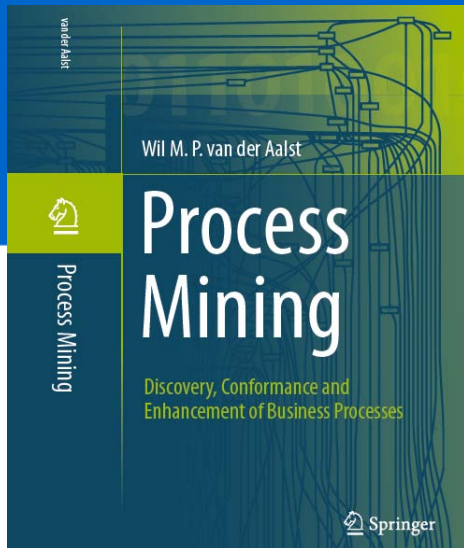


Chapter 6

Advanced Process Discovery Techniques

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



TU/e Technische Universiteit
Eindhoven
University of Technology

Where innovation starts

Overview

Chapter 1
Introduction

Part I: Preliminaries

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”

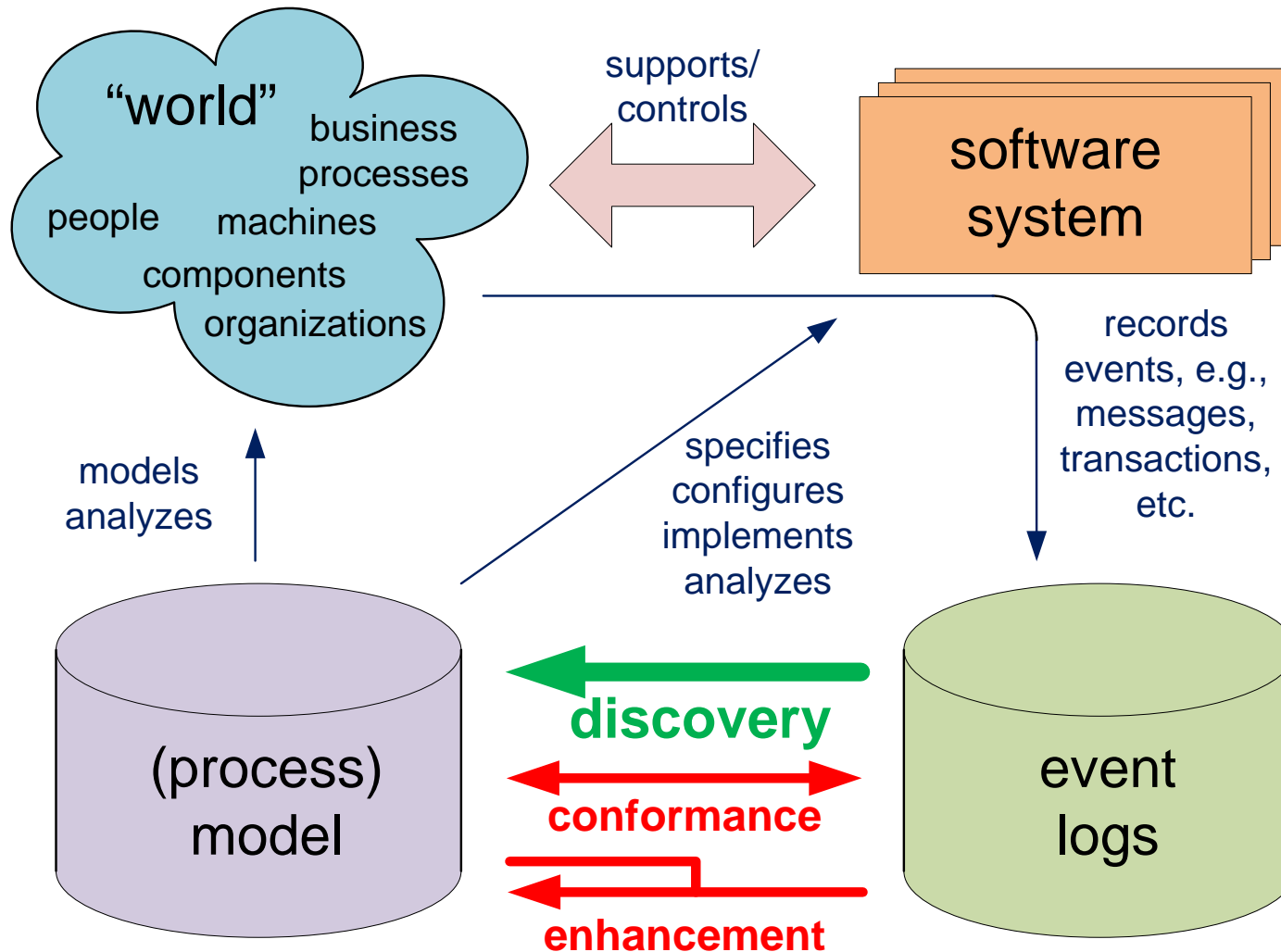
Chapter 12
Analyzing “Spaghetti
Processes”

Part V: Reflection

Chapter 13
Cartography and
Navigation

Chapter 14
Epilogue

Process discovery



Challenge

“able to replay event log”

“Occam’s razor”

fitness

simplicity

process
discovery

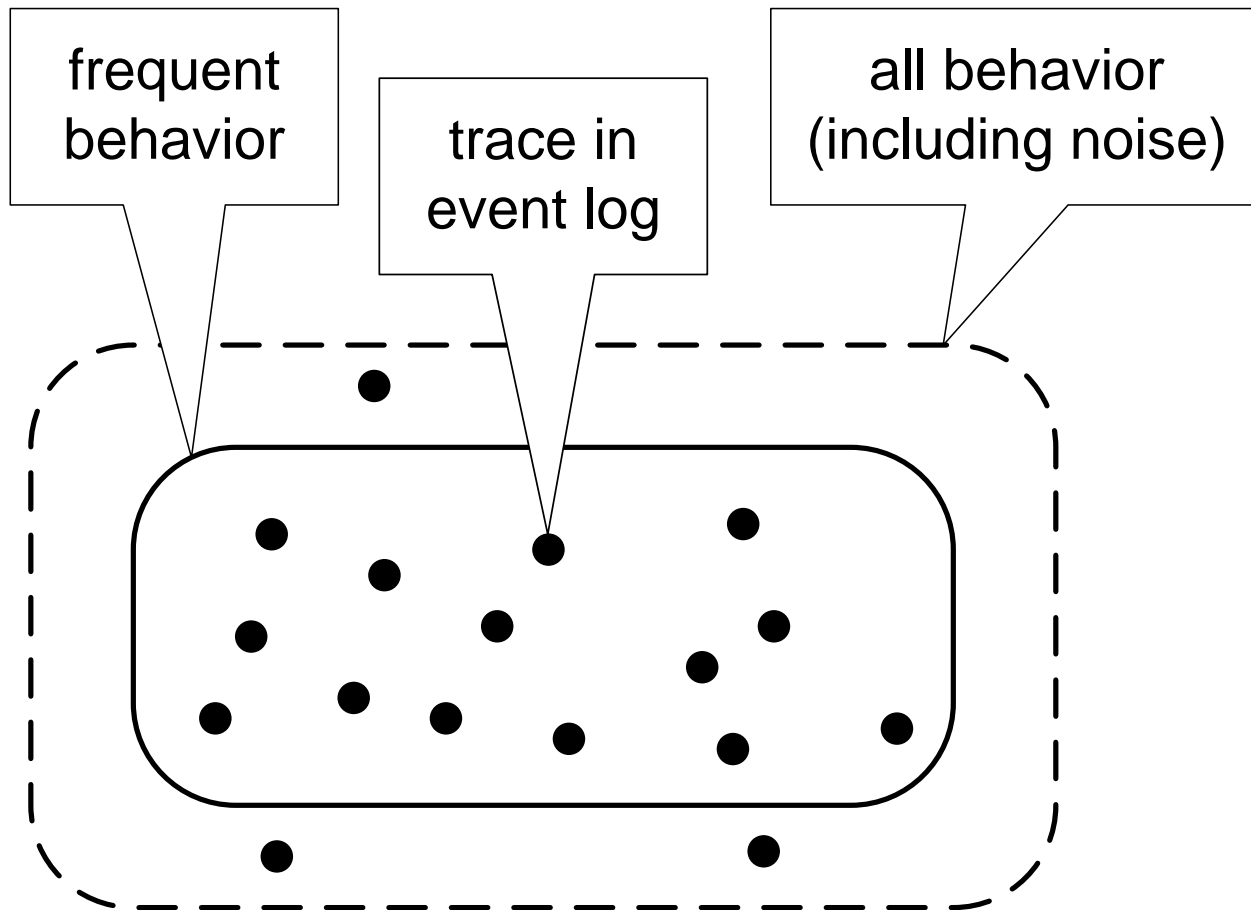
generalization

precision

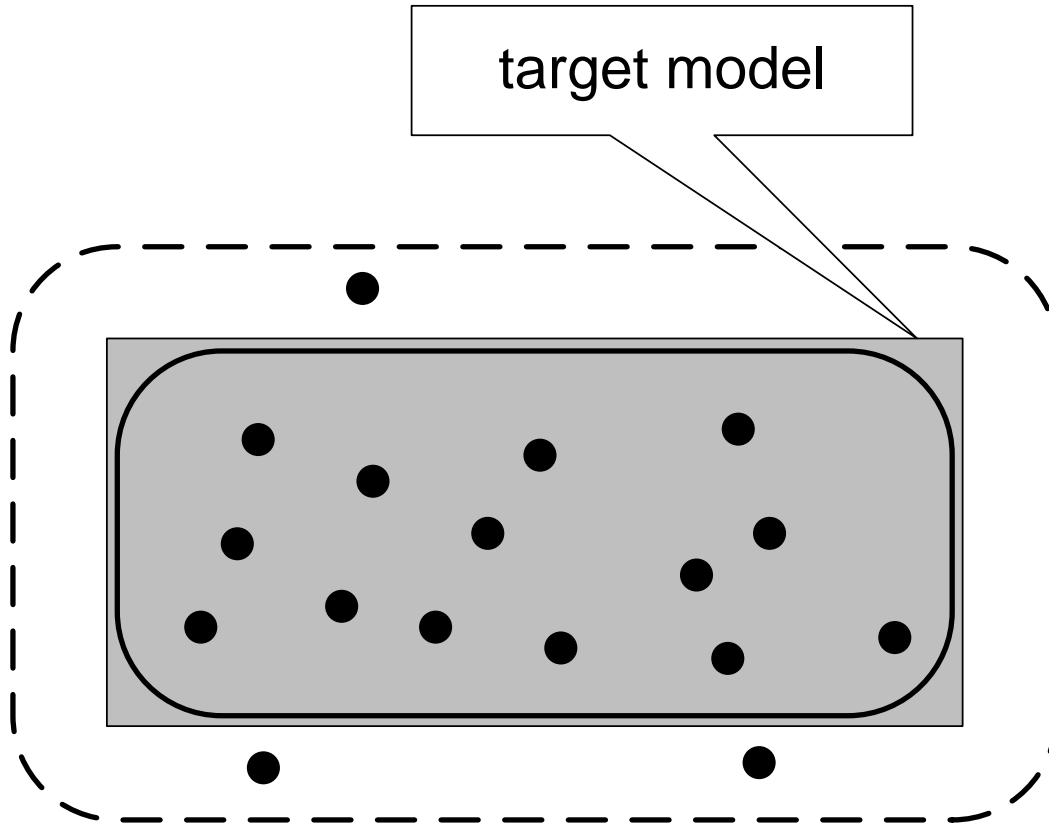
“not overfitting the log”

“not underfitting the log”

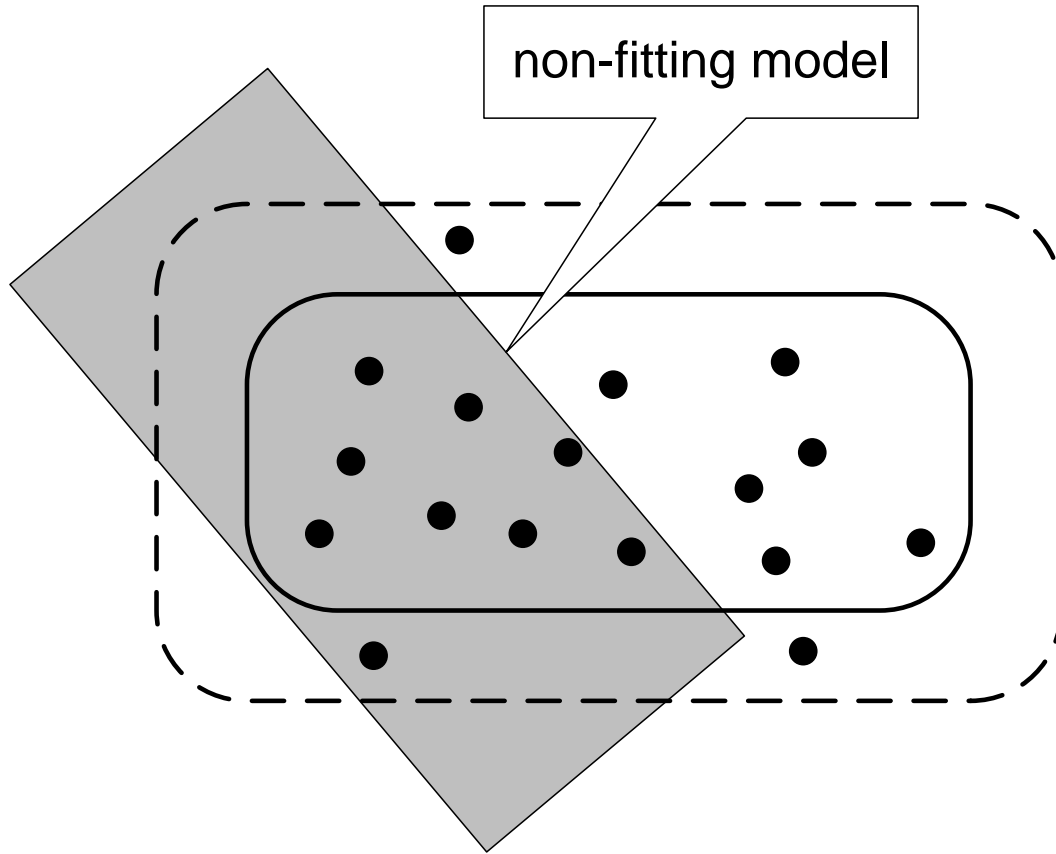
Observing a stable process infinitely long



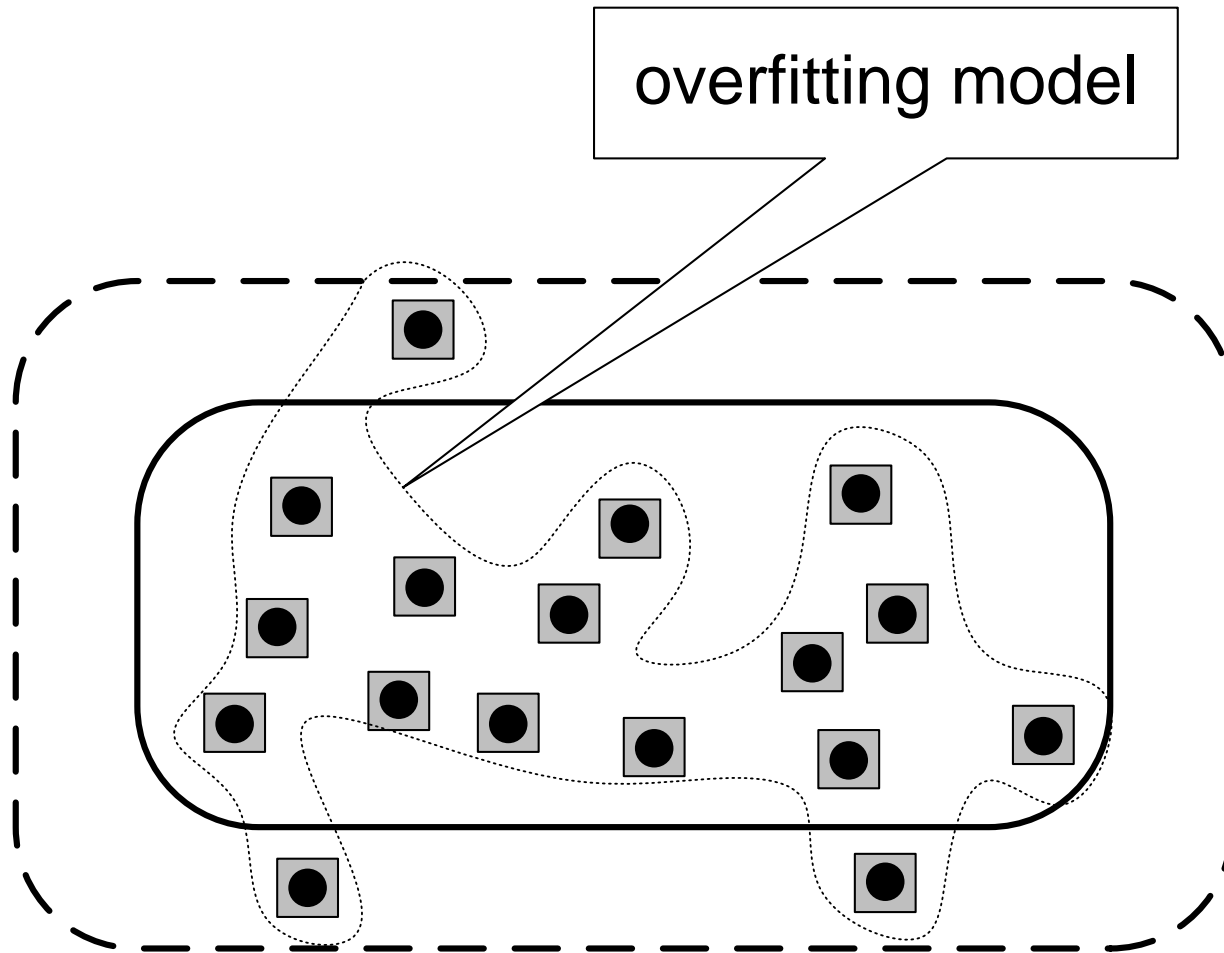
Target model



Non-fitting model

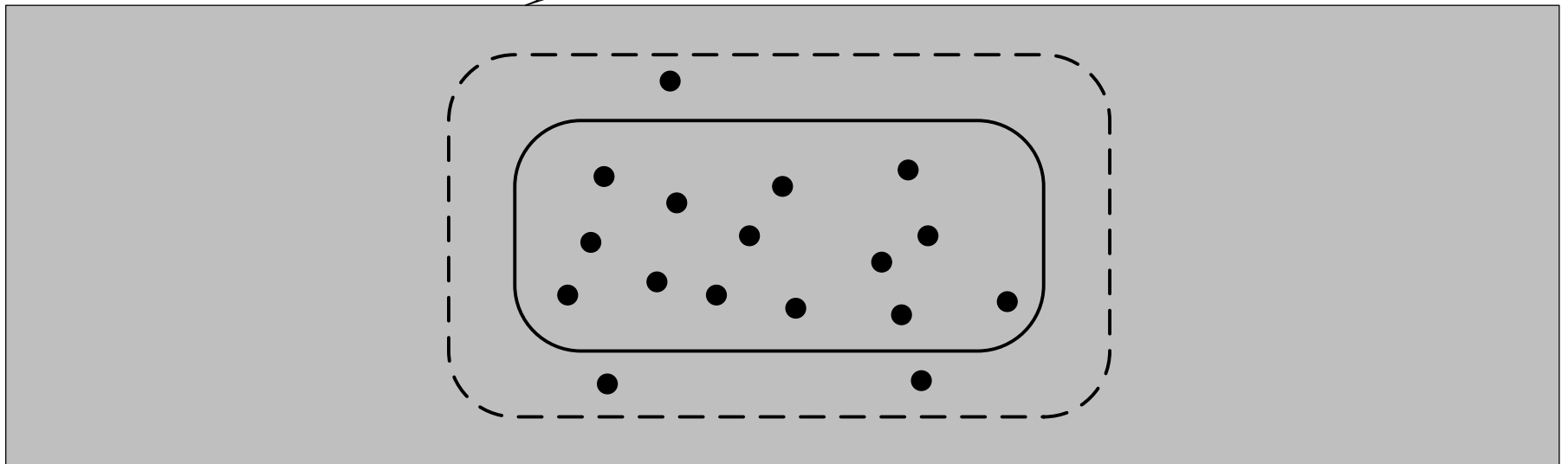


Overfitting model



Underfitting model

underfitting model



Characteristics of process discovery algorithms

- **Representational bias**
 - Inability to represent concurrency
 - Inability to deal with (arbitrary) loops
 - Inability to represent silent actions
 - Inability to represent duplicate actions
 - Inability to model OR-splits/joins
 - Inability to represent non-free-choice behavior
 - Inability to represent hierarchy
- **Ability to deal with noise**
- **Completeness notion assumed**
- **Approach used** (direct algorithmic approaches, two-phase approaches, computational intelligence approaches, partial approaches, etc.)

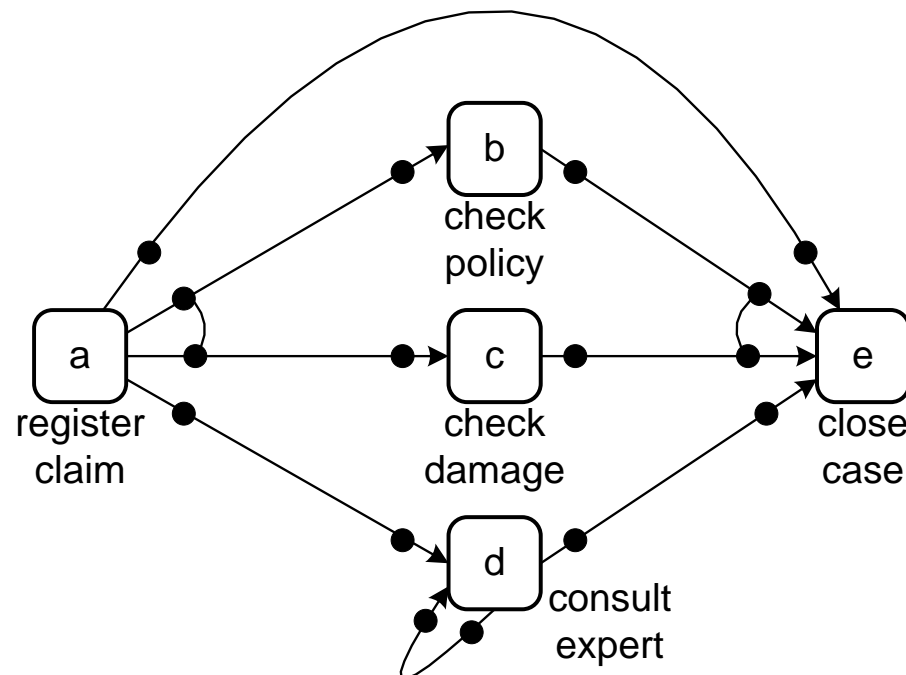
Examples

- **Algorithmic techniques**
 - Alpha miner
 - Alpha+, Alpha++, Alpha#
 - FSM miner
 - Fuzzy miner
 - Heuristic miner
 - Multi phase miner
- **Genetic process mining**
 - Single/duplicate tasks
 - Distributed GM
- **Region-based process mining**
 - State-based regions
 - Language based regions
- **Classical approaches not dealing with concurrency**
 - Inductive inference (Mark Gold, Dana Angluin et al.)
 - Sequence mining



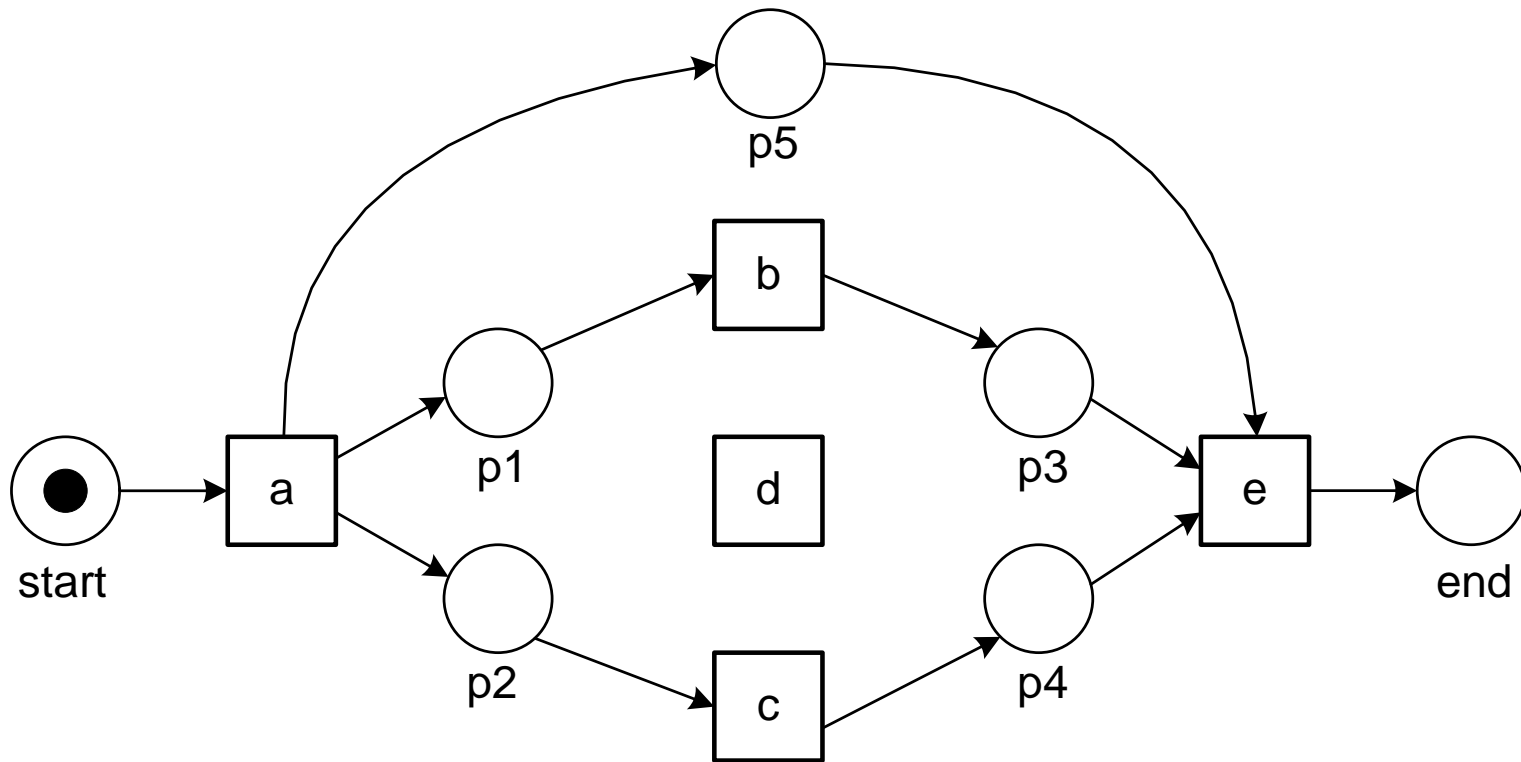
Heuristic mining

- To deal with noise and incompleteness.
- To have a better representational bias than the α algorithm (AND/XOR/OR/skip).
- Uses C-nets.



Example log; problem α algorithm

$$L = [\langle a, e \rangle^5, \langle a, b, c, e \rangle^{10}, \langle a, c, b, e \rangle^{10}, \langle a, b, e \rangle^1, \langle a, c, e \rangle^1, \langle a, d, e \rangle^{10}, \langle a, d, d, e \rangle^2, \langle a, d, d, d, e \rangle^1]$$



Taking into account frequencies

$$L = [\langle a, e \rangle^5, \langle a, b, c, e \rangle^{10}, \langle a, c, b, e \rangle^{10}, \langle a, b, e \rangle^1, \langle a, c, e \rangle^1, \langle a, d, e \rangle^{10}, \langle a, d, d, e \rangle^2, \langle a, d, d, d, e \rangle^1]$$

$$|a \rangle_L b| = \sum_{\sigma \in L} L(\sigma) \times |\{1 \leq i < |\sigma| \mid \sigma(i) = a \wedge \sigma(i+1) = b\}|$$

$ \rangle_L $	a	b	c	d	e
a	0	11	11	13	5
b	0	0	10	0	11
c	0	10	0	0	11
d	0	0	0	4	13
e	0	0	0	0	0

Dependency measure

$$|a \succ_L b| = \sum_{\sigma \in L} L(\sigma) \times |\{1 \leq i < |\sigma| \mid \sigma(i) = a \wedge \sigma(i+1) = b\}|$$

$|a \Rightarrow_L b|$ is the value of the dependency relation between a and b :

$$|a \Rightarrow_L b| = \begin{cases} \frac{|a \succ_L b| - |b \succ_L a|}{|a \succ_L b| + |b \succ_L a| + 1} & \text{if } a \neq b \\ \frac{|a \succ_L a|}{|a \succ_L a| + 1} & \text{if } a = b \end{cases}$$

Example

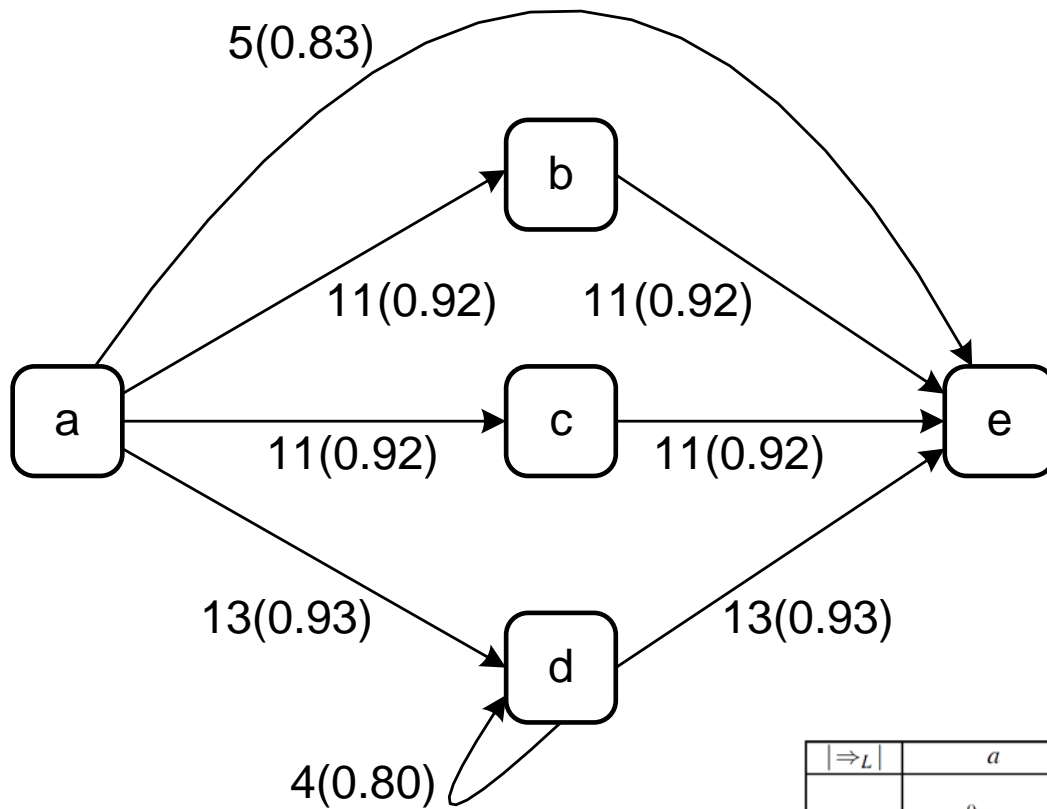
$ \Rightarrow_L $	a	b	c	d	e
a	$\frac{0}{0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{13-0}{13+0+1} = 0.93$	$\frac{5-0}{5+0+1} = 0.83$
b	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0}{0+1} = 0$	$\frac{10-10}{10+10+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
c	$\frac{0-11}{0+11+1} = -0.92$	$\frac{10-10}{10+10+1} = 0$	$\frac{0}{0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
d	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0-0}{0+0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{4}{4+1} = 0.80$	$\frac{13-0}{13+0+1} = 0.93$
e	$\frac{0-5}{0+5+1} = -0.83$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0}{0+1} = 0$

$|a \Rightarrow_L b|$ is the value of the dependency relation between a and b :

$$|a \Rightarrow_L b| = \begin{cases} \frac{|a >_L b| - |b >_L a|}{|a >_L b| + |b >_L a| + 1} & \text{if } a \neq b \\ \frac{|a >_L a|}{|a >_L a| + 1} & \text{if } a = b \end{cases}$$

$ \>_L $	a	b	c	d	e
a	0	11	11	13	5
b	0	0	10	0	11
c	0	10	0	0	11
d	0	0	0	4	13
e	0	0	0	0	0

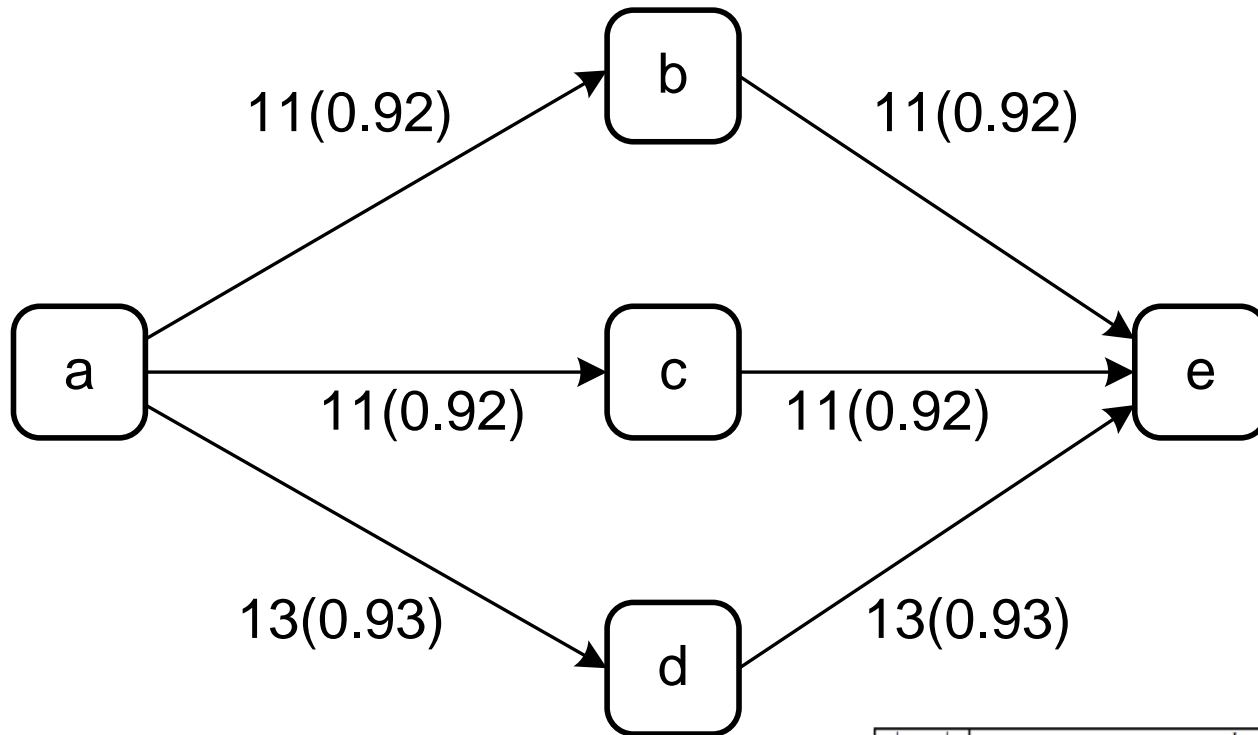
Lower threshold (2 direct successions and a dependency of at least 0.7)



$ \succ_L $	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>
<i>a</i>	0	11	11	13	5
<i>b</i>	0	0	10	0	11
<i>c</i>	0	10	0	0	11
<i>d</i>	0	0	0	4	13
<i>e</i>	0	0	0	0	0

$ \Rightarrow_L $	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>
<i>a</i>	$\frac{0}{0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{13-0}{13+0+1} = 0.93$	$\frac{5-0}{5+0+1} = 0.83$
<i>b</i>	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0}{0+1} = 0$	$\frac{10-10}{10+10+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
<i>c</i>	$\frac{0-11}{0+11+1} = -0.92$	$\frac{10-10}{10+10+1} = 0$	$\frac{0}{0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
<i>d</i>	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0-0}{0+0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{4}{4+1} = 0.80$	$\frac{13-0}{13+0+1} = 0.93$
<i>e</i>	$\frac{0-5}{0+5+1} = -0.83$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0}{0+1} = 0$

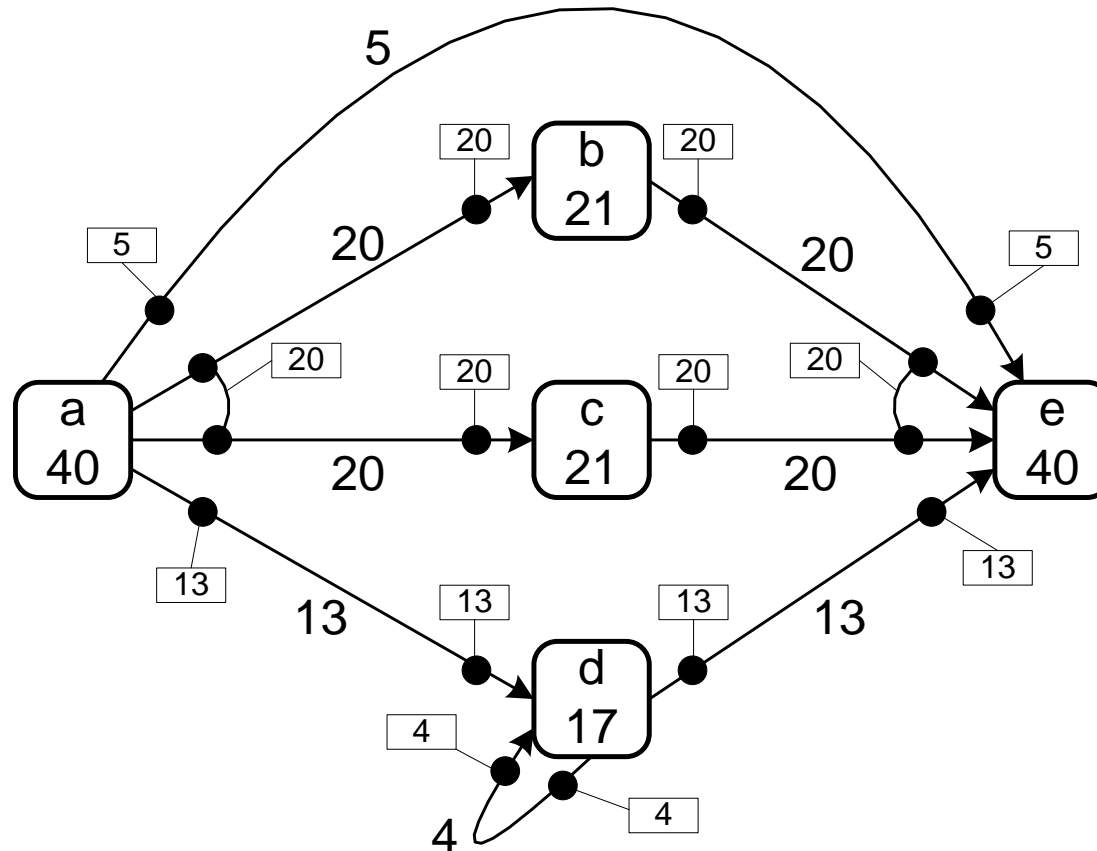
Higher threshold (5 direct successions and a dependency of at least 0.9)



$ \succ_L $	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>
<i>a</i>	0	11	11	13	5
<i>b</i>	0	0	10	0	11
<i>c</i>	0	10	0	0	11
<i>d</i>	0	0	0	4	13
<i>e</i>	0	0	0	0	0

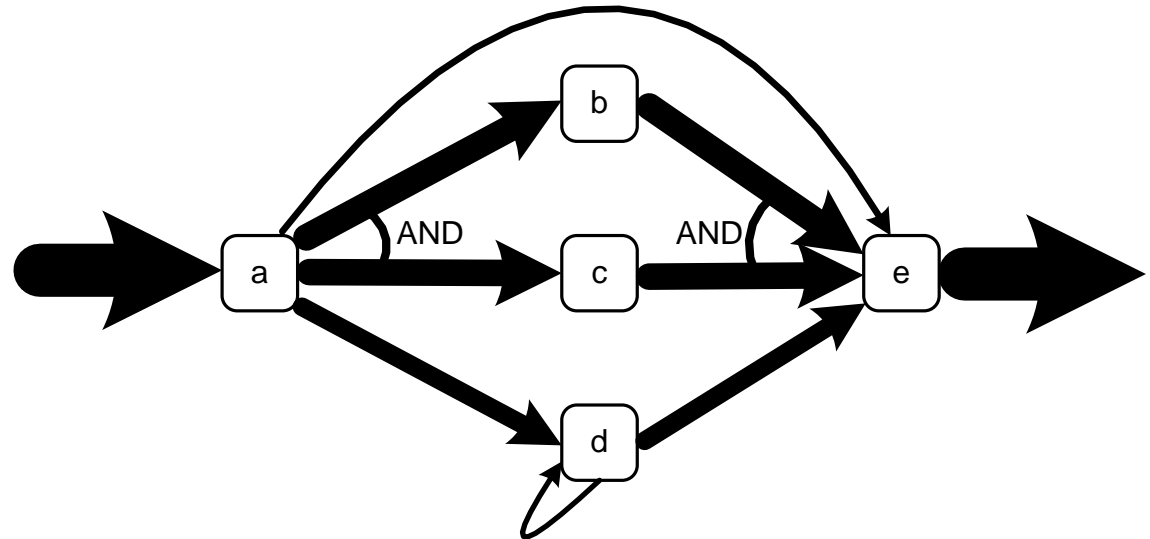
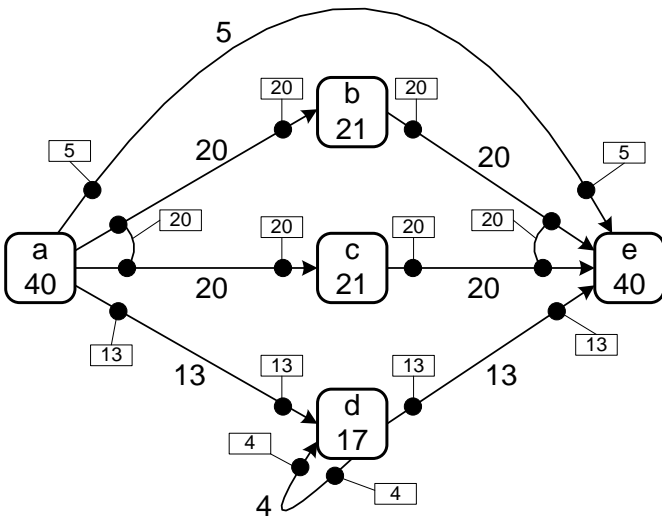
$ \Rightarrow_L $	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>
<i>a</i>	$\frac{0}{0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{13-0}{13+0+1} = 0.93$	$\frac{5-0}{5+0+1} = 0.83$
<i>b</i>	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0}{0+1} = 0$	$\frac{10-10}{10+10+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
<i>c</i>	$\frac{0-11}{0+11+1} = -0.92$	$\frac{10-10}{10+10+1} = 0$	$\frac{0}{0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
<i>d</i>	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0-0}{0+0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{4}{4+1} = 0.80$	$\frac{13-0}{13+0+1} = 0.93$
<i>e</i>	$\frac{0-5}{0+5+1} = -0.83$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0}{0+1} = 0$

Learning splits and joins



$$L = [\langle a, e \rangle^5, \langle a, b, c, e \rangle^{10}, \langle a, c, b, e \rangle^{10}, \langle a, b, e \rangle^1, \langle a, c, e \rangle^1, \langle a, d, e \rangle^{10}, \langle a, d, d, e \rangle^2, \langle a, d, d, d, e \rangle^1]$$

Alternative visualization

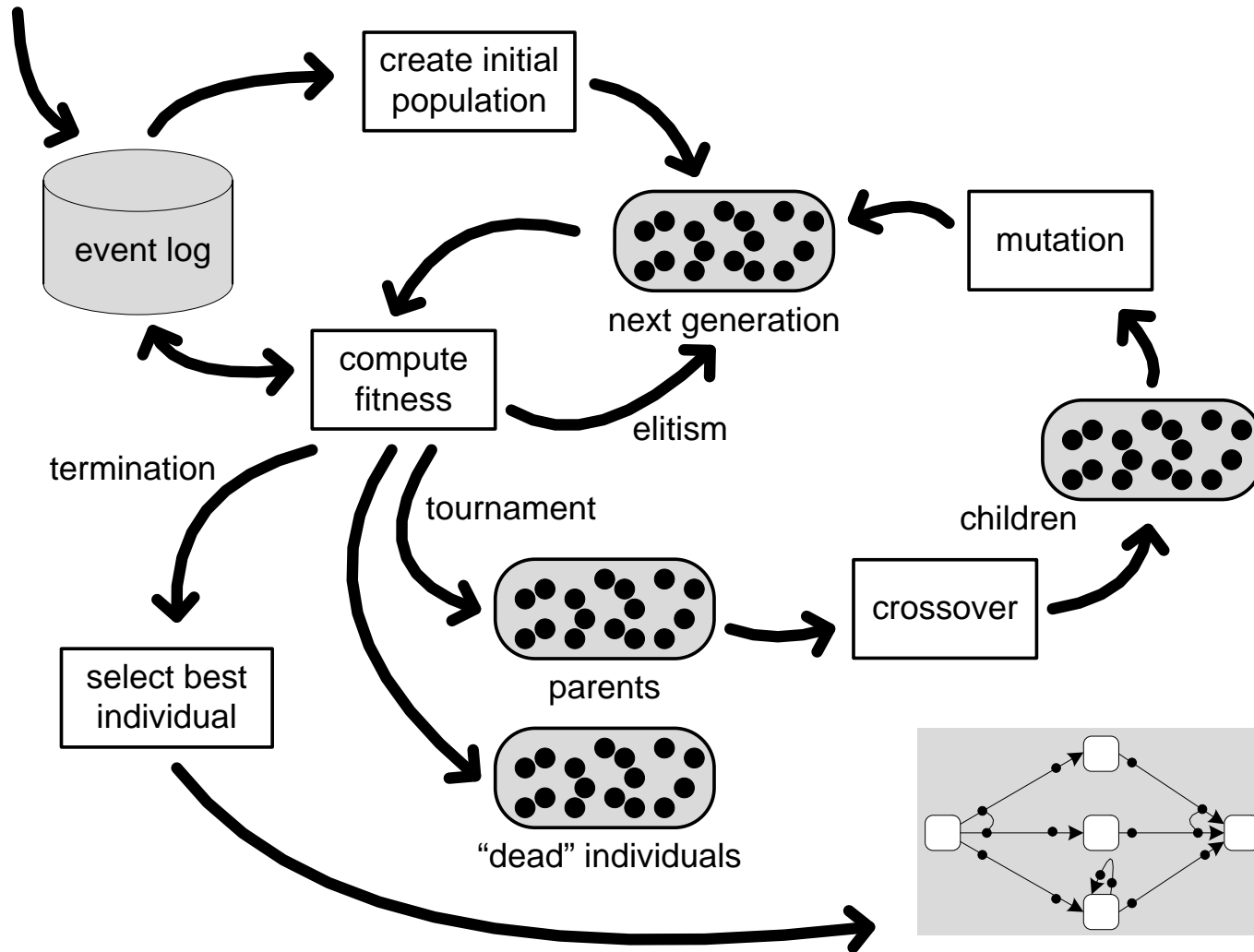


$$L = [\langle a, e \rangle^5, \langle a, b, c, e \rangle^{10}, \langle a, c, b, e \rangle^{10}, \langle a, b, e \rangle^1, \langle a, c, e \rangle^1, \langle a, d, e \rangle^{10}, \langle a, d, d, e \rangle^2, \langle a, d, d, d, e \rangle^1]$$

Characteristics of heuristic mining

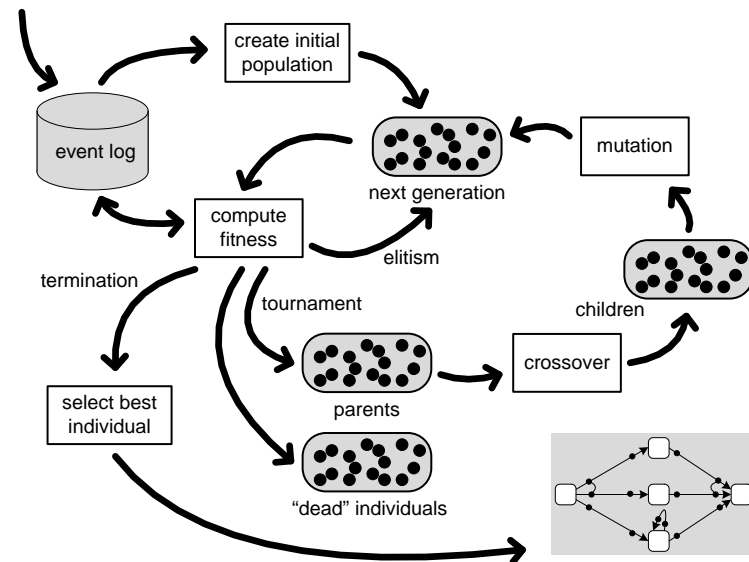
- **Can deal with noise and therefore quite robust.**
- **Improved representational bias.**
- **Split and join rules are only considered locally (therefore most of the discovered model are not sound and require repair actions).**

Genetic process mining

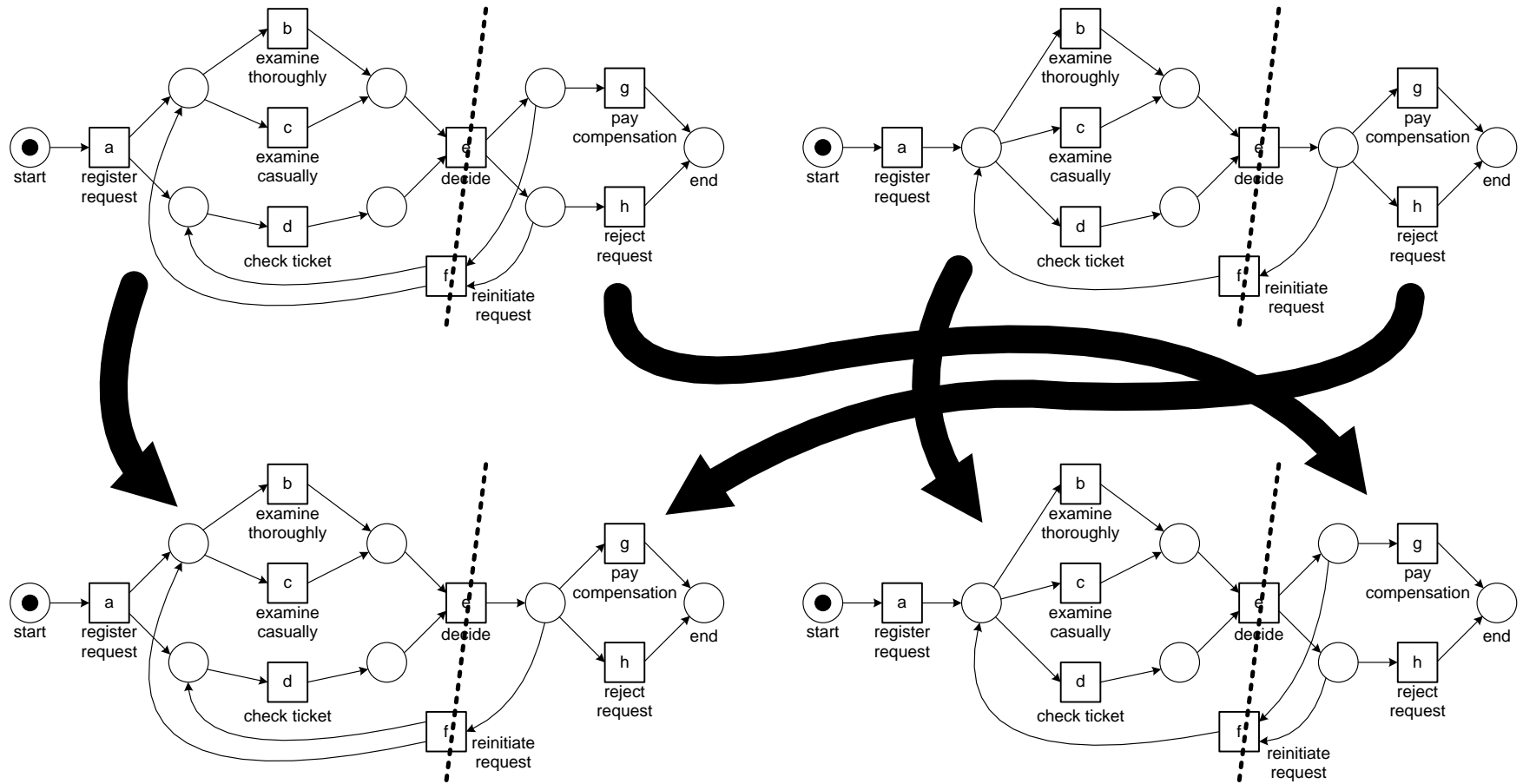


Design decisions

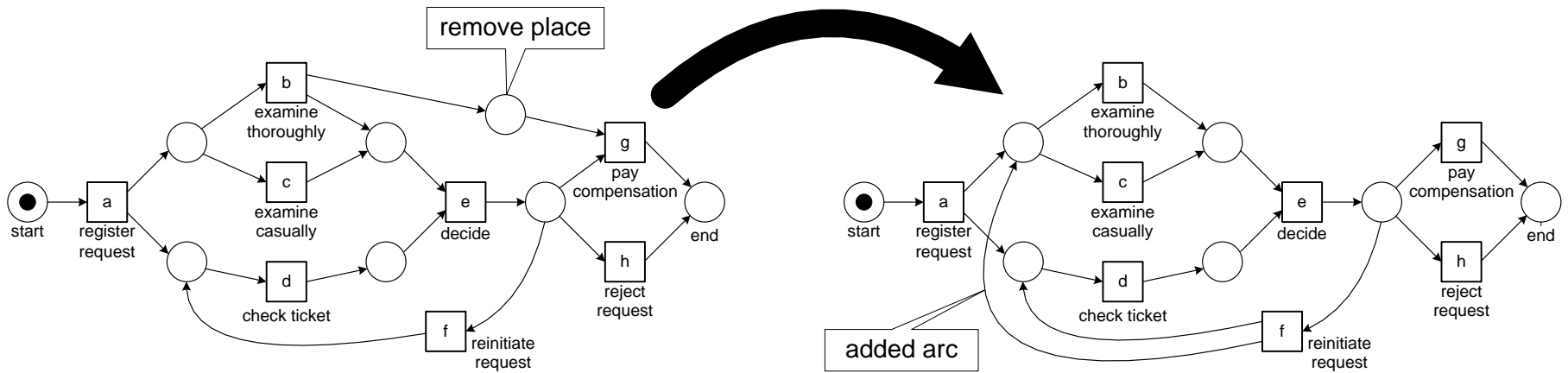
- Representation of individuals
- Initialization
- Fitness function
- Selection strategy (tournament and elitism)
- Crossover
- Mutation



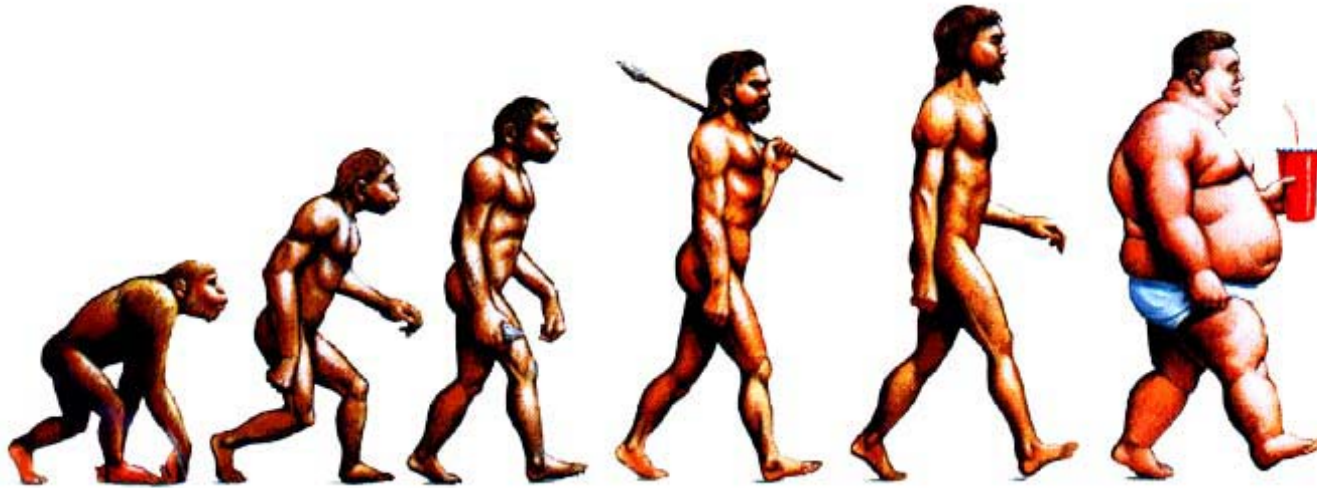
Example: crossover



Example: mutation



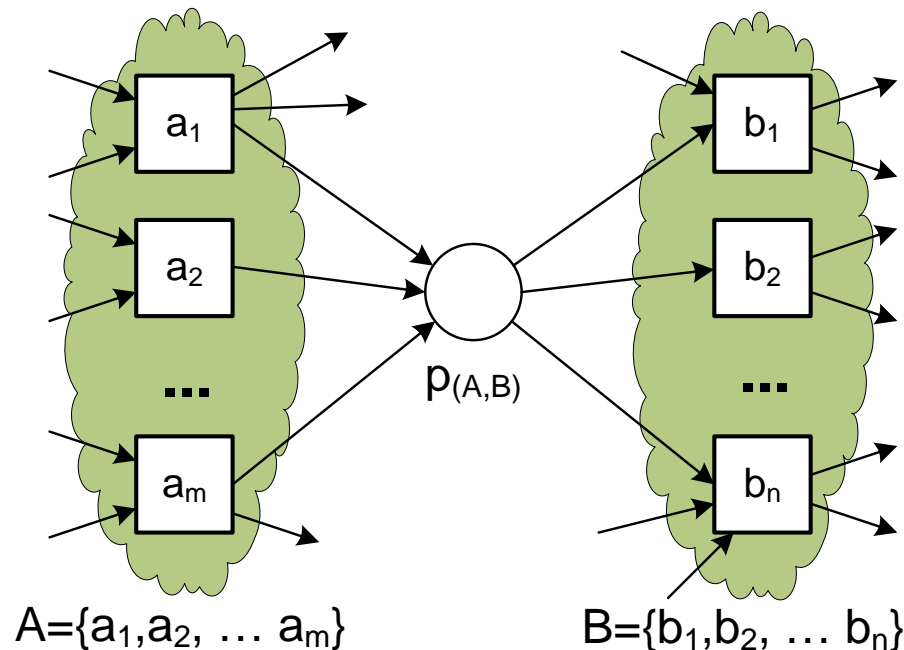
Characteristics of genetic process mining



- Requires a lot of computing power.
- Can be distributed easily.
- Can deal with noise, infrequent behavior, duplicate tasks, invisible tasks, etc.
- Allows for incremental improvement and combinations with other approaches (heuristics post-optimization, etc.).

Region-based mining

- **Two types of regions theory:**
 - State-based regions
 - Language-based regions
- **All about discovering places (like in the α algorithm)!**

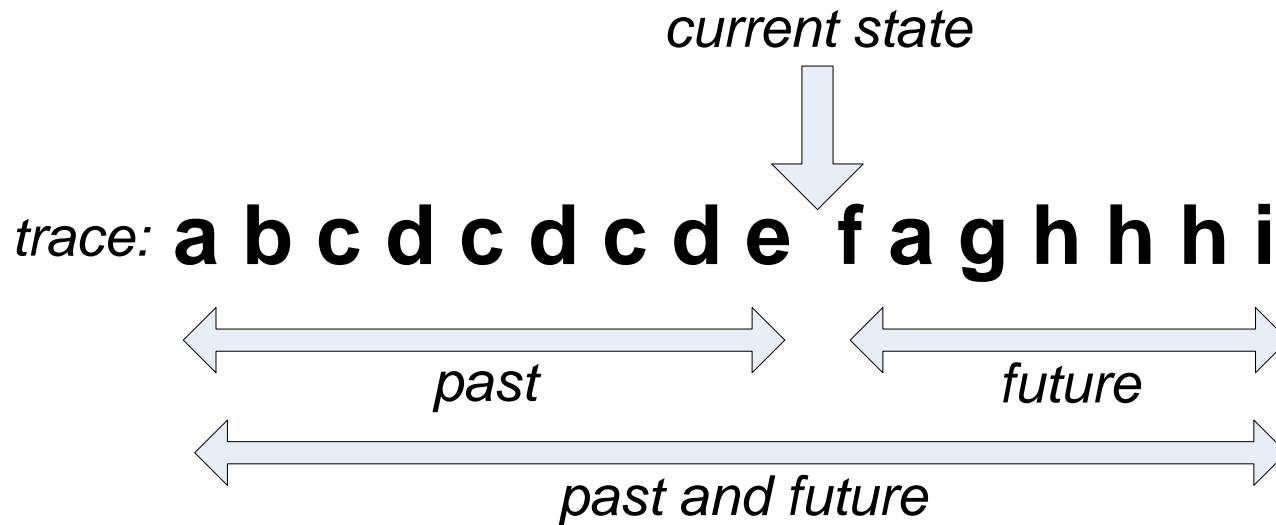


State-based regions

Two steps:

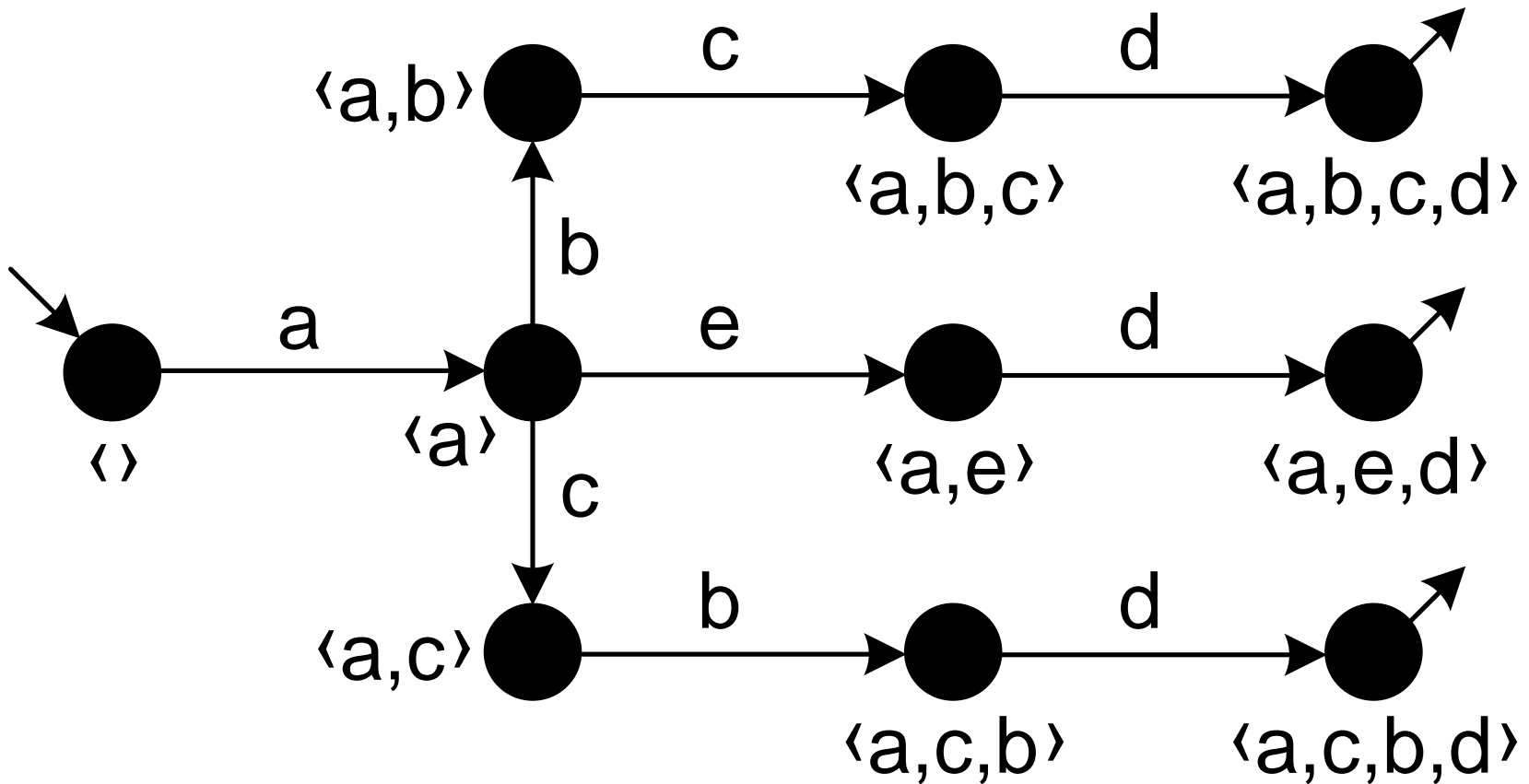
- 1. Discover a transition system (different abstractions are possible)**
- 2. Convert transition system into an “equivalent” Petri net.**

Step 1: learning a transition system



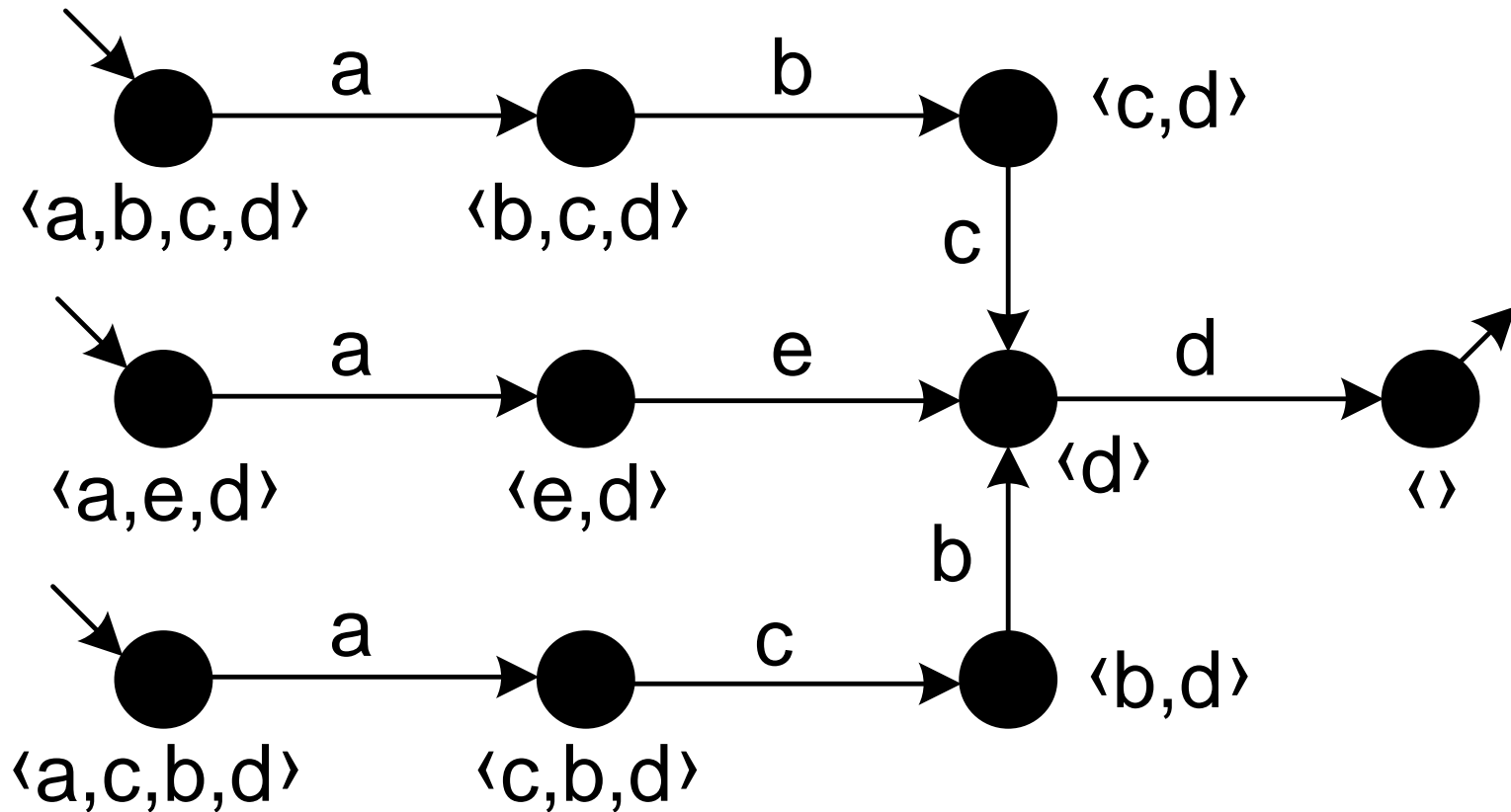
- **past, future, past+future**
- **sequence, multiset, set abstraction**
- **limited horizon to abstract further**
- **filtering e.g. based on transaction type, names, etc.**
- **labels based on activity name or other features**

Past without abstraction (full sequence)



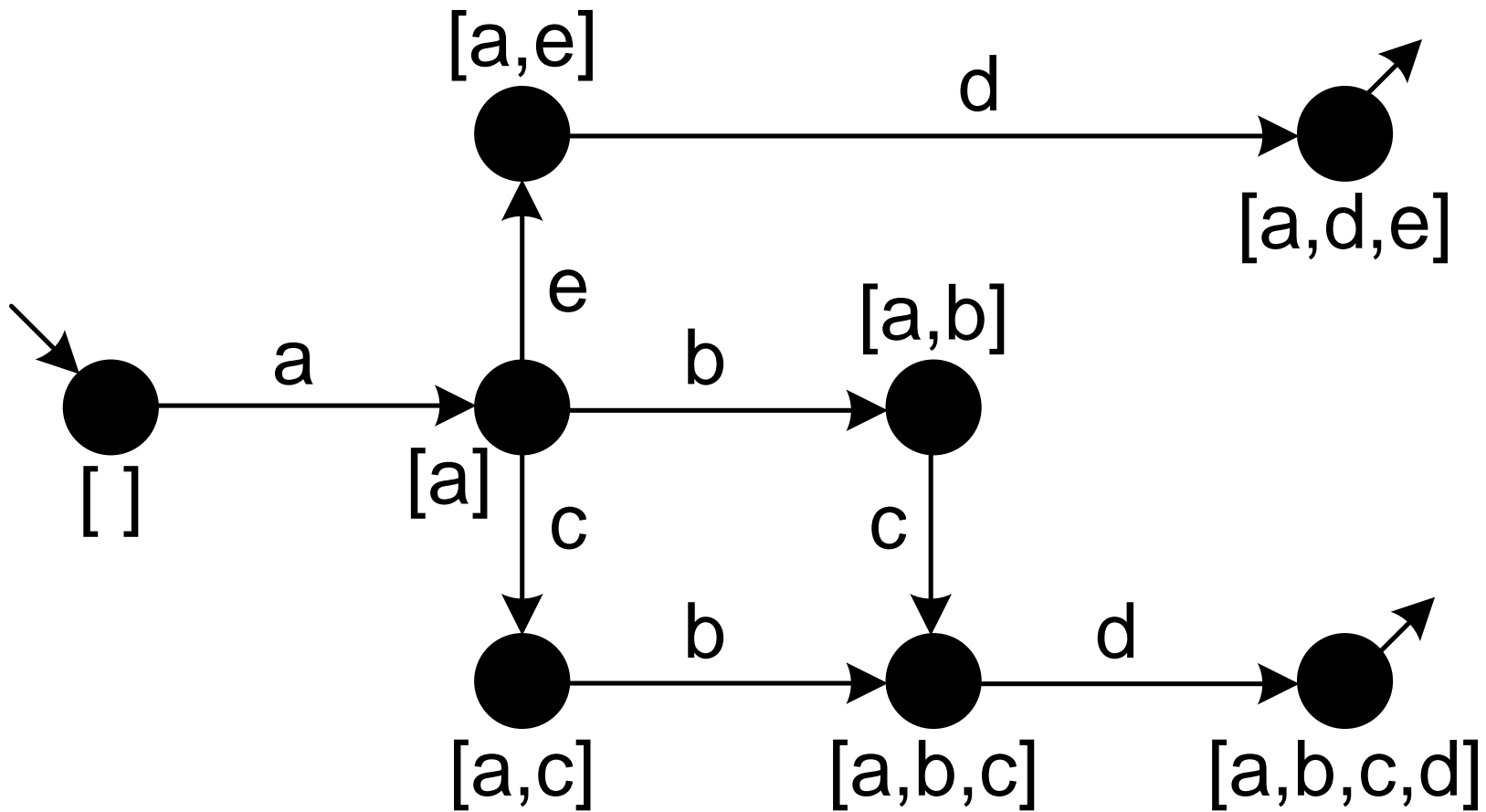
$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

Future without abstraction



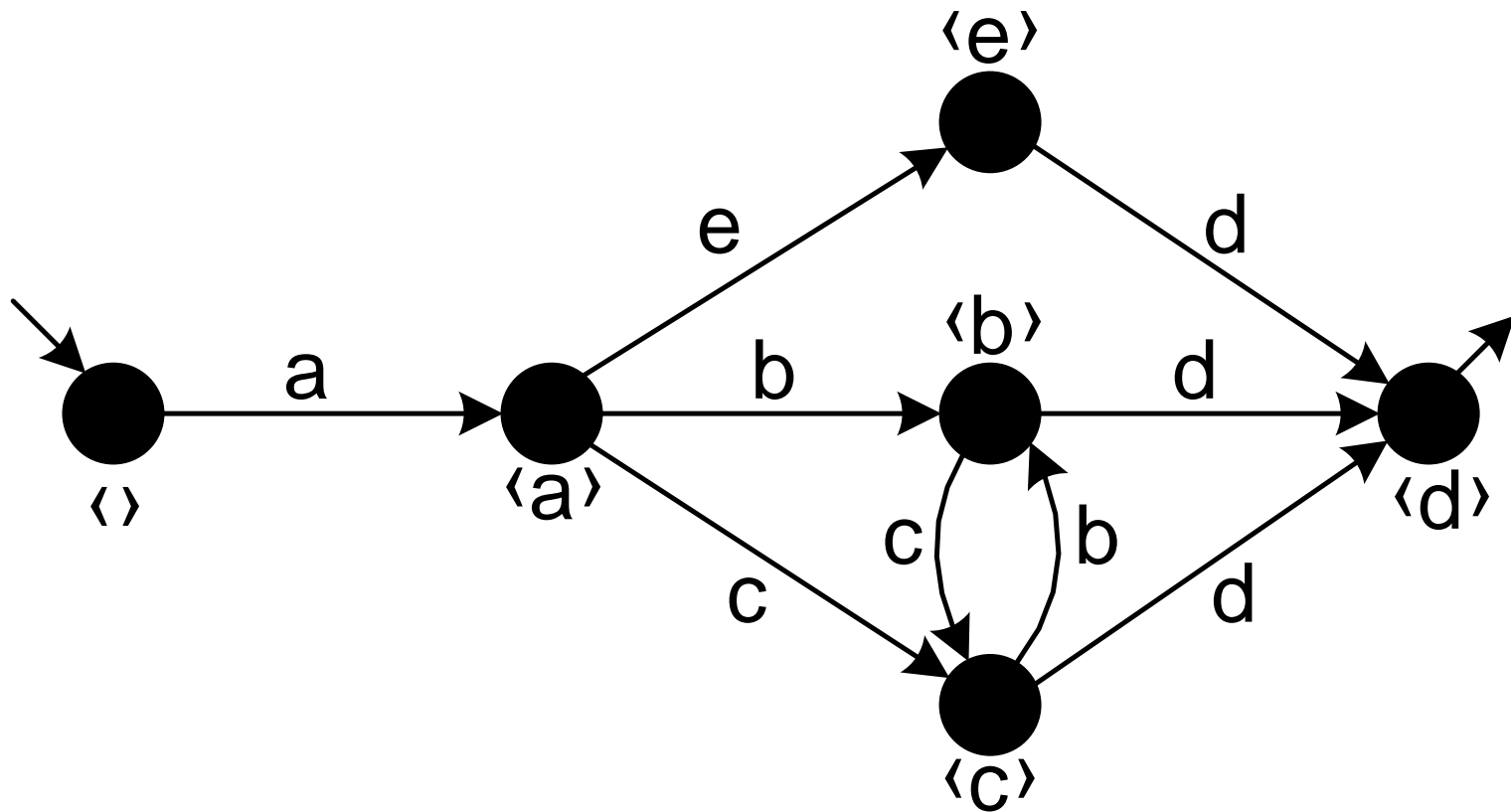
$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

Past with multiset abstraction



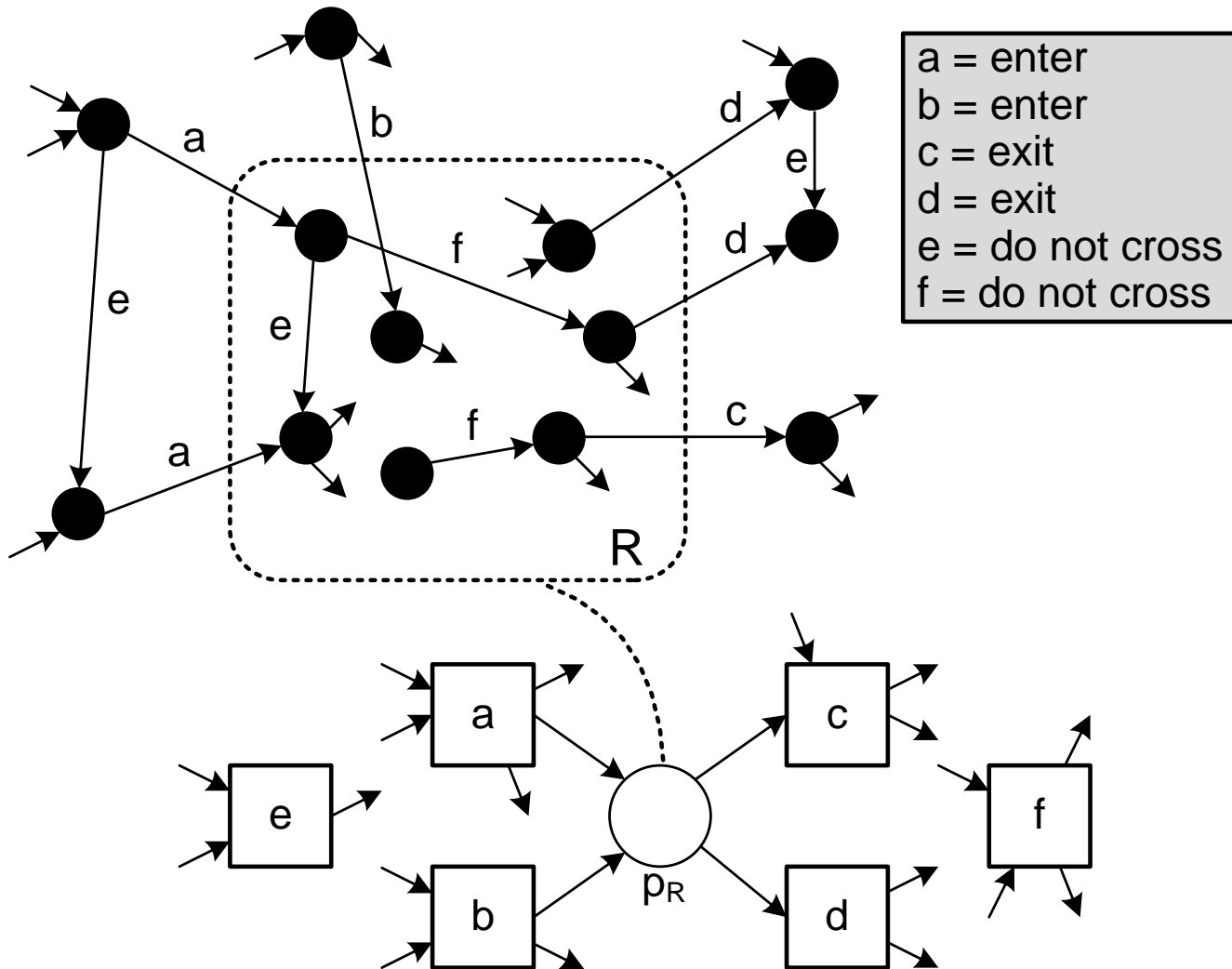
$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

Only last event matters for state

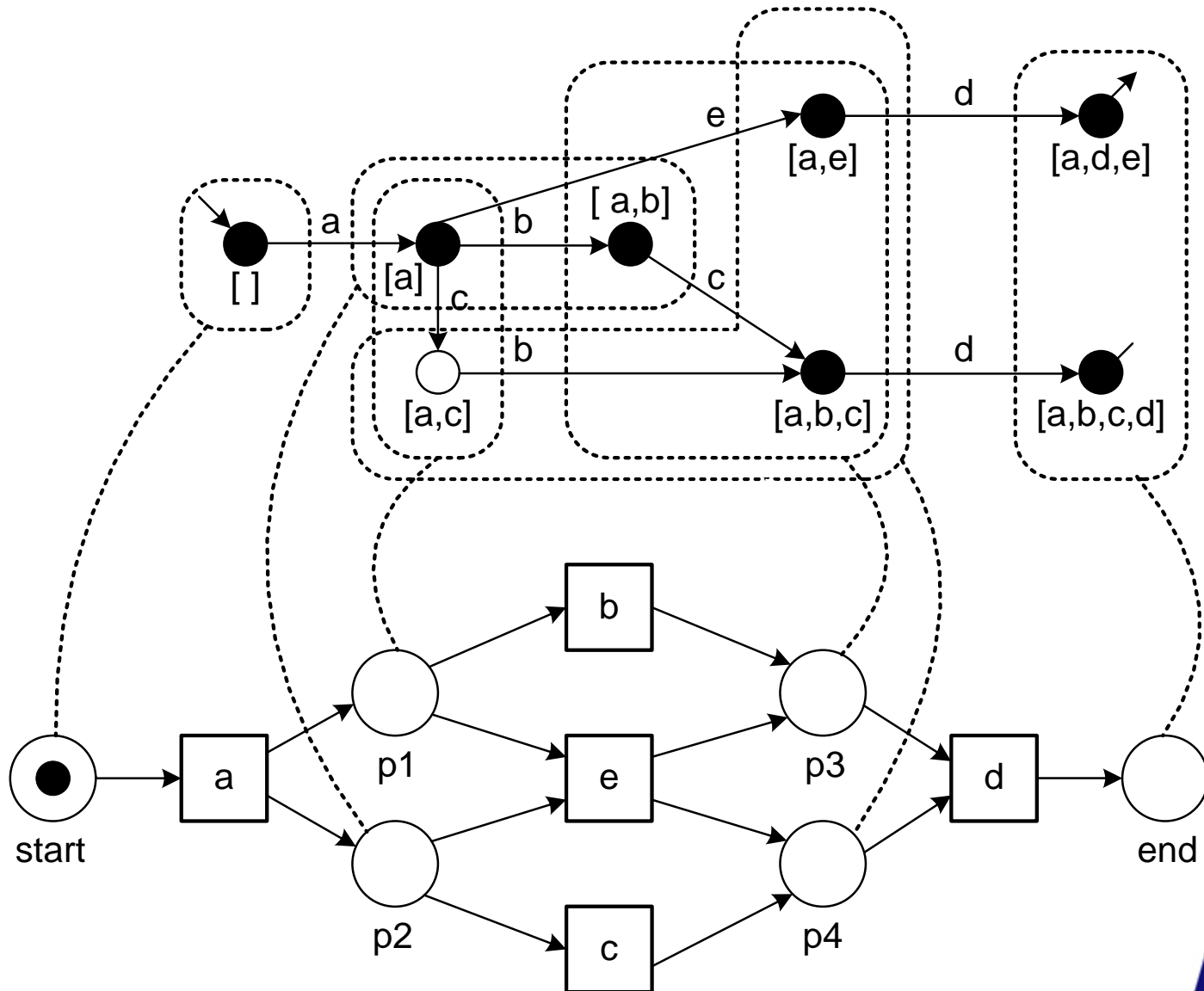


$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

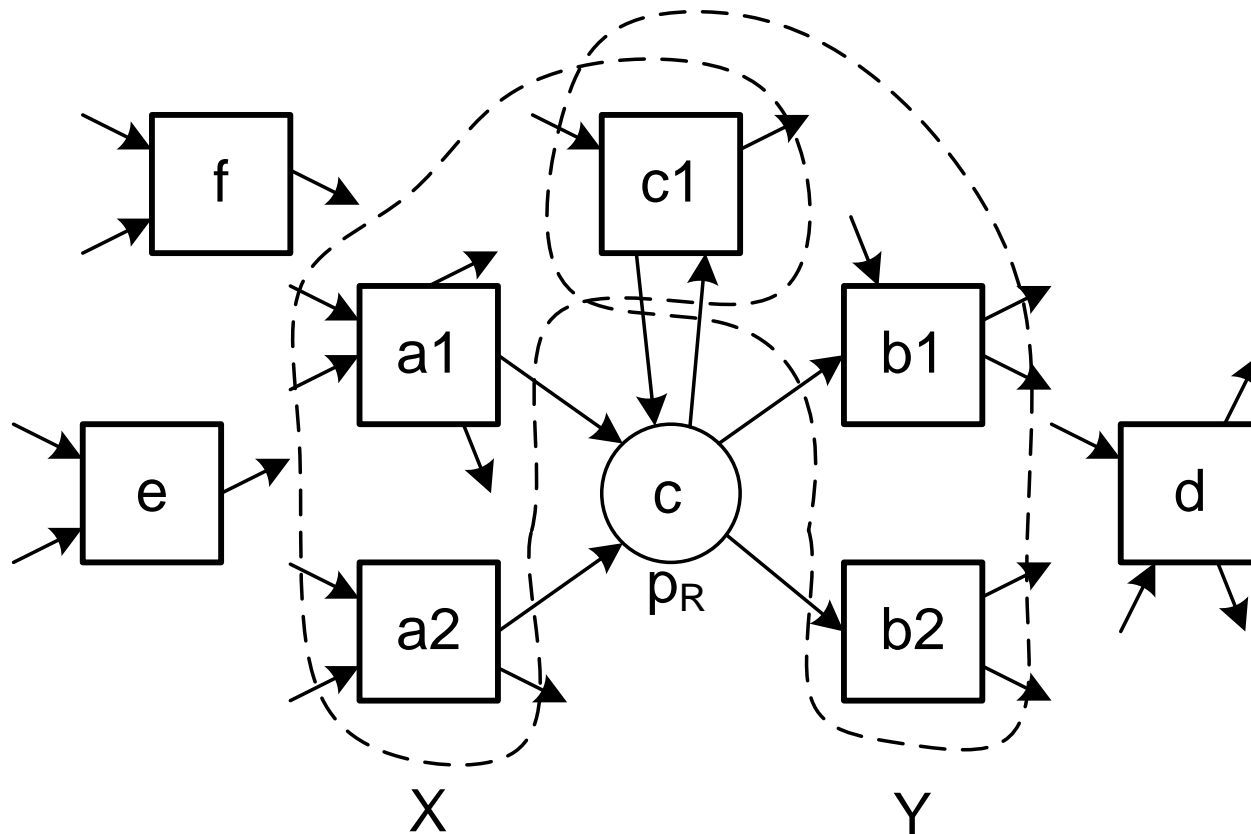
Step 2: constructing a Petri net using regions



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

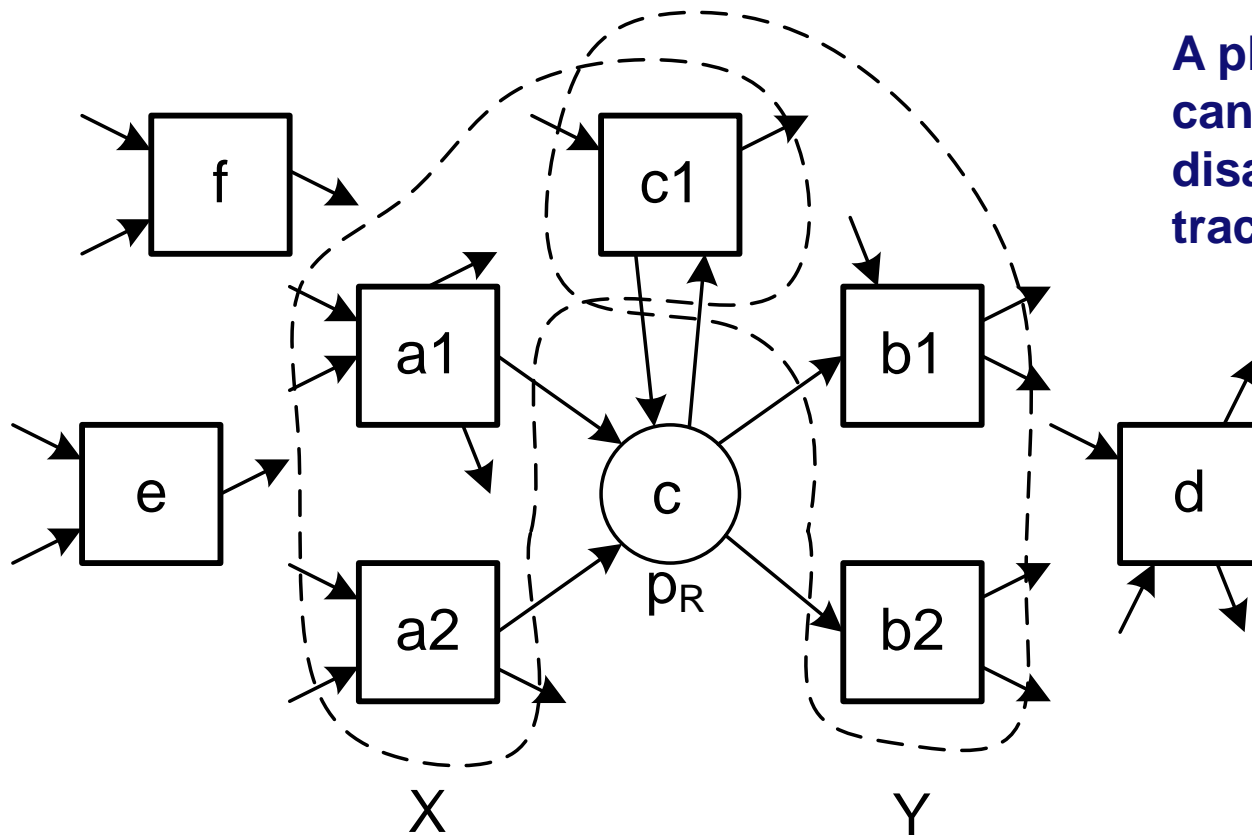


Language based regions



Region $R = (X, Y, c)$ corresponding to place p_R : $X = \{a1, a2, c1\}$ = transitions producing a token for p_R , $Y = \{b1, b2, c1\}$ = transitions consuming a token from p_R , and c is the initial marking of p_R .

Based idea: enough tokens should be present when consuming



A place is **feasible** if it can be added without disabling any of the traces in the event log.

for any $\sigma \in L, k \in \{1, \dots, |\sigma|\}, \sigma_1 = hd^{k-1}(\sigma), a = \sigma(k), \sigma_2 = hd^k(\sigma) = \sigma_1 \oplus a$:

$$c + \sum_{t \in X} \partial_{multiset}(\sigma_1)(t) - \sum_{t \in Y} \partial_{multiset}(\sigma_2)(t) \geq 0.$$

Example

$$L_9 = [\langle a, c, d \rangle^{45}, \langle b, c, e \rangle^{42}]$$

$$c - y_a \geq 0$$

$$c + x_a - (y_a + y_c) \geq 0$$

$$c + x_a + x_c - (y_a + y_c + y_d) \geq 0$$

$$c - y_b \geq 0$$

$$c + x_b - (y_b + y_c) \geq 0$$

$$c + x_b + x_c - (y_b + y_c + y_e) \geq 0$$

$$c, x_a, \dots, x_e, y_a, \dots, y_e \in \{0, 1\}$$

Regions

$$L_9 = [\langle a, c, d \rangle^{45}, \langle b, c, e \rangle^{42}]$$

$$R_1 = (\emptyset, \{a, b\}, 1)$$

$$c = y_a = y_b = 1, \quad x_a = x_b = x_c = x_d = x_e = y_c = y_d = y_e = 0$$

$$R_2 = (\{a, b\}, \{c\}, 0)$$

$$x_a = x_b = y_c = 1, \quad c = x_c = x_d = x_e = y_a = y_b = y_d = y_e = 0$$

$$R_3 = (\{c\}, \{d, e\}, 0)$$

$$x_c = y_d = y_e = 1, \quad c = x_a = x_b = x_d = x_e = y_a = y_b = y_c = 0$$

$$R_4 = (\{d, e\}, \emptyset, 0)$$

$$x_d = x_e = 1, \quad c = x_a = x_b = x_c = y_a = y_b = y_c = y_d = y_e = 0$$

$$R_5 = (\{a\}, \{d\}, 0)$$

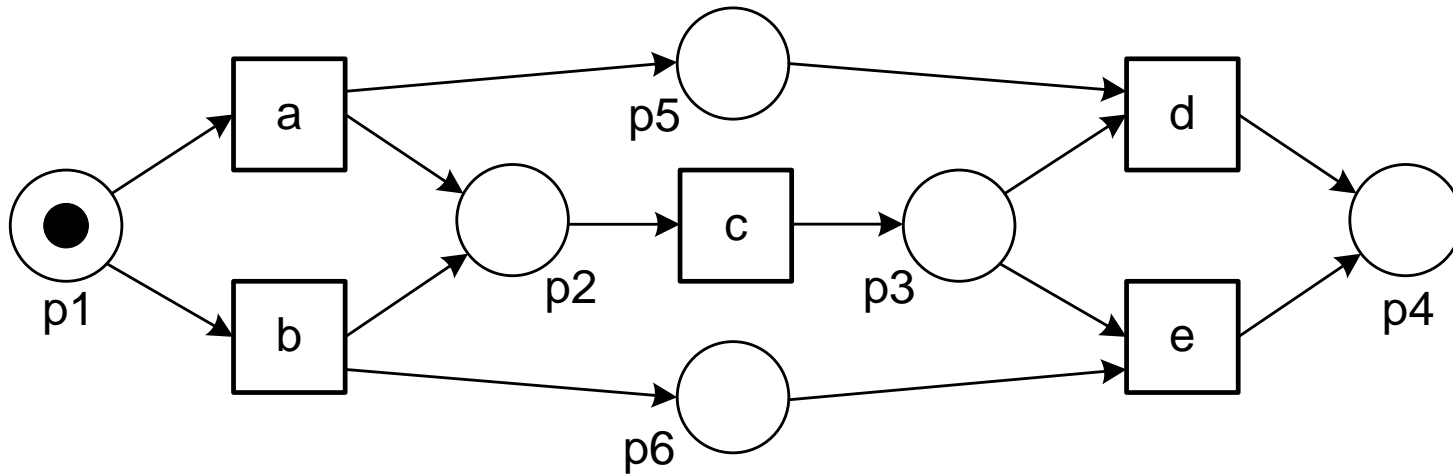
$$x_a = y_d = 1, \quad c = x_b = x_c = x_d = x_e = y_a = y_b = y_c = y_e = 0$$

$$R_6 = (\{b\}, \{e\}, 0)$$

$$x_b = y_e = 1, \quad c = x_a = x_c = x_d = x_e = y_a = y_b = y_c = y_d = 0$$

Model

$$L_9 = [\langle a, c, d \rangle^{45}, \langle b, c, e \rangle^{42}]$$



$$R_1 = (\emptyset, \{a, b\}, 1)$$

$$c = y_a = y_b = 1, \quad x_a = x_b = x_c = x_d = x_e = y_c = y_d = y_e = 0$$

$$R_2 = (\{a, b\}, \{c\}, 0)$$

$$x_a = x_b = y_c = 1, \quad c = x_c = x_d = x_e = y_a = y_b = y_d = y_e = 0$$

$$R_3 = (\{c\}, \{d, e\}, 0)$$

$$x_c = y_d = y_e = 1, \quad c = x_a = x_b = x_d = x_e = y_a = y_b = y_c = 0$$

$$R_4 = (\{d, e\}, \emptyset, 0)$$

$$x_d = x_e = 1, \quad c = x_a = x_b = x_c = y_a = y_b = y_c = y_d = y_e = 0$$

$$R_5 = (\{a\}, \{d\}, 0)$$

$$x_a = y_d = 1, \quad c = x_b = x_c = x_d = x_e = y_a = y_b = y_c = y_e = 0$$

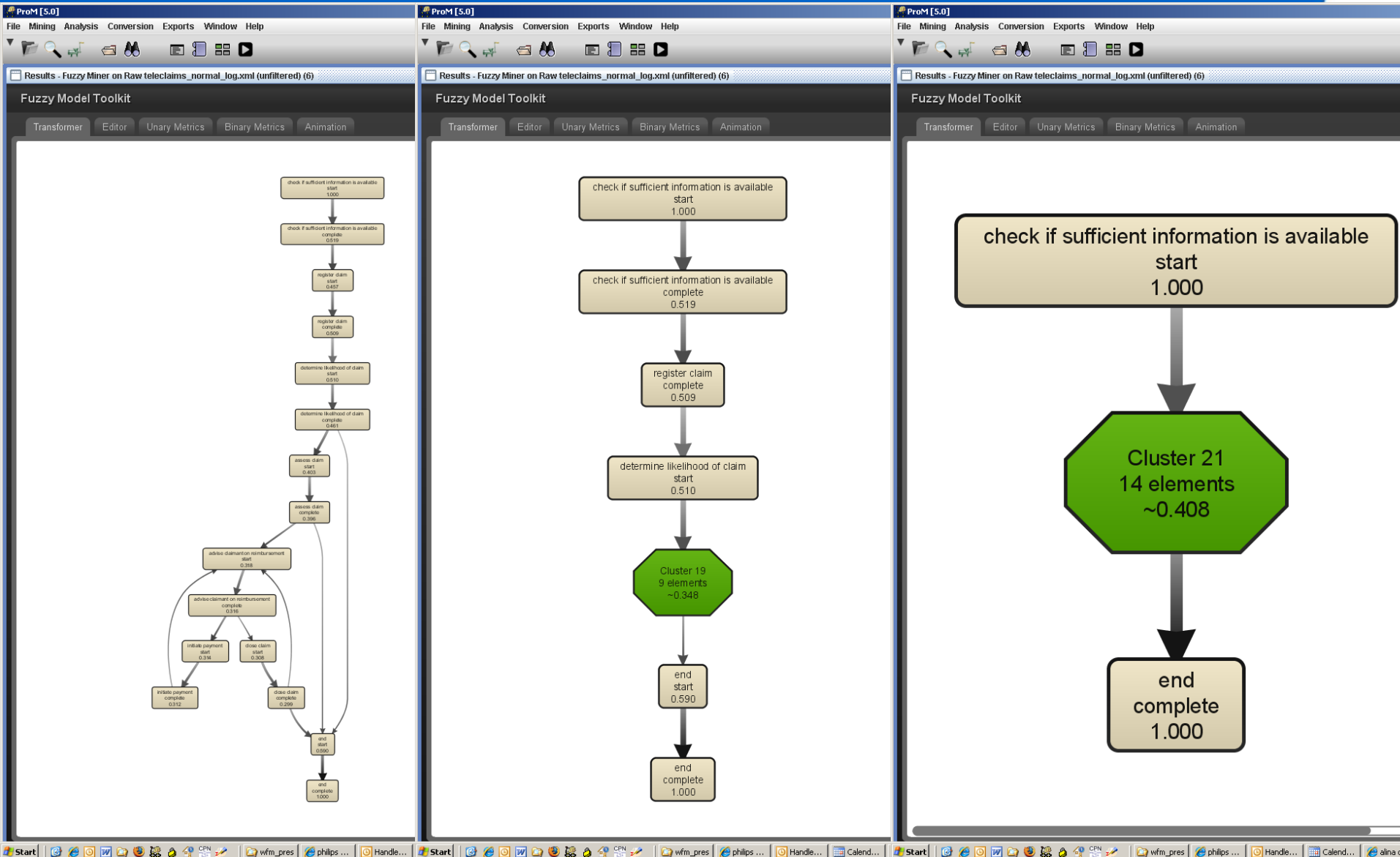
$$R_6 = (\{b\}, \{e\}, 0)$$

$$x_b = y_e = 1, \quad c = x_a = x_c = x_d = x_e = y_a = y_b = y_c = y_d = 0$$

Characteristics of region-based mining

- **Can be used to discover more complex control-flow structures.**
- **Classical approaches need to be adapted (overfitting!).**
- **Representational bias can be parameterized (e.g., free-choice nets, label splitting, etc.).**
- **Problems dealing with noise.**

Other approaches, e.g. fuzzy mining



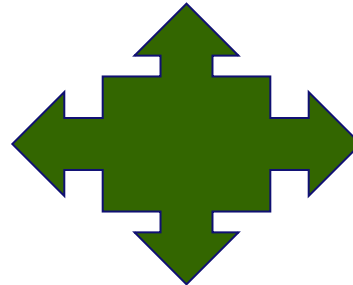
Evaluating the discovered process



Fitness: Is the event log possible according to the model?



Precision: Is the model not underfitting (allow for too much)?



Generalization: Is the model not overfitting (only allow for the “accidental” examples)?



Structure: Is this the simplest model (Occam's Razor)?

