Processes @ your Service

Using Process Mining to Turn Big Data into Real Value

prof.dr.ir. Wil van der Aalst

ICWE '13
AALBORG

Wil M. P. van der Aalst

Process Mining

Discovery, Conformance and Enhancement of Business Processes

Springer

Web Engineering

model, specify, configure, implement processes

observe behavior (e.g., event data)

Big … or fast and efficient?

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?

BPEL

WSAH

not solved by new notations or standards

notation/standards are not the key issue

- **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

problem #2

Straightjacketing smaller interacting processes into one monolithic model

What is the process instance?

Order
- OrderID : OrderID
- Customer : CustID
- Amount : Euro
- Created : DateTime
- Paid : DateTime
- Completed : DateTime

1 — 1..*

Orderline
- OrderLineID : OrderLineID
- OrderID : OrderID
- Product : ProdID
- NofItems : PosInt
- TotalWeight : Weight
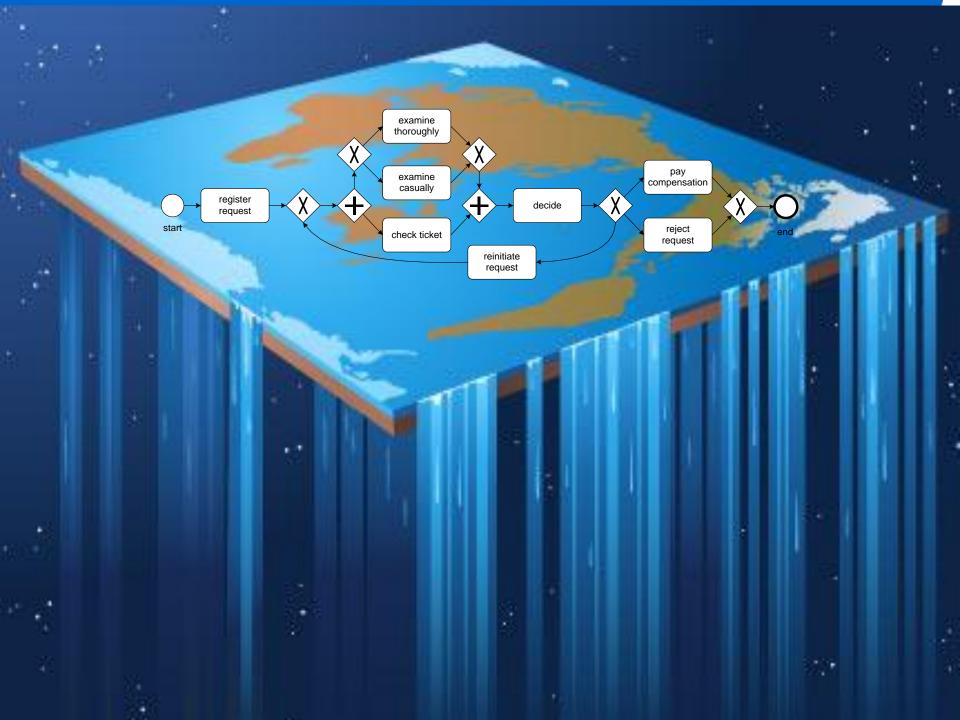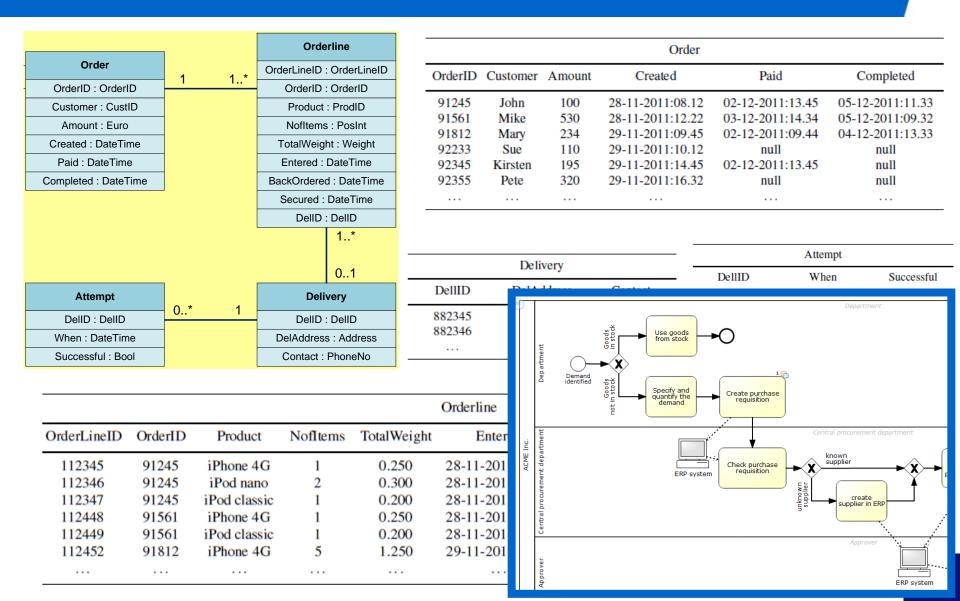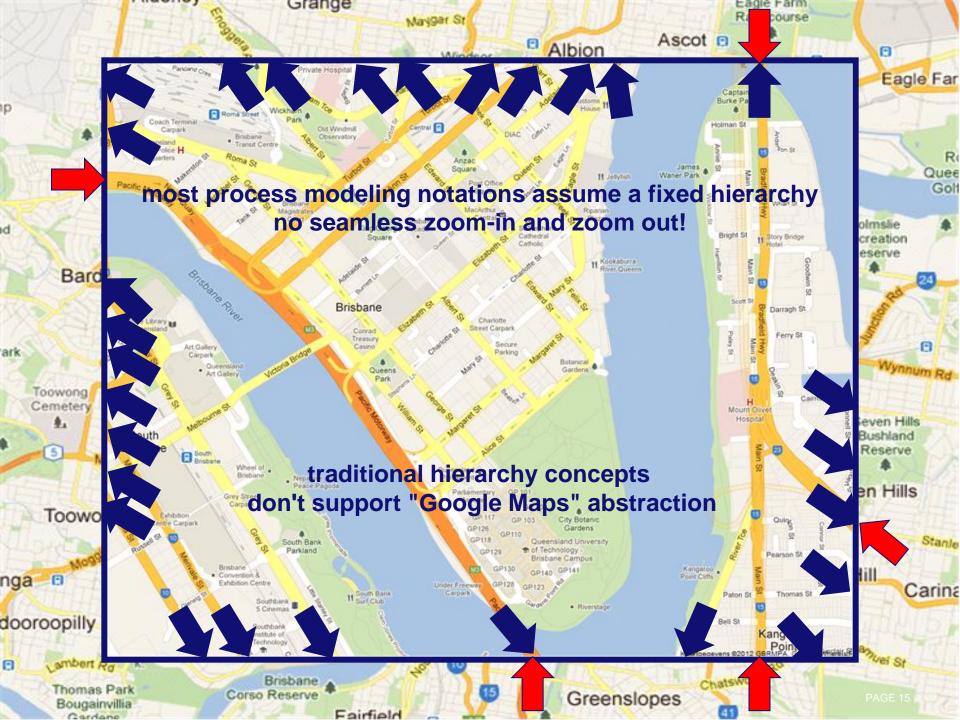- Entered : DateTime
- BackOrdered : DateTime
- Secured : DateTime
- DelID : DelID

1..*

0..1

Attempt
- DelID : DelID
- When : DateTime
- Successful : Bool

0..* — 1

Delivery
- DelID : DelID
- DelAddress : Address
- Contact : PhoneNo

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
…	…	…	…	…	…

Attempt		
DelID	When	Successful

Delivery		
DelID	DelAddress	Contact
882345		
882346		
…		

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

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

Yes

Timed Out [1 week]

Prepare Results

Results Web Site

E-Mail Results of Vote

Issues w/o Majority?

No

Did Enough Members Vote?

Yes

Change Vote Message

Vote Results

Have the members been warned?

Yes

No

Reduce number of Voting Members and Recalculate Vote

Re-announce Vote with

No

ISC Sup

Vendor

No

Washington product?

No

Yes

IT Service Center WN addresses

customer

Resolves problem?

No

Yes

Positioning Process Mining

process model analysis

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

performance-oriented questions, problems and solutions

process mining
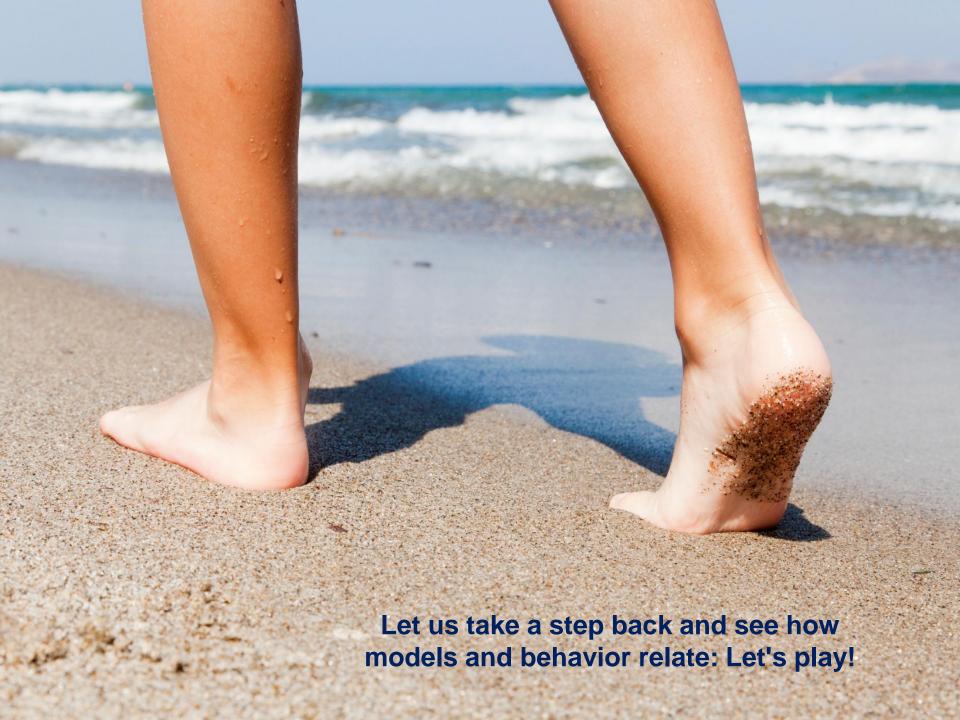
compliance-oriented questions, problems and solutions

data-oriented analysis
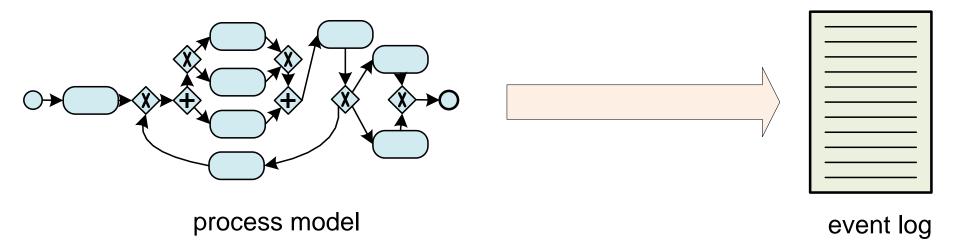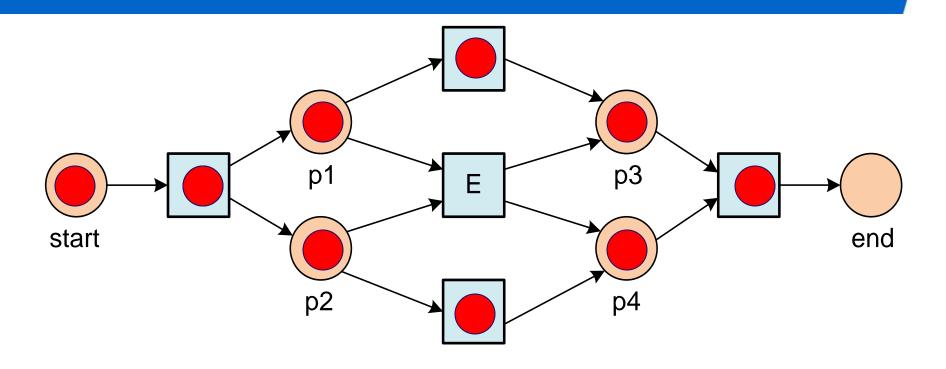
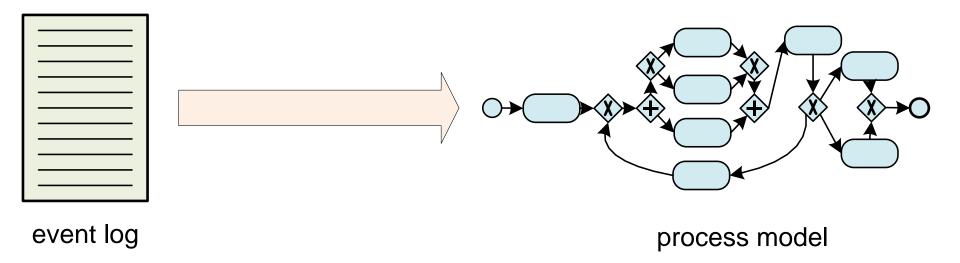(data mining, machine learning, business intelligence)

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

Play-Out

process model

event log

Play-In

event log

process model
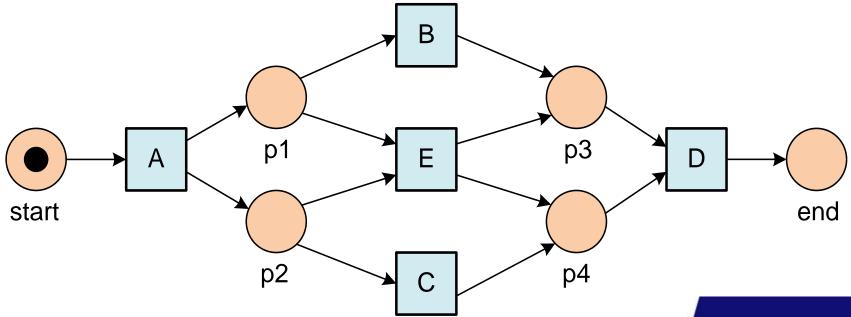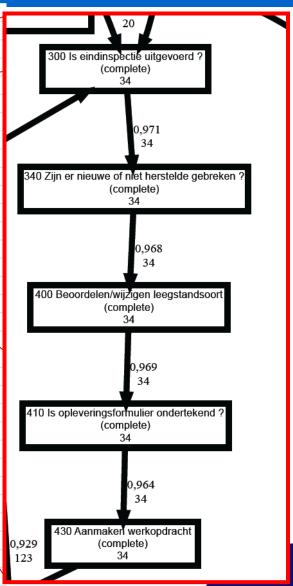
Example Process Discovery
(AMC, 627 gynecological oncology patients, 24331 events)

Replay

event log   process model

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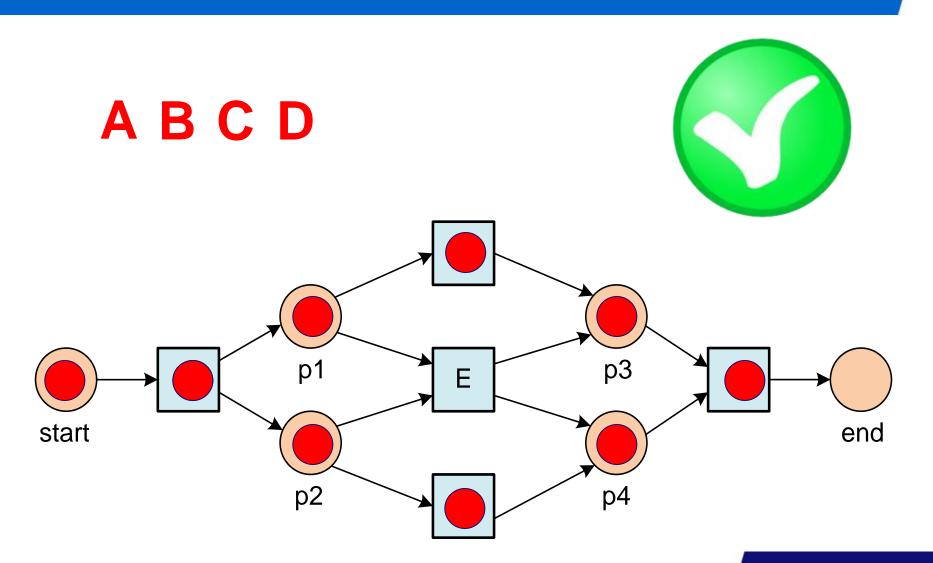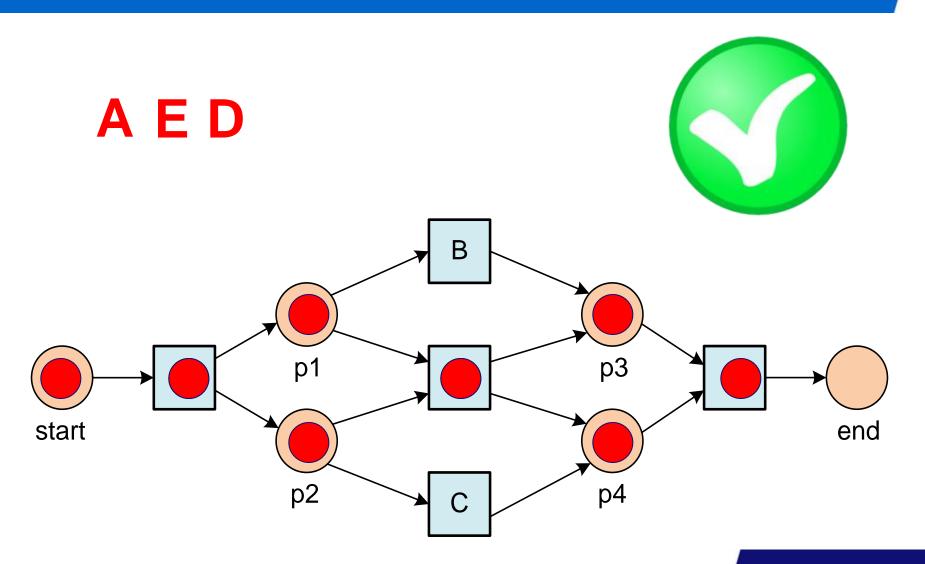- recommendations

Replay

A B C D

Replay
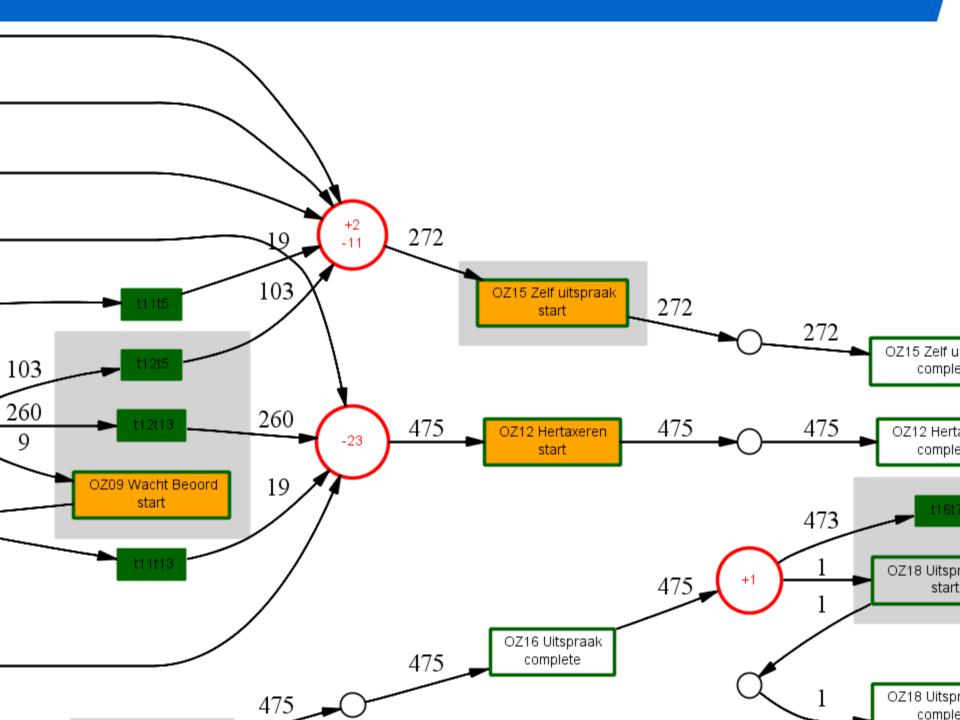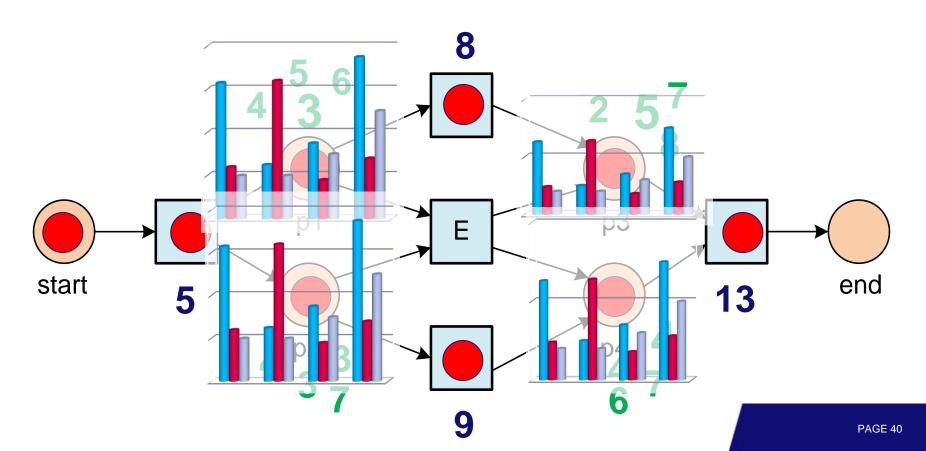
A E D
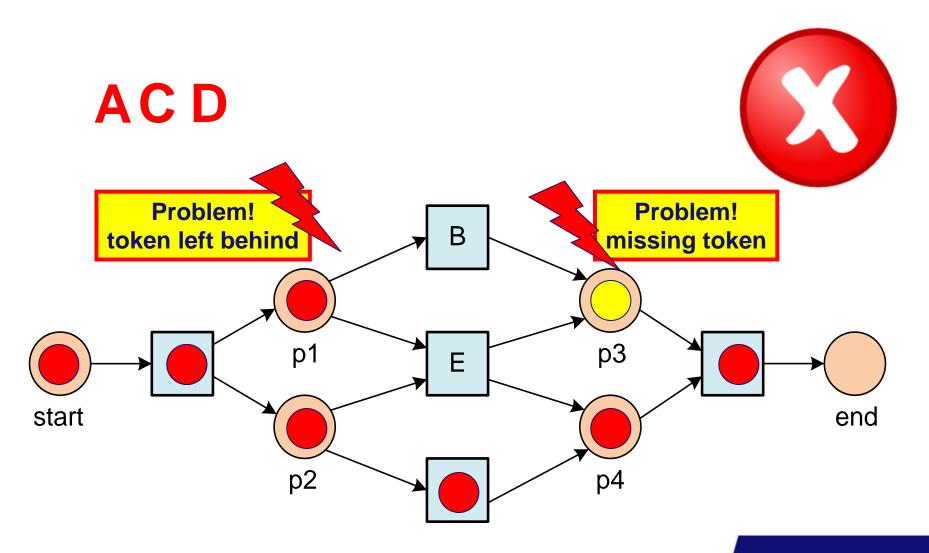
Replay can detect problems

PAGE 38
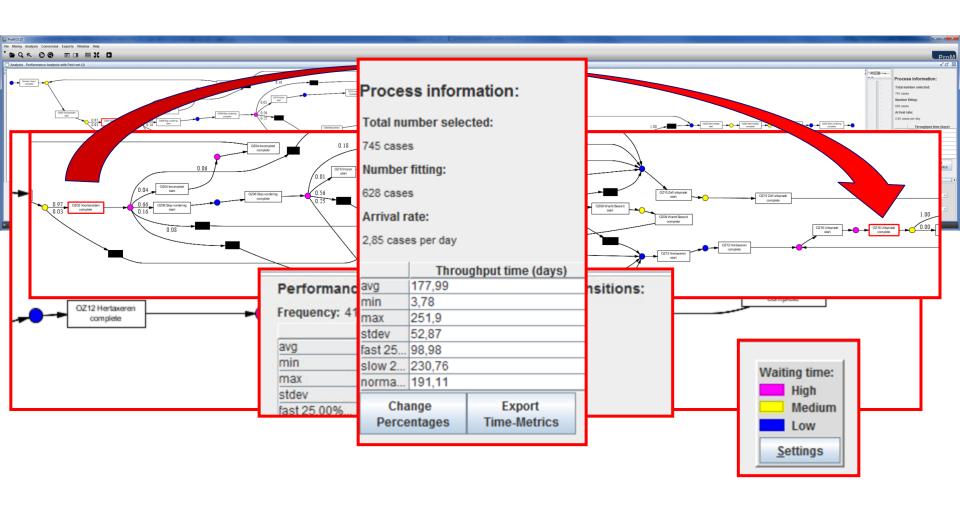
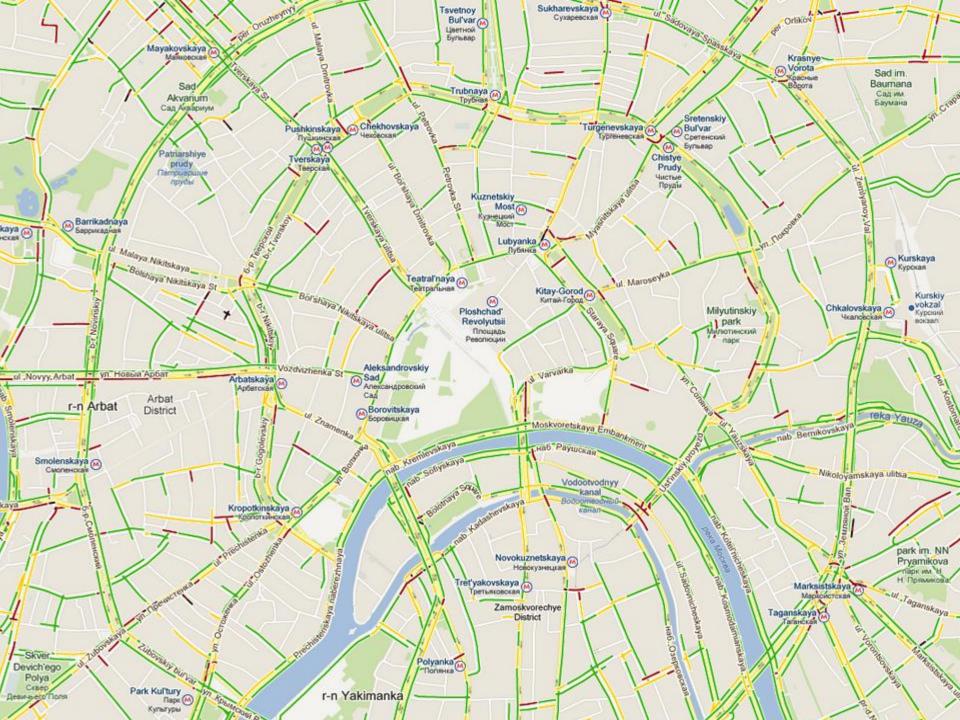# Performance Analysis Using Replay
(WOZ objections Dutch municipality, 745 objections, 9583 event, f= 0.988)

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

Tsvetnoy Bul'var
Цветной бульвар

Sukharevskaya
Сухаревская

ul. Sadovaya-Spasskaya

per. Orlikov

Mayakovskaya
Маяковская

per. Oruzheynyy

Krasnye Vorota
Красные Ворота

Sad im. Baumana
Сад им. Баумана

Sad Akvarium
Сад Аквариум

ul. Malaya Dmitrovka

Trubnaya
Трубная

ul. Petrovka

Tverskaya St

Pushkinskaya
Пушкинская

Chekhovskaya
Чеховская

Turgenevskaya
Тургеневская

Sretenskiy Bul'var
Сретенский бульвар

ul. Staraya

Patriarshiye prudy
Патриаршие пруды

Tverskaya
Тверская

ul. Bol'shaya Dmitrovka

Chistye Prudy
Чистые Пруды

Petrovka St

Kuznetskiy Most
Кузнецкий Мост

ul. Zemlyanoy Val

Barrikadnaya
Баррикадная

Tverskaya ulitsa

Tverskaya Dmitrovka

Myasnitskaya ulitsa

Kurskaya
Курская

ul. Malaya Nikitskaya

b-r Tverskoy

Lubyanka
Лубянка

ul. Pokrovka

Kurskiy vokzal
Курский вокзал

"Bolshaya" Nikitskaya St

b-f. Tverskoy

ul. Maroseyka

Bol'shaya Nikitskaya ulitsa

Teatral'naya
Театральная

Kitay-Gorod
Китай-Город

Milyutinskiy park
Милютинский парк

Chkalovskaya
Чкаловская

b-r Nikitskiy

Ploshchad' Revolyutsii
Площадь Революции

Staraya Square

b-r Novinskiy

Vozdvizhenka St

Aleksandrovskiy Sad
Александровский Сад

ul. Varvarka

Arbatskaya
Арбатская

ul. Solyanka

r-n Arbat

Arbat District

ul. Znamenka

Borovitskaya
Боровицкая

Moskvoretskaya Embankment

nab. Bernikovskaya

reka Yauza

b-r Gogolevskiy

ul. Novyy Arbat

ul. "Новый Арбат"

nab. Kremlevskaya

nab. Sofiyskaya

nab. Raushskaya

Ust'inskiy proyezd

nab. Yauzskaya

Nikoloyamskaya ulitsa

Smolenskaya
Смоленская

ul. Volkhonka

Vodootvodnyy kanal
Водоотводный канал

reka Moskva

park im. NN Pryamikova
парк им. Н. Пряникова

Kropotkinskaya
Кропоткинская

Bolotnaya Sq.

nab. Kadashevskaya

nab. Kotel'nicheskaya

ul. Prechistenka

Novokuznetskaya
Новокузнецкая

ul. Sadovnicheskaya

nab. Kosmodamianskaya

Marksistskaya
Марксистская

ul. Ostozhenka

Tret'yakovskaya
Третьяковская

Zamoskvorechye District

ul. Taganskaya

Prechistenskaya naberezhnaya

Taganskaya
Таганская

Skver Devich'ego Polya
Сквер Девичьего Поля

Zubovskiy bul'var

Polyanka
Полянка

ul. Ozerkovskaya

ul. Vorontsovskaya

Park Kul'tury
Парк Культуры

ul. Krymskiy
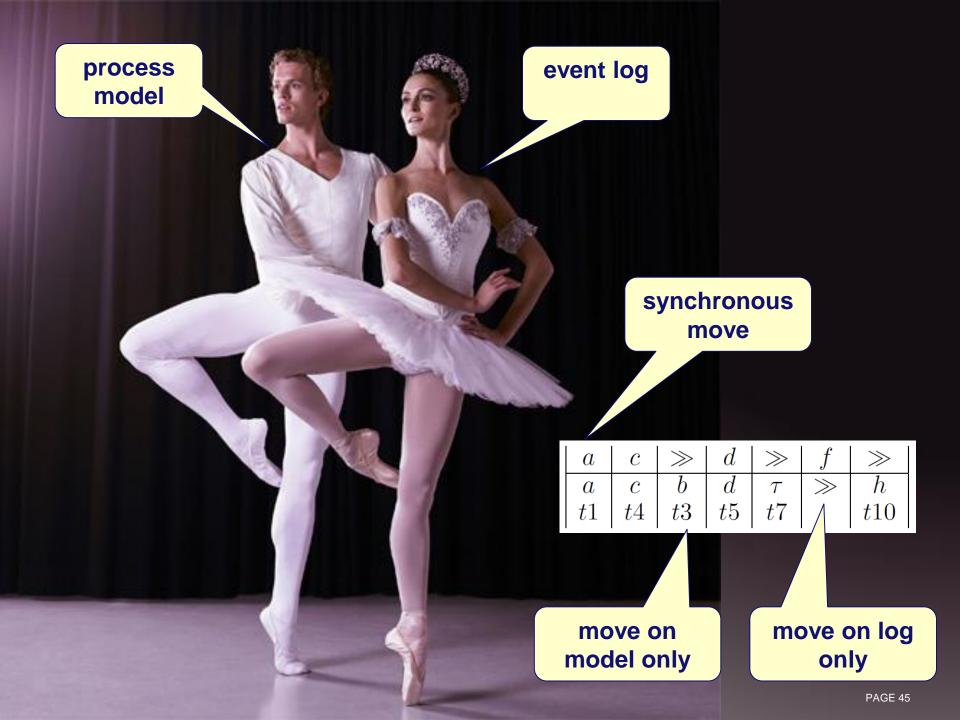
r-n Yakimanka

Markistskaya ul.

per. Kolomenskiy

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
$t1$	$t4$	$t3$	$t5$	$t7$		$t10$
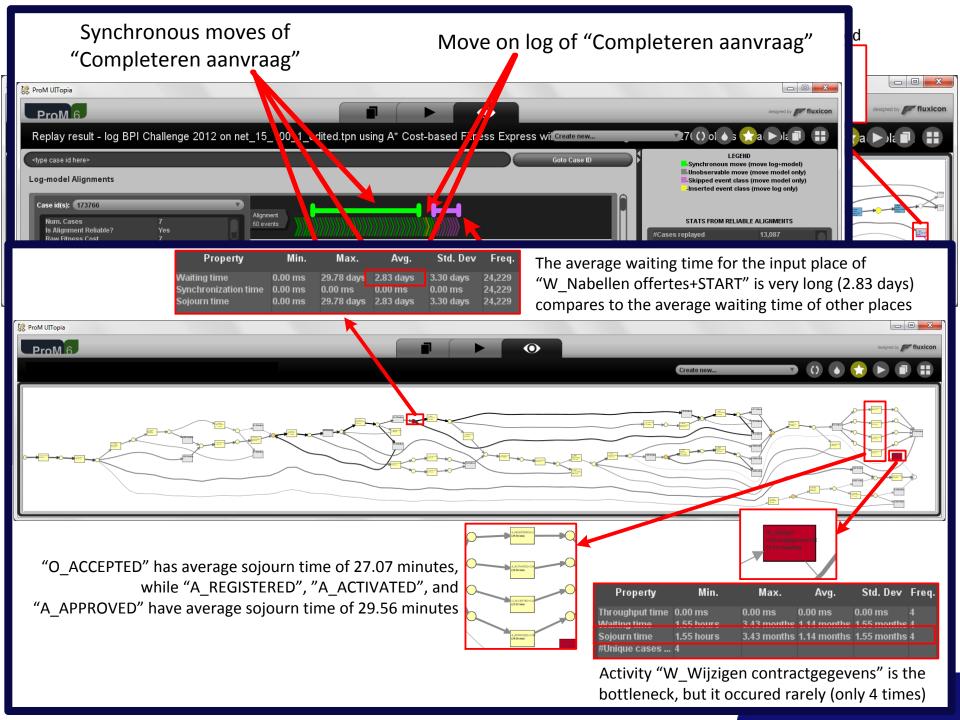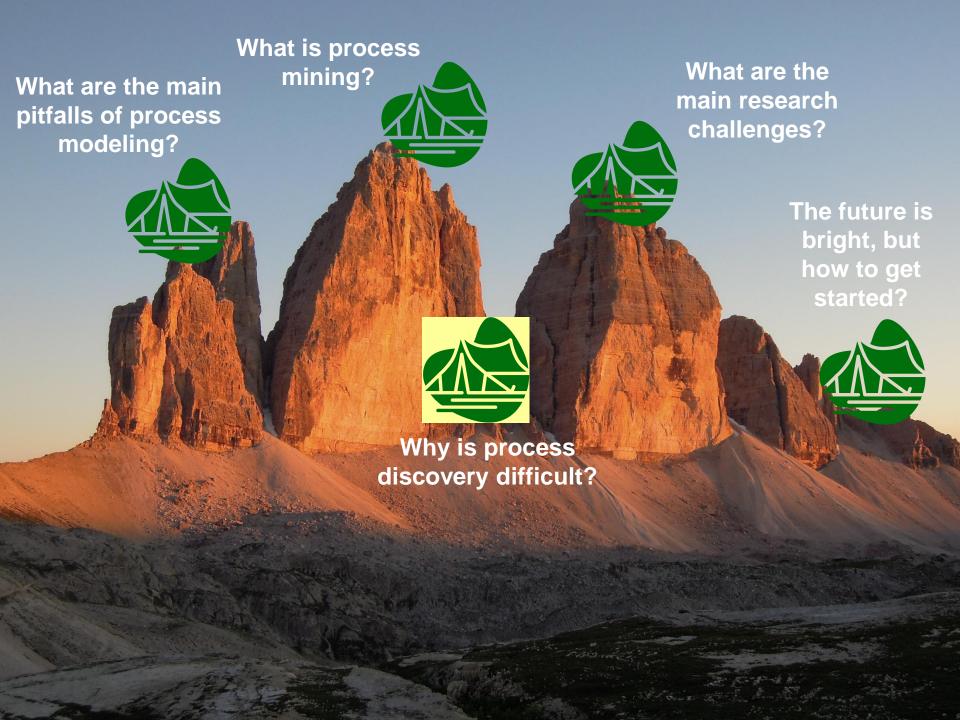
Example: BPI Challenge 2012
(Dutch financial institute, doi:10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f)

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

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"

LEGEND
- ■ Synchronous move (move log+model)
- ■ Unobservable move (move model only)
- ■ Skipped event class (move model only)
- ■ Inserted event class (move log only)

STATS FROM RELIABLE ALIGNMENTS
#Cases replayed   13,087

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)

Language identification in the limit (Mark Gold 1967)

abc

abd

ad

abbc

ac

…

abc ?

ab(c|d) ?

(ad)|(ab(c|d)) ?

ab*(c|d) ?

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

Learning is not easy …

- **Even simple language** ~~regular language~~ **learnable** ~~
- **M** ~~... with or ... tive examples, ... les, etc.**

... tence ≅ trace in event log
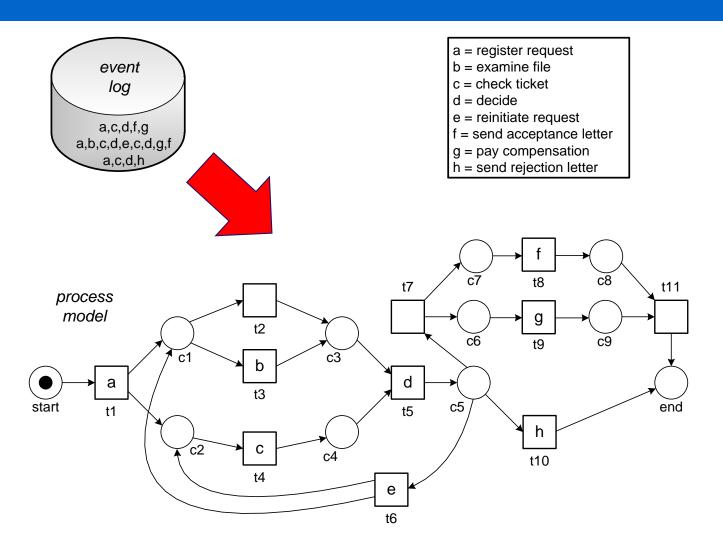
language ≅ process model

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

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

remarkable progress over a relatively short period

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

Process discovery algorithms
(small selection)

distributed genetic mining

automata-based learning

language-based regions

heuristic mining

state-based regions

genetic mining

LTL mining

stochastic task graphs

neural networks

fuzzy mining

hidden Markov models

mining block structures

α algorithm

conformal process graph

multi-phase mining
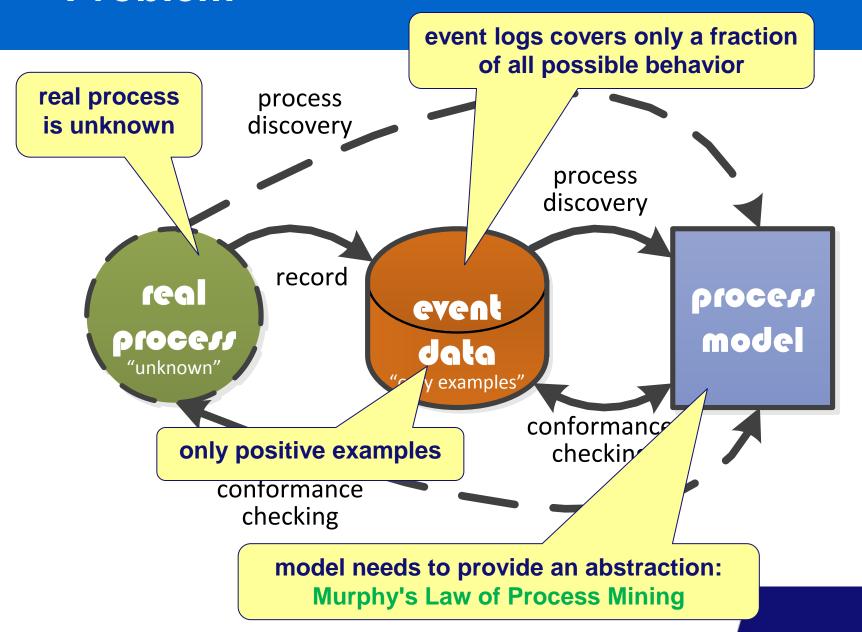
partial-order based mining

α# algorithm
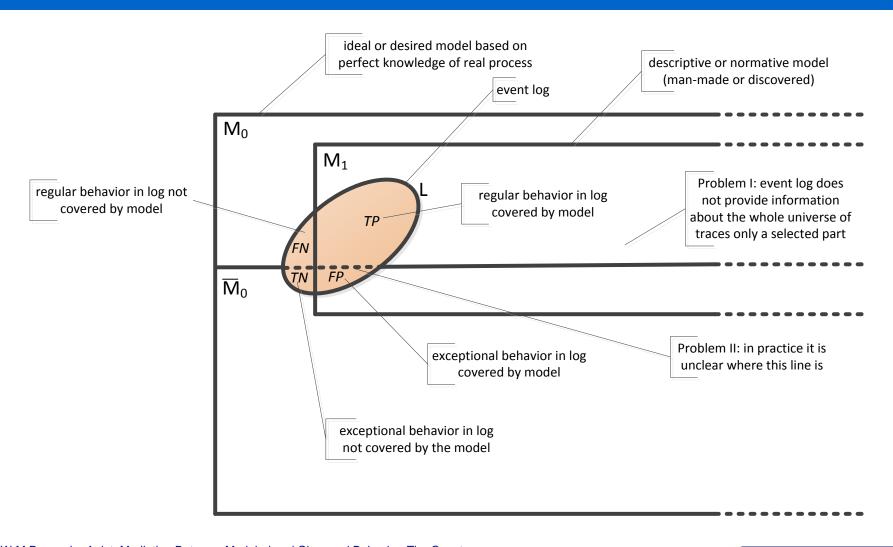
ILP mining

α++ algorithm
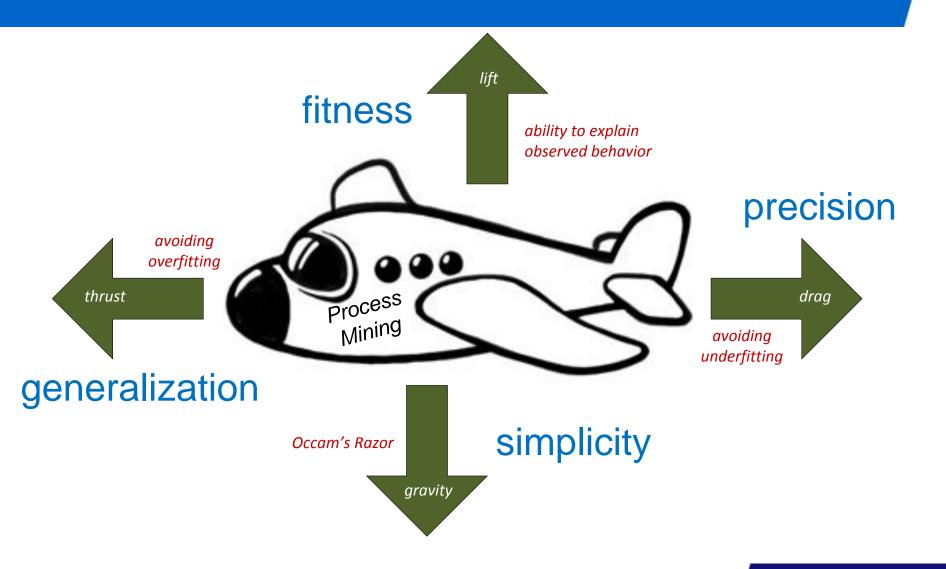
Problem

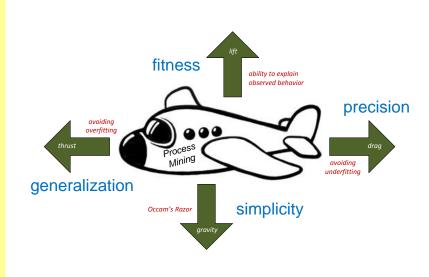# We only have example behavior (event log) and do not know the real process …

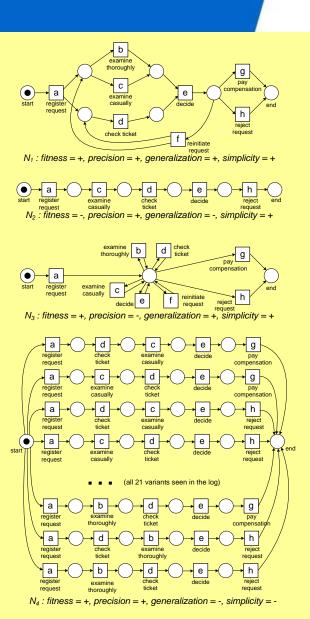W.M.P. van der Aalst. Mediating Between Modeled and Observed Behavior: The Quest for the "Right" Process. In *IEEE International Conference on Research Challenges in Information Science (RCIS 2013)*, pages 31-43. IEEE Computing Society, 2013.

Balance four forces

fitness

lift

ability to explain observed behavior

precision

avoiding overfitting

thrust

Process Mining

drag

avoiding underfitting

generalization

Occam's Razor

simplicity

gravity

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	
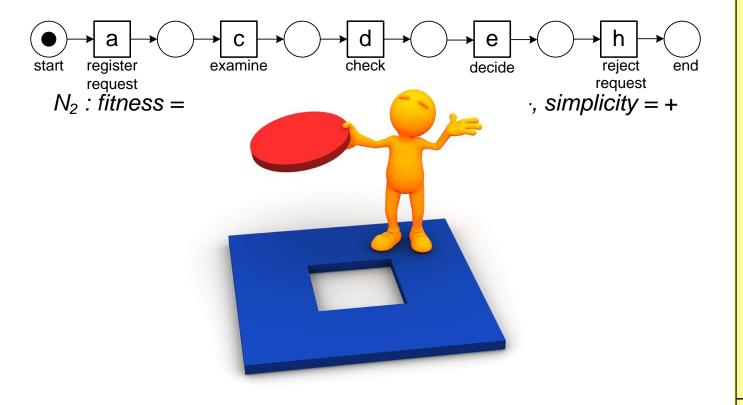
Model N₁

N₁ : 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$_2$

start | a register request | c examine | d check | e decide | h reject request | end

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$_3$

examine thoroughly **b** **d** check ticket

g pay

sta... ...d

#	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$_4$

a — register request
d — check ticket
c — examine casually
e — decide
g — pay compensation

a — register request
c — examine casually
d — check ticket
e — decide
g — pay compensation

a — register request
d — check ticket
c — examine casually
e — decide
h — reject request

a — register request
c — examine casually
d — check ticket
e — decide
h — reject request

start

end

• • •   (all 21 variants seen in the log)

a — register request
g — pay compensation

a — register request
h — reject request

a — register request
h — reject request

thoroughly   ticket

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)**

3 **ability to balance all conformance dimensions (fitness, precision, generalization, and simplicity) incl. noise**

4 **sound (result should at least be free of deadlocks, etc.)**

5 **provide guarantees (not just a best effort)**

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?

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?

BRIGHT
FUTURE
AHEAD

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

1965

How to get started?
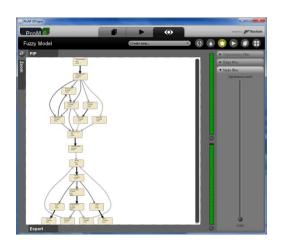
600+ plug-ins available covering the whole process mining spectrum
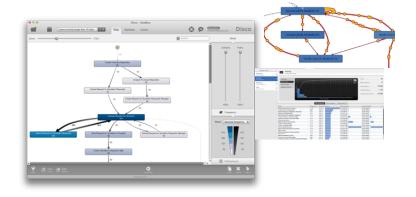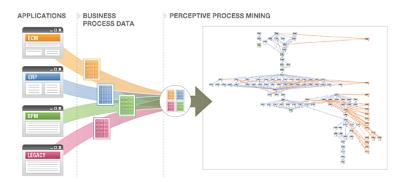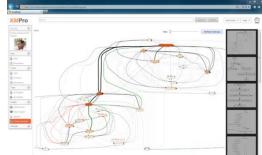

Download from: www.processmining.org

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)**
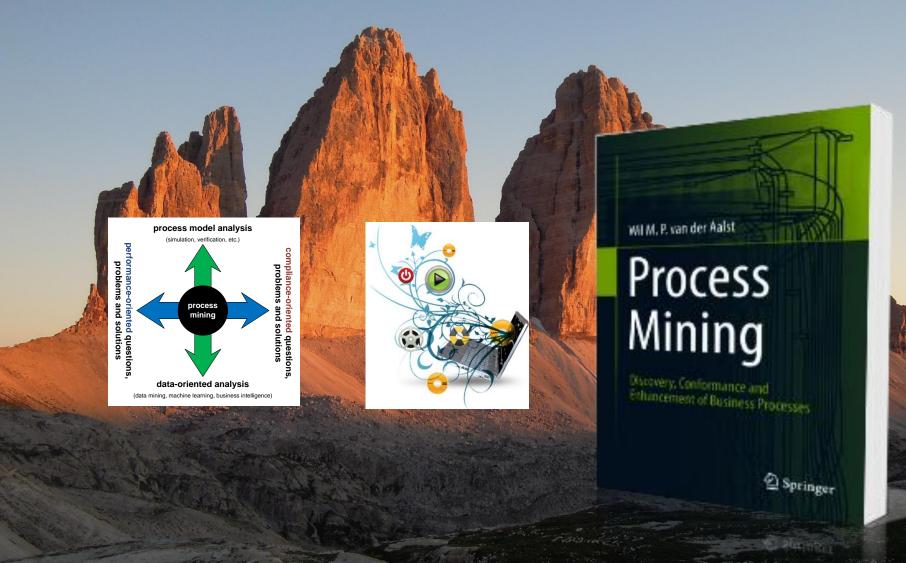- **…**

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.**

Conclusion

processmining.org