DATA PRODUCTS AND SERVICES (DPS)

ISABELLE LÜTHI

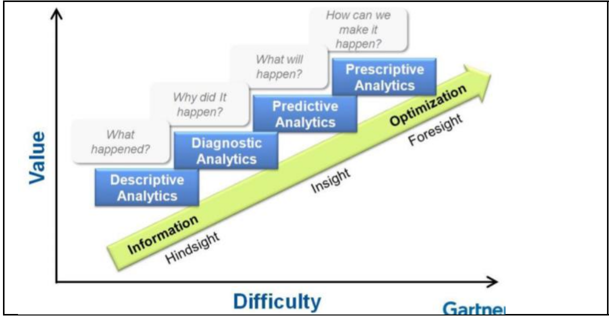
WOCHE 1 – MANAGING DATA PRODUCTS

DATA PRODUCT DEFINITION: ASSET [ONLY EXAMPLES] + CAPABILITY [ONLY AREAS] + ADDED VALUE

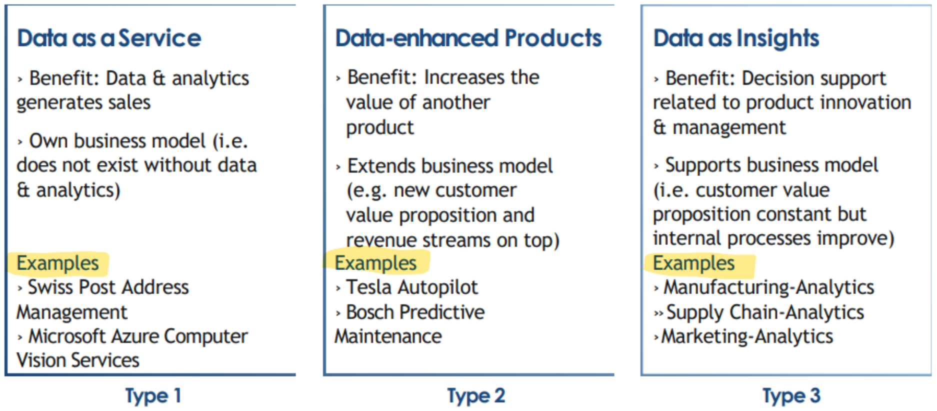
Data Assets belong to a multitude of different data types:
Transactional, demographic, times series, location, log, open source, clickstream, external web, internal text, mobile app, social media text, event, sensor, audio, video

Data related **capabilities** span a wide variety of **areas**:
Data Acquisition & Management, Data preparation, Data analysis, Visualization & Reporting, Deployment, Privacy & Security

Added Value:

	Descriptive Analytics Diagnostic Analytics Predictive Analytics Prescriptive Analytics

3 TYPES OF DATA PRODUCTS: DATA AS A SERVICE, DATA-ENHANCED PRODUCTS, DATA AS INSIGHT



Data as a Service

- Benefit: Data & analytics generates sales
- Own business model (i.e. does not exist without data & analytics)

Examples
- Swiss Post Address Management
- Microsoft Azure Computer Vision Services

Type 1

Data-enhanced Products

- Benefit: Increases the value of another product
- Extends business model (e.g. new customer value proposition and revenue streams on top)

Examples
- Tesla Autopilot
- Bosch Predictive Maintenance

Type 2

Data as Insights

- Benefit: Decision support related to product innovation & management
- Supports business model (i.e. customer value proposition constant but internal processes improve)

Examples
- Manufacturing-Analytics
- Supply Chain-Analytics
- Marketing-Analytics

Type 3

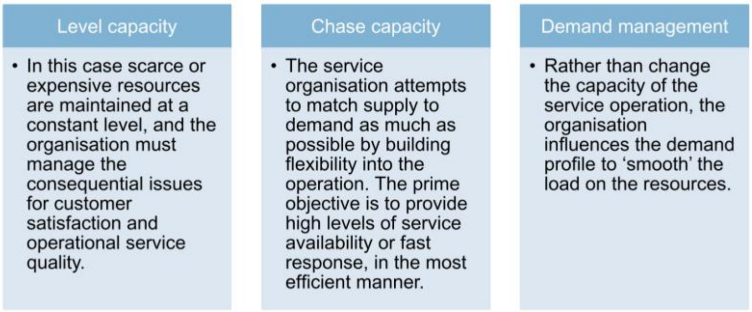
DATA PRODUCT DESIGN PROCESS: CRISP-DM [ONLY PHASES AND TASKS]

Phase	Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Generic Tasks in Phase	• Determine Business Objectives • Assess Situation • Determine Data Mining Goals • Produce Project Plan	• Collect initial Data • Describe Data • Explore Data • Verify Data Quality	• Select Data • Clean Data • Construct Data • Integrate Data • Format Data	• Select Modeling Technique • Generate Test Design • Build Model • Assess Model	• Evaluate Results • Review Process • Determine Next Steps	• Plan Deployment • Plan Monitoring and Maintenance • Produce Final Report • Review Project

WOCHE 2 – WAITING QUEUES

HOW TO DEAL WITH DEMAND VARIABILITY?

Most services are characterized by the IHIP properties:
I – Intangible: Services cannot be touched. Services are experienced!
H – Heterogeneous: customer involvement in delivery process results in variability
I – Instantaneous: Production, distribution and consumption happen at the same time
P – Perishable: services cannot be stored



Level capacity

- In this case scarce or expensive resources are maintained at a constant level, and the organisation must manage the consequential issues for customer satisfaction and operational service quality.

Chase capacity

- The service organisation attempts to match supply to demand as much as possible by building flexibility into the operation. The prime objective is to provide high levels of service availability or fast response, in the most efficient manner.

Demand management

- Rather than change the capacity of the service operation, the organisation influences the demand profile to 'smooth' the load on the resources.

WHAT ARE WAITING QUEUES? [ONLY M/M/1]

Queues = system in which objects or units (people, tasks or products) are waiting to be served or processed.

- Examples: Waiting at a counter, for a machine to become free, for a free doctor, for a green light
- Questions: How long do I have to wait? Ho likely is it, that I won't have to wait?

Characteristics of **MM1 - Queues**:

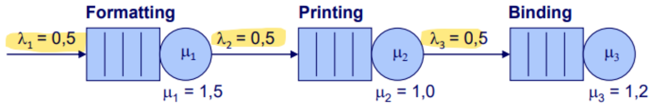
- MM1 = System with **a single server**
- **Poisson-distributed arrivals** often realistic
- **Exponentially distributed handling times** often out of touch with reality
- Very **nice analytical properties**: E. g. distribution of is Poisson.
- Very **simple formulas for performance indicators**

PERFORMANCE INDICATORS OF M/M/1 WAITING QUEUES [NOT DERIVATIONS]

➔ Application of all formulas for MM1 Queues on the formula sheet

PERFORMANCE INDICATORS OF M/M/1 WAITING QUEUES NETWORKS

«The output of a M/M/1 queue again follows a poisson distribution ($\lambda_{i+1} = \lambda_i$)»



$\rho_1 = \frac{0,5}{1,5} = 33,3\%$ $\rho_2 = \frac{0,5}{1} = 50\%$ $\rho_3 = \frac{0,5}{1,2} = 41,7\%$

$WS_1 = \frac{\rho_1}{\lambda_1(1-\rho_1)} = 1,0$ $WS_2 = \frac{\rho_2}{\lambda_2(1-\rho_2)} = 2,0$ $WS_3 = \frac{\rho_3}{\lambda_3(1-\rho_3)} = 1,4$

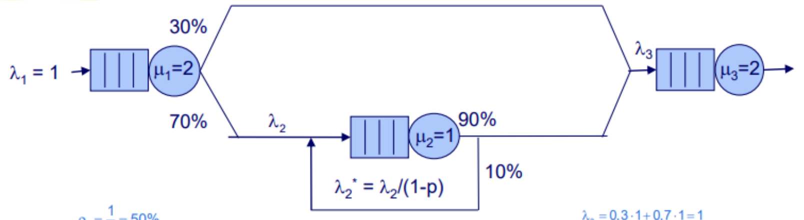
$$WS = WS_1 + WS_2 + WS_3 = 1,0 + 2,0 + 1,4 = 4,4 \text{ Stunden}$$

$$S_1 + S_2 + S_3 = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} = \frac{1}{1,5} + \frac{1}{1,0} + \frac{1}{1,2} = 2,5 \text{ Stunden}$$

Parameters
λ_i Order arrival rate [orders per hour]
μ_i Service rate of orders at station i [orders per hour]

Performance indicators
WS_i Expected waiting time of an order in station i [hours]
WS Expected waiting time in the complete network (stations 1, 2, 3) [hours]

How long does an order take on average?



$\rho_1 = \frac{1}{2} = 50\%$

$LS_1 = \frac{0,5}{1-0,5} = 1$

$WS_1 = \frac{\rho_1}{\lambda_1(1-\rho_1)} = \frac{0,5}{1(1-0,5)} = 1$

$\lambda_2 = 0,7 \cdot 1 = 0,7$

$\lambda_2^* = \frac{0,7}{1-0,1} = 0,78$

$\rho_2 = \frac{0,78}{1} = 78\%$

$LS_2 = \frac{0,78}{1-0,78} = 3,55$

$WS_2 = \frac{\rho_2}{\lambda_2^*(1-\rho_2)} = \frac{0,78}{0,78(1-0,78)} = 4,545$

$\lambda_3 = 0,3 \cdot 1 + 0,7 \cdot 1 = 1$

$\rho_3 = \frac{1}{2} = 50\%$

$LS_3 = \frac{0,5}{1-0,5} = 1$

$WS_3 = \frac{\rho_3}{\lambda_3(1-\rho_3)} = \frac{0,5}{1(1-0,5)} = 1$

$$\boldsymbol{WS = WS_1 + 0,7 * WS_2 + WS_3} = 1 + \boldsymbol{0,7 * 4,545} + 1 = \mathbf{5,182}$$

WOCHE 3 – REVENUE MANAGEMENT

WHAT IS REVENUE MANAGEMENT?

“Revenue Management embraces techniques to allocate limited resources to different types of customers at different prices in order to maximize company revenues.”

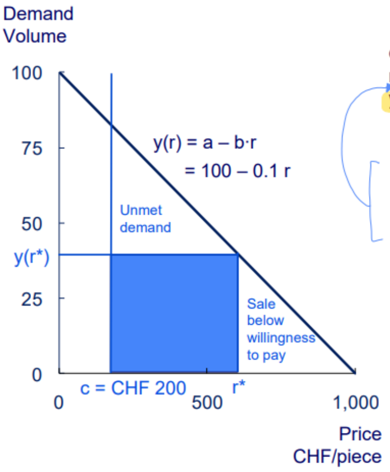
Alternative terms: Yield Management, Prive optimization, Demand Management

Revenue Management is most effective when:

- Product has limited shelf live and can be sold in advance
- Capacity is limited, can only be increased with lots of effort
- Market / customers can be divided into segments
- Variable costs are low
- Demant fluctuates over time, is unknown at the time of the decision
- Prices can be adjusted

Fields of application for revenue management: Airlines, Hotels, Car rental, freight, energy, health services, …

HOW TO DIFFERENTIATE PRICES [NOT DERIVATIONS]



Notation

c = variable cost

r = price

y(r) = a – b·r = linear demand function

Goal:

Determine the price r*, so that the profit is maximized

a= (theoretical) demand at r=0

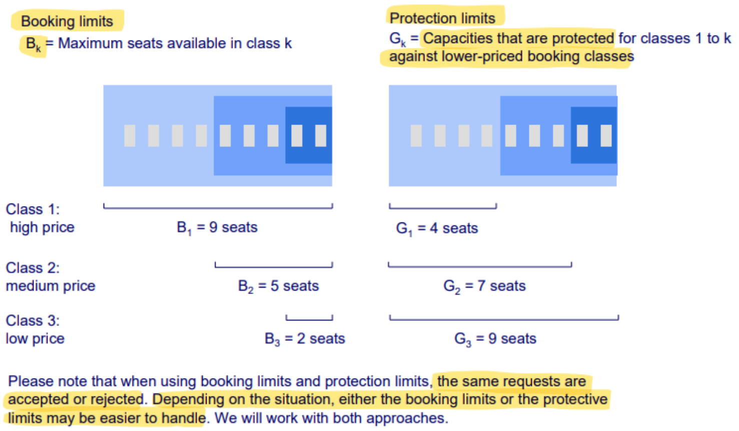
b= negative slope of demand function

r_0= prohibitive price

➔ Formulas on formula sheet

WOCHE 4 – REVENUE MANAGEMENT 2

HOW TO IMPLEMENT BOOKING CONTROL?



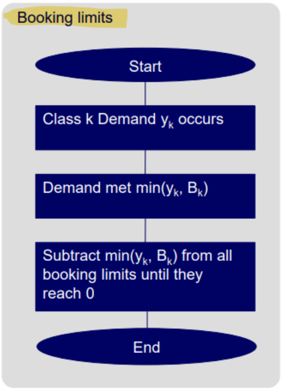
Booking limits

B_k = Maximum seats available in class k

Protection limits

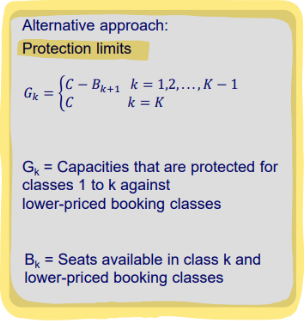
G_k = Capacities that are protected for classes 1 to k against lower-priced booking classes

Please note that when using booking limits and protection limits, the same requests are accepted or rejected. Depending on the situation, either the booking limits or the protective limits may be easier to handle. We will work with both approaches.



Booking limits

Start
Class k Demand y_k occurs
Demand met min(y_k, B_k)
Subtract min(y_k, B_k) from all booking limits until they reach 0
End



Alternative approach:

Protection limits

$$G_k = \begin{cases} C - B_{k+1} & k = 1,2,\ldots,K-1 \\ C & k = K \end{cases}$$

G_k = Capacities that are protected for classes 1 to k against lower-priced booking classes

B_k = Seats available in class k and lower-priced booking classes

HOW TO INTERPRET THE CUMULATIVE DISTRIBUTION FUNCTION FOR NORMALLY DISTRIBUTED VARIABLES?

- ➔ Formula sheet

HOW TO DETERMINE THE BOOKING LIMITS IN THE CASE OF TWO BOOKING CLASSES?

- ➔ Formula sheet

WOCHE 5 – EXPECTED MARGINAL SEAT REVENUE

HOW DO YOU DETERMINE BOOKING LIMITS FÜR MORE THAN TWO BOOKING CLASSES BY AGREGATING PROTECTION LIMITS?

= Expected Marginal Seat Revenue – a = ESMR-a

Assumptions:	Solution by using heuristics:
• K Booking Classes • Maximal Capacity C given • Classes are determined, so that r1 > r2 applies • Demand is met according to the sequential order of the booking classes • Nested protection limits • At each level, a two-class problem is solved	• ESMR-a ○ Expected marginal seat revenue – Version a ○ **Aggregated protection limits** • ESMR-b ○ Expected marginal seat revenue – Version b ○ **Aggregated demand**

HOW DO YOU DETERMINE BOOKING LIMITS FOR MORE THAN TWO BOOKING CLASSES BY AGGREGATING DEMAND?

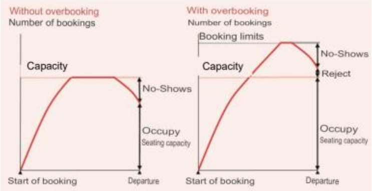
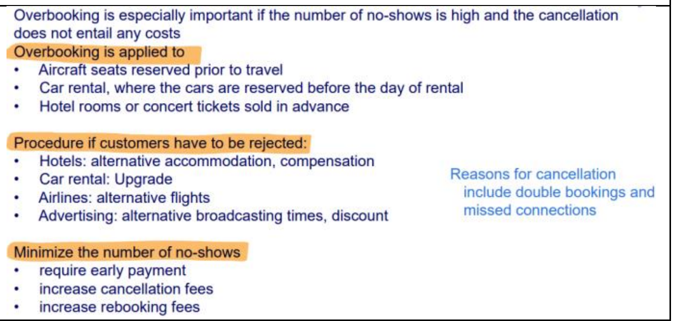
= Expected Marginal Seat Revenue – b (ESMR-b) → Formula sheet

Summary:

- An optional solution requires dynamic programming
- Both approaches are heuristic
- **ESMR-a aggregates the protection limits. ESMR-b aggregates the demand**
- Calculating the weighted average price is a critical assumption
- In practice, ESMR-b is more widespread
- Experimental studies show that neither method dominates the other

WOCHE 6A – OVERBOOKING

(ALL THINGS ABOUT OVERBOOKING)

Booking volume over time with and without overbooking	Managing Overbooking
	Overbooking is especially important if the number of no-shows is high and the cancellation does not entail any costs Overbooking is applied to • Aircraft seats reserved prior to travel • Car rental, where the cars are reserved before the day of rental • Hotel rooms or concert tickets sold in advance Procedure if customers have to be rejected: • Hotels: alternative accommodation, compensation • Car rental: Upgrade • Airlines: alternative flights • Advertising: alternative broadcasting times, discount Reasons for cancellation include double bookings and missed connections Minimize the number of no-shows • require early payment • increase cancellation fees • increase rebooking fees

- → Formula sheet

WOCHE 6B – SERVICE PROCESS MODELING

WHAT IS THE DATABASE FOR PROCESS MINING?

Database: Event Log

- An event log consists of events that relate to a specific case, an activity, and a point in time.
- A case describes an individual instance of the sequence of activities.
- A simplified event log is a multiset of traces (i.e., the same trace can occur multiple times).
- A trace is the activity sequence of a case (i.e. we only look at the order of the activities of each case).

$$L_1 = [\langle a,b,c,d\rangle^3, \langle a,c,b,d\rangle^2, \langle a,e,d\rangle]$$

Log Log ID Trace number of occurrences Activity

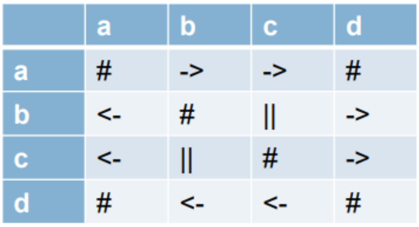
WHAT IS A PROCESS FOOTPRINT?

	a	b	c	d
a	#	->	->	#
b	<-	#	||	->
c	<-	||	#	->
d	#	<-	<-	#

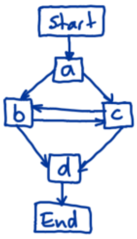
# no relation between a and d
| | parallel: b/c are followed by each other
-> a is directly followed by b
<- b follows after a

WOCHE 7 – DIRECTLY FOLLOWS GRAPH AND PETRI NETS

HOW DO YOU CREATE A DIRECT FOLLOWS GRAPH?



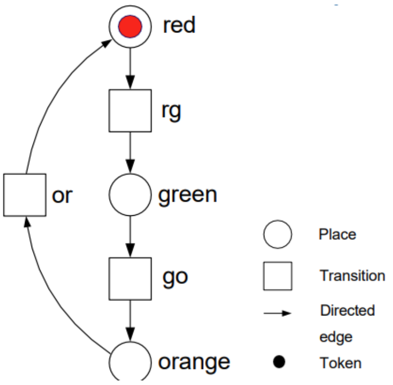
	a	b	c	d
a	#	->	->	#
b	<-	#	||	->
c	<-	||	#	->
d	#	<-	<-	#



There needs to be a “start” and an “end” point!

WHAT IS A PETRI NET?

- The Petri network is static and consists of places and transitions.
- Places mark potential states of a process, transitions model the dynamics between states.
- Places and transitions are connected by directed edges, where an edge connects exactly one place to a transition or vice versa.
- Places can contain tokens.
- Tokens indicate the state of a process.
- Transitions generate and/or consume tokens



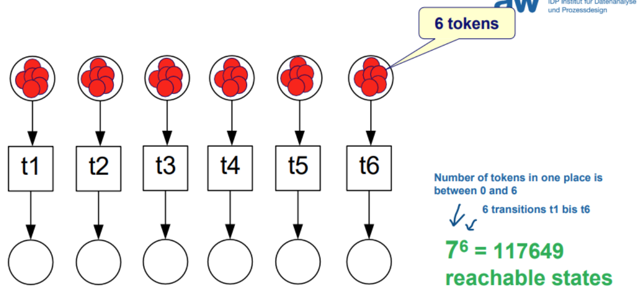
WHAT IS THE MARKING OF A PETRI NET?

- A **Marking** is the distribution of tokens across places in the petri net.
- **Initial marking**: Marking of the petri net prior to firing transitions.
- **Reachable marking**: Marking that is reachable from the initial marking via firing transitions.
- **Unreachable marking**: Marking that is nor reachable from the initial marking (Initial marking = 1 Token, Tokens can never be doubled in net, unreachable marking:3 Tokens)
- **Final marking**: Marking of the Petri network after all possible transitions have fired. (**Note: there may be none, one or multiple**.)

HOW TO FIRE MULTIPLE TRANSITIONS?

- Firing is atomic (consumption & production at the same time)
- If there are several identical transitions (e.g. 3), which are independent of each other, and each transition has 4 reachable states, then the total number of reachable states is 3**4 = 27.
- The same thing can be calculated if there are amounts of tokens involved (picture below).

State explosion



$7^6 = 117649$ reachable states

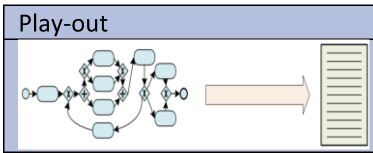
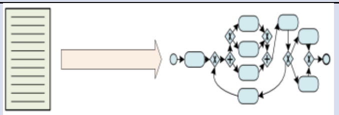
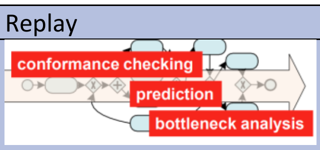
HOW DO YOU CALCULATE DIRECT FOLLOWS GRAPHS IN PYTHON?

➔ Jupiter Notebooks

WOCHE 8 – PROCESS DISCOVERY

WHAT ARE OBJECTIVES PROCESS MINING?

Process discovery, conformance checking, performance prediction based on event data

Play-out	Play-in	Replay
		
Simulation, Workflow automation	Process discovery	Aligning modeled/discovered and observed behavior: **The most important of process mining!** Confrontation between model and reality.

HOW DOES THE ALPHA ALGORITHM FOR PROCESS DISCOVERY WORK?

➔ Formula sheet

WHAT ARE THE LIMITATIONS OF THE ALPHA ALGORITHM?

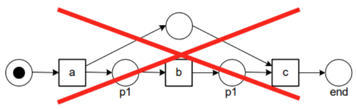
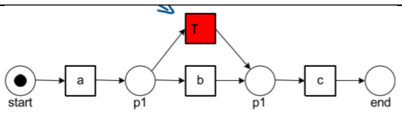
Implicit places: der Algorithmus kann diese im Workflow-Prozess nicht erkennen. Implizite Stellen sind solche, die notwendig sind, um die Korrektheit des Prozessmodells zu gewährleisten, aber nicht direkt aus den Log-Dateien abgeleitet werden können.

Loops:

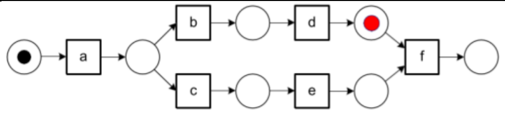
- **length 1** (e.g. b||b) are displayed separately (= inaccuracy / incomplete)
- **length 2** (e.g. b||c) c will be left out, left out transition can always fire (= inaccuracy / incomplete)

Non-local dependencies L=[<a,c,d> , <b,c,e>] in one petri net makes <a,c,e> possible, even though that trace doesn't exist. One would need two separate c-transitions, which isn't possible with alpha algorithm. Nicht effektiv bei Erkennung von nicht-lokalen Abhängigkeiten (Beziehungen zwischen Aktivitäten, die nicht direkt aufeinander folgen, aber dennoch voneinander abhängig sind).

Representational Bias L=[<a,b,c>,<a,c>] cannot be displayed without a SILENT TRANSITION (=TAU)
Der Algorithmus neigt dazu, bestimmte Muster oder Strukturen in den Daten zu bevorzugen und andere zu vernachlässigen. Dies kann zu einer Verzerrung im erstellten Prozessmodell führen, indem bestimmte Aspekte des tatsächlichen Prozesses unter- oder überpräsentiert werden.

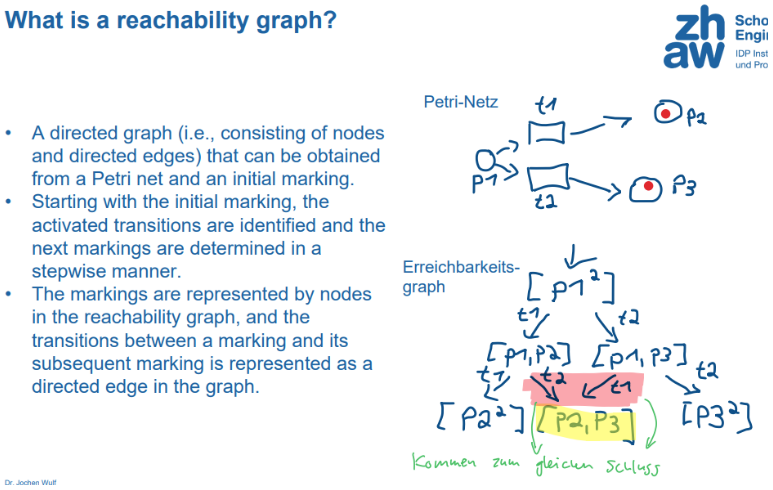
Wrong	Correct
 Transition a would produce 2 tokens... incorrect!	 This way, transition b becomes optional and there is still only 1 token in the net.

Soundness (Petri net with deadlocks): net is correct, but f can't fire = deadlock
Der Alpha-Algorithmus kann nicht immer garantieren, dass das generierte Prozessmodell korrekt ist, das heisst, dass es frei von Fehlern wie Deadlocks oder unerreichbaren Pfaden ist.

$L = [\langle a, b, d, e, f\rangle^{10}, \langle a, c, e, d, f\rangle^{10}]$	

WOCHE 9 – PLAYOUT A PROESS MODEL

HOW TO CONSTURCT A REACHABILITY GRAPH?



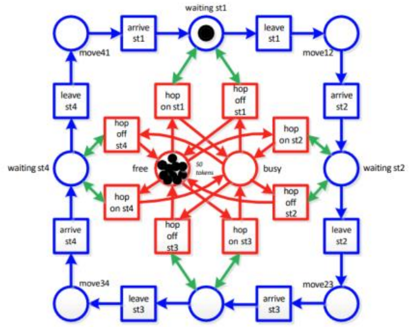
What is a reachability graph?

- A directed graph (i.e., consisting of nodes and directed edges) that can be obtained from a Petri net and an initial marking.
- Starting with the initial marking, the activated transitions are identified and the next markings are determined in a stepwise manner.
- The markings are represented by nodes in the reachability graph, and the transitions between a marking and its subsequent marking is represented as a directed edge in the graph.

Dr. Jochen Wulf

Important! **A Marking can only be displayed once** on a reachability graph! ([p2, p3])

HOW TO MODEL COMPLEY PROCESS MODELS WITH PETRI NETS?

	Separate Processes if possible, combine them step by step. Train station-Example 1. State of Train 2. Passengers 3. Combination between Train & Passengers Reachability graph is already too large to draw manually!

HOW TO GENERATE PETRI NETS, REACHABILITY GRAPHS AND PLAYOUTS IN PYTHON?

➔ Jupiter Notebooks

WOCHE 10 – CONDORMANCE CHECKING

WHY AN HOW TO ECALUATE THE CONFORMANCE BETWEEN TRACES AND PROCESS MODELS?

Fitness	Simplicity	Precision
– perfect fitness if all traces in the log can be replayed by the model from beginning to end. – Case level: fraction of traces in the log that can be fully replayed – Event level: fraction of events in the log that are indeed possible according to the model	– The simplest model that can explain the behavior seen in the log, is the best model. – Complexity of the model could be defined by the number of nodes and arcs.	– A model is precise if it does not allow for “too much” behavior (“underfitting”). – fraction of traces in the log to all traces that can be played out from the process model

HOW TO USE FOOTPRINTS TO MEASURE CONFORMANCE?

Count of footprint differences

→ Formula sheet

Limitations of footprint-based conformance: frequencies are not used, behavior is only considered indirectly, aims to capture fitness and precision in single metric

HOW TO MEASURE TRACE-LEVEL FITNESS WITH TOKEN-BASED REPLAY?

Consumed, produced, missing, remaining tokens

→ Formula sheet

CALCULATION AND PROBLEM WITH LOG-LEVEL TOKEN-BASED REPLAY?

Aggregation of tokens, tokens flooding, local decision making

→ Formula sheet

WOCHE 11 – ORGANIZATIONAL MINING

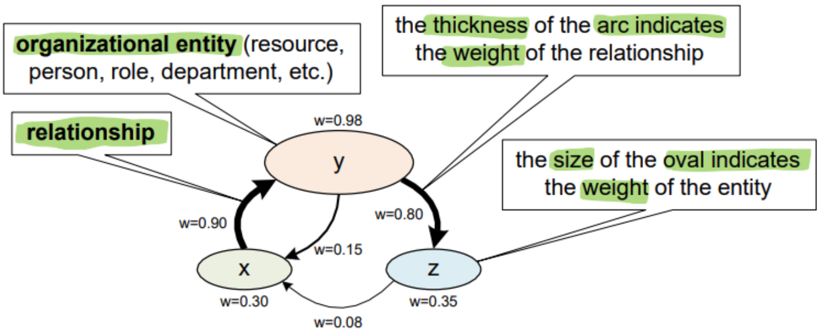
HOW TO CONSTRUCT A RESOURCE ACTIVITY MATRIX?

Volume of an activity carried out by a resource in the event log

- Challenge 1: Identification of roles from data
- Normalize resource-activity matrix: divide absolute frequencies by number of cases
- Resource-activity matrix provides basic insight into "who is doing what"

HOW TO INTERPRET A SOCIAL NETWORK?

Nodes represent social entities, edges relationships, edge thickness relationship strength


organizational entity (resource, person, role, department, etc.)
relationship
the thickness of the arc indicates the weight of the relationship
the size of the oval indicates the weight of the entity
w=0.98
y
w=0.90
w=0.80
w=0.15
x
z
w=0.30
w=0.35
w=0.08

HOW TO MEASURE RESOURCE SIMILARITY [PERSON EXCLUDED]?

→ Formula sheet

HOW TO IDENTIFY RESOURCE ROLES?

→ Formula sheet (clusters of similar resources)

HOW TO IDENTIFY TEAMS OF RESOURCES?

→ Formula sheet

WOCHE 12 – ORGANIZATIONAL MINING 2

HOW DO YOU CALCULATE THE HAND-OVER MATRIX?

Challenge: identification of potential resource bottlenecks

Identification of handovers

- ➔ Formula sheet

HOW DO YOU CALCULATE THE DEGREE CENTRALITY OF RESOURCES BASED ON THE HAND OVER MATRIX?

HOW DO YOU CALCULATE THE BETWEENNESS CENTRALITY OF A NODE?

HOW DO YOU CALCULATE THE BETWEENNESS CENTRALITY OF RESOURCE BASED ON THE HAND-OVER MATRIX?

- ➔ All three above: Formula sheet!

WHAT IS THE OBJECTIVE OF DECISION POINT MINING?

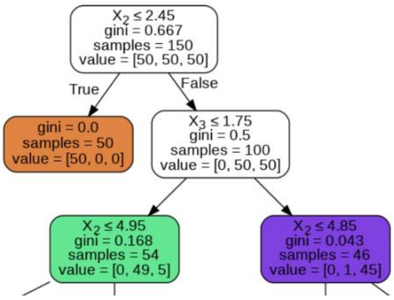
Identification of guards for the routing of cases

Guards ensure that the right path is taken (assuming that all cases have a data attribute with a value of blue or red: if red, then B+C, if blue, then E)

HOW DO YOU CONSTRUCT A CLASSIFICATION PROBLEM FOR A DECISION POINT?

Response Variable Activity, Predictor Variables Context of Decision Point

HOW DO YOU CALCULATE A DECISION TREE?



Selection of the split criterion with the maximum reduction of the Gini Impurity Mass → Formula sheet

Important: all info within one node (X, gini, sample, value, class)

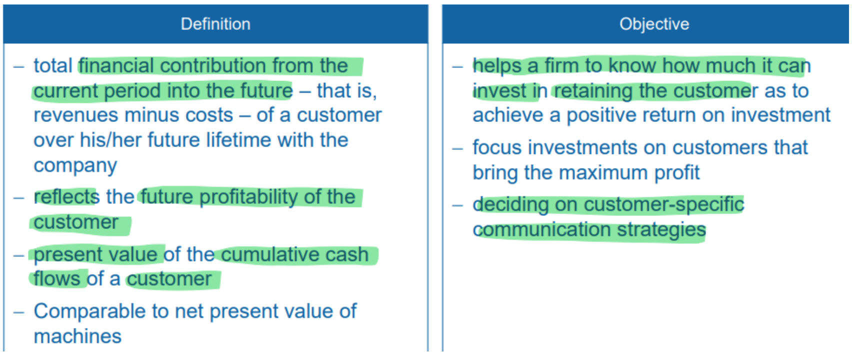
HOW DO YOU PERFORM A DECISION POINT ANALYSIS?

Identification of guards from the decision tree → Formula sheet

WOCHE 13 – CUSTOMER LIFETIME VALUE

WHAT IS THE CUSTOMER LIFETIME VALUE (CLV) AND WHY IS IT IMPORTANT?

Present value of the cumulative cash flows of a customer over complete lifetime



Definition	Objective
– total financial contribution from the current period into the future – that is, revenues minus costs – of a customer over his/her future lifetime with the company – reflects the future profitability of the customer – present value of the cumulative cash flows of a customer – Comparable to net present value of machines	– helps a firm to know how much it can invest in retaining the customer as to achieve a positive return on investment – focus investments on customers that bring the maximum profit – deciding on customer-specific communication strategies

WHAT IS THE BASIC APPROACH TO MEASURING CLV?

Present value of recurring revenues and cost → Formula sheet

HOW TO INCORPORATE CUSTOMER RETENTION, INFINITE TIME HORIZONS AND CUSTOMER AQUISITION COST?

- → Formula sheet

HOW TO CONSIDER FREQUENCY AND RECENCY EFFECTS IN CLV CALCULATIONS?

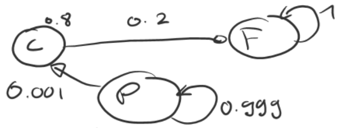
Probability of being active in subsequent time periods → Formula sheet

HOW TO MODEL STATE-DEPENDENT PROBABILITIES SUCH AS LOWER CUSTOMER CHURN OF LOYAL CUSTOMERS? [NOT INFINITE PERIODS]

Markov Chain Approach to CLV

- → Formula sheet

- A Markov chain is a stochastic process. (e.g. purchase, customer retention)
- The aim of applying Markov chains is to give probabilities for future events to occur. (e.g. Probability of Purchase -> Expected Profit)
- A Markov chain is defined by the fact that knowledge of only a limited prehistory (in this case: only the current state of the system) makes it possible to predict future development.
- The entire prehistory of the process is not relevant for prediction.
- Modeling with discrete time (e.g. months or years)
- Modeling with finite state set (e.g. customer status as potential, active or lost customer)



C = customer
P = potential customer
F = lost customer

$$\text{From}\ \begin{array}{c|ccc} & \overset{\text{To}}{P} & C & F \\ \hline P & 0.999 & 0.001 & 0 \\ C & 0 & 0.8 & 0.2 \\ F & 0 & 0 & 1 \end{array}\Bigg\} = P$$

"Transition Matrix"

$$\begin{bmatrix} -A \\ NC-M \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 30 \\ 0 \end{bmatrix} = \text{Revenue-Vector}$$

PYTHON PROGRAMMING

- ○ YOU ARE ABLE TO INTERPRET SOFTWARE CODE
- ○ YOU CAN PROVIDE THE OUTPUT OF SOFTWARE CODE
- ○ YOU WILL NOT BE ASKED TO WRITE CODE OR TO DETECT ERRORS
- ○ ONLY THE CODE IN THE JUPITER NOTEBOOKS WILL BE RELEVANT
- ○ YOU CAN BRING PRINOUTS OF THE GITHUB NOTEBOOKS [NO ANNOTATIONS]