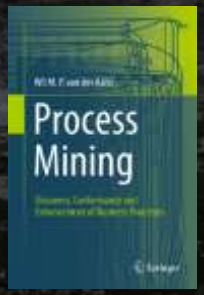
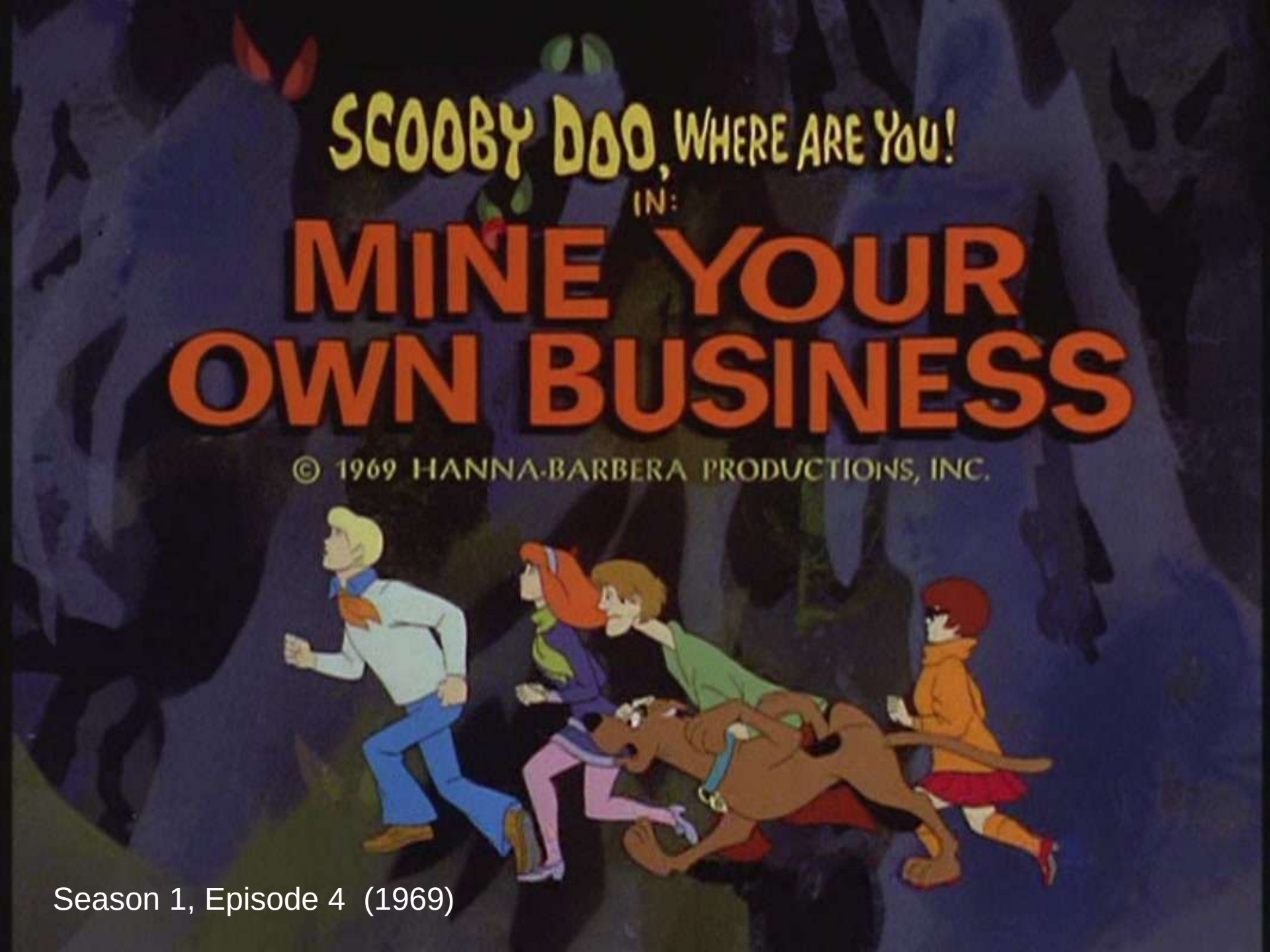
Mine Your Own Business

Using process mining to turn big data into better processes and systems

prof.dr.ir. Wil van der Aalst

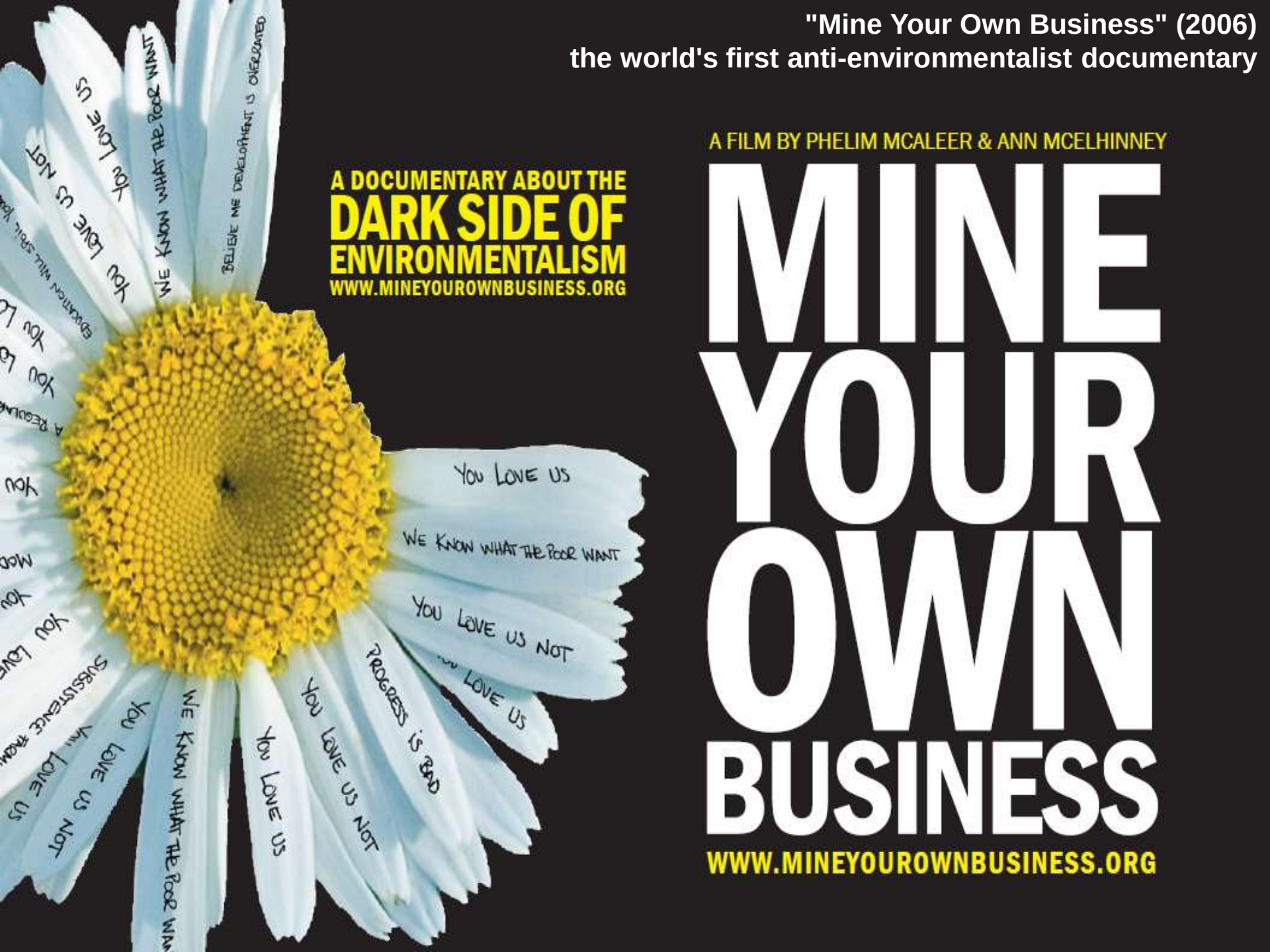


Technische Universiteit
Eindhoven
University of Technology

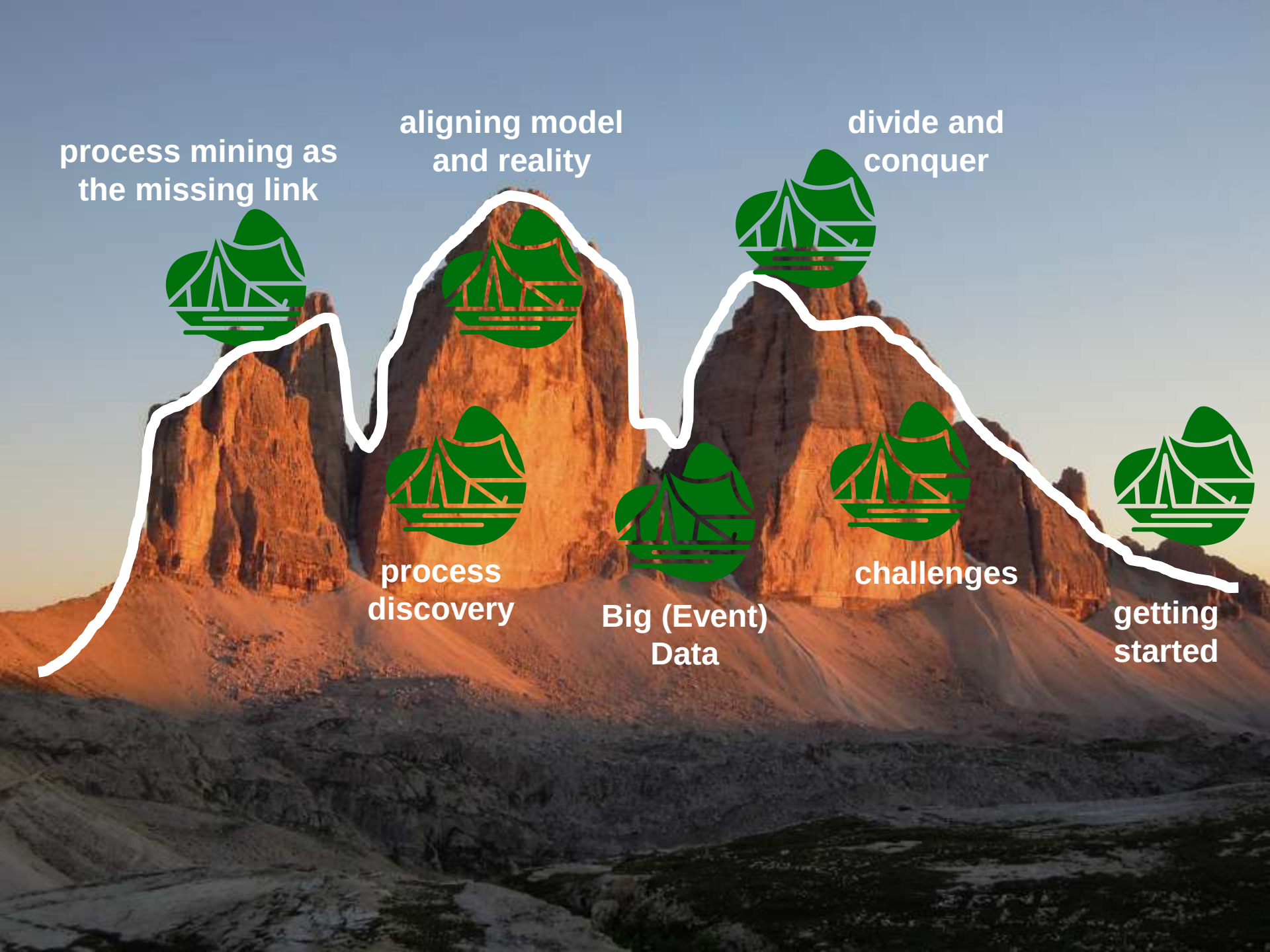


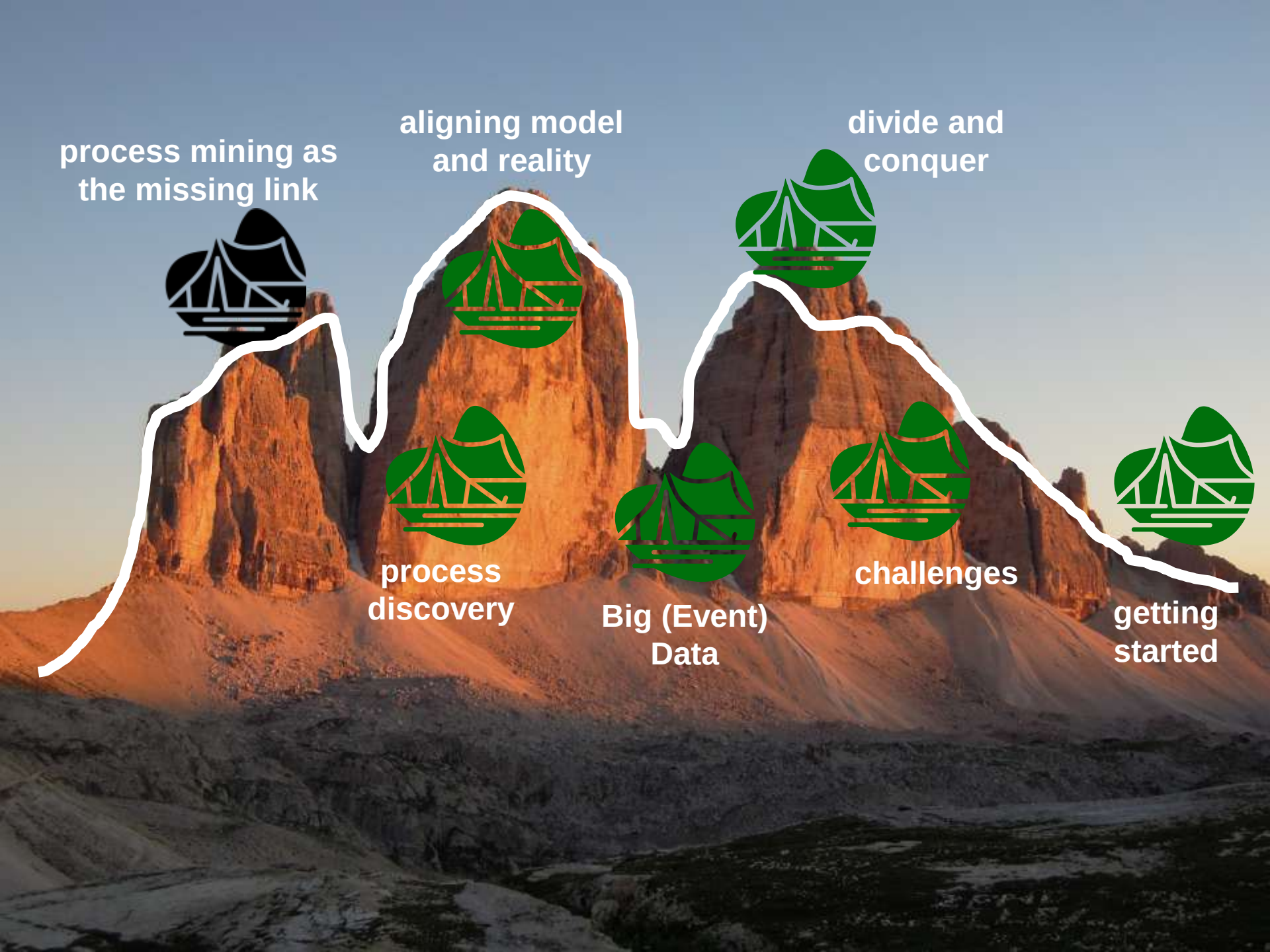
Season 1, Episode 4 (1969)

"Mine Your Own Business" (2006)
the world's first anti-environmentalist documentary



Mine your own business: Turning big data into real value


process mining as the missing link
aligning model and reality
divide and conquer
process discovery
Big (Event) Data
challenges
getting started

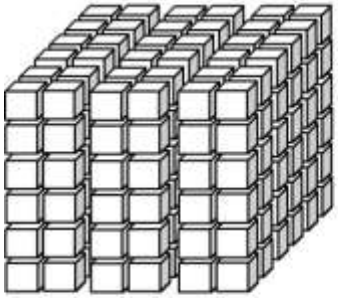

process mining as the missing link
aligning model and reality
divide and conquer
process discovery
Big (Event) Data
challenges
getting started

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)

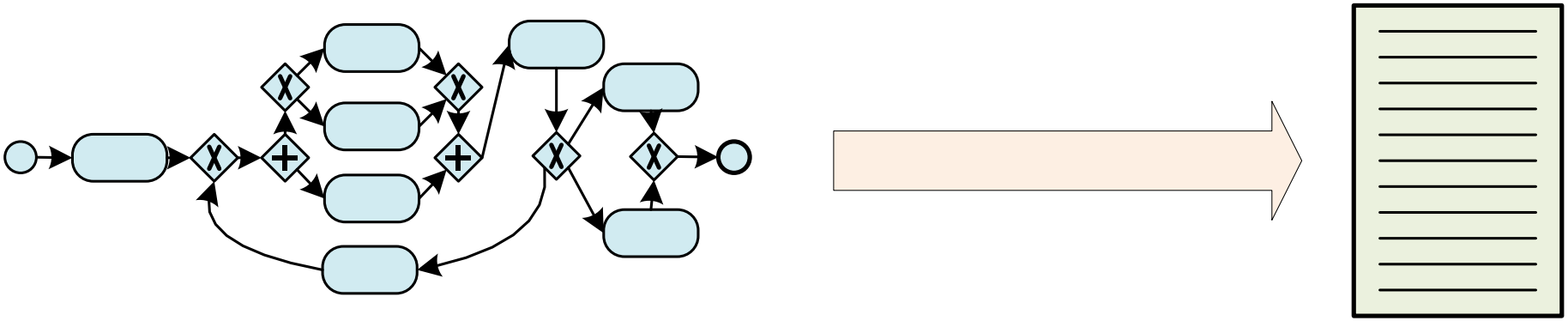

www.olifantenpaadjes.nl
60


0101010101010
0100110010


小草对您微微笑
请您把路绕一绕
KEEP OFF GRASS
绿色大学办公室
绿化中心园林科


let's play

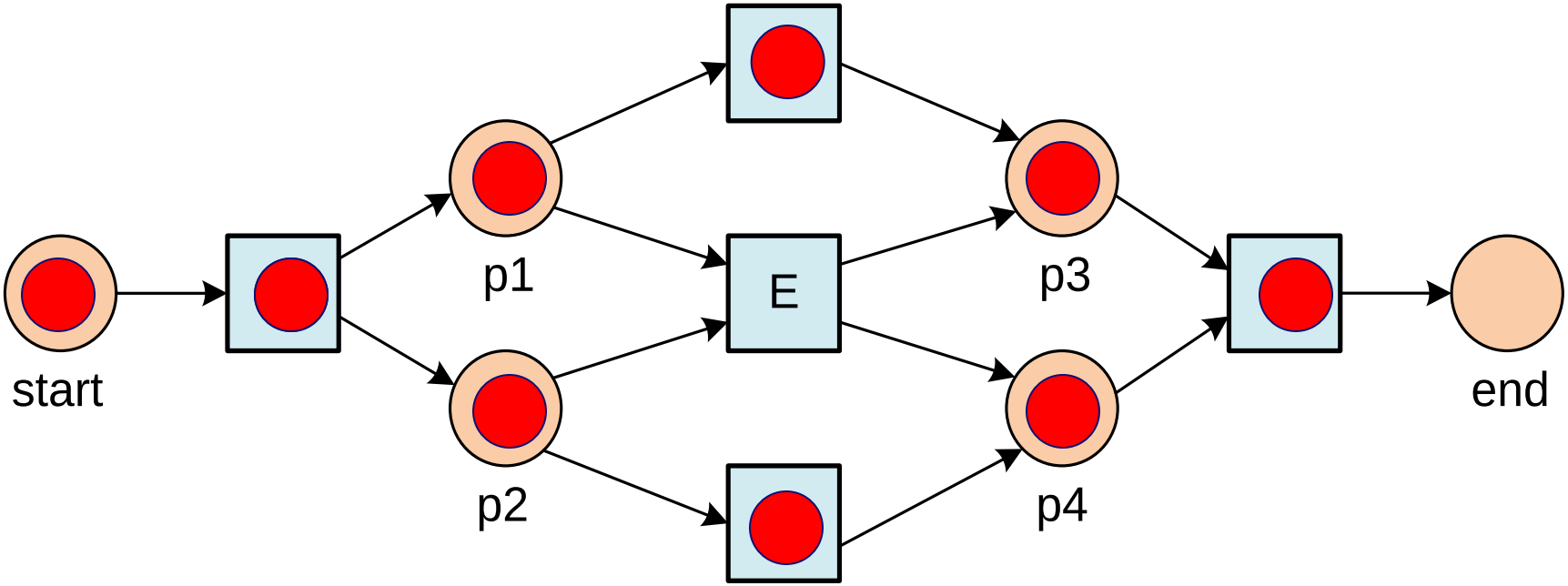
Play-Out



process model

event log

Play-Out (Classical use of models)



A B C D

A E D

A E D

A B C D

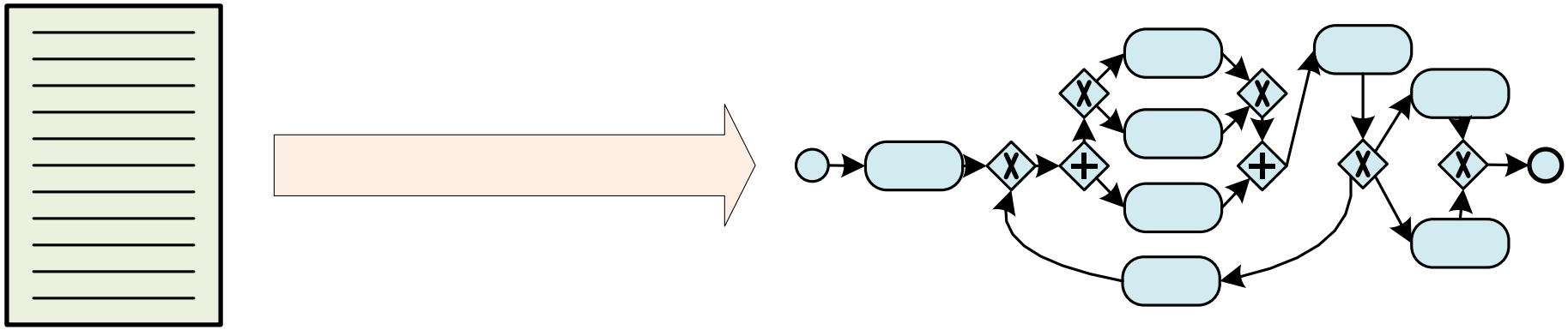
A C B D

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

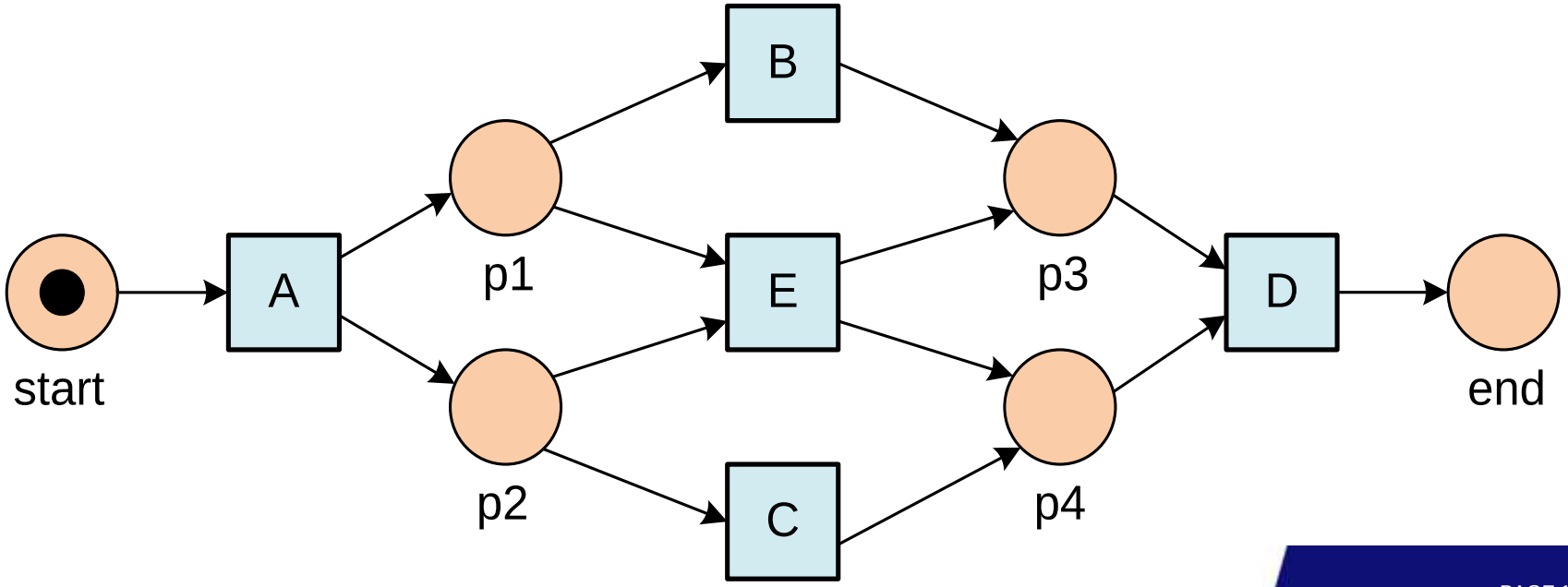
A B C D

A C B D

A C B D

A E D

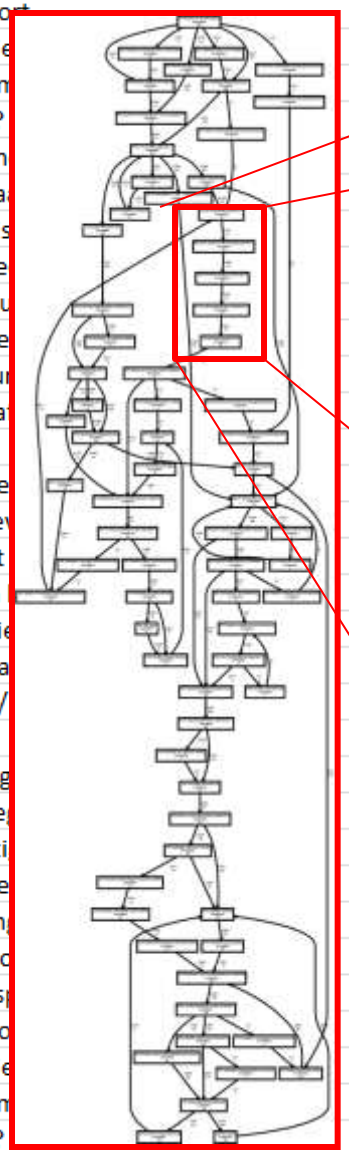
A C B D

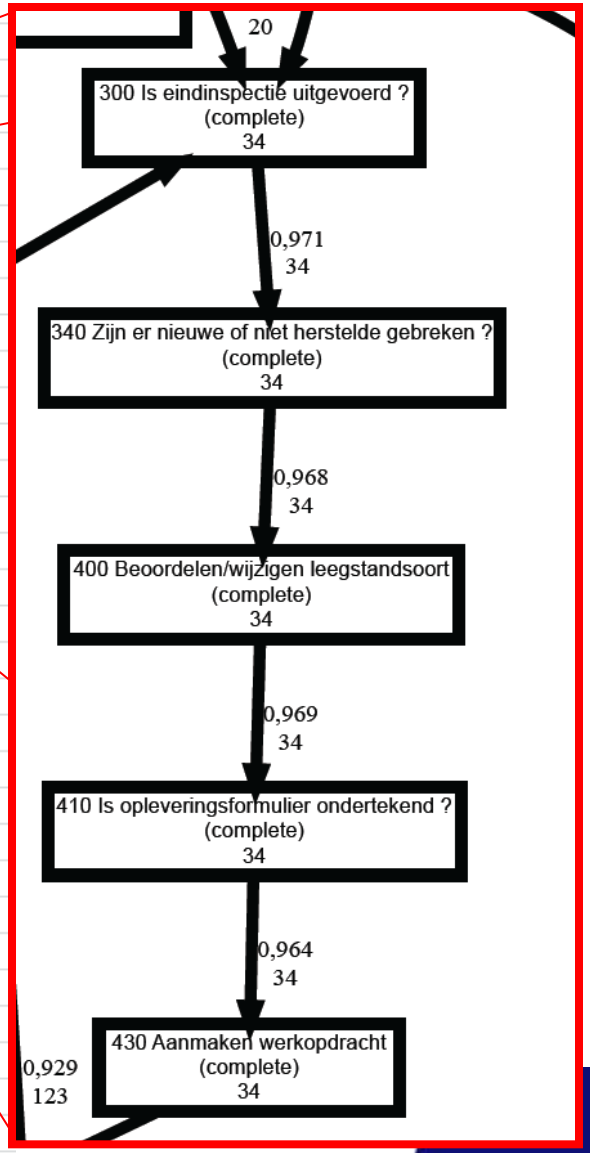


Example Process Discovery

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

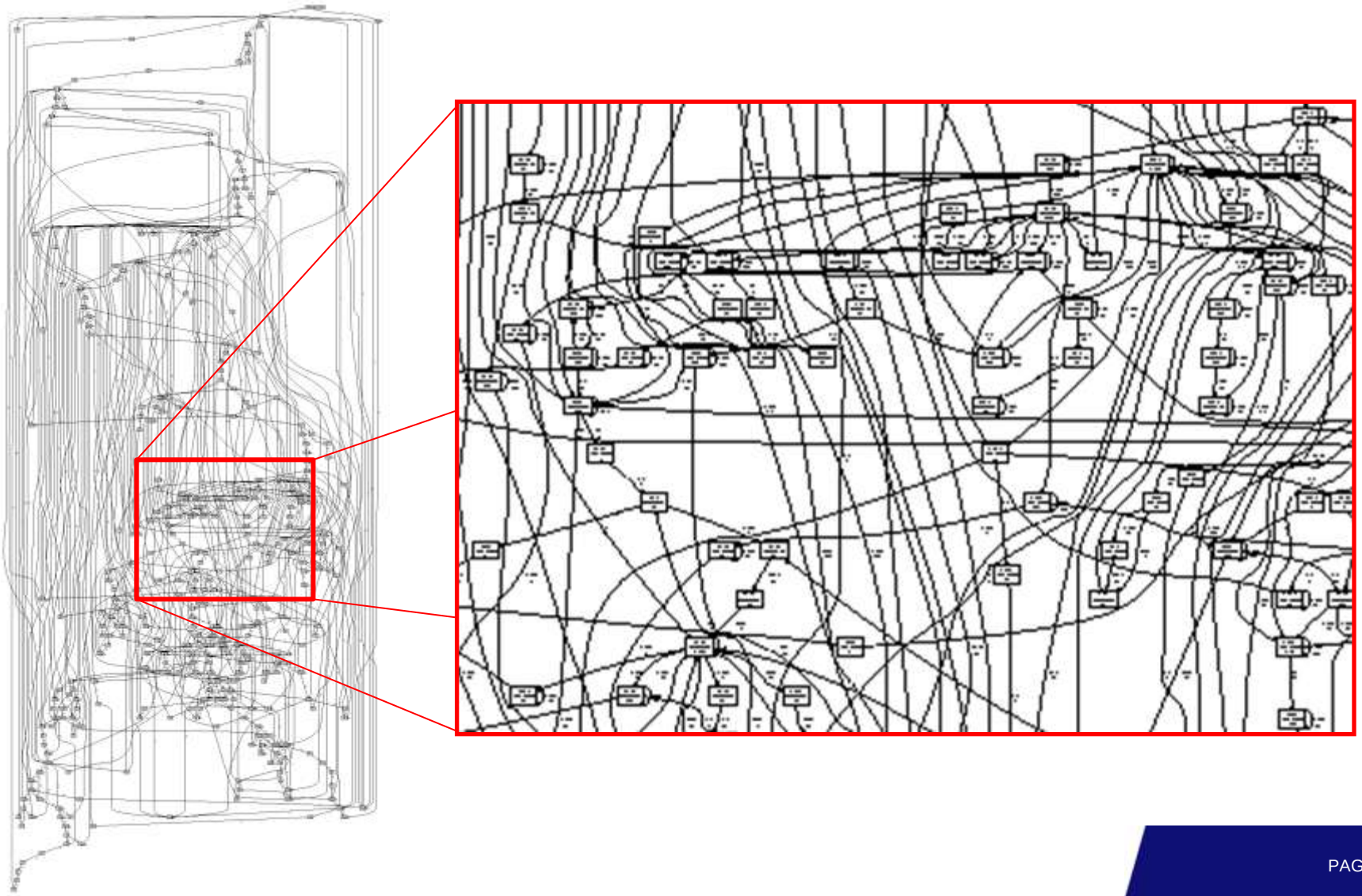
117315	110 Bepalen leegstandsoort	16.05.2007 14:06:23
117315	120 Plannen eindinspectie	16.05.2007 14:36:01
117315	130 Is het opleveringsform	23.05.2007 09:41:40
117315	150 Is er sprake van ZAV ?	23.05.2007 09:41:51
117315	170 Aanpassen plattegron	23.05.2007 11:57:18
117315	180 Aanpassen woningwa	23.05.2007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





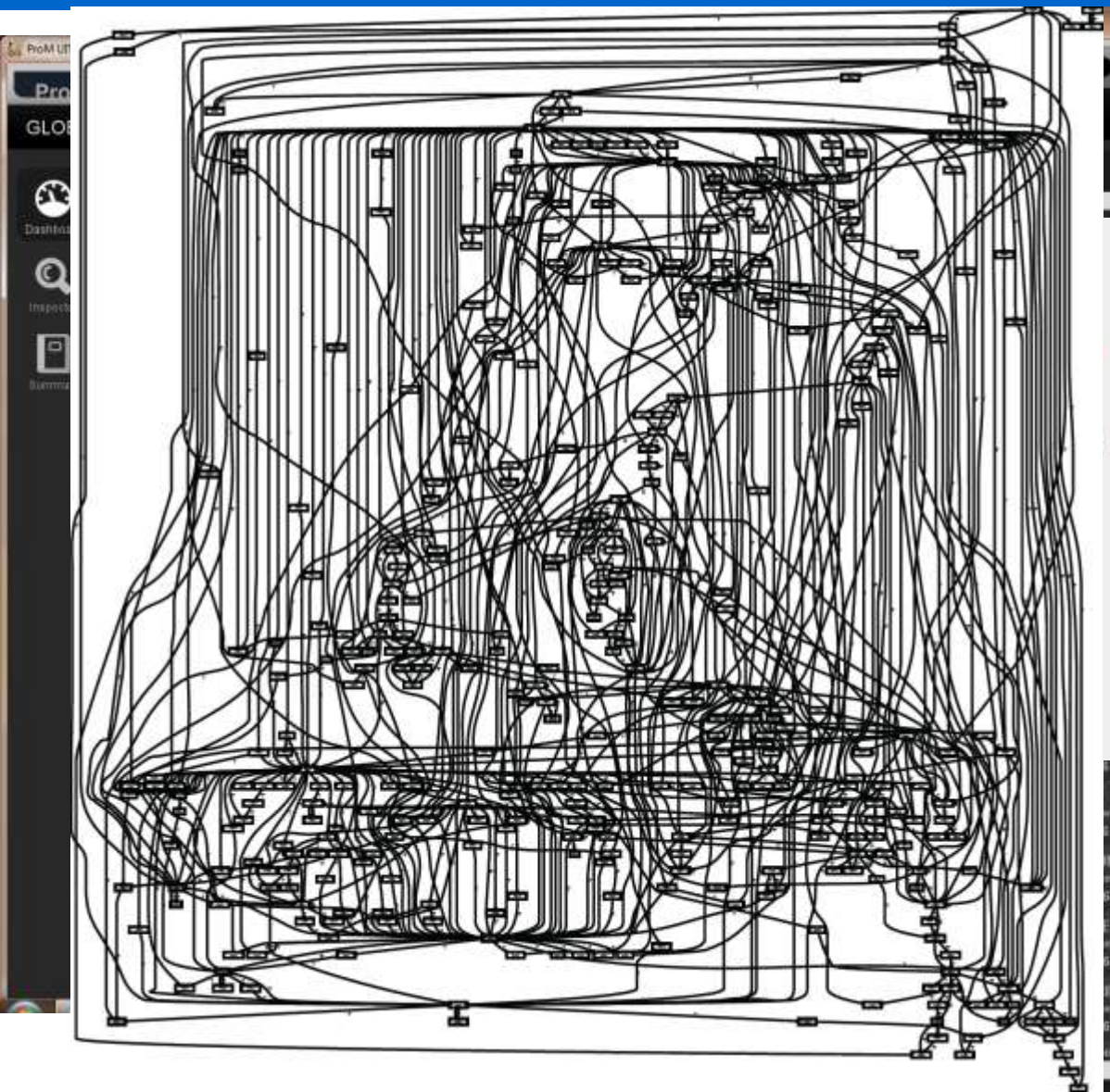
Example Process Discovery

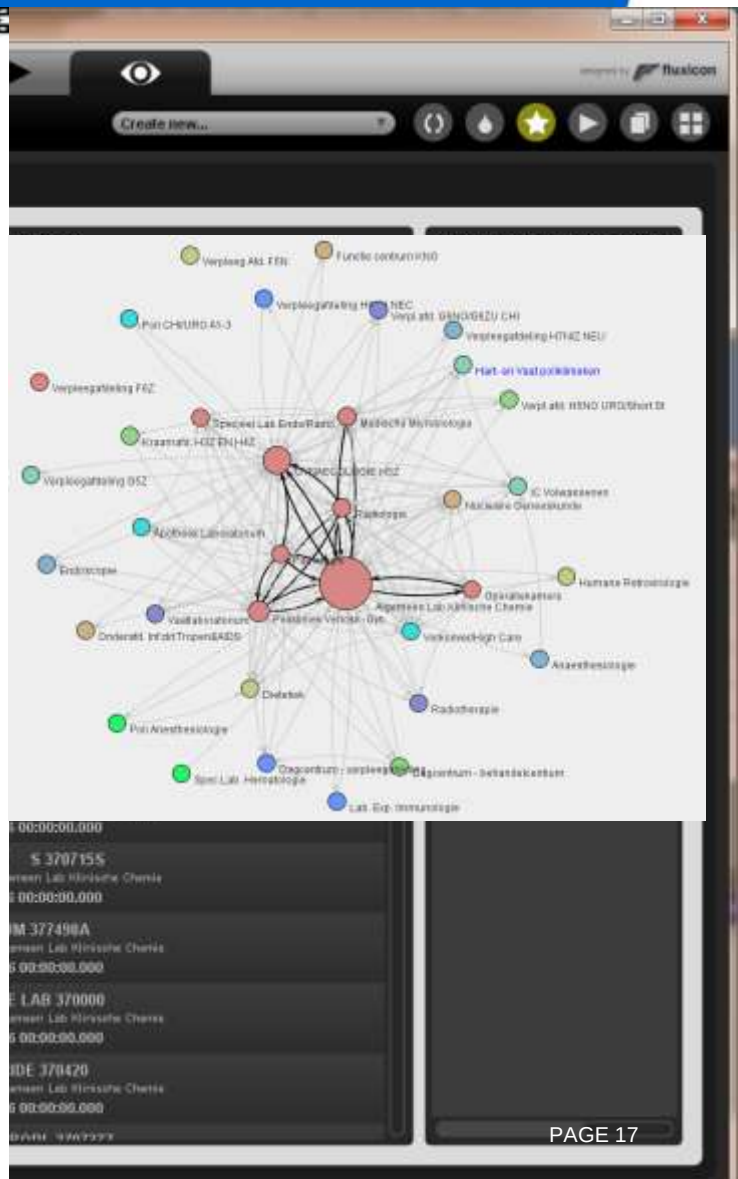
(ASML, test process lithography systems, 154966 events)



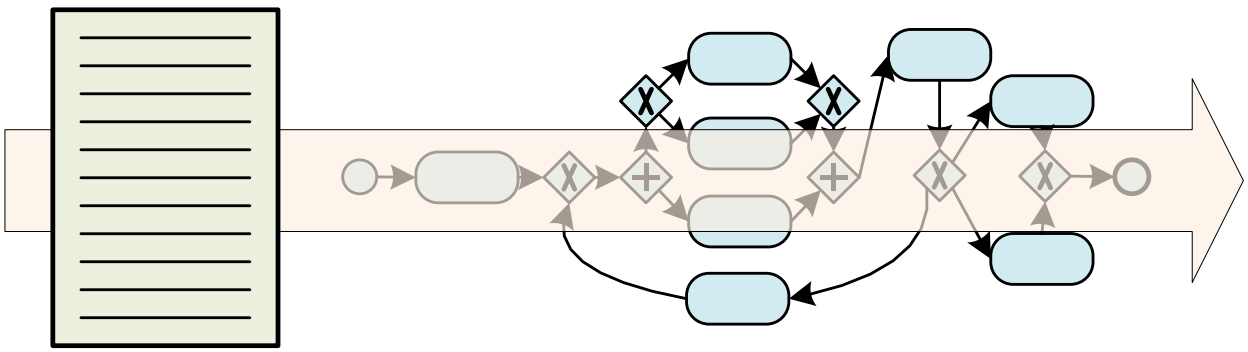
Example Process Discovery

(AMC, 627 gynecological oncology patients, 24331 events)





Replay



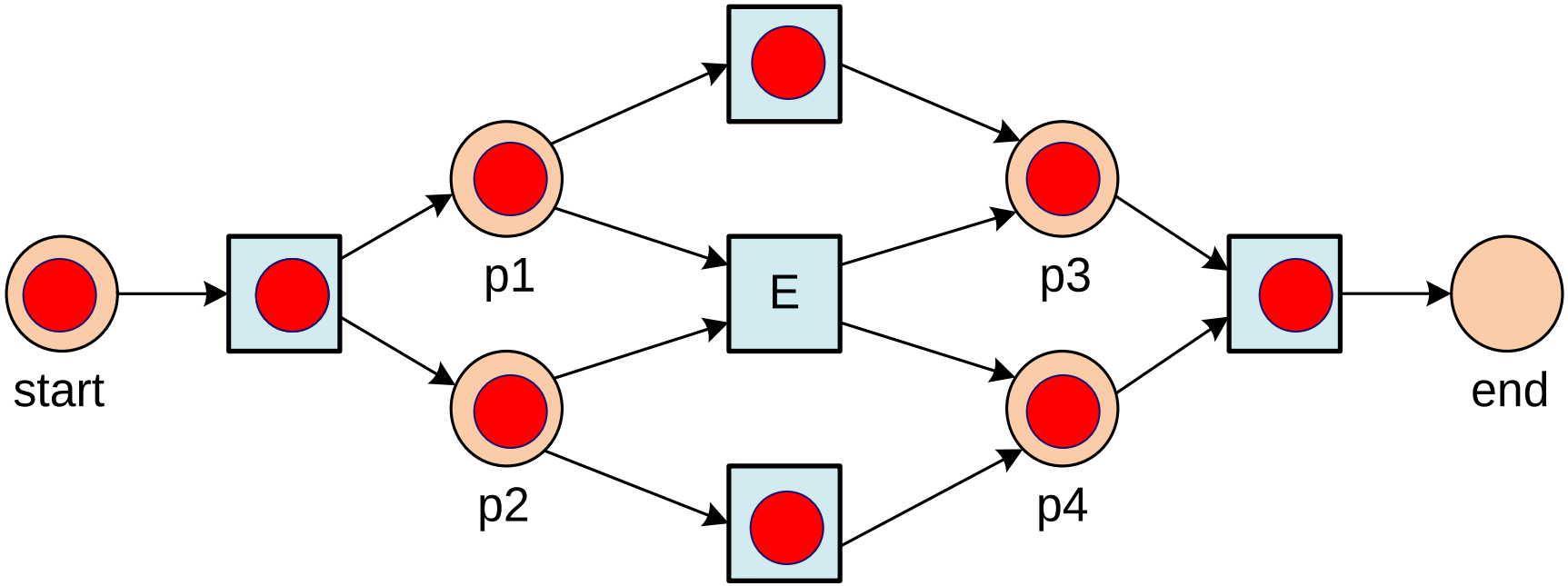
event log

process model

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

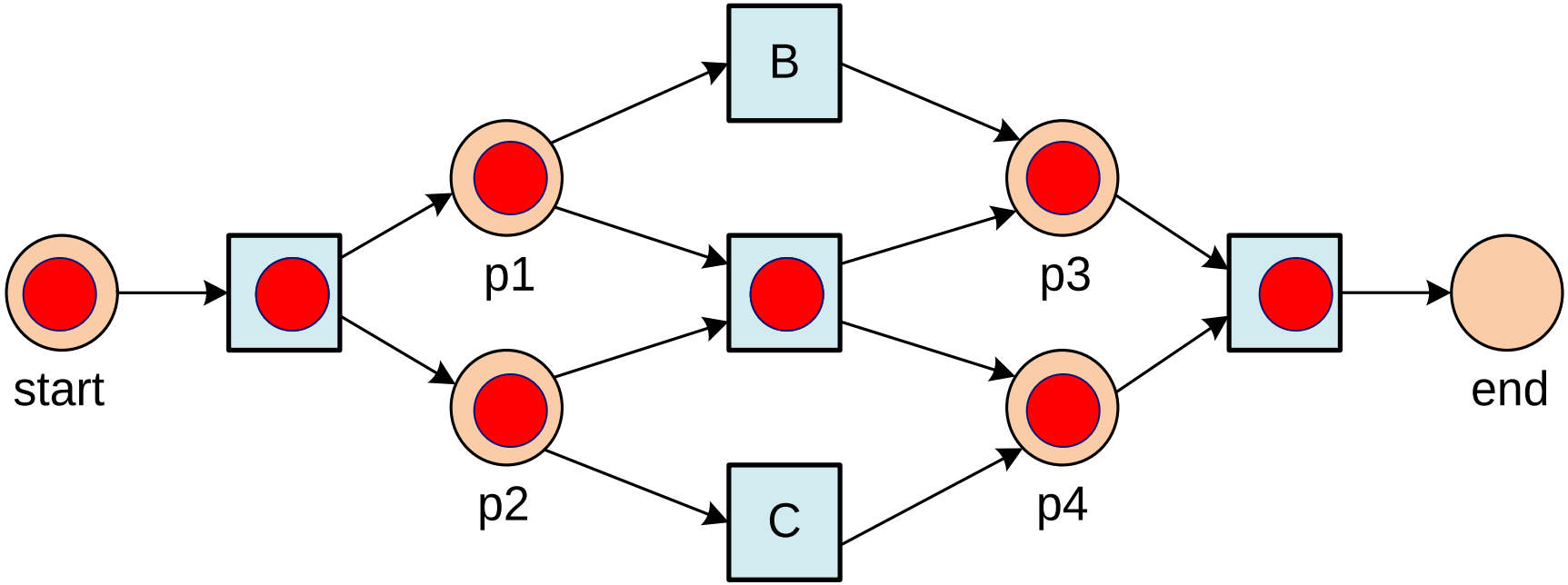
Replay

A B C D



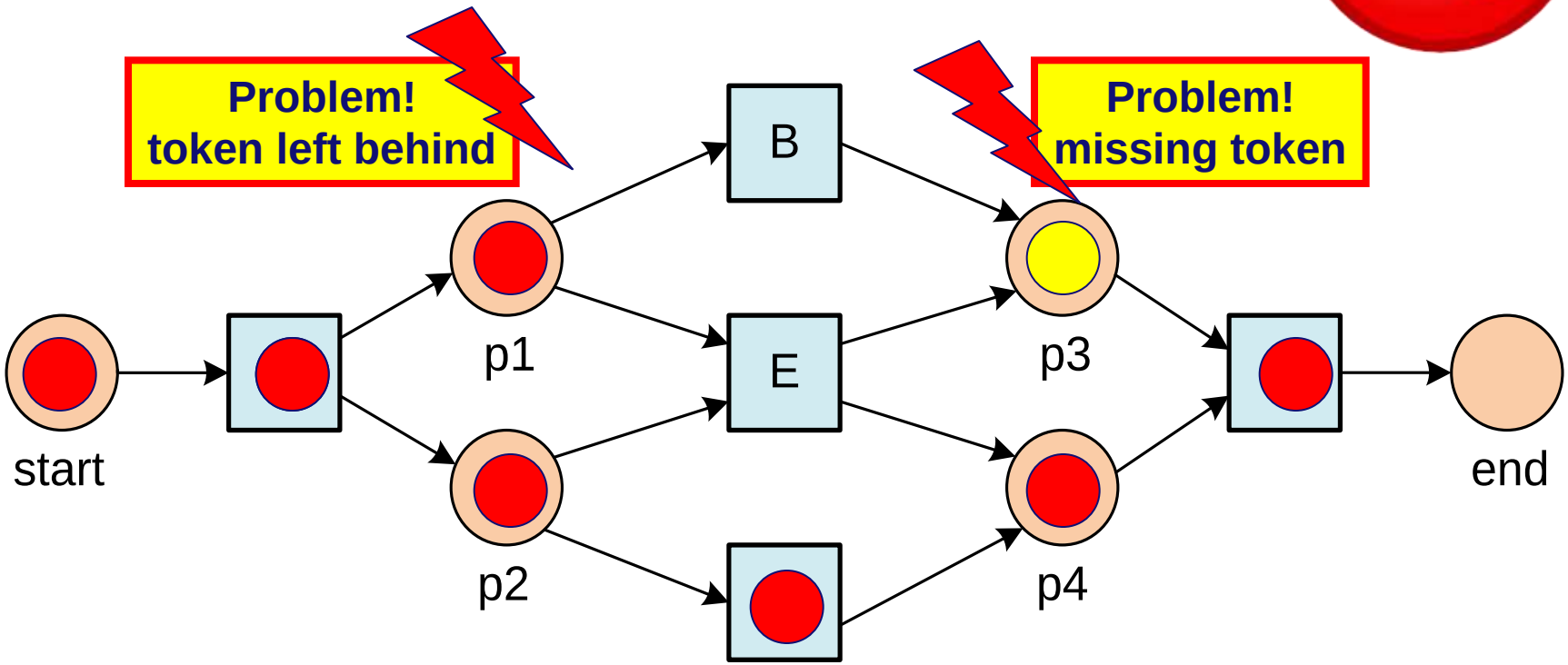
Replay

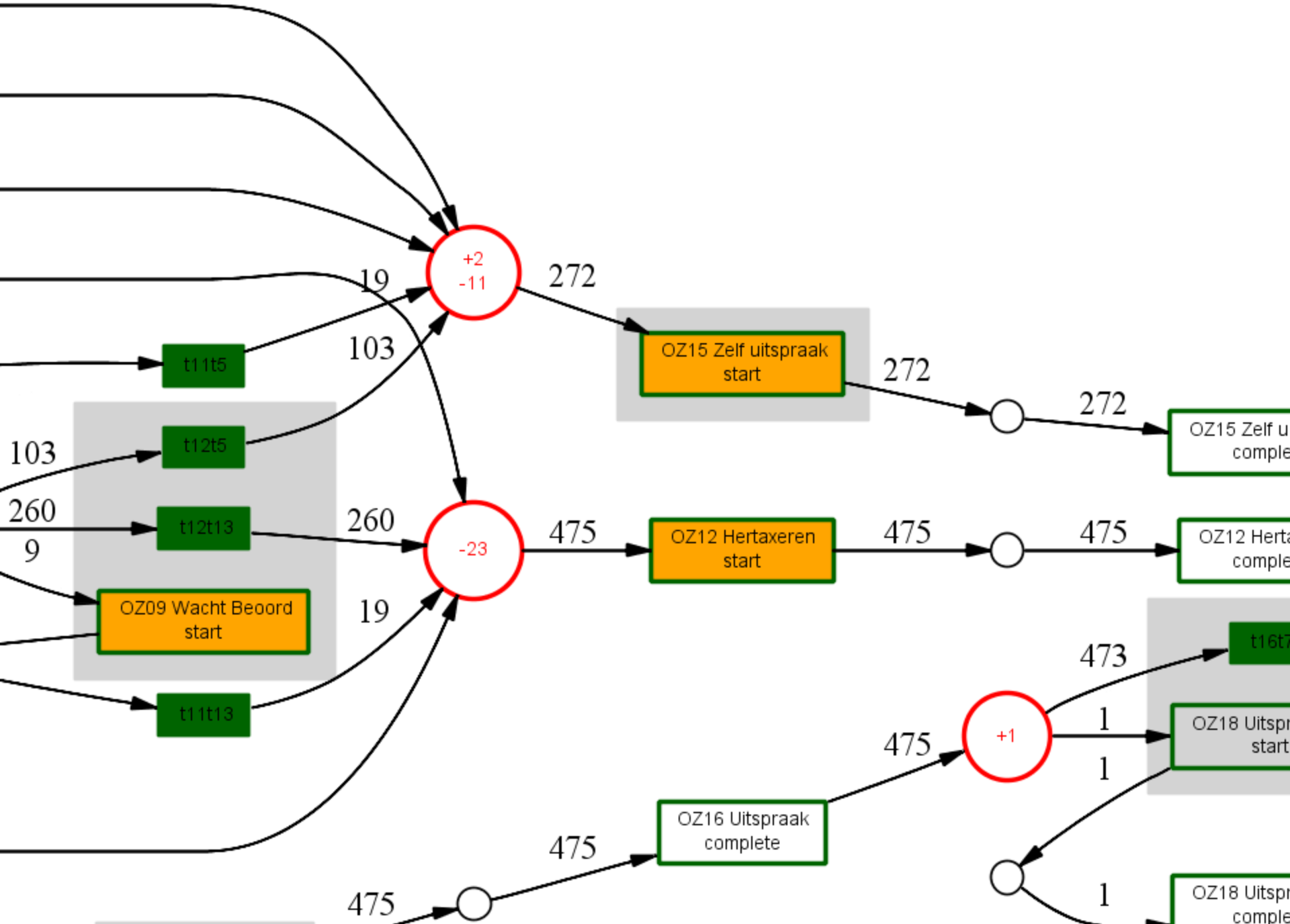
A E D



Replay can detect problems

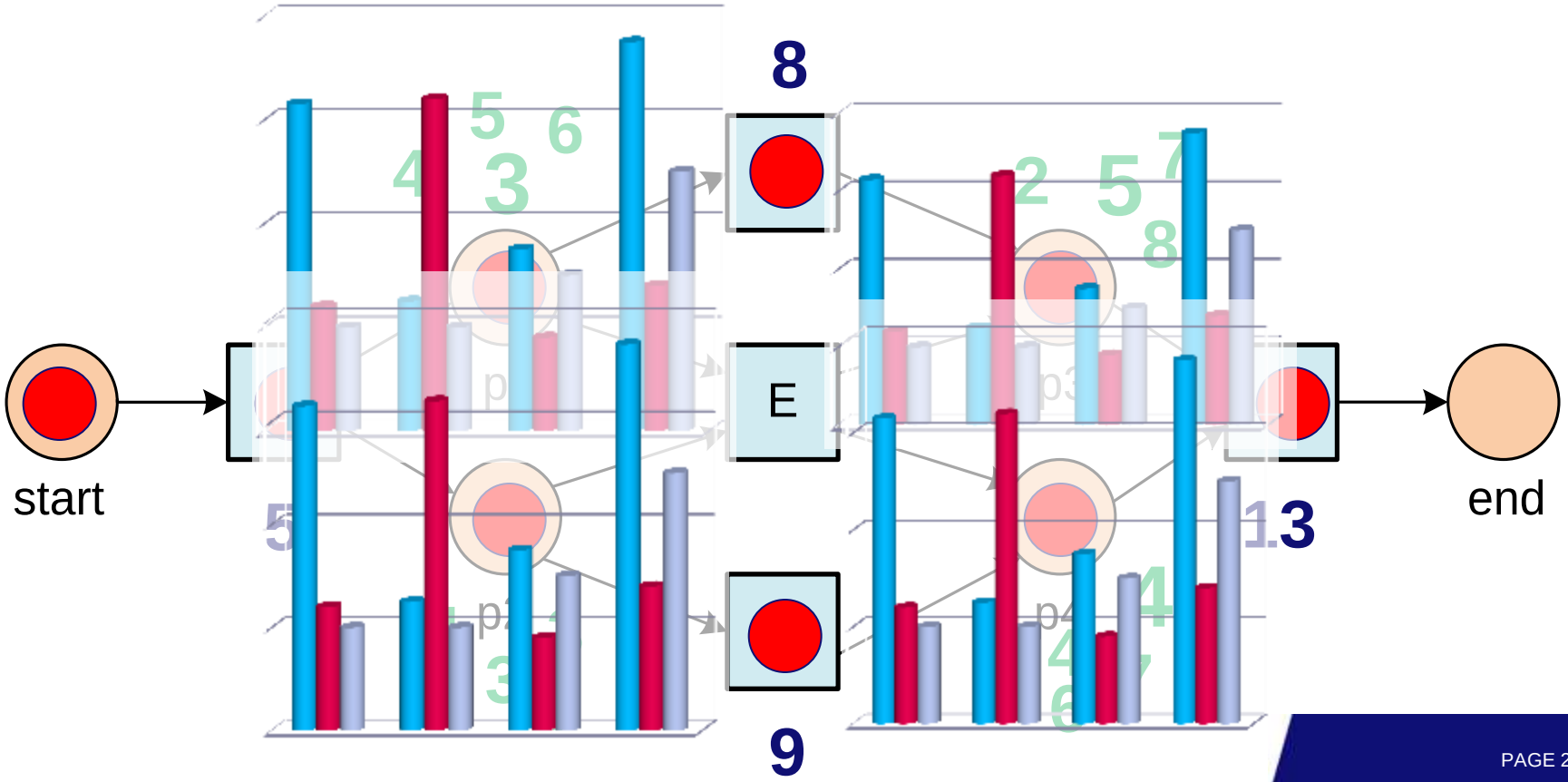
A C D




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

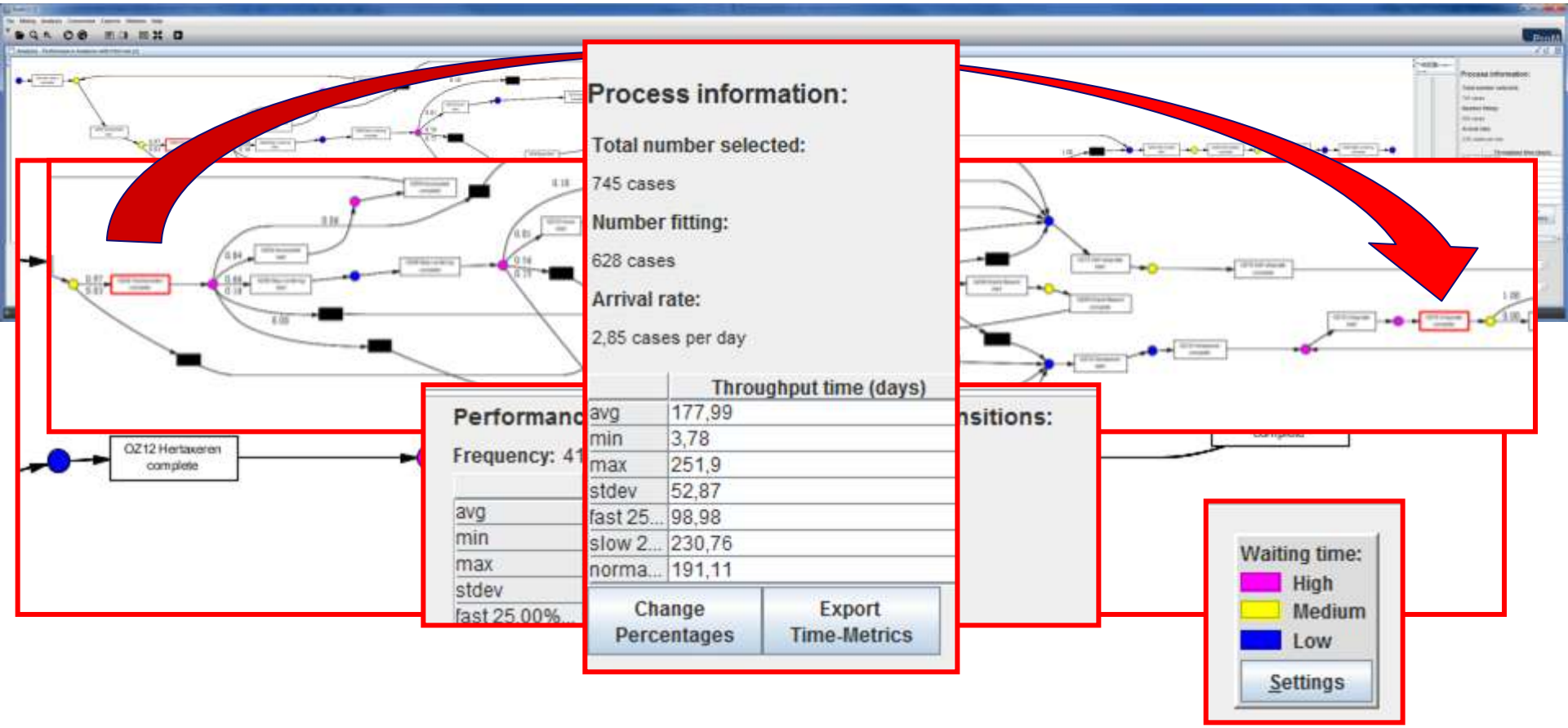
Replay can extract timing information

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

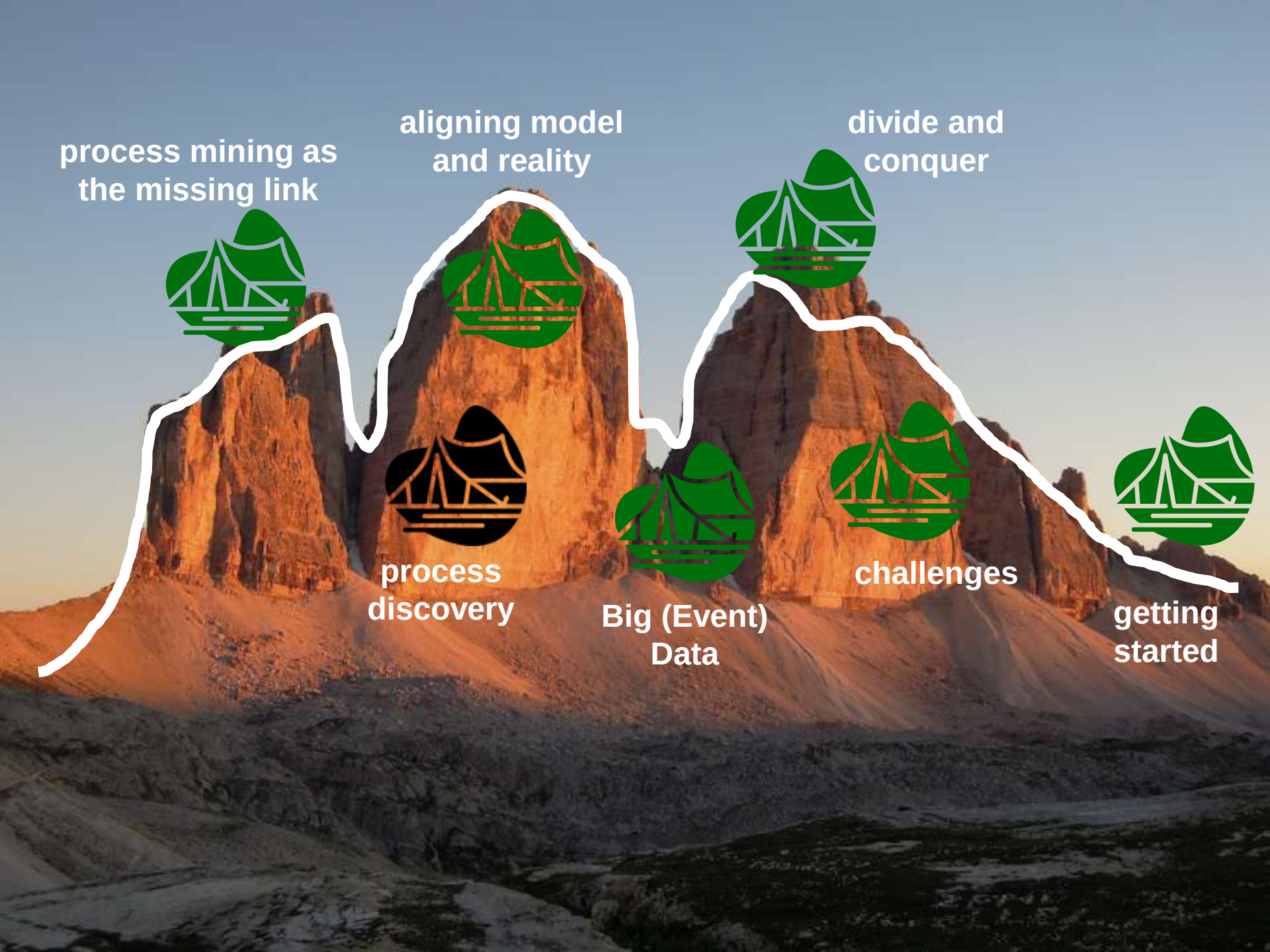


Performance Analysis Using Replay

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



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


process mining as the missing link
aligning model and reality
divide and conquer
process discovery
Big (Event) Data
challenges
getting started

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

Process discovery algorithms
(small selection)

distributed genetic mining

automata-based learning

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

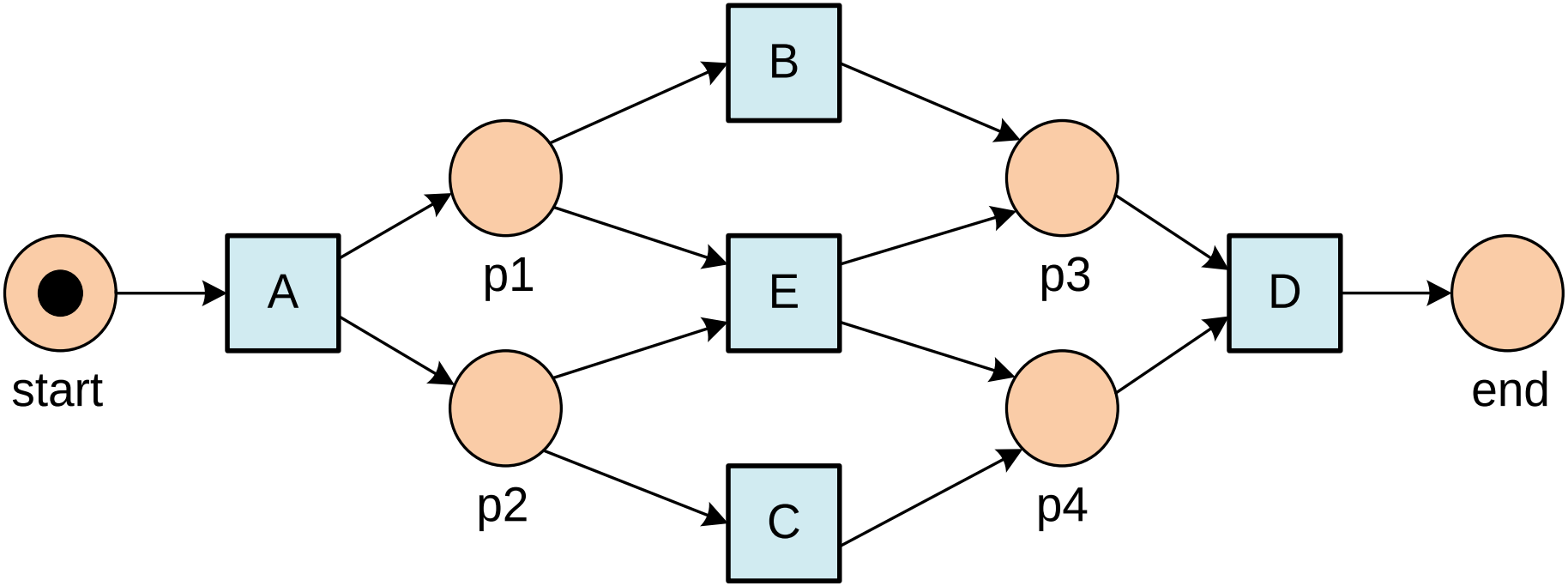
partial-order based mining

α# algorithm

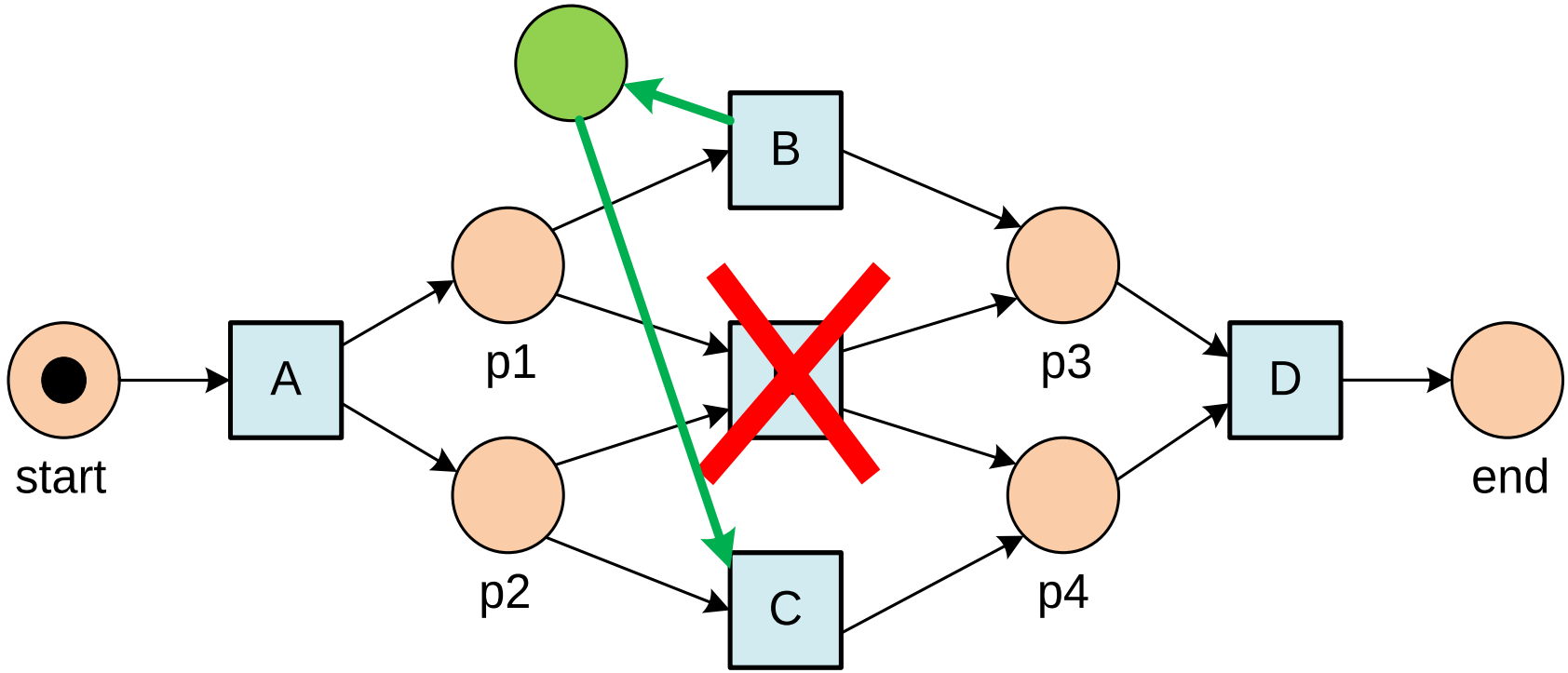
ILP mining

α++ algorithm

Quiz Question:
How to remove behavior?

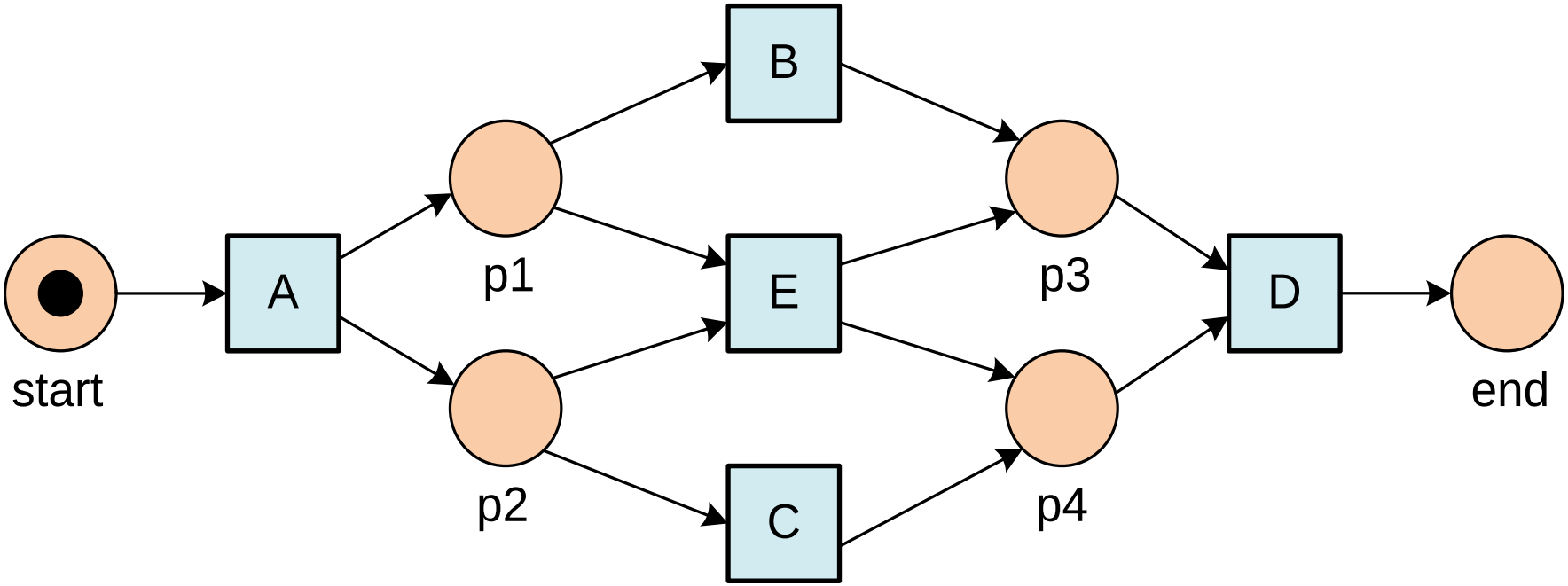


Quiz Question:
How to remove behavior?

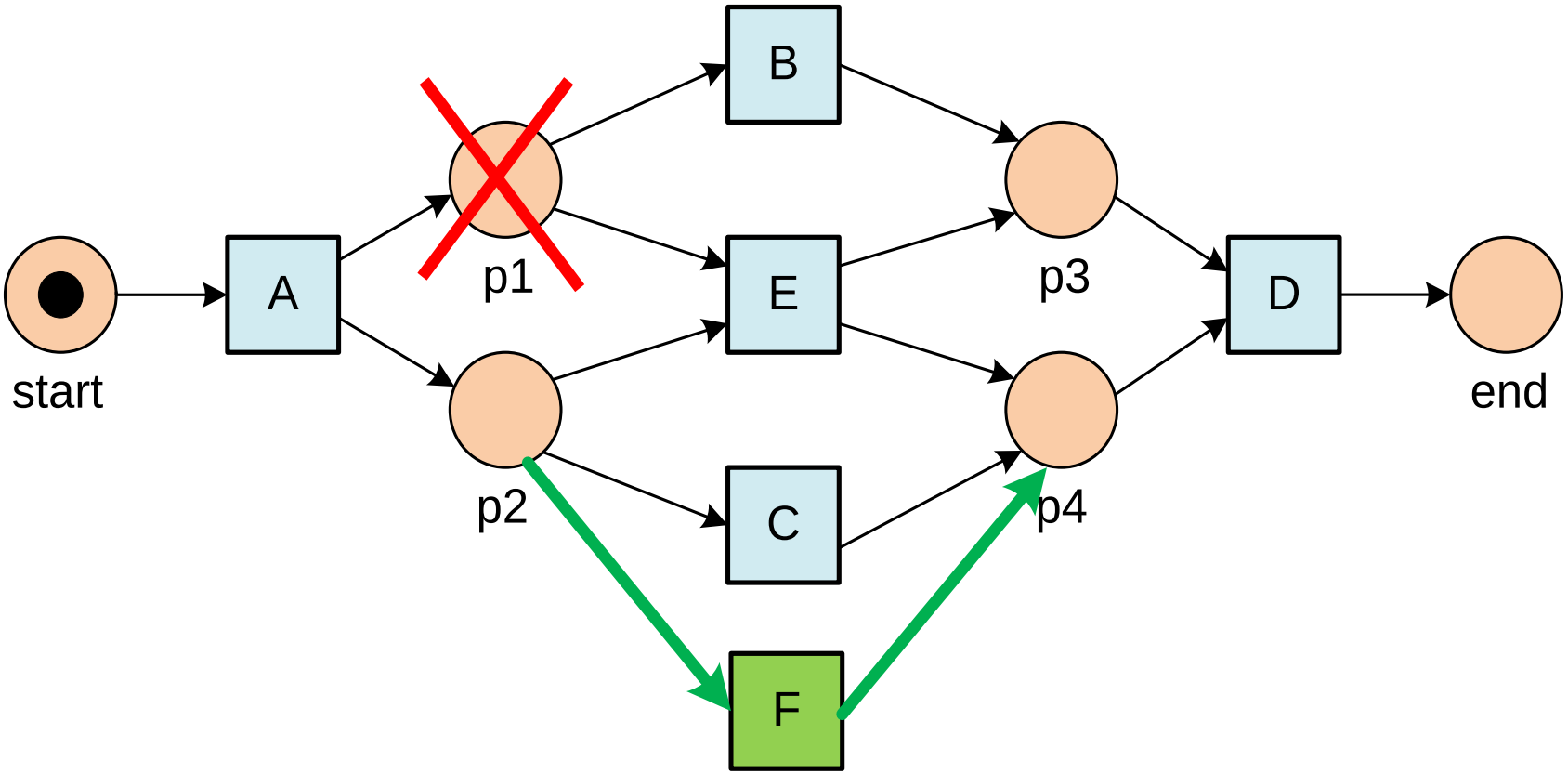


Add places or remove transitions!

Quiz Question:
How to add behavior?

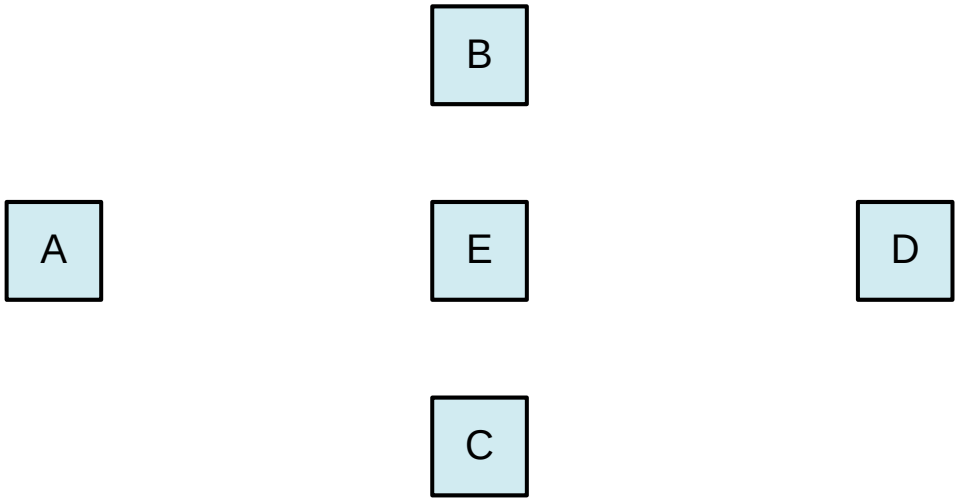


Quiz Question:
How to add behavior?



Add transition or remove places!

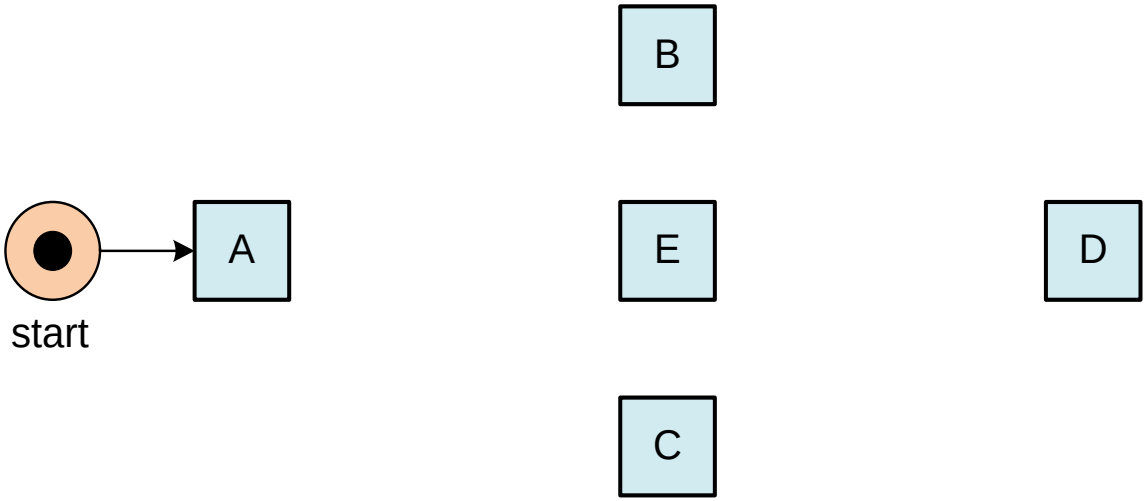
Places limit behavior



- **abcd**
- **ad**
- **abed**
- **abccd**
- **acbd**

- **aebcd**
- **aed**
- **aad**
- **caed**
- **aded**

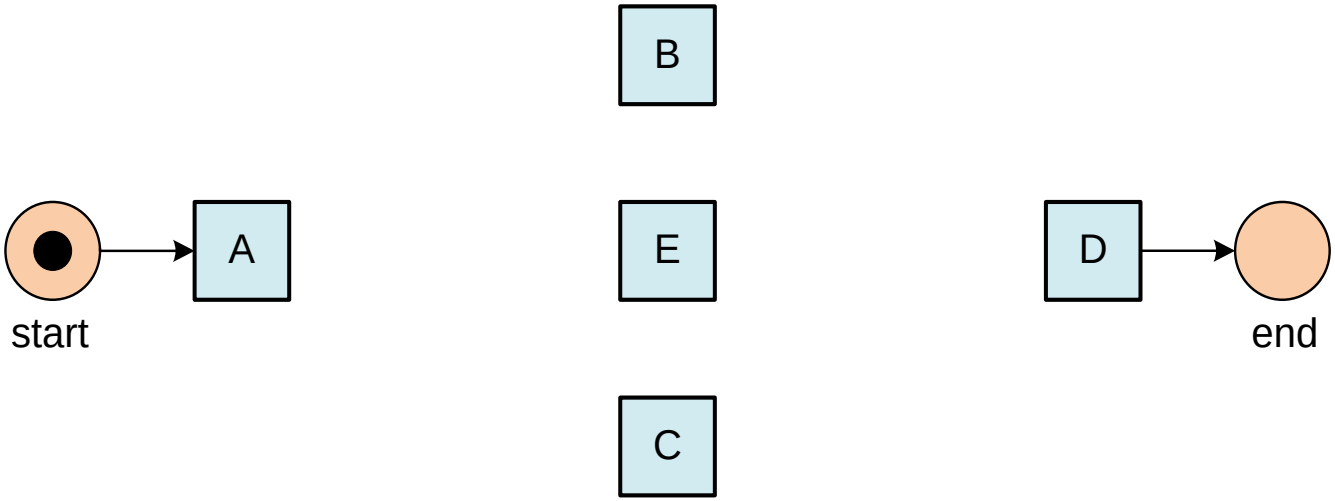
Places limit behavior



- abcd
- ad
- abed
- abccd
- acbd

- aeccd
- aed
- aad
- caed
- aded

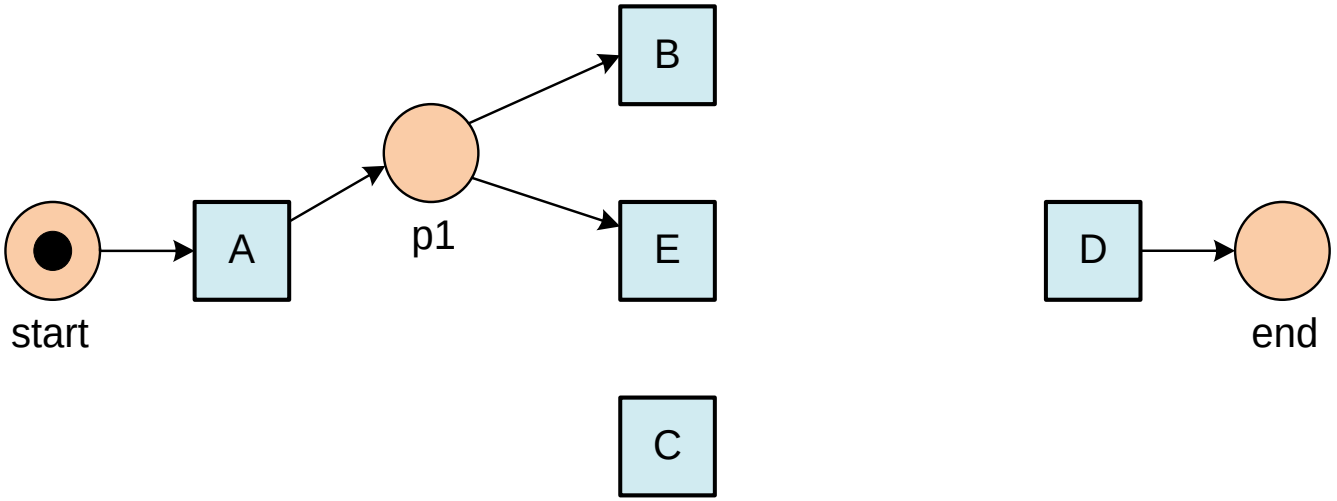
Places limit behavior



- abcd
- ad
- abed
- abccd
- acbd

- aebcd
- aed
- aad
- caed
- aded

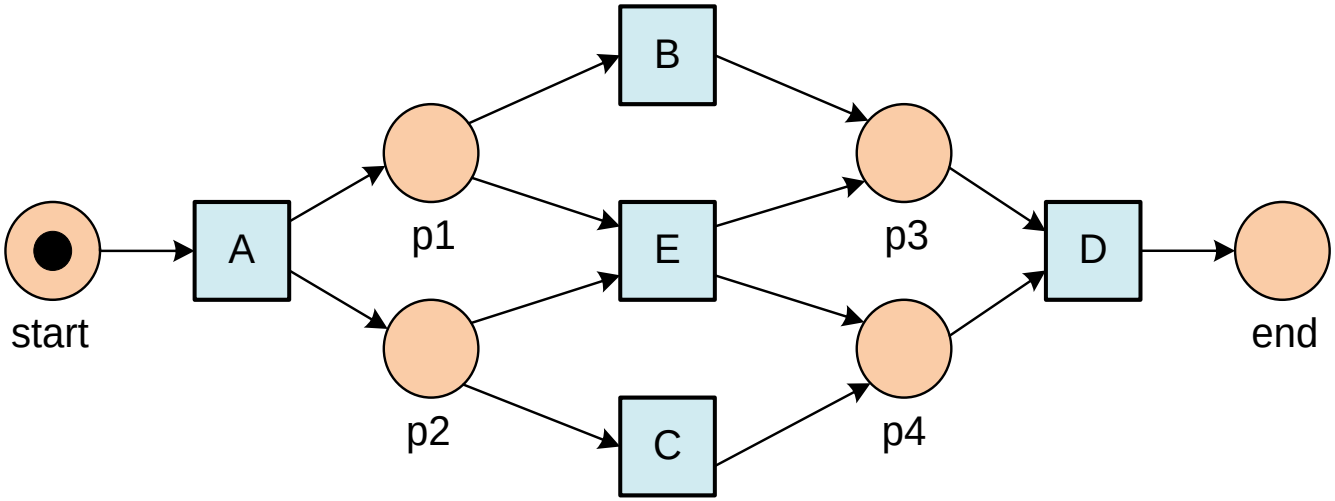
Places limit behavior



- abcd
- ad
- abed
- abccd
- acbd

- aebcd
- aed
- aad
- caed
- aded

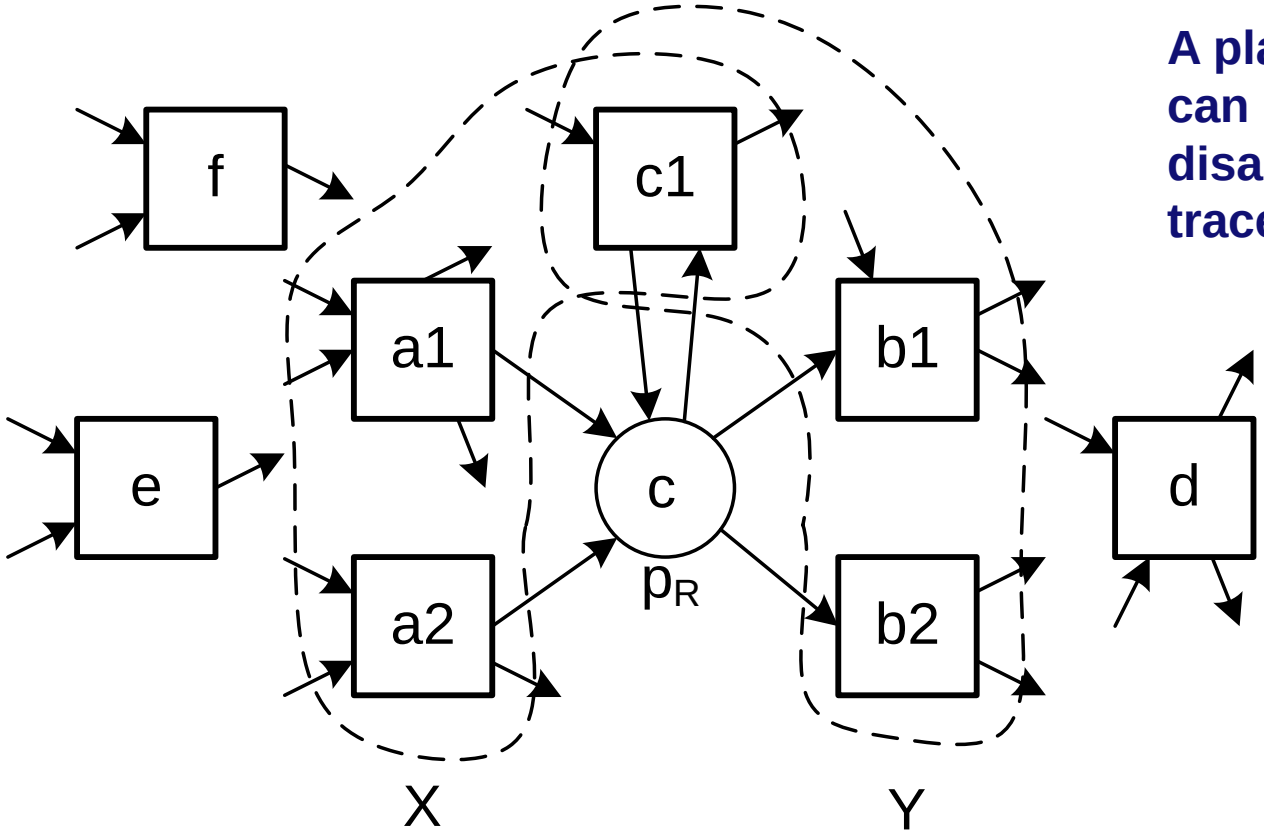
Places limit behavior



- abcd
- ad
- abed
- abccd
- acbd

- aebcd
- aed
- aad
- caed
- aded

Example: Process Discovery Using Language-Based Regions

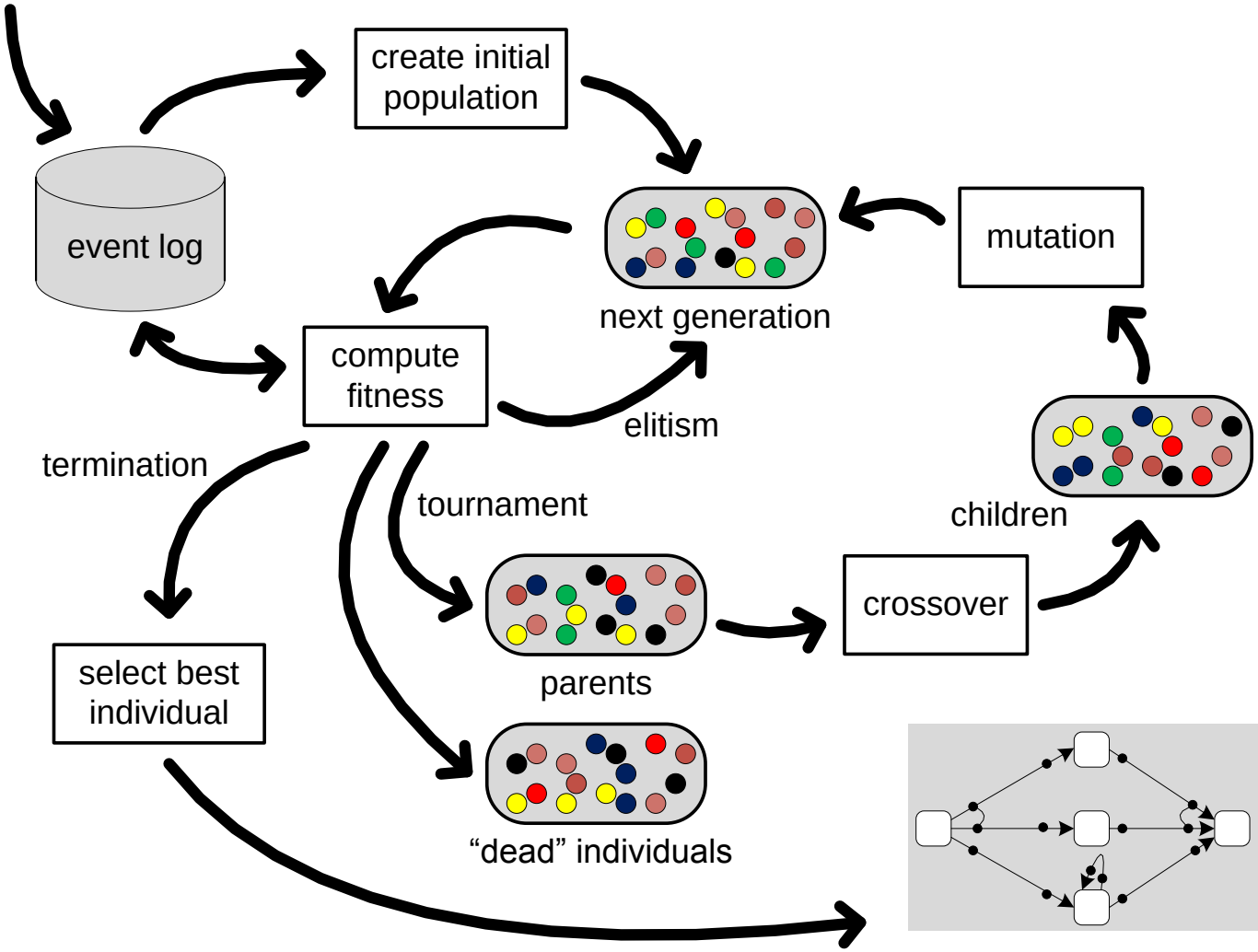


A place is feasible if it can be added without disabling any of the traces in the event log.

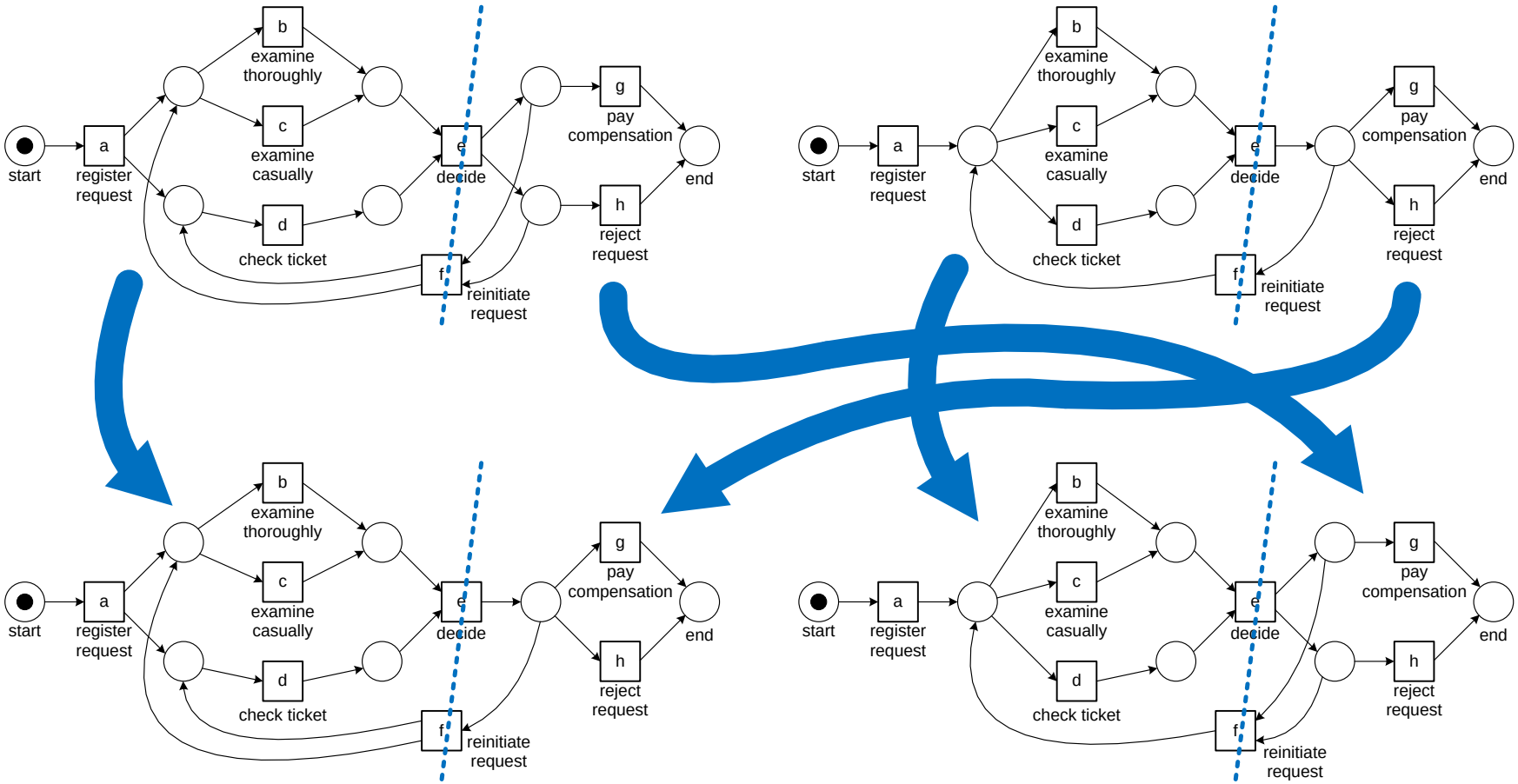
for any $\sigma \in L, k \in \{1, \ldots, |\sigma|\}, \sigma_1 = hd^{k-1}(\sigma), a = \sigma(k), \sigma_2 = hd^k(\sigma) = \sigma_1 \oplus a$:

$$c + \sum_{t \in X} \partial_{multiset}(\sigma_1)(t) - \sum_{t \in Y} \partial_{multiset}(\sigma_2)(t) \geq 0.$$

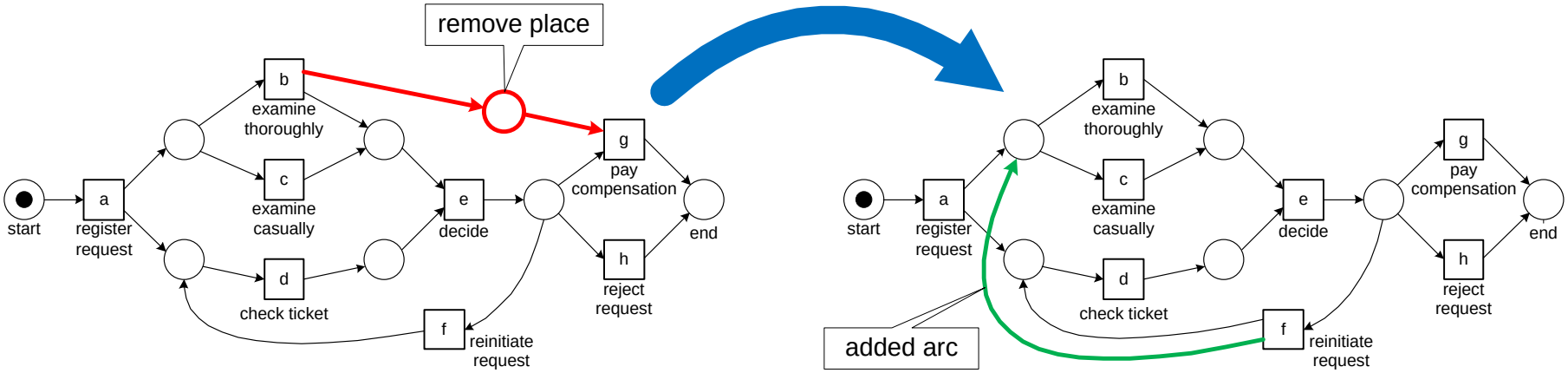
Genetic process mining: Overview

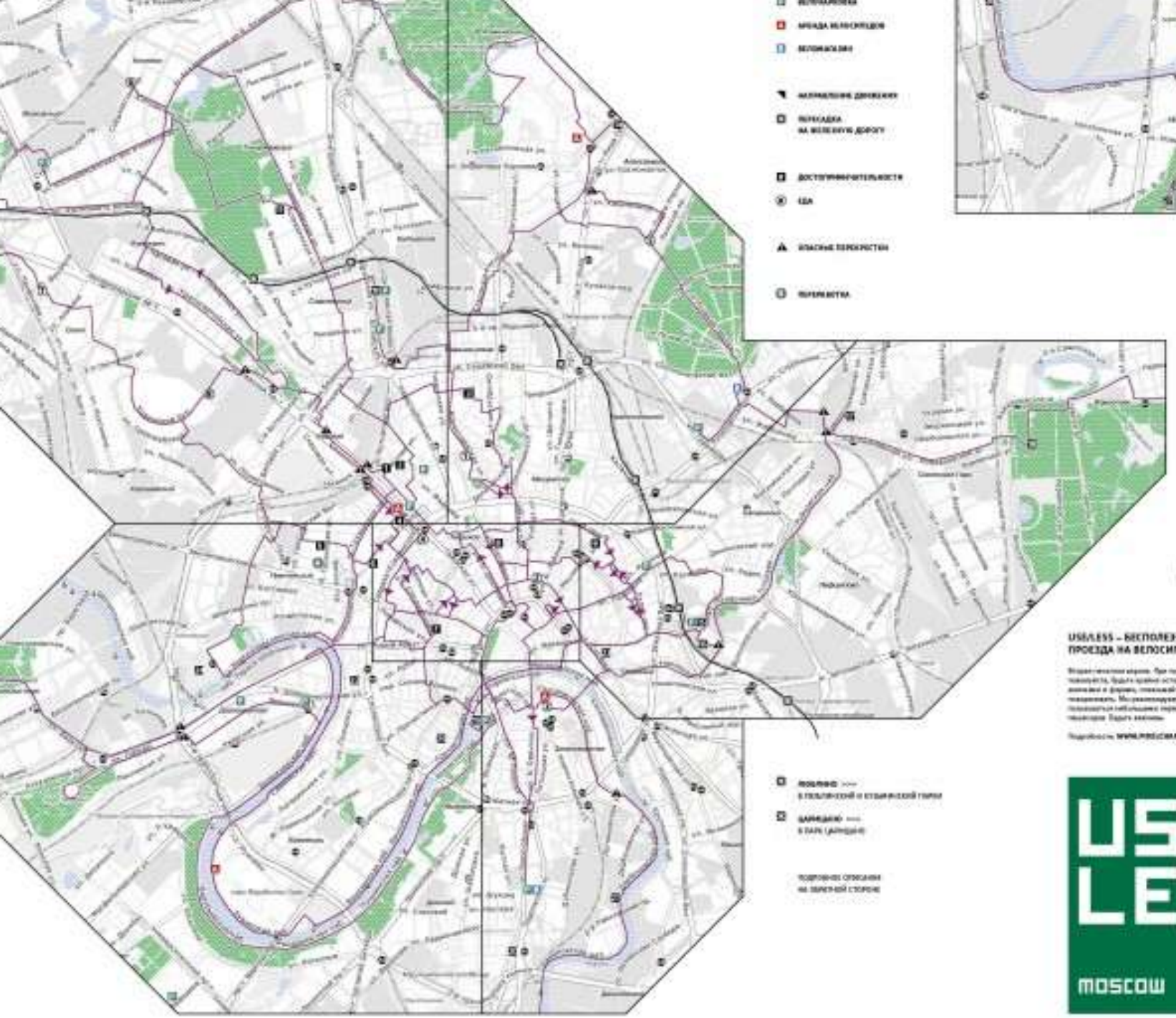


Example: crossover



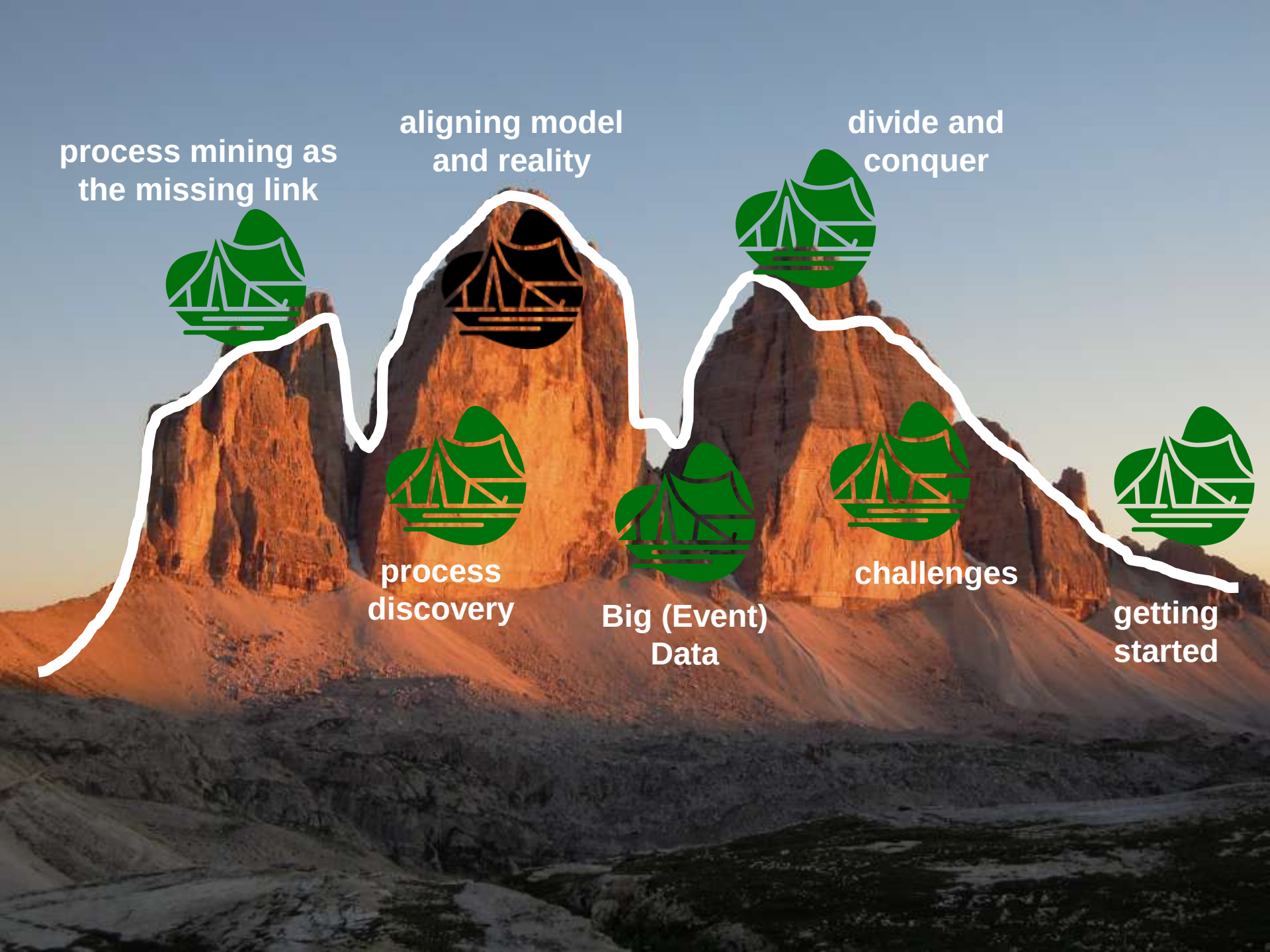
Example: mutation



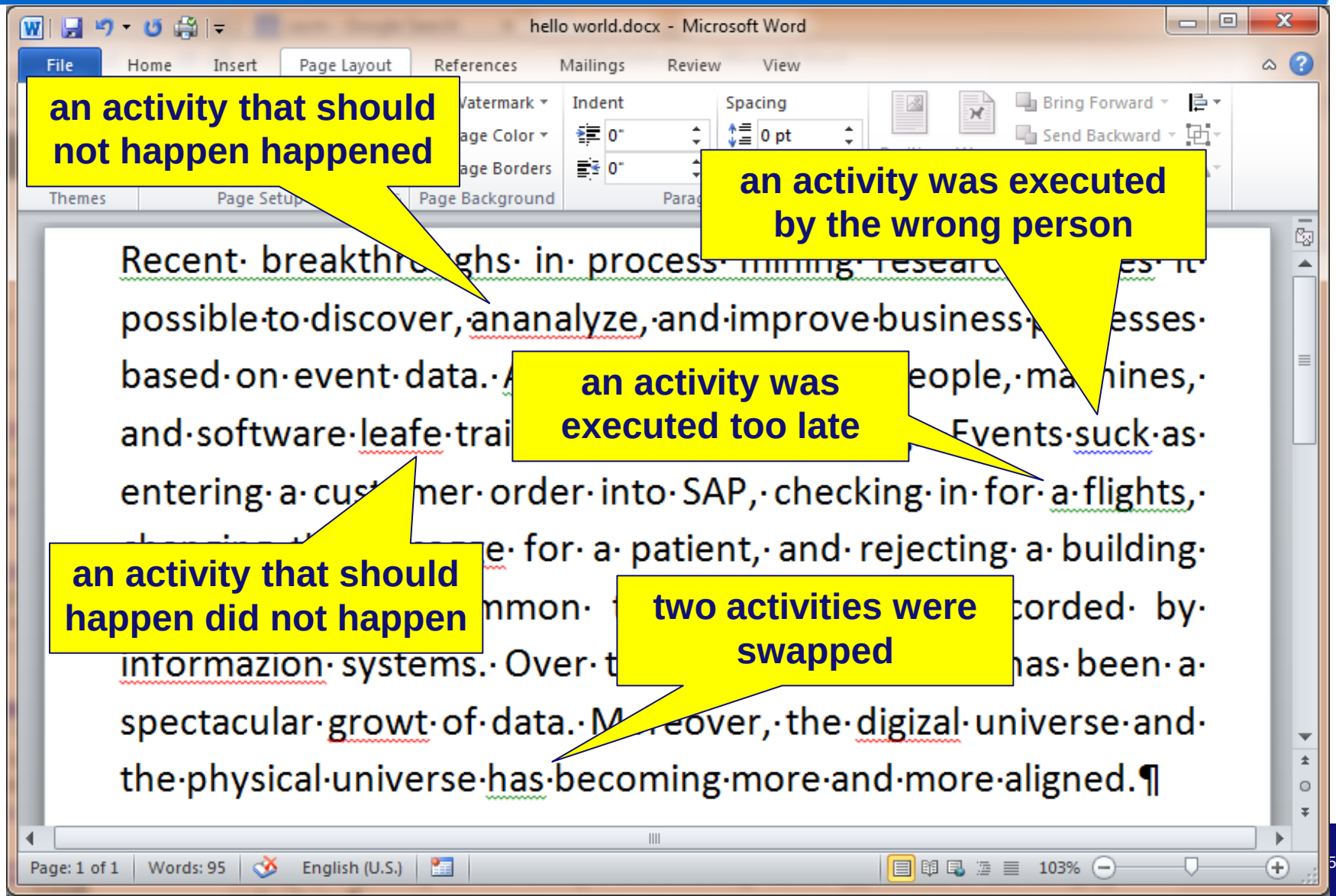


USELESS – БЕСПОЛЕЗНАЯ СХЕМА ПРОЕЗДА НА ВЕЛОСИПЕДЕ ПО МОСКВЕ

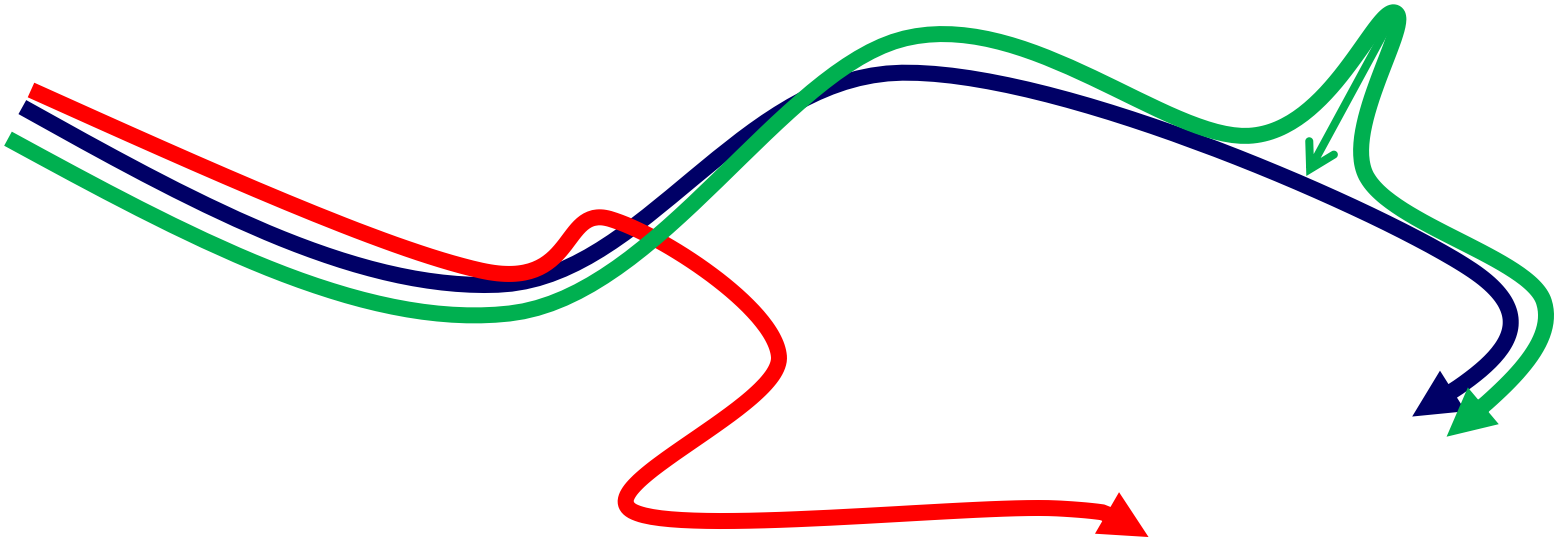
USE
LESS
MOSCOW


process mining as the missing link
aligning model and reality
divide and conquer
process discovery
Big (Event) Data
challenges
getting started

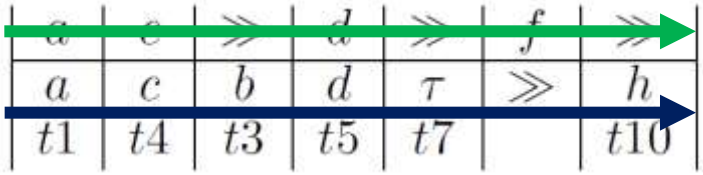
Conformance checking

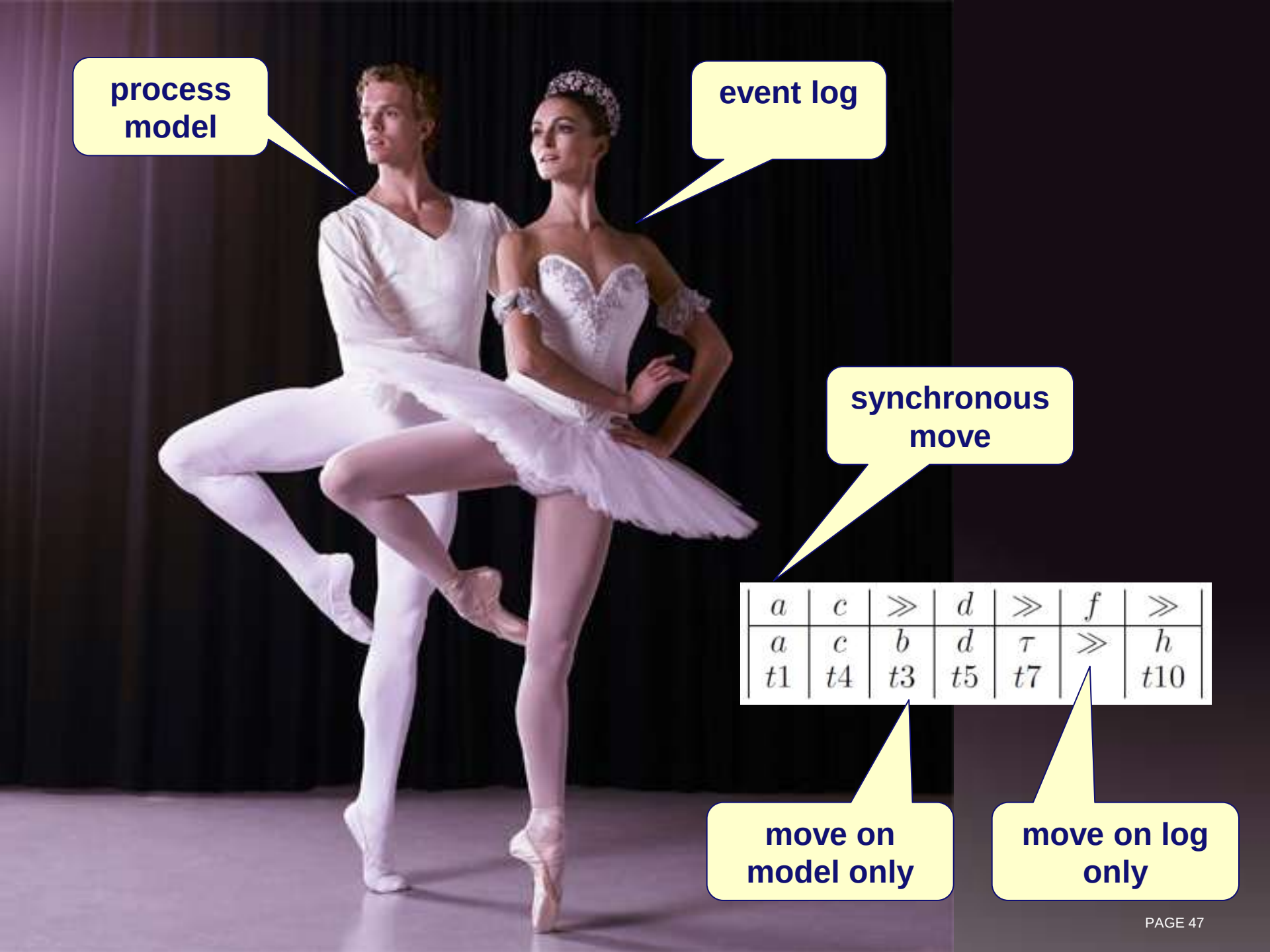


Alignments are essential!



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





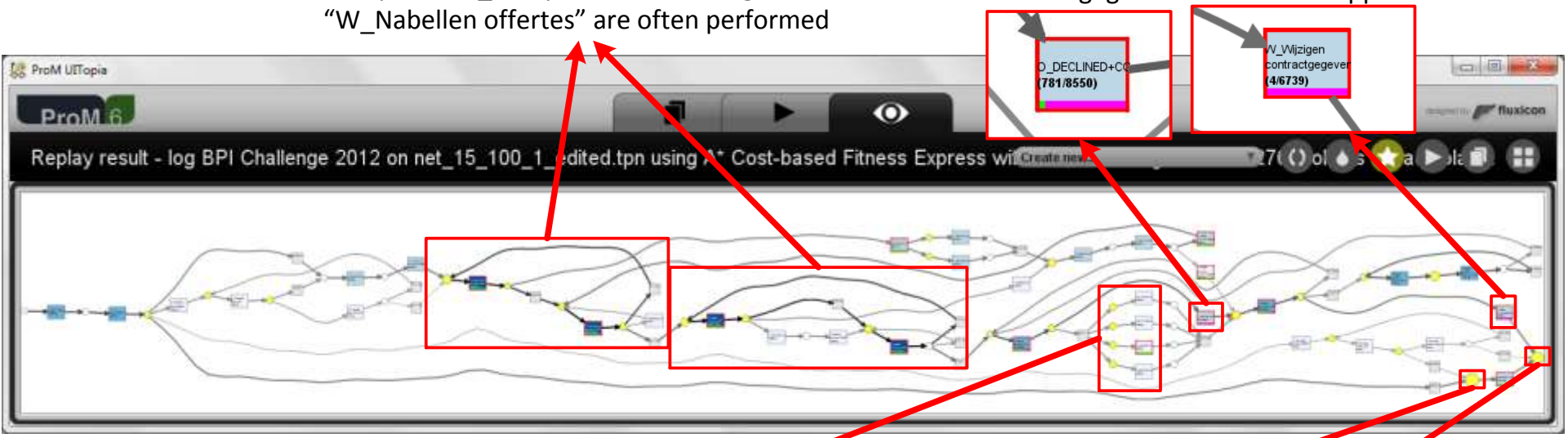
a	c	$\gg$	d	$\gg$	f	$\gg$
a	c	b	d	τ	$\gg$	h
$t1$	$t4$	$t3$	$t5$	$t7$		$t10$

Example: BPI Challenge 2012

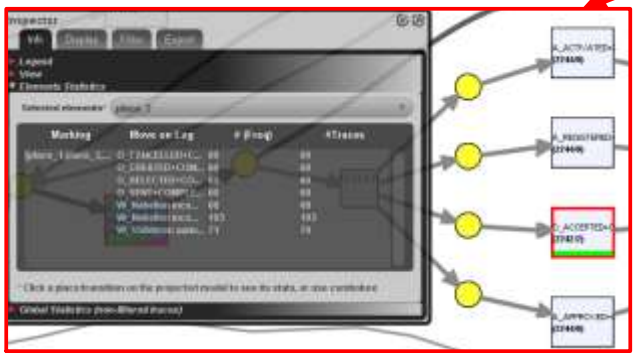
(Dutch financial institute, doi:10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f)

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

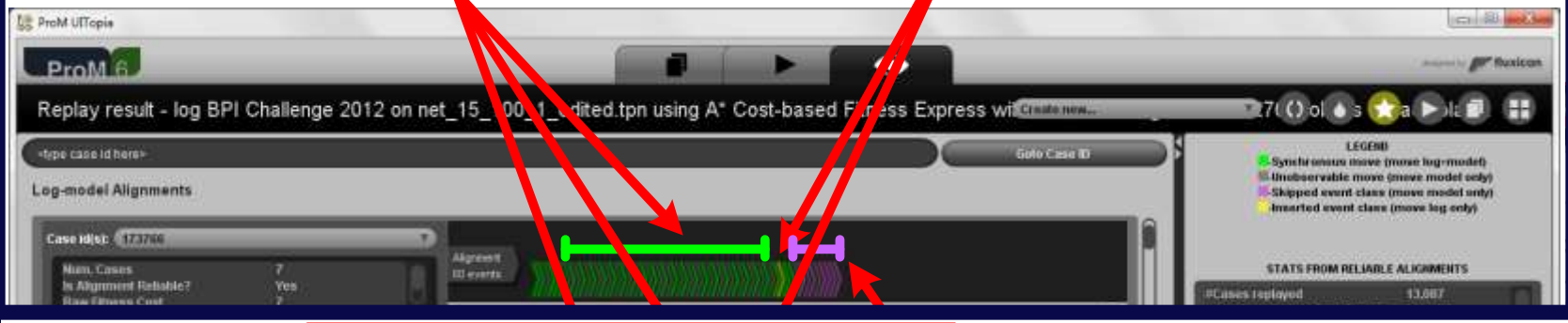


Many moves on log of “W_Afhandelen leads” (> 2200 times) occurred in the end of traces

Work of Arya Adriansyah (Replay project)

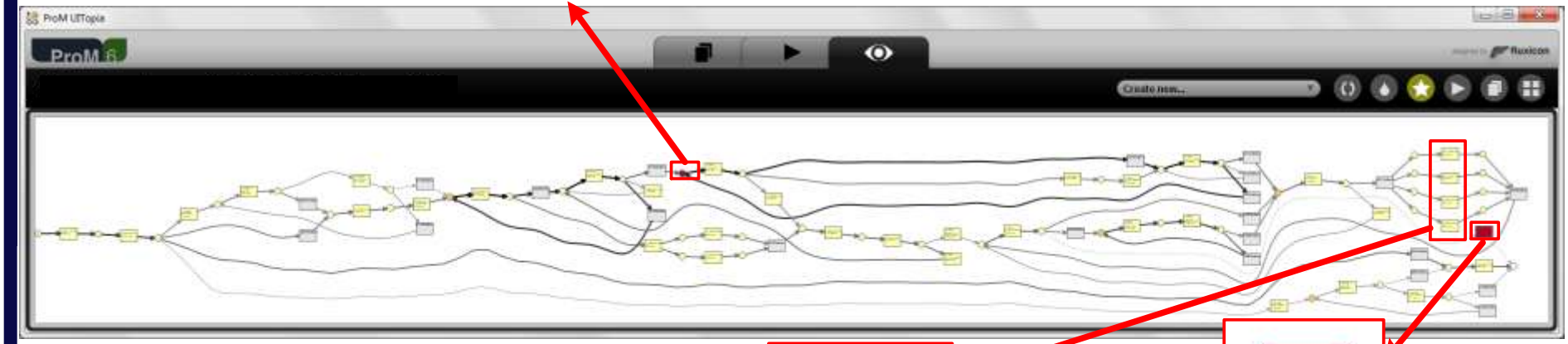
Synchronous moves of "Completeren aanvraag"

Move on log of "Completeren aanvraag"

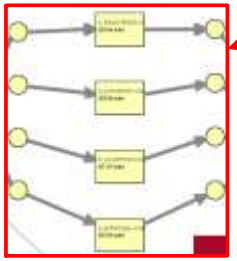


Property	Min.	Max.	Avg.	Std. Dev	Freq.
Waiting time	0.00 ms	29.78 days	2.83 days	3.30 days	24,229
Synchronization time	0.00 ms	0.00 ms	0.00 ms	0.00 ms	24,229
Sojourn time	0.00 ms	29.78 days	2.83 days	3.30 days	24,229

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

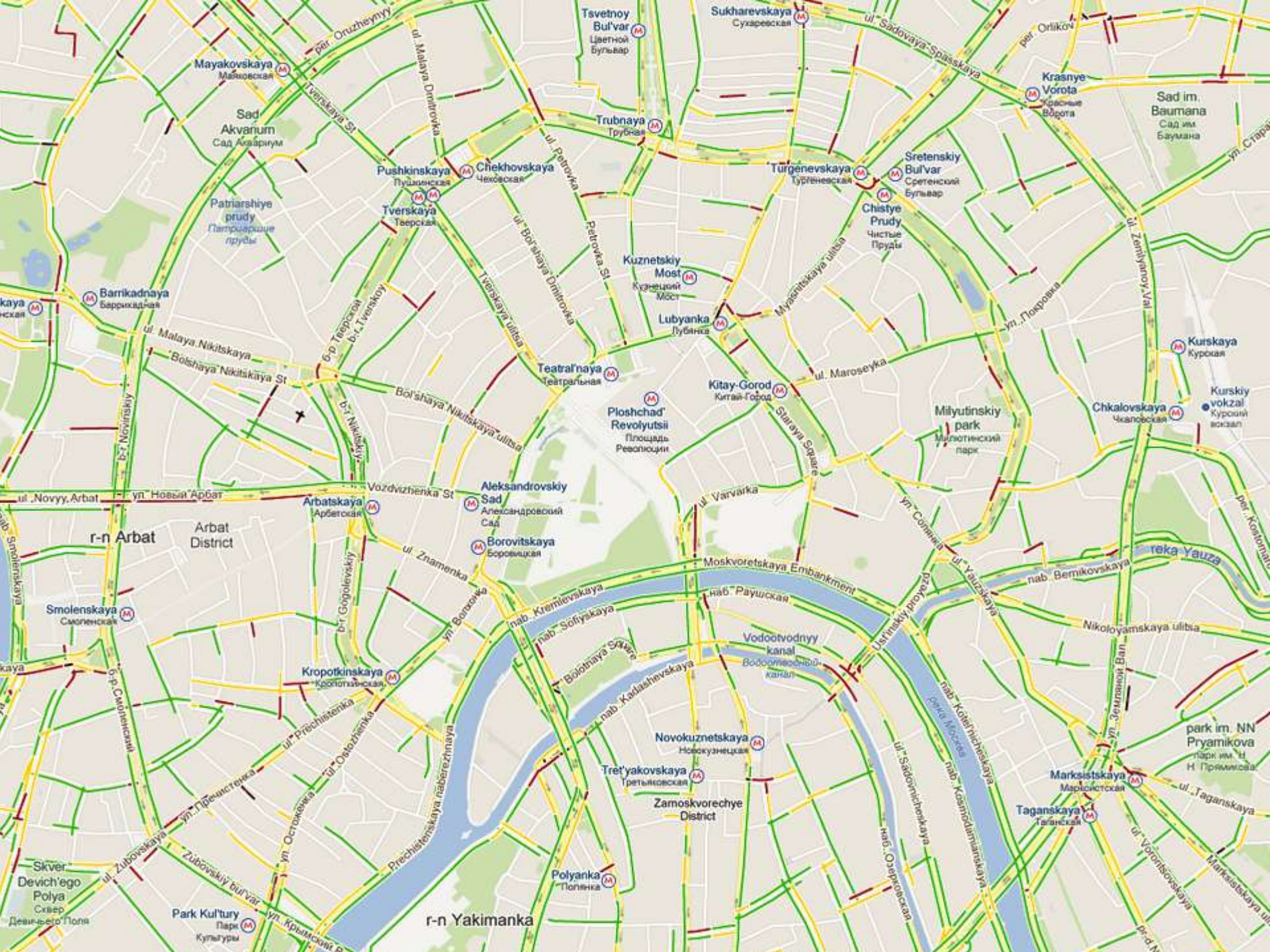


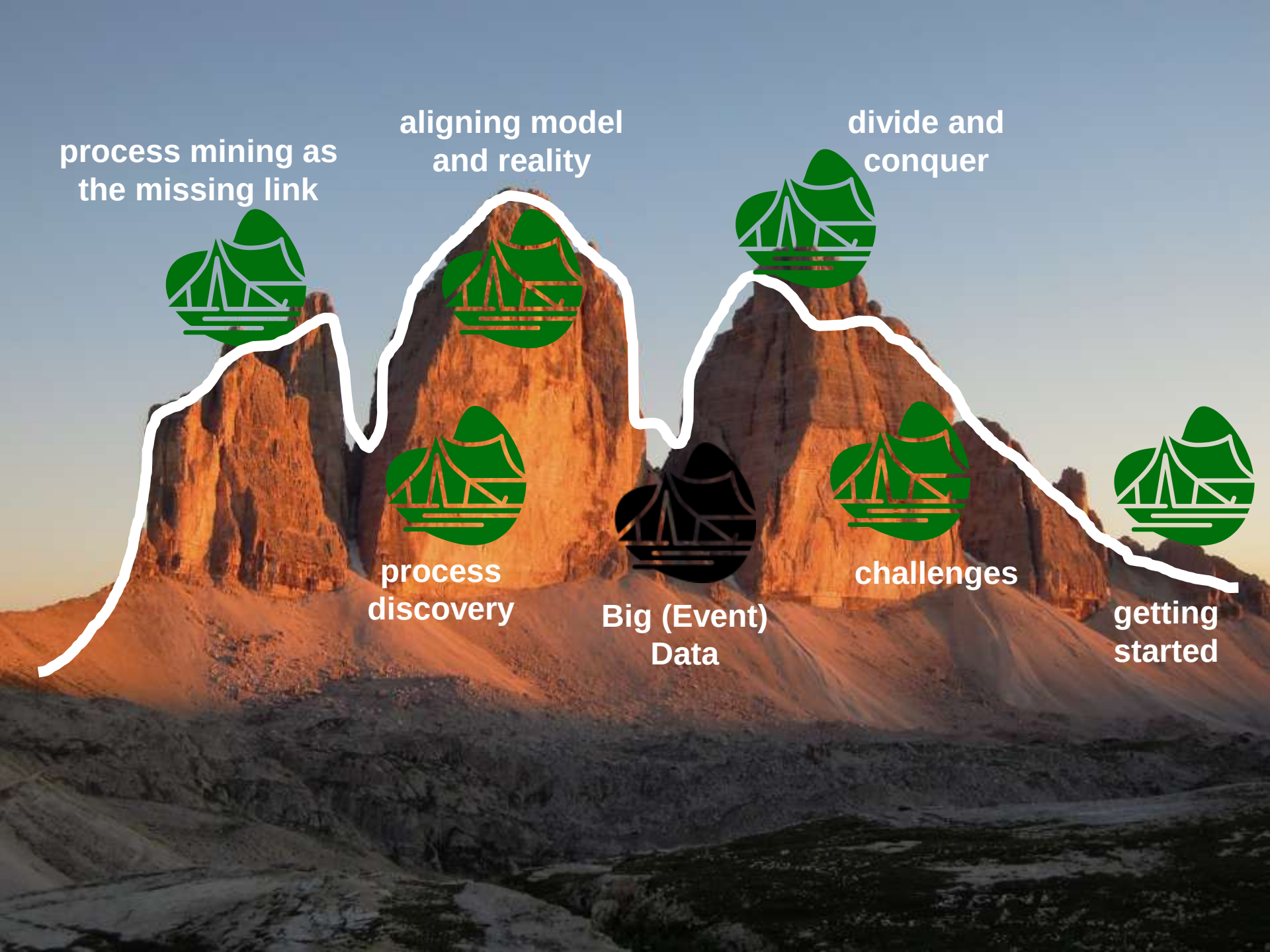
"O_ACCEPTED" has average sojourn time of 27.07 minutes, while "A_REGISTERED", "A_ACTIVATED", and "A_APPROVED" have average sojourn time of 29.56 minutes

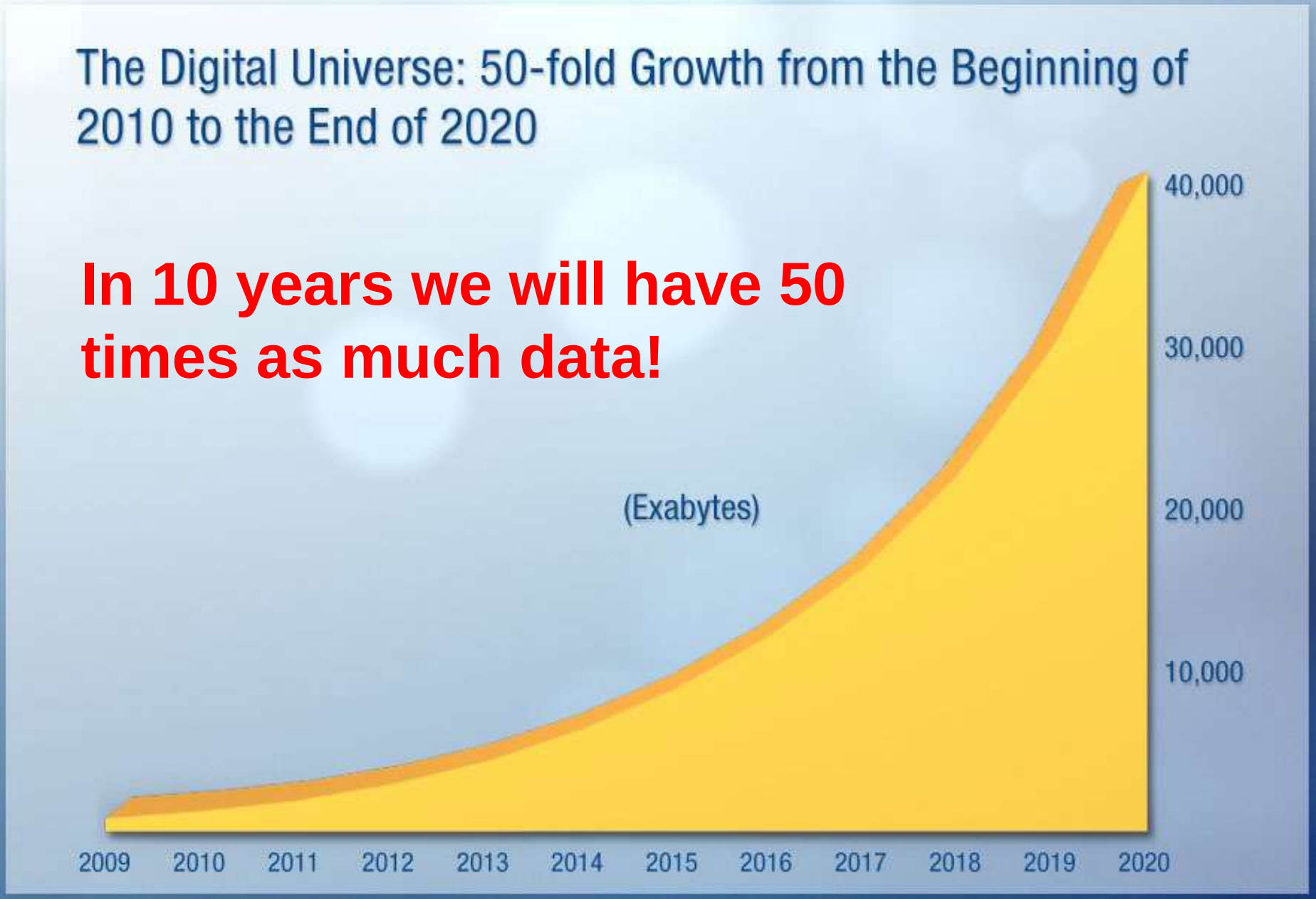


Property	Min.	Max.	Avg.	Std. Dev	Freq.
Throughput time	0.00 ms	0.00 ms	0.00 ms	0.00 ms	4
Waiting time	1.55 hours	3.43 months	1.14 months	1.55 months	4
Sojourn time	1.55 hours	3.43 months	1.14 months	1.55 months	4
#Unique cases ...	4				

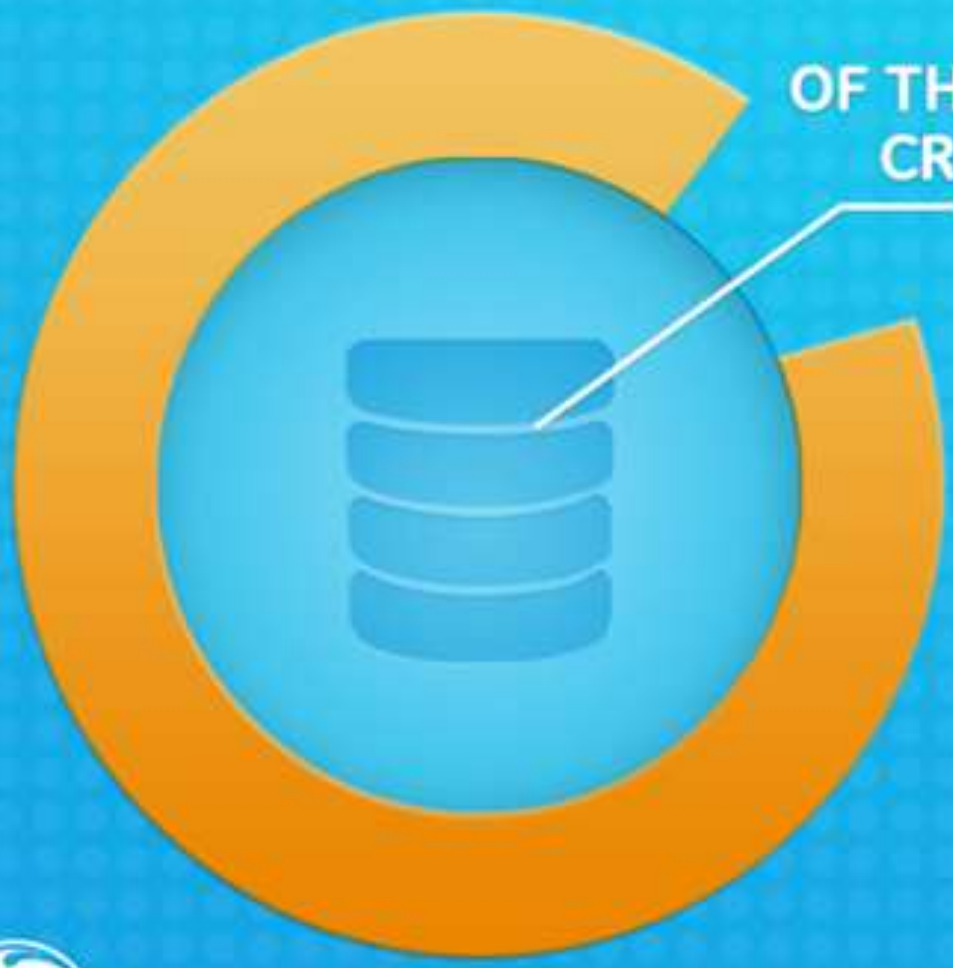
Activity "W_Wijzigen contractgegevens" is the bottleneck, but it occured rarely (only 4 times)


Tsvetnoy Bul'var
Цветной Бульвар
Sukharevskaya
Сухаревская
ul. Sadovaya-Spasskaya
per. Orlikov
Mayakovskaya
Маяковская
per. Oruzheynyy
ul. Malaya Dmitrovka
Tverskaya St
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
Myasnitskaya ulitsa
ul. Zemlyanoy Val
Barrikadnaya
Баррикадная
Tverskaya ulitsa
Lubyanka
Лубянка
ul. Pokrovka
ul. Malaya Nikitskaya
Bol'shaya Nikitskaya St
b-r. Tverskoy
Teatral'naya
Театральная
ul. Maroseyka
Kitay-Gorod
Китай-Город
Kurskaya
Курская
Ploshchad' Revolyutsii
Площадь Революции
Staraya Square
Milyutinskiy park
Милютинский парк
Chkalovskaya
Чкаловская
Kurskiy vokzal
Курский вокзал
Bol'shaya Nikitskaya ulitsa
b-r. Novinskiy
b-r. Nikitskiy
ul. Novyy Arbat
ул. Новый Арбат
Vozdvizhenka St
Aleksandrovskiy Sad
Александровский Сад
ul. Varvarka
Arbatskaya
Арбатская
r-n Arbat
Arbat District
Borovitskaya
Боровицкая
ul. Znamenka
Moskvoretskaya Embankment
nab. Bersenevskaya
nab. Bersenevskaya
nab. Bersenevskaya
Smolenskaya
Смоленская
b-r. Gogolevskiy
nab. Kremlevskaya
nab. Sofiyskaya
nab. Raushskaya
Vodootvodnyy kanal
Bolotnaya Square
Kropotkinskaya
Кропоткинская
nab. Kadashevskaya
Nikoloyamskaya ulitsa
ul. Prechistenka
ul. Ostozhenka
Novokuznetskaya
Новокузнецкая
Tret'yakovskaya
Третьяковская
Zamoskvorechye District
Prechistenskaya naberezhnaya
nab. Kotel'nicheskaya
nab. Kosmodamianskaya
park im. NN Pryamikova
Marksistskaya
Марксистская
Taganskaya
Таганская
ul. Taganskaya
Sad Devich'ego Polya
Сквер Девичьего Поля
ul. Zubovskaya
Zubovskiy bul'var
Polyanka
Park Kul'tury
Парк Культуры
r-n Yakimanka
Moskva River
Yauza River


process mining as the missing link
aligning model and reality
divide and conquer
process discovery
Big (Event) Data
challenges
getting started



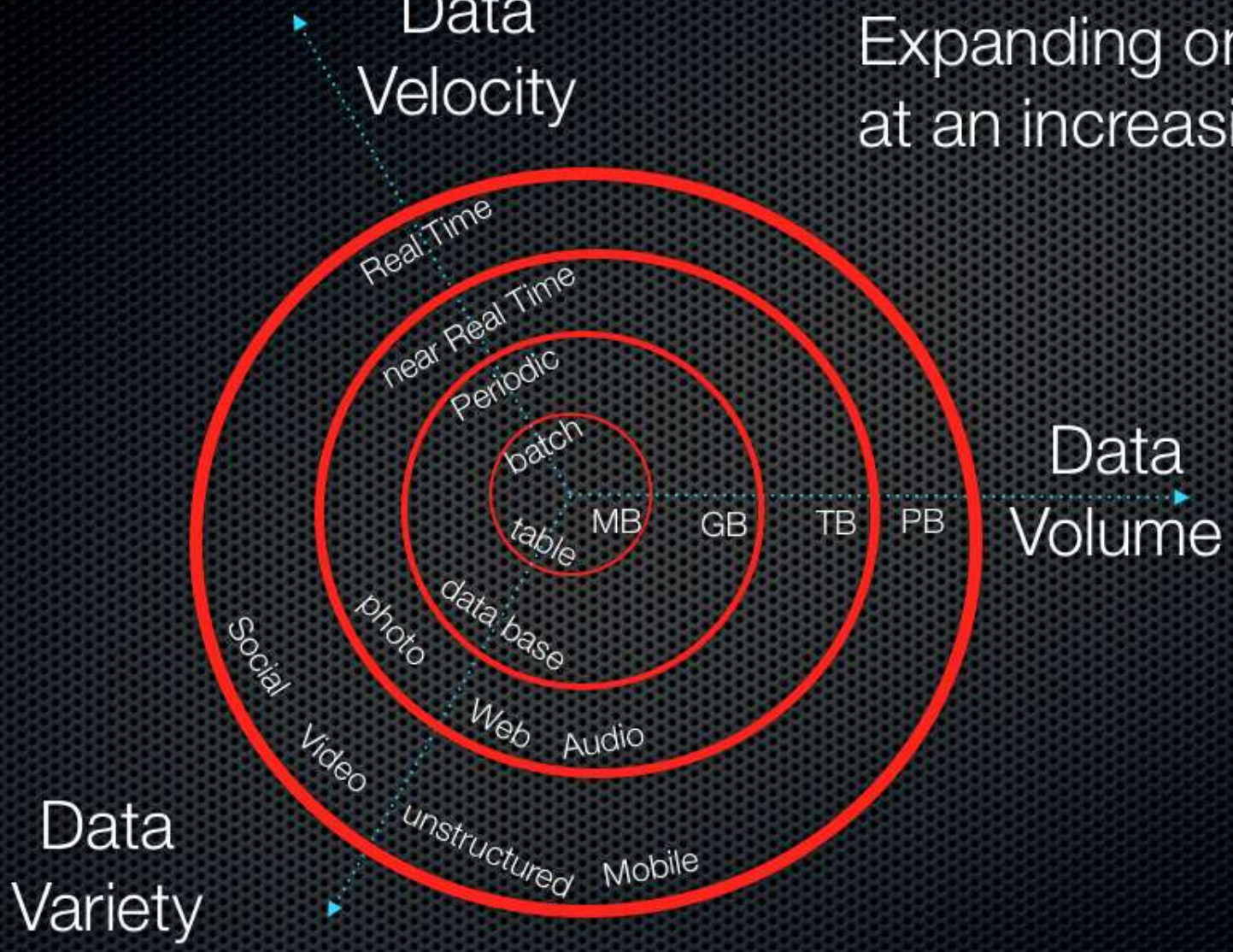
Source: IDC's Digital Universe Study, sponsored by EMC, December 2012



90%

OF THE WORLD'S DATA WAS CREATED IN THE LAST **TWO YEARS**

Big Data: Expanding on 3 fronts at an increasing rate.



Acquisition → Marshalling → Analysis → Action

Data Acquisition

IBM, SAP, sas, TIBCO, hp, Ab Initio, Microsoft, KALIDO, ORACLE, talend, INFORMATICA, splunk>, Numenta, syncsort

VLDW and BI Appliances

sas, ORACLE, IBM, SAP, hp, kognitio, Actian, Microsoft, ParAccel, EMC², TERADATA

Analytics

EMC², sas, hp, IBM, R, hadoop, Adobe, SAP, MZinga, salesforce, GoodData, KARMASPHERE, Microsoft, kxen, TERADATA, ORACLE

BPM & Action

TIBCO, EMC², Adobe, ORACLE, PROGRESS SOFTWARE, Pega Pegasystems Inc., software AG, IBM, SAP, OPEN TEXT, Microsoft, salesforce, TERADATA, ELOQUA, iGrafx

Including Complex Event Processing (CEP) tools

Data Providers

LexisNexis, comScore, Windows Azure Marketplace DataMarket, INRIX, nielsen, REUTERS, factual, SymphonyIRI Group, kaggle, GNIP, OPEN DATA, GfK, salesforce, radian6, Experian A world of insight, ims, DATASIFT, knoema

And all your own data

And your partners data

No SQL

hadoop, ORACLE, cloudera, EMC², IBM, Hortonworks, Google, hp, MAPR TECHNOLOGIES, Microsoft, mongoDB, snapLogic, CouchDB, amazon web services, Palantir, splunk>, Neo4j the graph database, MarkLogic, [kx], sparsity, Cassandra, TERADATA

Content Management

Adobe, ORACLE, Alfresco, EMC², IBM, OPEN TEXT, hp, Microsoft

Data Virtualization

COMPOSITE SOFTWARE, IBM, Microsoft, denodo, Tejid, SAP, ORACLE, hp, INFORMATICA

Capgemini - Capping IT off

Manuel Sevilla - 2012

BI Tools

ORACLE, Microsoft, GoodData, PANORAMA, IBM, SAP, panopticon Real-Time Visual Data Analysis, LogiXML, Information Builders, hp, Roambi, sas, pentaho, MicroStrategy, Google, SpagoBI, tableau, QlikView, DOMO, ESRI, TIBCO, DASSAULT SYSTEMES, neutrinoBI, ACTUATE

Data Governance

IBM, SAP, KALIDO, Microsoft, ORACLE, adaptive, INFORMATICA, TIBCO, hp, talend open data solutions, sas, Orchestra NETWORKS

The Economist

FEBRUARY 27TH–MARCH 5TH 2010 Economist.com

Obama the warrior

Misgoverning Argentina

The economic shift from West to East

Genetically modified crops blossom

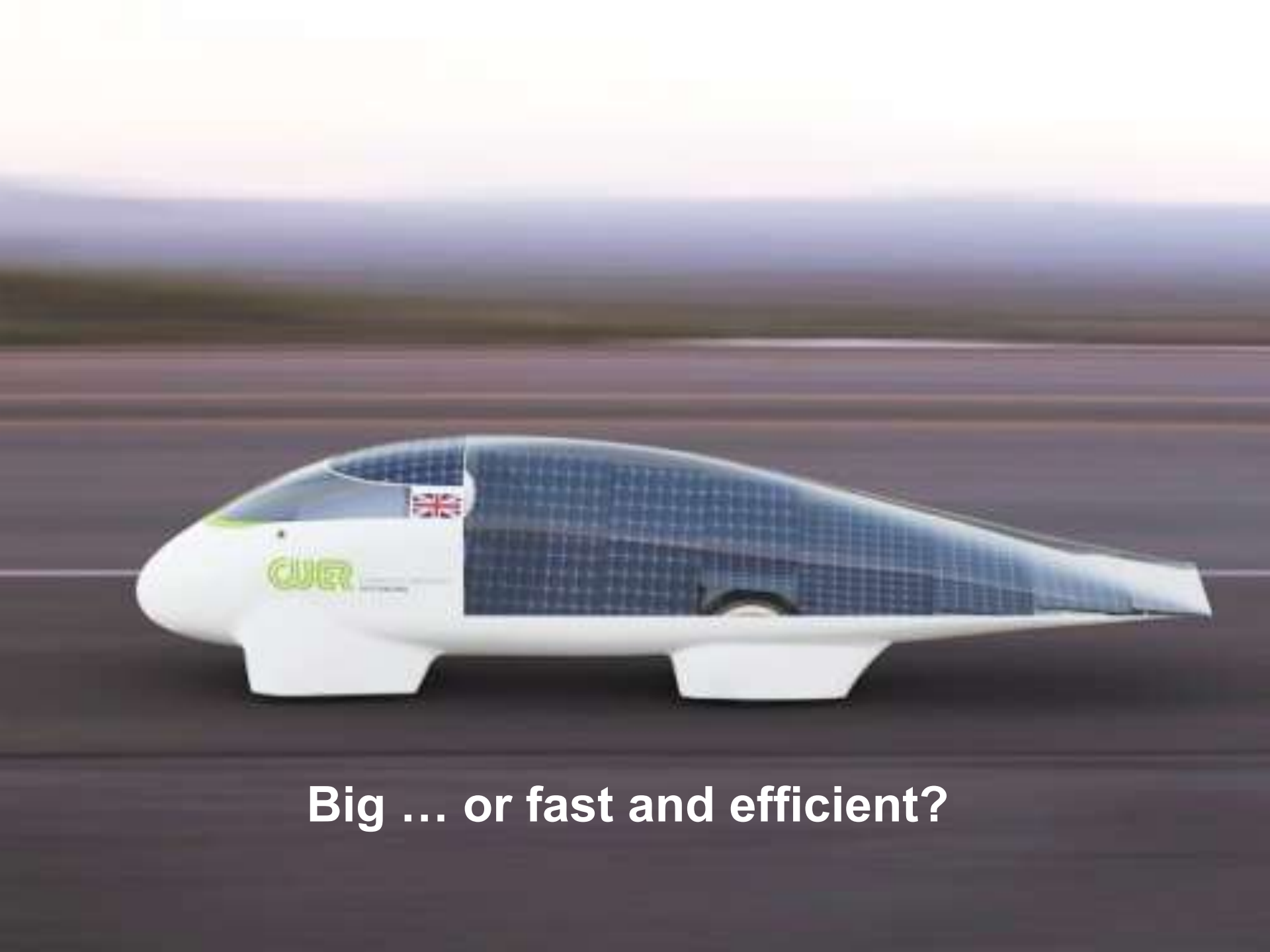
The right to eat cats and dogs

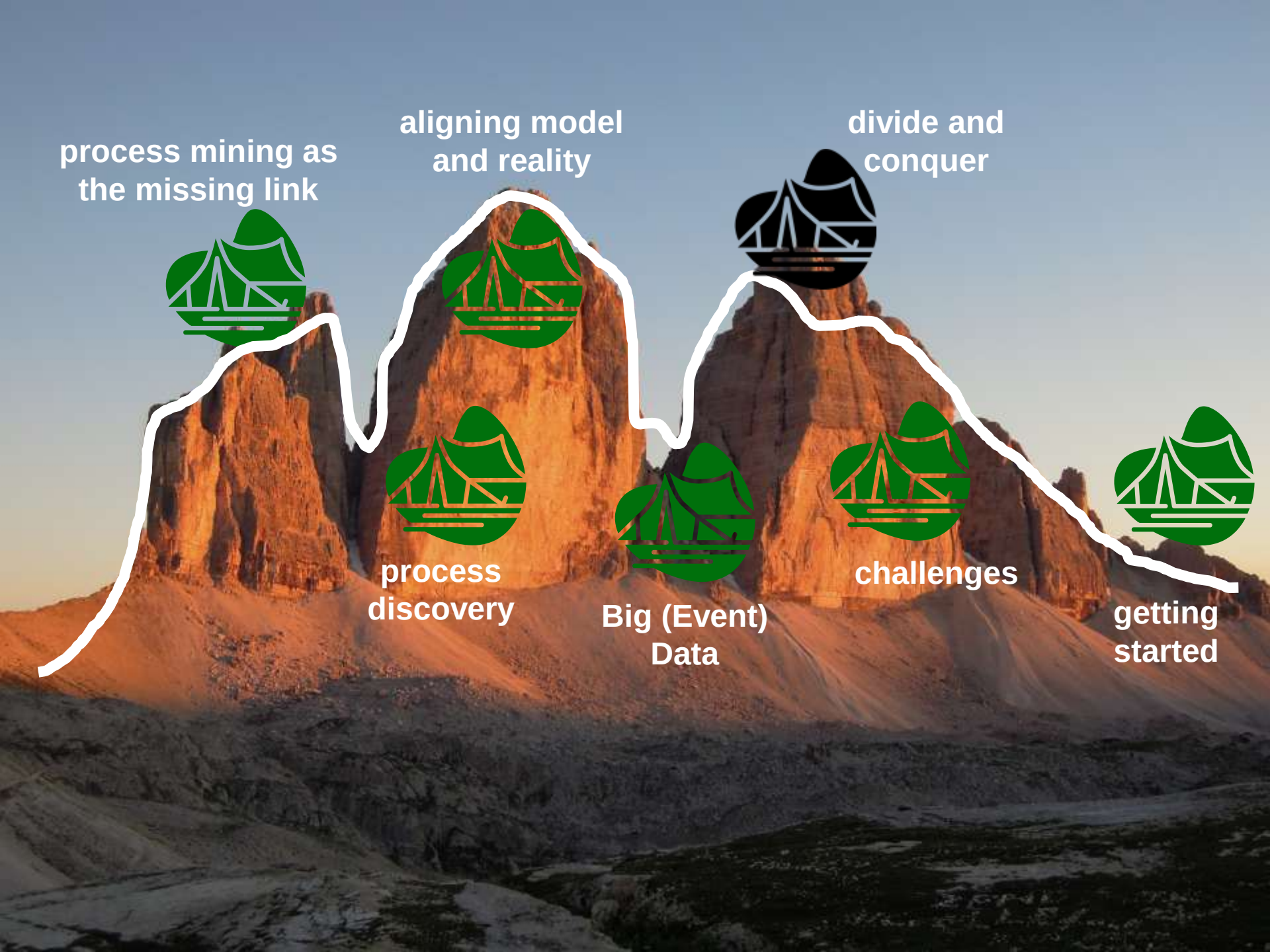
The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT


U.S. AIR FORCE
AFTERBURNER
CROSS INTO THE BLUE
U.S. AIR FORCE
1-800-423-USAF
AIRFORCE.COM

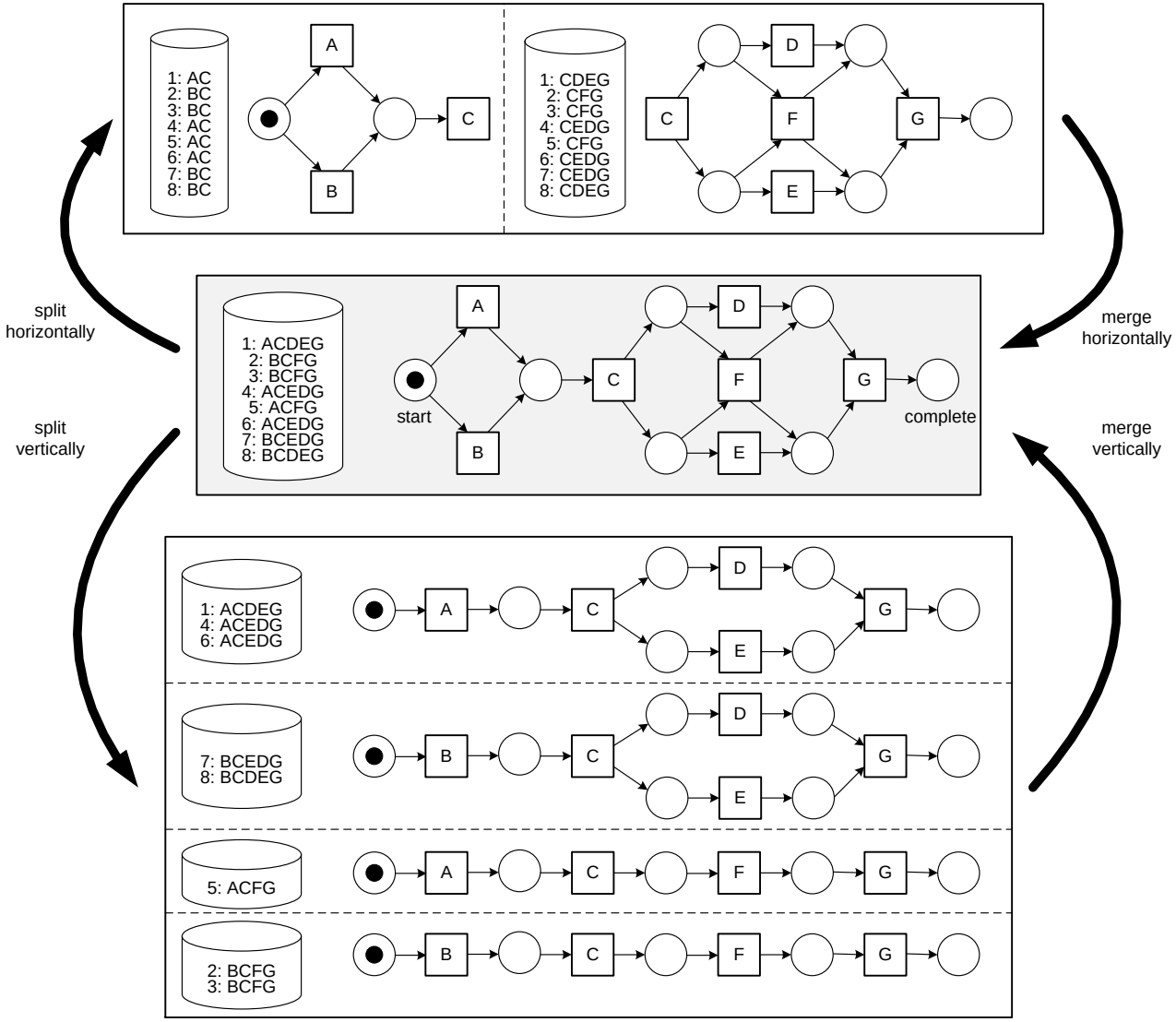
Big Data ?


Big … or fast and efficient?

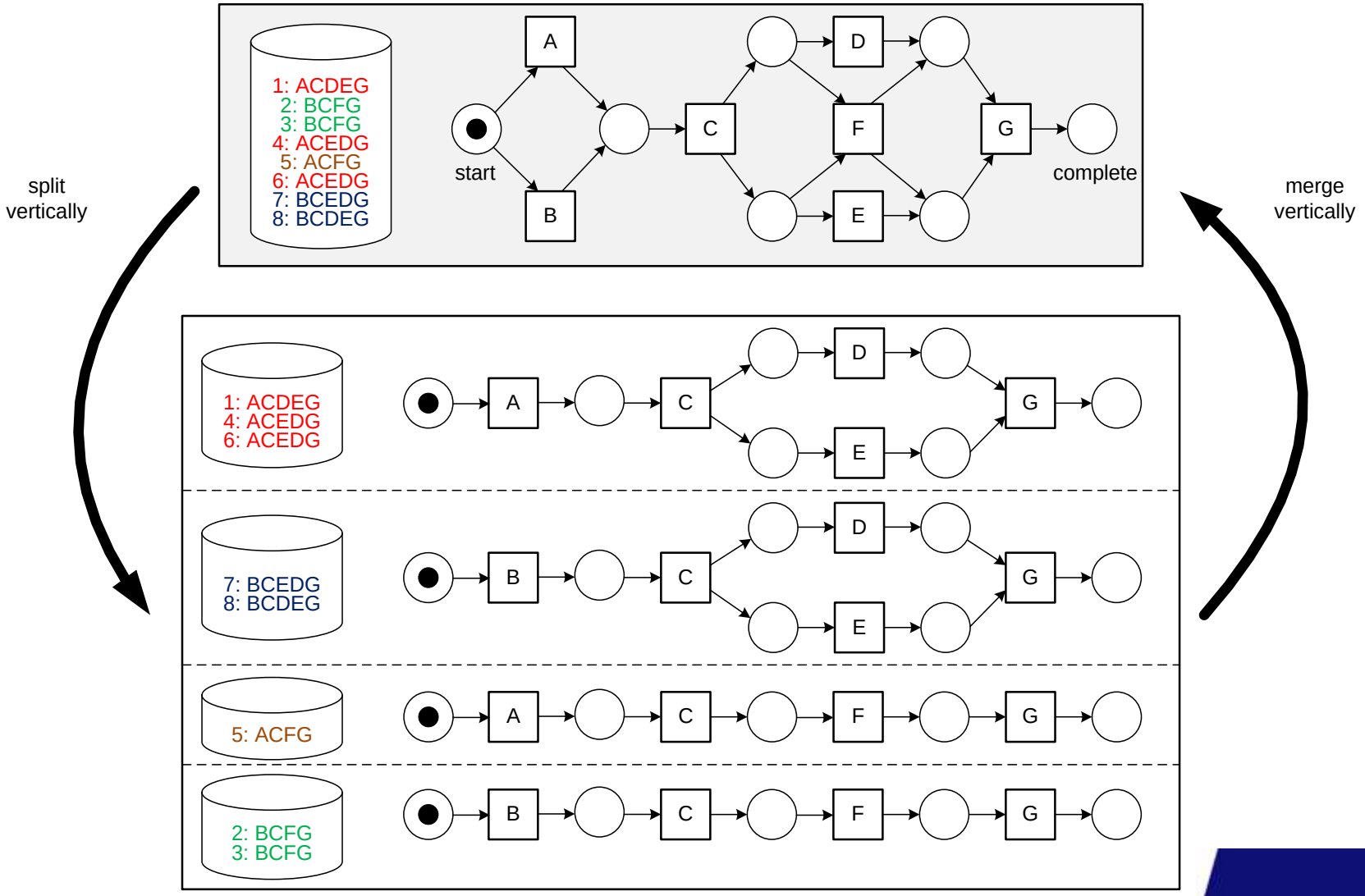

process mining as the missing link
aligning model and reality
divide and conquer
process discovery
Big (Event) Data
challenges
getting started

Big Data: Opportunities and Challenges

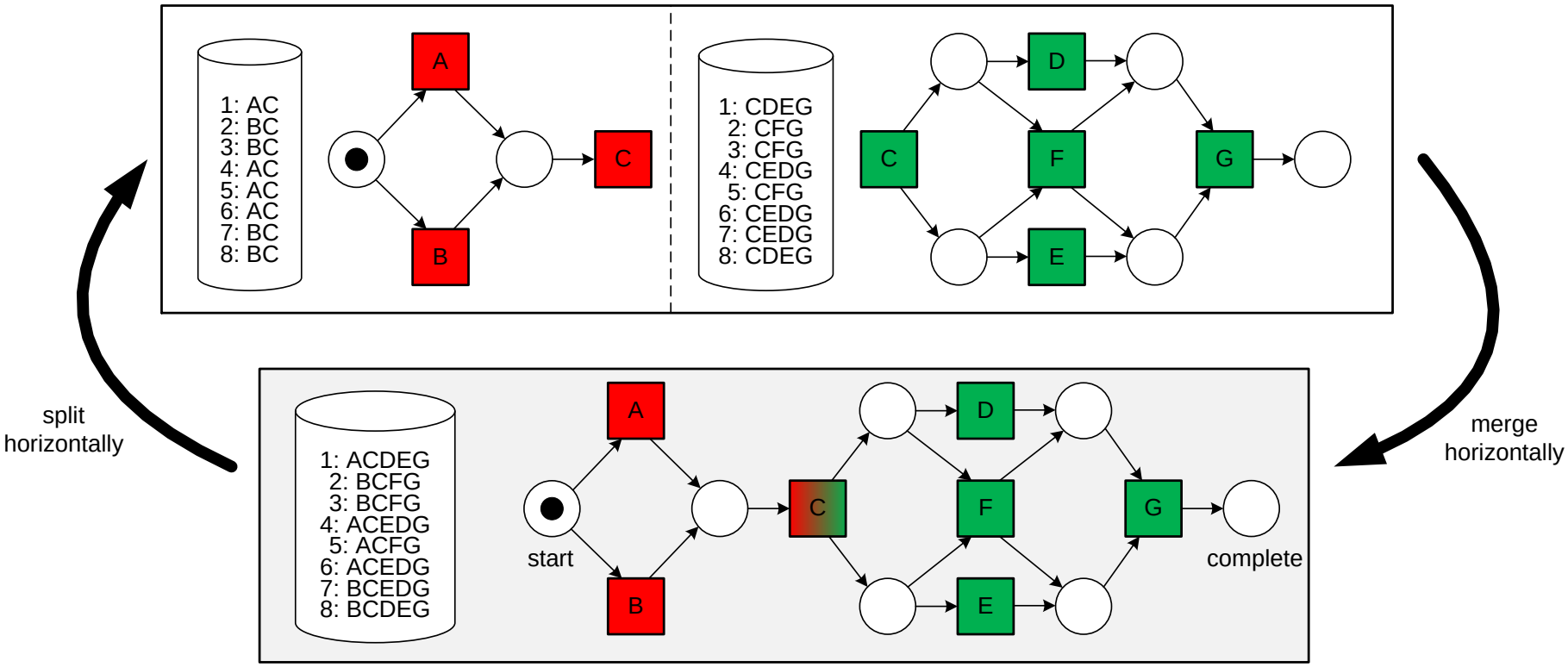
Divide and Conquer



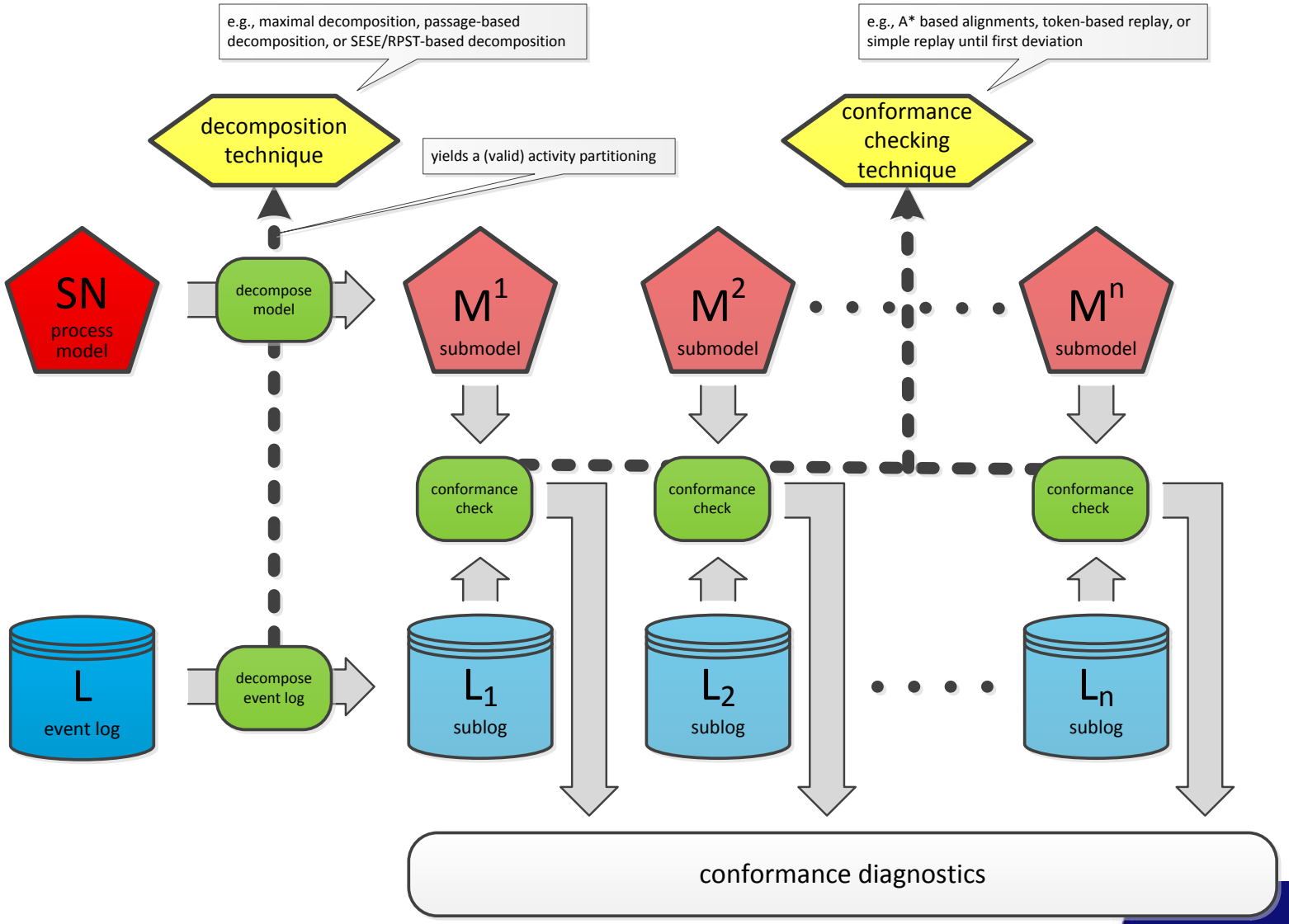
Vertical Decomposition



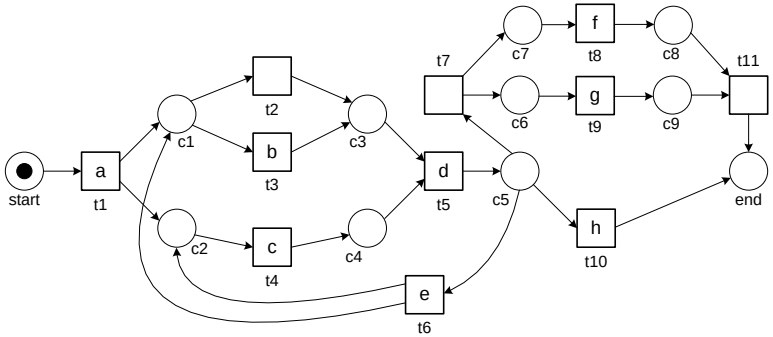
Horizontal Decomposition



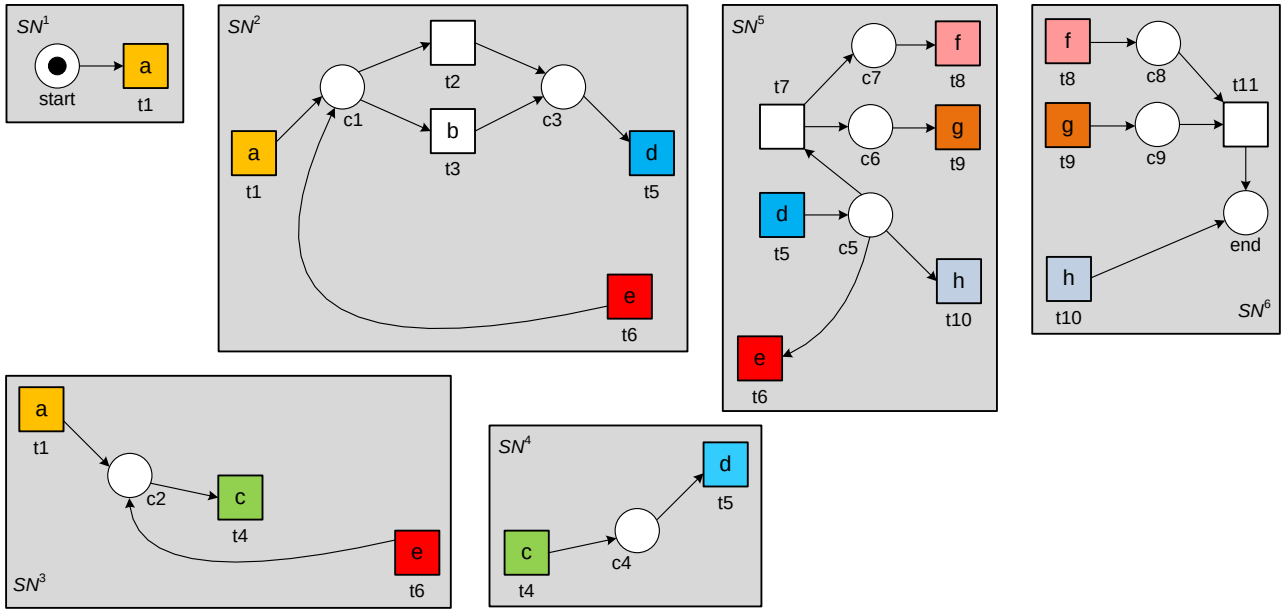
Decomposing Conformance Checking



Example of a valid decomposition

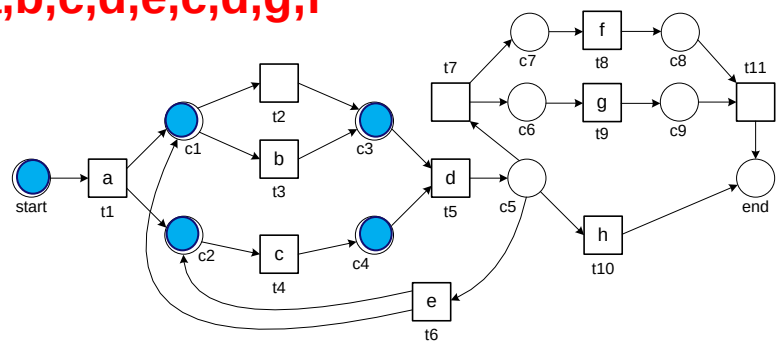


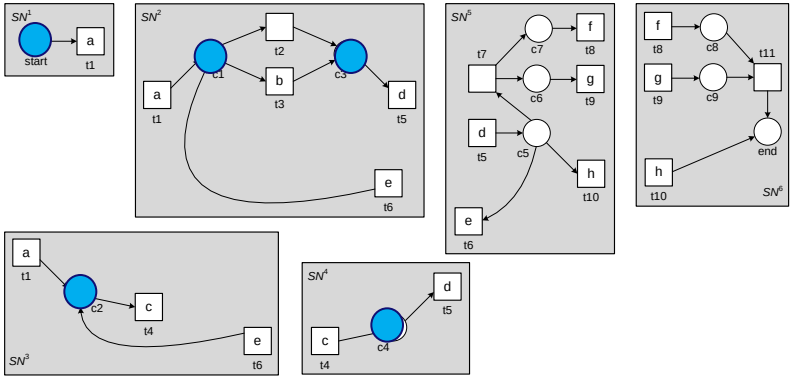
Log can be split in the same way!



Example of alignment for observed trace a,b,c,d,e,c,d,g,f

a,b,c,d,e,c,d,g,f





$$\gamma_3 = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|} \hline 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ \hline a & b & c & d & e & c & \gg & d & \gg & g & f & \gg \\ \hline a & b & c & d & e & c & \tau & d & \tau & g & f & \tau \\ t1 & t3 & t4 & t5 & t6 & t4 & t2 & t5 & t7 & t9 & t8 & t11 \\ \hline \end{array}$$

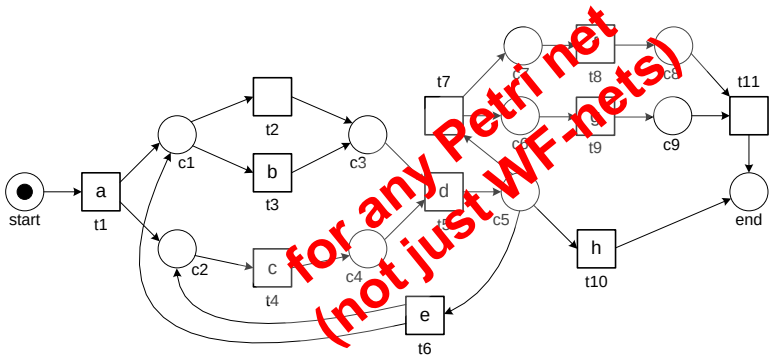
Etc.

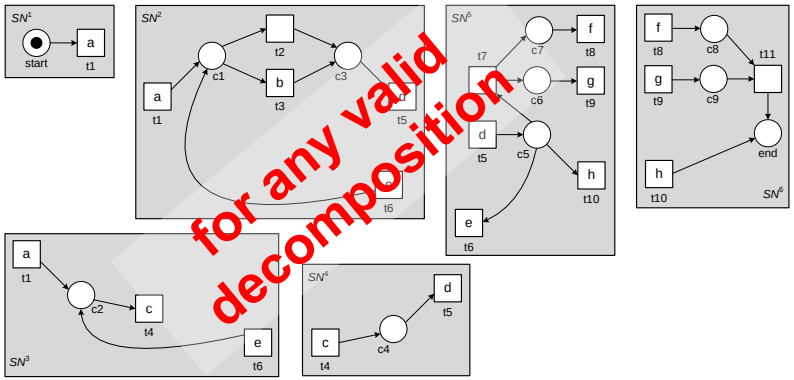
$$\gamma_3^1 = \begin{array}{|c|} \hline 1 \\ \hline a \\ \hline a \\ t1 \\ \hline \end{array} \qquad \gamma_3^2 = \begin{array}{|c|c|c|c|c|c|} \hline 1 & 2 & 4 & 5 & 7 & 8 \\ \hline a & b & d & e & \gg & d \\ \hline a & b & d & e & \tau & d \\ t1 & t3 & t5 & t6 & t2 & t5 \\ \hline \end{array} \qquad \gamma_3^3 = \begin{array}{|c|c|c|c|} \hline 1 & 3 & 5 & 6 \\ \hline a & c & e & c \\ \hline a & c & e & c \\ t1 & t4 & t6 & t4 \\ \hline \end{array}$$

$$\gamma_3^4 = \begin{array}{|c|c|c|c|} \hline 3 & 4 & 6 & 8 \\ \hline c & d & c & d \\ \hline c & d & c & d \\ t4 & t5 & t4 & t5 \\ \hline \end{array} \qquad \gamma_3^5 = \begin{array}{|c|c|c|c|c|c|} \hline 4 & 5 & 8 & 9 & 10 & 11 \\ \hline d & e & d & \gg & g & f \\ \hline d & e & d & \tau & g & f \\ t5 & t6 & t5 & t7 & t9 & t8 \\ \hline \end{array} \qquad \gamma_3^6 = \begin{array}{|c|c|c|} \hline 10 & 11 & 12 \\ \hline g & f & \gg \\ \hline g & f & \tau \\ t9 & t8 & t11 \\ \hline \end{array}$$

Conformance checking can be decomposed !!!

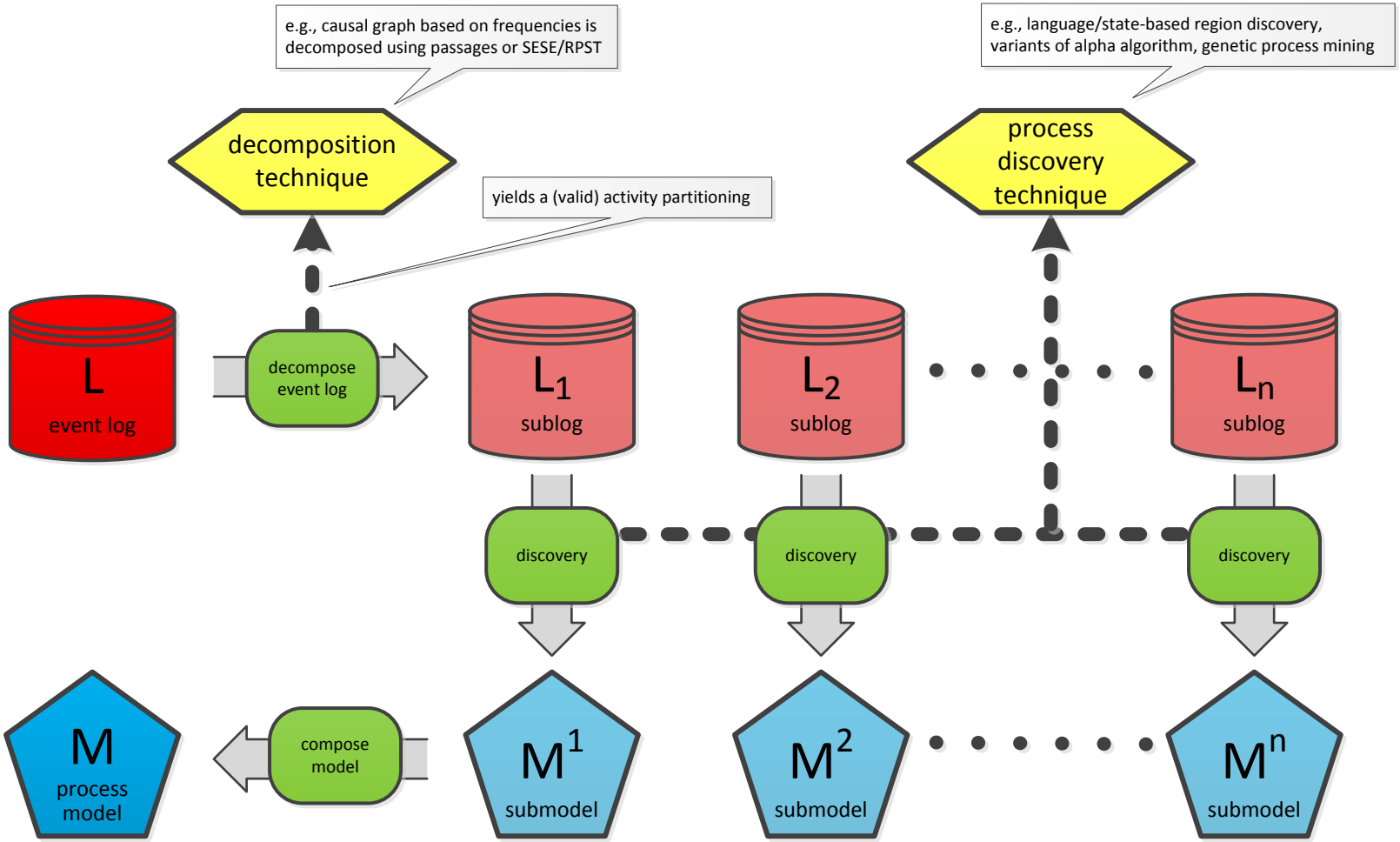
- **General result for any valid decomposition: Any event log or trace is perfectly fitting the overall model if and only if it is also fitting all the individual fragments**





Wil van der Aalst, Decomposing Petri nets for process mining: A generic approach. Distributed and Parallel Databases, Volume 31, Issue 4, pp 471-507, 2013

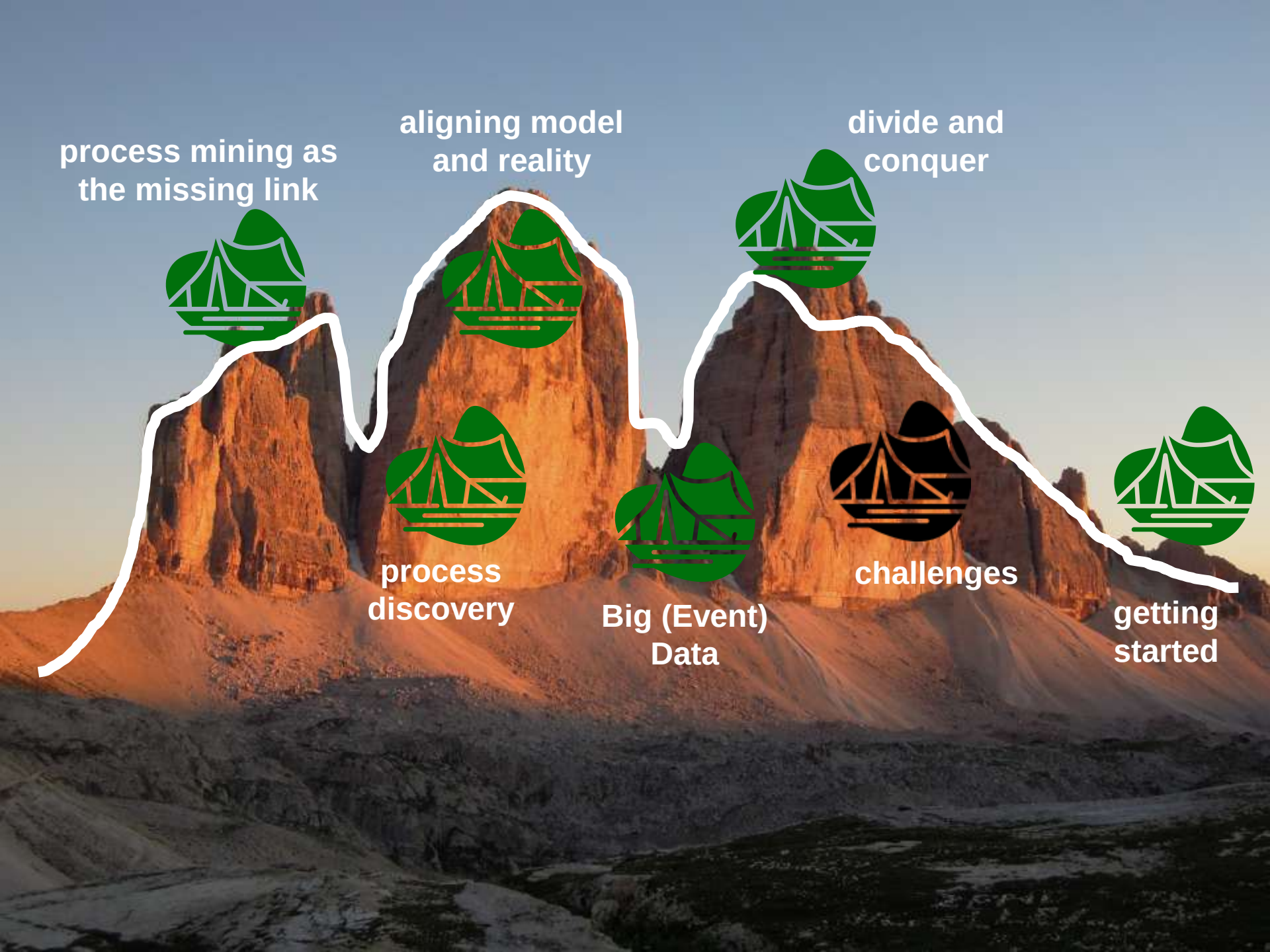
Decomposing Process Discovery



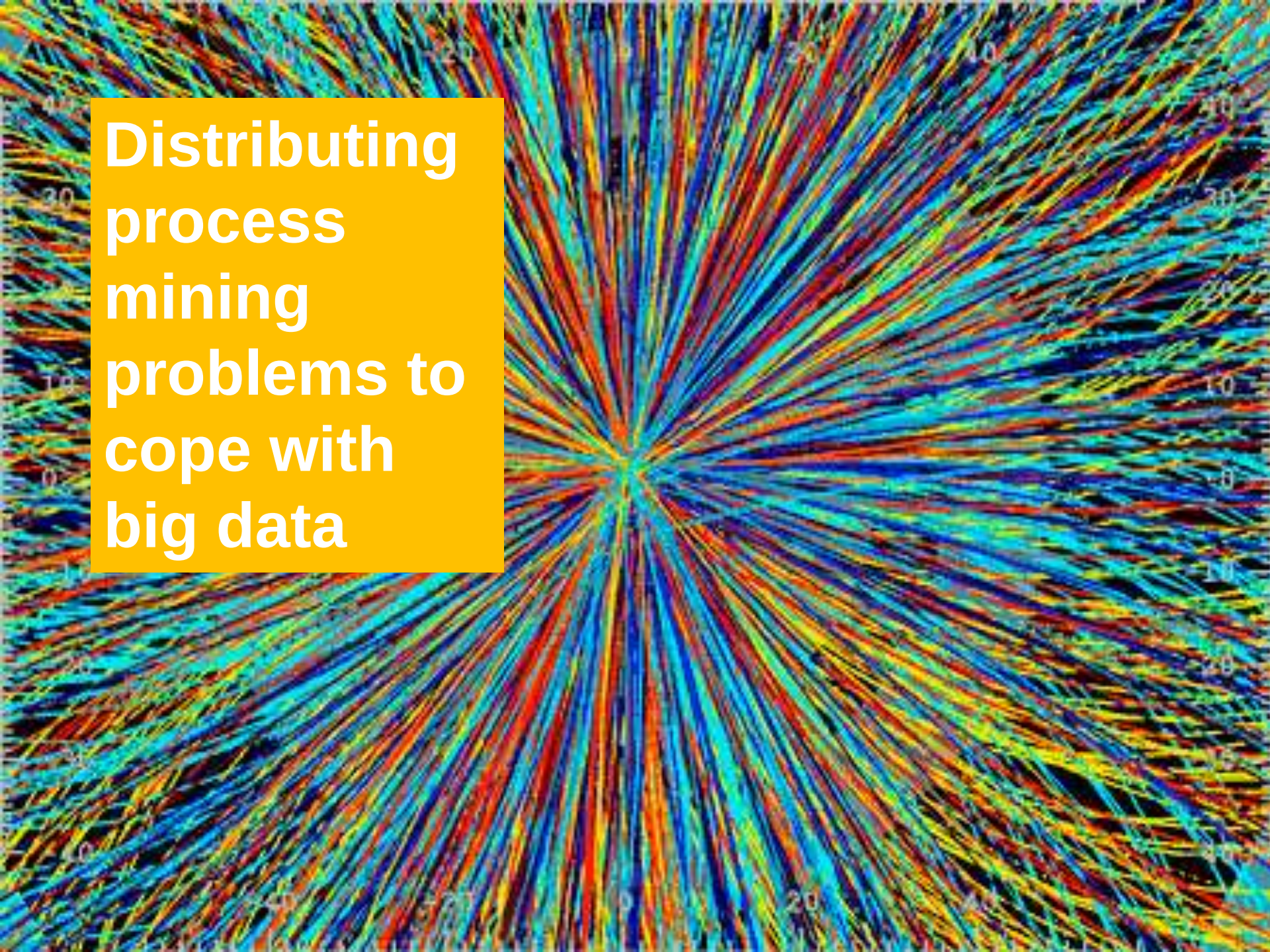
Learn more about decomposing process mining problems?

- **W.M.P. van der Aalst. Decomposing Petri Nets for Process Mining: A Generic Approach. *Distributed and Parallel Databases,* 31(4):471-507, 2013.**

- W.M.P. van der Aalst. A General Divide and Conquer Approach for Process Mining. In M. Ganzha, L. Maciaszek, and M. Paprzycki, editors, *Federated Conference on Computer Science and Information Systems (FedCSIS 2013)*, pages 1-10. IEEE Computer Society, 2013.
- W.M.P. van der Aalst. Decomposing Process Mining Problems Using Passages. In S. Haddad and L. Pomello, editors, *Applications and Theory of Petri Nets 2012*, volume 7347 of *Lecture Notes in Computer Science*, pages 72-91. Springer-Verlag, Berlin, 2012.
- J. Munoz-Gama, J. Carmona, and W.M.P. van der Aalst. Hierarchical Conformance Checking of Process Models Based on Event Logs. In J.M. Colom and J. Desel, editors, *Applications and Theory of Petri Nets 2013*, volume 7927 of *Lecture Notes in Computer Science*, pages 291-310. Springer-Verlag, Berlin, 2013.
- J. Munoz-Gama, J. Carmona, and W.M.P. van der Aalst. Conformance Checking in the Large: Partitioning and Topology. In F. Daniel, J. Wang, and B. Weber, editors, *International Conference on Business Process Management (BPM 2013)*, volume 8094 of *Lecture Notes in Computer Science*, pages 130-145. Springer-Verlag, Berlin, 2013.
- E. Verbeek and W.M.P. van der Aalst. Decomposing Replay Problems: A Case Study. In D. Moldt and H. Roelke, editors, *Proceedings of the International Workshop on Petri Nets in Software Engineering (PNSE 2013)*, volume 989 of *CEUR Workshop Proceedings*, pages 213-232. CEUR-WS.org, 2013.


process mining as the missing link
aligning model and reality
divide and conquer
process discovery
Big (Event) Data
challenges
getting started

Distributing process mining problems to cope with big data



streaming event data

(sensors, RFID, messages, etc.)


process discovery: finding sheep with five or more legs
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)

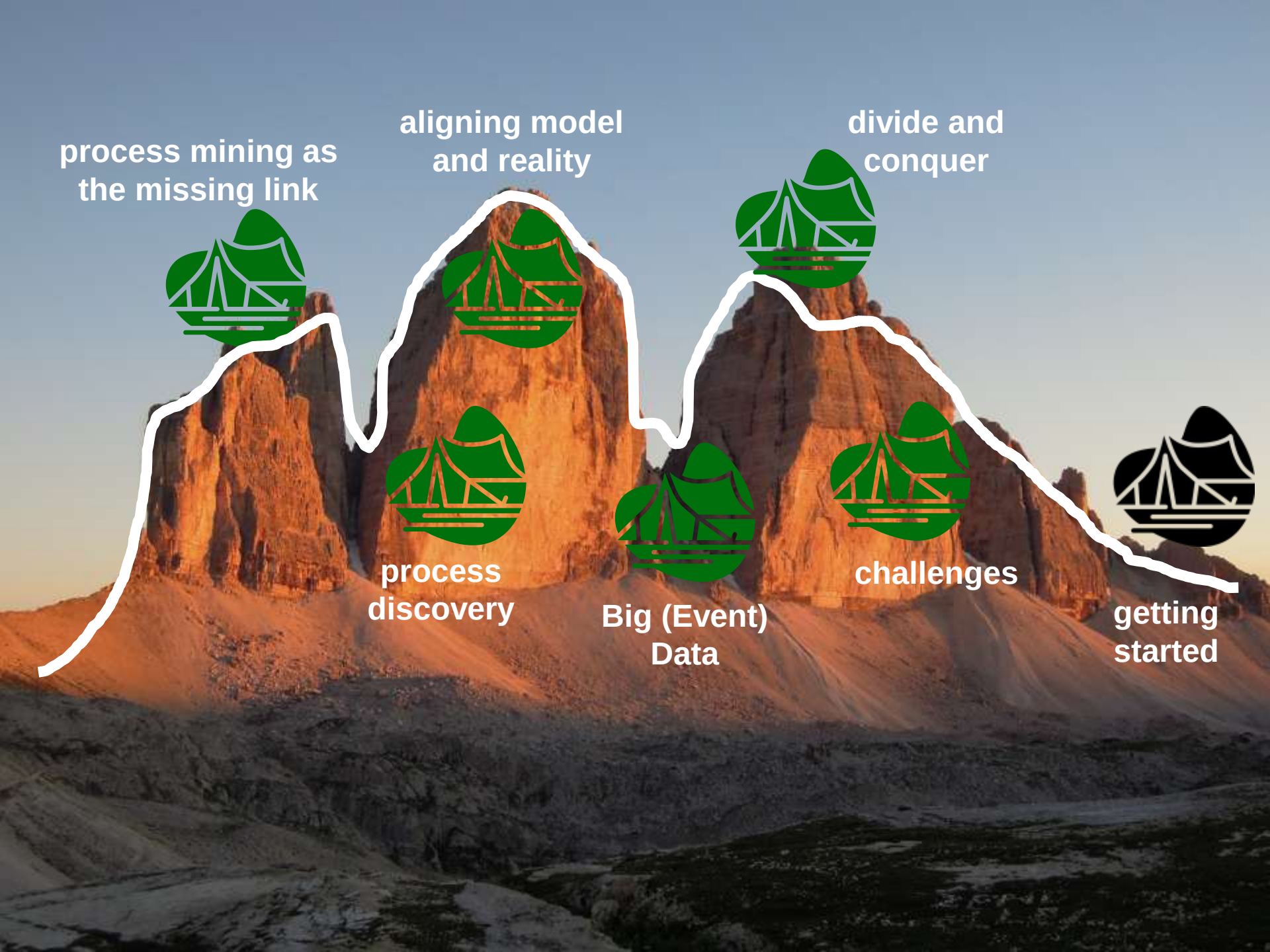
On-the-fly process mining

Operational support


Concept drift

cross-organizational / comparative process mining

Supporting the process of process mining

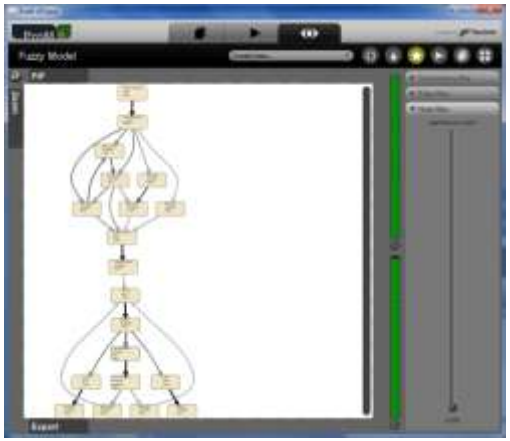

process mining as the missing link
aligning model and reality
divide and conquer
process discovery
Big (Event) Data
challenges
getting started

How to get started?

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

ProM
process mining workbench

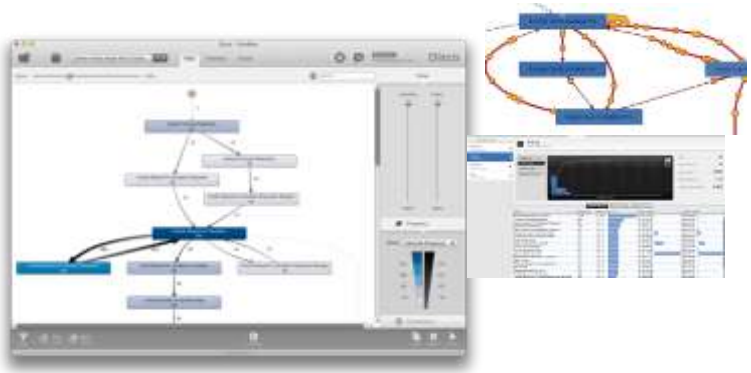
L)

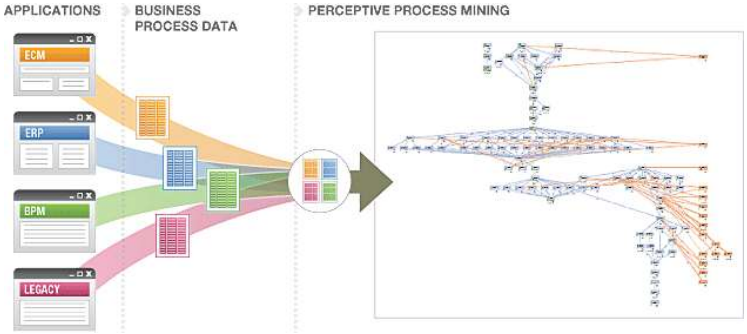


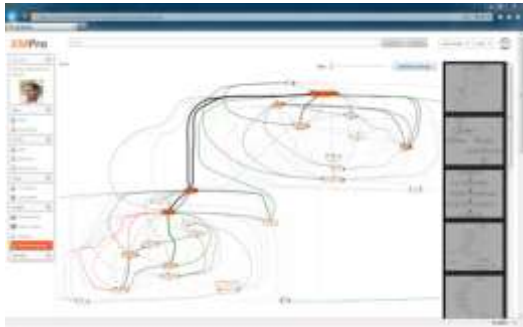
Download from: www.processmining.org

Commercial Alternatives

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







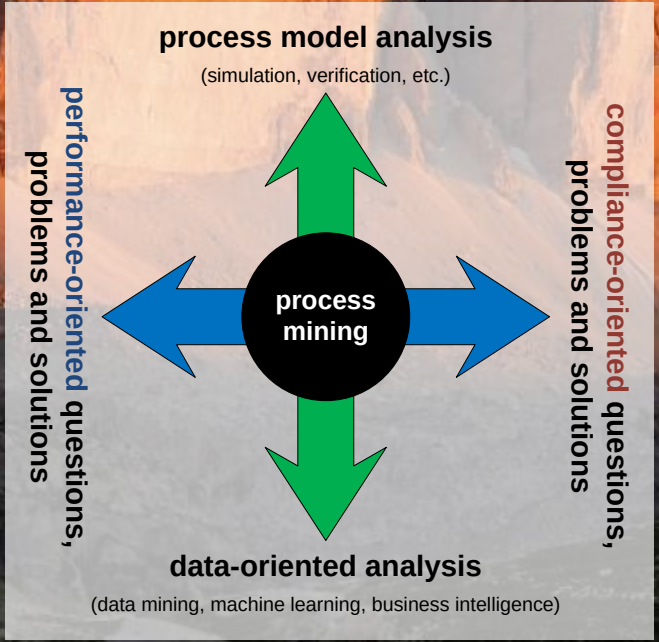
How to Get Started?

Collect event data ⟷ **Collect questions**

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

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

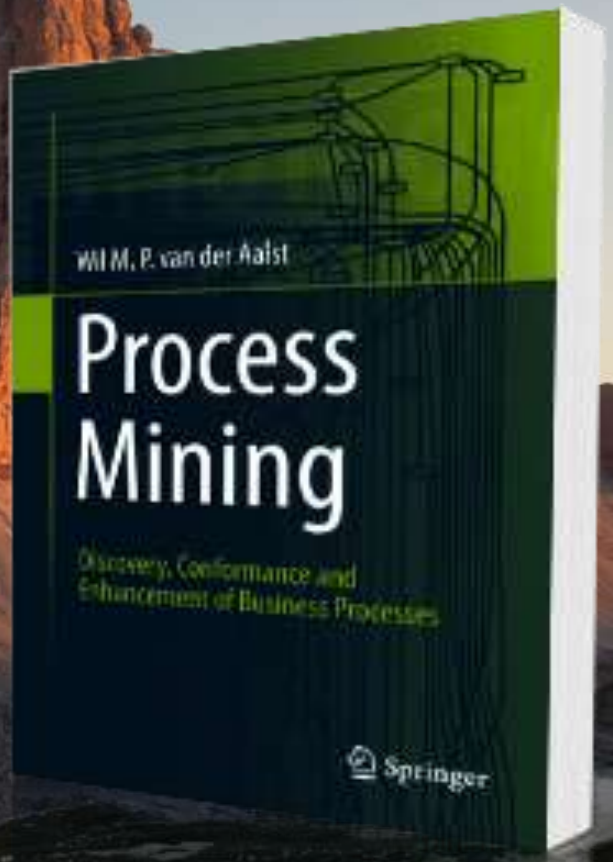
Join our expedition: Mine your processes!



Learn more?

http://www.youtube.com/watch?v=7oat7MatU_U

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



Informal PM Meeting (15.50 today).
Thanks to Krzysztof Kluza!
Building C2 (entrance through buildings C1 or C3), room no. 316 (3rd floor).

processmining.org