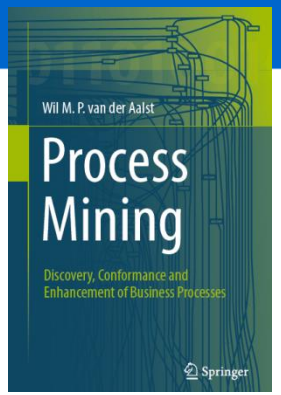
Towards a Process Scientist: Dealing with Big Data and Processes in a Comprehensive Manner

prof.dr.ir. Wil van der Aalst
www.processmining.org



TU/e Technische Universiteit Eindhoven University of Technology

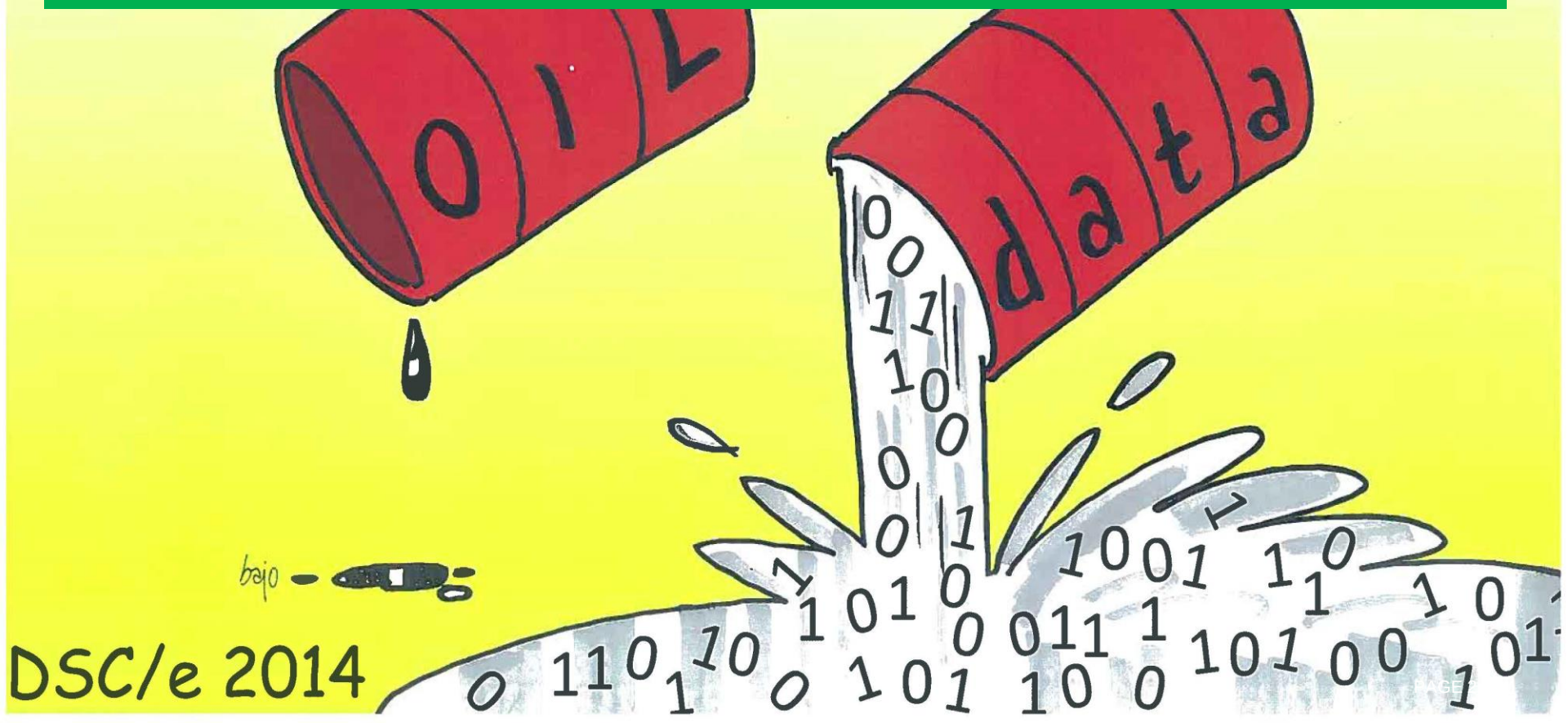
Where innovation starts

this seminar

previous seminars

21st Century

In the last 10 minutes we generated more data than from prehistoric times until 2003!



We are all generating event data!

taking the train

refueling your car

buying a coffee

adjusting the temperature in your home

making a phone call

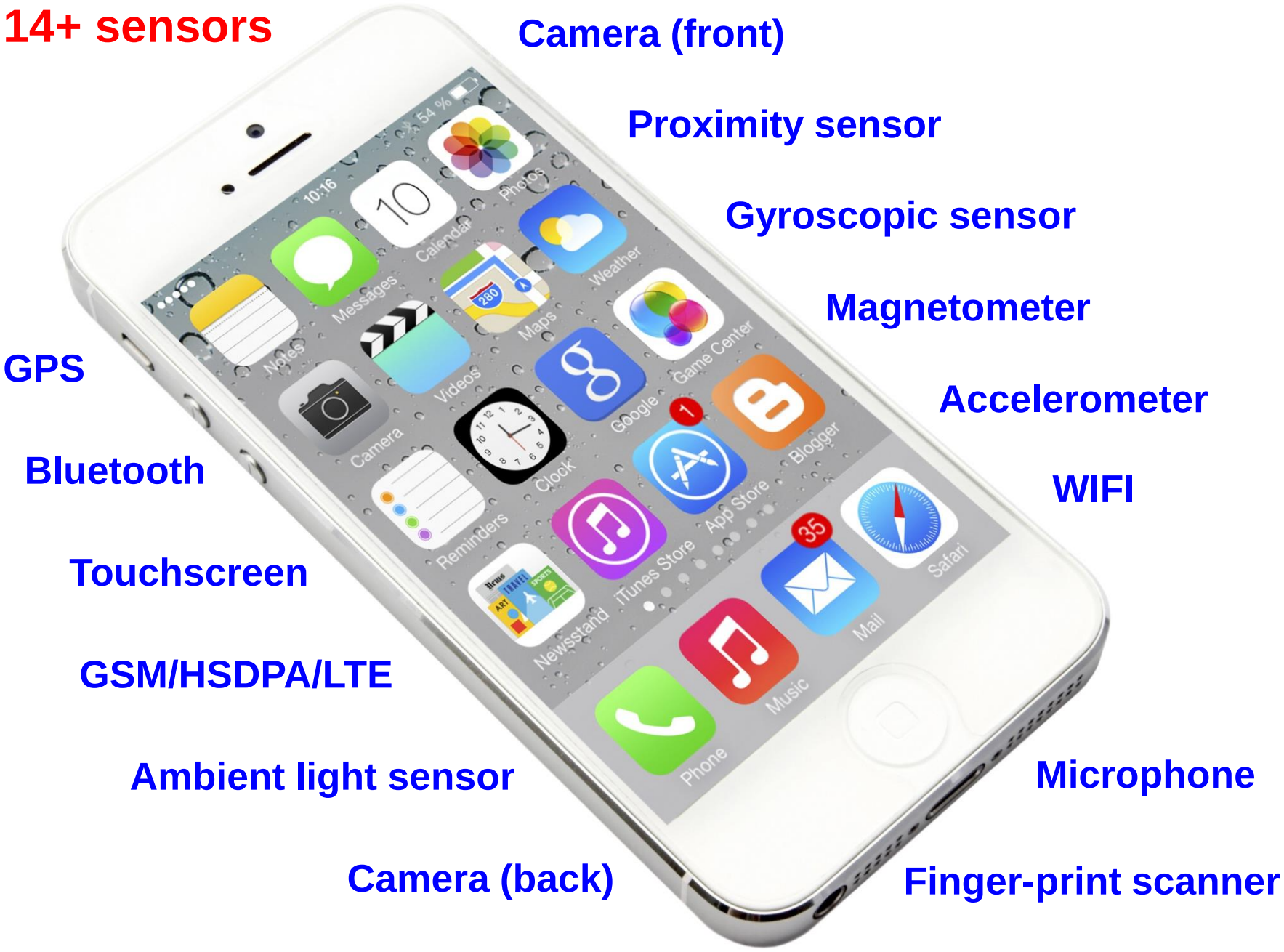
getting a speeding ticket

making a phone call

sending an e-mail

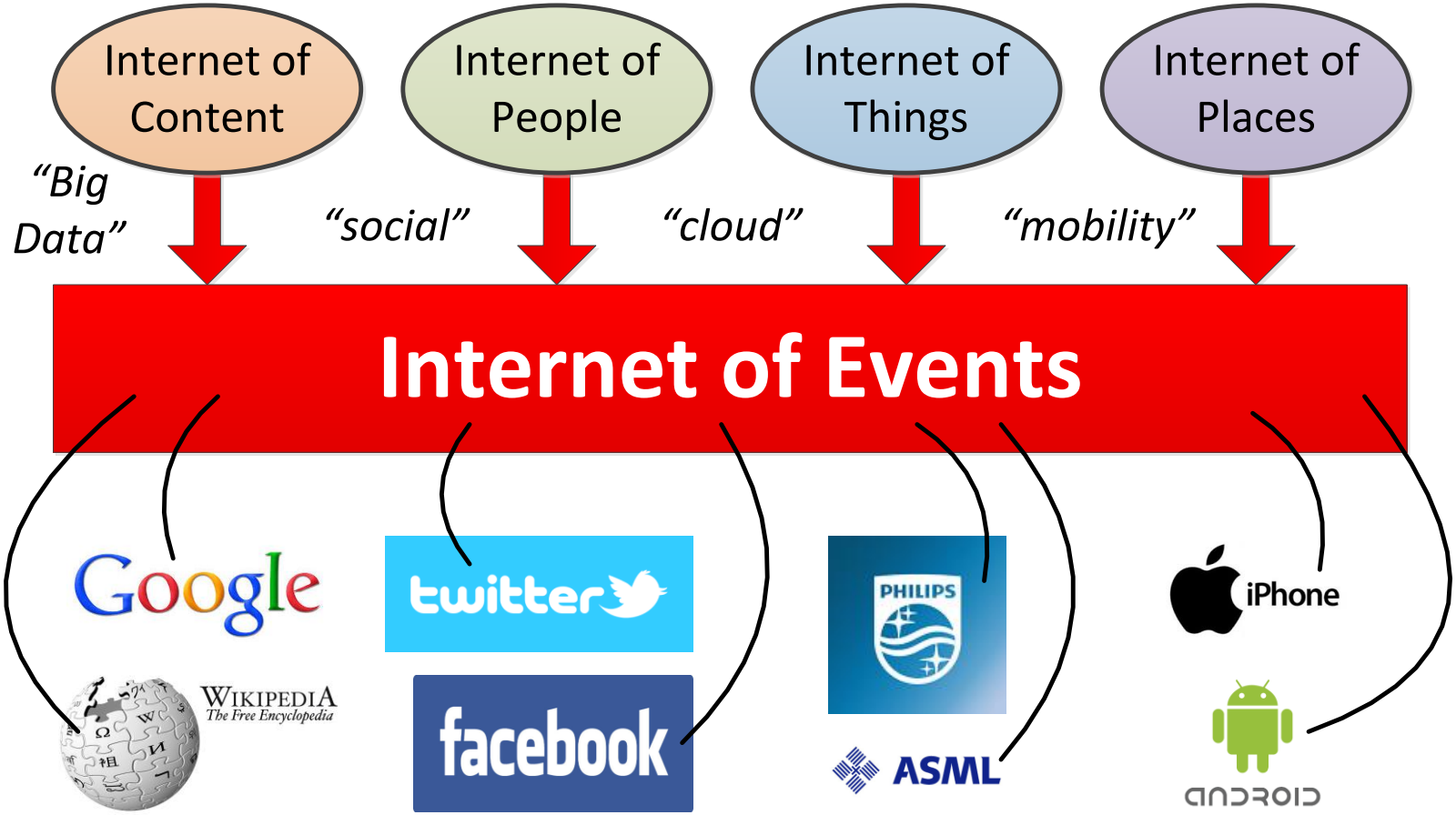
making an appointment

watching this lecture


14+ sensors
Camera (front)
Proximity sensor
Gyroscopic sensor
Magnetometer
GPS
Accelerometer
Bluetooth
WIFI
Touchscreen
GSM/HSDPA/LTE
Ambient light sensor
Microphone
Camera (back)
Finger-print scanner

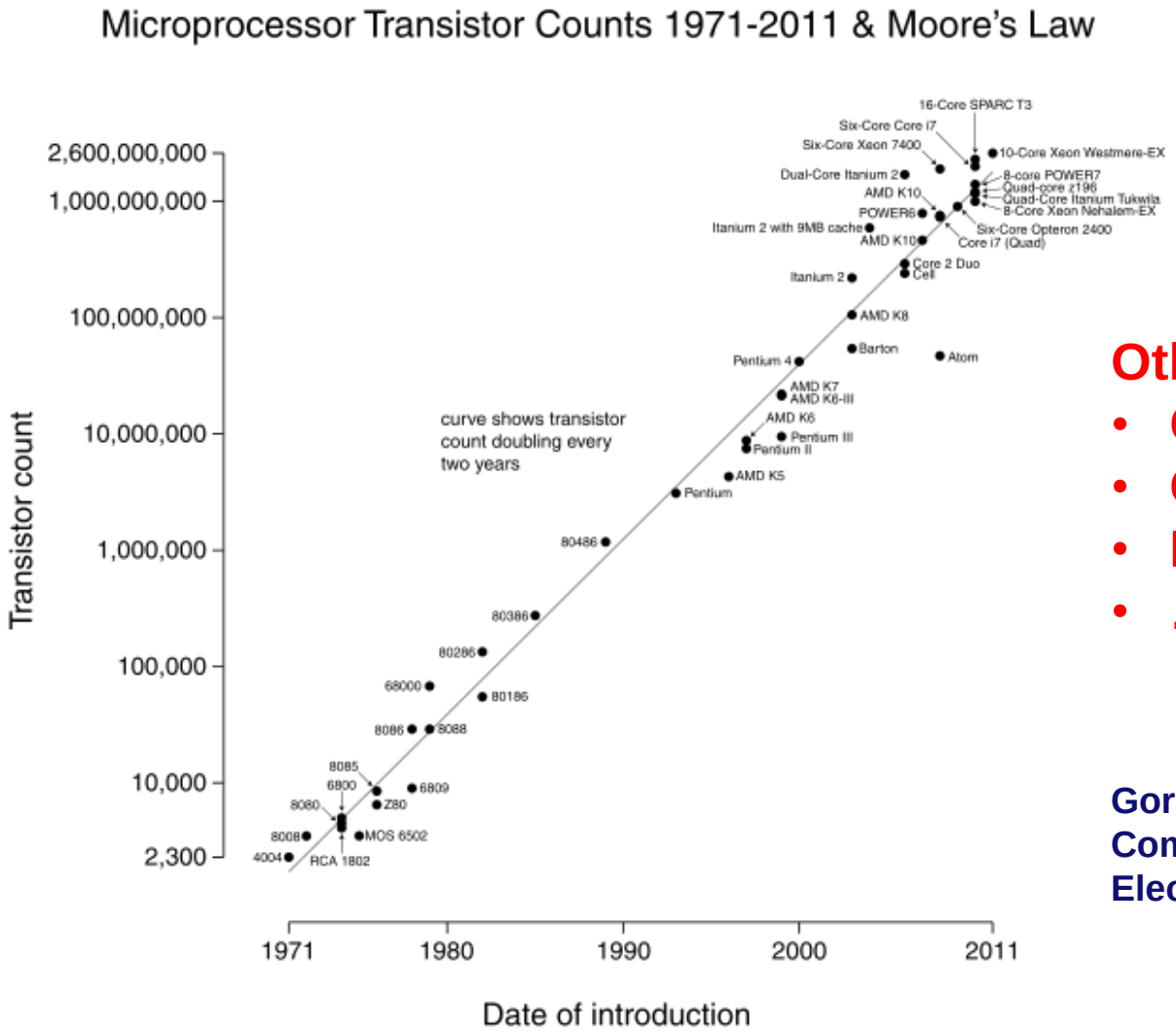

Internet of Events

Always On
Anything, Anytime, Anywhere




DATA
BIG

Moore's law: 2^{20} = 1.048.576x in 40 years



Other examples:

- **Computing power**
- **Capacity of disks**
- **Bytes per dollar**
- **…**

Gordon E. Moore, Cramming More Components onto Integrated Circuits, Electronics, pp. 114–117, 1965.

Diagram by Wgsimon/CC BY 3.0

Question

40 years ago it took approximately 7 hours to go from Amsterdam to New York by airplane.

How long would it take today if transportation technology would have followed Moore's law?

Answer

0.0240 seconds

Question

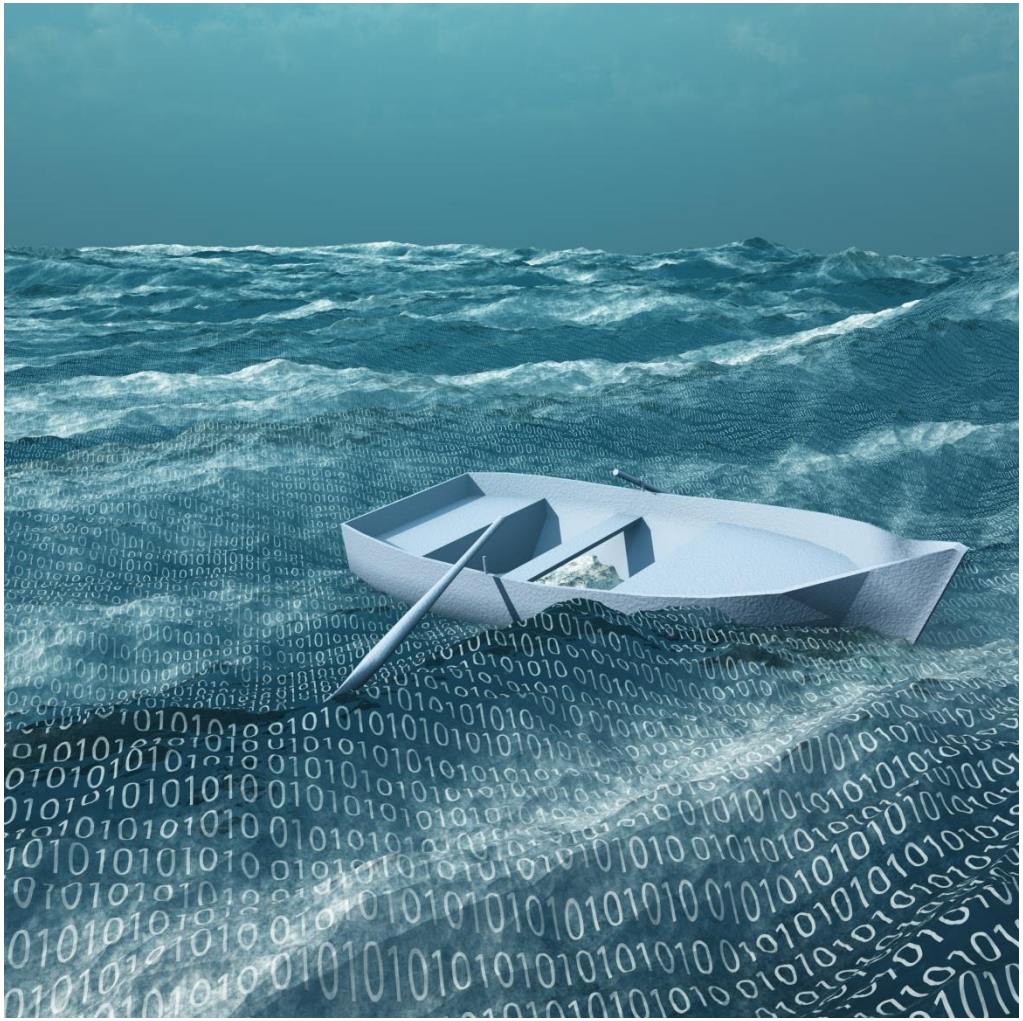
40 years ago it took approximately 4000 liters of petrol to drive around the world.

How much petrol would it take today if transportation technology would have followed Moore's law?

Answer

0.0038 liters

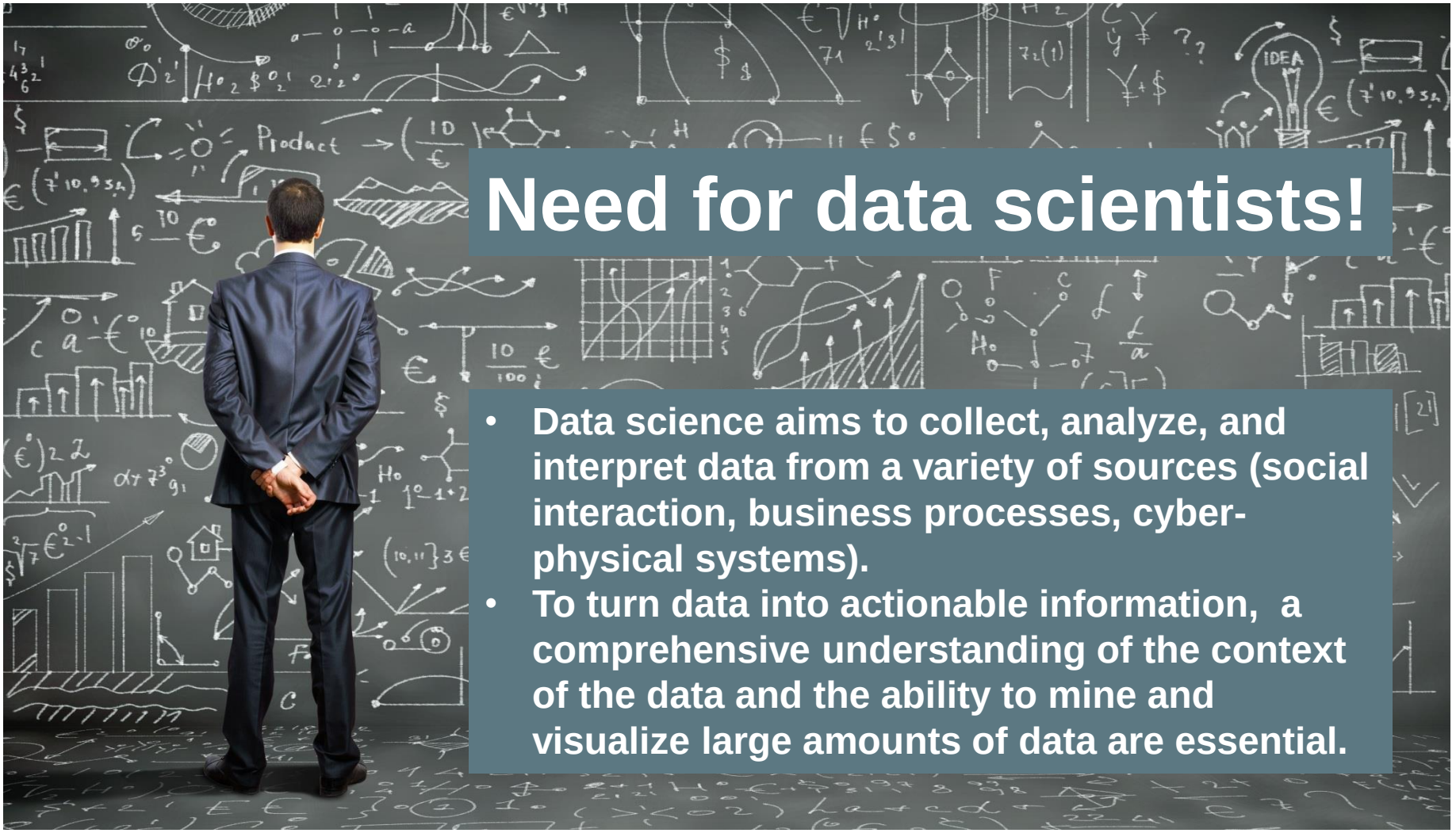
Drowning in data



How to extract real value from event data?

The four V's of Big Data

Data does not have to be "Big" to be challenging!



Need for data scientists!

- Data science aims to collect, analyze, and interpret data from a variety of sources (social interaction, business processes, cyber-physical systems).
- To turn data into actionable information, a comprehensive understanding of the context of the data and the ability to mine and visualize large amounts of data are essential.

generic data science question 1/4

generic data science question 2/4

generic data science question 3/4

generic data science question 4/4

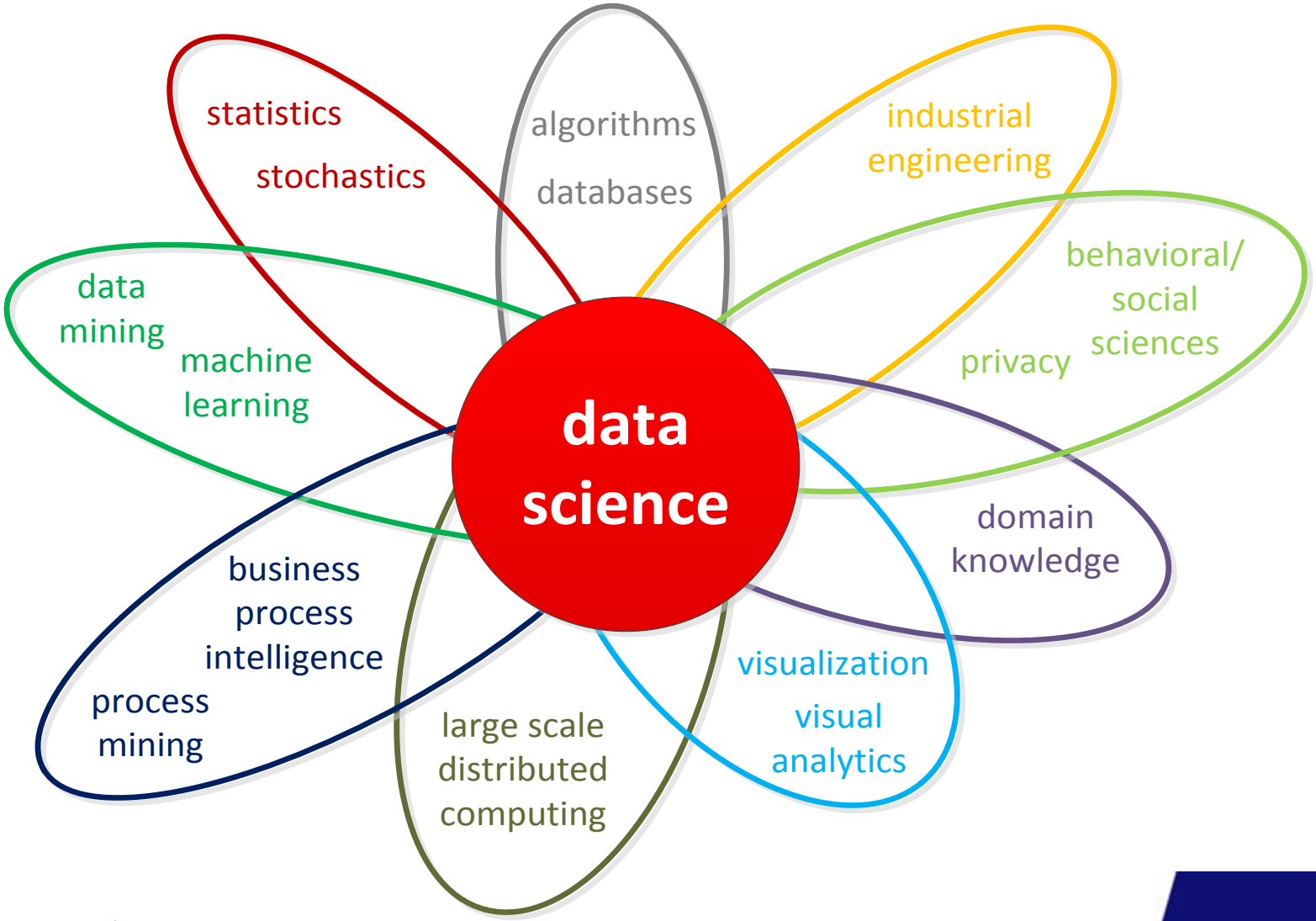
Imagine a hospital treating patients with lung cancer:

- **Patients complain about long waiting times.**
- **Staff complains about unbalanced workloads.**
- **There seem to be many deviations from the official medical guideline.**
- **Costs need to be reduced without endangering quality.**

- **What happened?**
- **Why did it happen?**
- **What will happen?**
- **What is the best that can happen?**


How are X-ray machines really used?
Can we predict that the machine will break down next week?
Why and when do X-ray machines malfunction?
Which parts need to be improved?
Which components should be replaced?
from the organizational level to the hardware/software level

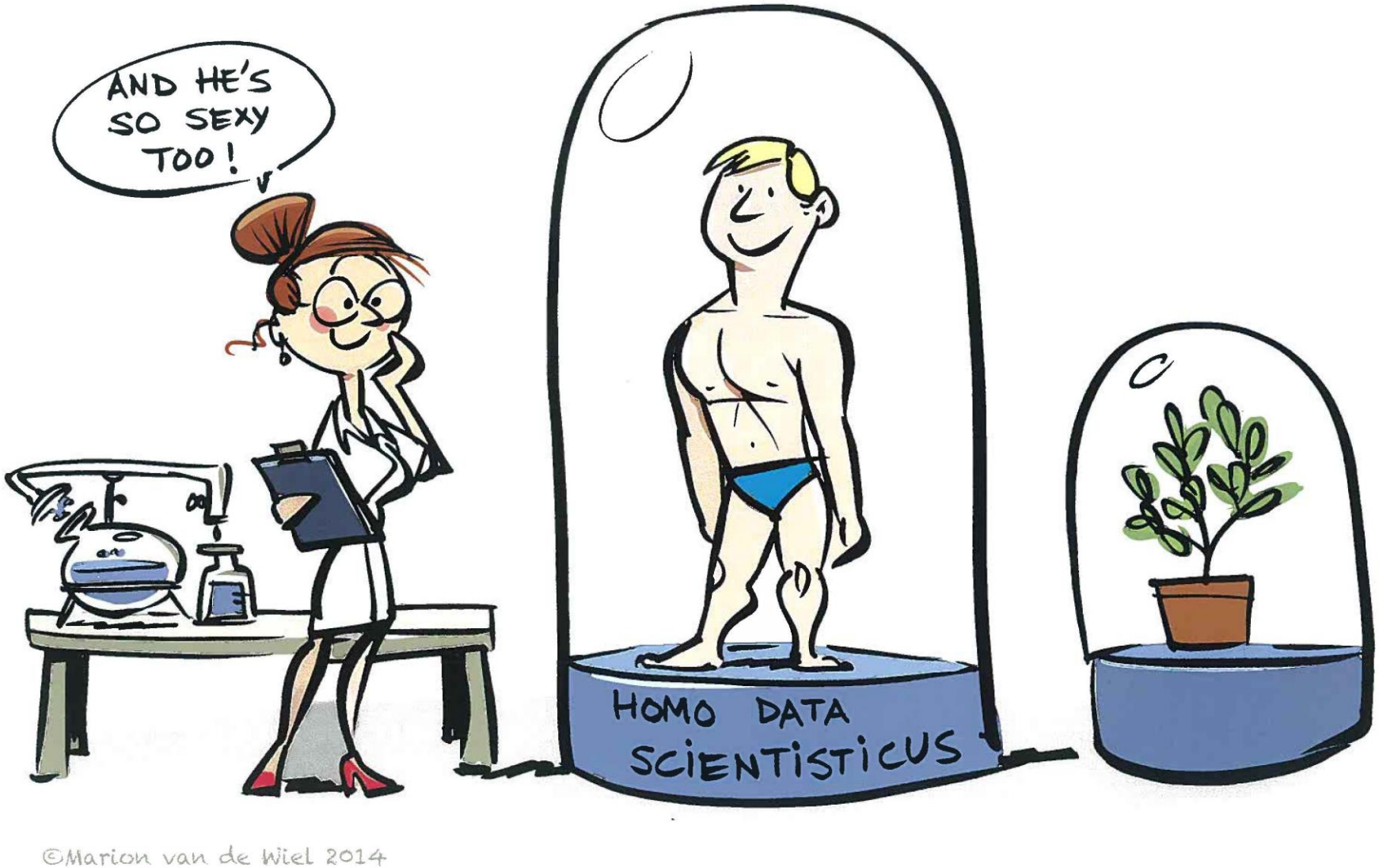
Data science skills needed to answer such questions




DSC/e
http://www.tue.nl/dsce/


Protocols; LOS in Hospitals


DSC/e
TU/e
Data Science Center Eindh
DATA SCIENTIST IS A NEW BREED....
AND HE'S SO SEXY TOO!
HOMO DATA SCIENTISTICUS

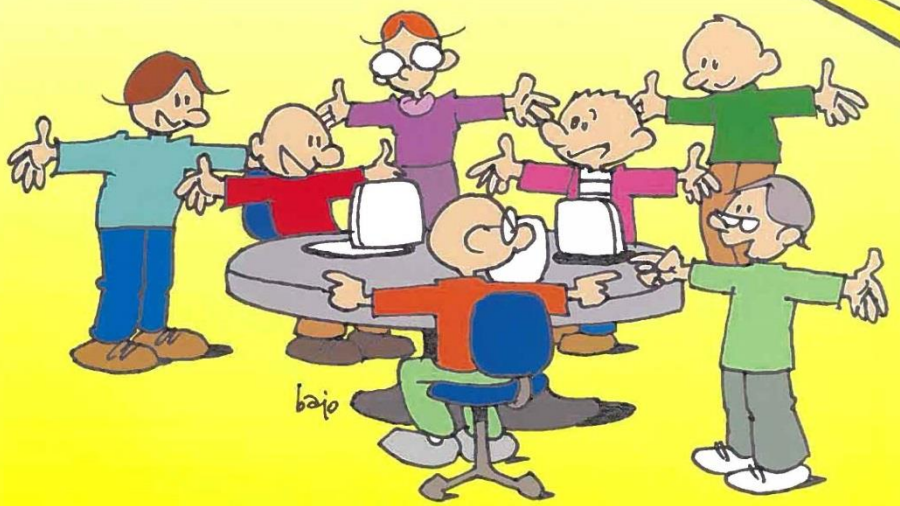


DSC/e 2014

Need:

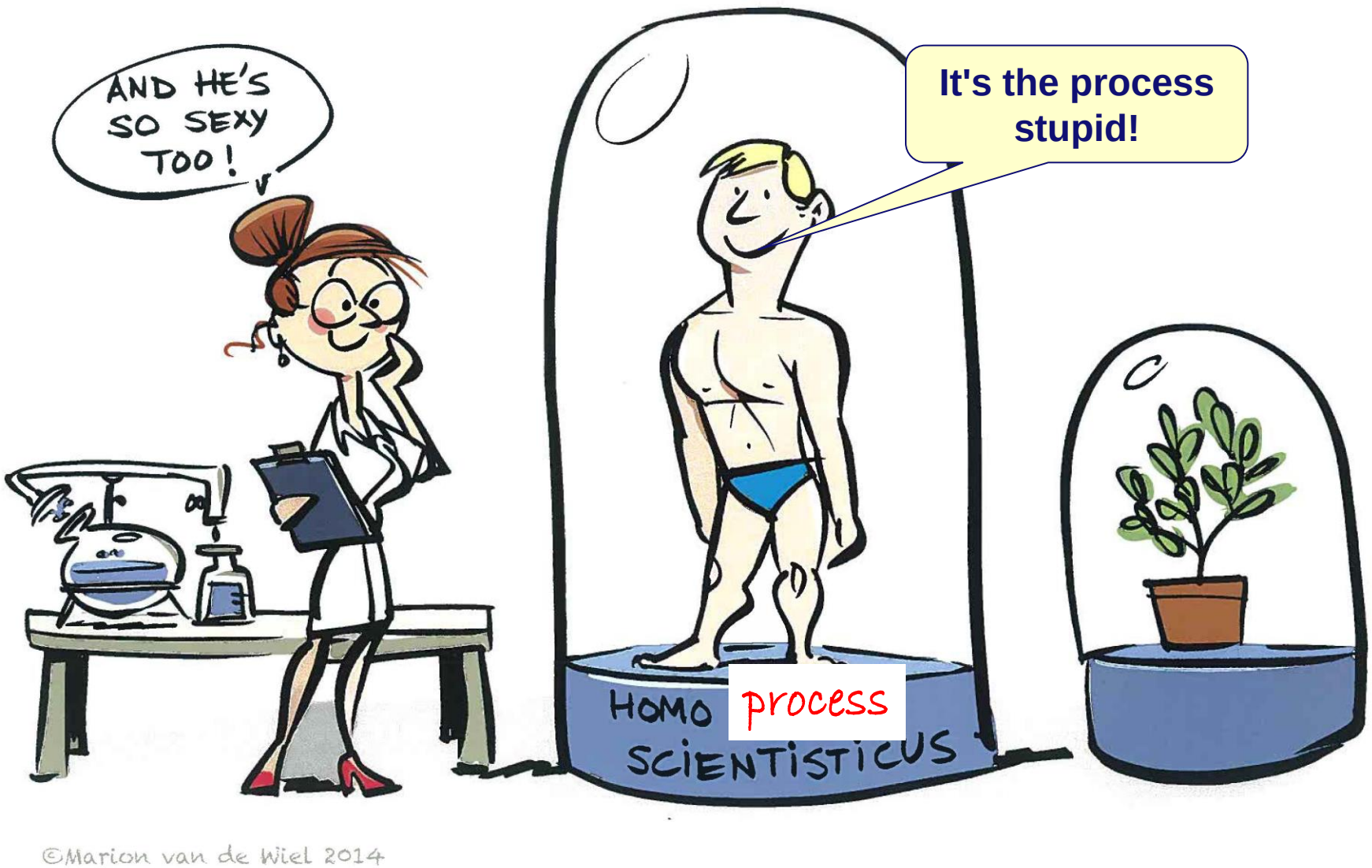
Data scientist

(T-shaped) teams

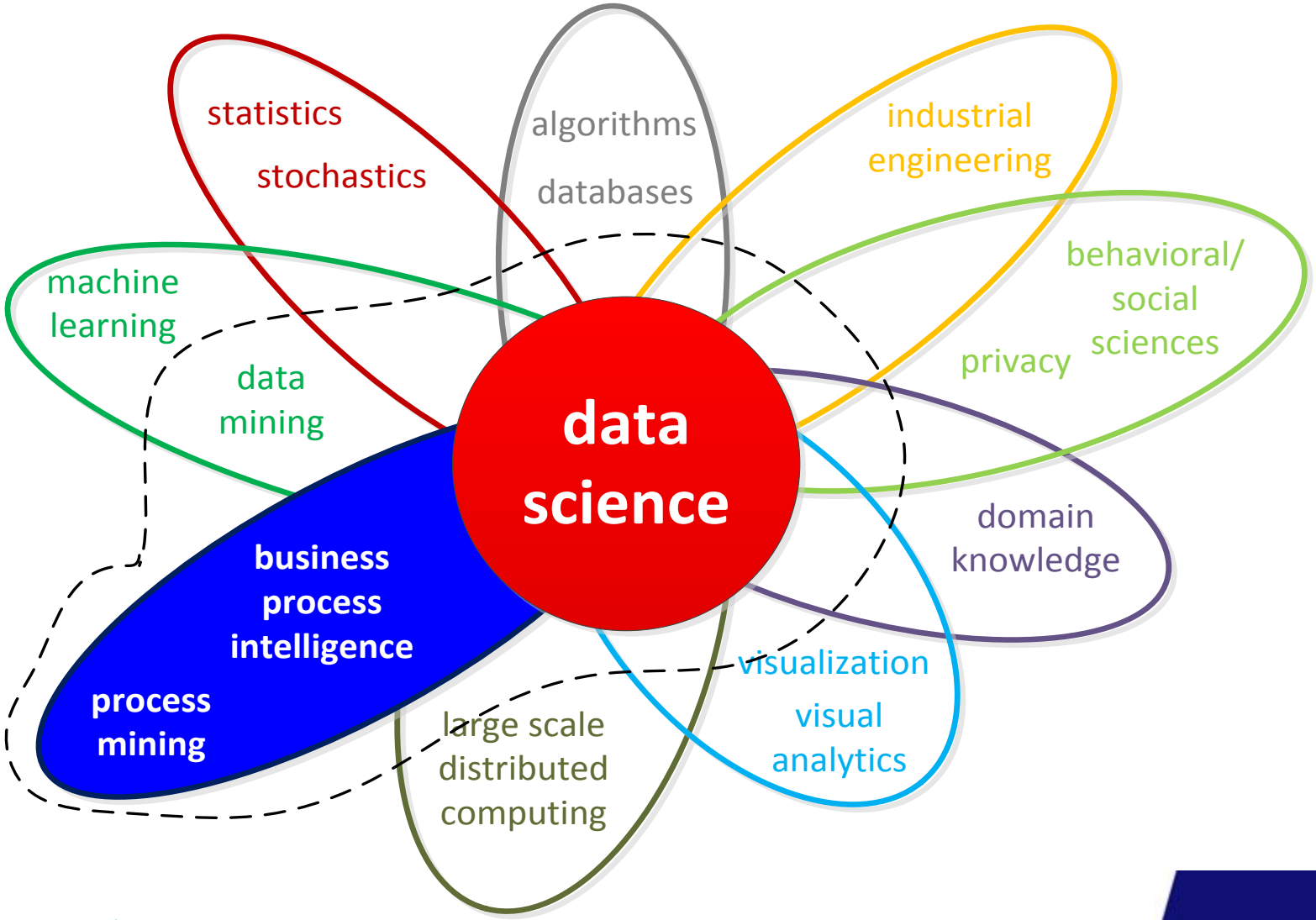


DSC/e 2014

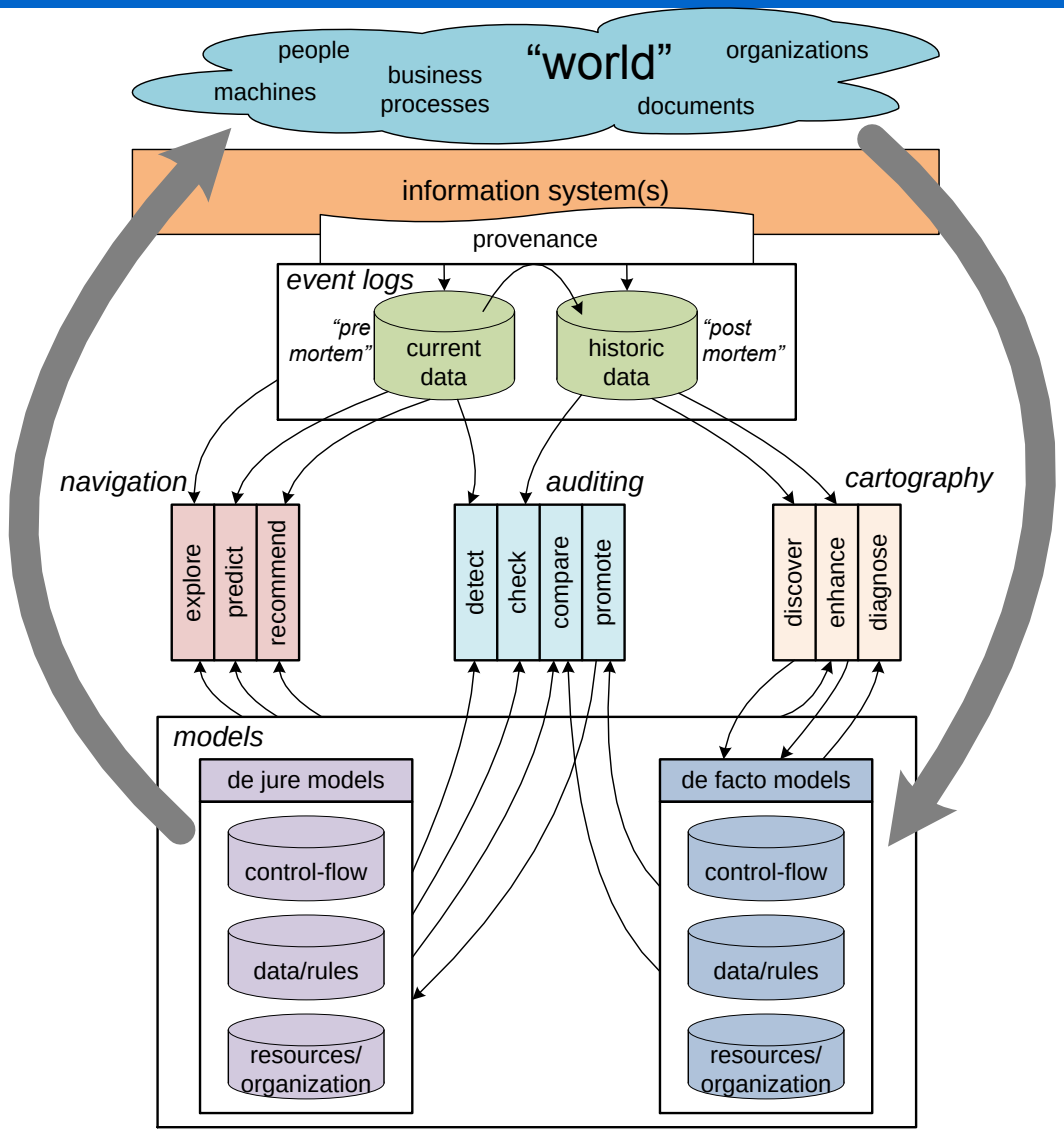
not just data …



Focus of process scientist

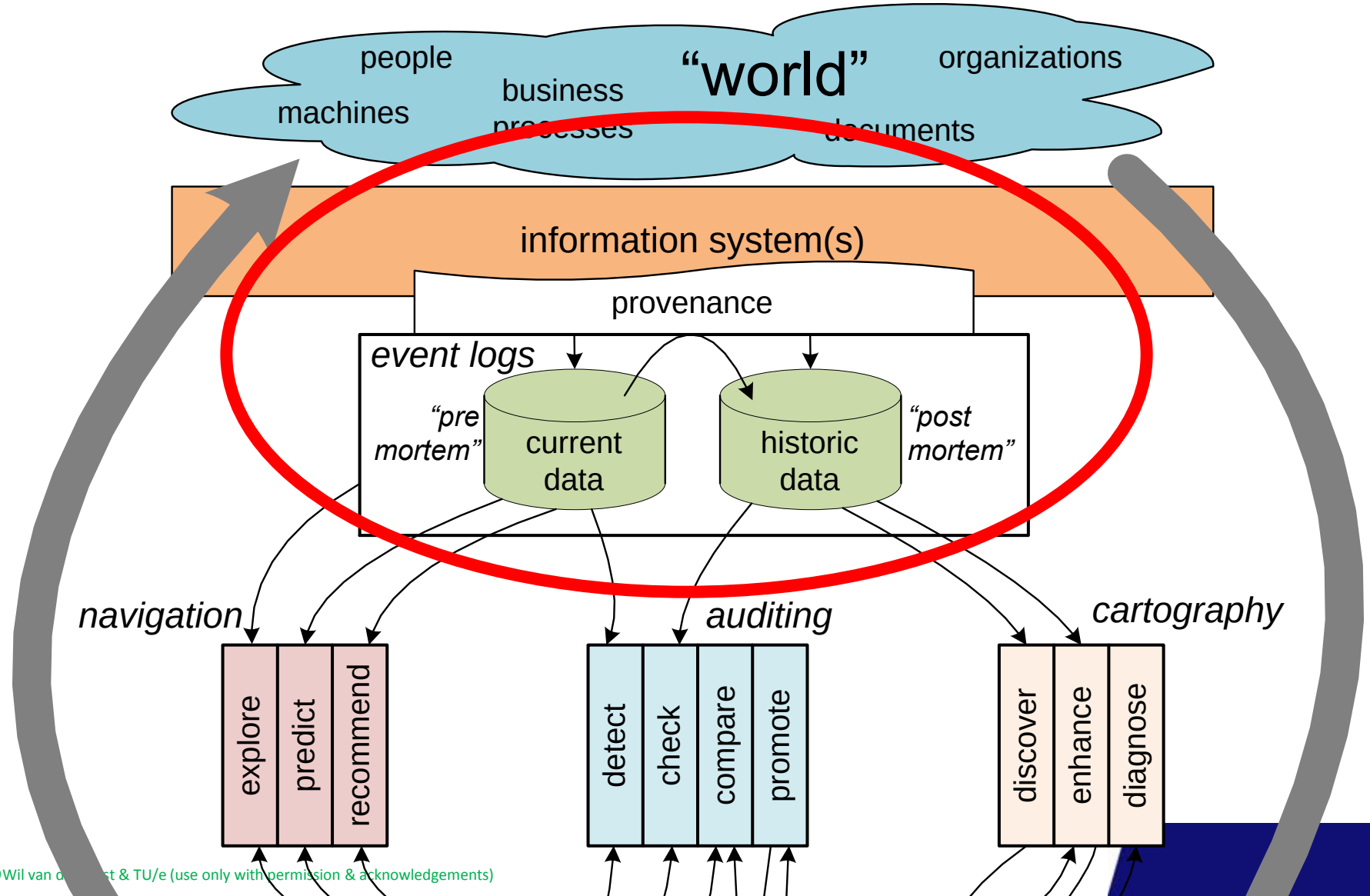


Refined process mining framework

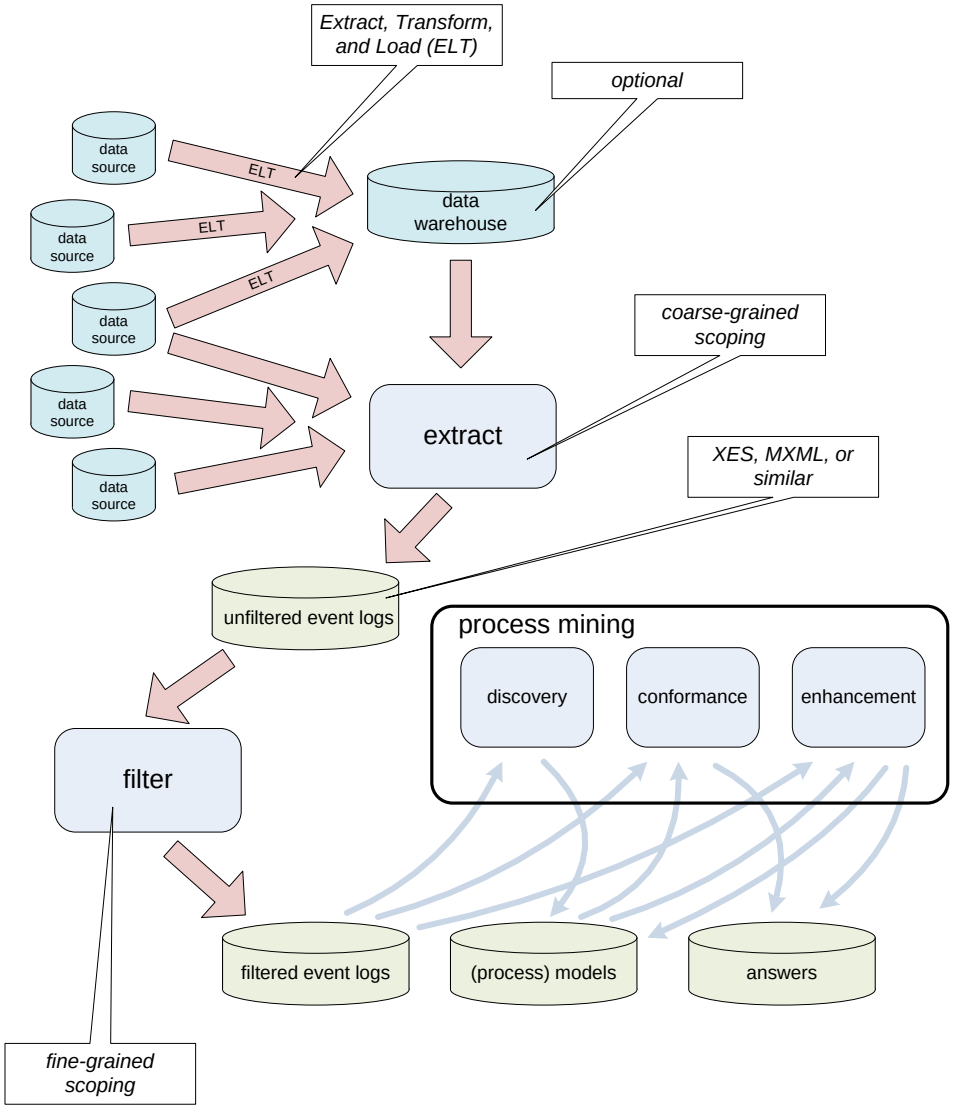


Much more than process discovery and conformance checking!

Today's focus: The Data!



Getting event logs




XES
data quality problems
data structure problems
guidelines for logging


XES
data quality problems
data structure problems
guidelines for logging

XES (eXtensible Event Stream)

- Adopted by the IEEE Task Force on Process Mining.
- The format is supported by tools such as ProM and Disco (used in this course).
- Predecessors: MXML and SA-MXML.
- Conversion from other formats (CSV) is easy if the right data are available.
- XML syntax and OpenXES library available.
- See www.xes-standard.org.

Event log

- We assume the existence of an event log where each event refers to a case, an activity, and a point in time.
- An event log can be seen as a collection of cases.
- A case can be seen as a trace/sequence of events.

Event data may come from …

- a database system (e.g., patient data in a hospital),
- a comma-separated values (CSV) file or spreadsheet,
- a transaction log (e.g., a trading system),
- a business suite/ERP system (SAP, Oracle, etc.),
- a message log (e.g., from IBM middleware),
- an open API providing data from websites or social media, …

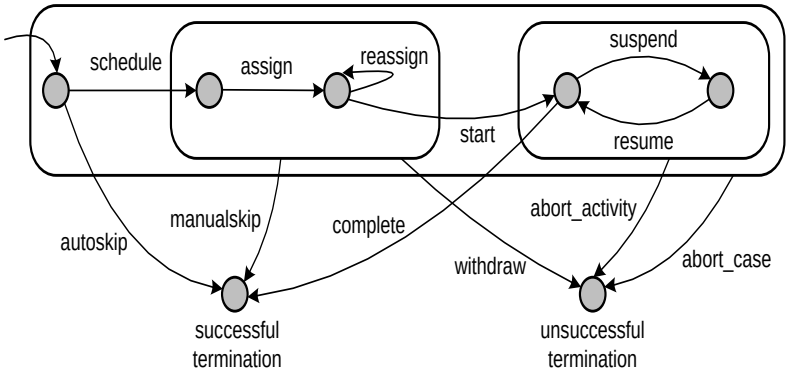
An example log

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**

Extensions

- **Transactional information on activity instances:**
  An event can represent a start, complete, suspend, resume, abort, etc.



- **Case versus event attributes:**
  - **case attributes do not change, e.g., the birth date or gender of a patient,**
  - **event attributes are related to a particular step in the process.**


XES
data quality problems
data structure problems
guidelines for logging

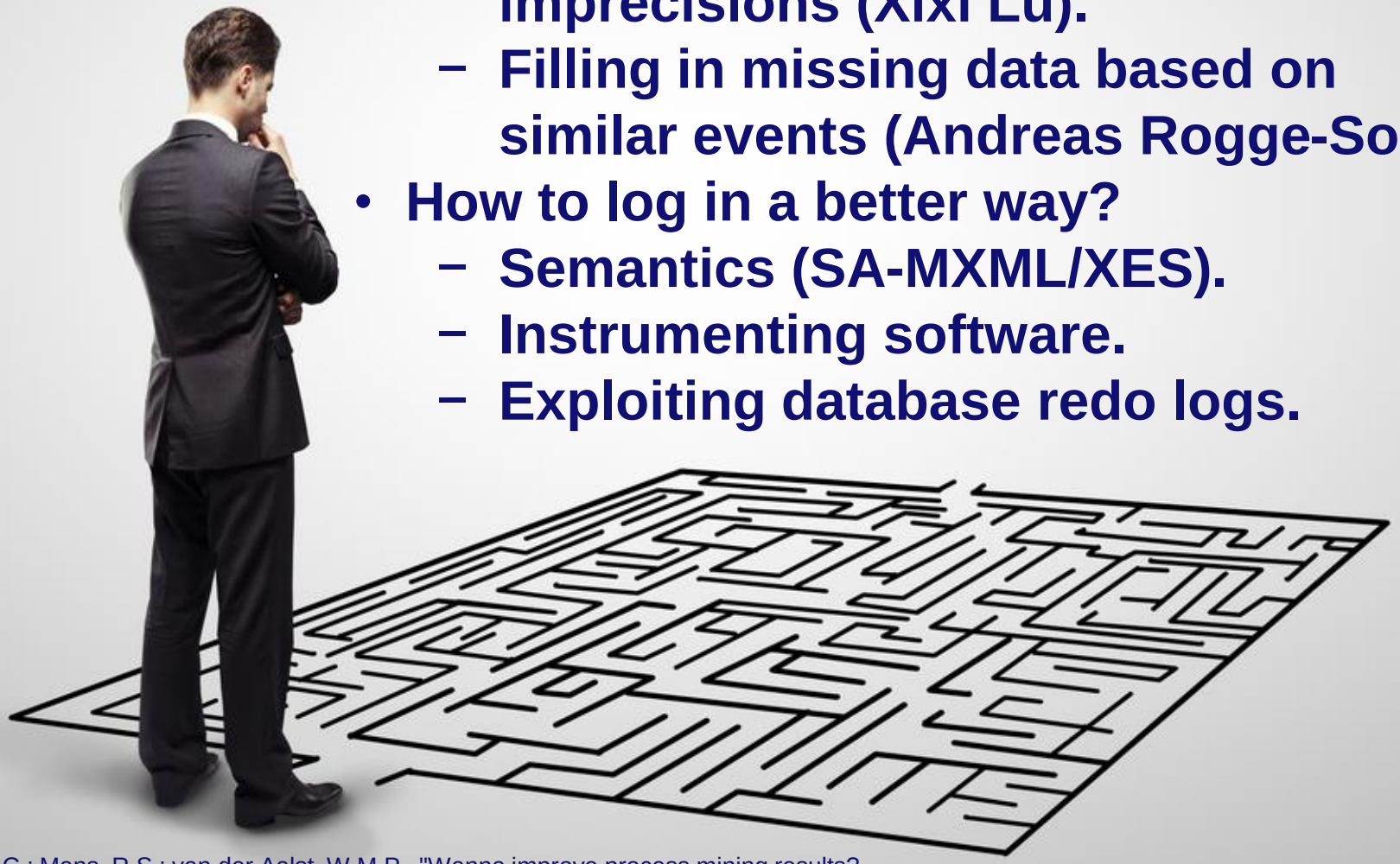
Data quality matrix

	case	event	belongs to	c attribute	position	activity name	timestamp	resource	e attribute
missing data	In reality a case has been executed but it has not been recorded in the log	Events are missing within the trace although they occurred in reality.	Association between events and cases is lost (correlation problem)	Case attribute was not recorded.	Ordering of events in the trace is lost.	Activity names of events are missing.	Timestamps of events are missing.	Resources that executed an activity have not been recorded.	Event attribute was not recorded.
incorrect data	Some cases in the log belong to a different process.	Events that were not actually executed for some cases are logged	Association between events and cases are logged incorrectly.	Values corresponding to case attributes are logged incorrectly.	Order is mixed up.	Wrong activity names are recorded.	Incorrect timestamps.	Incorrect resource assigned to event.	Attributes of events are recorded incorrectly.
imprecise data			Difficult to correlate events to specific cases (too coarse).	Provided value is too coarse, e.g., city but no address.	For example concurrent events may have become been totally ordered.	Activity names are too coarse.	Days rather than minutes or seconds. Hence, precise order cannot be derived.	Just role or department is recorded.	Provided value is too coarse.
irrelevant data	Irrelevant cases are included and cannot be removed easily.	Events may be irrelevant and difficult to remove							

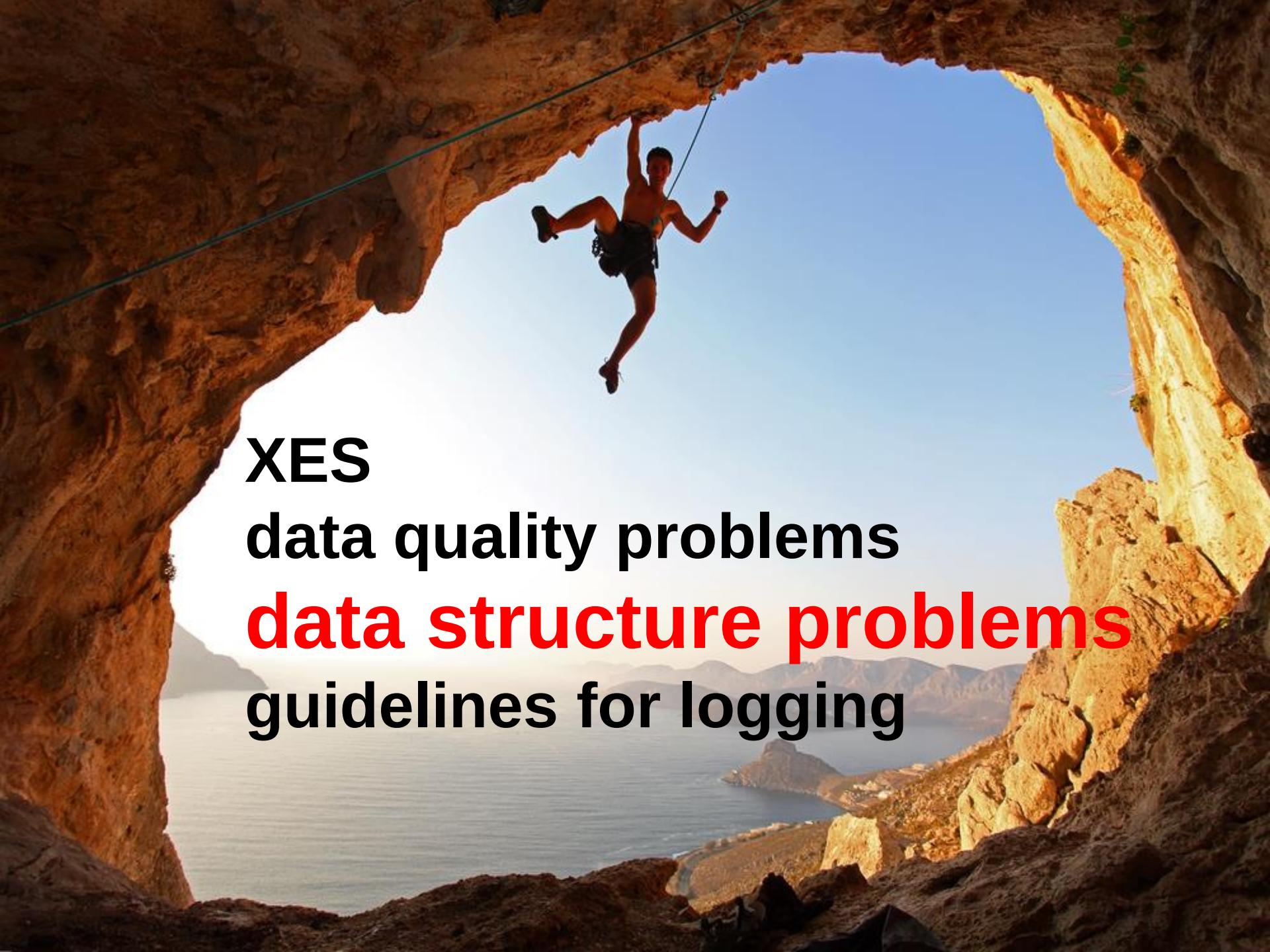
Bose, R.P.J.C.; Mans, R.S.; van der Aalst, W.M.P., "Wanna improve process mining results?," *Computational Intelligence and Data Mining (CIDM 2013)* , doi: 10.1109/CIDM.2013.6597227

Research directions:

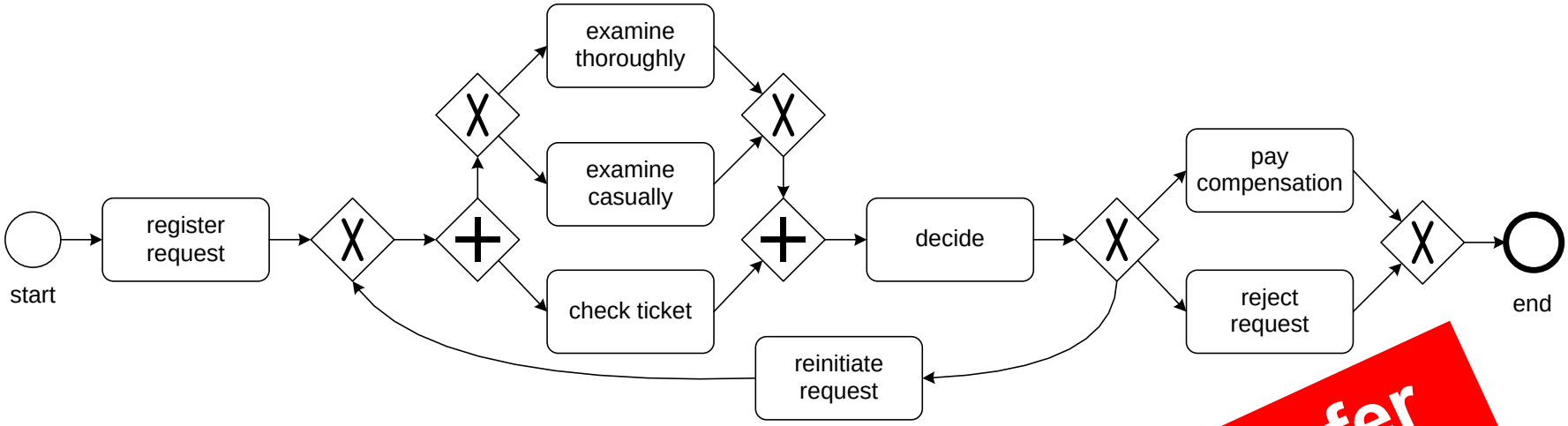
- How to handle imperfections in data?
  - Using partially ordered logs to handle imprecisions (Xixi Lu).
  - Filling in missing data based on similar events (Andreas Rogge-Solti).
- How to log in a better way?
  - Semantics (SA-MXML/XES).
  - Instrumenting software.
  - Exploiting database redo logs.



Bose, R.P.J.C.; Mans, R.S.; van der Aalst, W.M.P., "Wanna improve process mining results?, " *Computational Intelligence and Data Mining (CIDM 2013)* , doi: 10.1109/CIDM.2013.6597227

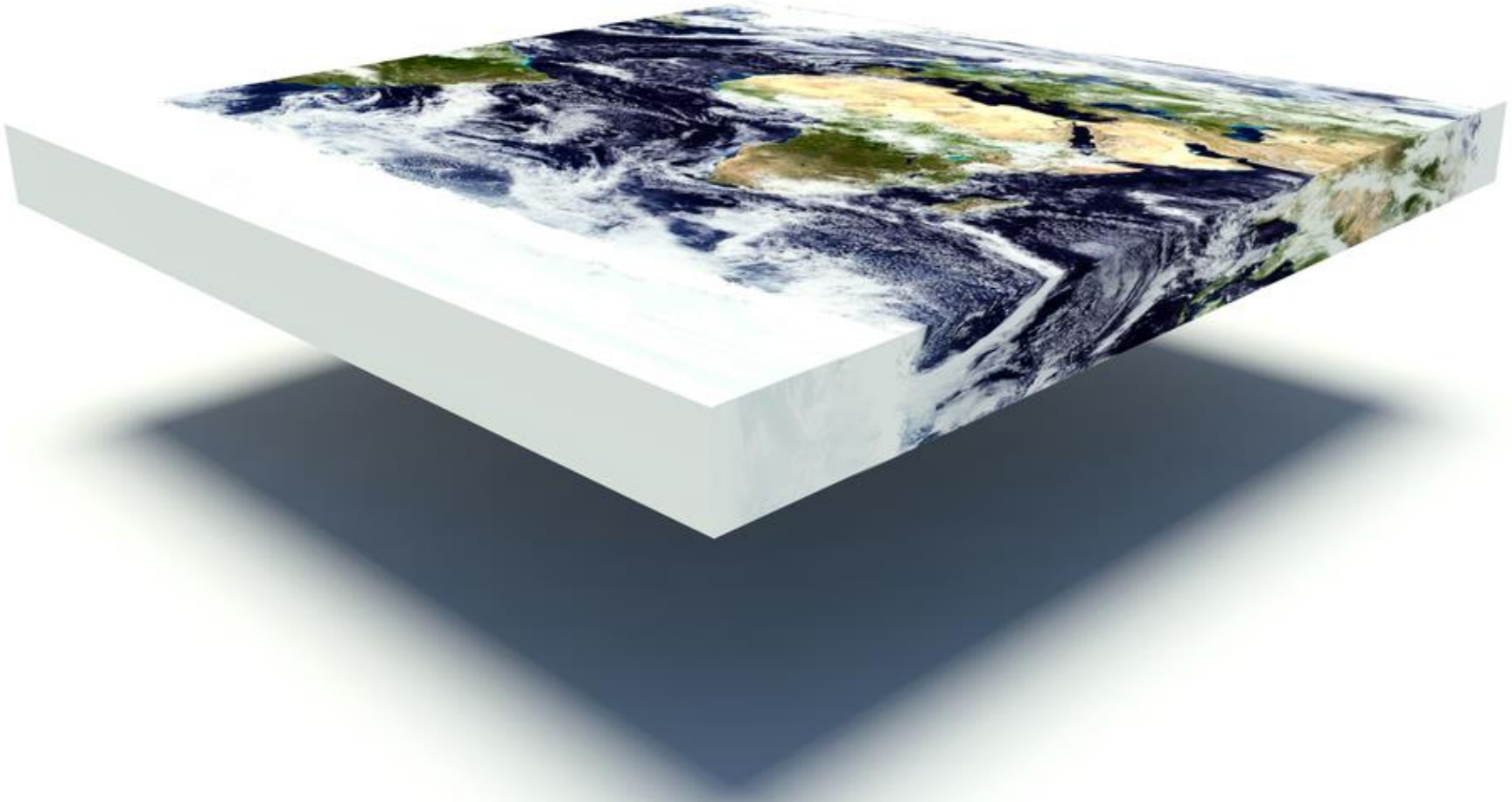

XES
data quality problems
data structure problems
guidelines for logging

Mainstream process models describe the lifecycle of instances in isolation !!!

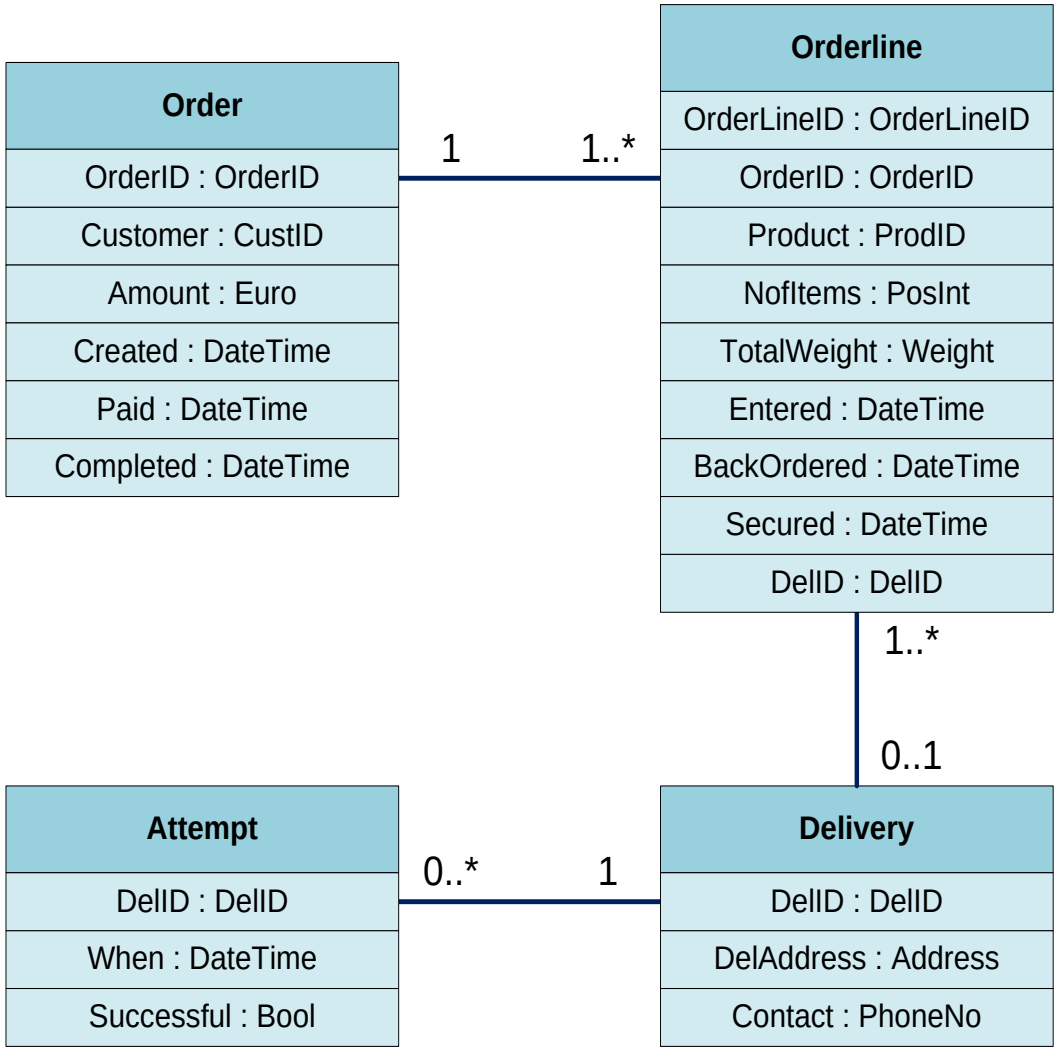


- **BPMN (Business Process Model and Notation) diagrams**
- **UML activity diagrams**
- **Event-driven process chains**
- **Petri net variants**
- **Causal nets (C-nets)**
- **…**

Events should refer to a process instance and activity!!!

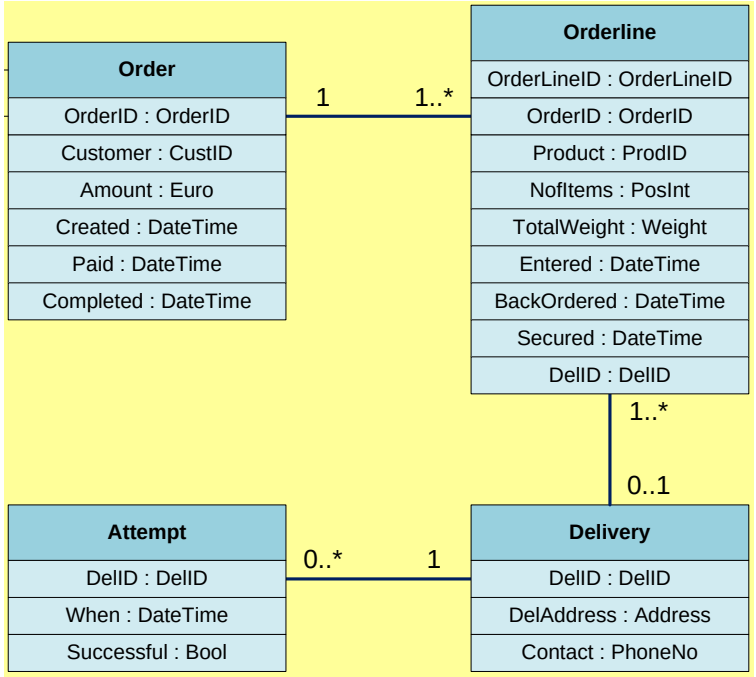


Example from the process mining book



What is the process instance?

Tables



Order

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

Delivery

DelIID	DelAddress	Contact
882345	5513VJ-22a	0497-2553660
882346	5513XG-45	040-2298761
...	...	...

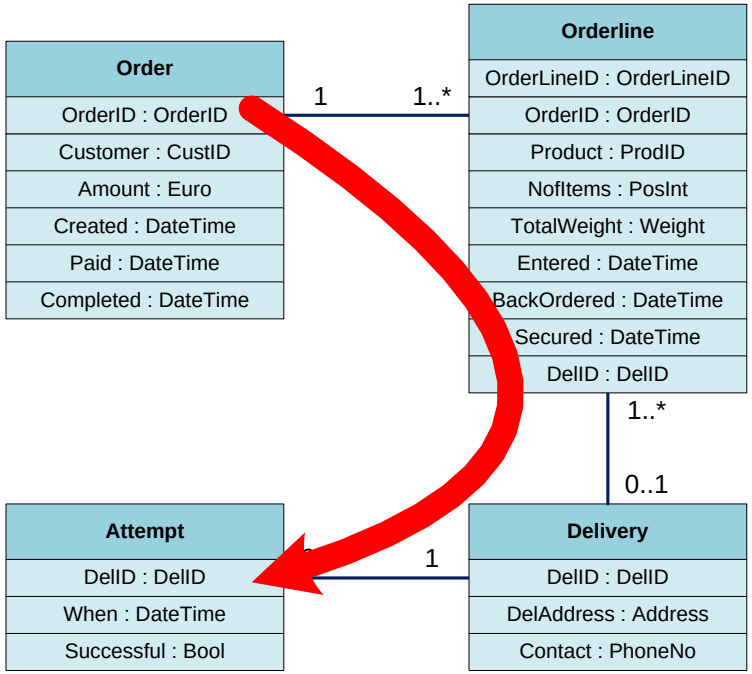
Attempt

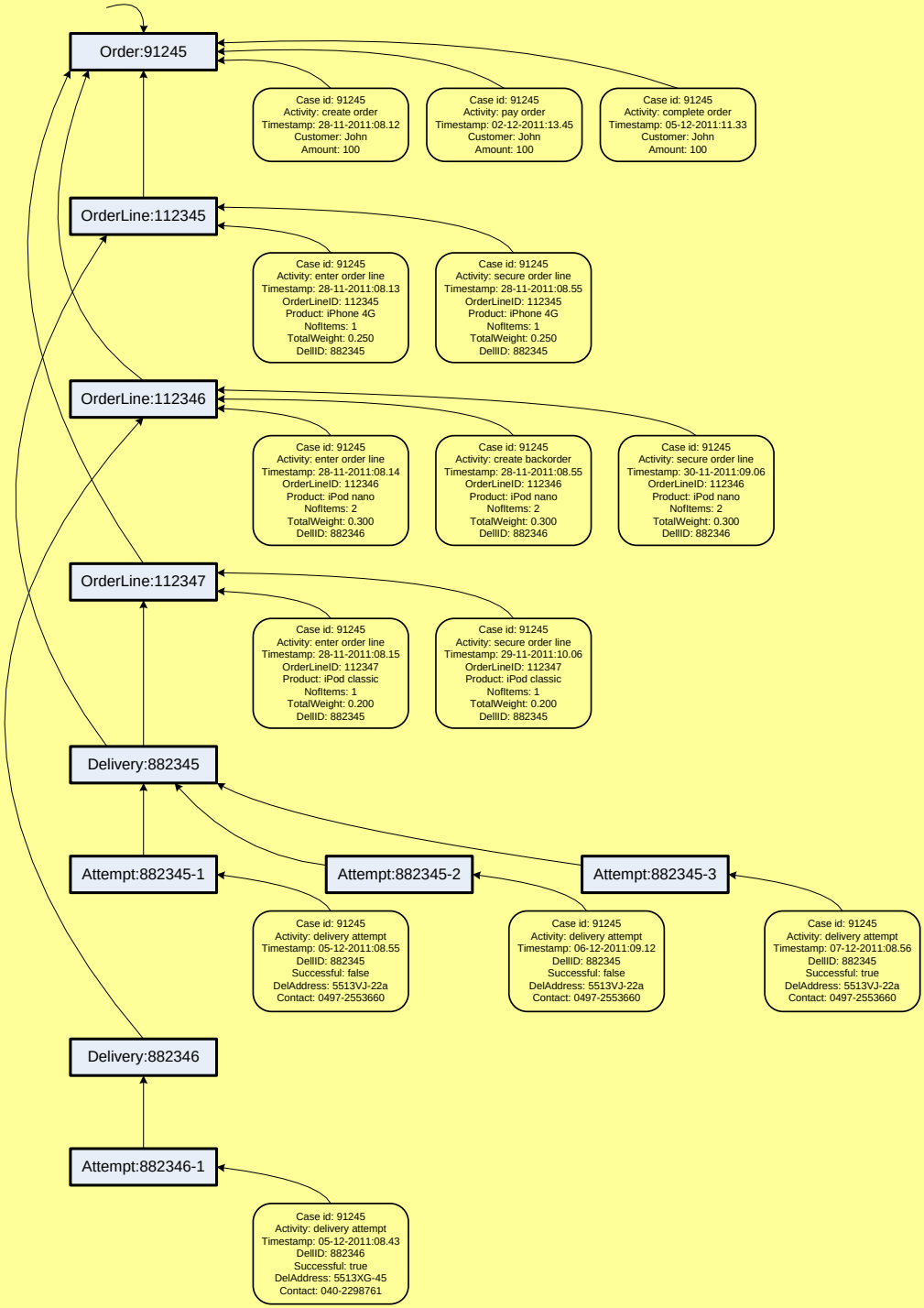
DelIID	When	Successful
882345	05-12-2011:08.55	false
882345	06-12-2011:09.12	false
882345	07-12-2011:08.56	true
882346	05-12-2011:08.43	true
...	...	...

Orderline

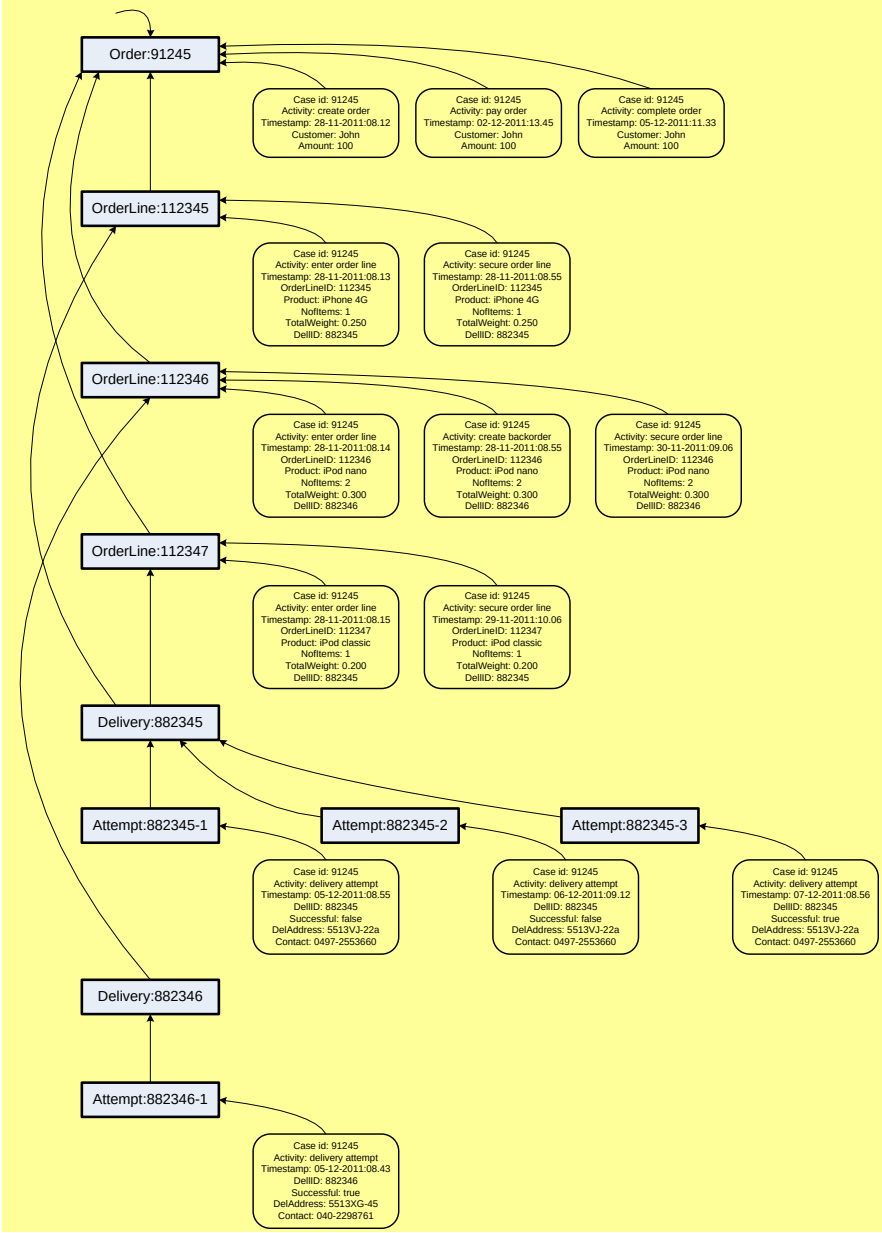
OrderLineID	OrderID	Product	NofItems	TotalWeight	Entered	BackOrdered	Secured	DelIID
112345	91245	iPhone 4G	1	0.250	28-11-2011:08.13	null	28-11-2011:08.55	882345
112346	91245	iPod nano	2	0.300	28-11-2011:08.14	28-11-2011:08.55	30-11-2011:09.06	882346
112347	91245	iPod classic	1	0.200	28-11-2011:08.15	null	29-11-2011:10.06	882345
112448	91561	iPhone 4G	1	0.250	28-11-2011:12.23	null	28-11-2011:12.59	882345
112449	91561	iPod classic	1	0.200	28-11-2011:12.24	28-11-2011:16.22	null	null
112452	91812	iPhone 4G	5	1.250	29-11-2011:09.46	null	29-11-2011:10.58	882346
...	...	...	...	...	...	...	...	...

Order instance

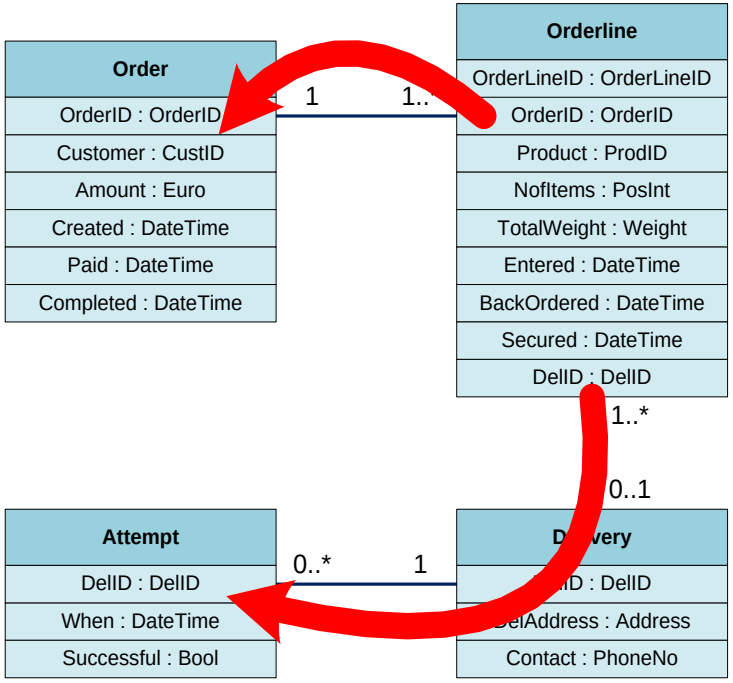


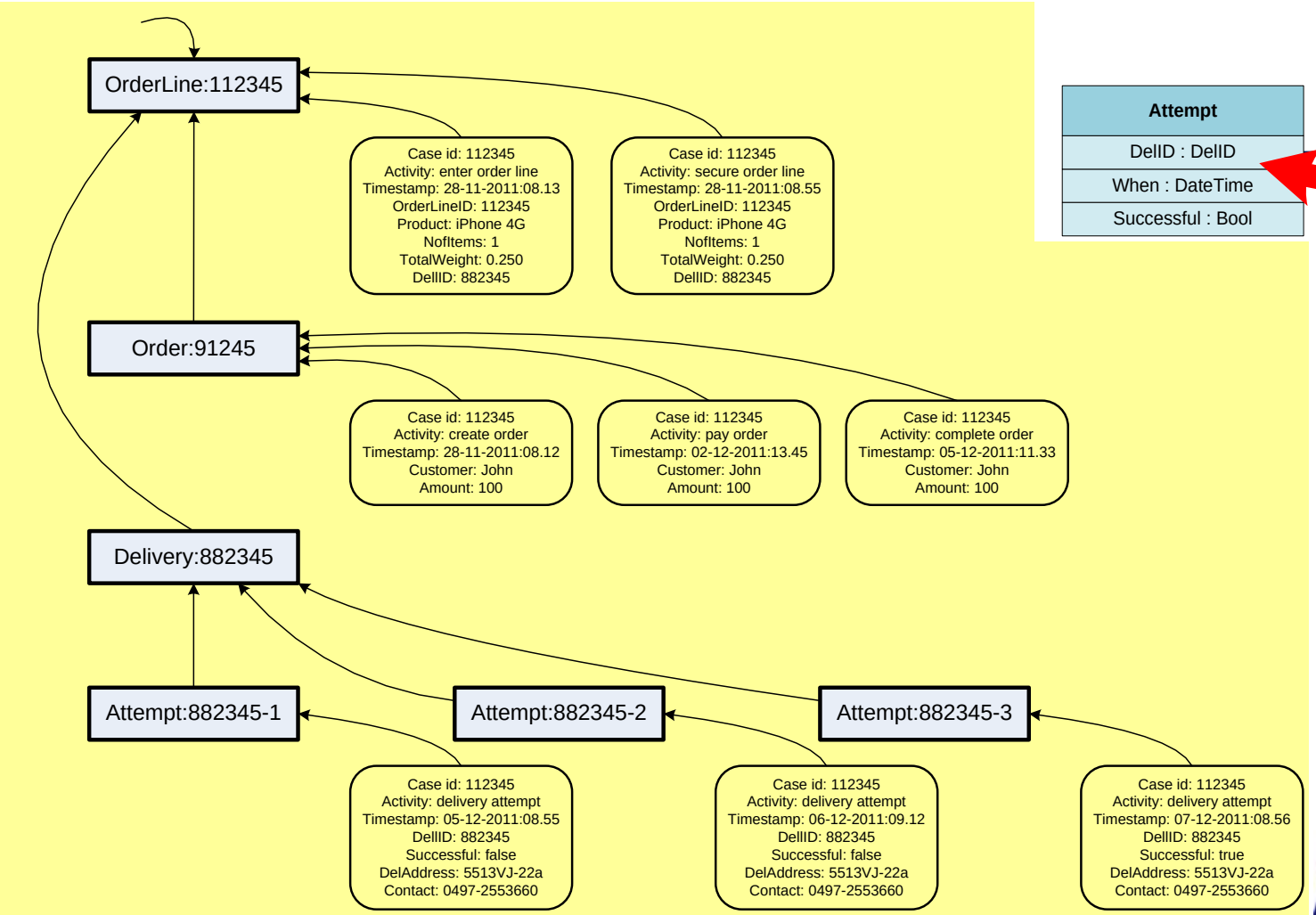


case id	activity	timestamp	other attributes
91245	create order	28-11-2011:08.12	Customer: John, Amount: 100
91245	enter order line	28-11-2011:08.13	OrderLineID: 112345, Product: iPhone 4G, NofItems: 1, TotalWeight: 0.250, DellID: 882345
91245	enter order line	28-11-2011:08.14	OrderLineID: 112346, Product: iPod nano, NofItems: 2, TotalWeight: 0.300, DellID: 882346
91245	enter order line	28-11-2011:08.15	OrderLineID: 112347, Product: iPod classic, NofItems: 1, TotalWeight: 0.200, DellID: 882345
91245	secure order line	28-11-2011:08.55	OrderLineID: 112345, Product: iPhone 4G, NofItems: 1, TotalWeight: 0.250, DellID: 882345
91245	create backorder	28-11-2011:08.55	OrderLineID: 112346, Product: iPod nano, NofItems: 2, TotalWeight: 0.300, DellID: 882346
91245	secure order line	29-11-2011:10.06	OrderLineID: 112347, Product: iPod classic, NofItems: 1, TotalWeight: 0.200, DellID: 882345
91245	secure order line	30-11-2011:09.06	OrderLineID: 112346, Product: iPod nano, NofItems: 2, TotalWeight: 0.300, DellID: 882346
91245	pay order	02-12-2011:13.45	Customer: John, Amount: 100
91245	delivery attempt	05-12-2011:08.43	DellID: 882346, Successful: true, DelAddress: 5513XG-45, Contact: 040-2298761
91245	delivery attempt	05-12-2011:08.55	DellID: 882345, Successful: false, DelAddress: 5513VJ-22a, Contact: 0497-2553660
91245	complete order	05-12-2011:11.33	Customer: John, Amount: 100
91245	delivery attempt	06-12-2011:09.12	DellID: 882345, Successful: false, DelAddress: 5513VJ-22a, Contact: 0497-2553660
91245	delivery attempt	07-12-2011:08.56	DellID: 882345, Successful: true, DelAddress: 5513VJ-22a, Contact: 0497-2553660
91561	create order	28-11-2011:12.22	Customer: Mike, Amount: 530
91561	enter order line	28-11-2011:12.23	OrderLineID: 112448, Product: iPhone 4G, NofItems: 1, TotalWeight: 0.250, DellID: 882345
...	...	...	...
...	...	...	...



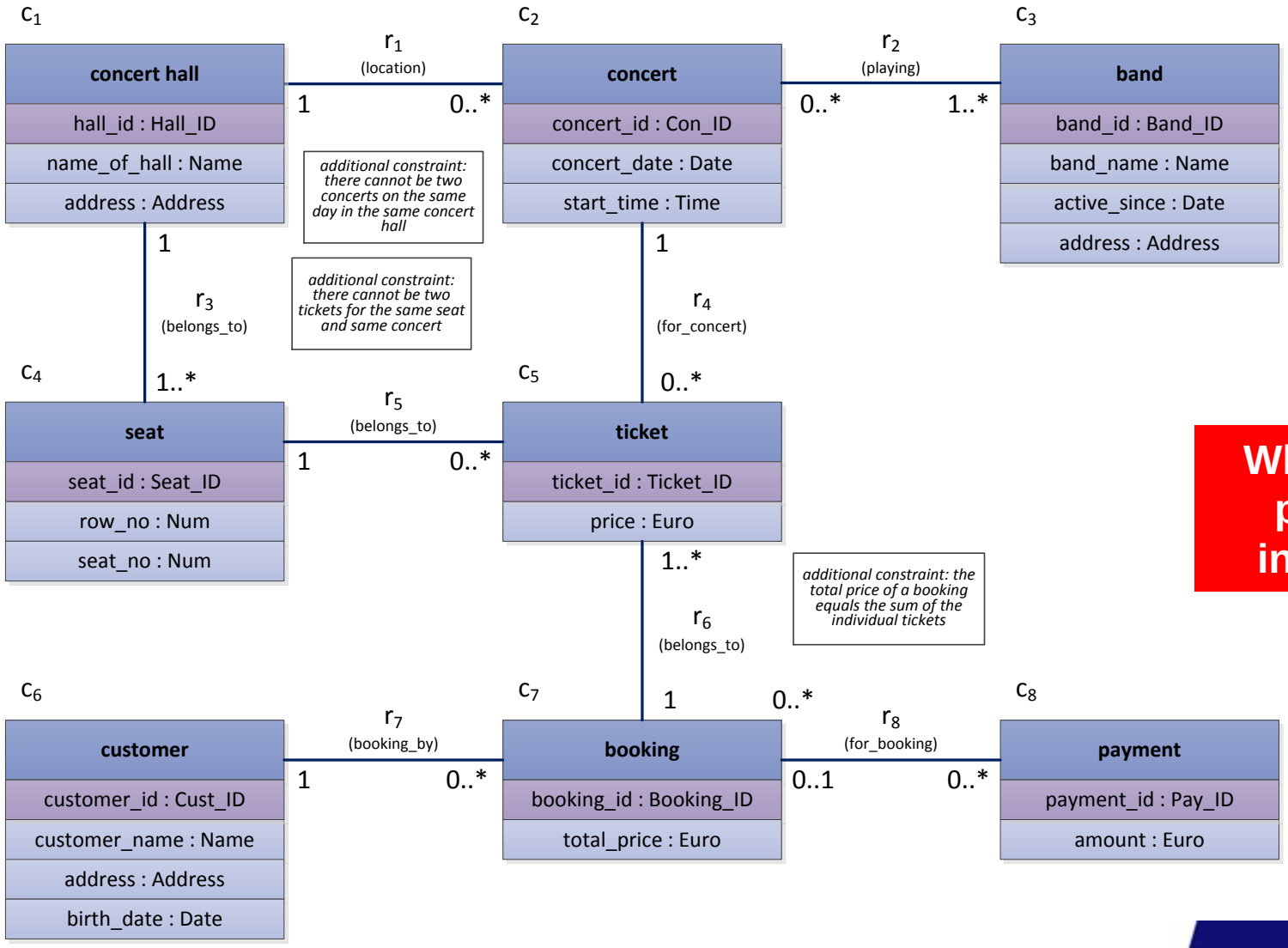
Orderline instance





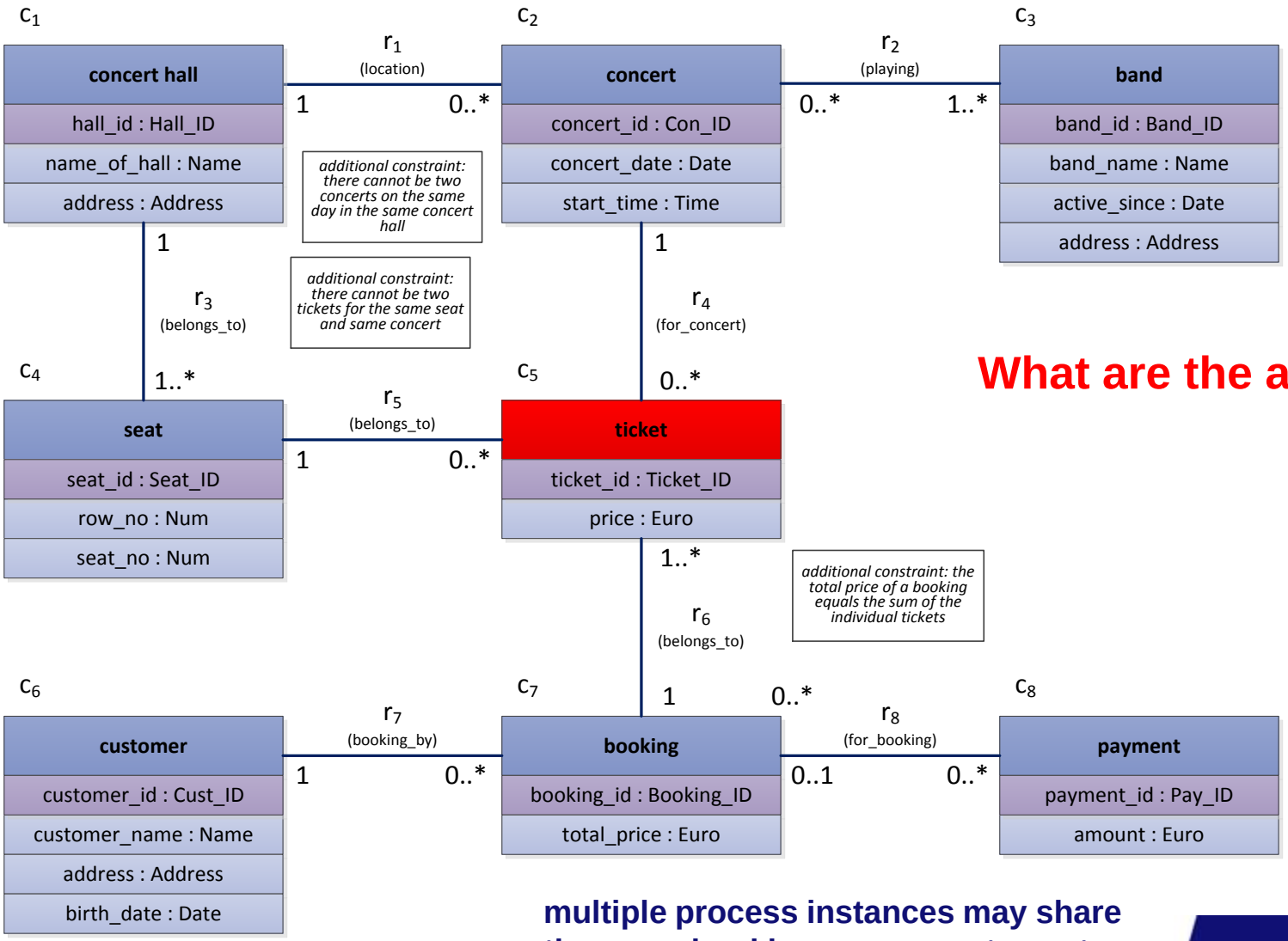
Another example: Booking concert seats

(assuming we have the redo logs)



What is the process instance?

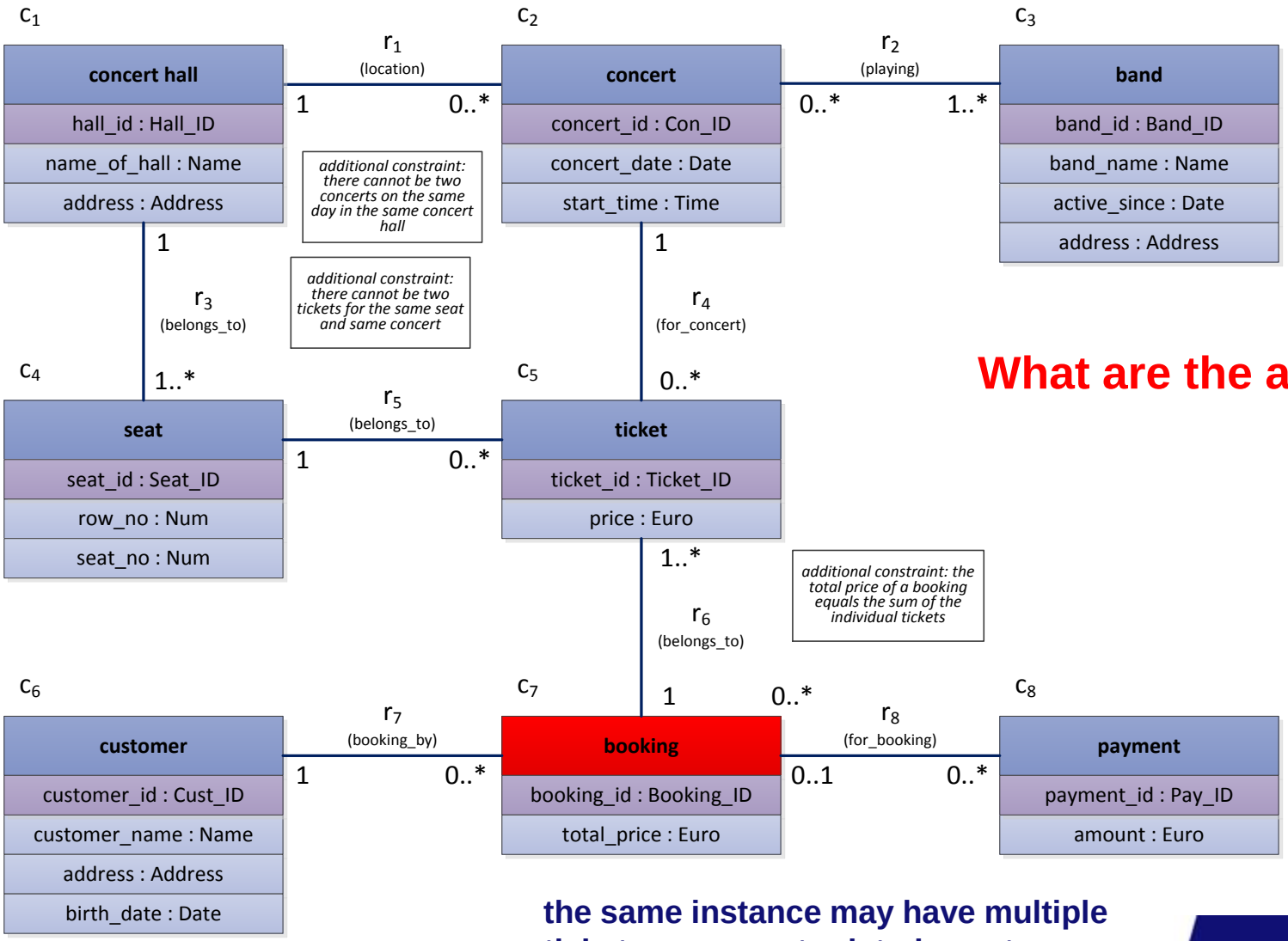
Lifecycle of ticket?



What are the activities?

multiple process instances may share the same booking or payment event

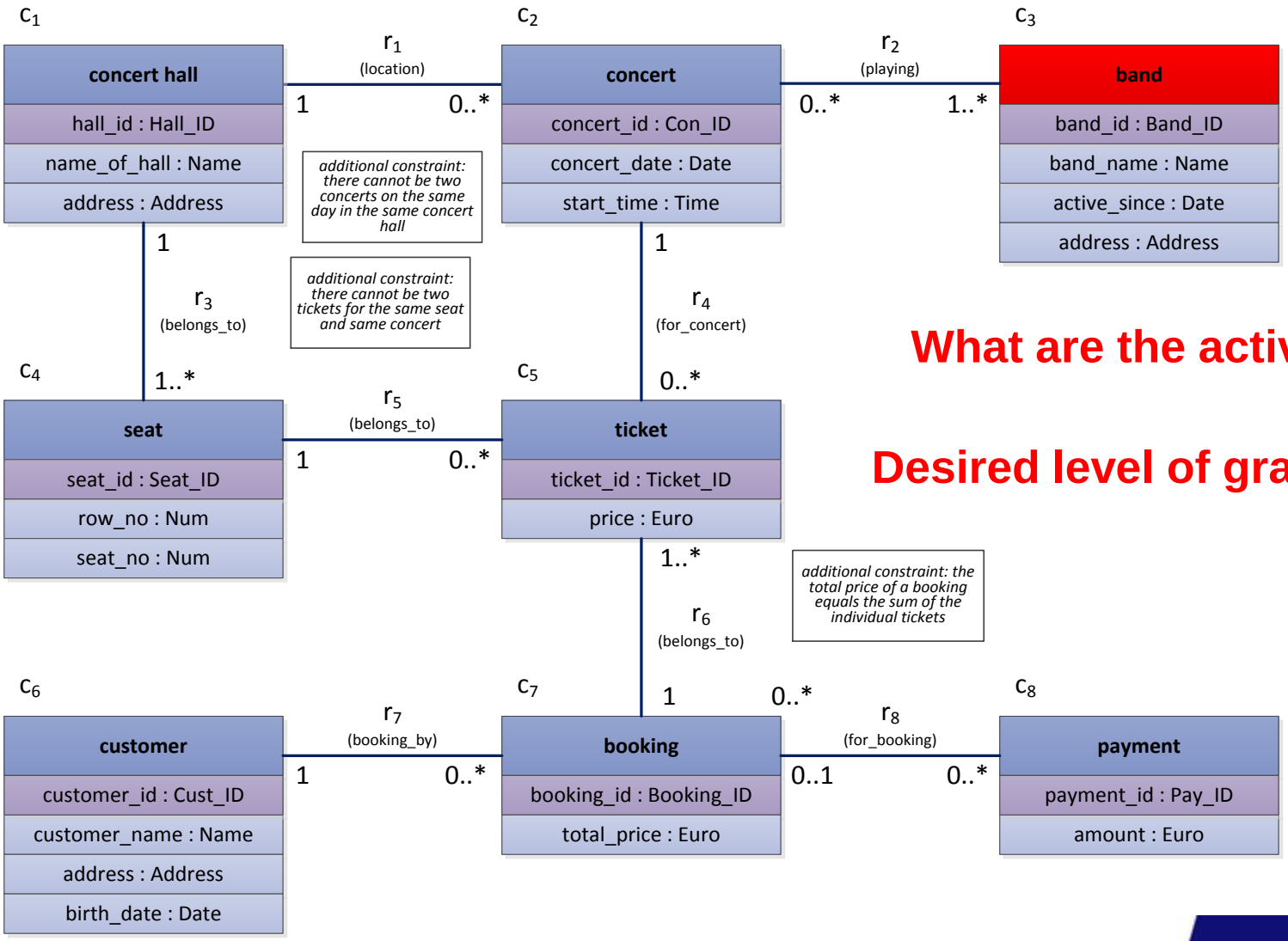
Lifecycle of booking?



What are the activities?

the same instance may have multiple ticket or payment related events

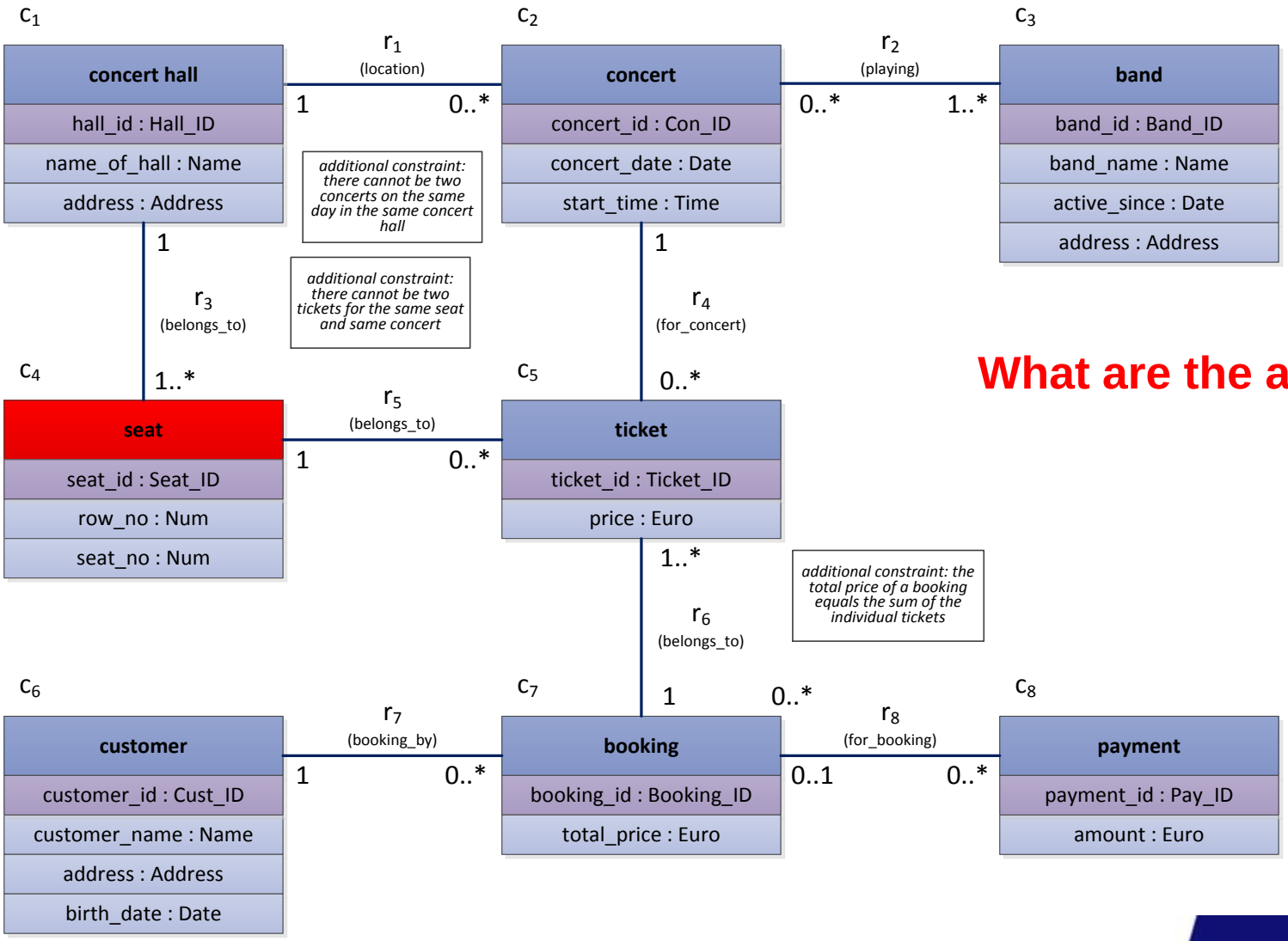
Lifecycle of band?



What are the activities?

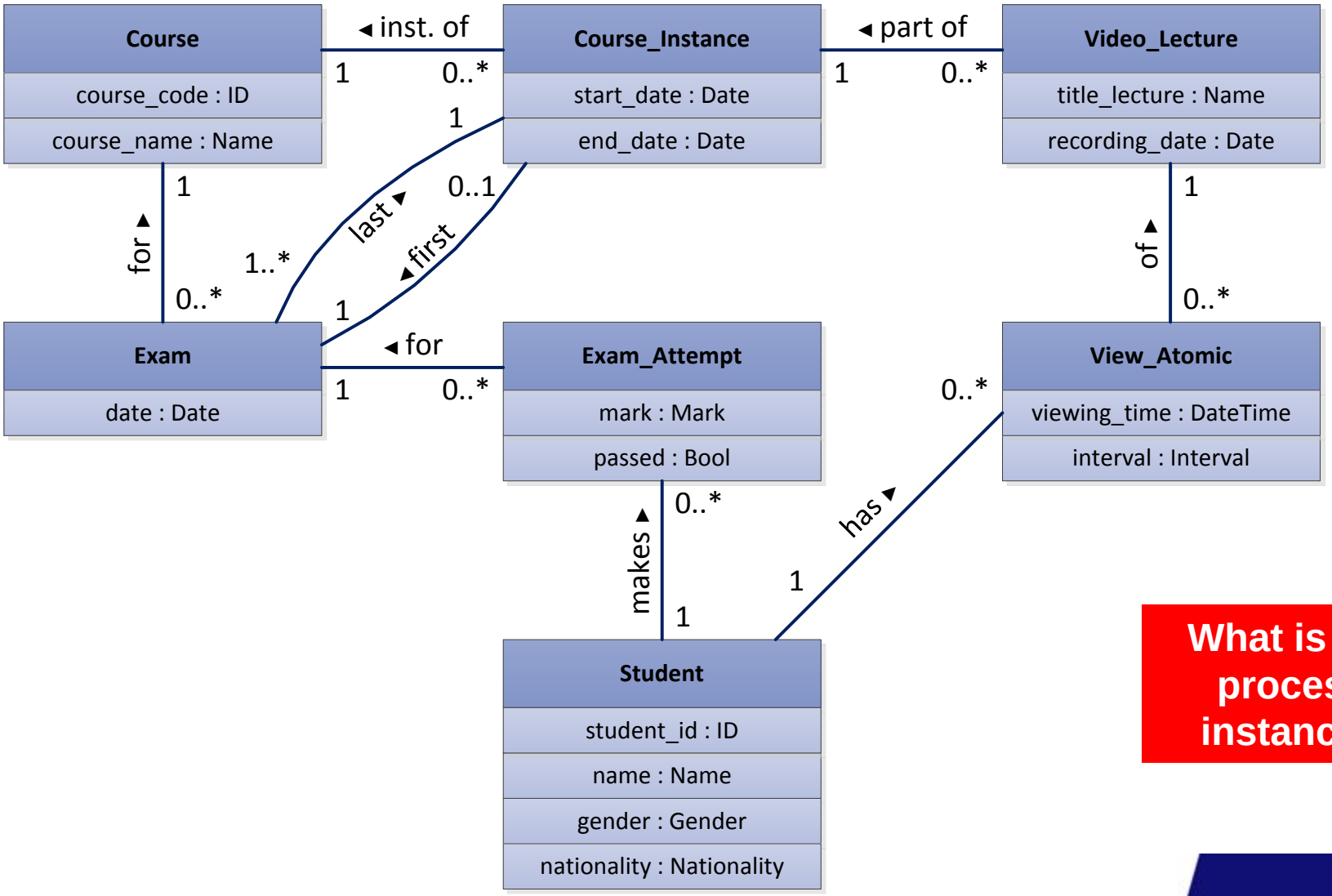
Desired level of granularity?

Lifecycle of seat?



What are the activities?

Another example

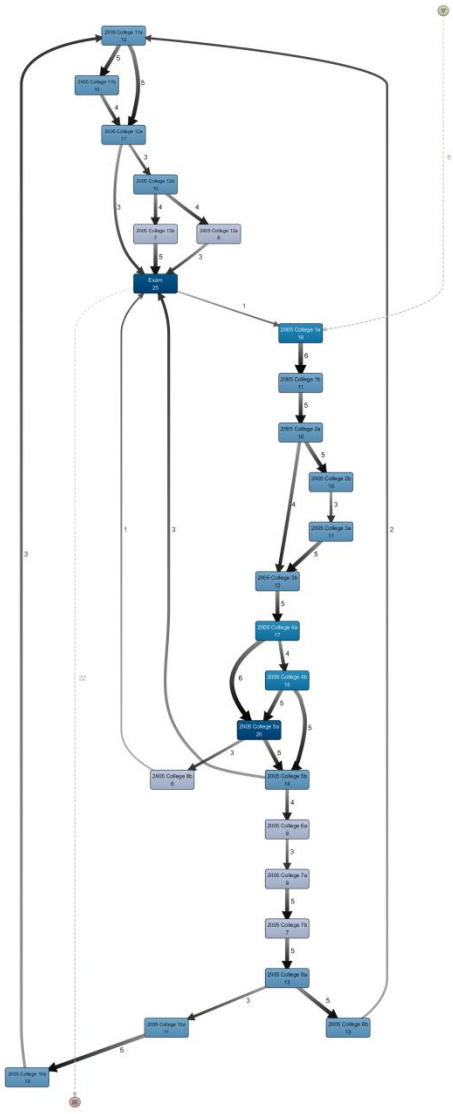


What is the process instance?

Case: Student taking the BIS course

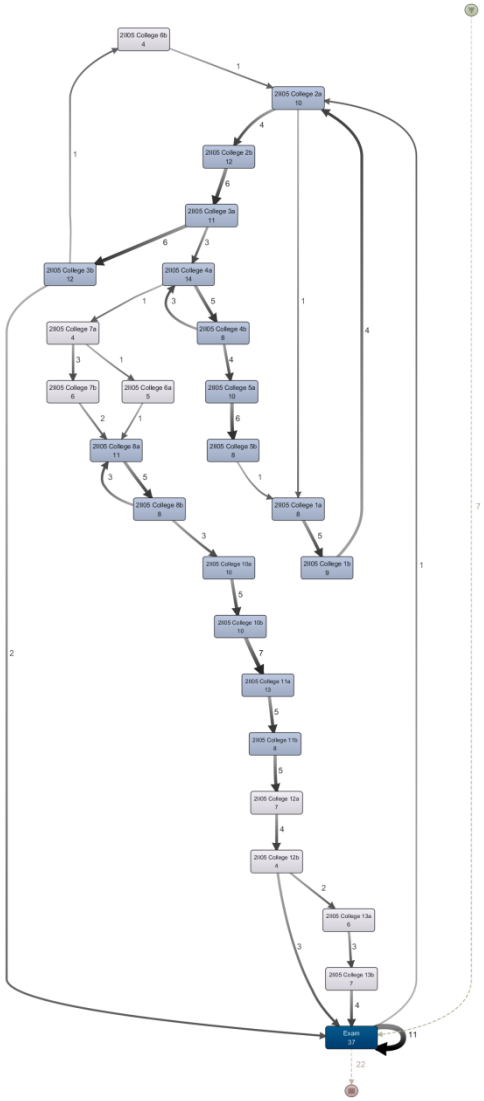
Example results: Lifecycle of BIS students (year: 2011)

PASSED



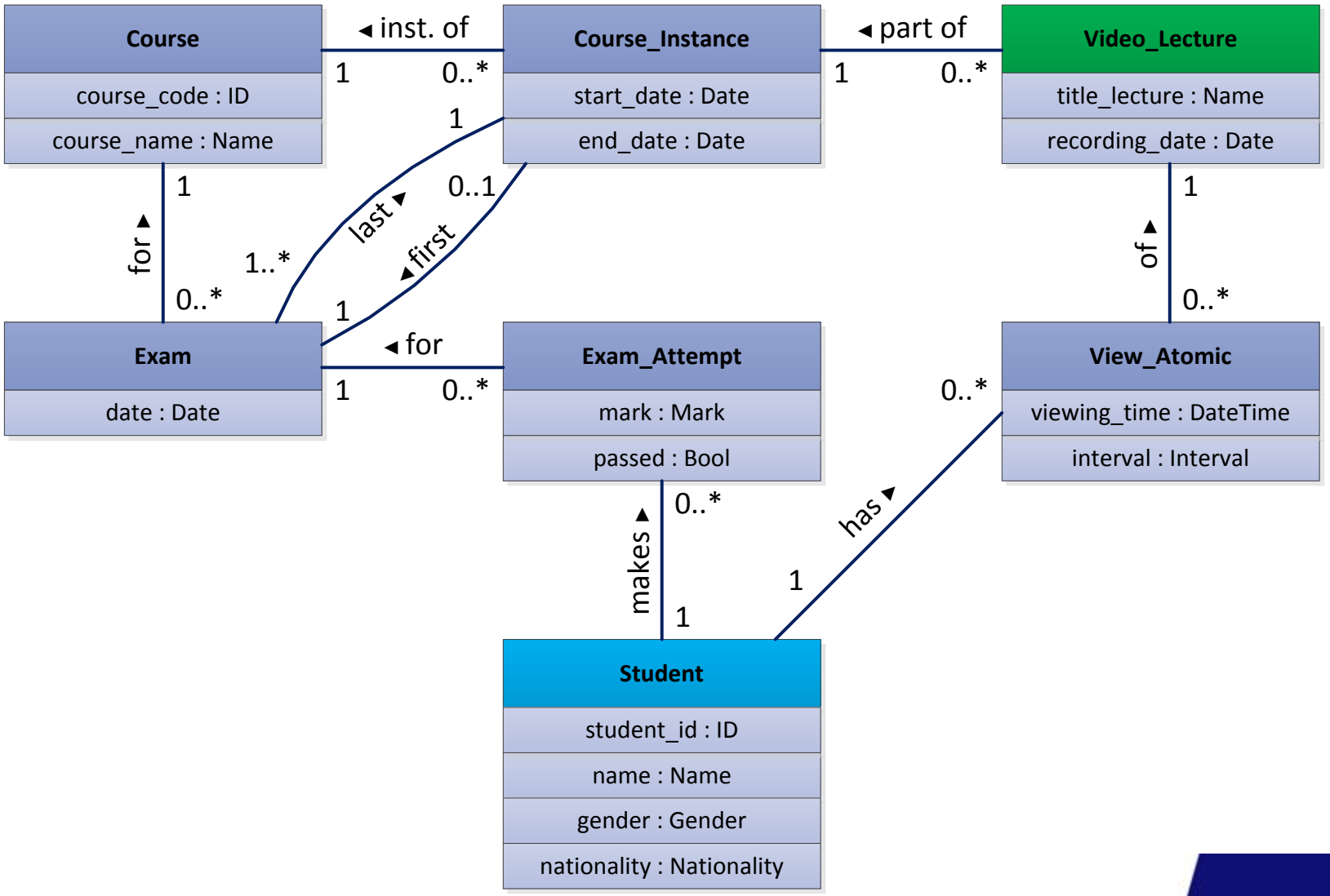
Fitness of event log wrt idealized model is 0.37

FAILED



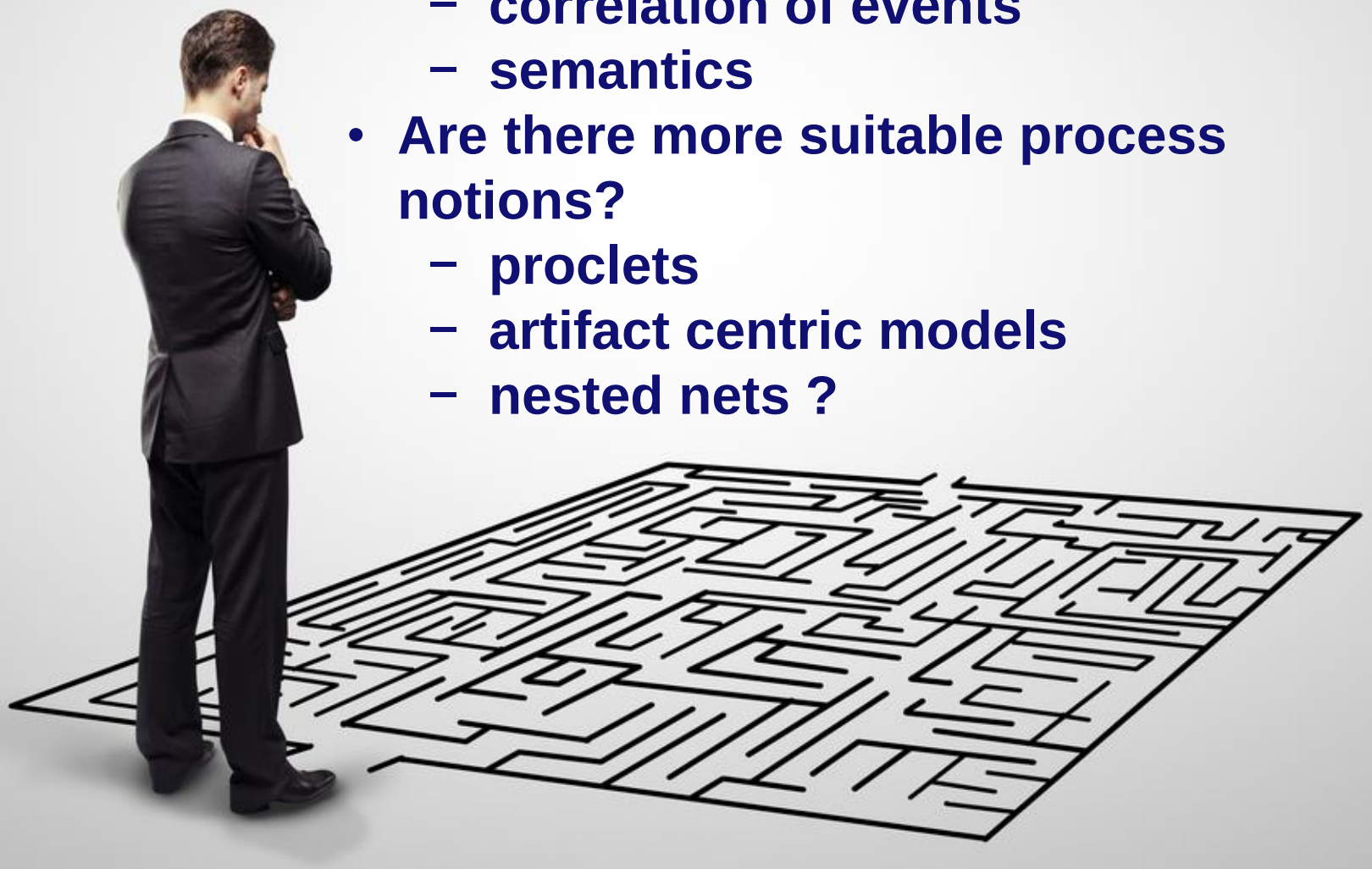
Fitness of event log wrt idealized model is 0.28

Other case notions: lifecycle of a lecture or the lifecycle of a student across courses



Research directions:

- How to preprocess data?
  - analyze table structure of database
  - correlation of events
  - semantics
- Are there more suitable process notions?
  - proclets
  - artifact centric models
  - nested nets ?

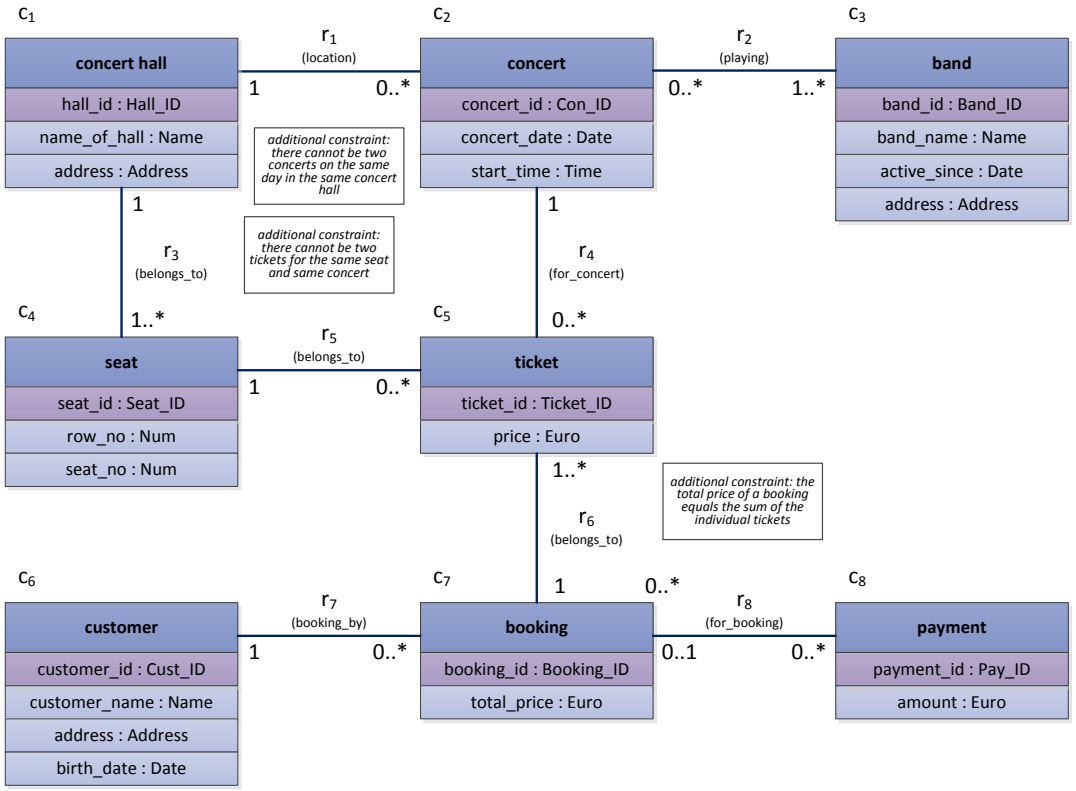


Some pointers

- W.M.P. van der Aalst, P. Barthelmess, C.A. Ellis, and J. Wainer. Proclets: A Framework for Lightweight Interacting Workflow Processes. *International Journal of Cooperative Information Systems*, 10(4):443-482, 2001.
- W.M.P. van der Aalst. Extracting Event Data from Databases to Unleash Process Mining. European BPM Roundtable, Liechtenstein, 2007.
- R.S. Mans, N.C. Russell, W.M.P. van der Aalst, A.J. Moleman, P.J.M. Bakker, and M. Jaspers. Proclets in Healthcare. *Journal of Biomedical Informatics*, 43(4):632-649, 2010.
- D. Fahland, M. De Leoni, B. van Dongen, and W.M.P. van der Aalst. Behavioral Conformance of Artifact-Centric Process Models. In A. Abramowicz, editor, *Business Information Systems (BIS 2011)*, volume 87 of *Lecture Notes in Business Information Processing*, pages 37-49. Springer-Verlag, Berlin, 2011.
- D. Fahland, M. De Leoni, B. van Dongen, and W.M.P. van der Aalst. Many-to-Many: Some Observations on Interactions in Artifact Choreographies. In D. Eichhorn, A. Koschmider, and H. Zhang, editors, *Proceedings of the 3rd Central-European Workshop on Services and their Composition (ZEUS 2011)*, CEUR Workshop Proceedings, pages 9-15. CEUR-WS.org, 2011.
- W.M.P. van der Aalst. Service Mining: Using Process Mining to Discover, Check, and Improve Service Behavior. *IEEE Transactions on Services Computing*, 6(4):525-535, 2013.
- H.M.W. Verbeek, J.C.A.M. Buijs, B.F. van Dongen, and W.M.P. van der Aalst. XES, XESame, and ProM 6. In P. Soffer and E. Proper, editors, *Information Systems Evolution*, volume 72 of *Lecture Notes in Business Information Processing*, pages 60-75. Springer-Verlag, Berlin, 2010.

- See also ACSI project!

Possible view on the World of Event Data (WoED)

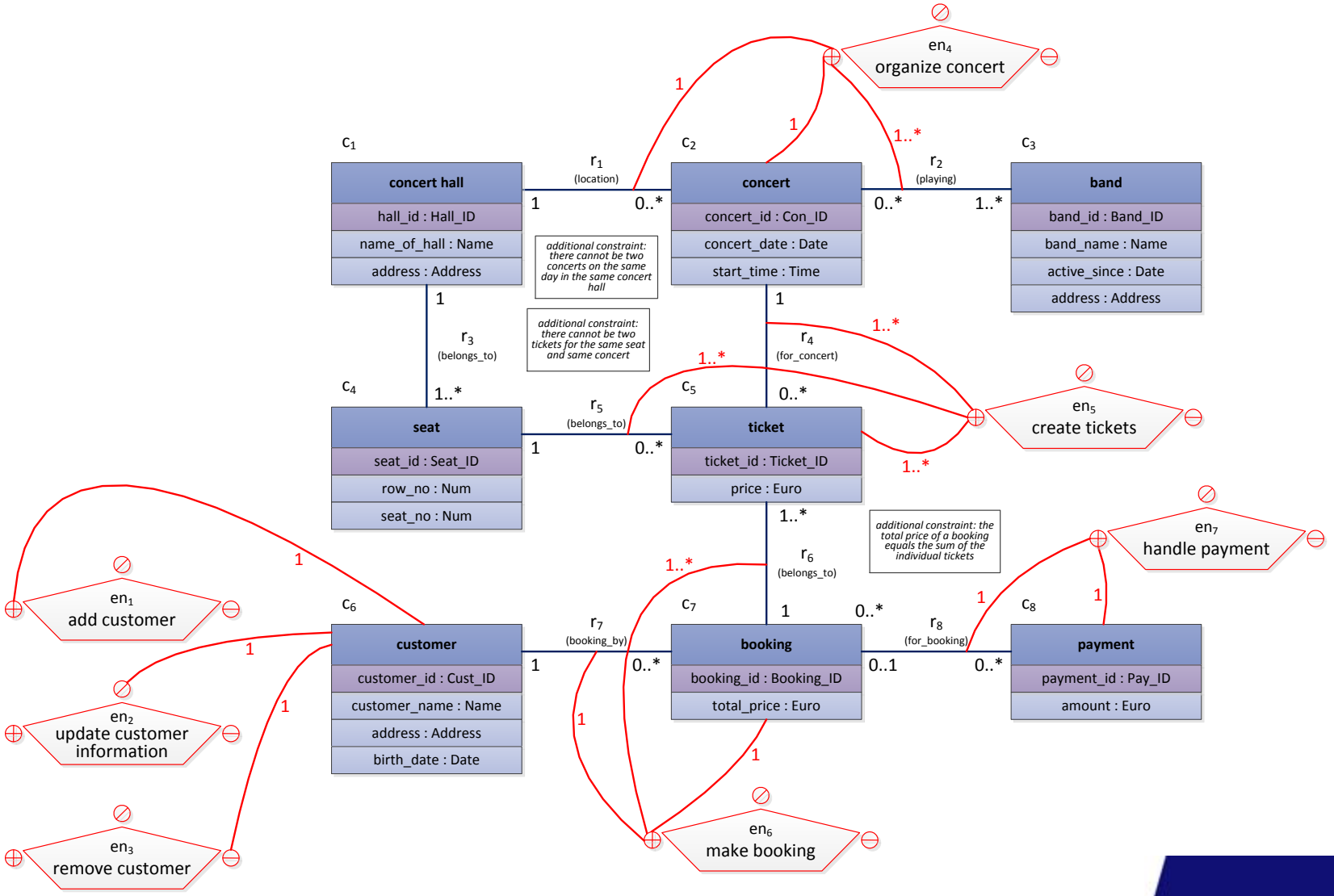


- **state** = content database
- **event** = change of state, i.e., (combinations) of database updates

Possible?

Yes, see e.g. Oracle redo logs!!

Events as combinations of low-level database updates



W.M.P. van der Aalst. Extracting Event Data from Databases to Unleash Process Mining. European BPM Roundtable, Liechtenstein, 2007.


XES
data quality problems
data structure problems
guidelines for logging

Background: Guidelines for making better process models (e.g., understandable and useful)

Guidelines of Business Process Modeling

Jörg Becker[1], Michael Rosemann[2], Christoph von Uthmann[1]

[1] Westfälische Wilhelms-Universität Münster
Department of Information Systems
Steinfurter Str. 109, 48149 Münster, Germany
Phone: +49 (0)251/83-38100, Fax: +49 (0)251/83-38109
{isjobe|ischut}@wi.uni-muenster.de
[2] Queensland University of Technology
School of Information Systems
2 George Street, Brisbane QLD 4001, Australia
Phone: +61 (0)7 3864 1117, Fax: +61 (0)7 3864 1969
m.rosemann@qut.edu.au

Abstract. Process modeling becomes more and more an important task not only for the purpose of software engineering, but also for many other purposes besides the development of software. Therefore it is necessary to evaluate the quality of process models from different viewpoints. This is even more important as the increasing number of different end users, different purposes and the availability of different modeling techniques and modeling tools leads to a higher complexity of information models. In this paper the Guidelines of Modeling (GoM)[1], a framework to structure factors for the evaluation of process models, is presented. Exemplary, Guidelines of Modeling for workflow management and simulation are presented. Moreover, six general techniques for adjusting models to the perspectives of different types of user and purposes will be explained.

1 Complexity and Quality of Business Process Models

The popularity of different process management approaches like Lean Management [58], Activity-based Costing [52], Total Quality Management [21, 35], Business Process Reengineering [16, 17], Process Innovation [7, 8], Workflow Management [14], and Supply Chain Management [39] has two main effects concerning the requirements on process models. First, the number and variety of model designers

Jörg Becker, Michael Rosemann, Christoph von Uthmann: Guidelines of Business Process Modeling. Business Process Management 2000: 30-49

W. van der Aalst et al. (Eds.): Business Process Management, LNCS 1806, pp 30-49, 2000

Information and Software Technology 52 (2010) 127–136

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

Seven process modeling guidelines (7PMG)

J. Mendling [a,*], H.A. Reijers [b], W.M.P. van der Aalst [b]

[a] Humboldt University, Unter den Linden 6, 10099 Berlin, Germany
[b] Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

ARTICLE INFO

Article history:
Received 21 December 2008
Received in revised form 29 July 2009
Accepted 17 August 2009
Available online 23 August 2009

Keywords:
Business process modeling
Model quality
Guidelines

ABSTRACT

Business process modeling is heavily applied in practice, but important quality issues have not been addressed thoroughly by research. A notorious problem is the low level of modeling competence that many casual modelers in process documentation projects have. Existing approaches towards model quality might be of benefit, but they suffer from at least one of the following problems. On the one hand, frameworks like SEQUAL and the Guidelines of Modeling are too abstract to be applicable for novices and non-experts in practice. On the other hand, there are collections of pragmatic hints that lack a sound research foundation. In this paper, we analyze existing research on relationships between model structure on the one hand and error probability and understanding on the other hand. As a synthesis we propose a set of seven process modeling guidelines (7PMG). Each of these guidelines builds on strong empirical insights, yet they are formulated to be intuitive to practitioners. Furthermore, we analyze how the guidelines are prioritized by industry experts. In this regard, the seven guidelines have the potential to serve as an important tool of knowledge transfer from academia into modeling practice.

1. Introduction

Since the 1970s and 1980s, conceptual modeling is a major research area in the IS field. The main motivation to engage in conceptual modeling is to reduce the chances on developing faulty requirements in the early phases of system development [1]. A recent empirical study has shown that *business processes* have become the central objects in many conceptual modeling efforts, e.g. to support their documentation, improvement and automated enactment [2]. This development can be explained by an increased focus of enterprises on those same business processes: they are perceived as the most relevant entities to be managed towards enhanced organizational performance [3].

Usability is an important quality issue of process documentations [4]. As understanding the process is a crucial task in any pro-

analyze and understand. Adequate guidance is of particular importance as large projects on process documentation heavily rely on novices and non-expert modelers [6]. To appreciate the impact of a model that is difficult to assess, it should be realized that in the execution of a single project dozens, hundreds or even thousands of process models may be developed [7,8]. This clarifies why a process model that is immediately usable towards its purpose is of great economic benefit.

Even though some theoretical frameworks and guidelines are available in the area of process modeling, for instance SEQUAL or the Guidelines of Modeling [9,10], these typically require a certain level of modeling competence. They distinguish the major quality categories, but remain too abstract to be directly applicable by non-experts. In other words, such guidelines are hardly related to the concrete actions that process modelers undertake in capturing

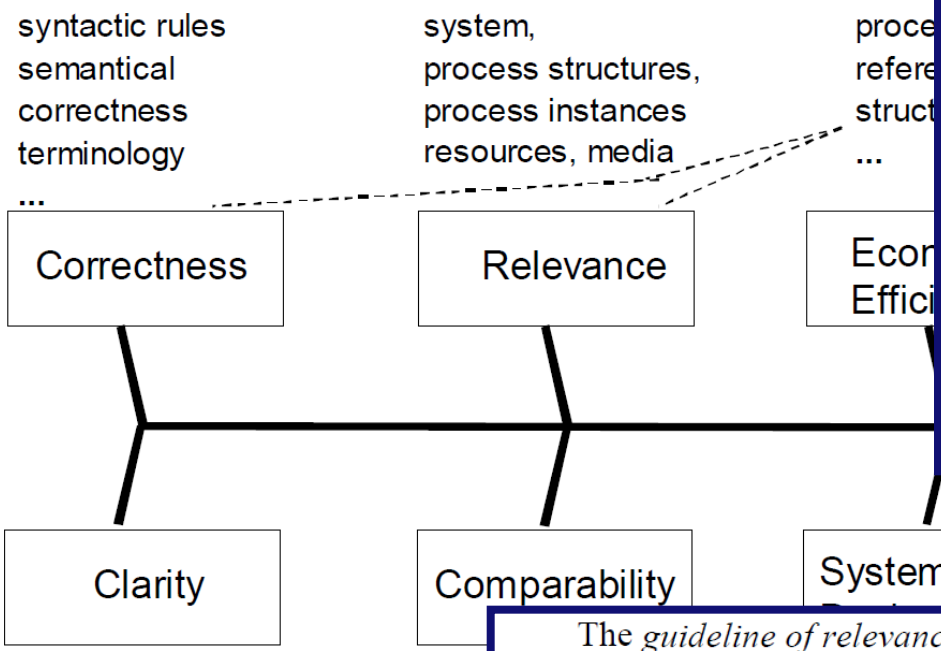
Jan Mendling, Hajo A. Reijers, Wil M. P. van der Aalst: Seven process modeling guidelines (7PMG). Information & Software Technology 52(2): 127-136 (2010)

* Corresponding author. Tel.: +49 30 2093 5805; fax: +49 30 2093 5741.
E-mail addresses: jan.mendling@wiwi.hu-berlin.de (J. Mendling), h.a.reijers@tue.nl (H.A. Reijers), w.m.p.v.d.aalst@tue.nl (W.M.P. van der Aalst).

0950-5849/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.infsof.2009.08.004

called 7PMG. This set is thought to be helpful in guiding users towards improving the quality of their models, in the sense that these are likely (1) to become comprehensible to various

Guidelines of Business P



The pragmatic aspect of the semiotic theory [27] is integrated in the GoM by the *guideline of clarity*. Without a readable, understandable, useful model all other efforts become obsolete. This guideline is extremely subjective and postulates exactly, that the model is understood by the model user. It is not sufficient, if a model designer regard the model as understandable (see also understandability in the GAAP ([9], p. 52). "Construct overload", the situation in the framework of WAND and WEBER in which one object type of an information modeling technique map to at least two ontological constructs is an example for missing clarity as additional knowledge outside the modeling technique is required ([56], p. 211). Mainly layout conventions put this guideline in concrete terms.

The *guideline of comparability* demands the consistent use of all guidelines within a modeling project. It is one of the guidelines which corresponds directly with one GAAP principle, the comparability principle ([19], pp. 551-552). Like the GAAP which aims to increase the comparability between businesses and periods (e.g. avoid different inventory methods like LIFO and FIFO), this guideline includes e. g. the conform application of layout or naming conventions. Otherwise, two models would follow certain, but different rules. The necessity to compare information models is obvious if as-is-models and to-be-models or enterprise-specific and reference models have to be compared.

The *guideline of systematic design* postulates well-defined relationships between information models, which belongs to different views, e.g. the integration of process models with data models. Every input and output data within a process model has to be specified in a corresponding data model. Further interdependencies exist, following for example the ARIS-approach [45, 46, 47], concerning the functions (function view), the organizational units (organizational view), the results of a process (output view) and the involved applications and databases (resource view). One demand is to use a meta model which integrates all relevant views.

Perspectives

The *guideline of correctness* l
correctness. A model is syntactic
meta model (see for definitions
model is based on. For the eval
indispensable to have an explici
postulates that the structure and t
world. Finally, the consistence b
correctness of the model [59].

The *guideline of relevance* postulates

- to select a relevant object system (universe of discourse),
- to take a relevant modeling technique or to configure an existing meta model adequately, and
- to develop a relevant (minimal) model system.

A model includes elements without relevance, if they can be eliminated without loss of meaning for the model user.

The *guideline of economic efficiency* is a constraint to all other guidelines. In the GAAP-context it is called the cost/benefit constraint ([9], p. 51). It is comparable to the criteria "feasibility" of LINDLAND ET AL. [27] and restricts e.g. the correctness or the clarity of a model. Approaches to support the economic efficiency are reference models, appropriate modeling tools or the re-use of models.

Seven process modeling guidelines (7PMG)

G1	Use a
G2	Mini
G3	Use c
G4	Mode
G5	Avoi

G1: Use as few elements in the model as possible
model has undesirable effects on understand
lihood of errors: Larger models tend to be n
understand [31] and have a higher error p
small models [41,44].

G2: Minimize the routing paths per element.
degree of an element in the process model,
of input and output arcs together, the harde
understand the model [31]. As shown in
strong correlation between the number of n
and the average or maximum degree of elem

G3: Use one start and one end event. The numb
end events is positively connected with an i
probability [44]. Most workflow engines r
start and end node [46]. Moreover, models
requirement are easier to understand and all
or analysis (e.g., soundness checks).

G4: Model as structured as possible. A process model is structured if every split connector matches a respective join connector of the same type. Structured models can be seen as formulas with balanced brackets, i.e., every opening bracket has a corresponding closing bracket of the same type. Unstructured models are not only more likely to include errors [44], people also tend to understand them less easily [31].

G5: Avoid OR routing elements. Models that have only AND and XOR connectors are less error-prone [44]. Furthermore, there are some ambiguities in the semantics of the OR-join leading to paradoxes and implementation problems [47].

G6: Use verb-object activity labels. A wide exploration of labeling styles that are used in actual process models, discloses the existence of two popular styles and a rest category [48]. From these, people consider the verb-object style, like "Inform complainant", as significantly less ambiguous and more useful than action-noun labels (e.g. "Complaint analysis") or labels that follow neither of these styles (e.g. "Incident agenda") [40].

G7: Decompose the model if it has more than 50 elements. This guideline relates to **G1** that is motivated by a positive correlation between size and errors. For models with more than 50 elements the error probability tends to be higher than 50% [44]. Therefore, large models should be split up into

Guidelines for logging

initial proposal

not about a specific syntax

Your input is welcome!

- Starting point: events refer to "things" that happen and events are described by **references** and **attributes**.
- References have a **reference name** and an **identifier** that refers to some object (person, case, ticket, machine, room, etc.) in the universe of discourse.
- **Attributes** have a **name** and a **value**, e.g., age=48 or time="28-6-2014 03:14:07".

[GL1]

Reference and variable names should have clear semantics, i.e., they should have the same meaning for all people involved in creating and analyzing event data.

[GL2]

There should be a structured and managed collection of reference and variable names.

Ideally, names are grouped hierarchically (like a taxonomy or ontology). A new reference and variable name can only be added after there is consensus on its value and meaning. Also consider adding domain or organization specific extensions (see extension mechanism of XES).

[GL3]

References should be stable (e.g., identifiers should not be reused or rely on the context).

For example, references should not be time or language dependent.

[GL4]

Attribute values should be as precise as possible. If the value does not have the desired precision, this should be indicated explicitly (e.g., through a qualifier).

For example, if for some events only the date is known but not the exact timestamp, then this should be stated explicitly.

[GL5]

Uncertainty with respect to the occurrence of the event or its references or attributes should be captured through appropriate qualifiers.

For example, due to communication errors, some values may be less reliable than usual. Note that uncertainty is different from imprecision.

[GL6]

Events should be at least partially ordered. The ordering of events may be stored explicitly (e.g., using a list) or implicitly through a variable denoting the event's timestamp.

[GL7]

If possible, also store transactional information about the event (start, complete, abort, schedule, assign, suspend, resume, withdraw, etc.).

Having start and complete events allows for the computation of activity durations. Store activity references to be able to relate events belonging to the same activity instance. Without activity references it may not always be clear which events belong together.

[GL8]

Perform regularly automated consistency and correctness checks to ensure the syntactical correctness of the event log.

Check for missing references or attributes, and reference/attribute names not agreed upon. Event quality assurance is a continuous process.

[GL9]

Ensure comparability of event logs over time and different groups of cases or process variants.

The logging itself should not change over time (without being reported). For comparative process mining, it is vital that the same logging principles are used. If for some groups of cases, some events are not recorded even though they occur, then this may suggest differences that do not actually exist.


[GL10]
Do not aggregate events in the event log used as input for the analysis process.
Aggregation should be done during analysis and not before (since it cannot be undone).
Event data should be as "raw" as possible.

[GL11]

Do not remove events and ensure provenance. Reproducibility is key for process mining.

For example, do not remove a student from the database after he dropped out since this may lead to misleading analysis results. Mark objects as not relevant (a so-called soft delete) rather than deleting them: concerts are not deleted - they are canceled, employees are not deleted - they are fired, etc.

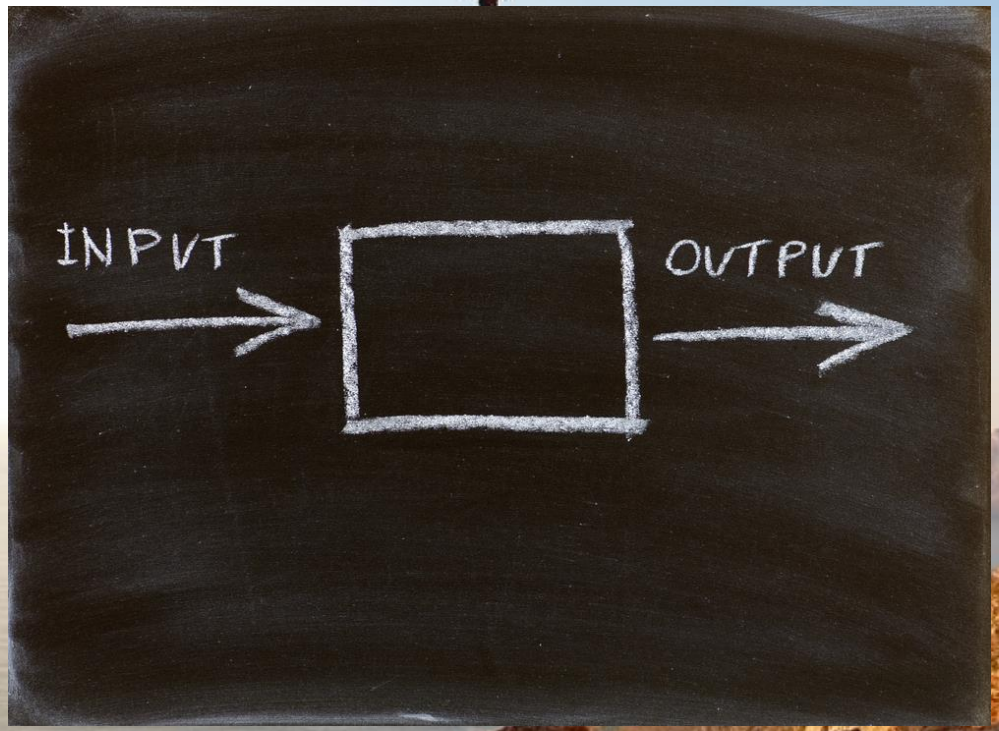
[GL12]

Ensure privacy without losing meaningful correlations.

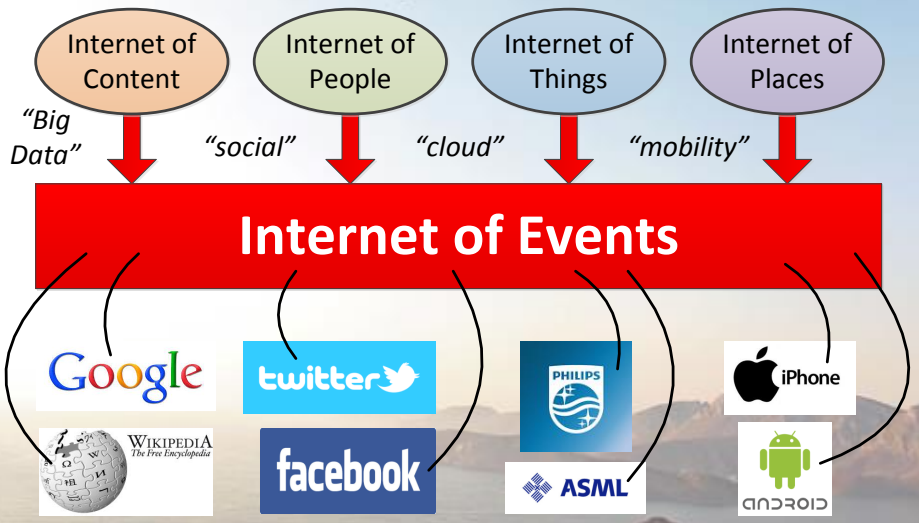
Sensitive or private data should be removed as early as possible (i.e., before analysis). However, if possible, one should avoid removing correlations. For example, it is often not useful to know the name of a student, but it may be important to still be able to use his high school marks and know what other courses he failed. Hashing can be a value tool in the trade-off between privacy and analysis.

- [GL1] Reference and variable names should have clear semantics.
- [GL2] There should be a structured and managed collection of reference and variable names.
- [GL3] References should be stable.
- [GL4] Attribute values should be as precise as possible.
- [GL5] Uncertainty with respect to the occurrence of the event or its references or attributes should be captured.
- [GL6] Events should be at least partially ordered.
- [GL7] If possible, also store transactional information about the event.
- [GL8] Perform regularly automated consistency and correctness checks.
- [GL9] Ensure comparability of event logs over time and different groups of cases or process variants.
- [GL10] Do not aggregate events in the event log used as input for the analysis process.
- [GL11] Do not remove events and ensure provenance.
- [GL12] Ensure privacy without losing meaningful correlations

Conclusion

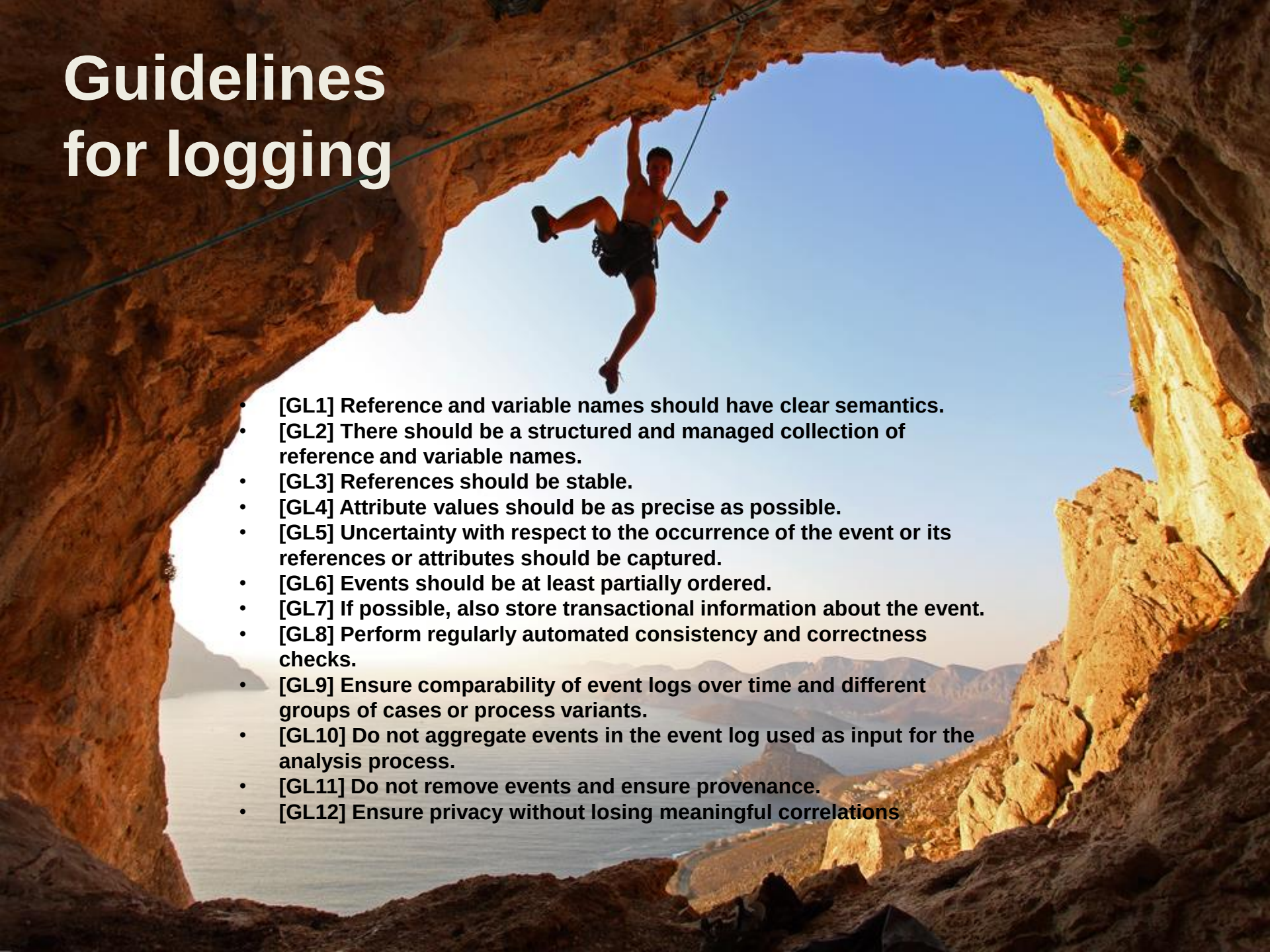


needed: data/process scientists !!!




Problems:
• data quality
• data structure

Guidelines for logging



- [GL1] Reference and variable names should have clear semantics.
- [GL2] There should be a structured and managed collection of reference and variable names.
- [GL3] References should be stable.
- [GL4] Attribute values should be as precise as possible.
- [GL5] Uncertainty with respect to the occurrence of the event or its references or attributes should be captured.
- [GL6] Events should be at least partially ordered.
- [GL7] If possible, also store transactional information about the event.
- [GL8] Perform regularly automated consistency and correctness checks.
- [GL9] Ensure comparability of event logs over time and different groups of cases or process variants.
- [GL10] Do not aggregate events in the event log used as input for the analysis process.
- [GL11] Do not remove events and ensure provenance.
- [GL12] Ensure privacy without losing meaningful correlations