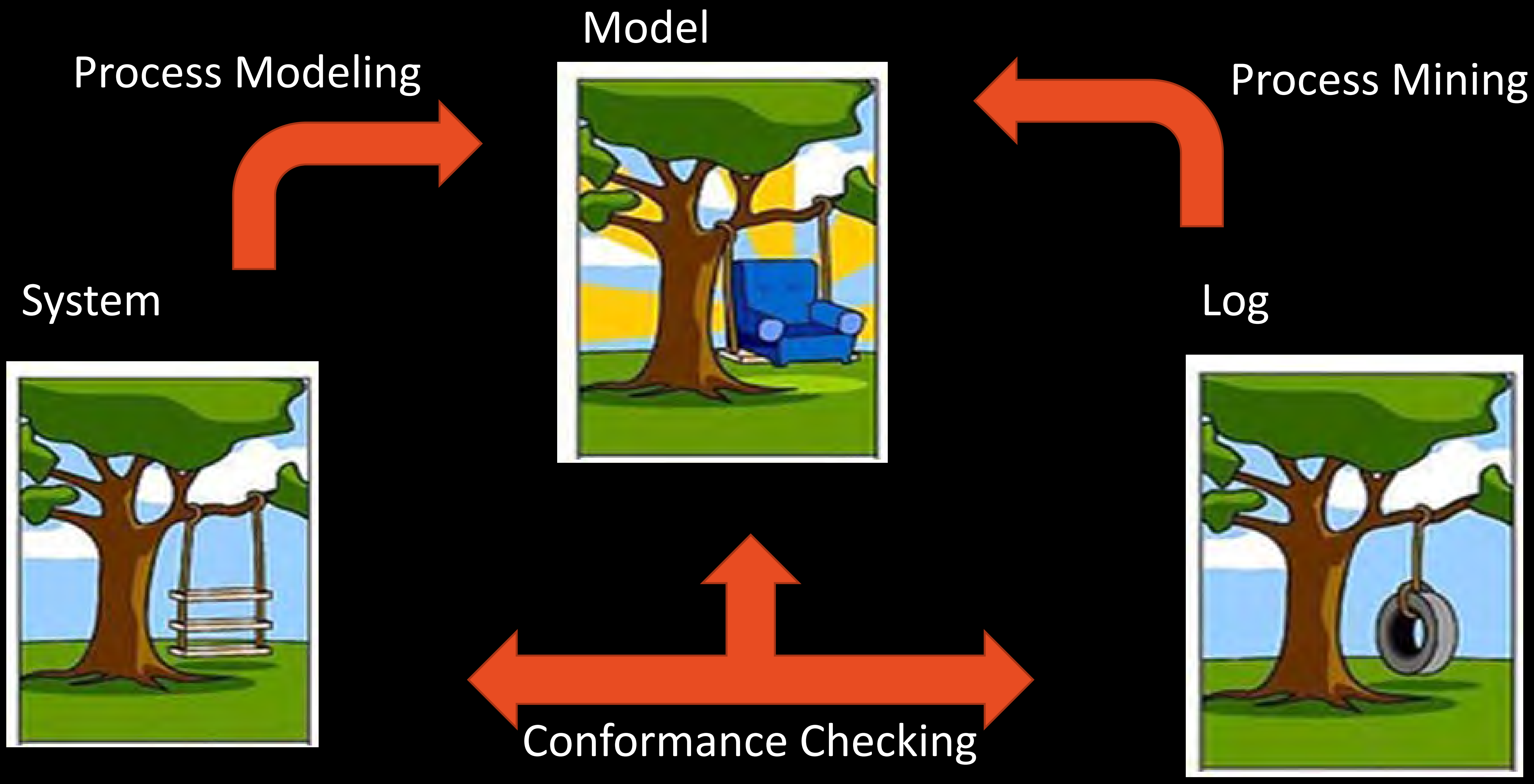
A Unified Approach for Measuring Precision and Generalization Based on Anti-Alignments

Boudewijn van Dongen

Josep Carmona

Thomas Chatain

Conformance in Process Mining



Model

Process Modeling

Process Mining

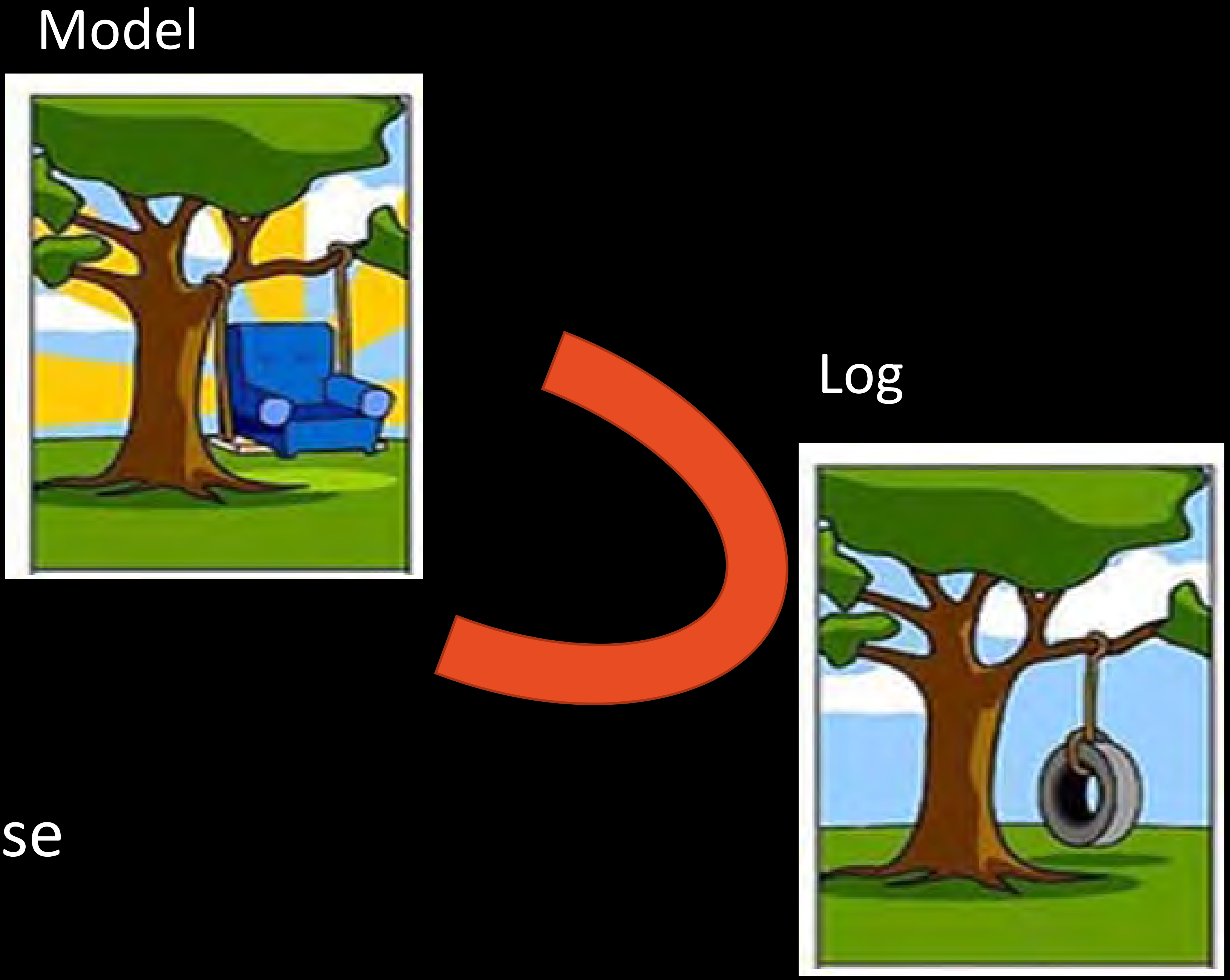
System

Log

Conformance Checking

Alignments & Fitness

- *Fitness* is a measure for the amount of behavior shown in the log that fits the model
- Alignments provide the basis for computing fitness
- An *alignment* shows where deviations occurred and why these deviation are considered as such



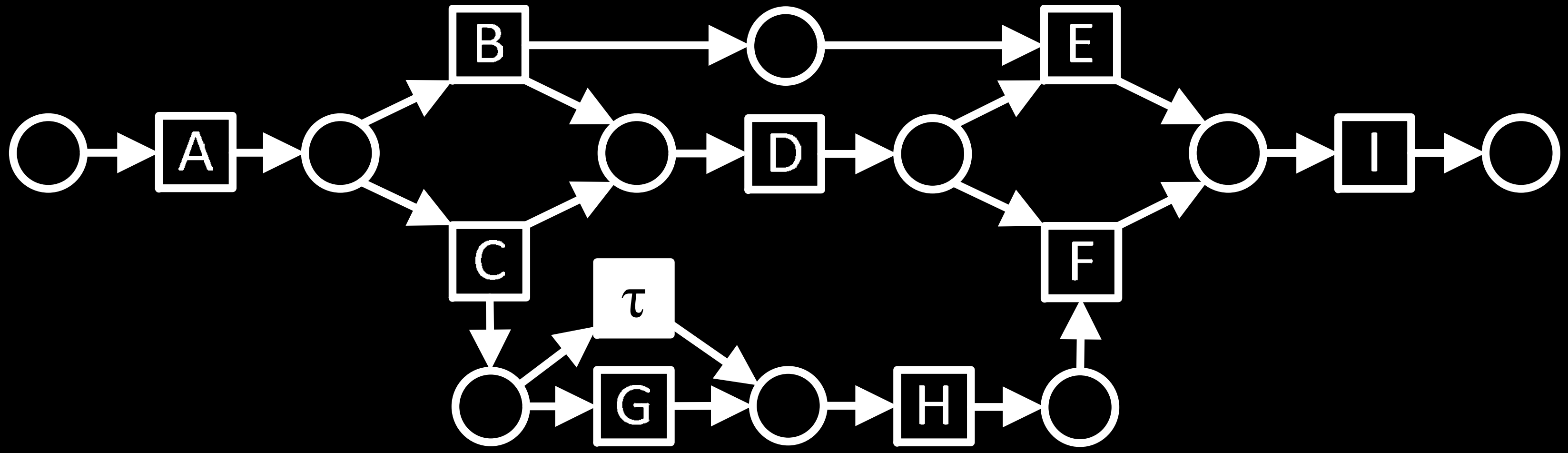
Anti-Alignments

- Anti-alignments show how far the model allows you to deviate from observed behavior
- Given a model *M*, a finite log *L* and a distance function *d*, an *(n,m)* anti-alignment is a firing sequence *s* of the model of length *n*, such that for each trace *t* in the log holds that *d(s,t) ≥ m*

Anti-Alignments - basics

- Given a model *M*, a finite log *L* and a distance function *d*, an *(n,m)* anti-alignment is a firing sequence *s* of the model of length *n*, such that for each trace *t* in the log holds that *d(s,t) ≥ m*
- A maximal complete anti alignment of length *n* reaches the final marking and maximizes the distance *m.*

Trace	Frequency
<A,B,D,E,I>	1207
<A,C,D,G,H,F,I>	145
<A,C,G,D,H,F,I>	56
<A,C,H,D,F,I>	23
<A,C,D,H,F,I>	28



Maximal complete AA:<A,C,G,H,D,F,I> with *n=7* and *m=$^1/_7$*

Anti-Alignments - properties

- No anti alignments have to exist, if *L* is the language of *M*
- If *M* has a loop, infinitely many anti-alignments exist and their distances typically go to 1.
- Finding an anti-alignment with maximal *m*, given *n* can be translated into a SAT problem (when using hamming distances)
- Finding an anti-alignment with minimal *n*, given *m* can be translated into a SAT problem (when using hamming distances)

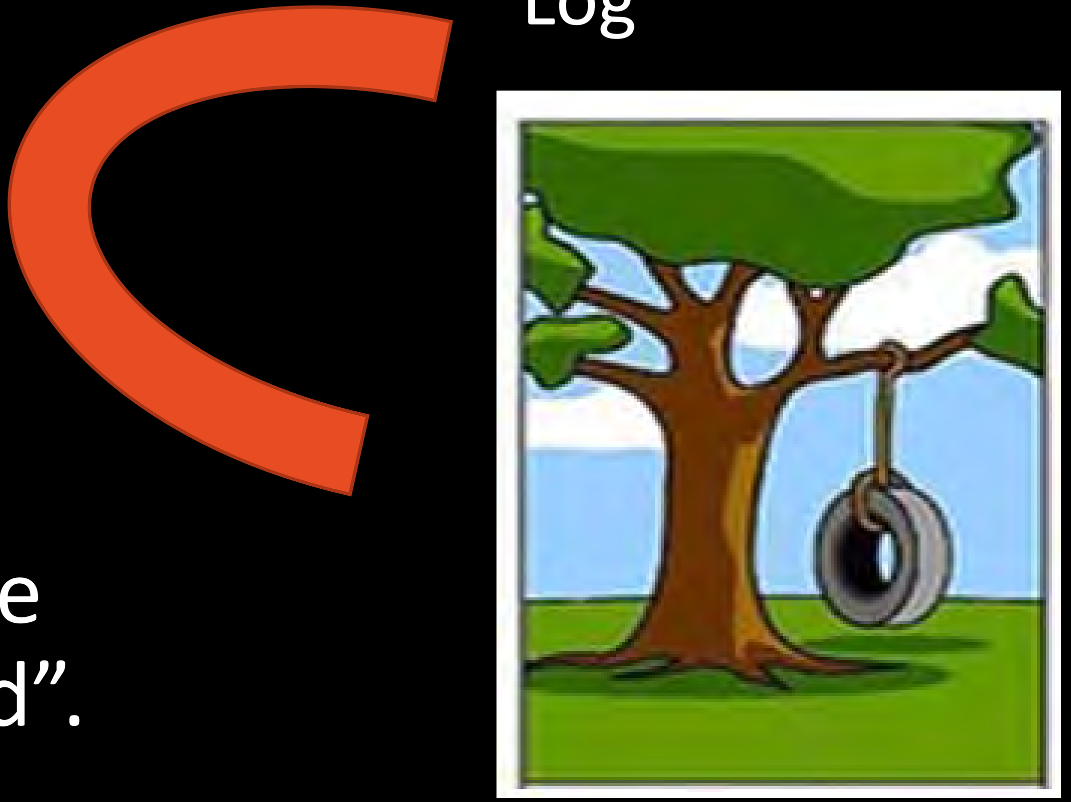
- *No smart way exists yet for computing anti-alignments using edit-distances.*

Conformance Checking: Precision

- Precision is a measure for the fraction of the behavior of the model that is not in the log
- Simply comparing the (possibly infinite) behavior is infeasible and not very informative.
- Most precision metrics estimate the size of the “unseen” behavior by looking “one step ahead”.

Model

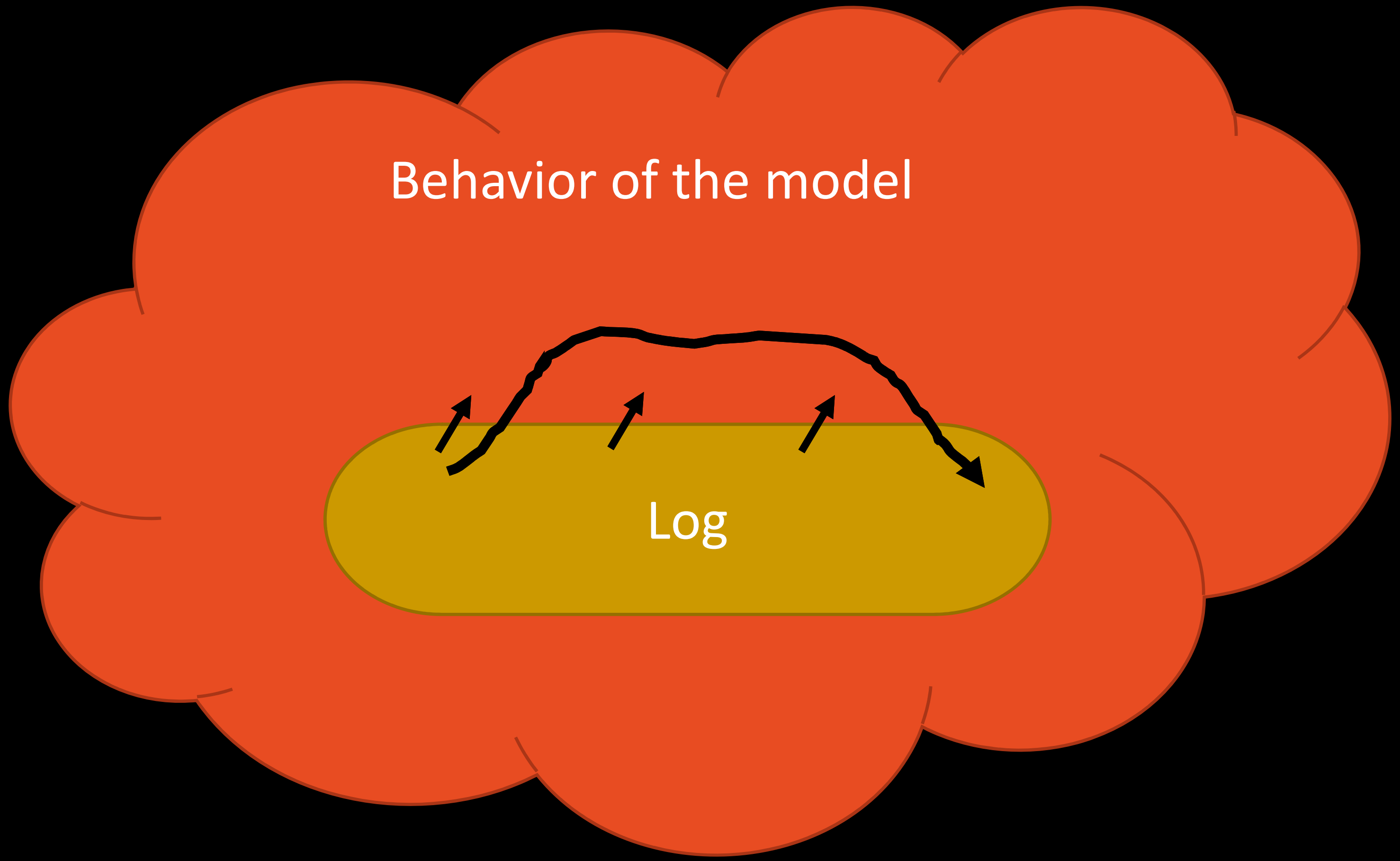
Log



Anti-Alignment Based Precision

- Anti-alignments show behavior of the model as different as possible from behavior in the log

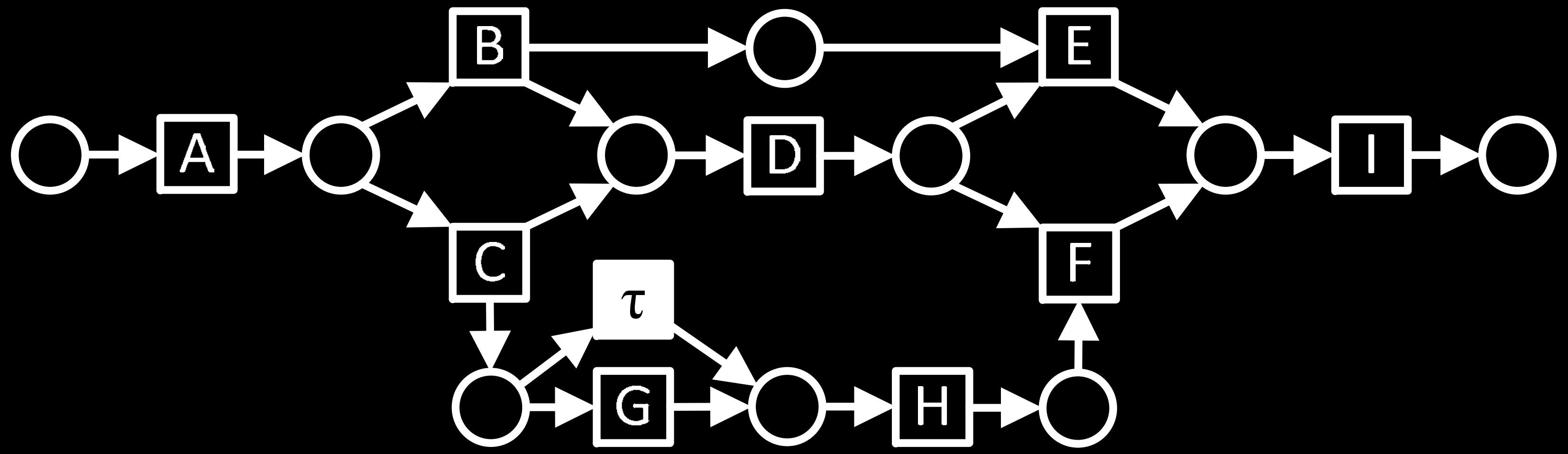
- With anti-alignments, you look at entire paths!



Anti-Alignment Based Precision

- Consider a model *M* and a log *L*.
- Now remove a trace *t* from the log to get L^t.
- Compute a maximal complete anti alignment *s* of length |*t*| for log L^t.
- In a very precise model, *s* = *t*, hence 1-*d(s,t)* is a precision measure.

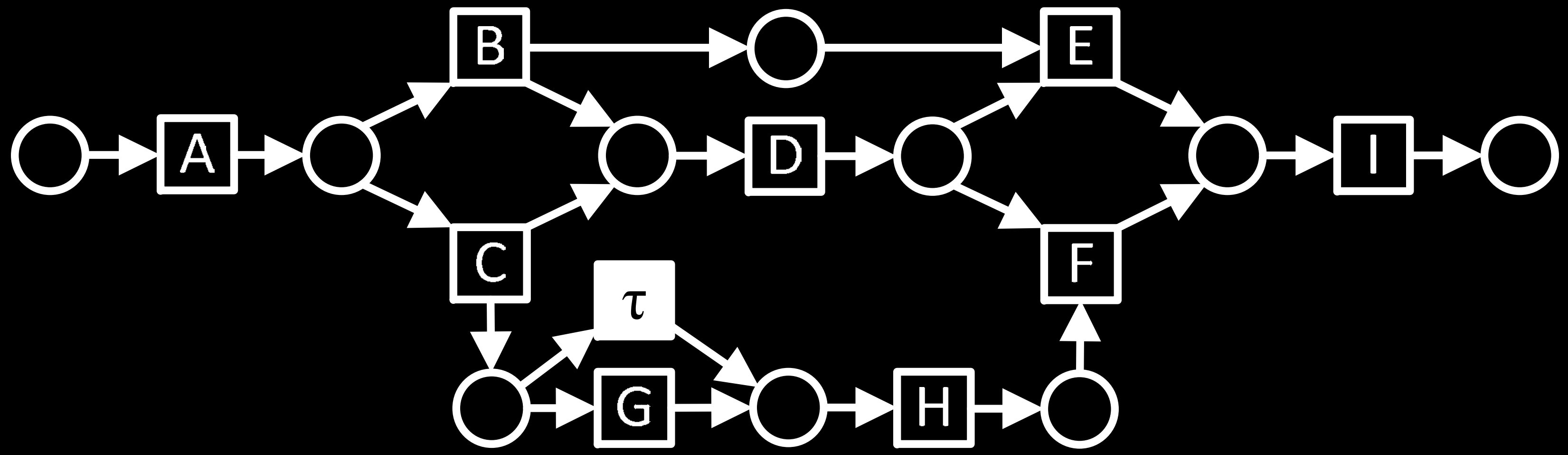
Trace *t*	**Freq.**	**AA for L^t *(s)***	***d(s,t)***
<A,B,D,E,I>	1207		
<A,C,D,G,H,F,I>	145		
<A,C,G,D,H,F,I>	56		
<A,C,H,D,F,I>	23		
<A,C,D,H,F,I>	28		



Anti-Alignment Based Precision

- Consider a model *M* and a log *L*.
- Now remove a trace *t* from the log to get L^t.
- Compute a maximal complete anti alignment *s* of length |*t*| for log L^t.
- In a very precise model, *s* = *t*, hence 1-*d(s,t)* is a precision measure.

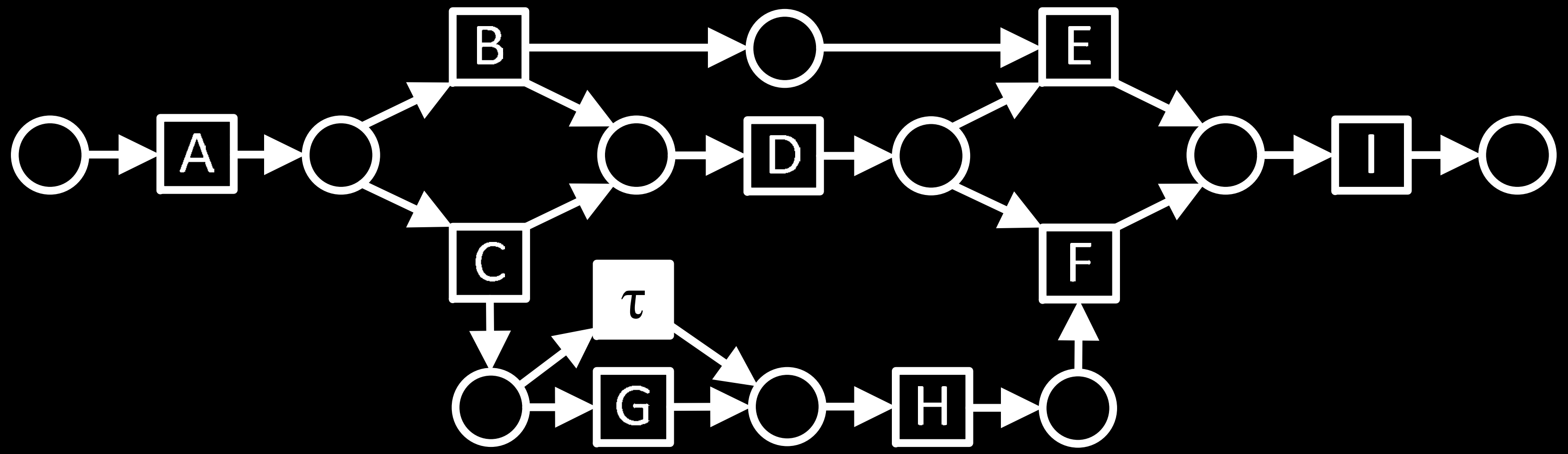
Trace *t*	**Freq.**	**AA for L^t *(s)***	***d(s,t)***
<A,B,D,E,I>	1207	<A,B,D,E,I>	0
<A,C,D,G,H,F,I>	145		
<A,C,G,D,H,F,I>	56		
<A,C,H,D,F,I>	23		
<A,C,D,H,F,I>	28		



Anti-Alignment Based Precision

- Consider a model *M* and a log *L*.
- Now remove a trace *t* from the log to get L^t.
- Compute a maximal complete anti alignment *s* of length |*t*| for log L^t.
- In a very precise model, *s* = *t*, hence 1-*d(s,t)* is a precision measure.

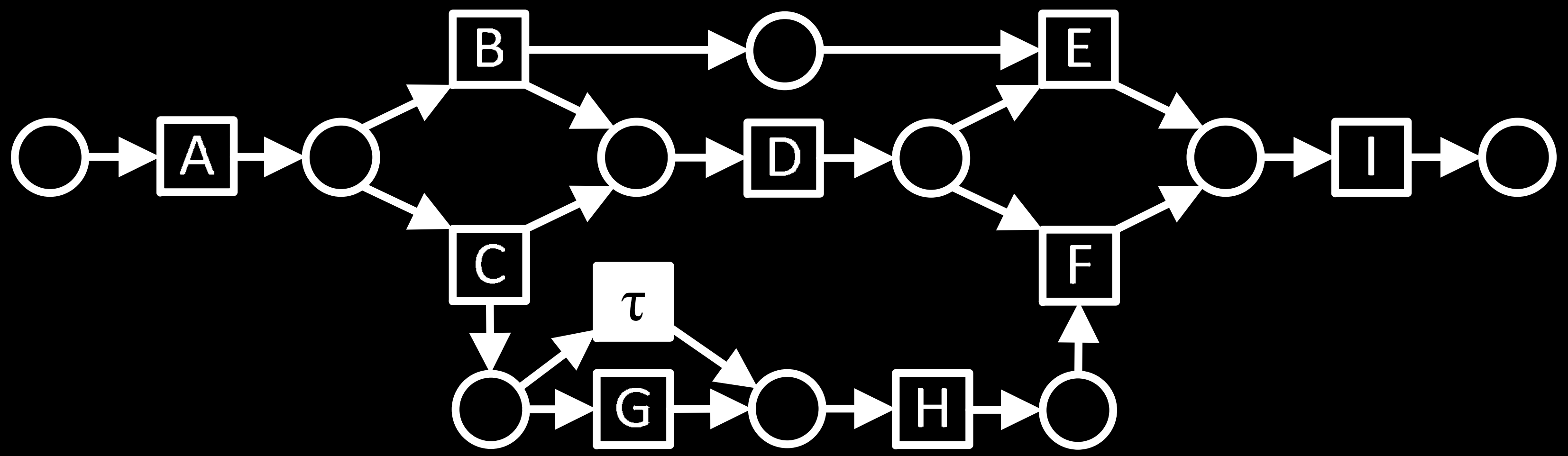
Trace *t*	**Freq.**	**AA for L^t *(s)***	***d(s,t)***
<A,B,D,E,I>	1207	<A,B,D,E,I>	0
<A,C,D,G,H,F,I>	145	<A,C,G,H,D,F,I>	$^2/_7$
<A,C,G,D,H,F,I>	56		
<A,C,H,D,F,I>	23		
<A,C,D,H,F,I>	28		



Anti-Alignment Based Precision (trace based)

- Consider a model *M* and a log *L*.
- Now remove a trace *t* from the log to get L^t.
- Compute a maximal complete anti alignment *s* of length |*t*| for log L^t.
- In a very precise model, *s* = *t*, hence 1-*d(s,t)* is a precision measure.

Trace *t*	**Freq.**	**AA for L^t : *s***	***d(s,t)***
<A,B,D,E,I>	1207	<A,B,D,E,I>	0
<A,C,D,G,H,F,I>	145	<A,C,G,H,D,F,I>	$^2/_7$
<A,C,G,D,H,F,I>	56	<A,C,G,H,D,F,I>	$^2/_7$
<A,C,H,D,F,I>	23	<A,C,H,D,F,I>	0
<A,C,D,H,F,I>	28	<A,C,D,H,F,I>	0

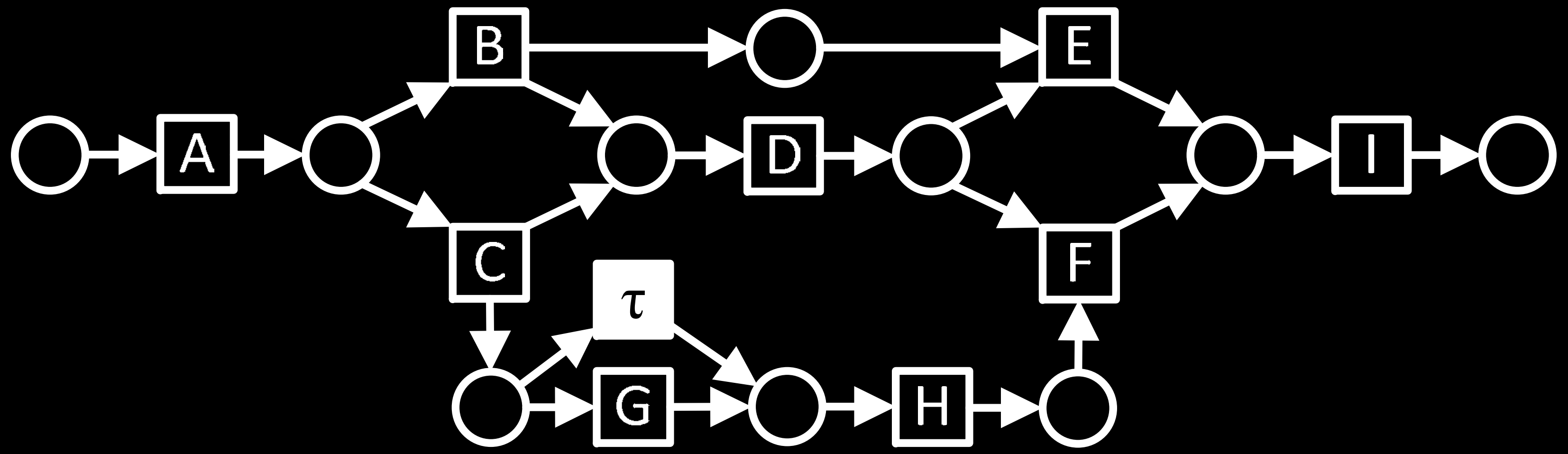


Precision $P_t = (1+{}^5/_7+{}^5/_7+1+1)/5 = 0.886$

Anti-Alignment Based Precision (log based)

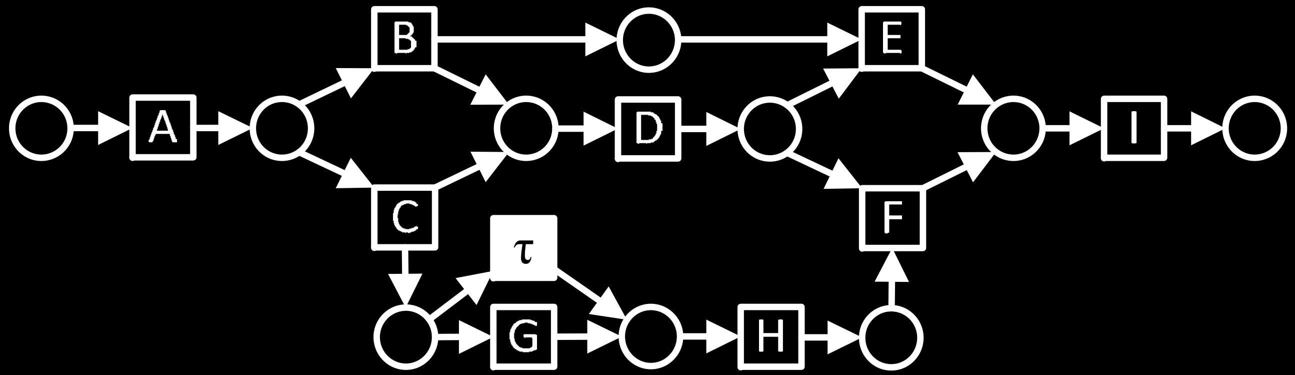
- Consider a model *M* and a log *L*.
- Compute a maximal complete anti alignment *s* of length $x{\cdot}|t|^{mx}$ for log *L*.
- In a very precise model, $s \in L$ and hence the minimal *d(s,t)* will be 0.

Trace *t*	**Freq.**	**AA for L^t : *s***	***d(s,t)***
<A,B,D,E,I>	1207	<A,B,D,E,I>	0
<A,C,D,G,H,F,I>	145	<A,C,G,H,D,F,I>	$^2/_7$
<A,C,G,D,H,F,I>	56	<A,C,G,H,D,F,I>	$^2/_7$
<A,C,H,D,F,I>	23	<A,C,H,D,F,I>	0
<A,C,D,H,F,I>	28	<A,C,D,H,F,I>	0
-	-	<A,C,G,H,D,F,I>	$^1/_7$

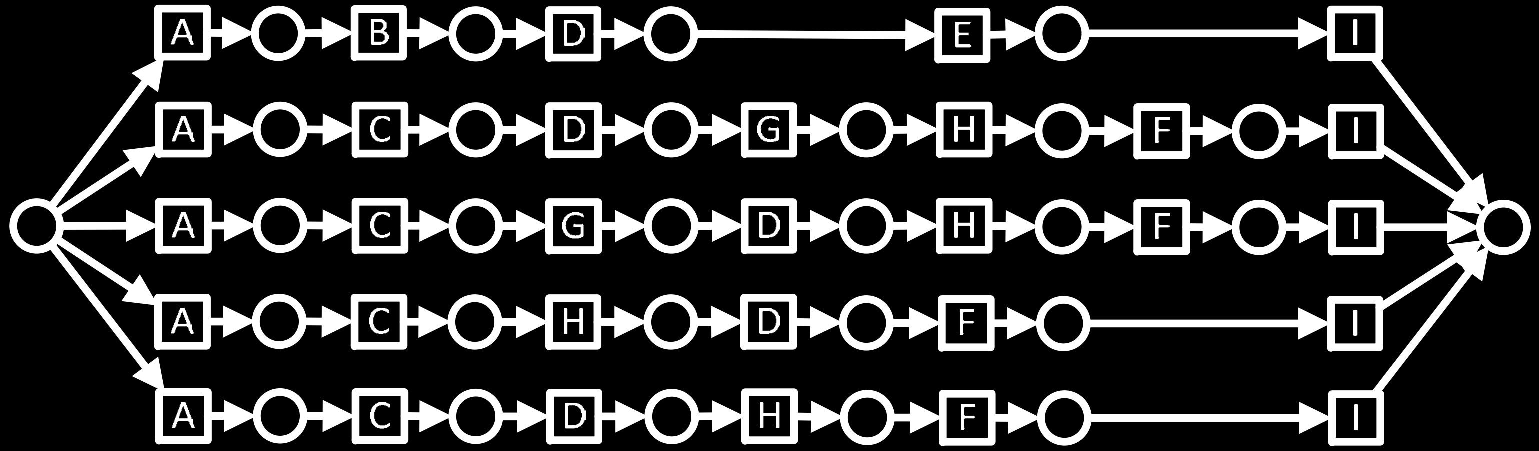


Precision $P_l^2 = 1 - {}^1/_7 = 0.857$

Precision

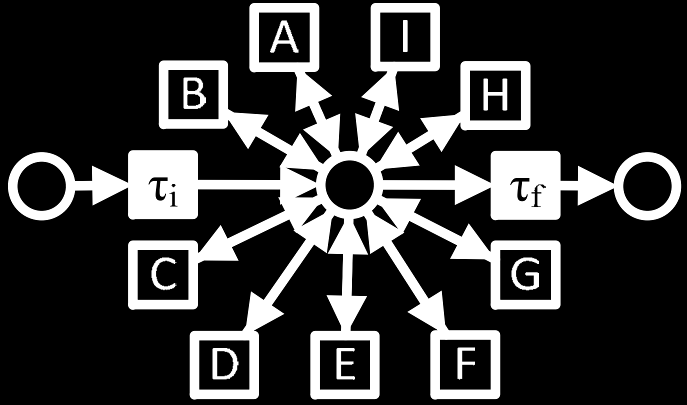


$P_{ETC} = 0.994$, $P_a = 0.982$, $P_t = 0.886$, $P_l = 0.857$

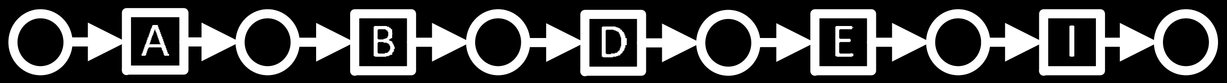


$P_{ETC} = 0.359$, $P_a = 1$, $P_t = 1$, $P_l = 1$

Trace	Frequency
<A,B,D,E,I>	1207
<A,C,D,G,H,F,I>	145
<A,C,G,D,H,F,I>	56
<A,C,H,D,F,I>	23
<A,C,D,H,F,I>	28

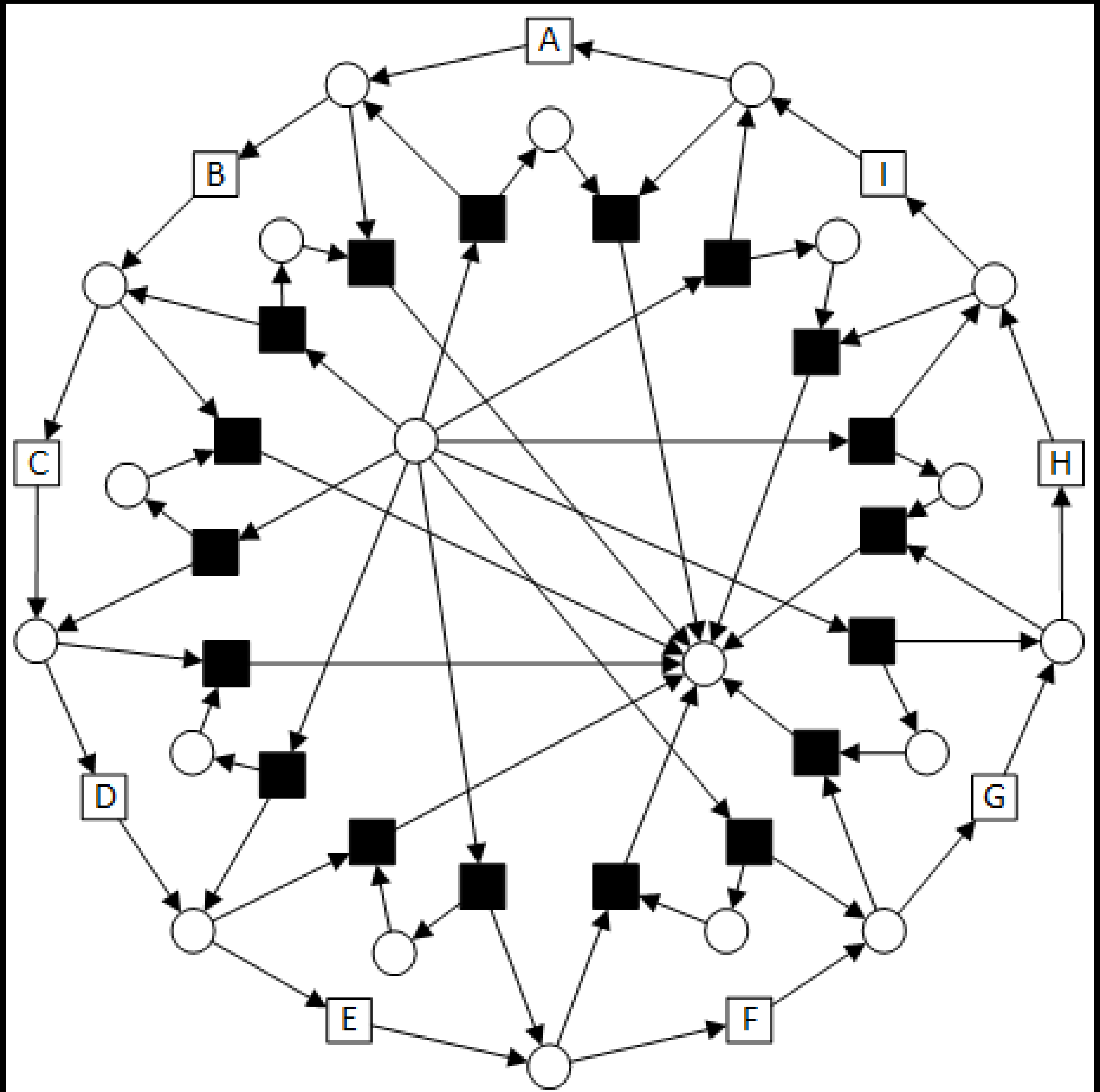


$P_{ETC} = 0.119$, $P_a = 0.142$, $P_t = 0$, $P_l = 0$



$P_{ETC} = 1$, $P_a = 1$, $P_t = 1$, $P_l = 1$

Precision



$P_{ETC} = 0.185$, $P_a = 0.889$, $P_t = 0$, $P_l = 0$

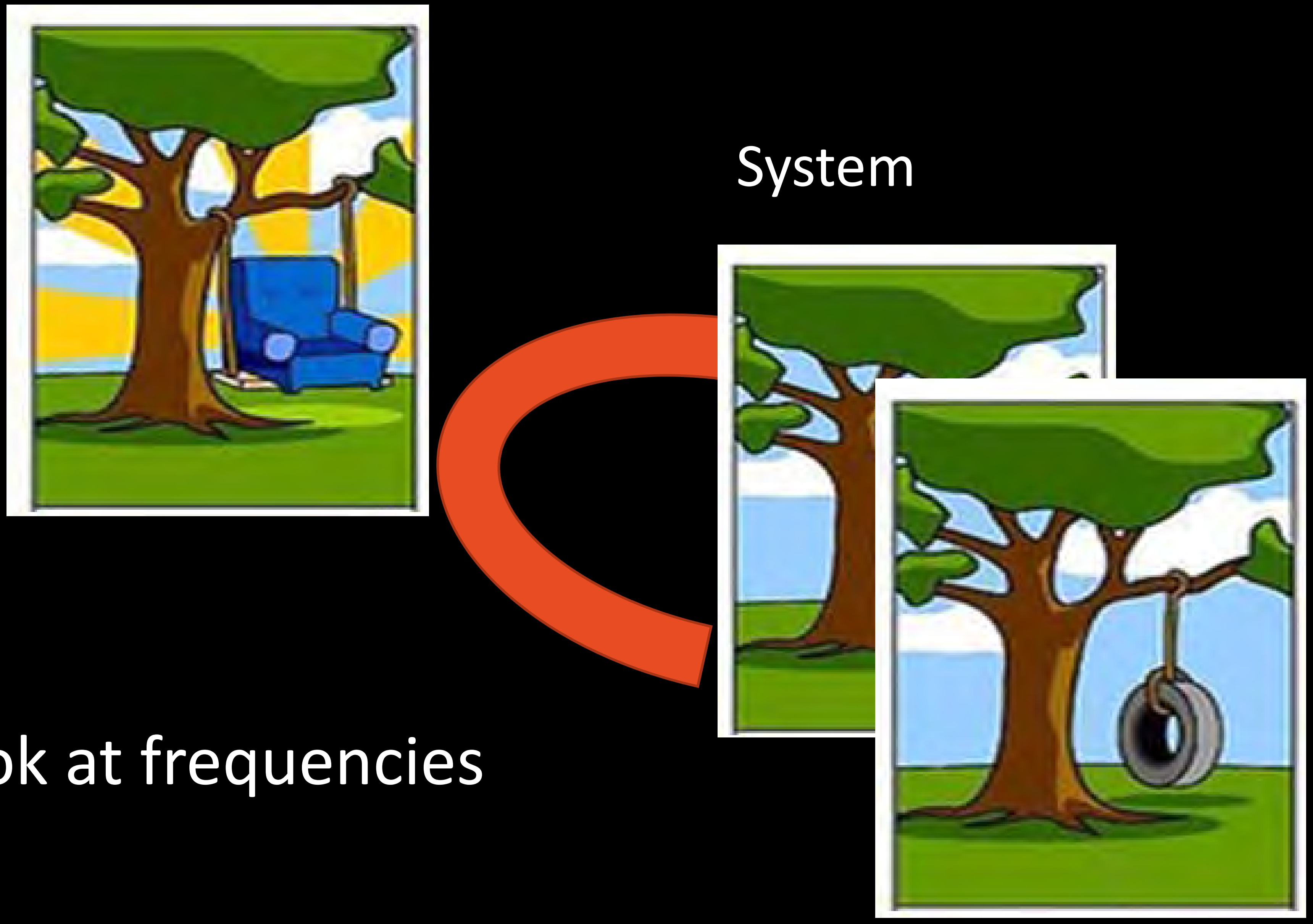
Trace	Frequency
<A,B,D,E,I>	1207
<A,C,D,G,H,F,I>	145
<A,C,G,D,H,F,I>	56
<A,C,H,D,F,I>	23
<A,C,D,H,F,I>	28

Conformance Checking: Generalization

- Generalization is a measure for the ability of a model to predict unseen, but correct behavior

- Behavior is correct if it is part of the unknown system

- Most generalization metrics look at frequencies of model elements

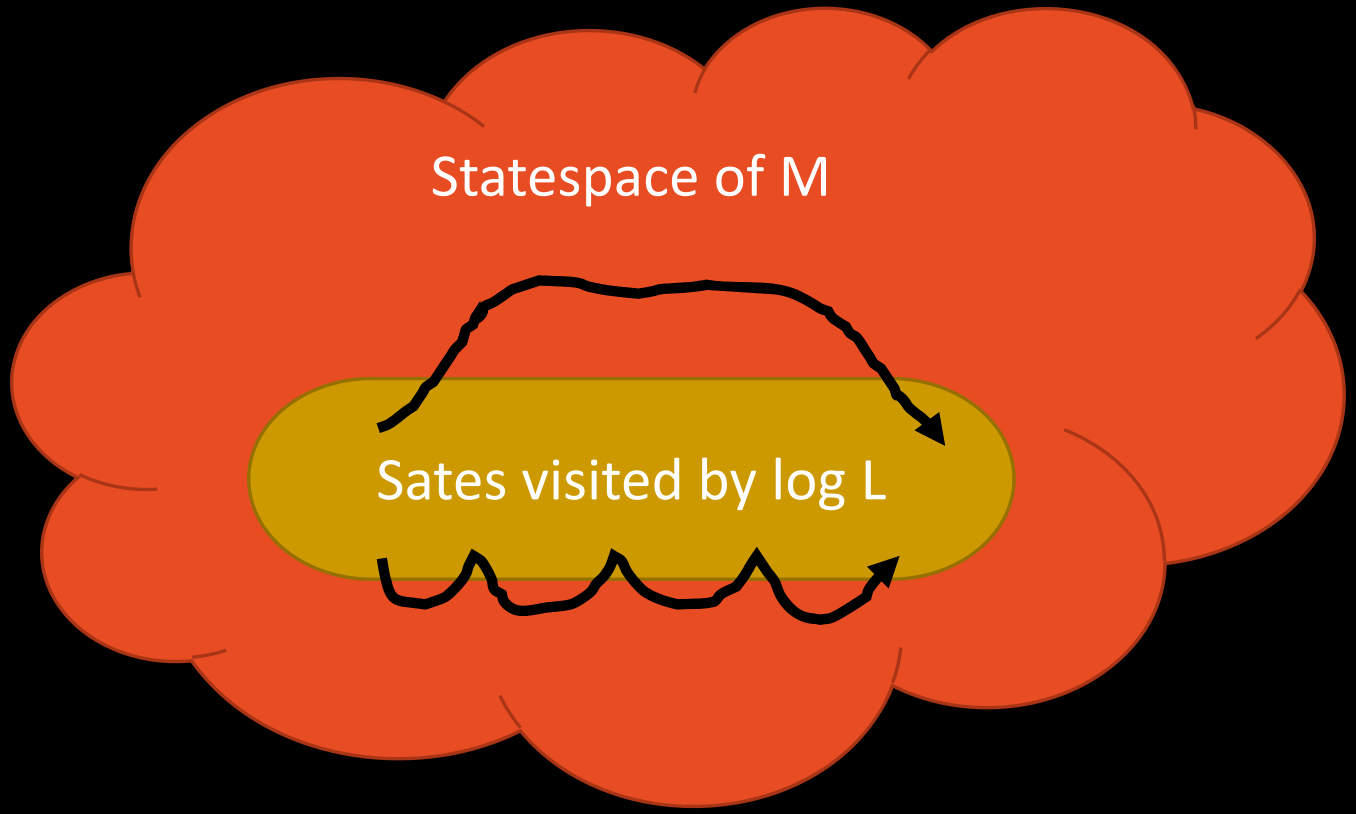


Model

System

Generalization

- Assumption: A model generalizes if it allows for new sequences of behavior while not introducing too many new states.
- Again, we use anti-alignments!
- Recovery distance is a measure for the maximal number of steps to get from the anti-alignment back into the statespace covered by the log.

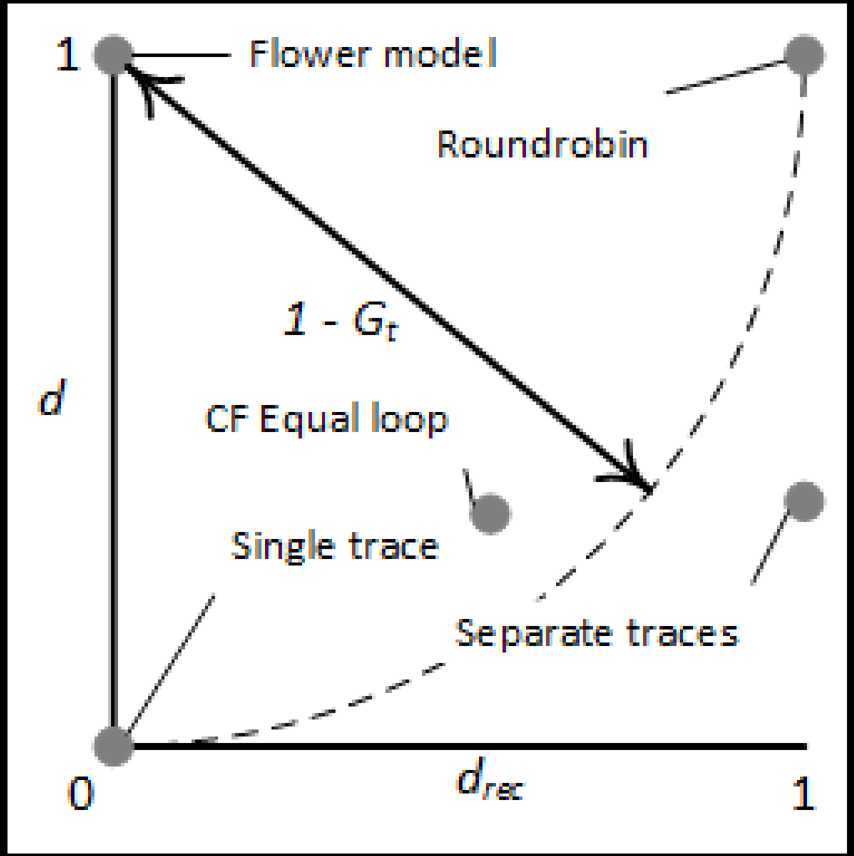


Anti-Alignment Based Generalization

- Consider a model *M* and a log *L*

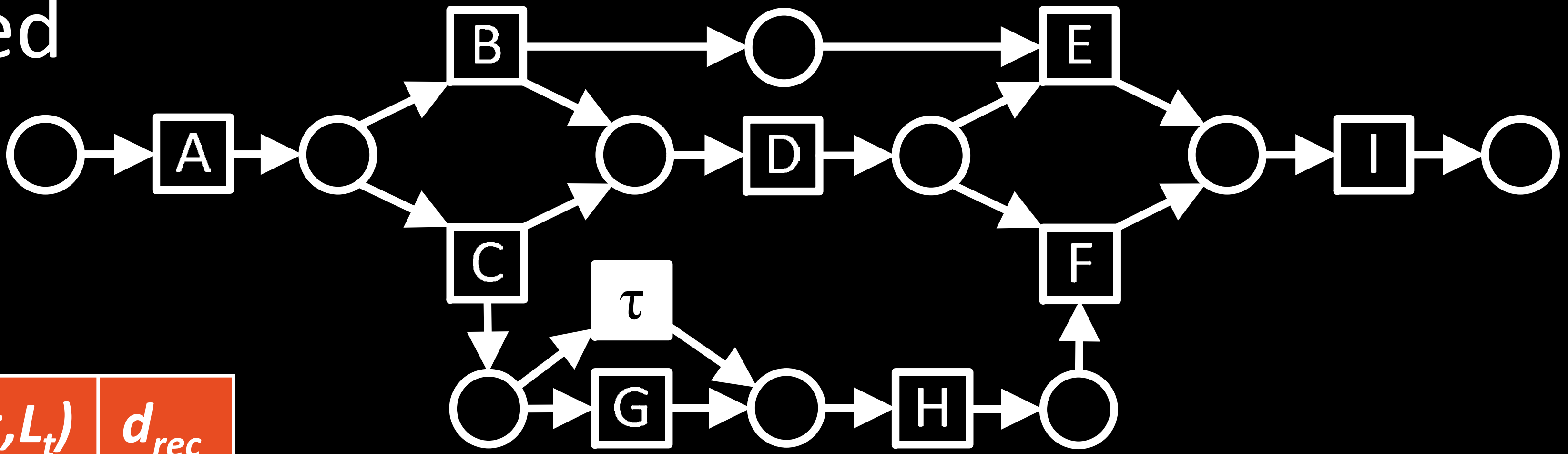
- Now remove a trace *t* from the log to get L^t and compute the corresponding anti-alignment *s*

- A model generalizes if s is very difference from the traces in L^t is high, and the recovery distance is low



Anti-Alignment Based Generalization

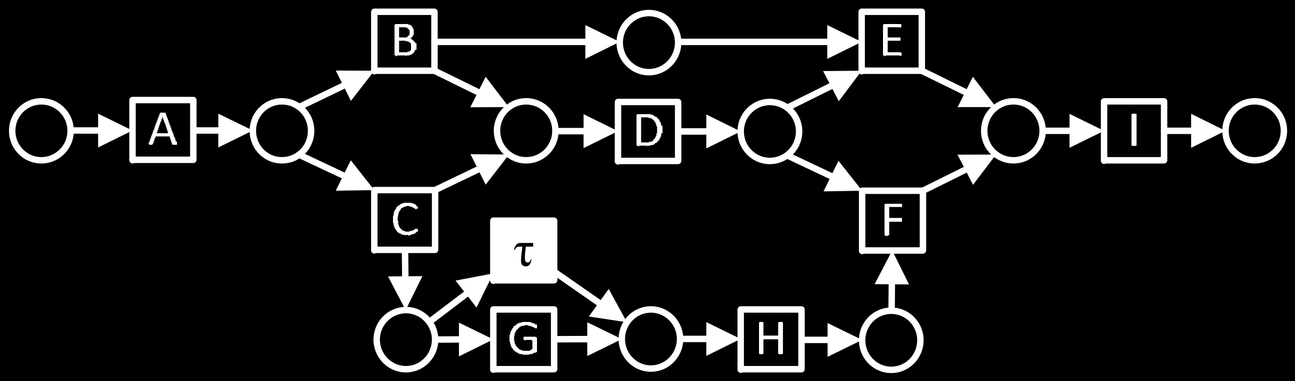
- Same procedure is used as for precision
- Distance to the log L^t is considered
- Recovery distance is considered
- Trace frequency is considered!



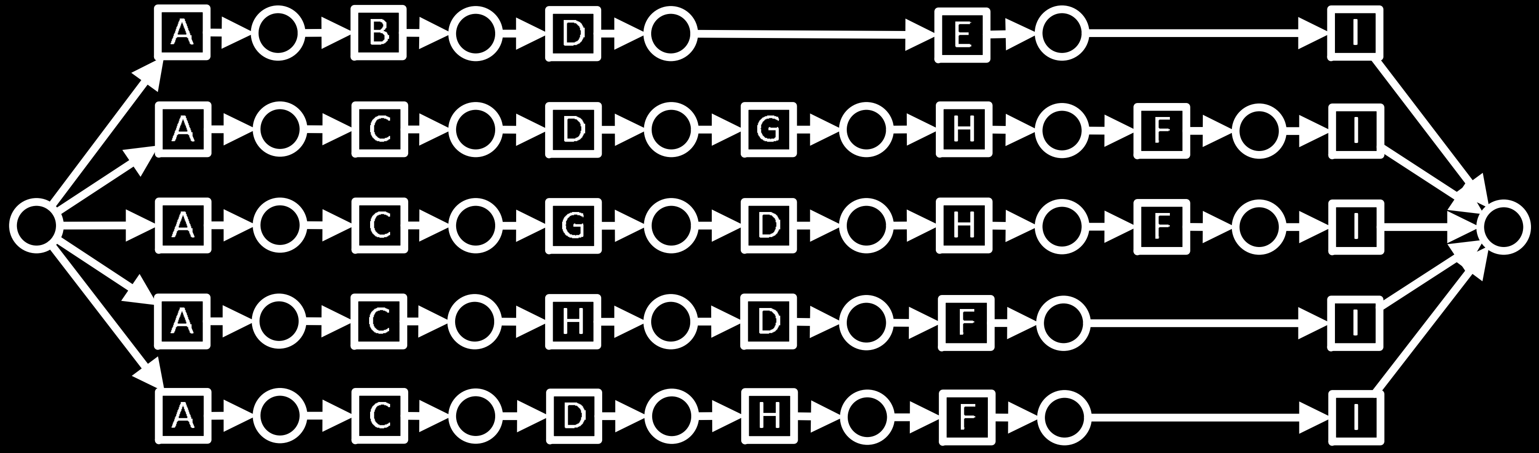
Trace *t*	**Freq.**	**AA for L^t : *s***	***d(s,t)***	***d(s,L_t)***	d_{rec}
<A,B,D,E,I>	1207	<A,B,D,E,I>	0	$^3/_6$	$^2/_4$
<A,C,D,G,H,F,I>	145	<A,C,G,H,D,F,I>	$^2/_7$	$^1/_7$	0
<A,C,G,D,H,F,I>	56	<A,C,G,H,D,F,I>	$^2/_7$	$^1/_7$	0
<A,C,H,D,F,I>	23	<A,C,H,D,F,I>	0	$^2/_6$	$^1/_6$
<A,C,D,H,F,I>	28	<A,C,D,H,F,I>	0	$^1/_6$	0
-	-	<A,C,G,H,D,F,I>		$^1/_7$	0

Precision P_t = 0.886
Precision P_l = 0.857
Generalization G_t= 0.270
Generalization G_l = 0.143

Generalization

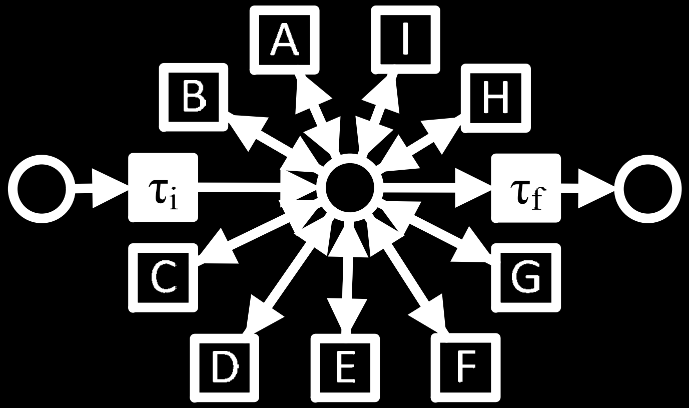


G= 0.585, G_t = 0.270, G_l = 0.143

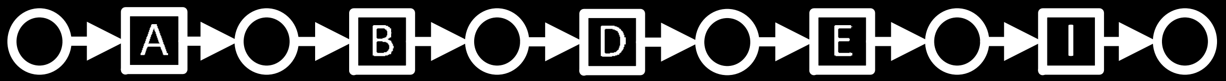


G_a = 0.145, G_t = 0, G_l = 0

Trace	Frequency
<A,B,D,E,I>	1207
<A,C,D,G,H,F,I>	145
<A,C,G,D,H,F,I>	56
<A,C,H,D,F,I>	23
<A,C,D,H,F,I>	28



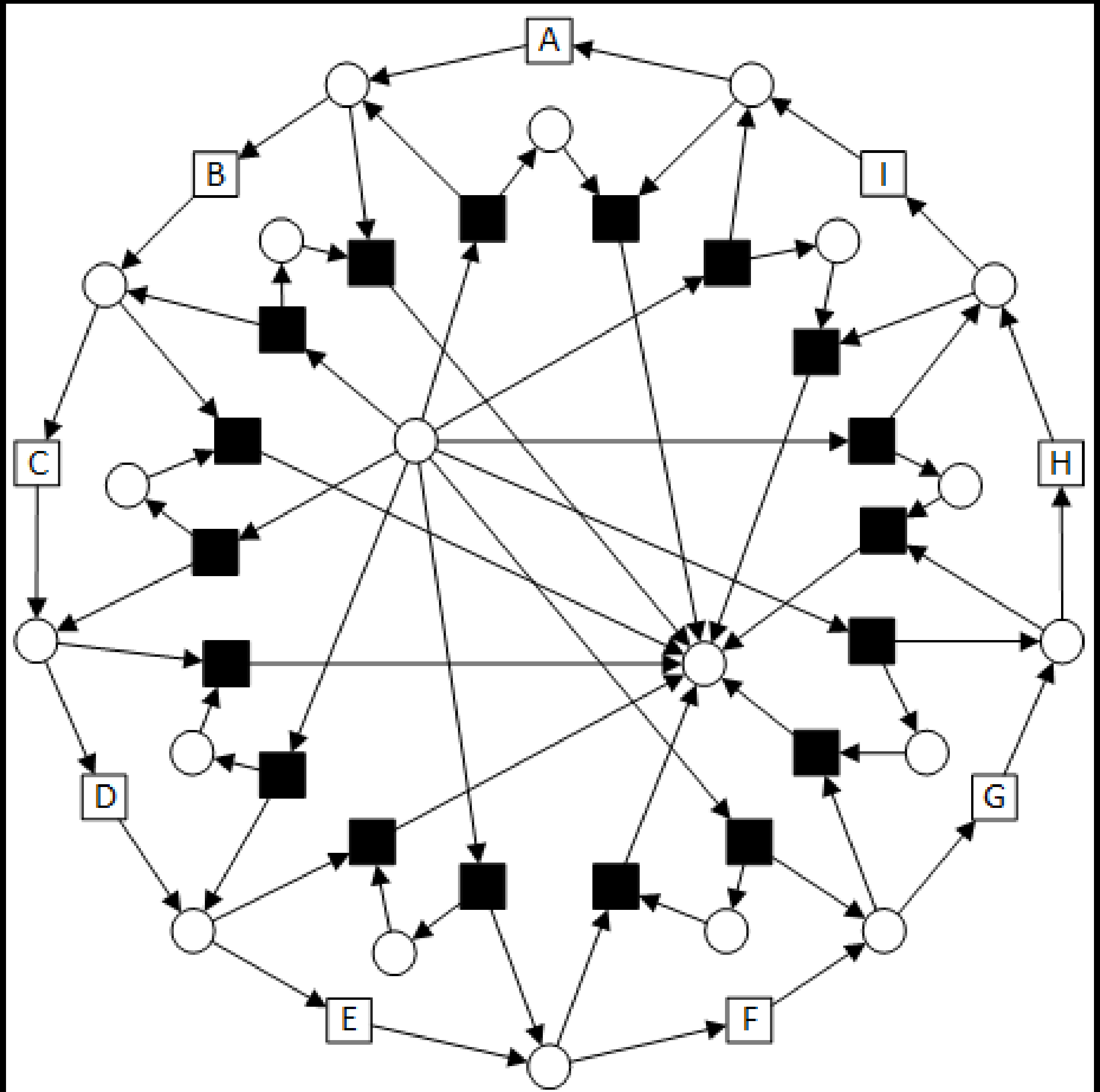
G_a = 0.903, G_t = 1, G_l = 1



G_a = 0.900, G_t = 0, G_l = 0

Generalization

Trace	Frequency
<A,B,D,E,I>	1207
<A,C,D,G,H,F,I>	145
<A,C,G,D,H,F,I>	56
<A,C,H,D,F,I>	23
<A,C,D,H,F,I>	28



$G_a = 0.4$, $G_t = 0$, $G_l = 0$

Conclusions & Future Work

- Anti-alignments provide insights into the behavior of a model outside of the log
- Using anti-alignments, we provide metrics for precision and generalization
- Our precision metric consistently ranks models with more traces as less precise as opposed to other metrics
- Efficient implementations are underway:
  - For various distance functions (hamming, edit, etc.)
  - For various model types (Petri nets / process trees)