Process mining: Leveraging event data to understand and improve organizations

Henrik Leopold
Han van der Aa

December 9, 2019
IEEE Big Data 2019, Los Angeles, USA

Overview

- This tutorial consists of four main parts:
  1. **Prerequisites:** What is a process? What is process mining?
  2. **Process Discovery:** How can we discover processes from event logs?
  3. **Conformance Checking:** How can we check whether an actual process matches the expected process?
  4. **Demo & Outlook:** What tool support is available? What are the current trends in research?

Part 1: Prerequisites

What is a Process?

"a collection of activities that take one or more kinds of input and create an output that is of value to the customer"
[Hammer & Champy 1993]

- Processes are everywhere.

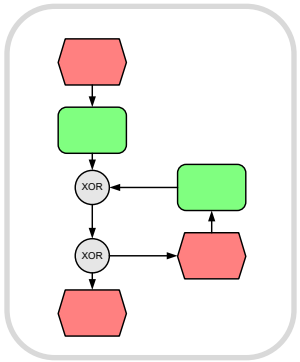
Car production

Mortgage application

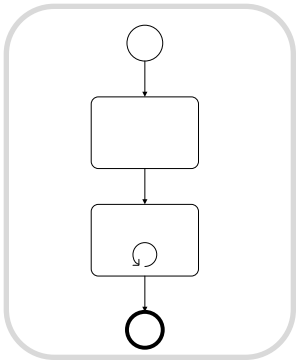
Medical treatment

How to Capture a Process?

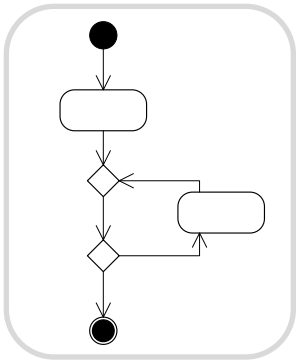
- Processes are typically captured using ***process models***.
- A process model essentially consists of a set of activities and execution constraints between them.
- There are several ***process model notations***:



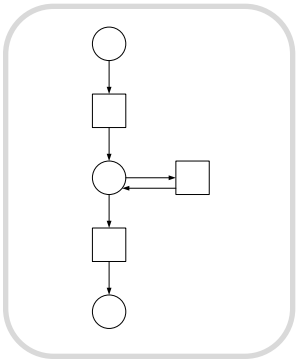
Event-driven Process Chains



Business Process Modeling and Notation



UML Activity Diagrams

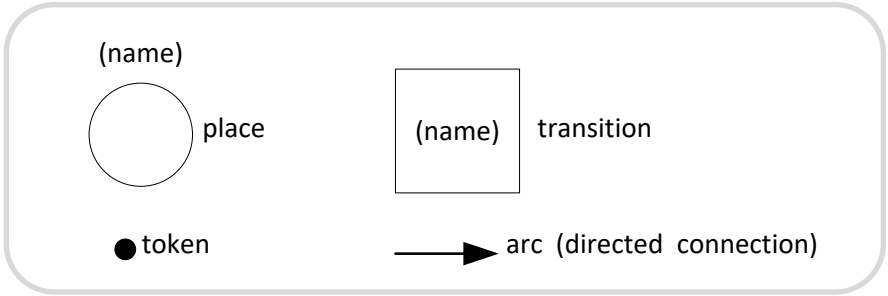


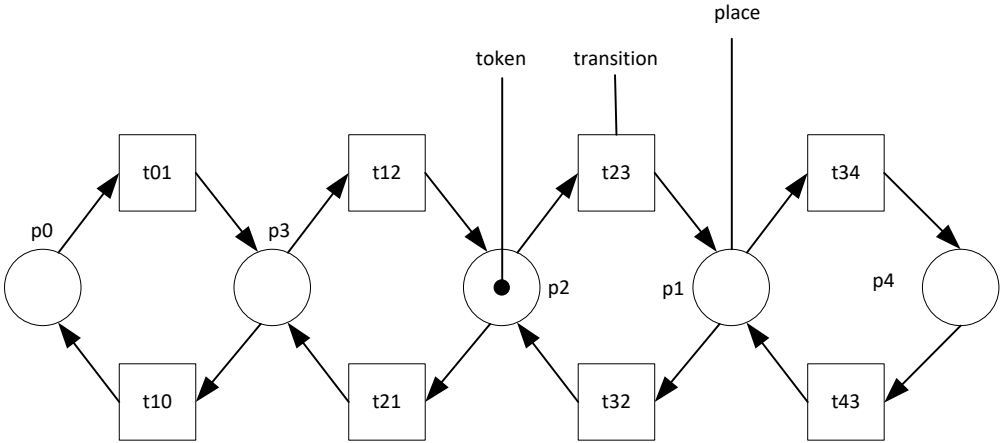
Petri Nets

Focus in This Tutorial: Petri Nets

- Petri nets are
  - based on strong mathematical foundation.
  - allow for formal analysis.
  - are “hidden” in high-level notations such as EPCs and BPMN models.
- Background
  - Foundations developed by Carl Adam Petri in 1962.
  - There are a variety of variants and extensions.
  - Today: Modelling and analysis of business processes.

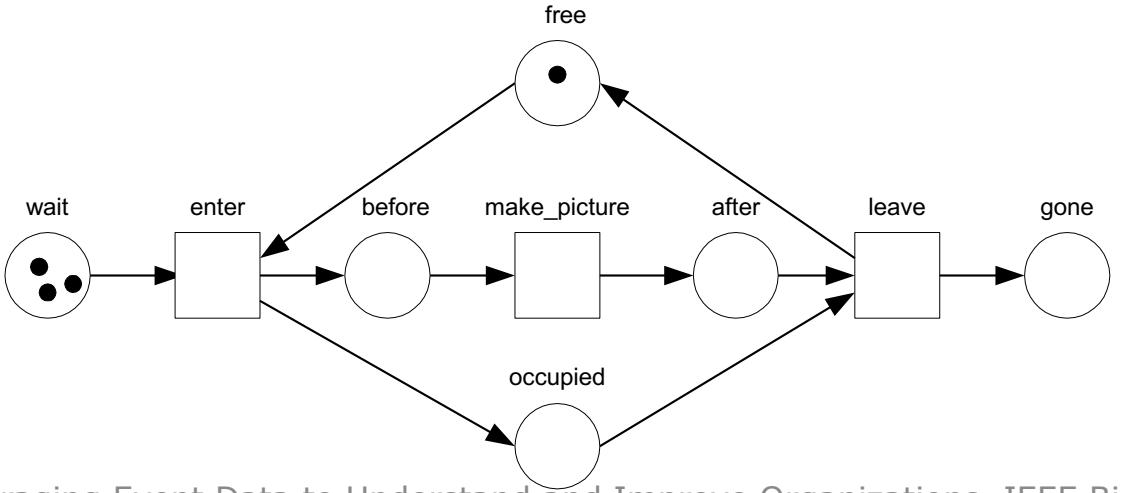
ELEMENTS


(name)
place
(name)
transition
token
arc (directed connection)


place
token
transition
t01
t12
t23
t34
p0
p3
p2
p1
p4
t10
t21
t32
t43

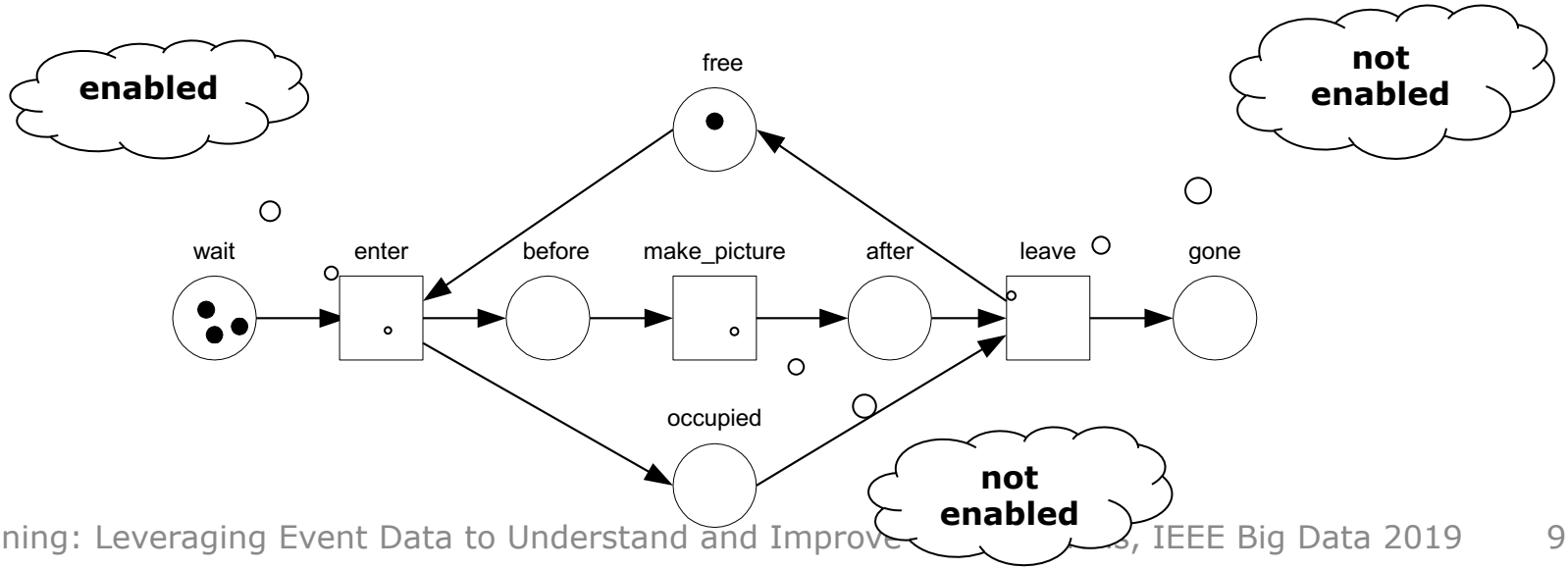
RULES

- Connections are directed.
- No connections between two places or two transitions.
- Places may hold zero or more tokens.



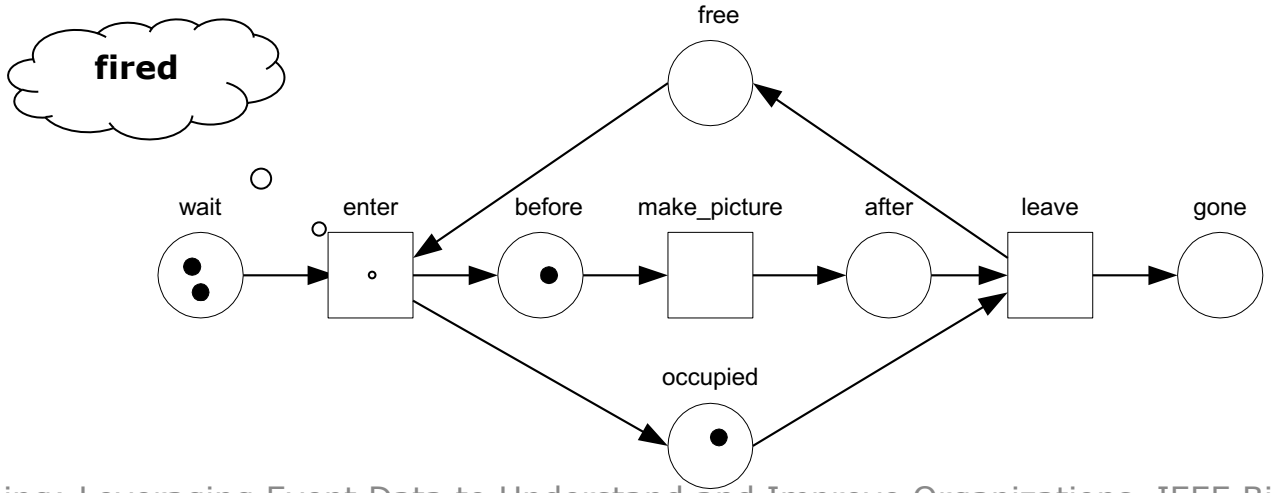
ENABLEMENT

- A transition is enabled if each of its input places contains at least one token.



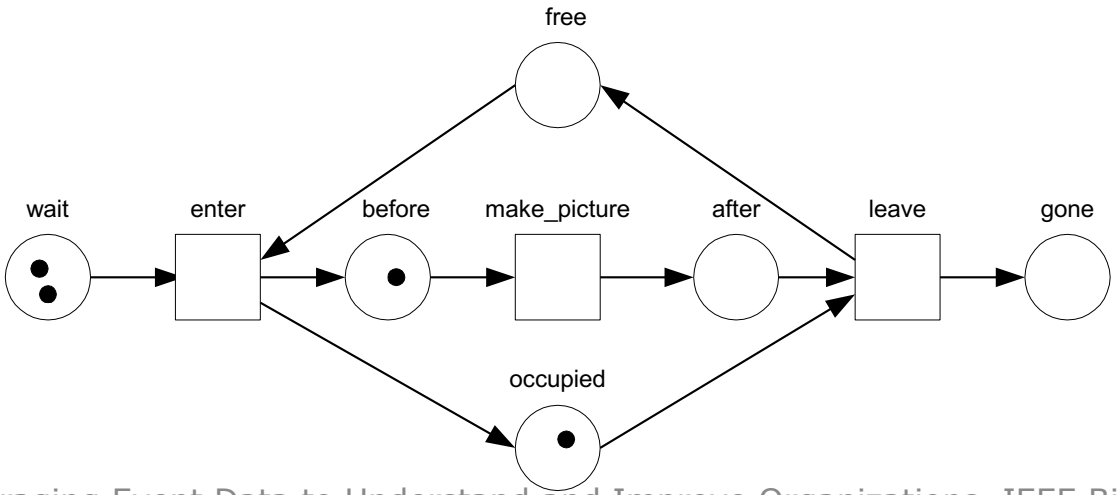
FIRING

- An enabled transition can fire (i.e., it occurs).
- When it fires it consumes a token from each input place and produces a token for each output place.



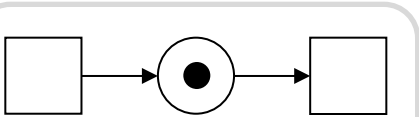
PLAY "TOKEN GAME"

- In the new state, make_picture is enabled. It will fire, etc.
- In this way, a Petri net defines how the process can be executed.

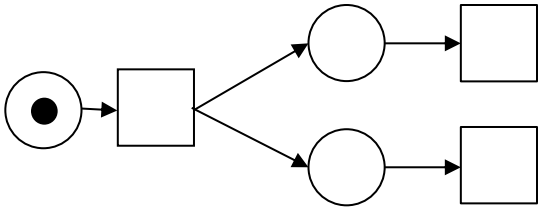


PROCESS PATTERNS

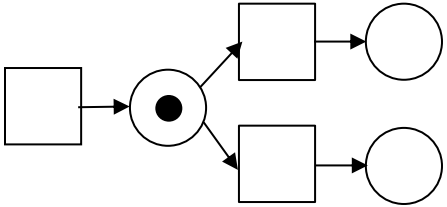
- Petri nets allow us to capture typical process patterns.



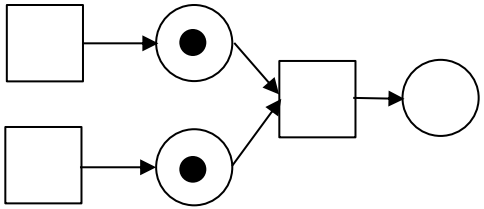
Sequence



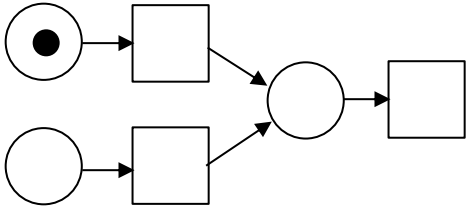
AND-Split



XOR-Split



AND-Join



XOR-Join

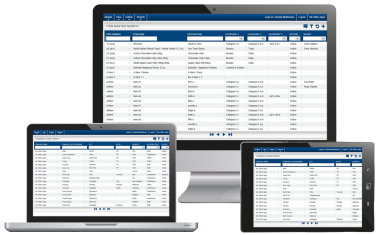
What is Process Mining?

What is Process Mining?

Business Processes

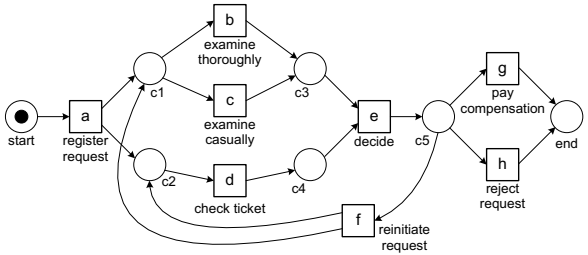
support

Software Systems



models / represents

record



Process Model

Discovery

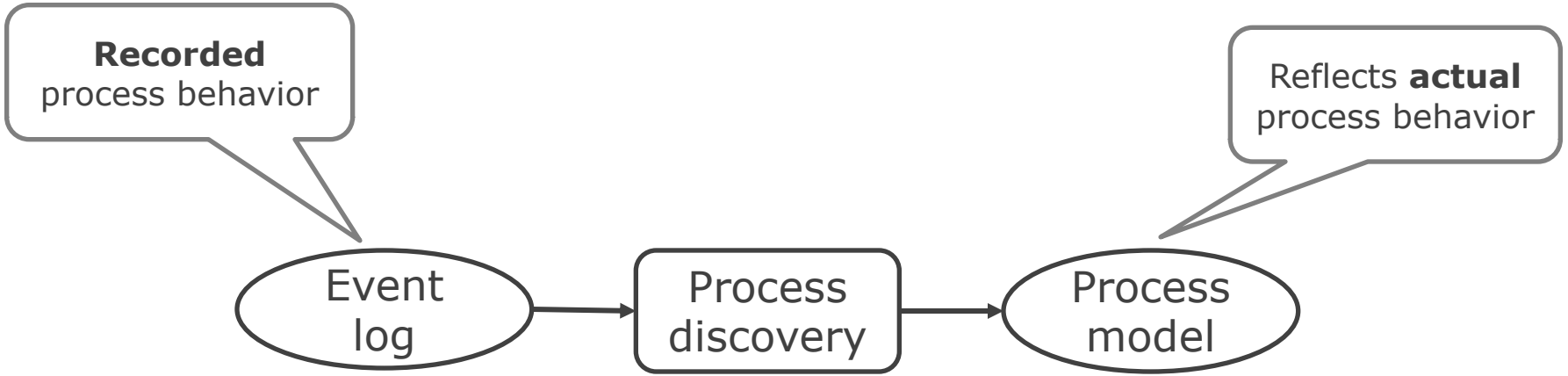
Conformance

Performance

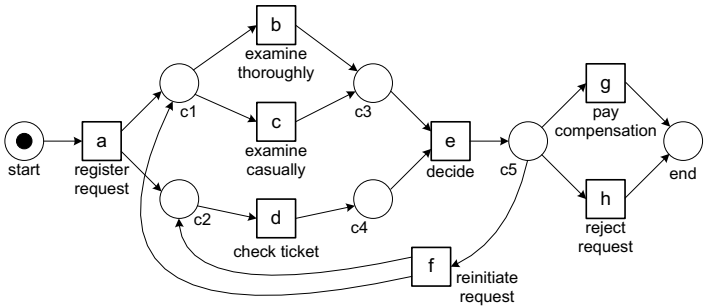
Event Log

Part 2: Process Discovery

PROCESS DISCOVERY: GOAL



Case ID	Activity ID	Activity	Timestamp
1	A	Register request	11.07.2018 – 13:53
2	A	Register request	11.07.2018 – 14:29
1	B	Examine thoroughly	11.07.2018 – 14:33
3	A	Register request	12.07.2018 – 08:45
1	D	Check ticket	12.07.2018 – 12:29
...	...	...	...



INPUT: EVENT LOG

Software Systems

record

Event Log

Case ID	Activity ID	Activity	Timestamp
1	A	Register request	11.07.2018 – 13:53
2	A	Register request	11.07.2018 – 14:29
1	B	Examine thoroughly	11.07.2018 – 14:33
3	A	Register request	12.07.2018 – 08:45
1	D	Check ticket	12.07.2018 – 12:29
2	D	Check ticket	12.07.2018 – 15:22
1	E	Decide	13.07.2018 – 15:22
3	B	Examine thoroughly	13.07.2018 – 19:18
1	G	Pay compensation	14.07.2018 – 08:15
2	C	Examine casually	14.07.2018 – 08:18
3	D	Check ticket	15.07.2018 – 07:21
3	E	Decide	15.07.2018 – 09:52
2	E	Decide	15.07.2018 – 10:01
2	G	Pay compensation	16.07.2018 – 17:12
3	H	Reject request	16.07.2018 – 18:08
…	…	…	…

From Event Log to Traces

Trace - Case 1	
a	Register request
b	Examine thoroughly
d	Check ticket
e	Decide
g	Pay compensation

Trace - Case 2	
a	Register request
d	Check ticket
c	Examine casually
e	Decide
g	Pay compensation

Trace - Case 3	
a	Register request
b	Examine thoroughly
d	Check ticket
e	Decide
h	Reject request

...

Trace - Case *N*	
a	Register request
c	Examine casually
d	Check ticket
e	Decide
f	Reinititate request
d	Check ticket
b	Examine thoroughly
e	Decide
g	Pay compensation

Abstract to sequences

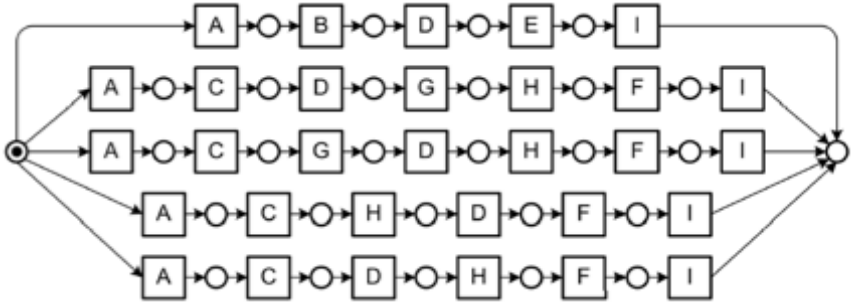
Log L = [$\langle a,b,d,e,g \rangle^{5}$, $\langle a,d,c,e,g \rangle^{10}$, $\langle a,b,d,e,h \rangle^{3}$ $\langle a,c,d,e,h \rangle^{8}$, $\langle a,c,d,e,f,d,b,e,g \rangle$]

Process Discovery: Model Quality

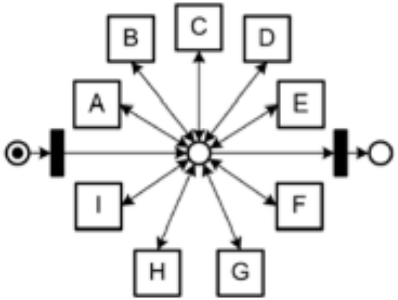
Goal is to discover a model that:

1. Allows for behavior seen in log (*fitness*)
2. Is as simple as possible (*simplicity)*
3. Balances between overfitting and underfitting (*precision* vs. *generalization*)

Sequence	Freq.
ABDEI	1207
ACDGHFI	145
ACGDHFI	56
ACGDHFI	56
ACHDFI	23
ACDHFI	28



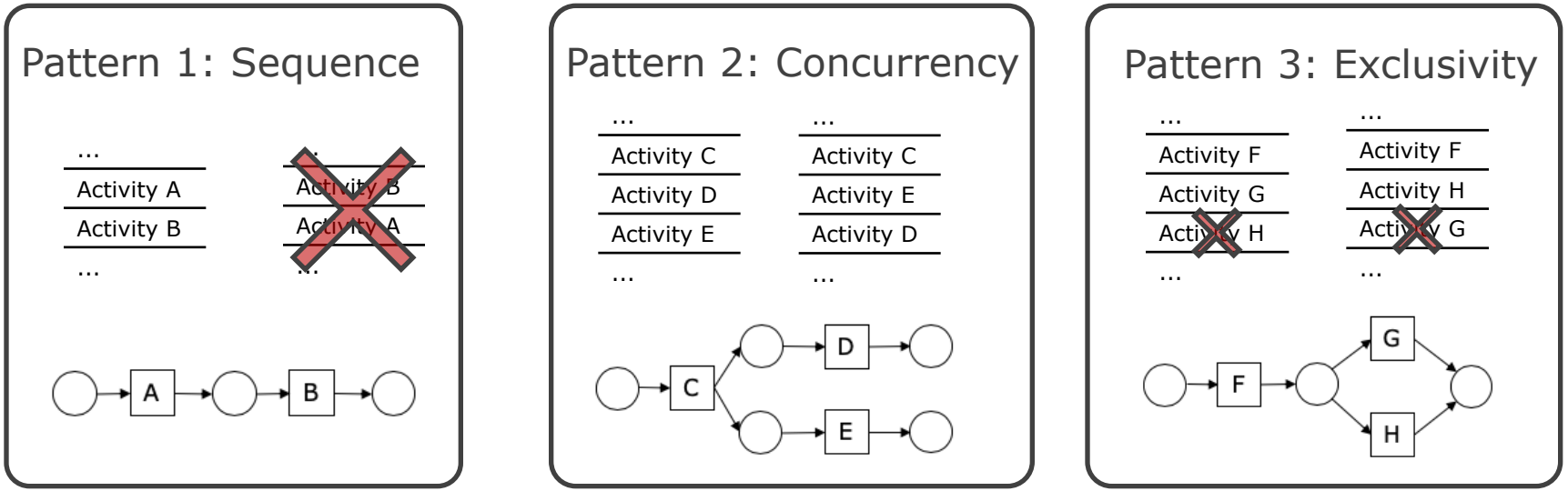
Model A: Overfitting and complex



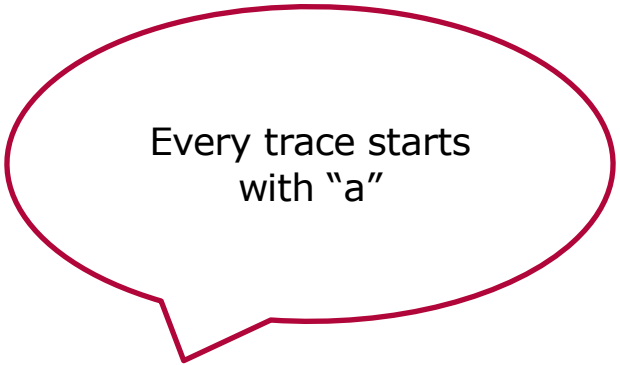
Model B: Underfitting

Key Idea of Discovery

Recognize patterns in the order relations between activities

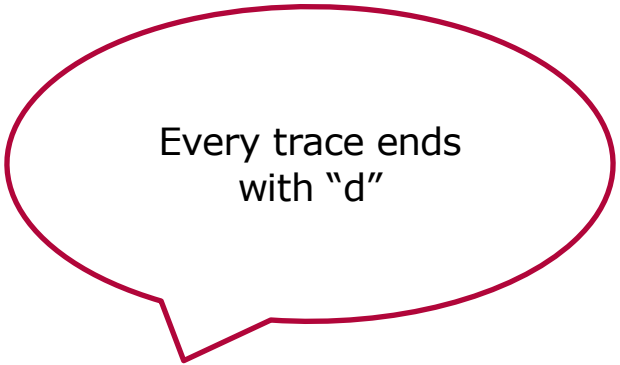


Discovery Example



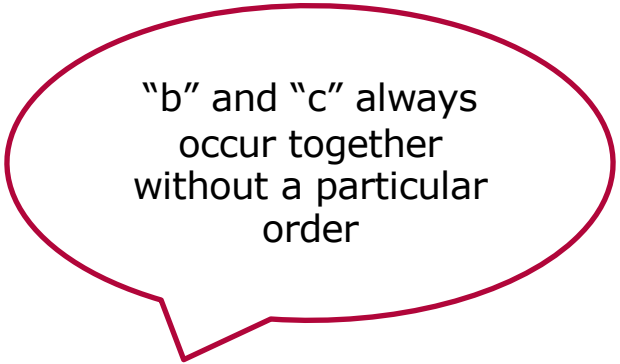
$$L_2 = [\langle a,b,c,d\rangle^3, \langle a,c,b,d\rangle^4, \langle a,b,c,e,f,b,c,d\rangle^2, \langle a,b,c,e,f,c,b,d\rangle,$$
$$\langle a,c,b,e,f,b,c,d\rangle^2, \langle a,c,b,e,f,b,c,e,f,c,b,d\rangle]$$

Discovery Example



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

Discovery Example



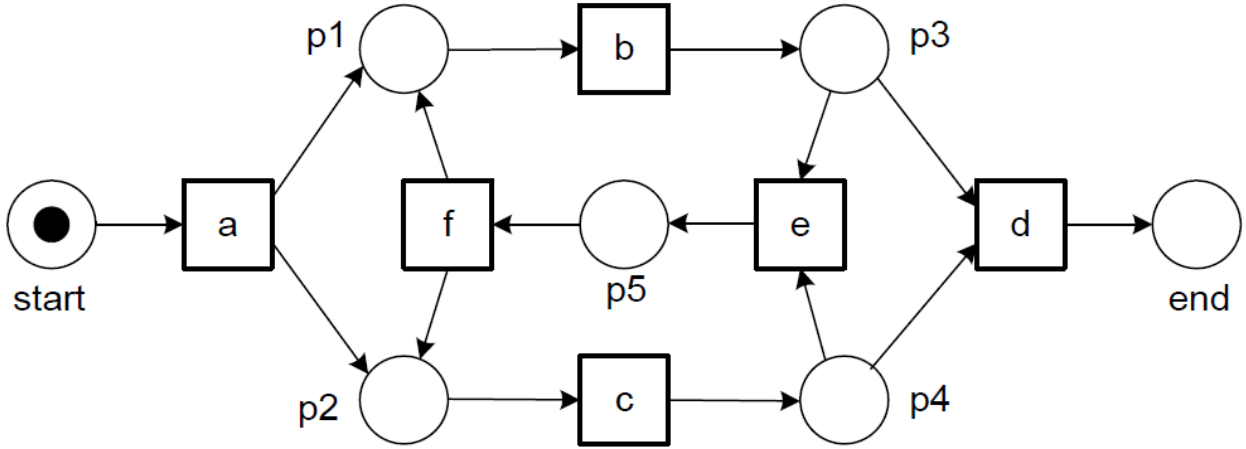
$$L_2 = [\langle a,b,c,d\rangle^3, \langle a,c,b,d\rangle^4, \langle a,b,c,e,f,b,c,d\rangle^2, \langle a,b,c,e,f,c,b,d\rangle,$$
$$\langle a,c,b,e,f,b,c,d\rangle^2, \langle a,c,b,e,f,b,c,e,f,c,b,d\rangle]$$

Discovery Example

Every “e” is followed by an “f”,
every “f” is preceded by an “e”

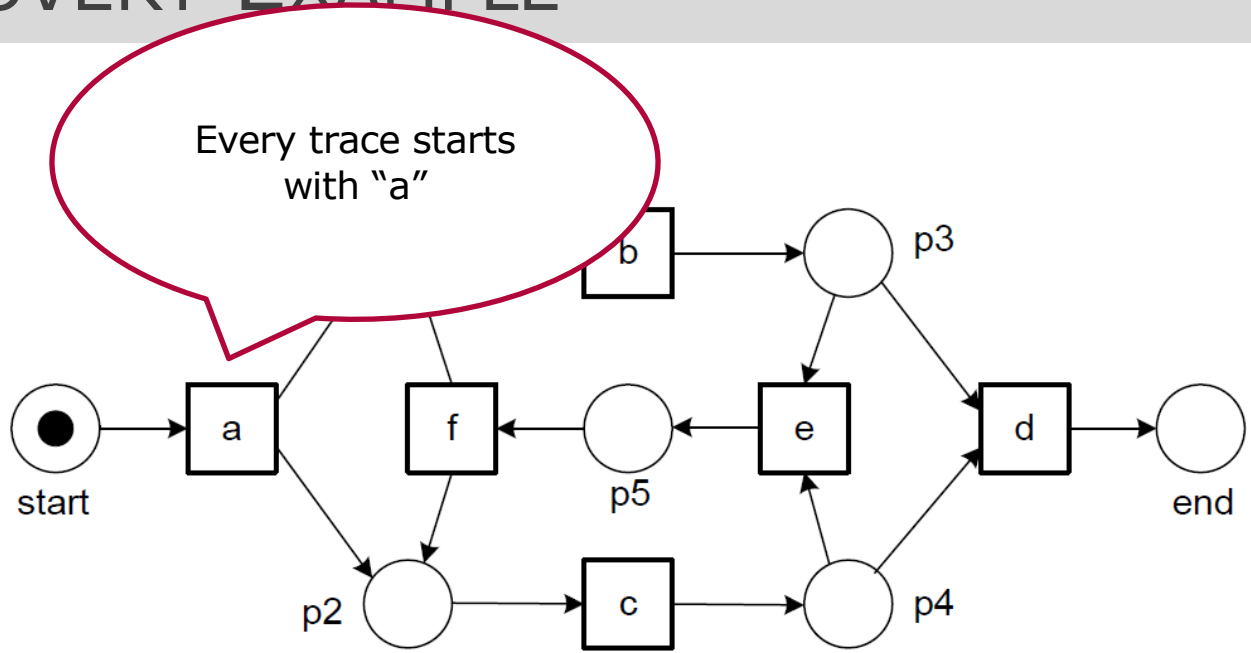
$$L_2 = [\langle a,b,c,d\rangle^3, \langle a,c,b,d\rangle^4, \langle a,b,c,e,f,b,c,d\rangle^2, \langle a,b,c,e,f,c,b,d\rangle,$$
$$\langle a,c,b,e,f,b,c,d\rangle^2, \langle a,c,b,e,f,b,c,e,f,c,b,d\rangle]$$

DISCOVERY EXAMPLE



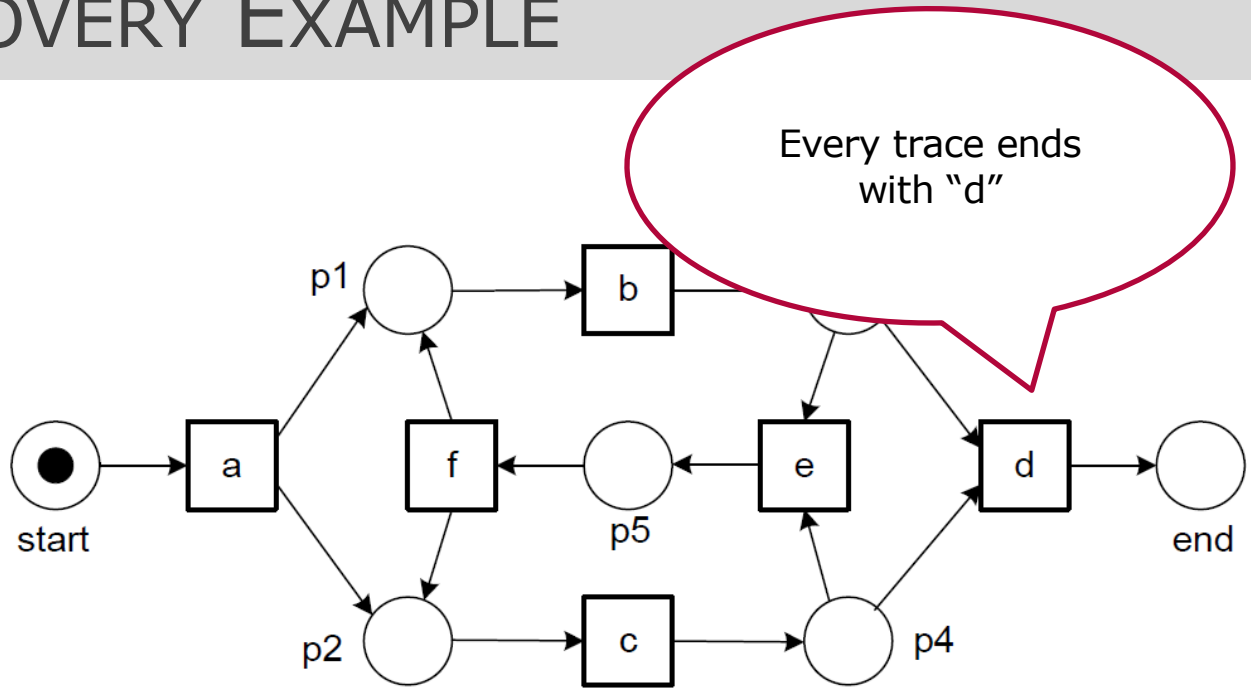
$$L_2 = [\langle a,b,c,d\rangle^3, \langle a,c,b,d\rangle^4, \langle a,b,c,e,f,b,c,d\rangle^2, \langle a,b,c,e,f,c,b,d\rangle,$$
$$\langle a,c,b,e,f,b,c,d\rangle^2, \langle a,c,b,e,f,b,c,e,f,c,b,d\rangle]$$

Discovery Example



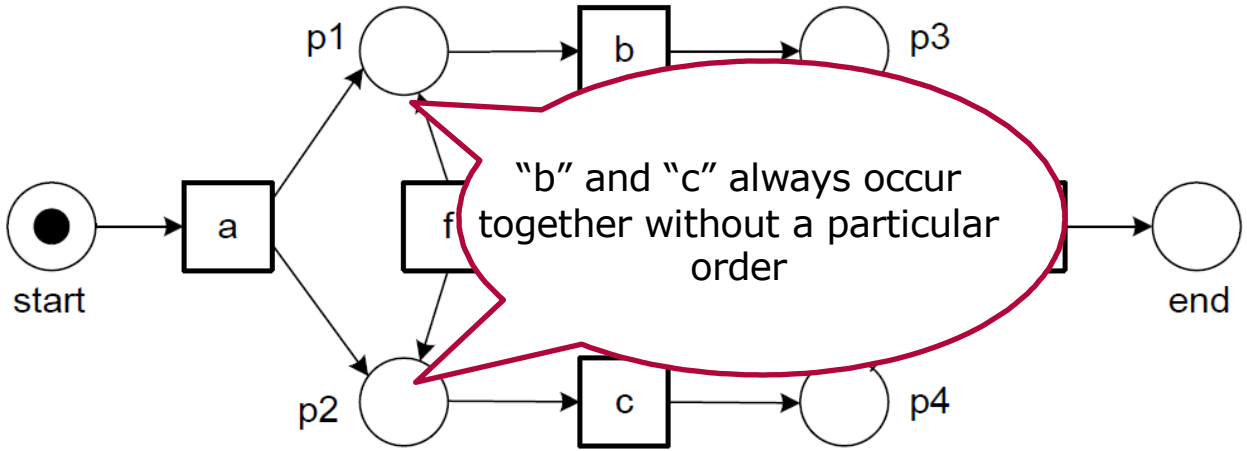
$$L_2 = [\langle a,b,c,d\rangle^3, \langle a,c,b,d\rangle^4, \langle a,b,c,e,f,b,c,d\rangle^2, \langle a,b,c,e,f,c,b,d\rangle,$$
$$\langle a,c,b,e,f,b,c,d\rangle^2, \langle a,c,b,e,f,b,c,e,f,c,b,d\rangle]$$

Discovery Example



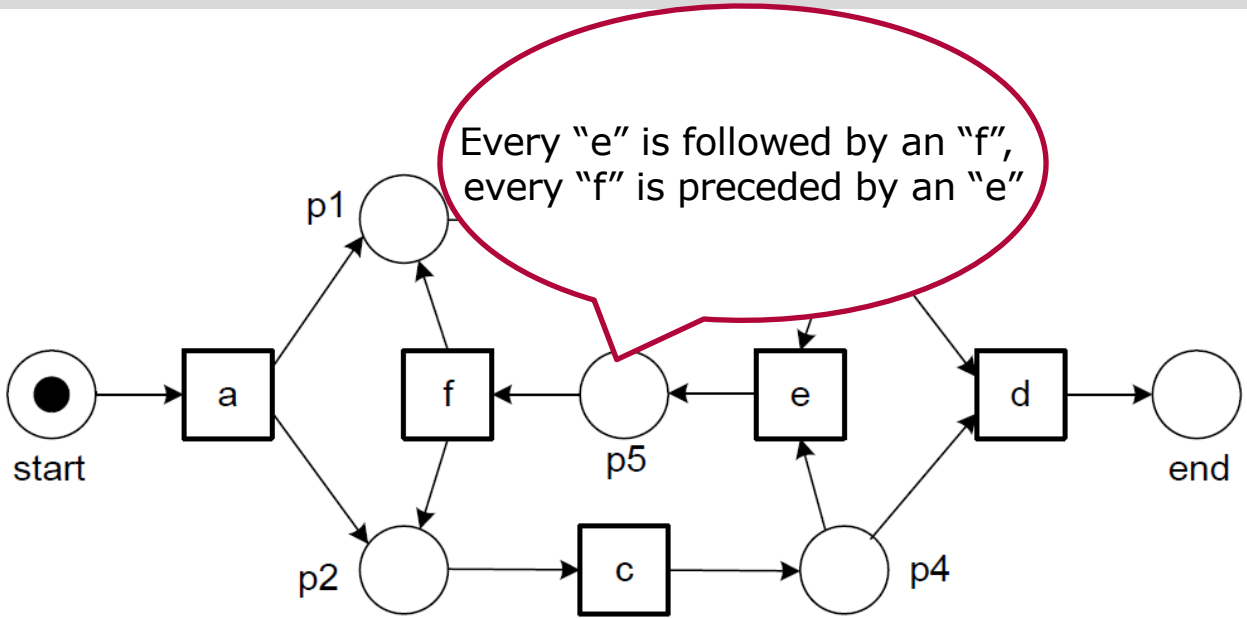
$$L_2 = [\langle a,b,c,d\rangle^3, \langle a,c,b,d\rangle^4, \langle a,b,c,e,f,b,c,d\rangle^2, \langle a,b,c,e,f,c,b,d\rangle,$$
$$\langle a,c,b,e,f,b,c,d\rangle^2, \langle a,c,b,e,f,b,c,e,f,c,b,d\rangle]$$

Discovery Example



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

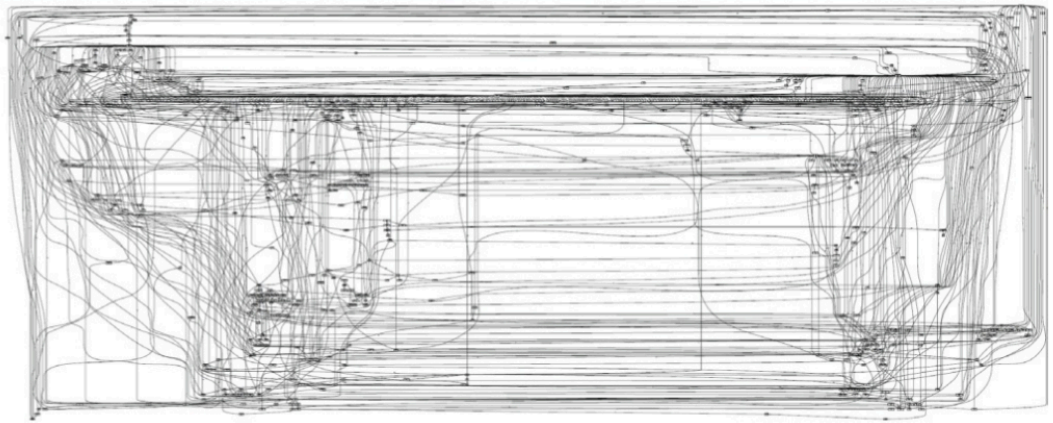
Discovery Example



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

Discovery Challenges

- Data quality issues
- Balancing over- and underfitting
- Spaghetti models



Discovery Algorithms

Key differences:

- Representation format
- How to tackle challenges
- Unique pros and cons
- No silver bullet

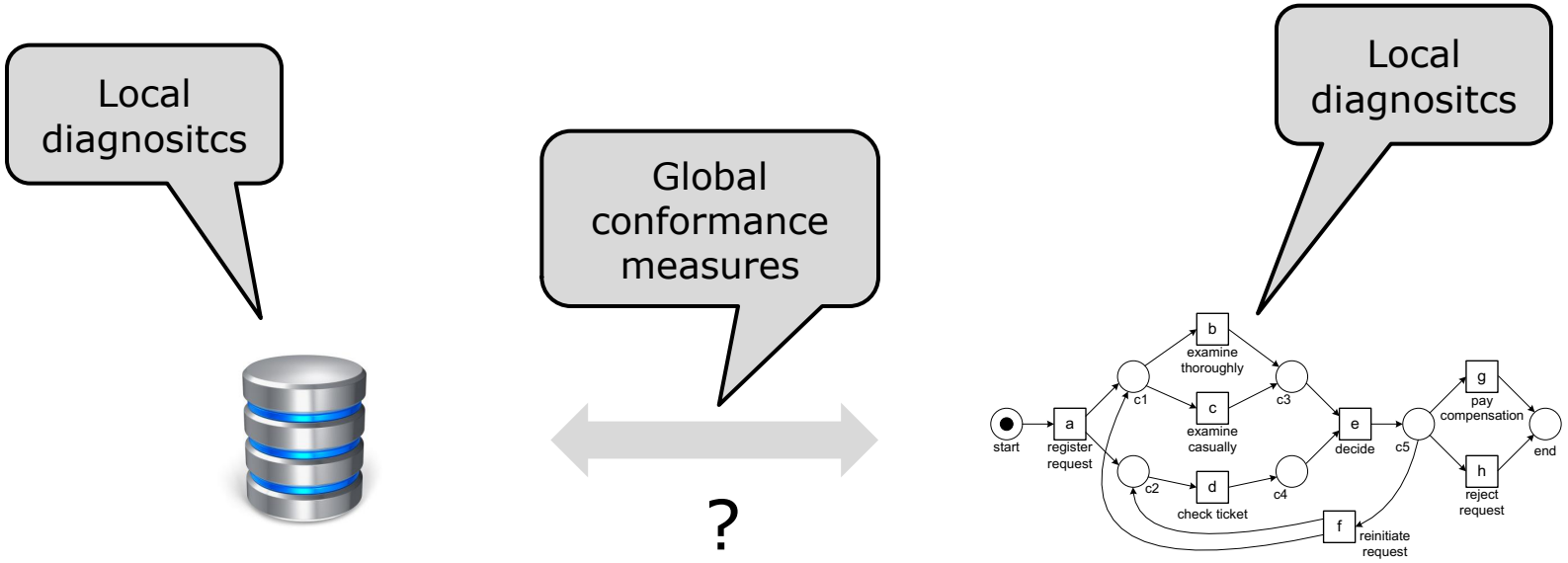
Part 3: Conformance Checking

MOTIVATION

Various drivers for conformance assessment

- Corporate governance, risk, compliance, and legislation such as the Sarbanes-Oxley (US), Basel II/III (EU)
- ISO 9001:2008 requires organizations to model their operational processes
- Business alignment: make sure that the information systems and the real business processes are well aligned

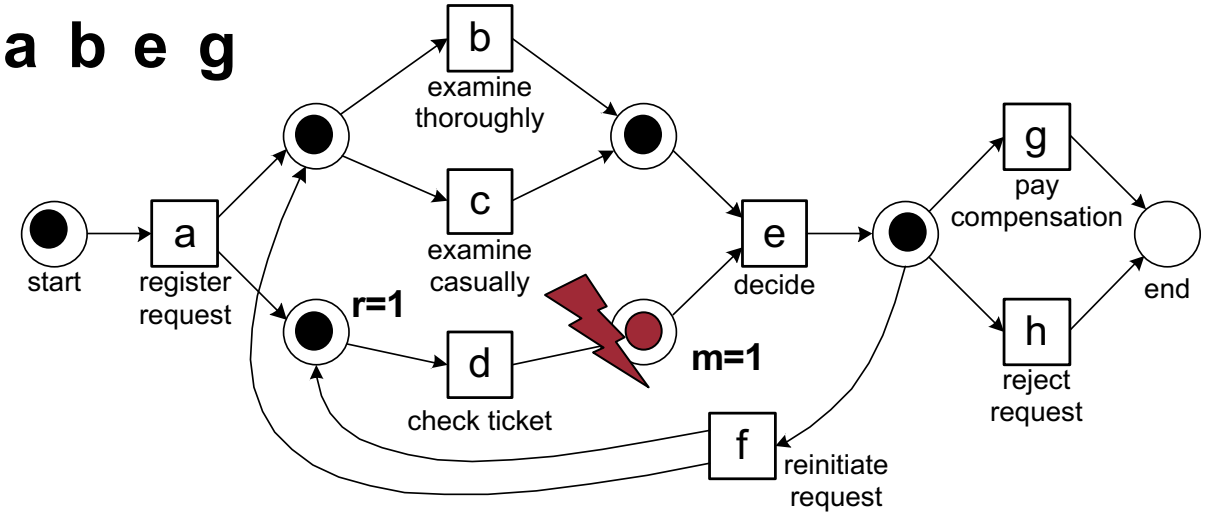
BASIC IDEA OF CONFORMANCE CHECKING



Event Log

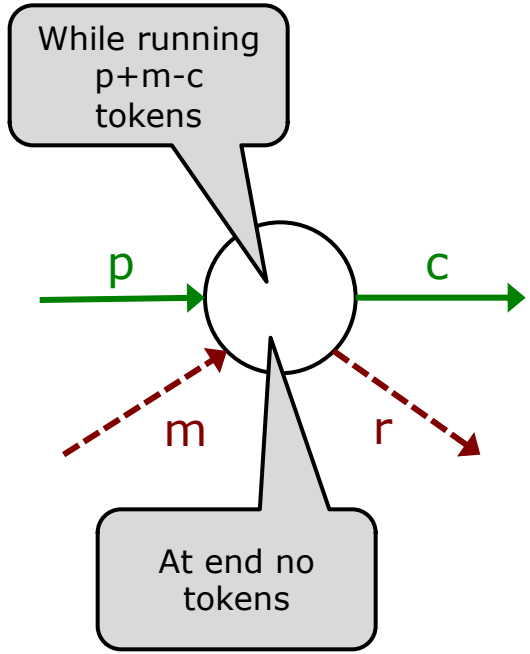
Process Model

REPLAYING TRACE "ABEG"



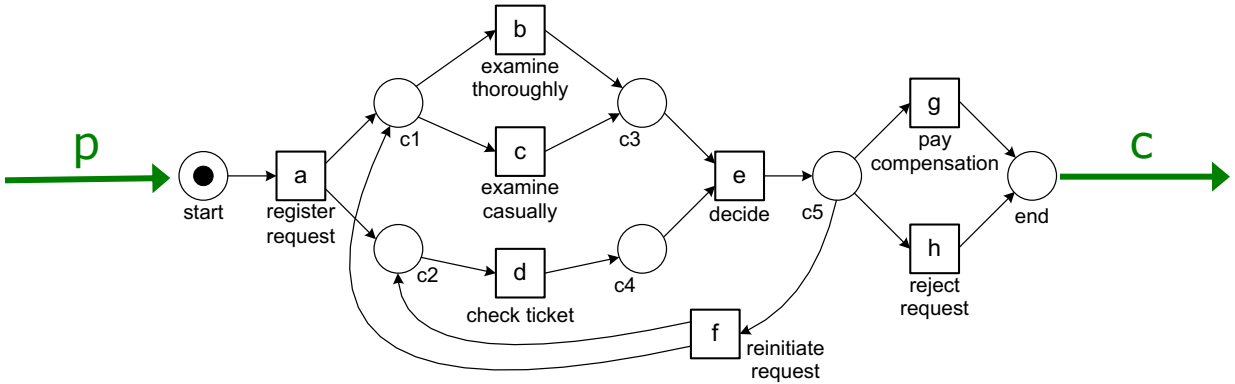
APPROACH

- Use four counters:
  - p = produced tokens
  - c = consumed tokens
  - m = missing tokens (consumed while not there)
  - r = remaining tokens (produced but not consumed)

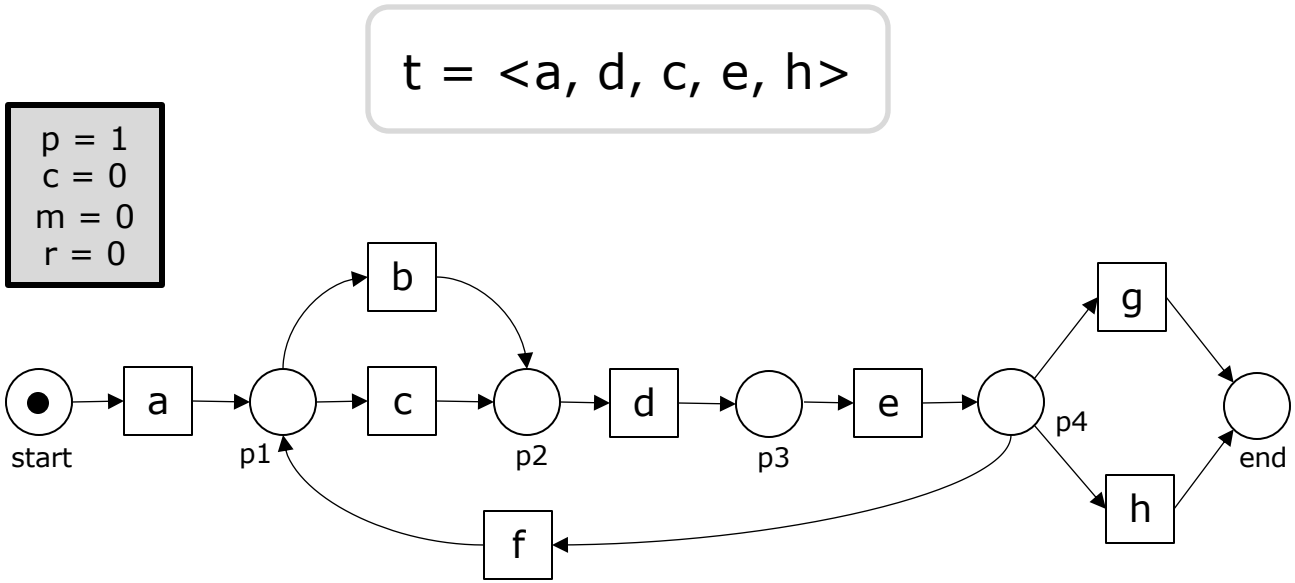


APPROACH (CONT.)

- Initialization and finalization:
  - In the beginning a token is produced for the source place: p = 1
  - At the end a token is consumed from the sink place (also if not there): c’ = c =1.



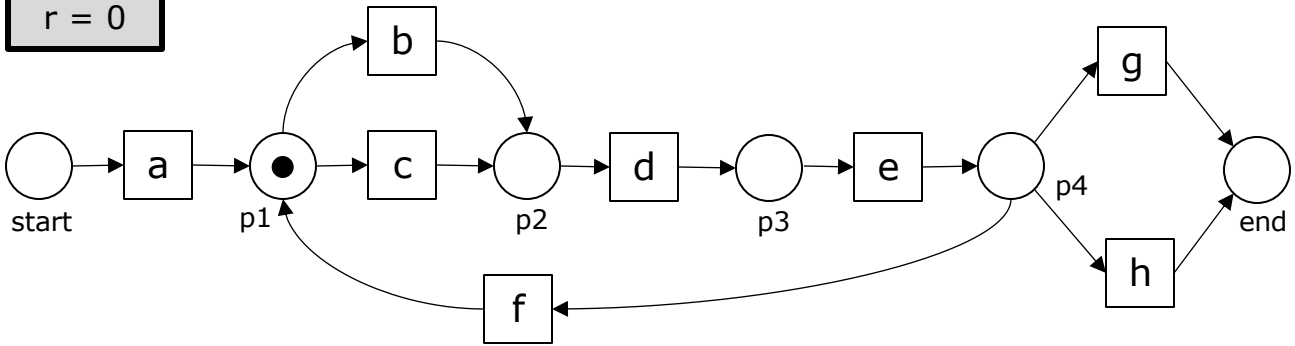
REPLAY EXAMPLE



REPLAY EXAMPLE (CONT.)

t = <a, d, c, e, h>

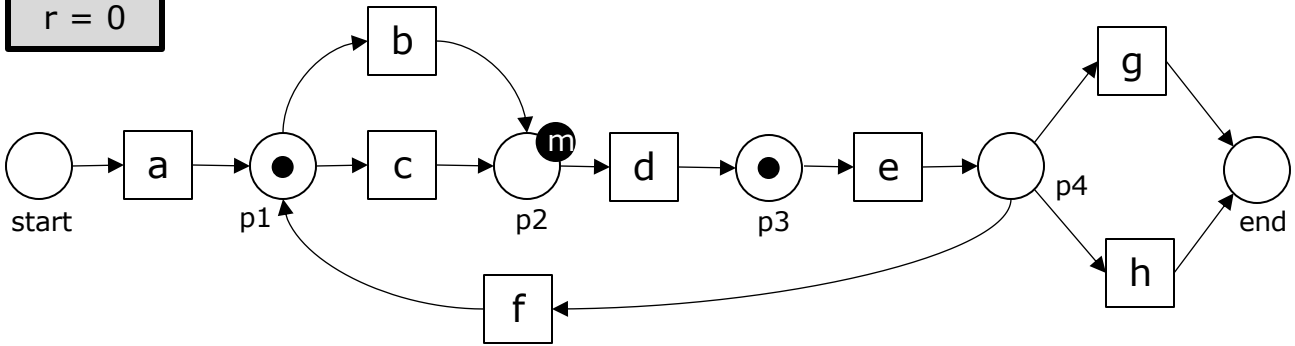
p = 2
c = 1
m = 0
r = 0



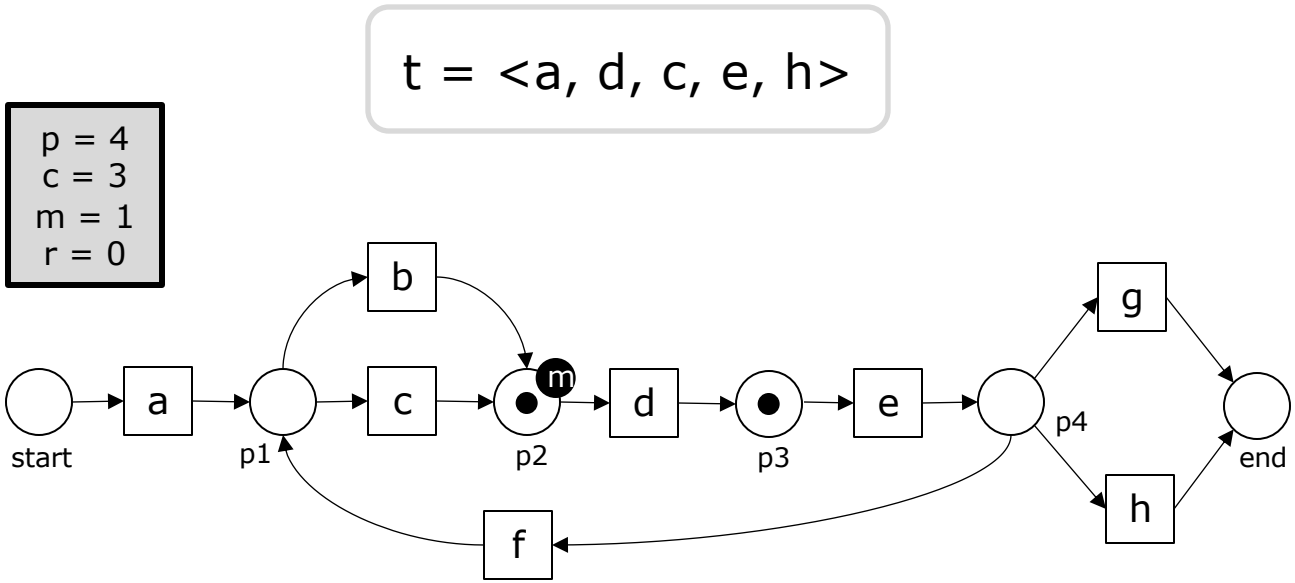
REPLAY EXAMPLE (CONT.)

t = <a, d, c, e, h>

p = 3
c = 2
m = 1
r = 0



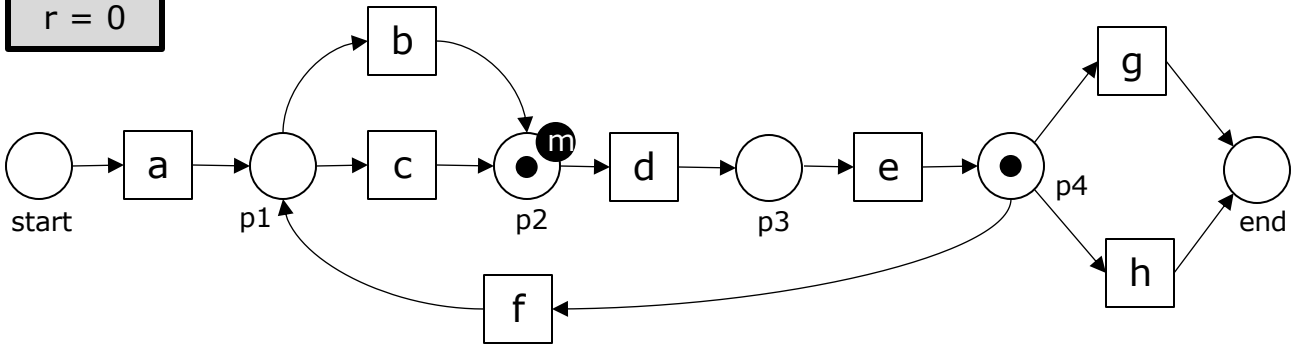
REPLAY EXAMPLE (CONT.)



REPLAY EXAMPLE (CONT.)

t = <a, d, c, e, h>

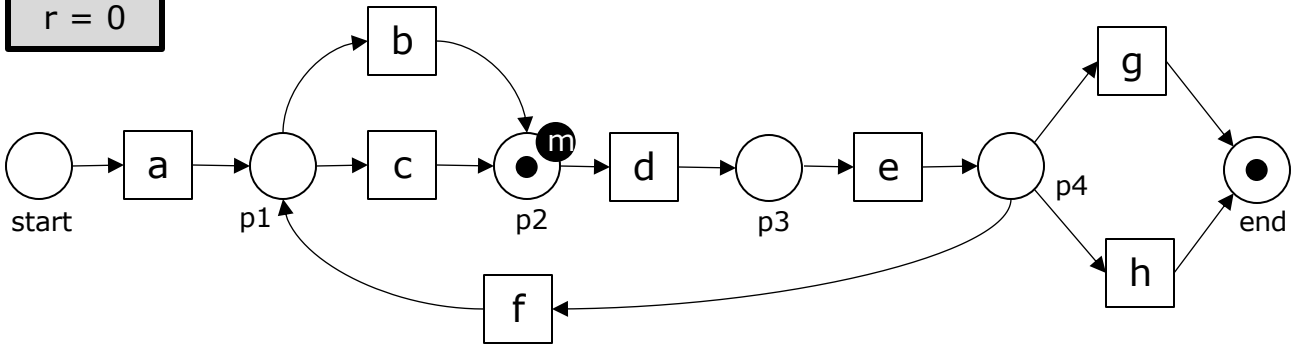
p = 5
c = 4
m = 1
r = 0



REPLAY EXAMPLE (CONT.)

t = <a, d, c, e, h>

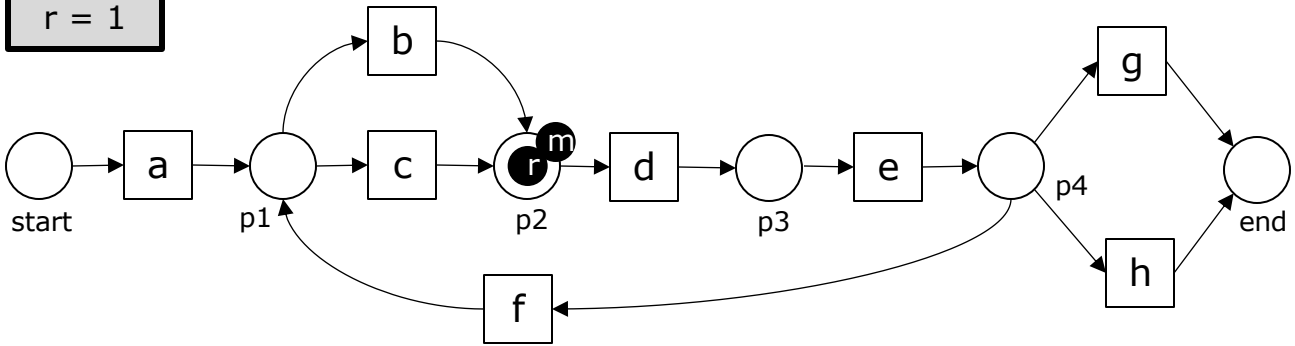
p = 6
c = 5
m = 1
r = 0



REPLAY EXAMPLE (CONT.)

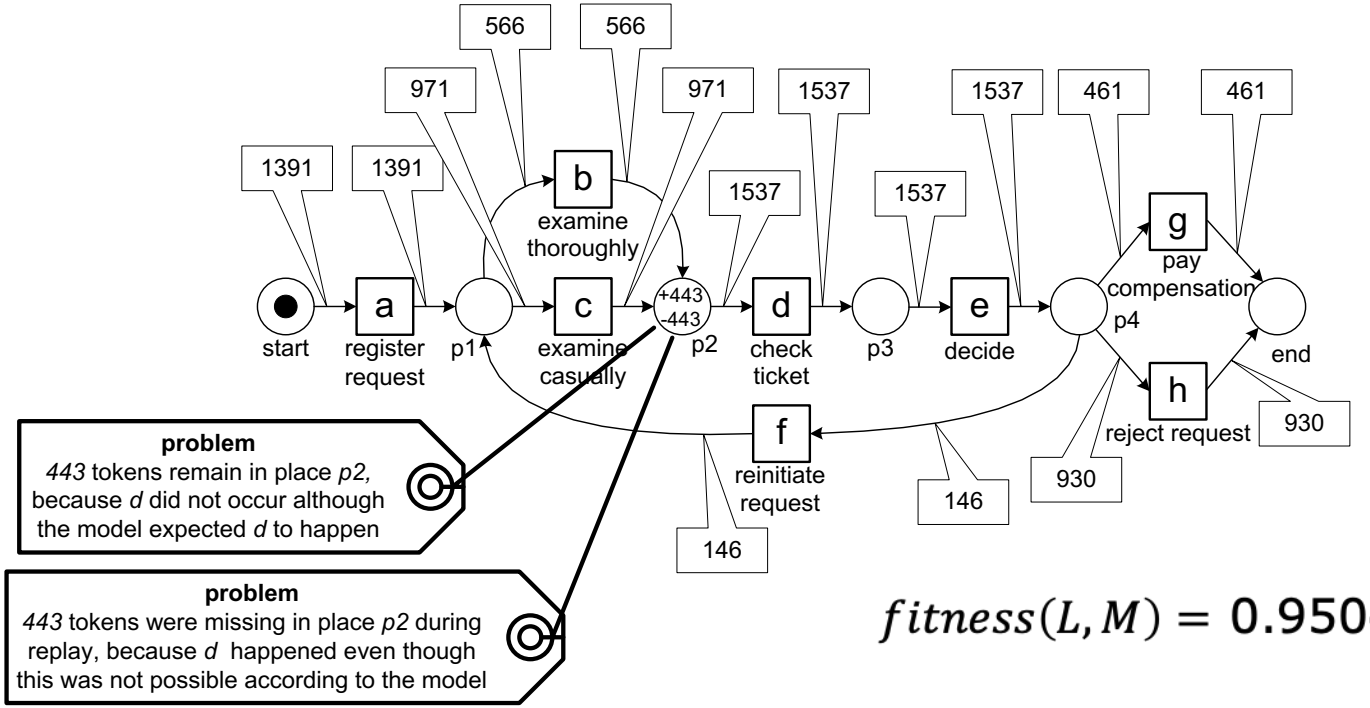
t = <a, d, c, e, h>

p = 6
c = 6
m = 1
r = 1



$$fitness(t, M) = \frac{1}{2}\left(1 - \frac{1}{6}\right) + \frac{1}{2}\left(1 - \frac{1}{6}\right) = 0.8333$$

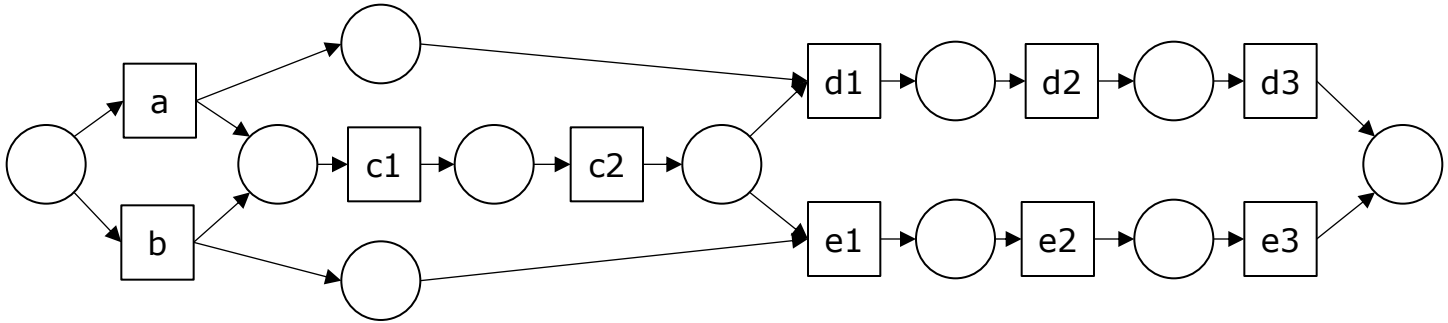
DIAGNOSTICS



BEYOND REPLAY

- Main limitation of token replay is that it does not take a wholistic perspective.
- Token replay is the basis for state-of-the-art alignment techniques.

t = <a,c1,c2,e1,e2,e3>



Part 4: Demo & Outlook

Process Mining Tools

Open source

- ProM6 (Java): www.promtools.org
- PM4Py (Python) www.pm4py.org

- Many state-of-the-art algorithms
- Stand-alone, plugin architecture

Commercial

celonis

Disco by Fluxicon

SIGNAVIO

LANA LABS

- User friendly
- Integrated analyses
- High performance
- Academic licenses available

Demo

OUTLOOK

- Beyond control-flow analysis

- Dealing with data quality issues

- Privacy-aware process mining

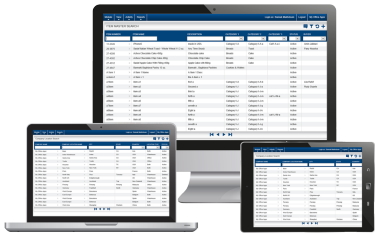
- From reactive to predictive analysis

CONCLUSION

Business Processes

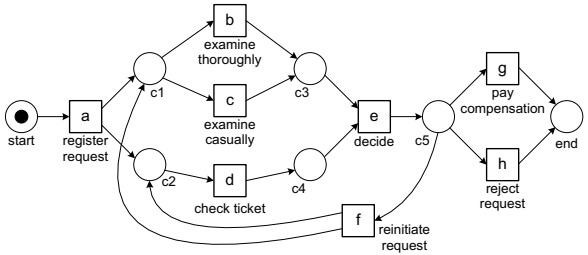
support

Software Systems



models / represents

record



Process Model

Discovery

Conformance

Performance

Event Log

Prof. Dr. Henrik Leopold

henrik.leopold@the-klu.org
www.henrikleopold.com

Dr. Han van der Aa

han.van.der.aa@hu-berlin.de
www.hanvanderaa.com