CS 4072 - Topics in CS
Process Mining

Lecture # 23

May 24, 2022

Spring 2022

FAST - NUCES, CFD Campus

Dr. Rabia Maqsood

rabia.maqsood@nu.edu.pk

Today’s Topics

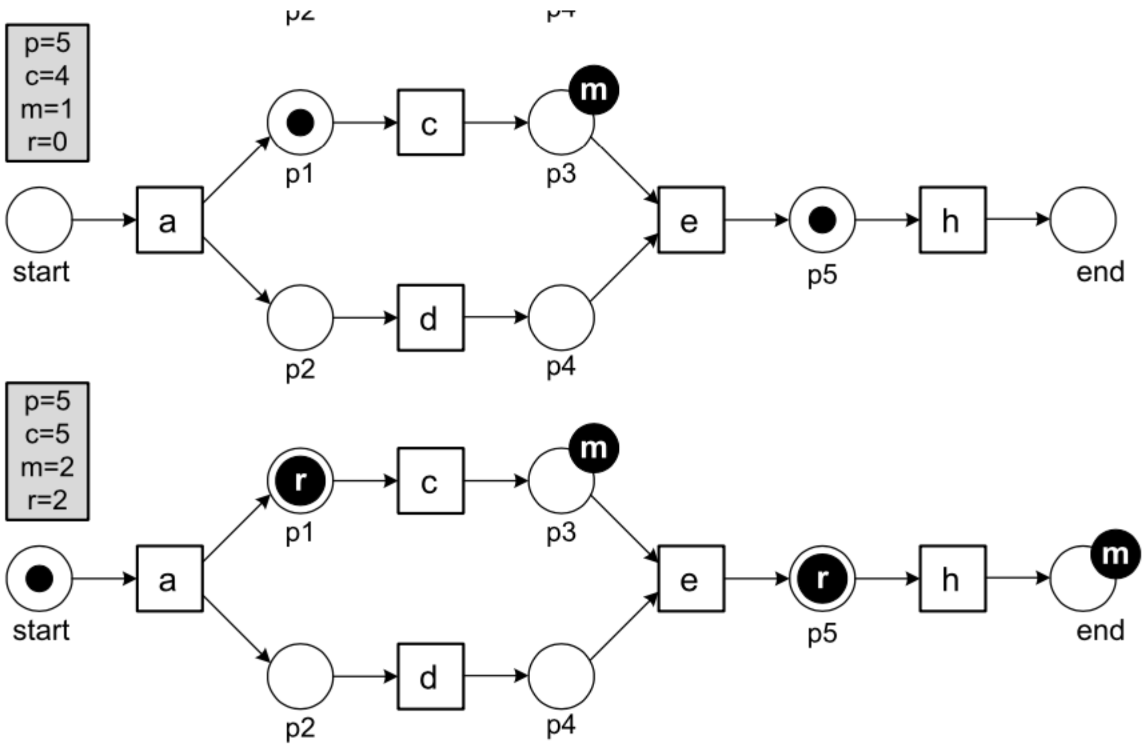
- Conformance Checking
  - Token-based replay: a quick recap
  - Sequence Alignment

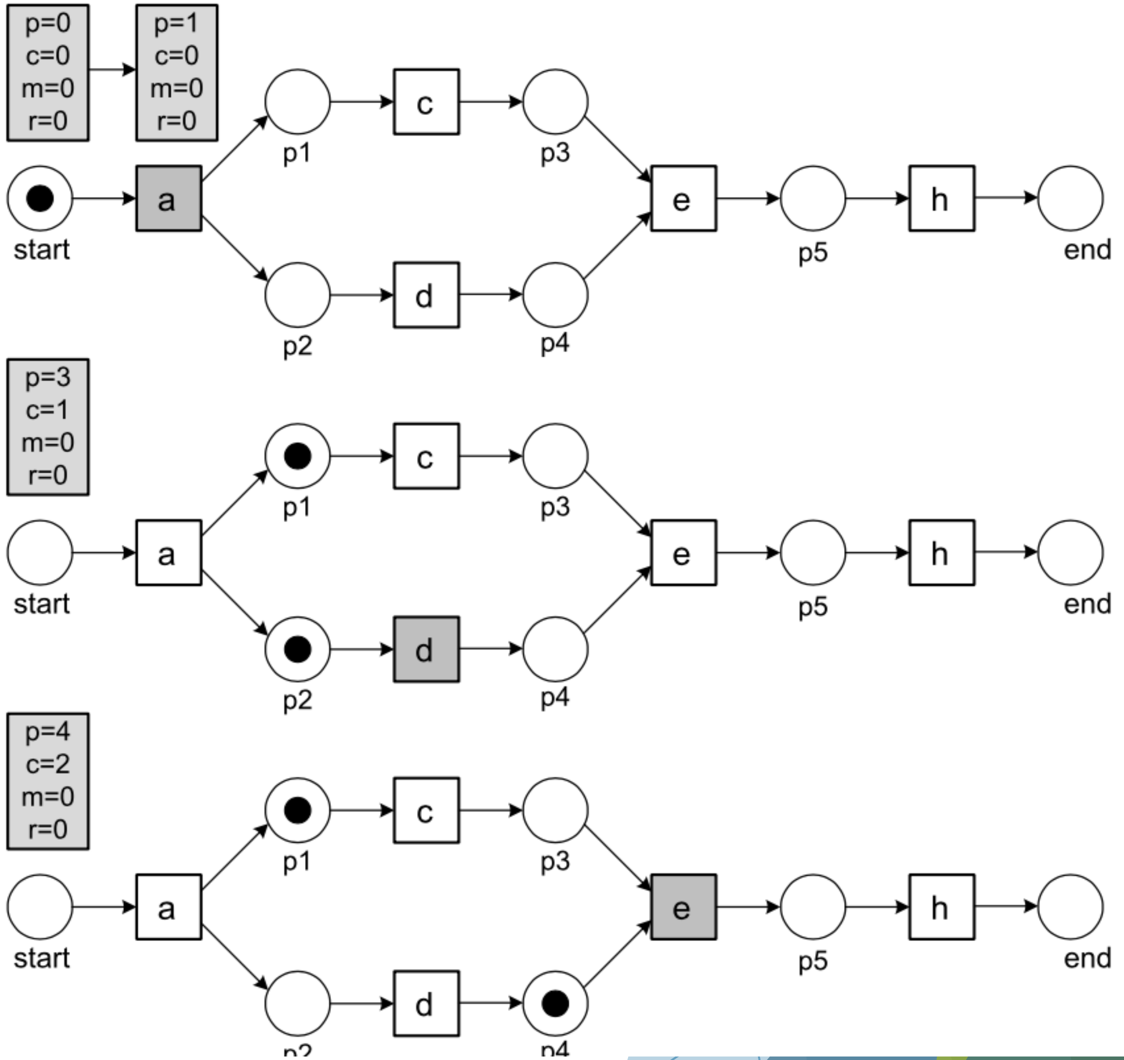
Approaches for Conformance Checking

Model and Log Fitness

Token-based replay: recap

Let the trace $\sigma_2' = \langle a,d,e \rangle$
and, WF-net N_3:



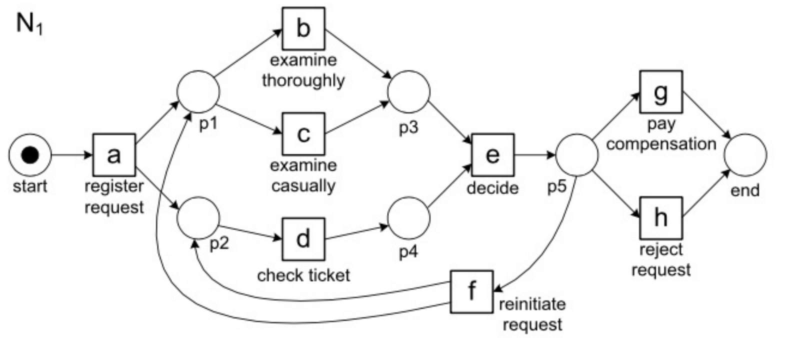


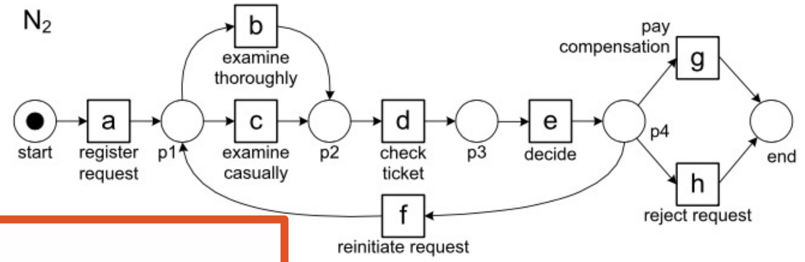
$$fitness(\sigma_2, N_3) = \frac{1}{2}\left(1 - \frac{2}{5}\right) + \frac{1}{2}\left(1 - \frac{2}{5}\right) = 0.6$$

Token-based replay: recap

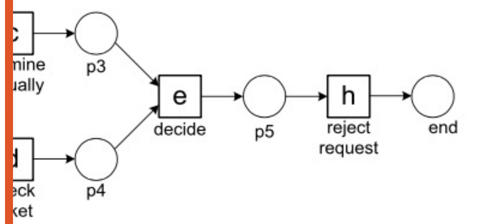
Frequency	Reference	Trace
455	σ_1	$\langle a, c, d, e, h \rangle$
191	σ_2	$\langle a, b, d, e, g \rangle$
177	σ_3	$\langle a, d, c, e, h \rangle$
144	σ_4	$\langle a, b, d, e, h \rangle$
111	σ_5	$\langle a, c, d, e, g \rangle$
82	σ_6	$\langle a, d, c, e, g \rangle$
56	σ_7	$\langle a, d, b, e, h \rangle$
47	σ_8	$\langle a, c, d, e, f, d, b, e, h \rangle$
38	σ_9	$\langle a, d, b, e, g \rangle$
33	σ_{10}	$\langle a, c$
14	σ_{11}	$\langle a, c$
11	σ_{12}	$\langle a, c$
9	σ_{13}	$\langle a, d$
8	σ_{14}	$\langle a, d$
5	σ_{15}	$\langle a, d$
3	σ_{16}	$\langle a, c$
2	σ_{17}	$\langle a, d$
2	σ_{18}	$\langle a, d$
1	σ_{19}	$\langle a, d$
1	σ_{20}	$\langle a, d$
1	σ_{21}	$\langle a, d$

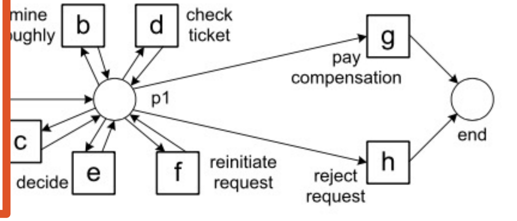
?





$$fitness(L_{full}, N_1) = 1$$
$$fitness(L_{full}, N_2) = 0.9504$$
$$fitness(L_{full}, N_3) = 0.8797$$
$$fitness(L_{full}, N_4) = 1$$





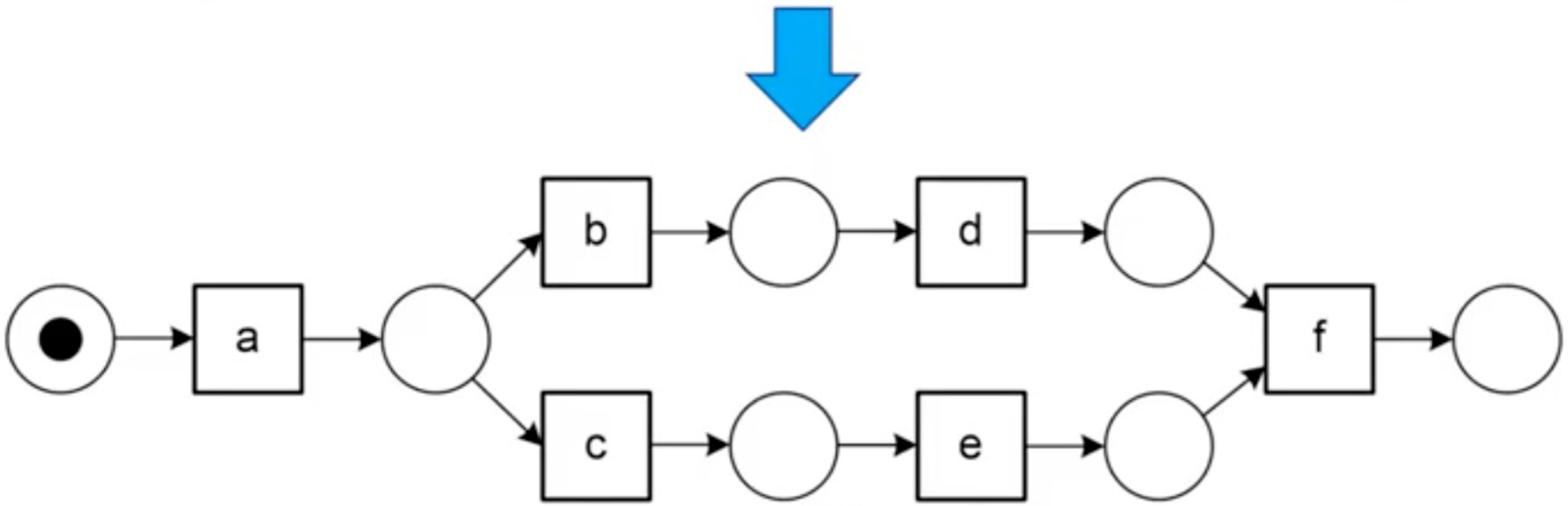
Question

- Consider the model generated by the Alpha algorithm
- Compute fitness using missing and remaining tokens (a.k.a. token-based replay)
- Share your findings

$$L = [\langle a, b, d, e, f \rangle^{10}, \langle a, c, e, d, f \rangle^{10}]$$

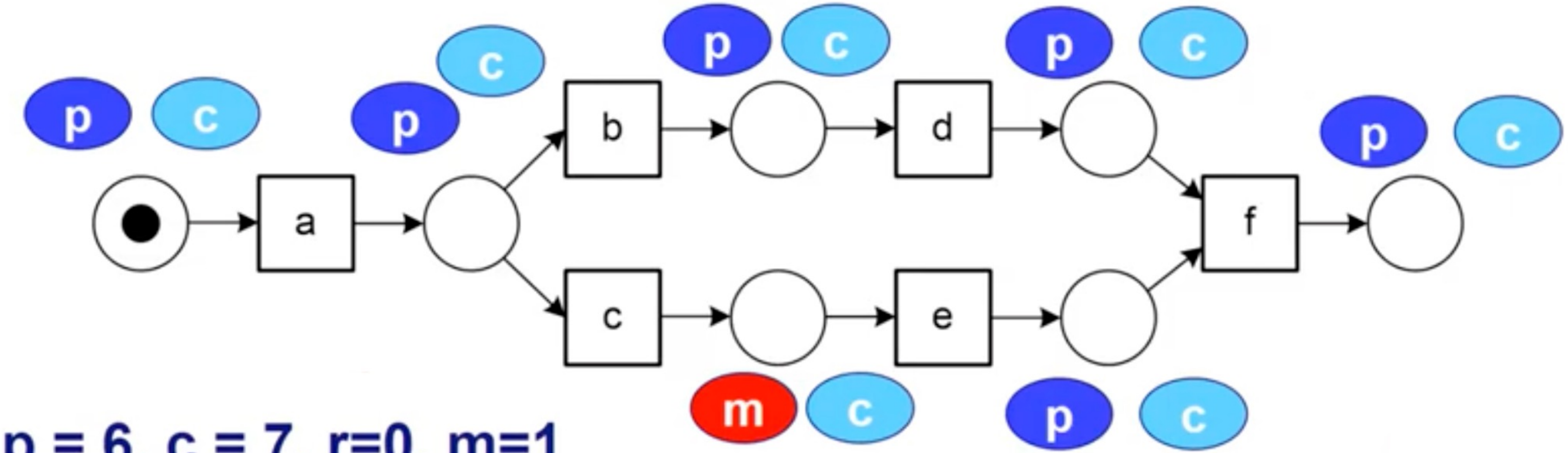
Model generated by Alpha algorithm

$$L = [\langle a, b, d, e, f\rangle^{10}, \langle a, c, e, d, f\rangle^{10}]$$



Token-based replay

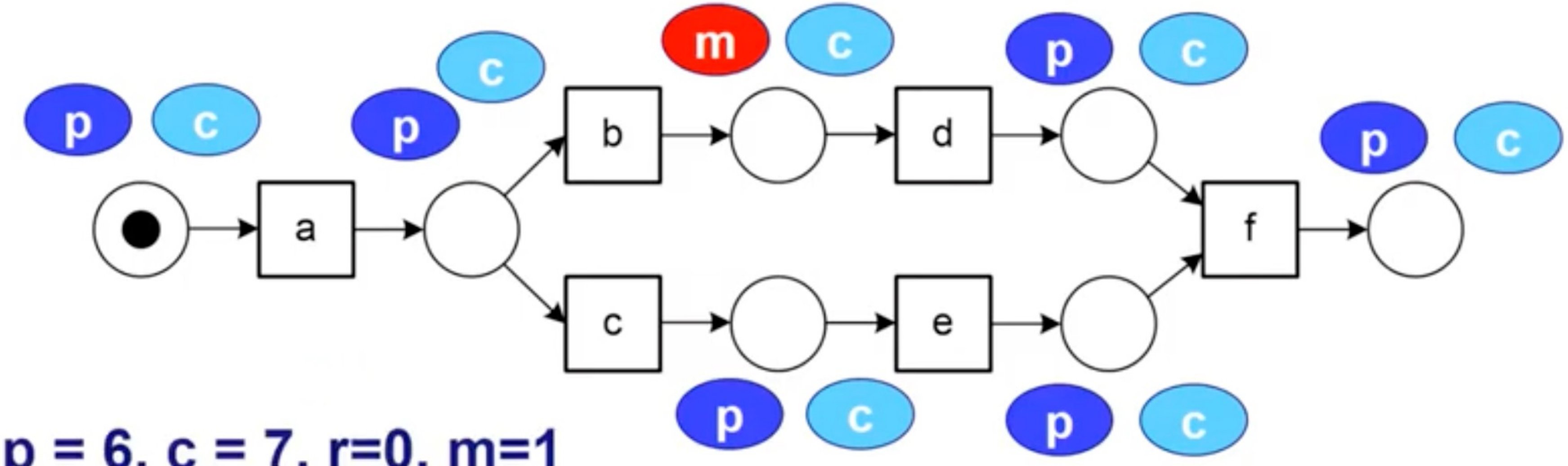
$$L = [\boxed{\langle a, b, d, e, f \rangle}^{10}, \langle a, c, e, d, f \rangle^{10}]$$



p = 6, c = 7, r=0, m=1

Token-based replay

$$L = [\langle a, b, d, e, f\rangle^{10}, \boxed{\langle a, c, e, d, f\rangle}^{10}]$$

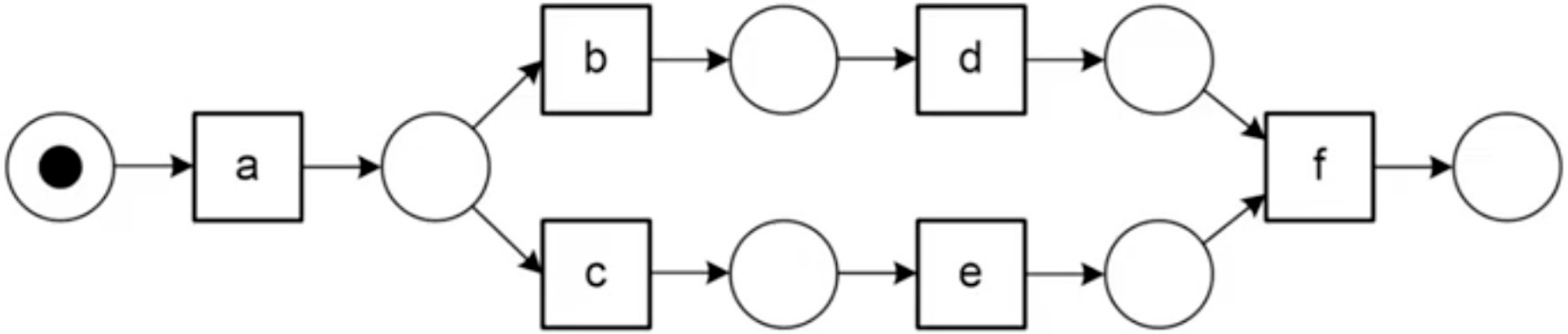


p = 6, c = 7, r=0, m=1

Overall log fitness

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

$$L = [\langle a,b,d,e,f\rangle^{10}, \langle a,c,e,d,f\rangle^{10}]$$

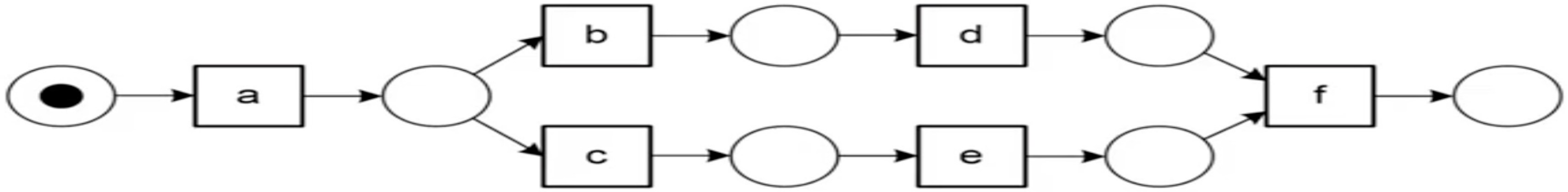


p = 2*10*6 =120, c = 2*10*7=140, r=2*10*0=0, m=2*10*1=20

$$\frac{1}{2}\left(1-\frac{20}{140}\right)+\frac{1}{2}\left(1-\frac{0}{120}\right)=\frac{13}{14}\approx 0.93$$

Findings

$$L = [\langle a, b, d, e, f\rangle^{10}, \langle a, c, e, d, f\rangle^{10}]$$



p = 2*10*6 =120, c = 2*10*7=140, r=2*10*0=0, m=2*10*1=20

$$\frac{1}{2}\left(1-\frac{20}{140}\right)+\frac{1}{2}\left(1-\frac{0}{120}\right)=\frac{13}{14}\approx 0.93$$

- The **model is not sound!**
  - In fact there is no firing sequence leading to the target marking.
  - Difficult to interpret the conformance results for an unsound model. Hence, we need a 'relaxed notion of soundness'.

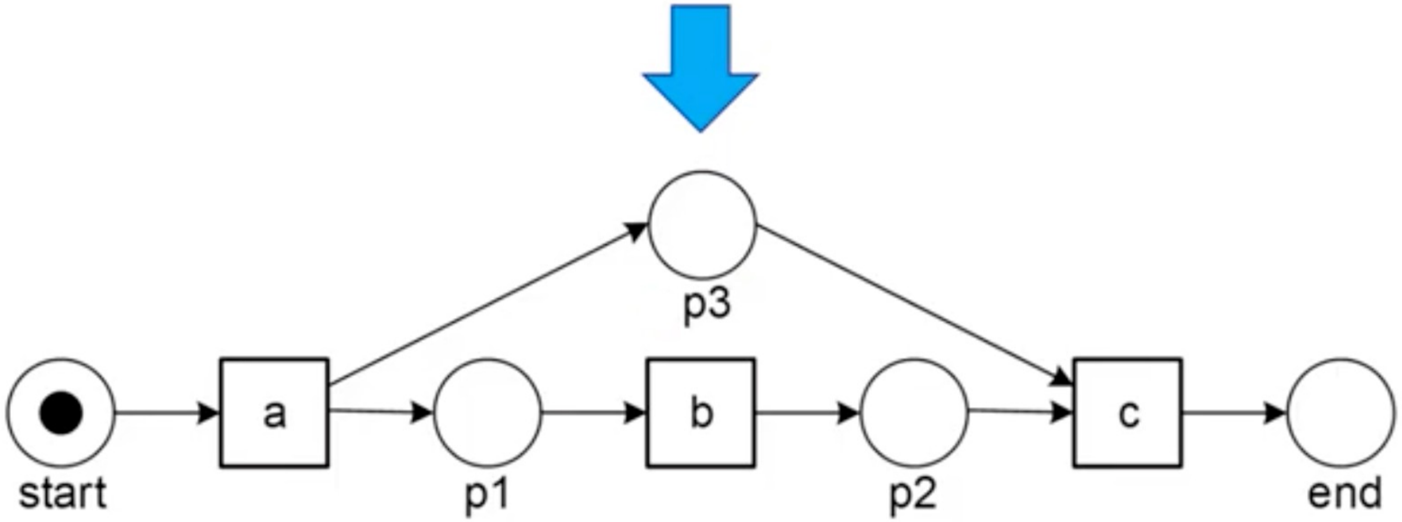
Question

- Consider the model generated by the Alpha algorithm
- Compute fitness using missing and remaining tokens (a.k.a. token-based replay)
- Share your findings

$$L_{11} = [\langle a,b,c\rangle^{20}, \langle a,c\rangle^{30}]$$

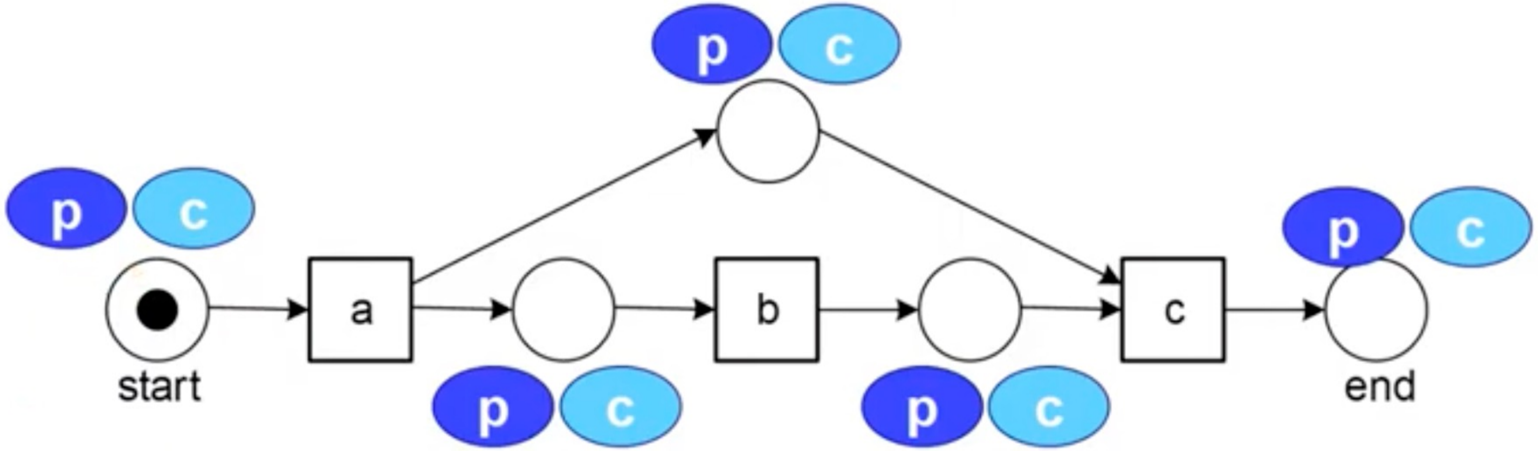
Model generated by Alpha algorithm

$$L_{11} = [\langle a,b,c\rangle^{20}, \langle a,c\rangle^{30}]$$



Token-based replay

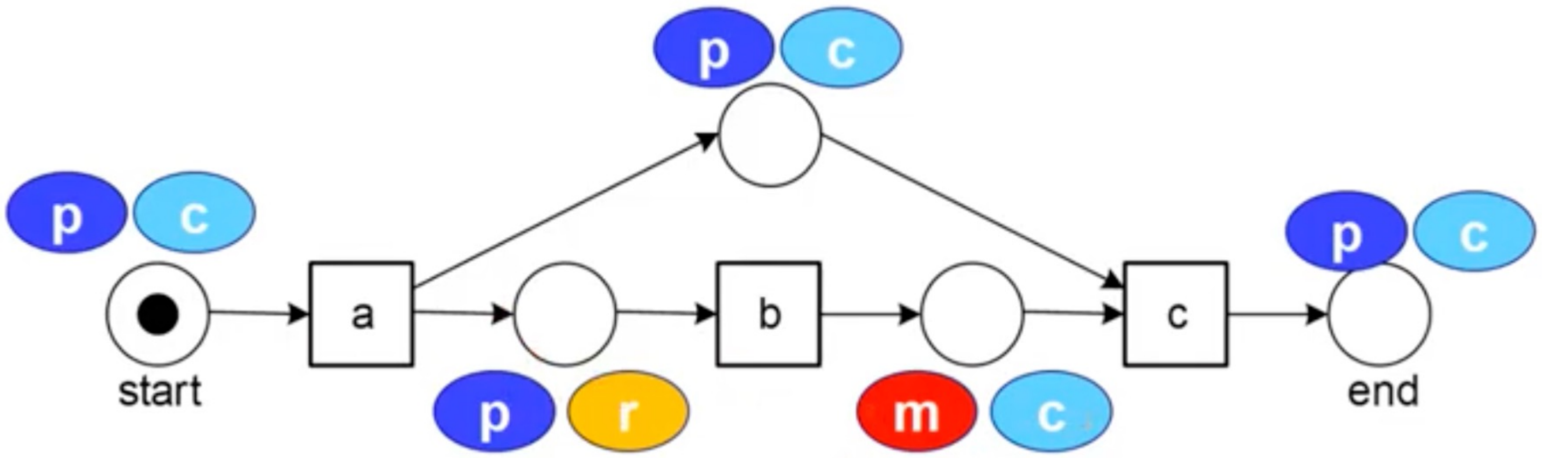
$$L_{11} = [\boxed{\langle a,b,c\rangle}^{20}, \langle a,c\rangle^{30}]$$



p = 5, c = 5, r=0, m=0

Token-based replay

$$L_{11} = [\langle a,b,c\rangle^{20}, \boxed{\langle a,c\rangle}^{30}]$$



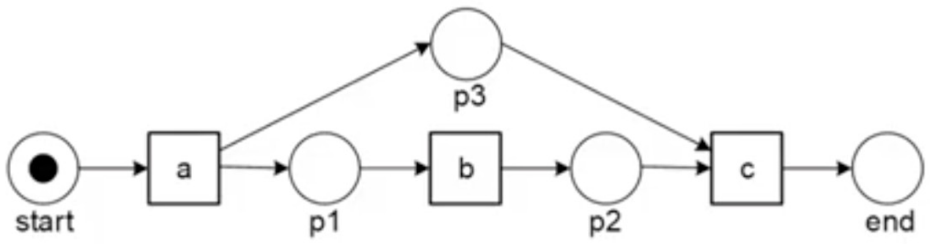
p = 4, c = 4, r=1, m=1

Overall log fitness

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

$$L_{11} = [\langle a,b,c\rangle^{20}, \langle a,c\rangle^{30}]$$

p = 5, c = 5, r=0, m=0 p = 4, c = 4, r=1, m=1



p = 20*5+30*4 =220, c = 20*5+30*4=220,
r = 20*0+30*1=30, m=20*0+30*1=30

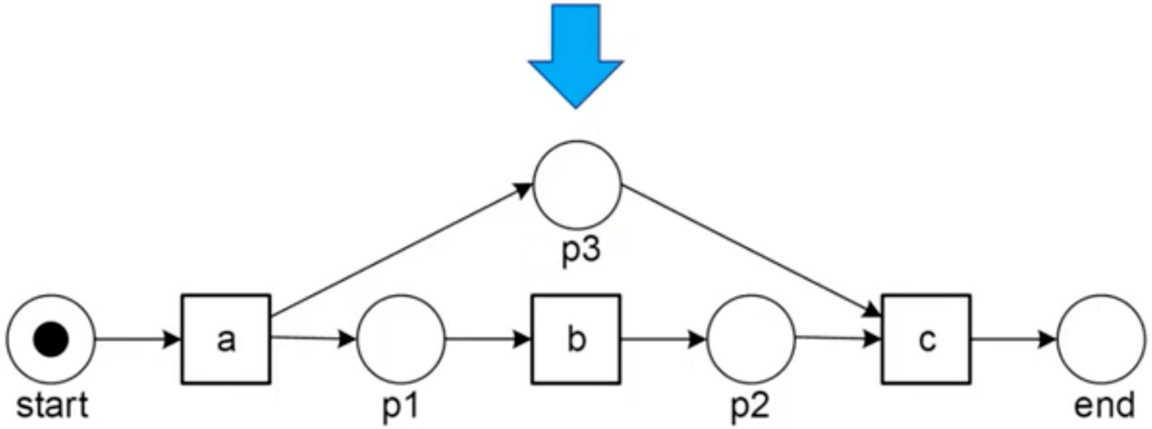
$$\frac{1}{2}\left(1-\frac{30}{220}\right)+\frac{1}{2}\left(1-\frac{30}{220}\right)=\frac{19}{22}\approx 0.86$$

Our findings: the second trace cannot be replayed as it has missing and remaining tokens

Redundant places impact on fitness

- Does the redundant places impact the fitness of a log?

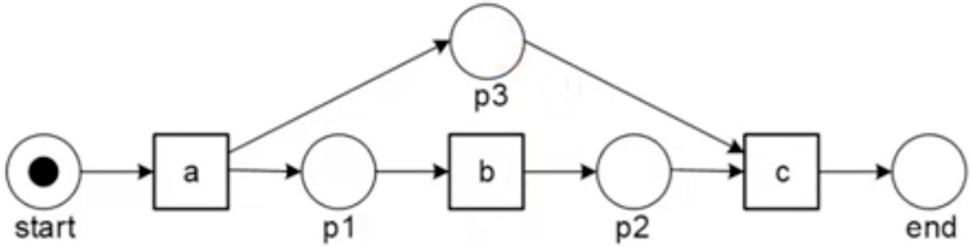
$$L_{11} = [\langle a, b, c\rangle^{20}, \langle a, c\rangle^{30}]$$



Redundant places impact on fitness

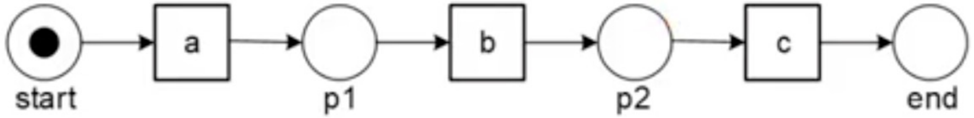
Redundant places increase the fitness of a log

$$L_{11} = [\langle a,b,c\rangle^{20}, \langle a,c\rangle^{30}]$$



$$\frac{1}{2}\left(1-\frac{30}{220}\right)+\frac{1}{2}\left(1-\frac{30}{220}\right)=\frac{19}{22}\approx 0.86$$

$$L_{11} = [\langle a,b,c\rangle^{20}, \langle a,c\rangle^{30}]$$

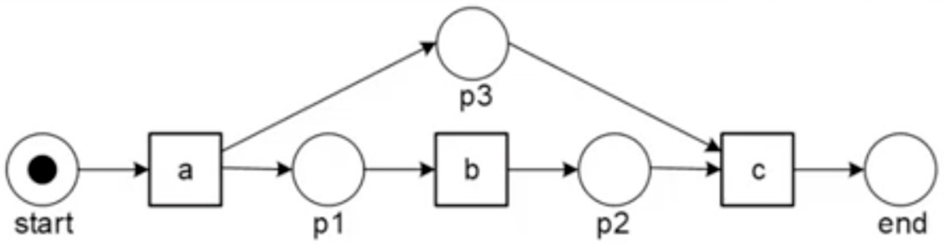


$$\frac{1}{2}\left(1-\frac{30}{170}\right)+\frac{1}{2}\left(1-\frac{30}{170}\right)=\frac{14}{17}\approx 0.82$$

Redundant places impact on fitness

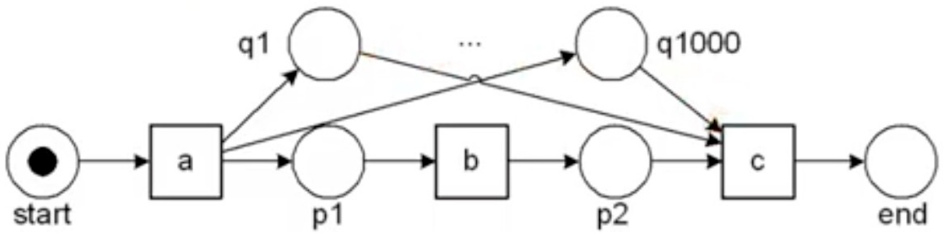
This can go very high even if we add many redundant places

$$L_{11} = [\langle a,b,c\rangle^{20}, \langle a,c\rangle^{30}]$$



$$\frac{1}{2}\left(1-\frac{30}{220}\right)+\frac{1}{2}\left(1-\frac{30}{220}\right)=\frac{19}{22}\approx 0.86$$

$$L_{11} = [\langle a,b,c\rangle^{20}, \langle a,c\rangle^{30}]$$



$$\frac{1}{2}\left(1-\frac{30}{50170}\right)+\frac{1}{2}\left(1-\frac{30}{50170}\right)=\frac{5014}{5017}\approx 0.999$$

Aligning the model and observed behavior

Requirements for an ideal conformance checking

- Conformance checking should not impose restrictions on the process notation (e.g., silent transitions and duplicate transitions should be possible).
- Two semantically equivalent models should have the same conformance value.
- Should provide a “closest matching path” through the process model for any trace in the event log.
  - Also required for performance analysis!
  - Beyond the analysis of replay fitness (advanced diagnostics, precision, generalization, etc.)

Alignments

- **Alignments** were introduced to overcome the limitations of token-based replay.

- The objective is to find the ***optimal*** sequence alignment between two traces.

Alignments

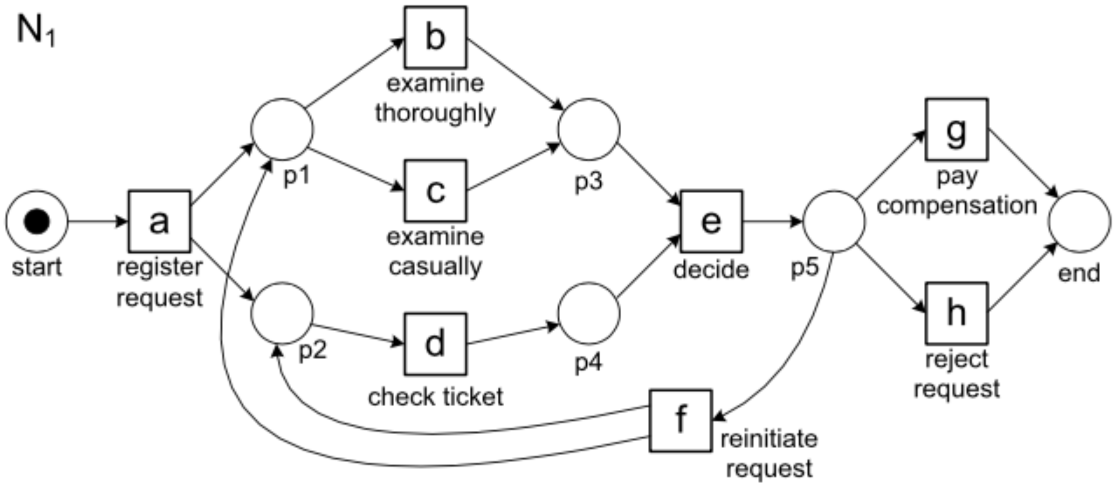
- Consider a trace: σ = ⟨a, d, b, e, h⟩ and model $\mathbf{N_1}$

$$\gamma_1 = \begin{array}{|c|c|c|c|c|} a & d & b & e & h \\ \hline a & d & b & e & h \end{array}$$

Trace log

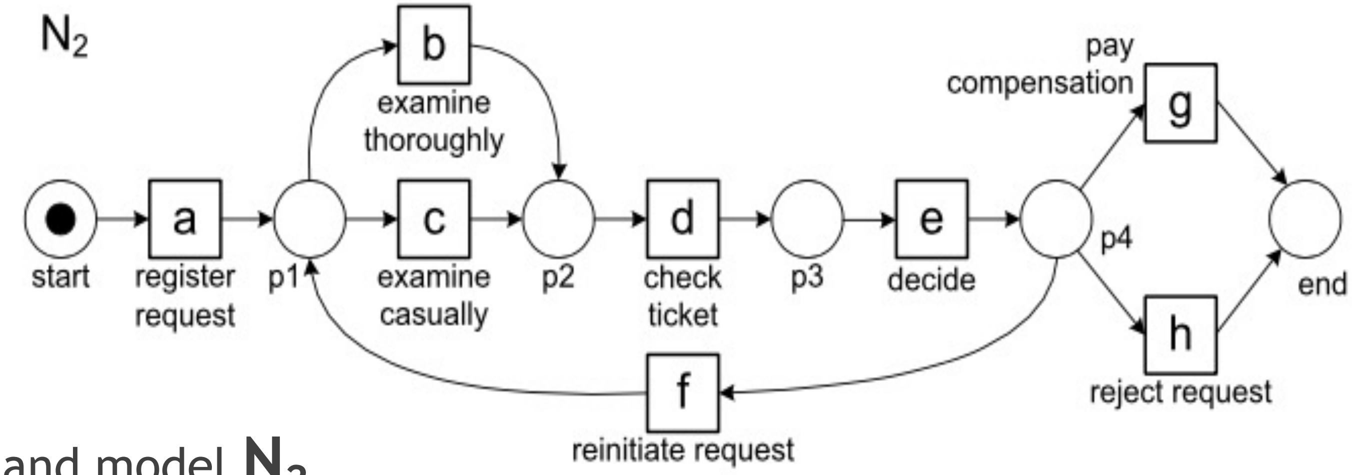
Model path
from initial to end

A possible alignment



All activities in the trace and model **match** perfectly

Alignments



- Consider a trace: $\sigma = \langle a, d, b, e, h \rangle$ and model $\mathbf{N_2}$

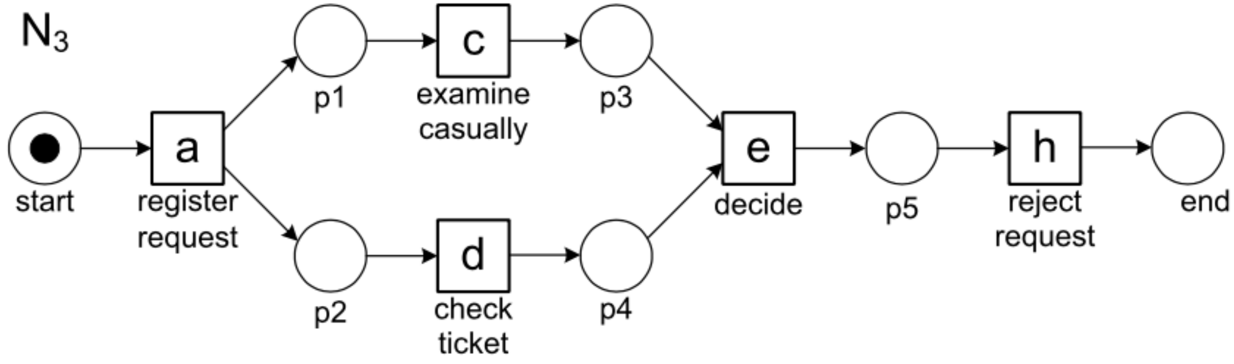
- Followings are the possible alignments

$$\gamma_{2a} = \begin{array}{|c|c|c|c|c|c|} a & \gg & d & b & e & h \\ \hline a & b & d & \gg & e & h \end{array} \qquad \gamma_{2b} = \begin{array}{|c|c|c|c|c|c|} a & \gg & d & b & e & h \\ \hline a & c & d & \gg & e & h \end{array} \qquad \gamma_{2c} = \begin{array}{|c|c|c|c|c|c|} a & d & b & \gg & e & h \\ \hline a & \gg & b & d & e & h \end{array}$$

Matched

Misalignment

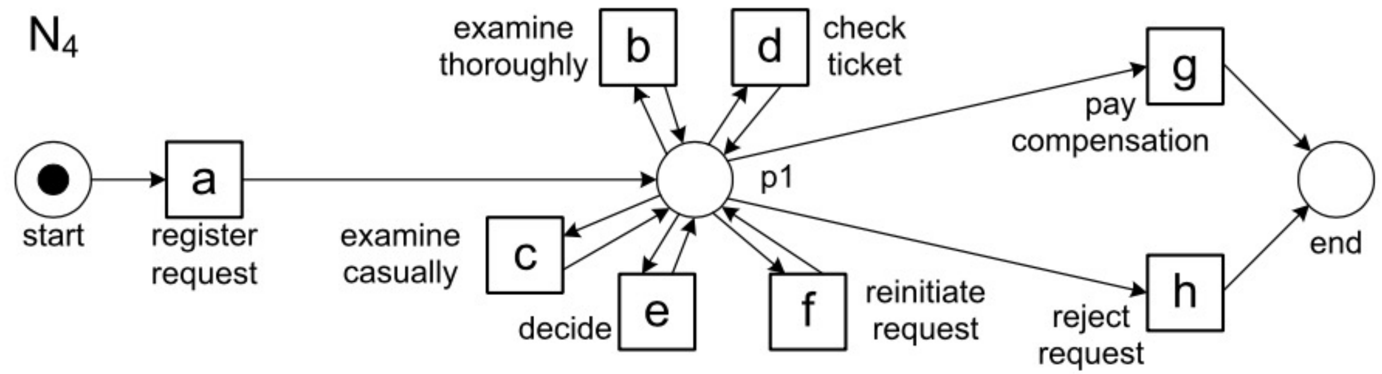
Alignments



- Consider a trace: $\sigma = \langle a, d, b, e, h \rangle$ and model $\mathbf{N_3}$
- Followings are the possible alignments

$$\gamma_{3a} = \begin{array}{|c|c|c|c|c|c|} a & \gg & d & b & e & h \\ \hline a & c & d & \gg & e & h \end{array} \qquad \gamma_{3b} = \begin{array}{|c|c|c|c|c|c|} a & d & \gg & b & e & h \\ \hline a & d & c & \gg & e & h \end{array} \qquad \gamma_{3c} = \begin{array}{|c|c|c|c|c|c|} a & d & b & \gg & e & h \\ \hline a & d & \gg & c & e & h \end{array}$$

Alignments



- Consider a trace: $\sigma = \langle a, d, b, e, h \rangle$ and model $\mathbf{N_4}$

- Following is the only possible alignment

$$\gamma_4 = \begin{array}{|c|c|c|c|c|} a & d & b & e & h \\ \hline a & d & b & e & h \end{array}$$

To be continued…

Reading Material

- Chapter 8: Aalst