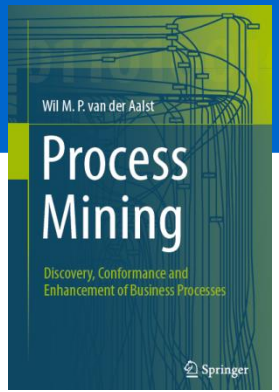
Process Mining: Data Science in Action

Token-Based Replay: Some Examples

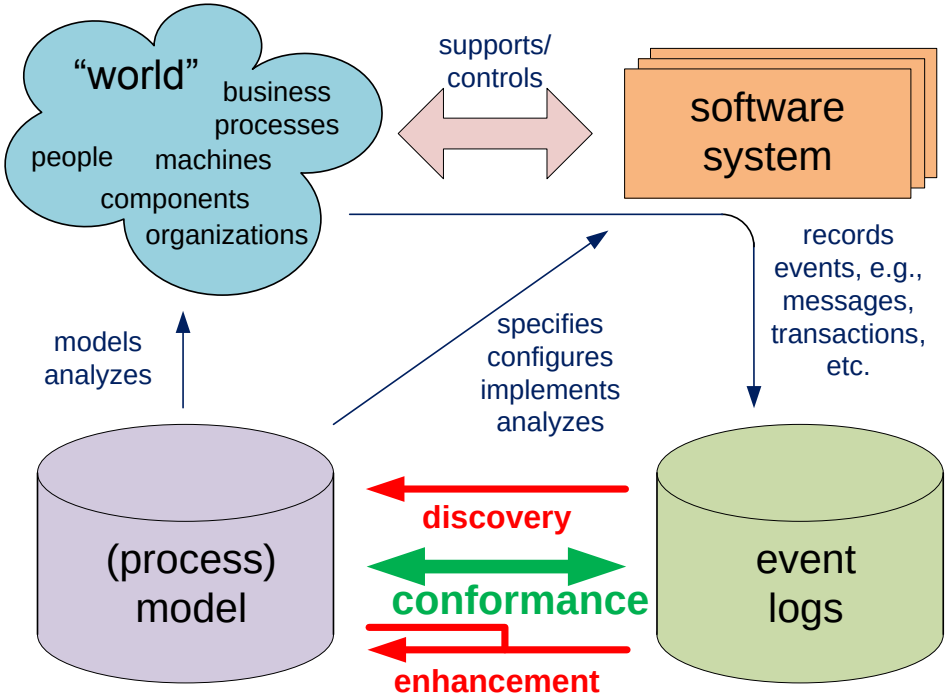


prof.dr.ir. Wil van der Aalst
www.processmining.org

TU/e Technische Universiteit Eindhoven University of Technology

Where innovation starts

Conformance checking



1. Conformance checking using causal footprints.
2. **Conformance checking based on token-based replay.**
3. Alignment-based conformance checking.

TU/e

Last lecture: Token-based replay

#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

missing tokens

?

remaining tokens

$$fitness(L,N) = \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}}\right) + \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}}\right)$$

consumed tokens

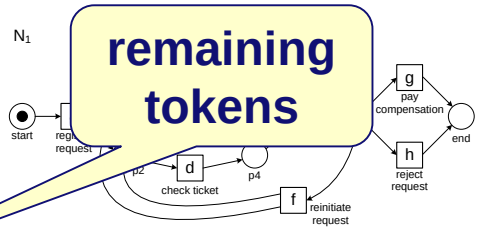
produced tokens

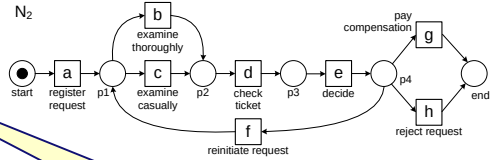
$$fitness(L_{full}, N_1) = 1$$

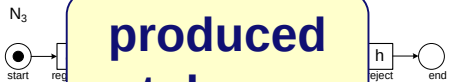
$$fitness(L_{full}, N_2) = 0.9504$$

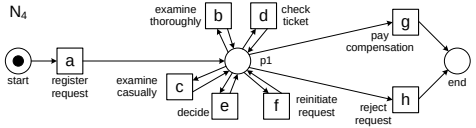
$$fitness(L_{full}, N_3) = 0.8797$$

$$fitness(L_{full}, N_4) = 1$$







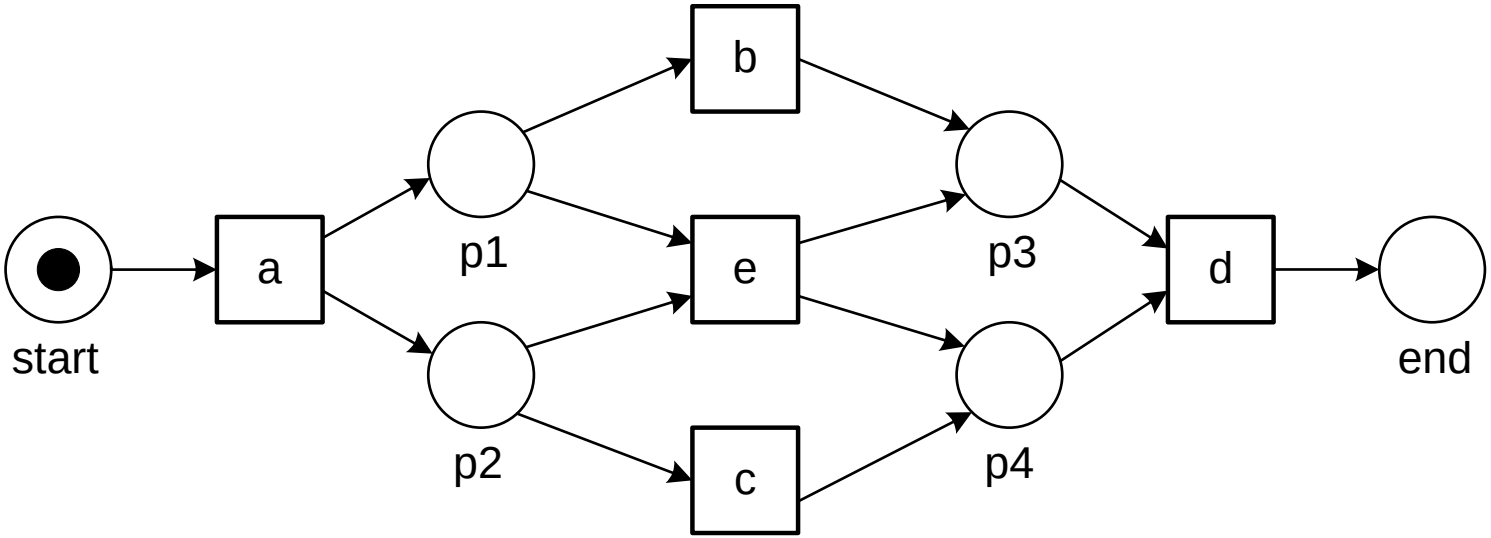


TU/e

Question (may take some time)

Compute fitness using missing and remaining tokens

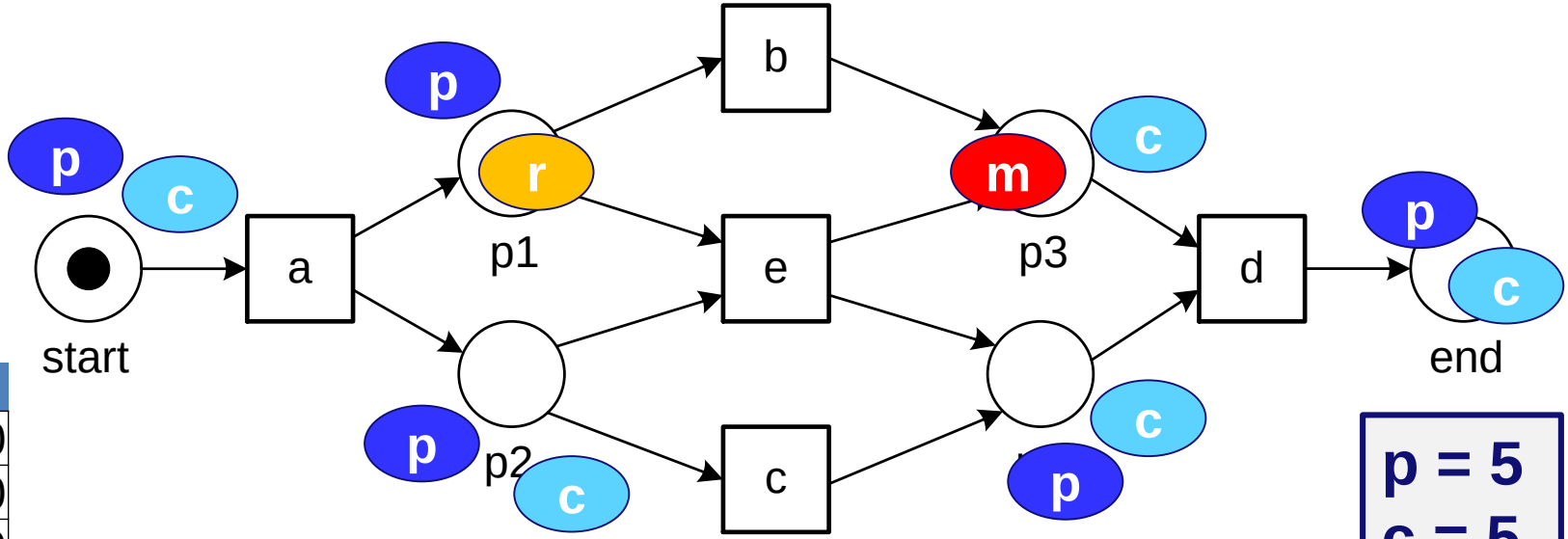
trace	frequency
abcd	10
acbd	10
aed	10
abd	2
acd	1
ad	1
abbd	1



- **Consider the event log containing 35 cases.**
- **What is the fitness?**

TU/e

Let us pick one trace: acd



trace	frequency
abcd	10
acbd	10
aed	10
abd	2
acd	1
ad	1
abbd	1

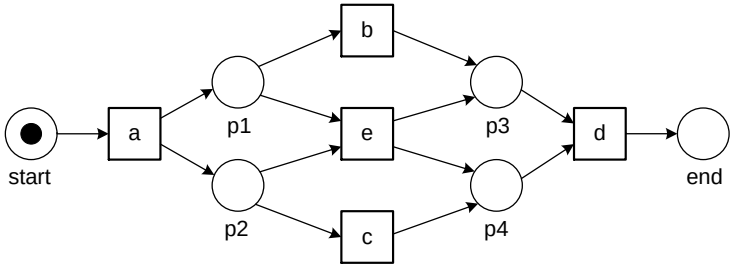
$$fitness(L,N) = \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}}\right) + \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}}\right)$$

p = 5
c = 5
m = 1
r = 1

TU/e

Fitness = 0.9658

trace	frequency	produced tokens (p)	remaining tokens (r)	consumed tokens (c)	missing tokens (m)	produced tokens (all)	remaining tokens (all)	consumed tokens (all)	missing tokens (all)
abcd	10	6	0	6	0	60	0	60	0
acbd	10	6	0	6	0	60	0	60	0
aed	10	6	0	6	0	60	0	60	0
abd	2	5	1	5	1	10	2	10	2
acd	1	5	1	5	1	5	1	5	1
ad	1	4	2	4	2	4	2	4	2
abbd	1	6	2	6	2	6	2	6	2



205	7	205	7
sum p	sum r	sum c	sum m

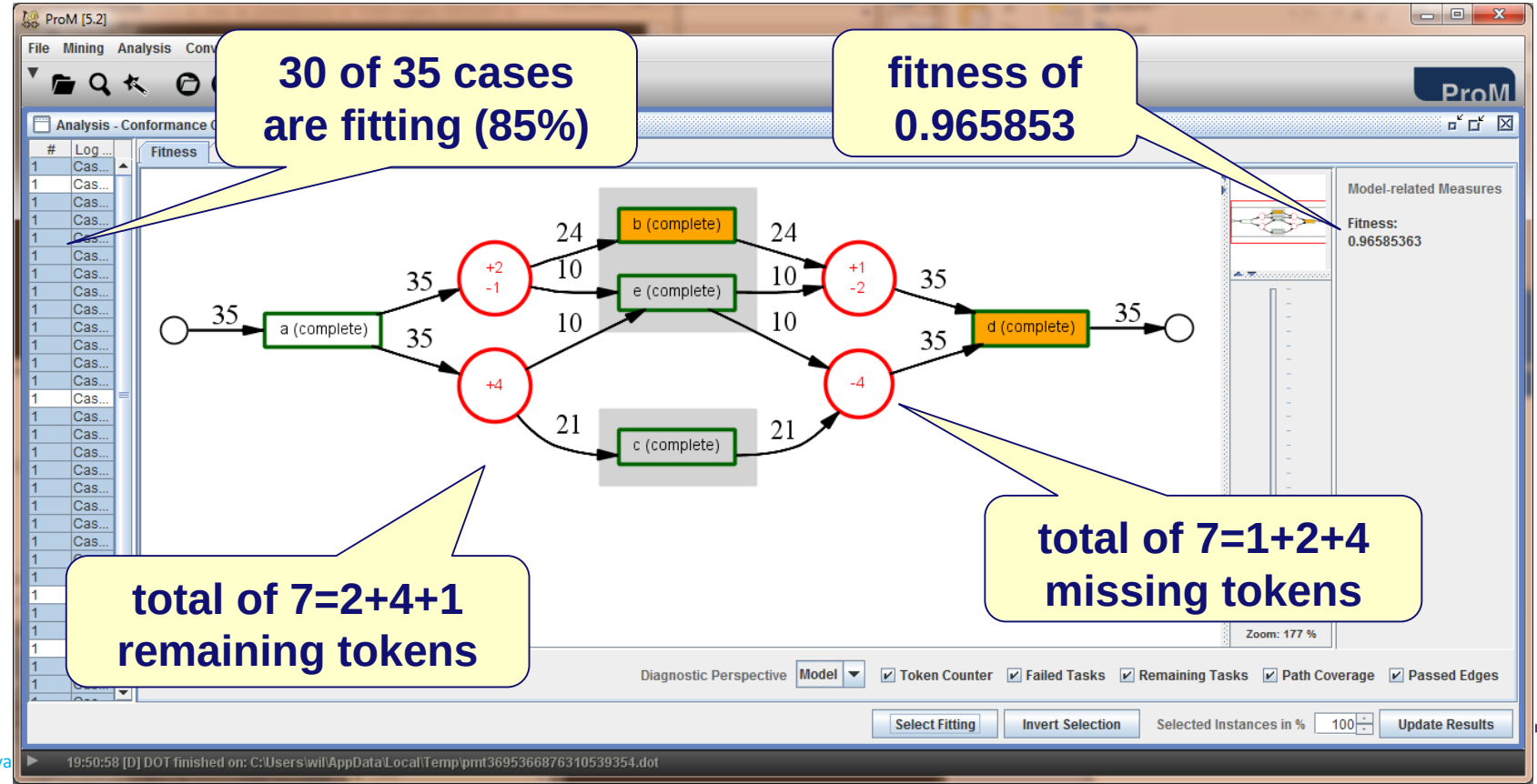
fitness	0.965853659

$$fitness(L,N) = \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}}\right) + \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}}\right)$$

TU/e

ProM 5.2 output

(ProM 6 only supports more advanced conformance checking techniques)

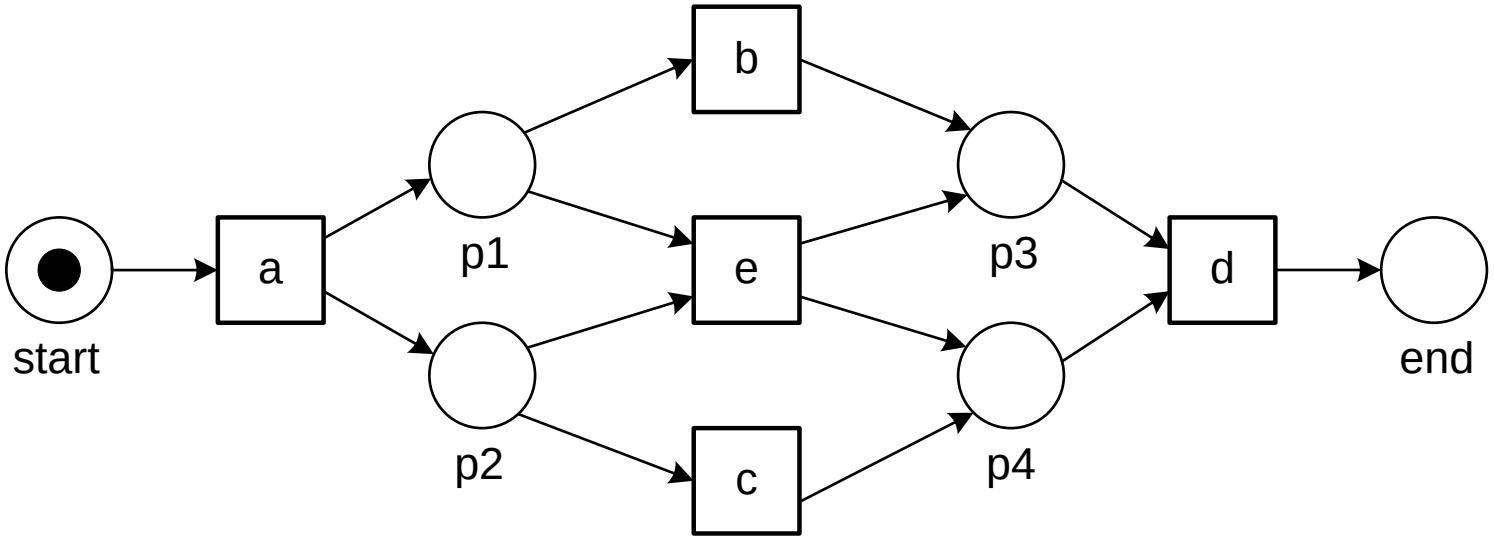


TU/e

Question

Compute fitness using missing and remaining tokens

trace	frequency
e	1

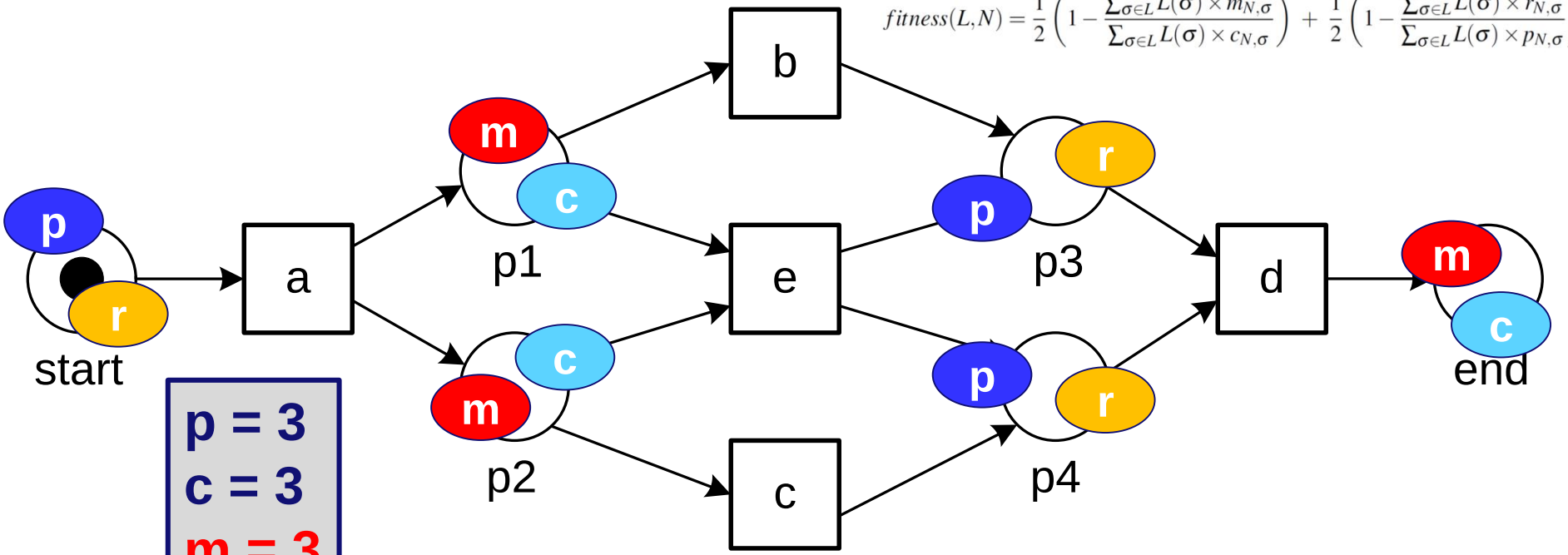


- **Consider the event log containing just one case:** $L = [\langle e \rangle]$.
- **What is the fitness (using token-based replay)?**

TU/e

Answer obtained by replaying ⟨e⟩

$$fitness(L,N) = \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}}\right) + \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}}\right)$$



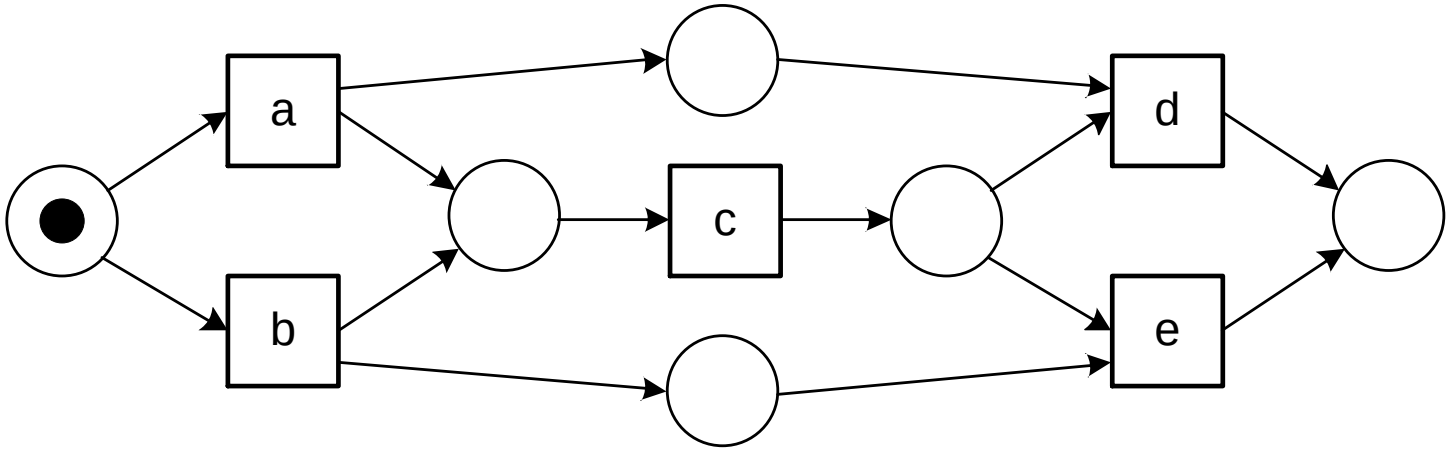
p = 3
c = 3
m = 3
r = 3

fitness = 0 !

TU/e

Another example

trace	frequency
acd	10
bce	10
ace	5
bcd	5
dca	1
abd	1
d	1

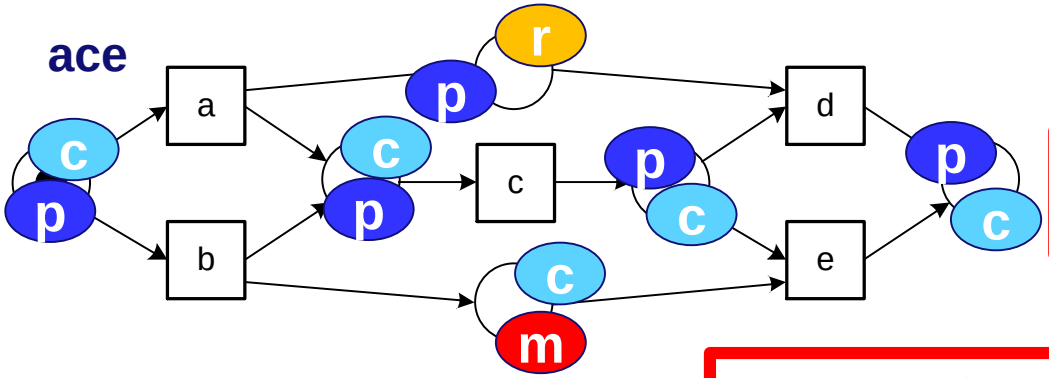


- **Consider the event log containing 33 cases.**
- **What is the fitness?**

TU/e

Fitness = 0.895705521

trace	frequency	produced tokens (p)	remaining tokens (r)	consumed tokens (c)	missing tokens (m)	produced tokens (all)	remaining tokens (all)	consumed tokens (all)	missing tokens (all)
acd	10	5	0	5	0	50	0	50	0
bce	10	5	0	5	0	50	0	50	0
ace	5	5	1	5	1	25	5	25	5
bcd	5	5	1	5	1	25	5	25	5
dca	1	5	3	5	3	5	3	5	3
abd	1	6	3	5	2	6	3	5	2
d	1	2	1	3	2	2	1	3	2



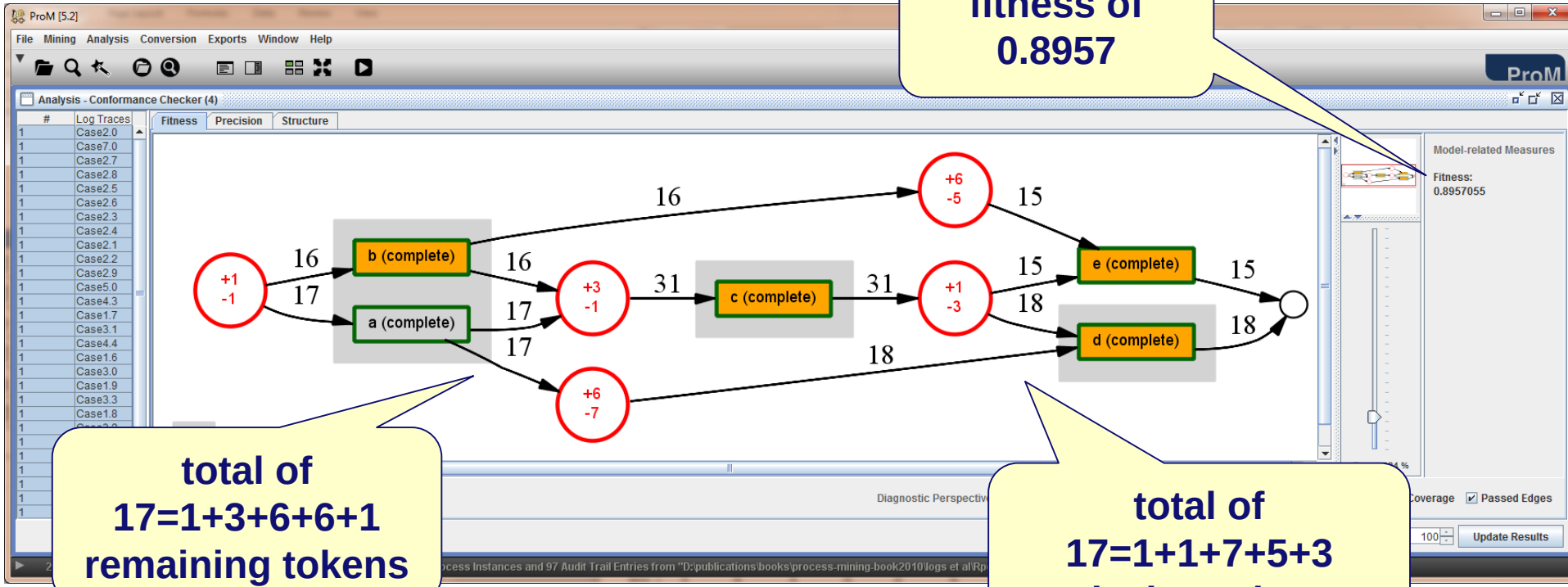
163	17	163	17
sum p	sum r	sum c	sum m

fitness **0.895705521**

$$fitness(L,N) = \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}}\right) + \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}}\right)$$

TU/e

ProM 5.2 diagnostics

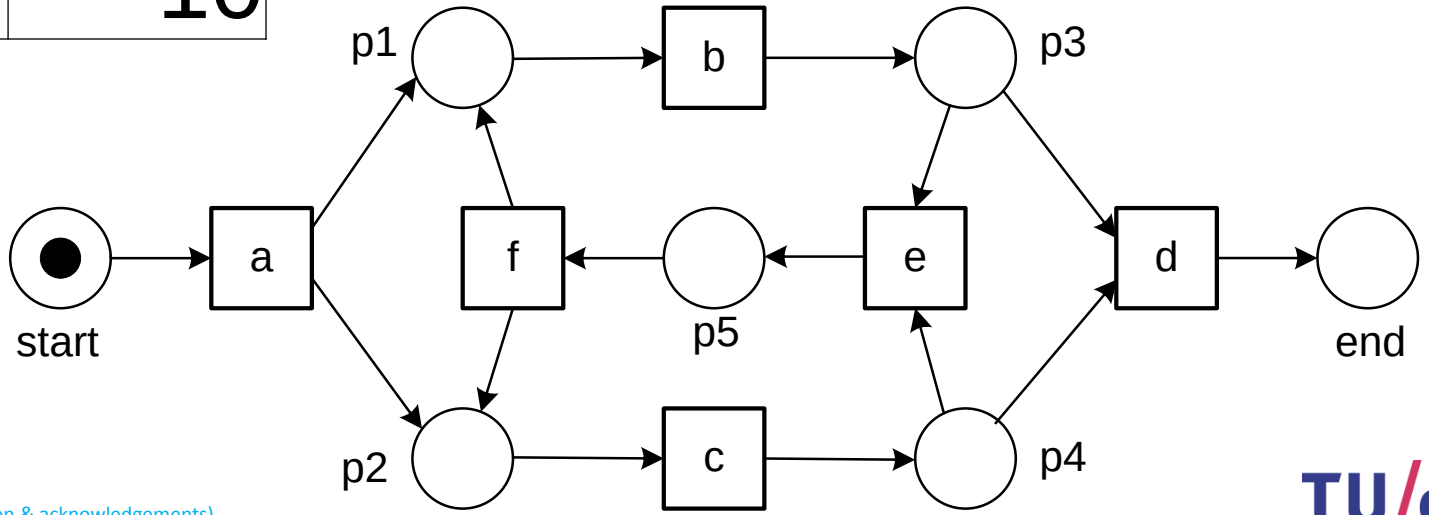


TU/e

Another example

trace	frequency
abefcd	10
abbefccd	10

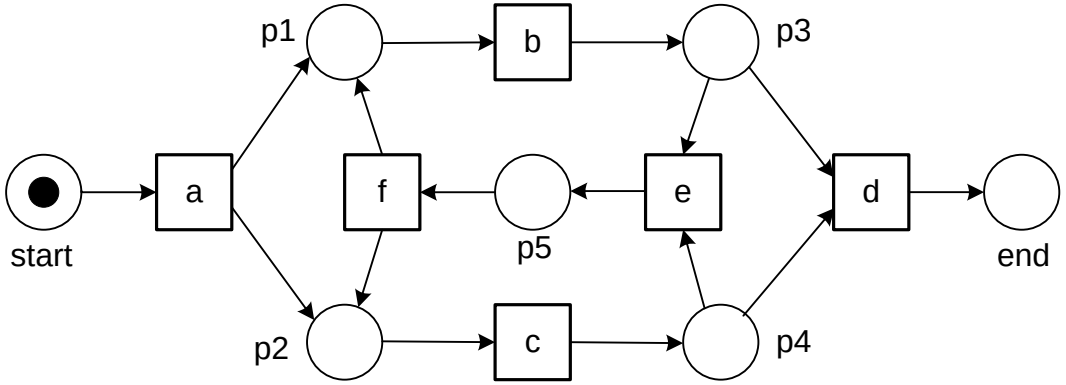
- **Consider the event log containing 20 cases.**
- **What is the fitness?**



TU/e

Fitness = 0.8

trace	frequency	produced tokens (p)	remaining tokens (r)	consumed tokens (c)	missing tokens (m)	produced tokens (all)	remaining tokens (all)	consumed tokens (all)	missing tokens (all)
abefcd	10	9	2	9	2	90	20	90	20
abbefccd	10	11	2	11	2	110	20	110	20



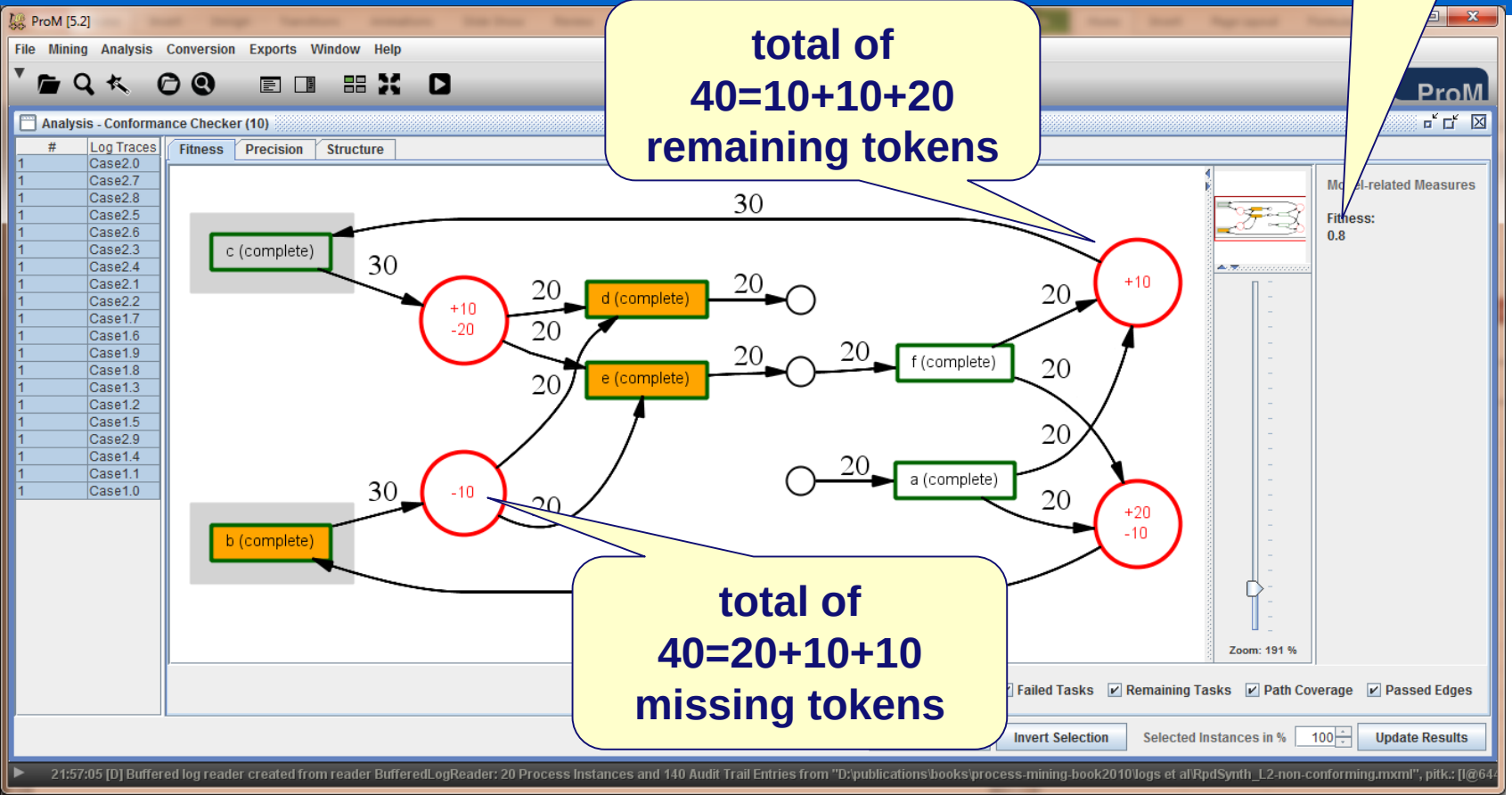
200	40	200	40
sum p	sum r	sum c	sum m

fitness **0.8**

$$fitness(L,N) = \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}}\right) + \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}}\right)$$

TU/e

ProM 5.2 diagnostics



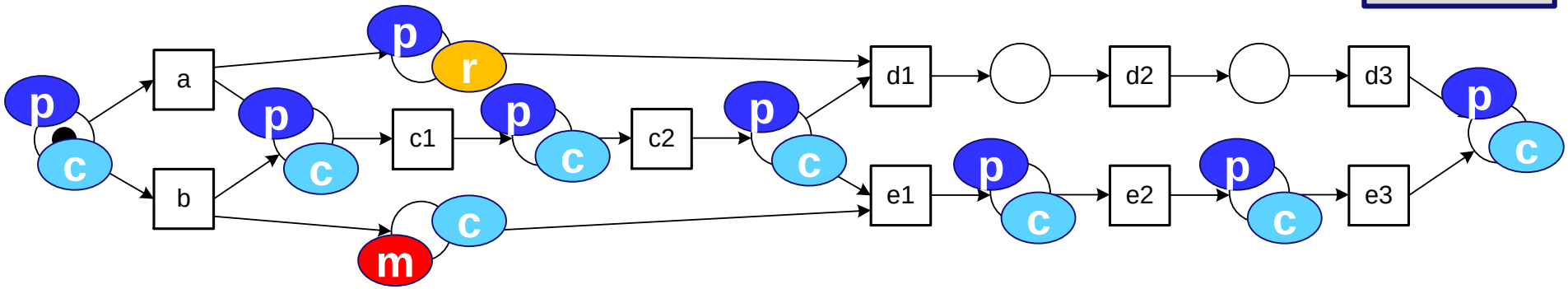
Limitations

- Basic replay approach assumes visible & uniquely labeled transitions.
- ProM implementation uses heuristics to deal with silent transitions and multiple transitions having the same label.
- Conformance values sometimes too optimistic due to "token flooding".
- Local decision making may lead to misleading results.

Local decision making is not enough …

p = 8
c = 8
m = 1
r = 1
f = 0.875

⟨a,c1,c2,e1,e2,e3⟩

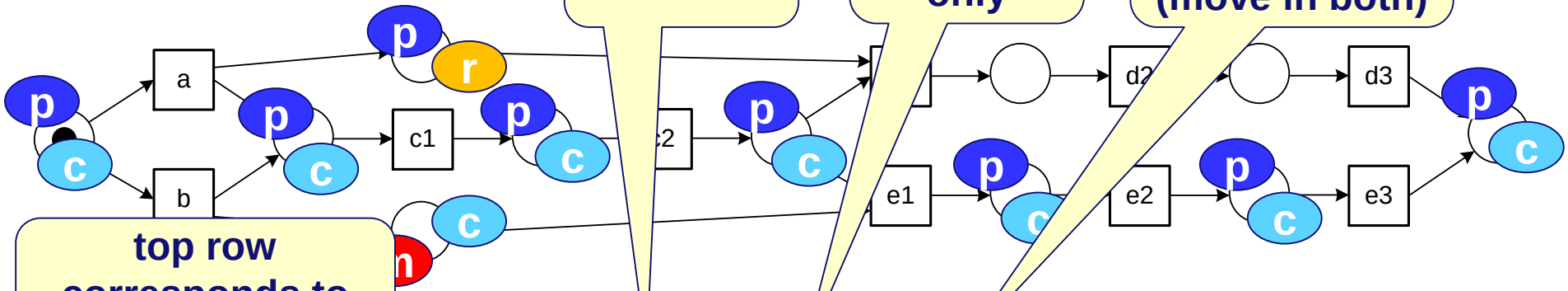


- Replay technique does not provide a corresponding path through the model (vital for conformance/performance analysis and other diagnostics).
- We would like to see the "closest path", i.e., ⟨b,c1,c2,e1,e2,e3⟩.

TU/e

Next: alignments

⟨a,c1,c2,e1,e2,e3⟩



move in log only

move in model only

synchronous move (move in both)

top row corresponds to trace in event log

bottom row corresponds to path in model

a	»	c1	c2	e1	e2	e3
»	b	c1	c2	e1	e2	e3

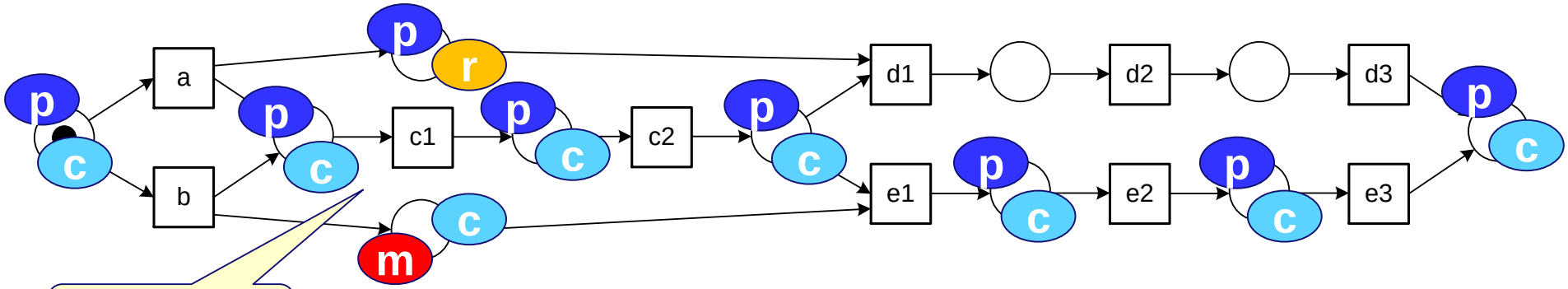
"no move"

"no move"

TU/e

Alignments provide better diagnostics

⟨a,c1,c2,e1,e2,e3⟩



???

b was replaced by a

a	»	c1	c2	e1	e2	e3
»	b	c1	c2	e1	e2	e3

TU/e

Part I: Preliminaries

- Chapter 1: Introduction
- Chapter 2: Process Modeling and Analysis
- Chapter 3: Data Mining

Part II: From Event Logs to Process Models

- Chapter 4: Getting the Data
- Chapter 5: Process Discovery: An Introduction
- Chapter 6: Advanced Process Discovery Techniques

Part III: Beyond Process Discovery

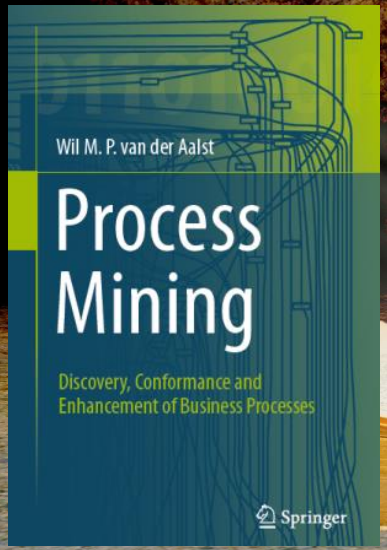
- Chapter 7: Conformance Checking
- Chapter 8: Mining Additional Perspectives
- Chapter 9: Operational Support

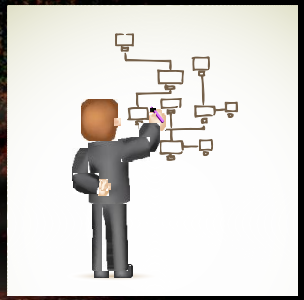
Part IV: Putting Process Mining to Work

- Chapter 10: Tool Support
- Chapter 11: Analyzing "Lasagna Processes"
- Chapter 12: Analyzing "Spaghetti Processes"

Part V: Reflection

- Chapter 13: Cartography and Navigation
- Chapter 14: Epilogue





TU/e