Process Oriented Data Science

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Campus d'Excel·lència Internacional

Josep Carmona
Computer Science Department

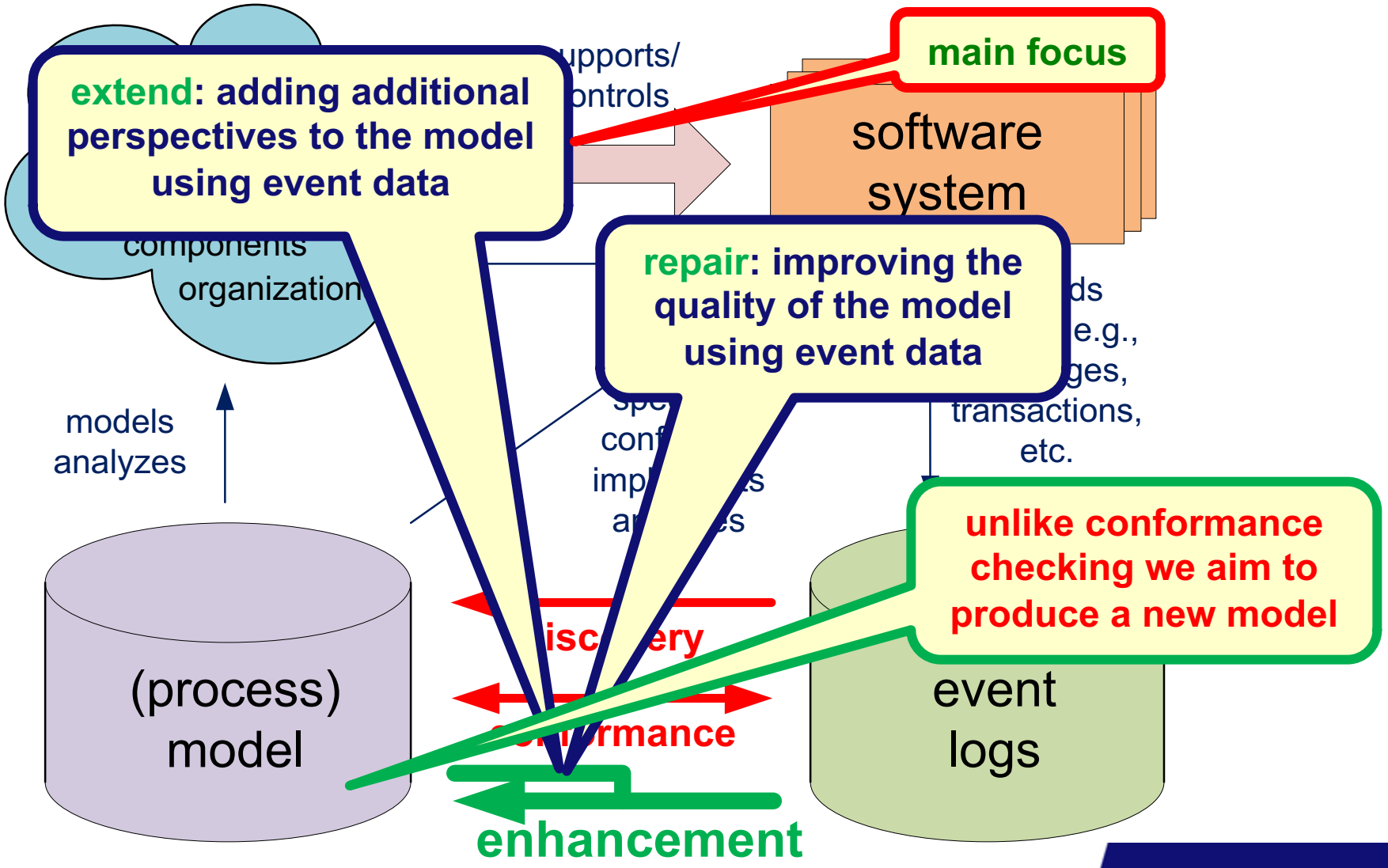
Outline

- M1: Process Mining Overview, Positioning & Preliminaries (Event data & Process Models)
- M2: Process Discovery
- M3: Conformance Checking
- **M4: Process Enhancement**

Disclaimer

- Most of the material of this course is taken from my colleagues:
  - **RWTH Aachen (Prof. Wil van der Aalst)**
  - Humboldt University zu Berlin (Prof. Matthias Weidlich)
  - Technishe Universiteit Eindhoven (Prof. Boudewijn van Dongen)
  - University of Tartu (Prof. Marlon Dumas)
  - University of Melbourne (Prof. Marcello La Rosa)
  - Technical University of Denmark (Prof. Andrea Burattin)
- Hence, this material is only provided for your learning, please do not share nor publish

Enhancement: Extension and Repair



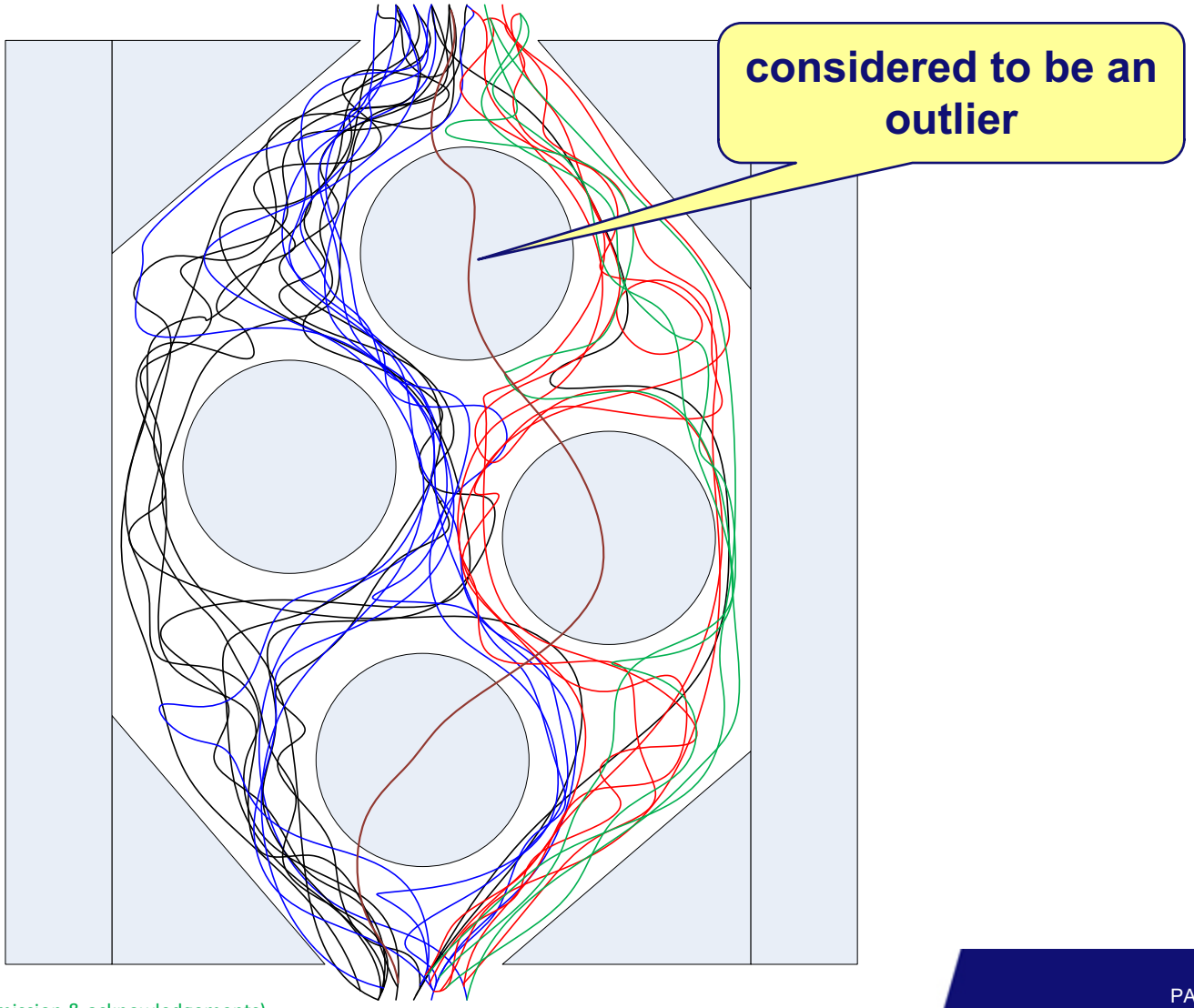
©Wil van der Aalst TU/e (use only with permission & acknowledgements)

repair
"paving the cow paths"

Repairing a model based on event data

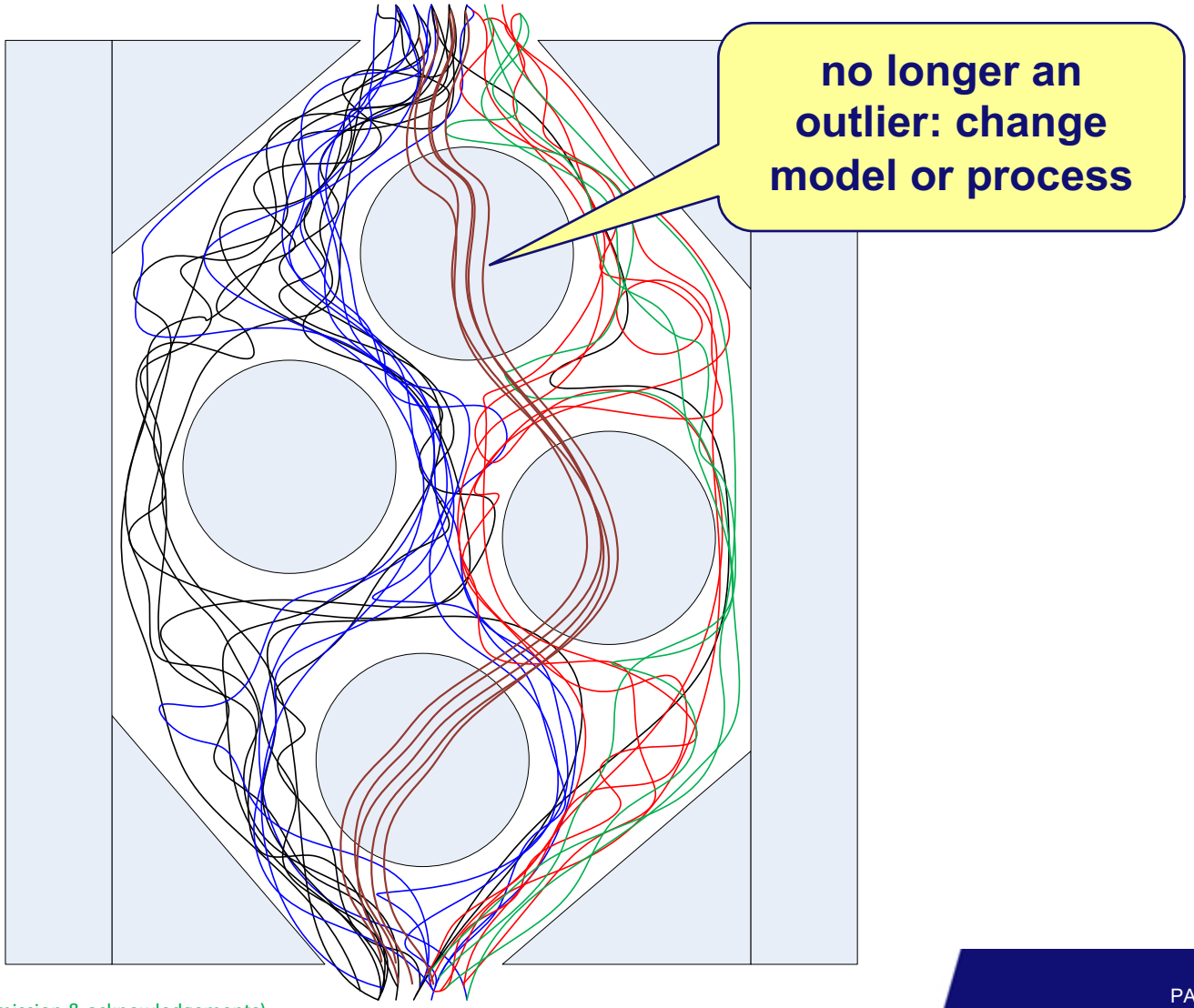
repair = pave the cow paths

No need to repair



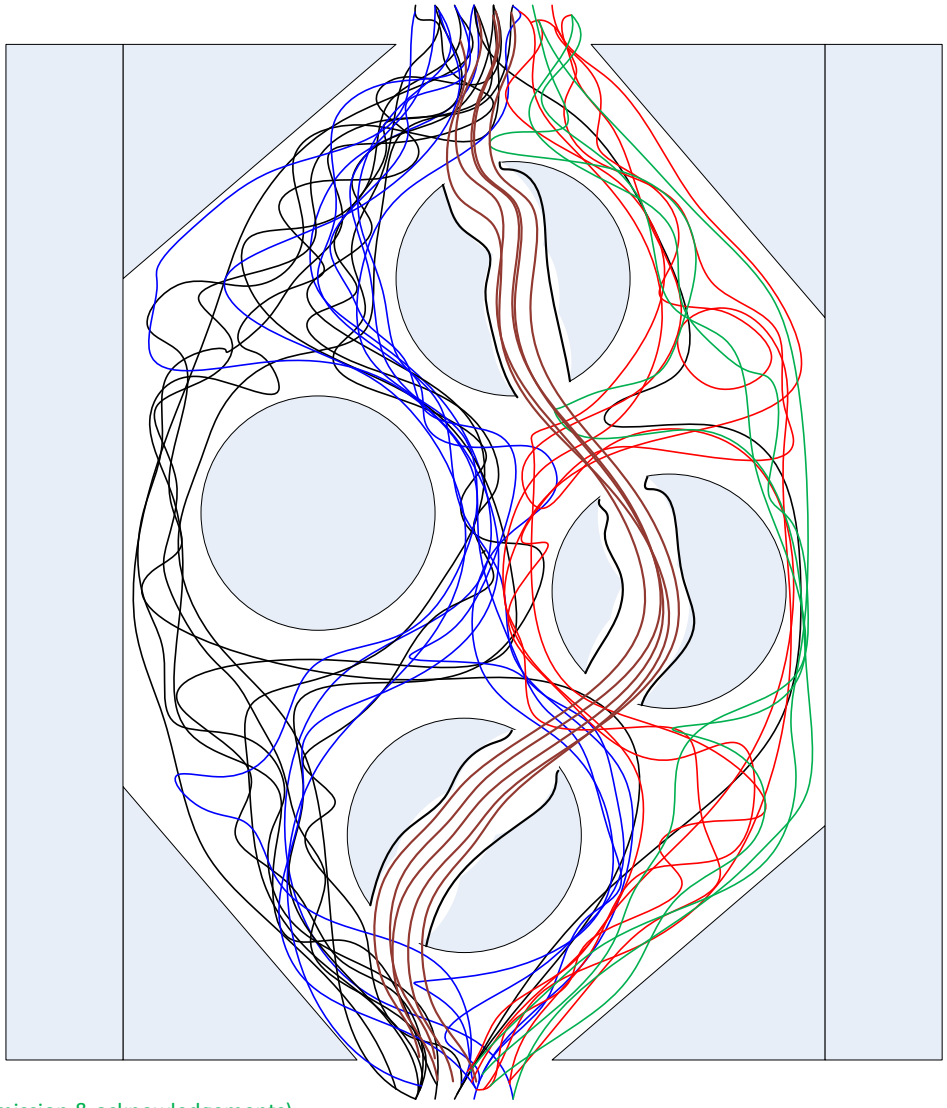
©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Repair model or influence reality ???



©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Repaired model

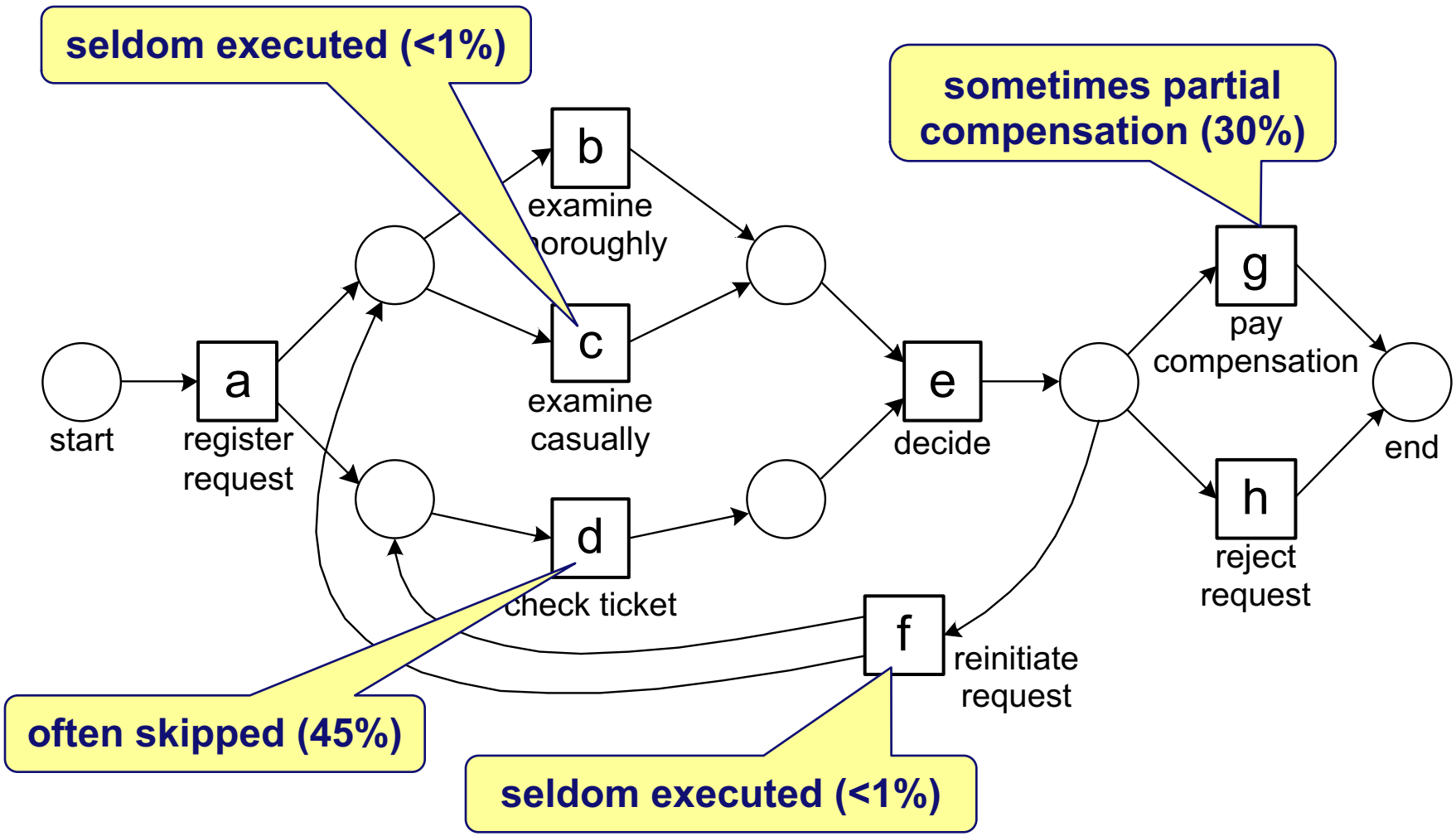


©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Metaphor: Improving passageways
(what are the minimal changes in m^2 asphalt)

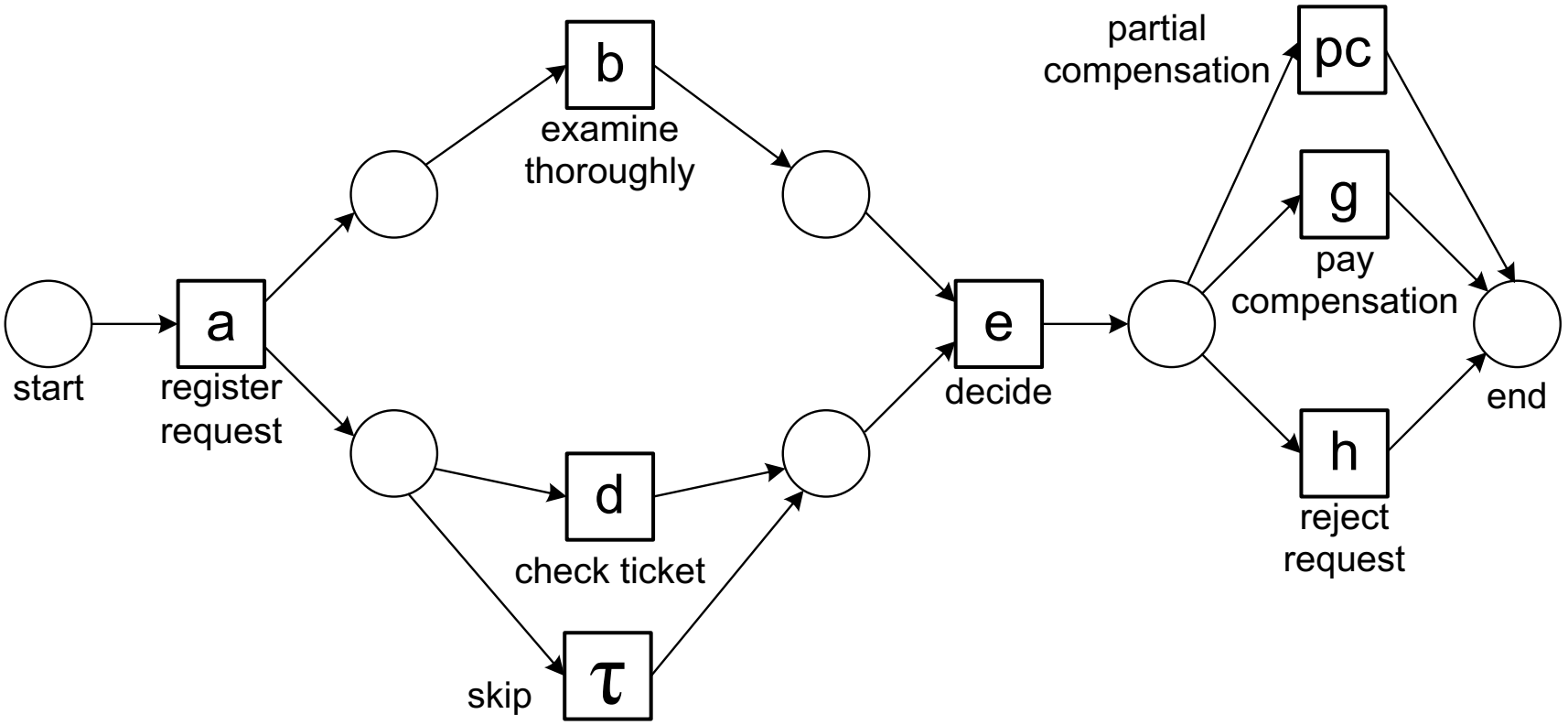
©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Assume the following replay results



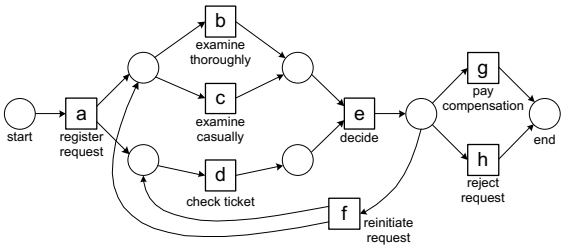
©Wil van der Aalst TU/e (use only with permission & acknowledgements)

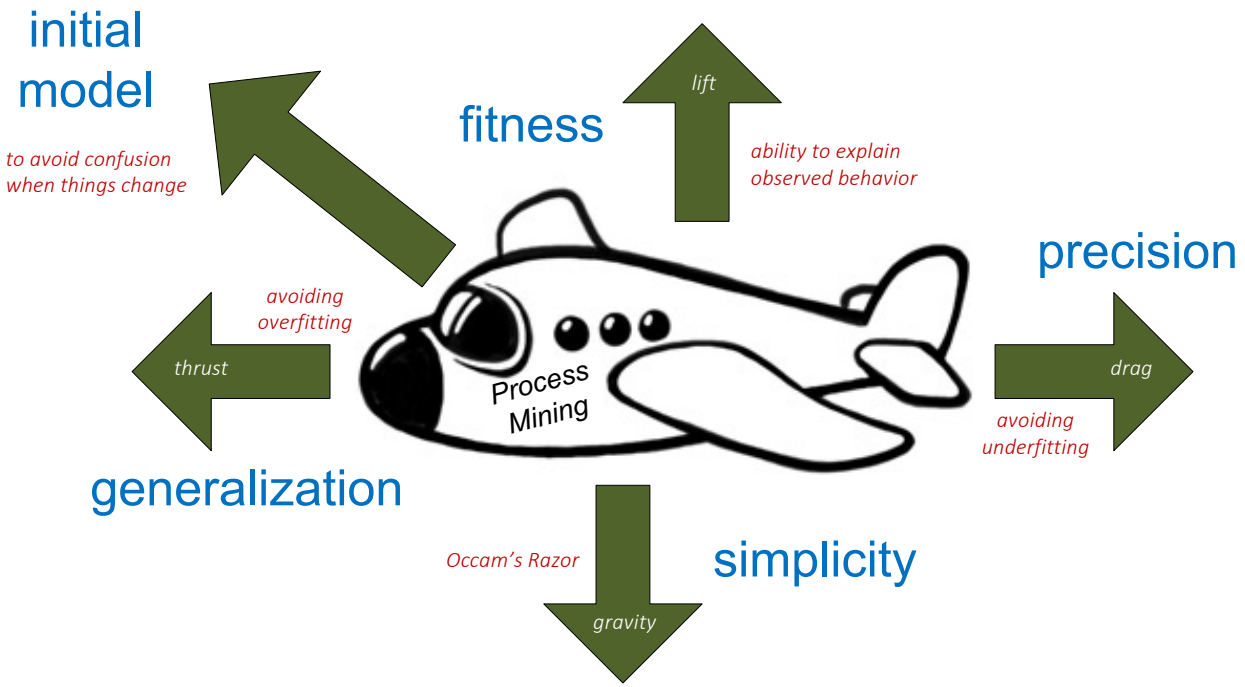
Repaired model



©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Original model is like a 5th force

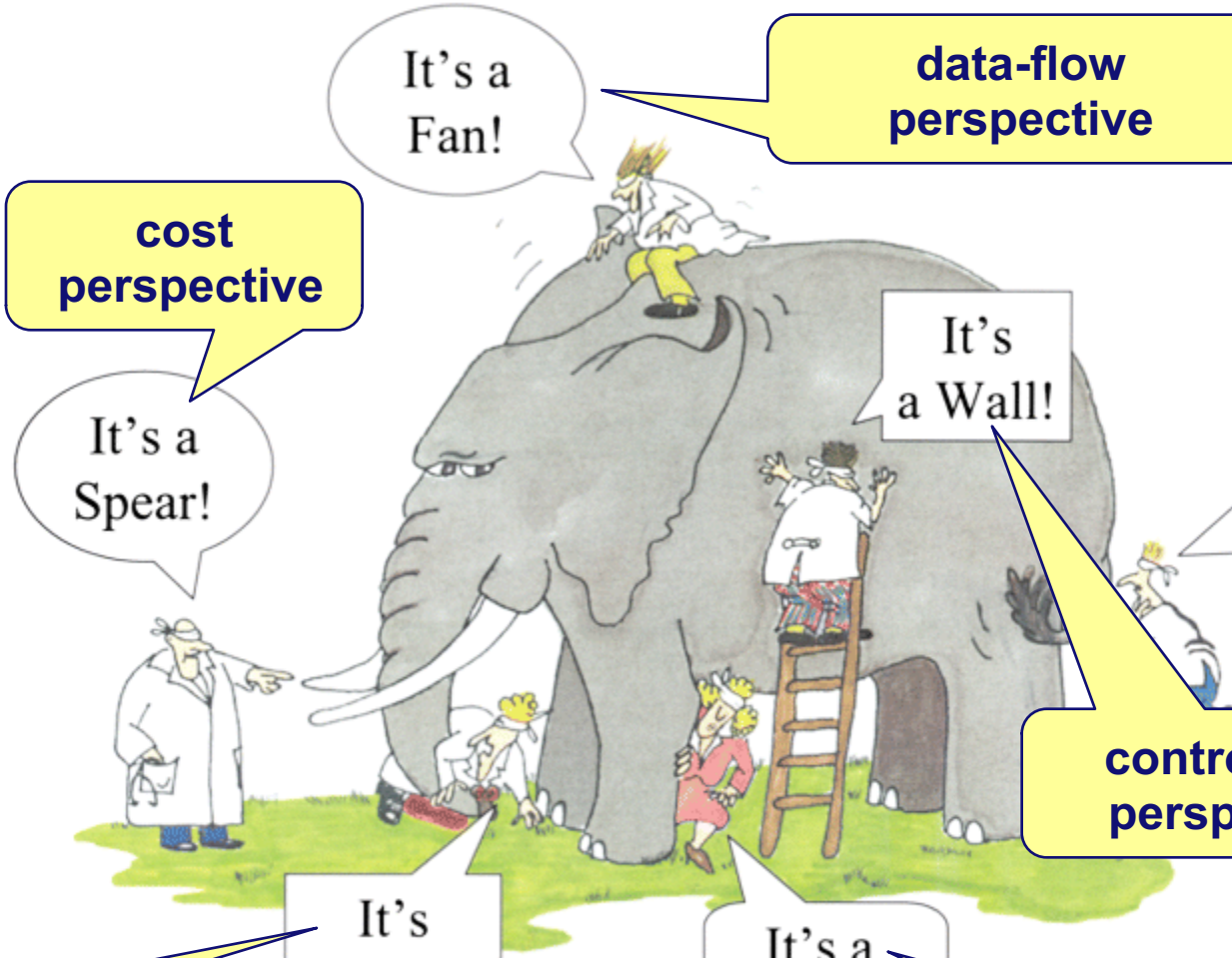




©Wil van der Aalst TU/e (use only with permission & acknowledgements)

extending process models with additional perspectives

Process are not just about control-flow!



data-flow perspective

time perspective

cost perspective

control-flow perspective

energy perspective

resource perspective

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Attributes in event logs

case id	event id	properties				
		time	activity	trans	resource	cost
	35654423	30-12-2010:11.02	register request	start	Pete	
	35654424	30-12-2010:11.08	register request	complete	Pete	50
	35654425	31-12-2010:10.06	examine thoroughly	start	Sue	
	35654427	31-12-2010:10.08	check ticket	start	Mike	
1	35654428	31-12-2010:10.12	examine thoroughly	complete	Sue	400
	35654429	31-12-2010:10.20	check ticket	complete	Mike	100
	35654430	06-01-2011:11.18	decide	start	Sara	
	35654431	06-01-2011:11.22	decide	complete	Sara	200
	35654432	07-01-2011:14.24	reject request	start	Pete	
	35654433	07-01-2011:14.32	reject request	complete	Pete	200
	35654483	30-12-2010:11.32	register request	start	Mike	
	35654484	30-12-2010:11.40	register request	complete	Mike	50
	35654485	30-12-2010:12.12	check ticket	start	Mike	
	35654486	30-12-2010:12.24	check ticket	complete	Mike	100
2	35654487	30-12-2010:14.16	examine casually	start	Pete	
	35654488	30-12-2010:14.22	examine casually	complete	Pete	400
	35654489	05-01-2011:11.22	decide	start	Sara	
	35654490	05-01-2011:11.29	decide	complete	Sara	200
	35654491	08-01-2011:12.05	pay compensation	start	Ellen	
	35654492	08-01-2011:12.15	pay compensation	complete	Ellen	200
...	...	...	...	...	...	...

As discussed before:

- **A process consists of cases.**
- **A case consists of events such that each event relates to precisely one case.**
- **Events within a case are ordered.**
- **Events can have attributes.**
- **Examples of typical attribute names are activity, time, costs, and resource.**

XES
Extensible Event Stream

©W

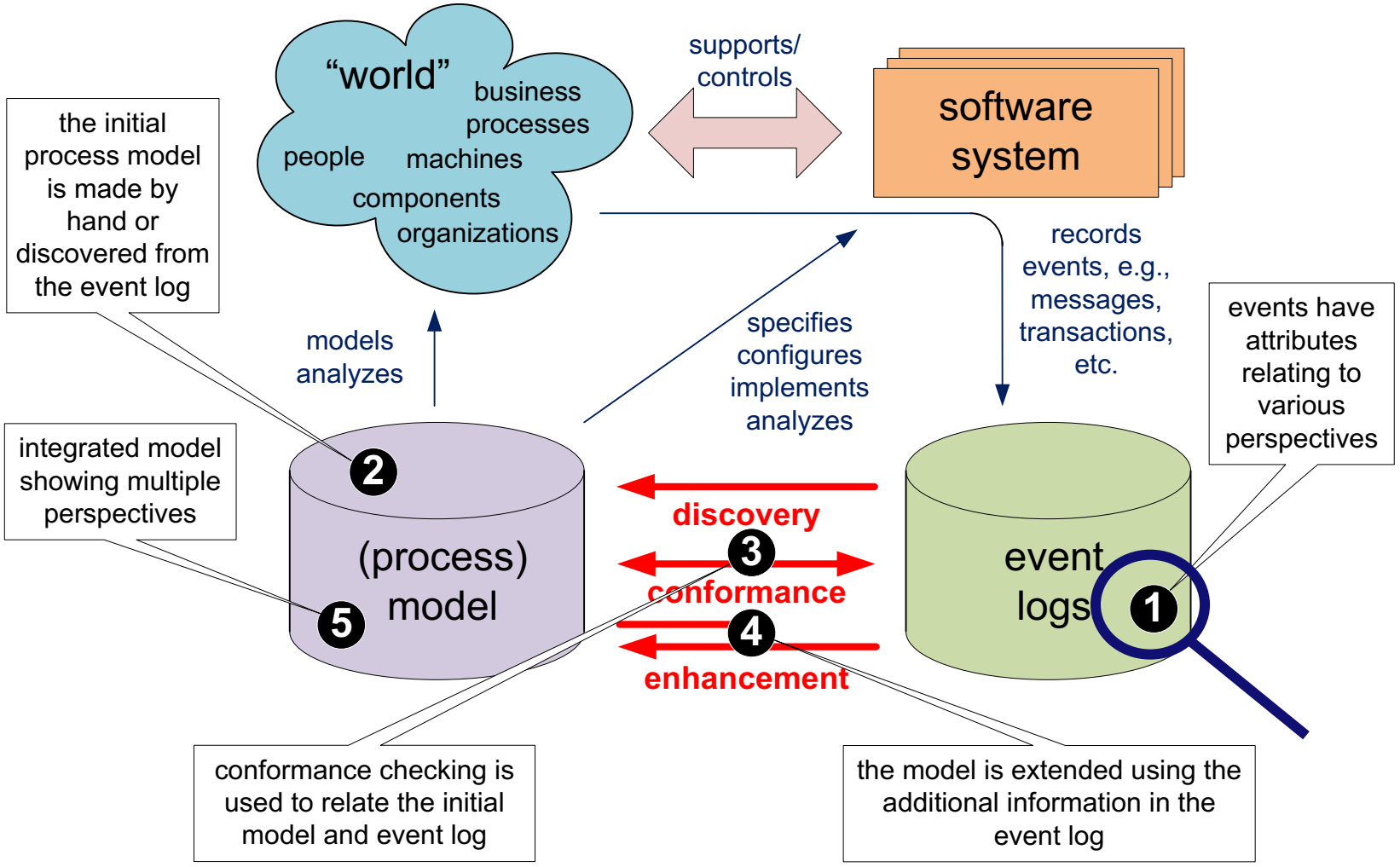
Cases may also have attributes

case id	custid	name	type	region	amount
1	9911	Smith	gold	south	989.50
2	9915	Jones	silver	west	546.00
3	9912	Anderson	silver	north	763.20
4	9904	Thompson	silver	west	911.70
5	9911	Smith	gold	south	812.10
6	9944	Baker	silver	east	788.00
7	9944	Baker	silver	east	792.80
8	9911	Smith	gold	south	544.70
...	...	...	...	...	...

XES
Extensible Event Stream

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Extending process models using event data

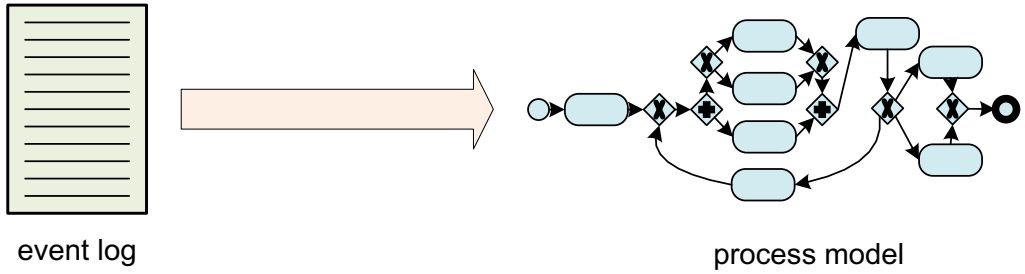


©Wil van der Aalst TU/e (use only with permission & acknowledgements)

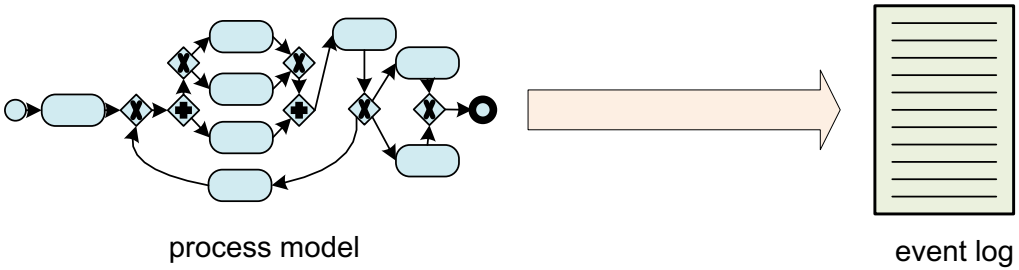
replay revisited

Connecting events to model elements is essential for model extension

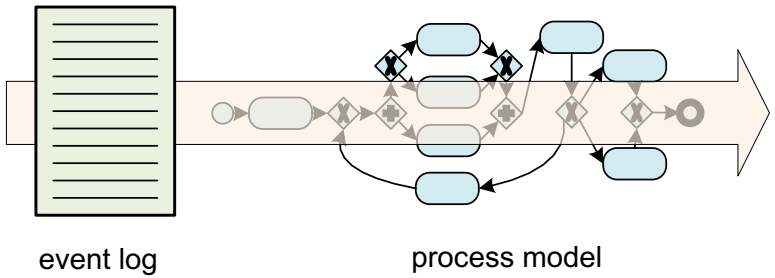
Play-In



Play-Out



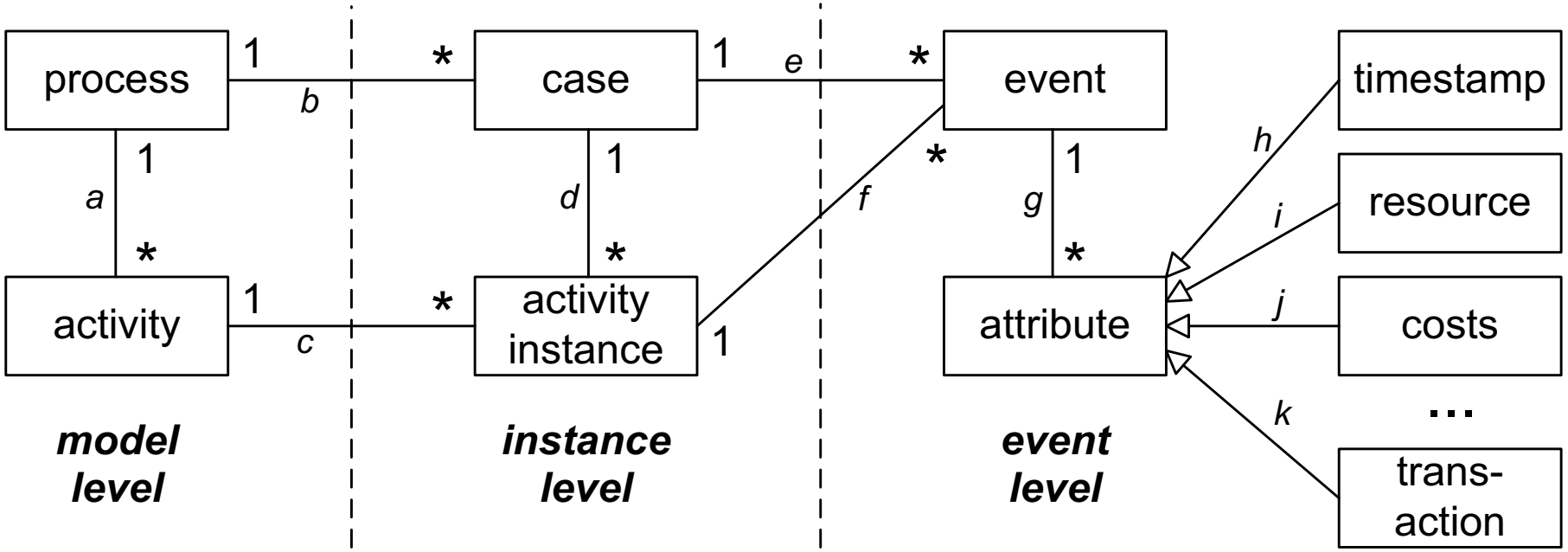
Replay



- extended model showing times, frequencies, etc.
- diagnostics
- predictions
- recommendations

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Connected event log and model through replay or alignments

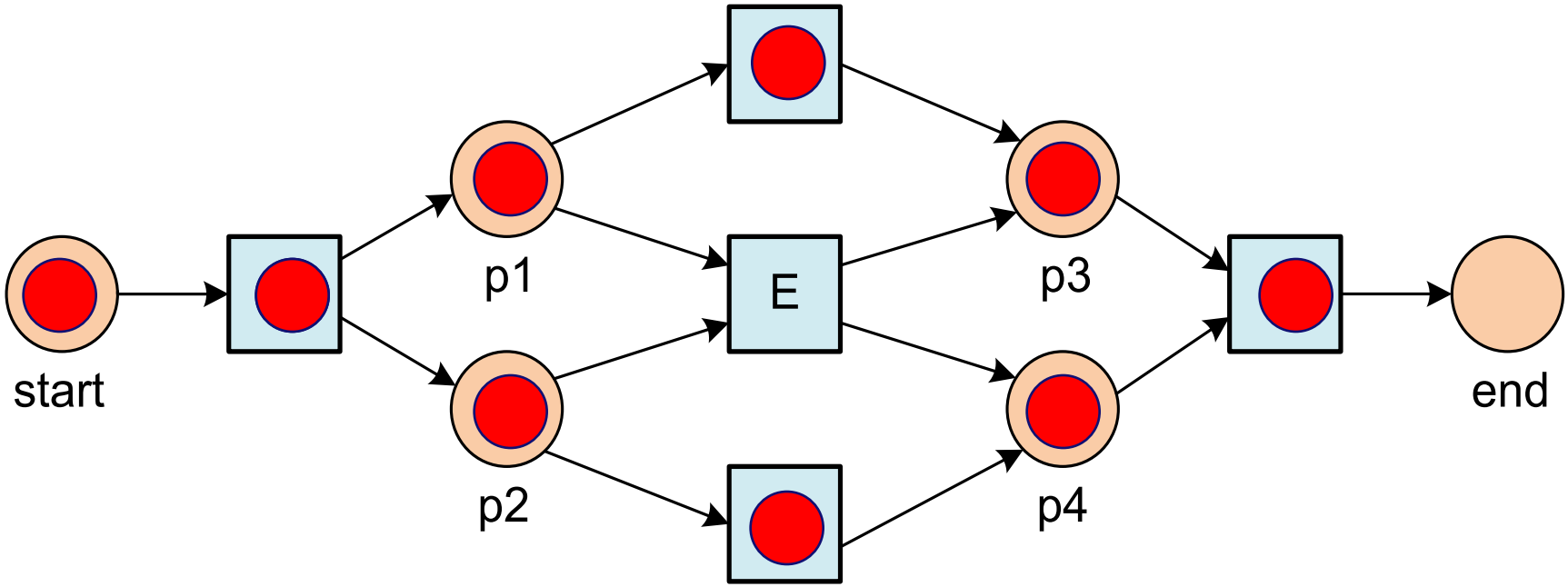


trace in the event log is related to a path in the model

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Remember: Replay!

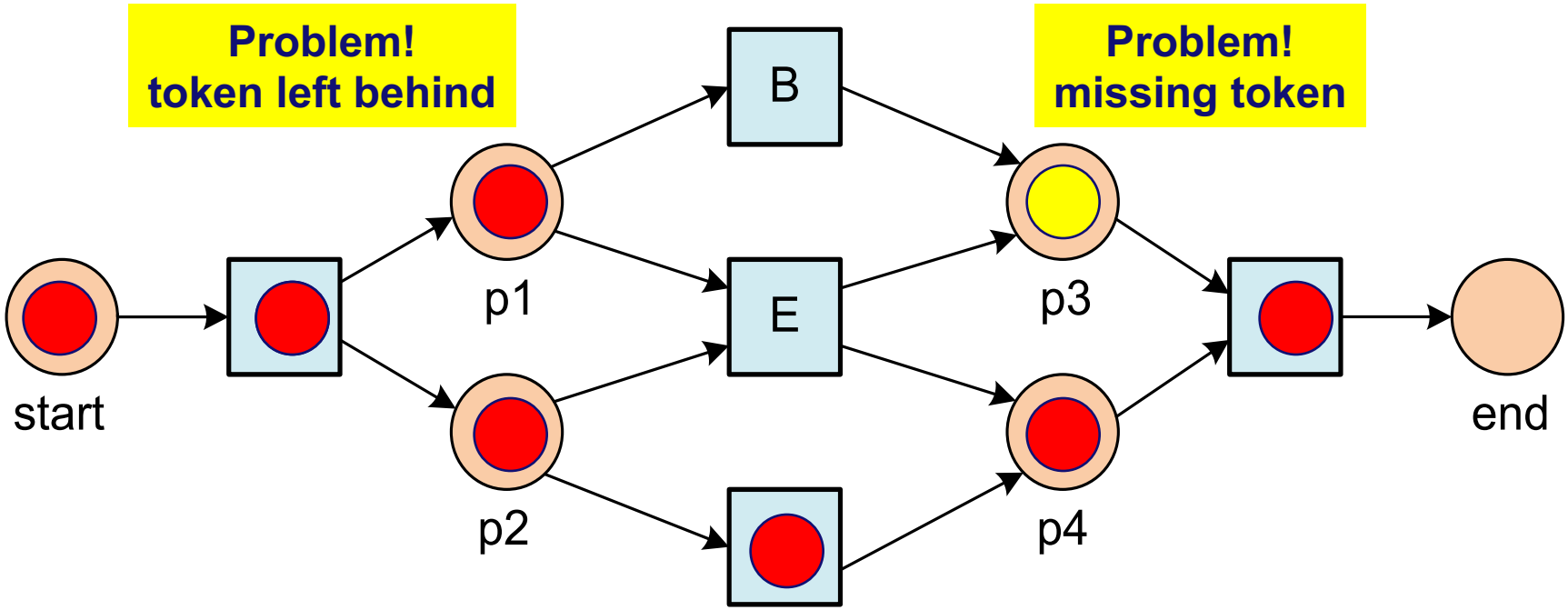
A B C D



©Wil van der Aalst TU/e (use only with permission & acknowledgements)

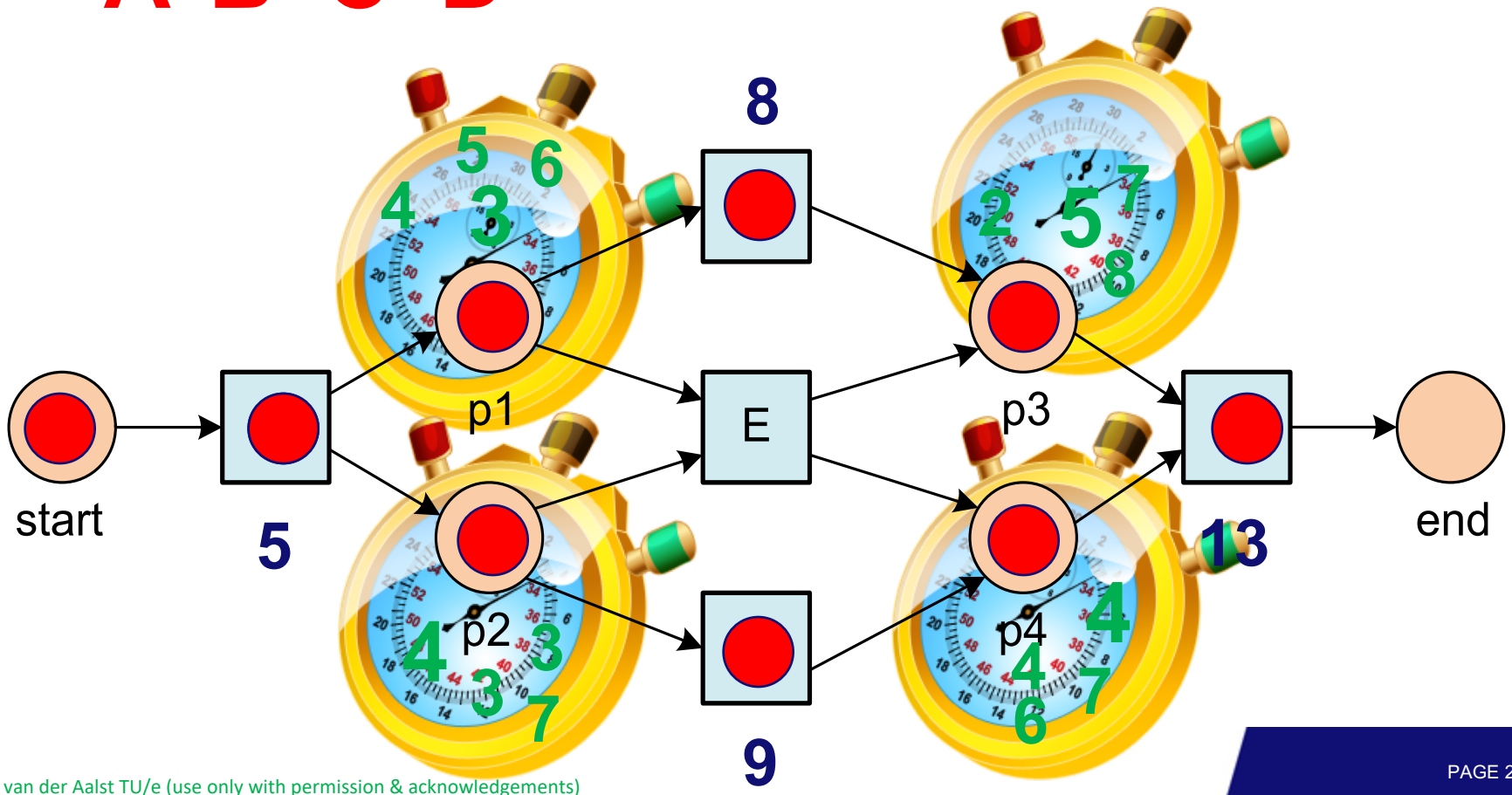
Replay can detect problems

A C D



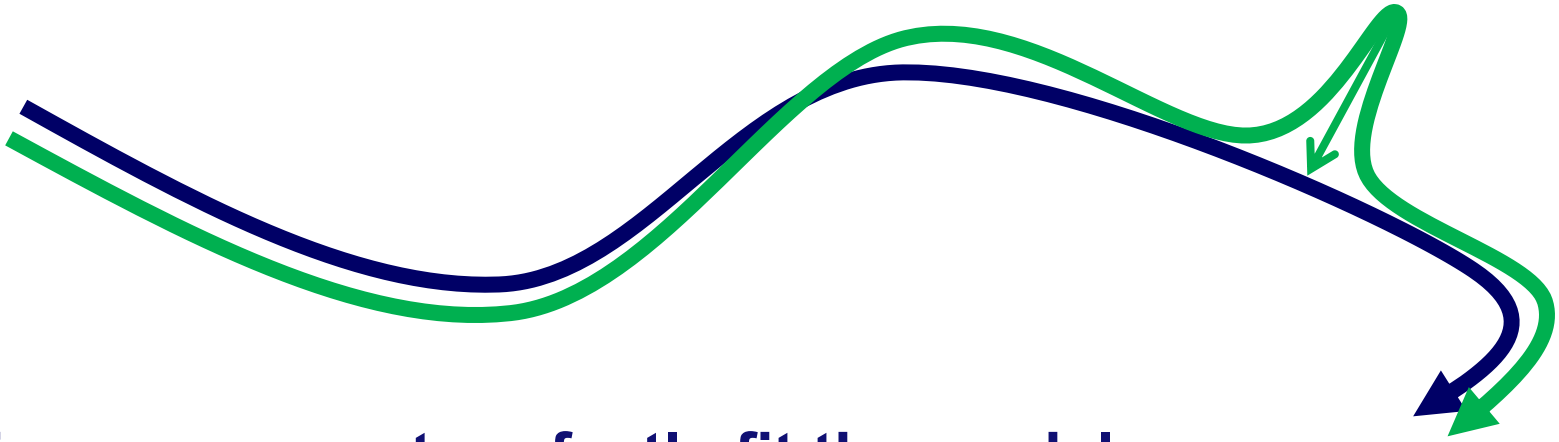
©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Replay can extract timing information

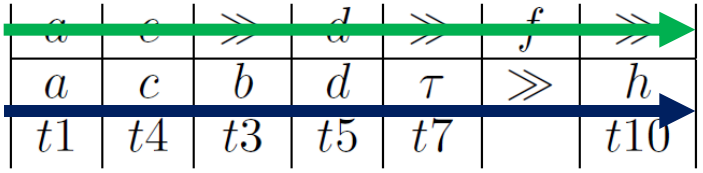


©Wil van der Aalst TU/e (use only with permission & acknowledgements)

When there are deviation we need to squeeze reality into the model



- Traces may not perfectly fit the model.
- Often we cannot throw non-fitting traces (loosing most of the data and/or introducing a hidden bias).
- Conformance checking techniques help us to map traces onto the "nearest" path in the model.



©Wil van der Aalst TU/e (use only with permission & acknowledgements)

In the remainder we assume that through prepressing model and log are aligned

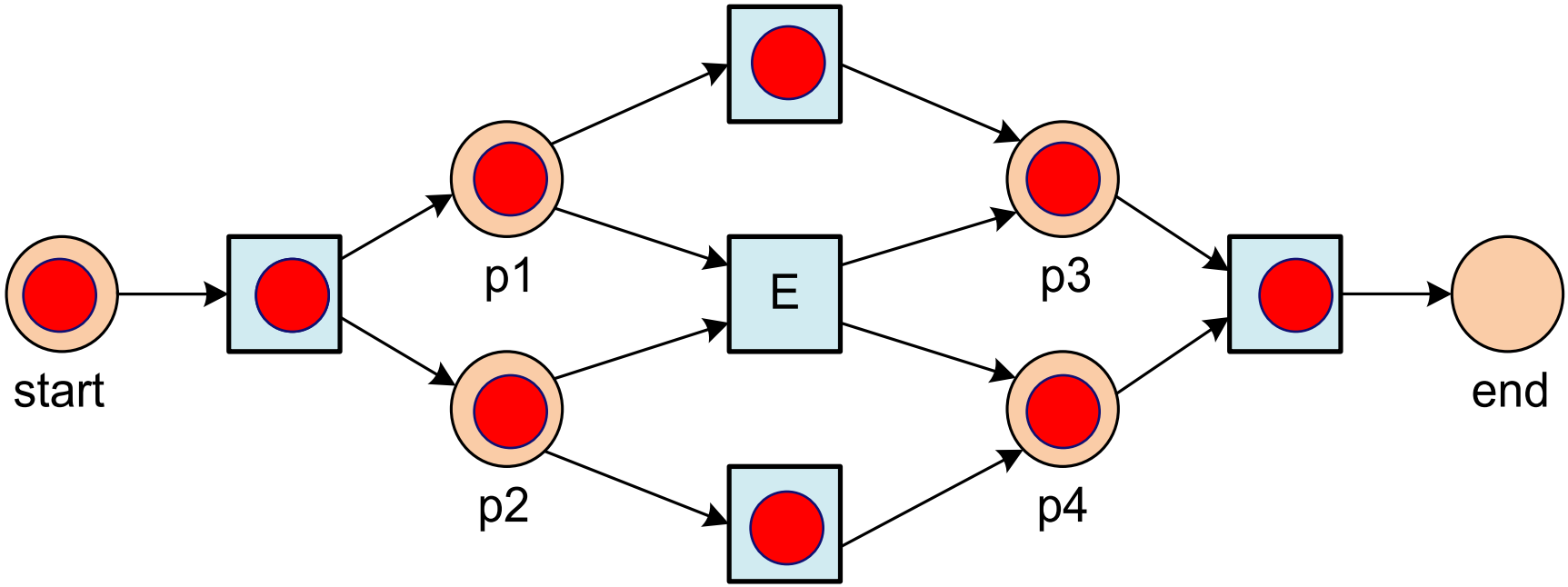
Assumption: every complete trace corresponds to a complete path through the model.

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

mining decision points

Decision mining: “Red” cases

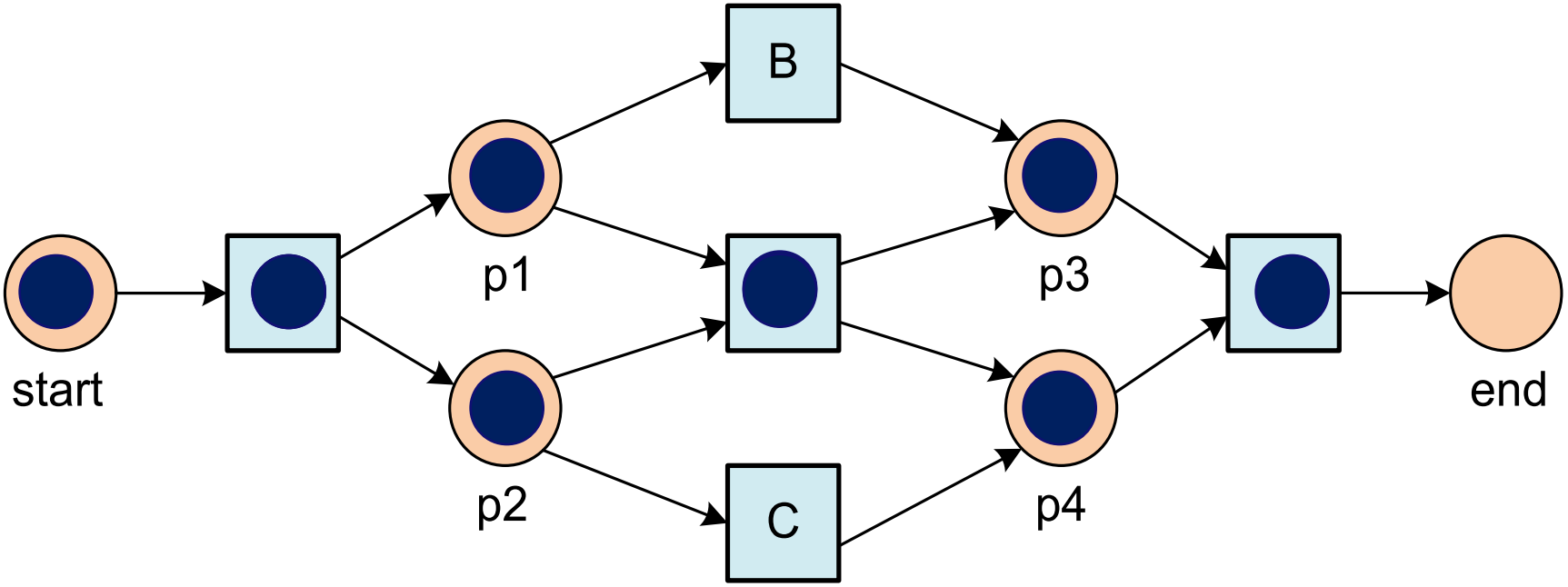
A B C D



©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Decision mining: “Blue” cases

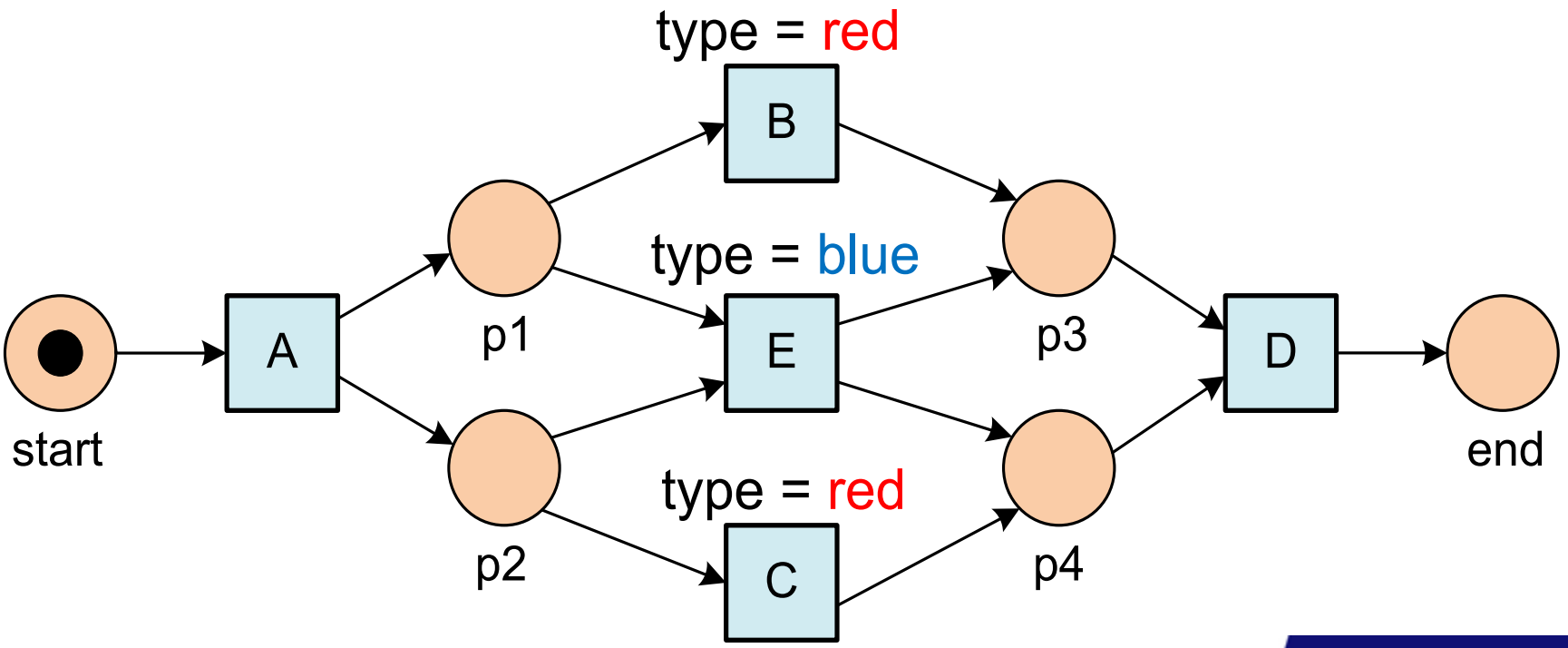
A E D



©Wil van der Aalst TU/e (use only with permission & acknowledgements)

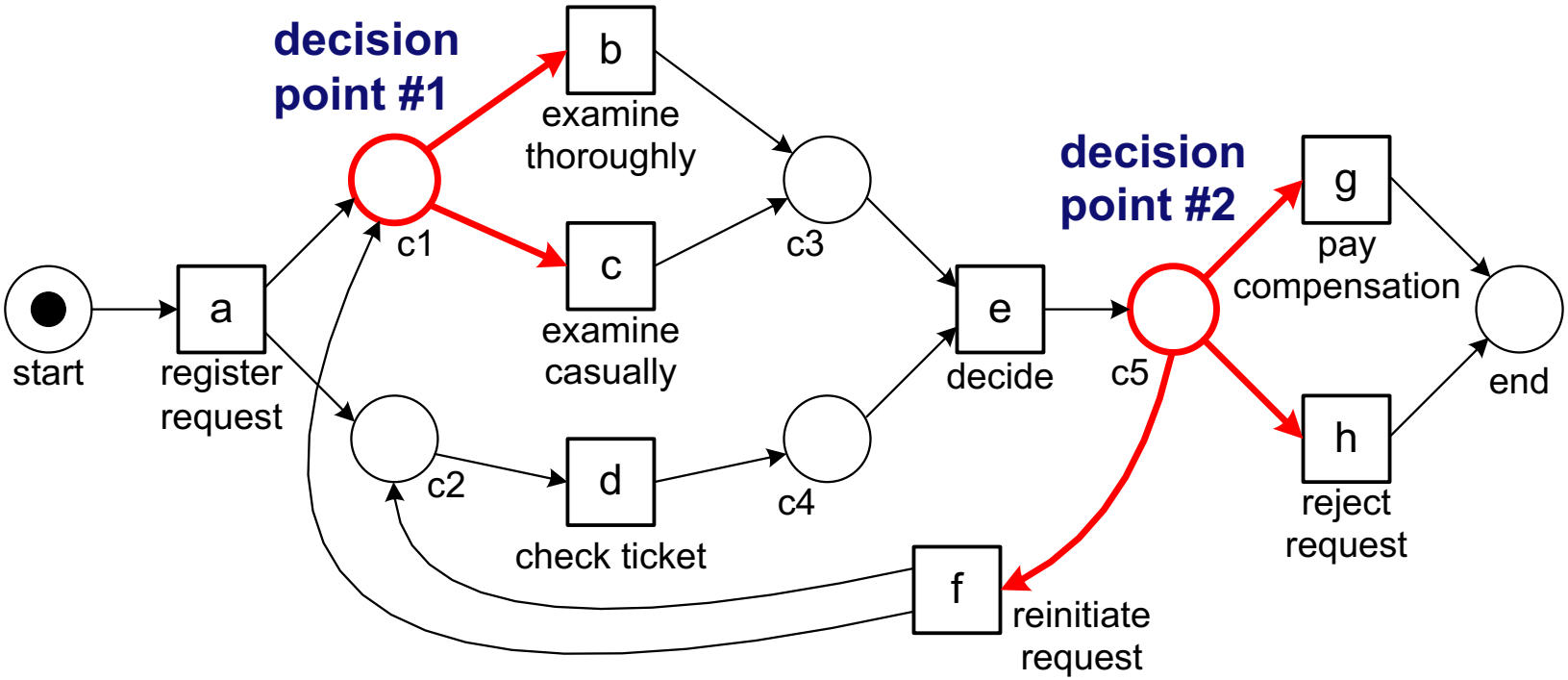
Guards ensure that the right path is taken

If red then B+C;
If blue then E;



©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Decision mining: places with multiple output arcs form decision points



In other notations often specific building blocks, e.g., XOR/OR gateways.

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

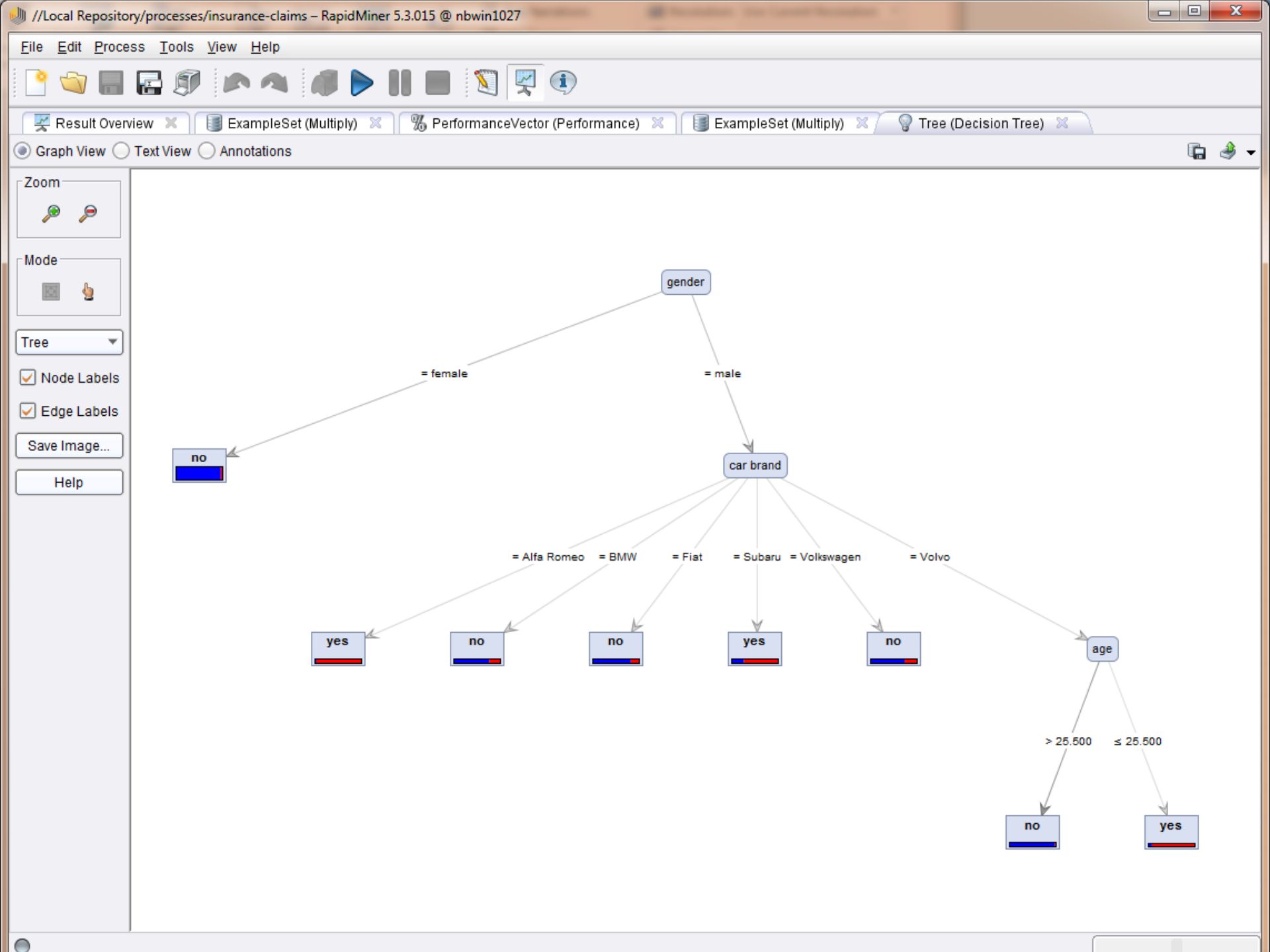
Remember: classification using decision trees

- **Response variable (dependent variable): claim (yes/no).**
- **Predictor variables (independent variables): gender, age, smoker, car brand.**

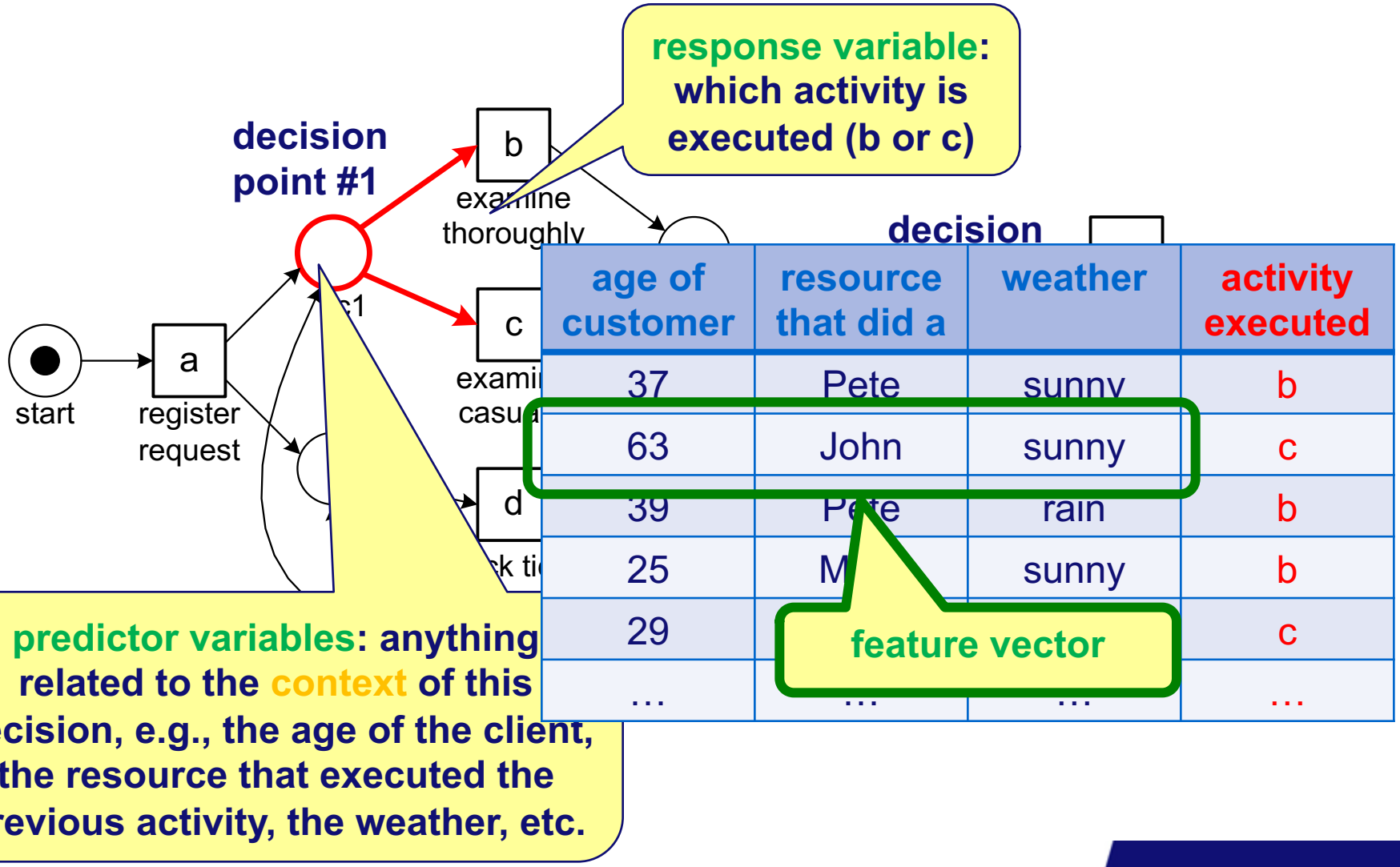
gender	age	smoker	car brand	claim
female	47	yes	Volvo	no
male	31	no	Alfa Romeo	yes
male	59	no	Alfa Romeo	yes
male	28	no	Fiat	no
male	44	no	BMW	no
female	27	no	Fiat	no
male	29	no	Subaru	no
male	44	yes	Subaru	yes
male	39	no	BMW	no
male	35	no	Subaru	yes

Goal: explain response variable in terms of relevant predictor variables.

©Wil van der Aalst TU/e (use only with permission & acknowledgements)


//Local Repository/processes/insurance-claims – RapidMiner 5.3.015 @ nbwin1027
File Edit Process Tools View Help
Result Overview
ExampleSet (Multiply)
PerformanceVector (Performance)
ExampleSet (Multiply)
Tree (Decision Tree)
Graph View
Text View
Annotations
Zoom
Mode
Tree
Node Labels
Edge Labels
Save Image...
Help
gender
= female
= male
no
car brand
= Alfa Romeo
= BMW
= Fiat
= Subaru
= Volkswagen
= Volvo
yes
no
no
yes
no
age
> 25.500
≤ 25.500
no
yes

Creating a classification problem

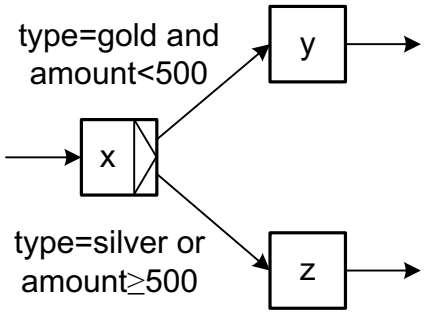


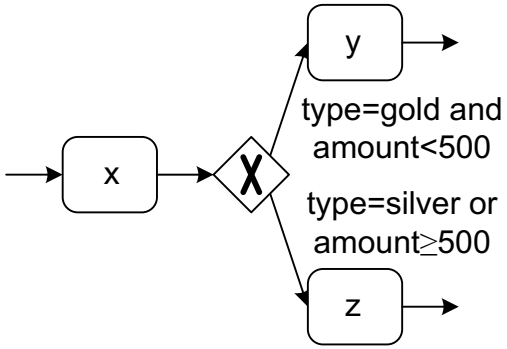
age of customer	resource that did a	weather	activity executed
37	Pete	sunny	b
63	John	sunny	c
39	Pete	rain	b
25	M	sunny	b
29			c
…	…	…	…

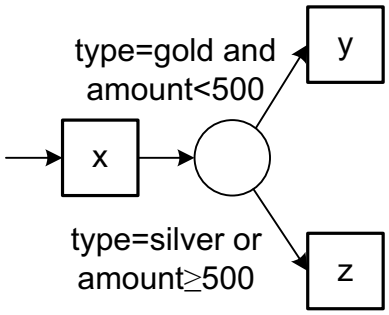
©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Example: XOR-split

type	region	amount	activity
gold	south	987.30	z
silver	north	178.70	z
gold	south	211.50	y
silver	west	587.70	z
silver	east	224.70	z
silver	south	278.50	z
gold	north	488.50	y
silver	west	443.20	z
silver	south	673.70	z
gold	west	413.50	y
silver	south	687.70	z
gold	south	987.30	z
silver	north	378.80	z
gold	south	314.50	y
silver	north	537.70	z
silver	west	158.70	z
gold	east	344.50	y
...	...	...	...







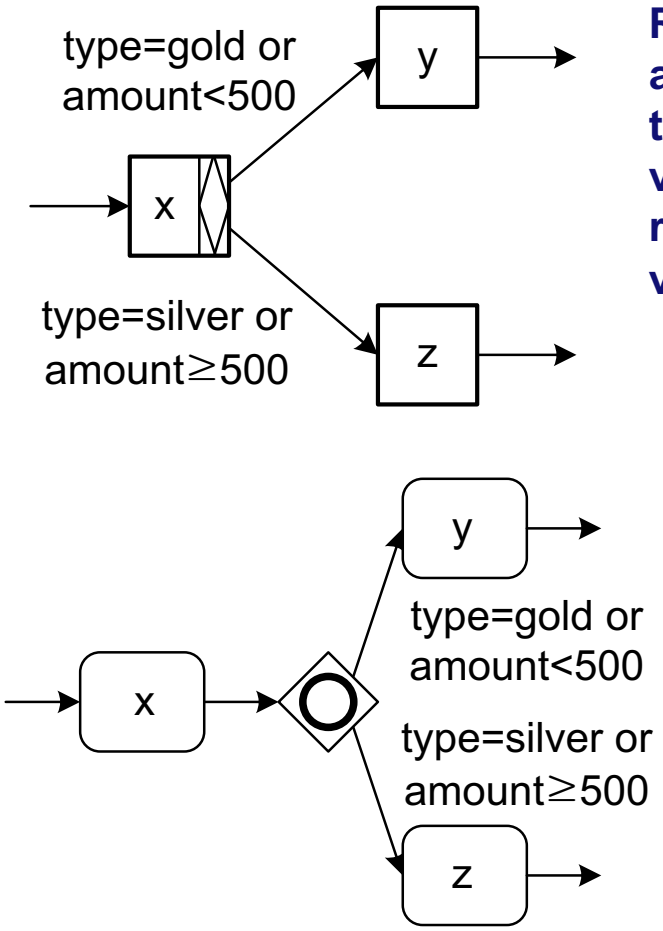
What are the "features" (predictor variables) influencing the decision?

A classification technique like decision tree learning can be used to find such rules: explain response variable (dependent variable) in terms of predictor variables (independent variables).

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Example: OR-split

type	region	amount	activity
gold	south	987.30	y and z
silver	north	178.70	y and z
gold	south	211.50	just y
silver	west	587.70	just z
silver	east	224.70	y and z
silver	south	278.50	y and z
gold	north	488.50	just y
silver	west	443.20	y and z
silver	south	673.70	just z
...	...	...	...

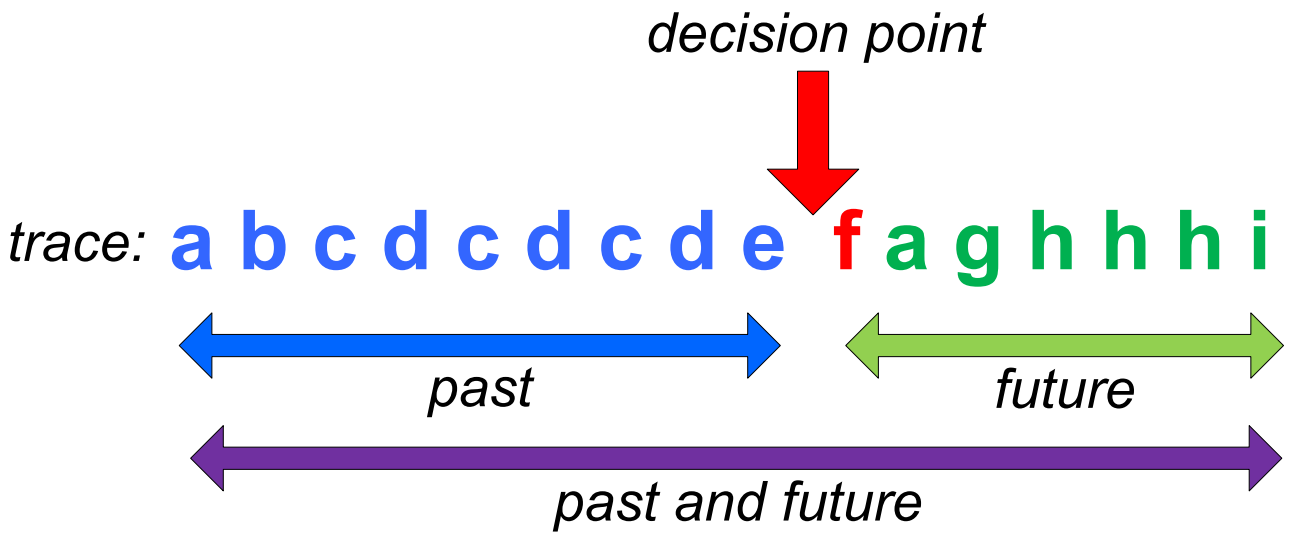


Replay/ alignment will tell what the value of the response variable is.

Response variable (dependent variable) is now an activity set rather than a single activity.

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

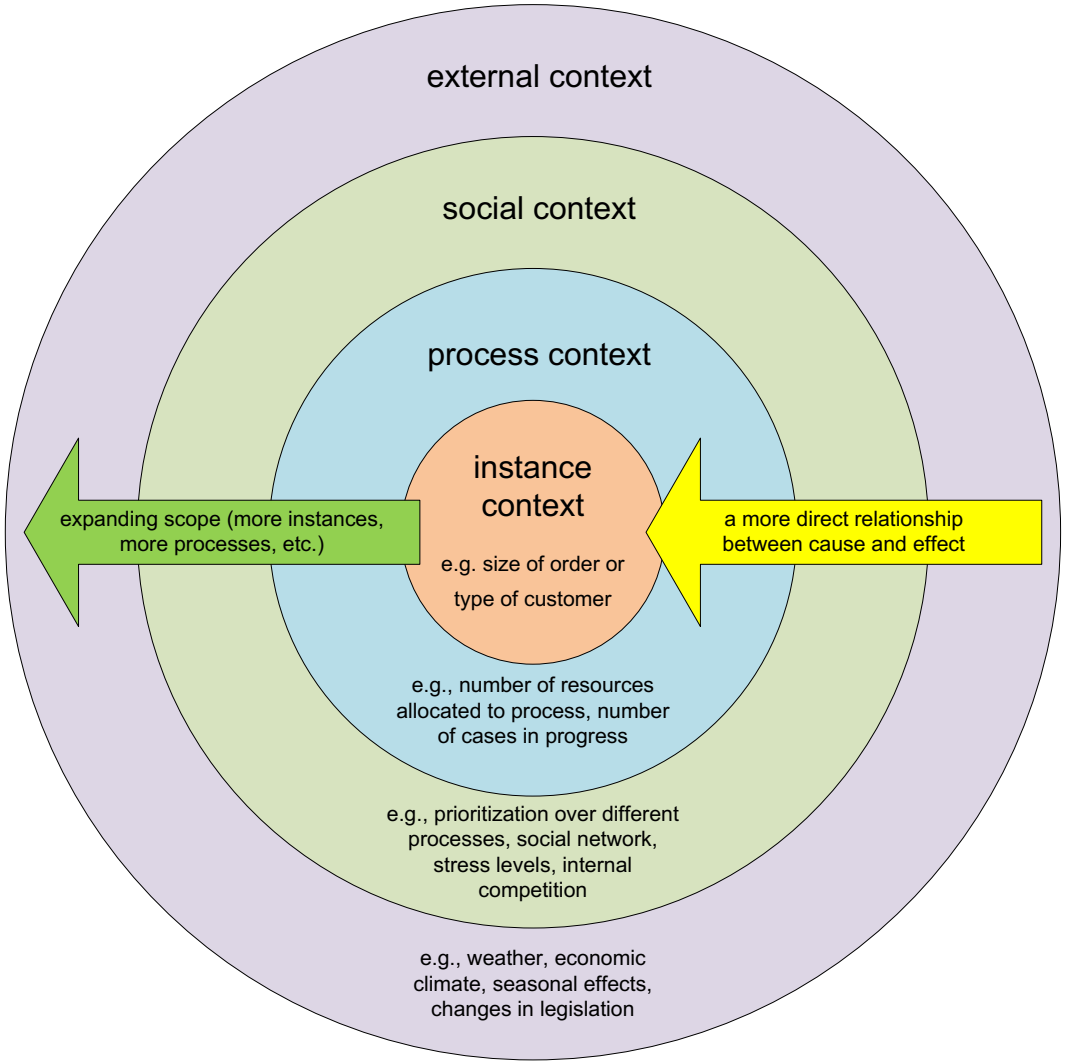
Predictor variables based on process instance only



- **Case variables**
- **Attributes (data, resource, etc.) of all past events**
- **Attributes (data, resource, etc.) of last event only**
- **Attributes (data, resource, etc.) of all events (including future).**

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Predictor variables based on context of the process instance

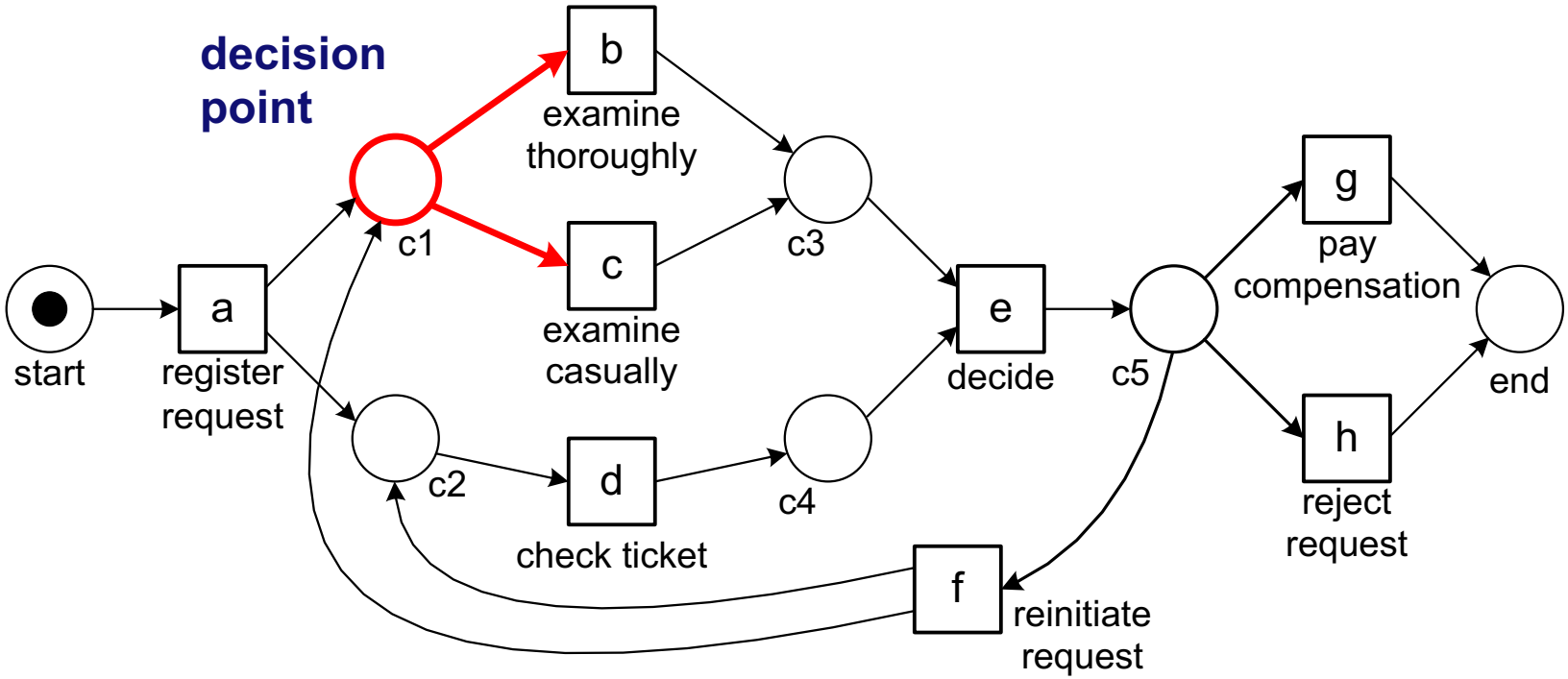


- Number of cases running (e.g. skip check if busy).
- Number of resources present.
- Day of the week.
- Weather.

Problem: curse of dimensionality!

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

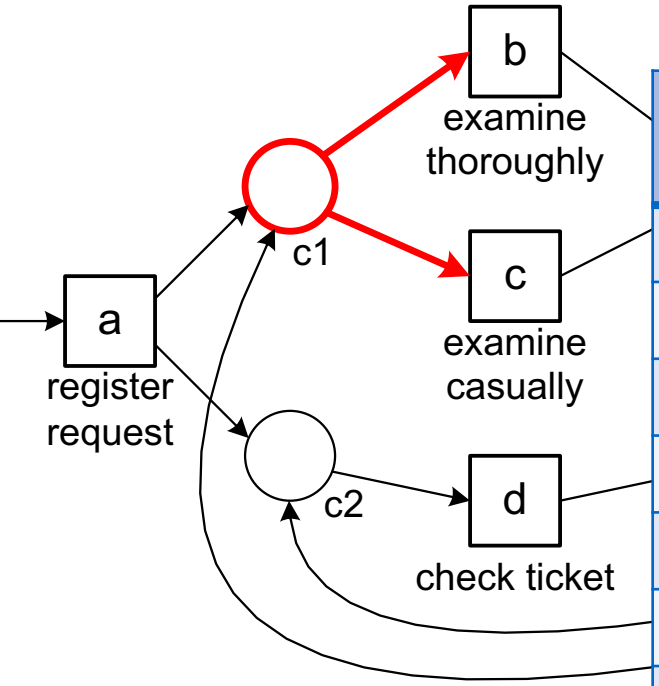
Question: Build classification problem to learn the decision point



case	activity	resource	time	customer	amount
1	a	John	8.11	silver	500
2	a	Mary	8.12	gold	800

©Wil van der Aalst TU/e (use only with permission & acknowle

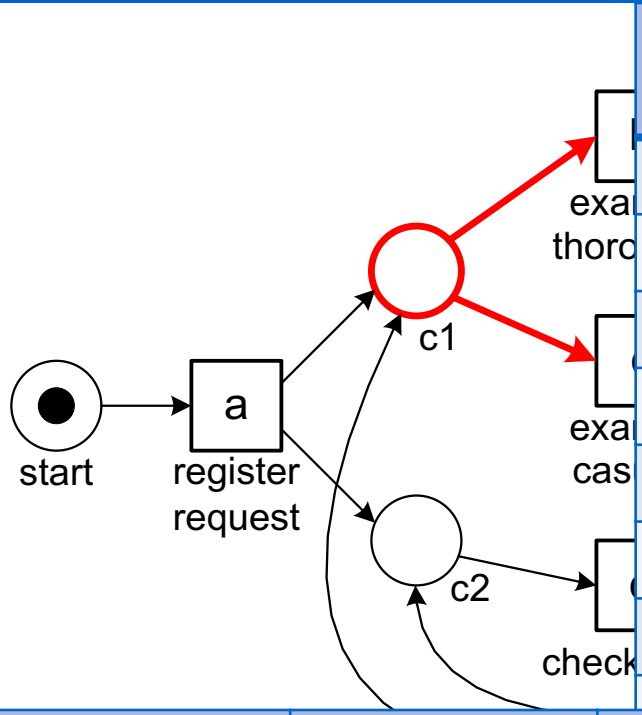
Question: Build classification problem to learn the decision point



case	activity	resource	time	customer	amount
1	a	John	8.11	silver	500
2	a	Mary	8.12	gold	800
2	d	Sue	8.32	gold	800
1	b	John	9.12	silver	500
3	a	John	9.45	silver	300
3	c	Mary	9.56	silver	300
1	d	John	9.45	silver	500
2	c	Mary	9.56	gold	800
3	d	Mary	10.43	silver	300
4	a	John	11.34	gold	850
4	c	John	11.57	gold	850
…	…	…	…	…	…

©Wil van der Aalst TU/e (use only with permission & acknowledg

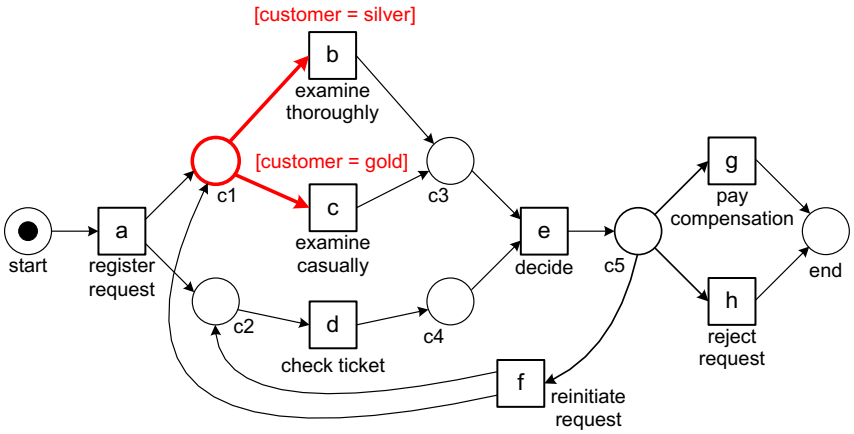
Classification problem (example, not all possible attributes included)

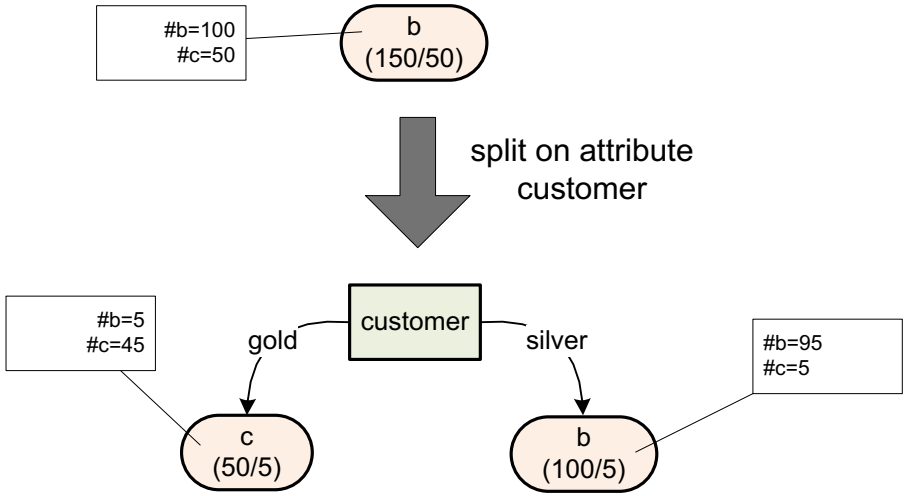


case	activity	resource	time	customer	amount
1	a	John	8.11	silver	500
2	a	Mary	8.12	gold	800
2	d	Sue	8.32	gold	800
1	b	John	9.12	silver	500
3	a	John	9.45	silver	300
3	c	Mary	9.56	silver	300
1	d	John	9.45	silver	500
				gold	800
				silver	300
				gold	850
				gold	850
				…	…

case	resource executing a	customer	amount	class
1	John	silver	500	b
2	Mary	gold	800	c
3	John	silver	300	b
4	John	gold	850	c

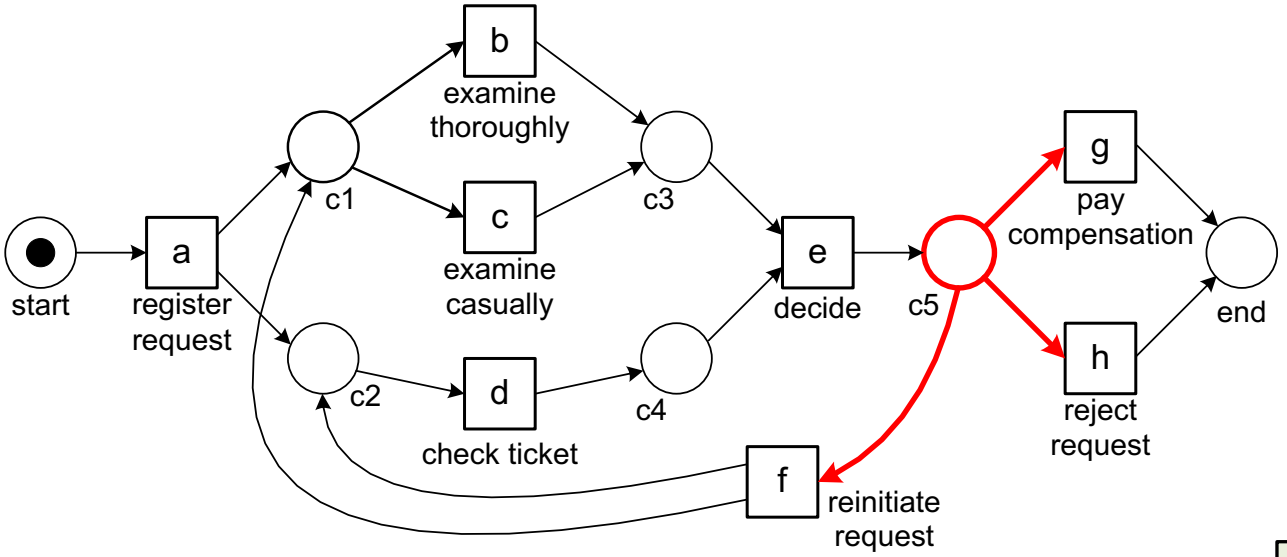
Decision tree and resulting guards



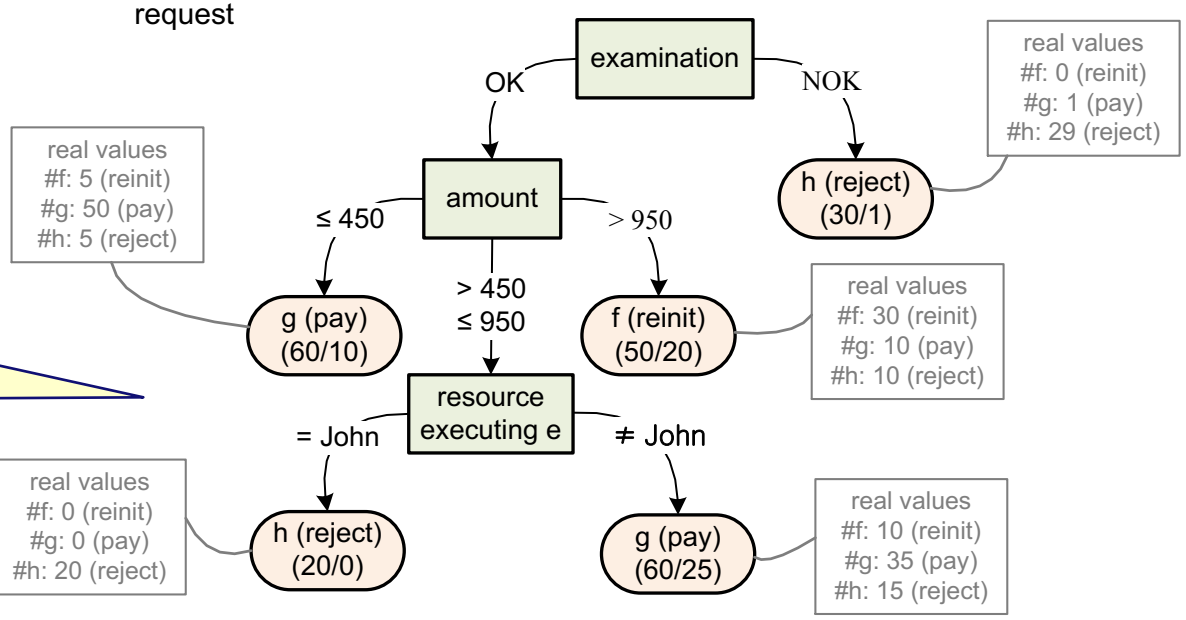


case	resource executing a	customer	amount	class
1	John	silver	500	b
2	Mary	gold	800	c
3	John	silver	300	b
4	John	gold	850	c

Question: create guards based on decision tree



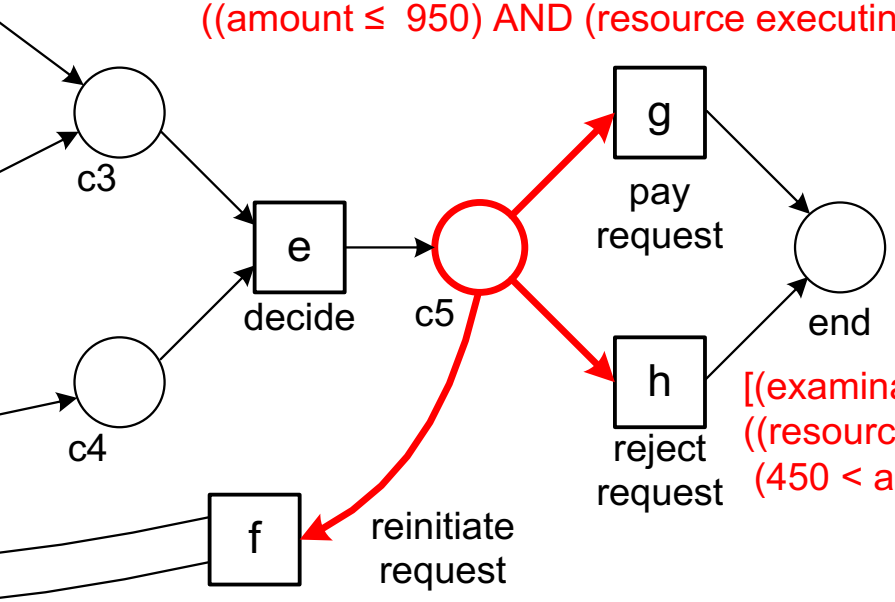
assume that this decision tree was learned for decision point



©Wil van der Aalst TU/e (use only with permission & acknowledg

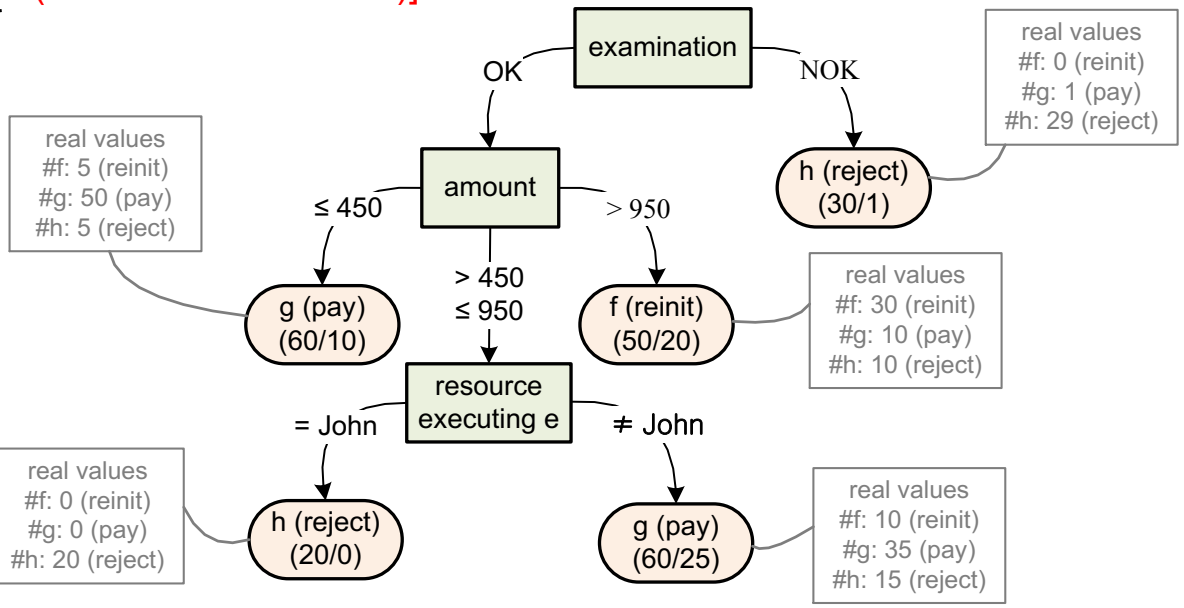
Deterministic guards

[(examination = OK) AND ((amount <= 450) OR ((amount ≤ 950) AND (resource executing e ≠ John)))]



[(examination = NOK) OR ((resource executing e = John) AND (450 < amount ≤ 950)]

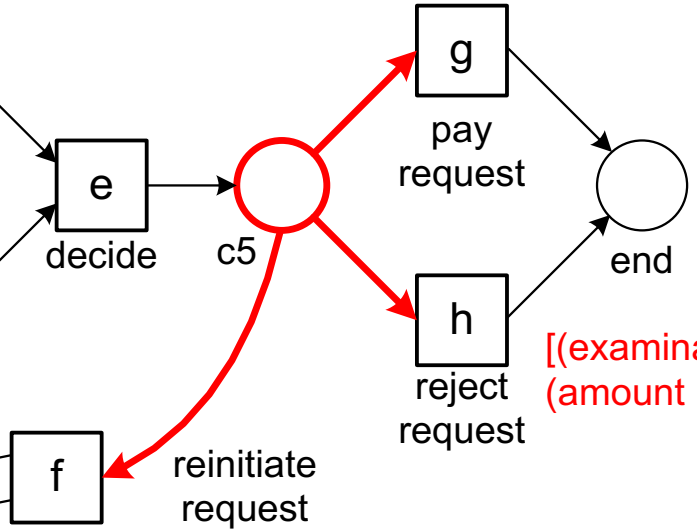
[(examination = OK) AND (amount > 950)]



©Wil van der Aalst TU/e (use only with permission & acknowledg

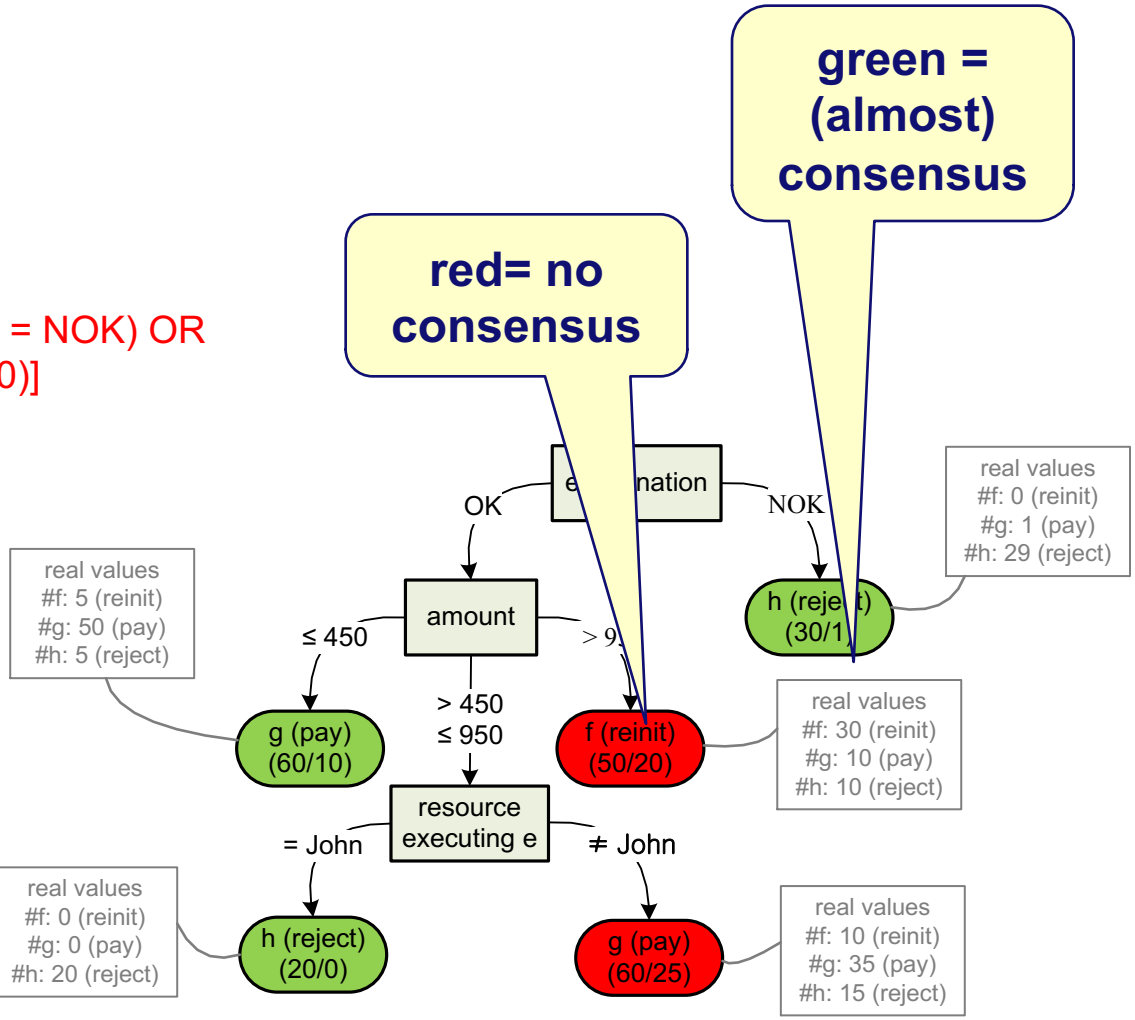
Non-deterministic guards

[(examination = OK) AND ((amount ≤ 450) OR (amount > 950) OR (resource executing e ≠ John)))]



[(examination = NOK) OR (amount > 450)]

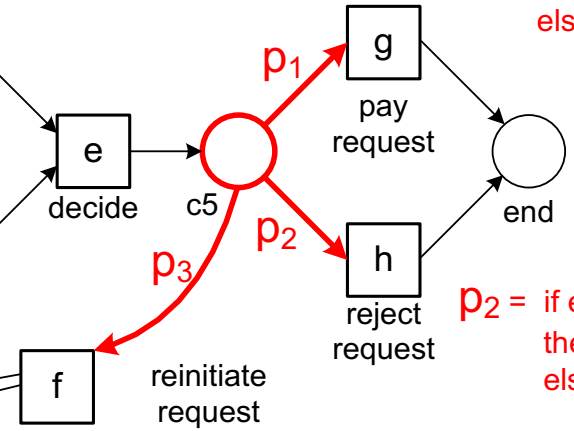
[(examination = OK) AND ((amount > 950) OR ((amount > 450) AND (resource executing e ≠ John)))]

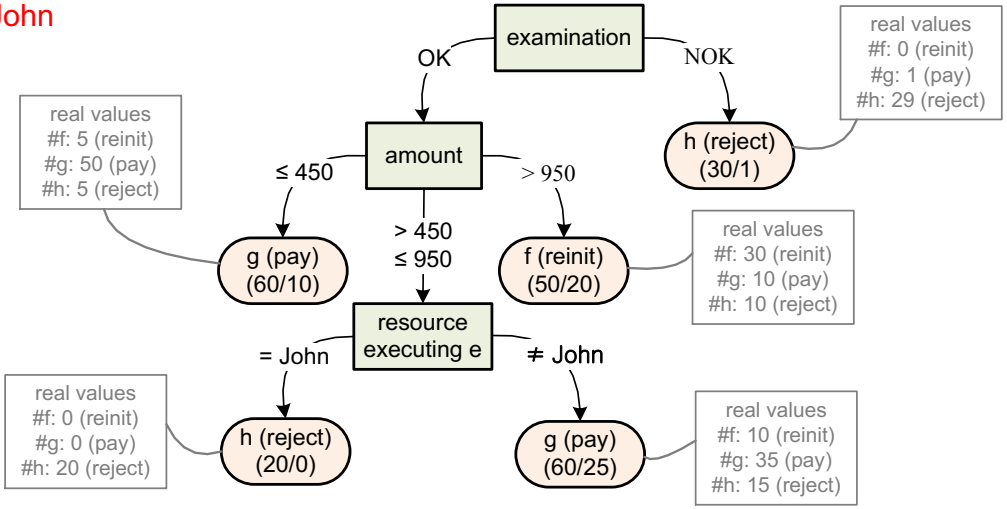


©Wil van der Aalst TU/e (use only with permission & acknowledgemen

Data-dependent probabilities rather than guards

p_1 = if examination = NOK
then 1/30
else if amount ≤ 450
then 50/60
else if amount > 950
then 10/50
else if resource executing a = John
then 0/20
else 35/60





p_2 = if examination = NOK
then 29/30
else if amount ≤ 450
then 5/60
else if amount > 950
then 10/50
else if resource executing a = John
then 20/20
else 15/60

p_3 = if examination = NOK
then 0/30
else if amount ≤ 450
then 5/60
else if amount > 950
then 30/50
else if resource executing a = John
then 0/20
else 10/60

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

So we can mine decision points …

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

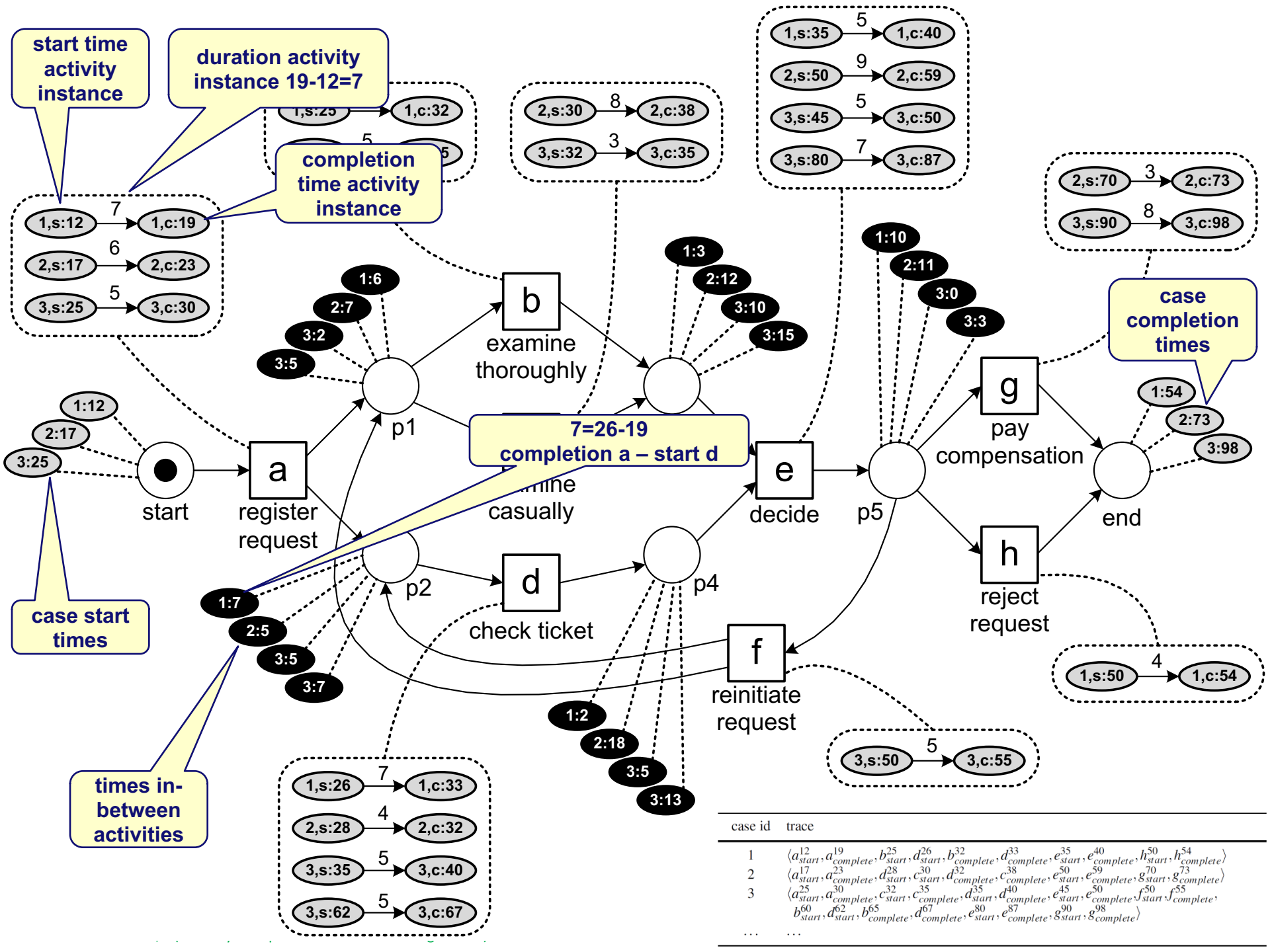
mining bottlenecks

Learning time and probabilities

case id	trace
1	$\langle a^{12}_{start}, a^{19}_{complete}, b^{25}_{start}, d^{26}_{start}, b^{32}_{complete}, d^{33}_{complete}, e^{35}_{start}, e^{40}_{complete}, h^{50}_{start}, h^{54}_{complete} \rangle$
2	$\langle a^{17}_{start}, a^{23}_{complete}, d^{28}_{start}, c^{30}_{start}, d^{32}_{complete}, c^{38}_{complete}, e^{50}_{start}, e^{59}_{complete}, g^{70}_{start}, g^{73}_{complete} \rangle$
3	$\langle a^{25}_{start}, a^{30}_{complete}, c^{32}_{start}, c^{35}_{complete}, d^{35}_{start}, d^{40}_{complete}, e^{45}_{start}, e^{50}_{complete}, f^{50}_{start}, f^{55}_{complete}, b^{60}_{start}, d^{62}_{start}, b^{65}_{complete}, d^{67}_{complete}, e^{80}_{start}, e^{87}_{complete}, g^{90}_{start}, g^{98}_{complete} \rangle$
...	...

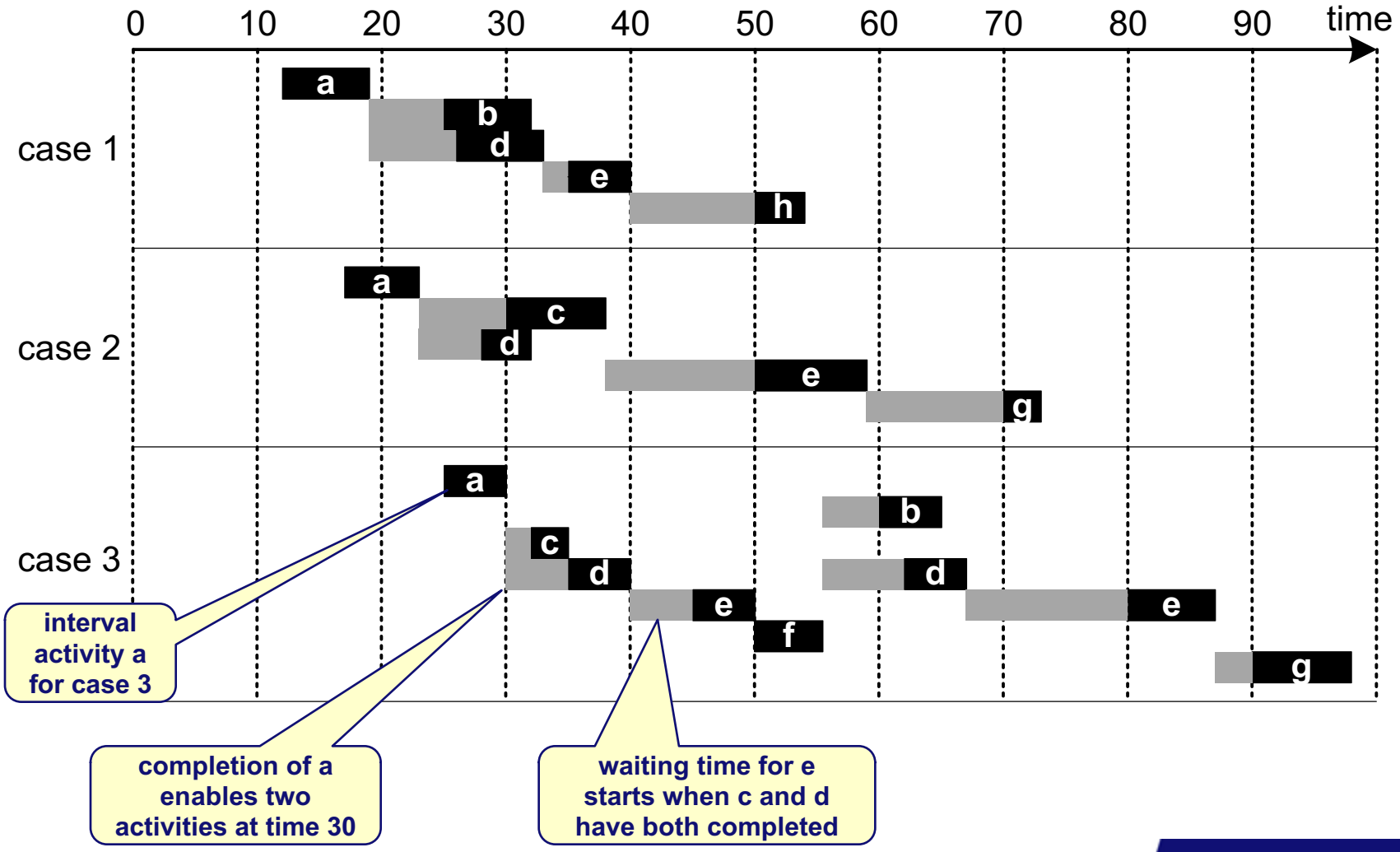
- Replay, as before, but now considering timestamps.
- Let us replay the first three cases in the event log:
  - case 1 starts at time 12 and ends at time 54,
  - case 2 starts at time 17 and ends at time 73,
  - case 3 starts at time 25 and ends at time 98.

©Wil van der Aalst TU/e (use only with permission & acknowledgements)



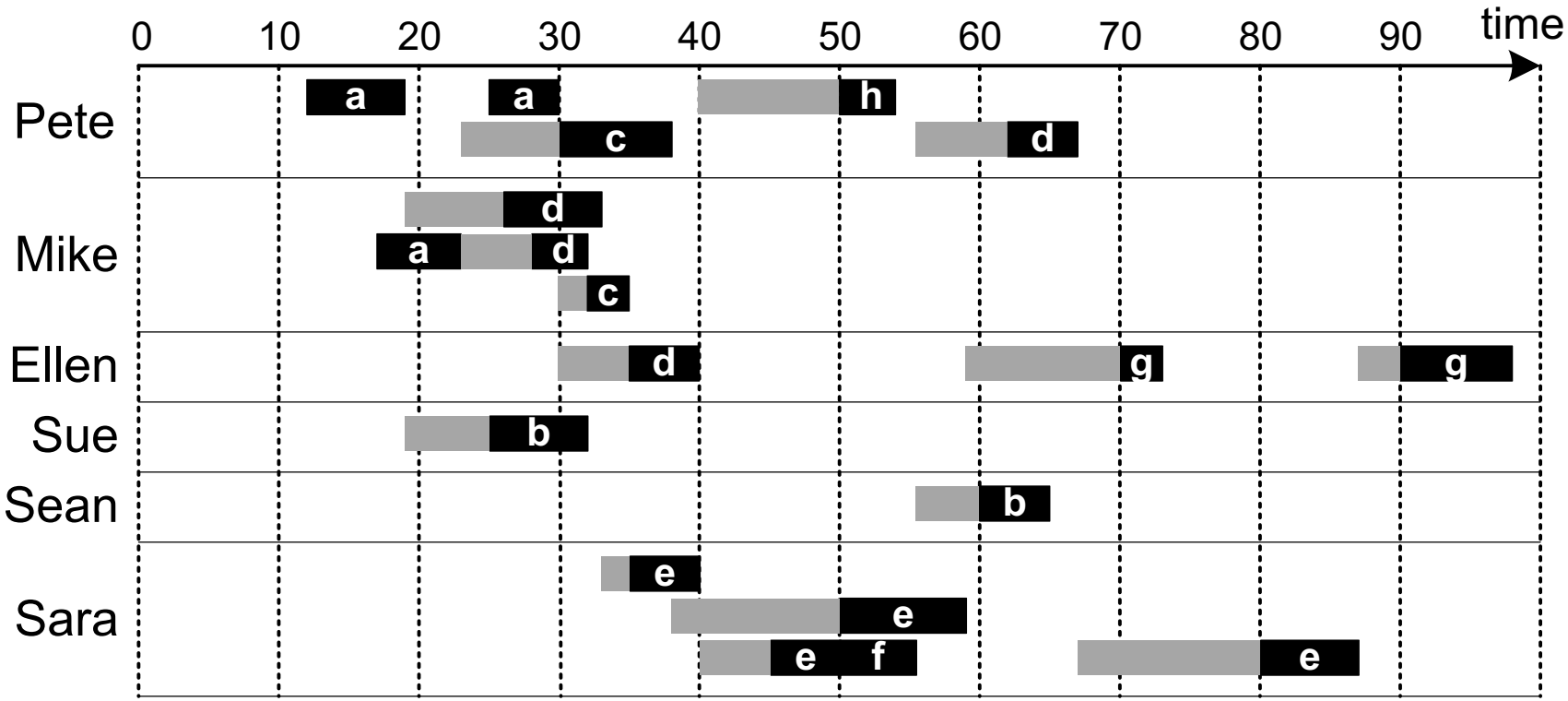
case id	trace
1	$\langle a^{12}_{start}, a^{19}_{complete}, b^{25}_{start}, d^{26}_{start}, b^{32}_{complete}, d^{33}_{complete}, e^{35}_{start}, e^{40}_{complete}, h^{50}_{start}, h^{54}_{complete} \rangle$
2	$\langle a^{17}_{start}, a^{23}_{complete}, d^{28}_{start}, c^{30}_{start}, d^{32}_{complete}, c^{38}_{complete}, e^{50}_{start}, e^{59}_{complete}, g^{70}_{start}, g^{73}_{complete} \rangle$
3	$\langle a^{25}_{start}, a^{30}_{complete}, c^{32}_{start}, c^{35}_{complete}, d^{35}_{start}, d^{40}_{complete}, e^{45}_{start}, e^{50}_{complete}, f^{50}_{start}, f^{55}_{complete}, b^{60}_{start}, d^{62}_{start}, b^{65}_{complete}, d^{67}_{complete}, e^{80}_{start}, e^{87}_{complete}, g^{90}_{start}, g^{98}_{complete} \rangle$
...	...

Another view on the timed replay of the first three cases



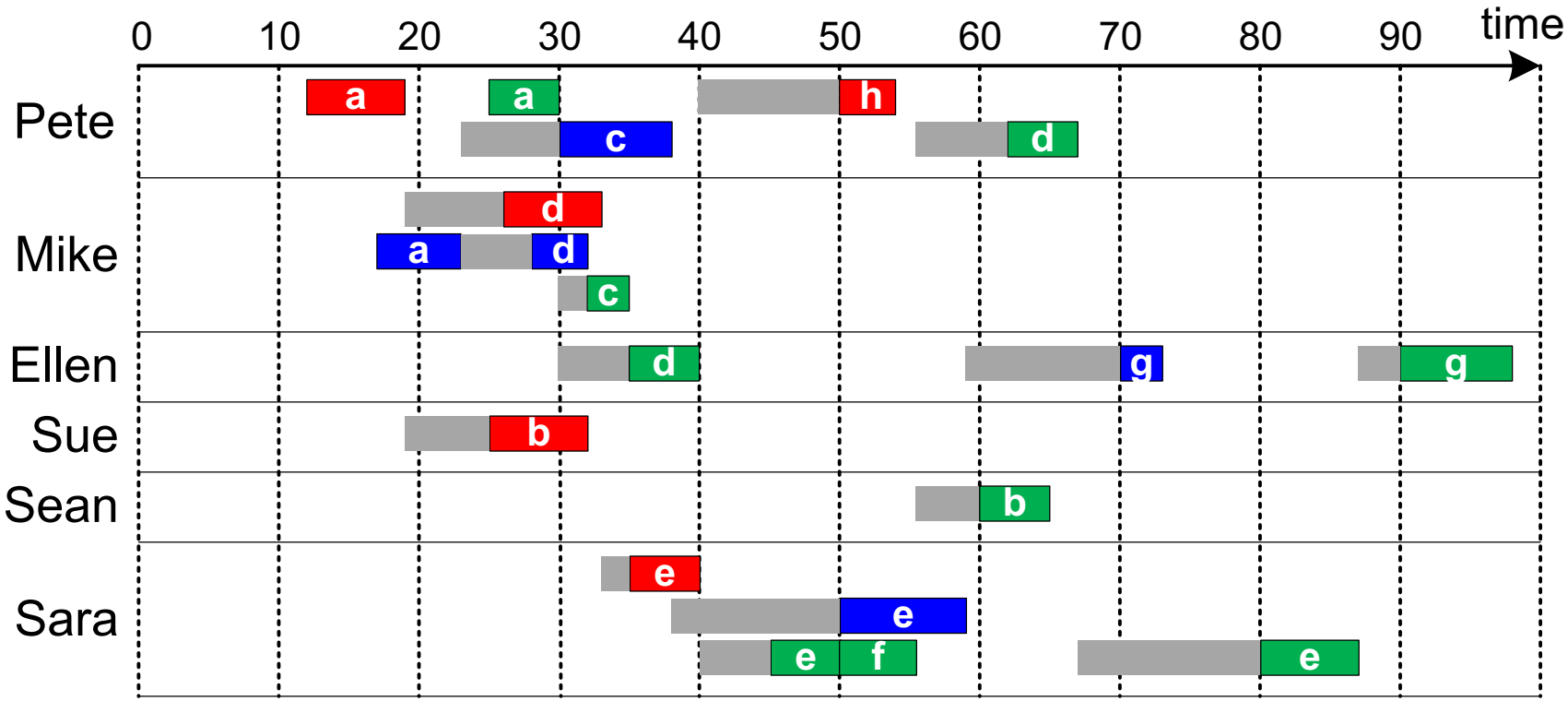
©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Timed replay projected onto resources



©Wil van der Aalst TU/e (use only with permission & acknowledgements)

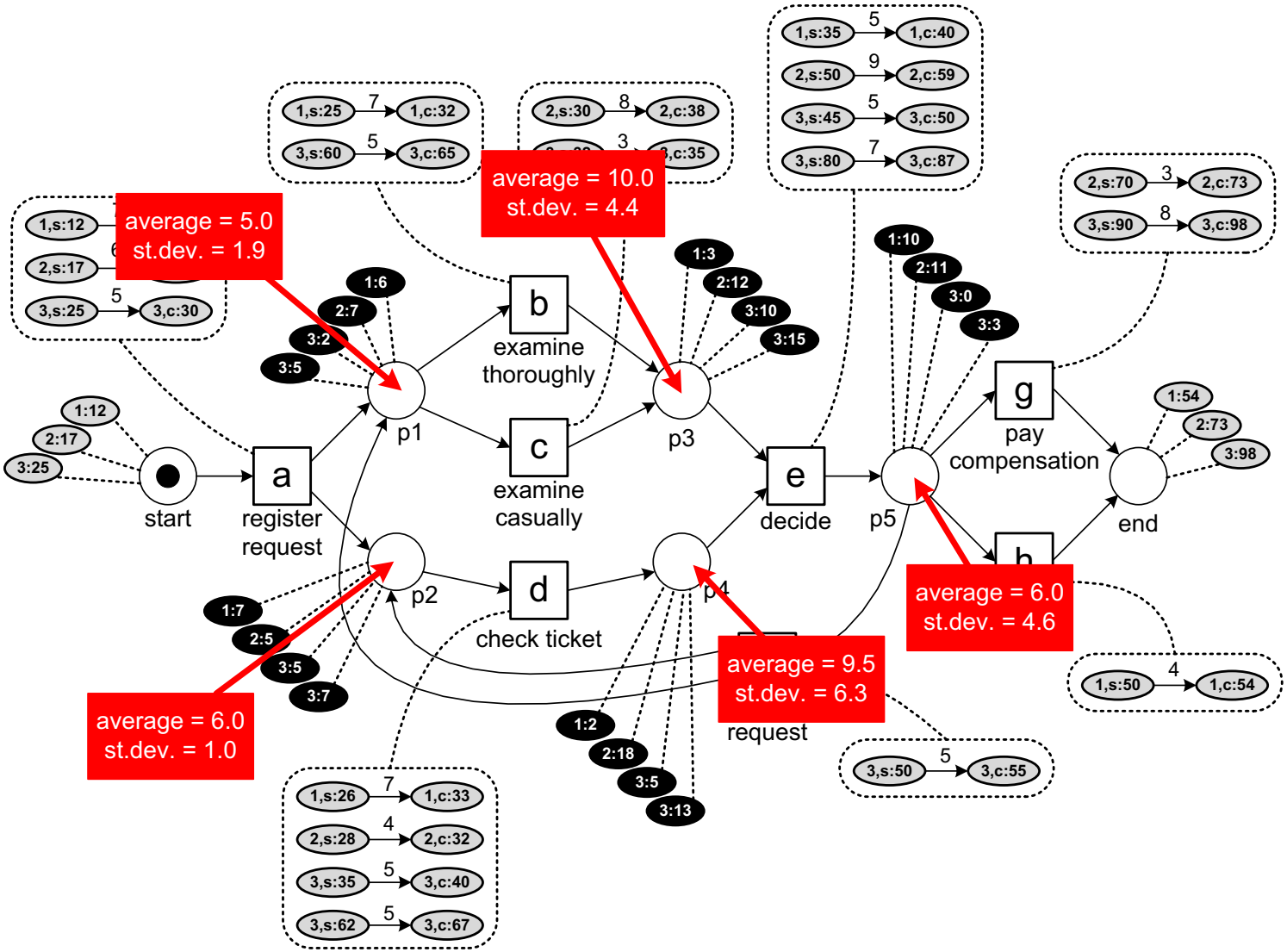
Timed replay projected onto resources
(activities colored by case)



case id	trace
1	$\langle a^{12}_{start}, a^{19}_{complete}, b^{25}_{start}, d^{26}_{start}, b^{32}_{complete}, d^{33}_{complete}, e^{35}_{start}, e^{40}_{complete}, h^{50}_{start}, h^{54}_{complete} \rangle$
2	$\langle a^{17}_{start}, a^{23}_{complete}, d^{28}_{start}, c^{30}_{start}, d^{32}_{complete}, c^{38}_{complete}, e^{50}_{start}, e^{59}_{complete}, g^{70}_{start}, g^{73}_{complete} \rangle$
3	$\langle a^{25}_{start}, a^{30}_{complete}, c^{32}_{start}, c^{35}_{complete}, d^{35}_{start}, d^{40}_{complete}, e^{45}_{start}, e^{50}_{complete}, f^{50}_{start}, f^{55}_{complete}, b^{60}_{start}, d^{62}_{start}, b^{65}_{complete}, d^{67}_{complete}, e^{80}_{start}, e^{87}_{complete}, g^{90}_{start}, g^{98}_{complete} \rangle$
...	...

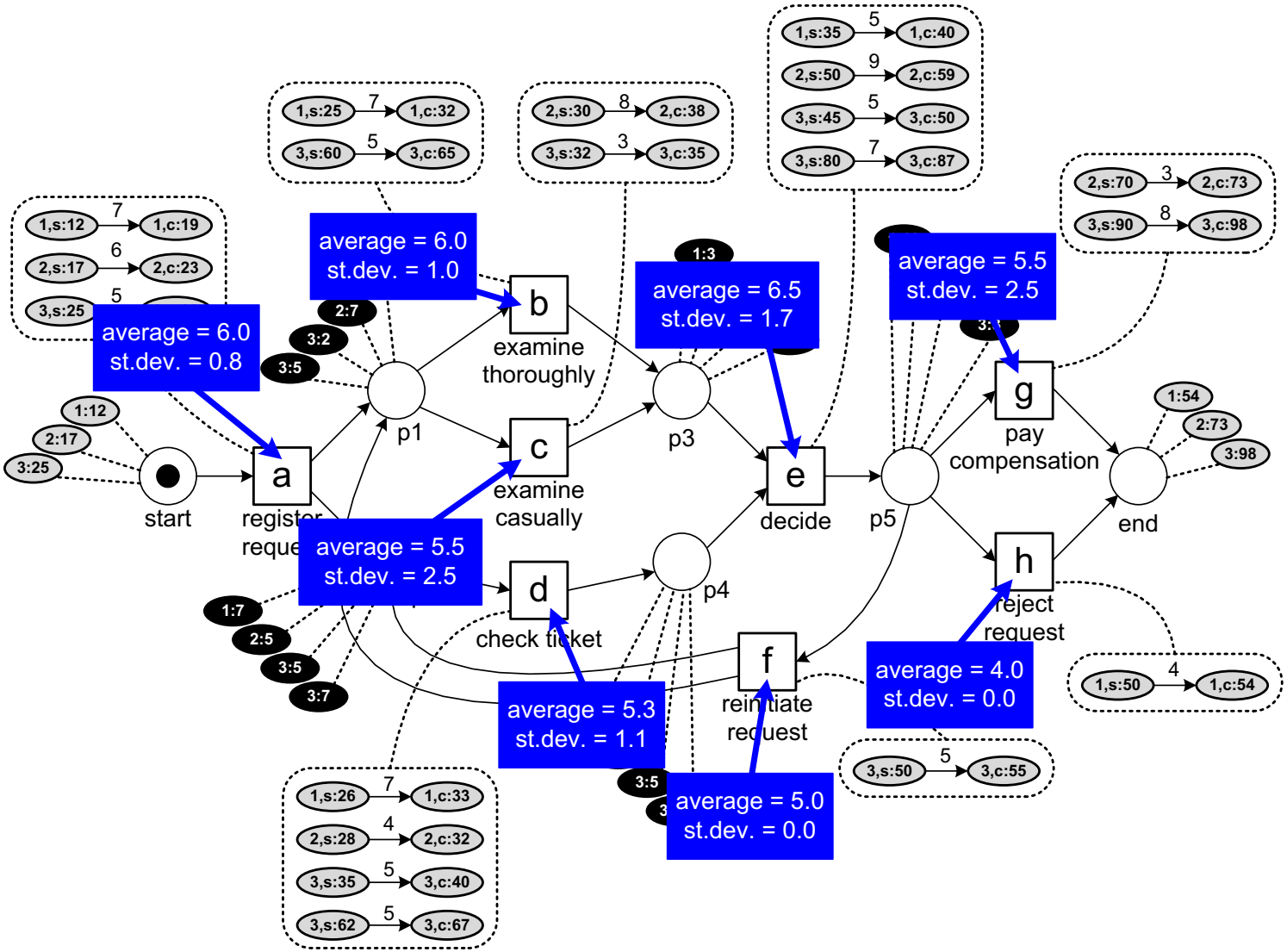
©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Waiting times



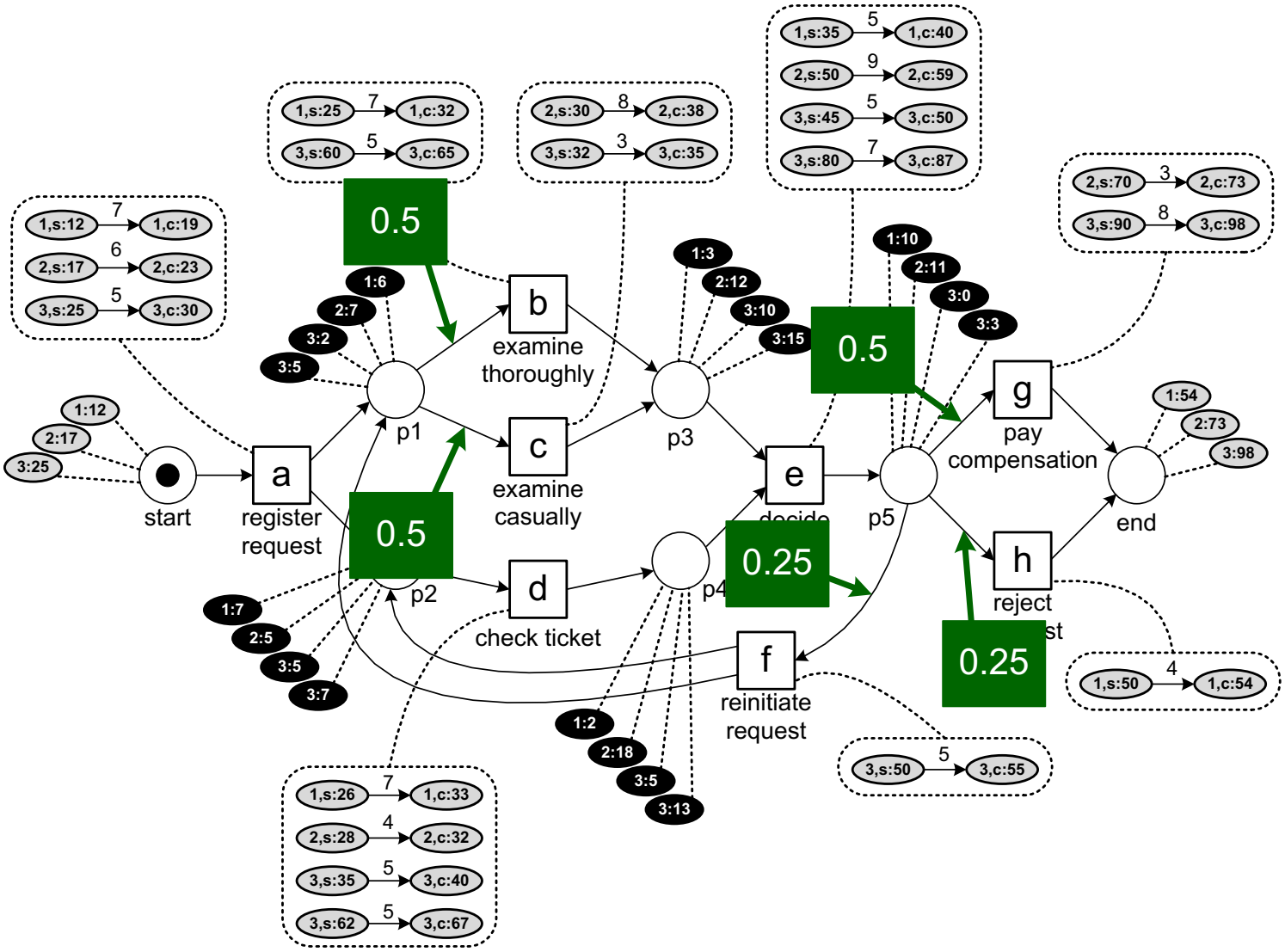
©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Service times



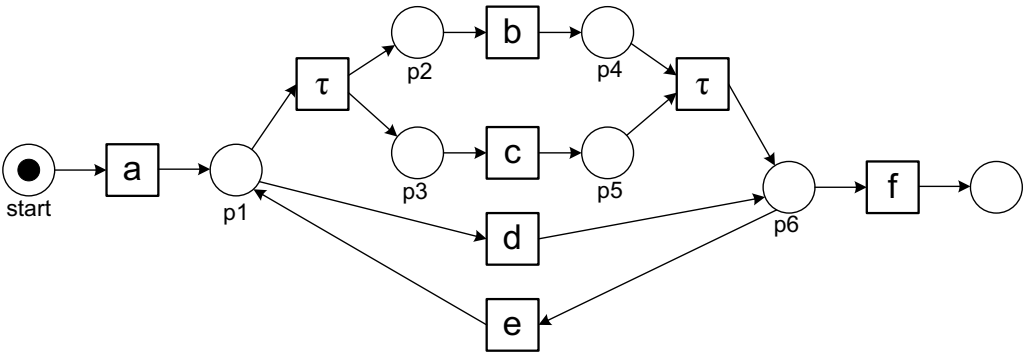
©Wil van der Aalst TU/e (use only with permission & acknowledgements)

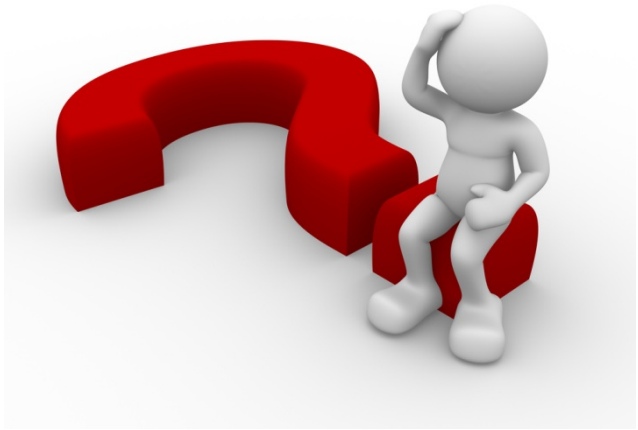
Routing probabilities



©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Estimate service times, waiting times, and routing probabilities

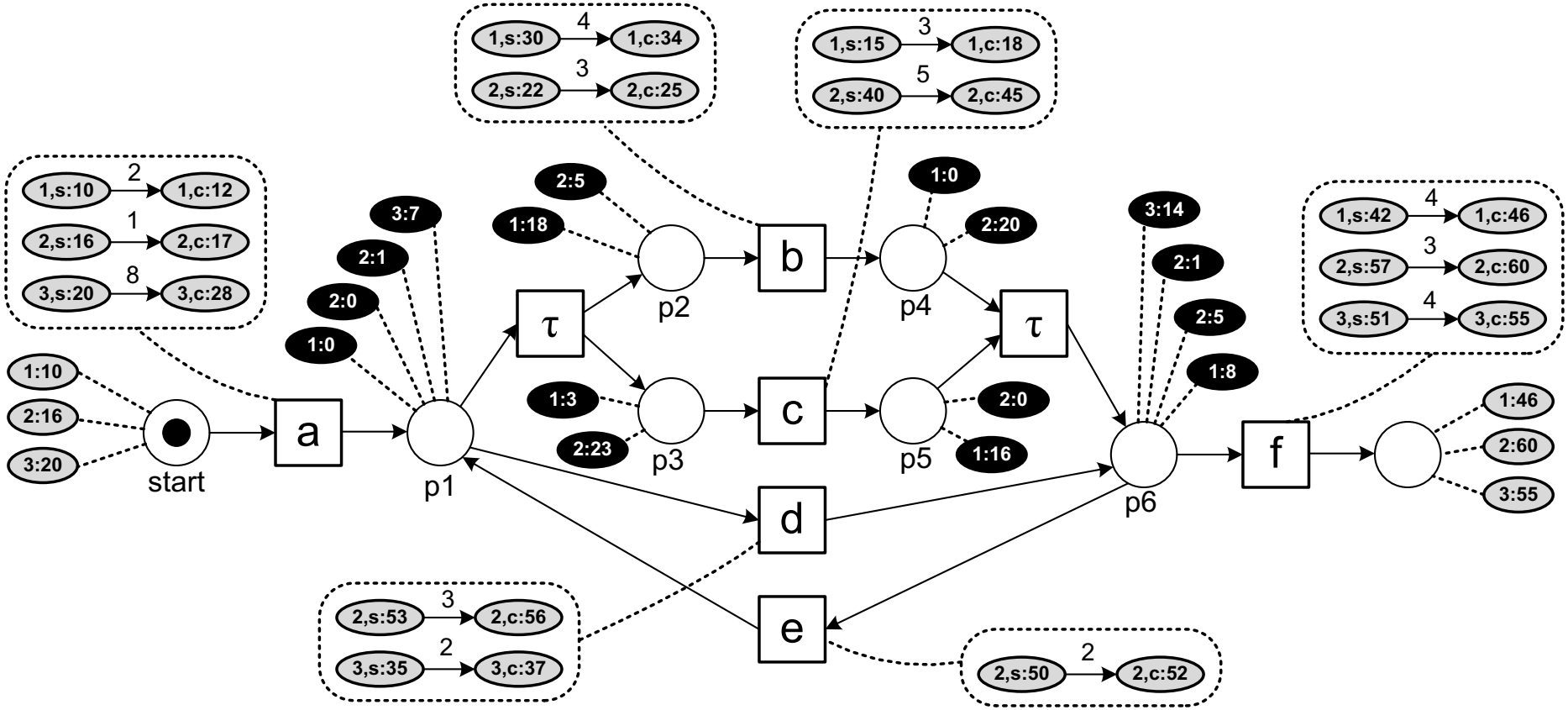




case id	activity	type	time	resource
1	a	start	10	Pete
1	a	complete	12	Pete
1	c	start	15	Sue
2	a	start	16	Pete
2	a	complete	17	Pete
1	c	complete	18	Sue
3	a	start	20	Pete
2	b	start	22	Mary
2	b	complete	25	Mary
3	a	complete	28	Pete
1	b	start	30	Mary
1	b	complete	34	Mary
3	d	start	35	Mary
3	d	complete	37	Mary
2	c	start	40	Sue
1	f	start	42	Carol
2	c	complete	45	Sue
1	f	complete	46	Carol
2	e	start	50	Kirsten
3	f	start	51	Carol
2	e	complete	52	Kirsten
2	d	start	53	Mary
3	f	complete	55	Carol
2	d	complete	56	Mary
2	f	start	57	Carol
2	f	complete	60	Carol

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

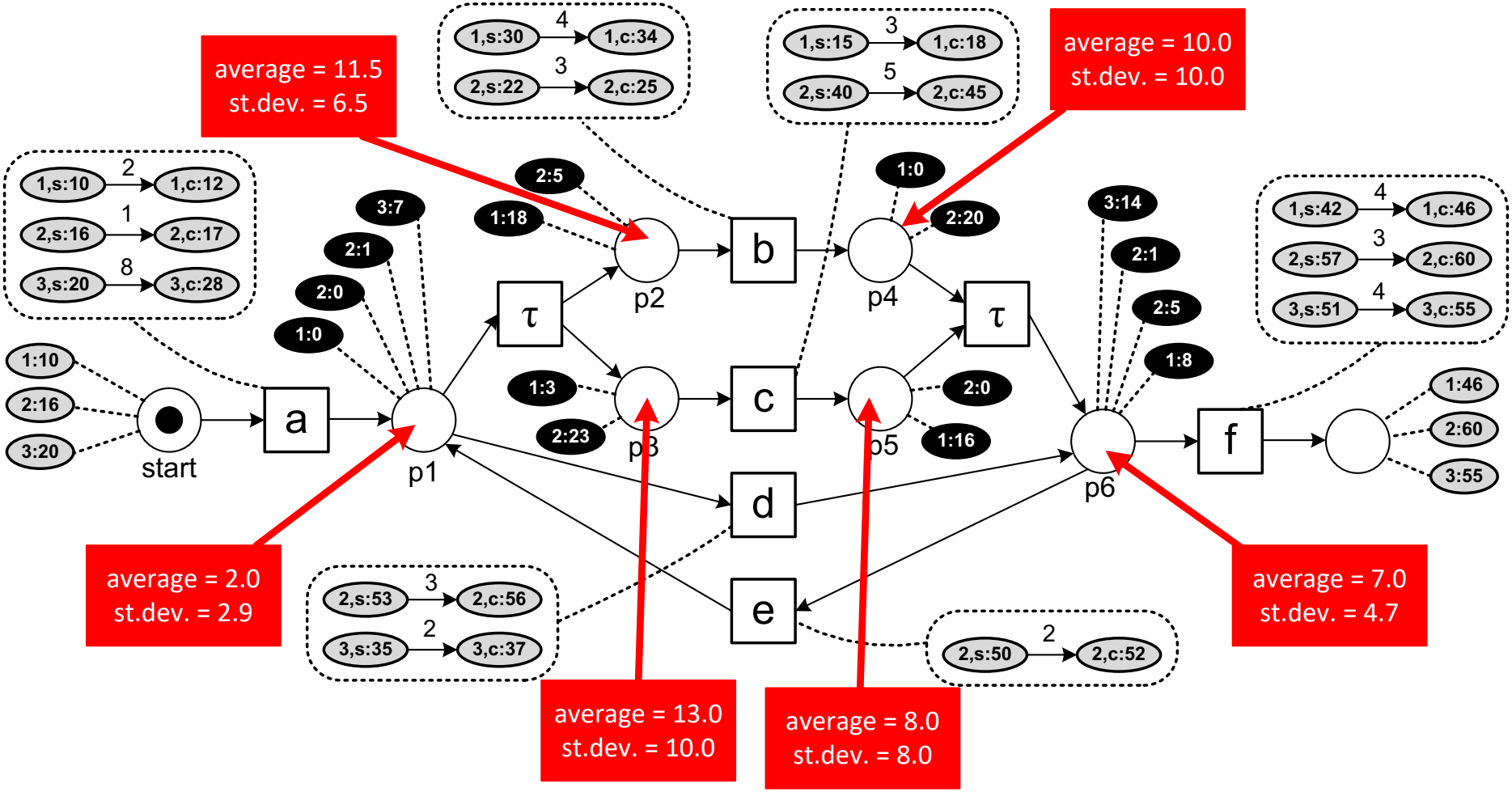
Times recorded during replay



case id	activity	type	time	resource
1	a	start	10	Pete
1	a	complete	12	Pete
1	c	start	15	Sue
2	a	start	16	Pete
2	a	complete	17	Pete

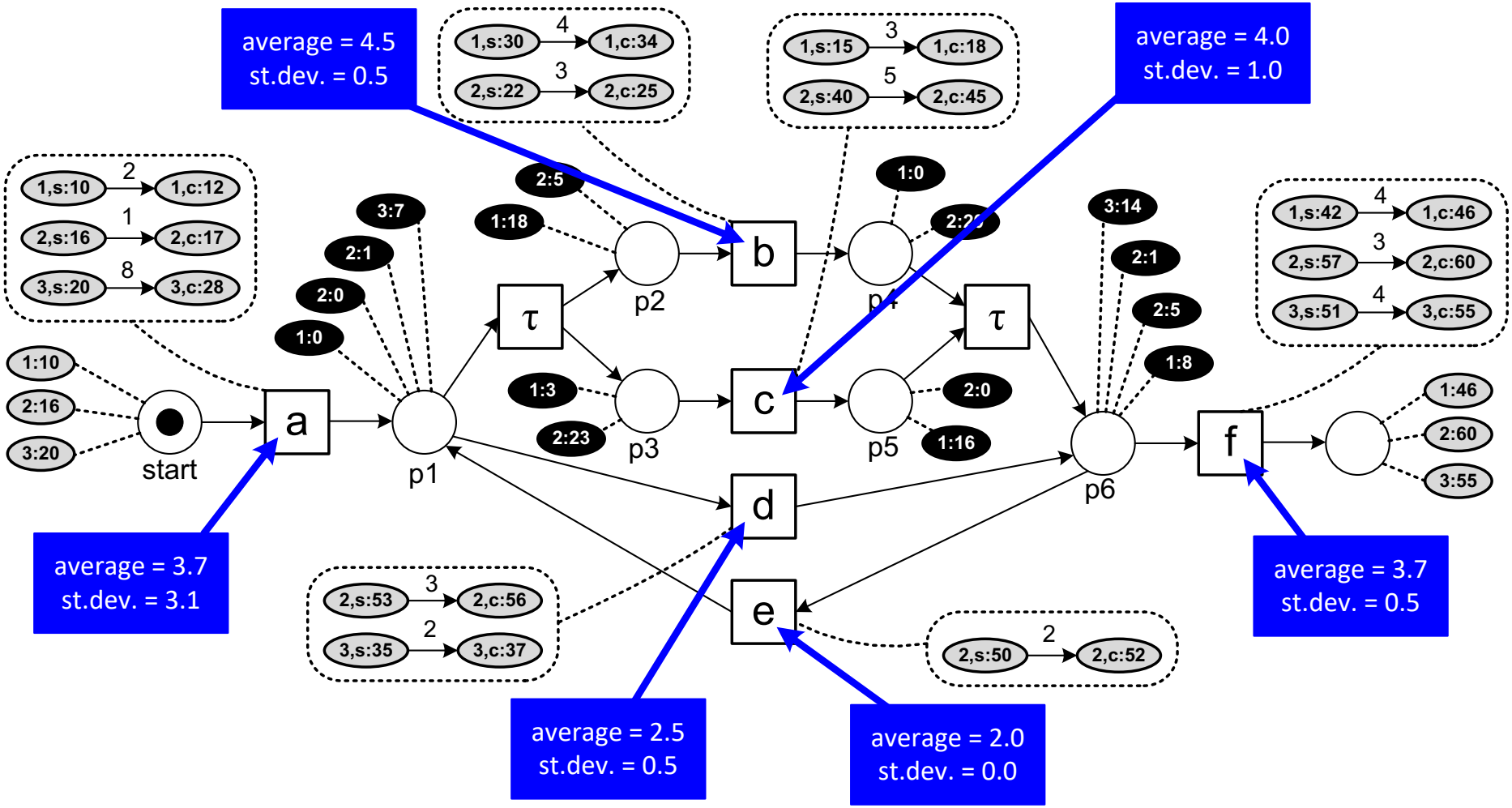
©Wil van der Aalst TU/e (use only with permission &

Waiting times



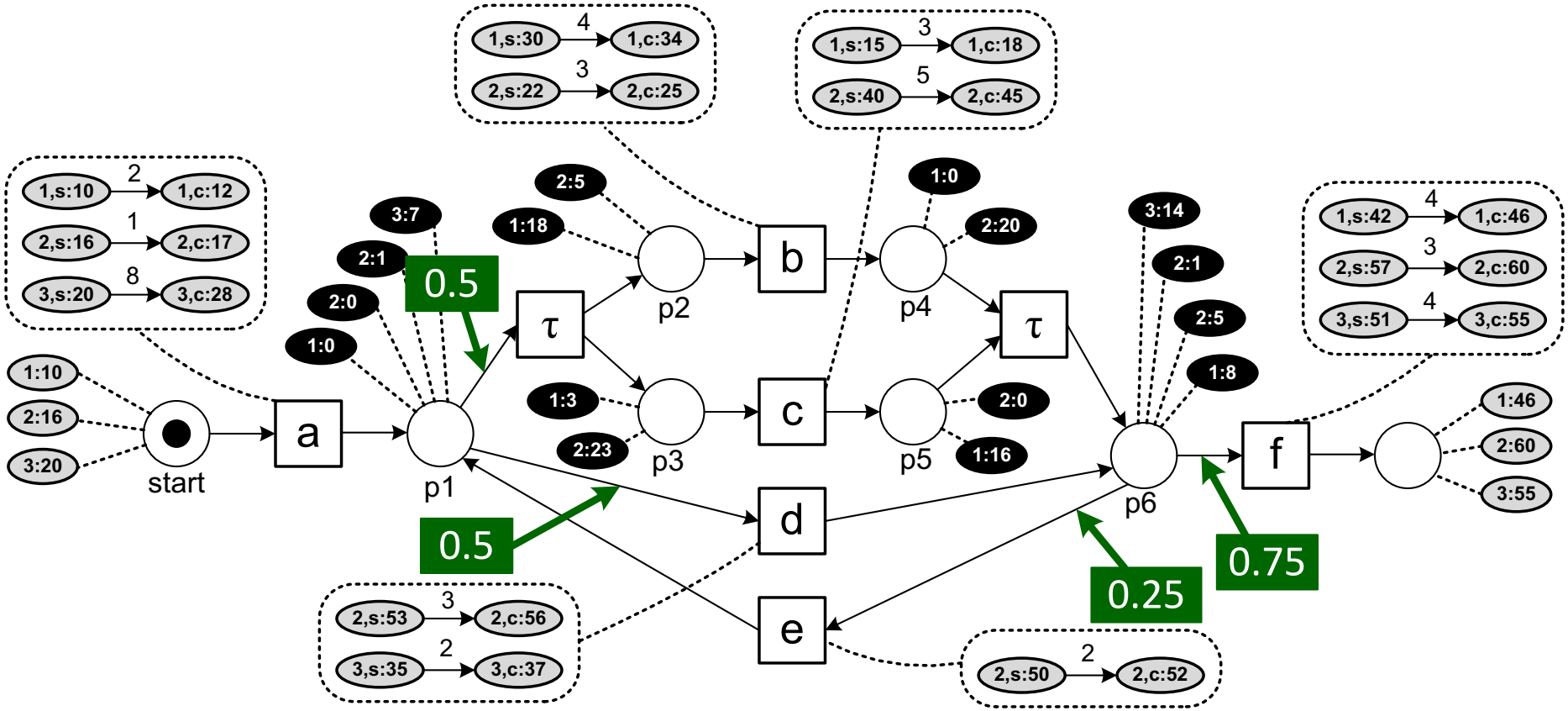
©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Service times



©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Routing probabilities



©Wil van der Aalst TU/e (use only with permission & acknowledgements)

So we can mine for bottlenecks and other performance related properties …

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

mining social networks

Organizational mining

case id	trace
1	$\langle a^{Pete}, b^{Sue}, d^{Mike}, e^{Sara}, h^{Pete} \rangle$
2	$\langle a^{Mike}, d^{Mike}, c^{Pete}, e^{Sara}, g^{Ellen} \rangle$
3	$\langle a^{Pete}, c^{Mike}, d^{Ellen}, e^{Sara}, f^{Sara}, b^{Sean}, d^{Pete}, e^{Sara}, g^{Ellen} \rangle$
4	$\langle a^{Pete}, d^{Mike}, b^{Sean}, e^{Sara}, h^{Ellen} \rangle$
5	$\langle a^{Ellen}, c^{Mike}, d^{Pete}, e^{Sara}, f^{Sara}, d^{Ellen}, c^{Mike}, e^{Sara}, f^{Sara}, b^{Sue}, d^{Pete}, e^{Sara}, h^{Mike} \rangle$
6	$\langle a^{Mike}, c^{Ellen}, d^{Mike}, e^{Sara}, g^{Mike} \rangle$
...	...

(a = *register request*, b = *examine thoroughly*, c = *examine casually*, d = *check ticket*, e = *decide*, f = *reinitiate request*, g = *pay compensation*, and h = *reject request*)

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Resource-activity matrix

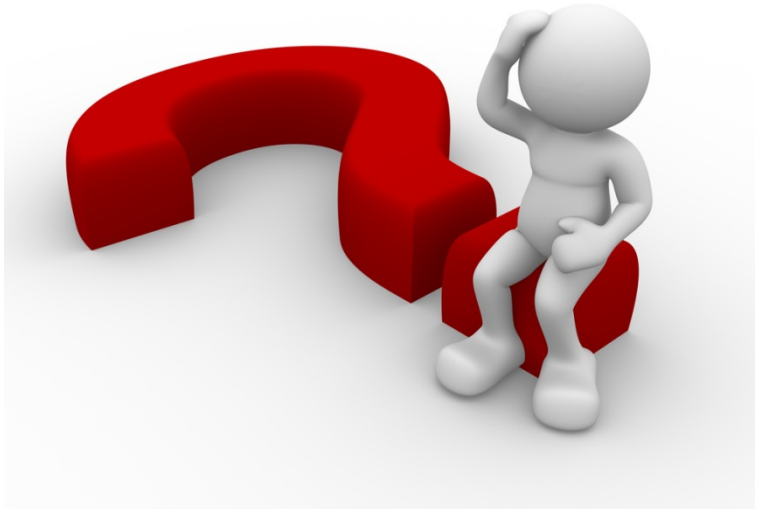
mean number of times a resource performs an activity per case

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*
Pete	0.3	0	0.345	0.69	0	0	0.135	0.165
Mike	0.5	0	0.575	1.15	0	0	0.225	0.275
Ellen	0.2	0	0.23	0.46	0	0	0.09	0.11
Sue	0	0.46	0	0	0	0	0	0
Sean	0	0.69	0	0	0	0	0	0
Sara	0	0	0	0	2.3	1.3	0	0

Activity a is executed exactly once for each case (hence the sum of the first column is 1). Pete, Mike, and Ellen are the only ones executing this activity. In 30% of the cases, a is executed by Pete, 50% is executed by Mike, and 20% is executed by Ellen. Activities e and f are always executed by Sara. Activity e is executed, on average, 2.3 times per case. Etc.

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Create resource-activity matrix



case id	activity	type	time	resource
1	a	start	10	Pete
1	a	complete	12	Pete
1	c	start	15	Sue
2	a	start	16	Pete
2	a	complete	17	Pete
1	c	complete	18	Sue
3	a	start	20	Pete
2	b	start	22	Mary
2	b	complete	25	Mary
3	a	complete	28	Pete
1	b	start	30	Mary
1	b	complete	34	Mary
3	d	start	35	Mary
3	d	complete	37	Mary
2	c	start	40	Sue
1	f	start	42	Carol
2	c	complete	45	Sue
1	f	complete	46	Carol
2	e	start	50	Kirsten
3	f	start	51	Carol
2	e	complete	52	Kirsten
2	d	start	53	Mary
3	f	complete	55	Carol
2	d	complete	56	Mary
2	f	start	57	Carol
2	f	complete	60	Carol

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Resource-activity matrix

case id	activity	type	time	resource
1	a	start	10	Pete
1	a	complete	12	Pete
1	c	start	15	Sue
2	a	start	16	
2	a	complete	17	
1	c	complete	18	
3	a	start	20	
2	b	start	22	
2	b	complete	25	
3	a	complete	28	
1	b	start	30	
1	b	complete	34	
3	d	start	35	
3	d	complete	37	Mary
2	c	start	40	Sue
1	f	start	42	Carol
2	c	complete	45	Sue
1	f	complete	46	Carol
2	e	start	50	Kirsten
3	f	start	51	Carol
2	e	complete	52	Kirsten
2	d	start	53	Mary
3	f	complete	55	Carol
2	d	complete	56	Mary
2	f	start	57	Carol
2	f	complete	60	Carol

	a	b	c	d	e	f
Pete	1.00	0.00	0.00	0.00	0.00	0.00
Mary	0.00	0.67	0.00	0.67	0.00	0.00
Sue	0.00	0.00	0.67	0.00	0.00	0.00
Kirsten	0.00	0.00	0.00	0.00	0.33	0.00
Carol	0.00	0.00	0.00	0.00	0.00	1.00

mean number of times a resource performs an activity per case

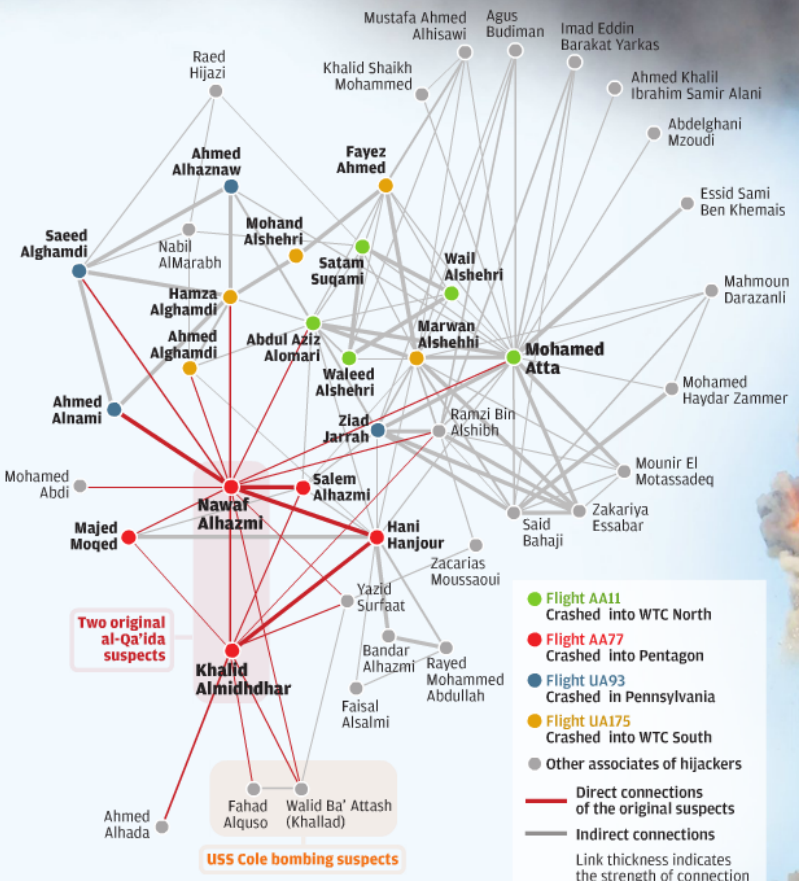
Social network analysis

How two names and a sheaf of newspaper cuttings revealed the 9/11 team

This social network of the 19 hijackers behind the 9/11 attacks in the United States, and their associates, was drawn up at the end of 2001. Valdis Krebs, a commercial consultant in network analysis, started with newspaper reports of the two original terrorist suspects, Nawaf Alhazmi and Khalid Almihdhar. He then plotted the position of the other hijackers and associates. His analysis highlighted the central role played by Mohamed Atta. It also shows the close associations between the "Hamburg cell" that Atta set up, as well as the close links with the two original suspects – critical information that may have helped to avert an attack had it been known.

Mohamed Atta | Nawaf Alhazmi | Khalid Almidhdhar

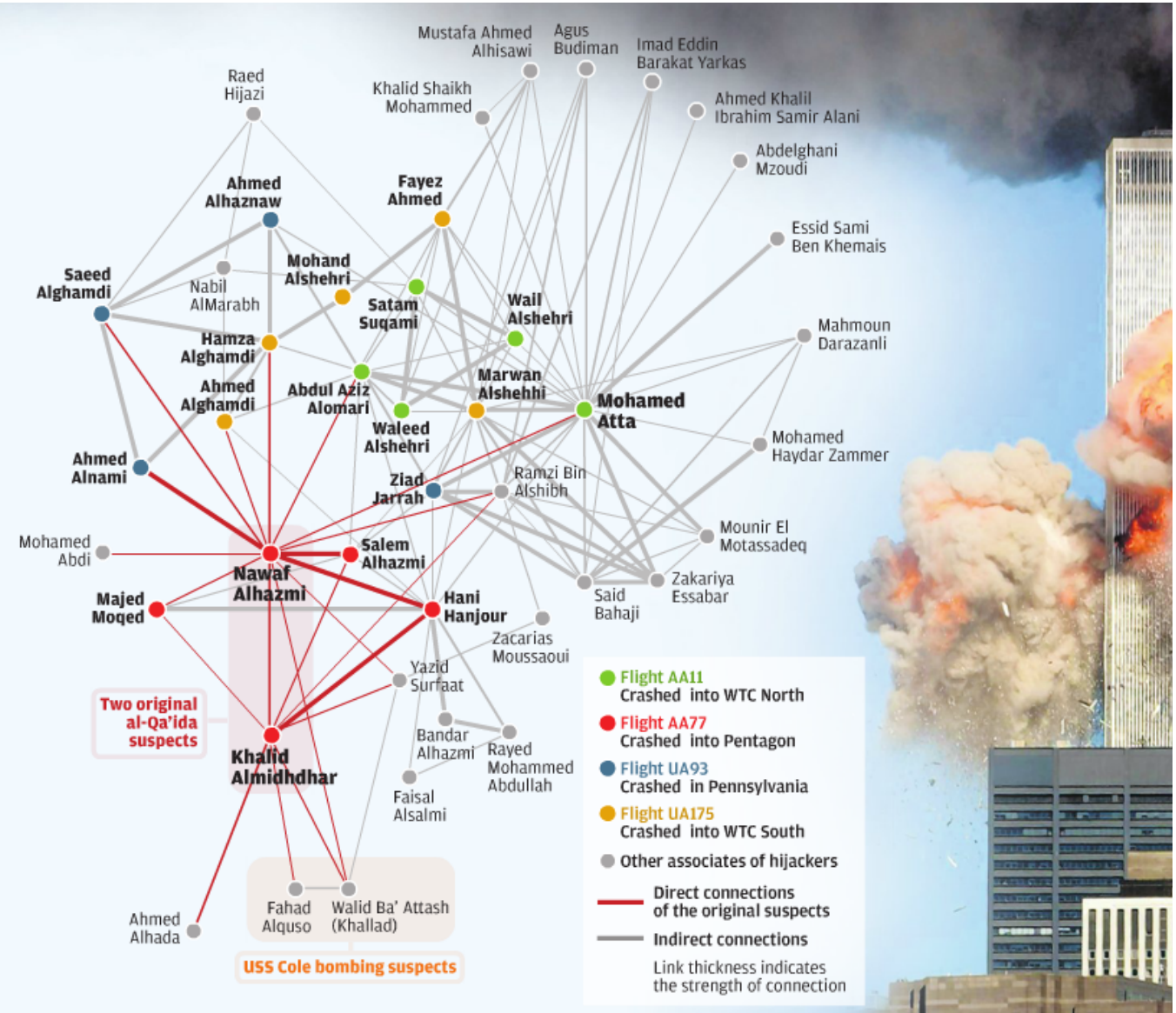
Emergency services attend the scene after Flight AA77 crashes into the Pentagon



The second plane, Flight UA175, crashes into the South Tower of the World Trade Centre

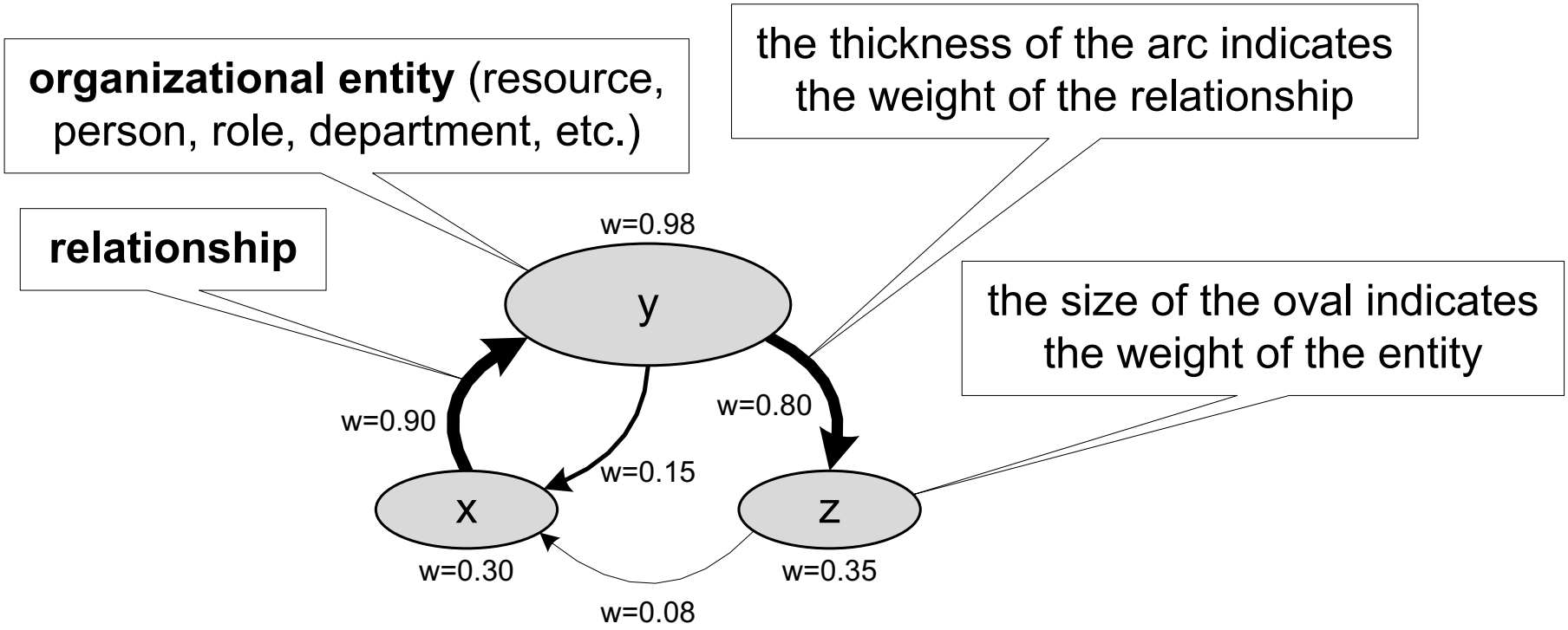
©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Social network analysis



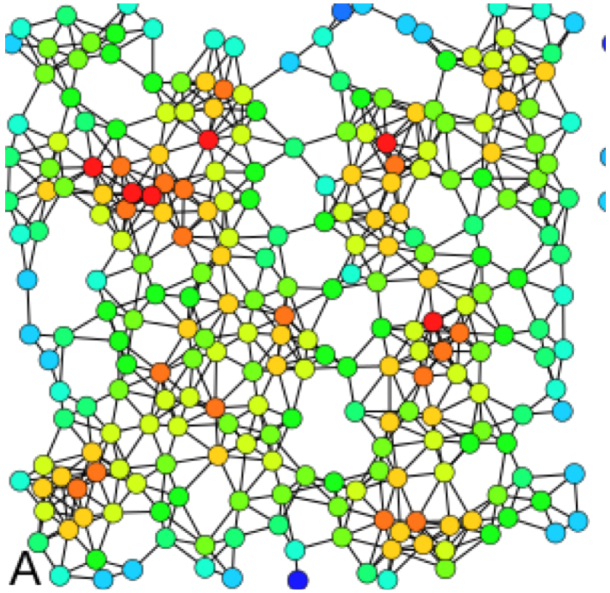
- *Sociometry*: present data on interpersonal relationships in graph or matrix form.
- Jacob Levy Moreno used such techniques in the 1930s to better assign students to residential cottages.
- Arcs: weights or (inverted) distance.
- Metrics to denote importance:
  - centrality,
  - closeness,
  - betweenness.
  - …
- Identification of cliques.

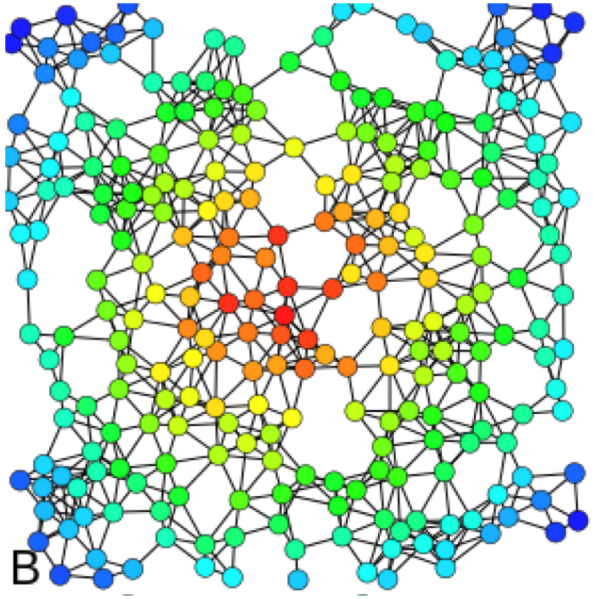
Social network

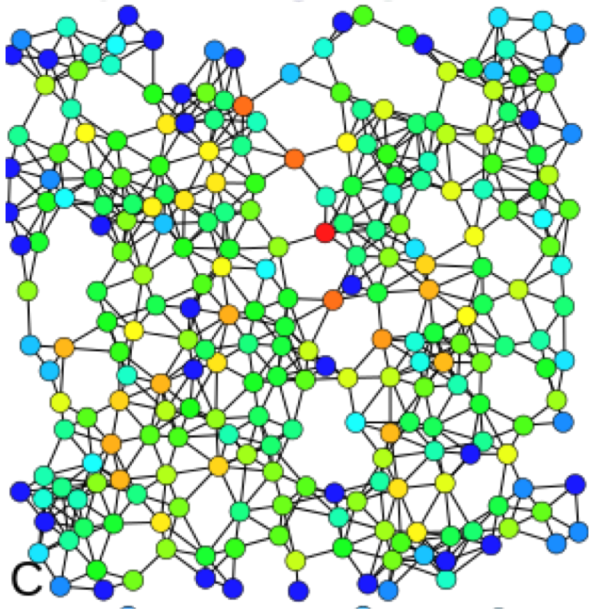


©Wil van der Aalst TU/e (use only with permission & acknowledgements)

"Importance" of nodes in a social network







degree centrality: number of connections a particular node has

closeness centrality: 1 divided by the sum of all shortest paths to a particular node

betweenness centrality: fraction of shortest paths between any two nodes passing a particular node

A wide variety of definitions exist for "importance".

Figures by Claudio Rocchini

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Handover of work matrix

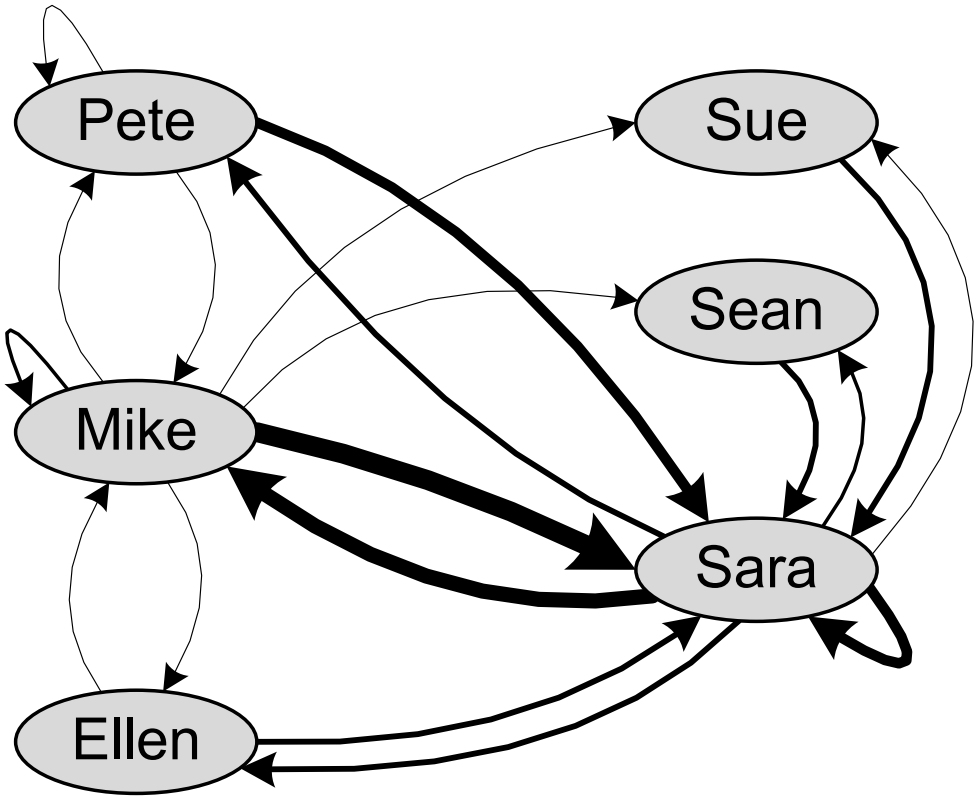
	Pete	Mike	Ellen	Sue	Sean	Sara
Pete	0.135	0.225	0.09	0.06	0.09	1.035
Mike	0.225	0.375	0.15	0.1	0.15	1.725
Ellen	0.09	0.15	0.06	0.04	0.06	0.69
Sue	0	0	0	0	0	0.46
Sean	0	0	0	0	0	0.69
Sara	0.885	1.475	0.59	0.26	0.39	1.3

Count the number of times work is handed over from one resource to another (on average per case).

The causal dependencies in the process model are used to count handovers in the event log.

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Social network based on handover of work (threshold of 0.1)



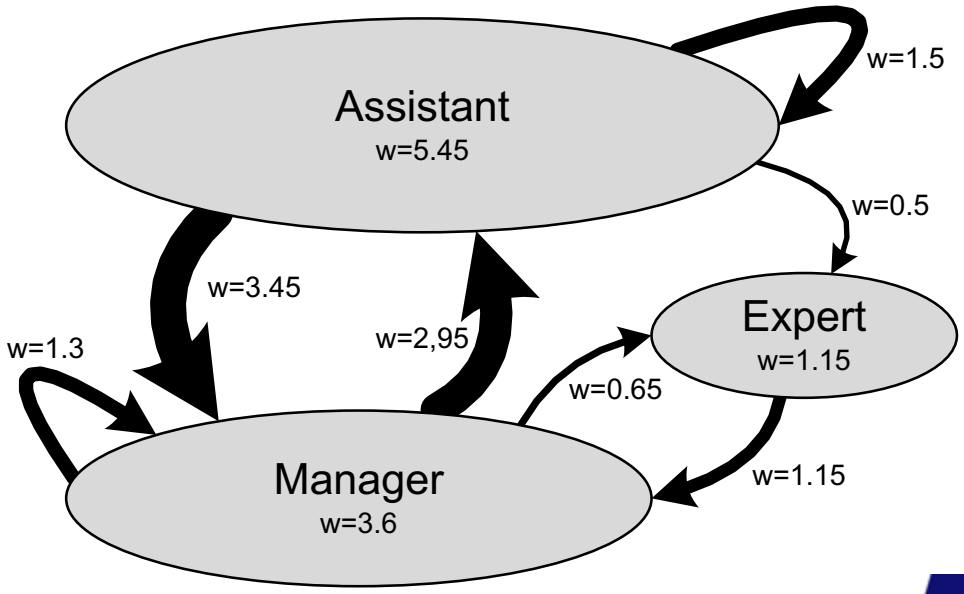
In this figure only the thickness of the arcs is based on frequencies.

	Pete	Mike	Ellen	Sue	Sean	Sara
Pete	0.135	0.225	0.09	0.06	0.09	1.035
Mike	0.225	0.375	0.15	0.1	0.15	1.725
Ellen	0.09	0.15	0.06	0.04	0.06	0.69
Sue	0	0	0	0	0	0.46
Sean	0	0	0	0	0	0.69
Sara	0.885	1.475	0.59	0.26	0.39	1.3

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Handover of work at role level

	Assistant	Expert	Manager
Assistant	1.5	0.5	3.45
Expert	0	0	1.15
Manager	2.95	0.65	1.3

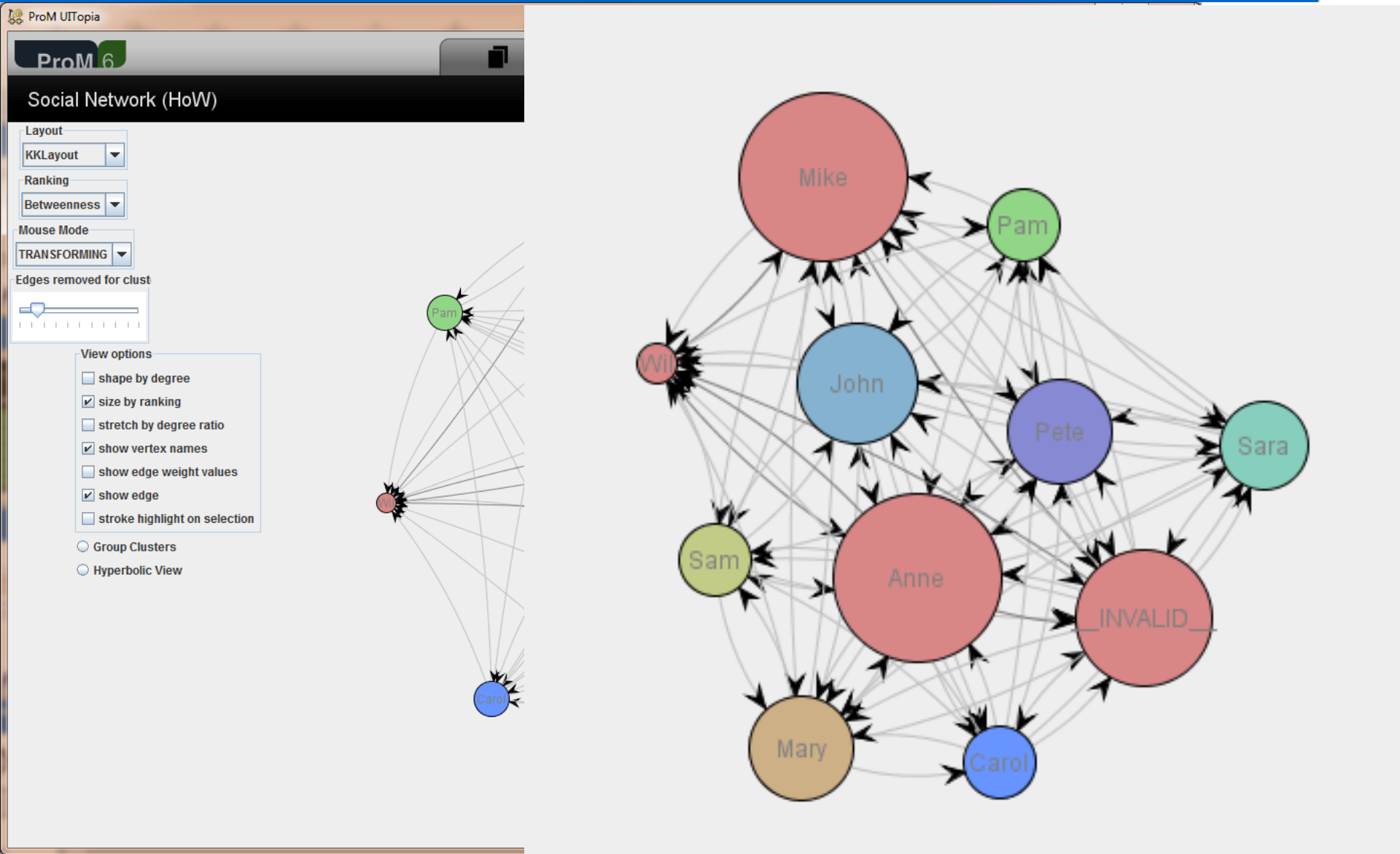


In this figure also the size of each node is based on frequencies.

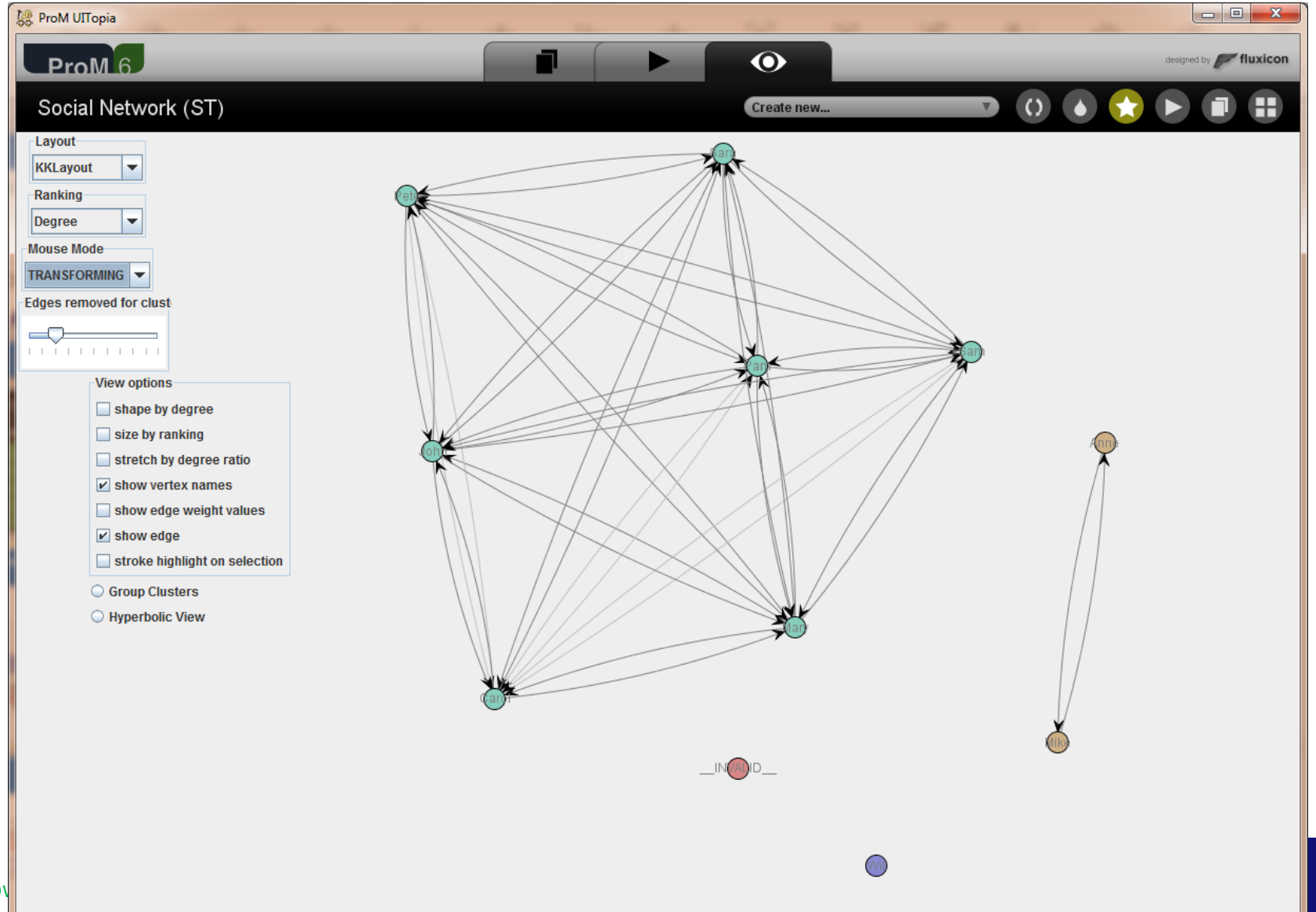
©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Social network miner in ProM

Social network based on hand-over of work



Social network based on similar tasks



Create handover of work matrix

case id	activity	type	time	resource
1	a	start	10	Pete
1	a	complete	12	Pete
1	c	start	15	Sue
2	a	start	16	Pete
2	a	complete	17	Pete
1	c	complete	18	Sue
3	a	start	20	Pete
2	b	start	22	Mary
2	b	complete	25	Mary
3	a	complete	28	Pete
1	b	start	30	Mary
1	b	complete	34	Mary
3	d	start	35	Mary
3	d	complete	37	Mary
2	c	start	40	Sue
1	f	start	42	Carol
2	c	complete	45	Sue
1	f	complete	46	Carol
2	e	start	50	Kirsten
3	f	start	51	Carol
2	e	complete	52	Kirsten
2	d	start	53	Mary
3	f	complete	55	Carol
2	d	complete	56	Mary
2	f	start	57	Carol
2	f	complete	60	Carol

- **Count the number of times work is handed over from one resource to another (on average per case).**
- **Ignore process model, no need to consider causalities and concurrency.**

Handover of work matrix

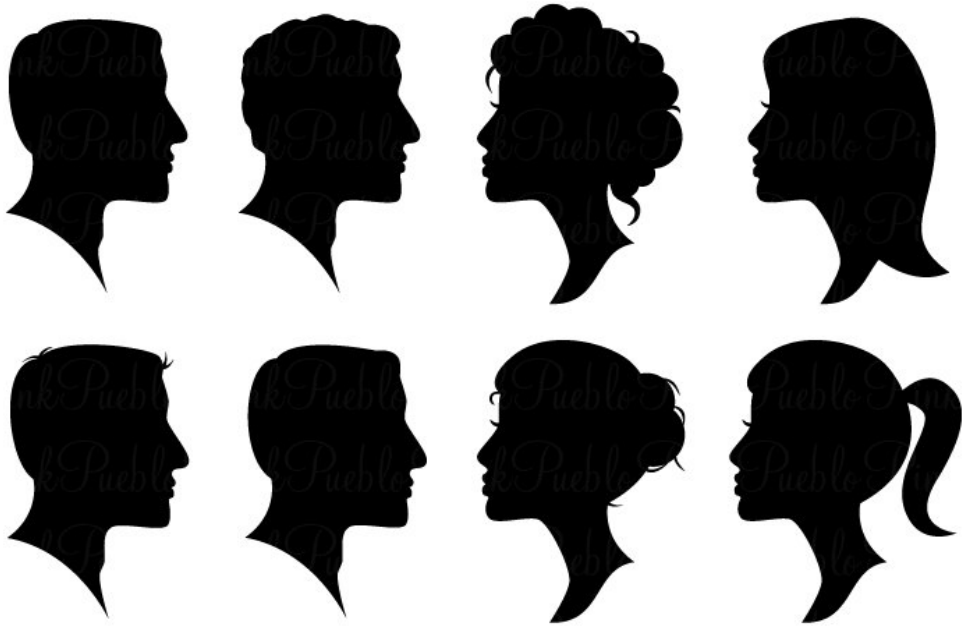
case id	activity	type	time	resource
1	a	start	10	Pete
1	a	complete	12	Pete
1	c	start	15	Sue
2	a	start	16	Pete
2	a	complete	17	Pete
1	c	complete	18	Sue
3	a	start	20	Pete
2	b	start	22	Mary
2	b	complete	25	Mary
3	a	complete	28	Pete
1	b	start	30	Mary
1	b	complete	34	Mary
3	d	start	35	Mar
3	d	complete	37	Mar
2	c	start	40	Sue
1	f	start	42	Car
2	c	complete	45	Sue
1	f	complete	46	Car
2	e	start	50	Kirst
3	f	start	51	Car
2	e	complete	52	Kirst
2	d	start	53	Mar
3	f	complete	55	Carol
2	d	complete	56	Mary
2	f	start	57	Carol
2	f	complete	60	Carol

1. **Pete, Sue, Mary, Carol**
2. **Pete, Mary, Sue, Kirsten, Mary, Carol**
3. **Pete, Mary, Carol**

	Pete	Mary	Sue	Kirsten	Carol
Pete	0.00	0.66	0.33	0.00	0.00
Mary	0.00	0.00	0.33	0.00	1.00
Sue	0.00	0.33	0.00	0.33	0.00
Kirsten	0.00	0.33	0.00	0.00	0.00
Carol	0.00	0.00	0.00	0.00	0.00

Profiles based on resource-activity matrix

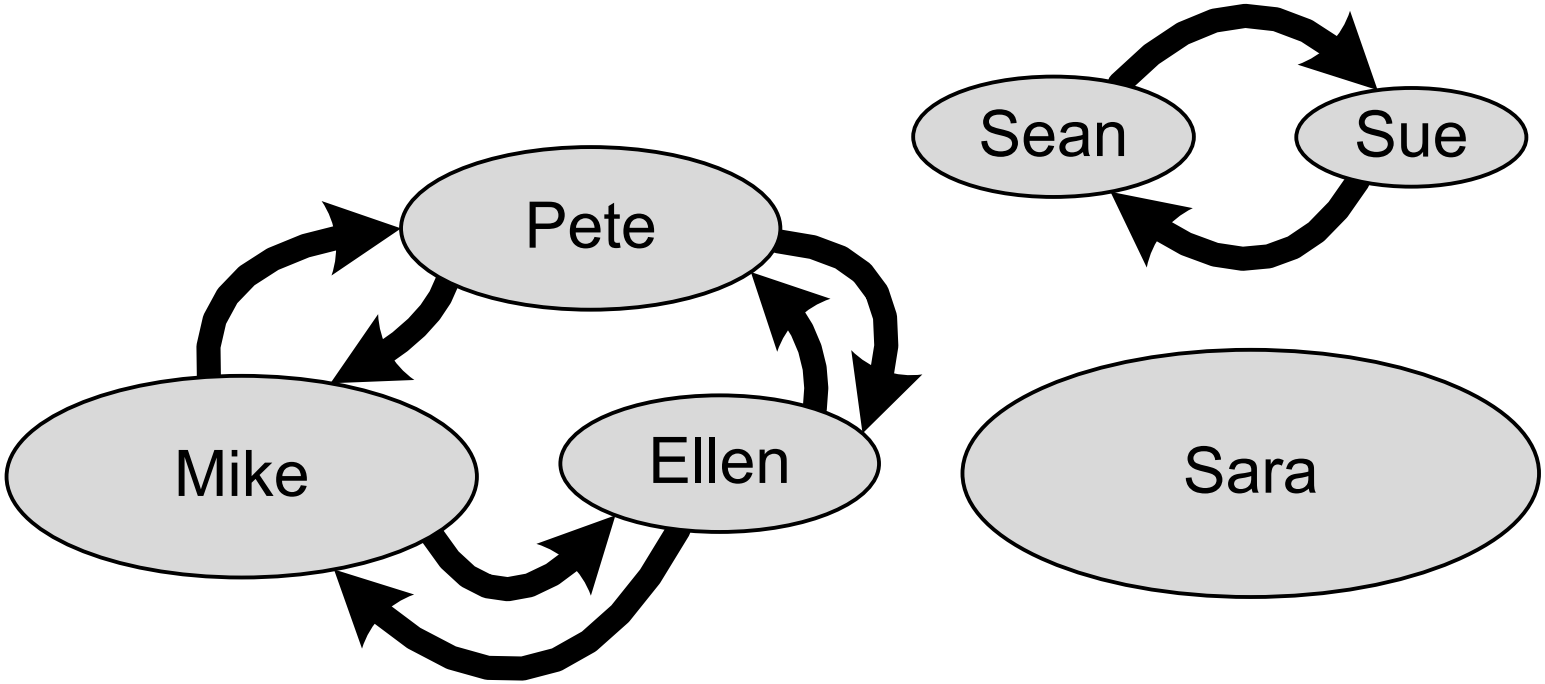
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*
Pete	0.3	0	0.345	0.69	0	0	0.135	0.165
Mike	0.5	0	0.575	1.15	0	0	0.225	0.275
Ellen	0.2	0	0.23	0.46	0	0	0.09	0.11
Sue	0	0.46	0	0	0	0	0	0
Sean	0	0.69	0	0	0	0	0	0
Sara	0	0	0	0	2.3	1.3	0	0



find the three profiles

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Social network based on similarity of profiles



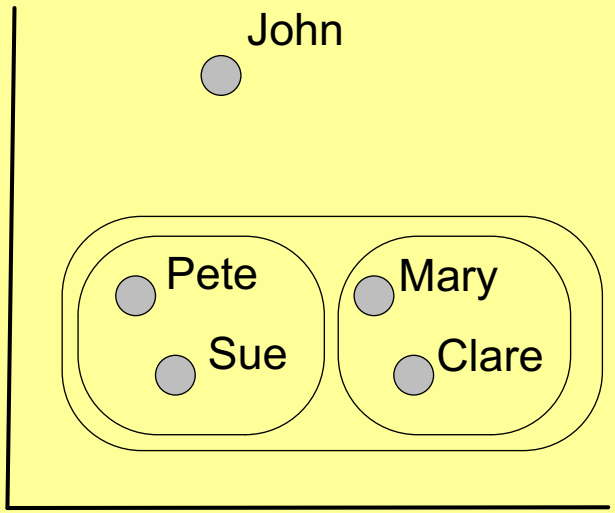
Resources that execute similar collections of activities are related. Sara is the only resource executing e and f. Therefore, she is not connected to other resources. Self-loops are suppressed as they contain no information (self-similarity).

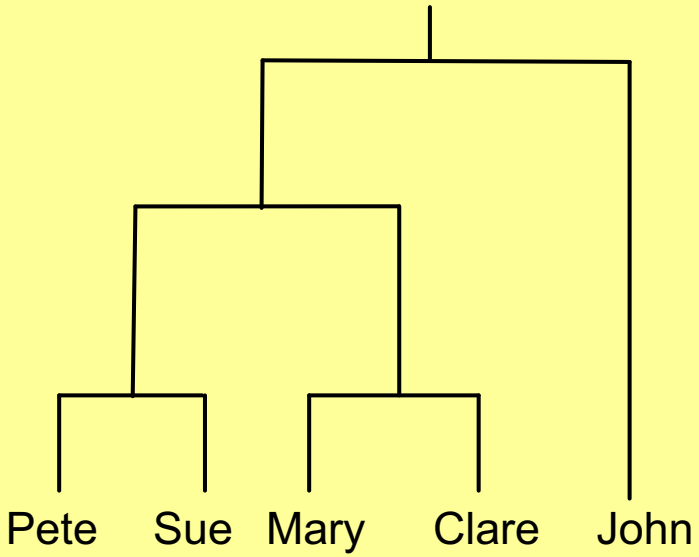
©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Clustering of resources

resource name	average nof times a was executed	average nof times b was executed	average nof times c was executed
Pete	0.15	0.00	0.05
Sue	0.16	0.00	0.04
Mary	0.07	0.01	0.14
Clare	0.09	0.02	0.13

feature vector





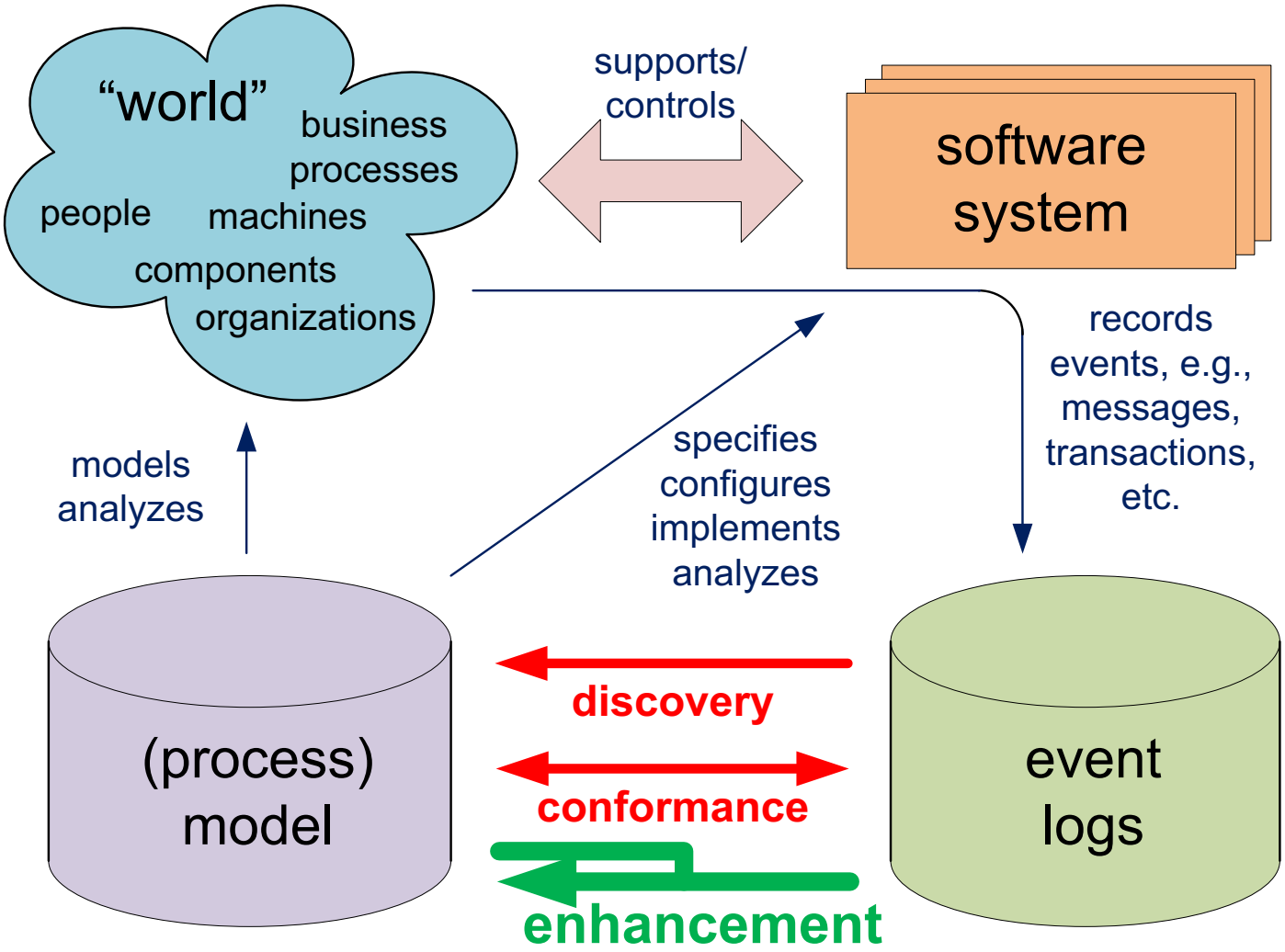
©W

So we can mine social networks based on handover of work matrix or resource-activity matrix …

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

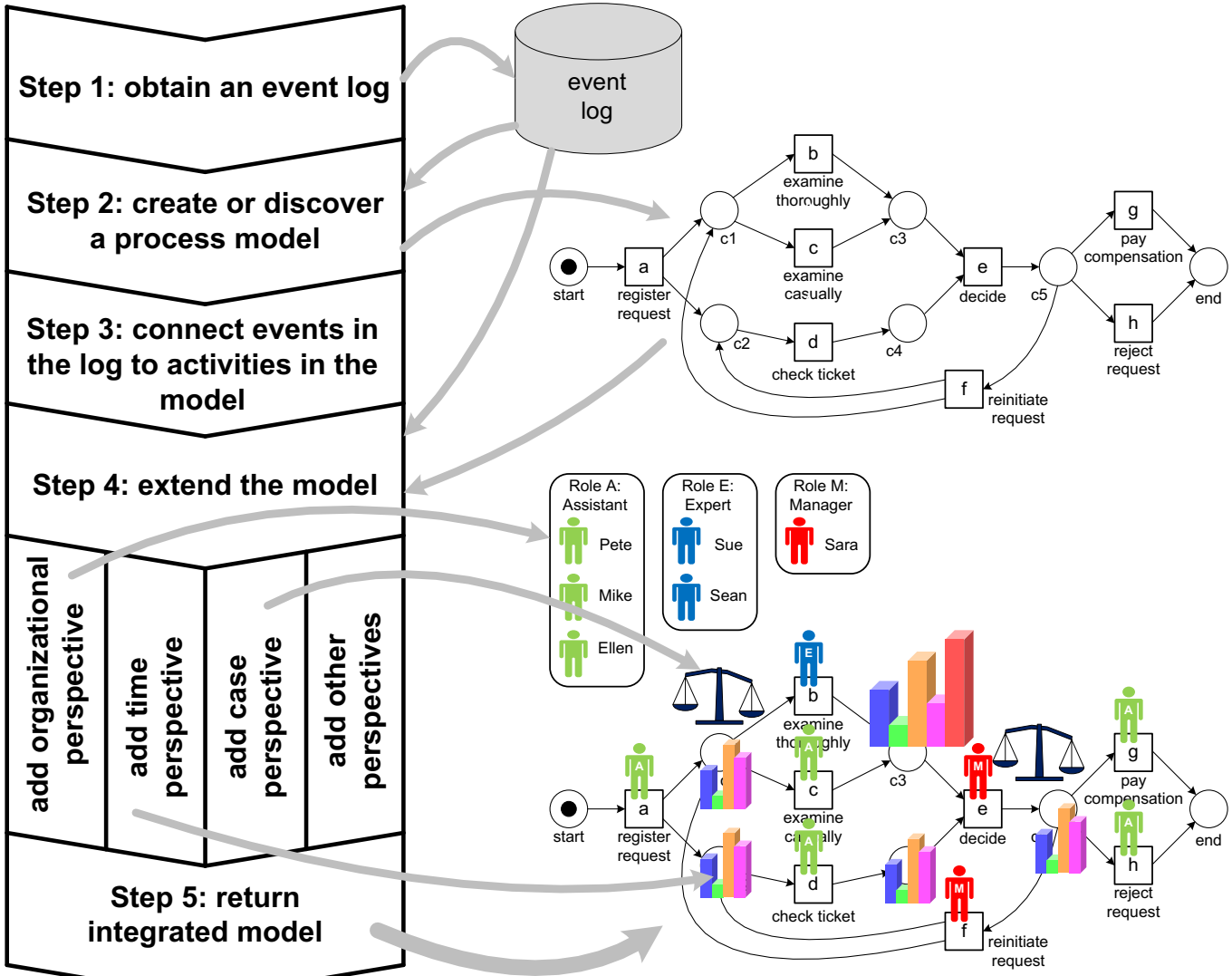
conclusion
and outlook

Enhancement: making better /more useful process models



©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Bringing it all together



High-level languages (e.g., BPMN) cover all these perspectives.

Even possible to automatically learn simulation models!

©Wil van der Aalst TU/e (use only with permission & acknowledgements)

Take Away

Process models can be repaired, enhanced to provide a better understanding of the underlying process

Decisions, performance, work collaboration, etc, ... can be mined...

Nice link between data mining and process mining