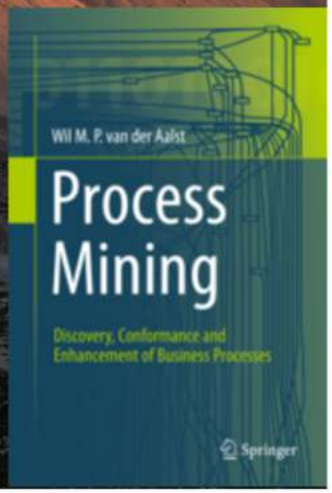
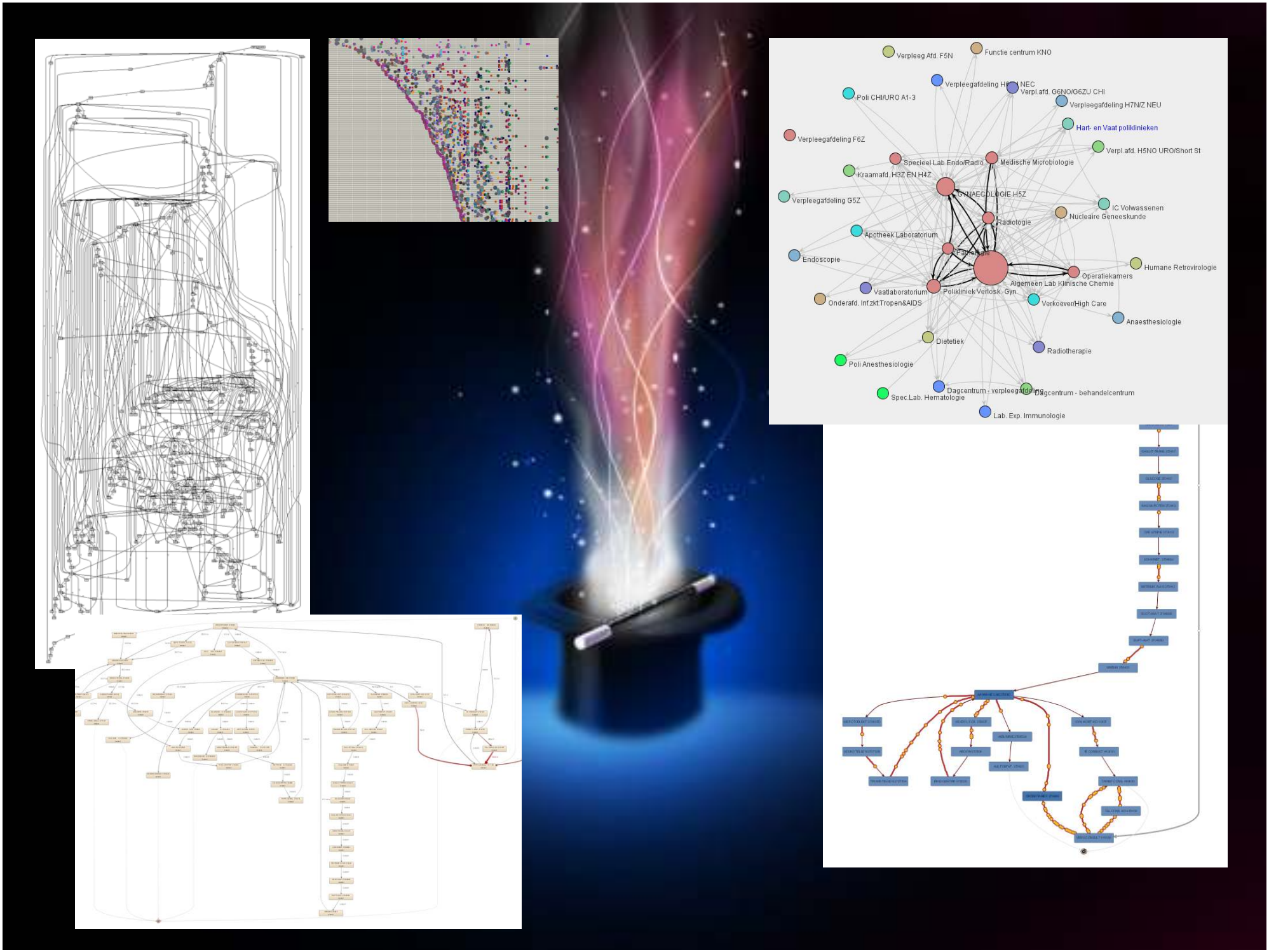
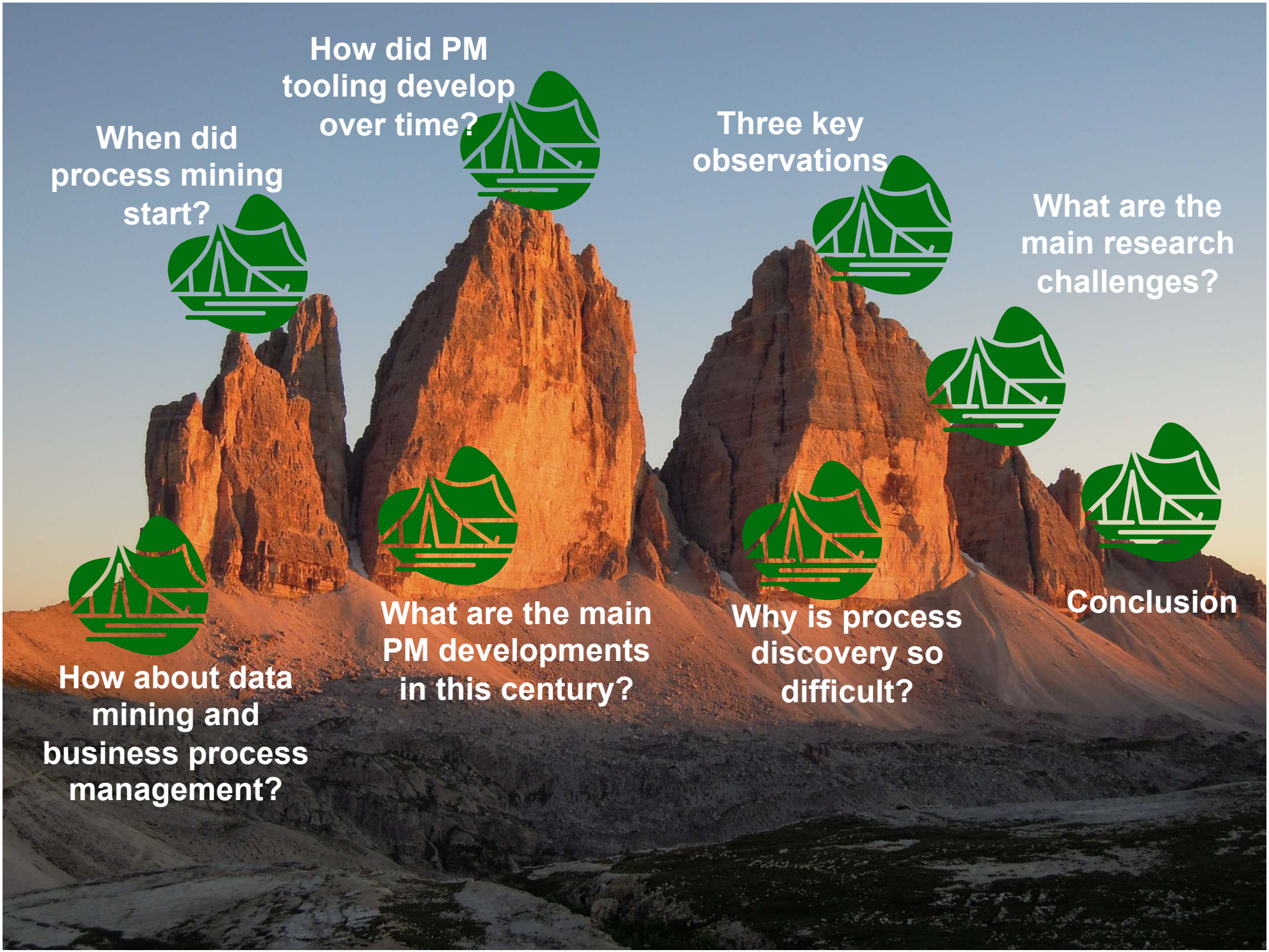
Process Mining:
A historical perspective

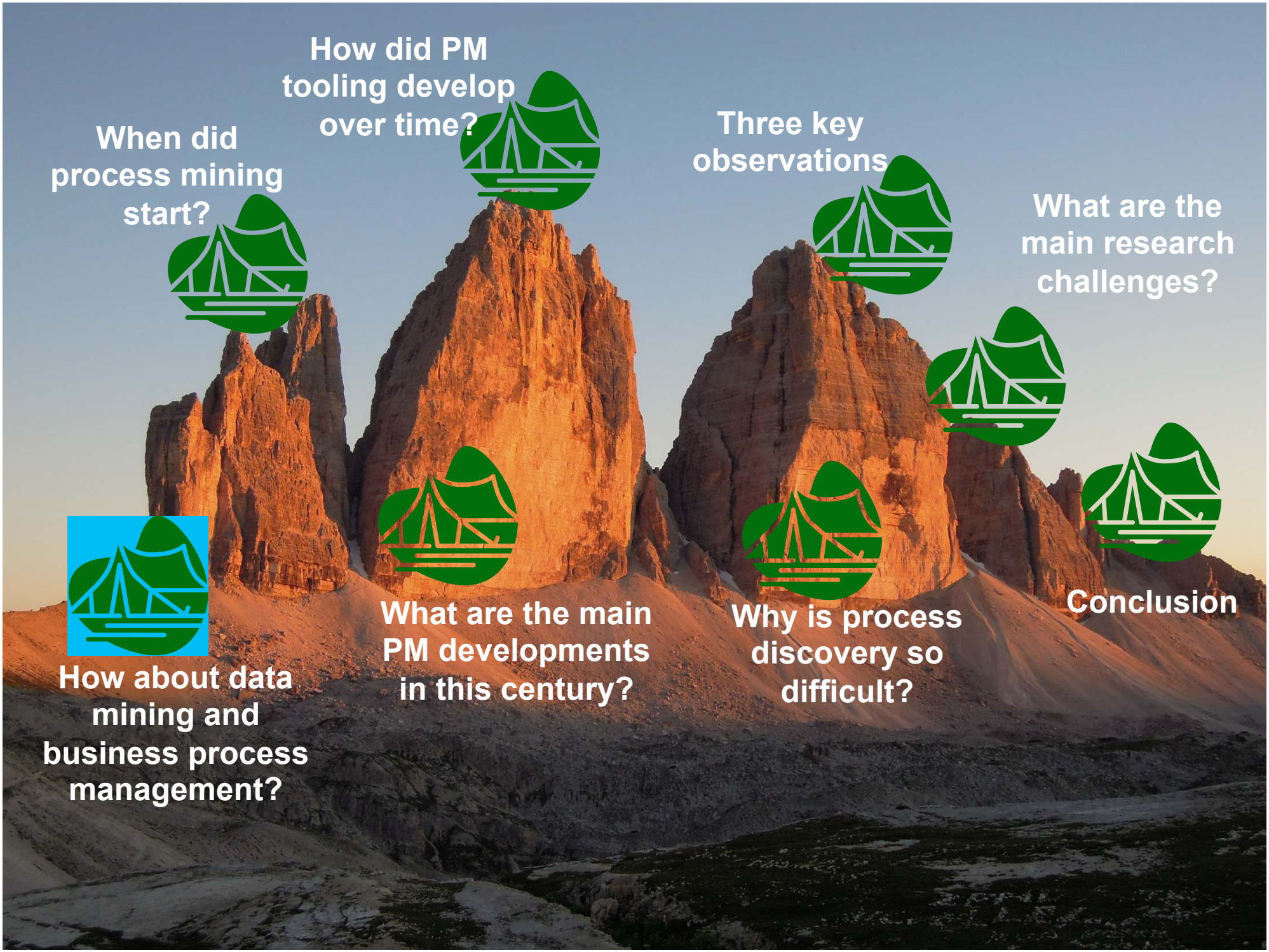
prof.dr.ir. Wil van der Aalst

2013
PROCESS MINING CAMP

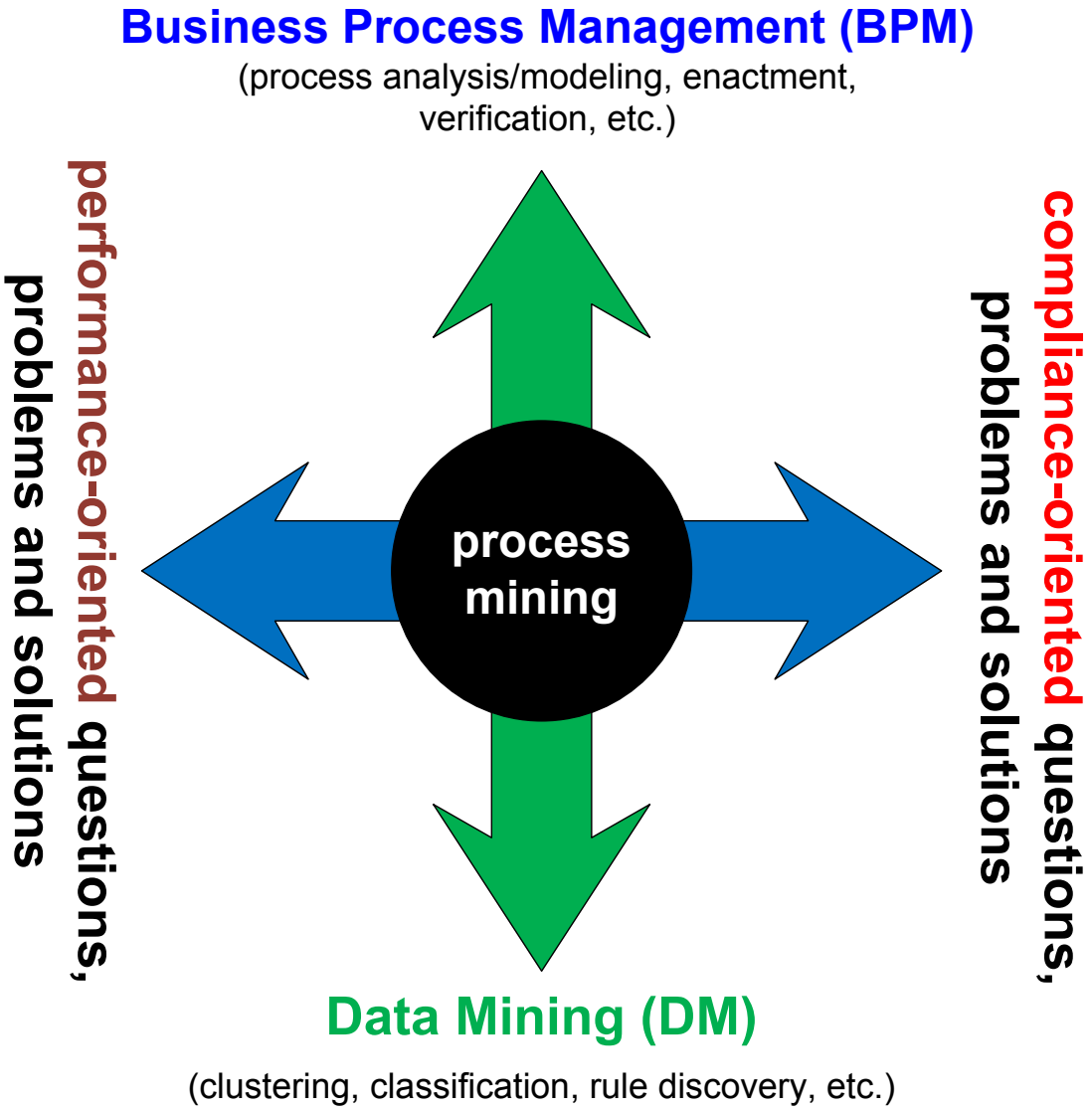



Verpleeg Afd. F5N
Functie centrum KNO
Verpleegafdeling H6N NEC
Verpl.afd. G6NO/G6ZU CHI
Poli CHI/URO A1-3
Verpleegafdeling H7N/Z NEU
Hart- en Vaat poliklinieken
Verpleegafdeling F6Z
Verpl.afd. H5NO URO/Short St
Specieel Lab Endo/Radio
Medische Microbiologie
Kraamafd. H3Z EN H4Z
GYN/AECOLOGIE H5Z
Verpleegafdeling G5Z
IC Volwassenen
Nucleaire Geneeskunde
Radiologie
Apotheek Laboratorium
Endoscopie
Humane Retrovirologie
Operatiekamers
Algemeen Lab Klinische Chemie
Vaatlaboratorium
Polikliniek Verlosk.-Gyn.
Onderafd. Inf.zkt:Tropen&AIDS
Verkoever/High Care
Anaesthesiologie
Dietetiek
Radiotherapie
Poli Anesthesiologie
Dagcentrum - verpleegafdeling
Dagcentrum - behandelcentrum
Spec.Lab. Hematologie
Lab. Exp. Immunologie

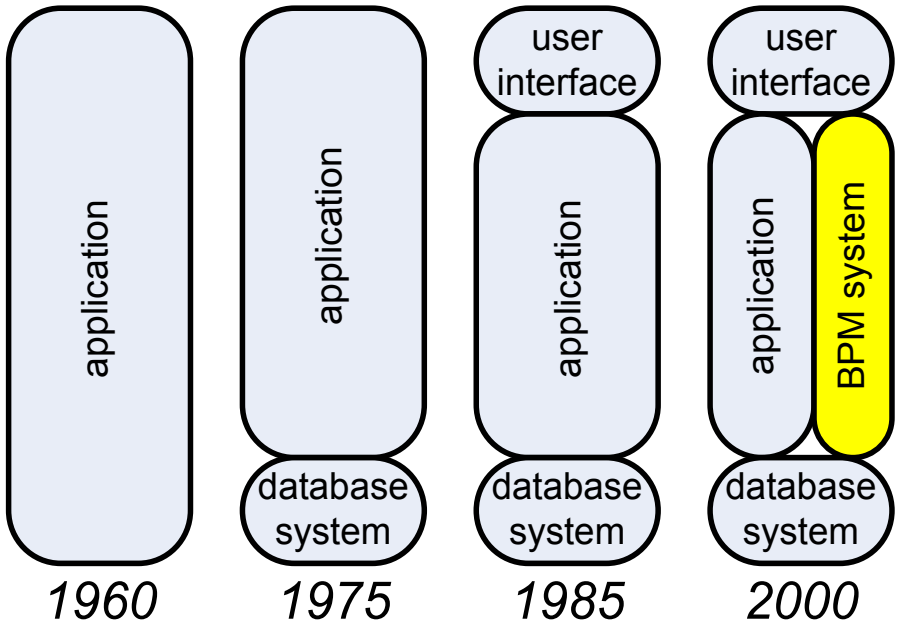

When did process mining start?
How did PM tooling develop over time?
Three key observations
What are the main research challenges?
How about data mining and business process management?
What are the main PM developments in this century?
Why is process discovery so difficult?
Conclusion

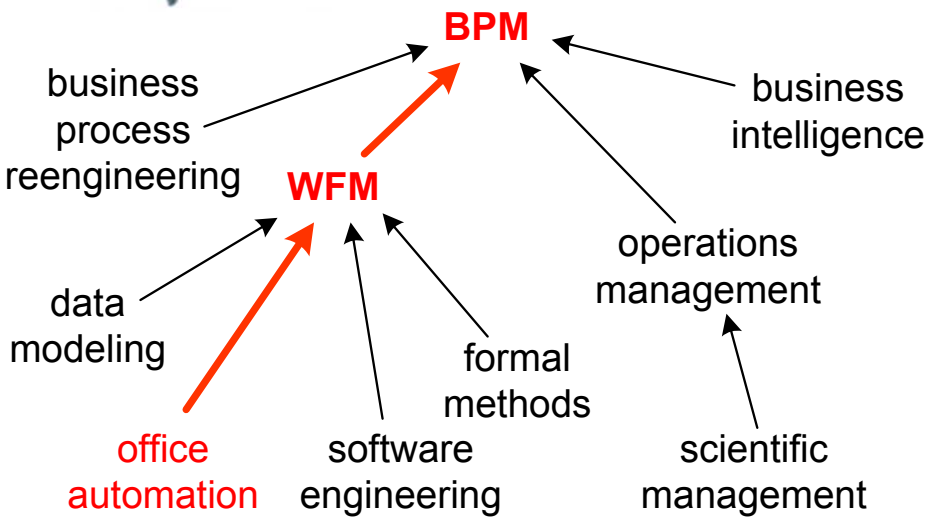

When did process mining start?
How did PM tooling develop over time?
Three key observations
What are the main research challenges?
How about data mining and business process management?
What are the main PM developments in this century?
Why is process discovery so difficult?
Conclusion

Positioning Process Mining



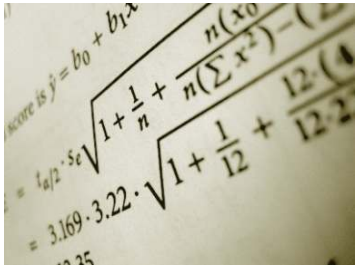
History and Origins of BPM



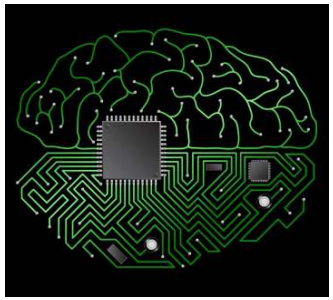


History and Origins of Data Mining

Classical statistics (since 500 BC): descriptive statistics (e.g., sample mean) statistical inference (e.g., confidence interval, regression, hypothesis testing).



data dredging, data fishing, data snooping



Artificial intelligence (since 1950): making intelligent machines by applying human-thought-like processing to statistical problems.

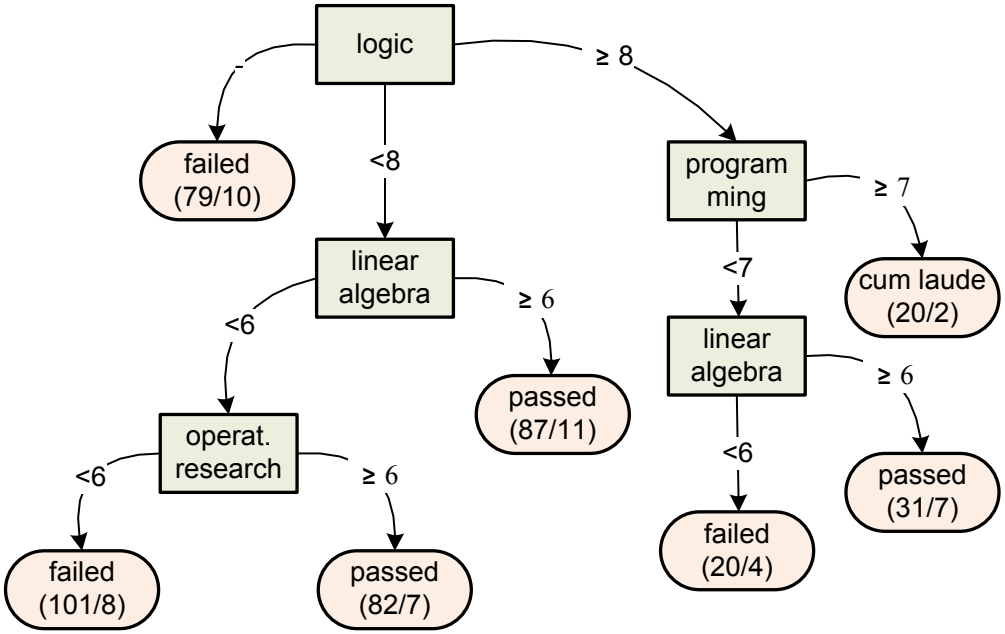
Machine learning (since 1950): construction and study of systems that can learn from data.

Many other terms: knowledge discovery, (predictive) analytics, …

Data Mining: Supervised Learning

- Labeled data, i.e., there is a response variable that labels each instance.
- Goal: explain response variable (dependent variable) in terms of predictor variables (independent variables).
- Classification techniques (e.g., decision tree learning) assume a categorical response variable and the goal is to classify instances based on the predictor variables.
- Regression techniques assume a numerical response variable. The goal is to find a function that fits the data with the least error.

Example: Decision tree learning



linear algebra	logic	program-ming	operations research	workflow systems	...	duration	result
9	8	8	9	9	...	36	cum laude
7	6	-	8	8	...	42	passed
-	-	5	4	6	...	54	failed
8	6	6	6	5	...	38	passed
6	7	6	-	8	...	39	passed
9	9	9	9	8	...	38	cum laude
5	5	-	6	6	...	52	failed
...	...	...	...	...	...	...	...

Unsupervised Learning

- Unsupervised learning assumes unlabeled data, i.e., the variables are not split into response and predictor variables.
- Examples: clustering (e.g., k-means clustering and agglomerative hierarchical clustering) and pattern discovery (association rules)

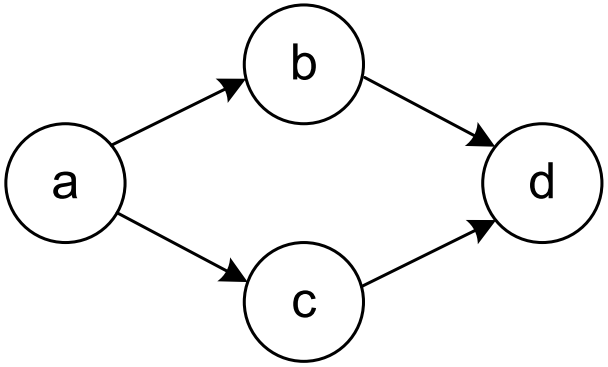
Example: Association rules

cappuccino	latte	espresso	americano	ristretto	tea	muffin	bagel
1	0	0	0	0	0	1	0
0	2	0	0	0	0	1	1
0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	0	0	1	2	0
0	0	0	1	1	0	0	0
…	…	…	…	…	…	…	…

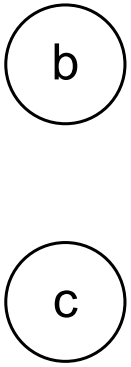
$$tea \wedge latte \Rightarrow muffin$$

$$tea \Rightarrow muffin \wedge bagel$$

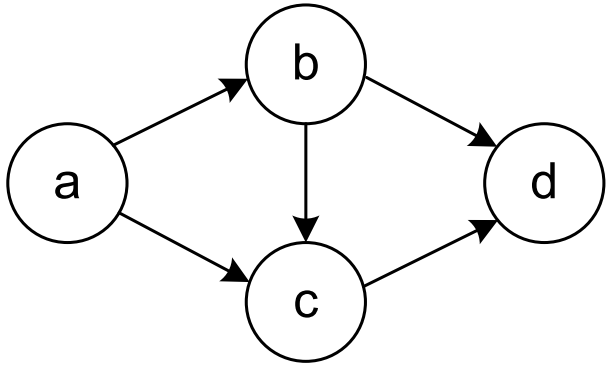
Example: Episode Mining



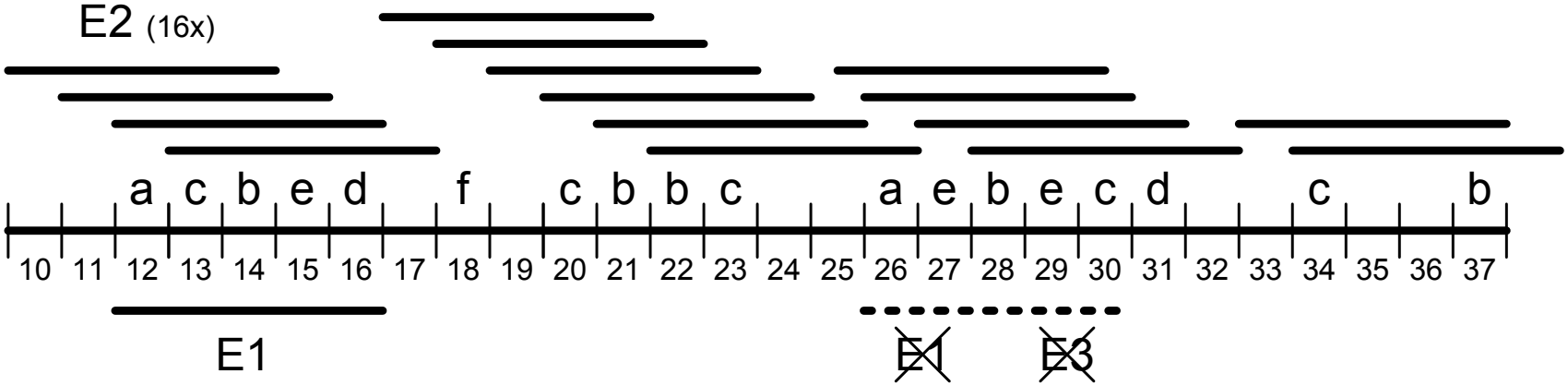
E1



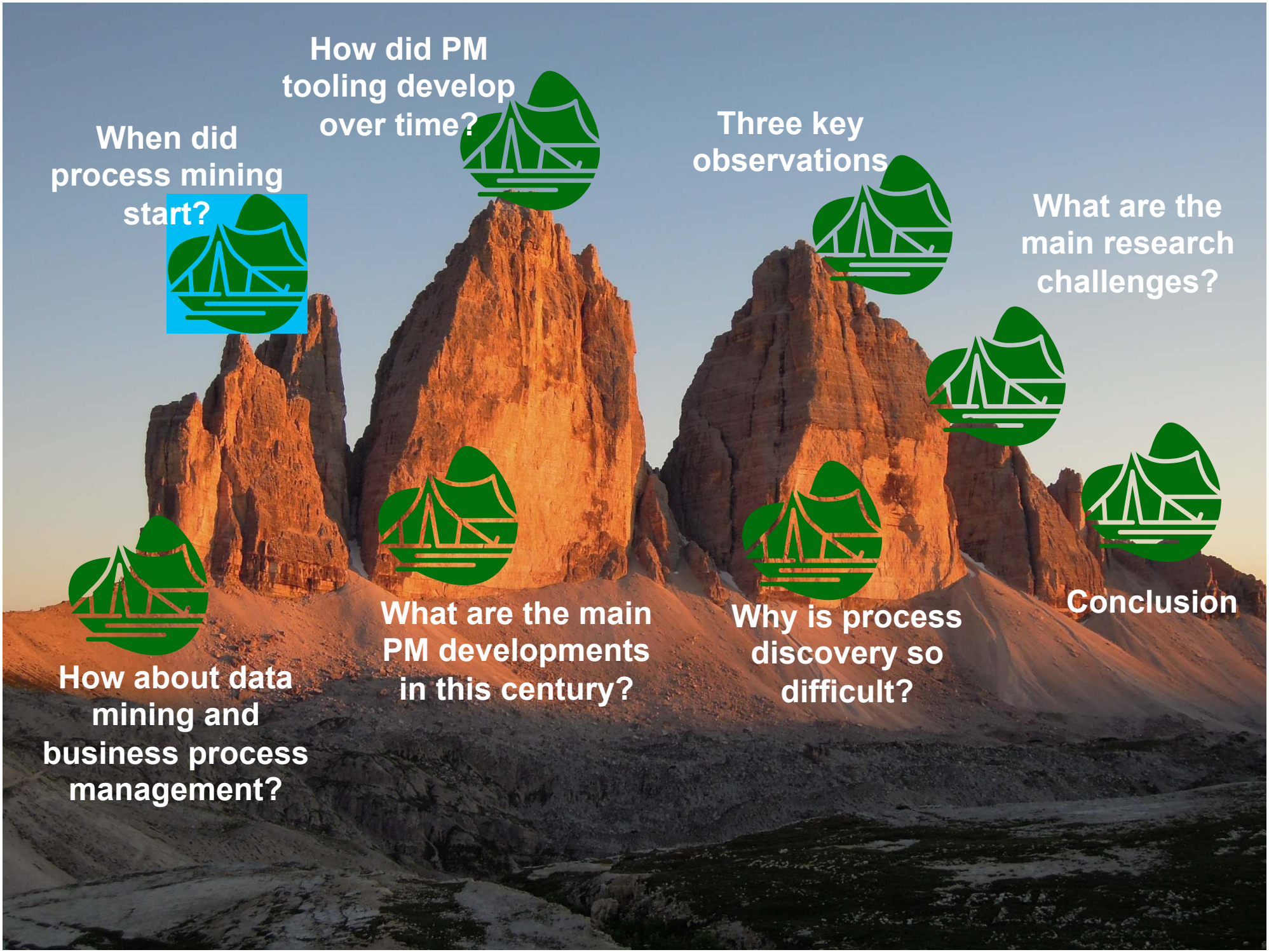
E2



E3



$E2 \Rightarrow E1$ has a confidence of $1/16$


When did process mining start?
How did PM tooling develop over time?
Three key observations
What are the main research challenges?
How about data mining and business process management?
What are the main PM developments in this century?
Why is process discovery so difficult?
Conclusion

Language identification in the limit (Mark Gold 1967)

- Mother uses sentences from some language {aab, ab, ab, abc, …}.
- "Perfect child" listens to mother and hypothesizes what the full language is like (given all sentences so far).
- Eventually the perfect child's hypothesis is correct and never changes again (without knowing), i.e., only finitely many wrong hypotheses are generated.
- A language is learnable in the limit if such a perfect child exists.

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

Language identification in the limit (E. Mark Gold 1967)

- Gold showed that most languages cannot be learned in the limit (including the most simple ones like regular languages ($ab^*(c|d)$).
- He noted that it matters whether the child gets positive and negative examples (corrections), whether the mother is evil, etc.
- Frequencies matter!
- Representational bias matters!

sentence ≅ trace in event log

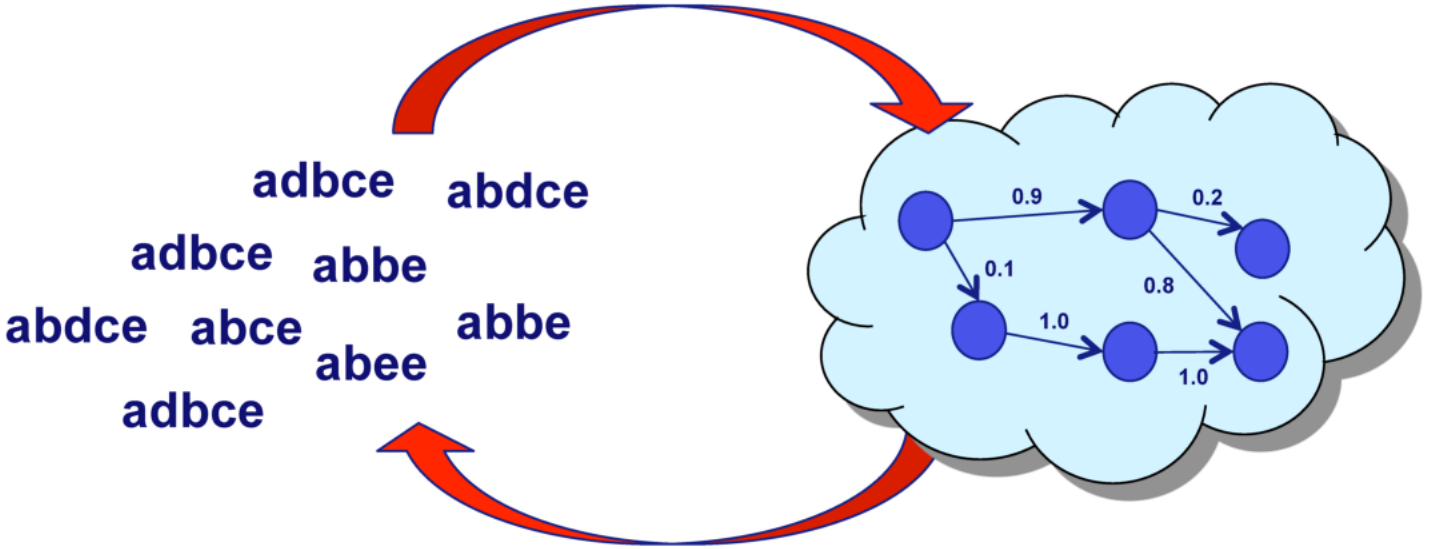
language ≅ process model

Myhill-Nerode Theorem (1958) and the Biermann/Feldman Algorithm (1972)

- There is a unique minimal deterministic finite automaton recognizing a regular language L (shown by John Myhill and Anil Nerode in 1958).
- The equivalence classes defined by $\cong$ determine the states of the automaton: $x \cong y$ if there is no z such that $xz \notin L$ and $yz \in L$.
- Cannot be applied to example traces: overfitting and no generalization.
- Alan W. Biermann and Jerome A. Feldman propose in 1972 techniques to learn finite state machines from examples (e.g., considering k-tails).

Nerode. Linear automaton transformations. Proc. Amer. Math. Soc. 9 1958 541-544.
Biermann and Feldman. On the synthesis of finite-state machines from samples of their behaviour. IEEE Transactions on Computers, 21:592–597, 1972.

Baum–Welch (1970) and Viterbi (1967) Algorithms to learn Hidden Markov Models



- The Viterbi algorithm finds the most likely sequence of hidden states – called the Viterbi path – that results in a sequence of observed events (Andrew Viterbi, 1967).
- The Baum–Welch algorithm is an expectation-maximization algorithm that constructs a HMM (Leonard E. Baum and Lloyd R. Welch, 1970).

Where/when did process mining start?

- Myhill/Nerode (1958)?
- Gold (19)?
- Ba /Welch (19)?
- Bierman dman (197 ?
- Ra Agrawal (1)?
  - Apriori algorith quent patterns, l ded to sequences, episodes, …
- Jonath ook and Alex der Wolf (1998)?
  - vering Models of Software Processes from Event-Based Data"
  - using techniques similar to Biermann/Feldman (k-tails) and Baum/Welch (Markov models)
- Rakesh Agrawal, Dimitrios Gun los, Frank Leymann?
  - "Mining Process Models from Workflo 998)
  - Flowmark process models witho ering type of splits and joins, no loops, etc.
- Anindya Datta (19
  - Automating the y of AS-IS Business Process Models
  - Biermann/Feldm style work, embedded in BPM

no concurrency

unable to handle noise

unable to handle incompleteness

no end-to-end models

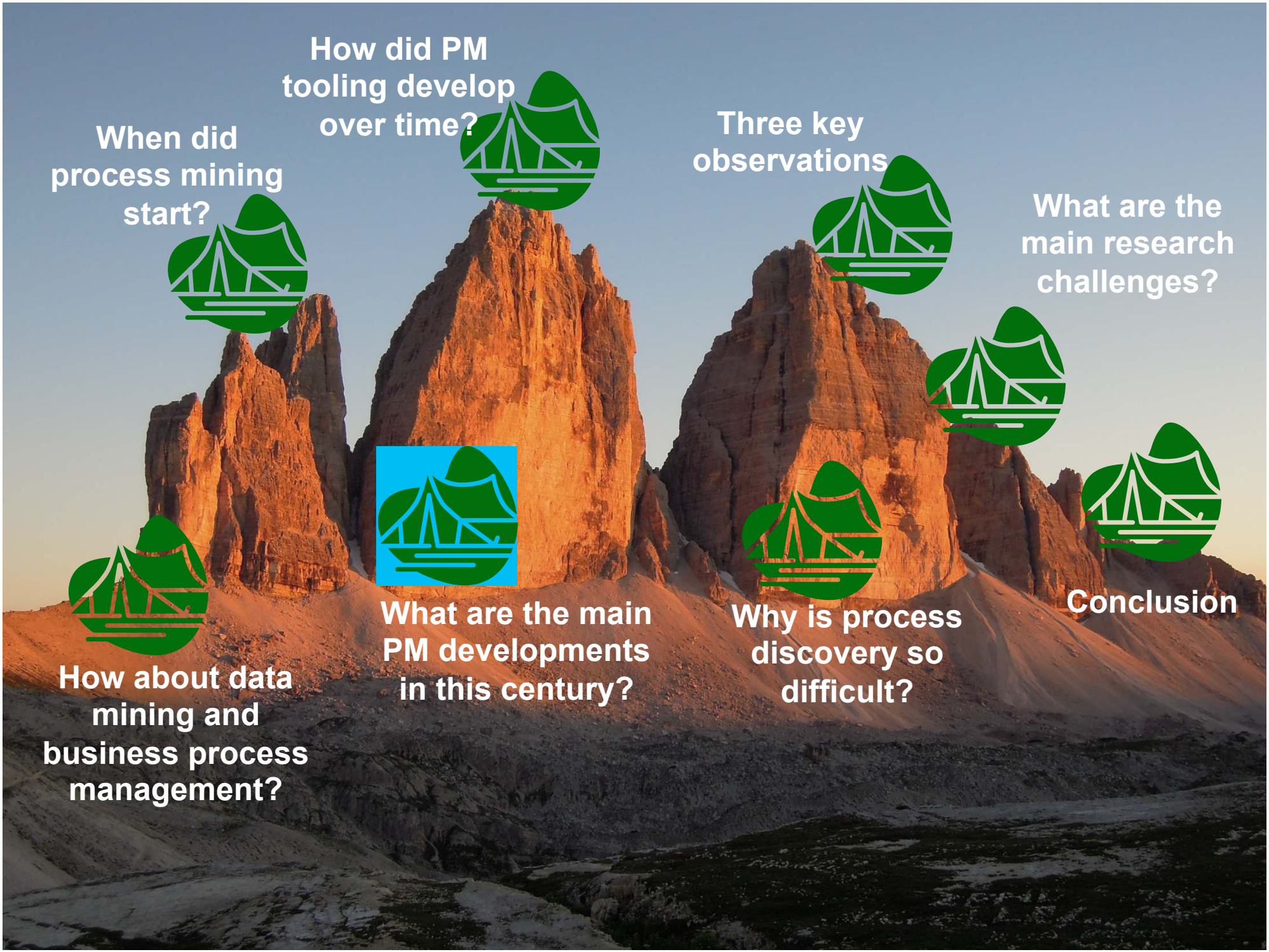
informal, no precise semantics

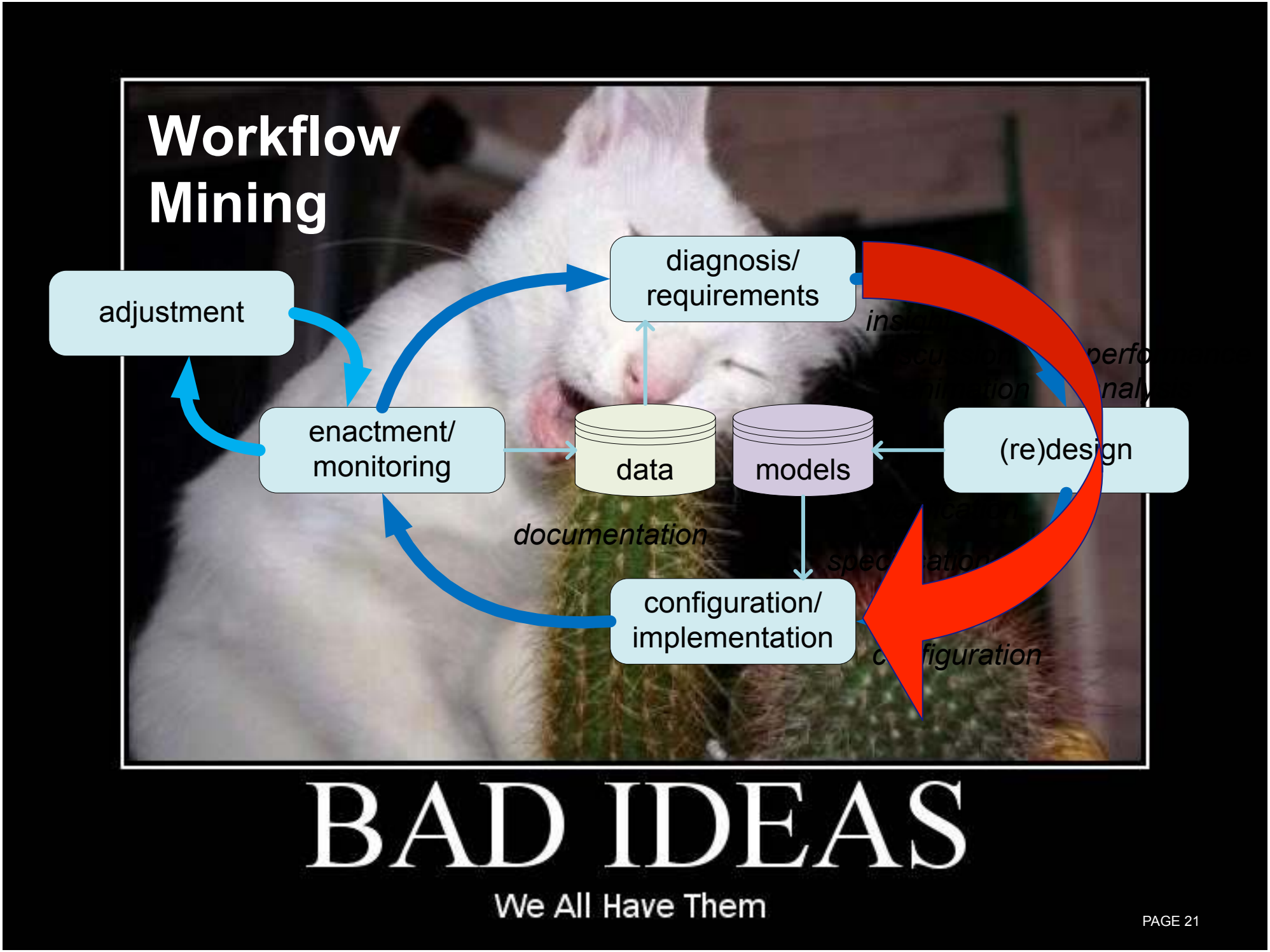
How did process mining start at TU/e?

- Paper and research proposal: "Process Design by Discovery: Harvesting Workflow Knowledge from Ad-hoc Executions" (1999)
  - Upcoming move to Technology Management department to lead the IS group (working at CU-Boulder at the time).
  - Collaboration with Ton Weijters stimulated by BETA (linking Petri nets and workflow to Ton's expertise in machine learning).
- First PhDs on process mining (many followed):
  - Laura Maruster
  - Ana Karla Alves de Medeiros
  - Boudewijn van Dongen
- Initial work on alpha algorithm (formal limits) and heuristic and genetic mining (dealing with noise).

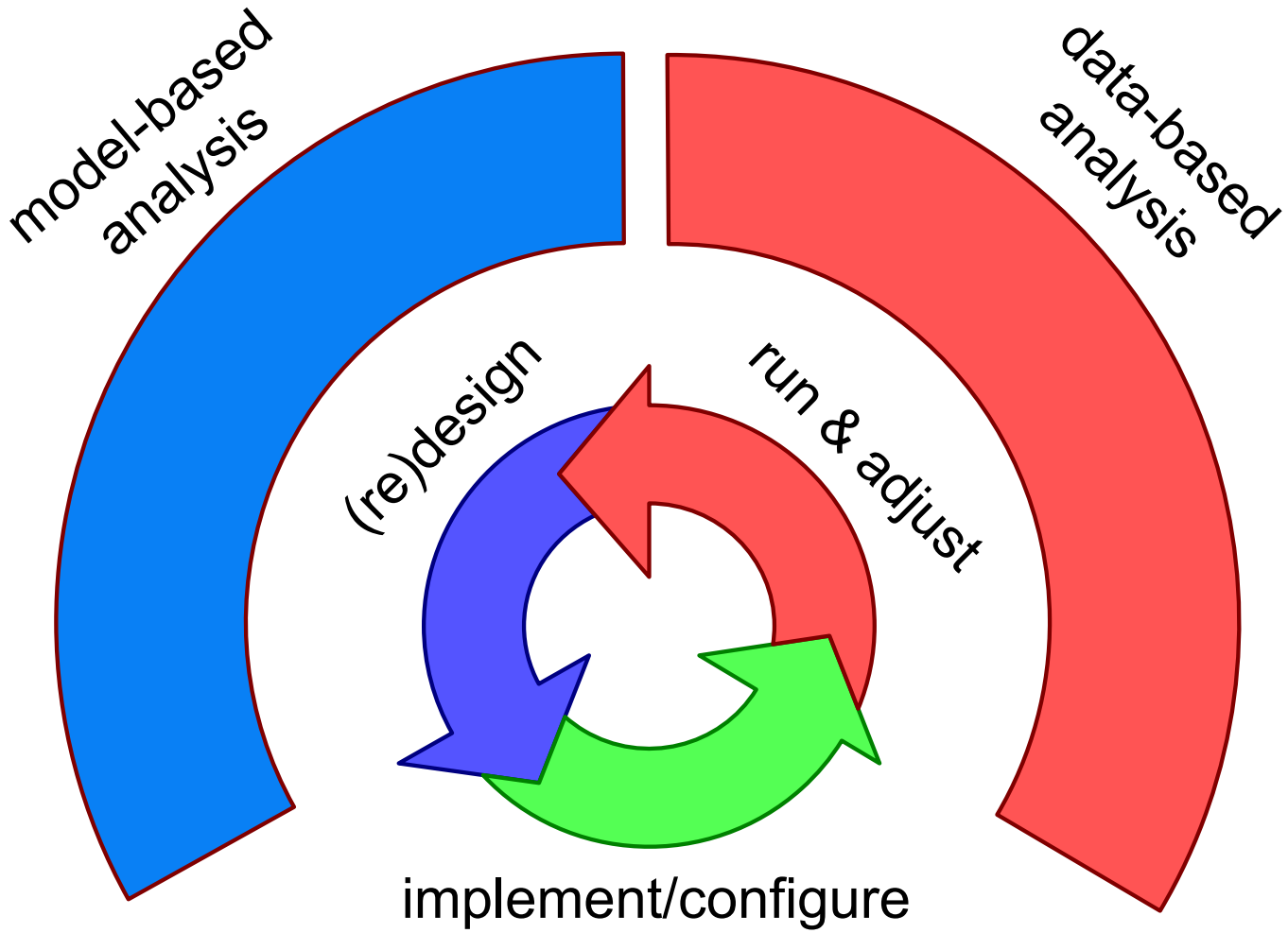
TU/e

Initial team

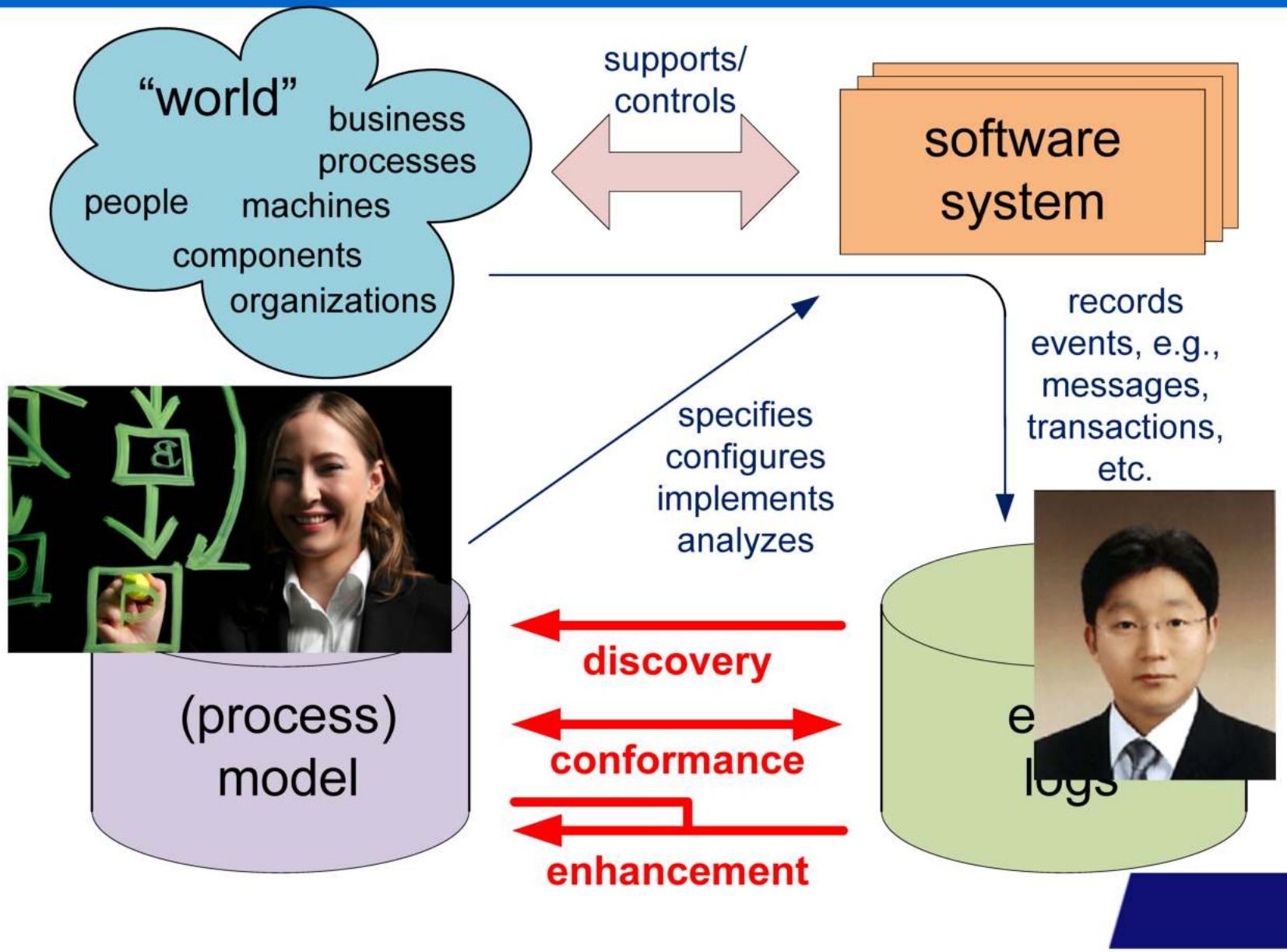

When did process mining start?
How did PM tooling develop over time?
Three key observations
What are the main research challenges?
How about data mining and business process management?
What are the main PM developments in this century?
Why is process discovery so difficult?
Conclusion


Workflow Mining
adjustment
diagnosis/ requirements
enactment/ monitoring
data
models
(re)design
documentation
configuration/ implementation
BAD IDEAS
We All Have Them

Models, data, and systems coexist



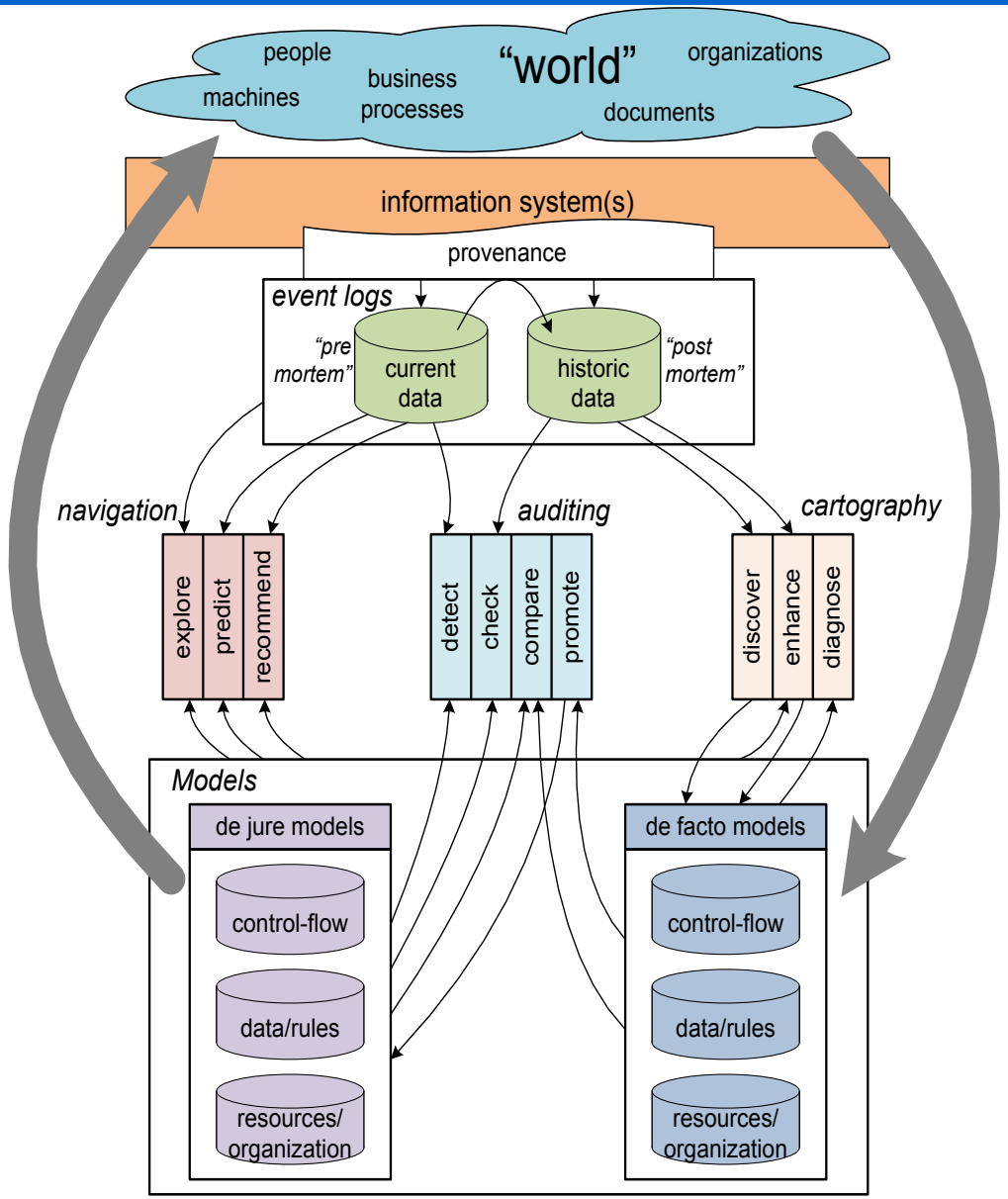
Process mining spectrum in 2007: Beyond control-flow discovery

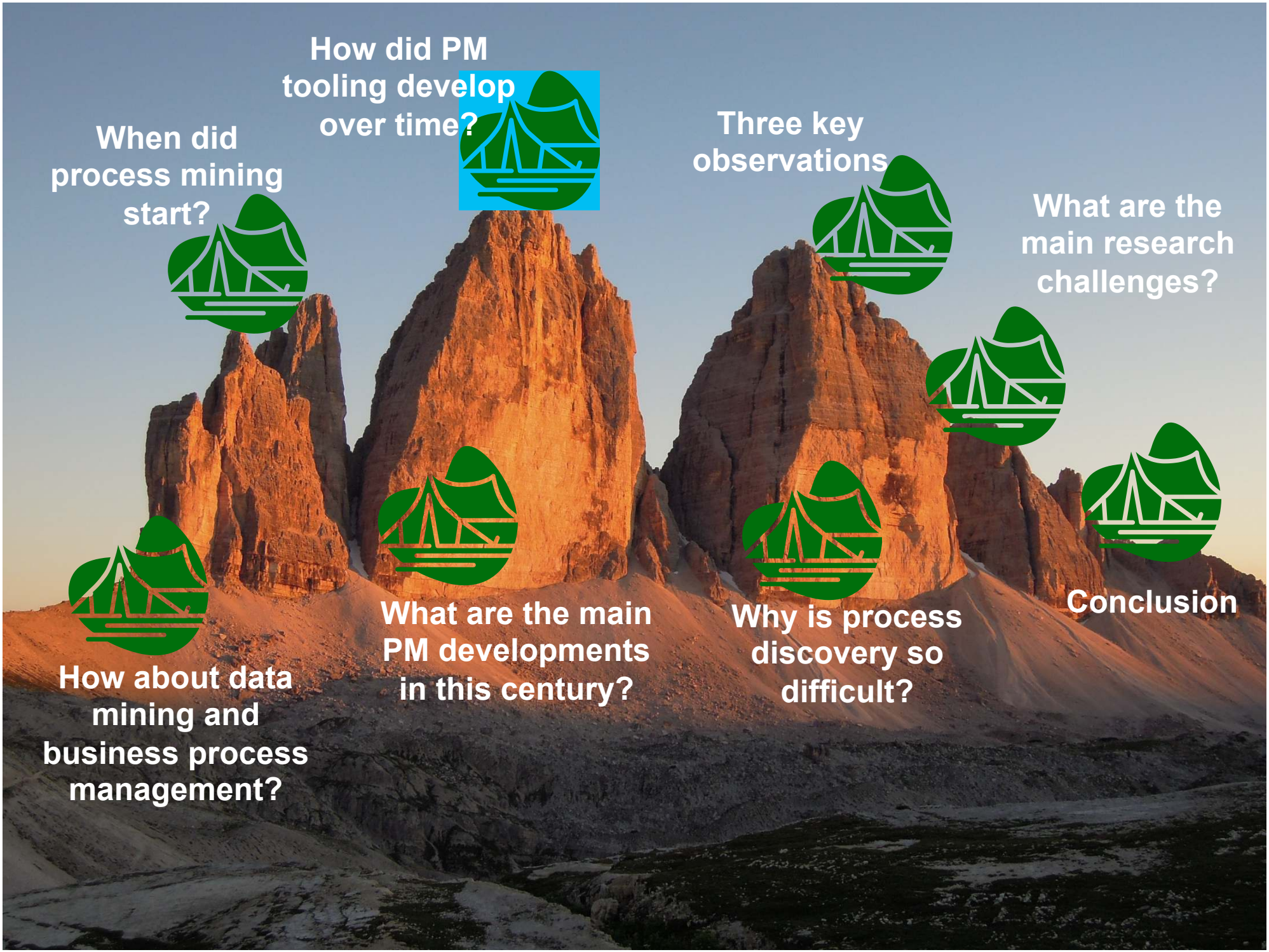


Team in November 2007

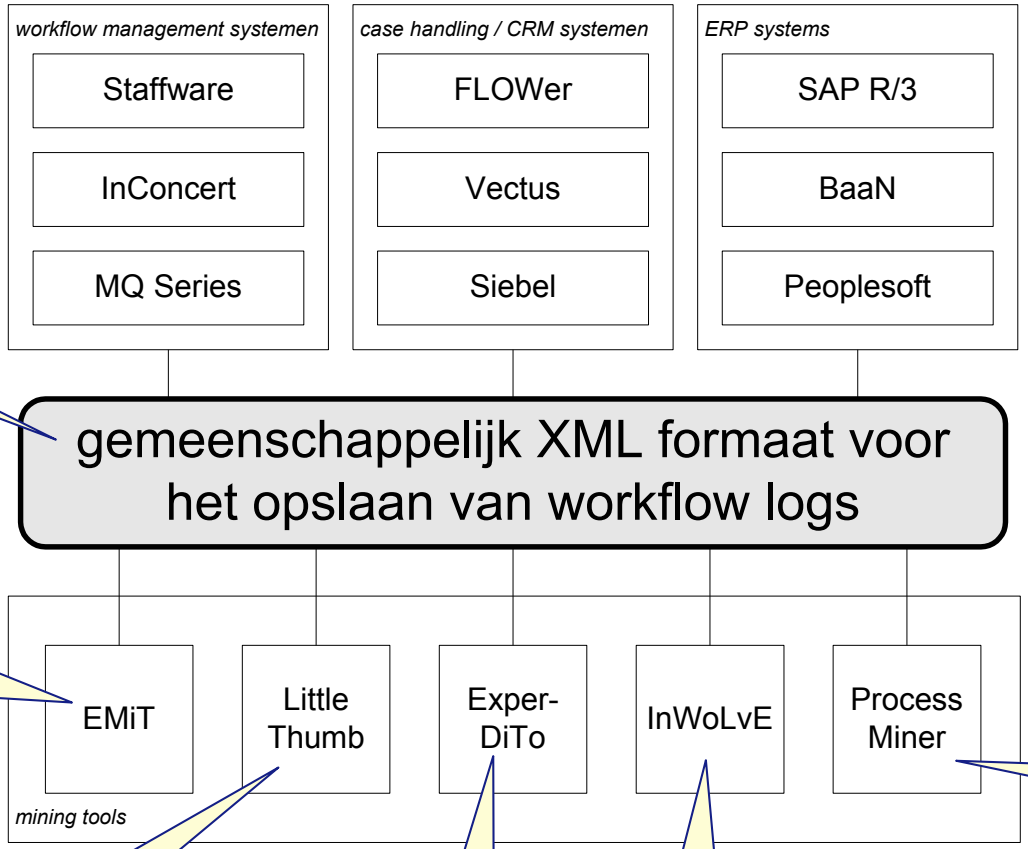
Some people are missing, e.g., Peter van den Brand.

Current process mining spectrum
(including alignments, operational support, and multiple perspectives)




When did process mining start?
How did PM tooling develop over time?
Three key observations
What are the main research challenges?
How about data mining and business process management?
What are the main PM developments in this century?
Why is process discovery so difficult?
Conclusion

Pre-ProM
(figure from March 2002!)



predecessor of MXML format

alpha algorithm including time analysis (BvD)

predecessor of ProM's heuristic miner (TW)

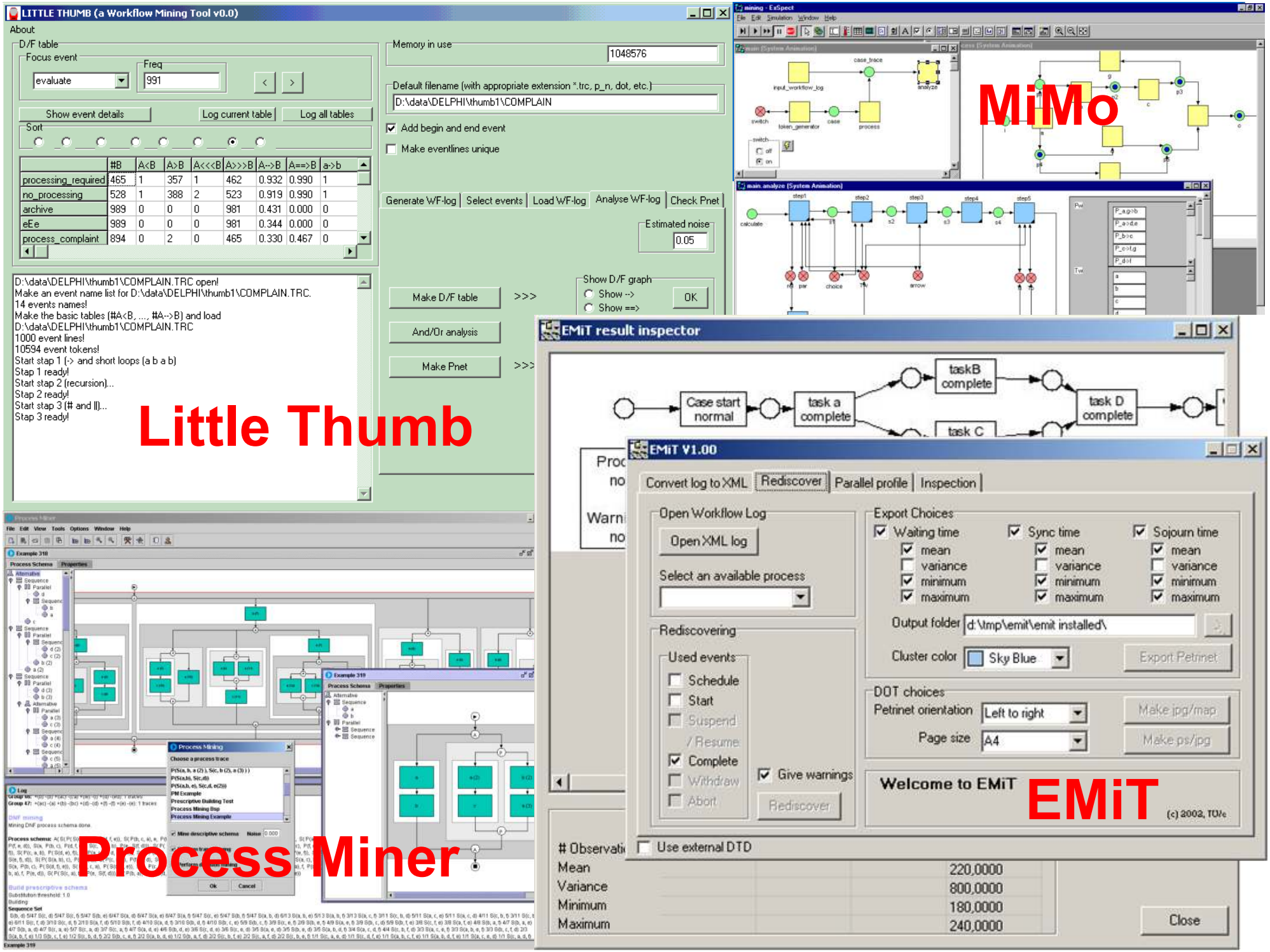
evaluation tool (Laura Maruster)

mining with duplicate tasks (Joachim Herbst)

mining block structured models (Guido Schimm)

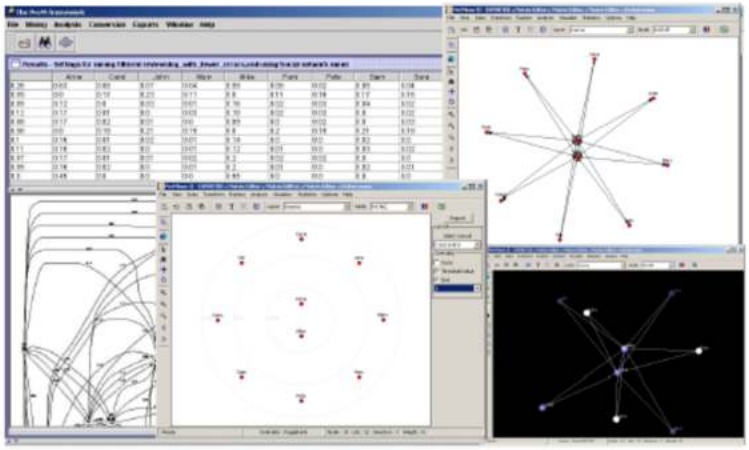
Tobias Blickle (ARIS PPM)

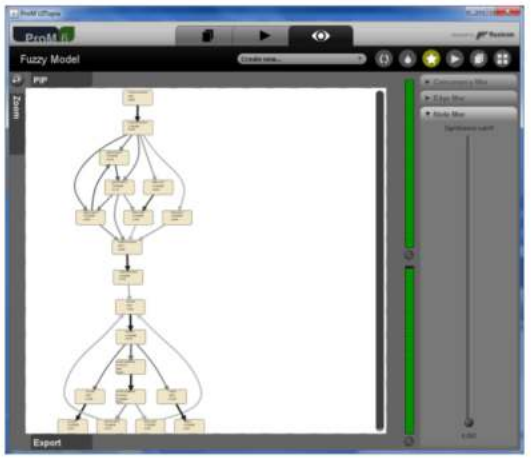
The first tool to support the alpha algorithm for process mining was the MiMo (Mining Module) tool based on ExSpect. Later it was implemented in EMiT and ProM.


LITTLE THUMB (a Workflow Mining Tool v0.0)
About
D/F table
Focus event
evaluate
Freq
991
Show event details
Log current table
Log all tables
Sort
#B A<B A>B A<<<B A>>>B A-->B A==>B a>b
processing_required 465 1 357 1 462 0.932 0.990 1
no_processing 528 1 388 2 523 0.919 0.990 1
archive 989 0 0 0 981 0.431 0.000 0
eEe 989 0 0 0 981 0.344 0.000 0
process_complaint 894 0 2 0 465 0.330 0.467 0
D:\data\DELPHI\thumb1\COMPLAIN.TRC open!
Make an event name list for D:\data\DELPHI\thumb1\COMPLAIN.TRC.
14 events names!
Make the basic tables (#A<B, ..., #A-->B) and load
D:\data\DELPHI\thumb1\COMPLAIN.TRC
1000 event lines!
10594 event tokens!
Start stap 1 (-> and short loops (a b a b)
Stap 1 ready!
Start stap 2 (recursion)...
Stap 2 ready!
Start stap 3 (# and ||)...
Stap 3 ready!
Memory in use
1048576
Default filename (with appropriate extension *.trc, p_n, dot, etc.)
D:\data\DELPHI\thumb1\COMPLAIN
Add begin and end event
Make eventlines unique
Generate WF-log
Select events
Load WF-log
Analyse WF-log
Check Pnet
Estimated noise
0.05
Make D/F table
>>>
Show D/F graph
Show -->
Show ==>
OK
And/Or analysis
Make Pnet
>>>
Little Thumb
MiMo
EMiT result inspector
Case start normal
task a complete
taskB complete
task C
task D complete
EMiT V1.00
Convert log to XML
Rediscover
Parallel profile
Inspection
Open Workflow Log
Open XML log
Select an available process
Rediscovering
Used events
Schedule
Start
Suspend
/ Resume
Complete
Withdraw
Abort
Give warnings
Rediscover
Export Choices
Waiting time
Sync time
Sojourn time
mean
variance
minimum
maximum
Output folder
d:\tmp\emit\emit installed\
Cluster color
Sky Blue
Export Petrinet
DOT choices
Petrinet orientation
Left to right
Page size
A4
Make jpg/map
Make ps/jpg
Welcome to EMiT
(c) 2002, TU/e
EMiT
Use external DTD
Mean 220,0000
Variance 800,0000
Minimum 180,0000
Maximum 240,0000
Close
Process Miner

ProM (2004 – now)

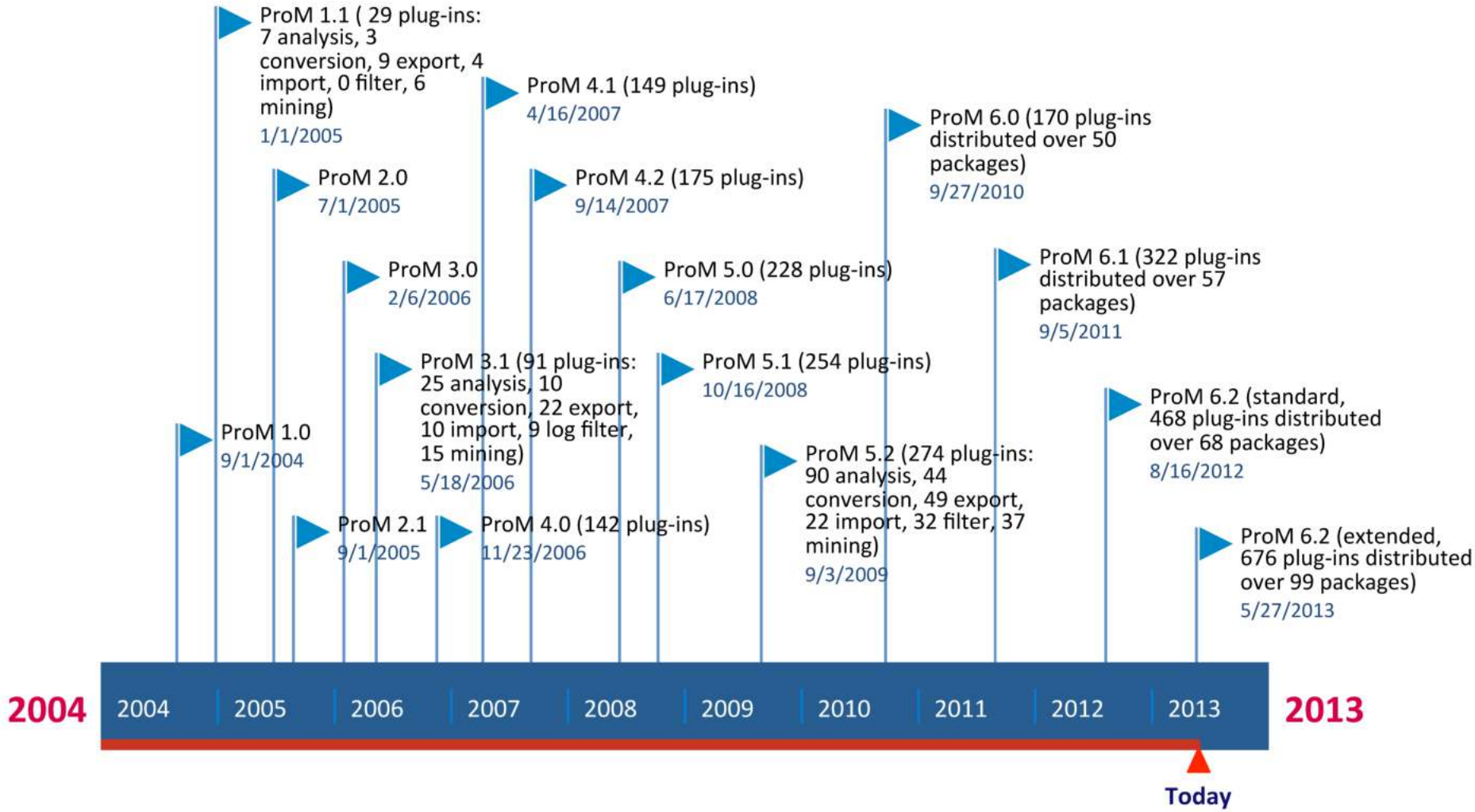
ProM
process mining workbench



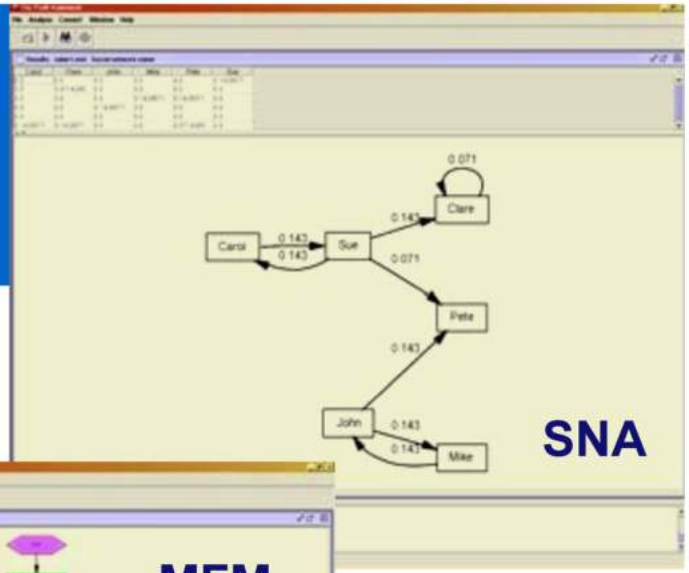


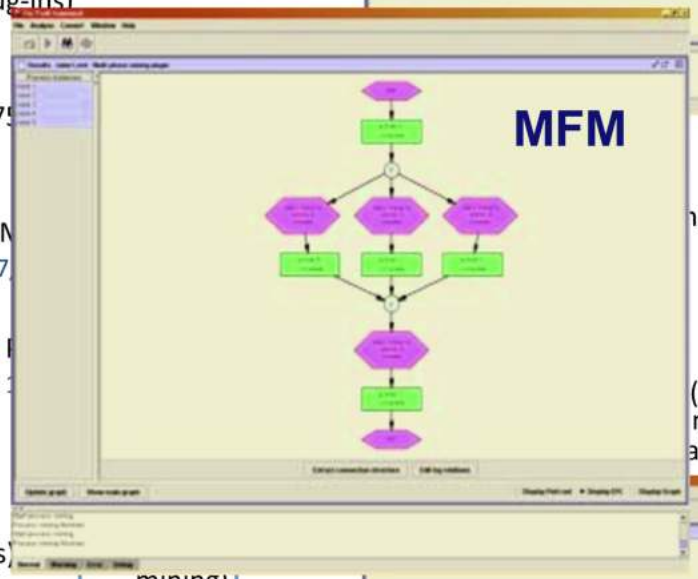
See www.processmining.org

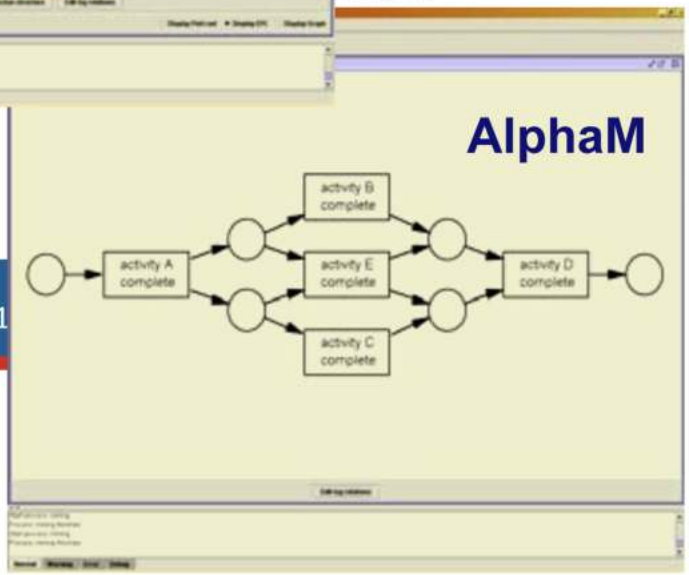
Overview of ProM releases



ProM 1.1







ProM 1.1 (29 plug-ins: 7 analysis, 3 conversion, 9 export, 4 import, 0 filter, 6 mining)
1/1/2005

ProM 1.0
9/1/2004

ProM 2.0
7/1/2005

ProM 3.0
2/6/2006

ProM 3.1 (91 plug-ins: 25 analysis, 10 conversion, 22 export, 10 import, 9 log filter, 15 mining)
5/18/2006

ProM 2.1
9/1/2005

ProM 4.0 (142 plug-ins)
11/23/2006

ProM 4.1 (149 plug-ins)
4/16/2007

ProM 4.2 (175
9/14/2007

ProM
6/17/

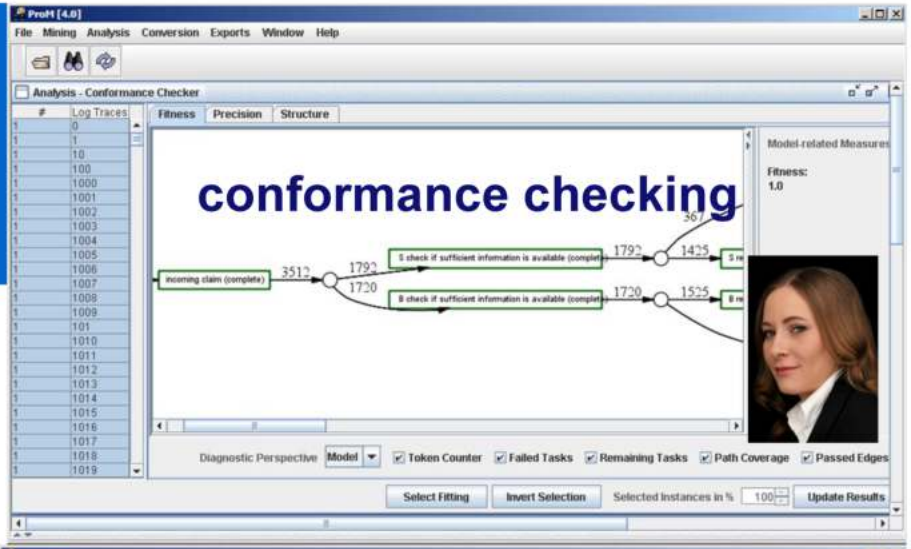
mining)
9/3/2009

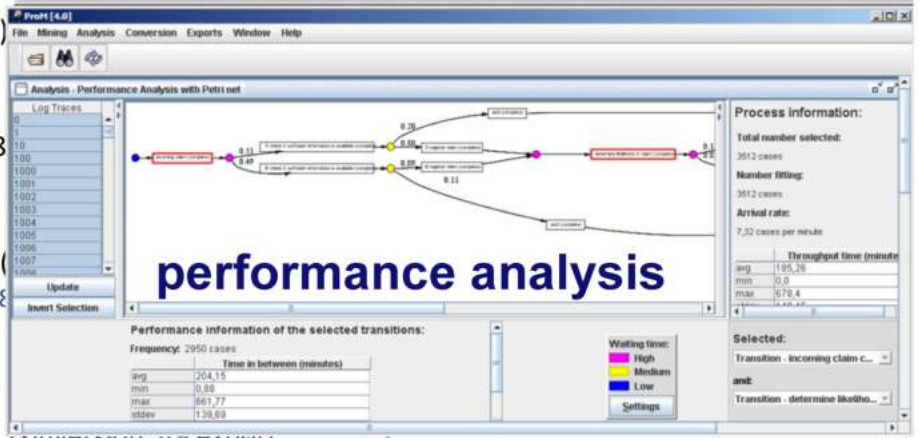
ns

(standard,
ns distributed
ackages)

2004 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011

ProM 4.0





ProM 1.1 (29 plug-ins: 7 analysis, 3 conversion, 9 export, 4 import, 0 filter, 6 mining)
1/1/2005

ProM 4.1 (149 plug-ins)
4/16/2007

ProM 2.0
7/1/2005

ProM 4.2 (175 plug-ins)
9/14/2007

ProM 3.0
2/6/2006

ProM 5.0 (228
6/17/2008

ProM 3.1 (91 plug-ins: 25 analysis, 10 conversion, 22 export, 10 import, 9 log filter, 15 mining)
5/18/2006

ProM 5.1 (
10/16/2008

ProM 1.0
9/1/2004

ProM 2.1
9/1/2005

ProM 4.0 (142 plug-ins)
11/23/2006

conversion, 49 export, 22 import, 32 filter, 37 mining)
9/3/2009

ProM 6.2 (extended, 676 plug-ins distributed over 99 packages)
5/27/2013

2004 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2013

Today

ProM 5.2

ProM 1.0
9/1/2004

ProM 1.1 (29 plug-ins: 7 analysis, 3 conversion, 9 export, 4 import, 0 filter, 6 mining)
1/1/2005

ProM 2.0
7/1/2005

ProM 2.1
9/1/2005

ProM 3.0
2/6/2006

ProM 3.1 (91 plug-ins: 25 analysis, 10 conversion, 22 export, 10 import, 9 log filter, 15 mining)
5/18/2006

ProM 4.0 (142 plug-ins)
11/23/2006

ProM 4.1 (149 plu
4/16/2007

ProM 4.2 (17
9/14/2007

ProM 5.0 (228 plug-ins)
6/17/2008

ProM 5.1 (254 plug-ins)
10/16/2008

ProM 5.2 (274 plug-ins: 90 analysis, 44 conversion, 49 export, 22 import, 32 filter, 37 mining)
9/3/2009

ProM 6.1 (322 plug-ins distributed over 57 packages)
9/5/2011

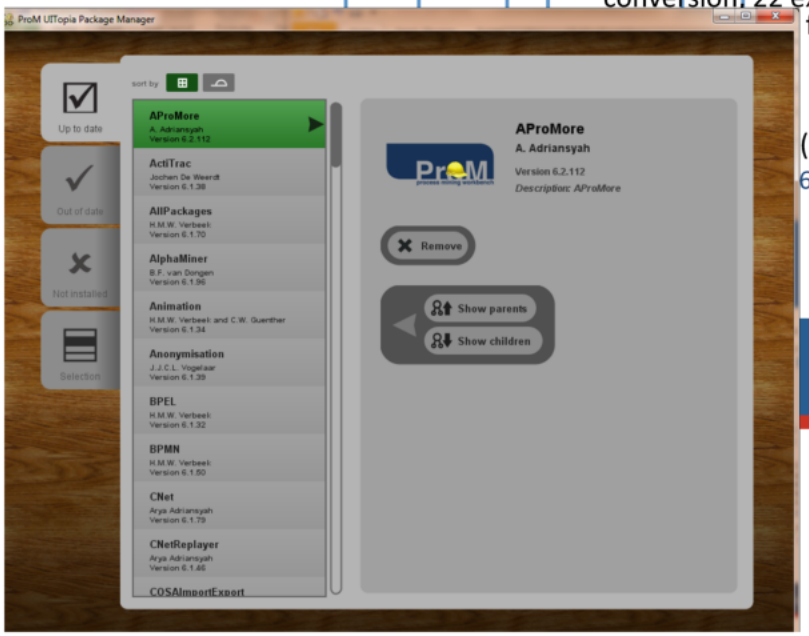
ProM 6.2 (standard, 468 plug-ins distributed over 68 packages)
8/16/2012

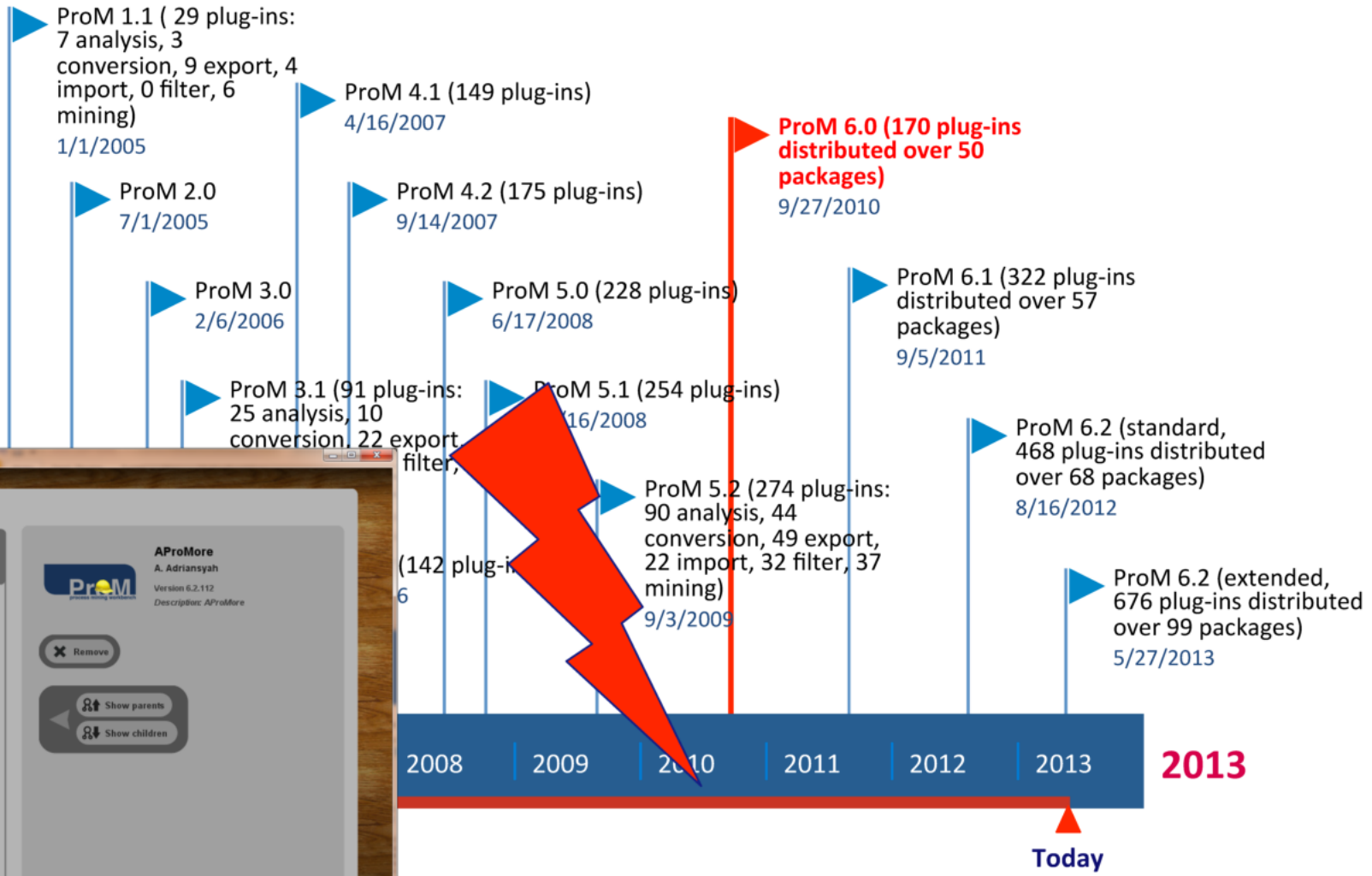
ProM 6.2 (extended, 676 plug-ins distributed over 99 packages)
5/27/2013

2004 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2013

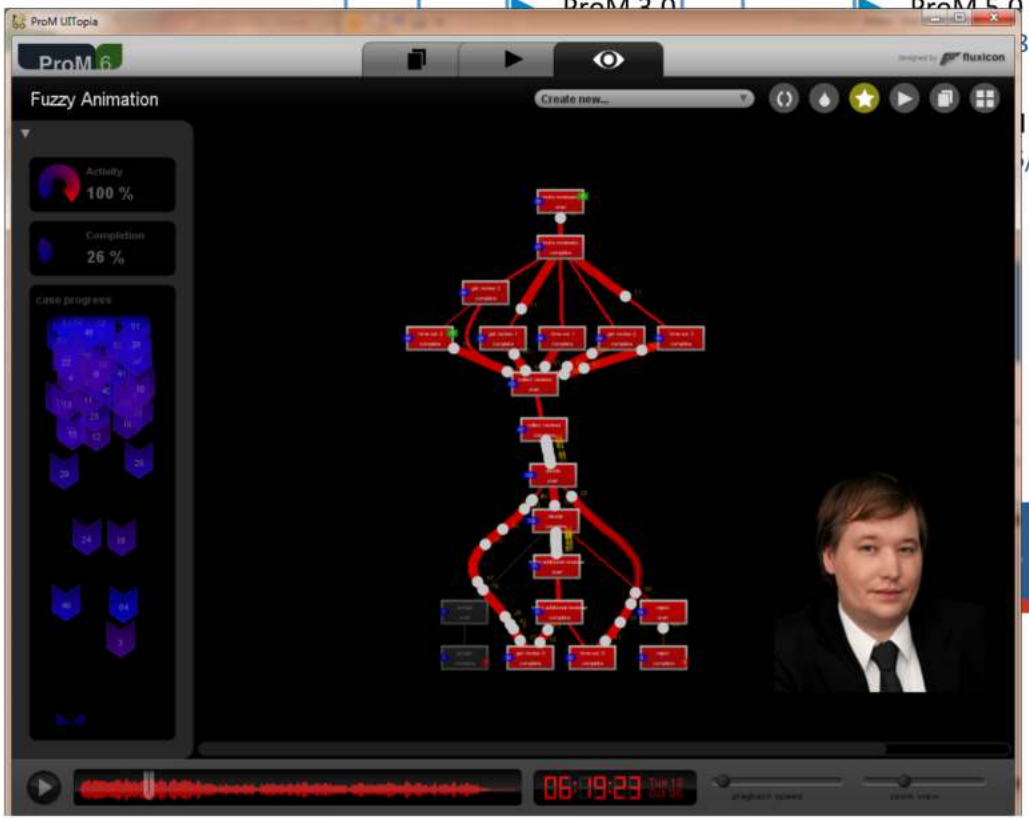
Today

ProM 6.0: A new start …





ProM Today



ProM 1.1 (29 plug-ins: 7 analysis, 3 conversion, 9 export, 4 import, 0 filter, 6 mining)
1/1/2005

ProM 2.0
7/1/2005

ProM 4.1 (149 plug-ins)
4/16/2007

ProM 4.2 (175 plug-ins)
9/14/2007

ProM 5.1 (254 plug-ins)
/2008

ProM 5.2 (274 plug-ins: 90 analysis, 44 conversion, 49 export, 22 import, 32 filter, 37 mining)
9/3/2009

ProM 6.0 (170 plug-ins distributed over 50 packages)
9/27/2010

ProM 6.1 (322 plug-ins distributed over 57 packages)
9/5/2011

ProM 6.2 (standard, 468 plug-ins distributed over 68 packages)
8/16/2012

ProM 6.2 (extended, 676 plug-ins distributed over 99 packages)
5/27/2013

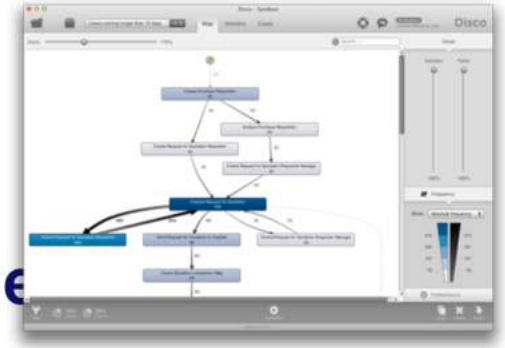
2010 | 2011 | 2012 | 2013

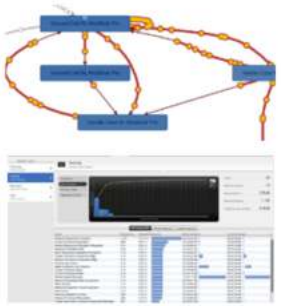
2013

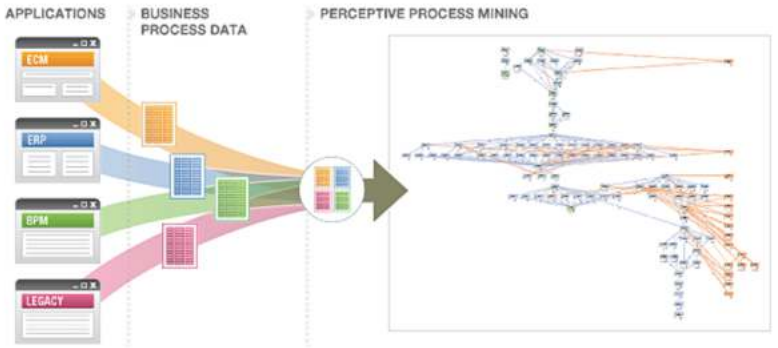
Today

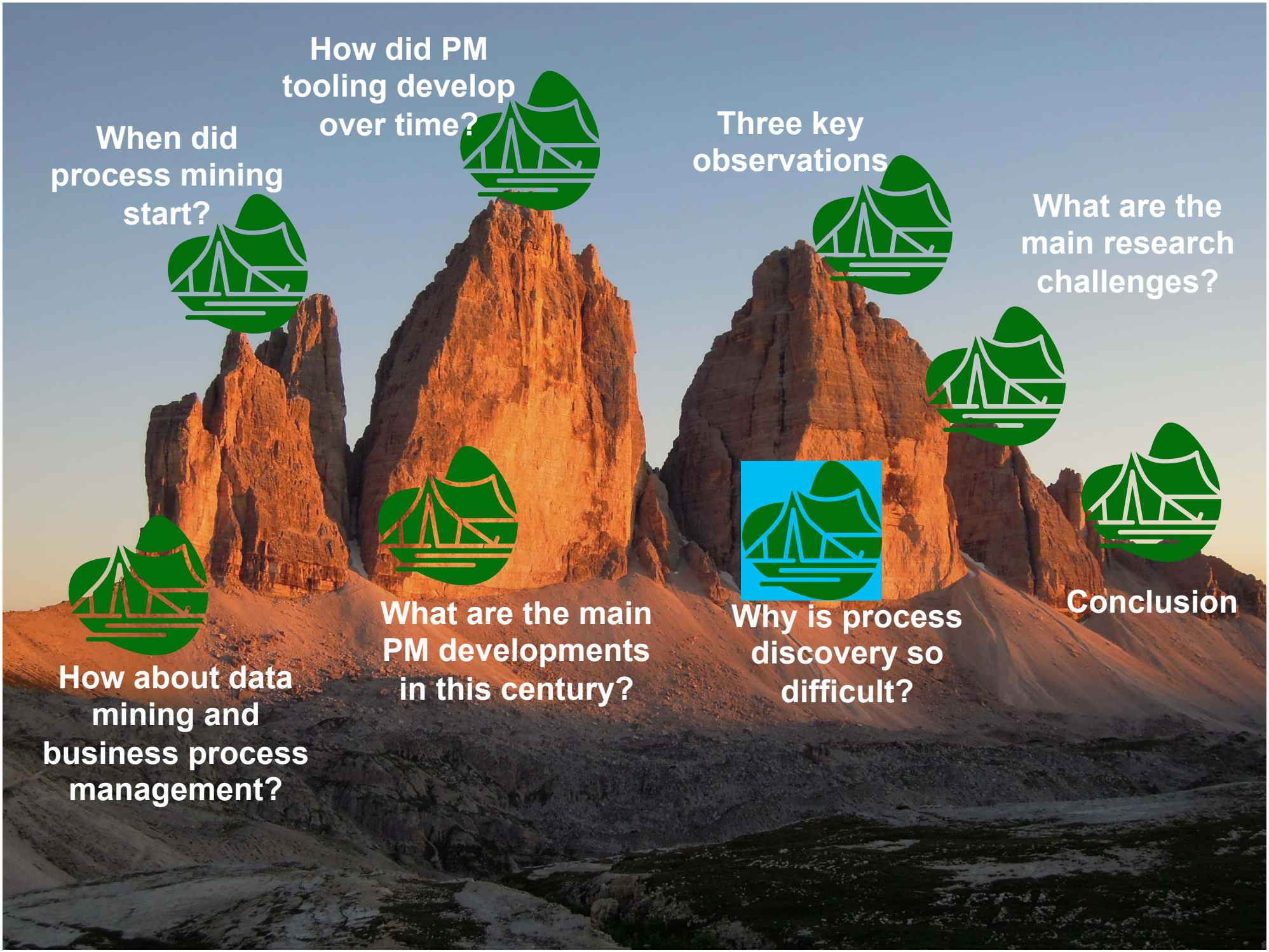
Commercial PM tools

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

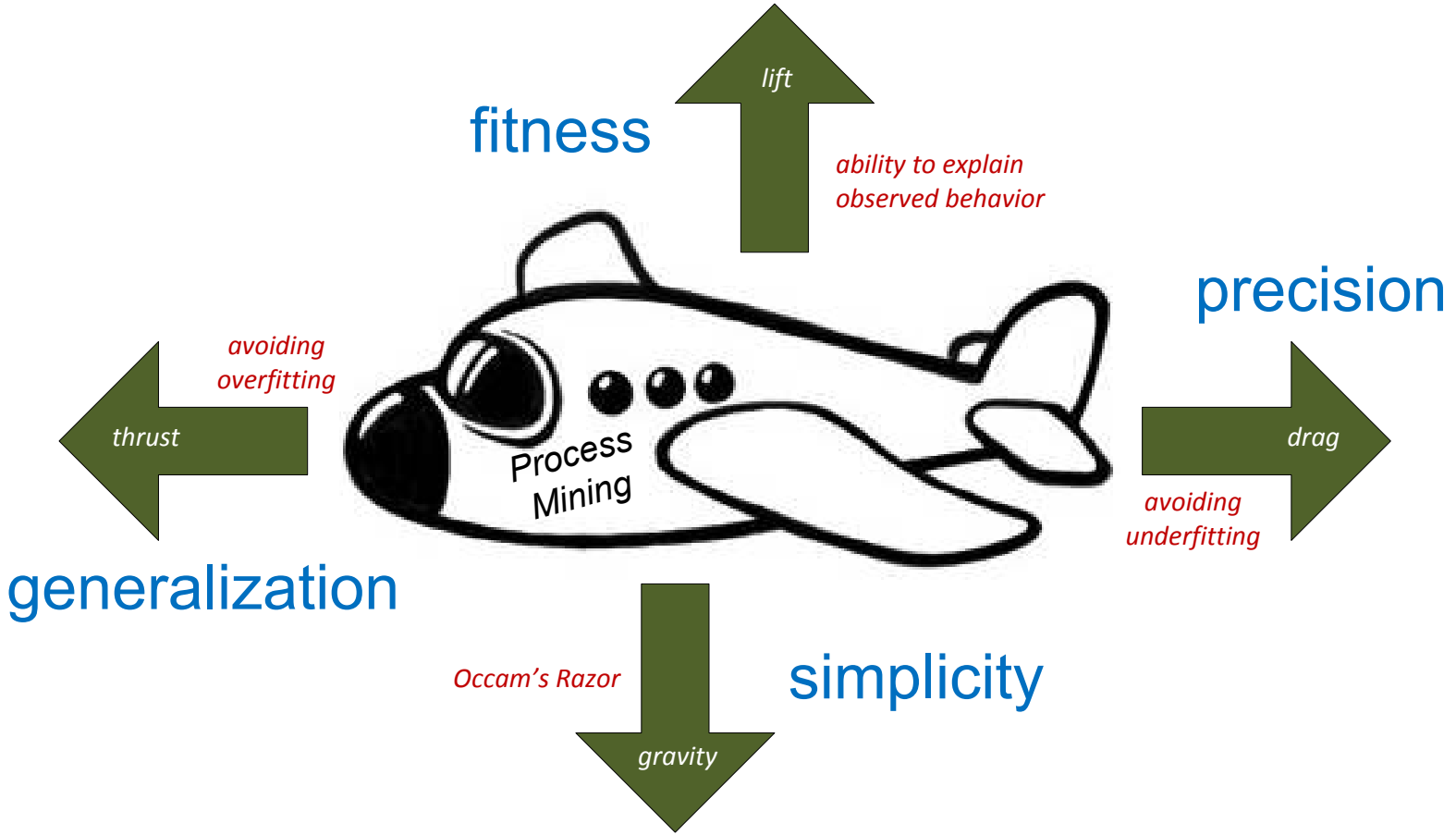






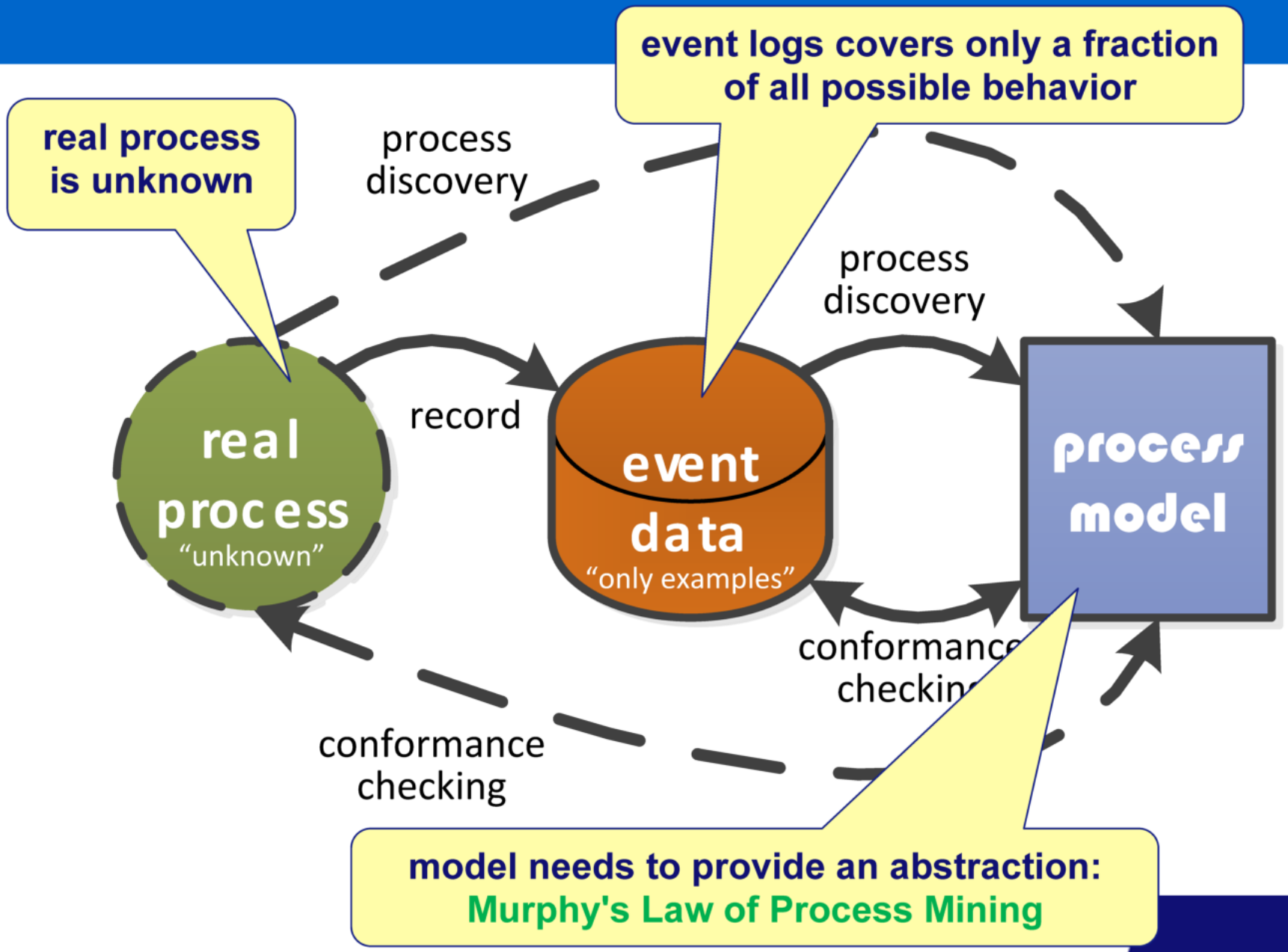

When did process mining start?
How did PM tooling develop over time?
Three key observations
What are the main research challenges?
How about data mining and business process management?
What are the main PM developments in this century?
Why is process discovery so difficult?
Conclusion

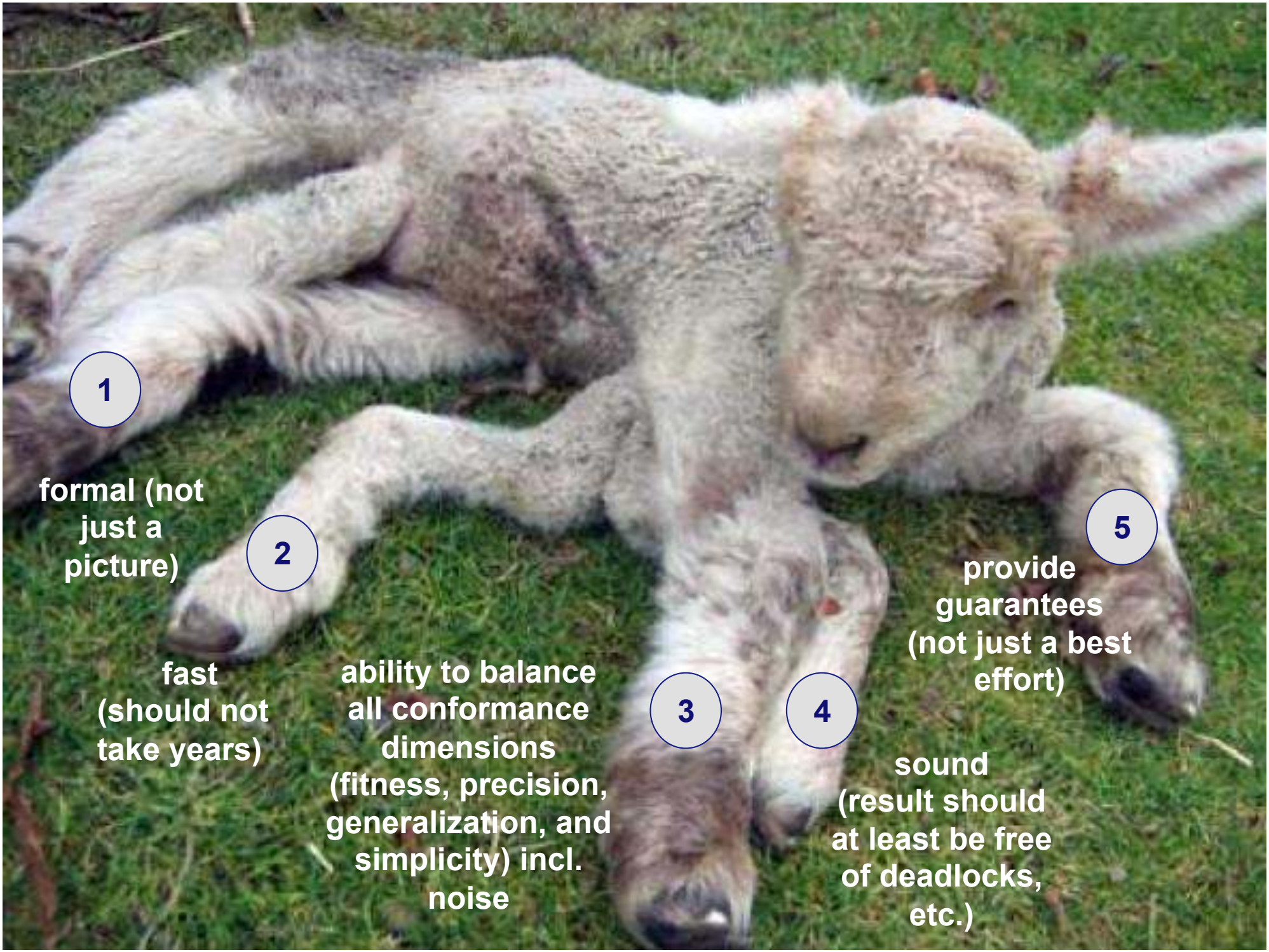
How good is my model: Four forces

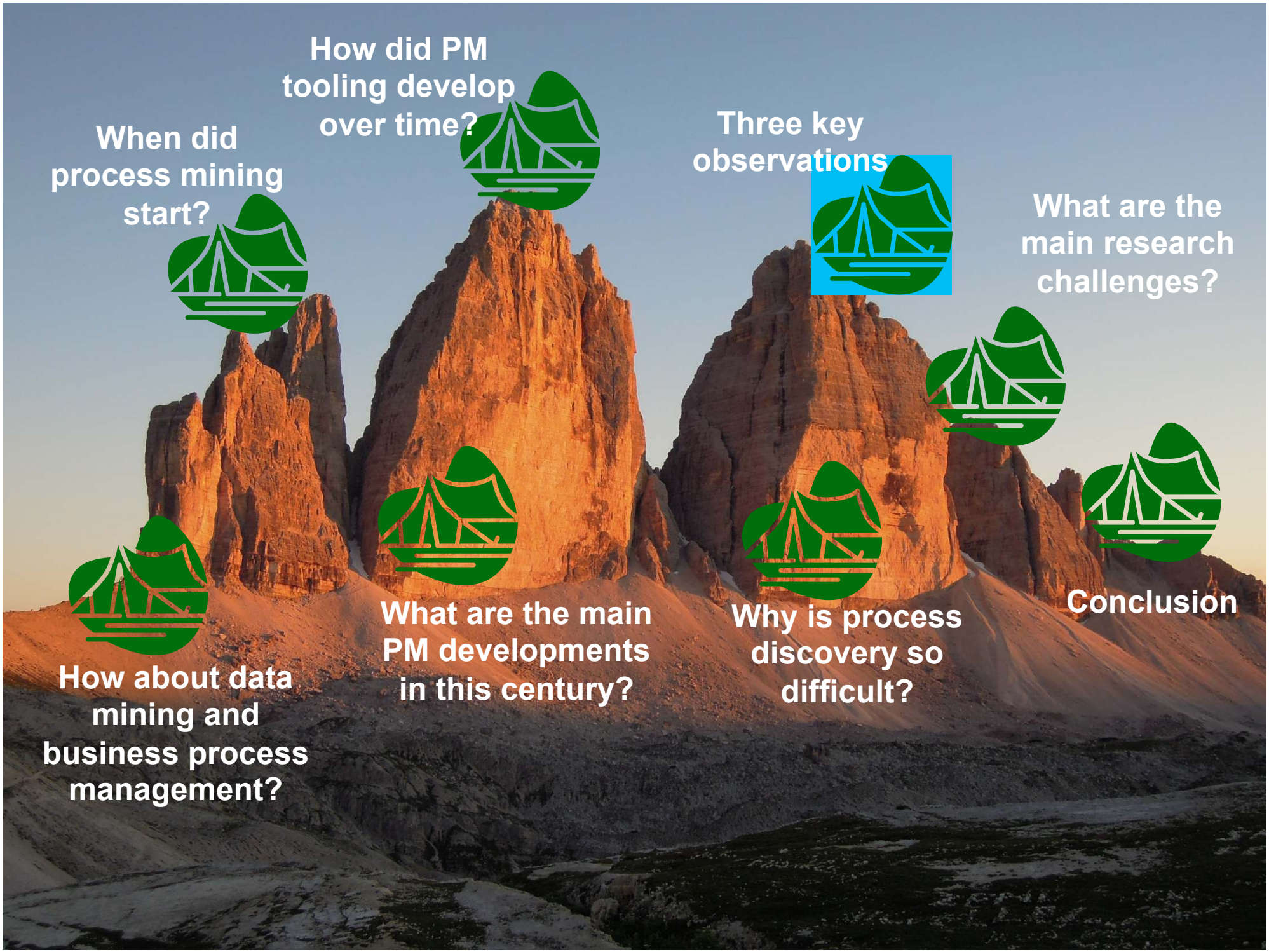


Leaving out one of these dimensions during discovery will lead to degenerate cases!

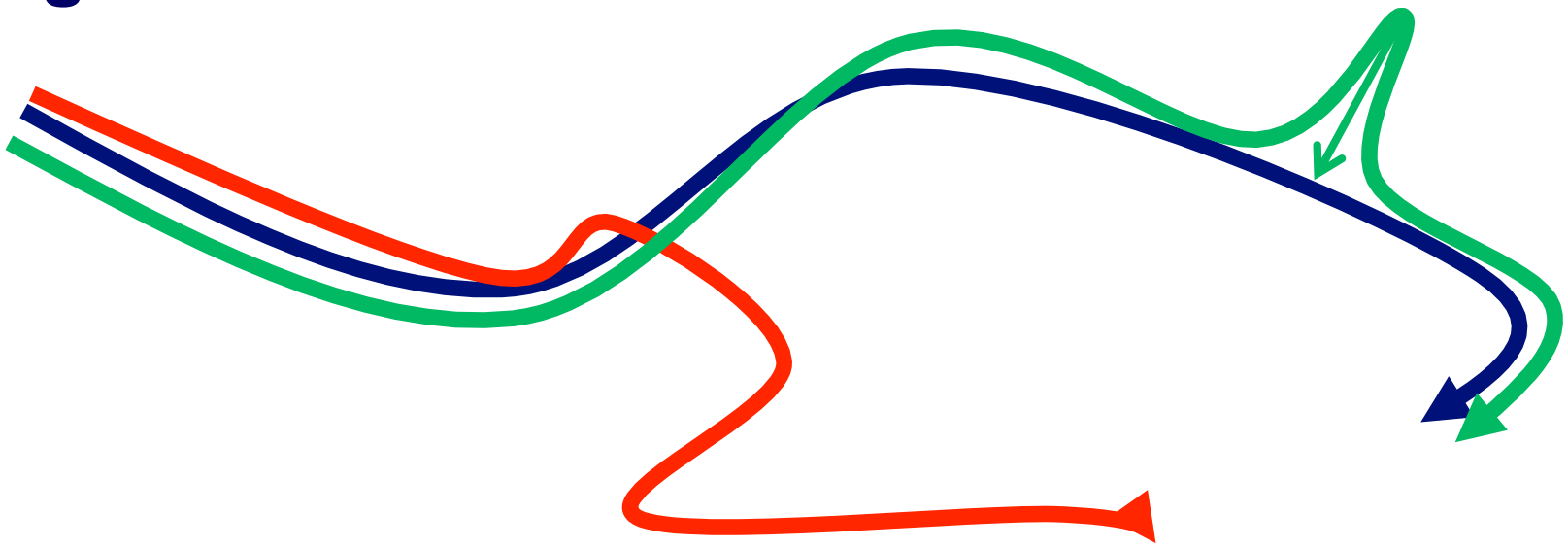
Problem


real process is unknown
event logs covers only a fraction of all possible behavior
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)


When did process mining start?
How did PM tooling develop over time?
Three key observations
What are the main research challenges?
How about data mining and business process management?
What are the main PM developments in this century?
Why is process discovery so difficult?
Conclusion

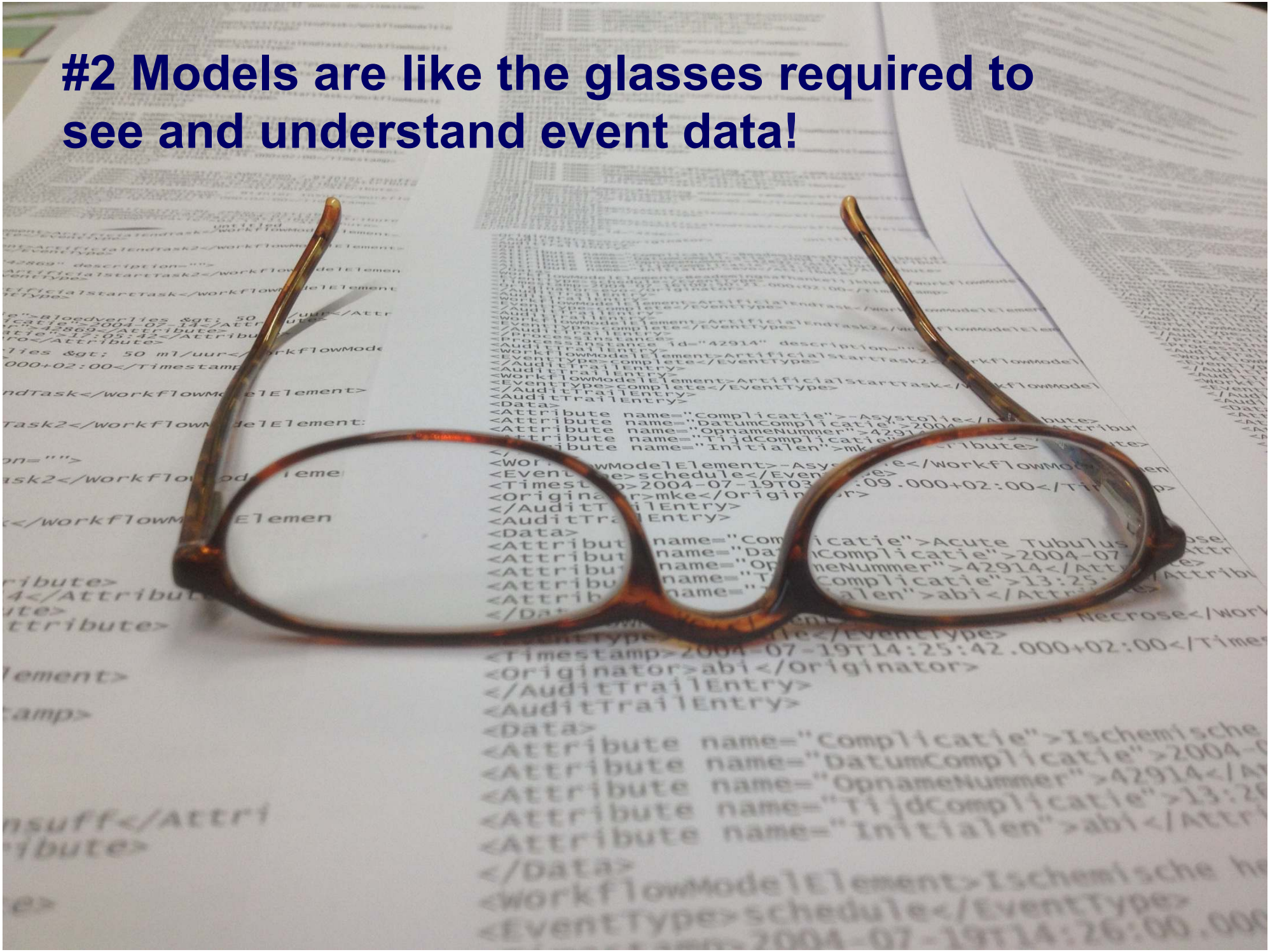
#1 Alignments are essential!

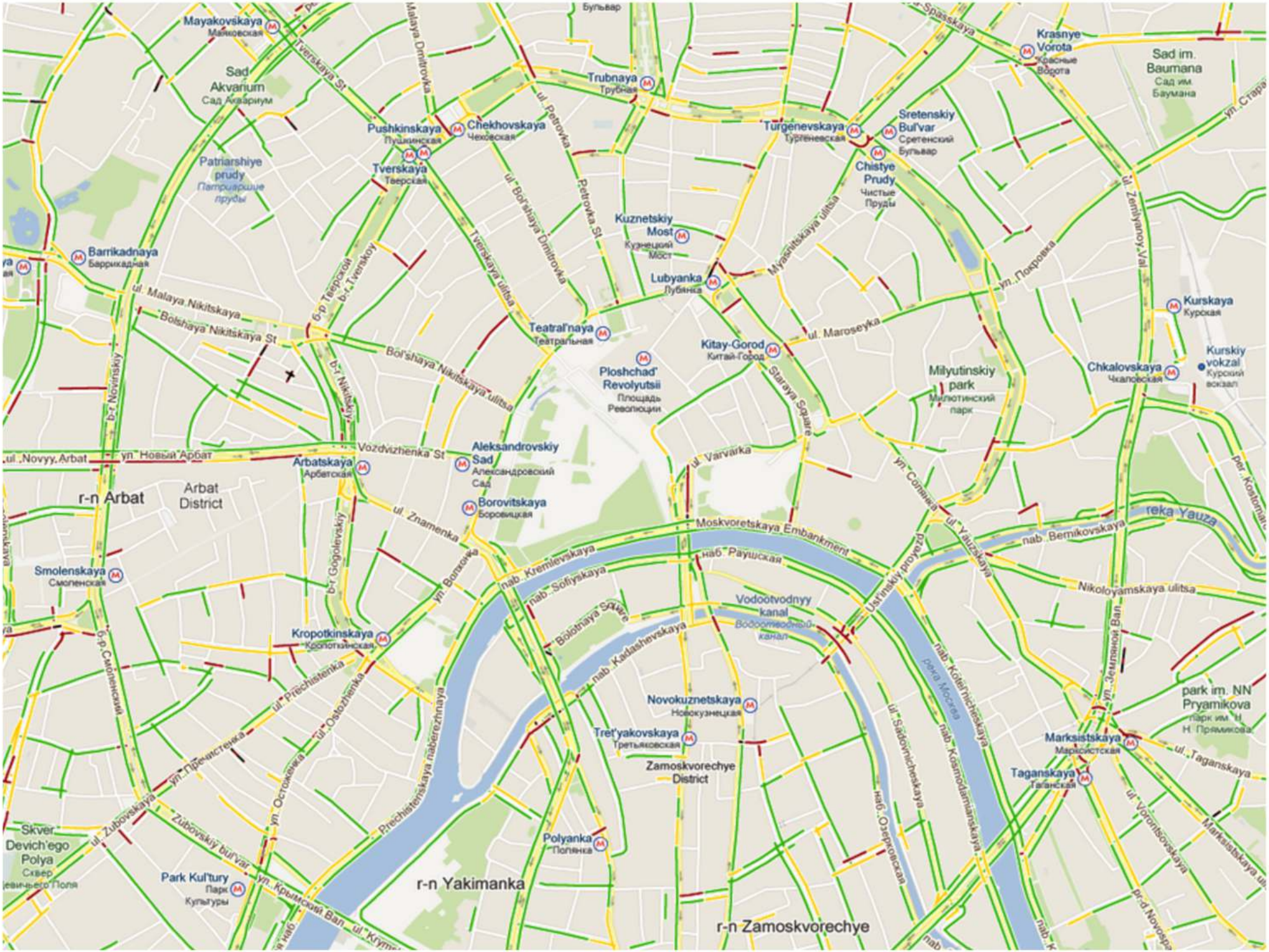


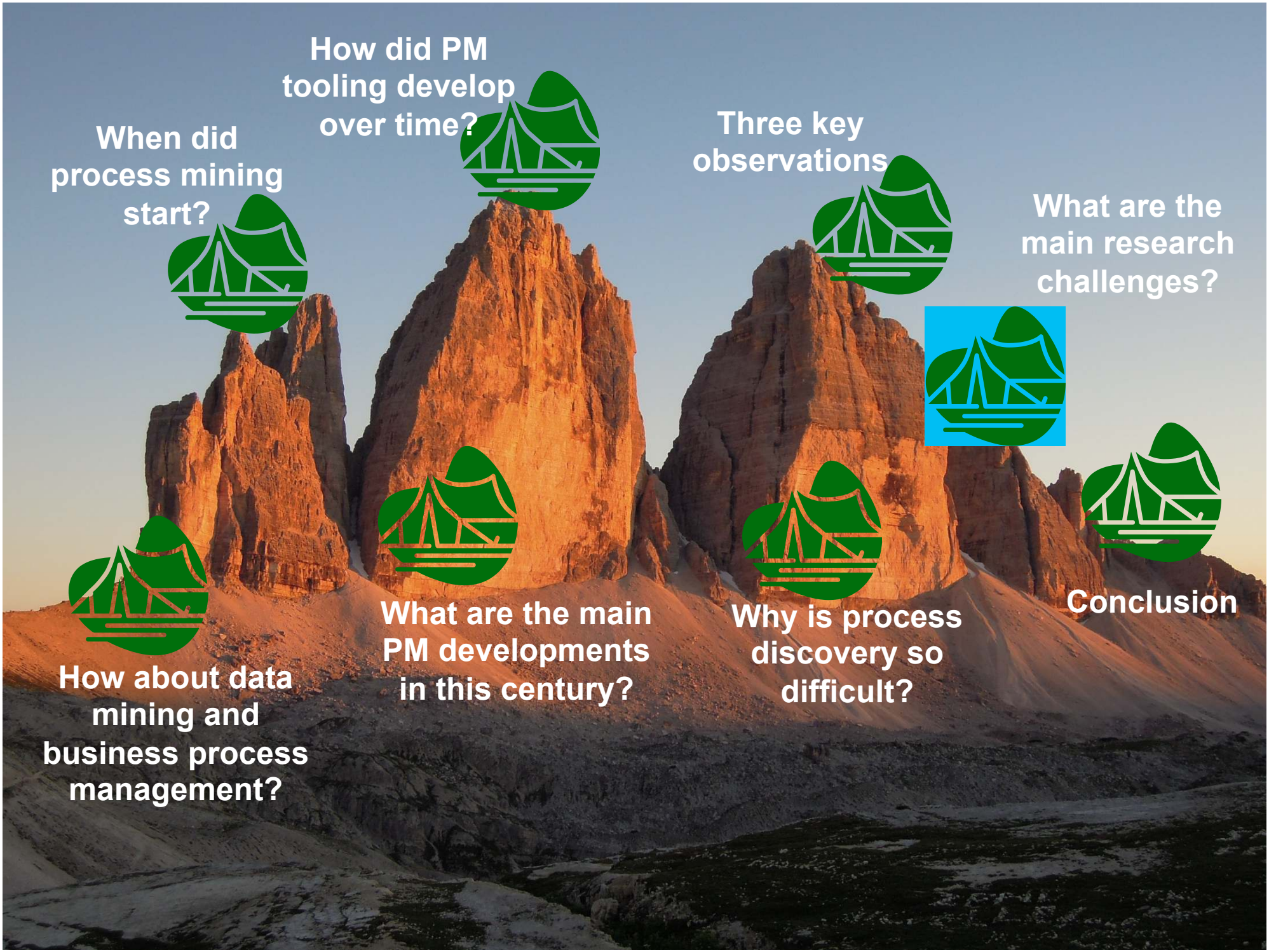
- conformance checking to diagnose deviations
- squeezing reality into the model to do model-based analysis

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

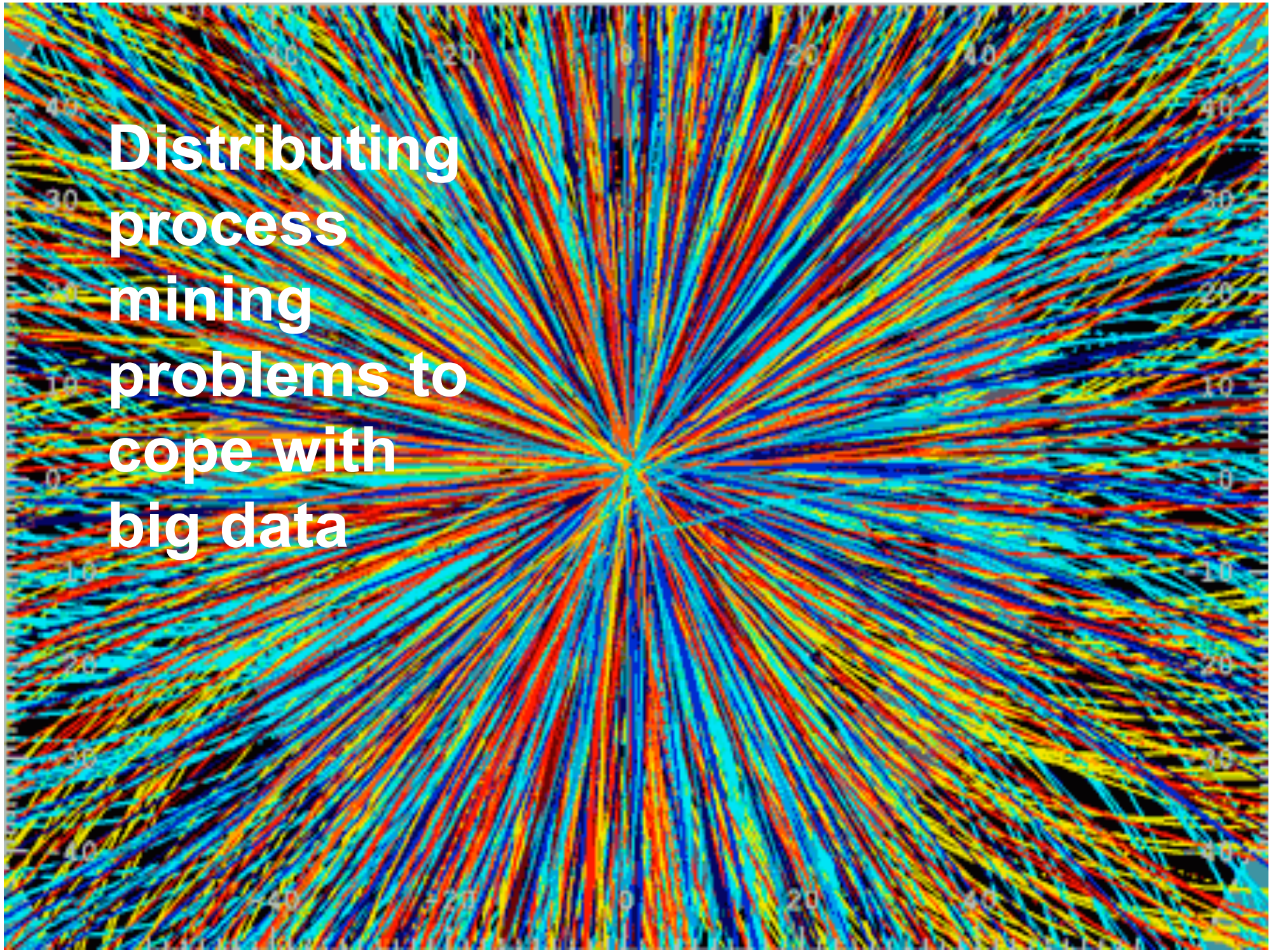
#2 Models are like the glasses required to see and understand event data!




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


When did process mining start?
How did PM tooling develop over time?
Three key observations
What are the main research challenges?
How about data mining and business process management?
What are the main PM developments in this century?
Why is process discovery so difficult?
Conclusion

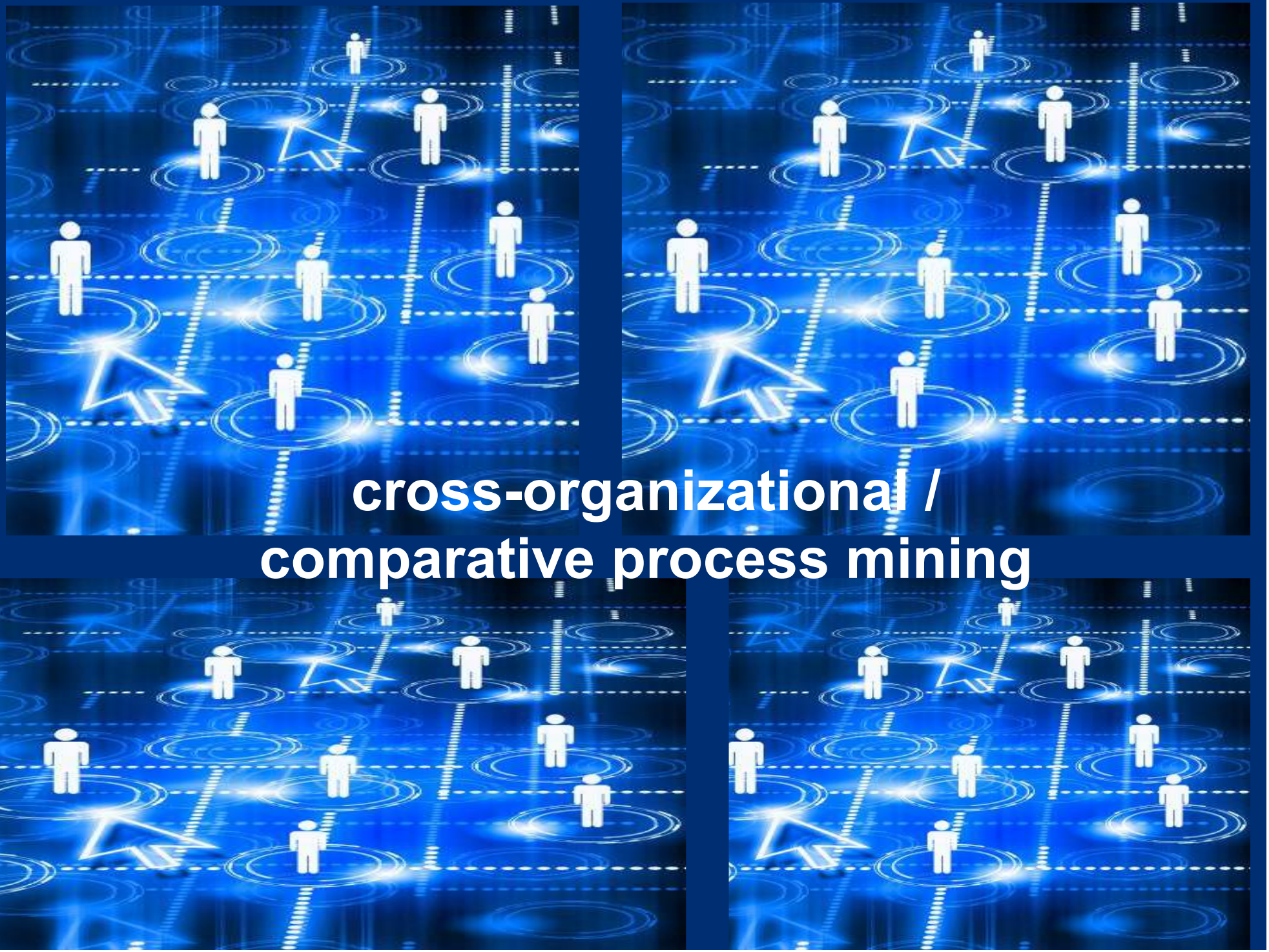

Finding sheep with five legs
we are getting close…


Distributing process mining problems to cope with big data

On-the-fly process mining

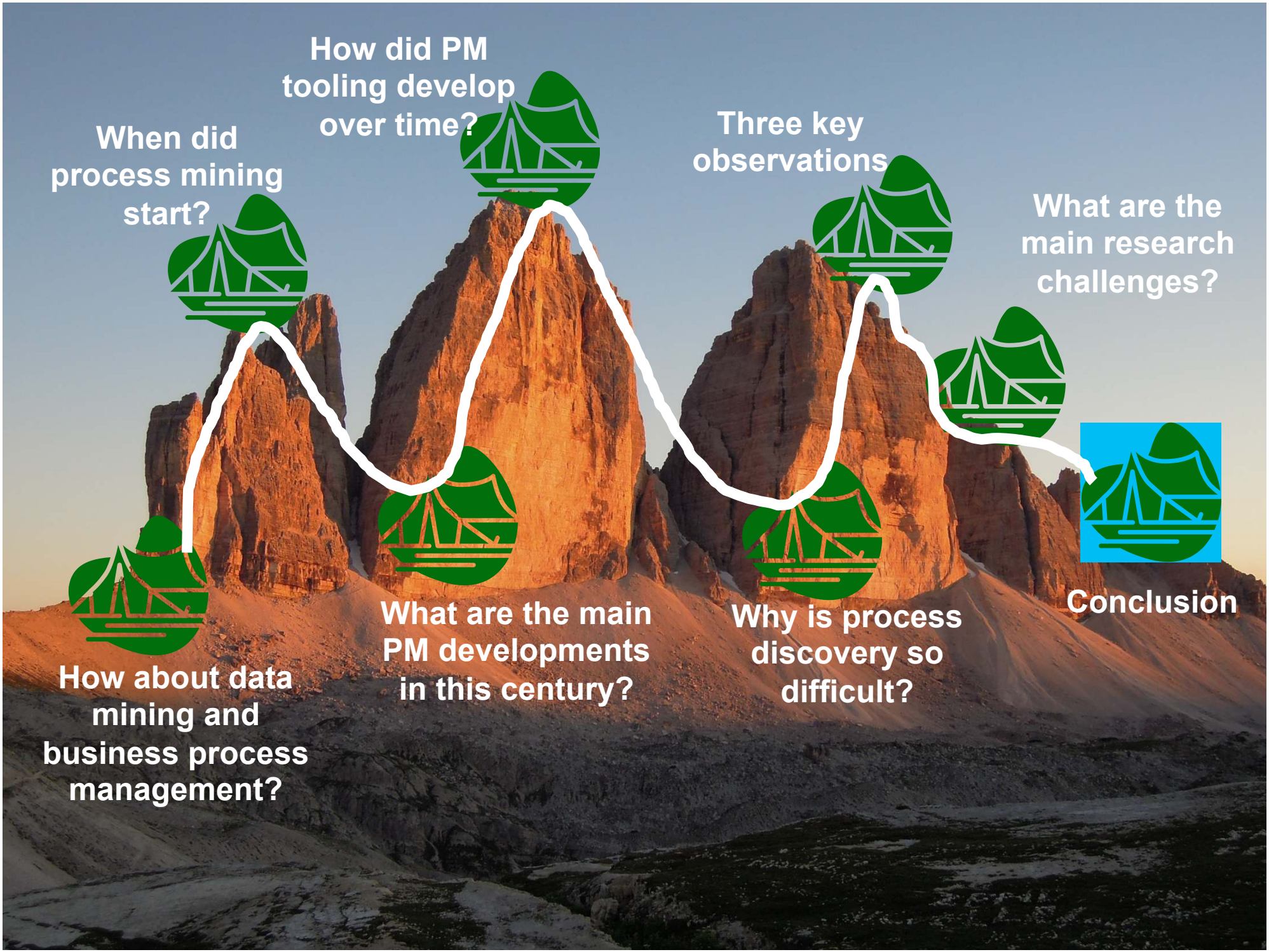
Operational support


Concept drift

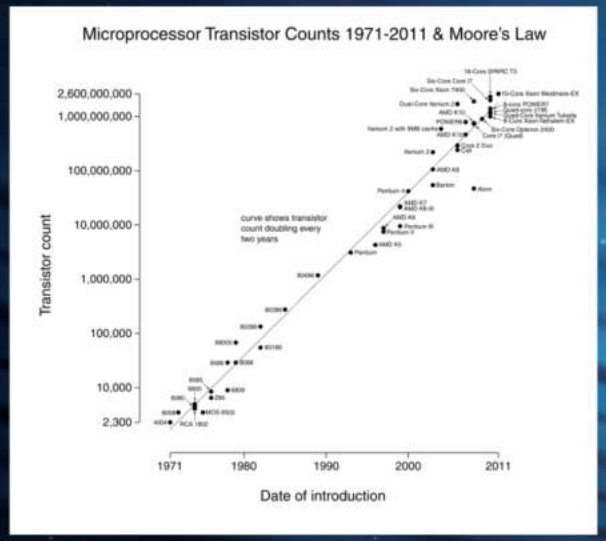

cross-organizational /
comparative process mining

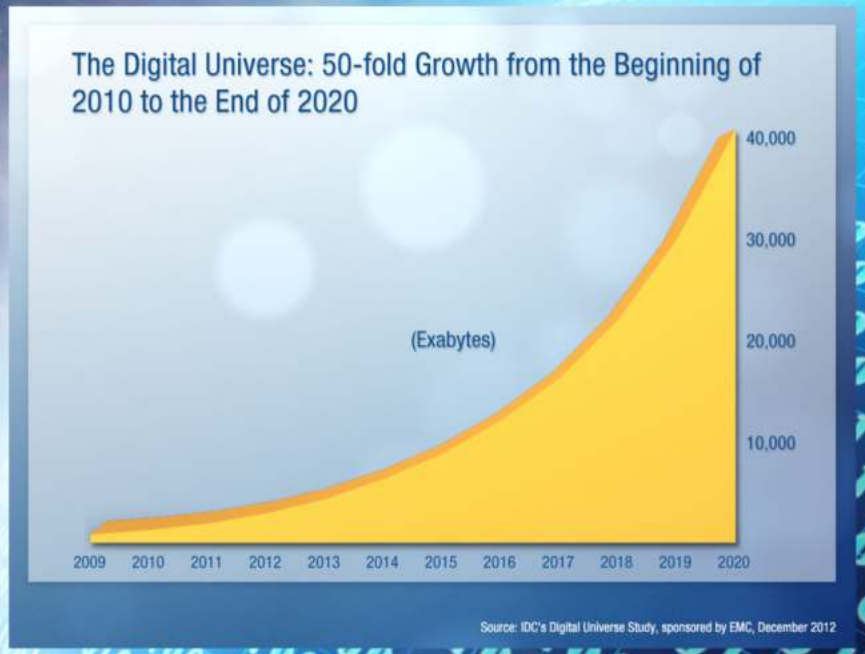

context aware
process mining

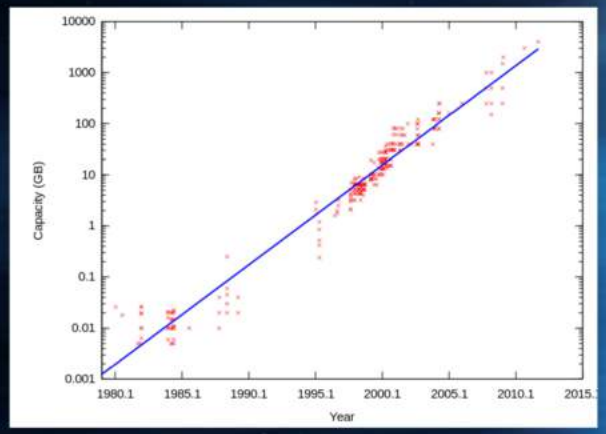
Supporting the process of process mining


When did process mining start?
How did PM tooling develop over time?
Three key observations
What are the main research challenges?
How about data mining and business process management?
What are the main PM developments in this century?
Why is process discovery so difficult?
Conclusion

Moore's Law


Microprocessor Transistor Counts 1971-2011 & Moore's Law
Transistor count
2,600,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
2,300
curve shows transistor count doubling every two years
1971
1980
1990
2000
2011
Date of introduction


The Digital Universe: 50-fold Growth from the Beginning of 2010 to the End of 2020
(Exabytes)
40,000
30,000
20,000
10,000
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Source: IDC's Digital Universe Study, sponsored by EMC, December 2012


Capacity (GB)
10000
1000
100
10
1
0.1
0.01
0.001
1980.1
1985.1
1990.1
1995.1
2000.1
2005.1
2010.1
Year

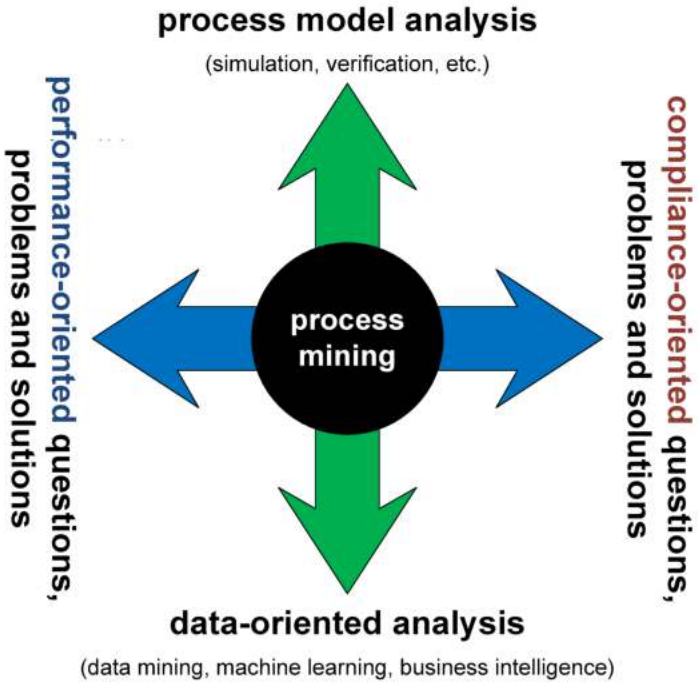

2060

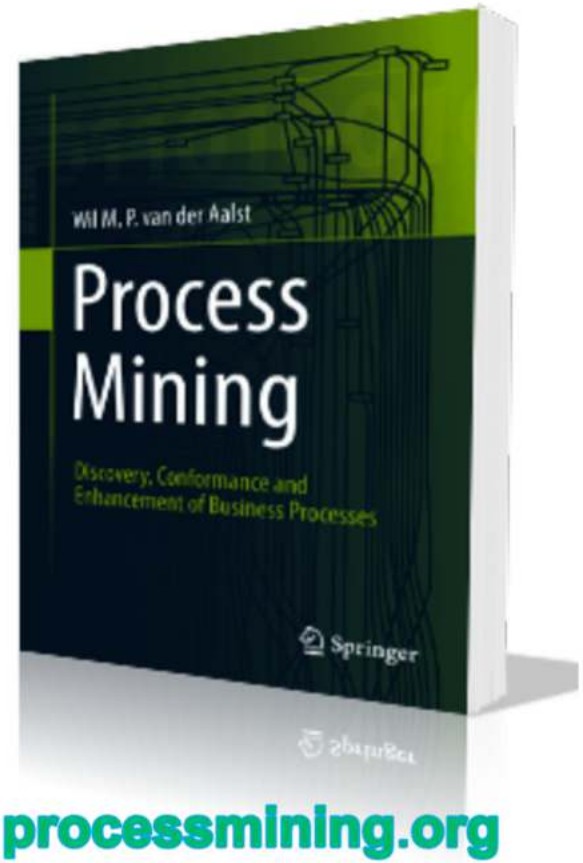
Turning Event Data into Real Value

Harvard Business Review

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil





processmining.org

IEEE

http://www.win.tue.nl/ieeetfpm/