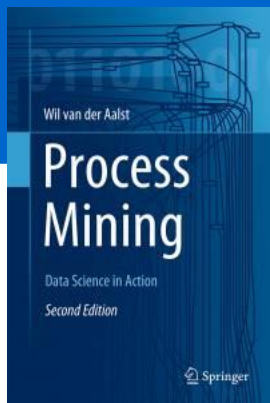


Process Mining: Data Science in Action

Alpha Algorithm: A Process Discovery Algorithm

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

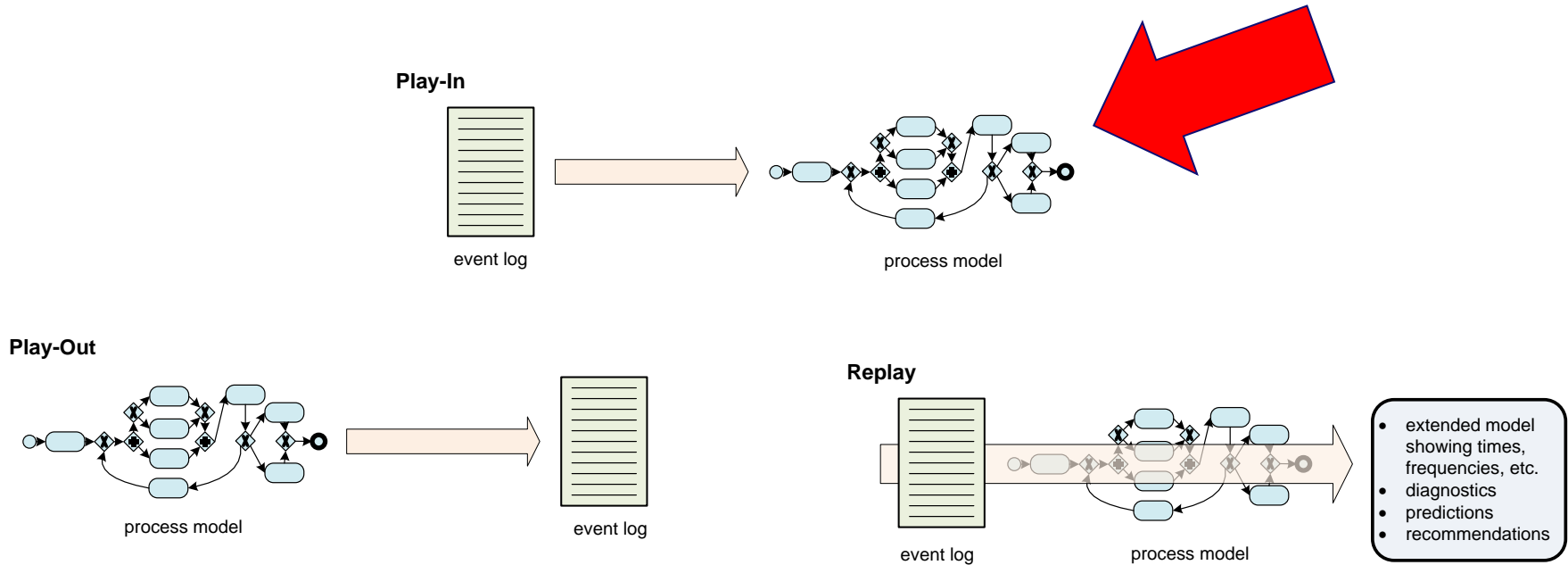


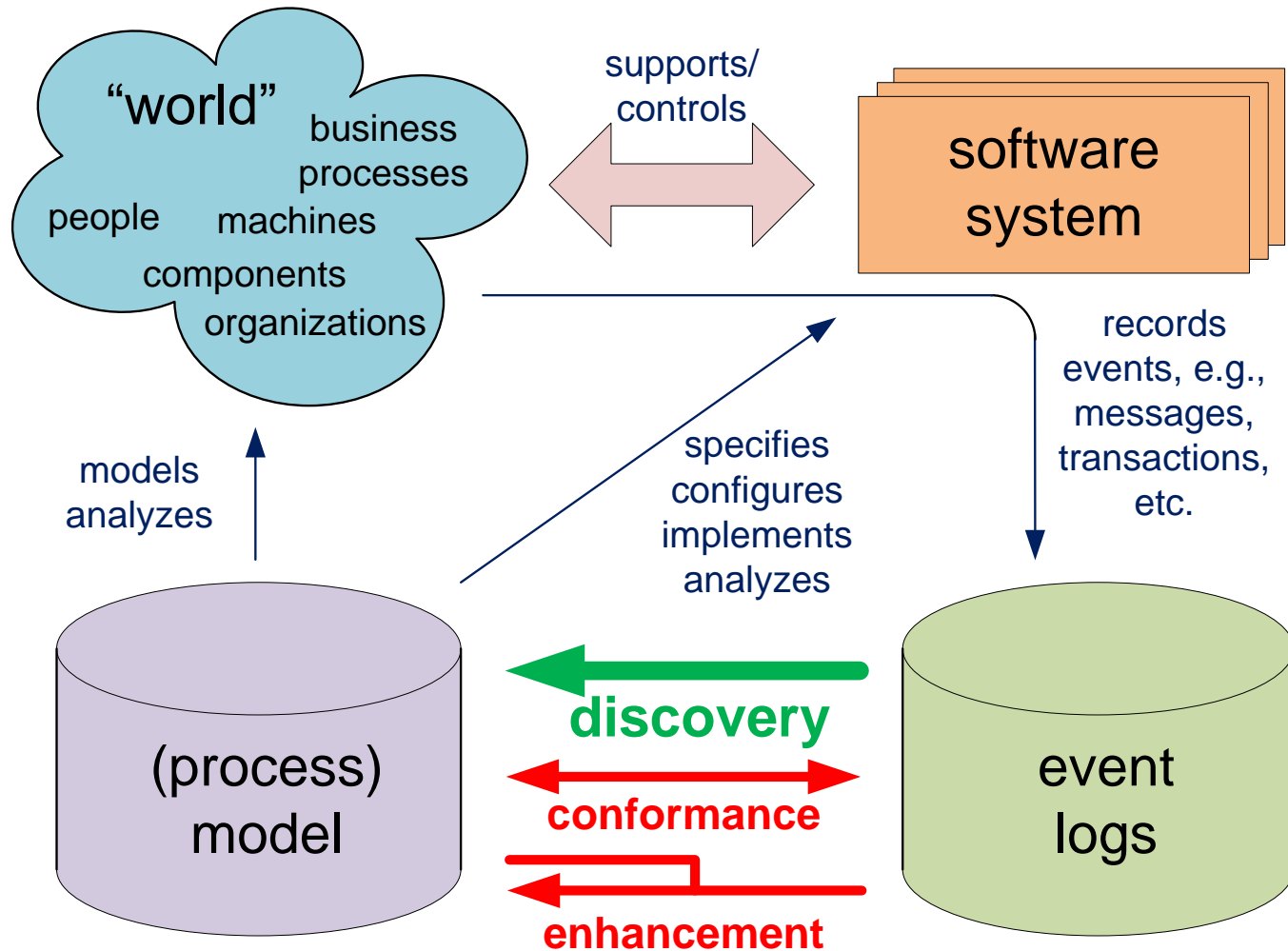
TU/e

Technische Universiteit
Eindhoven
University of Technology

Where innovation starts

Process discovery = Play-In







Simplifying event logs when focusing on control-flow

order number	activity	timestamp	user	product	quantity
9901	register order	22-1-2014@09.15	Sara Jones	iPhone5S	1
9902	register order	22-1-2014@09.18	Sara Jones	iPhone5S	2
9903	register order	22-1-2014@09.27	Sara Jones	iPhone4S	1
9901	check stock	22-1-2014@09.49	Pete Scott	iPhone5S	1
9901	ship order	22-1-2014@10.11	Sue Fox	iPhone5S	1
9903	check stock	22-1-2014@10.34	Pete Scott	iPhone4S	1
9901	handle payment	22-1-2014@10.41	Carol Hope	iPhone5S	1
9902	check stock	22-1-2014@10.57	Pete Scott	iPhone5S	2

[<register_order, check_stock, ship_order, handle_payment>,
<register_order, check_stock, cancel_order>,
<register_order, check_stock> , ...]

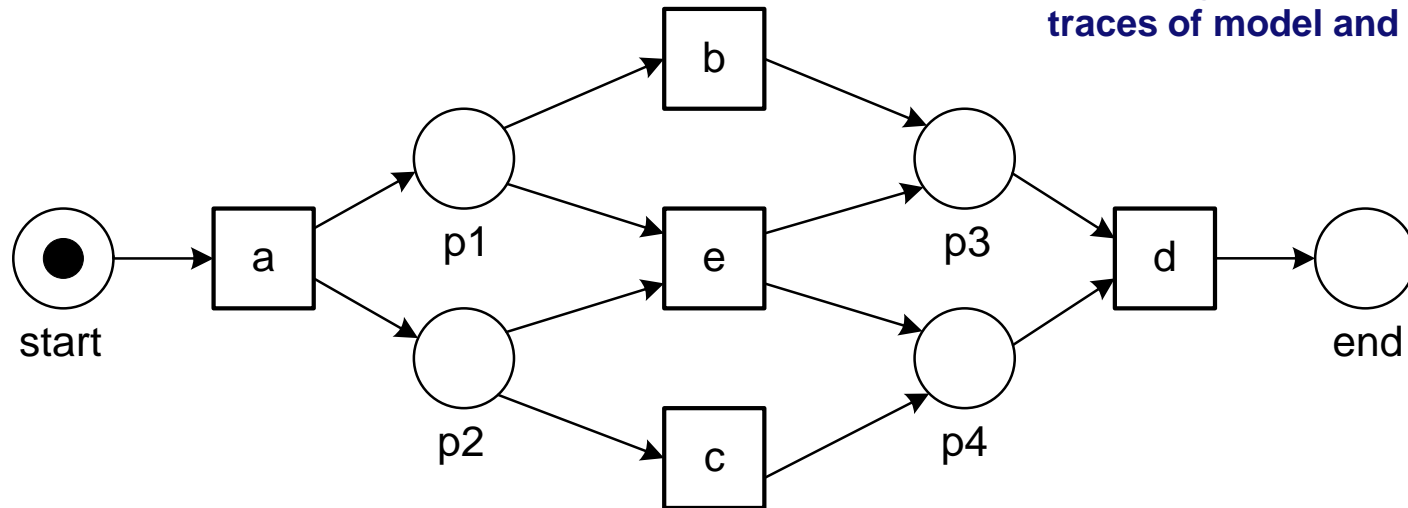
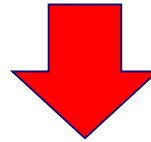
Simple event log

$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

- An **event log** is a **multiset of traces** (same trace may appear multiple times).
- A **trace** is a sequence of activity names (we abstract from all other attributes, but events are ordered).

Goal of Alpha algorithm

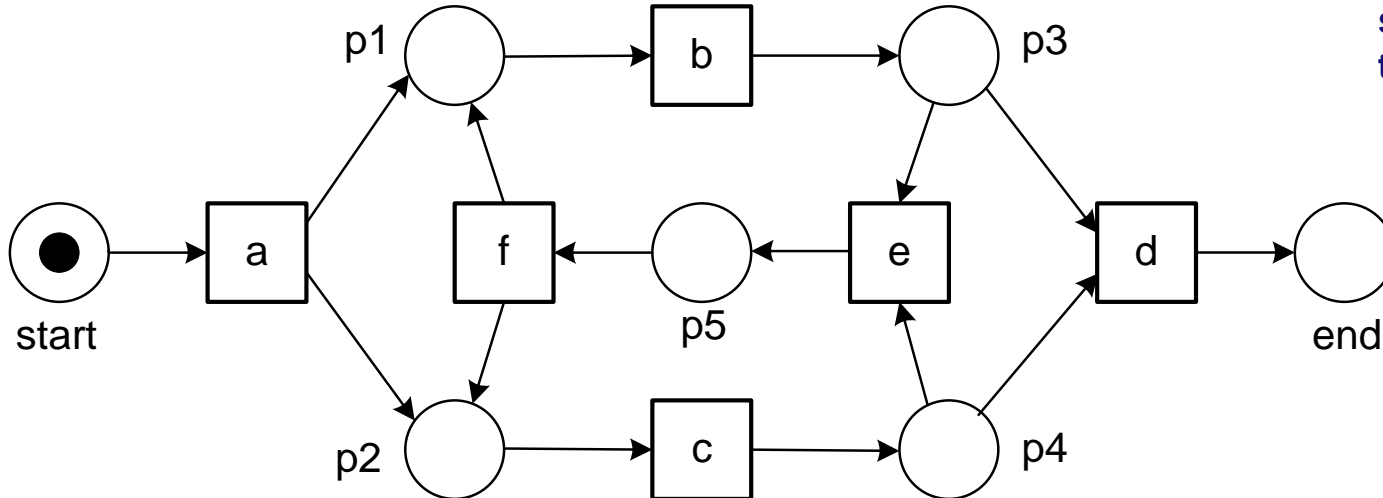
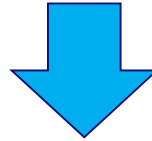
$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$



Event log contains all possible traces of model and vice versa.

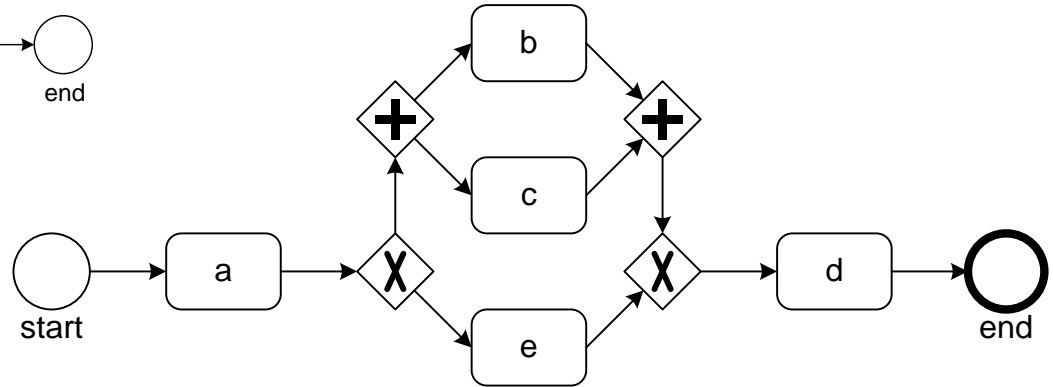
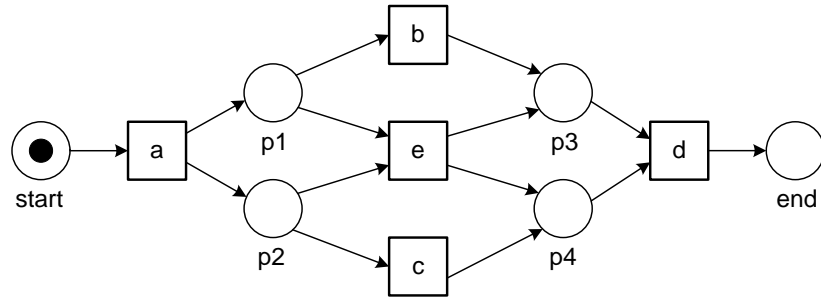
Another example

$$L_2 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^4, \langle a, b, c, e, f, b, c, d \rangle^2, \langle a, b, c, e, f, c, b, d \rangle, \\ \langle a, c, b, e, f, b, c, d \rangle^2, \langle a, c, b, e, f, b, c, e, f, c, b, d \rangle]$$



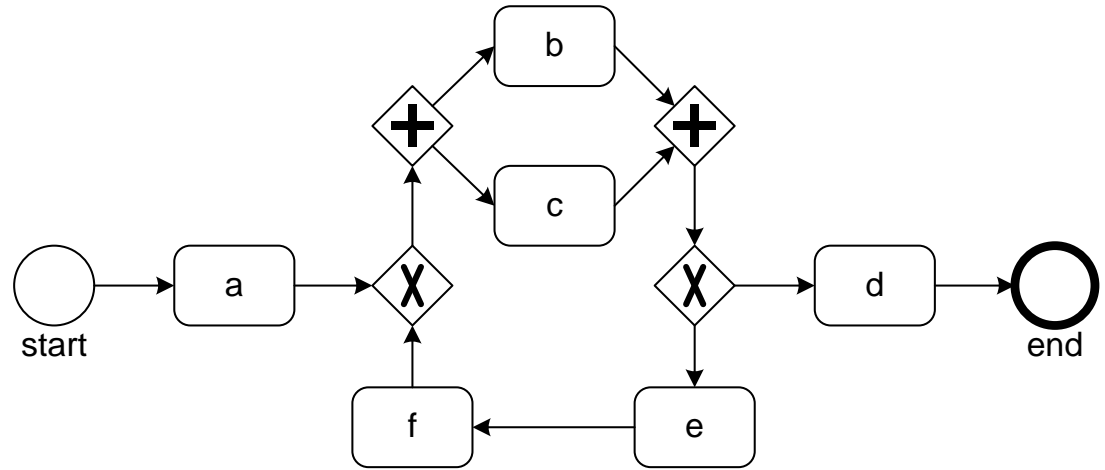
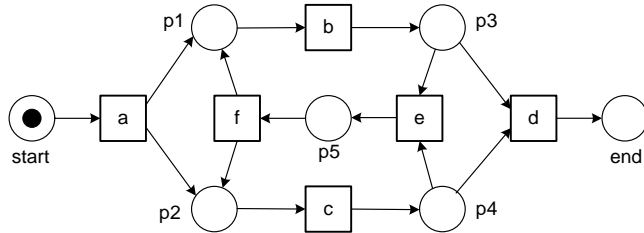
Generalization: event log contains only subset of all possible traces of model.

Notation is less relevant (e.g. BPMN)



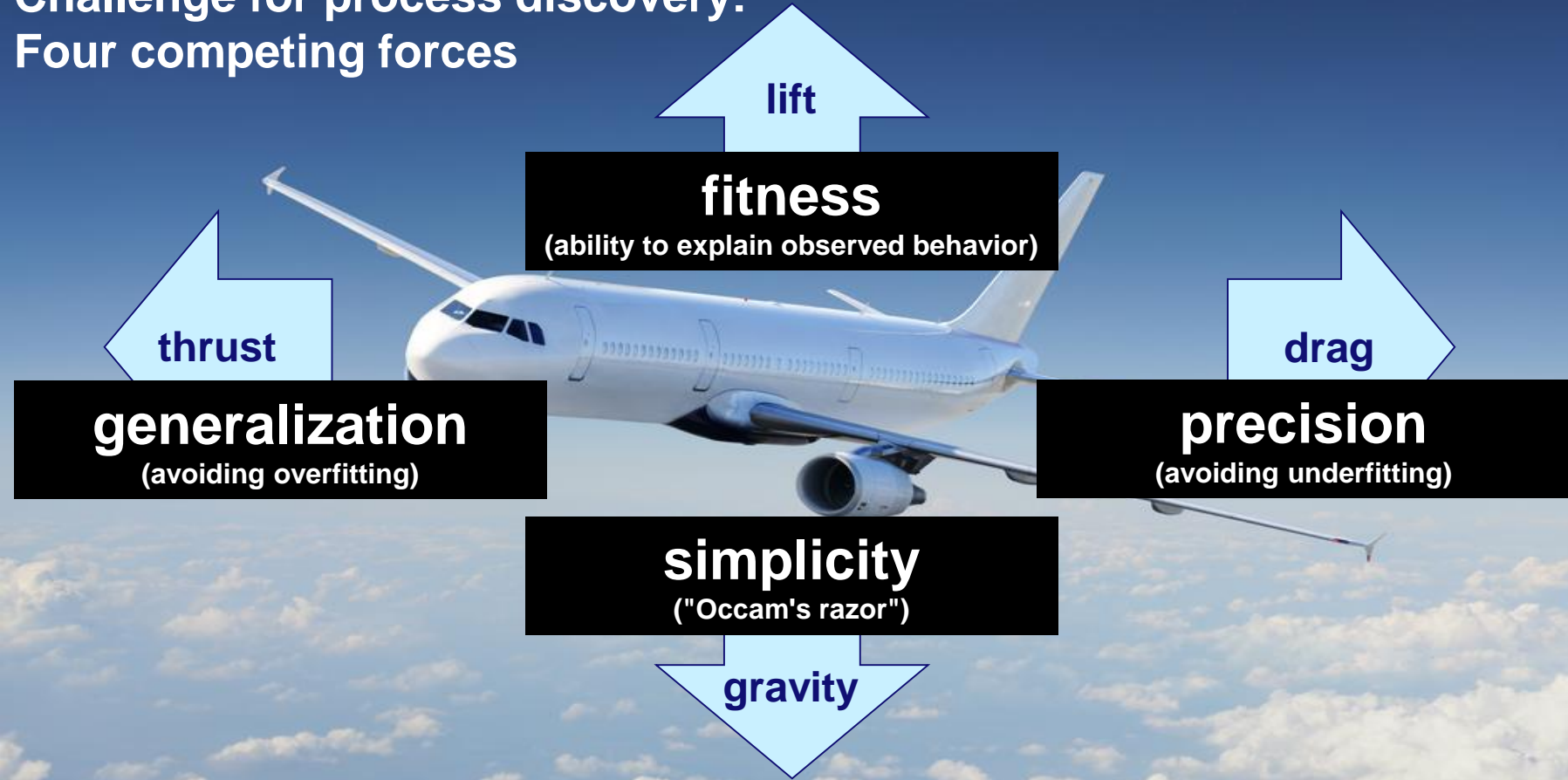
$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

Another BPMN example



$$L_2 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^4, \langle a, b, c, e, f, b, c, d \rangle^2, \langle a, b, c, e, f, c, b, d \rangle, \\ \langle a, c, b, e, f, b, c, d \rangle^2, \langle a, c, b, e, f, b, c, e, f, c, b, d \rangle]$$

Challenge for process discovery: Four competing forces

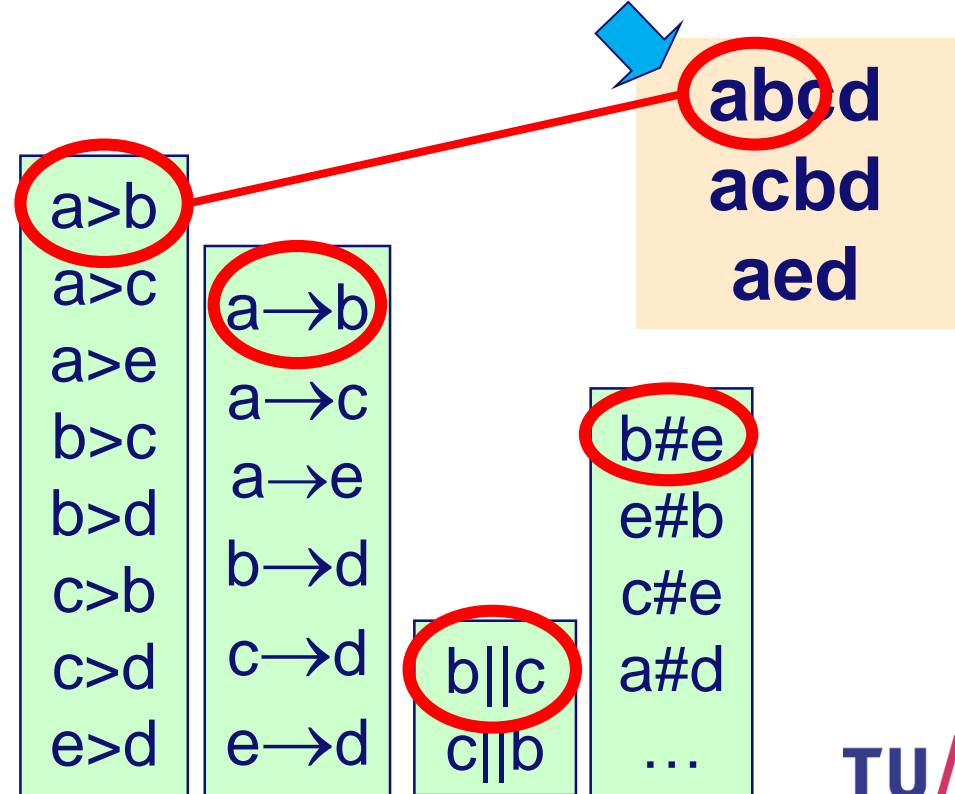


Will be discussed later ...

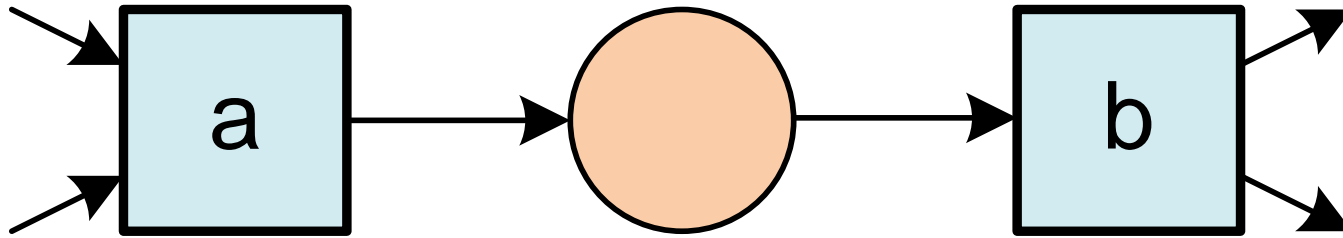
$>, \rightarrow, ||, \#$ relations

- Direct succession: $x > y$ iff for some case x is directly followed by y .
- Causality: $x \rightarrow y$ iff $x > y$ and not $y > x$.
- Parallel: $x || y$ iff $x > y$ and $y > x$
- Choice: $x \# y$ iff not $x > y$ and not $y > x$.

$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

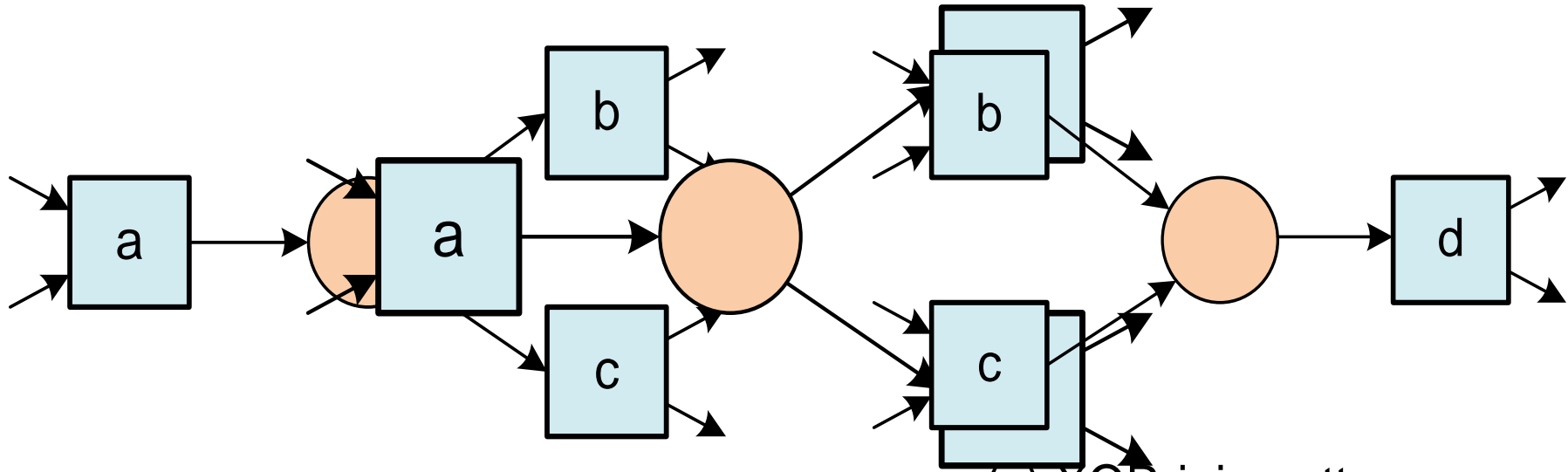


Basic Idea Used by Alpha Algorithm (1)



(a) sequence pattern: $a \rightarrow b$

Basic Idea Used by Alpha Algorithm (2)



(b) XOR-split pattern:

$a \rightarrow b$, $a \rightarrow c$, and $b \# c$

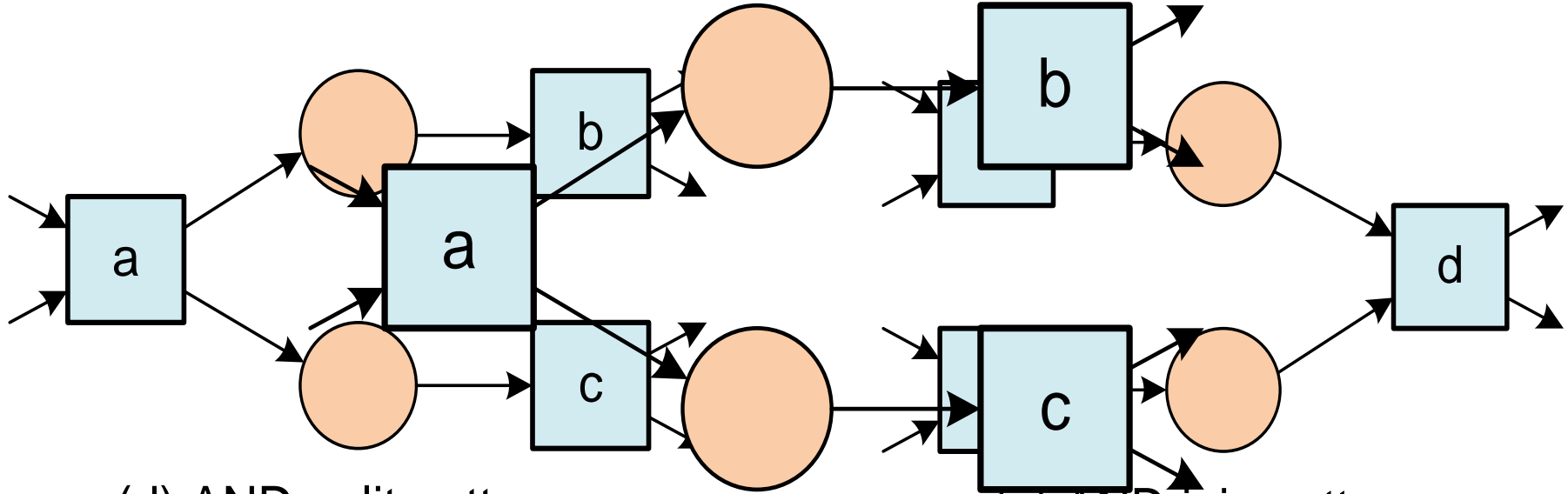
(b) XOR-split pattern:

$a \rightarrow b$, $a \rightarrow c$, and $b \# c$

(c) XOR-join pattern:

$b \rightarrow d$, $c \rightarrow d$, and $b \# c$

Basic Idea Used by Alpha Algorithm (3)



(d) AND-split pattern:

$a \rightarrow b$, $a \rightarrow c$, and $b \parallel c$

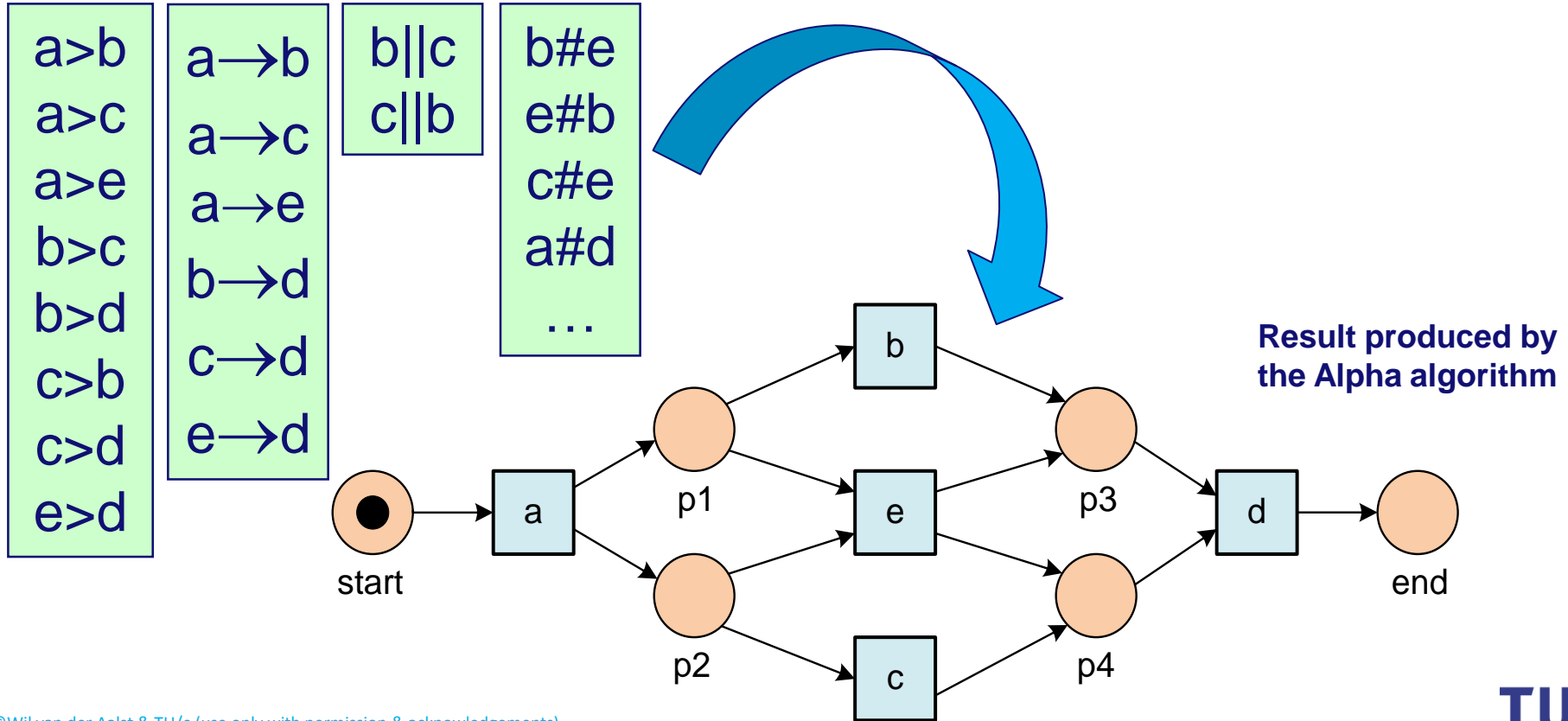
(e) AND-join pattern:

$b \rightarrow d$, $c \rightarrow d$, and $b \parallel c$

$a \rightarrow b$, $a \rightarrow c$, and $b \parallel c$

Example Revisited

$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$



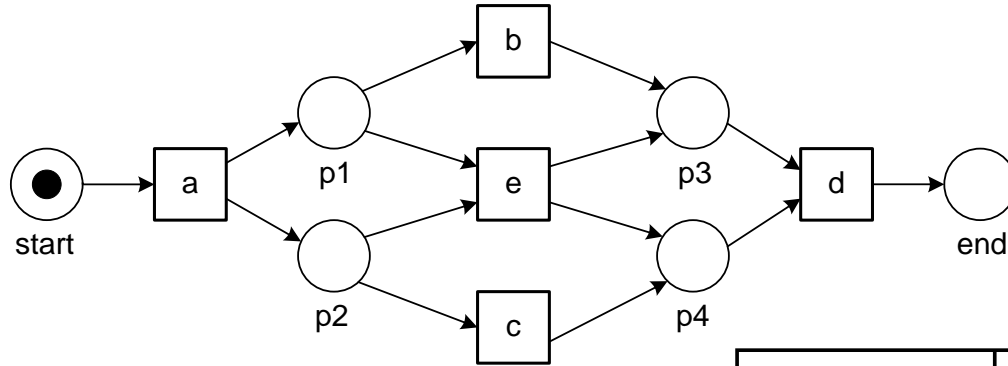
Footprint of L_1

One of the
following:
 $\rightarrow, \leftarrow, \#, \parallel$

$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

	a	b	c	d	e
a	$\#_{L_1}$	\rightarrow_{L_1}	\rightarrow_{L_1}	$\#_{L_1}$	\rightarrow_{L_1}
b	\leftarrow_{L_1}	$\#_{L_1}$	\parallel_{L_1}	\rightarrow_{L_1}	$\#_{L_1}$
c	\leftarrow_{L_1}	\parallel_{L_1}	$\#_{L_1}$	\rightarrow_{L_1}	$\#_{L_1}$
d	$\#_{L_1}$	\leftarrow_{L_1}	\leftarrow_{L_1}	$\#_{L_1}$	\leftarrow_{L_1}
e	\leftarrow_{L_1}	$\#_{L_1}$	$\#_{L_1}$	\rightarrow_{L_1}	$\#_{L_1}$

Discovered model has the same footprint



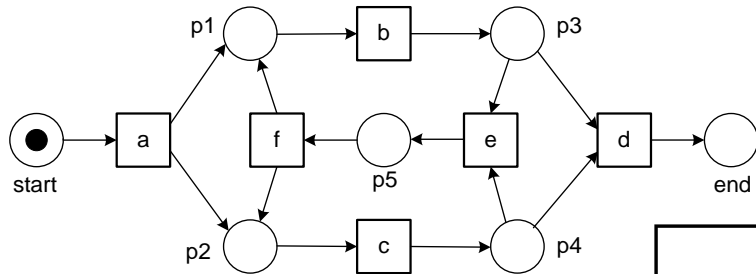
$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>
<i>a</i>	$\#_{L_1}$	\rightarrow_{L_1}	\rightarrow_{L_1}	$\#_{L_1}$	\rightarrow_{L_1}
<i>b</i>	\leftarrow_{L_1}	$\#_{L_1}$	\parallel_{L_1}	\rightarrow_{L_1}	$\#_{L_1}$
<i>c</i>	\leftarrow_{L_1}	\parallel_{L_1}	$\#_{L_1}$	\rightarrow_{L_1}	$\#_{L_1}$
<i>d</i>	$\#_{L_1}$	\leftarrow_{L_1}	\leftarrow_{L_1}	$\#_{L_1}$	\leftarrow_{L_1}
<i>e</i>	\leftarrow_{L_1}	$\#_{L_1}$	$\#_{L_1}$	\rightarrow_{L_1}	$\#_{L_1}$

Log and model agree on footprint

Footprint of L_2

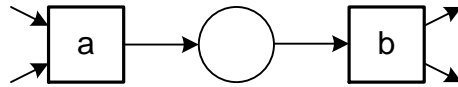
$$L_2 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^4, \langle a, b, c, e, f, b, c, d \rangle^2, \langle a, b, c, e, f, c, b, d \rangle, \\ \langle a, c, b, e, f, b, c, d \rangle^2, \langle a, c, b, e, f, b, c, e, f, c, b, d \rangle]$$



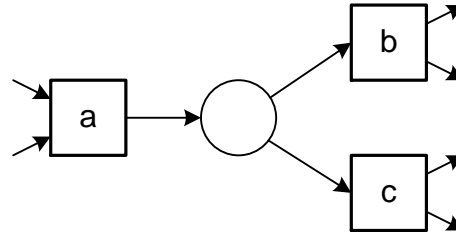
	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>
<i>a</i>	#	→	→	#	#	#
<i>b</i>	←	#		→	→	←
<i>c</i>	←		#	→	→	←
<i>d</i>	#	←	←	#	#	#
<i>e</i>	#	←	←	#	#	→
<i>f</i>	#	→	→	#	←	#

Log and model agree on footprint

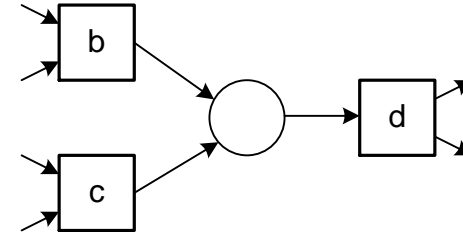
Summary: Simple process patterns can be discovered from event logs



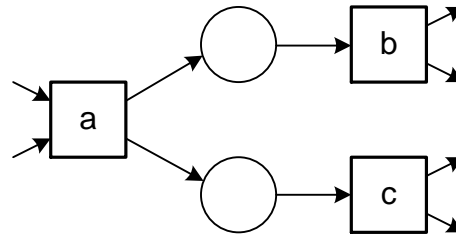
(a) sequence pattern: $a \rightarrow b$



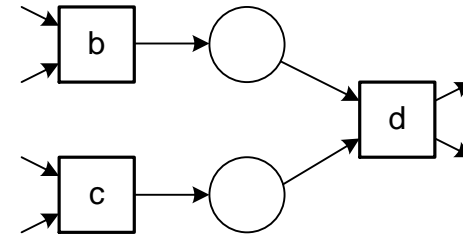
(b) XOR-split pattern:
 $a \rightarrow b$, $a \rightarrow c$, and $b \# c$



(c) XOR-join pattern:
 $b \rightarrow d$, $c \rightarrow d$, and $b \# c$



(d) AND-split pattern:
 $a \rightarrow b$, $a \rightarrow c$, and $b \parallel c$



(e) AND-join pattern:
 $b \rightarrow d$, $c \rightarrow d$, and $b \parallel c$



Let L be an event log over T . $\alpha(L)$ is defined as follows.

$$1. T_L = \{ t \in T \mid \exists_{\sigma \in L} t \in \sigma \},$$

$$2. T_I = \{ t \in T \mid \exists_{\sigma \in L} t = \text{first}(\sigma) \},$$

$$3. T_O = \{ t \in T \mid \exists_{\sigma \in L} t = \text{last}(\sigma) \},$$

$$4. X_L = \{ (A, B) \mid A \subseteq T_L \wedge A \neq \emptyset \wedge B \subseteq T_L \wedge B \neq \emptyset \wedge \\ \forall_{a \in A} \forall_{b \in B} a \rightarrow_L b \wedge \forall_{a_1, a_2 \in A} a_1 \#_L a_2 \wedge \forall_{b_1, b_2 \in B} b_1 \#_L b_2 \},$$

$$5. Y_L = \{ (A, B) \in X_L \mid \forall_{(A', B') \in X_L} A \subseteq A' \wedge B \subseteq B' \Rightarrow (A, B) = (A', B') \},$$

$$6. P_L = \{ p_{(A, B)} \mid (A, B) \in Y_L \} \cup \{ i_L, o_L \},$$

$$7. F_L = \{ (a, p_{(A, B)}) \mid (A, B) \in Y_L \wedge a \in A \} \cup \{ (p_{(A, B)}, b) \mid (A, B) \in Y_L \wedge b \in B \} \\ \cup \{ (i_L, t) \mid t \in T_I \} \cup \{ (t, o_L) \mid t \in T_O \}, \text{ and}$$

$$8. \alpha(L) = (P_L, T_L, F_L).$$

The α -algorithm

Let L be an event log over T . Then, $\alpha(L)$ is defined as follows:

1. $T_L = \{ t \in T \mid \exists_{\sigma \in L} t \in \sigma \},$

Each activity in L corresponds to a transition in $\alpha(L)$.

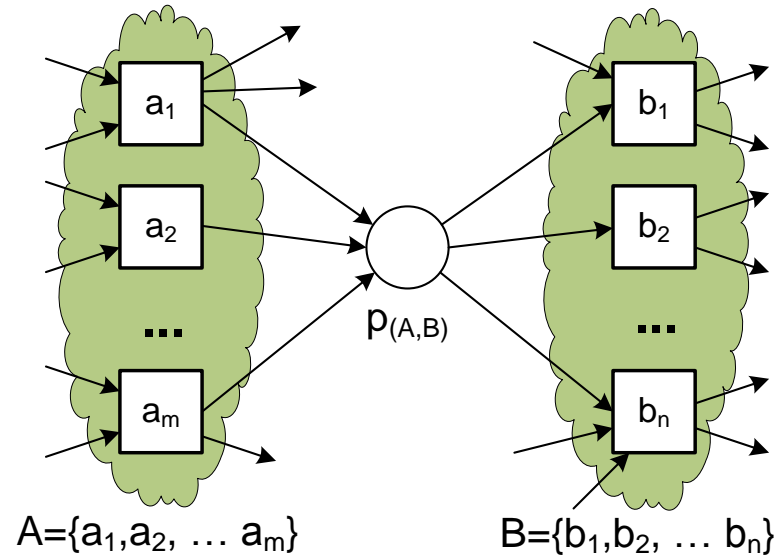
2. $T_I = \{ t \in T \mid \exists_{\sigma \in L} t = \text{first}(\sigma) \}$

Fix the set of start activities – that is, the **first** elements of each trace: $\langle t_1, \dots, t_n \rangle, \dots, \langle t'_1, \dots, t'_m \rangle$

3. $T_O = \{ t \in T \mid \exists_{\sigma \in L} t = \text{last}(\sigma) \}$

Fix the set of end activities – that is, elements that appear **last** in a trace : $\langle t_1, \dots, t_n \rangle, \dots, \langle t'_1, \dots, t'_m \rangle$

Next steps aim at finding places



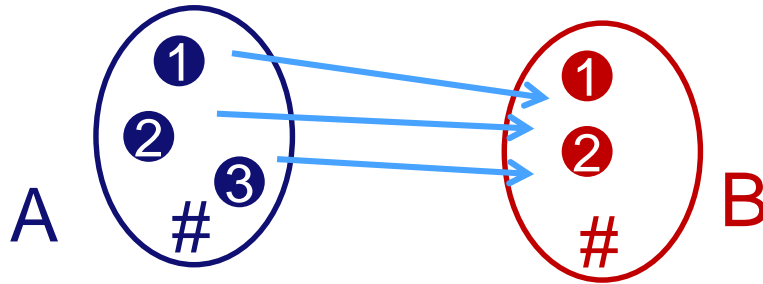
Step 4: Calculate pairs (A, B)

Step 5: Delete non-maximal pairs (A, B)

Step 6: Determine places $p_{(A, B)}$ from pairs (A, B)

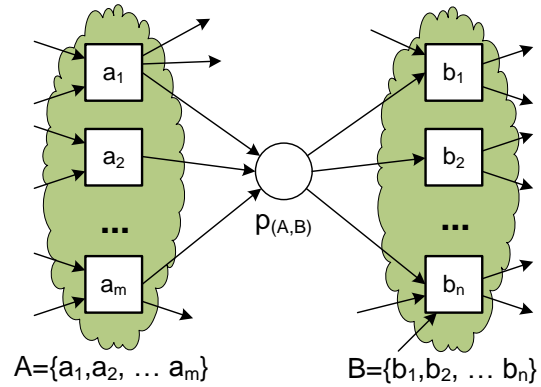
The α -algorithm (cont.)

$$\begin{aligned} 4. \quad X_L = \{ (A, B) \mid & A \subseteq T_L \wedge A \neq \emptyset \wedge B \subseteq T_L \wedge B \neq \emptyset \\ & \wedge \forall_{a \in A} \forall_{b \in B} a \rightarrow_L b \\ & \wedge \forall_{a_1, a_2 \in A} a_1 \#_L a_2 \\ & \wedge \forall_{b_1, b_2 \in B} b_1 \#_L b_2 \}, \end{aligned}$$



Find pairs (A, B) of sets of activities such that every element $a \in A$ and every element $b \in B$ are causally related (i.e., $a \rightarrow_L b$), all elements in A are independent ($a_1 \#_L a_2$), and all elements in B are independent ($b_1 \#_L b_2$).

Places as footprints

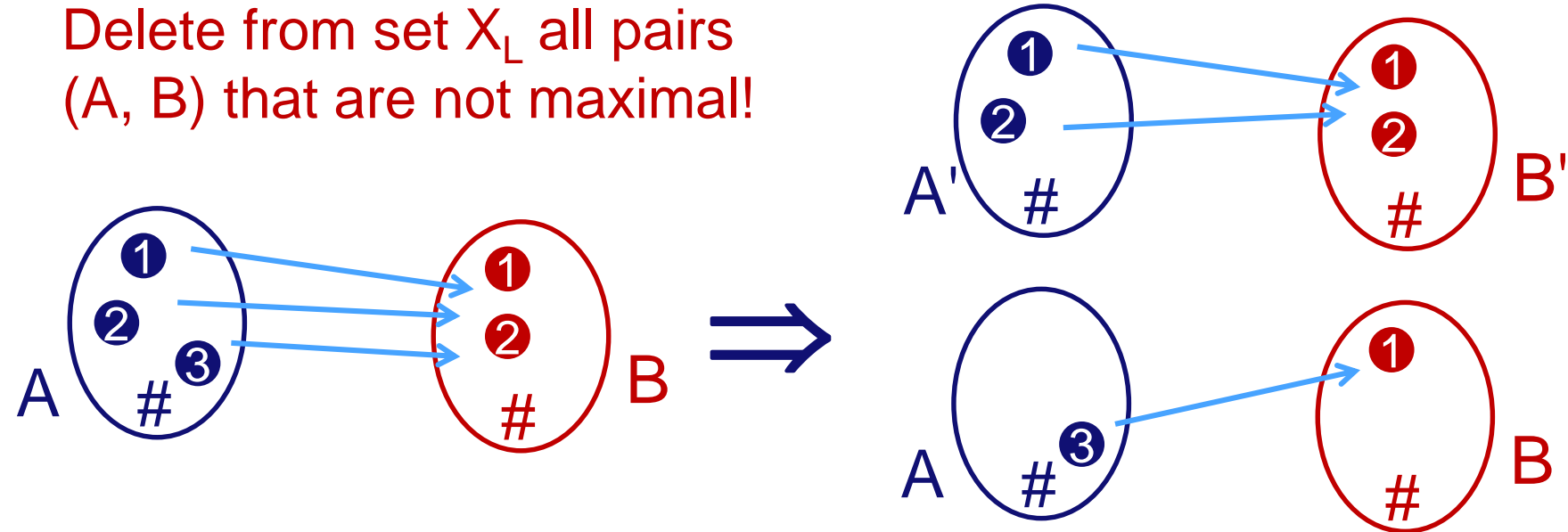


	a_1	a_2	...	a_m	b_1	b_2	...	b_n
a_1	#	#	...	#	→	→	...	→
a_2	#	#	...	#	→	→	...	→
...
a_m	#	#	...	#	→	→	...	→
b_1	←	←	...	←	#	#	...	#
b_2	←	←	...	←	#	#	...	#
...
b_n	←	←	...	←	#	#	...	#

The α -algorithm (cont.)

5. $Y_L = \{ (A,B) \in X_L \mid \forall_{(A',B') \in X_L} A \subseteq A' \wedge B \subseteq B' \Rightarrow (A,B) = (A',B') \}$

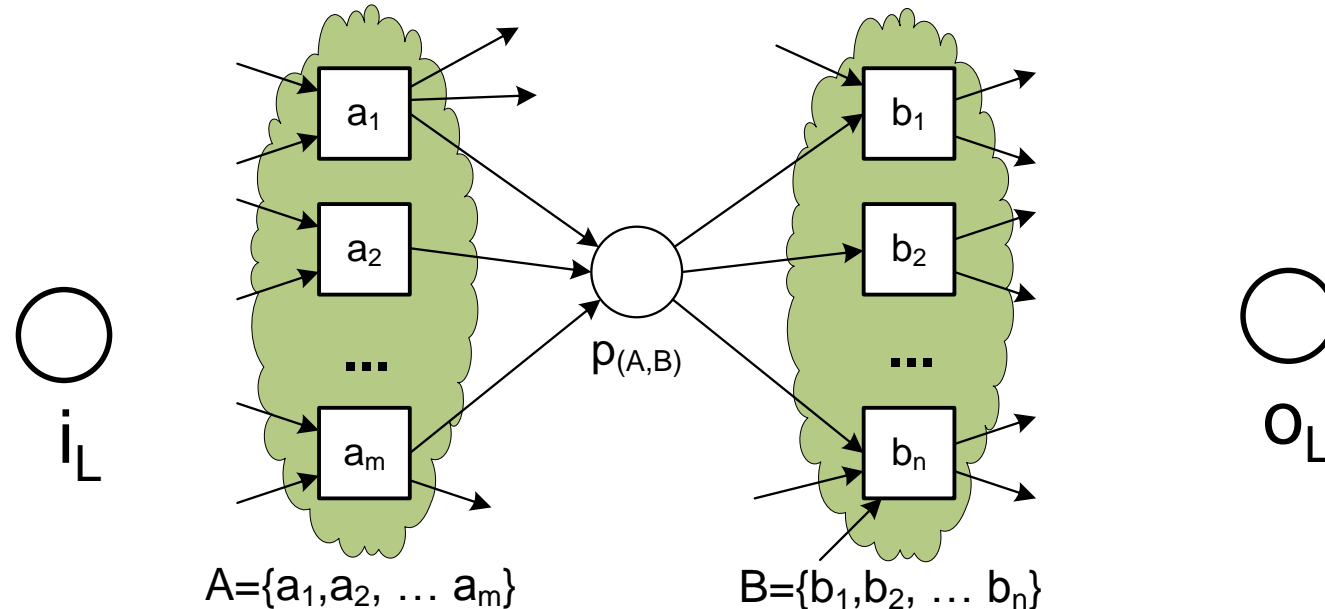
Delete from set X_L all pairs (A, B) that are not maximal!



The α -algorithm (cont.)

6. $P_L = \{ p_{(A,B)} \mid (A,B) \in Y_L \} \cup \{i_L, o_L\},$

Determine the place set:
Each element (A, B) of Y_L is a place. To ensure the workflow structure, add a source place i_L and a target place o_L



The α -algorithm (cont.)

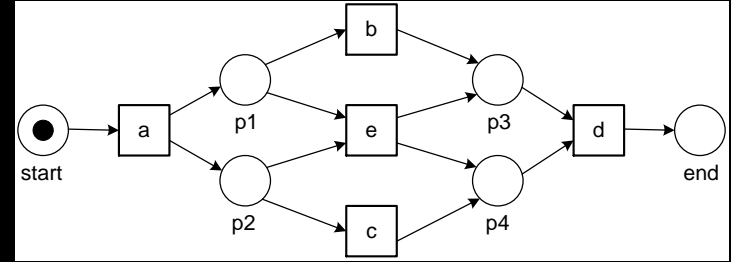
$$\begin{aligned} 7. \quad F_L = & \{ (a, p_{(A,B)}) \mid (A,B) \in Y_L \wedge a \in A \} \\ & \cup \{ (p_{(A,B)}, b) \mid (A,B) \in Y_L \wedge b \in B \} \\ & \cup \{ (i_L, t) \mid t \in T_I \} \cup \{ (t, o_L) \mid t \in T_O \} \end{aligned}$$

Determine the flow relation: Connect each place $p_{(A,B)}$ with each element a of its set A of source transitions and with each element of its set B of target transitions. In addition, draw an arc from the source place i_L to each start transition $t \in T_I$ and an arc from each end transition $t \in T_O$ to the sink place o_L .

$$8. \quad \alpha(L) = (P_L, T_L, F_L)$$

$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

	a	b	c	d	e
a	$\#_{L_1}$	\rightarrow_{L_1}	\rightarrow_{L_1}	$\#_{L_1}$	\rightarrow_{L_1}
b	\leftarrow_{L_1}	$\#_{L_1}$	\parallel_{L_1}	\rightarrow_{L_1}	$\#_{L_1}$
c	\leftarrow_{L_1}	\parallel_{L_1}	$\#_{L_1}$	\rightarrow_{L_1}	$\#_{L_1}$
d	$\#_{L_1}$	\leftarrow_{L_1}	\leftarrow_{L_1}	$\#_{L_1}$	\leftarrow_{L_1}
e	\leftarrow_{L_1}	$\#_{L_1}$	$\#_{L_1}$	\rightarrow_{L_1}	$\#_{L_1}$

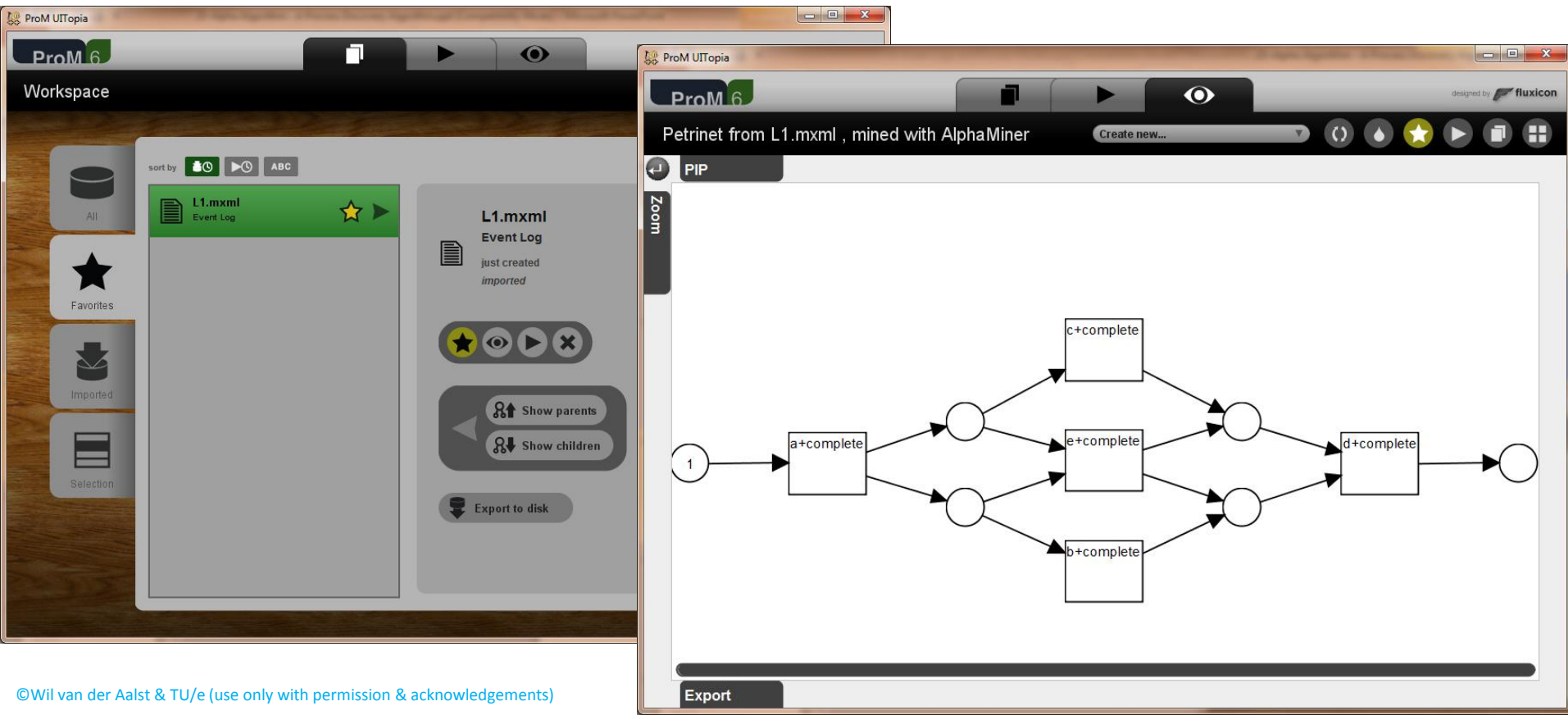


$$X_{L_1} = \{(\{a\}, \{b\}), (\{a\}, \{c\}), (\{a\}, \{e\}), (\{a\}, \{b, e\}), (\{a\}, \{c, e\}),$$

$$(\{b\}, \{d\}), (\{c\}, \{d\}), (\{e\}, \{d\}), (\{b, e\}, \{d\}), (\{c, e\}, \{d\})\}$$

$$Y_{L_1} = \{(\{a\}, \{b, e\}), (\{a\}, \{c, e\}), (\{b, e\}, \{d\}), (\{c, e\}, \{d\})\}$$

ProM's output for event log L₁



Question:

Give footprint matrix for event log L_3

$$L_3 = [\langle a, b, c, d, e, f, b, d, c, e, g \rangle, \\ \langle a, b, d, c, e, g \rangle^2, \\ \langle a, b, c, d, e, f, b, c, d, e, f, b, d, c, e, g \rangle]$$

Answer:

Footprint matrix for event log L_3

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>
<i>a</i>	#	→	#	#	#	#	#
<i>b</i>	←	#	→	→	#	←	#
<i>c</i>	#	←	#		→	#	#
<i>d</i>	#	←		#	→	#	#
<i>e</i>	#	#	←	←	#	→	→
<i>f</i>	#	→	#	#	←	#	#
<i>g</i>	#	#	#	#	←	#	#

$L_3 = [\langle a, b, c, d, e, f, b, d, c, e, g \rangle,$

$\langle a, b, d, c, e, g \rangle^2,$

$\langle a, b, c, d, e, f, b, c, d, e, f, b, d, c, e, g \rangle]$

Question:

Apply the 8 steps of the Alpha algorithm.

$$L_3 = [\langle a, b, c, d, e, f, b, d, c, e, g \rangle, \\ \langle a, b, d, c, e, g \rangle^2, \\ \langle a, b, c, d, e, f, b, c, d, e, f, b, d, c, e, g \rangle]$$

Let L be an event log over T . $\alpha(L)$ is defined as follows.

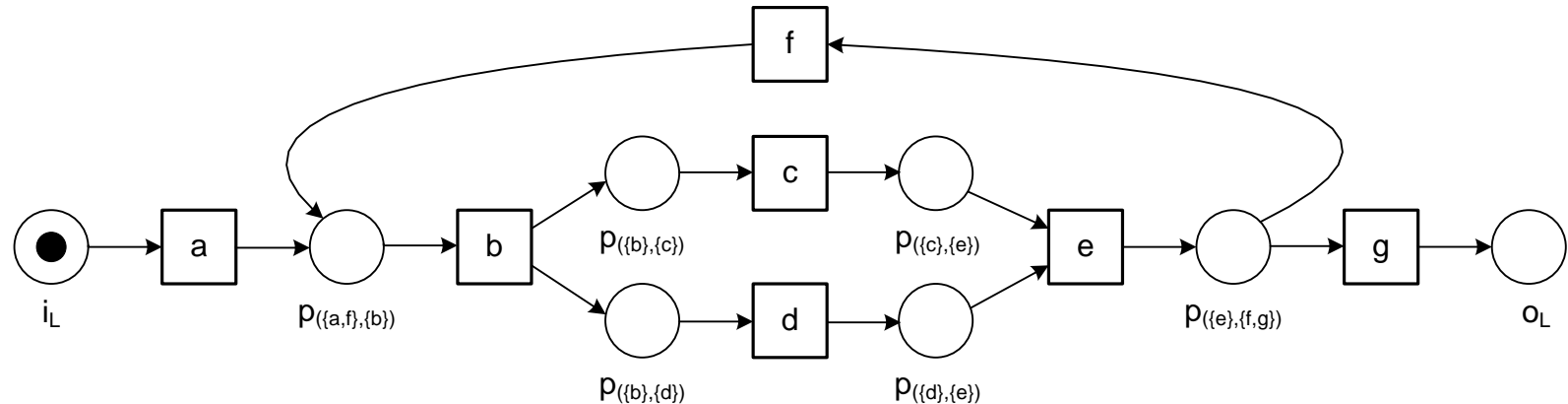
1. $T_L = \{t \in T \mid \exists \sigma \in L, t \in \sigma\}$,
2. $T_I = \{t \in T \mid \exists \sigma \in L, t = \text{first}(\sigma)\}$,
3. $T_O = \{t \in T \mid \exists \sigma \in L, t = \text{last}(\sigma)\}$,
4. $X_L = \{(A, B) \mid A \subseteq T_L \wedge A \neq \emptyset \wedge B \subseteq T_L \wedge B \neq \emptyset \wedge \forall a \in A, \forall b \in B, a \rightarrow_L b \wedge \forall a_1, a_2 \in A, a_1 \#_L a_2 \wedge \forall b_1, b_2 \in B, b_1 \#_L b_2\}$,
5. $Y_L = \{(A, B) \in X_L \mid \forall (A', B') \in X_L, A \subseteq A' \wedge B \subseteq B' \Rightarrow (A, B) = (A', B')\}$,
6. $P_L = \{p_{(A, B)} \mid (A, B) \in Y_L\} \cup \{i_L, o_L\}$,
7. $F_L = \{(a, p_{(A, B)}) \mid (A, B) \in Y_L \wedge a \in A\} \cup \{(p_{(A, B)}, b) \mid (A, B) \in Y_L \wedge b \in B\} \cup \{(i_L, t) \mid t \in T_I\} \cup \{(t, o_L) \mid t \in T_O\}$, and
8. $\alpha(L) = (P_L, T_L, F_L)$.

	a	b	c	d	e	f	g
a	#	\rightarrow	#	#	#	#	#
b	\leftarrow	#	\rightarrow	\rightarrow	#	\leftarrow	#
c	#	\leftarrow	#	\parallel	\rightarrow	#	#
d	#	\leftarrow	\parallel	#	\rightarrow	#	#
e	#	#	\leftarrow	\leftarrow	#	\rightarrow	\rightarrow
f	#	\rightarrow	#	#	\leftarrow	#	#
g	#	#	#	#	\leftarrow	#	#

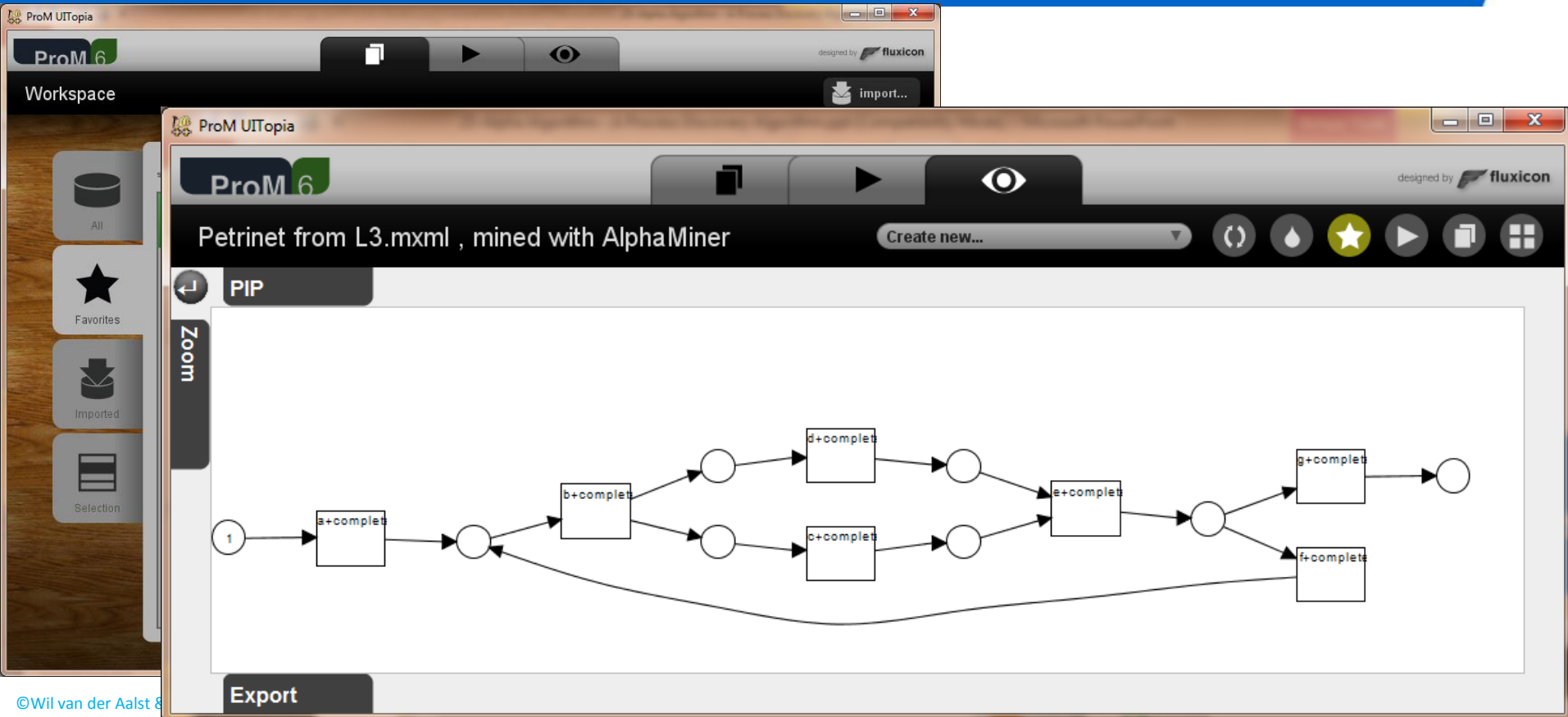
Model for L_3 discovered by the Alpha algorithm

$L_3 = [\langle a, b, c, d, e, f, b, d, c, e, g \rangle,$
 $\langle a, b, d, c, e, g \rangle^2,$
 $\langle a, b, c, d, e, f, b, c, d, e, f, b, d, c, e, g \rangle]$

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>
<i>a</i>	#	→	#	#	#	#	#
<i>b</i>	←	#	→	→	#	←	#
<i>c</i>	#	←	#		→	#	#
<i>d</i>	#	←		#	→	#	#
<i>e</i>	#	#	←	←	#	→	→
<i>f</i>	#	→	#	#	←	#	#
<i>g</i>	#	#	#	#	←	#	#

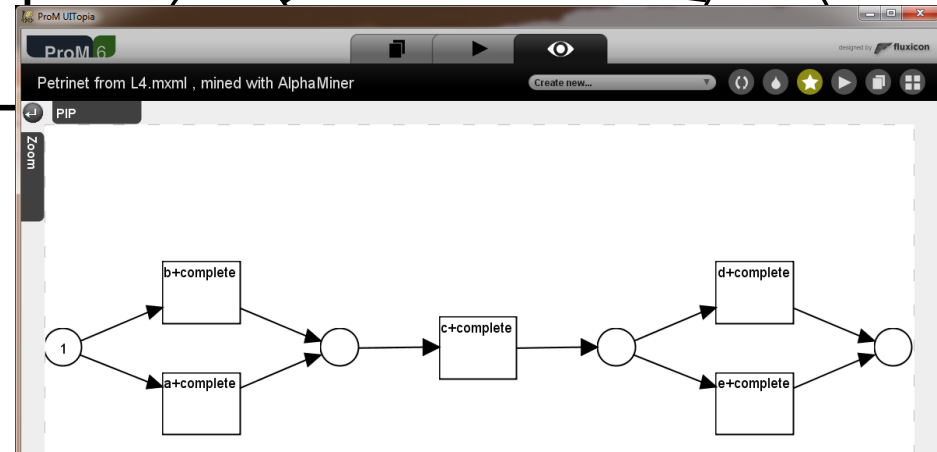
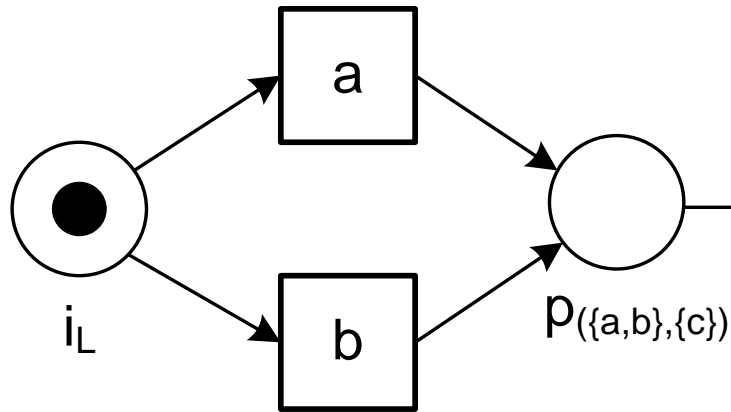


ProM's output for event log L₃



Another event log L_4

$$L_4 = [\langle a, c, d \rangle^{45}, \langle b, c, d \rangle^{42}, \langle a, c, e \rangle^{38}, \langle b, c, e \rangle^{22}]$$



Event log L_5

$$L_5 = [\langle a, b, e, f \rangle^2, \langle a, b, e, c, d, b, f \rangle^3, \langle a, b, c, e, d, b, f \rangle^2, \langle a, b, c, d, e, b, f \rangle^4, \langle a, e, b, c, d, b, f \rangle^3]$$

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>
<i>a</i>	#	→	#	#	→	#
<i>b</i>	←	#	→	←		→
<i>c</i>	#	←	#	→		#
<i>d</i>	#	→	←	#		#
<i>e</i>	←				#	→
<i>f</i>	#	←	#	#	←	#

$$T_L = \{a, b, c, d, e, f\}$$

$$T_I = \{a\}$$

$$T_I = \{f\}$$

$$X_L = \{(\{a\}, \{b\}), (\{a\}, \{e\}), (\{b\}, \{c\}), (\{b\}, \{f\}), (\{c\}, \{d\}), \\ (\{d\}, \{b\}), (\{e\}, \{f\}), (\{a, d\}, \{b\}), (\{b\}, \{c, f\})\}$$

$$Y_L = \{(\{a\}, \{e\}), (\{c\}, \{d\}), (\{e\}, \{f\}), (\{a, d\}, \{b\}), (\{b\}, \{c, f\})\}$$

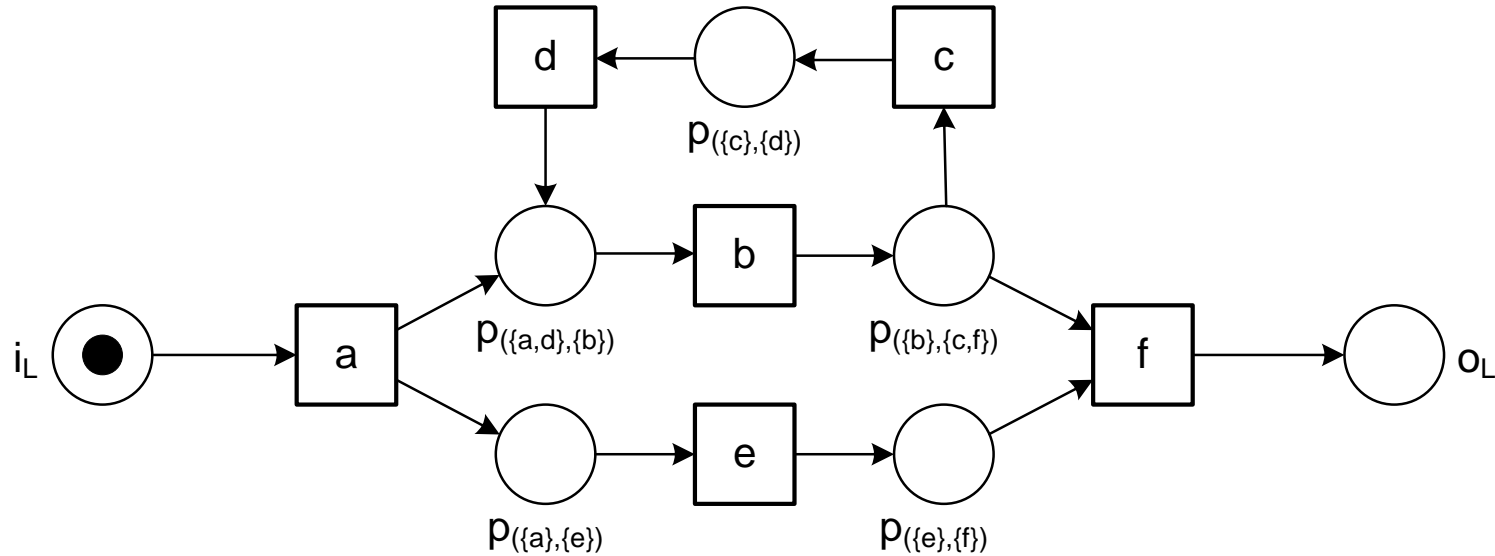
$$P_L = \{p(\{a\}, \{e\}), p(\{c\}, \{d\}), p(\{e\}, \{f\}), p(\{a, d\}, \{b\}), p(\{b\}, \{c, f\}), i_L, o_L\}$$

$$F_L = \{(a, p(\{a\}, \{e\})), (p(\{a\}, \{e\}), e), (c, p(\{c\}, \{d\})), (p(\{c\}, \{d\}), d), \\ (e, p(\{e\}, \{f\})), (p(\{e\}, \{f\}), f), (a, p(\{a, d\}, \{b\})), (d, p(\{a, d\}, \{b\})), \\ (p(\{a, d\}, \{b\}), b), (b, p(\{b\}, \{c, f\})), (p(\{b\}, \{c, f\}), c), (p(\{b\}, \{c, f\}), f), \\ (i_L, a), (f, o_L)\}$$

$$\alpha(L) = (P_L, T_L, F_L)$$

Discovered model

$$L_5 = [\langle a, b, e, f \rangle^2, \langle a, b, e, c, d, b, f \rangle^3, \langle a, b, c, e, d, b, f \rangle^2, \langle a, b, c, d, e, b, f \rangle^4, \langle a, e, b, c, d, b, f \rangle^3]$$



$$X_L = \{(\{a\}, \{b\}), (\{a\}, \{e\}), (\{b\}, \{c\}), (\{b\}, \{f\}), (\{c\}, \{d\}), (\{d\}, \{b\}), (\{e\}, \{f\}), (\{a, d\}, \{b\}), (\{b\}, \{c, f\})\}$$

$$Y_L = \{(\{a\}, \{e\}), (\{c\}, \{d\}), (\{e\}, \{f\}), (\{a, d\}, \{b\}), (\{b\}, \{c, f\})\}$$

Summary

- **The Alpha algorithm provides a basic process discovery approach.**
- **It has many limitations. These will be discussed later.**
- **However, it nicely illustrates the key ingredients of process discovery.**
- **Hence, it is important to understand the algorithm and practice using concrete examples.**

Part I: Introduction

Chapter 1

Data Science
in Action

Chapter 2

Process Mining:
The Missing Link

Part II: Preliminaries

Chapter 3

Process Modeling
and Analysis

Chapter 4

Data Mining

Part III: From Event Logs to Process Models

Chapter 5

Getting the Data

Chapter 6

Process Discovery:
An Introduction

Chapter 7

Advanced Process
Discovery Techniques

Part IV: Beyond Process Discovery

Chapter 8

Conformance
Checking

Chapter 9

Mining Additional
Perspectives

Chapter 10

Operational Support

Part V: Putting Process Mining to Work

Chapter 11

Process Mining
Software

Chapter 12

Process Mining in the
Large

Chapter 13

Analyzing “Lasagna
Processes”

Chapter 14

Analyzing “Spaghetti
Processes”

Part VI: Reflection

Chapter 15

Cartography and
Navigation

Chapter 16

Epilogue

