Analysen (Erfassen, Vorhersagen und Empfehlen von Pfaden zur Ausführungszeit).