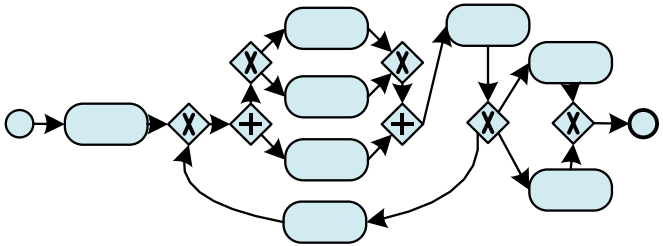
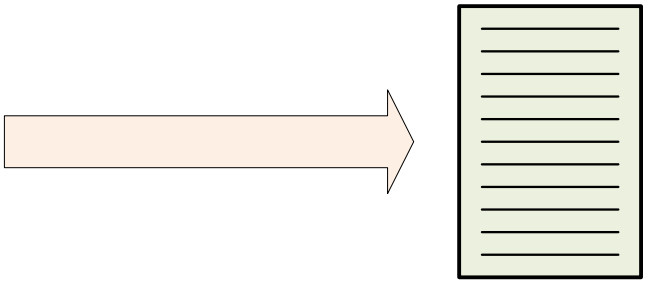

let's play

Play-Out

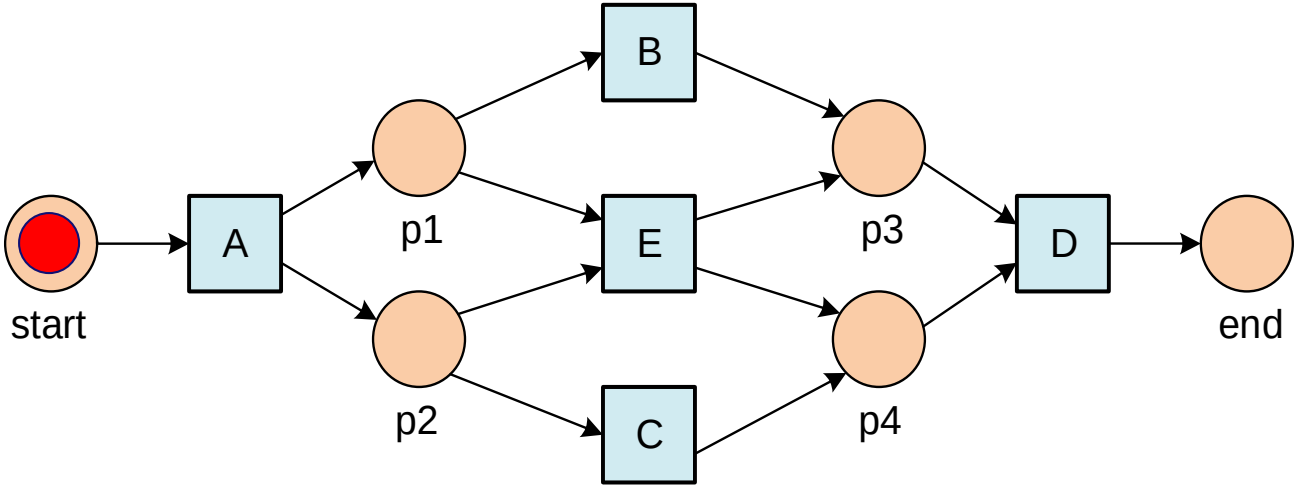


process model

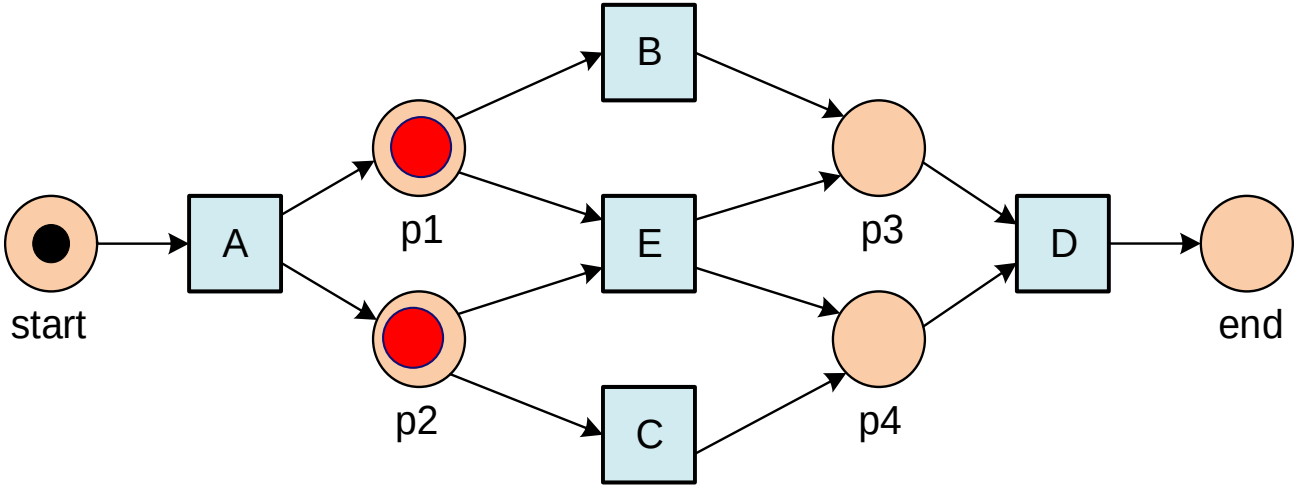


event log

Play-Out (Classical use of models)

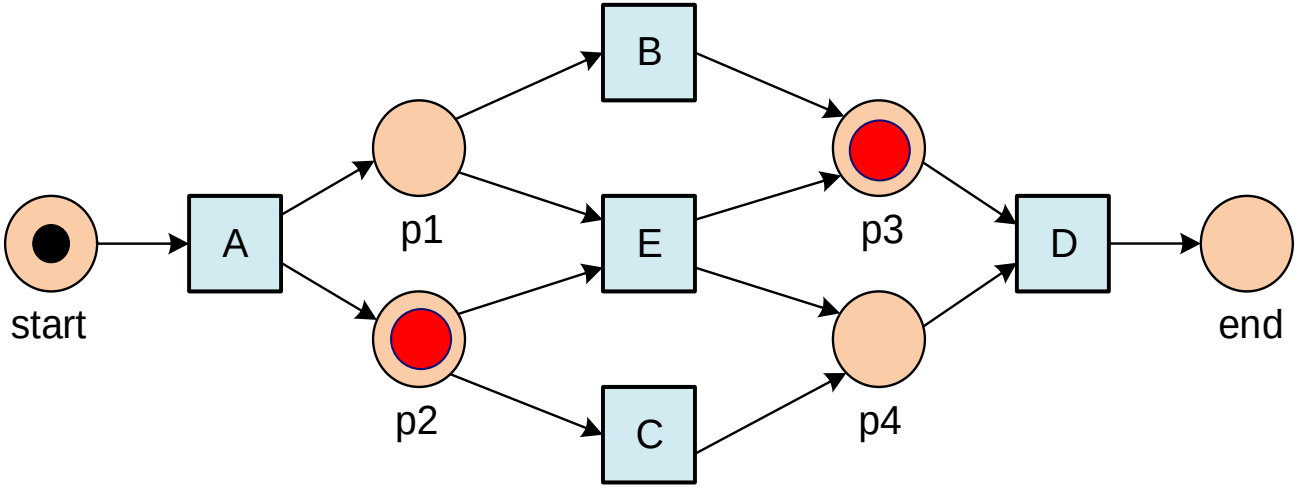


Play-Out (Classical use of models)



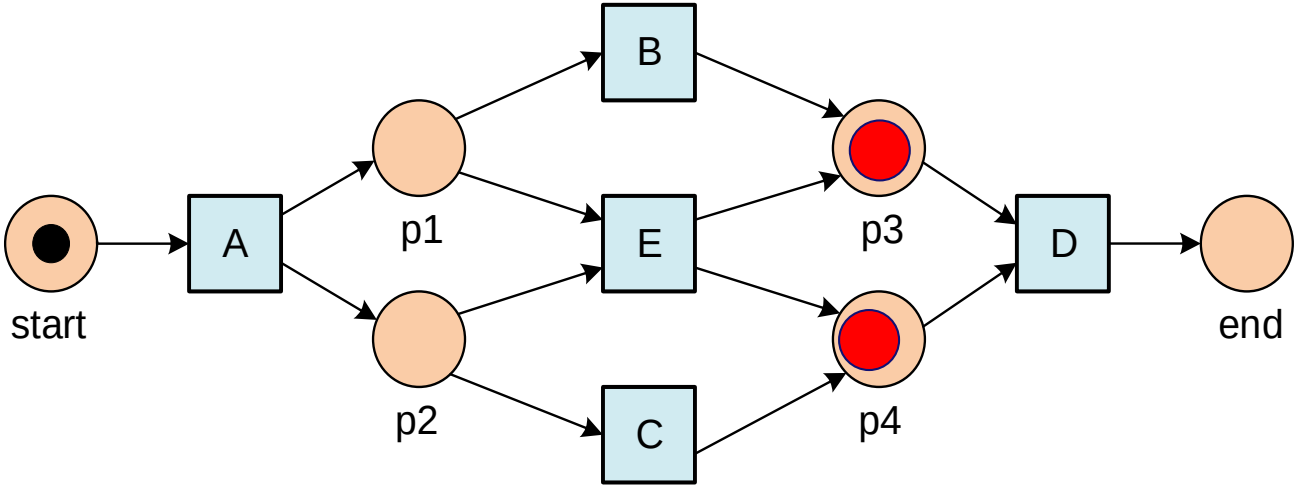
A

Play-Out (Classical use of models)



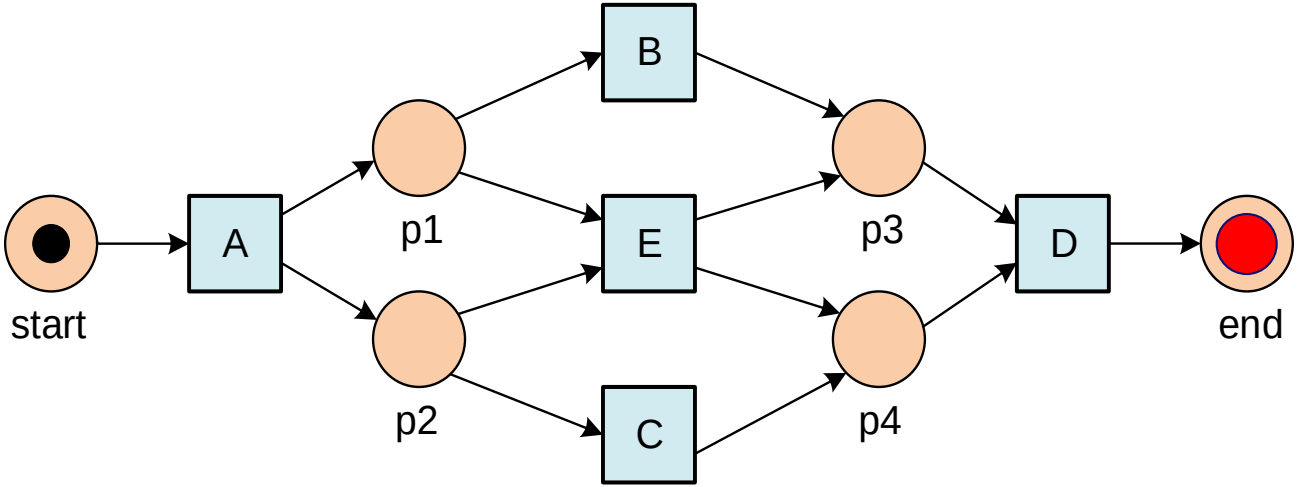
A B

Play-Out (Classical use of models)



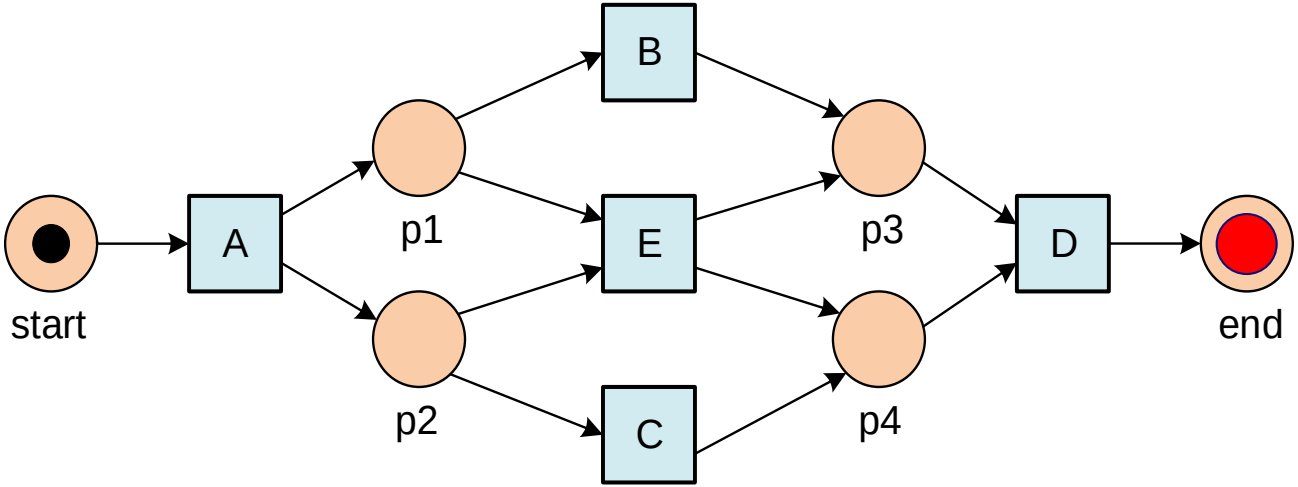
A B C

Play-Out (Classical use of models)



A B C D

Play-Out (Classical use of models)

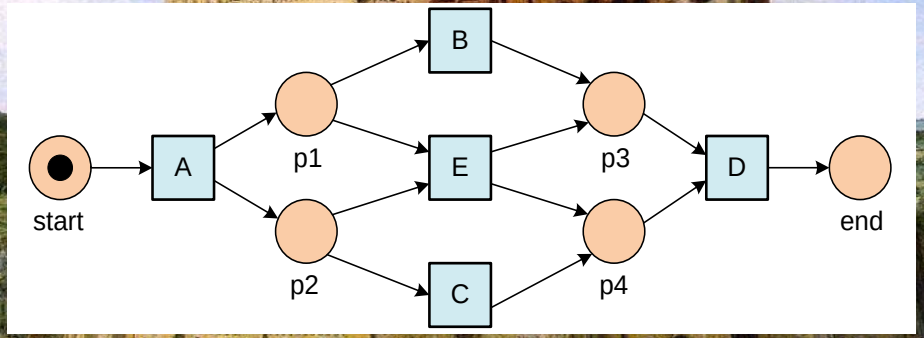


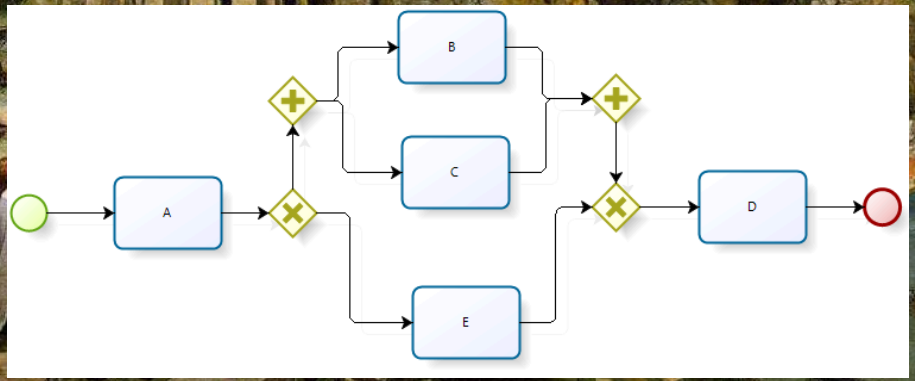
A B C D A E D A E D

A B C D A C B D

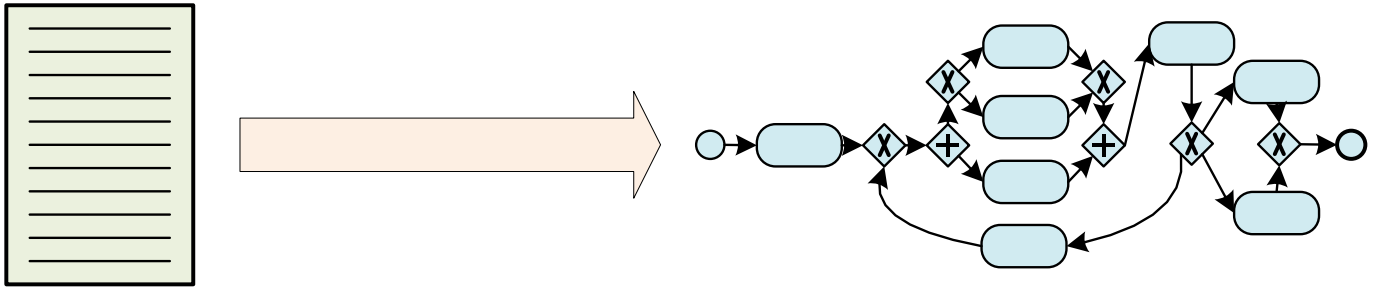
A C B D A E D A C B D

Let's not worry about syntax (there is difference between analysis and presentation)





Play-In



event log

process model

Play-In (Process Discovery, dude!)

A B C D

A E D

A E D

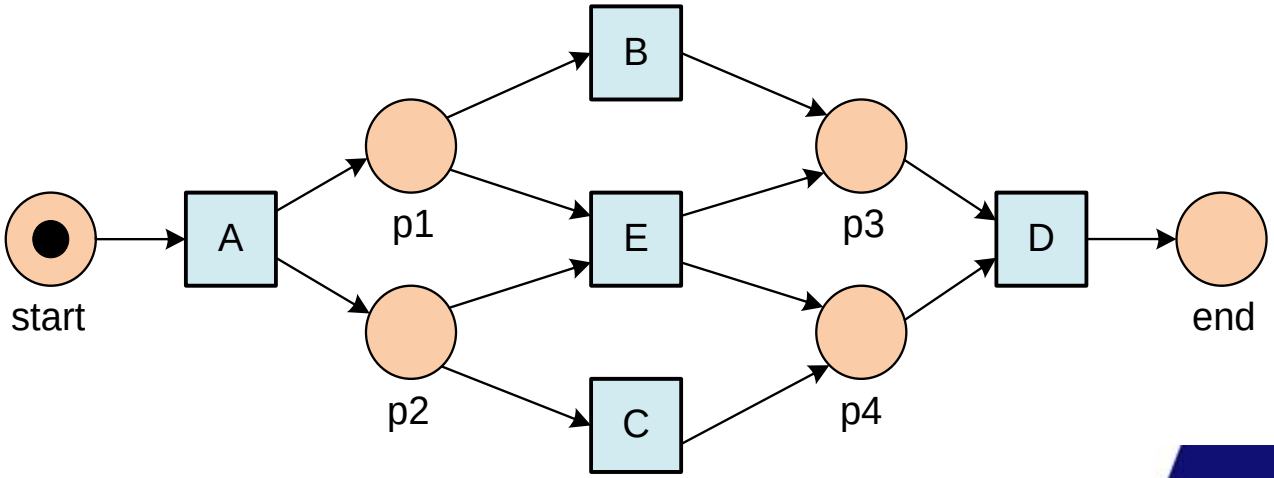
A B C D

A C B D

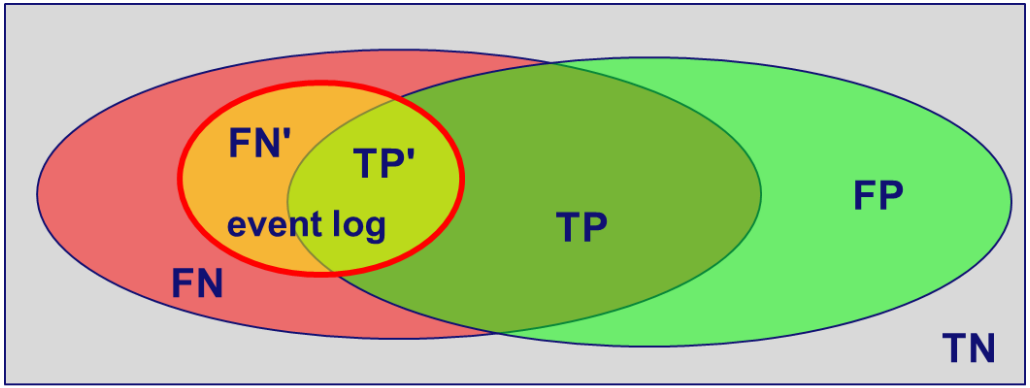
A C B D

A E D

A C B D

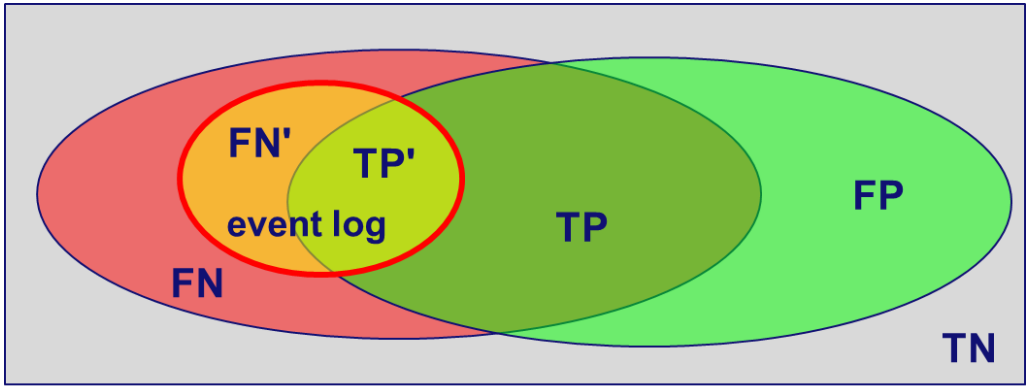


Challenges



Challenges

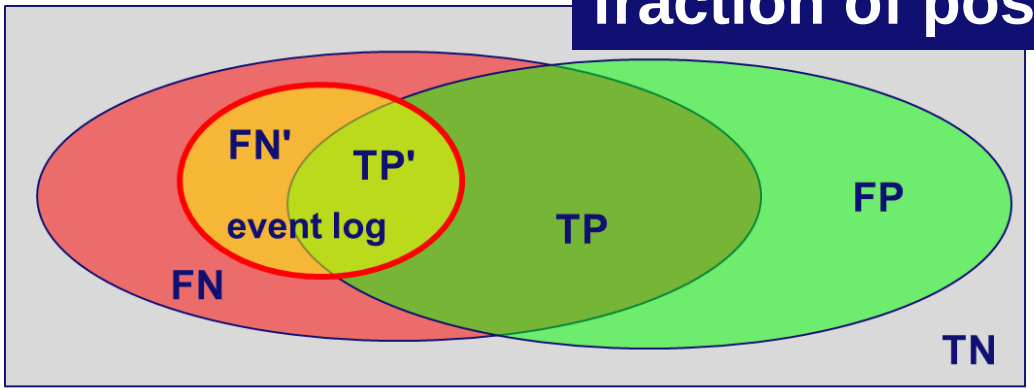
No negative examples
(cannot see what cannot happen)



Challenges

No negative examples
(cannot see what cannot happen)

Log contains only a fraction of possible traces



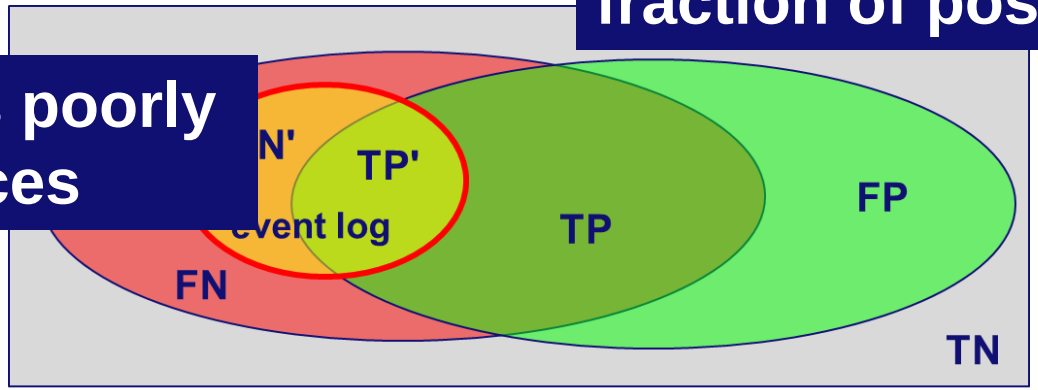
TU/e

Challenges

No negative examples
(cannot see what cannot happen)

Log contains only a fraction of possible traces

Almost vs poorly fitting traces



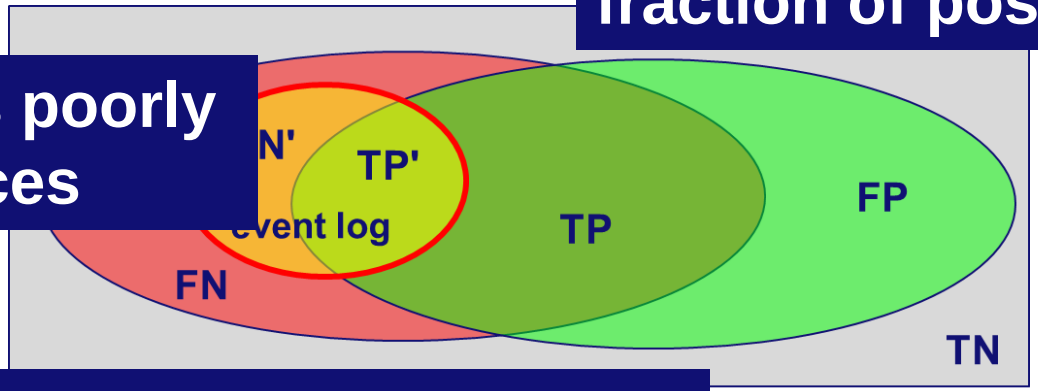
TU/e

Challenges

No negative examples
(cannot see what cannot happen)

Log contains only a fraction of possible traces

Almost vs poorly fitting traces



In case of loops often infinitely many possible traces

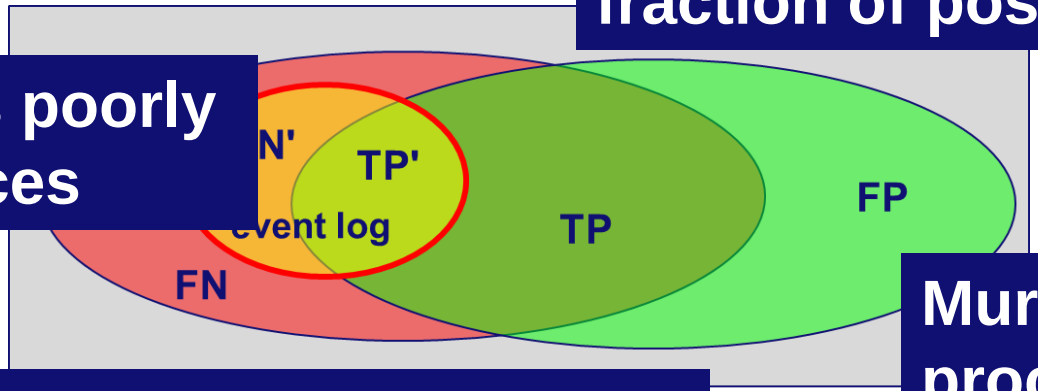
TU/e

Challenges

No negative examples
(cannot see what cannot happen)

Log contains only a fraction of possible traces

Almost vs poorly fitting traces



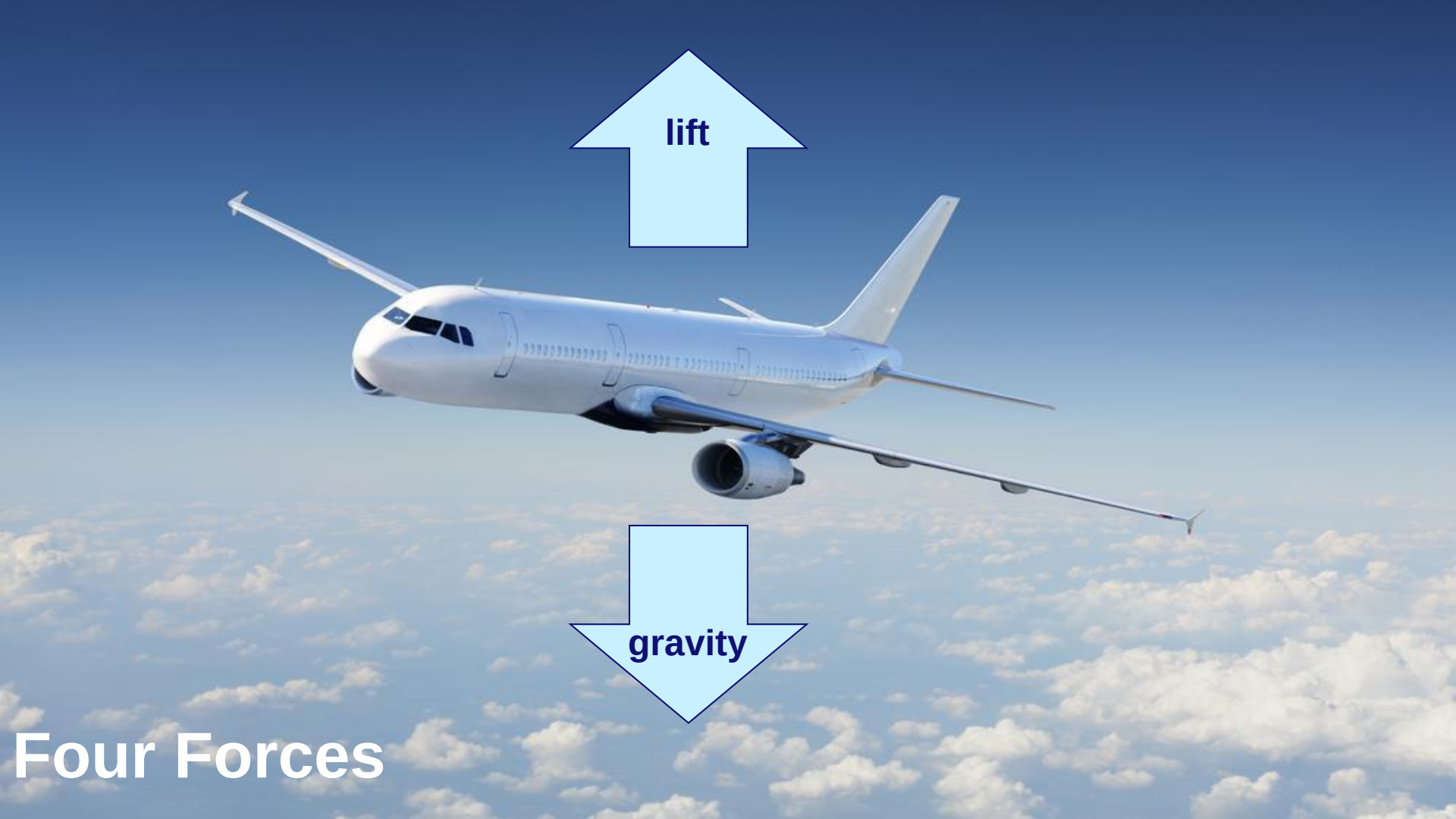
In case of loops often infinitely many possible traces

Murphy's law for process mining
(anything is possible, so probabilities matter)

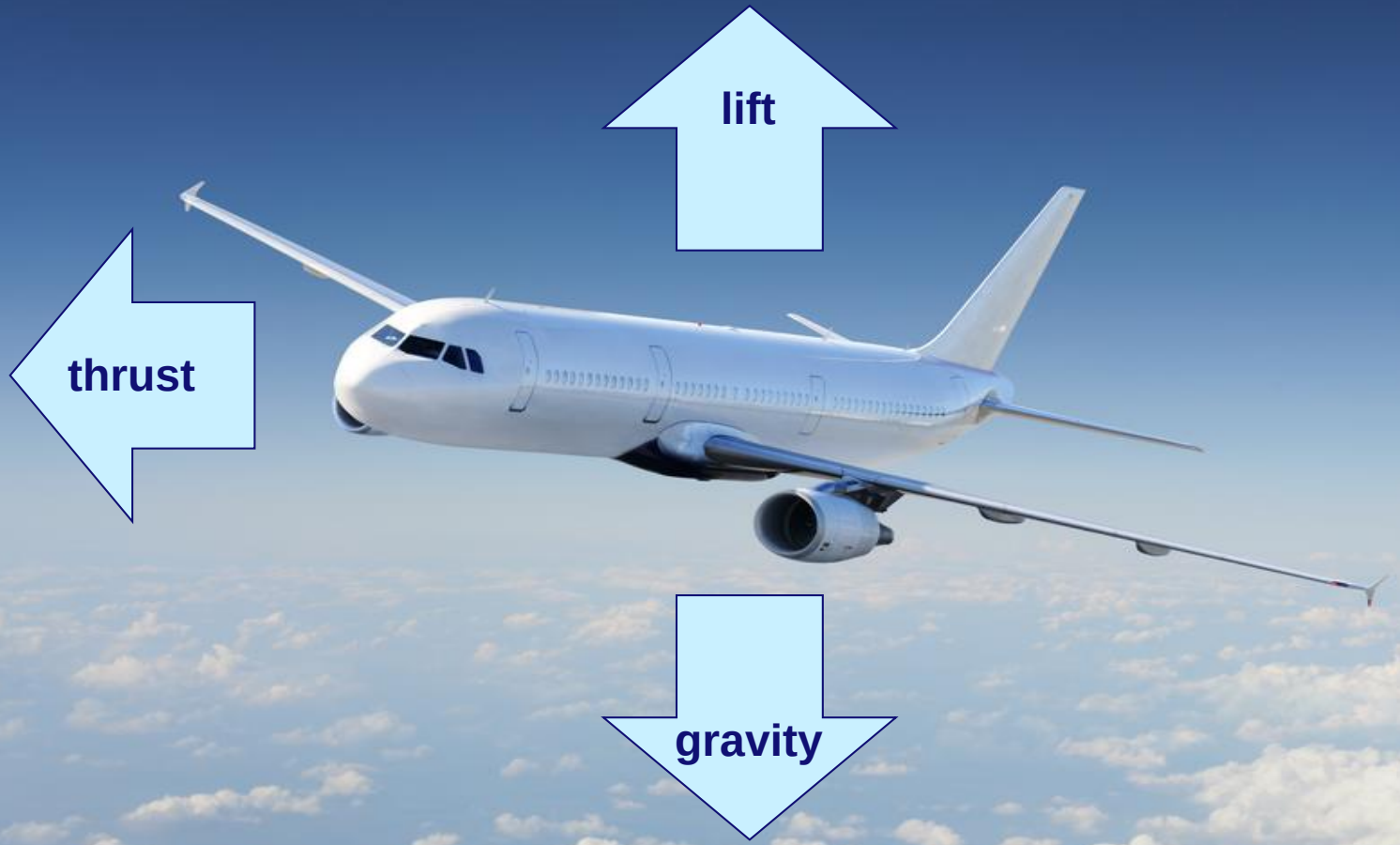
TU/e

Four Forces

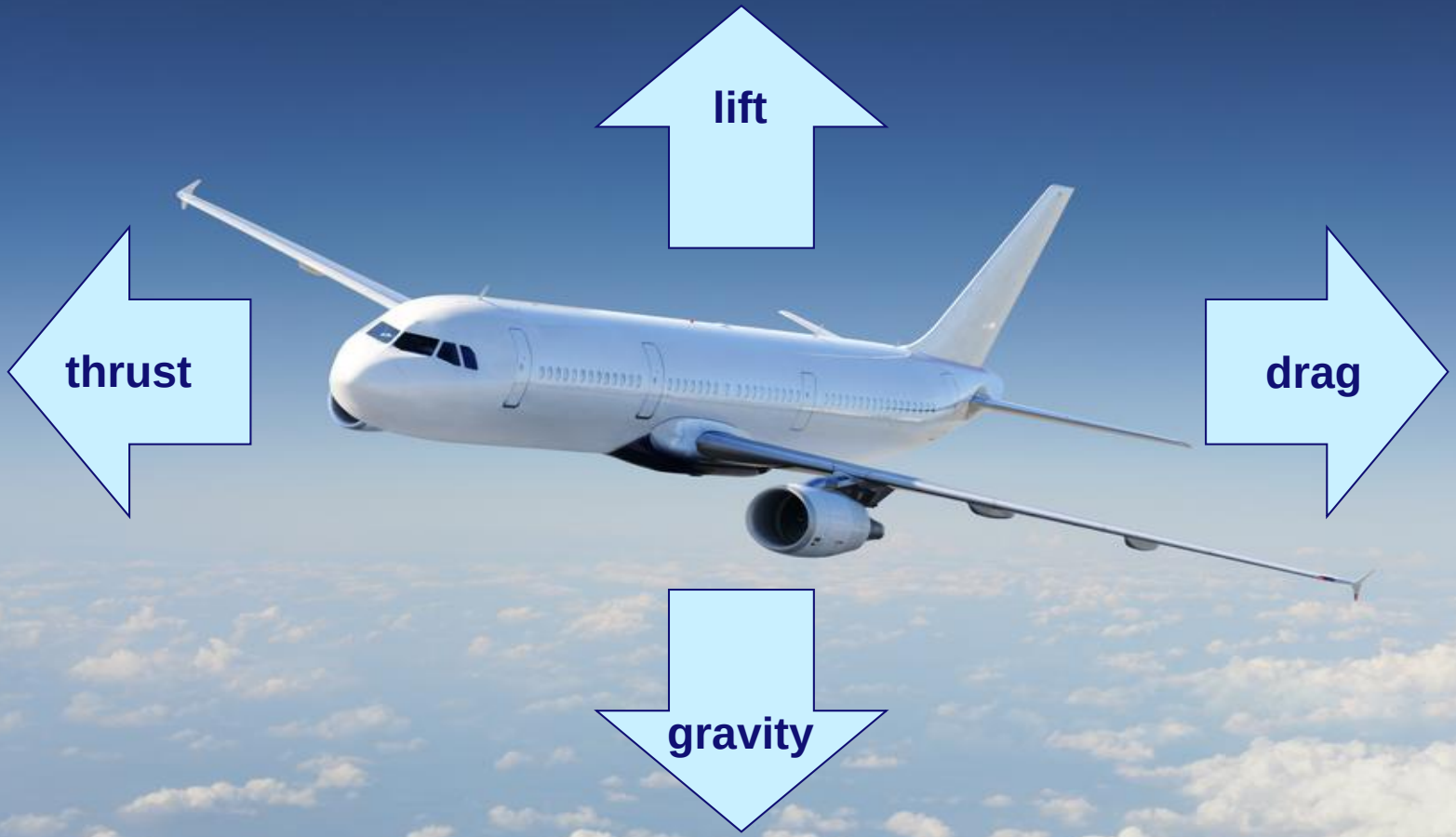
Four Forces



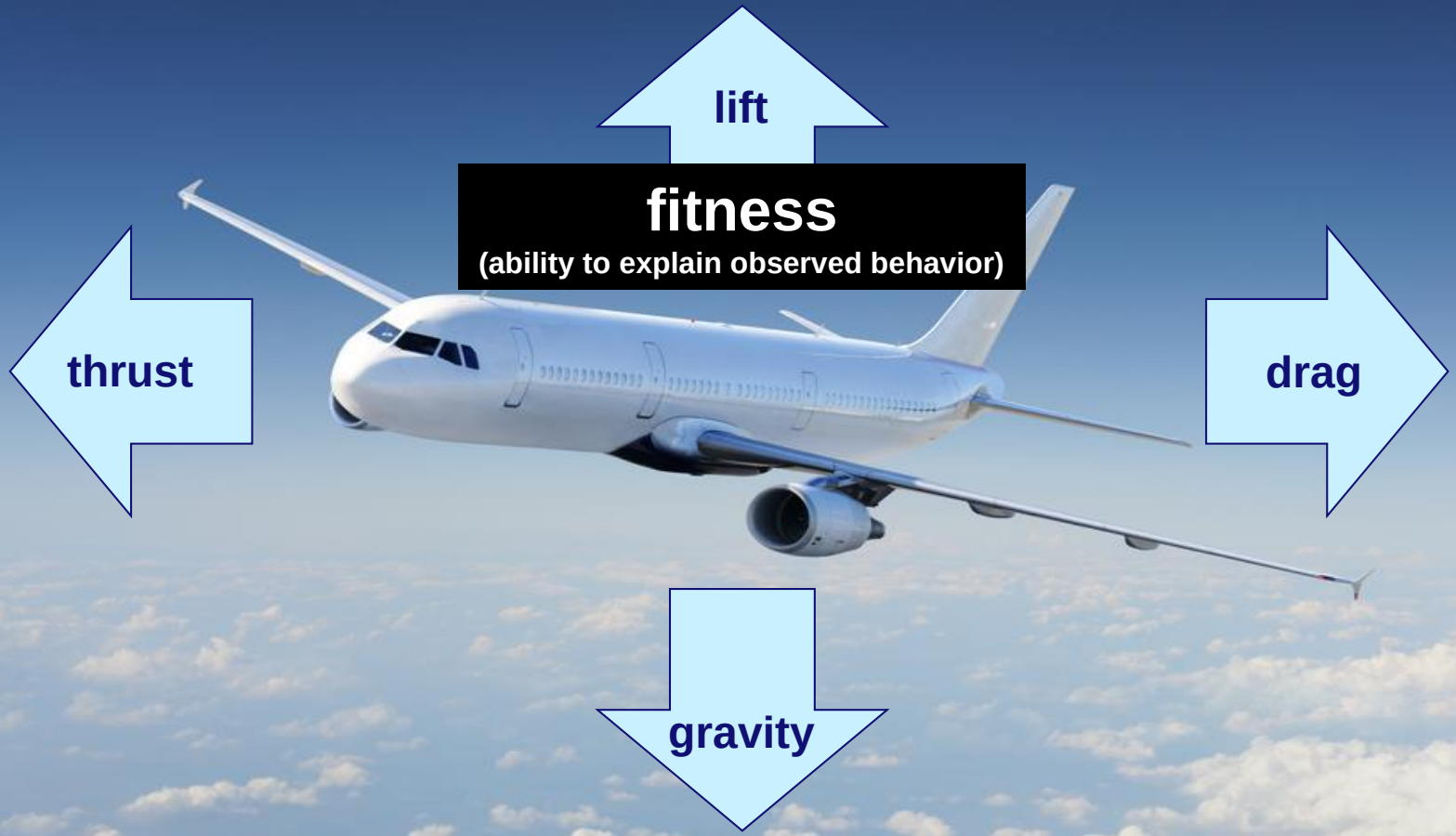
Four Forces



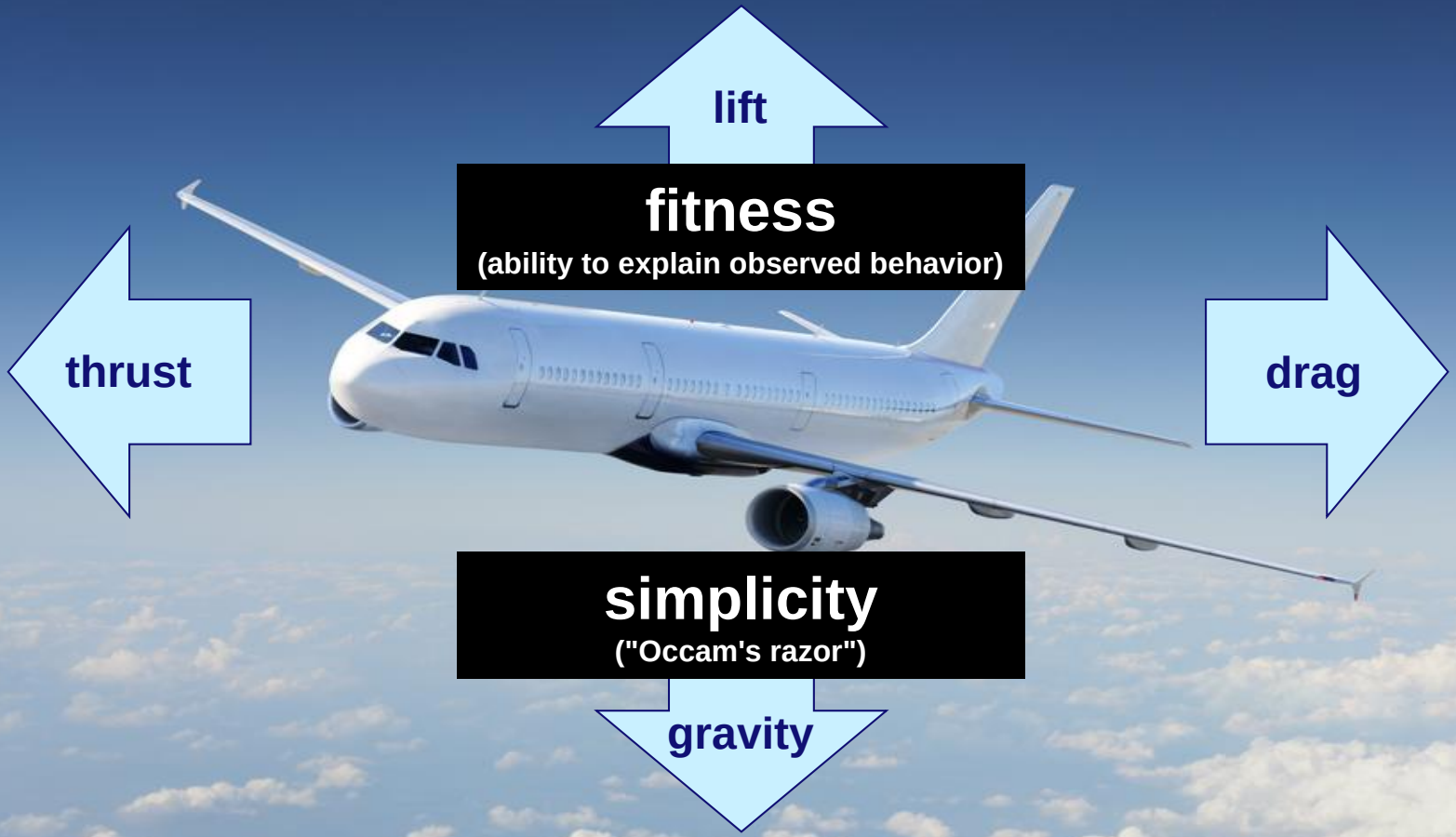
Four Forces



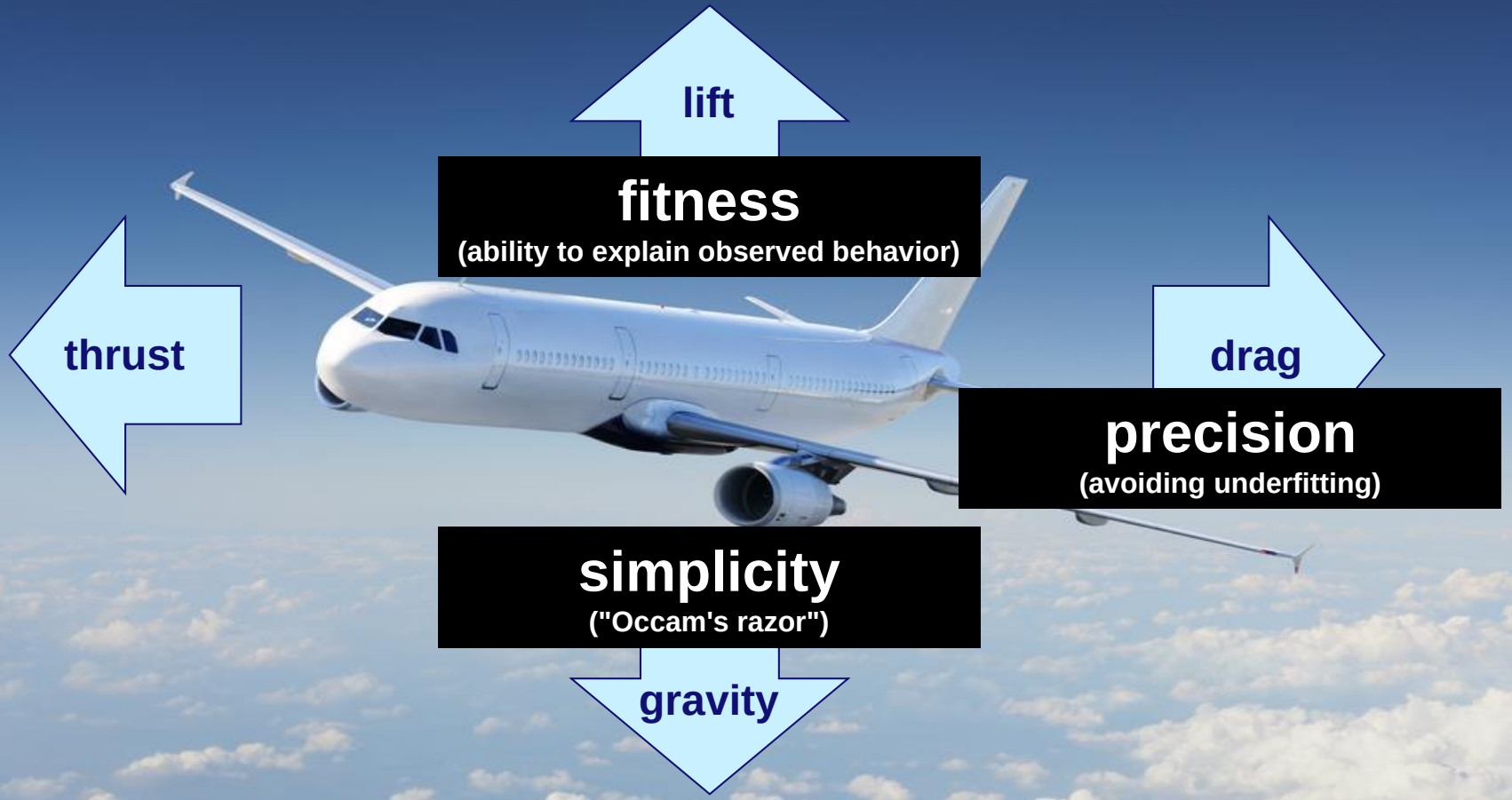
Four Forces


lift
fitness
(ability to explain observed behavior)
thrust
drag
gravity

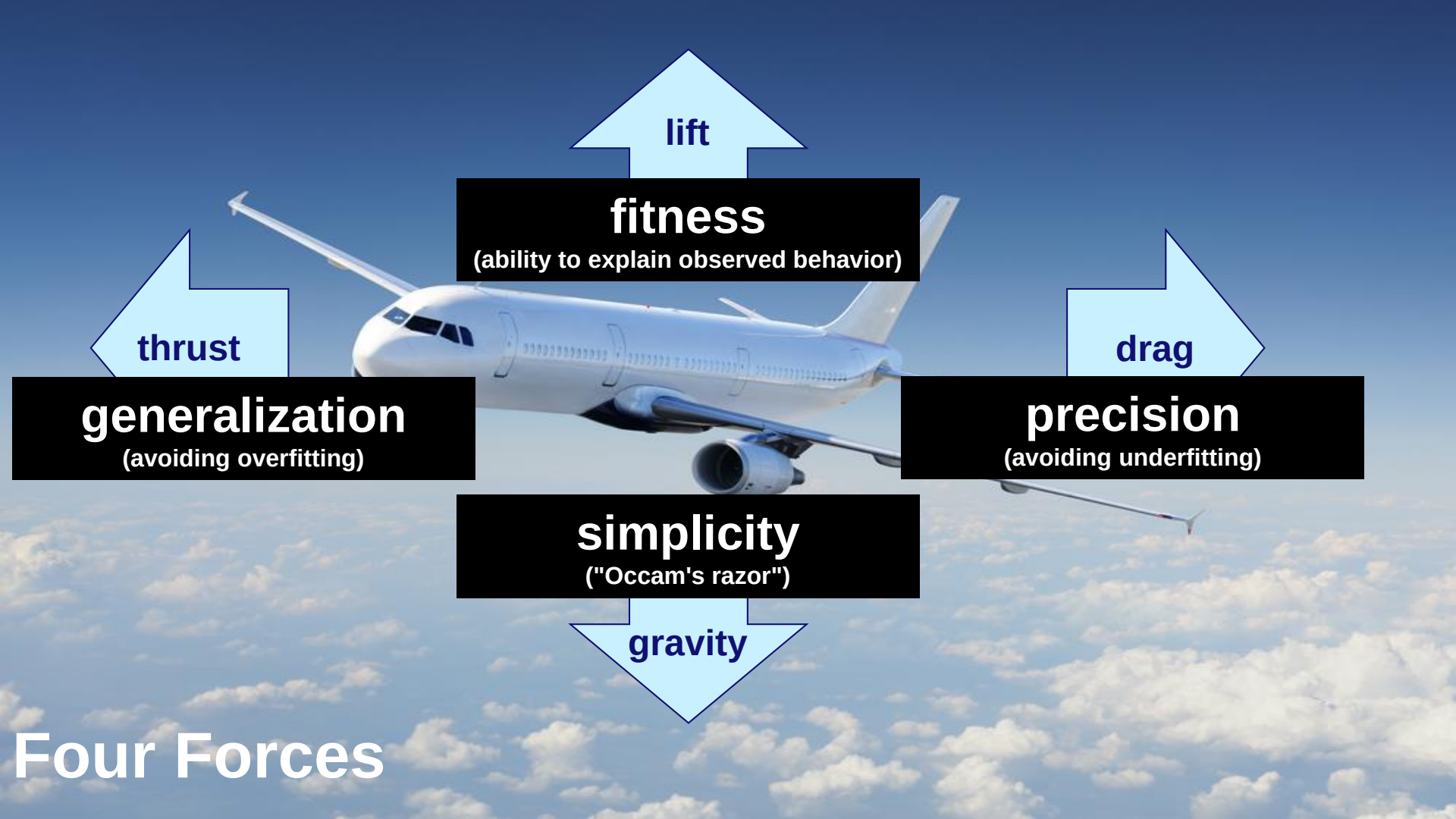
Four Forces


lift
fitness
(ability to explain observed behavior)
thrust
drag
simplicity
("Occam's razor")
gravity

Four Forces


lift
fitness
(ability to explain observed behavior)
thrust
drag
precision
(avoiding underfitting)
simplicity
("Occam's razor")
gravity

Four Forces


lift
fitness
(ability to explain observed behavior)
thrust
drag
generalization
(avoiding overfitting)
precision
(avoiding underfitting)
simplicity
("Occam's razor")
gravity

Four Forces

Example log

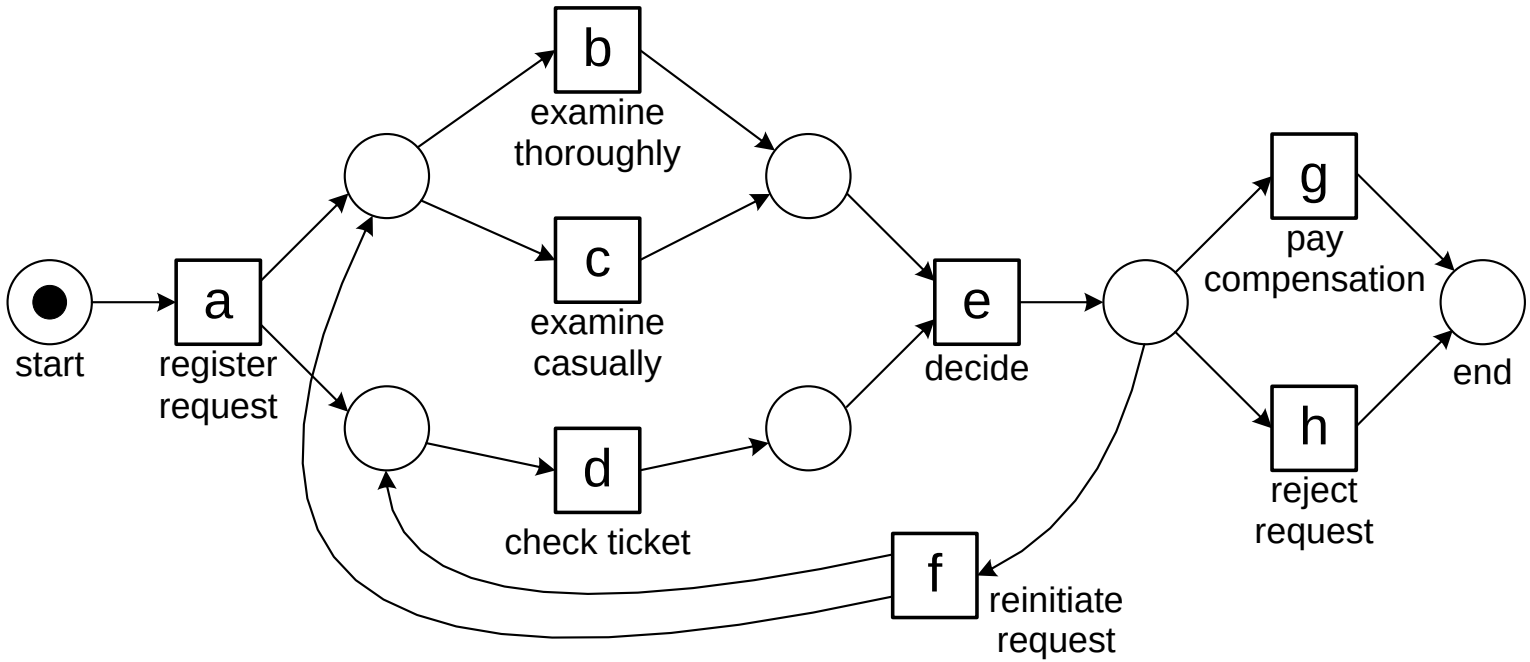
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh

Exar

2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

TU/e

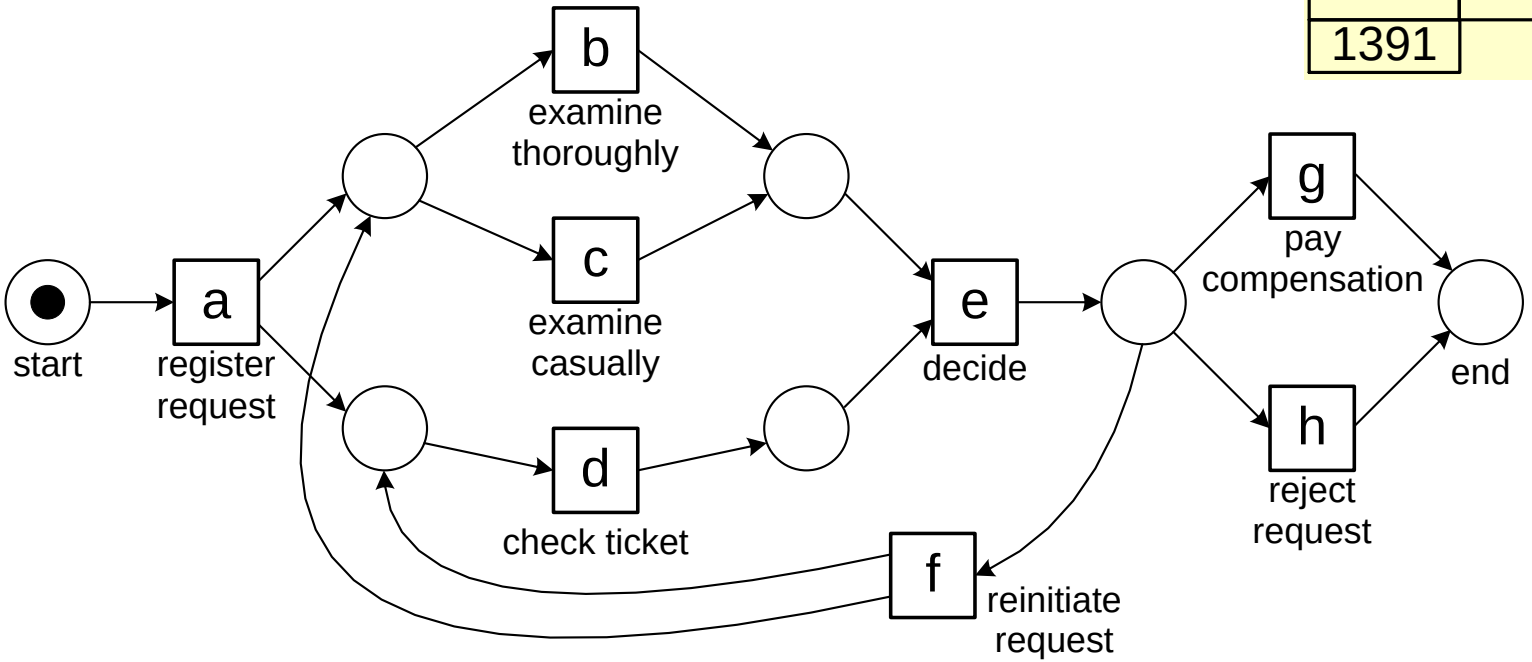
Model that seems to be OK …



#	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	

Model that seems to be OK …

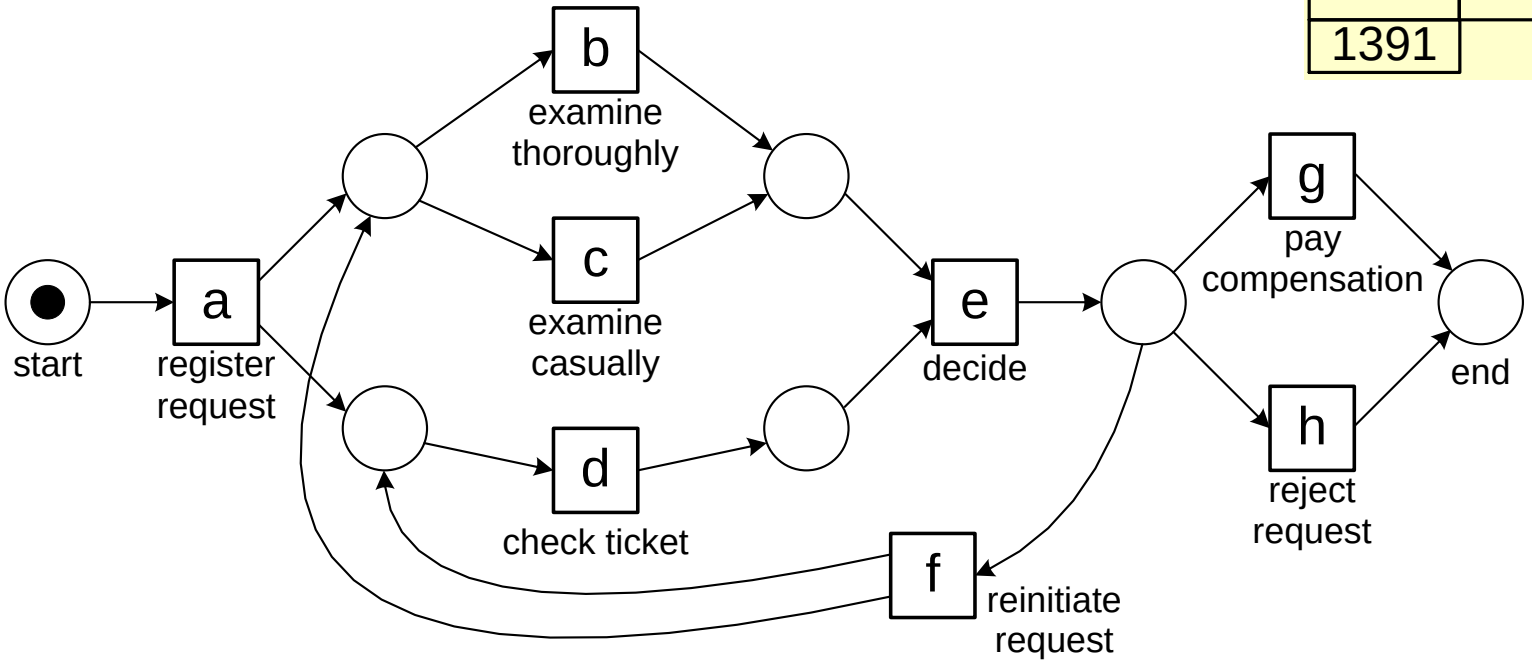
#	trace
455	acdeh
191	abdeg
...	...
1	adcefdbefcdefdbeg
1391	



TU/e

Model that seems to be OK …

#	trace
455	acdeh
191	abdeg
...	...
1	adcefdbefcdefdbeg
1391	

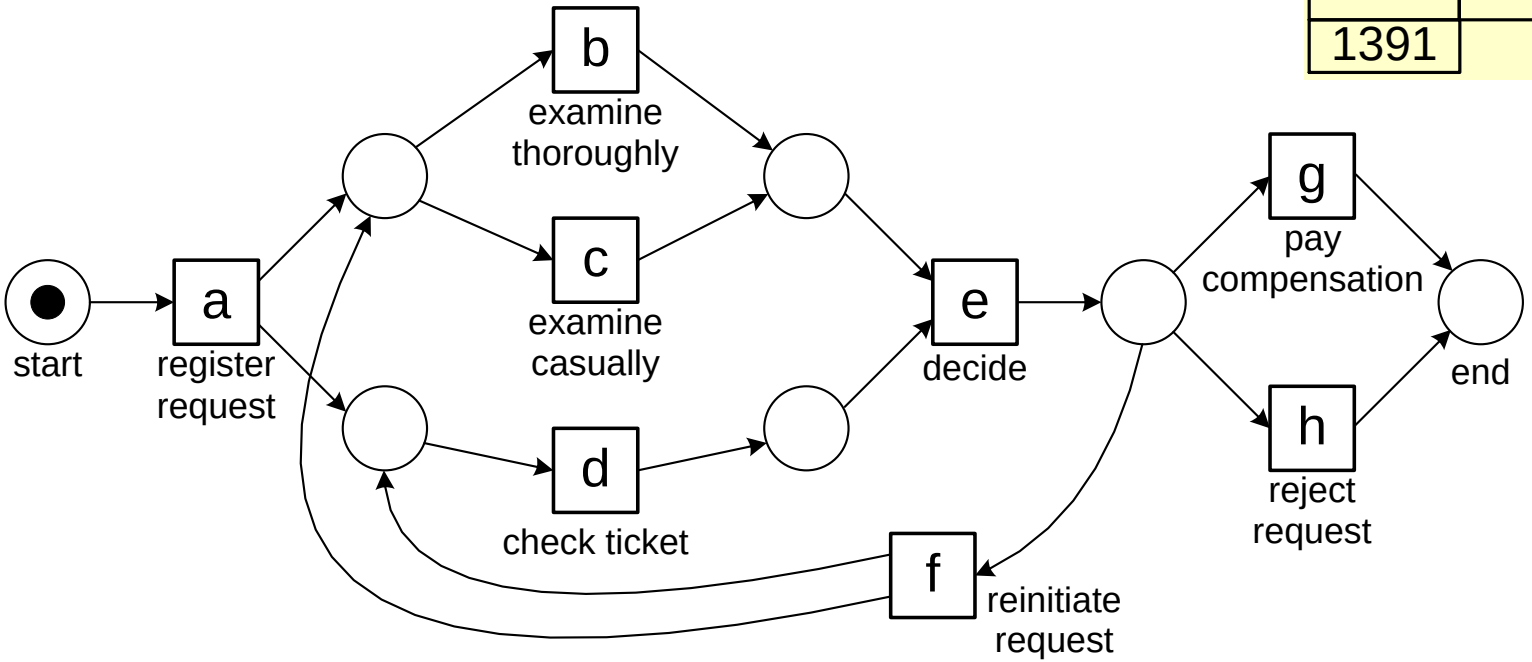


fitness
(observed behavior fits)

TU/e

Model that seems to be OK …

#	trace
455	acdeh
191	abdeg
...	...
1	adcefdbefcdefdbeg
1391	



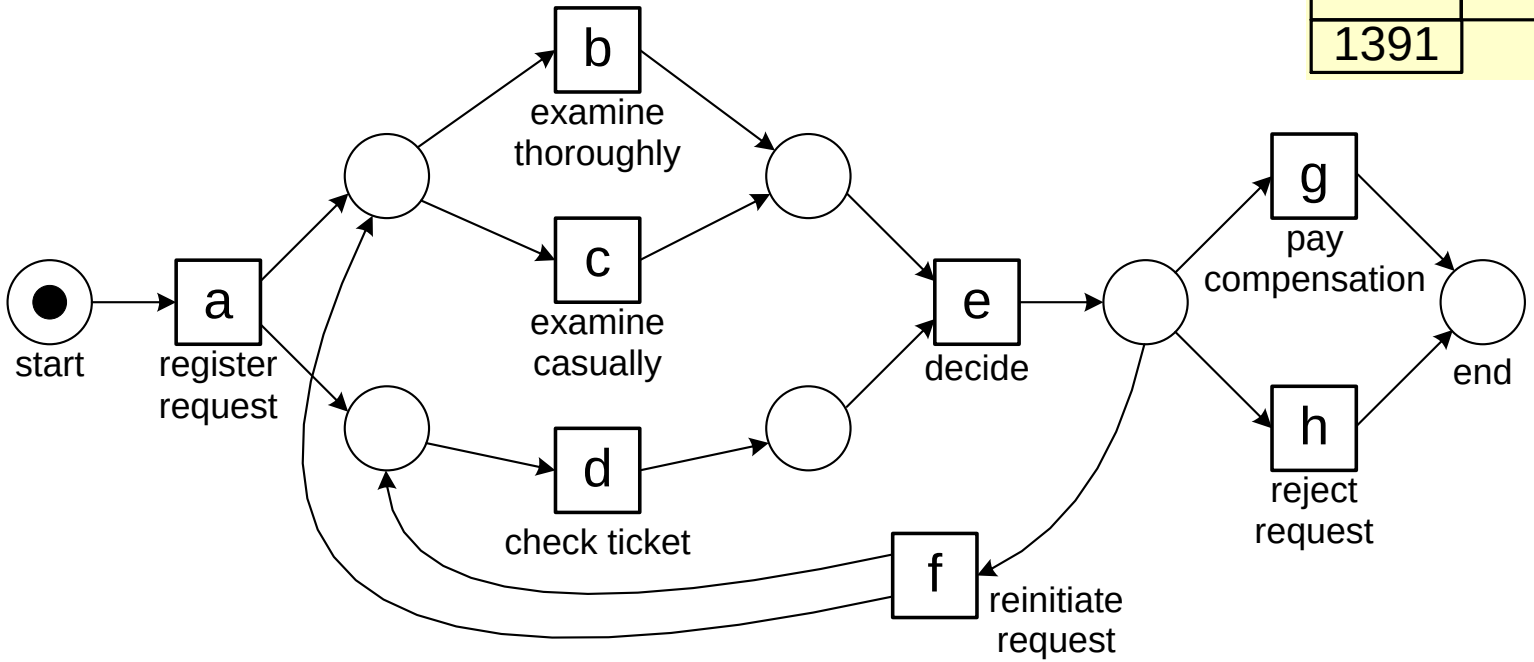
fitness
(observed behavior fits)

simplicity
("Occam's razor")

TU/e

Model that seems to be OK …

#	trace
455	acdeh
191	abdeg
...	...
1	adcefdbefcdefdbeg
1391	



fitness (observed behavior fits)

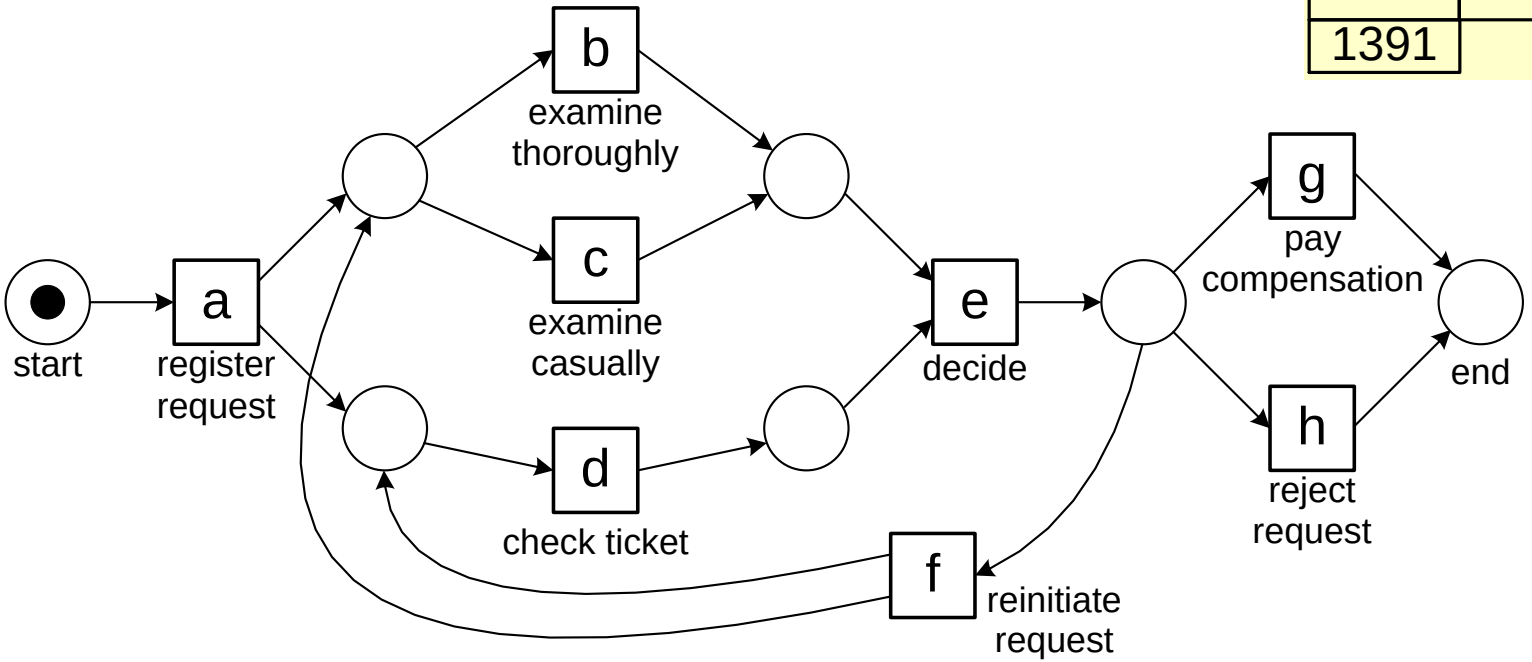
simplicity ("Occam's razor")

precision (avoiding underfitting)

TU/e

Model that seems to be OK …

#	trace
455	acdeh
191	abdeg
...	...
1	adcefdbefcdefdbeg
1391	



fitness (observed behavior fits)

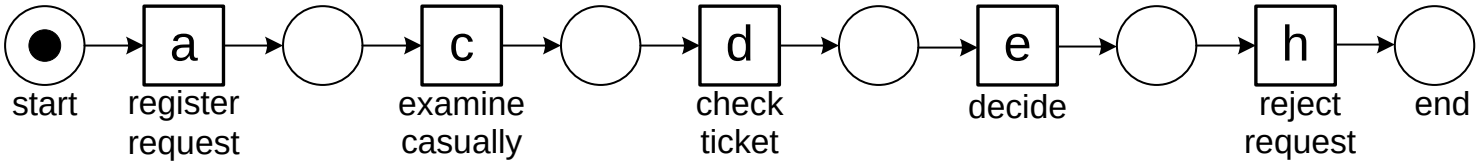
simplicity ("Occam's razor")

precision (avoiding underfitting)

generalization (avoiding overfitting)

TU/e

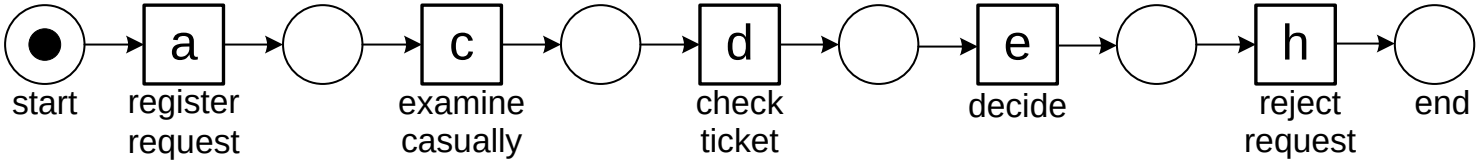
Non-fitting model



#	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	

Non-fitting model

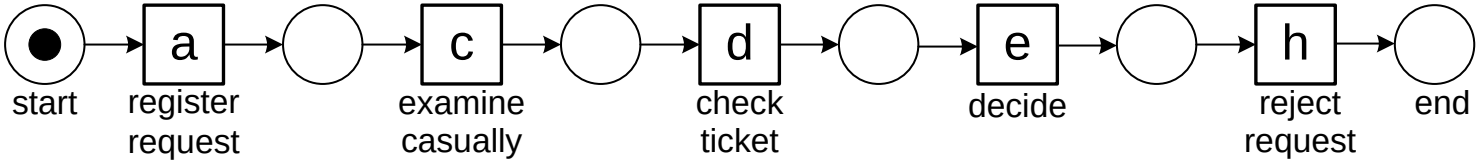
#	trace
455	acdeh
191	abdeg
...	...
1	adcefdbefcdefdbeg
1391	



TU/e

Non-fitting model

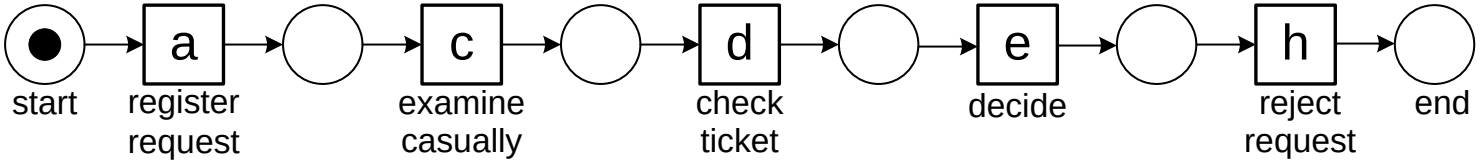
#	trace
455	acdeh
191	abdeg
...	...
1	adcefdbefcdefdbeg
1391	



fitness
(observed behavior fits)

Non-fitting model

#	trace
455	acdeh
191	abdeg
...	...
1	adcefdbefcdefdbeg
1391	



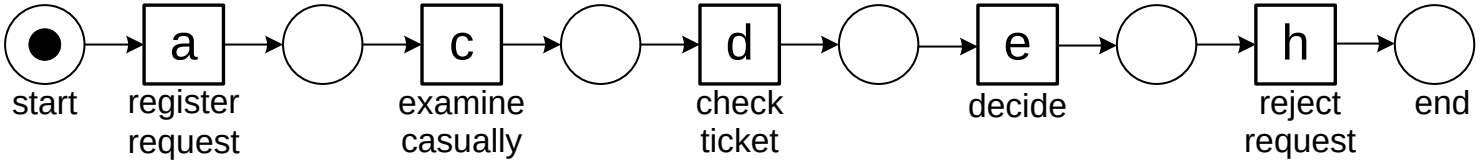
fitness
(observed behavior fits)

simplicity
("Occam's razor")

TU/e

Non-fitting model

#	trace
455	acdeh
191	abdeg
...	...
1	adcefdbefcdefdbeg
1391	



fitness
(observed behavior fits)

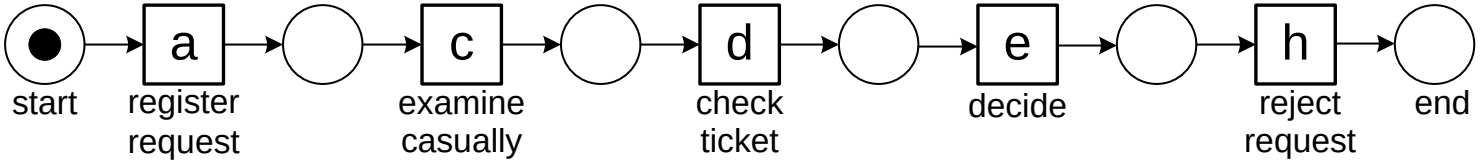
simplicity
("Occam's razor")

precision
(avoiding underfitting)

TU/e

Non-fitting model

#	trace
455	acdeh
191	abdeg
...	...
1	adcefdbefcdefdbeg
1391	



fitness (observed behavior fits)

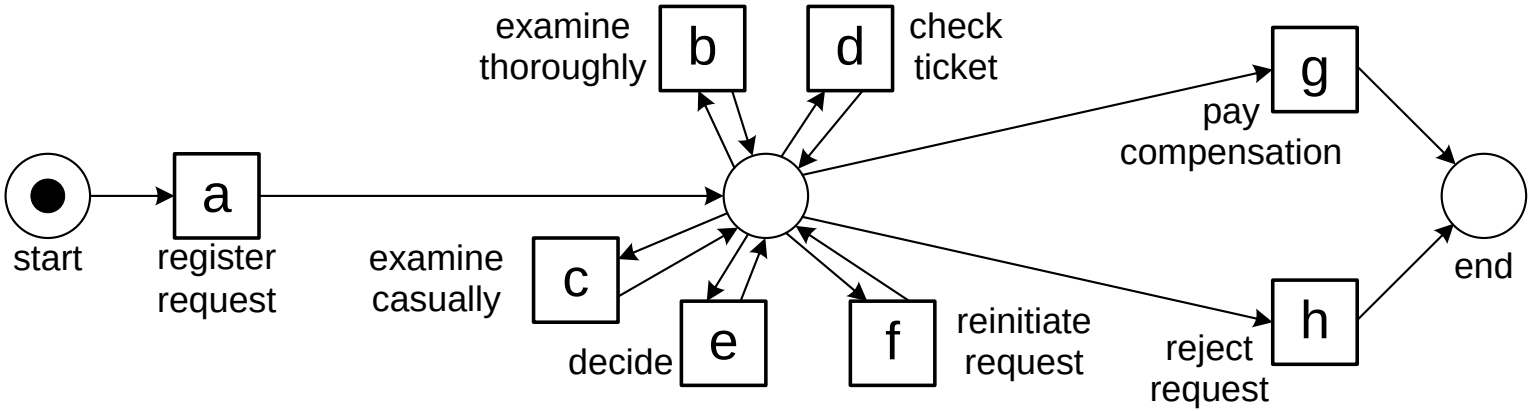
simplicity ("Occam's razor")

precision (avoiding underfitting)

generalization (avoiding overfitting)

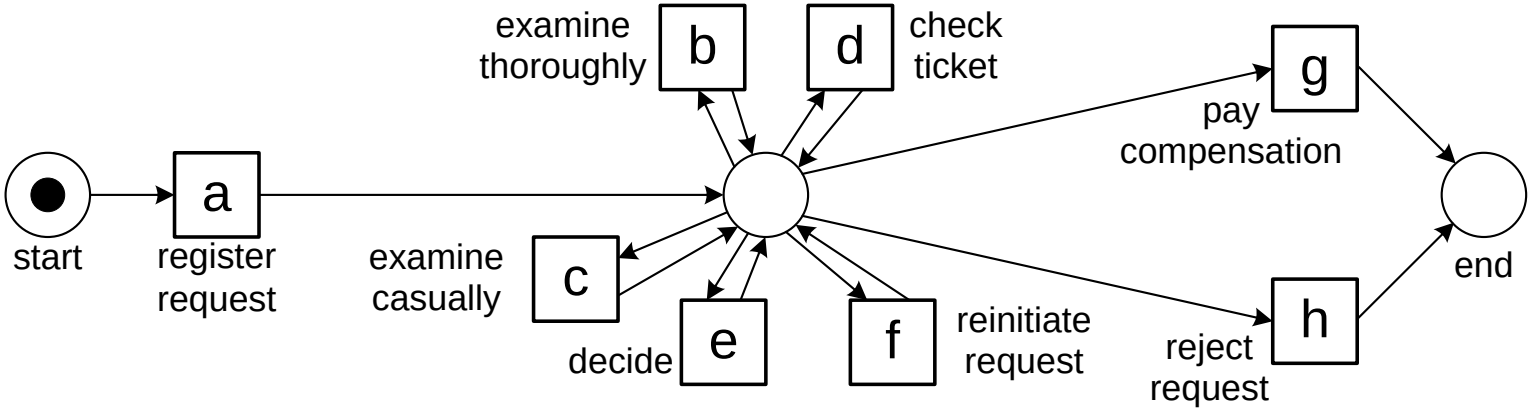
TU/e

Underfitting model



#	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	

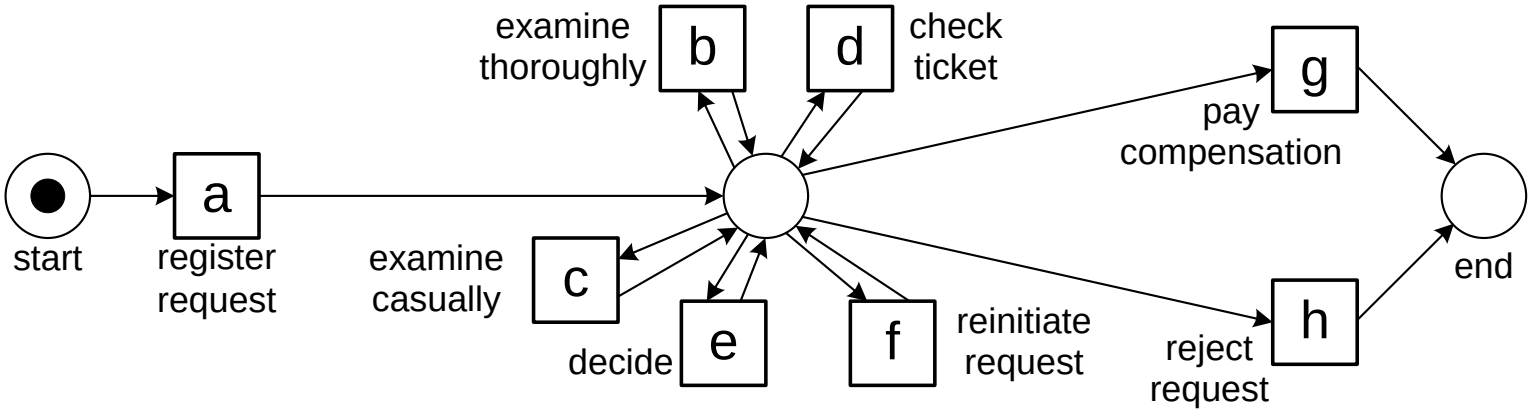
Underfitting model



#	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	

fitness
(observed behavior fits)

Underfitting model

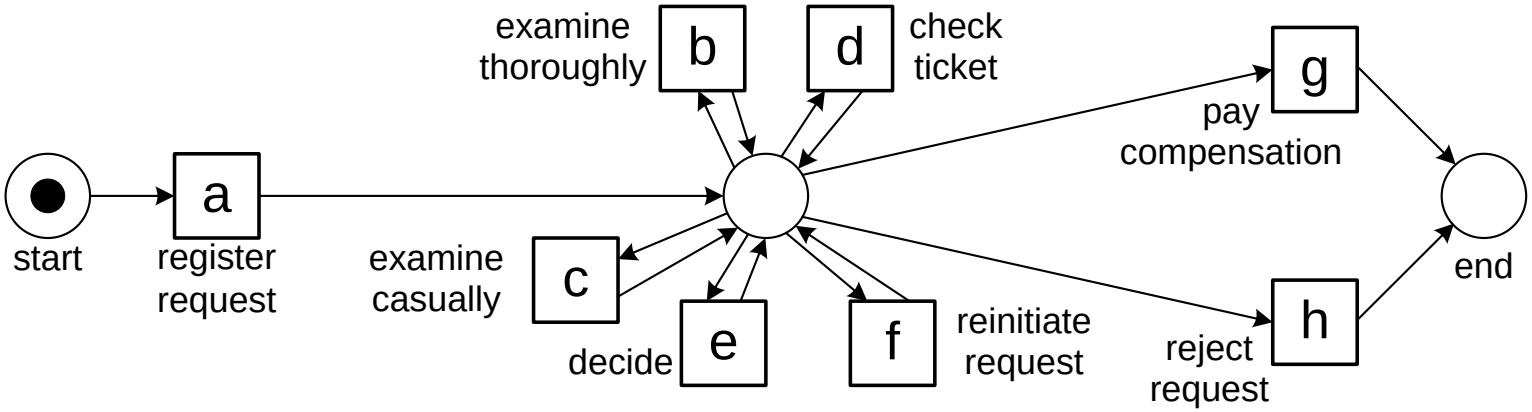


#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

fitness
(observed behavior fits)

simplicity
("Occam's razor")

Underfitting model



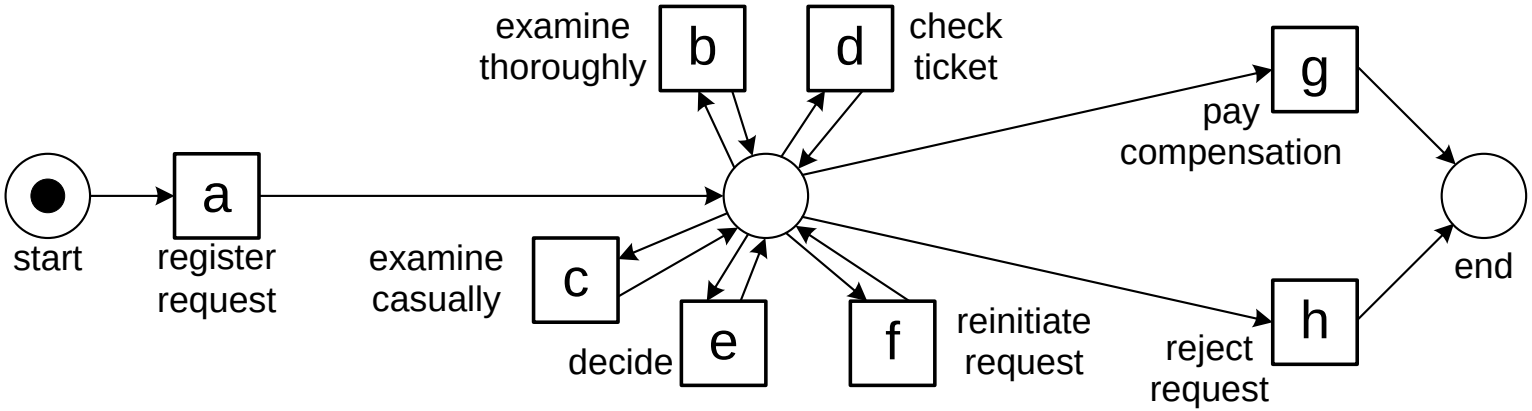
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

fitness
(observed behavior fits)

simplicity
("Occam's razor")

precision
(avoiding underfitting)

Underfitting model



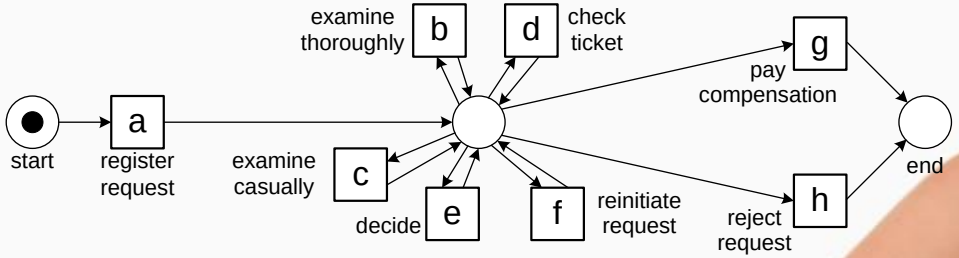
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
391	

fitness
(observed behavior fits)

simplicity
("Occam's razor")

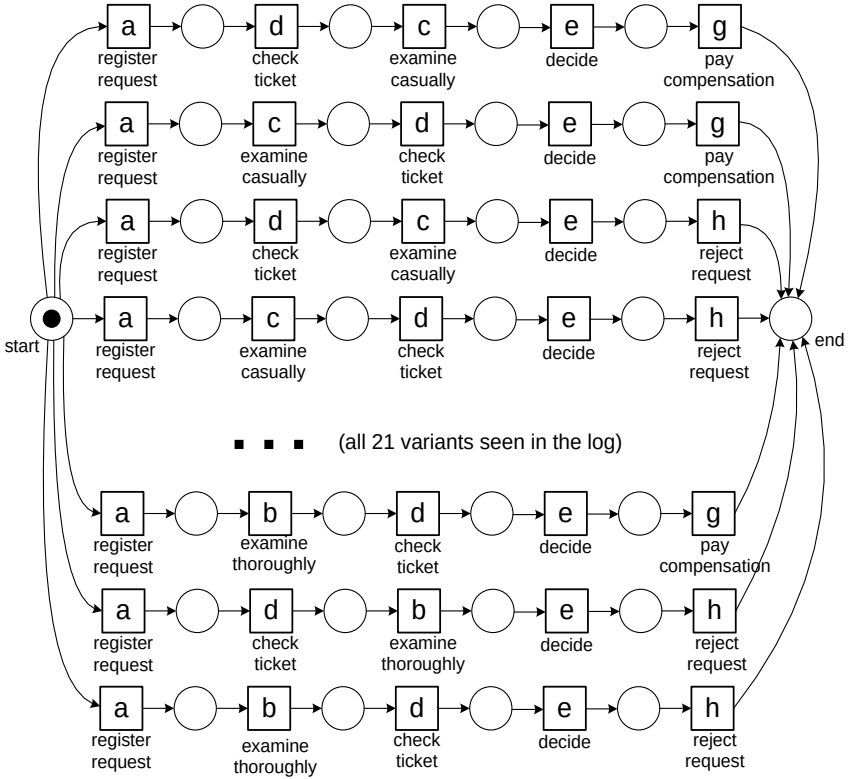
precision
(avoiding underfitting)

generalization
(avoiding overfitting)



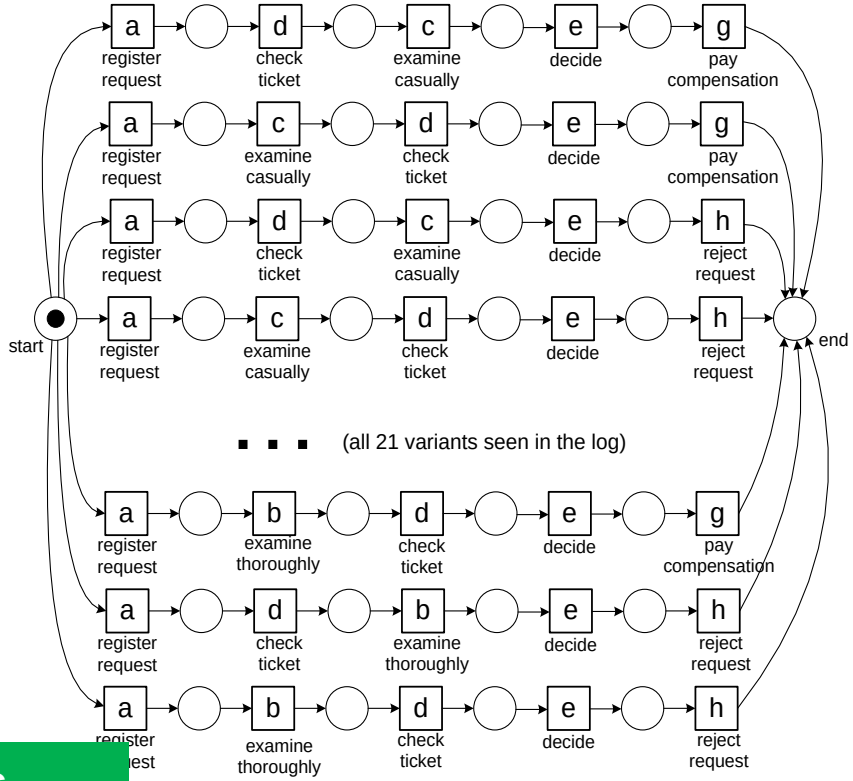
underfitting

Overfitting model



#	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	

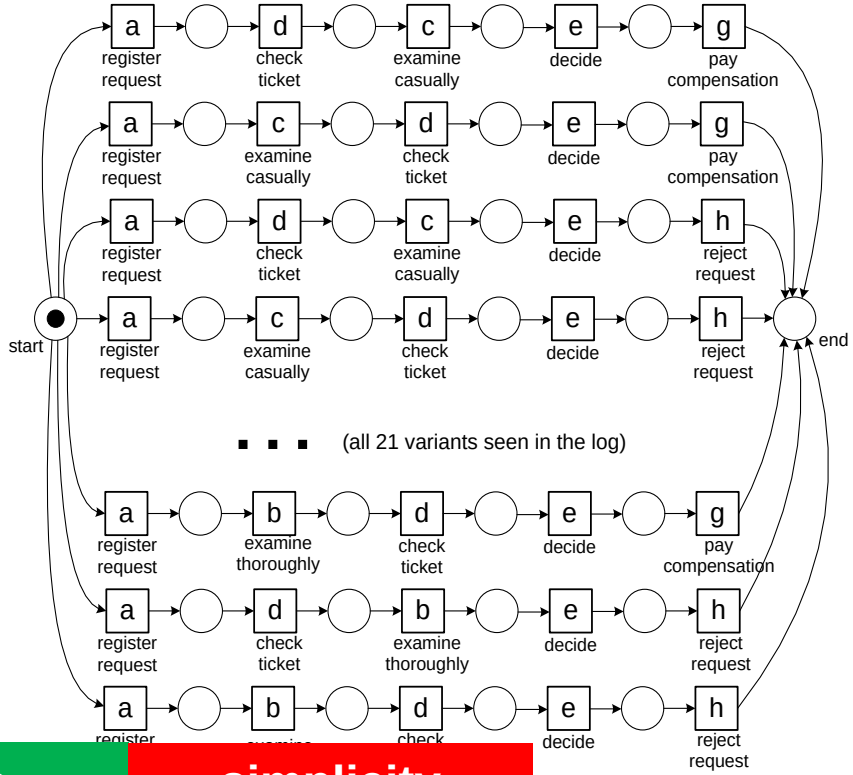
Overfitting model



fitness
(observed behavior fits)

#	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	

Overfitting model

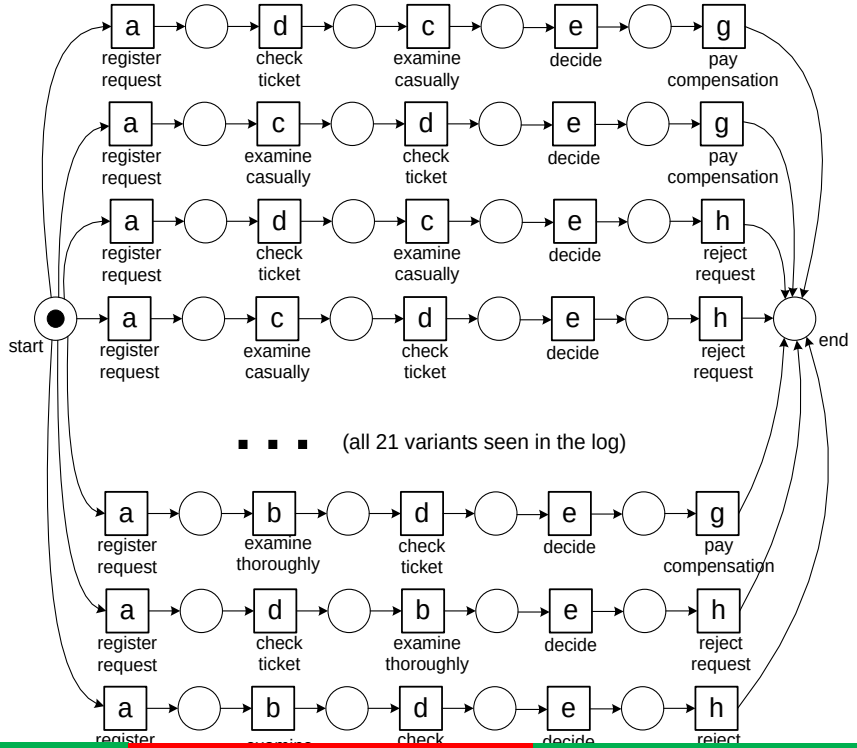


#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

fitness
(observed behavior fits)

simplicity
("Occam's razor")

Overfitting model



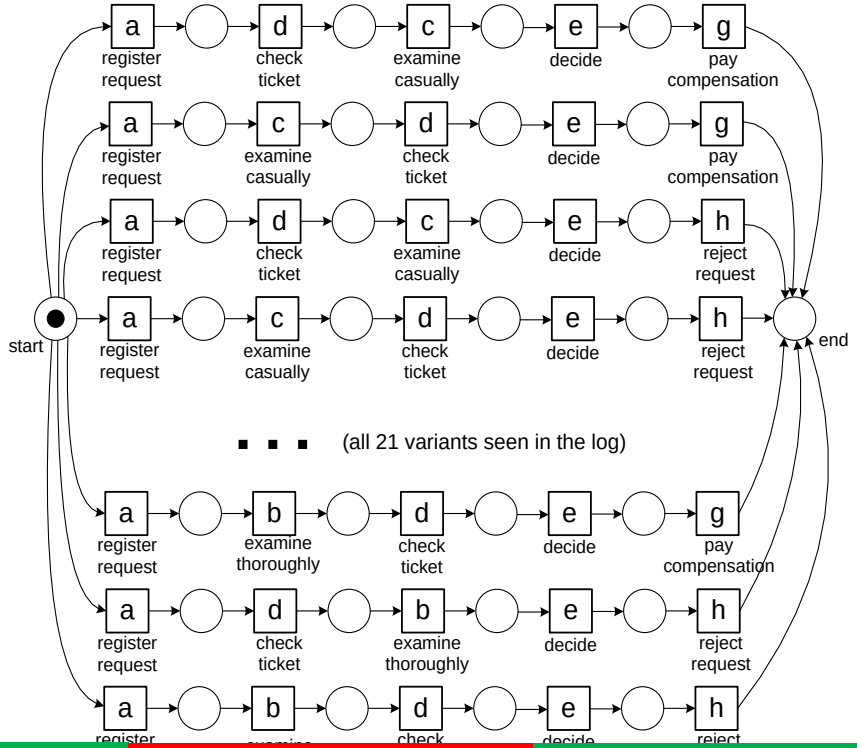
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

fitness
(observed behavior fits)

simplicity
("Occam's razor")

precision
(avoiding underfitting)

Overfitting model



#	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
391	

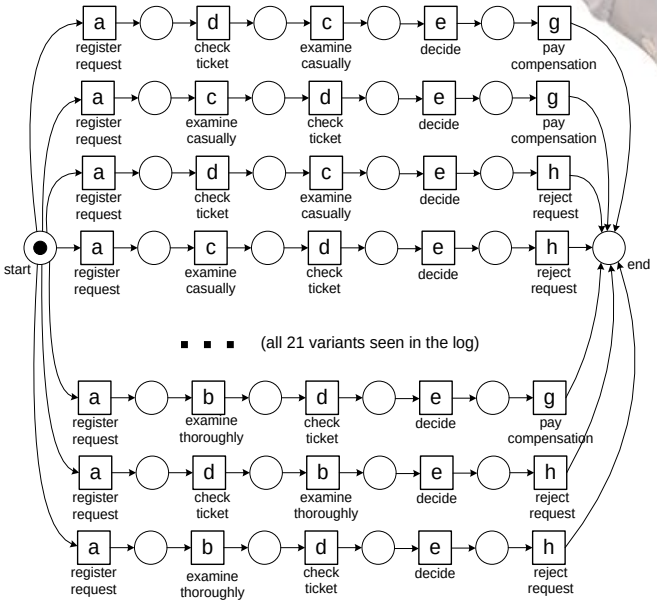
fitness
(observed behavior fits)

simplicity
("Occam's razor")

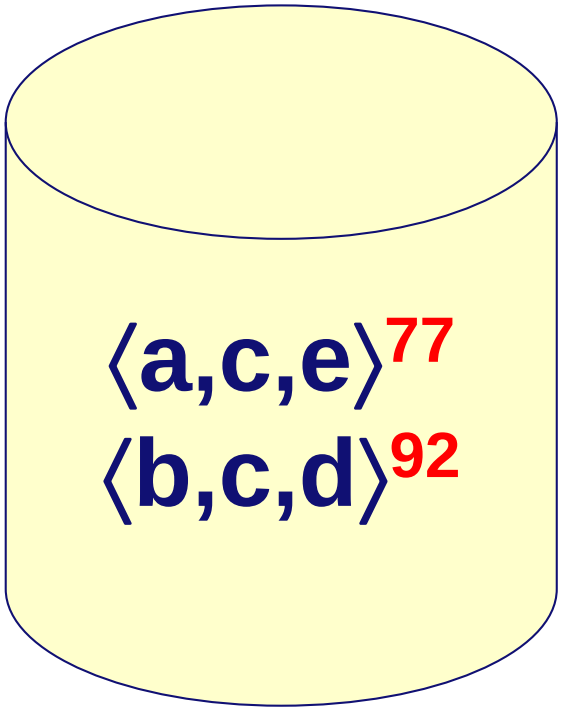
precision
(avoiding underfitting)

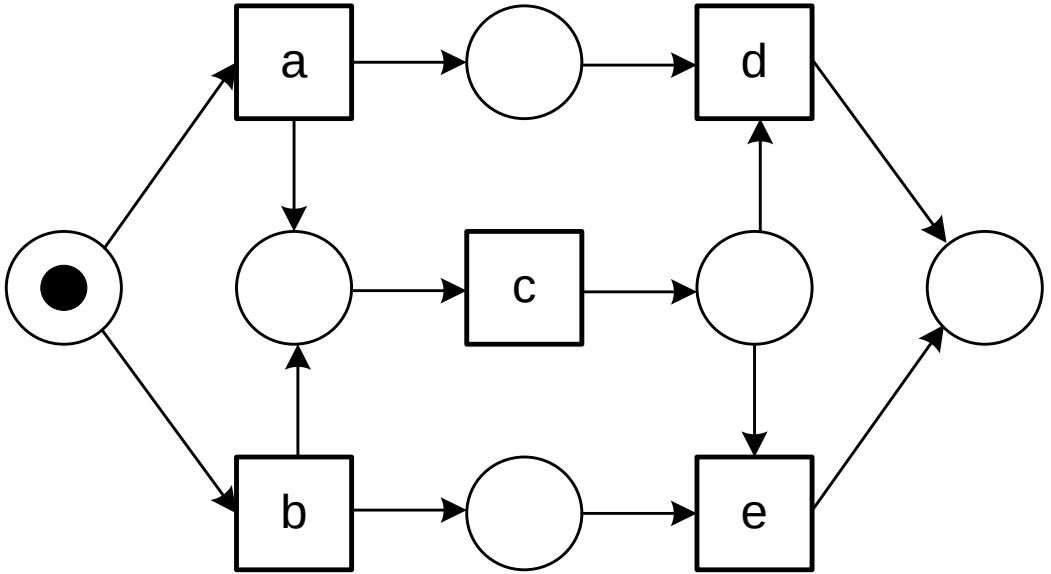
generalization
(avoiding overfitting)

overfitting

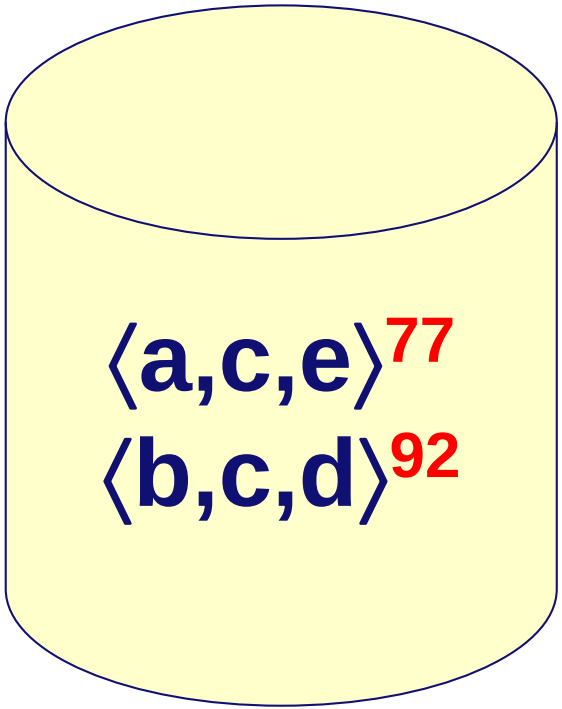


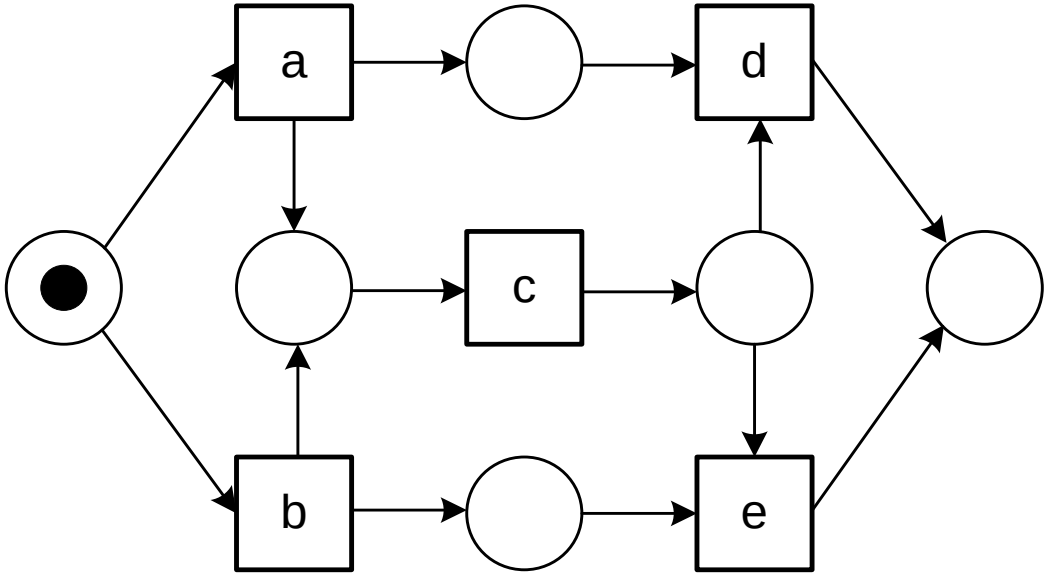
Fitness: good or bad?





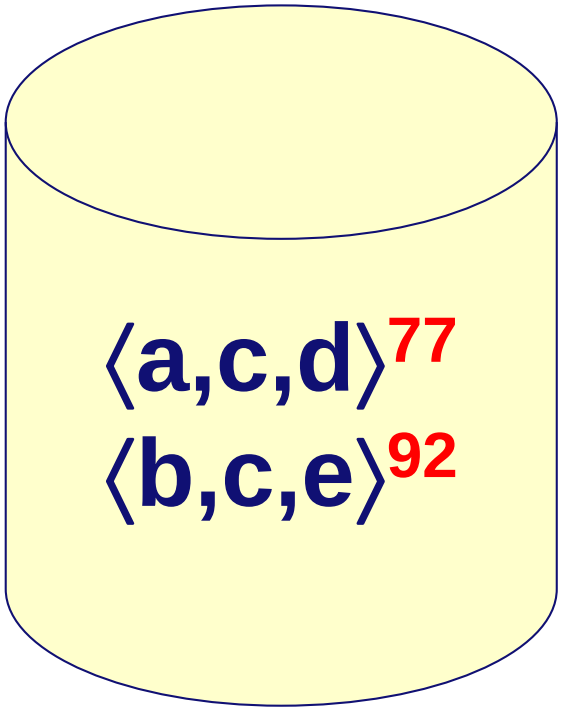
Fitness: bad!

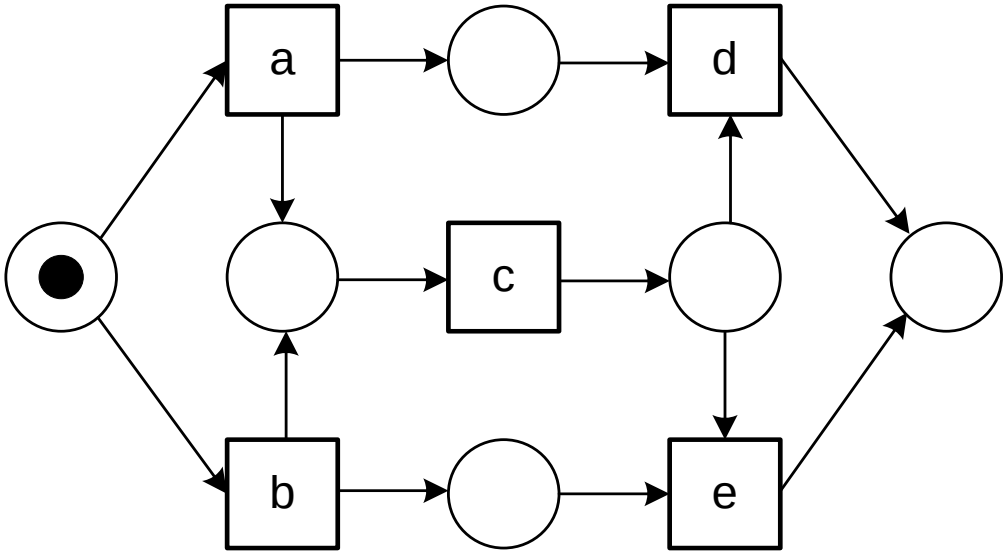




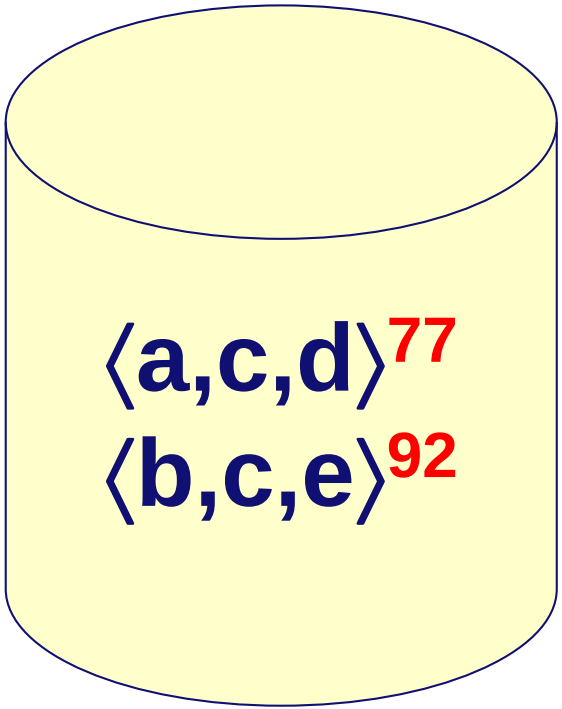
both traces do not fit …

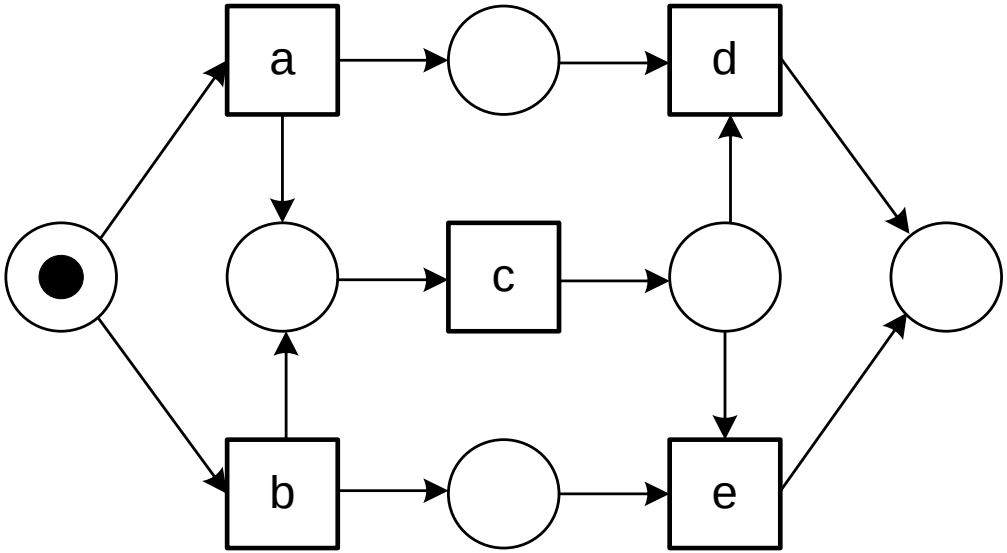
Precision: good or bad?





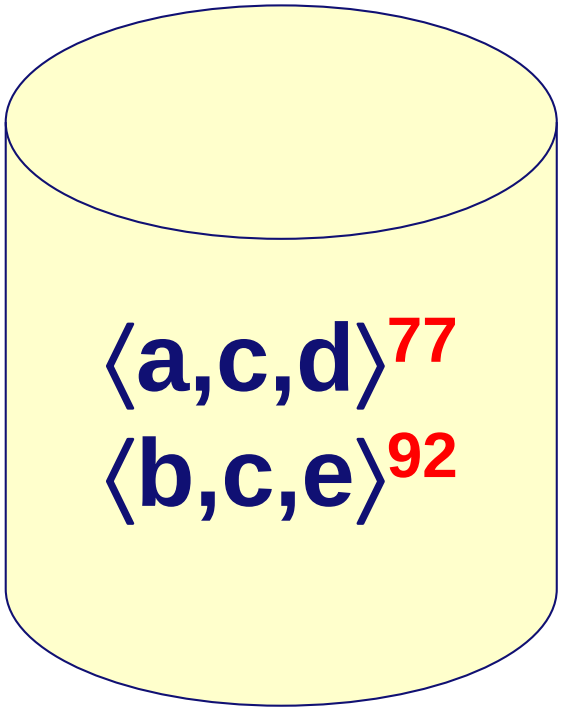
Precision: good!

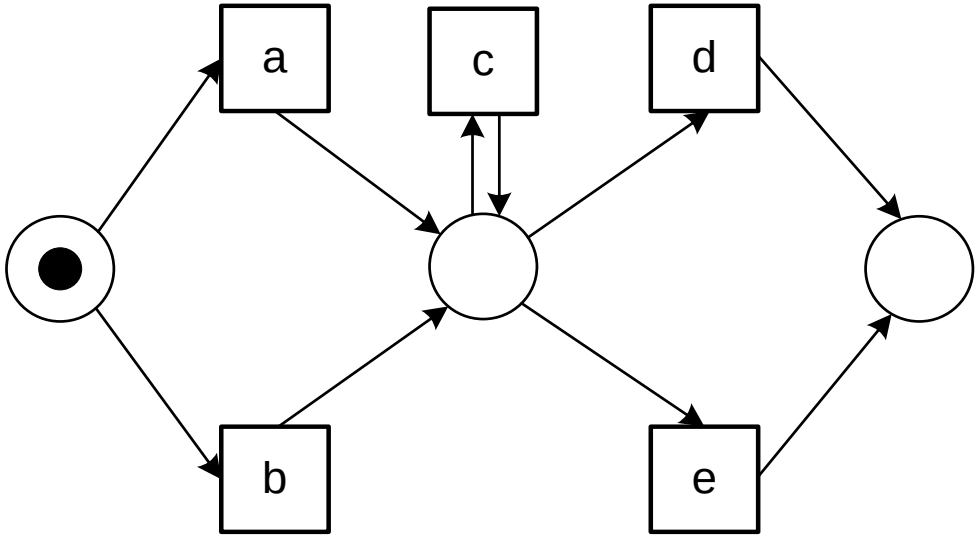




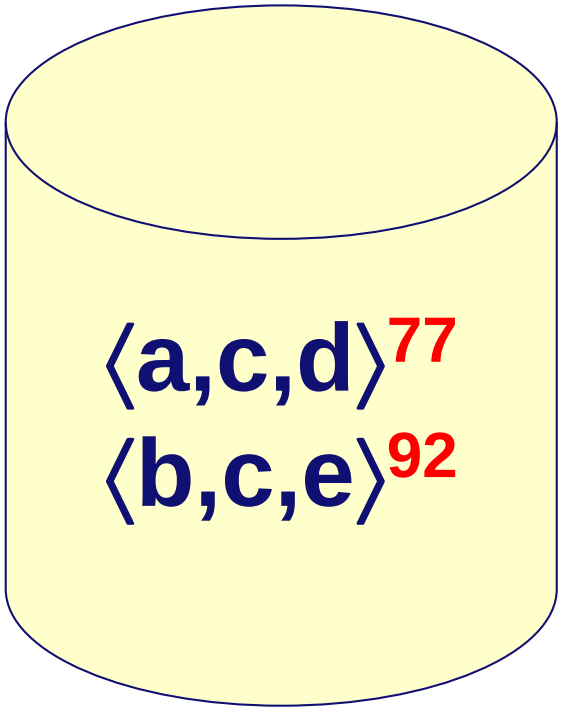
not underfitting…

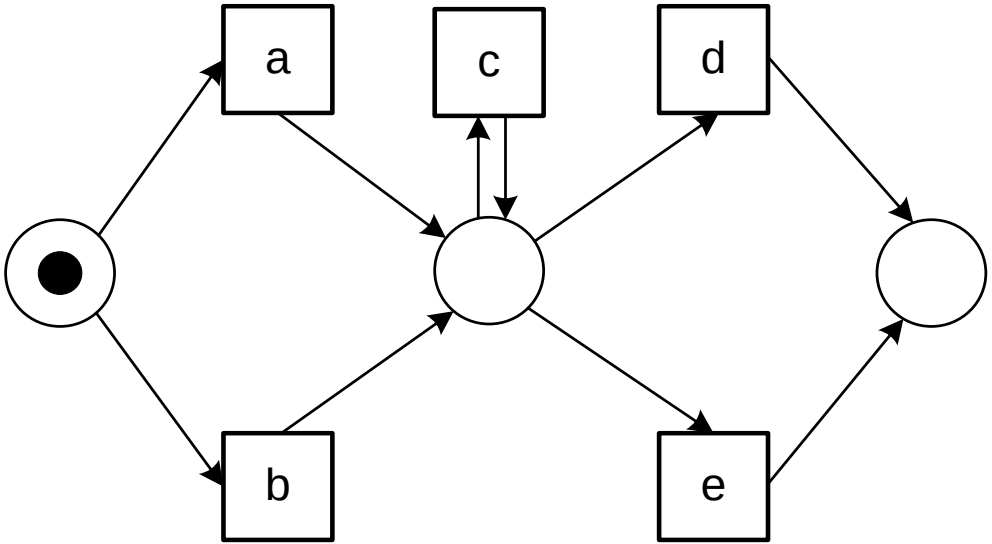
Precision: good or bad?





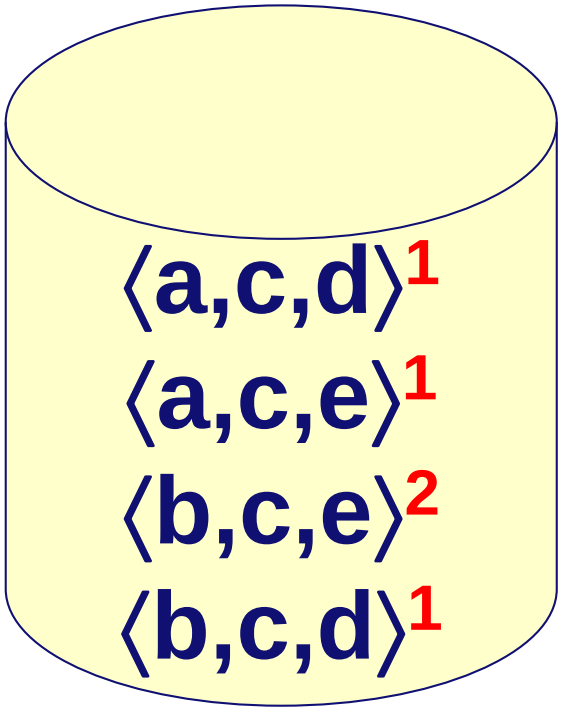
Precision: bad!

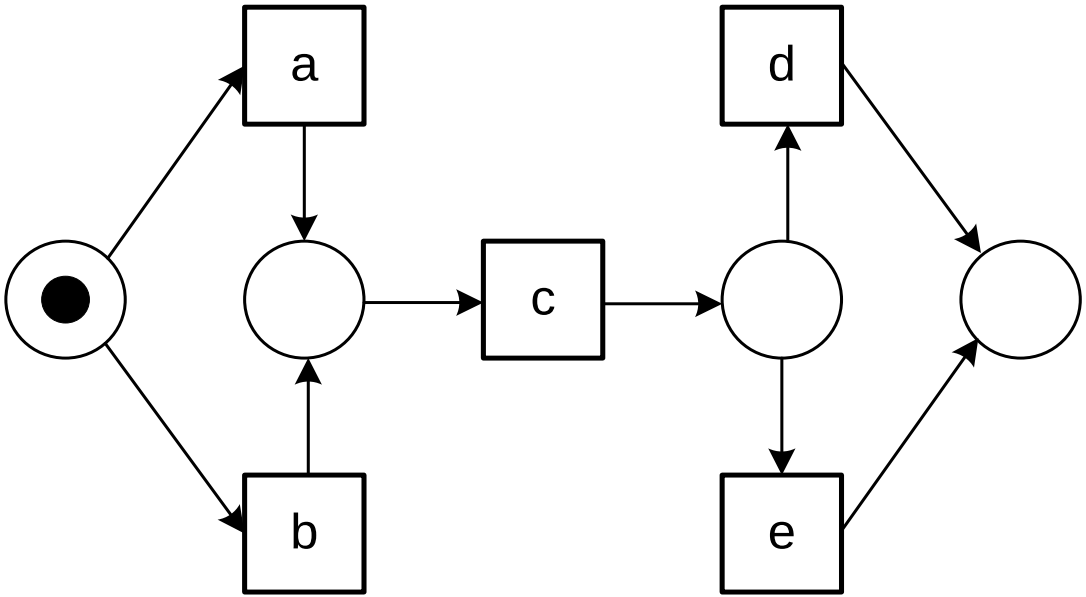




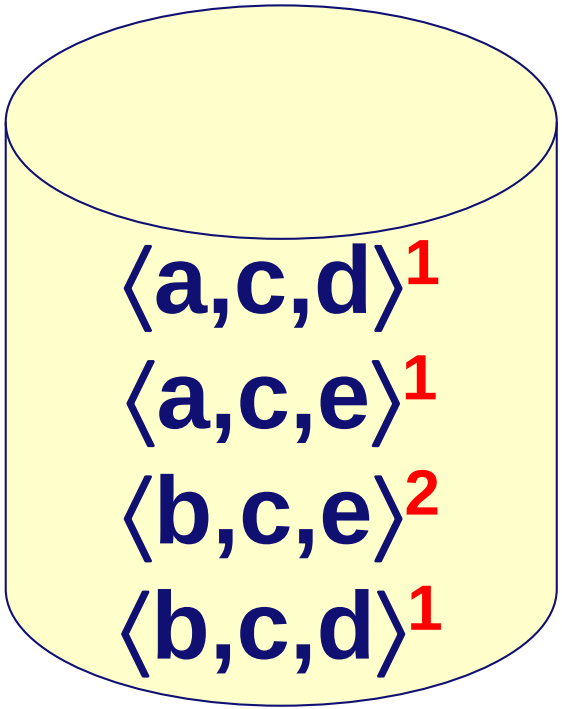
underfitting (allows for highly unlikely behavior) …

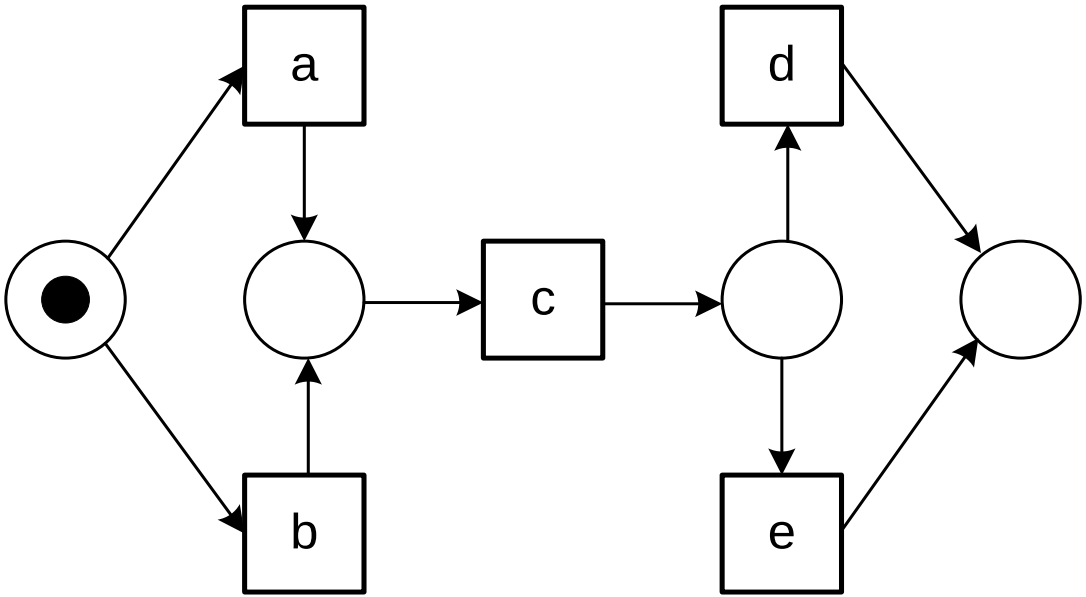
Generalization: good or bad?





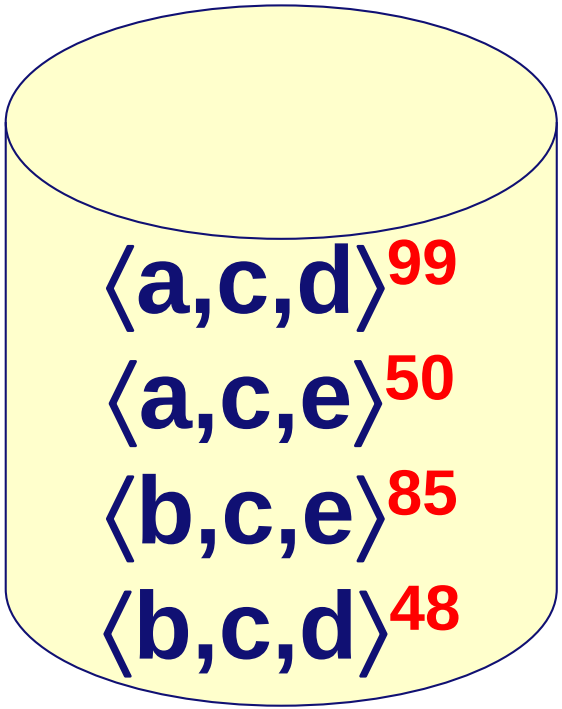
Generalization: bad!

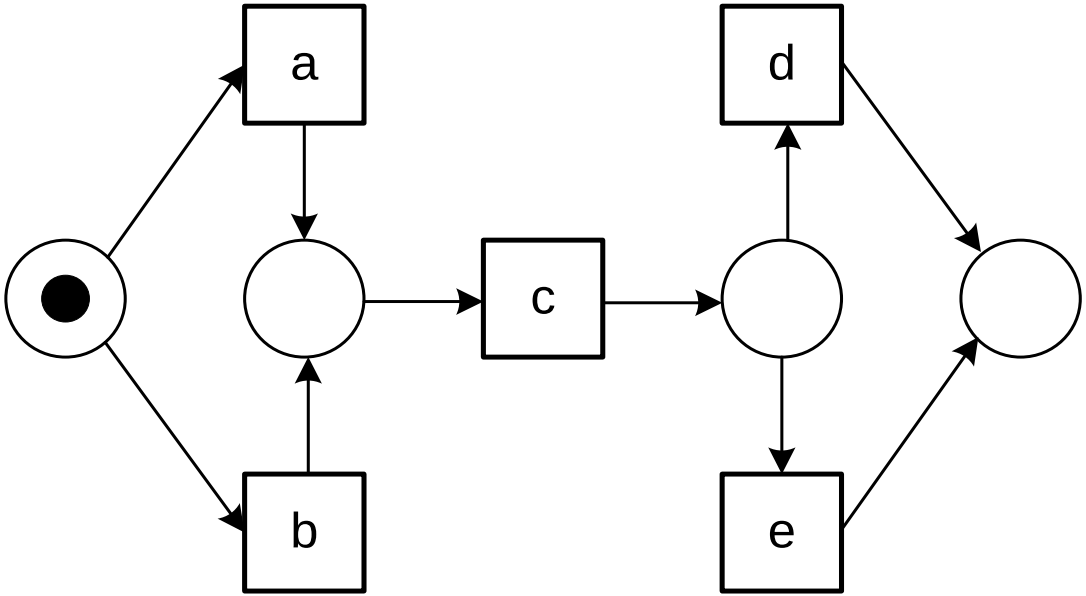




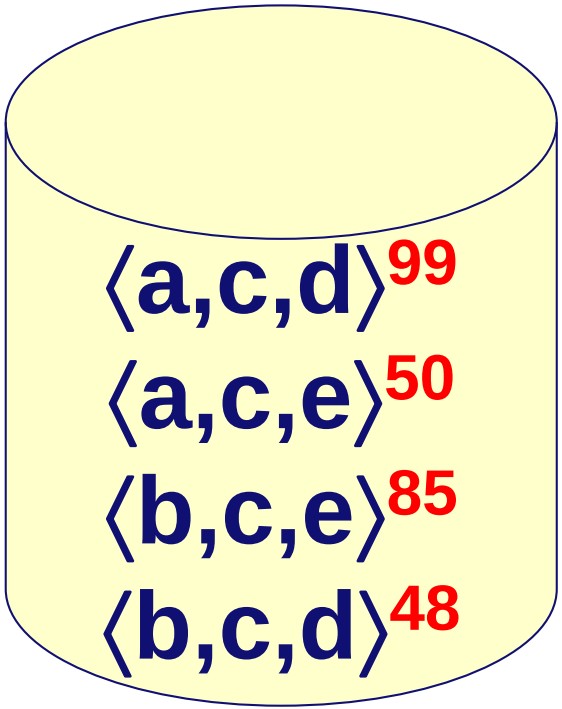
risk of overfitting on 5 example traces …

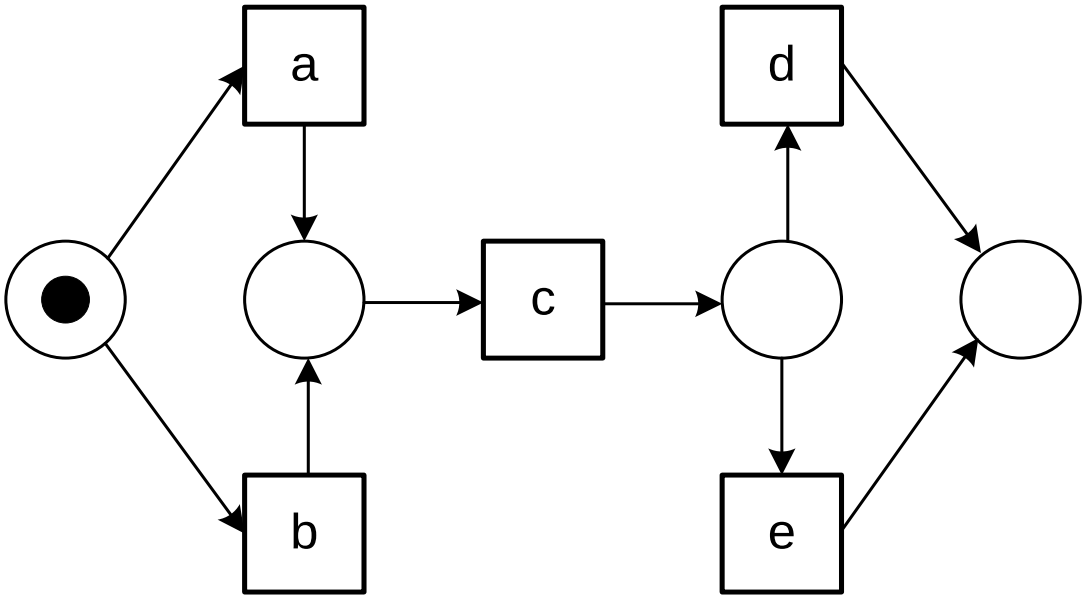
Generalization: good or bad?





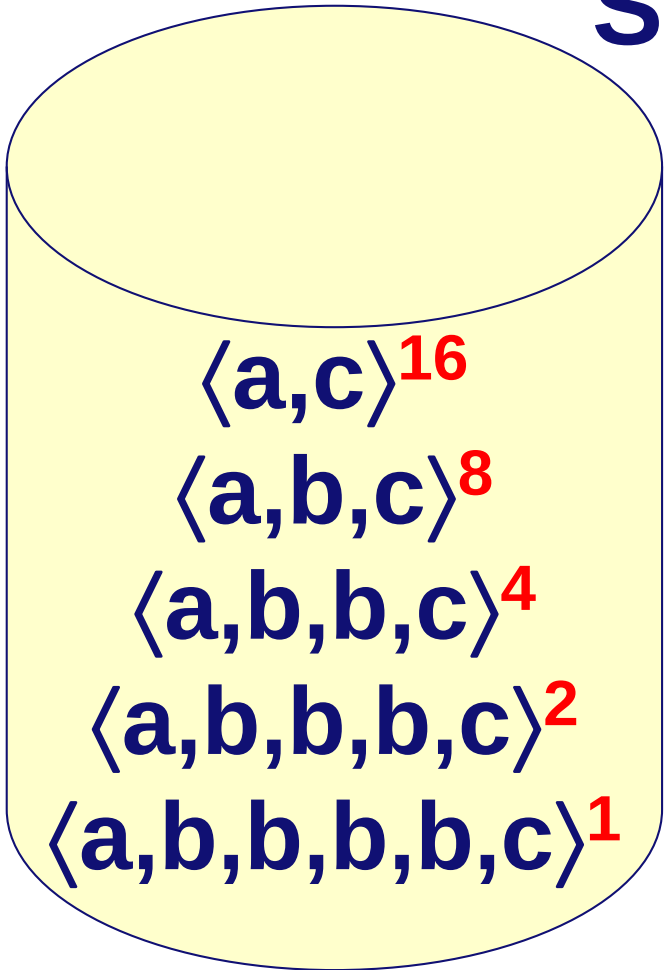
Generalization: good!

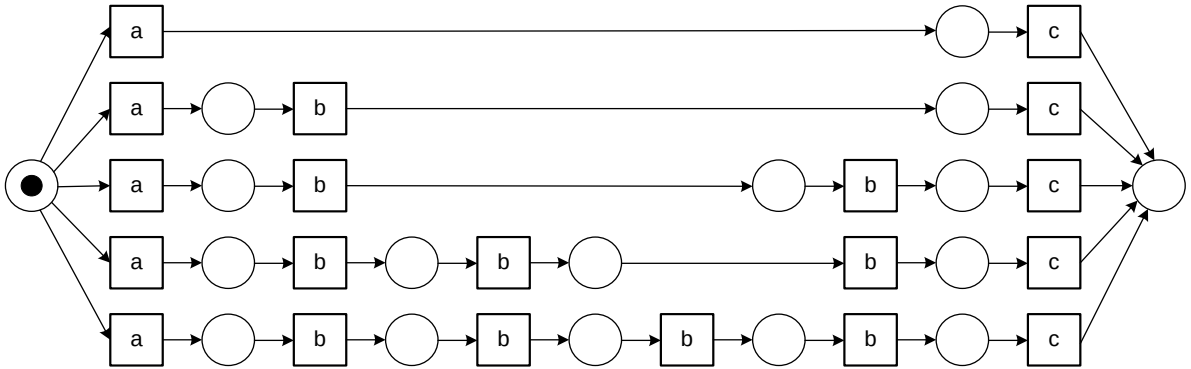




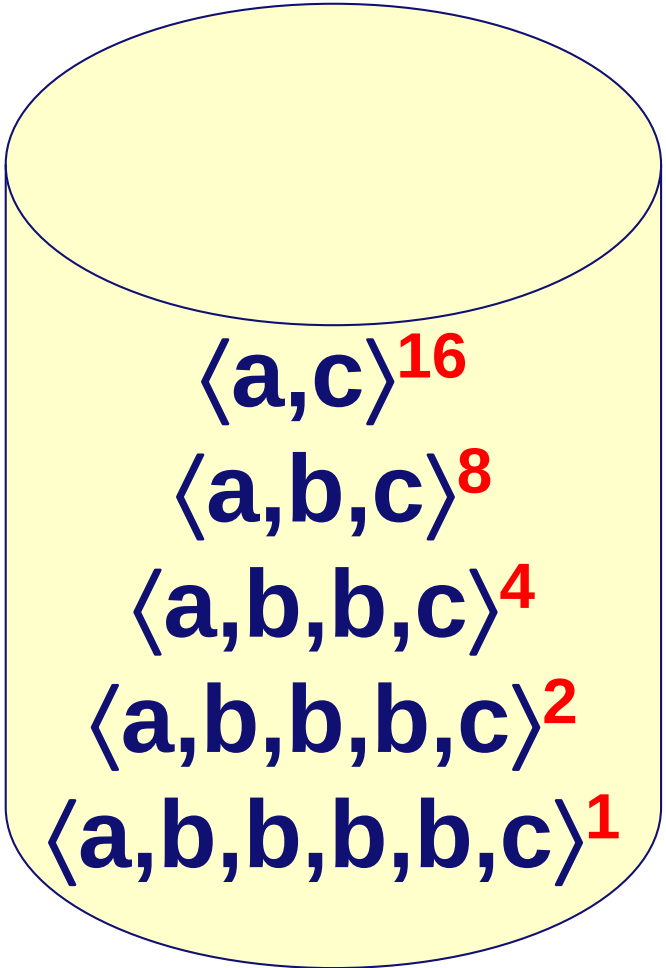
not overfitting…

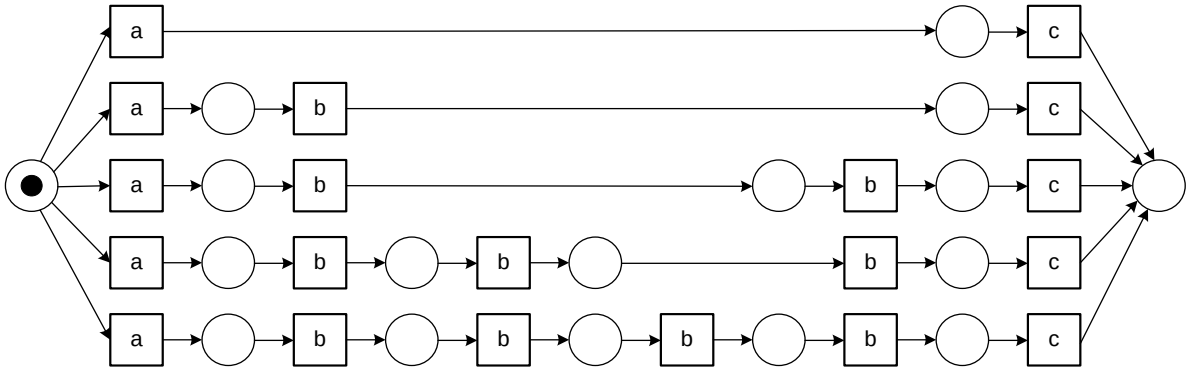
Simplicity: good or bad?





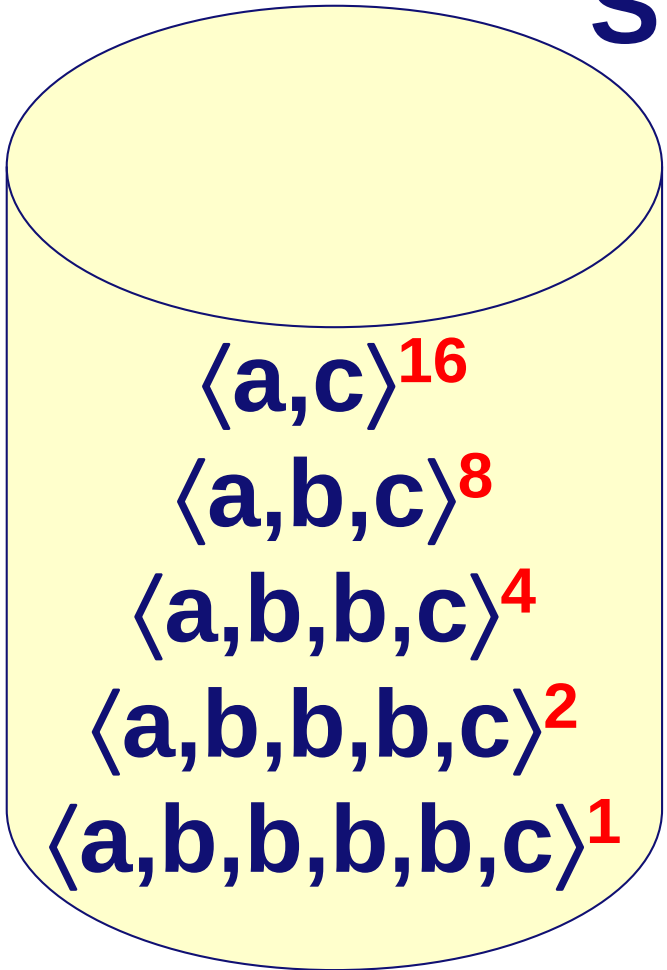
Simplicity: bad!

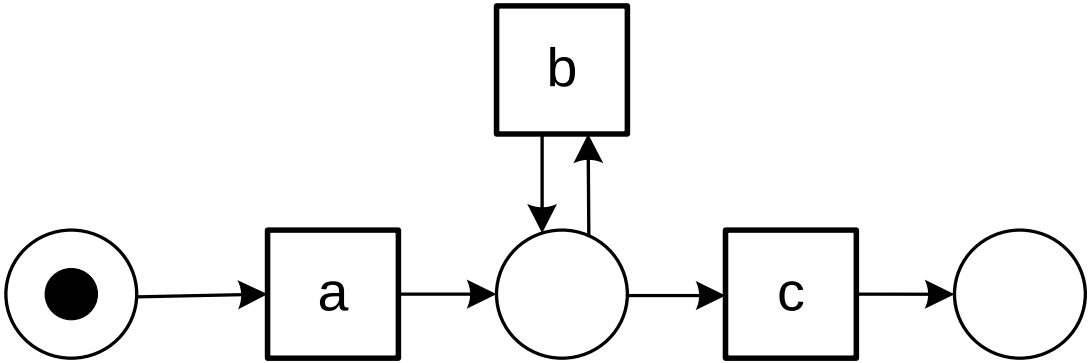




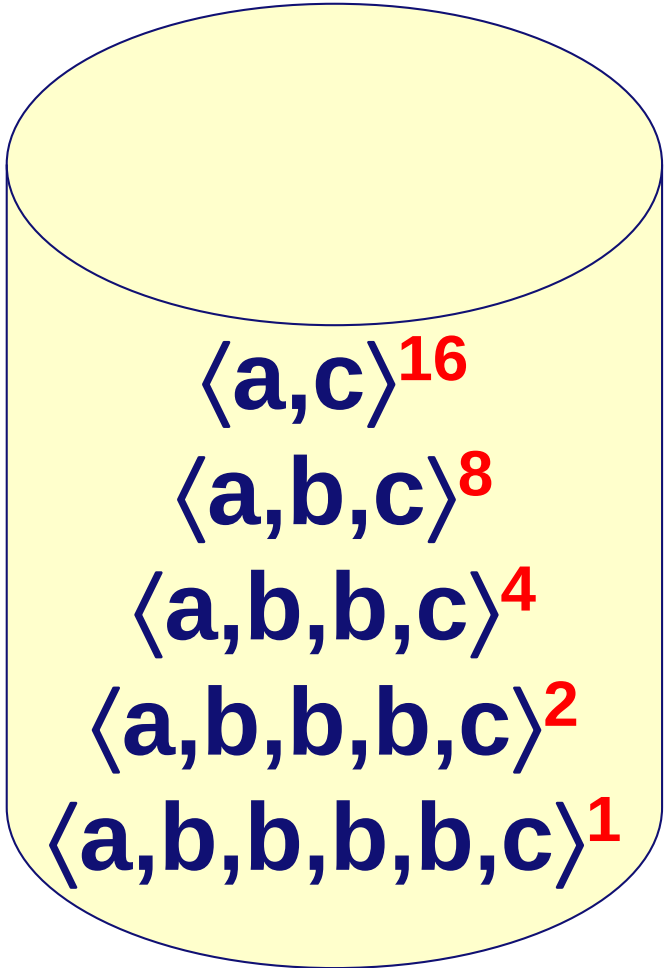
too complex/specific…

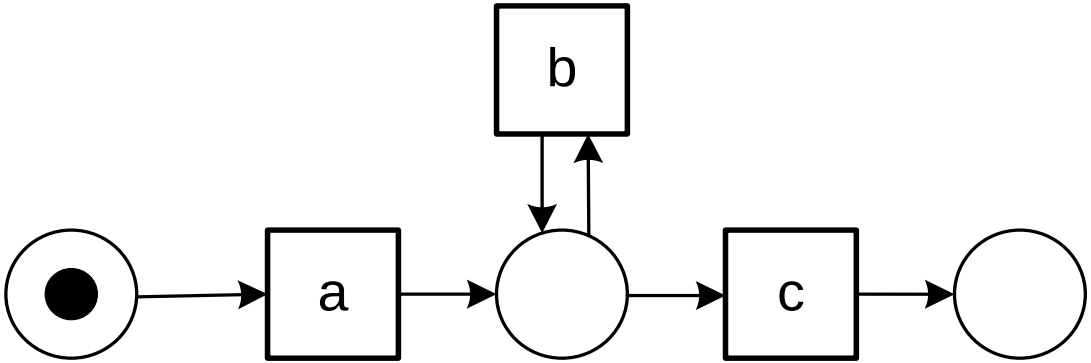
Simplicity: good or bad?

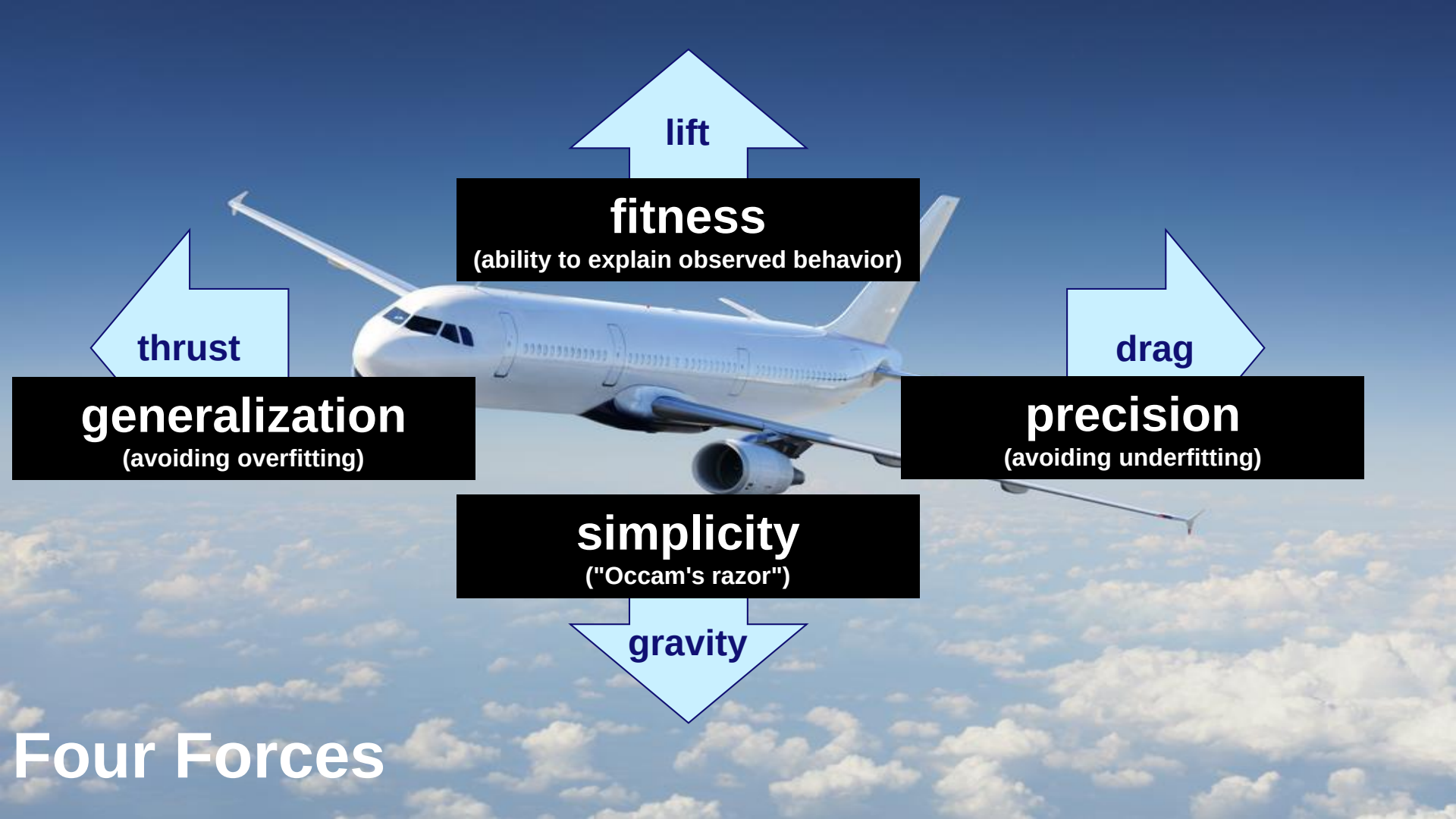




Simplicity: good!






lift
fitness
(ability to explain observed behavior)
thrust
drag
generalization
(avoiding overfitting)
precision
(avoiding underfitting)
simplicity
("Occam's razor")
gravity

Four Forces


THIS WAY
THAT WAY
ANOTHER WAY

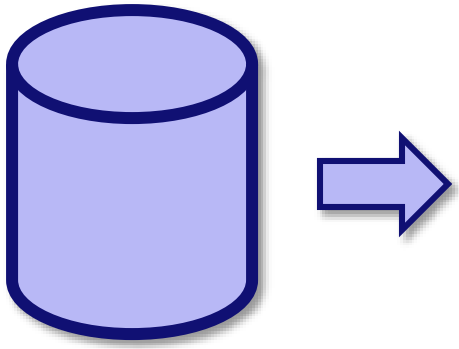

THIS WAY
THAT WAY
ANOTHER WAY

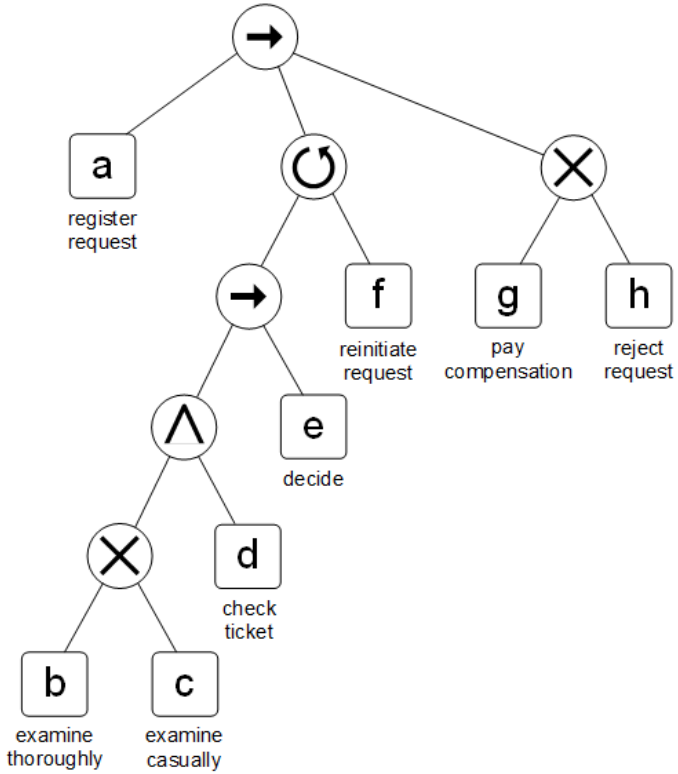
Characteristics

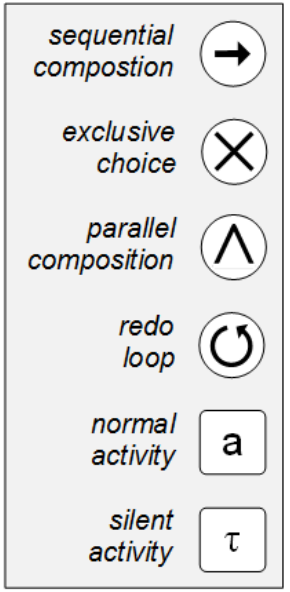
1. Representational bias (class of target models)
2. Ability to deal with noise/infrequent/incomplete behavior
3. Formal guarantees (in the limit, rediscoverability)
4. Scalability
5. Approach used:
   - Direct algorithmic (alpha-family, heuristic/fuzzy miner)
   - Region-based (language/state-based)
   - Generic/evolutionary
   - Inductive

Inductive mining

Process trees (to ensure soundness)

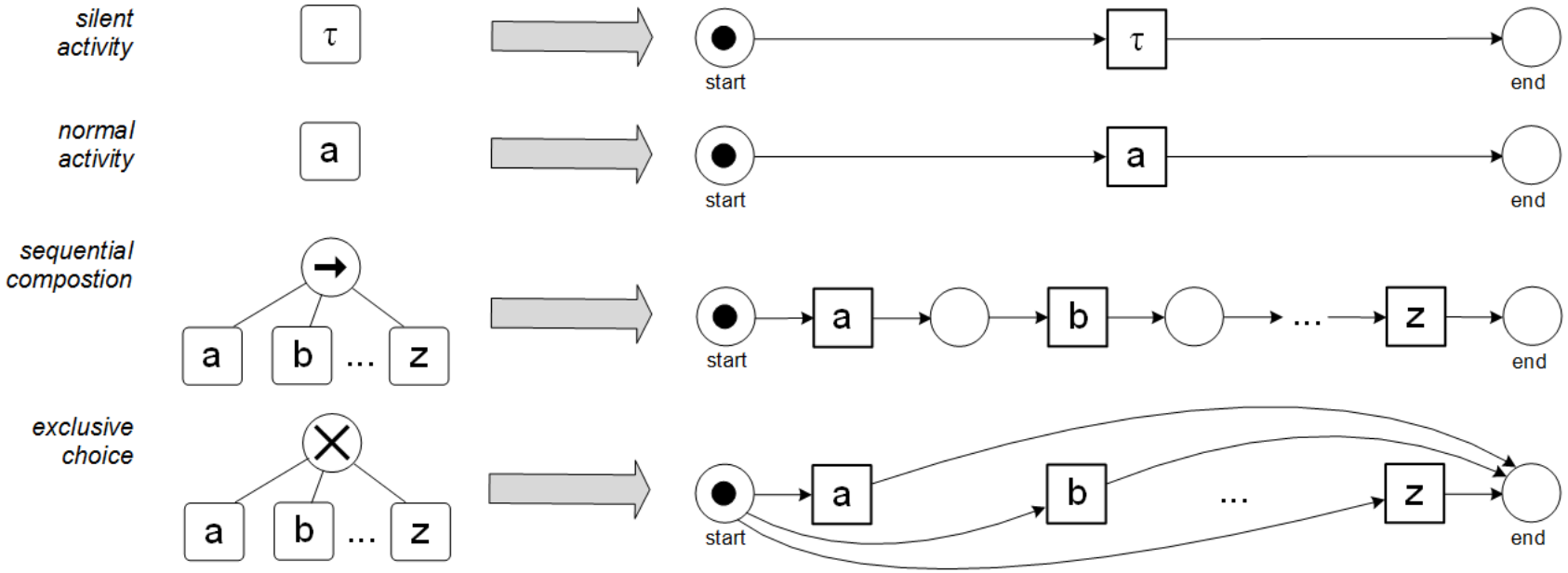






TU/e

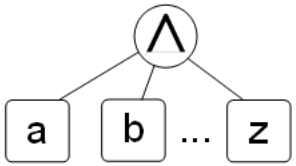
Process trees (semantics)

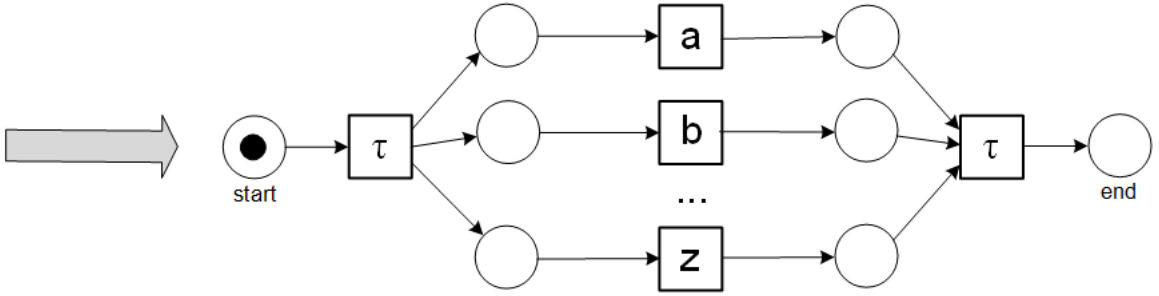


TU/e

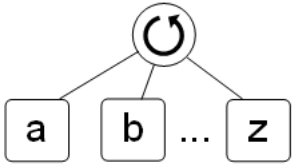
Process trees (semantics)

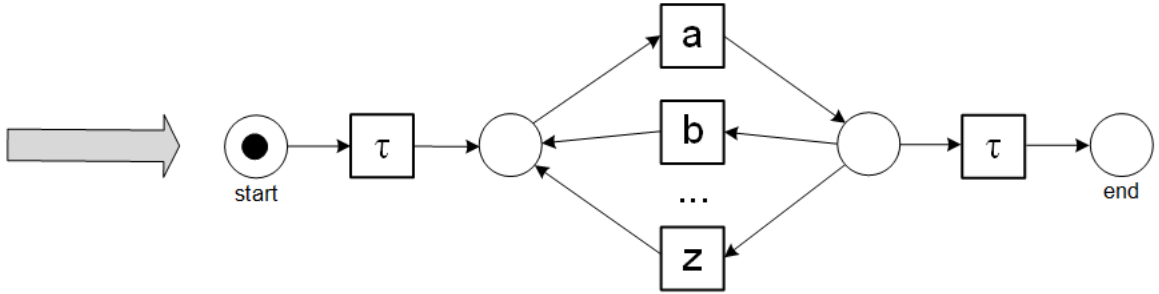
parallel composition





redo loop





TU/e

Split event logs based on activity labels

abdef
acdef
adbef
adcef
abdeg
acdeg
adbeg
adceg

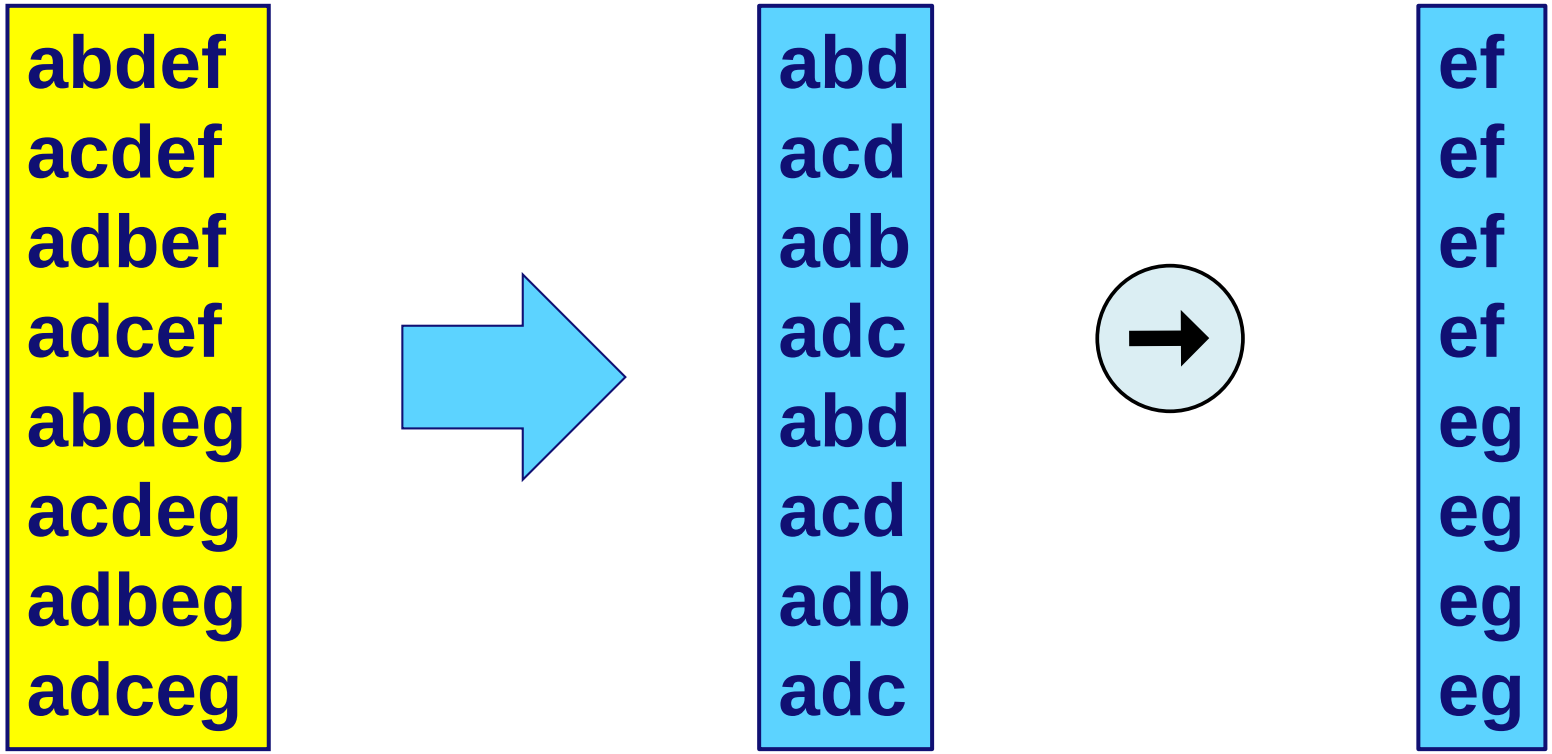
TU/e

Split {a,b,c,d,e,f,g,h} into {a,b,c,d} and {e,f,g} using sequence decomposition

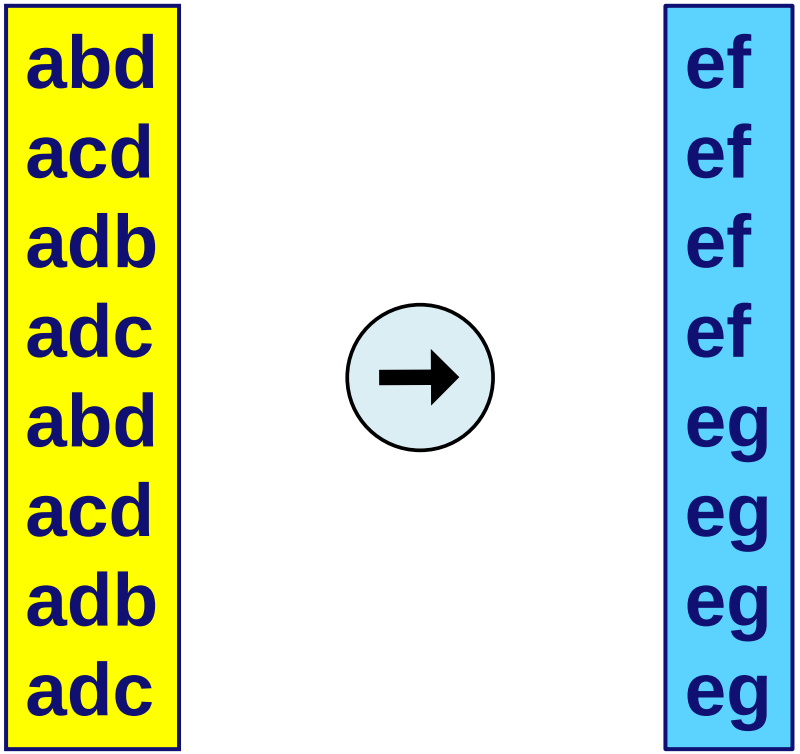
abdef
acdef
adbef
adcef
abdeg
acdeg
adbeg
adceg

TU/e

Result

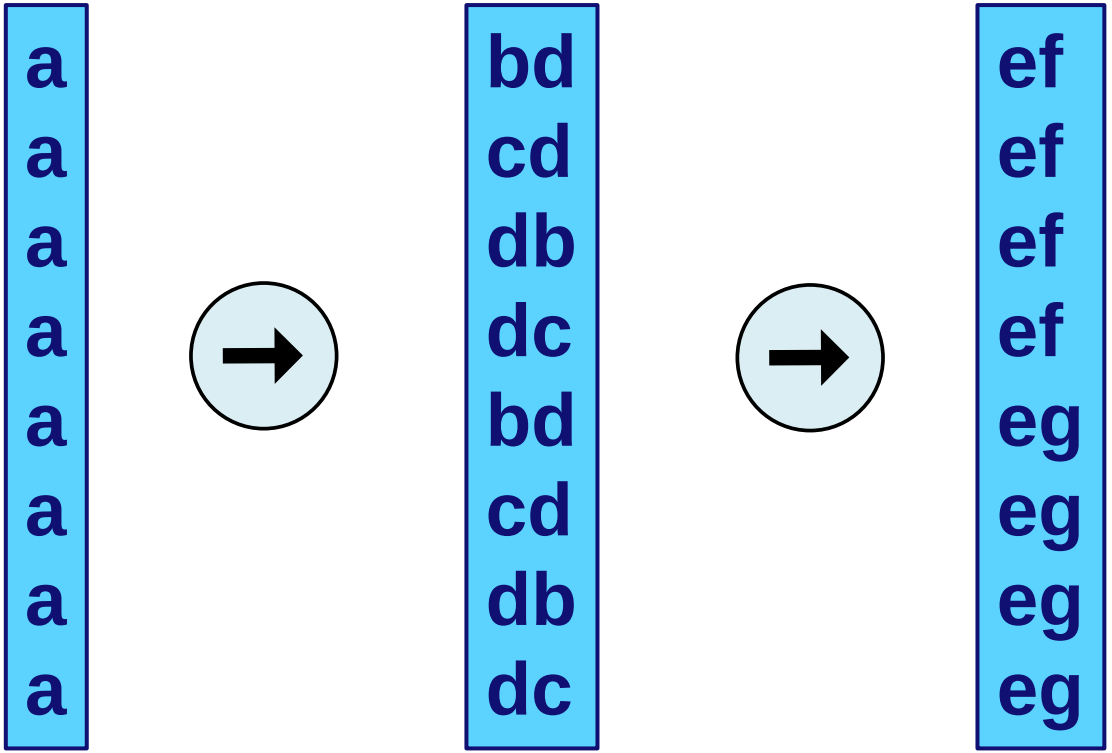


Split {a,b,c,d} into {a} and {b,c,d} using sequence decomposition

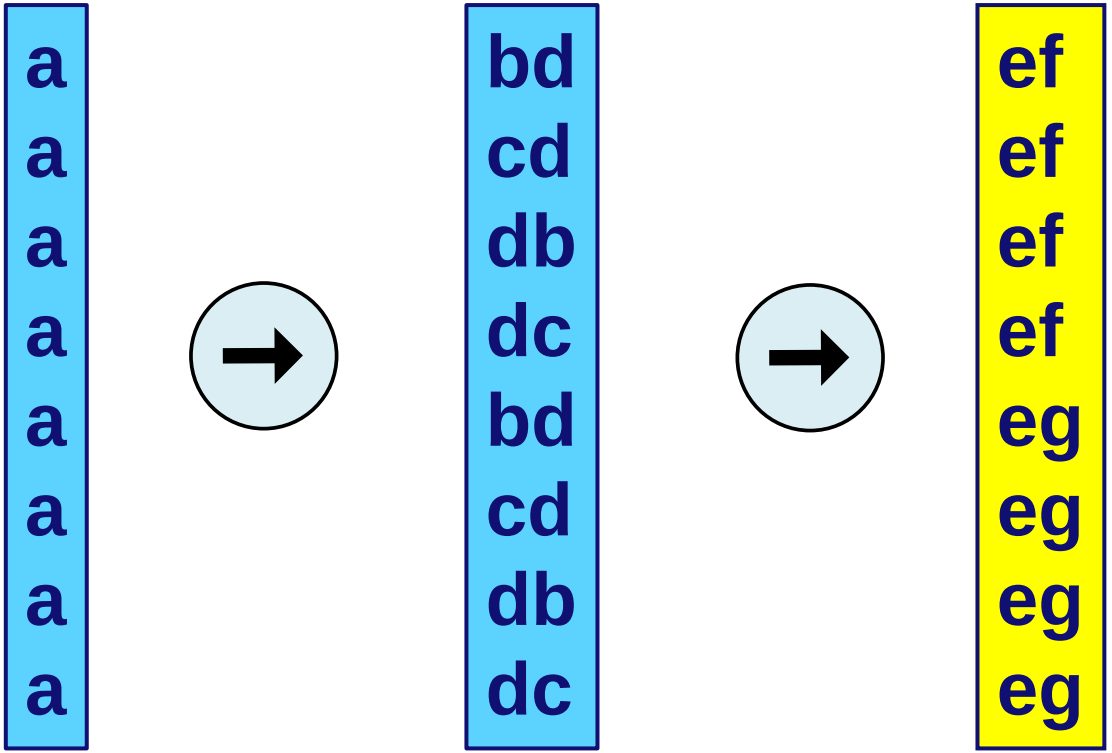


TU/e

Result

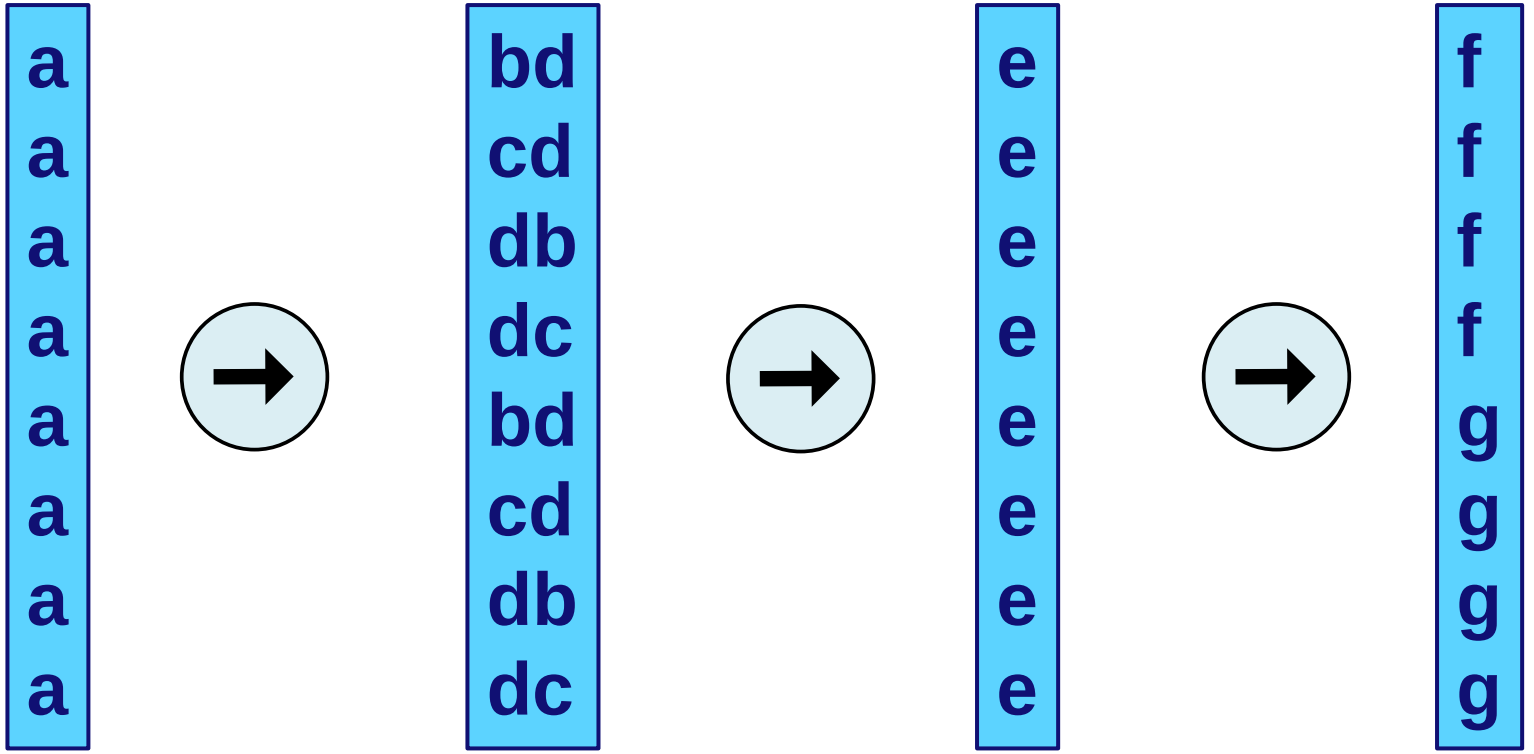


Split {e,f,g} into {e} and {f,g} using sequence decomposition



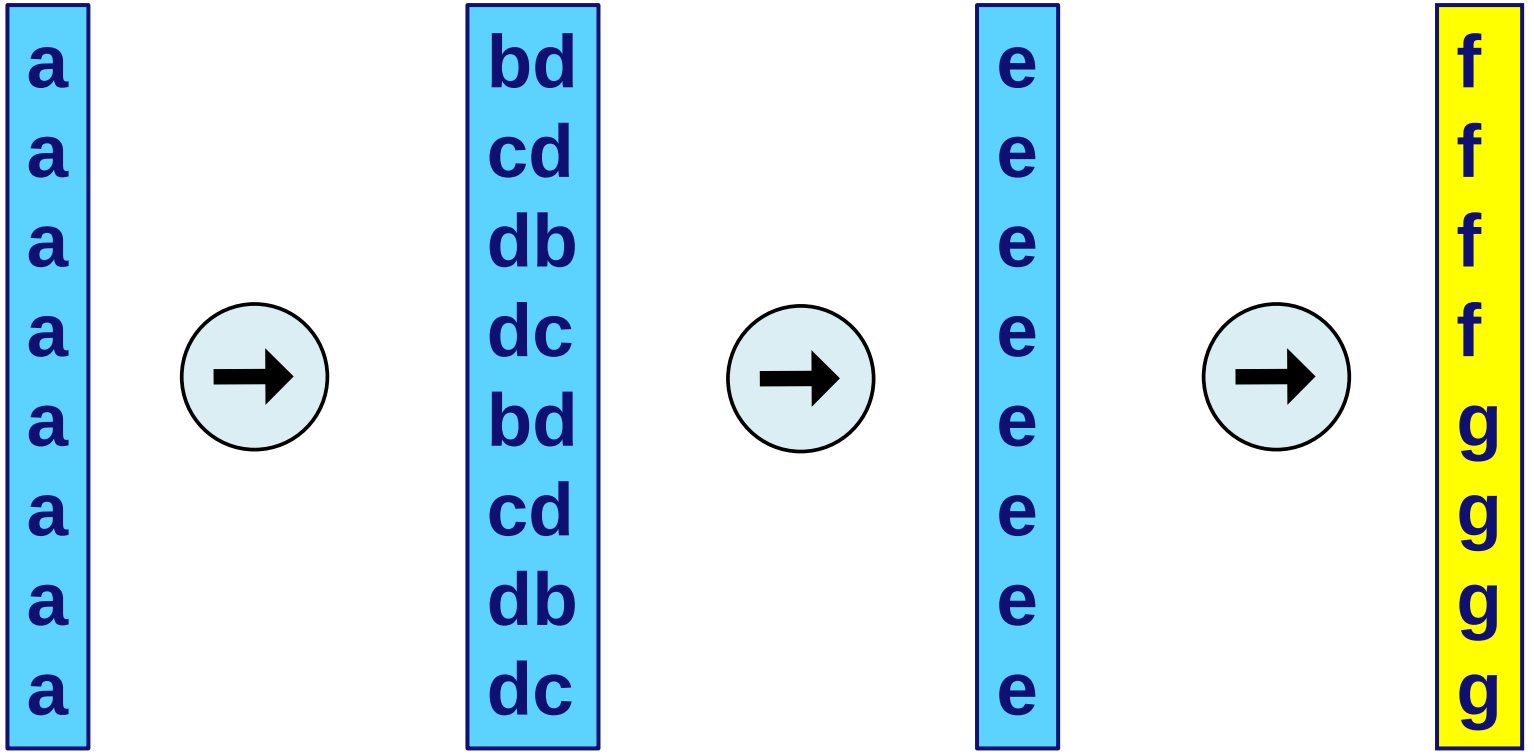
TU/e

Result



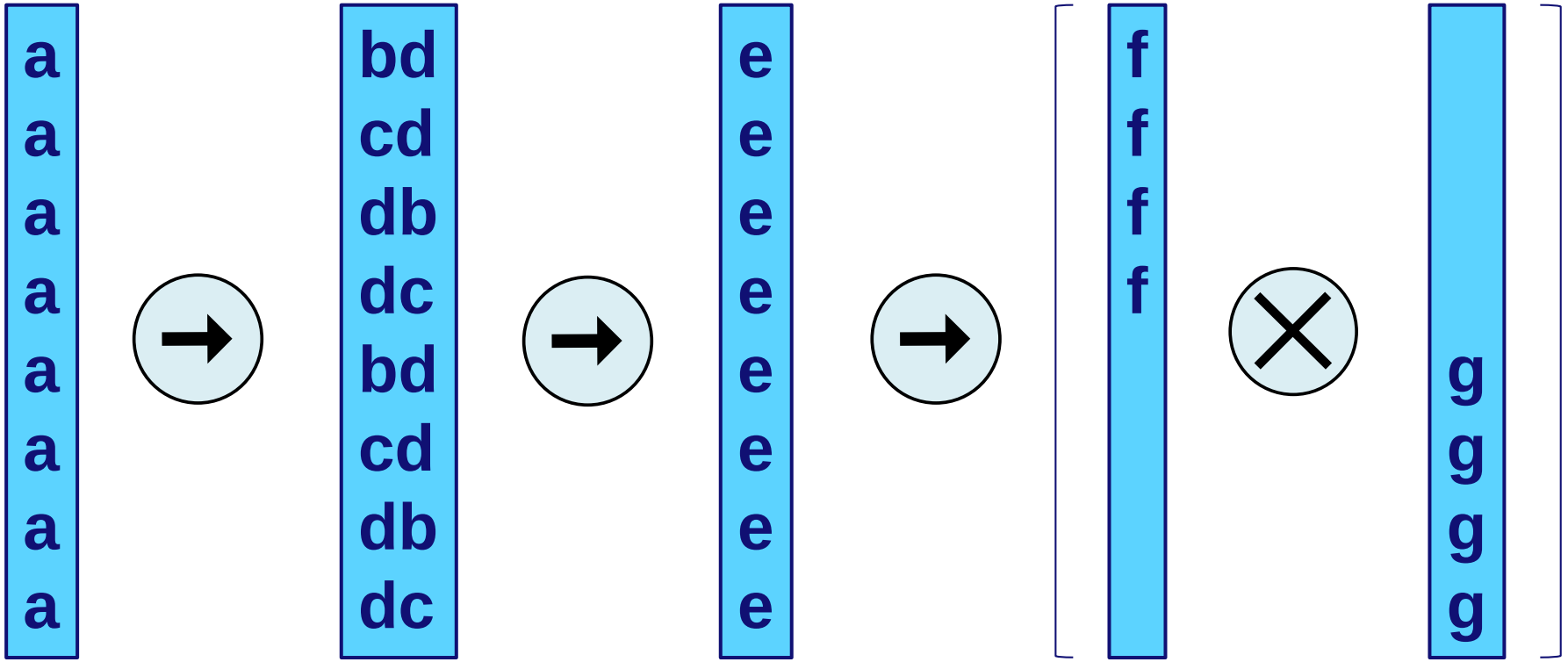
TU/e

Split {f,g} into {f} and {g} using XOR decomposition



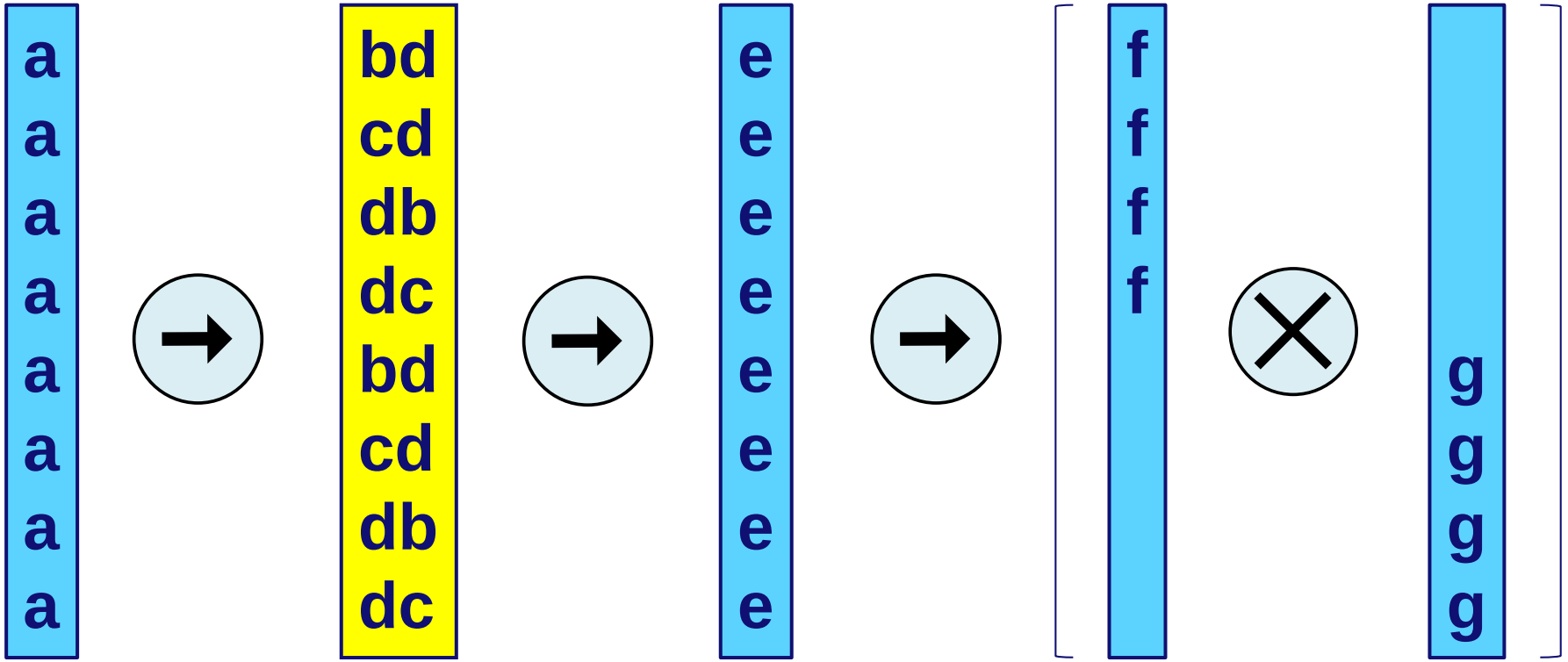
TU/e

Result



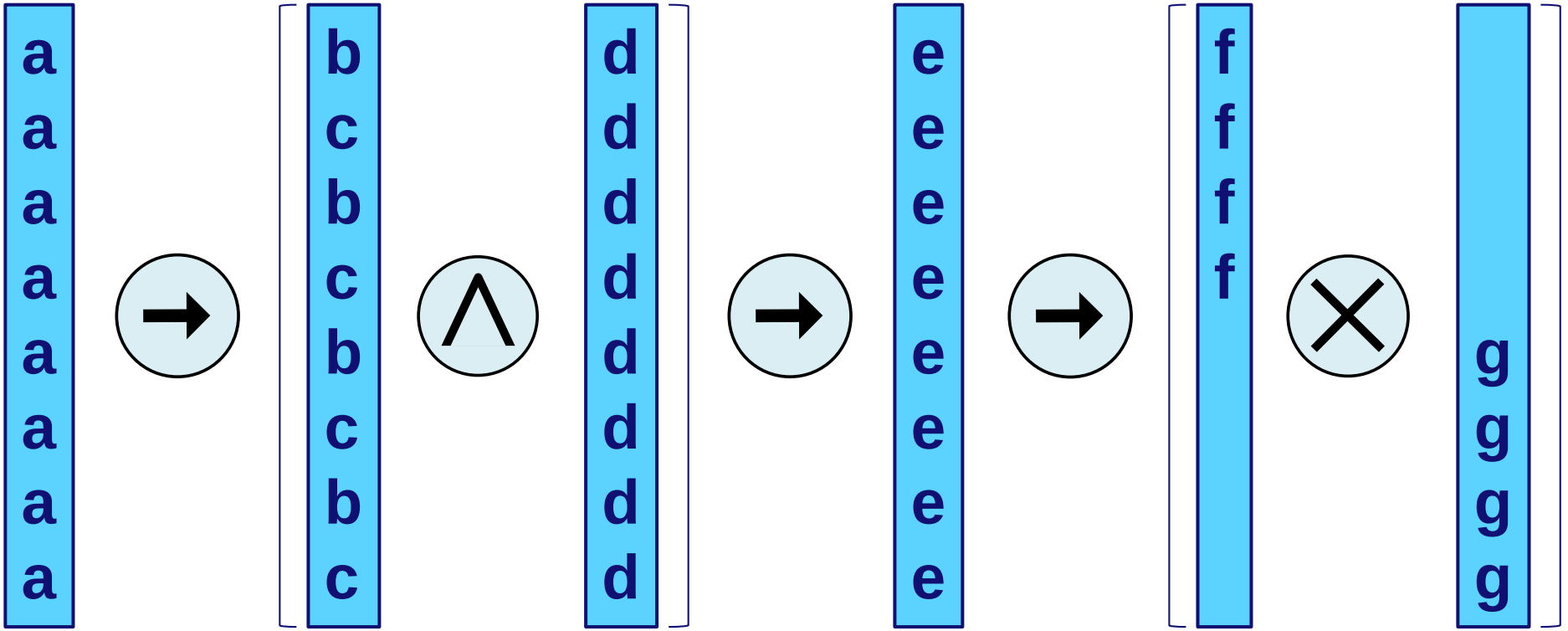
TU/e

Split {b,c,d} into {b,c} and {d} using AND decomposition



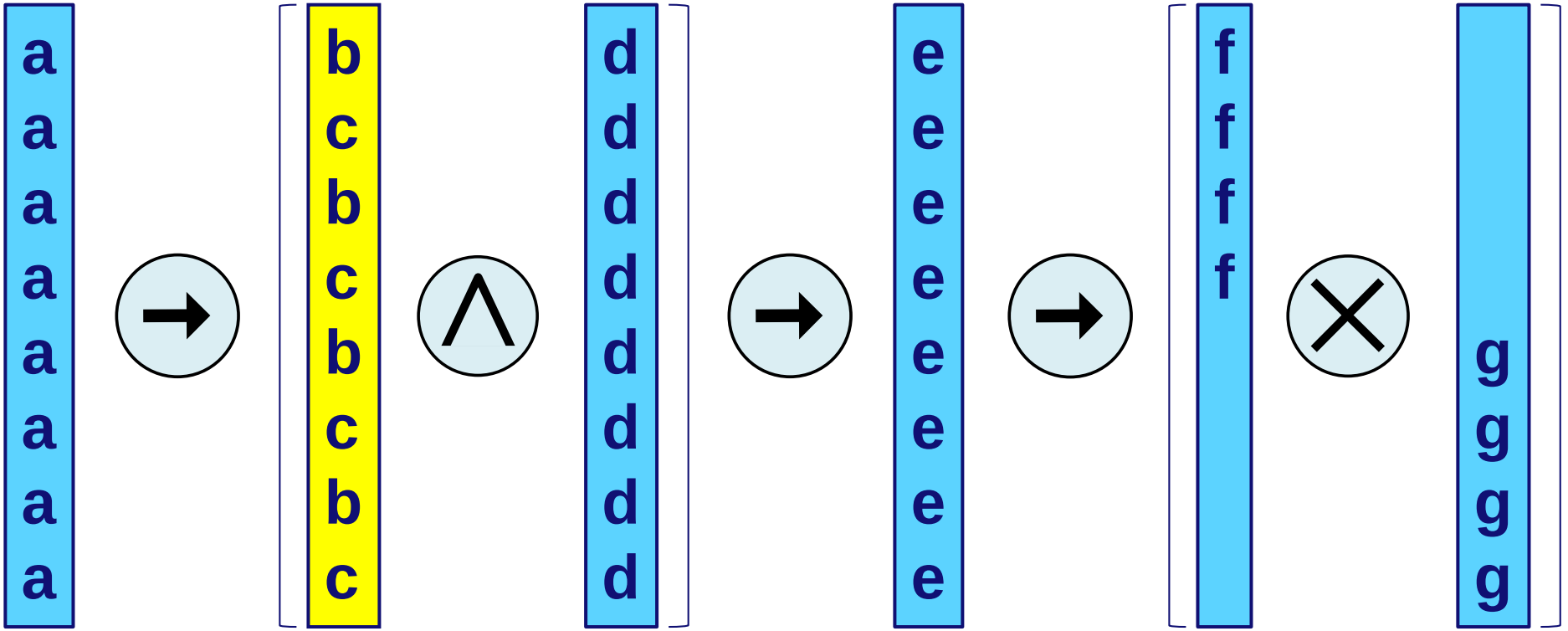
TU/e

Result



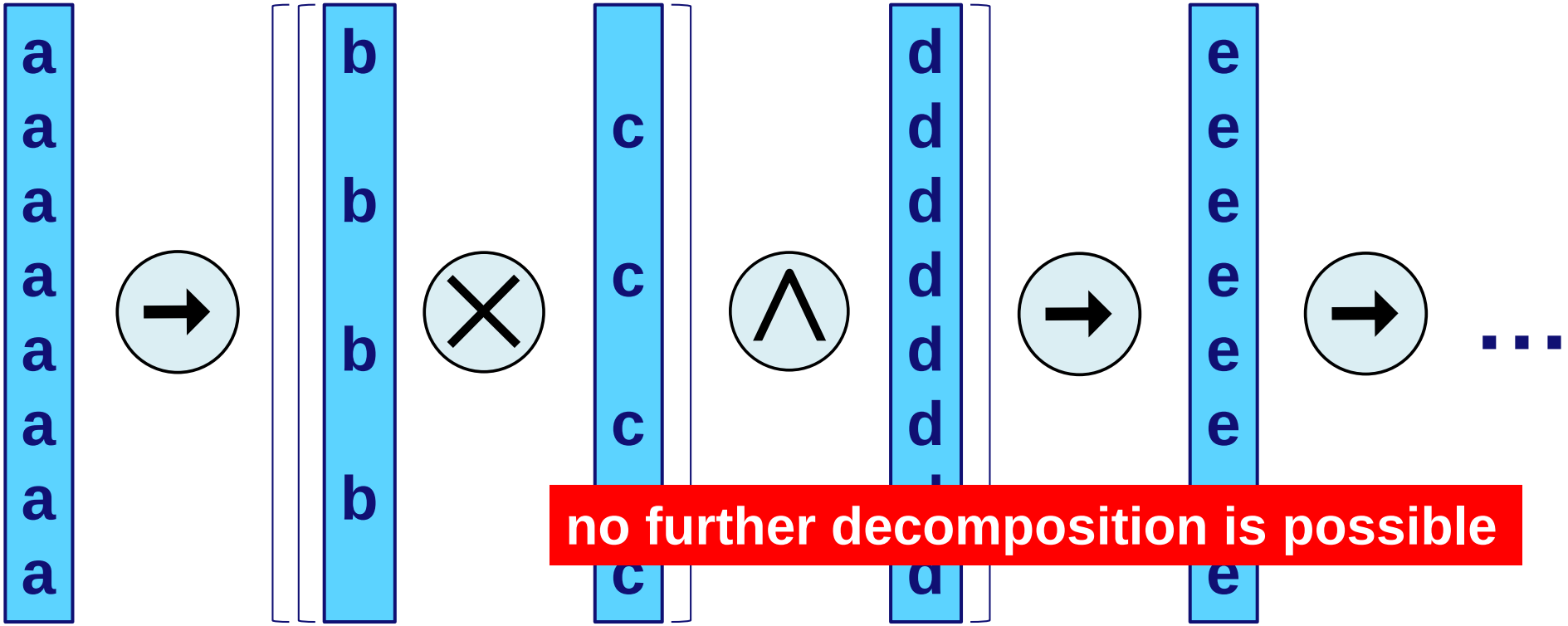
TU/e

Split {b,c} into {b} and {c} using XOR decomposition



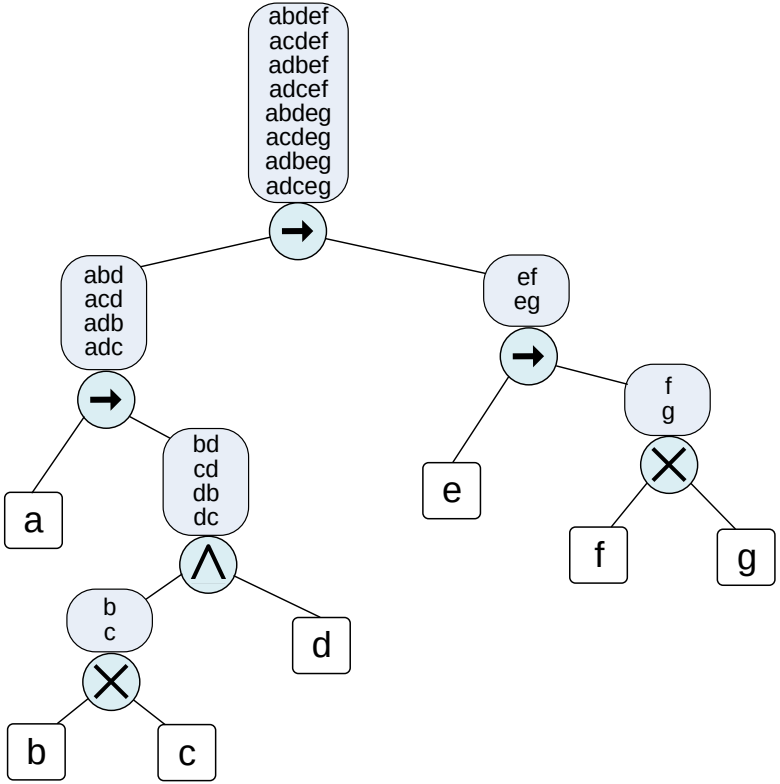
TU/e

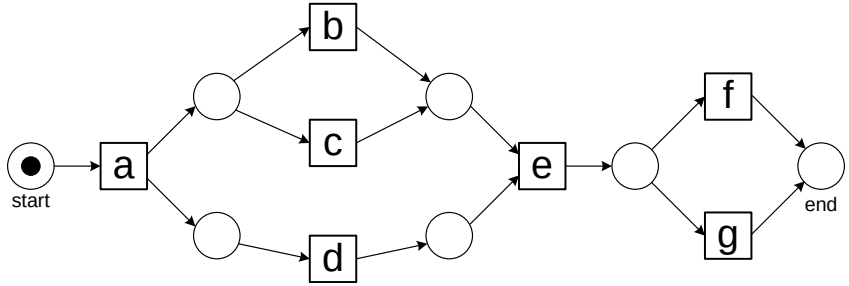
Result

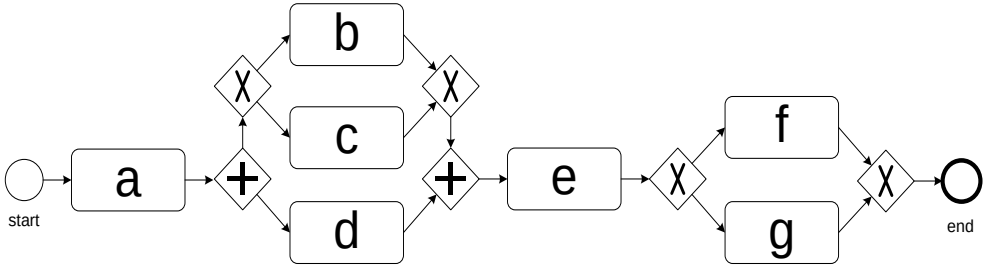


TU/e

Process tree





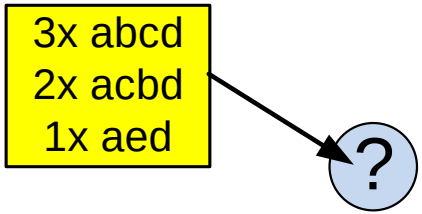


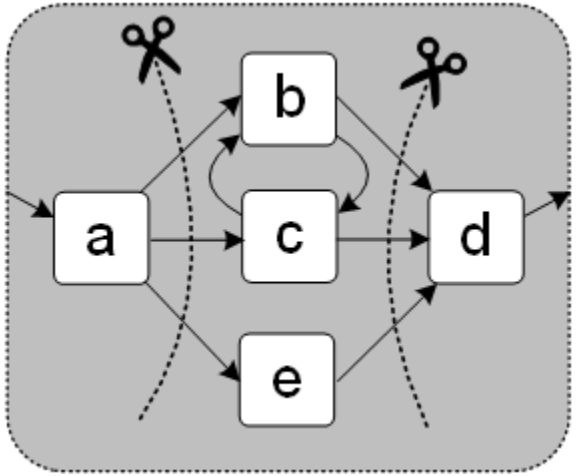
An example log (6 traces, 23 events)

3x abcd
2x acbd
1x aed

TU/e

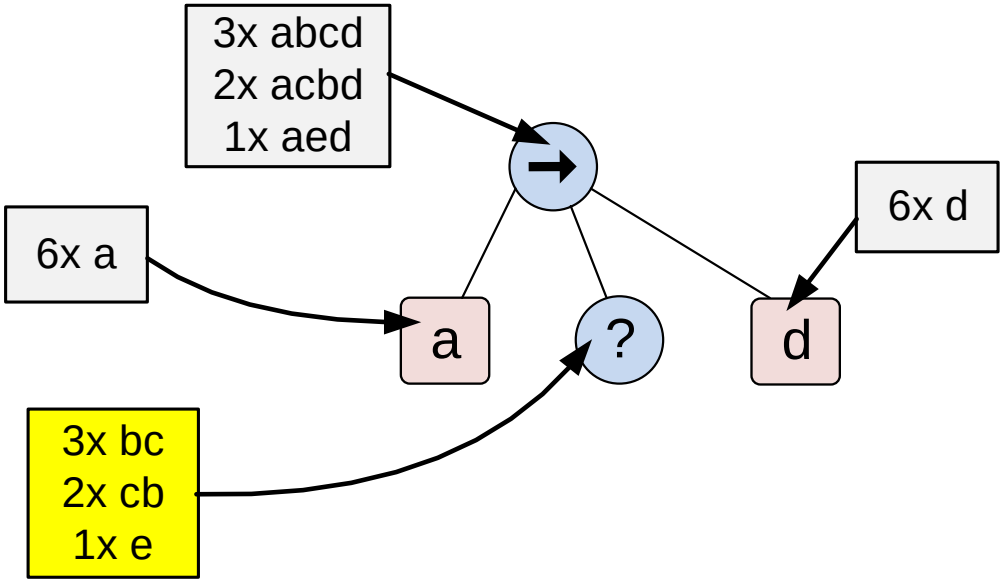
How to split this event log?

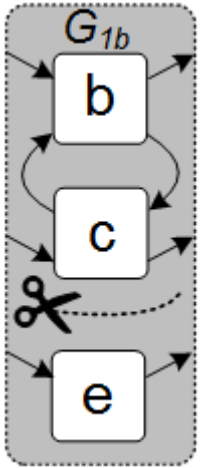




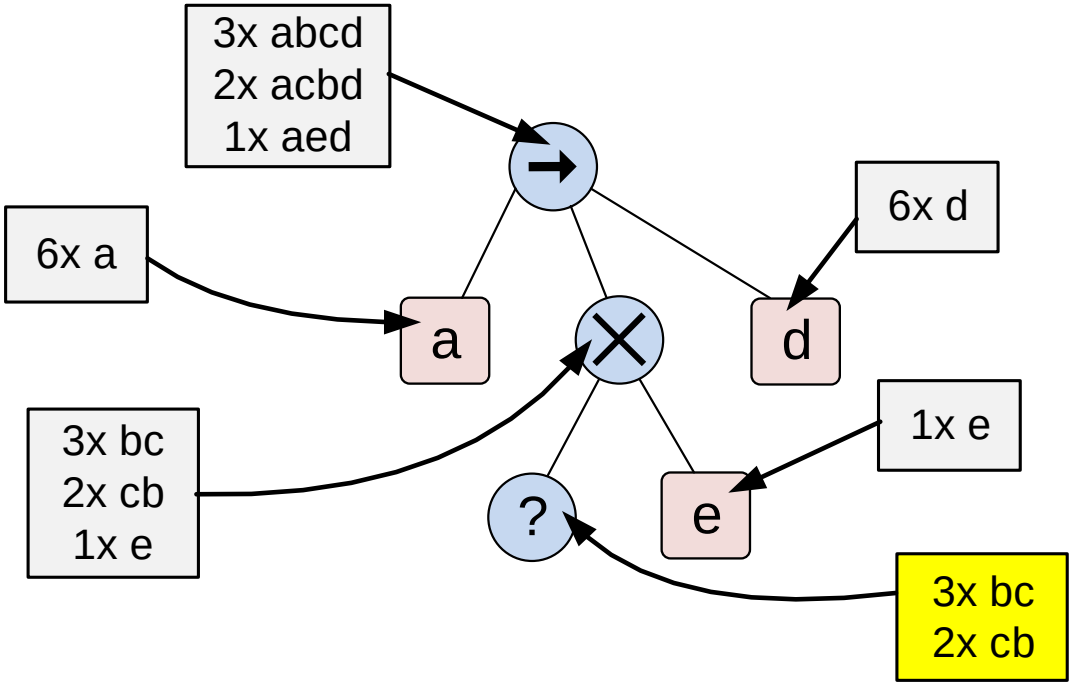
TU/e

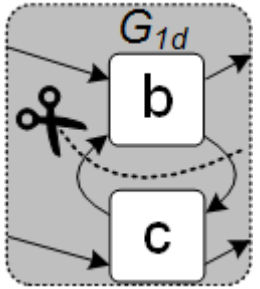
How to split this event log?





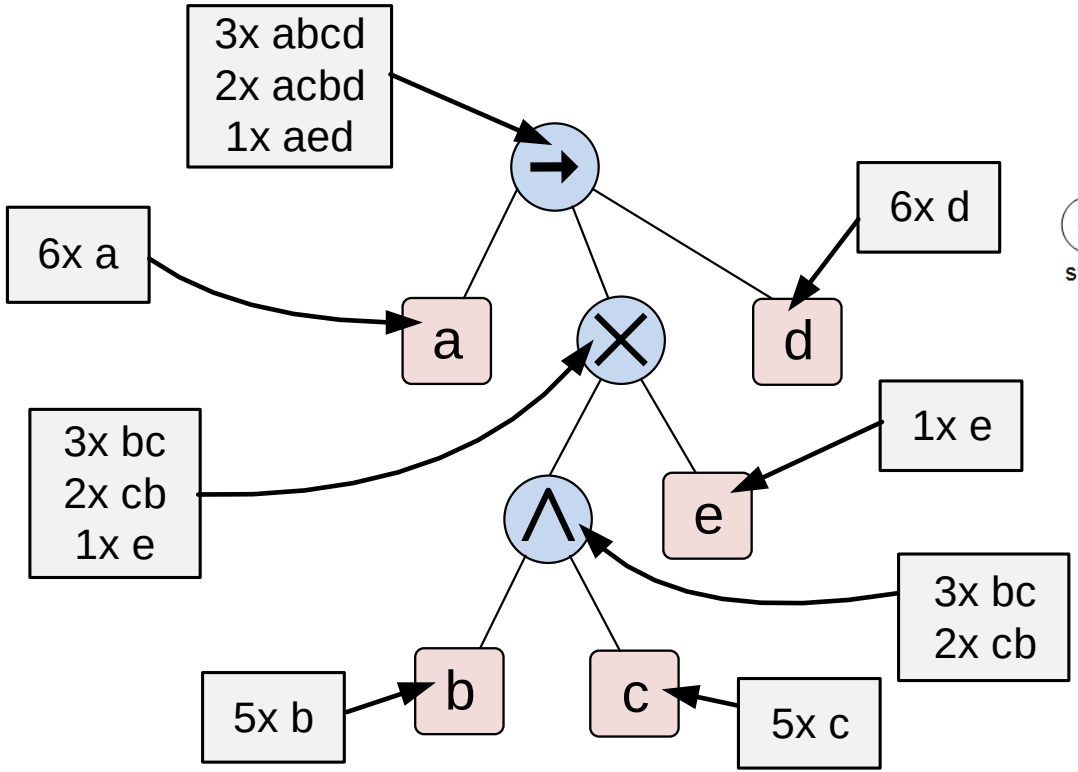
How to split this event log?

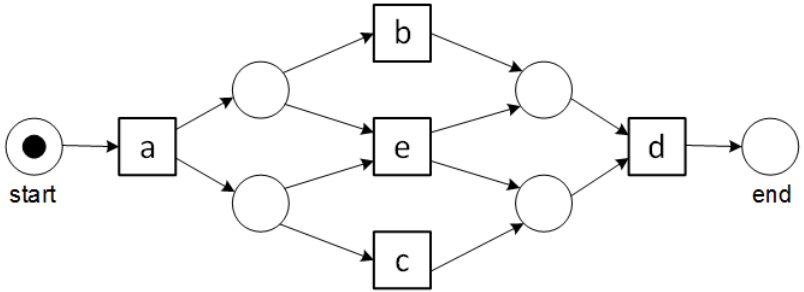




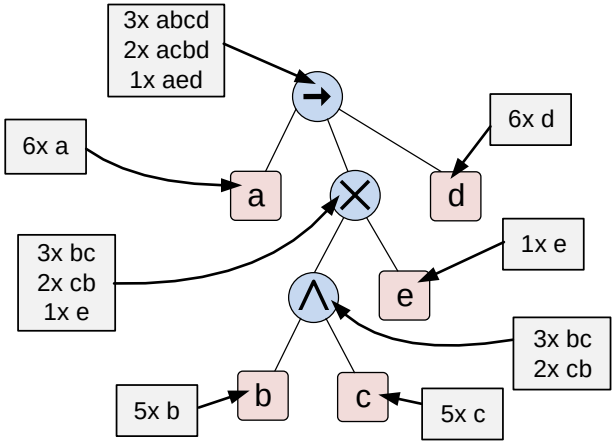
TU/e

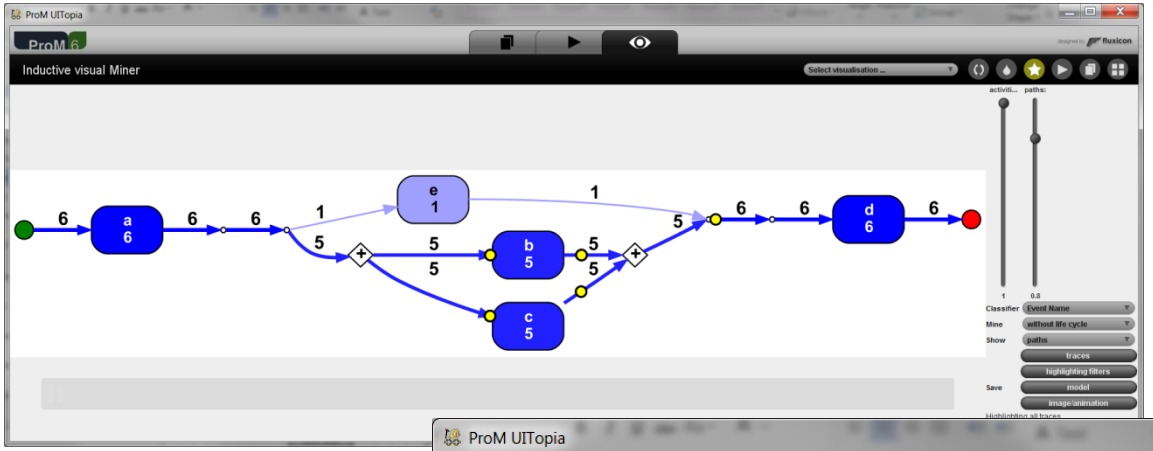
Final result

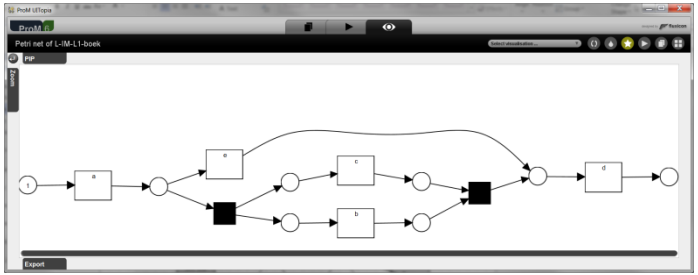


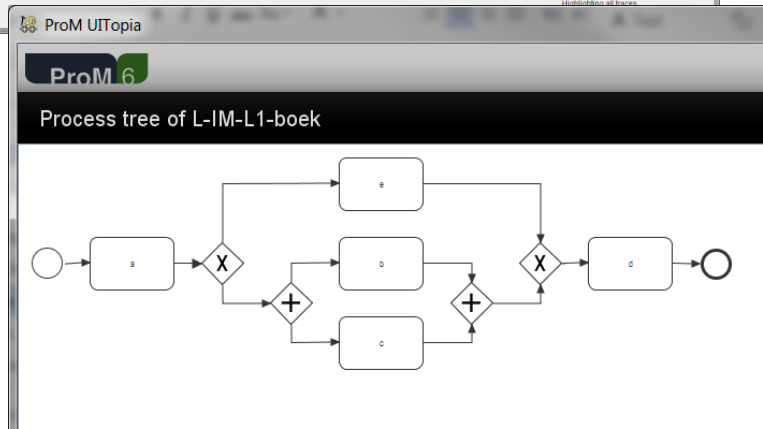


In ProM









Example log with loops (13 traces, 80 events)

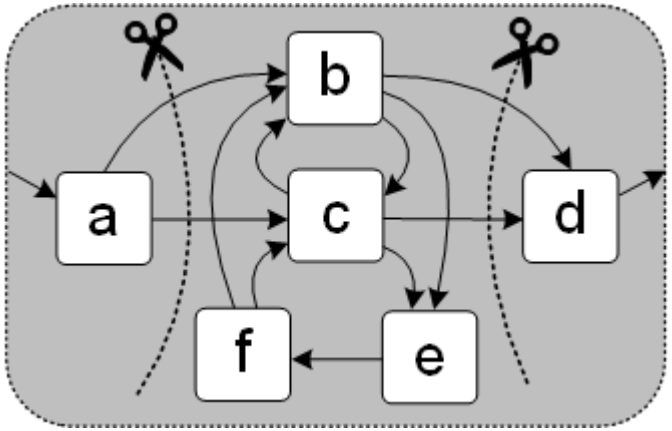
3x abcd
4x acbd
2x abcefbcd
2x acbefbcd
1x abcefcbd
1x acbefbcefcbd

TU/e

How to split this event log?

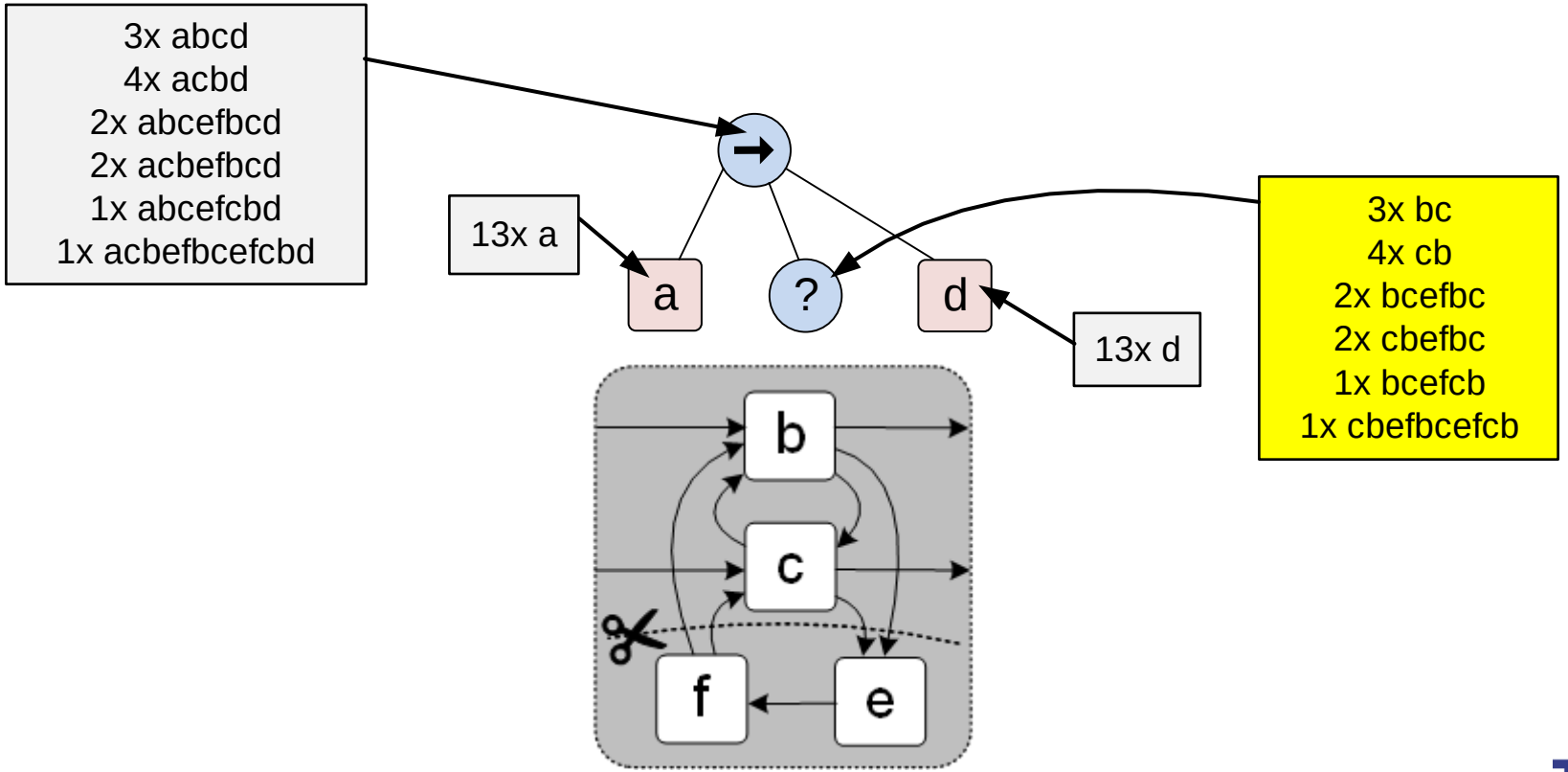
3x abcd
4x acbd
2x abcefbcd
2x acbefbcd
1x abcefcbd
1x acbefbcefcbd

?



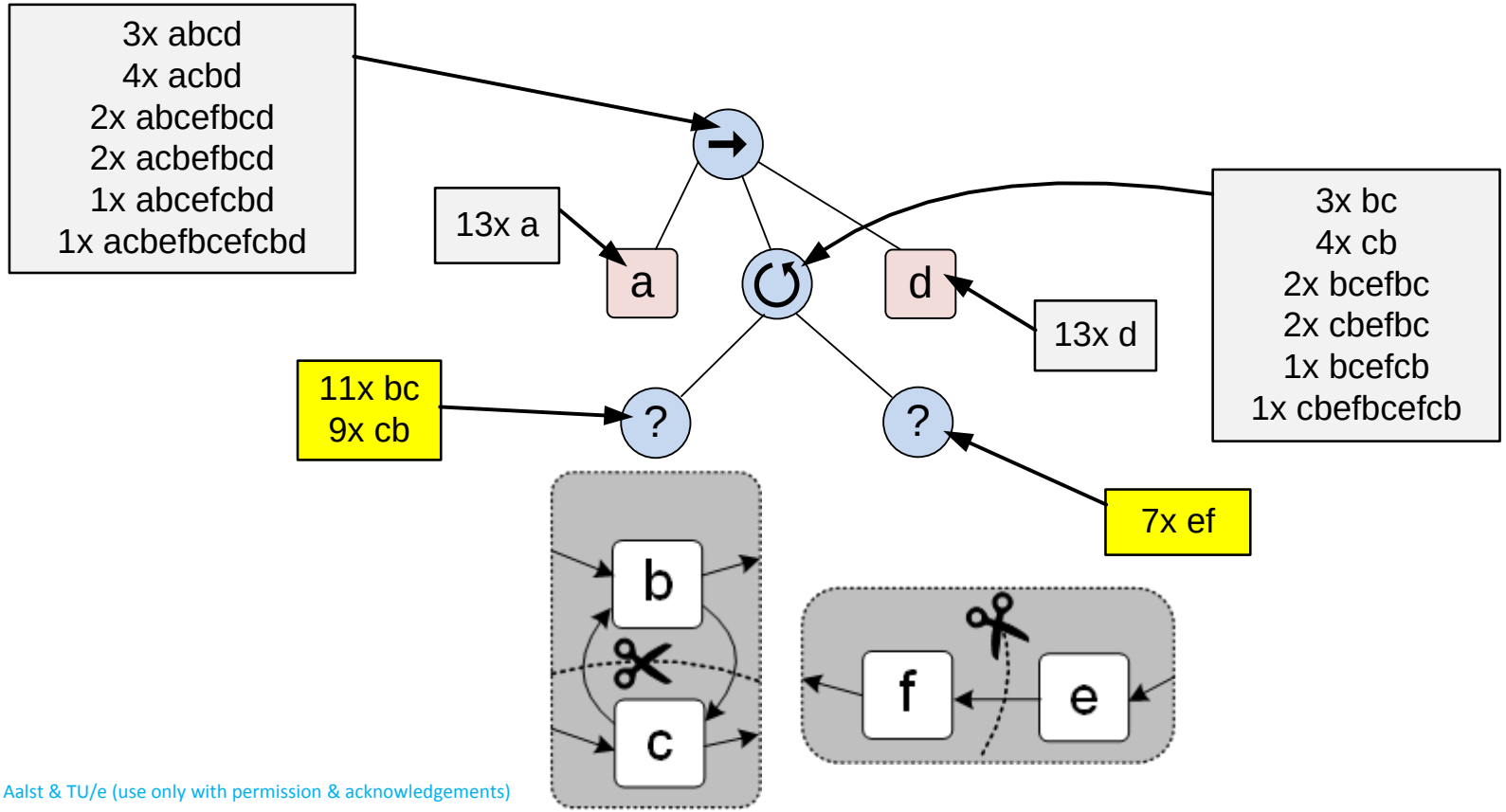
TU/e

How to split this event log?



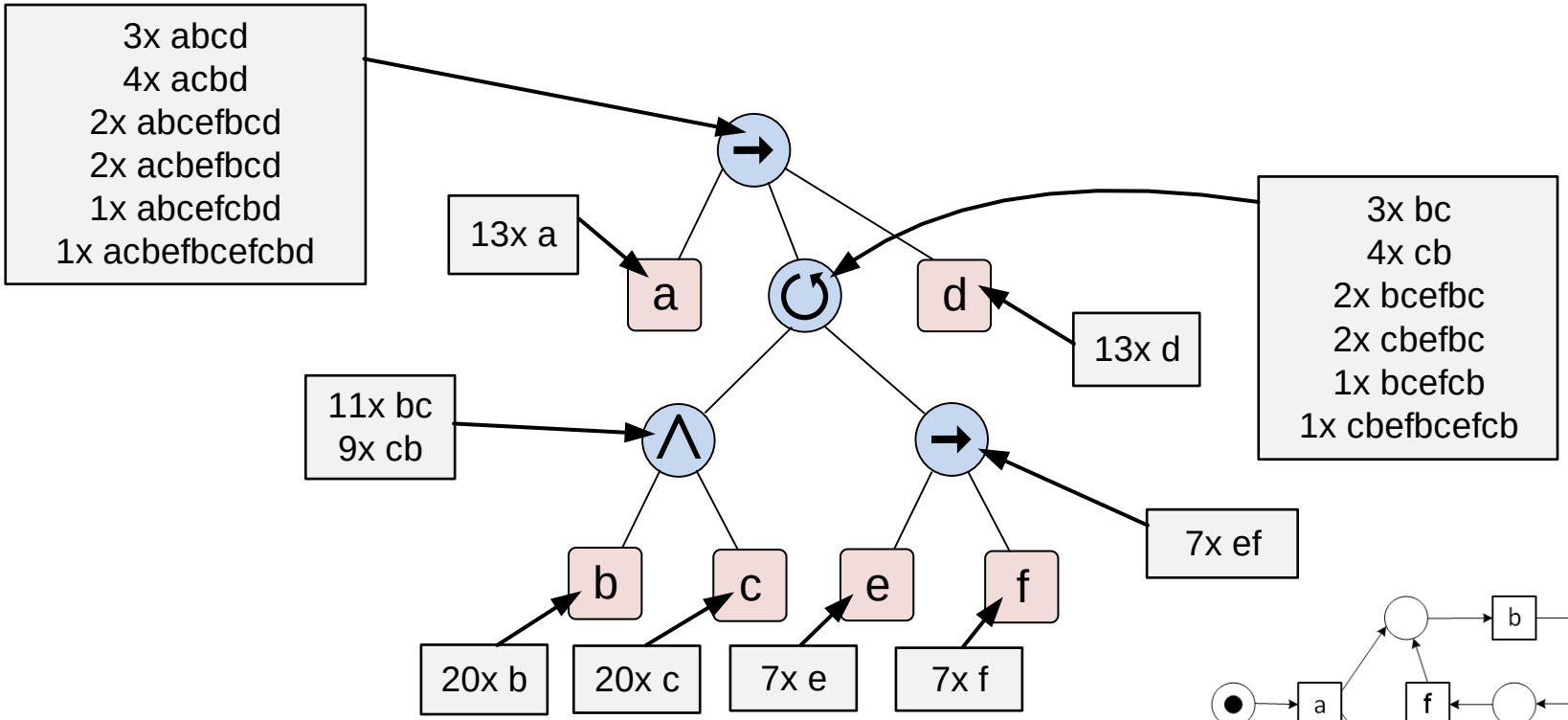
TU/e

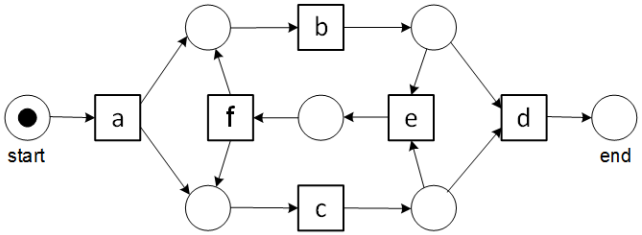
How to split this event log?



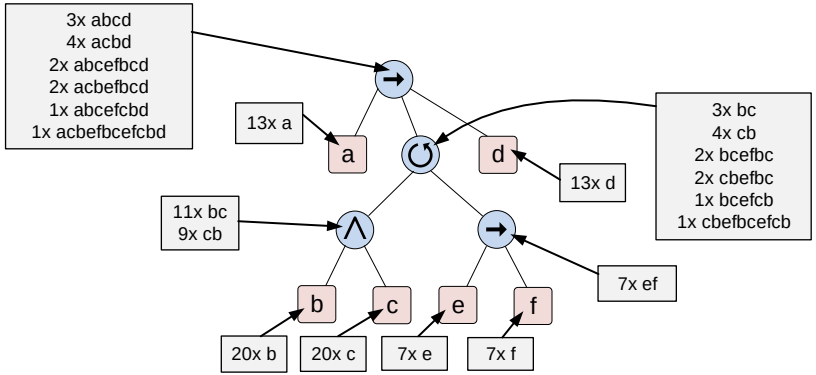
TU/e

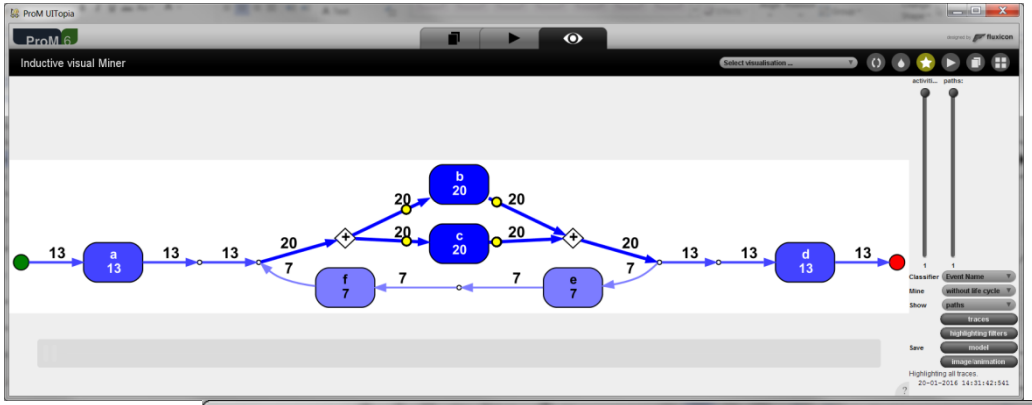
Result

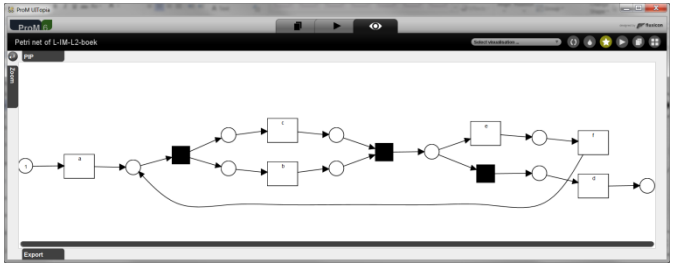


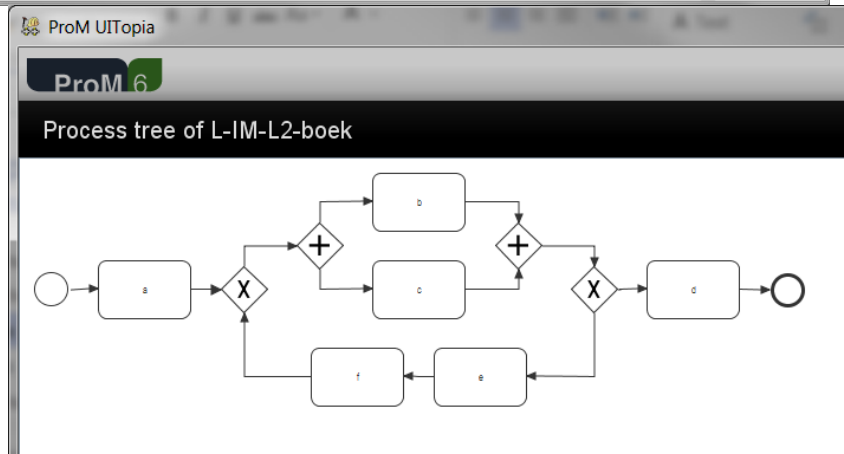


In ProM

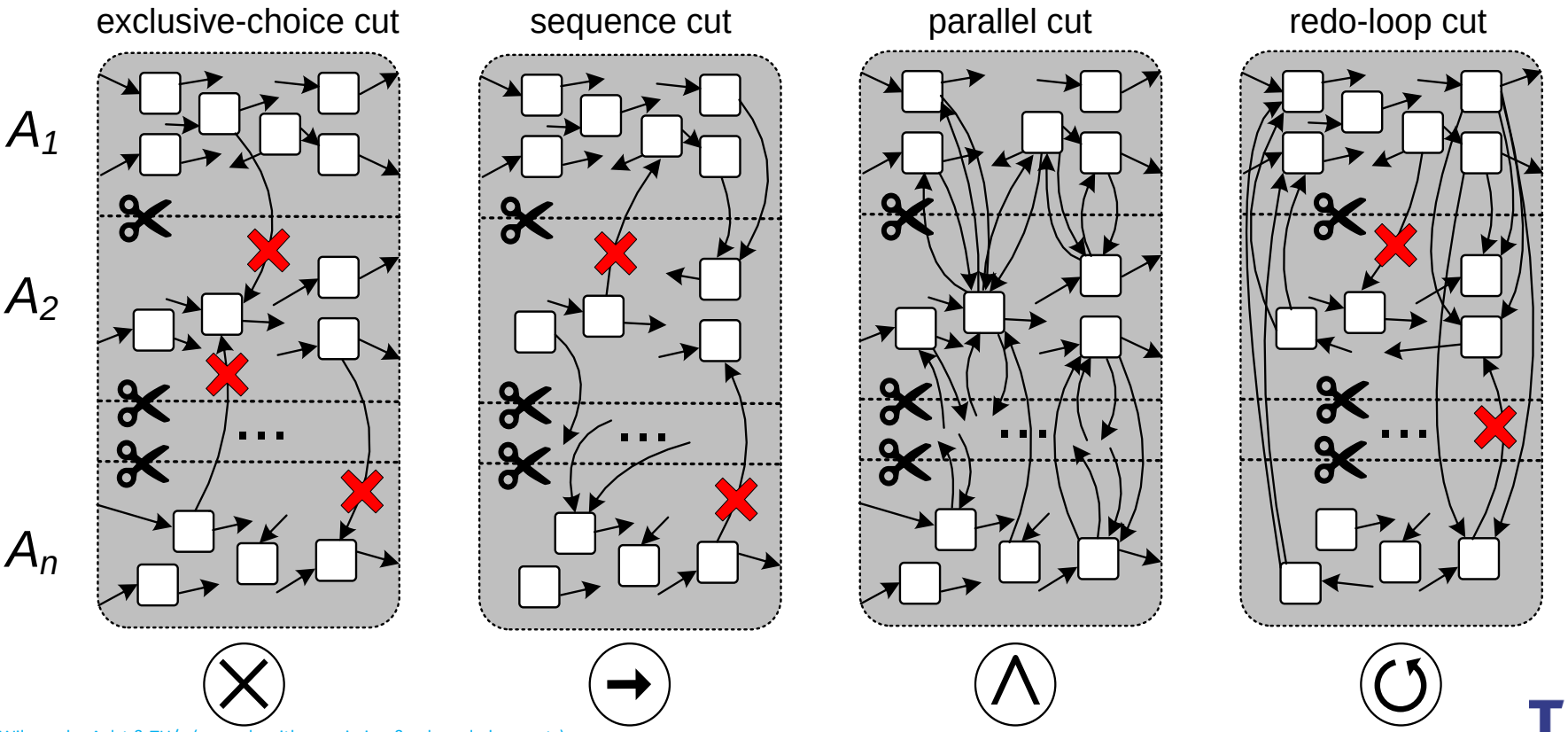








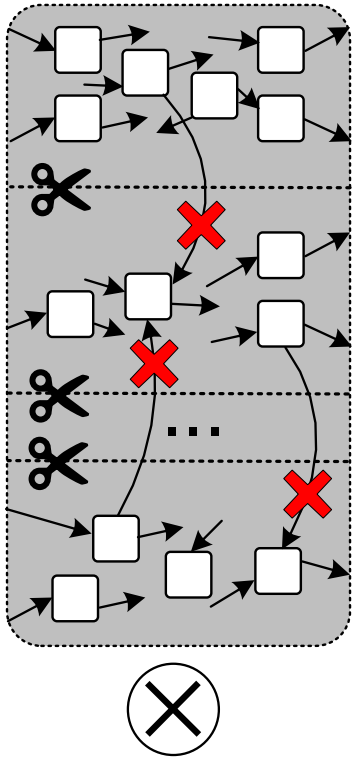
How to cut the event log?



TU/e

How to cut the event log?

exclusive-choice cut



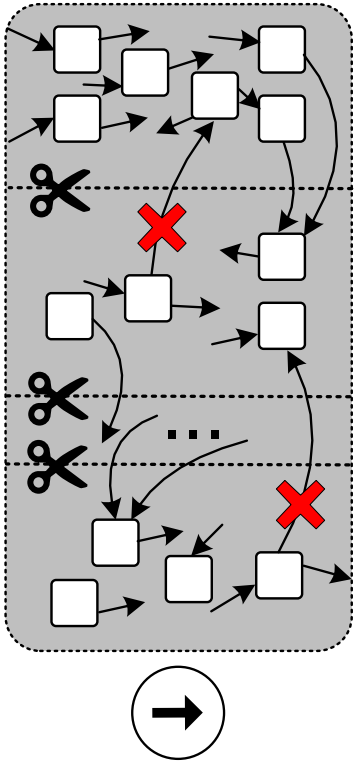
An *exclusive-choice cut* of $G(L)$ is a cut $(\times, A_1, A_2, \ldots, A_n)$ such that

– $\forall_{i,j \in \{1,\ldots,n\}} \forall_{a \in A_i} \forall_{b \in A_j} \; i \neq j \;\Rightarrow\; a \not\mapsto_L b.$

TU/e

How to cut the event log?

sequence cut



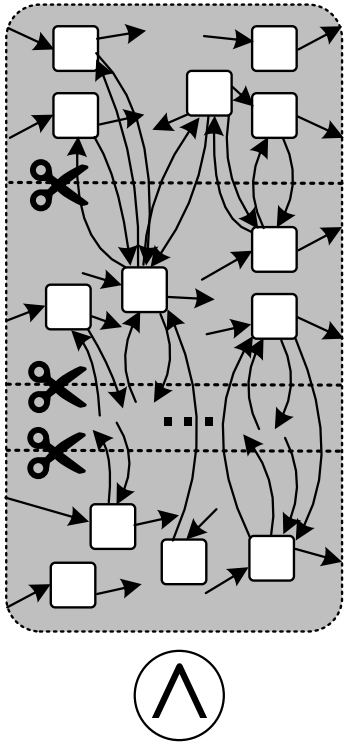
A *sequence cut* of $G(L)$ is a cut $(\rightarrow, A_1, A_2, \ldots, A_n)$ such that

– $\forall_{i,j \in \{1,\ldots,n\}} \forall_{a \in A_i} \forall_{b \in A_j} \; i < j \;\Rightarrow\; (a \mapsto_L^+ b \;\wedge\; b \not\mapsto_L^+ a)$.

TU/e

How to cut the event log?

parallel cut



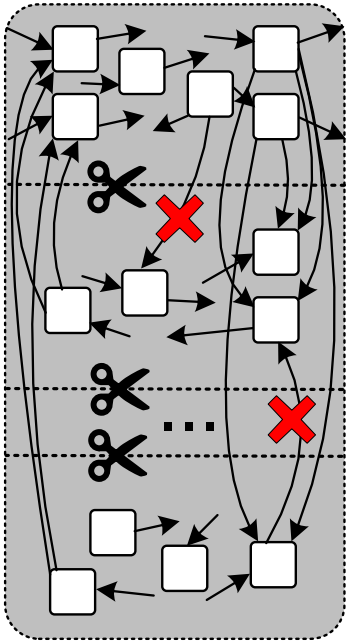
A *parallel cut* of $G(L)$ is a cut $(\wedge, A_1, A_2, \ldots, A_n)$ such that

- $\forall_{i \in \{1,\ldots,n\}} A_i \cap A_L^{start} \neq \emptyset \wedge A_i \cap A_L^{end} \neq \emptyset$ and
- $\forall_{i,j \in \{1,\ldots,n\}} \forall_{a \in A_i} \forall_{b \in A_j} i \neq j \Rightarrow a \mapsto_L b$.

TU/e

How to cut the event log?

redo-loop cut

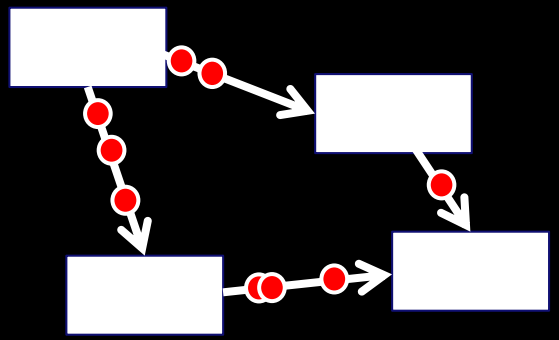


A *redo-loop cut* of $G(L)$ is a cut $(\circlearrowleft, A_1, A_2, \ldots, A_n)$ such that

- $n \geq 2$,
- $A_L^{start} \cup A_L^{end} \subseteq A_1$,
- $\{a \in A_1 \mid \exists_{i \in \{2,\ldots,n\}} \exists_{b \in A_i}\ a \mapsto_L b\} \subseteq A_L^{end}$,
- $\{a \in A_1 \mid \exists_{i \in \{2,\ldots,n\}} \exists_{b \in A_i}\ b \mapsto_L a\} \subseteq A_L^{start}$,
- $\forall_{i,j \in \{2,\ldots,n\}} \forall_{a \in A_i} \forall_{b \in A_j}\ i \neq j \Rightarrow a \not\mapsto_L b$,
- $\forall_{i \in \{2,\ldots,n\}} \forall_{b \in A_i} \exists_{a \in A_L^{end}}\ a \mapsto_L b \Rightarrow \forall_{a' \in A_L^{end}}\ a' \mapsto_L b$, and
- $\forall_{i \in \{2,\ldots,n\}} \forall_{b \in A_i} \exists_{a \in A_L^{start}}\ b \mapsto_L a \Rightarrow \forall_{a' \in A_L^{start}}\ b \mapsto_L a'$.

C11 (L) TOTAL C1
25

	A	B	C	D
1	ITEM	NO.	UNIT	COST
2	----	---	----	----
3	MUCK RAKE	43	12.95	556.85
4	BUZZ CUT	15	6.75	101.25
5	TOE TONER	250	49.95	12487.50
6	EYE SNUFF	2	4.95	9.90
7				----------
8			SUBTOTAL	13155.50
9			9.75% TAX	1282.66
10				----------
11			TOTAL	14438.16
12				
13				
14				
15				
16				



More info