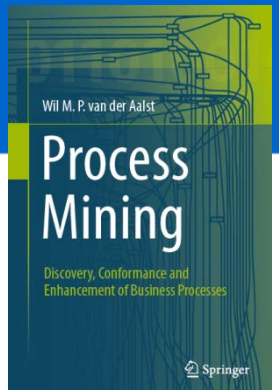
Process Mining: Data Science in Action

Getting the Right Event Data

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



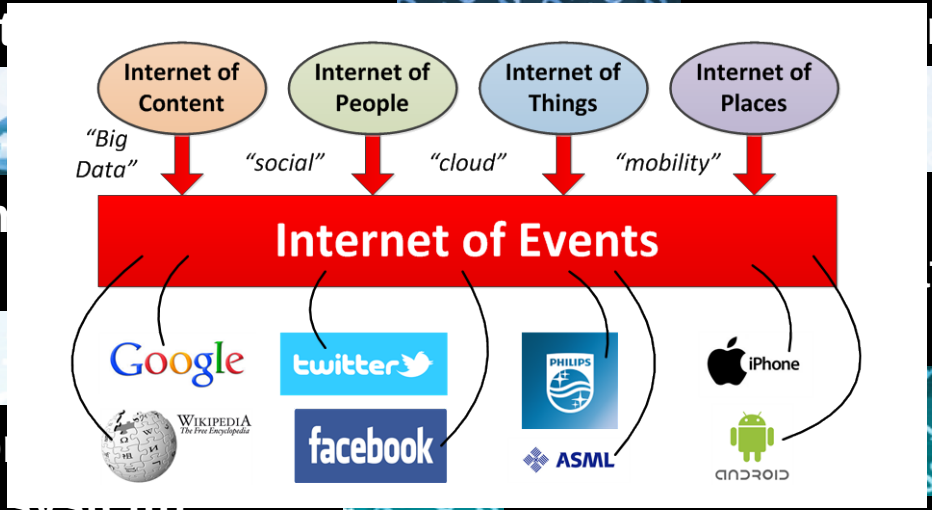
Technische Universiteit Eindhoven University of Technology

Where innovation starts

Event data

a database system (e.g., patient dat

a message log (e.g., from middleware)



a con

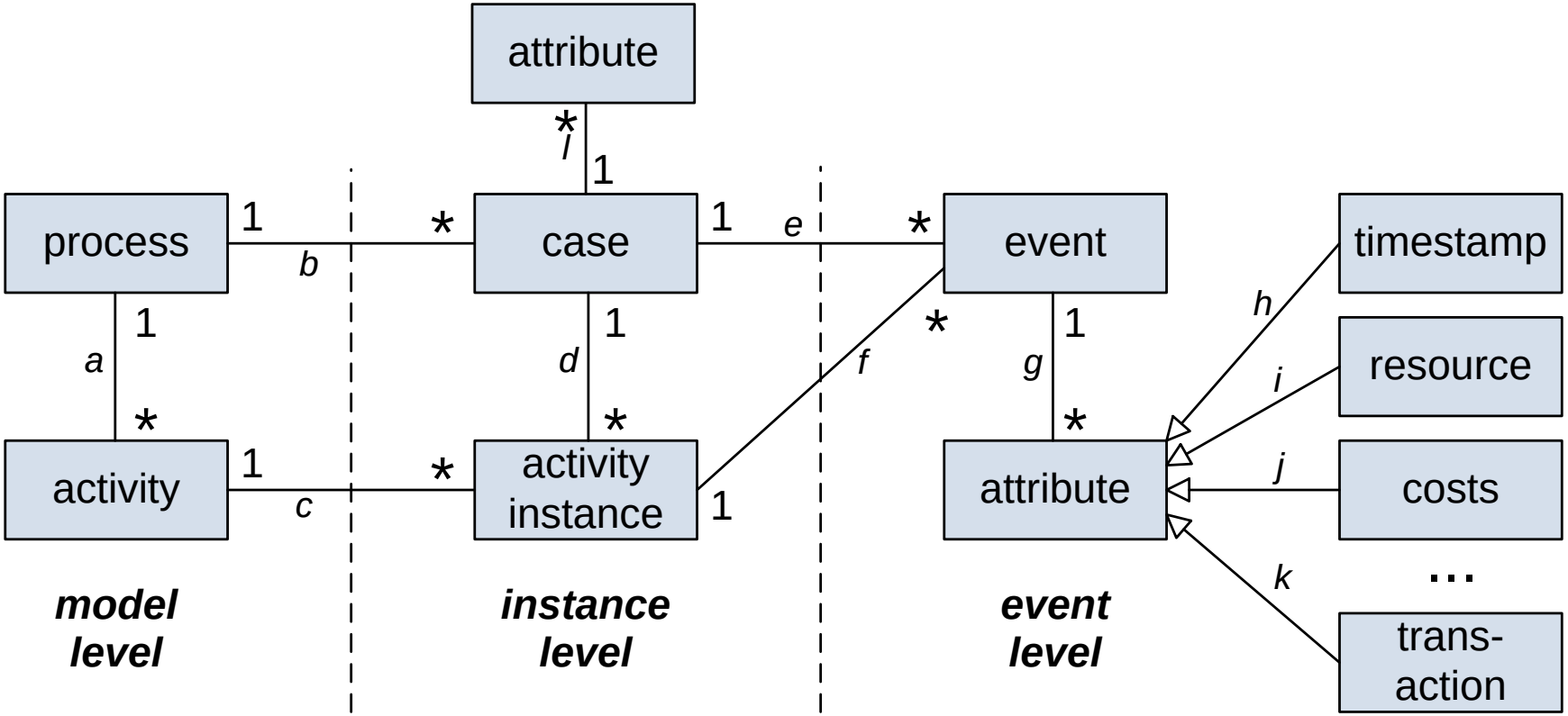
open API providing ta from websites or social media

a transactio trading system)

a business suite/ERP system (SAP, Oracle, etc.)

Conceptual model of event logs

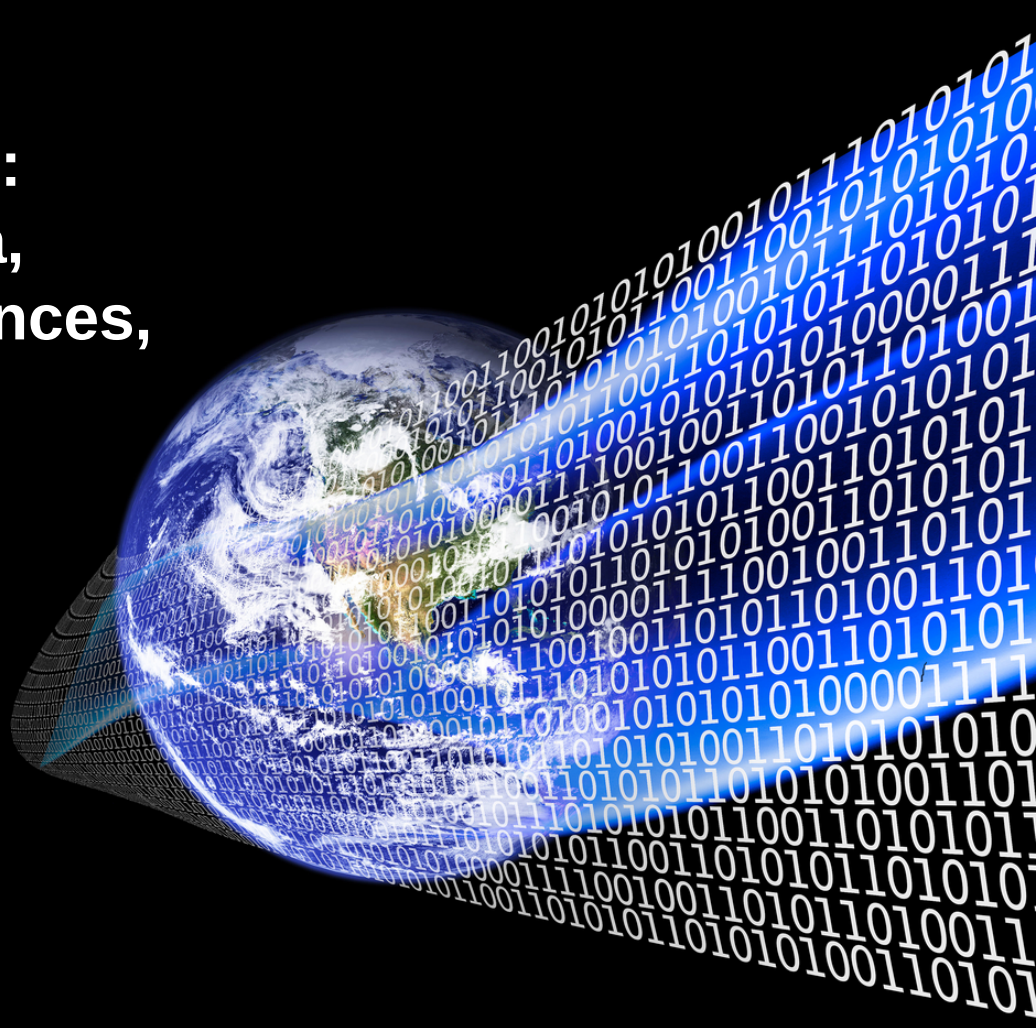
e.g. in XES format



TU/e

The challenge is not the syntactical conversion, but:

- **locating the relevant data,**
- **identifying process instances,**
- **scoping,**
- **…**



A typical SAP installation has tens of thousands of tables …

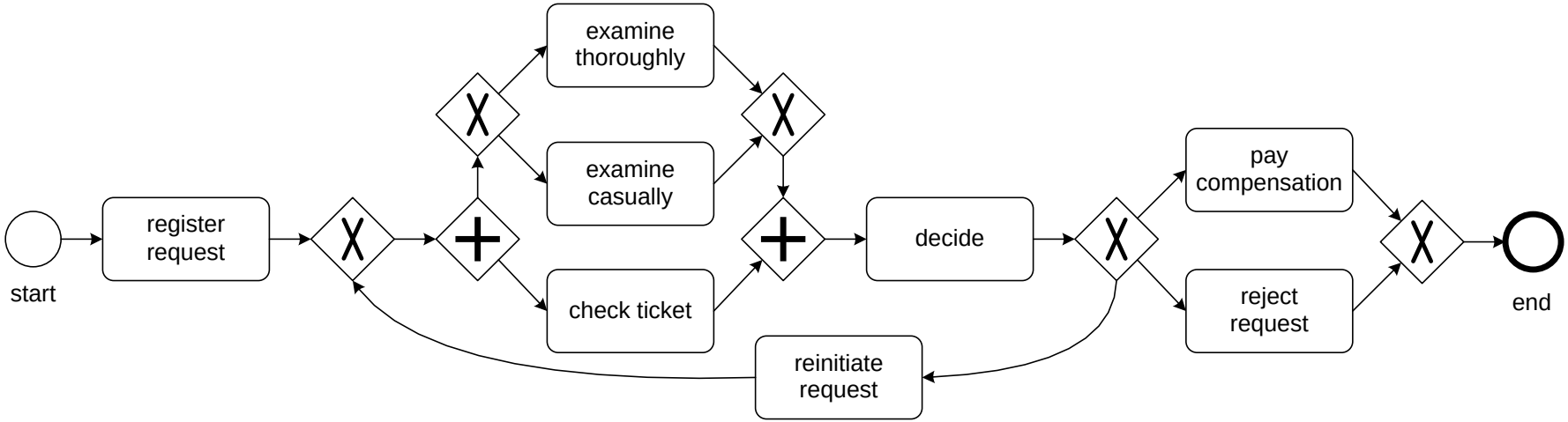
Need to understand the underlying data model to scope and select event data!

In a hospital one can easily find one thousand patient-related tables …

Flatten event data

case id	event id	properties				
		timestamp	activity	resource	cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
	35654427	07-01-2011:14.24	reject request	Pete	200	...
2	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket	Mike	100	...
	35654487	30-12-2010:14.16	examine casually	Pete	400	...
	35654488	05-01-2011:11.22	decide	Sara	200	...
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	...
3	35654521	30-12-2010:14.32	register request	Pete	50	...
	35654522	30-12-2010:15.06	examine casually	Mike	400	...
	35654524	30-12-2010:16.34	check ticket	Ellen	100	...

Lifecycle of process instance (case)

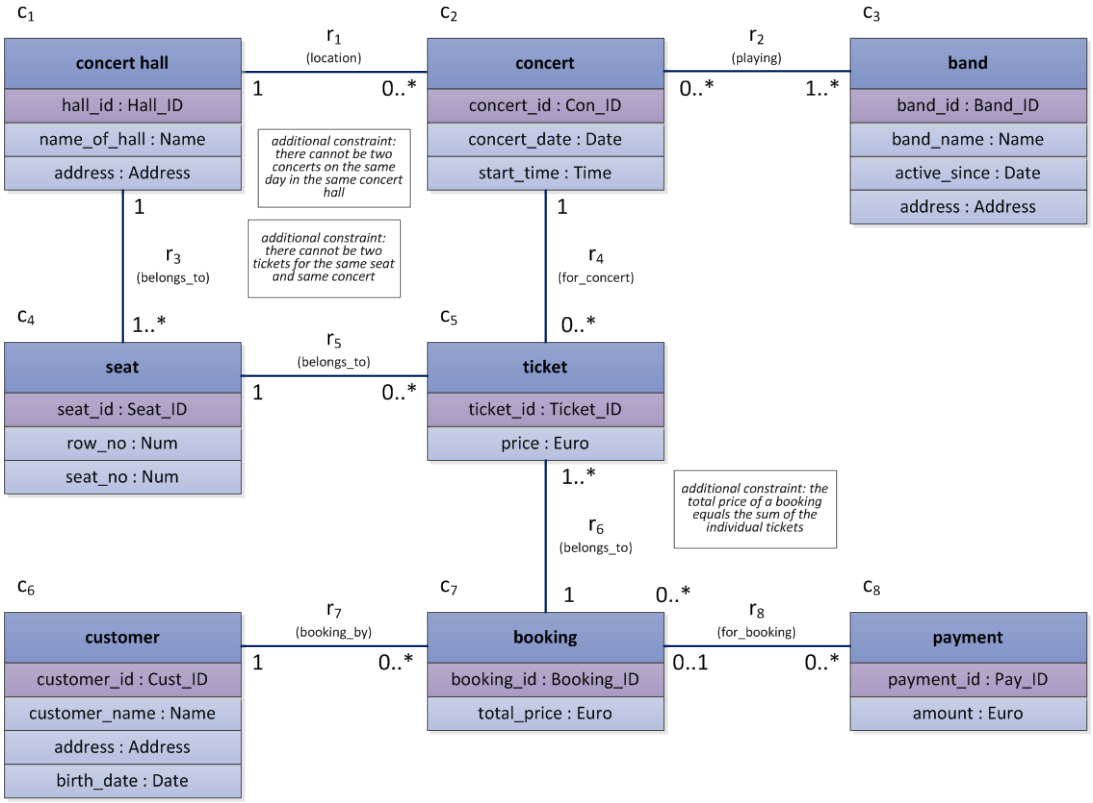


All mainstream notations model the lifecycle of an instance in isolation.

Notable exceptions: proclets and other artifact-centric models.

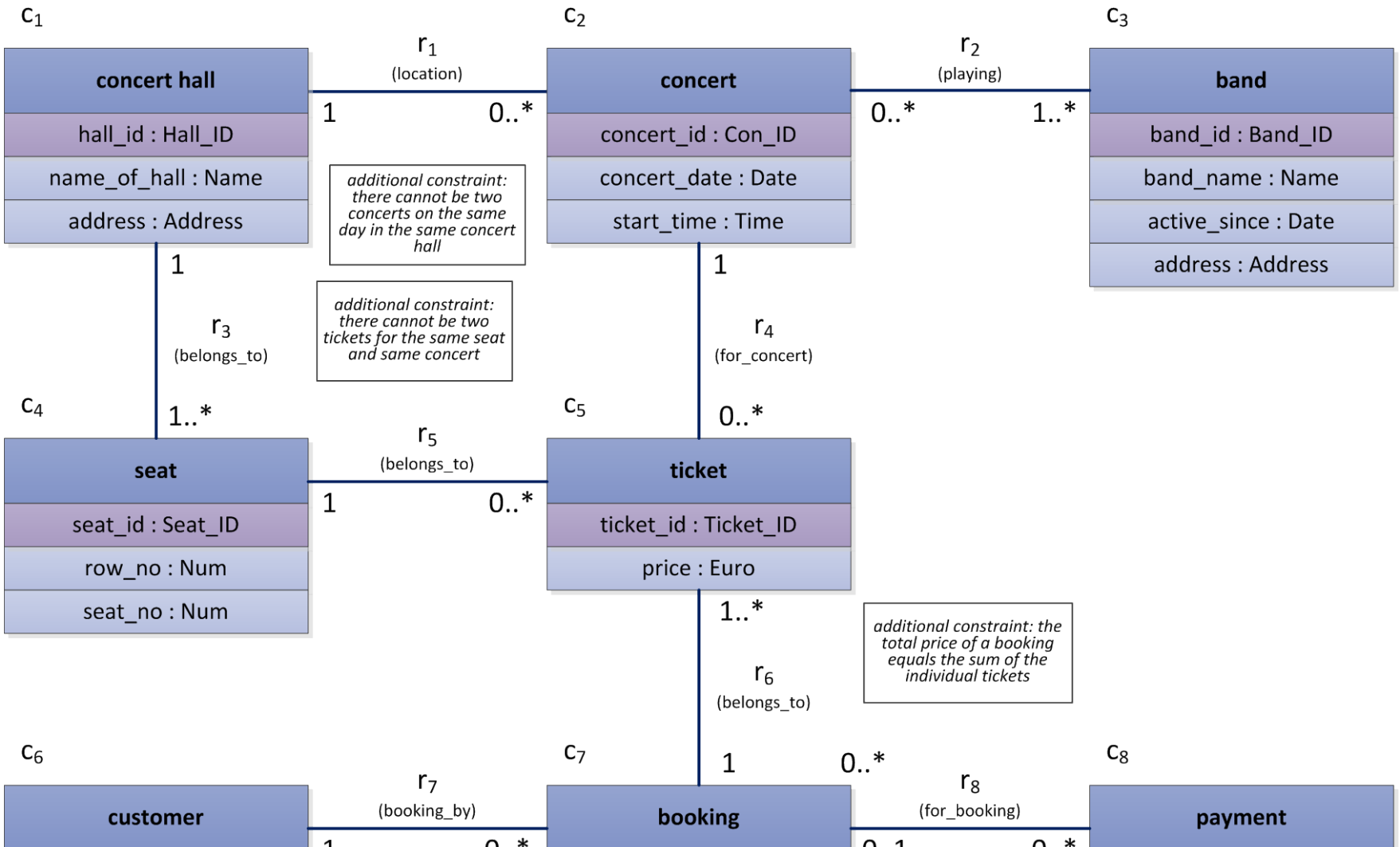
TU/e

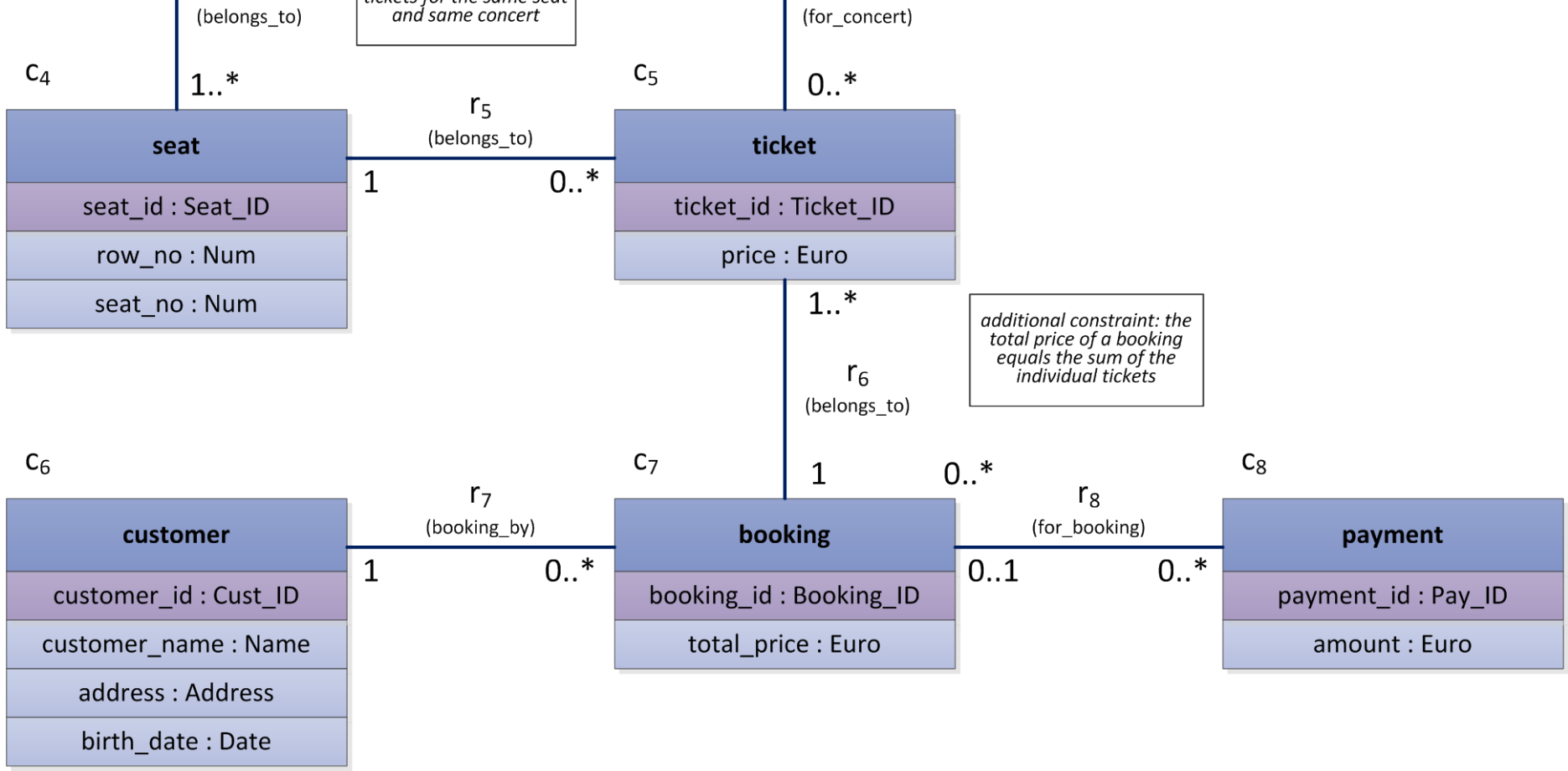
Consider a database for booking tickets



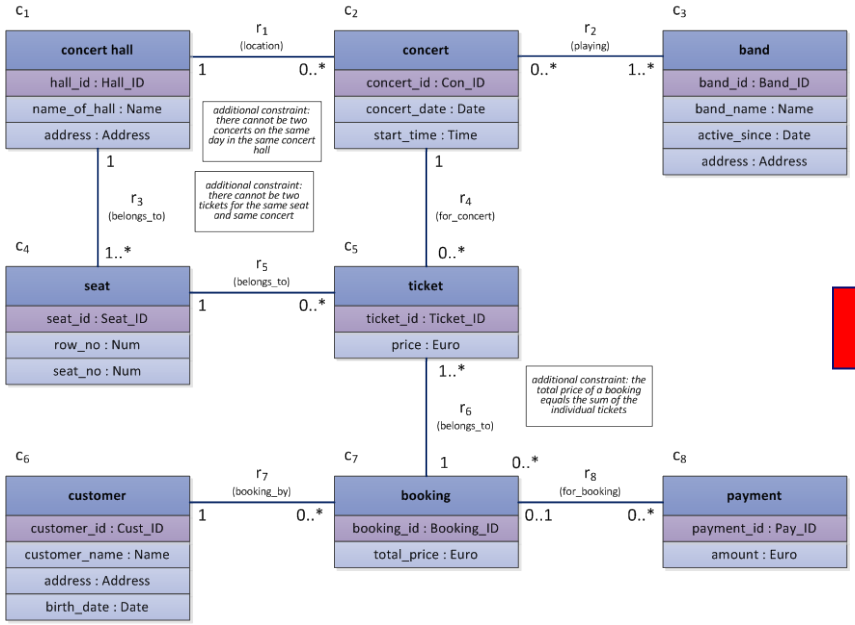
Assume all insertions and updates are recorded in a so-called redo log.

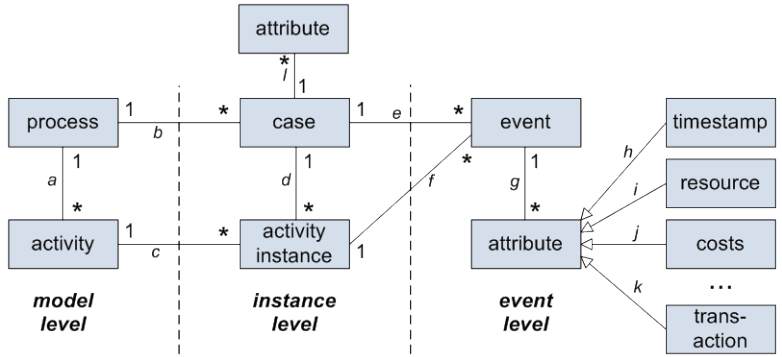
TU/e


c1
concert hall
hall_id : Hall_ID
name_of_hall : Name
address : Address
r1
(location)
1
0..*
c2
concert
concert_id : Con_ID
concert_date : Date
start_time : Time
r2
(playing)
0..*
1..*
c3
band
band_id : Band_ID
band_name : Name
active_since : Date
address : Address
additional constraint: there cannot be two concerts on the same day in the same concert hall
additional constraint: there cannot be two tickets for the same seat and same concert
1
r3
(belongs_to)
1..*
1
r4
(for_concert)
0..*
c4
seat
seat_id : Seat_ID
row_no : Num
seat_no : Num
c5
ticket
ticket_id : Ticket_ID
price : Euro
r5
(belongs_to)
1
0..*
1..*
r6
(belongs_to)
additional constraint: the total price of a booking equals the sum of the individual tickets
c6
customer
c7
booking
c8
payment
r7
(booking_by)
1
0..*
1
0..*
r8
(for_booking)
0..1
0..*

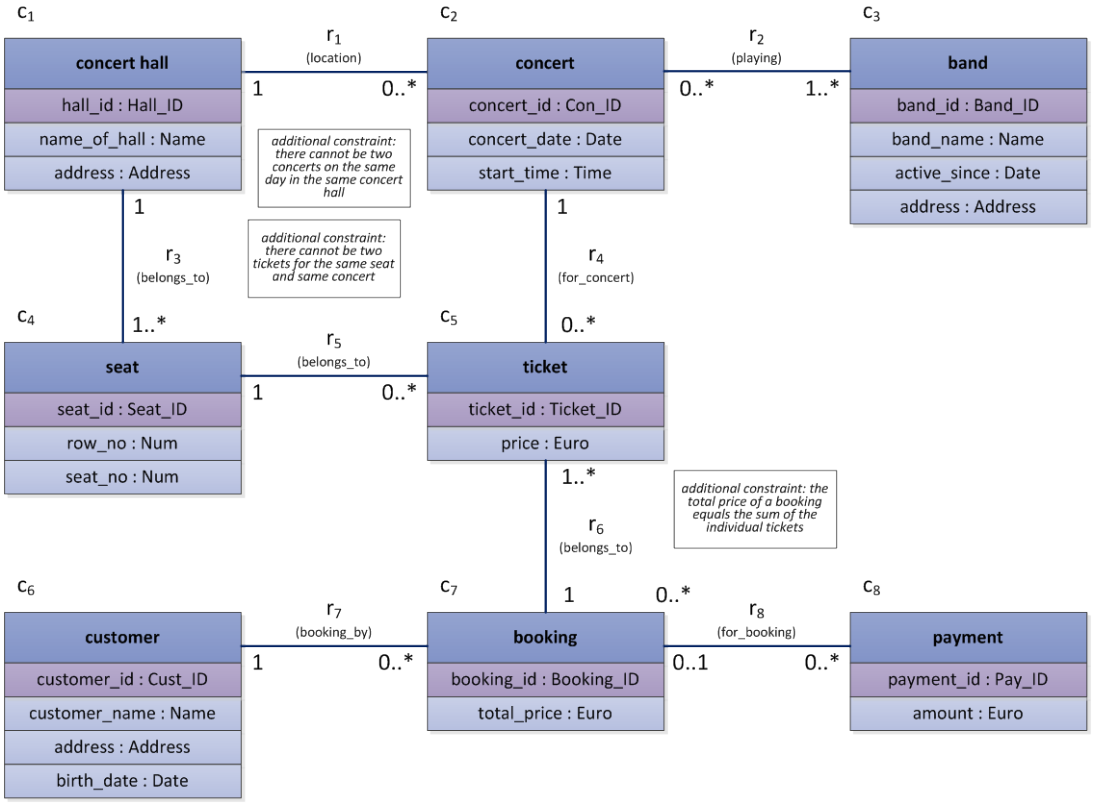

(belongs_to)
tickets for the same seat and same concert
(for_concert)
c4
1..*
r5
(belongs_to)
c5
0..*
seat
seat_id : Seat_ID
row_no : Num
seat_no : Num
1
0..*
ticket
ticket_id : Ticket_ID
price : Euro
1..*
additional constraint: the total price of a booking equals the sum of the individual tickets
r6
(belongs_to)
c6
c7
c8
customer
customer_id : Cust_ID
customer_name : Name
address : Address
birth_date : Date
r7
(booking_by)
1
0..*
1
0..*
booking
booking_id : Booking_ID
total_price : Euro
r8
(for_booking)
0..1
0..*
payment
payment_id : Pay_ID
amount : Euro

Mapping needed …



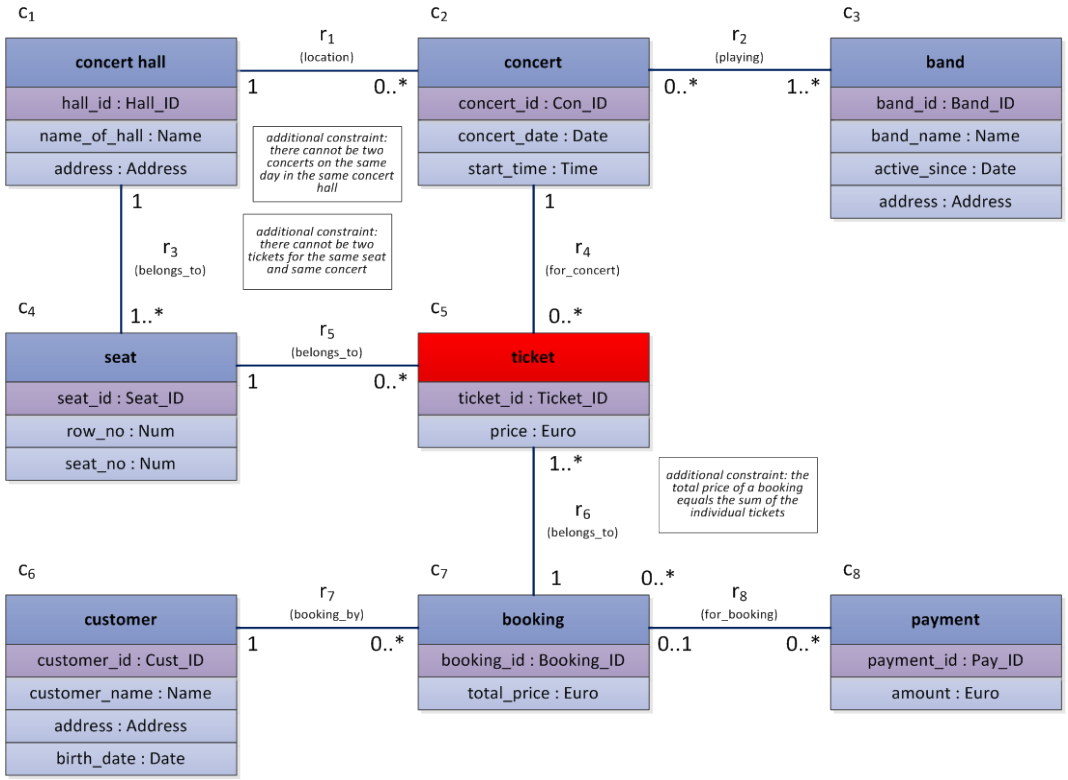


What is the process instance?



- Ticket?
- Booking?
- Band?
- Concert?
- Concert hall?
- Seat?
- Customer?
- Payment?

Lifecycle of ticket?

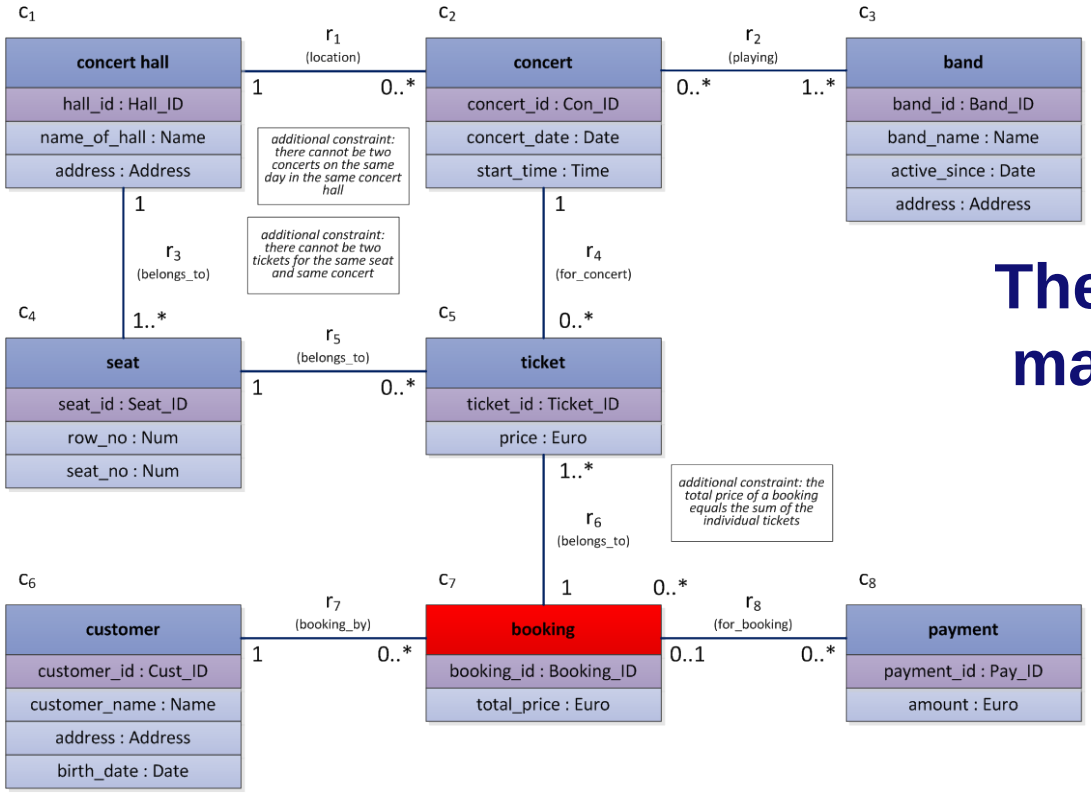


What are the activities?

Multiple process instances may share the same booking or payment event.

TU/e

Lifecycle of booking?



What are the activities?

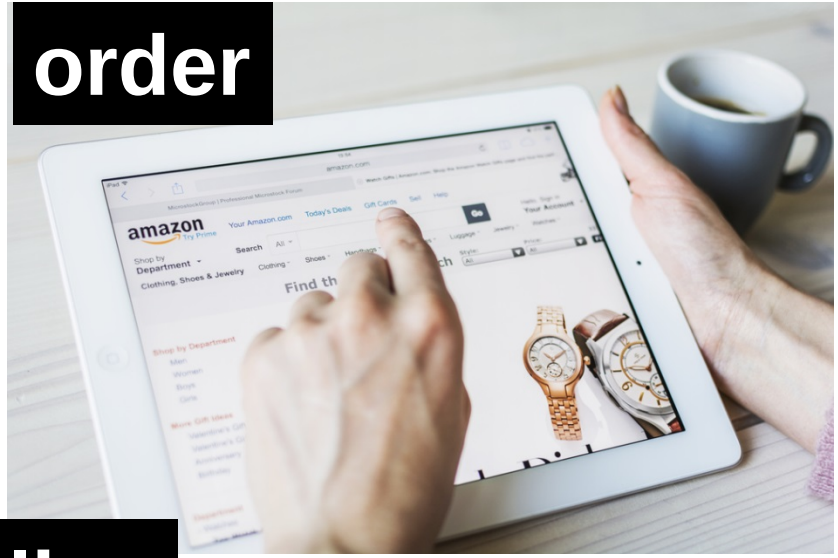
The same booking instance may have multiple ticket or payment related events.

Cancellation of a concert may impact many bookings.

TU/e

Ordering books from Amazon

order



delivery

orderline

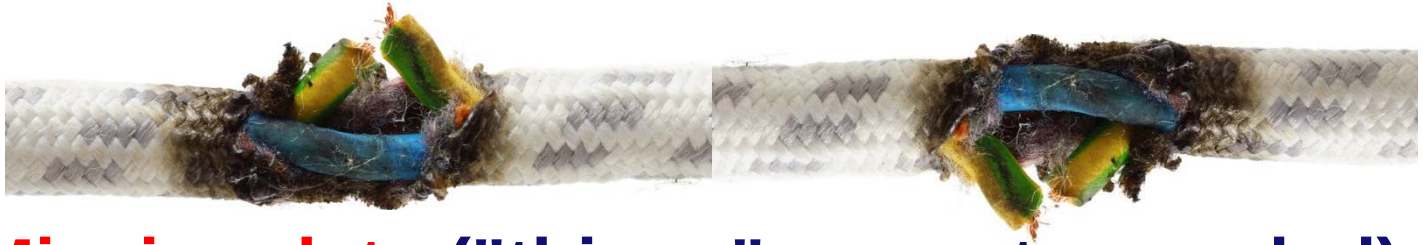
What are the process instances and corresponding activities?

TU/e

Next to selection and mapping problems …

… there may be data quality issues!

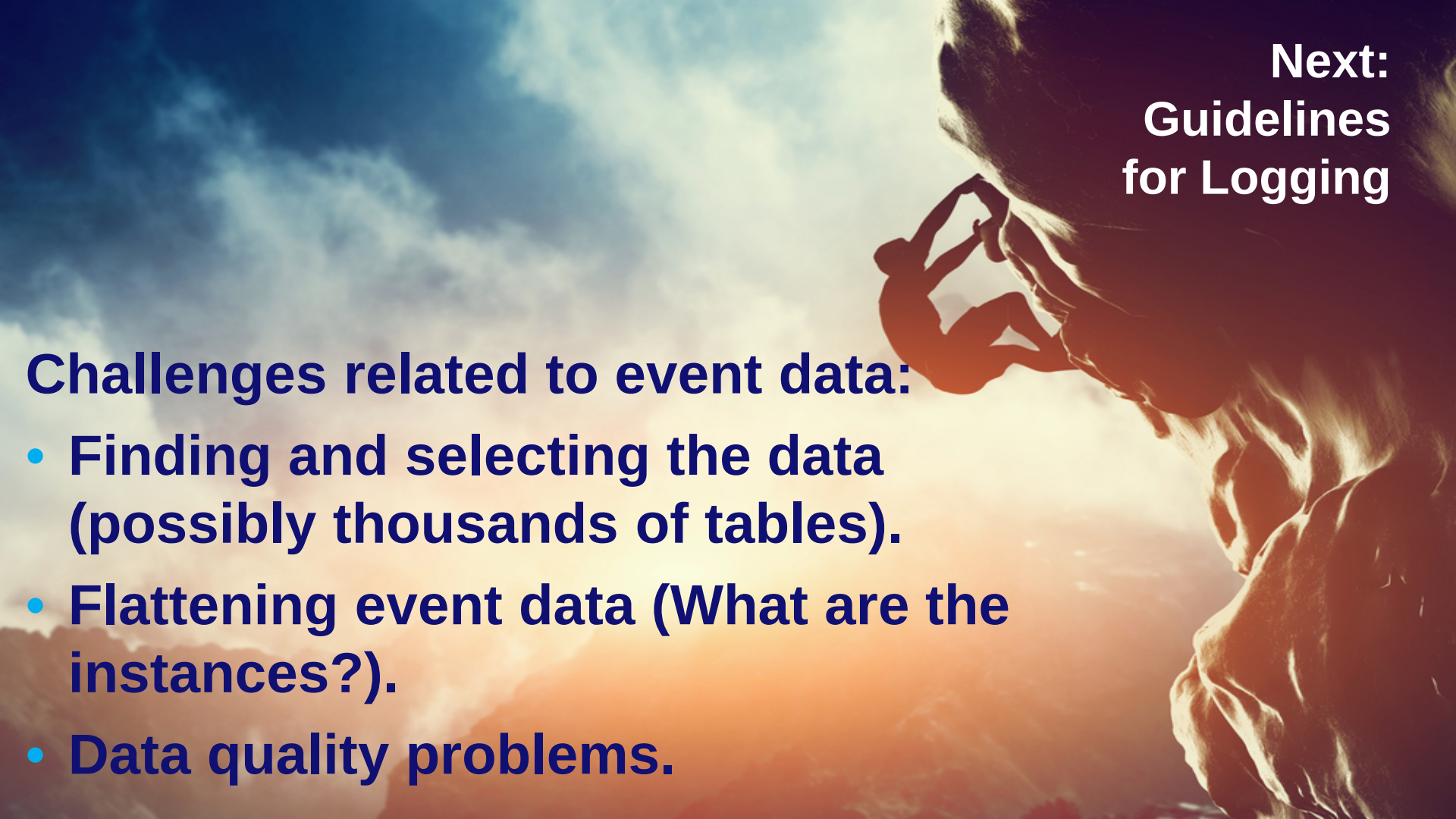
Data quality problems



- **Missing data** ("things" are not recorded).
- **Incorrect data** ("things" are recorded incorrectly).
- **Imprecise data** ("things" are not recorded at the desired level of granularity).
- **Irrelevant data** (relevant "things" are submerged in poorly structured data).

TU/e

	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 correspondin g 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							



Next: Guidelines for Logging

Challenges related to event data:

- **Finding and selecting the data (possibly thousands of tables).**
- **Flattening event data (What are the instances?).**
- **Data quality problems.**

Part I: Preliminaries

Chapter 1
Introduction

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

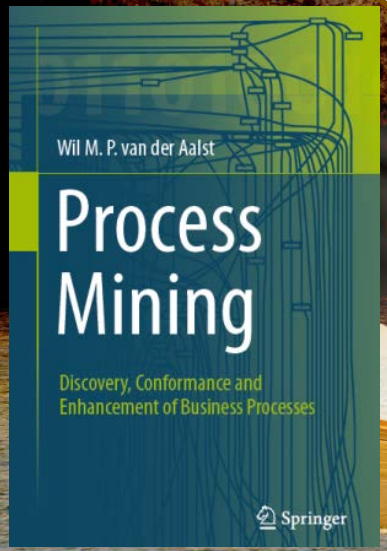
Chapter 11
Analyzing "Lasagna Processes"

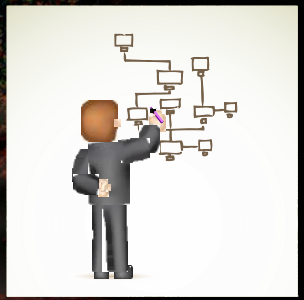
Chapter 12
Analyzing "Spaghetti Processes"

Part V: Reflection

Chapter 13
Cartography and Navigation

Chapter 14
Epilogue





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

TU/e