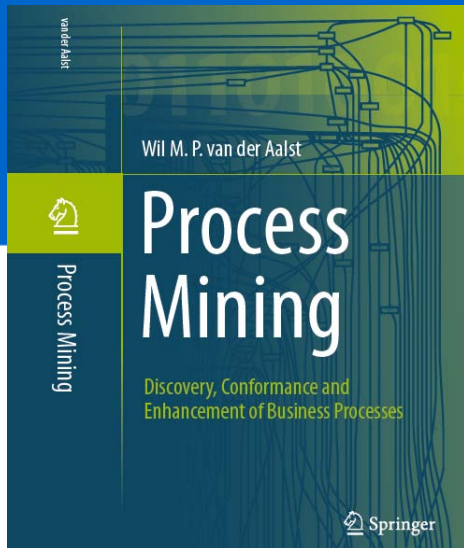


# Chapter 4

## Getting the Data

prof.dr.ir. Wil van der Aalst  
[www.processmining.org](http://www.processmining.org)



**TU** / **e** Technische Universiteit  
**Eindhoven**  
University of Technology

Where innovation starts

# Overview

Chapter 1  
Introduction

---

## *Part I: Preliminaries*

Chapter 2  
Process Modeling and  
Analysis

Chapter 3  
Data Mining

---

## *Part II: From Event Logs to Process Models*

Chapter 4  
Getting the Data

Chapter 5  
Process Discovery: An  
Introduction

Chapter 6  
Advanced Process  
Discovery Techniques

---

## *Part III: Beyond Process Discovery*

Chapter 7  
Conformance  
Checking

Chapter 8  
Mining Additional  
Perspectives

Chapter 9  
Operational Support

---

## *Part IV: Putting Process Mining to Work*

Chapter 10  
Tool Support

Chapter 11  
Analyzing “Lasagna  
Processes”

Chapter 12  
Analyzing “Spaghetti  
Processes”

---

## *Part V: Reflection*

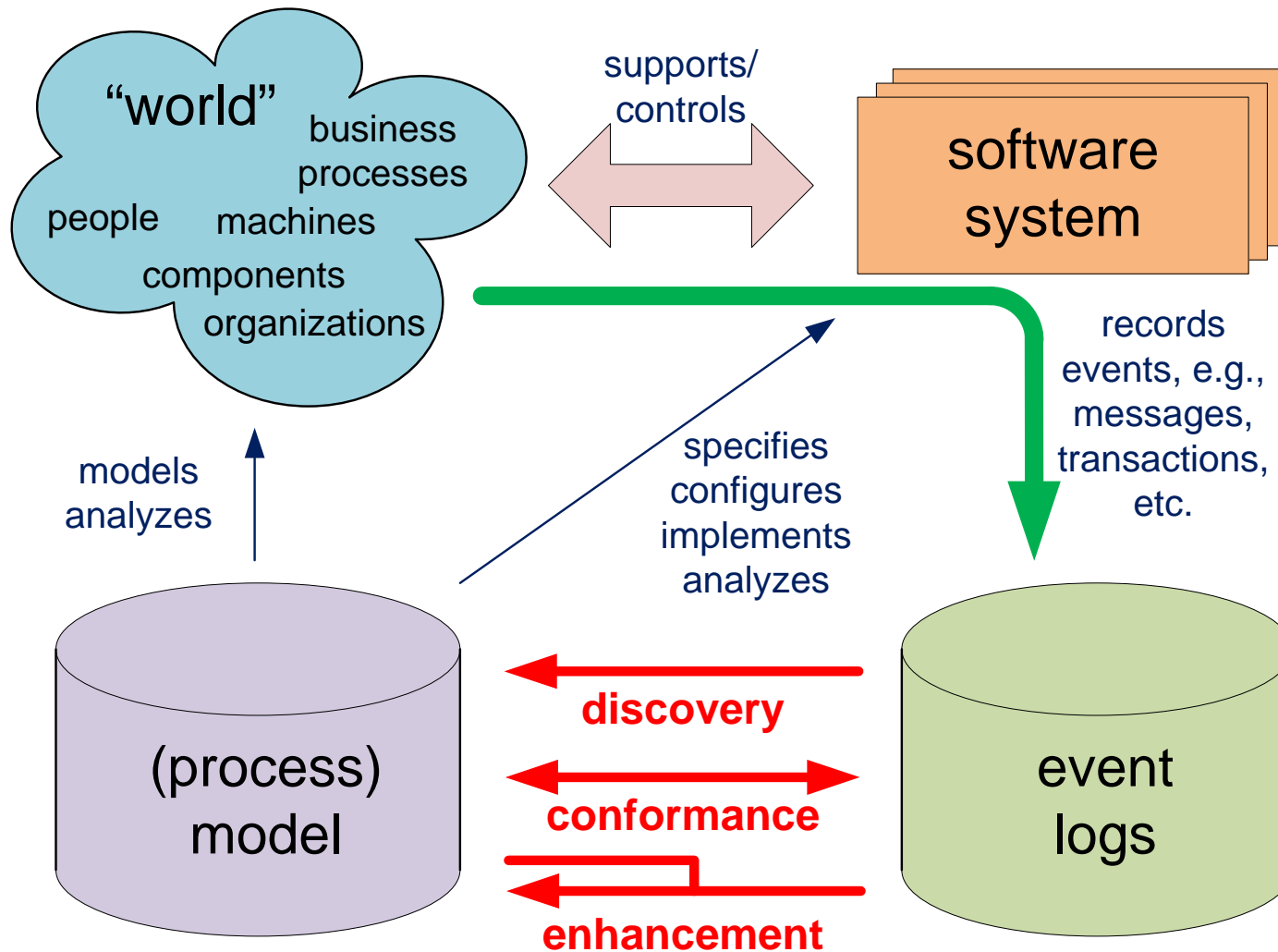
Chapter 13  
Cartography and  
Navigation

Chapter 14  
Epilogue

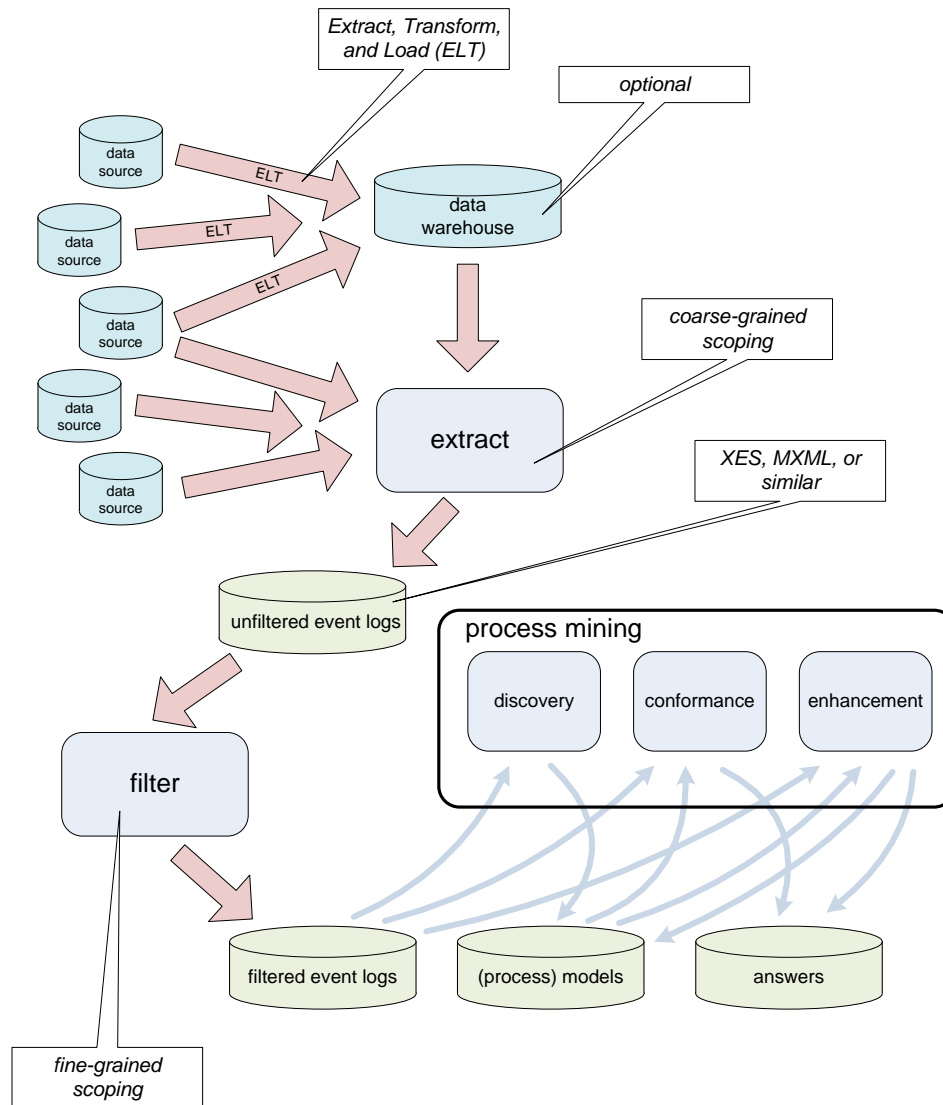
# Goal of process mining

- **What really happened in the past?**
- **Why did it happen?**
- **What is likely to happen in the future?**
- **When and why do organizations and people deviate?**
- **How to control a process better?**
- **How to redesign a process to improve its performance?**

# Getting the data



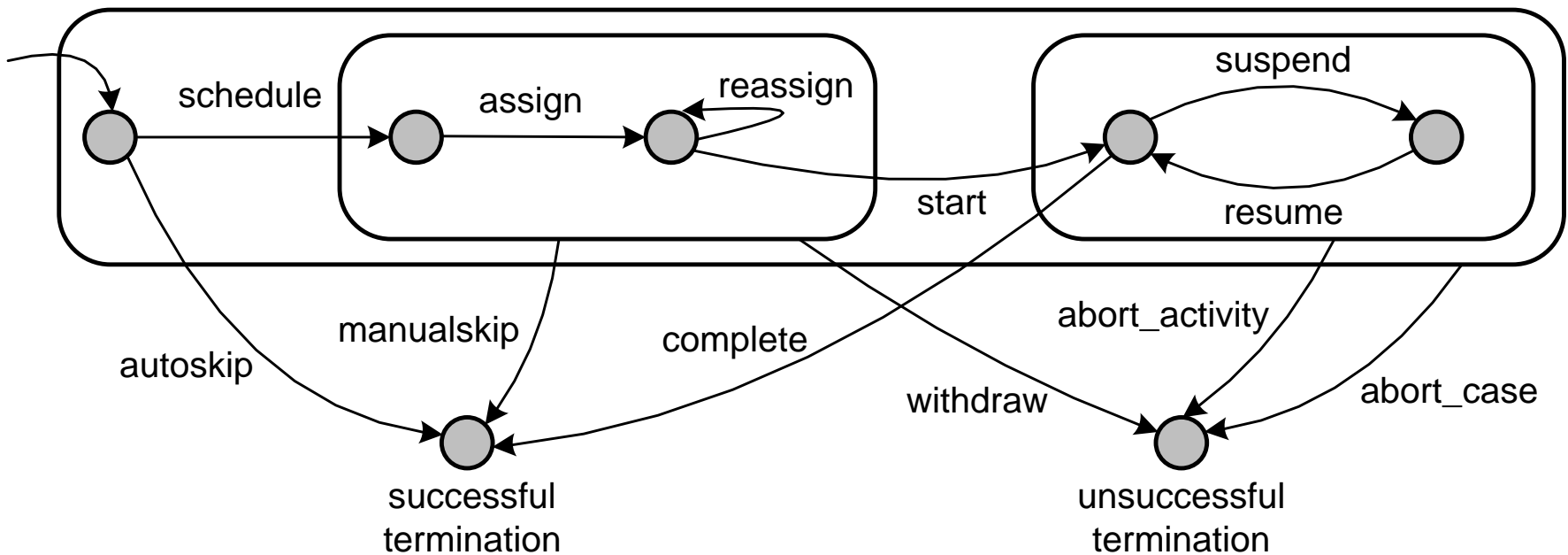
# From heterogeneous data sources to process mining results



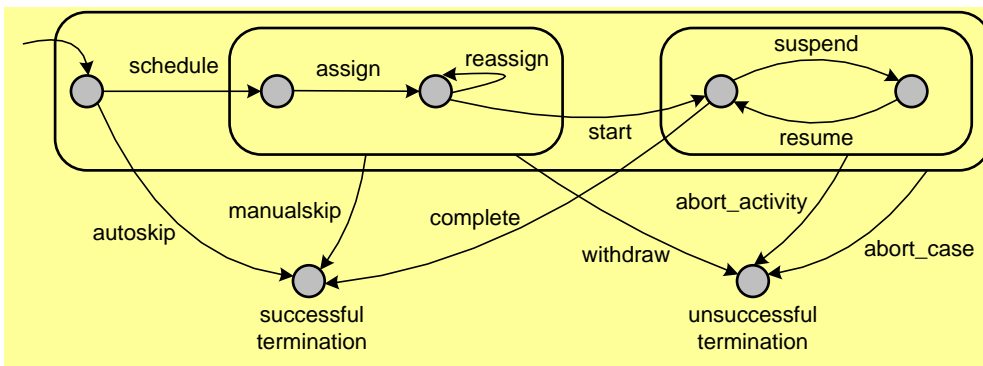
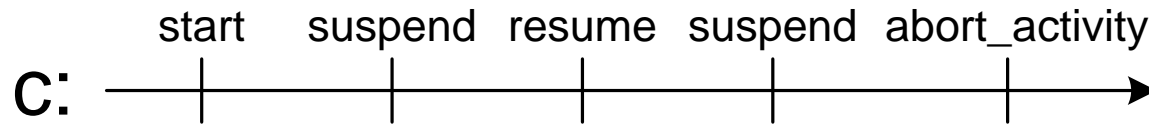
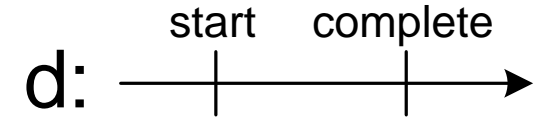
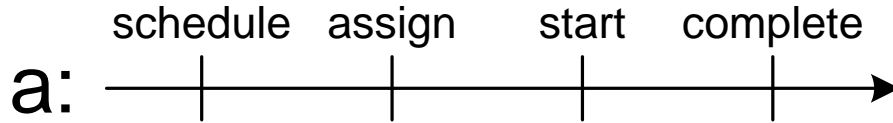




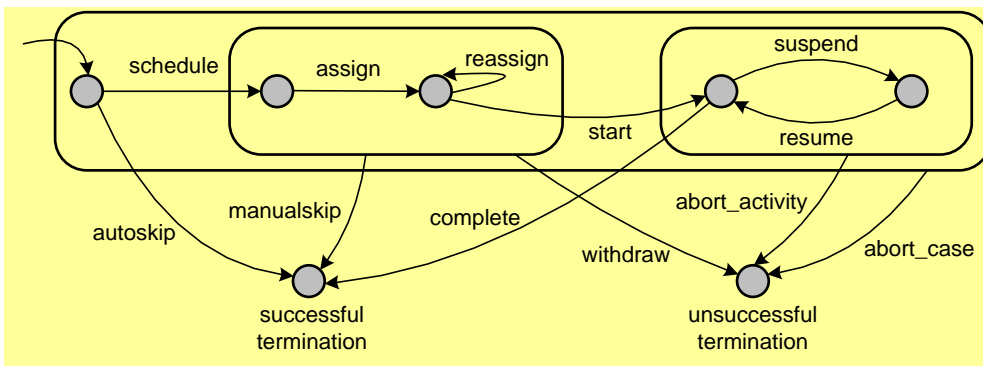
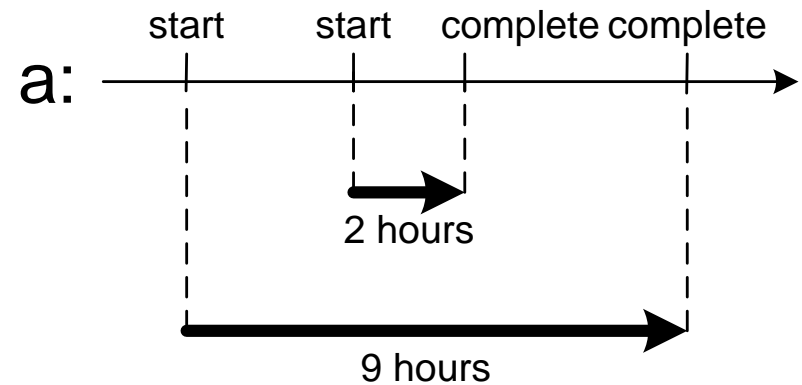
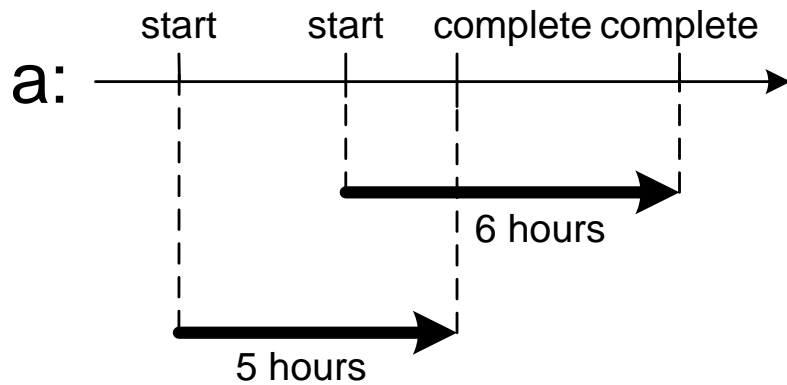
# Standard transactional life-cycle model



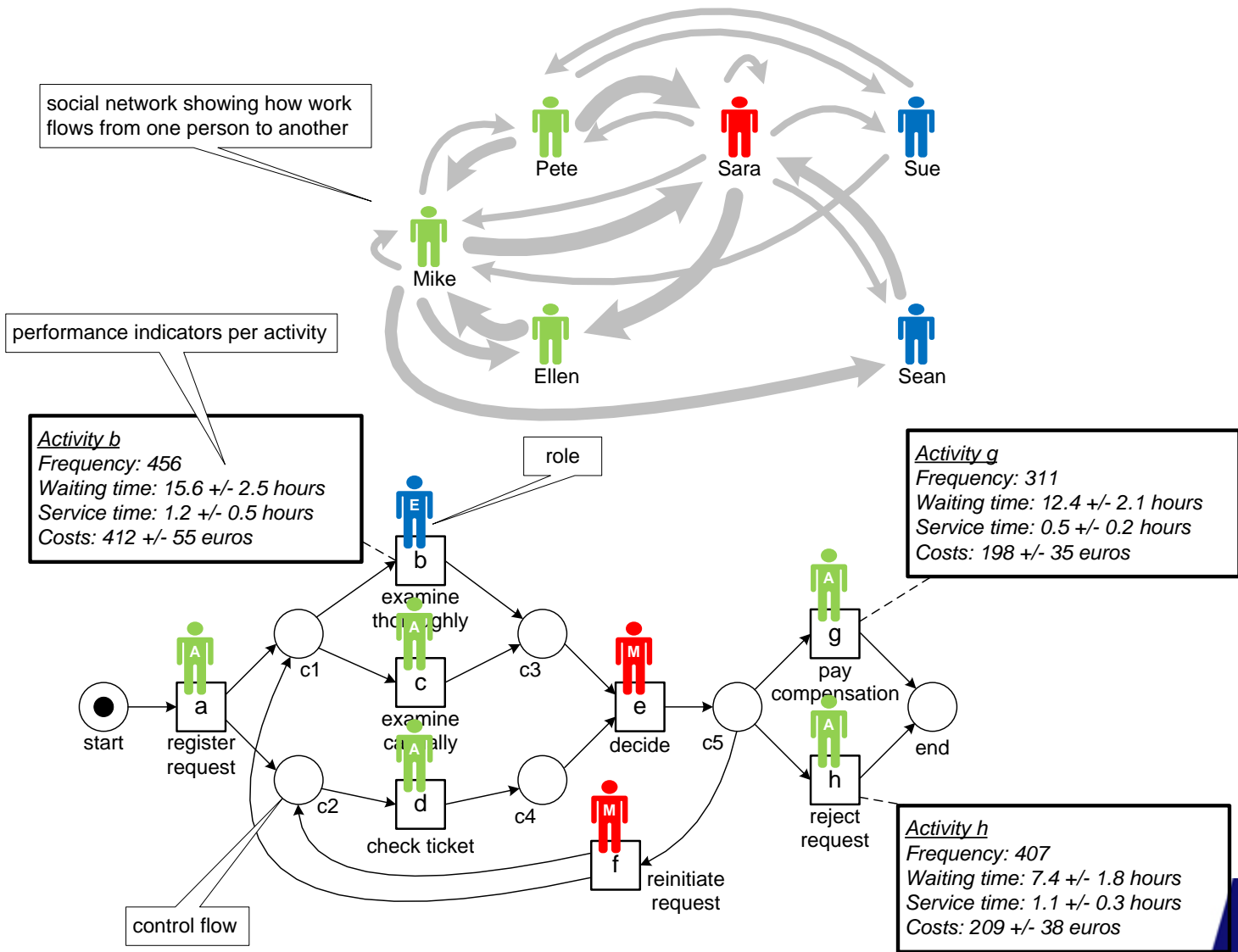
# Five activity instances



# Overlapping activity instances

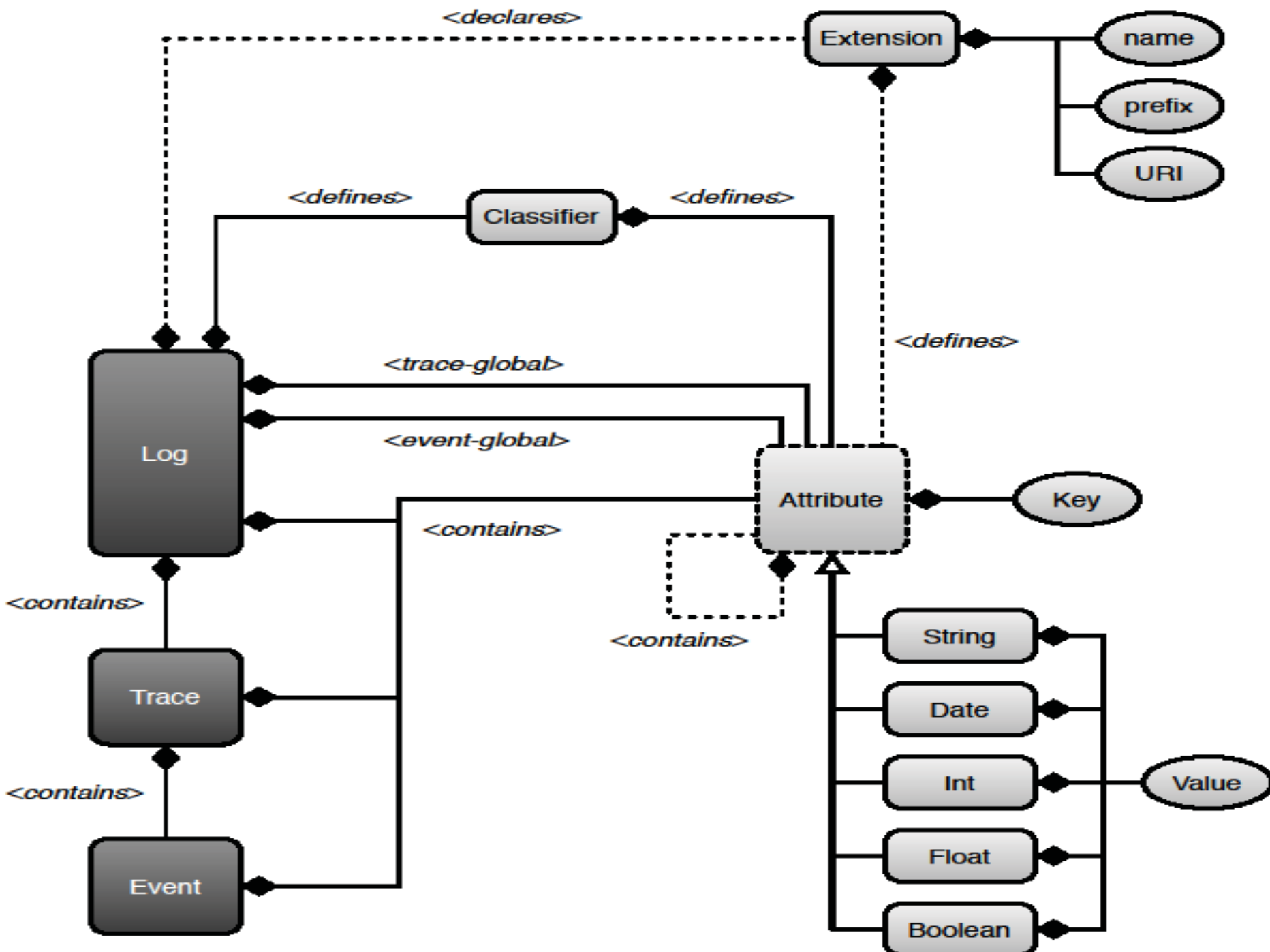


# Using attributes



# XES (eXtensible Event Stream)

- See [www.xes-standard.org](http://www.xes-standard.org).
- Adopted by the IEEE Task Force on Process Mining.
- Predecessor: MXML and SA-MXML.
- The format is supported by tools such as ProM (as of version 6), Nitro, XESame, and OpenXES.
- ProMimport supports MXML.



# XES (compatible with MXML)

## Event log consists of:

- **traces (process instances)**
  - events
- **Standard extensions:**
  - **concept (for naming)**
  - **lifecycle (for transactional properties)**
  - **org (for the organizational perspective)**
  - **time (for timestamps)**
  - **semantic (for ontology references)**

```
<log xes.version="1.0" xes.features="nested-attributes" openxes.version="1.0RC7">
  <extension name="Concept" prefix="concept" uri="http://code.deckfour.org/xes/concept.xesext"/>
  <extension name="Semantic" prefix="semantic" uri="http://code.deckfour.org/xes/semantic.xesext"/>
  <extension name="Time" prefix="time" uri="http://code.deckfour.org/xes/time.xesext"/>
  <extension name="Organizational" prefix="org" uri="http://code.deckfour.org/xes/org.xesext"/>
  <extension name="Lifecycle" prefix="lifecycle" uri="http://code.deckfour.org/xes/lifecycle.xesext"/>
- <global scope="trace">
  <string key="conceptname" value="__INVALID__"/>
</global>
- <global scope="event">
  <string key="conceptname" value="__INVALID__"/>
  <string key="lifecycle:transition" value="complete"/>
</global>
<classifier name="MXML Legacy Classifier" keys="conceptname lifecycle:transition"/>
<classifier name="Event Name" keys="conceptname"/>
<classifier name="Resource Name" keys="conceptname"/>
<string key="source" value="http://code.deckfour.org/xes/trace_more_data.zip"/>
<string key="conceptname" value="Simulated process"/>
<string key="lifecycle:mode" value="standard"/>
<string key="description" value="Simulated process"/>
- <trace>
  <string key="conceptname" value="1"/>
  <string key="description" value="Simulated process instance"/>
- <event>
  <string key="orgresource" value="Mike"/>
  <date key="time:timestamp" value="2006-01-01T00:00:00.000+01:00"/>
  <string key="conceptname" value="invite reviewers"/>
  <string key="lifecycle:transition" value="start"/>
</event>
- <event>
  <string key="orgresource" value="Mike"/>
  <date key="time:timestamp" value="2006-01-06T00:00:00.000+01:00"/>
  <string key="conceptname" value="invite reviewers"/>
  <string key="lifecycle:transition" value="complete"/>
</event>
- <event>
```

**extensions loaded**

**every trace has a name**

**every event has a name and a transition**

**start of trace (i.e. process instance)**

**classifier = name + transition**

**name of trace**

**resource**

**timestamp**

**name of event (activity name)**

**transition**

```
<event>
  <string key="orgresource" value="Anne"/>
  <date key="time.timestamp" value="2009-06-23T01:00:00.000+02:00"/>
  <string key="conceptname" value="accept"/>
  <string key="lifecycle.transition" value="start"/>
</event>
```

end of trace (i.e. process instance)

```
<event>
  <string key="orgresource" value="Anne"/>
  <date key="time.timestamp" value="2009-06-28T01:00:00.000+02:00"/>
  <string key="conceptname" value="accept"/>
  <string key="lifecycle.transition" value="complete"/>
</event>
```

start of trace

```
</trace>
<trace>
  <string key="conceptname" value="68"/>
  <string key="description" value="Simulated process instance"/>
```

name of trace

```
<event>
  <string key="orgresource" value="Mike"/>
  <date key="time.timestamp" value="2006-10-14T01:00:00.000+02:00"/>
  <string key="conceptname" value="invite reviewers"/>
  <string key="lifecycle.transition" value="start"/>
</event>
```

resource

timestamp

name of event (activity name)

```
<event>
  <string key="orgresource" value="Mike"/>
  <date key="time.timestamp" value="2006-10-14T01:00:00.000+02:00"/>
  <string key="conceptname" value="invite reviewers"/>
  <string key="lifecycle.transition" value="complete"/>
</event>
```

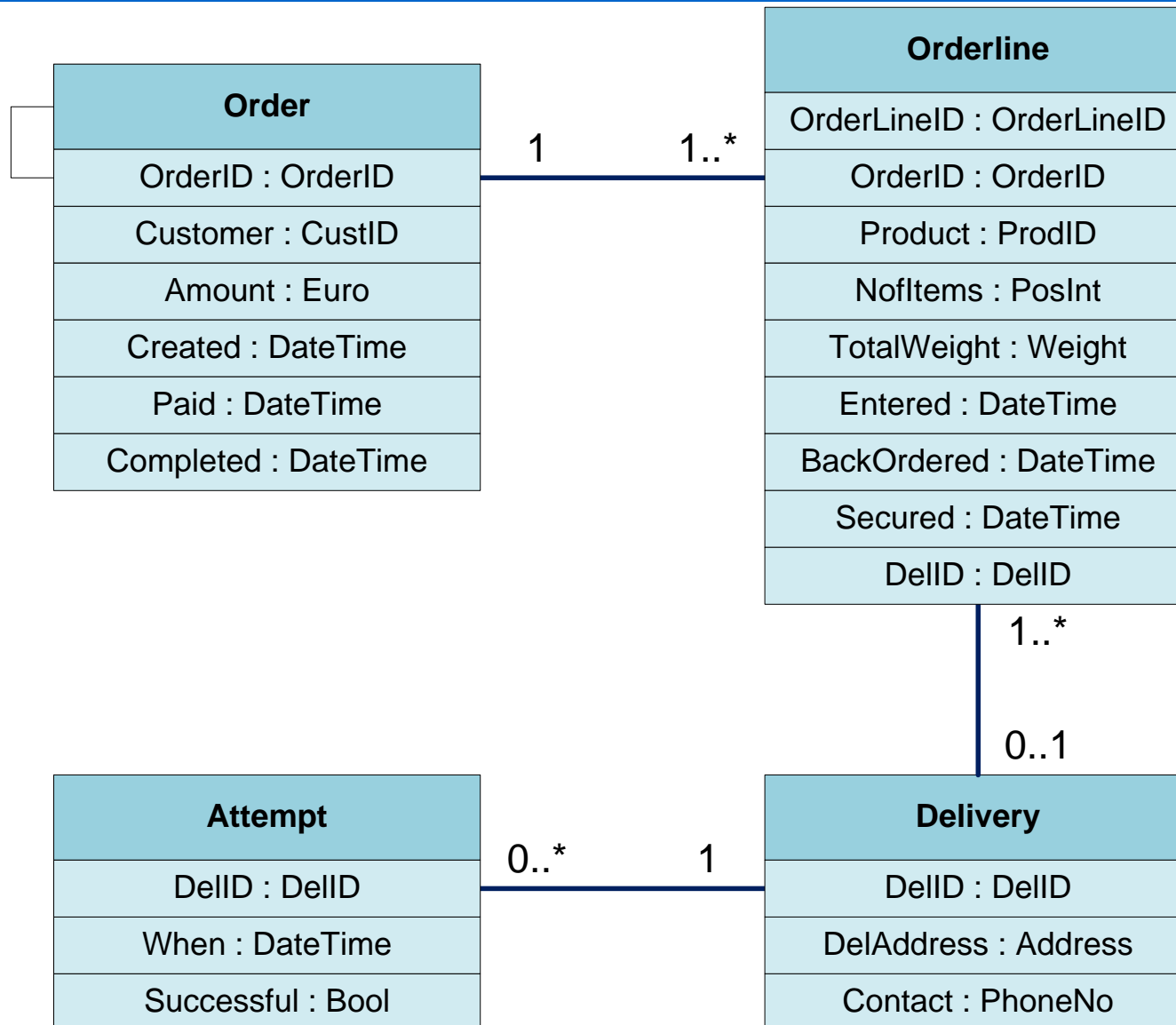
data associated to event

```
<event>
  <string key="orgresource" value="Pam"/>
  <date key="time.timestamp" value="2006-10-14T01:00:00.000+02:00"/>
  <string key="lifecycle.transition" value="complete"/>
  <string key="Result by Reviewer A" value="reject"/>
  <string key="conceptname" value="get review 1"/>
</event>
```

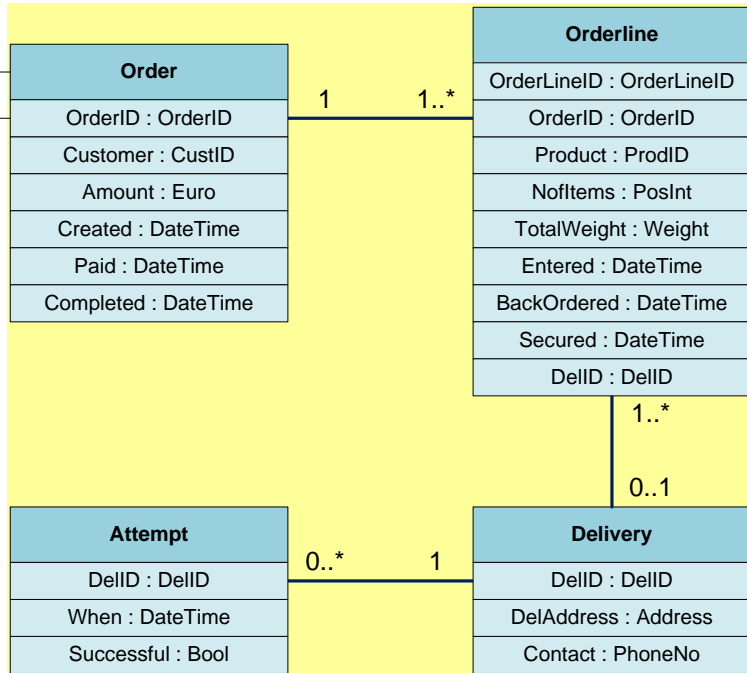
# Challenges when extracting event logs

- **Correlation:** Events in an event log are grouped per case. This simple requirement can be quite challenging as it requires event correlation, i.e., events need to be related to each other.
- **Timestamps:** Events need to be ordered per case. Typical problems: only dates, different clocks, delayed logging.
- **Snapshots.** Cases may have a lifetime extending beyond the recorded period, e.g., a case was started before the beginning of the event log.
- **Scoping.** How to decide which tables to incorporate?
- **Granularity:** the events in the event log are at a different level of granularity than the activities relevant for end users.

# Flattening reality into event logs



# Tables

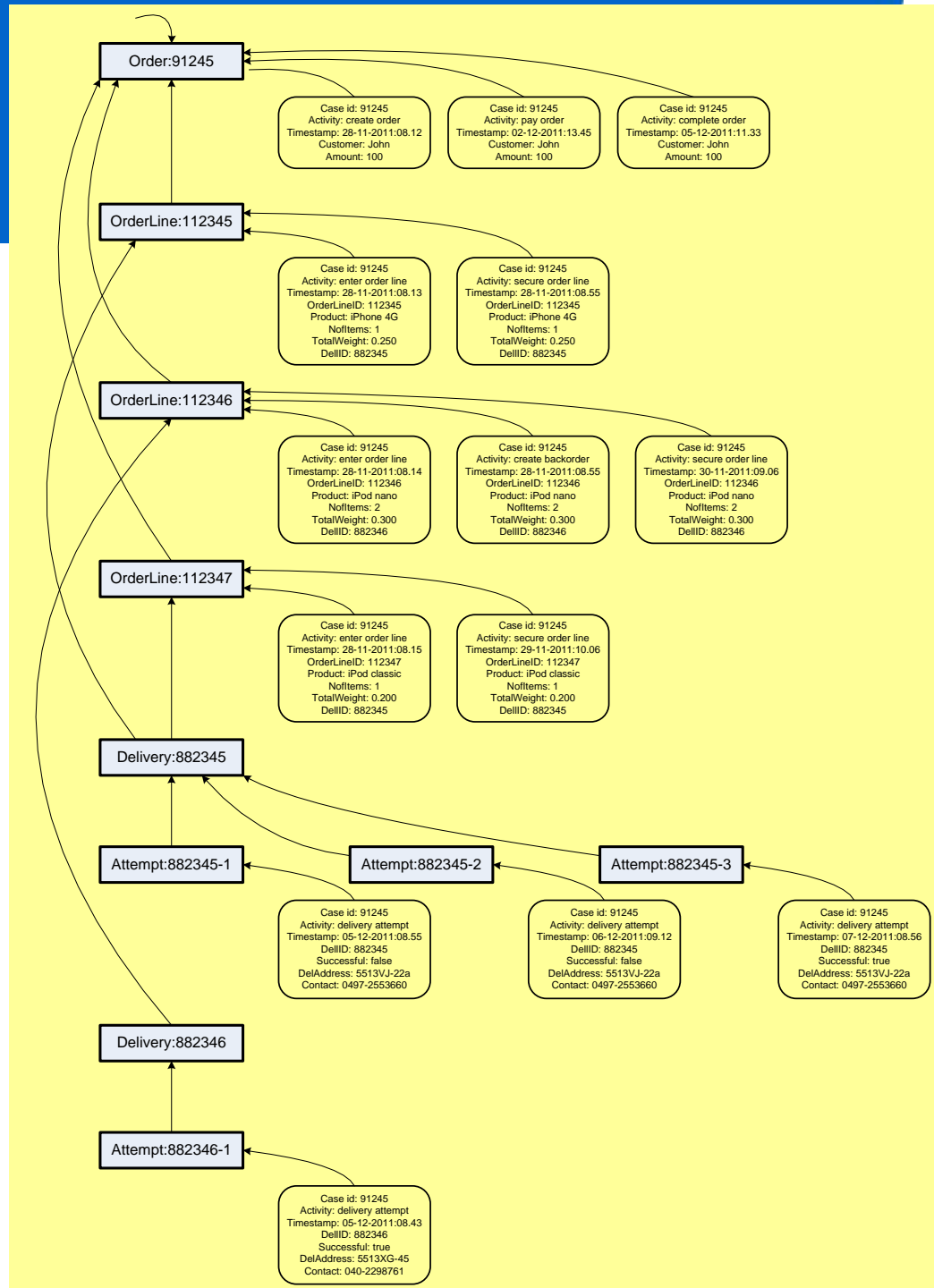
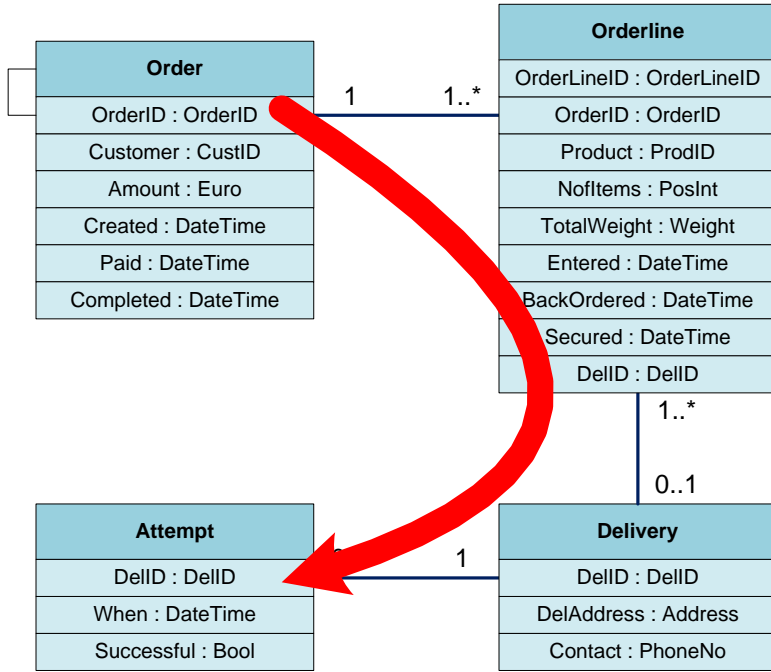


Order					
OrderID	Customer	Amount	Created	Paid	Completed
91245	John	100	28-11-2011:08.12	02-12-2011:13.45	05-12-2011:11.33
91561	Mike	530	28-11-2011:12.22	03-12-2011:14.34	05-12-2011:09.32
91812	Mary	234	29-11-2011:09.45	02-12-2011:09.44	04-12-2011:13.33
92233	Sue	110	29-11-2011:10.12	null	null
92345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

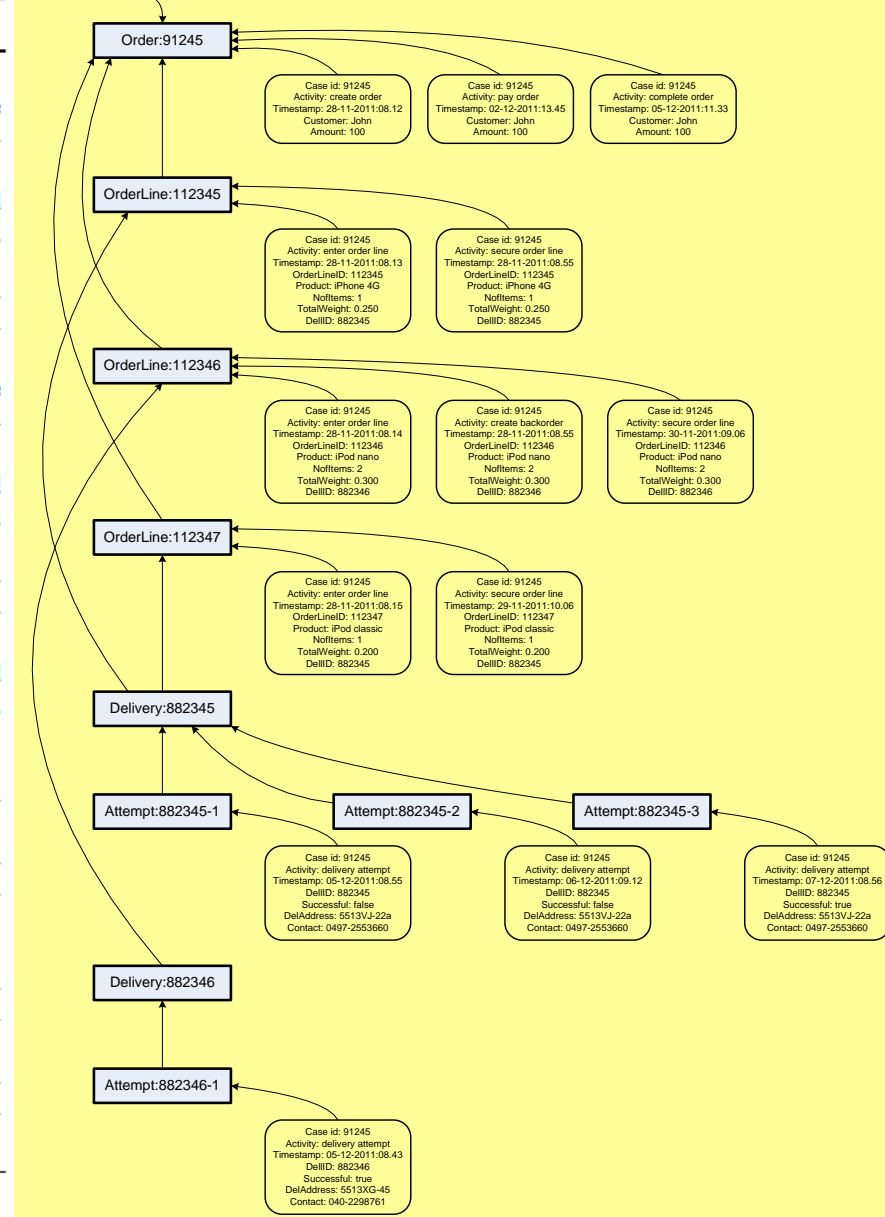
Delivery			Attempt		
DelIID	DelAddress	Contact	DelIID	When	Successful
882345	5513VJ-22a	0497-2553660	882345	05-12-2011:08.55	false
882346	5513XG-45	040-2298761	882345	06-12-2011:09.12	false
...	...	...	882345	07-12-2011:08.56	true
			882346	05-12-2011:08.43	true
...	...	...	...	...	...

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, DelIID: 882345
91245	enter order line	28-11-2011:08.14	OrderLineID: 112346, Product: iPod nano, NofItems: 2, TotalWeight: 0.300, DelIID: 882346
91245	enter order line	28-11-2011:08.15	OrderLineID: 112347, Product: iPod classic, NofItems: 1, TotalWeight: 0.200, DelIID: 882345
91245	secure order line	28-11-2011:08.55	OrderLineID: 112345, Product: iPhone 4G, NofItems: 1, TotalWeight: 0.250, DelIID: 882345
91245	create backorder	28-11-2011:08.55	OrderLineID: 112346, Product: iPod nano, NofItems: 2, TotalWeight: 0.300, DelIID: 882346
91245	secure order line	29-11-2011:10.06	OrderLineID: 112347, Product: iPod classic, NofItems: 1, TotalWeight: 0.200, DelIID: 882345
91245	secure order line	30-11-2011:09.06	OrderLineID: 112346, Product: iPod nano, NofItems: 2, TotalWeight: 0.300, DelIID: 882346
91245	pay order	02-12-2011:13.45	Customer: John, Amount: 100
91245	delivery attempt	05-12-2011:08.43	DelIID: 882346, Successful: true, DelAddress: 5513XG-45, Contact: 040-2298761
91245	delivery attempt	05-12-2011:08.55	DelIID: 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	DelIID: 882345, Successful: false, DelAddress: 5513VJ-22a, Contact: 0497-2553660
91245	delivery attempt	07-12-2011:08.56	DelIID: 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, DelIID: 882345
...	...	...	...



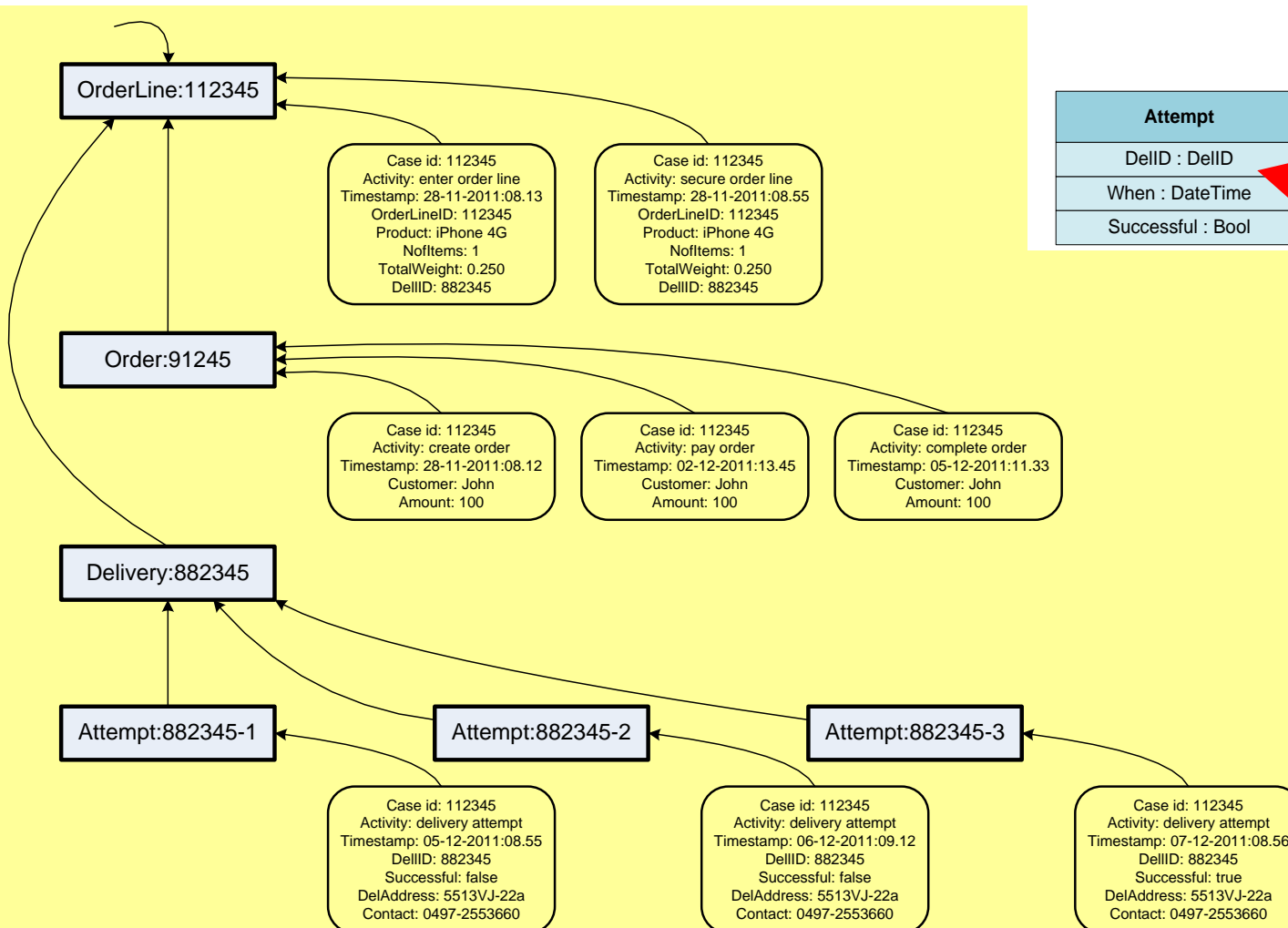
# Orderline instance

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

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

Attempt
DelID : DelID
When : DateTime
Successful : Bool

Delivery
DelID : DelID
DelAddress : Address
Contact : PhoneNo



# Other examples

- **The life cycles of reviewers, authors, papers, reviews, PC chairs, etc.**
- **The life cycles of job applications and vacancies.**
- **X-ray machine logs: machine, machine day, patient, treatment, routine, etc.?**
  
- **Therefore, the selection and scoping of instances is needed.**
- **Like making deciding on the elements to be put on map; there may be many maps covering partially overlapping areas.**

# Extracting event logs

- **Not just a syntactical issue.**
- **Different views are possible.**
- **Important:**
  - **Selecting the right instance notion.**
  - **Ordering of events.**
  - **Selection of events.**
- **Proclets: the true fabric of real-life processes.**