TECHNISCHE UNIVERSITÄT DRESDEN

CHAIR OF BUSINESS INFORMATION SYSTEMS
ESP. INTELLIGENT SYSTEMS AND SERVICES

DRESDEN concept

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Data Science: Advanced Analytics
Process Analytics

Process Warehouse and Process Mining

Process Analytics and Process Mining

"Process mining aims to **exploit event data** in a meaningful way"

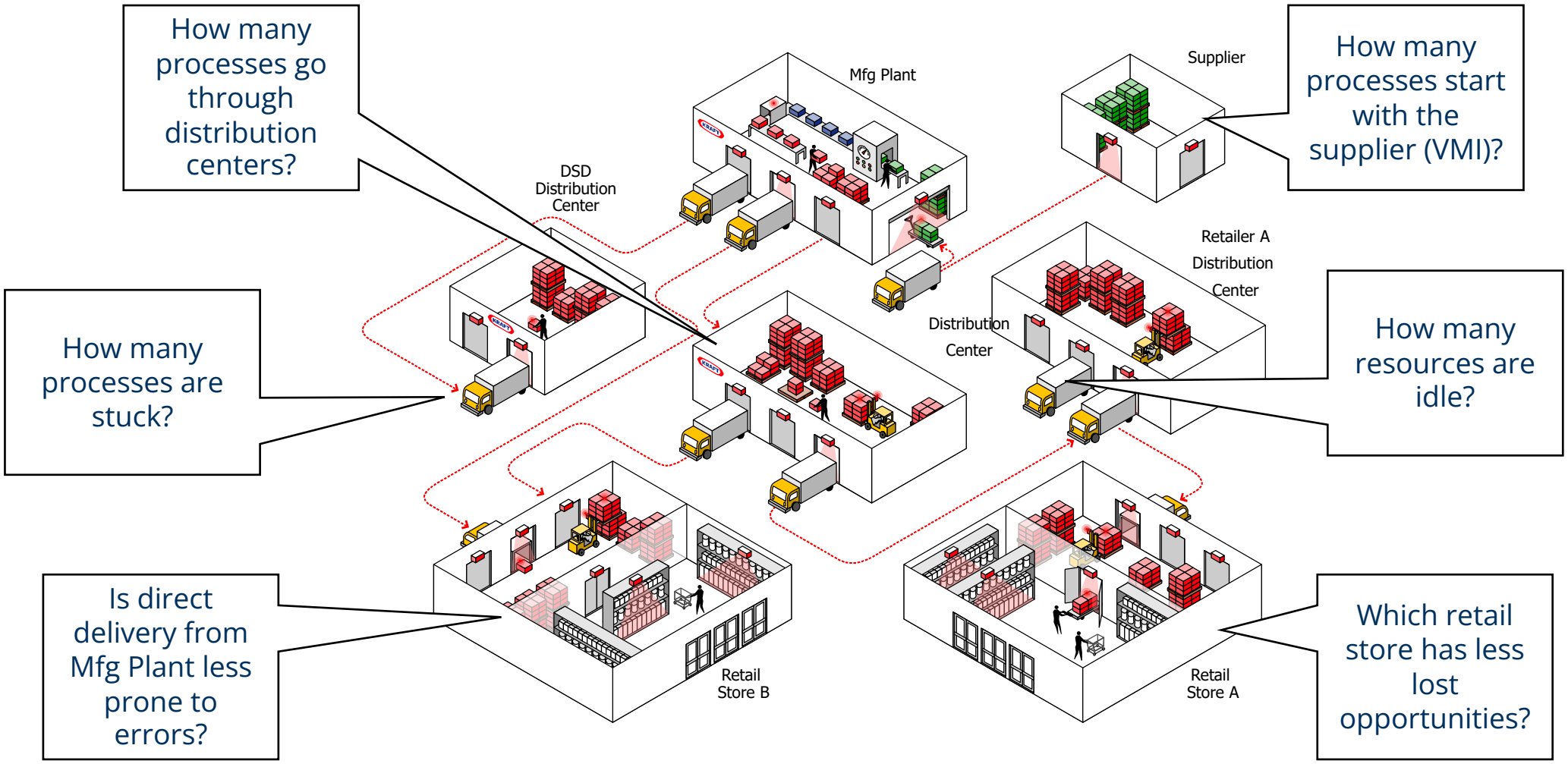
"Process mining techniques use event data to **discover** processes, check **compliance**, analyze **bottlenecks**, compare process **variants**, and suggest improvements"

"Process mining adds the **process perspective** to machine learning and data mining"

"Process mining brings together **traditional model-based process analysis** and **data-centric analysis techniques**"

van der Aalst (2017)

Analyzing Processes



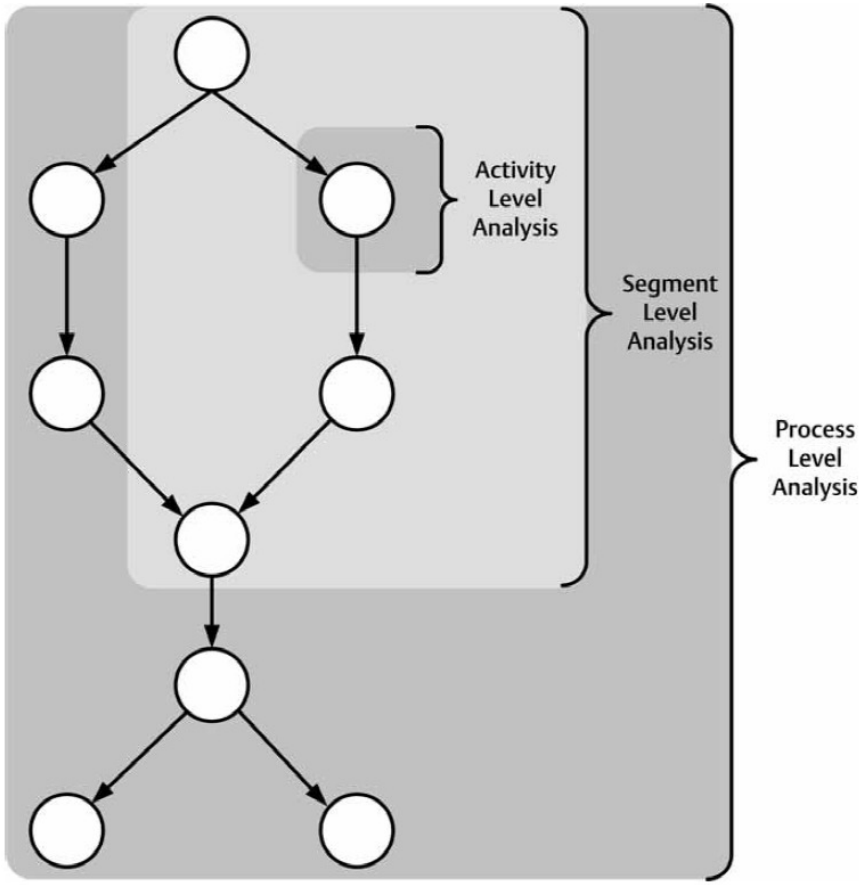
SAP (2008)

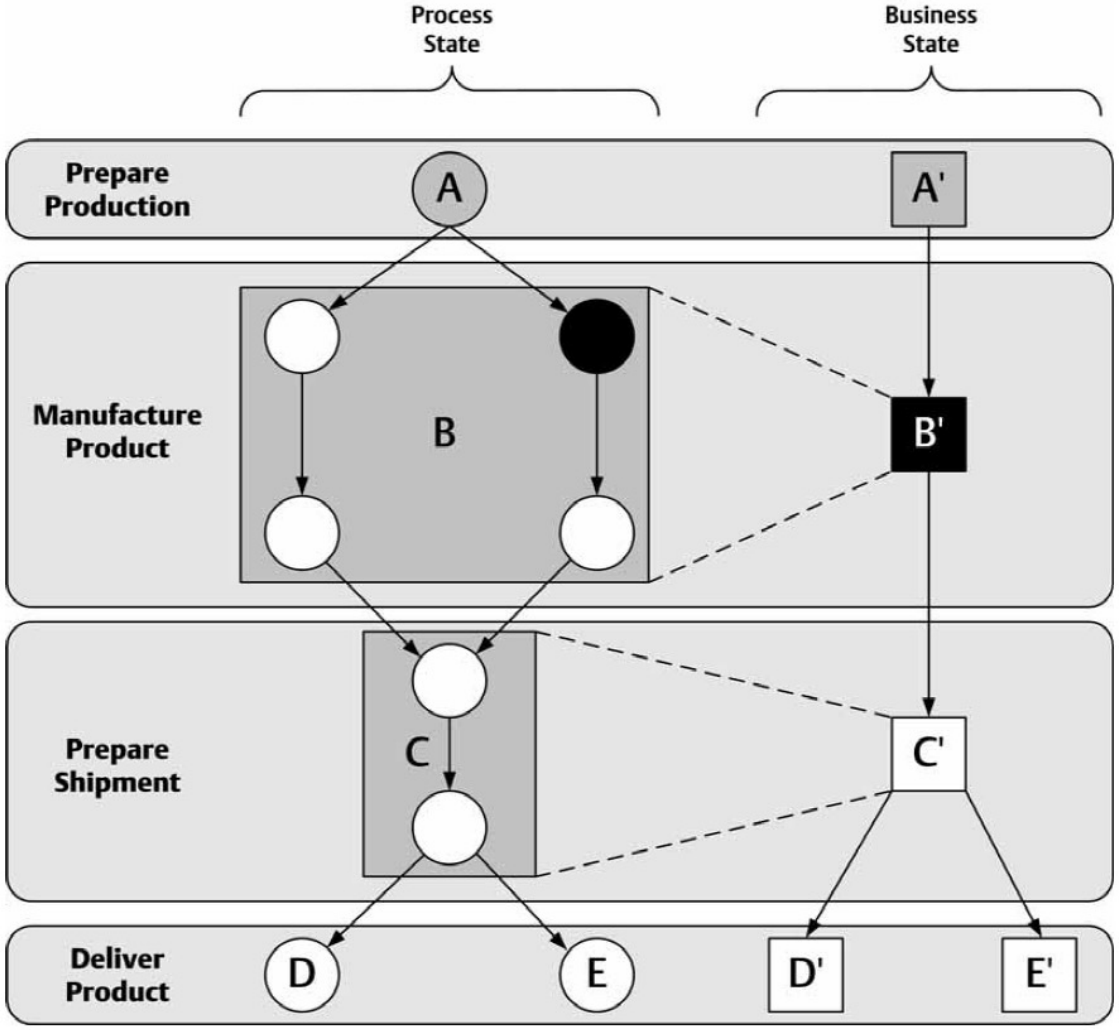
Event Logs

Case ID	Event ID	Timestamp	Activity	Resource
1	Ch-4680555556-1	2012-07-30 11:14	Check stock availability	SYS1
1	Re-5972222222-1	2012-07-30 14:20	Retrieve product from warehouse	Rick
1	Co-6319444444-1	2012-07-30 15:10	Confirm order	Chuck
1	Ge-6402777778-1	2012-07-30 15:22	Get shipping address	SYS2
1	Em-6555555556-1	2012-07-30 15:44	Emit invoice	SYS2
1	Re-4180555556-1	2012-08-04 10:02	Receive payment	SYS2
1	Sh-4659722222-1	2012-08-05 11:11	Ship product	Susi
1	Ar-3833333333-1	2012-08-06 09:12	Archive order	DMS
2	Ch-4055555556-2	2012-08-01 09:44	Check stock availability	SYS1
2	Ch-4208333333-2	2012-08-01 10:06	Check materials availability	SYS1
2	Re-4666666667-2	2012-08-01 11:12	Request raw materials	Ringo
2	Ob-3263888889-2	2012-08-03 07:50	Obtain raw materials	Olaf
2	Ma-6131944444-2	2012-08-04 14:43	Manufacture product	SYS1
2	Co-6187615741-2	2012-08-04 14:51	Confirm order	Conny
2	Em-6388888889-2	2012-08-04 15:20	Emit invoice	SYS2
2	Ge-6439814815-2	2012-08-04 15:27	Get shipping address	SYS2
2	Sh-7277777778-2	2012-08-04 17:28	Ship product	Sara
2	Re-3611111111-2	2012-08-07 08:40	Receive payment	SYS2
2	Ar-3680555556-2	2012-08-07 08:50	Archive order	DMS
3	Ch-4208333333-3	2012-08-02 10:06	Check stock availability	SYS1
3	Ch-4243055556-3	2012-08-02 10:11	Check materials availability	SYS1
3	Ma-6694444444-3	2012-08-02 16:04	Manufacture product	SYS1
3	Co-6751157407-3	2012-08-02 16:12	Confirm order	Chuck

Dumas et al. (2018)

Process Analytics Dimensions





zur Mühlen (2004)

Process Analytics Dimensions

Attribute	Possible Values					
Focus	Technical Information	Business Information				Presentation
Presentation	Active	Passive				
Timeframe	Running Processes	Completed Processes				
Aggregation	Single Instance	Multiple Instances				
Data Scope	Process	Process + Business Objects	Enterprise			Information
Object	Event	Activity	Process	Resource	Business Object	
Process Scope	Activity	Segment	Process	Process Chain		

zur Mühlen (2004)

Process Mining Challenges

Correlation challenge

- Identify the case an event belongs to

Timestamps challenge

- Logging is often delayed until the system has idle time: sequential events with the same timestamp, logs from different Business Process Management Software (BPMS)

Longevity challenge

- Long running processes might be too slow for snapshot window

Scoping challenge

- IS does not directly produce event logs. Logs are synthesized

Granularity challenge

- Abstraction of model might be different from log

Dumas et al. (2013, 2018)

Process Analytics Questions

What is the **performance** of the process?

What **is** the actual process model?

To which extent are the **rules** of the process model **followed**?

TECHNISCHE UNIVERSITÄT DRESDEN

CHAIR OF BUSINESS INFORMATION SYSTEMS
ESP. INTELLIGENT SYSTEMS AND SERVICES

DRESDEN concept

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Process Analytics

Process Warehouse

Process Warehouse Functionality

"A data warehouse that caters for the need of process analytics"

Collection and integration of **historic and current data** for **predefined** process performance indicators

Transformation, **calculation** and **aggregation** of process performance indicators

Provisioning of **tools** for **process-based** multidimensional **analysis** and **navigation**

Distribution and **presentation** of analysis results

Kueng et al. (2001), List et al. (2002), Becker et al. (2006)

Process Warehouse Challenges

Needs data from **many sources** (many of which are **not databases**)

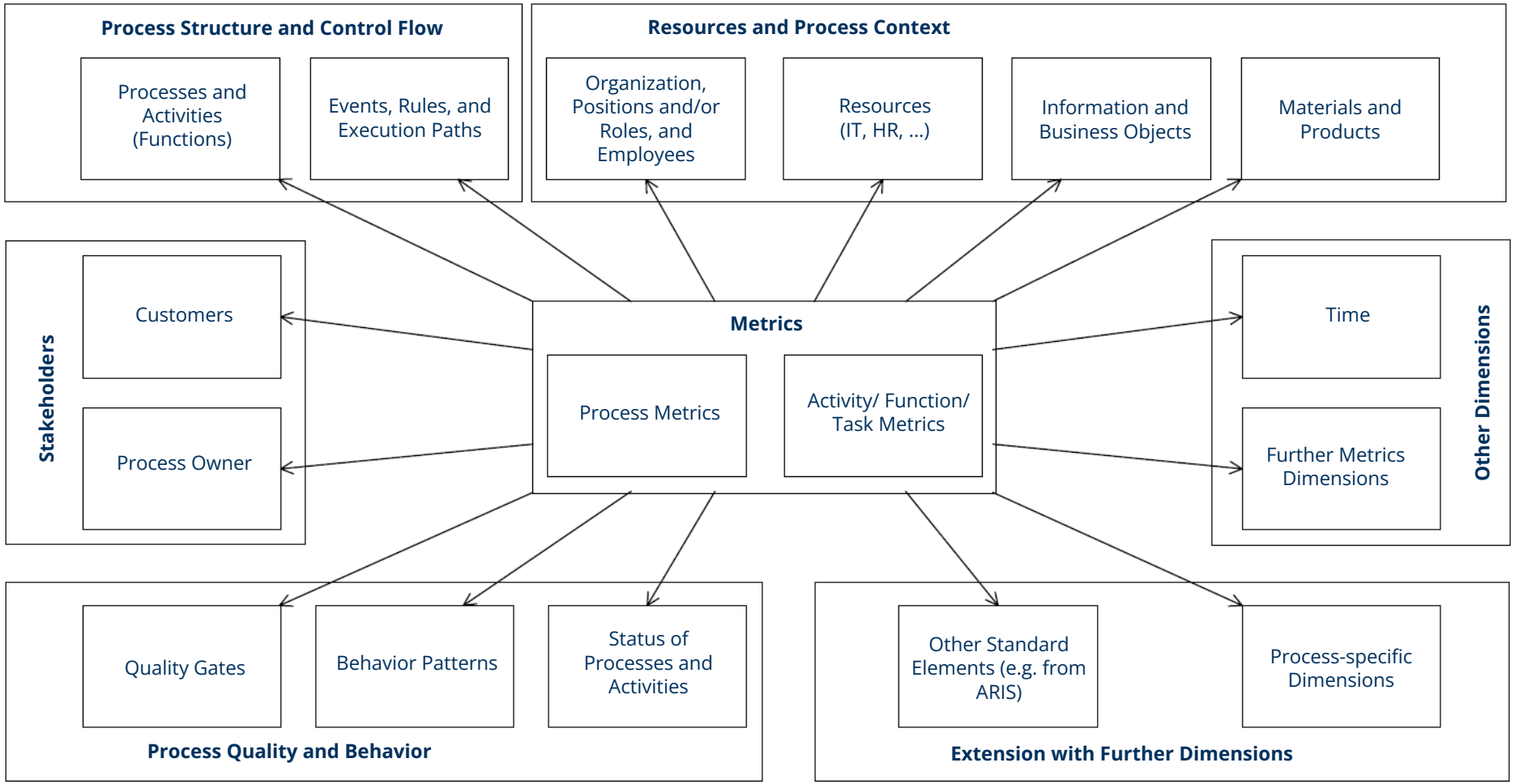
Needs **historical data** in addition to data about **current process instances**

Needs complex transformations (e.g., **map system events into abstract process progression** – but structures are complex and may contain loops)

Many reporting and analytic tools already work with data warehouses

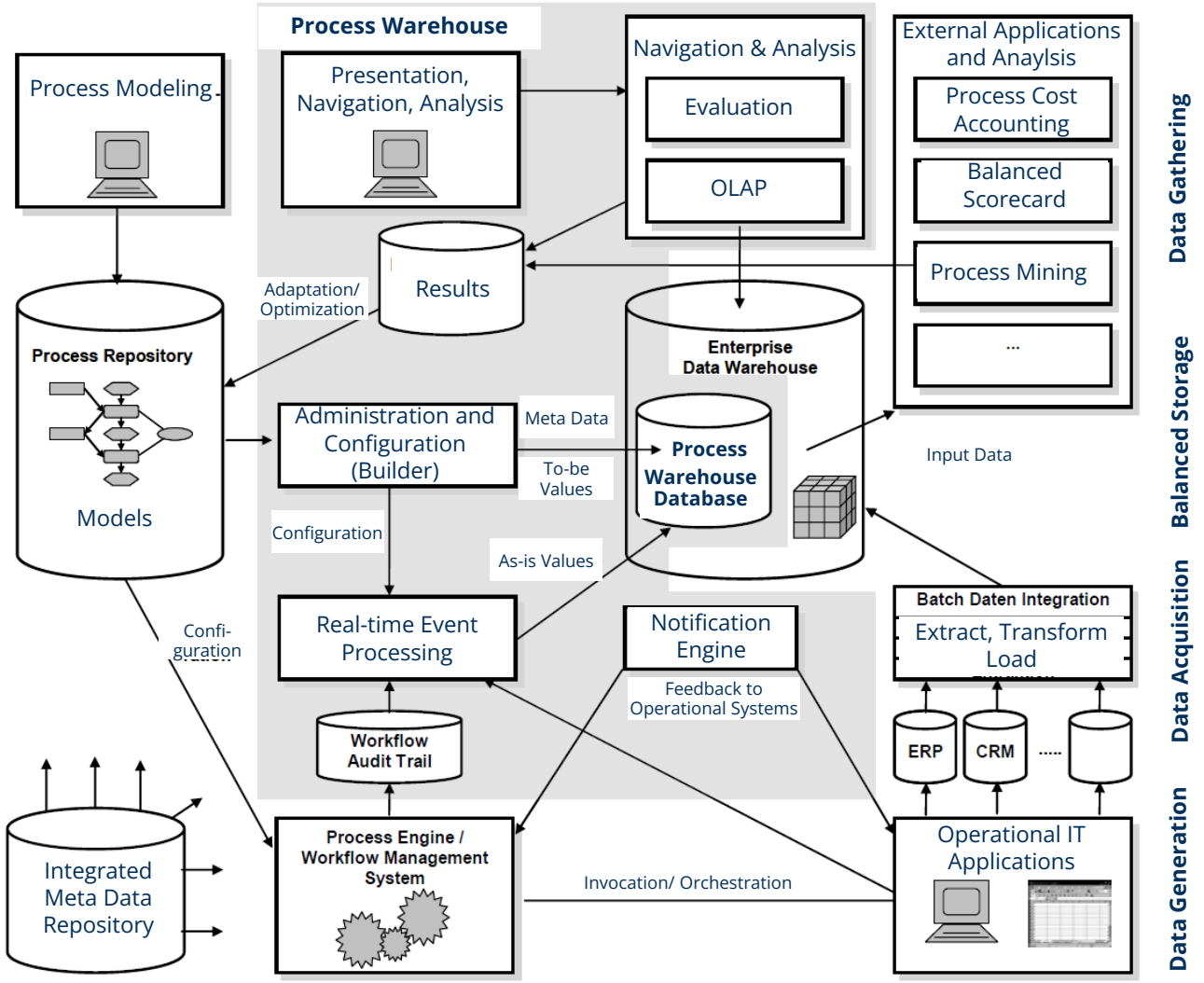
Casati et al. (2017)

Process Warehouse Reference Dimensions



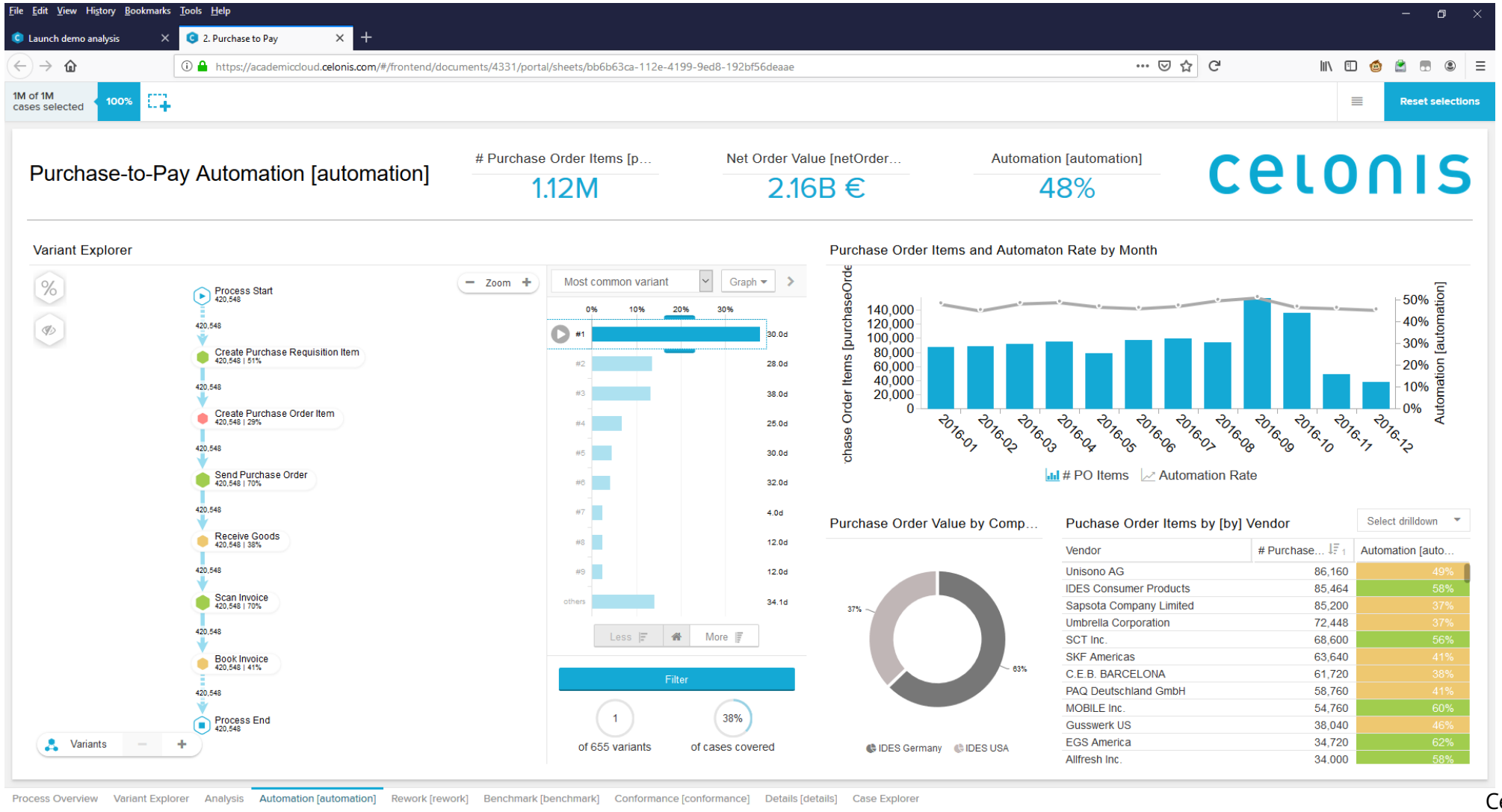
Becker (2007), Becker, Chamoni (2008)

Process Warehouse Reference Architecture



Becker (2007), Becker, Chamoni (2008)

Example: Celonis Viewer & Analyst



Celonis SE (2019)

Process Analytics Questions

What is the **performance** of the process? 

What **is** the actual process model? 

To which extent are the **rules** of the process model **followed**? 

TECHNISCHE UNIVERSITÄT DRESDEN

CHAIR OF BUSINESS INFORMATION SYSTEMS
ESP. INTELLIGENT SYSTEMS AND SERVICES

DRESDEN concept

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Process Analytics

Process Discovery

Automatic Process Discovery

Goal

– construct a process model that **captures the behavior** of an event log in a representative way

Construction

– **automatically and generically** using algorithms
– should make minimal assumptions about properties of the log and the resulting process model

Result

– constructed process model should be able to **replay** the cases of the event log and
– **forbid behavior** not found in the logs

Dumas et al. (2018)

α-algorithm

The *α*-algorithm is a **basic** algorithm for discovering process models from event logs

It is less complex than other advanced algorithms

It will not be sufficient for practical application

Idea is to identify the relations between **all pairs of tasks** from the workflow log to construct a process model

Dumas et al. (2018), van der Aalst (2017)

Assumptions

Order of Events

– The events in the log are chronologically ordered (e.g. timestamps)

Case Reference

– Each event refers to a single case

Activity Reference

– Each event relates to a specific activity (task) of the process

Activity Completeness

– Each activity of the process is included in the log

Behavioral Completeness

– If *b* can follow *a*, then there is at least one case where one can observe *ab*

Dumas et al. (2013)

Preparation of Workflow Log

Case ID	Event ID	Timestamp	Activity
1	Ch-468	2012-07-30 11:14	Check stock availability
1	Re-597	2012-07-30 14:20	Retrieve product from warehouse
1	Co-631	2012-07-30 15:10	Confirm order
1	Ge-640	2012-07-30 15:22	Get shipping address
1	Em-655	2012-07-30 15:44	Emit invoice
1	Re-418	2012-08-04 10:02	Receive payment
1	Sh-465	2012-08-05 11:11	Ship product
1	Ar-383	2012-08-06 09:12	Archive order
2	Ch-405	2012-08-01 09:44	Check stock availability
2	Ch-420	2012-08-01 10:06	Check materials availability
2	Re-466	2012-08-01 11:12	Request raw materials
2	Ob-326	2012-08-03 07:50	Obtain raw materials
2	Ma-613	2012-08-04 14:43	Manufacture product
2	Co-618	2012-08-04 14:51	Confirm order
2	Em-638	2012-08-04 15:20	Emit invoice
2	Ge-643	2012-08-04 15:27	Get shipping address
2	Sh-727	2012-08-04 17:28	Ship product
2	Re-361	2012-08-07 08:40	Receive payment
2	Ar-368	2012-08-07 08:50	Archive order

Letter	Actvities
a	Check stock availability
b	Retrieve product from warehouse
c	Check materials availability
d	Request raw materials
e	Obtain raw materials
f	Manufacture product
g	Confirm order
h	Get shipping address
i	Ship product
j	Emit invoice
k	Receive payment
l	Archive order

Workflow Log

a,b,g,h,j,k,i,l

a,c,d,e,f,g,j,h,i,k,l

...

Dumas et al. (2018)

α-Algorithm Phases

Phase 1

– different **order relations are extracted** from the workflow log

Phase 2

– the **process model is constructed** in a stepwise fashion from these identified relations

Dumas et al. (2018)

α-Algorithm Order Relations

Basic ($a > b$)

Task a is directly followed by b

Basis for the definition of the other three order relations

Causality ($a \rightarrow b$)

$a > b$ and that $b \not> a$

Parallelism ($a \parallel b$)

$a > b$ and $b > a$

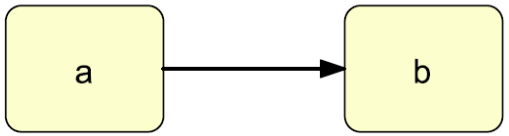
Non-succession ($a \# b$)

$a \not> b$ and $b \not> a$

Dumas et al. (2018)

Order Relation Examples

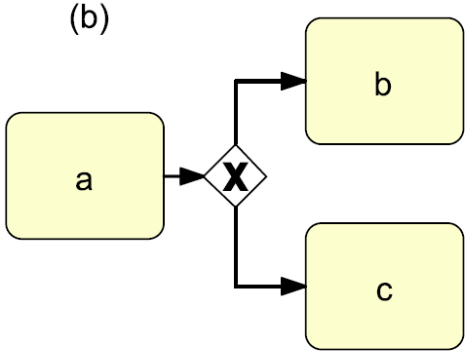
(a)



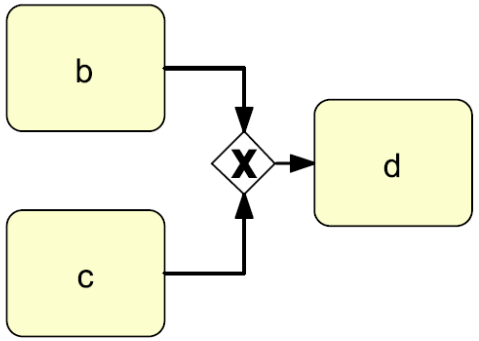
Causality ($a \rightarrow b$)

Causality ($a \rightarrow b$)
Causality ($a \rightarrow c$)
Non-succession ($b \# c$)

(b)



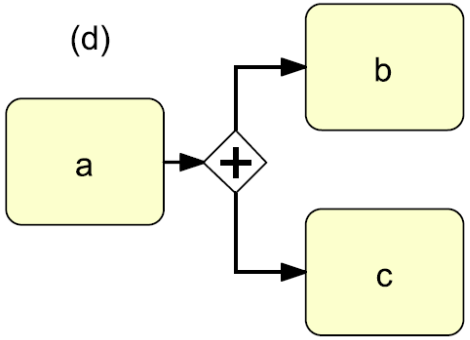
(c)



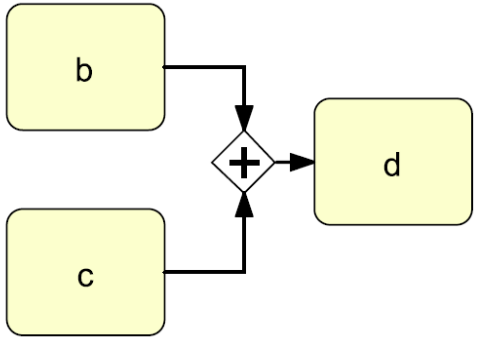
Causality ($b \rightarrow d$)
Causality ($c \rightarrow d$)
Non-succession ($b \# c$)

Causality ($a \rightarrow b$)
Causality ($a \rightarrow c$)
Non-succession ($b \# c$)
Parallelism ($b \parallel c$)

(d)



(e)



Causality ($b \rightarrow d$)
Causality ($c \rightarrow d$)
Non-succession ($b \# c$)
Parallelism ($b \parallel c$)

Dumas et al. (2018)

Extract Order Relations

Traces (a, b, g, h, j, k, i, l) & (a, c, d, e, f, g, j, h, i, k, l)

Basic relations

$a > c$

$c > d$

	a	b	c	d	e	f	g	h	i	j	k	l
a	#	→	→	#	#	#	#	#	#	#	#	#
b	←	#	#	#	#	#	→	#	#	#	#	#
c	←	#	#	→	#	#	#	#	#	#	#	#
d	#	#	←	#	→	#	#	#	#	#	#	#
e	#	#	#	←	#	→	#	#	#	#	#	#
f	#	#	#	#	←	#	→	#	#	#	#	#
g	#	←	#	#	#	←	#	→	#	→	#	#
h	#	#	#	#	#	#	←	#	→	‖	#	#
i	#	#	#	#	#	#	#	←	#	#	‖	→
j	#	#	#	#	#	#	←	‖	#	#	→	#
k	#	#	#	#	#	#	#	#	‖	←	#	→
l	#	#	#	#	#	#	#	#	←	#	←	#

k > i

i > l

i > k

k > l

Parallelism

$h \parallel j$

$k \parallel i$

Causality

$a \rightarrow b$

$b \rightarrow g$

$g \rightarrow h$

$j \rightarrow k$

$i \rightarrow l$

$a \rightarrow c$

$c \rightarrow d$

$d \rightarrow e$

$e \rightarrow f$

$f \rightarrow g$

$g \rightarrow j$

$h \rightarrow l$

$k \rightarrow l$

Non-sucession

all other relations

Dumas et al. (2018)

Process Footprint Matrix

	a	b	c	d	e	f	g	h	i	j	k	l
a	#	→	→	#	#	#	#	#	#	#	#	#
b	←	#	#	#	#	#	→	#	#	#	#	#
c	←	#	#	→	#	#	#	#	#	#	#	#
d	#	#	←	#	→	#	#	#	#	#	#	#
e	#	#	#	←	#	→	#	#	#	#	#	#
f	#	#	#	#	←	#	→	#	#	#	#	#
g	#	←	#	#	#	←	#	→	#	→	#	#
h	#	#	#	#	#	#	←	#	→	‖	#	#
i	#	#	#	#	#	#	#	←	#	#	‖	→
j	#	#	#	#	#	#	←	‖	#	#	→	#
k	#	#	#	#	#	#	#	#	‖	←	#	→
l	#	#	#	#	#	#	#	#	←	#	←	#

Dumas et al. (2018)

α-Algorithm Steps (1)

1. Identify the **set of all tasks** in the log

2. Identify the set that has been observed as the **first task**

3. Identify the set that has been observed as the **last task**

4. Identify the **set of all connections** to be potentially represented in the process
    1. Pattern (a): all pairs for which holds $a \rightarrow b$
    2. Pattern (b): all triples for which holds $a \rightarrow (b \# c)$
    3. Pattern (c): all triples for which holds $(b \# c) \rightarrow d$

    Note that triples for which Pattern (d) $a \rightarrow (b \parallel c)$ or Pattern (e) $(b \parallel c) \rightarrow d$ hold are not included

Dumas et al. (2018)

α-Algorithm Steps (2)

5. Construct the a **subset** by
    1. Eliminating $a \rightarrow b$ and $a \rightarrow c$ if there exists some $a \rightarrow (b \# c)$
    2. Eliminating $b \rightarrow c$ and $b \rightarrow d$ if there exists some $(b \# c) \rightarrow d$

6. Connect **start and end events**
    1. If there are multiple first tasks, then draw a start event leading to a split (XOR or AND) which connects to all first tasks; otherwise, directly **connect the start event with the** only **first task**
    2. For each last task, add an end event and **draw an arc from the task to the end event**

Dumas et al. (2018)

α-Algorithm Steps (3)

7. Construct the **flow arcs** in the following way
    1. Pattern (a): For each $a \rightarrow b$, draw an arc a to b
    2. Pattern (b): For each $a \rightarrow (b \# c)$, draw an arc from a to an **XOR-split**, and from there to b and c
    3. Pattern (c): For each $(b \# c) \rightarrow d$, draw an arc from b and c to an **XOR-join**, and from there to d
    4. Pattern (d) and (e): If a task in the so constructed process model has multiple incoming or multiple outgoing arcs, bundle these arcs with an **AND-split** or **AND-join**, respectively

8. Return the newly constructed process model

Dumas et al. (2018)

Example Discovery

(a, b, g, h, j, k, i, l)

(a, c, d, e, f, g, j, h, i, k, l)

(1) All tasks *(a, b, c, d, e, f, g,h, i, j, k, l)*

(2) Start task ***a***

(3) End task ***l***

(4a) All relations (see footprint matrix)

(4b) Rows ***a→(b#c)*** as XOR-split

(4c) Columns ***(b#f)→g*** as XOR-join

(5) Remove ***a→b***, ***a→c***, ***b→g***, and ***f→g***

(6a+b) Connect start *a* and end event *l*

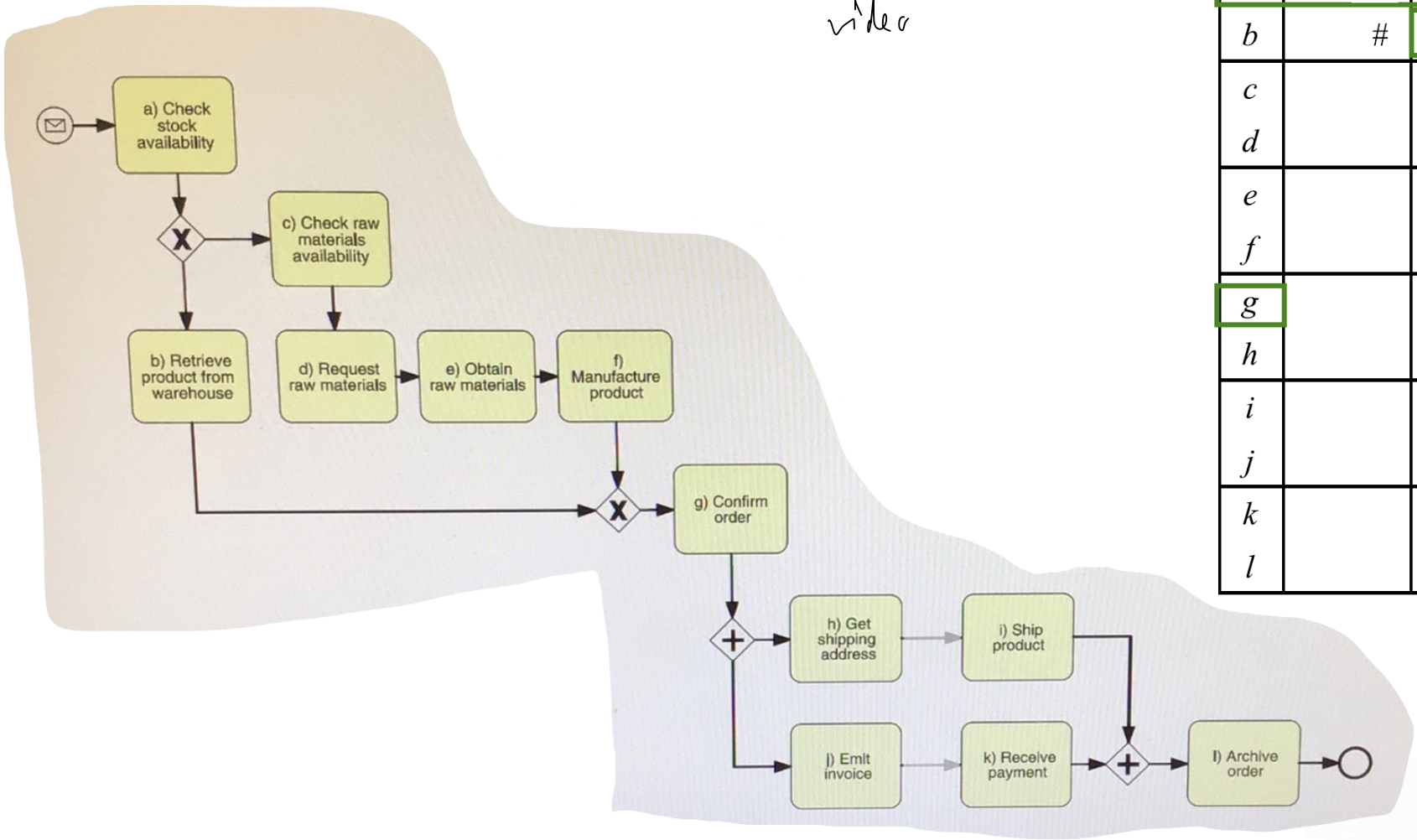
(7) Add arcs and AND-gateways

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*
a	#			#	#	#	#	#	#	#	#	#
b		#	#	#	#	#		#	#	#	#	#
c			#	→	#	#	#	#	#	#	#	#
d				#	→	#	#	#	#	#	#	#
e					#	→	#	#	#	#	#	#
f						#		#	#	#	#	#
g							#	→	#	→	#	#
h								#	→	‖	#	#
i									#	#	‖	→
j										#	→	#
k											#	→
l												#

Dumas et al. (2018)

Returned Model

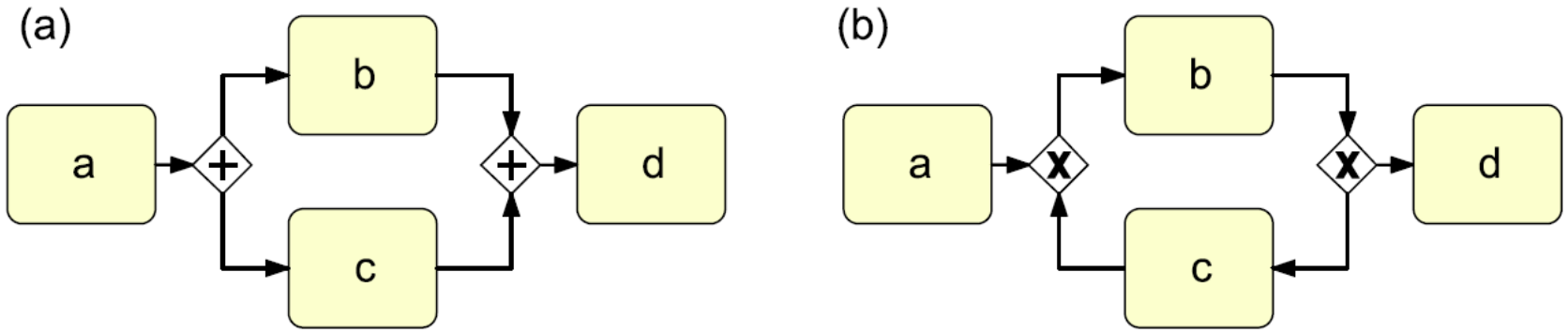
→ full model on lecture video



	a	b	c	d	e	f	g	h	i	j	k	l
a	#			#	#	#	#	#	#	#	#	#
b		#	#	#	#	#		#	#	#	#	#
c			#	→	#	#	#	#	#	#	#	#
d				#	→	#	#	#	#	#	#	#
e					#	→	#	#	#	#	#	#
f						#		#	#	#	#	#
g							#	→	#	→	#	#
h								#	→	‖	#	#
i									#	#	‖	→
j										#	→	#
k											#	→
l												#

Dumas et al. (2013)

Problems with the *α*-Algorithm



Short loops

$a \rightarrow b$, $b \rightarrow a$ is not always $(a \parallel b)$

Incompleteness and **noise**

For 10 concurrent tasks, we need 10! = 3,628,800 cases

Missing tasks

Dumas et al. (2018)

Four Quality Criteria

Approaches to fix these issues use

Fitness

– ability to replay

Simplicity (Occam's razor)

– simplest model to explain behavior, is the best model

Precision

– degree of behavior allowed by the model, but not in the logs

– do not **underfit**

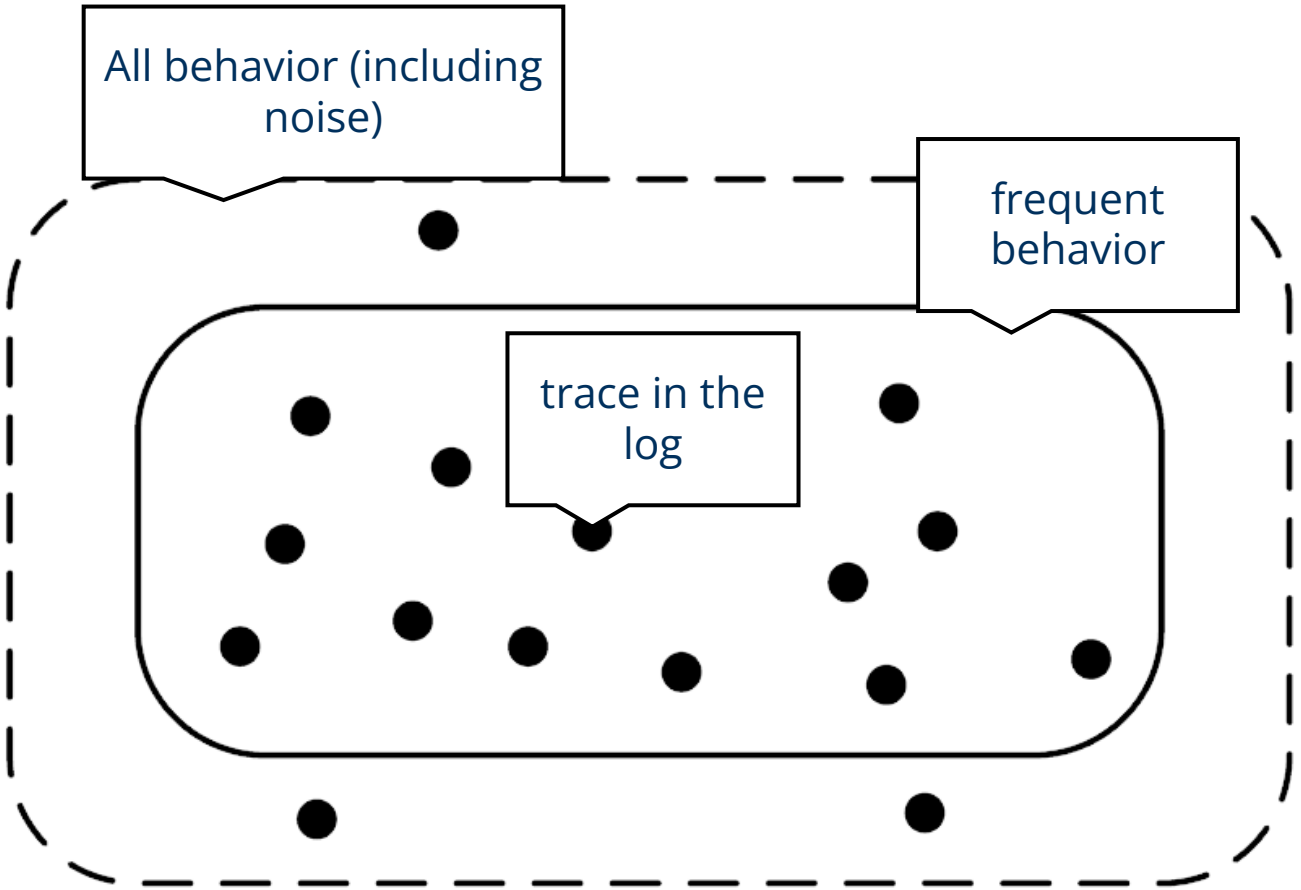
Generalization

– work with incomplete behavior

– do not **overfit**

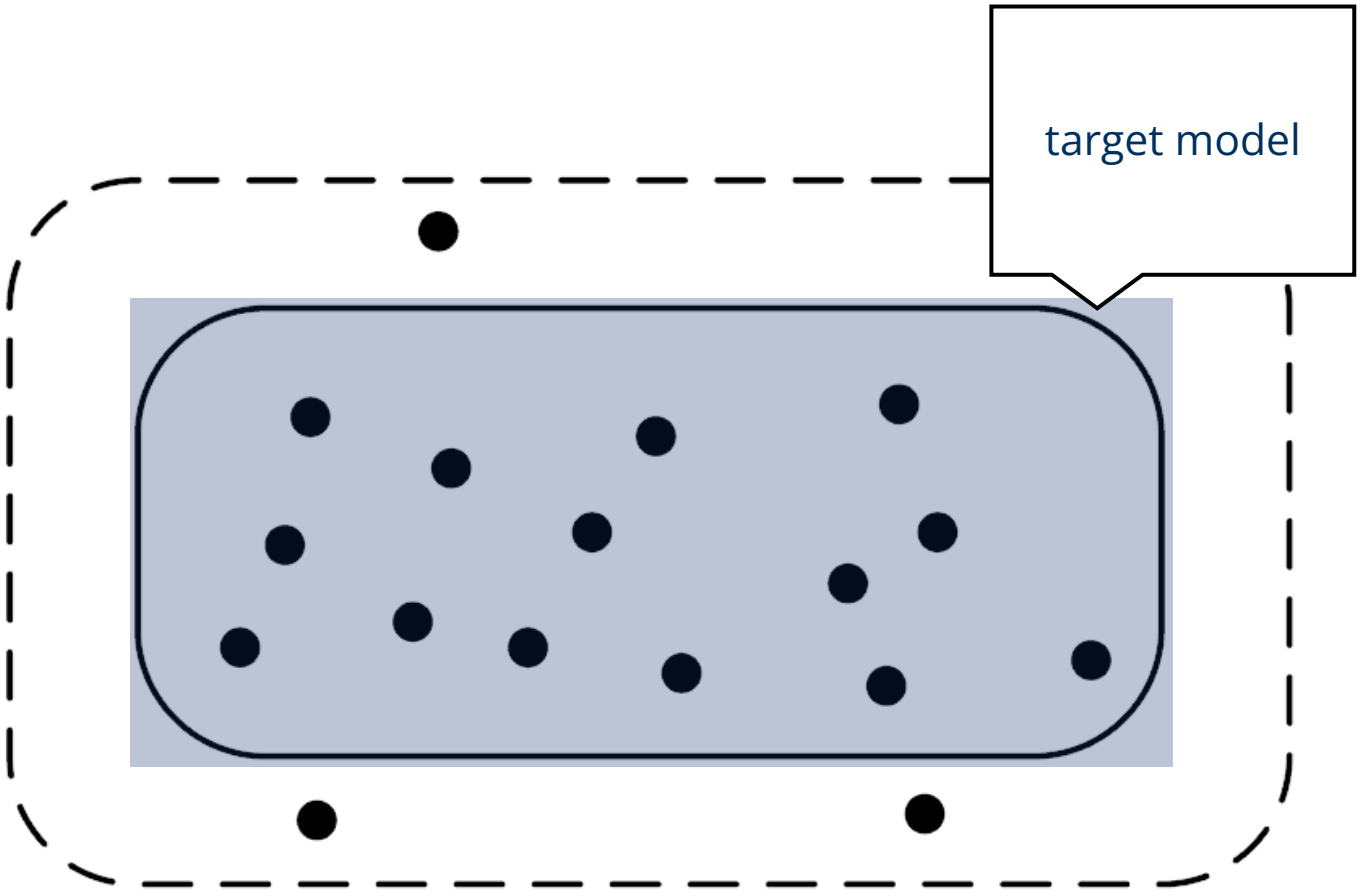
Dumas et al. (2018)

Challenges in Process Mining



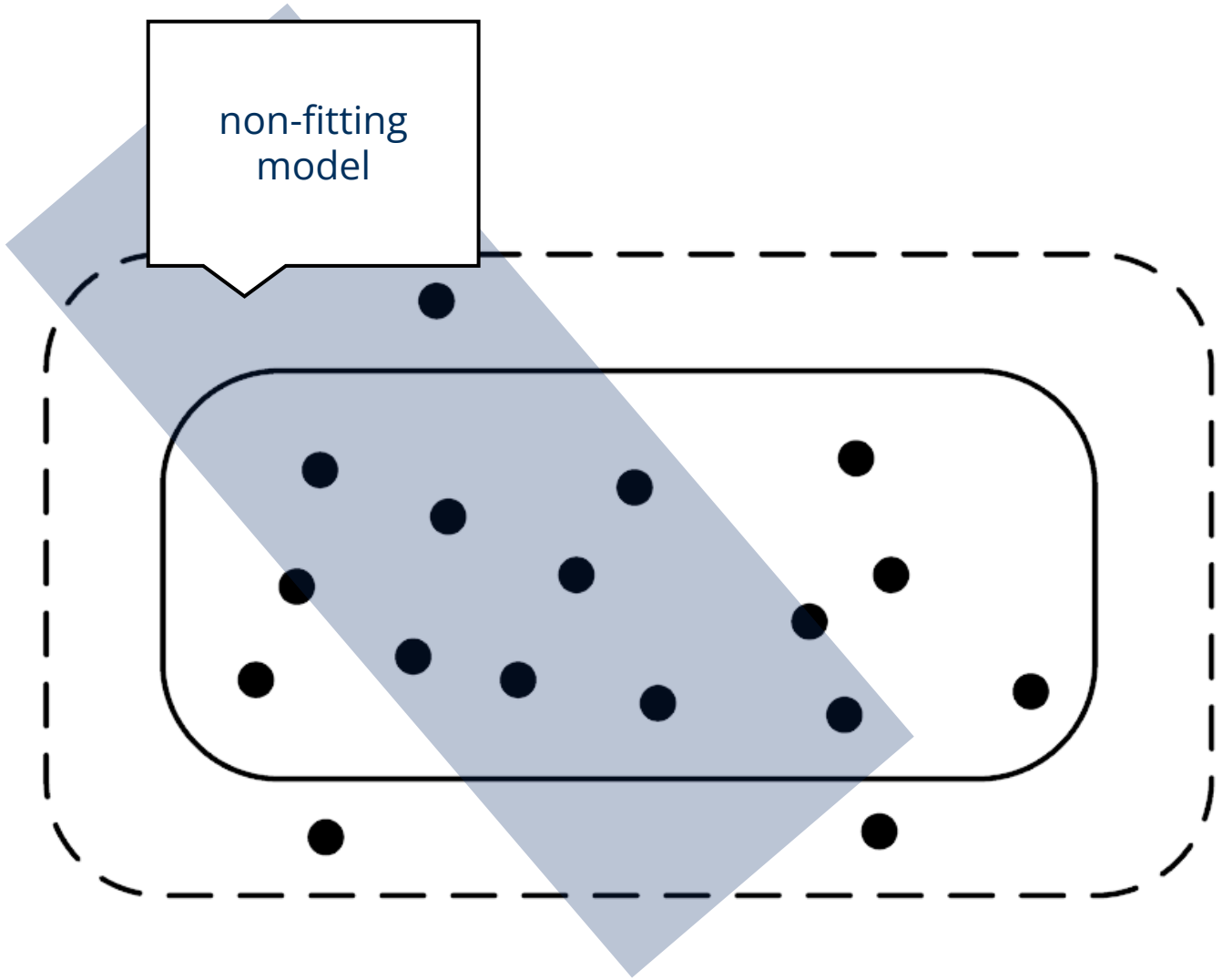
van der Aalst (2017)

Challenges in Process Mining



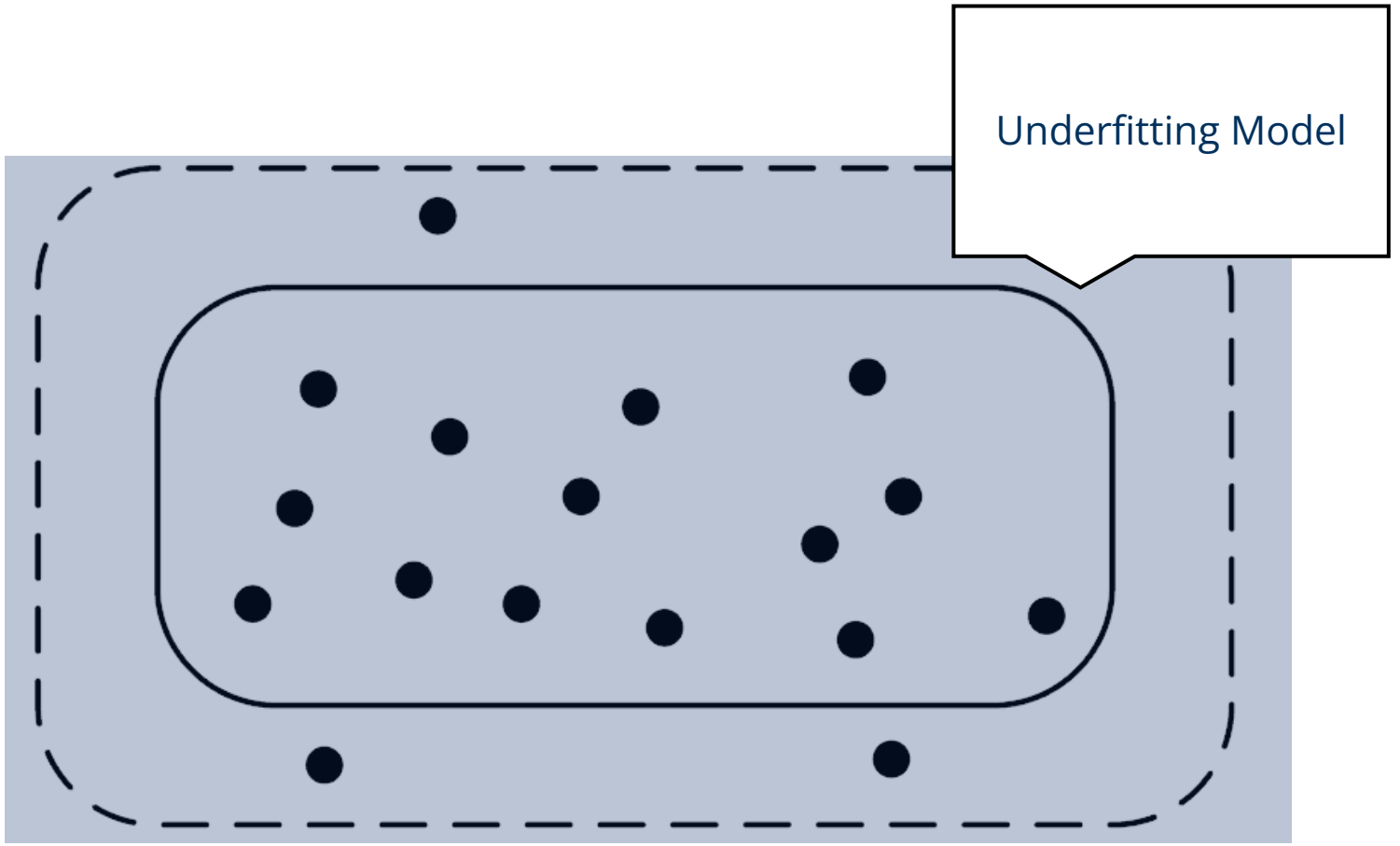
van der Aalst (2017)

Challenges in Process Mining



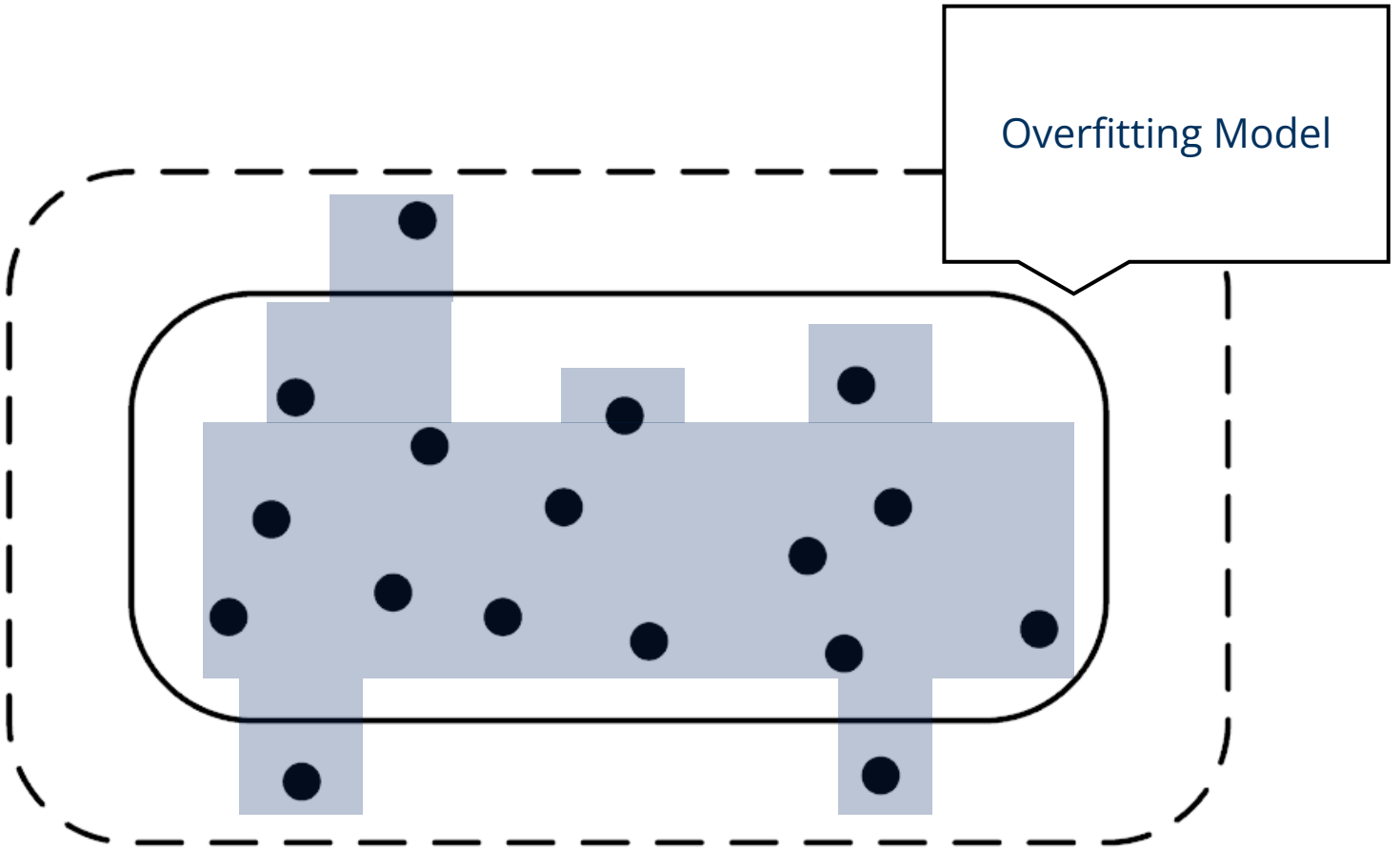
van der Aalst (2017)

Challenges in Process Mining



van der Aalst (2017)

Challenges in Process Mining



van der Aalst (2017)

Process Analytics Questions

What is the **performance** of the process?

What **is** the actual process model?

To which extent are the **rules** of the process model **followed**?

TECHNISCHE UNIVERSITÄT DRESDEN

CHAIR OF BUSINESS INFORMATION SYSTEMS
ESP. INTELLIGENT SYSTEMS AND SERVICES

DRESDEN concept

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Process Analytics

Process Conformance Checking

Conformance Checking

"Does the execution of a process **follows predefined rules or constraints** or does it **violate** them?"

Violations relate to one of the three process perspectives

- **Control flow**

Explicit constraints

Normative process model

- **Data**
- **Resources**

in isolation or in combination

Dumas et al. (2018)

Explicit Constraints

Mandatoriness

- Tasks that are **required from a control perspective**
- **Violations** can be found by searching for **traces without them**
- e.g., review tasks

Exclusiveness

- Tasks that relate to a **decision**
- **Violations** can be found by searching for **traces with both/all of them**
- e.g., accept vs. reject task

Ordering

- Tasks that have to be in a **specific succession**
- Violations can be found by searching for **traces with the activities appearing in the wrong order**
- e.g., pay -> deliver

Dumas et al. (2018)

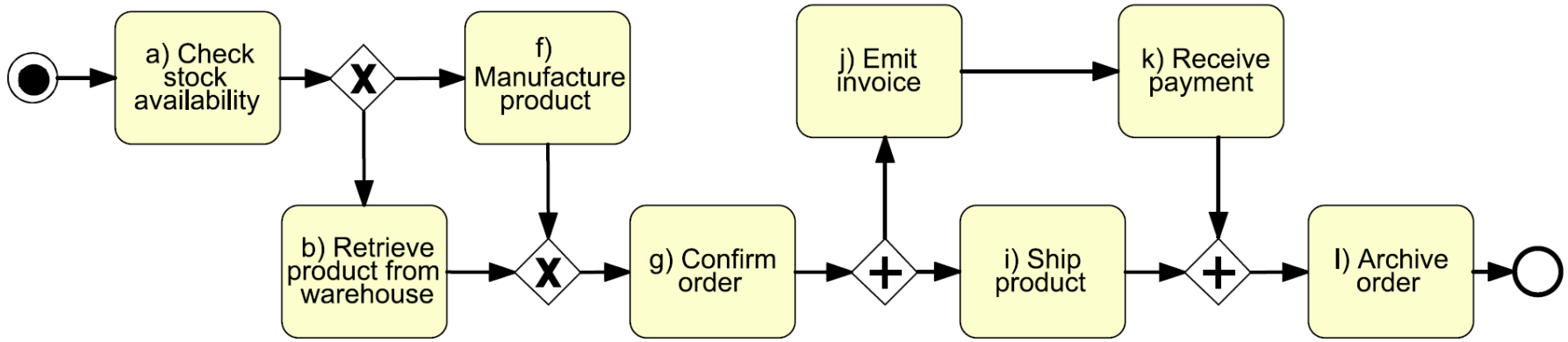
Normative Process Model

For each trace in the workflow log

replay with tokens and

record at each step whether an activity was **allowed** to be executed according to the rules of the model

Is (a, b, g, j, i, k, l) valid?

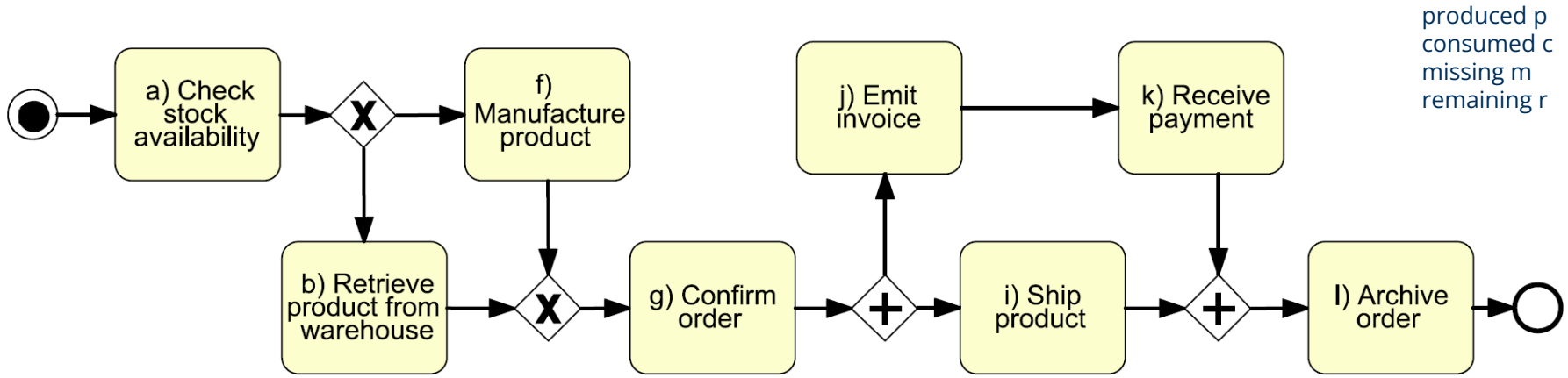


Is (a, b, i, j, k, l) valid, too? If not, what does it violate?

Dumas et al. (2018)

Tokens in a Normative Process Model

How many misplaced tokens in (a, b, i, j, k, l)?



the number of tokens that are **correctly produced *p***

the number of tokens that are **correctly consumed *c***

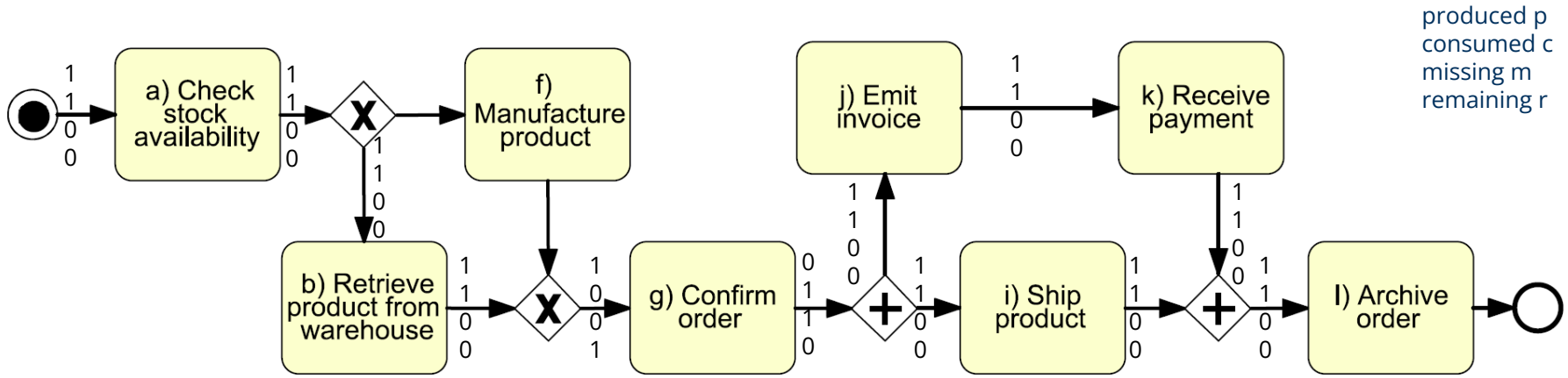
the number of tokens that are **missing for executing *m***
the next activity in the log, and

the number of tokens **remaining unconsumed *r*** after executing the final activity in the log

Dumas et al. (2018)

Tokens in a Normative Process Model

How many misplaced tokens in (a, b, i, j, k, l)?



the number of tokens that are **correctly produced *p***

the number of tokens that are **correctly consumed *c***

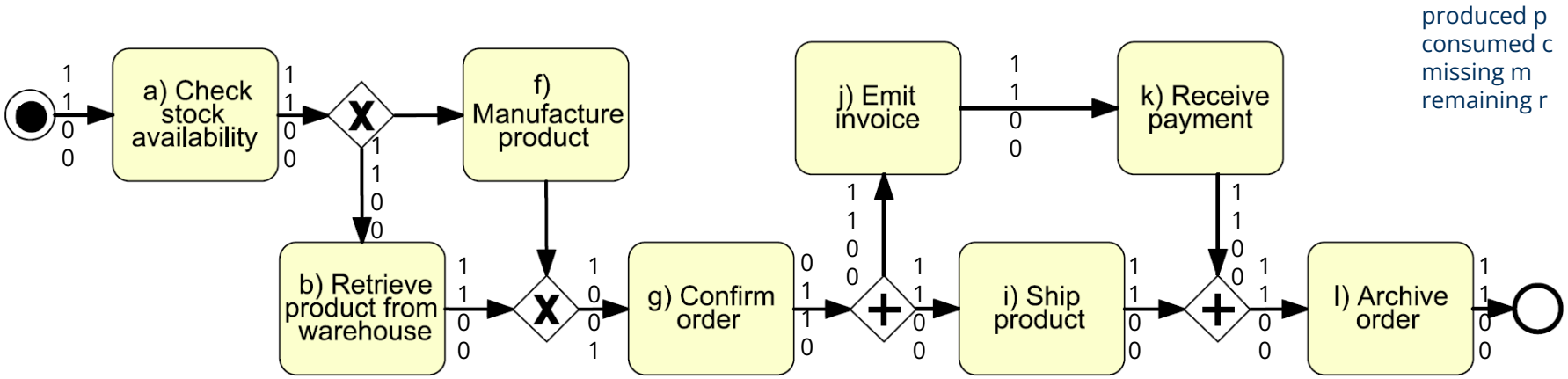
the number of tokens that are **missing for executing *m***
the next activity in the log, and

the number of tokens **remaining unconsumed *r*** after executing the final activity in the log

Dumas et al. (2018)

Tokens in a Normative Process Model

How many misplaced tokens in (a, b, i, j, k, l)?



the number of tokens that are **correctly produced *p (13)***

the number of tokens that are **correctly consumed *c (13)***

the number of tokens that are **missing for executing *m* (1)**
the next activity in the log, and

the number of tokens **remaining unconsumed *r (1)*** after executing the final activity in the log

Dumas et al. (2018)

Calculating Fitness for $n = 1$

$$fitness = \frac{1}{2}\left(1 - \frac{m}{c}\right) + \frac{1}{2}\left(1 - \frac{r}{p}\right)$$

$p = 13$

$c = 13$

$m = 1$

$r = 1$

$$fitness = \frac{1}{2}\left(1 - \frac{1}{13}\right) + \frac{1}{2}\left(1 - \frac{1}{13}\right)$$

$fitness = 0{,}9231$

Dumas et al. (2018)

Interpretation

Overall fitness

- is a **high-level measure**
- gives an idea of how accurately the process model **matches the actually observed behavior** as reflected by the set of cases
- does **not** help us to **analyze the deviations** in detail

Arc analysis

- Inspect arcs of the process model that have **missing or remaining tokens**
- Interview process participants why *x* has been omitted for some cases and identify cause and desirable behavior
- **CAVEAT:** Consider multiple opinions on this!

Dumas et al. (2018)

Conformance of Data and Resources

Participants usually require **permissions** to execute certain activities

Permissions are bundled for specific **roles**

Violations of permissions can be checked by searching for each **activity conducted** by a participant whether or not an **appropriate role or permission** existed

Specific control rules which require two different persons to approve a business transaction are called **separation of duties** constraints

These rules do not necessarily involve supervisors

Dumas et al. (2018)

Process Analytics Questions

What is the **performance** of the process?

What **is** the actual process model?

To which extent are the **rules** of the process model **followed**?

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Thank you for your attention