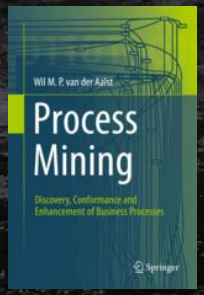
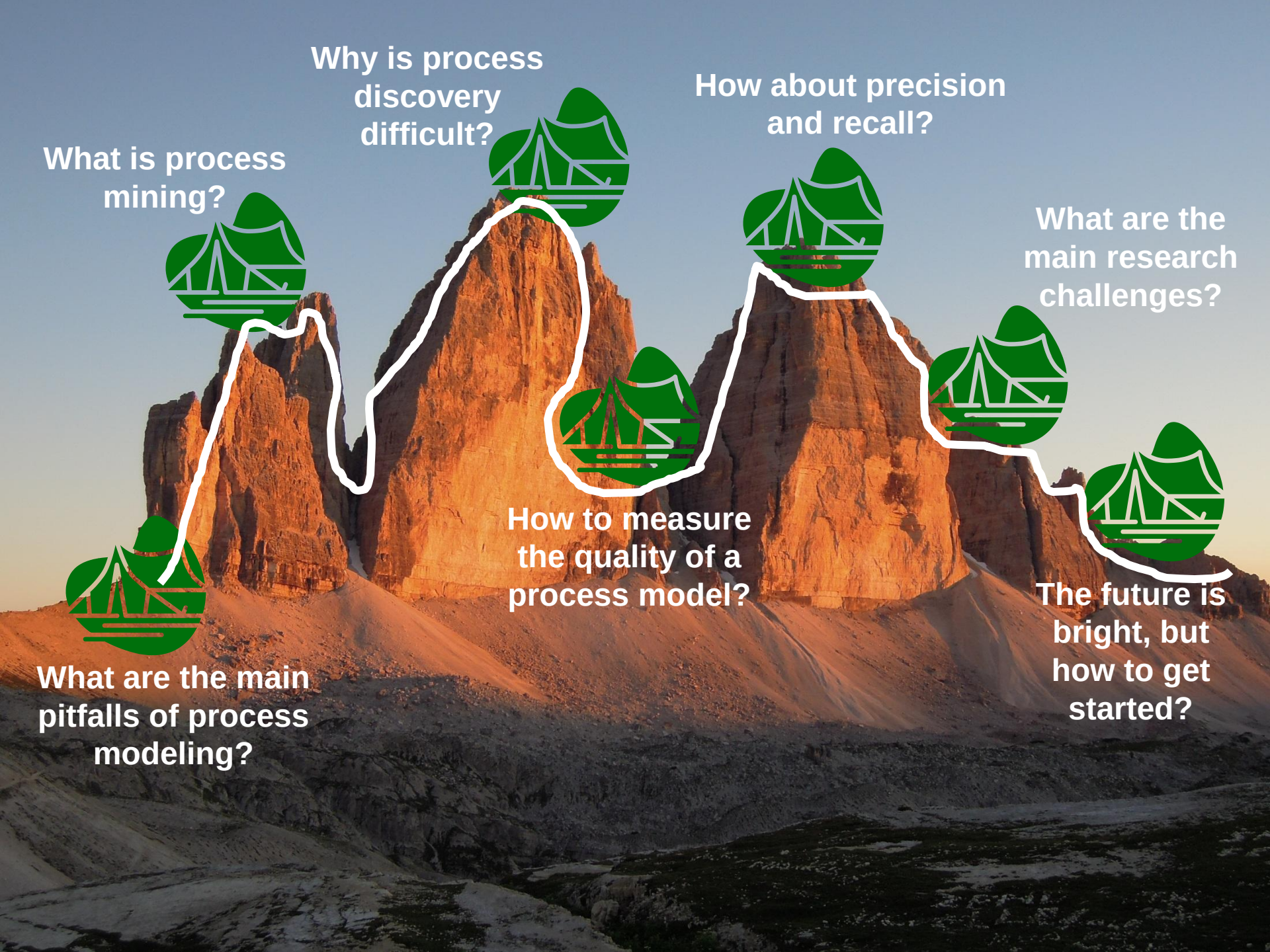
Mediating Between Modeled and Observed Behavior

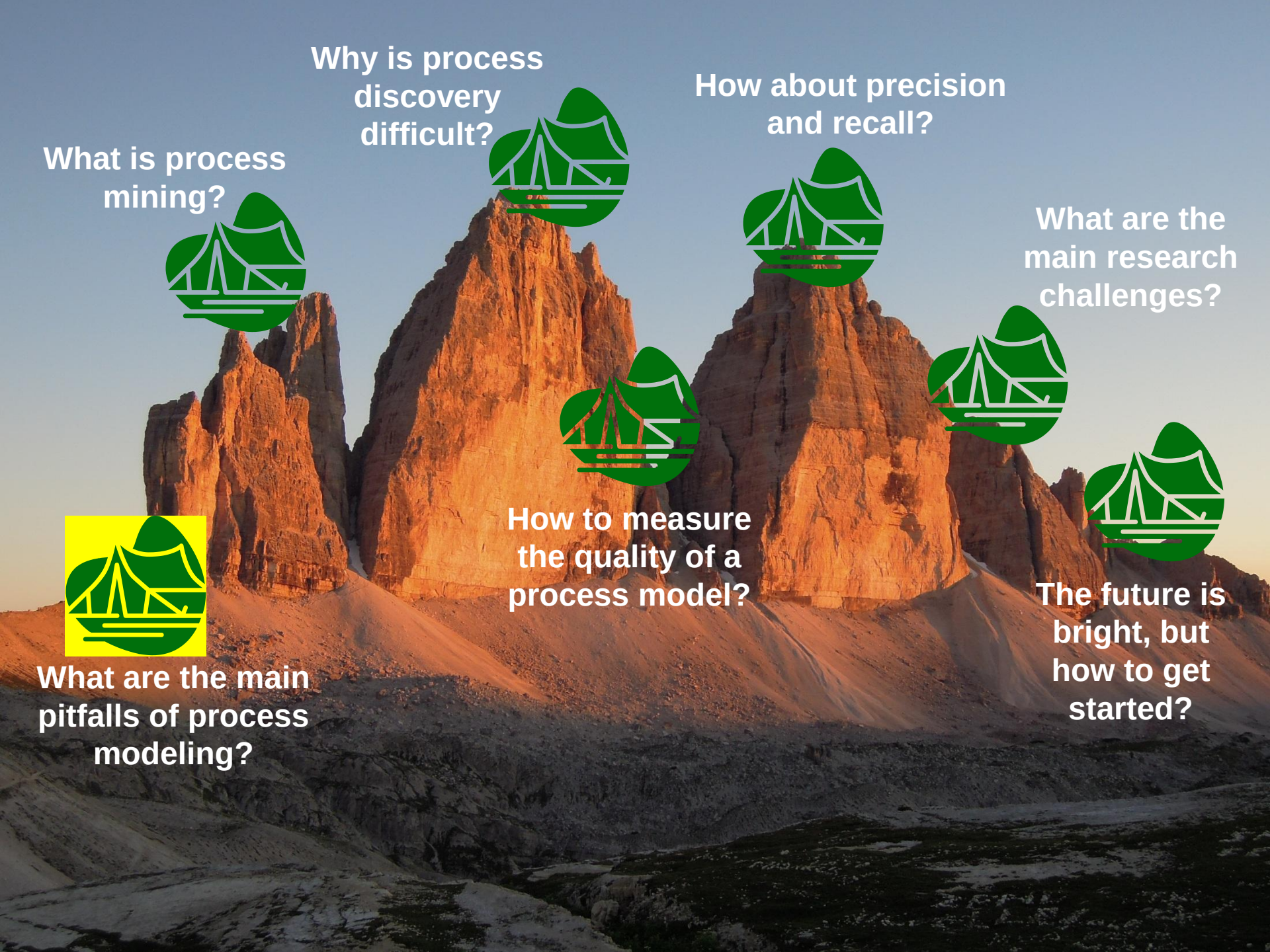
The Quest for the "Right" Process

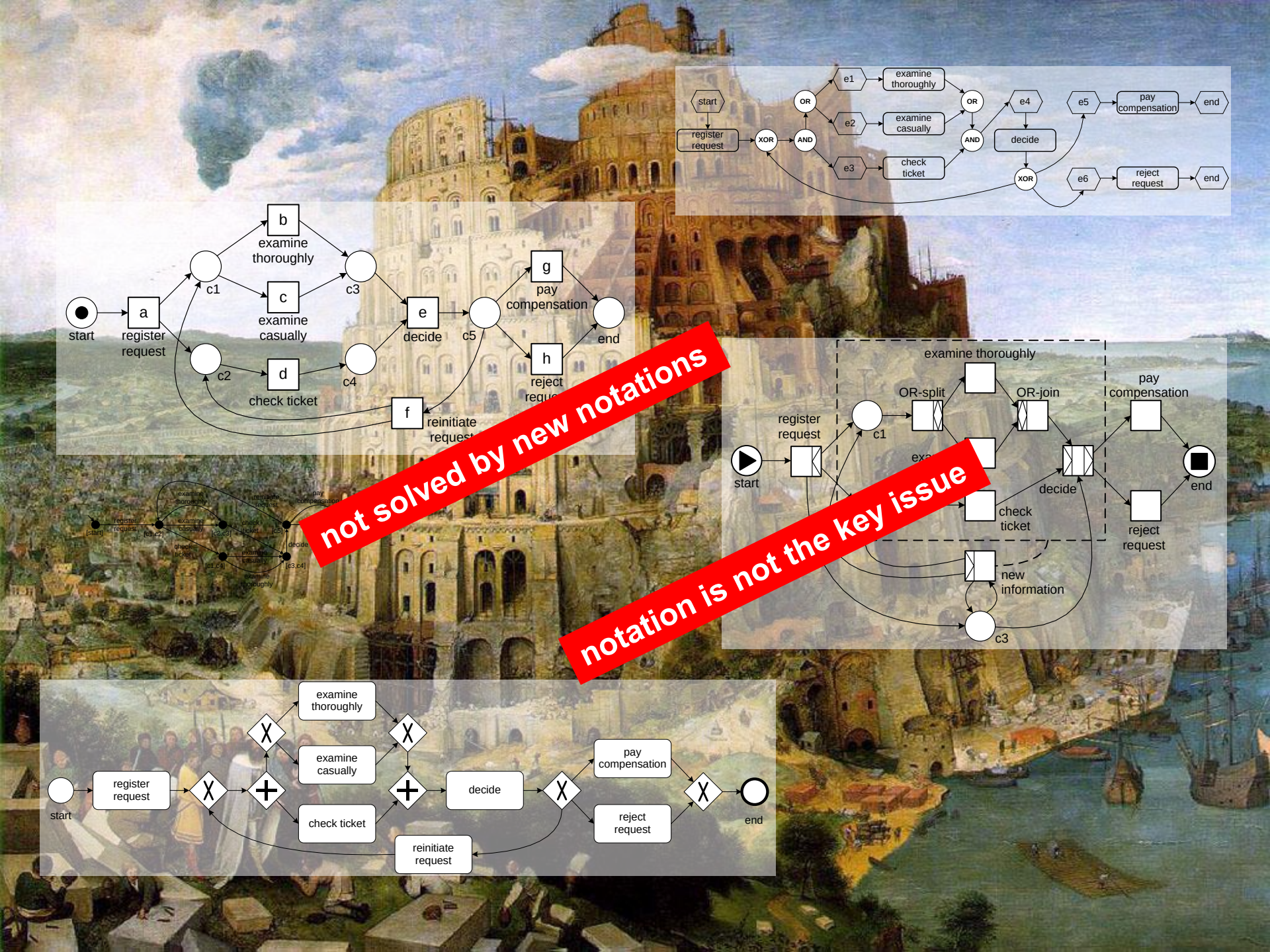
ROIS 2013

prof.dr.ir. Wil van der Aalst

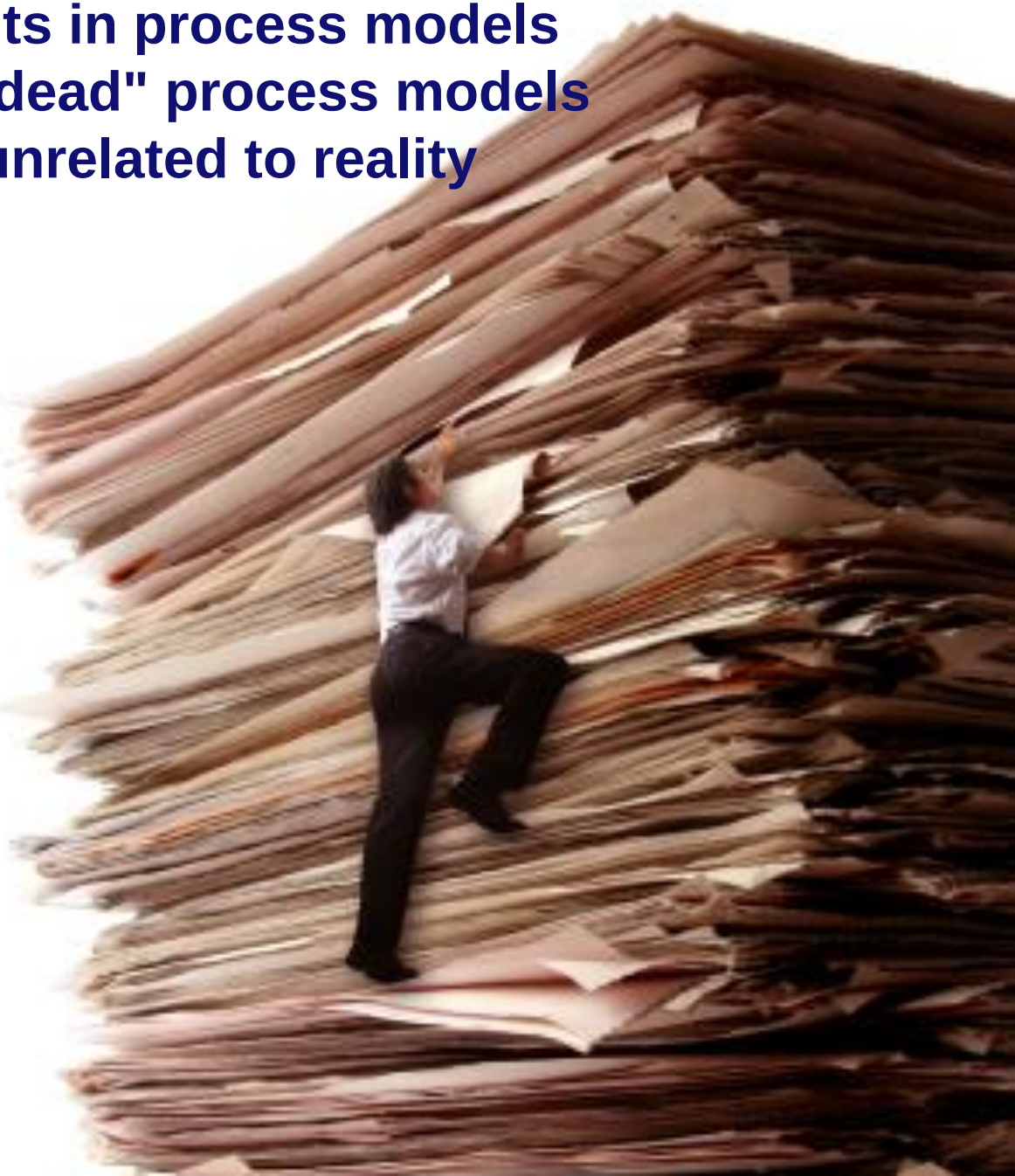



What is process mining?
Why is process discovery difficult?
How about precision and recall?
What are the main research challenges?
What are the main pitfalls of process modeling?
How to measure the quality of a process model?
The future is bright, but how to get started?


What is process mining?
Why is process discovery difficult?
How about precision and recall?
What are the main research challenges?
How to measure the quality of a process model?
What are the main pitfalls of process modeling?
The future is bright, but how to get started?

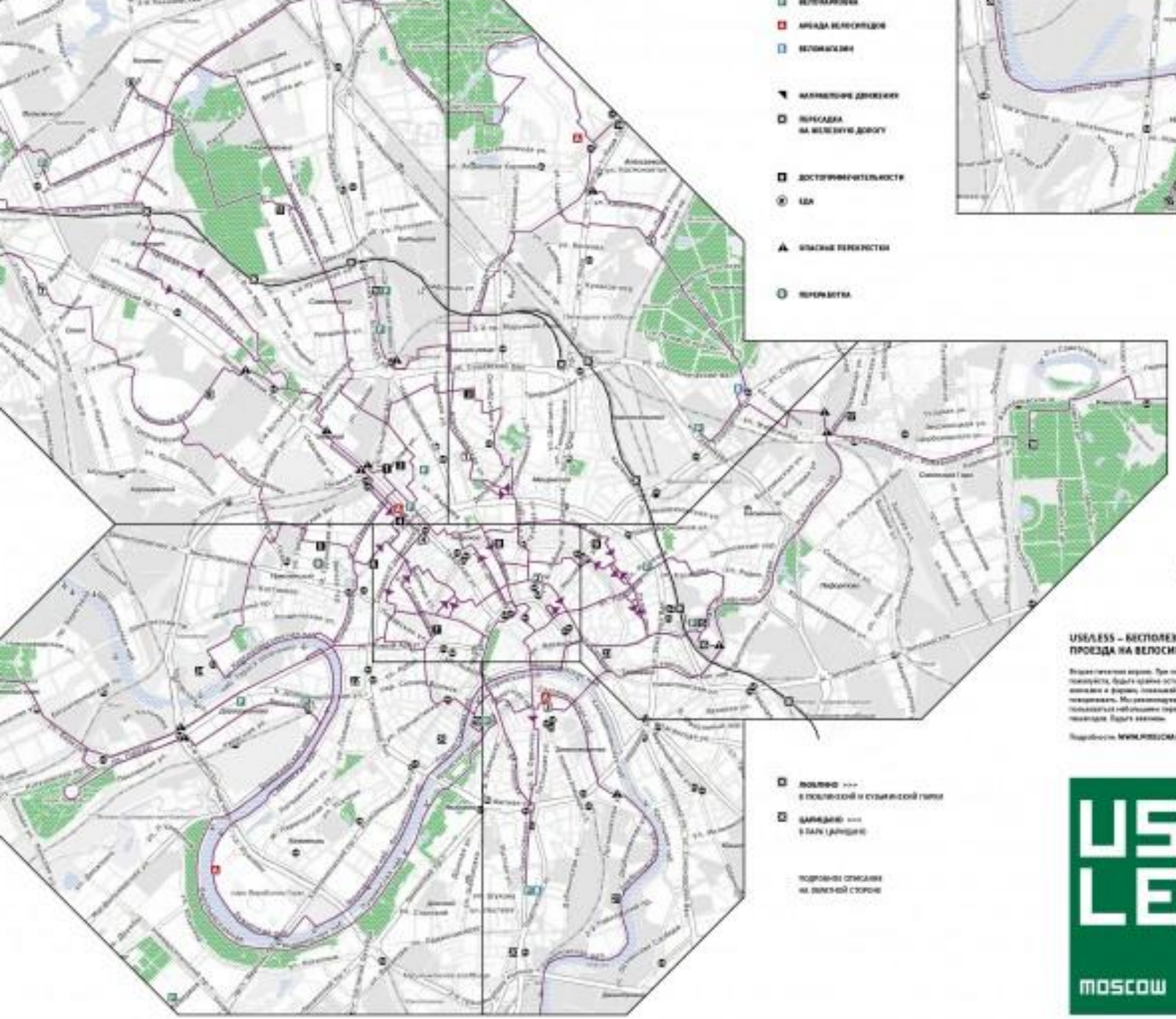

start
register request
XOR
AND
OR
e1
examine thoroughly
e2
examine casually
e3
check ticket
OR
AND
e4
decide
XOR
e5
pay compensation
end
e6
reject request
end
start
a
register request
c1
b
examine thoroughly
c
examine casually
c2
d
check ticket
c3
c4
e
decide
c5
g
pay compensation
h
reject
end
f
reinitiate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

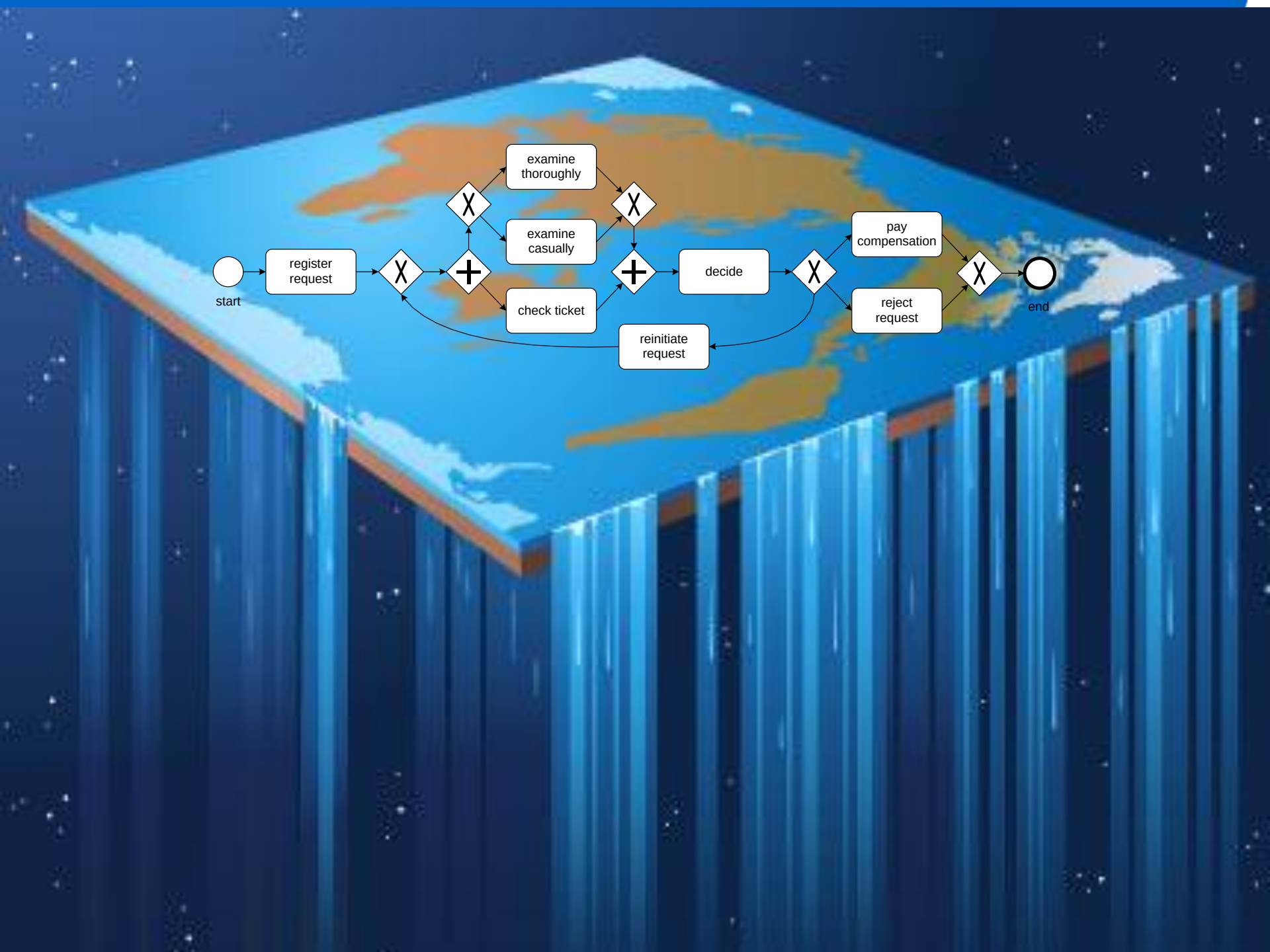


USELESS – БЕСПОЛЕЗНАЯ СХЕМА
ПРОЕЗДА НА ВЕЛОСИПЕДЕ ПО МОСКВЕ

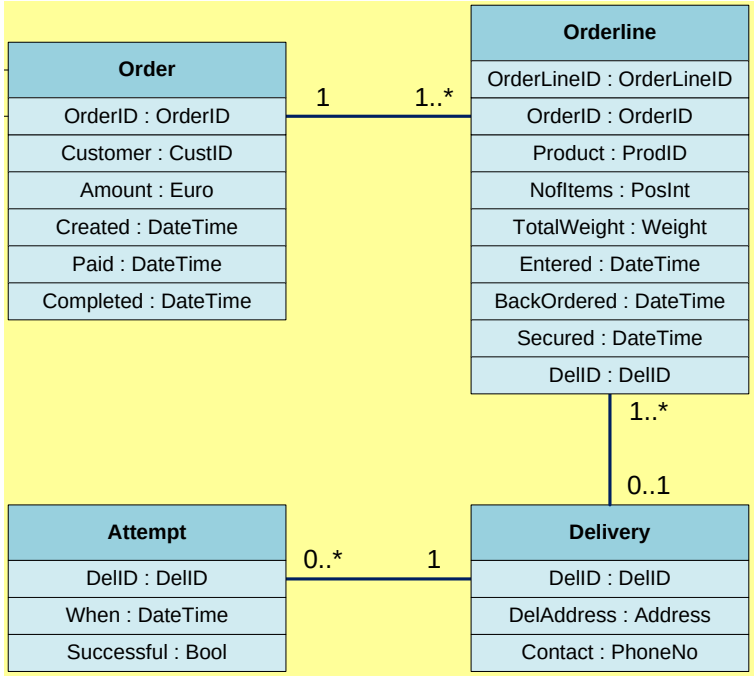
USE
LESS
MOSCOW

problem #2

Straightjacketing smaller interacting processes into one monolithic model


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

What is the process instance?



Order

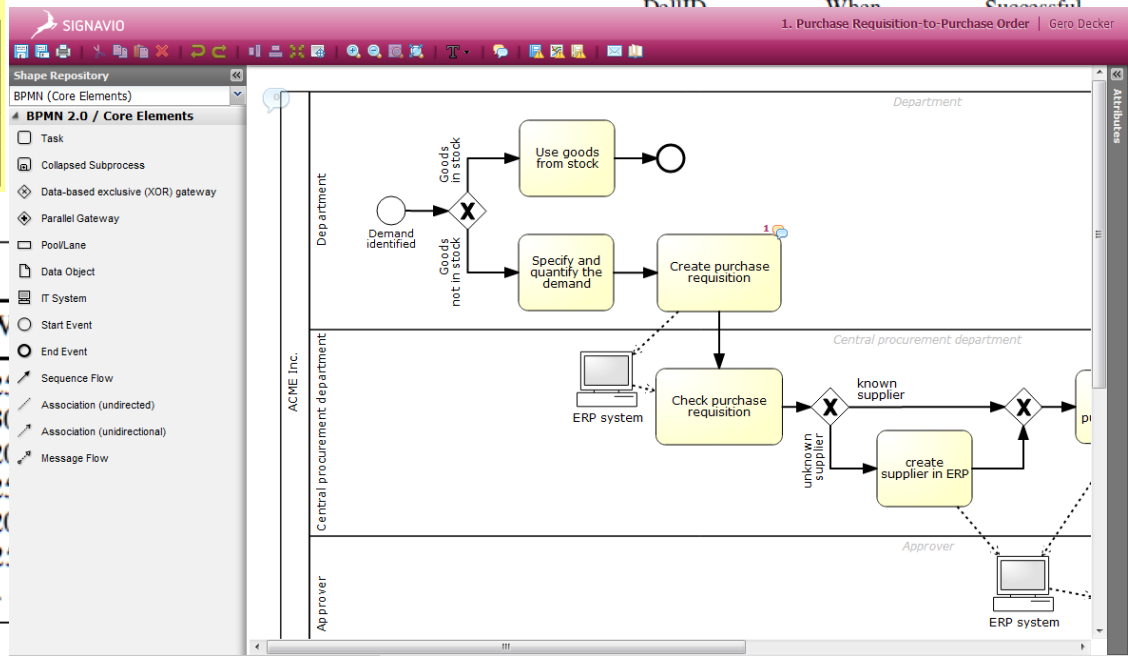
OrderID	Customer	Amount	Created	Paid	Completed
91245	John	100	28-11-2011:08.12	02-12-2011:13.45	05-12-2011:11.33
91561	Mike	530	28-11-2011:12.22	03-12-2011:14.34	05-12-2011:09.32
91812	Mary	234	29-11-2011:09.45	02-12-2011:09.44	04-12-2011:13.33
92233	Sue	110	29-11-2011:10.12	null	null
92345	Kirsten	195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

Delivery

Attempt

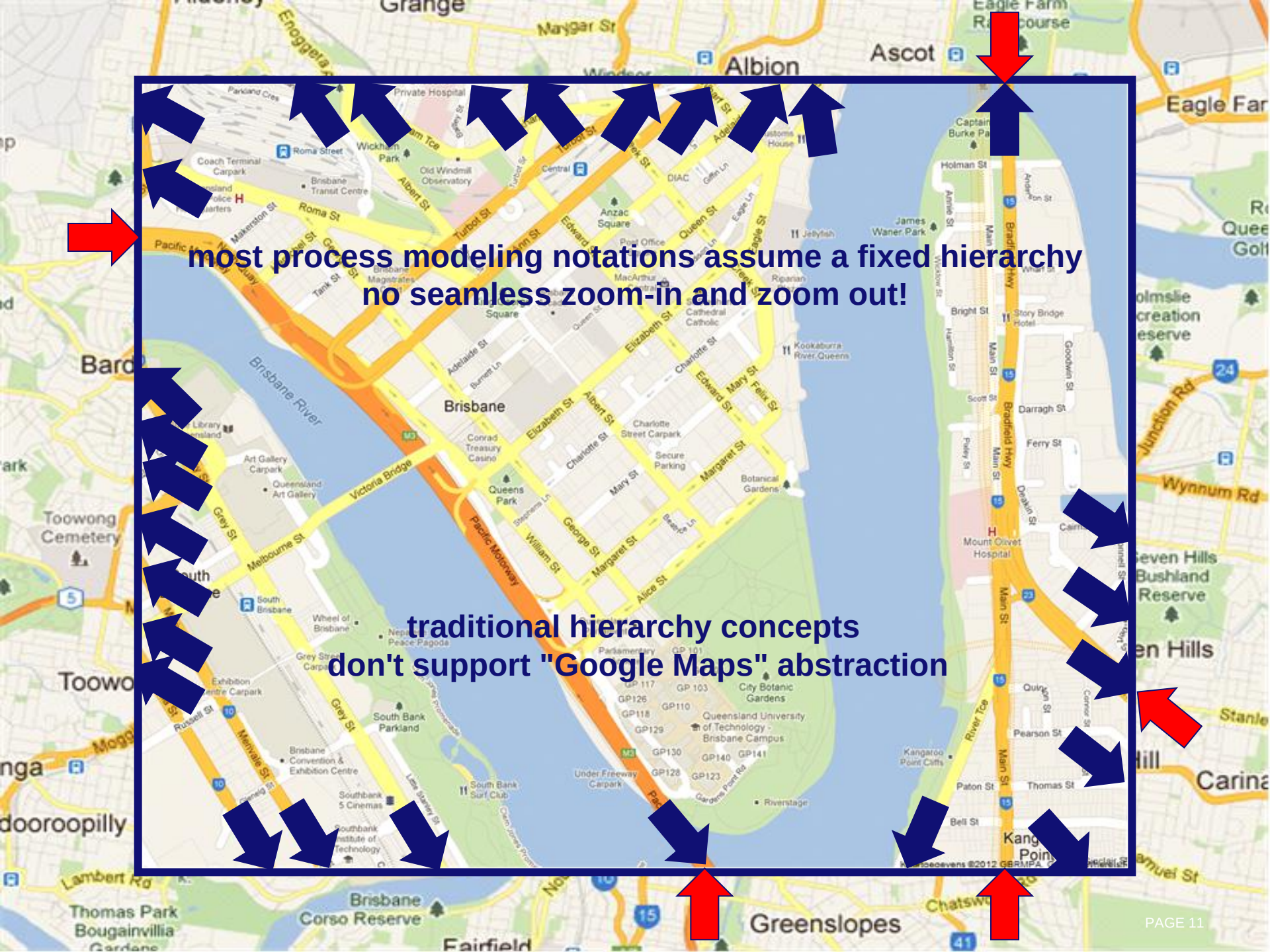
DelID	When	Successful

OrderLineID	OrderID	Product	NofItems	TotalW
112345	91245	iPhone 4G	1	0.2
112346	91245	iPod nano	2	0.3
112347	91245	iPod classic	1	0.2
112448	91561	iPhone 4G	1	0.2
112449	91561	iPod classic	1	0.2
112452	91812	iPhone 4G	5	1.2
...	...	...	...	..



problem #3

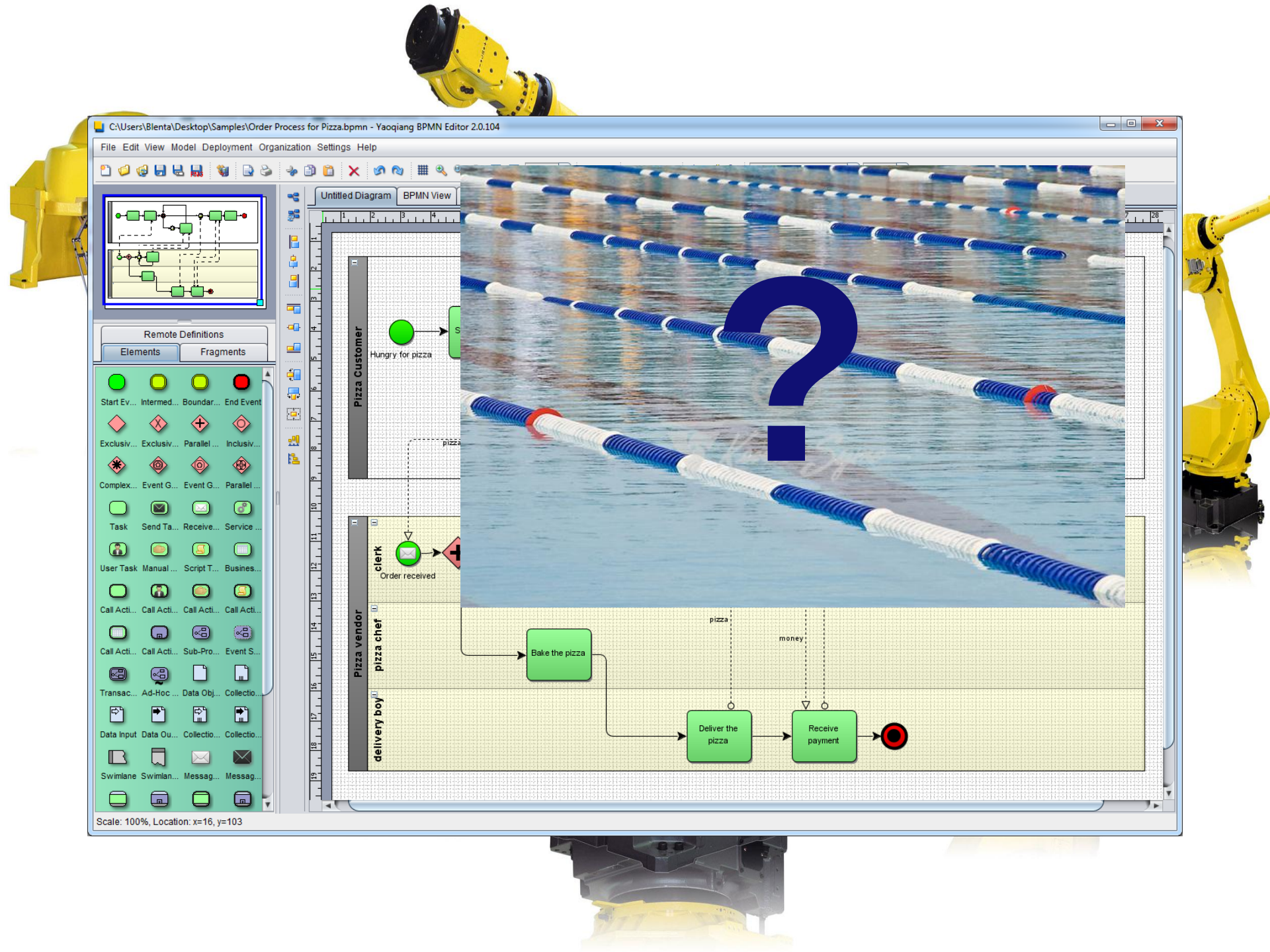
Using a static hierarchical decomposition as the only abstraction mechanism


most process modeling notations assume a fixed hierarchy
no seamless zoom-in and zoom out!
traditional hierarchy concepts
don't support "Google Maps" abstraction
Brisbane
Brisbane River
Albion
Ascot
Greenslopes
Fairfield
Toowong
Bardon
Indooroopilly
Carina
Eagle Farm
Kangaroo Point

problem #4

Modeling humans as if they are machines doing a single task


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

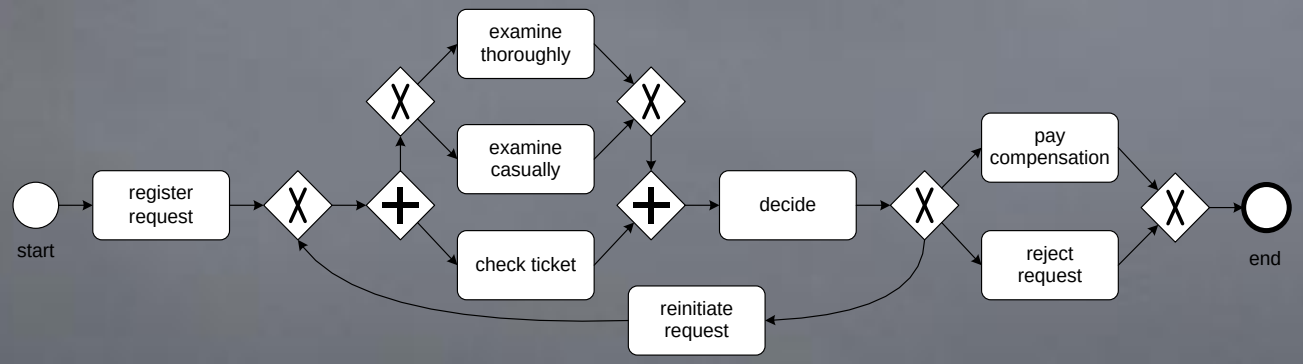

C:\Users\Blenta\Desktop\Samples\Order Process for Pizza.bpmn - Yaoqiang BPMN Editor 2.0.104
File Edit View Model Deployment Organization Settings Help
Untitled Diagram
BPMN View
Remote Definitions
Elements
Fragments
Pizza Customer
Hungry for pizza
Pizza vendor
clerk
Order received
pizza chef
Bake the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

problem #5

Being vague about vagueness

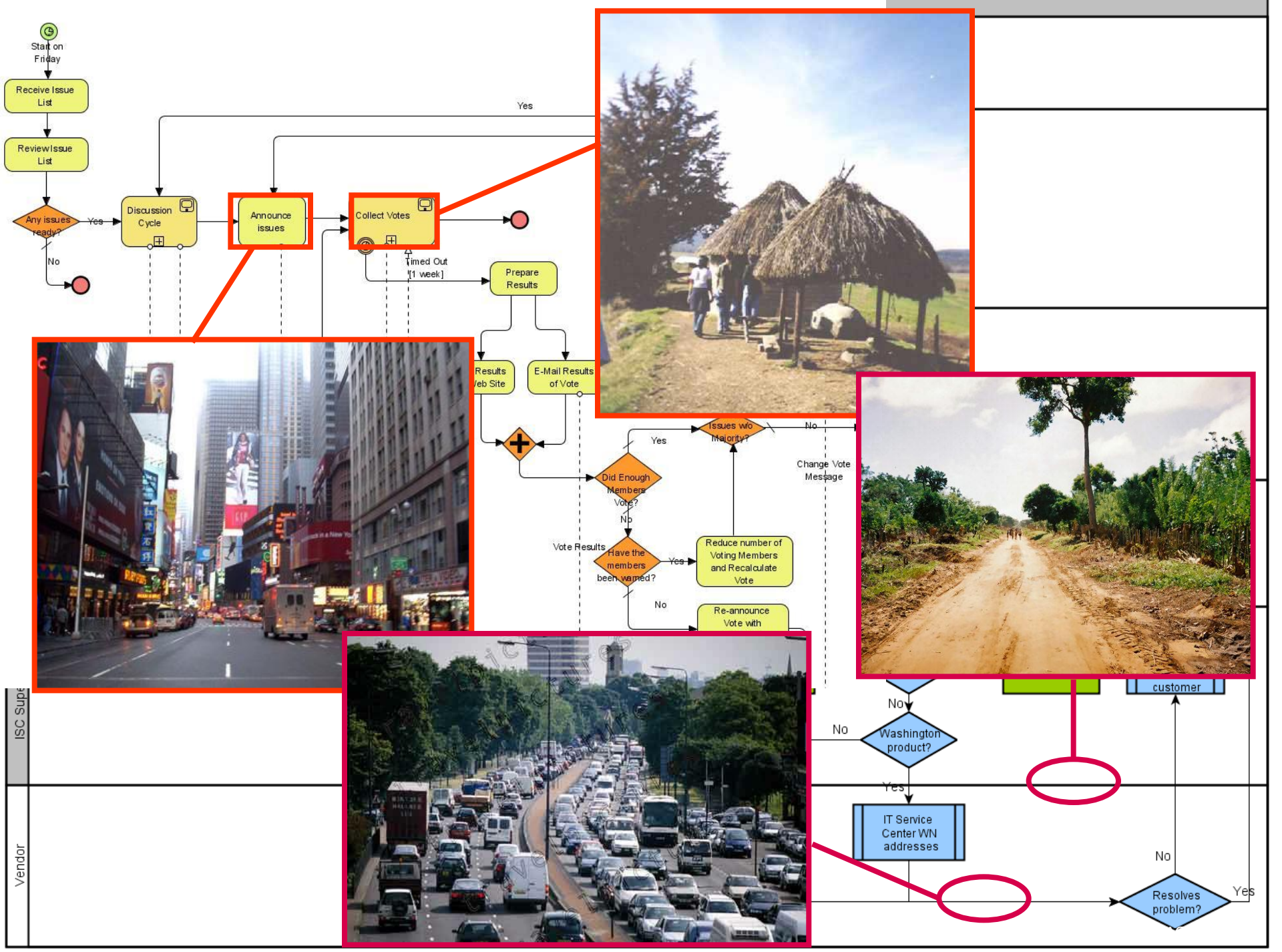
Make vagueness explicit!

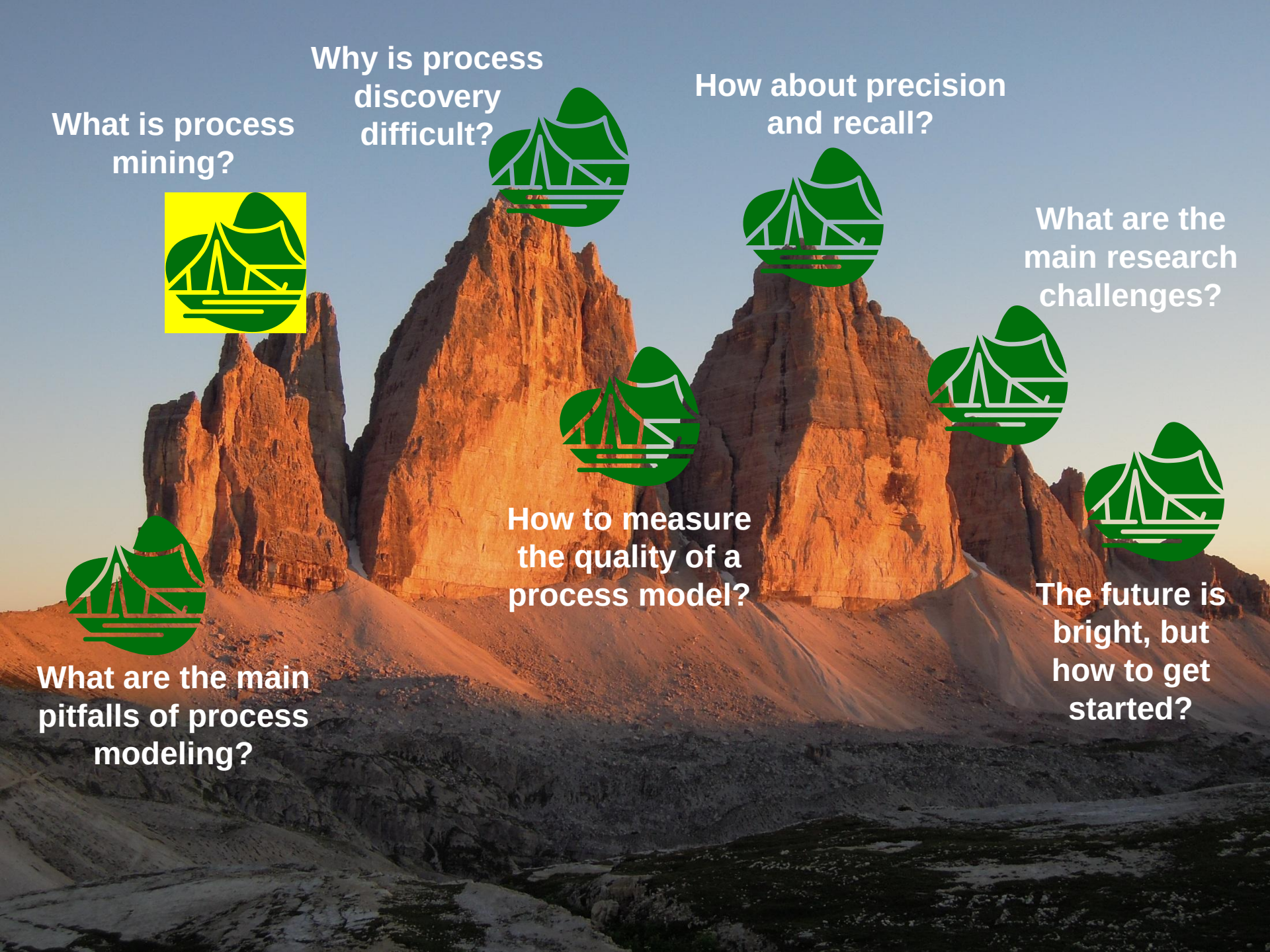
(rather than having multiple notations for the same thing)



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
Yes
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
Issues w/o Majority?
No
Change Vote Message
customer
No
Washington product?
Yes
IT Service Center WN addresses
No
Resolves problem?
Yes
ISC Supp
Vendor


What is process mining?
Why is process discovery difficult?
How about precision and recall?
What are the main research challenges?
How to measure the quality of a process model?
What are the main pitfalls of process modeling?
The future is bright, but how to get started?

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

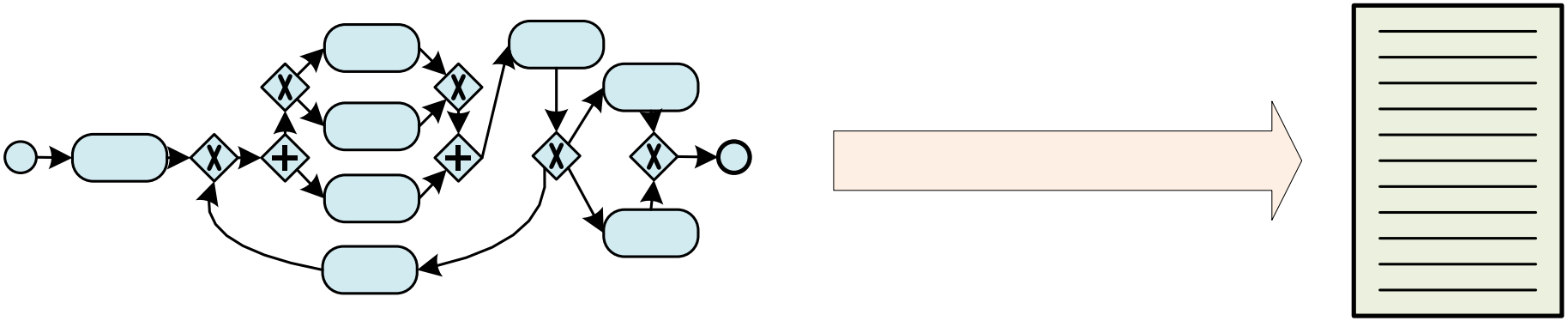
(data mining, machine learning, business intelligence)


www.olifantenpaadjes.nl
60
0100110010101010
0100110101010


0100110101010101010

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

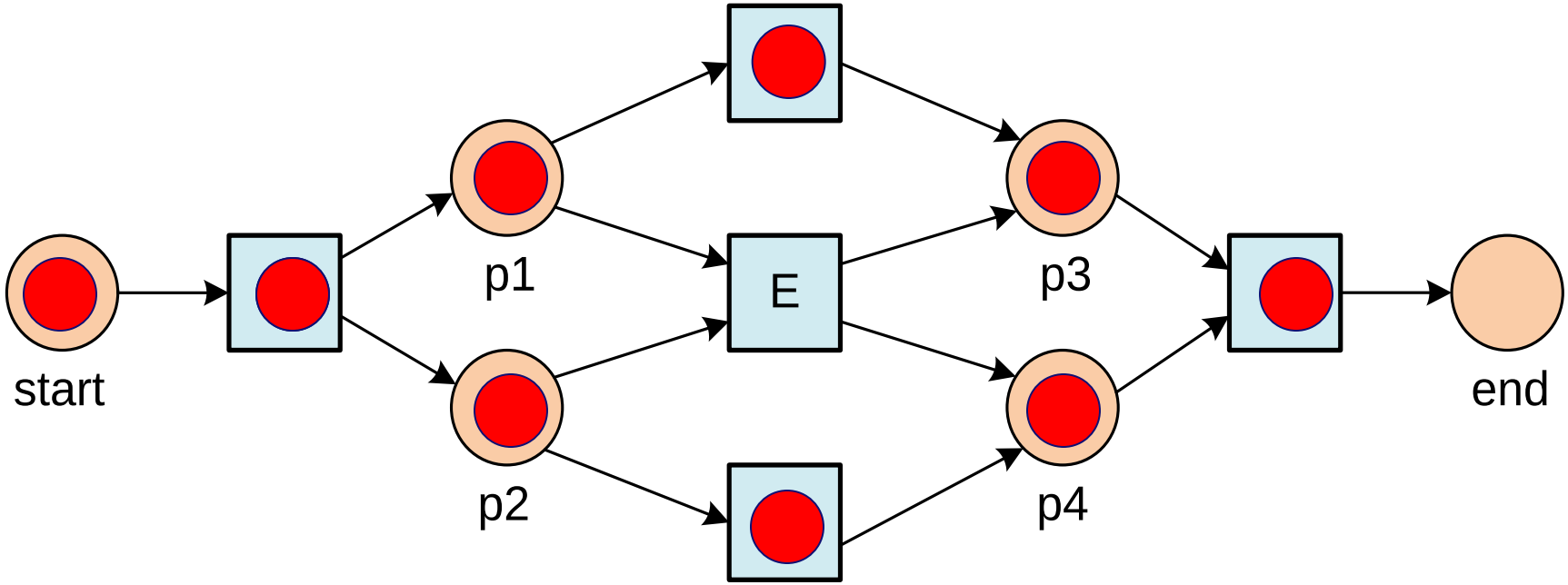
Play-Out



process model

event log

Play-Out (Classical use of models)



A B C D

A E D

A E D

A B C D

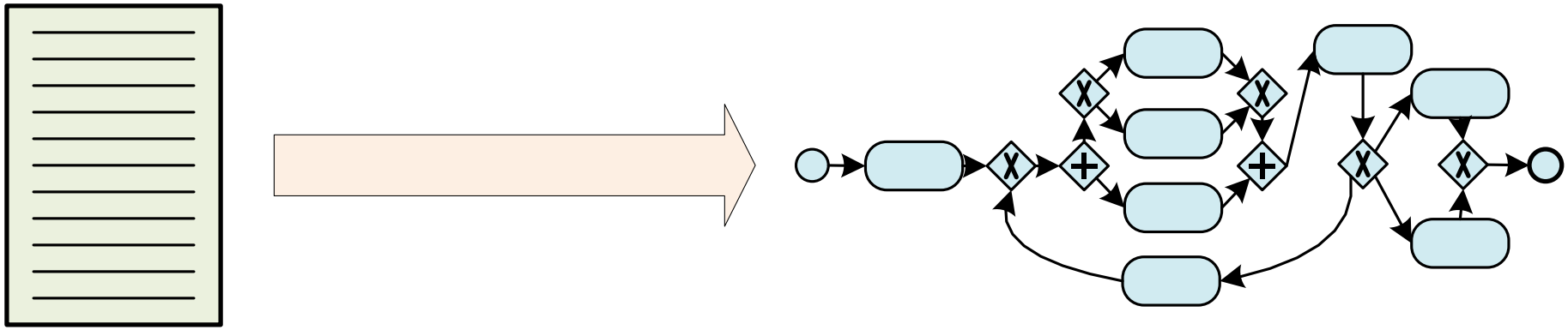
A C B D

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

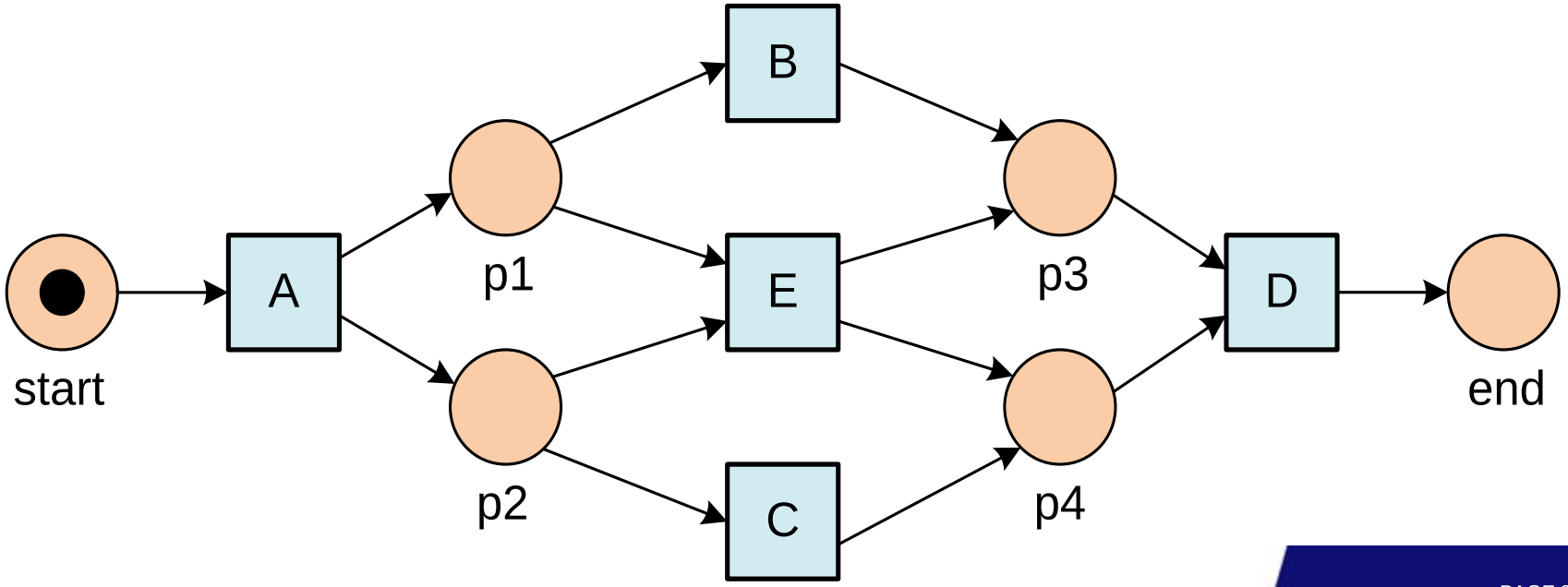
A B C D

A C B D

A C B D

A E D

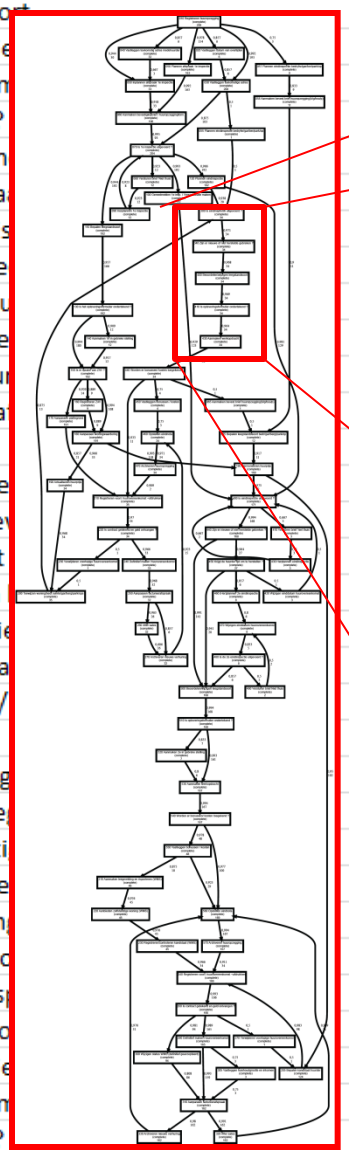
A C B D

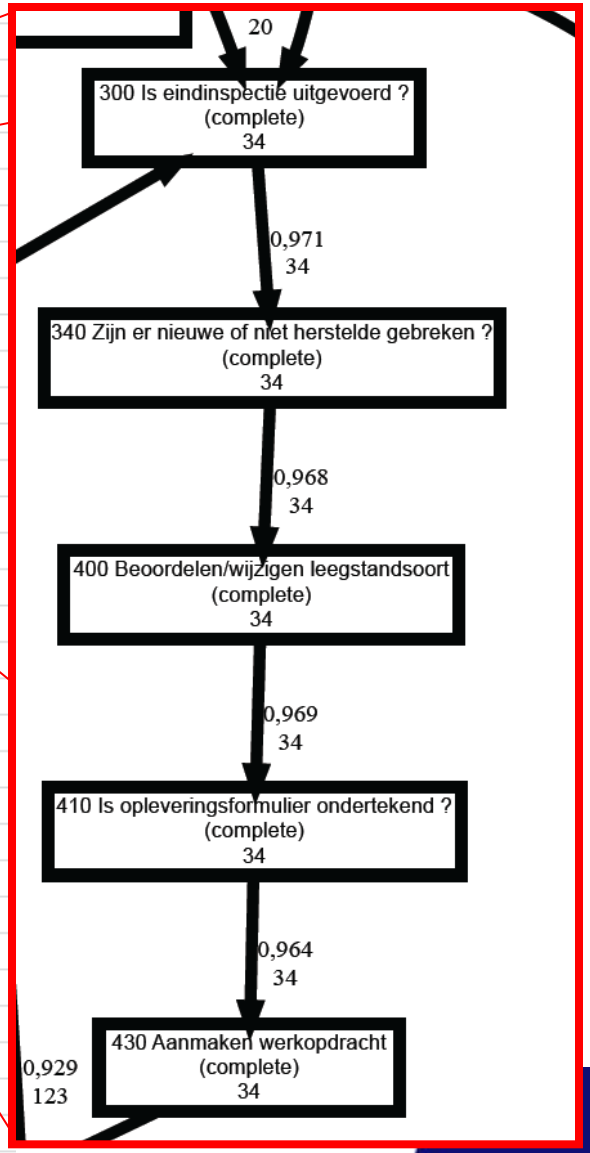


Example Process Discovery

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

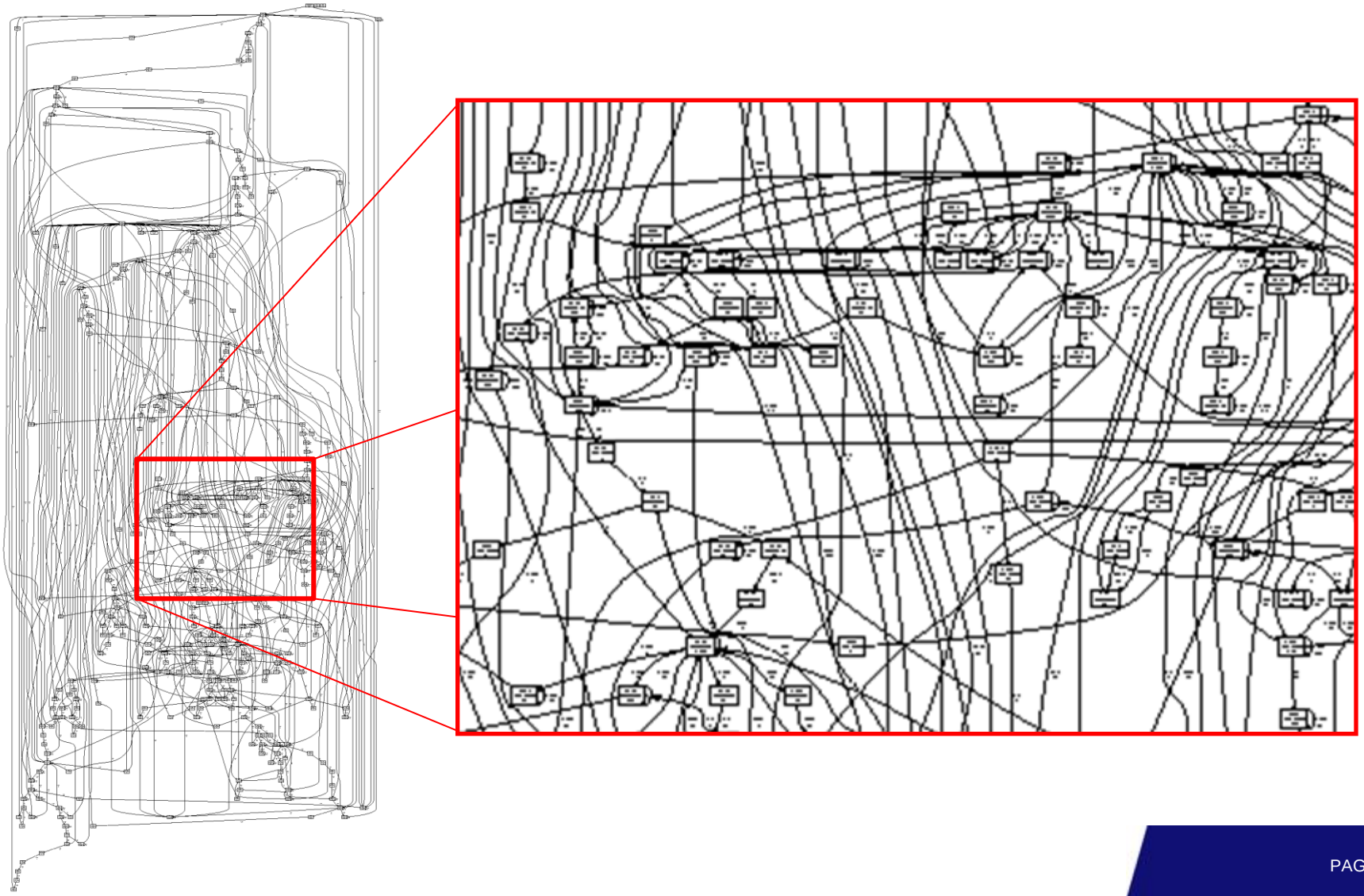
117315	110 Bepalen leegstandsoort	16.05.2007 14:06:23
117315	120 Plannen eindinspectie	16.05.2007 14:36:01
117315	130 Is het opleveringsform	23.05.2007 09:41:40
117315	150 Is er sprake van ZAV ?	23.05.2007 09:41:51
117315	170 Aanpassen plattegron	23.05.2007 11:57:18
117315	180 Aanpassen woningwa	23.05.2007 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	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





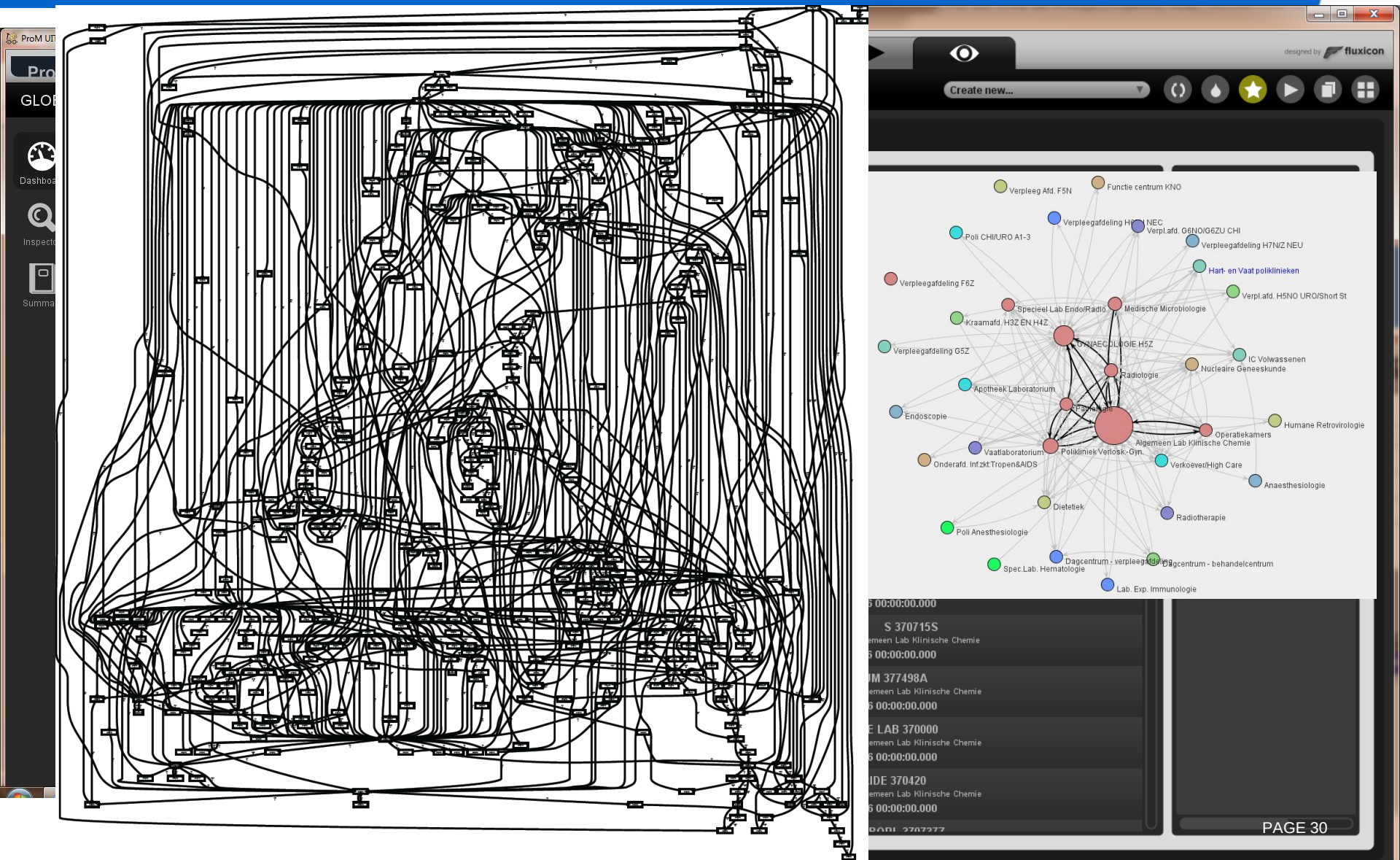
Example Process Discovery

(ASML, test process lithography systems, 154966 events)

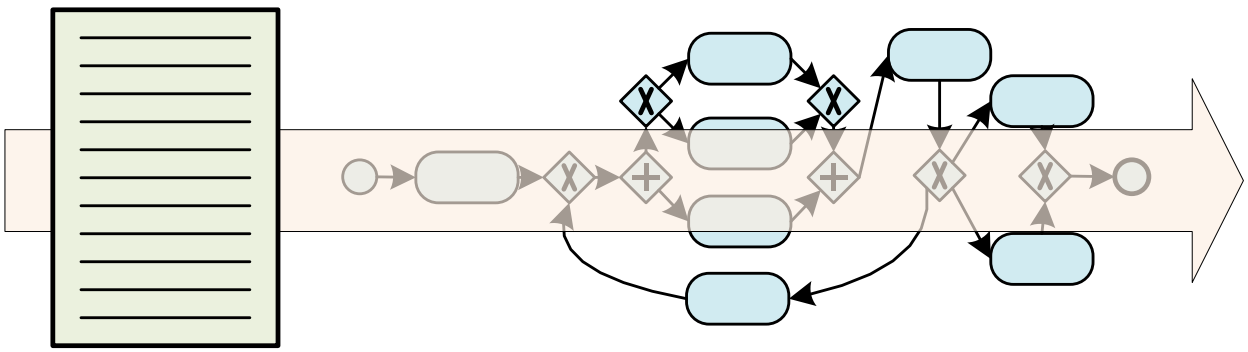


Example Process Discovery

(AMC, 627 gynecological oncology patients, 24331 events)



Replay



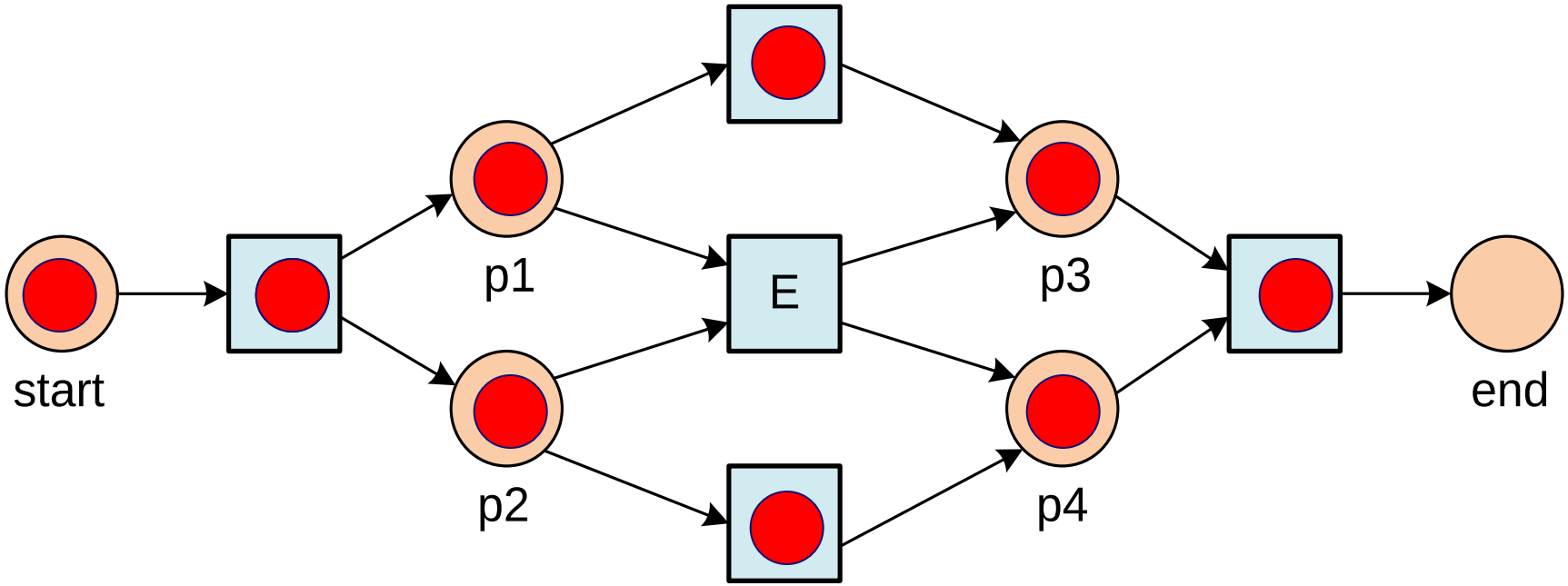
event log

process model

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

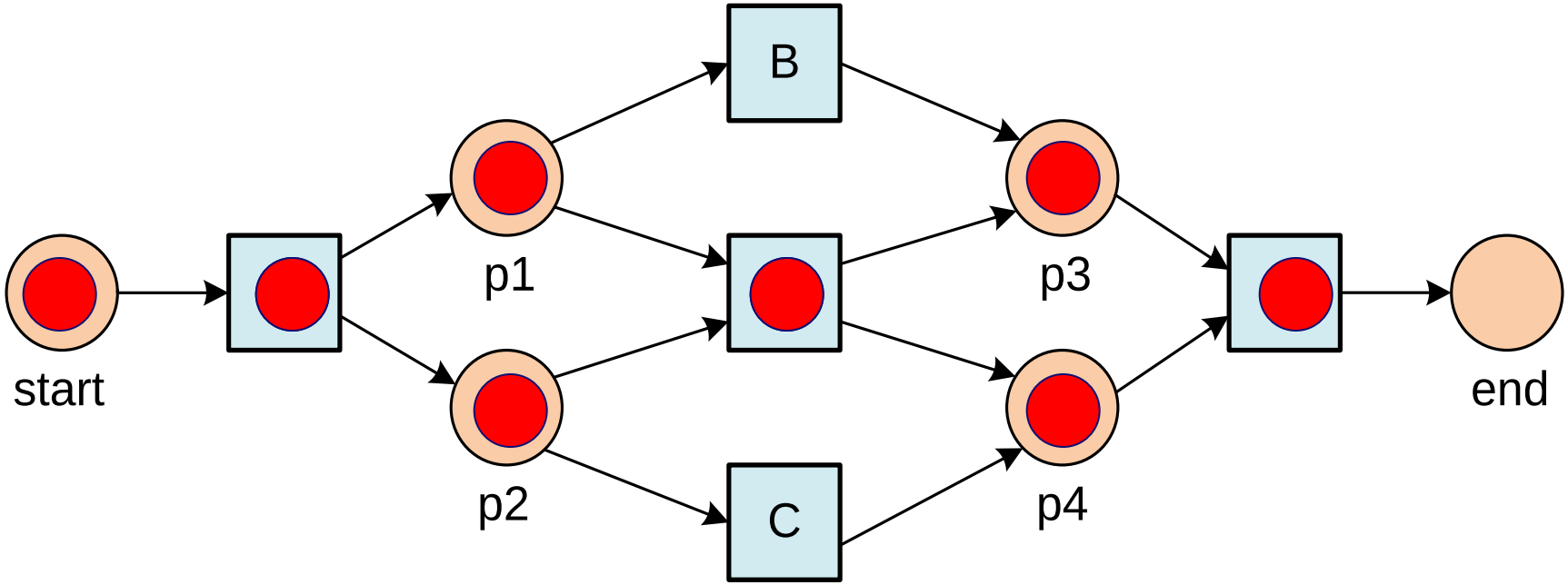
Replay

A B C D



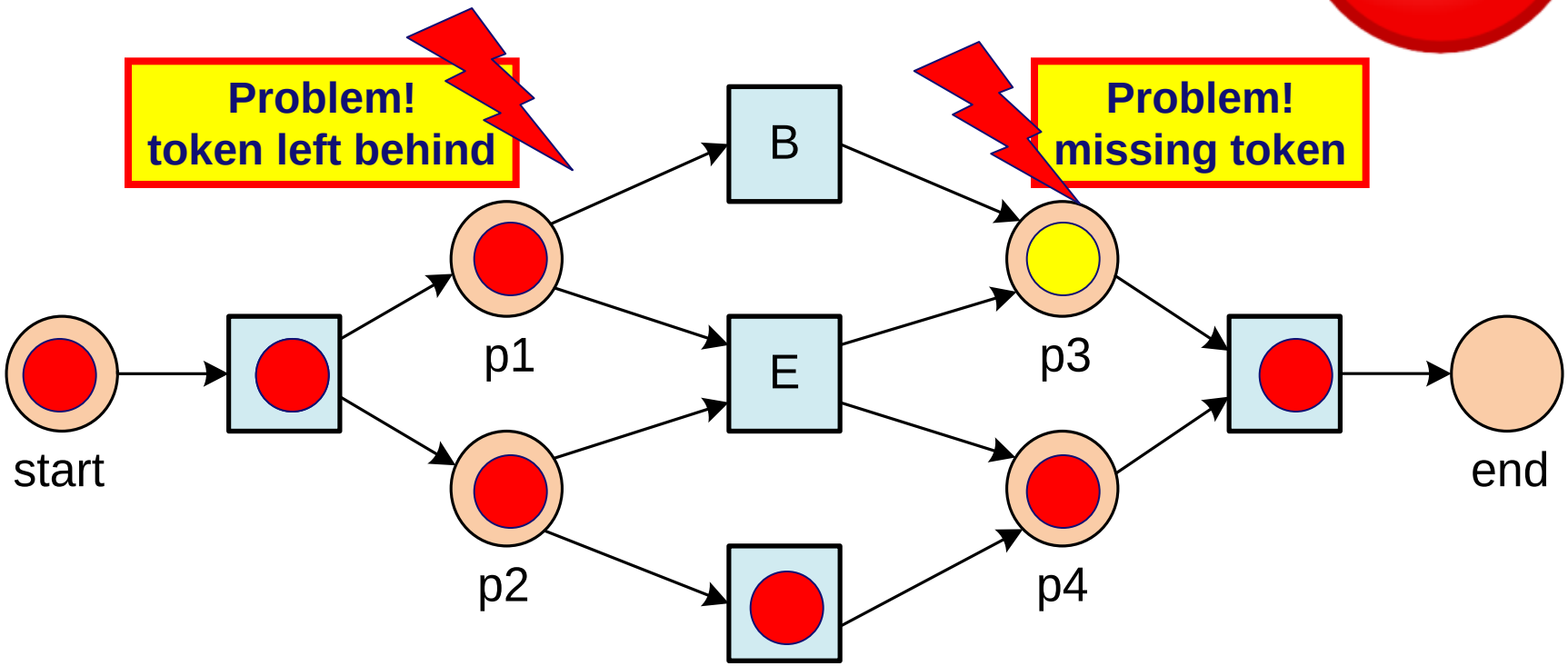
Replay

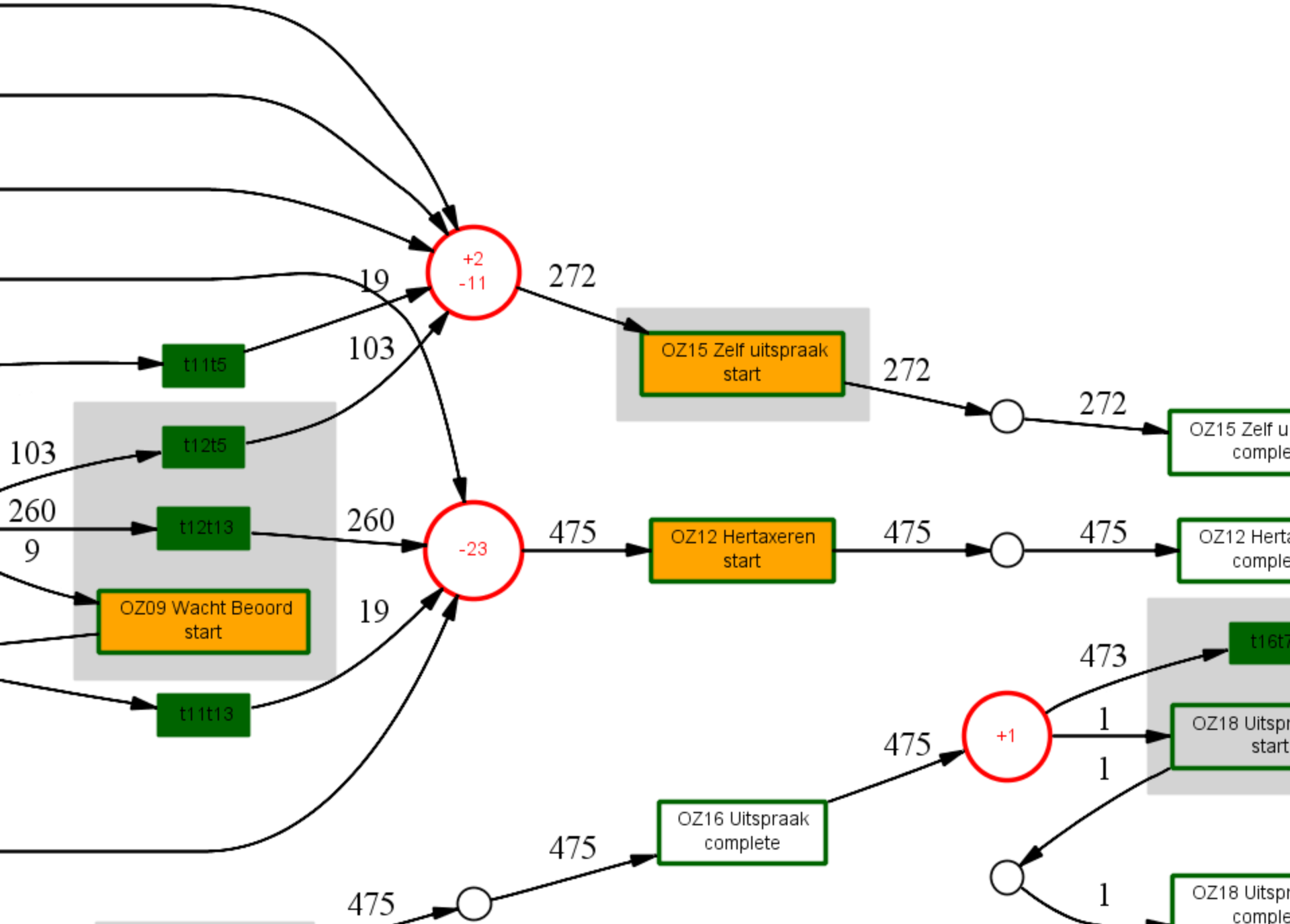
A E D



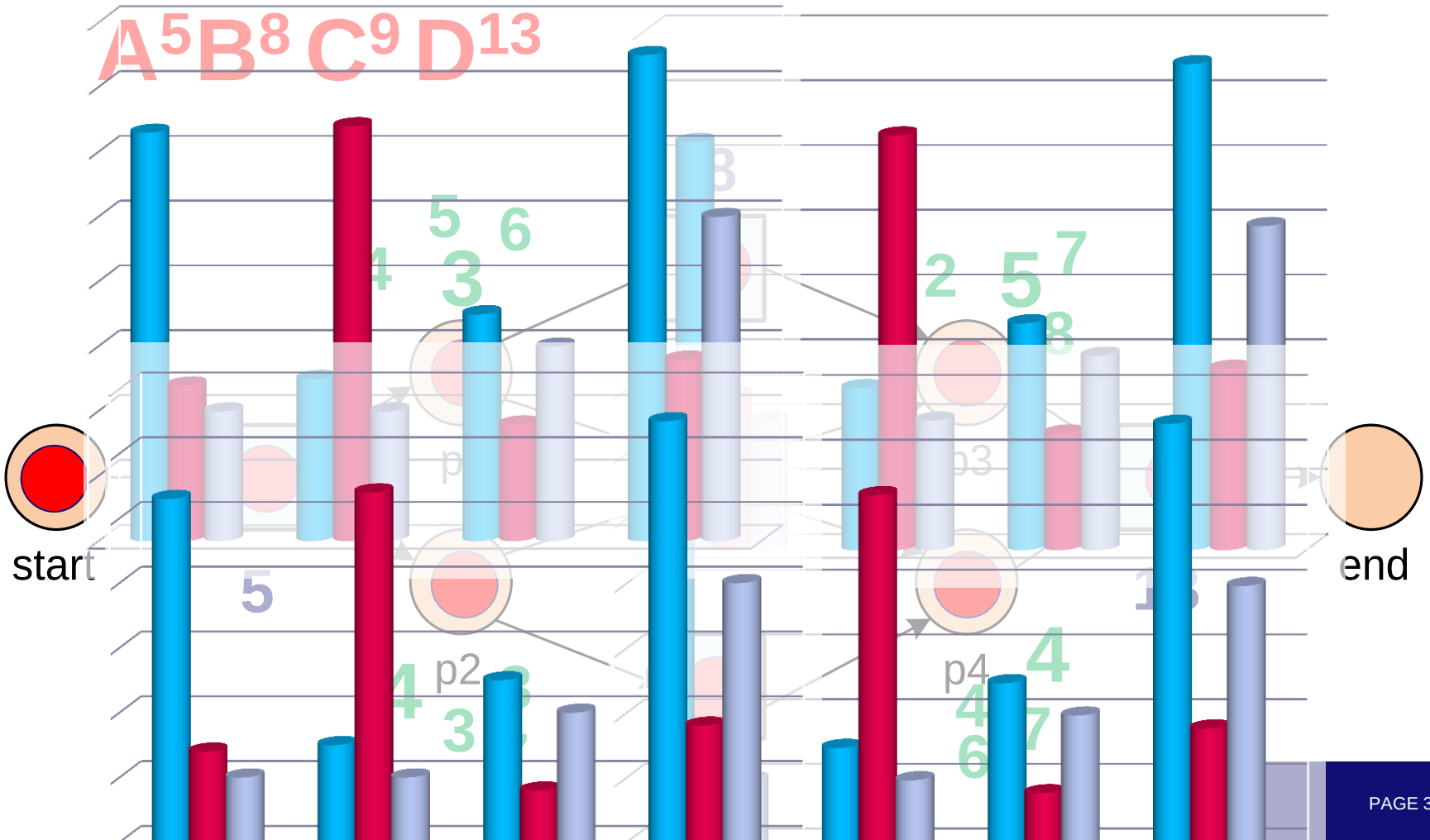
Replay can detect problems

A C D



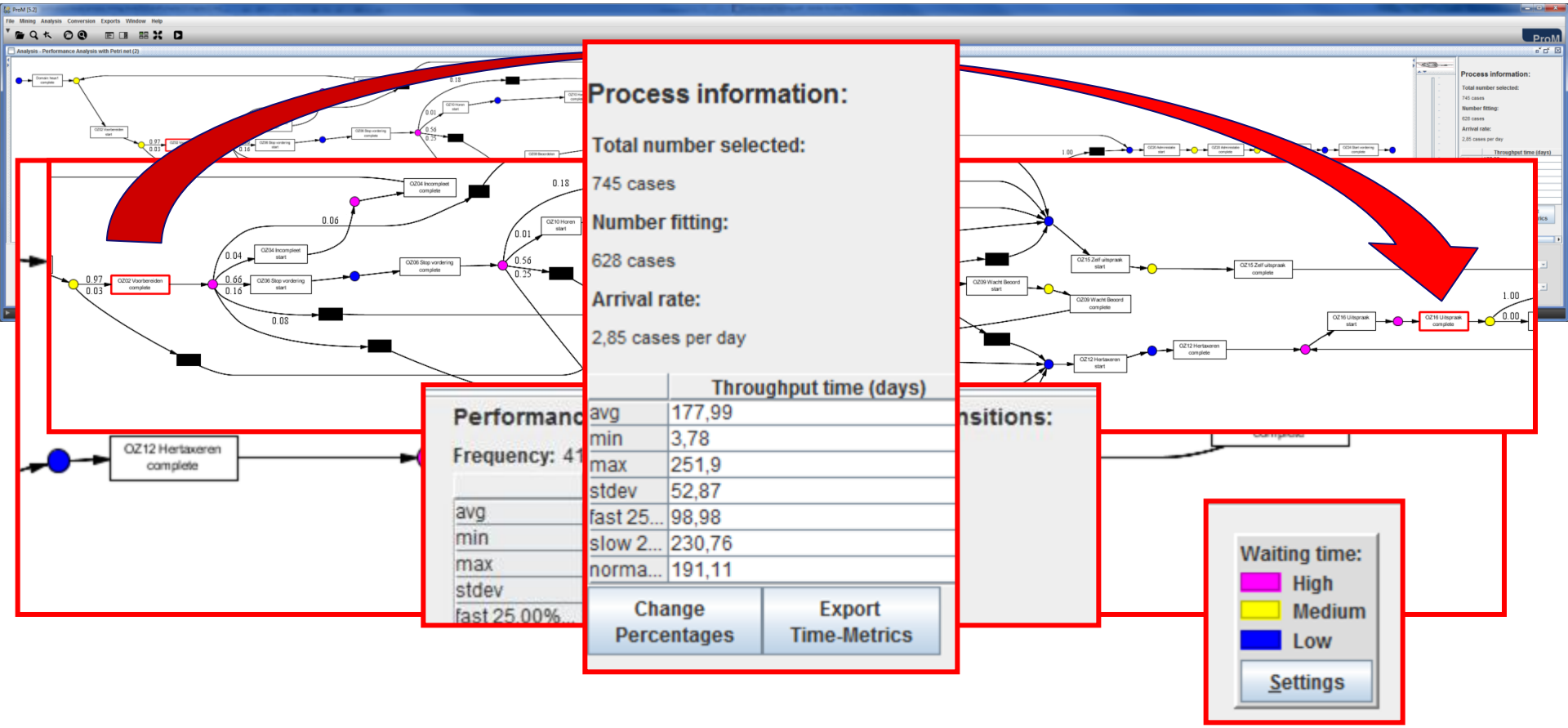

+2
-11
19
272
t11t5
103
OZ15 Zelf uitspraak start
272
272
OZ15 Zelf u
comple
103
t12t5
260
t12t13
260
9
-23
475
OZ12 Hertaxeren start
475
475
OZ12 Herta
comple
OZ09 Wacht Beoord start
19
t11t13
473
t16t7
1
OZ18 Uitspr
start
475
+1
1
OZ16 Uitspraak complete
475
475
1
OZ18 Uitspr
comple

Replay can extract timing information



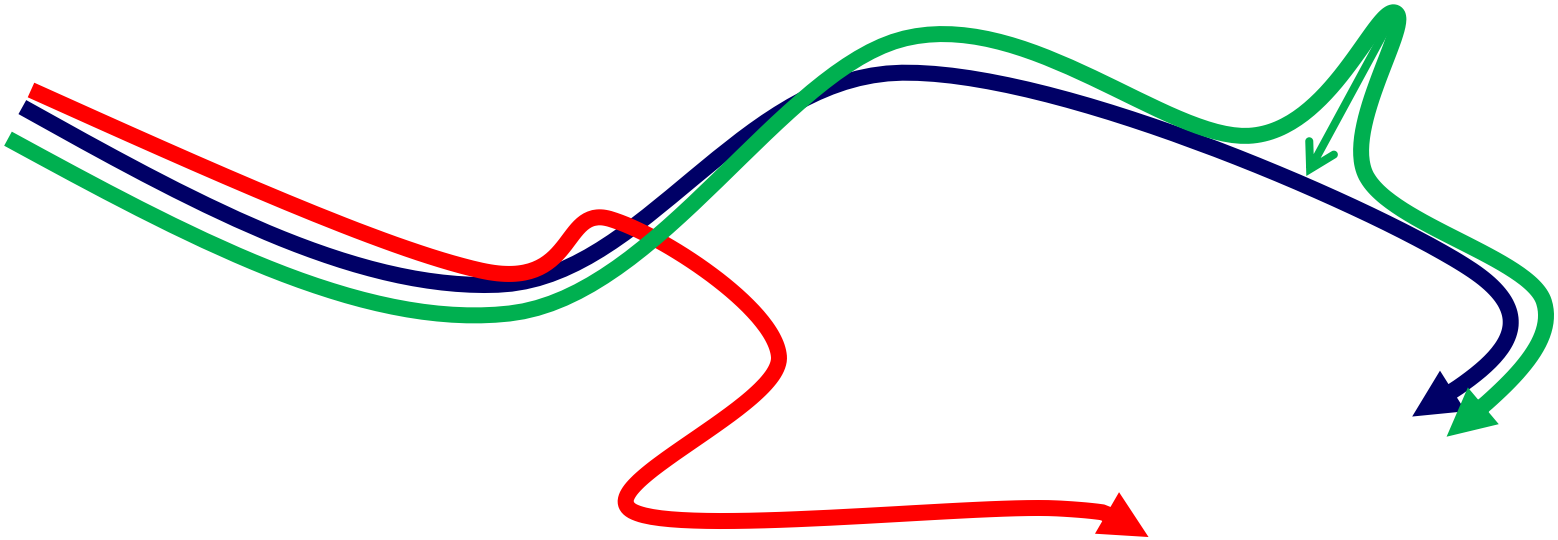
Performance Analysis Using Replay

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



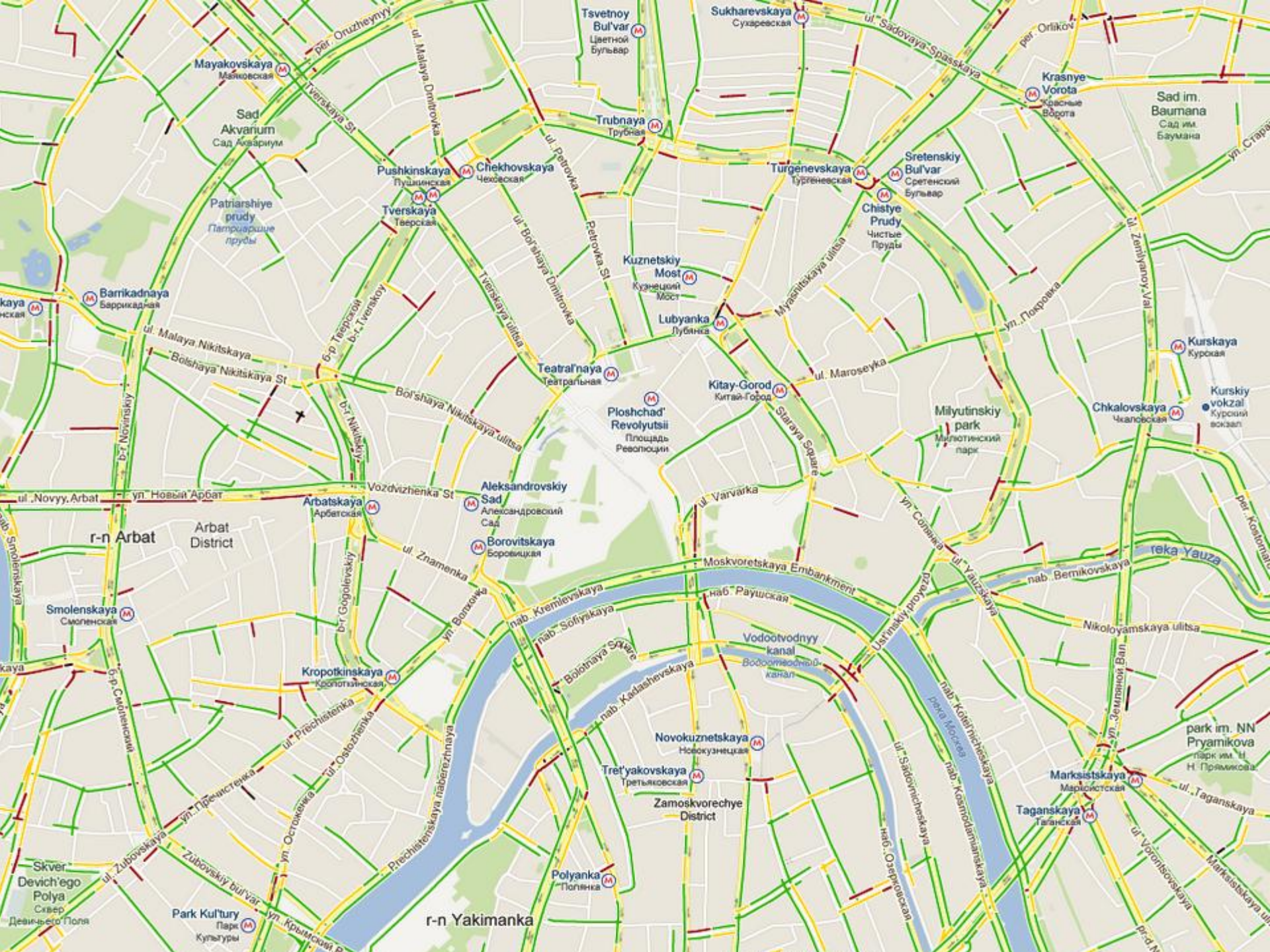
Models are like the glasses required to see and understand event data!

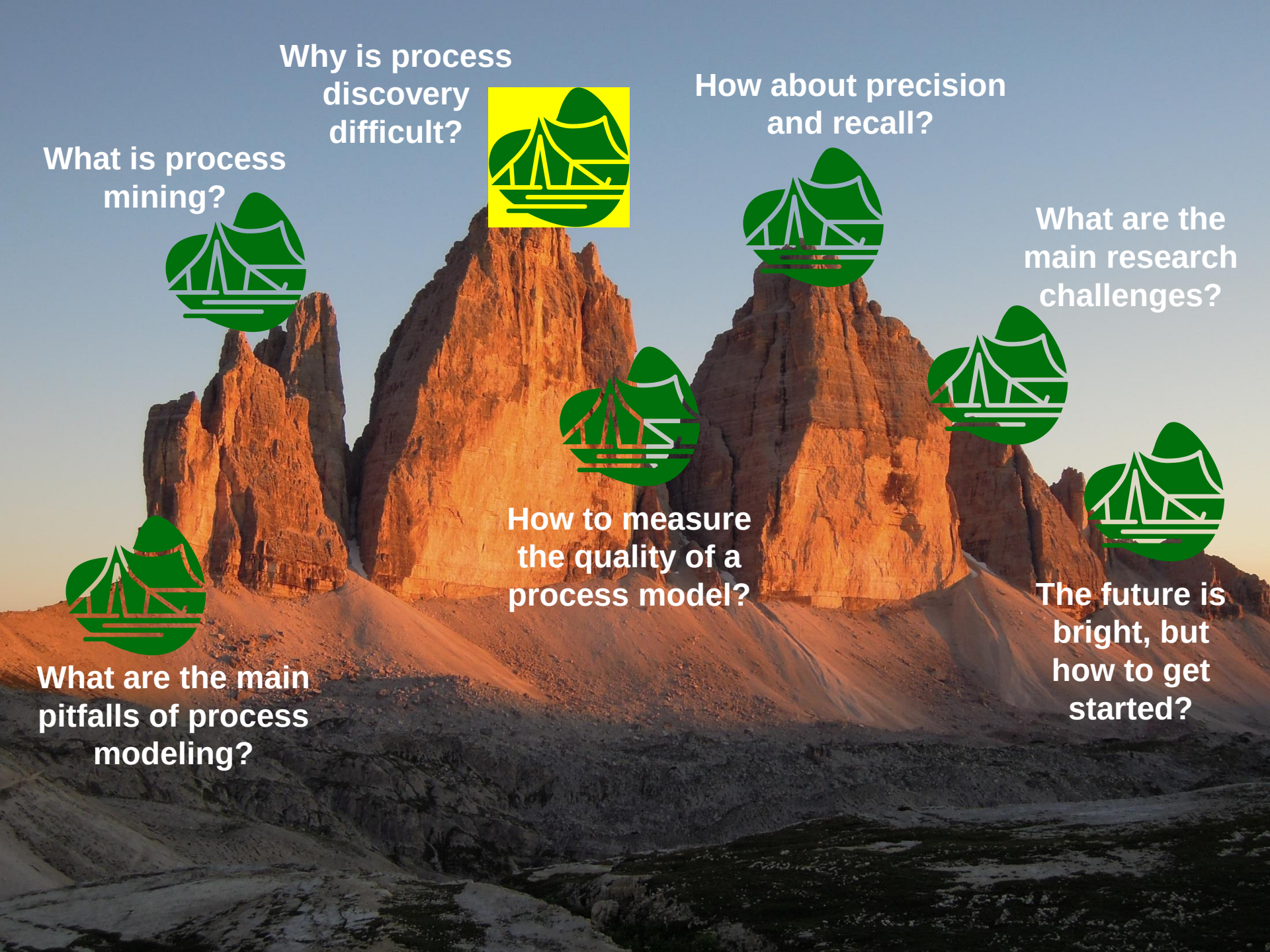
Alignments are essential!



- **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
$t1$	$t4$	$t3$	$t5$	$t7$		$t10$


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


Why is process discovery difficult?
How about precision and recall?
What is process mining?
What are the main research challenges?
How to measure the quality of a process model?
The future is bright, but how to get started?
What are the main pitfalls of process modeling?

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.

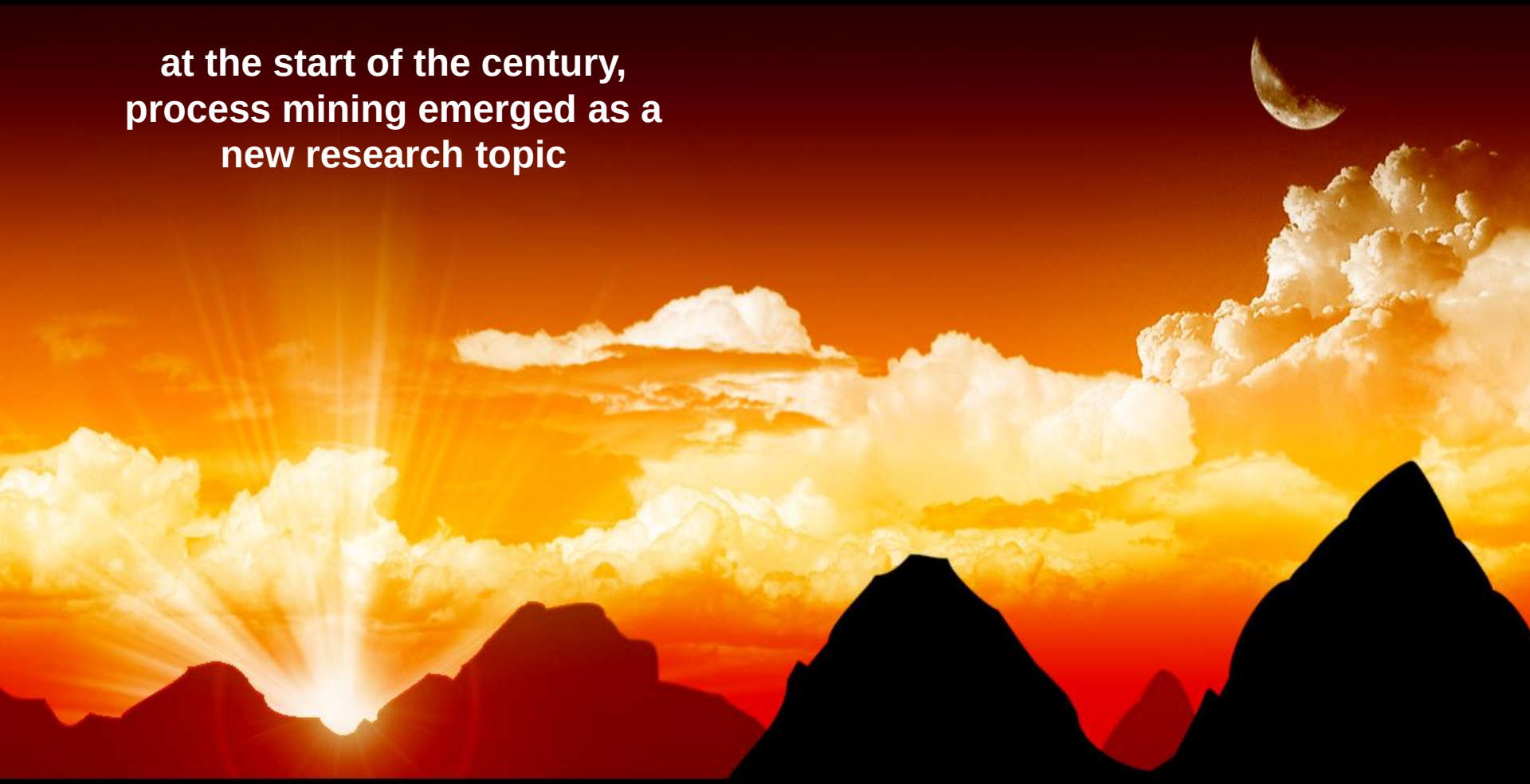
Learning is not easy …

- Even simple languag
  regular languag
  learnable i
- M
  with or
  ive examples,
  cles, 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

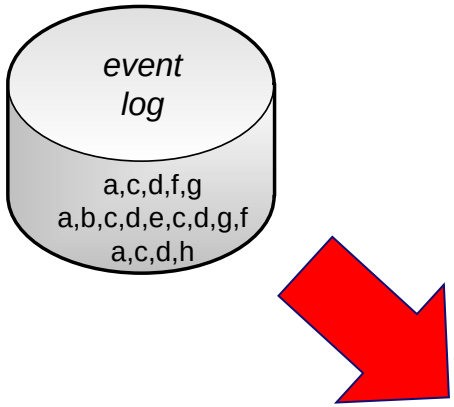


at the start of the century, process mining emerged as a new research topic

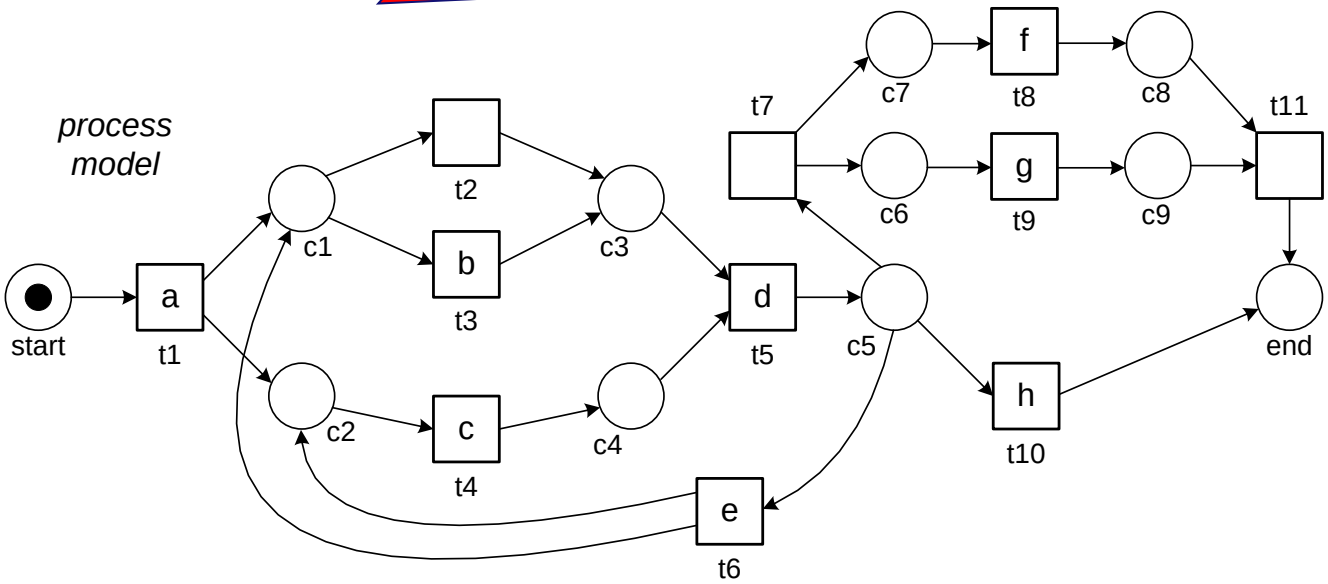
remarkable progress over a relatively short period

See my keynote at http://fluxicon.com/camp/2013/ earlier this week.

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



a = register request
b = examine file
c = check ticket
d = decide
e = reinitiate request
f = send acceptance letter
g = pay compensation
h = send rejection letter



Process discovery algorithms
(small selection)

automata-based learning

distributed genetic mining

language-based regions

heuristic mining

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

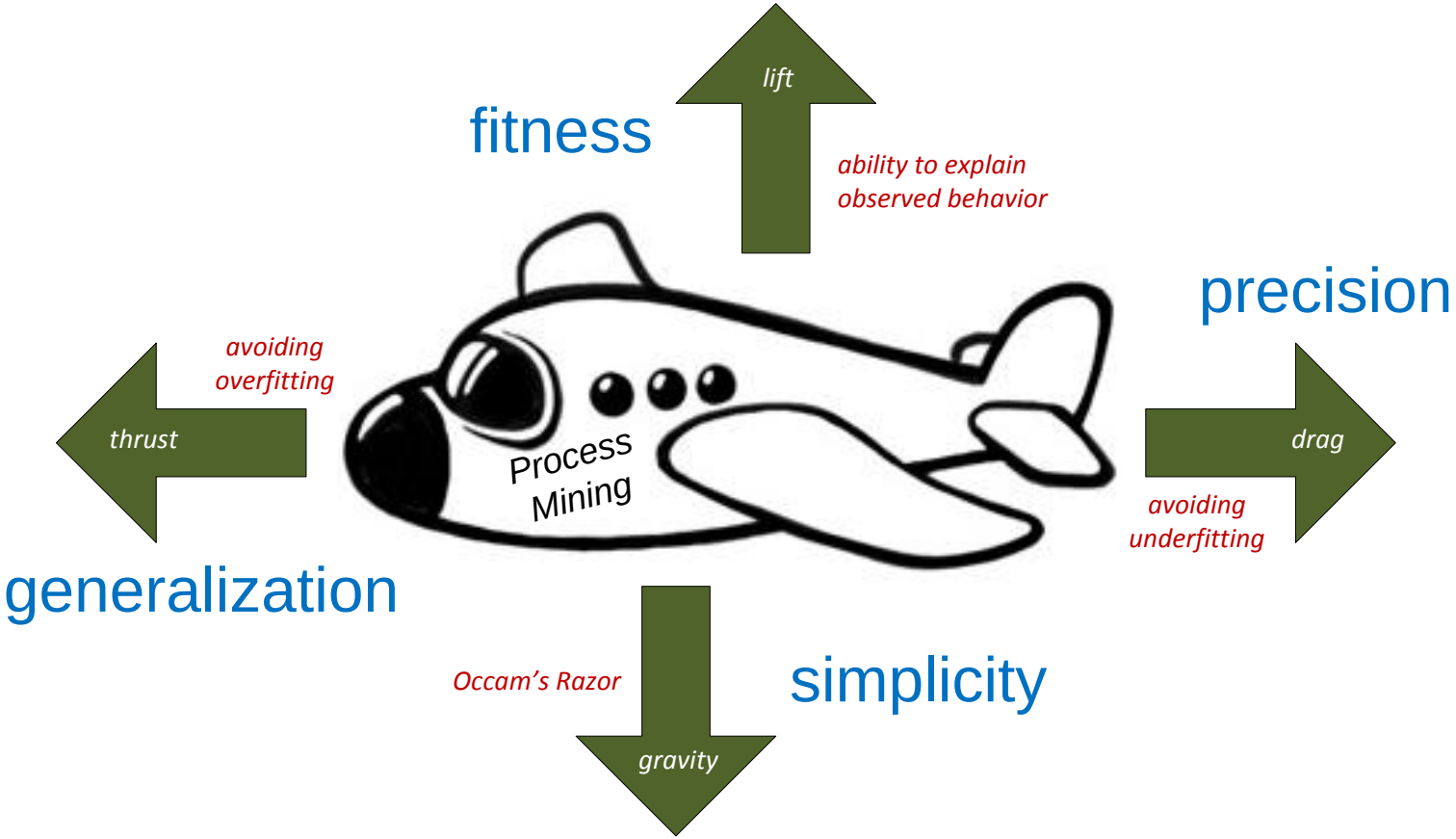
partial-order based mining

α# algorithm

ILP mining

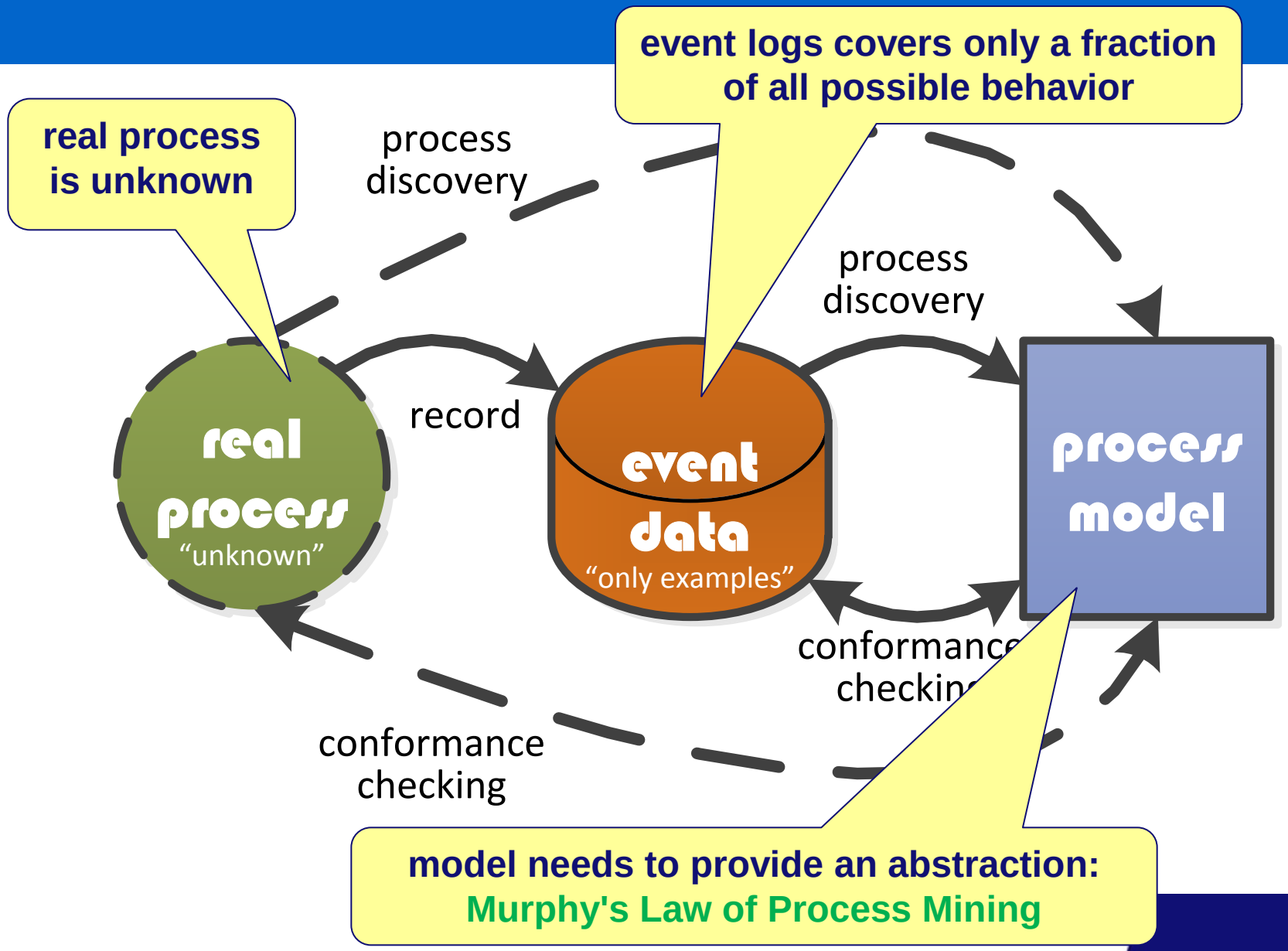
α++ algorithm

How good is my model: Four forces

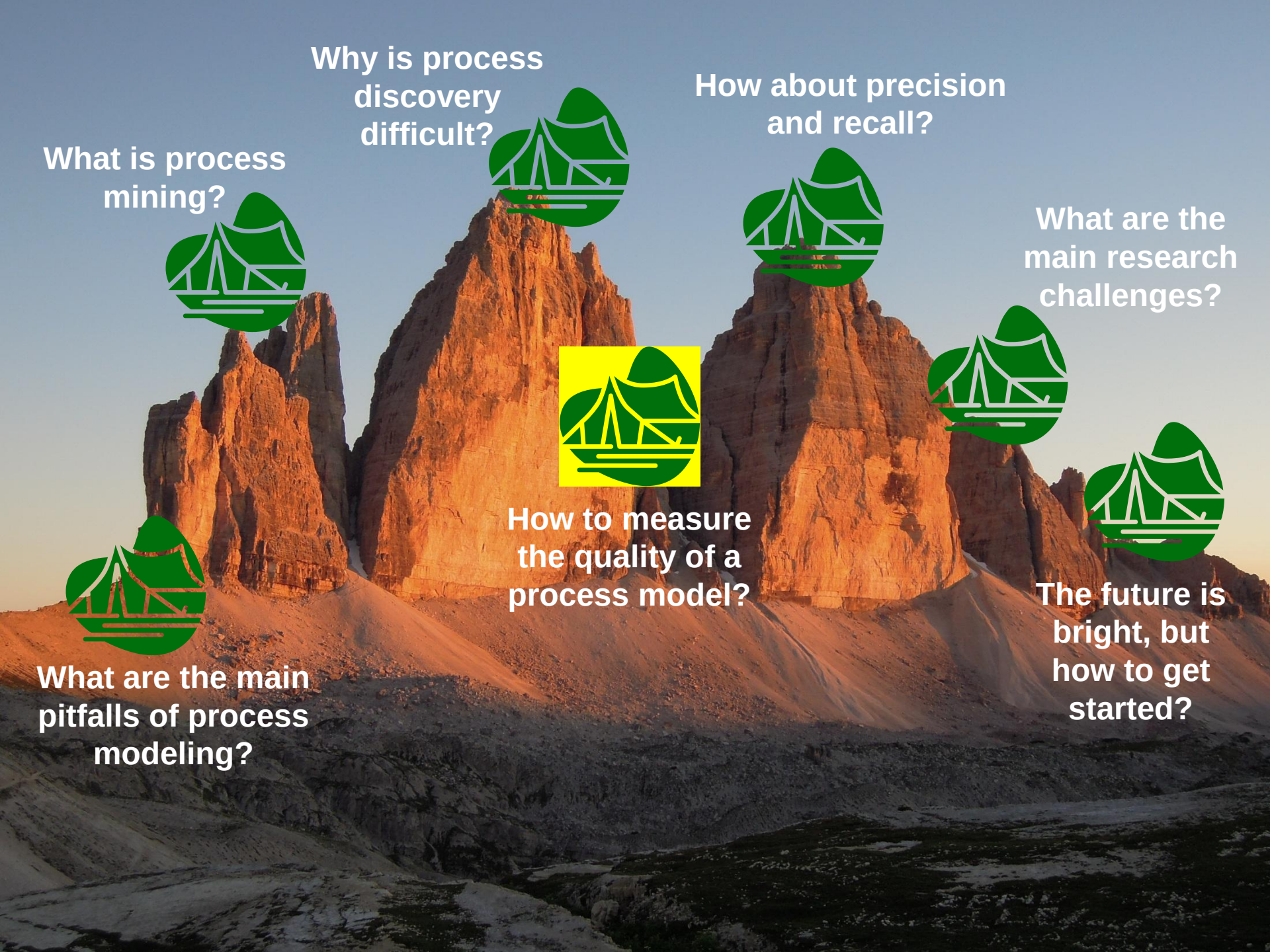


Leaving out one of these dimensions during discovery will lead to degenerate cases!

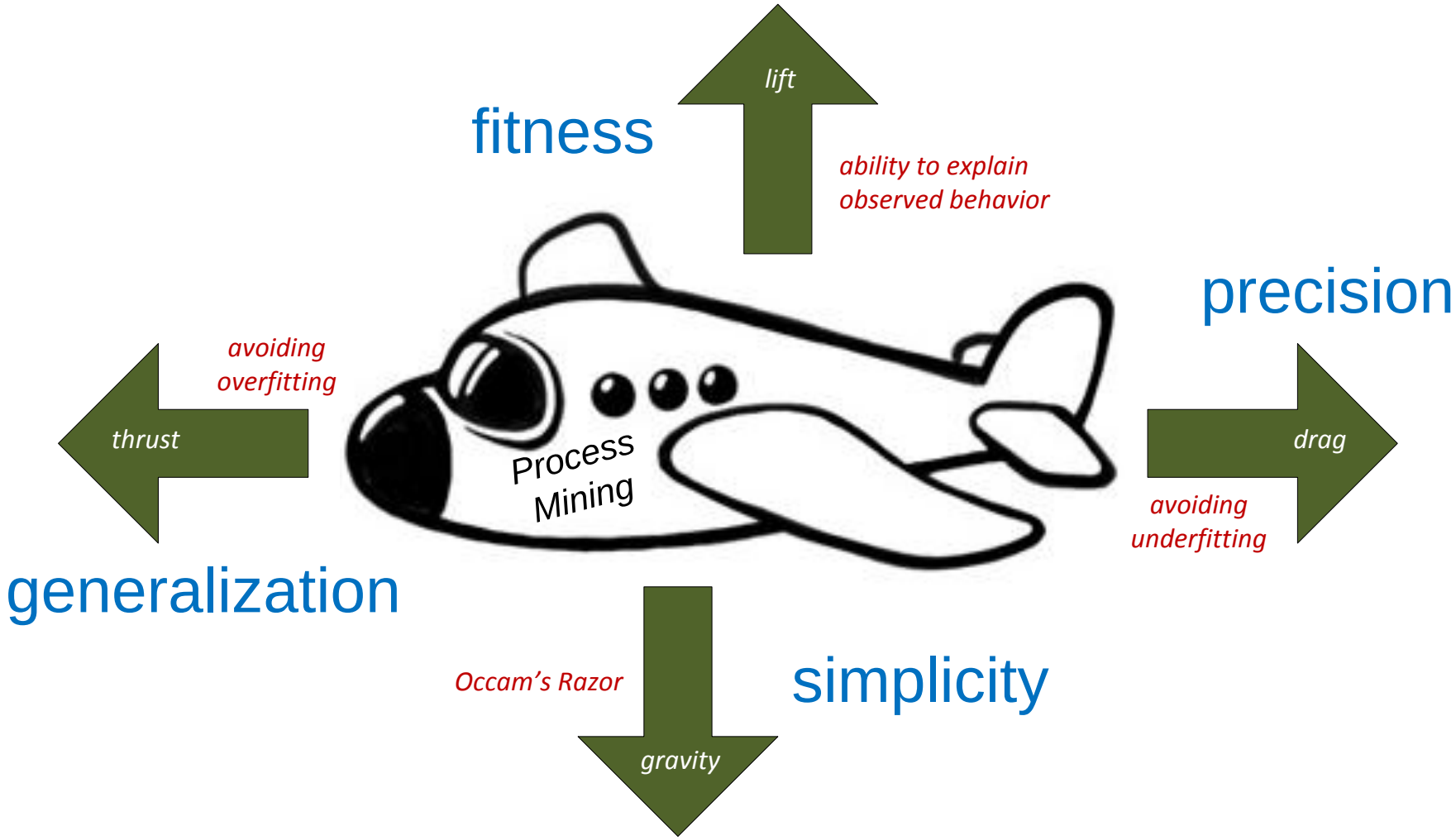
Problem


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


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)

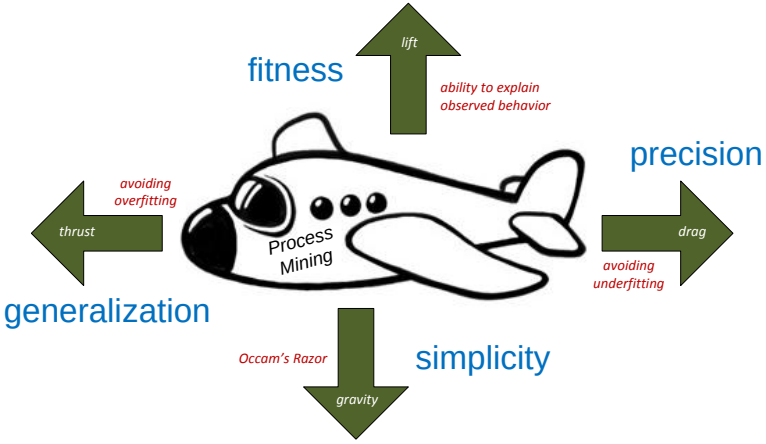

What is process mining?
Why is process discovery difficult?
How about precision and recall?
What are the main research challenges?
How to measure the quality of a process model?
What are the main pitfalls of process modeling?
The future is bright, but how to get started?

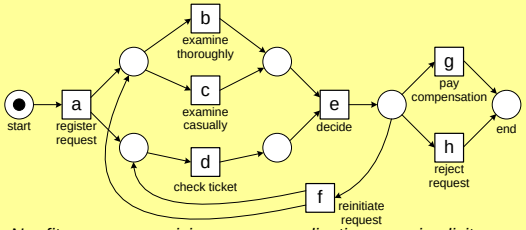
Remember the four forces



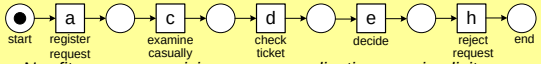
Example: one log four models

#	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	

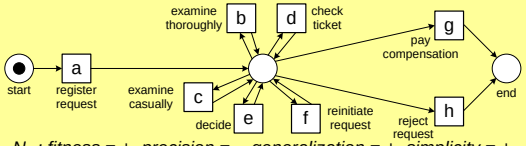




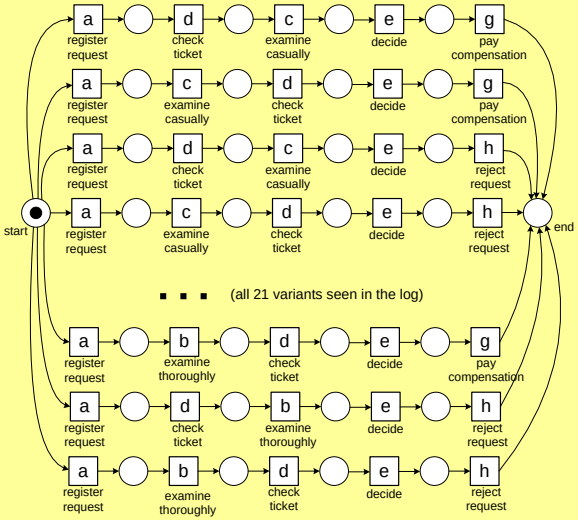
N_1 : fitness = +, precision = +, generalization = +, simplicity = +



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

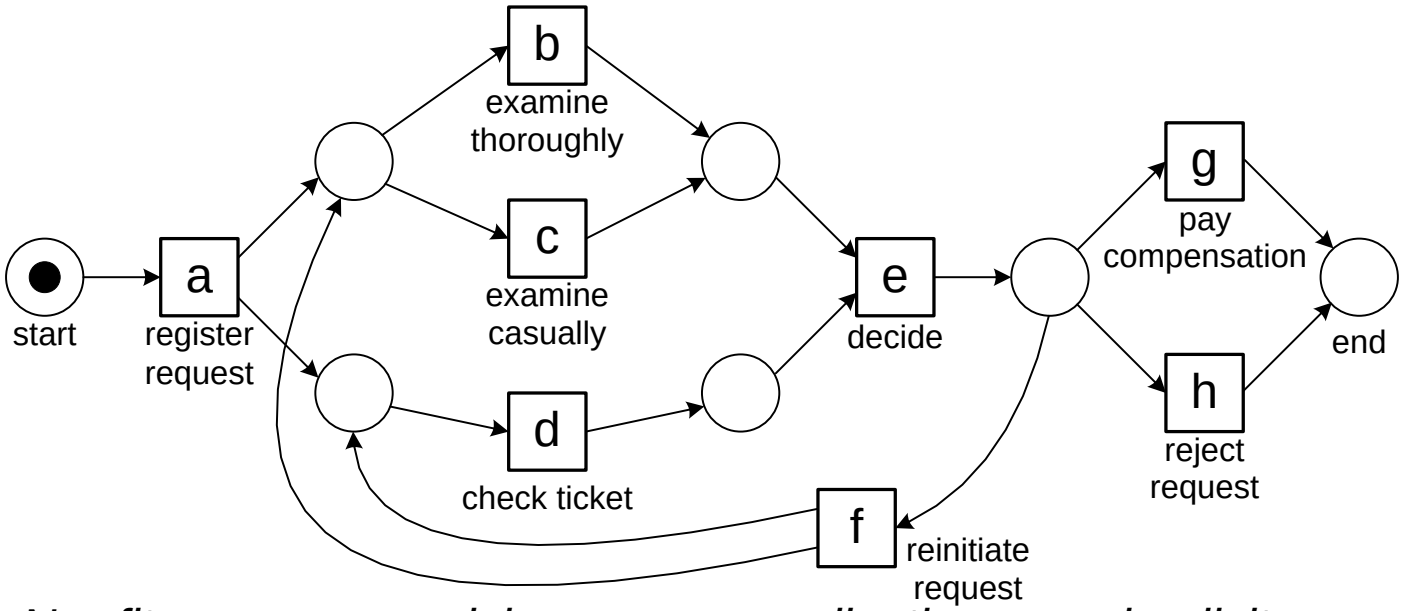


N_3 : fitness = +, precision = -, generalization = +, simplicity = +



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

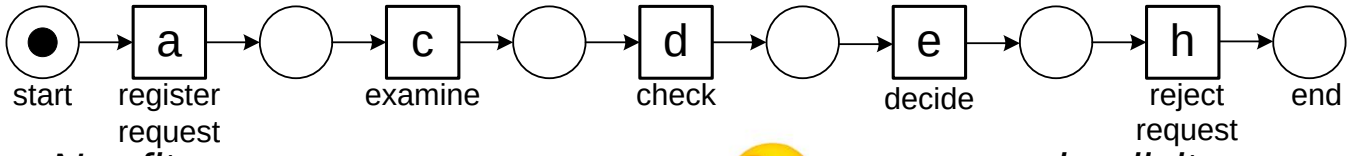
Model N_1



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	

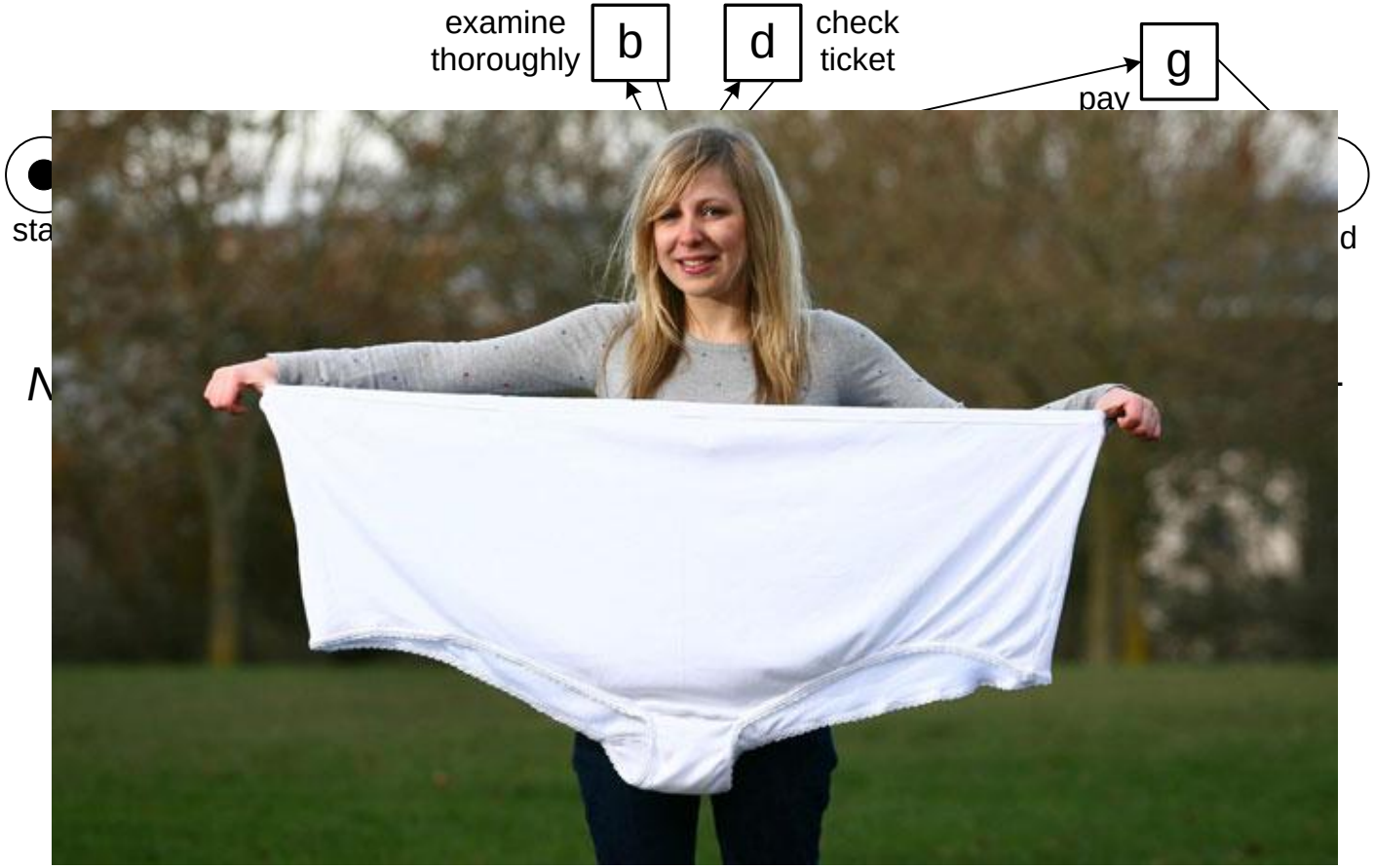
Model N_2



N_2 : fitness = *, simplicity = +*

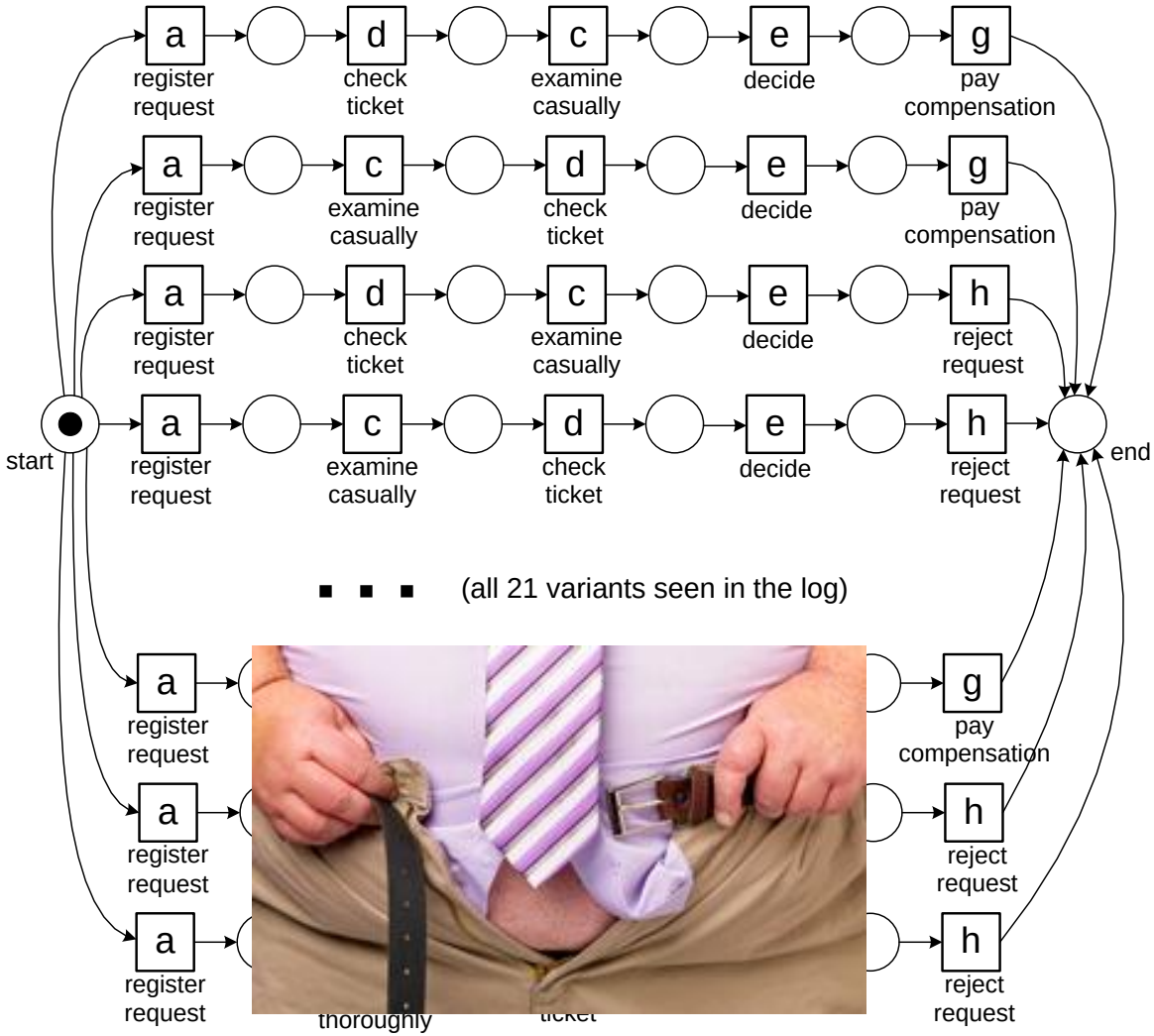
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

Model N_3



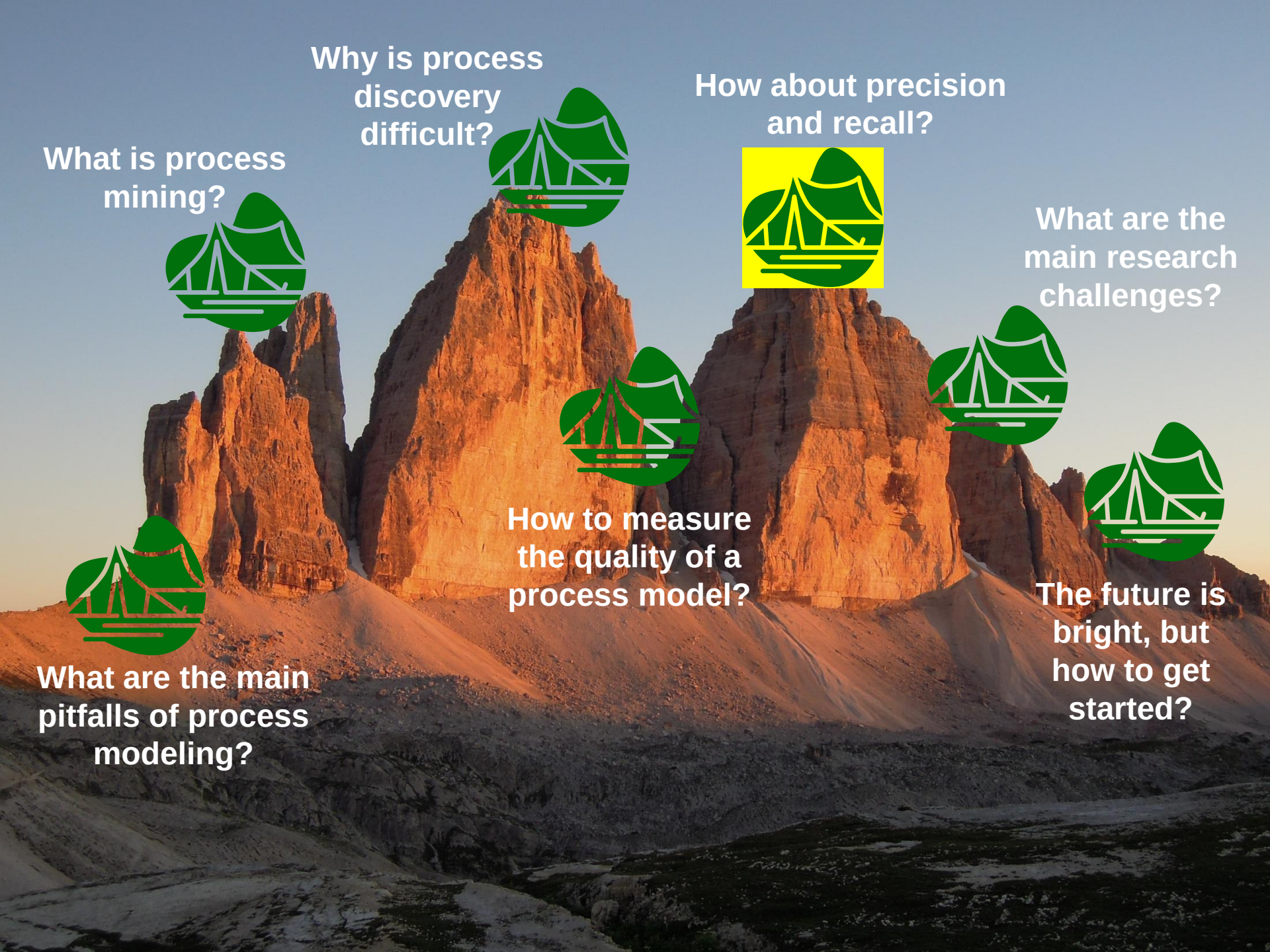
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

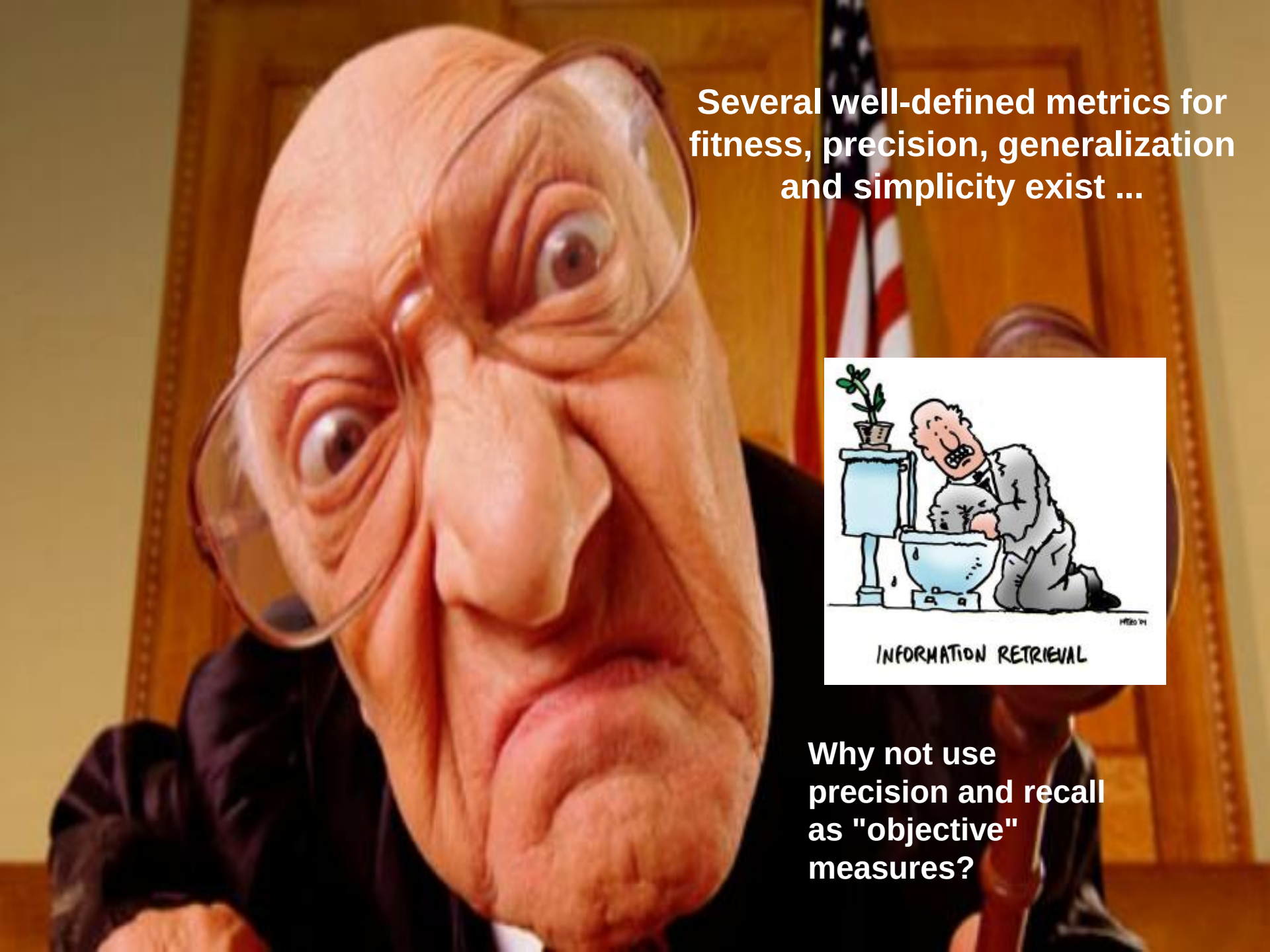
Model N_4



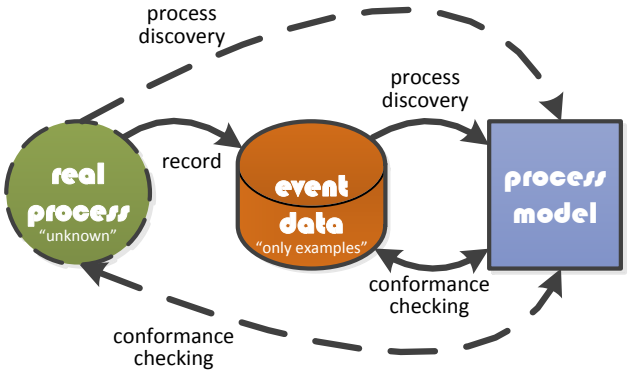
N_4 : fitness = +, precision = +, generalization = -, simplicity = -

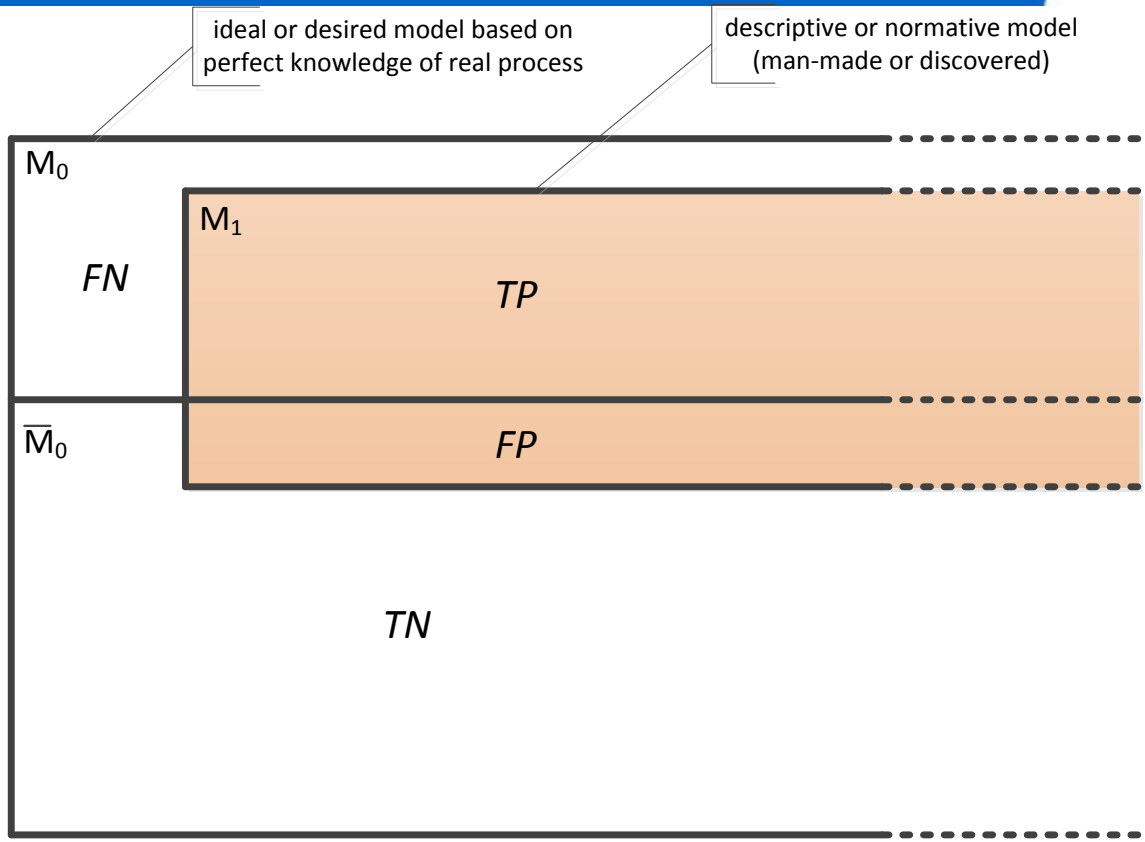
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	


What is process mining?
Why is process discovery difficult?
How about precision and recall?
What are the main research challenges?
How to measure the quality of a process model?
What are the main pitfalls of process modeling?
The future is bright, but how to get started?


Several well-defined metrics for fitness, precision, generalization and simplicity exist ...
INFORMATION RETRIEVAL
Why not use precision and recall as "objective" measures?

Suppose we know the "real model"

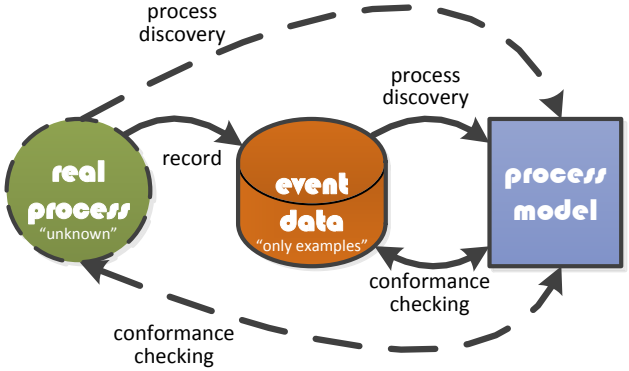




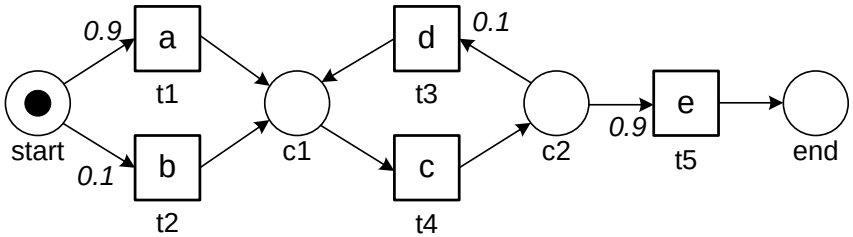
$$precision = \frac{TP}{M_1} = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{M_0} = \frac{TP}{TP + FN}$$

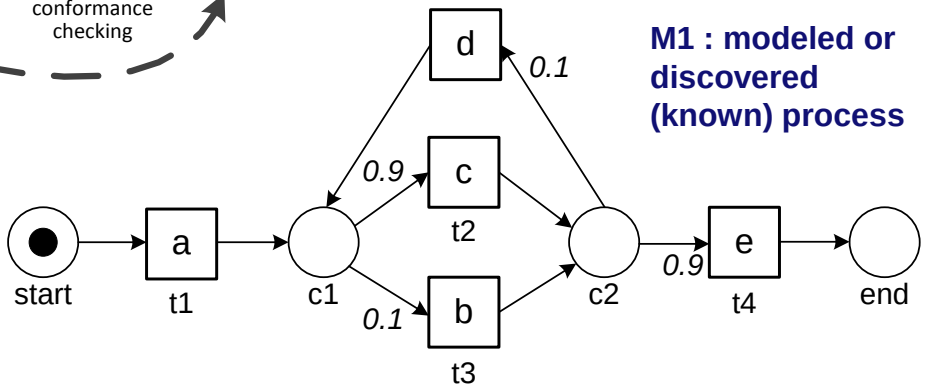
Let us try …



M_0 : unknown real process

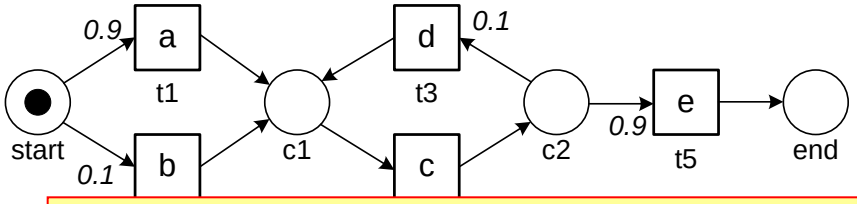


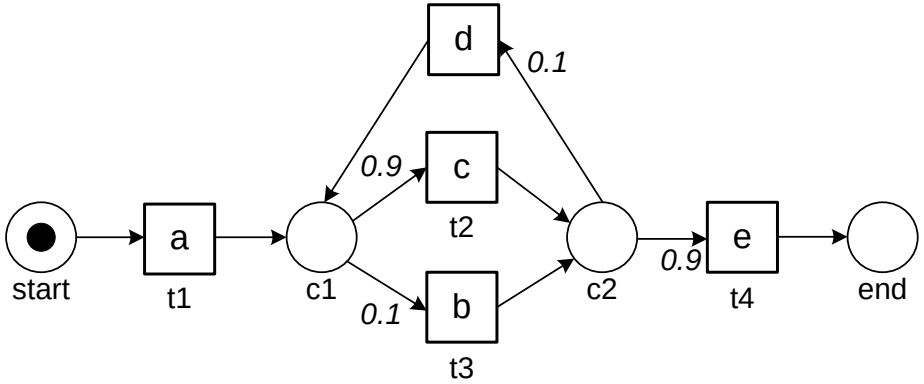
M1 : modeled or discovered (known) process



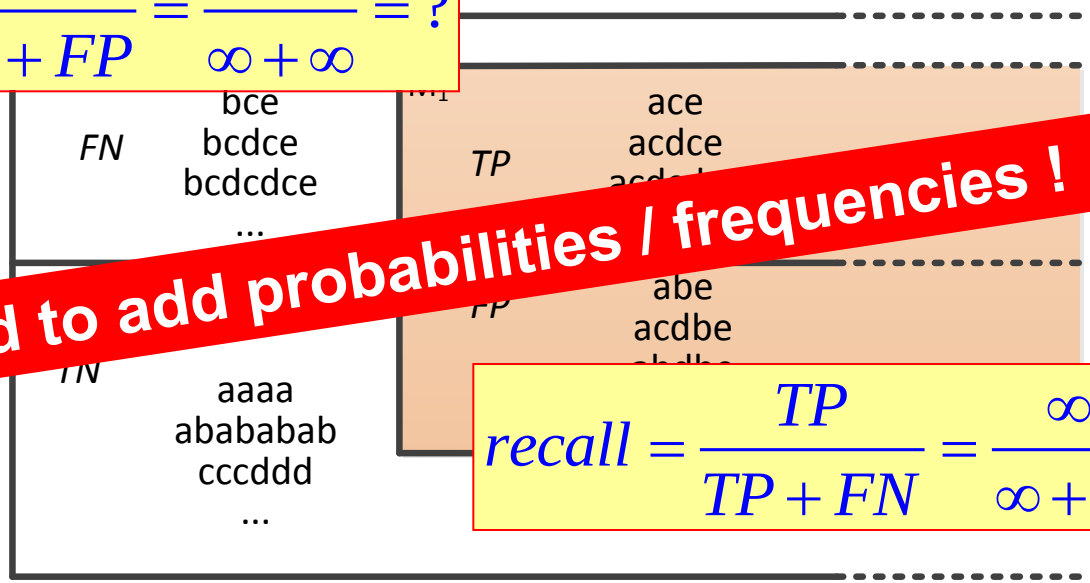
M_0		M_1	
FN	bce bcdce bcdcdce …	*TP*	ace acdce acdcdce …
$\overline{M}_0$		*FP*	abe acdbe abdbe …
TN	aaaa abababab cccddd …		

Oops …





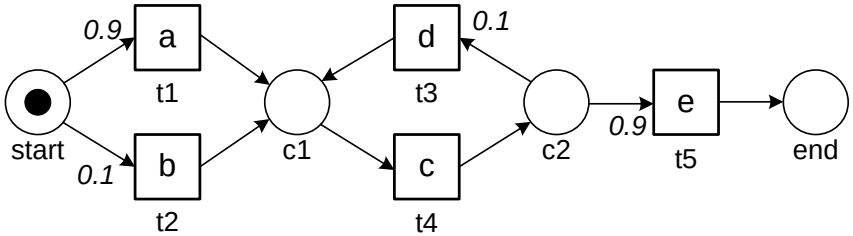
$$precision = \frac{TP}{TP + FP} = \frac{\infty}{\infty + \infty} = ?$$

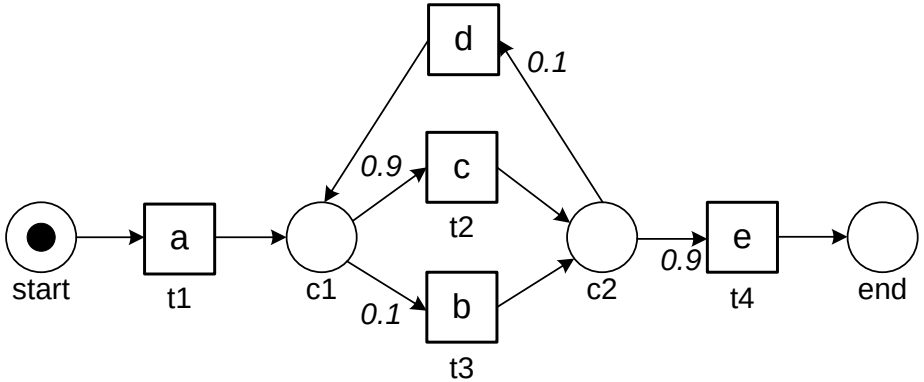


Need to add probabilities / frequencies !

$$recall = \frac{TP}{TP + FN} = \frac{\infty}{\infty + \infty} = ?$$

Let's compute precision and recall taking a specific viewpoint (M_0 or M_1)





$$precision_0 = \frac{TP}{TP+FP} = \frac{0.9}{0.9+0.0} = 1$$

ace	0.81
bce	0.09
acdce	0.081

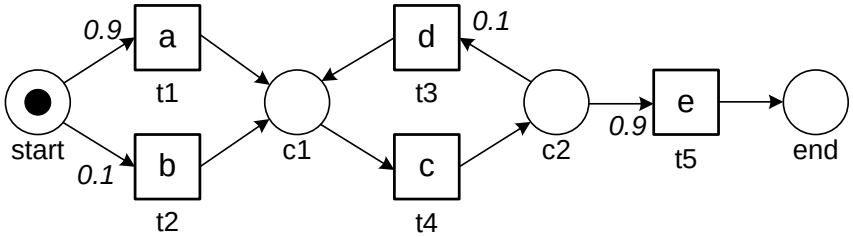
$$recall_0 = \frac{TP}{TP+FN} = \frac{0.9}{0.9+0.1} = 0.9$$

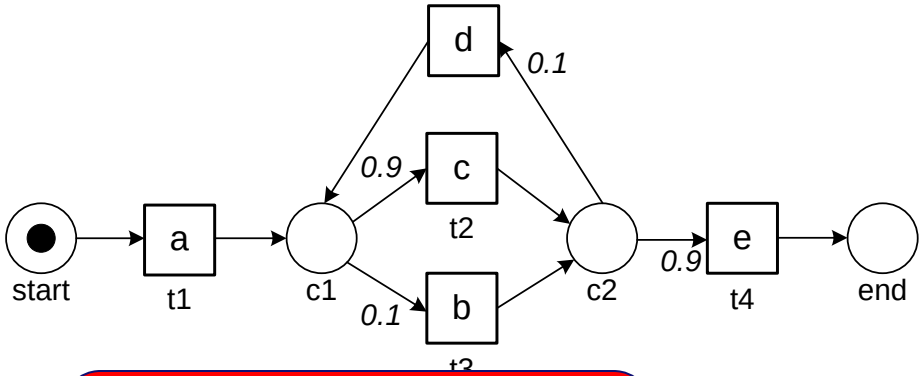
trace	probability in M_0
ace	0.81
acdce	0.0729
…	…
bce	0.0

$$precision_1 = \frac{TP}{TP+FP} = \frac{0.888}{0.888+0.111} = 0.888$$

$$recall_1 = \frac{TP}{TP+FN} = \frac{0.888}{0.888+0.0} = 1$$

Only two dimensions rather than four?

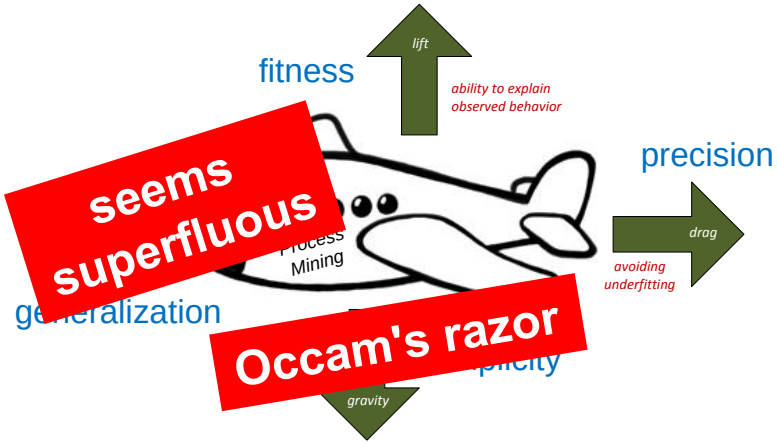




$$recall_0 = \frac{TP}{TP + FN} = \frac{0.9}{0.9 + 0.1} = 0.9$$

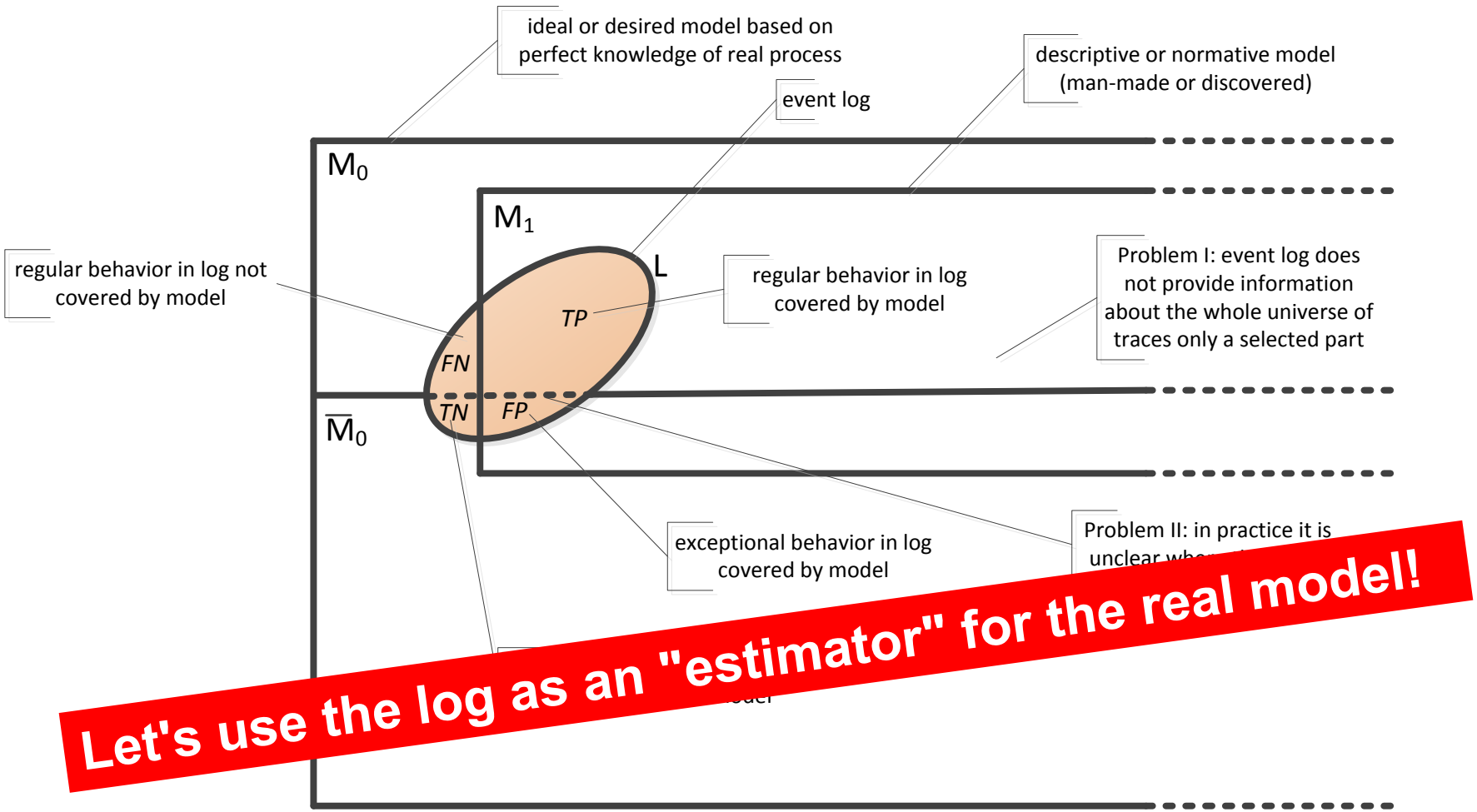
small if there are many likely traces in the real model that do not fit the discovered model

small if there are many likely traces in the discovered model that are impossible in the real model



$$precision_1 = \frac{TP}{TP + FP} = \frac{0.888}{0.888 + 0.111} = 0.888$$

But we have event logs and do not know the real process …



Let's use the log as an "estimator" for the real model!

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.

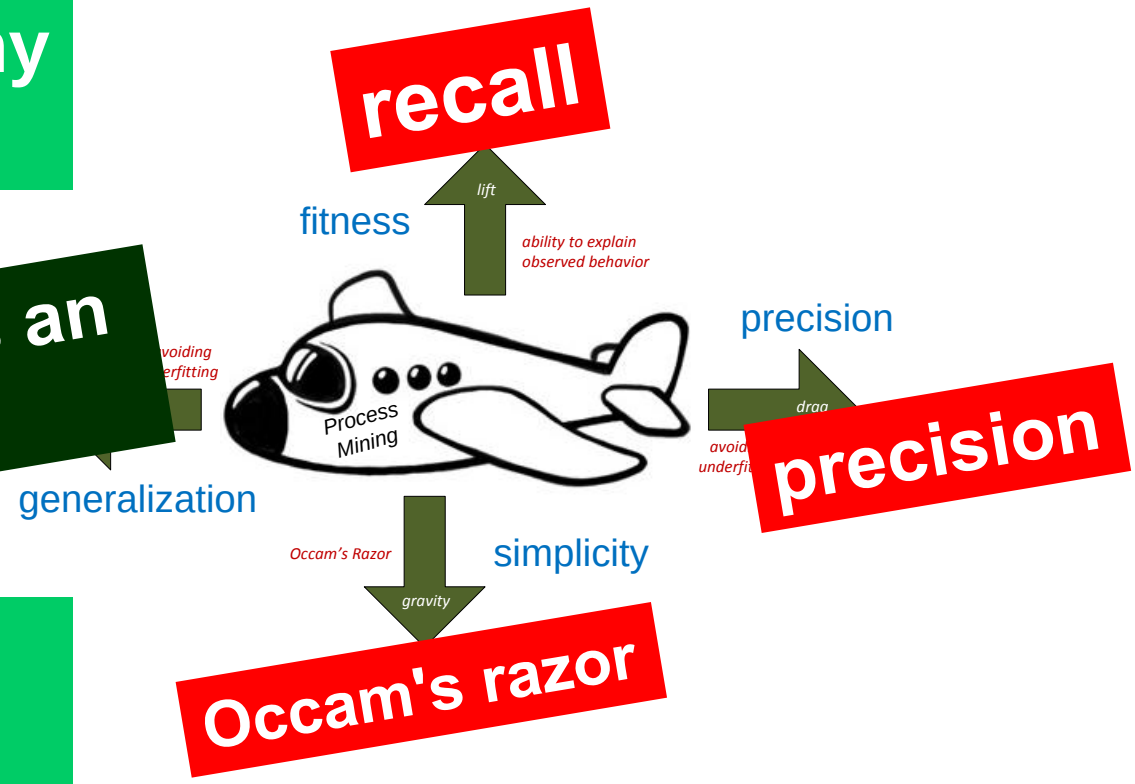
See paper for details!

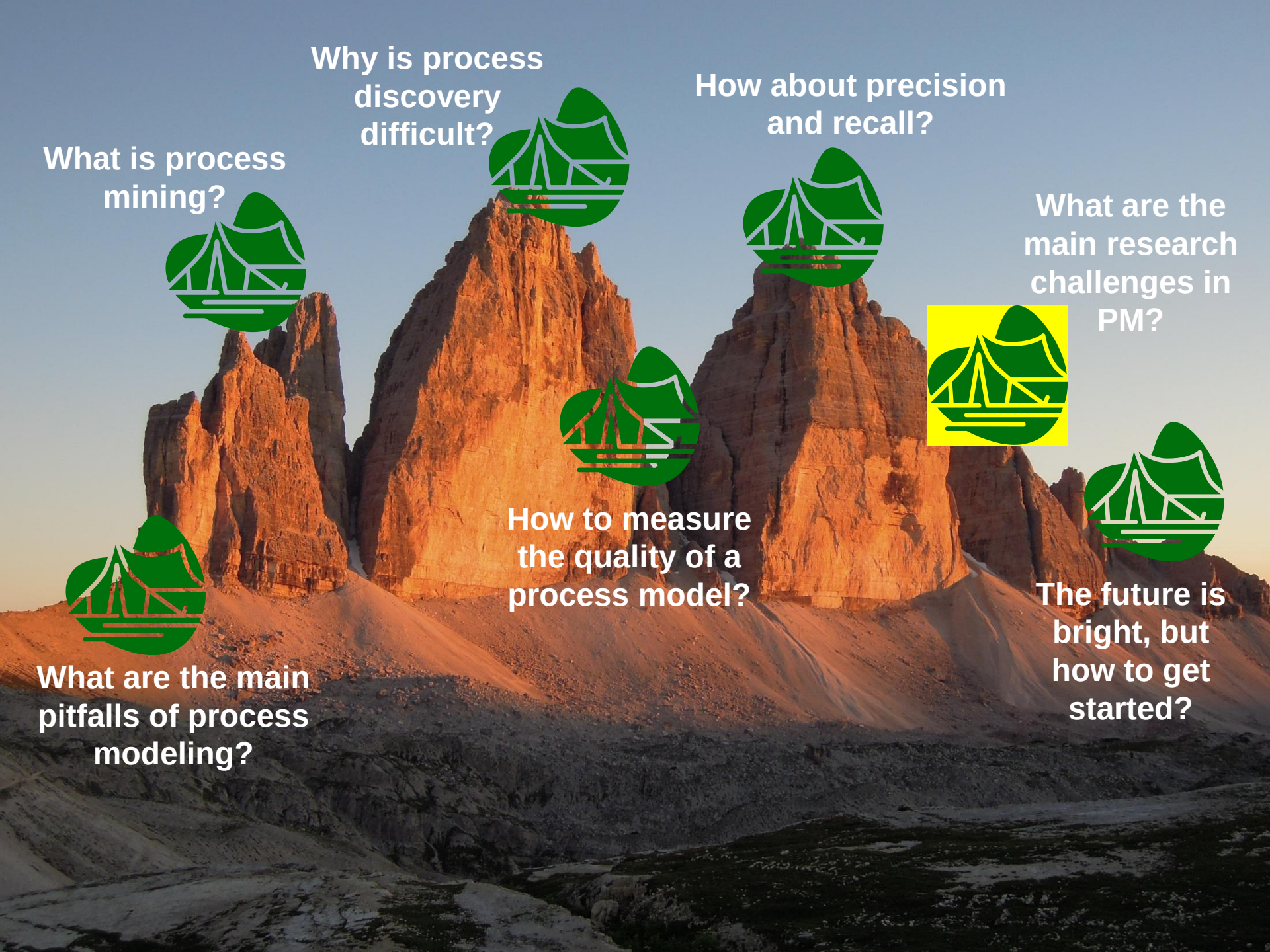
Considering event logs rather than the unknown model …

How complete is my event log?

quality of log as an estimator

Have we seen all characteristic behaviors?




What is process mining?
Why is process discovery difficult?
How about precision and recall?
What are the main research challenges in PM?
How to measure the quality of a process model?
What are the main pitfalls of process modeling?
The future is bright, but how to get started?


Finding sheep with five legs
we are getting close…

Distributing process mining problems to cope with big data

On-the-fly process mining

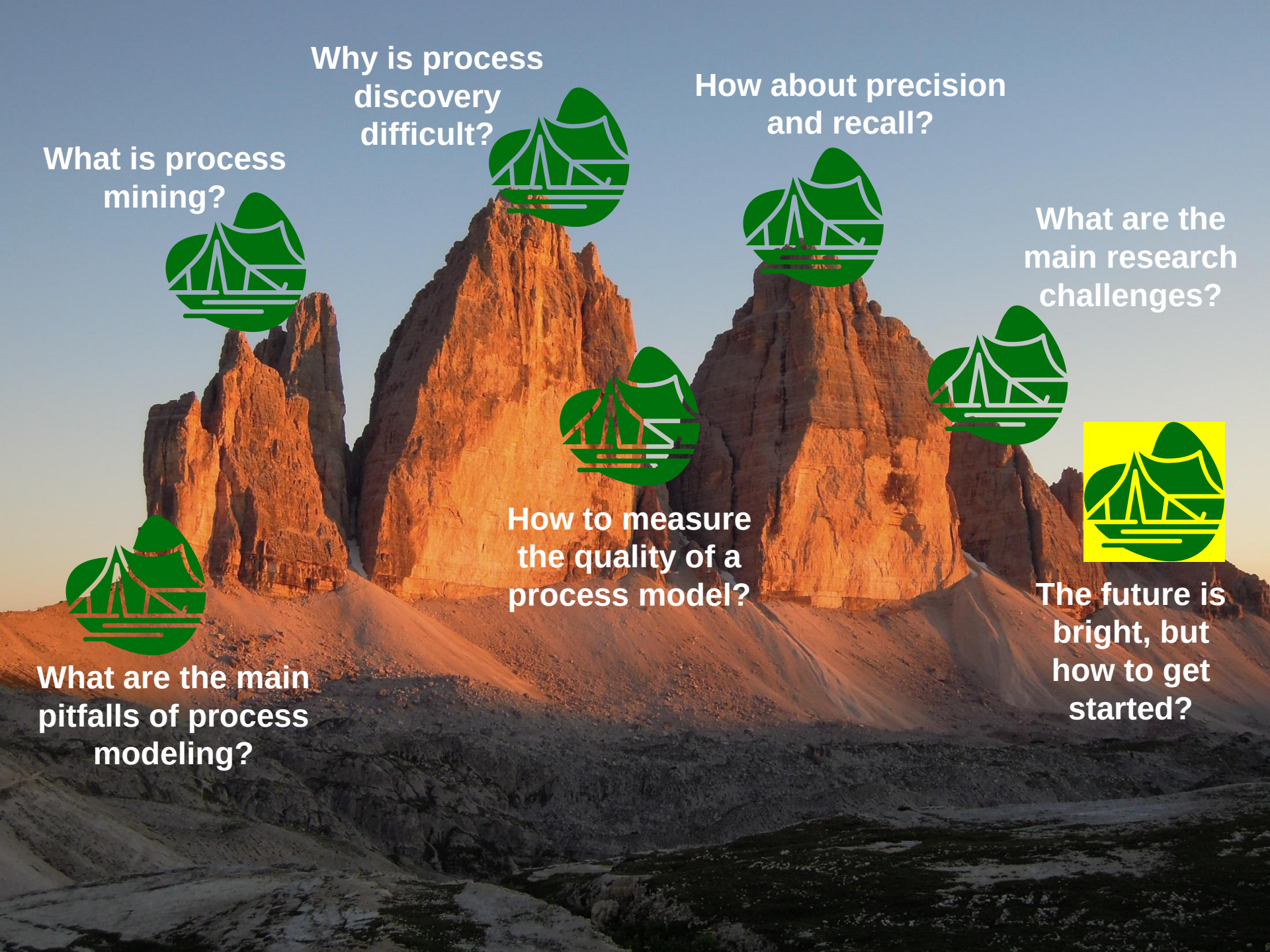
Operational support


Concept drift


cross-organizational / comparative process mining

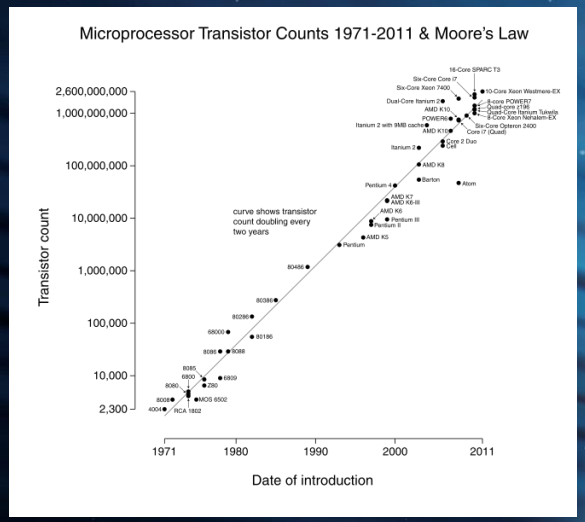

context aware process mining

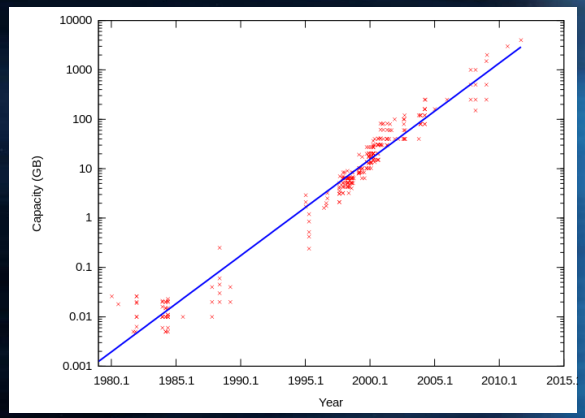
Supporting the process of process mining

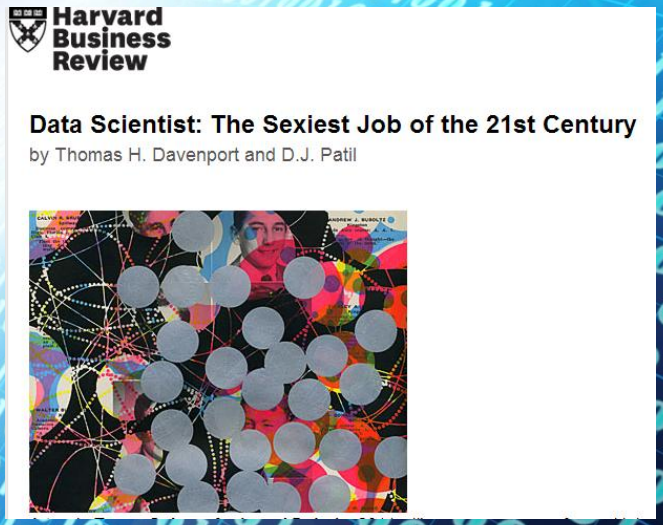

What is process mining?
Why is process discovery difficult?
How about precision and recall?
What are the main research challenges?
How to measure the quality of a process model?
What are the main pitfalls of process modeling?
The future is bright, but how to get started?

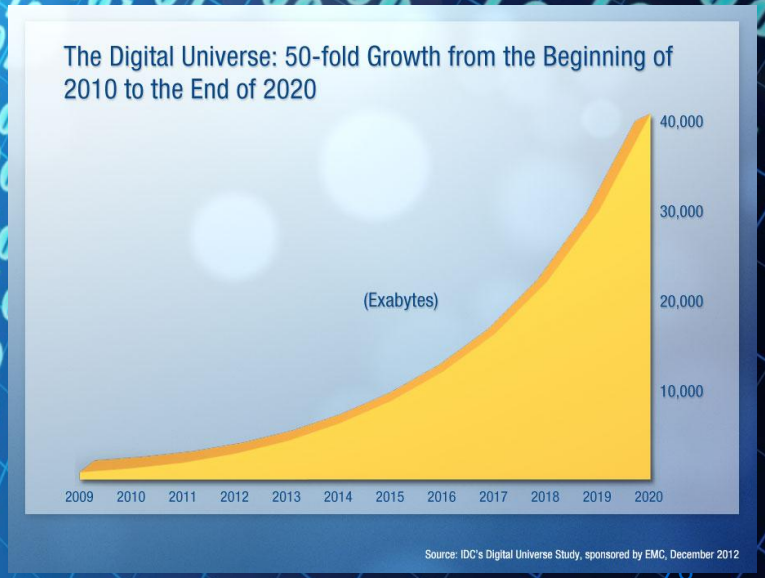

BRIGHT
FUTURE
AHEAD

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





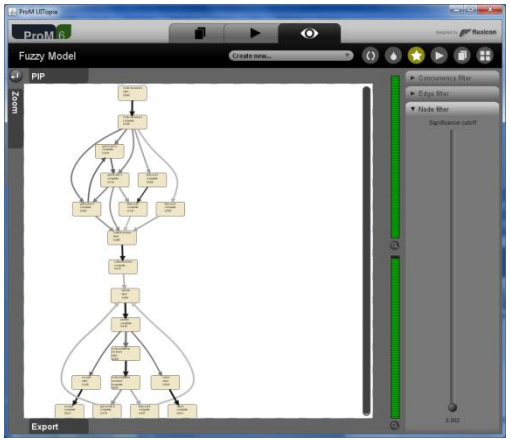


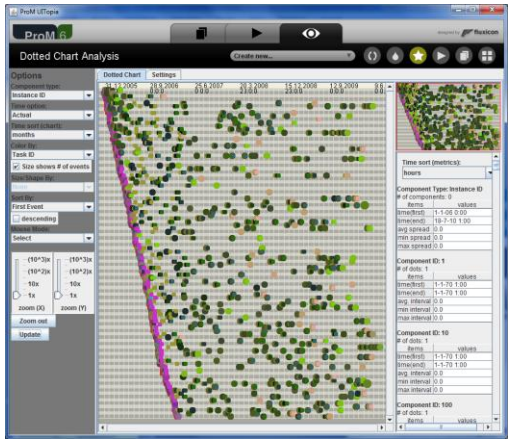


How to get started?

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

ProM
process mining workbench

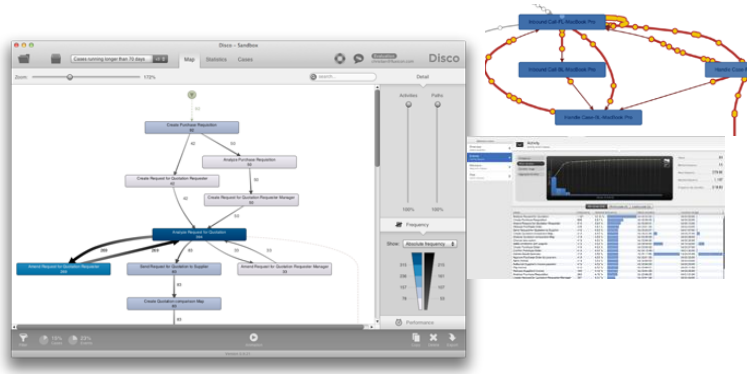


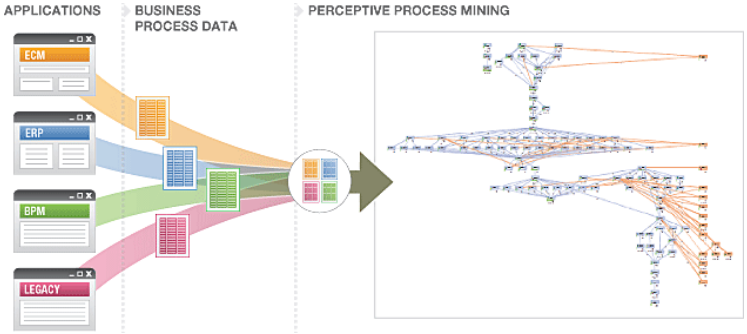


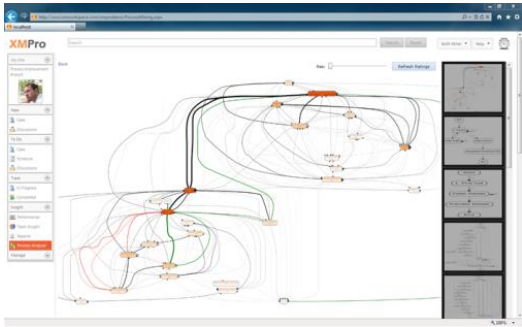
Download from: www.processmining.org

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)**
- **…**







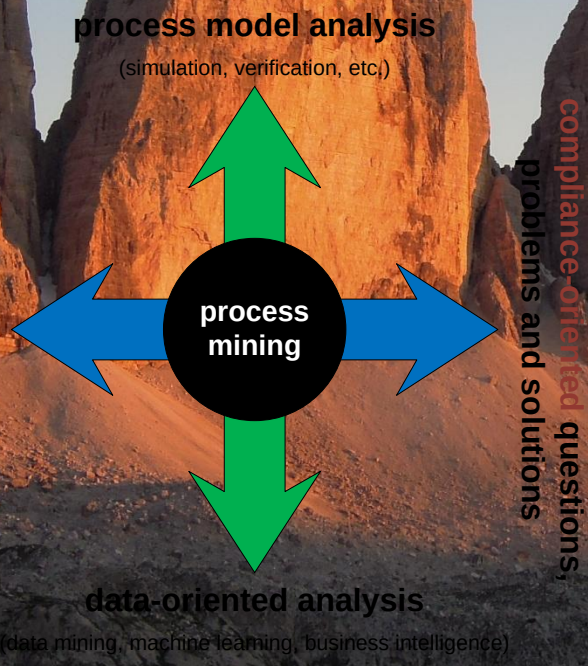
How to Get Started?

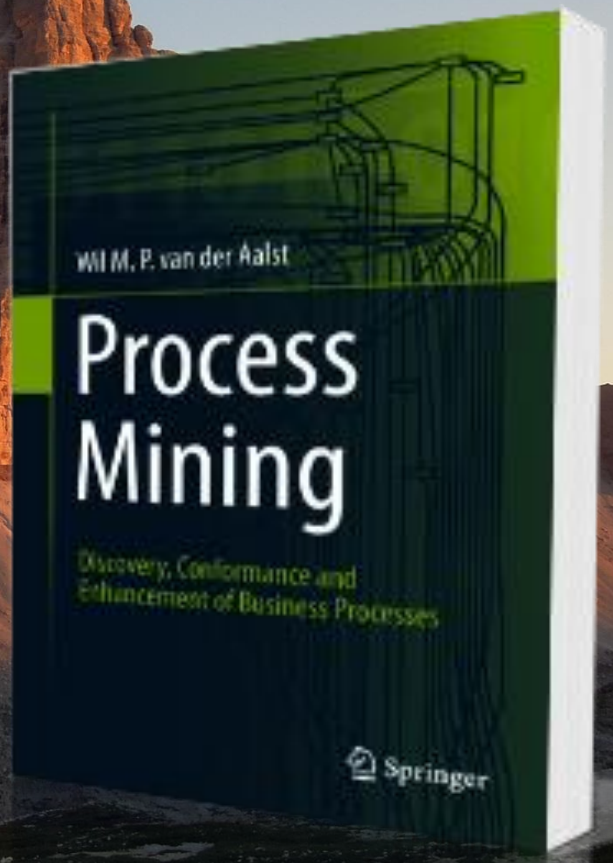
Collect event data ⟷ **Collect questions**

- **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:
The Quest for the "Right" Process





processmining.org