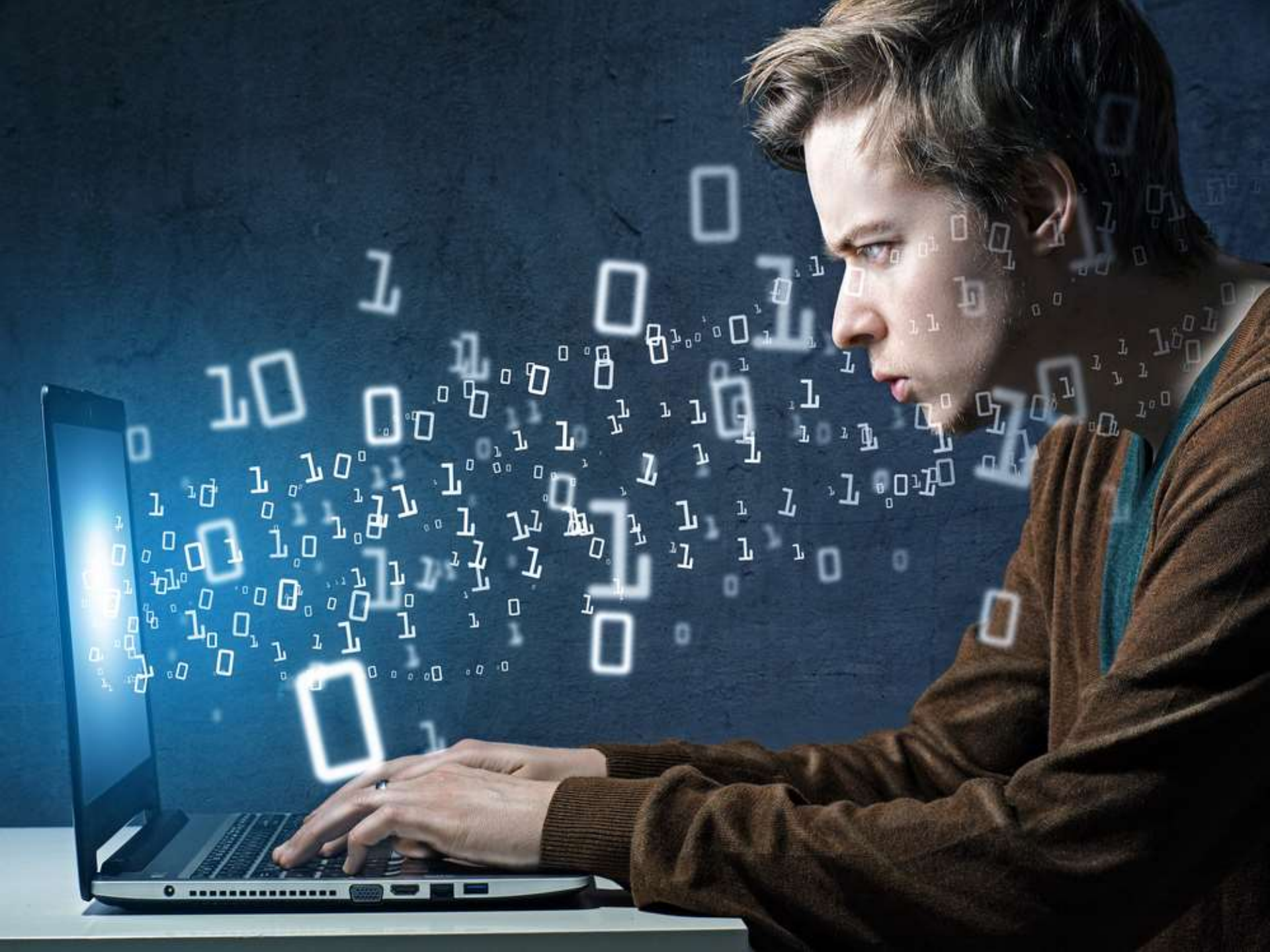
How People Really (Like To) Work

Comparative Process Mining to Unravel Human Behavior

prof.dr.ir. Wil van der Aalst

5th International Conference on Human-Centered Software Engineering (HCSE 2014), Paderborn, 16-9-2014.

TU/e Technische Universiteit Eindhoven University of Technology
Data Science Center Eindhoven



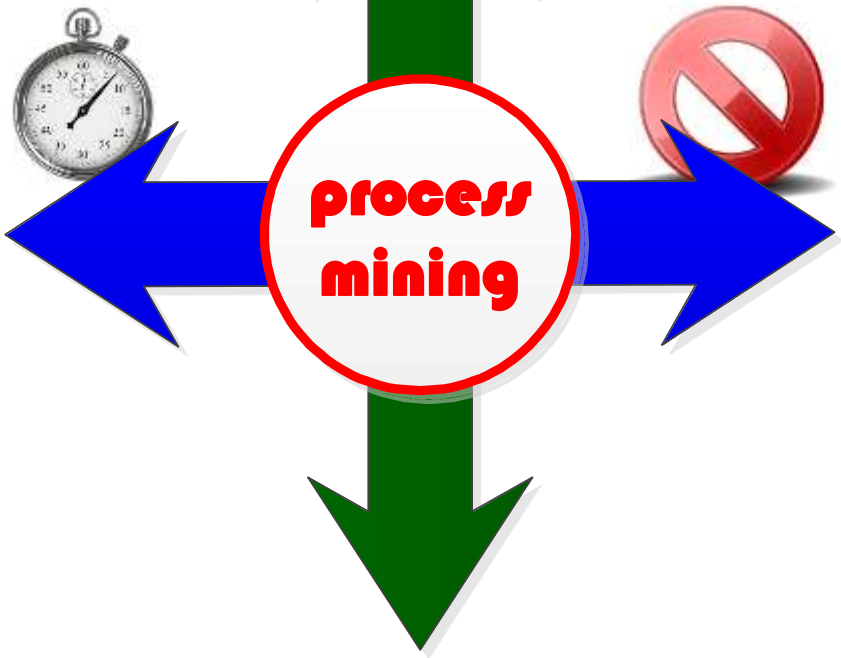
process mining

Process mining: The missing link

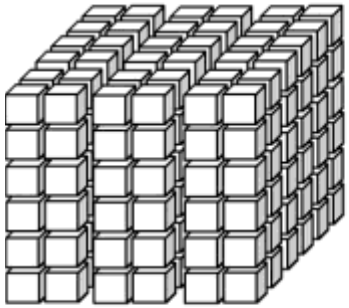
process model analysis

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

performance-oriented questions, problems and solutions



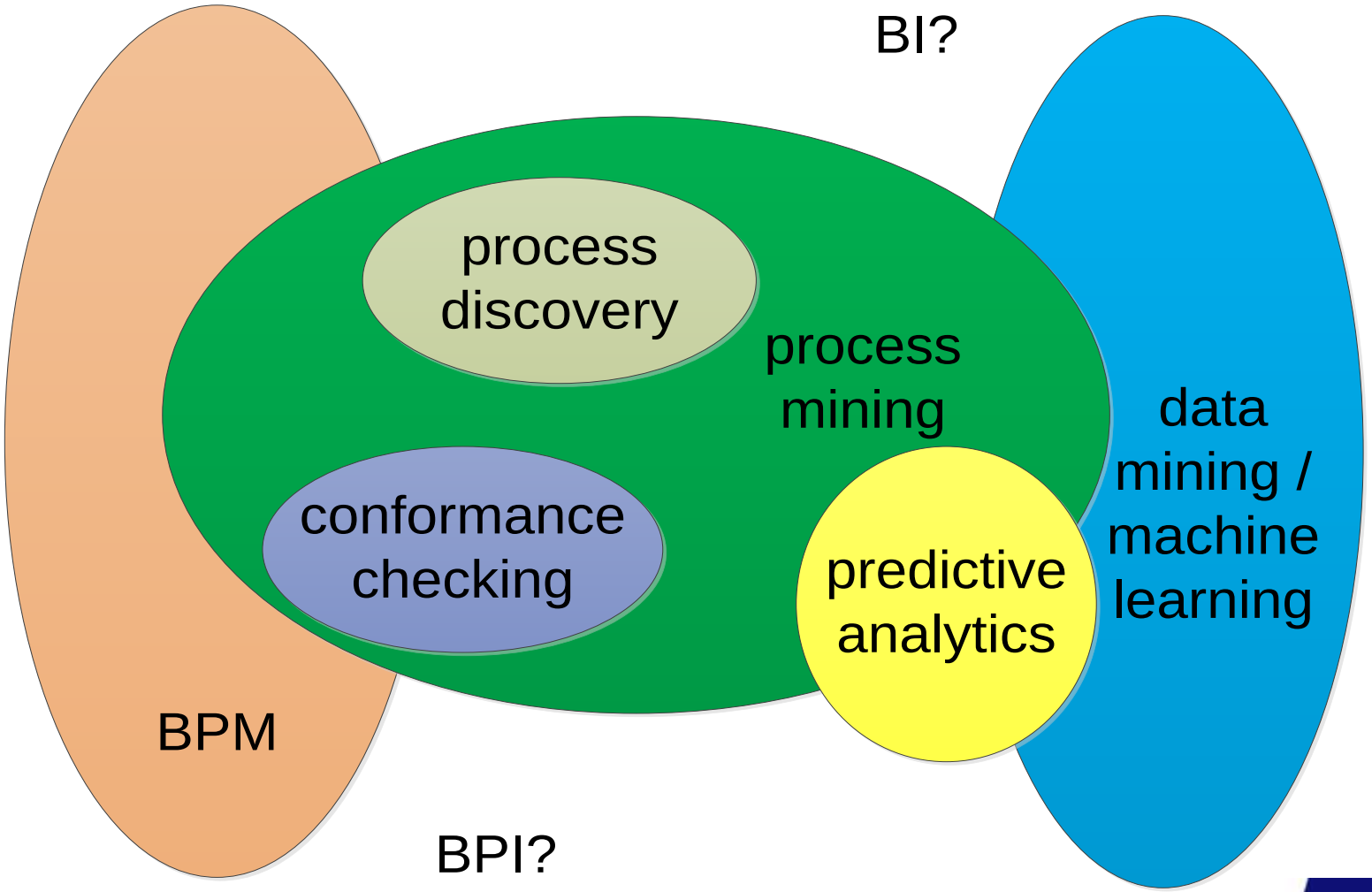
compliance-oriented questions, problems and solutions



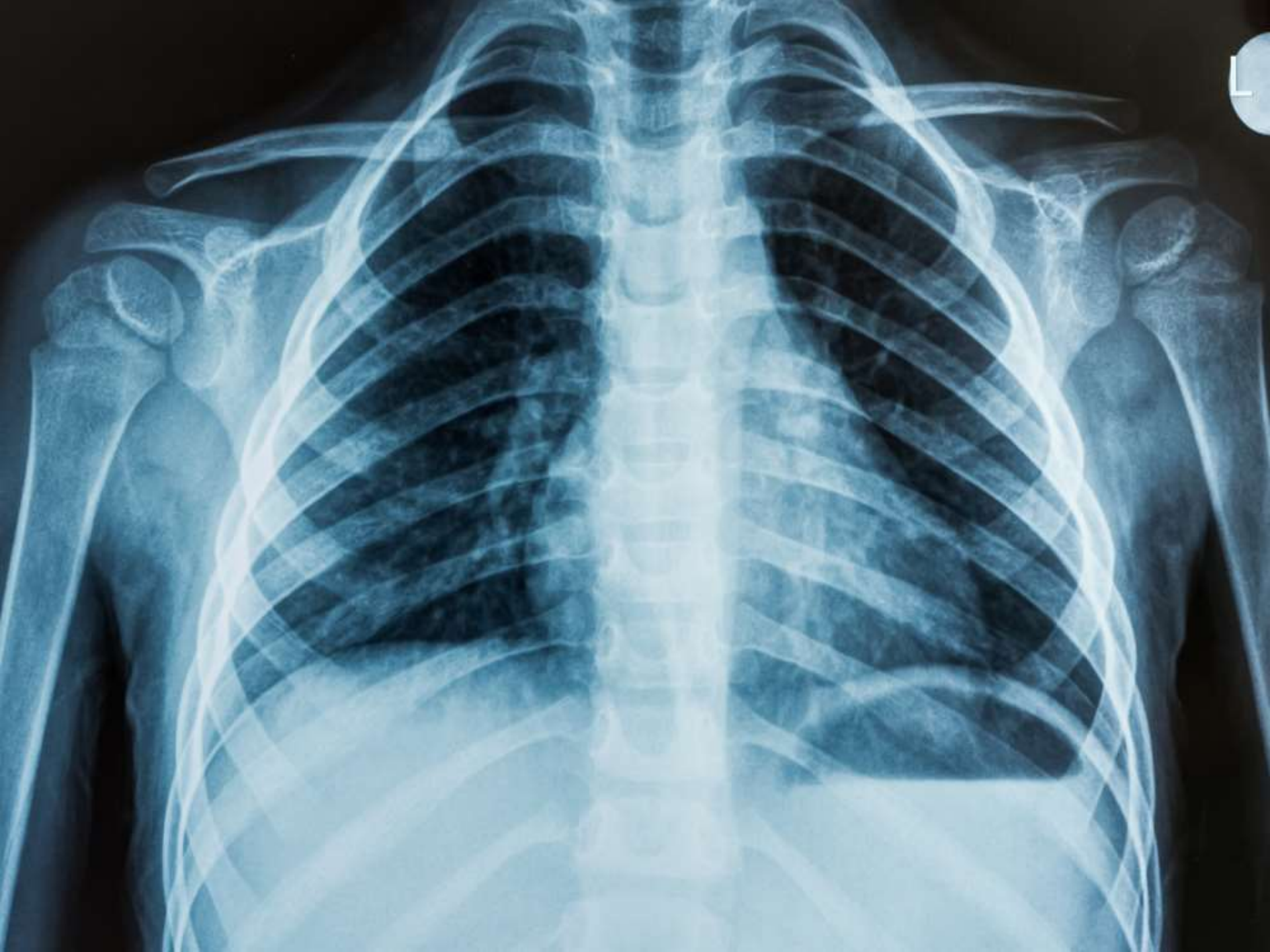
data-oriented analysis

(data mining, machine learning, business intelligence)

Positioning Process Mining




36
37
38
39
40
41



Starting point for process mining: Event data

every row is an event (here: an exam attempt)

student name	course name	exam date	mark
Peter Jones	Business Information systems	16-1-2014	8
Sandy Scott	Business Information systems	16-1-2014	5
Bridget White	Business Information systems	16-1-2014	9
John Anderson	Business Information systems	16-1-2014	8
Sandy Scott	BPM Systems	17-1-2014	7
Bridget White	BPM Systems	17-1-2014	8
Sandy Scott	Process Mining	20-1-2014	5
Bridget White	Process Mining	20-1-2014	9
John Anderson	Process Mining	20-1-2014	8
…	…	…	…

case id

activity name

timestamp

other data

Another event log: order handling

order number	activity	timestamp	user	product	quantity
9901	register order	22-1-2014@09.15	Sara Jones	iPhone5S	1
9902	register order	22-1-2014@09.18	Sara Jones	iPhone5S	2
9903	register order	22-1-2014@09.27	Sara Jones	iPhone4S	1
9901	check stock	22-1-2014@09.49	Pete Scott	iPhone5S	1
9901	ship order	22-1-2014@10.11	Sue Fox	iPhone5S	1
9903	check stock	22-1-2014@10.34	Pete Scott	iPhone4S	1
9901	handle payment	22-1-2014@10.41	Carol Hope	iPhone5S	1
9902	check stock	22-1-2014@10.57	Pete Scott	iPhone5S	2
9902	cancel order	22-1-2014@11.08	Carol Hope	iPhone5S	2
…	…	…	…	…	…

case id

activity name

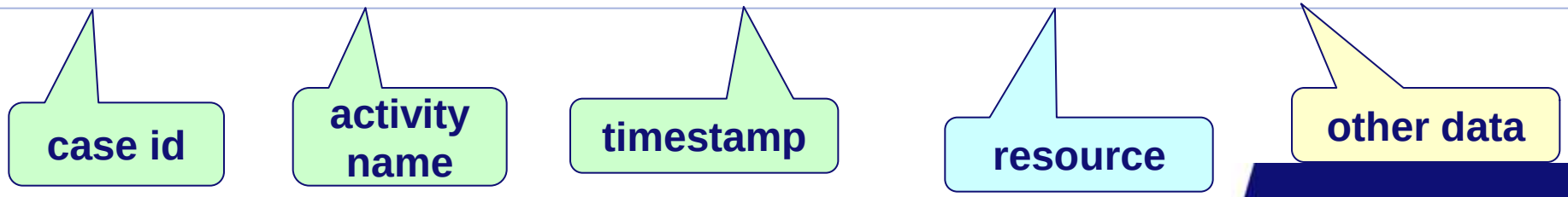
timestamp

resource

other data

Another event log: patient treatment

patient	activity	timestamp	doctor	age	cost
5781	make X-ray	23-1-2014@10.30	Dr. Jones	45	70.00
5541	blood test	23-1-2014@10.18	Dr. Scott	61	40.00
5833	blood test	23-1-2014@10.27	Dr. Scott	24	40.00
5781	blood test	23-1-2014@10.49	Dr. Scott	45	40.00
5781	CT scan	23-1-2014@11.10	Dr. Fox	45	1200.00
5833	surgery	23-1-2014@12.34	Dr. Scott	24	2300.00
5781	handle payment	23-1-2014@12.41	Carol Hope	45	0.00
5541	radiation therapy	23-1-2014@13.57	Dr. Jones	61	140.00
5541	radiation therapy	23-1-2014@13.08	Dr. Jones	61	140.00
…	…	…	…	…	…




www.olifantenpaadjes.nl
60

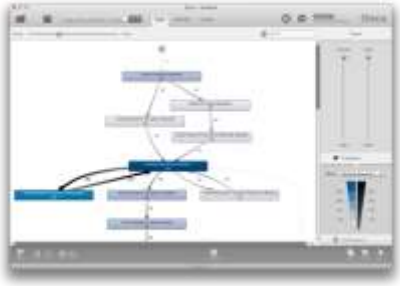
demo

ProM
process mining workbench

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

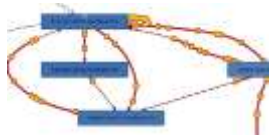
Disco
by Fluxicon.

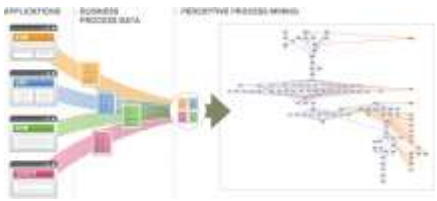
perceptivesoftware
a Lexmark company



QPR
Quality. Processes. Results.

celonis
process mining





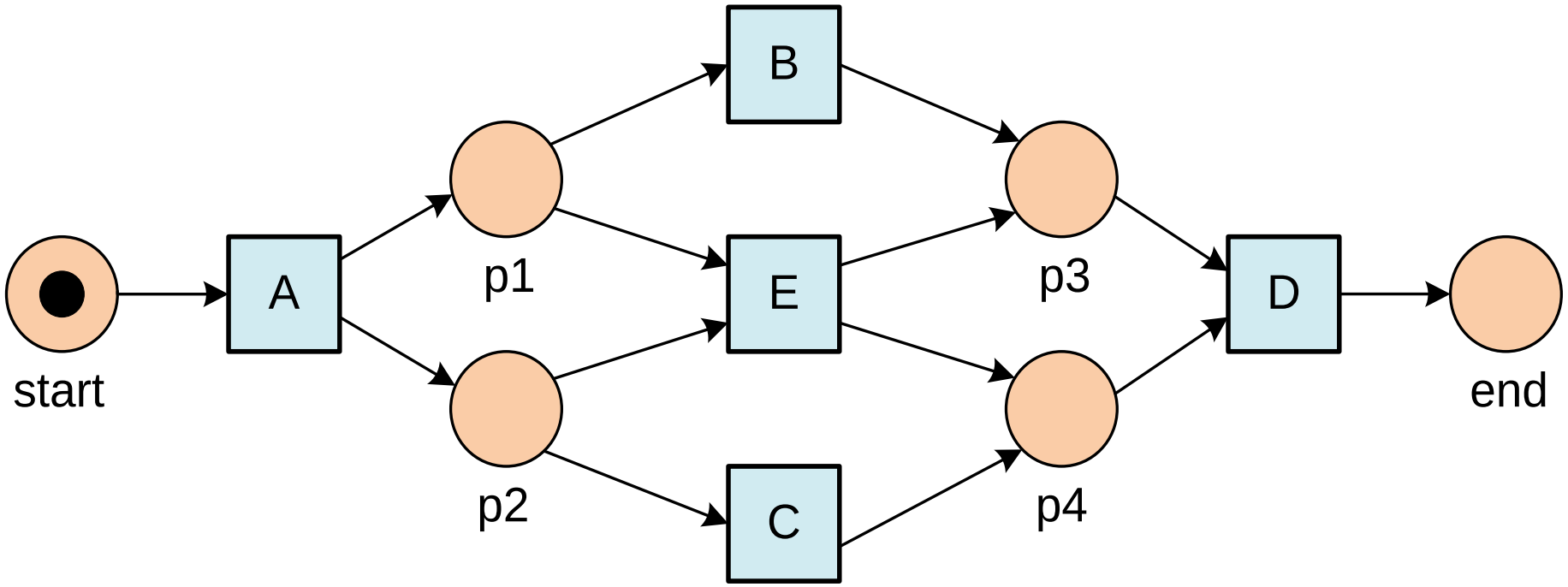
XMPRO
GET BETTER AT GETTING WORK DONE

StereoLogic

ARIS
Process Performance Manager

FUJITSU

Quiz Question:
How to remove behavior?



ABCD

ACBD

AED

ACBD

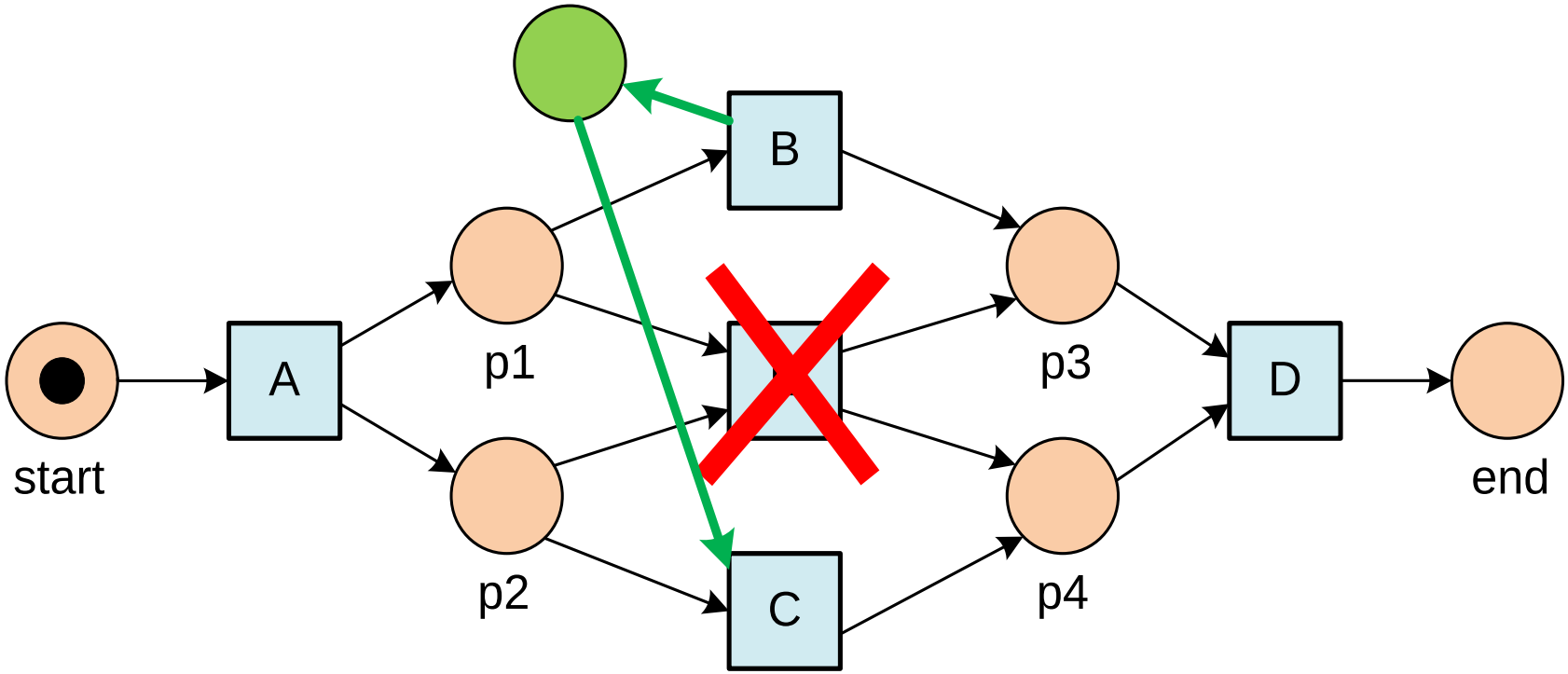
ABCD

AED

ABCD

ABCD

Answer:
Add places or remove transitions



ABCD

~~**ACBD**~~

~~**AED**~~

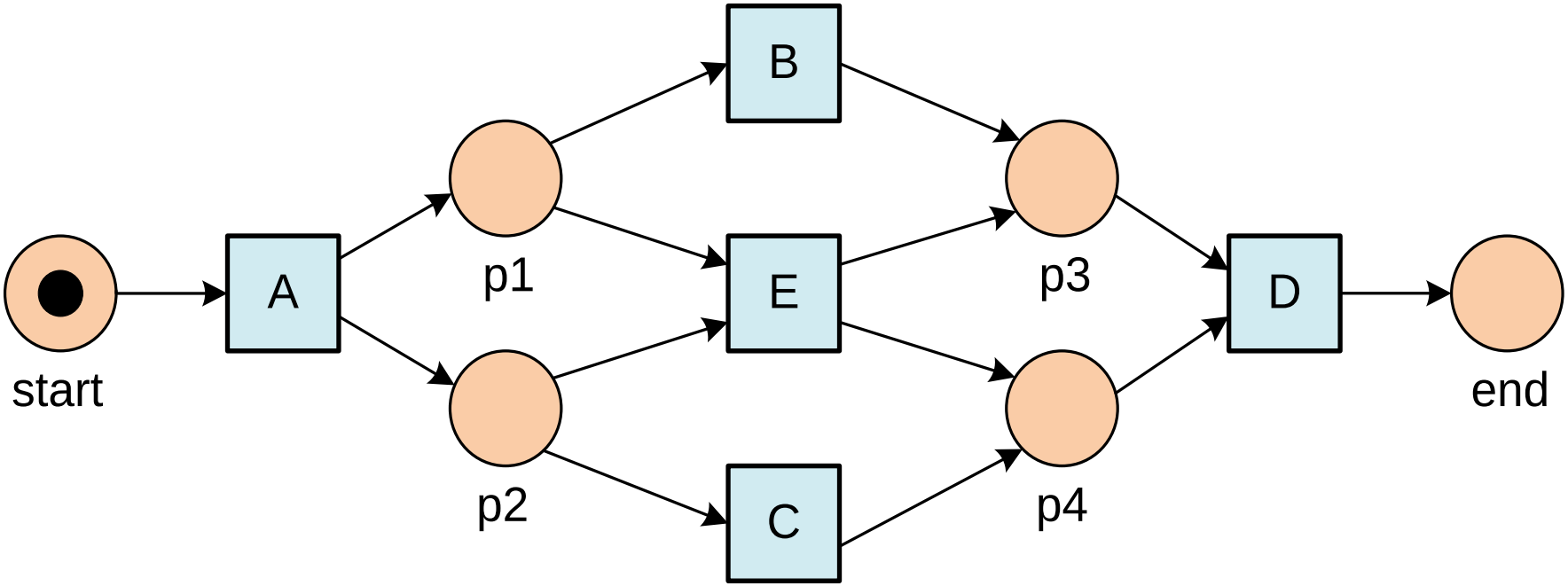
ABCD

~~**ACBD**~~

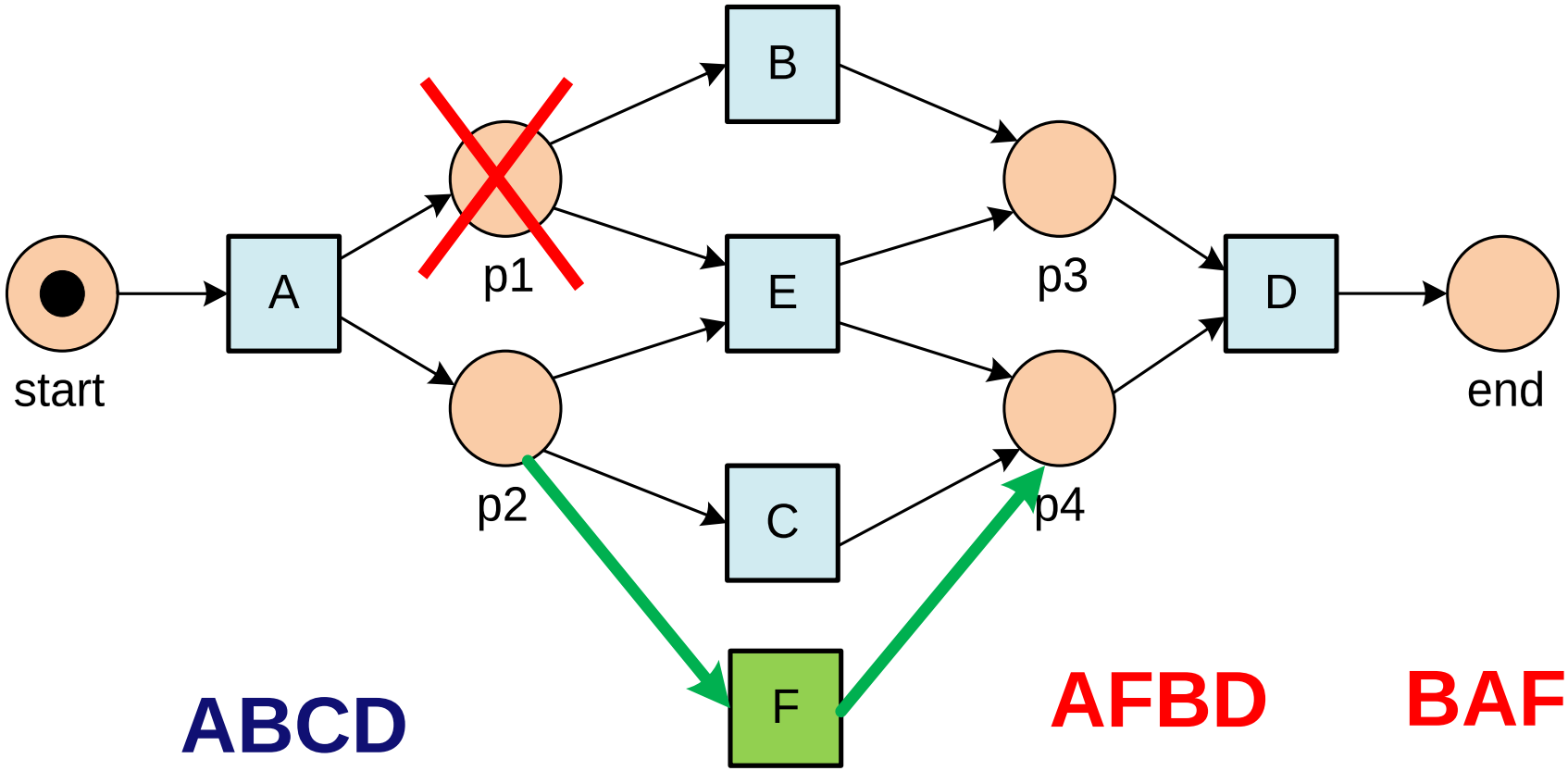
~~**AED**~~

~~**ABCD**~~

Quiz Question:
How to add behavior?



Answer:
Add transitions or remove places!



ABCD

ACBD

AED

ABCD

ABCD

AFBD

BAFD

BACD

ABFD

Process discovery algorithms
(small selection)

automata-based learning

distributed genetic mining

heuristic mining

language-based regions

genetic mining

state-based regions

stochastic task graphs

LTL mining

ETM genetic algorithm

Inductive Miner (infrequent)

fuzzy mining

neural networks

mining block structures

hidden Markov models

α algorithm

conformal process graph

multi-phase mining

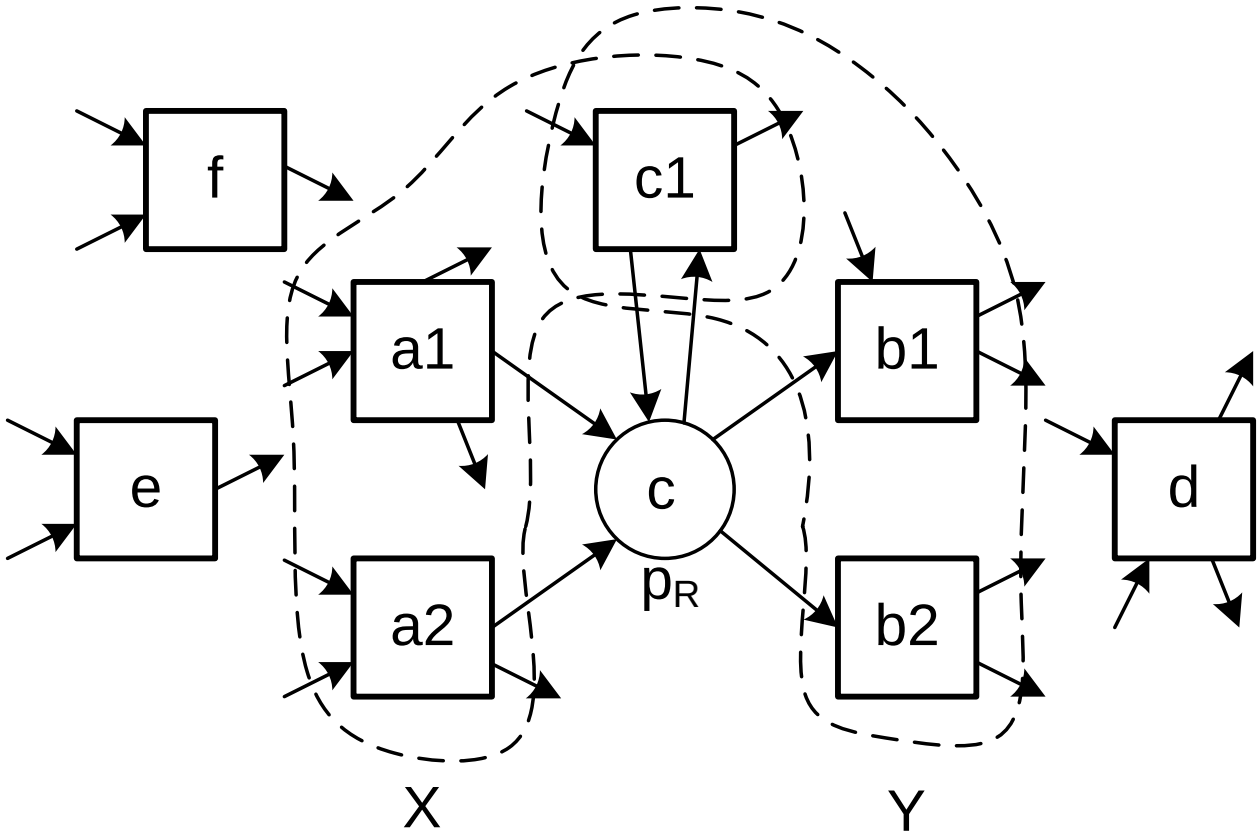
partial-order based mining

α# algorithm

ILP mining

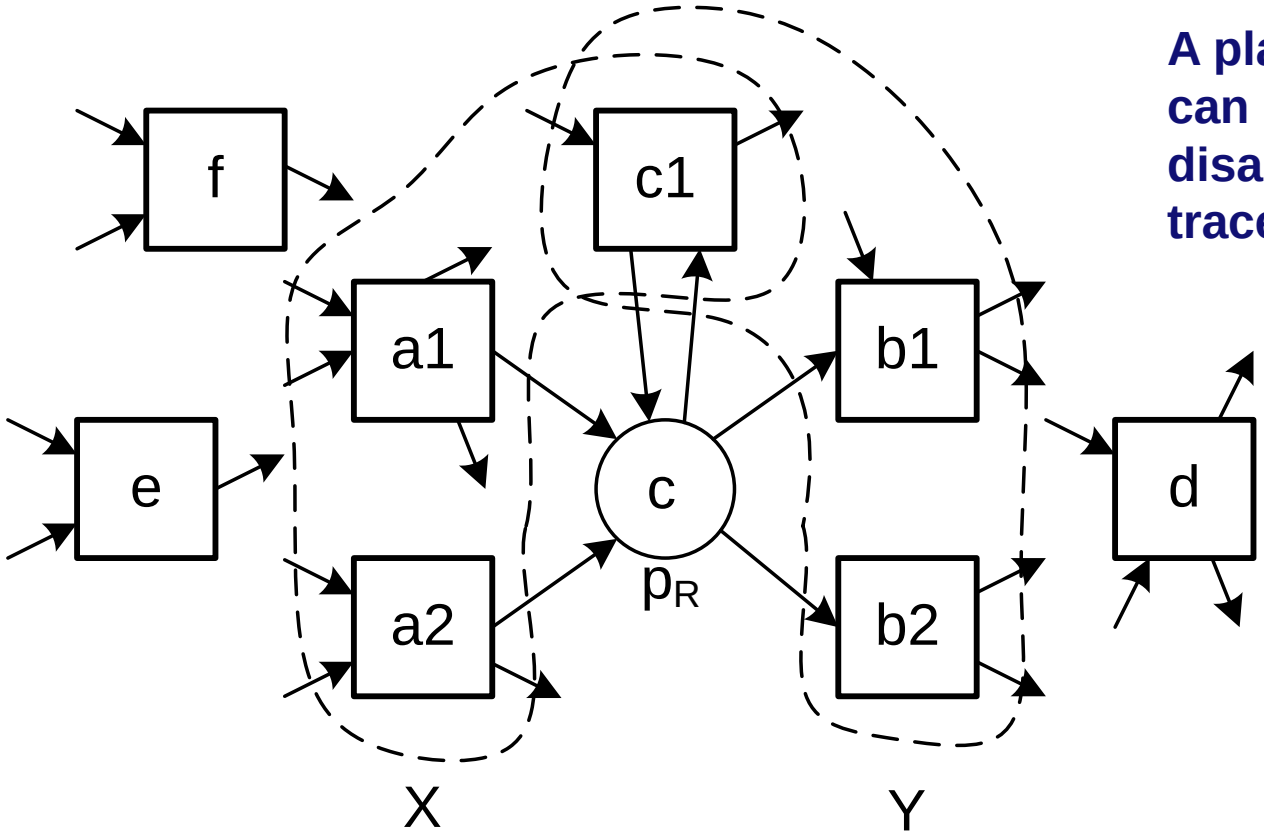
α++ algorithm

Language based regions



Region R = (X,Y,c) corresponding to place p_R: X = {a1,a2,c1} = transitions producing a token for p_R, Y = {b1,b2,c1} = transitions consuming a token from p_R, and c is the initial marking of p_R.

Basic idea: enough tokens should be present when consuming



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

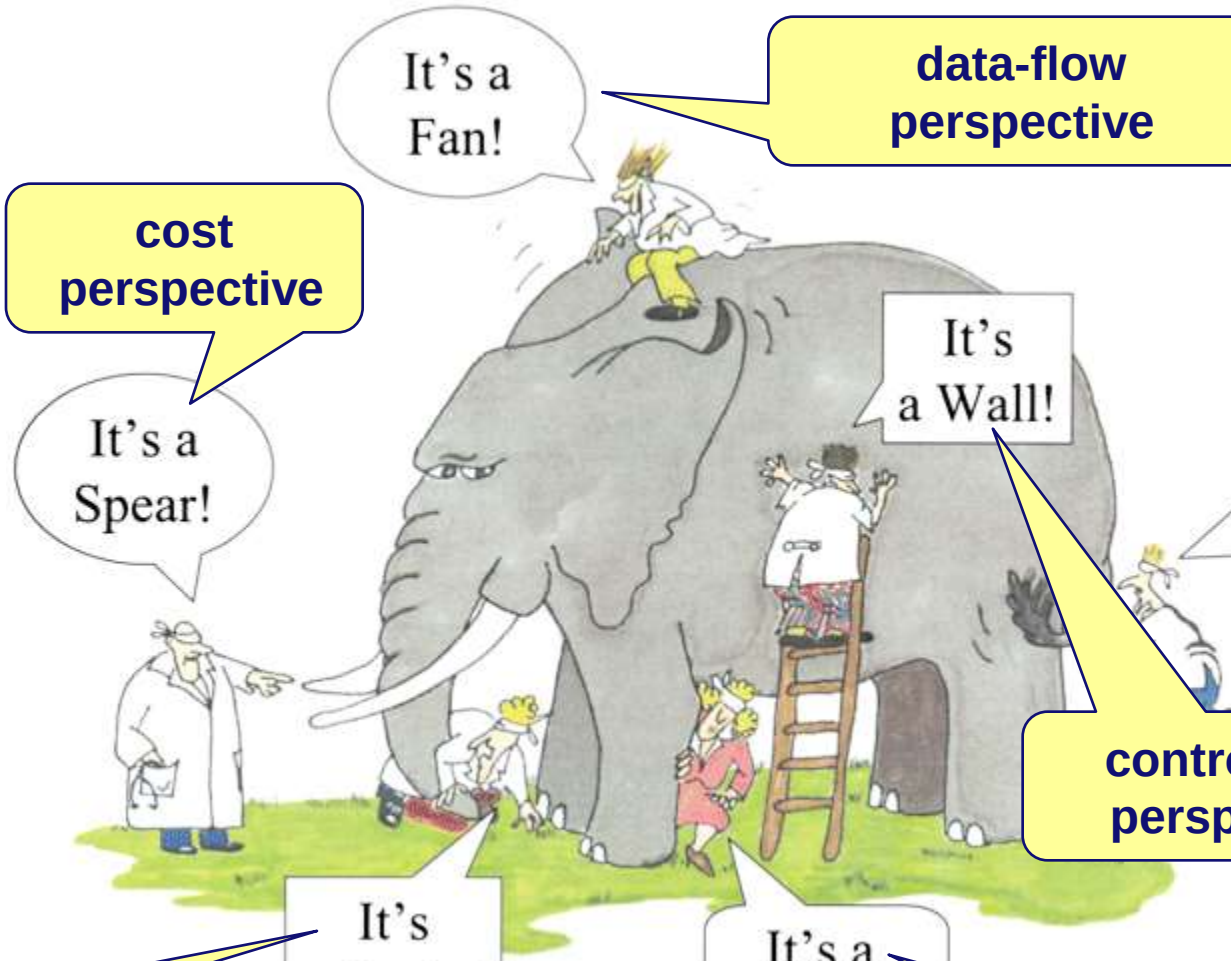
for any $\sigma \in L, k \in \{1, \ldots, |\sigma|\}, \sigma_1 = hd^{k-1}(\sigma), a = \sigma(k), \sigma_2 = hd^{k}(\sigma) = \sigma_1 \oplus a$:

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

Process mining is about connecting things

- **Data – Process**
- **Business – IT**
- **Business Intelligence – Business Process Management**
- **Performance – Compliance**
- **Runtime – Design time**
- **…**

Processes are not just about control-flow!



data-flow perspective

time perspective

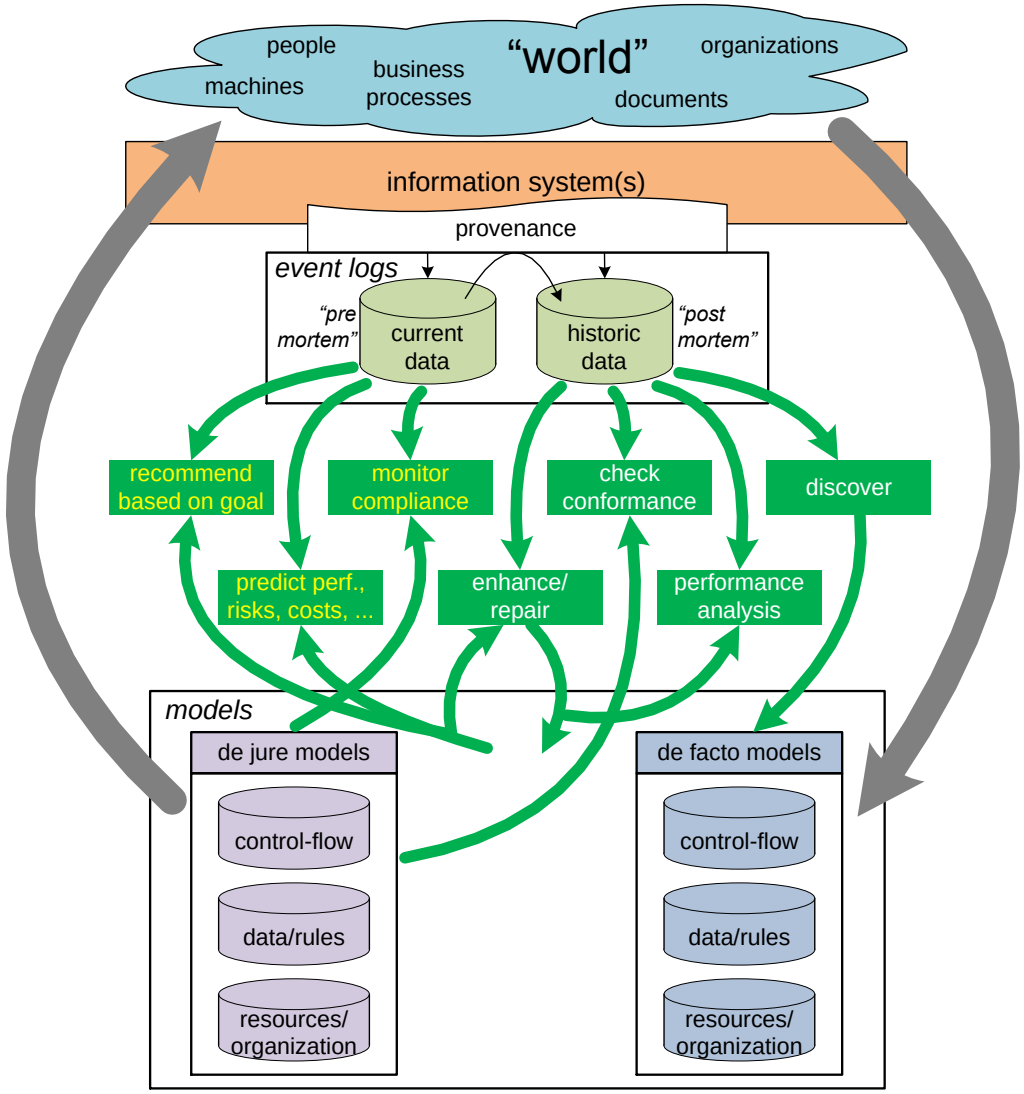
cost perspective

control-flow perspective

energy perspective

resource perspective

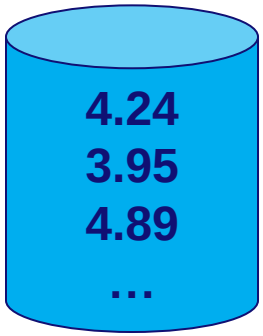
Process mining spectrum



- **Online and offline (near realtime).**
- **All perspectives.**
- **De facto models are descriptive/predictive.**
- **De jure models are normative/prescriptive.**
- **Process discovery is just one element: Aligning model and reality is the key thing.**

process cubes

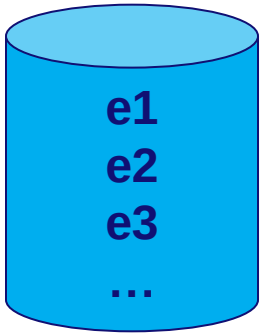
Process discovery is like applying a function



compute average

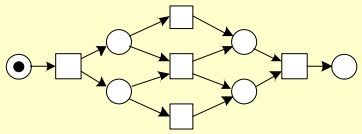
(freq., sum, variance, mode, …)

4.56



create model

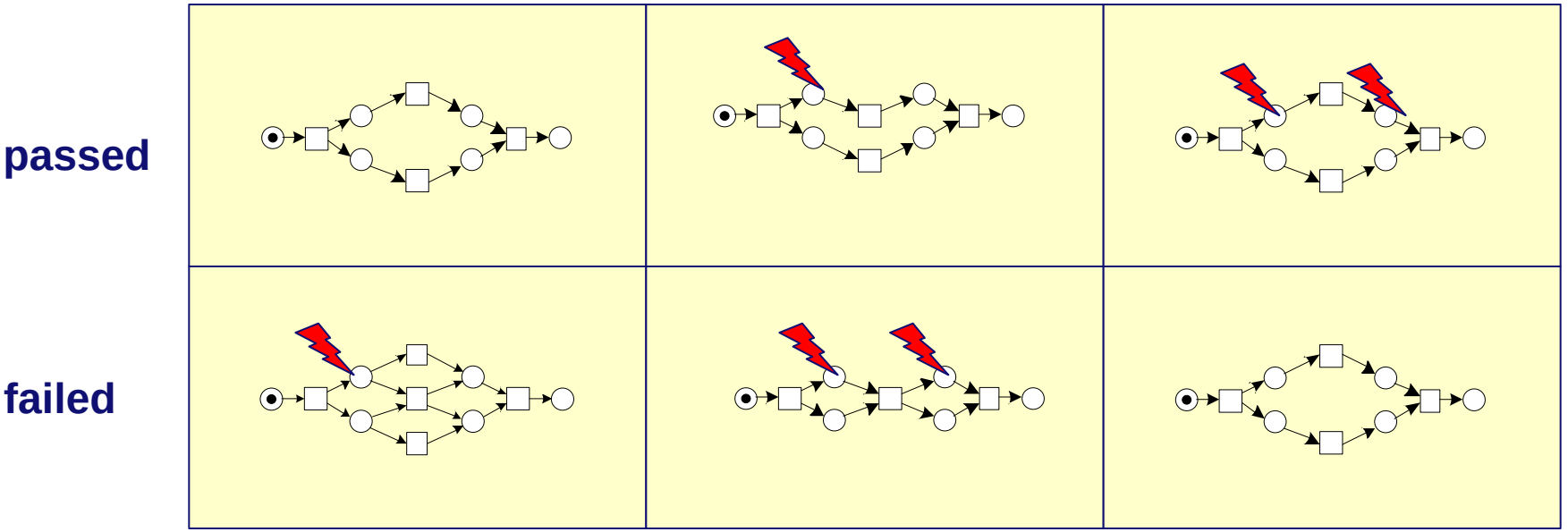
(alpha, heuristic, fuzzy, social, dotted, compliance, etc., etc.)



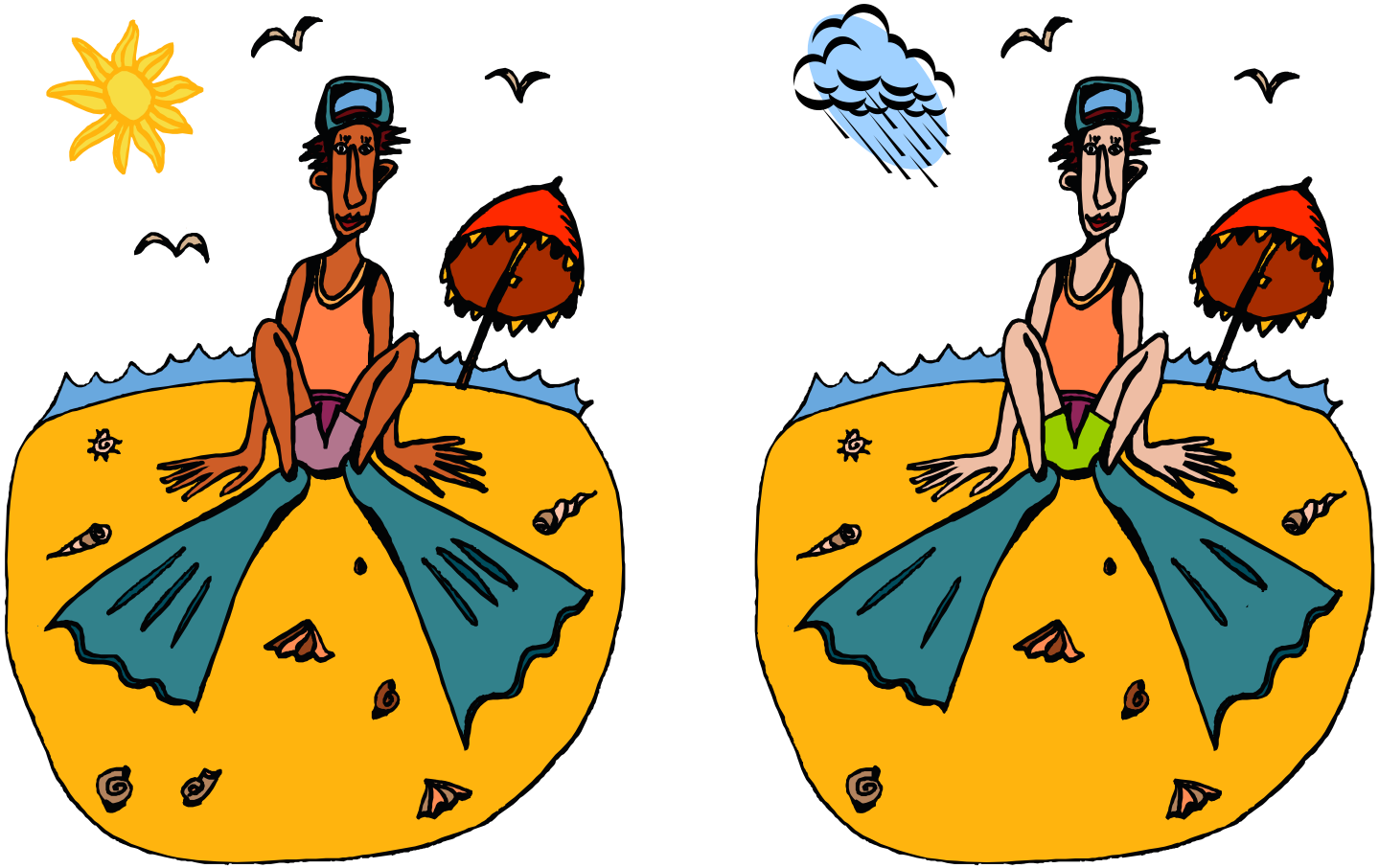
2 dimensions: model = number

	TU/e	HSE	QUT
passed	hours: 50.5 number: 32	hours: 48.5 number: 55	hours: 23.2 number: 49
failed	hours: 23.5 number: 23	hours: 10.5 number: 4	hours: 24.5 number: 8

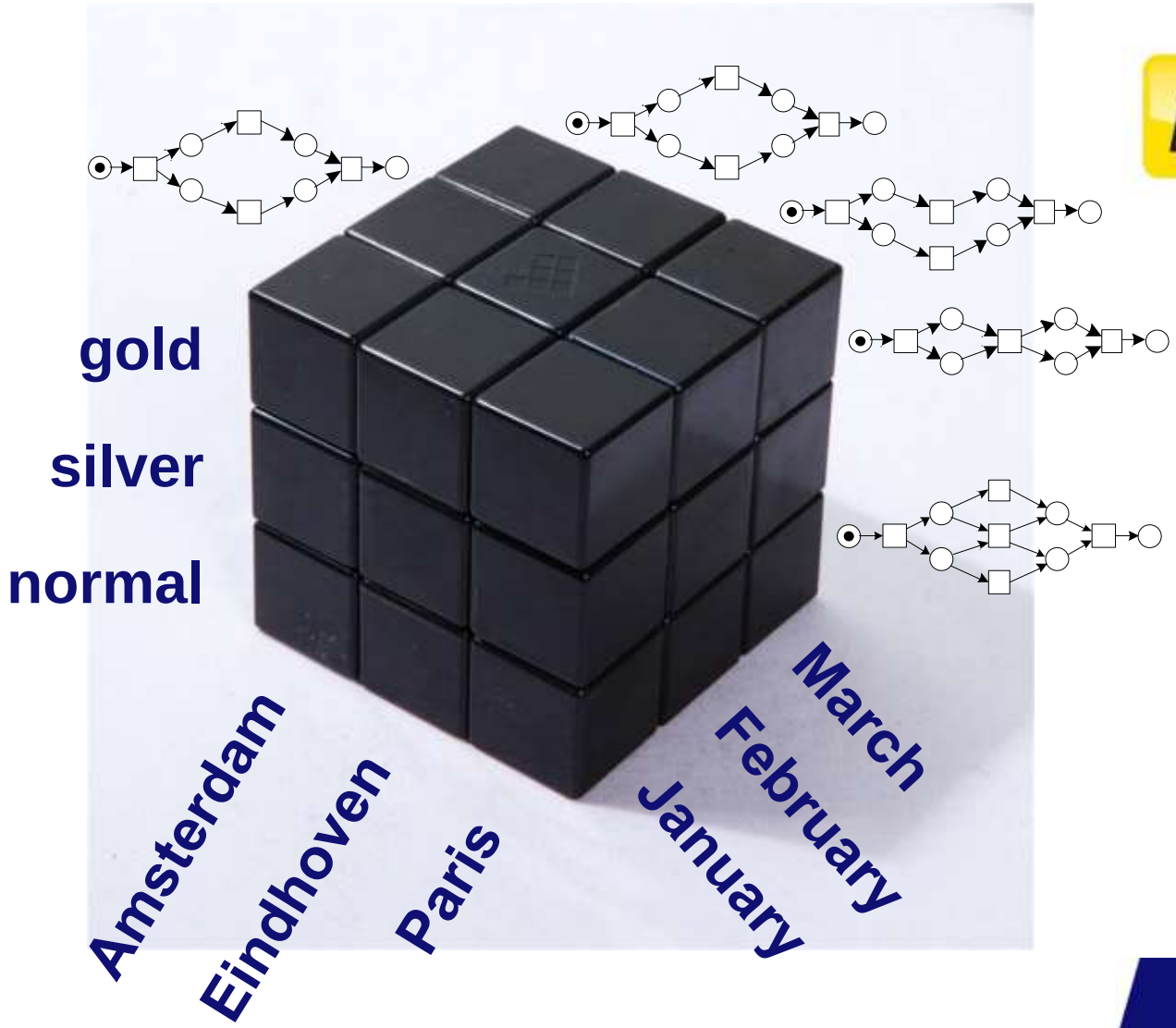
Process models are computed on two dimensional event data



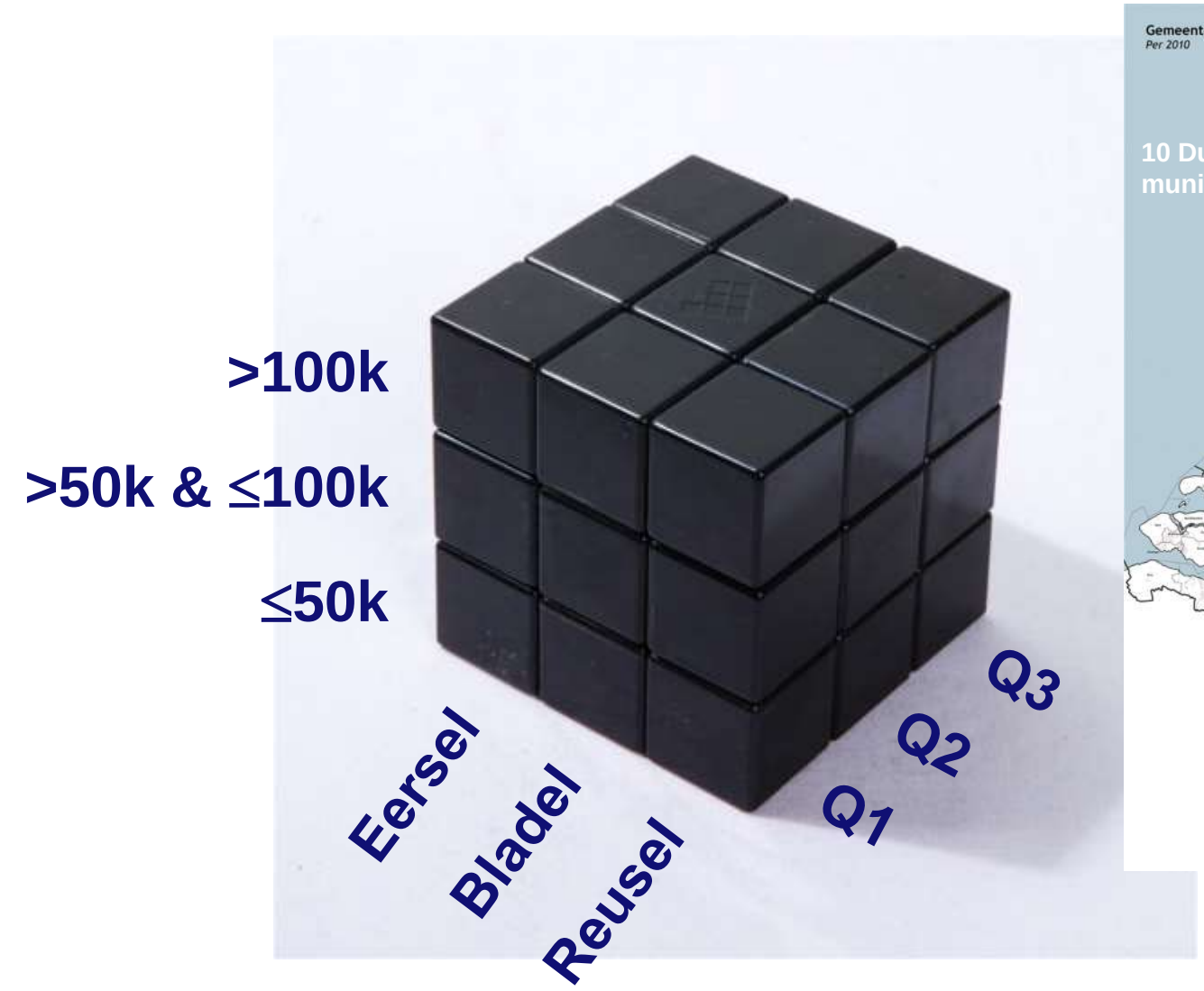
What are the differences?



Example: Hertz has 8,650 rental locations and different types of customers



Example: All Dutch municipalities need to handle building permits



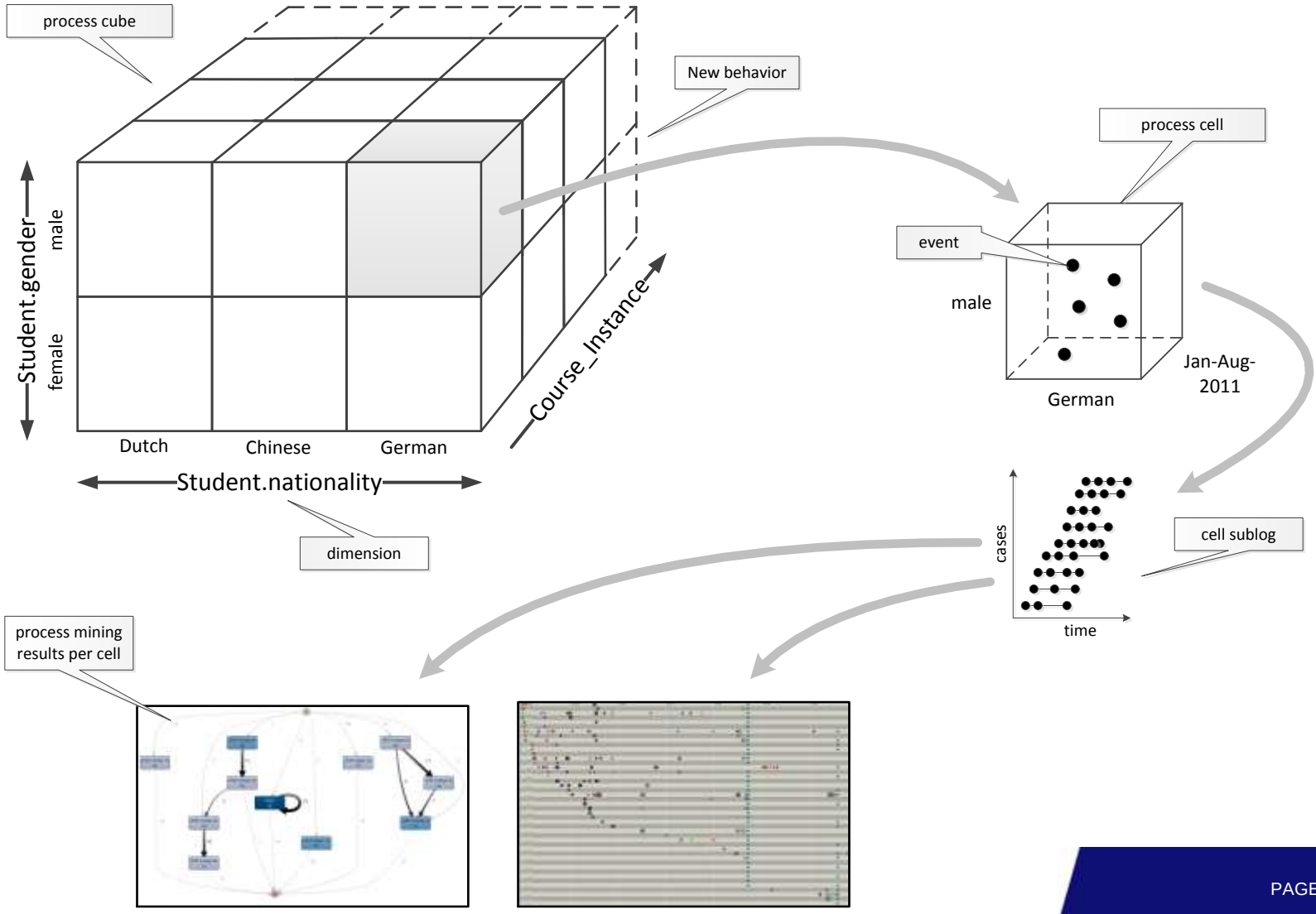
Example: Suncorp has different brands and different types of insurance

Example: students watching recorded video lectures and making exams

TU/e Technische Universiteit Eindhoven University of Technology

Where innovation starts

Process Cubes (OLAP for processes)



Overview data for a particular course

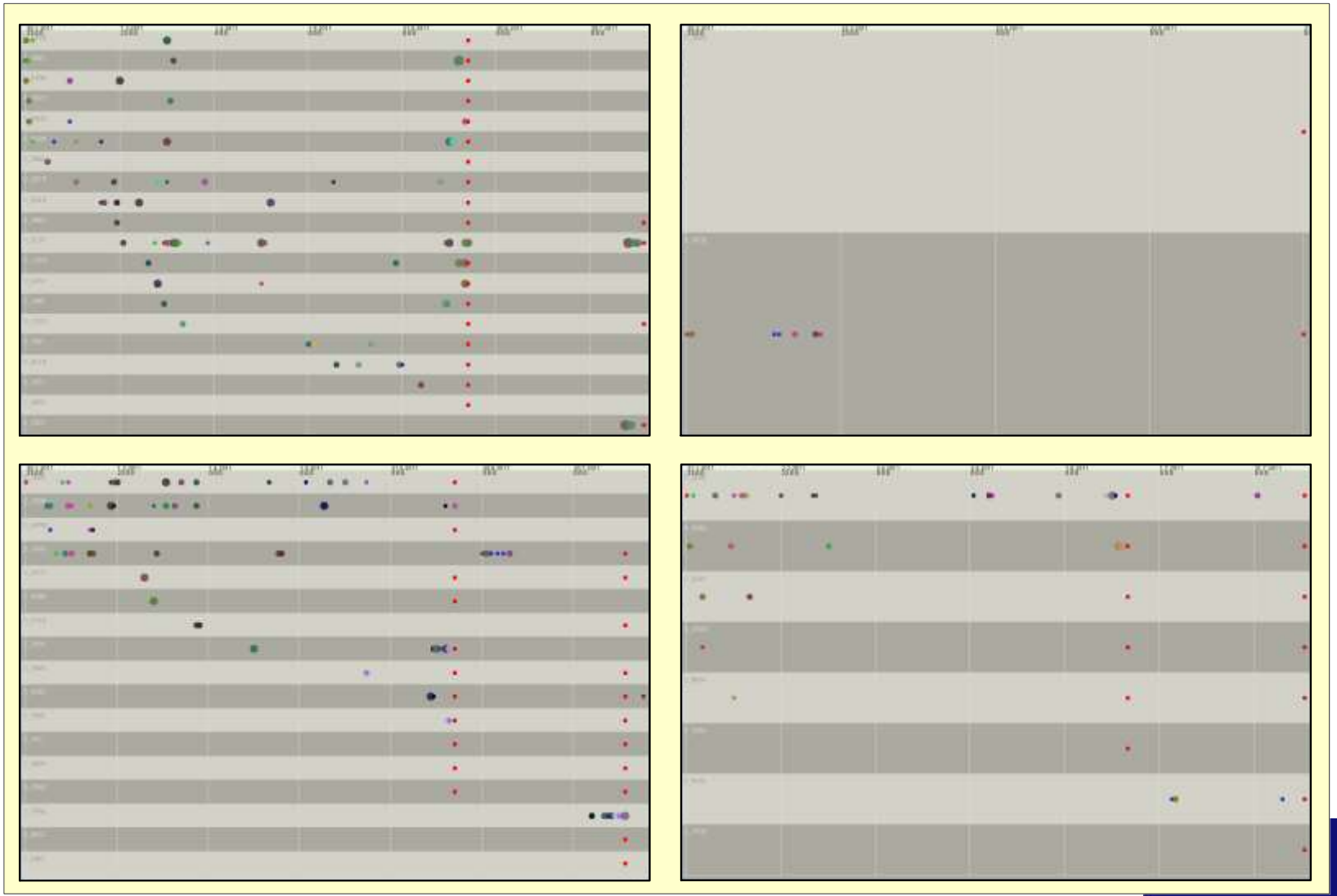
Drilling-down into a course instance

Dutch students

International students

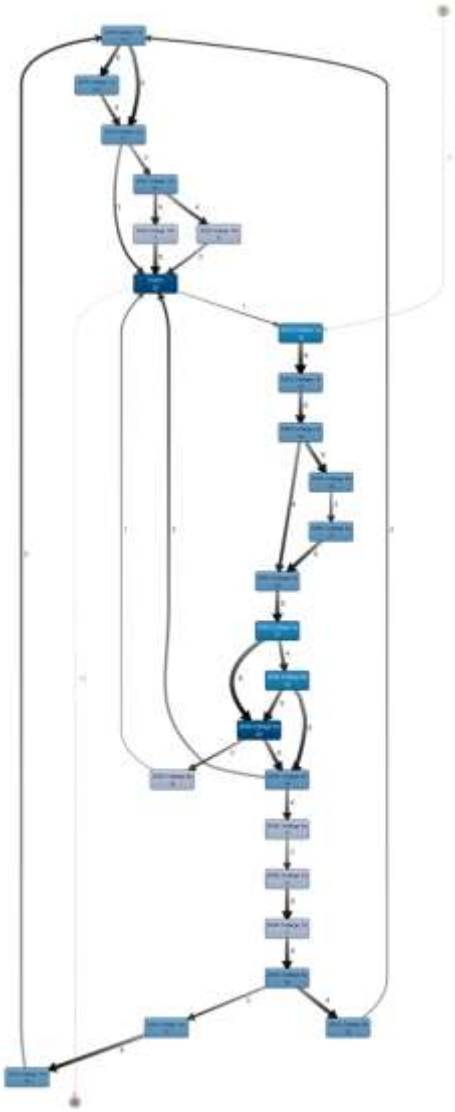
Students that passed

Students that did not pass

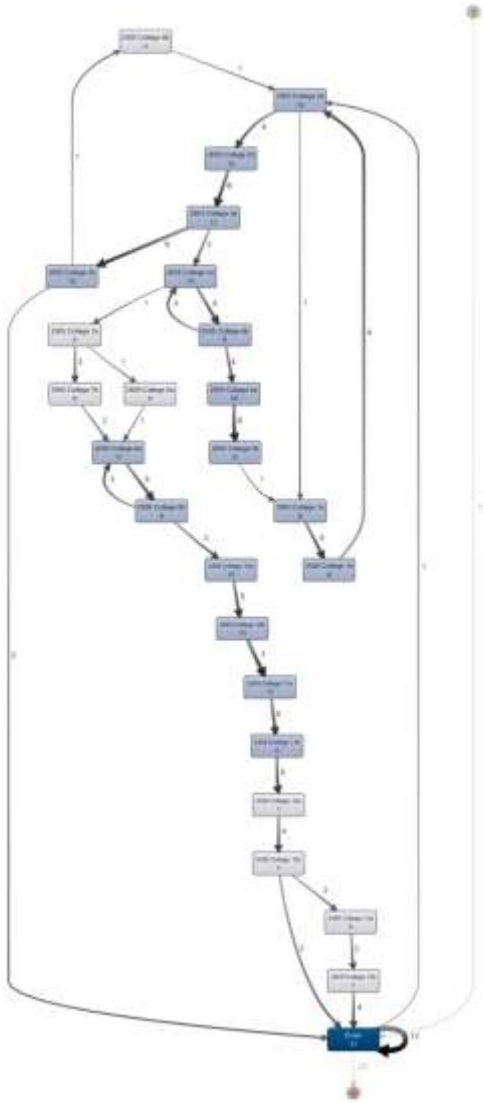


Comparing processes

PASSED



Fitness of event log wrt idealized model is 0.37

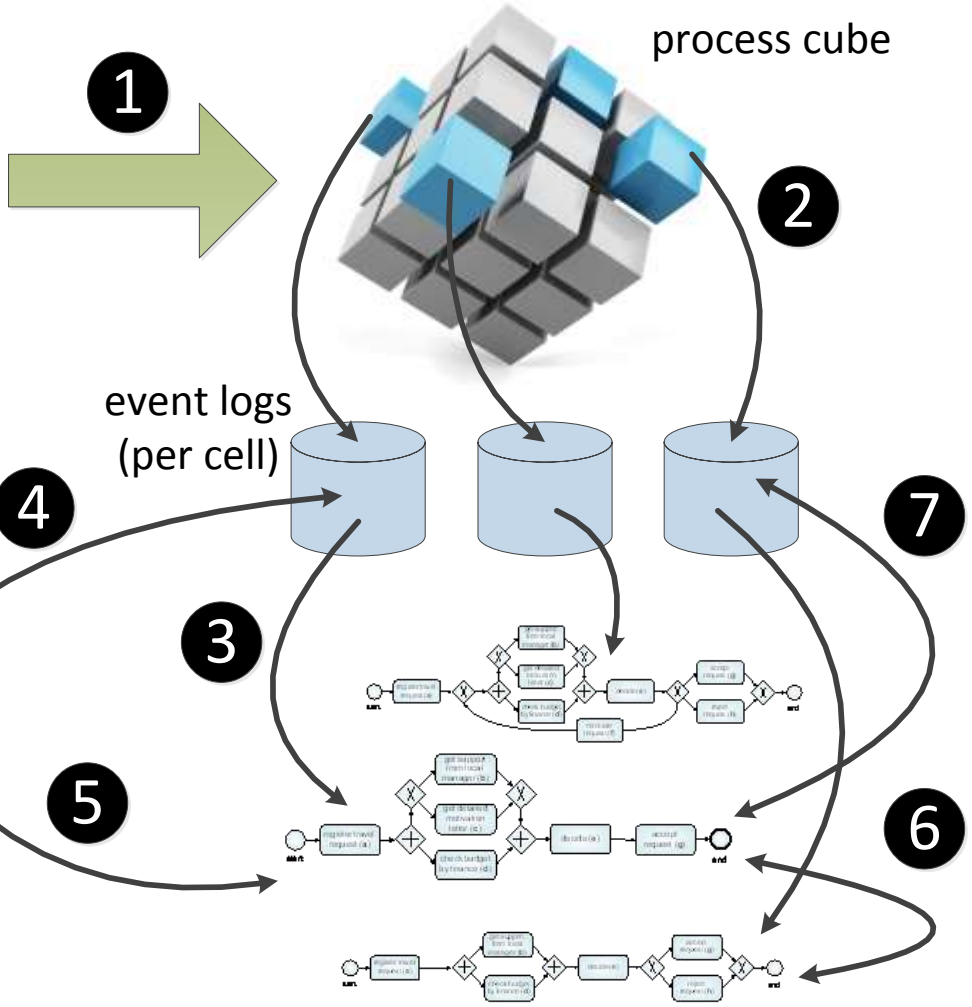


FAILED

Fitness of event log wrt idealized model is 0.28

Overview of approach

event data



1. Store events in the process cube.
2. Materialize the events in a cell as an event log that can be analyzed.
3. Automatically discover models per cell (e.g., a BPMN or UML model).
4. Check conformance by replaying event data on normative (process) models.
5. Compare discovered models and normative models.
6. Compare discovered models corresponding to different cells.
7. Compare different behaviors by replaying event data of one cell on another cell's model.

splitting event logs

Star Trek

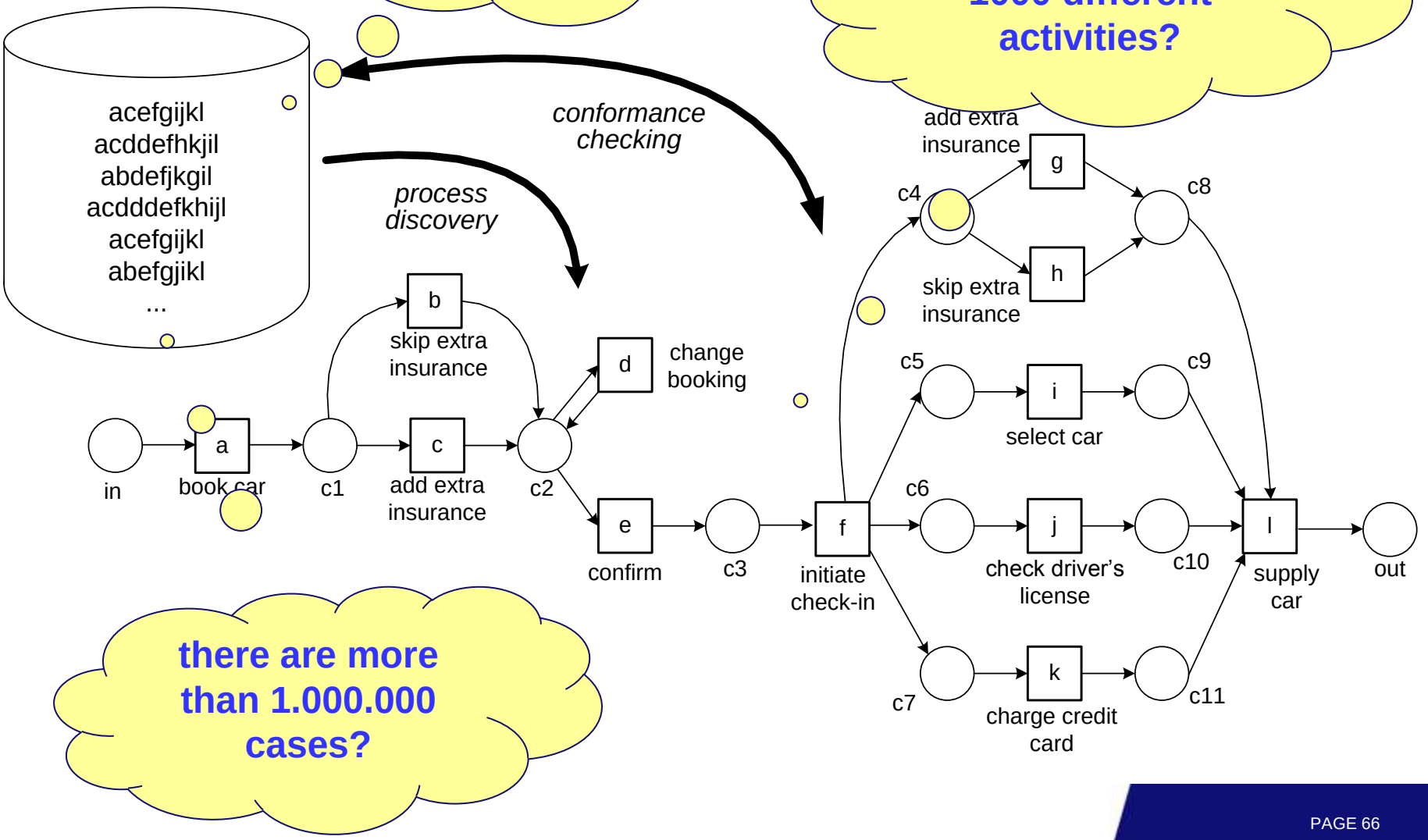
Big data

What if?

there are more than 100.000.000 events?

there are more than 1000 different activities?

there are more than 1.000.000 cases?



Decompose event log!

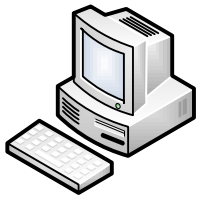
vertical or horizontal

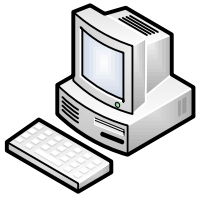
sets of cases

sets of activities

Vertical distribution I: Split cases arbitrarily

sets of cases





abcdeg
abdcefbcdeg
abdceg
abcdefbcdeg
abdcefbdceg
abcdefbdceg
abcdeg
abdceg
abdcefbdcefbdceg
abcdeg
abcdefbcdefbdceg
abcdefbdceg
abcdeg
abdceg
abdcefbcdeg
abcdeg

abcdeg
abdcefbcdeg
abdceg
abcdefbcdeg
abdcefbdceg
abcdefbdceg
abcdeg
abdceg

abdcefbdcefbdceg
abcdeg
abcdefbcdefbdceg
abcdefbdceg
abcdeg
abdceg
abdcefbcdeg
abcdeg

Vertical distribution II: Split cases based on a specific feature

abcdeg
abdcefbcdeg
abdceg
abcdefbcdeg
abdcefbdceg
abcdefbdceg
abcdeg
abdceg
abdcefbdcefbdceg
abcdeg
abcdefbcdefbdceg
abcdefbdceg
abcdeg
abdceg
abdcefbcdeg
abcdeg

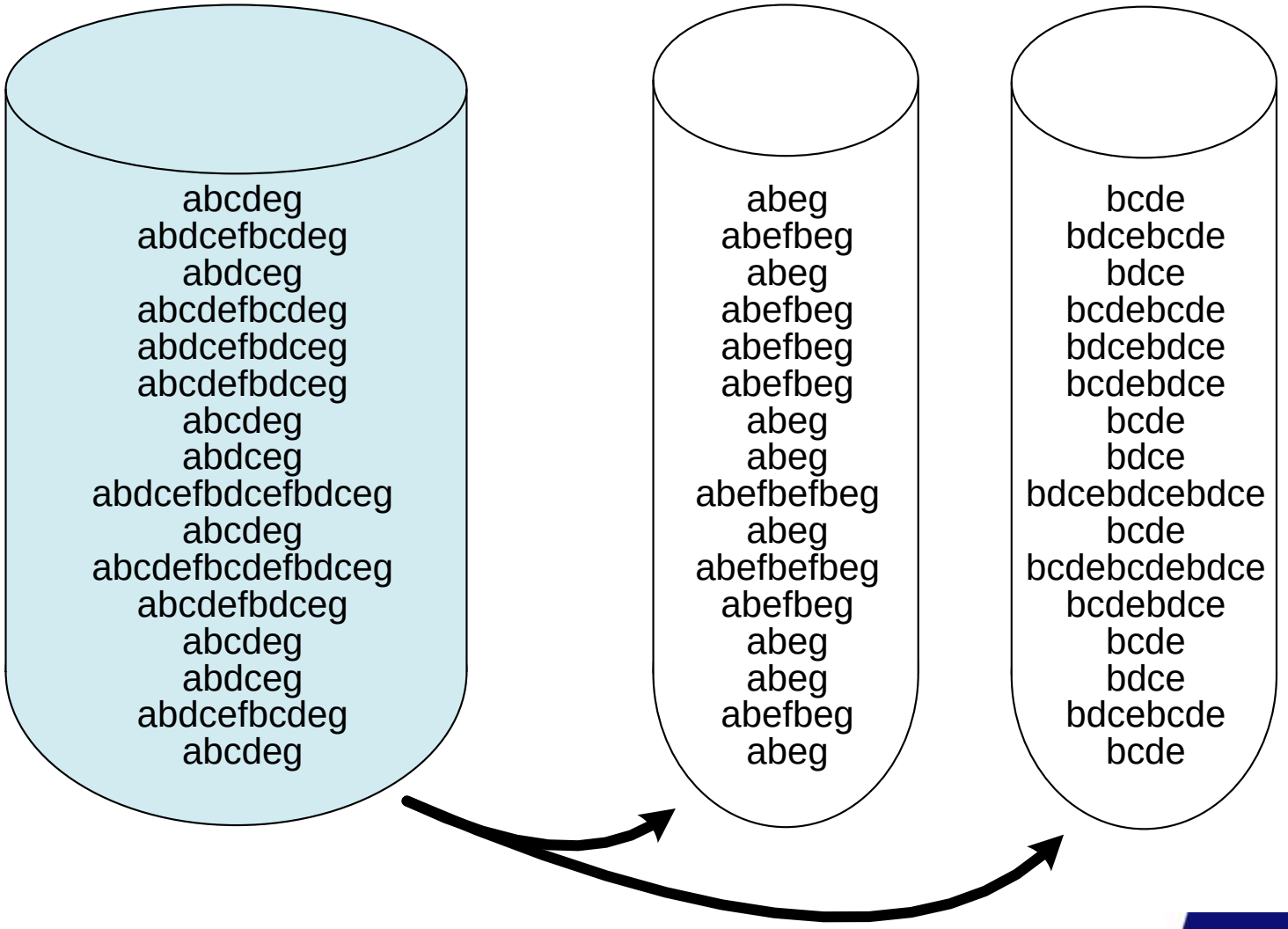
abcdeg
abdceg
abcdeg
abdceg
abcdeg
abcdeg
abdceg
abcdeg

abdcefbcdeg
abcdefbcdeg
abdcefbdceg
abcdefbdceg
abcdefbdceg
abdcefbcdeg

abdcefbdcefbdceg
abcdefbcdefbdceg

Horizontal distribution

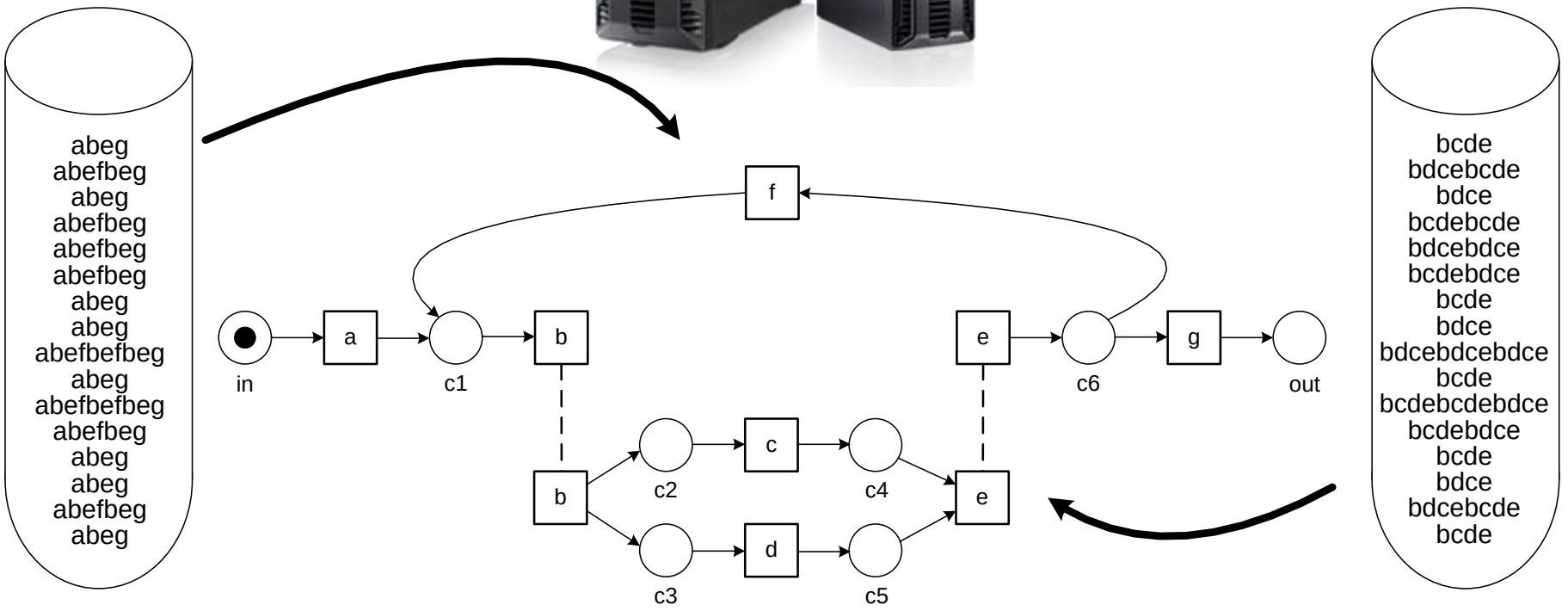
sets of activities



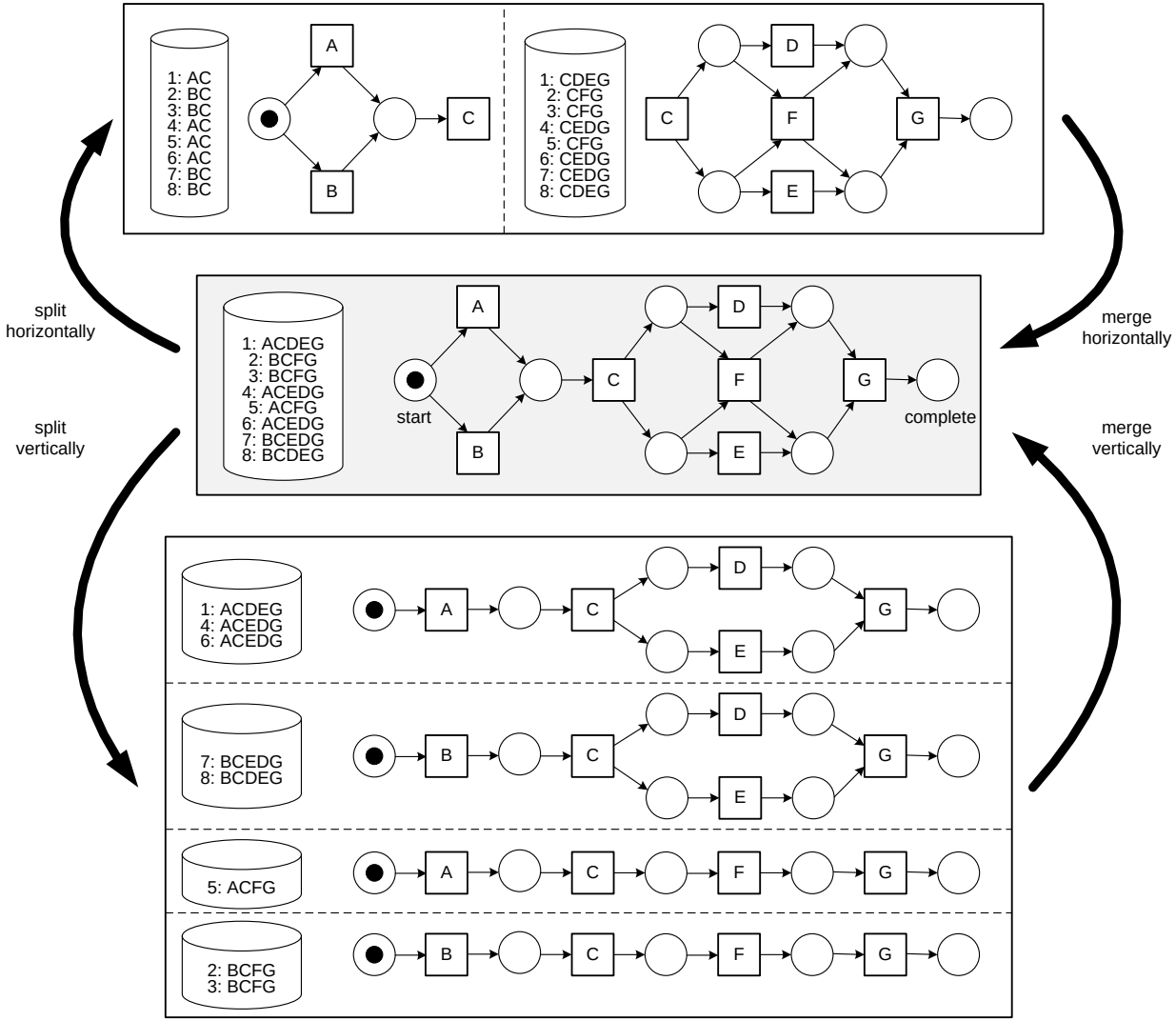
Horizontal distribution: The key idea

projected on a,b,e,f,g

projected on b,c,d,e

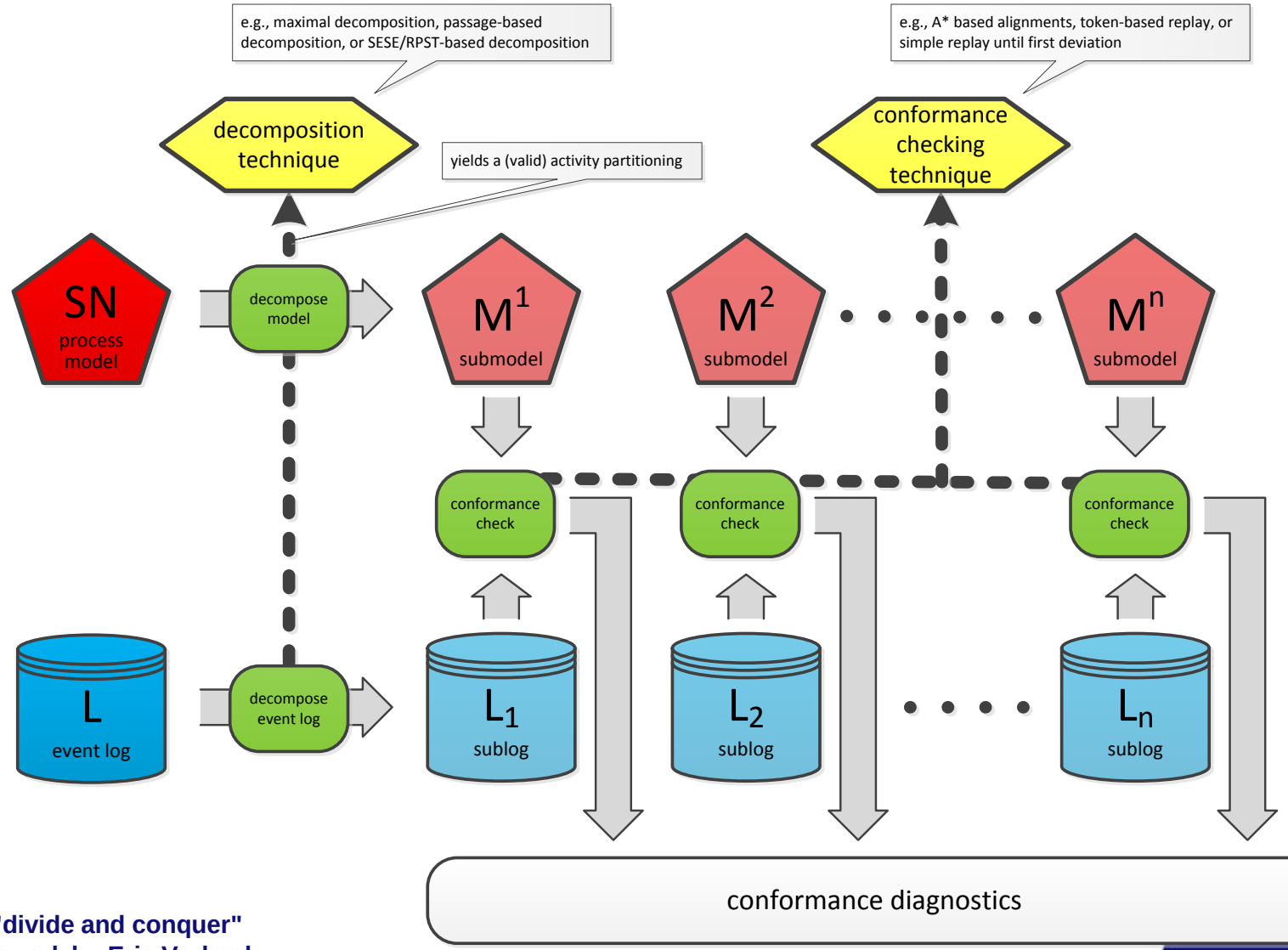


Two foundational ways of spitting event data: horizontal or vertical



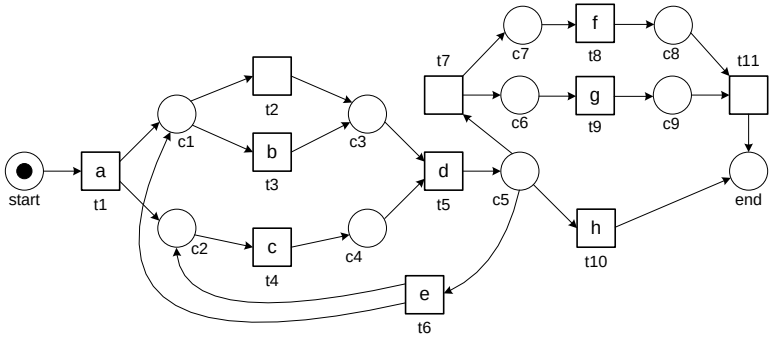
decomposed process mining

Decomposing Conformance Checking

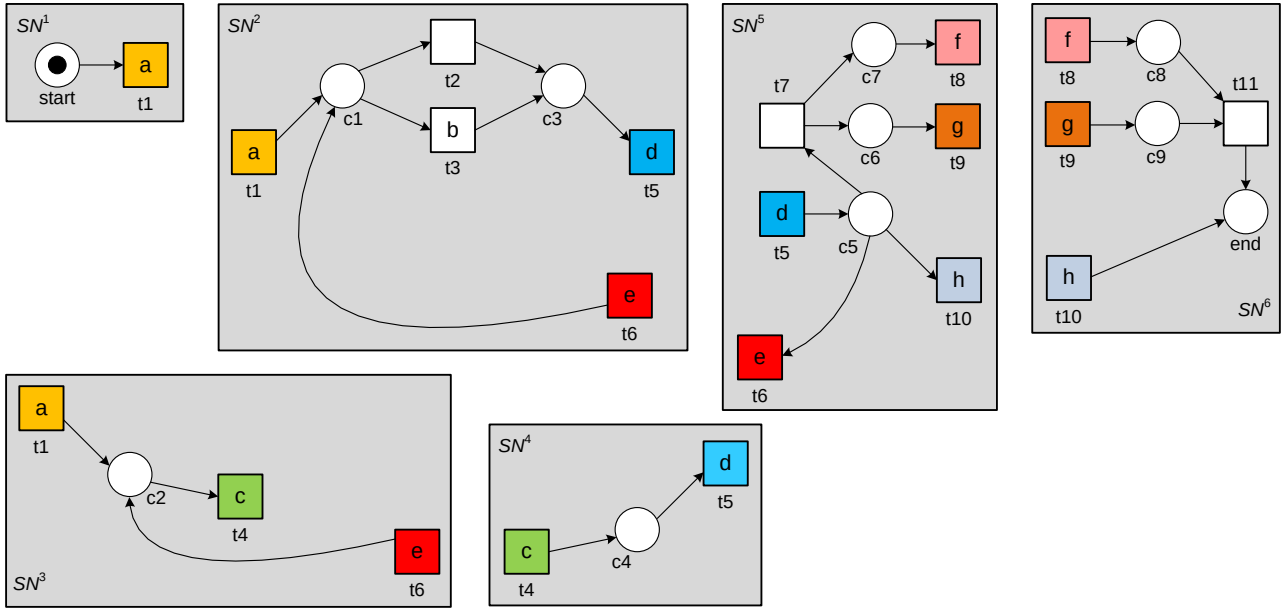


See "divide and conquer" framework by Eric Verbeek.

Example of a valid decomposition

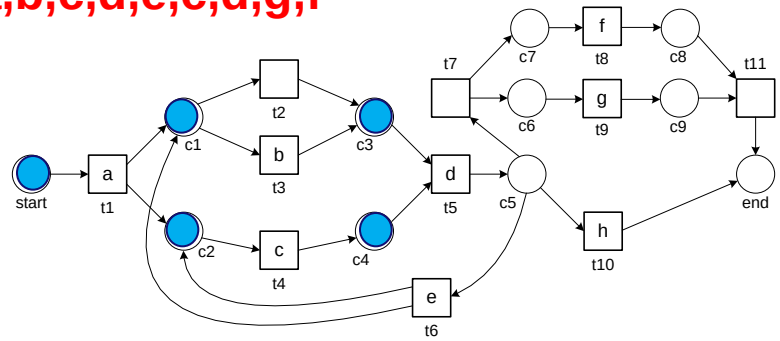


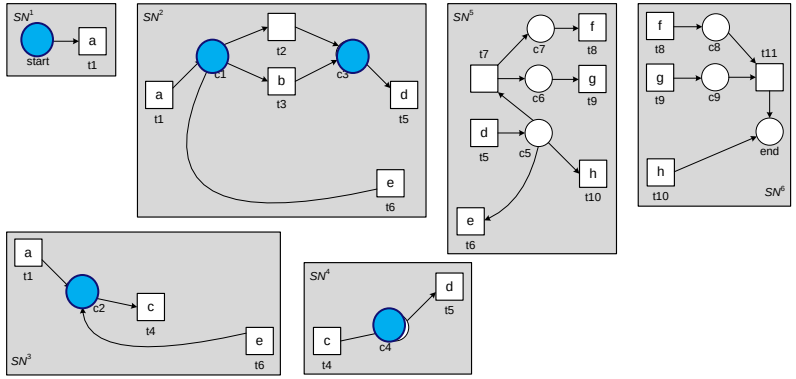
Log can be split in the same way!



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

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





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

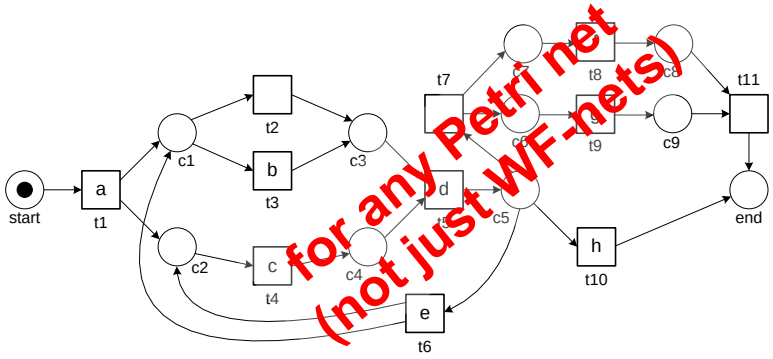
Etc.

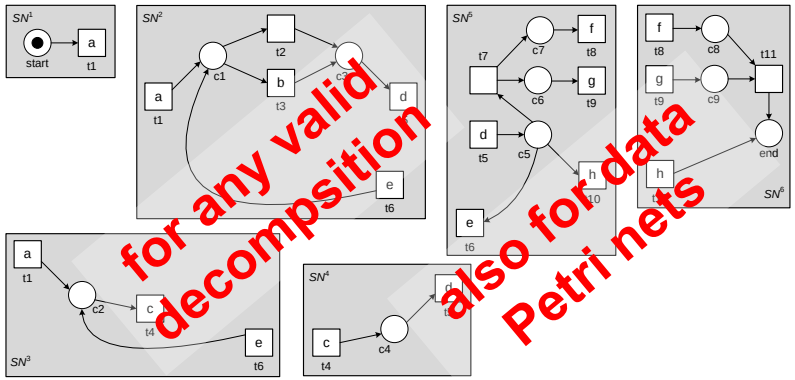
$$\gamma_3^1 = \begin{array}{|c|} \hline 1 \\ \hline a \\ \hline a \\ t1 \\ \hline \end{array} \qquad \gamma_3^2 = \begin{array}{|c|c|c|c|c|c|} \hline 1 & 2 & 4 & 5 & 7 & 8 \\ \hline a & b & d & e & \gg & d \\ \hline a & b & d & e & \tau & d \\ t1 & t3 & t5 & t6 & t2 & t5 \\ \hline \end{array} \qquad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

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

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**

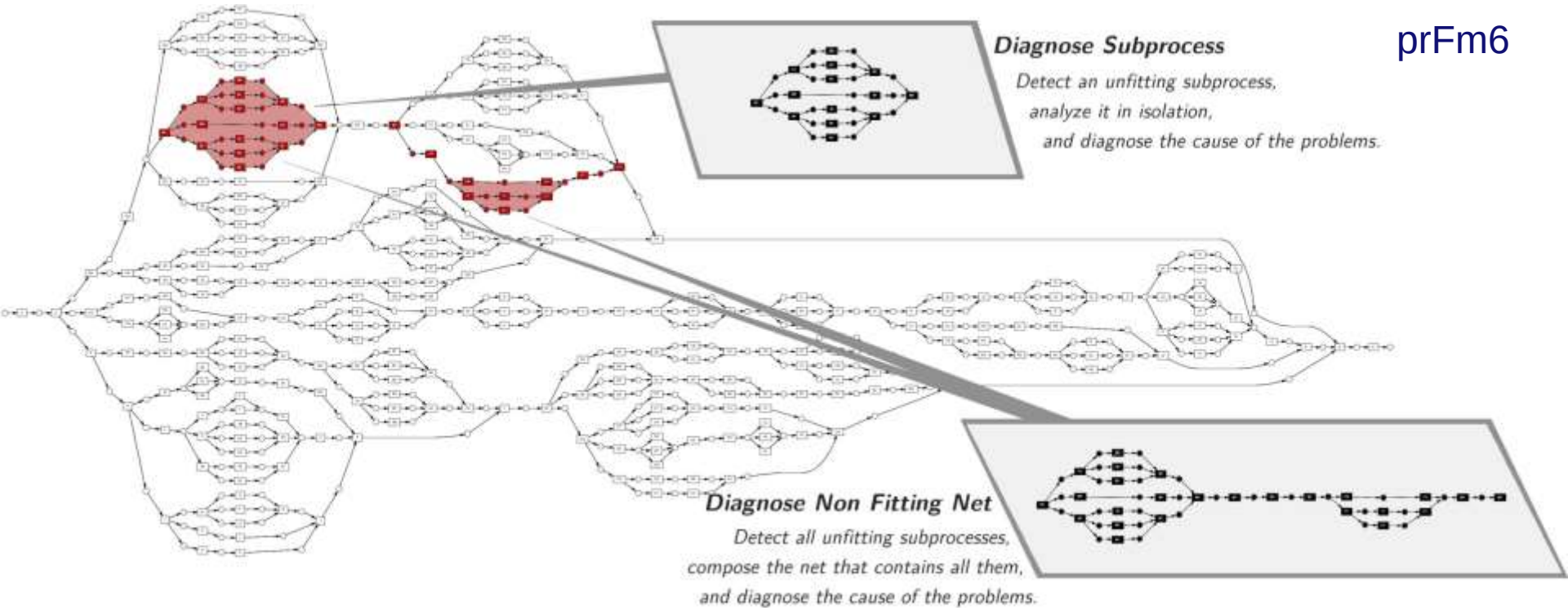




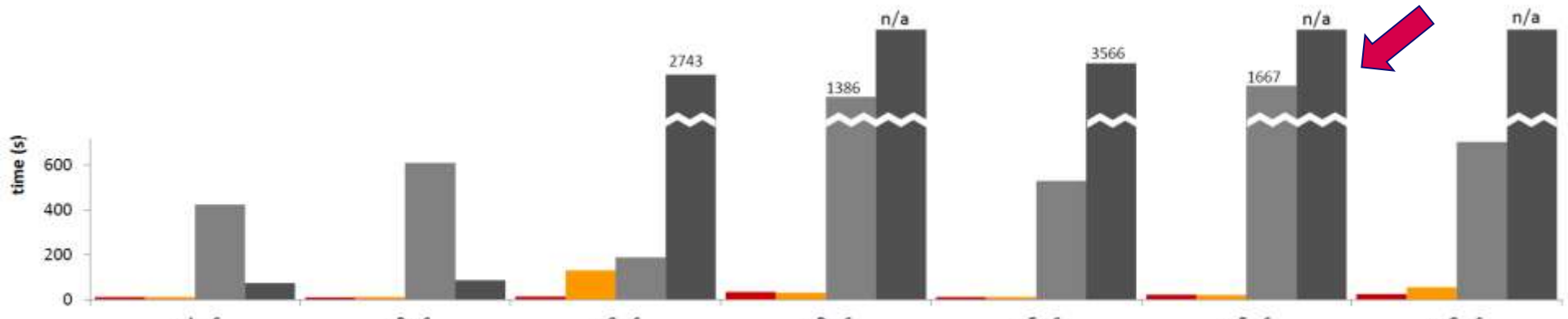
Wil van der Aalst, Decomposing Petri nets for process mining: A generic approach. Distributed and Parallel Databases, Volume 31, Issue 4, pp 471-507, 2013

Example

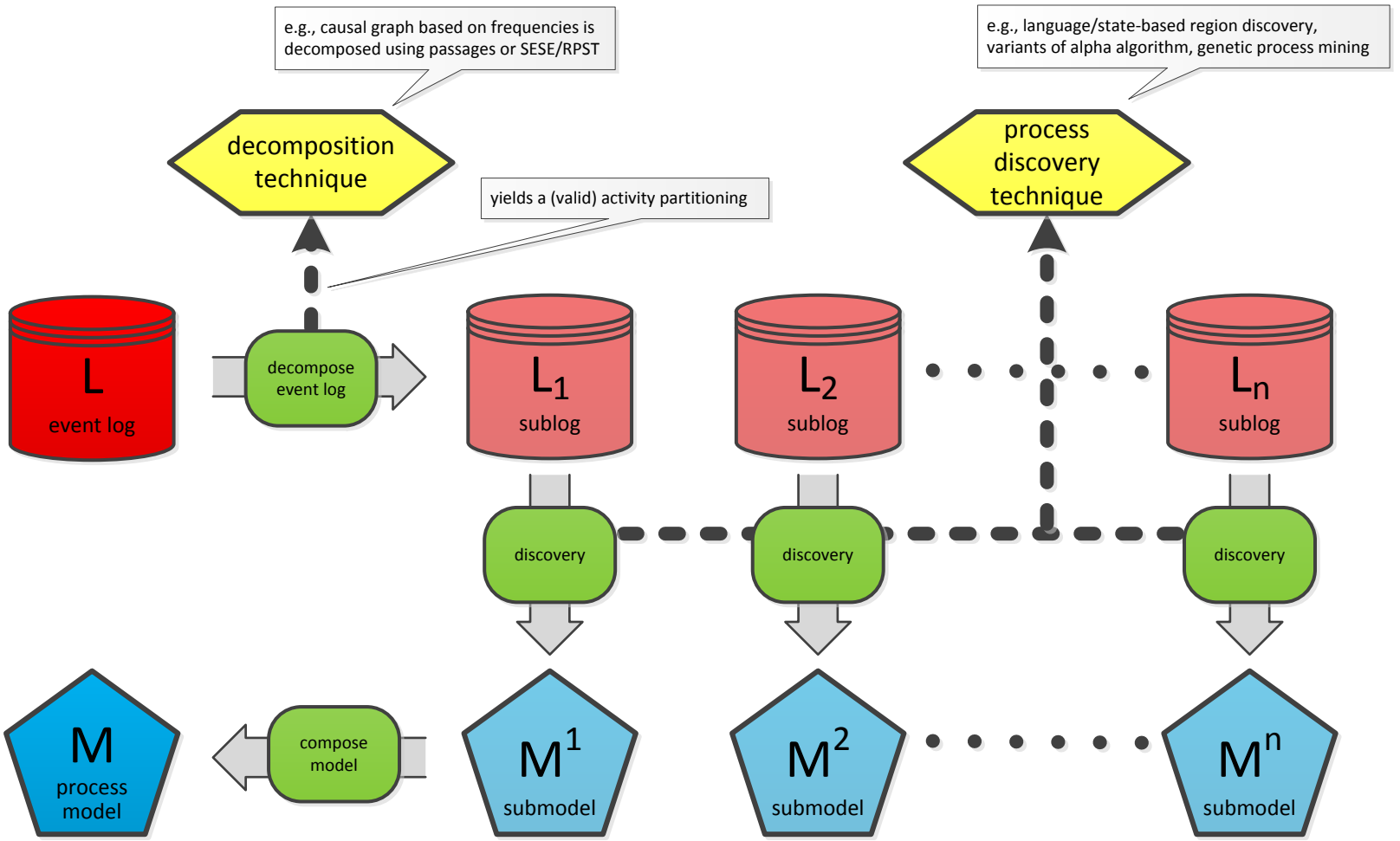
(work with Jorge Munoz-Gama and Josep Carmona)



prFm6



Decomposing Process Discovery



See "divide and conquer" framework by Eric Verbeek.

conclusion


Event data is everywhere!
Process are everywhere!
Why not connect them?
Process mining!
Challenges:
• comparative process mining
• Big event data, Big processes

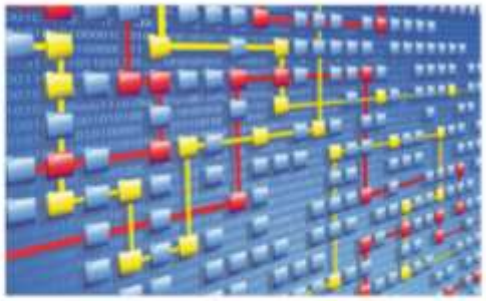
Process Mining: Data Science in Action

https://www.coursera.org/course/procmin

TU/e Technische Universiteit Eindhoven University of Technology

Process Mining: Data science in Action

Process mining is the missing link between model-based process analysis and data-oriented analysis techniques. Through concrete data sets and easy to use software the course provides data science knowledge that can be applied directly to analyze and improve processes in a variety of domains.



About the Course

Data science is the profession of the future, because organizations that are unable to use (big) data in a smart way will not survive. It is not sufficient to focus on data storage and data analysis. The data scientist also needs to relate data to process analysis. **Process mining bridges the gap between traditional model-based process analysis (e.g., simulation and other business process management techniques) and data-centric analysis techniques such as machine learning and data mining.** Process mining seeks the confrontation between event data (i.e., observed behavior) and process models (hand-made or discovered automatically). This technology has become available only recently, but it can be applied to any type of operational processes (organizations and systems). Example applications include: analyzing treatment processes in hospitals, improving customer service processes in a multinational, understanding the browsing behavior of customers using a booking site, analyzing failures of a baggage handling system, and improving the user interface of an X-ray machine. All of these applications have in common that dynamic behavior needs to be related to process models. Hence, we refer to this as "data science in action".

The course explains the key analysis techniques in process mining. Participants will learn various process discovery algorithms. These can be used to automatically learn process models from raw event data. Various other process analysis techniques that use event data will be presented. Moreover, the course will provide **easy-to-use software**, **real-life data sets**, and **practical skills** to **directly apply the theory** in

Sessions

Nov 12th 2014 - Dec 24th 2014

Starts in 3 months

Eligible for

Statement of Accomplishment

Course at a Glance

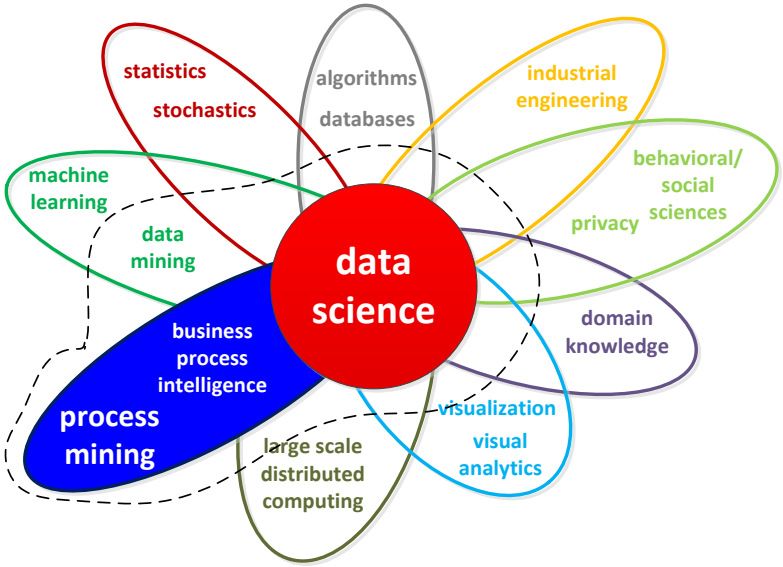
- 6 weeks of study
- 4-6 hours of work / week
- English
- English subtitles

Instructors

Wil van der Aalst
Eindhoven University of Technology

Categories

First Massive Open Online Course (MOOC) on Process Mining



coursera

TU/e Technische Universiteit Eindhoven University of Technology

Data Science Center Eindhoven

Wil M. P. van der Aalst

Process Mining

Discovery, Conformance and Enhancement of Business Processes

More and more information about business processes is recorded by information systems in the form of so-called "event logs". Despite the omnipresence of such data, most organizations diagnose problems based on fiction rather than facts. Process mining is an emerging discipline based on process model-driven approaches and data mining. It not only allows organizations to fully benefit from the information stored in their systems, but it can also be used to check the conformance of processes, detect bottlenecks, and predict execution problems.

Wil van der Aalst delivers the first book on process mining. It aims to be self-contained while covering the entire process mining spectrum from process discovery to operational support. In Part I, the author provides the basics of business process modeling and data mining necessary to understand the remainder of the book. Part II focuses on process discovery as the most important process mining task. Part III moves beyond discovering the control flow of processes and highlights conformance checking, and organizational and time perspectives. Part IV guides the reader in successfully applying process mining in practice, including an introduction to the widely used open-source tool ProM. Finally, Part V takes a step back, reflecting on the material presented and the key open challenges.

Overall, this book provides a comprehensive overview of the state of the art in process mining. It is intended for business process analysts, business consultants, process managers, graduate students, and BPM researchers.

Features and Benefits:

- First book on process mining, bridging the gap between business process modeling and business intelligence.
- Written by one of the most influential and most-cited computer scientists and the best-known BPM researcher.
- Self-contained and comprehensive overview for a broad audience in academia and industry.
- The reader can put process mining into practice immediately due to the applicability of the techniques and the availability of the open-source process mining software ProM.

Computer Science

ISBN 978-3-642-19344-6

springer.com

van der Aalst Process Mining

ProM process mining workbench

Wil M. P. van der Aalst

Process Mining

Discovery, Conformance and Enhancement of Business Processes

www.processmining.org

www.win.tue.nl/ieeetfpm/

Springer