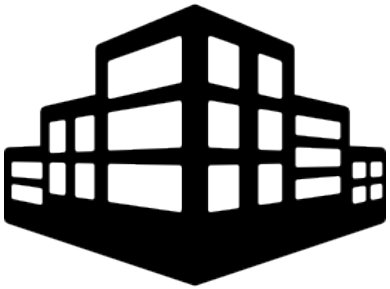
Avoiding Over-Fitting in Region-Based Process Discovery

AIS Meeting
S.J. van Zelst
March 19, 2015

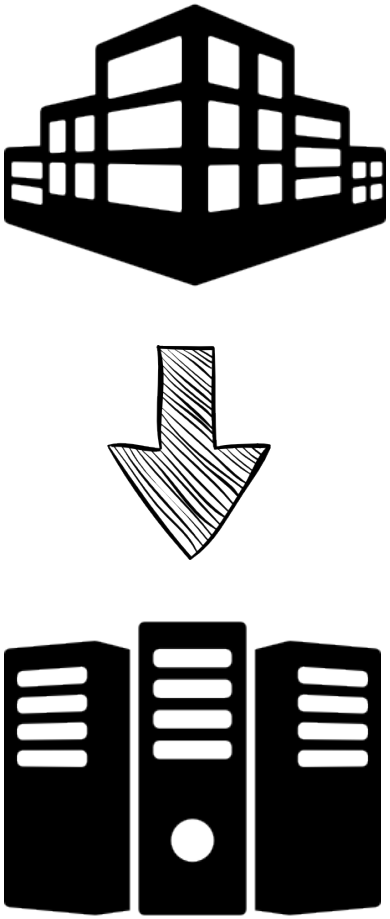
TU/e Technische Universiteit Eindhoven University of Technology

Where innovation starts

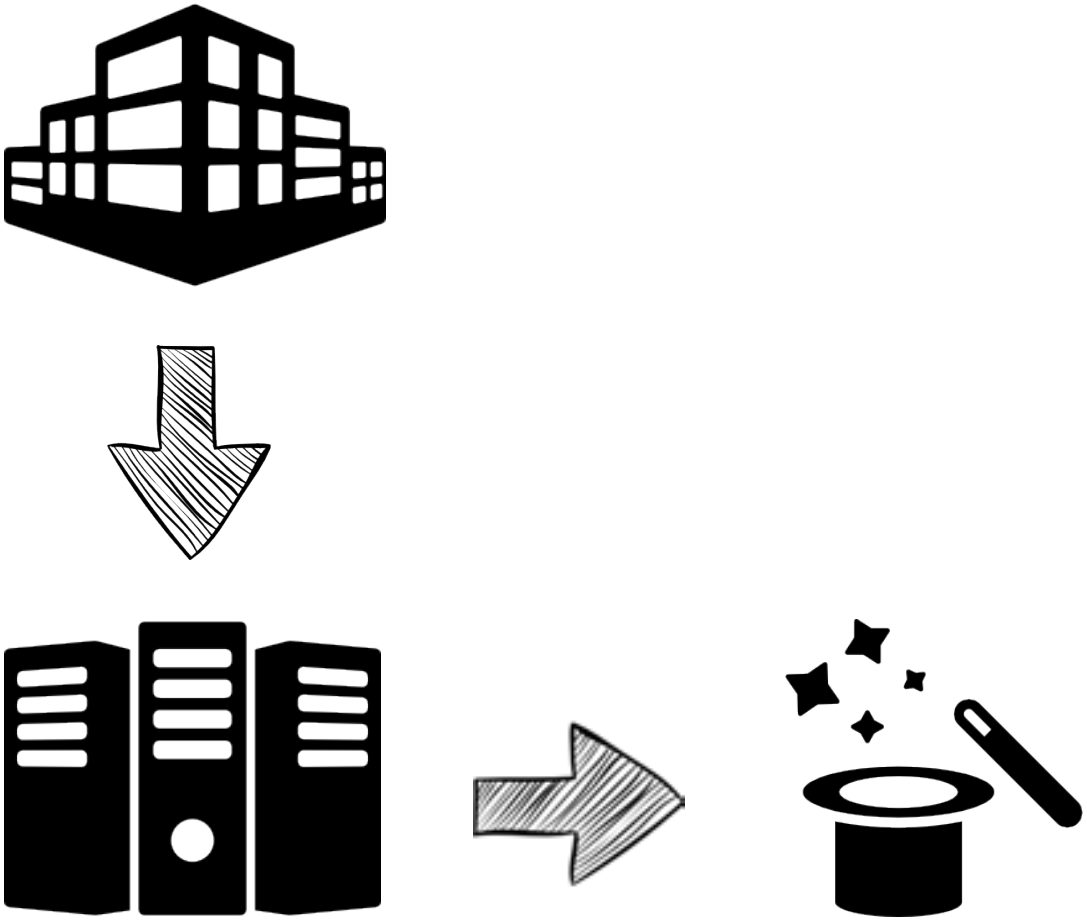
Setting the Scene
Process Discovery



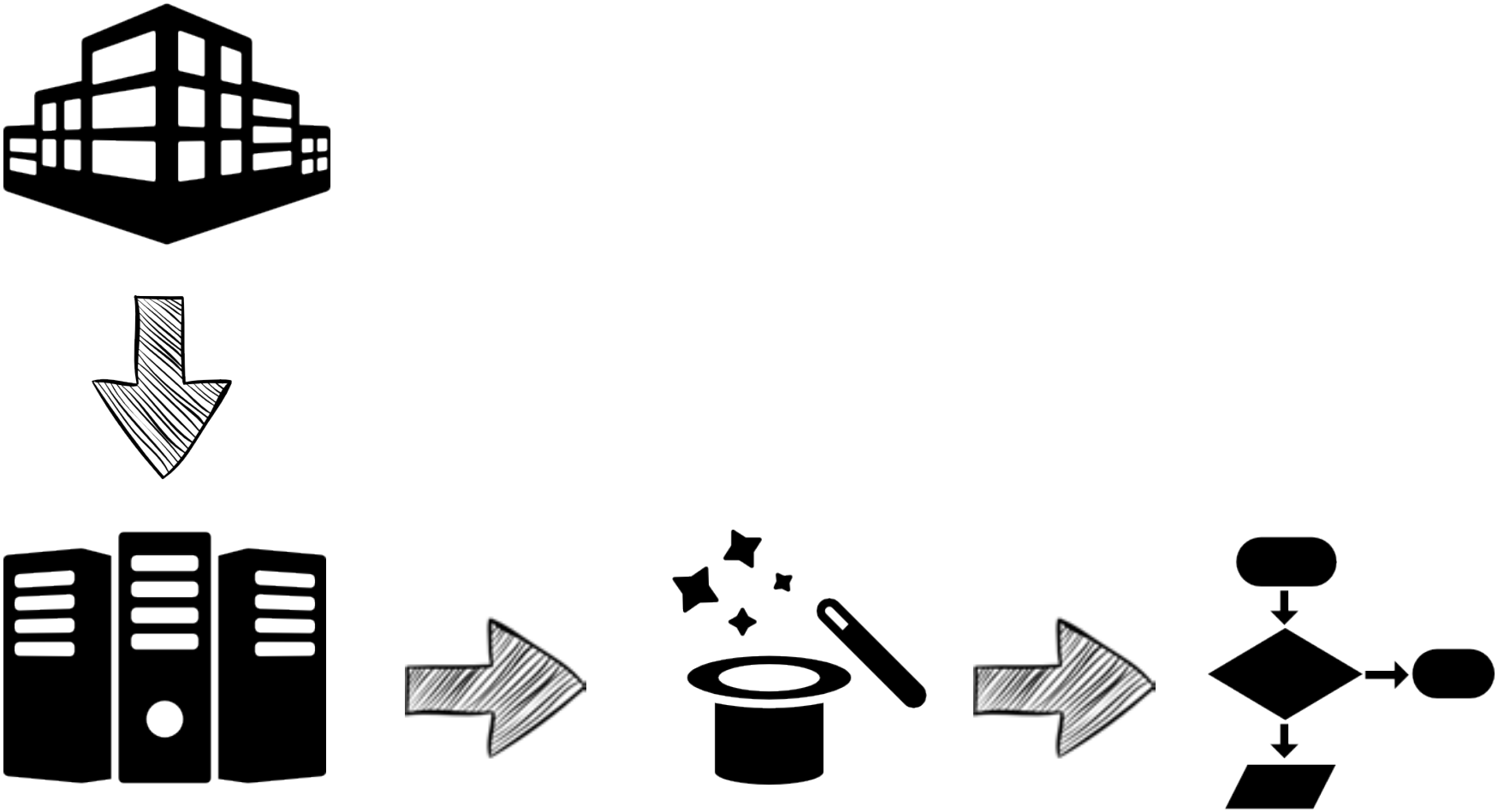
Setting the Scene
Process Discovery



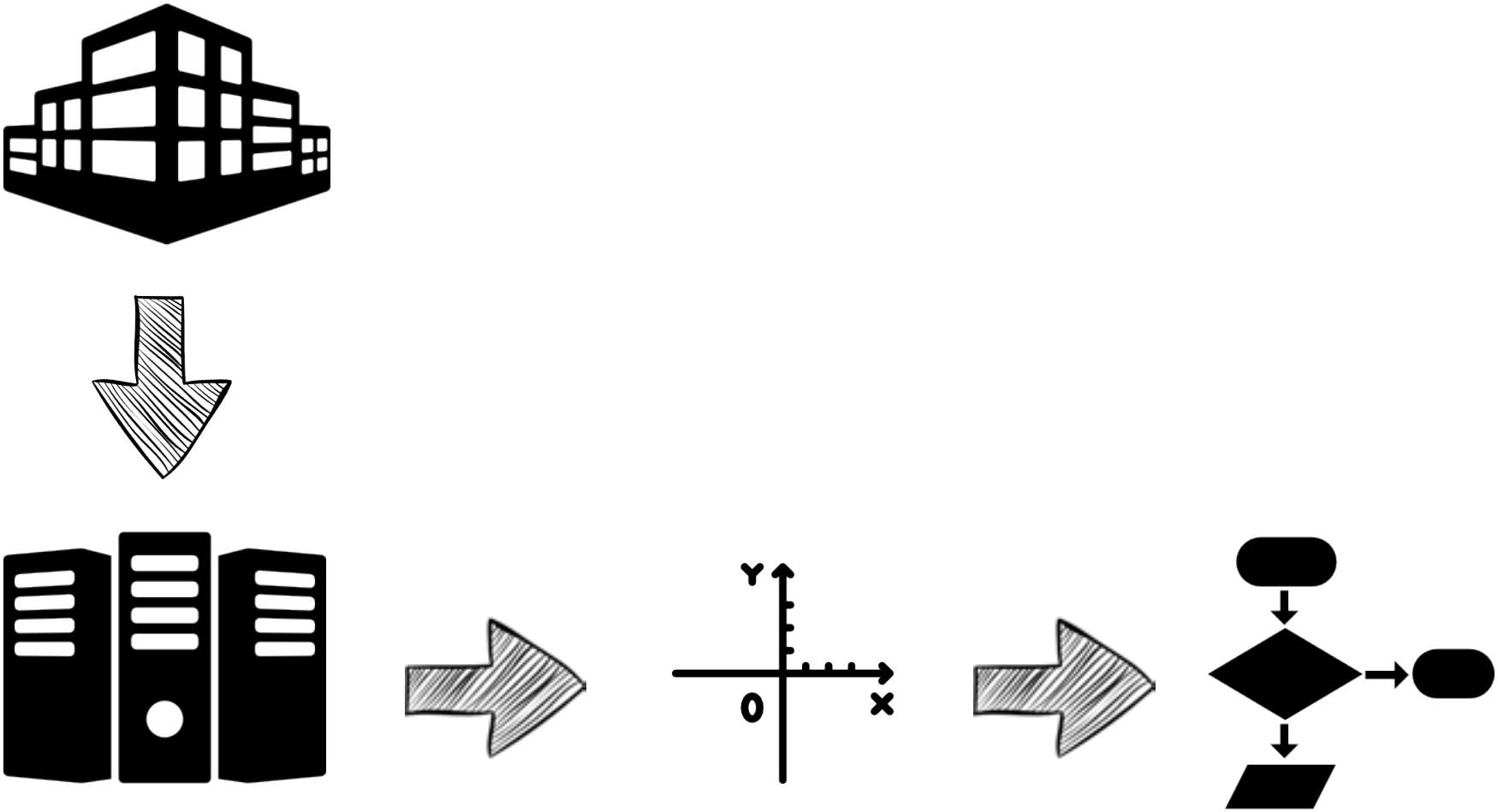
Setting the Scene
Process Discovery



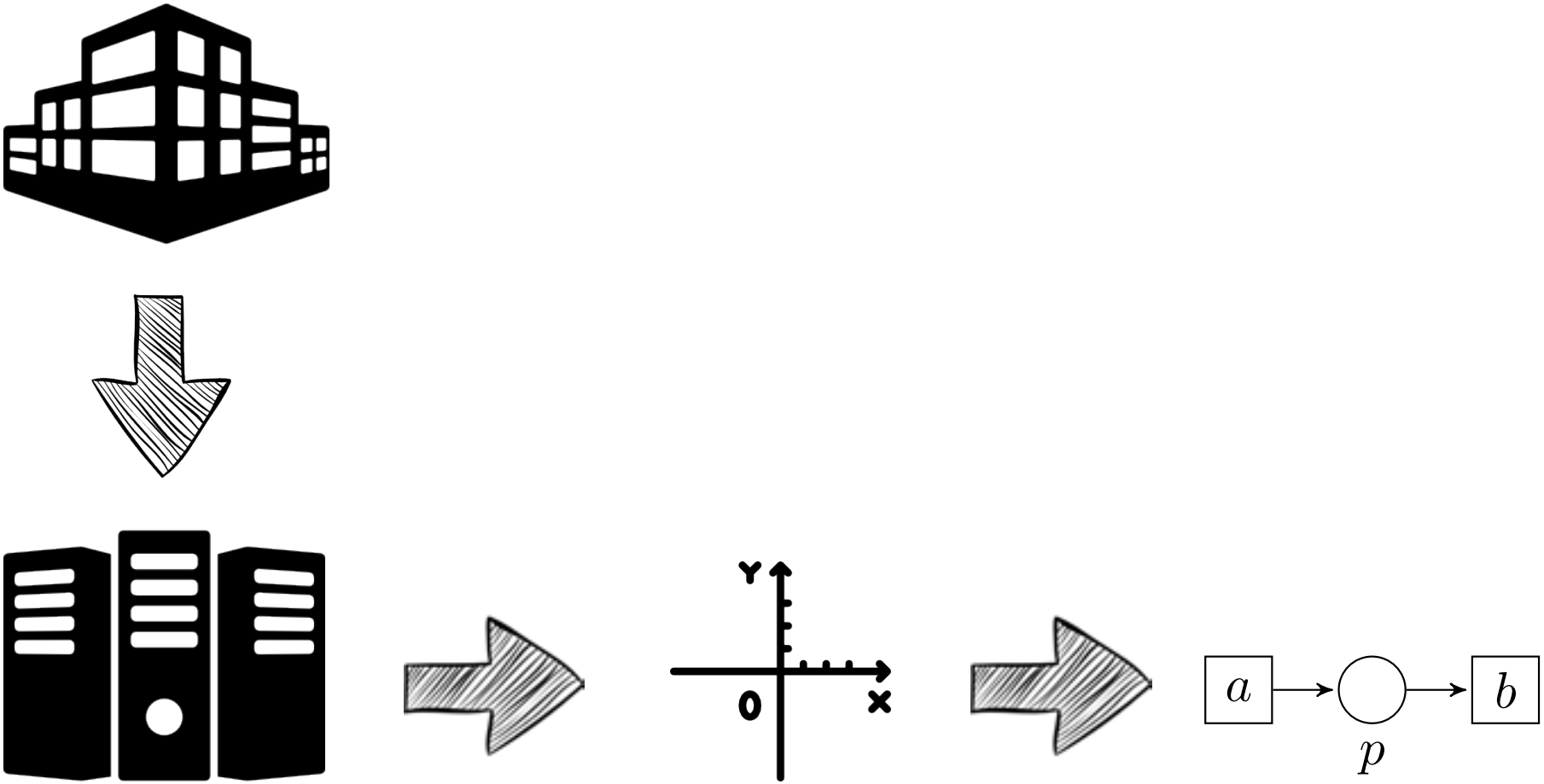
Setting the Scene
Process Discovery



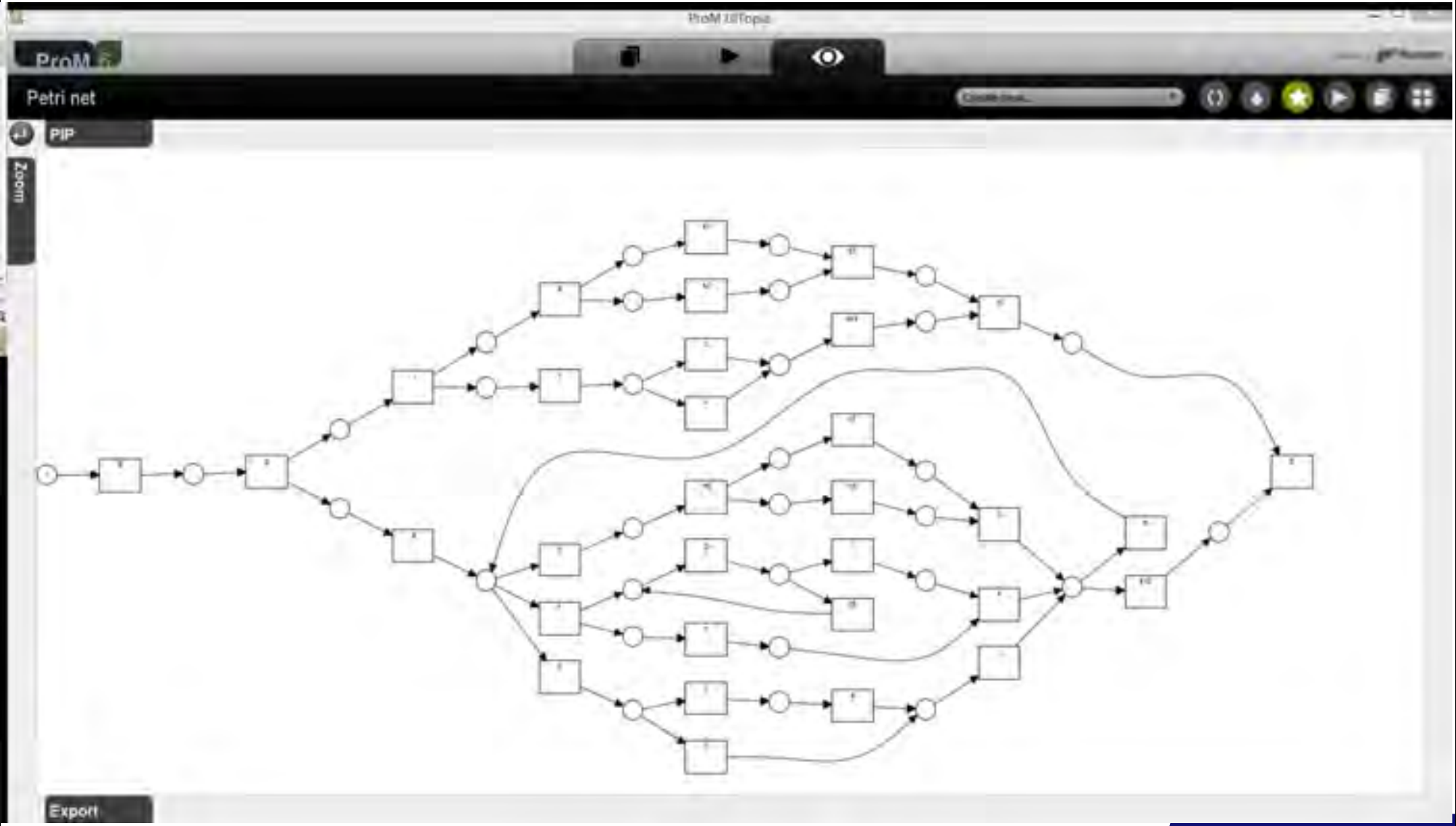
Setting the Scene
ILP-Based Process Discovery



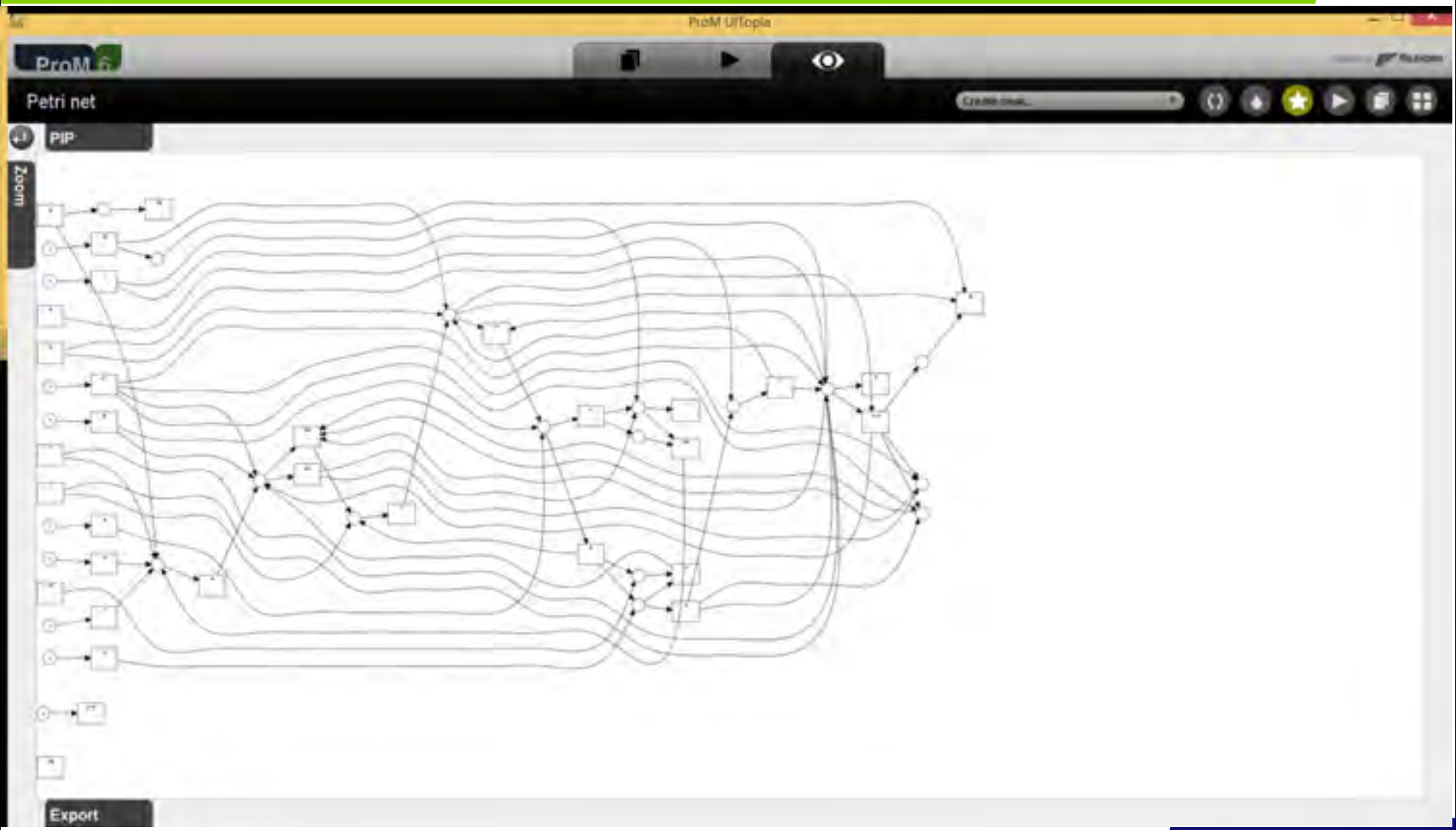
Setting the Scene
ILP-Based Process Discovery



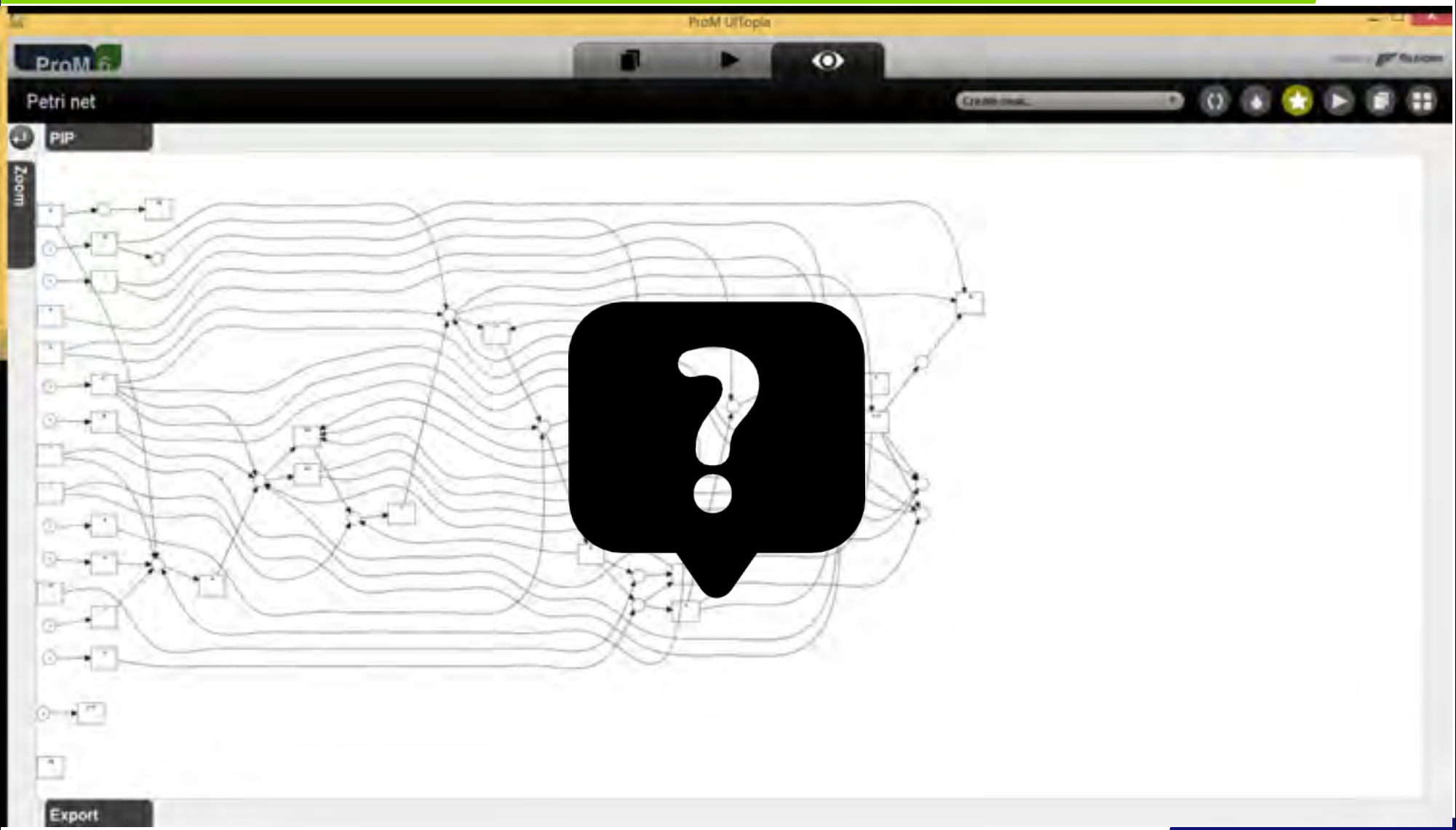
Motivation
The Impact of Exceptional Behavior



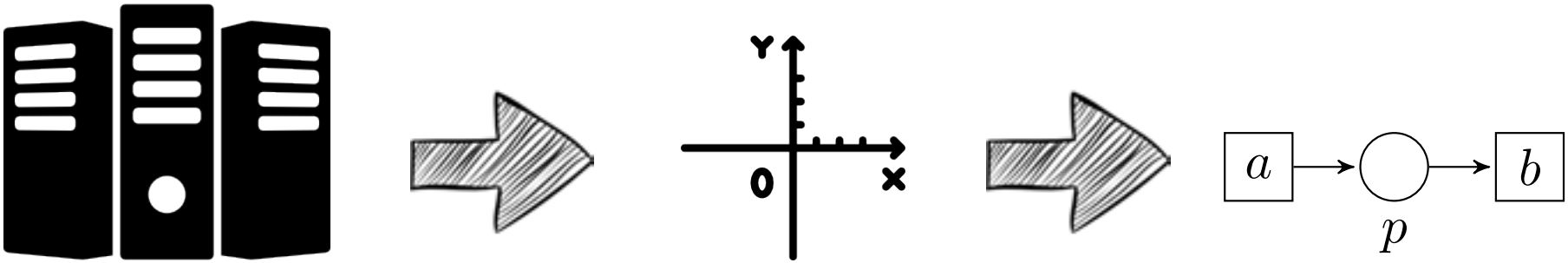
Motivation
The Impact of Exceptional Behavior



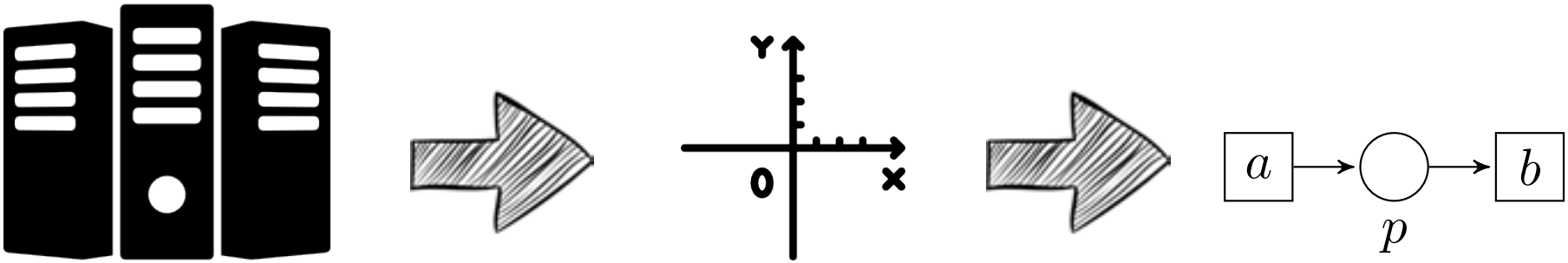
Motivation
The Impact of Exceptional Behavior



Historical Perspective
Petri net Synthesis

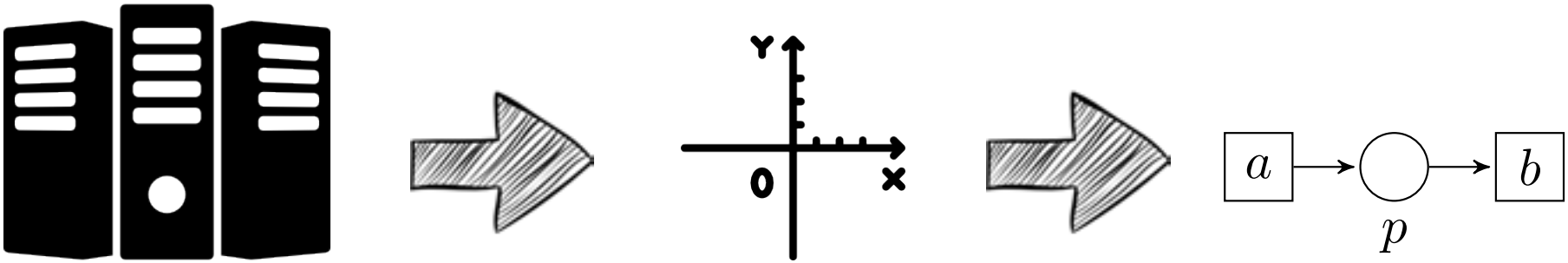


Historical Perspective
Petri net Synthesis

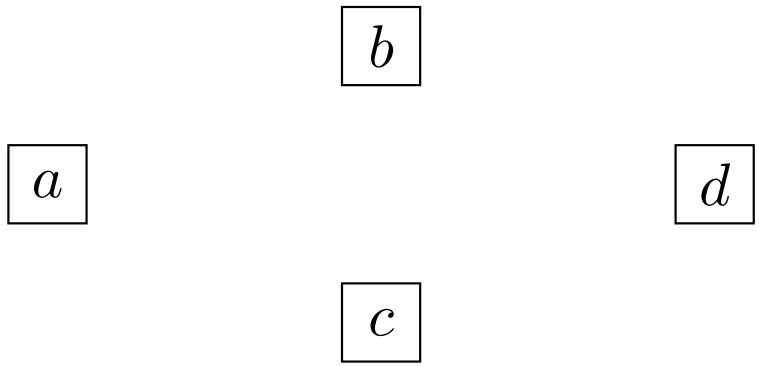


- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}]$

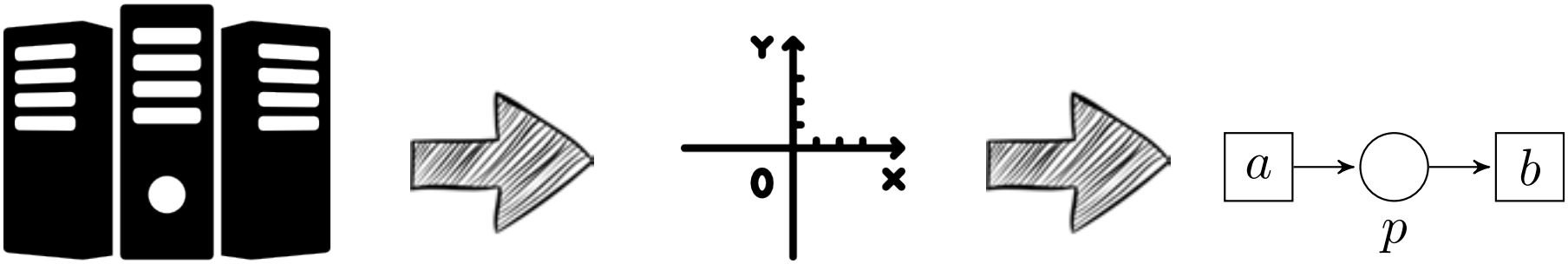
Historical Perspective
Petri net Synthesis



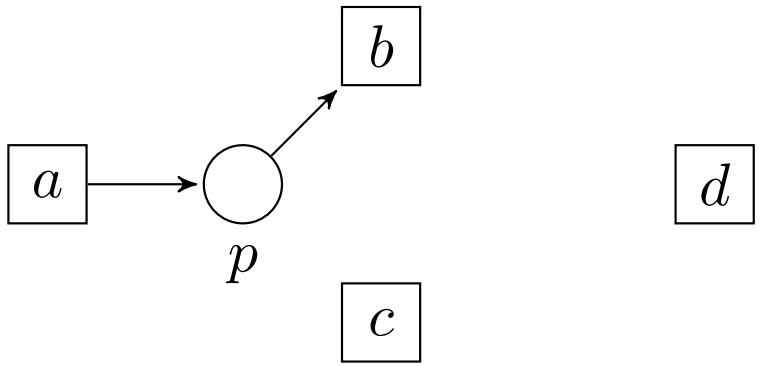
- $L = [\langle a, b, d\rangle^{10}, \langle a, c, d\rangle^{10}]$
- $\overline{L} = \{\epsilon, \langle a\rangle, \langle a, b\rangle, \langle a, c\rangle, \langle a, b, d\rangle, \langle a, c, d\rangle\}$



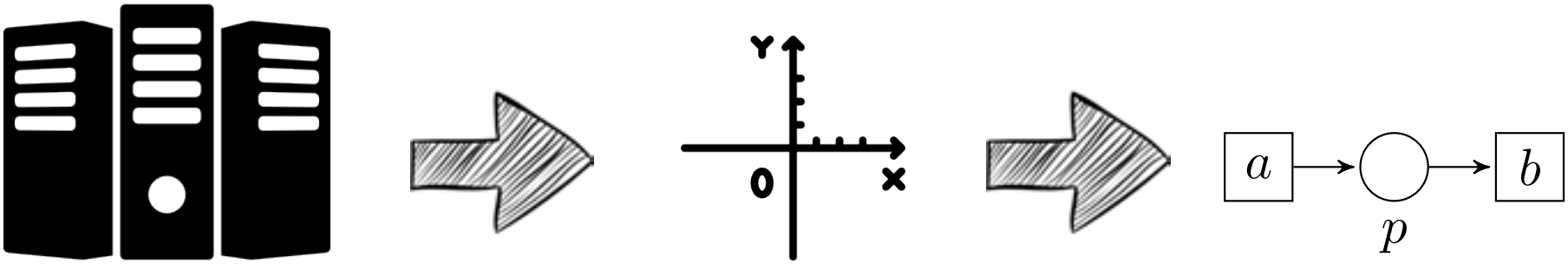
Historical Perspective
Petri net Synthesis



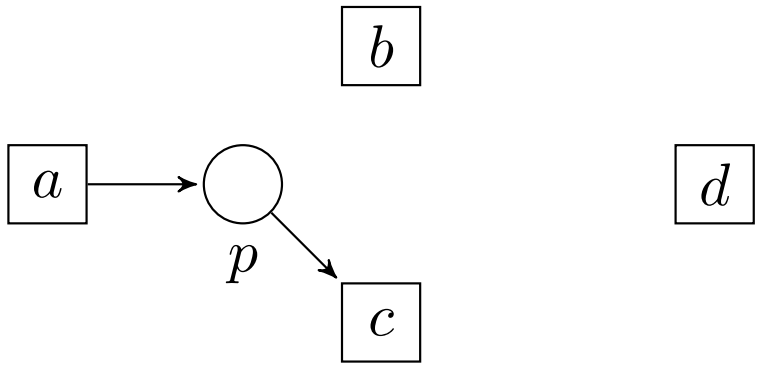
- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle\}$



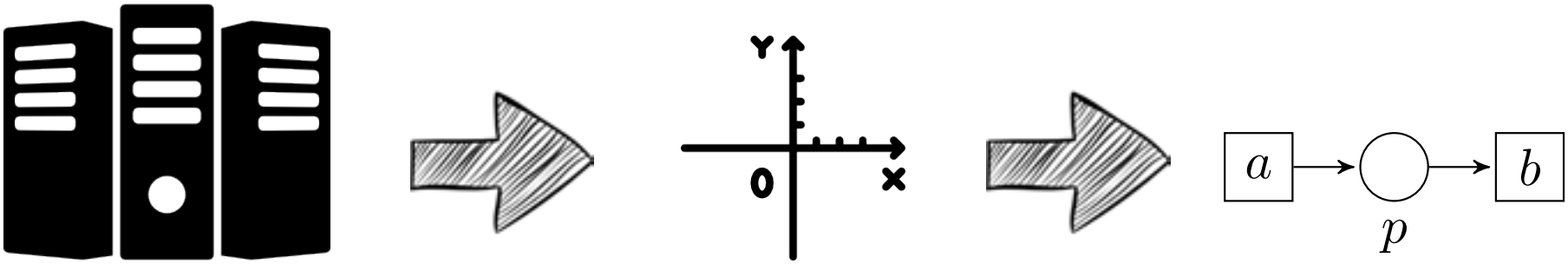
Historical Perspective
Petri net Synthesis



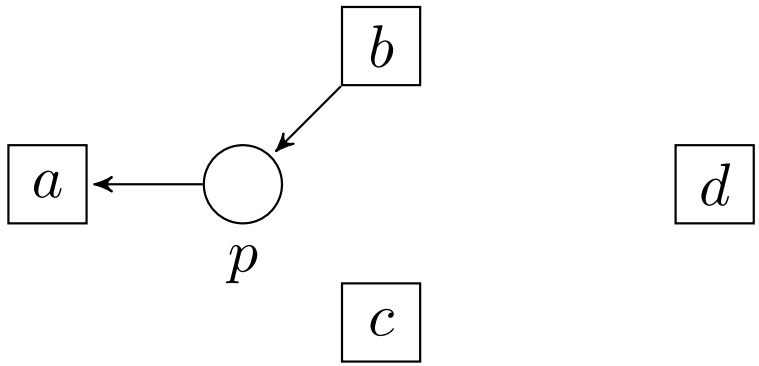
- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle\}$



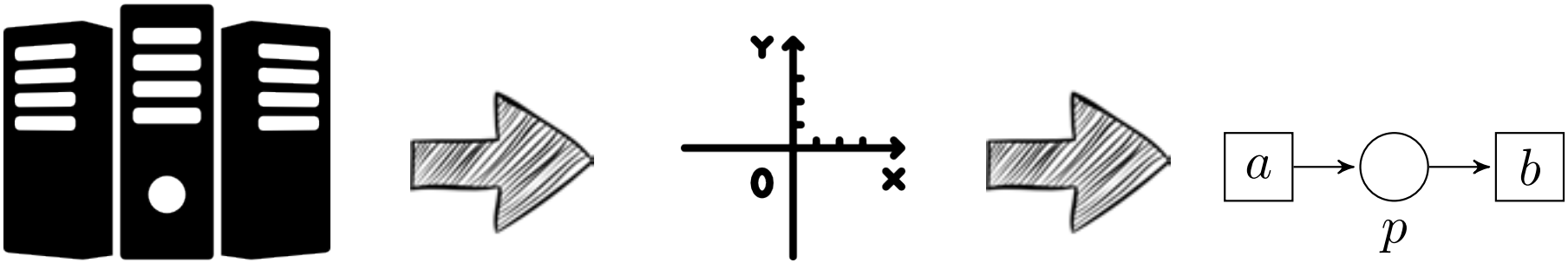
Historical Perspective
Petri net Synthesis



- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle\}$

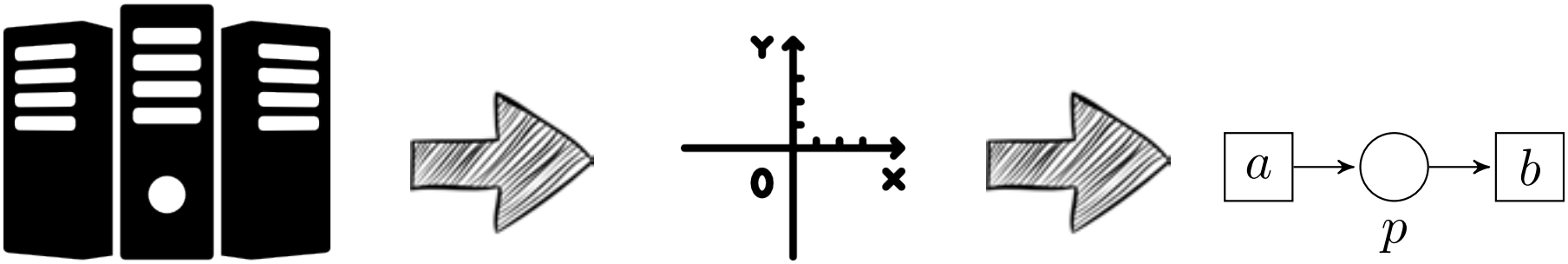


Process Discovery
Integer Linear Programming

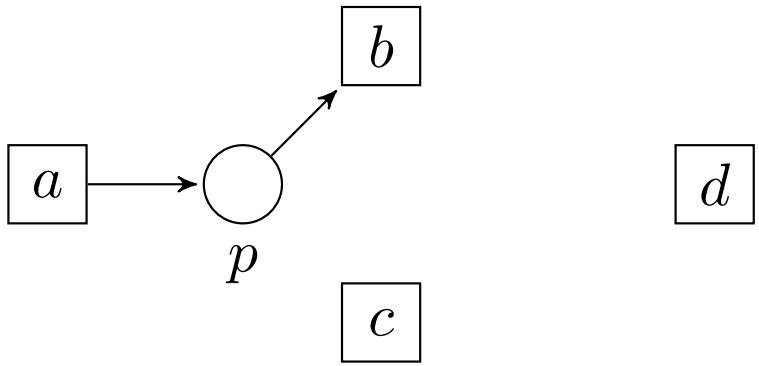


- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle\}$

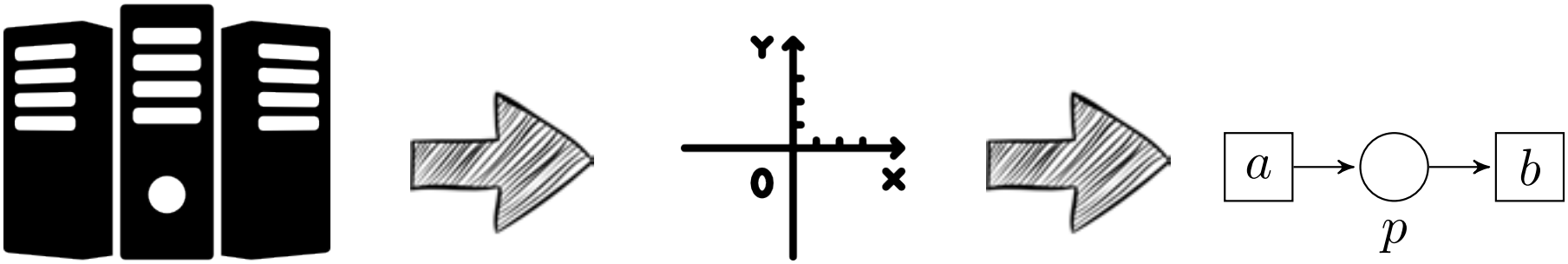
Process Discovery
Integer Linear Programming



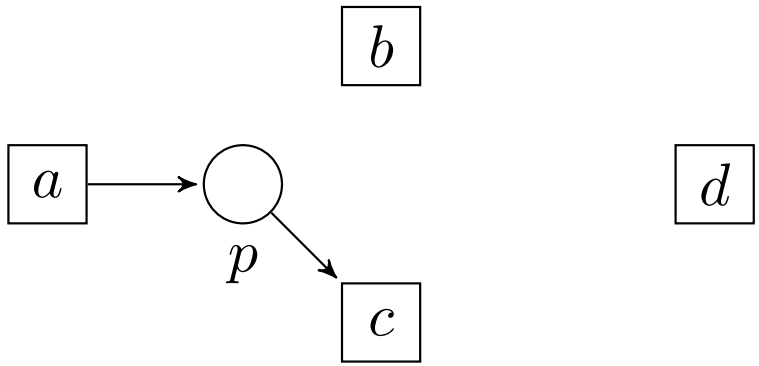
- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle\}$



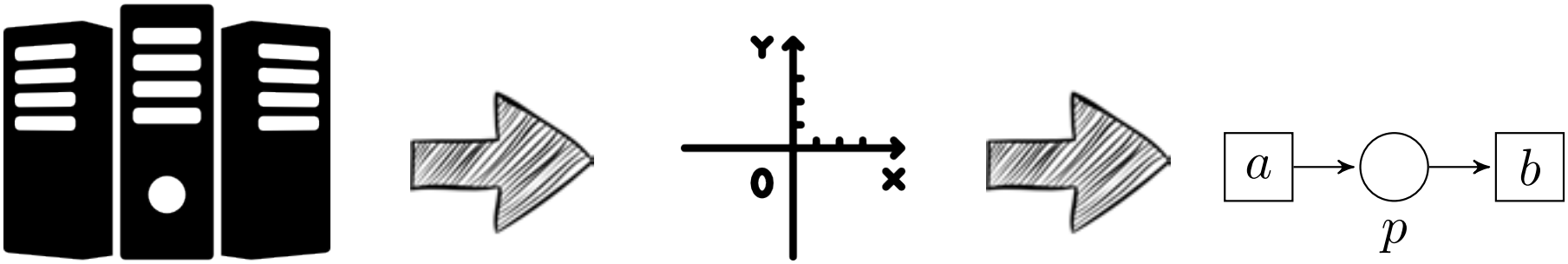
Process Discovery
Integer Linear Programming



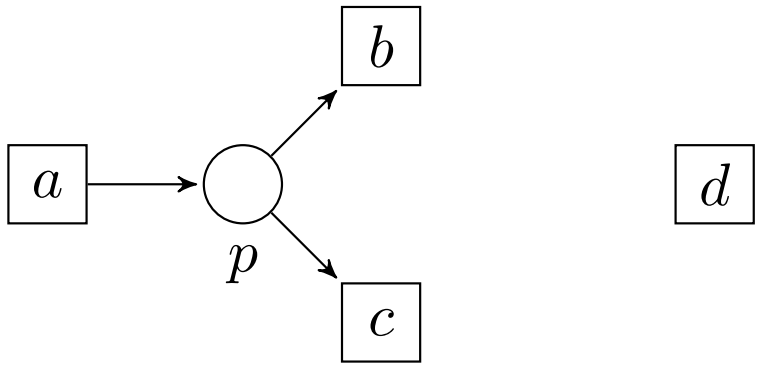
- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle\}$



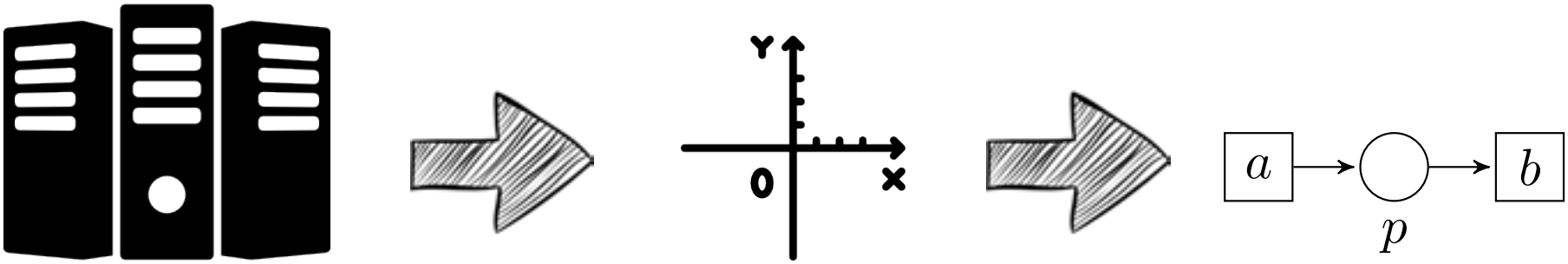
Process Discovery
Integer Linear Programming



- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle\}$

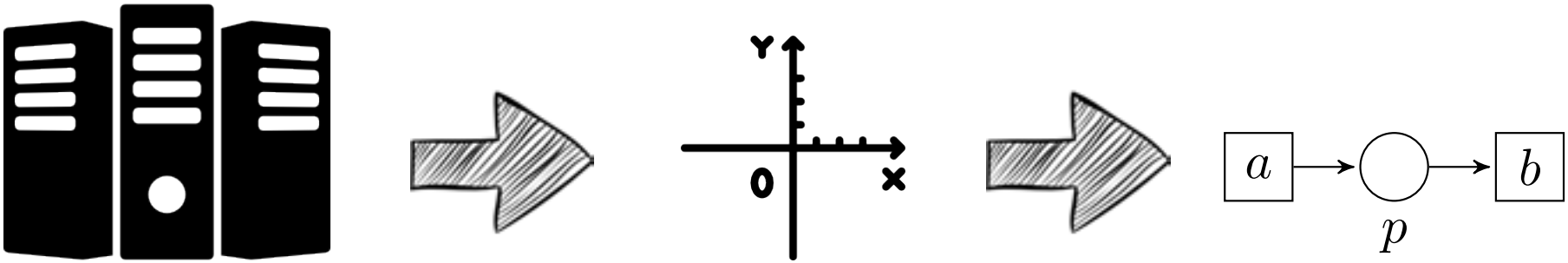


Process Discovery
Integer Linear Programming



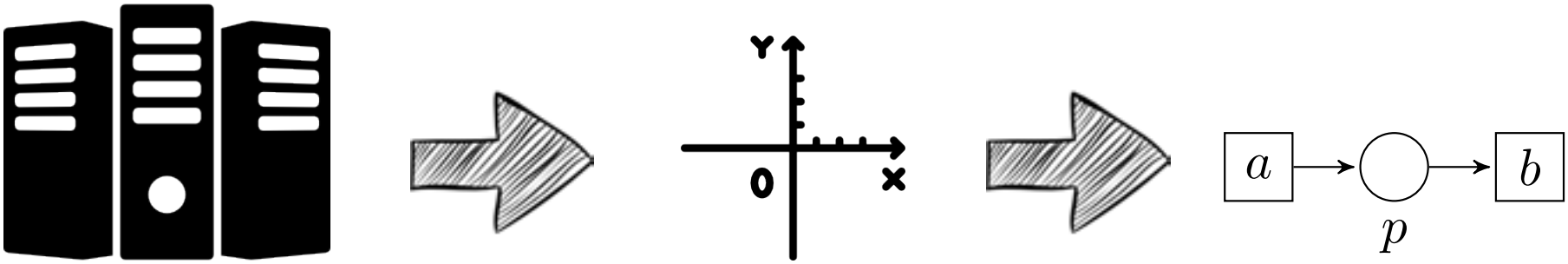
- Each $\sigma \in \overline{L} \rightarrow$ 1 constraint

Process Discovery
Integer Linear Programming



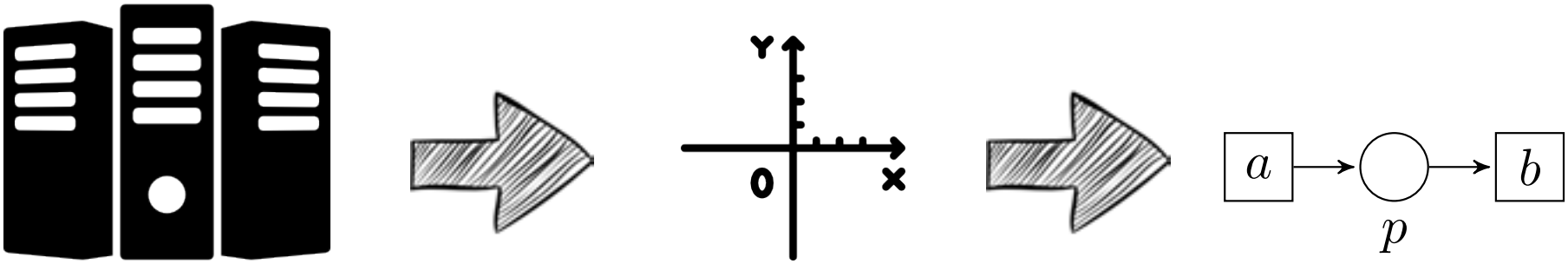
- Each $\sigma \in \overline{L} \rightarrow$ 1 constraint
- Each event class $\rightarrow$ 2 variables

Process Discovery
Integer Linear Programming

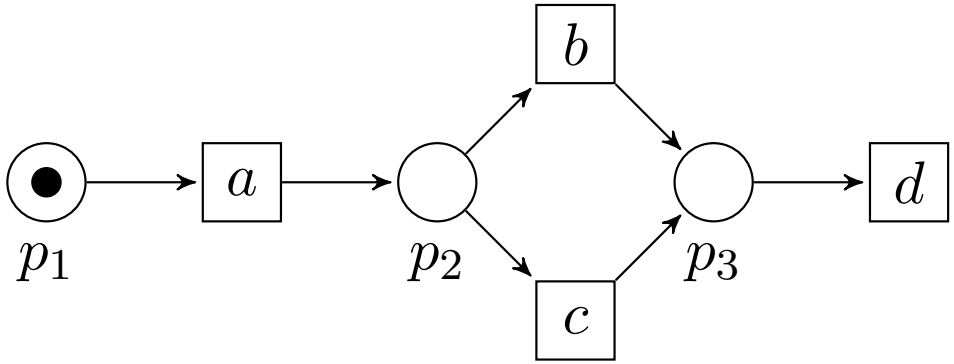


- Each $\sigma \in \overline{L} \rightarrow$ 1 constraint
- Each event class $\rightarrow$ 2 variables
- Each causal relation $\rightarrow$ 1 ILP

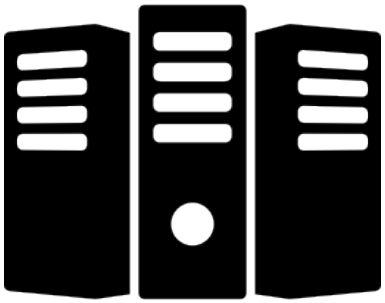
Process Discovery
Integer Linear Programming

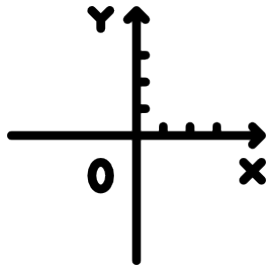


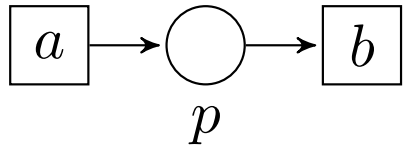
- Each $\sigma \in \overline{L} \rightarrow$ 1 constraint
- Each event class $\rightarrow$ 2 variables
- Each causal relation $\rightarrow$ 1 ILP



Process Discovery
Integer Linear Programming





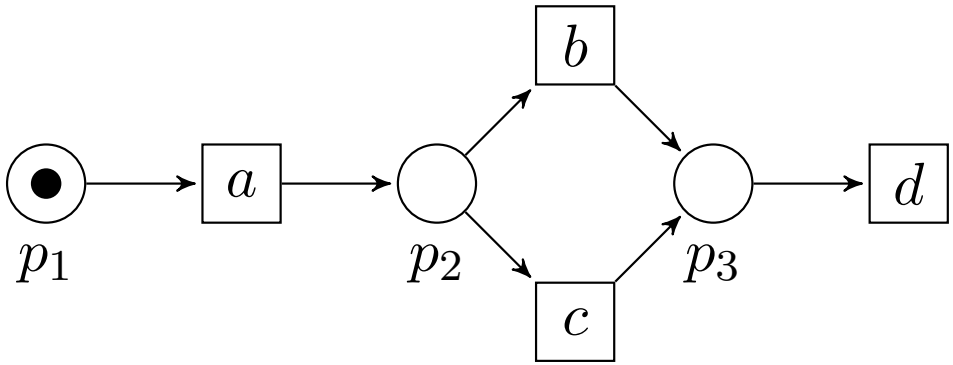


Process Discovery
Integer Linear Programming

Process Discovery
Integer Linear Programming

Process Discovery
Integer Linear Programming

- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle\}$



Language-Based Regions
Impact of Exceptional Behavior

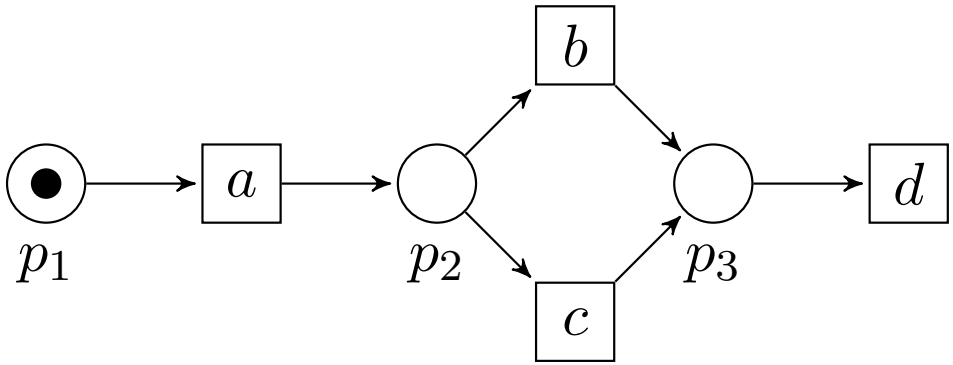
- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$

Language-Based Regions
Impact of Exceptional Behavior

- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle, \langle a, b, c \rangle\}$

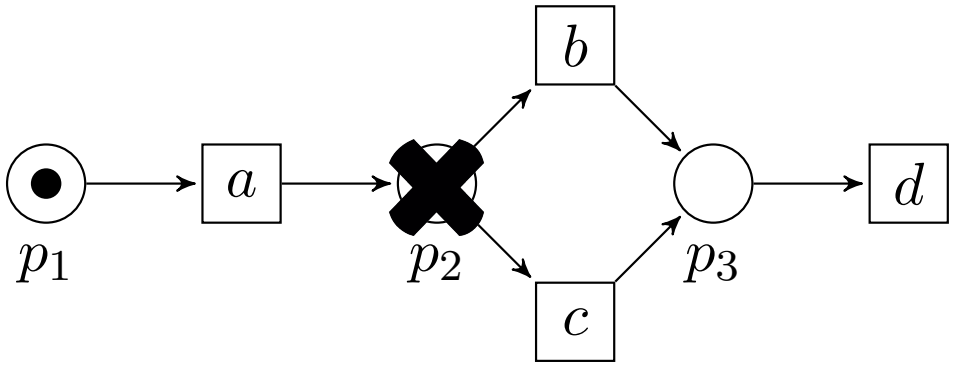
Language-Based Regions
Impact of Exceptional Behavior

- $L = [\langle a, b, d\rangle^{10}, \langle a, c, d\rangle^{10}, \langle a, b, c\rangle]$
- $\overline{L} = \{\epsilon, \langle a\rangle, \langle a, b\rangle, \langle a, c\rangle, \langle a, b, d\rangle, \langle a, c, d\rangle, \langle a, b, c\rangle\}$



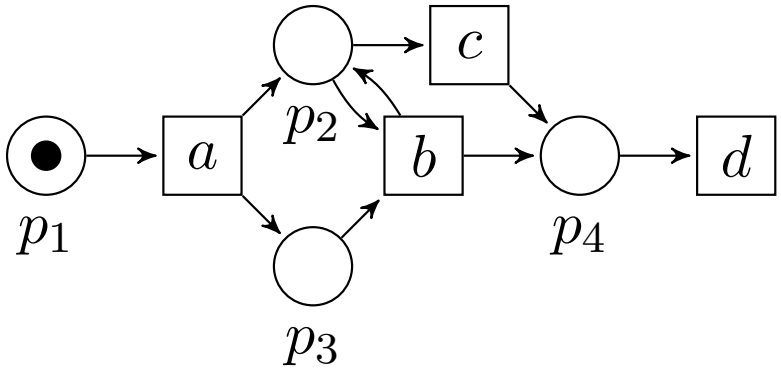
Language-Based Regions
Impact of Exceptional Behavior

- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle, \langle a, b, c \rangle\}$



Language-Based Regions
Impact of Exceptional Behavior

- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle, \langle a, b, c \rangle\}$



ILP-Based Process Discovery
Slack Variable Filtering

ILP-Based Process Discovery
Slack Variable Filtering

- Each event class $\rightarrow$ 2 variables
- Each $\sigma \in \overline{L} \rightarrow$ 1 constraint

ILP-Based Process Discovery
Slack Variable Filtering

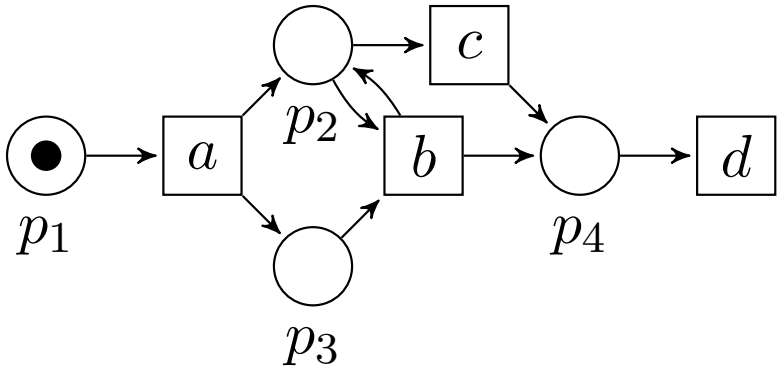
- Each event class $\rightarrow$ 2 variables
- Each $\sigma \in \overline{L} \rightarrow$ 1 constraint

- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle, \langle a, b, c \rangle\}$

ILP-Based Process Discovery
Slack Variable Filtering

- Each event class $\rightarrow$ 2 variables
- Each $\sigma \in \overline{L} \rightarrow$ 1 constraint

- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle, \langle a, b, c \rangle\}$

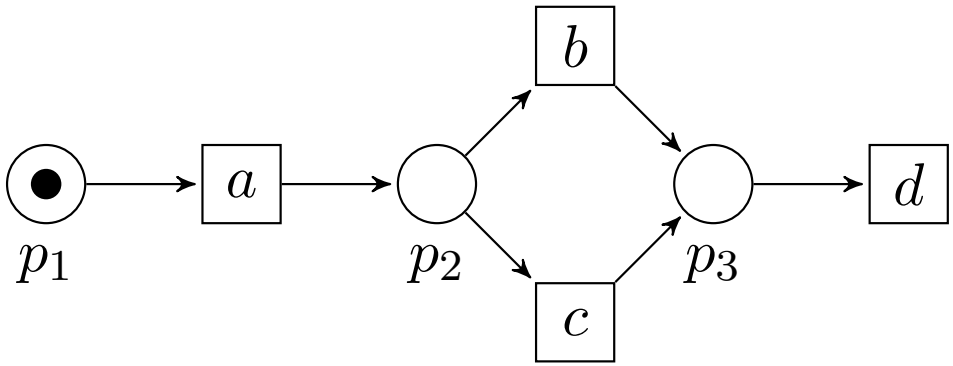


ILP-Based Process Discovery
Slack Variable Filtering

- Each event class $\rightarrow$ 2 variables
- Each $\sigma \in \overline{L} \rightarrow$ 1 constraint
- Each $\sigma \in \overline{L} \rightarrow$ 1 variable
- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle, \langle a, b, c \rangle\}$

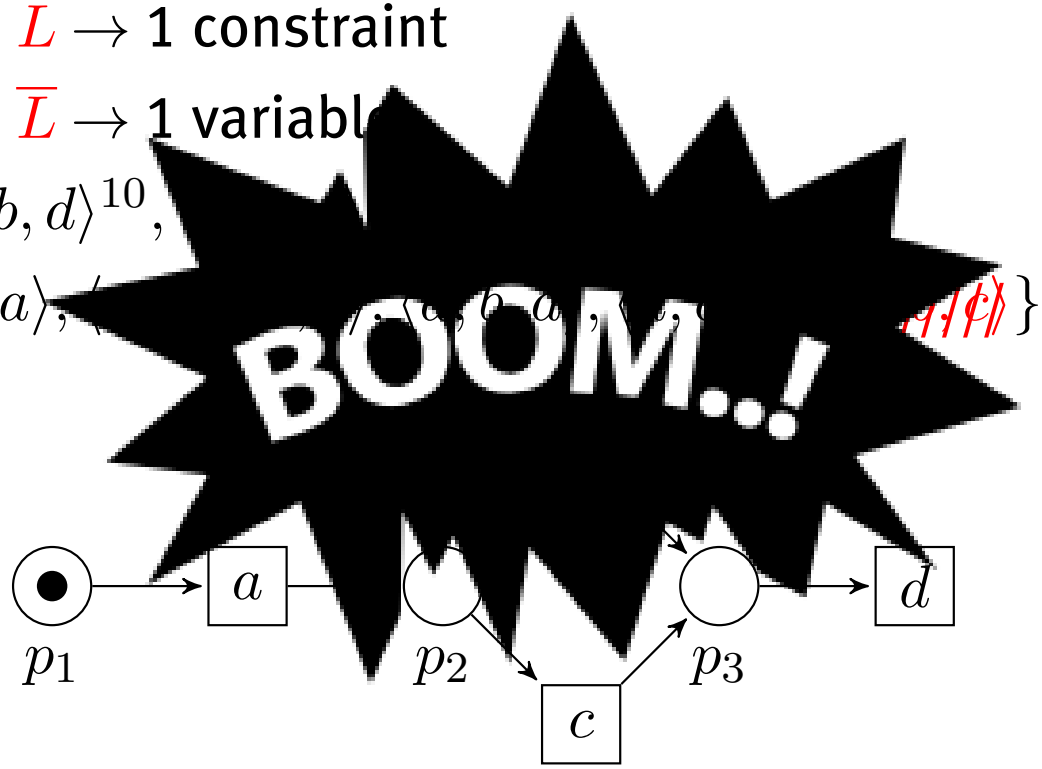
ILP-Based Process Discovery
Slack Variable Filtering

- Each event class $\rightarrow$ 2 variables
- Each $\sigma \in \overline{L} \rightarrow$ 1 constraint
- Each $\sigma \in \overline{L} \rightarrow$ 1 variable
- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$
- $\overline{L} = \{\epsilon, \langle a \rangle, \langle a, b \rangle, \langle a, c \rangle, \langle a, b, d \rangle, \langle a, c, d \rangle, \langle a, b, c \rangle\}$



ILP-Based Process Discovery
Slack Variable Filtering

- Each event class $\rightarrow$ 2 variables
- Each $\sigma \in \overline{L} \rightarrow$ 1 constraint
- Each $\sigma \in \overline{L} \rightarrow$ 1 variable
- $L = [\langle a, b, d \rangle^{10},$
- $\overline{L} = \{\epsilon, \langle a \rangle,$ $\}$



ILP-Based Process Discovery Sequence Encoding Filtering

ILP-Based Process Discovery
Sequence Encoding Filtering

- $\langle a \rangle = ([], a)$

ILP-Based Process Discovery
Sequence Encoding Filtering

- $\langle a \rangle = ([], a)$
- $\langle a, b \rangle = ([a], b)$

ILP-Based Process Discovery
Sequence Encoding Filtering

- $\langle a \rangle = ([], a)$
- $\langle a, b \rangle = ([a], b)$
- $\langle a, c \rangle = ([a], c)$

ILP-Based Process Discovery
Sequence Encoding Filtering

- $\langle a \rangle = ([], a)$
- $\langle a, b \rangle = ([a], b)$
- $\langle a, c \rangle = ([a], c)$

$\vdots$

- $\langle a, b, c \rangle = ([a, b], c)$

ILP-Based Process Discovery
Sequence Encoding Filtering

- $\langle a \rangle = ([], a)$
- $\langle a, b \rangle = ([a], b)$
- $\langle a, c \rangle = ([a], c)$

$\vdots$

- $\langle a, b, c \rangle = ([a, b], c)$
- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$

ILP-Based Process Discovery
Sequence Encoding Filtering

- $\langle a \rangle = ([], a)$
- $\langle a, b \rangle = ([a], b)$
- $\langle a, c \rangle = ([a], c)$

$\vdots$

- $\langle a, b, c \rangle = ([a, b], c)$
- $L = [\langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$
- $\overline{L} = [\epsilon^{21}, \langle a \rangle^{21}, \langle a, b \rangle^{11}, \langle a, c \rangle^{10}, \langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$

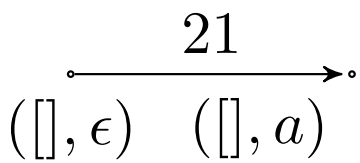
ILP-Based Process Discovery
Sequence Encoding Filtering

- $\overline{L} = [\epsilon^{21}, \langle a \rangle^{21}, \langle a, b \rangle^{11}, \langle a, c \rangle^{10}, \langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$

$$([\overset{\circ}{]}, \epsilon)$$

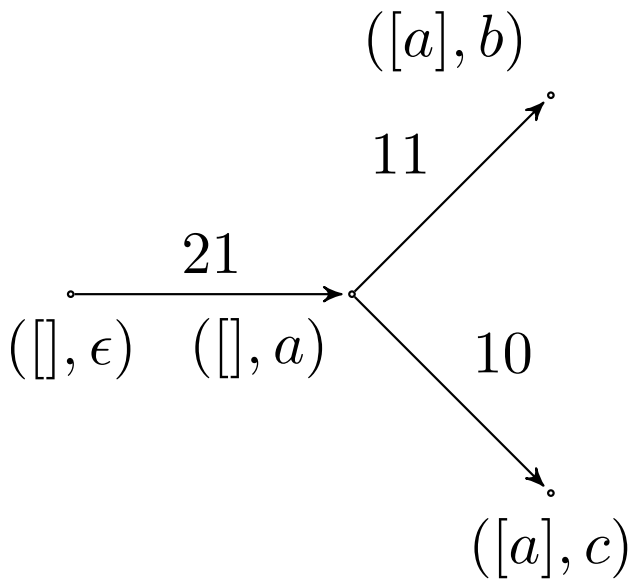
ILP-Based Process Discovery
Sequence Encoding Filtering

- $\overline{L} = [\epsilon^{21}, \langle a \rangle^{21}, \langle a, b \rangle^{11}, \langle a, c \rangle^{10}, \langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$



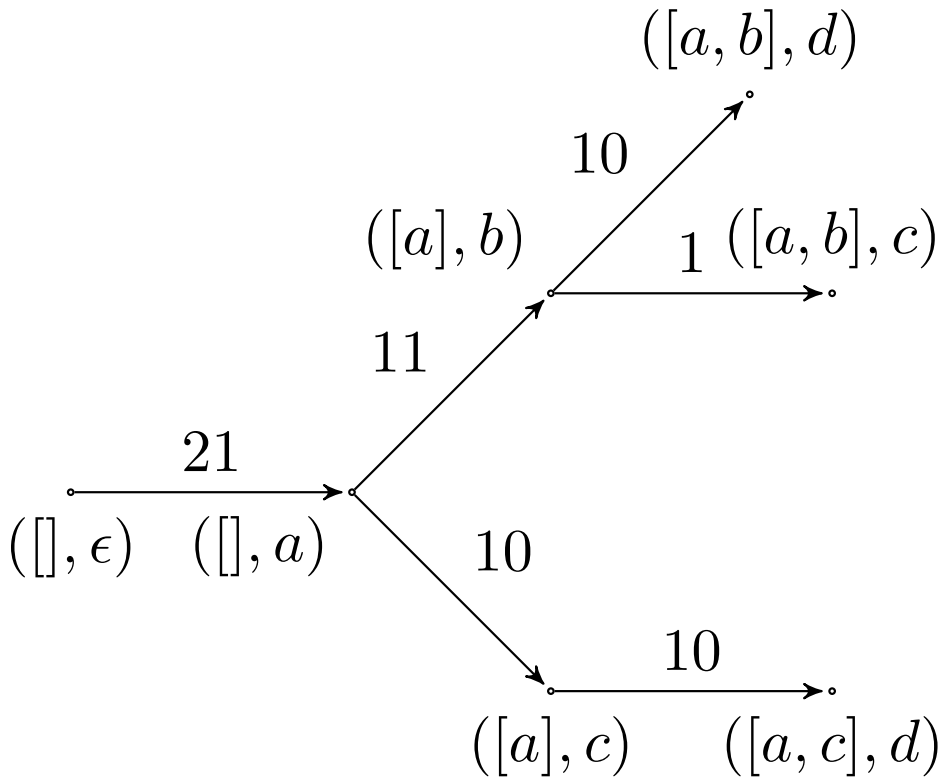
ILP-Based Process Discovery
Sequence Encoding Filtering

- $\overline{L} = [\epsilon^{21}, \langle a \rangle^{21}, \langle a, b \rangle^{11}, \langle a, c \rangle^{10}, \langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$

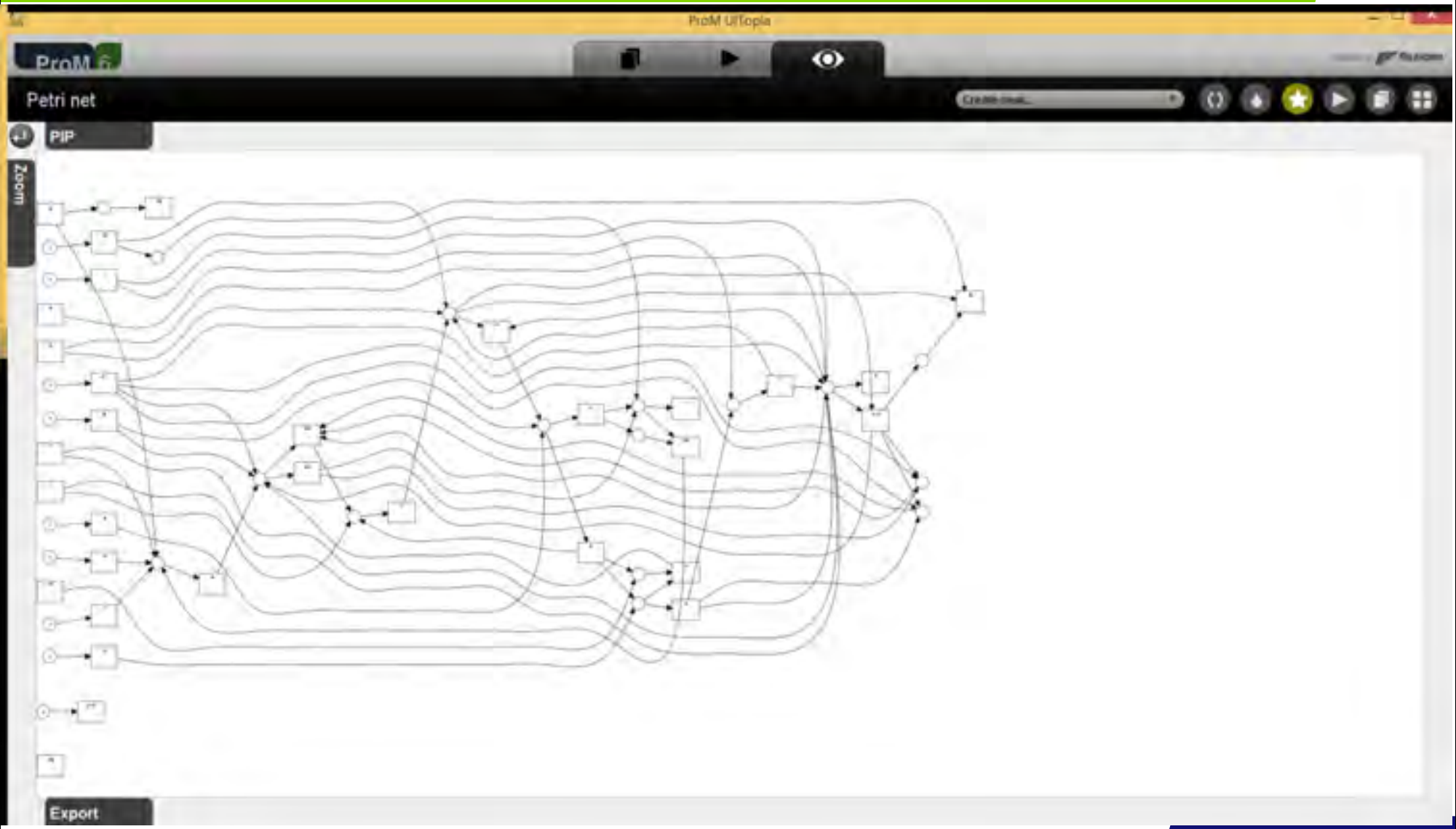


ILP-Based Process Discovery
Sequence Encoding Filtering

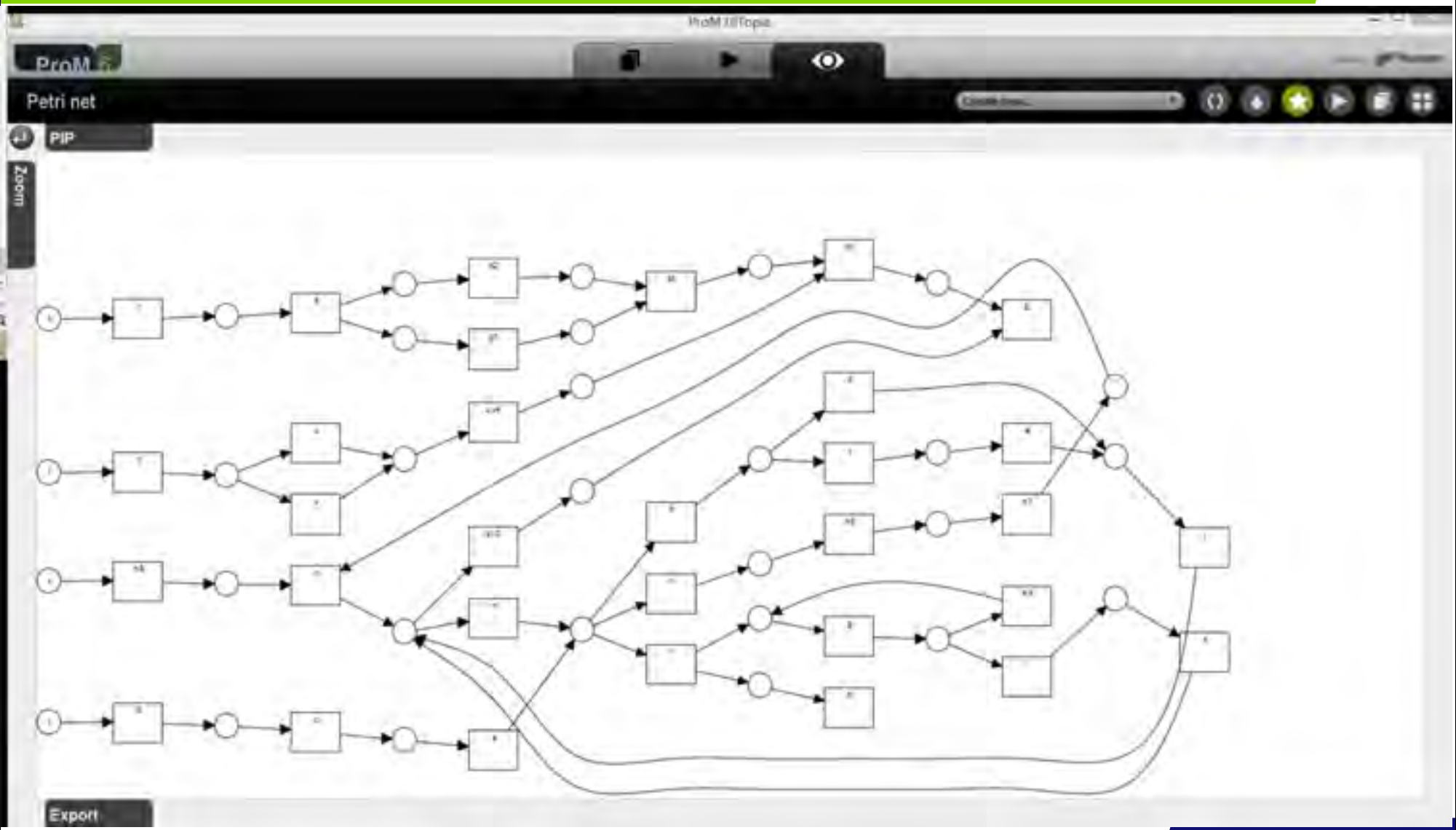
- $\overline{L} = [\epsilon^{21}, \langle a \rangle^{21}, \langle a, b \rangle^{11}, \langle a, c \rangle^{10}, \langle a, b, d \rangle^{10}, \langle a, c, d \rangle^{10}, \langle a, b, c \rangle]$



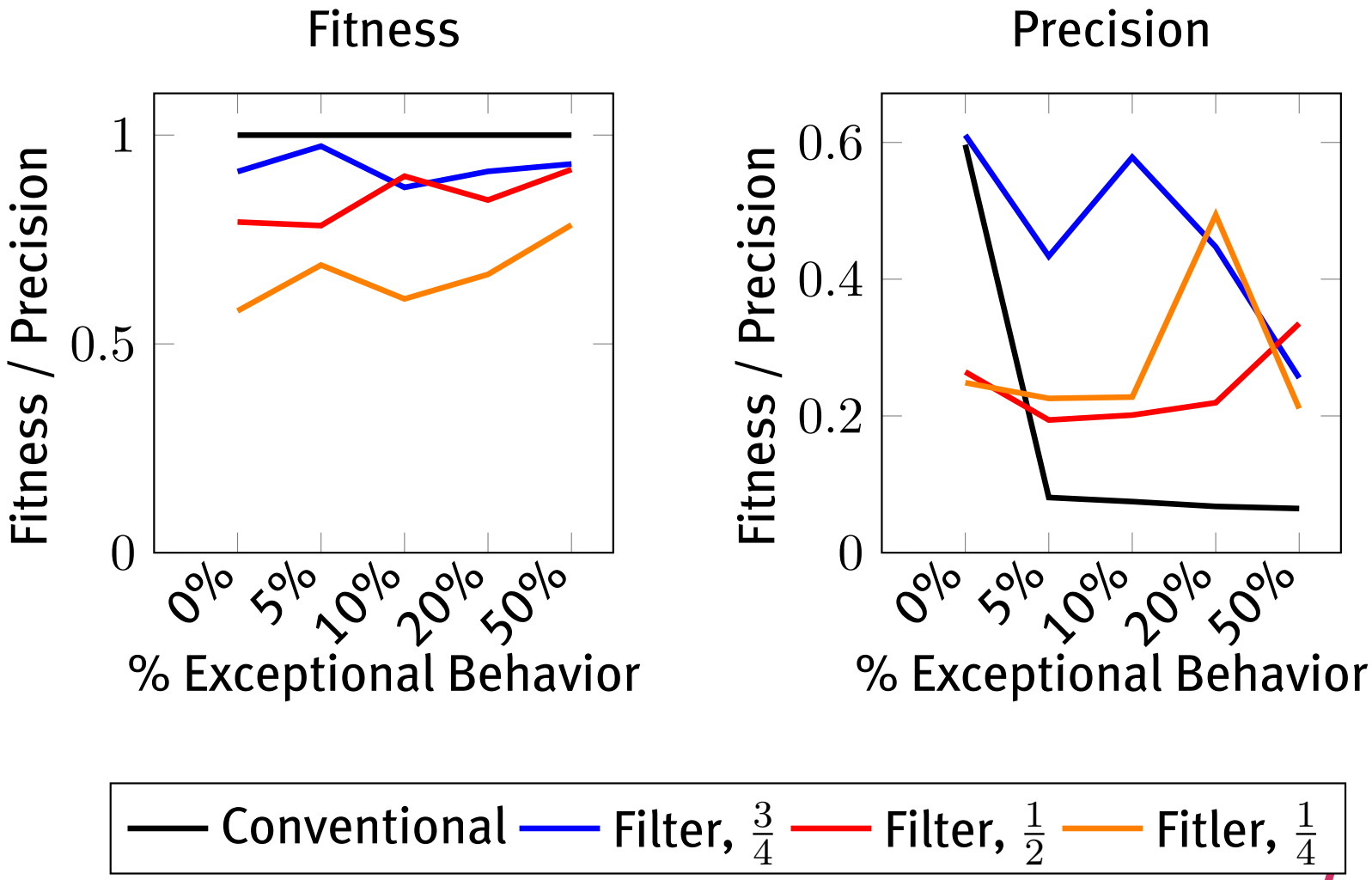
Evaluation
Qualitative



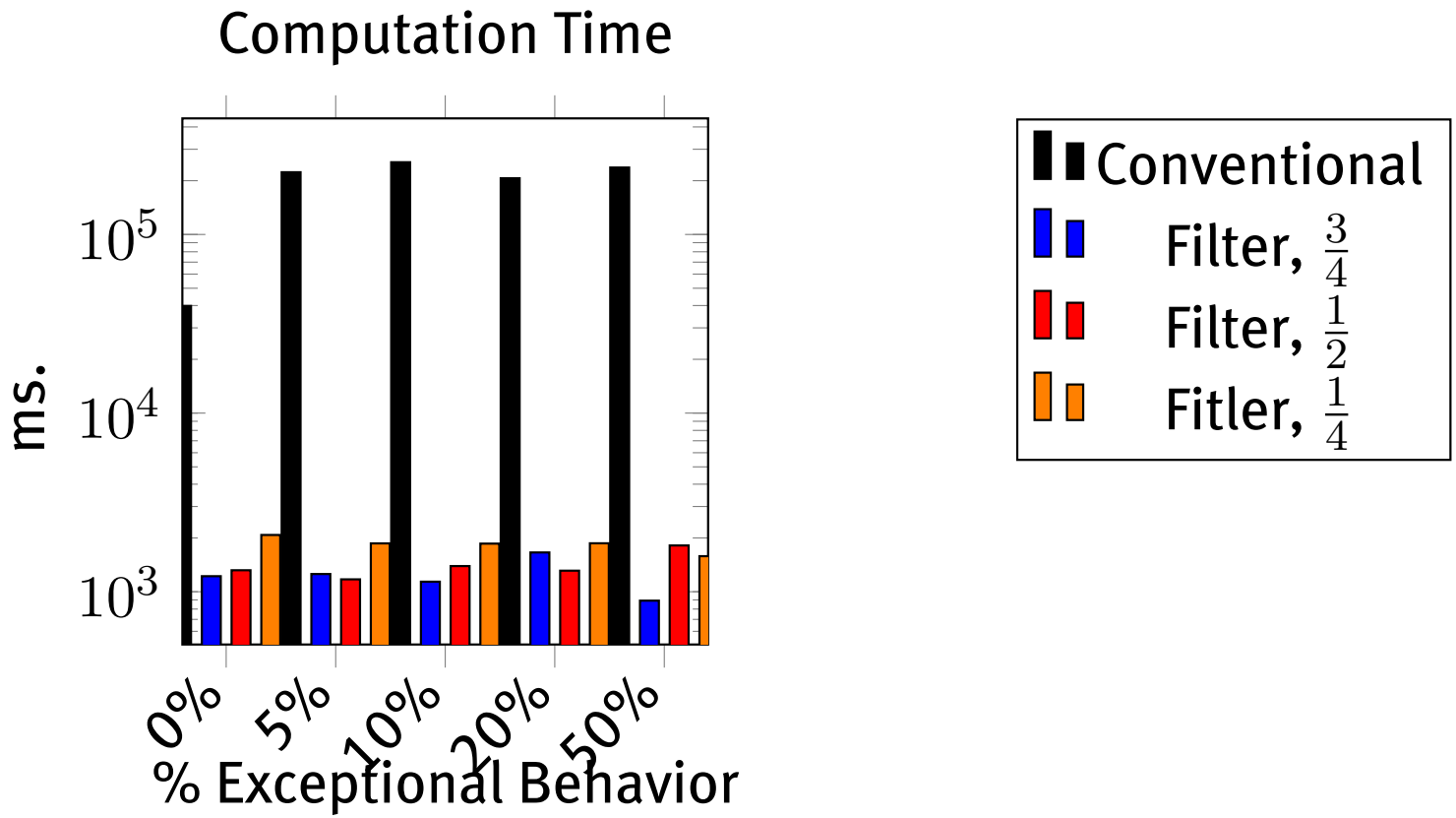
Evaluation
Qualitative



Evaluation
Quantitative - Replay Fitness / Precision



Evaluation
Quantitative - Computation Time



Conclusion

Conclusion

Conclusion

Conclusion

Questions