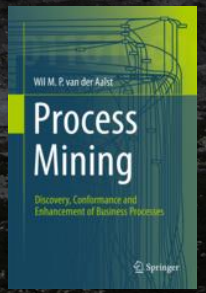
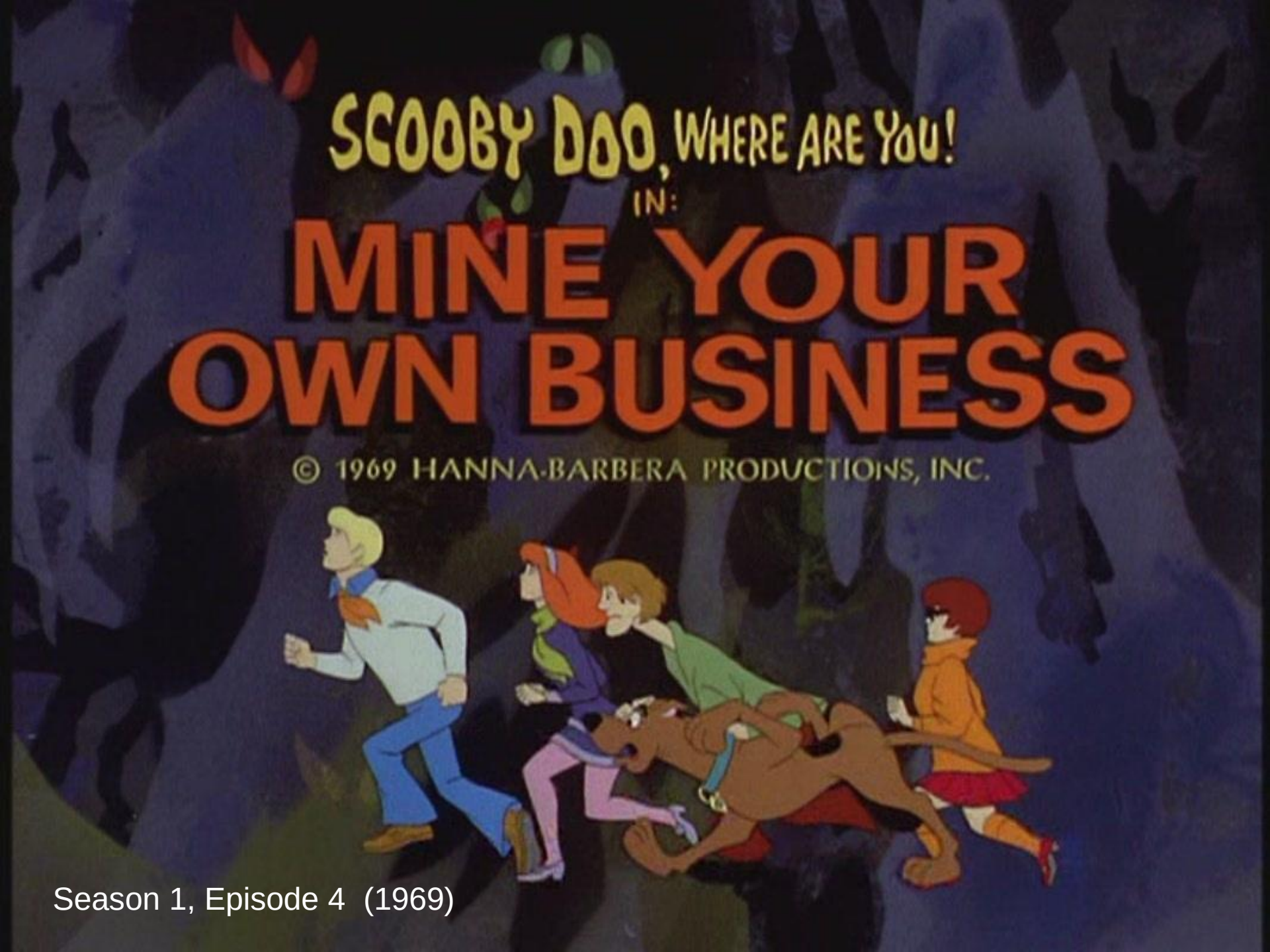
Mine Your Own Business

Using Process Mining to Turn Big Data into Real Value

prof.dr.ir. Wil van der Aalst

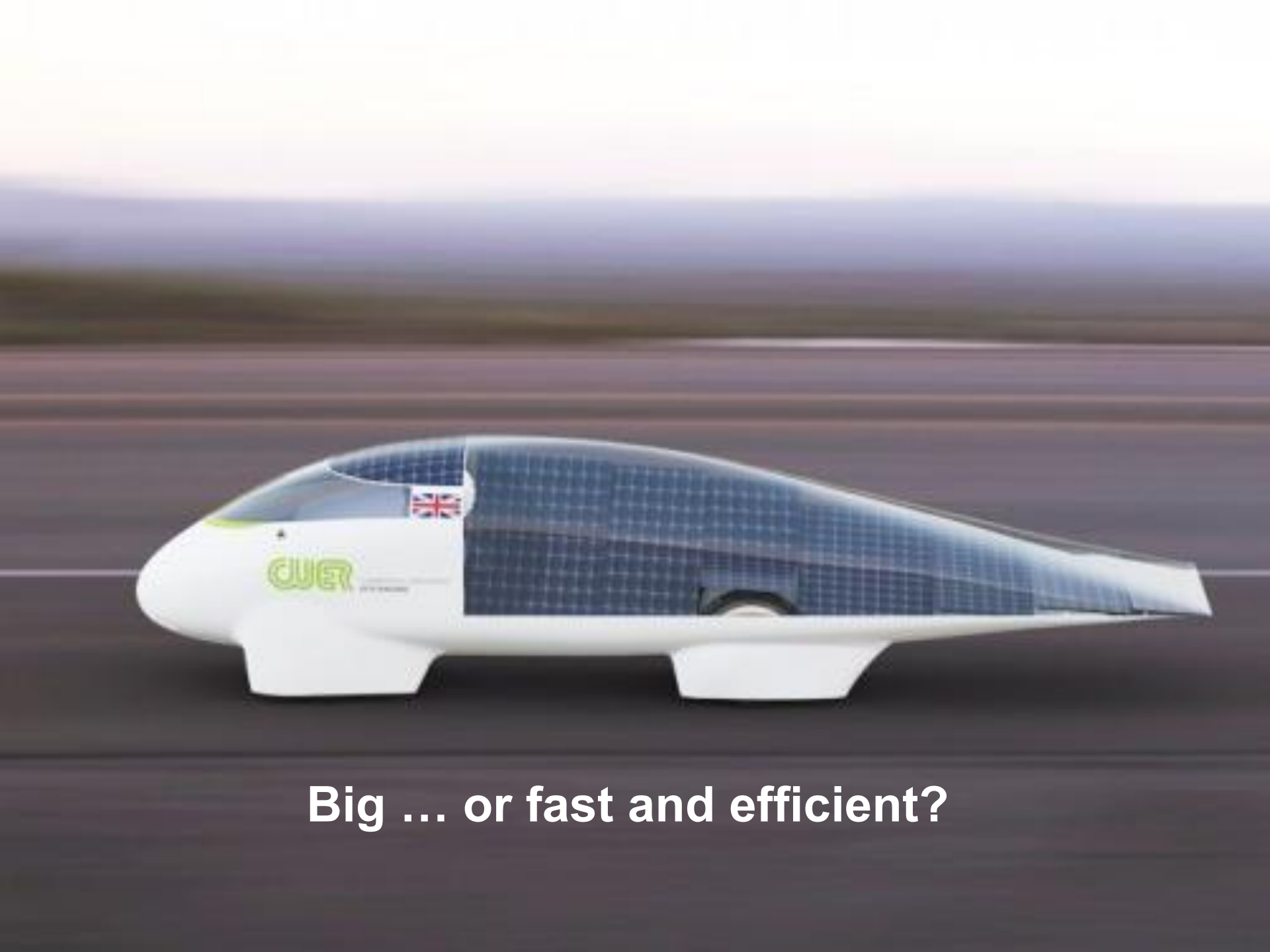
ECIS 2013 Utrecht





Season 1, Episode 4 (1969)


U.S. AIR FORCE
AFTERBURNER
Cross Into The Blue
U.S. AIR FORCE
1-800-423-USAF
AIRFORCE.COM
Big Data ?


Big … or fast and efficient?


LinkedIn
foursquare
skype
YouTube
facebook
twitter
tumblr.
Blogger

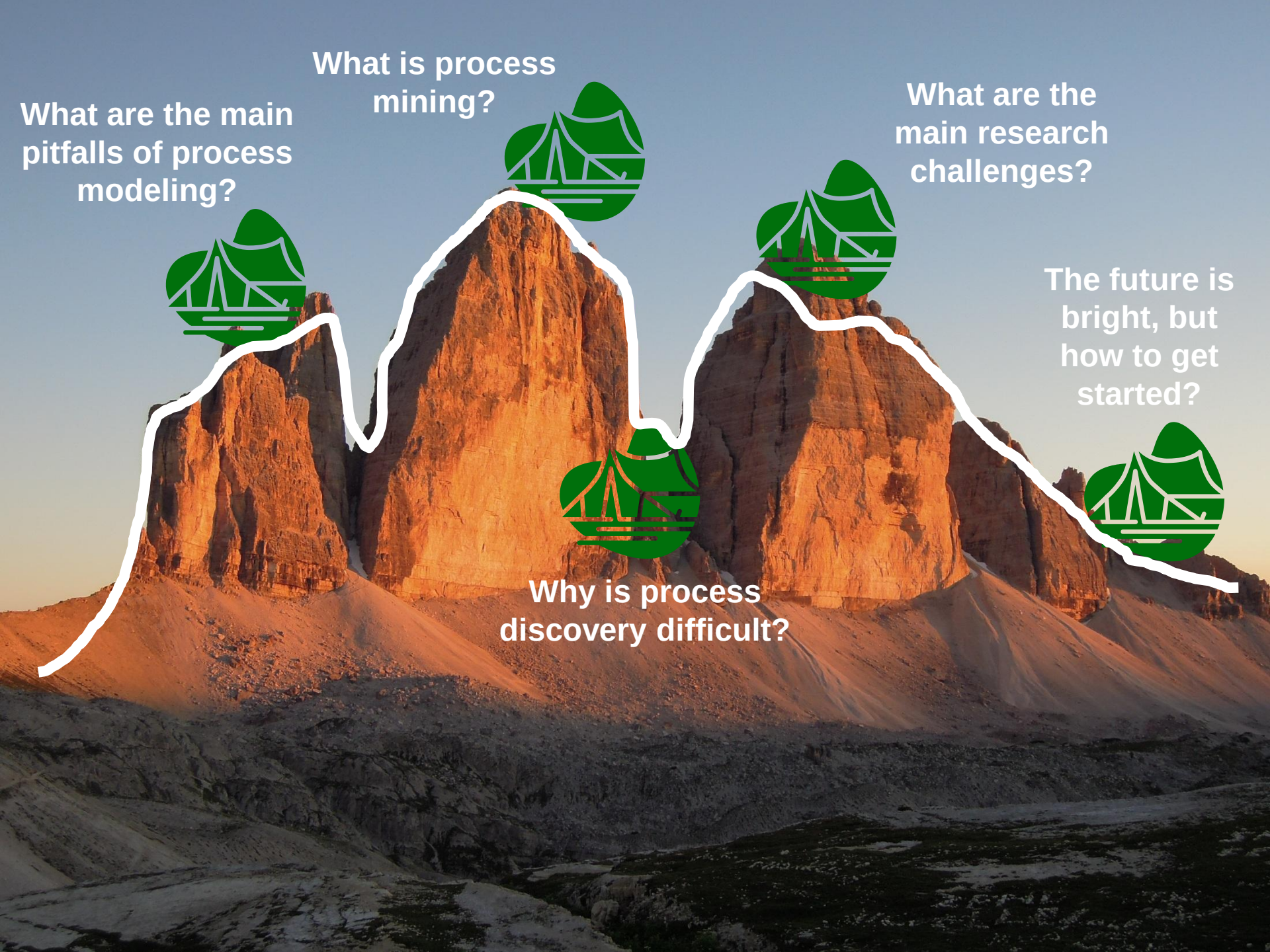

Changes
NEXT EXIT

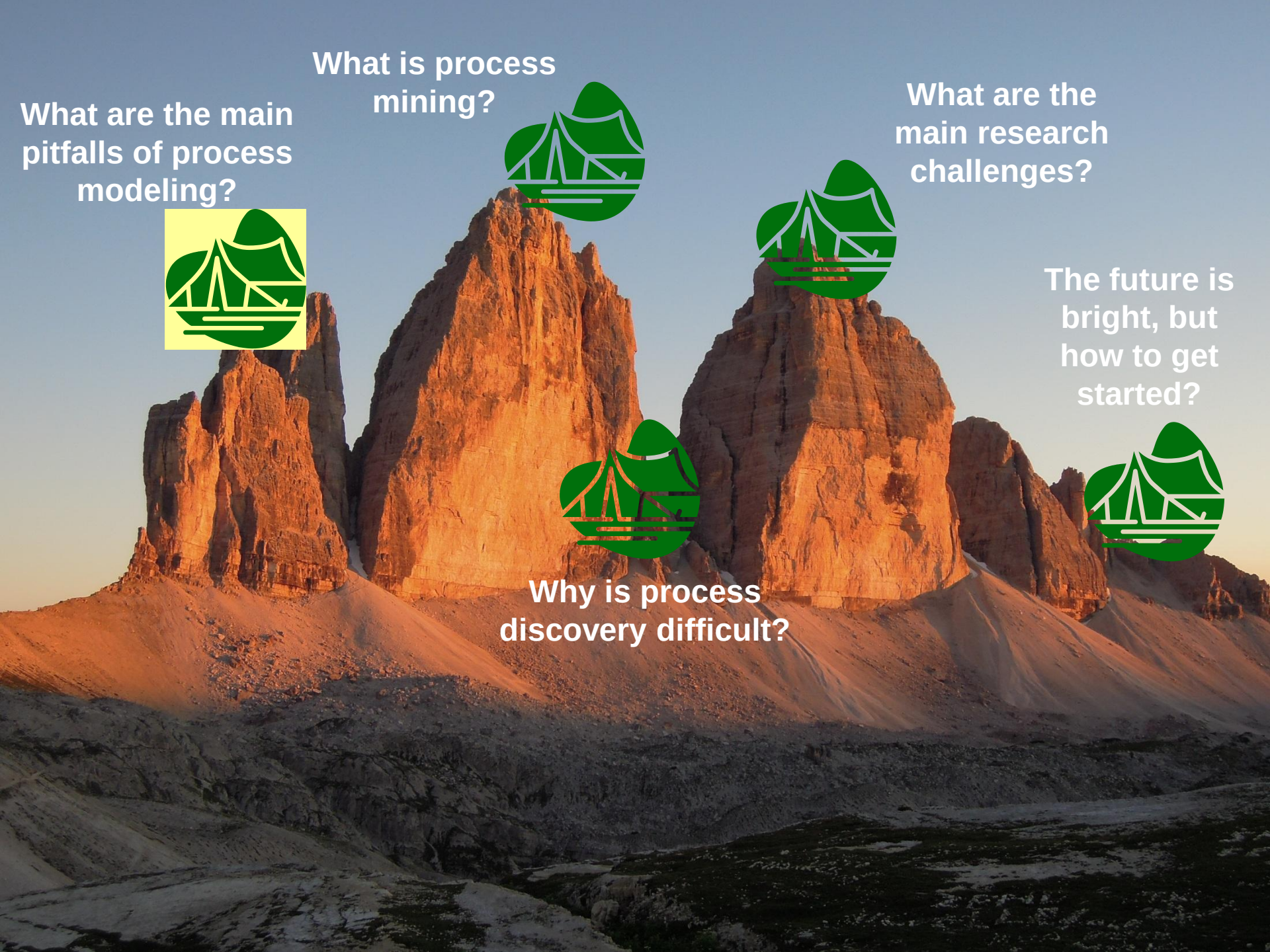

The
INTERNET
of THINGS

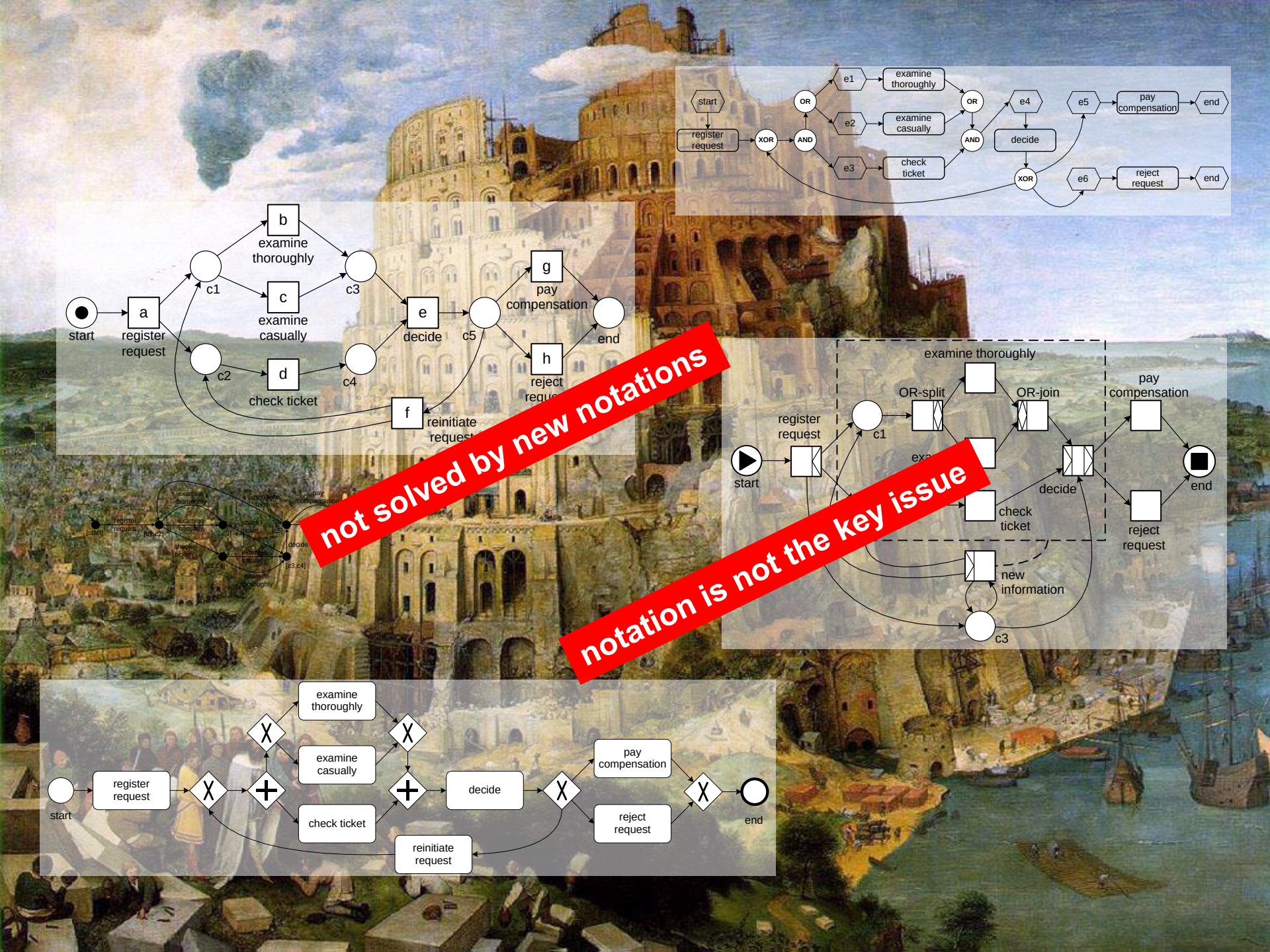

Big Data
Straight Ahead

No more (process) models?

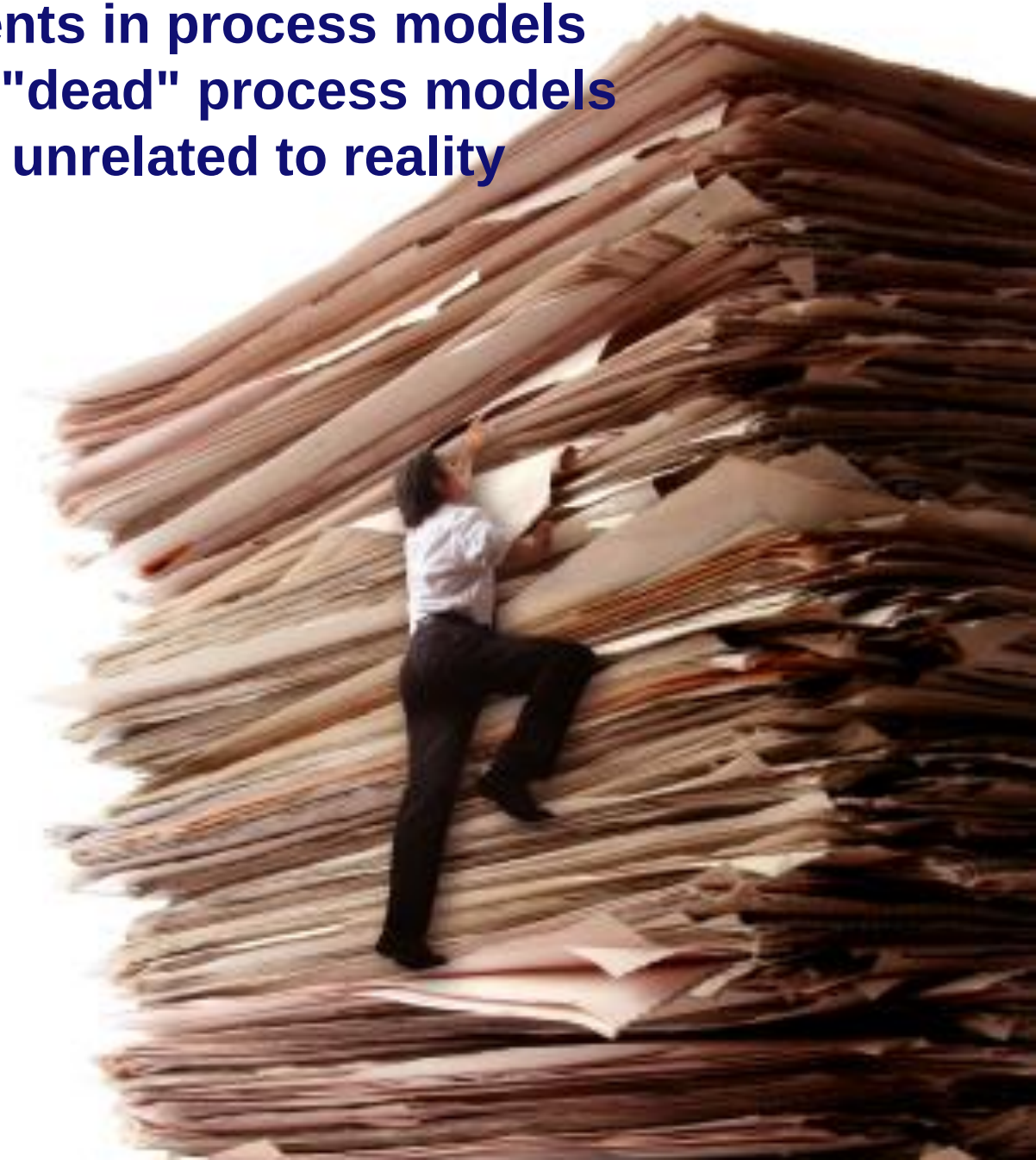

Don't think so!
but
models
need to add
real value …


What are the main pitfalls of process modeling?
What is process mining?
What are the main research challenges?
The future is bright, but how to get started?
Why is process discovery difficult?

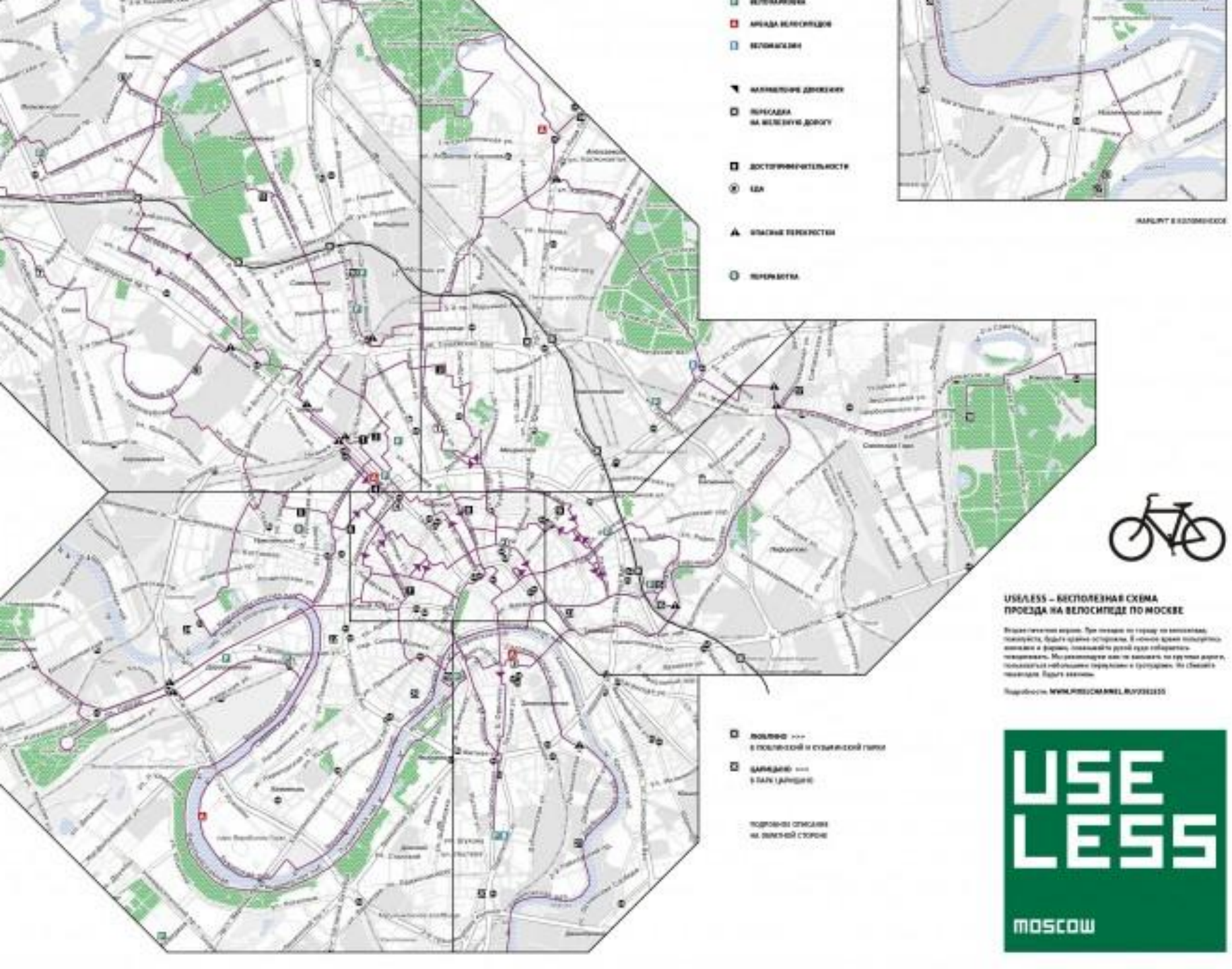

What are the main pitfalls of process modeling?
What is process mining?
What are the main research challenges?
The future is bright, but how to get started?
Why is process discovery difficult?


start
register request
XOR
AND
OR
e1
examine thoroughly
e2
examine casually
e3
check ticket
OR
AND
e4
decide
XOR
e5
pay compensation
e6
reject request
end
end
a
register request
b
examine thoroughly
c
examine casually
d
check ticket
c1
c2
c3
c4
c5
e
decide
f
reinitiate request
g
pay compensation
h
reject request
end
not solved by new notations
notation is not the key issue
examine thoroughly
register request
OR-split
OR-join
pay compensation
start
c1
decide
end
check ticket
reject request
new information
c3
start
register request
examine thoroughly
examine casually
check ticket
decide
pay compensation
reject request
reinitiate request
end

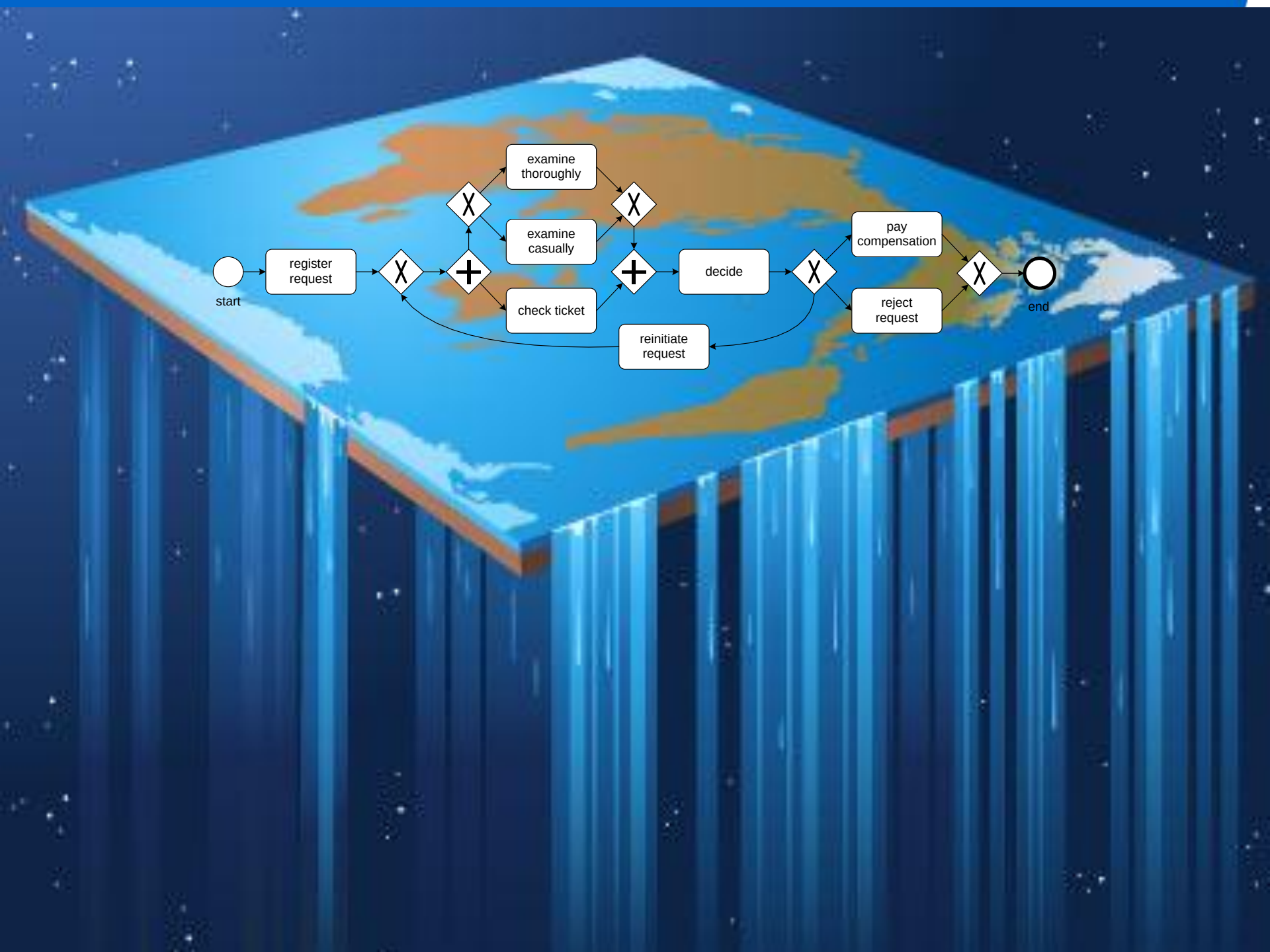
- **enormous investments in process models**
- **large collections of "dead" process models**
- **not taken seriously, unrelated to reality**



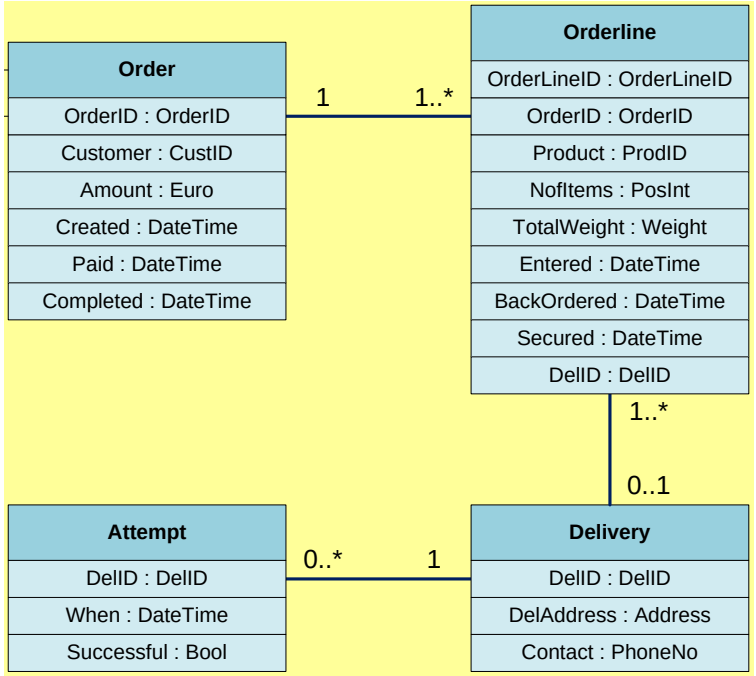

problem #1
Aiming for one model that suits all purposes


USE
LESS
moscow


problem #2
Straightjacketing smaller interacting processes into one monolithic model


examine thoroughly
examine casually
check ticket
register request
start
decide
pay compensation
reject request
reinitiate request
end

What is the process instance?



Order

OrderID	Customer	Amount	Created	Paid	Completed
91245	John	100	28-11-2011:08.12	02-12-2011:13.45	05-12-2011:11.33
91561	Mike	530	28-11-2011:12.22	03-12-2011:14.34	05-12-2011:09.32
91812	Mary	234	29-11-2011:09.45	02-12-2011:09.44	04-12-2011:13.33
92233	Sue	110	29-11-2011:10.12	null	null
92345	Kirsten	195	29-11-2011:14.45	02-12-2011:13.45	null
92355	Pete	320	29-11-2011:16.32	null	null
…	…	…	…	…	…

Delivery

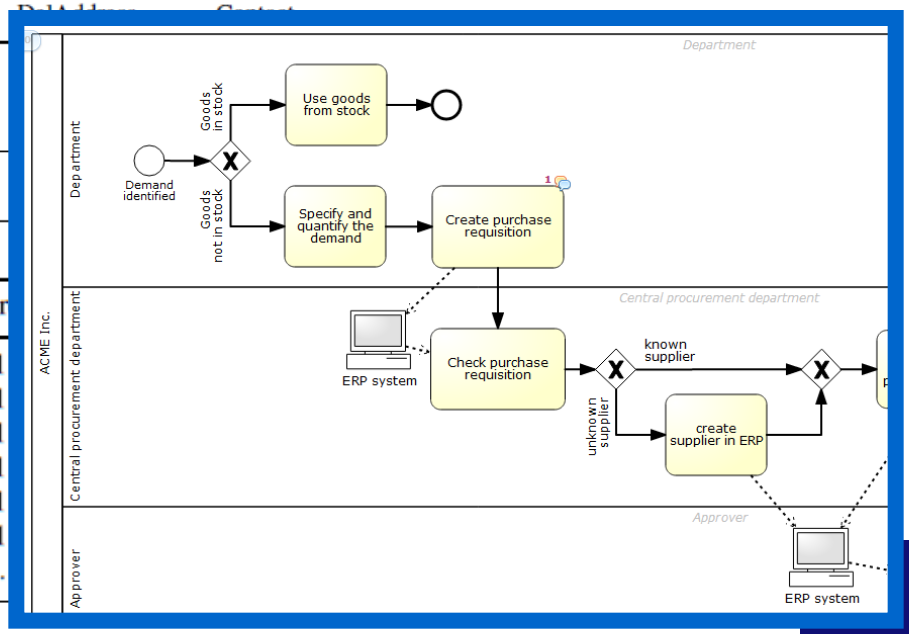
DelID	…
882345	
882346	
…	

Attempt

DelID	When	Successful

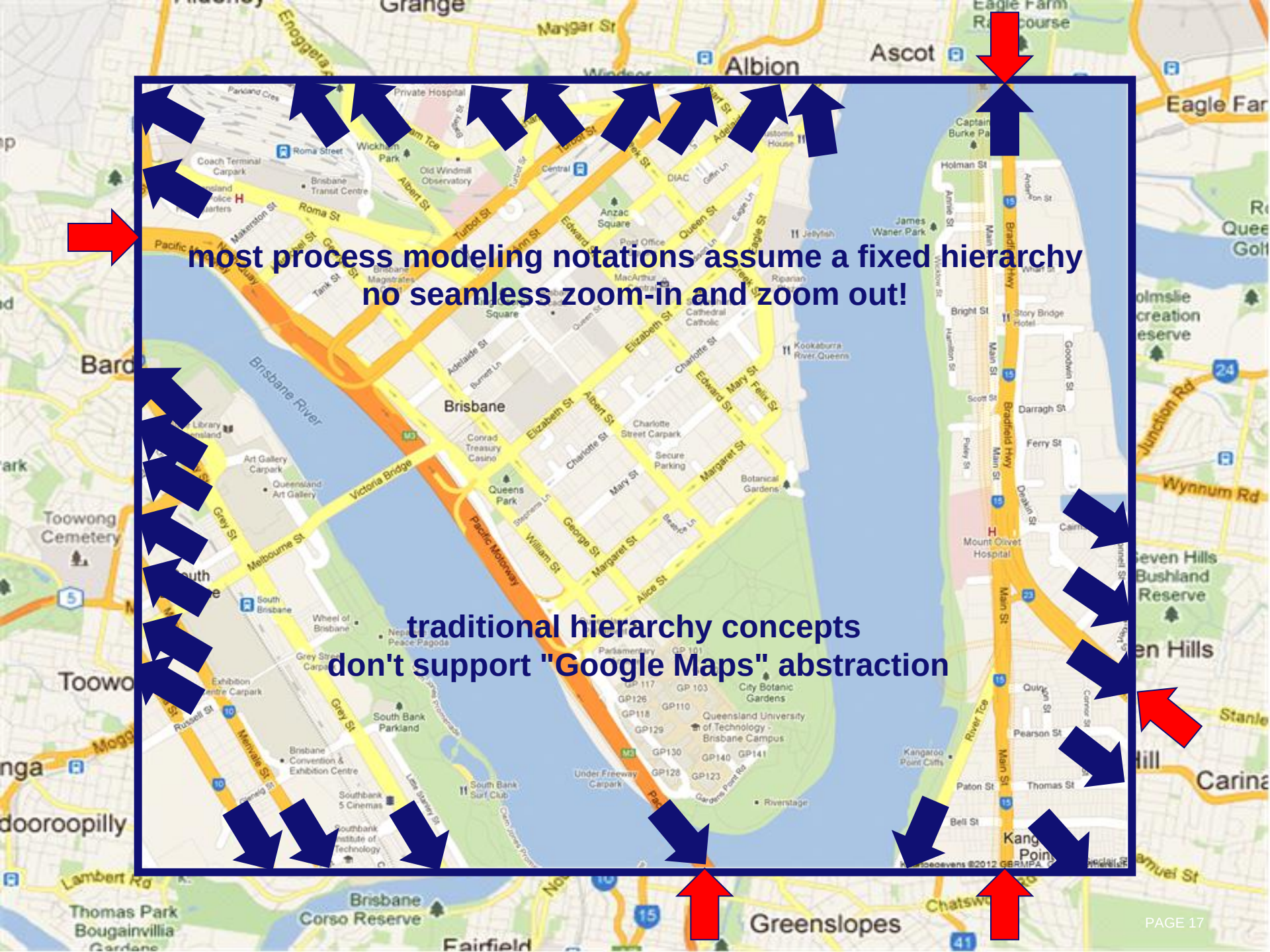
Orderline

OrderLineID	OrderID	Product	NofItems	TotalWeight	Enter
112345	91245	iPhone 4G	1	0.250	28-11-201
112346	91245	iPod nano	2	0.300	28-11-201
112347	91245	iPod classic	1	0.200	28-11-201
112448	91561	iPhone 4G	1	0.250	28-11-201
112449	91561	iPod classic	1	0.200	28-11-201
112452	91812	iPhone 4G	5	1.250	29-11-201
…	…	…	…	…	…



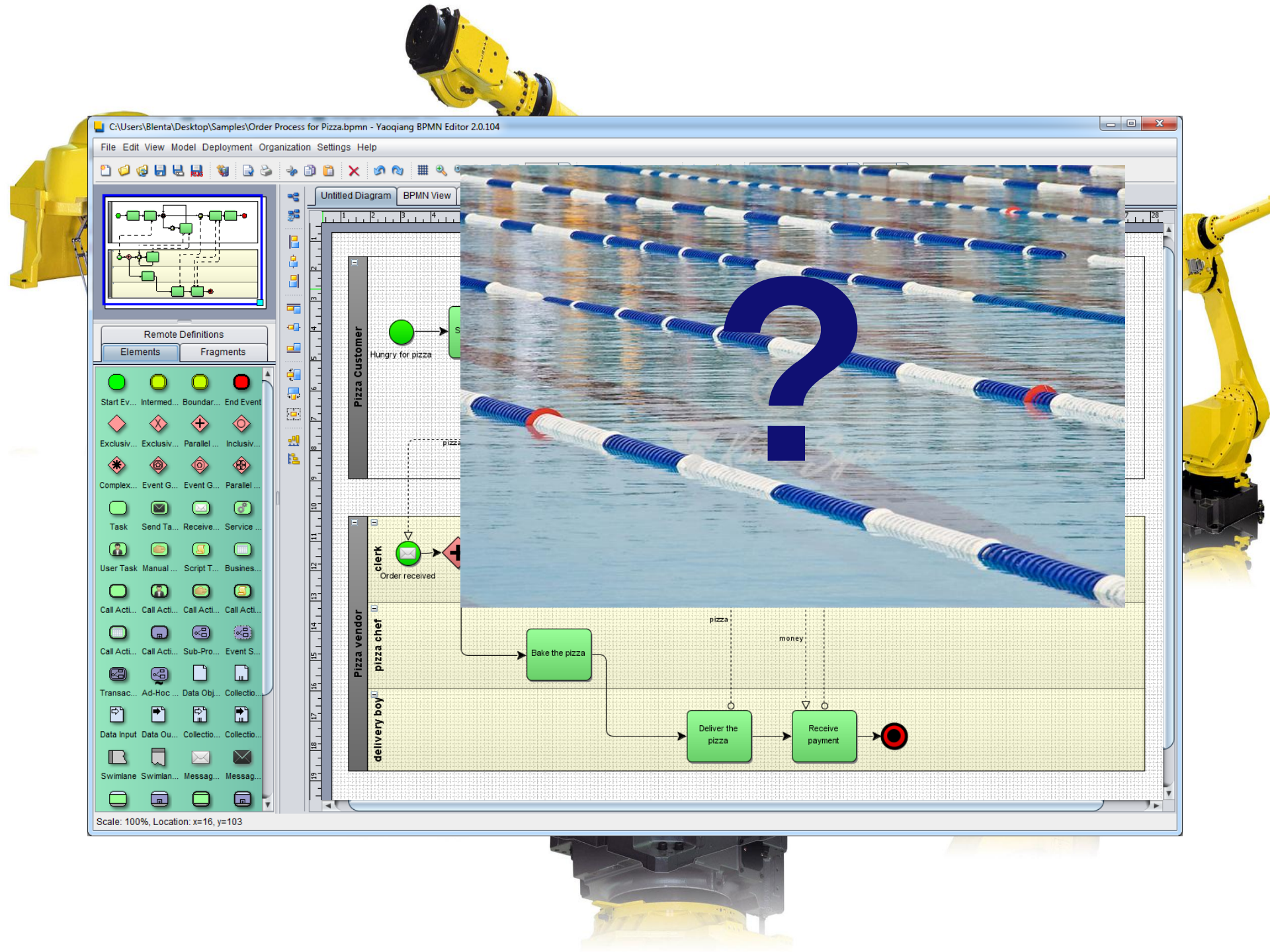
problem #3

Using a static hierarchical decomposition as the only abstraction mechanism


most process modeling notations assume a fixed hierarchy
no seamless zoom-in and zoom out!
traditional hierarchy concepts
don't support "Google Maps" abstraction


problem #4
Modeling humans as if they are machines doing a single task

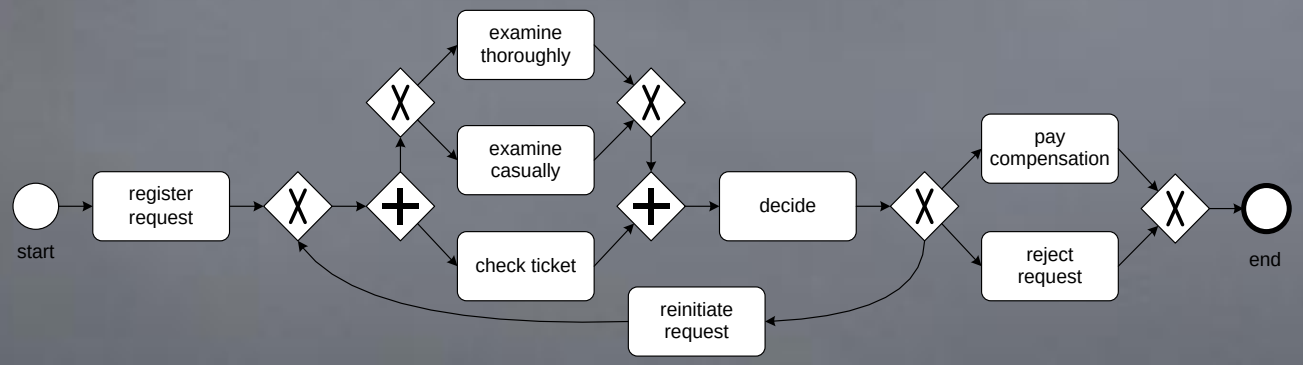

"My processes are unique,
my people are artists!"


C:\Users\Blenta\Desktop\Samples\Order Process for Pizza.bpmn - Yaoqiang BPMN Editor 2.0.104
File Edit View Model Deployment Organization Settings Help
Untitled Diagram
BPMN View
Remote Definitions
Elements
Fragments
Pizza Customer
Hungry for pizza
clerk
Order received
Pizza vendor
pizza chef
Bake the pizza
pizza
money
delivery boy
Deliver the pizza
Receive payment
?
Scale: 100%, Location: x=16, y=103

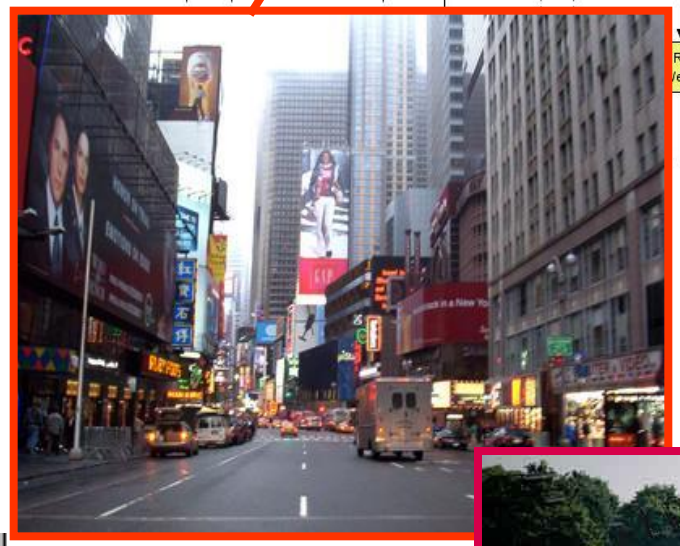

problem #5
Being vague about vagueness

Make vagueness explicit!

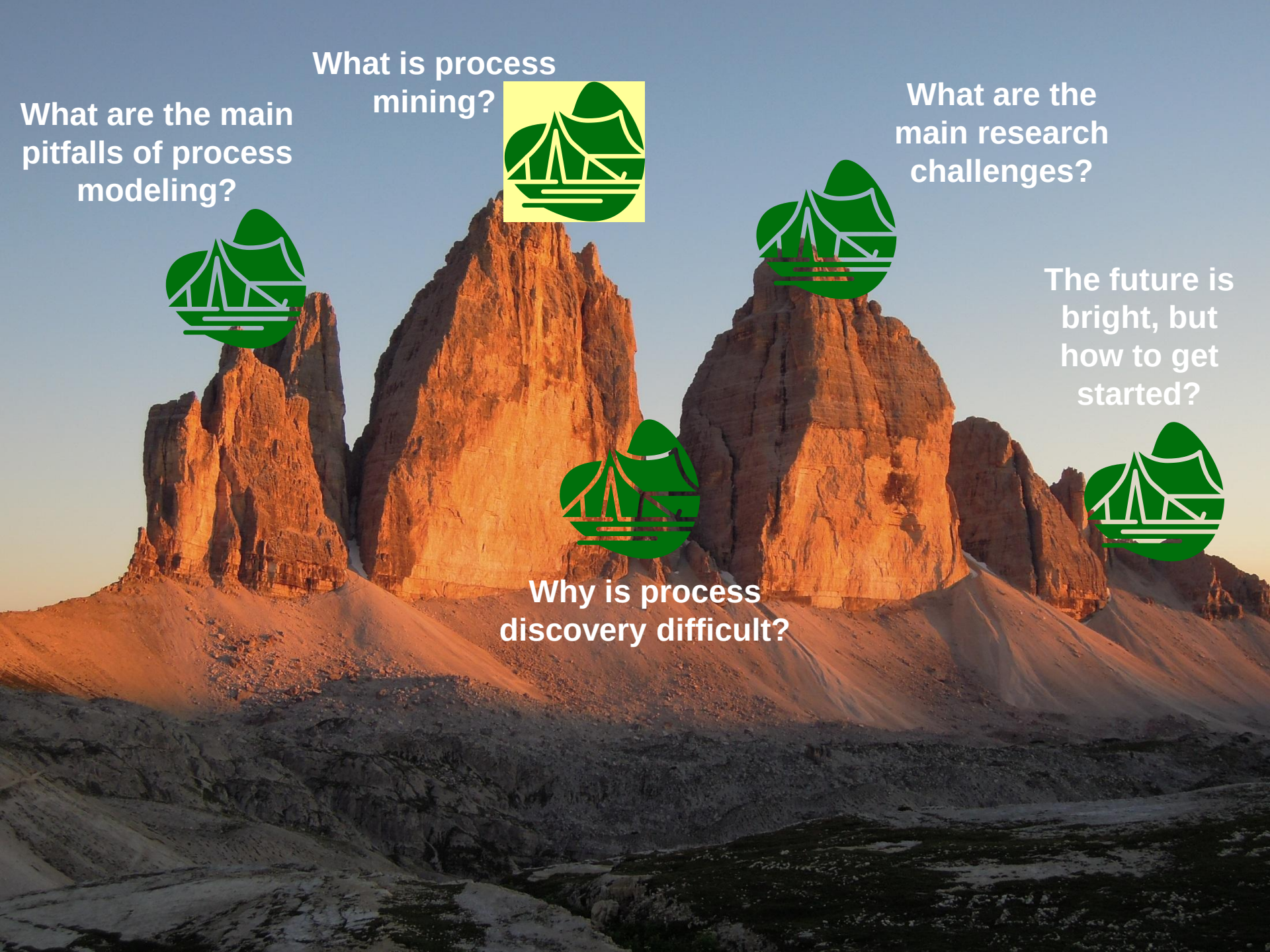
(rather than having multiple notations for the same thing)




problem #6
Abstracting from the things that matter



Start on Friday
Receive Issue List
Review Issue List
Any issues ready?
Yes
No
Discussion Cycle
Announce issues
Collect Votes
Timed Out [1 week]
Prepare Results
Results Web Site
E-Mail Results of Vote
Issues w/o Majority?
Yes
No
Change Vote Message
Did Enough Members Vote?
No
Vote Results
Have the members been warned?
Yes
Reduce number of Voting Members and Recalculate Vote
No
Re-announce Vote with
customer
No
No
Washington product?
Yes
IT Service Center WN addresses
No
Resolves problem?
Yes
ISC Supp
Vendor


What is process mining?
What are the main pitfalls of process modeling?
What are the main research challenges?
The future is bright, but how to get started?
Why is process discovery difficult?

Positioning Process Mining

process model analysis

(simulation, verification, optimization, gaming, etc.)

performance-oriented questions, problems and solutions

compliance-oriented questions, problems and solutions

data-oriented analysis

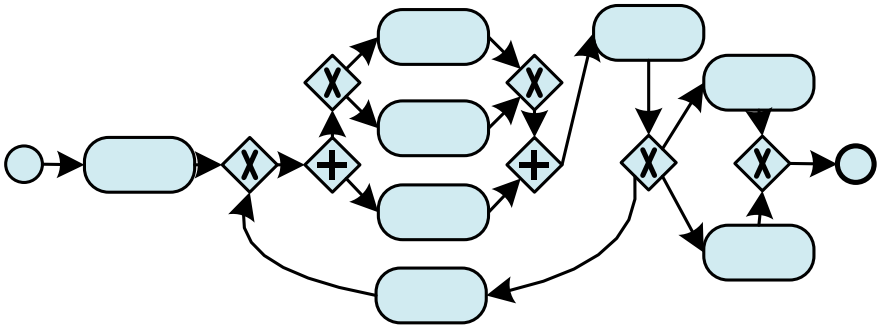
(data mining, machine learning, business intelligence)


www.olifantenpaadjes.nl
01001100110101010
0100110101010

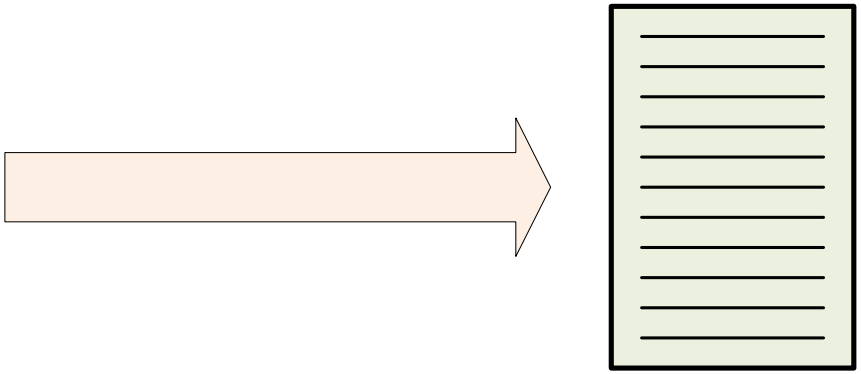

0100110101010


Let us take a step back and see how
models and behavior relate: Let's play!

Play-Out

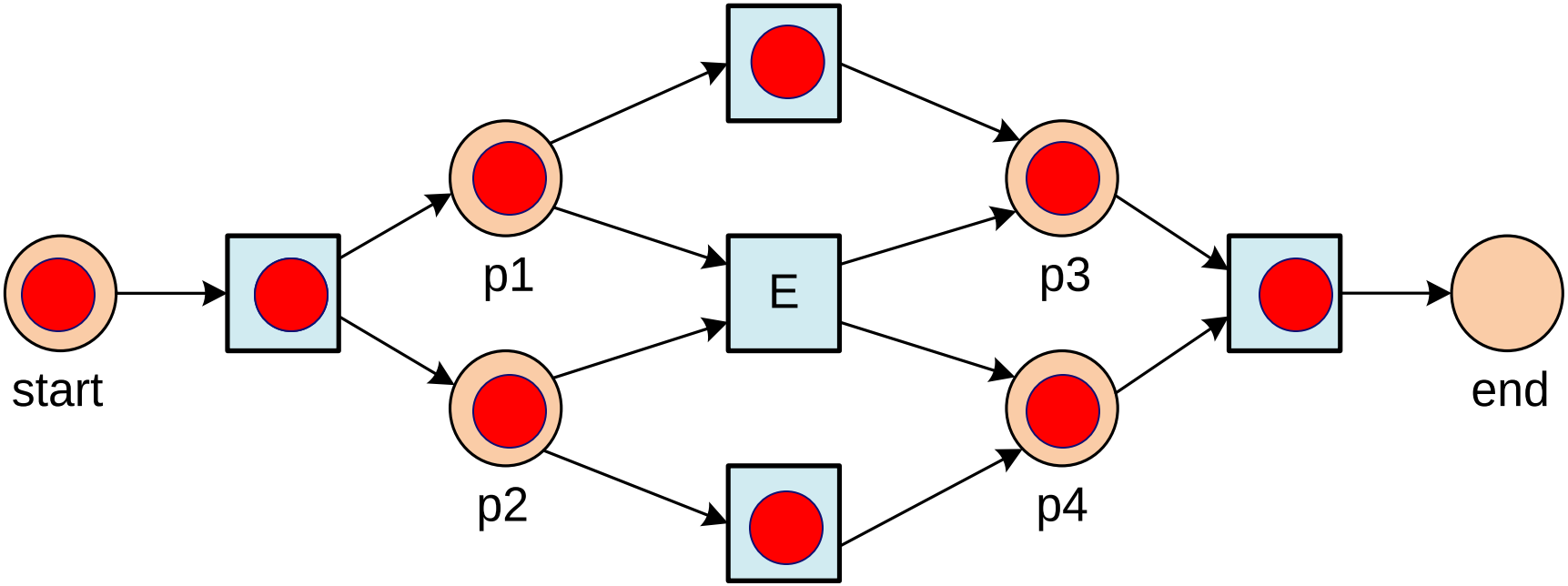


process model



event log

Play-Out (Classical use of models)

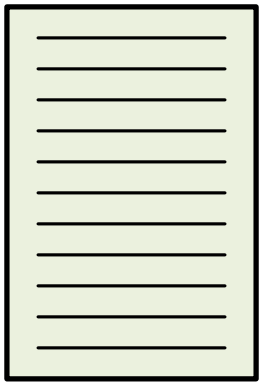


A B C D A E D A E D

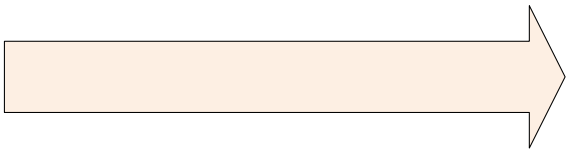
A B C D A C B D

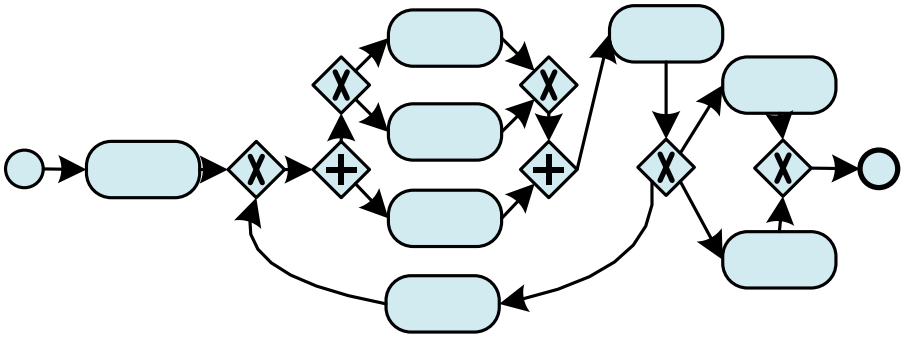
A C B D A E D A C B D

Play-In



event log





process model

Play-In

A B C D

A E D

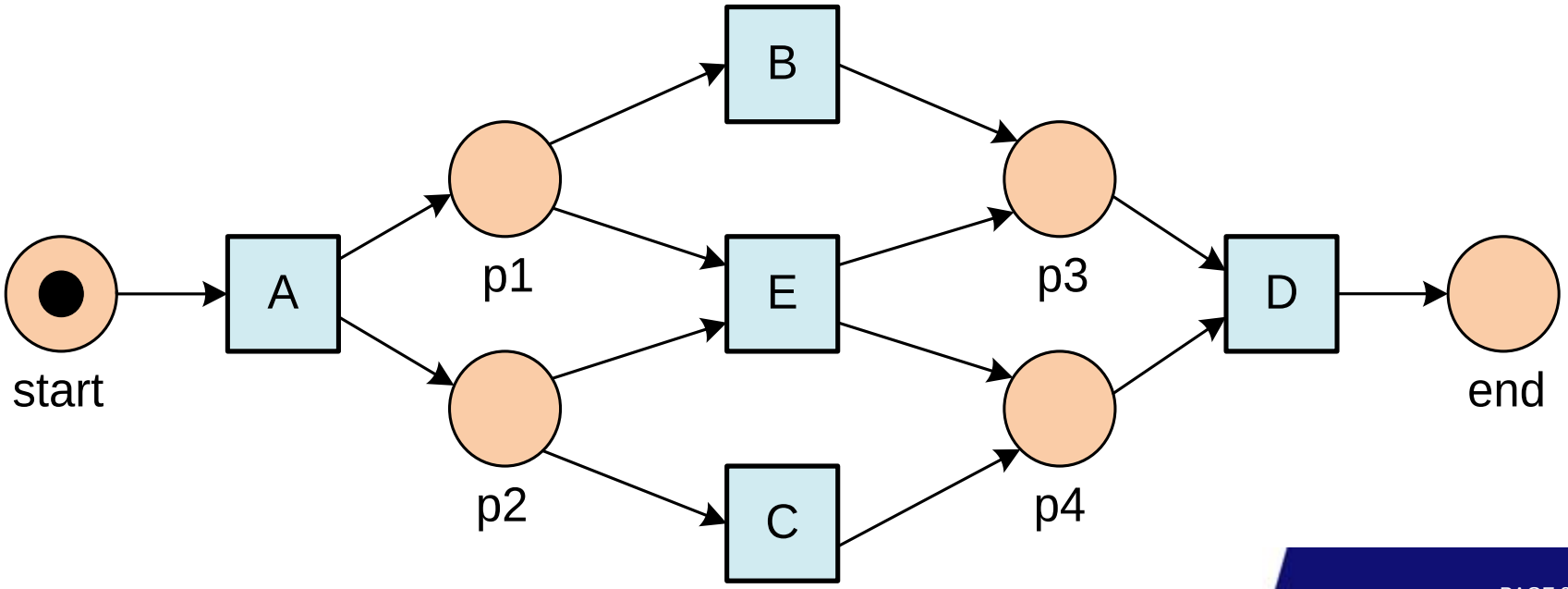
A E D

A B C D

A C B D

A C B D

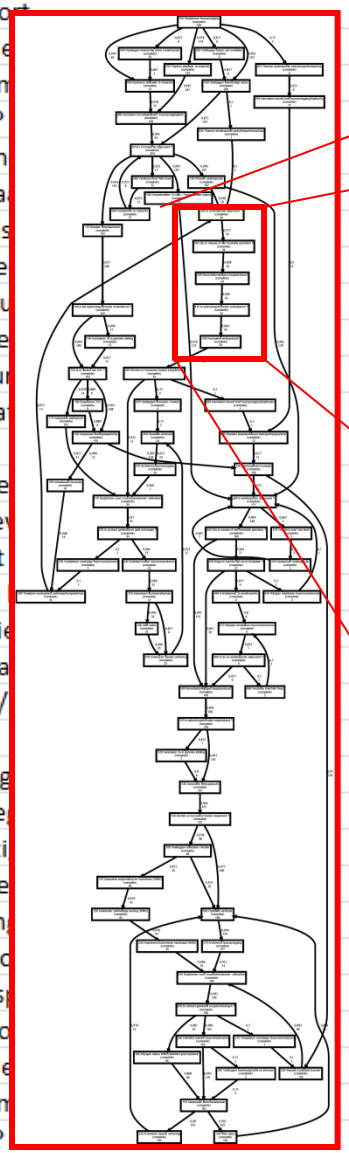
A E D

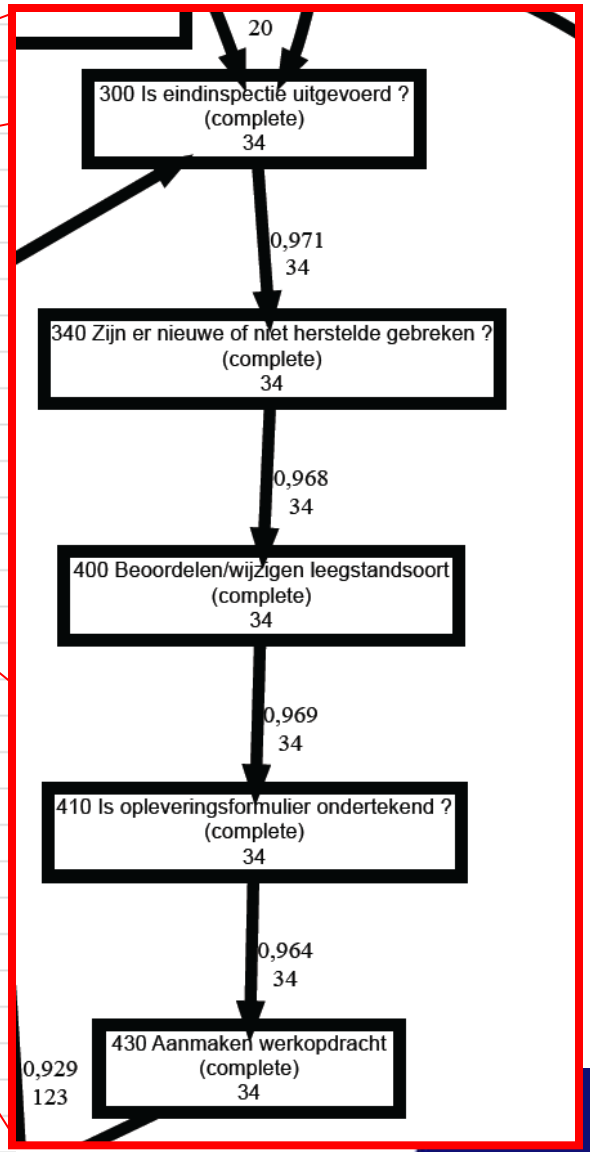


Example Process Discovery

(Vestia, Dutch housing agency, 208 cases, 5987 events)

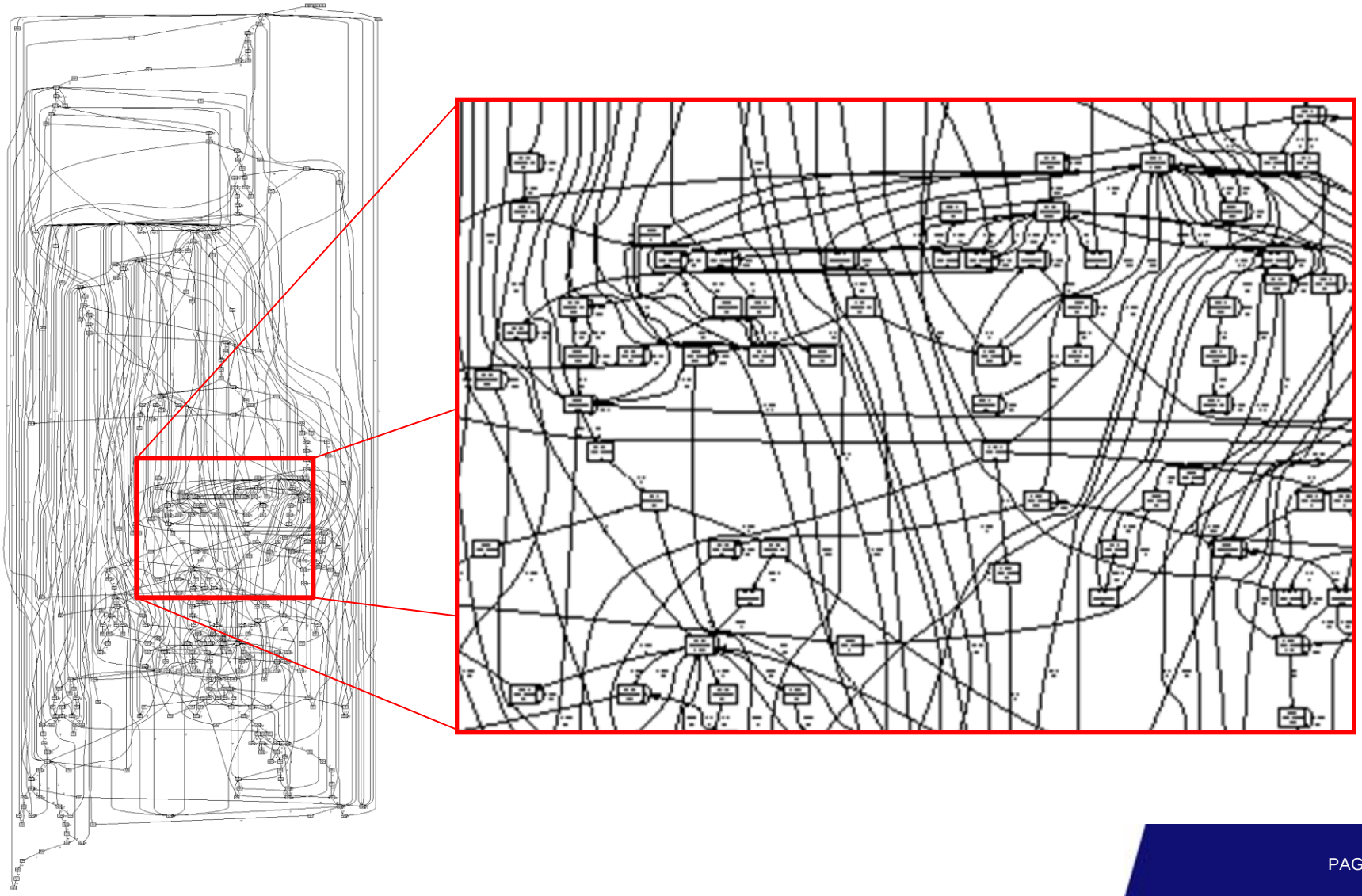
117315	110 Bepalen leegstandsoo	16.05.2007 14:06:23
117315	120 Plannen eindinspectie	16.05.2007 14:36:01
117315	130 Is het opleveringsform	23.05.2007 09:41:40
117315	150 Is er sprake van ZAV ?	23.05.2007 09:41:51
117315	170 Aanpassen plattegron	23.05.2007 11:57:18
117315	180 Aanpassen woningwa	23.05.2007 09:42:37
117315	190 Actualiseren huurprijs	23.05.2007 09:48:23
117315	200 Toewijzen woning/be	23.05.2007 09:48:29
117315	210 Registreren voorl. huu	10.09.2007 16:24:36
117315	220 Is contract getekend e	11.09.2007 14:56:18
117315	240 Definitief maken Huur	31.03.2008 16:17:12
117315	250 Aanpassen factureera	09.09.2008 15:39:59
117315	260 After sales	09.09.2008 16:51:24
117315	270 Archiveren nieuwe ve	10.09.2008 07:52:08
117315	300 Is eindinspectie uitgev	07.06.2007 14:47:04
117315	340 Zijn er nieuwe of niet	07.06.2007 14:47:06
117315	400 Beoordelen/wijzigen l	07.06.2007 14:51:16
117315	410 Is opleveringsformulie	07.06.2007 14:51:26
117315	430 Aanmaken werkopdra	11.06.2007 09:21:39
117315	440 Worden er bonussen/	11.06.2007 09:21:49
117315	460 Opstellen eindnota	08.08.2007 16:18:26
117315	470 Archiveren huuropzeg	09.08.2007 14:42:23
119763	010 Registreren huuropze	09.05.2007 11:19:14
119763	030 Vastleggen toekomsti	09.05.2007 12:25:01
119763	050 Inplannen afspraak 1e	09.05.2007 11:59:52
119763	060 Aanmaken bevestigin	09.05.2007 12:31:57
119763	070 Is 1e inspectie uitgevo	16.05.2007 13:04:26
119763	100 Gereedmelden 1e insp	16.05.2007 13:43:39
119763	110 Bepalen leegstandsoo	16.05.2007 13:43:28
119763	120 Plannen eindinspectie	16.05.2007 13:42:58
119763	130 Is het opleveringsform	16.05.2007 13:34:49
119763	150 Is er sprake van ZAV ?	16.05.2007 13:34:56





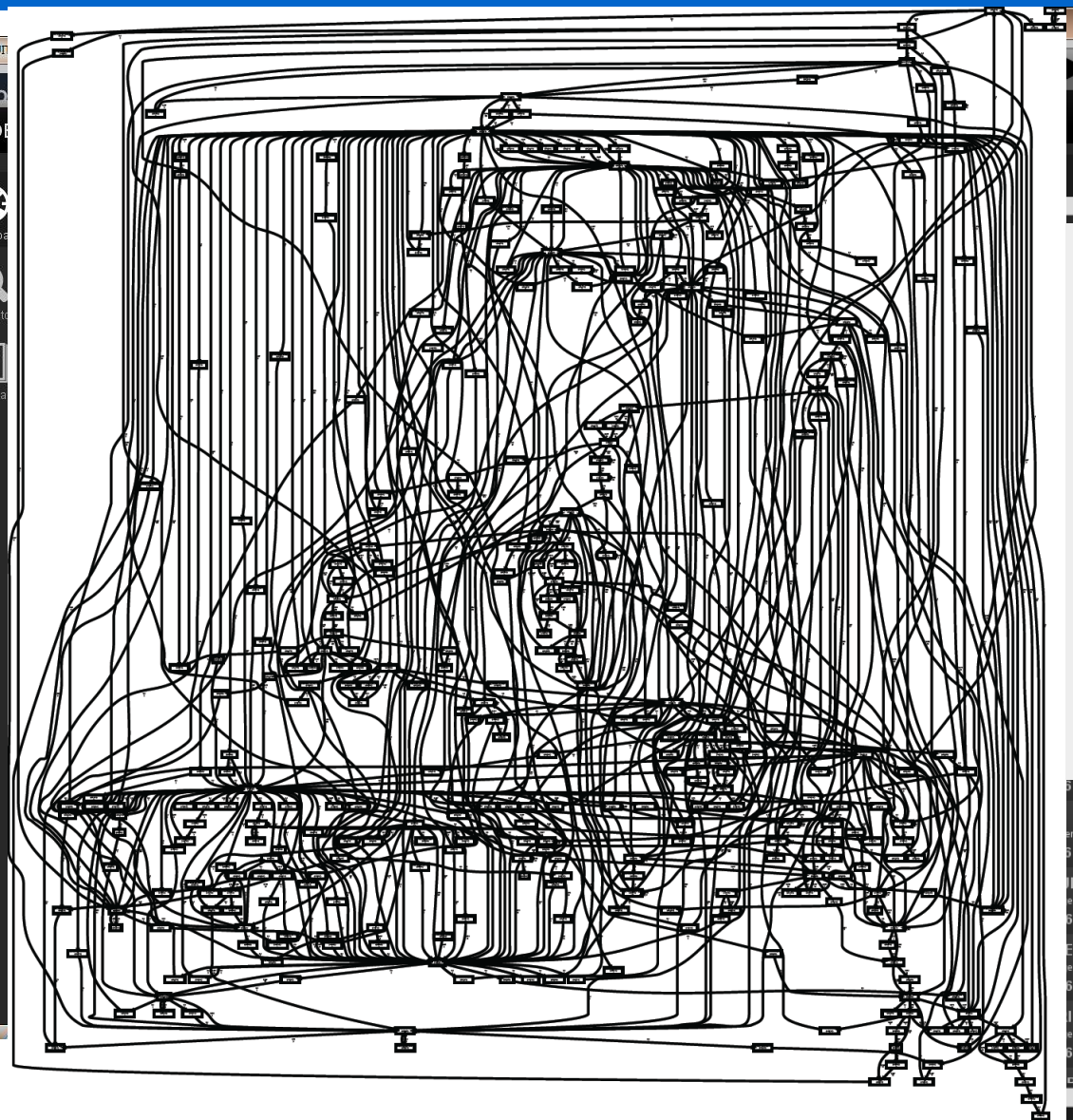
Example Process Discovery

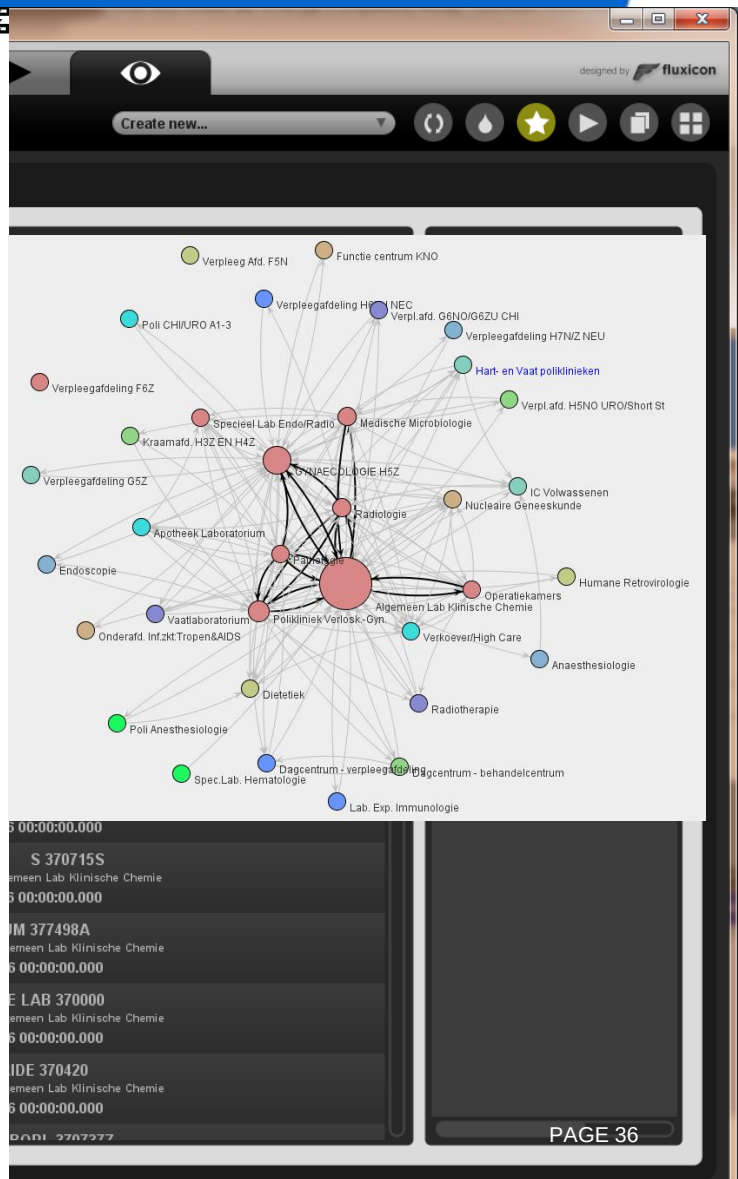
(ASML, test process lithography systems, 154966 events)



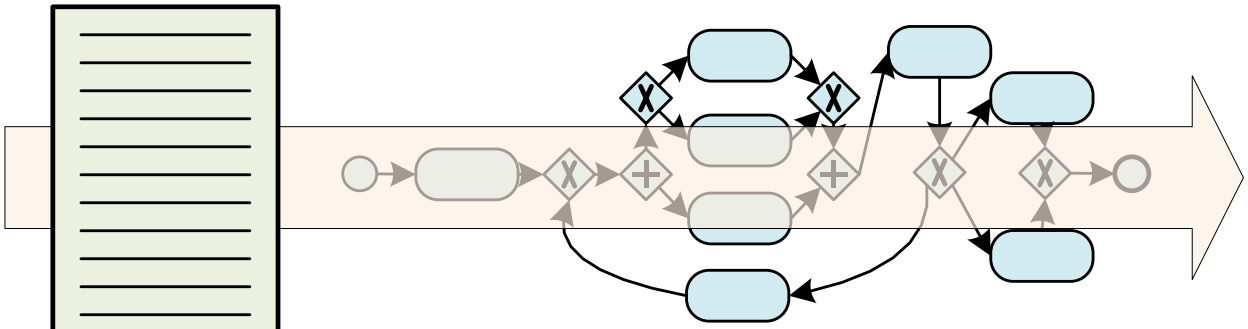
Example Process Discovery

(AMC, 627 gynecological oncology patients, 24331 events)





Replay



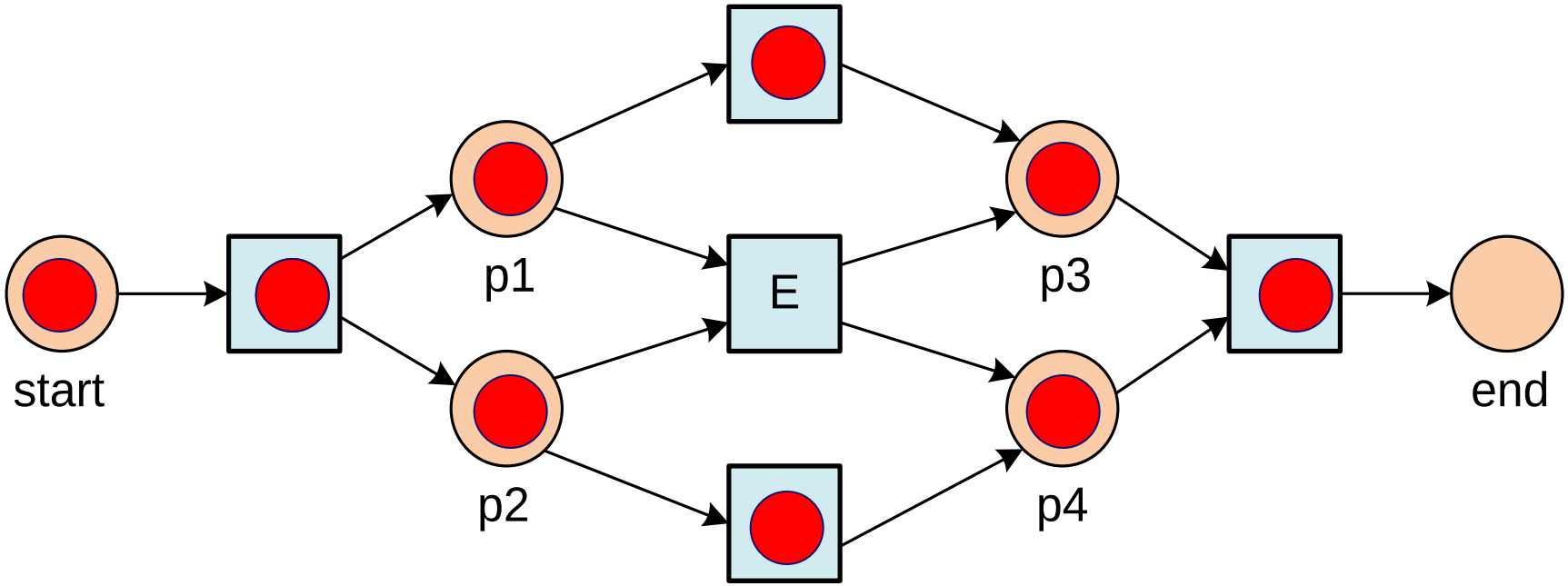
event log

process model

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

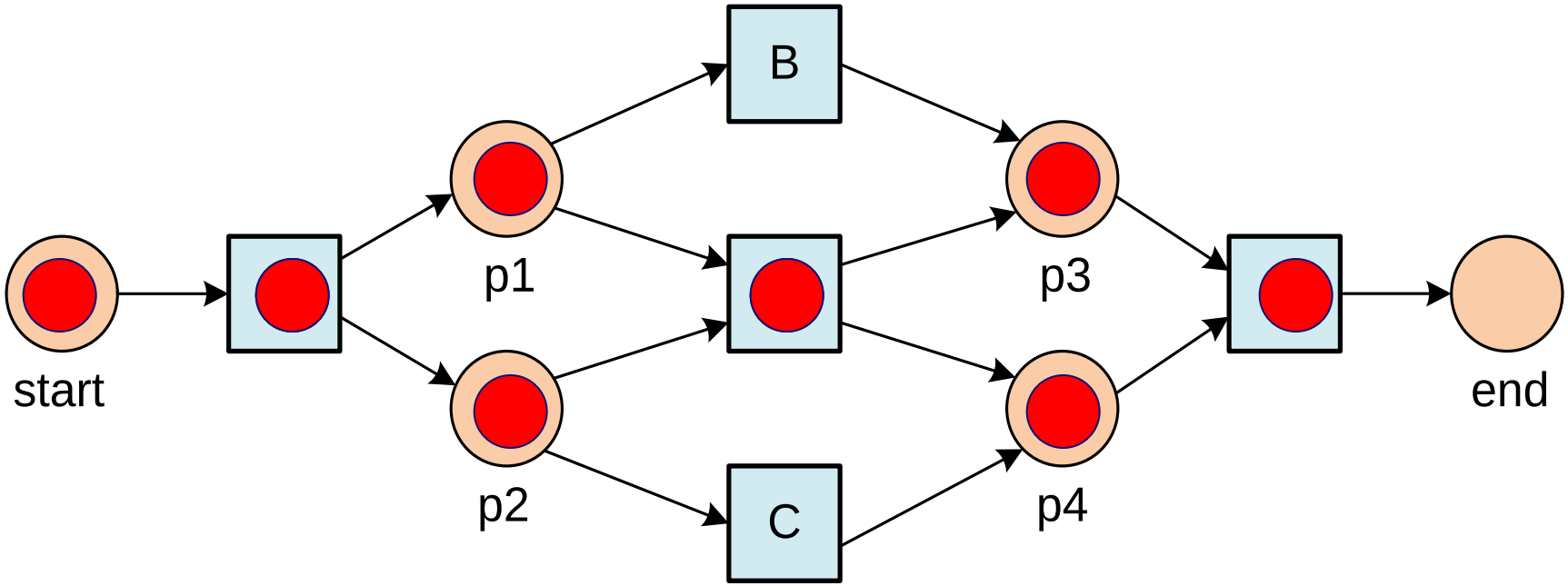
Replay

A B C D



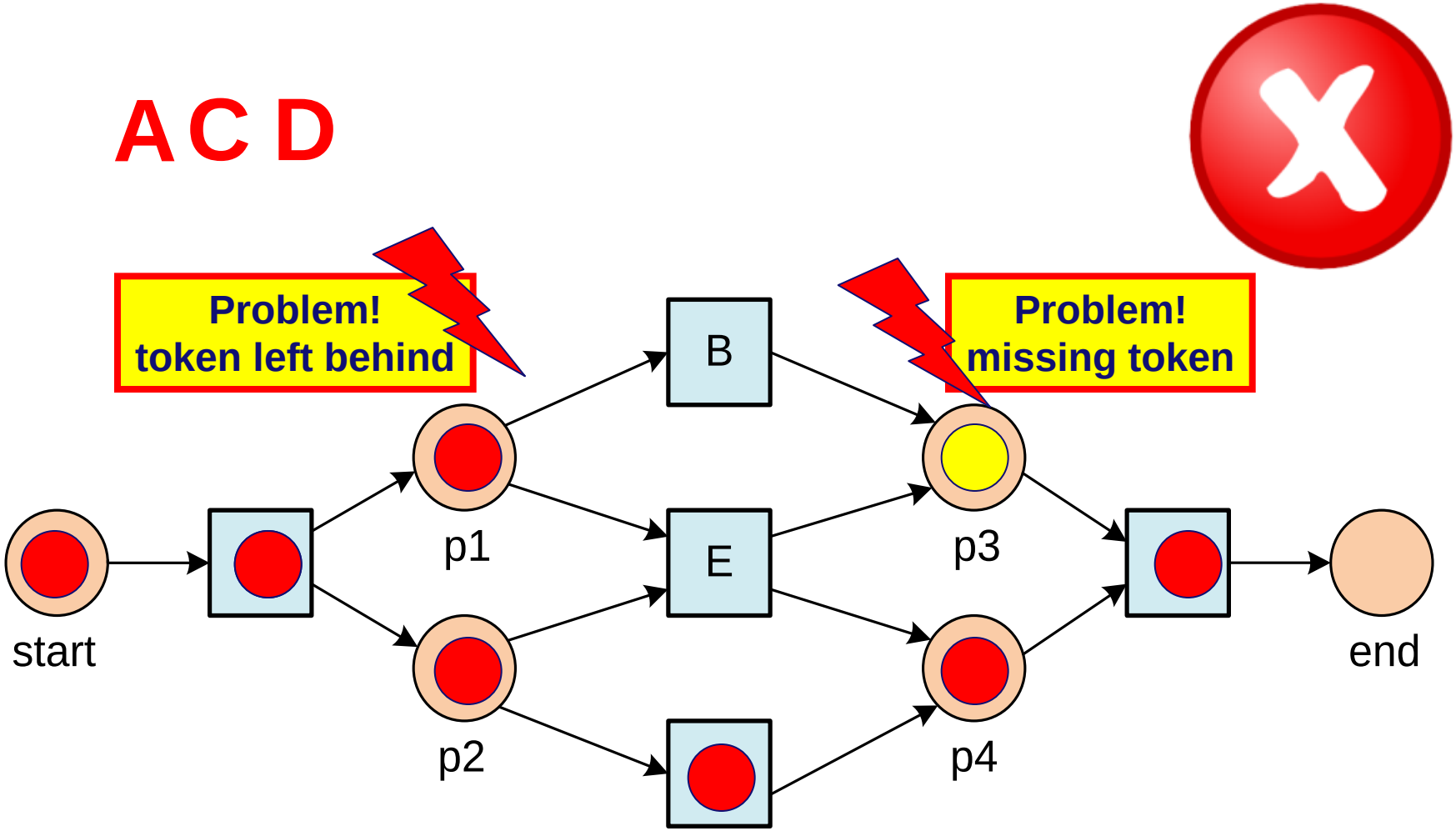
Replay

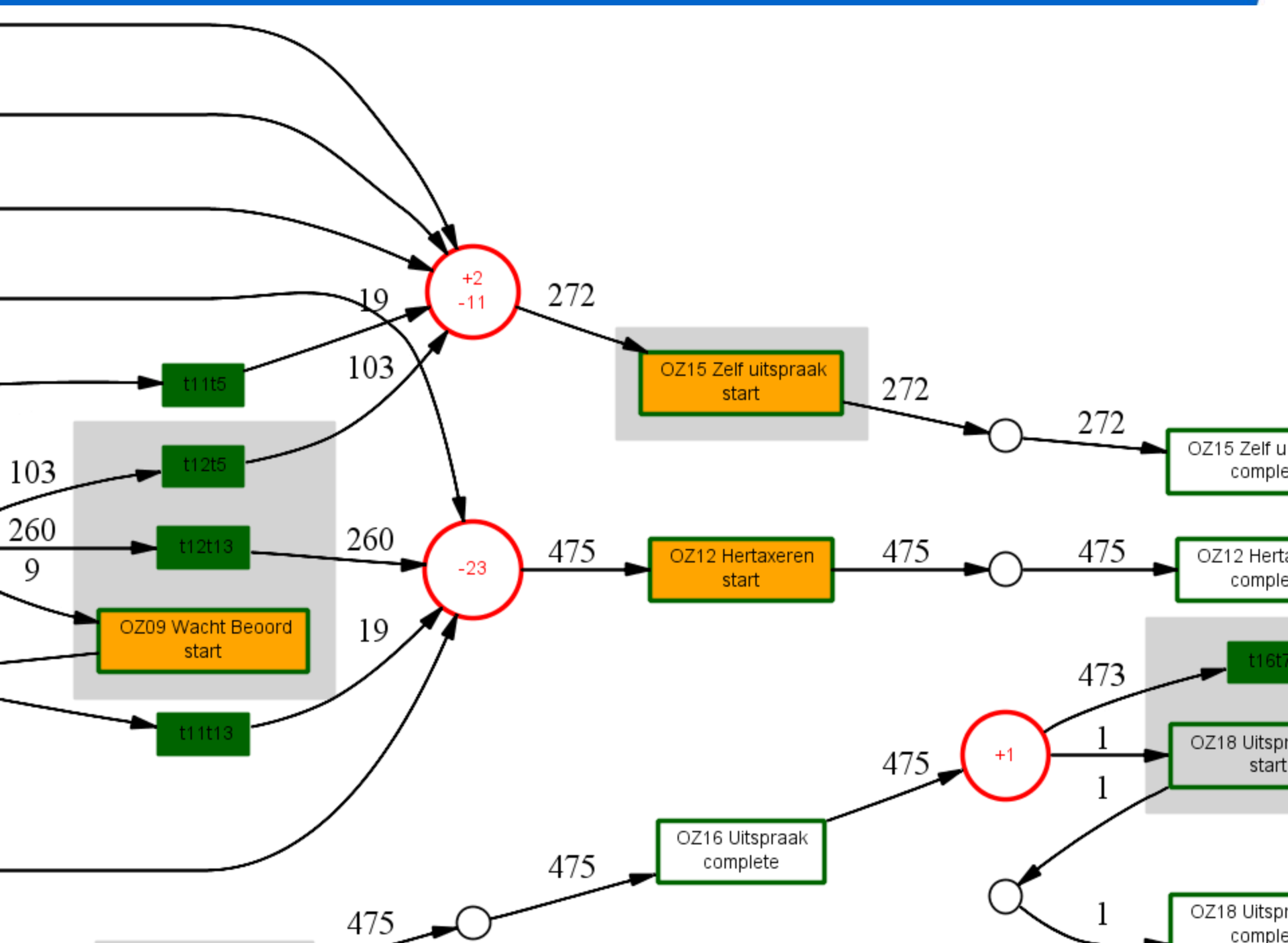
A E D



Replay can detect problems

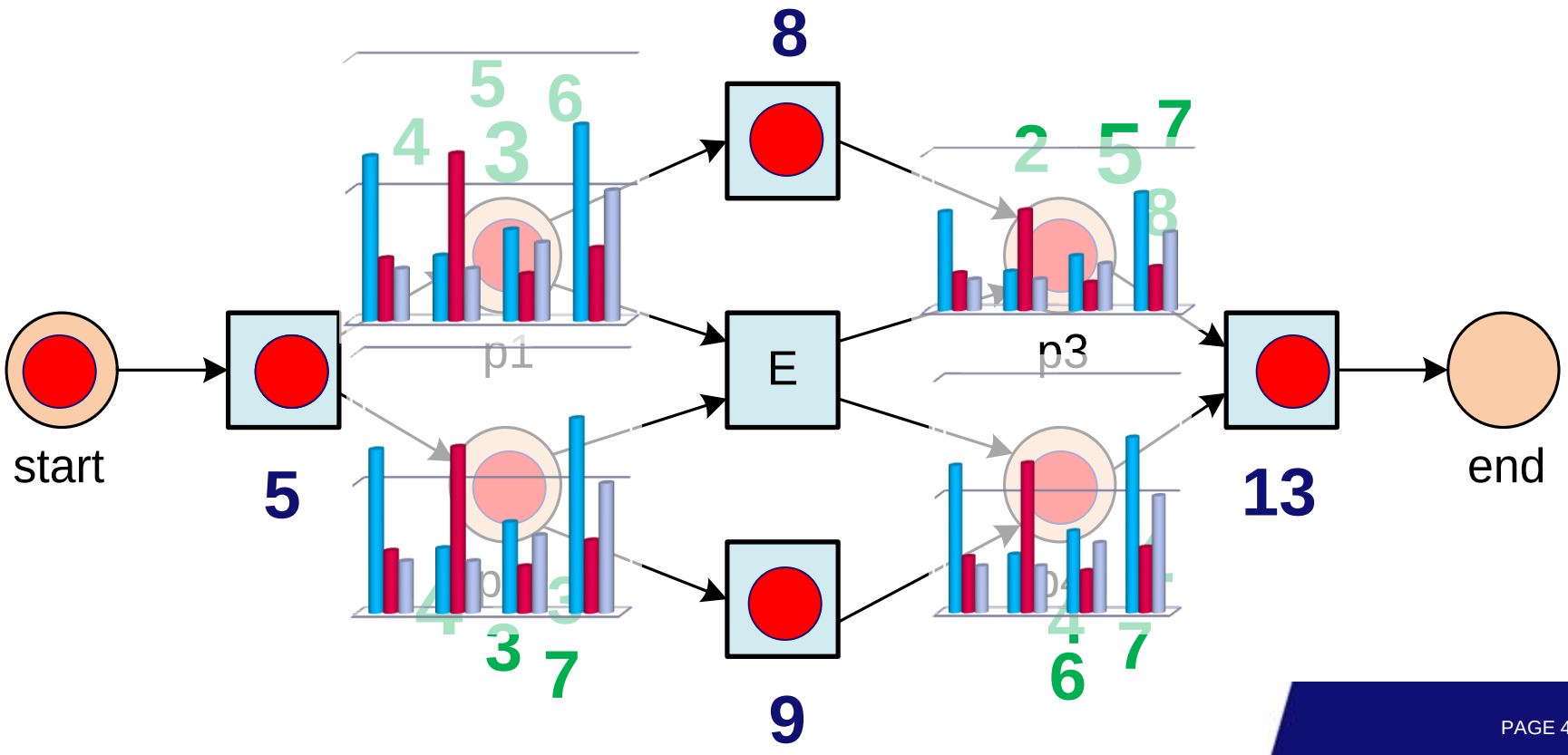
A C D




+2
-11
19
103
272
t11t5
t12t5
103
260
9
t12t13
OZ09 Wacht Beoord start
t11t13
260
19
-23
475
OZ12 Hertaxeren start
OZ15 Zelf uitspraak start
272
272
272
OZ15 Zelf u
comple
475
475
OZ12 Herta
comple
473
1
1
+1
475
OZ16 Uitspraak complete
475
475
OZ18 Uitspr
start
1
OZ18 Uitspr
comple

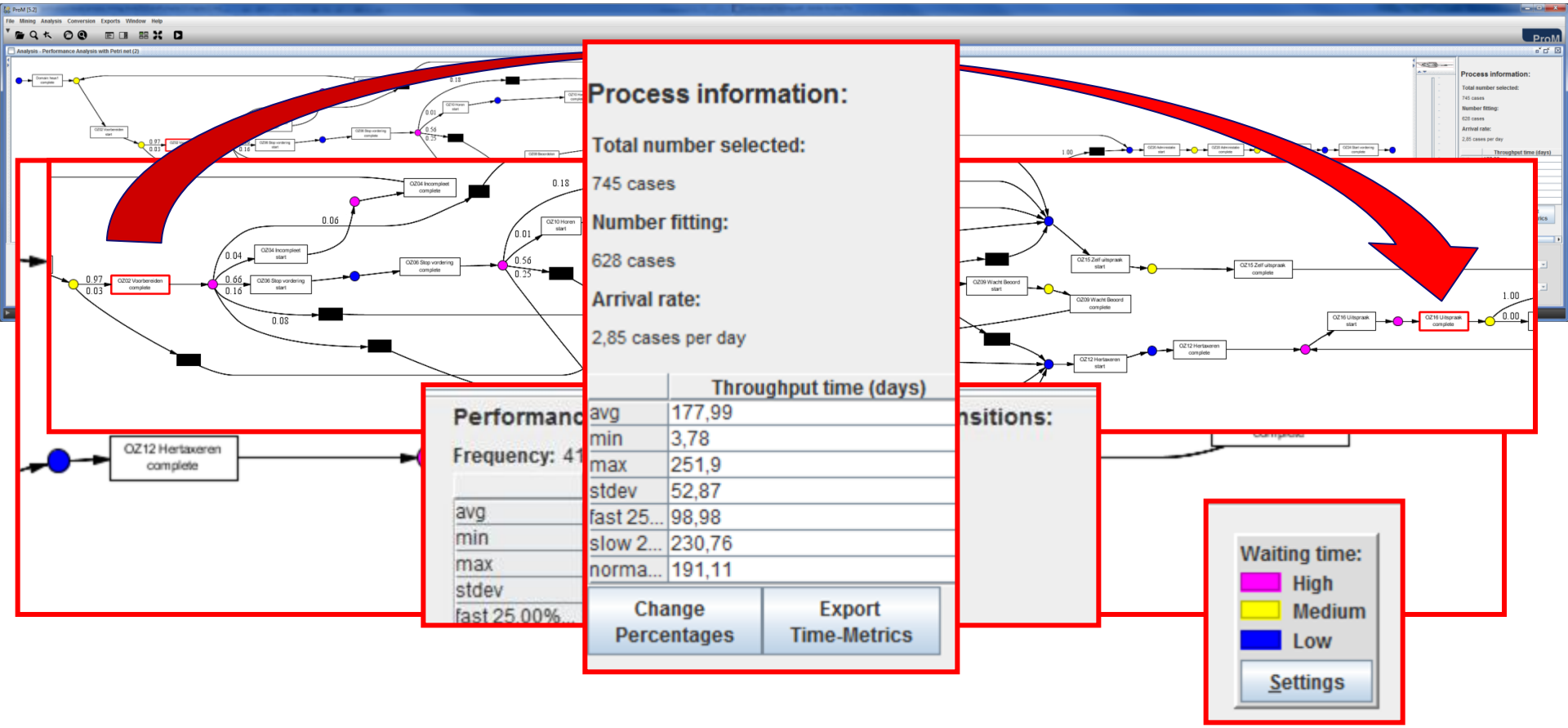
Replay can extract timing information

$A^{5}B^{8}C^{9}D^{13}$

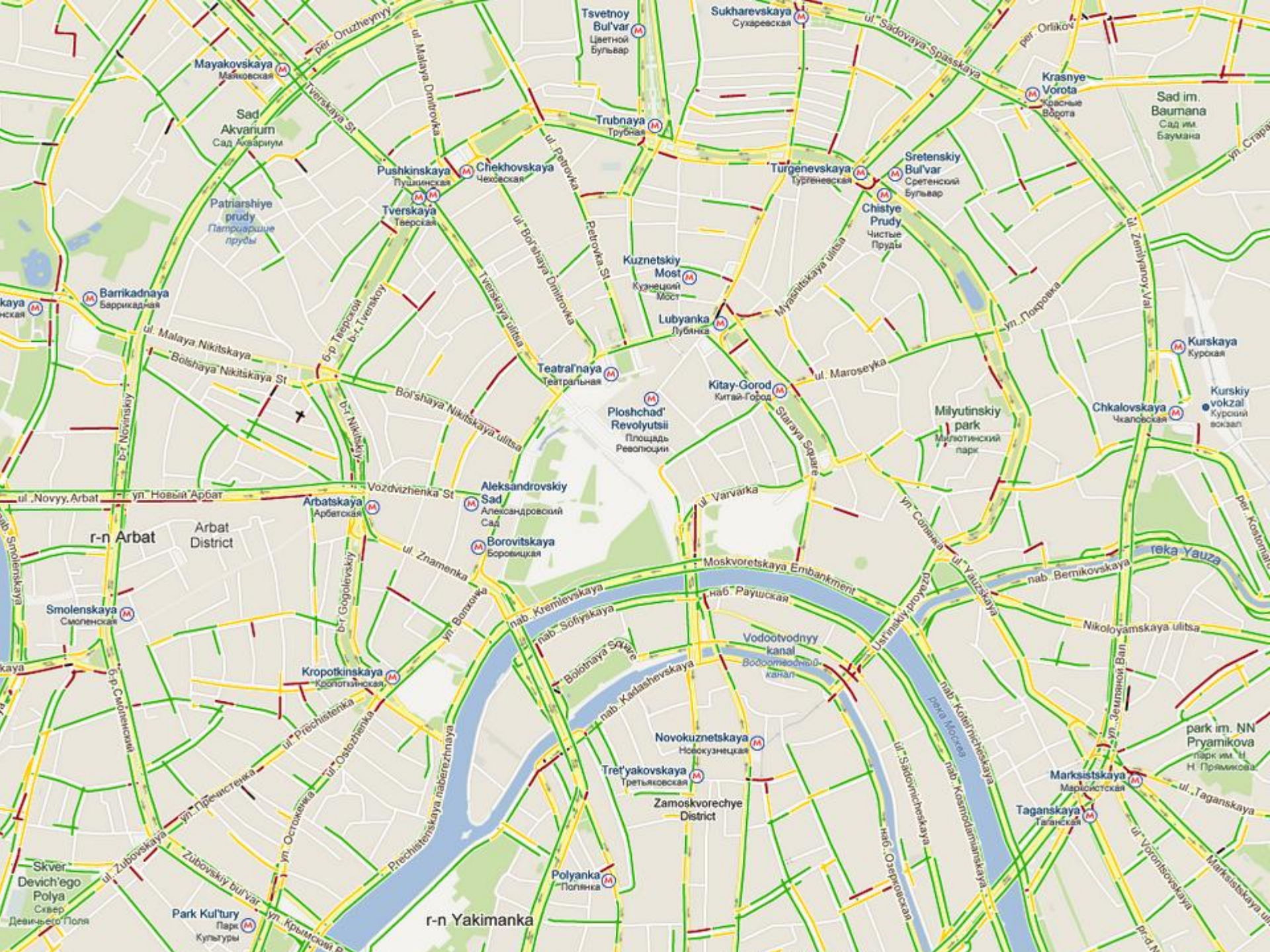


Performance Analysis Using Replay

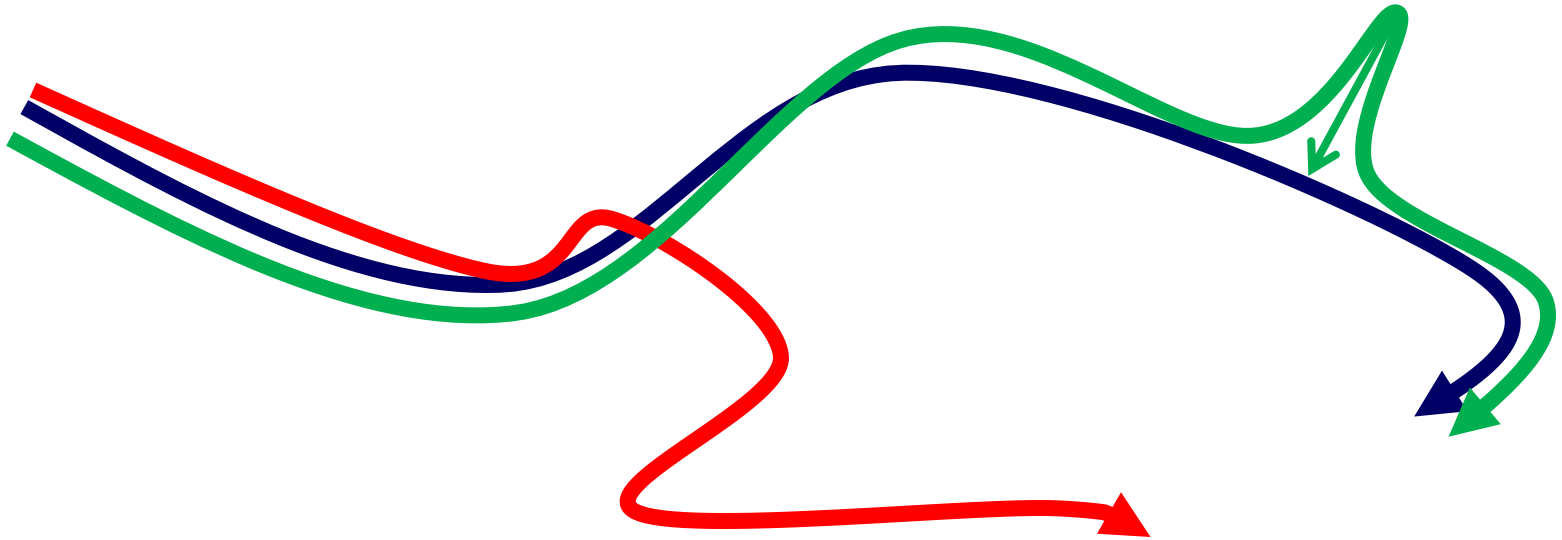
(WOZ objections Dutch municipality, 745 objections, 9583 event, f= 0.988)



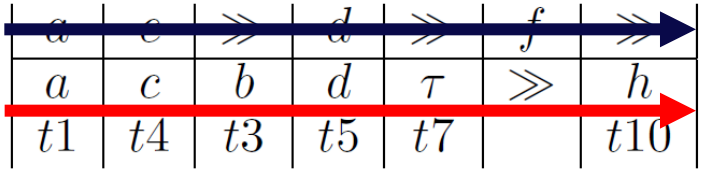
Models are like the glasses required to see and understand event data!

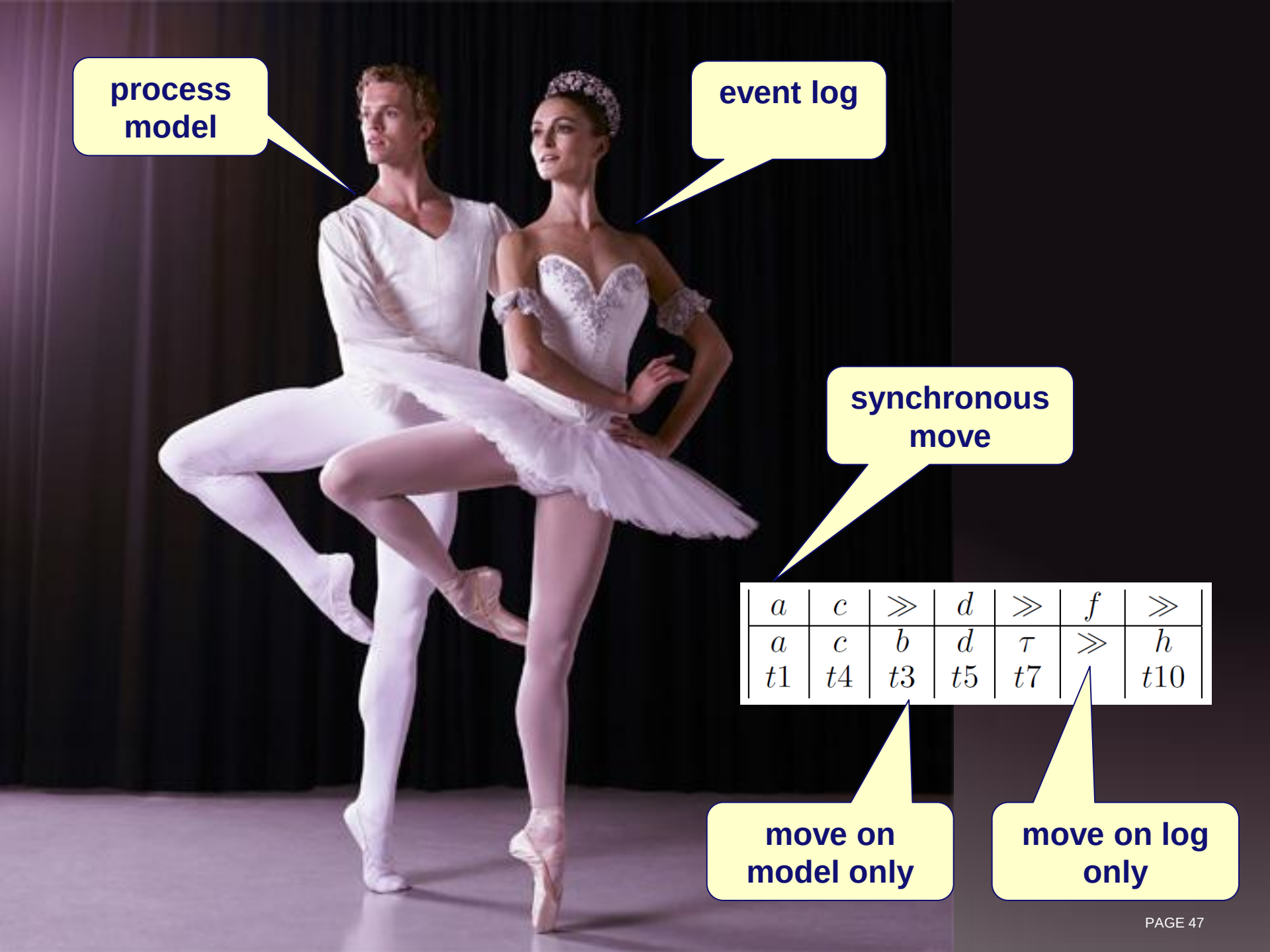

Tsvetnoy Bul'var
Цветной бульвар
Sukharevskaya
Сухаревская
ul. Sadovaya-Spasskaya
per. Orlikov
Krasnye Vorota
Красные Ворота
Sad im. Baumana
Сад им. Баумана
Mayakovskaya
Маяковская
per. Oruzheynyy
Tverskaya St
ul. Malaya Dmitrovka
Sad Akvarium
Сад Аквариум
Trubnaya
Трубная
ul. Petrovka
Petrovka St
Turgenevskaya
Тургеневская
Sretenskiy Bul'var
Сретенский бульвар
Chistye Prudy
Чистые Пруды
Pushkinskaya
Пушкинская
Chekhovskaya
Чеховская
Patriarshiye prudy
Патриаршие пруды
Tverskaya
Тверская
ul. Bol'shaya Dmitrovka
Tverskaya ulitsa
Kuznetskiy Most
Кузнецкий Мост
Myasnitskaya ulitsa
Barrikadnaya
Баррикадная
b-r Tverskoy
b-r. Tverskoy
Lubyanka
Лубянка
ul. Malaya Nikitskaya
Bolshaya Nikitskaya St
Teatral'naya
Театральная
ul. Maroseyka
Kitay-Gorod
Китай-город
Kurskaya
Курская
Kurskiy vokzal
Курский вокзал
Ploshchad' Revolyutsii
Площадь Революции
Milyutinskiy park
Милютинский парк
Chkalovskaya
Чкаловская
b-r Novinskiy
b-r Nikitskiy
Bol'shaya Nikitskaya ulitsa
Staraya Square
ul. Novyy Arbat
ул. Новый Арбат
Vozdvizhenka St
Aleksandrovskiy Sad
Александровский Сад
Arbatskaya
Арбатская
ul. Varvarka
r-n Arbat
Arbat District
Borovitskaya
Боровицкая
ul. Znamenka
Moskvoretskaya Embankment
nab. Kremlevskaya
nab. Sofiyskaya
nab. Raushskaya
reka Yauza
nab. Bernikovskaya
Smolenskaya
Смоленская
b-r Gogolevskiy
ul. Volkhonka
Vodootvodnyy kanal
Водоотводный канал
Nikoloyamskaya ulitsa
Kropotkinskaya
Кропоткинская
Bolotnaya Square
nab. Kadashevskaya
ul. Prechistenka
ul. Ostozhenka
Novokuznetskaya
Новокузнецкая
park im. NN Pryamikova
парк им. Н. Н. Прямикова
Tret'yakovskaya
Третьяковская
Zamoskvorechye District
Marksistskaya
Марксистская
Taganskaya
Таганская
ul. Taganskaya
Prechistenskaya naberezhnaya
reka Moskva
nab. Kotel'nicheskaya
ul. Sadovnicheskaya
Skver Devich'ego Polya
Сквер Девичьего Поля
Zubovskiy bul'var
Park Kul'tury
Парк Культуры
Polyanka
Полянка
r-n Yakimanka

Alignments are essential!



- **conformance checking to diagnose deviations**
- **squeezing reality into the model to do model-based analysis**





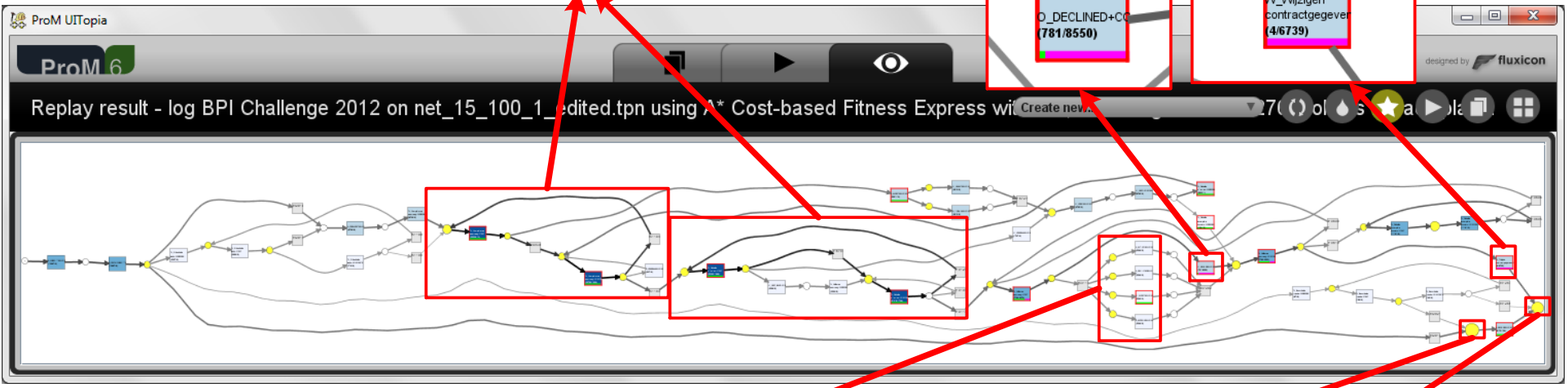
a	c	$\gg$	d	$\gg$	f	$\gg$
a	c	b	d	τ	$\gg$	h
$t1$	$t4$	$t3$	$t5$	$t7$		$t10$

Example: BPI Challenge 2012

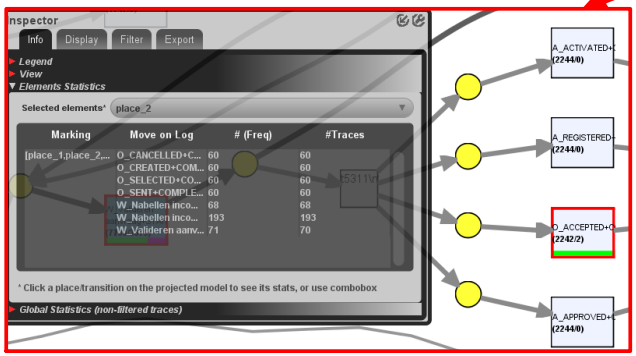
(Dutch financial institute, doi:10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f)

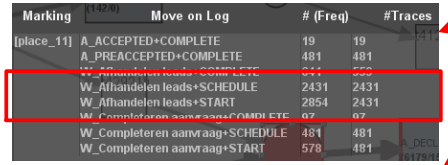
Loops of “W_Completeren aanvraag” and “W_Nabellen offertes” are often performed

“O_DECLINED” and “W_Wijzigen contractgegevens” are often skipped

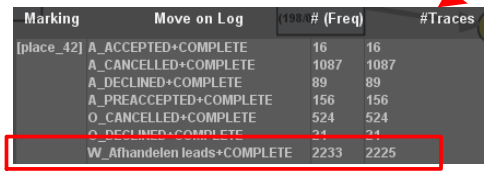


Many moves on log of “O_CANCELLED”, ”O_CREATED”, ”O_SELECTED”, “O_SENT” occurred with the same frequency value (i.e. 60) before parallel branch





Marking	Move on Log	# (Freq)	#Traces
[place_11]	A_ACCEPTED+COMPLETE	19	19
	A_PREACCEPTED+COMPLETE	481	481
	W_Afhandelen leads+SCHEDULE	2431	2431
	W_Afhandelen leads+START	2854	2431
	W_Completeren aanvraag+SCHEDULE	481	481
	W_Completeren aanvraag+START	578	481



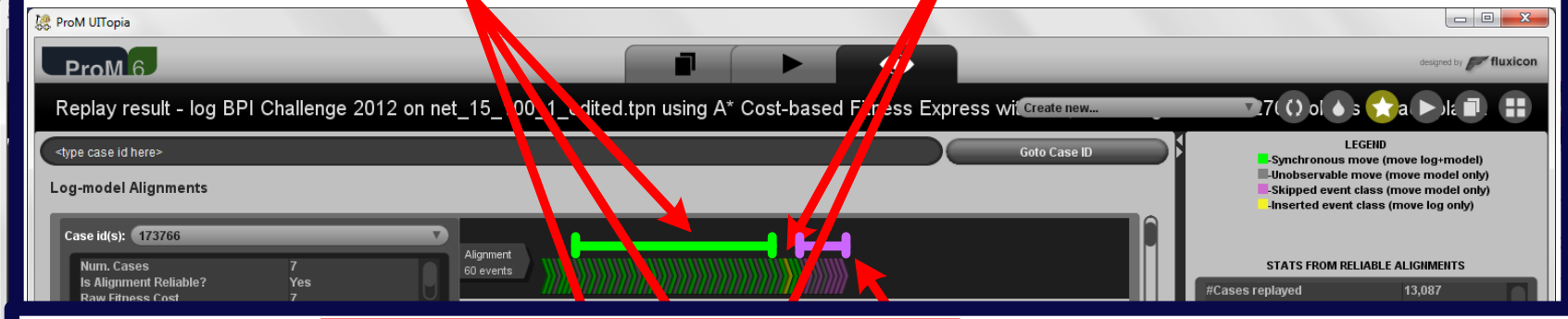
Marking	Move on Log	# (Freq)	#Traces
[place_42]	A_ACCEPTED+COMPLETE	16	16
	A_CANCELLED+COMPLETE	1087	1087
	A_DECLINED+COMPLETE	89	89
	A_PREACCEPTED+COMPLETE	156	156
	O_CANCELLED+COMPLETE	524	524
	W_Afhandelen leads+COMPLETE	2233	2225

Many moves on log of “W_Afhandelen leads” (> 2200 times) occurred in the end of traces

Work of Arya Adriansyah (Replay project)

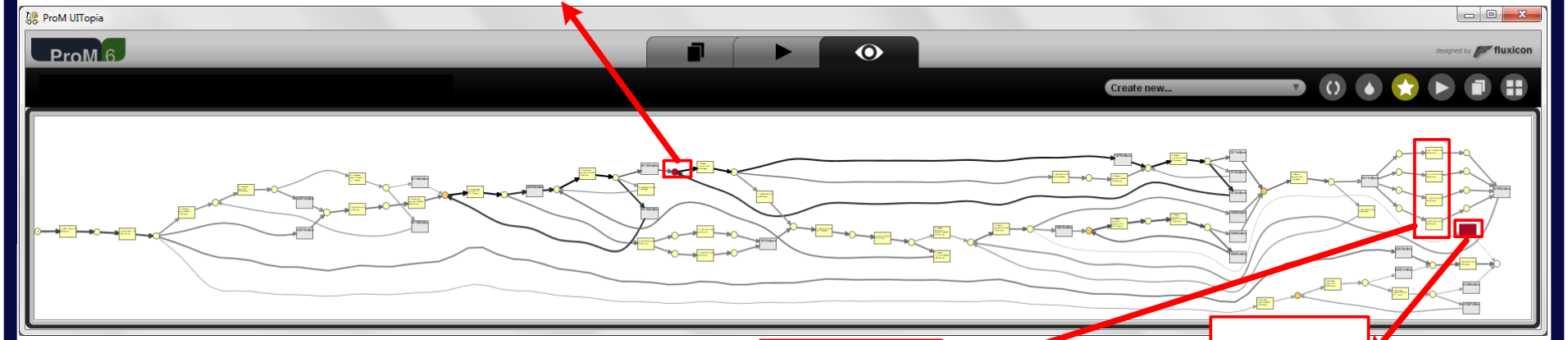
Synchronous moves of “Completeren aanvraag”

Move on log of “Completeren aanvraag”

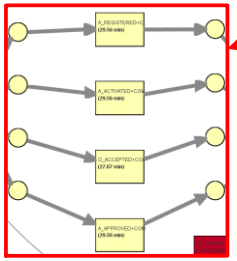


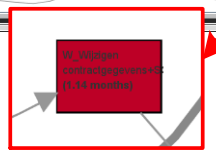
Property	Min.	Max.	Avg.	Std. Dev	Freq.
Waiting time	0.00 ms	29.78 days	2.83 days	3.30 days	24,229
Synchronization time	0.00 ms	0.00 ms	0.00 ms	0.00 ms	24,229
Sojourn time	0.00 ms	29.78 days	2.83 days	3.30 days	24,229

The average waiting time for the input place of “W_Nabellen offertes+START” is very long (2.83 days) compares to the average waiting time of other places



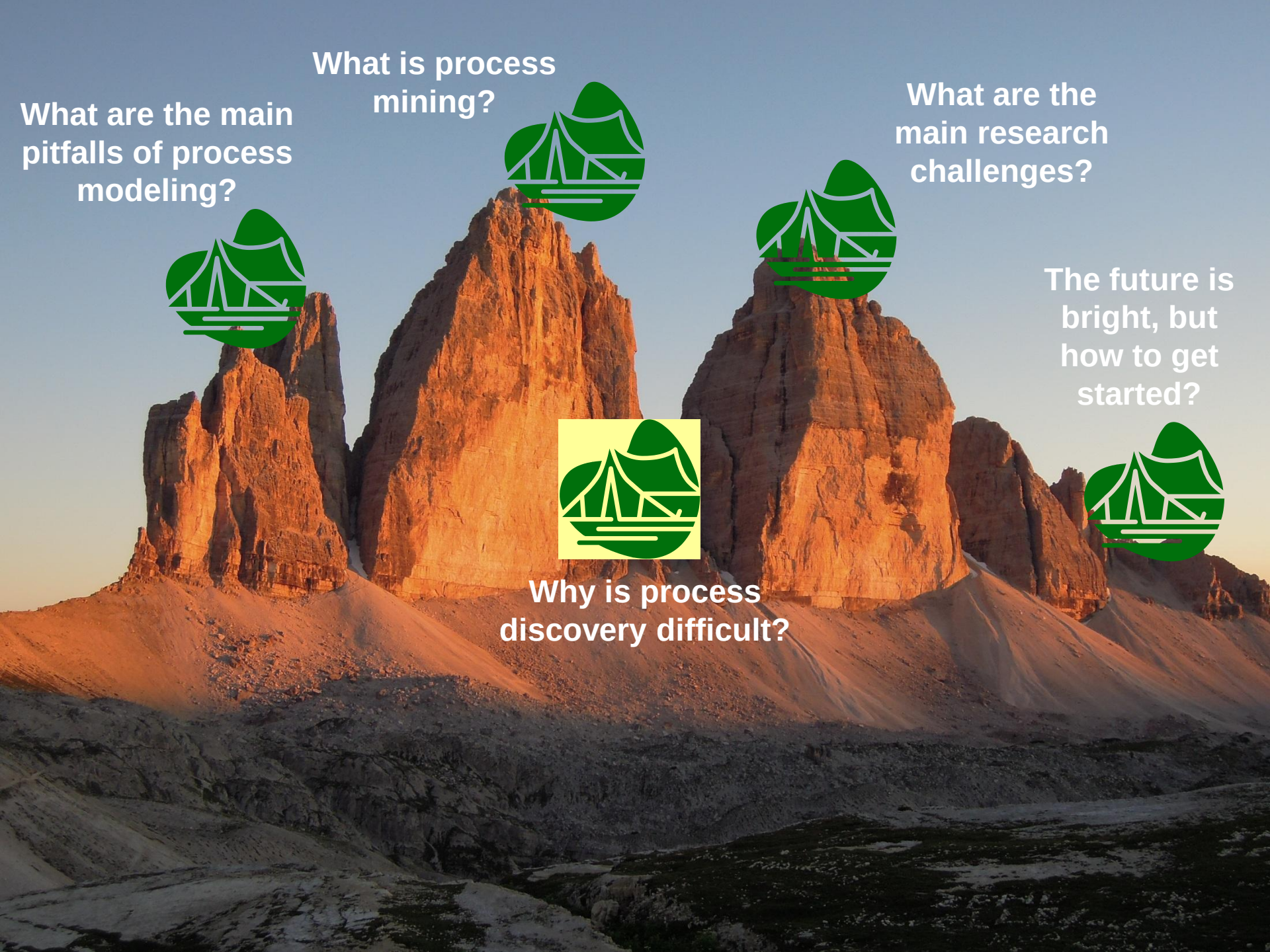
“O_ACCEPTED” has average sojourn time of 27.07 minutes, while “A_REGISTERED”, ”A_ACTIVATED”, and “A_APPROVED” have average sojourn time of 29.56 minutes





Property	Min.	Max.	Avg.	Std. Dev	Freq.
Throughput time	0.00 ms	0.00 ms	0.00 ms	0.00 ms	4
Waiting time	1.55 hours	3.43 months	1.14 months	1.55 months	4
Sojourn time	1.55 hours	3.43 months	1.14 months	1.55 months	4
#Unique cases ...	4				

Activity “W_Wijzigen contractgegevens” is the bottleneck, but it occured rarely (only 4 times)


What are the main pitfalls of process modeling?
What is process mining?
What are the main research challenges?
The future is bright, but how to get started?
Why is process discovery difficult?

Language identification in the limit (Mark Gold 1967)

A language is learnable in the limit if there exists a perfect child that generates only finitely many hypotheses.

Language identification in the limit by E Mark Gold, Information and Control, 10(5):447–474, 1967.

Learning is not easy …

- Even simple languag…
  regular languag…
  learnable i…
- M… …with or
  …ive examples,
  …icies, etc.

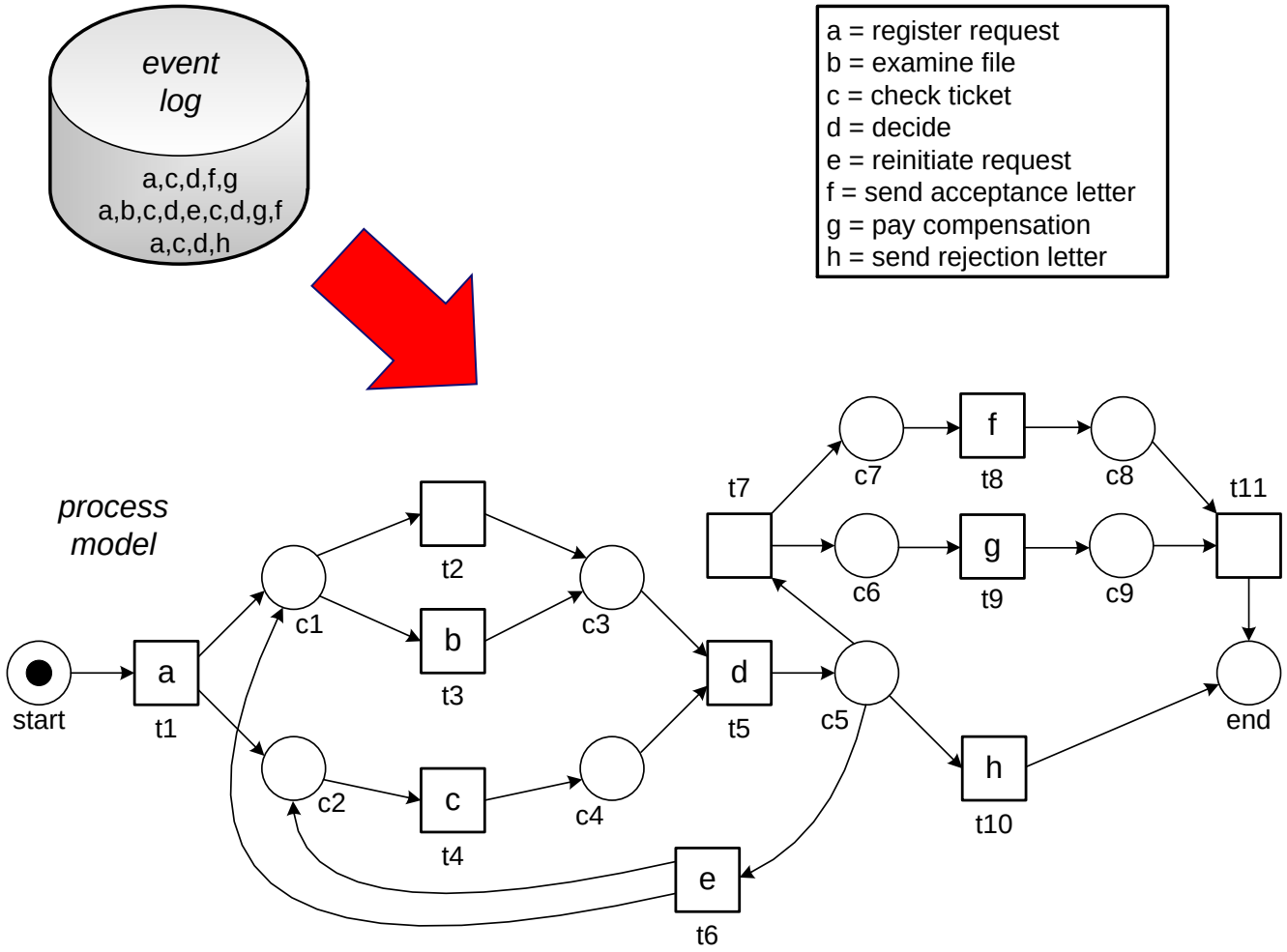
Classical approaches (before 1998) did not consider concurrency and definitely not end-to-end business process models.

…tence ≅ trace in event log

language ≅ process model


at the start of the century,
process mining emerged as a
new research topic
remarkable progress over a
relatively short period
See keynote at Process Mining Camp 2013, http://fluxicon.com/camp/2013/

Process discovery challenge
(oversimplied no resources, data, etc.)



Process discovery algorithms
(small selection)

automata-based learning

distributed genetic mining

heuristic mining

language-based regions

genetic mining

state-based regions

stochastic task graphs

LTL mining

neural networks

fuzzy mining

hidden Markov models

mining block structures

α algorithm

conformal process graph

multi-phase mining

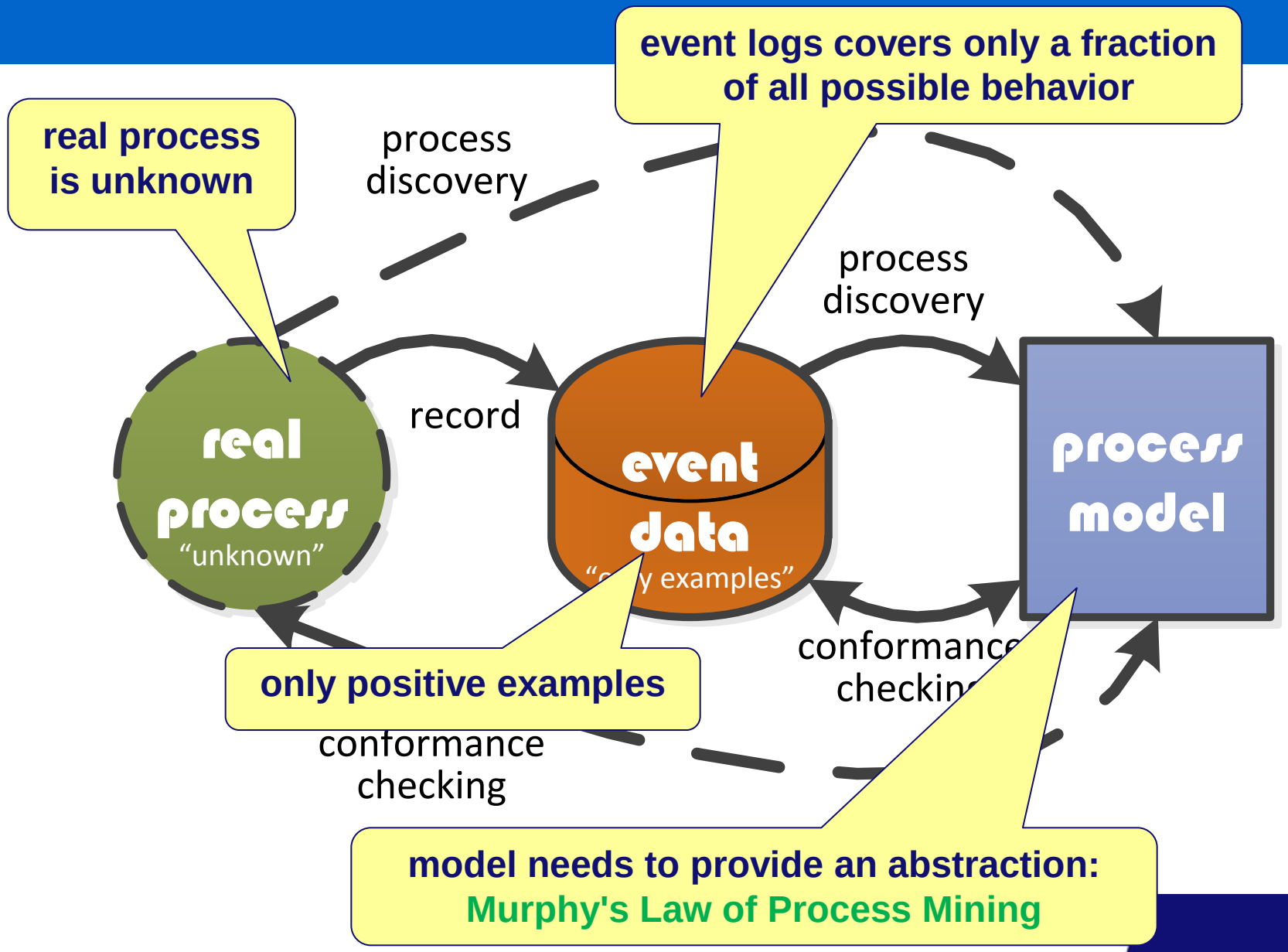
partial-order based mining

α# algorithm

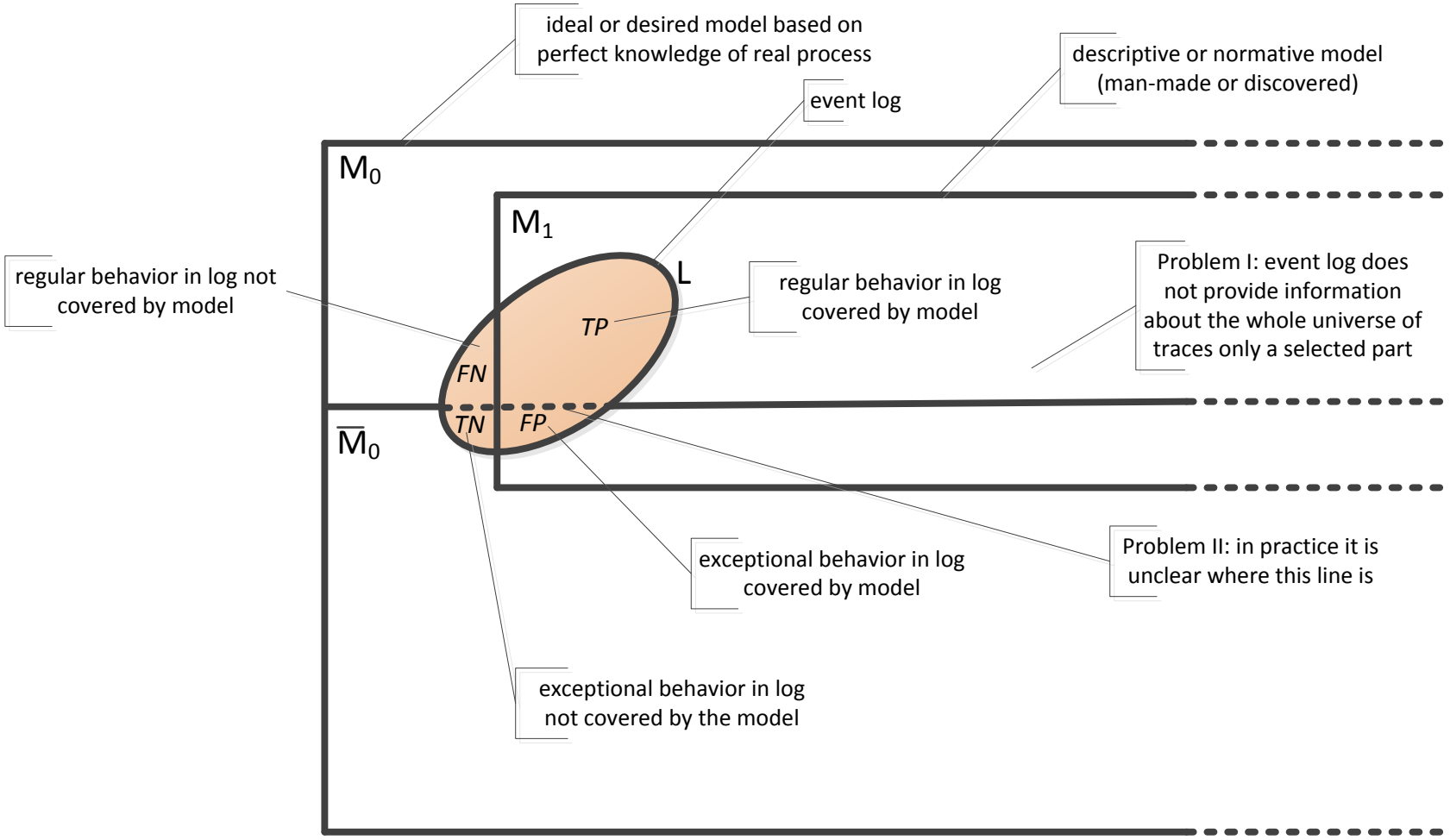
ILP mining

α++ algorithm

Problem

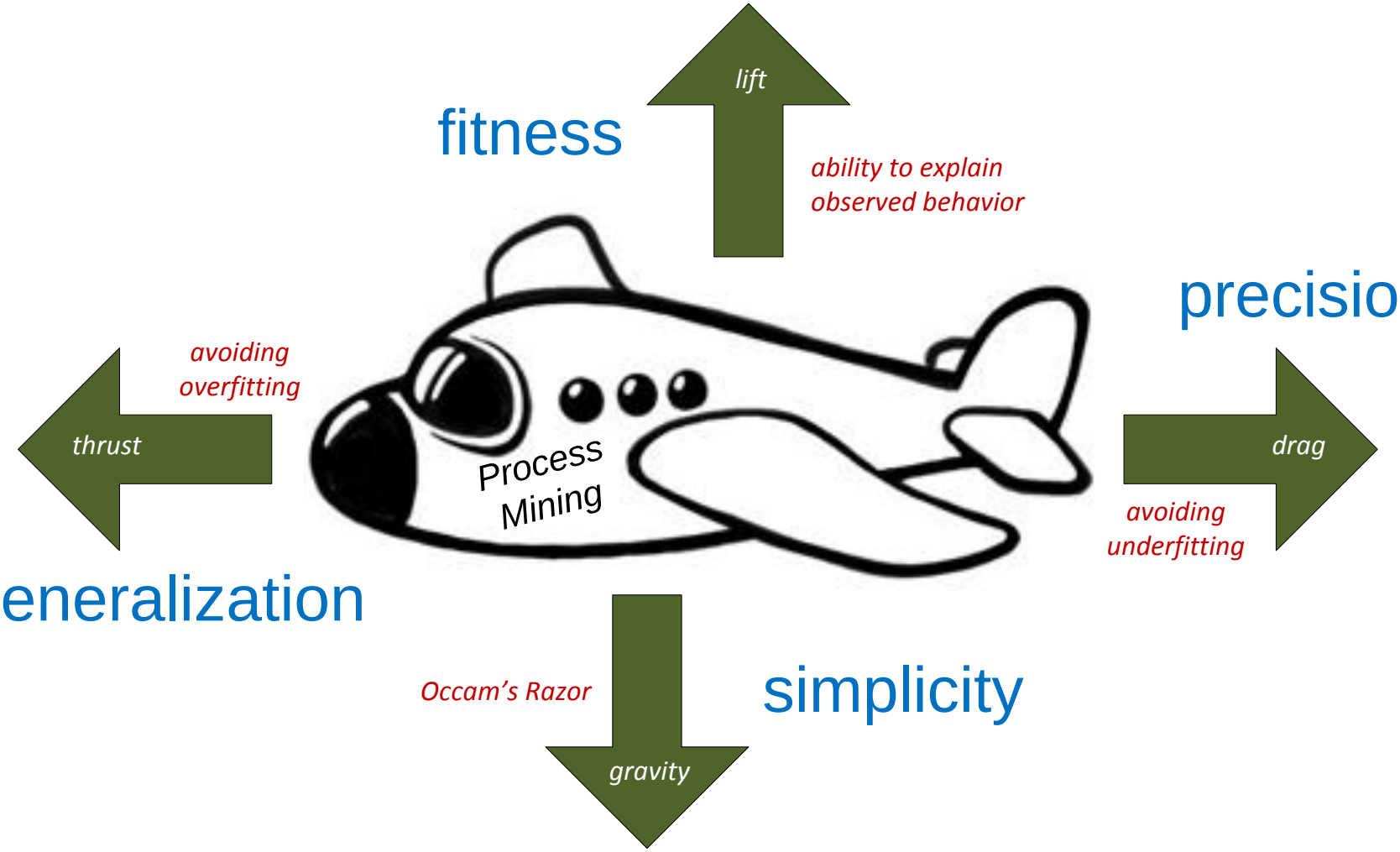

event logs covers only a fraction of all possible behavior
real process is unknown
process discovery
process discovery
record
real process
"unknown"
event data
process model
only positive examples
conformance checking
conformance checking
model needs to provide an abstraction:
Murphy's Law of Process Mining

We only have example behavior (event log) and do not know the real process …



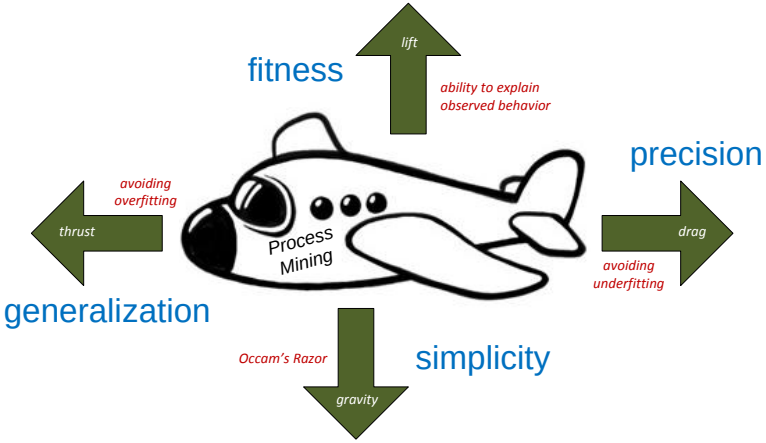
W.M.P. van der Aalst. Mediating Between Modeled and Observed Behavior: The Quest for the "Right" Process. In *IEEE International Conference on Research Challenges in Information Science (RCIS 2013)*, pages 31-43. IEEE Computing Society, 2013.

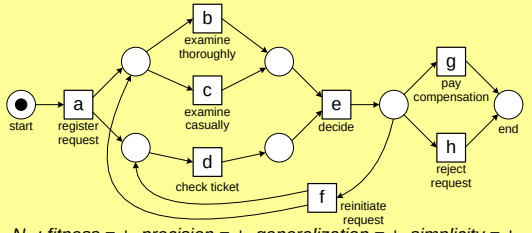
Balance four forces



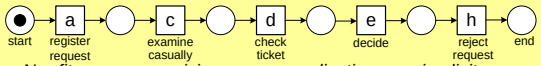
Example: one log four models

#	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	

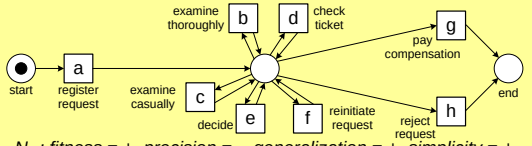




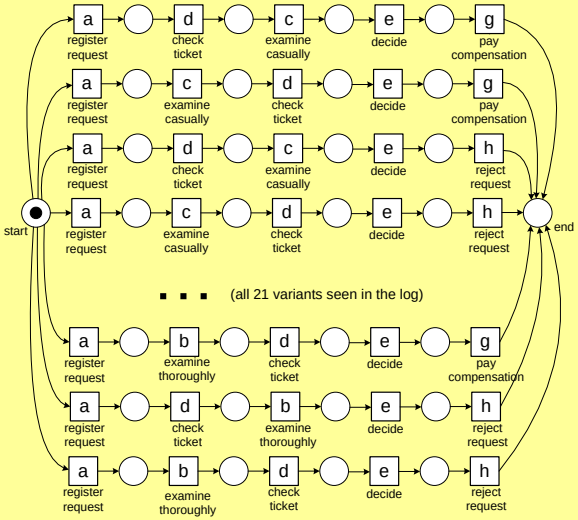
N_1 : *fitness = +, precision = +, generalization = +, simplicity = +*



N_2 : *fitness = -, precision = +, generalization = -, simplicity = +*

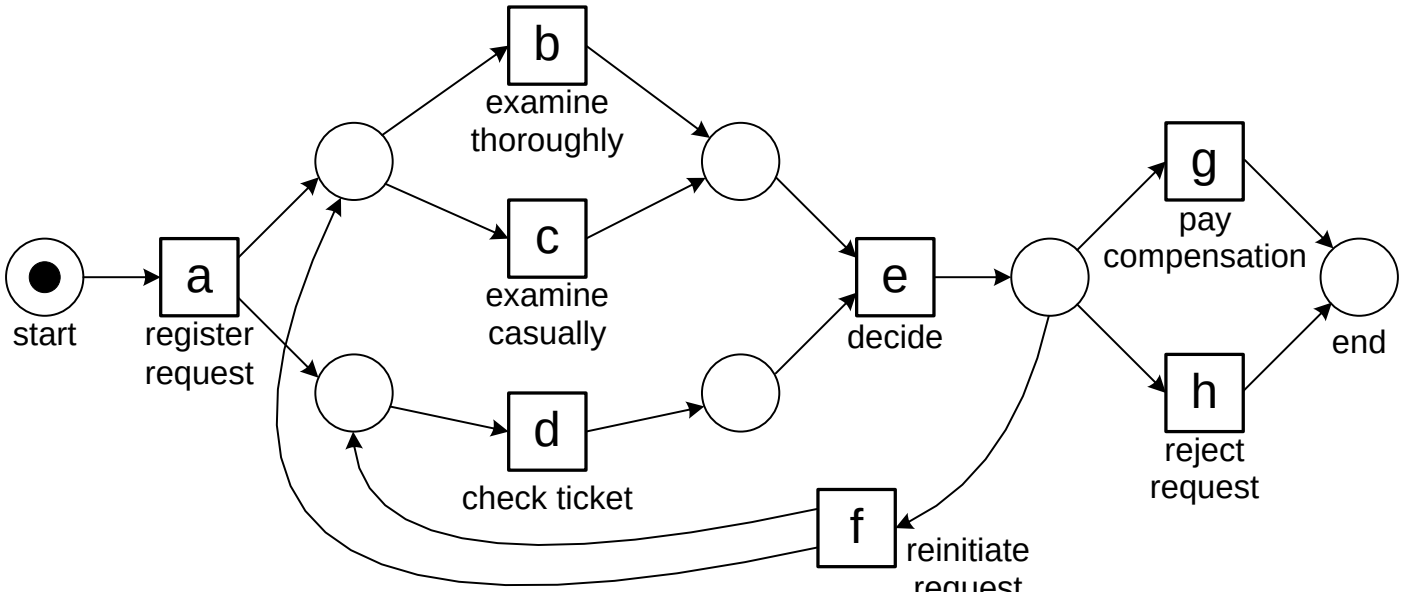


N_3 : *fitness = +, precision = -, generalization = +, simplicity = +*



N_4 : *fitness = +, precision = +, generalization = -, simplicity = -*

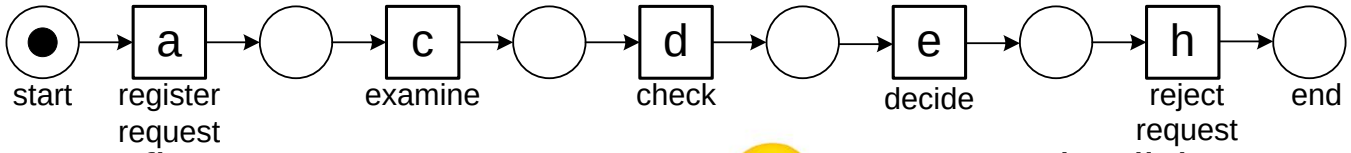
Model N_1



N_1 : fitness = +, precision = +, generalization = +, simplicity = +

#	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	

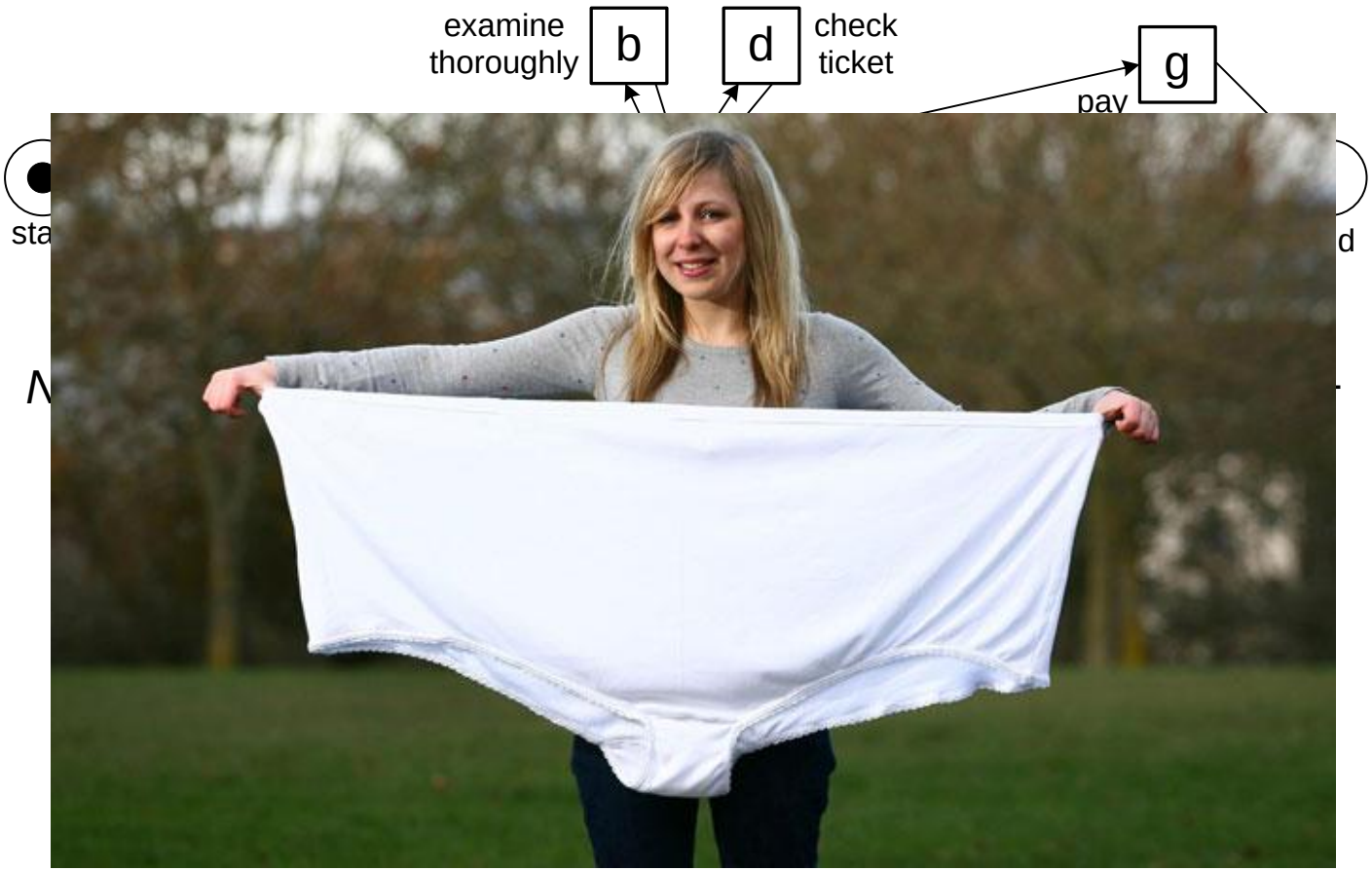
Model N_2



N_2 : fitness = *, simplicity = +*

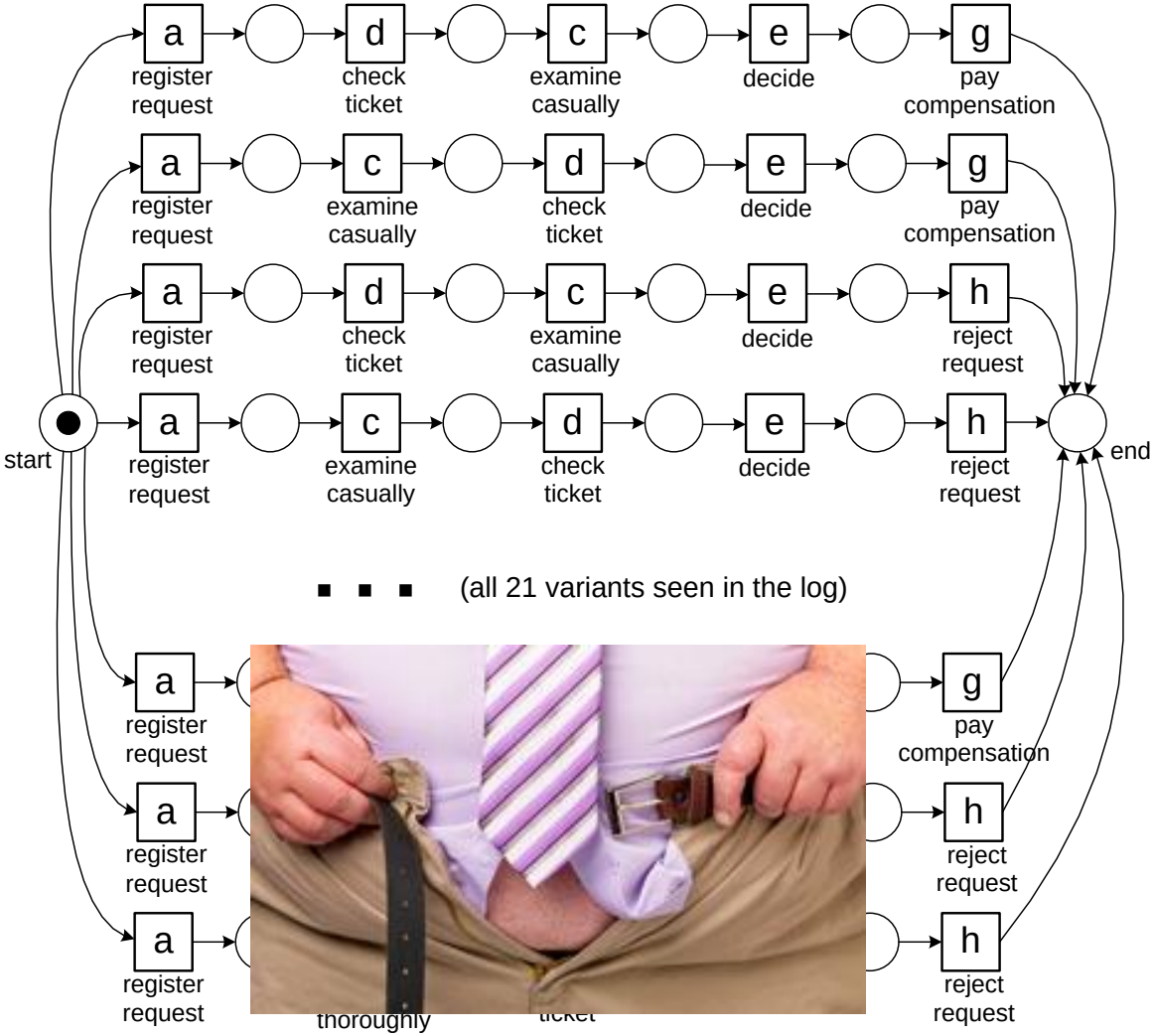
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

Model N_3



#	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	

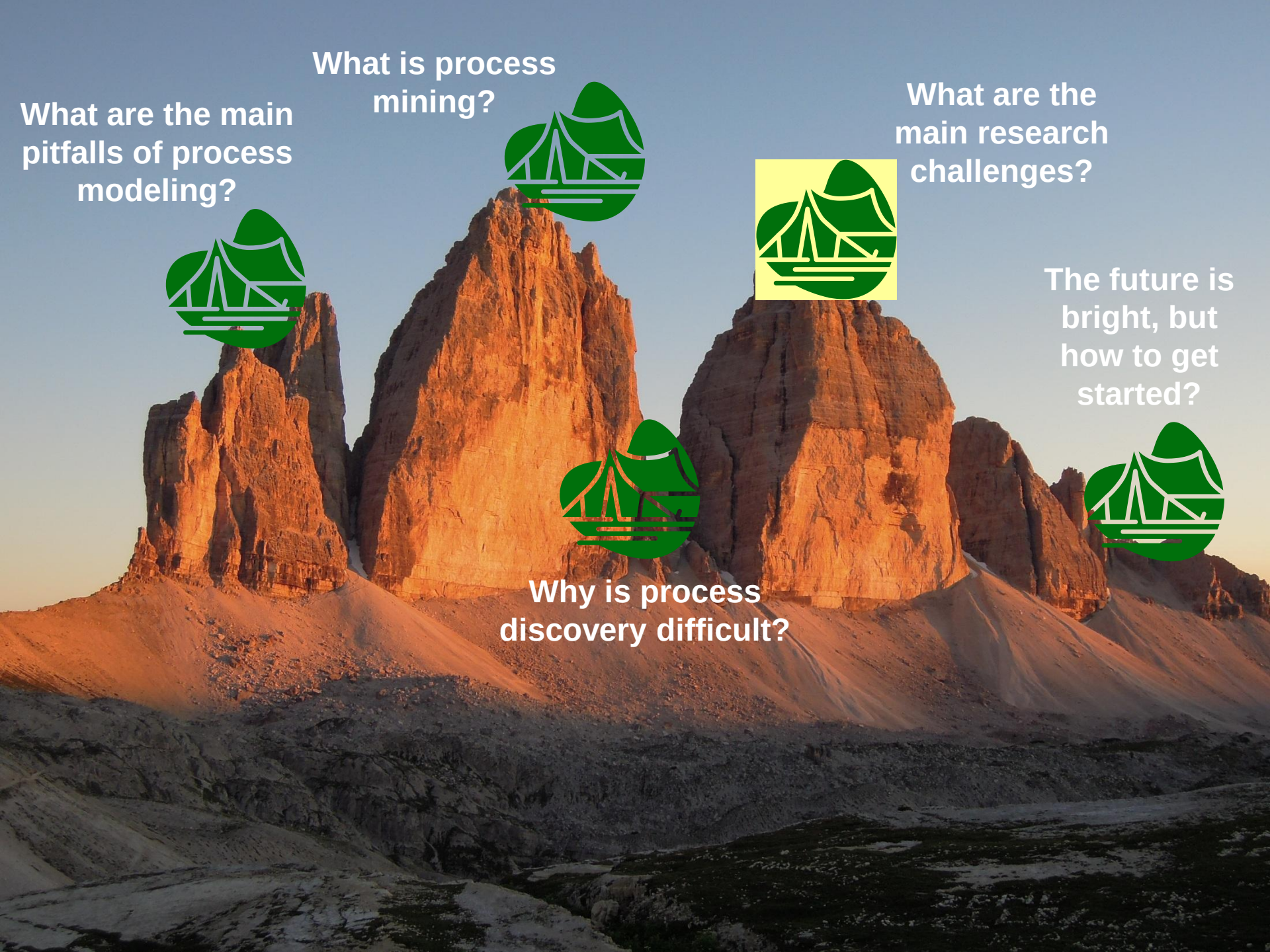
Model N_4



N_4 : fitness = +, precision = +, generalization = -, simplicity = -

#	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	


1
formal (not just a picture)
2
fast (should not take years)
ability to balance all conformance dimensions (fitness, precision, generalization, and simplicity) incl. noise
3
4
sound (result should at least be free of deadlocks, etc.)
5
provide guarantees (not just a best effort)


What are the main pitfalls of process modeling?
What is process mining?
What are the main research challenges?
The future is bright, but how to get started?
Why is process discovery difficult?


Finding sheep with five legs
we are getting close…


Distributing process mining problems to cope with big data

On-the-fly process mining

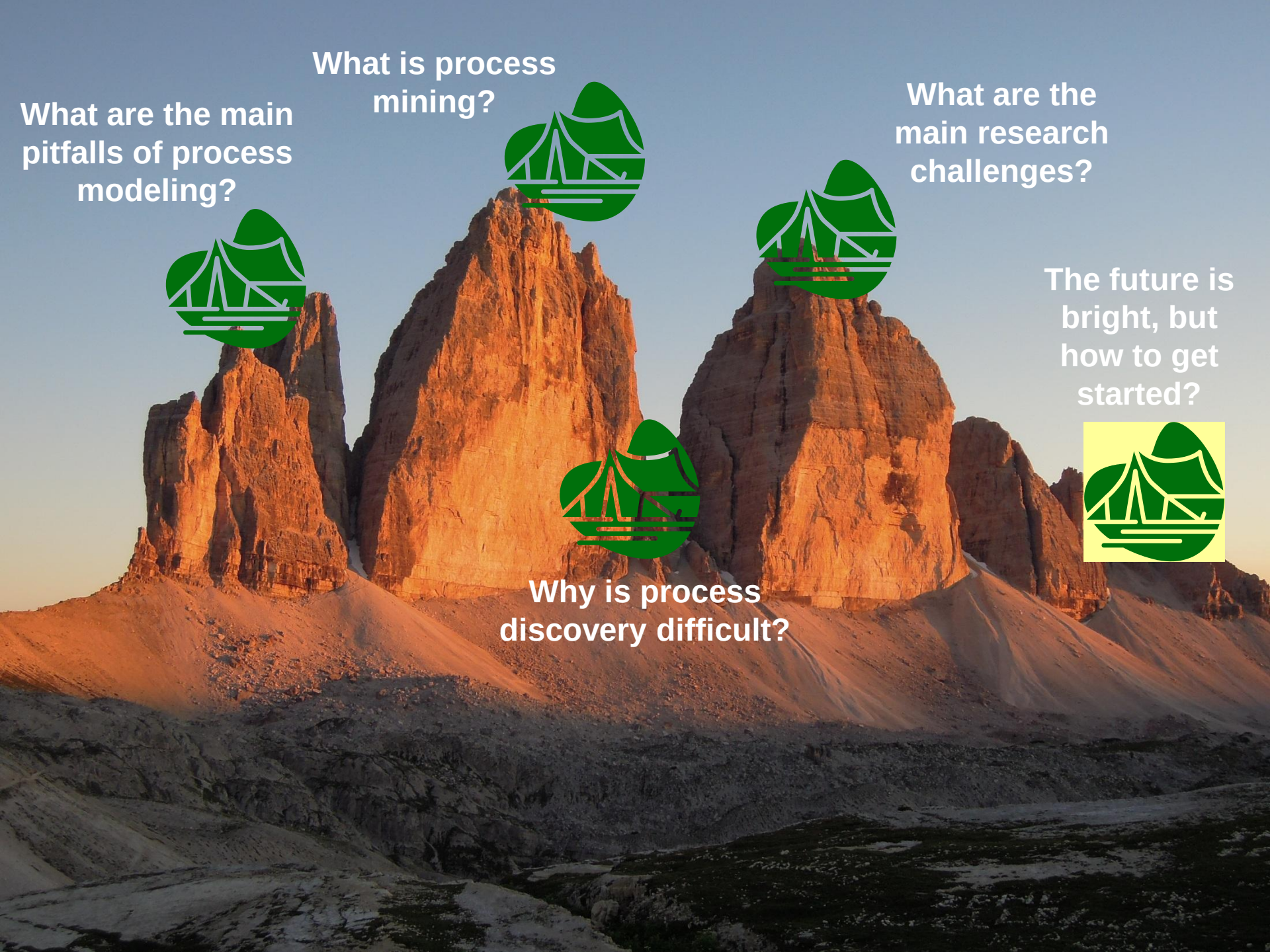
Operational support


Concept drift

cross-organizational / comparative process mining

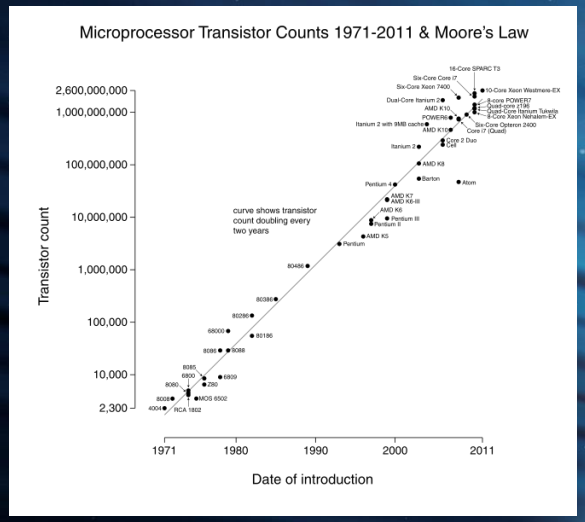

context aware
process mining

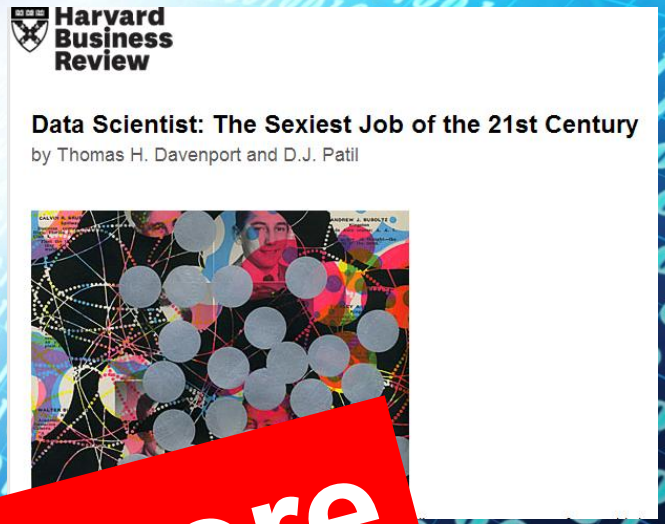
Supporting the process of process mining

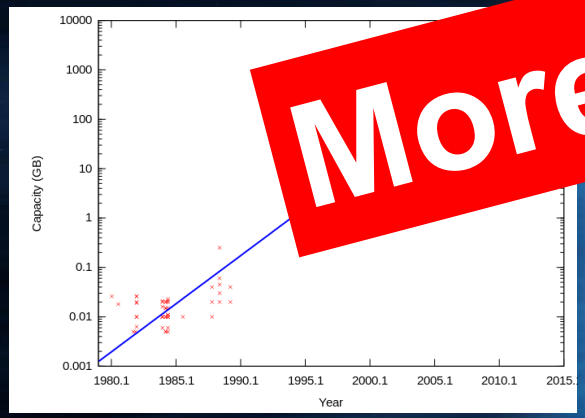

What are the main pitfalls of process modeling?
What is process mining?
What are the main research challenges?
The future is bright, but how to get started?
Why is process discovery difficult?


BRIGHT
FUTURE
AHEAD

The Sexiest Job of the 21st century (thanks to Moore's Law)







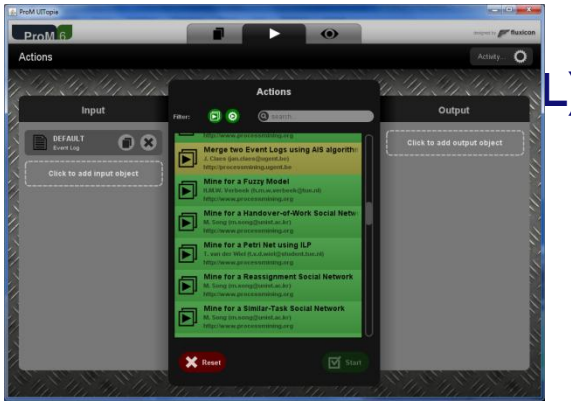
More from Moore

1965

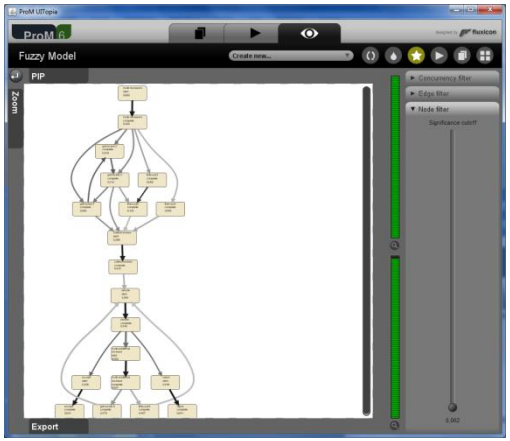
How to get started?

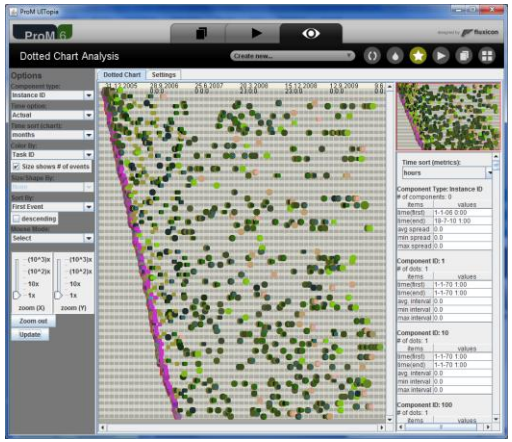
600+ plug-ins available covering the whole process mining spectrum

ProM
process mining workbench



L)

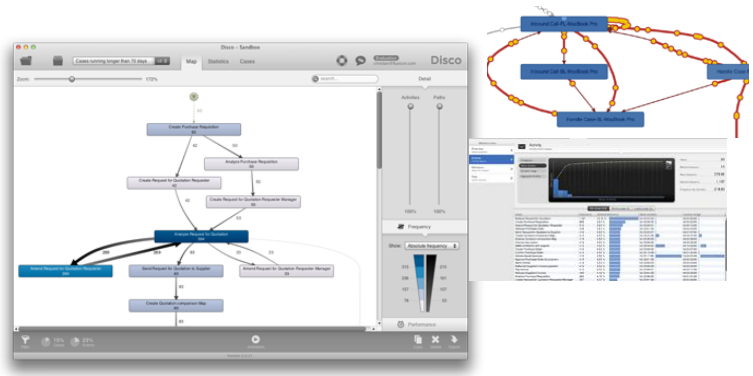


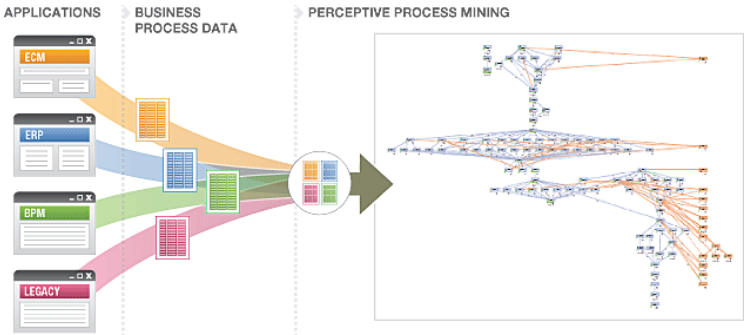


Download from: www.processmining.org

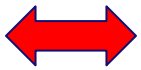
Commercial Alternatives

- **Disco** (Fluxicon)
- **Perceptive Process Mining** (before Futura Reflect and BPM|one)
- ARIS Process Performance Manager
- QPR ProcessAnalyzer
- Interstage Process Discovery (Fujitsu)
- Discovery Analyst (StereoLOGIC)
- XMAnalyzer (XMPro)
- …





How to Get Started?

Collect event data  **Collect questions**

- **Minimal requirement: events referring to an activity name and a process instance.**
- **Good to have: timestamps, resource information, additional data elements.**
- **Challenges: scoping and sometimes correlation.**

- **What kind problems would you like to address (cost, time, risk, compliance, service, etc.)?**
- **Related to discovery, conformance, enhancement?**
- **Iterative process: can be "curiosity driven" initially.**

Join our expedition: Mine your processes

processmining.org